

Statistique et probabilité au lycée : quelques compléments de formation à l'usage des enseignants

Il n'est pas question ici de faire un cours de probabilité, ni même un cours de statistique : il en existe d'excellents (voir la bibliographie). L'objectif de ces quelques pages est de montrer qu'il est possible de démarrer avec quelques éléments... sachant qu'il y faudra bien évidemment un temps d'approfondissement...

Les programmes de collège et de lycée font aujourd'hui la part belle à l'enseignement de la statistique.... alors il faut bien s'y mettre ! Voici quelques citations qui doivent nous aider à comprendre les enjeux de cette formation.

Quelques citations officielles

1. Lu dans le rapport de la commission Kahane (2002)

« Pour comprendre l'actualité, une formation à la statistique est aujourd'hui indispensable ; c'est une formation qui développe des capacités d'analyse et de synthèse et exerce le regard critique. Le langage élémentaire de la statistique (avec ses mots tels que moyenne, dispersion, estimation, fourchette de sondage, différence significative, corrections saisonnières, espérance de vie, risque, etc.) est, dans tous les pays, nécessaire à la participation aux débats publics : il convient donc d'apprendre ce langage, ses règles, sa syntaxe, sa sémantique ; l'enseignement de la statistique étant, par nature, associé à celui des probabilités, il s'agit en fait d'une " formation à l'aléatoire ". »

2. Lu dans les programmes officiels

- ... l'interprétation fréquentiste est très importante pour les applications des probabilités dans des situations de la vie courante. Elle permet en outre de donner une justification des calculs de probabilités dans des expériences à deux épreuves. L'approche fréquentiste exige que des fréquences soient observées expérimentalement (*programme de 3^e*).
- L'esprit statistique naît lorsque l'on prend conscience de l'existence de fluctuation d'échantillonnage (*programme de lycée*).
- Le choix est d'aller de l'observation vers la conceptualisation et non d'introduire d'abord le langage probabiliste pour constater ensuite que tout se passe comme le prévoit la théorie (*idem*).
- Le lycée a pour perspectives d'acquérir une expérience de l'aléatoire et d'ouvrir le champ du questionnement statistique, de voir dans un cas simple ce qu'est un modèle probabiliste et d'aborder le calcul des probabilités (*accompagnement des programmes de seconde*).
- Modéliser une expérience aléatoire, c'est lui associer une loi de probabilité (*accompagnement des programmes de première ES et S*).
- L'élève devra être capable de poser le problème de l'adéquation à une loi équirépartie (*idem*).

Former à l'aléatoire ! Voilà les enjeux ! Le hasard, c'est quoi ?

Le lecteur curieux pourra lire l'introduction que faisait à son cours de probabilité en 1912, H. Poincaré (facile à trouver en tapant "cours de probabilité de Poincaré")...

En ce qui nous concerne nous allons simplement évoquer quelques éléments qui doivent permettre à tout enseignant de mathématiques de s'attaquer à cette partie du cours.

Le statisticien, pour répondre à une question concernant une population, ne dispose, en général, que des informations relatives à un échantillon de cette population. Comment peut-il, à partir de cette information partielle, apporter des éléments de réponse pour qu'il soit possible de prendre une décision pour l'ensemble de la population ? Répondre à cette dernière question, tel est l'enjeu de la statistique inférentielle dont les outils sont la théorie de l'estimation et celle des tests d'hypothèses.

I. Généralités

À partir d'un **échantillon**, on veut « estimer » la valeur $\hat{\theta}$ (inconnue) d'un paramètre θ de la population de laquelle est issu l'échantillon (supposé obtenu par un échantillonnage aléatoire avec remise).

Pour cela, on modélise la situation, en supposant que tous les individus de la population ont la même probabilité d'appartenir à l'échantillon, et on construit un **estimateur** noté T .

T est donc une variable aléatoire qui, à chaque échantillon, associe la valeur notée $\hat{\theta}$ du paramètre θ correspondant à cet échantillon : cette valeur est une **estimation ponctuelle** de θ (c'est la réalisation de l'estimateur T obtenue à partir de l'échantillon observé).

On étudie alors les propriétés des estimateurs : biais, écart quadratique moyen, convergence...

II. Propriétés d'un estimateur

1. Biais d'un estimateur

On dit que T , estimateur d'un paramètre θ , est sans biais si son espérance est θ .

