Statistique descriptive

BCPST 1 — 27 février 2017

La statistique est l'étude de la collecte de données, de leur traitement, de leur interprétation et de leur représentation. C'est à la fois une science (qui s'appuie sur la théorie des Probabilités) et une technique employée dans une large gamme de domaines.

I — Position du problème

I.1 — Où est l'aléatoire ?

L'aléatoire est présent dans toute expérience scientifique. Les deux grandes explications en sont

- L'aléatoire « intrinsèque » lié à la complexité des individus et des phénomènes étudiés et au manque d'information dans le domaine.
- L'aléatoire « expérimental » lié au protocole lui-même, aux moyens limités, etc. Cet aléatoire expérimental peut provenir
  - de la diversité des individus étudiés et de la variété des environnements dans lesquels ces individus ont été placés (manque d'information) ;
  - de l'échantillonnage (moyens limités) ;
  - des erreurs de mesures expérimentales.

La première source d'aléatoire est le manque d'informations, ou ignorance, la seconde est matérielle et peut donc être contrôlée et donnée elle-même lieu à une étude plus poussée. Il n'y a pas de raison de penser qu'une source d'aléatoire peu complètement être supprimée.

En Biologie (ainsi qu'en Médecine), étant donné l'extrême complexité des systèmes étudiés, l'aléatoire est très présent

Exemple On étudie l'influence d'un champignon sur une population d'hêtres dans deux parcelles de forêt. Plusieurs sources d'aléatoires vont être prise en compte.

- Plusieurs individus de la même espèce, de la même origine et du même âge ont des biomasses différentes : aléatoire « intrinsèque » à l'expérience) ;
- les parcelles ne sont pas identiques (déclivité du sol, présence d'autres espèces, etc.) : aléatoire « extrinsèque » ;

– les échantillons de population ne possèdent pas le même nombre d'individus, les calculs vont donc être source d'aléatoire : aléatoire « expérimental », lié au protocole expérimental ;
– la grandeur mesurée ne peut l'être qu'avec une précision finie, difficilement réductible.

Souvent, on peut au mieux faire un graphe, un nuage de points représentant les données mesurées en fonction des paramètres que l'on a fait varier. La question que l'on pose est alors « Comment mettre en évidence la corrélation entre deux grandeurs ? »

Cette analogie aléatoire pousse à interpréter les mesures effectuées comme des réalisations d'une variables aléatoire X « cachée ». Cette variable aléatoire admet une espérance μ et un écart-type σ_x.

I.2 — Vocabulaire

La statistique utilise un vocabulaire spécifique :

– Une étude statistique porte sur des individus formant une population statistique.
– sur les individus de la population on étudie un caractère qui peut être qualitatif (couleur, sexe, positif/négatif, etc.) ou quantitatif (un quantité mesurée : poids, taille, pH, etc.) ;
– on extrait de cette population un échantillon. La manière de choisir cet échantillon donne elle-même lieu à une étude :
  • L'échantillon doit être homogène : les individus de l'échantillon sont similaires relativement à tous les paramètres extérieurs à l'étude ;
  • l'échantillon ne doit pas être biaisé : la méthode de sélection ne favorise pas un des caractères étudiés.
– les résultats de l'étude donne lieu à des séries statistiques $x_1, x_2, \dots, x_n$, puis $y_1, y_2, \dots, y_n$, etc où l'indice représente l'individu et la lettre x, y, etc. représente le caractère.

On dira qu'un échantillon est biaisé, s'il fournit une dispersion de la VA systématiquement différente de ce qu'elle ait dans la population ciblée.

Par exemple les myopes dans un amphi : le premier rang est un échantillon biaisé... Ou encore le piégeage en écologie : pensez à la taille et comportement des individus) D'où l'intérêt de la théorie de l'échantillonnage. Nous partirons sur le principe simplificateur que les individus de l'échantillon ont été choisi parfaitement au hasard.

Exemple Le contrôle des populations de mustangs (chevaux sauvages) a été l'objet de nombreuses controverses aux États-Unis. Des scientifiques américains ont étudié le contrôle des populations de mustang par stérilisation des mâles [1].

1. Eagle, T. C., Asa, C., and Garrott, R. et al. (1993), Efficacy of Dominant Male Sterilization To Reduce Reproduction in Feral Horses, Wildlife Society Bulletin , 21(2), 116-121.

Ils localisaient d'abord plusieurs troupeaux de mustangs qu'ils encerclaient. À l'intérieur de ce troupeau, ils dénombraient tous les adultes de plus de 3 ans, et parmi ceux-ci ils procédaient à la stérilisation des mâles. Puis ces mâles étaient marqués par un collier et le troupeau était relâché.

Entre juin 1986 et juillet 1988, les scientifiques procédaient à une étude des troupeau de mustangs dans les zones étudiées. Les troupeaux étaient reperés par hélicoptère, et on dénombrait le nombre d'adultes, le nombre de poulains et le nombre de mâles castrés (identifiables à leurs colliers) dans le troupeau. Les résultats de cette expérience figure dans le tableau ci-dessous.

Adultes	Mâles stérilisés	Poulains	Traitement	Ratio poulain/adultes
232	0	28	0	0,12
172	0	18	0	0,10
136	0	16	0	0,12
127	0	20	0	0,16
118	0	20	0	0,17
115	0	20	0	0,17
226	0	39	0	0,17
197	0	34	0	0,17
143	0	26	0	0,18
159	0	17	0	0,11
139	0	22	0	0,16
169	0	28	0	0,17
173	0	32	0	0,18
243	0	28	0	0,12
240	0	27	0	0,11
180	0	24	0	0,13
192	0	27	0	0,14
170	0	27	0	0,16
178	0	37	0	0,21
52	9	7	1	0,13
36	5	3	1	0,08
25	5	6	1	0,24
69	13	17	1	0,25
65	10	14	1	0,22
60	12	14	1	0,23
35	7	1	1	0,03
31	3	1	1	0,03
63	9	2	1	0,03
53	8	7	1	0,13
57	9	6	1	0,11
40	7	1	1	0,03

Adultes	Mâles stérilisés	Poulains	Traitement	Ratio poulain/adultes
36	8	3	1	0,08
30	5	1	1	0,03
36	5	1	1	0,03
65	5	2	1	0,03
44	8	2	1	0,05
45	8	2	1	0,04
38	8	5	1	0,13

Dans ce tableau

- Chaque ligne correspond à un individu de la population statistique : un troupeau étudié.
- Chaque colonne correspond à un caractère : nombre d'adultes, de poulains, etc.
- Le nombre de mesure pour chaque caractère est la taille de la série statistique. Ici elle est la même pour chaque caractère : 38.
- Il existe deux types de caractères : quantitatifs (le nombre d'adultes ou de poulains) et qualitatifs (le troupeau a-t-il été traité ou non ?)

