ΑΝΝΑΣ ΝΙΚΟΛΑΟΥ

Η ΧΡΗΣΗ ΤΩΝ EDGEWORTH ΑΝΑΠΤΥΓΜΑΤΩΝ ΓΙΑ ΤΗΝ ΚΑΤΑΣΚΕΥΗ ΔΙΑΣΤΗΜΑΤΩΝ ΕΜΠΙΣΤΟΣΥΝΗΣ

ΔΙΑΓΡΑΜΜΑ

ΠΕΡΙΛΗΨΗ

Η θεωρία μέγιστης πιθανοφάνειας παρέχει προσεγγιστικά διαστήματα εμπιστοσύνης για μια παράμετρο ενδιαφέροντος θ, τα τυπικά διαστήματα $\hat{\theta} \pm z_{1-\alpha}\hat{\sigma}$, όπου $\hat{\theta}$ είναι ο εκτιμητής μέγιστης πιθανοφάνειας και $\hat{\sigma}$ είναι ένας εκτιμητής του τυπικού λάθους βασισμένος σε παραγώγιση του λογαρίθμου της συνάρτησης πιθανοφάνειας. Πρόσφατη δουλειά έχει παράγει διαστήματα εμπιστοσύνης, των οποίων η πιθανότητα κάλυψης έχει μεγαλύτερη ακρίβεια κατά μία τάξη μεγέθους. Σ' αυτή την εργασία συζητούμε τη χρήση των Edgeworth αναπτυγμάτων στην κατασκευή βελτιωμένων προσεγγιστικών διαστημάτων εμπιστοσύνης.

1. ΕΙΣΑΓΩΓΗ ΚΑΙ ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ

Ας υποθέσουμε ότι $X_1, ..., X_n$ είναι ανεξάρτητες τυχαίες μεταβλητές ισόνομα κατανεμημένες με συνάρτηση κατανομής F, μέσο μ και διακύμανση σ^2. Προκειμένου να θέσουμε διαστήματα εμπιστοσύνης για το μ, θεωρούμε το στατιστικό $T_n = \sqrt{n}\dfrac{(\overline{X}-\mu)}{\sigma}$ και μελετάμε τη δειγματική του κατανομή. Τυπικά μια πρώτης τάξης προσέγγιση δίνεται από την τυποποιημένη κανονική κατανομή· σε πολλές ενδιαφέρουσες περιπτώσεις η κατανομή του T_n δέχεται ένα Edgeworth ανάπτυγμα της μορφής

$$P(T_n \le x) = \Phi(x) + \sum_{i=1}^{k-2} \frac{P_i(F,x)}{n^{i/2}}\varphi(x) + o\left(\frac{1}{n^{(k-2)/2}}\right), \qquad (1)$$

όπου τα P_i είναι πολυωνυμικοί διορθωτικοί όροι, που καθορίζονται από τις ροπές της κατανομής πληθυσμού F (Feller, 1966). Υπό την προϋπόθεση ότι η F είναι γνωστή, το ασυμπτωτικό ανάπτυγμα του T_n μπορεί να αντιστραφεί δίνοντας

$$P\left(T_n \le + \sum_{i=1}^{k-2} \frac{S_i(F,x)}{n^{i/2}}\right) = \Phi(x) + o\left(\frac{1}{n^{(k-2)/2}}\right), \qquad (2)$$

όπου τα πολυώνυμα S_i διαφέρουν από τα P_i (Pfanzagl, 1973). Το αποτέλεσμα της αντιστροφής είναι ένα τροποποιημένο στατιστικό, η χρήση του οποίου έχει ως συνέπεια την αύξηση της ασυμπτωτικής ακρίβειας. Θέτοντας, για παράδειγμα, στη (2) $x=\Phi^{-1}(z_{1-\alpha})$, αποκτούμε ένα διάστημα εμπιστοσύνης $\{\mu: \overline{X} < C(\mu)\}$, όπου

$$C(\mu) = \mu + \sigma\frac{z_{1-\alpha}}{n^{1/2}} + \sigma\sum_{i=1}^{\kappa-2}\frac{S_i(F, z_{1-\alpha})}{n^{(i+1)/2}} \quad \text{με βαθμό} \quad 1-\alpha+o\left(n^{-(k-2)/2}\right).$$

