Valeurs non représentatives.[1]

1. Valeurs extrêmes, valeurs non représentatives

La plupart des distributions statistiques ont la majeure partie de leurs réalisations comprises dans un intervalle d'une largeur de 6 écarts types autour de la moyenne μ. Par exemple dans le cas de la loi normale de paramètres μ et σ, la probabilité d'obtenir une valeur dans un intervalle $[\mu - 3\sigma, \mu + 3\sigma]$ est de 99,8 %. Il ne faut pas déduire de cette propriété que l'on n'observera jamais de réalisations se trouvant en dehors de cet intervalle dans la pratique. En effet, bien que la soit probabilité de se trouver face à une telle situation soit faible, elle n'est pas nulle. De surcroît, plus l'effectif de l'échantillon est important, plus ce cas de figure est susceptible de se produire.

En effet, comme vous le savez, en notant X une variable aléatoire d'espérance μ et d'écart type σ et n le nombre de réalisations de X que l'on considère, la fréquence théorique de l'évènement considéré est

$$n \times (1 - \mathbb{P}[-3\sigma \leqslant X - \mu \leqslant +3\sigma]).$$

En considérant le cas d'un loi normale de moyenne $\mu = 0$ et d'écart type $\sigma = 1$, on obtient :

$$n \times (1 - \mathbb{P}[-3 \leqslant X \leqslant +3]) = n \times 0,00270.$$

Ainsi pour un effectif de 400 la fréquence attendue est de $1,080$.

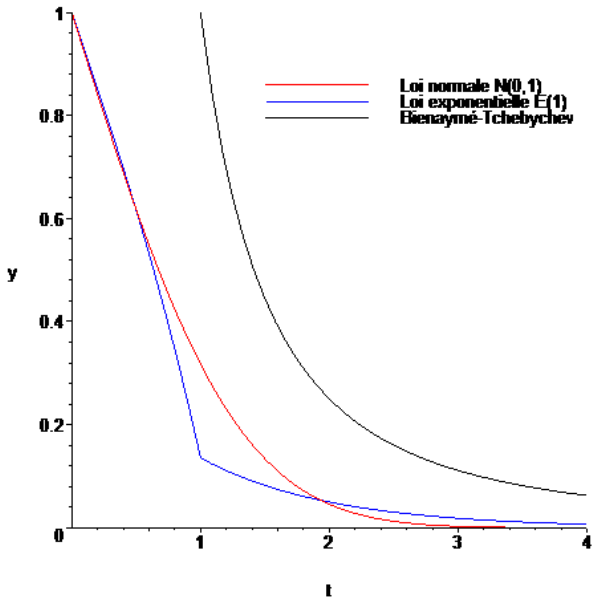
Plus généralement, l'inégalité de Bienaymé-Tchebychev, valable pour toute variable aléatoire Y admettant une variance σ_Y^2, et par conséquent une moyenne μ_Y, permet d'obtenir la relation non paramétrique suivante :

$$\Phi_Y(\lambda) \;=\; 1 - \mathbb{P}[\mu - \lambda\sigma < Y < \mu + \lambda\sigma] \;\leqslant\; \frac{1}{\lambda^2}.$$

Ainsi pour toute variable aléatoire Y dont la loi est une distribution de probabilité admettant une variance l'évènement $\mathbb{P}[|Y - \mu| \geqslant 2\sigma]$ a au plus une probabilité de $1/4 = 0,25$ et l'évènement $\mathbb{P}[|Y - \mu| \geqslant 3\sigma]$ a au plus une probabilité de $1/9 \approx 0,11$. On note que cette estimation est trop pessimiste pour un emploi pratique et est donc exclusivement réservé au cas on l'on ne connaît rien sur la loi de Y.

[1] Les références [1], [2] ayant servi à l'élaboration de ce document sont mentionnées dans la bibliographie.

Exemple 1.1.



Dans le graphique ci-contre, la valeur de la probabilité Φ_Y est réprésentée sur l'axe des ordonnées en fonction de la valeur de λ qui varie selon l'axe des abscisses.

Trois courbes ont été tracées, l'une associée à l'inégalité de Bienaymé-Tchebychev, les deux autres étant les valeurs exactes des probabilités Φ_Y pour une loi normale centrée-réduite $\mathcal{N}(0,1)$ et pour une loi exponentielle de moyenne 1, $\mathcal{E}(1)$.

2. Que vérifier ?

Lorsque que vous rencontrez ce type de valeurs dans un échantillon vous devez avoir la réaction suivante :

- L'hypothèse qui est faite sur la loi de ma variable aléatoire est-elle fondée ? Une alternative non paramétrique ou robuste n'est-elle pas préférable dans ma situation ? Certains des phénomènes que vous étudiez sont connus pour avoir des modélisations qui ont déjà été étudiées. Vous devez alors vous référer à la bibliographie pour déterminer si la modélisation que l'on fait est en accord avec les connaissances a priori que l'on a du phénomène.
- Si l'on utilise un modèle statistique, les autres hypothèses sous-jacentes au modèle sont-elles toutes vérifiées ?
- Les données sont-elles fiables ? Si vous les avez récoltées, les conditions expérimentales étaient-elles semblables à celles que vous avez fixées ou observées lors des autres essais ?
- Ne s'agit-il pas d'une simple erreur de copie et donc alors d'une valeur aberrante ?

