

Σημειακή Εκτίμηση και ιδιότητες σημειακών εκτιμητριών

Το πρόβλημα της Εκτιμητικής είναι η εκτίμηση (προσέγγιση) κάποιου άγνωστου χαρακτηριστικού ενός υπό μελέτη πληθυσμού που περιγράφεται από μία παράμετρο θ .

Επιθυμούμε να εκτιμήσουμε την πραγματική αλλά άγνωστη σε εμάς τιμή της παραμέτρου θ . Η σημειακή εκτίμηση της παραμέτρου θ γίνεται με τη

βοήθεια ενός τυχαίου δείγματος n παρατηρήσεων

$X_1, X_2, \dots, X_n$ από τον πληθυσμό. Με άλλα λόγια,

η εκτίμηση της άγνωστης τιμής της παραμέτρου

θ του πληθυσμού γίνεται μέσω "πληροφοριών" που

παρέχονται από ένα τυχαίο δείγμα. Μία στατιστική

συνάρτηση (σ.σ.) $\hat{\theta}_n$ οποία είναι συνάρτηση μόνο

των παρατηρήσεων του δείγματος, δηλ. $\hat{\theta}_n = h(X_1, \dots, X_n)$

χρησιμοποιείται για να εκτιμήσουμε την παράμετρο θ .

Πριν την επιλογή του δείγματος, οι $X_1, \dots, X_n$

είναι τ.ρ. Επομένως, η σ.σ. $\hat{\theta}_n$ είναι επίσης μία τ.ρ.

Η κατανομή της σ.σ. $\hat{\theta}_n$ καλείται δειγματική κατανομή της $\hat{\theta}_n$.

Εάν χρησιμοποιήσουμε μία σ.σ. για να εκτιμήσουμε μία παράμετρο θ , ενδεχομένως, η εκτίμηση αυτή να διαφέρει σημαντικά από την πραγματική τιμή της παραμέτρου θ .

Επομένως, θα πρέπει να καθορίσουμε κάποιον κανόνα ή κάποιο κριτήριο που θα δίνει "κανονιστικά"

αποτελέσματα στην ευτίρηση της άγνωστης τιμής της παραμέτρου, είτε "ραυροπρόθεσμα", είτε "κατά μέσο όρο" είτε με "υψηλή πιθανότητα".

Οποιοδήποτε κανόνα ή κριτήριο (μέθοδο) ευτίρησης και αν χρησιμοποιήσουμε θα μας οδηγήσει στην κατασκευή κατάλληλων κατανομών για τις ευτιρήτριες της παραμέτρου. Συνεπώς, το πρόβλημά μας δεν είναι να βρούμε ευτιρήσεις αλλά να βρούμε κατάλληλες ευτιρήτριες. Η ευτίρηση θα διαφέρει σημαντικά από την πραγματική τιμή της παραμέτρου. Η ευτιρήτρια θα απορρίπτεται όταν δίνει άσχημα αποτελέσματα ευτίρησης ραυροπρόθεσμα δηλαδή όταν η κατανομή όλων των πιθανών τιμών της ευτιρήτριας αποκλίνει σημαντικά από την πραγματική τιμή της παραμέτρου θ.

Η ευτιρήτρια (estimator) είναι μία τ.φ. που χρησιμοποιείται για να ευτιρήσει ένα χαρακτηριστικό του πληθυσμού το οποίο περιγράφεται συνήθως από μία (ή περισσότερες) παραμέτρους. Η αριθμητική τιμή που λαμβάνει η ευτιρήτρια για συγκεκριμένο δείγμα ονομάζεται ευτίρηση (estimate).

Ιδιότητες σημαντικών ευτιρητριών

Τα βασικά κριτήρια επιλογής ευτιρητριών είναι:
(α) Αμεροληψία (β) Ελάχιστη διασπορά - Αποτελεσματικότητα (γ) Συνέπεια (δ) Επάρκεια.

Η τ.ρ. $\hat{\theta}_n = h(X_1, X_2, \dots, X_n)$ καλείται ευτιμήτρια και η τιμή της $h(x_1, \dots, x_n)$ καλείται ευτιμήτης.