Bien souvent le caractère pertinent doit être déduit des autres caractères. Par exemple ici on aimerait avoir une mesure du taux de natalité des différents groupes. Comme cette mesure est impossible, on choisit d'utiliser le ratio nombre de poulain/nombre d'adultes.

L'objet essentiel des statistiques est d'étudier la corrélation entre deux caractères. Ici on veut établir un lien entre le traitement des troupeaux et le taux de natalités.

I.3 — Inférence statistique

Les questions essentielles qui se posent sont :

- En étudiant un caractère sur l'échantillon, que peut-on en déduire concernant la population tout entière ?
- Si on considère un même caractère mesuré sur deux échantillons différents, que peut-on dire de la différence entre ces deux ensembles de mesures ? Est-elle le reflet d'une différence « réelle » entre les deux échantillons ou n'est-elle explicable qu'en faisant appel au caractère aléatoire de la quantité ?
- Si on considère deux ou plusieurs caractères mesurés sur un seul échantillon, peut-on mettre en évidence une « corrélation » entre ces deux caractères ? En quoi cette apparente « corrélation « est-elle fiable ?

Ces questions sont au coeur du problème de l'inférence statistique.

II — Statistique univariée

Dans cette partie, on s'intéresse à un unique caractère mesuré sur une population.

II.1 — Descriptif – Représentation graphique

Le premier paramètre, essentiel, est l'effectif de l'échantillon.

La représentation de la série $(x_i)_{1\leqslant i\leqslant n}$ n'a pas d'intérêt en soi. C'est plutôt la répartition des valeurs des x qui est représentée, à l'aide d'un histogramme.

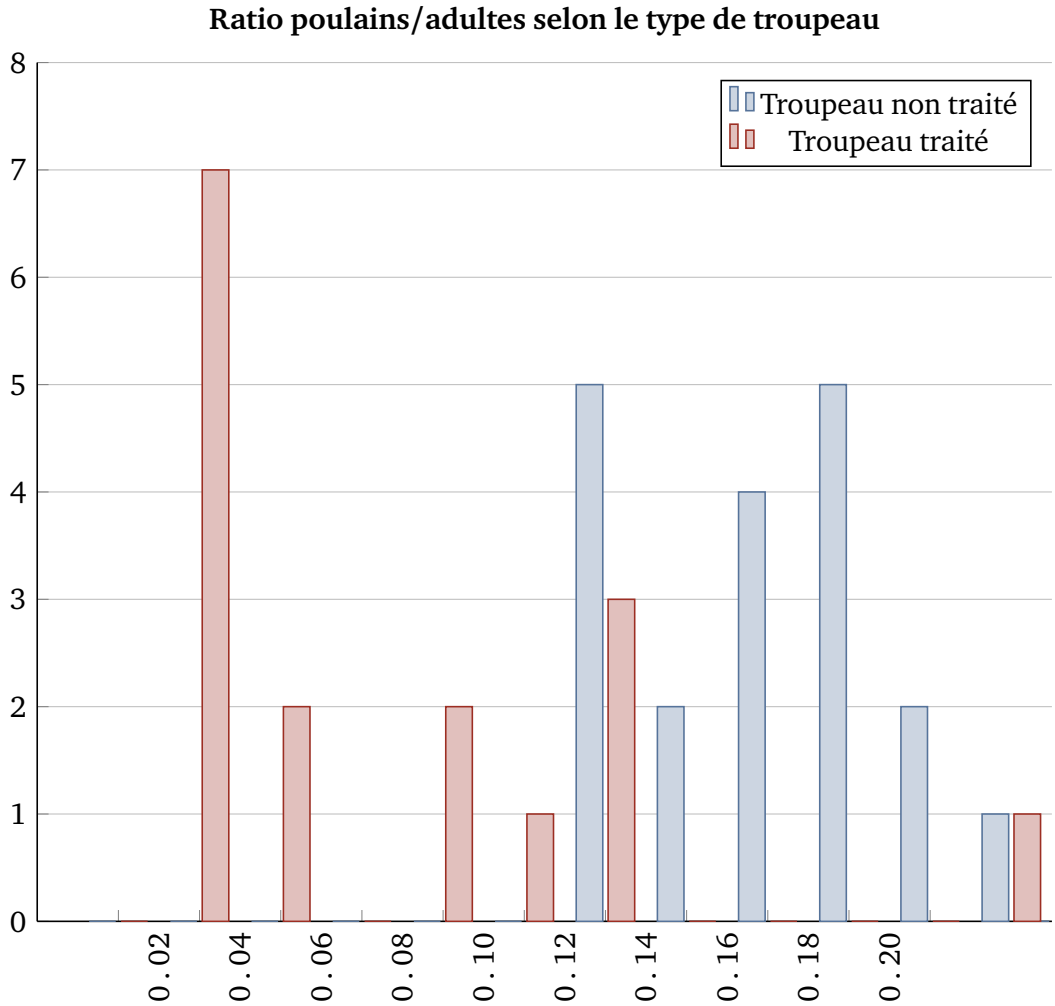
Pour construire un histogramme, on doit d'abord choisir les classes de valeurs : des intervalles pour des caractères quantitatifs ou des ensembles de valeurs pour des caractères qualitatifs. Ensuite pour chaque classe on compte les valeurs de l'intervalle.

Dans le cas d'un caractère quantitatif, un histogramme cumulé peut être construit directement [2] : on en déduit ensuite l'histogramme simple.

Exemple Dans l'étude sur les mustangs, les ratios s'étalent de 0,03 à 0,21. Afin que chaque classe contienne un nombre décents d'individus, on va choisir 10 classes de 0 à 0.25. Pour chaque classe on compte le nombre de troupeaux dans la classe et on représente graphiquement le résultat obtenu.