3. Que faire avec une valeur potentiellement non-représentative ?

Si vous soupçonnez une valeur d'être une valeur non représentative, il existe plusieurs procédures de tests possibles. Attention, il ne faut pas abuser de ces pratiques et ne les utiliser que pour des cas légitimes. De plus si vous soupçonnez plusieurs valeurs d'être des valeurs non représentatives, on verra par la suite comment il faut procéder : il ne s'agit

pas simplement de répéter le traitement d'une seule valeur non reprséentative plusieurs fois de suite. Outre les problèmes d'évaluation du risque de première espèce, la présence d'autre valeurs non représentatives peut fausser les résultats des tests.

3.1. Notations

Soit $\boldsymbol{x} = (x_1, \ldots, x_n)$ un $n-$échantillon d'une variable aléatoire X.
On note $x_{(1)}, \ldots, x_{(n)}$ les valeurs $x_1, \ldots, x_n$ **ordonnées** dans l'ordre croissant ($x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}$). On note $\boldsymbol{x}_{(\bullet)}$ ce nouvel échantillon.
Le classement des valeurs de $\boldsymbol{x}$ ne change ni la valeur de la moyenne de l'échantillon, notée $\bar{\boldsymbol{x}}$, ni celle de l'écart type de l'échantillon, noté $s(\boldsymbol{x})$. En effet on montre que :

$$\begin{aligned} \bar{\boldsymbol{x}} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n} x_{(i)} = \bar{\boldsymbol{x}}_{(\bullet)}, \\ s(\boldsymbol{x}) &= \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(x_{(i)} - \bar{\boldsymbol{x}}_{(\bullet)}\right)^2 = s(\boldsymbol{x}_{(\bullet)}). \end{aligned}$$

On appelle **variation** d'un échantillon $\boldsymbol{x}$ la quantité $\mathrm{VA}(\boldsymbol{x})$:

$$\begin{aligned} \mathrm{VA}(\boldsymbol{x}) &= \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ &= (n-1)\, s^2(\boldsymbol{x}) = (n-1)\, s^2(\boldsymbol{x}_{(\bullet)}) = \mathrm{VA}(\boldsymbol{x}_{(\bullet)}). \end{aligned}$$

La variation n'est pas affectée par le classement des valeurs dans l'ordre croissant.

On appelle **somme des carrés** d'un échantillon $\boldsymbol{x}$ la quantité $\mathrm{SC}(\boldsymbol{x})$:

$$\mathrm{SC}(\boldsymbol{x}) = \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_{(i)}^2 = \mathrm{SC}(\boldsymbol{x}_{(\bullet)}).$$

La somme des carrés n'est pas affectée par le classement des valeurs dans l'ordre croissant.

On considèrera **systématiquement** dans ce cours que X suit une **loi normale**. Il est possible de réaliser les mêmes types de tests pour des variables aléatoires qui suivent des lois autres que la loi normale ; il faut alors utiliser d'autres valeurs critiques que celles qui vous ont été fournies.

3.2. Test de Grubbs pour une seule valeur non représentative

On construit les trois statistiques suivantes :

$$T = \max\left(T_1, T_n\right),$$

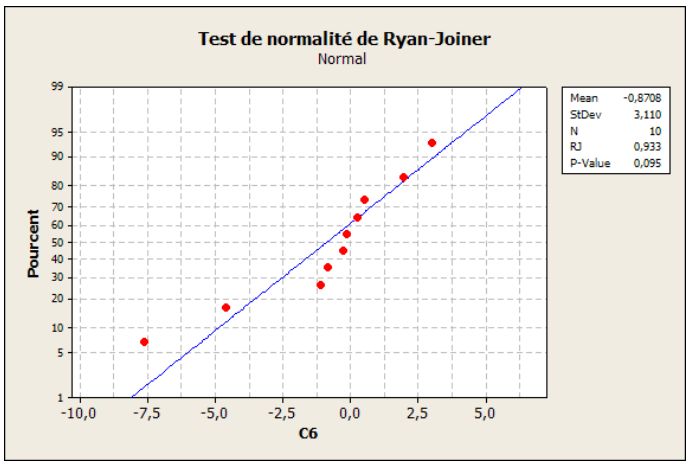
avec

$$T_1 = \frac{(\bar{x} - x_{(1)})}{s},$$
$$T_n = \frac{(x_{(n)} - \bar{x})}{s},$$

où $\bar{x}$ est la moyenne de l'échantillon $\boldsymbol{x} = (x_1, \ldots, x_n)$ et s est son écart type, donc corrigé. On utilise T_1 ou T_n si l'on suspecte la présence d'une valeur non représentative dans une direction donnée et T si l'on n'a pas d'idée a priori sur la direction dans laquelle il pourrait y avoir une valeur non représentative. En effet T_1 mesure la déviation de la plus petite valeur de l'échantillon par rapport à la moyenne standardisée par l'écart type corrigé de l'échantillon. De même T_n mesure la déviation de la plus grande valeur de l'échantillon par rapport à la moyenne standardisée par l'écart type corrigé de l'échantillon.

