La guerre des étoiles
Distinguer le signal et le bruit

Arthur Charpentier

Professeur d'actuariat à l'Université du Québec, Montréal

La grande difficulté dans la modélisation et la construction de modèles prédictifs est de réussir à distinguer « le signal et le bruit » – pour reprendre le titre du classique de Nate Silver [2012]. La réponse statistique est la notion de significativité et la recherche des « étoiles » dans les sorties de régression. Avec l'explosion du nombre de données, il est devenu crucial de faire cette distinction, de savoir quelles sont les interactions qui sont significatives.

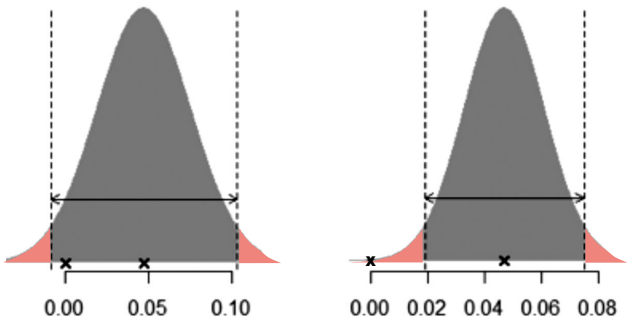
Approches historiques de la notion de « significativité »

Le débat sur la significativité est ancien, même si sa formulation s'est faite historiquement dans des termes assez vagues. Par exemple, dès 1710, le médecin et mathématicien John Arbuthnot s'était interrogé sur le ratio du nombre de naissances de garçons et de filles, se demandant si la différence était « statistiquement significative ». Pour être plus précis, en utilisant des statistiques sur près de quatre-vingt-dix ans, il avait noté [Arbuthnot, 1710] : "*There seems no more probable Cause to be assigned in Physics for this Equality of the Births, than that in our 'first Parents Seed there were at first formed an equal Number of both Sexes.*" John Arbuthnot pose le premier la question en termes probabilistes. Un siècle plus tard, Pierre Simon de Laplace a présenté ce que l'on peut interpréter comme un « test de significativité » (avec notre terminologie actuelle). Il avait en effet noté, en prenant des mesures sur des baromètres, que les observations à 9 heures du matin et à 4 heures de l'après-midi étaient différentes. Significativement différentes. Et, là encore, il avait posé la question en termes probabilistes, en se demandant s'il est « extrêmement probable » qu'il y ait une différence entre les deux mesures. Il avait alors introduit, le premier, un test de comparaison entre des valeurs moyennes : ayant noté que la différence excédait plusieurs écarts types – ce qu'il jugeait alors significativement important –, il en conclut que les séries sont significativement différentes. En 1885, Francis Edgeworth avait repris ces idées, en comparant la taille des criminels à celle des gens ordinaires. Mais il faudra attendre les travaux de William Gosset (plus connu sous le nom Student), de Karl Pearson et de Ronald Fisher surtout pour avoir une définition plus rigoureuse de la significativité.

La notion de significativité est cruciale dans la construction de modèles prédictifs. En assurance automobile, l'âge du conducteur est une variable « significative » quand il s'agit d'expliquer la fréquence de sinistres. Formellement, cela signifie que l'âge et la fréquence de sinistres sont corrélés, et que cette corrélation, notée *R*, est « significativement non nulle ». En 1922, Ronald Fisher, en proposant la construction d'un intervalle de confiance, propose ainsi un test de significativité.

Comme il l'écrit [Fisher, 1925], "*from these values, we obtain the difference 0,0471 ± [0,0142] which might well be regarded as significant*". Dans la figure 1, la courbe de droite correspond au cas où la valeur est significativement non nulle (car l'intervalle de confiance à 95 % ne contient pas 0), alors que la courbe de gauche correspond au cas où la valeur est non significative (car l'intervalle de confiance à 95 % contient 0). Ronald Fisher passera plusieurs années à formaliser et expliquer cette idée de significativité statistique.

Figure 1 - Distribution théorique de l'estimateur de la corrélation, *R*, et intervalle de confiance



Source : auteur.

Tests, prise de décision et erreurs

Une des contributions majeures de la statistique des années 1920 a été de formaliser la prise de décision. Le tableau 1 montre ainsi le mécanisme binaire de la prise de décision et les deux types d'erreur : rejeter à tort une hypothèse ou accepter à tort une hypothèse.

Tableau 1 - Schéma de la prise de décision

	On accepte H_0	On rejette H_0
H_0 est vraie	Bonne décision	*Erreur (première espèce)*
H_0 est fausse	*Erreur (seconde espèce)*	Bonne décision

Source : auteur.

