Sujet 14. Statistiques

1. Le sujet

Exercice

L'observatoire météorologique de Paris-Montsouris relève en permanence depuis 1872 la température extérieure et fournit des moyennes annuelles à partir de ces relevés. Une analyse des températures moyennes annuelles entre 1881 et 1980 montre que ce sont des données gaussiennes de moyenne m = 11,49°C et d'écart-type σ = 0.54°C.

Le tableau ci dessous donne la série des moyennes des températures annuelles en degrés Celsius des années 1981 à 2000.

Année	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
temp moyennes	11,50	12,40	12,30	11,85	11,10	11,25	11,15	12,40	12,95	13,10
Année	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
temp moyennes	11,75	12,30	11,85	13,10	12,85	11,40	12,90	12,40	13,05	12,90

1.1. Déterminer la médiane ainsi que les premier et troisième quartiles de la série des températures durant la période 1981-2000.

1.2. Construire pour cette série le diagramme en boîte. On fera figurer la médiane, les premier et troisième quartile, le minimum et le maximum de la série de températures.

1.3. Déterminer la moyenne de la série des températures annuelles de 1981 à 2000 (on arrondira le résultat au dixième).

2. Déterminer la plage de normalité à 68% de la série des températures moyennes annuelles entre 1881 et 1980.

3. Comparer les températures moyennes observées à Paris dans les vingt dernières années du XXe siècle à celles observées au cours des cent années précédentes.

(*D'après baccalauréat série L septembre 2006, Polynésie*)

Le travail à exposer devant le jury

1. Un professeur propose l'exercice ci-dessus en supprimant les questions 1 et 2. Quelles compétences cherche-t-il selon vous à développer chez ses élèves ?

2. Proposez une correction de la question 3 telle que vous la présenteriez à des élèves de première.

3. Présentez deux ou trois exercices de statistique descriptive à une ou deux variables dont l'un au moins amène à comparer plusieurs séries statistiques.

2. Eléments de correction

L'exercice original est un exercice de baccalauréat (épreuves anticipées, fin de première L). Il est uniquement destiné à évaluer les connaissances et compétences des candidats. Le critère majeur dans la construction de l'énoncé est de proposer aux élèves une succession graduée d'items : savoir déterminer une médiane et des quartiles, construire un diagramme en boîte, calculer une moyenne puis un écart-type, savoir interpréter un résultat. La pertinence de l'enchaînement des questions posées passe au second plan.

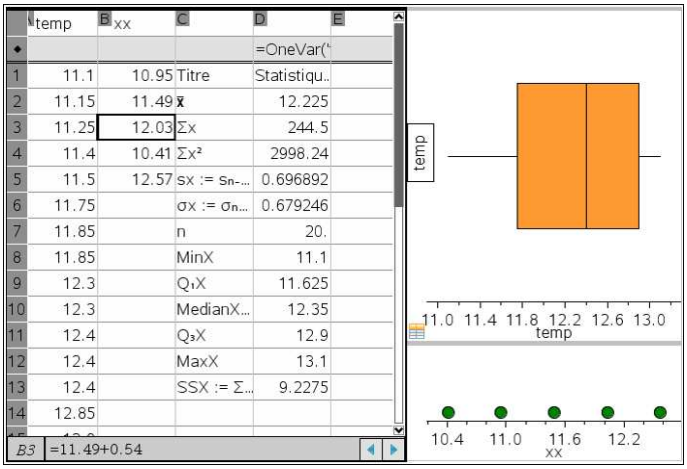
La façon dont est présentée la situation présente de ce fait quelques inconvénients :

- La construction d'un diagramme en boîte représentant la série n'est en aucune façon motivée.
- Il en est de même de la détermination de la plage de normalité.
- La détermination de la moyenne est mal placée dans la question 1 : il s'agit d'une donnée gaussienne, indépendante de la notion de médiane ou de quartiles. Elle aurait dû faire partie de la question 2 (question 2.1 la moyenne et question 2.2 la plage de normalité).

La copie d'écran ci-dessous résume les résultats importants de l'exercice :

La série **temp** est celle des températures moyennes enregistrées de 1981 à 2000. Elle a été triée par ordre croissant.
La série **xx** : $\{m-\sigma, m, m+\sigma\}$ permet de préciser les plages de normalité à 68 % de la série des températures moyennes annuelles entre 1881 et 1980. On a fait figurer également dans cette liste $m-2\sigma$ et $m+2\sigma$, pour préciser le cas échéant la plage de normalité à 95 %.

On constate que sur les 20 années observées à partir de 1981, 12 valeurs des températures moyennes sont à l'extérieur de la plage de normalité à 68 % du siècle précédent, toutes supérieures à la normale.



