STATISTIQUES

FRANCE IMAGE LOGICIEL

SOMMAIRE

PRÉSENTATION

L'idée de « statistiques » est sans doute d'origine immémoriale : le décompte est en effet l'une des plus vieilles activités exercées par l'être humain. Elle s'est complexifiée avec l'évolution de celui-ci : simple énumération d'hommes ou d'objets au départ, comme des listes de butin qu'on trouve chez Homère, elle progresse assez tardivement (mais radicalement), dès lors qu'on se met à analyser et non plus simplement énumérer les collections d'objets décomptés. La « population » étudiée est alors divisée en sous-groupes dont on évalue l'importance relative, et dont les individus se caractérisent par une propriété commune : lieu d'habitation, activité... Ce niveau est déjà atteint dans les recensements commandés par les empereurs romains au début de notre ère.

Mais ce n'est qu'avec les progrès effectués par la science mathématique à partir de la Renaissance que le concept moderne de « statistiques » a pu pleinement se réaliser ; l'impossibilité de constituer une information complète sur une population dès lors que celle-ci devient par trop importante, est palliée par les lois arithmétiques qui garantissent la validité des résultats obtenus par l'examen d'une vaste collection d'objets dont on n'étudie qu'une fraction, (un échantillon), sélectionnée selon des critères bien définis.

Parallèlement à ce perfectionnement des techniques mathématiques d'étude des populations, les statistiques ont vu s'étendre leur domaine d'application : limitées originellement au champ socio-politique (évaluation des biens, gouvernement), elles sont devenues au XIX^{e} siècle un moyen d'investigation pour la science physique.

L'informatique, pour sa part, si elle n'a pas ce privilège de l'ancienneté, peut néanmoins revendiquer une place égale dans le domaine des techniques modernes. L'association de ces deux techniques, dans le logiciel STATISTIQUES que nous vous présentons, constitue un nouveau progrès pour chacune d'elles. L'informatisation des techniques statistiques permet en effet d'automatiser tous les calculs qu'il faut effectuer pour l'étude d'une population ; ceci permet de ne plus exiger de l'utilisateur de cette technique, des compétences particulières en mathématiques. Votre seule tâche, avec le logiciel STATISTIQUES, consiste en effet à introduire les données concernant la population étudiée, suivant des critères que vous fournit le logiciel : le programme vous propose ensuite plusieurs types de calculs à effectuer sur la base de ces données.

Toutefois, la maîtrise complète de l'outil efficace qu'est STATISTIQUES interdit une méconnaissance totale des procédés mathématiques mis en œuvre par le logiciel pour effectuer l'étude des données. C'est pourquoi cette notice comprend deux parties : un « manuel utilisateur » qui vous permet de maîtriser les opérations nécessaires à l'exécution de ce pro-

gramme, et un « précis de statistiques », qui vous sera indispensable d'une part pour constituer correctement vos échantillons, d'autre part pour interpréter les résultats des calculs effectués par le logiciel.

Le MANUEL UTILISATEUR vous renverra au PRECIS DE STATISTIQUES dès qu'un problème théorique se posera.

Cependant, si vous n'avez que peu de notions de statistiques, il est préférable de vous reporter au précis avant toute utilisation du logiciel afin de rafraîchir votre mémoire.
Enfin, les annexes disposées à la fin de la notice constituent des compléments utiles à la lecture du précis ou du manuel ; les EXEMPLES illustrent les notions statistiques et vous proposent en même temps des modèles d'utilisation du logiciel. Les définitions du GLOSSAIRE faciliteront votre lecture de l'une ou l'autre des deux parties de cette notice.

SOMMAIRE

N.B. — *Les définitions des termes de la statistique font l'objet de normes (ISO 3534 1977* [*E/F*] et NF X 06.03).

I. — GÉNÉRALITÉS

1. POPULATION - ÉCHANTILLON

Aujourd'hui, l'outil statistique, indispensable à l'exploitation et à l'interprétation des résultats expérimentaux, présente des intérêts économiques et scientifiques qui justifient ses utilisations nombreuses (gestion, laboratoire, contrôle industriel, etc.).

Du point de vue scientifique, les méthodes statistiques permettent la formulation de lois en dégageant les facteurs prépondérants d'un phénomène, en éliminant les résultats aberrants.

Du point de vue économique, elles permettent d'optimiser un programme d'essais par la détermination du nombre minimal d'expériences nécessaires. L'étude d'un phénomène expérimental peut également permettre de réduire considérablement l'expérimentation d'un phénomène analogue.

Les statistiques, entièrement gouvernées par les lois mathématiques, permettent donc d'optimiser les résultats en fonction des effectifs et des modalités de l'expérimentation. Elles peuvent être définies comme l'ensemble des techniques qui permettent d'étudier une vaste collection d'éléments en se limitant à une fraction bien choisie de cette collection.

« Population » désignera la collection d'éléments que l'on envisage d'étudier et « échantillon », la fraction d'éléments qui va permettre l'étude d'un caractère, qualitatif ou quantitatif, de la population.

Un « prélèvement » permettra de constituer l'échantillon. La « taille de l'échantillon » ou son « effectif » est le nombre d'éléments de cet échantillon.

2. DISTRIBUTION STATISTIQUE

L'échantillon prélevé dans la population doit être le plus possible représentatif de celle-ci. Pour ce faire, le prélèvement doit obéir à certaines règles. La plus importante est que le prélèvement soit tel qu'il offre à chaque élément de la population une chance égale d'être prélevé.
Ce prélèvement est appelé prélèvement au hasard. Ceci est une opération difficile.
Citons un exemple de cette simulation du hasard : aux différents objets d'un lot, on attribue un numéro. On effectue ensuite le prélèvement des objets dont les numéros correspondent aux nombres donnés par une table de nombres au hasard. (Une table de nombres au hasard présente des lignes successives de chiffres obtenus au hasard, par des procédés que nous ne développerons pas ici.)

a) *Mise en ordre des résultats.*

En possession d'un échantillon représentatif d'une population donnée, on détermine sur chacun de ses éléments la valeur de la caractéristique X étudiée.

Par exemple : Notes données par vingt professeurs différents à une même dissertation de philosophie.
(Caractéristique X = note)

Plusieurs possibilités peuvent se présenter :

■ La série des résultats est laissée intacte. Elle peut être exploitable telle quelle.
Cette série est constituée de données que nous appellerons *données indépendantes* ou *données appariées* suivant le cas (voir le chapitre I. 3.a pour plus de détails).

■ Les résultats de la série sont mis en ordre. A chaque réalisation x_i de la variable X, on associe l'effectif n_i qui la caractérise. Les valeurs de la variable étant rangées dans l'ordre croissant, le tableau obtenu constitue la donnée de la distribution expérimentale observée.

Voici la distribution de l'exemple cité précédemment :

Notes (x_i)	7	8	9	10	11	12	13
Effectifs (n_i)	1	5	9	2	1	1	1

Cette série est constituée d'éléments que nous appellerons *classes-variable discrète*.

Il faut remarquer cependant que la représentation de la série statistique par un tableau (x_i, n_i) est uniquement une façon simple et claire de la résumer et qu'en réalité une distribution comprend autant de termes qu'il y a d'éléments dans l'échantillon, certains de ces termes pouvant être égaux.

Ce que nous regroupons sous le vocable *classes-variable discrète* correspond en fait à deux choses :

- la première est la représentation simplifiée d'un échantillon qui comporte plusieurs termes égaux, la variable X n'étant alors pas nécessairement discrète.

- la seconde est celle qui est explicitée dans le paragraphe I.3.b, c'est-à-dire l'opposée d'une variable continue.

■ Quand les résultats obtenus sont nombreux, on les regroupe en classes d'intervalles égaux ou variables selon les cas. On ne consi-

dère plus la valeur exacte d'une réalisation, mais le nombre de réalisations dans chaque classe.
Cette série est alors constituée d'éléments que nous appellerons *classes-variable continue.*

Nous avons choisi de définir arbitrairement les classes par leurs valeurs limites inférieures et leurs valeurs limites supérieures. Il est préférable de choisir la limite supérieure de la classe numéro i égale à la limite inférieure de la classe numéro i + 1.

Bien entendu, les classes sont rangées dans l'ordre croissant.

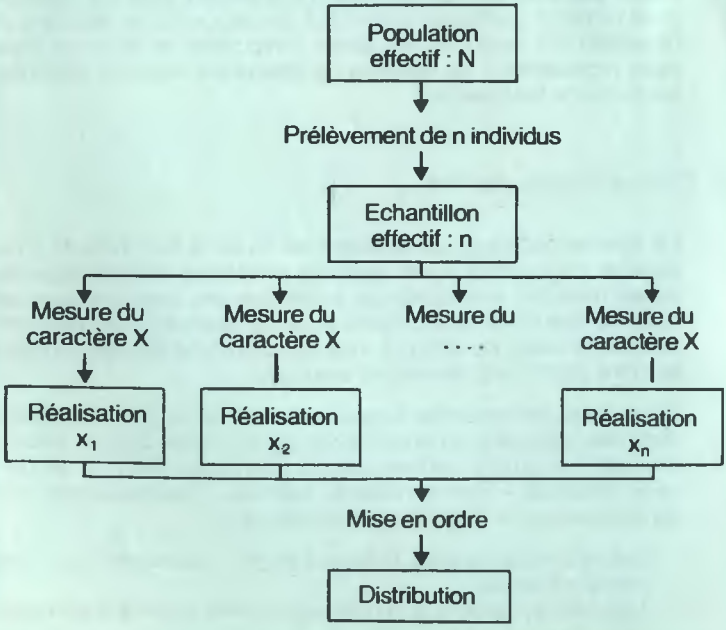
Attention !

Il faut bien noter que le regroupement en classes provoque toujours une perte d'information.

Nous emploierons le terme échantillon ou le terme distribution (échantillon ordonné) sans distinction. Toutefois, lorsque le rangement dans l'ordre croissant de l'échantillon est primordial, nous le signalerons.

b) *Traitement statistique général.*

En résumé, le traitement statistique s'effectue de la façon suivante :



3. LES DIFFÉRENTS TYPES DE DONNÉES

Nous venons de spécifier quatre types de données dans le chapitre précédent :

- Données indépendantes.
- Données appariées.
- Classes-variable discrète.
- Classes-variable continue.

a) *Données indépendantes - Données appariées.*

Rappelons que ces données n'ont subi ni regroupement, ni mise en ordre. La différence entre ces deux types est surtout importante lorsqu'il s'agit de comparer les données de deux ou plusieurs échantillons (étant entendu que tous ces échantillons ont le même type de données).

Les séries composées de données indépendantes peuvent représenter par exemple les résultats d'une mesure faite sur plusieurs prélèvements (à chaque échantillon correspond donc une série de résultats). En revanche, les séries composées de données appariées représentent les résultats de différentes mesures effectuées sur le même prélèvement.

b) *Classes-variable discrète.*

La mise en évidence de la continuité ou de la discontinuité d'une variable peut parfois poser quelques problèmes. Par exemple, certaines mesures sont continues théoriquement, mais pratiquement discontinues du fait que certains résultats (s'appliquant à des grandeurs continues) ne diffèrent entre eux que d'une quantité inférieure à la plus petite unité de mesure employée.

Il faut donc faire attention lorsqu'on décide de ranger des données dans des « classes-variable discrète », car comme nous le préciserons ultérieurement, certains calculs sont impossibles sur les données intitulées « classes-variable discrète ». Heureusement, certains exemples ne posent aucun problème :

voitures immatriculées en France en 19 . ., classées d'après leur puissance fiscale.
Il est bien entendu que la puissance fiscale ne peut prendre que des valeurs bien déterminées (2 CV, 4 CV...).

c) *Classes-variable continue.*

Nous avons déjà décrit la structure de ces classes :

- Borne inférieure.
- Borne supérieure.
- Effectif de la classe.

II. — PARAMÈTRES CARACTÉRISTIQUES D'UNE DISTRIBUTION

Un échantillon rassemble en général un nombre élevé d'informations (toutes les réalisations ou mesures de la variable X).

Pour caractériser la population d'où il est prélevé, il est préférable de définir un certain nombre de paramètres qui doivent posséder les propriétés suivantes :

- Tenir compte de toutes les données de l'échantillon.
- Varier très peu entre deux échantillons prélevés dans une même population.
- Etre simple à obtenir et pouvoir se prêter au calcul algébrique.

Deux types de paramètres vont répondre à ces conditions :

- Les paramètres de position ou caractéristiques de tendance centrale.
- Les paramètres de dispersion ou caractéristiques de répartition autour de la tendance centrale.

1. PARAMÈTRES DE POSITION

Cette première catégorie de paramètres rend compte de la tendance centrale de la population.
Ceci laisse donc supposer que ces valeurs sont choisies de telle façon que les termes de l'échantillon se groupent autour de ces valeurs.

Les paramètres les plus utilisés sont :

- La médiane
- La moyenne (arithmétique)
- Les quartiles.

a) *La médiane*

Tous les termes de l'échantillon étant rangés par ordre de grandeur croissant, la médiane est une quantité telle qu'il existe autant de termes de l'échantillon qui lui sont inférieurs que de termes qui lui sont supérieurs.

Ou encore, pour parler en termes de distribution, la médiane est la réalisation de la variable X qui sépare la distribution en deux sous-ensembles de même effectif.

Considérons les échantillons suivants :

- 3 ; 6 ; 11 ; 19 ; 23
  Médiane = 11.
- 3 ; 6 ; 11 ; 13 ; 19 ; 23

  la médiane, dans ce cas, serait tout nombre compris entre 11 et 13. En général, on prend la moyenne des deux termes médians.
  Médiane = 12.

La médiane se détermine donc facilement. Elle a de plus l'avantage de ne pas être influencée par les termes anormaux de la distribution.

Avant d'aborder le calcul de la médiane pour les données de types classes-variable continue, il nous faut voir le problème posé par les données de type classes-variable discrète.

Etudions ceci sur un exemple :

300 lancers de dé sont répartis de la façon suivante :

x_i	1	2	3	4	5	6
y_i	40	49	72	50	42	47

Classes-variable discrète.

La médiane est donc un nombre compris entre le 150e et 151e terme.

Celui-ci appartient à la troisième classe. Or, le nombre 3 ne répond pas à la définition de la médiane car parmi les 150 termes écrits à sa gauche (resp. droite), certains lui sont inférieurs (resp. supérieurs), mais d'autres lui sont égaux.

Dans le cas très particulier où la somme des termes de la première et deuxième classes serait égales à la somme des quatrième et cinquième classes, « 3 » pourrait être considéré comme médiane.

Dans tous les autres cas, un échantillon à caractère discontinu, c'est-à-dire formé de classes-variable discrète, ne possède pas de médiane. Celle-ci ne sera donc pas calculée.

Revenons au mode de calcul pour les données de type *classes-variable continue*. Rappelons que ces classes sont spécifiées grâce à leurs bornes inférieures et supérieures.

Soit la distribution : (110 termes)

x_i	0-2	2-4	4-6	6-8	8-10
y_i	8	23	54	15	10

Classes-variable continue

La médiane se trouve entre les 55e et le 56e termes. Ceci correspond donc à un terme de la troisième classe.
Nous pouvons calculer le rang de la médiane à l'intérieur de cette classe.

Le premier terme de la troisième classe est le 32e terme de la distribution. Donc, la médiane occupe le 55 − 31 = 24e rang dans la classe.

Les termes de la distribution sont censés varier de façon continue. Nous pouvons donc supposer que les 54 termes de la troisième classe se succèdent par intervalles égaux à 2/54.

Par suite, le 24e terme de cette classe se trouve égal à :

$4 + 2/54 \times 24 = 4{,}89.$

Médiane = 4,89.

b) *La moyenne*

La moyenne arithmétique d'un échantillon est égale à la somme des termes de cet échantillon divisée par le nombre de ces termes.
Ou encore, soit x_i la valeur de la distribution prise avec l'effectif n_i. La moyenne $\bar{x}$ est telle que

$$\bar{x} = \frac{\sum_i n_i \, x_i}{\sum_i n_i} = \frac{\sum_i n_i \, x_i}{n}$$

$n = \sum_i n_i$: la taille de l'échantillon.

Dans le cas de données de type classes-variable continue, on fait l'hypothèse qu'à l'intérieur d'une classe, tous les termes sont égaux à la valeur centrale de cette classe, et on applique à ces valeurs centrales la formule ci-dessus.

c) *Les quartiles*

Les quartiles sont des quantiles d'ordre 4. Ce sont donc les trois valeurs du caractère X qui partagent la distribution en quatre sous-ensembles d'effectifs égaux.

Il faut bien noter que pour le calcul de ces valeurs, l'échantillon doit être rangé dans l'ordre croissant.

Ces paramètres sont très utilisés dans les statistiques démographiques.

Par le même raisonnement que pour la médiane, il faut remarquer qu'on ne peut pas calculer en général les quartiles pour un échantillon dont les données sont du type classes-variable discrète.

Le premier quartile est encore appelé quartile inférieur. Le troisième quartile est encore appelé quartile supérieur. Le quartile central est en fait la médiane.

Considérons la distribution :

11 ; 13 ; 14 ; 15 ; 16.

Quartile inférieur = 13.
Quartile supérieur = 15.

2. PARAMÈTRES DE DISPERSION

Cette deuxième catégorie de paramètres caractérise l'étalement ou la répartition des termes de la distribution autour de la valeur centrale.

Les paramètres les plus utilisés sont :

- La variance
- L'écart-type
- L'écart-moyen
- L'écart interquartile

a) *Variance et Ecart-Type*

On appelle variance ou fluctuation d'un échantillon la moyenne des carrés des écarts à la moyenne.
L'écart-type est la racine carrée de la variance. Il est encore appelé écart quadratique moyen. C'est le paramètre le plus employé pour caractériser la dispersion.

Soit un échantillon de taille n, soit $\bar{x}$ sa moyenne : sa variance est notée σ^2 :

$$\sigma^2 = \sum_i \frac{(x_i - \bar{x})^2}{n} = \sum_i \frac{x_i^2}{n} - \bar{x}^2$$

On peut encore écrire, si à la réalisation x_i de la variable X il correspond l'effectif n_i :

$$\sigma^2 = \frac{\sum_i n_i (x_i - \bar{x})^2}{\sum_i n_i} = \frac{\sum_i n_i x_i^2}{n} - \bar{x}^2$$

(Voir le paragraphe IV. ESTIMATION.)

De même que pour la moyenne, s'il s'agit de données de type classes-variable continue, on prend pour chaque classe la valeur centrale et on applique la formule ci-dessus.

b) *L'écart moyen*

L'écart moyen est la moyenne des valeurs absolues des écarts des mesures à la moyenne $\bar{x}$.

$$EM = \frac{\sum_i |x_i - \bar{x}|}{n}$$

n : taille de l'échantillon.

Ou encore :

$$EM = \frac{\sum_i n_i |x_i - \bar{x}|}{\sum_i n_i}$$

Ce paramètre est d'une utilisation de moins en moins répandue, car les valeurs absolues limitent beaucoup son exploitation mathématique.

c) *L'écart interquartile*

L'écart interquartile est égal à la différence entre le troisième et le premier quartile.

Comme nous l'avons déjà précisé dans le paragraphe II.2, les quartiles ne sont pas calculés pour les données de type classes-variable discrète. Il en est forcément de même pour l'écart interquartile.

D'une manière générale, l'écart interquartile est peu précis. car il ne tient compte que de la moitié des termes de la distribution.

III. — MODÈLE THÉORIQUE DE DISTRIBUTION

Nous venons de décrire les paramètres d'une distribution, sans avoir fait d'hypothèse sur sa forme.

Nous savons que nous ne pouvons en général connaître une population qu'au travers d'échantillons prélevés dans celle-ci.

Il sera donc nécessaire de faire une hypothèse sur la forme de la population, c'est-à-dire de choisir un modèle théorique de distribution permettant de remonter de la connaissance de l'échantillon à la connaissance de la population.

En fonction des résultats obtenus sur l'échantillon qui est considéré par hypothèse, comme représentatif de la population, nous choisirons le modèle le plus adapté.

Ce choix qui paraît a priori assez arbitraire peut être justifié par les tests que nous aborderons dans le paragraphe V.

Le choix du modèle théorique permet d'obtenir à partir des résultats de l'échantillon, des données beaucoup plus élaborées sur la population.

D'une façon générale, la grande majorité des variables aléatoires continues suit une loi de probabilité normale.

Nous nous contenterons donc d'étudier celle-ci.