Η μελέτη των ιδιοτήτων των ευτιμήτριών βασίζεται στη μελέτη της δειγματικής κατανομής τους. Η κατανομή αυτή και οι ιδιότητές της, θα μας δείξει κατά πόσο η ευτιμήτρια $\hat{\theta}_n$ από ένα τυχαίο δείγμα είναι πλησιέστερα προς την αληθινή τιμή της παραμέτρου θ .

(α) Αμεροληψία (unbiasedness)

Το κριτήριο αμεροληψίας βασίζεται στην έννοια της μέσης (αναμενόμενης) τιμής της ευτιμήτριας.

Μία ευτιμήτρια $\hat{\theta}_n$ καλείται αμεροληπτική (unbiased) αν η αναμενόμενη (μέση) τιμή της δειγματικής της κατανομής ισούται με την υπό ευτίμηση παράμετρο θ του πληθυσμού δηλ. $E(\hat{\theta}_n) = \theta$.

Αν μία ευτιμήτρια $\hat{\theta}_n$ δεν είναι αμεροληπτική καλείται μεροληπτική και η ποσότητα μεροληψίας της είναι:

$$\text{bias}(\hat{\theta}_n) = E(\hat{\theta}_n) - \theta$$

Παραδείγματα (1) τ.δ. $X_1, X_2, \dots, X_n$ από ένα πληθυσμό με κατανομή $B(l, p)$.

Ο $\hat{p} = \frac{\sum_{i=1}^n X_i}{n}$ είναι αμεροληπτος ευτιμήτης του p ,

$$\begin{aligned} \text{δύτι: } E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] &= \frac{1}{n} E\left[\sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n E[X_i] \\ &= \frac{1}{n} n p = p. \end{aligned}$$

Άρα, ο $\hat{\mu} = \bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$ είναι αμερόληπτος εκτιμητής της παραμέτρου μ . -4-

(2) Οι δειγματικές ροπές κ-τάξης $\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^n X_i^k$

για ένα τ.δ. είναι αμερόληπτοι εκτιμητές των αντίστοιχων ροπών του πληθυσμού $\mu_k = E[X^k]$, διότι:

$$E[\hat{\mu}_k] = \frac{1}{n} \sum_{i=1}^n E[X_i^k] = \frac{1}{n} n \mu_k = \mu_k.$$

(3) Ο δειγματικός μέσος $\bar{X}_n$ ενός τ.δ. μεγέθους n είναι πάντοτε (για κάθε κατανομή), ένας αμερόληπτος εκτιμητής του μέσου μ του πληθυσμού, διότι:

$$\begin{aligned} E[\bar{X}_n] &= E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n E[X_i] \\ &= \frac{1}{n} n \mu = \mu. \end{aligned}$$

Αντίθετα, η δειγματική διασπορά

$$s_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \quad \underline{\text{δεν}} \text{ είναι αμερόληπτη}$$

εκτιμήτρια της διασποράς σ^2 του πληθυσμού,

διότι

$$\begin{aligned} s_n^2 &= \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n} \left(\sum_{i=1}^n (X_i^2 - 2X_i\bar{X}_n + \bar{X}_n^2) \right) \\ &= \frac{1}{n} \left(\sum_{i=1}^n X_i^2 - 2\bar{X}_n \sum_{i=1}^n X_i + n\bar{X}_n^2 \right) \end{aligned}$$

$$= \frac{1}{n} \left(\sum_{i=1}^n X_i^2 - 2 \bar{X}_n n \bar{X}_n + n \bar{X}_n^2 \right)$$

$$= \frac{1}{n} \left(\sum_{i=1}^n X_i^2 - 2n \bar{X}_n^2 + n \bar{X}_n^2 \right)$$

$$= \frac{1}{n} \left(\sum_{i=1}^n X_i^2 - n \bar{X}_n^2 \right) = \frac{\sum_{i=1}^n X_i^2}{n} - \bar{X}_n^2$$

