Gilles HUNAULT

2005

quelques Repères en STATISTIQUES pour Doctorants

Université d'Angers

http ://www.info.univ-angers.fr/pub/gh/

Le principe consistant à ne pouvoir affirmer que des différences et jamais des identités s'applique aussi à la vie courante. Une ressemblance peut toujours être fortuite ; seule une différence permet une conclusion (négative) sûre.

Ainsi j'ai perdu un bouton de ma blouse en faisant mon cours et, me penchant, j'en aperçois un par terre. S'il n'est pas de la même couleur que les miens, il est certain que ce n'est pas le bouton qui me manque ("test couleur" négatif). S'il est de la bonne couleur (test positif), le bouton est peut-être le mien, mais ce n'est pas certain. En effet, s'il n'est pas de la bonne taille ("test taille" négatif), ce n'est pas le mien. S'il est de la bonne taille, c'est peut-être le mien, mais ce n'est pas encore certain et ainsi de suite...

S. Frontier, D. Davoult, V. Gentilhomme, Y. Lagadeuc
Statistique pour les sciences de la vie et de l'environnement
Ch. 5 : Tests d'hypothèses sur les moyennes, p. 141

Table des matières

Chapitre 1.

Introduction

1.1 Des chiffres et des lettres

La réalisation et la rédaction d'une analyse statistique pour un article de recherche ou dans le simple but de rendre compte du travail effectué mettent en jeux deux domaines de compétence :

- le calcul statistique avec ses termes techniques et ses formules, leurs conditions d'applications, les conclusions mathématiques licites auxquelles elles aboutissent,
- l'écriture du rapport d'analyse via la présentation du protocole et des données (voire des hypothèses sous-jacentes), la mise en forme de tout ou partie des résultats et la rédaction des conclusions, interprétations et commentaires pour les spécialistes du domaine, pour l'équipe de recherche ou pour le grand public.

Une analyse statistique ne se réduit donc pas à une suite de calculs, même justes et justifiés. La rédaction est un art difficile. Elle est souvent bâclée par les [pseudo]scientifiques qui confondent phrases, littérature, verbiage et production littéraire. Or, la qualité d'un article de recherche, d'un rapport d'expérimentation transparait au fil des paragraphes. Le choix des termes employés renforce chez le lecteur ou le correcteur la conviction que le travail fait a été bien fait, que les méthodes statistiques sont maitrisées, que le passage des chiffres (comme $m = 12.3$ jours) aux lettres (”*un durée moyenne aussi faible qu'à l'habitude*”) est le fruit d'un mûre réflexion...

1.2 Contenu du manuel

On trouvera dans ces quelques pages un guide pour conduire des analyses statistiques usuelles avec au passage quelques conseils pour la rédaction et la mise en forme des résultats. La partie purement mathématique (formules, démonstrations...) a été réduite au minimum afin de focaliser l'attention sur les concepts et méthodes. Nous avons donc fait le pari d'écrire un texte lisible sans équation ni intégrale. Les principales formules statistiques, notamment pour les intervalles de confiance sont toutefois fournies en fin de manuel. On pourra toujours retrouver les formules et les démonstrations manquantes dans les ouvrages cités dans la bibliographie ou dans nos cours.

Une remarque similaire s'applique aux loix probabilistes utilisées, comme la loi normale de Laplace-Gauss, la loi du χ^2 ou la loi de Student. Nous supposons que le lecteur et la lectrice savent que ces lois correspondent à des modèles théoriques, à des cas "parfaits" ou "idéaux" "et que leurs fonctions de répartition directe ou inverse servent de référence au même titre que les équations de droite, parabole, sinus ou exponentielle servent de référence dans l'étude des fonctions réelles (sans que l'on sache forcément écrire les équations de ces fonctions de répartition).

Le chapitre 2 est centré sur la notion de variable statistique et de traitement *descriptif* associé pour les deux grandes classes de variables que sont les QT (quantitatives) et les QL (qualitatives). Le chapitre 3 passe en revue les divers intervalles de confiance fournis par la théorie de l'estimation ainsi que les principaux *tests* de la statistique inférentielle, leurs utilisations et la lecture de leurs résultats, qu'ils soient paramétriques ou non. Enfin, le chapitre 4 reprend les divers *graphiques* qu'il est bon de savoir aujourd'hui maitriser (sans oublier de rappeler les "erreurs graphiques" à ne pas commettre) avant de rappeler les incontournables et autres garde-fous liés aux *protocoles* d'expérimentation et à la *rédaction*.

Nous avons également ajouté en fin de manuel une bibliographie courte plutôt orientée cours traditionnel et ouvrages avec exercices corrigés que nous avons complétée par une liste – volontairement courte aussi – de références *Web* plus générales. En particulier celle de SMEL, qui signifie *Statistique Médicale En Ligne* et qui correspond à l'URL

http ://www.math-info.univ-paris5.fr/∼smel/

devrait être profitable à tous ceux et celles qui ont à inclure des rapports d'études statistiques dans des articles médicaux.

Chapitre 2.

Statistiques descriptives

2.1 Données, variables et calculs

Le poids d'un individu exprimé en kilogrammes, la présence ou l'absence d'une tumeur ne sont pas des données de même nature. C'est pourquoi on désigne par le vocable "variable quantitative" ou QT toute série de chiffres se rapportant à une quantité mesurable pour laquelle la notion de moyenne a un sens. De même, on désigne par le vocable variable qualitative ou QL toute série de chiffres* se rapportant à des qualités c'est à dire à des états distincts et en nombre fini. Il est important de reconnaitre les types de variables car le vocabulaire et les traitements statistiques sont différents pour ces deux types de variables.

Il est plus imagé et sans doute plus facile à mémoriser d'utiliser les appellations *variables à unités sommables* pour les QT et *variables à codes arbitraires* pour les QL. Il faut noter au passage qu'on nomme aussi "catégorie" ou "modalité" chacune des valeurs possibles pour les états des QL.

Il existe bien sûr d'autres types de variables. Ainsi les variables hiérarchiques ou "rangs" dont les valeurs indiquent un ordre de préférence ou de classement, les variables textuelles qui fournissent des phrases plutôt que des nombres, les variables multi-réponses qui généralisent les QL...

En Biologie comme en Médecine, de nombreuses variables ne sont ni des QT ni des QL et il serait imprudent de les traiter comme telles. Rentrent dans cette catégorie tous les indices, index et autres proportions ou pourcentages **non sommables** comme les densités et l'IMC ("indice de masse corporelle").

* éventuellement obtenus par recodage, comme les classiques 0 et 1 pour "oui" et "non".

L'usage veut qu'en statistiques descriptives on analyse séparément les variables avant de les traiter conjointement c'est à dire par paires. Ce qui signifie qu'on commence par traiter les diverses colonnes de chiffres dans les fichiers comme s'il s'agissait de colonnes séparées, le regroupement de ces colonnes ne se faisant qu'au travers d'un tableau récapitulatif de leurs caractéristiques avant de les passer en revue comme si les fichiers étaient constitués de deux colonnes seulement. Il ne s'agit pas d'une vue de l'esprit mais d'une démarche progressive : on effectue les calculs en dimension 1 (analyse "univariée") avant de passer à la dimension 2 (analyse "bivariée"). Nous leurs préférons les termes moins conventionnels mais plus explicites d'analyse séparée et d'analyse conjointe. Les calculs en dimensions supérieures font appel à des calculs plus généraux, souvent vectoriels, regroupés sous les termes d'analyse multidimensionnelle ou d'analyse des données ("à la française") ou encore analyse multivariée, ce qui inclut les méthodes factorielles, les techniques de classification, régression, discrimination, segmentation...

Il est tout à fait naturel de vouloir comparer les résultats de ces analyses à une, deux dimensions ou plus, de vouloir quantifier le degré de confiance qu'on peut accorder à ces résultats d'où la notion d' *intervalle de confiance* et de *test statistique* qui donne un cadre mathématique rigoureux à ces notions de confiance et de comparaison. Nous présenterons les tests dans le chapitre suivant.

Pour ne pas encombrer la lecture de formules mathématiques, nous avons regroupé celles-ci en fin de manuel. Rappelons que l'utilisation de logiciels ne dispense pas de connaitre et de savoir interpréter ces formules. Elle autorise seulement à ne pas les connaitre par coeur et remplace le calcul manuel – le calcul, avons-nous écrit, pas la réflexion pas plus que l'interprétation et la rédaction.

2.2 Analyse séparée (univariée)

Comme les valeurs numériques ou "codes" d'une QL sont arbitraires, les calculs associés se résument à des comptages. Pour faire "compliqué" là où on peut faire simple, les statisticiens ont inventé un vocabulaire précis et pas toujours intuitif. Ainsi on nomme *effectif absolu* de la modalité i pour la variable j le nombre de fois où on trouve le code numéro i de cette variable. Ce simple dénombrement est aussi appelé comptage ou fréquence de la modalité. La somme des effectifs absolus pour la variable j est nommée *effectif total* de la variable et le rapport effectif absolu/effectif total pour chaque modalité est nommé *effectif relatif* ou proportion ou pourcentage de la modalité. Le regroupement de ces calculs est nommé *tri à plat* de la variable en français (et table de fréquences en anglais). Le *mode* (masculin) d'une QL est alors la modalité de plus grand effectif. En principe à chaque modalité de chaque QL est associé un "label" ou "intitulé" plus ou moins court qui doit figurer à la place de chaque code dans les tableaux de résultats.

Lorsqu'on s'intéresse à la distribution ou répartition théorique des valeurs on calcule aussi les *effectifs cumulés* qui induisent la fonction de répartition empririque de la variable.

Ces calculs sont en général doublés de graphiques comme les histogrammes et polygones de fréquences que nous traiterons au chapitre 4.

Une erreur classique (parfois volontaire) est d'oublier d'indiquer la taille totale n des données traitées : 20 % de 5 personnes ne signifient pas la même chose que 20 % pour 500 personnes. Et que dire de "*10 % des patients...*" quand on ne sait pas combien il y a de patients en tout ?

La taille des données joue un role important pour les analyses de type enquête. Par exemple en France, en-dessous de 1000 personnes une enquête ne peut se dire nationale au sens "représentative de l'ensemble de la population française". Dans le cas d'études sur des pathologies, il arrive qu'on ne dispose que d'un nombre de données faible ou très faible soit parce qu'elles sont [très] peu disponibles soit parce que les obtenir coute cher. Il faut en tenir compte au niveau des conclusions et de la rédaction qui devront être plus nuancées qu'avec un grand nombre de valeurs.

De nombreux logiciels se contentent d'un affichage des effectifs absolus par ordre croissant de numéro de modalité. C'est bien sûr insuffissant : il vaut mieux fournir un affichage par effectif relatif décroissant de façon à favoriser la comparaison des variables, quitte à "élaguer" les modalités de faible effectif (qu'on fournira éventuellement sous forme de document annexe).

Il y a un *ordre* statistique de présentation pour un tableau récapitulatif des tris à plat de variables QL : c'est celui où les variables sont présentées à raison d'une variable par ligne, chaque ligne contenant les effectifs relatifs avec leur label rangés par ordre décroissant, les lignes étant triées par mode décroissant. Cet affichage fait ressortir les variables les plus "marquées", voire à modalités majoritaires de celles plus faiblement ou plus uniformément réparties.

Lors de calculs sous *Excel*, il est courant de voir des tableaux de résultats sans aucune référence aux noms des variables et des modalités et donc aussi incompréhensibles que

```
V1  0  3 %  1  6 %  2  30 % 3 21 %  4  39 %
V2  0 35 %  1 65 %
```

C'est impardonnable. Un tableau récapitulatif des QL se doit d'être presque totalement auto-descriptif comme par exemple le tableau suivant qui présente les mêmes résultats que le tableau précédent :

```
Tableau 1 : récapitulatif des tris à plat pour les 99 individus

  Variable   Mode Pourcent | 2ème mod.   Pourcent  | 3ème mod. Pourcent
   SEXE      Homme   65 %  | Femme           35 %  |
   ETUDES    Sup.    39 %  | Bepc            30 %  |  Bac.         21 %  ...
```

L'analyse des valeurs numériques d'une QT demande des calculs beaucoup plus techniques que pour une QL et aboutit à de nombreuses valeurs nommées *paramètres*. Commençons par supposer que nos valeurs correspondent à l'*ensemble de la population* observée. Le premier paramètre à fournir concerne la taille des données et s'exprime simplement comme le nombre n de valeurs mises en jeu. Comme on le verra au chapitre suivant, ce nombre est important car certains calculs sont différents suivant qu'on dispose d'un "petit" ensemble de valeurs ou d'un "grand". Les paramètres de position (ou encore "de tendance", de "tendance centrale") sont principalement la moyenne [et la médiane] alors que les paramètres principaux de dispersion sont l'écart-type (dispersion absolue) et le coefficient de variation (dispersion relative) [et la distance interquartile].

La moyenne m de n valeurs x_i est un résumé très imparfait : elle remplace ces valeurs par un seul nombre via la somme et le nombre de ces valeurs puisque la "bonne" définition de la moyenne m est : $n \times m = \Sigma\ x_i$.

La médiane permet de séparer l'ensemble trié par ordre croissant des valeurs en deux sous-ensembles avec le même nombre (50 %) de valeurs. La variance V est le carré de l'écart-type σ et quantifie la dispersion ou "variation moyenne autour de la moyenne" définie comme la moyenne des carrés des différences entre les valeurs x_i et leur moyenne m.

Il est d'usage de fournir au minimum le nombre de valeurs, la moyenne et l'écart-type lors de l'analyse d'une QT sans oublier de rappeler l'unité de mesure [sommable]. Nous conseillons très fortement d'adjoindre le coefficient de variation σ/m exprimé en % (voire la médiane lorsque c'est possible). L'affichage de la variance peut parfois prêter à confusion : elle ne s'exprime pas avec la même unité que la variable mais avec celle de son carré.

Il faut se rappeler qu'une moyenne seule ne décrit pas suffisamment les données. Ainsi les quatre séries A, B, C et D ci-dessous ont la même moyenne à savoir la valeur 10 alors qu'elles sont trés différentes en termes de dispersion :

```
A   10 10 10 10 10 10 10 10 10 10
B    9 11  9 11  9 11  9 11  9 11
C    9  9  9  9  9 11 11 11 11 11
D    2 18  2 18  2 18  2 18  2 18
```

L'écart-type permet de quantifier globalement la variation absolue autour de la moyenne. Ainsi $\sigma(A) = 0$, $\sigma(B) = \sigma(C) = 1$, $\sigma(D) = 8$. La série A est donc constante, les données de la série D varient plus que les données de la série B qui, elle, varie autant que la série C.

L'écart-type ne permet pas de comparer la dispersion de deux séries dont les moyennes sont différentes. De plus l'écart-type est lié à l'unité de msure et à l'ordre de grandeur des valeurs. On pourrait imaginer qu'à un grand écart-type correspond une grande dispersion. C'est bien sur faux dans l'absolu. C'est pourquoi le coefficient de variation σ/m est un bon indicateur de dispersion relative : si E correspond à la série D multipliée par 10 alors la moyenne et l'écart-type de E sont dix fois plus grands que la moyenne et l'écart-type de D. Par contre, les coefficients de variation sont les mêmes.

N'oublions pas non plus que ni la moyenne ni l'écart-type ne rendent compte de l'évolution des valeurs lorsque celles-ci sont ordonnées par exemple chronologiquement : m et σ sont invariantes par permutation et ne rendent donc pas compte de la "progression" (ou "évolution") des valeurs de la série C.

De même qu'il y avait un ordre "intelligent" pour l'affichage des résultats des QL il y a un "bon" ordre pour les QT : c'est celui qui présente les variables par ordre décroissant de coefficient de variation.