Όταν οι ροπές του πληθυσμού είναι άγνωστες, είναι φυσικός ο πειρασμός της αντικατάστασής τους με τις αντίστοιχες δειγματικές ροπές. Κάτι τέτοιο όμως θα έχει ως αποτέλεσμα το λάθος να είναι της τάξης $O_p(1/n)$, ανεξάρτητα από το πόσο μικρό ήταν στο αρχικό ανάπτυγμα. Μια διαδικασία τροποποίησης του αρχικού pivotal στατιστικού σε πολλαπλά στάδια, έτσι ώστε να διατηρείται ο βαθμός ασυμπτωτικής ακρίβειας του αναπτύγματος, μελετήθηκε από τους Hall (1983) και Abramovitch & Singh (1985). Ο πρώτος διορθωτικός όρος σκο-

πεύει στην εξάλειψη της επίδρασης της λοξότητας του πληθυσμού. Για παράδειγμα, το Student t–στατιστικό τροποποιείται ως εξής:

$$P\left(\sqrt{n}\,\frac{\overline{X}-\mu}{s} \le x - \frac{\lambda_3\left(1+2x^2\right)}{\sigma\sqrt{n}}\right) = \Phi(x) + o\left(\frac{1}{\sqrt{n}}\right)$$

όπου $s = \sqrt{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2/n-1}$ και $\lambda_3 = \sum_{i=1}^{n}\left(X_i-\overline{X}\right)^3/ns^3$ (Hall, 1983, p. 571).

Γενικά η κατανομή οποιουδήποτε pivotal στατιστικού $\left(\hat{\theta}-\theta\right)/\sigma_n$, όπου σ_n είναι η τυπική απόκλιση του $\hat{\theta}$, του οποίου το cumulants είναι της ίδιας τάξης, όπως αυτά του στανταρισμένου μέσου T_n, προσεγγίζεται από μια Edgeworth σειρά της μορφής (1). Οι Bhattacharya and Ghosh (1978) συνεισέφεραν σημαντικά στη θεωρία των Edgeworth αναπτυγμάτων προσεγγίζοντας με αυστηρότητα την κατανομή μιας τάξης στατιστικών, που μπορούν να εκφραστούν ως "ομαλές" συναρτήσεις αθροισμάτων ανεξαρτήτων τυχαίων διανυσμάτων. Πιο συγκεκριμένα, αν $\{X_n\}_{n\ge1}$ είναι μία ακολουθία ανεξαρτήτων και ισόνομων κ–διάστατων τυχαίων διανυσμάτων, το στατιστικό του ενδιαφέροντος $\sqrt{n}\left(\hat{\theta}-\theta\right)$ έχει τη μορφή $T_n = \sqrt{n}\left(g\left(\overline{X}\right)-g(\mu)\right)$, όπου $\overline{X}$ είναι ο αριθμητικός μέσος των $\{X_n\}_{n\ge1}$ και g είναι μια συνάρτηση ορισμένη στο R^k με συνεχείς παραγώγους σε μια γειτονιά του $\mu=EX$. Στην ειδική περίπτωση μιας πραγματικής συνάρτησης g απέδειξαν ότι το ασυμπτωτικό ανάπτυγμα, που αποκτούμε ανάγοντας το πολυμεταβλητό Edgeworth ανάπτυγμα για το μέσο $\overline{X}$ σε ένα μονομεταβλητό για το $g(\overline{X})$, ταυτίζεται με το ανάπτυγμα που παίρνουμε χρησιμοποιώντας τη δ–μέθοδο για τον υπολογισμό ροπών. Αυτό το δεύτερο ανάπτυγμα είναι απλώς η Edgeworth σειρά βασισμένη στις προσεγγιστικές ροπές του στατιστικού T_n, οι οποίες υπολογίζονται υψώνοντας το περικομμένο Taylor ανάπτυγμα του $g(\overline{X})$ γύρω από το $\theta=g(\mu)$ στην κατάλληλη δύναμη και στη συνέχεια ολοκληρώνοντας κάθε όρο.