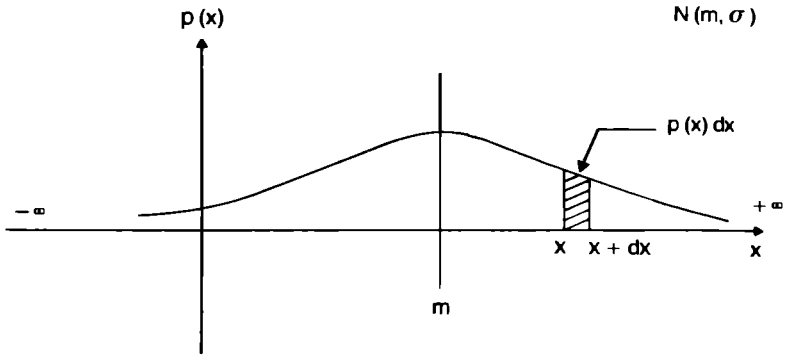
LOI NORMALE

Celle-ci est encore appelée loi de Gauss ou loi de Laplace-Gauss.
Cette loi définit la densité de probabilité p(x) telle que :

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}$$

m : moyenne de la population.
σ : écart-type de la population.

Cela signifie que la probabilité qu'une valeur x_i de la variable X soit située entre x et x + dx est p(x) dx.



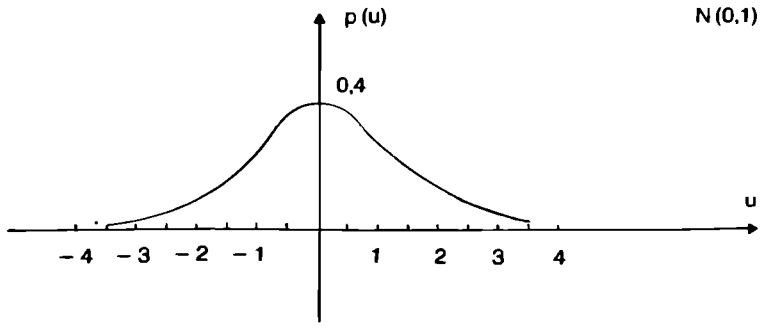
En fait, la loi utilisée est la loi normale centrée réduite de moyenne nulle et d'écart-type égale à 1 :

$$N\ (0,1) \qquad p(u) = \frac{1}{\sqrt{2\pi}}\ e^{-\frac{1}{2}u^2}$$

Des tables fournissent $p(u) = \int_{-\infty}^{u} p(u)$, en fonction de u, c'est-à-dire la probabilité que des réalisations u_i de u soient comprises entre $-\infty$ et u.

Pour obtenir cette loi par rapport à la précédente, il a suffi de faire le changement de variable suivant :

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



IV. — ESTIMATION

Nous venons de voir que pour étudier une population, il fallait choisir un modèle de distribution théorique. Il faut maintenant fixer ses paramètres. Or, comme nous ne pouvons pas prendre en compte la population complète, car il est impossible de connaître ses paramètres réels, il faut les approcher par l'intermédiaire d'un échantillon.

Bien sûr, on peut considérer que les paramètres de l'échantillon sont égaux à ceux de la population. Ce n'est pas forcément la meilleure représentation.
Nous allons donc faire subir aux paramètres de l'échantillon certaines opérations appelées « estimations », afin de mieux cerner les paramètres de la population. Les résultats obtenus s'appellent également estimations.

Nous dirons simplement que les estimations tiennent compte des aléas de l'échantillonnage et nous ne rentrerons pas plus avant dans les détails. Nous énoncerons quelques résultats, dans le cas de la loi normale.

- La moyenne « m » de la population étant inconnue, une estimation de sa variance est :

$$\frac{n\sigma^2}{n-1} = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$$

La taille de l'échantillon est $n = \sum_i n_i$.
$\bar{x}$ est la moyenne de l'échantillon.
σ^2 est la variance de l'échantillon.

Ou encore

$$\frac{\sum_i n_i\ (x_i - \bar{x})^2}{\sum_i n_i - 1} = \frac{\sum n_i\ x_i^2}{n-1} - \frac{n}{n-1}\bar{x}^2$$

- La variance de la population étant inconnue, une estimation de sa moyenne est :

$\bar{x}$ (moyenne de l'échantillon) ;

$$\bar{x} = \frac{\sum_i n_i\ x_i}{\sum_i n_i}$$

En fait, dans les calculs de moyennes et de variances qui vous sont proposés dans ce logiciel, l'estimation est déjà effectuée.

Donc, les moyenne et variance d'un échantillon sont déjà les estimations des moyenne et variance de la population.

De même, les paramètres de la loi normale du paragraphe 3 a. sont en fait les paramètres estimés.

En résumé, il faut bien faire la distinction entre les trois notions suivantes :

- *Les paramètres mesurés sur l'échantillon qui sont des chiffres précis :*

$$\bar{x} = \sum_i x_i / n$$

$$\overline{\sigma^2} = \sum_i (x_i - \bar{x})^2 / n.$$

- *Les paramètres exacts de la population totale qui sont inconnus puisqu'on n'en a observé qu'un échantillon.*
- *Les estimations $\hat{x}$ et $\hat{\sigma}^2$ que l'on peut faire de ces divers paramètres à l'aide des observations faites sur l'échantillon :*

$$\hat{x} = \bar{x}.$$

$$\bar{\sigma}^2 = n\,\bar{\sigma}^2 / (n-1)$$

V. — TESTS

Il est possible d'estimer les paramètres d'une population à partir des paramètres de l'échantillon.
La statistique permet de résoudre d'autres problèmes qui sont toutefois intimement liés à la théorie de l'estimation comme celui de la comparaison de deux populations au travers de leurs échantillons ou celui de la comparaison du modèle statistique théorique choisi pour représenter la population (cf. paragraphe III), à l'échantillon prélevé.

La démarche suivie pour résoudre ces problèmes est de faire une hypothèse H. Dans le premier problème, par exemple, on supposera que les deux populations sont identiques.

Dans le cadre de cette hypothèse, l'écart entre leurs paramètres (paramètres estimés) suivra une certaine loi de probabilité qui permettra de déterminer un intervalle dans lequel le paramètre (la moyenne ou la variance) aura une probabilité $(1 - \alpha)$ de se trouver, si l'hypothèse est exacte.

La stratégie du statisticien consiste donc à choisir un ensemble D (complémentaire de l'intervalle cité ci-dessus) et à rejeter H ou à l'accepter, au contraire, selon que l'écart tombe dans D ou non.

D porte aussi le nom de *région critique du test.* C'est la région de rejet, son complémentaire étant la région d'acceptation de l'hypothèse.

- α représente la probabilité de rejeter H dans le cas où H est vraie. C'est la probabilité d'erreur de premier type.
- 1- α représente la probabilité d'accepter H dans le cas où elle est fausse. C'est la probabilité d'erreur de deuxième type.

Cette démarche porte le nom de *test d'hypothèse.* Il faut bien remarquer que la réponse obtenue n'est pas une affirmation absolue, que ce soit pour infirmer ou confirmer l'hypothèse. La réponse comporte toujours un risque.

Nous allons étudier deux catégories de tests :

- les tests d'ajustement ou tests du χ^2,
- les tests paramétriques.

Les premiers consistent à comparer le modèle théorique choisi pour représenter la population, à l'échantillon prélevé dans celle-ci.

Les seconds permettent la comparaison des paramètres estimés sur deux ou plusieurs échantillons afin de savoir si leurs différences sont

significatives ou si on peut les considérer comme issus d'une même population.

1. TEST DU χ^2 (KHI-DEUX)

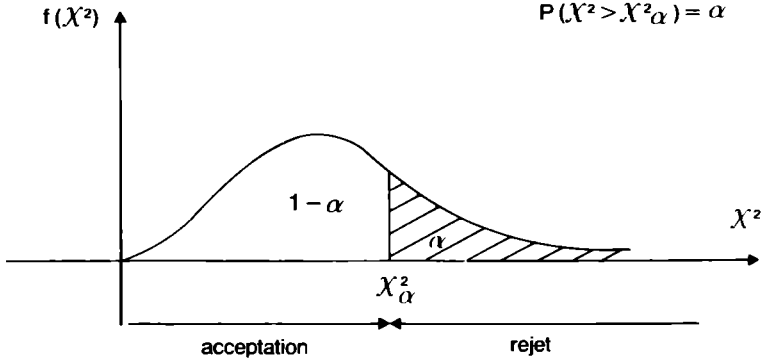
Le choix du modèle théorique et le calcul des paramètres estimés étant effectués, il est nécessaire de tester la légitimité de ce choix.

On fait donc l'hypothèse H, que le modèle théorique choisi représente bien la population étudiée. Le principe du test consiste alors à s'assurer que l'échantillon est compatible avec ce modèle.

Le test du χ^2 que nous allons développer permet de se raccorder à n'importe quelle loi de distribution théorique ; nous insisterons plus particulièrement sur le cas d'une loi théorique de répartition connue mais qui nécessite l'utilisation d'un échantillon de référence donné par l'utilisateur, et sur celui d'une loi théorique connue normale (variable continue).

Il faut noter avant toute chose que ces tests ne s'appliquent qu'aux données de type classes, car ils se calculent sur des effectifs de classe.

Fonction de répartition de la loi du χ^2 :



Principe

Chaque classe i de l'échantillon a un effectif n_i.

On construit un échantillon idéal de MEME TAILLE que l'échantillon prélevé. On l'ordonne suivant des classes identiques avec des effectifs n_i, obtenus à partir du modèle théorique.

On utilise la somme des écarts quadratiques relatifs entre les effectifs qui n'est autre que le χ^2 observé.

$$\chi^2_{obs} = \sum_{i} \frac{(n_i - n'_i)^2}{n'_i}$$

Attention !

On peut admettre que χ^2_{obs} suit une loi du χ^2, quel que soit Σn_i, sous réserve que les effectifs des classes soient au moins égaux à 4 ou 5.

Les valeurs de α en fonction de χ^2_α sont tabulées :

α représente la probabilité pour une valeur de $\chi^2_{obs.}$ de dépasser la valeur χ^2_α théorique :

- Si $\chi^2_{obs} < \chi^2\alpha$, l'hypothèse peut être retenue (mais le test ne justifie pas l'hypothèse).

- Si $\chi^2_{obs} > \chi^2\alpha$, l'hypothèse doit être rejetée.

Soit ν, le nombre de degrés de liberté. $\nu = n - p - 1$

avec n = nombre de classes de l'échantillon
p = nombre de paramètres estimés

Lorsqu'on désire consulter une table (pour α donné), le risque d'erreur (de premier type) que l'on se donne en général est $\alpha = 0{,}05$.
Notons que le choix d'un α plus faible conduit à augmenter χ^2_α.
Comme nous l'avons déjà fait remarquer, le test du χ^2 permet uniquement de justifier *avec un risque d'erreur choisi* le refus de l'hypothèse.

Le programme qui vous est proposé ne nécessite pas l'utilisation d'une table du χ^2.

En effet, la probabilité P calculée, représente la probabilité d'accepter H alors que celle-ci est fausse. P correspond donc à $1 - \alpha$. Plus P est petite, meilleur est le résultat.

1.1. Loi théorique de répartition connue — ajustement à une référence

Ce test s'applique aux données de type classes-variable continue ou classes-variable discrète.

Les effectifs théoriques se trouvent dans un échantillon appelé échantillon de référence. L'effectif total de l'échantillon de référence est réajusté pour être égal à celui de l'échantillon à tester.

Traitons l'exemple suivant :

D'une table de nombres au hasard, on extrait 250 chiffres et on observe ainsi :

x_i	0	1	2	3	4	5	6	7	8	9
n_i	32	22	23	31	21	23	28	25	18	27

L'hypothèse est que chaque chiffre a une probabilité de 0,1 (1/10) d'apparaître à un endroit quelconque de la table.

L'échantillon de référence va donc être ainsi constitué :

x_i	0	1	2	3	4	5	6	7	8	9
n_i	25	25	25	25	25	25	25	25	25	25

$$\chi^2_{obs} = \frac{1}{25}\,[(32-25)^2 + (22-25)^2 + \ldots]$$

$\chi^2_{obs} = 7{,}2$ $\qquad$ $\nu = n - p - 1 = 9$

car dans ce cas il n'y a pas de paramètres estimés, donc $p = 0$.

$P = 0{,}38$

Si nous utilisons une table de χ^2 *:*

pour $\alpha = 0{,}05$ *et* $\nu = 9$ $\quad \chi^2\,\alpha = 16{,}92$

$\chi^2_{obs} < \chi^2\alpha$ *on décide donc d'accepter l'hypothèse.*

Résultat prévisible grâce à la probabilité P puisque la probabilité d'accepter H, dans le cas où celle-ci est fausse, n'est que de 38 %.

1.2. Loi théorique de répartition normale

L'ajustement à une loi normale s'opère sur les données de type *classes-variable continue*.

On fait donc l'hypothèse que la loi de probabilité de la population est une loi normale dont les paramètres sont des estimations obtenues sur l'échantillon prélevé, soit $\bar{x}$ et σ la moyenne et l'écart-type.

Rappelons que les classes sont définies par leurs bornes inférieures et supérieures.

La méthode employée utilise les bornes supérieures (une autre méthode consiste à prendre le centre de classe, mais les calculs sont plus approximatifs).

Procédure pratique :

Il nous faut connaître les valeurs n_i' des effectifs théoriques. Il faut pour cela connaître les valeurs de p_i, c'est-à-dire les valeurs de la probabilité de prélever un élément de la population appartenant à la classe i.

Pour cela, nous allons considérer la variable réduite :

$u_i = (x_i - \bar{x}) / \sigma$

Les bornes supérieures des classes sont recalculées en variable réduite. Les p_i sont alors calculées (la méthode employée par le programme ne nécessite pas d'être explicitée — une autre méthode consiste à regarder les tables de la loi normale réduite).

Les effectifs théoriques sont alors : $n_i' = n\ p_i / \Sigma p_i$ avec n, la taille de l'échantillon et Σp_i, la somme des probabilités qui n'est pas toujours 1, car ce test ne prend pas en compte les classes extrêmes.

ν, le nombre de degrés de liberté est n-3 puisque dans ce cas, le nombre de paramètres estimés, p est 2.

Etudions ceci sur un exemple :

On désire vérifier la normalité d'une variable continue à partir de la distribution observée suivante :

x_i	1-3	3-5	5-7	7-9	9-11	11-13	13-15
n_i	8	23	54	128	114	65	18

$n = \sum_i n_i$ = *taille de l'échantillon* = 410.

moyenne estimée $\bar{x} = 8{,}85$

écart-type estimé $\sigma = 2{,}58$

La variable réduite à considérer est donc :

$u_i = (x_i - 8{,}85) / 2{,}58$

Les résultats sont regroupés dans le tableau ci-dessous :

$p_i = P(u_i) - P(u_{i-1})$

Borne sup u_i	n_i	$P(u_i)$	p_i	n'_i
− 2,27	8	0,0105	0,0105	4,35
− 1,49	23	0,0681	0,0562	23,25
− 0,72	54	0,2358	0,1689	69,95
− 0,06	128	0,5238	0,2865	118,61
− 0,83	114	0,7967	0,2746	113,68
− 1,61	65	0,9463	0,1484	61,44
− 2,38	18	0,9913	0,0452	18,72

$$\chi^2_{obs} = \sum_i \frac{(n_i - n'_i)^2}{n'_i} = 7{,}68$$

$$\nu = 7 - 1 - 2 = 4$$

Si nous prenons $\alpha = 5\ \%$ et que nous regardons dans une table : $\chi^2 \alpha = 9{,}49$.

Rien ne s'oppose au rejet de l'hypothèse, bien que la différence entre χ^2_{obs} et χ^2 théorique ne soit pas énorme.

La probabilité P proposée par le programme est : P = 0.89, c'est-à-dire qu'il y a un risque de 89 % d'accepter H dans le cas où celle-ci est fausse.

A vous de décider !

2. TESTS PARAMETRIQUES

Deux types de problèmes sont résolus au moyen de ces tests :

- la comparaison de deux échantillons normaux, dans le but de faire un test d'identité de deux populations,
- la recherche d'un ou plusieurs facteurs contrôlés sur la variable X, autrement dit une analyse de variance.

2.1. Test d'identité de deux populations

On considère que la loi de distribution théorique des deux populations dont on a prélevé un échantillon est normale.

La comparaison des échantillons se limite à celle des variances et des moyennes estimées.

L'hypothèse H que l'on se propose de vérifier est l'identité des deux populations. Pour ce faire, on peut procéder à :

- une comparaison des variances,
- une comparaison des moyennes.

Les tests utilisés dépendent de la taille de l'échantillon et parfois du type de données, bien que valables pour tous les types.

Nous appellerons échantillon de taille élevée, un échantillon dont l'effectif total est supérieur à 100 (N.B. : dans notre logiciel, il ne peut s'agir que de classes, puisque l'effectif des échantillons constitués de données indépendantes ou appariées est limité à 100).

2.1.1. *Comparaison de variances*

Considérons deux échantillons de taille respective n_1 et n_2 et d'écart-type σ_1 et σ_2.

H = les variances des deux populations sont égales à σ.

a) *Echantillons de taille élevée.*

n_1 et n_2 étant élevés, σ_1 et σ_2 suivent approximativement une loi normale, de moyenne m et d'écarts-types respectifs $\sigma / \sqrt{2n_1}$ et $\sigma / \sqrt{2n_2}$.

σ étant inconnue, on utilise l'estimation basée sur les deux échantillons, à savoir :

$$\sigma^2 = \frac{n_1 \sigma_1^2 + n_2 \sigma_2^2}{n_1 + n_2 - 2} \qquad (\sigma_1^2, \sigma_2^2 : \text{variances estimées})$$

La somme ou la différence de deux variables normales étant une variable normale, $\sigma_1 - \sigma_2$ est une variable normale de moyenne nulle et d'écart-type :

$$\sigma \sqrt{\frac{1}{2n_1} + \frac{1}{2n_2}}$$

soit u, la variable aléatoire normale réduite : $u = \dfrac{\sigma_1 - \sigma_2}{\sqrt{1/2\, n_1 + 1/2\, n_2}}$

On a une probabilité $1 - \alpha$ que u soit comprise dans $[-u_\alpha, u_\alpha]$; si u n'est pas comprise dans cet intervalle, on pourra repousser H, avec un risque α.

La donnée de u vous permet donc de vous servir d'une table, mais le programme vous fournit une probabilité P qui représente le risque d'accepter H dans le cas où elle est fausse.

Les tables statistiques vous sont alors inutiles.

Prenons un exemple :

Les résultats obtenus sur deux échantillons sont les suivants :

1er échantillon $n_1 = 100$; $\sigma_1 = 1$
2e échantillon $n_2 = 250$; $\sigma_2 = 4$

$$\sigma = \sqrt{\frac{100 + 250 \times 16}{100 + 250 - 2}} = 3{,}48$$

$$d'où\ u = -10{,}30 = \frac{1 - 4}{3{,}48 \times (1/200 + 1/500)}$$

Si l'on veut se servir d'une table et que l'on choisit $\alpha = 5\ \%$, *alors* $u = 1{,}96$.

Par conséquent, on peut rejeter l'hypothèse d'égalité des variances puisque

$-10{,}30$ *n'appartient pas à l'intervalle* $[-1{,}96 ; 1{,}96]$

La probabilité P donnée par le programme est « 1 », c'est-à-dire que le risque d'accepter H dans le cas où celle-ci est fausse est de 100 %. On rejette donc l'hypothèse sans scrupule.

b) *Echantillons de taille réduite.*

Soit $\mathcal{F}$ le rapport des deux estimations indépendantes de σ^2

$\mathcal{F} = \frac{{\sigma_1}^2}{{\sigma_2}^2}$ le rapport considéré est celui de la plus grande estimation sur la plus petite.

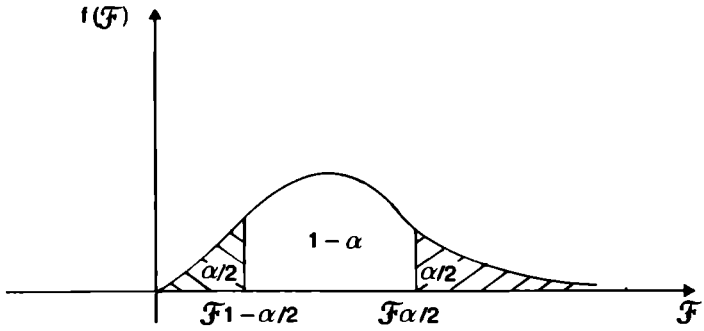
Cette variable suit une loi de Snédécor à
$r_1 = (n_1 - 1)$, $r_2 = (n_2 - 1)$ degrés de liberté.

