REVUE DE STATISTIQUE APPLIQUÉE

J. BERNIER

Les méthodes statistiques de comparaison de deux séries de valeurs voisines

Revue de statistique appliquée, tome 12, n° 2 (1964), p. 15-25

<http://www.numdam.org/item?id=RSA_1964__12_2_15_0>

LES MÉTHODES STATISTIQUES DE COMPARAISON DE DEUX SÉRIES DE VALEURS VOISINES

J. BERNIER
Centre de Recherches et d'Essais de Chatou (E.D.F.)

La comparaison constitue un des instruments fondamentaux de toute démarche scientifique. Dans le domaine de l'Hydrométéorologie notamment elle intervient de façon constante.

- Etude de l'homogénéité des conditions d'occurence d'un certain phénomène entre deux périodes de relevés de débits ou de précipitations à une même station.

- Comparaison des débits observés à deux stations de jaugeages voisines, situées ou non sur le même cours d'eau.

- Confrontation, en vue d'une étude de régime, des débits observés à deux époques différentes de l'année.

- Etude des écarts observés sur des précipitations recueillies sur un même bassin dans des conditions opératoires différentes , par exemple avec et sans insémination préalable des nuages par l'iodure d'argent.

Pour résoudre ces divers problèmes de comparaison dont l'inventaire précédent ne donne d'ailleurs pas une liste exhaustive, on sait bien maintenant que la seule prise en compte sans précautions de moyennes empiriques calculées sur deux séries de grandeurs voisines ne suffit pas.

Il importe de considérer les fluctuations aléatoires des observations et, pour en mesurer l'incidence, de les décrire par un modèle probabiliste adéquat.

I - LA NOTION DE MODELE

Soient deux séries d'observations d'un certain phénomène hydrométéorologique :

$$x_1, x_2, \ldots\ldots\ldots x_n$$

$$y_1, y_2, \ldots\ldots\ldots y_p$$

Pour fixer les idées nous supposerons que les x et y sont des hauteurs de précipitations recueillies sur un certain bassin.

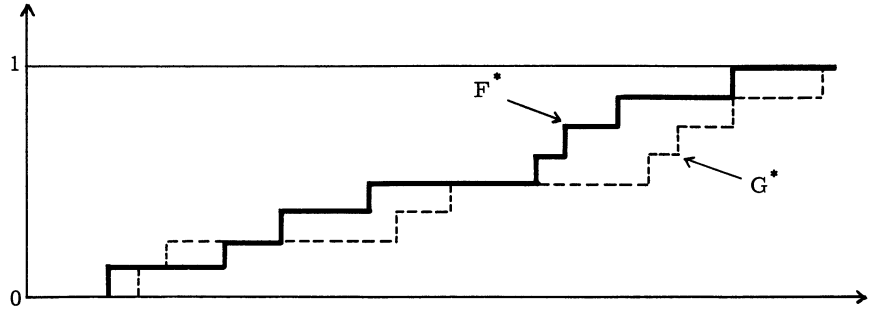
Définir un modèle, c'est admettre un certain nombre d'hypothèses concernant les lois de probabilités des variables aléatoires dont les observations constituent des échantillons et c'est représenter ces hypothèses sous une certaine forme mathématique. La considération à priori d'un tel modèle est absolument nécessaire. Cette nécessité ne répond pas seulement à un souci de rigueur mathématique dont l'hydrologue ou le météorologue n'auraient que faire. C'est aussi la seule façon d'apprécier la portée et les limites des méthodes de comparaison compte tenu des écarts qui existent toujours entre les hypothèses de base et la réalité. Pour l'avoir oublié certaines personnes ont accordé parfois une confiance exagérée et dangereuse à des calculs statistiques dont la précision mathématique peut être fallacieuse.