Déclarer qu'une variable, ou une différence, est significative peut conduire à deux types d'erreur : la déclarer comme non significative, alors qu'elle l'était (prendre un signal pour du bruit), et la déclarer comme significative, alors qu'elle ne l'était pas (prendre du bruit pour un signal). Quand on construit un test, on va essayer de contrôler la probabilité de commettre de telles erreurs.

Mais de faibles taux d'erreur ne veulent pas forcément dire qu'un test est efficace, et le principal danger des tests médicaux est lié à une mauvaise interprétation de ces taux d'erreur. Supposons qu'un test détecte 1 personne sur 1 000 dans une population. Supposons aussi que le test soit relativement « efficace » au sens où 90 % des cas sont des cas positifs correspondent effectivement à des personnes malades, et où le test est négatif dans 99 % des cas quand on n'est pas touché par la maladie.

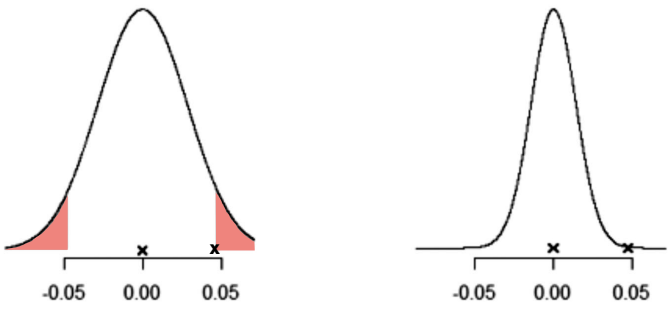
Ces chiffres sont bien au-dessus de la plupart des principaux tests couramment utilisés. Si le test est effectué sur 10 000 personnes, il sera positif sur 10 personnes (en moyenne), parmi lesquelles 9 sont effectivement malades, mais 1 est saine. À côté, le test sera négatif pour 9 990 personnes. Et, parmi ces 9 990 personnes négatives, 9 890 sont effectivement saines, et une centaine sont pourtant malades. Donc, au final, sur les 109 personnes malades, 9 ont été détectées, mais 100 sont passées inaperçues, soit un peu plus de 90 % des malades ! Ce test supposé efficace ne l'est peut-être pas tant que ça.

Le concept de *p*-value et le mythe des 5 %

Le concept de *p*-value est lié justement à l'erreur de rejeter à tort une hypothèse. Si Ronald Fisher en parle abondamment, il faut toutefois attribuer la paternité du concept à son collègue Karl Pearson. Ce dernier, pour définir la notion de « significativité », utilise la formulation suivante : "P = *0,1227 or the odds are now only 8 to 1 against a system of deviations as improbable as or more improbable than this one.*"

Si l'on reprend l'exemple de la corrélation, Pearson nous dit qu'il y a 1 chance sur 8 d'obtenir une valeur aussi improbable. Comme l'illustre la figure 3, la *p*-value est la probabilité d'avoir une statistique aussi grande ou aussi petite que celle obtenue sur l'échantillon si effectivement la corrélation était nulle. Ou encore $p = P[|R| > r|H_0]$.

Figure 3 - Distribution théorique de l'estimateur de la corrélation sous l'hypothèse où cette dernière serait nulle et probabilité que $|R|$ dépasse la valeur empirique observée



Source : auteur.

Sur la courbe de gauche, la *p*-value correspondant à l'aire rouge est de l'ordre de 10 %, alors que sur la courbe de droite l'aire est de 0,1 %. En fait, si la *p*-value est inférieure à 5 %, alors 0 n'est pas dans l'intervalle de confiance à 95 % de l'estimateur de corrélation. Donner la *p*-value est alors suffisant pour juger de la significativité d'une statistique.

Ronald Fisher [1925], dans le chapitre 4 de son livre, pose les bases de la pratique (toujours en vigueur aujourd'hui) des tests statistiques : "*If* p *is between 10% and 90% there is certainly no reason to suspect the hypothesis tested. If it is below 2% it is strongly indicated that the hypothesis fails to account for the whole of facts.*"

Autrement dit, avec une *p*-value supérieure à 10 %, on peut accepter notre hypothèse (souvent, l'hypothèse que l'on cherche à tester est que la variable n'est pas significative), et, avec une *p*-value inférieure à 2 %, on va la rejeter (la variable sera alors significative si l'on fait un test de nullité d'une corrélation). Mais entre les deux ?

Ronald Fisher clôt le débat sans vraiment s'en rendre compte en affirmant que "*we whall not often be astray if we draw a conventional line at 5%*". Le choix – aujourd'hui dogmatique – d'un seuil à 5 % repose sur l'idée de rejeter à tort une hypothèse avec 1 chance sur 20, ou bien il correspond au fait de s'éloigner de deux écarts types de la moyenne d'une loi normale centrée-réduite (ce qui arrive avec 1 chance sur 22).

