L'APPLICATION DE LA REPRÉSENTATIVITÉ À L'AIDE A LA DÉCISION

Jean-Louis Pineau, Nathalie Valentin, Isabelle Barale*
LEM - LCE

La représentativité est généralement une notion qualitative de la propriété de représentation d'un lot pour une épreuve par un échantillon. Seule l'approche globale de GY permet d'en faire une grandeur quantifiable. Elle est égale au coefficient de variation cv pour un échantillonnage juste et elle dépend des trois opérations que sont l'échantillonnage, l'épreuve et la mesure. En particulier si les conditions d'application de ces opérations ne sont pas vérifiées, la représentativité n'est plus significative et l'aide à la décision est sans fondement.

The representativeness is usually a qualitative property of a sample to represent a whole in such a way that the value of a studied property is the same in both the whole and the sample. Only the global approach proposed by GY allows the representativeness to be a quantitative notion, equal to a coefficient of variation cv for a correct sampling. The representativeness depends on the three following operations : the sample preparation, the test and the measurement. If the operating conditions are not correctly checked, the representativeness is not significant and can't be used for further decision.

Qualifier un échantillon de représentatif ou lui adjoindre la qualité de représentativité, c'est garantir à la fois le résultat obtenu sur l'échantillon et le professionnalisme du praticien. Mais il apparaît souvent que cette qualité puise sa signification dans une intuition au lieu d'être l'expression d'un état physique parfaitement défini. L'explication se trouve dans le schéma décisionnel élémentaire (figure 1) qui montre que l'objet expérimental, le résultat expérimental, le résultat final et l'aide à la décision sont l'image des maillons de la chaîne qui relie l'objet initial à la décision.

De fait garantir le résultat obtenu sur l'échantillon, revient à assurer la validité des opérations intermédiaires et, en se référant à l'image de la chaîne à en garantir la solidité. Cette dernière étant dépendante du maillon le plus faible, est assurée par la qualité physique de chacun de ses maillons. C'est ainsi que la représentativité globale dépend de la représentativité de chacune des opérations et des opérateurs.

La représentativité n'est pas une notion statistique comme le montre l'étude des définitions présentées dans la première partie. C'est une notion expérimentale généralement qualitative, exceptée celle de GY qui est une grandeur physique. Pour mener à bien une telle application, les trois opérations que sont la mesure, l'épreuve et la préparation de l'objet expérimental doivent respecter certaines conditions dont le détail est donné dans la deuxième partie.

DÉFINITIONS DE LA REPRÉSENTATIVITÉ

La représentativité est une notion expérimentale. Elle est quasi-absente des ouvrages de mathématiques appliquées.

La représentativité du statisticien

Le domaine mathématique dans lequel la représentativité de l'échantillon est citée, est celui des statistiques au niveau de l'introduction à l'échantillonnage dans quelques ouvrages pour qualifier l'échantillon sur lequel vont porter toutes les définitions et démonstrations.

– Pour que les informations collectées puissent être étendues à l'ensemble de la population (Grais,1984) ; ou pour assurer une bonne qualité de l'information (Abboud, Audroing, 1989), il faut que l'échantillon soit représentatif,

– la représentativité est assurée par le mode de tirage équiprobable et indépendant (Saporta, 1990), (Spiegel, 1984),

– un échantillon stratifié issu d'une population stratifiée est représentatif si le taux de sondage est uniforme. Le taux de

Réalisation de l'objet	**Objet initial** *Préparation en vue de l'expérience retenue pour décrire la propriété*	L'opérateur Le préparateur
La propriété	**Objet expérimental** *Épreuve servant à caractériser la propriété*	L'expérimentateur
La quantification	*Le résultat expérimental* *Mesure*	L'analyste
	Le résultat final ou donnée	
La probabilité de la validité du classement référence/résultat	*Comparaison référence/résultat*	L'analyseur des données
	Aide à la décision	
	Décision	

Figure 1 : Schéma décisionnel élémentaire

sondage est le rapport entre le nombre d'individus de l'échantillon et celui de la strate dans laquelle l'échantillon a été prélevé (Mothes, 1968).
La représentativité de la donnée est plus rarement définie. Elle correspond à des valeurs caractéristiques qui donnent une image la plus fidèle de la distribution. Yule définit ce type de valeur à l'aide de 5 conditions indépendantes de toutes formulations mathématiques (Pace, Cluzel, 1982). La moyenne arithmétique est à priori la donnée représentative exceptés les cas où apparaissent des valeurs disparates ou ceux où la taille de l'échantillon est faible. Dans ces cas, la médiane est la meilleure valeur centrale (Morice, Chartier, 1954).