Les observations d'une part, le modèle c'est-à-dire les lois de probabilités du phénomène d'autre part, constituent une partie des données de base du problème de comparaison. Pour achever de décrire celui-ci il importe d'en préciser les objectifs. Il peut s'agir uniquement de savoir si les deux échantillons sont issus de la même variable aléatoire -problème du test d'homogénéité- ou d'estimer les différences si elles existent effectivement -problème d'estimation-. Dans chaque cas il arrive fréquemment que des conséquences économiques découlent des résultats de la comparaison. La considération à priori des éléments économiques conditionne alors le choix de la méthode de comparaison. Nous y reviendrons plus loin.

L'inventaire complet de ces problèmes nécessiterait des développements qui ne pourraient être présentés dans les limites du présent exposé. Nous traiterons seulement le cas des tests de l'homogénéité de deux séries entre lesquelles n'existent aucune dépendance. Il n'est pas douteux que c'est là une restriction importante pour certaines applications hydrologiques (comparaison des débits moyens de deux mois voisins par exemple). Mais il s'agit ici de dégager quelques principes généraux de l'examen de certains tests d'homogénéité.

Enfin pour simplifier l'exposé plus complètement encore, nous nous bornerons à ne considérer que des échantillons de tailles identiques, soit : $n = p$.

II - TESTS PARAMETRIQUES et NON PARAMETRIQUES



Considérons une fonction $F^*(x)$ qui croit par sauts d'amplitude constante égale à $\frac{1}{n}$ pour chaque valeur de l'échantillon des x. $F^*(x)$ donne ainsi la fréquence des observations inférieures à chaque valeur de l'échantillon, fréquence qui estime la probabilité d'obtenir une observation inférieure à cette valeur. Cet "histogramme de fréquences" constitue donc une estimation de la fonction de répartition F(x) de la variable aléatoire x. Une fonction semblable $G^*(y)$ peut être construite pour estimer la fonction de répartition G(y) des y.

L'hypothèse d'homogénéité mise à l'épreuve des observations s'exprime ici par :

$$F(x) = G(x).$$

Il semble assez intuitif qu'un test possible peut être construit à partir d'une certaine distance entre les deux histogrammes de fréquences, soit, par exemple :

$$D_n = \max_x . \left| F^*(x) - G^*(x) \right|$$

On sera conduit à rejeter l'hypothèse si cette distance est trop grande. Mais les fluctuations d'échantillonnage sont telles que les histogrammes peuvent être différents alors même que les lois F et G sont identiques. Dans ce cas, il importe donc de prendre en compte la loi de probabilité de Dn, calculée dans le cadre de l'hypothèse d'homogénéité. Cette loi, déterminée pour les grandes valeurs de n par les mathématiciens russes Kolmogoroff et Smirnoff, permet le calcul d'un seuil d_α correspondant à une probabilité α faible, tel que :

$$\text{Prob.} [D_n > d_\alpha] = \alpha$$

La technique du test consiste à rejeter l'hypothèse d'homogénéité si :

$$D_n(\text{observé}) > d_\alpha$$

* * *

Supposons maintenant que les lois de probabilités des précipitations sont gaussiennes et que leurs variances sont identiques et égales à σ. Nous admettons ainsi que la seule hétérogéneité possible réside dans une différence des espérances mathématiques que nous supposons de plus ne pouvoir être que positive en faveur de la distribution des y.

Avec ces hypothèses la mailleure mesure de distance pour détecter l'hétérogénéité éventuelle est la différence des moyennes empiriques. Dans ce qui précède l'adjectif meilleur a un sens bien précis que nous expliciterons plus loin.

L'hypothèse d'homogénéité est donc rejetée si : (*)

$$\Delta_n = \frac{\sqrt{n}\,(\bar{y} - \bar{x})}{\sqrt{s_1^2 + s_2^2}} > \delta_\alpha \qquad (2)$$

(*) La différence des moyennes est ici normée par une estimation de son écart type sur les variances empiriques s_1^2 et s_2^2 des deux échantillons.

Le seuil de signification σ_α est calculé de telle sorte que, en supposant l'homogénéité, on ait :

$$\text{Prob.}\,[\,\Delta_n > \delta_\alpha\,] = \alpha$$

La loi de probabilité de Δ_n se rattache à la loi dite de Student. Si n est grand la loi de Δ_n est voisine de la distribution de Gauss de moyenne nulle et de variance égale à 1.

