STATISTIQUES

Johann PAVIS (selon source : B. PERROT)

Table des matières

Chapitre 1

Généralités

1.1 But général

Ce cours est un cours sur les statistiques inductives. Celles-ci ont pour but d'induire les propriétés d'un caractère d'une population, à partir de l'étude de ce caractère sur un échantillon de la population.

1.2 Terminologie et notations

Dans la terminologie usuelle en statistique, 'caractère' correspond à 'variable aléatoire' dans le cours de probabilités.

X étant un caractère, $X_1, X_2, ..., X_N$ désigne un échantillon de modèle X.

Une **réalisation** de cet échantillon (obtenue par expérience physique, enquête, mesure, *etc* ...) donne une suite de valeurs numériques $x_1, x_2, ...x_N$ appelée *série statistique* (si ces valeurs n'étaient pas numériques, par exemple couleurs, réponse oui ou non, on les numérise symboliquement).

Pour un couple (triplet, n-uplet), on dit *série statistique double (triple, n-uple)*.

On note	On l'appelle
$\overline{X} = \frac{1}{N}.\sum_n X_n$	estimateur de moyenne
$\overline{x} = \frac{1}{N}.\sum_n x_n$	moyenne de la série statistique
$C^*_{X,X} = \frac{1}{N}.\sum_n \left(X_n - \overline{X}\right)^2$	estimateur de variance
$c_{x,x} = \frac{1}{N}.\sum_n \left(x_n - \overline{x}\right)^2$	variance de la série statistique
$C^*_{X,Y} = \frac{1}{N}.\sum_n \left(X_n - \overline{X}\right).\left(Y_n - \overline{Y}\right)$	estimateur de covariance
$c_{X,Y} = \frac{1}{N}.\sum_n \left(x_n - \overline{x}\right).\left(y_n - \overline{y}\right)$	covariance de la série statistique
$R^*_{X,Y} = \frac{C^*_{X,Y}}{\sqrt{C^*_{X,X}.C^*_{Y,Y}}}$	estimateur du coefficient de corrélation
$r_{x,y} = \frac{c_{x,y}}{\sqrt{c_{x,x}.c_{y,y}}}$	coefficient de corrélation de la série statistique

Voici quelques formules utilisant la notation compacte avec le symbole (celles ci se démontrent simplement en utilisant la linéarité de) :

$$\begin{array}{lcccc}
C^*_{X,X} & = & \overline{X^2} - \left(\overline{X}\right)^2 & = & \overline{\left(X - \overline{X}\right)^2} \\
c_{x,x} & = & \overline{x^2} - \left(\overline{x}\right)^2 & = & \overline{\left(x - \overline{x}\right)^2} \\
C^*_{X,Y} & = & \overline{X.Y} - \overline{X}.\overline{Y} & = & \overline{\left(X - \overline{X}\right).\left(Y - \overline{Y}\right)} \\
c_{x,y} & = & \overline{x.y} - \overline{x}.\overline{y} & = & \overline{\left(x - \overline{x}\right).\left(y - \overline{y}\right)}
\end{array}$$

1.3 Représentation graphique

1.3.1 Diagramme en bâtons

Soient X un caractère et $X_1, ...X_N$ un échantillon de modèle X. Soit $x_1, ..., x_N$ la série statistique associée (par une réalisation de cet échantillon).

Dans le cas où X est un caractère suivant une loi discrète et où le nombre de valeurs distinctes dans la série statistique, soient $a_1, a_2, ..., a_q$, est petit, on représente graphiquement la série statistique par un diagramme en bâtons.

Pour $i \in [1, q]$, on appelle *fréquence de la valeur a_i* dans la série statistique **le nombre de termes de la série statistique qui sont égaux à** a_i

divisé par N. On note p_i cette fréquence ($p_i = Card\{x_k/k \in [1,q]$ et $x_k = a_i\}$).

Remarque : $\sum_{i=1}^{q} p_i = 1$

Remarque fondamentale : De la même manière, on définit la fréquence de la valeur a_i dans l'échantillon et on la note P_i. Ainsi, P_i est une *variable aléatoire* qui, au cours de la réalisation considérée de l'échantillon, donne la valeur numérique p_i.

On note $p_i^{'}$ la probabilité $p(X = a_i)$.

On considère la variable aléatoire "$1a_i$ caractérise l'événement '$X =$ 'a_i" où $1a_i$ est définie comme suit : $\begin{cases} 1a_i = 1 \text{ , si } X = a_i \\ 0 \text{ , sinon} \end{cases}$

Les variables aléatoire X_n , $n \in [1, N]$ donnent ainsi des variables aléatoires caractéristiques qui forment un échantillon de modèle $1a_i$ lorsque les X_n parcourent $X_1, X_2, ..., X_N$.

L'estimateur de moyenne de cet échantillon est : $\overline{1a_i} = P_i$. L'espérance de $1a_i$ est : $E(1a_i) = p_i^{'}$.

La loi des grands nombres (vue dans le cours de probabilité) nous dit donc :

Pour tout $i \in [1,q]$, les variables aléatoire P_i convergent en probabilité vers les valeurs numériques $p_i^{'}$ lorsque N tend vers l'infini.

C'est cette remarque qui fait le lien heuristique entre le diagramme en bâtons et l'allure de la loi de X.

1.3.2 Histogramme

Soient X un caractère et $X_1, X_2, ..., X_N$ un échantillon de modèle X. On note $x_1, x_2, ..., x_N$ la série statistique associée (par la réalisation de l'échantillon).

Dans le cas où X est un caractère suivant une loi diffuse et où le nombre de valeurs distinctes obtenues dans la série statistique, soit $a_1, a_2, ..., a_q$, est grand, on représente graphiquement la série statistique par un histogramme.

On considère une succession d'intervalles, soit A_i , où $i \in [1, q]$ telle que les valeurs de la série statistique soient contenues dans la réunion des A_i. Par la suite, on désignera ces intervalles A_i plutôt avec la terminologie "*classe* A_i", car il s'agit bien des classes de valeurs pour la série statistique.

Les classes A_i vont maintenant jouer le rôle que jouaient les a_i dans le paragraphe précédent (I.3.1).

Pour $i \in [1, q]$, on appelle *fréquence de la classe A_i* dans la série statistique **le nombre de termes de la série statistique appartenant à la classe A_i divisé par N**. On note p_i cette fréquence ($p_i = Card\{a_k \backslash k \in [1, q]$ et $a_k \in A_i\}$).

Remarque : $\sum_{i=1}^{q} p_i = 1$

Remarques : Les p_i représentent **l'aire** des rectangles de base A_i (pas leur hauteur). Pour obtenir un histogramme significatif, il faut respecter quelques règles heuristiques lors du partage en classes :

- il faut que chaque classe contienne suffisamment de termes (au moins 5, sauf éventuellement les deux classes des extrémités) ;
- il faut suffisamment de classes pour que le dessin des rectangles donne l'allure de la densité de la loi de X

Ces conditions imposent que N (taille de l'échantillon, soit taille de la série statistique) soit suffisamment grand.

Remarque fondamentale : De la même manière, on définit la fréquence de la classe A_i dans l'échantillon et on la note P_i. Ainsi, P_i est une variable aléatoire qui, au cours de la réalisation considérée de l'échantillon, donne la valeur numérique p_i.

On note $p_i^{'}$ la probabilité : $p(X \in A_i)$.

On considère la variable aléatoire $1A_i$ caractéristique de l'événement "$X \in A_i$" où $1A_i$ est définie par : $\begin{cases} 1A_i = 1 \text{ , si } X \in A_i \\ 0 \text{ , sinon} \end{cases}$

Les variables aléatoires X_n où $n \in [1, N]$ donnent ainsi des variables aléatoires caractéristiques qui forment un échantillon de modèle $1A_i$ lorsque les X_n parcourent $X_1, X_2, ..., X_N$.

L'estimateur de moyenne de cet échantillon est : $\overline{1A_i} = P_i$.

L'espérance de $1A_i$ vaut : $E(1A_i) = p_i^{'}$.

La loi des grands nombres (vue dans le cours de probabilité) nous dit donc :

Pour tout $i \in [1, q]$, les variables aléatoire P_i convergent en probabilités vers les valeurs numériques $p_i^{'}$ lorsque N tend vers l'infini.

C'est cette remarque qui fait le lien heuristique entre l'histogramme et l'allure de la densité de la loi de X.

1.4 Régression

1.4.1 Généralités

Étant donnée une série statistique double, soit $\begin{cases} x_1, x_2, \ldots, x_N \\ y_1, y_2, \ldots, y_N \end{cases}$, on considère le nuage de points (dans le plan) formé par les N points M_n ce coordonnées (x_n, y_n). On cherche une courbe modèle sur laquelle ces N points du nuage sont "à peu près bien situés".

Cette courbe modèle est appelée *courbe de régression*. Elle sert, en quelque sorte, de '*courbe idéalisée*' d'interpolation entre les points M_n.

Ceci se généralise au cas de séries triples (on cherche alors une surface modèle sur laquelle les triplets M sont "à peu près bien placés") et plus généralement au cas des séries statistiques n-uples (on dit qu'il s'agit d'une *régression multivariée*).

La série statistique est obtenue comme **réalisation d'un échantillon d'un couple (ou d'un n-uplet) de caractères**. La régression revient donc à chercher une relation modèle que vérifie "à peu près" les caractères.

1.4.2 Droite de régression

Le cas le plus courant est celui où la courbe de régression recherchée est une droite. C'est à dire que l'on cherche une ***relation affine*** $Y = a.X + b$ que vérifient "à peu près" les caractères X et Y.

Remarque : Souvent, on est amené à chercher la relation affine entre X et $\log(Y)$ ou entre $\log(X)$ et Y plutôt qu'entre X et Y.

Chercher une droite qui passe au mieux par N points du plan est un problème classique de *moindre carré* (vu dans l'UV Algorithmique numérique).

Cela consiste *à chercher les paramètres a et b qui **minimisent** l'expression* :

$$f(a,b) = \sum_{n=1}^{N} (a.x_n + b - y_n)^2$$

Cette expression est le carré de la norme euclidienne du *vecteur erreur d'interpolation*.

Les équations normales (écrites en utilisant les notations statistiques compactes vues au 1.2) sont : $\left| \begin{array}{l} a.\overline{x^2} + b.x = \overline{x.y} \\ a.\overline{x} + b.\overline{y} \end{array} \right. \Leftrightarrow \left| \begin{array}{l} \dfrac{\partial f}{\partial a}(a,b) = 0 \\ \dfrac{\partial f}{\partial b}(a,b) = 0 \end{array} \right.$

Le déterminant de ce système est : $\Delta = \overline{x^2} - (\overline{x})^2 = C_{x,x} > 0$ (sauf dans le cas trivial où tous les x_n sont égaux).

La solution de ce système est : $\left\{ \begin{array}{lcl} a & = & \dfrac{C_{x,y}}{C_{x,x}} \\ b & = & \overline{y} - a.\overline{x} \end{array} \right.$

Remarque : La seconde équation ($a.\overline{x} + b = \overline{y}$) montre que le centre de gravité du nuage de points se trouve sur la droite de régression.

