COURS

B-A-BA de statistique : moyenne, incertitude et test d'hypothèse

G. Paturel, Observatoire de Lyon

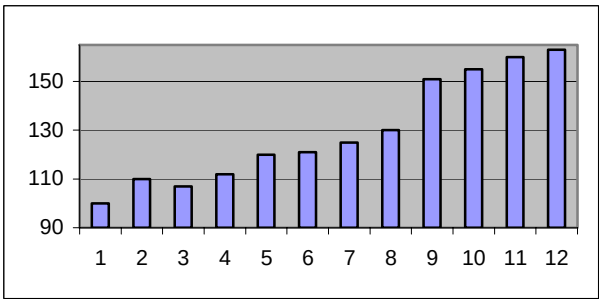
Résumé : *Les statistiques sont très souvent utilisées en science. Nous présentons d'abord le calcul de moyenne, d'écart type et d'erreur moyenne. Puis nous expliquons le test de significativité (dit test de Student), qui doit nous renseigner sur le degré de confiance que l'on doit accorder à tel ou tel résultat.*

Introduction

Vous avez, sans doute, déjà entendu quelqu'un vous affirmer qu'il ne faut pas croire aux statistiques, car "*on peut leur faire dire ce que l'on veut !*". Nous allons nous appliquer à montrer que cette croyance n'est vraie que si les statistiques sont mal employées ou pire, employées par des gens malhonnêtes.

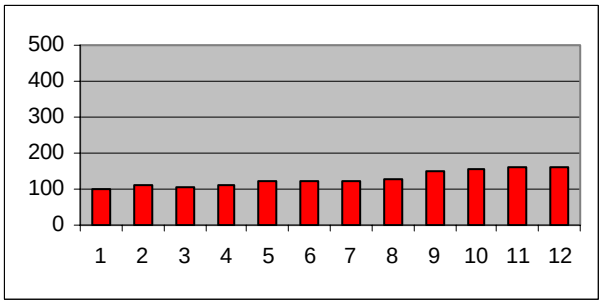
Nous allons justement commencer par quelques affirmations qui cachent un piège.

Ainsi, après un sondage une personne déclare : "90% des personnes qui ont provoqué un accident de voiture avaient absorbé de l'eau dans la journée de l'accident. L'eau, au volant, serait-elle dangereuse ?". Dans le même style, il y a la corrélation réelle entre le nombre de cigognes et le nombre de bébés, en Alsace. Au vu de cette corrélation, il serait facile de conclure que ce sont les cigognes qui apportent les bébés.



Donnons un exemple réel moins trivial. Un histogramme publié par une banque vous incite à investir sur la valeur en bleu qui subit une forte hausse, alors que la valeur en rouge a une tendance à la stagnation. Quelle valeur boursière choisir ?

Vous n'hésitez pas et optez pour la valeur bleue. Erreur ! Les deux diagrammes sont obtenus avec exactement les mêmes données, mais le premier dilate l'échelle verticale (l'origine n'est pas à zéro). Cet exemple est inspiré d'un exemple réel.



Moyenne, écart-type et erreur moyenne

Commençons par la quantité la plus simple : le calcul d'une moyenne. Tout le monde sait faire : on ajoute toutes les valeurs et on divise par le nombre de valeurs.

$$moyenne = \overline{X} = \frac{\sum_{i=1}^{N} X_i}{N} \qquad (1)$$

Les valeurs des X_i sont les résultats des différentes mesures de la grandeur X. Ces mesures doivent être indépendantes les unes des autres. Dans la grande majorité des cas, en physique ou en astronomie, on suppose que les causes qui influent sur la valeur d'une mesure X_i sont très nombreuses. On peut appliquer la statistique de Gauss. La distribution des valeurs de X obéit à la célèbre courbe en cloche de Gauss (voir à la fin de l'article). La position du maximum est la valeur moyenne.

L'*écart type* permet d'apprécier la dispersion des mesures X_i. Si la courbe en cloche est large, l'écart type est grand. Comment calculer sa valeur ? L'expression est simple et très logique :

$$\text{écart-type} = \sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N-1}} \qquad (2)$$

On fait la somme des carré des écarts à la moyenne (on élève au carré pour éviter que les écarts positifs compensent les écarts négatifs) et on divise par le nombre de mesures "moins un". Puis on prend la racine carré du total, pour que l'unité de mesure soit la même que celle des valeurs observées.

Pourquoi "moins un" ? C'est un petit détail, négligeable quand le nombre N est grand. Cette correction prend en compte le fait que le calcul de la moyenne $\overline{X}$ se fait avec les mêmes données que celles utilisées dans le calcul de σ. On dit qu'il y a un degré de liberté en moins. Si on divise par N au lieu de N-1, le résultat, peu différent, s'appelle *l'écart quadratique moyen*.