* * *

Chacun des deux tests ci-dessus est caractéristique des deux grandes classes de méthodes statistiques de comparaison : d'une part les méthodes paramétriques pour lesquelles il importe de caractériser de façon précise les lois de probabilités des observations -c'est le cas du test sur les moyennes-, d'autre part les méthodes non paramétriques qui ne demandent que le minimum d'hypothèses concernant la forme des lois de probabilité ; c'est le cas du test de Kolmogoroff Smirnoff pour lequel la seule continuité des fonctions de répartition est nécessaire.

III - PUISSANCE ET ROBUSTESSE DES TESTS

Eu égard à la nature aléatoire des phénomènes étudiés, les conclusions que nous pouvons tirer des tests peuvent nous amener à commettre certaines erreurs.

- D'une part on peut rejeter l'hypothèse d'homogénéité alors qu'elle est vraie. Il résulte du principe même de la construction des tests que la fréquence de cette erreur est précisément égale à la probabilité α du seuil de signification. Le choix d'une valeur faible pour α permet ainsi de limiter cette erreur dite de première espèce.

- Mais d'autre part on peut commettre l'erreur inverse, dite de seconde espèce, à savoir : accepter l'hypothèse d'homogénéité des deux séries alors qu'elles sont hétérogènes. Pour un test donné on peut calculer la fréquence β de ce type d'erreur en fonction du seuil de signification α choisi à priori et de la forme de l'hétérogénéité éventuelle. La puissance du test définie par $1-\beta$ donne la probabilité de rejeter l'hypothèse d'homogénéité alors qu'elle est effectivement fausse. C'est donc une mesure du pouvoir de détection du test.

Avec les hypothèses de normalité des observations et d'hétérogénéité éventuelle portant sur les espérances mathématiques seules, on peut démontrer que le test basé sur les différences de moyennes possède une puissance maximale.

Le tableau ci-dessous donne la puissance de ce test obtenue pour diverses valeurs de n et diverses valeurs de la différence éventuelle μ des espérances mathématiques exprimée en pourcentage de l'écart type σ, et pour un risque de première espèce égal à 5 %.

TABLEAU I

μ \ n	25	50	100	500
10	0,10	0,13	0,18	0,48
40	0,41	0,64	0,88	$> 0,99$

Ce tableau illustre deux des propriétés importantes de la puissance,

- la croissance de la puissance lorsque la taille des échantillons augmente.

- la croissance de la puissance avec la distance éventuelle entre les lois de probabilités des deux échantillons.

Aucune méthode propre à calculer avec assez d'exactitude la puissance du test Kolmogoroff Smirnoff n'est actuellement connue. Mais dans le cadre des hypothèses de normalité développées plus haut, la puissance de ce test est certainement inférieure à celle du test sur les moyennes puisque celui-ci possède une puissance maximale. Il semblerait même qu'elle soit assez nettement inférieure. Etant donné que les distributions empiriques proches de la loi normale se rencontrent assez fréquemment on pourrait penser que le test sur les moyennes présente une utilité pratique plus grande que le test de Kolmogoroff Smirnoff.

Cependant un modèle, aussi précis soit-il, ne prétend pas représenter exactement la vraie loi du phénomène. Il existe toujours une distance plus ou moins grande entre la réalité et ce modèle et le test sera plus ou moins robuste selon la plus ou moins grande sensibilité de ses résultats à une déviation des hypothèses de base.

Par construction le test de Kolmogoroff Smirnoff est très robuste puisqu'il ne demande aucune hypothèse précise, autre que la continuité des distributions théoriques. Par contre, la robustesse du test sur les moyennes reste à prouver. Cette robustesse peut d'ailleurs prendre plusieurs aspects :

- Le risque de première espèce α peut varier si la loi de prebabilité réelle des observations n'est pas celle qui a servi à construire le test.

- La puissance $1-\beta$ peut varier si l'hétérogénéité éventuelle n'est pas celle qui a été prise en compte dans la construction du test.