Uai

$$E[S_n^2] = E \left[\frac{\sum_{i=1}^n X_i^2}{n} \right] - E[\bar{X}_n^2]$$

$$= \frac{1}{n} \sum_{i=1}^n E[X_i^2] - E[\bar{X}_n^2]$$

$$= \frac{1}{n} \sum_{i=1}^n (\text{var}(X_i) + (E[X_i])^2) - (\text{var}(\bar{X}_n) + (E[\bar{X}_n])^2)$$

$$\stackrel{(*)}{=} \frac{1}{n} \sum_{i=1}^n (\sigma^2 + \mu^2) - \left(\frac{\sigma^2}{n} + \mu^2 \right)$$

$$= \frac{1}{n} n(\sigma^2 + \mu^2) - \frac{\sigma^2}{n} - \mu^2$$

$$= \cancel{\sigma^2 + \mu^2} - \frac{\sigma^2}{n} - \cancel{\mu^2} = \sigma^2 - \frac{\sigma^2}{n} = \sigma^2 \left(1 - \frac{1}{n} \right)$$

$$= \sigma^2 \left(\frac{n-1}{n} \right) = \left(\frac{n-1}{n} \right) \sigma^2 \neq \sigma^2$$

$$(*) \text{var}(\bar{X}_n) = \text{var} \left(\frac{\sum_{i=1}^n X_i}{n} \right) = \frac{1}{n^2} \sum_{i=1}^n \text{var}(X_i)$$

$$= \frac{1}{n^2} n \sigma^2 = \frac{\sigma^2}{n}$$

Η μη-αμερόληπτη (μεροληπτική) εκτιμήτρια S_n^2 έχει -6- μεροληψία

$$\begin{aligned} \text{bias}(S_n^2) &= E(S_n^2) - \sigma^2 = \left(\frac{n-1}{n}\right)\sigma^2 - \sigma^2 = \\ &= \sigma^2\left(\frac{n-1}{n} - 1\right) = \sigma^2\left(\frac{n-1-n}{n}\right) = -\frac{\sigma^2}{n} \end{aligned}$$

Όσο το μέγεθος δείγματος n αυξάνει η μεροληψία της S_n^2 μειώνεται και καθώς $n \rightarrow \infty$ έπεται ότι:

$\text{bias}(S_n^2) \xrightarrow{n \rightarrow \infty} 0$. Συνεπώς, μπορούμε να πούμε ότι η εκτιμήτρια S_n^2 δεν είναι αμερόληπτη εκτιμήτρια της διασποράς του πληθυσμού, είναι όμως ασυμπτωτικά αμερόληπτη.

Ως αμερόληπτη εκτιμήτρια της διασποράς σ^2 λαμβάνουμε την εκτιμήτρια $S_n^{*2} = \frac{n}{n-1} S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$

δύο:

$$E(S_n^{*2}) = E\left(\frac{n}{n-1} S_n^2\right) = \frac{n}{n-1} E(S_n^2) =$$

$$= \frac{n}{n-1} \left(\frac{n-1}{n}\right) \sigma^2 = \sigma^2$$

Συνήθως, η παραπάνω εκτιμήτρια S_n^{*2} ορίζεται ως δειγματική διασπορά (αντί της S_n^2).

(6) Αποτελεσματικότητα (efficiency)

Αν και η αμεροληψία είναι μία χρήσιμη ιδιότητα, δεν μας δίνει καμία πληροφορία για την μεταβλητότητα των εκτιμητή και την αβεβαιότητα της εκτίμησης.

Αυτό επιτυγχάνεται με την αρχή της ελάχιστης διακύμανσης (ή διασποράς).

Κάτω από αριστά γενικές συνθήκες, υπάρχει ένα κατώτερο φράγμα για τη διακύμανση μίας αμερόληπτης εκτιμήτριας το οποίο προκύπτει από την ανισότητα Cramér-Rao.

