

Διάστημα Εμπιστοσύνης για σύγκριση ποσοστών ανάμεσα σε δυο πληθυσμούς: περίπτωση δυο ανεξάρτητων τυχαίων δειγμάτων

Στα τελευταία μαθήματα είδαμε πως μπορούμε να ελέγξουμε τη σχέση ανάμεσα στις μέσες τιμές δυο πληθυσμών, όταν στη διάθεση μας έχουμε δυο ανεξάρτητα τυχαία δείγματα.

Εξίσου ενδιαφέρουσα είναι και η περίπτωση στην οποία θέλουμε να εξετάσουμε τα ποσοστά των ατόμων με κάποιο συγκεκριμένο χαρακτηριστικό, ανάμεσα σε δυο διαφορετικούς πληθυσμούς.

Για παράδειγμα:

- είναι το ποσοστό των φοιτητριών που επιτυγχάνουν σε κάποιες εξετάσεις, ίδιο με αυτό των φοιτητών;
 - είναι το ποσοστό των καπνιστών ίδιο ανάμεσα στις γυναίκες και τους άντρες;
 - το ποσοστό των υλικών που αντέχουν μια συγκεκριμένη πίεση, είναι το ίδιο ανάμεσα σε δυο διαφορετικές μεθόδους παρασκευής;
-

Έστω λοιπόν ότι θέλουμε να απαντήσουμε σε τέτοιας φύσεως ερωτήματα για μια τ.μ. X η οποία μπορεί να πάρει **μόνο δυο διαφορετικές τιμές** (π.χ. το φύλο ενός ατόμου).

Ας θεωρήσουμε ότι η τ.μ. X παίρνει τις τιμές 0 ή 1 και ότι στον ένα πληθυσμό, το ποσοστό των ατόμων με τιμή 1 (αλλιώς, η πιθανότητα κάποιος να πάρει την τιμή 1) είναι ίσο με

$$P(X=1) = p_1$$

ενώ στο δεύτερο, το αντίστοιχο ποσοστό είναι

$$P(X=1) = p_2.$$

Τα p_1, p_2 είναι προφανώς άγνωστα και θέλουμε να εξετάσουμε τη σχέση ανάμεσά τους. Δηλαδή, τι ισχύει

$$p_1 > p_2 \text{ ή } p_1 < p_2 \text{ ή } p_1 = p_2;$$

Για το σκοπό αυτό, θεωρούμε ότι υπάρχουν δυο **ανεξάρτητα** τυχαία δείγματα, από τους δυο πληθυσμούς.

Ένα τυχαίο δείγμα X_1, X_2, \dots, X_{v_1} μεγέθους v_1 από τον ένα πληθυσμό (για τη μεταβλητή που μελετάμε) και ένα ακόμη τυχαίο δείγμα Y_1, Y_2, \dots, Y_{v_2} μεγέθους v_2 , από το δεύτερο πληθυσμό.

Ας συμβολίσουμε με \hat{p}_1, \hat{p}_2 τα δειγματικά ποσοστά. Δηλαδή, το \hat{p}_1 συμβολίζει πόσα άτομα στο δείγμα X_1, X_2, \dots, X_{v_1} έχουν το χαρακτηριστικό που μελετάμε (αντίστοιχα το \hat{p}_2). Οπότε,

$$\hat{p}_1 = \frac{X_1 + X_2 + \dots + X_{v_1}}{v_1} \text{ και } \hat{p}_2 = \frac{Y_1 + Y_2 + \dots + Y_{v_2}}{v_2}$$

(αφού οι X_i, Y_i παίρνουν την τιμή 1 όταν κάποιος έχει το χαρακτηριστικό και 0 διαφορετικά).

Οδυσμζια σνάρμσν

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{v_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{v_2}}} \sim N(0,1)$$

η πρσγγισις
 v_1, v_2 μεγάλσ

$$P\left(-z_{\alpha/2} \leq \text{οδυσμζια} \leq z_{\alpha/2}\right) = 1 - \alpha$$

$$P\left(\hat{p}_1 - \hat{p}_2 - z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{v_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{v_2}} \leq p_1 - p_2 \leq \hat{p}_1 - \hat{p}_2 + z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{v_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{v_2}}\right) = 1 - \alpha$$

Δ.Ε (1-α)·100% :

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{v_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{v_2}}$$

Παράδειγμα 22.1

Από τις εισαγωγικές εξετάσεις στα πανεπιστήμια, καταγράψαμε από τα δεδομένα των προηγούμενων ετών ότι από τα 4500 κορίτσια που συμμετείχαν σ' αυτές, οι 2550 εισήχθησαν σε κάποιο πανεπιστήμιο, ενώ από τα 4200 αγόρια τα 2300 ήταν αυτά που τελικά πέρασαν σε κάποια σχολή. Να κατασκευασθεί διάστημα εμπιστοσύνης με συντελεστή εμπιστοσύνης 99% για τη διαφορά των ποσοστών επιτυχίας κοριτσιών και αγοριών.