1.4.3 Formule donnant l'interprétation du coefficient de corrélation de la série statistique

On s'intéresse à la norme du vecteur erreur d'interpolation.

Plus précisément, nous allons calculer le carré de cette norme et le diviser par N pour rendre cette estimation de l'importance de l'erreur d'interpolation insensible au nombre N de valeurs de la série statistique.

$$S = \frac{1}{N}.\sum_{n=1}^{N} (a.x_n + b - y_n)^2 = \frac{1}{N}.\sum_{n=1}^{N} [a.x_n + (\overline{y} - a.\overline{x}) - y_n]^2$$

$$S = \frac{1}{N}.\sum_{n=1}^{N} [a.(x_n - \overline{x}) - (y_n - \overline{y})]^2 = a^2.C_{x,x} + C_{y,y} - 2.a.C_{x,y}$$

$$S = \frac{C_{x,y}^2}{C_{x,x}} + C_{y,y} - 2.\frac{C_{x,y}}{C_{x,x}} = C_{y,y}.\left(1 - \frac{C_{x,y}^2}{C_{x,x}.C_{y,y}}\right)$$

$$\frac{1}{N}.\sum_{n=1}^{N}(a.x_n + b - y_n)^2 = C_{y,y}.(1 - r_{x,y}^2)$$

Cette formule montre que $|r_{x,y}| \leq 1$ et donne l'interprétation de $r_{x,y}$.

Chapitre 2

Estimation

2.1 Estimation ponctuelle

2.1.1 But de l'estimation

Il s'agit d'estimer une grandeur a avec des observations $x_1, x_2, \ldots, x_N$. Pour l'étude de ce problème, on considère que les observations sont la réalisation de variables aléatoires $X_1, X_2, \ldots, X_N$ (qui est le plus souvent un échantillon de modèle X) avec a un paramètre de la loi du caractère X.

Une variable aléatoire diffuse, soit A (fonction des variables aléatoires $X_1, X_2, \ldots, X_N$) sera appelée **estimateur de a** si l'on a de bonnes raisons de penser que, lors de la réalisation des variables aléatoires $X_1, X_2, \ldots, X_N$ (qui donne les observations $x_1, x_2, \ldots, x_N$), la valeur donnée par A est proche de a.

Les raisons évoquées ci-dessus sont appelées **qualités de l'estimateur**.

2.1.2 Qualités d'un estimateur

Sans biais : c'est un estimateur A dont l'espérance $E(A)$ est égale à a.

Exemple : $\overline{X}$ est un estimateur sans biais de $E(X)$

Convergent : c'est un estimateur de A converge en probabilité vers a lorsque N tend vers l'infini.

Exemple : $\overline{X}$ est un estimateur de $E(X)$ convergent (c'est exactement ce que dit la loi des grands nombres).

Propriété 1 *Un estimateur sans biais et dont la variance tend vers 0 pour N tendant vers l'infini est convergent.*

Soit ε fixé. L'inégalité de Chebitchev appliquée à A donne $p(|A-a| > \frac{1}{b^2}$. On choisis $b = \dfrac{\varepsilon}{\sigma}$. On obtient alors : $(|A-a| > \varepsilon) \leq \left(\dfrac{\sigma}{\varepsilon}\right)^2$

Correct : Un estimateur est dit *correct* s'il est sans biais et convergent.

Exemple : $\overline{X}$ est un estimateur correct de $E(X)$.

De variance minimum : c'est un estimateur A tel que : $\forall B$ variable aléatoire, $E\left((A-a)^2\right) \leq E\left((B-a)^2\right)$

Propriété 2 *Un estimateur de variance minimum est sans biais.*

Preuve : Soit $B = A - E(A-a)$. On a donc : $(B-a)^2 = [(A-a) - E(A-a)]^2$.

D'où : $E\left((B-a)^2\right) = E\left((A-a)^2\right) - E\left((A-a)^2\right)$

Donc l'inégalité de l'hypothèse entraîne que $E(A-a) = 0$ soit : $E(A) = a$.

Exhaustif : c'est un estimateur A telle que *la loi de probabilité conditionnelle du N-uplet $X_1, X_2, \ldots, X_N$ sachant A* ne dépend pas de a.

Robuste : Lorsque l'on a un très grand nombre d'observations pour faire l'estimation on peut craindre la présence de valeurs aberrantes. On dit qu'un estimateur est **robuste** s'il élimine les valeurs aberrantes avant de procéder à l'estimation.

2.1.3 Méthode du maximum de vraisemblance

C'est la méthode la plus courante pour obtenir les estimateurs.

On définit d'abord la fonction de vraisemblance d'observations par :

$$\begin{cases} vraisemblance(x_1, \ldots, x_N) = p(X_1 = x_1, \ldots, X_N = x_n) \\ \text{si les } X_n \text{ suivent une loi discrète} \\ vraisemblance(x_1, \ldots, x_N) = \textit{densité du N-uplet}(X_1, \ldots, X_N) \\ \text{si les } X_n \text{ suivent une loi diffuse avec densité} \end{cases}$$

Cette fonction de vraisemblance dépend aussi du paramètre a. On cherche alors la valeur de $a = \varphi(x_1, \ldots, x_n)$ qui rend maximum cette vraisemblance. De ceci on déduit naturellement l'estimateur $A = \varphi(X_1, \ldots, X_N)$.

Exemple : Cherchons l'estimateur au maximum de vraisemblance du paramètre a d'une loi de Poisson (obtenu avec $X_1, \ldots, X_N$ un échantillon de modèle X qui suit la loi de Poisson de paramètre a). Notons v cette vraisemblance.

$$v(x_1, \ldots, x_N) = p(X_1 = x_1, \ldots, X_N = x_N) = \prod_{n=1}^{N} p(X_n = x_n)$$

$$v(x_1, \ldots, x_N) = \prod_{n=1}^{N} \frac{a^{x_n}}{x_n!}.\mathbf{e}^{-a}$$

$$v(x_1, \ldots, x_N) = \mathbf{e}^{-N.a}.a^{\sum_{n=1}^{N} X_n}.\prod_{n=1}^{N} \frac{1}{x_n!}$$

Pour déterminer le maximum par rapport à a, on va annuler la dérivée partielle de v par rapport à a : $\dfrac{\partial v}{\partial a} = \text{fonction ne s'annulant pas} . \displaystyle\sum_{n=1}^{N} x_n - N.a$.

D'où : $a = \overline{x}$, ce qui donne l'estimateur de $\overline{X}$.

Remarquer que ce résultat est attendu car pour une loi de Poisson de paramètre a on sait que $a = E(X)$.

2.1.4 Estimateurs classique de variance

Les observations $x_1, \ldots, x_N$ sont une réalisation de $X_1, \ldots, X_N$ (échantillon d'un caractère X) avec lesquelles on veut estimer la variance de X.

Cas où on connaît $m = E(X)$:

On a : $E(V^*) = Var(X)$, où : $V^* = \frac{1}{N}.\sum_{n=1}^{N}(X_n - m)^2$

V^* est donc un estimateur sans biais de $Var(X)$. On note σ^* la racine carrée de V^* ($\sigma^* = \sqrt{V^*}$).

Remarque : Si X suit une loi Normale, alors on peut chercher l'estimateur de $Var(X)$ avec le principe du maximum de vraisemblance. On trouve V^* et peut aussi vérifier (en utilisant la propriété vue en II.2.2) qu'il est convergent.

Cas où on ne connaît pas m :

Dans ce cas, on ne peut utiliser m dans la formule donnant V^*. On le remplace alors par son estimateur $\overline{X}$. On obtient alors la formule : *estimateur de variance* $=$ $C^*_{X,X}$ (déjà appelé estimateur de variance au I.1.2).

Cependant (d'après un calcul effectué en probabilité) :

$$E\left(C^*_{X,X}\right) = \frac{1}{N}.\sum_{n=1}^{N} E\left((X_n - \overline{X})^2\right) = \frac{N-1}{N}.Var(X)$$

Ainsi $C^*_{X,X}$ est un biais, qui est négligeable pour les grandes valeurs de N mais que l'on préfère corriger pour les petites valeurs de N en prenant pour estimateur de variance : $V^*_{N-1} = \frac{N}{N-1}.C^*_{X,X} = \frac{1}{N-1}.\sum_{n=1}^{N}(X_n - \overline{X})^2$

(l'indice $N-1$ dans cette notation a pour but de différencier V^* et V^*_{N-1} et sert de moyen mnémotechnique pour se rappeler que l'on divise par $N-1$ au lieu de diviser par N).

On note $\sigma^*_{N-1} = \sqrt{V^*_{N-1}}$.

Si X suit une loi Normale, on peut aussi vérifier (en utilisant la première propriété du II.1.2) que $C^*_{X,X}$ est un estimateur convergent.

2.1.5 Inégalité de FRECHET, CRAMER et RAO

Soient $X_1, X_2, \ldots, X_N$ un échantillon de modèle X dont la loi est diffuse de densité $f(x)$ et A un estimateur sans biais de a, un paramètre intervenant dans la loi de X.

On a alors : $Var(X) \leq \dfrac{1}{N.E\left[\left(\dfrac{\partial}{\partial a}\log\left(f(X)\right)\right)^2\right]}$

Cette valeur minorante est appelée **borne de FRECHET, CRAMER, RAO**

Preuve : Notons $v(x_1, x_2, \ldots, x_N)$ la densité du n-uple $X_1, X_2, \ldots, X_N$. On a donc : $v(x_1, x_2, \ldots, x_N) = \prod_{n=1}^{N} f(x_n)$

A est sans biais : $a = E(A)$.

On a donc : $E(A) = \int A(x_1, x_2, \ldots, x_N).v(x_1, x_2, \ldots, x_N).dx_1dx_2\ldots dx_N$

Donc : $\int A(x_1, x_2, \ldots, x_N).\dfrac{\partial}{\partial a}v(x_1, x_2, \ldots, x_N).dx_1dx_2\ldots dx_N = 1$ (*)

Considérons la variable aléatoire $B = \dfrac{\partial}{\partial a}\log[v(X_1, X_2, \ldots, X_N)) = \dfrac{1}{v}.\dfrac{\partial}{a}v.$ Alors on a : $\dfrac{\partial}{\partial a}v = B.v.$

B est centrée. En effet :

$E(B) = \int B(x_1, x_2, \ldots, x_N).v(x_1, x_2, \ldots, x_N).dx_1dx_2\ldots dx_N$

$E(B) = \int \dfrac{\partial}{\partial a}v(x_1, x_2, \ldots, x_N).dx_1dx_2\ldots dx_N = 0$

En effet : $\int v(x_1, x_2, \ldots, x_N).dx_1dx_2\ldots dx_N = 1$ D'où la valeur de $E(B)$)

$E(A.B) = \int A(x_1, x_2, \ldots, x_N).B(x_1, x_2, \ldots, x_N).v(x_1, x_2, \ldots, x_N).dx_1dx_2\ldots dx_N$

Or, on a vu : $\dfrac{\partial}{\partial a}v = B.v.$

Donc : $E(A.B) = \int A(x_1, x_2, \ldots, x_N).\dfrac{\partial}{\partial a}v(x_1, x_2, \ldots, x_N).dx_1dx_2\ldots dx_N = 1$ (cf *)

On a également : $(\rho_{A,B})^2 = \dfrac{(E(A.B) - E(A).E(B))^2}{Var(A).Var(B)} \leq 1$

D'où : $\dfrac{1-0}{Var(A).Var(B)} \leq 1$. Ainsi : $\dfrac{1}{Var(B)} \leq Var(A)$.

