REVUE DE STATISTIQUE APPLIQUÉE

A. VESSEREAU

Intervalles de confiance et tests dans le cas de changement de variable cas de la loi log-normale

Revue de statistique appliquée, tome 21, n° 1 (1973), p. 59-66

<http://www.numdam.org/item?id=RSA_1973__21_1_59_0>

INTERVALLES DE CONFIANCE ET TESTS DANS LE CAS DE CHANGEMENT DE VARIABLE CAS DE LA LOI LOG-NORMALE

A. VESSEREAU

Le calcul des intervalles de confiance et les tests classiques de comparaison de moyennes et de variances s'effectuent de façon particulièrement simple lorsque la variable étudiée est distribuée suivant une loi normale.

Lorsqu'il n'en est pas ainsi il est souvent conseillé de rechercher un changement de variable $Y = \varphi(X)$ permettant de "normaliser" la distribution : les méthodes classiques pourront alors être appliquées à la variable Y. Mais on s'abstient trop souvent de dire si et comment les résultats obtenus sur Y peuvent être appliqués à la variable initiale X. Par exemple, les limites d'un intervalle de confiance d'un paramétre de Y (moyenne, écart-type) étant L_i et L_s, peut-on en déduire que pour le paramètre correspondant dans la loi de X, les limites sont $\varphi^{-1}(L_i)$ et $\varphi^{-1}(L_s)$? Quelques articles traitant de cette question sont cités dans la bibliographie figurant à la fin de cette note.

Dans celle-ci on se propose d'étudier ce problème lorsque la loi de X est une loi log-normale.

1. POSITION DU PROBLEME

La variable X est distribuée en loi log-normale, avec une espérance mathématique $E(X) = m_x$, et un écart-type σ_x.

Les estimations ponctuelles (sans biais) de m_x et σ_x^2, calculées à partir d'un échantillon de n valeurs x_i sont $\bar{x} = \frac{\Sigma x_i}{n}$ et $s_x^2 = \frac{1}{n-1}\Sigma(x_i - \bar{x})^2$.

Si l'on a deux échantillons, correspondant à des distributions de paramètres (m_x, σ_x) et (m'_x, σ'_x) :

l'estimation (sans biais) de $m - m'$ est $\bar{x} - \bar{x}'$

l'estimation de $\frac{\sigma_x^2}{\sigma_x'^2}$ est $\frac{s_x^2}{s_x'^2}$.

La variable $Y = \log_e X$ est distribuée en loi normale, avec l'espérance mathétique $E(Y) = m_y$ et l'écart-type σ_y.

En posant $C_x = \frac{\sigma_x}{m_x}$ (coefficient de variation de X), on a, entre les paramètres des lois de X et de Y, les relations suivantes :

$$m_x = e^{m_y + \frac{\sigma_y^2}{2}} \qquad (1) \qquad\qquad \sigma_x^2 = e^{2(m_y + \frac{\sigma_y^2}{2})} (e^{\sigma_y^2} - 1) \qquad (2)$$

$$m_y = \log_e \frac{m_x}{\sqrt{1 + C_x^2}} \qquad (3) \qquad\qquad \sigma_y^2 = \log_e (1 + C_x^2) \qquad (4)$$

L'échantillon "associé" aux x_i : $y_1, y_2, \ldots\ldots, y_n$ (avec $y_i = \log x_i$) a pour moyenne arithmétique $\bar{y}$ et la variance estimée est $s_y^2 = \frac{1}{n-1} \sum_i (y_i - \bar{y})^2$

Dans quelles conditions les informations apportées, par l'échantillon des y_i, sur les paramètres de la loi de Y sont-elles utilisables pour les paramètres de la variable initiale X ?

2. INTERVALLES DE CONFIANCE

On ne considérera que les intervalles bilatéraux, symétriques en probabilité, au niveau de confiance $1 - \alpha$.

2.1. Intervalle de confiance de la moyenne

Lorsque l'écart-type σ_y est connu (ce qui implique, d'après (4), que le coefficient de variation C_x est connu, ce qui paraît être un cas très exceptionnel), les limites L_i et L_s de l'intervalle de m_y sont :

$$L_i = \bar{y} - u_{1-\alpha/2} \frac{\sigma_y}{\sqrt{n}} \qquad\qquad L_s = \bar{y} + u_{1-\alpha/2} \frac{\sigma_y}{\sqrt{n}}$$

$u_{1-\alpha/2}$ étant le quantile d'ordre $1 - \frac{\alpha}{2}$ de la variable normale réduite.

