GEOMETRIE EUCLIDIENNE ET STATISTIQUE

Position du problème

On considère une variable aléatoire X définie sur une population. On suppose que X est de loi normale, $N(\mu,\sigma)$.

On prélève dans cette population un échantillon aléatoire et simple de taille n. Il en résulte n variables aléatoires, $X_1, X_2, ...X_n$ indépendantes et de même loi que X.

On note : $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ et $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

Nous allons démontrer que la variable aléatoire $\frac{nS^2}{\sigma^2}$ suit la loi de χ^2 à (n – 1) degrés de liberté.

Rappel

Par définition une variable aléatoire suit la loi de χ^2 à ν degrés de liberté si elle est la somme des carrés de ν variables aléatoires normales, centrées, réduites et indépendantes.

Principe

On va travailler dans l'espace euclidien $\mathbb{R}^n$ muni de son produit scalaire usuel.

On considère le vecteur aléatoire $V = \begin{pmatrix} \frac{X_1 - \overline{X}}{\sigma} \\ \frac{X_2 - \overline{X}}{\sigma} \\ . \\ . \\ \frac{X_n - \overline{X}}{\sigma} \end{pmatrix}$ et le vecteur $I = \begin{pmatrix} 1 \\ 1 \\ . \\ . \\ 1 \end{pmatrix}$

$$V.I = \frac{1}{\sigma}\left((X_1 - \overline{X}) + (X_2 - \overline{X}) + ... + (X_n - \overline{X})\right) = \frac{1}{\sigma}\left(X_1 + X_2 + ... + X_n - n\overline{X}\right) = 0$$

V est donc orthogonal à I. Il en résulte que V appartient à l'hyperplan (sous-espace vectoriel de dimension n-1) de $\mathbb{R}^n$ orthogonal à I.

D'autre part $\|V\|^2 = \frac{1}{\sigma^2}\left((X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + ... + (X_n - \overline{X})^2\right) = \frac{nS^2}{\sigma^2}$

La démonstration consiste à utiliser une base orthonormale, $(e_i)_{1\le i\le n}$, de $\mathbb{R}^n$ telle que $e_1 = \frac{1}{\sqrt{n}} I$. On sait qu'une telle base existe et, dans une telle base, V a pour coordonnées $(0, X'_2, X'_3, ..., X'_n)$ et $\|V\|^2 = \frac{nS^2}{\sigma^2} = \sum_{i=2}^{n} X'^2_i$.

Il restera à démontrer que les $X'_i \ (2 \le i \le n)$ sont de loi N(0, 1) et indépendantes.

Quelques rappels d'algèbre linéaire

Pour obtenir une matrice de changement de base on place en <u>colonnes</u> les coordonnées, dans la première base, des vecteurs de la deuxième base. Cette matrice permet de calculer les coordonnées d'un vecteur dans la première base en fonction de celles de ce vecteur dans la deuxième base.

Ce qui nous intéresse ici est la réciproque : calculer les nouvelles coordonnées en fonction des anciennes. Nous avons donc besoin de la matrice inverse de la matrice de changement de base.

Lors du passage d'une base orthonormale à une autre base orthonormale, la matrice de passage est une matrice orthogonale, son inverse est alors égale à sa transposée. Nous utiliserons donc une matrice dont les <u>lignes</u> seront les coordonnées des vecteurs de la nouvelle base.

Exemple de matrice convenant ici :

Considérons la matrice suivante :

$$\begin{pmatrix} 1 & 1 & 1 & 1 & ... & ... & 1 & 1 \\ 1 & -1 & 0 & 0 & ... & ... & 0 & 0 \\ 1 & 1 & -2 & 0 & ... & ... & 0 & 0 \\ 1 & 1 & 1 & -3 & 0 & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... & ... & ... & ... \\ 1 & 1 & 1 & ... & ... & 1 & -n+2 & 0 \\ 1 & 1 & 1 & ... & ... & 1 & 1 & -n+1 \end{pmatrix}$$

Ses vecteurs-lignes sont 2 à 2 orthogonaux.. pour obtenir une matrice répondant à notre problème il suffit de multiplier chacun d'eux par l'inverse de sa norme.

Démonstration

- Soit une base orthonormale, $(e_i)_{1\le i\le n}$, de $\mathbb{R}^n$ telle que $e_1 = \frac{1}{\sqrt{n}} I$

Pour $1 \le i \le n$ et $1 \le j \le n$ on notera a_{ij} la $j^{ième}$ coordonnée de e_i.

Pour tout i, $1 \le i \le n$, on a : $\|e_i\| = 1$ d'où $\sum_{j=1}^{n} a_{ij}^2 = 1.$ (1)

D'autre part, pour $i \neq k$, $e_i.e_k = 0$ d'où : $\sum_{j=1}^{n} a_{ij}.a_{kj} = 0.$ (2)

En particulier, pour tout i, $2 \le i \le n$, $e_i.e_1 = 0$ d'où : $\sum_{j=1}^{n} a_{ij} = 0.$ (3)

- Avec les notations définies précédemment on obtient :

$$\begin{pmatrix} X_1' \\ X_2' \\ . \\ . \\ X_n' \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \dots & \dots & \frac{1}{\sqrt{n}} \\ a_{21} & a_{22} & \dots & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} \frac{X_1 - \overline{X}}{\sigma} \\ \frac{X_2 - \overline{X}}{\sigma} \\ . \\ . \\ \frac{X_n - \overline{X}}{\sigma} \end{pmatrix}$$

$X_1' = \frac{1}{\sqrt{n}} V.I = 0$, d'où, puisque la nouvelle base est aussi orthonormale, l'égalité :

$\|V\|^2 = \sum_{i=2}^{n} X_i'^2$ c'est-à-dire $\frac{nS^2}{\sigma^2} = \sum_{i=2}^{n} X_i'^2$.

