QUANTERRA SHORT COURSE – 01F 24.01.2003

QUANTERRA

INTERNATIONAL INDEPENDENT CENTER OF CLIMATE CHANGE IMPACT ON NATURAL RISK ANALYSIS IN MOUNTAINOUS AREA

Short course

Considérations sur les droites de régression appliquées aux données XRD et lien avec les fonctions utilisées pour les profils de diffraction (analyse semi-quantitative et largeur de Scherrer)

Par M. Jaboyedoff

www.quanterra.org

Quanterra
Chemin de la Tour-Grise 28
1007-Lausanne
Tel. + 41 79 752 35 15
E-mail: mail@quanterra.org

Tables des matières

Considérations sur les droites de régression appliquées aux données XRD et lien avec les fonctions utilisées pour les profils de diffraction (analyse semi-quantitative et largeur de Scherrer)

Ce chapitre à pour but d'approfondir les fondements théoriques des régressions linéaires et l'erreur d'estimation qui les accompagne. D'autre part, lors de ces développements certaines fonctions gamma sont utilisées, il se trouve qu'elles sont souvent utilisées pour ajuster des pics de diffraction.

Les lignes qui suivent sont souvent directement inspirées d'ouvrages classiques. C'est pourquoi chaque titre est suivi des références concernant les paragraphes qui suivent. D'autre part, certains développements ou remarques sont faits en petits caractères car ils ne sont pas primordiaux pour la compréhension du texte. Les formules importantes sont signalées par (X).

XVI.1 Introduction [12]

Pour bien comprendre ce qu'est une erreur, il faut d'abord avoir en tête que nous n'avons généralement accès qu'à l'estimation de paramètres statistiques (essentiellement la moyenne et la variance), qui, à l'aide d'un modèle connu, à priori, permettent d'établir la confiance qu'on peut apporter aux résultats. On suppose donc toujours que plus le nombre "d'expériences" est grand, plus l'estimation de paramètres statistiques s'améliore. Nous allons voir qu'un estimateur ne prend pas nécessairement la même forme que son expression théorique.

> Il est important de tester la "viabilité" du modèle sur les données à tester.
>
> Par confiance on entend la probabilité qu'une mesure appartienne à l'intervalle défini par l'erreur.

On notera l'estimation d'un paramètre statistique **a** et noté **â**. Une variable aléatoire **X** dont on appellera les réalisations $\mathbf{x_1}$, $\mathbf{x_1}$,$\mathbf{x_2}$, ... $\mathbf{x_n}$. Si l'on effectue **N** tirages d'ensembles $\mathbf{x_i}$, on peut classer et former un nouvel ensemble de variables aléatoires $\mathbf{X_i}$, qui possèdent les mêmes propriétés que **X**. Donc on peut aussi estimer **a** à l'aide des $\mathbf{X_i}$ variables indépendantes, ainsi on comprend mieux que **â** est aussi une variable aléatoire.

> Par exemple si on effectue n tirages de boules dans une urne "infinie" contenant des boules noires et rouges, dont on calcule le pourcentage de boules rouges, est qu'on répète cette opération N fois on obtient une distribution des pourcentages dont la moyenne tend vers la proportion de boules rouges se trouvant dans l'urne.

On exige de l'estimateur qu'il soit consistant, c'est-à-dire que **â** tende vers **a** lorsque le nombre de réalisations augmentent et que l'espérance mathématique de **â** soit égale à **a**:

$E[\hat{a}] = a$

Si l'estimateur satisfait cette condition il est dit non-biaisé. De plus si la variance est minimale:

$E[(\hat{a}-a)^2]$ minimum

l'estimation est dite effective.

Formellement c'est la loi des grands nombres qui implique la convergence. En fait une variable aléatoire dont on attend que sa moyenne et sa variance existe, a un comportement tel que plus on connaît des valeurs de cette variable plus la moyenne arithmétique tend vers l'espérance mathématique de cette variable.

XVI.1.a Moyenne et variance [12]

Il est souvent important de connaître la variabilité d'une mesure. Pour se faire on répète plusieurs fois la même mesure afin d'en extraire la moyenne et la variance pour obtenir une valeur plus sûre et une incertitude expérimentale de mesure. Pour se faire une idée de l'estimation, nous allons calculer l'espérance mathématique et la variance d'une variable **X** connaissant n de ces réalisations $\mathbf{x_i}$.

Rappel:

$$\mathrm{Var}(X \pm Y) = E\left[\left((X+Y) \pm \left(\overline{X+Y}\right)\right)^2\right] = E\left[\left((X-\overline{X})^2 + (Y-\overline{Y})^2\right)\right] \pm 2E\left[(X-\overline{X})(Y-\overline{Y})\right]$$

$$= \mathrm{Var}(X) + \mathrm{var}(Y) + 2\mathrm{Cov}(X,Y)$$

L'espérance mathématique d'un paramètre et la valeur de ce paramètre a(X) si l'on connaissait le processus à décrire parfaitement. Elle s'écrit:

$E[a(X)] = \int_{-\infty}^{+\infty} a(X)\ f(X)\ dx$ ou $\int_{\text{espace accecible}} a(X)\ f(X)\ dx$ ou f(X) est la densité de probabilité de la variable X. Il est préférable d'utiliser cette notation continue car elle ne prête pas à confusion. L'espérance mathématique n'est connue que lorsqu'on connaît complètement le processus.

Il est naturel de prendre la moyenne arithmétique des valeurs observées comme estimateur de la moyenne:

$$\hat{m} = m_{arithm} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (1)$$

Il est facile de comprendre que plus **n** augmente plus on se rapprochera de l'espérance mathématique.

L'espérance mathématique de cette estimateur peut être vu comme la somme des espérances mathématiques de la variable x_i:

$$E[\hat{m}] = \frac{\sum_{i=1}^{n} E[x_i]}{n} = \frac{nE[X]}{n} = E[X] = \mu$$

de sorte que la moyenne arithmétique est non biaisée. La variance d'estimation est:

$$\mathrm{Var}(\hat{m}) = \mathrm{E}\left[(\hat{m}-\mu)^2\right] = \frac{1}{n}\sigma_x^2$$

La variance de l'estimation de la moyenne est déduite de:

$$\hat{m} = \frac{\Sigma x_i}{n} = \frac{\Sigma X_i}{n} \quad \text{et} \quad \mathrm{Var}(\hat{m}) = \frac{1}{n^2}\Sigma\,\mathrm{var}(X_i) = n\left(\frac{1}{n^2}\sigma_x^2\right) = \frac{1}{n}\sigma_x^2$$

où σ_x^2 est la variance théorique de la variable **X**. A l'aide de la moyenne et de la variance, en supposant que la distribution de la moyenne est gaussienne, ce qui est une approximation, on sait que par exemple celle ci à 68% de chance de se trouver dans l'intervalle $\hat{m} \pm \sigma_x / \sqrt{n}$.

XVI.1.b Estimation de la variance [12]

On appelle souvent la variance de l'échantillon:

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \hat{m})^2}{n}$$

s est consistant car

$s_x^2 = \frac{(\Sigma x_i^2)}{n} - \hat{m}^2$ converge vers σ_x. La démonstration de cette expression est la même que celle de la covariance ou en remplace les variables et espérances par les estimateurs.

or on peut montrer qu'il ne s'agit pas d'un estimateur non biaisé de la variance, en écrivant:

$$s_x^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \frac{\sum_{i=1}^{n} x_i^2}{n^2} - 2\frac{\sum_{i<j} x_i x_j}{n^2} = \frac{n-1}{n^2}\sum_{i=1}^{n} x_i^2 - 2\frac{\sum_{i<j} x_i x_j}{n^2}$$

dont on prend l'espérance mathématique. La variance par définition ne dépend pas de l'origine des coordonnées, donc on a le droit de se placer en la moyenne théorique de la population $\boldsymbol{\mu}$. De sorte que l'espérance de chaque x_i^2 est égale à la variance de la variable **X**. D'autre part les termes croisés prenne la forme de la covariance et sont nuls car les expériences sont indépendantes. Ainsi on obtient:

$$\mathrm{E}\left[s_x^2\right] = \frac{(n-1)n}{n^2}\sigma_x^2 = \frac{(n-1)}{n}\sigma_x^2$$

De sorte que $\mathbf{s_X}$ est biaisé, c'est pourquoi lorsque n est petit on utilise comme estimateur de la variance:

$$\hat{\sigma}_x^2 = \frac{\sum_{i=1}^{n}(x_i - \hat{m})^2}{n-1} \qquad (2)$$

Ainsi cet estimateur de la variance est non biaisé, consistant car quand **n** tend vers l'infini, **n/(n-1)** tend vers 1 et que les termes de la variance converge (voir première expression de s). Mais on peut sentir qu'il n'est pas effectif. De la même manière on obtient l'estimateur de la covariance:

$$\sigma_{xy} = \frac{\sum(x_i - \hat{m}_x)(y_i - \hat{m}_y)}{n-1} \qquad (3)$$

Comme pour l'estimateur de la variance on développe, on place l'origine aux espérance mathématique de x et y et on évalue l'espérance mathématique de tous les termes:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{m}_x)(y_i - \hat{m}_y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \frac{\sum x_j}{n})(y_i - \frac{\sum y_k}{n}) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i y_i - x_i\frac{\sum y_k}{n} - y_i\frac{\sum y_j}{n} + \frac{\sum x_j y_k}{n^2}\right)$$

$$\text{avec} \quad \text{cov}(x,y) = \sigma_{xy} = E\left[(X-\mu_X)(Y-\mu_Y)\right] = E\left[XY - \mu_X Y - \mu_Y X - \mu_x\mu_Y\right] = E\left[XY\right] - E\left[X\right]E\left[Y\right]$$

De sorte que si

$$E\left[X\right] = E\left[Y\right] = 0, \quad \sigma_{XY} = E\left[XY\right]$$

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{m}_x)(y_i - \hat{m}_y)\right] = \frac{1}{n}\sum_{i=1}^{n}\left(\sigma_{XY} - \frac{x_i y_i}{n} - \frac{x_i y_i}{n} - \frac{\sum x_i y_i}{n^2}\right)$$

$$= \frac{1}{n}\left(n\sigma_{XY} - \sigma_{XY} - \sigma_{XY} + \sigma_{XY}\right) = \frac{n-1}{n}\sigma_{XY}$$

On s'aperçoit que dans l'estimation du coefficient de corrélation le facteur **(n-1)** ne joue pas de rôle.

XVI.1.c Le coefficient de corrélation [12]

Ce coefficient noté **r** est fréquemment utilisé comme mesure de la qualité d'une droite de régression. En ces termes cette assertion est fausse comme nous le verrons. C'est la covariance qui caractérise le lien qu'il peut exister entre deux variables. Cependant pour pouvoir comparer un jeu de données à un autre, on norme la covariance par les écarts types des deux variables afin que ce coefficient soit égal à ±1 dans le cas d'une droite:

$$(y - \bar{y}) = b(x - \bar{x})$$

$$\text{et} \quad r = \frac{\sigma_{xy}}{\sqrt{\sigma_x^2\sigma_y^2}} = \frac{\sum\left((y_i - \bar{y})(x_i - \bar{x})\right)}{\sqrt{\sum(x_i - \bar{x})^2\sum(y_i - \bar{y})^2}} = \frac{b\sum(x_i - \bar{x})^2}{|b|\sum(x_i - \bar{x})^2} = \pm 1$$

Ce qui montre bien que pour tous jeux de données de ce type le résultat et équivalent. On va encore montrer que **r** varie de +1 à -1. Si on choisit une variable aléatoire **Z** tel que:

$Z = \left(\frac{X}{\sigma_x} \pm \frac{Y}{\sigma_y} \right)$ ou **X** et **Y** sont des variables centrées sur leur moyenne. Ainsi X et Y sont sur une droite, on choisit un signe tel que **Z** soit nul et ainsi sa variance est nulle. Par contre si X et Y sont indépendantes alors la variance est égale à 2.

$$\mathrm{var}(Z) = \frac{1}{\sigma_x^2}\mathrm{var}(X) + \frac{1}{\sigma_y^2}\mathrm{var}(Y) \pm \frac{2}{\sigma_x \sigma_y}\mathrm{cov}(X,Y) \geq 0$$

Une variance est toujours positive ou nulle. On obtient:

$$\pm \frac{\mathrm{cov}(X,Y)}{\sigma_x \sigma_y} \geq -1 \quad \text{donc} \quad |r| = \left| \frac{\mathrm{cov}(X,Y)}{\sigma_x \sigma_y} \right| \leq 1 \qquad (4)$$

Le coefficient **r** qualifie le lien qui existe entre deux variables, nul si aucun lien n'existe pas. C'est-à-dire que si la probabilité de $\mathbf{Y_i}$ connaissant $\mathbf{X_i}$ est contrainte par cette valeur ou l'inverse. Ce paramètre, comme la covariance, caractérise la qualité du lien de ces deux variables, et non pas la qualité de la droite de régression. En effet un coefficient de corrélation moyen peut être inhérent à des mesures et très peu varié avec le nombre de mesures. Par contre la fiabilité de la régression va augmenter.