Cette règle des 5 % – 1 chance sur 20 – est encore en vigueur aujourd'hui. C'est elle que l'on utilise dans tous les modèles économétriques en regardant les « étoiles » associées à chacune des variables explicatives (3 étoiles si la *p*-value est inférieure à 0,1 %, 1 étoile si elle est inférieure à 5 %, et rien au-delà de 10 %). Si la *p*-value est une mesure continue de la distance entre l'hypothèse que l'on cherche à tester et les données, ces étoiles ont instauré des seuils, malheureusement supposés infranchissables. Avec cette méthode, comme le notait Gelman [2015], "*it seems impressive to see multiple independent findings that are statistically significant but with enough effort it is possible to find statistical significance anywhere*".

Pour illustrer la difficulté de la prise de décision, Power [2014] prend un exemple tiré de la survie à des séances de chimiothérapie, pour guérir un cancer (voir tableau 2 page ---).

Tableau 2 - Schéma de la prise de décision

	Survie	Décès	
Research Hospital	48 (86 %)	8 (14 %)	*56 (100 %)*
Non Research Hospital	54 (81 %)	13 (19 %)	67 (100 %)
	102 (83 %)	21 (17 %)	123 (100 %)

Source : Williams *et al.* (1987).

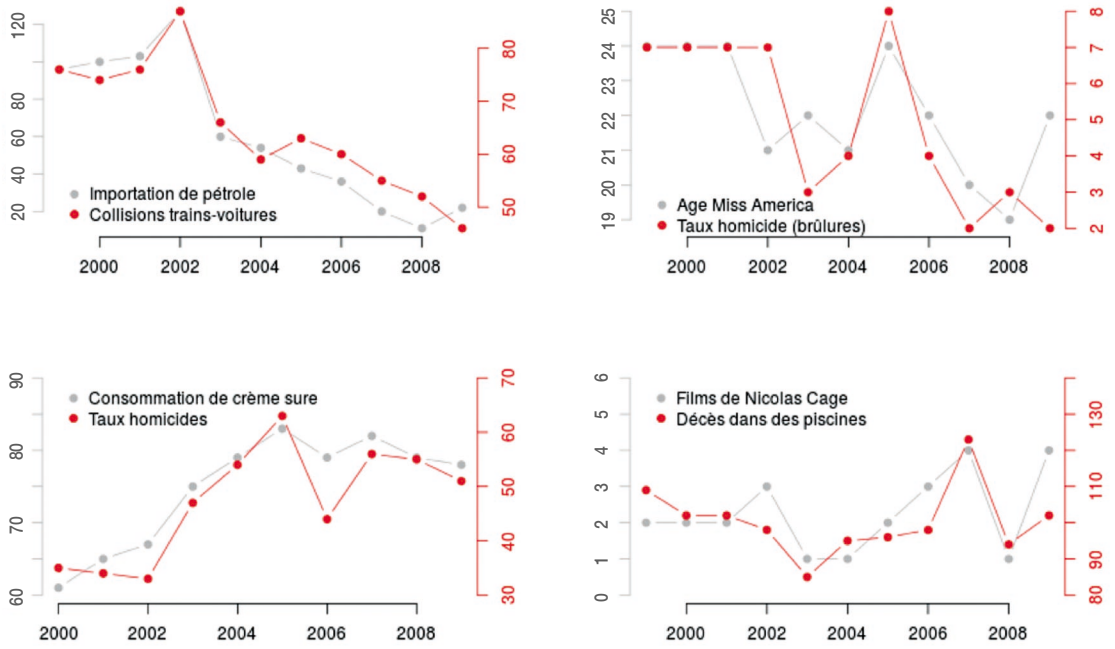
Dans un "research hospital" le taux de survie est de 86 %, alors qu'il est de 81 % dans un "non-research hospital". Un test du chi-deux, mesurant l'indépendance entre le choix de l'hôpital et la survie, donne une statistique de test de 0,5635, correspondant à une *p*-value de 45,29 %. Autrement dit, on devrait être indifférent à l'endroit où on va se faire traiter, puisque l'hôpital est non significativement corrélé à la survie. Mais lorsque la vie est en jeu, est-on prêt à dire qu'une différence de 5 points (86 % vs. 81 %) n'est pas "statistiquement significative" ? Même avec une *p*-value proche de 50 % !

Le problème des tests multiples

Si cette méthode pour juger de la significativité d'une variable dans un modèle prédictif a eu beaucoup de succès, il convient d'insister sur ses limites quand on se retrouve face à des données massives. Si l'on cherche à expliquer une variable (comme la sinistralité) par cent variables, possiblement indépendantes de notre variable d'intérêt, l'analyse précédente devrait nous pousser à retenir cinq variables comme significativement corrélées avec la sinistralité (voir figure 4).