Ainsi les résultats pour nos 4 séries devraient être présentés comme suit

Série	*Moyenne*	*Ecart-type*	*Coefficient de variation*
D	10 g	8.00	80 %
B	10 g	1.00	10 %
C	10 g	1.00	10 %
A	10 g	0.00	NaN

Résultats triés par *cdv* pour les $n = 10$ valeurs de poids de l'exemple

Cet ordre est bien sur le seul légal lorsque les variables sont exprimées dans des unités différentes : on ne met pas la taille d'indidivus adultes exprimée en centimètres avant leur poids en kilogrammes sous prétexte que la moyenne des tailles est plus grande que la moyenne des poids ! L'ordre chronologique d'entrée des variables lorsqu'il n'est pas bien pensé n'est pas spécialement intéressant : on ne commence pas par commenter la variable "hauteur des pissenlits" seulement parce que c'est la première variable du fichier.

Il faut parfois compléter cet affichage par des extraits triés des résultats notamment lorsque plusieurs variables comparables mettent en jeu les mêmes unités. Ou respecter l'ordre imposé par un protocole ou par un ordre alphabétique lorsqu'on manipule de trés nombreuses variables...

Si les valeurs analysées ne représentent pas toute la population mais seulement un *échantillon* de la population, alors on reprend les mêmes indicateurs de taille et de moyenne mais le calcul de la variance est légèrement modifié : on utilise la variance estimée qui correspond à la variance précédente (dite exacte ou "empirique") multipliée par $n/(n-1)$. Cela induit de grandes différences pour n petit mais cela ne change pas grand chose pour les grandes valeurs de n comme le montre le tableau suivant où r vaut $n/(n-1)$

n	5	10	20	50	100	500
r	1.250000	1.111111	1.052632	1.020408	1.010101	1.002004
$\sqrt{r}$	1.118034	1.054093	1.025978	1.010153	1.005038	1.001002

Il y a malheureusement toute une "floppée" d'autres paramètres statistiques disponibles et parfois indispensables pour étudier une série de valeurs QT, notamment lorsqu'on suppose que la distribution est normale. Par exemple nous voulons traiter les valeurs suivantes `1.440 1.115 1.620 1.272 1.121 1.039` exprimées en g correspondant au poids absolu du coeur pour des rats témoins agés de 10 à 12 semaines ([$cf.$ G. JOHNSON]).

Le minimum statistique à fournir pour analyser ces données dans l'absolu est sans doute (abstraction faite pour l'instant du choix du nombre de décimales)

n	m	med	σ	σ/m
6	1.2678333 g	1.1965000 g	0.2240370 g	17.6708539 %

alors que pour un article on fournira seulement (et en anglais), à l'intérieur d'un tableau plus complet les valeurs

...	$Name$	$mean \pm\ SEM$	...
...	Heart (g)	1.3 ± 0.1	...

La plupart des logiciels statistiques fournissent systèmatiquement de nombreux paramètres comme *Statbox* qui affiche pour les mêmes données :

Indicateur	*Valeur*
Nbr de valeurs utilisées	6
Nbr de valeurs ignorées	0
Nbr de val. min.	1
% de val. min.	16,67
Minimum	1,04
1er quartile	1,12
Médiane	1,20
3ème quartile	1,44
Maximum	1,62
Etendue	0,58
Total	7,61
Moyenne	1,27
Moyenne géométrique	1,25
Moyenne harmonique	1,24
Aplatissement (Pearson)	-1,68
Asymétrie (Pearson)	0,45
Aplatissement	-0,69
Asymétrie	0,80
CV (écart-type/moyenne)	0,18
Variance d'échantillon	0,04
Variance estimée	0,05
Ecart-type d'échantillon	0,20
Ecart-type estimé	0,22
Ecart absolu moyen	0,18
Ecart-type de la moyenne	0,09

Le grand logiciel *SAS* n'est pas en reste non plus pour si peu de données :

```
The SAS System

    The UNIVARIATE Procedure

    Moments

    N                   6          Sum Weights         6
    Mean                1.26783333 Sum Observations    7.607
    Std Deviation       0.22403698 Variance            0.05019257
    Skewness            0.80198676 Kurtosis           -0.689441
    Uncorrected SS      9.895371   Corrected SS        0.25096283
    Coeff Variation    17.6708539  Std Error Mean      0.09146271


    Basic Statistical Measures

    Location                     Variability

    Mean     1.267833     Std Deviation            0.22404
    Median   1.196500     Variance                 0.05019
    Mode      .           Range                    0.58100
                          Interquartile Range      0.32500


    Tests for Location: Mu0=0

    Test           -Statistic-    -----p Value------

    Student's t    t  13.86175    Pr > |t|    <.0001
    Sign           M   3          Pr >= |M|   0.0313
    Signed Rank    S  10.5        Pr >= |S|   0.0313
```

A la vue de ces divers calculs, il est facile de comprendre que souvent *l'usage fait loi* pour les articles de recherche : on met en général les mêmes paramètres que ceux fournis dans les autres articles du domaine, même si parfois les habitudes sont à la limite du discutable statistique... Nous conseillons cependant de mettre en annexe ce genre de résultats détaillés pour les experts...

2.3 Analyse conjointe (bivariée)

Traiter deux QL ensemble se dit "effectuer un tri croisé des variables". Le résultat est un tableau croisant toutes les modalités d'une variable avec toutes les modalités de l'autre variable. La valeur obtenue pour chaque case peut être absolue ou relative. Si elle est relative, ce peut être de trois façons : exprimée comme pourcentage du total général, du total par ligne, du total par colonne. Le choix de l'un de ces trois types de division est en général arbitraire, à moins que l'une des variables soit considérée *a priori* comme déterminante. C'est pourquoi certains logiciels comme *SAS* les fournissent systématiquement :

```
Table of UI (rows) by HT (columns) for Low Birth Rate Study

http://www-unix.oit.umass.edu/~statdata/statdata/data/lowbwt.txt

 UI (Presence of Uterine Irritability)
 HT (History of Hypertension)


                    |  No     |  Yes    |  Total
               ---------+---------+---------+
 Frequency      No       |    149  |     12  |    161
 Percent                 |  78.84  |   6.35  |  85.19
 Row Pct                 |  92.55  |   7.45  |
 Col Pct                 |  84.18  | 100.00  |
               ---------+---------+---------+
               Yes      |     28  |      0  |     28
                        |  14.81  |   0.00  |  14.81
                        | 100.00  |   0.00  |
                        |  15.82  |   0.00  |
               ---------+---------+---------+
               Total    |    177  |    12   |   189
                        |   93.65 |    6.35 |   100.00
```

Lorsqu'on dispose de beaucoup de QL on renonce en général à faire tous les tris croisés : n variables QL représentent $n(n-1)/2$ tableaux de croisements ce qui devient vite fastidieux. Ainsi pour 10 variables QL et ce, quelque soit le nombre de lignes traitées, il y a en tout 45 tableaux de tris croisés. On conçoit que pour une page de données (disons 50 ou 60 lignes), présenter 45 tableaux de tris croisés n'est pas un résumé des plus concis !

La situation est toute autre pour les QT car à chaque couple de QT on associe non pas un tableau mais un seul nombre, noté ρ et nommé coefficient de corrélation linéaire. Ce nombre qui par construction varie entre -1 et 1 réflète la tendance des données à se présenter sous forme d'une droite : plus $|\rho|$ est proche de 1, plus les données sont linéaires. Le signe de ρ indique alors la croissance réciproque (ρ positif) ou la décroissance (ρ négatif).

Il est très facile de représenter l'ensemble des coefficients de corrélation linéaire sous forme d'une matrice triangulaire inférieure pour avoir une idée de l'ensemble des liaisons linéaires possibles mais il faut passer en revue les valeurs décroissantes de $|\rho|$ pour savoir quelles relations linéaires il faut retenir. On ne s'étonnera donc pas de voir par exemple des sorties numériques où seules les relation linéaires les plus fortes sont accompagnées des équations correspondantes comme par exemple pour cette étude d'importation de 18 catégories de vins ciblée sur 8 pays :

```
Matrice des coefficients de corrélation linéaire pour les 8 pays
(ou "Matrice des Corrélations") pour les importations

        BELG        NEDE       RFA     ITAL  UK      SUIS  USA     CANA
BELG    1,00
NEDE    0,87    1,00
RFA     0,87    0,58    1,00
ITAL    0,59    0,29    0,70    1,00
UK      0,94    0,70    0,97    0,69    1,00
SUIS    0,34    0,52    0,20    0,31    0,25    1,00
USA     0,87    0,68    0,85    0,72    0,89    0,47    1,00
CANA    0,81    0,46    0,95    0,66    0,93    -0,02   0,75    1,00

Liste triée des corrélations  |  Equations des meilleurs corrélations
                              |
RFA    UK     0,9693          |  0,9693  : RFA  = 1,831 * UK  - 5921,349
RFA    CANA   0,9476          |  0.9693    UK   = 0,513 * RFA + 3903,587
BELG   UK     0,9416          |
UK     CANA   0,9256          |  0,9476  : RFA  = 4,857 * CANA + 6596,411
UK     USA    0,8935          |            CANA = 0,185 * RFA  -  932,386
BELG   NEDE   0,8702          |
BELG   USA    0,8699          |  0,9416  : BELG = 0,398 * UK   + 1772,722
BELG   RFA    0,8692          |            UK   = 2,225 * BELG - 2322,591
RFA    USA    0,8477          |
BELG   CANA   0,8143          |  0,9256  : UK   = 2,511 * CANA + 7233,245
USA    CANA   0,7469          |            CANA = 0,341 * UK   - 2064,845
ITAL   USA    0,7172          |
RFA    ITAL   0,6998          |
...                           |
RFA    SUIS   0,1984          |
SUIS   CANA  -0,0246          |
```

Le coefficient ρ ou plus exactement $\rho(X, Y)$ si l'on veut faire explictement référence aux noms X et Y des deux variables QT mises en jeu résulte d'un calcul mathématique assez simple basé sur des sommes, des produits et des divisions : on divise la covariance par le produit des écart-types (la covariance étant l'analogue de la variance mais pour deux variables, soit la différence entre la moyenne du produit des variables et le produit des moyennes). La relation mathématique binaire entre X et Y induite par $|\rho(X, Y)| = 1$ est transitive et symétrique ce qui peut avoir de lourdes conséquences :

- deux variables corrélées linéairement à une même troisième variable sont corrélées linéairement, ce qui est très délicat à détecter et à commenter lorsque la troisième variable n'est pas présente dans le dossier d'étude,
- si deux variables sont corrélées linéairement, rien n'indique s'il y a un sens de liaison de l'une à l'autre.

Si l'on ajoute l'adage célèbre "*corrélation* [*linéaire*] *n'est pas causalité*" et si l'on sait que "*liaison ne signifie pas systématiquement corrélation* [*linéaire*]" on aura compris que la corrélation linéaire est simple pour les formules mais pas forcément pour la pratique statistique et qu'elle ne couvre pas, loin s'en faut, tous les cas de "liaison".

Ainsi trouver une liaison linéaire entre le prix d'un carnet de ticket de métro et celui d'un paquet de couches-culottes ou celui d'une voiture est "banal" car la plupart des prix sont liés à une variable cachée nommée "inflation".

De même on ira chercher une causalité entre "le nombre hebdomadaire de chomeurs saisonniers en Anjou au mois de juillet" et "le nombre de *mm* de pluie tombée" dans le "bon sens" si on connait un peu les cultures ligériennes de mais et de tabac mais on ne dira surtout pas "*plus il y a de chomeurs, plus il pleut*" !

2.4 Analyse multidimensionnelle (multivariée)

Il est bien sur possible de faire des calculs à plus de deux variables, comme les tris dits multicritères pour les QL et les régressions partielles pour les QT mais rien n'empêche de panacher, par exemple de refaire les analyses univariées et bivariées des QT pour chaque modalité de toutes les QL mais alors pour une page de chiffres en entrée, on obtiendra une cinquantaine ou plus de pages de résumé !

C'est pourquoi des méthodes vectorielles multidimensionnellles ont vu le jour. On les nommes classifications (CAH, centres mobiles, nuées dynamiques...), analyses factorielles (AFC, ACP, AFCM...). Leurs principes sont simples et la mise en oeuvre logicielle souvent facile mais la compréhension des méthodes, la vérification des conditons d'application et l'interprétation des résultats tant numériques que graphiques demande un certain apprentissage. Notamment pour les méthodes qui mettent en jeu des inerties car cette notion ne se lit pas directement sur un graphique : seule la composante distance, soit d dans la formule $I = md^2$ est représentée sur un graphique.

Au passage, on notera la similarité des formules entre inertie $I = md^2$ et variance $V = pX^2$, ce qui permet de comprendre l'analogie sous-jacente à de nombreuses méthodes multivariées entre mécanique et statistique.

Lorsque le nombre de variables commence à être élevé (disons plus de quatre), il est nécessaire de recourir à ces méthodes pour avoir une bonne description des relations générales entre variables, ce qui n'est pas forcément le but des expérimentations cliniques souvent plus orientées vers la comparaisons des variables.

Chapitre 3.

Statistiques inférentielles et tests

3.1 Estimation et intervalles de confiance

Lorsque les valeurs étudiées ne correspondent pas à la population entière mais seulement à un échantillon, il est judicieux de se demander quels résultats induire sur la population à partir des résultats sur l'échantillon. La théorie de l'estimation permet de répondre : il faut prendre

$$m_{pop} \ = \ m_{ech} \ \text{ et } \ V_{pop} \ = \ V_{ech} \times n/n-1$$

Pour savoir quel degré de confiance on peut accorder à ces résultats, la théorie statistique permet d'obtenir des encadrements probabilistes des résultats.

Ainsi à l'aide du calcul de la moyenne m et de l'écart-type σ pour n valeurs on sait fournir un intervalle centré autour de la moyenne $[m - \varepsilon \ ; \ m + \varepsilon]$ nommé *intervalle de confiance* de la moyenne pour un "risque" d'erreur α qui vaut généralement 5 % [soit encore pour un niveau de confiance 1-α].

Cela signifie que sur la base de l'échantillon étudié, la probabilité que l'intervalle contienne la "vraie" moyenne de la population est 1-α, qu'on exprime généralement en pourcent.

Il faut retenir que les formules ne sont pas les mêmes

- selon que l'échantillonnage s'effectue sans remise ou avec remise,
- selon que n est petit ou grand.

Ainsi pour $n \leqslant 30$ la demi-longueur ε est calculée à l'aide de la loi de *Student* au seuil $\alpha/2$ alors que pour $n > 30$ on la calcule à l'aide de la loi normale centrée réduite (toujours au seuil $\alpha/2$).

Par exemple avec une moyenne $m = 158.86\ mm$ et un écart-type $\sigma = 6.09$ pour la longueur de la rectrice centrale des gélinotes huppées ([Scherrer, p.335]), on pourra fournir les intervalles ($L =$"lower", $U =$ "upper")

α	ε	L_m	U_m
1	7.22462	151.635	166.085
5	4.96515	153.895	163.825
10	4.00387	154.856	162.864

pour $n = 9$

α	ε	L_m	U_m
1	2.21845	156.642	161.078
5	1.68803	157.172	160.548
10	1.41664	157.443	160.277

pour $n = 50$

Il existe aussi des formules pour l'intervalle de confiance de la médiane (rarement utilisées), pour l'intervalle de confiance de la variance et pour l'intervalle de confiance de l'écart-type. Pour ces deux derniers intervalles, c'est la loi du χ^2 qui intervient (on s'en douterait si on savait qu'une loi de χ^2 ce n'est jamais qu'une somme de carrés de lois normales !).