Συμβολίζουμε με p_1 το ποσοστό επιτυχίας των κοριτσιών και με p_2 το αντίστοιχο των αγοριών.

Ταυτόχρονα, από την εκφώνηση του παραδείγματος γνωρίζουμε ότι

$$n_1 = 4500, n_2 = 4200, \hat{p}_1 = 2550 / 4500 = 0.57, \hat{p}_2 = 2300 / 4200 = 0.55,$$

$$\alpha = 0.01,$$

οπότε,

Δ. Ε:

$$\frac{2550}{4500}$$

$$\frac{2300}{4200}$$

$$\hat{p}_1 - \hat{p}_2 = 0.57 - 0.55 = 0.02$$

$$\alpha = 0.01, \quad \frac{\alpha}{2} = 0.005$$

$$\sqrt{\frac{0.55(1-0.55)}{4200} + \frac{0.57(1-0.57)}{4500}} =$$

$$0.02 \pm \underset{2.57}{Z_{0.005}} \cdot 0.0106 = (-0.0072, 0.0472)$$

Το 0 οριακά ανήκει στο διάστημα
φαίνεται με εμβυσότητα 99% το $p_1 > p_2$
αλλά δεν μπορεί να αποκλειστεί ότι το p_1 μπορεί
να είναι ίσο με το p_2

Παράδειγμα 22.2

Σε μια έρευνα που αποσκοπούσε στο να συγκρίνει τα ποσοστά καπνιστών σε άντρες και γυναίκες με ηλικία κάτω των 25 ετών, συμμετείχαν 150 γυναίκες και 180 άντρες. Από τις 150 γυναίκες οι 90 απάντησαν πως καπνίζουν (και οι υπόλοιπες όχι) και από τους άντρες οι 80 δήλωσαν καπνιστές. Μπορούμε να υποστηρίξουμε με συντελεστή εμπιστοσύνης 95% ότι η διαφορά των ποσοστών γυναικών και ανδρών είναι 10%;

$$p_r = \frac{90}{150} = 0.6 \quad p_A = \frac{80}{180} = 0.44, \quad \alpha = 5\%$$

$$p_r - p_A = 0.6 - 0.44 = 0.16$$

$$\sqrt{\frac{0.6(1-0.6)}{150} + \frac{0.44(1-0.44)}{180}} = \sqrt{0.0016 + 0.0013} = 0.0538$$

$$0.16 \pm \underset{\substack{0.025 \\ 1.96}}{Z} \cdot 0.0538 : (0.054, 0.265)$$

5.4% vs 26%

Αρα με εμπιστοσύνη 95% τα νοσήρια ανδρών
και γυναικών προοιῶν να διαφέρουν κατὰ 10%
0 \notin διάστημα αρα τα νοσήρια δεν είναι ίσα
και φαίνεται $p_1 > p_2$

Παράδειγμα 22.3

Από μια έρευνα θέλουμε να εξετάσουμε εάν το ποσοστό επιτυχίας μιας θεραπείας σε άτομα ηλικίας άνω των 45 ετών, είναι διαφορετικό με το αντίστοιχο ποσοστό σε άτομα με ηλικία μικρότερη των 45 ετών. Από 100 άτομα ηλικίας άνω των 45 ετών, οι 85 θεραπεύτηκαν ακολουθώντας τη συγκεκριμένη θεραπεία, ενώ από τα 150 άτομα ηλικίας μικρότερης των 45 ετών, οι 125 ήταν εκείνοι που θεραπεύτηκαν. Μπορούμε να υποστηρίξουμε με συντελεστή εμπιστοσύνης 90% ότι τα δυο ποσοστά είναι διαφορετικά;

Διάστημα Εμπιστοσύνης για τη διαφορά των μέσων τιμών δυο τυχαίων μεταβλητών που ακολουθούν κανονική κατανομή, με γνωστές διασπορές: περίπτωση δυο εξαρτημένων τυχαίων δειγμάτων (confidence interval for the difference between the means of normal populations with known variances: paired random samples)

Στα τελευταία διαστήματα εμπιστοσύνης που μελετήσαμε, βασιστήκαμε στο ότι από τους δυο διαφορετικούς πληθυσμούς, έχουμε στη διάθεσή μας δυο ανεξάρτητα τυχαία δείγματα.

Υποθέταμε δηλαδή ότι η επιλογή των δυο δειγμάτων γινόταν με τέτοιο τρόπο ώστε οι τιμές που παρατηρούσαμε στο ένα δείγμα ή αλλιώς, τα άτομα (αντικείμενα) που αποτελούσαν το ένα δείγμα, δεν επηρέαζαν τις τιμές που θα παρατηρήσουμε στο άλλο δείγμα.