Il ne reste donc plus qu'à estimer $Var(B)$.

$$B = \frac{\partial}{\partial a} \log\left[v(X_1, X_2, \ldots, X_N)\right] = \frac{\partial}{\partial a} \log\left(\prod_{n=1}^{N} f(x_n)\right) = \frac{\partial}{\partial a} \sum_{n=1}^{N} \log[f(x_n)]$$

$$Var(B) = \sum_{n=1}^{N} Var\left(\frac{\partial}{\partial a}\log[f(x_n)]\right) = \sum_{n=1}^{N} E\left(\left[\frac{\partial}{\partial a}\log[f(x_n)]\right]^2\right)$$

En effet, $\dfrac{\partial}{\partial a}\log[f(X)]$ est centrée :

$$N.E\left(\frac{\partial}{\partial a}\log[f(X)]\right) = \sum_{n=1}^{N} E\left(\frac{\partial}{\partial a}\log[f(X_n)]\right)$$

$$N.E\left(\frac{\partial}{\partial a}\log[f(X)]\right) = E\left(\sum_{n=1}^{N}\frac{\partial}{\partial a}\log[f(X_n)]\right) = E(B) = 0$$

Donc : $Var(B) = N.e\left(\left[\dfrac{\partial}{\partial a}\log[f(X)]\right]^2\right)$.

2.1.6 Coefficient d'efficacité d'un estimateur sans biais

Par définition $e(A)$ **est le coefficient d'efficacité d'un estimateur** A ***sans biais*** : $e(A) = \dfrac{\text{borne de FRECHET, CRAMER, RAO}}{Var(A)}$

Ainsi : $0 < e(A) < 1$. Un estimateur sans biais est dit **efficace** si son coefficient d'efficacité est égal à 1 (*i.e.* $e(A) = 1$).

Remarque : Très souvent, les estimateurs construits avec le principe du maximum de vraisemblance sont efficaces.

Exemple : Soit X une variable aléatoire suivant une loi $N(m,\sigma)$. On a alors : $Var\left(\overline{X}\right) = \dfrac{1}{N}.\sigma^2$.

La loi de densité de X est f est telle que : $f(X) = \dfrac{1}{\sqrt{2.\pi}.\sigma}\mathbf{e}^{\left[-\frac{1}{2}.\left(\frac{X-m}{\sigma}\right)^2\right]}$

D'où : $\log[f(X)] = -\frac{1}{2}.\left(\frac{X-m}{\sigma}\right).\log\left(\frac{1}{\sqrt{2.\pi.\sigma}}\right)$

Ainsi, en dérivant et en élevant au carré : $\left[\frac{\partial}{\partial a}(\log[f(X)])\right]^2 = \left[0+\frac{X-m}{\sigma^2}\right]^2 = \frac{(X-m)^2}{\sigma^4}$

Ainsi : $E\left[\frac{(X-m)^2}{\sigma^4}\right] = \frac{1}{\sigma^4}.E\left[(X-m)^2\right] = \frac{1}{\sigma^4}$

La borne de FRECHET, CRAMER, RAO est donc : $\frac{1}{N}.\sigma^2$. Cet estimateur est donc efficace.

2.2 Lois du Chi-deux et de Student

2.2.1 Calculs auxiliaires

Soit $X_1, X_2, \ldots, X_N$ un échantillon de modèle X qui suit une loi normale $N(m,\sigma)$. On considère la somme : $\sum_{n=1}^{N}\left(X_n - \overline{X}\right)^2$

On a alors : *$\exists Z_1, Z_2, \ldots, Z_{N-1}$ un échantillon de la loi $N(0,\sigma)$ tel que :* $\sum_{n=1}^{N}\left(X - \overline{X}\right)^2 = \sum_{n=1}^{N-1} Z_n^2$

De plus*, en posant : $Z_0 = \sqrt{N}.\left(\overline{X} - m\right)$, on obtient que $Z_0, Z_1, \ldots, Z_{N-1}$ est un échantillon de taille N de la loi normale $N(0,\sigma)$, les Z_i étant* ***indépendants****.*

Le but de ces calculs est de montrer que l'on peut remplacer *une somme de N carrés* ***non stochastiquement indépendants*** *(car : $\sum\left(X - \overline{X}\right) = 0$) par une somme de $N-1$ carrés* ***stochastiquement indépendants****.*

Preuve : Pour $n \in [1, N-1]$, considérons les vecteurs lignes 1_n de dimension N définis par :

$$\left\{\begin{array}{lcll} 1_n(i) & = & 1 & \text{si } i \leq n \\ 1_n(n+1) & = & -n & \\ 1_n(i) & = & 0 & \text{si } i > n+1 \end{array}\right.$$

Le vecteur 1_N est le vecteur ligne à N composantes, toutes égales à 1.

Par construction, ces vecteurs sont orthogonaux. De plus, on a :

$\forall n \in [1, N-1]$, $||1_n|| = \sqrt{n+n^2}$ et $||1_N|| = \sqrt{N}$

On considère alors le système orthogonal obtenu en divisant ces vecteurs par leur norme. On forme ainsi une matrice Q dont les N lignes sont ces vecteurs orthonormés. Par construction, Q est donc orthogonale.

On définit le vecteur Z de composantes $Z_1, Z_2, \ldots, Z_N$ par la formule : $Z = Q.X$, avec X le vecteur de composantes $X_1, X_2, \ldots, X_N$.

Remarquer que l'on a : $Z_N = \dfrac{1}{\sqrt{N}}.\displaystyle\sum_{n=1}^{N} X_n = \sqrt{N}.\overline{X} = Z_0 + m.\sqrt{N}$.

Ainsi, comme Q est orthogonale, on a : $||Z||^2 = ||X||^2$.

Ce qui s'écrit aussi : $\displaystyle\sum_{n=1}^{N} X_n^2 = \sum_{n=1}^{N} Z_n^2 = \sum_{n=1}^{N-1} Z_n^2 + Z_N = \sum_{n=1}^{N-1} Z_n^2 + N.\left(\overline{X}\right)^2$

Donc : $\displaystyle\sum_{n=1}^{N-1} Z_n^2 = \sum_{n=1}^{N} X_n^2 - N \cdot \left(\overline{X}\right)^2 = N\left(\overline{X^2} - \overline{X}^2\right) = \sum_{n=1}^{N} \left(X_n - \overline{X}\right)^2$

(en effet : $\overline{X^2} - \overline{X}^2 = \overline{\left(X - \overline{X}\right)^2}$)

L'égalité souhaitée est donc démontrée.

Les formules : $\forall n \in [1, N-1]$, $Z_n = \dfrac{1}{\sqrt{n+n^2}}.\left(\displaystyle\sum_{j=1}^{N} X_j - n.X_{n+1}\right)$ montrent que les Z_n suivent la loi normale $N(0,\sigma)$.

Démontrons à présent l'indépendance stochastique des Z_n ($n \in [1, N]$).

$Z_1, Z_2, \ldots, Z_N$ est l'image de $X_1, X_2, \ldots, X_N$ (échantillon de modèle X qui suit la loi $N(0,\sigma)$) par l'application linéaire dont la matrice est Q. Donc, d'après le III.1.9 du cours de probabilité, la densité du n-uple $Z_1, Z_2, \ldots, Z_N$ est : $f\left(Q^{-1}.|detJacobienne\left(Q^{-1}\right)|\right)$, où f désigne la densité du n-uple $X_1, X_2, \ldots, X_N$.

Comme $|det\ Jacobienne\,(Q^{-1})| = 1$, on obtient comme densité f de la forme : `coeff`.**exp** $\left[-\dfrac{1}{2}.\displaystyle\sum_{n=1}^{N}\left(\frac{x-n-m}{\sigma}\right)^2\right]$ où x est le vecteur de composantes $x_1, x_2, \ldots, x_N$ est l'image du vecteur z (de composantes $z_1, z_2, \ldots, z_N$) par Q^{-1} ($x = Q^{-1}(z)$).

En développant, on obtient que cette densité est de la forme :

$$\texttt{coeff}.\mathbf{exp}\left[-1\ \frac{1}{2.\sigma^2}.\left(N.m^2+\sum_{n=1}^{N}x_n^2-2.m.\sum_{n=1}^{N}x_n\right)\right]$$

$$\texttt{nv-coeff}.\mathbf{exp}\left[-\frac{1}{2.\sigma^2}.\left(\sum_{n=1}^{N}z_n^2-2.m.\sqrt{N}.z_N\right)\right]$$

$$\texttt{nv-coeff}.\mathbf{exp}\left[-\frac{1}{2}.\left(\sum_{n=1}^{N-1}\left(\frac{z_n}{\sigma}\right)^2+\left(\frac{z_N}{\sigma}\right)^2-2.\frac{m\sqrt{N}}{\sigma^2}.z_N\right)\right]$$

Soit donc une densité de la forme :

$$\texttt{nv-coeff}.\prod_{n=1}^{N-1}\mathbf{exp}\left[-\frac{z_n^2}{2.\sigma^2}.\mathbf{exp}\left(-\frac{z_N^2-2.m\sqrt{N}.z_N}{(2.\sigma)^2}\right)\right]$$

Ceci montre l'indépendance des Z_n ($n\in[1,N-1]$). Or $Z_0=Z_N-m.\sqrt{N}$. Donc, on peut remplacer Z_N dans l'expression précédente par Z_0, ce qui prouve le résultat annoncé.

2.2.2 Loi du Chi-deux

$X_1,\ldots,X_N$ étant un un échantillon de modèle X qui suit la loi N(0,1), on considère la variable aléatoire : $U_n=\displaystyle\sum_{k=1}^{n}K_k^2$

On dit que <u>U_n suit la loi du Chi-deux à n degrés de libertés.</u>

Dans la pratique, on aura à calculer des probabilités telles que $p(U_n>u)$. Ces valeurs sont données dans des tables (*cf* Annexe1).

<u>Principe de calcul des valeurs</u> 3 *Cela revient à calculer la fonction de répartition suivante :* $F(u)=p(U_n\leq u)=\dfrac{1}{(2.\pi)^2}.\displaystyle\int_A \boldsymbol{e}^{-\frac{||X||^2}{2}}dx_1\ldots dx_n$ *où :* $A=\{X=x_1,\ldots,x_n/||X||^2\leq u\}$.

On ne peut en général effectuer un calcul numérique direct. En effet, il s'agit d'une intégrale en dimension n, dont il faut faire un calcul analytique (pour se ramener à une intégrale en dimension 1).

On effectue à cet effet un passage en coordonnées sphériques (en dimension n) :

$$\begin{aligned}
x_1 &= r.\prod_{i=1}^{n-1}\cos(\theta_i) \\
x_2 &= r.\prod_{i=1}^{n-2}\cos(\theta_i).\sin(\theta_{n-1}) \\
x_3 &= r.\prod_{i=1}^{n-3}\cos(\theta_i).\sin(\theta_{n-2}) \\
&\vdots \\
x_{n-1} &= r.\cos(\theta_1).\sin(\theta_2) \\
x_n &= \sin(\theta_1)
\end{aligned}$$

où r est un réel positif et $\forall i \in [1, n-2]$, $\theta_i \in [-\frac{\pi}{2}, \frac{\pi}{2}[$ et $\theta_{n-1} \in [0, 2.\pi[$.