Il est également possible de déterminer l'intervalle de confiance d'une proportion lorsqu'on sait si l'échantillonnage s'effectue avec remise ou sans remise à condition de disposer de grands échantillons. Pour les petits échantillons, comme il n'existe pas de calcul pour la fonction de répartition inverse de la loi binomiale, les calculs sont moins automatisables (sauf à savoir programmer).

Ainsi pour les $N_{tot} = 1000$ cerfs de Virginie morts en 1975 ([McConell et coll.]) avec un échantillon de $n = 146$ cerfs morts dont $i = 41$ males (soit une proportion de $p = 28.08$ %), au risque de $\alpha = 5$ % on peut affirmer que l'intervalle de confiance pour le pourcentage de males morts est [20.98 % ; 35.18 %].

Lorsqu'on retourne la relation entre ε, n et σ ou entre ε, n et p il est possible de déterminer l'effectif n de l'échantillon pour une précision ε donnée de la moyenne ou du pourcentage.

Par exemple si nous reprenons les gélinottes pour avoir une précision relative de α% sur la longueur moyenne de la rectrice, il faut les effectifs suivants :

α	%	n (précis)	n (arrondi)
1	%	97.7495	98
5	%	56.5948	57
10	%	39.8598	40

Un *test d'hypothèse* est une procédure statistique permettant d'aboutir, en fonction de certaines règles de décision, au rejet [ou à l'acceptation] d'une *hypothèse* statistique de départ nommée hypothèse "nulle" et notée classiquement H_0 au dépend [ou au profit] de l'autre hypothèse ("hypothèse alternative").

Un test paramétrique suppose qu'on connait ou qu'on sait modéliser les paramètres (m, V, σ, p...) des distributions liées aux populations sous-jacentes alors que pour un test non paramétrique il n'est pas nécessaire de les spécifier. On utilise les tests non paramétriques lorsque les échantillons sont trés petits ($n \leqslant 7$) ou lorsque les prérequis des tests (variances bien estimées, distributions normales, variances égales...) ne sont pas satisfaits. Il faut certainement considérer le test du χ^2 d'indépendance et le test du χ^2 d'adéquation comme des tests non paramétriques (bien qu'ils s'adressent à des variables QL).

Il est important de préciser si le test est unilatéral comme $\theta_1 > \theta_2$ ou bilatéral comme $\theta_1 \neq \theta_2$ car dans le premier cas les tables sont lues au seuil α alors que dans le deuxième elles le sont au seuil $\alpha/2$.

Les tests permettent de comparer un échantillon (on emploie aussi le terme de "série") à une distribution théorique, deux échantillons indépendants ou appariés (deux séries de valeurs pour les mêmes individus). Il y a sans doute en tout une centaine de tests, certains connus uniquement par une ultraminorité de surexperts (!). Voici un tout petit tableau de quelques tests célébres :

Type Nom	pour
paramétriques	
test t (*Student*) / test z (loi $\mathcal{N}$)	deux moyennes petits / grands éch.
analyse de la variance	plus de deux moyennes
test F de Fisher	deux variances
test de Bartlett	plus de deux variances
non-paramétriques	
Kolgomorov-Smirnov	fonctions de répartition ($\Rightarrow$ normalité)
Wilcoxon-Mann-Whitney	égalité des rangs pour deux populations
test des signes	égalité de deux moyennes appariées
Kruskal-Wallis	l'égalité de plus de deux populations

En général un test statistique est accompagné d'une p-value ou *probabilité associée*. Plus cette p-value est faible et plus le test est significatif car on l'interpréte comme la probabilité d'obtenir "au hasard" un résultat "aussi extrême". En pratique, on rejete l'hypothèse nulle lorsque cette p-value est inférieure au risque de première espèce α.

Un test statistique et l'hypothèse statistique qui en découle est la traduction de l'hypothèse biologique. Il faut bien séparer les deux. Vouloir regarder si "le produit A induit une dégénérescence de..." signifie que l'hypothèse statistique nulle sera un test unilatéral ($m_{avec} < m_{sans}$). Penser que telle sous-population réagit différemment d'une autre signifie que l'hypothèse statistique débouche sur un test bilatéral ("la moyenne est significativement différente de 0").

Les logiciels fournissent en général deux résultats distincts pour les tests : un écart (ou "distance") que l'on compare à un seuil théorique en fonction du risque choisi et la p-value (ou *probabilité critique*) liée au dépassement de cette valeur. Au vu de ces deux indicateurs la décision est la même mais la p-value donne une meilleure idée de la comparaison. Prenons une analogie avec le sport en gériartie. Courir le 100 mètres à 70 ans en 25 secondes, c'est bien si la moyenne théorique est de 30 secondes. Par contre savoir que seul 3 % des personnes sont capables de faire mieux que 25 secondes quantifie autrement ce résultat.

Il est parfois difficile de s'y retrouver dans les tests statistiques

- parce qu'il y a beaucoup de tests en tout, certains pour un même type de comparaison,
- parce qu'il faut faire des choix sur les hypothèses avant d'utiliser les tests,
- parce que les conditions d'application sont variées et pas toujours simples à vérifier,
- parce que les résultats sont probabilistes ("on refuse au seuil de ...%") et non pas binaires ("on refuse").

Pour un travail de recherche, il est en général bon de se fier à la communauté et d'utiliser un test connu, qui sera interprétable rapidement par les experts du domaine.

Lorsqu'on apprend un nouveau test, un "truc" simple pour se familiariser avec le test est de l'appliquer

- à des données quasiment identiques comme `155 165 160 150 150` *vs.* `154 165 160 150 150` pour connaitre les "bons" cas,
- à des données très différentes comme `155 165 160 150 150` *vs.* `54 65 900 800 50` pour connaitre les cas extrêmes.

3.2 Tests paramétriques

Un *test de conformité* compare les paramétres observés de l'échantillon aux paramètres connus de la population ("de référence") comme par exemple la comparaison de la moyenne observée à la moyenne théorique, la comparaion d'une fréquence observée à une fréquence théorique alors qu'un *test d'homogénéité* viendra comparer les moyennes, fréquences... de deux échantillons.

Pour la comparaison en conformité de moyennes, on calcule en général un "écart réduit" qui correspond à la division de la différence des moyennes par un terme de variation. Le cas où la variance de la population de référence est connue se traite différemment du cas où elle est inconnue. Dans ce dernier cas, suivant que l'effectif est petit ($n < 30$) ou grand, on utilise une loi normale (test Z) ou une loi de Student à $n - 1$ ddl (test t).

C'est à peu près la même chose pour un test d'homogénéité de moyennes. On distingue cependant les échantillons indépendants des échantillons appariés pour lesquels les données sont liées, comme les valeurs "avant" et "après" pour un même individu, ou la partie "droite" et "gauche" d'un même organe. Dans le premier cas, suivant que les variances sont supposées égales (donc connues) ou non, le dénominateur de l'écart réduit est "pooled" ou estimé.

On trouvera sur les pages qui suivent des exemples de tests avec l'énoncé précis et la rédaction de la solution, sans le détail des calculs. Les formules utilisées sont en fin de manuel.

Exemples de tests de conformité

Comparaison à une moyenne théorique (variance connue)

On relève chez 9 patients une glycémie moyenne m de 1,12 g/l. Ces patients font partie d'une population pour laquelle la glycémie moyenne suit une loi normale de moyenne $m_0 = 1{,}0$ g/l et d'écart-type 0,1 g/l. L'échantillon est-il représentatif de la population ?

Réponse : l'hypothèse nulle est ici bilatérale : $m = m_0$. Nous sommes en variance connue, l'écart-réduit vaut $\varepsilon = 3.6$; au risque $\alpha = 5$ %, le seuil est 1.96 donc on rejette l'hypothèse H_0 à 5 %. La p-value associée à 3.6 est 0,0003 ce qui signifie que dans 3 cas sur 10 000 la différence est 3.6 ou plus, ce qui est "très rare". On rejete donc H_0 en meilleure connaissance de cause.

Comparaison à une moyenne théorique (variance inconnue)

Pour étudier un lot de fabrication de comprimés, on prélève au hasard 10 comprimés parmi les 30 000 comprimés fabriqués dans la journée et on les pèse. On obtient ainsi les valeurs de poids en grammes fournis par ce tableau :

```
0,81  0,84  0,83 0,80 0,85 0,86 0,85 0,83 0,84 0,80
```

Le poids moyen observé est-il conforme à la moyenne de production standard pour le poids qui est 0,83 g ?

Réponse : là encore l'hypothèse nulle est ici bilatérale : $m = m_0$. Nous sommes par contre en variance inconnue mais avec une loi normale pour le poids (seule supposition minimaliste "raisonnable").

La moyenne de l'échantillon est 0.831 g. L'écart-réduit lié à une estimation de la variance aboutit à la valeur $t_{obs} = 0.14834$. Pour un risque de 5 % avec 9 ddl, le seuil t_{seuil} vaut 2.26216 donc on accepte l'hypothèse H_0 à 5 %. La p-value associée à 0.14834 est trés fortement supérieure à 0,05 donc là encore on accepte H_0 en meilleure connaissance de cause.

Biologiquement, on dira que l'échantillon est conforme à la production standard.

Comparaison à une fréquence théorique

Une anomalie génétique touche en France 1 individu sur 1000. On a constaté pour une région donnée 57 personnes atteintes sur 50 000 naissances. Cette région est-elle représentative de la France entière ?

Réponse : l'hypothèse statistique nulle est ici l'hypothèse bilatérale : $f = f_0$.

f vaut 57/50000 et f_0 1/10000. Si la loi de l'anomalie génétique est une loi binomale (ce qui est la seule supposition minimaliste "raisonnable") de paramètres $n = 50000$ et $p = 1/1000$ alors l'écart réduit vaut 0.99044 et au risque de 5 % le seuil est 1.95996 donc on accepte H_0. La p-value associée à 0.99044 est 0.32 qui est largement supérieure à 0.05 donc là encore on accepte H_0 en meilleure connaissance de cause.

Exemples de tests d'homogénéité

Comparaison de moyennes (échantillons indépendants)

L'étude du cornicule gauche de *Myzus persicae* pour deux échantillons A et B dans des conditions d'élevage différentes fournit les valeurs suivantes en unités micrométriques :

```
Ech. A 313 257 322 332 302 ... 290          (30 valeurs)
Ech. B 346 279 228 306 246 ... 250 334      (29 valeurs)
```

Les conditions d'élevage ont-elles une influence sur la longueur du cornicule ?

Réponse : l'écart réduit est de 1.03 donc au risque de 5 % soit le seuil 1.96 on accepte l'hypothèse que les conditions d'élevage n'ont pas d'influence sur la longueur du cornicule.

Comparaison de moyennes (échantillons appariés)

Lors d'une étude sur un lot de 20 moules communes (*Mytilus galloprovincialis*) on mesure les valves gauches et droites :

```
V. gauche   89.0  109.0  101.9 ...  121.6   (20 valeurs)
V. droite   88.0  105.6  102.7 ...  124.6   (20 valeurs)
```

Les valves sont-elles symétriques ?

Réponse : on a affaire ici à des échantillons appariés puis la valve gauche numéro i et la valve droite numéro i appartiennent à la même moule. A l'aide de la variable "différence des longueurs" on effectue un test t qui aboutit à une valeur 1,16 ; comme le seuil au risque de 5 % est de 2,1 on peut conclure qu'il n'y a pas de différences significatives entre les parties droite et gauche des valves.

Comparaison de variances (petits échantillons)

Pour étudier l'action de la digitonine sur des embryons de *Rana platyrrhina* on prépare deux séries d'échantillons : une série témoin et une série traitée.

Les volumes en mm^3 des embryons sont comme suit :

```
Temoin      3.36 3.52 4.10 4.02 4.40 4.30 4.36 3.94
Traitement  3.36 3.82 4.46 4.52 5.08 5.20 5.50 5.06
```

Peut-on dire que la digitonine agit sur le volume ?

Réponse : le rapport de variance vaut 3,65 ; le seuil F au risque 5 % est 3,8. On accepte donc l'hypothèse qu'il n'y a pas de différence.

Comparaison de fréquences

Un étudiant de Maitrise (soit aujourd'hui le niveau "Master 1" dans le cadre de la réforme LMD) étudie un lot de cocons de *Bombyx mori* sous rayonnement X (2000 roentgen). Il a consigné sur deux périodes l'éclosion des papillons normaux et mutants :

	Mutants	Normaux	
Janvier	30	220	(p1 = 0.136)
Juillet	12	150	(p7 = 0.080)

Toutefois, comme il trouve que cela ne fait pas beaucoup, il décide de multiplier tous les résultats par 10. Obtient-on la même conclusion ?

Réponse : Bien sûr que non ! Une lecture soignée des formules permet de voir que si les données sont multipliées par 10, les fréquences ne sont pas modifiées mais que l'écart-réduit est multiplié par $\sqrt{10}$. Avec les données originales, cet écart vaut 1,68 donc inférieur au seuil 1.96 et la décision est "les fréquences de mutants sont égales" alors qu'avec les données multipliées par 10 l'écart vaut 5,31 et donc la décision est "les fréquences de mutants sont différentes".

3.3 Tests non paramétriques

Le premier test non paramétrique à connaitre est le test de Kolmogorov-Smirnov. Il permet de comparer deux distributions de fréquences relatives cumulées. On peut s'en servir pour comparer deux échantillons par exemples pour des QT découpées en classes, mais on peut aussi l'utiliser comme *test de normalité* si on compare la distribution de la QT avec la fonction de répartition de la loi normale.

Pour des petits échantillons d'effectif total n_1 et n_2 inférieurs à 25, le test consiste à comparer $n_1 n_2 D_{obs}$ et une valeur seuil nommée $n_1 n_2 D_\alpha$ lue dans une table où D_{obs} est la plus grande différence entre effectifs relatifs cumulés. Pour les grands échantillons (n_1 ou n_2 supérieur à 25), on compare directement D_{obs} au seuil $K_\alpha \sqrt{(n_1 + n_2)/n_1 n_2}$ où le facteur de correction K_α se calcule par $\sqrt{(-log(\alpha/2)/2}$.

Soient à traiter les données suivantes concernant la superficie du domaine vital des ours noirs (*Ursus americanus*) mâles et femelles, superficies en km^2

Males	94	504	173	560	274	168			
Femelles	37	72	60	49	18	50	102	49	20

Après avoir défini des classes de surfaces (comme [18,20],]20,37]...]504,560]) on ordonne les données et on compte le nombre d'observations par classe pour chaque sexe et on cumule les effectifs relatifs obtenus pour chaque sexe. On cherche ensuite la plus grande différence obtenue entre effectifs relatifs cumulés. On trouve ici $D_{obs} = 0.888$; comme il s'agit de petits échantillons on multiplie par 6×9 ce qui nous donne 47.952. Au risque α=5 % la table fournit le seuil 39 et on refuse donc l'hypothèse que le domaine vital des males ne différe pas de celui des femelles.

Un deuxièmes test non paramétrique à connaitre [ou plutot à savoir utiliser via un logiciel statistique]) est celui de Wilcoxon, Mann et Whitney. Il se base sur l'analyse des rangs globaux des observations (rangs éventuellement fractionnaires) pondérés par un score qui est fonction du nombre d'observations différentes et égales. Le cumul des scores aboutit à une valeur nommée U que l'on compare à une valeur seuil dans une table. Ce qui complique un peu l'utilisation de ce test c'est le fait qu'il faut séparer

- les très petits échantillons pour n_1 et n_2 inférieurs à 8,
- les petits échantillons pour n_1 et n_2 inférieurs à 20,
- les grands échantillons pour n_1 ou n_2 supérieur à 20,

Ainsi pour "nos" ours, on se pose la question de savoir si le domaine vital des ours noirs est plus étendu que celui des femelles. On formule donc l'hypothèse $H_0 : p(\text{"}x_M > y_F\text{"}) = 1/2$ que la probabilité qu'un domaine vital pris au hasard pour un mal soit supérieur à celui d'une femelle est égale à 0,5.

