2ο online μάθημα / 1|4|2020

ΚΕΝΤΡΙΚΟ ΟΡΙΑΚΟ ΘΕΩΡΗΜΑ (Κ.Ο.Θ)

Το Κ.Ο.Θ. μας δίνει την κατά προσέγγιση κατανομή του αθροίσματος πολλών όρων ανεξάρτητων και ισόνομων τ.μ.
Αποδεικνύονται (δεν είναι στους σκοπούς αυτού του μαθήματος τα ακόλουθα):

Κ.Ο.Θ.:
Έστω $X_1, X_2, \dots, X_n$ n το πλήθος ανεξάρτητες και ισόνομες τ.μ. με $EX_i = \mu < +\infty$ και $Var X_i = \sigma^2 < +\infty$. Τότε αν n μεγάλο

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\,\sigma} \overset{\text{προσεγγιστικά}}{\sim} N(0,1)$$

ή ισοδύναμα $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \overset{\text{προσεγγιστικά}}{\sim} N(0,1)$

Ποια είναι η αξία του θεωρήματος;
Ακόμα κι αν η κατανομή του πληθυσμού δεν είναι κανονική η δειγματική μέση τιμή ή το άθροισμα n το πλήθος ανεξάρτητων τ.μ. από αυτόν προσεγγίζεται από την κανονική.

Τι πλεονεκτήματα μας δίνει αυτό;
Μπορούμε να χρησιμοποιήσουμε την θεωρία που στηρίζεται στην κανονική κατανομή αρκεί το n μεγάλο.

Πόσο μεγάλο;;; Αυτό χρήζει μεγάλης συζήτησης παρότι στη βιβλιογραφία λέμε $n \geq 30$!!!

Επομένως το Κ.Ο.Θ. δικαιολογεί την προσέγγιση της κατανομής του $\bar{X}$ με την $N(\mu, \sigma^2/n)$ όταν n μεγάλο. Έτσι φαίνεται για μια ακόμη φορά η σπουδαιότητα της Normal. Επίσης μας βοηθά στην προσέγγιση του $\sum_{i=1}^{n} X_i$ όπου π.χ. X_i είναι σφάλμα στην i-οστή μέτρηση, χρόνος ζωής i-οστού ασθενή, μετατόπιση i-οστού σωματιδίου

Προσέγγιση Διωνυμικής $B(n,p)$

οποιοδήποτε πείραμα που έχει 2 δυνατά αποτελέσματα Επιτυχία, Αποτυχία.

Έστω $X \sim B(n,p)$. Τότε η τ.μ. παριστάνει τον αριθμό των επιτυχιών σε n ανεξάρτητες δοκιμές Bernoulli με πιθανότητα επιτυχίας p. Τότε (το αποτέλεσμα της μιας δοκιμής δεν επηρεάζει το αποτέλεσμα της άλλης δοκιμής) παραμένει σταθερή

$X = X_1 + \dots + X_n$ όπου $X_i = \begin{cases} 1, & \text{αν επιτυχία στην i-οστή δοκιμή} \\ 0, & \text{αν αποτυχία στην i-οστή δοκιμή} \end{cases}$

Γιατί θέλησα να γράψω την $X = \sum_{i=1}^{n} X_i$??
Να για να μπορώ να εφαρμόσω Κ.Ο.Θ. που είναι για άθροισμα ή δειγματική μέση τιμή

Είναι τώρα X_i τ.μ. με

$EX_i \xlongequal{\text{ορισμός}} 1 \cdot P(X_i = 1) = 1 \cdot P(\text{επιτυχία στην i-οστή δοκιμή})$

δηλαδή $EX_i = P$

$$\mu = \begin{cases} \sum x f(X=x) & \text{, διακριτή } X \\ \int x f_X(x)dx & \text{, συνεχής } X \end{cases}$$

Επίσης $Var\, X_i = EX_i^2 - (EX_i)^2$

με $EX_i^2 = 1^2 \cdot P(X_i = 1) = p$

Άρα $Var\, X_i = p(1-p)$

Παρατήρηση: Κάποιος μπορούσε να αναγνωρίσει ότι $X_i \sim B(1,p)$

Άρα $EX_i = \mu = p$, $Var\, X_i = p(1-p) = \sigma^2$

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} \overset{\text{προσεγγιστικά}}{\sim} N(0,1)$$

$\sigma^2 \to p(1-p)$

ή

$$\frac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}} \overset{\text{προσεγγιστικά}}{\sim} N(0,1)$$