De façon plus générale : le réel $E(T) - \theta$ est appelé **biais de T** , noté $B(T)$: $B(T) = E(T) - \theta$ où $E(T)$ est l'espérance de T .

Autrement dit, un estimateur est sans biais ou de biais nul) si, en « moyenne » le résultat est θ (autrement dit, si la moyenne des estimations obtenues à partir de tous les échantillons est θ).

Remarques :

- Le biais d'un estimateur peut dépendre de la taille n de l'échantillon ; dans ce cas si la limite de $B(T_n)$ est 0 lorsque n tend vers l'infini on dit **que T_n est asymptotiquement sans biais**.
- Lorsque l'on a le choix, il vaut mieux un estimateur sans biais... quoique... un estimateur ayant un petit biais mais une « variabilité » réduite est « meilleur » qu'un estimateur sans biais mais avec une grande « variabilité », d'où la nécessité de creuser un peu ...

2. Écart quadratique moyen

Si T est un estimateur d'un paramètre θ , le réel $E((T - \theta)^2)$, noté $Eqm(T)$ est appelé **écart quadratique moyen de T** : $Eqm(T) = E((T - \theta)^2) = V(T) + (B(T))^2$ (cette dernière égalité s'obtient par un calcul utilisant la linéarité de l'espérance, la définition de la variance et du biais ; $V(T)$ est la variance de T).

Autrement dit $Eqm(T)$ est « la moyenne des carrés des écarts entre la *réalité* et l'observation sur les échantillons ». Ce réel positif mesure l'efficacité d'un estimateur : plus il est petit et plus l'estimateur donnera des résultats proches de la réalité...

On choisira, en règle générale, parmi tous les estimateurs possibles, celui qui a le plus petit écart quadratique moyen.

3. Estimateurs convergents

Un estimateur dépend, en général, de la taille n de l'échantillon ; dans ce cas on dit qu'il est convergent si :

$\lim_{n \rightarrow +\infty} P(|T_n - \theta| < \varepsilon) = 1$. Ceci nécessite la notion de convergence en probabilité et c'est pourquoi nous nous contenterons du cas particulier des estimateurs sans biais ou asymptotiquement sans biais et, dans ce cas, la convergence de T est équivalente à : $\lim_{n \rightarrow +\infty} (V(T_n)) = 0$.

III. Construction d'estimateurs

Pour construire un estimateur d'un paramètre θ , il est possible d'utiliser des méthodes plus ou moins empiriques (l'expérience permet d'avoir une bonne idée de l'efficacité de cet estimateur). Cependant, pour pouvoir mesurer les qualités d'un estimateur il vaut mieux utiliser des méthodes plus mathématiques ; nous allons en particulier citer ici la **méthode du maximum de vraisemblance**.

Soit une variable aléatoire X dépendant d'un paramètre θ , dont je veux « estimer » la valeur.

Soit (X_1, X_2, \dots, X_n) un échantillon aléatoire de taille n de la variable aléatoire X et (x_1, x_2, \dots, x_n) une de ses réalisations. Je veux définir un estimateur T_n de θ , qui me donnera, en l'appliquant à (x_i) , une estimation $\hat{\theta}$ de θ . Le principe repose sur l'idée que ce que j'ai observé est ce qui avait le plus de « chance » de se produire...

Soit la fonction notée ν , appelée **fonction de vraisemblance** définie par :

$$\nu : (x_1, x_2, \dots, x_n) \mapsto \begin{cases} P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) & \text{si } X \text{ est discrète} \\ f(x_1) \times f(x_2) \times \dots \times f(x_n) & \text{si } X \text{ est continue de densité } f \end{cases}$$

L'estimateur de θ noté L_n (L comme Likelihood = vraisemblance) est celui qui, à l'échantillon obtenu, associe $\hat{\theta}$ la valeur de θ qui rend maximale la fonction ν (ou $\ln \nu$, car \ln est croissante) ; il suffit donc d'étudier la fonction ν de θ (ou plus souvent $\ln(\nu)$) pour trouver la valeur en laquelle elle admet un maximum...

IV. Estimations ponctuelles des paramètres usuels

Supposons la situation modélisée : soit X une variable aléatoire avec $E(X) = \mu$ et $V(X) = \sigma^2$.

On dispose d'un échantillon (X_1, X_2, \dots, X_n) de taille n (n variables aléatoires indépendantes de même loi que X) et d'une réalisation (x_1, x_2, \dots, x_n) de cet échantillon.