Trait.	Classe	0,02	0,04	0,06	0,08	0,1	0,12	0,14	0,16	0,18	0,2	0,22
non	Cumulé	0	0	0	0	0	5	7	11	16	18	19
non	Simple	0	0	0	0	0	5	2	4	5	2	1
oui	Cumulé	0	7	9	9	11	12	15	15	15	15	16
oui	Simple	0	7	2	0	2	1	3	0	0	0	1

2. avec la fonction `NB.SI()` d'Excel par exemple



Ce graphique est un élément de discussion important sur la distribution des données. Même si on peut penser que le traitement des troupeau a une influence sur sa fécondité, cette distribution semble fortement aléatoire. Les caractéristiques de positions permettent de quantifier d'éventuelle différence.

II.2 — Caractéristiques de position

Définition 2.1 — *La moyenne d'un caractère quantitatif x est l'équivalent de l'espérance d'une variable aléatoire. Elle est définie par*

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

La médiane d'une série est une valeur qui divise la population en deux ensembles égaux. C'est donc un réel m tel que les deux ensembles

$$\{x_i \ \text{ tel que } \ x_i \leqslant m\} \qquad \text{et} \qquad \{x_i \ \text{ tel que } \ x_i \geqslant m\}$$

contiennent le même nombre de valeurs. Il peut y avoir plusieurs médianes possibles. La médiane est un paramètre de dispersion plus pertinent que la moyenne dans le cas des séries asymétriques (revenus en économie, notes à un DS...)

Les modes d'une série sont les valeurs les plus représentées et les moins représentées. Cette notion s'étend aux caractères qualitatifs.

Le minimum et le maximum des valeurs mesurées peuvent également être intéressants, notamment dans les séries de faibles effectifs.

II.3 — Caractéristique de dispersion

L'écart-type d'un caractère quantitatif x *devrait* être défini par la formule

$$\overline{s_x} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

où μ est l'espérance de X, la variable aléatoire sous-jacente à l'étude. Toutefois on souhaite plutôt définir cet écart-type à partir de la moyenne calculée précédemment. La formule est alors un peu différente [3] :

Définition 2.2 — *L'écart-type d'un caractère quantitatif x est défini par*

$$\overline{s_x} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

La raison du facteur $1/n-1$ au lieu de $1/n$ est que la grandeur $\overline{s_x}^2$, qui est une grandeur aléatoire, admet bien pour espérance $V(X)$, la variance théorique de X.

Définition 2.3 — *Les quartiles et les déciles sont des généralisations des médianes vu précedemment. Les quartiles sont 3 valeurs $x_{1/4}$, $x_{2/4}$ et $x_{3/4}$ qui divisent les valeurs mesurées en 4 ensembles égaux :*

$$\text{card}\left\{x_i \ \text{ tel que } \ x_i \leqslant x_{1/4}\right\} = \text{card}\left\{x_i \ \text{ tel que } \ x_{1/4} \leqslant x_i \leqslant x_{2/4}\right\} =$$

3. la fonction ECARTYPE d'Excel tient compte de ce facteur, alors que la fonction ECARTYPEP est définie avec le facteur $1/n$

$$\operatorname{card}\left\{x_i \ \textit{tel que} \ x_{2/4} \leqslant x_i \leqslant x_{3/4}\right\} \operatorname{card} \operatorname{card}\left\{x_i \ \textit{tel que} \ x_{3/4} \leqslant x_i\right\} \simeq \frac{n}{4}$$

Les déciles se définissent de la même façon.

Exemple Dans l'étude sur les troupeaux de mustang, Excel nous calcule toutes les valeurs demandées. Retenons en particulier :

	Moyenne	Écart-type
Troupeau non traité	0,15	0,031
Troupeau traité	0,1	0,08

Les moyennes des deux intervalles sont différentes. Mais cette différence peut-être dû complètement à l'aléatoire intrinsèque à l'expérience. Comment peut-on affirmer que cet écart n'est pas dû au hasard ?

II.4 — Écart-type et précision expérimentale

La précision expérimentale est la part de hasard dûe à l'imperfection des moyens matériels. On l'estime en étalonnant ses appareils, en réalisant une expérience témoin, en utilisant les relations entre les paramètres, en lisant la notice des appareils de mesures.

L'écart-type mesure la part d'imprécision dû au hasard intrinsèque à l'expérience, la variabilité de la grandeur étudiée.

Il est impératif d'indiquer l'un et l'autre dans votre TIPE ! Aucune grandeur expérimentale ne peut être discuté sans ces deux valeurs !

Si l'imprécision expérimentale est très petite devant l'écart-type, il est raisonnable de l'écarter des discussions.

Dans le cas contraire, aucune discussion statistique sérieuse ne peut être menée : le protocole expérimental est à revoir.

II.5 — L'intervalle de confiance

L'intervalle de confiance permet de répondre grossièrement à la question

Considérant un même caractère mesuré sur deux échantillons différents, la différence de moyennes observée est-elle le reflet d'une différence « réelle » entre les deux échantillons ou n'est-elle explicable que par l'aléatoire ?

Tout d'abord il faut comprendre qu'en Statistique il n'y a pas de réponse binaire oui/non. Tout est aléatoire, donc tout est possible. On remplace la notion de vérité par une notion de seuil de vraisemblance. Il y a trois seuils admis :

– 5 % : un évènement observé dont la probabilité théorique est inférieur à 5 % sera considéré comme statistiquement significatif, puisqu'il est peu probable d'être dû au hasard. C'est le seuil le plus communément employé.
– 1 % : on parle d'évènement hautement significatif ;
– 0,1 % : on parle d'évènement très hautement significatif.

Le principe de l'intervalle de confiance est d'interpréter la moyenne $\overline{x}$ comme une variable aléatoire $\overline{X}$. Son espérance est alors μ et son écart-type est $\frac{1}{\sqrt{n}}\sigma(X)$. Ainsi, lorsque n augmente la variance tend vers 0 et donc la valeur observée de $\overline{X}$ se rapproche de la valeur théorique μ.

Toujours lorsque n tend vers $+\infty$, la loi de la variable aléatoire $\overline{X}$ se rapproche d'une loi de Gauss. On peut dans ce cas affirmer qu'une valeur observée en dehors de l'intervalle

$$\left[\,\mu - 1{,}96\,\sigma\,;\,\mu + 1{,}96\,\sigma\,\right]$$

a moins de 5% de chances d'être observée : elle est statistiquement significative.