On a alors : $|det\ Jacobienne| = r^{n-1}.\prod_{i=1}^{n-2}\cos^{n-1-i}(\theta_i)$

Le coefficient $\frac{1}{(2.\pi)^n}$ avec les intégrations angulaires donnent C_n.

On trouve : $C_n = \begin{cases} [2^m.(m-1)!]^{-1} & \text{, si } n = 2.m \\ 2.\left[\sqrt{2.\pi}.3.5.7\ldots(n-2)\right]^{-1} & \text{, si } n = 2.m+1 \end{cases}$

Il reste alors à calculer numériquement l'intégrale décrite précédemment : $\int_{[0,\sqrt{u}]} r^{n-1}.\mathbf{e}^{-\dfrac{r^2}{2}}\mathrm{d}r$

On constate que les tables de la loi du Chi-deux vont jusqu'à $n = 30$ degrés de liberté. Comme $E(U_n) = n$ et $Var(U_n) = 2.n$ (calculs aisés), il découle directement du théorème central limite (vu en probabilité) que $\dfrac{(U_n - n)}{\sqrt{2.n}}$ **converge en loi vers la loi normale réduite** $N(0,1)$, lorsque n tend vers l'infini. On en déduit, par le principe d'approximation, que, pour $n > 30$, *on peut approcher dans les calculs la loi de* $\dfrac{(U_n - n)}{\sqrt{2.n}}$ *par la loi normale réduite* $N(0,1)$.

Théorème 4 *Soit $X_1, \ldots, X_N$ un échantillon de modèle X suivant la loi normale $N(m,\sigma)$. Alors :*

1. $N.\dfrac{V^*}{Var(X)} = N.\left(\dfrac{\sigma^*}{\sigma}\right)^2$ *suit la loi du Chi-deux à N degrés de liberté*
2. $\dfrac{(N-1).V^*_{N-1}}{Var(X)} = (N-1).\left(\dfrac{\sigma^*_{N-1}}{\sigma}\right)^2$ *suit la loi du Chi-deux à $N-1$ degrés de liberté.*

Preuve : Le premier point découle immédiatement des définitions.

Pour le second point, on va utiliser les calculs auxiliaires précédents (vus en 2.2.1) : $\dfrac{(N-1).V^*_{N-1}}{Var(X)} = (N-1).\left(\dfrac{\sigma^*_{N-1}}{\sigma}\right)^2 = \displaystyle\sum_{n=1}^{N}\left(\dfrac{X_n - \overline{X}}{\sigma}\right)^2 = \sum_{n=1}^{N-1}\left(\dfrac{Z_n}{\sigma}\right)^2$

Le fait que les $\dfrac{Z_n}{\sigma}$ forment un échantillon de la loi $N(0,1)$ termine la démonstration.

2.2.3 Loi de STUDENT

Soit $X_0, \ldots, X_N$ un échantillon de modèle X qui suit la loi N(0,1). On considère la variable aléatoire suivante : $St_n = X_0.\left[\dfrac{1}{n}.\displaystyle\sum_{k=1}^{n} K_k^2\right]^{-1}$

On dit que St_n suit la loi de Student à n degrés de liberté.

Comme dans le cas de la loi du Chi-deux, les calculs donnant la fonction de répartition donnent lieu à une intégrale en dimension multiple pour laquelle un calcul préliminaire (utilisant le même passage en coordonnées sphériques) permet de se ramener à un calcul en dimension 1. L'intégrale obtenue après ce changement de variable est alors numériquement calculable : les résultats sont consignés en Annexe 3.

Comme dans le cas de la loi du Chi-deux, les tables sont données jusqu'à $N = 30$ degrés de liberté. Pour $n > 30$, on utilise le *principe d'approximation* : la loi de Student peut s'approcher par une loi normale réduite N(0,1).

En effet : $\sqrt{\dfrac{1}{n}.\displaystyle\sum_{k=1}^{n} X_k^2}$ donne des valeurs* proches de $\sqrt{E\left(X^2\right)} = 1$ (* lors des réalisations). Donc St_n est proche de X_0.

<u>Théorème</u> 5 *Soit $X_1, \ldots, X_N$ un échantillon de modèle X suivant la loi normale $N(m,\sigma)$.*

Alors : $\dfrac{\sqrt{N}.\left(\overline{X}-m\right)}{\sigma^*_{N-1}}$ *suit la loi de Student à $N-1$ degrés de liberté.*

<u>Preuve</u> :

$$S = \frac{\sqrt{N}.\left(\overline{X}-m\right)}{\sigma^*_{N-1}} = Z_0.\left(\sqrt{\frac{1}{n}\sum_{n=1}^{N-1}\left(X_n-\overline{X}\right)^2}\right)^{-1}$$

$S = \sqrt{N-1}.Z_0.\sqrt{\sum_{n=1}^{N-1} Z_n^2} = \sqrt{N-1}.\dfrac{Z_0}{\sigma}.\sqrt{\sum_{n=1}^{N-1}\left(\dfrac{Z_n}{\sigma}\right)^2}$ (d'après les notations des calculs auxiliaires (cf 2.2.1)).

On a donc bien : $\dfrac{\sqrt{N-1}.\left(\overline{X}-m\right)}{\sigma^*_{N-1}}$ suit la loi de Student à $N-1$ degrés de liberté.

Autre principe d'approximation : Soit $X_1, \ldots, X_N$ un échantillon de modèle X (on ne suppose plus que X suit une loi normale). Alors, pour $N > 30$ on peut approcher la loi de : $\dfrac{\sqrt{N-1}.\left(\overline{X}-m\right)}{\sigma^*_{N-1}}$ par la loi normale réduite $N(0,1)$.

En effet : $\dfrac{\sqrt{N-1}.\left(\overline{X}-m\right)}{\sigma^*_{N-1}} \cong \dfrac{\sqrt{N-1}.\left(\overline{X}-m\right)}{\sigma}$ (car σ est proche de $\sigma N-1^*$).

Ceci donne donc : $\dfrac{1}{\sqrt{N}}.\sum_{k=1}^{n}\dfrac{X_k-m}{\sigma}$. Le théorème central limite justifie alors ce principe d'approximation.

2.3 Estimation par intervalle de confiance

2.3.1 But de l'estimation par intervalle de confiance

Dans l'estimation ponctuelle, on a une variable aléatoire A (fonction des variables aléatoire $X_1, \dots, X_N$) appelée *estimateur de a* telle que l'on a de bonnes raisons de penser que, lors de la réalisation des variables aléatoires $X_1, \dots, X_N$, qui donne les observations $x_1, \dots, x_N$, **la valeur donnée par A est proche de a**.

Dans l'estimation par intervalle de confiance, il s'agit de donner ***une évaluation de la précision de l'estimateur ponctuel*** A.

On cherche un intervalle $I = [\alpha(X_1, \dots, X_N), \beta(X_1, \dots, X_N)]$ *entourant l'estimateur* A et tel qu'on aura confiance - dans l'hypothèse $a \in I$ - avec un niveau de confiance donné (99%, 95% ou encore 90%). En général, le niveau de confiance utilisé est 95%).

On écrit : $\boxed{p(a \in [\alpha(X_1, \dots, X_N), \beta(X_1, \dots, X_N)]) \leq \textit{niveau de confiance.}}$

Ceci peut se récrire en utilisant la notion de seuil $\varepsilon = 1 - \textit{niveau de confiance}$ (en général, $\varepsilon = 5\%$) : $p(a \notin [\alpha(X_1, \dots, X_N), \beta(X_1, \dots, X_N)]) < \varepsilon$.

Naturellement, on cherche l'intervalle de confiance le plus précis et vérifiant la condition de seuil imposé.

Lors de la réalisation des variables aléatoires $X_1, \dots, X_N$, qui donne les observations $x_1, \dots, x_N$, $[\alpha(X_1, \dots, X_N), \beta(X_1, \dots, X_N)]$ fournit l'intervalle de confiance : $[\alpha(x_1, \dots, x_N), \beta(x_1, \dots, x_N)]$.

2.3.2 Un cas standard

On cherche l'intervalle entourant $A(X_1, \dots, X_N)$ sous la forme $[A(X_1, \dots, X_N) - r, A(X_1, \dots, X_N) + r]$. Il s'agit alors de déterminer le rayon minimum r_{min} tel que l'on ait la condition de seuil : $\boxed{p(|A(X_1, \dots, X_N) - a| > r_{min}) < \varepsilon}$

<u>**Cas pratique**</u> : Soit X une variable aléatoire suivant la loi $N(a,\sigma)$. On cherche le rayon r tel que : $p\left(|a - \overline{X}| \leq r\right) \geq 0.95$.

On sait : $\overline{X}$ suit la loi $N\left(a, \dfrac{\sigma}{\sqrt{N}}\right)$.

On cherche donc r tel que : $p\left(\left|\dfrac{a - \overline{X}}{\sigma}\right| . \sqrt{N} < r . \dfrac{\sqrt{N}}{\sigma}\right) \geq 0.95$

On utilise alors la table pour trouver r_max.

Exemple : Intervalle de confiance dans l'estimation de l'espérance $E(X)$ d'un caractère X suivant la loi normale $N(m,\sigma)$.

Dans ce cas, on a : $A(X_1, \ldots, X_N) = \overline{X}$.

On cherche alors la borne $\alpha(X_1, \ldots, X_N)$ sous la forme $\overline{X} - r.\sigma^*_{N-1}$ et la borne $\beta(X_1, \ldots, X_N)$ sous la forme $\overline{X} + r.\sigma^*_{N-1}$ (naturellement, dans un cas où σ est connu, on l'utilise directement au lieu de se servir de son estimateur σ^*_{N-1}).

La condition de seuil s'écrit : $p\left(\left|\overline{X} - m\right| > r.\sigma^*_{N-1}\right) < \varepsilon$

C'est à dire : $p\left(\left|\sqrt{N}.\dfrac{\overline{X} - m}{\sigma^*_{N-1}}\right| > \sqrt{N}.r\right) < \varepsilon$.

Or, d'après le théorème 2.2.3 : $\sqrt{N}.\dfrac{\overline{X} - m}{\sigma^*_{N-1}}$ suit la loi de Student à $N-1$ degrés de liberté. On peut donc trouver le réel minimal $\sqrt{N}.r$ tel que la condition de seuil sera réalisée. On trouve donc le réel r cherché.

Ainsi, lorsque l'on a les observations $x_1, \ldots, x_N$ - obtenues comme réalisation de l'échantillon $X_1, \ldots, X_N$ du caractère X - l'intervalle de confiance pour $E(X)$ obtenu avec le seuil ε est l'intervalle : $[\overline{x} - r.\sigma^*_{N-1}, \overline{x} + r.\sigma^*_{N-1}]$

2.3.3 Un cas particulier de l'estimation d'un écart type

Dans ce cas particulier, compte tenu de la nature de σ et de son rôle, on cherche l'intervalle de confiance entourant σ^* sous la forme $[0, r.\sigma^*]$, avec $r > 1$.