On compare alors la valeur de T_1, T_n ou T avec la valeur de référence correspondante de la table adéquate (Table de Grubbs). Ainsi pour T_1 et T_n on prendra la valeur du $(1-\alpha)$ % quantile et pour T celle du $(1-\alpha/2)$ % quantile.

Exemple 3.1.



x_i	$x_{(i)}$
0,26787	-7,61567
3,01367	-4,60385
-0,27047	-1,11060
-7,61567	-0,82072
-4,60385	-0,27047
0,54445	-0,10821
-0,10821	0,26787
1,99539	0,54445
-1,11060	1,99539
-0,82072	3,01367

$T_1 = 2,85558$ et $T_{10} = 0,79458$ donc $T = 2,85558$. La valeur de T est supérieure à celle de la table pour $\alpha = 5$ % (2,290). On a donc détecté une valeur non représentative. On en déduit que $x_{(1)} = -7,61567$ ou que $x_{(10)} = 3,01367$ est une valeur non représentative avec un risque de première espèce de 5 %. On aurait pu procéder uniquement au test de $x_{(1)}$ ou respectivement de $x_{(10)}$ à l'aide de la statistique T_1 ou respectivement de T_{10}. Pour un risque de première espèce de 5 %, la valeur critique à utiliser avec les statistiques T_1 et T_{10} est de 2,176. On en déduit que $x_{(1)} = -7,61567$ est une valeur non représentative avec un risque de première espèce de 5 %.

. . . . . .

3.3. Test de Dixon pour une seule valeur non représentative

La statistique du test de Dixon pour détecter une valeur non représentative parmi les grandes valeurs de l'échantillon, $r_{1,0}$, est le rapport entre l'écart entre les deux observations les plus grandes et l'étendue de l'échantillon.

$$r_{1,0} = \frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}}.$$

La statistique du test de Dixon pour détecter une valeur non représentative parmi les petites valeurs de l'échantillon, $r'_{1,0}$, est le rapport entre l'écart entre les deux observations les plus petites et l'étendue de l'échantillon.

$$r'_{1,0} = \frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}}.$$

En posant $r = \max(r_{1,0}, r'_{1,0})$, on obtient une statistique de test, r, permettant de tester la présence d'une valeur non repésentative dans l'une des deux directions. On compare alors la valeur de $r_{1,0}, r'_{1,0}$ ou r avec la valeur de référence, au niveau α, de la table.

On note qu'il s'agit du test le plus simple à réaliser en pratique puisqu'il ne nécessite que peu de calculs.

Exemple 3.2.
On reprend les données de l'exemple ci-dessus. On a alors :

$$\begin{aligned} x_{(1)} &= -7,61567, \\ x_{(2)} &= -4,60385, \\ x_{(9)} &= 1,99539, \\ x_{(10)} &= 3,01367, \\ x_{(10)} - x_{(1)} &= 10,62934. \end{aligned}$$

On calcule :

$$\begin{aligned} r_{1,0} &= 0,09580, \\ r'_{1,0} &= 0,28335, \\ r &= 0,28335. \end{aligned}$$

La valeur critique, pour $\alpha = 5$ % et $n = 10$, est de 0,412 pour $r_{1,0}$ et $r'_{1,0}$. Celle pour r est égale à 0,469.

On constate qu'aucune des réalisations des statistiques de test n'est supérieure à la valeur critique au seuil $\alpha = 5\ \%$ qui lui est associée. Les tests ne sont donc pas significatifs. Ainsi pour ces procédures de tests il n'y a pas de valeurs non représentatives.
Cette conclusion n'est pas la même que celle du test de Grubbs. Afin de savoir quel crédit on peut lui accorder on devrait calculer le risque de deuxième espèce associé à une telle décision. Malheusement sa détermination n'est pas aisée et on ne pourra le faire ici.

. . . . . .

On note que le test de Dixon proposé ci-avant ne parvient pas à détecter ni la valeur non-représentative $x_{(10)}$ de la partie supérieure de l'échantillon, ni la valeur non-représentative $x_{(1)}$ de la partie inférieure de l'échantillon. Ceci peut être dû à un effet masquant lié aux valeurs de $x_{(1)}$, $x_{(2)}$ et $x_{(9)}$.

On introduit alors les statistiques $r_{j,k}$ suivantes qui permettent de ne pas tenir compte des $j-1$ valeurs les plus fortes, $x_{(n-1)}, \ldots, x_{(n-j+1)}$, et des k valeurs les plus faibles de l'échantillon, $x_{(1)}, \ldots, x_{(k+1)}$, lors du test de la valeur $x_{(n)}$. En effet si celles-ci sont elles aussi non représentatives, cela peut fausser le résultat du test.

$$r_{j,k} = \frac{x_{(n)} - x_{(n-j)}}{x_{(n)} - x_{(k+1)}}.$$

Symétriquement, on introduit les statistiques $r'_{j,k}$ pour la détermination de la représentativité de $x_{(1)}$ sans tenir compte des valeurs les plus fortes ou les plus faibles.

$$r'_{j,k} = \frac{x_{(j+1)} - x_{(1)}}{x_{(n-k)} - x_{(1)}}.$$

Quelles valeurs choisir pour les valeurs de j et de k ?
Dixon a formulé les recommandations d'utilisation suivantes. Si l'effectif n de l'échantillon est inférieur ou égal à 7, il faut utiliser $j = 1$ et $k = 0$. Si $8 \leqslant n \leqslant 14$, il faut utiliser $j = 2$ et $k = 1$. Enfin si $n \geqslant 15$, il faut utiliser $j = 2$ et $k = 2$.

