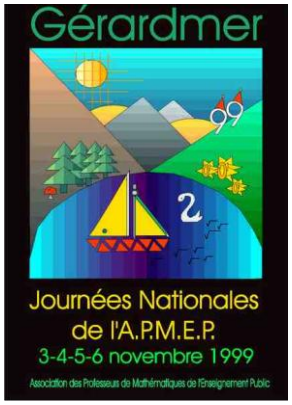


JOURNÉES NATIONALES A.P.M.E.P.
GÉRARDMER
3-6 novembre 1999

Atelier VA13
QUE PEUT-ON FAIRE AVEC
LE PROGRAMME DE STATISTIQUE DE COLLÈGE ?

Bernard PARZYSZ[1]

L'introduction relativement récente d'un chapitre de statistique dans les programmes de tous les niveaux du collège a perturbé un certain nombre d'enseignants de mathématiques, qui ont rechigné à l'enseigner en avançant deux raisons principales:

1° ***Ce n'est pas des mathématiques***. Cela tient peut-être à ce qu'ils ne voient pas clairement comment ce qu'ils ont à enseigner se rattache à leur discipline, les séries statistiques étudiées de façon descriptive au collège conduisant, au lycée, à l'introduction et à l'étude (mathématique) des probabilités, et les modèles ainsi dégagés servant ensuite de fondement à la statistique inférentielle qui permet, entre autres, d'estimer des paramètres à partir d'un échantillon et de comparer entre eux divers échantillons.

2° ***Je n'ai pas été formé pour l'enseigner***. C'est sans doute en partie vrai, car beaucoup d'étudiants de mathématiques terminent leur licence sans avoir jamais rencontré de statistiques à l'université, et la formation continue reste numériquement insuffisante, malgré les stages proposés au niveau académique depuis un certain nombre d'années. La situation peut également s'améliorer grâce à la formation initiale, un nombre croissant d'IUFM incluant un module "Statistique" obligatoire dans la formation de leurs stagiaires de mathématiques.

Au cours de l'atelier nous avons étudié plusieurs exemples montrant comment le manque de culture statistique amène certains journalistes, soit à écrire des affirmations inexactes, soit à présenter des résultats exacts sous une forme inadéquate (voire inacceptable). Pour des raisons de place, je me contenterai ici de traiter, sur un exemple, une démarche statistique possible avec les outils du collège (et d'autres qu'on pourrait y trouver, mais qui n'y figurent pas).

1) La question initiale :

On entend souvent dire que les femmes vivent plus longtemps que les hommes. Que faut-il en penser ?

Pour tenter de répondre à cette question encore vague, il va falloir se renseigner sur l'âge au décès d'hommes et de femmes d'une "même population", et pour cela:

1° déterminer précisément quelle sera cette population

[1] parzysz.bernard@wanadoo.fr

2° aller chercher les données.

Ces deux problèmes sont en fait liés: soit on dispose de données déjà recueillies dans une population donnée (par l'INSEE, l'INED, etc) et il s'agira de travailler à partir de ces données, soit on n'a aucune donnée, et il faudra alors faire le travail initial soi-même (définition de la population étudiée, données à recueillir, saisie...) avant d'entreprendre l'analyse. Il faut qu'au moins une fois les élèves aient entrepris cette démarche de bout en bout, afin qu'ils se rendent compte des problèmes qu'elle pose, mais je je supposerai ici qu'on dispose d'un corpus de données. En l'occurrence l'âge au décès de 180 Canadiens en 1988 (source: *La Presse*, Montréal, édition du 27 juin 1988).

Hommes (effectif: 92)

25	65	79	59	58	77	73	33	75	68	66	75	71	63	89
66	77	84	84	79	77	59	60	81	45	65	84	31	67	82
74	78	86	76	56	59	56	78	84	62	48	79	41	60	41
79	71	72	34	43	59	63	68	65	89	54	59	83	77	77
75	61	73	76	52	68	87	62	72	52	71	79	81	80	58
86	75	83	63	63	80	59	84	64	64	79	79	63	75	63
68	63													

Femmes (effectif: 88)

94	82	63	56	90	88	72	76	94	89	67	88	82	82	61
97	88	68	89	91	56	85	77	68	81	88	45	73	97	78
83	74	91	80	76	80	95	62	98	73	66	90	71	73	66
55	53	94	57	69	62	81	80	90	81	43	76	91	88	87
73	80	87	87	84	64	78	77	64	54	81	88	84	85	74
44	84	97	54	76	87	59	80	72	90	90	70	87		

2) Étude qualitative:

