TEST DE DIXON
RECHERCHE DE VALEURS ABERRANTES

Extrait du référentiel du BTSA ANABIOTEC, module M53 :

Objectif 4.5 : Repérer des valeurs aberrantes, test de Dixon.
Recommandation pédagogique : ce test permet d'écarter des valeurs aberrantes. On traitera le cas d'une valeur aberrante ou de plusieurs.

Préambule

Quiconque voulant découvrir le test de Dixon va vite se trouver confronté à un obstacle : la multiplicité des sources, des méthodes, notations et tables.

L'objectif de cet article est de proposer une méthode simple à comprendre et à utiliser au niveau BTSA, afin d'uniformiser les pratiques pédagogiques à ce niveau.

Un petit peu d'histoire

En 1951, R. B. DEAN, and W. J. DIXON dans leur article *Simplified Statistics for small Numbers of Observations* s'intéressent à ce qu'ils appellent les "extraneous values". Traduisons "extraneous" : "sans grande portée", "superflu", "étranger". Ces "extraneous values" sont ce que nous appelons de nos jours les valeurs aberrantes. Quelques années plus tard (1969), dans les travaux de Grubbs, nous pouvons trouver une définition de cette notion, "outlier " dans le texte :

Valeur aberrante : observation qui semble dévier de façon marquée par rapport à l'ensemble des autres membres de l'échantillon dans lequel elle apparaît.

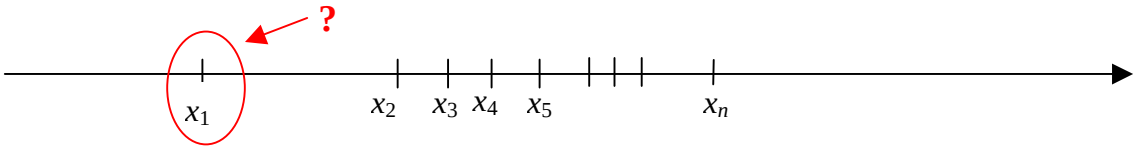
Le contexte

Au cours d'une expérimentation, il peut arriver qu'un des résultats semble s'écarter notablement des autres. Un graphique peut être d'une grande utilité pour s'en apercevoir. Une attitude classique, que l'on rencontre trop souvent, consiste à éliminer cette valeur en la considérant comme aberrante. Une bonne attitude à avoir est d'essayer de trouver la cause de l'écart (erreur de lecture, faute de calcul, etc) ; dans ce cas, il est tout à fait normal de l'éliminer. En revanche, si aucune cause accidentelle n'a pu être détectée, on s'abstiendra d'éliminer brutalement la valeur incriminée. Pour cela, il faut avoir recours à un test statistique permettant de justifier l'élimination de la valeur aberrante avec un risque de se tromper choisi au préalable. Le test de Dixon, que nous allons exposer, permet de réaliser cela, sous condition de normalité du caractère.

Principe du test

Notons tout d'abord qu'il peut s'appliquer aussi bien pour une série statistique à une variable (x_i) que pour une série statistique bivariée $(x_i ; y_i)$.

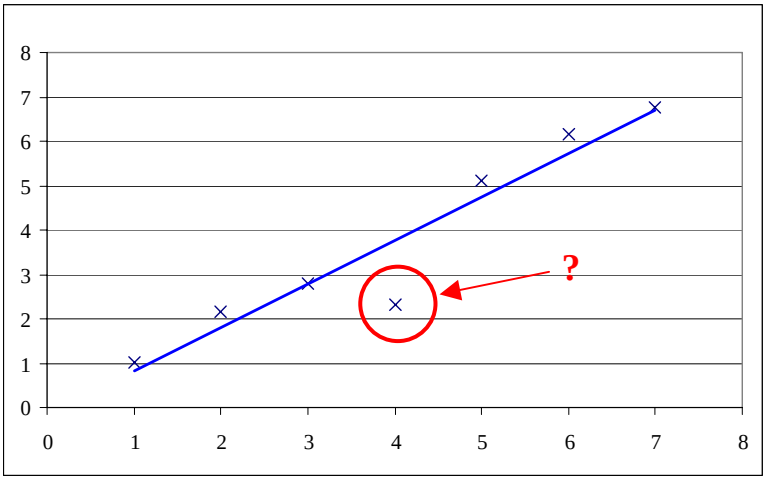
Dans le premier cas, les valeurs x_i étant rangées dans l'ordre croissant, le test de Dixon va détecter la (ou les) valeur(s) aberrante(s), aux extrémités de la distribution.



Si la valeur aberrante suspectée est très supérieure aux autres (à droite du graphique), les valeurs peuvent être alors classées dans l'ordre décroissant.

