Sensibilisation à la Statistique

Yves Escoufier
Laboratoire de Probabilités et Statistiques, Université Montpellier 2
Place E. Bataillon, 43095 Montpellier CEDEX 5

Le texte qui suit donne la trame d'une séance de sensibilisation à la Statistique faite à l'invitation de l'IUFM et de l'IREM de Dijon pour des professeurs et élèves professeurs de mathématiques des lycées et collèges de l'académie de Bourgogne. L'objectif annoncé aux auditeurs était de parcourir les grandes étapes d'une démarche d'analyse statistique en tentant pour chacune d'entre elles d'inviter à des prolongements personnels.

I) Un thème d'étude et des données

La proposition faite aux auditeurs est de prendre pour thème d'étude la morphologie de la main gauche. Dans ce but, chacun des participants est invité à placer sa main gauche sur une feuille de papier, bien à plat et doigts écartés le plus possible, à tracer le contour de la main avec un crayon et à marquer dans un coin de la feuille H ou F selon qu'il est de sexe masculin ou féminin.
Plusieurs commentaires peuvent être faits à ce moment pour dépasser le thème d'étude lui même . On peut espérer que l'intérêt que les auditeurs portent au thème de l'étude alimentera l'intérêt qu'ils manifesteront pour les éléments de Statistique qui vont être présentés. En ce sens , le choix du thème n'est pas neutre : plus il concerne les auditeurs, plus grandes sont les chances de retenir leur attention pour la Statistique. L'aspect en partie ludique du tracé du contour de la main est aussi un élément susceptible de faciliter l'intérêt. Il faut cependant être prudent : quelle serait ici la réaction d'un auditeur dont la main gauche serait déformée ? Il peut y avoir des réactions négatives au thème proposé ; elles ne doivent pas être négligées.
On peut imaginer que des statisticiens expérimentés s'engagent directement dans l'étude des courbes particulières que sont les contours des mains recueillis. Ce n'est bien sûr pas le cas dans cette séance de sensibilisation. On va donc devoir choisir quelques caractéristiques qui deviendront les objets des études statistiques ultérieures. Des kinésithérapeutes, des ergonomes, des chirurgiens spécialisés dans les soins de la main auraient à coup sûr des suggestions à faire pour le choix de ces caractéristiques : ils ont des connaissances sur le sujet ; ils connaissent les résultats d'études antérieures qui pourraient servir de guide. En bref, l'étude qui sera faite, comme toute étude, ne portera que sur des aspects restreints du phénomène étudié ; la restriction dépend pour une part de nos connaissances antérieures sur le sujet , pour une autre part des outils de mesure et d'enregistrement des données dont nous disposons et de nos capacités à les maîtriser.
Pour l'étude présente, nous retiendrons une mesure de la largeur de la paume et une mesure de la longueur du majeur. A ces deux caractéristiques numériques, s'ajoute la caractéristique qualitative « sexe » dont les deux valeurs sont « homme » et « femme ». On peut ici élargir l'exposé aux différents types de caractéristiques(numériques continues, numériques discrètes, ordinales, qualitatives, dichotomiques), passer du temps sur le concept de variable statistique, sur ses valeurs ou modalités, sur ces fréquences. Une ouverture peut être faite sur les variables statistiques multidimensionnelles.
Les lancés d'une pièce de monnaie ou d'un dé sont souvent pris comme thèmes d'études. Ils permettent effectivement sans grand investissement de disposer de données qui peuvent interpeller des élèves et donc faciliter leur attention en faveur des éléments de Statistique qu'ils permettront d'introduire. Ils présentent un autre avantage. Le nombre des expériences qui peuvent être faites avec une pièce de monnaie ou un dé reste limité. Il devient alors assez naturel de rechercher un procédé technologique qui réagisse comme un dé ou une pièce de monnaie. Le simulateur de nombres uniformes sur (0,1] s'introduit ici naturellement et il est facile en commençant par la simulation de dés pipés ou de pièces truquées de montrer qu'il va permettre de simuler des expériences complexes.

II) La descriptions des données

Afin de permettre aux auditeurs de reproduire eux mêmes les démarches qui vont être faites, le choix est fait d'utiliser « l'utilitaire d' analyse » que l'on trouve dans la rubrique « outil » du logiciel Excel ainsi que la fonction « graphique » de la rubrique « insertion » . Les données recueillies en séance sont données dans le tableau 1. Elles permettent de mettre en œuvre un certain nombres de méthodes de description de données offertes par le logiciel, d'évoquer les problèmes implicites éventuels de leur mise en œuvre, de commenter les résultats.