XVI.2 Lois qui peuvent régir des données

XVI.2.a Ce qui justifie l'emploi de la loi normale: le théorème central limite [11].

On montre ici de façon simplifiée que lorsqu'une variable **Y** est la somme de **n** variables indépendantes **X** dont les moyennes et les variances sont définies, quand **n** tend vers l'infini la variable **Y** suit une loi normale:

$$P(x < a) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{u^2}{2\sigma^2}} du$$

Avant de démontrer l'assertion précédente il faut introduire ce qu'on appelle la fonction génératrice de moments. Il s'agit du développement de Taylor d'une fonction qui induit les différents moments centrés par rapport à l'origine.

La série de Taylor pour e^x quelque soit x est

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots$$

En effet si on prend l'espérance mathématique de $\mathbf{e^{tx}}$ alors on obtient:

$$E\left[e^{tx}\right] = E\left[1 + tx + \frac{t^2}{2!}x^2 + ...\right] = 1 + tE[x] + \frac{t^2}{2!}E\left[x^2\right] + ...$$

ou en notation continue $E\left[e^{tx}\right] = \int_{-\infty}^{+\infty} e^{tx} f(x)dx = f^*(t)$

où le développement autour de 0 de **f*(t)** comparé à celui de e^{tx} donne la valeur des différents moments centrés à l'origine.

> Il serait plus élégant d'utiliser la fonction caractéristique qui est la transformée de Fourier de la distribution., dont les propriétés d'inversion sont plus intéressantes que celles de la fonction génératrice de moments.

La fonction génératrice de moments de la somme de deux variables indépendantes est égale au produit de ses deux fonctions génératrices de moments. En effet:

$$E\left[e^{t(x+y)}\right] = E\left[e^{tx}e^{ty}\right] = E\left[e^{tx}\right]E\left[e^{ty}\right]$$

On pose les hypothèses suivantes:

$$Y = \sum_{i=1}^{n} X_i \quad E[X_i] = \mu \quad E[y] = n\mu \quad var(X_i) = \sigma^2$$

Il est préférable de choisir une variable centrée réduite pour Y de sorte que

$$Z = \frac{Y - n\mu}{\sigma\sqrt{n}}$$

La fonction génératrice des moments de la somme de variables indépendantes est leur produit de sorte que:

$E\left[e^{tZ}\right] = \left(E\left[e^{t\frac{X-\mu}{\sigma\sqrt{n}}}\right]\right)^n$ où X est l'une des variables Xi puisqu'elles ont toutes la même distribution. Si on effectue le développement de Taylor sur une variable on obtient, lorsque n est grand:

$$E\left[e^{t\frac{X-\mu}{\sigma\sqrt{n}}}\right] = E\left[1 + t\frac{X-\mu}{\sigma\sqrt{n}} + \frac{1}{2}t^2\frac{(X-\mu)^2}{\sigma^2 n} + ...\right] \approx 1 + \frac{1}{2n}t^2 + ...$$

On obtient le résultat, pour n variable $\mathbf{X_i}$ à l'aide de la formule du binôme:

$$\left\{ E\left[e^{t\frac{X-\mu}{\sigma\sqrt{n}}} \right] \right\}^n \approx \left(1+\frac{1}{2n}t^2\right)^n = 1+\frac{n!}{(n-1)!}\frac{t^2}{2n}+\frac{n!}{(n-2)!2!n^2}\left(\frac{t^2}{2}\right)^2+\ldots$$

$$=1+\frac{t^2}{2}+\frac{1}{2!}\left(\frac{t^2}{2}\right)^2+\ldots=e^{t^2/2}$$

qui est la fonction génératrice de moments d'une loi normale, sans montrer sont caractère unique, on obtient en effet:

$$f(t)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{tx}e^{-\frac{x^2}{2}}dx=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-\frac{1}{2}\left(x^2-2tx+t^2\right)}e^{\frac{t^2}{2}}dx$$

$$=\frac{1}{\sqrt{2\pi}}e^{\frac{t^2}{2}}\int_{-\infty}^{+\infty} e^{-\frac{1}{2}(x-t)^2}dx=\frac{1}{\sqrt{2\pi}}e^{\frac{t^2}{2}}\int_{-\infty}^{+\infty} e^{-\frac{1}{2}(u)^2}du=\frac{\sqrt{2\pi}}{\sqrt{2\pi}}e^{\frac{t^2}{2}}=e^{\frac{t^2}{2}}$$

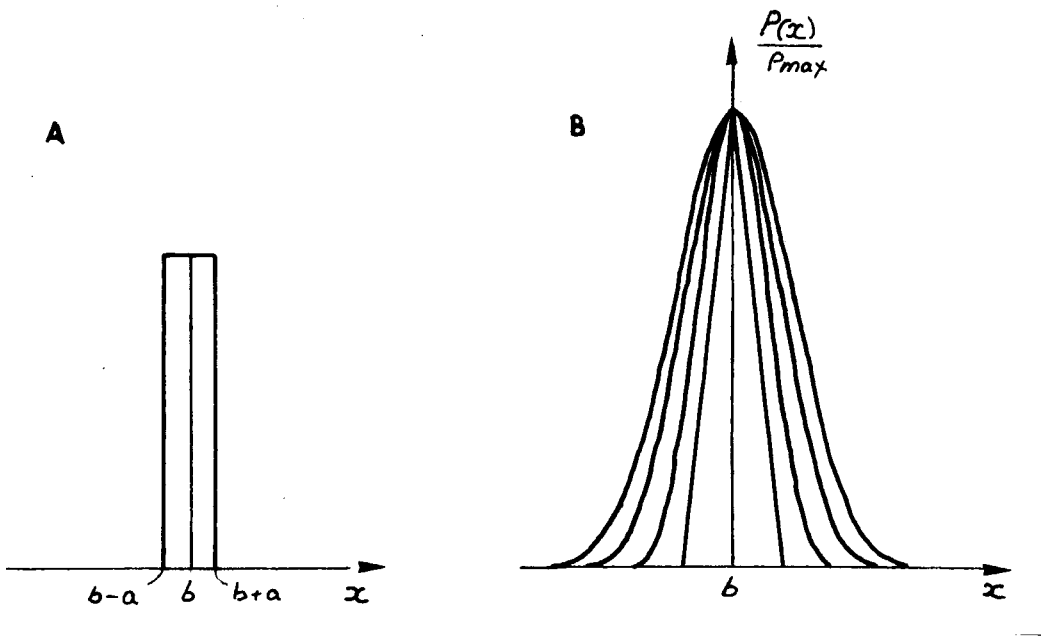

Figure 1:Iillustration du théorème central limite. A) un processus qui engendre des valeurs de type tout ou rien est une loi de distribution en forme de créneau. Une variable b ayant subi ce processus peut prendre indifféremment des valeurs entre b-a et b+a. B) distribution qu'on obtient pour une valeur d'origine ayant subi 2,5,10 et 15 fois le processus A). On voit qu'on tend très vite vers la loi de Gauss.

Ainsi on a montré l'intérêt de la loi normale. Comme il a été dit plus haut, ce théorème a une portée plus générale, c'est-à-dire que la somme de variables $\mathbf{X_i}$ indépendantes, voire faiblement liées lorsque leur moyenne et leur variance

existent, est une variable normale. Cette assertion justifie souvent l'utilisation de la loi normale comme distribution de variables qui sont le résultat d'une somme de phénomènes. Alors si:

$\overline{Y} = \sum_{i=1}^{n} \overline{X}_i$ et $\sigma_Y^2 = \sum_{i=1}^{n} \sigma_X^2$ existent **Y** tend vers une loi normale.

Lorsqu'une variable est la somme de cinq processus on peut déjà s'approcher de la loi normale dans la région de la moyenne (figure 1).

XVI.2.b La loi du χ2 [10, 11]

La distribution du $\chi 2$ est une distribution gamma dont la fonction de répartition est fonction gamma normée. Une variable aléatoire $\chi^2 = X_1^2 + X_2^2 + ... + X_n^2$ où les X_i sont des variables aléatoires indépendantes gaussiennes de moyenne nulle et de variance 1, suit une lois du χ^2:

$$P\left(\chi^2 \le x\right) = \frac{1}{2^{\upsilon/2}\Gamma(\upsilon/2)} \int_0^x t^{(\upsilon/2-1)} e^{-t/2} dt \qquad (6)$$

La fonction gamma représente, à un termes près, le calcul de factorielles pour les nombres réels positifs elle est définie comme: $\Gamma(p) = \int_0^\infty x^{p-1} e^{-x} dx$ avec $p > 0$

Ces propriétés essentielles sont [2]:

$$\Gamma(n) = \int_0^\infty x^{n-1} e^{-x} dx = (n-1)!,\ \Gamma(p+1) = p\Gamma(p), \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

Où **ν** est le nombre de degré de liberté, ici égal à **n** (le nombre de variable X). Dans le cas où $\mathbf{Y=X^2}$ et **X** une variable centrée réduite est normalement distribuée la probabilité que Y soit inférieur à y s'écrit:

$$P(Y \le y) = P(X^2 \le y) = P(-\sqrt{y} \le x \le +\sqrt{y})$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{y}}^{+\sqrt{y}} e^{-x^2/2} dx = \frac{2}{\sqrt{2\pi}} \int_0^{+\sqrt{y}} e^{-x^2/2} dx$$

Comme on veut retrouver la borne d'intégration y on effectue le changement de variable $x = \sqrt{u}$ de sorte que

$$dx = \frac{du}{2\sqrt{u}} \quad \text{et}$$

$$\frac{1}{\sqrt{2\pi}} \int_0^y u^{-1/2} e^{-u/2} du = \int_0^y \left[\frac{1}{2^{1/2}\Gamma(1/2)} u^{-1/2} e^{-u/2} \right] du$$

Ainsi on a montré que **Y** suit une loi du $\chi 2$. Il faut maintenant montrer que la somme de deux variables distribuées selon une loi du $\chi 2$ suit aussi une telle distribution.

Soit **X** et **Y** deux variables distribuées selon des lois du $\chi 2$. La probabilité qu'une variable **Z=X+Y** soit inférieure à **z** est donnée par:

$$F(z) = \frac{1}{2^{r/2}2^{s/2}\Gamma(r/2)\Gamma(s/2)} \iint_{x+y<z} e^{-x/2} x^{r/2-1} e^{-y/2} y^{s/2-1} dxdy$$

On élimine **y** de l'intégrale en posant que **y=x-z** ainsi **dx=dy** est les bornes d'intégration =0 x=z et pou y=z x=0.

$$F(z) = \frac{1}{2^{r+s/2}\Gamma(r/2)\Gamma(s/2)} \int_0^z e^{-x/2} x^{r/2-1} e^{-(z-x)/2} (z-x)^{s/2-1} dx$$

Afin de retrouver une fonction connue (fonction bêta) on pose **x=tz** ainsi pour x=0, t=0 et pour x=z, t=1:

$$F(z) = \frac{e^{-z/2} z^{\left((r+s)/2\right)-1}}{2^{(r+s)/2}\Gamma(r/2)\Gamma(s/2)} \int_0^1 t^{r/2-1} (1-t)^{s/2-1} dt$$

La fonction bêta est définie comme: $B(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1} dx$

avec p>0 et q>0. On peut montré que $B(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$

L'intégrale est égale à $\frac{\Gamma\left(r/2\right)\Gamma\left(s/2\right)}{\Gamma\left((r+s)/2\right)}$ de sorte que la distribution de **Z** est bien une distribution du $\chi 2$. En vertu de ce qui précède, toutes sommes du carré de variables réduites distribuées normalement possèdent une distribution du $\chi 2$. On observe que cette densité de probabilité présente un graphe asymétrique car les croissances et décroissances des deux termes qui composent cette fonction n'ont pas le même comportement. On peut calculer à l'aide des fonctions gamma que la moyenne d'une distribution du $\chi 2$ est égale à ν et sa variance $\sigma^2=\nu$.