1. Estimation d'une moyenne

On veut estimer, à partir des données (x_1, x_2, \dots, x_n) , la moyenne μ inconnue.

La variable aléatoire \bar{X}_n définie par $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ est un estimateur sans biais et convergent de μ .

Ceci permet donc de prendre : $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$. En effet, il est facile de montrer que : $E(X_n) = E(X) = \mu$ et

$$\lim_{n \rightarrow +\infty} (V(\bar{X}_n)) = 0 \text{ car } V(\bar{X}_n) = \frac{\sigma^2}{n}.$$

2. Estimation d'une variance

On veut estimer, à partir des données (x_1, x_2, \dots, x_n) , la variance σ^2 inconnue.

On montre que la variable aléatoire S_n^2 définie par $S_n^2 = \frac{1}{n} \left(\sum_{i=1}^n (X_i - \bar{X}_n)^2 \right) = \left(\frac{1}{n} \sum_{i=1}^n X_i^2 \right) - \bar{X}_n^2$ est un

estimateur biaisé de σ^2 : en effet, $E(S_n^2) = \frac{n-1}{n} \sigma^2$.

En utilisant la linéarité de l'espérance on peut montrer que la variable aléatoire $S_n'^2$ définie par

$$S_n'^2 = \frac{n-1}{n} S_n^2 \text{ est un estimateur sans biais de } \sigma^2 ; \text{ ce qui permet donc de prendre : } \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

On montre que $S_n'^2$ est aussi un estimateur convergent.

3. Estimation d'une fréquence

Nous voulons estimer la proportion τ inconnue, des individus d'une population possédant une certaine propriété. Nous disposons d'un échantillon de taille n dans lequel la proportion des individus possédant la propriété est f . Soit X_i la variable aléatoire qui prend la valeur 1 si l'individu numéro i a la propriété et 0 sinon (variable de Bernoulli de paramètre τ).

La variable aléatoire F_n définie par : $F_n = \frac{1}{n} \sum_{i=1}^n X_i$ est un estimateur sans biais et convergent de τ ; ce qui permet donc de prendre $\boxed{\hat{\mu} = f}$.

En effet il est facile de montrer que : $E(F_n) = \tau$ et $\lim_{n \rightarrow +\infty} V(F_n) = \lim_{n \rightarrow +\infty} \left(\frac{\tau(1-\tau)}{n} \right) = 0$.

V. Intervalles de confiance des estimations

Nous venons de déterminer des estimateurs d'une moyenne, d'une variance et d'une fréquence. Chacun de ces estimateurs donne, pour l'échantillon considéré, une valeur ponctuelle du paramètre cherché (estimation) ; mais les résultats sur la fluctuation d'échantillonnage nous permettent d'affirmer que, même avec des estimateurs sans biais et convergents, il est possible d'obtenir des estimations assez éloignées de la réalité... Bien sûr, nous savons que la probabilité que cette estimation soit très éloignée de la réalité est faible, mais cela reste vague (que veut dire une valeur approchée sans le calcul de l'erreur ?).

Est-il possible de préciser cela ? Oui, par la notion *d'intervalle de confiance*.

Le but est, à partir de l'estimation $\hat{\theta}$ obtenue, en appliquant l'estimateur à l'échantillon et en acceptant un risque d'erreur (*risque α* , souvent 5 % ou 0,05 ou *seuil de confiance $1 - \alpha$*), de construire un intervalle (aléatoire) qui contiendra presque sûrement (avec la probabilité de $1 - \alpha$) la valeur inconnue θ .

Nous allons donc construire, à partir de l'estimation obtenue, un intervalle qui contiendra ou non (et nous ne le savons pas, mais nous avons confiance !) la vraie valeur de θ ... la seule chose que nous savons, c'est que si nous faisons un très grand nombre d'estimations différentes, seulement α % des intervalles de confiance fabriqués ne contiendront pas θ !!!

Nous allons uniquement donner quelques résultats pour une moyenne et une fréquence ; dans ces deux cas, il s'agit de déterminer un intervalle I_c (et donc un réel positif ε) de la forme $I_c =]\hat{\theta} - \varepsilon, \hat{\theta} + \varepsilon[$ tel que l'intervalle I_c contienne presque sûrement (très probablement) θ .

