JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

DOMINIQUE LADIRAY

Discussion et commentaires. Approche graphique en analyse des données. Graphiquez vos données !

Journal de la société française de statistique, tome 141, n° 4 (2000), p. 61-67

<http://www.numdam.org/item?id=JSFS_2000__141_4_61_0>

DISCUSSION ET COMMENTAIRES

Approche graphique en analyse des données

Graphiquez vos données ! [1]

Dominique LADIRAY [2]

Jean-Paul VALOIS nous offre, avec son article « Approche graphique en analyse des données », une présentation synthétique et très bien documentée de l'utilisation des graphiques dans le travail journalier du statisticien. En filigrane de son exposé transparaît un regret que je partage avec lui, celui du « conflit » qui perdure entre méthodes numériques et méthodes graphiques. Ce sont à ces rapports difficiles entre « statistiques » et « graphiques » que sont consacrées les lignes qui suivent. La première partie traite de l'objectivité, ou de la subjectivité, des graphiques et des statistiques. La seconde partie est quant à elle consacrée à la notion de « bon » ou de « mauvais » graphique. A travers quelques exemples simples, et souvent bien connus, nous verrons qu'à mon avis un graphique n'est ni plus ni moins objectif qu'une statistique et qu'il n'existe a priori ni bon ni mauvais graphique. Mieux encore, il me semble que nous reprochons souvent aux graphiques des pratiques jugées « optimales » dans le cadre de la statistique mathématique. A l'heure actuelle, les méthodes graphiques se développent et se diffusent bien plus vite que les méthodes statistiques : l'enjeu est donc bien pour le statisticien de se rapprocher de ces outils et de les inclure dans ses façons de penser et d'enseigner, faute de quoi... c'est ce que nous verrons en conclusion.

1. « SUBJECTIVITÉ » DES GRAPHIQUES ET « OBJECTIVITÉ » DES STATISTIQUES ?

L'enseignement de la statistique en France, très mathématisé il faut bien le reconnaître, reste profondément marqué par l'idée « qu'un graphique n'a jamais rien prouvé ». Cette croyance se répercute jusque dans la pratique

1. Ne cherchez pas dans le dictionnaire le verbe « grafiquer », vous ne le trouverez pas ! L'analyse exploratoire des données est pleine de ces mots nouveaux : celui-ci est dû à Jacques Vanpoucke de l'Université Paul Sabatier de Toulouse (VANPOUCKE, 1991).

2. Administrateur de l'INSEE, actuellement en poste à EUROSTAT, et professeur d'analyse exploratoire des données à l'école nationale de la statistique et de l'administration économique (ENSAE) ; e-mail : Dominique.Ladiray@cec.eu.int

quotidienne du statisticien qui rechigne à utiliser des graphiques dans ses analyses, leur préférant des méthodes « optimales » souvent beaucoup plus complexes. Les graphiques sont alors réservés à la communication des résultats auprès d'utilisateurs jugés « peu au fait de la beauté et de la pureté de la statistique mathématique ».

Comme le note Jean-Paul VALOIS, cette opposition « nombre – graphique » existe depuis longtemps et c'est sans doute DESCARTES qui en est à l'origine, probablement à son corps défendant. Car il est bon de remarquer que ce grand philosophe et mathématicien est à la fois le créateur de graphiques fameux, les graphiques cartésiens, et le père d'une définition de l'objectivité [3] ! Il est donc fort probable qu'il n'associait pas automatiquement graphique et subjectif. Un siècle plus tard, un autre philosophe, VOLTAIRE, déclarait au contraire : « Qu'est-ce qu'une idée ? C'est une image qui se peint dans mon cerveau », propos partagé par le génial mathématicien EULER qui écrivait : « Je n'ai des idées que parce que j'ai des images ».

Mais comment juger du caractère objectif d'une statistique ou d'un graphique ? A priori, nous pouvons rechercher cette objectivité dans leur définition, dans leur calcul ou dans leur interprétation.