Exemple 3.3.
L'effectif de l'échantillon utilisé dans les exemples ci-dessus est de 10. On calcule donc $r_{2,1}$ et $r'_{2,1}$:

$$\begin{aligned} r_{2,1} &= \frac{x_{(10)} - x_{(8)}}{x_{(10)} - x_{(2)}} \\ &= \frac{3,01367 - 0,54445}{3,01367 - (-4,60385)} = 0,32415, \\ r'_{2,1} &= \frac{x_{(3)} - x_{(1)}}{x_{(9)} - x_{(1)}} \\ &= \frac{-1,11060 - (-7,61567)}{1,99539 - (-7,61567)} = 0,67683. \end{aligned}$$

Les valeurs critiques du test sont : pour $\alpha = 10$ %, 0,551, pour $\alpha = 5$ %, 0,612, et pour $\alpha = 1$ %, 0,726. Ainsi, même au seuil $\alpha = 10$ %, le test basé sur $r_{2,1}$ n'est pas significatif : $x_{(10)}$ n'est pas une valeur non-représentative. Par contre, aux seuils $\alpha = 10$ % et $\alpha = 5$ %, le test basé sur $r'_{2,1}$ est significatif. Ainsi après avoir éliminé l'effet masquant de $x_{(2)}$, on peut à nouveau mettre en évidence la non représentativité de $x_{(1)}$.

. . . . . .

L'exemple que l'on vient de traiter pose le problème pratique suivant : en utilisant r ou $r'_{1,0}$ les tests ne sont pas significatifs. Tandis qu'en utilisant $r'_{2,1}$, le test est significatif. Cette situation ne serait-elle pas due au fait que la valeur $x_{(2)}$ n'est pas, elle non plus, représentative. Or si, comme dans cet exemple, l'on soupçonne non pas la présence d'une valeur non représentative mais de plusieurs, il existe d'autres procédures de tests à appliquer qui seront détaillées à la section 4 et à la section 6.

3.4. Test basé sur l'étendue

La statistique du test basé sur l'étendue de l'échantillon est :

$$u = \frac{x_{(n)} - x_{(1)}}{s}.$$

où l'on rappelle que l'étendue e d'un échantillon est la longueur de l'intervalle séparant la plus petite valeur de la plus grande valeur de l'échantillon, entre d'autres termes, $e = x_{(n)} - x_{(1)}$. Ce test permet de détecter une valeur non représentative dans l'une quelconque des directions, c'est-à-dire à la fois une valeur trop faible ou trop forte. Il est néanmoins plus naturel de s'en sevir pour tester la paire de valeurs potentiellement non représentative $(x_{(1)}, x_{(n)})$ puisque l'on compare l'étendue e de l'échantillon à son écart type s.

Exemple 3.4.
En reprenant les données ci-dessus, on calcule u pour cet exemple :

$$u = \frac{3,01367 - (-7,61567)}{3,10974} = 3,41808.$$

Les valeurs critiques pour $\alpha = 5$ % et $\alpha = 10$ % sont respectivement 3,68 et 3,57. Ainsi à aucun de ces niveaux le test n'est significatif. La paire $(x_{(1)}, x_{(n)})$ n'est pas constituée de valeurs toutes les deux non représentatives, ce qui est bien cohérent avec les résultats précédents.

. . . . . .

4. Test simultané de k valeurs non représentatives

Le nombre k est ici fixé avant la procédure de test. Le cas où k serait également à déterminer sera abordé aux sections 5 et 6.

4.1. Test de Grubbs pour k valeurs non représentatives dans une direction donnée.

La statistique exposée ci-après a été proposée par Grubbs en 1950 pour $k = 2$ puis étendue par Tietjen and Moore en 1972 au cas $1 < k < n$. À la fois la queue de la distribution (valeurs fortes ou valeurs faibles) dans laquelle on procède au test et le nombre de valeurs testées k doivent être fixés à l'avance. On note L_k les statistiques associées à la détection de valeurs non représentatives parmi les fortes valeurs de l'échantillon et $L_k^\star$ celles associées à la recherche parmi les faibles valeurs de l'échantillon.

$$L_k = \frac{\sum_{i=1}^{n-k} \left(x_{(i)} - \bar{x}_{n-k}\right)^2}{(n-1)s^2}$$

où $\bar{x}_{n-k}$ est la moyenne des $n-k$ plus petites valeurs de l'échantillon, c'est-à-dire $\bar{x}_{n-k} = \frac{1}{n-k}\sum_{i=1}^{n-k} x_{(i)}$ et s^2 est toujours la variance corrigée de l'échantillon de taille n.

$$L_k^\star = \frac{\sum_{i=k+1}^{n} \left(x_{(i)} - \bar{x}^\star_{n-k}\right)^2}{(n-1)s^2}$$

où $\bar{x}^\star_{n-k}$ est la moyenne des $n-k$ plus grandes valeurs de l'échantillon, c'est-à-dire $\bar{x}^\star_{n-k} = \frac{1}{n-k}\sum_{i=k+1}^{n} x_{(i)}$.