- D'autre part pour $2 \le i \le n$ on a :

$X_i' = \frac{1}{\sigma}\sum_{j=1}^{n} a_{ij}(X_j - \overline{X}) = \frac{1}{\sigma}\sum_{j=1}^{n} a_{ij}X_j - \frac{\overline{X}}{\sigma}\sum_{j=1}^{n} a_{ij} = \frac{1}{\sigma}\sum_{j=1}^{n} a_{ij}X_j$ d'après (3).

X_i' est donc de loi normale comme combinaison linéaire de variables aléatoires normales indépendantes. De plus :

$E(X_i') = \frac{1}{\sigma}\sum_{j=1}^{n} a_{ij}E(X_j) = \frac{\mu}{\sigma}\sum_{j=1}^{n} a_{ij} = 0$ d'après (3).

et, les X_i étant indépendantes,

$V(X_i') = \frac{1}{\sigma^2}\sum_{j=1}^{n} a_{ij}^2 V(X_j) = \frac{\sigma^2}{\sigma^2}\sum_{j=1}^{n} a_{ij}^2 = 1$ d'après (1).

Il en résulte que, pour tout i, $2 \le i \le n$, X_i' est de loi normale centrée et réduite.

- Il reste à montrer que les $X_i^{'}$, pour $2 \le i \le n$, sont indépendantes.

Toute combinaison linéaire des $X_i^{'}$ est une combinaison linéaire des X_i et est donc de loi normale. La définition donnée en annexe permet donc de dire que le vecteur aléatoire $\left(X_2^{'}, X_3^{'}, ...X_n^{'}\right)$ est gaussien. Pour que les $X_i^{'}$ $\left(2 \le i \le n\right)$ soient indépendantes il suffit, d'après le théorème donné en annexe, qu'elles soient 2 à 2 non corrélées.

Calculons donc, pour h différent de k (h et k compris entre 2 et n), la covariance de $X_h^{'}$ et $X_k^{'}$:

$$\mathrm{Cov}(X_h^{'}, X_k^{'}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_{hi}.a_{kj}\mathrm{Cov}(X_i, X_j)$$

Or, pour $i \neq j$, $\mathrm{Cov}(X_i, X_j) = 0$ puisque les X_i sont indépendantes et, pour tout i compris entre 1 et n, $\mathrm{Cov}(X_i, X_i) = V(X_i) = \sigma^2$ .Il en résulte :

$$\mathrm{Cov}(X_h^{'}, X_k^{'}) = \frac{1}{\sigma^2} \sum_{j=1}^{n} a_{hj}.a_{kj}.\sigma^2 = \sum_{j=1}^{n} a_{hj}a_{kj} = 0 \text{ d'après (2).}$$

Les $X_i^{'}$ $\left(2 \le i \le n\right)$ sont donc indépendantes.

Conclusion : La variable aléatoire $\frac{nS^2}{\sigma^2}$ est la somme des carrés de n - 1 variables aléatoires normales centrées réduites et indépendantes, elle est donc de loi de χ^2 à n - 1 degrés de liberté.

***Prolongements possibles* :**

- On peut également montrer que $\frac{nS^2}{\sigma^2}$ est indépendante de $\overline{X}$:

$$Cov(X_i^{'}, \overline{X}) = Cov(\frac{1}{\sigma}\sum_{j=1}^{n} a_{ij}X_j, \frac{1}{n}\sum_{k=1}^{n} X_k) = \frac{1}{\sigma n}\sum_{j=1}^{n}\sum_{k=1}^{n} a_{ij}Cov(X_j, X_k)$$

$$Cov(X_i^{'}, \overline{X}) = \frac{1}{\sigma n}\sum_{j=1}^{n} a_{ij}Cov(X_j, X_j) = \frac{\sigma^2}{\sigma n}\sum_{j=1}^{n} a_{ij} = 0 \text{ d'après (3).}$$

Il en résulte que, pour tout i compris entre 2 et n, $X_i^{'}$ et $\overline{X}$ sont non corrélées et donc indépendantes d'après le théorème cité précédemment. On en déduit que $\frac{nS^2}{\sigma^2}$ qui est la somme des carrés de ces variables est indépendante de $\overline{X}$.

- En analyse de variance à un facteur :

On démontre de la même manière que, lorsque les moyennes des p populations sont égales, la variable aléatoire $\frac{SCE_F}{\sigma^2}$ suit la loi de χ^2 à p - 1 ddl (le même type de matrice est utilisable avec n = p lorsque les échantillons sont de même taille sinon on doit avoir $a_{1j} = \sqrt{\frac{n_j}{n}}$, les n_j étant les effectifs des échantillons et n l'effectif total).

- La matrice donnée en exemple (ou toute autre construite de la même manière)

est utilisée pour les comparaisons multiples de moyennes par une méthode appelée méthode des contrastes lorsque les effectifs des échantillons sont égaux.

Mais tout cela sera pour une autre fois si certains en manifestent le souhait.

Annexe

Définition : Une variable aléatoire définie sur un espace de probabilité et à valeurs dans $\mathbb{R}^n$ est dite normale ou gaussienne si toute combinaison linéaire de ses coordonnées est de loi normale. (remarque : une variable aléatoire constante est considérée comme étant de loi normale d'écart type 0 ou dégénérée)

Théorème : Pour que la suite des coordonnées d'un vecteur normal soit indépendante, il faut et il suffit que la covariance de ce vecteur (matrice des covariances de ses coordonnées) soit une matrice diagonale.

- = - = - = - = - = - = - = -