INCERTITUDES et ERREURS.

Préambule.

Il m'a semblé utile de rédiger ce papier dans le cadre d'une formation au CNAM qui nécessite une bonne connaissance de certaines notions fondamentales. Je ne prétends pas être exhaustif, mais j'essayerai d'être le plus précis possible.
Par la suite, je l'ai étendu à des notions, soit plus précises, soit plus générales, et qui donc débordent par rapport au titre initial.

Etant donné que la finalité est la même, je me suis inspiré principalement du cours de J.J. LEVALLOIS, Ingénieur en Chef Géographe (IGN 1960). Parallèlement, je recommande vivement le cours du professeur Mathieu ROUAUD http://www.incertitudes.fr/proba-stat-acp/livre.pdf.
On pourra lire aussi le livre de John Hartong "Probabilités et Statistiques".

Il me paraît important de préciser quelques termes de vocabulaire.

Valeur vraie : toute quantité, qu'elle soit un nombre entier (comptage) ou un nombre réel a une valeur unique. Dans le sujet qui nous intéresse, cette valeur est généralement inconnue, on l'appelle "valeur vraie".

Mesure et observation : ces deux termes représentent à peu près la même chose. On emploiera plutôt "mesure" lorsque le résultat est une valeur réelle, longueur par exemple, et "observation" lorsque le résultat est un comptage.

Incertitude et erreur : ces deux termes désignent la même chose, c'est à dire la différence entre la valeur mesurée ou adoptée et la valeur vraie, généralement inconnue. J'utiliserai de préférence le terme "erreur" et naturellement, il n'a rien à voir avec le terme "faute".

Moyenne : sans qualificatif, il s'agit de la moyenne arithmétique.

Ecart-type : c'est le terme employé actuellement à la place de "écart moyen quadratique". C'est une unité de mesure des erreurs accidentelles.

Expérience : Quelque soit le contexte, le domaine étudié, on appelle expérience un ensemble d'observations, d'actions ou de mesures répétées et comptabilisées. On peut citer la pêche (industrielle), le tirage de dés , le jeu de pile ou face, le tir au pistolet ou au canon.

Les deux types d'erreur.

Prenons l'exemple du tir au fusil.
Il y a une cible constituée de cercles concentriques, on admet que le tireur vise correctement à chaque tir le centre de la cible, mais ne voit pas le résultat des tirs précédents.
On constate que les impacts sont groupés, mais pas forcément au centre de la cible. La raison est naturellement le réglage imparfait du viseur. On dit que la distance entre le centre de gravité des impacts et le centre de la cible est une erreur systématique.
On appelle erreur accidentelle la mesure de la dispersion des différents impacts.

Les erreurs systématiques doivent être évitées, ou connues, ou corrigées. Il y a plusieurs méthodes selon l'outil de mesure, par exemple l'étalonnage. Ceci n'entre pas dans cadre d'un présent papier.

Postulat de la moyenne.

On a constaté, à l'aide nombreuses expériences, que si on augmente le nombre de mesures, la moyenne de ces mesures tend vers une valeur. On dit aussi "converge vers …".
On en a tiré la conclusion logique que la moyenne des mesures doit être très proche de la valeur vraie, généralement inconnue. C'est le postulat de la moyenne. On a démontré que ce choix était justifié.

Constatation statistique.

En géodésie, on mesure les angles des triangles formés par les sommets de triangulation. Compte tenu de la correction de dv (différence entre les triangles sphériques et les triangle plans), on appelle "erreur de fermeture d'un triangle" la différence entre la somme des 3 angles mesurés et l'angle plat (180°).

Sur un grand nombre de triangles on a réparti les erreurs par tranches, dans le tableau suivant :

Signe	0" à 5"	5" à 10"	10" à 15"	15" à 20"	20" à 25"	25" à 30"
+	105	84	40	9	3	2
-	103	86	34	13	3	2

Si on reporte en abscisse les limites de tranche et en ordonnée le nombre d'écarts constatés, et que l'on joint les milieux des sommets de barres, on observe une courbe en cloche bien connue sous le nom de courbe de Gauss.

Citation : "Cette expérience a été faite bien souvent, sur des objets très différents, ses résultats sont constants, la courbe de répartition a toujours la même allure. Il y a plus, les courbes obtenues dans l'étude des différents cas sont superposables par un simple changement d'échelle des abscisses et des ordonnées, […]".

La conclusion de ceci est importante : étant donné une expérience quelconque, on est sûr que la répartition des écarts sera conforme à celle connue. Donc dans tous les cas, la répartition des écarts est la même, il en résulte que si on constate une répartition différente de la répartition connue, il y a une anomalie.