Σε πολλές περιπτώσεις όμως, μια τέτοια υπόθεση δεν είναι καθόλου ρεαλιστική ή ορθή.

Για παράδειγμα, ας υποθέσουμε ότι θέλουμε να εξετάσουμε την ψυχολογική κατάσταση των ασθενών πριν και μετά από μια σοβαρή επέμβαση, ή την αντοχή ενός υλικού πριν και μετά από κάποιο γεγονός.

Ένας τρόπος να κάνουμε κάτι τέτοιο είναι οι ασθενείς να απαντήσουν σε ένα σύνολο ερωτήσεων πριν από την επέμβαση (από τις οποίες θα πάρουμε ένα συνολικό σκορ) και αφού ολοκληρωθεί αυτή, οι ίδιοι ασθενείς να απαντήσουν εκ νέου, στις προηγούμενες ερωτήσεις.

Στη δεύτερη περίπτωση, από ένα σύνολο υλικών μετράμε την αντοχή τους πριν από κάποιο γεγονός, και ξανέ μετά την πραγματοποίησή του.

Στην πρώτη περίπτωση θα έχουμε δύο τυχαία δείγματα:

- τις απαντήσεις n ατόμων πριν την επέμβαση (X_1, X_2, \dots, X_n) και
- τις απαντήσεις των ίδιων ατόμων μετά την επέμβαση (Y_1, Y_2, \dots, Y_n).

Αυτό που θα μας ενδιέφερε π.χ. είναι να συγκρίνουμε τις μέσες τιμές των απαντήσεων πριν και μετά την επέμβαση.

Είναι όμως λογικό να υποθέσουμε ότι τα δυο δείγματα είναι ανεξάρτητα;
Μάλλον, όχι!

Ας θεωρήσουμε ένα δεύτερο παράδειγμα όπου ενδιαφερόμαστε κατά πόσο μια τηλεοπτική εμφάνιση δυο πολιτικών αρχηγών επηρέασε την άποψη/στάση των ψηφοφόρων.

Έτσι σ' ένα τυχαίο δείγμα ατόμων ρωτήσαμε πριν από την τηλεοπτική εμφάνιση, ποιον από τους δύο θεωρούν καταλληλότερο πρωθυπουργό. Στα ίδια άτομα και μετά από την τηλεοπτική εμφάνιση κάναμε την ίδια ερώτηση.

Με τον τρόπο αυτό, έχουμε πάλι στη διάθεσή μας δύο τυχαία δείγματα:

- τις απαντήσεις n ατόμων πριν την τηλεοπτική εμφάνιση (X_1, X_2, \dots, X_n) και
- τις απαντήσεις των ίδιων ατόμων μετά την τηλεοπτική εμφάνιση (Y_1, Y_2, \dots, Y_n) .

Τι μπορούμε να πούμε για τα ποσοστά που πήρε ο ένας από τους δυο αρχηγούς;

Είναι ίδια πριν και μετά από την τηλεοπτική εμφάνιση;

Είναι λογικό να υποθέσουμε και εδώ ότι έχουμε ανεξάρτητα δείγματα; Μάλλον, όχι!

Ο λόγος είναι απλός.

Η τιμή που θα πάρουμε από ένα άτομο π.χ. το πρώτο άτομο του δείγματος μας (δηλ. η X_1) είναι λογικό να υποθέσουμε ότι επηρεάζει την τιμή που θα μας δώσει το ίδιο άτομο μετά από την τηλεοπτική εμφάνιση (δηλ. την Y_1) ή την επέμβαση.

Το γεγονός αυτό διαφοροποιεί τον τρόπο που πρέπει να κάνουμε τους ελέγχους για τις μέσες τιμές ή τα ποσοστά των δυο πληθυσμών, σε σχέση με τη μεθοδολογία που ακολουθήσαμε στην περίπτωση των δυο ανεξάρτητων δειγμάτων.

Δυο δείγματα από δυο πληθυσμούς, τα οποία δεν είναι ανεξάρτητα αλλά παρουσιάζουν ένα είδος εξάρτησης/συσχέτισης, θα τα καλούμε **συσχετισμένα δείγματα** (ζευγαρωτές παρατηρήσεις, paired samples).

Ας υποθέσουμε επιπλέον, ότι η τ.μ. που μελετάμε ακολουθεί την **κανονική κατανομή** και στους δυο πληθυσμούς, με **άγνωστες** διασπορές

Συγκεκριμένα, θεωρούμε ότι το τυχαίο δείγμα X_1, X_2, \dots, X_n προέρχεται από μια $N(\mu_1, \sigma_1^2)$ και το Y_1, Y_2, \dots, Y_n από μια $N(\mu_2, \sigma_2^2)$, όπου τα σ_1^2, σ_2^2 θεωρούνται άγνωστα.