La représentativité de l'expérimentateur

Comme la représentativité n'est pas une notion statistique et qu'elle est souvent citée en science expérimentale, il est logique de considérer qu'elle est une notion expérimentale. L'application montre que ce n'est pas une notion simple parce que l'expérimentateur rencontre de nombreuses difficultés lors des études. L'objet initial sur lequel devrait porter l'expérience, n'est pas l'objet expérimental et l'expérience correspond à une épreuve pour laquelle interviennent d'autres paramètres souvent non maîtrisés. C'est ainsi que la perception que l'expérimentateur a de la représentativité, dépend de la priorité qu'il donne à l'objet ou à l'épreuve. Aussi n'est - elle qu'une notion qualitative. A l'inverse, GY, avec une approche globale, en fait une grandeur quantifiable.

La représentativité, notion qualitative

L'analyse des normes Afnor montre qu'il n'y a pas une représentativité mais plusieurs telles que la représentativité de l'échantillon, la représentativité en relation avec la composition de l'objet :
– « un échantillon est représentatif lorsque pour une propriété ou des propriétés que l'on veut mesurer, il manifeste les mêmes caractéristiques que la matière dont il est issu » (X31 210)
– « un échantillon représentatif est supposé avoir la même composition que la matière échantillonnée quand celle-ci est considérée comme un tout homogène » (T20 080).
– « l'essai de lixiviation… met en œuvre… des opération de broyage de certains déchets qui peuvent avoir pour conséquence un comportement de l'échantillon lors de l'essai très peu représentatif de l'évolution réelle du déchet dans l'environnement » (X31 210 -sept 88)
Ces définitions expriment intuitivement que la représentativité correspond à une qualité de l'échantillon, voire une propriété mais n'arrivent pas à la quantifier contrairement à GY.

La représentativité, notion quantitative

Gy (GY, 1988) définit la représentativité à partir de deux notions, la justesse et la reproductibilité et d'une grandeur, l'erreur d'échantillonnage ou EE. EE est la différence relative entre le résultat final a_e, obtenu sur l'objet expérimental ou échantillon et le résultat attendu a_L : $EE = (a_e - a_L)/a_L$.
L'échantillon est juste si l'espérance de EE définie par E (EE) est nulle ou quasiment nulle : E (EE) = 0 ou E (EE) = e avec e un biais de faible valeur.
La justesse est vérifiée par l'équiprobabilité des prélèvements des constituants de l'échantillon, l'échantillonnage correspondant est dit correct.
L'échantillon est reproductible si la variance de EE ou V (EE) est inférieure à un seuil préalablement défini par l'expérimentateur.
Il est représentatif si il est juste et reproductible, ce que résume la moyenne quadratique de EE : $E(EE^2) = V(EE) + [E(EE)]^2$ qui après développement, donne la relation suivante : $E(EE^2) = V(a_e/m(a_e)) \cdot [m(a_e)/a_L]^2 + [(m(a_e)/a_L)-1]^2$
où $[(m(a_e)/a_L)^{-1}]^2$ est le biais e.
En effet, avec GY, l'objet expérimental est équivalent à l'individu moyen. L'individu moyen est soit l'individu dont le résultat expérimental est égal à la moyenne des différents résultats expérimentaux, soit l'individu dont les constituants sont issus d'un prélèvement équiprobable dans l'ensemble reconstitué et homogénéisé des différents objets expérimentaux sur lesquels aurait du être effectuée l'épreuve. Aussi le résultat obtenu sur l'échantillon, ae, est équivalent à m (ai), ai étant le résultat expérimental de l'objet expérimental en tant qu'individu i si l'épreuve avait été effectuée sur cet objet. Par voie de conséquence, ae est l'estimateur de aL. Suivant ce principe, EE est l'estimateur de 0 (zéro), expliquant à la fois la relation de la justesse de l'échantillon et la représentativité par le biais de l'erreur quadratique moyenne simplifiée.
La relation précédente montre que la représentativité est d'autant meilleure que le biais est faible et elle l'est d'autant plus que la variance relative sur le résultat expérimental est petite. Comme l'application de la représentativité sert à la comparaison référence/résultat expérimental, il est important de tenir compte de la relativité de ces données. Par exemple le cas décrit par la figure 2 pose le choix entre une opération biaisée et une sans biais. Il est évident qu'il vaut mieux choisir l'opération biaisée.