L'intérêt d'une telle distribution est d'estimer la vraisemblance d'une distribution théorique ajustée à des valeurs expérimentales. On choisit une "distance" entre les valeurs de la distribution expérimentale et les valeurs théoriques comme mesure de l'écart entre les deux distributions on prend:

$$\sqrt{\frac{(n_i - Np_i)^2}{Np_i}}$$

où $\mathbf{n_i}$ et le nombre d'observations qui se trouvent dans un l'intervalle i (histogramme), **N** le nombre total d'observations et $\mathbf{p_i}$ la probabilité théorique

d'occurrence de cet intervalle. On considère que ces écarts sont distribués normalement (0,1). En fait on reconnaît les caractéristiques d'une loi de Poisson, la moyenne étant égale à la variance. Lorsque $\mathbf{Np_i}$ est grand, $\mathbf{p_i}$ petit et que les écarts à cette moyenne sont faibles, la distribution de Poisson tend vers la loi normale car dans ce cas alors $\left((x-\lambda)/\sqrt{\lambda}\right)$ apparaît bien comme une variable réduite. La somme de telles variables s'approche de la loi du $\chi 2$:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - Np_i)^2}{Np_i} \qquad (7)$$

Principe et exemple [12]

On possède 500 valeurs dont on connaît la distribution par intervalle (figure 2). On calcule la moyenne et la variance. et on suppose cette distribution normale. On calcule d'abord les $\mathbf{n_i}$, puis les fréquences relatives $\mathbf{n_i/N}$, ainsi la moyenne et la variance sont calculées facilement. Les valeurs théoriques sont simplement calculées pour la loi normale, par exemple pour la classes -2.5:

$$p_i = \frac{1}{\sigma\sqrt{2\pi}} \Delta_i \times e^{-\frac{(I_i - \hat{m})^2}{2\sigma^2}} \text{ par exemple } = \frac{1}{\sigma\sqrt{2\pi}} \times e^{-\frac{(-(-2.5)-0.168)^2}{2\times 2.098}} = 0.05$$

Centre de classe Ii =1	**n_i**	**n_i/N**	**I_i*n_i/N**	**$n_i*(I_i$-moy)2/N**	**Val théorique p_i**	**Fréq. théorique**	**$(n_i-np_i)^2/np_i$**
-3.5	6	0.012	-0.042	0.161	0.011	5.6	0.032
-2.5	25	0.05	-0.125	0.356	0.050	25.2	0.002
-1.5	72	0.144	-0.216	0.401	0.142	71.0	0.015
-0.5	133	0.266	-0.133	0.119	0.248	123.8	0.680
0.5	120	0.24	0.120	0.026	0.268	134.2	1.493
1.5	88	0.176	0.264	0.312	0.180	90.2	0.055
2.5	46	0.092	0.230	0.500	0.075	37.7	1.839
3.5	10	0.02	0.070	0.222	0.020	9.8	0.006
	Nb Total		Moyenne	Variance	Total poids	Tot.	$\chi 2$
	500	1	0.168	2.098	0.995	497.4	4.122

Tableau D1: calcul nécessaire au test du χ^2.

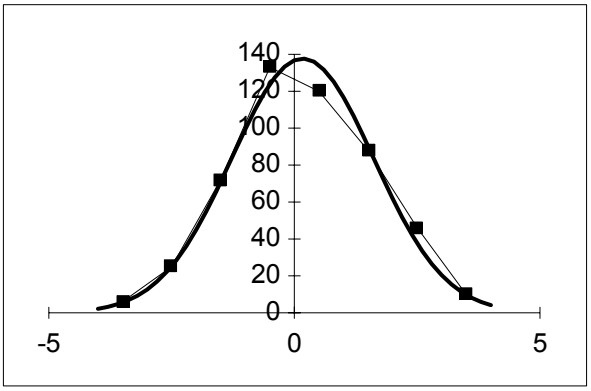

Figure 2: graphe de la distribution théorique et expérimentale (d'après [12]).

La somme $\chi 2$ et égale à 4.12. Pour un nombre de variables (intervalle) égale à 8, mais il y a trois contraintes, la moyenne, la variance, et le nombre total de points, ce qui donne 5 degrés de liberté. Selon les tables la probabilité d'occurrence d'une telle valeur est de plus de 50 % ce qui est considéré comme très bon, car on se trouve au centre de tous intervalles de confiances centrés. Cela n'infirme pas là une hypothèse de loi normale.

XVI.2.c La loi de Student [11]

On montre que des variables aléatoires indépendantes, l'une **Y** qui possède une distribution normale centrée réduite, et l'autre **Z** distribuée selon la loi du $\chi 2$ avec ν degrés de liberté, alors la variable aléatoire **t** définie comme

$t = \dfrac{Y}{\sqrt{Z/\nu}}$ suit une loi de Student à ν degrés de liberté.

Les distributions de **Y** et de **Z** sont données par

$$f(y) = \left(\frac{1}{\sqrt{2\pi}}\right) e^{-y^2/2} \quad \text{et} \quad f(z) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} z^{(\nu/2-1)} e^{-z/2}$$

Comme les variables **X** et **Y** sont indépendantes, la distribution jointe est le produit des deux. On cherche la fonction de répartition en effectuant un changement de borne d'intégration car **X** et **Y** sont liés via **t**. Ainsi

$$F(x) = P(t \le x) = P(Y \le x\sqrt{Z/\nu})$$

$$= \frac{1}{\sqrt{2\pi}2^{\nu/2}\Gamma(\nu/2)} \iint_{y \le x\sqrt{z/\nu}} z^{(\nu/2-1)} e^{-(y^2+z)/2} \, dydz$$

on fixe **z** et on intègre par rapport à **y** de $\left]-\infty, x\sqrt{z/\nu}\right]$ ainsi

$$F(x)=\frac{1}{\sqrt{2\pi}2^{\nu/2}\Gamma(\nu/2)}\int_{z=0}^{z=\infty} z^{(\nu/2-1)}e^{-z/2}\left[\int_{y=-\infty}^{y=x\sqrt{z/\nu}} e^{-y^2/2}\,dy\right]dz$$

Comme on est intéressé à retrouver x comme borne d'intégration, on va choisir une variable de telle sorte que $u=y/\sqrt{z/\nu}$ et ainsi la borne supérieure de u devient x, puis on inverse à nouveau l'ordre d'intégration

$$F(x)=\frac{1}{\sqrt{2\pi}2^{\nu/2}\Gamma(\nu/2)}\int_{z=0}^{z=\infty}\int_{u=-\infty}^{u=x} z^{(\nu/2-1)}e^{-z/2}\sqrt{z/\nu}\,e^{-u^2z/2\nu}\,du\,dz$$

$$F(x)=\frac{1}{\sqrt{2\pi\nu}2^{\nu/2}\Gamma(\nu/2)}\int_{u=-\infty}^{u=x}\left[\int_{z=0}^{\infty} z^{((\nu-1)/2)}e^{-z(1+u^2/\nu)/2}\,dz\right]du$$

Afin d'évaluer la valeur numérique de l'intégrale entre crochet, en l'occurrence on obtient une fonction gamma, on pose:

$$w=\frac{z}{2}\left(1+\frac{u^2}{\nu}\right)\quad\text{et}\quad dz=\frac{1}{2}\left(1+\frac{u^2}{\nu}\right)^{-1}dw$$

ainsi on obtient:

$$F(x)=\frac{1}{\sqrt{2\pi\nu}2^{\nu/2}\Gamma(\nu/2)}\int_{u=-\infty}^{u=x}\left[\int_{w=0}^{\infty}\frac{2^{((\nu-1)/2)-1}w^{((\nu-1)/2)}e^{-w}}{\left(1+u^2/\nu\right)^{((\nu-1)/2)+1}}dw\right]du$$

Ou on reconnaît la densité de probabilité de Student:

$$f(x)=\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma(\nu/2)}\frac{du}{\left(1+u^2/\nu\right)^{((\nu+1)/2)}}$$

$$F(x)=\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma(\nu/2)}\int_{u=-\infty}^{u=x}\frac{du}{\left(1+u^2/\nu\right)^{((\nu+1)/2)}}\qquad(8)$$

On remarque que cette loi est paire ce qui signifie que sa distribution est symétrique par rapport à 0. Sa moyenne vaut 0 et sa variance $\sigma^2=\nu/(\nu-2)$. On peut aussi démontrer que lorsque ν dépasse 30 alors la distribution de Student tend vers la loi normale réduite. Là encore on s'aperçoit que la loi normale peut être utilisée.

Lorsque n est grand la fonction gamma peut être remplacée par l'approximation de Stirling

$$\Gamma(n+1)=n!\approx n^n e^{-n}\sqrt{2\pi n}$$

A l'aide de cette formule on développe le terme constant de sorte que:

$$\frac{\sqrt{2}}{\sqrt{2\pi}}\frac{\sqrt{2\pi\left(\nu/2-1/2\right)}}{\sqrt{\nu}\sqrt{2\pi\left(\nu/2-1\right)}}\frac{\left(\nu/2-1/2\right)^{\left(\nu/2-1/2\right)}e^{-\left(\nu/2-1/2\right)}}{\left(\nu/2-1\right)^{\left(\nu/2-1\right)}e^{-\left(\nu/2-1\right)}}$$

$$\approx \frac{1}{\sqrt{2\pi}}\sqrt{\frac{\left(\nu/2-1/2\right)}{\left(\nu/2-1\right)}}\left(\frac{\left(\nu/2-1/2\right)}{\left(\nu/2-1\right)}\right)^{\left(\nu/2-1/2\right)}\frac{\sqrt{2}\left(\nu/2-1\right)^{1/2}}{\sqrt{\nu}}(1)\approx\frac{1}{\sqrt{2\pi}}(1)(1)(1)(1)=\frac{1}{\sqrt{2\pi}}$$

D'autre part, si on développe le logarithme de:

$$\ln\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2}=-\frac{(\nu+1)}{2}\ln\left(1+\frac{t^2}{\nu}\right)$$

Comme ν est grand la série de Taylor de ln(x) donne

$$-\frac{(\nu+1)}{2}\ln(1+x)=-\frac{(\nu+1)}{2}\left(x+\frac{x^2}{2}+\frac{x^3}{3}+\ldots\right)$$

$$=-\frac{(\nu+1)}{2}\left(\frac{t^2}{\nu}+\frac{t^4}{2\nu^2}+\frac{t^6}{3\nu^2}+\ldots\right)=-\left(\frac{t^2}{2}+\frac{t^2}{2\nu}+\frac{t^4}{4\nu}+\frac{t^4}{4\nu^2}+\ldots\right)\approx-\frac{t^2}{2}$$

On trouve le résultat en prenant l'exponentielle de ce dernier résultat et la limite du terme constant, on obtient: $\frac{1}{\sqrt{2\pi}}e^{-t^2/2}$

XVI.2.d Estimateur et intervalles de confiance [9, 10, 12]

Pour connaître l'intervalle dans lequel une valeur a un certain nombre de chances de se trouver, on fait appel à un modèle **à priori**. C'est-à-dire qu'on suppose un certain comportement, comme par exemple une loi normale. Un exemple situera le problème. On sait qu'un type de mesure se distribue de façon normale si on en effectue un grand nombre. Dans notre cas on connaît la variance, mais une seule mesure. Quelle est la probabilité **β** que l'espérance mathématique **μ** de cette mesure se trouve dans un intervalle donné **2ε** tel que:

$$P\left(\left|x-\mu\right|<\varepsilon\right)=\beta$$

En général on choisit **β** et on en déduit **ε** à l'aide de la loi choisie. Ainsi dans ce cas une probabilité de 95% équivaut à **μ±1.96σ,**. ou bien

$$P\left(\left|x-\mu\right|<1.96\sigma\right)=95\%\quad\text{d'où}\quad\frac{\left|x-\mu\right|}{\sigma}<1.96$$

et donc $\mu-1.96\sigma<x<1.96\sigma+\mu$ est l'intervalle dans lequel on a 95% de chance de trouver l'espérance mathématique de x. Dans ce cas on a choisi un intervalle centré sur la moyenne et donc symétrique. Mais il s'agit aussi de l'intervalle à plus forte densité de probabilité. Ce choix est donc légitime. Dans le cas général on choisit un intervalle de confiance de sorte que la probabilité **α** que **x** n'appartienne

pas à un intervalle soit répartie de façon symétrique, c'est-à-dire qu'on choisit un intervalle tel que **(1-β)/2=α/2**. Donc comme borne **α/2** et **1-α/2**.

$$F(a) = \int_{-\infty}^{x} f(u)du \quad \text{et} \quad \int_{-\infty}^{+\infty} f(u)du = 1 \quad \text{où f(u) est la distribution de x.}$$

Plus formellement si **F(x)** est la fonction de répartition d'une variable, alors l'incertitude **[a,b]** est telle que **F(a)=α/2 et 1-F(b)=α/2**.

