REVUE DE STATISTIQUE APPLIQUÉE

A. GHOUILA HOURI

J. MOTHES

Estimation de la moyenne et de la variance d'une population normale Dans l'hypothèse de prélèvements exhaustifs

Revue de statistique appliquée, tome 9, nº 3 (1961), p. 27-31

<http://www.numdam.org/item?id=RSA_1961__9_3_27_0>

ESTIMATION DE LA MOYENNE ET DE LA VARIANCE D'UNE POPULATION NORMALE DANS L'HYPOTHÈSE DE PRÉLÈVEMENTS EXHAUSTIFS

A. GHOUILA HOURI et J. MOTHES

I - INTRODUCTION

L'estimation sur échantillon d'une proportion de déchets et l'estimation également sur échantillon de la moyenne et de la variance d'une population normale sont deux problèmes classiques relevant de la statistique mathématique élémentaire.

En ce qui concerne les proportions de déchets, 2 solutions sont généralement fournies, l'une concernant des prélèvements exhaustifs (on se réfère alors à la loi hyper-géométrique) l'autre concernant des prélèvements non exhaustifs (on se réfère alors à la loi binomiale ou à la loi de Poisson).

En ce qui concerne d'autre part les estimations de la moyenne et de la variance d'une population normale, les solutions classiques concernent toujours la moyenne et la variance de la population de référence. En d'autres termes, si nous désignons par m_d et σ_d^2 la moyenne et la variance de la population mère (finie) soumise à examen, en admettant que la distribution de cette population se raccorde à une population de référence normale de moyenne m et de variance σ^2, les techniques usuelles conduisent, à partir des résultats échantillonnés dans la population (m_d, σ_d^2), à l'estimation de m et de σ^2. En pratique, les techniques classiques sont amplement suffisantes, les écarts entre m_d et m d'une part, entre σ_d^2 et σ^2 d'autre part étant négligeables. Toutefois, il est à titre d'exercice intéressant de se proposer l'estimation sur échantillon de m_d et de σ_d^2, autrement dit de définir les tests tenant compte de l'exhaustivité des prélèvements. C'est ce que nous ferons au cours des pages suivantes.

II - ESTIMATION DE m_d

Ayant prélevé un échantillon de n éléments dans la population, échantillon de moyenne $\bar{x}$, nous savons que $\bar{x}$ est distribuée suivant une loi normale $(m, \frac{\sigma^2}{n})$.

De même, la moyenne $\bar{\bar{x}}$ des N - n pièces résiduelles est distribuée suivant une loi normale $(m, \frac{\sigma^2}{N - n})$.

Dans ces conditions, $\bar{\bar{x}} - \bar{x}$ est distribuée suivant une loi normale de moyenne O et de variance $\sigma^2 \left| \frac{1}{n} + \frac{1}{N - n} \right| = \sigma^2 \frac{N}{n(N - n)}$.

Désignant par m_d la moyenne des N éléments,

$$m_d = \frac{n\,\bar{x} + (N - n)\,\bar{\bar{x}}}{N} = \bar{x} + \frac{N - n}{n}(\bar{\bar{x}} - \bar{x})$$

et dans ces conditions $m_d - \bar{x}$ suit une loi normale de moyenne O et de variance $\frac{N - n}{N} \cdot \frac{\sigma^2}{n}$. D'où l'estimation de m_d :

$$\hat{m}_d = \bar{x}$$

L'intervalle de confiance de $\hat{m}_d$ est de la forme :

$$\bar{x} \pm t_\alpha \sqrt{\frac{N - n}{N}} \cdot \frac{\sigma}{\sqrt{n}}$$

avec t_α découlant des tables de la loi normale. Ce dernier résultat est toutefois de peu d'intérêt puisqu'on ne connaît pas σ.

Pour tourner la difficulté il suffit de remarquer :

- d'une part que

$$\vartheta = \frac{m_d - \bar{x}}{\frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N - n}{N}}}$$

est distribuée suivant une loi normale réduite

- d'autre part que

$$\chi^2 = \frac{n\,\sigma_1'^2}{\sigma^2} \quad \text{(avec } \sigma_1'^2 = \frac{\sum (x - \bar{x})^2}{n} \text{ variance de l'échantillon observé)}$$

est distribuée suivant une loi de χ^2 à n - 1 degrés de liberté.