- le tracé d'un histogramme pose le problème du choix des classes

- il est intéressant de vérifier comment le logiciel calcule effectivement la médiane, les quartiles et plus généralement les quantiles
- Obtenir des graphiques agréables à lire suppose de réaliser des translations des données
- Le logiciel propose un écart type résultant d'une division par n et un autre que l'on ne peut que renvoyer à plus tard divisé par n-1.
- Les analyses faites sur l'ensemble des données peuvent facilement être reprises pour les hommes seuls ou les femmes seules : voilà une bonne introduction pour l'adjectif « conditionnel » en Statistique.

A titre d'illustration, les tableaux 2 et 3 donnent les histogrammes de la variable « paume » pour les femmes et pour les hommes. Les classes ont été choisies identiques pour les deux groupes. Les graphiques mettent bien en évidence que pour les données recueillies, les paumes les plus larges sont masculines et les paumes les moins larges féminines.

III) Statistique et Probabilités

S'il ne s'agissait que de décrire les cinquante mains observées, nous pourrions nous arrêter ici. Mais le statisticien veut le plus souvent étendre à des données potentielles non observées les résultats de ses observations limitées. Pour cela , il a besoin d'un cadre qui le guide et donne un sens aux données partielles qu'il recueille. La théorie des probabilités va le lui fournir.
La question est parfois posée de savoir pourquoi ce sont les enseignants de mathématiques qui doivent enseigner les statistiques dans les lycées. La réponse vient de la nécessité d'utiliser les concepts et les résultats des probabilités, branche des mathématiques , pour donner un sens à la démarche statistique.
En effet, toute extension des résultats obtenus sur un fini observé à un ensemble plus grand qui l'englobe, suppose que l ' observé puisse être considéré comme une réalisation d'un échantillon représentatif de l'ensemble plus grand. On ne peut échapper à la théorie des probabilités pour donner un sens à ces mots : échantillon, représentatif, réalisation.
Un enseignement classique devrait traverser les étapes de l'espace probabilisable et des opérations sur les événements ; il continuerait par une présentation axiomatique des probabilités avec éclairage sur les probabilités conditionnelles et les événements indépendants ; viendraient alors l'espace probabilisé , la notion de variable aléatoire et le théorème de transfert de la probabilité. A ce stade l'étude des modèles classiques de variables aléatoires peut être faite ce qui introduit dans le vocabulaire les mots de « lois et distributions de probabilité ».Le concept de variables aléatoires indépendantes conduit alors à la notion d'échantillon de variables aléatoires indépendantes de même loi. Reste ensuite à parler des réalisations d'un échantillon puis des fonctions définies sur un échantillon et des valeurs prises par ces fonctions sur les réalisations de l'échantillon.
Bien sûr, on ne peut pas faire tout cela de manière rigoureuse dans le secondaire. Mais celui ou celle qui l'enseigne doit le posséder d'autant mieux qu'il doit en donner une version simplifiée mais juste. La simulation peut – elle aider dans cet enseignement ? Certainement. La simulation d'une certaine loi peut être assimilée à la variable aléatoire qui obéit à cette loi ; le résultat de chacun des appels du simulateur est alors considéré comme une réalisation de cette variable. On peut donc visualiser les fluctuations de ces réalisations mais aussi calculer des fonctions de ces réalisations telles que la moyenne ou la variance. Cette approche expérimentale se substitue à l'étude analytique des variations de la fonction de densité.
Dans le même esprit, n appels successifs du simulateur peuvent être assimilés à un échantillon de taille n de la variable qu'il simule. Une réalisation de ces n appels fournit une réalisation de l'échantillon de taille n. Les fluctuations de toutes les fonctions de ces réalisations peuvent être étudiées en répétant les n appels.
Nous venons de dire que la Statistique avait besoin des probabilités : elles lui donne son vocabulaire, ses concepts et les propriétés des objets qu'elle manipule . En retour, par son contact avec les différents domaines dans lesquels elle intervient, la Statistique est sans cesse confrontée à des problèmes nouveaux et à des données de types nouveaux. Elle demande aux probabilités de construire les outils dont elle a besoin et d'en préciser les propriétés. Par là , la Statistique alimente le développement des probabilités. Pour donner deux exemples actuels, évoquons les données spatiales et celles liées au génome. Le statisticien les a rencontrées avant d'avoir les outils adaptés à leur étude. Peu à peu avec les probabilistes, il dégage les concepts nécessaires et met en évidence leurs propriétés.

IV) Estimation et tests

Revenons à l'étude proposée en I. Les mesures ont été faites sur des femmes et des hommes adultes, étudiants ou enseignants de mathématiques. Peut – on d'une manière ou d'une autre étendre les résultats constatés sur les quinze femmes et les trente cinq hommes observés à l'ensemble des femmes et hommes adultes , étudiants ou enseignants de mathématiques ? C'est là tout le champ de l'inférence statistique.

