ΚΕΦΑΛΑΙΟ 15

ΔΙΑΣΤΗΜΑΤΑ ΕΜΠΙΣΤΟΣΥΝΗΣ ΓΙΑ ΑΝΑΛΟΓΙΕΣ

Α. Περίπτωση Ενός Πληθυσμού

Έστω ότι μελετάμε μια ακολουθία n ανεξαρτήτων δοκιμών κάθε μία από τις οποίες οδηγεί είτε σε επιτυχία είτε σε αποτυχία με σταθερή πιθανότητα επιτυχίας p.

Τότε, όπως γνωρίζουμε, ο αριθμός των επιτυχιών ακολουθεί την διωνυμική κατανομή με παραμέτρους n και p ($X \sim b(n,p)$).

Έστω ότι επιθυμούμε να εκτιμήσουμε την πραγματική αναλογία των επιτυχιών p στον πληθυσμό.

Μια σημειακή εκτιμήτρια του p είναι βέβαια η

$$\hat{p} = \frac{X}{n} \qquad \left(E(\hat{p}) = p\right)$$

δηλαδή η δειγματική αναλογία επιτυχιών.

Έστω ότι θέλουμε να κατασκευάσουμε ένα 100(1-α)% διάστημα εμπιστοσύνης για το p.

Για να το κάνουμε αυτό θα πρέπει, φυσικά, να καθορίσουμε την κατανομή του $\hat{p}$.

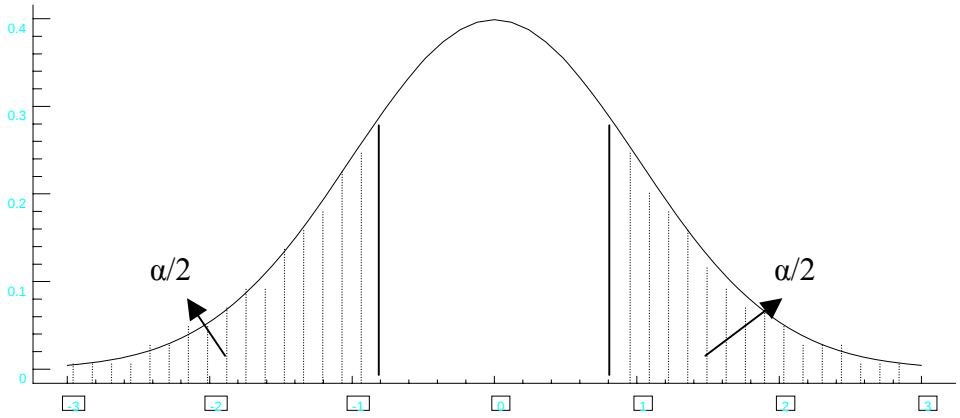
Ένα γνωστό αποτέλεσμα από τις πιθανότητες είναι ότι, για μεγάλο n, η X ακολουθεί κατά προσέγγιση την κανονική κατανομή με μέση τιμή np και διασπορά npq ($X \underset{appr.}{\sim} N(np,npq)$).

Επομένως,

$$\frac{X - np}{\sqrt{npq}} = \frac{\frac{X}{n} - p}{\sqrt{\frac{pq}{n}}} \underset{appr.}{\sim} N(0, 1)$$

Επομένως έχουμε ότι,

$$P\left(-Z_{1-\alpha/2} \le \frac{\frac{X}{n} - p}{\sqrt{\frac{pq}{n}}} \le Z_{1-\alpha/2} \right) = 1\text{-}\alpha$$



$-Z_{1-\alpha/2}$ $Z_{1-\alpha/2}$

Η ανισότητα που εμφανίζεται στην παρένθεση της παραπάνω σχέσης είναι ισοδύναμη με την ανισότητα

$$K(p) = \left(\frac{X}{n} - p \right)^2 - Z^2_{1-\alpha/2} \frac{pq}{n} \le 0$$

Η συνάρτηση K(p) είναι μία τετραγωνική μορφή ως προς p και είναι εύκολο να προσδιορισθούν οι τιμές του p που ικανοποιούν την σχέση $K(p) \le 0$ με καθορισμό των λύσεων της $K(p) = 0$ έστω p_1 και p_2. (Τα p_1 και p_2 είναι συναρτήσεις του X/n και του $Z_{1-\alpha/2}$).
Τότε, η τελευταία ανισότητα παίρνει την μορφή,

$$p_1 \le p \le p_2$$

και, επομένως, θα έχουμε ότι,

$$P\ (p_1 \le p \le p_2\ )\ = 1\text{-}\alpha$$

Πολλοί στατιστικοί προτιμούν μια απλουστευμένη προσέγγιση η οποία δεν είναι βέβαια εξίσου ακριβής. Αυτό που κάνουν είναι ότι