Ainsi dans chacun des deux cas on fait le rapport de la somme des carrés des déviations à la moyenne de l'échantillon privé des k valeurs que l'on suspecte d'être non représentatives par la somme des carrés des déviations à la moyenne de l'échantillon tout entier. Ces rapports sont toujours inférieurs ou égaux à 1. Une valeur proche de 0 indiquera que la présence des valeurs testées augmente de manière conséquente la variation de l'échantillon. On comparera la réalisation du test au quantile à α %.

Exemple 4.1.
On reprend toujours les mêmes données. Les résultats des tests de Dixon ont confirmé le fait que $x_{(1)}$ est une valeur non représentative et ont amené à envisager que $x_{(2)}$ en

était aussi une. On cherche à tester deux valeurs, donc $k = 2$, dans la partie inférieure de l'échantillon et on calcule ainsi $L_2^\star$.

$$\begin{aligned} \bar{x}_8^\star &= \frac{1}{8}\sum_{i=3}^{10} x_{(i)} = 0,43892 \\ L_2^\star &= \frac{\sum_{i=3}^{10}\left(x_{(i)} - \bar{x}_8^\star\right)^2}{9s^2} \\ &= \frac{13,8826}{87,0345} = 0,15951. \end{aligned}$$

On compare ce résultat avec les valeurs critiques (associées aux quantiles inférieurs de la distribution cette fois-ci) pour un niveau de $\alpha = 1\%$ et de $\alpha = 5\%$. Après lecture dans la table on trouve respectivement $0,142$ et $0,233$. Puisque $L_2^\star = 0,160 < 0,233$ le test est significatif au niveau $\alpha = 5\%$. Il ne l'est pas au niveau $\alpha = 1\%$ car $L_2^\star = 0,160 > 0,142$. Ainsi avec un risque de première espèce de 5%, on peut conclure que les deux valeurs $x_{(1)} = -7,61567$ et $x_{(2)} = -4,60385$ sont non-représentatives.

. . . . . .

4.2. Test de Grubbs pour deux valeurs non représentatives, une de chaque côté

On utilise la même idée que celle sur laquelle repose les statistiques de test L_k et $L_k^\star$. On compare la variation totale de l'échantillon où l'on a retiré les deux valeurs $x_{(1)}$ et $x_{(n)}$ à celle de l'échantillon complet en faisant le rapport suivant :

$$\frac{S_{1,n}^2}{S^2} = \frac{\sum_{i=2}^{n-1}\left(x_{(i)} - \bar{x}_{1,n}\right)^2}{(n-1)s^2}$$

où $\bar{x}_{1,n}$ est la moyenne des $n-2$ valeurs centrales de l'échantillon initial, c'est-à-dire $\bar{x}_{1,n} = \dfrac{1}{n-2}\displaystyle\sum_{i=2}^{n-1} x_{(i)}$.

Exemple 4.2.
On continue avec les mêmes données :

$$\begin{aligned}
\bar{x}_{1,10} &= \frac{1}{8}\sum_{i=2}^{9} x_{(i)} = -0,513268 \\
\frac{S^2_{1,n}}{S^2} &= \frac{\sum_{i=2}^{9}\left(x_{(i)} - \bar{x}_{1,10}\right)^2}{9s^2} \\
&= \frac{25,4295}{87,0345} = 0,292177.
\end{aligned}$$

La valeur critique pour $\alpha = 10$ % est de 0,246. Le test n'est donc pas significatif à ce niveau. Le minimum et le maximum de l'échantillon ne sont pas tous les deux des valeurs non représentatives.

.....

Il est bien entendu possible de généraliser cette approche au cas où l'on considérerait $2k$ valeurs, k de chaque côté de l'échantillon. Malheureusement vous ne disposez pas des valeurs critiques associées à ces tests.

4.3. Test de Tietjen-Moore pour k valeurs non représentatives de l'un ou des deux côtés

Là encore on reprend l'idée qui a permis de construire L_k. On construit r l'échantillon formé des valeurs absolues des déviations par rapport à la moyenne :

$$r_i = |x_i - \bar{x}|.$$

Puis on classe cet échantillon : $r_{(1)}, \ldots, r_{(n)}$. En conservant l'ordre ainsi obtenu on multiplie chaque $r_{(i)}$ par le signe de $x_{(i)} - \bar{x}$ et on note $z_{(i)}$ les valeurs ainsi construites. Les statistiques E_k sont alors :

$$E_k = \frac{\sum_{i=1}^{n-k}\left(z_{(i)} - \bar{z}_{n-k}\right)^2}{\sum_{i=1}^{n}\left(z_{(i)} - \bar{z}\right)^2},$$

où $\bar{z}$ est la moyenne de $z_{(1)}, \ldots, z_{(n)}$ et $\bar{z}_{n-k}$ est la moyenne de $z_{(1)}, \ldots, z_{(n-k)}$.
Les $z_{(i)}$ sont simplement les déviations par rapport à la moyenne $\bar{z}$ ordonnées par valeur absolue croissante.