Naturellement, dans le cas où m n'est pas connu (ce qui le cas en général), on remplace σ^* par σ^*_{N-1}.

La condition de seuil s'écrit : $p(r.\sigma^* < \sigma) < \varepsilon$ (ou, dans l'autre cas : $p(r.\sigma^*_{N-1} < \sigma) < \varepsilon$).

Ceci donne : $p\left(N.\left(\dfrac{\sigma^*}{\sigma}\right)^2 < \dfrac{N}{r^2}\right) < \varepsilon$ ou, dans l'autre cas :

$p\left(N.\left(\dfrac{\sigma^*_{N-1}}{\sigma}\right)^2 < \dfrac{N}{r^2}\right) < \varepsilon$

D'après le théorème 2.2.2 et en supposant que le caractère X étudié suit une loi normale $N(m,\sigma)$, on trouve la valeur minimale de r en utilisant *la loi du Chi-deux à N degrés de liberté* (ou à $N-1$ degrés de liberté dans l'autre cas).

Chapitre 3

Tests d'hypothèses

3.1 Test de Bayes

3.1.1 Principe général

Dans un test d'hypothèse, il s'agit de tester une hypothèse H contre une hypothèse K. Le plus souvent, K n'est pas mentionnée et est sous entendue comme l'hypothèse contraire de H.

Il faut déterminer une condition de rejet de H de manière à ce que l'on puisse, grâce à des observations $x_1, \ldots, x_N$ (qui sont une réalisation de $X_1, \ldots, X_N$), éventuellement pouvoir rejeter H.

Naturellement, il existe deux types de risque d'erreur :

– le risque de rejeter H si H est vraie, qui correspond à la probabilité $p(\text{rejet de } H|H \text{ vraie})$ et qui a un coût C_H que l'on estime
– le risque de rejeter K si K est vraie, qui correspond à la probabilité $p(\text{rejet de } H|K \text{ vraie})$ et qui a un coût C_K que l'on estime

Il s'agit de déterminer une condition naturelle de rejet de H qui <u>minimise le coût moyen du risque d'erreur</u> :

$$C = C_H.p(\text{rejet de } H|H \text{ vraie}) + C_K.p(\text{rejet de } K|K \text{ vraie})$$

3.1.2 Exemple

Le principe du test de Bayes es utilisé pour la détection d'un signal en présence *d'un bruit blanc gaussien additif.*

On reçoit les observations $x_1, \ldots, x_N$ qui sont une réalisation du processus $X_1, \ldots, X_N$, échantillon de modèle X.

On note : $\begin{cases} X = 1 + B \text{ , si le signal est présent} \\ X = B \text{ , sinon} \end{cases}$

avec B le bruit blanc, B étant supposé suivant la loi normale $N(0,\sigma)$.

L'hypothèse H est : 'le signal est présent'. Le risque correspondant au rejet de H s'appelle ***défaut de détection*** $p(\text{rejet de} H|H \text{ vraie})$ et a un coût C_H que l'on estime.

L'hypothèse K est 'le signal est absent' ($i.e$ le contraire de H). Le risque correspondant au rejet de K s'appelle ***fausse alarme***$p(\text{rejet de } K|K \text{ vraie})$ et a un coût C_K que l'on estime.

Une condition naturelle de rejet de H est $\overline{X} < r$. On doit donc déterminer r pour minimiser le coût C, coût moyen du risque d'erreur.

$$C = C_H.p(\text{rejet de } H|H \text{ vraie}) + C_K.p(\text{rejet de } K|K \text{ vraie})$$

$$C = C_H.p(\overline{X} < r|X = 1 + B) + C_K.p(\overline{X} > r|X = B)$$

$$C = C_H.p(1 + \overline{B} < r) + C_K.p(\overline{B} > r)$$

$$C = C_H.p\left(\sqrt{N}.\frac{\overline{B}}{\sigma} < \sqrt{N}.\frac{r-1}{\sigma}\right) + C_K.p\left(\sqrt{N}.\frac{\overline{B}}{\sigma} > \sqrt{N}.\frac{r}{\sigma}\right)$$

Or : $\sqrt{N}.\dfrac{\overline{B}}{\sigma}$ suit la loi normale réduite.

On note alors F la fonction de répartition de la loi normale réduite et f sa densité.

$$C = C_H.F\left(\sqrt{N}.\frac{r-1}{\sigma}\right) + C_K.F\left(-\sqrt{N}.\frac{r}{\sigma}\right)$$

On dérive, pour trouver le minimum :

$$C' = C_H.\frac{\sqrt{N}}{\sigma}.f\left(\sqrt{N}.\frac{r-1}{\sigma}\right) + C_K.\frac{\sqrt{N}}{\sigma}.f\left(\sqrt{N}.\frac{r}{\sigma}\right)$$

$$C' = 0 \Leftrightarrow \log(C_H) - \left[\frac{\sqrt{N}.(r-1)}{\sqrt{2}.\sigma}\right]^2 = \log(C_K) - \left[\frac{-\sqrt{N}.r}{\sqrt{2}.\sigma}\right]^2$$

Donc : $\dfrac{2}{N}.\log\left(\dfrac{C_H}{C_K}\right) = \left(\dfrac{r-1}{\sigma}\right)^2 - \left(\dfrac{r}{\sigma}\right)^2 = \dfrac{1-2r}{\sigma^2}$

Ainsi : $2.r = 1 + \dfrac{2.\sigma^2}{N}.\log\left(\dfrac{C_H}{C_K}\right) \Rightarrow r = \dfrac{1}{2} + \dfrac{\sigma^2}{N}.\log\left(\dfrac{C_H}{C_K}\right)$

Vérifier que ce résultat est bien conforme à l'intuition (par exemple si $C_K = C_H$, si $C_K > C_H$) et regarder la limite du coût moyen lorsque N tend vers l'infini.

3.2 Test de Neymann-Pearson

3.2.1 Principe général

Dans ces tests d'hypothèse, il s'agit aussi de tester une hypothèse H contre une hypothèse K qui, le plus souvent, est l'hypothèse contraire de H et est sous-entendue comme telle.

Il fat déterminer ***une condition de rejet de*** H de manière à ce que l'on puisse, grâce à des observations $x_1, \ldots, x_N$ (qui sont une réalisation de $X_1, \ldots, X_N$), éventuellement pouvoir rejeter H avec une grande confiance. Ceci se traduit par la condition de seuil : $\boxed{p(\text{rejet de } H|H \text{ vraie}) \leq \varepsilon}$

On appelle **région de rejet** *l'ensemble des N-uples d'observations $x_1, \ldots, x_N$ qui vérifient la condition de rejet.* Naturellement on cherche à pouvoir rejeter H le plus facilement possible, $\boxed{\text{tout en conservant la condition de seuil}}$ (c'est à dire en gardant la même confiance dans le rejet de H). On est alors *conduit à chercher la région de rejet* $\boxed{\textit{la plus grande possible}}$ qui vérifie la condition de seuil.

Deux cas peuvent se présenter :

- l'hypothèse H détermine complètement $p(\text{rejet de } H|H \text{ vraie})$. Dans ce cas on peut calculer directement cette probabilité. On dit qu'il s'agit d'une hypothèse simple (c'est, par exemple, le cas dans l'étude d'un caractère qui suit la loi $N(m,1)$ pour l'hypothèse : '$m = 2$').
- l'hypothèse H ne détermine pas complètement $p(\text{rejet de } H|H \text{ vraie})$. Dans ce cas on ne peut calculer directement cette probabilité. On parle alors d'hypothèse composée (c'est le cas, par exemple, dans l'étude d'un caractère suivant la loi $N(m,1)$ pour l'hypothèse '$m < 2$').
  Une hypothèse composée nécessite donc un calcul préliminaire de majoration de $p(\text{rejet de } H|H \text{ vraie})$.

Lorsque l'on a déterminé la région de rejet la plus grande possible qui vérifie la condition de seuil, il ne reste plus qu'à regarder si les observations sont cette région de rejet.

Bien remarquer le rôle dissymétrique des hypothèses H et K.

Exemple : On choisit les hypothèses : $H \rightarrow$ '$m = 0$' et $K \rightarrow$ '$m = 1$'. On suppose que $\overline{X}$ suit la loi $N\left(m, \frac{2}{\sqrt{N}}\right)$.

On décide de rejeter H si $\overline{X} > r$. La condition de seuil est donc la suivante :

$p\left(\overline{X} > r | m = 0\right) < 0.05 \Leftrightarrow p\left(\frac{\sqrt{N}.\overline{X}}{2} > \frac{\sqrt{N}.r}{2}\right) < 0.05$. On rappelle que : $\frac{\sqrt{N}.\overline{X}}{2}$ suit la loi $N\left(0, \frac{2}{\sqrt{N}}\right)$.

Pour $N = 4$, la table donne : $\frac{2.\sqrt{N}}{r} = 1.65 \Rightarrow r = 1.65 \Rightarrow r \geq 1.65$

3.2.2 Test de comparaison des moyennes de deux caractères

Il s'agit ici de tester l'hypothèse de l'égalité des espérances (que l'on note m_X et m_Y) de deux caractères X et Y.

On dispose d'observations $x_1, \ldots, x_M$ et $y_1, \ldots, y_N$ qui sont des réalisations des variables aléatoires $X_1, \ldots, X_M$ et $Y_1, \ldots, Y_N$ qui sont des échantillons des deux caractères X et Y (supposées globalement indépendantes).

Deux cas peuvent se présenter.

3.2.2.1 cas des grands échantillons ($M > 30$ et $N > 30$ en pratique)

Dans ce cas, on utilisera <u>le principe d'approximation suivant</u> : "La loi de $\frac{\left(\overline{Y} - \overline{X}\right) - (m_Y - m_X)}{\sqrt{\frac{C^*_{Y,Y}}{N} + \frac{C^*_{X,X}}{N}}}$ s'approche par la loi normale réduite $N(0, 1)$".

On peut justifier ce principe d'approximation de la manière suivante : dans le numérateur, $\left(\overline{Y} - m_y\right)$ se présente comme une somme de variables aléatoires centrées et de même variance. Donc le principe d'approximation issu du théorème centrale limite nous dit que la loi de $\left(\overline{Y} - m_y\right)$ s'approche par la loi normale $N\left(0, \dfrac{\sigma_y}{\sqrt{N}}\right)$. Il en est de même pour $\left(\overline{X} - m_X\right)$.

La loi du numérateur s'approche donc par la loi normale centrée de variance $\dfrac{\sigma_Y^2}{N} + \dfrac{\sigma_Y^2}{M}$. Or, on sait que $C^*_{Y,Y}$ approche σ_Y^2 et que $C^*_{X,X}$ approche σ_X^2. Donc, $\sqrt{\dfrac{C^*_{Y,Y}}{N} + \dfrac{C^*_{Y,Y}}{N}}$ approche l'écart type du numérateur.

L'hypothèse H est '$m_X = m_Y$'.