Or on peut montrer, tout au moins pour de grands échantillons, que le risque α est peu sensible à un écart entre la loi réelle et la loi normale.

Par contre la puissance est très sensible à une déviation de l'hypothèse d'hétérogénéité éventuelle portant uniquement sur les espérances mathématiques. Si en fait l'écart entre des distributions des x et des y

est imputable aux seules variances, il est d'ailleurs clair que la différence des moyennes est un très mauvais test d'homogénéité. Quelle que soit la déviation et si grande soit elle, la puissance est dans ce cas pratiquement constante et égale au risque α.

Il existe un test d'égalité des variances des deux échantillons et basé sur le rapport des variances empiriques des deux échantillons :

$$F = \frac{s_1^2}{s_1^2}$$

Cependant la robustesse (sous son premier aspect) de ce test est très faible et si la loi réelle présente des écarts même faibles à la loi normale, le risque de première espèce α réel peut-être très différent de la valeur choisie. On ne peut donc plus contrôler les risques d'erreurs.

Dans ces circonstances, il n'est pas douteux que le test de Kolmogoroff Smirnoff plus robuste présente des avantages certains malgré sa puissance plus faible.

On voit donc par les considérations précédentes que le choix d'un test d'homogénéité ne doit pas être basé sur des règles absolues telles qu'on pourrait les trouver dans un manuel mais il doit surtout résulter d'une connaissance pratique des conditions d'occurence du phénomène étudié. Ce choix résultera toujours d'un compromis entre la robustesse et la puissance qui se présentent en fait comme deux objectifs contradictoires.

IV - ACCROISSEMENT DE PUISSANCE ET ECONOMIE d'OBSERVATIONS

Prenons l'exemple des opérations d'inséminations de nuages destinés à accroitre les précipitations. Il y a tout lieu de penser que les augmentations, si elles existent, sont relativement faibles.

L'important est donc d'utiliser des tests ayant une puissance suffisante pour permettre la détection de tels écarts sans qu'il soit nécessaire pour cela de prendre en compte des échantillons de taille prohibitive.

Il existe un moyen propre à augmenter de façon sensible la puissance des tests : c'est l'utilisation de variables témoins comme la hauteur des précipitations mesurés sur une zone "témoin" conjointement aux hauteurs de précipitations inséminées et non inséminées par l'iodure d'argent et relevées sur la zone "cible". Dans ces circonstances les tests possibles sont, dans leur principe, équivalents à une comparaison entre les écarts des hauteurs de précipitation de la cible à leurs moyennes liées, conditionnées par les précipitations observées sur le témoin. La puissance des tests est alors d'autant meilleure que la corrélation cible-témoin est plus grande.

Pour préciser cela nous remarquerons que, dans le cadre d'un test donné et pour un seuil α fixé, il existe une taille n de l'échantillon permettant d'atteindre une puissance fixée. Si on se fixe un risque β d'erreur de seconde espèce, de deux tests le plus puissant demandera moins d'observations que l'autre. On peut montrer que le nombre d'observations nécessaires pour atteindre une puissance quelconque est proportionnel à $1-\rho^2$ (ρ étant le coefficient de corrélation entre les hauteurs de précipitations de la cible et du témoin).

Le tableau ci-dessous donne l'économie (en % du nombre d'observations) obtenue pour différents coefficients de corrélation par rapport au cas ou ρ est nul (absence de témoin).

TABLEAU II

ρ	économie en %
0,50	25
0,90	81
0,95	90
0,99	98

V - LE CONTROLE DES RISQUES d'ERREURS

L'intérêt des tests statistiques est de controler les risques d'erreur d'interprétation des résultats d'une expérimentation. Mais encore faut-il que les prémisses, sur lesquels repose le modèle probabiliste qui permet le contrôle, soient vérifiés. Le contrôle des opérations de pluie provoquée repose sur la comparaison de deux séries d'observateurs soumises ou non au traitement préalable de l'insémination par l'iodure d'argent. Il est donc nécessaire d'être assuré que les différences significatives éventuelles sont imputables au traitement effectué. De nombreux contrôles ont été effectués en comparant des hauteurs de précipitation (période d'expérience) aux hauteurs de précipitations relevées au cours d'une période antérieure dite période historique. Or il est apparu qu'une telle comparaison peut mettre en évidence des augmentations significatives de précipitations dues à une combinaison de facteurs n'ayant rien de commun avec l'insémination des nuages (implantation, mode de relevés des pluviographes, choix des épisodes pluvieux soumis à l'insémination, types de temps favorables à la pluviosité sur la cible, défavorable sur le témoin etc...).

