ΕΚΤΙΜΗΤΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

Θα ασχοληθούμε τώρα με το ιδιαίτερο κεφάλαιο της Στατιστικής Συμπερασματολογίας που αφορά την εκτίμηση μιας ή περισσοτέρων παραμέτρων της κατανομής ενός ποσοτικού χαρακτηριστικού X των στοιχείων του πληθυσμού που μελετάμε.

Έστω f(x;θ) η συνάρτηση πιθανότητας του χαρακτηριστικού X. Το θ είναι η παράμετρος της κατανομής, όπως π.χ. το p της Διωνυμικής, ή το λ της Poisson, η τιμή της οποίας μας είναι άγνωστη. Γενικότερα το $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ μπορεί να είναι μια άγνωστη m-διάστατη παράμετρος, όπως η διδιάστατη $\boldsymbol{\theta} = (\mu, \sigma^2)$ της κανονικής. Το πρόβλημα της εκτιμητικής είναι πώς με βάση ένα τυχαίο δείγμα $X_1,\ldots, X_n$ είναι δυνατόν να προσδιορίσουμε κατά βέλτιστο τρόπο την παράμετρο **θ**. Ο προσδιορισμός της καλείται **εκτίμηση κατά σημείο.**

Θα αναφερθούμε μόνο σε ένα από τα κριτήρια επιλογής εκτιμητριών, αυτό της αμεροληψίας, και εν συνεχεία θα αναπτύξουμε δύο μεθόδους κατασκευής εκτιμητριών.

Ξεκινάμε με κάποιους θεμελιώδεις ορισμούς:

<u>Ορισμοί:</u>

i) Τυχαίο δείγμα μεγέθους n από την σ.π. $f(x;\boldsymbol{\theta})$ θα καλείται ένα σύνολο ανεξάρτητων και ισόνομων τ.μ. $X_1, X_2,\ldots, X_n$ που έχουν σ.π. $f(x;\boldsymbol{\theta})$.

ii) Δειγματοληπτικός χώρος θα καλείται το σύνολο των δυνατών τιμών του δείγματος (π.χ. αν $X_i \in \mathbb{R}$, τότε ο δειγματοληπτικός χώρος θα είναι ο $\mathbb{R}^n$).

iii) Παραμετρικός χώρος (*Θ*) θα καλείται το σύνολο των επιτρεπτών τιμών της παραμέτρου **θ**.

Έστω λοιπόν $X_1, X_2,\ldots, X_n$ ένα τυχαίο δείγμα από την $f(x;\boldsymbol{\theta})$. Η εκτίμηση της παραμέτρου $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_m)$ γίνεται μέσω κάποιων συναρτήσεων των $X_1,X_2,\ldots,X_n$. Πιο συγκεκριμένα θα έχουμε ότι:

iv) Στατιστική (ή δειγματική) συνάρτηση θα λέγεται κάθε συνάρτηση $T = t(\mathbf{X}) = t(X_1,X_2, \ldots,X_n)$ των τ.μ. του δείγματος $X_1, X_2,\ldots, X_n$. Προφανώς, κάθε στατιστική συνάρτηση είναι και αυτή μία τ.μ. Για παράδειγμα, γνωστές στατιστικές συναρτήσεις είναι οι:

- $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (μέση τιμή δείγματος)
- $s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2$ (διασπορά δείγματος)

Τέλος,

v) Εκτιμήτρια συνάρτηση μίας παραμέτρου **θ** θα καλείται μία στατιστική συνάρτηση $T(X_1, X_2,..., X_n)$ η οποία χρησιμοποιείται για την εκτίμηση της **θ**.

Τον υπό μελέτη πληθυσμό μπορούμε να τον συμβολίσουμε με το παρακάτω μαθηματικό μοντέλο:

$$(X, \mathcal{X}, f(x;\boldsymbol{\theta}), \boldsymbol{\theta}= (\theta_1, ..., \theta_m) \in \Theta),$$

όπου X είναι η τυχαία μεταβλητή που συμβολίζει το χαρακτηριστικό των στοιχείων του πληθυσμού που μελετάμε, $\mathcal{X}$ το πεδίο τιμών της, $f(x;\boldsymbol{\theta})$ η σ.π. της τ.μ. X με $\boldsymbol{\theta}= (\theta_1, ..., \theta_m)$ η άγνωστη m-διάστατη παράμετρο με τιμές στον παραμετρικό χώρο $\Theta \subset \mathbb{R}^m$.