Exemple 4.3.
Toujours avec le même exemple :

x_i	$x_{(i)}$	r_i	$r_{(i)}$	$z_{(i)}$
0,26787	-7,61567	1,13868	0,05009	0,05009
3,01367	-4,60385	3,88448	0,23979	-0,23979
-0,27047	-1,11060	0,60034	0,60034	0,60034
-7,61567	-0,82072	6,74486	0,76260	0,76260
-4,60385	-0,27047	3,73304	1,13868	1,13868
0,54445	-0,10821	1,41526	1,41526	1,41526
-0,10821	0,26787	0,76260	2,86620	2,86620
1,99539	0,54445	2,86620	3,73304	-3,73304
-1,11060	1,99539	0,23979	3,88448	3,88448
-0,82072	3,01367	0,05009	6,74486	-6,74486

$$\begin{aligned}
\bar{z}_9 &= 0,749428,\\
E_1 &= \frac{\sum_{i=1}^{9}\left(z_{(i)}-\bar{z}_9\right)^2}{\sum_{1}^{10}\left(z_{(i)}-\bar{z}\right)^2}\\
&= \frac{36,4867}{87,0345} = 0,41922.\\
\bar{z}_8 &= 0,357546,\\
E_2 &= \frac{\sum_{i=1}^{8}\left(z_{(i)}-\bar{z}_8\right)^2}{\sum_{1}^{10}\left(z_{(i)}-\bar{z}\right)^2}\\
&= \frac{25,4295}{87,0345} = 0,29218.\\
\bar{z}_7 &= 0,941915,\\
E_3 &= \frac{\sum_{i=1}^{7}\left(z_{(i)}-\bar{z}_7\right)^2}{\sum_{1}^{10}\left(z_{(i)}-\bar{z}\right)^2}\\
&= \frac{6,30625}{87,0345} = 0,072457.
\end{aligned}$$

pour $\alpha = 1$ % et dans l'ordre $k = 2, 3, 4$, les valeurs critiques sont, 0,235, 0,101, 0,048. Ainsi aucun des tests n'est significatif. Au seuil $\alpha = 5$ %, on a, toujours dans le même ordre, les valeurs critiques 0,356, 0,172, 0,083.

Les choses se compliquent puisque maintenant seul le test basé sur E_3 est significatif. Ceci est assez cohérent avec ce qui précède puisque les deux valeurs les plus faibles de l'échantillon sont associées à la première et à la troisième déviation par rapport à la moyenne les plus fortes en valeurs absolues. En d'autres termes $z_{(i)}$ est associé à $x_{(1)}$ et $z_{(3)}$ à $x_{(2)}$. Or si l'on avait vu que $x_{(1)}$ et $x_{(2)}$ étaient non représentatives, aucun test jusqu'alors n'avait permis de conclure à la non représentativité de $x_{(10)}$.

. . . . . .

5. Procédures pour détecter un nombre de valeurs non représentatives non fixé à l'avance

5.1. La boîte à moustaches

On base une mesure de la dispersion de l'échantillon sur la longueur de l'intervalle interquartile. En procédant ainsi, on construit un indicateur qui ne dépend que des valeurs centrales de l'échantillon, la moitié du nombre totale d'entre elles pour être précis. De ce fait, l'étendue interquartile est un indicateur robuste de la dispersion de l'échantillon.

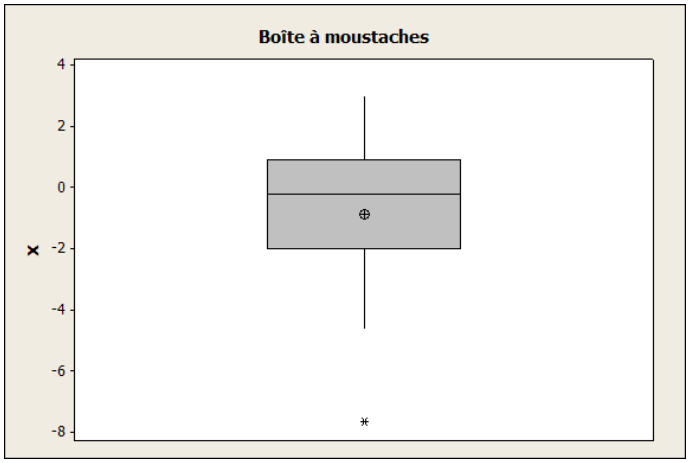
On note Q_1 le premier quartile de l'échantillon et Q_3 le troisième quartile, $IQR = Q_3 - Q_1$ est alors l'étendue interquartile. On dit qu'une valeur x de l'échantillon est une valeur extrême si $x \notin [Q_1 - 3/2IQR, Q_3 + 3/2IQR]$ et qu'il s'agit d'un type particulier de valeur extrême, une valeur très extrême, si $x \notin [Q_1 - 3IQR, Q_3 + 3IQR]$.