Du point de vue de la définition, graphique et statistique sont tout aussi objectifs l'un que l'autre. Prenons l'exemple d'un simple nuage de points : si je précise correctement le type de graphique, les échelles, les couleurs, les dimensions, la typographie ..., il n'y a aucune ambiguïté possible sur le graphique. Bien sûr la définition précise d'un graphique est toujours complexe mais que dire alors de celle d'une droite calculée par les moindres carrés ordinaires ? De même, il existe de nombreuses façons de représenter la même chose mais repensez un instant seulement aux multiples facettes de la « moyenne » : arithmétique, géométrique, harmonique, tronquée, winsorisée, pondérée, etc.

Croire a priori à l'objectivité du calcul d'une statistique est quelque peu exagéré : c'est ignorer l'importance et les particularités de l'ordinateur, instrument pourtant « objectif » s'il en est. Les logiciels statistiques, même les plus sophistiqués, ne vous mettent pas à l'abri de belles surprises. Ainsi, McCULLOUGH (1998, 1999) montre qu'avec SAS, SPSS ou S plus, vous pouvez dans certains cas calculer des moyennes et variances fausses ! Quant à la résolution de certains problèmes non linéaires, les résultats peuvent être simplement fantaisistes et bien entendu différents d'un logiciel à l'autre.

Il est difficile voire impossible, par nature, de parler d'objectivité dans l'interprétation d'une statistique ou d'un graphique. Pour un graphique, cela va de soi mais c'est tout aussi vrai d'une statistique : repensez à toutes ces soirées électorales où le candidat A explique que ses 32% de voix traduisent une immense victoire et que les 68% de voix de son adversaire B, élu lui, représentent la plus humiliante des défaites !

3. Objectivité : 1) Chez Descartes, qui n'est que conceptuel ; 2) Qui existe hors de l'esprit, comme un objet indépendant de l'esprit (Petit ROBERT, 1993).

Enfin, pour terminer sur ce point, notons qu'il n'y a aucune raison pour qu'un graphique d'analyse, celui que vous utilisez pour étudier des données, soit objectif. L'analyse des données est en effet un acte personnel, une interaction directe entre le statisticien et ses données; dans ces conditions, choisissez les outils qui vous conviennent, même s'ils ne sont pas « convenables » !

2. « BONS » ET « MAUVAIS » GRAPHIQUES ?

Même parmi les statisticiens convaincus de l'intérêt des graphiques, et ils sont nombreux, cette méfiance « culturelle » persiste sous une forme cependant un peu différente. C'est ainsi que de nombreux ouvrages et articles par ailleurs excellents véhiculent cette étrange idée qu'il y aurait de « bons » et de « mauvais » graphiques : « Le vice et la vertu au pays des graphiques » (DROESBECKE, 1994), « How to Display Data Badly » (WAINER, 1984) etc. Pour démontrer leurs propos, ces auteurs donnent des exemples de « mauvais » graphiques, créant ainsi un très joli paradoxe. Car enfin, un « mauvais » graphique ne devient-il pas un « bon » graphique dès lors qu'il s'agit de montrer ce qu'est un « mauvais » graphique ?

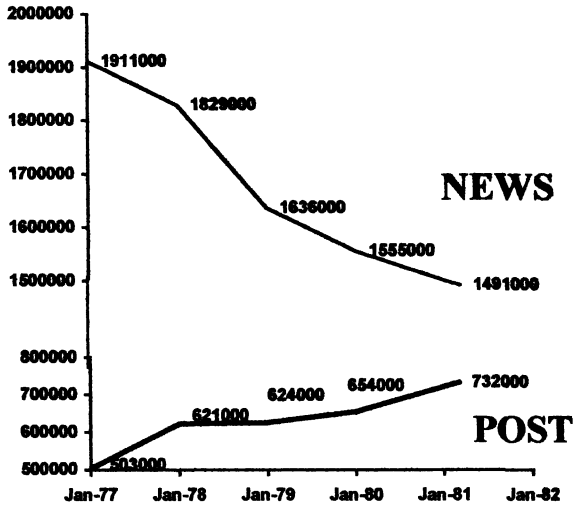
Le graphique 1, repris d'ailleurs dans de nombreux cours de statistique, est ainsi présenté comme un exemple de ce qu'il ne faut pas faire (WAINER, 1984). Ce graphique compare les tirages de deux journaux, le *New York Post* et le *Daily News*. WAINER fait remarquer que, grâce à un changement d'échelle au milieu de l'axe, on a l'impression que le *Post* rattrape le *News* en escamotant le fait que le tirage du *Post* reste moitié de celui du *News*. C'est exact, comme il est aussi exact que le *News* s'effondre (−25 % en 4 ans) et que le *Post* est en pleine expansion (+50 % en 4 ans). De plus, le *News* pourrait utiliser ce même graphique pour stimuler ses employés en leur montrant le danger de la situation actuelle et le *Post* pour féliciter et encourager ses propres troupes !

