ΣΤΑΤΙΣΤΙΚΗ ΙΙΙ

Θεωρία 05

Δήμητρα Κυριακοπούλου

Τμήμα Οικονομικών Επιστημών

Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

- Η **έννοια της επάρκειας** έχει εισαχθεί από τον Fisher και παίζει βασικό ρόλο σε διάφορα προβλήματα της Στατιστικής επαγωγής.
- Χαρακτηρίζει εκείνες τις στατιστικές συναρτήσεις που έχουν την ιδιότητα να περιέχουν όλες τις πληροφορίες για την άγνωστη παράμετρο θ.
- Κάποιες φορές το να εξάγουμε συμπεράσματα ή να πάρουμε χρήσιμες πληροφοριές για τον πληθυσμό μέσω των διαθέσιμων δειγμάτων είναι εξαιρετικά δύσκολο
  - μήπως όμως είναι δυνατόν να απλοποιήσουμε (ελαττώσουμε) τα δεδομένα που έχουμε συλλέξει, χωρίς να χάσουμε την πληροφορία που μας ενδιαφέρει;

- Επομένως, η επάρκεια συνεπάγεται ότι γνώση της παρατηρηθείσας τιμής της τυχαίας μεταβλητής που μας ενδιαφέρει δεν προσφέρει τίποτα περισσότερο, όσον αφορά στην εξαγωγή συμπερασμάτων για το θ, από ότι προσφέρει η γνώση μόνο της τιμής μιας επαρκούς στατιστικής συνάρτησης.

- Επιπλέον, μια επαρκής στατιστική συνάρτηση παρουσιάζει το σημαντικό πλεονέκτημα έναντι των παρατηρήσεων ότι συνήθως έχει πολύ μικρότερη διάσταση από τη διάσταση n του τυχαίου δείγματος.

Η Από Κοινού Κατανομή του Τυχαίου Δείγματος

- Έστω τυχαίο δείγμα $(X_1, X_2, ..., X_n)$ από πληθυσμό X με κατανομή $f(x;\theta)$ όπου θ άγνωστη παράμετρος.
- Η από κοινού κατανομή του τυχαίου δείγματος γράφεται ως:

$$f(x_1, x_2, ..., x_n) := f_{X_1, X_2, ..., X_n}(x_1, x_2, ..., x_n).$$

- Αν η κατανομή $f(x;\theta)$ είναι διακριτή, τότε:

$$f(x_1, x_2, ..., x_n) = P(X_1 = x_1, ..., X_n = x_n).$$

- Αν η κατανομή $f(x;\theta)$ είναι συνεχής, τότε η $f(x_1, x_2, ..., x_n)$ δίνει την από κοινού πυκνότητα υπολογισμένη στο σημείο $x_1, x_2, ..., x_n$.
- Επειδή πρόκειται για τυχαία μεταβλητή έχουμε:

$$f(x_1, x_2, ..., x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

Ορισμός: Επαρκές Στατιστικό (sufficient)

- Έστω στατιστική συνάρτηση:

$$T = T\left(x_1, x_2, ..., x_n\right),$$

από την οποία μπορούμε να αντλήσουμε τις πληροφορίες που είναι σχετικές με την παράμετρο θ.

- Αν η δεσμευμένη κατανομή του τυχαίου δείγματος δοθέντος ότι $T = t$, δηλαδή ότι είναι μια σταθερά, είναι **ανεξάρτητη** της παραμέτρου θ για κάθε τιμή t και παραμένει η ίδια για όλα τα μέλη της οικογένειας κατανομής, και η τιμή του T περιέχει όλες τις πληροφορίες που περιέχει το τυχαίο δείγμα, τότε το στατιστικό T καλείται επαρκές για την παράμετρο θ.

- Παράδειγμα:

  Έστω $(X_1, X_2, ..., X_n)$ τυχαίο δείγμα από την κατανομή Poisson $P(\lambda)$ με συνάρτηση πιθανότητας:

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

  Θα δείξουμε ότι η στατιστική συνάρτηση $T = X_1 + X_2 + ... + X_n$ είναι επαρκής για την οικογένεια κατανομών $\{f(x;\lambda) : \lambda > 0\}$ ή ισοδύναμα για την παράμετρο λ.

