ΣΤΑΤΙΣΤΙΚΗ ΙΙΙ

Θεωρία 01

Δήμητρα Κυριακοπούλου

Τμήμα Οικονομικών Επιστημών

Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Βασική Ορολογία

- Ένα πεπερασμένο σύνολο από i.i.d. τυχαίες μεταβλητές $X_1, X_2, ..., X_n$ λέγεται τυχαίο δείγμα και ο αριθμός n ονομάζεται μέγεθος του δείγματος.
- Το τυχαίο δείγμα είναι μια πολυδιάστατη τυχαία μεταβλητή με συνιστώσες ανεξάρτητες και ισόνομες τυχαίες μεταβλητές (λόγω της ιδιότητας i.i.d.).
- Κάθε μετρήσιμη συνάρτηση που μπορεί να υπολογισθεί από τα διαθέσιμα δεδομένα του δείγματος (χωρίς γνώση της άγνωστης παραμέτρου θ) καλείται στατιστική συνάρτηση, η οποία βοηθά να οριστούν τα διάφορα στατιστικά μεγέθη:
  - για παράδειγμα ο δειγματικός μέσος $\bar{X}$
  - πεδίο ορισμού μιας στατιστικής συνάρτησης είναι ο δειγματοχώρος ενώ πεδίο τιμών είναι ένα υποσύνολο του συνόλου των πραγματικών αριθμών.
- Μια στατιστική που παίρνει τιμές από εναν παραμετρικό χώρο Θ καλείται εκτιμήτρια (συνάρτηση) της άγνωστης παραμέτρου θ.

- Είναι **σημαντικό** να κάνουμε την παρακάτω διευκρίνηση και διαχωρισμό:
  - οι εκτιμήτριες είναι και οι ίδιες τυχαίες μεταβλητές
  - μία εκτίμηση είναι η πραγμάτωση μιας τυχαία μεταβλητής
    - για παράδειγμα, τα δεδομένα ενός δείγματος είναι μια πραγμάτωση (εκτίμηση) των τυχαίων μεταβλητών
    - η τιμή μιας εκτιμήτριας είναι η εκτίμηση της άγνωστης παραμέτρου
    - οι τυχαίες μεταβλητές και η στατιστική του δειγματικού μέσου είναι πραγμάτωση της εκτιμήτριας του μέσου ενός πληθυσμού και ο δειγματικός μέσος θεωρείται απλώς μια εκτίμηση.

- Έχουμε ήδη διαπιστώσει ότι οι κατανομές εξαρτώνται από παραμέτρους οι οποίες στην πράξη είναι άγνωστες και θα πρέπει να προσδιοριστούν, δηλαδή να εκτιμηθούν.
- Αυτό επιτυγχάνεται με την βοήθεια του τυχαίου δείγματος.

Παραδείγματα:

- Στην κατανομή Poisson $P(\lambda)$ με συνάρτηση πιθανότητας:

$$P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 1, 2, ..., \quad \lambda > 0$$

η παράμετρος είναι η λ.

- Στην κανονική κατανομή $N\left(\mu, \sigma^2\right)$ με συνάρτηση πυκνότητας πιθανότητας:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad x \in \mathbb{R}$$

υπάρχουν δύο παράμετροι, μ και σ^2.

Εκτιμήτρια (Estimator)

- Έστω $X_1, X_2, ..., X_n$ τυχαίο δείγμα από πληθυσμό με κατανομή:

$$f(x;\theta),\ \theta \in \Theta \subseteq \mathbb{R},$$

με γνωστή συναρτησιακή μορφή που εξαρτάται από την άγνωστη παράμετρο θ.

- Το παραπάνω είναι γνωστό στη Στατιστική ως παραμετρικό μοντέλο.

- Καλούμε **εκτιμήτρια** της παραμέτρου θ μια κατάλληλη συνάρτηση:

$$\hat{\theta} := \hat{\theta}(X_1, X_2, ..., X_n).$$

- Παράδειγμα 1:

$$\hat{\theta} := \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

αν

$$\theta = E(X) = \mu.$$

- Παράδειγμα 2:

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

αν

$$\theta = E\left(X-\mu\right)^2 = Var\left(X\right) = \sigma^2.$$

- Η εκτιμήτρια $\hat{\theta}$ ως συνάρτηση των τυχαίων μεταβλητών $X_1, X_2, ..., X_n$ είναι και αυτή μια τυχαία μεταβλητή με κάποια κατανομή που έχει μέση τιμή:

$$E\left(\hat{\theta}\right)$$

και διακύμανση:

$$Var\left(\hat{\theta}\right) = E\left[\hat{\theta} - E\left(\hat{\theta}\right)\right]^2.$$

- Το εύλογο ερώτημα είναι πώς αυτά συσχετίζονται με την κατανομή από την οποία προέρχεται το δείγμα, δηλαδή την κατανομή του πληθυσμού.