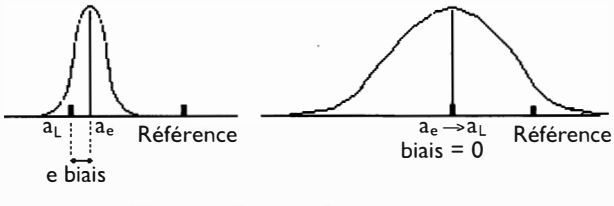

Figure 2 : Choix de l'opération et de sa représentativité

En général, e n'est pas connu. Il est donc préférable d'opérer suivant un schéma de prélèvement équiprobable. En effet l'inégalité de Bienayme-Tchebicheff démontre que ae tend vers aL.
En conclusion, la représentativité généralisée de GY appliquée au schéma décisionnel correspond au coefficient de variation de la statistique obtenu sur une opération qui donne un résultat juste.

L'APPLICATION DE L'ANALYSE DES DONNÉES A LA REPRÉSENTATIVITÉ

Le fait que la représentativité soit égale au coefficient de variation cv en fait une grandeur importante. En effet si la loi

de distribution des résultats expérimentaux est connue, si la moyenne des résultats expérimentaux est égale ou supposée égale au résultat attendu et si la variance des résultats est finie, alors la représentativité permet de définir des limites de proximité en fonction d'une probabilité d'acceptation du résultat attendu par rapport au résultat expérimental à partir de la précision relative demandée. La probabilité d'acceptation Pa correspond à la surface relative comprise entre les deux limites définies par la précision relative comme le montre la figure 3 dans laquelle sont reportées trois distributions. Il est évident que plus l'écart type s sera faible, plus la probabilité d'acceptation sera élevée.

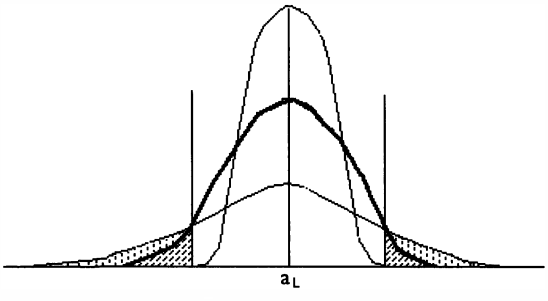

Figure 3 : **Probabilité d'acceptation et limites**

La représentativité peut être abordée de deux façons. Revenons à l'image du maillon, nous avons 2 possibilités, ou nous fabriquons les maillons ce qui nous permet d'avoir une maîtrise de la représentativité, ou on nous impose les maillons et la représentativité est alors imposée.
La représentativité est maîtrisée : Nous pouvons réduire autant que faire ce peut l'écart-type donc avoir une meilleure représentativité de l'échantillon en agissant sur les caractéristiques de l'objet. La maîtrise de la représentativité se fait par le biais d'une étude préalable consistant à déterminer la relation liant cv aux paramètres dont il dépend. Nous avons deux approches pour calculer cv. L'une des deux consiste à effectuer une étude statistique sur l'ensemble des opérations et à établir la relation probable entre cv et les paramètres expérimentaux accessibles qualitativement et quantitativement. L'autre possibilité est celle développée par GY avec la teneur des minerais et étendue à l'échantillonnage des ordures ménagères (Pineau, 1995). Dans ce cas, les caractéristiques de l'objet initial qui permettent de formaliser la propriété et de calculer la représentativité correspondante, sont connues et quantifiées ou au moins quantifiables facilement.
– La représentativité est imposée : Nous subissons la dispersion des résultats. Nous devons alors à la fois établir la relation entre la représentativité et la probabilité d'acceptation et préciser sa dépendance vis à vis de certains paramètres.
De même qu'un maillon de la chaîne a une probabilité d'avoir la résistance annoncée, la représentativité est liée à une probabilité d'acceptation. Sa mesure et la détermination de sa probabilité d'acceptation sont effectuées à l'aide de l'analyse des données.
Le résultat final est égal à la valeur attendue aL modifiée par la somme des résultats des actions des trois opérations indépendantes que sont la préparation de l'échantillon (ech), l'épreuve (épr) et la mesure (mes) :
$a_e = a_L + b_{ech} + e_{ech} + b_{épr} + e_{épr} + b_{mes} + e_{mes}$.
Chaque opération est caractérisée par son erreur e mais aussi par son biais b éventuel.