Il est fréquent d'utiliser la loi normale, en vertu du théorème central limite. En effet si la variance est due à une somme d'effet indépendant alors on se trouve proche de la loi normale. Cependant fréquemment on suppose que certaines variables auxiliaires suivent la loi du $\chi 2$ ou la loi de Student.

On va analyser les propriétés de l'estimation de la moyenne. Notons que

$$\sum_{i=1}^{n}(X_i-\mu)^2 = \sum_{i=1}^{n}\left((X_i-\overline{X})+(\overline{X}-\mu)\right)^2$$

$$= \sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 + \sum_{i=1}^{n}\left(\overline{X}-\mu\right)^2 + 2\left(\overline{X}-\mu\right)\sum_{i=1}^{n}\left(X_i-\overline{X}\right)$$

$$= \sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 + n\left(\overline{X}-\mu\right)^2$$

En divisant ce dernier résultat par σ^2 on obtient:

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \frac{ns^2}{\sigma^2} + \left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2 \quad \text{où} \quad s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i-\hat{m}\right)^2$$

En supposant que la distribution de $\mathbf{X_i}$ soit gaussienne et par conséquent la distribution de $\overline{X}$ l'est aussi puisqu'il s'agit d'une somme de variables normales. Le terme de gauche est une variable du $\chi 2$ de degré **n** puisqu'il s'agit de **n** variables centrées réduites. D'autre part le second terme du membre de gauche est une variable normale centrée réduite. On en déduit que le premier terme de gauche est une variable du $\chi 2$ à **n-1** degrés de liberté. On peut former une variable de Student à cause des propriétés énoncées plus haut:

$$T = \frac{\frac{\overline{X}-\mu}{\sigma}\sqrt{n}}{\sqrt{\frac{ns^2}{(n-1)\sigma^2}}} = \frac{\overline{X}-\mu}{s}\sqrt{n-1}$$

Ainsi de cette expression on peut tirer l'intervalle de confiance de la moyenne. En effet on sait que:

$$-t_{\alpha/2} < T < t_{\alpha/2}$$

ainsi .....(9)

$$\overline{X} - t_{n-1}\frac{s}{\sqrt{n-1}} < \mu < \overline{X} + t_{n-1}\frac{s}{\sqrt{n-1}}$$

D'autre part l'estimation de l'incertitude sur la variance se déduit sachant que

$\frac{ns^2}{\sigma^2}$ suit une loi du $\chi 2$ à **n-1** degré de liberté.

De sorte que l'intervalle de confiance se construit comme suit

$$P\left(a < \frac{ns^2}{\sigma^2} < b\right) = 1-\alpha$$

$$P\left(\frac{ns^2}{\sigma^2} < b\right) = 1-\alpha/2 \Rightarrow \sigma^2 > \frac{ns^2}{b} \text{.....(10)}$$

$$P\left(a < \frac{ns^2}{\sigma^2}\right) = \alpha/2 \Rightarrow \sigma^2 < \frac{ns^2}{a}$$

On a ainsi déterminer les intervalles de confiance de la moyenne et de la variance de variables normales. On ne doit pas perdre de vue cette hypothèse.

XVI.3 La régression linéaire (standardisation) [1,5]

Lors de procédures d'étalonnage de mesure, il est fréquent de corréler linéairement des mesures. Généralement on connaît l'une des mesures (x) mieux que l'autre (y), de sorte qu'on écrit:

$$y_i = a + bx_i + \Delta_i$$

où **a** et **b** sont les paramètres de la droite et Δ_i les écarts à cette droite dus aux erreurs (figure 3). On choisit généralement la méthode des moindres carrés qui consiste à minimiser la distance:

$$\delta = \sum \Delta_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

En fait on peut supposer que les Δ_i possèdent une distribution gaussienne telle que la probabilité que l'occurrence de y_i $e^{-\sum (y_i - a - bx_i)^2 / 2\sigma^2}$ soit maximum, ce qui implique la minimisation de la somme des écarts à la droite.

On cherche alors un minimum pour les deux variables **a** et **b**

$$\frac{\partial \delta}{\partial a} = -2\sum_{i=1}^{n}\left(y_i - a - bx_i\right) = 0$$

$$\frac{\partial \delta}{\partial b} = -2\sum_{i=1}^{n}\left(x_i y_i - a x_i - b x_i^2\right) = 0$$

d'où on tire:

$$b = \frac{n\sum x_i y_i - \sum x_i \sum x_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)} = \frac{\sigma_{xy}}{\sigma_x^2} \quad \text{et} \quad \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2} = \hat{b} \qquad (11)$$

$$a = \frac{\sum y_i}{n} - b\frac{\sum x_i}{n} \qquad (12)$$

C'est la forme habituelle de la droite de régression. On voit dans l'expression de **a** que la droite passe par les moyennes de **x** et **y**.

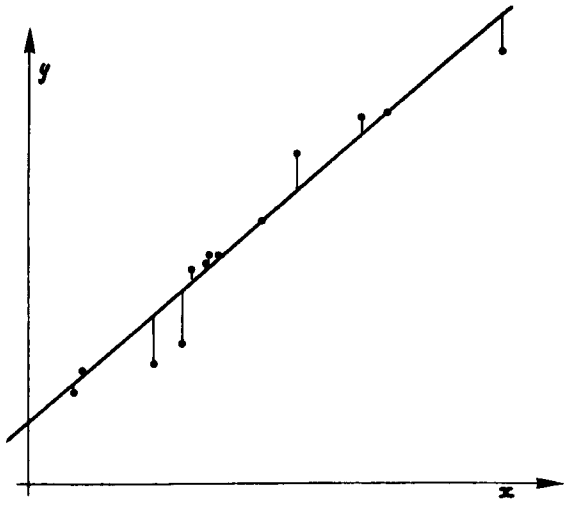

Figure 3:Iillustration de la distance minimisée lors d'une régression sur y, sachant que l'incertitude sur x est très inférieure à celle sur y.

Pour faciliter les calculs d'erreurs il est préférable d'utiliser la moyenne de x comme origine de sorte que:

$$\xi = x - \hat{m}_x \quad \text{où} \quad \hat{m} = \frac{\sum x_i}{n} \quad \text{ainsi} \quad \sum \xi_i = 0$$

et on notera l'estimation de **y** en fonction de ξ:

$$u(\xi) = a_1 + b_1 \xi$$

La pente $\mathbf{b_1}$ reste inchangée (b) car la variance et la covariance sont invariantes. D'autre part $\mathbf{a_1}$ est égal à la moyenne des **y** car la droite passe par les moyennes et l'origine et justement la moyenne des **x**. On obtient:

$$b_1 = \frac{\sum \xi_i y_i}{\sum \xi^2} \quad \text{et} \quad a_1 = \frac{\sum y_i}{n}$$

On peut montrer que $\mathbf{a_1}$ et $\mathbf{b_1}$ sont non biaisés. On se contentera de remarquer qu'on ne peut pas faire mieux pour $\mathbf{b_1}$ car on utilise les estimateurs non biaisés de la variance et de la covariance, les termes (n-1) disparaissant. Il en va de même pour $\mathbf{a_1}$ qui est égal à l'estimateur non biaisé de la moyenne de $\mathbf{y_i}$.

XVI.3.a L'estimation des erreurs sur une droite de régression

Pour estimer les erreurs le long d'une droite de régression on travaille sur les résidus $v_i = y_i - u(\xi_i)$. En fait on a fait l'hypothèse qu'ils possèdent une distribution gaussienne. Cela implique qu'un grand nombre d'expériences successives pour une valeur $\mathbf{x_i}$ donnée s'approcherait d'une distribution gaussienne de variance σ. Ceci est vrai pour n'importe quel $\mathbf{x_i}$. C'est pourquoi lors d'une calibration il est préférable de faire un histogramme des résidus et d'effectuer un test du $\chi 2$ pour vérifier la vraisemblance d'une gaussienne.

Remarques

Il est important d'étudier la forme de l'histogramme et surtout sa symétrie même si le test n'est pas très concluant, car l'hypothèse de symétrie est fondamentale. En effet dans le cas qui nous préoccupe, si la distribution des écarts à une droite de régression présente une asymétrie, elle a de forte chance de signifier la non-linéarité de la relation entre les points x, y, ou une erreur systématique. Il toujours possible alors de restreindre la zone de régression.

XVI.3.b Les erreurs sur une droite de régression standard [5,9]

Lorsqu'on effectue une régression, non seulement la distribution des écarts doit être symétrique, mais en plus il est préférable d'avoir une distribution de points assez homogène. Il est en effet peu recommandé de se caler sur des points esseulés.

Le but de ce paragraphe est d'estimer les écarts-types le long d'une droite de régression. Tout d'abord il faut trouver un estimateur de la variance des résidus v_i:

$$v_i = y_i - u(\xi_i) = y_i - a_1 - b_1\xi_i$$

et $\quad E[v_i] = 0$

$$E[v_i^2] = \mathrm{var}(v_i) = \mathrm{var}(y_i) + \mathrm{var}(a_1) + \xi_i^2\,\mathrm{var}(b_1) - 2\,\mathrm{cov}(y_i, a_1) - 2\,\mathrm{cov}(y_i, b_1) + 2\,\mathrm{cov}(a_1, b_1)$$

on va estimer terme par terme cette expression. On pose que $\mathbf{var(y_i)=\sigma^2}$. D'autre part $\mathbf{Y_i}$, $\mathbf{A_1}$ et $\mathbf{B_1}$ sont les valeurs exactes mais non connues de la droite recherchée

$$\mathrm{var}(b_1) = \mathrm{var}\left(\frac{\sum \xi_i y_i}{\sum \xi_i^2}\right) = \frac{\sum \xi_i^2\,\mathrm{var}(y_i)}{\left(\sum \xi_i^2\right)^2} = \frac{\sigma^2}{\sum \xi_i^2}$$

$$\mathrm{var}(a_1) = \mathrm{var}\left(\frac{\sum y_i}{n}\right) = \frac{\sum \mathrm{var}(y_i)}{n^2} = \frac{\sigma^2}{n}$$

$$\mathrm{cov}(y_i, b_1) = E[(y_i - Y_i)(a_1 - A_1)] = E\left[(y_i - Y_i)\sum (y_j - Y_j)/n\right]$$

$$= \frac{1}{n}\mathrm{var}(y_i) + \frac{1}{n}E\left[\sum_{j\neq i}(y_i - Y_i)(y_j - Y_j)/n\right] = \frac{\sigma^2}{n} = \mathrm{var}(a_1)$$