On en déduit, en effet, immédiatement que :

$$t = \frac{\vartheta}{\sqrt{\frac{\chi}{n - 1}}} = \frac{m_d - \bar{x}}{\hat{\sigma}/\sqrt{n}} \cdot \sqrt{\frac{N}{N - n}} \quad (1)$$

est distribuée suivant une loi de Student-Fisher à n - 1 degrés de liberté.

L'intervalle de confiance de $\hat{m}_d$ est ainsi défini par :

$$\bar{x} \pm t_\alpha \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N - n}{N}}$$

t_α découlant de la table de Student-Fisher (avec $\nu = n - 1$ degré de liberté).

(1) avec $\hat{\sigma}^2 = \frac{n}{n - 1}\,\sigma'^2$.

III - ESTIMATION DE σ_d^2

Désignons par

$$x_{11}, x_{12}, \ldots, x_{1n}$$

les résultats échantillonnés et par

$$x_{21}, x_{22}, \ldots, x_{2,N-n}$$

les valeurs de N - n éléments résiduels.

Par définition

$$\sigma_d^2 = \frac{\sum (x_{1i} - m_d)^2}{N} + \frac{\sum (x_{2j} - m_d)^2}{N}$$

Nous pouvons d'autre part écrire :

$$\sum (x_{1i} - m_d)^2 = \sum (x_{1i} - \bar{x})^2 + n(\bar{x} - m_d)^2$$

$$\sum (x_{2j} - m_d)^2 = \sum (x_{2j} - \bar{\bar{x}})^2 + (N - n)\ (\bar{\bar{x}} - m_d)^2$$

De son côté, compte-tenu de la relation :

$$n\bar{x} + (N - n)\ \bar{\bar{x}} = N m_d$$

la somme

$$n(\bar{x} - m_d)^2 + (N - n)\ (\bar{\bar{x}} - m_d)^2 = \frac{n(N - n)}{N}\ (\bar{x} - \bar{\bar{x}})^2$$

En définitive, considérant $N\sigma_d^2$ après division par σ^2, variance de la population de référence, on peut écrire :

$$\frac{N\sigma_d^2}{\sigma^2} = \frac{\sum (x_{1i} - \bar{x})^2}{\sigma^2} + \frac{\sum (x_{2j} - \bar{\bar{x}})^2}{\sigma^2} + \frac{n(N - n)}{N}\ \frac{(\bar{x} - \bar{\bar{x}})^2}{\sigma^2}$$

Compte tenu du raccordement à la loi normale

$\frac{N\sigma_d^2}{\sigma^2}$ suit une loi de χ^2 à N - 1 degré de liberté

$\frac{\sum (x_{1i} - \bar{x})^2}{\sigma^2}$ suit une loi de χ^2 à n - 1 degrés de liberté

$\frac{\sum (x_{2j} - \bar{\bar{x}})^2}{\sigma^2}$ suit une loi de χ^2 à N - 1 degrés de liberté

$\frac{n(N - n)}{N}\ \frac{(\bar{x} - \bar{\bar{x}})^2}{\sigma^2}$ suit une loi de χ^2 à 1 degré de liberté(1)

----

(1) Rappelons en effet qu'il revient au même de dire que $\frac{\bar{x} - \bar{\bar{x}}}{\sigma\sqrt{\frac{N}{n(N - n)}}}$ suit une loi normale réduite (cf. le paragraphe consacré à l'estimation de m_d) ou que son carré suit une loi de χ^2 à 1 degré de liberté.

En vertu du théorème de Cochran, les 3 dernières expressions sont indépendantes. Dans ces conditions :

$$\frac{\sum (x_{2i} - \bar{\bar{x}})^2}{\sigma^2} + \frac{n(N - n)}{N} \frac{(\bar{x} - \bar{\bar{x}})^2}{\sigma^2} = \frac{N\sigma_d^2}{\sigma^2} - \frac{\sum (x_{1i} - \bar{x})^2}{\sigma^2}$$

(distribuée suivant une loi de χ^2 à N - n degrés de liberté) et $\frac{\sum (x_{1i} - \bar{x})^2}{\sigma}$ (distribuée suivant une loi de χ^2 à n - 1 degrés de liberté) *sont deux expressions indépendantes.*