La valeur de σ mesure, de manière standardisée, la largeur de la courbe de Gauss. Attention, cette valeur n'est pas l'erreur que l'on commet sur l'estimation de $\overline{X}$, ce n'est que la dispersion des mesures individuelles.

On conçoit bien que pour améliorer l'estimation de $\overline{X}$ il faut faire un grand nombre de mesures. Plus il y aura de mesures et plus précise sera l'estimation de la position du maximum de la courbe de Gauss. L'incertitude sur cette position, varie comme l'inverse de la racine carré du nombre de mesures. Nous désignerons cette incertitude par *erreur moyenne*[1]. Son expression est :

$$\text{erreur-moyenne} = \overline{\sigma}_{\cdot} = \frac{\sigma}{\sqrt{N}} = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N(N-1)}}$$

Test d'hypothèse

Un problème se pose souvent. Je fais une mesure d'une grandeur X en effectuant plusieurs mesures individuelles. Par exemple, je mesure le temps de chute d'un corps lâché d'une hauteur donnée (on se distrait comme on peut !). Au bout d'un certain nombre de mesures, je calcule le temps moyen $\overline{T}$ et son erreur moyenne $\overline{\sigma}$. J'utilise ces mesures pour estimer le module de l'accélération de la pesanteur g, et je me demande si la valeur que je trouve, $\overline{g}$, diffère significativement de la vraie valeur g_o=9,81m.s^{-2}.

Il faut effectuer un test d'hypothèse. Il existe plusieurs types de tests. Nous verrons le plus simple et le plus courant, le test de Student.

Calculons la valeur $t = \dfrac{|g_o - \overline{g}|}{\overline{\sigma}}$.

On comprend aisément en voyant la définition que plus t est grand et plus ma mesure diffère de la valeur vraie. De combien ma valeur est-elle autorisée à s'écarter de la valeur g_o pour que je puisse dire qu'elle est *compatible,* avec une probabilité donnée ? Je fais l'hypothèse que ma mesure est *compatible* et si t est inférieur à une certaine limite, $t_{0,01,}$ je conclurai que mes mesures sont compatibles avec la véritable valeur.

La limite théorique $t_{0,01}$ à ne pas dépasser, pour une probabilité d'erreur de 1%, est calculée en fonction du nombre N de mesures. Le tableau ci-dessous donne ces limites :

N	Probabilité1%
10	3,17
20	2,85
30	2,75
∞	2,58

Il est d'usage d'accepter l'hypothèse de compatibilité si $t < t_{0,01}$. La probabilité d'erreur est alors inférieure ou égale à 1%. Au-delà la compatibilité n'est pas prouvée (mais elle est peut-être vraie).

On voit que la limite, pour un échantillon assez large est autour de 2. C'est ce qui justifie la règle empirique des "rejets à 2-sigma". Si dans le calcul d'une moyenne, une valeur s'écarte de la moyenne de plus de 2 fois la valeur de $\overline{\sigma}$, cette valeur est douteuse et doit être rejetée. Nous avions vu un exemple d'un tel rejet à $2\overline{\sigma}$ quand nous avions traité les mesures de la distance Terre Soleil par différents auteurs (CC107, p25)[2]

Nous verrons la prochaine fois les subtilités de la régression linéaire et les pièges qui y sont attachés. Bref, il ne faut pas utiliser un tableur sans savoir ce que l'on calcule.

[1] Il serait plus juste de dire l'erreur sur la moyenne. En anglais cette erreur est appelée "*mean error*", souvent abrégée en *m.e.*.

[2] Dans l'article, j'avais écrit que le rejet était à 2σ en disant que σ était l'écart-quadratique-moyen. Mais j'utilisais bien l'erreur-moyenne et non l'écart-quadratique-moyen. Vous pouvez vous amuser à le vérifier en calculant la moyenne de la dernière colonne du Tableau B, page 26 du CC107.

Astuce de calcul

Quand vous devez calculer une moyenne $\overline{X}$ et son écart type σ, vous pensez qu'il faut calculer d'abord la moyenne puis l'écart-type, puisque la moyenne figure dans l'expression de l'écart-type. Ceci peut s'avérer fastidieux avec une calculette, car il faut saisir deux fois les données individuelles X_i.

Nous allons montrer qu'il n'en est rien. Il suffit de saisir les données une fois en calculant la somme des X_i (que nous désignerons par *SX*) et la somme des X_i^2 (que nous désignerons par *SX2*).