$$\mathrm{cov}(a_1, b_1) = E[(b_1 - B_1)(a_1 - A_1)]$$

$$= \frac{\sum (y_i - Y_i)\sum \xi_j (y_j - Y_j)}{n\sum \xi_k^2}$$

L'espérance mathématique des termes où les i=j sont nuls, car indépendants, donc

$$\mathrm{cov}(a_1, b_1) = \frac{\sum \xi_i\,\mathrm{var}(y_i)}{n\sum \xi_k^2} = \frac{\mathrm{var}(y_i)\sum \xi_i}{n\sum \xi_k^2} = 0$$

$$\mathrm{cov}(y_i, b_1) = E\left[(y_i - Y_i)\frac{\sum \xi_j (y_j - Y_j)}{\sum \xi_k^2}\right]$$

$$= \frac{\xi_i\,\mathrm{var}(y_i)}{\sum \xi_k^2} = \xi_i\,\mathrm{var}(b_1)$$

On peut maintenant réunir tous les termes et on obtient:

$$E[v_i^2] = \mathrm{var}(y_i) + \mathrm{var}(a_1) + \xi_i^2\,\mathrm{var}(b_1) - 2\,\mathrm{var}(a_1) - 2\xi_i^2\,\mathrm{var}(b_1)$$

$$= \mathrm{var}(y_i) - \mathrm{var}(a_1) - \xi_i^2\,\mathrm{var}(b_1)$$

$$= \sigma^2\left\{1 - \frac{1}{n} - \frac{\xi_i^2}{\sum \xi_j^2}\right\}$$

Si on prend l'espérance mathématique de la somme des $\mathbf{v_i}$, on obtient:

$$E\left[\sum v_i^2\right] = \sigma^2 (n-2)$$

Ainsi l'estimateur de la variance des écarts résiduels est

$$\hat{\sigma}^2_{v_i} = \frac{\sum v_i^2}{n-2} \qquad (13)$$

Le facteur **n-2** provient du fait que la droite est contrainte par deux paramètres. Maintenant connaissant $\hat{\sigma}^2_{v_i}$ on peut calculer la variance d'estimation le long de la droite de régression. On rappelle qu'on a supposé que l'erreur n'entachait que la variable **y**. On peut poser le problème comme suit. Quelle est la variance d'un point qu'on observe si on connaît **x** (ou ξ), sur la base des mesures qui ont généré la droite de régression. On peut écrire:

$$y_{attendu} = y_{att} = a_1 + b_1 \xi + v$$

où v est l'écart aléatoire à la droite normalement distribué, de moyenne nulle et de variance $\hat{\sigma}^2_{vi}$. On va donc évaluer la variance de cette expression:

$$var(y_{att}) = var(a_1 + b_1 \xi + v)$$

$$= var(a_1) + \xi^2 \, var(b_1) + var(v) + 2\xi \, cov(a_1, b_1) + 2\, cov(a_1, v) + 2\xi \, cov(v, b_1)$$

L'estimation de chaque terme donne:

$$var(v) = \sigma_v^2$$

$$var(a_1) = var\left(\frac{\sum y_i}{n}\right) = \frac{1}{n^2}\left(var(y_1) + ... + var(y_n)\right) = \frac{\sigma_v^2}{n}$$

$$var(b_1) = var\left(\frac{\sum \xi_i y_i}{\sum \xi_j^2}\right) = \frac{1}{\left(\sum \xi_j^2\right)^2} \sum \xi_i^2 \, var(y_i) = \frac{\sigma_v^2}{\sum \xi_j^2}$$

Les covariances contenant **v** sont nulles car ils sont indépendants de $\mathbf{b_1}$ et $\mathbf{a_1}$. D'autre part comme on l'a montré plus haut la covariance de $\mathbf{b_1}$ et $\mathbf{a_1}$ est nulle. Ainsi la variance d'une valeur **y** à estimer est:

$$var(y_{att/\xi}) = \sigma^2_{att/\xi} = \sigma_v^2 \left\{1 + \frac{1}{n} + \frac{\xi^2}{\sum \xi_i^2}\right\}$$

Il est intéressant de constater que cette méthode calcule l'erreur due aux paramètres de la droite, mais on y ajoute l'erreur inhérente au phénomène étudié ou à la mesure elle-même. **Pour une seule mesure on ne peut en effet pas attendre mieux que ce que l'on observe.**

Lors du traitement on a supposé que l'écart $\left(y_{att/\xi} - Y_\xi\right)$ entre **y** attendu et la valeur exacte **Y** pour ξ donné, possède une distribution gaussienne de variance $\sigma^2_{att/\xi}$ centrée en $\mathbf{Y_{exacte}}$. La meilleur estimation de la valeur exacte de **Y** est

celle de la droite de régression. D'autre part l'estimateur $\hat{\sigma}^2_{att}$ de $\sigma^2_{att/\xi}$ possède une distribution du **χ2** à **(n-2)** degrés de liberté car $\hat{\sigma}^2_{att}$ est une homotetie de $\hat{\sigma}^2_{v}$ qui est une somme de **(n-2)** termes "gaussiens" indépendants. Ainsi comme le montre le rapport d'une variable gaussienne et de la racine d'une variable distribuée selon un $\chi 2$ suit une lois de Student (ceci est vrai car les variances de chaque terme sont équivalentes.)

$$\frac{(y_{att/\xi} - Y_{exacte/\xi})}{\sqrt{\hat{\sigma}^2_{att/\xi}}} = t_{n-2}$$

Ainsi les intervalles de confiance autour d'une droite de régression (figure 4), en remplaçant ξ par sa valeur sont donnés par:

$$\Delta_{\%} y_{att/x} = \pm t_{\%} \sqrt{\left(\frac{\sum v_i^2}{n-2}\right)} \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum (x_i - \bar{x})^2}} \qquad (14)$$

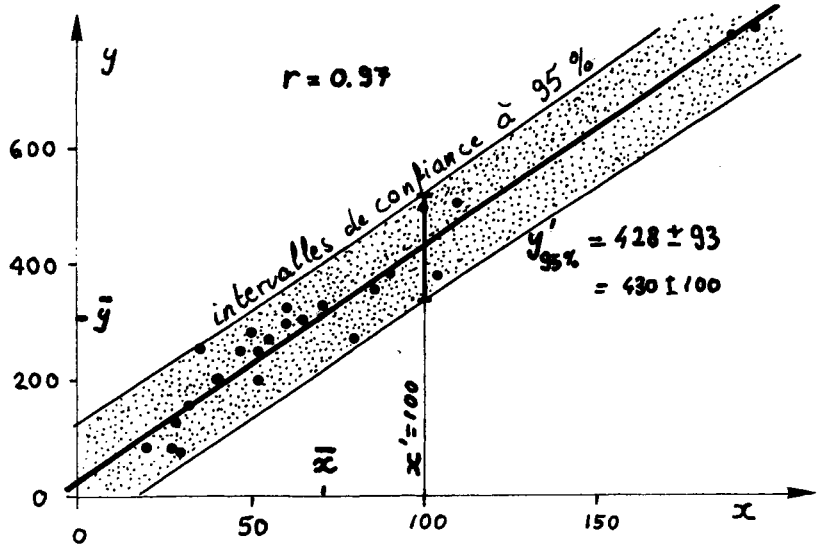

Figure 4: Exemple tiré de [10]. Positions des intervalles de confiances (95%) sur une droite de régression pour l'estimation d'une valeur unique (14). On note que deux points sont éloignés des autres, il faut être conscient qu'ils ont un grand poids. Le coefficient de corrélation par exemple s'en trouve amélioré. Mais dans ce cas, il y a lieu de penser que ces deux points sont assez sûrs. En effet sur 24 points aucun ne sort de l'intervalle de confiance alors qu'on pourrait s'attendre à ce qu'au moins un (4%) se trouve hors de cet intervalle à 95%. Ce type de remarques aide à valider une régression. On observe aussi que pour une valeurs x=100 l'estimation est grande. Les limites des intervalles de confiance sont des arcs d'hyperboles.

On constate que cet intervalle dépend de l'erreur des deux paramètres de la droite et de l'erreur inhérente à la mesure. Cette intervalle diminue avec le nombre de mesures **n**, et dépend essentiellement des écarts quadratiques résiduels. Lorsqu'on cherche l'estimation d'une moyenne on peut supprimer le terme de l'erreur inhérente à la mesure. Ainsi on obtient:

$$\Delta_{\%} y_{\text{moyen.att}/x} = \pm t_{\%} \sqrt{\left(\frac{\sum v_i^2}{n-2}\right)} \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum (x_i - \bar{x})^2}} \qquad (14 \text{ bis})$$

XVI.4 Droite de régression passant par l'origine [5].

Dans le cas d'une régression contrainte à l'origine il n'y a plus qu'un paramètre à estimer la pente **b**. On procède comme pour le cas général. On cherche le minimum de la distance à la droite:

$$\frac{\partial}{\partial b_1} \sum (y_i - b_1 x_i)^2 = \sum 2 b_1 x_i^2 - 2 \sum y_i x_i = 0$$

$$\Rightarrow b_1 = \frac{\sum y_i x_i}{\sum x_j^2}$$

La variance de $\mathbf{b_1}$ se déduit de la même manière que précédemment:

$$\text{var}(b_1) = \frac{\sigma_v^2}{\sum x_j^2}$$

Quant à l'estimation des erreurs sur cette droite on peut faire l'analogie suivante avec la droite de régression normale. On double le nombre de points sur lesquels on effectue la régression en ajoutant aux points $\{x_i, y_i\}$ les points $\{-x_i, -y_i\}$. Ainsi les moyennes de **x** et **y** sont nulles, d'autre part $\mathbf{a_1 = 0}$ et $\mathbf{\text{var}(a_1) = 0}$, par hypothèse. Dans ce cas l'estimation des résidus est

$$\hat{\sigma}_v^2 = \frac{\sum_{i=1}^{i=2n} v_i^2}{(2n-2)} = \frac{\sum_{i=1}^{i=n} (-v_i)^2 + \sum_{i=1}^{i=n} v_i^2}{2(n-1)} = \frac{\sum_{i=1}^{i=n} v_i^2}{(n-1)}$$

car les résidus apparaissent tous à double. Comme il n'y a qu'une contrainte le numérateur est **(n-1)**.

Ainsi l'estimation des intervalles de confiance devient

$$\Delta_{\%} y_{\text{att}/x} = \pm t_{\%} \sqrt{\left(\frac{\sum v_i^2}{n-1}\right)} \sqrt{1 + \frac{x^2}{\sum x_i^2}} \qquad (15)$$

Comme pour le cas précédent si on cherche une valeur moyenne la formule (15) devient:

$$\Delta_{\%} y_{\text{moyen.att}/x} = \pm t_{\%} \sqrt{\left(\frac{\sum v_i^2}{n-1}\right)} \sqrt{\frac{x^2}{\sum {x_i}^2}} \qquad (15\text{bis})$$

XVI.5 Droite de régression prenant en compte des erreurs qui apparaissent sur les deux axes

XVI.5.a Reduced major axis (RMA) [4, 6, 7]

Lorsque des incertitudes apparaissent sur les deux directions **x** et **y**, on peut chercher à rendre équivalent les deux directions d'un point de vue statistique (figure 5) de sorte que:

$$\frac{y-\bar{y}}{\sigma_y} = \frac{x-\bar{x}}{\sigma_x}$$

et comme la moyenne de points alignés sur une droite doit s'y trouver, on écrit:

$$y = \frac{\sigma_y}{\sigma_x} x + \left(\bar{y} - \frac{\sigma_y}{\sigma_x}\bar{x}\right) \qquad (15)$$

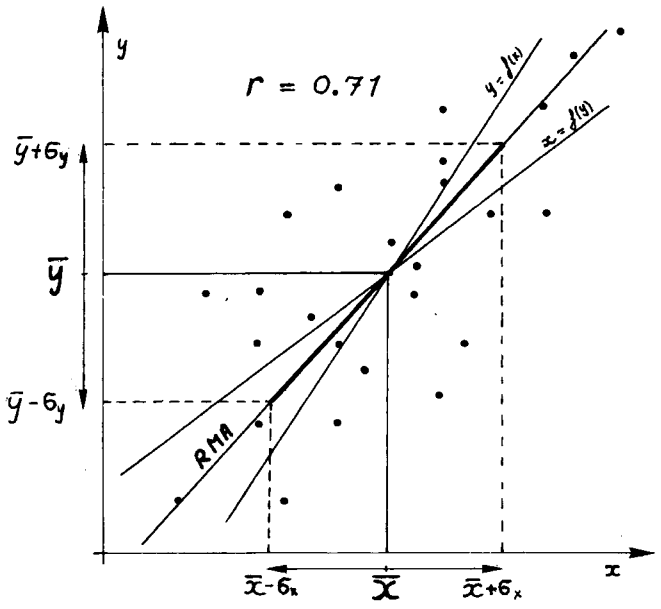

Figure 5: illustration de la droite RMA (d'après [4]). La pente est égale au rapport des écarts types, et la droite passe par les moyenne. Dans le cas d'une corrélation moyenne (r=0.71) la droite RMA est un meilleur compromis que les deux droites de moindres carrés classiques, car elle se trouve dans une position médiane.