Σ' αυτή την εργασία εφαρμόζουμε τη μεθοδολογία των Bhattacharya & Ghosh (1978), για να προσεγγίσουμε τη δειγματική κατανομή του εκτιμητή μέγιστης πιθανοφάνειας ενός στοιχείου μιας p–διάστατης παραμέτρου και μελετάμε τροποποιήσεις του αντίστοιχου pivotal στατιστικού που θα δώσουν ακριβή προσεγγιστικά διαστήματα εμπιστοσύνης.

2. ΤΟ EDGEWORTH ΑΝΑΠΤΥΓΜΑ ΤΟΥ WALD ΣΤΑΤΙΣΤΙΚΟΥ ΚΑΙ BAYESIAN ΤΡΟΠΟΠΟΙΗΣΕΙΣ

Ας υποθέσουμε ότι X, X_1, X_2, ... είναι ανεξάρτητες ισόνομα κατανεμημένες τυχαίες μεταβλητές με συνάρτηση πυκνότητας που εξαρτάται από μια p–διάσταση παράμετρο $\theta=(\theta_1, \ldots, \theta_p)$. Δεδομένου ενός συνόλου παρατηρήσεων $\mathbf{x}_n=(x_1, \ldots, x_n)$, γράφουμε $\ell(\theta,\mathbf{x}_n)$ για τον λογάριθμο της συνάρτησης πιθανοφάνειας και $L(\theta,\mathbf{x}_n)=\ell(\theta,\mathbf{x}_n)/n$ για την κανονικοποιημένη μορφή του. Οι μερικές παράγωγοι της $\ell(\theta,X)$ και οι αντίστοιχες ροπές τους συμβολίζονται με

$$\ell_r(\theta,X)=(\partial/\partial\theta_r)\,\ell(\theta,X) \qquad \ell_{rs}(\theta,X)=(\partial^2/\partial\theta_r\partial\theta_s)\ell(\theta,X)$$
$$\mu_r=E\ell_s, \qquad \mu_{rs}=E\ell_{rs}, \qquad \mu_{r,st}=E\ell_r\ell_{st}, \qquad \mu_{rst}=E\ell_{rst}.$$

Ας υποθέσουμε, για απλότητα, ότι η παράμετρος έχει μόνο δύο στοιχεία, $\theta=(\theta_1,\theta_2)$, και ότι το ενδιαφέρον μας συγκεντρώνεται στο θ_1, θεωρώντας το θ_2 ως μια ενοχλητική παράμετρο. Σε ό,τι ακολουθεί προσεγγίζουμε την κατανομή του στατιστικού $z_{wald}=\sqrt{n}\left(\hat\theta_1-\theta_{01}\right)/\hat\sigma$ με το Edgeworth ανάπτυγμά της. $\hat\sigma^2=-\hat L^{11}$ είναι η ασυμπτωτική διακύμανση του $\sqrt{n}\left(\hat\theta_1-\theta_{01}\right)$ και L^{ij} είναι το αντίστροφο στοιχείο πίνακα του L_{ij}. Το καπελάκι δηλώνει υπολογισμό των συναρτήσεων στον εκτιμητή $\hat\theta$.