Le hasard.

Cette notion, le hasard, n'a pas été évoquée pour l'instant.
Dans le cas cité des 484 triangles, étant donné que les mesures ont été réalisées par plusieurs personnes différentes et sur plusieurs années, mais avec des méthodes identiques, seul le hasard intervient dans la répartition des écarts.
John Hartong définit très bien et en détail cette notion de "hasard", je ne reviendrai donc pas sur ce point.

Petite vérification facile.

On a vu que les écarts se calculent par rapport à la moyenne des observations.
Prenons un dé à jouer (6 faces) numérotées 1 à 6 ou du 9 à l'As ou n'importe quel graphisme.
On fait 30 tirages et on note le numéro de la face sortie à chaque tirage.
On les compte, la moyenne doit faire 5.
Pour chaque face, on calcule la différence du nombre de sorties par rapport à la moyenne (5 dans le cas présent), et on l'élève au carré. On en fait la somme que l'on divise par 6, on en prend la racine carrée : c'est l'écart-type.
On calcule l'écart probable ep= 2/3 écart-type.
On définit 4 classes limitées par –1ep, 0, +1ep
On compte le nombre de sorties dans chaque classe et on trouve forcément 25% pour chaque classe.
Ce test est très simplifié, puisqu'il n'y a que 6 issues possibles, mais il a l'intérêt de pouvoir être fait indépendamment de tout système parasite.
Il est bien évident que le même test avec un dé à 24 faces et un plus grand nombre de tirages donnera un résultat d'une plus grande finesse.

L'écart type.

Soit un ensemble N de mesures indépendantes d'une même quantité X.
On peut calculer la moyenne arithmétique Xm, et pour chaque quantité l'écart vi à la moyenne : vi=Xm - xi.
Supposons que l'on connaisse la valeur vraie de la quantité X, alors l'erreur vraie de chaque mesure est ei = X – xi, différente de l'erreur apparente vi.
L'écart type emq = sqrt(somme(ei^2/N))
Strictement parlant l'emq est "plus ou moins". Dans la pratique, on considère qu'une emq est positive, sachant que par symétrie elle peut aussi être négative.

Si on ne connaît pas la valeur vraie de X, alors
Emq = sqrt(somme(vi^2/(N-1)))

Explications :
Dans le cas de la fermeture des triangles géodésiques, on connaît la valeur vraie de la somme des angles intérieurs, dans le cas du tirage de dé aussi. Ces cas ne correspondent pas aux cas les plus fréquents, en effet le but étant le plus souvent de définir la mesure la plus probable pour X, on ne la connaît pas. Donc, dans le cas général, tout au moins le plus fréquent, le dénominateur de la formule sera (N-1).
Imaginons que l'on ait effectué une mesure d'une chose, et une seule. On a donc Xm=x1 et e1=Xm-x1 ; e1=0.
Si on appliquait le calcul de l'écart type en prenant N comme dénominateur, on obtiendrait un écart type égal à 0, ce qui signifierait que la valeur mesurée est égale à la valeur vraie, ce qui serait vraiment très intéressant !
En fait le bon calcul donne emq=0/(1-1)=0/0 qui est une valeur indéterminée.
J'insiste sur ce point, puisqu'on trouve quelquefois des calculs d'écarts type faits à partir d'une seule observation.

La loi normale.

La courbe de Gauss (appelée aussi courbe en cloche) a une importance considérable dans le monde réel. Par exemple le seuil d'une maison ancienne a pris la forme de la courbe de Gauss.
La courbe de Gauss est le courbe représentative de la loi normale. Etudions ses caractéristiques.
On appelle "densité" l'aire de la zone qui se trouve entre la courbe et l'axe des X.
Il existe des tables qui donnent la valeur de la densité en fonction de la variable x. Ces tables sont à utiliser dans le cas (rare) de recherche de précision sur la répartition des écarts, ce qui n'a rien à voir avec la gestion des écarts eux-mêmes.