Ανάλογα για το τυχαίο δείγμα $\mathbf{X} = (X_1,\dots, X_n)$ από τον παραπάνω ορισμό έχουμε το μαθηματικό μοντέλο:

$$(\mathbf{X}, \mathcal{X}^n, \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta}), \boldsymbol{\theta}= (\theta_1, ..., \theta_m) \in \Theta).$$

<u>Αμεροληψία:</u>

<u>Ορισμός:</u> Αν $T_1 = t(\mathbf{X})$ είναι η εκτιμήτρια της πρώτης, έστω συνιστώσας θ_1 της παραμέτρου $\boldsymbol{\theta} = (\theta_1, ..., \theta_m)$ τέτοια ώστε:

$$E(T_1 \mid \boldsymbol{\theta}) = E\{t(\mathbf{X}) \mid \boldsymbol{\theta}\} = \theta_1 \qquad \forall\ \boldsymbol{\theta} \in \Theta,$$

τότε η εκτιμήτρια T_1 καλείται αμερόληπτη εκτιμήτρια της θ_1.

Αν η εκτιμήτρια T_1 δεν είναι αμερόληπτη τότε η ποσότητα $b(\boldsymbol{\theta}) = E(T_1 \mid \boldsymbol{\theta}) - \theta_1$ ονομάζεται μεροληψία της T_1.

<u>Παράδειγμα:</u> Ας θεωρήσουμε δυο διαφορετικές εκτιμήτριες της διασποράς σ^2 ενός πληθυσμού:

$$T_1 = t_1(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

και

$$T_2 = t_2(\mathbf{X}) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 .$$

Παρατηρούμε ότι:

$$E[\sum_{i=1}^{n}(X_i-\bar{X})^2] = E[\sum_{i=1}^{n}X_i^2-n\bar{X}^2] = \sum_{i=1}^{n}E[X_i^2]-nE[\bar{X}^2]$$

και εύκολα προκύπτει ότι:

$$E[X_i^2] = \sigma^2 + \mu^2, \quad E[\bar{X}^2] = \sigma^2/n + \mu^2,$$

δηλαδή $E[\sum_{i=1}^{n}(X_i-\bar{X})^2] = (n-1)\sigma^2$. Συνεπώς $E[T_1 \mid \sigma^2] = \frac{n-1}{n}\sigma^2$ και $E[T_2 \mid \sigma^2] = \sigma^2$.

Έτσι η T_2 είναι αμερόληπτη εκτιμήτρια της διασποράς σ^2.

Μέθοδοι Εκτίμησης:

Α) Μέθοδος Ροπών:

Έστω $\mathbf{X} = (X_1,\ldots, X_n)$ το τυχαίο δείγμα από τον πληθυσμό $(X, \mathcal{X}, f(x;\boldsymbol{\theta}), \boldsymbol{\theta}= (\theta_1, \ldots, \theta_m) \in \Theta)$.
Λαμβάνουμε τις m πρώτες δειγματικές ροπές περί την αρχή:

$$m_k' = \frac{1}{n}\sum_{i=1}^{n} X_i^k,\ k = 1,..,m$$

και τις εξισώνουμε με τις αντίστοιχες ροπές του πληθυσμού:

$$\mu_k' = E(X^k) = \int x^k f(x;\boldsymbol{\theta})dx = g_k(\boldsymbol{\theta}),\ k = 1,..,m.$$

Τότε έχουμε το σύστημα:

$$g_k(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} X_i^k,$$

η επίλυση του οποίου μας δίνει τις εκτιμήτριες με την μέθοδο των ροπών $\theta_1',\ldots,\theta_m'$.

Παράδειγμα: Έστω τυχαίο δείγμα $\mathbf{X} = (X_1,\ldots, X_n)$ προερχόμενο από κανονικό πληθυσμό $N(\mu, \sigma^2)$. Να βρεθούν με την μέθοδο των ροπών εκτιμήτριες για τα μ, σ^2.