Θεώρημα. Κάτω από ορισμένες συνθήκες κανονικότητας (Bhattacharya & Ghosh, 1978, p. 439):

$$P\left(\sqrt{n}\,\frac{\hat\theta_1-\theta_{01}}{\hat\sigma}\le x\right)=\Phi(x)-\frac{1}{\sqrt{n}}\left(\xi_1+\xi_3\,\frac{\left(x^2-1\right)}{6}\right)\varphi(x)+o\left(\frac{1}{\sqrt{n}}\right)$$

ομοιόμορφα στο x όπου:

$$\xi_1=\frac{\mu^{1,s}}{2\sqrt{\mu^{1,1}}}\left(\mu^{t,u}\left(2\mu_{u,st}-\mu_{s,ut}\right)-\frac{\mu_{s,22}+\mu_{s22}}{\mu_{22}}\right)$$

$$\xi_3=\frac{3}{\sqrt{\mu^{1,1}}^3}\left(\mu^{1,t}\mu^{1,k}\mu^{1,s}\left(2\mu_{k,st}+\frac{\mu_{k,s,t}}{3}\right)+\mu^{1,k}\mu^{1,s}\mu^{1,t}\mu_{kst}-\mu^{1,1}\mu^{1,k}\left(\begin{matrix}\mu^{s,t}\left(\mu_{k,st}+\mu_{kst}\right)+\\+\dfrac{\mu_{k,22}+\mu_{k22}}{\mu_{22}}\end{matrix}\right)\right).$$

Οι δείκτες παίρνουν τις τιμές 1 και 2 και προσθέτουμε ως προς κάθε δείκτη που επαναλαμβάνεται.

Απόδειξη (Nicolaou, 1990). Ορίζουμε τις ασυμπτωτικά κανονικές τυχαίες μεταβλητές με μέσο μηδέν και σταθερή διακύμανση

$$Z_r = \frac{\sum_{i=1}^{n} \ell_r(\theta_0, x_i)}{\sqrt{n}}, \qquad Z_{rs} = \frac{\sum_{i=1}^{n} \left(\ell_{rs}(\theta_0, x_i) - \mu_{rs}\right)}{\sqrt{n}}.$$

Η απόκλιση του εκτιμητή μέγιστης πιθανοφάνειας $\hat{\theta}_1$ από την αληθινή τιμή της παραμέτρου έχει το ακόλουθο ασυμπτωτικό ανάπτυγμα (McGullagh, 1987)

$$\sqrt{n}\left(\hat{\theta}_1 - \theta_{01}\right) = \mu^{1,i} Z_i + \frac{2\mu^{1,i}\mu^{j,k} Z_{ij} Z_k + \mu^{1,k}\mu^{i,s}\mu^{j,t}\mu_{kst} Z_i Z_j}{2\sqrt{n}} + O_p\left(\frac{1}{n}\right) \qquad (3)$$

Πολλαπλασιάζοντας το (3) με το Taylor ανάπτυγμα του $\left(-\hat{L}^{11}\right)^{-1/2}$ γύρω από την αληθινή τιμή της παραμέτρου

$$\left(-\hat{L}^{11}\right)^{-1/2} = \left(\mu^{1,1}\right)^{-1/2}\left(1 - \frac{1}{2\sqrt{n}}\left(\mu^{i,j}\left(Z_{ij} + \mu_{rij}\mu^{r,s}Z_s\right) + \frac{Z_{22} + \mu_{r22}\mu^{r,s}Z_s}{\mu_{22}}\right)\right) + O_p\left(\frac{1}{n}\right)$$

προκύπτει το ασυμπτωτικό ανάπτυγμα του z_{wald}. Αυτό έχει τη φόρμα του Taylor αναπτύγματος ενός στατιστικού, $\sqrt{n}\,(g(\Lambda) - g(\mu))$, όπου $g(\,.\,)$ είναι κάποια πραγματική συνάρτηση του τυχαίου διανύσματος Λ, που έχει για στοιχεία του μέσους ανεξαρτήτων και ισόνομων τυχαίων μεταβλητών

$$\Lambda = \left(\frac{Z_1}{\sqrt{n}}, \ldots, \frac{Z_p}{\sqrt{n}}, \frac{Z_{11}}{\sqrt{n}}, \ldots, \frac{Z_{pp}}{\sqrt{n}}\right), \quad p=2$$