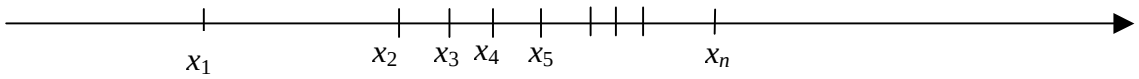
Dans le second cas, les observations sont représentées par un nuage de points dispersés autour de la droite de régression de y en x d'équation $y = a\,x + b$ (obtenue par la méthode des moindres carrés), le test est basé sur la distribution des résidus.

Ces derniers sont notés, pour tout entier i, $e_i = y_i - \hat{y}_i$, c'est-à-dire $e_i = y_i - (a\,x_i + b)$.



I. Cas d'une seule valeur aberrante

Les valeurs observées sont classées par ordre croissant et notées x_1, x_2, ..., x_n.



Hypothèses

H_0 : "La valeur douteuse n'est pas une valeur aberrante."

H_1 : "La valeur douteuse est une valeur aberrante."

Variable de décision utilisée

Il faut comparer la distance entre la valeur suspectée aberrante et une valeur des plus proches, avec la distance entre la valeur suspectée aberrante et une des valeurs les plus éloignées de l'échantillon.

Notons R la variable aléatoire prenant pour valeur le rapport de ces distances. Sa valeur observée est donnée dans le tableau ci-dessous selon la valeur de n et la position de la valeur suspectée aberrante :

	la valeur suspectée aberrante est x_1	la valeur suspectée aberrante est x_n
$n \leq 10$	$R_{obs} = \dfrac{x_2 - x_1}{x_n - x_1}$	$R_{obs} = \dfrac{x_n - x_{n-1}}{x_n - x_1}$
$n > 10$	$R_{obs} = \dfrac{x_3 - x_1}{x_{n-2} - x_1}$	$R_{obs} = \dfrac{x_n - x_{n-2}}{x_n - x_3}$

Remarque

- Plus la valeur observée de R est élevée, plus la valeur suspectée est aberrante.
- On distingue $n \leq 10$ et $n > 10$ pour détecter les cas où il y a plus d'une valeur aberrante (voir troisième exemple suivant).

Valeur critique

On se fixe un seuil de risque α. La valeur critique est notée $r_{1-\alpha}$, elle est définie par : $P(R \leq r_{1-\alpha}) = 1 - \alpha$ et elle est donnée par la table en fin d'article.

Exemple d'utilisation de la table : $n = 8$ et $\alpha = 0{,}01$.

Dans le cas de la recherche d'une valeur aberrante, la table de Dixon indique que pour $n = 8$ et $\alpha = 0{,}01$, la valeur critique est $r_{0,99} = 0{,}59$.

Cela signifie que si l'on prélève aléatoirement un échantillon de taille 8 dans une population dans laquelle les données sont distribuées normalement alors la probabilité que R prenne une valeur inférieure ou égal à 0,59 est 0,99.

Règle de décision

Si $R_{obs} > r_{1-\alpha}$, on rejette H_0, donc la valeur suspectée est aberrante.

Si $R_{obs} \leq r_{1-\alpha}$, on n'est pas en mesure de rejeter H_0.

II. Un peu de pratique

Voici trois exemples d'application.

- Un premier sur une situation classique dans laquelle la valeur la plus élevée apparaît aberrante.
- Un second montrant un point aberrant au sein d'un nuage.
- Puis un troisième exemple dont le but est de montrer une situation dans laquelle on justifie la distinction entre $n \leq 10$ et $n > 10$ et qui montre qu'il peut exister deux valeurs aberrantes (cas traité dans la seconde partie de l'article).

Exemple 1

Dans la fabrication de comprimés effervescents, il est prévu que chaque comprimé doit contenir 1 625 mg de bicarbonate de sodium. Afin de contrôler la fabrication de ces médicaments, on a prélevé un échantillon de 10 comprimés et on a mesuré la quantité de bicarbonate de sodium en mg pour chacun d'eux. Les résultats obtenus sont résumés dans le tableau suivant:

1 620 1 621 1 623 1 628 1 633 1 635 1 637 1 641 1 643 1 659

On peut demander aux étudiants de réaliser un graphique sur un axe gradué pour détecter quelle(s) valeur(s) semble(nt) aberrante(s).

On effectue un test de Dixon au seuil de risque 0,05 pour tester si la valeur supérieure 1 659 est aberrante.

On teste les deux hypothèses :

H_0 : "1 659 n'est pas une valeur aberrante."

H_1 : "1 659 est une valeur aberrante."

$n = 10$ donc on utilise la variable aléatoire R qui prend comme valeur observée $R_{obs} = \frac{x_n - x_{n-1}}{x_n - x_1}$, soit $R_{obs} = \frac{x_{10} - x_9}{x_{10} - x_1}$ qui est égale à 0,410.

