Université de Lorraine - Site de Metz - UFR MIM **Année 2012/2013**
Master de Mathématiques, année M1
Module *Introduction à l'analyse des données*

Exercices - Feuille 6

Tests statistiques

1- Traduction d'une hypothèse linéaire
On considère une distribution $N_2(\mu, \Sigma)$. Formuler l'hypothèse $\mu_1 = \mu_2 = \mu_3$ sous la forme $A\mu = a$, avec $A \in \mathbb{M}_3(\mathbb{R})$ et $a \in \mathbb{R}^3$.

2- Simulation de la loi multinormale
On s'intéresse à la simulation d'une loi $N_p(\mu, \Sigma)$.
1) Soit Y une variable aléatoire de loi $N_q(0, I_q)$.
Rappeler la transformation de Mahalanobis de matrice $A \in \mathbb{M}_{p,q}(\mathbb{R})$ et de vecteur $\mu \in \mathbb{R}^p$.
2) Rappeler la définition de la matrice $\Sigma^{1/2}$.
3) Vérifier que la variable $X = \mu + \Sigma^{1/2}Y$ a pour loi $N_p(\mu, \Sigma)$.
4) On considère le script matlab suivant:

```
% DISTRIBUTION NORMALE VECTORIELLE
clear all;
mu = [1 2];
Sigma = [1 .5; .5 2];
Sigmas=Sigma^0.5;
z = repmat(mu,1000,1) + randn(1000,2)*Sigmas;
m=mean(z);
c=cov(z);
```

Que fait ce programme ?
5) Déduire de ce qui précède un script qui simule un nombre N de fois une loi $N_p(\mu, \Sigma)$.
6) On considère à présent la decomposition de Choleski de Σ sous la forme $\Sigma = R^T R$ avec R triangulaire supérieure. Vérifier que la variable aléatoire $Z = \mu + R^T X$ suit également une loi $N_p(\mu.\Sigma)$.
7) Vérifier numériquement sur quelques exemples (on choisira p, μ, Σ) que la moyenne empirique m_n et que la matrice covariance empirique C_n satisfont bien $m_n \to \mu$ et $C_n \to \Sigma$ lorsque $n \to +\infty$.

3- Test sur la loi normale (1)
1) Simuler un échantillon multinormal avec $\mu = (1, 2)^T$ et $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix}$.
2) Tester l'hypothèse (H_0) : $2\mu_1 - \mu_2 = 0$, en considérant d'abord le cas Σ connu et ensuite Σ inconnu.

4- Test sur la loi normale (2)
On considère un échantillon de taille $n = 5$ provenant d'une distribution

$$X \sim N_2\left(\mu, \begin{pmatrix} 3 & \rho \\ \rho & 1 \end{pmatrix}\right) \tag{1}$$

où ρ est un paramètre connu. On suppose que la moyenne empirique de l'échantillon est $\bar{x}^T = (1, 0)$. Pour quelle valeur de ρ a-t-on que l'hypothèse $(H_0) : \mu = (0, 0)$ est à rejeter en faveur de l'hypothèse $(H_1) : \mu^T \neq (0, 0)$ (au risque 5%) ?

5- Test multivarié, Σ connu
On considère une variable aléatoire $X \sim N_2(\mu, \Sigma)$, où Σ est supposé connu avec

$$\Sigma = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}. \tag{2}$$

On suppose que l'on dispose d'un échantillon de taille $n = 6$ dont la moyenne est $m = (1, 1/2)^T$. Effectuer les tests statistiques suivants (avec un risque $\alpha = 0.05$.)
a)

$$(H_0) : \ \mu = (2, 2/3)^T, \quad (H_1) : \ \mu \neq (2, 2/3)^T \tag{3}$$

b)

$$(H_0) : \ \mu_1 + \mu_2 = 7/3, \quad (H_1) : \ \mu_1 + \mu_2 \neq 7/3 \tag{4}$$

c)

$$(H_0) : \ \mu_1 - \mu_2 = 1/2, \quad (H_1) : \ \mu_1 - \mu_2 \neq 1/2 \tag{5}$$

d)

$$(H_0) : \ \mu_1 = 2, \quad (H_1) : \ \mu_1 \neq 2 \tag{6}$$

Dans chacun des cas, représenter graphiquement la région de rejet.

6- Test multivarié, Σ inconnu
Répéter les tests précédents dans le cas où la matrice de covariance Σ est inconnue et où la matrice de covariance empirique (avec biais) C_b est donné par

$$C_b = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}. \tag{7}$$

Comparer les résultats.