Un graphique n'est, a priori, ni bon ni mauvais : c'est le statisticien ou l'utilisateur qui en fait une bonne ou une mauvaise utilisation et cette réflexion de bon sens doit rester présente à l'esprit. C'est d'ailleurs la même chose en statistique. Ainsi, pour estimer la moyenne d'une distribution, la moyenne empirique est selon le cas, et malgré toutes ses excellentes propriétés, le meilleur ou le pire des estimateurs : c'est évidemment le meilleur lorsque la loi sous-jacente est gaussienne, mais c'est l'un des pires lorsque la loi sous-jacente est de Cauchy ou plus généralement à queue épaisse (HOAGLIN, MOSTELLER et TUKEY, 1983).

3. GRAPHIQUES ET STATISTIQUES : DE GRANDES SIMILITUDES

Une statistique et un graphique ont en effet de forts points communs : ils ont le même support, les données, et le même objectif, les résumer. Mieux encore,

comme nous allons le voir sur quelques exemples, les statistiques souffrent parfois des mêmes travers que les graphiques et, inversement, on reproche souvent aux graphiques des choses jugées « optimales » dans le cadre de la statistique mathématique.



GRAPHIQUE 1. — Un exemple de « mauvais » graphique (repris de WAINER, 1984).

Statistiques et graphiques « en trompe l'œil »

Qu'un graphique puisse nous abuser est évident ; nous avons tous en mémoire ces fameux « trompe l'œil ». Le graphique 2 nous montre deux représentations de la même série de chiffres : un simple changement d'échelle peut nous masquer, ou au contraire nous révéler, la structure sinusoïdale de la série.

Mais les statistiques ont parfois le même défaut. Ainsi, le fameux « effet de structure », qui permet qu'un indice de prix baisse alors que le prix de toutes ses composantes augmente, a de quoi rendre fou ! De même, en 1973, ANSCOMBE présentait un exemple célèbre, lui aussi repris dans de nombreux cours de statistique. Les quatre distributions bivariées (Y_1, X_1), (Y_2, X_1), (Y_3, X_1), (Y_4, X_2) du tableau 1 conduisent exactement à la même droite de régression, au même coefficient de corrélation linéaire, aux mêmes tests de Fisher et de Student. De quoi croire à l'identité des distributions. Et pourtant, un simple graphique montre que c'est loin d'être le cas (graphique 3).

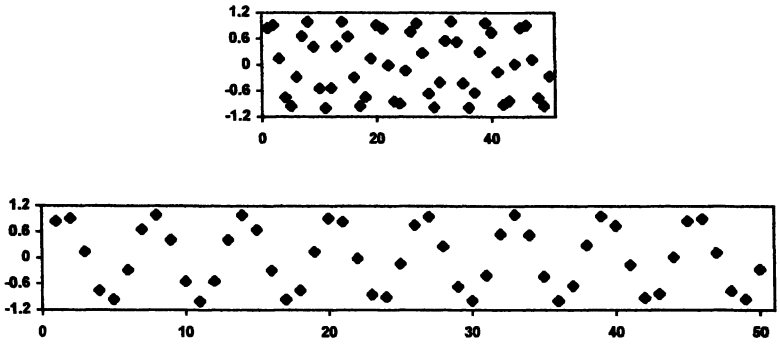
Déformations, zooms, caricatures

Les représentations graphiques utilisent souvent des effets visuels pour mettre l'accent sur une information ou une structure intéressantes, effets qui ont le don d'exaspérer certains statisticiens. Et pourtant, il est facile de trouver

dans la statistique mathématique de telles « manipulations » dont on démontre qu'elles sont alors optimales.