Ce résultat est invariant si on change d'échelle. On peut aussi y parvenir en minimisant l'aire qui se trouve entre un point et la droite. Son intérêt est aussi la relative simplicité de l'expression de l'erreur d'estimation.

On peut estimer la variance d'une fonction à plusieurs variables h(x) ou $x=\{x_1,x_2,....,x_n\}$ avec

$$h(x)_a \approx h(a) + \sum_{i=1}^{n} h(a)'_i\,(x_i - a_i) + \ldots \text{ et}$$

$$\operatorname{var}\left(h(x)_a\right) \approx E\left[\left(\sum_{i=1}^{n} h'_i\,(a)(x_i - a_i)\right)^2\right] = \sum_{i=1}^{n} h'_i\,(a)\operatorname{var}(x_i) + 2\sum_{i=2}^{n}\sum_{j=1}^{n-1} h'_i\,(a)h'_j\,(a)\operatorname{cov}(x_i, x_j)$$

Ainsi pour une variable, on obtient: $\operatorname{var}\left(h(x)\right) = \left(\frac{dh}{dx}\right)_a^2 \operatorname{var}(x)$

On évalue d'abord la différentielle du rapport des deux variances estimées à l'aide des moments centrés (sur la moyenne) d'ordre 2 notés $\mathbf{m_{ij}}$ où **ij** sont les ordres des variables **X** et **Y.** μ_{ij} est l'espérance de ces moments:

$$\delta\left(\frac{\sigma_y^2}{\sigma_x^2}\right) \approx \delta\left(\frac{m_{02}}{m_{20}}\right) = \frac{\delta m_{02}\mu_{20} - \delta m_{20}\mu_{02}}{\mu_{20}^2} = \left(\frac{\mu_{02}}{\mu_{20}}\right)\left(\frac{\delta m_{02}}{\mu_{02}} - \frac{\delta m_{20}}{\mu_{20}}\right)$$

En prenant l'espérance mathématique du carré de cette expression on obtient:

$$\operatorname{var}\left(\frac{m_{02}}{m_{02}}\right) = \left(\frac{\mu_{02}^2}{n\mu_{20}^2}\right)\left(\frac{\mu_{04}}{\mu_{02}^2} + \frac{\mu_{40}}{\mu_{20}^2} - \frac{2\mu_{22}}{\mu_{20}\mu_{02}}\right)$$

Ainsi par une dernière transformation on obtient la variance du rapport des écarts types, en utilisant à nouveau la formule approchée de la variance:

$$\operatorname{var}\left(\sqrt{\frac{\sigma_y^2}{\sigma_x^2}}\right) \approx \frac{1}{4}\frac{\sigma_x^2}{\sigma_y^2}\operatorname{var}\left(\frac{\sigma_y^2}{\sigma_x^2}\right) = \frac{1}{4}\frac{\mu_{20}}{\mu_{02}}\operatorname{var}\left(\frac{m_{02}}{m_{20}}\right)$$

$$= \left(\frac{\mu_{02}}{4n\mu_{20}}\right)\left(\frac{\mu_{04}}{\mu_{02}^2} + \frac{\mu_{40}}{\mu_{20}^2} - \frac{2\mu_{22}}{\mu_{20}\mu_{02}}\right)$$

Si on suppose que les couples **XY** suivent une loi normale bivariée alors:

$$\mu_{20} = \sigma_x^2,\ \ \mu_{40} = 3\sigma_x^4,\ \ \mu_{22} = \sigma_x^2\sigma_y^2\left(1 + 2r^2\right)$$

$$\mu_{02} = \sigma_y^2,\ \ \mu_{04} = 3\sigma_y^4$$

où **r** est le coefficient de corrélation, et ainsi on obtient:

$$\operatorname{var}(b) = \operatorname{var}\left(\frac{\sigma_y}{\sigma_x}\right) = \frac{\sigma_y^2}{\sigma_x^2}\frac{(1 - r^2)}{n} \qquad (17)$$

On calcule la variance du terme constant en différenciant:

$$\delta\left(\bar{y}-\bar{x}(\sigma_y/\sigma_x)\right)=\delta\bar{y}-\delta\bar{x}(\sigma_y/\sigma_x)-\mu_x\delta(\sigma_y/\sigma_x)$$

$$\begin{aligned}\mathrm{var}(a)=\mathrm{var}\left(\bar{y}-\bar{x}(\sigma_y/\sigma_x)\right)&=\left(\delta\bar{y}\right)^2+\left(\delta\bar{x}(\sigma_y/\sigma_x)\right)^2+\left(\mu_{10}\delta(\sigma_y/\sigma_x)\right)^2\\&-2\delta\bar{y}\delta\bar{x}(\sigma_y/\sigma_x)-2\delta\bar{y}\mu_x\delta(\sigma_y/\sigma_x)+2\delta\bar{x}(\sigma_y/\sigma_x)\mu_x\delta(\sigma_y/\sigma_x)\end{aligned}\qquad(18)$$

$$=\frac{\sigma_y^2}{n}+\frac{\sigma_x^2}{n}\frac{\sigma_y^2}{\sigma_x^2}+\mu_x^2\,\mathrm{var}\left(\frac{\sigma_y}{\sigma_x}\right)-2\,\mathrm{cov}(x,y)\frac{\sigma_y}{\sigma_x}$$

$$=\frac{\sigma_y^2}{n}\left(2-2r+\frac{\mu_x^2(1-r^2)}{\sigma_x^2}\right)$$

On utilise ici les estimateurs de tous les moments. Il s'agit maintenant d'estimer la variance le long de la droite RMA. On ne trouve pas dans la littérature d'expression pour cette erreur, cependant on peut la calculer par analogie avec la droite de régression standard. Dans le cas présent, on connaît les variances des 2 paramètres de la droite pour n'importe quelle coordonnée. On remarquera que l'erreur sur la pente se propage de part et d'autre de la moyenne, car le point le mieux connu est la moyenne. Si on note $\mathbf{\Delta x_i}$ et $\mathbf{\Delta y_i}$ les incertitudes par rapport à la droite, alors

$$\mathrm{var}\left(b(x_i+\Delta x_i)\right)=\left(x_i-\bar{x}\right)^2\mathrm{var}(b)+b^2\,\mathrm{var}(\Delta x_i)+\underbrace{\Delta x_i^2\,\mathrm{var}(b)}_{\text{petit}}$$

$$\approx\left(x_i-\bar{x}\right)^2\mathrm{var}(b)+b^2\,\mathrm{var}(\Delta x_i)$$

Ainsi comme précédemment on estime l'erreur de la valeur attendue:

$$\begin{aligned}\mathrm{var}(y_{att})&=\mathrm{var}(a+b(x_i-\bar{x}+\Delta x_i)+\Delta y_i\\&\approx\mathrm{var}(a)+\left(x_i-\bar{x}\right)^2\mathrm{var}(b_1)+b^2\,\mathrm{var}(\Delta x_i)+\mathrm{var}(\Delta y_i)+2\,\mathrm{cov}(a,b(x_i-\bar{x}+\Delta x_i))\\&+2\,\mathrm{cov}(a,\Delta y_i)+2\,\mathrm{cov}(\Delta y_i,b(x_i-\bar{x}+\Delta x_i))\end{aligned}$$

Les covariances mettant en jeux les $\mathbf{\Delta x_i}$ et $\mathbf{\Delta y_i}$ sont nulles car ces variables sont aléatoires et indépendantes des autres. D'autre part **a** et **b** sont indépendants car les moyennes ne dépendent pas des variances. Ainsi on se place dans le cas où toutes les variables ont une distribution gaussienne indépendante. Dans ce cas il n'y a pas d'évidence pour utiliser une distribution de Student, on se "contentera" des incertitudes déduites d'une loi normale:

$$\frac{y_{att}-Y_{exacte}}{\sqrt{\mathrm{var}(a)+(x_i-\bar{x})^2\,\mathrm{var}(b)+b^2\,\mathrm{var}(\Delta x_i)+\mathrm{var}(\Delta y_i)}}\qquad(19)$$

Cette expression suit une loi normale réduite, à partir de laquelle ont déduit les intervalles de confiance. On remplacera dans les équations l'estimation de $b^2\,\mathrm{var}(\Delta x_i)+\mathrm{var}(\Delta y_i)$ par $\mathrm{var}\left(y_{estimé}-y_{mesuré}\right)$. En effet l'erreur sur x est

reportée sur y. Comme on l'a vu précédemment si l'on estime une moyenne on peut supprimer les deux derniers termes qui concernent la mesure elle-même:

$$\frac{y_{moyen.att} - Y_{exacte}}{\sqrt{var(a) + (x_i - \bar{x})^2\, var(b)}} \quad (19bis)$$

XVI.5.b Matrice de corrélation et distribution normale à plusieurs dimensions [3, 10]

Soit une variable aléatoire $\mathbf{X[X_1, X_2.... X_n]}$ à n dimension et $\mathbf{Y_j}$ une nouvelle variable, dans un nouveau repère, ou simplement une variables dans le repère de **X**, le lien entre ces deux variables s'écrit

$$Y_j = a_{1j}X_1 + a_{2j}X_2 + ... + a_{nj}X_n = \sum_{i=1}^{n} a_{ij}X_i$$

On notera qu'il s'agit de la projection sur un vecteur unitaire de direction j dans le repère des **X**. Maintenant on aimerait connaître la variance de $\mathbf{Y_j}$. On peut montrer que la moyenne d'un nuage de points dans un repère est égale à la moyenne transformée provenant de l'autre repère. Si on explicite la variance, on obtient

$$var(Y_j) = var\left(\sum_{i=1}^{n} a_{ij}X_i\right) = E\left[\left(\sum_{i=1}^{n} a_{ij}X_i - \sum_{i=1}^{n} a_{ij}\mu_i\right)^2\right] = E\left[\left(\sum_{i=1}^{n} a_{ij}(X_i - \mu_i)\right)^2\right]$$

$$= E\left[\left(\sum_{k=1}^{n}\sum_{i=1}^{n} a_{ij}(X_i - \mu_i)(X_k - \mu_k)a_{kj}\right)\right] = \sum_{k=1}^{n}\sum_{i=1}^{n} a_{ij}\sigma_{ik}^2 a_{kj}$$

Où $\mathbf{\sigma_{ik}^2}$ représente les covariances entre chaque composante de **X** dont l'ensemble forme la matrice des covariances.. Sous forme matricielle on écrit

$$\sum_{k=1}^{n}\sum_{i=1}^{n} a_{ij}\sigma_{ik}^2 a_{kj} = a^T \Sigma\ a = \begin{bmatrix} a_{1j} & \dots & a_{nj} \end{bmatrix} \begin{bmatrix} \sigma_{11}^2 & \dots & \sigma_{1n}^2 \\ . & \dots & . \\ \sigma_{n1}^2 & \dots & \sigma_{nn}^2 \end{bmatrix} \begin{bmatrix} a_{1j} \\ \dots \\ a_{nj} \end{bmatrix}$$

Cette opération n'est autre chose que l'estimation dans la direction $\mathbf{a_j}$. La matrice des covariances est symétrique. On exige qu'elle soit régulière, c'est-à-dire qu'elle soit inversible. On s'aperçoit que si l'on trouve une transformation qui diagonalise la matrice, seules les variances sont non nulles (la trace). En fait on peut les trouver en cherchant les valeurs et les vecteurs propres. Les vecteurs propres sont alors les axes principaux de l'ellipse de dispersion des points.