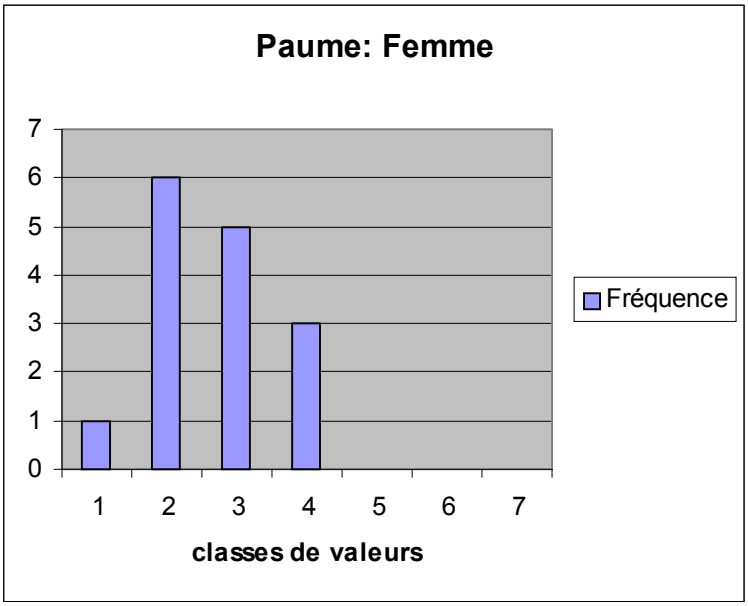
Rien ne pourra être fait sans pouvoir considérer que , par exemple, les quinze femmes étudiées constituent un échantillon représentatif de la population des femmes adultes, étudiantes ou enseignantes en mathématiques ce qui veut dire que les quinze mesures obtenues sont bien une réalisation d'un échantillon de taille quinze de la variable paume pour cette population. C'est donc la procédure de constitution de l'échantillon qui doit être interrogée. Il y a des livres entiers sur ce sujet : échantillons aléatoires ; échantillons stratifiés ; méthodes des quota. Nous n'irons pas plus loin ici sur ce thème.
Cette question traitée, l'inférence statistique se partitionne en deux domaines : l'estimation et les tests.
L'estimation consiste à induire des observations faites sur l'échantillon des informations valables pour la population. Dans sa forme dite paramétrique, elle consiste à postuler le modèle qui régit la variable aléatoire objet de l'étude (on postulera par exemple qu'elle suit une loi normale) et à obtenir de l'échantillon des estimations des paramètres qui déterminent cette loi (la moyenne et la variance pour la loi normale). Ce faisant, on a ajouté aux données disponibles , l'hypothèse d'un modèle pour la variable aléatoire. Plusieurs procédures sont disponibles pour estimer les paramètres. C'est par exemple la méthode dite du maximum de vraisemblance qui conduit au diviseur n – 1 pour l'écart type , proposition du logiciel Excel que nous évoquions en II. Il faut comprendre l'intérêt que représente une telle association d'un modèle théorique à une variable. Si l'association est possible (condition qui sera mise à l'épreuve d'un test), tout ce qui est connu pour le modèle devient applicable à la variable.
L'estimation peut être conduite de manière non paramétrique. Cette approche s'est développée avec la généralisation des ordinateurs. Elle ne demande pas hypothèse sur la forme du modèle qui régit la variable (ou seulement des hypothèses très générales) mais nécessite plus de calculs aussi bien pour obtenir le résultat qui se présente sous la forme d'un graphique que pour l'utiliser.
L'autre grand domaine de l'inférence statistique est celui des tests. Il consiste toujours à se demander si une valeur calculée à partir des données observées est ou non une réalisation anormalement grande ou anormalement petite d'une variable aléatoire dont cette valeur est une réalisation. L'idée sous-jacente est que les probabilités des valeurs anormalement grandes ou anormalement petites sont elles mêmes très petites et qu'un événement de petite probabilité ne doit pas se produire dans une expérience unique, celle d'observer la réalisation d'un seul échantillon. Sur ce sujet aussi , la littérature est foisonnante. Restons proche de notre exemple. La moyenne de la variable paume pour les quinze femmes est 76,6mm. Elle est de 87,1mm chez les trente cinq hommes. Peut – on considérer que cette différence est compatible avec l'hypothèse de l'égalité des moyennes dans les deux populations ? Autrement dit, peut – on considérer que 76,6mm et 87,1mm sont deux fluctuations possibles pour une même variable aléatoire ?
Pour être légitime, le test de comparaison des moyennes doit être précédé d'un test de comparaison des variances. Le tableau 4 fourni par le logiciel Excel montre que l'hypothèse de l'égalité des variances n'est pas rejetée : la variable F objet du test dépasse la valeur trouvée (1,227)avec une probabilité de 0,35 . La valeur trouvée (1,227)n'est donc pas anormalement grande.
Au contraire, dans le tableau 5, la variable T, objet du test, n'est en deçà de la valeur trouvée (-7,26) qu'avec une probabilité extrêmement faible (1,47x10-7). Cette valeur (-7,26) est donc anormalement petite ; elle conduit à rejeter l'hypothèse de l'égalité des moyennes dans la population des femmes et des hommes.
On pourrait ici faire d'autres développements sur les tests et les risques qui leur sont associés ; on sortirait du cadre de cette sensibilisation à la Statistique.
Notons pour terminer qu'un simulateur de la loi de la variable F ou de la variable T, permettrait de construire un histogrammes des valeurs prises par ces variables dans un grand nombre d'appels du simulateur. On pourrait alors situer les valeurs observées dans ces histogrammes et conduire les tests de façon approchée. Cette remarque introduit naturellement certaines pratiques de tests pour des quantités dont on ne connaît pas la forme analytique de la distribution : la forme est approchée par simulation et la valeur observée est située dans cette distribution. De nombreuses méthodes statistiques (tests de permutation, Jackknife, Bootstrap) utilisent aujourd'hui l'ordinateur pour obtenir par des calculs nombreux ce qui ne peut être atteint analytiquement.