On a montré que si l'effectif n de l'échantillon est compris entre 20 et 75 on doit s'attendre à ce que 1 à 2 % des observations soient des valeurs extrêmes. Si l'effectif n est supérieur à 100, alors on doit s'attendre à ce que moins de 1 % des observations soient des valeurs extrêmes, la valeur limite, lorsque n tend vers l'infini convergeant vers $0,7$ %.

En ce qui concerne les valeurs très extrêmes, leur proportion est de moins de 1 % si $6 \leqslant n \leqslant 19$ et de moins de $0,1$ % si $n \geqslant 20$, la valeur limite lorsque n tend vers l'infini convergeant vers une proportion de $0,00023$ %.

L'intérêt de la boîte à moustaches est de proposer un outil visuel d'exploration des données, résistant aux effets d'écrantage car basé sur les valeurs centrales de l'échantillon et pour lequel on n'a pas à faire d'hypothèses sur la localisation ou le nombre de valeurs non représentatives présentes dans l'échantillon. Son principal défaut est qu'il ne permet pas de faire de test et donc de quantifier les risques associés aux décisions qui découlent de son utilisation. En tant que technique exploratoire, la boîte à moustaches permet de choisir des valeurs plausibles pour k et d'utiliser alors la panoplie de tests présentés dans les sections 3 et 4 précédentes et dans la section 6 à venir.

Exemple 5.1.
On continue avec les mêmes données.



On constate la présence d'une valeur extrême. Minitab ne distinguant pas les valeurs extrêmes des valeurs très extrêmes on va faire le calcul séparément.

$$\begin{aligned} Q_1 &= -1,98391 \\ Q_3 &= 0,90719 \\ IQR &= 2,89110 \\ Q_1 - 3/2IQR &= -6,32056 \geqslant x_{(1)} \\ Q_1 - 3IQR &= -10,6572 \leqslant x_{(1)} \end{aligned}$$

Ainsi $x_{(1)}$ n'est pas une valeur très extrême mais seulement une valeur extrême. De cette représentation graphique on tirerait l'information $k = 1$. Or on a vu qu'il ne s'agit pas de la seule situation qui soit envisageable.

. . . . . .

5.2. Test basé sur le coefficient d'asymétrie

La statistique b_1 du test basé sur le coefficient d'asymétrie est :

$$\sqrt{b_1} = \frac{m_3}{m_2^{\frac{3}{2}}},$$

où m_3 est le moment centré d'ordre 3 de la variable étudiée et m_2 est le moment centré d'ordre 2, c'est-à-dire la variance, de la variable étudiée.
Si $\sqrt{b_1} > 0$ les données sont plus localisées au delà de la moyenne. Si $\sqrt{b_1} < 0$ les données sont plus localisées en deçà de la moyenne.
On utilise généralement un test bilatéral sauf si l'on sait que les valeurs non représentatives ne sont présentes que d'un côté de l'échantillon.

Exemple 5.2.

$$\sqrt{b_1} = -1,25505$$

On conclut à l'aide des valeurs de la table.

. . . . . .

5.3. Test basé sur le coefficient d'aplatissement

La statistique b_2 du test basé sur le coefficient d'aplatissement est :

$$b_2 = \frac{m_4}{m_2^2}.$$

On utilise un test unilatéral dont la zone de rejet est située au voisinage de $+\infty$.

Exemple 5.3.

$$b_2 = 1,64728$$

On conclut à l'aide des valeurs de la table.

. . . . . .

6. Procédures séquentielles de détections de valeurs non représentatives

Comme le montre l'exemple 4.3, il est difficile de déterminer a priori le nombre de valeurs non représentatives ainsi que leur répartition de chaque côté. On se tourne alors vers des algorithmes itératifs pour détecter lesquelles parmi celles que l'on soupçonne sont des valeurs non représentatives. L'utilisation d'une boîte à moustaches, détaillée à la section 5, ou d'une droite de Henry, il s'agit du type de graphique qui apparaît dans l'exemple 3.1, permet généralement d'avoir une idée du nombre potentiel, donc maximal, de valeurs non représentatives présentes dans l'échantillon étudié.

6.1. Procédure séquentielle de Prescott

L'idée ici est de calculer sucessivement les rapports de sommes de carrés ce qui ressemble à ce que l'on a déjà utilisé pour les tests de Grubbs en retirant les valeurs potentiellement non représentatives l'une après l'autre. On spécifie le nombre maximal de valeurs non représentatives k à détecter ce qui permet de trouver les valeurs $\lambda_j(\beta)$ en utilisant une table. On s'en sert de la manière suivante. On commence par calculer les D_j :

$$D_j = \frac{S^2_{(j)}}{S^2_{(j-1)}} \quad \text{pour } 1, \ldots, k,$$

où $S^2_{(j)}$ est la somme des carrés de l'échantillon privé des j observations les plus éloignées de la moyenne, c'est-à-dire, en concervant les notations du 4.3, privé de $z_{(n)}, \ldots, z_{(n-j+1)}$.