Les données ont été recueillies en vrac. Pour y voir plus clair, le premier travail va consister à les *classer* (les tableurs le font très bien), ce qui permettra de déterminer les âges extrêmes de chaque sous-population *(étendue)*: pour les hommes, 25-89 ans; pour les femmes, 43-98 ans. On pourra également, sans calcul, visualiser la répartition des âges à l'aide de diagrammes en "tige et feuilles":

Hommes	Femmes
2 5	
3 134	
4 11358	**4** 345
5 22346688999999	**5** 34456679
6 00122333333445556678888	**6** 122344667889
7 11122334555556677777889999999	**7** 012233334466667788
8 00112334444466799	**8** 000001111222344455777778888899
	9 0000011144457778

La simple comparaison visuelle montre immédiatement:

1- que l'âge au décès est très variable (de 25 à 98 ans)
2- qu'aucun homme n'a atteint 90 ans
3- qu'aucune femme n'est décédée avant 40 ans
4- que les femmes ont vécu plus longtemps que les hommes
5- que la plus grande mortalité des hommes est celle des septuagénaires, alors que pour les femmes il s'agit des octogénaires

On peut également, à partir de cet exemple, faire des remarques d'ordre général, telles que: l'étendue de la population totale est la réunion des étendues des populations partielles.

[N.B.: la présentation en "tige et feuilles" n'est en quelque sorte qu'un cas particulier de diagramme en bâtons].

3) Étude quantitative:

(1)

Si maintenant on désire préciser (quantifier) cette plus grande longévité des femmes constatée qualitativement, on peut chercher à déterminer, comme le fit Christian Huygens en son temps [Parzysz 1998], *l'âge pour lequel il y a autant de gens qui sont morts avant qu'après,* c'est-à-dire l'âge médian. Par simple comptage, on s'aperçoit alors:

1- que l'âge médian des femmes est de 80 ans

2- qu'il n'y pas un âge médian pour les hommes, puisque le 46ème est mort à 68 ans et le 47ème à 71 ans: il y a en fait tout un intervalle médian:]68; 71[(pratique usuelle: on prend par convention le centre de cet intervalle, soit 69,5 ans)

3- que l'âge médian des hommes est de toute façon inférieur à celui des femmes.

On peut également constater que l'âge médian de la population totale (75 ans) est compris entre celui des hommes (69,5 ans) et celui des femmes (80 ans): *(est-ce toujours le cas?)*

Conclusion: ce qui précède permet de voir que la présentation en "tige et feuilles" (constituée à partir de la série classée), facile à réaliser, se révèle fort utile: sans

presque aucun calcul (à l'exception d'un simple comptage), on a obtenu des renseignements chiffrés sur la population étudiée.

(2)

Si on veut à présent poursuivre l'étude, on peut (les tableurs le font également très bien) déterminer l'*âge moyen* au décès pour les hommes, les femmes et la population totale; on trouve alors:

- hommes: 67,8 ans
- femmes: 77, 3 ans
- série entière: 72,4 ans

On s'aperçoit alors que:
1° l'âge moyen, comme l'âge médian, appartient à l'étendue, ce qui peut être utile pour détecter une éventuelle erreur dans le calcul de la moyenne *(est-ce toujours le cas?)*
2° l'âge moyen n'est pas une valeur observée
3° l'âge moyen n'est pas égal à l'âge médian, et dans les trois cas l'âge moyen est inférieur à l'âge médian *(ce résultat peut-il s'interpréter sur le diagramme en "tige et feuilles"? sur un histogramme?)*
4° l'âge moyen de la population totale est compris entre celui des hommes et celui des femmes *(est-ce toujours les cas?)*, et qu'il est même presque exactement au milieu *(est-ce toujours le cas?).*
5° il est possible d'observer de grands écarts par rapport à la moyenne.

Remarque: Dans ce qui précède, nous avons considéré l'âge comme un caractère statistique discret ne prenant que des valeurs entières. Mais on peut aussi le considérer comme un caractère *continu* (temps écoulé depuis la naissance). Que signifie alors "avoir 56 ans"? Est-ce avoir entre 55 ans 1/2 et 56 ans 1/2, ou est-ce avoir 56 ans révolus (c'est-à-dire entre 56 et 57 ans)? Dans un cas comme dans l'autre, les données recueillies résultent d'un groupement en classes d'amplitude un an, mais le centre de ces classes n'est pas le même:
- dans le premier cas, le centre de la classe "56 ans" est effectivement 56 ans;
- dans le second cas, le centre de la classe "56 ans" est en réalité 56 ans 1/2.