  Για τη δεσμευμένη κατανομή του τυχαίου δείγματος δοθέντος ότι $T(X) = t$, όπου $t = \sum_{i=1}^{n} X_i$ έχουμε ότι:

$$f\left(x_1, x_2, ..., x_n \mid t = \sum_{i=1}^{n} X_i\right) = \frac{f(x_1, x_2, ..., x_n)}{f_T\left(t = \sum_{i=1}^{n} X_i; \theta\right)},$$

  δηλαδή η δεσμευμένη συνάρτηση πυκνότητας είναι ο λόγος της συνάρτησης πυκνότητας του τυχαίου δείγματος προς την συνάρτηση πυκνότητας της στατιστικής συνάρτησης T.

Όμως η τυχαία μεταβλητή $T = \sum_{i=1}^{n} X_i$ ακολουθεί την κατανομή Poisson με παράμετρο $n\lambda$ και

$$f\left(t;\theta\right) = e^{-n\lambda}\frac{\left(n\lambda\right)^{\sum x_i}}{\left(\sum x_i\right)!}$$

και αντικαθιστώντας στη δεσμευμένη κατανομή έχουμε ότι:

$$\begin{aligned} f\left(x_1, x_2, ..., x_n \mid t = \sum_{i=1}^{n} X_i\right) &= \frac{\prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} \times 1}{e^{-n\lambda}\frac{\left(n\lambda\right)^{\sum x_i}}{\left(\sum x_i\right)!}} \\ &= \frac{e^{-n\lambda}\frac{\lambda^{\sum x_i}}{\prod x_i!}}{e^{-n\lambda}\frac{\left(n\lambda\right)^{\sum x_i}}{\left(\sum x_i\right)!}} \\ &= \left(\frac{1}{n}\right)^{\sum x_i}\frac{\left(\sum x_i\right)!}{\prod x_i!} \end{aligned}$$

- Επομένως, η δεσμευμένη κατανομή του τυχαίου δείγματος δοθέντος ότι $T\left(X\right) = t$ προκύπτει ότι είναι πολυωνυμική με παραμέτρους t και $\left(\frac{1}{n}, \frac{1}{n}, ..., \frac{1}{n}\right)$ δηλαδή είναι ανεξάρτητη της παράμετρου λ.
- Άρα η στατιστική συνάρτηση $T\left(X\right) = \sum_{i=1}^{n} X_i$ είναι επαρκής για το λ.

- Γενικά, η απόδειξη της επάρκειας μιας στατιστικής συνάρτησης από τον ορισμό παρουσιάζει δυσκολίες, ειδικά στις συνεχείς κατανομές, επειδή απαιτείται ο υπολογισμός της δεσμευμένης κατανομής, προκειμένου να διαπιστωθεί η μη εξάρτησή της από την άγνωστη παράμετρο θ.

- Επίσης μια άλλη δυσκολία είναι ότι πριν εφαρμόσουμε τον ορισμό, πρέπει πρώτα να «μαντέψουμε» ποια στατιστική συνάρτηση είναι υποψήφια για επαρκής, κάτι που γενικά δεν είναι εύκολο.

- Οι δύο αυτές δυσκολίες μπορούν να ξεπεραστούν με την εφαρμογή μιας απλής ικανής και αναγκαίας συνθήκης που συνήθως αναφέρεται στην βιβλιογραφία ως παραγοντικό κριτήριο των Neyman-Fisher.

Παραγοντικό Κριτήριο των Neyman-Fisher (factorization theorem)

- Έστω τυχαίο δείγμα $(X_1, X_2, ..., X_n)$ από πληθυσμό X με κατανομή $f(x;\theta)$ όπου θ άγνωστη παράμετρος.
- Κι έστω και η στατιστική συνάρτηση $T = T(x_1, x_2, ..., x_n)$.
- Τότε αυτή καλείται επαρκής για το θ, αν και μόνο αν η από κοινού κατανομή του τυχαίου δείγματος γράφεται (παραγοντοποιείται) στη μορφή:

$$f(x;\theta) = g(T;\theta)\, h(x), \quad \forall\theta$$

όπου:

  - η μη αρνητική συνάρτηση $g(\cdot)$ εξαρτάται από το τυχαίο δείγμα μόνο μέσω της στατιστικής συνάρτησης T,
  - η μη αρνητική συνάρτηση $h(\cdot)$ είναι ανεξάρτητη του θ.