7- Etude sur les "pull-overs bleus"

On souhaite étudier différentes stratégies marketing pour la vente d'un certain type de pull-overs. Les données dont on dispose sont les suivantes. Il s'agit de 10 mesures de 4 caractéritiques. Chaque mesure correspond à une séquence de vente particulière (un magasin différent ou une période de vente différente). Trois méthodes de marketing distinctes ont été utilisées et on souhaite interpréter le résultat des ventes en fonction de ces données.

séquence	**prix**	**ventes**	**publicité**	**temps (en h.)**
1	230	125	200	109
2	181	99	55	107
3	165	97	105	98
4	150	115	85	71
5	97	120	0	82
6	192	100	150	103
7	181	80	85	111
8	189	90	120	93
9	172	95	110	86
10	170	125	130	78

1) Lire les données du tableau `pullover.dat` dans 4 tableaux `vente`, `prix`, `publicite`, `heures`. Ces données correspondent respectivement au nombre de pull-overs vendus, au prix de vente unitaire (en), au coût en publicité et au nombre d'heures de travail effectuées par les vendeurs.
2) Calculer la moyenne et l'écart-type de chaque caractéristique X_1, X_2, X_3, X_4.
3) Calculer la matrice de covariance empirique correspondant aux données.
4) Le coût salarial horaire est de 10 €. Quelle est en fonction de X la variable Y qui donne la dépense totale (en €) ?
5) On pense qu'il y a une dépendance du nombre de ventes en fonction du prix. Ceci correspond à postuler une hypothèse de type moindres carrés de la forme

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{8}$$

Combien y-a-t-il de paramètres à déterminer dans le modèle (8) ?
6) Calculer les estimations $\hat{\alpha}$ et $\hat{\beta}$, ainsi que le coefficient de détermination r^2 défini par

$$r^2 = \frac{\sum_{i=1}^n (\bar{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \tag{9}$$

Quelle est la conclusion pratique ?
7) Combien de pull-overs peut-on espérer vendre avec un prix unitaire de 105 €?
8) On souhaite analyser à présent l'impact de trois stratégies marketing différentes qui correspondent aux trois types d'investissement suivants:

1. Publicité dans les journaux locaux.
2. Présence d'assistants de vente.
3. Mise en valeur des produits en vitrine.

Le tableau donnant le nombre d'articles vendus au cours des 10 séquences de ventes est le suivant (`pullover2.dat`).

magasin	**strat. 1**	**strat. 2**	**strat. 3**
1	9	10	18
2	11	15	14
3	10	11	17
4	12	15	9
5	7	15	14
6	11	13	17
7	12	7	16
8	10	15	14
9	11	13	17
10	13	10	15

L'analyse de variances de ces données consiste à observer le rapport entre écart à la moyenne sans tenir compte des p catégories (ici $p = 3$) et ce même écart en tenant compte des p catégories. De façon précise, il s'agit d'un F test de la forme ($m = 10$):

$$F = \frac{(SCTR)/(p-1)}{SCE/(n-p)} \tag{10}$$

avec

$$SCTR = m \sum_{l=1}^{p} \sum_{k=1}^{m} (\bar{y}_l - \bar{y})^2 \tag{11}$$

et

$$SCE = \sum_{l=1}^{p} \sum_{k=1}^{m} (y_{kl} - \bar{y}_l)^2 \tag{12}$$

On peut montrer que la distribution de F suit une loi de Fisher $F_{p-1,n-p}$. Effectuer le test statistique proposé avec $\alpha = 0.05$ et dire si il y a une différence entre les trois stratégies.
9) On reprend une analyse moindres carrés des données `pullover.dat`. Effectuer une analyse moindres carrés multivariée exprimant les ventes X_1 en fonction des trois caractères X_2, X_3, X_4,

$$X_1 = \alpha + \beta_1 X_2 + \beta_2 X_3 + \beta_3 X_4 + \epsilon. \tag{13}$$

On donnera les estimations α, β_1, β_2, β_3. Quel est le coefficient r^2 de détermination ?
10) On suppose que (X_1, X_2, X_3, X_4) suit une loi multinormale $N_4(\mu, \Sigma)$ avec

$$\mu = \begin{pmatrix} 172.7 \\ 104.6 \\ 104.0 \\ 93.8 \end{pmatrix} \quad et \quad \Sigma = \begin{pmatrix} 1037.21 & & & \\ -80.02 & 219.84 & & \\ 1430.70 & 92.10 & 2624.00 & \\ 271.44 & -91.58 & 210.30 & 177.36 \end{pmatrix} \tag{14}$$

Quelle est la densité de la distribution de (X_1, X_2, X_3, X_4) ? Quelle est la loi de X_1 sachant X_2, X_3, X_4?
11) On revient au modèle moindres carrés global (13). On observe que $\beta_1 \simeq -\frac{1}{2}\beta_2$. Quel test suggère cette identité ?