Οι δύο πρώτες δειγματικές ροπές είναι:

$$m_1' = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X} \ \text{και}\ m_2' = \frac{1}{n}\sum_{i=1}^{n} X_i^2 .$$

Οι αντίστοιχες ροπές του πληθυσμού είναι:

$$\mu_1' = E(X) = \mu \quad \text{και} \quad \mu_2' = E(X^2) = \sigma^2 + \mu^2 .$$

Άρα έχουμε το σύστημα:

$$\left\{ \begin{array}{l} \mu = \overline{X} \\ \sigma^2 + \mu^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \end{array} \right. \Rightarrow \left\{ \begin{array}{l} \mu' = \overline{X} \\ (\sigma^2)' = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right) \end{array} \right. .$$

Β) Μέθοδος Μεγίστης Πιθανοφάνειας:

Έστω $\mathbf{X} = (X_1,\ldots, X_n)$ το τυχαίο δείγμα από τον πληθυσμό $(X, \mathcal{X}, f(x;\boldsymbol{\theta}), \boldsymbol{\theta}= (\theta_1, \ldots, \theta_m) \in \Theta)$. Η συνάρτηση πιθανότητας του $\mathbf{X}$ τότε είναι $\prod_{i=1}^{n} f(x_i;\boldsymbol{\theta})$. Η συγκεκριμένη συνάρτηση όταν οι τιμές x_i των X_i (i=1,..,n) είναι γνωστές μπορεί να θεωρηθεί ως συνάρτηση της παραμέτρου **θ**. Έτσι μπορούμε να θέσουμε:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta}),\ \boldsymbol{\theta}\in \Theta.$$

Η $L(\boldsymbol{\theta})$ ονομάζεται **συνάρτηση πιθανοφάνειας** καθότι εκφράζει πόσο πιθανοφανείς, ή διαφορετικά πόσο σύμφωνες με το συγκεκριμένο δείγμα **X**=**x**, είναι οι διάφορες τιμές της παραμέτρου **θ**.

Μεγιστοποιώντας την $L(\boldsymbol{\theta})$ ως προς $\boldsymbol{\theta}\in \Theta$ προκύπτουν οι εκτιμήτριες μέγιστης πιθανοφάνειας (Ε.Μ.Π.) της **θ**, συμβολιζόμενες με $\widehat{\theta_1},....,\widehat{\theta_n}$.

Δηλαδή:

$$\hat{\boldsymbol{\theta}}=(\widehat{\theta_1},....,\widehat{\theta_n})\ :\ L(\hat{\boldsymbol{\theta}}) = \sup L(\boldsymbol{\theta}),\ \ \boldsymbol{\theta}\in\Theta \Leftrightarrow \hat{\boldsymbol{\theta}}=(\widehat{\theta_1},....,\widehat{\theta_n})\ :\ l(\hat{\boldsymbol{\theta}}) = \sup l(\boldsymbol{\theta}),\ \ \boldsymbol{\theta}\in\Theta,$$

όπου $l(\boldsymbol{\theta}) = \log(L(\boldsymbol{\theta}))$, αύξουσα συνάρτηση της $L(\boldsymbol{\theta})$, με log να είναι ο φυσικός λογάριθμος.

Λύνοντας το σύστημα $\dfrac{\partial l(\boldsymbol{\theta})}{\partial\theta_j}=0\ \ \forall$ j=1,.., m προκύπτει η $\hat{\boldsymbol{\theta}}=(\widehat{\theta_1},....,\widehat{\theta_n})$, αφού πρώτα βέβαια βεβαιωθούμε ότι οι αντίστοιχοι πίνακες των δευτέρων μερικών παραγώγων της $l(\boldsymbol{\theta})$ είναι γνήσια αρνητικοί για $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, έτσι ώστε το $\hat{\boldsymbol{\theta}}$ να είναι πράγματι σημείο μεγίστου.

<u>Παράδειγμα:</u> Έστω τυχαίο δείγμα **X** = (X_1,…, X_n) προερχόμενο από πληθυσμό P(λ). Να βρεθεί η Ε.Μ.Π. της λ.

Η συνάρτηση πιθανοφάνειας είναι:

$$L(\lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda}\frac{\lambda^{\sum_{i=1}^{n}x_i}}{\prod_{i=1}^{n}x_i!} \Rightarrow l(\lambda)=-n\lambda+\sum_{i=1}^{n}x_i\log(\lambda)-\sum_{i=1}^{n}\log(x_i!)\,.$$

Οπότε:

$$\frac{\partial l(\lambda)}{\partial\lambda}=0\Rightarrow -n+\frac{1}{\lambda}\sum_{i=1}^{n}X_i=0\Rightarrow\hat{\lambda}=\frac{1}{n}\sum_{i=1}^{n}X_i=\bar{X}\,.$$

<u>**Θεώρημα:**</u> Εάν η απεικόνιση g: $\Theta\rightarrow\Xi$ είναι αμφιμονοσήμαντη και επί τότε η Ε.Μ.Π. της $\boldsymbol{\xi} = g(\boldsymbol{\theta})$ είναι η $\hat{\boldsymbol{\xi}} = g(\hat{\boldsymbol{\theta}})$.