Il en résulte des conséquences immédiates, en particulier:
- l'étendue n'est pas la même: pour la population totale, elle devient maintenant [24,5 ; 98,5[(premier cas) ou [25 ; 99[(second cas), au lieu de [25 ; 98];
- la moyenne (comme la médiane) est inchangée dans le premier cas, mais, dans le second, elle est plus élevée d'une demi-année.

Il en résulte une règle impérative: la définition précise du caractère étudié doit figurer nécessairement dans le texte accompagnant les données statistiques.

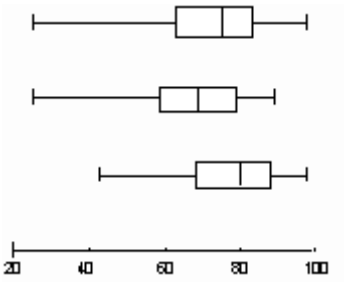
(3)

En allant au-delà du programme actuel de Troisième, on peut poursuivre la dichotomie commencée avec l'âge médian, et chercher *l'âge pour lequel il y a un quart de gens qui sont morts plus jeunes et trois quarts de gens qui sont morts plus vieux* (premier quartile, Q1), ainsi que *l'âge pour lequel trois quarts de gens sont morts plus jeunes et un quart de gens sont morts plus vieux* (troisième quartile, Q3). On obtient alors:

série entière: Q1 = 63 Q3 = 83
hommes: Q1 = 59,5* Q3 = 79
femmes: Q1 = 68,5* Q3 = 88
(* obtenu par interpolation)

On peut alors observer que (comme pour la médiane), le premier quartile de la série entière se situe entre les premiers quartiles des séries partielles, et qu'il en va de même pour le troisième quartile.

Les trois distributions peuvent maintenant être représentées par des *diagrammes de dispersion* (ou boites à moustaches, ou boites à pattes) :



série entière

hommes

femmes

échelle des âges

On est alors à même de retrouver graphiquement les observations faites précédemment à propos de l'étendue, de la médiane et des quartiles, et retrouver l'allure des distributions à partir des diagrammes. Et on peut répondre - au moins qualitativement - à la question posée initialement : pour la population considérée, toutes les études entreprises font apparaitre que les femmes vivent plus longtemps que les hommes (environ 10 ans de différence pour la moyenne comme pour la médiane). Ce qu'on ne peut pas encore dire, c'est si cette différence constatée est significative (même si le bon sens suggère que son ampleur plaide en faveur de la significativité), c'est-à-dire si elle n'est pas simplement liée à la taille relativement faible de la population étudiée: c'est justement là qu'intervient la statistique inférentielle.

En conclusion, ce simple exemple permet de voir que, si les notions spécifiques qui interviennent sont en petit nombre, elles permettent néanmoins de réinvestir d'autres notions mathématiques (classement de décimaux, intervalles réels... sans parler de la proportionnalité à propos des représentations graphiques). De plus, il intervient ici une grande variété de registres de représentation, et on est amené à passer de l'un à l'autre en se posant la question de ce qui est conservé et de ce qui disparait, ou est transformé (problème de la congruence sémantique). Enfin, il s'agit d'un lieu privilégié de travail en équipe, où même la mise en commun des résultats partiels obtenus conduit à se poser des questions. Certains aficionados des probabilités et/ou de la statistique reprocheront sans doute au programme actuel:

- de ne pas aller assez loin, et en particulier de n'introduire ni la notion de quantile, ni celle d'écart-type, et de s'en tenir, pour la dispersion, à la seule étendue (qui est effectivement un indicateur bien fruste).

- de ne pas préparer suffisamment les élèves à l'aléatoire ni à la statistique inférentielle (étude de la stabilisation de la fréquence et des fluctuations d'échantillonnage).

Quoiqu'il en soit, malgré les imperfections qu'on peut lui trouver, ce programme, s'il n'est pas traité à la sauvette, permet au moins d'aborder plusieurs éléments essentiels dans la formation des élèves, les conduisant entre autres à réfléchir sur de vraies questions d'aujourd'hui, à faire preuve d'esprit critique quant à la présentation qu'en donnent les médias, à travailler en équipe, à utiliser conjointement des connaissances issues de diverses disciplines ou de divers domaines des mathématiques, à rechercher une gestion optimale entre les résultats qu'on cherche à obtenir et l'ampleur des moyens qu'on utilisera pour le faire, à synthétiser ces résultats de façon claire et opératoire. Ce qui n'est déjà pas si mal....

Bibliographie

Girard Jean-Claude (1997): *Pourquoi il ne faut pas laisser de côté les chapitres de statistique au collège*, in Repères-IREM n° 23.

Parzysz Bernard (1996): *L'espérance de vie des frères Huygens*, in Bulletin de l'APMEP n° 416.