

Εκτίμηση

Βασικός σκοπός και χρησιμότητα της Στατιστικής: **εκτίμηση** χαρακτηριστικών ενός πληθυσμού βάσει πληροφορίας από ένα δείγμα.

Κύρια σημεία:

- εκτίμηση μ , σ και κατανομής πιθανότητας P
- **διαστήματα εμπιστοσύνης** για P
- διαστήματα εμπιστοσύνης για μ

Ως βάση της εκτίμησης χρησιμοποιούμε τα *στατιστικά*. Αυτά δεν ταυτίζονται με τις παραμέτρους αλλά είναι οι καλύτερες *εκτιμήσεις* των παραμέτρων που μπορούμε να κάνουμε.

Υπενθύμιση αντιστοιχιών: $\bar{x} \leftarrow \rightarrow \mu$, $s \leftarrow \rightarrow \sigma$

Σημαντική παρατήρηση για το στατιστικό της τυπικής απόκλισης, s : αποδεικνύεται ότι αν χρησιμοποιήσουμε τον τύπο για την τυπική απόκλιση “ως έχει” για *μικρά δείγματα*, θα τείνει να υποεκτιμήσει το σ σε σχέση με την πραγματική τιμή του. Αντί γι' αυτόν χρησιμοποιούμε μια τροποποιημένη μορφή όπου ο παρονομαστής έχει $n-1$ αντί για n . Όταν το δείγμα είναι πολύ μεγάλο, τότε η διαφορά τείνει να εξαλειφθεί.

Διαστήματα Εμπιστοσύνης

Το πρόβλημα είναι ότι δε μπορούμε ποτέ να πούμε ότι μια τιμή που προβλέπουμε για τα χαρακτηριστικά ενός πληθυσμού είναι απόλυτα σωστή και βέβαιη. Αλλά τουλάχιστον, μπορούμε να εκφράσουμε ποσοτικά την αβεβαιότητα που χαρακτηρίζει την εκτίμησή μας.

Π.χ. για τη μέση τιμή, αυτό γίνεται λέγοντας ότι είναι “από A μέχρι B με $X\%$ εμπιστοσύνη”

Όσο πιο μεγάλο το X (πιο κοντά στο 100%) τόσο πιο μεγάλο το διάστημα (A,B) . Όσο πιο στενό το διάστημα γύρω από μια προβλεπόμενη τιμή, τόσο πιο μικρό το X .

Τα A , B προκύπτουν προσθέτοντας και αφαιρώντας από την προβλεπόμενη τιμή, μια τιμή που δείχνει το εύρος της αβεβαιότητάς μας.

Τα συνήθη ποσοστά εμπιστοσύνης είναι 95%, 98, 99, 99.5%

Παράδειγμα 1:

Μηχανές A και B παράγουν κυλινδρικές μεταλλικές ράβδους με συγκεκριμένο επιθυμητό μήκος.

Τυπική Απόκλιση $A = 0.1$ εκ

Τυπική Απόκλιση $B = 1$ εκ.

Κάνουμε δειγματοληψία ανά 36 και σχηματίζουμε την κατανομή δειγματοληψίας.

Π.χ. για τη μηχανή A

$$\mu_{\bar{x}} = \mu = 1, \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.1}{\sqrt{36}} = 0.0167$$

Από πίνακες της τυπικής κανονικής κατανομής ή συναρτήσεις λογιστικών φύλλων μπορούμε να δούμε ότι, βάσει ΚΟΘ, η πιθανότητα να είναι το $\bar{x}$ μεγαλύτερο ή μικρότερο από 1 κατά $\sigma = 0.0167$ θα είναι $0.3413 + 0.3413 = 0.6826$

Βρίσκουμε $z = (x - \mu)/\sigma$, δηλαδή με κέντρο το μ , ανάγουμε το x σε μονάδες σ .

Σημαντικό: το παραπάνω διάστημα εμπιστοσύνης ισχύει *ακόμη και αν το μ είναι άγνωστο*. Θα βρούμε μία μέση τιμή μ' από το δείγμα και η πιθανότητα να είναι στα όρια $\mu' - 0.0167$ και $\mu' + 0.0167$ είναι 68%.

Παράδειγμα 2:

Να βρεθεί η μέση τιμή των παραγόμενων ράβδων με 95% εμπιστοσύνη, αν η μέση τιμή του δείγματος είναι 0.5 μέτρα.