Cette loi est tabulée.

Si H est vérifiée, $\mathcal{F}_{1-\alpha/2} < \mathcal{F} < \mathcal{F}_{\alpha/2}$ dans $(1-\alpha)$ des cas.

Donc, si $\mathcal{F} \in]\mathcal{F}_{1-\alpha/2} ; \mathcal{F}_{\alpha/2}[$ rien ne permet de rejeter H.

Si $\mathcal{F} \notin]\mathcal{F}_{1-\alpha/2} ; \mathcal{F}_{\alpha/2}[$ on rejette H.



La probabilité P calculée par le logiciel représente la probabilité d'accepter H alors que celle-ci est fausse.

Exemple :
Soient deux lots de poudre à fusil. On effectue sept tirs au fusil avec chacun de ces lots et on relève les vitesses initiales de la balle.

Le problème est de savoir si ces deux lots ont été fabriqués suivant le même procédé.

On commence donc par tester l'égalité des variances.

Ech. 1	Ech. 2
801	809
803	801
804	805
798	803
805	800
797	808
802	801

$\nu_1 = \nu_2 = 6$

$\sigma_1 = 3{,}09$

$\sigma_2 = 3{,}69$

$$\mathcal{F} = \frac{(3{,}69)^2}{(3{,}09)^2} = 1{,}43$$

$P = 0{,}33$

La probabilité d'accepter H, alors que celle-ci est fausse, étant de 33 %, l'hypothèse d'égalite des variances est acceptée.

Nous allons pouvoir tester l'hypothèse d'égalité des moyennes.

Remarque :
La comparaison des moyennes peut toutefois être effectuée même si la comparaison des variances a donné un résultat négatif.

2.1.2. *Comparaison des moyennes*

Considérons deux échantillons de taille n_1 et n_2 et de moyenne $\overline{x_1}$ et $\overline{x_2}$.

H : les moyennes des deux populations sont identiques.

a) *Echantillons de taille élevée.*

$\overline{x_1}$ et $\overline{x_2}$ suivent une loi normale de moyenne m et d'écart-type $\sigma_1 / \sqrt{n_1}$ et $\sigma_2 / \sqrt{n_2}$

avec σ_1 et σ_2 les écarts-type estimés des échantillons.

L'estimation de σ est alors :

$$\sigma = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

La différence de deux variables normales étant une variable normale. $\overline{x_1} - \overline{x_2}$ est une variable normale de moyenne nulle et d'écart-type : σ.

Soit u, la variable aléatoire normale réduite

$$u = \frac{\overline{x_1} - \overline{x_2}}{\sigma}$$

On a une probabilité $(1 - \alpha)$ que u soit comprise dans $[-u_\alpha ; +u_\alpha]$. Si u n'est pas comprise dans cet intervalle, on pourra repousser H avec un risque α.

La donnée de u vous permet de vous servir d'une table, mais le programme vous fournit une probabilité P, qui représente le risque d'accepter H dans le cas où elle est fausse.

Prenons un exemple :

Les résultats obtenus sur deux échantillons sont les suivants :

1er échantillon : $n_1 = 150$; $\overline{x_1} = 80,5$; $\sigma_1 = 2,45$.
2e échantillon : $n_2 = 2,50$; $\overline{x_2} = 81,7$; $\sigma_2 = 2,53$.

$$u = \frac{80{,}5 - 81{,}7}{\sqrt{\frac{(2{,}45)^2}{150} + \frac{(2{,}53)^2}{250}}} = -4{,}68$$

$P = 1$

L'hypothèse d'égalité des moyennes est à rejeter puisque le risque d'accepter H dans le cas où elle est fausse est de 100 %.

b) *Echantillons de taille réduite.*

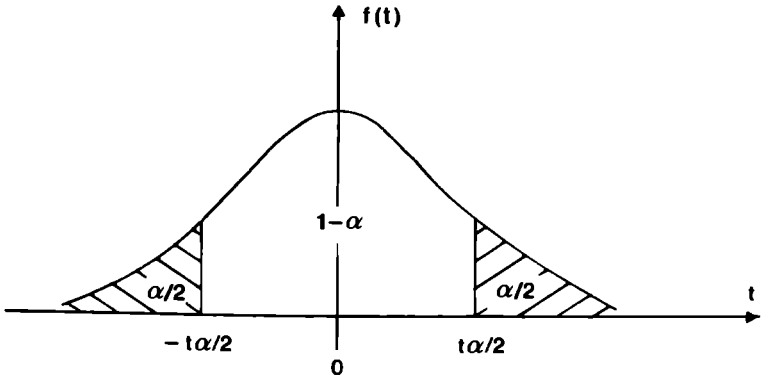
La variable $t = \dfrac{\overline{x_1} - \overline{x_2}}{\sigma}$

avec $\sigma^2 = \dfrac{\sigma_1^{\,2}}{n_1} + \dfrac{\sigma_2^{\,2}}{n_2}$

($\sigma_1^{\,2}$ et $\sigma_2^{\,2}$ sont les variances estimées)

suit une loi de Student-Fisher à $\nu = n_1 + n_2 - 2$ degrés de liberté.

Cette loi est tabulée ; voici sa fonction de répartition :



Si H est vérifiée, $t \in [-t_{\alpha/2} ; t_{\alpha/2}]$ dans $(1 - \alpha)$ des cas

— si $|t| > t_{\alpha/2}$, on rejette l'hypothèse

— si $|t| < t_{\alpha/2}$, on admet que les deux échantillons proviennent de la même population (car ils ont même variance et même moyenne).

La probabilité P calculée par le logiciel représente la probabilité d'accepter H dans le cas où celle-ci est fausse.

Etudions un exemple :

Dans une usine, on cherche à savoir si un changement de l'environnement peut modifier un rendement mesuré ici par le nombre moyen de pièces produites à l'heure par chaque ouvrier.

On note pour chacun des quinze ouvriers observés, son rendement avant et après l'introduction de ces changements.

Avant	Après	Avant	Après
45	48	50	60
36	40	40	40
47	53	40	40
40	40	30	35
45	46	45	50
35	30	30	40
36	40	45	50
50	60		

Les deux rendements obtenus avant et après transformation par le même ouvrier sont en relation. On veut savoir si ces rendements sont significativement différents. Si c'est le cas, l'hypothèse H n'est pas vérifiée.
On trouve : $\nu = 28$
$t = 1{,}38$ $P = 0{,}82$

La probabilité d'accepter H dans le cas où celle-ci est fausse est de 82 %.

Donc, l'hypothèse doit être rejetée, ce qui signifie que les rendements des ouvriers ne sont pas indépendants des modifications de l'environnement.

GENERALISATION

— *Tests de comparaison de plusieurs variances.*

Il en existe, mais ils sont en général peu efficaces et assez approximatifs. Nous ne les étudierons donc pas.

— *Tests de comparaison de plusieurs moyennes.*

On peut tester l'égalité des moyennes de plusieurs populations par une technique appelée « analyse de la variance » que nous allons développer dans le paragraphe suivant.

2.2. Analyse de la variance

Cette analyse permet de rechercher l'influence d'un ou plusieurs facteurs contrôlés sur des éléments de base (nous nous limiterons à deux dans notre exposé).

Le problème est que l'influence de ces facteurs peut avoir un caractère aléatoire et les éléments de base peuvent être dispersés. L'analyse de la variance permet de discerner la variation due aux facteurs, et celle due à la dispersion des éléments.

Une hypothèse fondamentale est que la caractéristique X étudiée suit une loi normale de moyenne m et d'écart-type σ .

Nous allons étudier l'influence d'un et de deux facteurs contrôlés.

Ces tests ne s'utilisent que sur les *données indépendantes* ou *appariées*.

Des données indépendantes conduiront à une analyse à un facteur contrôlé ou encore à une voie ; des données appariées conduiront à une analyse à deux voies.

2.2.1. *Analyse de la variance à un facteur contrôlé*

On veut déterminer l'influence éventuelle d'un facteur A sur la variable X. On dispose de k échantillons, qui représentent les k modalités du facteur A.

A_1	A_2	...	A_k
x_{11} x_{1n1}	x_{21} x_{2n2}		x_{k1} x_{knk}

Les échantillons n'ont pas forcément le même effectif.

Rappelons qu'il s'agit de données indépendantes.

L'hypothèse H est que le facteur A est sans influence sur la variable X.

Notations et principe de la méthode

$\overline{x_i}$: moyenne de l'échantillon i.

$\overline{x}..$: moyenne sur tous les échantillons.

St : somme totale des écarts quadratiques pour l'ensemble des valeurs de x.

$$St = \sum_i \sum_j x_{ij}^2 - \sum_i n_i \, \overline{x}_{..}^2$$

Sa : somme des écarts quadratiques dus aux différentes modalités de A (et au hasard). A peut en effet avoir une influence sur les écarts entre les moyennes $\overline{x}_i$. et la moyenne globale $\overline{x}_{..}$

$$Sa = \sum_i n_i \, \overline{x}_{i.}^2 - \sum_i n_i \, \overline{x}_{..}^2$$

Sr : somme des écarts quadratiques résiduels (ils sont dus au seul hasard, puisque A ne peut être responsable des écarts entre les valeurs de X d'une même colonne, puisqu'il est fixé par chaque échantillon).

au total : $St = Sa + Sr$

Soient Va, Vt et Vr des variances calculées à partir de ces écarts.

Va : variance entre échantillons calculée à partir de Sa (représente la dispersion entre modalités).

$$Va = \frac{Sa}{k-1}$$

Vr : variance résiduelle calculée à partir de Sr (représente la dispersion intra-modalités).

$$Vr = \frac{Sr}{\sum_i n_i - k}$$

Vt : variance totale calculée à partir de St (représente la dispersion totale).

$$Vt = \frac{St}{\sum_i n_i - 1}$$

Va et Vr représentent deux estimations indépendantes de σ^2. Nous ne développerons pas les calculs plus avant et nous dirons simplement que le test de comparaison des variances Va et Vr revient à tester la non-influence du facteur contrôlé.

$$\mathcal{F} = \frac{Va}{Vr}$$

suit une loi de Snédécor à $\nu_1 = k - 1$, $\nu_2 = \sum_i n_i - k$ degrés de liberté
(se reporter au paragraphe 2.1.2. b)

Rappelons que la probabilité calculée par le logiciel représente la probabilité d'accepter H alors que celle-ci est fausse.

Remarque :
Le problème que nous venons d'exposer est identique au problème de comparaison des moyennes de plusieurs échantillons, où le facteur de variation possible est la différence des populations dont sont extraits les échantillons considérés.

Traitons un exemple :

On observe les dépenses annuelles en vêtements de n = 23 individus répartis dans quatre grandes villes de population identique de quatre pays industrialisés :

Pays 1	*Pays 2*	*Pays 3*	*Pays 4*
250	*270*	*250*	*300*
250	*270*	*275*	*305*
270	*290*	*280*	*310*
280	*300*	*295*	*325*
300	*330*	*295*	*325*
	336	*300*	
	345		

Ces observations permettent-elles de conclure à l'existence d'un effet « pays » sur le montant des dépenses en vêtements ?

H = hypothèse de non-influence.

Nous présentons les calculs sous forme d'un tableau regroupant les résultats importants :

Ecart	*DDL*	*Variance*	*F*	*P*
Entre-modalités	*3*	*2144,270*	*4,119*	*0,84*
Intra-modalités	*19*	*520,47*		
Total	*22*	*741,898*		

La probabilité d'accepter H alors que celle-ci est fausse étant de 84 %, nous pouvons conclure qu'il y a un effet « pays » sur le montant des dépenses en vêtements.

2.2.2. *Analyse de la variance à deux facteurs contrôlés*

Cette fois, on veut déterminer l'influence éventuelle de deux facteurs A et B sur une variable.

Le facteur A a k_1 modalités, le facteur B a k_2 modalités. Ceci correspondant en fait à k_1 échantillons dont l'effectif est k_2. Rappelons que ce sont des données appariées.

B \ A	A_1	A_2		A_{k_1}
B_1	x_{11}	x_{21}		x_{k1}
B_2	x_{12}	x_{22}		x_{k12}
.	.	.		.
.	.	.		.
.	.	.		.
B_{k2}	x_{1k2}	x_{2k2}		x_{k1k2}

Deux hypothèses sont faites :

H1 : la non-influence de A sur X.
H2 : la non-influence de B sur X.

Notations et principe de la méthode

$\overline{x}_{i\,.}$: moyenne des résultats relatifs à A_i

$\overline{x}_{.\,j}$: moyenne des résultats relatifs à B_j

$\overline{x}_{..}$: moyenne totale de tous les résultats.

St : somme totale des écarts quadratiques pour l'ensemble des valeurs de X

$$St = \sum_i \sum_j x_{ij}^2 - \frac{\overline{x_{..}}^2}{k_1 k_2}$$

Sa : somme des écarts quadratiques dus aux différentes modalités de A

$$Sa = k_2 \sum_i \overline{x_{i\,.}}^2 - k_1 k_2 \overline{x_{..}}^2$$

Sb : somme des écarts quadratiques dus aux différentes modalités de B

$$Sb = k_1 \sum_j \overline{x_{.\,j}}^2 - k_1 k_2 \overline{x_{..}}^2$$

Sr : somme des écarts quadratiques résiduels

au total : $St = Sa + Sb + Sr$

Soient Va, Vb, Vr et Vt les variances calculées à partir de ces écarts.

Va : variance entre les différentes modalités de A (représente la dispersion entre colonnes)

$$Va = \frac{Sa}{k_1 - 1}$$

Vb : variance entre les différentes modalités de B (représente la dispersion entre lignes)

$$Vb = \frac{Sb}{k_2 - 1}$$

Vr : variance résiduelle

$$Vr = \frac{Sr}{(k_1 - 1)\ (k_2 - 1)}$$

Vt : variance totale

$$Vt = \frac{St}{(k_1 - 1)\ (k_2 - 1) + k_1 + k_2 - 2}$$

Va, Vb et Vr représentent trois estimations indépendantes de σ^2. Le test de comparaison des variances Va et Vr, et Vb et Vr revient à tester la non-influence de A et la non-influence de B.

$\mathcal{F}_1 = \frac{Va}{Vr}$ suit une loi de Snédécor à $\nu 1 = k_1 - 1$, et $\nu 2 = (k_1 - 1)\ (k_2 - 1)$ degrés de liberté.

$\mathcal{F}_2 = \frac{Vb}{Vr}$ suit une loi de Snédécor à $\nu 1 = k_2 - 1$, et $\nu 2 = (k_1 - 1)\ (k_2 - 1)$ degrés de liberté.

(Se reporter au paragraphe 2.1.2. b)

On rappellera simplement que P représente la probabilité d'accepter H alors que celle-ci est fausse.

Traitons un exemple :

Imaginons que l'on dispose de dix observations classées selon deux critères :

La critère A a cinq modalités — le critère B en a quatre

B \ A	A_1	A_2	A_3	A_4	A_5
B_1	8	7	5	9	2
B_2	11	13	9	7	9
B_3	2	5	4	3	2
B_4	6	1	2	4	1

Ces observations permettent-elles de conclure à l'existence d'une influence due à A sur X ou d'une influence due à B sur X ?

H_1 et H_2 : hypothèses de non-influence.

Les résultats sont regroupés dans le tableau suivant :

Ecart	DDL	Variance	F	P
Entre-lignes	3	53,6000	12,7	0,96
Entre-colonnes	4	6,8750	1,659	0,4
Résiduel	12	4,1417		
Total	19	12,3684		

Nous pouvons donc conclure qu'il n'y a pas d'influence du facteur A sur la variable X puisque la probabilité d'accepter H_1 dans le cas où elle est fausse est de 40 %.

Par contre, l'influence du facteur B est significative, puisque le risque d'accepter H_2 dans le cas où elle est fausse est de 96 %.

VI. — CORRÉLATION

La plupart des populations auxquelles nous nous sommes intéressés jusqu'à maintenant ne contenaient que des éléments repérés par la mesure d'un seul caractère.

Mais dans de nombreux cas, il est nécessaire de prendre en compte les valeurs de plusieurs caractères pour individualiser un élément.

Par exemple, il est plus intéressant, pour connaître une population d'individus, d'étudier la distribution de la taille et du poids plutôt que d'étudier la distribution d'un seul de ces paramètres.

Il peut exister entre les valeurs prises par ces caractères, des liaisons de nature différente :

— les valeurs peuvent être liées par une relation linéaire. La donnée d'une ou plusieurs de ces valeurs détermine parfaitement les autres ;

— les valeurs prises par les caractères étudiés sont totalement indépendantes ;

— les caractères ne sont ni indépendants ni liés par une relation linéaire. Les deux caractères sont en corrélation, mais la donnée de l'un d'eux ne permet pas de définir parfaitement les autres.

Nous allons envisager deux types de coefficients de corrélation. Un coefficient de corrélation linéaire et deux coefficients de corrélation non paramétriques calculés sur les rangs des données et non pas sur les données elles-mêmes.

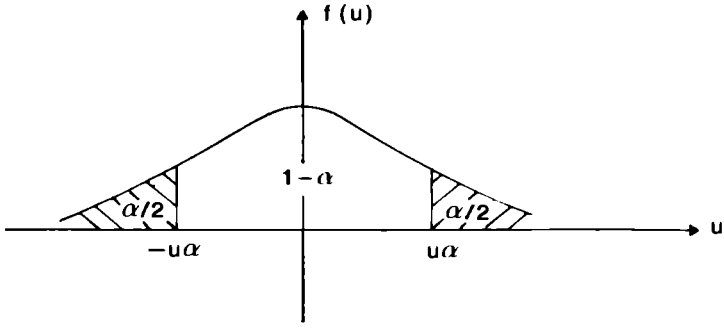
Le but de ces coefficients est toutefois le même : déceler une éventuelle dépendance entre deux variables au vu de n couples d'observations.

L'hypothèse H est l'indépendance de ces deux variables.

Notons, avant de commencer leur étude, que ces coefficients et les probabilités associées ne sont vraiment significatifs que pour des échantillons d'au moins dix observations.

La probabilité utilisée étant la loi normale, on pourra se reporter au paragraphe III pour plus de précisions.

Rappelons tout de même que la probabilité P calculée par le logiciel représente le risque d'accepter H dans le cas où celle-ci est fausse.



1. « r » DE BRAVAIS-PEARSON

Soient deux variables aléatoires X et Y.

Le coefficient de corrélation est défini par la covariance C et les écarts-types σx et σy

$$\rho = \frac{C}{\sigma x \ \sigma y} \qquad \rho \in [-1 ; +1]$$

On peut énoncer quelques propriétés :

- Quelle que soit la loi de probabilité suivie par (X, Y),
  - si X et Y sont indépendantes, alors $|\rho| = 0$;
  - si X et Y sont liées par une relation fonctionnelle, alors $|\rho| = 1$.
- Réciproquement, si la loi de probabilité de X et Y est normale, alors :
  - si $|\rho| = 0$, alors X et Y sont indépendantes ;
  - si $|\rho| = 1$, alors X et Y sont en relation fonctionnelle.

Le coefficient de corrélation « r » que nous calculons est une estimation de $|\rho|$ dans le cas où la loi de probabilité de X et Y est normale.

$$r = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{n \sigma x \ \sigma y}$$

n : effectif de chaque échantillon
σx et σy écarts-types ; $\bar{x}$ et $\bar{y}$ moyennes.

La loi de probabilité de « r » étant complexe, on utilise pour des valeurs de n inférieures à 100 le procédé de la corrélation transformée de Fisher (uniquement dans le cas où $r \neq 1$)

$$z = \text{Argth } r = \frac{1}{2} \text{ Ln } \frac{1+r}{1-r}$$

(Argth : arc tangente hyperbolique)

z suit une loi normale :

de moyenne 0
d'écart-type $1 / \sqrt{n-3}$

Nous considérerons donc la variable normale réduite

$u = \sqrt{n-3}\ z$

Si $u \in [-u\alpha ; +u\alpha]$, on accepte H dans $1-\alpha$ des cas.

Exemple :

Les variables X et Y étant normales, on veut vérifier l'existence d'une relation fonctionnelle entre les deux variables.