La mesure

Le cas idéal serait qu'à chaque répétition de la mesure, le résultat final soit toujours le même. Mais avec la mesure réelle, il y a plusieurs résultats finaux différents (figure 4).
Le résultat final est une variable aléatoire caractérisée par une loi de distribution.

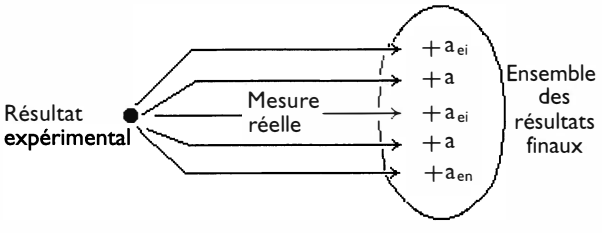

Figure 4 : La mesure réelle

Avec ce schéma, l'épreuve est répétée sur le même objet représentant le résultat expérimental mais il est des cas où l'épreuve ne peut pas l'être parce que ce dernier n'est plus réutilisable à cause du principe expérimental de la mesure.

Cas où la mesure est répétée sur le même objet représentant le résultat expérimental

Il est admis que la loi de distribution des résultats finaux est une loi de Laplace Gauss (LG) parce que la distribution expérimentale de ces résultats est semblable dans de nombreux cas à cette loi. Exceptée l'expérience de Galton (Leconte, Deltheil, 1937) qui se trouve dans des ouvrages anciens, aucune simulation de la mesure n'est proposée. Pour accepter la modélisation de la mesure par l'expérience de Galton, il faut partir de l'unité expérimentale de mesure et considérer que cette unité définit la taille de l'intervalle définie ci-après.
L'expérience de Galton consiste à faire glisser une boule sur un plan incliné parsemé de pointes réparties en quinconce (schéma a de la figure 5). En fin de parcours la boule se loge dans une case réceptrice ou intervalle.
La probabilité de rencontre (schéma b fig 5) d'une boule dans la $i^{ème}$ case est le $i^{ème}$ terme de la relation suivante une fois développée correspondant à la loi binomiale correspondante, avec ω égal 1/2 : $[\omega + (1- \omega)]^{2n}$
Comme les fonctions caractéristiques de la loi binomiale et de la loi LG convergent quand n est grand, l'hypothèse de la loi LG comme loi de distribution des résultats finaux se justifie.

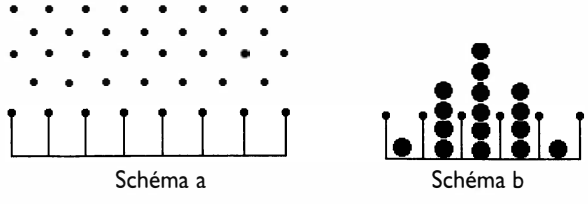

Figure 5 : L'expérience de Galton

Cas où la mesure ne peut pas être répétée sur l'objet représentant le résultat expérimental