En pratique, bien sûr, μ est remplacée par la moyenne $\overline{x}$ et σ est remplacé par l'écart-réduit, qui est l'écart-type de $\overline{X}$

$$\overline{t_x} = \frac{1}{\sqrt{n}} s_x = \frac{1}{\sqrt{n}} \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Notez le facteur $1/\sqrt{n}$ qui est apparu, et qui fait toute la différence ! L'intervalle de confiance à 5% est donc

$$[\,\overline{x} - 1{,}96\;t_x\;;\,\overline{x} + 1{,}96\;t_x\,]$$

Pour répondre grossièrement à la question précédente, on rapporte les deux moyennes mesurées $\overline{x}_1$ et $\overline{x}_2$ aux intervalles de confiance l'une de l'autre.

Le raisonnement statistique est le suivant :

Si l'expérience « extra » a été réalisée entièrement au hasard, la valeur x_{N+1} est alors entièrement dûe au hasard.

Si l'écart entre x_{N+1} et $\overline{x}$ à moins de 5 % de chances de se produire, alors on estimera que cet écart est « significativement » peu probable, c'est-à-dire que l'écart entre x_{N+1} et $\overline{x}$ est significatif.

On peut élever le niveau d'exigence : si on exige que l'écart ait moins de 1% de chance de se produire, alors on parle d'écart « très significatif ». Pour un seuil de 0,1 %, on parle d'écart « hautement significatif ».

Grosso-modo, en comparant les deux intervalles de confiance, on demande à ce que la moyenne d'un échantillon ne puisse pas être interprété comme une valeur « au hasard » de l'autre échantillon.

Remarquez qu'une telle étude permet d'affirmer qu'une différence est significative, c'est-à-dire qu'elle a peu de chance d'être dû au hasard. En aucun cas elle permet d'affirmer que cette différence n'est effectivement pas dûe au hasard !

Limite de la méthode Le principal écueil est qu'en général le caractère étudié n'a aucune raison de suivre une loi de Gauss. On peut montrer que la loi de la moyenne tend, lorsque n tend vers l'infini, vers une loi de Gauss. Quelles valeurs de n sont raisonnables ? Tout dépend de l'étude, mais en général avec une moyenne sur moins de 10/15 valeurs, on est loin d'une loi de Gauss.

Pour cela, il existe d'autres tests plus précis.

Risque des conclusions statistiques : comparaison d'hypothèses

Une hypothèse nulle H_0 étant émise : on l'accepte alors qu'elle est vraie, on la repousse alors qu'elle est fausse.

On la rejette alors qu'elle est vraie (risque de première espèce) on l'accepte alors qu'elle est fausse (seconde espèce).

Exemple Dans l'étude sur les troupeaux de mustang, les populations sont de tailles qui permettent d'utiliser ce test. Complétons donc le tableau précédent.

	Moyenne	Écart-type	Nb	Écart-réduit	Intervalle de confiance	
Troupeau non traité	0,15	0,031	19	0,007	0,136	0,164
Troupeau traité	0,1	0,08	19	0,018	0,064	0,136

Comme les deux intervalles de confiances sont disjoints, on peut en déduire que la différence entre les moyennes est significative : le traitement a bien eu une influence sur les troupeaux de mustang.

III — Statistique bivariée

III.1 — Représentation graphique

Dans cette partie, on s'intéresse à deux caractères quantitatifs mesurés sur une population. On dispose donc d'une série statistique double $(x_1, y_1), \ldots, (x_n, y_n)$.

Ainsi à chaque individu i correspond un point (x_i, y_i)., ce qui définit un nuage de points.

Nom	Relation attendue	Redressement affine	Variables à étudier	
Exponentielle	$Y = C\exp(\alpha X)$	$\ln(Y) = \alpha X + \ln(C)$	$X' = X$	$Y' = \ln(Y)$
Algébrique	$Y = CX^{\alpha}$	$\ln(Y) = \alpha \ln(X) + C$	$X' = \ln(X)$	$Y' = \ln(Y)$
Logistique	$Y = \dfrac{Y_{\max}}{1 + a\mathrm{e}^{-\alpha X}}$	$\ln\left(\dfrac{Y_{\max}}{Y} - 1\right) = -\alpha X + \ln(a)$	$X' = X$	$Y' = \ln\left(\dfrac{Y_{\max}}{Y} - 1$

FIGURE I.1 — ***Redressement affine :*** *Il arrive fréquemment que l'on ne s'attende pas à une dépendance linéaire entre les deux variables. On peut alors procéder à un redressement affine. La fonction logistique est employée dans tous les phénomènes de croissance. Le paramètre $Y_{\max}$ est introduit à la main.*

III.2 — Caractéristique de position

Dans ce nuage de point, le point $(\overline{x}, \overline{y})$ et le point moyen du nuage.

III.3 — Caractéristique de dispersion

Par analogie avec le cours de Probabilités on définit la covariance et le coefficient de corrélation linéaire de la série par

$$\overline{\mathrm{cov}_{x,y}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})\,(y_i - \overline{y}) \qquad \text{et} \qquad \overline{\rho_{x,y}} = \frac{\overline{\mathrm{cov}_{x,y}}}{\overline{s_x}\,\overline{s_y}}$$

III.4 — Ajustement affine

L'idée de l'ajustement affine est de trouver une droite qui passe le plus près possible des points du nuage. On cherche donc deux réels a et b tels que la somme des distances des points à la droite soit minimale… enfin, non, pas vraiment !

En pratique, un problème légèrement différent admet une solution facile à calculer. Très souvent, X n'est pas vraiment une variable aléatoire : sa variabilité est très faible et facile à contrôler. On étudie donc uniquement Y, dont on étudie la dépendance en X. La grandeur

$$f(a,b) = \sum_{i=1}^{n}(ax_i + b - y_i)^2$$

mesure cette dépendance, et on cherche a et b de façon à ce qu'elle soit minimale.