Lorsque σ_y inconnu est estimé par s_y, on a :

$$L_i = \bar{y} - t_{1-\alpha/2}\frac{s_y}{\sqrt{n}} \qquad L_s = \bar{y} + t_{1-\alpha/2}\frac{s_y}{\sqrt{n}}$$

$t_{1-\alpha/2}$ étant le quantile d'ordre $1 - \frac{\alpha}{2}$ de la variable t à $\nu = n - 1$ degrés de liberté.

De la relation

$$\Pr[L_i < m_y < L_s] = 1 - \alpha$$

on déduit, compte tenu de (3) :

$$\Pr[e^{L_i} < \frac{m_x}{\sqrt{1 + C_x^2}} < e^{L_s}] = 1 - \alpha.$$

Il en résulte que e^{L_i}, e^{L_s} ne constituent (approximativement) des limites de confiance de m_x – d'ailleurs non symétriques – qu'à la condition que le coefficient de variation C_x soit petit.

2.2. Intervalle de confiance de la variance (de l'écart-type)

Les limites de l'intervalle de confiance de σ_y^2 sont :

$$L_i = \frac{(n-1)s_y^2}{\chi^2_{1-\alpha/2}} \qquad L_s = \frac{(n-1)s_y^2}{\chi^2_{\alpha/2}}$$

$\chi^2_{\alpha/2}$ étant le quantile d'ordre $\alpha/2$ de la variable χ^2 à $\nu = n - 1$ degrés de liberté.

De la relation

$$\Pr[L_i < \sigma_y^2 < L_s] = 1 - \alpha$$

on déduit, compte tenu de (4) :

$$\Pr[e^{L_i} - 1 < C_x^2 < e^{L_s} - 1] = 1 - \alpha$$

$e^{L_i} - 1$, $e^{L_s} - 1$ sont des limites de confiance, non de σ_x^2, mais de $C_x^2 = \frac{\sigma_x^2}{m_x^2}$;

$$\sqrt{e^{L_i} - 1} \qquad \text{et} \qquad \sqrt{e^{L_s} - 1}$$

sont des limites de confiance du coefficient de variation C_x.

2.3. Intervalle de confiance de la différence de deux moyennes

a) Si l'on suppose connus les écarts-types σ_y, σ'_y (ce qui entraîne que les coefficients de variation C_x, C'_x sont connus) on a, pour la variable Y, en posant

$$\sigma_d^2 = \frac{\sigma_y^2}{n} + \frac{\sigma_y'^2}{n'}$$

$$L_i = (\bar{y} - \bar{y}') - u_{1-\alpha/2}\,\sigma_d \qquad L_s = (\bar{y} - \bar{y}') + u_{1-\alpha/2}\,\sigma_d$$

$$\Pr[L_i < m_y - m'_y < L_s] = 1 - \alpha$$

entraîne :

$$\Pr[e^{L_i} < \frac{e^{m_y}}{e^{m'_y}} < e^{L_s}] = 1 - \alpha$$

d'où, compte tenu de (3)

$$\Pr[e^{L_i} < \frac{m_x}{m'_x}\sqrt{\frac{1 + C'^2_x}{1 + C_x^2}} < e^{L_s}] = 1 - \alpha$$

Si les coefficients de variation C_x et C'_x sont égaux, (d'où $\sigma_y = \sigma'_y$), e^{L_i}, e^{L_s} sont des limites de confiance exactes, non de la différence $m_x - m'_x$ mais du rapport m_x/m'_x. Si, sans être égaux, ils sont petits tous les deux, ce sont des limites de confiance approximatives.

b) Si les écarts-types σ_y et σ'_y sont inconnus et supposés égaux, cela revient à supposer que les coefficients de variation C_x, C'_x sont égaux. Dans les expressions de L_i et L_s on remplace u par t ($\nu = n + n' - 2$ degrés de liberté) et σ_d par son estimation s_d. On a alors :

$$\Pr\left[e^{L_i} < \frac{m_x}{m'_x} < e^{L_s}\right] = 1 - \alpha$$

de sorte que e^{L_i}, e^{L_s} sont des limites de confiance exactes pour le rapport $\frac{m_x}{m'_x}$.