On ordonne nos données et on calcule les scores, soit le tableau :

$Valeur$	$Sexe$	$Score$
18	F	0
20	F	0
37	F	0
49	F	0
50	F	0
60	F	0
72	F	0
94	M	8
102	F	1
168	M	9
173	M	9
274	M	9
504	M	9
560	M	9

Tous calculs faits, le test fournit les valeurs U_F=1, U_M=53. On prend donc U=1 et comme la valeur de la table est 12 au risque de 5 % (car le test est unilatéral), on rejette donc l'hypothèse d'où : le domaine vital des ours noirs est plus étendu que celui des femelles.

Le "grand" test non paramétrique pour les échantillons appariés est le *test des signes*. Au lieu, comme en paramétrique de reposer sur l'hypothèse que la différence entre données appariées suit une loi normale, ce test suppose que les différences entre données appariées ont autant de chances d'être positives que négatives et l'hypothèse nulle est donc $H_0 : p(\text{"}d > 0\text{"}) = p(\text{"}d < 0\text{"}) = 1/2$ et on se ramène donc à une loi binomiale de paramètres n et $p = 1/2$. Atttention : ici le nombre n ne désigne pas l'effectif commun mais le nombre de couples pour lesquels la différence est non nulle.

Pour les grands échantillons ($n \geqslant 30$) on utilise l'approximation normale de la loi binomiale alors que pour les petits échantillons il faut faire des calculs de "bouts de chandelle" (voir [Scherrer] page 525).

Il existe aussi :

- un test non paramétrique de comparaison de médianes, qui aboutit à un calcul dont le seuil maximal autorisé est une valeur de χ^2,
- un test non paramétrique de Wilcoxon pour échantillons appariés, conseillé lorsque les conditions d'application du test paramétrique de comparaison de moyennes ne sont pas respectées.

Un autre test qui n'est pas traditionnellement considéré comme un test non paramétrique est le test du χ^2. Il se décline en deux versions :

- le χ^2 d'adéquation (ou χ^2 d'ajustement) qui permet de comparer une distribution de fréquences observées à une distribution de fréquences théoriques,
- le χ^2 d'indépendance qui compare les distributions relatives à deux caractères (quantitatifs groupés en classe ou qualitatifs) présentant plusieurs modalités et définis sur une même population qui sont comparées ; les données utilisées correspondent au tableau des effectifs observés pour les deux caractères comparés qui est donc le tableau du tri-croisé ou "table de contingence".

Il y a des conditions d'applications à respecter pour appliquer un χ^2 sous peine d'obtenir des résultats non cohérents ou infinis : l'effectif total doit être au moins de 50, il faut que chaque effectif de distribution ou de croisement soit supérieur à 5...

Prenons un exemple classique de χ^2 d'adéquation, celui lié au nombre de filles dans une famille de 5 enfants, qui se transpose facilement à toute observation binaire (absence/présence, oui/non, plus/moins...) pour une série de 5 sujets. Pour 200 familles de 5 enfants interrogées, on dresse le tableau suivant du nombre n_i de familles avec x_i filles :

x_i	0	1	2	3	4	5
n_i	20	30	70	60	15	5

Dans une famille, chaque enfant a la probabilité $p = 0.5$ d'être une fille. La loi de l'évènement "on compte 1 si l'enfant est une fille" est donc modélisée par la loi de bernouilli $b(0.5)$. La loi du comptage du nombre de filles dans un famille de 5 enfants suit alors une loi binomiale $\mathcal{B}(5, 0.5)$ ce qui permet de calculer le nombre théorique t_i de familles avec x_i filles :

x_i	0	1	2	3	4	5
t_i	7	31	62	62	31	7

La valeur du χ^2 observé qui est une somme pondérée de carrés des différences entre les t_i et les n_i a pour valeur 34.1 ; la valeur "maximale autorisée" lue dans la table à 5 % est 9.5 pour 4 *ddl* donc au risque de 5 % on refuse l'hypothèse que la distribution du nombre de filles dans les 200 familles correspond à une distribution binomiale.

Passons maintenant au χ^2 d'indépendance. Voici un tableau de contingence pour un naufrage célèbre, celui du *Titanic*. Ce tableau fournit en effectifs absolus le résultat du tri croisé entre la variable "classe de cabine" des personnes présentes (passagers et membres d'équipage) et variable de "survie".

	équipage	1ère classe	seconde classe	troisième classe
Décédés	673	122	167	528
Survivants	212	203	118	178

La question posée ici est la dépendance éventuelle entre ces deux variables*.

L'*hypothèse d'indépendance* consiste à dire que les effectifs absolus ne doivent dépendre que des totaux du tableau, soient les sommes

	équipage	1ère classe	seconde classe	troisième classe	
Décédés					1490
Survivants					711
	885	325	285	706	2201

Un calcul simple permet de prendre comme effectif absolu le produit de la somme en ligne par la somme en colonne pondéré par le total général soient les effectifs théoriques (arrondis)

	équipage	1ère classe	seconde classe	troisième classe
Décédés	599	220	193	478
Survivants	286	105	92	228

La valeur du χ^2 d'indépendance qui là aussi est une somme pondérée de carrés des différences entre les effectifs théoriques et réels a pour valeur 190.4 ; la valeur "maximale autorisée" lue dans la table à 5 % est 7.8 pour (4-1)*(2-1) *ddl* donc au risque de 5 % on refuse l'hypothèse qu'il y a indépendance. L'analyse des *contributions signées* met principalement en évidence la dépendance entre la modalité "1ère classe" et la modalité "Survivants" sous la forme d'une surabondance (contribution de + 91.5 soit 48 % du χ^2) ainsi qu'une sousabondance entre "1ère classe" et "Décédés" (contribution de -43.7 soit 23 % du χ^2).

* on remarquera que pour deux QT on cherche une liaison [linéaire] alors que pour deux QL on recherche une indépendance.

Chapitre 4.

Graphiques, protocoles, rédaction et logiciels

4.1 Courbes et graphiques

On se pose souvent la question de savoir s'il faut mettre des graphiques et lequels. Plus finement, la question est de savoir s'il faut mettre des graphiques à la place des résultats chiffrés. La réponse n'est pas la même suivant que l'on veuille être exhaustif ou seulement démonstratif.

Dans le cadre d'une enquête statistique ou d'une étude clinique, tous les résultats chiffrés devront être "muris" et accompagné des graphiques pertinents. Cela signifie qu'il faudra penser

- à l'ordre d'affichage des résultats,
- au nombre de chiffres après la virgule,
- à la lisibilité (si, si) des tableaux de chiffres,
- au schéma de lecture (tous les tableaux puis tous les graphiques ou tableaux et graphiques en alternance),
- aux couleurs des graphiques si le document est destiné à circuler (et donc à être photocopié sans couleur).

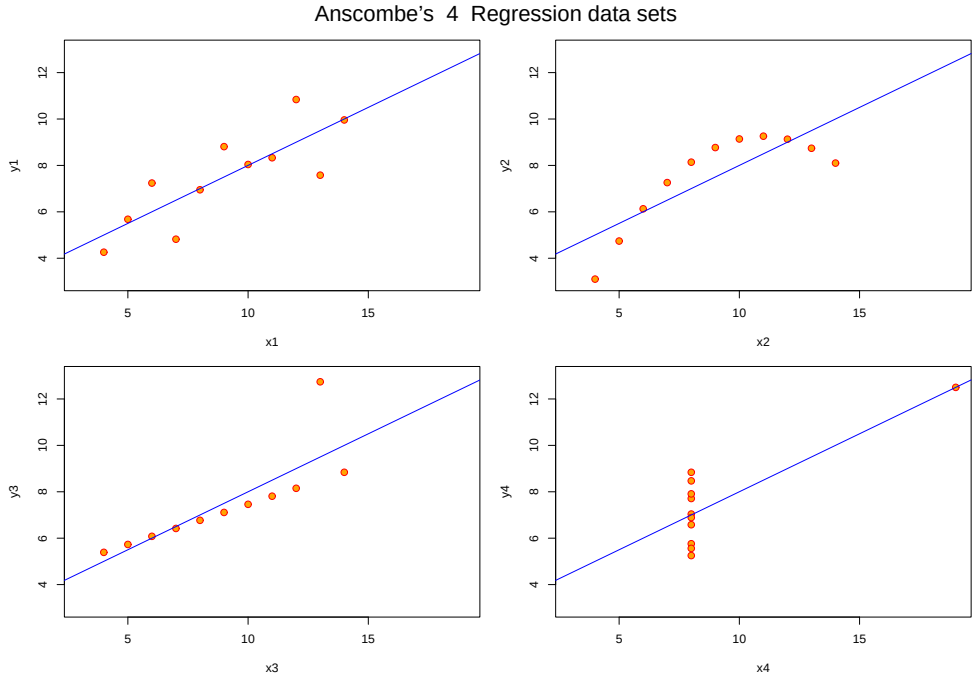
Pour un article où le nombre de pages est souvent limité, il faut choisir soigneusement quels graphiques vont illustrer les tableaux chiffrés car il est hors de question de fournir des graphiques sans les valeurs numériques associées ; par contre on peut penser à préparer des graphiques complémentaires si on doit exposer les résultats de l'article.

Penser que les graphiques sont inutiles, ou la "cerise sur le gateau" est une grave erreur. Dans son article célèbre de 1973, F. J. Anscombe a proposé le jeu de données suivant

ID	X	Y1	Y2	Y3	X4	Y4
a	4	4.26	3.10	5.39	19	12.50
b	5	5.68	4.74	5.73	8	6.89
c	6	7.24	6.13	6.08	8	5.25
d	7	4.82	7.26	6.42	8	7.91
e	8	6.95	8.14	6.77	8	5.76
f	9	8.81	8.77	7.11	8	8.84
g	10	8.04	9.14	7.46	8	6.58
h	11	8.33	9.26	7.81	8	8.47
i	12	10.84	9.13	8.15	8	5.56
j	13	7.58	8.74	12.74	8	7.71
k	14	9.96	8.10	8.84	8	7.04

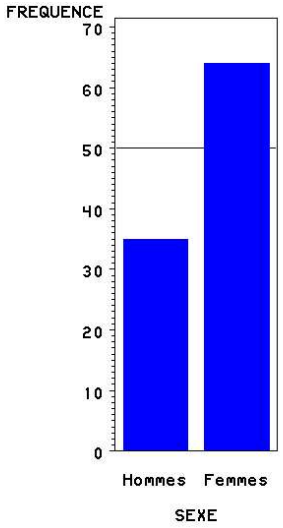
Si on étudie (rapidement) les séries Y on arrive à la conclusion qu'elles se ressemblent trés fortement puisqu'elles ont toutes 7.5 comme moyenne et 1.94 comme écart-type.

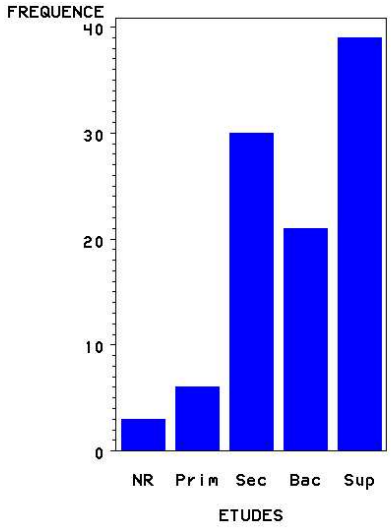
Hélas, le tracé des séries Y en fonction des séries X prouve le contraire :



Les graphiques usuels liés aux QL sont les histogrammes (de fréquences, pas de valeurs !) qu'on utilisera plutot sous forme de batons verticaux dans des cadres de même échelle ce qui permet une comparaison des graphiques.

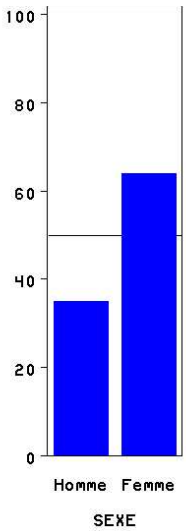
Par exemple il serait très mauvais de fournir les histogrammes suivants

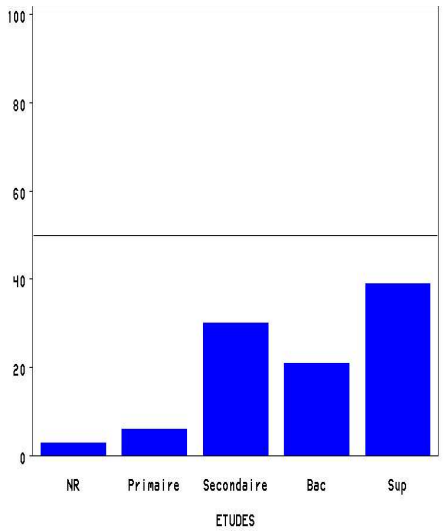




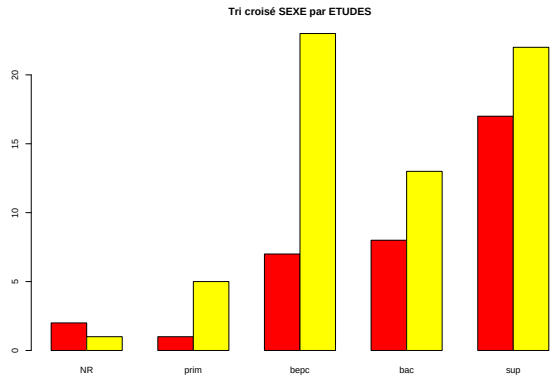
car une lecture attentive des échelles montre que les effectifs des modalités de droite ne sont pas aussi importants que semblent indiquer les histogrammes.

C'est flagrant si on utilise les "bons" histogrammes avec des axes normalisés :



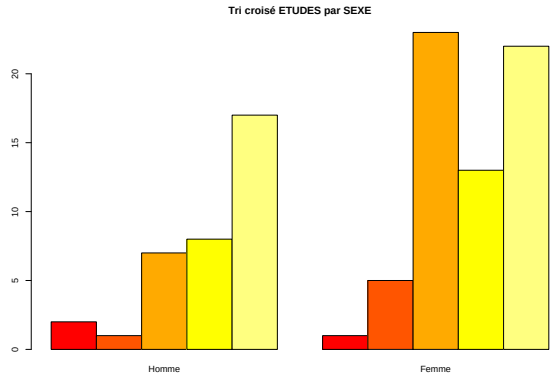


L'intérêt des histogrammes en batons (contrairement aux diagrammes circulaires) est de pouvoir être cumulés, empilés et ventilés. Il faut à nouveau se poser la question du sens du cumul, de la redondance des histogrammes de tris croisés par rapport à ceux des tris à plat. Ainsi un "mauvais" histogramme de tri croisé du niveau d'études "versus" la variable sexe est



car il ne fait que répéter l'information vue au niveau du tri à plat du sexe à savoir *il y a deux fois plus de femmes* (en jaune et à droite) *que d'hommes* (en rouge et à gauche).