Cette fonction de deux variables a et b s'écrit aussi

$$f(a,b) = n\overline{x^2}a^2 + nb^2 + n\overline{y} + 2n\overline{x}ab - 2n\overline{y}b - 2an\overline{xy}$$

On montre qu'elle est minimale lorsque les dérivées partielles par rapport à a et b sont nulles, c'est-à-dire pour a et b tels que

$$\begin{cases} \dfrac{\partial f}{\partial a} = 2n\left(\overline{x^2}a + \overline{x}b - \overline{xy}\right) = 0 \\ \dfrac{\partial f}{\partial b} = 2n\left(b + \overline{x}a - \overline{y}\right) = 0 \end{cases} \iff \begin{cases} \overline{x^2}a + \overline{x}b = \overline{xy} \\ \overline{x}a + b = \overline{y} \end{cases}$$

$$\iff \quad a = \frac{\overline{\mathrm{cov}_{x,y}}}{\overline{s_x}} \qquad \text{et} \qquad b = \overline{y} - a\overline{x}$$

Le rapport entre la variance de y et la variance de $ax + b$ est le coefficient de corrélation linéaire r^2 définie par

$$\overline{r}^2 = \frac{a^2\overline{s_x}^2}{\overline{s_y}^2} = \frac{\overline{\mathrm{cov}_{x,y}}^2}{\overline{s_x}^2\,\overline{s_y}^2}$$

C'est la partie de la variabilité de y qui est expliquée par la régression linéaire. On peut voir les choses ainsi : la variable aléatoire X est supposée avoir une faible variance, alors que Y est apparement « plus incertaine ». Après application de la régression linéaire, il s'avère qu'une partie de la variabilité de Y s'explique par sa dépendance par rapport à X.

Il demeure une partie de la variabilité de Y qui et dûe, en l'absence d'autres interprétation, « au hasard ». Cette partie est de l'ordre de $1 - r^2$.

IV — Moyens de modélisation

IV.1 — Principe général

On se place maintenant dans la situation suivante : dans l'expérience, deux paramètres varient, dont l'un est « parfaitement » connu (du point de vue statistique) noté X et l'autre subit une variation aléatoire, on le note Y.

Typiquement X est le temps, la concentration, le pH, etc.

La question est d'évaluer l'hypothèse statistique « Y dépend linéairement de X ».

On va exposer une méthode très simple, quoique imparfaite.

IV.2 — La droite des moindres carrés

Calcul de a et de b. On trouve :

$$b = \frac{\sum_{k=1}^n x_k^2 \sum_{k=1}^n y_k - \sum_{k=1}^n x_k \sum_{k=1}^n x_k y_k}{n\sum_{k=1}^n x_k^2 - \left(\sum_{k=1}^n x_k^2\right)^2}$$

$$a = \frac{n\sum_{k=1}^n x_k y_k - \sum_{k=1}^n x_k \sum_{k=1}^n y_k}{n\sum_{k=1}^n x_k^2 - \left(\sum_{k=1}^n x_k^2\right)^2}$$

$$r^2 = \frac{\sum_{k=1}^{n}(ax_k + b - \overline{y})}{\sum_{k=1}^{n}(y_k - \overline{y})}$$

r^2 est le rapport de : l'écart entre la droite prédite et la moyenne théorique et l'écart entre les valeurs observées et les valeurs prédites.

C'est la fraction des variations de y qui peuvent être expliqué par le modèle linéaire. Par exemple, $r^2 = 0,89$ signifie que la droite des moindres carrées n'expliquent pas 11% de l'écart.

Tout comme précédemment, il y a plusieurs seuils de signification.

Le seuil de tolérance de l'hypothèse H_0 statistique : « la variable Y dépend linéairement de X » est donc $1 - r$. Si $1 - r > 5\%$, alors l'hypothèse est rejetée.

IV.3 — Modèle linéarisable

On peut parfois avoir à faire à des modèles linéarisables. Il s'agit des cas où la littérature nous fait penser que Y à une certaine dépendance par rapport à X, d'une forme connue. Dans ce cas, on remplace Y par une nouvelle fonction Y' et X par une variable X'. On s'arrange pour que Y' dépendant linéaire de X'.

Cette méthode a un avantage : elle permet de valider la dépendance entre Y' et X'. Mais il y a des difficultés pour passer au lien entre les X et les Y.

- Dépendance exponentielle : $Y = a\exp(\alpha X)$. Dans ce cas, $\ln Y = \alpha X + \ln a$. On prend $Y' = \ln Y$ et $X' = X$.
- Dépendance algébrique : $Y = aX^{\alpha}$. Dans ce cas, $\ln Y = \alpha \ln X + \ln a$. On prend $Y' = \ln Y$ et $X' = \ln X$.
- Dépendance logarithmique : $Y = aln(X/X_0)$. Dans ce cas, $\exp(Y/a) = X/X_0$.
- Modèle logistique (croissance de plantes et d'animaux) : $Y = Y_{max}(1-\exp(-\alpha X)$. On prend $Y' = \ln(Y/Y_{max} - 1)$ et $X' = X$. Problème difficile.

V — Complément : tests statistiques

V.1 — Test d'appartenance d'une moyenne d'échantillon à une population dont moyenne μ et variance σ^2 sont connues

Ca n'arrive jamais ! Mais on peut faire comme si !

On va supposer que X suit $\mathcal{N}(\mu, \sigma^2)$.

Alors X_N suit $\mathcal{N}(\mu, \sigma^2/N)$. On définit donc l'écart réduit :

$$k = \frac{\left|X_N - \mu\right|}{\sigma/\sqrt{N}}$$

On voudrait savoir si l'écart entre X_N et μ est significatif. Quel sens donne-t-on à ce mot ?

On divise la droite réelle en plusieurs intervalles :

– le premier est centré autour de la moyenne. La probabilité associé est 95%. On parlera d'écart non significatif.
– Ensuite écart significatif, très significatif, hautement significatif.

Il suffit donc de comparer l'écart réduit aux trois valeurs précédentes.

Exemple Dans un population, une variable aléatoire est distribuée normalement autour d'une moyenne $\mu = 10$, $\sigma^2 = 25$.

On mesure une valeur $X = 15$ cela vous surprend-il beaucoup ?

Donner les intervalles dans lesquelles un écart est significatif, très significatif, etc.

$[2,8\ ;\ 14\,.\,3]\ [-1,1\ ;\ 20,1]\ [-3,9\ ;\ 24]\ [-7,5\ ;\ 27,5]$

V.2 — Test d'appartenance d'une moyenne d'échantillon à une population dont moyenne μ et variance inconnue

Dans ce cas, on calcul l'écart avec la variance mesurée $s^2(X)$:

$$k = \frac{\left|X_N - \mu\right|}{s(X)/\sqrt{N}}$$

Cela introduit un biais qui interdit d'utiliser une loi normale comme loi de référence.