Connaissant la matrice des covariances, on doit chercher à trouver la forme de la loi multinormale. Le problème est d'estimer la variance, on le contourne en cherchant une loi normale réduite. Il faut trouver un produit scalaire normé à la

variance. Pour ceci connaissant la matrice de covariance en son inverse $\left(\Sigma^{-1}\right)$ on va recalculer le produit scalaire sous cette forme:

$$1=\sum_{k=1}^{n}\sum_{i=1}^{n}a_{ij}\left(\sum_{l=1}^{n}\sigma_{il}^{2}\left(\Sigma_{lk}^{-1}\right)\right)a_{kj}=E\left[\left(\sum_{k=1}^{n}\sum_{i=1}^{n}a_{ij}\sum_{l=1}^{n}\left((X_i-\mu_i)\left(\Sigma_{lk}^{-1}\right)(X_l-\mu_l)\right)a_{kj}\right)\right]$$

Ainsi on a obtenu une variable normée réduite. Pour obtenir la distribution il faut encore normer l'élément de volume, c'est-à-dire calculer le jacobien de la transformation. Il s'agit ici d'une homotétie on peut en effet voir que la variable normée réduite s'écrit sous forme matricielle comme:

$$Y=\Sigma^{-1/2}(X-\mu)\quad\text{et}\quad\text{ainsi}\quad dY=\frac{1}{\det\left(\Sigma^{1/2}\right)}dX$$

De sorte que la distribution s'écrit:

$$f(X_1,....,X_n)=\frac{1}{(2\pi)^{n/2}(\det\Sigma)^{1/2}}e^{-\left(\frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n}(x_j-\mu_j)\Sigma_{ij}^{-1}(x_i-\mu_i)\right)}$$

Le terme $(2\pi)^{n/2}$ provient du fait qu'il s'agit du produit de n variables, on le vérifie bien si la matrice est diagonale, alors la loi se résume au produit de n lois normales. Dans le cas où n=2 on a:

$$\Sigma=\begin{bmatrix}\sigma_x^2 & \left(r^2\sigma_y\sigma_x\right)\\ \left(r^2\sigma_y\sigma_x\right) & \sigma_y^2\end{bmatrix}\qquad\det(\Sigma)=\sigma_x^2\sigma_y^2\left(1-r^2\right)$$

$$\Sigma^{-1}=\frac{1}{\det(\Sigma)}\begin{bmatrix}\sigma_y^2 & -\left(r^2\sigma_y\sigma_x\right)\\ -\left(r^2\sigma_y\sigma_x\right) & \sigma_x^2\end{bmatrix}$$

Ainsi le terme constant est $\left(2\pi\sigma_x\sigma_y\sqrt{1-r^2}\right)^{-1}$ et l'exposant devient:

$$-\frac{1}{2}\frac{1}{\sigma_x^2\sigma_y^2(1-r^2)}\begin{bmatrix}(x-\mu_x) & (y-\mu_y)\end{bmatrix}\begin{bmatrix}\sigma_y^2 & -\left(r^2\sigma_y\sigma_x\right)\\ -\left(r^2\sigma_y\sigma_x\right) & \sigma_x^2\end{bmatrix}\begin{bmatrix}(x-\mu_x)\\ (y-\mu_y)\end{bmatrix}$$

$$=-\frac{1}{2(1-r^2)}\left[\left(\frac{(x-\mu_x)}{\sigma_x}\right)^2-2r^2\left(\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right)+\left(\frac{(y-\mu_y)}{\sigma_y}\right)^2\right]$$

Une ellipse centrée à l'origine peut être vue comme une forme bilinéaire du type:

$$\begin{bmatrix}x & y & 1\end{bmatrix}\begin{bmatrix}a_{11} & a_{12} & 0\\ a_{12} & a_{22} & 0\\ 0 & 0 & c\end{bmatrix}\begin{bmatrix}x\\ y\\ 1\end{bmatrix}=0$$

En fait c'est équivalent à un produit scalaire dans le plan dont le résultat est une constante égale à -c>0. En partant du cas particulier ou $a_{11}=a_{22}=1$ on obtient un cercle dont le rayon et donné par $r^2=-c$. Partant de cette constatation on sait qu'une ellipse est une dilatation d'un cercle suivant une direction. Intuitivement toute transformation linaires ayant des valeurs propres et vecteurs propres réelles induit une ellipse. Dans notre cas cette matrice est symétrique donc les valeurs propre sont réelles.

Ainsi une probabilité donnée suit une ellipse dans le plan xy. On peut ici remplacer les moyennes et les variances par leurs estimateurs. On voit ici tout le problème de l'application des méthodes statistiques. Chercher les axes principaux (voir plus loin) d'une telle distribution revient à supposer que la distribution des points respecte une telle distribution. C'est ce que nous avons fait lorsqu'on a simplifié l'erreur de la droite RMA. Il est donc important de constater que certaines hypothèses sont très fortes. En effet des droites de régression sur les deux axes dont le nuage de points n'est pas "régulier" peut donner des résultats assez erronés au niveau de l'estimation des erreurs, qui sont elles contraintes par des hypothèses de normalité. Il est possible comme on l'a vu pour la méthode RMA de travailler sans hypothèses gaussiennes, mais les calculs sont beaucoup plus longs et font appel à des moments d'ordre supérieur dont l'estimation est très sensible aux valeurs éloignées. Il est donc parfois préférable d'utiliser la méthode des moindres carrés standards, qui elle, ne nécessite pas des hypothèses aussi fortes ou ces ennuis. Cependant on peut à ce moment-là pratiquer une moyenne entre les deux régressions (x et y), de même pour les erreurs.

XVI.5.c La droites des "axes principaux" [4, 7]

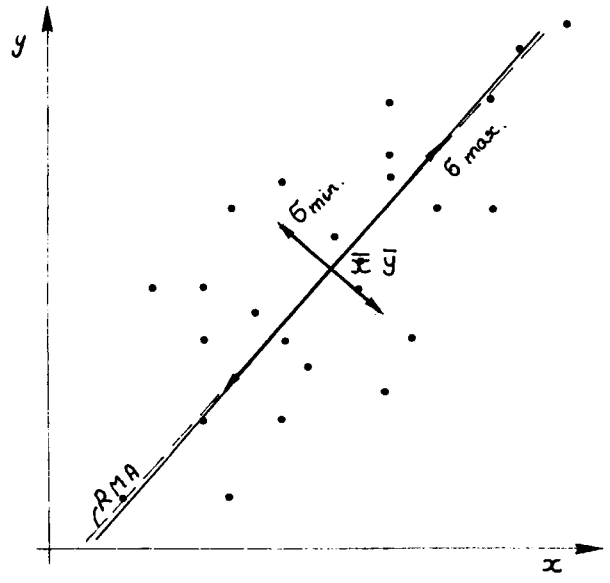

Figure 6: (tiré de [4]) La droite est alignée sur le vecteur d'écart type maximum (σ_{max}). La direction perpendiculaire (σ_{max}) à la droite et la direction de plus faible variance. Il y a une grande similitude entre les droites RMA et axes principaux. Comparer avec la figure 5.

On a vu qu'on pouvait connaître la variance dans n'importe quelle direction de l'espace. Or il existe, dans le cas idéal, des directions privilégiées: ce sont les axes principaux de l'ellipse de dispersion. On va donc chercher les directions ou la variance est maximale ou minimale (figure 6).

> Ces directions sont les vecteurs propres de la matrice Σ, on peut le montrer en annulant les premières dérivées par rapport aux à directions choisies. Et ceci sous la contrainte que les vecteurs de la bases cherchée soient normés. On l'effectue par la méthode des multiplicateurs de Lagrange, qui ne sont autres chose que les valeurs propres.

Mais ici on va déduire les deux directions de façon plus simple. Soit **X** et **Y** des variables aléatoires centrées sur les moyennes. On peut choisir de nouvelles coordonnées normées perpendiculaires telles que

$X' = X \cos\alpha + Y \sin\alpha$

$Y' = -X \sin\alpha + Y \cos\alpha$

Dans les directions principales la covariance est nulle, par conséquent on calcule la covariance de **X' Y'** et on l'annule:

$$\begin{aligned}
&\operatorname{cov}(X', Y') = E[X'Y'] \\
&= -E[X^2]\sin(\alpha)\cos(\alpha) + E[XY]\left(\cos^2(\alpha) - \sin^2(\alpha)\right) + E[Y^2]\sin(\alpha)\cos(\alpha) \\
&= \frac{1}{2}\sin(2\alpha)\left(\operatorname{var}(Y) - \operatorname{var}(X)\right) + 2\operatorname{cov}(X, Y)\cos(2\alpha)
\end{aligned}$$

de sorte que $$\operatorname{tg}(2\alpha) = \frac{2\operatorname{cov}(X, Y)}{\operatorname{var}(X) - \operatorname{var}(Y)}$$

On obtient ainsi deux valeurs pour α perpendiculaire. Avec les estimateurs on peut écrire:

$$\alpha = \frac{1}{2}\operatorname{arctg}\left(\frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2 - \hat{\sigma}_y^2}\right) \qquad (20)$$

Comme la variance et la covariance sont indépendantes de l'origine, on peut généraliser au cas où les moyennes et l'origine ne coïncident pas. Ainsi on a trouvé la pente **b=tgα**. On trouve le terme constant avec la condition de passer par les moyennes:

$$\bar{y} = a + \bar{x}\ \operatorname{tg}(\alpha) \qquad (21)$$

L'erreur sur a et b peuvent être obtenus après des calculs longs et fastidieux en suivant le même principe que la dérivation des erreurs de la droite RMA. Les formules générales étant longues nous ne reproduirons ici que les résultats pour des distributions normales bivariées:

$$\operatorname{var}(b) = \frac{(1 - r^2)\operatorname{tg}^2\alpha}{nr^2} \qquad (22)$$

$$\mathrm{var}(a) = \frac{1}{n}\left[\left(\sigma_y - \sigma_x \mathrm{tg}\alpha\right)^2 + (1-r)\mathrm{tg}\alpha\left(2\sigma_x\sigma_y + \frac{\bar{x}^2(1+r)\mathrm{tg}\alpha}{r^2} \right)\right] \quad (23)$$

Etant entendu que le nuage de points doit être proche d'une ellipse pour utiliser ces résultats. L'erreur le long de la droite peut être estimer selon la même formule que pour la droite des RMA.

XVI.6 Les droites de régressions appliquées aux standardisations XRD

Il y deux cas principaux qui nous intéressent. Le premier est celui de la recherche d'une quantité d'un minéral connaissant ses intensités de diffractions considéré proportionnel à la quantité de minéral. On utilise soit les intensités brutes ou corrigées par un coefficient d'absorption, ou encore les rapporter à l'intensité d'un minéral ajouté en quantité connue. Le second cas concerne la standardisation en laboratoire d'un paramètre, dans le cas qui nous préoccupe la largeur de Scherrer.

XVI.6.a Erreur dans l'estimation quantitative

La méthode des standards internes ou externes qui prend en compte les rapports de l'intensité de certains pics de diffraction, nécessitent au préalable l'établissement d'une droite d'étalonnage. On considère en effet que le rapport des intensités est proportionnel au rapport des masses. Pour établir une telle droite on prépare de nombreux échantillons de substance dans des rapports différents. On considère qu'il n'y a pas d'erreurs sur les masses, car en effet on est en dessous de 1% d'erreur. On ne fait donc pas une grosse erreur en effectuant la régression que sur le rapport des intensités. En général on utilise une droite de régression contrainte à l'origine, qui met en relation les rapports des masses (M) et des intensités (I):

$$\frac{I_a}{I_b} = b\frac{M_a}{M_b}$$

> Il est intéressant d'effectuer des tester avec des droites non contraintes à l'origine de même que, des régressions sur les deux variables en même temps, car l'examen de tels résultats peut donner des renseignements précieux.