La moyenne se calcule par : $\overline{X} = \frac{SX}{N}$, et l'écart type par : $\sigma = \sqrt{\frac{SX2 - \frac{(SX)^2}{N}}{N-1}}$

La démonstration est facile. D'après la relation (2) :

$$(N-1)\sigma^2 = \sum_{i=1}^{N}(X_i - \overline{X})^2$$

Développons le second membre :

$$\sum_{i=1}^{N}(X_i - \overline{X})^2 = \sum_{i=1}^{N}(X_i^2 + \overline{X}^2 - 2\overline{X}X_i)$$

$\overline{X}$ est une valeur constante qui peut sortir du signe de sommation. On a donc :

$$\sum_{i=1}^{N}(X_i - \overline{X})^2 = \sum_{i=1}^{N}\left(X_i^2\right) + N\overline{X}^2 - 2\overline{X}\sum_{i=1}^{N}X_i$$

Dans la dernière sommation, je peux multiplier par N et diviser par N,ce qui ne change rien, mais fait apparaître $\overline{X}$.

$$\sum_{i=1}^{N}(X_i - \overline{X})^2 = \sum_{i=1}^{N}\left(X_i^2\right) + N\overline{X}^2 - 2\overline{X}N\frac{\sum_{i=1}^{N}X_i}{N}$$

On obtient donc :

$$\sum_{i=1}^{N}(X_i - \overline{X})^2 = \sum_{i=1}^{N}\left(X_i^2\right) + N\overline{X}^2 - 2N\overline{X}^2$$

qui se simplifie donc en :

$$\sum_{i=1}^{N}(X_i - \overline{X})^2 = \sum_{i=1}^{N}\left(X_i^2\right) - N\overline{X}^2$$

et avec nos notations :

$$\sum_{i=1}^{N}(X_i - \overline{X})^2 = SX2 - N\left(\frac{SX}{N}\right)^2$$

donc finalement on a :

$$(N-1)\sigma^2 = \sum_{i=1}^{N}(X_i - \overline{X})^2 = SX2 - \frac{(SX)^2}{N}$$

ce qui conduit bien à l'expression cherchée :

La courbe de Gauss

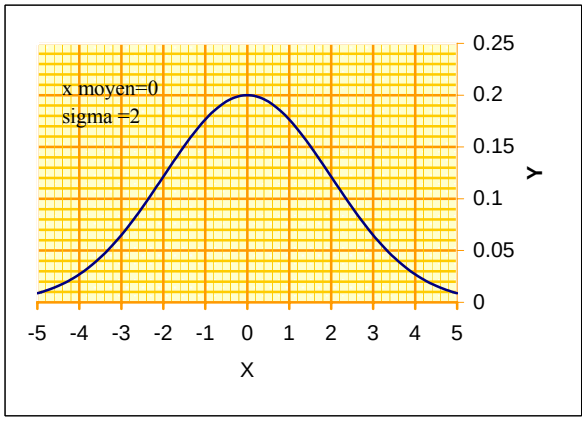
Dans les calculs de cet article nous avons admis que la statistique des mesures suivait la loi de Gauss (ou loi *normale*). Est-ce le bon choix ?

Pour des phénomènes régis par un très grand nombre de variables indépendantes, la statistique de Gauss est bien adaptée. Il existe d'autres lois statistiques, comme la loi de Poisson ou la loi binomiale, mais la loi de Gauss reste la plus appropriée dans les mesures physiques affectées par de multiple petits effets imprévisibles.

L'équation de la loi de Gauss est la suivante :

$$y = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[\frac{-(X-\overline{X})^2}{2\sigma^2}\right]$$

Voici sa forme, par exemple, pour $\overline{X} = 0$ et $\sigma = 2$:



L'aire de la courbe dans un intervalle, de part et d'autre de la moyenne, donne la probabilité d'obtenir une mesure dans cet intervalle. L'aire totale de la courbe vaut 1 (voir ci-dessous).

■

L'intégrale de la courbe de Gauss

Dans l'expression de la courbe de Gauss on voit apparaître $\sqrt{\pi}$. D'où vient-il ?

Calculons $I = \int_{-\infty}^{\infty} e^{-x^2}\,dx$.

Pour cela écrivons I^2 et passons en "polaire".

$I^2 = \int_{-\infty}^{\infty} e^{-y^2}\,dy \int_{-\infty}^{\infty} e^{-x^2}\,dx = \int_0^{2\pi}\int_0^{\infty} e^{-r^2}\,r\,dr\,d\theta$, qui

donne $I^2 = \int_0^{2\pi}\int_0^{\infty} e^{-r^2}\,r\,dr\,d\theta = 2\pi\int_0^{\infty} e^{-u}\frac{du}{2} = \pi$

D'où I=$\sqrt{\pi}$.

C'est ainsi que l'on trouve : $\int_{-\infty}^{\infty} e^{-\frac{(x-\overline{x})^2}{2\sigma^2}} = \sigma\sqrt{2\pi}$.