Le nombre de valeurs non représentatives m est alors le plus grand entier $m \leqslant k$ tel que :

$$D_j < \lambda_j(\beta).$$

Exemple 6.1.
Attention, on calcule ici **la somme des carrés et non la variation**.

$$\begin{aligned} S^2_{(0)} &= S^2 = 94,61771 \\ S^2_{(1)} &= 36,61928 \\ S^2_{(2)} &= 27,53707 \\ S^2_{(3)} &= 6,34164 \\ D_1 &= 0,38702 \\ D_2 &= 0,75198 \\ D_3 &= 0,23029. \end{aligned}$$

Les tables disponibles ne contiennent les valeurs que pour un effectif de 10 que si $k = 2$. Alors même au niveau $\alpha = 10$ %, il n'y a pas de valeurs non représentatives, $\lambda_1(\beta) = 0,360$ et $\lambda_2(\beta) = 0,385$.
Par contre, en étudiant les tables pour $k = 2$ et $k = 3$ on se rend compte que même si les valeurs critiques pour $k = 2$ sont plus élévées que pour $k = 3$, elles restent semblables. Ainsi il est fort probable que $D_3 < \lambda_3(\beta)$ au seuil $\alpha = 2,5$ % et assurément au seuil $\alpha = 5$ %. De ce fait on détecte trois valeurs non représentatives sur le maximum de trois que l'on avait fixé ($k = 3$).

. . . . . .

6.2. Procédure RST de Rosner

À nouveau, on utilise une procédure séquentielle basée sur des statistiques R_i de Grubbs. On choisit à nouveau un nombre k, le nombre maximal de valeurs non repésentatives présentes dans l'échantillon. La différence avec la procédure de Prescott exposée au 6.1 est que l'on va utiliser une moyenne et une variance tronquée basée sur l'échantillon dans lequel on a supprimé les k valeurs les plus fortes et les k valeurs les plus faibles.

$$\begin{aligned} a &= \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)}, \\ b^2 &= \frac{1}{n-2k-1} \sum_{i=k+1}^{n-k} \big(x_{(i)} - a\big)^2. \end{aligned}$$

Maintenant si on note E_0 l'échantillon complet on appelle x^0 l'élément tel que :

$$R_1 = \max_{E_0} \left| \frac{x_i - a}{b} \right| = \left| \frac{x^0 - a}{b} \right|.$$

Puis on considère $E_1 = E_0 - \{x^0\}$ et on cherche x^1 tel que :

$$R_2 = \max_{E_1} \left| \frac{x_i - a}{b} \right| = \left| \frac{x^1 - a}{b} \right|.$$

et ainsi de suite... On remarque que les R_i sont les déviations par rapport à la moyenne tronquée a et normalisées par b classées dans l'ordre décroissant.
On compare alors les valeurs de R_i avec les valeurs critiques qui dépendent de α, n et k, i observations étant non représentatives si R_i est supérieur à la valeur critique.

Exemple 6.2.

$$\begin{aligned} a &= -0,23288 \\ b &= 0,45213 \\ R_1 &= \frac{|-7,61567 - (-0,23288)|}{0,45213} = 16,32894 \\ R_2 &= \frac{|-4,60385 - (-0,23288)|}{0,45213} = 9,66752 \\ R_3 &= \frac{|3,01367 - (-0,23288)|}{0,45213} = 7,18059. \end{aligned}$$

Les tables disponibles ne sont utilisables que si $n \geqslant 20$.

. . . . . .

7. Conclusion

Avant tout il s'agit de réaliser les deux représentations graphiques des données que sont la boîte à moustaches et la droite de Henry. Ceci fait, on aura une idée de la situation dans laquelle l'on se trouve.
Même alors, les procédures proposées restent variées. Il serait intéressant de disposer d'études de puissance afin de comparer les différents choix qui s'offrent au praticien au sein d'une même problématique.
Lorsque l'on se demande si une valeur et une seule ($k = 1$) est une valeur non représentative, le test T de Grubbs se comporte toujours le mieux. Si $k > 1$, le mieux est de confronter les résultats des différentes procédures. Si l'on souhaite seulement savoir s'il y a des valeurs non représentatives dans une direction donnée, on utilisera un test sur le coefficient

d'asymétrie. Si au contraire on suspecte la présence de valeurs non représentatives dans chacune des directions, on utilsera un test basé sur le coefficient d'aplatissement. Si l'on cherche à déterminer le nombre de valeurs non représentatives on utilisera une procédure séquentielle.
D'autre part, il est conseillé, par Grubbs, de toujours procéder au test des valeurs non représentatives à un niveau α plus conservatif que le niveau communément utilisé de 5 %, c'est-à-dire avec $\alpha < 0,05$, par exemple $\alpha = 0,01$.

Références

[1] G. Parreins. *Techniques Statistiques : moyens rationnels de choix et de décision*. Dunod technique, Paris, 1974.

[2] H.C. Thode. *Testing for Normality*. Number 164 in Statistics : textbooks and monographs. Marcel Dekker, New-York, 2002.