Ακολουθούμε τα εξής βήματα:

- διαιρούμε την επιθυμητή εμπιστοσύνη διά 2 ώστε να δούμε το μέρος του εμβαδού που είναι κάτω από την καμπύλη στο ένα μισό. Η καμπύλη είναι συμμετρική. Βρίσκουμε 47.5%
- Βρίσκουμε από πίνακες ή συναρτήσεις λογιστικών φύλλων τη z_0 τιμή της τυπικής κανονικής κατανομής που αντιστοιχεί σε αυτό.
- υπολογίζουμε το εύρος γύρω από τη μέση τιμή ως z_0 επί τυπική κατανομή δειγματοληψίας (βλ. ΚΟΘ)
- βρίσκουμε το άνω και κάτω όριο προσθαφαιρώντας αυτό που βρήκαμε από τη μέση τιμή

Συγκεκριμένα:

$$95\%/2 = 47,5\% = 0,475$$

$$z_0(0,475) = 1,96 \text{ (από πίνακες)}$$

Για την κατανομή δειγματοληψίας βρήκαμε προηγουμένως ότι $\sigma=0.0167$ επομένως

$$\bar{x} = 0.5 \pm 1.96 \times 0.0167 = 0.5 \pm 0.033$$

Αν θέλουμε να αυξήσουμε την εμπιστοσύνη στο 99% θα βρούμε

$$99/2 = 49,5 = 0.495$$

$$z_0(0,495) = 2,58 \text{ (από πίνακες)}$$

$$\bar{x} = 0.5 \pm 2.58 \times 0.0167 = 0.5 \pm 0.043$$

Βλέπουμε ότι, όπως θα θεωρούσαμε λογικό, όσο μεγαλύτερο το εύρος γύρω από τη μέση τιμή, τόσο μεγαλύτερη η εμπιστοσύνη (αλλά μικρότερη η ακρίβεια).

Γενικά, λοιπόν, το διάστημα εμπιστοσύνης δίνεται από

$$\mu_{\bar{x}} = \bar{x} \pm z_0 \left(\frac{\sigma}{\sqrt{n}} \right)$$

Άγνωστο σ πληθυσμού

Αν δε γνωρίζω την τυπική απόκλιση του πληθυσμού, θα βασιστώ στη στατιστική του δείγματος.

Παίρνω τη “διορθωμένη” σχέση για την τυπική απόκλιση (με το $n-1$ για να αντισταθμίσω την υποεκτίμηση του σ για μικρά δείγματα) του δείγματος

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

Τότε η διαδικασία είναι ίδια όπως πριν, μόνο που γράφουμε s αντί για σ :

$$\mu_{\bar{x}} = \bar{x} \pm z_0 \left(\frac{s}{\sqrt{n}} \right)$$

Επίπτωση μεγέθους δείγματος:

- Η διαδικασία που περιγράψαμε προϋποθέτει αρκετά μεγάλα δείγματα ($n > 30$) για να έχει εφαρμογή το ΚΟΘ
- Αν το σ είναι άγνωστο, χρησιμοποιούμε το s ως εκτίμηση
- Αν το δείγμα είναι μικρό, τότε
 - αν γνωρίζω ή περιμένω, θεωρώ λογικό, ότι ακολουθεί κανονική κατανομή, χρησιμοποιώ **κατανομή του t κατά Student** (βλ. σχετικά, παρακάτω) και s για την τυπική απόκλιση.
 - αλλιώς, δε μπορώ να κάνω ασφαλή εκτίμηση

Για την τελευταία περίπτωση

$$\mu_{\bar{x}} = \bar{x} \pm t_0 \left(\frac{s}{\sqrt{n}} \right)$$

Πίνακες κατανομής Student

Η κατανομή του t μοιάζει με την κανονική κατανομή, είναι και αυτή συμμετρική και για μεγάλα δείγματα τείνει να ταυτιστεί με την κανονική.

Στους πίνακες δίνονται:

- αριστερή στήλη: **βαθμοί ελευθερίας** = $n - 1$, π.χ. για δείγμα των 15, $df = 14$
- πρώτη οριζόντια γραμμή (επικεφαλίδες): πιθανότητα της τυπικής ή ανηγμένης μεταβλητής t

(όπως το z για την τυπική κανονική κατανομή), να *υπερβεί* την τιμή που αναγράφεται στο πεδίο για δεδομένο αριθμό βαθμών ελευθερίας.

Παράδειγμα 3:

Έχουμε δείγμα με $n = 9$ και θέλουμε να βρούμε τη μέση τιμή με διάστημα εμπιστοσύνης 95%.