X		Y	
1	7	5	10
1	7	6	11
2	8	6	11
2	9	7	12
3	9,5	7	13
3	11	8	13
4	11	8	14
4	11,5	8,5	14
5	12	9	14
5		9,5	
6		10	

On observe les résultats suivants :

r = 0,99 *z = 2,65* *u = 10,93*
P = 1

Conclusion : la probabilité d'accepter H dans le cas où celle-ci est fausse étant de 100 %, on peut conclure qu'il existe une relation fonctionnelle entre X et Y.

2. « τ » DE KENDALL

On peut appeler ce coefficient de corrélation, ainsi que celui de Spearman, des coefficients non paramétriques, car ils ne font intervenir que les rangs des données dans les échantillons.

Principe

Pour chacun des couples (y_i, y_j) du deuxième échantillon, on compte une concordance $(+1)$ si (x_i, x_j) du premier échantillon sont rangés dans le même ordre et une discordance (-1) s'ils sont rangés en sens contraire.

Soient C la somme des concordances et D la somme des discordances.

$K = C - D$ est appelée statistique de Kendall.

S'il n'y a pas de dépendance monotone, donc si H est vérifiée, la valeur de K est proche de 0.

Il est plus simple d'utiliser la statistique suivante, appelée coefficient de corrélation des rangs de Kendall :

$$\tau = \frac{2K}{n(n-1)} \qquad \tau \in [-1 ; +1]$$

Dans le cas d'échantillons dont l'effectif est au moins de 8, τ suit une loi normale de :

moyenne 0

d'écart-type $\sqrt{\dfrac{2\ (2n+5)}{9n\ (n-1)}}$

on utilisera donc la variable aléatoire normale réduite

$$u = \tau \sqrt{\frac{9n\ (n-1)}{2\ (2n+5)}}$$

Si $u \in [-u\alpha ; +u\alpha]$, on accepte H dans $1 - \alpha$ des cas.

Remarque : le traitement des ex-aequo.

Certaines données d'un échantillon peuvent être égales, elles auront donc un rang identique.

On remplacera ce rang par le rang moyen des observations.

Etudions ceci sur un exemple :

Soit l'échantillon X : 9,4 ; 8,2 ; 11,4 ; 10,2 ; 8,2.

Les données n° 1 et n° 5 étant identiques, nous leur attribuons le même rang, soit :

$$(1 + 2) / 2 = 1,5$$

Nous obtenons alors :

X	9,4	8,2	11,4	10,2	8,2
Rang	3	1,5	5	4	1,5

Etudions maintenant un exemple complet

Dans une enquête relative aux ressemblances des goûts entre conjoints, on a demandé à une femme de classer sept tableaux dans l'ordre de ses préférences. Le mari a ensuite, et indépendemment, classé les mêmes tableaux, ce qui donne les résultats suivants :

Tableau	a	b	c	d	e	f	g
Classement du mari	2	3	1	5	4	7	6
Classement de la femme	1	4	2	6	3	5	7

$\tau = 0,62$
$P = 0,975$

La probabilité d'accepter H dans le cas où elle est fausse étant de 97 %, on peut conclure qu'il y a une dépendance monotone entre les deux classements.

3. « ρ » DE SPEARMAN

Ce coefficient est donc lui aussi calculé sur les rangs des données.

Soient deux échantillons dont les observations sont $(x_1, \ldots x_n)$ et $(y_1 ., \ldots y_n)$:

on remplace les x_i par leurs rangs r_i

on remplace les y_i par leurs rangs s_i

(Le traitement des ex-aequo se fait comme expliqué précédemment dans le paragraphe VI.2.)

La statistique ou coefficient de corrélation de Spearman s'écrit :

$$\rho = \frac{\sum_i (r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_i (r_i - \bar{r})^2 \sum_i (s_i - \bar{s})^2}}$$

où r et s sont les moyennes des rangs r_i et des rangs s_i.

Sachant que $r = s = (n + 1) / 2$ et en posant $d_i = r_i - s_i$

on peut écrire plus simplement

$$\rho = 1 - \frac{6}{n(n^2 - 1)} \sum d_i^2$$

on va considérer la variable aléatoire normale réduite :

$$u = \sqrt{n - 1}\,\rho$$

car pour $n > 10$, ρ suit une loi normale :

de moyenne 0
d'écart-type $1/\sqrt{n-1}$

si u est compris dans $[-u\alpha ; +u\alpha]$, on accepte H dans $(1 - \alpha)$ des cas.

Remarque : Ce calcul peut être fait avec le « r » de Bravais-Pearson. Mais les calculs sont plus pénibles pour sensiblement le même résultat.

Etudions un exemple :

Quinze élèves ont été classés une première fois par une épreuve de français, une seconde fois par une épreuve de mathématiques :

Elèves	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*	*M*	*N*	*O*
Français	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
Mathématiques	*9*	*3*	*1*	*11*	*2*	*5*	*8*	*13*	*4*	*10*	*7*	*14*	*15*	*6*	*12*

On obtient : $\tau = 0{,}507$
$P = 0{,}94$

On rejette donc l'hypothèse d'indépendance entre les deux variables. Il y a donc une dépendance entre les notes obtenues en français et celles obtenues en mathématiques.

MANUEL UTILISATEUR

SOMMAIRE

CHAPITRE TROISIEME : CALCULS STATISTIQUES

I. STATISTIQUE DESCRIPTIVE

II. TEST DU χ^2

1. Ajustement loi normale.
2. Ajustement référence.

III. TESTS PARAMETRIQUES

1. Comparaison de variances et de moyennes.
2. Analyse de la variance.

IV. CORRELATION

CHAPITRE QUATRIEME : UTILITAIRES

I. LES COMMANDES DOS

1. Catalogue.
2. Blocs libres sur disquette.
3. Initialisation de la disquette.
4. Copie Fichier.
5. Destruction de Fichier.

II. CARACTERISTIQUES SYSTEME

CHAPITRE PREMIER : GENERALITES

I. PRINCIPAUX ASPECTS DE L'UTILISATION DU SYSTEME TO7 POUR STATISTIQUES

Le double aspect du fonctionnement du logiciel, constitution des données (indication des échantillons utilisés pour l'étude statistique d'une population), et calculs statistiques, confère un rôle important à l'utilisation du lecteur de disquette.
Vous devrez en effet « stocker » sur disquette les échantillons, puis, pour permettre au logiciel d'effectuer les calculs que vous lui commanderez, vous devrez sélectionner tel ou tel échantillon sur la disquette.
Par ailleurs, ceci impose une certaine complexité dans l'organisation des fonctions que met en œuvre le logiciel. Nous vous conseillons donc, pour mieux maîtriser celui-ci, d'assimiler rapidement l'arbre fonctionnel, que vous propose la mémo carte jointe à la notice, et notamment de bien reconnaître les divers « menus » qui articulent cet arbre : ce sont eux qui vous permettent de circuler dans le logiciel pour sélectionner ses différentes fonctions.
Les chapitres qui suivent vous proposent des informations plus détaillées sur tous les aspects de l'utilisation de STATISTIQUES. Nous vous recommandons d'en prendre connaissance avant d'aborder l'utilisation du logiciel.

1. CONFIGURATION

Afin de mettre en œuvre le logiciel STATISTIQUES, votre système doit comprendre au minimum les éléments suivants :

- un micro-ordinateur TO7 avec extension 16K, ou TO7 70, une cartouche BASIC, un téléviseur,
- un lecteur de disquettes,
- une disquette STATISTIQUES.
- une disquette vierge, destinée à contenir les données que vous utiliserez (disquette FICHIER).

Attention :

Si vous possédez deux lecteurs, vous devez placer la disquette STATISTIQUES dans le premier lecteur (lecteur 0), et la disquette FICHIER dans le second lecteur (lecteur 1).

Si vous ne possédez qu'un lecteur, vous devrez y placer alternativement l'une ou l'autre des deux disquettes : en effet, le programme n'est pas chargé intégralement en début d'utilisation ; mais il est divisé en plusieurs sections qui ne sont chargées que lorsque l'utilisation de STATISTIQUES le requiert.
Des messages vous indiqueront aux moments adéquats laquelle des deux disquettes doit être placée dans le lecteur :

- *Lorsque la disquette FICHIER est requise, le message ci-dessous apparaît :*
« Mettre la disquette FICHIER puis presser ENTREE ».

- *De même lorsque la disquette STATISTIQUES est nécessaire, vous verrez sur votre écran :*
« Mettre la disquette STATISTIQUES puis presser ENTREE ».

- *Il convient d'être attentif à :*

— *introduire la disquette requise dans le lecteur, étiquette vers le haut,*

— *bien fermer le loquet du lecteur,*

— *les disquettes étant tout de même fragiles, il faut les manipuler avec douceur, sans précipitation,*

— *faire attention aux messages d'ERREUR que le programme affiche en rouge en bas de l'écran lors d'une manipulation incorrecte.*
(ex. : « vérifier le lecteur », dans le cas où celui-ci n'est pas vérouillé).

Par ailleurs, vous pouvez compléter ce système par une imprimante thermique ou à impact.

2. QUELQUES PRINCIPES D'UTILISATION DES FICHIERS DE DONNEES

PRECIS I.2.
Distributionstatistique

2.1. Capacités

Chaque fichier peut contenir 30 échantillons :

— si les données sont indépendantes ou appariées, chaque échantillon peut contenir 100 données.

— si les données sont des classes, chaque échantillon peut contenir 30 classes et chaque classe un effectif égal au maximum à 1 000.

Le nombre maximal de caractères utilisables pour chaque type d'information constituant le fichier est le suivant :

titre de l'étude : 38 caractères
nombre d'échantillons : 2 caractères
libellé de l'échantillon : 9 caractères
effectif de l'échantillon : 3 caractères

2.2. Echelles

Les données composant vos échantillons sont exprimées en nombre réels, qui peuvent être écrits en puissance de 10 ; dans ce cas, l'exposant de la puissance doit être précédé de « E », et peut varier de + 38 à − 38 ; la mantisse ne peut avoir plus de sept chiffres significatifs.

Exemple :

9.10^{-4} s'écrira 9 E − 4.

Si vos données sont toutes de l'ordre de 10E + 10 ou plus, vous avez avantage à effectuer un changement d'échelle sur vos données avant de les rentrer. En effet, vous risqueriez, avec de telles données, de dépasser les capacités de la machine lors des calculs.

Exemple :

Prenons le cas d'un échantillon X, composé des données suivantes :

Echantillon X :	*6,9*	*E + 09*
	7,56	*E + 10*
	4,09	*E + 08*
	3	*E + 10*
	1,872	*E + 10*

Pour modifier l'échelle, on va utiliser le changement de variable suivant :
$Y = X/E + 10$
ce qui va permettre de traiter les données suivantes :

Echantillon Y :	*6,9*	*E – 01*
	7,56	
	4,09	*E – 02*
	3	
	1,872	

Attention :

la variance et l'écart-type de l'échantillon restent inchangés, mais pour obtenir sa moyenne, il faut remultiplier la moyenne obtenue par E + 10.

2.3. Nom du fichier

Chacun des fichiers de données que vous allez constituer doit posséder un nom, qui permet son identification, et donc sa recherche sur la disquette :

— un nom de fichier se compose de huit caractères au plus (chiffres ou lettres) et ne doit comporter ni point, ni virgule ;

— une fois le nom rentré, n'oubliez pas de le valider en appuyant sur la touche ENTREE ;

— en cours de frappe, vous pouvez corriger l'information en appuyant sur la touche [↵] . Le curseur revient alors au début de la zone, qui est complètement effacée.

2.4. Sauvegarde des données

La sauvegarde s'effectue automatiquement dès que vous avez achevé de rentrer les données constituant un échantillon. Le fichier se constitue donc sur la disquette, échantillon par échantillon, jusqu'au nombre d'échantillons que vous avez entré au départ. (cf. paragraphe 2.1.)

Attention :

— *si vous abandonnez le programme en cours de description d'un échantillon, les échantillons que vous aviez constitués précédemment sont effacés. Vous devrez reprendre complètement l'élaboration de votre fichier ;*

— *la même chose se produira si vous interrompez le programme après avoir décrit un échantillon, mais sans avoir décrit la totalité des échantillons, dont vous avez indiqué le nombre lors de l'ouverture du fichier.*

2.5. Place libre sur disquette

Toute opération sur les données, saisie, correction, ou transformation, implique une sauvegarde sur disquette.

Avant de procéder à de telles opérations, vous devrez donc vérifier qu'il vous reste suffisamment de place sur la disquette : vous utiliserez pour cela l'option « blocs libres sur disquettes », depuis le menu « commandes D.O.S. ».

Note :

S'il ne reste plus de place sur le disque pour sauvegarder un fichier, le programme vous le signalera.

3. UTILISATION DES MENUS

Au cours de l'utilisation du logiciel, vous devrez sélectionner telle ou telle de ses fonctions : cette sélection s'opère par le déplacement dans les divers sous-menus du programme. Vous trouverez ci-dessous toutes les commandes effectuables depuis un « écran menu ».

— Pour sélectionner l'un des programmes proposés par le menu affiché à l'écran, placez la fenêtre lumineuse indiquée par la main devant le programme de votre choix, en utilisant :

- les touches ↑ , ↓ et ← ,
  validez ensuite votre choix en appuyant sur la touche [ENTREE],
- la touche-chiffre correspondant au numéro du programme,
- le crayon optique : déplacez le crayon sur les numéros des rubriques. Pour valider la rubrique choisie, appuyez le crayon sur le numéro correspondant.

— Dans chaque écran-menu, la dernière option proposée correspond au menu précédent.

4. L'IMPRIMANTE

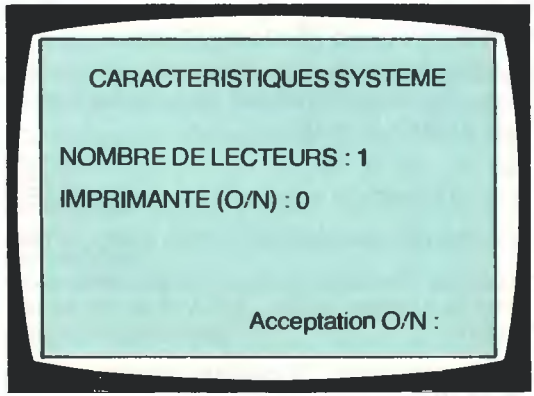
Une imprimante constituera un complément parfait pour l'utilisation de STATISTIQUES : en effet, vous pourrez obtenir une liste de vos données, et les résultats de tous les tests effectués sur celles-ci.
Ceci vous permet de visualiser sur papier tous ces éléments, et donc de ne pas « bloquer » l'écran à cette fin.

II. MISE EN SERVICE

Vous devez tout d'abord charger le programme STATISTIQUES : pour cela, sélectionnez l'option « 2 » du menu du micro-ordinateur.

— Au bout de quelques instants, la page en-tête de STATISTIQUES est affichée. Appuyez sur une touche quelconque pour relancer le chargement.

— Lors de la première utilisation de la disquette, l'écran ci-dessous est affiché : il vous permet d'indiquer les caractéristiques de votre configuration telles que le nombre de lecteurs et l'existence d'une imprimante.



Le programme va demander successivement les informations nécessaires à la bonne marche du produit, et il les contrôlera au fur et à mesure.

Il convient d'être attentif au curseur clignotant qui indique l'endroit où l'on va saisir un caractère (chiffre ou lettre).

Vous devez donc indiquer :

1. le nombre de lecteurs que vous possédez.

Celui-ci doit être compris entre « 1 » et « 4 ». Frappez au clavier le numéro choisi. Si ce nombre est correct, il s'affiche (notons que le logiciel STATISTIQUES est conçu pour une utilisation avec un ou deux lecteurs. Si vous tapez « 3 » ou « 4 », le nombre qui s'affiche sera « 2 ») sinon le programme attend toujours une réponse correcte de votre part.

2. la présence ou non d'une imprimante dans votre système :

Tapez O (oui) si vous possèdez une imprimante, (vérifiez qu'elle est en état de marche et qu'il y a du papier !), et N (non) dans le cas contraire.

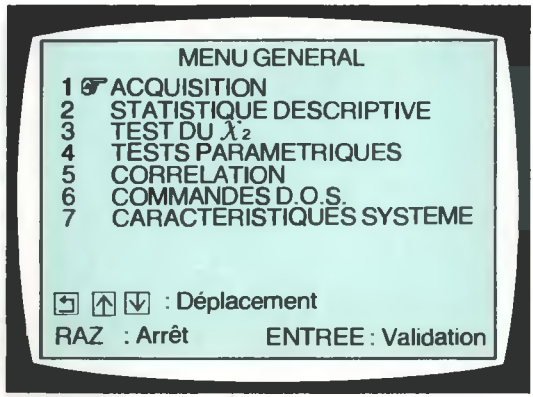
Si les renseignements que vous avez sous les yeux vous satisfont, vous pouvez valider ces informations en répondant O (oui) à : « Acceptation O/N ». Dans le cas contraire, N (non), le curseur se replace sur la première information.

— Si cette première information est correcte, appuyez sur la touche [ENTREE] ; le curseur se place alors sur la deuxième information.

— Si l'information doit être corrigée, tapez directement la valeur correcte qui remplace la précédente.

— Procédez de même pour la seconde information.

Dès que vous répondez O (oui) à la question « Acceptation O N », le MENU GENERAL du logiciel, qui vous propose le choix entre ses principaux programmes, apparaît.



Rappel :

L'appel d'un programme s'opère en plaçant la fenêtre lumineuse indiquée par la main, sur le nom du programme désiré, ceci par :

— *action des touches* ↑ , ↓ *et* ← , *validation du choix par la touche* [ENTREE],

ou par :

— *action des touches 1 à 7.*

ou par :

— *appui du crayon optique sur le numéro du programme désiré.*

Pour sortir du programme, tapez [RAZ] *; vous avez alors à votre disposition le BASIC DOS du TO7.*

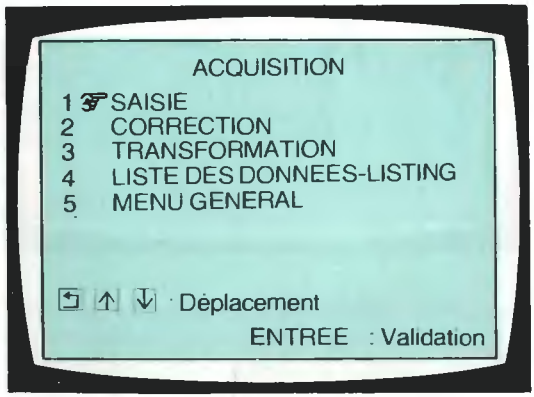
Dans la suite de çette notice, trois chapitres vous commentent dans l'ordre toutes les opérations auxquelles vous conduit la sélection d'une option du MENU GENERAL.

Nous allons ainsi passer successivement en revue l'ACQUISITION DES DONNEES préalable à tous les CALCULS STATISTIQUES, qui font l'objet d'un deuxième chapitre. Enfin, nous terminerons par les options UTILITAIRES du logiciel, qui vous permettront l'une, d'opérer sur votre disquette FICHIER (copie, destruction), l'autre, de réviser la caractérisation de votre système.

CHAPITRE DEUXIEME : ACQUISITION DES DONNEES

Dans cette partie de la notice, nous allons passer en revue toutes les procédures qui vous permettront d'indiquer au logiciel les données sur lesquelles il va effectuer divers calculs.

Pour accéder à ces procédures, sélectionnez l'option 1 du MENU GENERAL, « Acquisition ». Un nouveau menu apparaît alors :



Chacune des options de ce sous-menu représente un aspect de la constitution des données.

Pour introduire de nouvelles données, vous devrez sélectionner l'option SAISIE. Par ailleurs, deux options vous permettront d'intervenir sur des données déjà saisies :

— l'option CORRECTION vous permet de corriger les valeurs de ces données ;

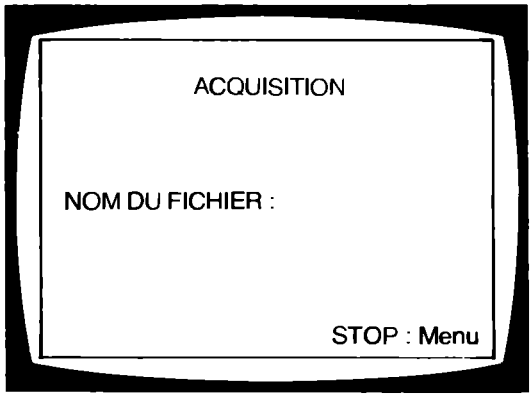
— l'option TRANSFORMATION vous permet de modifier (à l'aide de diverses fonctions algébriques) les valeurs d'un ensemble de données.