Αν $X_1, X_2, \dots, X_n$ είναι ανεξάρτητες παρατηρήσεις πάνω σε μία τ.ρ. X με συνάρτηση πυκνότητας (ή συνάρτηση πιθανότητας) $P_X(x; \theta)$ και $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$ είναι μία αμερόληπτη εκτιμήτρια της παραμέτρου θ , τότε:

$$\text{Var}(\hat{\theta}_n) \geq \frac{1}{n E \left[\left(\frac{\partial}{\partial \theta} \ln P_X(x; \theta) \right)^2 \right]} \quad (*)$$

Η ποσότητα στο δεξιό μέλος της ανισότητας καλείται κάτω φράγμα των Cramér-Rao.

Η ποσότητα: $E \left[\left(\frac{\partial}{\partial \theta} \ln P_{X_1, \dots, X_n}(x_1, \dots, x_n; \theta) \right)^2 \right]$ καλείται

πληροφορία του Fisher σχετικά με την παράμετρο θ , η οποία περιέχεται στις παρατηρήσεις $X_1, \dots, X_n$ και συμβολίζεται με I_θ . Η ποσότητα πληροφορίας για μία μοναδική παρατήρηση X είναι: $E \left[\left(\frac{\partial}{\partial \theta} \ln P_X(x; \theta) \right)^2 \right] = i_\theta$

Άρα η ανισότητα (*) γράφεται:

$$\text{Var}(\hat{\theta}_n) \geq I_\theta^{-1} = (n i_\theta)^{-1}$$

Μία ευτιμήτρια $\hat{\theta}_n$ που επιτυγχάνει αυτό το κατώτερο φράγμα Cramer- Rao για όλες τις τιμές του θ καλείται

αμερόληπτη ευτιμήτρια οβιόμορφα ελάχιστης διασποράς (Α.Ο.Ε.Δ. ευτιμήτρια).

Η ανισότητα Cramer- Rao δίνει ένα τρόπο για την αναζήτηση μιας Α.Ο.Ε.Δ. ευτιμήτριας. Η σύγκριση της διασποράς μιας αμερόληπτης ευτιμήτριας $\hat{\theta}_n$ με την ποσότητα I_{θ}^{-1} οδηγεί στην έννοια της αποτελεσματικότητας (efficiency) η οποία ορίζεται ως εξής:

$$eff(\hat{\theta}_n) = \frac{I_{\theta}^{-1}}{var(\hat{\theta}_n)}$$

Εάν $eff(\hat{\theta}_n)$ είναι κοντά στην μονάδα, η επιλογή της $\hat{\theta}_n$ μοιάζει ικανοποιητική. Οι υπολογισμοί στην ανισότητα Cramer- Rao γίνονται απλούστεροι με τη χρησιμοποίηση της σχέσης: $E\left[\frac{\partial}{\partial \theta} \ln P_X(x; \theta)\right]^2 = -E\left(\frac{\partial^2}{\partial \theta^2} \ln P_X(x; \theta)\right)$ όπου $X = (X_1, \dots, X_n)$.

Για δύο αμερόληπτες ευτιμήτριες $\hat{\theta}_1, \hat{\theta}_2$ που προέρχονται από τ.δ. ίδων μεγέθους λείβε ότι η $\hat{\theta}_1$ έχει μεγαλύτερη σχετική αποτελεσματικότητα σε σχέση με την $\hat{\theta}_2$ στην εκτίμηση της παραμέτρου θ , αν έχει μικρότερη διασπορά δηλ. αν $var(\hat{\theta}_1) < var(\hat{\theta}_2)$ και $E(\hat{\theta}_1) = E(\hat{\theta}_2) = \theta$.

Η αναζήτηση ενός Α.Ο.Ε.Δ. ευτιμότη με την ανισότητα Cramer-Rao γίνεται ως εξής:

1^η βήμα Βρίσκουμε το κάτω φράγμα δια της διασποράς όλων των αμερόληπτων ευτιμητριών (για τους οποίους ισχύουν οι γενικές συνθήκες).

2^η βήμα Ζητάμε έναν αμερόληπτο ευτιμητή, του οποίου η διασπορά συμπίπτει με το κάτω φράγμα που προσδιορίσαμε. Εάν το πετύχουμε αυτό έχουμε βρει τον Α.Ο.Ε.Δ. ευτιμότη.