Βρίσκουμε $df = n - 1 = 8$

Επειδή αναζητούμε πιθανότητες να υπερβεί κάποια τιμή, είτε από δεξιά είτε από αριστερά, αφαιρούμε $100 - 95 = 5\%$ και μετά διαιρούμε διά 2, το οποίο δίνει $2.5\% = 0.025$

Δηλαδή, η πιθανότητα να βγει παραπάνω από το άνω όριο εμπιστοσύνης θα είναι 2.5% και άλλο τόσο να βγει λιγότερο από το κάτω όριο.

Άρα, πάμε στη στήλη 0.025 και αναζητούμε τη γραμμή 8 και στην τομή τους βρίσκουμε την τιμή 2.306 που σημαίνει ότι $t > 2.306$ με πιθανότητα 2.5% και $t < -2.306$ με πιθανότητα επίσης 2.5%.

Άρα, το διάστημα εμπιστοσύνης θα είναι

$$(\bar{x} - 2.306s/\sqrt{9}, \bar{x} + 2.306s/\sqrt{9}) = (\bar{x} - 0.769s, \bar{x} + 0.769s) \quad \text{με πιθανότητα } 95\%.$$

Παρατήρηση: η κατανομή του t κατά Student μπορεί να χρησιμοποιηθεί και για $n > 30$ δίνοντας *ακριβέστερα* αποτελέσματα από την κανονική όταν το σ είναι άγνωστο και καταφεύγουμε στο s . Αλλά αυτή και η κανονική κατανομή τείνουν να συμπίπτουν με αυξανόμενο μέγεθος δείγματος ώστε η διαφορά να είναι ασήμαντη.

Παράδειγμα 3:

Θέλουμε να βρούμε διάστημα εμπιστοσύνης 98% για τη μέση τιμή μιας κατανομής:

α) διτροπική κατανομή πληθυσμού με γνωστή $\sigma = 7.85$ από δείγμα $n = 50 > 30$.

β) κατανομή πληθυσμού γνωστή ως κανονική, άγνωστο σ και δείγμα των $15 < 30$.

γ) μονοτροπικό αλλά έντονα ασύμμετρο δείγμα και κατανομή πληθυσμού γνωστή ως μη κανονική, άγνωστο σ και δείγμα $36 > 30$.

δ) Κανονική κατανομή πληθυσμού με γνωστό $\sigma = 8$ και δείγμα $25 < 30$.

ε) Κανονική κατανομή πληθυσμού με άγνωστο σ και δείγμα $45 > 30$.

στ) Μη κανονική κατανομή, διτροπικό δείγμα $24 < 30$.

Απαντήσεις:

α) γνωστή σ και $n > 30$ συνεπάγονται χρήση τυπικής κανονικής κατανομής (z)

β) δείγμα μικρότερο του 30 και άγνωστο σ συνεπάγονται χρήση κατανομής Student και s αντί για σ

γ) δείγμα μεγαλύτερο του 30 αλλά άγνωστο σ συνεπάγεται χρήση του z αλλά με s αντί για σ .

δ) δείγμα μικρότερο του 30 συνεπάγεται χρήση του t (Student) με s παρά το γνωστό σ .

ε) δείγμα άνω των 30 αλλά άγνωστο σ συνεπάγεται χρήση z ή χρήση t με s αντί για σ , τόσο με το z όσο και με το t .

στ) δείγμα μικρότερο του 30 για μη κανονική κατανομή δεν επιτρέπει χρήση z ή t για διάστημα εμπιστοσύνης μέσης τιμής.

Εμπιστοσύνη αναλογίας πληθυσμού – διωνυμική κατανομή

Χρησιμοποιούμε μια παρόμοια με τις προηγούμενες, σχέση:

$$P = p \pm z_0 \frac{\sqrt{pq}}{\sqrt{n}}$$

Παράδειγμα 4

Εμβολιάζονται 200 άνθρωποι για το νέο της γρίππης και οι 185 αποκτούν ανοσία, άρα $p = 0.925$.

Θέλουμε να ορίσουμε ένα διάστημα εμπιστοσύνης 99% για το παραπάνω αποτέλεσμα. Το μικρότερο υποσύνολο πρέπει να έχει τουλάχιστον 10 παρατηρήσεις. Εδώ έχουμε $15 > 10 = \text{OK}$.

Εκτιμούμε τις πιθανότητες βάσει του δείγματος: $p = 0.925$, $q = 1 - p = 0.075$.

Βρίσκουμε το z_0 όπως κάναμε πιο πάνω και είναι ίσο με 2.58.

Βρίσκουμε διάστημα εμπιστοσύνης 0.925 ± 0.048 .