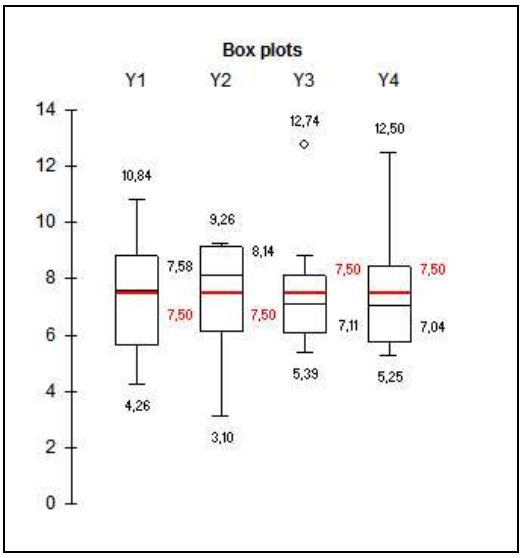
C'est pourquoi ici "le" bon histogramme est



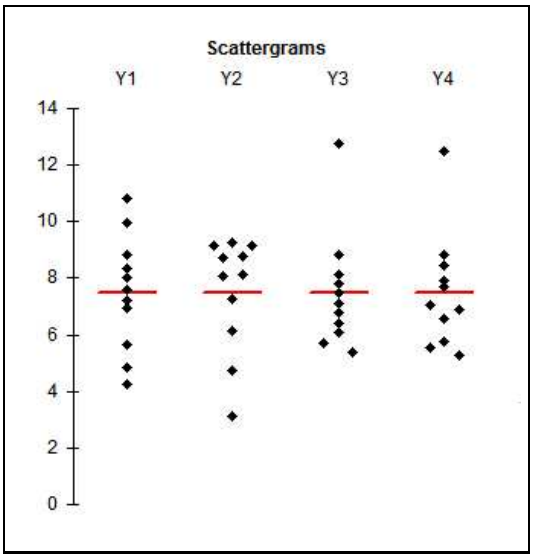
car il montre à gauche une progression croissante des effectifs pour les hommes alors qu'à droite il y a un "trou" dans les effectifs des femmes, ce qui indique qu'il n'y a pas la même distribution du niveau d'études pour les hommes et les femmes.

Aux courbes en "nuages de points" il est d'usage aujourd'hui d'ajouter des nouveaux graphiques nommés "scattergrams", "boxplots" etc. traduits en principe par... scattergrammes (!) et "boites à moustaches" (re!) dont le but est d'aider à visualiser la tendance et la dispersion des données. Voici par exemple les graphiques fournis par Statbox pour les séries Y d'Ancombe.

A l'aide des "boites à moustaches" on voit légèrement que la série 4 différe des autres



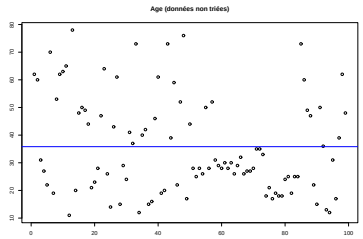
mais avec les scattergrammes on se rend mieux compte de la répartition des points :



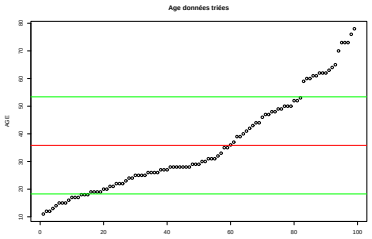
Pour des données entières, comme des ages (ans), des tailles (cm) les diagrammes de tige et feuilles (*stem and lead*) peuvent se révéler plus parlants que les histogrammes classiques, comme pour l'age des hommes et des femmes de notre dossier ELF :

```
AGES DES HOMMES                 AGES DES FEMMES

Tige Nb | Feuilles              Tige Nb | Feuilles
 1 ( 4) | 2678                   1 (15) | 123455577889999
 2 (14) | 00225667888899         2 (17) | 11234455566778889
 3 ( 4) | 0235                   3 ( 9) | 011156799
 4 ( 5  | 23779                  4 ( 8) | 01446889
 5 ( 2) | 22                     5 ( 5) | 00039
 6 ( 5) | 02245                  6 ( 5) | 01123
 7 ( 1) | 8                      7 ( 5) | 03336

 Total 35                        Total 64
```

On n'oubliera pas que des données triées sont parfois plus explicites que les données originales. Ainsi nos 99 ages présentés dans l'ordre du fichier peuvent faire croire à une répartition "au hasard"



alors que ces mêmes données triés par ordre croissant font apparaitre une répartition presque uniforme :



4.2 Protocoles et Rédaction

Rédiger est un travail pénible et long (relectures) surtout s'il faut écrire en anglais, ce qui n'est pas une mince affaire (à ne pas traduire par "*mean affair*" !). N'étant pas compétent en médecine les quelques remarques qui suivent proviennent de mon expérience de relecture d'articles scientifiques d'autres domaines et se focalisent sur quelques points fondamentaux.

La description du protocole est trés importante et se fait en général selon des normes en vigueur dans la communauté de recherche. Décrire un protocole ne se fait donc pas à la légère ni dans n'importe quel ordre.

La ponctuation est une "fourche caudine" car les "p'tits français" mettent souvent des espaces doubles aprés **:!** et **?** alors que la norme anglaise et américaine mettent un espace simple après chaque symbole de ponctuation. Ainsi **"Results : we have ..."** ira forcément "énerver" un correcteur alors que la phrase **"Results: we have..."** passera sans anicroche.

La rédaction est souvent limitée en nombre de pages et on sacrifie parfois le détail des méthodes statistiques au profit du domaine d'étude ce qui est parfois pénalisé ("method not specified","statistical test not defined"...).

Lors de la rédaction, il faut se rappeler que les variables sont vues au travers d'indicateurs et mettent en jeu des populations, des sous-populations. De nombreux adjectifs éventuellement qualifiés par des adverbes sont disponibles mais attention au champ sémantique. Ainsi "la population est très hétérogéne pour ..." a un sens et correspond à une forte dispersion de la variable X mais on ne dit pas "qu'une population est dispersée" ou "qu'une variable est hétérogéne".

La tendance et la dispersion des QT induisent des termes comme grand, trés grand, relativement grand ainsi que dispersé, fluctant ce qui permet de parler d'un [fort] effet de taille, d'une grande homogénité... Pour les QL on se référera à l'équirépartition ou au contraire à la disproportion, on mettra en évidence une modalité majoritaire ou seulement des consensus, reflétant une disparité ou une conjonction d'avis, de propriétés...

Lors de la rédaction des conclusions sur des tests statistiques, on distinguera l'hypothèse biologique ("le produit X a une influence sur Y") de l'hypothèse statistique (la moyenne des différences est significativement nulle) et la conclusion statistique ("au risque de x % on accepte l'hypothèse") de la conclusion biologique qu'on viendra souvent nuancer en fonction de la taille de l'échantillon ("il semblerait donc qu'on puisse dire que...").

4.3 Logiciels

Une question qui ne manque pas de se poser lorsqu'on doit réaliser une étude statistique est celle du logiciel. Lequel utiliser ? et surtout, pourquoi ? La plupart du temps, il y a déjà au moins un logiciel de statistiques (ou utilisé comme tel) dans le service, dans le laboratoire et la question ne se pose plus : on utilisera celui-là.

Il faut toutefois savoir que si l'on veut publier ses résultats certains logiciels font loi (SAS, SPSS, Statistica, R, S...) alors que d'autres sont déconseillés car "dangereux" (dont Excel et la plupart des "add-on" pour Excel). Il suffit de lire les articles de recherche de son domaine pour voir quels logiciels sont utilisés et si possible les faire acheter. On trouvera dans les références Web fournies en annexe une explication des "méfaits" d'Excel.

Excel n'est pas en soi un mauvais logiciel. De plus il est souvent disponible sur toutes les machines. Il permet à la fois de saisir les données, de faire des calculs, des graphiques. Nous conseillons de l'utiliser pour la saisie et la vérification des données, le survol des calculs (sans trop en attendre), l'affichage graphique rapide (avec ses imprécisions).

Que reproche-t-on à Excel ? Tout d'abord de ne pas être un logiciel de statistiques. Excel ne fournit aucun moyen de définir une variable statistique avec la *contrainte de type* que cela impose : les données doivent se suivre (sans trou ou "cellule vide", ni graphique entre deux données) et être de même nature (toutes numériques ou toutes caractères). Excel est de plus incomplet et incorrect car s'il fournit quelques fonctions et graphiques orientés statistiques il ne fournit pas toute la panoplie de base (pas de coefficient de variation ni de test, pas de boite à moustaches ni de tige et feuilles) et certaines fonctions sont incorrectes pour des "petites variations de grandes valeurs".

Par exemple, tout le monde sait que si on ajoute la valeur c à toutes les données de X, la moyenne de X augmente de c alors que sa variance ne change pas soit les formules $moy(X + c) = moy(X) + c$ et $var(X + c) = var(X)$. Prenons pour X les valeurs de 1 à 4 et mettons dans Y les valeurs de X augmentées de 99999996.

Demandons à Excel la variance de X et de Y : on obtient respectivement 1.25 et 0 !

Pour bien choisir un logiciel, il faut (indépendamment du prix) savoir si on doit faire des calculs "en routine" ou "au petit bonheur". Dans le premier cas on veut utiliser la même démarche, faire les mêmes calculs sur plusieurs groupes, plusieurs variables similaires (par exemple une même variable ventilée sous différentes conditions expérimentales) puis comparer les résultats. Dans le second cas on veut réalisee une analyse statistique sur le coup, sans qu'on ait à reproduire les manipulations ni sans avoir à optimiser les clics de souris.

Les "bons" logiciels cités (*Sas*, *Spss*, *Statistica*, *R*, *S*...) fournissent des moyens d'automatiser une suite de traitements où seul change le nom des données – que ce soit le nom des variables ou le nom du fichier des données. Au prix d'un effort d'apprentissage et de [re]codage des données et traitements, on obtient une automatisation des calculs très efficace qui permet de ne pas hésiter sur les calculs à effectuer.

Prenons comme exemple le traitement des poids des coeurs pour les rats soumis à une restriction alimentaire. Il y a 30 données réparties en 5 groupes de 6 rats définis comme suit

Groupe	*% de restriction*	Commentaire
1	0	(témoin)
2	25	
3	50	
4	75	
5	100	(jeûne)

L'étude statistique devra certainement à faire l'étude globale et par groupe, le tracé des graphiques associés, la recherche de corrélation linéaire entres groupes et entre le degré de restriction et la perte de poids, la comparaison des moyennes voir une analyse de la variance...

Si *Excel* permet de "bricoler" à la main et de transférer rapidement des formules d'une série à l'autre avec sa fameuse "poignée de recopie", *Excel* n'a pas toutes les fonctionnalités ni la vraie "batterie de tests" que peuvent fournir *Sas*, *Statistica*, *Spss*...

L'automatisation des tâches prend tout son sens pour de nombreuses variables. Ainsi l'étude menée sur le coeur des rats soumis à restriction alimentaire doit être reproduite pour le poids total final, le foie, la rate, les reins. De plus une analyse similaire doit être menée sur les muscles, sur la nourriture ingérée, le poids total...

Avec *Sas* par exemple, on écrira la suite des traitements à effectuer dans une *macro* et on l'utlisera pour chacun des organes. La production de l'ensemble des calculs et des résultats se résumera sans doute alors à

```
%macro traiteRats(fichier) ;
  ...  <== ici les instructions de traitement
%mend traiteRats ;

%traiteRats(coeur) ;
%traiteRats(foie) ;
%traiteRats(rate) ;
...
```

Un bon logiciel de statistiques se doit de

- mettre à disposition toutes les grandes méthodes statistiques usuelles,
- fournir une aide à la démarche statistique, à la compréhension et à l'utilisation des méthodes,
- prévoir des jeux d'essais pour tester rapidement les concepts, méthodes et types de graphiques,
- permettre une automatisation des traitements.

C'est pourquoi (sans avantage commercial et sans publicité) nous recommandons *Sas*, *Statistica*, *Spad*, *Spss*, *S* et sa version gratuite nommée *R* qui sont les seuls à réunir toutes ces qualités.

ANNEXES

Bibliographie

B. Scherrer
Biostatistique
Gaetan Morin éditeur, 1984.

S. Frontier, D. Davoult, V. Gentilhomme, Y. Lagadeuc
Statistique pour les sciences de la vie et de l'environnement, cours et exercices corrigés
Dunod, Paris 2001.

Y. Dodge
Statistique, dictionnaire encyclopédique
Dunod, Paris 1993.

J.M. Legay
Exercices de Statistique pour Biologistes
Flammarion, 1966.

D. Foata, A. Fuchs
Calcul des probabilités
Dunod, 1998.

D. C. Hoaglin, F. Mosteller, J. W. Tukey
Understanding robust and exploratory data analysis
John Wiley & Sons, 2000.

L. Lebart, A. Morineau, M. Piron
Statistique exploratoire multidimensionnelle
2^{eme} édition, *Dunod, 1997.*

Références Web

http ://spiral.univ-lyon1.fr/mathsv/

Dans la partie gauche (Cours) cliquer sur "Probabibilité-Statistique". On trouve dans ces pages Web le rappel de cours et des formules ainsi que des exercices corrigés détaillés. Ce site est plutot à considérer comme un aide-mémoire rapide de niveau L1.

http ://www.lsp.ups-tlse.fr/Besse/enseignement.html

L'URL indiquée est la page principale des cours du Pr. Philippe BESSE. Suite logique et approfondie de l'URL précédente, il s'agit encore de cours disons "académiques" détaillés.On y trouve aussi une initiation aux logiciesl SAS er R et un cours intéressant sur les des données d'expression génomique fournies par les biopuces ("$microarrays$").

http ://www.math-info.univ-paris5.fr/ smel/

Ce site est plus particulièrement destiné au milieu médical. Le cours en ligne de *Statistique Médicale En Ligne* est particuliérement bien fait, même s'il est un peut succint à mon gout. Le site comprend de plus un lexique des termes statistiques, des articles médicaux plubliés, des données réelles qui servent pour les exemples.

http ://tecfa.unige.ch/staf/staf-d/merino/UDO/

Ce site est en français, limité à certains tests mais il est bien détaillé.

http ://www.psychstat.smsu.edu/sbk00.htm

En anglais, ce site est assez complet. En particulier on y retrouve assez facilement le vocabulaire anglais utilisé en statistiques. Il correspond au livre de *Stockburger*.

http ://www.agro-montpellier.fr/cnam-lr/statnet/

Ce cours, très "propre", d'une collaboration Cnam, Agro Montpellier et Université de Montpellier se lit très bien. De plus certaines séquences vidées (Real Player) permettent de "lire" différement le cours et le lexique, à l'adresse

```
http ://www.agro-montpellier.fr/cnam-lr/statnet/mod6/mod6lx.htm
```

permet de retrouver rapidement une notion ou une formule oubliée.

http ://members.aol.com/johnp71/javastat.html

On trouve à cette adresse plus de 600 liens sur des pages Web qui effectuent des calculs statistiques en ligne, que ce soit en java, javascript ou autre langage. On y trouve notamment trois références pour savoir quel test choisir :

```
http ://www.graphpad.com/www/Book/Choose.htm
http ://www.socialresearchmethods.net/selstat/ssstart.htm
http ://members.aol.com/statware/pubpage.htm#HERE009
```

Les autres liens permettent, après quelques essais, de trouver sans utiliser de logiciel particulier (ou pour vérifier les résultats d'un logiciel) de faire "en ligne" les calculs, souvent par simple copie/coller des données, comme par exemple le site

```
http ://www.statlets.com/log_in.htm
```

http ://www.info.univ-angers.fr/pub/gh/vitrine/Democgi/loisStatp.htm

Cette page permet de calculer rapidement les effectifs théoriques absolus pour une loi discrète dont on connait les paramétres. Par exemple je m'en suis servi pour le calcul du χ^2 de l'exemple "filles dans une famille de 5 enfants".