Dans ce cas, à un objet ne correspond qu'un et unique résultat final. Ainsi l'ensemble des résultats finaux dépend à la fois de l'ensemble des objets représentant le résultat expérimental et de la mesure parce que la mesure effectuée sur l'objet n'est pas la mesure idéale. Bien qu'il soit à priori logique d'admettre que la distribution des résultats finaux suit une loi différente de celle du cas précédent, nous montrons néanmoins que les deux distributions peuvent être confondues. En effet leurs lois de distribution et leurs paramètres sont semblables parce que la loi de distribution de la moyenne suit une loi LG puisque chaque résultat de l'ensemble des résultats finaux correspondant au cas précédent suit une loi LG, parce que la moyenne des résultats finaux de ce cas est l'estimation de la moyenne des résultats finaux du cas précédent puisque chaque objet sert à estimer la même valeur, parce que la variance de la moyenne est égale à la variance des mesures puisque toutes les variances des mesures des différents résultats du cas précédent ne peuvent être qu'identiques.

Application à la précision

Comme il est admis classiquement que 100 données suffisent à décrire une loi LG, cela signifie que 9 intervalles de mesures suffisent largement pour décrire la distribution d'après l'expérience de Galton. Par conséquent le nombre de chiffre du résultat doit se limiter au dixième de la précision absolue donnée par le constructeur de l'appareil.

Application au seuil de détection

Les autres problèmes posées par la mesure sont liées aux conditions limites d'utilisation des appareils, en particulier pour les basses valeurs de concentrations et de teneurs. Le cas extrême est celui où la probabilité d'être hors mesure dépasse les 50 %, la méthode de mesure est à remettre en cause parce qu'il est impossible d'établir la distribution des résultats.

L'épreuve

En général la taille de l'objet initial n'est pas compatible avec l'épreuve. Il est décomposé en objets expérimentaux (figure 6) et est décrit par l'ensemble des résultats.

L'interrogation posée par ce schéma est celle de la correspondance entre l'ensemble des résultats expérimentaux décrit par l'ensemble des résultats finaux, indépendamment de l'erreur sur la mesure d'une part et la valeur attendue de la propriété de l'objet initial d'autre part. La réponse est dans l'analyse du schéma qui nous permet de définir les conditions dont dépend la représentativité de l'épreuve.

Le schéma montre que l'objet initial est l'union des différents objets expérimentaux, disjoints, sur chacun desquels est effectuée une seule épreuve, indépendante des autres. Dans de nombreux cas l'objet expérimental n'est pas remis dans l'objet initial soit parce qu'il n'est plus réutilisable, soit parce que cela n'est pas nécessaire pour la détermination de la propriété, à l'inverse de certaines propriétés pour lesquelles la remise est nécessaire, par exemple le nombre de poissons vivant dans un lac. C'est ainsi que la condition première pour que le schéma soit applicable, est l'indépendance de la propriété par rapport à la décomposition de l'objet initial. Cette condition dépend pour l'essentiel de l'état suivant lequel se trouve l'objet initial. Elle est vérifiée pour la lixiviation, si l'objet initial est suffisamment fragmenté. Elle ne l'est pas si l'objet initial est compact et qu'il faille le fragmenter. L'épreuve est alors biaisée. A l'inverse s'il n'y pas de biais, l'épreuve étant juste, le résultat attendu est la somme des résultats si la propriété est extensive. Si la propriété est intensive, la somme est effectuée sur les grandeurs extensives dont elle dépend et le résultat final est établi à partir de la composition des sommes des grandeurs extensives.

Si nous suivons ce raisonnement, la détermination du résultat attendu passe par une décomposition intégrale de l'objet initial en différents objets expérimentaux et par la réalisation de l'épreuve sur chacun des objets. En pratique, une telle étude n'est pas possible. Elle est remplacée par une estimation statistique du résultat attendu à partir d'un nombre restreint de résultats expérimentaux. Pour que cette application soit possible, il faut que soient vérifiées certaines conditions propres à la statistique.

La première condition est liée au facteur de pondération qui doit être le même pour tous les objets expérimentaux sur lesquels est effectuée l'épreuve. Le facteur de pondération est une grandeur extensive dont dépend la propriété étudiée et qui peut être différent suivant l'approche envisagée de la propriété, exemple la teneur. Le facteur de pondération de la teneur est la masse de l'échantillon quand celle-ci est la variable aléatoire sur laquelle nous appliquons directement la statistique, le facteur est le volume de l'échantillon avec la théorie de l'échantillonnage de GY.