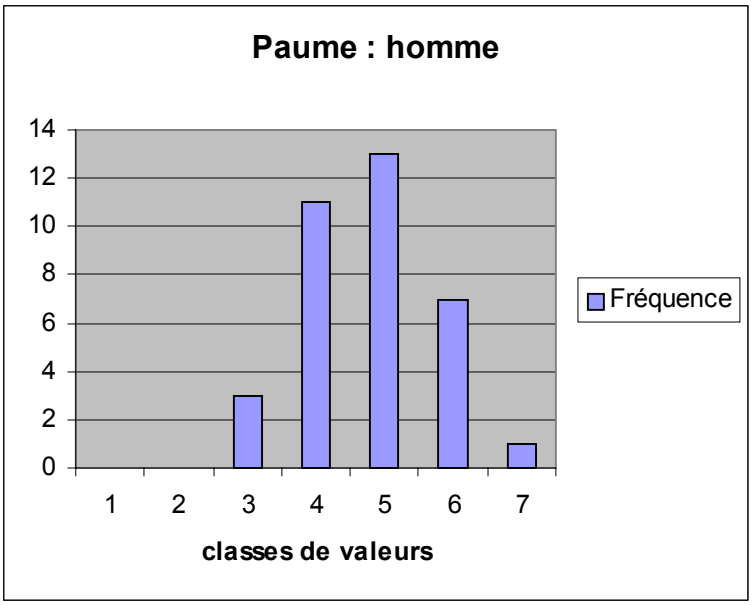
Références : Gilbert Saporta : Probabilités, Analyse des Données et Statistique, Edition Technip, 1990.

paume	doigt	sexe		paume	doigt	sexe
80	80	F		94	89	M
75	78	F		88	84	M
71	88	F		87	85	M
69	67	F		90	95	M
75	75	F		94	95	M
75	74	F		90	84	M
77	85	F		85	79	M
78	85	F		90	84	M
76	70	F		85	83	M
84	83	F		84	91	M
78	85	F		93	87	M
74	75	F		82	88	M
83	85	F		90	83	M
82	81	F		89	83	M
72	87	F		82	76	M
81	88	M		86	94	M
97	95	M		86	83	M
77	80	M		87	84	M
91	85	M		83	83	M
89	83	M		85	86	M
90	83	M		93	90	M
81	80	M		94	76	M
80	76	M		91	90	M
85	80	M		79	75	M
87	78	M		84	84	M

Tableau 1

Classes	*Fréquence*
70	1
75	6
80	5
85	3
90	0
95	0
ou plus...	0





Test d'égalité des variances (F-Test)		
paume		
	homme	*femme*
Moyenne	87,1142857	76,6
Variance	23,2806723	18,9714286
Observations	35	15
Degré de liberté	34	14
F	1,22714387	
P(F<=f) unilatéral	0,35150245	
Valeur critique pour F (unilatéral)	2,28869368	

Test d'égalité des espérances: deux observations de variances égales		
paume		
	femme	*homme*
Moyenne	76,6	87,1142857
Variance	18,9714286	23,2806723
Observations	15	35
Variance pondérée	22,0238095	
Différence hypothétique des moyennes	0	
Degré de liberté	48	
Statistique t	-7,25985928	
P(T<=t) unilatéral	1,4708E-09	
Valeur critique de t (unilatéral)	1,67722419	
P(T<=t) bilatéral	2,9415E-09	
Valeur critique de t (bilatéral)	2,01063358	