Il va de soi que ε va dépendre de n , du risque α et de la loi suivie par la variable de départ X et aussi de la distribution d'échantillonnage.

1. Intervalle de confiance d'une moyenne

Soit X une variable aléatoire avec $E(X) = \mu$ et $V(X) = \sigma^2$; on veut estimer μ à partir d'un échantillon de taille n (n variables aléatoires indépendantes de même loi que $X : (X_1, X_2, \dots, X_n)$) et d'une réalisation (x_1, x_2, \dots, x_n) .

Si la variable aléatoire X suit une loi normale de paramètres μ et σ alors \bar{X}_n suit aussi une loi normale :

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \text{ et donc } Z = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1).$$

On peut alors calculer ε : $\varepsilon = t_\alpha \times \left(\frac{\sigma}{\sqrt{n}}\right)$ où t_α vérifie : $P(-t_\alpha \leq Z \leq t_\alpha) = 1 - \alpha$; malheureusement ceci suppose que l'on connaisse σ , ce qui n'est pas très fréquent.

• Que se passe-t-il si on ne connaît pas σ ?

On estime alors σ^2 par $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ et on démontre que la variable aléatoire T définie par

$$T = \frac{\bar{X}_n - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}} \text{ suit une loi de Student à } n - 1 \text{ degrés de liberté ; on peut alors calculer } \varepsilon : \varepsilon = t_\alpha \times \left(\frac{\sigma}{\sqrt{n}}\right) \text{ où } t_\alpha$$

vérifie : $P(-t_\alpha \leq T \leq t_\alpha) = 1 - \alpha$.

• Si n est assez grand ($n \geq 30$), alors quelle que soit la loi de X , \bar{X}_n suit "quasiment" une loi normale et pour déterminer ε (ou plus exactement t_α) on procède comme précédemment.

• Si $n < 30$ et que l'on ne connaît pas la loi de X , alors on ne dispose pas de méthode générale : on fait une nouvelle étude avec un échantillon plus grand !

2. Intervalle de confiance d'une proportion

Soit une population dans laquelle une proportion τ , inconnue, d'individus possède une certaine propriété. On y prélève un échantillon aléatoire, de taille n . On veut estimer, à partir de cet échantillon la proportion τ . On détermine la proportion f des individus de cet échantillon qui possèdent la propriété.

Soit la variable aléatoire F_n définie par : $F_n = \frac{1}{n} \sum_{i=1}^n X_i$ où la variable X_i prend la valeur 1 si l'individu n° i a la propriété et 0 sinon (l'image de l'échantillon par F_n est donc f).

Chacune des variables aléatoires X_i suit une loi de Bernoulli de paramètre τ et donc $\sum_{i=1}^n X_i$ suit une loi binomiale de paramètres (n, τ) ; on connaît bien son espérance ($n\tau$) et sa variance ($n\tau(1 - \tau)$) ; il est donc facile de déduire que : $E(F_n) = \tau$ et $V(F_n) = \frac{\tau(1 - \tau)}{n}$.

F_n est un estimateur sans biais, convergent de τ ; ce qui permet donc de prendre : $\hat{\tau} = f$.

En résumé : la fréquence f d'individus possédant la propriété dans l'échantillon est une "bonne" (la "meilleure" ?) estimation ponctuelle de τ car c'est la réalisation de F_n , estimateur sans biais, convergent, de variance minimale, de τ .

Ici, il va être relativement facile de mesurer la qualité de cette estimation. En effet, dans tous les cas, la variable aléatoire nF_n suit une loi binomiale de paramètres (n, τ) : déterminer un réel positif ε tel que l'intervalle $]\hat{\theta} - \varepsilon, \hat{\theta} + \varepsilon[$ contienne très probablement la valeur τ ne relève que du calcul de probabilité.

En réalité, on calcule le réel t_α vérifiant l'égalité $\varepsilon = t_\alpha \times \left(\frac{\sigma}{\sqrt{n}} \right)$ où $\sigma^2 = V(F_n)$.

- Si n est petit, il est possible de faire le calcul de t_α directement en utilisant la loi binomiale.
- Si n est grand ($n \geq 30$) et $nf > 10$ et $nf(1 - f) > 10$ le calcul de t_α (et donc de ε) se fait en utilisant l'approximation normale.
- Si n est grand ($n \geq 30$) mais f petit, le calcul de t_α (et donc de ε) se fait en utilisant l'approximation de la loi binomiale par une loi de Poisson.