A ce stade, nous disposons de tous les éléments nécessaires pour calculer l'intervalle de confiance de σ_d^2. Il suffit de se rappeler que le rapport

$$F = \frac{\chi_1^2 / \nu_1}{\chi_2^2 / \nu_2},$$

χ_1^2 et χ_2^2 étant des variables indépendantes distribuées suivant des lois de χ^2 à ν_1 et ν_2 degrés de liberté respectivement, est lui-même distribué suivant une loi de Snedecor à (ν_1, ν_2) degrés de liberté. Par application directe de ce résultat le rapport

$$F = \frac{\dfrac{N\sigma_d^2 - \sum (x_{1i} - \bar{x})^2}{\sigma^2} / N - n}{\dfrac{\sum (x_{1i} - \bar{x})^2}{\sigma^2} / n - 1} = \frac{(N\sigma_d^2 - n\sigma_1'^2) / N - n}{n\sigma_1'^2 / n - 1}$$

est en effet distribué suivant une loi de Snedecor avec $\nu_1 = N - n$ et $\nu_2 = n - 1$ degrés de liberté.

Au risque α choisi, on peut donc déduire pour le rapport

$$\frac{(N\sigma_d^2 - n\sigma_1'^2) / N - n}{n\sigma_1'^2 / n - 1}$$

un intervalle de confiance tel que

$$\Pr\left\{\frac{1}{F_{\frac{\alpha}{2}}(\nu_2, \nu_1)} < \frac{(N\sigma_d^2 - n\sigma_1'^2) / N - n}{n\sigma_1'^2 / n - 1} < F_{1-\frac{\alpha}{2}}(\nu_1 \ \nu_2)\right\} = 1 -$$

De là, $\sigma_1'^2$ étant connu, découle l'intervalle de confiance pour σ_d^2.

On peut comparer les résultats que donnent :

<u>Observation</u> :

1) l'estimation de σ^2 par la méthode classique.

2) l'estimation de σ_d^2 par la méthode ici décrite.

La quantité

$$\chi^2 = \frac{n\sigma_1'^2}{\sigma^2} = \frac{(n - 1)\sigma^2}{\sigma^2}$$

est distribuée comme χ^2 ($\nu = n - 1$).

La quantité

$$F = \frac{(N\sigma_d^2 - n\,\sigma_1'^2)/(N - n)}{n\,\sigma_1'^2/(n - 1)} = \frac{[N\,\sigma_d^2 - (n - 1)\ \hat{\sigma}^2)/(N - n)}{\hat{\sigma}^2}$$

est distribuée comme $F(\nu_1 = N - n \quad \nu_2 = n - 1)$. Si N est grand, n étant négligeable par rapport à N sans être trop faible, F devient voisin de $\hat{\sigma}_d^2/\sigma^2$ et sa distribution est voisine de la distribution de $F(\infty, \nu_2)$, c'est-à-dire de la distribution de $\frac{\nu_2}{\chi^2(\nu_2)} = \frac{n - 1}{\chi^2(n - 1)}$.

La distribution de $\frac{(n - 1)\,\hat{\sigma}^2}{\sigma_d^2}$ est pratiquement celle de $\chi^2 = \frac{(n - 1)\,\hat{\sigma}^2}{\sigma^2}$ comme on pouvait s'y attendre.

Exemple I - $N = 1010 \quad n = 10 \quad (n - 1)\,\hat{\sigma}^2 = 400$

On aura :

1) $0,95 = \Pr\left[2,70 < \frac{400}{\sigma^2} < 19\right]$

2) $0,95 = \Pr\left[\frac{1}{2,13} < \frac{1010\ \sigma_d^2 - 400}{400} \times \frac{9}{1000} < 3,34\right]$

soit

$0,95 = \Pr\,[21,05 < \sigma^2 < 148,1]$

$0,95 = \Pr\,[21,03 < \sigma < 147,4]$

Exemple II - $N = 100 \quad n = 20 \quad (n - 1)\,\hat{\sigma}^2 = 400$

1) $0,95 = \Pr\left[8,91 < \frac{400}{\sigma^2} < 32,9\right]$

2) $0,95 = \Pr\left[\frac{1}{1,90} < \frac{100\ \sigma_d^2 - 400}{400} \times \frac{19}{80} < 2,24\right]$

soit

1) $0,95 = \Pr\,[12,2 < \sigma^2 < 44,8]$

2) $0,95 = \Pr\,[12,8 < \sigma_d^2 < 41,7]$