Les « déformations » des observations initiales sont ainsi fréquentes en statistique robuste et non paramétrique où les estimateurs à noyaux attribuent des poids différents aux données (voir par exemple HÄRDLE (1990) pour un exposé complet des méthodes non paramétriques de régression).

De même, en analyse de séries temporelles, on n'hésite pas, pour rendre stationnaire la série, à la différencier (pour stabiliser la moyenne) ou à la transformer en prenant son logarithme (pour stabiliser la variance).



GRAPHIQUE 2. — Deux représentations de la même série de chiffres.

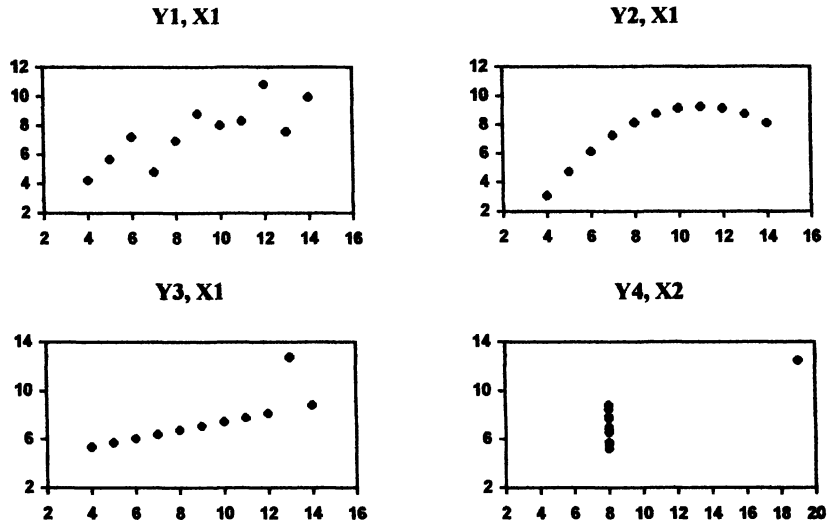
TABLEAU 1. — Les données d'ANSCOMBE légèrement modifiées [4] et statistiques associées pour la régression.

X_1	Y_1	Y_2	Y_3	X_2	Y_4	**Résultats**	
4	4.260	3.100	5.386	19	12.500	N	11
5	5.683	4.734	5.731	8	6.890	$\overline{x}$	9
6	7.240	6.140	6.077	8	5.245	σ_x	3.3166248
7	4.818	7.260	6.423	8	7.910	$\overline{y}$	7.5
8	6.950	8.134	6.769	8	5.760	σ_y	2.0309322
9	8.801	8.770	7.114	8	8.842	$Cov(xy)$	0.0989806
10	8.038	9.131	7.460	8	6.580	b	3
11	8.330	9.254	7.806	8	8.463	Student (b)	2.6688025
12	10.840	9.137	8.151	8	5.560	a	0.5
13	7.580	8.740	12.740	8	7.710	Student (a)	4.2431255
14	9.960	8.100	8.843	8	7.040	R^2	0.6667175

4. Les données présentées ici ont été légèrement modifiées par l'auteur, par rapport à celles d'ANSCOMBE, pour qu'elles fournissent exactement les mêmes résultats, quel que soit le nombre de décimales requis.

Certaines observations sont parfois simplement « oubliées » pour obtenir de meilleurs estimateurs : c'est le cas des moyennes tronquées et en particulier de la médiane qui est un excellent estimateur du centre d'une distribution symétrique à queue épaisse pour de petits échantillons (HOAGLIN, MOSTELLER et TUKEY, 1983).

On se concentre aussi parfois sur des aspects très particuliers des données comme, par exemple, lorsqu'on utilise pour comparer une distribution observée à une distribution théorique, le test de Kolmogorov-Smirnov basé sur le maximum des différences absolues des fonctions de répartition empiriques.