Παράδειγμα Δίνεται τ.δ. $X_1, \dots, X_n$ από ένα πληθυσμό με κατανομή Poisson (θ), δηλ. η συνάρτηση πιθανότητας κάθε τ.ρ. $X_i, i=1, \dots, n$ είναι $p(x; \theta) = e^{-\theta} \frac{\theta^x}{x!}, x=0, 1, 2, \dots$. Να βρεθεί ο Α.Ο.Ε.Δ. ευτιμότης του θ .

Λύση Για την κατανομή Poisson ισχύουν οι γενικές συνθήκες για κάθε αμερόληπτο ευτιμητή της παραμέτρου θ .

Το κάτω φράγμα από την ανισότητα Cramer-Rao είναι:

$(n i_{\theta})^{-1}$, όπου

$$i_{\theta} = E \left[\frac{\partial \ln p(X; \theta)}{\partial \theta} \right]^2 = E \left[\frac{\partial}{\partial \theta} (-\theta + X \ln \theta - \ln X!) \right]^2$$

$$= E \left[-1 + \frac{X}{\theta} \right]^2 = \frac{1}{\theta^2} E [X - \theta]^2 = \frac{\text{Var}(X)}{\theta^2} = \frac{\theta}{\theta^2}$$

$$= \frac{1}{\theta} \text{ Άρα } i_{\theta} = \frac{1}{\theta}. \text{ Έστω } T \text{ μία αμερόληπτη ευτιμή-}$$

$$\text{τρια της } \theta. \text{ Τότε } \text{Var}(T) \geq \frac{1}{n i_{\theta}} = \frac{1}{n (\frac{1}{\theta})} = \frac{\theta}{n}$$

Για τον αμερόληπτο ευτιμητή $\bar{X}_n$ έχουμε:

$$\text{var}(\bar{X}_n) = \frac{1}{n^2} \sum_{i=1}^n \text{var}(X_i) = \frac{\theta}{n}$$

Συνεπώς, η διασπορά του $\bar{X}_n$ συμπίπτει με το κάτω φράγμα από την ανισότητα Cramer-Rao. Άρα, ο $\bar{X}_n$ είναι Α.Ο.Ε.Δ. ευτιμότης της παραμέτρου θ .

Παράδειγμα Έστω τ.δ. $X_1, X_2, \dots, X_n$ από έναν πληθυσμό με κατανομή $B(1, p)$, δηλ. η συνάρτηση πιθανότητας κάθε τ.φ. $X_i, i=1, \dots, n$ είναι $p(x; \theta) = \theta^x (1-\theta)^{1-x}, x=0,1$. Να βρεθεί ο Α.Ο.Ε.Δ. ευτιμότης της παραμέτρου $p=\theta$.

Λύση Για την $B(1, p)$ ισχύουν οι γενικές συνθήκες για κάθε αφερόμενο ευτιμότης της παραμέτρου θ . Για το κάτω φράγμα Cramer-Rao, έχουμε ($p=\theta$):

$$\ln p(x; \theta) = x \ln \theta + (1-x) \ln(1-\theta)$$

$$\frac{\partial \ln p(x; \theta)}{\partial \theta} = \frac{x}{\theta} - \frac{1-x}{1-\theta}$$

$$I_{\theta} = E \left[\frac{X}{\theta} - \frac{1-X}{1-\theta} \right]^2$$

$$= E \left[\frac{X^2}{\theta^2} + \frac{(1-X)^2}{(1-\theta)^2} - 2 \frac{X(1-X)}{\theta(1-\theta)} \right]$$

$$= \frac{1}{\theta^2} E[X^2] + \frac{1}{(1-\theta)^2} E(1-X)^2 - \frac{2}{\theta(1-\theta)} E[X(1-X)]$$

Από τη σχέση:

$$\text{var}(X) = E[X^2] - (E[X])^2 \Rightarrow E[X^2] = \theta^2 + \theta(1-\theta) = \theta.$$

Επίσης,

$$E[1-X]^2 = 1 + E[X^2] - 2E[X] = 1 + 0 - 2 \cdot 0 = 1 - 0$$

και

$$E[X(1-X)] = E[X] - E[X^2] = 0 - 0 = 0$$

Άρα, από τις τρεις τελευταίες σχέσεις, έπεται ότι:

$$i_0 = \frac{1}{0(1-0)} \quad \text{και} \quad (n i_0)^{-1} = \frac{0(1-0)}{n}$$