2.4. Intervalle de confiance du rapport de deux variances

Les limites de confiance de $\sigma_y^2/\sigma_y'^2$ sont :

$$L_i = \frac{1}{F_{1-\alpha/2}(n-1,\, n'-1)} \frac{s_y^2}{s_y'^2} \qquad L_s = F_{1-\alpha/2}(n'-1,\, n-1)\frac{s_y^2}{s_y'^2}$$

$F_{1-\alpha/2}(n-1,\, n'-1)$ étant le quantile d'ordre $1-\frac{\alpha}{2}$ de la variable F à $\nu = n-1$, $\nu' = n'-1$ degrés de liberté.

$$\Pr\left[L_i < \frac{\sigma_y^2}{\sigma_y'^2} < L_s\right] = 1-\alpha$$

entraîne, d'après (4) :

$$\Pr\left[L_i < \frac{\log(1+C_x^2)}{\log(1+C_x'^2)} < L_s\right] = 1-\alpha$$

Si les coefficients de variation C_x, C'_x sont petits tous les deux, on a approximativement :

$$\Pr\left[L_i < \frac{C_x^2}{C_x'^2} < L_s\right] = 1-\alpha$$

L_i, L_s sont des limites de confiance (approximatives), non du rapport des variances $\sigma_x^2/\sigma_x'^2$ mais du rapport des carrés des coefficients de variation $C_x^2/C_x'^2$; $\sqrt{L_i}$ et $\sqrt{L_s}$ sont des limites de confiance (approximatives) du rapport C_x/C'_x.

On remarquera que $C_x = C'_x$ entraînerait que l'on a de façon certaine $\sigma_y^2 = \sigma_y'^2$, d'où $L_i = L_s = 1$ pour le rapport $\sigma_y^2/\sigma_y'^2$.

3. TESTS D'HYPOTHESE

Là encore on ne considèrera que des tests bilatéraux, symétriques en probabilité, au risque de 1ère espèce α.

3.1. Test de comparaison d'une moyenne à une valeur donnée

L'hypothèse (H_o) $m_x = a$, a étant un nombre donné, ne peut être testée directement à partir des résultats x_i.

Par contre, on sait (cf. (3) et (4)) que, sous l'hypothèse (H_o) y est distribuée normalement avec

$$\text{la moyenne} \qquad \log_e \frac{a}{\sqrt{1 + C_x^2}}$$

$$\text{la variance} \qquad \frac{1}{2} \log_e (1 + C_x^2)$$

Cette distribution dépend de C_x^2. Le test ne peut donc être effectué que si l'on connaît le coefficient de variation de X. S'il en est ainsi, la condition de rejet de l'hypothèse (H_o) est

$$\sqrt{n} \left| \frac{\bar{y} - \log_e (a/\sqrt{1 + C_x^2}}{\sqrt{\log_e (1 + C_x^2)}} \right| > u_{1-\alpha/2}$$

3.2. Test de comparaison d'une variance à une valeur donnée

L'hypothèse (H_o) est $\sigma_x^2 = A^2$, A^2 étant une valeur donnée. Sous cette hypothèse (cf. (4)) la quantité

$$\frac{(n-1)\, s_y^2}{\sigma_y^2} = \frac{(n-1)\, s_y^2}{\log_e (1 + C_x^2)} = \frac{(n-1)\, s_y^2}{\log_e \left(1 + \dfrac{A^2}{m_x^2}\right)}$$

est distribuée en loi de χ^2 à $\nu = n - 1$ degrés de liberté. Comme elle dépend de m_x, le test ne peut être effectué que si la moyenne de la distribution de X est connue. S'il en est ainsi, l'hypothèse (H_o) est rejetée si cette quantité est inférieure à $\chi^2_{\alpha/2}(n-1)$ ou supérieure à $\chi^2_{1-\alpha/2}(n-1)$.