Enfin, vous pourrez consulter vos données en utilisant l'option : LISTE DES DONNEES-LISTING.

I. IDENTIFICATION DU FICHIER

Dans les quatre options du menu ACQUISITION, vous serez conduit à utiliser votre disquette FICHIER, puisque c'est sur elle que sont conservées vos données. Familiarisez-vous donc dès maintenant avec la procédure d'identification du fichier, que vous devrez employez chaque fois que vous selectionnerez une de ces options.

Dès qu'une de ces options est selectionnée depuis le menu ACQUISITION, l'écran ci-dessous apparaît :



- Si vous n'avez qu'un seul lecteur, le programme va vous demander d'introduire la disquette adéquate dans le lecteur. Un fois ceci effectué, il vous faut donner le nom du FICHIER que vous voulez saisir, corriger, transformer ou dont vous désirez obtenir la liste ou l'impression.

- Si vous accédez à cet écran depuis l'option SAISIE, le nom que vous tapez ne doit pas être celui d'un fichier déjà existant sur la disquette. Si c'est la cas, un message d ERREUR vous le signale.

  Si vous accédez à cet écran depuis les autres options du menu ACQUISITION, par contre, le nom tapé doit être celui d'un fichier existant. Dans le cas contraire, ceci vous est également signalé.

La touche [STOP] vous permet de revenir au menu ACQUISITION, en cours de saisie du nom du fichier.

II. CARACTERISTIQUES DE LA SAISIE

Nous vous indiquons ici les règles que vous devrez respecter pour remplir les écrans de données auxquels vous accèderez en sélectionnant l'une des trois premières options du menu ACQUISITION :
la saisie consiste simplement à remplir les différentes zones visualisées sur l'écran par des rectangles bleus qui repèrent l'emplacement et le nombre de caractères disponibles pour saisir une information, en utilisant les touches du clavier.

a) La saisie d'une information dans la zone qui lui correspond n'est prise en compte par le micro-ordinateur, que si elle est validée par la touche ENTREE.

b) Les caractères disponibles au clavier ne sont pas tous accessibles dans chaque zone : voici, pour chaque rubrique, la liste des caractères accessibles dans la zone qui lui correspond :

- TITRE et LIBELLE :
  tous les chiffres, signes divers, lettres majuscules et minuscules, sauf lettres accentuées.
- EFFECTIF et NOMBRE d'ECHANTILLONS :
  les chiffres de 0 à 9.
- DONNEES :
  les chiffres de 0 à 9 ; . ; + ; − ; E

Rappel :

Pour écrire un nombre avec exposant, utilisez la lettre E et les signes + −, par exemple 45 E + 10 pour « 45 multiplié par 10 puissance 10 », ou 45 E − 10 pour « 45 multiplié par 10 puissance − 10 ».

c) Si la donnée est incorrecte, différents messages peuvent apparaître en rouge sous la zone après validation de la saisie :
— « erreur syntaxe » : dans ce cas, le curseur se positionne sous la première erreur détectée,
— « nombre trop grand » ou « nombre trop petit » : ceci étant dû en général à la puissance, le curseur se place après le caractère E.

d) Lorsque toutes les zones de l'écran sont remplies, le message « Validation O/N » apparaît :

— la réponse O (oui) entraîne la validation de cet écran.

Si cet écran est le dernier d'un échantillon, celui-ci est alors sauvegardé sur la diquette FICHIER.

— si vous répondez N (non), le curseur revient dans la première zone de saisie ; en appuyant sur les touches vous pouvez le déplacer d'une zone à l'autre afin de corriger telle ou telle donnée. N'oubliez pas de valider les zones que vous aurez corrigées.

Remarque :
Vous pouvez à tout instant revenir au Menu ACQUISITION en appuyant sur la touche STOP ; mais attention : toutes les données saisies dans l'écran que vous quittez seront alors perdues.

En bref :

— La touche [ENTREE] permet de valider la frappe d'une information,

— la réponse O (oui) à « Validation O/N » vous permet de valider globalement un écran,

— la réponse N (non) vous permet de corriger chacune des données saisies dans l'écran.

— le curseur se déplace :

- à l'intérieur d'une zone par action sur les touches → et ←,
- d'une zone à la suivante (uniquement en mode correction) par action sur ↓ ou [ENTREE],
- d'une zone à la précédente (uniquement en mode correction) par action sur ↑,
- d'une zone au début de l'écran de saisie par ⮌ (uniquement en mode corection).

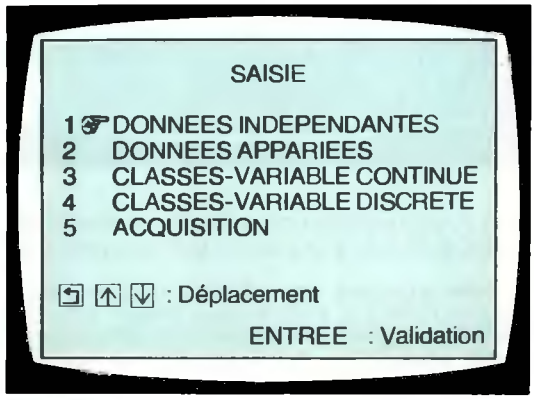
— d'autres commandes permettent :

- d'insérer un caractère [INS]
- d'effacer un caractère [EFF]
- d'abandonner en cours [STOP]

— Attention aux messages d'ERREUR qui apparaissent en rouge.

III. MENU SAISIE

En choisissant SAISIE dans le menu ACQUISITION, vous obtenez :



Chacune des options de ce menu SAISIE vous permet d'accéder à la saisie d'un type particulier de données : pour la différence de signification entre ces types de données reportez-vous au document PRECIS DE STATISTIQUES.

Quelque soit le type de données que vous sélectionnez, vous accédez à un premier écran, dans lequel vous devez indiquer successivement :

— le titre de l'étude,

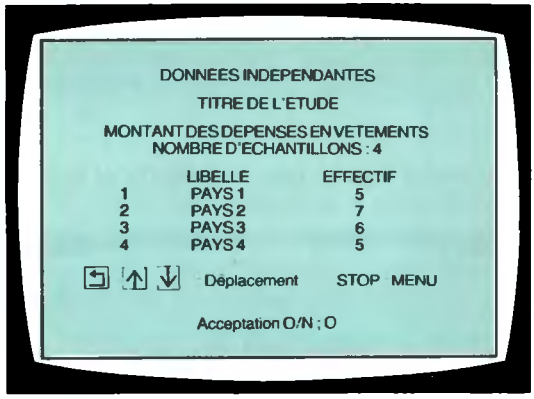
— le nombre d'échantillons à sauvegarder dans le fichier,

— le libellé de chacun de ces échantillons ainsi que leur effectif (pour les classes, cet effectif est en fait le nombre de classes de chaque échantillon).

Dans tous les écrans de saisie, il vous est rappelé le type des données et bien entendu le nom de l'étude.

Remarque :

Le titre de l'étude et le nombre d'échantillons ne peuvent être corrigés directement ; pour cela, il faut sélectionner l'option CORRECTION depuis le menu ACQUISITION.



Dans un écran, vous pouvez remplir les libellés et effectifs de six échantillons.

Pour des données appariées, une précision importante est à ajouter :

ce type de données implique que tous les échantillons d'un même fichier soient de tailles égales. En effet, rappelons que des échantillons de données appariées sont composés du même ensemble de n sujets sur lesquels on a effectué des mesures différentes.

Pour vous éviter de remplir les zones EFFECTIF pour chacun de ces échantillons, celles-ci se remplissent automatiquement dès que l'effectif n° 1 est rentré.

Pour corriger ces effectifs, (après avoir répondu N (non) à « Validation O/N », ou sélectionné l'option CORRECTION), il vous suffira de corriger le premier, afin que tous les autres se corrigent automatiquement.

Une fois ces informations saisies, vous allez rentrer les données proprement dites de chaque échantillon.

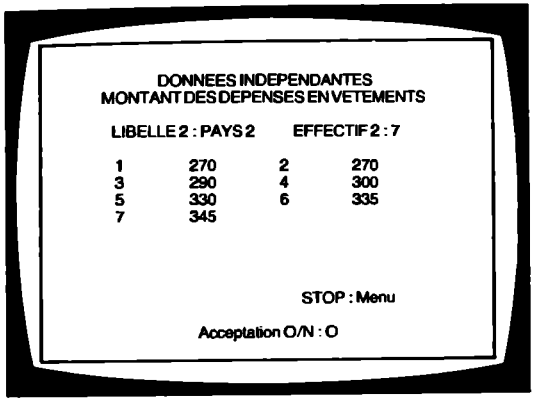
Le libellé et l'effectif de chacun d'entre eux vous est rappelé.

1. DONNEES INDEPENDANTES ET APPARIEES

Dans un écran, vous pouvez indiquer douze données.

Exemple :

— un échantillon de quatorze données se répartit de la façon suivante :
- *le premier écran contient les douze premières données ;*
- *dès que vous avez validé le premier écran, un nouvel écran est disponible, dans lequel vous inscrivez les deux dernières données.*



2. CLASSES-VARIABLE CONTINUE OU VARIABLE DISCRETE

Comme nous vous l'avons indiqué dans le « PRECIS DE STATISTIQUES », nous avons choisi de caractériser chaque classe par :

— sa borne inférieure

— sa borne supérieure

— son effectif

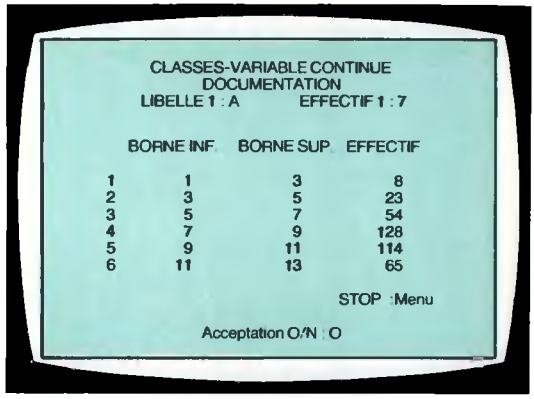
Les valeurs « borne inférieure » et « borne supérieure » doivent être rentrées en ordre croissant. Dans le cas contraire, un message d'erreur apparaît sous la zone de saisie incorrecte.

- Dans le cas d'une variable continue, il est recommandé de faire des classes consécutives, à bornes communes.
- Dans le cas d'une variable discrète, la borne inférieure et la borne supérieure sont confondues. Pour éviter à l'utilisateur de rentrer deux fois le même nombre, le programme recopie automatiquement dans la zone « borne supérieure », la valeur entrée pour la « borne inférieure ».

Un écran contient six classes (soit dix-huit données).

Exemple :

— un échantillon de huit classes se répartit de la façon suivante :
- *le premier écran contient les six premières classes ;*
- *dès que le premier écran est validé, un nouvel écran est disponible, dans lequel vous inscrivez les deux dernières données.*



Une fois tout ceci effectué, vos échantillons sont sauvegardés sur la disquette FICHIER, sous le nom que vous avez indiqué en début de SAISIE.

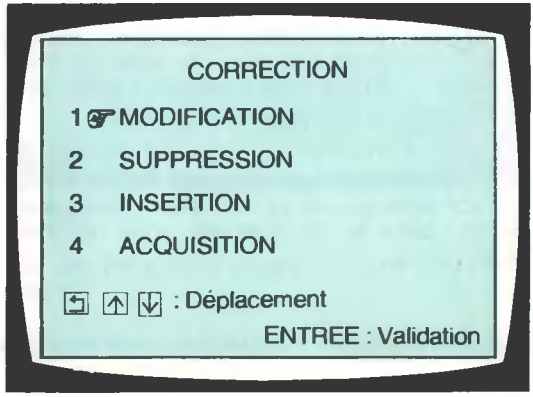
Si ces données sont importantes n'hésitez pas, à l'aide de COPIE FICHIER dans la rubrique COMMANDES DOS du MENU GENERAL à les sauvegarder sur plusieurs disquettes FICHIER (voir la quatrième partie de cette notice).

Si vous désirez apporter certaines modifications à vos données, supprimer ou au contraire ajouter un échantilon ou une donnée dans un échantillon, lisez attentivement le chapitre ci-dessous avant de sélectionner l'option CORRECTION du menu ACQUISITION.

IV. MENU CORRECTION

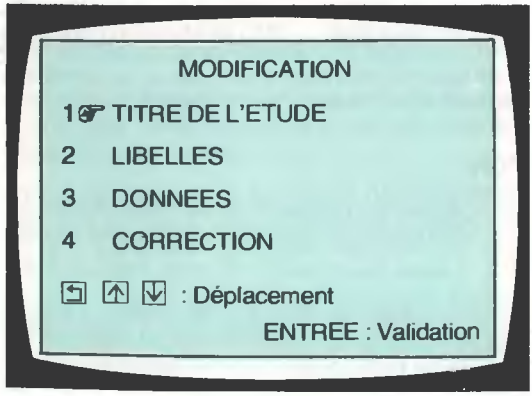
Vous avez, comme nous vous l'avons indiqué plus haut, entré le nom du fichier que vous désirez corriger.

Le menu CORRECTION vous propose le choix entre plusieurs types de correction :



1. MODIFICATION

Celles-ci peuvent porter sur le titre de l'étude, les libellés des échantillons ou les données elles-mêmes.



a) Titre de l'Etude

L'ancien titre vous est représenté dans une zone de saisie, le curseur en première position, il vous suffit de corriger ce titre grâce aux touches habituelles de la saisie.

b) Libellés

Un écran contient six libellés. Pour chaque libellé, il vous est rappelé l'effectif de l'échantillon.
Pour corriger un libellé dans cet écran, répondez N (non) à la question « Validation O/N » : le curseur se place en début de zone du premier libellé ; positionnez-le dans la zone à corriger à l'aide des touches de déplacement du clavier et effectuez vos corrections.

Par contre, si vous répondez O (oui), vous confirmez les dorinées présentes.
Si votre fichier comporte plus de six échantillons, vous obtenez alors les libellés suivants ; dans le cas contraire, vous retournez au menu CORRECTION.

c) Données

Avant de corriger une donnée, il faut préciser dans quel échantillon elle se situe.
Le programme vous pose donc la question. Si vous indiquez un numéro d'échantillon qui n'existe pas, un message d'erreur vous rappelle le nombre d'échantillons contenu dans ce fichier.

Ensuite les données de l'échantillon choisi vous sont proposées par écran de douze pour les données indépendantes ou appariées, ou par dix-huit (ou six classes) pour les classes.

Opérez de la même manière que dans la rubrique LIBELLE.

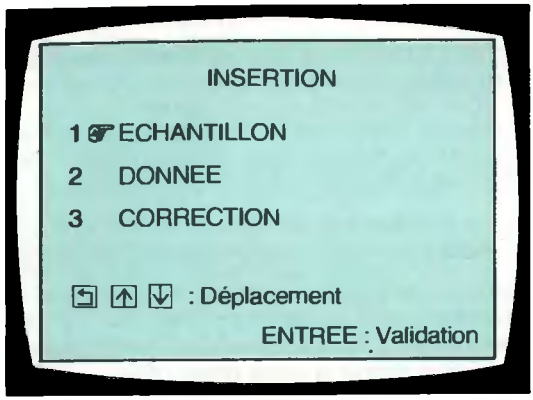
Les corrections dans les classes doivent conserver l'ordre croissant entre les valeurs, sinon un message d'erreur apparaît.

Notons encore, que pour les classes-variable discrète, il n'y a que la modification de la borne inférieure d'une classe qui entraîne la modification du couple borne inférieure — borne supérieure.

2. INSERTION

Vous pouvez, soit insérer un échantillon complet entre deux échantillons déjà existants, soit insérer une donnée entre deux autres dans un échantillon du fichier.

Attention :
Vous ne pouvez effectuer une insertion que si votre fichier (ou votre échantillon) n'est pas plein. Dans le cas contraire, un message d'erreur apparaît.



Dans les deux cas, il vous faut indiquer le numéro de l'échantillon qui vous intéresse.

a) Echantillon

- Le numéro que vous indiquez est celui que va avoir votre nouvel échantillon.

Par exemple :

Votre fichier comporte déjà trois échantillons, vous désirez en insérer un entre le premier et le deuxième, vous donnez donc comme numéro : 2.

- Vous devez maintenant préciser le libellé et l'effectif de votre nouvel échantillon.

Tous les libellés de votre fichier vous sont présentés par six (vous devez commencer à avoir l'habitude de cette présentation : cf. SAISIE — MODIFICATION).

Au numéro du nouvel échantillon correspond une zone vide pour le libellé, ainsi que pour l'effectif (sauf pour les données appariées où l'effectif est automatiquement rempli).

Si votre nouvel échantillon ne se trouve pas dans le premier écran qui vous est présenté, répondez O (oui) au message « Validation O/N » pour passer au suivant. Dans le cas contraire, répondez N (non), grâce aux touches de déplacement, allez remplir les zones vides.

- Une fois libellé et effectif saisis, l'écran de saisie de données de votre échantillon apparaît. Vous n'avez plus qu'à le remplir. Pour plus de précisions, reportez-vous à la rubrique SAISIE de DONNEES.
- Sitôt l'opération achevée, le menu CORRECTION est affiché.

b) Données

- Le numéro de l'échantillon que vous indiquez est celui dans lequel vous désirez insérer une donnée.
- Où voulez-vous insérer cette nouvelle donnée ?
  Le numéro que vous allez préciser est celui que va prendre celle-ci.
- Toutes les données de votre échantillon vous sont présentées (par douze ou dix-huit, suivant le type de vos données).

Au numéro de la nouvelle donnée correspond une ou trois zones vides (une pour données appariées ou indépendantes, trois pour classes).

Si votre donnée ne se trouve pas dans le premier écran qui vous est présenté, répondez O (oui) au message « Validation O/N » pour passer au suivant. Dans le cas contraire, répondez N (non) et grâce aux touches de déplacement allez remplir la ou les zones vides.
(Attention : respectez l'ordre croissant pour les classes).

Bien entendu, si vous vous apercevez en même temps que certaines données sont erronées, n'hésitez pas à les corriger (pour plus de précisions, reportez-vous au chapitre CORRECTION). Il vous suffit de répondre N (non) à la question « Validation O/N » de l'écran où vous avez décelé une erreur et de la corriger en vous aidant des touches de déplacement.

- Sitôt l'opération achevée, le menu CORRECTION est affiché.

Attention :

N'oubliez pas que les échantillons d'un même fichier dont les données sont appariées doivent tous avoir le même effectif.
Donc, si vous décidez d'insérer une donnée dans un fichier de ce type, procédez à une insertion dans chaque échantillon du fichier, afin que les effectifs restent égaux. Sinon vos résultats seront erronés.

3. SUPPRESSION

Vous pouvez supprimer tout un échantillon ou simplement une donnée dans un échantillon.

— Si un fichier ne contient qu'un seul échantillon et que vous décidiez de le supprimer, l'opération ne se fera pas.
Il est en effet plus simple dans ce cas de se servir de l'instruction DESTRUCTION FICHIER de la rubrique COMMANDES DOS.

— Si un échantillon ne contient qu'une donnée et que vous décidiez de la supprimer, c'est l'échantillon lui-même qui sera supprimé, sauf dans le cas cité ci-dessus.

a) Echantillon

Le numéro de l'échantillon à supprimer vous est demandé. L'opération s'effectue. Puis on revient au menu SUPPRESSION.

b) Donnée

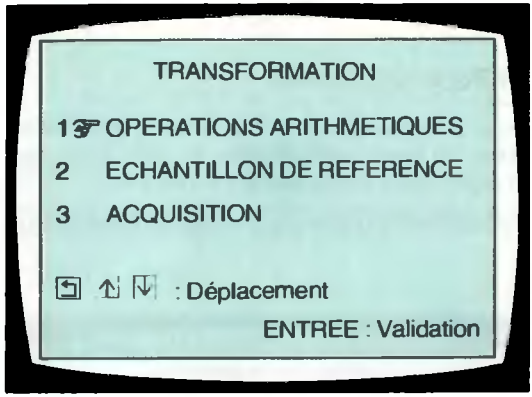
Le numéro de l'échantillon, puis celui de la donnée vous sont demandés, l'opération s'effectue, puis on revient au menu SUPPRESSION.