Or on s'aperçoit qu'on cherche l'inverse. En effet on veut les rapports de masses connaissant les rapports d'intensités, donc on utilise:

$$\frac{M_a}{M_b} = \frac{1}{b}\frac{I_a}{I_b}$$

L'erreur lors de l'établissement de la régression se trouve sur le rapport des intensités. Une fois qu'on a estimé le rapport des masses on calcule les intervalles de confiance qui y correspondent suivant la régression. On obtient un intervalle de confiance sur $\mathbf{I_a/I_b}$, en supposant que l'erreur sur ce rapport est inhérent aux

mesures, on calcule les erreurs sur le rapport $\mathbf{M_a/M_b}$ en négligeant les termes de second ordre:

$$\Delta_{\%}\left(\frac{M_a}{M_b}\right)=\left(b+\frac{\Delta_{\%}b}{b}\right)\Delta_{\%}\left(\frac{I_a}{I_b}\right)\approx b\Delta_{\%}\left(\frac{I_a}{I_b}\right)=\Delta_{M\%}$$

Dans le cas des standards internes on doit trouver l'erreur sur le minéral inconnu, deux cas se présentent selon que la masse inconnue **x** est au numérateur ou au dénominateur, en différenciant, on obtient:

$$\left|\frac{d}{dx}\left(\frac{a}{x}\right)\right|\delta x=\frac{a}{x^2}\delta x=\Delta_{M\%}\quad \delta x=\frac{x^2}{a}\Delta_{M\%}\qquad (22)$$

$$\text{et}\ \left|\frac{d}{dx}\left(\frac{x}{a}\right)\right|\delta x=\frac{1}{a}\delta x=\Delta^{*}_{M\%}\quad \delta x=a\Delta^{*}_{M\%}\qquad (23)$$

Ces résultats indiquent que quand **x<a** l'erreur diminue lorsque **x** est au dénominateur. Par contre lorsque x est plus grand il est préférable d'utiliser l'autre cas. Il faut cependant faire attention dans le choix de l'un ou de l'autre de ces solutions, car la valeur de $\mathbf{\Delta_M}$ n'est plus la même. C'est à l'examen visuel et à la qualité des régressions qu'on choisira l'opportunité des différents cas.

> Le cas où l'inconnue est au numérateur trace une hyperbole. On s'aperçoit tout de suite qu'une grande erreur sur y pour des petites valeurs de x est petite. A l'inverse, si x est grand, une petite erreur sur y induit une grosse erreur sur x. En fait en utilisant x au dénominateur on dilate le segment [0,1] du cas linéaire en à l'infini.

Dans le cas des standards externes on a une incertitude sur les deux termes de la fraction. Par conséquent on effectue la différentielle totale:

$$\frac{\partial}{\partial x}\left(\frac{x}{y}\right)\delta x+\frac{\partial}{\partial y}\left(\frac{x}{y}\right)\delta y=\frac{y\delta x+x\delta y}{y^2}=\Delta_{M\%}$$

On peut à priori connaître la répartition du poids des incertitudes, tel que

$$w\frac{\delta x}{x}=\frac{\delta y}{y}\quad \delta y=w\frac{y}{x}\delta x$$

Ainsi on obtient pour l'intervalle de confiance δx sur x

$$\frac{y\delta x+xw(y/x)\delta x}{y^2}=(1+w)\frac{\delta x}{y}=\Delta_{M\%}\qquad (24)$$

Souvent on peut choisir **w=1**. Les remarques précédentes s'appliquent aussi à ce cas.

XVI.6.b Standardisation d'un laboratoire à l'autre

Le cas qui nous intéresse est celui de la calibration de la largeur de Scherrer. Dans ce cas les valeurs sur les deux axes sont comparables, il en va de même pour les erreurs. Il est donc nécessaire d'utiliser une régression sur les deux variables. On

choisira donc la droite RMA et/ou des axes principaux de dispersion. Il n'est pas inutile d'effectuer des régressions simples sur x et sur y pour se faire une idée de la variabilité. Comme l'estimation de l'erreur sur les limites est primordiale, il est nécessaire de choisir un domaine de valeurs dans lequel le comportement est linéaire. En effet on peut s'attendre à ce que sur un trop grand champ de valeurs le comportement soit non linéaire. De ce fait une droite contrainte à l'origine n'est pas adéquate.

XVI.7 Les chiffres significatifs

On parle ici indifféremment d'erreurs et d'incertitudes.

Il est préférable d'écrire les résultats en valeurs absolues, car cela évite des bévues au niveau de la forme des résultats, par exemple on n'écrit pas

$\underbrace{753.4521 \pm 1\%}_{\text{faux}}$ mais 753 ± 8.

En effet l'incertitude absolue ne doit pas avoir de décimale plus élevée que celle du résultat, puisque on a pas pu mesurer au-delà. On se limite au maximum à une erreur sur les deux derniers chiffres significatifs. On en utilise qu'un seul si le premier chiffre de l'erreur n'est pas du même ordre de grandeur que le dernier chiffre significatif de la mesure, autrement dit si on s'arrange avec des puissances de dix pour que les deux termes significatifs soient juste après la virgule on a:

$a.bc \pm 0,b'c'$ alors si 0.b'>10 0.0c ou 0.01 si c=0 on a:

a.b ou a.(b+1) si c≥5

et l'intervalle de confiance est remplacé par

0.(b'+1)

Un exemple:

4.11±0.93 peut être remplacé par 4.1±1.0

On note que le zéro est un chiffre significatif. Cependant ces règles de cuisine ne marchent que lorsque qu'on utilise les puissances de dix. Mais on peut toujours s'y ramener, et il est même préférable de s'y ramener pour des nombres grands avec de grandes erreurs, car les zéros de l'erreur peuvent être pris comme chiffres significatifs...

On considère ici que les erreurs de lecture et de mesure sont inclues dans l'estimation statistique, ce qui n'est pas forcément le cas. Il arrive qu'il faille ajouter des erreurs entres elles.

On prendra toujours garde d'éviter les erreurs systématiques telles les dérives etc...

XVI.8 Tentative d'interprétation des formes de pics de diffraction par la statistique seulement

Sans faire appel à des notions de la théorie de la diffraction classique, dans la pratique on utilise fréquemment pour simuler des pics de diffraction des fonctions telles que la loi normale, la loi de Cauchy (Lorentz) ou encore Pearson-VII. Or on peut toutes les déduire de la loi de Student sous certaines conditions. C'est cette remarque qui motive ces quelques lignes.

Soit des cristallites de taille moyenne **Y** comportant **n** fois un plan atomique. On suppose que la distance interréticulaire n'est pas constante. On définit la distance interréticulaire de la population de cristallite $\mathbf{d_m=Y/n}$ On écrit la taille d'une cristallite quelconque $\mathbf{Y_i}$

$$Y_i = \sum_{i=0}^{n} d_i$$

On suppose que le d-spacing apparent d'une cristallite est égal à $\mathbf{Y_i/n}$. On cherche à connaître la distribution des espacement interréticulaires moyens de chaque cristallite, ou plus commodément l'écart à la moyenne. Ainsi l'écart et la variance de cet écart s'écrivent:

$$\frac{(Y_i - Y)}{n} = \frac{1}{n}\sum_{i=1}^{n}(d_i - d) \quad \text{et} \quad \text{var}(Y_i - Y) = \sum \text{var}(d_i) = \sum \Delta d_i^2$$

Lorsque **n** est raisonnablement grand, en vertu du théorème central limite on peut supposé que la somme des $\mathbf{d_i}$ suit une loi normale, et par conséquent la variable

$\dfrac{(Y_i - Y)}{\sum_{j=1}^{n} \Delta d_j^2}$ est une variable de Student à un degré de liberté, qui n'est autre qu'un distribution de Cauchy.

Par contre, si à priori on suppose que les $\Delta\mathbf{d_i}$ suivent une loi gaussienne et donc a fortiori $\mathbf{Y_i}$. Cette fois on est en présence d'une variable de Student à n degrés de liberté. Lorsque n est grand on approche d'une loi normale. Si **n** est petit on se trouve dans un cas intermédiaire aux deux précédents, il s'agit d'une distribution de Pearson-VII. En effet il existe une relation entre la fonction de Pearson-VII et la loi de Student. On la note [6]:

$$P(x) = \frac{\Gamma(m)}{\sqrt{\pi}\Gamma(m-1/2)} \frac{c^{2m-1}}{\left[c^2 + (x-a)^2\right]^m} = \frac{\Gamma(m)}{\sqrt{\pi}\Gamma(m-1/2)} \frac{1}{c\left[1 + \dfrac{(x-a)^2}{c^2}\right]^m}$$

Si a=0, c=√ν et m=(ν+1)/2 on retrouve une loi de Student. Cette transformation n'est pas très contraignante, à un terme constant près on retrouve dans tous les cas une distribution identique à celle de Student.

Si les défauts d'empilement n'ont pas une distribution gaussienne, alors on tend vers un profil de Cauchy, ceci est vrai pour des lois différentes ou identiques. Par contre lorsque l'écart à la position idéale d'un plan atomique est de type gaussien on obtient un profil gaussien.

Une poudre est composé de cristallites d'épaisseurs différentes, cependant la courbe ne changera que peu, car ν sera soit très grand et variera en fonction de **n** et sera donc gaussien, et sinon ν sera petit et ne variera que peu.

Reynolds [14] et Ergun [13] montre que plus un empilement est désordonné, plus il tend vers un profil de Cauchy.

Un profil sans défaut est proche d'un profil de Gauss. Ce qui est dû à l'effet de taille. Mais on peut envisager que le même type de profil avec des défauts de type gaussien. Bien sûr nous n'avons pas tenu compte de l'effet de taille (fonction d'interférence) qui est proche d'une gaussienne. Or si l'on convolue une courbe de Cauchy avec un gaussienne on obtient une fonction proche de Pearson-VII. Donc cette fonction à l'avantage d'englober facilement les effets de taille et de désordre. Pour des défauts gaussiens, la convolution de deux courbes gaussiennes donne une gaussienne.

Bibliographie du chapitre XVI

[1] Arbenz, K., Wohlhauser, A. (1986): Analyse numérique. Lausanne: Presse polytechnique et universitaires romandes.

[2] Boas, M. L. (1983): Mathematical methods in the physical Sciences. New York, Jhon Wiley & sons.

[3]Chatfield, C., Collins, A.J. (1980): Introduction to multivariate analysis. New York: Chapman and Hall.

[4] Davis, J. C. (1986): Statistics ans data analysis in geology. Second edition. New York: Jhon Wiley & Sons.

[5] Guest, P. G. (1961): Numerical methods of curve fitting. Cambridge University Press.

[6] Johnson, N. L., Kotz, S. (1969): Continuous univariate distributions. New York: Wiley Interscience.

[7] Kendall, M. G., Stuart, A. (1966): The advanced theory of statistics. London: Griffin.

[8] Kermak, K. A., Haldane, J.B.S (1950): Organic correlation and allometry. Biometrika, *37, 30-41.*

[9] Mood, A. M., Graybill, F.A. (1973): Introduction à la statistique théorique (Maugen, P.Y., Burle, Y., Trans.). Paris: Dunod.

[10] Saporta, G. (1990): Probabilités analyse des données et statistique (Technip ed.). Paris:

[11] Spiegel, M. R. (1983): Probabilités et statistique (R. Jacoud, Trans.). Paris: Mc Graw-Hill.

[12] Ventsel, H. (1973): Théorie des probabilités (A. Sokova, Trans.). (Edition MIR ed.). Moscou:

[13] Ergun, S. (1970): X-ray scattering by very defective lattices. Phys. Rev. B., *131, 3371-3380.*

[14] Reynolds, R.C. Jr (1989): Diffraction by small and disordered crystals. In Modern powder diffraction, Bisch D.L. & Post J.E., eds. Mineralogical Society of America: Reviews in Mineralogy vol. 20, *145-181.*