La randomination permet de tourner cette difficulté. Les unités d'observations éventuelles (situations météorologiques favorables) définies et décelées à priori par une prévision météorologiques sont soumises ou non au traitement insémination selon le résultat d'un tirage au sort auxiliaire. Par ce procédé on n'élimine pas les facteurs pertubateurs mais on leur fait jouer des rôles symétriques sur les observations inséminées et non inséminées de sorte qu'ils n'interviennent plus pour fausser la comparaison.

VI - LES INCIDENCES ECONOMIQUES des CONCLUSIONS TIREES des TESTS de COMPARAISON.

Pour fixer les idées nous nous plaçons maintenant dans le cadre des hypothèses servant de base au test sur les moyennes ; nous supposons notamment que l'effet des insémina'ions porte sur une augmentation des espérances mathématiques c'est-à-dire des moyennes réelles des précipitations, si tant est que la notion de vraie moyenne ait un sens.

Compte tenu, d'une part du bénéfice escompté d'une augmentation éventuelle des précipitations, d'autre part du montant des investissements destinés à l'installation de réseaux de générateurs de fumées d'iodure d'argent, il existe un seuil de rentabilité m pour l'accroissement μ.

Le problème de la comparaison revient donc à choisir entre les deux hypothèses :

- H_o: $\mu < m$ auquel cas on n'investit pas dans l'installation des générateurs.

- H_1: $\mu > m$ auquel cas l'installation des générateurs est entreprise.

On obtient ainsi une partition en deux classes de l'ensemble des valeurs du paramètre μ. Pour simplifier l'exposé nous admettrons que chaque classe peut être représentée par une valeur unique et que le problème revient à confronter les deux hypothèses.

$$H_o : \mu = \mu_o$$

$$H_1 : \mu = \mu_1 > \mu_o$$

Nous voudrions souligner qu'une telle simplification ne retire rien au réalisme du modèle construit. Il est bien évident qu'eu égard à l'imprécision des éléments économiques entrant en jeu la détermination d'un seuil de rentabilité m précis est largement théorique. En pratique on conçoit qu'il puisse exister une zone d'indifférence pour les décisions à prendre, zone qui correspond ici aux valeurs de μ comprises entre μ_o et μ_1.

Nous supposons de plus que $\mu_o = 0$

Dans ces conditions à chaque décision prise, on peut associer un coût fonction de l'augmentation réelle des précipitations. La situation est résumée de façon simplifiée dans le tableau des coûts ci-dessous :

TABLEAU III

effet réel / décision	$\mu = 0$	$\mu = \mu_1$
accepter H_o	O	B
accepter H_1	A	O

La donnée de ce tableau conjointement à l'ensemble des décisions, l'ensemble des observations et l'ensemble des lois de probabilités qui les régissent achève la description des éléments formels de la théorie des fonctions de décision statistique de Wald qui a été exposée dans une autre communication.

Revenons au test sur les moyennes ; on accepte H_1 si :

$$\bar{y} - \bar{x} > k$$

Dans la terminologie de Wald ce test est la fonction de décision. Le choix de la limite k qui, précédemment, a résulté d'un risque d'erreur de première espèce α fixé à priori, peut maintenant être basé sur des considérations économiques.