Attention

N'oubliez pas que les échantillons d'un fichier dont les données sont appariées doivent tous avoir le même effectif.
Donc, si vous décidez de supprimer une donnée dans un fichier de ce type, supprimez une donnée dans chaque échantillon, afin que les effectifs restent égaux, sinon vos résultats seront erronés.

V. MENU TRANSFORMATION

Maintenant que votre fichier est correct, vous pourrez faire subir à vos données diverses transformations, ceci en sélectionnant l'option 4 du menu ACQUISITION :



Comme d'habitude, le nom du fichier sur lequel vous désirez effectuer des transformations vous est demandé. Nous appellerons ce fichier le fichier des données brutes.

Vos données brutes ne seront pas détruites par les données transformées, car toute transformation donne lieu à la constitution d'un nouveau fichier, le fichier des données transformées, dont vous devez donner le nom.

Si celui-ci est le nom d'un fichier déjà existant sur la disquette, un message d'erreur vous le signale, et il ne vous reste plus qu'à en donner un autre.

Rappelons que, comme pour la saisie, la sauvegarde va se faire après la transformation de chaque échantillon, ceci pour vous permettre de constituer un plus grand nombre d'échantillons par fichier.

Pour contrôler le résultat de vos transformations, sélectionnez la rubrique LISTE DES DONNEES — LISTING (voir ci-après).

Attention :

Le fichier des données brutes et le fichier des données transformées sont nécessairement sur la même disquette FICHIER.

Assurez-vous, avant d'effectuer une transformation, qu'il y a assez de place sur la disquette. Si ce n'est pas le cas, en cours de transformation,

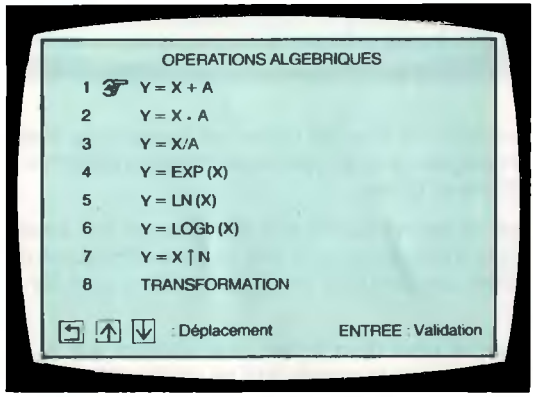
un message d'erreur vous signalera le manque de place et le fichier ne sera pas transformé.

Vous n'aurez plus qu'à faire de la place sur la disquette en détruisant des fichiers dont vous n'avez plus l'usage ou encore en recopiant le fichier de données brutes sur une disquette moins chargée, ceci grâce aux différentes rubriques du menu COMMANDES DOS.

1. OPERATIONS ALGEBRIQUES

Après que vous avez indiqué le nom du fichier à transformer et celui du nouveau fichier, un menu vous propose le choix entre plusieurs types d'opérations algébriques à effectuer sur les données.

Les opérations sont effectuées sur toutes les données de tous les échantillons.



- Comme nous vous en avons déjà informé dans la rubrique ECHELLES, p. 51, l'ordinateur a des limites, et les réels calculés doivent être compris entre − 1,7 E − 38 et 1,7 E + 38. Or une opération telle que l'exponentielle, par exemple, peut atteindre assez vite ces limites.

  Dans ce cas, un message d'erreur apparaît pour vous annoncer que la transformation que vous souhaitez ne peut se faire. Le fichier n'est donc pas créé.

- D'autre part, il faut se souvenir que les classes d'effectifs doivent être rangées en ordre croissant. Les bornes d'une classe sont des réels positifs en général, donc si vous divisez par un nombre négatif ou si vous élevez à une puissance négative, votre ordre se trouve changé. Les classes sont rangées dans l'ordre décroissant.

Si vous voulez faire de telles opérations sur des classes, un message d'erreur apparaîtra pour vous rappeler qu'il faut préserver l'ordre croissant et que la transformation est donc possible. Vous retournez alors au menu ACQUISITION.

■ Voici maintenant la liste des fonctions que vous pouvez utiliser pour transformer vos données :

a) Y = X + A ; Y = X * A ; Y = X/A

La constante A est un réel de 12 caractères. Elle sera donc ajoutée, multipliée ou elle divisera toutes les données de tous les échantillons du fichier, sans exception.

A vous de la préciser.

Une fois la constante rentrée, la transformation s'effectue et ceci vous est précisé par le message CALCULS EN COURS.

b) Y = EXP (X) ; Y = Ln (X)

Cette fois aucune constante à rentrer puisque vous venez de choisir de calculer l'exponentielle ou le logarithme népérien de vos données.

c) Y = LOGb (X)

b désignant la base du logarithme, c'est à vous de la choisir.

Evidemment b doit être supérieur à 0, sinon la transformation sera impossible.

d) Y = X ↑ N

Toutes vos données sont élevées à la puissance N, N étant un réel.

Bien entendu, si certaines de vos données sont négatives et que vous essayiez de les élever à la puissance 0.5 (racine carrée), un message d'erreur apparaîtra.

■ Une fois les données transformées et stockées sur disquette dans un fichier, vous devrez créer un troisième fichier si vous désirez transformer à nouveau ces données. Ainsi toutes les étapes de vos transformations sont sauvegardées ; vous avez la possibilité de détruire celles qui ne vous intéressent plus, en utilisant l'option DESTRUCTION FICHIER dans le menu COMMANDES DOS.

2. ECHANTILLON DE REFERENCE

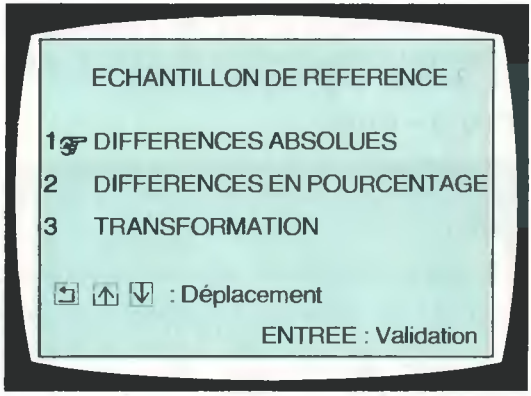
Cette option vous permet d'utiliser un échantillon que vous aurez sélectionné dans le fichier à transformer. Ses données opéreront sur les données des autres échantillons du fichier, pour les transformer.

Le fichier des données transformées contiendra un échantillon de moins (l'échantillon de référence), que le fichier des données brutes.

Cette transformation ne s'effectue que sur les données indépendantes ou les données appariées. Si vous essayez d'effectuer cette transformation sur des échantillons de classes, un message d'erreur apparaît et le menu ACQUISITION est affiché. De même, si vos échantillons n'ont pas tous le même effectif.

Si votre fichier ne contient qu'un seul échantillon, il y a peu de chances que la transformation s'effectue.

Par contre, si votre fichier respecte toutes les conditions requises, deux types d'opérations vous sont proposés.



Dans les deux cas, le numéro de l'échantillon de référence vous est demandé. Si vous demandez un numéro d'échantillon inexistant dans le fichier, un message d'erreur vous donne le nombre d'échantillons contenus dans le fichier. Il ne vous reste plus qu'à entrer un numéro correct.

a) Différences absolues

Les échantillons ayant tous le même nombre de données, chaque donnée de l'échantillon de référence est retranchée à la donnée de rang correspondant dans chacun des autres échantillons du fichier.

b) Différences en pourcentage

Les nouvelles données sont égales au rapport de la différence entre les anciennes données et les données de l'échantillon de référence, par les données de l'échantillon de référence, le tout multiplié par 100, d'où le nom de pourcentage.

$$\frac{\text{donnée} - \text{donnée éch. réf.}}{\text{donnée éch. réf.}} \; 100$$

VI. LISTE DES DONNEES — LISTING

Vous pouvez à tout moment visualiser sur écran la totalité des données d'un fichier ou même les imprimer sur listing (dans le cas, bien sûr, où vous avez une imprimante et si vous avez rempli en conséquence la rubrique CARACTERISTIQUES SYSTEME).

Lorsque vous sélectionnez cette rubrique, vous obtenez tout d'abord l'écran où vous sont rappelés le type des données, le titre de l'étude, le nombre d'échantillons ainsi que leurs libellés et effectifs.

Trois commandes vous sont alors proposées (vous les retrouverez dans chaque nouvel écran).

- [ACC] : Acceptation

Lorsque vous avez fini de consulter les données affichées dans cet écran, cette commande vous permet de passer à l'écran suivant. Arrivé au dernier, vous obtenez la possibilité de tirer un listing.

- [RAZ] : Listing

Peut-être ne désirez-vous pas voir toutes vos données défiler à l'écran, mais par contre obtenir un listing rapidement.
Bien entendu, si vous avez spécifié que vous ne possédiez pas d'imprimante, vous retournerez directement au menu ACQUISITION.

- [STOP] : Menu

Cette commande vous permet de revenir à un menu, sans possibilité d'obtenir un listing.

Libellés, effectifs et données sont visualisés de la manière suivante :
Peuvent être visualisés simultanément :

— six libellés

— six effectifs

— douze données (données indépendantes ou appariées)

— dix-huit données (classes).

Attention :

Avant de lancer un listing, assurez-vous que votre imprimante est bien connectée à l'ordinateur et sous tension.

Pensez également à régler le cadrage du papier (ceci consiste à placer manuellement une pliure de papier juste au-dessus de la tête d'impression de l'imprimante).

Vous obtenez le listing complet de votre fichier et pas seulement une copie de l'écran à partir duquel vous avez appelé le listing.

CHAPITRE TROISIÈME :
CALCULS STATISTIQUES

Nous allons aborder la partie STATISTIQUES proprement dite, puisque vous avez maintenant sur votre disquette FICHIER des données prêtes à être traitées. Rappelons encore que tous les résultats de ces tests peuvent être imprimés sur listing.

Pour toutes les parties qui vont suivre, reportes-vous au PRECIS DE STATISTIQUES pour de plus amples informations mathématiques. Certains tests ne s'appliquent pas à tous les types de données. Dans ce cas, celles-ci vous seront précisées.

• La procédure d'identification du fichier sur lesquels les calculs vont être effectués est la même que dans le programme ACQUISITION :

— indiquez le nom du fichier sur lequel vous désirez effectuer des tests.
— une fois le fichier retrouvé par le logiciel, confirmez sa sélection en répondant O (oui), ou indiquez un autre nom de fichier après avoir répondu N (non).

Si les calculs dépassent les capacités de la machine, le message CALCULS IMPOSSIBLES apparaît et un nouveau nom de fichier vous est demandé.

• Dans chacun des écrans de présentation des résultats de calcul, vous trouverez les mêmes commandes, [STOP], [ACC], [RAZ], qui vous permettront respectivement :

[STOP] : le retour au MENU GENERAL.

[ACC] : le passage à l'écran suivant.

[RAZ] : l'accès au listing dont vous devez confirmer ensuite l'exécution en répondant au message : « Désirez-vous un listing O/N ».

Remarques :

— *Le listing vous est proposé automatiquement lorsque vous avez achevé la visualisation (par [ACC]) de la totalité des écrans présentés pour un type de calcul.*

— *La commande [RAZ] peut avoir une fonction particulière, qui sera détaillée dans le commentaire de l'option concernée.*

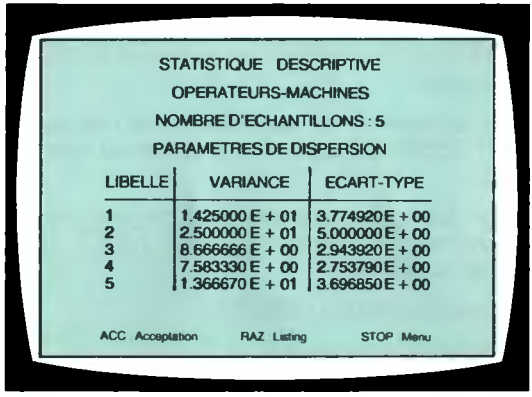
PRECIS II et IV
— **Paramètres caractéristiques d'une distribution**
— **Estimation**

I. STATISTIQUE DESCRIPTIVE

Ce programme calcule pour chaque échantillon de votre fichier : moyenne, médiane, variance, écart-type, écart-moyen, quartiles inférieur et supérieur, écart inter-quartile, quelque soit le type des données, (huit rubriques).

Les résultats vous sont donnés sous forme réelle, c'est-à-dire qu'ils sont exprimés par des nombres avec ou sans partie fractionnaire, avec ou sans exposant (celui-ci est précédé de la lettre E).

Ils sont présentés dans des tableaux. Chacun est constitué de cinq échantillons au plus et propose les résultats de deux rubriques à la fois, par exemple : variance et écart-type.



COMMANDES

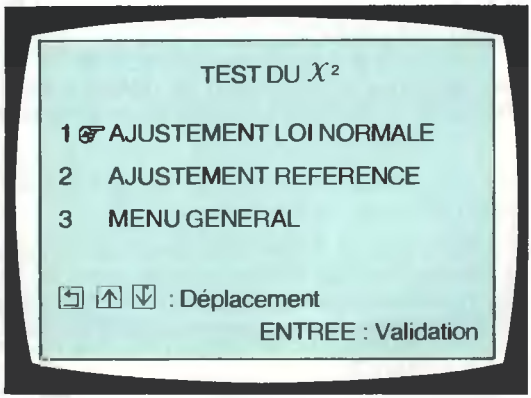
RAZ : Dans cette option, la commande RAZ vous permet d'accéder aux rubriques suivantes, sans passer systématiquement en revue les tableaux de la rubrique précédente (par la touche ACC) Lorsque vous avez passé en revue toutes les rubriques, le listing vous est proposé.

Attention :

Dans le cas des classes-variable discrète, médiane, quartiles inférieur et supérieur, écart inter-quartile n'ayant aucun sens en général, ils ne sont donc pas calculés. Dans vos tableaux de résultats, vous ne relirez que des 0.

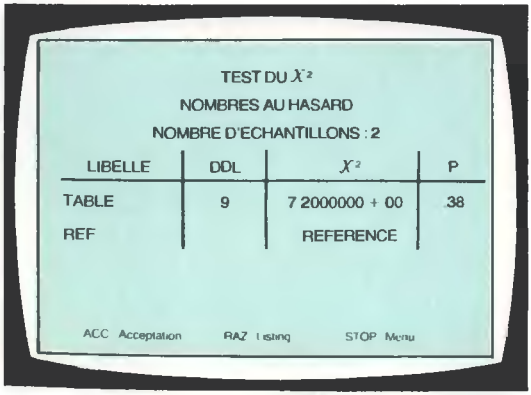
II. TEST DU χ^2

Le test du χ^2 permet de se raccorder à deux types de loi de distribution théorique différents, que vous aurez sélectionnés à partir d'un menu :



Dans les deux cas, les résultats sont présentés sous forme de tableau.

Pour chaque échantillon est indiqué : son degré de liberté (DDL), son χ^2, et la probabilité correspondante. Chaque tableau contient les indications concernant six échantillons au maximum. Le χ^2 est exprimé sous forme réelle.



1. AJUSTEMENT LOI NORMALE

Ce test, comme son nom l'indique, vous permet d'infirmer ou de confirmer la normalité d'une variable continue à partir de la distribution observée, pour chacun des échantillons.

Ce test s'applique donc uniquement aux classes-variable continue : si votre fichier contient un autre type de données, un message d'erreur apparaît et le menu est de nouveau présenté.

2. AJUSTEMENT REFERENCE

Dans le fichier se trouve un échantillon de référence dans lequel la distribution théorique est décrite. Le numéro de cet échantillon vous est demandé.

Ce test s'applique aux deux types de classes. C'est-à-dire classes-variable continue et classes-variable discrète.

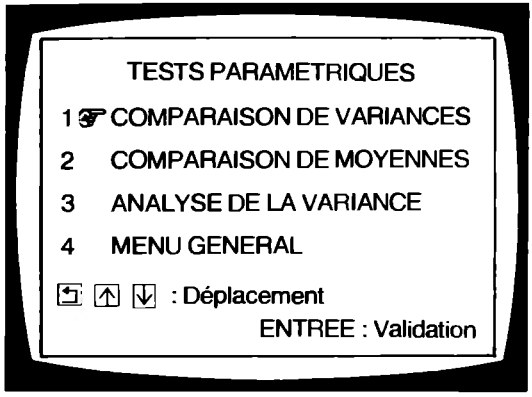
Puisque les échantillons du fichier sont comparés à l'échantillon de référence, ils doivent avoir les mêmes bornes, sinon les calculs sont impossibles, le message d'erreur CLASSES DIFFERENTES apparaît.

Dans le tableau des résultats, l'échantillon de référence vous est rappelé par la mention REFERENCE.

PRECIS V.2. et V.3
— Tests parametriques
— Analyse de la variance

III. TESTS PARAMETRIQUES

Deux sortes de tests vous sont proposés : des tests de comparaison et des analyses de variance.

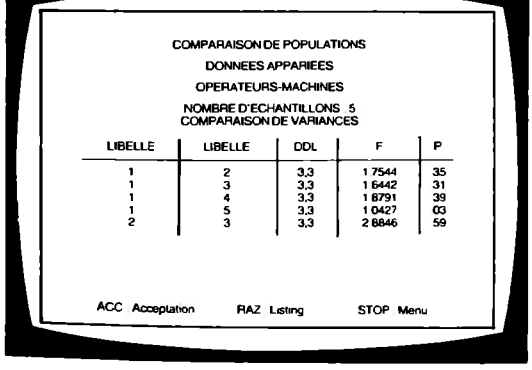


1. COMPARAISON DE VARIANCES ET DE MOYENNES

Ces deux tests s'appliquent à tous les types de données. Rappelons que la comparaison de variances revient à faire un test de Snédécor et la comparaison de moyennes, à faire un test de Student.

Tous les échantillons de votre fichier sont comparés deux à deux. Donc si votre fichier contient n échantillons, il y a (n (n − 1))/2 comparaisons.

Les tableaux de résultats vous proposent cinq comparaisons à la fois.



2. ANALYSE DE LA VARIANCE

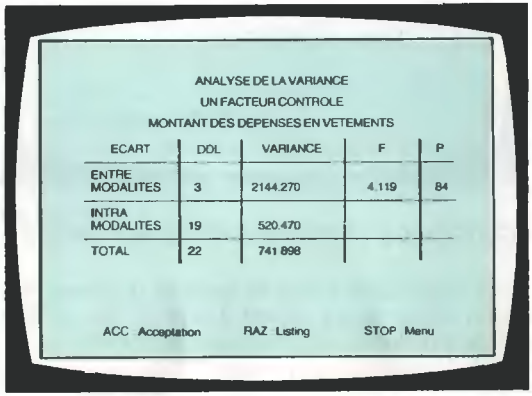
Ce test permet de déterminer l'influence d'un ou de deux facteurs sur une variable X.

Dans le cas de données indépendantes, l'analyse est à une voie (ou un facteur contrôlé). Dans le cas de données appariées, l'analyse est à deux voies (ou deux facteurs contrôlés).

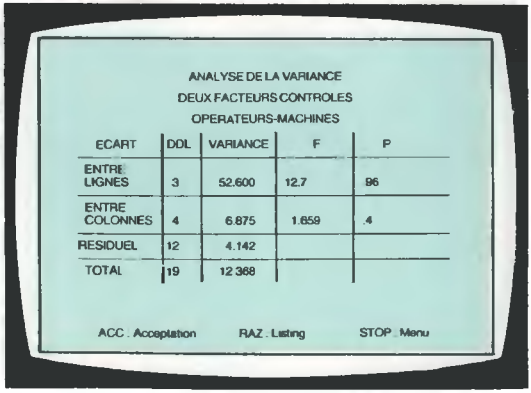
Ce test ne s'applique donc pas aux classes.

