Διαστήματα Εμπιστοσύνης για αναλογίες

Αναλογία Πληθυσμού

- Η ***αναλογία πληθυσμού***, p , είναι η αναλογία τον ατόμων στον πληθυσμό που έχουν κάποιο συγκεκριμένο χαρακτηριστικό,
  - για παράδειγμα, η αναλογία των χρηστών Internet στον ελληνικό πληθυσμό,
  - η αναλογία των φοιτητών του Πανεπιστημίου Μακεδονίας που κατάγονται από την Κοζάνη, κλπ.
  - **Είναι, με άλλα λόγια, ο αριθμός των ατόμων του πληθυσμού που έχουν το συγκεκριμένο χαρακτηριστικό δια του συνολικού μεγέθους του πληθυσμού.**

- Η ***αναλογία στο δείγμα***, $\hat{p}$, είναι η αναλογία των ατόμων στο δείγμα που έχουν κάποιο συγκεκριμένο χαρακτηριστικό.
- Είναι, με άλλα λόγια, ο αριθμός των ατόμων του δείγματος που έχουν το συγκεκριμένο χαρακτηριστικό δια του συνολικού μεγέθους του δείγματος.
- Εάν πάρουμε πολλά δείγματα θα βρούμε πολλά διαφορετικά $\hat{p}$.
- Επομένως, η αναλογία στο δείγμα ακολουθεί μια κατανομή, η οποία λέγεται ***κατανομή δειγματοληψίας της αναλογίας*** $\hat{p}$.

,

- Ο μέσος, $\mu_{\hat{p}}$, είναι ίσος με την αναλογία στον πληθυσμό.
- Δηλαδή, εάν 30% των ελλήνων έχει πρόσβαση στο Internet και πάρουμε διάφορα δείγματα τότε θα βρούμε σε αυτά διάφορες αναλογίες, π.χ. 28%, 33%, 31%, κλπ.
- Ο μέσος όλων αυτών των αναλογιών είναι ίσος με την αναλογία του πληθυσμού, 30%.

’

- Το τυπικό σφάλμα, $\sigma_{\hat{p}}$, είναι ίσο με:

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

- Το Κεντρικό Οριακό Θεώρημα ισχύει και για τις αναλογίες, επομένως η κατανομή δειγματοληψίας του $\hat{p}$ προσεγγίζεται από την κανονική κατανομή εάν το δείγμα είναι αρκετά μεγάλο.

- Συνεπώς, για να προσεγγίσουμε πιθανότητες για το $\hat{p}$ χρησιμοποιούμε την τυπική κανονική κατανομή. Η προσέγγιση γίνεται όλο και μεγαλύτερη καθώς αυξάνει το μέγεθος του δείγματος.

- Η μετατροπή σε z τιμές γίνεται με τον τύπο:

$$z_{\hat{p}} = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

- Προσέξτε ότι λέγοντας αναλογία p είναι για διακριτό πληθυσμό το ίδιο με τις ***“επιτυχίες”*** που είχαμε στη **διωνυμική κατανομή**.
- Επομένως, εάν μας ενδιαφέρει ο αριθμός των ατόμων στο δείγμα με το χαρακτηριστικό που επιθυμούμε, τότε προσεγγίζουμε τις πιθανότητες με τη διωνυμική κατανομή.
- Για να έχει εφαρμογή το Κεντρικό Οριακό Θεώρημα σε αναλογίες στο δείγμα θα πρέπει να ισχύει ταυτόχρονα ότι
  - **np>10 και**
  - **n(1-p)>10** .

Διάστημα εμπιστοσύνης για αναλογίες

- Όταν δεν διαθέτουμε τις αναλογίες του πληθυσμού, που είναι το σύνηθες, τις εκτιμούμε. Το τυπικό σφάλμα του δειγματικού ποσοστού είναι

$$\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Διάστημα εμπιστοσύνης για αναλογίες

- Ο αριθμός των τυπικών σφαλμάτων – περιθώριο σφάλματος

$$z\left(\sigma_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Για να χρησιμοποιήσουμε τις z τιμές θα πρέπει να ισχύει $n\hat{p} > 10$ και $n(1 - \hat{p}) > 10$.

Στις περισσότερες έρευνες το μέγεθος του δείγματος είναι αρκετά μεγάλο, έτσι ώστε ο παραπάνω περιορισμός ικανοποιείται σχεδόν πάντα.