1. Les compétences que le professeur cherche à développer en supprimant les questions 1 et 2 :

Ce sont celles figurant dans l'objectif général des programmes (et non des compétences particulières liées aux statistiques qui, de toute façon, figurent à un moment donné dans la résolution) :

- Mettre en œuvre une recherche de façon autonome.
- Mener des raisonnements.
- Avoir une attitude critique vis-à-vis des résultats obtenus.

La décision de l'enseignant paraît surtout motivée par un choix pédagogique en vigueur depuis le collège : « *La compréhension et l'appropriation des connaissances mathématiques reposent sur l'activité de chaque élève qui doit donc être privilégiée. Pour cela, et lorsque c'est possible, sont choisies des situations créant un problème dont la solution fait intervenir des « outils », c'est-à-dire des techniques ou des notions déjà acquises, afin d'aboutir à la découverte ou à l'assimilation de notions nouvelles ... Ainsi, les connaissances peuvent prendre du sens pour l'élève à partir des questions qu'il se pose et des problèmes qu'il résout* ».

L'enseignant veut manifestement travailler sur le thème : « *Utiliser de façon appropriée les deux couples usuels qui permettent de résumer une série statistique : (moyenne, écart-type) et (médiane, écart interquartile)* ». Plus particulièrement, il veut expliciter comment ces couples apportent des moyens de comparaison de deux séries.

Si l'enseignant pose le problème original, les notions mathématiques enseignées précèdent leur exemple d'application, l'enseignant se positionne dans un dispositif transmissif.

En supprimant les questions 1 et 2, l'enseignant cherche à provoquer un questionnement. Il s'attend à la *conjecture* : « Il y a un réchauffement », et il s'agira alors d'apporter des outils mathématiques susceptibles de confirmer ou d'infirmer la conjecture. Les couples usuels résumant une série statistiques fourniront autant d'arguments alimentant le débat. Quant à elle, la notion de « plage de normalité » prend son sens à partir du problème posé. Elle n'est plus enseignée *a priori*.
On ne peut qu'approuver cette décision dès lors que cet exercice est exploité en cours d'apprentissage.

2. Le fait de supprimer les questions 1 et 2 implique une correction en deux temps. Il ne faut pas que les élèves s'arrêtent au fait que « l'on voit qu'il y a beaucoup de températures supérieures à 12°*C*, la température a augmenté ».

Premier temps : Que sait-on des deux séries ? Quels sont les moyens de comparaison dont on dispose ?

Série 1881/1980. On en connaît la moyenne et l'écart type et l'on sait qu'il s'agit de « données gaussiennes ». Il faut donc expliquer le terme. Cela signifie que la série est à peu près symétrique autour de sa moyenne, et que environ 68 % (respectivement 95 %) des données se trouvent dans l'intervalle $[m-\sigma,\ m+\sigma]$ (respectivement $[m-2\sigma,\ m+2\sigma]$). Cet intervalle est la plage de normalité pour le niveau de confiance 68 % (respectivement 95 %).

Série 1981/2000. Nous en connaissons les données brutes. Il nous faut donc organiser et résumer les informations. La correction se bornera à faire l'inventaire des résumés pertinents. Il est probable que, puisque la série du siècle précédent est représentée par sa moyenne et son écart type, des élèves vont aussi calculer la moyenne et l'écart type de la série 1981/2000. Il conviendra d'inciter à trouver un autre jeu de représentants (médiane et quartiles), de façon à diversifier l'information dont on dispose.

À partir de là, les élèves sont invités à produire un argument justifiant leur comparaison.

Deuxième temps : la comparaison proprement dite
Elle s'appuie sur le calcul des paramètres résumant une série statistique. On peut attendre plusieurs arguments pertinents en faveur d'un réchauffement, dépendant des paramètres représentatifs des séries que l'on veut utiliser, par exemple :

1. La température moyenne sur les 20 dernières années est nettement plus élevée que la température moyenne du siècle précédent (12,25°*C* contre 11,49°*C*, soit 0,76°*C* de plus)
2. La moitié des températures des 20 dernières années sont supérieures à $Q_2 = 12{,}35°C$, et elles sont toutes en dehors et au dessus de la plage de normalité à 68 % du siècle précédent.
3. Le quart des températures des 20 dernières années sont supérieures à $Q_3 = 12{,}9°C$, et elles sont toutes en dehors et au dessus de la plage de normalité à 95 % du siècle précédent.
4. A peu près la moitié des températures du siècle précédent étaient inférieures à 11,5°*C* (médiane = moyenne pour une série gaussienne), alors qu'il y en a eu moins du quart pendant les 20 dernières années : $Q_1 = 11{,}625°C$.

Le premier argument porte sur le fait que la moyenne des températures est sensiblement plus élevée.
Les trois autres arguments portent sur le fait que l'on observe un décalage des températures vers le « plus chaud » (arguments 2 et 3) et vers le « moins froid » pour les plus froides (argument 4).