En fait, t suit une loi différente, la loi de Student. Cette loi dépend notamment du « nombre de degré de libertés »: nombre de grandeurs observé moins le nombre de relations qui les lient. Une relation les lie, puisque la moyenne doit faire μ. Il y a $N - 1$ degré de libertés.

Exemple

On pèse plusieurs souris d'un élevage et on trouve les valeurs suivantes :

20,4 ; 18,4 ; 19,2 ; 19,5 ; 22 ; 21,1 ; 19,2 ; 18,6

Donner une estimation de μ. (19.8)

Construire un intervalle de sécurité de la moyenne au niveau 0,95. (18,4 ; 21,1)

On pèse une souris à 23g. Est-elle obèse ?

V.3 — Comparaison de deux moyennes d'échantillons dont les variances sont connues.

Dans cette problématique, il n'y a plus de populations de référence. Les deux sont « équivalentes ».

Cette fois, on considérera que c'est la variables $X_1 - X_2$ qui est normale.

Son écart type est alors $\sqrt{\dfrac{\sigma_1^2}{N} + \dfrac{\sigma_2^2}{N}}$.

L'écart réduit s'écrit donc

$$t = \frac{|X_1 - X_2|}{\sqrt{\frac{\sigma_1^2}{N} + \frac{\sigma_2^2}{N}}}$$

Exemple On mesure une même substance A et B par deux procédés, donc la variabilité statistique est connues, valant $\sigma_A^2 = 105$ et $\sigma_B^2 = 26$.

En faisant une mesure sur un même échantillon par ces deux procédés, on trouve deux valeurs différentes, et on veut savoir si l'écart est significatif.

On effectue 50 dosages avec A et on trouve $X_A = 24,8$, et 30 dosages avec B et on trouve $X_B = 27,5$.

Qu'en pensez-vous ? (non significative)

A partir de quelle différence en valeur absolue peut-on en déduire que l'écart est significatif, très, etc. ?

V.4 — Comparaison de deux moyennes d'échantillons dont les variance sont inconnues.

Si $N_1 + N_2 > 20$ aucun des deux n'étant plus petit que 5, alors on prend l'écart réduit

$$t = \frac{|X_1 - X_2|}{\sqrt{\frac{s_1^2}{N} + \frac{s_2^2}{N}}}$$

et on le compare à une loi de Student à $N_1 + N_2 - 2$ degré de libertés.

Pratique d'un test statistique pour comparer deux échantillons

Pourquoi faire ? Comparer les moyennes inconnues de deux échantillons.

Dans quels cas ? Les conditions suivantes doivent être vérifiées ou supposées :

1. Les échantillons sont statistiquement homogènes.
2. Les échantillons ont été sélectionnés indépendamment l'un par rapport à l'autre (indépendance d'un point de vue statistique).
3. Soit n_1 le nombre de mesure du premier échantillon et n_2 celui du second échantillon. Si $n_1 + n_2 < 15$, ce test ne fonctionne que dans des cas très sûrs (à voir au cas par cas). Si $15 \leqslant n_1 + n_2 < 40$, ce test fonctionne, mais la conclusion est à amener avec beaucoup de précautions. Si $n_1 + n_2 \geqslant 40$, on peut faire confiance à la conclusion statistique de ce test.

Test détaillé Ce test est à pratiquer pour la ou les conclusions les plus importantes de votre TIPE. Il nécessite un peu de soin, et surtout de bien en comprendre la démarche statisitique car vous pourrez être interrogé dessus.

1. Faire une hypothèse statistique H_0 parmi l'une des quatre suivantes :

$$\begin{array}{ll} (1) \quad \mu_1 - \mu_2 = 0\,; & (2)\ \mu_1 - \mu_2 \neq 0\,; \\ (3) \quad \mu_1 - \mu_2 > 0\,; & (4)\ \mu_1 - \mu_2 < 0\,; \end{array}$$

   Les hypothèses (1) et (2) sont dites « bilatérales » et les hypothèses (3) et (4) sont dites « unilatérales ».

2. Choisir ensuite un niveau de signification (d'habitude 5 %). On en déduit un seuil de tolérance, sur la table de la loi normale.

Niveau de signification	5%	1%	0,1%
Seuil de tolérance, test unilatéral	1,96	2,6	3,3
Seuil de tolérance, test bilatéral	1,64	2,33	3,09

3. Calculer l'écart réduit

$$t^* = \frac{\left|\overline{X_1} - \overline{X_2}\right|}{\sqrt{\dfrac{(\overline{\sigma_1})^2}{n_1} + \dfrac{(\overline{\sigma_2})^2}{n_2}}}$$

4. Comparer l'écart réduit t^* au seuil de tolérance s.
5. Rejeter l'hypothèse H_0 si t^* est supérieur au seuil de tolérance s. Dans le cas contraire ne pas la rejeter (ne pas l'accepter pour autant !).

   Rédiger d'abord une conclusion en termes statistiques.

   Rédiger ensuite une conclusion dans les termes de la problématique. Elle fera logiquement référence à l'hypothèse contraire à H_0 : en effet la conclusion « l'hypothèse H_0 est rejetée » sera « statistiquement significative ».

Test simplifié Dans de nombreuses situations, on veut pouvoir affirmer rapidement que deux expériences ont eu des résultats différents. Voici un raisonnement simple qui permet de répondre grossièrement à cette problématique.

Soit un échantillon 1. Si on suppose que la moyenne calculée $\overline{X}_1$ suit une loi normale de paramètres $\overline{X}_1$ (moyenne) et $\overline{\sigma}_1$ (écart-type), alors seulement 13% des mesures seront en dehors de l'intervalle $\left[\overline{X}_1 - 2\overline{\sigma}_1 \,;\, \overline{X}_1 + 2\overline{\sigma}_1\right]$ et 30% des mesures sont en dehors de $\left[\overline{X}_1 - \overline{\sigma}_1 \,;\, \overline{X}_1 + \overline{\sigma}_1\right]$.

Ainsi, si la moyenne $\overline{X}_2$ d'un autre échantillon est en dehors de l'intervalle $\left[\overline{X}_1 - 2\overline{\sigma}_1 \,;\, \overline{X}_1 + 2\overline{\sigma}_1\right]$, alors on pourra penser que les deux expériences ont amené des résultats différents.