D'après la table, la valeur critique est $r_{0,95} = 0{,}412$. Comme $0{,}41 < 0{,}412$: on n'est pas en mesure de rejeter H_0. La valeur 1 659 ne peut pas être considérée comme aberrante, au seuil de 0,05.

Exemple 2

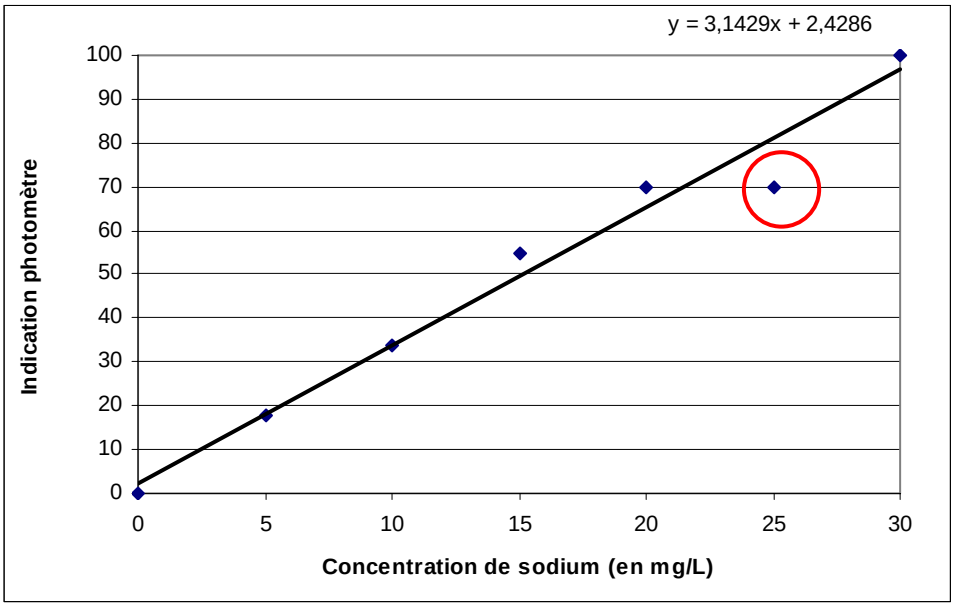
Lors d'un dosage de sodium par photométrie de flamme, on a procédé à un étalonnage (fond de flamme à 0 et solution concentrée à 100).

Les mesures figurent dans le tableau suivant :

Concentration de sodium (en mg/L) : *X*	0	5	10	15	20	25	30
Indication du photomètre : *Y*	0	18	34	55	70	70	100

La valeur observée pour une concentration de 25 mg/L peut-elle être considérée comme aberrante ?

Un petit coup d'œil sur le graphique :



On détermine l'équation de la droite d'ajustement de Y en X par la méthode des moindres carrés : $y = 3{,}1\,x + 2{,}4$.

X	0	5	10	15	20	25	30
Y	0	18	34	55	70	70	100
Estimation : $\hat{Y}$	2,4	17,9	33,4	48,9	64,4	79,9	95,4
Résidus : *e*	– 2,4	0,1	0,6	6,1	5,6	– 9,9	4,6

Classons les résidus par ordre croissant :

i	1	2	3	4	5	6	7
e_i	– 9,9	– 2,4	0,1	0,6	4,6	5,6	6,1

Valeur observée de *R* : $R_{obs} = \dfrac{e_2 - e_1}{e_7 - e_1} \simeq 0{,}75$.

Valeur critique au seuil de 0,05 : $r_{0,95} = 0{,}507$.

Décision : $0{,}75 > 0{,}507$, on rejette H_0 au seuil de 0,05 ce qui justifie que la valeur suspectée est aberrante.

Exemple 3

Une entreprise étudie la possibilité de lancer sur le marché un yaourt à la rhubarbe. Elle réalise des mesures de pH sur un échantillon de 11 pots. Les mesures observées sont les suivantes :

5,40 5,70 6,15 6,16 6,18 6,25 6,43 6,45 6,45 6,60 6,75

Existe-il une valeur aberrante ?

Dans un premier temps, nous allons effectuer un test de Dixon au seuil de risque 0,05 sur la valeur $x_1 = 5{,}40$ de manière ensuite à justifier la distinction qui doit être faite entre $n \leq 10$ et $n > 10$ pour la valeur observée de R.

Le nombre d'observations est ici 11 qui est supérieur à 10, que se passerait-il si on utilisait la valeur observée du cas $n \leq 10$?

$\frac{x_2 - x_1}{x_{11} - x_1} \simeq 0{,}222$ (à 10^{-3} près).