Πόσο μεγάλο πρέπει να είναι το μέγεθος του δείγματος;

- Το ερώτημα αυτό είναι πολύ σημαντικό για κάθε είδους έρευνα.
- Η απάντηση δεν είναι εύκολη και οι στατιστικοί έχουν αναπτύξει αρκετά περίπλοκες σε ορισμένες περιπτώσεις τεχνικές για το σκοπό αυτό.
- Διαισθητικά θα έλεγε κανείς ότι όσο πιο μεγάλο είναι το δείγμα τόσο το καλύτερο.
- Αλλά μεγάλο δείγμα σημαίνει και μεγάλο κόστος σε χρόνο και σε χρήμα.

Μέγεθος δείγματος n	Περιθώριο σφάλματος $z\left(\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right)$	Περιθώριο σφάλματος %
100	$1{,}96\left(\sqrt{\frac{(0{,}54)(0{,}46)}{100}}\right) = 0{,}09769$	9,77%
500	$1{,}96\left(\sqrt{\frac{(0{,}54)(0{,}46)}{500}}\right) = 0{,}04369$	4,37%
1.000	$1{,}96\left(\sqrt{\frac{(0{,}54)(0{,}46)}{1.000}}\right) = 0{,}03089$	3,09%
2.000	$1{,}96\left(\sqrt{\frac{(0{,}54)(0{,}46)}{2.000}}\right) = 0{,}02184$	2,18%
4.000	$1{,}96\left(\sqrt{\frac{(0{,}54)(0{,}46)}{4.000}}\right) = 0{,}01545$	1,55%

Διαστήματα εμπιστοσύνης για αναλογίες

- Εάν έχουμε αναλογίες ανθρώπων ή πραγμάτων με ένα συγκεκριμένο χαρακτηριστικό,
  - **τότε δημιουργούμε διαστήματα εμπιστοσύνης για την αναλογία στον πληθυσμό.**
- Το διάστημα εμπιστοσύνης είναι:

$$\left(\hat{p} - z\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right), \qquad \hat{p} + z\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)\right)$$

- Έστω ότι σε δείγμα 120 ατόμων η εξέταση αίματος έδειξε ότι δεν υπάρχει κανένα πρόβλημα στους 85. Να δημιουργηθεί το 95% διάστημα εμπιστοσύνης για την εξέταση αίματος στον πληθυσμό.
- Απάντηση
- Η αναλογία είναι $\hat{p} = \frac{0{,}85}{120} = 0{,}71$

$$1 - \hat{p} = 0{,}29$$

$$np = 120 \cdot (0{,}71) = 85.2 > 10$$

$$n(1-p) = 24{,}7 > 10$$

- Η τιμή για διάστημα εμπιστοσύνης 95% είναι 1,96, οπότε έχουμε:

$$\left(\hat{p} - z\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right), \quad \hat{p} + z\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)\right)$$

$$\hat{p} - z\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) < p < \hat{p} + z\left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

$$\left(0{,}71 - 1{,}96\left(\sqrt{\frac{(0{,}71)(0{,}29)}{120}}\right), \quad 0{,}71 + 1{,}96\left(\sqrt{\frac{(0{,}71)(0{,}29)}{120}}\right)\right)$$

$$= (0{,}71 - 1{,}96 \cdot (0{,}041493), \quad 0{,}71 + 1{,}96 \cdot (0{,}041493)\ ) =$$

$$= (0{,}71 - 0{,}081326, \quad 0{,}71 + 0{,}081326\ ) = (0{,}63, \quad 0{,}79\ )$$

- Επομένως, μπορούμε να πούμε με 95% εμπιστοσύνη ότι το ποσοστό των εξετάσεων αίματος που δεν θα έχουν πρόβλημα στον πληθυσμό είναι ανάμεσα σε 63% και 79% με βάση το δείγμα μας.

- Επομένως, μπορούμε να πούμε με 95% εμπιστοσύνη ότι το ποσοστό των εξετάσεων αίματος που δεν θα έχουν πρόβλημα στον πληθυσμό είναι ανάμεσα σε 63% και 79% με βάση το δείγμα μας.

- Σε μια πειραματική μελέτη μια νέα κλινική θεραπεία ‘ΑΑΑ’ δόθηκε σε τυχαίο δείγμα 100 ασθενών. Στο τέλος της μελέτης 65 ασθενείς αποθεραπεύτηκαν. Βρείτε το 90% διάστημα εμπιστοσύνης της μέσης τιμής του ποσοστού αποθεραπείας της ‘ΑΑΑ’ (αποτελεσματικότητα της ‘ΧΤΥΖ’).

- Σε μια πειραματική μελέτη μια νέα μέθοδος διδασκαλία εφαρμόστηκε σε τυχαίο δείγμα 300 φοιτητών. Στο τέλος της μελέτης 270 φοιτητές έλαβαν προβιβάσιμο βαθμό. Βρείτε το 99% διάστημα εμπιστοσύνης της μέσης τιμής του ποσοστού επιτυχίας της εν λόγω μεθόδου.