- Απόδειξη:

  Έχουμε ότι:

$$f\left(x_1, x_2, ..., x_n \mid T\right) = h\left(x_1, x_2, ..., x_n\right)$$

  και

$$f\left(x_1, x_2, ..., x_n \mid T\right) = \frac{f\left(x;\theta\right) f\left(T \mid x_1, x_2, ..., x_n\right)}{g\left(T;\theta\right)},$$

  όπου είτε

$$f\left(x_1, x_2, ..., x_n \mid T\right) = \frac{f\left(x;\theta\right)}{g\left(T;\theta\right)} \quad \text{ή} \quad 0$$

  Συνεπώς:

$$f\left(x;\theta\right) = g\left(T;\theta\right) h\left(x_1, x_2, ..., x_n\right).$$

- Παραγοντοποιώντας την κατανομή του τυχαίου δείγματος, προσπαθούμε να ενσωματώσουμε σε μια συνάρτηση $h(\cdot)$ όρους που δεν περιέχουν την άγνωστη παράμετρο θ και εξαρτώνται μόνο από το τυχαίο δείγμα.
- Ότι απομένει μετά την ενσωμάτωση, δηλαδή $\frac{f(x;\theta)}{h(x_1,x_2,...,x_n)}$, είναι ακριβώς ο όρος $g(T;\theta)$ ο οποίος εξαρτάται από το θ και έμμεσα από το τυχαίο δείγμα μέσω κάποιας τιμής, έστω T.
- Αυτή η τιμή ταυτοποιεί τη στατιστική συνάρτηση ως επαρκή.

Επαρκή Στατιστικά

- $X_1, X_2, ..., X_n \sim B(p) \quad \rightarrow \sum_{i=1}^{n} X_i = T$ είναι επαρκές στατιστικό για το p.
- $X_1, X_2, ..., X_n \sim U(0, \theta) \quad \rightarrow \max_{1 \leq i \leq n} X_i = T$ είναι επαρκές στατιστικό για το θ.
- Επί της ουσίας, αυτό που προσπαθούμε να κάνουμε κάθε φορά είναι να παραγοντοποιήσουμε την από κοινού κατανομή του τυχαίου δείγματος, για παράδειγμα:

$$p^{\sum_{i=1}^{n} X_i} (1-p)^{n-\sum_{i=1}^{n} X_i} := p^T (1-p)^{n-T} h(x_1, x_2, ..., x_n),$$

όπου $p^T (1-p)^{n-T} \equiv g(T; p)$ και $h(x_1, x_2, ..., x_n) = 1$.

- Ή, για παράδειγμα (κατανομή Poisson(λ)):

$$e^{-n\lambda} \frac{\lambda^{\sum X_i}}{\prod_{i=1}^{n} X_i} := e^{-n\lambda} \lambda^T \frac{1}{\prod_{i=1}^{n} X_i},$$

όπου $e^{-n\lambda} \lambda^T \equiv g(T; \lambda)$ και $h(x_1, x_2, ..., x_n) = \frac{1}{\prod_{i=1}^{n} X_i}$.

- Ένα άλλο παράδειγμα είναι μέσω της Κανονικής κατανομής (ας θεωρήσουμε ότι $\sigma^2 = 1$):

$$\begin{aligned} f\left(x_1, x_2, ..., x_n; \mu\right) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{\left(x_i-\mu\right)^2}{2}} \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\sum\left(x_i-\mu\right)^2} \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\sum\left(x_i^2-2\mu x_i+\mu^2\right)} \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\sum x_i^2-2\mu\sum x_i+n\mu^2\right)} \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\sum x_i^2} e^{-\frac{1}{2}\left(n\mu^2-2\mu\sum x_i\right)} \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\sum x_i^2} e^{-\frac{1}{2}\left(n\mu^2-2\mu T\right)}, \end{aligned}$$

όπου $e^{-\frac{1}{2}(n\mu^2-2\mu T)} \equiv g\left(T = \sum x_i; \mu\right)$ και $h\left(x_1, x_2, ..., x_n\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\sum x_i^2}$.

- Επομένως, η στατιστική συνάρτηση $T = \sum_{i=1}^{n} X_i$ είναι επαρκής.