A k fixé correspond une valeur de α et une valeur de β fixes. Dans ces conditions la fonction de risque, qui représent l'espérance des coûts encourus, s'écrit :

$$r = A\,\alpha \qquad \text{si} \qquad \mu = 0$$

$$r = B\,\beta \qquad \text{si} \qquad \mu = \mu_1$$

Au sens économique le meilleur test est celui qui minimise la fonction de risque r. Or cette fonction dépend de μ inconnu à priori. Pour tourner la difficulté le principe de Bayes peut être utilisé qui repose sur le choix de probabilités à priori pour les deux hypothèses : par exemple p pour H_o et 1 - p pour H_1. On peut alors calculer un coût moyen : par exemple ρ pour H et $1-\rho$ pour H_1. On peut alors calculer un coût moyen :

$$\bar{r} = p\,\alpha\,A + (1 - p)\,\beta\,B$$

Le meilleur test sera celui pour lequel r est minimum.

Un calcul simple montre que la condition nécessaire et suffisante pour cela est :

$$k = \frac{\mu_1}{2} + \frac{\sigma^2}{2\,n\,\mu_1}\ \text{Log}\ \frac{p\,A}{(1 - p)\,B}$$

Cette formule est instructive car elle met en évidence le rôle important joué par les probabilités à priori dans le choix du seuil de signification qui est une fonction croissante de p.

L'intérêt essentiel de la théorie de Wald est de fournir un cadre formel pour représenter les problèmes pratiques de décision mieux adapté que la théorie classique basée sur le choix à priori de la probabilité α du risque de premier espèce. L'arbitraire du choix de α est en fait reporté sur le choix des probabilités à priori. C'est ainsi que dans le problème du contrôle des opérations de pluie provoquée, les résultats obtenus ont très

souvent reflétés les opinions à priori de leurs auteurs. Une analyse basée sur la théorie de Wald les aurait certainement enclains à plus de prudence dans leurs conclusions. Dans l'état actuel des choses, il apparaît que si l'efficacité de l'iodure d'argent a été décelé en laboratoire, les physiciens de l'atmosphère sont loin d'avoir donné des preuves de son action dans la Nature. Le contrôle statistique doit donc être basé sur une probabilité à priori 1 - p assez prudente telle qu'aucune conclusion en faveur de l'efficacité ne peut présentement être avancée.

VII - CONCLUSIONS

Les considérations développées dans cet exposé et que nous avons présentées à l'occasion du problème du contrôle des opérations de pluie provoquée sont transposables à tout problème de comparaison statistique. Par suite de leur présentation et du matériel mathématique mis en oeuvre les méthodes statistiques bénéficient souvent du préjugé de rigueur qui s'attache, avec raison, à tout raisonnement mathématique. La rigueur est en fait relative au calcul précis des risques d'erreurs, dans le cadre d'un modèle probabiliste retenu. Elle permet de contrôler ces erreurs et d'éclairer le choix des décisons à prendre. Elle ne permet pas d'éliminer ces erreurs. Bien que faisant constamment appel aux mathématiques, la statistique n'est qu'une science appliquée.

Il importe donc de tenir compte à tout instant de la distance entre la réalité et les modèles plus ou moins schématiques que propose le calcul des probabilités.

DISCUSSION

(Président : M. MORLAT)

M. le Président remercie M. Bernier de son exposé.

M. Sneyers demande s'il existe une mesure chiffrée de la robustesse d'un test.

M. Bernier répond que cette robustesse est assez difficile à chiffrer parce qu'elle peut prendre des aspects très divers selon l'écart entre la loi exacte des phénomènes (si tant est que cette expression ait un sens) et les modèles que l'on utilise. On pourrait essayer de chiffrer, par exemple, une certaine mesure des variations des risques α et β, mais rien n'a encore été fait dans ce domaine.

M. Le Président confirme que la considération de la robustesse d'un test est assez récente, malgré son importance que M. Sneyers a eu raison de souligner. On a déjà, dans un test, l'occasion, dès le départ, de plonger le modèle probabiliste envisagé dans une famille plus large ; ensuite, pour mesurer la robustesse, il faut généraliser le modèle initial d'une autre façon, moyennant quoi on peut certainement donner une réponse satisfaisante à la question de M. Sneyers. Mais M. Morlat ne connaît pas de littérature sur la question.