Για τον αερόβηπτο ευτιμήτη $\bar{X}_n$ έχουμε ότι η διασπορά του είναι:

$$\text{var}(\bar{X}_n) = \frac{1}{n^2} \sum_{i=1}^n \text{var}(X_i) = \frac{1}{n^2} n \cdot 0(1-0) = \frac{0(1-0)}{n}$$

δηλ. $\text{var}(\bar{X}_n) =$ κάτω φράγμα Cramer-Rao και συνεπώς ο $\bar{X}_n$ είναι ο Α.Ο.Ε.Δ. ευτιμήτης της παραμέτρου $\theta = p$.

(δ) Συνέπεια (Consistency)

Στην ευτίρηση της άγνωστης τιμής ρίας παραμέτρου χρειάζεται να λάβουμε υπόψη και τις πιθανότητες οι τιμές των ευτιμήτριών να προσεγγίζουν τις άγνωστες τιμές των παραμέτρων που ευτιρούν. Επιπλέον, είναι απαραίτητο να μελετήσουμε τη συμπεριφορά των ευτιμήτριών όταν το μέγεθος του δείγματος αυξάνεται.

Μία ευτιμήτρια $\hat{\theta}_n$ είναι μία συνεπής ευτιμήτρια ρίας παραμέτρου θ του πληθυσμού αν για μία οποιαδήποτε ποσό μικρή θετική τιμή ε , $\varepsilon > 0$ ισχύει ότι:

$$\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| < \varepsilon) = 1.$$

Με άλλα λόγια, η $\hat{\Theta}_n$ είναι συνεπής, εάν η πιθανότητα $-12-$ ότι ο $\hat{\Theta}_n$ θα λάβει τιμή που διαφέρει αριθμητικά από την τιμή του Θ περισσότερο από μία αυθαίρετη σταθερή ποσότητα ϵ , τείνει στο μηδέν, όταν $n \rightarrow \infty$. Η συνέπεια είναι μία ασυμπτωτική ιδιότητα. Είναι η σύχναση κατά πιθανότητα της ευτιμήτριας $\hat{\Theta}_n$ στο Θ . Ο αρχικός ορισμός δόθηκε από τον Fisher. Ισχύει ότι αν T_n είναι μία ευτιμήτρια μιας παραμέτρου Θ και (α) η T_n είναι αμερόληπτη ευτιμήτρια της Θ και (β) $\text{var}(T_n) \rightarrow 0$, όταν $n \rightarrow \infty$, τότε η T_n είναι συνεπής ευτιμήτρια της Θ .

Παράδειγμα Έστω $X_1, X_2, \dots, X_n$ τ.δ. από έναν πληθυσμό με κατανομή $N(\mu, \sigma^2)$. Ναδειχθεί ότι ο ευτιμητής S_n^{*2} είναι συνεπής ευτιμητής της διασποράς σ^2 .

Αρκεί ναδειχθεί ότι: $E(S_n^{*2}) = \sigma^2$ και $\lim_{n \rightarrow \infty} \text{Var}(S_n^{*2}) = 0$.

Ισχύει ότι η τ.ρ. $\frac{(n-1)S_n^{*2}}{\sigma^2} \sim \chi_{n-1}^2$

Άρα $E\left[\frac{(n-1)S_n^{*2}}{\sigma^2}\right] = n-1$ και $\text{var}\left[\frac{(n-1)S_n^{*2}}{\sigma^2}\right] = 2(n-1)$

Άρα, $E[S_n^{*2}] = \sigma^2$ και $\frac{(n-1)^2}{\sigma^4} \text{var}(S_n^{*2}) = 2(n-1)$

Συνεπώς, $\text{var}(S_n^{*2}) = \frac{2(n-1)\sigma^4}{(n-1)^2} = \frac{2\sigma^4}{n-1} \xrightarrow{n \rightarrow \infty} 0$.