Oui mais concrètement... ?

On se trouve, la plupart du temps, dans le deuxième cas, celui où la loi binomiale peut être approchée par une

loi normale et l'intervalle est à peu près : $\left[f - t_\alpha \sqrt{\frac{f(1-f)}{n}} ; f + t_\alpha \sqrt{\frac{f(1-f)}{n}} \right]$ où t_α est le quantile $1 - \alpha$ de

la loi normale centrée réduite.

On peut retenir que : si $\alpha = 5\%$, alors $t_\alpha = 1,96$; si $\alpha = 1\%$, alors $t_\alpha = 2,58$; si $\alpha = 10\%$, alors $t_\alpha = 1,64$.

La marge d'erreur, $t_\alpha \sqrt{\frac{f(1-f)}{n}}$, valeur approchée de $t_\alpha \sqrt{\frac{\tau(1-\tau)}{n}}$ peut être majorée par $\frac{t_\alpha}{2\sqrt{n}}$.

Cette majoration permet de calculer, *a priori*, la taille d'échantillon nécessaire pour être "sûr", au niveau $1 - \alpha$ d'atteindre une précision donnée pour τ (document d'accompagnement des programmes de seconde).

Remarques :

- Dans le cas d'un tirage sans remise dans une population finie de taille N , on obtient des intervalles en utilisant les mêmes formules que dans le cas avec remise, mais en remplaçant $\sqrt{\frac{f(1-f)}{n}}$ par $\sqrt{\frac{f(1-f)}{n} \times \frac{N-n}{N-1}}$. On constate, bien entendu, que, si n est "petit" par rapport à N , le facteur d'exhaustivité $\frac{N-n}{N-1}$ est très proche de 1, et donc, dans ce cas, les intervalles de confiance sont les "mêmes", que le tirage se fasse avec ou sans remise.

- Lorsque l'échantillon n'est pas aléatoire (mais fabriqué, par exemple, à partir de la méthode des quotas) il n'est pas possible de déterminer un intervalle de confiance ; en réalité on fait, dans ce cas, comme si...

ANNEXES

Annexe 1: quelques éléments bibliographiques (liste subjective et non exhaustive...)

- Le jeu de la science et du hasard. La statistique et le vivant. D. Schwartz (Flammarion).
- Chemins de l'aléatoire. Le hasard et le risque dans la société moderne. D. Dacunha-Castelle (Flammarion).
- Simulation et statistique en seconde. Collectif (Université Paris Nord) Inter-IREM.
- Enseigner les probabilités au lycée, Commission INTER IREM Statistique et probabilités.
- Les trois brochures de la Commission Inter-IREM « Statistique et probabilités » publiées par l'APMEP (n° 143, n° 156 et n° 167).
- Enseigner la Statistique au lycée : des enjeux aux méthodes (Commission inter-IREM « lycées technologiques » n° 112).
- Des statistiques à la pensée statistique (IREM de Montpellier).
- Quelques ouvrages de cours dont : Statistique inférentielle. Idées, démarches et exemples (Daudin – Robin – Vuillet SFDS - P.U. Rennes) et Éléments de statistique de JJ Dreesbeke – Ellipses...
- L'induction statistique au lycée, P. Dutarte, éditions Didier.

Annexe 2 : quelques sites

- T3 France : <http://education.ti.com/>
- Utilisateurs de TI-Nspire : <http://www.univers-ti-nspire.fr/>
- Statistix : <http://www.statistix.fr/>
- SfdS : <http://www.sfds.asso.fr/>
- IREM : <http://www.univ-irem.fr/> avec ses liens sur les différents IREM dont celui de Lille <http://irem-old.univ-lille1.fr/activites/>
- APMEP : <http://www.apmep.asso.fr/> avec ses liens sur les régionales et en particulier la nôtre (<http://apmeploiraine.free.fr/>)...
- Cours de Poincaré : <http://www.futuretgc.com/FTHumanEvolutionCourse/FTFreeLearningKits/01-MA-Mathematics/013-MA13-UN04-12-Probability%20and%20Statistics/Additional%20Resources/Henri%20Poincare%20-%20Calcul%20des%20Probabilites.pdf>
- Travail de l'Irem de Basse Normandie sur la loi de Moivre-Laplace : <http://www.math.ens.fr/culturemath/histoire%20des%20maths/pdf/LoidesGrandsNombres.pdf>