Bien entendu, il faut mener le raisonnement de façon symétrique car rien ne permet de favoriser statistiquement un échantillon par rapport à l'autre. Il faut donc aussi vérifier que $\overline{X}_1$ est en dehors de l'intervalle $\left[\overline{X}_2 - 2\overline{\sigma}_2 \,;\, \overline{X}_2 + 2\overline{\sigma}_2\right]$.

Bref, le test consiste à vérifier que que $\left|\overline{X}_1 - \overline{X}_2\right|$ est supérieur à la fois à $2\overline{\sigma}_1$ et $2\overline{\sigma}_2$. Le niveau de signification est médiocre : 13%.

On peut aussi se limiter à comparer $\left|\overline{X}_1 - \overline{X}_2\right|$ à $\overline{\sigma}_1$ et $\overline{\sigma}_2$ mais le niveau de signification est encore plus mauvais : 30%.

VI — Compléments : autres exemples

Exemple Des dosages de calcium sur deux échantillons de yaourt ont donné les résultats suivants :

1ère échantillon : 11 yaourts, moyenne 3,92 variance 0,3130

2ère échantillon : 9 yaourts, moyenne 4,18 variance 0,4231

L'écart des moyennes est-il significatif ? non

Exemple Des truites sont mesurées sur deux échantillons.

Le premier est composé de 50 truites d'élevage et donne $X_a = 158{,}86$ mm avec $s_a^2 = 37,18$.

Le premier est composé de 67 truites de rivières et donne $X_a = 134{,}36$ mm avec $s_b^2 = 25,92$.

L'écart des moyennes est-il significatif ? oui (loi normale)

Petite subtilité : on ne doit pas faire de test bilatéral, mais plutôt latéral gauche. Quelle sont les nouvelles bornes à prendre en compte ?

Exemple

On se pose la question de savoir si une nourriture carencée à une influence sur la croissance d'une population de souris.

Pour cela, on observe quatre populations de souris. Une population de mâle et une population de femelles, toutes deux nourries normalement, une population de mâle et une population de femelle reçevant une nourriture carencée (vitamines et fer). On observe l'évolution du poids des souris en fonctions du temps.

Commentaires sur ce protocole :

– Une souris est un système très compliqué. De l'observation du poids de deux souris (une souris témoin et une souris « carencée ») on ne pourrait rien déduire. L'éventuelle différence de poids pourrait être dû à une multitude de paramètres incontrôlables (complexité) ou inconnus (ignorance) ;
– en observant toute une population, on fera une moyenne. L'hypothèse est qu'il existe un état « moyen » (on devrait dire « témoin »), et que tous les écarts incontrôlables vont se compenser. La moyenne des observations effectuées sera supposée significatives.
– L'équivalent de la question que se posait Darwin est de savoir si l'écart entre les moyennes mesurées est significatif.

Exemple Les coucous sont réputés pour pondre leurs œufs dans le nid d'une espèce hôte. Les œufs sont adoptés par les hôtes et les poussins sont ensuite nourris. Une étude menée en 1940 a démontré que les coucous reviennent année après année sur le même territoire, et laissent leurs œufs dans le nid d'une espèce hôte particulière, uniquement dans leur territoire d'adoption.

On émet l'hypothèse que des sous-espèces géographiques de coucous pourraient apparaître, chacune d'entre elles avec une espèce hôte. La sélection naturelle aurait favorisé les coucous les plus capables de faire adopter leus œufs par une espèce hôte particulière.

Les données suivantes vous permettent-elles d'affirmer ou d'infirmer cette hypothèse ?

VII — Complément : Test χ^2

Il s'agit de comparer les fréquences mesurées d'une modalité à des fréquences théoriques. Une modalité est un phénotype, l'apparition d'un symptôme d'une maladie, etc.

L'exemple classique d'application du test est l'expérience de Mendel. Chez les pois, le caractère couleur est codé par un gène présentant deux formes allèles C et c, correspondant aux couleurs jaune et vert. Le jaune est dominant, le vert récessif. La forme, rond ou ridé, est portée par un autre gène à deux allèles R (dominant) et r (récessif). Si on croise deux individus dont le génotype est CcRr, on peut obtenir 16 génotypes équiprobables. Les descendants seront jaunes et ronds dans 9 cas sur 16, jaunes et ridés dans 3 cas sur 16, verts et ronds dans 3 cas sur 16, verts et ridés dans 1 cas sur 16. Dans un de ses expériences, Mendel a obtenu les résultats suivants.

	Jaune Rond	Jaune Ridé	Vert Rond	Vert Ridé
Effectif	315	101	108	32
Fréquence mesurée f_i	0,567	0,182	0,194	0,058
Fréquence théorique p_i	9/16	3/16	3/16	1/16

La taille de l'échantillon est N, à chaque modalité i correspond une variable aléatoire X_i dont la fréquence $p_i = E(X_i)/N$ est supposée connue (par un modèle théorique, une étude préliminaire, une hypothèse, etc.). La question est de savoir dans quelle mesure les fréquences mesurées s'accordent avec les fréquences théoriques.

Faisons l'hypothèse que les modalités observées sont indépendantes et forment un système complet (elles sont incompatibles et exhaustives). Dans ce cas, chacune suit une loi binomilae $X_i \hookrightarrow \mathscr{B}(N, p_i)$. De plus

$$P(X_1 = N_1 \cap X_2 = N_2 \cap \cdots \cap X_n = N_n) = \frac{N!}{N_1!N_2!\cdots N_n!} p_1^{N_1} p_2 N_2 \cdots p_n^{N_n}$$

On parle de loi multinomiale. On appelle χ^2 à $N-1$ degré de liberté la v.a.r. mesurant l'écart entre les valeurs mesurées et les valeurs théoriques :

$$\chi^2 = \sum_{i=1}^{n} \frac{(N_i - Np_i)^2}{Np_i}$$

Dans le cas $n = 2$, on observe que $\chi^2 = \left[\frac{N_1 - Np_1}{\sigma_1}\right]^2$ qui est le carré d'une $\mathscr{N}(0,1)$.

Dans le cas général,

$$\chi^2 = \sum_{i=1}^{n-1} \left[\frac{N_i - Np_i}{\sigma_i}\right]^2$$

la somme des carrés de $n-1$ variables aléatoires indépendantes suivant la loi $\mathscr{N}(0,1)$. Il est donc aisé d'en établir une table.