Vous trouverez bien sûr cette pages de liens (cherchez le mot "repères" ou le mot "CHU") et bien d'autres références à l'adresse

http ://www.info.univ-angers.fr/pub/gh/wstat/statgen.htm

Formules mathématiques

Analyse univariée QT de X (soit n valeurs X_i)

Paramètre de	*Nom*	*Notation*	*Formule*
Taille	Effectif	n ou n_X	(nombre de valeurs)
Position	Moyenne	m ou m_X	$n \times m \ = \Sigma X_i$
(tendance)	Médiane	$q_2(X)$ ou $q_{0.50}$	$card(\ \{i\ ;\ X_i \leqslant q_2\}\ ) = n/2$
Dispersion	Variance	V ou V_X	$V = moy(\ (X-m)^2\ )$
(variation)	Ecart-type	σ ou σ_X	$\sigma^2 = V$
	Coefficient de variation	cdv	σ/m
	Ecart inter-quartiles	Q	$q_{0.75} - q_{0.25}$

La variance de la population est estimée par $V \times n/n-1$.

Analyse bivariée QT de X et Y

Remarque : les valeurs Y_i sont appariées à celles de X soit n couples (X_i, Y_i).

Nom	*Notation*	*Formule*
Covariance	cov_{XY}	$m_{XY} - m_X m_Y$
Coefficient de corrélation linéaire	ρ_{XY}	$cov_{XY}/\sigma_X\sigma_Y$

Equation de la régression linéaire de Y par rapport à X si $|\rho|$ proche de 1 :

$$Y = aX + b \ \text{ avec } \ a = \rho\sigma_Y/\sigma_X \ \text{ et } \ b = m_Y - am_X$$

Analyse univariée QL de X

Soient n valeurs de X correspondant à p modalités q_j.

Nom	*Notation*	*Formule*
Effectif absolu de la modalité j	n_j	$card(\ \{\ i\ ;\ X_i = q_j\}\ )$
Effectif total pour X	N	$\Sigma\ n_j$
Effectif relatif de la modalité j	f_j	n_j/N
Effectif cumulé (absolu) de la modalité j	c_j	$n_1 + n_2\ +\ ...\ + n_j$

Analyse bivariée QL de X et Y

Remarque : les valeurs Y_i sont appariées à celles de X soit n couples (X_i, Y_i).

Les n valeurs de Y correspondant à r modalités t_k.

On effectue le croisement des p modalités j de X mises en ligne et des r modalités k de Y mises en colonne.

Nom	*Notation*	*Formule*
Effectif absolu du croisement de q_j et t_k	$n_{j,k}$	$card(\{i; X_i = q_j\ et\ Y_i = t_k\})$
Effectif absolu de la modalité (ligne) q_j	$n_{j.}$	$\Sigma_k\ n_{j,k}$
Effectif absolu de la modalité (colonne) t_k	$n_{.k}$	$\Sigma_j\ n_{j,k}$
Effectif total	N ou $n_{..}$	$\Sigma_j\ n_{j.}$ ou $\Sigma_k\ n_{.k}$
Effectif relatif de q_j et t_k p.r. à la ligne j		$n_{j,k}/n_{j.}$
Effectif relatif de q_j et t_k p.r. à la colonne j		$n_{j,k}/n_{.k}$
Effectif relatif de q_j et t_k p.r. au total	$f_{j,k}$	$n_{j,k}/n_{..}$
Effectif relatif de la modalité (ligne) q_j	$f_{j.}$	$n_{j.}/n_{..}$
Effectif relatif de la modalité (colonne) t_k	$f_{.k}$	$n_{.k}/n_{..}$

Intervalle bilatéral de confiance d'une moyenne

Soit X un échantillon QT avec n valeurs, de moyenne m et d'écart-type σ.

Soit α le risque considéré et N l'effectif de la population totale.

Pour un échantillonnage sans remise $\quad s = \sigma/\sqrt{n} \times \sqrt{(N-n)/N}$
Pour un échantillonnage avec remise $\quad s = \sigma/\sqrt{n-1}$

Cas des petits effectifs ($n \leqslant 30$) :
$I_\alpha = [m - \varepsilon \ ; \ m + \varepsilon]$ avec $\varepsilon = s \ t_{alpha/2}$
t est la fonction de répartition inverse de la loi de Student avec $\nu = n - 1$ ddl.

Cas des "grands" effectifs ($n \geqslant 30$) :
$I_\alpha = [m - \varepsilon \ ; \ m + \varepsilon]$ avec $\varepsilon = s \ Z_{alpha/2}$
où Z est la fonction de répartition inverse de la loi normale soit 1.96 pour $\alpha = 5$ %.

Intervalle de confiance d'une proportion

Soit X un échantillon de n valeurs dont a sont marquées,
extrait d'une population de N valeurs dont A sont marquées.

Soit α le risque considéré.

Pour un échantillonnage avec remise $\quad s^2 = p(1-p)/(n-1)$
Pour un échantillonnage sans remise $\quad s^2 = p(1-p)/(n-1) \times \ (N-n)/N$

Cas des "grands" effectifs (n subordonné à p) :
$I_\alpha = [p - \delta \ ; \ p + \delta]$ avec $\delta = 1/2n + s \ Z_{alpha/2}$
où Z est la fonction de répartition inverse de la loi normale soit 1.96 pour $\alpha = 5$ %.

Détermination d'effectif pour une précision donnée

Sachant ε, m et σ on prend $\beta = t_{alpha/2}$ ou $\beta = Z_{alpha/2}$ comme au-dessus.

Puisque $\varepsilon = \beta\sigma/\sqrt{n}$, n vaut $E[\ (\beta\sigma/\varepsilon)^2\ ] + 1$.

Pour une proportion p et une précision relative δ dans le cadre d'une approximation normale, n vaut $E[\ p(1-p)(Z_{alpha/2}/\delta)^2\ ] + 1$.

Tests de conformité

Test de H_0 : $m = m_0$ pour un échantillon de moyenne m

Variance connue	écart réduit $\varepsilon_{obs} = \|m - m_0\|/\sqrt{V/n}$ à comparer avec $\varepsilon_{seuil} = Z_{alpha/2}$
Variance inconnue	écart réduit $t_{obs} = \|m - m_0\|/\sqrt{V_{ech}/n}$ à comparer avec $t_{seuil} = t_{alpha/2}$ pour $n-1$ ddl pour $n < 30$ l'échantillon doit suivre une loi normale $\mathcal{N}(m, \sigma)$.

Test de H_0 : $k/n = k_0/n_0$ pour un échantillon de n valeurs dont k valeurs sont marquées

Pour $p = k/n$ et $p_0 = k_0/n_0$, l'écart réduit est $\delta_{obs} = |p - p_0|/\sqrt{p_0 * (1 - p_0)/n}$

A comparer avec $\delta_{seuil} = Z_{alpha/2}$.

Tests d'homogénéité

Comparaison de moyennes H_0 : $m_1 = m_2$ et $N = n_1 + n_2 - 2$

On pose $Vp_d = \dfrac{(n_1 - 1)V_1 + (n_2 - 1)V_2}{(n_1 - 1) + (n_2 - 1)}$ et $V_d = \dfrac{\Sigma\ (\ (x_i - m_1) - (y_i - m_2)\ )^2}{n - 1}$.

Nature des échantillons	ε_{obs}	ε_{seuil}	dll
Indépendants			
grands échantillons	$\|m_1 - m_2\|/\sqrt{V_1/n_1 + V_2/n_2}$	$Z_{alpha/2}$	/
petits échantillons			
variances égales σ	$\|m_1 - m_2\|/\sqrt{V_p(1/n_1 + 1/n_2)}$	$t_{alpha/2}$	N
variances inégales σ_1, σ_2	$\|m_1 - m_2\|/\sqrt{V_1/n_1 + V_2/n_2}$	$t_{alpha/2}$	N
Appariés ($n_1 = n_2 = n$)	$\|m_1 - m_2\|/\sqrt{V_d/n}$	$t_{alpha/2}$	$n - 1$

Comparaison de deux variances V_1 et V_2

Comparer le rapport de variance R au F_β de Fisher-Snédécor pour $n_2 - 1$ et $n_1 - 1$ dll si R et β sont définis par

Hypothèse alternative	*Rapport de variance*	*Condition*	β
$V2 \neq V1$	$V1/V2$ $V2/V1$	$V1/V2$ $V2/V1$	$\alpha/2$ $\alpha/2$
$V1 > V2$	$V1/V2$	/	α
$V2 > V1$	$V2/V1$	/	α

Comparaison de deux fréquences

Si k_i individus sont marqués dans l'échantillon E_i de taille n_i et si $f_i = k_i/n_i$ alors, sachant la fréquence globale $f = (k_1 + k_2)/(n_1 + n_2)$ l'écart réduit est

$$\frac{|f_1 - f_2|}{\sqrt{f * (1 - f) * (1/n_1 + 1/n_2)}}$$

que l'on compare à $Z_{alpha/2}$.

χ^2 d'adéquation d'effectifs observés à une loi théorique

Si n effectifs théoriques th_i correspondent à n effectifs obs_i observés, la valeur du χ^2 d'adéquation est

$$\sum_{i=1}^{i=n} \frac{(obs_i - th_i)^2}{th_i}$$

que l'on compare au "χ^2 maximal autorisé" lu dans la table pour le nombre $\nu = n - 1$, $n - 2$ ou $n - 3$ degrés de liberté.

A condition : - qu'il y ait au moins 50 valeurs en tout,
- que chaque effectif soit supérieur à 5,
- que la somme des effectifs théoriques et observés soit la même.

Programmes et Sorties informatiques

Tri à plat de la variable SEXE avec Excel

Instructions

```
Ouvrir ELF.DBF
Faire
  Formats
  Mise en Forme automatique
  <Ok>
Faire
  Données
  Rapport de tableau croisé dynamique
  Données dans liste ou Base <suivant>
  Plage Base_de_données <suivant>
  Nouvelle feuille
  Disposition
  NUM  en "données"
  SEXE en "ligne"
  <Ok>
  <Terminer>
```

Résultats

```
Nombre de NUM
SEXE            Total
         0         35
         1         64
     Total         99
```

Tri à plat de la variable SEXE avec R

Instructions

```
source("statgh.r")
elfdata <- read.table("elf.dar",header=TRUE)
sexeElf <- elfdata[,2]
triAplat("Sexe de la personne",sexeElf,  c("homme","femme") )
```

Résultats

```
   R : Copyright 2004, The R Foundation for Statistical Computing
   Version 2.0.1  (2004-11-15), ISBN 3-900051-07-0

   R is free software and comes with ABSOLUTELY NO WARRANTY.
   You are welcome to redistribute it under certain conditions.
   Type 'license()' or 'licence()' for distribution details.

   R is a collaborative project with many contributors.
   Type 'contributors()' for more information and
   'citation()' on how to cite R or R packages in publications.

>   source("statgh.r")

    statgh.r, version 2.17

>   elfdata <- read.table("elf.dar",header=TRUE)
>   sexeElf <- elfdata[,2]
>   triAplat("Sexe de la personne",sexeElf,  c("homme","femme") )

    QUESTION :  Sexe de la personne

                      homme femme  Total
     Effectif            35    64     99
     Frequence (en %)    35    65    100
```

Analyse univariée de QT avec Statbox(Excel)

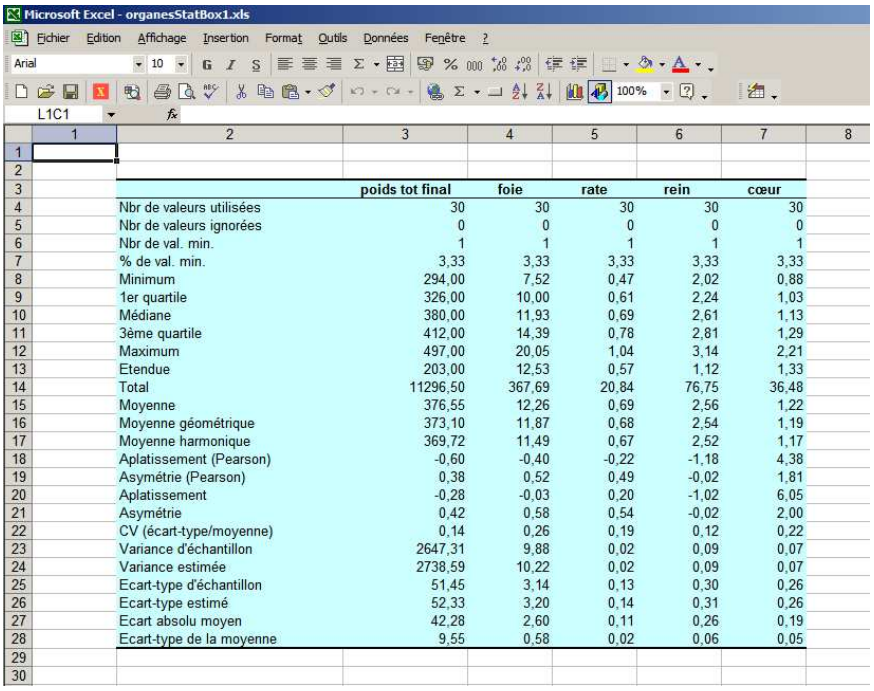
On part du fichier des données

restriction		poids	tot	final	foie	rate	rein	coeur
0	497	20.05	0.936		2.75	1.440		
0	392	15.171	0.575		2.622	1.115		
0	456	18.9	1.041		3.052	1.620		
0	425	16.832	0.8		2.808	1.272		
0	380	14.89	0.745		2.47	1.121		
0	361	13.255	0.608		2.263	1.039		
25	412	14.49	0.839		2.744	1.286		
25	414	13.87	0.724		2.986	1.390		
25	361	11.65	0.78		2.508	1.087		
25	359	12.706	0.649		2.695	1.130		
25	330	11.343	0.618		2.159	1.077		
25	350	13.051	0.706		2.529	1.033		
50	481	15.397	0.917		2.835	1.450		
50	395	13.62	0.565		2.579	1.206		
50	382	14.385	0.75		3.136	1.267		
...								
100	395	10.028	0.742		2.641	1.427		
100	363	8.7	0.673		2.599	1.265		
100	326	7.98	0.536		2.237	1.000		
100	323	8.13	0.468		2.144	1.116		
100	326	8.53	0.607		2.184	0.956		
100	294	7.52	0.499		2.212	0.879		

que l'on restructure en

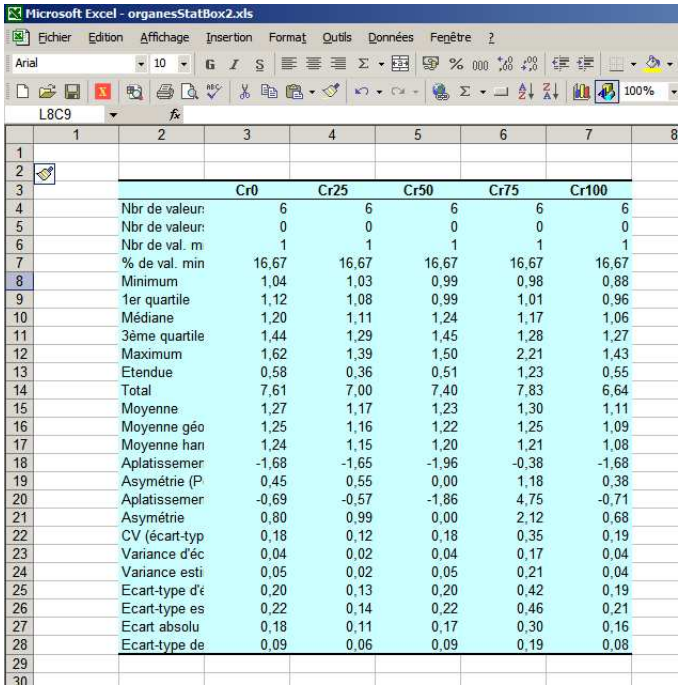
Num	Cr0	Cr25	Cr50	Cr75	Cr100	Pr0 Pr1...
1	1.440	1.286	1.450	1.280	1.427	...
2	1.115	1.390	1.206	2.210	1.265	
3	1.620	1.087	1.267	1.206	1.000	
4	1.272	1.130	1.497	1.139	1.116	
5	1.121	1.077	0.992	0.981	0.956	
6	1.039	1.033	0.991	1.009	0.879	