- Το ίδιο παράδειγμα με πριν μπορούμε να το δούμε και με τον εξής τρόπο όπως ακολουθεί:

- Έστω $X_1, X_2, ..., X_n \sim N\left(\mu, \sigma^2 = 1\right)$ και ας θεωρήσουμε επίσης ότι

$$T = \sum_{i=1}^{n} X_i \sim N\left(n\mu, n\right).$$

- Τότε:

$$\begin{aligned} f\left(x_1, x_2, ..., x_n \mid T\right) &= \frac{f\left(x_1, x_2, ..., x_n\right)}{f_T\left(t = \sum_{i=1}^{n} X_i; \theta\right)} \\ &= \frac{\left(1/\sqrt{2\pi}\right)^n e^{-1/2\sum\left(x_i-\mu\right)^2}}{\frac{1}{\sqrt{2\pi}} e^{-1/2\frac{(T-n\mu)^2}{n}}} \\ &= \sigma\tau\alpha\theta\varepsilon\rho\acute{\alpha} \times e^{-1/2\left(\sum x_i^2 - \left(\sum x_i\right)^2\right)}, \end{aligned}$$

που σημαίνει ότι η κατανομή του X δοθέντος ότι $T = t$ δεν εξαρτάται από το μ.

- Ας υποθέσουμε ότι έχουμε τυχαίο δείγμα από πληθυσμό με κατανομή την Ομοιόμορφη $U(0,\theta)$ με άγνωστη παράμετρο θ.

- Η από κοινού συνάρτηση πυκνότητας πιθανότητας δίνεται από:

$$f(x_1, x_2, ..., x_n; \theta) = \frac{1}{\theta^n} 0 \leq \max x_i \leq \theta,$$

όπου είναι η δείκτρια συνάρτηση (indicator function) που παίρνει την τιμή 1 εάν το όρισμά της ισχύει, ή την τιμή 0 εάν το ορισμά της δεν ισχύει.

- Από το παραπάνω αποτέλεσμα, προκύπτει ότι:

$$g(T; \theta) = \frac{1}{\theta^n} 0 \leq \max x_i \leq \theta$$

και

$$h(x_1, x_2, ..., x_n) = 1.$$

- Άρα $T = \max\{x_1, x_2, ..., x_n\}$ είναι ένα επαρκές στατιστικό για το θ της Ομοιόμορφης κατανομής.

- Θα μπορούσε ο δειγματικός μέσος να είναι επαρκές στατιστικό για την περίπτωση της Ομοιόμορφης κατανομής;
- Η απάντηση είναι όχι, επειδή μέσω του παραγοντικού κριτηρίου των Neyman-Fisher, ο δειγματικός μέσος θα μπορούσε να είναι επαρκές στατιστικό έαν μπορούμε να γράψουμε τον όρο $0 \leq \max x_i \leq \theta$ ως συνάρτηση του δειγματικού μέσου και του θ.
- Αυτό όμως δεν είναι εφικτό οπότε σε αυτό το πλαίσιο ο δειγματικός μέσος δεν είναι επαρκές στατιστικό.

- Έστω τυχαίο δείγμα $(x_1, x_2, ..., x_n)$ της τυχαίας μεταβλητής X από την κανονική κατανομή $N\left(0, \theta^2\right),\ \theta > 0$.

- Τότε, εφαρμόζοντας το παραγοντικό κριτήριο είναι εύκολο να δειχθεί ότι οι στατιστικές συναρτήσεις:

$$T_1 = (x_1, x_2, ..., x_n),$$
$$T_2 = \left(x_1^2, x_2^2, ..., x_n^2\right),$$
$$T_3 = \sum_{i=1}^{n} X_i^2$$

είναι επαρκείς.

- Παρατηρούμε ότι κάθε μία από τις στατιστικές αποτελεί συνάρτηση όλων των προηγούμενων.

- Η T_3 που είναι συνάρτηση όλων των άλλων, αποτελεί τη μεγαλύτερη δυνατή σύμπτυξη του τυχαίου δείγματος χωρίς απώλεια πληροφορίας για το θ^2.

- Γι' αυτόν το λόγο, η T_3 αναφέρεται ως **ελάχιστη επαρκής (minimal sufficient)** στατιστική συνάρτηση. Ελάχιστη διότι φέρει τη μικρότερη διάσταση.
- Γενικά, μια επαρκής στατιστική συνάρτηση λέγεται ελάχιστη επαρκής εάν είναι συνάρτηση οποιασδήποτε άλλης επαρκούς στατιστικής συνάρτησης.
- Ένας γραμμικός μετασχηματισμός ελάχιστης επαρκούς στατιστικής συνάρτησης είναι επίσης ελάχιστη επαρκής.
- Επίσης, αν δύο επαρκείς στατιστικές συναρτήσεις είναι ελάχιστες, τότε κάθε μία είναι συνάρτηση της άλλης.