Ερωτήματα Σχετικά με την Εκτιμήτρια

- Πώς γνωρίζουμε ότι ο δειγματικός μέσος $\bar{X}$ είναι μια καλή εκτιμήτρια του μέσου του πληθυσμού μ; Με άλλα λόγια, υπάρχει κάποιο σύνολο κριτηρίων τα οποία μας επιτρέπουν να αξιολογήσουμε την ποιότητα μεμονομένων εκτιμητριών;
- Χρησιμοποιώντας αυτά τα κριτήρια, μπορούμε να απαντήσουμε ποια είναι η καλύτερη εκτιμήτρια μιας παραμέτρου;
- Δεν θα ήταν προτιμότερο, αντί για μία εκτιμήτρια σημείου - που είναι ένας μόνο αριθμός και δεν μπορούμε ποτέ να συμφωνήσουμε με ακρίβεια ότι είναι η πραγματική τιμή της άγνωστης παραμέτρου - να χρησιμοποιήσουμε εναλλακτικά ένα διάστημα τιμών για το οποίο έχουμε αρκετά καλή πιθανότητα να περιλαμβάνει τη σωστή απάντηση για την άγνωστη παράμετρο;

Αμεροληψία

- Η εκτιμήτρια $\hat{\theta}$ καλείται αμερόληπτη (unbiased) για την παράμετρο θ εάν ισχύει:

$$E\left(\hat{\theta}\right) = \theta.$$

- Επίσης, καλείται ασυμπτωτικά αμερόληπτη, καθώς το μέγεθος του δείγματος n αυξάνεται, εάν:

$$E\left(\hat{\theta}\right) \to \theta, \quad n \to \infty.$$

- Η διαφορά $E\left(\hat{\theta}\right) - \theta$ καλείται μεροληψία (bias).

- Παράδειγμα 1:

$$\begin{aligned} E\left(\bar{X}\right) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu \\ &= \mu \end{aligned}$$

- Επομένως, ο δειγματικός μέσος $\bar{X}$ είναι αμερόληπτη εκτιμήτρια της μέσης τιμής του πληθυσμού μ.

- Επίσης, υπολογίζουμε τη διακύμανση:

$$\begin{aligned} Var\left(\bar{X}\right) &= \frac{1}{n^2}\sum_{i=1}^{n} Var\left(X_i\right) \\ &= \frac{1}{n^2} n\sigma^2 \\ &= \frac{\sigma^2}{n} \end{aligned}$$

και άρα έχουμε ότι:

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$$

- Όπως κάθε τυχαία μεταβλητή, έτσι και οι στατιστικές οι οποίες είναι τυχαίες μεταβλητές ακολουθούν κάποια κατανομή.

- Αν $X_1, X_2, ..., X_n$ είναι τυχαίο δείγμα από κανονική κατανομή $N\left(\mu, \sigma^2\right)$, τότε ο δειγματικός μέσος $\bar{X}$ ακολουθεί την κανονική κατανομή:

$$N\left(\mu, \frac{\sigma^2}{n}\right).$$

- Παράδειγμα 2:

  Να δειχθεί ότι η διασπορά S^2 είναι αμερόληπτη εκτιμήτρια της διακύμανσης του πληθυσμού:

$$E\left(S^2\right) = \sigma^2.$$

  Ξεκινάμε με τον αριθμητή της διασποράς:

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

- Άρα:

$$\begin{aligned} E\left(S^2\right) &= \frac{1}{n-1} E\left[\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] \\ &= \frac{1}{n-1} E\left[\sum_{i=1}^{n}\left(X_i - \mu\right)^2 - n\left(\bar{X} - \mu\right)^2\right] \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n} E\left(X_i - \mu\right)^2 - nE\left(\bar{X} - \mu\right)^2\right] \\ &= \frac{1}{n-1}\left[n\sigma^2 - n\frac{\sigma^2}{n}\right] \\ &= \sigma^2 \end{aligned}$$

- Αν $X_1, X_2, ..., X_n \sim f\left(\mu, \sigma^2\right)$ με σ^2 άγνωστο, τότε:

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu\right)^2$$

- Ισχύει ότι:

$$\begin{aligned} E\left(\hat{\sigma^2}\right) &= \frac{1}{n}\sum_{i=1}^{n} E\left(X_i - \mu\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} Var\left(X_i\right) \\ &= \frac{n\sigma^2}{n} \\ &= \sigma^2 \end{aligned}$$

- Δείξαμε ότι η δειγματική διασπορά που δίνεται από τον τύπο:

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

είναι ένας αμερόληπτος εκτιμητής της θεωρητικής διασποράς οποιασδήποτε κατανομής.