Résultats numériques globaux



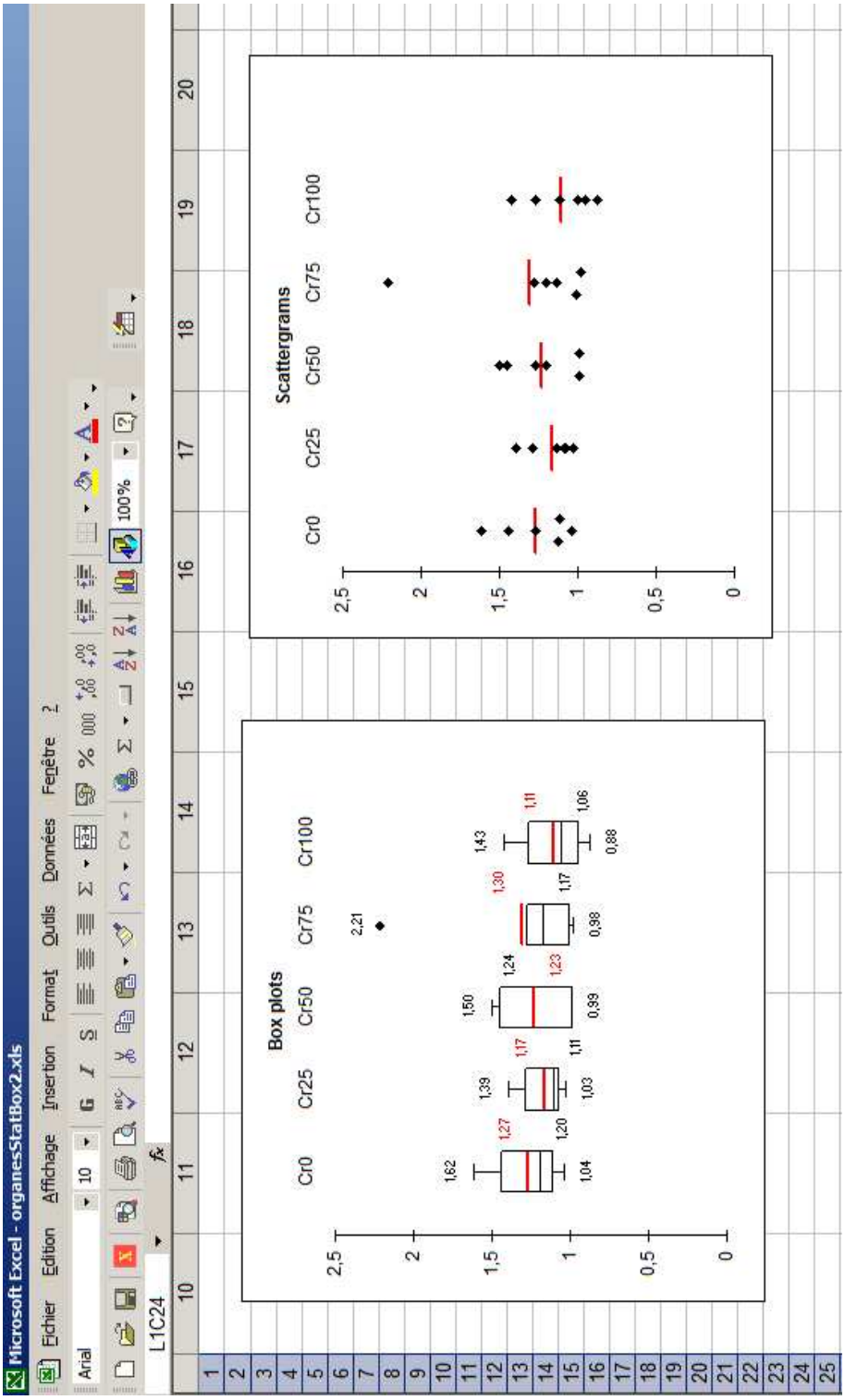
	poids tot final	foie	rate	rein	cœur
Nbr de valeurs utilisées	30	30	30	30	30
Nbr de valeurs ignorées	0	0	0	0	0
Nbr de val. min.	1	1	1	1	1
% de val. min.	3,33	3,33	3,33	3,33	3,33
Minimum	294,00	7,52	0,47	2,02	0,88
1er quartile	326,00	10,00	0,61	2,24	1,03
Médiane	380,00	11,93	0,69	2,61	1,13
3ème quartile	412,00	14,39	0,78	2,81	1,29
Maximum	497,00	20,05	1,04	3,14	2,21
Etendue	203,00	12,53	0,57	1,12	1,33
Total	11296,50	367,69	20,84	76,75	36,48
Moyenne	376,55	12,26	0,69	2,56	1,22
Moyenne géométrique	373,10	11,87	0,68	2,54	1,19
Moyenne harmonique	369,72	11,49	0,67	2,52	1,17
Aplatissement (Pearson)	-0,60	-0,40	-0,22	-1,18	4,38
Asymétrie (Pearson)	0,38	0,52	0,49	-0,02	1,81
Aplatissement	-0,28	-0,03	0,20	-1,02	6,05
Asymétrie	0,42	0,58	0,54	-0,02	2,00
CV (écart-type/moyenne)	0,14	0,26	0,19	0,12	0,22
Variance d'échantillon	2647,31	9,88	0,02	0,09	0,07
Variance estimée	2738,59	10,22	0,02	0,09	0,07
Ecart-type d'échantillon	51,45	3,14	0,13	0,30	0,26
Ecart-type estimé	52,33	3,20	0,14	0,31	0,26
Ecart absolu moyen	42,28	2,60	0,11	0,26	0,19
Ecart-type de la moyenne	9,55	0,58	0,02	0,06	0,05

Résultats numériques pour le coeur



	Cr0	Cr25	Cr50	Cr75	Cr100
Nbr de valeur	6	6	6	6	6
Nbr de valeur	0	0	0	0	0
Nbr de val. m	1	1	1	1	1
% de val. min	16,67	16,67	16,67	16,67	16,67
Minimum	1,04	1,03	0,99	0,98	0,88
1er quartile	1,12	1,08	0,99	1,01	0,96
Médiane	1,20	1,11	1,24	1,17	1,06
3ème quartile	1,44	1,29	1,45	1,28	1,27
Maximum	1,62	1,39	1,50	2,21	1,43
Etendue	0,58	0,36	0,51	1,23	0,55
Total	7,61	7,00	7,40	7,83	6,64
Moyenne	1,27	1,17	1,23	1,30	1,11
Moyenne géo	1,25	1,16	1,22	1,25	1,09
Moyenne har	1,24	1,15	1,20	1,21	1,08
Aplatissemer	-1,68	-1,65	-1,96	-0,38	-1,68
Asymétrie (P	0,45	0,55	0,00	1,18	0,38
Aplatissemer	-0,69	-0,57	-1,86	4,75	-0,71
Asymétrie	0,80	0,99	0,00	2,12	0,68
CV (écart-typ	0,18	0,12	0,18	0,35	0,19
Variance d'éc	0,04	0,02	0,04	0,17	0,04
Variance esti	0,05	0,02	0,05	0,21	0,04
Ecart-type d'é	0,20	0,13	0,20	0,42	0,19
Ecart-type es	0,22	0,14	0,22	0,46	0,21
Ecart absolu	0,18	0,11	0,17	0,30	0,16
Ecart-type de	0,09	0,06	0,09	0,19	0,08

Résultats graphiques pour le coeur



Analyse univariée de QT avec Sas

Instructions

```
filename forganes 'organes.dat' ;

data organes ;
     infile forganes ;
     input  restriction  poidstotfinal foie    rate    rein   coeur ;

proc print data=organes ;

/* analyse globale rapide */

proc means data=organes n mean stddev cv min max ;
     var poidstotfinal foie    rate    rein   coeur ;

/* analyse rapide par restriction pour le coeur */

proc means data=organes n mean stddev cv  ;
     var  coeur ;
     class restriction ;

/* analyse globale longue pour le coeur */

proc univariate data=organes all ;
     var  coeur ;

run ;
```

Extrait des Résultats

Nous ne donnons qu'un extrait des résultats car SAS fournit 8 pages de résultas détaillés...

The MEANS Procedure

Variable	N	Mean	Std Dev	Coeff of Variation
poidstotfinal	30	376.5500000	52.3315607	13.8976393
foie	30	12.2564667	3.1968759	26.0831769
rate	30	0.6948000	0.1352031	19.4592893
rein	30	2.5581667	0.3068745	11.9958746
coeur	30	1.2160333	0.2624360	21.5813163

The MEANS Procedure

Analysis Variable : coeur

restriction	N Obs	N	Mean	Std Dev	Coeff of Variation
0	6	6	1.2678333	0.2240370	17.6708539
25	6	6	1.1671667	0.1397962	11.9773949
50	6	6	1.2338333	0.2169382	17.5824528
75	6	6	1.3041667	0.4581691	35.1311806
100	6	6	1.1071667	0.2067263	18.6716521

The UNIVARIATE Procedure

Variable: coeur

Moments

N	30	Sum Weights	30
Mean	1.21603333	Sum Observations	36.481
Std Deviation	0.262436	Variance	0.06887265
Skewness	2.00091266	Kurtosis	6.05412684
Uncorrected SS	46.359419	Corrected SS	1.99730697
Coeff Variation	21.5813163	Std Error Mean	0.04791404

Basic Statistical Measures

Location		Variability	
Mean	1.216033	Std Deviation	0.26244
Median	1.134500	Variance	0.06887
Mode	1.206000	Range	1.33100
		Interquartile Range	0.25300

Basic Confidence Limits Assuming Normality

Parameter	Estimate	95% Confidence Limits	
Mean	1.21603	1.11804	1.31403
Std Deviation	0.26244	0.20901	0.35280
Variance	0.06887	0.04368	0.12447

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	25.37948	Pr > |t|	<.0001
Sign	M	15	Pr >= |M|	<.0001
Signed Rank	S	232.5	Pr >= |S|	<.0001

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.834838	Pr < W	0.0003
Kolmogorov-Smirnov	D	0.161553	Pr > D	0.0444
Cramer-von Mises	W-Sq	0.167274	Pr > W-Sq	0.0140
Anderson-Darling	A-Sq	1.134068	Pr > A-Sq	<0.0050

Trimmed Means

Percent Trimmed in Tail	Number Trimmed in Tail	Trimmed Mean	Std Error Trimmed Mean	95% Confidence Limits	
26.67	8	1.165714	0.041898	1.075200	1.256229

Winsorized Means

Percent Winsorized in Tail	Number Winsorized in Tail	Winsorized Mean	Std Error Winsorized Mean	95% Confidence Limits	
26.67	8	1.162400	0.042749	1.070047	1.254753

Robust Measures of Scale

Measure	Value	Estimate of Sigma
Interquartile Range	0.253000	0.187549
Gini's Mean Difference	0.267676	0.237222
MAD	0.136000	0.201634
Sn	0.202742	0.202742
Qn	0.244409	0.216931

Variable: coeur

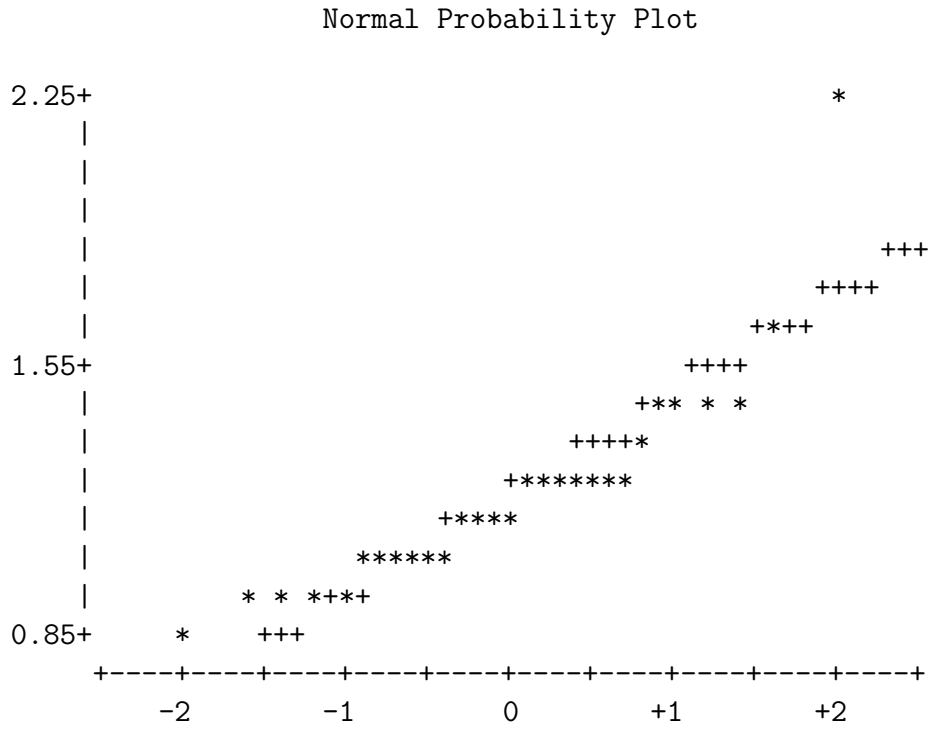
Stem	Leaf	#	Boxplot
22	1	1	*
21			
20			
19			
18			
17			
16	2	1	|
15	0	1	|
14	345	3	|
13	9	1	|
12	1167789	7	+--+--+
11	22234	5	*-----*
10	013489	6	+-----+
9	6899	4	|
8	8	1	|

----+----+----+----+

Multiply Stem.Leaf by 10**-1

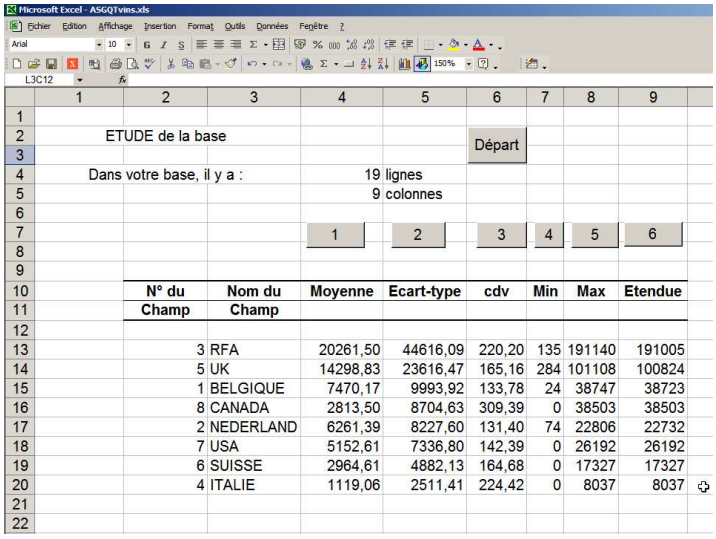
The UNIVARIATE Procedure

Variable: coeur



Analyse bivariée **QT** du dossier VINS

Résultats statistiques avec asgQT.xlt (gH)



Microsoft Excel - ASGQTvins.xls

ETUDE de la base

Départ

Dans votre base, il y a : 19 lignes

9 colonnes

1 2 3 4 5 6

N° du Champ	Nom du Champ	Moyenne	Ecart-type	cdv	Min	Max	Etendue
3	RFA	20261,50	44616,09	220,20	135	191140	191005
5	UK	14298,83	23616,47	165,16	284	101108	100824
1	BELGIQUE	7470,17	9993,92	133,78	24	38747	38723
8	CANADA	2813,50	8704,63	309,39	0	38503	38503
2	NEDERLAND	6261,39	8227,60	131,40	74	22806	22732
7	USA	5152,61	7336,80	142,39	0	26192	26192
6	SUISSE	2964,61	4882,13	164,68	0	17327	17327
4	ITALIE	1119,06	2511,41	224,42	0	8037	8037