γύρω από $\mu = E\Lambda$. Η ύπαρξη του Edgeworth αναπτύγματος του z_{wald} προκύπτει από το θεώρημα 2 των Bhattacharya & Ghosh (1978). Για να καθορίσουμε τη φόρμα των δύο πρώτων όρων του, προσεγγίζουμε τα cumulants του z_{wald}, υψώνοντας το ασυμπτωτικό του ανάπτυγμα στην κατάλληλη δύναμη και υπολογίζοντας την αναμενόμενη τιμή κάθε όρου. Λαμβάνοντας υπόψη ότι

$$E(Z_r Z_s) = \mu_{r,s}, \qquad E(Z_r Z_s Z_t) = \mu_{r,s,t}/\sqrt{n}, \qquad E(Z_r Z_{st} Z_u) = (\mu_{r,st,u} + \mu_{st}\mu_{ru})/\sqrt{n},$$

$$E(Z_r Z_s Z_t Z_u) = \mu_{r,s}\mu_{t,u}[3] + 0(1/n), \quad E(Z_r Z_s Z_t Z_{uv}) = \mu_{r,s}\mu_{t,uv}[3] + 0(1/n),$$

όπου οι αγκύλες δηλώνουν άθροιση ως προς τους τρεις συνδυασμούς δεικτών, αποδεικνύεται μετά από επίπονη αλλά στοιχειώδη άλγεβρα ότι τα τρία

πρώτα cumulants του z_{wald} είναι $\kappa_1=\xi_1+0(1/n)$, $\kappa_2=1+0(1/n)$, $\kappa_2=\xi_3+0(1/n)$. Η Edgeworth σειρά προκύπτει ευθέως.

Μέθοδοι τροποποίησης του στατιστικού z_{wald}, προκειμένου να κατασκευάσουμε βελτιωμένα διαστήματα εμπιστοσύνης, βασίζονται σε Bayesian επιχειρήματα. Ας θεωρήσουμε την posterior συνάρτηση κατανομής της θ_1, $\rho(./x_n)$. Μια πρώτης τάξης προσέγγιση του $1-\alpha$ εκατοστημορίου της είναι το όριο $\hat{\theta}_1 + z_{1-\alpha}\sqrt{-\hat{L}^{11}/n}$, που προκύπτει από το z_{wald}. Διάφοροι συγγραφείς (Welch & Peers (1963), Peers(1965), Stein (1982)) έχουν δείξει ότι προσδιορίζοντας κατάλληλα μια prior πυκνότητα, το αντίστοιχο $1-\alpha$ posterior εκατοστημόριο βελτιώνει το παραπάνω όριο με ένα διορθωτικό όρο τάξης $0(1/n)$, που εξαρτάται από την prior και τρίτης τάξης παραγώγους της συνάρτησης log–πιθανοφάνειας υπολογισμένες στο μέγιστο $\hat{\theta}$. Η γεννόμενη posterior κατανομή της θ_1, υπολογισμένη στην αληθινή τιμή της παραμέτρου θ_{01} προσεγγίζεται από την κανονική κατανομή (Nicolaou, 1990)

$$\rho(\theta_{01}/x_n)=\Phi(z_{awald})+0_p(1/n),$$

όπου

$$z_{awald} = -z_{wald}\frac{1}{\sqrt{n}}\left\{-\frac{\partial}{\partial\theta_i}\left(\hat{\mu}^{1,1}\right)^{-1/2}\hat{\mu}^{1,i}-\left(z_{wald}^2-1\right)\frac{\hat{L}^{i1}\hat{L}^{jk}\hat{L}_{ijk}}{6\sqrt{-\hat{L}^{11}}^3}+\frac{\hat{L}^{i1}\hat{L}^{j1}\hat{L}^{k1}\hat{L}_{ijk}}{2\sqrt{-\hat{L}^{11}}}\right\}.$$