La condition de rejet sera donc : $\dfrac{\left|\overline{Y} - \overline{X}\right|}{\sqrt{\frac{C^*_{Y,Y}}{N} + \frac{C^*_{X,X}}{M}}} > r$

Remarque : le dénominateur introduit dans cette condition de rejet a pour but de permettre d'utiliser le principe d'approximation donné ci-dessus.

La condition de seuil $p(\text{rejet de } H|H \text{ vraie})$ s'écrit :

$$p\left(\frac{\left|\overline{Y} - \overline{X}\right|}{\sqrt{\dfrac{C^*_{Y,Y}}{N} + \dfrac{C^*_{X,X}}{M}}} > r|m_Y = m_X\right) \leq \varepsilon$$

On lit donc, d'après le principe d'approximation donné ci-dessus, le réel r cherché qui détermine la condition de rejet dans *la table de la loi normale réduite.*

3.2.2.2 cas des petits échantillons

Dans ce cas, on ne peut plus utiliser le principe d'approximation précédent. On est alors obligé de faire l'hypothèse (**qui doit être justifiée dans chaque d'utilisation**) que les deux caractères X et Y suivent des lois normales de même variance, soit σ^2.

Dans ces conditions, on a la propriété suivante :

Propriété 6 *$c.\dfrac{\left(\overline{Y} - \overline{X}\right) - (m_Y - m_X)}{\sqrt{N.C^*_{Y,Y} + M.C^*_{X,X}}}$ avec $c = \sqrt{\dfrac{M.N.(M+N-2)}{M+N}}$ suit une loi de Student à $M + N + 2$ degré de liberté.*

Preuve : Les calculs auxiliaires du 2.2.1 permettent d'écrire $N.C^*_{Y,Y} + M.C^*_{X,X}$ comme suit : $\sum_{n=1}^{N-1}\left(Z'_n\right)^2 + \sum_{n=1}^{M-1}\left(Z''_n\right)^2$

Le dénominateur devient donc (en regroupant les Z' et les Z'' : $\sqrt{\sum_{n=1}^{M+N-2} Z_n^2}$

On pose : $Z_0 = \sqrt{\frac{M.N}{M+N}}.\left[\left(\overline{Y}-\overline{X}\right)(m_Y - m_X)\right]$. Z_0 est de variance σ^2.

Finalement, l'expression énoncée dans la propriété s'écrit :

$\frac{Z_0}{\sigma} \times \left(\sqrt{(M+N+2).\sum_{n=1}^{M+N+2}\left(\frac{Z_n}{\sigma}\right)^2}\right)^{-1}$ ce qui correspond à la loi de Student à $M+N-2$ degrés de liberté.

L'hypothèse H est "$m_x = m_y$".

La condition de rejet est donc sera donc : $\frac{c.\left|\overline{Y}-\overline{X}\right|}{\sqrt{N.C^*_{Y,Y} + M.C^*_{X,X}}} > r$ avec $c = \sqrt{\frac{M.N.(M+N-2)}{M+N}}$

La condition de seuil p(rejet de $H|H$ vraie) s'écrit :

$$p\left(\frac{c.\left|\overline{Y}-\overline{X}\right|}{\sqrt{N.C^*_{Y,Y} + M.C^*_{X,X}}} > r | m_Y = m_Y\right) \leq \varepsilon$$

On lit donc, d'après ma propriété donnée ci-dessus, le réel r cherché qui détermine la région de rejet dans la table de la loi de Student à $M+N-2$ degrés de liberté.

3.2.3 Test d'ajustement du Chi-deux

Dans ce test, on dispose d'observations $x_1, \ldots, x_N$ qui sont une réalisation des variables aléatoires $X_1, \ldots, X_N$ qui forment un échantillon du caractère X dont on étudie la loi.

On considère une loi P. L'hypothèse est "*le caractèreX suit la loi P*".

Ce test est dit ***non paramétrique*** car il ne concerne pas un paramètre de la loi du caractère X, mais la nature de sa loi.

Principe du test :
Comme dans le 1.3.2, on considère une succession d'intervalles $(A_i)_{i\in[1,q]}$ - qu'on appelle **classes** - qui forment une partition des valeurs possibles pour le caractère X.

Pour $i \in [1, q]$, on considère la variable aléatoire P_i, fréquence de la classe A_i dans l'échantillon. Alors, $N.P_i$ est une variable aléatoire discrète qui suit la loi binomiale de paramètre $p_i^{'} = p(X \in A_i)$ et d'ordre N. Ceci s'approche (pour N assez grand) par la loi de Poisson de paramètre $N.p_i^{'}$.

Des considérations déjà faîtes dans la remarque fondamentale du 1.3.2, on déduit que, pour N suffisamment grand, les P_i donnent des valeurs p_i proches de $p_i^{'}$ lors d'une réalisation.

La variable aléatoire $U = N.\sum_{i=1}^{q} \frac{(P_i - p_i^{'})^2}{p_i^{'}}$ prendra des valeurs d'autant plus petites que les P_i prendront des valeurs proches des $p_i^{'}$. On pourra donc rejeter l'hypothèse H si U est trop grande.

Donc, la condition de rejet de H est : $U > \beta$, où β est déterminée par la condition de seuil $p(\text{rejet de } H|Htextrmvraie)$. Ceci se traduit ici par : $p\,(U > \beta|X \text{ suit la loi } P) \leq \varepsilon$

Il faut donc connaître la loi de U, ce qui est l'objet du <u>*principe d'approximation*</u> suivant :

Pour N suffisamment grand ($N > 30$ en pratique), la loi de U s'approche par la loi du Chi-deux à $q - 1$ degrés de liberté

(d'où le nom du test)

Justification du principe d'approximation :

On écrit : $U = N.\sum_{i=1}^{q} \frac{(P_i - p_i^{'})^2}{p_i^{'}} = \sum_{i=1}^{q} \left(\frac{\sqrt{N}.(P_i - p_i^{'})}{\sqrt{p_i^{'}}} \right)^2$

Chacune des variables aléatoires $\frac{\sqrt{N}.(P_i - p_i^{'})}{\sqrt{p_i^{'}}}$ peut s'écrire sous la forme d'une somme de N variables aléatoires **indépendantes**, centrées et de même variance. Le principe d'approximation issu du théorème central limite nous dit que la loi de de ces variables s'approche par la une loi normale dont on vérifie aisément qu'elle centrée et de variance égale à 1.

U apparaît donc sous la forme d'une somme de carrés de variables aléatoires suivant la loi normale réduite. Cependant, ces variables **ne sont pas indépendantes** car $\sum P_i - p_i^{'} = 0$.

La méthode de calcul utilisée au 2.2.1 permet de les remplacer par des Z_i qui forment un échantillon de taille $q-1$ de la loi normale réduite. Ainsi, U apparaît sous la forme de la somme de carrés de $q-1$ variables aléatoires indépendantes qui suivent la loi normale réduite, ce qui correspond à la loi du Chi-deux à $q-1$ degrés de liberté.

Avec les observations données, il suffit alors de regarder si la valeur prise par U est supérieure à β.

Remarque importante : Dans cette table du Chi-deux à $q-1$ degrés de liberté, on trouve aussi un réel α tel que : $p(U_{q-1} < \alpha) = \varepsilon$. Si la valeur prise par U avec les observations est inférieure à α, c'est que les P_i sont **trop proches** des $p_i^{'}$. Dans ce cas, c'est la validité des observations que l'on rejette.

En résumé : Si la valeur u de U obtenue avec les observations est telle que : $\begin{cases} u > \beta, \text{ on rejette l'hypothèse } H \\ u < \alpha, \text{ on rejette la validité des observations} \end{cases}$.

Autre remarque importante : On imagine bien que pour que ce test soit fiable il faut respecter quelques règles heuristiques lors de la formation des classes A_i :

– chaque classe doit être assez grande ($N.p_i^{'}$, le nombre d'éléments attendus dans la classe A_i, doit être supérieur à 5 (sauf éventuellement pour la première et la dernière classe où l'on peut se contenter d'un nombre d'éléments attendus simplement supérieur à 1))

– il doit y avoir assez de classes (q, le nombre de classes, doit être supérieur à 4 fois le nombre de paramètres intervenant dans la loi de P)

3.2.4 Test des Longueurs

Il s'agit aussi d'un test non paramétrique. Il s'agit de tester l'hypothèse "*deux caractères X et Y ont la même loi*".

On dispose d'observations $x_1, \ldots, x_M$ et $y_1, \ldots, y_N$ qui sont une réalisation des variables aléatoires $X_1, \ldots, X_N$ et $Y_1, \ldots, Y_N$, deux échantillons des caractères X et Y supposés globalement indépendants.

L'hypothèse H est "*X et Y ont même loi*".

On va définir une variable aléatoire L qui nous permettra éventuellement de pouvoir rejeter l'hypothèse H.

Pour une réalisation quelconque de $X_1, \ldots, X_M$ et de $Y_1, \ldots, Y_N$, on effectue un rangement par valeurs croissantes (les x et les y étant mélangés). Dans un tel rangement, on appelle **longueur de** x une succession de lettres x telle qu'il y ait un y ou rien (s'il s'agit du début ou de la fin de la série statistique) avant et après. Même définition pour y.

L sera la variable aléatoire *nombre total de longueurs*. Si les deux caractères suivent la même loi, alors les x et les y se répartissent au hasard dans le rangement décrit précédemment. Au contraire, si les deux caractères ne suivent pas la même loi, alors les x et les y s'accumulent en des endroits distincts, provoquant une diminution du nombre des longueurs. On rejettera donc H si L est trop petit.

La condition de rejet de H est : $L < r$.

La condition de seuil $p(\text{rejet de } H | H \text{ vraie}) \leq \varepsilon)$ s'écrit ici : $p(L < r | H \text{ vraie}) \leq \varepsilon$

Le plus grand entier r (afin d'obtenir une région de rejet la plus grande possible) qui permet la condition de seuil se cherche donc avec la loi de L sachant H.

Donnons à présent cette loi : "*Sachant que H est vraie*" implique que tous les rangements sont équiprobables.

Donc on a : $p(L = k | H \text{ vraie})$, ce qui est donnée par la formule célèbre :
$\dfrac{\text{nombre de cas favorables}}{\text{nombre de cas possibles}}$

Résultat : On suppose $M \leq N$ (l'autre cas s'en déduit aisément).

- $\underline{k = 2.s}$, avec $1 \leq s \leq M \rightarrow p(L = k|H \text{ vraie}) = 2.C_{M-1}^{s-1} + \dfrac{C_{M-1}^{s-1}}{C_{M+N}^{M}}$
- $\underline{k = 2.s + 1}$, avec $1 \leq s \leq M$
  $\rightarrow p(L = k|H \text{ vraie}) = \dfrac{C_{M-1}^{s}.C_{M-1}^{s-1} + C_{M-1}^{s-1}.CN - 1s}{C_{M+N}^{N}}$
- $\underline{k = 2.M + 1}$, avec $M < N \rightarrow p(L = k|H \text{ vraie}) = \dfrac{C_{N-1}^{M}}{C_{M+N}^{M}}$

Preuve : Le nombre cas possibles est toujours C_{M+N}^{M}.

Pour compter le nombre de cas favorables, supposons $k = 2.s$ et considérons le cas où le la 1ère longeur est formée avec des x. Il y a donc s longueurs de x. Il y a C_{M-1}^{s-1} manières de remplir ces s longeurs avec des x. On le démontre par récurrence sur s.