La deuxième condition est l'équiprobabilité des prélèvements des différents constituants de l'objet expérimental. L'équiprobabilité est facilitée par l'homogénéisation de l'objet initial décomposé. Il arrive parfois que cette condition ne puisse pas être respectée parce qu'à la fois, l'objet initial ne peut être intégralement décomposé et qu'il existe des hétérogénéités de distribution de la propriété dans l'objet initial.

La dernière condition a trait aux paramètres expérimentaux dépendant du matériel utilisé et aux autres propriétés de l'objet qui interfèrent lors de l'épreuve. Il faut donc que

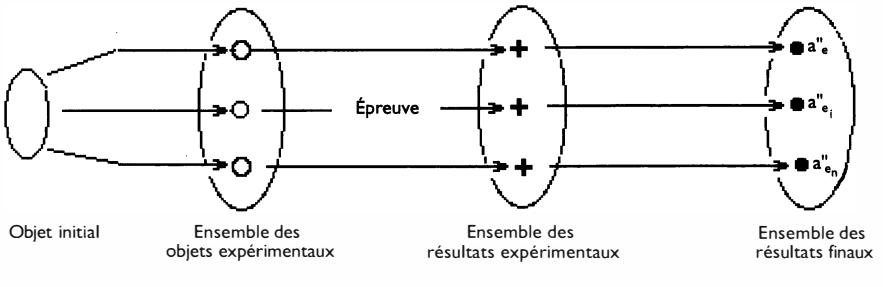

Figure 6 : L'épreuve

l'épreuve soit suffisamment ciblée, le matériel adapté et les propriétés perturbatrices connues pour que la distribution puisse permettre d'avoir une estimation fiable du résultat attendu et une variance exploitable.

L'exemple d'application a trait à l'étude de l'effet d'échelle en mécanique des roches.

Pour cette étude, la propriété est la résistance à la compression du sel estimée par la résistance à l'écrasement d'une éprouvette parallélipédique entre les deux plateaux d'une presse. La figure 7 (Mandzic, 1974) décrit les variations de la résistance de l'éprouvette en fonction de sa taille. Cette figure montre une convergence des résultats vers une limite et une diminution de la dispersion des résultats quand la taille de l'éprouvette augmente. C'est ainsi que les éprouvettes de plus grande taille sont plus représentatives que celles de taille inférieure sauf pour les éprouvettes de très petite taille puisque la dispersion est nulle. Il est évident que la représentativité des petites éprouvettes est faussée parce que les courbes ne tiennent pas compte de la rupture prématurée de certaines éprouvettes lors de leur préparation. Il manque la probabilité de rupture d'une éprouvette lors de son façonnage en fonction de sa taille.

La préparation de l'objet expérimental

Elle correspond au plan d'échantillonnage défini par la norme Afnor T 20 080 : « le plan d'échantillonnage est la marche à suivre, planifié par la sélection, le prélèvement et le traitement d'un ou plusieurs échantillons à partir d'un lot en vue d'obtenir à partir de l'échantillon final, l'information recherchée de façon qu'une décision sur le lot puisse être prise ». L'application du plan d'échantillonnage par l'Afnor est purement expérimental puisque les introductions de normes d'échantillonnage signalent que « les modes opératoires indiqués dans ce projet sont reconnus bons dans la pratique ». Il est évident qu'il est impossible d'établir un plan d'échantillonnage s'il n'y a pas un minimum de reconnaissance de l'objet initial. En effet l'estimation du résultat par l'épreuve peut être biaisée si les hétérogénéités liées à la répartition de la propriété dans l'objet initial n'ont pas été parfaitement définies, comme nous l'avons signalé précédemment avec l'épreuve. C'est ainsi que la préparation de l'objet expérimental est étroitement liée à la reconnaissance des hétérogénéités de distribution caractérisées par le niveau des covariances entre variables ; l'exemple le plus simple est celui de la somme pour laquelle la variance est égale à la somme de la variance correspondant à chaque objet expérimental et de la covariance entre objets. Par exemple pour deux objets expérimentaux X, Y, la variance de leur somme est :

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\,\text{covar}(X, Y)$$

Si les objets sont indépendants, la variance de la somme est égale à la somme des seuls variances. Néanmoins que les objets soient liés ou indépendants, la moyenne de la somme reste identique : moy (X + Y) = moy (X) + moy (Y).