- Να δείξετε ότι η δειγματική διασπορά που δίνεται από τον τύπο:

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

δεν είναι αμερόληπτος εκτιμητής της αντίστοιχης θεωρητικής διασποράς:

$$\begin{aligned} S'^2 &= \frac{n-1}{n}S^2 \Rightarrow \\ E\left(S'^2\right) &= \frac{n-1}{n}E\left(S^2\right) \\ &= \frac{n-1}{n}\sigma^2. \end{aligned}$$

- Στην περίπτωση αυτή, η μεροληψία του εκτιμητή S'^2 είναι ίση με:

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

Αποτελεσματικότητα

- Έστω οι εκτιμήτριες $\hat{\theta}_1$και $\hat{\theta}_2$ της παραμέτρου θ.
- Αν και οι δύο εκτιμήτριες είναι αμερόληπτες, δηλαδή αν ισχύει ότι:

$$E\left(\hat{\theta}_1\right) = E\left(\hat{\theta}_2\right) = \theta,$$

τότε η εκτιμήτρια με τη μικρότερη διακύμανση καλείται σχετικώς αποτελεσματική.

- Παράδειγμα:

Έστω X_1, X_2, X_3 τυχαίο δείγμα από πληθυσμό με κατανομή $P(\lambda)$. Προτείνονται οι εκτιμήτριες:

$$\hat{\lambda_1} = \frac{X_1 + X_2 + X_3}{3}$$
$$\hat{\lambda_2} = \frac{X_1 + 2X_2 + 3X_3}{6}$$

Ποια εκτιμήτρια πρέπει να προτιμηθεί και γιατί;

- Έχουμε ότι:

$$E\left(\hat{\lambda_1}\right) = \lambda = E\left(\hat{\lambda_2}\right)$$

και

$$\begin{aligned} Var\left(\hat{\lambda_1}\right) &= \frac{\lambda}{9} \\ Var\left(\hat{\lambda_2}\right) &= \frac{\lambda + 4\lambda + 9\lambda}{36} \\ &= \frac{7\lambda}{18} \end{aligned}$$

- Συνεπώς:

$$\begin{aligned} \frac{Var\left(\hat{\lambda_1}\right)}{ar\left(\hat{\lambda_2}\right)} &= \frac{\lambda/9}{7\lambda/18} \\ &= \frac{2}{7} < 1 \end{aligned}$$

- Άρα, προτιμάται η εκτιμήτρια $\hat{\lambda_1}$ ως σχετικώς αποτελεσματική.

- Γενίκευση:

$$\hat{\lambda_{1n}} = \bar{X}$$
$$\hat{\lambda_{2n}} = \frac{\sum_{k=1}^{n} kX_k}{\sum_{k=1}^{n} k}$$

- Να αποδειχθεί ότι:

$$Var\left(\hat{\lambda_{1n}}\right) < Var\left(\hat{\lambda_{2n}}\right)$$

για κάθε μέγεθος δείγματος n και να υπολογισθεί το

$$\lim_{n\to\infty} \frac{Var\left(\hat{\lambda_{1n}}\right)}{Var\left(\hat{\lambda_{2n}}\right)}.$$

- Παράδειγμα:

  Έστω μια παρατήρηση X από κατανομή Poisson με παράμετρο λ.

  Να βρεθεί αμερόληπτη εκτιμήτρια της ποσότητας e^{λ}. Πρέπει να ισχύει:

$$E\left(\phi\left(x\right)\right) = \sum_{x=0}^{\infty} \phi\left(x\right) e^{-\lambda} \frac{\lambda^{x}}{x!} = e^{\lambda} \Longrightarrow$$

$$\sum_{x=0}^{\infty} \phi\left(x\right) \frac{\lambda^{x}}{x!} = e^{2\lambda}.$$

Έχουμε ότι:

$$\sum_{x=0}^{\infty} \frac{\phi(x)}{x!}\lambda^x = \sum_{x=0}^{\infty} \frac{(2\lambda)^x}{x!} \Longleftrightarrow$$
$$\sum_{x=0}^{\infty} \frac{\phi(x)}{x!}\lambda^x = \sum_{x=0}^{\infty} \frac{2^x}{x!}\lambda^x \Longleftrightarrow$$
$$\frac{\phi(x)}{x!} = \frac{2^x}{x!} \Longleftrightarrow$$
$$\phi(x) = 2^x$$

Άρα:

$$\begin{aligned} E(\phi(x)) &= e^{-\lambda} \sum_{x=0}^{\infty} 2^x \frac{\lambda^x}{x!} \\ &= e^{-\lambda} e^{2\lambda} \\ &= e^{\lambda}, \quad \forall \lambda. \end{aligned}$$