$P(X=4)$ υπάρχει

$P(\text{συνεχής} = x) = 0$ συνεχής

$B(n,p)$ διακριτή

Προσοχή: Προσεγγίζουμε μια διακριτή κατανομή (την Διωνυμική) από μια συνεχή (την κανονική). Αυτό τι πρόβλημα δημιουργεί;
Σκεφτείτε ότι η πιθανότητα μιας συνεχής τ.μ. να πάρει συγκεκριμένη τιμή είναι μηδέν. Δηλαδή αν είχαμε τον υπολογισμό της

$$P\left(a \le \sum_{i=1}^{n} X_i \le b\right) \quad P\left(\sum X_i \le b\right)$$

τότε κατά την προσέγγιση δεν θα "παίρναμε" πιθανότητα στα a, b.
Για το λόγο αυτό έχει προταθεί η λεγόμενη ΔΙΟΡΘΩΣΗ ΣΥΝΕΧΕΙΑΣ.
Είναι μια διόρθωση έτσι ώστε να συμπεριλαμβάνονται οι ισότητες.
Έτσι για να βρω π.χ. $P\left(a \le \sum_{i=1}^{n} X_i \le b\right)$

$$= P\left(a - 0.5 \le \sum_{i=1}^{n} X_i \le b + 0.5\right)$$

Εφαρμόζεται μόνο όταν τα X_i είναι διακριτές ή $P(\Sigma x_i \leq \beta) = P(\Sigma x_i \leq \beta + 0{,}5)$

Παρατήρηση: Η προσέγγιση της $B(n,p)$ από την $N(0,1)$ είναι ικανοποιητική όσο το n μεγαλύτερο και p πλησιέστερο στο $1/2$.

Εμπειρικά

$$0 \leq np - 2\sqrt{npq} < np + 2\sqrt{npq} < n.$$

1/4/2020 / 2ο online μάθημα ①

Στατιστική συμπερασματολογία: είναι ο κλάδος της στατιστικής που ασχολείται με την εξαγωγή συμπερασμάτων από το Τυχαίο δείγμα για τον πληθυσμό. Σ' αυτό το πλαίσιο, ο στατιστικός συλλέγει δεδομένα με πειράματα ή δειγματοληψίες & επιδιώκει να βγάλει συμπεράσματα.

Η στατιστική συμπερασματολογία χωρίζεται σε 2 κλάδους

ΕΚΤΙΜΗΤΙΚΗ
- σε σημείο ή σημειοεκτιμητική
- σε διάστημα

ΕΛΕΓΧΟΙ ΥΠΟΘΕΣΕΩΝ
- μεθοδολογίες με τις οποίες θέλω να ελέγξω στατιστικά κάποια πράγματα.

Εκτίμηση παραμέτρων

Έστω $X_1, X_2, \dots X_n$ ένα τ.δ. (δηλ. οι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές από έναν πληθυσμό με συνάρτηση πυκνότητας πιθανότητας (σ.π.π.) ή συνάρτηση πιθανότητας (σ.π) $f(x, \theta)$ όπου θ άγνωστη/άγνωστες παράμετροι.
$\theta \in \Theta$, με Θ ο παραμετρικός χώρος

Τι σημαίνουν τα προηγούμενα;

Εκθ, Poisson

Σημαίνουν ότι μας είναι γνωστή η συναρτησιακή μορφή της σ.π.π. ή της σ.π. αλλά μας είναι άγνωστη η τιμή

θ λ $B(n,p)$

μιας ποσότητας-παραμέτρου που εμπλέκεται σ' αυτήν.
Για παράδειγμα ξέρουμε ότι ο χρόνος ζωής ακολουθεί Εκθ(λ) δηλ. $f_X(x;\lambda) = \lambda e^{-\lambda x}$, $x \geq 0$ αλλά το $\lambda \in (0, +\infty)$ μας είναι άγνωστο.

②

Εδώ $\vartheta = \lambda$ $\Theta = (0, +\infty)$

• Όταν $X \sim B(n, p)$, τότε $\vartheta = p$ και $\Theta = (0,1)$

Είναι προφανές ότι όταν έχουμε ως στόχο την εκτίμηση των αγνώστων παραμέτρων αυτή πρέπει να στηρίζεται στο δείγμα και πιο συγκεκριμένα σε στατιστικές συναρτήσεις (συναρτήσεις των $X_1 \ldots X_n$ και γνωστών ποσοτήτων)

Αυτές οι σ.σ. ονομάζονται <u>εκτιμήτριες συν/σεις</u> και οι τιμές (οι αριθμητικές) που προκύπτουν από την εφαρμογή τους στο δείγμα ονομάζονται εκτίμηση-εκτιμητής της άγνωστης παραμέτρου θ.