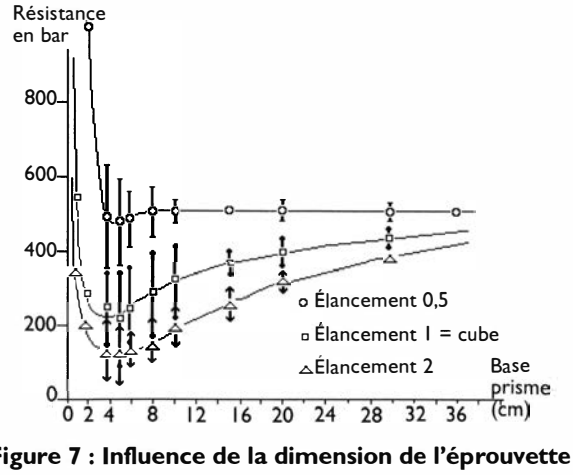

Figure 7 : Influence de la dimension de l'éprouvette sur la résistance mécanique du sel gemme (Mandzic)

CONCLUSION

La représentativité est une grandeur physique. Elle est égale au coefficient de variation pour un échantillonnage correct d'après la définition de GY. Cette correspondance statistique lui permet d'estimer la probabilité d'acceptation du résultat attendu pour un domaine fixé.

Étant égale au coefficient de variation, la représentativité dépend de la dispersion des résultats. Si cette dispersion a été reliée aux paramètres expérimentaux et aux caractéristiques de la matière, la valeur de la représentativité du futur échantillon peut être choisie. Dans le cas où la relation n'est connue soit parce qu'elle n'a pas été déterminée, soit parce que l'étude commence, la représentativité est imposée. Dans ce cas il est important d'avoir les conditions expérimentales de la préparation de l'échantillon, de l'épreuve et de la mesure avec en particulier celles liées à la taille des particules composant le lot vis à vis de l'épreuve, au facteur de pondération et à l'hétérogénéité. Si ces conditions ne permettent pas d'avoir un résultat juste ou un biais connu et une dispersion interprétable, le coefficient de variation et par voie de conséquence la représentativité n'auront aucune valeur et l'aide à la décision sera sans fondements.

*** Isabelle Barale, Jean-Louis Pineau**
LEM- UA 235 CNRS - ENS Géologie - 54500 Vandoeuvre

*** Nathalie Valentin**
LCE - Université de Provence - 13000 Marseille

Bibliographie

Abboud N, Audroing J.F. 1989 - *Probabilités et inférences statistiques*- Nathan/supérieur/économie 351 p.

Capera Ph, Van Cutsen B 1988 - *Méthodes et modèles en statistique non paramétrique, exposé fondamental*- presse de l'université Laval Dunod 358 p.

Grais B 1984 - *Méthodes statistiques* - économie module/Dunod 381 p.

Gy P 1988 - *Hétérogénéité, échantillonnage, homogénéisation, ensemble cohérent de théories*, collection mesures physiques Masson 599 p.

Leconte T., Deltheil R. 1937 - *Préparation à l'étude des probabilités* - Librairie Vuibert Paris

Mandzic E 1974 - *Effective strength of rock salt* - CR 3° cong. de la soc. intern. de mécanique des roches p 186,191.

Morice E., Chartier F 1954 - *Méthodes statistiques* deuxième partie *analyse statistique* - Insee Imprimerie nationale 555 p.

Moyhes J 1968 - *Prévisions et décisions statistiques dans l'entreprise* - 2° édit. Dunod 622 p.

Pace P., Cluzel R. 1982 - *Statistiques et probabilité* - Aide mémoire Delagrave Technor 125 p.

Pineau J.L. 1995 - *La masse de l'échantillon d'ordures ménagères en vue d'une étude descriptive quantitative* TSM n°12.

Saporta G 1990 - *Probabilités, analyse des données et statistiques* - édit. Technip 493 p.

Spiegel M.R. 1984 - *Théories et application de la statistique* - série Schaum.