Άρα, ο S_n^{*2} είναι συνεπής ευτιμητής της σ^2 .

(δ) Επάρκεια (Sufficiency) Είναι θεμελιώδους σημασίας στη Στατιστική Συμπερασματολογία. Χρησιμοποιήθηκε για πρώτη φορά από τον Fisher και χαρακτηρίζει εκείνες τις στατιστικές συναρτήσεις ή επιτηρητές παραμέτρων οι οποίοι έχουν την ιδιότητα να περιέχουν όλες τις πληροφορίες, οι οποίες περιέχονται στο αρχικό δείγμα σχετικά με τις παραμέτρους. Με άλλα λόγια, η γνώση των τιμών ενός δείγματος δεν μας δίνει τίποτα περισσότερο σχετικά με την επιτήρηση της θ , από ό,τι μας δίνει μία επαρκής στατιστική συνάρτηση.

Η έννοια της επάρκειας χρησιμοποιείται και από τη σκοπιά της "συμπύκνωσης" των τιμών του δείγματος δηλ. των δεδομένων. Συχνά στη Στατιστική είναι επιθυμητή η περίληψη, η σύμπτυξη των δείγματιών δεδομένων χωρίς, αν είναι δυνατόν, να χάσουμε πληροφορία, που περιέχεται στο αρχικό δείγμα, για τον άγνωστο πληθυσμό. Απορρίπτουμε το μέρος των δεδομένων τα οποία δεν περιέχουν πληροφορία για το άγνωστο θ και συνεπώς είναι άχρηστα για την εξαγωγή συμπερασμάτων για το θ . Αυτό επιτυγχάνεται με την έννοια της επάρκειας.

Παράδειγμα Έστω η δομικές τύπου Bernoulli, δηλ. ένα τ.δ. $X_1, X_2, \dots, X_n$ με $X_i \sim B(1, \theta)$, $i=1, \dots, n$.

Έστω $T = X_1 + X_2 + \dots + X_n$, όπου $T \sim B(n, \theta)$.

Το πείραμα επτελείται και έχουμε n τιμές $X_1, \dots, X_n$ με $X_1 + X_2 + \dots + X_n = t$. Θέλουμε να επιτηρήσουμε το θ βάσει των x_i .

Σε ποια περίπτωση γνωρίζουμε "περισσότερα" για το θ ; -14-

(α) Όταν γνωρίζουμε τον αριθμό των επιτυχιών και που έγιναν αυτές; ή

(β) όταν απλά γνωρίζουμε τον αριθμό επιτυχιών;

Με άλλα λόγια, είναι τα $X_1, \dots, X_n$ και T ισοδύναμα από πλευράς "ποσότητας πληροφορίας" που παρέχουν για το θ ;

Αν η πιθανότητα

$P(X_1=x_1, \dots, X_n=x_n | T=t)$ δεν εξαρτάται από το θ , και στις δύο περιπτώσεις (α) και (β) έχουμε την ίδια ποσότητα πληροφορίας για το θ (διότι η σειρά με την οποία εμφανίσθηκαν οι επιτυχίες δεν παίζει κανένα ρόλο και συνεπώς η ουσιαστική ποσότητα είναι ο συνολικός αριθμός t των επιτυχιών). Αν η πιθανότητα εξαρτάται από το θ τότε προφανώς στην περίπτωση (α) έχουμε περισσότερες πληροφορίες για το θ .

Όμως,

$$\begin{aligned} & P(X_1=x_1, X_2=x_2, \dots, X_n=x_n | T=t) \\ &= P(\underline{X}=\underline{x} | T=t) = \frac{P(\underline{X}=\underline{x}, T=t)}{P(T=t)}, \quad \begin{matrix} \underline{X}=(X_1, \dots, X_n) \\ \underline{x}=(x_1, \dots, x_n) \end{matrix} \\ &= \frac{P(\underline{X}=\underline{x})}{P(T=t)} = \frac{\prod_{i=1}^n P(X_i=x_i)}{P(T=t)} = \frac{\theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i}}{\binom{n}{t} \theta^t (1-\theta)^{n-t}} \end{aligned}$$

$$= \frac{\theta^t (1-\theta)^{n-t}}{\binom{n}{t} \theta^t (1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}} \text{ (ανεξάρτητη του } \theta \text{)}.$$

Επομένως, η στατιστική συνάρτηση $T = \sum_{i=1}^n X_i$ περιέχει όλες τις πληροφορίες για την παράμετρο θ που περιέχονται στο αρχικό δείγμα. Συνεπώς, στην περίπτωση (β) έχουμε περισσότερο "συμπυκνωμένες" πληροφορίες για το πείραμα.