Ένας εκτιμητής συμβολίζεται $\hat{\vartheta} = \hat{\vartheta}(X_1 \ldots X_n)$

<u>Εκτίμηση σε σημείο</u> $\tilde{\vartheta}$, $\hat{\hat{\vartheta}}$

Υπάρχουν πολλοί εκτιμητές σε σημείο για μια συγκεκριμένη ποσότητα

<u>Παράδειγμα</u>

$X_1 \ldots X_n$ τ.δ. από έναν πληθυσμό που περιγράφει τον δείκτη νοημοσύνης των ανδρών ηλικίας 20-29

Γνωρίζουμε ότι ο δείκτης νοημοσύνης ακολουθεί $N(\mu, \sigma^2)$, μ, σ^2 άγνωστα. Ποιος θα μπορούσε να 'ναι ένας εκτιμητής του μ;

Μια απάντηση είναι ότι ένας εκτιμητής της άγνωστης πληθυσμιακής μέσης τιμής δύναται να είναι η δειγματική μέση τιμή ($\bar{x}$) ή δειγματική διάμεσος, ενώ θα μπορούσαν να προταθούν και άλλοι εκτιμητές, π.χ $\frac{X_{min} + X_{max}}{2}$

Αυτό οδηγεί στην εισαγωγή κριτηρίων για την επιλογή του "καλύτερου" εκτιμητή με βάση αυτό το θεμελιώδες κριτήριο. Τέτοια κριτήρια υπάρχουν πολλά.

→ <u>Κριτήριο αμεροληψίας</u>

Ένας εκτιμητής $\hat{\theta}$ λέμε ότι είναι <u>αμερόληπτος</u> εκτιμητής της παραμέτρου θ αν $E(\hat{\theta})=\theta \quad \forall \theta \in \Theta$

Η έκφραση το κριτήριο αυτό θα συζητηθεί. $\hat{\theta}=\hat{\theta}(X_1,\dots,X_n)$

- Το ερώτημα που προκύπτει τώρα είναι αν το κριτήριο αυτό βοηθά στην επιλογή εκτιμητή; Ένας εκτιμητής είναι

Η απάντηση είναι όχι. Γιατί; $\hat{P}=\frac{\text{πλήθος επιτυχιών}}{\text{πλήθος δοκιμών}}$

Η απάντηση είναι όχι και η αιτιολόγηση είναι η εξής:

<u>Πρόταση</u> (εκτός ύλης, θα την δείτε στο 6ο εξάμηνο)

"Το σύνολο των αμερόληπτων εκτιμητών μιας θ είναι είτε κενό (∅) είτε μόνο σύνολο είτε περιέχει άπειρα στοιχεία!!!

Επίσης αν αιτιολογήσουμε το όχι με βάση το παράδειγμα μας είναι $X_1, X_2 \dots X_n$ τδ $N(\mu,\sigma^2)$

τότε $T_1=\frac{X_1+X_2}{2}$

$$T_2=\frac{X_1+X_2+X_3+X_4}{4}$$

$$T_3=\frac{\sum_{i=1}^{n} X_i}{n}=\bar{X}$$

Εύκολα προκύπτει: $ET_1=\frac{1}{2}(\overset{\mu}{EX_1}+\overset{\mu}{EX_2})=\mu$

$ET_2=\mu$

$ET_3=\mu$

Αυτό συμβαίνει ότι το κριτήριο της αμεροληψίας από μόνο του δεν αρκεί για τον καθορισμό του καλύτερου "εκτιμητή"

Λύση: Επιλογή του αμερόληπτου εκτιμητή που ελαχιστοποιεί (που έχει την μικρότερη) διακύμανση.

Δηλ. από τους αμερόληπτους να βρω εκείνον με την μικρότερη διακύμανση. Λογικό κριτήριο;

A.O.E.Δ. Ένας εκτιμητής λέγεται Αμερόληπτος Ομοιόμορφα Ελάχιστος Διακύμανσης όταν είναι αμερόληπτος και έχει τη μικρότερη διακύμανση από κάθε άλλο αμερόληπτο εκτιμητή.

- Πως βρίσκουμε A.O.E.Δ.? Δεν είναι στους σκοπούς του μαθήματος μας αλλά θα το δείτε στο 6º εξάμηνο
- Υπάρχουν και άλλα κριτήρια? ΝΑΙ → βλέπε 6º εξάμηνο

Πρόταση

- Ο δειγματικός μέσος $\overline{X}$ είναι πάντοτε αμερόληπτος εκτιμητής της πληθυσμιακής μέσης τιμής

  Θυμηθείτε: $E\overline{X} = \mu$ (έχει γίνει η απόδειξη)
- Η δειγματική διακύμανση είναι πάντοτε αμερόληπτη εκτιμήτρια της πληθυσμιακής διακύμανσης

  Θυμηθείτε: $ES^2 = \sigma^2$ (έχει γίνει η απόδειξη)
- Αν επιπλέον ο πληθυσμός είναι κανονικός τότε $\overline{X}$ είναι A.O.E.Δ. της μ (χωρίς απόδειξη)

$$\operatorname{Var} \overline{X} = \frac{\sigma^2}{n}$$