Le problème des tests multiples est d'autant plus important en imagerie médicale. Dans la figure 5 (voir p. ---), 100 échantillons gaussiens sont simulés, de moyenne nulle, et on teste la nullité des moyennes (seules les *p*-values sont présentées). Imaginons des tests sur des pixels d'une image d'IRM, par exemple. Sur une image obtenue sur un individu sain, si des tests sont effectués pixel par pixel, 10 % des points

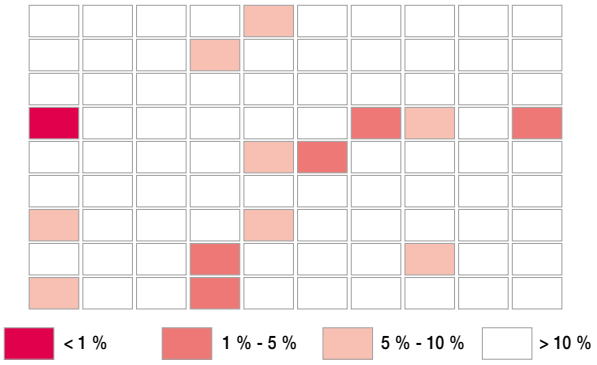
Figure 4 - Exemples de séries significativement (très) corrélées



Source : auteur, données Vigen [2015]. *http://tylervigen.com/spurious-correlations*

présenteront une *p*-value inférieure à 10 %, 5 % une *p*-value inférieure à 5 %, etc.

Figure 5 - 90 tests de nullité de moyenne sur 90 échantillons simulés (de moyenne nulle)



Source : auteur.

En l'occurrence, la vision bayésienne de la prise de décision, décrite dans Greenland et Poole [2013], peut s'avérer plus juste, en plus d'offrir une interprétation plus claire.

Les alternatives pour juger de la significativité

Les *p*-values sont un outil important, central, dans la construction de modèles prédictifs. Et la pratique des tests est souvent un exercice intéressant : si la *p*-value est petite, on se sent conforté dans l'idée que la variable est significative. Mais, si elle est grande (disons, si elle excède 10 %), on ose parfois se dire que l'on n'a pas eu de chance… Dans certains cas, on peut même s'offrir la chance de faire le test sur un autre échantillon, et – avec un peu de chance – la *p*-value sera plus faible.

Comme le note Briggs [2013], les *p*-values "*encourage magical thinking [...] they focus attention on the unobservable*". Même avec un regard critique, l'utilisation des *p*-values est dangereuse. Et les utiliser de manière automatique, dans un algorithme d'apprentissage, l'est encore plus. Sans bon sens, on verra des variables extrinsèques exotiques utilisées dans les modèles prédictifs, pour reprendre la terminologie de Cass et Shell [1983]. À quand un assureur qui utiliserait la pointure de chaussures, les résultats au brevet des collèges ou la couleur de la boîte à lettres dans son tarif d'assurance auto ?

Bibliographie

ARBUTHNOT J., "An Argument for Divine Providence, Taken from the Constant Regularity Observed in the Births of Both Sexes", *Philosophical Transactions of the Royal Society of London*, vol. 27, 1710.

BRIGGS W., "Everything wrong with *p*-values under one roof", wmbriggs.com, 6 octobre 2013. *http://wmbriggs.com/post/9338/*

CASS D. ; SHELL K., "Do Sunspots Matter?", *Journal of Political Economy*, vol. 91, n° 21, 1983, pp. 193-228.

FISHER R., *On the mathematical foundations of theoretical statistics.*, Philos. Trans. Roy. Soc. London Ser. A, 1922.

FISHER R., *Statistical Methods for Research Workers*, Oliver and Boyd, 1925.

GELMAN A., "*P*-values and statistical practice", andrewgelman.com, 4 septembre 2015. *http://andrewgelman.com/2015/09/04/p-values-and-statistical-practice-2/*

GREENLAND S. ; Poole C., "Living with *P*-Values: Resurrecting a Bayesian Perspective on Frequentist Statistics", *Epidemiology*, vol. 24, 2013, pp. 62-68.

PEARSON, K., "On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it Can be Reasonably Supposed to have Arisen from Random Sampling", *Philosophical Magazine*, vol. 50, n° 302, 1900, pp. 157-175.

POWER P.R., (2014). *Acts of God and Man Ruminations on Risk and Insurance*, Columbia University Press.

SILVER N., *The Signal and the Noise: Why So Many Predictions Fail – But Some Don't*, Penguin Press, 2012.

VIGEN T., *Spurious Correlations*, Hachette Books, 2015.

WILLIAMS S.D. *et al.*, Treatment of Disseminated Germ-Cell Tumors with Cisplatin, Bleomycin, and either Vinblastine or Etoposide. *New England Journal of Medecine*, vol. 316, 1987, pp. 1435-1440.