La propriété est vraie pour $s = 1$. Supposons la vraie pour s. Pour $s + 1$, pour la dernière longeur il peut y avoir :

- un x, il en reste $M - 1$ pour les s autres cases, soit C_{M-2}^{s-1} manières
- deux x, il en reste alors $M - 2$ pour les s autres case, soit C_{M-3}^{s-1} manières

$$\vdots$$

- $(M - s)$ x, il en reste alors s pour les s autres cases, soit C_{s-1}^{s-1} manières

D'après la formule du triangle de Pascal : $\sum i = s - 1^{M-2} C_{i}^{s-1} = C_{M-1}^{s}$

Comme on a C_{M-1}^{s-1} possibilités pour les x, il en reste C_{N-1}^{s} pour les y et il faut multiplier par deux pour tenir compte du cas où la $1^{ère}$ longeur est formée avec des y.

Pour $k = 2.s + 1$ (avec $1 \leq s \leq M$), le raisonnement est identique. Dans la somme $C_{M-1}^{s}.C_{N-1}^{s-1} + C_{M-1}^{s-1}.C_{N-1}^{s}$, le 1er terme correspond au cas où il y a une longueur de plus de x que de y et le secon à l'autre cas.

Pour $k = 2.M + 1$, c'est comme précédemment, mais il ne peut y avoir plus de longeurs de x que de y, d'où le résultat.

Mais ces résultats ne sont pas commodément utilisables si M et N sont asez grands. Dans ce cas, on préfère utiliser le principe d'approximation ci-dessous.

Principe d'approximation : Pour M et N assez grand (ici supérieur tous deux à 15), on considère que la loi de L s'approche par la loi normale $N(m,\sigma)$

avec : $\left\{ \begin{array}{l} m = 1 + 2.\dfrac{M.N}{M+N} \\ \sigma^2 = 2.\dfrac{M.N.(M^2+N^2-M-N)}{(M+N)^2.(M+N-1)} \end{array} \right.$

Justification : Pour $i \in [2, M+N]$, on définit la variable aléatoire C_i, caractéristique de l'évènement "changement de lettre au rang i", par :

$\left\{ \begin{array}{l} C_i = 1 \text{ , si la lettre au rang } i \text{ n'est pas la même qu'au rang } i-1 \\ C_i = 0 \text{ , sinon} \end{array} \right.$

On a alors :

$p(C_i = 1) = p(x \text{ au rang } i \text{ et } y \text{ au rang } i-1) + p(y \text{ au rang } i \text{ et } x \text{ au rang } i-1)$

$$p(C_i = 1) = p(x \text{ au rang } i | y \text{ au rang } i-1).p(y \text{ au rang } i-1) +$$
$$p(y \text{ au rang } i | x \text{ au rang } i-1).p(x \text{ au rang } i-1)$$

$$p(C_i = 1) = \frac{M}{M+N-1} \times \frac{N}{M+N} + \frac{N}{M+N-1} \times \frac{M}{M+N}$$

$p(C_i = 1) = 2.\dfrac{M.N}{(M+N-1).(M+N)}$ (en effet, C_i est une variable caractéristique d'un évènement)

La variance de C_i, notée $Var(C_i)$ et vaut :

$Var(C_i) = E\left(C_i^2\right) - (E(C_i))^2 = E(C_i) - (E(C_i))^2$

Ainsi : $Var(C_i) = 2.\dfrac{M.N.(M^2+N^2-N-M)}{(M+N)^2.(M+N-1)^2}$

On remarque que ni $E(C_i)$ ni $Var(C_i)$ ne dépendent de i.

On pose : $Nor = \dfrac{1}{\sqrt{M+N-1}} \times \displaystyle\sum_{i=2}^{M+N} \dfrac{C_i - E(C_i)}{\sqrt{Var(C_i)}}$

On a donc : $Nor \times \sqrt{Var(C_i) \times (M+N-1)} + (M+N-1).E(C_i) = \displaystyle\sum_{i=2}^{M+N} C_i$

Le principe d'approximation issu du théorème central limite nous dit que, pour M et N suffisamment grands, la loi de Nor s'approche par une loi normale réduite.

Or, on a :

$L = 1 + \sum i = 1^{M+N} C_i = 1 + Nor \times \sqrt{Var(C_i) \times (M+N-1)} + (M+N-1).E(C_i)$

Donc L suit une loi d'espérance :

$E(L) = 1 + (M+N-1).E(C_i) = 1 + 2 \times \dfrac{M.N}{M+N}$

La variance de L est :

$Var(L) = Var(C_i).(M+N-1) = 2 \times \dfrac{M.N(M^2+N^2-M-N)}{(M+N)^2.(M+N-1)}$

3.2.5 Test des rangs de X

Il s'agit aussi d'un test non paramétrique. Il s'agit de tester l'hypothèse H "*deux caractères X et Y ont la même loi*" contre l'hypothèse K "*Y est stochastiquement supérieure à X*", où Ystochastiquement supérieure à X signifie : $\forall z \in \mathbb{R}$, $p(Y \leq z) \leq p(X \leq z)$ et avec inégalité stricte **pour au moins une valeur de** z.

On dispose d'observations $x_1, \ldots, x_M$ et $y_1, \ldots, y_N$ qui sont une réalisation des variables aléatoires $X_1, \ldots, X_M$ et $Y_1, \ldots, Y_N$ qui sont des échantillons des caractères X et Y, supposés globalement indépendants.

Remarquer dans ce test la dissymétrie du rôle de X et de Y 5c'est ce qui fait la différence avec le test des longueurs vu en 3.2.4). On va définir une variable aléatoire R adaptée à cette dissymétrie et qui nous permettra éventuellement de pouvoir rejeter l'hypothèse H.

Pour une réalisation quelconque de $X_1, \ldots, X_M$ et de $Y_1, \ldots, Y_N$, on effectue un rangement par valeurs croissantes (les x et les y étant mélangés). Dans un tel rangement, on compte la somme des rangs des lettres x.

R sera la variable aléatoire *somme des rangs des lettres x*.

Si les deux caractères suivent la même loi, alors les x et les y se répartissent aléatoirement dans le rangement décrit précédemment. Au contraire, si les deux caractères suivent deux lois différentes, alors c'est K qui est vraie. Les valeurs de Y ont tendances à être supérieures à celle de X. Les x s'accumulent donc au début du rangement, ce qui a pour effet de diminuer R. On rejettera donc H si R est trop petit.

La condition de rejet de H est : $R < t$

La condition de seuil $p(\text{rejet de } H | H \text{ vraie}) \leq \varepsilon$ s'écrit : $p(R < t | H \text{ vraie}) \leq \varepsilon$

Le plus grand entier t (afin d'obtenir une région de rejet la plus grande possible) vérifiant la condition de seuil se cherche donc avec la loi de R sachant H.

Cependant, pour M et N suffisamment grands (en pratique supérieurs à 10), on utilise le principe d'approximation ci-dessous. Pour les petites valeurs de M et N, on lisse le soin à l'utilisateur de faire les calculs de dénombrement nécessaires à l'obtention de la loi de R sachant H.

Principe d'approximation : Pour M et N suffisamment grands (ici, supérieurs à 10), on considère que la loi de R sachant H s'approche par la loi normale $N(m,\sigma)$ avec : $\left\{\begin{array}{l} m = \dfrac{M.(M+N+1)}{2} \\ \sigma^2 = \dfrac{M.N.(M+N+1)}{12} \end{array}\right.$

Justification : Pour $i \in [1, M+N]$, on considère la variable aléatoire C_i définie par : $\left\{\begin{array}{l} C_i = i, \text{ si présence d'une lettre } x \text{ au rang } i \\ C_i = 0, \text{ sinon} \end{array}\right.$

On a : $R = \sum_{i=1}^{M+N} C_i$

D'après le principe d'approximation issu du théorème central limite, la loi de R s'approche par une loi normale dont il ne reste à calculer que l'espérance et la variance.

$$E(C_i) = i \times \frac{M}{M+N} \Rightarrow E(R) = \frac{M}{M+N} \times \sum_{i=1}^{M+N} i = \frac{M.(M+N+1)}{2}$$

$$Var(R) = E\left(R^2\right) - (E(R))^2$$

On connaît $E(R)$. Reste à calculer $E\left(R^2\right)$.

$$R^2 = \sum_{i=1}^{M+N} C_i^2 + \sum_{i \neq j}^{M+N} C_i.C_j$$

$$\text{D'où : } E(\left(R^2\right) = \sum_{i=1}^{M+N} i^2 \times \frac{M}{M+N} + \sum_{i \neq j}^{M+N} i.j \times \frac{M}{M+N} \times \frac{M-1}{M+N-1}$$

$$E\left(R^2\right) = \frac{M}{M+N} \times \sum_{i=1}^{M+N} i^2 + \frac{M.(M-1)}{(M+N).(M+N-1)} \times \sum_{i \neq j}^{M+N} i.j$$

Or : $\sum_{i=1}^{M+N} i^2 = \frac{(M+N).\left[2.(M+N)^2+3.(M+N)+1\right]}{6}$ et $\sum_{i \neq j}^{M+N} i.j = \left(\sum_{i=1}^{M+N} i\right)^2 - \sum_{i=1}^{M+N} i^2$

Ce qui donne donc : $E\left(R^2\right) = \frac{M.(M+N+1) \times (3.M+3.M.N+3.M+N)}{12}$

D'où : $Var(R) = \frac{M.N.(M+N+1)}{12}$

Annexe 1 : Valeurs de la fonction ϕ de répartition de la variable réduite de Gauss