λύνουν ως προς p την ανισότητα της αρχικής σχέσης η οποία και γράφεται ισοδύναμα με τον εξής τρόπο:

$$P\left(\frac{X}{n}-Z_{1-\alpha/2}\sqrt{\frac{pq}{n}}\le p\le\frac{X}{n}+Z_{1-\alpha/2}\sqrt{\frac{pq}{n}}\right)=1-\alpha$$

Στην συνέχεια, αντικαθιστούν τα p και q με τις εκτιμήτριές τους X/n και 1-(X/n), αντίστοιχα βρίσκοντας ως <u>100(1-α)% διάστημα εμπιστοσύνης για το p το,</u>

$$\frac{X}{n}\pm Z_{1-\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

ή ισοδύναμα το,

$$\frac{X}{n}\pm Z_{1-\alpha/2}\sqrt{\frac{(X/n)(1-X/n)}{n}}$$

Παράδειγμα: Μια εταιρεία δημοσκοπήσεων ενδιαφέρεται να εκτιμήσει το ποσοστό των ψηφοφόρων που προτιμούν ένα συγκεκριμένο υποψήφιο. Για τον λόγο αυτό, επιλέγει ένα τυχαίο δείγμα 100 ψηφοφόρων και βρίσκει ότι 55 από αυτούς προτιμούν τον συγκεκριμένο υποψήφιο. Τί μπορούμε να πούμε για την πιθανότητα του υποψηφίου αυτού να κερδίσει τις εκλογές με βάση το δείγμα αυτό;

Λύση: Θα κατασκευάσουμε ένα 95% διάστημα εμπιστοσύνης για το p την αναλογία δηλαδή των εκλογέων του πληθυσμού που προτίθεται να ψηφίσει για τον συγκεκριμένο υποψήφιο.

Από τα δεδομένα του προβλήματος,

$$n = 100 \qquad X = 55$$

και επομένως,

$$\hat{p} = .55 \qquad \hat{q} = .45$$

και το 95% διάστημα εμπιστοσύνης είναι,

$$.55 \pm 1.96\sqrt{\frac{(.55)(.45)}{100}}$$

ή ισοδύναμα,

(.45, .65)

Επομένως, με πιθανότητα 95% η αναλογία των ψηφοφόρων του υπό μελέτη πληθυσμού που θα ψήφιζαν τον συγκεκριμένο υποψήφιο, αν οι εκλογές γίνονταν την ημέρα της σφυγμομέτρησης, θα ήταν μεταξύ 45% και 65%.

Σημείωση: Η μεθοδολογία που αναπτύξαμε ισχύει με την προϋπόθεση ότι το μέγεθος του δείγματος είναι αρκετά μεγάλο και ο αριθμός των επιτυχιών ή των αποτυχιών είναι επίσης αρκετά μεγάλος, ώστε να μπορεί να χρησιμοποιηθεί η κανονική προσέγγιση της διωνυμικής κατανομής. Αν όμως ή το μέγεθος του δείγματος δεν είναι μεγάλο ή κάποιο από τα ποσοστά των αποτυχιών ή των επιτυχιών είναι πολύ μικρό, τότε θα πρέπει να χρησιμοποιηθεί η διωνυμική κατανομή, δηλαδή, η μεθοδολογία που αναπτύξαμε στην αρχή της ενότητας.

Για ένα δεδομένο μέγεθος δείγματος, τα διαστήματα εμπιστοσύνης για αναλογίες φαίνονται να είναι ευρύτερα σε σχέση με αυτά που αντιστοιχούν σε συνεχείς μεταβλητές. Αυτό γιατί σε συνεχείς τυχαίες μεταβλητές μετρήσεις σε κάθε παρατήρηση προσφέρουν περισσότερες πληροφορίες παρά σε μία δίτιμη μεταβλητή. Με άλλα λόγια, μια μεταβλητή που έχει μόνο δύο δυνατές τιμές είναι ένα πολύ προσεγγιστικό μέτρο σε σύγκριση με μια συνεχή μεταβλητή γι' αυτό και κάθε παρατήρηση στο δείγμα προσφέρει μια περιορισμένη ποσότητα πληροφορίας για την παράμετρο που ενδιαφερόμαστε να εκτιμήσουμε.