Dans les deux cas, tous les échantillons de votre fichier sont pris en compte pour le calcul, et les résultats sont réunis dans un même tableau :

Un facteur contrôlé



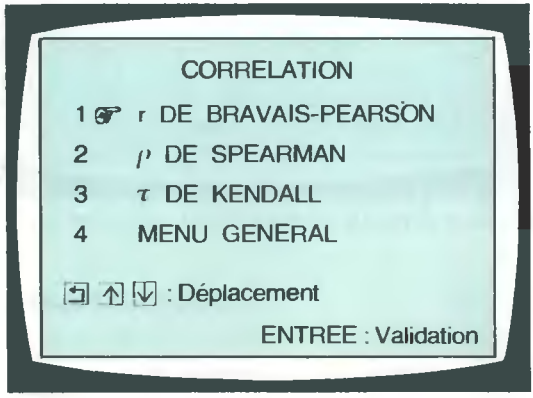
Deux facteurs contrôlés



IV. CORRELATION

PRECIS VI. Correlation

Ce programme vous propose trois calculs de corrélation. Le coefficient de corrélation linéaire usuel ou coefficient de Bravais-Pearson et deux coefficients de corrélation qui se calculent sur les rangs des données, le coefficient de Kendall et le coefficient de Spearman.



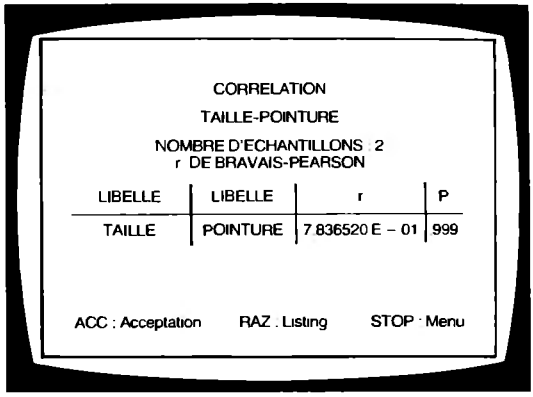
Ces trois tests ne s'appliquent pas aux classes.

Les échantillons du fichier sont comparés deux à deux. Donc si vous avez n échantillons, cela représente (n (n − 1))/2 comparaisons.

Les résultats (coefficients et probabilités associées) vous sont présentés par tableau de cinq.

Attention, tous les échantillons doivent avoir le même effectif. Si ce n'est pas le cas, les calculs ne se seront pas effectués et le message d'erreur « Effectifs différents » apparaîtra.

Rappelons encore que pour que les tests soient significatifs, il faut au moins dix données par échantillon.


CORRELATION
TAILLE-POINTURE
NOMBRE D'ECHANTILLONS 2
r DE BRAVAIS-PEARSON
LIBELLE
LIBELLE
r
P
TAILLE
POINTURE
7 836520 E – 01
999
ACC : Acceptation
RAZ : Listing
STOP : Menu

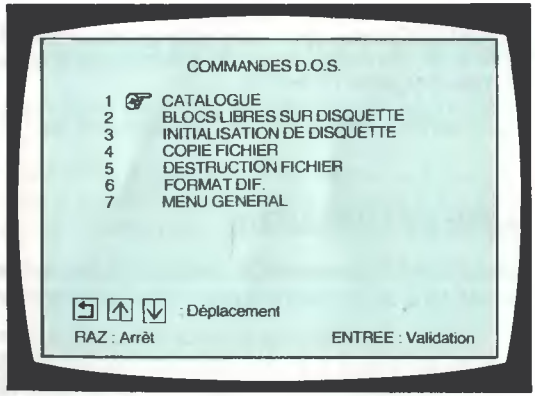
CHAPITRE QUATRIÈME : UTILITAIRES

Dans ce dernier chapitre du manuel, nous allons détailler les deux dernières options du MENU GENERAL, l'option COMMANDES DOS et l'option CARACTERISTIQUES SYSTEME.

L'option COMMANDE DOS rassemble toutes les opérations que vous pouvez effectuer sur la disquette FICHIER, depuis la visualisation de son contenu jusqu'à la destruction d'un fichier.

L'option CARACTERISTIQUES SYSTEME vous permet de modifier les indications que vous avez entrées lors de la première utilisation.

I. LES COMMANDES DOS



Rappel :

- Toutes ces commandes (mise à part la septième qui permet de revenir au MENU GENERAL) s'opèrent sur les disquettes FICHIER (si vous n'avez qu'un seul lecteur, il vous faudra mettre alternativement dans le lecteur l'une ou l'autre disquette).

Lorsque la disquette FICHIER est requise, le message ci-dessous apparaît :

Mettre la disquette FICHIER
puis presser ENTREE

De même, lorsque la disquette STATISTIQUES est nécessaire, vous verrez sur votre écran :

Mettre la disquette STATISTIQUES
puis presser [ENTREE]

Il convient d'être attentif à :

— *mettre la disquette requise ;*

— *bien fermer le loquet du lecteur ;*

— *les disquettes étant tout de même fragiles, il faut les manipuler avec douceur, sans précipitation ;*

— *faire attention aux messages d'ERREUR que le programme affiche en rouge en bas de l'écran lors d'une manipulation incorrecte. (Ex. : « Vérifier le lecteur » dans le cas où celui-ci n'est pas verrouillé.)*

1. CATALOGUE

Comme son nom l'indique, cette commande vous permet d'obtenir la liste des fichiers de données, issus du programme STATISTIQUES, présents sur votre disquette FICHIER.

Pour revenir au menu COMMMANDES DOS, appuyez sur [STOP].

2. BLOCS LIBRES SUR DISQUETTE

Votre disquette FICHIER contient 80 k-octets ou 80 blocs. Cette instruction vous permet donc de connaître la place libre sur votre disquette.

Voici un petit formulaire pour que vous puissiez calculer la taille de votre fichier, et ainsi voir si celui-ci peut être sauvegardé sur votre disquette :

— si votre fichier est composé de données indépendantes ou appariées (voir ACQUISITION), soit N l'effectif du plus grand échantillon. Votre fichier aura :

$$120 \times N + 1\,000 \text{ octets}$$

— si votre fichier est composé de classes d'effectifs (voir ACQUISITION), soit N l'effectif du plus grand échantillon. Votre fichier aura :

$$360 \times N + 1\,000 \text{ octets}$$

3. INITIALISATION DE LA DISQUETTE

Toute disquette vierge doit être initialisée avant d'enregistrer quoi que ce soit pour la première fois. Attention, cette opération détruit le contenu de la disquette si celle-ci n'est pas vierge. Il faut donc manipuler cette instruction avec précaution.

4. COPIE FICHIER

Cette commande vous permet de :

— copier sur la même disquette un fichier de données sous un autre nom ;
— copier sur une autre disquette un fichier sous le même nom.

Le nom du fichier à copier vous est demandé.

Celui-ci comprend huit caractères (chiffres ou lettres) au plus et ne doit comporter ni « . », ni « , ».

Une fois le nom rentré, validez par [ENTREE].

En cours de frappe, il est possible de corriger l'information en tapant sur la touche [↰] . Dans ce cas, le curseur se déplace en début de zone qui est redevenue vierge.

Une frappe en cours peut être abandonnée en appuyant sur la touche [STOP].

Si le nom du FICHIER que vous avez donné ne correspond pas à un fichier existant sur la disquette FICHIER, un message d'ERREUR apparaît : « Ce fichier n'existe pas », et un autre nom vous est demandé.

Lorsque le nom du FICHIER est correct, vous avez le choix entre les deux formes de « copie » :

— sur la même disquette sous un autre nom,

— sur une autre disquette sous le même nom,

ceci en répondant par OUI ou par NON à la question :

« Sur une autre disquette O/N »

● Si vous désirez copier sur une autre disquette votre FICHIER de données, répondez O (oui). Vous avez à portée de la main une disquette qui est appelée disquette DESTINATION, et qui aura été préalablement initialisée si elle est vierge. La disquette sur laquelle se trouve votre fichier est appelée disquette SOURCE.

Si le fichier est assez gros, il va vous falloir insérer dans le lecteur alternativement la disquette SOURCE et la disquette DESTINATION, un certain nombre de fois. Celles-ci vous sont demandées tour à tour.

Dans le cas d'un système à deux lecteurs, vous laissez la disquette STATISTIQUES dans le premier, et c'est dans le deuxième que vous insérez successivement les disquettes SOURCE et DESTINATION.

- Si vous désirez copier sur la même disquette votre fichier mais sous un autre nom, répondez N (non), celui-ci vous est alors demandé.

Si le nom que vous donnez alors est le nom d'un fichier déjà existant sur la disquette, le message d'ERREUR apparaît :

« Ce fichier existe déjà »

Un nouveau nom vous est alors demandé.

Lorsque tout est correct, la copie s'opère, vous pouvez alors redemander une autre copie. Sinon, pour revenir au menu précédent : COMMANDES DOS, appuyez sur [STOP].

5. DESTRUCTION DE FICHIER

Si un fichier de données ne vous est plus utile, vous pouvez le supprimer de la disquette et ainsi récupérer de la place.

Le nom du fichier à détruire vous est demandé (voir les spécifications du NOM dans copie FICHIER).

Attention, cette commande est irréversible. Une fois le fichier détruit, vous ne pouvez plus le récupérer.

6. FORMAT - DIF

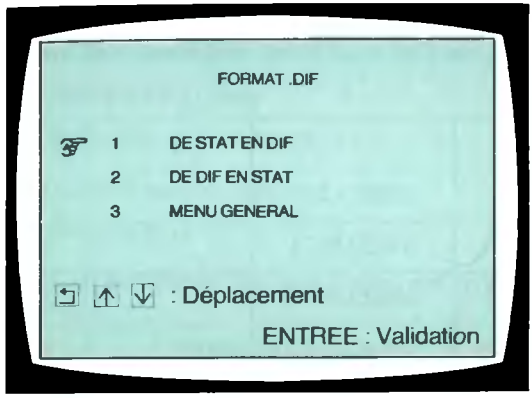
Cette rubrique vous permet de communiquer avec les autres logiciels de la collection professionnelle GRAPHIQUE, COLORCALC..., par l'intermédiaire de fichiers stockés sous format DIF.

Le transfert peut s'effectuer dans les deux sens :

— transformation d'un fichier issu de STATISTIQUES en un fichier DIF.

— transformation d'un fichier DIF en un fichier relisible par STATISTIQUES.

Ce choix vous est demandé par l'intermédiaire du menu ci-dessous :



Dans les deux cas, le nom du fichier à transformer vous est demandé (voir les spécifications du NOM dans copie FICHIER).

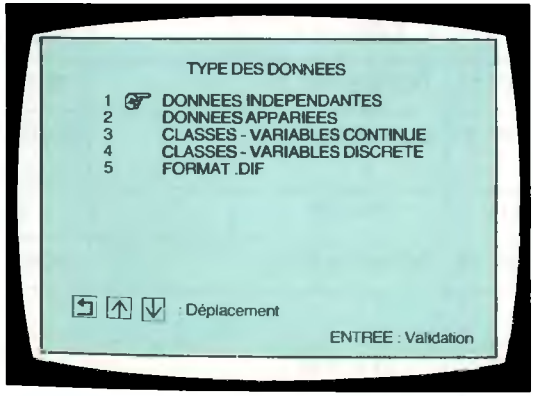
a. **DE STAT EN DIF**

Le tableau obtenu est tel que chaque colonne contient un échantillon (le libellé et les données). La première colonne contient en plus le titre de l'étude.

b. **DE DIF EN STAT**

Cette transformation vous impose de respecter la structure des échantillons qui sont traités par STATISTIQUES.

Pour cela vous devez tout d'abord spécifier le type des données qui se trouvent dans votre fichier DIF :



DONNEES INDEPENDANTES - DONNEES APPARIEES

Le fichier DIF,visualisé sous forme de tableau, doit avoir la structure suivante :

TITRE			
LIBELLE 1	LIBELLE 2	—	LIBELLE 30
VALEUR 11	VALEUR 21	—	VALEUR 301
VALEUR 12	VALEUR 22		VALEUR 302
VALEUR 13	VALEUR 23		VALEUR 303
• •	• •		• •
VALEUR 1100	VALEUR 2100		VALEUR 30100

NOMBRE MAXIMAL DE COLONNES : 30
NOMBRE MAXIMAL DE LIGNES : 101

CLASSES - VARIABLE CONTINUE OU VARIABLE DISCRETE

Le fichier DIF, visualisé sous forme de tableau, doit avoir la structure suivante :

TITRE			
LIBELLE 1	LIBELLE 2	—	LIBELLE 30
BORNE INF. 11	BORNE INF. 21	—	BORNE INF. 301
BORNE SUP. 11	BORNE SUP. 21		BORNE SUP. 301
EFF. 11	EFF. 21		EFF. 301
BORNE INF. 12	BORNE INF. 22		BORNE INF. 302
EFF. 130	EFF. 230		EFF. 3030

NOMBRE MAXIMAL DE COLONNES : 30
NOMBRE MAXIMAL DE LIGNES : 91

$0 \leqslant$ EFF. i j $\leqslant$ 1000 et EFF. i j entier

Borne inf. i j $\leqslant$ Borne Sup. i j $\leqslant$ Borne Inf. i j + 1

Dans le cas de classes-variable discrète :

Borne Inf. i j = Borne Sup. i j

Si le fichier DIF ne correspond pas à la structure du type de données choisi, le message :

« Transfert Irnpossible »

apparaît à l'écran et le menu FORMAT. DIF est de nouveau affiché.

II. — CARACTÉRISTIQUES SYSTÈME

Cet écran vous est déjà apparu lors de la première utilisation de votre logiciel STATISTIQUES.

Vous pouvez donc le rappeler à partir du MENU GENERAL pour vérifier la configuration que vous lui aviez indiquée ou pour la modifier. La modification peut être opérée en répondant N (non) à la question « Acceptation O/N ».

Pour de plus amples détails, reportez-vous à MISE EN SERVICE.

GLOSSAIRE

GLOSSAIRE

Ajustement loi normale : test de raccordement d'une distribution observée à une population, suivant une loi de probabilité gaussienne ou normale. Voir « Test du χ^2 ».

Ajustement référence : test de raccordement d'une distribution observée à une population, suivant une loi de probabilité théorique quelconque, grâce à un échantillon de référence. Voir « Test du χ^2 ».

Analyse de la variance à un facteur contrôlé : recherche de l'influence éventuelle d'un facteur A sur les mesures d'une variable X.

Analyse de la variance à deux facteurs contrôlés : recherche de l'influence éventuelle de deux facteurs A et B sur les mesures d'une variable X.

Bravais-Pearson : coefficient de corrélation linéaire. Coefficient de la droite de régression linéaire. Voir « Corrélation ».

Classes-variable continue : l'un des quatre types de données statistiques traités par STATISTIQUES. Les données de l'échantillon ont été regroupées en classes. La caractéristique mesurée est continue.

Classes-variable discrète : l'un des quatre types de données statistiques traitées par STATISTIQUES. Les données de l'échantillon ont été regroupées en classes. La caractéristique mesurée est discontinue.

Corrélation : les coefficients de corrélation permettent de déceler une éventuelle relation fonctionnelle entre deux variables.

Degré de liberté : nombre de variables aléatoires indépendantes.

Données appariées : l'un des quatre types de données statistiques traitées par STATISTIQUES.
On dit que des échantillons sont composés de données appariées ou sont appariés si ces données représentent les résultats de différentes mesures effectuées sur le même prélèvement.
D'autre part, les données n'ont subi ni regroupement, ni mise en ordre.

Données indépendantes : l'un des quatre types de données statistiques traitées par STATISTIQUES.
On dit que des échantillons sont composés de données indépendantes ou sont indépendants, si ces données sont par exemple les résultats d'une mesure effectuée sur plusieurs prélèvements.

Ecart interquartile : différence entre le quartile supérieur et le quartile inférieur.

Ecart moyen : moyenne des valeurs absolues des écarts des mesures à la moyenne $\bar{x}$.

Ecart type : appelé encore écart quadratique moyen, c'est la racine carrée de la variance. C'est le paramètre le plus employé pour caractériser la dispersion.

Echantillon : c'est la fraction d'éléments, prélevés dans la population que l'on veut étudier, qui va permettre l'étude d'un caractère de la population.

Estimation : ne pouvant connaître les paramètres réels d'une population, on est amené à les estimer à partir de ceux de l'échantillon prélevé.

Kendall : coefficient de corrélation ou « τ » de Kendall. Coefficient non paramétrique calculé sur les rangs qu'occupent les données dans l'échantillon. Voir « Corrélation ».

Médiane : tous les termes de l'échantillon étant rangés par ordre de grandeur croissant, la médiane est une valeur telle qu'il existe autant de termes de l'échantillon qui lui sont inférieurs que de termes qui lui sont supérieurs.

Moyenne : la moyenne arithmétique d'un échantillon est égale à la somme des termes de cet échantillon divisée par le nombre de ces termes.

Paramètres de dispersion : ils caractérisent la répartition des termes de la population autour de la valeur centrale. Voir « Variance » - « Ecart-type » - « Ecart-moyen » - « Ecart-interquartile ».

Paramètres de position : ils rendent compte de la tendance centrale de la population. Voir « Moyenne » - « Médiane » - « Quartile inférieur » - « Quartile supérieur ».

Population : représente la collection d'éléments que l'on envisage d'étudier par l'intermédiaire d'un échantillon prélevé dans celle-ci.

Probabilité : chaque test est accompagné d'une probabilité spécifiant le domaine de validité de l'hypothèse posée. Cette probabilité représente en général le risque d'acceptation de l'hypothèse alors que celle-ci est fausse.

Quartile inférieur : tous les termes de l'échantillon étant rangés par ordre de grandeur croissant, le quartile inférieur est une valeur telle qu'il existe 25 % des termes de l'échantillon qui lui sont inférieurs.

Quartile supérieur : tous les termes de l'échantillon étant rangés en ordre de grandeur croissant, le quartile supérieur est une valeur telle qu'il existe 25 % des termes de l'échantillon qui lui sont supérieurs.

Spearman : coefficient de corrélation ou «ρ» de Spearman. Coefficient non paramétrique calculé sur les rangs qu'occupent les données dans l'échantillon. Voir « Corrélation ».

Statistique descriptive : phase analytique de la démarche statistique. Les données sont réduites à un nombre limité de paramètres carac-

téristiques (paramètres de position, paramètres de dispersion) susceptibles de décrire la série statistique. Voir « Paramètres de position » - « Paramètres de dispersion ».

Test du χ^2 *:* test d'ajustement d'une distribution observée à une distribution théorique. Voir « Ajustement loi normale » - « Ajustement référence ».

Test de Snédécor : test de comparaison des variances de deux échantillons.

Test de Student : test de comparaison des moyennes de deux échantillons.

Tests paramétriques : comparaison de paramètres estimés sur deux ou plusieurs échantillons, afin de savoir si leurs différences sont significatives, ou si on peut les considérer comme issus d'une même population. Voir « Analyse de la variance à un facteur contrôlé » - « Analyse de la variance à deux facteurs contrôlés » - « Test de Snédécor » - « Test de Student ».

Variance : elle représente la fluctuation de la population autour de la valeur centrale. C'est la moyenne des carrés des écarts à la moyenne. La variance calculée est toujours la variance estimée.

EXEMPLES

EXEMPLES

EXEMPLE 1

Recherche de l'influence d'un facteur sur un ensemble de mesures :

Trois lots de poudre à fusil sont fabriqués suivant trois procédés : A, B, C.

On effectue dix tirs au fusil avec chacun de ces lots et on relève les vitesses initiales des balles.

Ces échantillons sont donc formés de données indépendantes.

Ils vont subir les tests suivants :

— statistique descriptive : calcul des paramètres de position et de dispersion afin de les caractériser avec un nombre restreint de paramètres ;

— test d'analyse de la variance à un facteur contrôlé, permettant de tester la différence entre les trois procédés.

Démarche à suivre dans le logiciel :

0) Caractéristiques-système.

1) Acquisition.

2) Statistique descriptive.

3) Tests paramétriques.

0) *Caractéristiques-système :*

Si vous vous servez de ce logiciel pour la première fois, certains renseignements vous sont demandés :

— le nombre de lecteurs à votre disposition : un ou deux,

— l'existence d'une imprimante : OUI ou NON.

1) *Acquisition :*

Ces trois échantillons doivent d'abord être saisis. Pour cela, choisissez dans le MENU GENERAL l'option ACQUISITION.

• Puisque vos données sont nouvelles, il s'agit donc d'une SAISIE. Les données vont être sauvegardées par la suite dans un fichier. Vous devez dès à présent donner son nom : par exemple « FUSIL ».

• Les données des fichiers sont indépendantes, donc dans le menu SAISIE qui vous est maintenant proposé, choisissez la rubrique numéro 1, c'est-à-dire DONNEES INDEPENDANTES.

● Le premier écran de saisie apparaît à l'écran. Celui-ci vous demande des renseignements généraux sur votre étude :

— le titre de l'étude,
— le nombre d'échantillons,

puis :

— le libellé et l'effectif de chacun de ces échantillons.