En résumé, on pourra insister sur le fait que les températures mesurées sur les dernières vingt années semblent *globalement* supérieures aux températures du siècle précédent. Il y a eu non seulement augmentation de la moyenne des températures mais aussi décalage général des températures vers le « plus chaud » (l'augmentation de la moyenne n'est pas due à des années isolées exceptionnellement chaudes).

3. Voir REDCM pages 183 à 187.

3. Pour aller plus loin

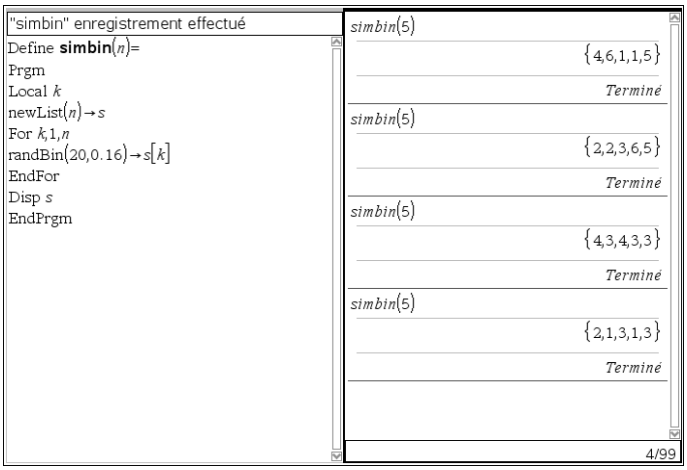
On attend des élèves de première L une comparaison des deux séries de températures basée sur l'observation des paramètres représentatifs.

On peut cependant se poser la question : qu'est-ce qui nous autorise à dire (par exemple) que le nombre de relevés situés au-delà de la plage de normalité à 68 % est « anormalement » élevé ? Le chapitre « **Echantillonnage »** du programme de TES offre une ouverture sur ce sujet :

« *Exploiter l'intervalle de fluctuation à un seuil donné, déterminé à l'aide de la loi binomiale, pour rejeter ou non une hypothèse sur une proportion. L'objectif est d'amener les élèves à expérimenter la notion de « différence significative » par rapport à une valeur attendue* ».

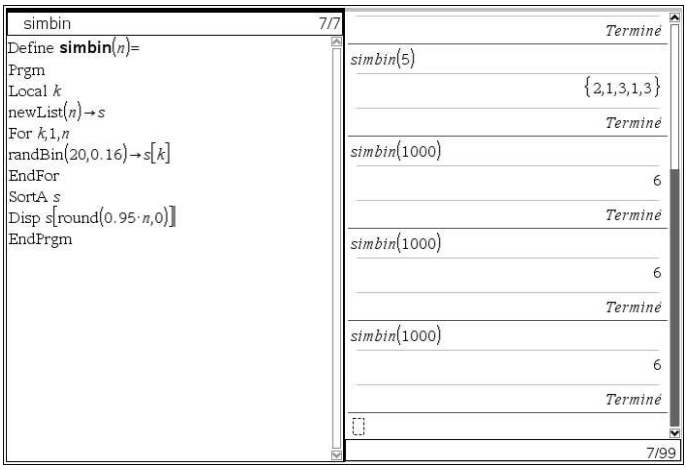
Compte tenu de la plage de normalité à 68 % obtenue, s'il n'y a pas eu de changement significatif du climat, on peut attendre une proportion de 16 % de températures moyennes qui sont supérieures ou égales au seuil $m+\sigma=12{,}03$. Dans cette hypothèse, l'évènement « la température moyenne observée est au dessus du seuil » a une probabilité égale à 0,16. Si on effectue 20 observations, le nombre de températures au dessus du seuil devrait suivre la loi binomiale $B(20,\ 0.16)$.

L'instruction randBin $(20,\ 0.16)$ simule un relevé du nombre de températures au dessus du seuil observées au cours d'une période de 20 ans. Le programme **simbin**, d'argument *n*, construit une série nommée **s** de *n* relevés.



Quelques unes de ces séries *s* obtenues en exécutant le programme sont affichées pour une petite valeur de *n*.

On a corrigé le programme en effectuant un tri croissant de la série **s** et en faisant afficher le 95ème centile, puis on l'exécute pour une valeur de *n* nettement plus grande (*n* = 1000 par exemple).



La série *s* obtenue sert de série de référence. Plusieurs exécutions du programme **simbin** amènent toutes à dire que dans 95 % des cas, le nombre de températures situés au dessus du seuil que l'on devrait relever en 20 ans d'observations devrait être inférieur ou égal à 6.

On peut affirmer au seuil de confiance 95 %, qu'obtenir 12 relevés supérieurs à 12,03°*C* au cours des années 1981/2000 est un phénomène anormal, qui n'est pas dû à la fluctuation d'échantillonnage.