Η κατανομή του τροποποιημένου στατιστικού z_{awald} μπορεί να προσεγγισθεί από μια Edgeworth σειρά με την τυποποιημένη κανονική κατανομή ως πρώτο όρο. Χάρη στην επιλεγμένη prior ο $1/\sqrt{n}$ όρος της σειράς μηδενίζεται και έτσι η κανονική προσέγγιση ισχύει μέχρι $0(1/n)$. Συνεπώς, η $\rho(\theta_{01}/x_n)$ έχει κάτω από επαναλαμβανόμενη δειγματοληψία την ομοιόμορφη κατανομή στο διάστημα $(0, 1)$ και το αντίστοιχο εκατοστημόριό της καλύπτει την αληθινή τιμή της παραμέτρου με πιθανότητα δειγματοληψίας $\alpha+O(1/n)$.

Παρατήρηση 1. Στην περίπτωση που η ενοχλητική παράμετρος θ_2 είναι ορθογώνια της παραμέτρου ενδιαφέροντος θ_1 ως προς τον αναμενόμενο πίνακα πληροφορίας του Fisher, δηλ. $\mu_{1,2}=0$, η prior πυκνότητα που οδηγεί στο τροποποιημένο pivot z_{awald} δίνεται από $\pi(\theta_1,\theta_2)=\sqrt{\mu_{1,1}(\theta_1,\theta_2)}g(\theta_2)$, όπου $g(.)$ είναι μια αυθαίρετη συνάρτηση της θ_2.

Παρατήρηση 2. Ας θεωρήσουμε το εναλλακτικό pivotal στατιστικό για τον έλεγχο μιας δοσμένης τιμής της θ_1,

$z_{dev} = \text{sgn}\left(\theta_{01} - \hat{\theta}_1\right)\sqrt{2\left(\ell\left(\hat{\theta}_1, \hat{\theta}_2\right) - \ell\left(\theta_{01}, \tilde{\theta}_2\right)\right)}$, όπου $\tilde{\theta}_2$ είναι ο περιορισμένος εκτιμητής μέγιστης πιθανοφάνειας, όταν το θ_1 παίρνει την τιμή θ_{01}. Το τροποποιημένο Wald στατιστικό μπορεί να εκφραστεί ως $z_{awald} = z_{dev} - E_n(z_{dev}) + 0(1/n)$, όπου $E_n(z_{dev})$ είναι ο posterior μέσος του z_{dev}.

ΒΙΒΛΙΟΓΡΑΦΙΑ

L. Abramovitch and K. Singh, «Edgeworth corrected pivotal statistics and boutstrap», *Annals of Statistics,* 13:116-32, 1985.

R. N. Bhattacharya and I. K. Ghosh, «On the validity of the formal edgeworth expansion», *Annals of Statistics,* 6:434-51, 1978.

W. Feller, *An Introduction to Probability and its Applications,* Vol. II, New York: John Wiley, 1966.

P. Hall, «Inverting an edgeworth expansion», *Annals of Statistics,* 11:569-76, 1983.

P. McGullagh, *Tenson Methods in Statistics,* London: Chapman and Hall, 1986.

A. Nicolaou, «Condidence intervals for a scalar parameter in the presence of nuisance parameters», *PhD Thesis,* Yale University, New Haven, 1990.

H. W. Peers, «On confidence points and Bayesian probability points in the case of several parameters», *J. R. Statist. Soc.* B, 27:9-16, 1965.

J. Pfanzagl, «Asymptotic expansions related to minimum contrast estimators», *Annals of Statistics,* 1:993-1026, 1973.

B. L. Welch and H. W. Peers, «On formulae for confidence points based on integrals of weighted likelihoods», *J. R. Statist. Soc.* B, 25:318-29, 1963.