Bien que nous ne disposions pas de la valeur tabulée pour $n = 11$, il semble évident que la valeur critique $r_{0,95}$ serait largement supérieure à 0,222. Il faudrait donc en conclure que 5,40 n'est pas une valeur aberrante.

Cependant, si on élimine cette valeur de l'échantillon et que l'on effectue un test de Dixon au seuil 0,05 sur la valeur $x_2 = 5{,}70$ en considérant les 10 valeurs restantes, on observe alors que 5,70 est une valeur aberrante ($R_{obs} \simeq 0{,}429$ et $r_{0,95} = 0{,}412$)

Cette situation invite les étudiants à s'interroger sur cette anomalie car il parait évident que si la deuxième valeur est aberrante, la première l'est tout autant. L'erreur de décision qui est faite en utilisant $\frac{x_2 - x_1}{x_{11} - x_1}$ se justifie par le fait que les deux premières valeurs sont proches et toutes deux aberrantes.

On vérifie alors que l'utilisation de $\frac{x_3 - x_1}{x_{10} - x_1}$ permet de conclure à l'aberration de la première valeur ($R_{obs} \simeq 0{,}714$ et $r_{0,95} = 0{,}637$).

Pour des échantillons de taille strictement supérieure à 10, le calcul de $R_{obs} = \frac{x_3 - x_1}{x_{n-2} - x_1}$ prend en compte la possibilité d'avoir deux valeurs aberrantes inférieures (x_1 et x_2).

Cette situation est plus *rare* avec des tailles d'échantillon faibles ($n \leq 10$).

III. Cas de deux valeurs aberrantes

Pour appliquer la méthode, il faut dans ce cas que $n > 10$.

Plusieurs situations sont possibles :

1) Si les résultats douteux sont x_1 et x_n, on applique successivement le test de Dixon aux deux valeurs séparément.

2) Si les deux résultats douteux sont "du même côté", on applique le test à l'avant dernière, après avoir éliminé provisoirement la dernière (comme dans l'exemple 3).

Concrètement, s'il s'agit de x_1 et x_2, après avoir éliminé x_1, on applique le test à x_2 en prenant $R_{obs} = \frac{x_4 - x_2}{x_{n-2} - x_2}$

S'il s'agit de x_{n-1} et x_n, après avoir éliminé x_n, on applique le test à x_{n-1}en prenant $R_{obs} = \frac{x_{n-1} - x_{n-3}}{x_{n-1} - x_4}$.

Si le test conduit à considérer x_2 (respectivement x_{n-1}) comme aberrantes, alors x_1 (respectivement x_n) l'est aussi. Sinon on lui applique le test à son tour.

Complément : Test de Grubbs (hors programme)

C'est un test beaucoup plus puissant dans le cas des petits échantillons.

Il permet de rejeter deux valeurs aberrantes dans une série de mesures, ou encore de rejeter une ou deux moyennes par rapport à la moyenne générale.

Il est basé sur le calcul des résidus normalisés : $G = \frac{\overline{x} - x_1}{s}$ ou $G = \frac{x_n - \overline{x}}{s}$.

Mais ceci est une autre histoire…

Une idée, pour finir

On peut proposer ce test dans le cadre de l'objectif 4.1 du module M42 : *Explorer et mettre en œuvre les fonctions avancées du tableur pour résoudre un problème, notamment dans le domaine professionnel de l'option du BTSA.*

Cette séance de TD pourrait être l'occasion d'utiliser les fonctions RECHERCHEV(), NBVAL et SI, ainsi que des commandes de tri.

En guise d'exemple, vous pouvez trouver le fichier nous ayant permis de faire les calculs dans cet article, à l'adresse suivante : http://www.enfa.fr/r2math

Bibliographie

Article de Dean et Dixon :

http://depa.pquim.unam.mx/amyd/archivero/ac1951_23_636_13353.pdf

Table de la loi de Dixon

Valeur de $r_{1-\alpha}$

n \ α	**0,01**	**0,05**
3	0,988	0,941
4	0,889	0,765
5	0,780	0,642
6	0,698	0,560
7	0,637	0,507
8	0,590	0,468
9	0,555	0,437
10	0,527	0,412
11	0,745	0,637
12	0,704	0,600
13	0,670	0,570
14	0,641	0,546
15	0,616	0,525
16	0,595	0,507
17	0,577	0,490
18	0,561	0,475
19	0,547	0,462
20	0,535	0,450
21	0,524	0,440
22	0,514	0,430
23	0,505	0,421
24	0,497	0,413
25	0,489	0,406
26	0,486	0,399
27	0,475	0,393
28	0,469	0,387
29	0,463	0,381
30	0,457	0,376