Dans notre cas, les renseignements suivants peuvent être fournis :

Titre de l'étude : Tirs au fusil
Nombre d'échantillons : 3
Libellé premier échantillon : LOT A
Effectif premier échantillon : 10
Libellé deuxième échantillon : LOT B
Effectif deuxième échantillon : 10
Libellé troisième échantillon : LOT C
Effectif troisième échantillon : 10

Le curseur se place automatiquement au début de la zone à saisir. Lorsque vous avez fini de rentrer votre donnée (les commandes d'édition usuelles sont à votre disposition : [INS] , [EFF] , ⇦ , ⇨ , vous validez la zone par appui sur [ENTREE] et le curseur se positionne sur la zone suivante.

Lorsque tous les libellés et effectifs ont été saisis et s'ils vous paraissent corrects, vous répondez « O » à « Validation O/N », ce qui permet de passer directement à l'écran de saisie du premier échantillon.

Si vous avez, par contre, omis le blanc dans le libellé du deuxième échantillon et que vous avez écrit « LOTB », vous répondez « N » à la question « Validation O/N ».

Vous pouvez alors, grâce aux commandes de déplacement ⇧, ⇩, ⇦, vous positionner sur la zone de saisie en question et grâce à la commande d'insertion INS, insérer un blanc entre « LOT » et « B ».

Ensuite, validez la page en répondant « O » à la question « Validation O/N ».

● Dans le deuxième écran, le libellé et l'effectif du premier échantillon vous sont rappelés.

Dix zones de saisie vierges attendent les données de votre échantillon. Ce sont les suivantes :

801 - 803 - 805 - 797 - 804
798 - 802 - 806 - 801 - 799

Vous allez opérer les mêmes opérations que pour l'écran précédent.

Lorsque votre écran est correctement rempli, vous répondez « O » à « Validation O/N ».

• Le troisième écran vous permet de donner à l'ordinateur les données du deuxième échantillon.

809 - 801 - 804 - 804 - 800
809 - 801 - 806 - 802 - 805

• Le quatrième et dernier écran de saisie attend les dix données du troisième échantillon :

795 - 798 - 803 - 800 - 803
801 - 803 - 805 - 799 - 796

L'acquisition des données est terminée. Le Menu « Acquisition » vous permet maintenant de sortir sur listing vos données, ceci, bien sûr, si vous avez précisé dans la rubrique « Caractéristiques Système » que vous possédez une imprimante.

Il vous suffit de choisir l'option numéro 4, LISTE DES DONNEES-LISTING.

2) *Statistique descriptive :*

Ce module va vous permettre de réduire les données de vos trois échantillons à un nombre restreint de paramètres susceptibles de décrire chacun de ces échantillons.

Vous devez préciser le nom de votre fichier, car il ne faut pas oublier que l'une quelconque des options du MENU GENERAL peut être appelée directement dès le lancement du programme. Tapez donc dans la zone réservée à cet effet : « FUSIL ».

Le programme va mettre quelques secondes pour calculer tous les paramètres pour chacun des échantillons.

Il vous présente ensuite les résultats dans quatre tableaux : deux tableaux pour les paramètres de position et deux pour les paramètres de dispersion.

Pour passer à l'écran suivant, tapez sur la touche ACC .

Lorsque les quatre tableaux ont été visualisés, vous pouvez demander le listing de ces résultats sur imprimante.

Conclusions : étudions plus en détail les résultats obtenus, par exemple, sur le premier échantillon. Tout d'abord, les quatre paramètres de position :

Paramètres de Position - LOT A

Moyenne = 801,6	Quartile inférieur = 798,5
Médiane = 801,51	Quartile supérieur = 802,5

Ceux-ci rendent compte de la tendance de la valeur centrale de la population.

La moyenne, à l'encontre de la médiane, tient compte de tous les termes de l'échantillon. Elle a donc une signification plus synthétique.

Mais nous pouvons voir sur cet échantillon que moyenne et médiane sont pratiquement confondues, ce qui est normal puisque l'échantillon ne contient pas de termes anormaux.

Les quartiles inférieur et supérieur nous indiquent que 25 % des vitesses de tir sont inférieures à 798,5 et 25 % sont supérieures à 802,5.

Paramètres de dispersion - LOT A

Variance = 8,93	Ecart moyen = 2,4
Ecart-type = 2,99	Ecart interquartile = 4

Les paramètres de position permettent de situer l'échantillon autour d'une valeur centrale, mais ne donnent aucune idée de la répartition des termes autour de cette valeur. Ceci est réalisé par les paramètres de dispersion.

En fait, l'écart-type est le paramètre le plus employé pour caractériser la dispersion.

Il présente un intérêt particulier dans les échantillons dits « normaux ».

En général, 95 % des termes d'un échantillon sont compris dans l'intervalle :

$(x - 2\ \sigma\ ; x + 2\ \sigma)$ $\quad$ x = moyenne
σ = écart-type

3) *Tests paramétriques :*

Parmi les tests paramétriques proposés, le plus intéressant dans le cas de nos trois échantillons est l'analyse de la variance.

Les données étant indépendantes, il s'agira d'une analyse de la variance à un facteur contrôlé.

Ce test va nous permettre de tester la différence entre les trois procédés de fabrication.

L'hypothèse posée est alors appelée hypothèse nulle.

H = non-influence du facteur contrôlé, c'est-à-dire non-influence du procédé de fabrication.

Ecart	DDL	VARIANCE	F	P
Entre modalités	2	37,3	3,71	0,73
Intra modalités	27	10,052		
Total	29	11,931		

La dispersion totale entre les différentes valeurs de la caractéristique se décompose en une dispersion due aux différentes modalités du facteur contrôlé (écart entre modalités) et en une dispersion due au seul hasard, puisque dans une même colonne, le facteur contrôlé ne peut pas être responsable des écarts entre les différentes valeurs (écart intramodalités).

Ce tableau nous indique que la probabilité d'accepter H dans le cas où elle est fausse est de 73 %. Nous pouvons donc conclure à l'influence du procédé de fabrication de la poudre sur la vitesse initiale des balles tirées au fusil.

EXEMPLE 2

Ajustement d'une loi statistique à une loi théorique Test du χ^2

On effectue 500 mesures de l'erreur de pointage en dérive, lors du tir à partir d'un avion, sur une cible terrestre. Les résultats des mesures sont exprimés en millièmes de radian.

Il est plus simple d'avoir les résultats sous forme compacte et ordonnée lorsque l'effectif d'un échantillon est supérieur à 100.

Les données sont donc regroupées en classes. La variable aléatoire sur laquelle ont été effectuées les observations est une variable continue.

Classes	(−4; −3)	(−3; −2)	(−2; −1)	(−1; 0)	(0; 1)	(1; 2)	(2; 3)	(3; 4)
Effectifs	6	25	72	133	120	88	46	10

On cherche à vérifier la conformité des répartitions théorique et statistique.

Pour cela, on va utiliser la loi théorique normale de moyenne m et d'écart-type σ .

m et σ sont les paramètres de l'échantillon que l'on va pouvoir connaître par les calculs proposés dans STATISTIQUE DESCRIPTIVE.

Démarche à suivre :

1) Acquisition.
   1.1) Saisie.
      1.1.1) Classes-variable continue.
2) Statistique descriptive.
3) Test du χ^2.
   3.1) Ajustement loi normale.

Statistique descriptive :

Moyenne	:	$1{,}68\ 10^{-1}$
Médiane	:	$1{,}17\ 10^{-1}$
Quartile inférieur	:	$-\ 8{,}35\ 10^{-1}$
Quartile supérieur	:	1,22
Variance	:	2,1
Ecart-type	:	1,45
Ecart-moyen	:	1,19
Ecart interquartile	:	2,05

Les deux paramètres qui nous intéressent sont les paramètres caractéristiques d'une loi normale, c'est-à-dire moyenne et écart-type.

Il ne faut pas oublier que le but de l'étude est la comparaison de la distribution statistique de l'échantillon à la distribution théorique choisie pour représenter la population.

Cette loi théorique est la loi normale de moyenne 0,168 et d'écart-type 1,45.

On peut tout de même tirer les renseignements suivants :

50 % des tirs se situent dans (−4 rd ; 0,117 rd).

25 % dans (−4 rd ; −0,835 rd).

25 % dans (1,22 rd ; 4 rd).

Test du χ^2 *:*

Comme nous venons de l'expliquer dans la recherche des paramètres caractéristiques, il s'agit d'effectuer l'ajustement de la distribution statistique à une distribution normale.

Le logiciel donne les résultats suivants :

$\chi^2 = 3{,}45$ D.D.L. = 5
P = 0,36

L'hypothèse H posée est que le modèle théorique choisi représente bien la population.

La probabilité P calculée indique qu'il y a un risque de 36 % d'accepter H dans le cas où elle est fausse.

Rien ne s'oppose donc à l'acceptation de l'hypothèse et on considère que la population d'où est extrait l'échantillon suit bien une loi normale :

N (0,168 ; 1,45)

EXEMPLE 3

Vérification de l'appartenance de deux échantillons à une même population

On a déjà étudié un exemple (Exemple 2) dans lequel il s'agissait de vérifier que la distribution statistique de l'échantillon était bien conforme à la distribution théorique choisie pour représenter la population.

Il s'agissait de mesurer sur l'erreur de pointage en dérive lors du tir à partir d'un avion sur une cible terrestre.

On dispose d'une série de 371 mesures du même type effectuées à partir d'un autre avion.

On veut vérifier que ces deux séries de résultats appartiennent à la même population.

Il s'agit donc de comparer leurs variances, puis leurs moyennes.

Mais avant de lancer la procédure de test, il faut saisir les données. Il ne faut pas oublier qu'un fichier contenant le premier échantillon a déjà été créé. Il reste donc à rajouter le second. Il s'agit alors d'insérer un nouvel échantillon dans le fichier. Celui-ci sera obligatoirement composé de classes-variable continue.

Classes	(−4 ; −3)	(−3 ; −2)	(−2 ; −1)	(−1 ; 0)	(0 ; 1)	(1 ; 2)	(2 ; 3)
Effectifs	3	18	52	112	96	65	25

Il sera également utile de vérifier que la loi de cet échantillon peut se raccorder à une loi normale, avant de le comparer à l'échantillon « normal ».

Ceci sera effectué par un test du χ^2.

Démarche à suivre :

1) Acquisition.
   1.1) Correction.
      1.1.1) Insertion.
         Numéro de l'échantillon à insérer : 1.
         (Le nouvel échantillon se trouve en première place.)

2) Test du χ^2.
   2.1) Ajustement loi normale.

3) Tests paramétriques.
   3.1) Comparaison variances.
   3.2) Comparaison moyennes.

Test du χ^2 *:*

Ajustement de la distribution statistique de cet échantillon à une distribution théorique normale de moyenne m et d'écart-type σ . Ceux-ci étant les paramètres estimés sur l'échantillon.

Le logiciel donne les résultats suivants :

$\chi^2 = 3{,}49$ D.D.L. = 4
P = 0,51

H = le modele théorique choisi représente bien la population d'où est extrait l'échantillon.

Il y a ici un risque de 51 % d'accepter H dans le cas où elle est fausse.

On considère donc que la population d'où est extrait l'échantillon suit bien une loi normale.

Tests paramétriques :

Les échantillons sont de taille élevée (effectifs supérieurs à 100). Les tests qui vont être effectués ne sont donc pas le test de Snédécor et le test de Student habituels pour comparer variances et moyennes.

Les tests sont des tests « normaux ».

Lorsqu'il s'agit de savoir si deux échantillons appartiennent à la même population, il faut effectuer la comparaison des variances et celles des moyennes. Mais il faut bien noter que les tests de comparaison des moyennes proposées ne s'appuient pas sur la conformité des variances.

En d'autres termes, même si le test de comparaison des variances s'est avéré négatif, on peut effectuer le test de comparaison des moyennes.

Comparaison des variances :

Le logiciel STATISTIQUES donne les résultats suivants :

u= − 2,2878 P = 0,97

L'hypothèse H posée était l'égalité des variances. Or, la probabilité P nous indique que le risque d'accepter H dans le cas où elle est fausse est de 97 %.

L'hypothèse d'égalité des variances doit donc être rejetée.

Comparaison des moyennes :

STATISTIQUES donne les résultats suivants :

u = − 1,2765 P = 0,79

L'hypothese H posée était l'égalité des moyennes.

La probabilité P calculée dans ce cas est légèrement inférieure à la précédente. Mais le risque d'accepter H alors qu'elle est fausse est tout de même de 79 %.

En conclusion, on peut rejeter l'hypothèse d'identité des deux populations d'où sont extraits respectivement les deux échantillons.

On peut toutefois, bien que le risque soit malgré tout élevé, accepter l'hypothèse d'égalité des moyennes des deux échantillons.

EXEMPLE 4

Le coefficient de corrélation de Bravais-Pearson

Soit un système de deux variables aléatoires : HOMME

x, la taille d'un homme en cm,
y, sa pointure.

Vingt expériences ont été effectuées. Chaque ième expérience a fourni un couple de valeurs (x_i, y_i).

On se propose d'effectuer :

— Le calcul des paramètres caractéristiques des deux échantillons.

— Le calcul du coefficient de corrélation de Bravais-Pearson qui va nous permettre de savoir s'il existe une relation fonctionnelle entre la taille d'un homme et sa pointure.

i	x_i	y_i	i	x_i	y_i	i	x_i	y_i	i	x_i	y_i
1	172	41	6	170	40	11	170	42	16	187	44
2	174	43	7	174	42	12	185	45	17	166	40
3	175	41	8	175	43	13	180	42	18	167	41
4	181	44	9	164	40	40	172	39	19	171	42
5	168	38	10	168	41	15	174	43	20	179	43

Démarche à suivre :

1) Acquisition.
   1.1) Saisie.
      1.1.1) Données appariées.
2) Statistique descriptive.
3) Corrélation.
   3.1) r de Bravais-Pearson.

Statistique descriptive :

Paramètres de position.

	Moyenne	Médiane	Quartile Inférieur	Quartile supérieur
Taille	173,6	173	169	177
Pointure	41,7	42	40,5	43

Paramètres de dispersion.

	Variance	*Ecart-type*	*Ecart-moyen*	*Ecart interquartile*
Taille	38,57	6,21	4,8	8
Pointure	3,17	1,78	1,43	2,5

• Examinons tout d'abord les paramètres caractéristiques de la « Taille ».

La taille moyenne de la population dans laquelle a été prélevé l'échantillon est de 173,6 cm avec un écart-type de 6,2 cm.

Ceci implique que la majeure partie de la population a une taille comprise entre :

(167,4 ; 179,8)

D'autre part, 50 % de la population mesure moins de 173 cm, 25 % moins de 169 cm et 25 % plus de 177 cm.

• Les paramètres caractéristiques de la « Pointure ».

La pointure moyenne de la population d'où a été extrait l'échantillon de 20 personnes est 41,7 avec un écart-type de 1,8.

Donc, la pointure de la population se trouve en grande partie comprise dans l'intervalle :

(39,9 ; 43,5)

50 % de la population chausse moins de 42 ; 25 % plus de 43 ; 25 % moins de 40,5.

Corrélation :

L'hypothèse H qui est testée est toujours l'indépendance des deux variables.

Les résultats obtenus sont les suivants :

$r = 0,784$

$P = 0,999$

La probabilité calculée indique donc qu'il y a 99 % de chances d'accepter H dans le cas où elle est fausse.

On en déduit donc aisément une relation fonctionnelle très forte entre la taille et la pointure d'un homme.

EXEMPLE 5

Utilisation des coefficients de corrélation non paramétriques

On a demandé à une femme de classer dix tableaux dans l'ordre de ses préférences. Indépendamment, le mari a ensuite classé les mêmes tableaux.

On se propose de décrire statistiquement le degré de concordance entre ces deux classements.

Les données de ces échantillons représentent en fait des rangs et il paraît évident que dans ce cas, la loi suivie n'est pas une loi normale.

L'utilisation des coefficients de corrélation non paramétriques apparaît comme la méthode idéale pour résoudre le problème.

Tableaux	a	b	c	d	e	f	g	h	i	j
Classement de la femme	1	4	2	6	3	5	7	9	8	10
Classement du mari	2	3	1	5	4	7	6	10	8	9

Démarche à suivre :

1) Acquisition.
   1.1) Saisie.
      1.1.1) Données appariées.
2) Corrélation.
   2.1) τ de Kendall.
   2.2) ρ de Spearman.

Corrélation :

Dans les deux calculs de coefficients de corrélation, l'hypothèse H posée est l'indépendance des deux variables.

τ de Kendall

Les résultats proposés par le logiciel sont les suivants :

$\tau = 7.77778\ 10^{-1}$
P = 0.999

La probabilité d'accepter H dans le cas où elle est fausse est presque de 100 %, ce qui indique que l'hypothèse est mauvaise et donc qu'il y a une dépendance très nette entre le classement du mari et celui de la femme.

ρ de Spearman

Une deuxième façon de vérifier la dépendance entre les deux classements est d'effectuer le calcul du coefficient de corrélation de Spearman.

$\rho = 9.27273\ 10^{-1}$
$P = 0.994$

La probabilité d'accepter H dans le cas où elle est fausse est de 99,4 %. On vérifie donc bien encore la forte dépendance entre les deux classements.

EXEMPLE 6

Recherche de l'influence de deux facteurs sur un ensemble de mesures

On cherche à déterminer si le sexe d'une part et la nature des études secondaires d'autre part ont une influence sur la note de mathématiques obtenue à un concours.

Se présentaient à ce concours :

300 filles - 300 garçons,
3 sections : C, D, E,

Dans chaque section, 100 filles et 100 garçons, ceci afin de ne favoriser aucun des deux sexes, ni aucune des trois sections.

Nous allons travailler directement sur les moyennes obtenues et non sur les notes elles-mêmes.

Section / Sexe	C	D	E
Masculin	15,2	12,9	13,1
Féminin	14,3	12,9	11,1

Le test à utiliser pour trouver la solution de ce problème est une analyse de la variance à deux facteurs contrôlés, qui sont ici :
— nature des études secondaires,
— sexe.

Démarche à suivre :

1) Acquisition.
   1.1) Saisie.
      1.1.1) Données appariées.

2) Tests paramétriques.
   2.1) Analyse de la variance.

Analyse de la variance à deux facteurs contrôlés :

Les deux hypothèses posées sont :

H1 : le facteur « section » ou « nature des études secondaires » n'a pas d'influence sur les notes obtenues au concours.

H2 : le facteur « sexe » n'a pas d'influence sur les notes obtenues au concours.

Les résultats sont regroupés dans le tableau ci-dessous :

ECART	DDL	VARIANCE	F	P
ENTRE LIGNES	1	3,08167	3,942	0,48
ENTRE COLONNES	2	1,92167	2,458	0,33
RESIDUEL	2	0,781667		
TOTAL	5	1,69767		

Au vu de ces résultats, nous pouvons conclure qu'il n'y a aucune influence du facteur « sexe » ni du facteur « section » sur les notes.

En effet, la probabilité d'accepter H1 dans le cas où elle est fausse est de 33 % et celle d'accepter H2 dans le cas où elle est fausse, de 48 %.

APPLICATION DES TESTS AUX TYPES DE DONNEES

LES QUATRE TYPES DE DONNEES ET LES TESTS QUI S'Y APPLIQUENT

TESTS	TYPES DE DONNEES
• Statistique descriptive moyenne variance écart-type écart-moyen	Les quatre types de données
quartile inférieur quartile supérieur écart interquartile médiane	Ne peuvent être calculés pour les données de types classes-variable discrète
• Test du χ^2 Ajustement loi normale Ajustement référence	Classes-variable continue. Les deux types de classes : Variable continue Variable discrète
• Tests paramétriques Comparaison moyennes Comparaison variances	Les quatre types de données
Analyse de la variance à un facteur contrôlé	Données indépendantes
Analyse de la variance à deux facteurs contrôlés	Données appariées
• Corrélation Bravais-Pearson Kendall Spearman	Données indépendantes et données appariées

Pour test du χ^2 - Echantillon de référence

L'échantillon courant et l'échantillon de référence doivent avoir :

— les mêmes bornes inférieures,

— les mêmes bornes supérieures,

— le même nombre de classes.

Pour tests de corrélation

Les échantillons à comparer doivent avoir le même effectif.

INDEX

INDEX

Imprimé par Timaroto
Imprimé en France