Μία σ.σ. T καλείται επαρκής (sufficient) για την παράμετρο θ ενός πληθυσμού αν η σ.σ. T περιέχει όλες τις πληροφορίες στο δείγμα για την παράμετρο θ (ή ισοδύναμα, αν η δεσφρευμένη κατανομή της X που εκφράζει τον πληθυσμό δοθέντος $T=t$ είναι ανεξάρτητη από την θ για όλα τα t).

Το θεώρημα Rao-Blackwell βελτιώνει μία αφερόληπτη ευτιμήτρια $\hat{\theta}$ με την έννοια της μικρότερης διασποράς και με τη βοήθεια μιας επαρκούς στατιστικής συνάρτησης T .

Θεώρημα Rao-Blackwell Αν $\tilde{\theta}$ είναι μία αφερόληπτη ευτιμήτρια της θ και T μία επαρκής σ.σ. για την θ , η ευτιμήτρια $\hat{\theta} \equiv \hat{\theta}(T) = E[\tilde{\theta} | T]$ είναι επίσης αφερόληπτη ευτιμήτρια της θ και επιπλέον ισχύει ότι: $\text{Var}(\hat{\theta}) \leq \text{Var}(\tilde{\theta})$.

Παράδειγμα Έστω $X_1, X_2, \dots, X_n$ τ.δ. από έναν
πληθυσμό με συνάρτηση πιθανότητας από τη διωνυμική
κατανομή $B(1, \theta)$.

Γνωρίζουμε ότι η σ.σ. $U = X_1$ είναι μία αμερόληπτη
σ.σ. του θ και $T = \sum_{i=1}^n X_i$ είναι μία επαρκής σ.σ.
για το θ .

Να βρεθεί ένας άλλος αμερόληπτος ευτιμήτης με
διακύμανση μικρότερη του ευτιμήτη U .

Λύση
Υπολογίζουμε την $E[X_1 | T=t]$

Έχουμε:

$$E[X_1 | T=t] = 0 \cdot P[X_1=0 | T=t] + 1 \cdot P[X_1=1 | T=t]$$

Όπως

$$P[X_1=0 | T=t] = \frac{P(X_1=0, X_1+\dots+X_n=t)}{P_T(t)}$$

$$= \frac{P(X_1=0) P(X_2+\dots+X_n=t)}{P_T(t)}$$

$$= \frac{(1-\theta) \binom{n-1}{t} \theta^t (1-\theta)^{n-1-t}}{\binom{n}{t} \theta^t (1-\theta)^{n-t}} = \frac{n-t}{n}$$

Παρόμοια, $P[X_1=1 | T=t] = \frac{t}{n}$

Άρα $E[X_1 | T=t] = 0 \cdot \frac{n-t}{n} + 1 \cdot \frac{t}{n} = \frac{t}{n}$

Άρα $U^* = \frac{T}{n} = \bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$ είναι αμερόληπτος

επιτητής του θ με $\text{var}(\bar{U}^*) < \text{var}(\bar{U})$ (βάσει του
 Θεωρήματος Rao-Blackwell), και φυσικά είναι
 καλύτερος από τον επιτητή $\bar{U}$.

$$\text{Πράγματι, } \text{var}(\bar{U}) = \text{var}(X_1) = \theta(1-\theta)$$

$$\text{και } \text{var}(\bar{U}^*) = \text{var}(\bar{X}_n) = \frac{\theta(1-\theta)}{n}$$

$$\text{άρα } \text{var}(\bar{U}^*) < \text{var}(\bar{U}).$$