Voici la méthode simple et facile à retenir pour utiliser la loi normale.
Etant donné le résultat d'une expérience, on peut calculer l'écart type (emq).
On calcule l'écart probable ep = 2/3 emq.
L'écart probable est tel que la moitié des écarts lui est inférieur.
On défini 10 classes dont les frontières seront, en plus de 0, 1ep, 2ep, 3ep et 4ep à droite et à gauche. Alors
0.35% des écarts seront inférieurs à –4ep
2% des écarts seront compris entre –4ep et –3ep
7% des écarts seront compris entre –3ep et –2ep
16% des écarts seront compris entre –2ep et –1ep
25% des écarts seront compris entre -1ep et 0
25% des écarts seront compris entre 0 et 1ep
16% des écarts seront compris entre 1ep et 2ep
7% des écarts seront compris entre 2ep et 3ep
2% des écarts seront compris entre 3ep et 4ep
0.35% des écarts seront supérieurs à 4ep
On admet généralement que les écarts supérieurs à –4ep ou 4ep sont anormaux, donc à éliminer.

Dans la littérature actuelle, on défini généralement 8 classes et non 10, l'unité de mesure étant l'écart type, mais le principe est strictement le même. De même on peut lire que un écart inférieur à 2 écarts types a une probabilité de 95%, ce qui est aussi identique à la répartition indiquée.

Il est important de vérifier que la répartition des écarts est conforme à la loi normale. Si ce n'est pas le cas, il y a une erreur, ou plus grave, une tricherie.
Avec les moyens informatiques actuels, il peut être considéré comme une faute d'avoir omis de vérifier la répartition des écarts, dans le cas où celle-ci ne serait pas conforme à loi normale.

Méthode de Monte-Carlo.

Le principe est très simple : on doit résoudre une équation compliquée qu'on ne sait pas résoudre par les méthodes habituelles. L'exemple classique est la résolution de certaines équations différentielles.
On fait une simulation qui consiste à partir de valeurs aléatoires en assez grand nombre et à "calculer" l'équation.
On va ainsi obtenir un couple ou triplet variable(s)-solution(s) qui va tendre vers la solution de l'équation. Le nom de la méthode est naturellement issu du casino du même nom, bien connu.
A l'expérience, on constate que le résultat converge très rapidement, ce qui est la meilleure des justifications. Pour ce genre de simulation, on utilise la fonction rand, bien connue.

Méthode des moindres carrée.

Considérons un ensemble d'observations.
Pour fixer les idées, voici deux exemples :

1- dans le cadre d'un levé topographique, on a mesuré des angles et des distances entre points de triangulation. Le valeurs sont en sur-nombre, comment mener le calcul ?
2- vous cherchez à mettre en superposition deux plans de la même zone issus de deux sources différentes. Vous allez identifier des points caractéristiques identifiables sur les deux documents et noter leurs coordonnées dans deux systèmes différents et n'ayant aucun rapport. Comment mener la calcul ?

On constate tout de suite que la formulation est difficile, et que mathématiquement parlant, il existe une infinité de solutions.
Citation :
"Puisque, pour la moyenne arithmétique, la somme des carrés des résidus est minima, on conviendra d'appliquer le principe, quelque soit la forme des relations d'observations. On démontre d'ailleurs que cette solution est << la plus probable >> au sens du calcul des probabilités".

En d'autres termes, on cherche à trouver le résultat le plus probable, c'est à dire celui qui sera le plus proche de la valeur vraie, inconnue. On trouve souvent l'expression "maximum de vraisemblance". On a vu que ceci était le cas lorsque l'écart moyen quadratique était minimum. La méthode des moindres carrés consiste donc à minimiser la somme des carrés des écarts. L'étude détaillée de la méthode dépasse le cadre de ce papier. Disons seulement que si S est la somme des carrés des écarts, S sera minimum pour la valeur qui annule sa dérivée. Le plus souvent S est fonction de plusieurs variables, l'annulation de chaque dérivée partielle conduira à un système linéaire de n équations à n inconnues.

Composition des erreurs.

Pour mémoire, il y a lieu de rappeler que les erreurs systématiques s'ajoutent, d'où l'importance d'éviter de type d'erreur, par des moyens appropriés.

Les erreurs accidentelles se composent quadratiquement.
Prenons deux exemples pour l'expliquer.

Soit à mesurer une longueur de 1km environ avec un ruban d'acier qu'on appelle chaîne d'arpenteur, de 20 mètres. On sait par des expériences précédentes que l'erreur accidentelle d'une mesure est emq=0.01. Quelle sera l'erreur moyenne

quadratique, c'est à dire l'écart type sur la mesure de cette distance, soit 50 portées ? La solution est 0.01 x sqrt(50), soit environ 7 cm.