Donner l'allure. L'espérance de χ^2 est $n-1$.

Le test du χ^2 s'applique quand les hypothèses suivantes sont vérifiées :

1. Les individus de l'échantillon sont choisis au hasard ;
2. les données sont calculées sous la forme de fréquence (notamment pas de pourcentage ;
3. les modalités observées sont indépendantes et forment un système complet (elles sont incompatibles et exhaustives) ;
4. les moyennes théoriques doivent être supérieures à 5.

Nombre de garçons dans une famille de 4 enfants. L'exemple concerne 10 000 familles.

Garçons	0	1	2	3	4
Fréquence mesurée	0.0572	0.2329	0.3758	0.2632	0.0709
Fréquence théorique (hypothèse 1)	1/16	4/16	6/16	4/16	1/16
Fréquence théorique (hypothèse 2)	0.0556	0.2356	0.3744	0.2644	0.0700

Dans l'hypothèse 1, on suppose les naissances indépendantes et la probabilité d'avoir un garçon de 1/2.

Le χ^2 est $34,47$.

Dans l'hypothèse 2, on suppose les naissances indépendantes et la probabilité d'avoir un garçon de $0,514425$ (obtenu en divisant le nombre de garçon par le nombre d'enfants).

Le χ^2 est $0,9883$. Le nombre de degré de liberté n'est plus que 3.

Pratique du test du χ^2.

Pourquoi faire ? Comparer des fréquences observées à des fréquences théoriques.

Dans quels cas ? Les conditions suivantes doivent être vérifiées ou supposées :

1. Les échantillons sont statistiquement homogènes.
2. Les échantillons ont été sélectionnés indépendamment (d'un point de vue statistique).
3. Les données sont calculées sous la forme de fréquence (notamment pas de pourcentage).
4. Les modalités observées sont indépendantes et forment un système complet (elles sont incompatibles et exhaustives).
5. Les moyennes théoriques Np_i doivent être supérieures à 5. Si ce n'est pas le cas, on peut procéder à des regroupements de modalités.
6. Le nombre de degré de liberté (en général $N-1$) n'est pas supérieure à 30 (sinon on procéde à des regroupements de modalités).

Comment ? 1. Choisir d'abord un niveau de signification (d'habitude 5 %).

2. Faire l'hypothèse statistique H_0 suivante : *« L' écart entre la répartition observée et la répartition théorique n'est pas significatif. »*
   Le test du χ^2 est toujours unilatéral.
3. En déduire le seuil de tolérance s, d'après la table donnée.
4. Calculer le χ^2 :
$$\boxed{\chi^2 = \sum_{i=1}^{n} \frac{(N_i - Np_i)^2}{Np_i}}$$
5. Comparer χ^2 au seuil de tolérance s.

Pot		1	2	3	4	5	6	7	8
Δ taille	cm (± 0,32 cm)	15,6	-21,3	2,5	5,1	1,9	7,3	8,9	13,0
Pot		9	10	11	12	13	14	15	
Δ taille	cm (± 0,32 cm)	4,4	9,2	17,8	7,6	23,8	19,1	-15,2	

Minimum	-21,3	Maximum	23,8	Médiane	7,62
Moyenne	6,65	Écart-type	11,98	Écart réduit	2,15

FIGURE I.2 — *Différences de taille pour 15 paires de plants de maïs (fertilisation croisée moins auto-fertilisé) relevées par Charles Darwin et données statistiques associées.*

6. Rejeter l'hypothèse H_0 si χ^2 est supérieur au seuil de tolérance s (hypothèse d'un écart trop important) ou inférieur au seuil de tolérance (hypothèse d'un écart trop faible qui correspond à une falsification des données ou à l'existence d'un phénomère de régulation).

   Rédiger d'abord une conclusion en termes statistiques.

   Rédiger ensuite une conclusion dans les termes de la problématique. Elle doit être toujours faite en référence à l'hypothèse contraire à H_0 : en effet si l'hypothèse H_0 est rejetée, ce résultat aura une « signification statistique ».

VIII — How a Student helps Charles Darwin.

Parmi les nombreux travaux qui ont occupé le grand Charles Darwin, il y en a un qu'il n'a hélas pas pu mené à bien.

Darwin a voulu mettre en évidence l'influence de l'auto-fertilisation sur la taille des plantes. Pour cela, il plantait dans un même pot deux graines d'une même plante. Les deux graines provenait de la même population de plantes, mais l'une des graines avait été obtenue par auto-fertilisation, l'autre par fertilisation croisée (normale).

Il a mené cette expérience chez lui, dans sa serre, onze années durant, avec plusieurs espèces de plantes. Mais l avait peu d'individus à observer, il a donc récolté peu de données. Vu le manque de résultat, Darwin était incapable de conclure son expérience, c'est-à-dire de répondre à la question : les différences de taille observées sont-elle significatives ?

Darwin fît parvenir ses résultats à Francis Galton, un éminent statisticien de l'époque, qui lui répondit proprement qu'aucune méthode statistique n'existait à ce jour pour exploiter rigoureusement ces données.

William Gosset était un employé de la célèbre brasserie de bière irlandaise Guinness. Son travail consistait à dépouiller les résultats des expériences menées pour améliorer la culture de l'orge. Gosset se trouvait dans la même situation que Darwin,

à savoir que du fait des contraintes liées à l'expérience, il ne pouvait récolter que peu de données. Gosset était un autodidacte qui a continué à étudier toute sa vie, suivant notamment à l'université des cours de statistiques, ceux de Karl Pearson. Il fît part à Pearson de ses difficultés, mais celui-ci, qui dépouillait régulièrement des expériences fournissant plusieurs centaines de mesures, ne trouvait pas cette problématique bien excitante.

Gosset s'entêta. Il publie enfin un article [4] portant sur le traitement statistique des expériences de petites tailles. Ces idées furent ensuite reprise par de nombreux statisticiens, et nous allons aujourd'hui profiter de ce travail.

4. Pour éviter qu'un secret industriel soit indûment éventé, Guinness interdisait strictement à ses employés de publier un quelconque travail scientifique. Pour contourner l'interdiction, Gosset publia son article en utilisant un pseudonyme : Student (étudiant).