GRAPHIQUE 3. — Nuages de points correspondant aux données d'ANSCOMBE modifiées.

4. EN GUISE DE CONCLUSION

Il y a donc beaucoup plus de similitudes qu'on ne le croit généralement entre les approches graphique et numérique qui dérivent selon moi de la même logique d'analyse des données. Certains graphiques sont d'ailleurs de simples présentations astucieuses des données brutes (listes codées comme dans le graphique 3 de Jean-Paul VALOIS, branchage de TUKEY [5], etc.) ou de statistiques particulières (boîtes à pattes). Simplement, il se trouve que nous sommes actuellement beaucoup plus à l'aise avec les nombres. Si nous sommes capables de définir des statistiques totales, exhaustives, complètes, de calculer la précision d'estimateurs et de vérifier qu'ils sont sans

5. Stem and Leaf display.

biais, convergents, efficaces, il est encore aujourd'hui difficile de définir et plus encore de mesurer les qualités et l'impact d'un graphique. Des travaux importants, souvent méconnus des statisticiens, ont déjà été accomplis ; parmi eux, citons les contributions fondamentales de BERTIN à la science graphique et les développements de la statistique graphique faits sous la direction de TUKEY autour du logiciel S. Il faut continuer à encourager toutes les études qui sont faites dans ce sens et ne pas refuser a priori d'utiliser des méthodes, à l'évidence excellentes, sous le prétexte qu'elles ne sont pas (encore) parfaitement compréhensibles.

L'informatique permet aujourd'hui l'accès à d'énormes quantités d'informations de toute sorte. Curieusement, la statistique classique inférentielle est un peu démunie devant la taille des fichiers et la nature nouvelle des données et le statisticien retourne aux vieilles méthodes d'observation pour y voir plus clair. Des méthodes se développent, des logiciels apparaissent et des mots nouveaux surgissent : *Data Mining, Data Warehouse*, etc. Dans toute cette évolution, la statistique n'a malheureusement pas la part qui devrait lui revenir et les statisticiens interviennent peu dans les choix logiciels stratégiques. L'analyse exploratoire des données (EDA) constitue à cet égard une exception remarquable. Branche de la statistique qui utilise en parallèle l'approche graphique et la statistique robuste et non paramétrique (TUKEY, 1977 ; DESTANDEAU, LADIRAY, LEGUEN, 1999), elle se développe très rapidement, tant sur le plan théorique que pratique et donne naissance à de nouveaux logiciels de plus en plus populaires.

BIBLIOGRAPHIE

ANSCOMBE F. J. (1973), *Graphs in statistical analysis.* The American Statistician, vol 27, 17-21.

DESTANDEAU S., LADIRAY D., LEGUEN M. (1999), Analyse Exploratoire de Données, *Courrier des Statistiques* n° 90, 3-43, INSEE, Paris.

DROESBECKE J.J. (1994), Le vice et la vertu au pays des graphiques, *Cahiers du CERO*, n° 36, 77-104.

HÄRDLE W. (1990), *Applied Nonparametric Regression*, Cambridge University Press.

HOAGLIN D.C., MOSTELLER F., TUKEY J.W. (1983), *Understanding Robust and Exploratory Data Analysis*, John Wiley, New York.

Le Petit ROBERT (1993), Dictionnaire de la langue française.

MCCULLOUGH B. D. (1998), Assessing the Reliability of Statistical Software : Part I, *The American Statistician*, vol. 52, n° 4, pp 358-366

MCCULLOUGH B. D. (1999), Assessing the Reliability of Statistical Software : Part II, *The American Statistician*, vol. 53, n° 2, pp 149-159

TUKEY J. W. (1977), Exploratory Data Analysis, Addison Wesley.

VANPOUCKE J. (1991), Grafiquer les données : trafiquer pour un meilleur confort graphique, Cours de l'école d'été EDA, *Association MIRAGE*, http ://www.unige.ch/ses/sococ/mirage/.

WAINER H. (1984), How to Display Data Badly, *The American Statistician*, vol 32, n° 2, 137-147.