On mesure un angle avec un théodolite. Son emq est 0.001 grade. Pour éliminer les erreurs systématiques, on fait plusieurs tours, c'est à dire qu'on mesure plusieurs fois le même angle en changeant l'origine du cercle.
Pour 2 tours, l'emq sera 0.001 / racine(2)
Pour 4 tours, l'emq sera 0.001 / racine(4)
Pour 16 tours, l'emq sera 0.001 / racine(16)

Il s'agit là de cas simples. Dans la pratique les mesures seront faites avec des outils et des méthodes différentes et non homogènes. Dans certains cas on pourra négliger certaines erreurs par rapport à d'autres, mais souvent il faudra composer ces différentes erreurs par des moyennes pondérées.

Conclusion.
En conclusion, le citerai cette phrase de M. de Lapalisse "Si une erreur était connue, ce ne serait pas une erreur". Cela signifie que une emq se calcule, mais cela ne permet que de vérifier que la valeur réelle d'une mesure se situe dans un intervalle précis.

Exercice 1.

Je suis face à un problème d'optimisation :
mon client me dit que je dois livrer la pièce de remplacement de la pièce défectueuse au bout de 30jrs.
Moi je ne peux livrer la pièce qu'au bout de 70jrs
sachant que par an j'ai 27 demandes
quel est le stock minimum que je dois mettre en place pour pouvoir assurer mon engagement
comment formaliser ce problème mathématiquement.
(Copie d'une question posée sur un forum).

Exercice 2

Un entreprise de pêche dispose de plusieurs chalutiers.
Pour des raisons évidentes de simplification cette entreprise a passé un contrat avec divers clients qui consiste à fournir dès le retour au port une ou plusieurs caisses.
On sait que les poissons les plus appréciés sont ceux qui sont ni trop petits, ni trop gros.
Le directeur de l'entreprise soupçonne certains patrons pêcheurs de réserver aux caisses marquées "Client" les meilleurs poissons. Comment peut-il le vérifier et le prouver ?.

Annexe A : Le hasard

Comme j'ai dit plus haut, il n'est pas question pour moi de chercher à compléter le livre de John Hartong, mais seulement de préciser quelques méthodes d'approches qui dépendent de cette notion : le hasard.

Tirage à pile ou face.
On sait que la probabilité de tirer pile, resp. face, est ½. Si on réalise un grand nombre de tirages, on observe que le nombre de pile est très proche du nombre de face.
On peut compliquer cette expérience, en comptant, non plus le nombre de pile, resp. face, mais le nombre de sorties de listes continues de pile, resp. face.
La probabilité de tirer 2 face en suivant est ¼, pour 3 faces c'est 1/8, pour 4 faces c'est 1/16 etc.
Cette expérience est relativement facile à faire avec un ordinateur. On utilise pour cela un générateur de nombres pseudo-aléatoires. Un tel générateur n'est pas vraiment aléatoire, dans le sens où le cycle de sortie des nombres est un nombre fini. C'est à dire que si on démarre le tirage toujours au même endroit, qu'on appelle la graine, on aura toujours la même liste. On voit donc l'importance d'initialiser le tirage de façon aléatoire, c'est à dire déterminer la graine, on utilise souvent l'horloge interne de la machine.
Généralement cette fonction, appelée rand() ou ALEA() est basée sur une multiplication puis un modulo. Certains auteurs de logiciels destinés aux mathématiciens ont cru bon de créer une liste de nombres identiquement répartis sur un intervalle, et c'est une fonction appelée rand() ! Lorsqu'elle est utilisée sans autre précision, c'est à dire avec l'option par défaut, la liste ne correspond pas au résultat d'un tirage aléatoire. Pour faire une simulation de tirage aléatoire, il faut préciser l'option "normal".
A l'opposé, probablement pour des raisons de jeux sur Internet, il existe une fonction appelée GenRand() qui garanti à chaque tirage une répartition conforme à la loi normale. Il s'agit là d'un excès inverse par rapport à la fonction rand() par défaut des logiciels à orientation mathématique.
Personnellement, je considère ces deux déviations comme des anomalies, puisque, dans le monde réel, la notion d'aléatoire est unique.

Annexe B : Les régressions

La problématique est la suivante : on dispose d'un certain nombre d'observations sous forme de couples de valeurs. On aura donc des couples (x1;y1), (x2;y2),…,(xn;yn).
On cherche une relation entre ces couples, telle qu'on puisse l'écrire sous la forme mathématique Y = f(X).