Matrice des corrélations détaillée avec Statistica

Correlations (vins)
Marked correlations are significant at p < ,05000
(Casewise deletion of missing data)

Var. X & Var. Y	Mean	Std.Dv.	r(X,Y)	r²	t	p	N	Constant dep: Y	Slope dep: Y	Constant dep: X	Slope dep: X
ITALIE	1119,06	2584,22	1,000000	1,000000			18	0,00	1,000	0,00	1,000
VIN	9,50	5,34	1,000000	1,000000			18	0,00	1,000	0,00	1,000
USA	5152,61	7549,50	1,000000	1,000000			18	0,00	1,000	0,00	1,000
SUISSE	2964,61	5023,67	1,000000	1,000000			18	0,00	1,000	0,00	1,000
UK	14298,83	24301,14	1,000000	1,000000			18	0,00	1,000	0,00	1,000
CANADA	2813,50	8956,99	1,000000	1,000000			18	0,00	1,000	0,00	1,000
RFA	20261,50	45909,58	1,000000	1,000000			18	0,00	1,000	0,00	1,000
NEDERLAND	6261,39	8466,13	1,000000	1,000000			18	0,00	1,000	0,00	1,000
BELGIQUE	7470,17	10283,65	1,000000	1,000000			18	0,00	1,000	0,00	1,000
RFA	20261,50	45909,58	0,969259	0,939463	15,75752	0,000000	18	-5921,35	1,831	3903,59	0,513
UK	14298,83	24301,14	0,969259	0,939463	15,75752	0,000000	18	3903,59	0,513	-5921,35	1,831
RFA	20261,50	45909,58	0,947598	0,897942	11,86483	0,000000	18	6596,41	4,857	-932,39	0,185
CANADA	2813,50	8956,99	0,947598	0,897942	11,86483	0,000000	18	-932,39	0,185	6596,41	4,857
BELGIQUE	7470,17	10283,65	0,941583	0,886579	11,18335	0,000000	18	1772,72	0,398	-2322,59	2,225
UK	14298,83	24301,14	0,941583	0,886579	11,18335	0,000000	18	-2322,59	2,225	1772,72	0,398
CANADA	2813,50	8956,99	0,925628	0,856788	9,78378	0,000000	18	-2064,84	0,341	7233,25	2,511
UK	14298,83	24301,14	0,925628	0,856788	9,78378	0,000000	18	7233,25	2,511	-2064,84	0,341
USA	5152,61	7549,50	0,893528	0,798392	7,96002	0,000001	18	1183,44	0,278	-521,02	2,876
UK	14298,83	24301,14	0,893528	0,798392	7,96002	0,000001	18	-521,02	2,876	1183,44	0,278
BELGIQUE	7470,17	10283,65	0,870166	0,757190	7,06364	0,000003	18	852,04	1,057	909,95	0,716
NEDERLAND	6261,39	8466,13	0,870166	0,757190	7,06364	0,000003	18	909,95	0,716	852,04	1,057
USA	5152,61	7549,50	0,869918	0,756757	7,05534	0,000003	18	381,94	0,639	1364,48	1,185
BELGIQUE	7470,17	10283,65	0,869918	0,756757	7,05534	0,000003	18	1364,48	1,185	381,94	0,639
BELGIQUE	7470,17	10283,65	0,869172	0,755460	7,03057	0,000003	18	3525,40	0,195	-8724,73	3,880
RFA	20261,50	45909,58	0,869172	0,755460	7,03057	0,000003	18	-8724,73	3,880	3525,40	0,195
RFA	20261,50	45909,58	0,847658	0,718524	6,39087	0,000009	18	-6298,79	5,155	2328,33	0,139
USA	5152,61	7549,50	0,847658	0,718524	6,39087	0,000009	18	2328,33	0,139	-6298,79	5,155
CANADA	2813,50	8956,99	0,814274	0,663042	5,61103	0,000039	18	-2484,54	0,709	4839,88	0,935
BELGIQUE	7470,17	10283,65	0,814274	0,663042	5,61103	0,000039	18	4839,88	0,935	-2484,54	0,709
USA	5152,61	7549,50	0,746946	0,557928	4,49368	0,000368	18	3381,31	0,630	-1752,75	0,886
CANADA	2813,50	8956,99	0,746946	0,557928	4,49368	0,000368	18	-1752,75	0,886	3381,31	0,630
USA	5152,61	7549,50	0,717239	0,514432	4,11717	0,000807	18	2807,82	2,095	-145,98	0,246

Intervalles de confiance en Sas

Instructions

```
/* scherrer */

data e10p1page335 ;
     input n m s alpha ;
     t=probit(1-alpha/200) ;
     u=s/sqrt(n) ;
     eps=t*u ;
     Lm=m-eps;
     Um=m+eps ;
     datalines ;
50 158.86 6.09 1
50 158.86 6.09 5
50 158.86 6.09 10
;

proc print data=e10p1page335 ;

/* ================================ */

data e10p3p337 ;
     input n m s alpha ;
     df=n-1 ;
     t= tinv(1-alpha/200,df) ;
     u=s/sqrt(n-1) ;
     eps=t*u ;
     Lm=m-eps;
     Um=m+eps ;
     datalines ;
9 23.5 4.5 1
9 23.5 4.5 5
9 23.5 4.5 10
9 158.86 6.09 1
9 158.86 6.09 5
9 158.86 6.09 10
;

proc print data=e10p3p337 ;

/* ================================ */
```

```
data e10p9p351 ;
     input n i Ntot alpha ;
     p=i/n ;
     q=1-p ;
     t=probit(1-alpha/200) ;
     s=sqrt(p*q*(Ntot-n)/( (n-1)*Ntot)) ;
     eps=t*s +1/(2*n) ;
     Lm=p-eps;
     Um=p+eps ;
     datalines ;
146 41 1000 5
;

/* ================================ */

proc print data=e10p9p351 ;

data e10p13p362 ;
     input m v alpha ;
     prec = 0.01 ;
     z=probit(1-alpha/200) ;
     n= (1/(prec*prec))*z*z*v/(m*m) ;
     datalines ;
158.86 37.18  1
158.86 37.18  5
158.86 37.18 10
;

proc print data=e10p13p362 ;


run ;
```

Résultats

```
The SAS System
```

Obs	n	m	s	alpha	t	u	eps	Lm	Um
1	50	158.86	6.09	1	2.57583	0.86126	2.21845	156.642	161.078
2	50	158.86	6.09	5	1.95996	0.86126	1.68803	157.172	160.548
3	50	158.86	6.09	10	1.64485	0.86126	1.41664	157.443	160.277

Obs	n	m	s	alpha	df	t	u
1	9	23.50	4.50	1	8	3.35539	1.59099
2	9	23.50	4.50	5	8	2.30600	1.59099
3	9	23.50	4.50	10	8	1.85955	1.59099
4	9	158.86	6.09	1	8	3.35539	2.15314
5	9	158.86	6.09	5	8	2.30600	2.15314
6	9	158.86	6.09	10	8	1.85955	2.15314

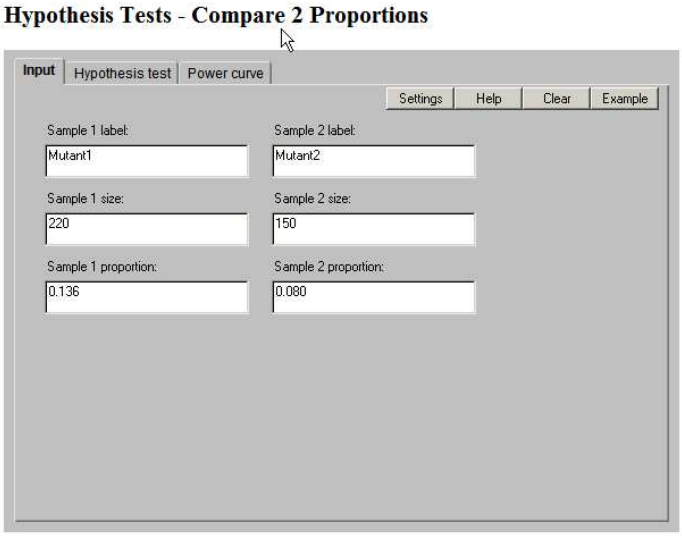
eps	Lm	Um
5.33839	18.162	28.838
3.66883	19.831	27.169
2.95852	20.541	26.459
7.22462	151.635	166.085
4.96515	153.895	163.825
4.00387	154.856	162.864

Obs	n	i	Ntot	alpha	p	q	t	s	eps	Lm	Um
1	146	41	1000	5	0.28082	0.71918	1.95996	0.034489	0.071022	0.20980	0.35184

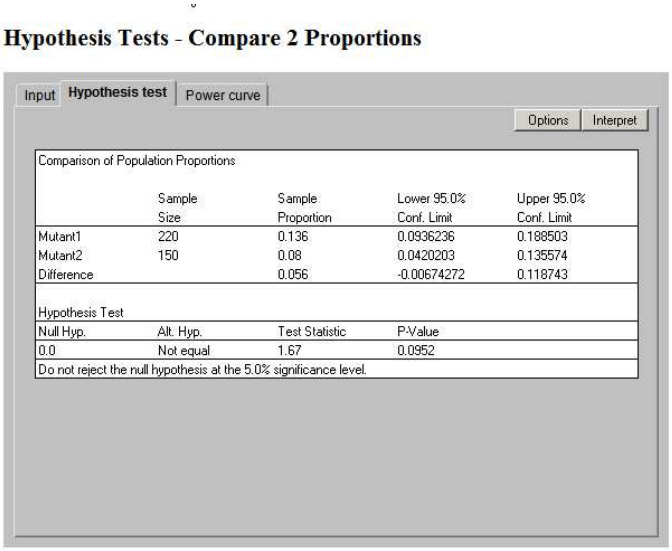
Obs	m	v	alpha	prec	z	n
1	158.86	37.18	1	0.01	2.57583	97.7495
2	158.86	37.18	5	0.01	1.95996	56.5948
3	158.86	37.18	10	0.01	1.64485	39.8598

Comparaison de fréquences via la page Web "Statlets"

Saisie des valeurs



Résultats



Comparaison de fréquences en ligne de commandes (gH)

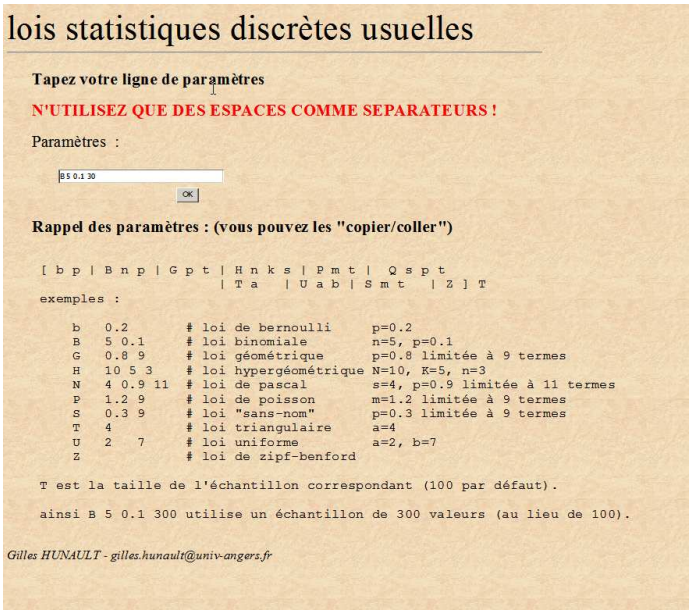
```
 compourc.rex (gH) : comparaison de pourcentages

ia               30  na              220  pa        0.136
ib               12  nb              150  pb        0.080
ii               42  nn              370  p         0.114

dp          0.05636  r2          0.00113  r   0.0335895151536240
eps         1.67801  soit en gros      1.68

 Au seuil de 5 % soit la valeur 1.96
 on peut  accepter  l'hypothèse que les pourcentages sont égaux.
```

Après multiplication des données par 10 :

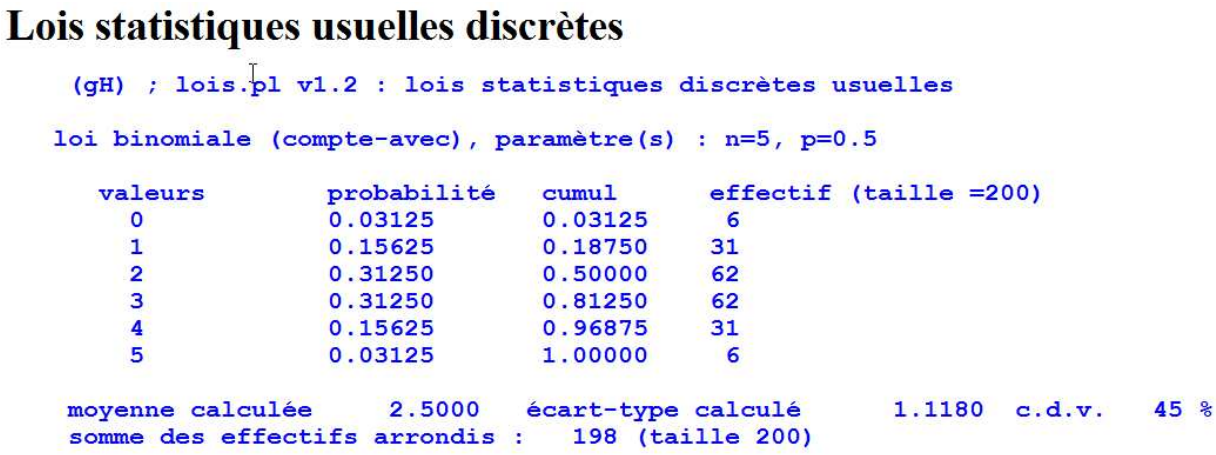
```
 compourc.rex (gH) : comparaison de pourcentages

ia              300  na             2200  pa        0.136
ib              120  nb             1500  pb        0.080
ii              420  nn             3700  p         0.114

dp          0.05636  r2          0.00011  r   0.0106219373386192
eps         5.30634  soit en gros      5.31

 Au seuil de 5 % soit la valeur 1.96
 on peut  refuser  l'hypothèse que les pourcentages sont égaux.
```

Effectifs d'une loi théorique par une page Web

Saisie des valeurs



Résultats



χ^2 d'indépendance avec une macro Excel

	A	B	C	D	E	F	G	H	I	J	K	L
1	**Données**	équipage	1ère_classe	seconde_classe	troisième_classe	**Théoriques**	équipage	1ère_classe	seconde_classe	seconde_classe	troisième_classe	
2	Décédés	673	122	167	528	Décédés	569,48	209,13	183,39	962,00	477,94	1490
3	Survivants	212	203	118	178	Survivants	315,52	115,87	101,61	533,00	228,06	711
4							885	325	285	1495	706	2201
5	Distance chi-deux		158,714									
6	Chi-deux table		5,991									
7												
8	Contributions											
9		+	65,519	Survivants	1ère_classe							
10		-	36,301	Décédés	1ère_classe							
11		-	33,965	Survivants	équipage							
12		+	18,819	Décédés	équipage							
13		+	5,244	Décédés	Théoriques							
14		+	2,644	Survivants	seconde_classe							
15		-	1,465	Décédés	seconde_classe							
16		-	3,486	Décédés	seconde_classe							
17												
18												
19												