Β. Περίπτωση Δύο Πληθυσμών

Έστω ότι ενδιαφερόμαστε να εκτιμήσουμε τη διαφορά $p_1 - p_2$ των αναλογιών "επιτυχιών" σε δύο ανεξάρτητους πληθυσμούς. Για τον λόγο αυτό, θεωρούμε δύο ανεξάρτητα τυχαία δείγματα μεγέθους n_1 και n_2 από τους πληθυσμούς αντίστοιχα και καταγράφουμε τους αντίστοιχους αριθμούς των επιτυχιών έστω X_1 και X_2 .

Τότε,

$$\frac{X_1}{n_1} \text{ και } \frac{X_2}{n_2}$$

είναι αμερόληπτες εκτιμήτριες των p_1 και p_2 αντίστοιχα.
Επομένως,

$$\frac{X_1}{n_1} - \frac{X_2}{n_2}$$

είναι μια αμερόληπτη εκτιμήτρια του $p_1 - p_2$.
Εξάλλου, γνωρίζουμε ότι

$$\frac{X_i}{n_i} \underset{appr.}{\sim} N\left(p_i, \frac{p_i q_i}{n_i}\right) \qquad i = 1, 2$$

Επομένως,

$$\frac{X_1}{n_1} - \frac{X_2}{N_2} \underset{appr.}{\sim} N\left(p_1 - p_2, \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right)$$

Επομένως, το κατά προσέγγιση 100(1-α)% διάστημα εμπιστοσύνης για τη διαφορά $p_1 - p_2$, κάτω από τις συνήθεις προϋποθέσεις θα είναι το,

$$\frac{X_1}{n_1} - \frac{X_2}{N_2} \pm Z_{1-\alpha/2} \sqrt{\frac{\left(\frac{X_1}{n_1}\right)\left(1 - \frac{X_1}{n_1}\right)}{n_1} + \frac{\left(\frac{X_2}{n_2}\right)\left(1 - \frac{X_2}{n_2}\right)}{n_2}}$$

Παράδειγμα: Προκειμένου να διαμορφώσει την στρατηγική της προεκλογικής του εκστρατείας, ένας υποψήφιος επιθυμεί να εκτιμήσει την διαφορά στην απήχηση που έχει στο εκλογικό σώμα μεταξύ ανδρών και γυναικών ψηφοφόρων. Προκειμένου να εκτιμήσει την διαφορά αυτή, αποφασίζει να κατασκευάσει ένα 99% διάστημα εμπιστοσύνης για την διαφορά $p_1 - p_2$, όπου p_1 είναι η αναλογία των γυναικών του πληθυσμού που τον υποστηρίζουν και p_2 είναι η αντίστοιχη αναλογία των ανδρών του εκλογικού σώματος. Προκειμένου να προχωρήσει στην μελέτη αυτή, ο υποψήφιος επιλέγει ένα τυχαίο δείγμα 1.000 ψηφοφόρων από κάθε μία κατηγορία

ανδρών και γυναικών. Μεταξύ των ανδρών βρίσκει ότι 388 θα τον υποστήριζαν και μεταξύ των γυναικών βρίσκει ότι 459 θα τον υποστήριζαν.
Επομένως,

$$\hat{p}_1 = \frac{459}{1000} = .459 \qquad \text{και} \qquad \hat{p}_2 = \frac{388}{1000} = .388$$

Επομένως, 99% διάστημα εμπιστοσύνης για τη διαφορά $p_1 - p_2$ θα είναι το,

$$(.459-.388) \pm 2.58\sqrt{\frac{(.459)(.541)}{1000} + \frac{(.388)(.612)}{1000}}$$

ή

$$.071 \pm .057$$

ή

$$(.014, .128)$$

Επομένως, ο υποψήφιος μπορεί με βεβαιότητα 99% να θεωρήσει ότι η διαφορά των αναλογιών στον πληθυσμό μεταξύ γυναικών και ανδρών που θα τον υποστήριζαν κυμαίνεται μεταξύ 1.4% και 12.8%.