Rappel : $\phi(t) = P_Z(]-\infty, t[= \frac{1}{\sqrt{2\pi}}.\int_{-\infty}^{t} \mathbf{e}^{-\ \frac{Z^2}{2}} dz$

t	$\phi(t)$	t	$\phi(t)$	t	$\phi(t)$	t	$\phi(t)$	t	$\phi(t)$	t	$\phi(t)$
0,00	0,5000	1,00	0,8413	2,00	0,9772	0,50	0,6915	1,50	0,9332	2,50	0,9938
0,01	0,5040	1,01	0,8437	2,01	0,9779	0,51	0,6950	1,51	0,9345	2,51	0,9940
0,02	0,5080	1,02	0,8461	2,02	0,9783	0,52	0,6985	1,52	0,9357	2,52	0,9941
0,03	0,5120	1,03	0,8485	2,03	0,9788	0,53	0,7019	1,53	0,9370	2,53	0,9943
0,04	0,5160	1,04	0,8508	2,04	0,9793	0,54	0,7054	1,54	0,9382	2,54	0,9945
0,05	0,5199	1,05	0,8531	2,05	0,9798	0,55	0,7088	1,55	0,9394	2,55	0,9946
0,06	0,5239	1,06	0,8554	2,06	0,9803	0,56	0,7123	1,56	0,9406	2,56	0,9948
0,07	0,5279	1,07	0,8577	2,07	0,9808	0,57	0,7157	1,57	0,9418	2,57	0,9949
0,08	0,5319	1,08	0,8599	2,08	0,9812	0,58	0,7190	1,58	0,9429	2,58	0,9951
0,09	0,5359	1,09	0,8621	2,09	0,9817	0,59	0,7224	1,59	0,9441	2,59	0,9952
0,10	0,5398	1,10	0,8643	2,10	0,9821	0,60	0,7257	1,60	0,9452	2,60	0,9953
0,11	0,5438	1,11	0,8665	2,11	0,9826	0,61	0,7291	1,61	0,9463	2,61	0,9955
0,12	0,5478	1,12	0,8686	2,12	0,9830	0,62	0,7324	1,62	0,9474	2,62	0,9956
0,13	0,5517	1,13	0,8708	2,13	0,9834	0,63	0,7357	1,63	0,9484	2,63	0,9957
0,14	0,5557	1,14	0,8729	2,14	0,9838	0,64	0,7389	1,64	0,9495	2,64	0,9959
0,15	0,5696	1,15	0,8749	2,15	0,9842	0,65	0,7422	1,65	0,9505	2,65	0,9960
0,16	0,5636	1,16	0,8770	2,16	0,9846	0,66	0,7454	1,66	0,9515	2,66	0,9961
0,17	0,5675	1,17	0,8790	2,17	0,9850	0,67	0,7486	1,67	0,9525	2,67	0,9962
0,18	0,5714	1,18	0,8810	2,18	0,9854	0,68	0,7517	1,68	0,9535	2,68	0,9963
0,19	0,5753	1,19	0,8830	2,19	0,9857	0,69	0,7549	1,69	0,9545	2,69	0,9964
0,20	0,5793	1,20	0,8849	2,20	0,9861	0,70	0,7580	1,70	0,9554	2,70	0,9965
0,21	0,5832	1,21	0,8869	2,21	0,9864	0,71	0,7611	1,71	0,9564	2,71	0,9966
0,22	0,5871	1,22	0,8888	2,22	0,9868	0,72	0,7642	1,72	0,9573	2,72	0,9967
0,23	0,5910	1,23	0,8907	2,23	0,9871	0,73	0,7673	1,73	0,9582	2,73	0,9968
0,24	0,5948	1,24	0,8925	2,24	0,9875	0,74	0,7703	1,74	0,9591	2,74	0,9969
0,25	0,5987	1,25	0,8944	2,25	0,9878	0,75	0,7734	1,75	0,9599	2,75	0,9970
0,26	0,6026	1,26	0,8962	2,26	0,9881	0,76	0,7764	1,76	0,9608	2,76	0,9971
0,27	0,6064	1,27	0,8980	2,27	0,9883	0,77	0,7794	1,77	0,9616	2,77	0,9972
0,28	0,6103	1,28	0,8997	2,28	0,9887	0,78	0,7823	1,78	0,9625	2,78	0,9973
0,29	0,6141	1,29	0,9015	2,29	0,9890	0,79	0,7852	1,79	0,9633	2,79	0,9974
0,30	0,6179	1,30	0,9032	2,30	0,9893	0,80	0,7881	1,80	0,9641	2,80	0,9974
0,31	0,6217	1,31	0,9049	2,31	0,9896	0,81	0,7910	1,81	0,9649	2,81	0,9975
0,32	0,6255	1,32	0,9066	2,32	0,9898	0,82	0,7939	1,82	0,9656	2,82	0,9976
0,33	0,6293	1,33	0,9082	2,33	0,9901	0,83	0,7967	1,83	0,9664	2,83	0,9977
0,34	0,6331	1,34	0,9099	2,34	0,9904	0,84	0,7995	1,84	0,9671	2,84	0,9977
0,35	0,6368	1,35	0,9115	2,35	0,9906	0,85	0,8023	1,85	0,9678	2,85	0,9978
0,36	0,6406	1,36	0,9131	2,36	0,9909	0,86	0,8051	1,86	0,9686	2,86	0,9979
0,37	0,6443	1,37	0,9147	2,37	0,9911	0,87	0,8078	1,87	0,9693	2,87	0,9979
0,38	0,6480	1,38	0,9162	2,38	0,9913	0,88	0,8106	1,88	0,9699	2,88	0,9980
0,39	0,6517	1,39	0,9177	2,39	0,9916	0,89	0,8133	1,89	0,9706	2,89	0,9981
0,40	0,6554	1,40	0,9192	2,40	0,9918	0,90	0,8159	1,90	0,9713	2,90	0,9981
0,41	0,6591	1,41	0,9207	2,41	0,9920	0,91	0,8186	1,91	0,9179	2,91	0,9982
0,42	0,6628	1,42	0,9222	2,42	0,9922	0,92	0,8212	1,92	0,9726	2,92	0,9982
0,43	0,6664	1,43	0,9236	2,43	0,9925	0,93	0,8238	1,93	0,9732	2,93	0,9983
0,44	0,6700	1,44	0,9251	2,44	0,9927	0,94	0,8264	1,94	0,9738	2,94	0,9984
0,45	0,6736	1,45	0,9265	2,45	0,9929	0,95	0,8298	1,95	0,9744	2,95	0,9984
0,46	0,6772	1,46	0,9279	2,46	0,9931	0,96	0,8315	1,96	0,9750	2,96	0,9985
0,47	0,6808	1,47	0,9292	2,47	0,9932	0,97	0,8340	1,97	0,9756	2,97	0,9985
0,48	0,6844	1,48	0,9306	2,48	0,9934	0,98	0,8365	1,98	0,9761	2,98	0,9986
0,49	0,6879	1,49	0,9319	2,49	0,9936	0,99	0,8389	1,99	0,9767	2,99	0,9986

Grandes valeurs absolues de t

t	$\phi(t)$
$4,5$	$0,999997$
$4,0$	$0,999968$
$3,8$	$0,999928$
$3,6$	$0,999841$
$3,5$	$0,99976$
$3,4$	$0,99966$
$3,3$	$0,99952$
$3,2$	$0,99931$
$3,1$	$0,99904$
$3,0$	$0,99865$

Annexe 2 : Table de K. PEARSON ou table de χ^2

Valeurs de χ^2 ayant une probabilité p d'être dépassée

Remarque : ν est le nombre de degré de liberté.

	0,990	0,975	0,950	0,900	0,100	0,050	0,025	0,010
1	0,0002	0,001	0,004	0,016	2,71	3,84	5,02	6,63
2	0,02	0,05	0,10	0,21	4,61	5,99	7,38	9,21
3	0,12	0,22	0,35	0,58	6,25	7,81	9,35	11,34
4	0,30	0,48	0,71	1,06	7,78	9,49	11,14	13,28
5	0,55	0,83	1,15	1,61	9,24	11,07	12,83	15,09
6	0,87	1,24	1,64	2,20	10,64	12,59	14,45	16,81
7	1,24	1,69	2,17	2,83	12,02	14,07	16,01	18,47
8	1,65	2,18	2,73	3,49	13,36	15,51	17,53	20,09
9	2,09	2,70	3,33	4,17	14,68	16,92	19,02	21,67
10	2,56	3,25	3,94	4,87	15,99	18,31	20,48	23,21
11	3,05	3,82	4,57	5,58	17,27	19,67	21,92	24,72
12	3,57	4,40	5,23	6,30	18,55	21,03	23,24	26,22
13	4,11	5,01	5,89	7,04	19,81	22,36	24,74	27,69
14	4,66	5,63	6,57	7,79	21,06	23,68	26,12	29,14
15	5,23	6,26	7,26	5,55	22,31	25,00	27,49	30,58
16	5,81	6,91	7,96	9,31	23,54	26,30	28,84	32,00
17	6,41	7,56	8,67	10,08	24,77	27,59	30,19	33,41
18	7,01	8,23	9,39	10,86	25,99	28,87	31,53	34,80
19	7,63	8,91	10,12	11,65	27,20	30,14	32,85	36,19
20	8,26	9,59	10,85	12,44	28,41	31,41	34,17	37,57
21	8,90	10,28	11,59	13,24	29,61	32,67	35,48	38,93
22	9,54	10,98	12,34	14,04	30,81	33,92	36,78	40,29
23	10,20	11,69	13,09	14,85	32,01	35,17	38,08	41,64
24	10,86	12,40	13,85	15,66	33,20	36,41	39,37	42,98
25	11,52	13,12	14,61	16,47	34,38	37,65	40,65	44,31
26	12,20	13,84	15,38	17,29	35,36	38,88	41,92	45,64
27	12,88	14,57	16,15	18,11	36,74	40,11	43,19	46,96
28	13,57	15,31	16,93	18,94	37,92	41,34	44,46	48,28
29	14,26	16,05	17,71	19,77	39,09	42,56	45,72	49,59
30	14,95	16,79	18,49	20,60	40,26	43,77	46,98	50,89

Annexe 3 : Table de distribution de la variable de FISHER-STUDENT T

Valeurs t **telles que** $p(|T_\nu > t) = p$

(ν représente le nombre de degrés de liberté)

	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.001
1	0.158	0.325	0.510	0.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63657	636.619
2	0.142	0.289	0.445	0.617	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	0.137	0.277	0.424	0.584	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.929
4	0.134	0.271	0.414	0.569	0.741	0.941	1.190	1.355	2.132	2.776	3.747	4.604	8.610
5	0.132	0.267	0.408	0.559	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.131	0.265	0.404	0.553	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.130	0.263	0.402	0.549	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.130	0.262	0.399	0.546	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.04
9	0.129	0.261	0.398	0.543	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.78
10	0.129	0.260	0.397	0.542	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.58
11	0.129	0.260	0.396	0.540	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.43
12	0.128	0.259	0.395	0.539	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.31
13	0.128	0.259	0.394	0.538	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.22
14	0.128	0.258	0.393	0.537	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.14
15	0.128	0.258	0.393	0.536	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.07
16	0.128	0.258	0.392	0.535	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.01
17	0.128	0.257	0.392	0.534	0.689	0.863	1.069	1.333	1.740	2.100	2.567	2.898	3.96
18	0.127	0.257	0.392	0.534	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.92
19	0.127	0.257	0.391	0.533	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.88
20	0.127	0.257	0.391	0.533	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.85
21	0.127	0.257	0.391	0.532	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.81
22	0.127	0.256	0.390	0.532	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.79
23	0.127	0.256	0.390	0.532	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.76
24	0.127	0.256	0.390	0.531	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.74
25	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.74
26	0.127	0.256	0.390	0.531	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.72
27	0.127	0.256	0.389	0.531	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.69
28	0.127	0.256	0.389	0.530	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.67
29	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.65
30	0.127	0.256	0.389	0.530	0.683	0.854	1.055	1.310	1.697	20.42	2.457	2.750	3.64
40	0.126	0.255	0.388	0.529	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.55
80	0.126	0.254	0.387	0.527	0.679	0.848	1.046	1.296	1.671	2.000	2.390	2.660	3.46
120	0.126	0.254	0.386	0.526	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.37
∞	0.126	0.253	0.385	0.524	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.29

Annexe 4 : Bibliographie

M. GAULTIER : *Analyse, probabilités et méthode statistique*, Tome 2, VUIBERT, ISBN : 2-7117-2179-5.

J. PELLAUMAIL : *Probabilités, statistiques, files d'attente*, DUNOD.

G. SAPORTA : *Probabilités, analyse de données et statistique*, TECHNIP, ISBN : 2-7108-0565-0.