On applique la méthode des moindres carrés : S est la somme des carrés des écarts.
$S= \Sigma (yi - f(xi))^2$ pour i de 1 à n.
S sera minimum lorsque sa dérivée s'annule.
Supposons que la fonction f soit de la forme f(x) = A + Bx, alors
$S = \Sigma (yi – A – Bxi)^2 = \Sigma(yi^2 - 2Ayi - 2Bxiyi + A^2 + 2ABxi + B^2xi^2)$
Les variables sont A et B. Dérivons par rapport à A et à B.
$dS/dA = - 2\Sigma yi + 2nA + 2\Sigma Bxi$
$dS/dB = - 2\Sigma xiyi + 2\Sigma Axi + 2\Sigma Bxi^2$
Le minimum pour S sera obtenu pour les valeurs de A et B qui annulent les dérivées.
$nA + B\Sigma xi = \Sigma yi$
$A\Sigma xi + B\Sigma xi^2 = \Sigma xiyi$
C'est un système de 2 équations linéaires à 2 inconnues qui admet une solution.

Ceci est la méthode et le principe général applicable dans tous les cas.

Cependant, le cas de la fonction f(x) = A + Bx n'est pas général. Par contre, on peut s'y ramener très souvent par un simple changement de variable, par exemple avec les fonctions Log ou Exp.
Lorsqu'on a l'occasion de calculer de nombreuses régressions, il est intéressant de développer un outil qui calcule les différentes fonctions possibles et qui indique la meilleure.
Avec Log et Exp on dispose de 4 fonctions possibles.
Régression linéaire y = a + bx
Ajustement par une courbe exponentielle $y = ae^{bx}$
Ajustement par une courbe logarithmique y = a + b ln x
Ajustement par une courbe de fonction puissance $y = ax^b$
Quel est le critère de choix ?
On utilise généralement le coefficient de détermination R^2.
$R^2 = (A\Sigma yi + B\Sigma xiyi – 1/n (\Sigma yi)^2) / (\Sigma (yi^2) – 1/n (\Sigma yi)^2)$
Plus ce coefficient est proche de 1, meilleure est la régression.
On peut aussi utiliser le calcul de l'écart-type. Cela a l'intérêt de fournir la précision numérique des données et de la régression.

Quelques généralisations de cette méthode de base.

1- On constate que les paramètres A et B ne suffisent pas pour résoudre le problème assez finement. Dans ce cas, il faudra trouver une fonction qui admette 3 paramètres, par exemple :
Y = A + B.C^X ; cad ln(Y-A) = lnB + X.ln(C)
Y = A + exp(D.X + E) ; cad ln(Y-A) = D.X + E

On aura donc à établir et à calculer un système de 3 équations à 3 inconnues.

2- Lorsque les observations sont sous le forme de triplets et non plus de couples, la fonction recherchée sera de la forme Z = f(X,Y). Compte tenu des changement de variable avec Log et/ou Exp, le nombre de fonctions possibles est plus important, mais le principe et la méthode restent les mêmes.

3- Si le nombres de variables est supérieur à 3, le nombre de fonctions à essayer devient trop important. J'ai adopté le principe que la fonction serait toujours de la forme.
Y = K . X1^a . X2^b . X3^c ….
Cette fonction a un intérêt au niveau écriture. Si on en prend le log, ça donne
log(Y)= log(K) + a.log(X1) + b.log(X2) + c.log(X3) +...
qui est une fonction linéaire très facile à traiter.

Par ailleurs, les paramètres a, b, c, suivant leur signe et la valeur par rapport à 1 donnent des courbes de variation qui peuvent prendre à peu près toutes les formes, par exemple a=1 donne une droite, a=-1 donne une hyperbole, a=2 donne une parabole etc. Une fonction de cette forme est utilisée en calcul de pluviométrie, non seulement personne ne l'a remise en cause et surtout elle a été officiellement confirmée vers les années 2000.

Il y a un autre avantage : les paramètres a, b, c, peuvent être eux-mêmes fonction d'autres paramètres, ce qui augmente encore la souplesse de cette fonction.

4- Un autre type de formule de régression : un polynôme de degré 4. Contrairement aux formules précédemment décrites, celle-ci peut être non monotone. Bien-sûr elle nécessite 5 paramètres, ce qui complique la mise en équation.

Concernant les régressions, il faut bien distinguer le cas où on connaît la forme de la fonction représentative du phénomène étudié, dans ce cas, on applique directement la méthode des moindres carrés et on calcule les paramètres par résolution d'un système d'équation.
Dans les cas plus fréquents, on ne dispose que d'observations et de mesures et il faut trouver une fonction qui convient, sachant que cette fonction n'est pas parfaite mais qu'elle suffit au point de vue précision.
Dans tous les cas, on restera très prudent lorsqu'il d'agira d'extrapoler, c'est à dire de calculer une valeur n'appartenant pas à l'intervalle réellement observé.

Pierre Dolez