

Master d'infochimie M1S1

Examen d'analyse statistique des données*

Nom:

Prénom:

Janvier 2009

Documents autorisés. Durée : 2h. La copie de l'étudiant sera constituée d'un support papier traditionnel et d'un ou plusieurs fichiers de calcul pour Excel et/ou Minitab. Les tests statistiques seront effectués en fixant le niveau de confiance à 95% sauf si la question en exige un autre. Le candidat pourra néanmoins en changer si cela lui semble utile à ses conclusions auquel cas celui-ci devra être clairement précisé. Toute réponse doit être motivée ou sera considérée comme nulle.

Exercice 1 : Intervalle de confiance

Le document `BinoDist.xls` contient une suite de 300 nombres obtenus à partir d'une loi binomiale. L'objectif de l'exercice est de comprendre comment le modèle probabiliste sous-jacent affecte un test statistique [2].

L'exercice suit le principe suivant : les données sont tout d'abord examinées en utilisant les méthodes paramétriques standard développées dans le cours. Par la suite la véritable distribution et ses paramètres sont donnés. La moyenne de l'échantillon est alors réexaminée.

Question 1

Expliquez succinctement (deux lignes) ce qu'est une loi binomiale.

Question 2

Estimez la moyenne, l'écart-type, le coefficient d'asymétrie d'aplatissement et la Kurtosis à partir de cet échantillon.

Tracez un histogramme de cette distribution.

Pensez vous qu'une distribution normale soit un bon choix pour approximer cette distribution ?

Question 3

Calculez un intervalle de confiance sur votre estimation de la moyenne à 95% et à 99% pour cet échantillon. Vous supposerez que la moyenne se distribue selon une loi de Student ou une loi normale (au choix).

La loi Binomiale possède deux paramètres, le nombre d'essais n et la probabilité de succès p . Le paramètre p se déduit de la moyenne m et de l'écart-type σ de la distribution à l'aide de la formule :

$$p = 1 - \frac{\sigma^2}{m} \quad (1)$$

Le paramètre n de la distribution se déduit alors par l'équation :

$$n = \frac{m}{1 - \frac{\sigma^2}{m}} \quad (2)$$

* Enseignants: G. Marcou, P. Jost, Université de Strasbourg, Faculté de Chimie, 4, rue Blaise Pascal, 67000 Strasbourg

Question 4

Calculez les paramètres p et n de la distribution de probabilité binomiale dont est issu l'échantillon.

En réalité les paramètres de la distribution sont $p = 0.3$ et $n = 10$. La valeur moyenne exacte de la distribution est donc $\mu = np = 3$.

Question 5

La valeur exacte est-elle dans l'intervalle de confiance à 95% de la moyenne de l'échantillon¹ ?

La valeur moyenne m de l'échantillon est une variable aléatoire, somme de N variables aléatoires suivant une loi binomiale de paramètres p et n , où N est la population de l'échantillon. Par conséquent, cette moyenne est reliée à une variable aléatoire X obéissant également à une loi binomiale de paramètres p et $n \times N$ selon la relation :

$$m = \frac{X}{N} \quad (3)$$

Question 6

Calculer l'intervalle de confiance « exact » à 95% et à 99% sur la moyenne de l'échantillon. Pour ce faire, il faut calculer la probabilité cumulative pour chaque valeur possible de la moyenne d'un échantillon de même taille que celui qui est présenté. Il vous est recommandé de suivre les étapes suivante :

1. construire une colonne contenant tous les nombres de 0 à 10 par pas de 1/300 -les valeurs possibles de la moyenne ;
2. construire une colonne contenant tous les nombres entiers de 0 à 3000 -les valeurs possibles de la somme des individus de l'échantillon, la variable aléatoire X ;
3. utiliser la seconde colonne comme argument pour calculer la probabilité cumulative selon une loi binomiale de paramètres $p = 0.3$ et $n = 3000$;
4. en déduire l'intervalle de confiance « exact » à 95% et à 99% sur la moyenne de l'échantillon.

Question 7

Donnez brièvement vos conclusions (5 lignes maximum).

Exercice 2 : Analyse en composantes principales et régression

Un récent « défis » de modélisation QSAR a proposé aux chercheurs du monde entier de modéliser la solubilité aqueuse d'une centaine de composés [1]. Le fichier `s100.xls` contient le nom du composé, le logarithme de la solubilité aqueuse et 61 descripteurs calculés sur les structures des molécules. L'objectif est de construire un modèle prédictif et explicatif de la propriété.

Question 8

Expliquez brièvement ce qu'est l'analyse en composantes principales (maximum 3 lignes).

¹ calculé à la question 3

Question 9

Calculez les 40 premières composantes principales du jeu de données. Stockez les projections (valeurs) des données sur ces composantes. Combien de composantes principales sont nécessaires pour décrire 99% de la variance des facteurs du jeu de données?

Question 10

Réalisez une régression multilinéaire pour modéliser la solubilité aqueuse en utilisant un sous-ensemble des 61 facteurs d'origine. Utilisez une méthode de sélection de variable pas à pas ascendante et descendante; faites attention au traitement de l'ordonnée à l'origine. Commentez la signification de votre modèle.

Question 11

Utilisez la même approche que précédemment, mais cette fois-ci, sélectionnez vos facteurs parmi les 40 composantes principales calculées précédemment. Commentez la signification du modèle et justifiez du traitement de l'ordonnée à l'origine.

Question 12

Comparez le nombre de facteurs utilisés pour construire un modèle basé sur les composantes principales par rapport aux modèles construits sur les facteurs d'origine. Concluez en quelques lignes (5 lignes maximum).

Barème indicatif

- Question 1 : 1 points
- Question 2 : 2 points
- Question 3 : 1 points
- Question 4 : 2 points
- Question 5 : 1 points
- Question 6 : 2 points
- Question 7 : 1 points
- Question 8 : 1 points
- Question 9 : 3 points
- Question 10 : 3 points
- Question 11 : 2 points
- Question 12 : 1 points

Références

- [1] A. Llinàs, R. C. Glen and J. M. Goodman, *Can You Predict Solubilities of Thirty-Two Molecules Using a Database of One Hundred Reliable Measurements?*; J. Chem. Inf. Modeling, **2008**, 48, 1289-1303.
- [2] P. H. Kvam and B. Vidakovic, *Nonparametric statistics with applications to science and engineering*; Wiley Series in Probability and Statistics, **2007**.