Par contre, s'il s'agit de tester l'hypothèse de l'égalité du coefficient de variation de X à une valeur donnée C, le test peut être effectué à partir de

$$\frac{(n-1)\, s_y^2}{\log_e (1 + C^2)} = \chi^2 (\nu = n - 1) \text{ sous l'hypothèse } C_x = C.$$

L'hypothèse $C_x = C$ est rejetée si $\frac{(n-1)\,s_y^2}{\log_e(1+C^2)}$ est extérieur à l'intervalle $\chi^2_{\alpha/2}\ (\nu = n-1),\ \chi^2_{1-\alpha/2}\ (\nu = n-1)$.

3.3. Test de comparaison de deux moyennes

L'hypothèse (H_o) est $m_x - m'_x = 0$, qui est équivalent à

$$\log_e m_x - \log_e m'_x = 0$$

ou encore, d'après (3) :

$$m_y - m'_y = \log_e \sqrt{\frac{1+C'^2_x}{1+C^2_x}}$$

Ce test ne peut donc être effectué que si C_x et C'_x sont connus.

S'il en est ainsi, en posant $\log_e \sqrt{\frac{1+C'^2_x}{1+C^2_x}} = D$, l'expression :

$$(\bar{y} - \bar{y}' - D) \Big/ \sqrt{\frac{1}{n}\log_e(1+C^2_x) + \frac{1}{n'}\log_e(1+C'^2_x)}$$

suit la loi normale réduite. L'hypothèse (H_o) est rejetée si cette expression est extérieure à l'intervalle $\pm\, u_{1-\alpha/2}$.

Lorsque les coefficients de variation C_x et C'_x sont égaux, l'expression précédente prend la forme plus simple :

$$\sqrt{\frac{nn'}{n+n'}}\,(\bar{y} - \bar{y}') \Big/ \sqrt{\log_e(1+C^2_x)}$$

3.4. Test de comparaison de deux variances

L'hypothèse (H_o) est $\sigma^2_x = \sigma'^2_x$.

$$\frac{s^2_y}{\sigma^2_y} \Big/ \frac{s'^2_y}{\sigma'^2_y} = \frac{s^2_y \cdot \log_e(1+C'^2_x)}{s'^2_y \cdot \log_e(1+C^2_x)} = \frac{s^2_y \cdot \log_e\,(1+\sigma^2_x/m^2_x)}{s'^2_y \cdot \log_e\,(1+\sigma'^2_x/m'^2_x)}$$

suit la loi de F à $\nu = n - 1, \nu' = n' - 1$ degrés de liberté.

Cette propriété ne peut être utilisée pour le test de l'hypothèse (H_o) que si les moyennes m_x, m'_x sont égales. Lorsqu'il en est ainsi, $s_y^2/s_y'^2$ suit, sous l'hypothèse (H_o), la loi de $F(n-1, n'-1)$. L'hypothèse (H_o) est rejetée si $s_y^2/s_y'^2$ est extérieur à l'intervalle $\frac{1}{F_{1-\alpha/2}(n'-1, n-1)}, F_{1-\alpha/2}(n-1, n'-1)$.

Par contre, s'il s'agit de tester l'hypothèse de l'égalité des coefficients de variation C_x, C'_x, le rapport des variances $s_y^2/s_y'^2$ suit la loi de F, si cette hypothèse est vraie, ce qui permet d'effectuer le test.

4. CONCLUSION

Les intervalles de confiance déterminés, et les tests effectués sur la variable transformée $Y = \log X$ s'appliquent aux paramètres de cette variable. Leur interprétation en terme des paramètres de la variable X est parfois possible (pas toujours), sous certaines conditions. Dans certains cas (par exemple, intervalle de confiance ou test relatif à une moyenne) celles-ci sont très restrictives et peuvent limiter fortement l'intérêt du changement de variable.

BIBLIOGRAPHIE

FINNEY. – On the distribution of a variate whose logarithm is normally distributed. Journal of the Royal Statistical Society (B) *7*, 1941 p. 155/61.

NEYMAN and SCOTT. – Correction for bias introduced by a transformation of variables. Annals of Mathematical Statistics *31*, 1960, p. 643/55.

DAGNELIE. – A propos des transformations de variables. Biométrie-Praximétrie *6*, 1965 p. 59/78.

DAGNELIE. – Relation entre le coefficient de variation d'une variable et l'écart-type de son logarithme. Biométrie-Praximétrie *11*, 1970 n° 4, p. 117/23.