Από τα δυο εξαρτημένα δείγματα που έχουμε στη διάθεσή μας, συμβολίζουμε με D_1, D_2, \dots, D_n τις διαφορές ανάμεσα στις τιμές των X_i και Y_i . Δηλαδή,

$$D_1 = X_1 - Y_1, D_2 = X_2 - Y_2, \dots, D_n = X_n - Y_n.$$

Τη δειγματική μέση τιμή των D_1, D_2, \dots, D_v και τη δειγματική τυπική απόκλιση, τις συμβολίζουμε με \bar{D} και S_d , αντιστοίχως. Άρα,

$$\bar{D} = \frac{D_1 + D_2 + \dots + D_v}{v}$$

$$S_d = \sqrt{\frac{1}{v-1} \left[(D_1 - \bar{D})^2 + (D_2 - \bar{D})^2 + \dots + (D_v - \bar{D})^2 \right]}.$$

Οδηγία συνάρτησης

Παράδειγμα 22.4

Το υπουργείο παιδείας μιας χώρας ήθελε να εξετάσει κατά πόσο μια σειρά σεμιναρίων βοηθάει στην απόκτηση χρήσιμων γνώσεων και γενικότερα στην κατάρτιση μιας επαγγελματικής ομάδας.

Για το λόγο αυτό σε 10 άτομα, που θα συμμετείχαν στα σεμινάρια, δόθηκε ένα τεστ γνώσεων και καταγράφηκαν οι επιδόσεις τους. Στα ίδια 10 άτομα και μετά από τη συμμετοχή τους στα σεμινάρια, δόθηκε ένα ισοδύναμο τεστ γνώσεων. Οι επιδόσεις από τα δυο τεστ φαίνονται στον επόμενο πίνακα.

Να κατασκευασθεί διάστημα εμπιστοσύνης με σ.ε. 95% για τη διαφορά των μέσων τιμών επίδοσης πριν και μετά τα σεμινάρια. Υποθέτουμε ότι η επίδοση πριν και μετά ακολουθεί κανονική κατανομή.

Πριν	50	62	62	45	80	90	55	65	88	56
Μετά	55	68	65	60	82	92	70	65	89	57

Ας συμβολίσουμε με μ_1 τη μέση τιμή της επίδοσης πριν τα σεμινάρια και με μ_2 μέση τιμή της επίδοσης μετά τα σεμινάρια.

Από τον πίνακα με τα δεδομένα παίρνουμε

Πριν (X_i)	50	62	62	45	80	90	55	65	88	56
Μετά (Y_i)	55	68	65	60	82	92	70	65	89	57
Διαφορά (D_i)	-5	-6	-3	-15	-2	-2	-15	0	-1	-1

Οπότε

$$\bar{D} = \frac{D_1 + D_2 + \dots + D_v}{v} = \frac{-5 - 6 - 3 - 15 - 2 - 2 - 15 + 0 - 1 - 1}{10} = -4.8$$

$$S_d = \sqrt{\frac{1}{9} \left[(-5 + 4.8)^2 + (-6 + 4.8)^2 + (-3 + 4.8)^2 + \dots + (-1 + 4.8)^2 \right]} = 5.77$$

$\Delta E.$

$$\alpha = 0.05$$

$$-4.8 \pm \underset{\substack{\text{"} \\ 2.262}}{t_{9, 0.025}} \frac{5.77}{\sqrt{10}} : (-8.92, -0.67)$$

\notin διάστημα αρα οι μέσες αξίες πριν και
μετα τα σεμινάρια διαφέρουν με
εμπιστοσύνη 95% και φαίνεται ότι $\mu_1 < \mu_2$

Παράδειγμα 22.5

Θέλοντας τα παιδιά να βλέπουν το σχολείο και ως ένα χώρο δημιουργίας, ένα χώρο χρήσιμων και εποικοδομητικών δραστηριοτήτων, επιχειρείται να εισαχθούν νέα μαθήματα και νέες δραστηριότητες. Για να αξιολογηθεί η αποτελεσματικότητα των αλλαγών, ένα δείγμα 8 μαθητών κλήθηκε να δώσει απαντήσεις για το πώς «βλέπει» το σχολείο στην αρχή και στο τέλος της χρονιάς. Τα σκορ από τις απαντήσεις τους φαίνονται στον επόμενο πίνακα (μεγαλύτερο σκορ σημαίνει καλύτερη «εικόνα» για το σχολείο).

Να κατασκευασθεί Δ.Ε. με σ.ε. 99% για τη διαφορά των μέσων τιμών των σκορ στο τέλος και την αρχή της χρονιάς. Υποθέτουμε ότι τα σκορ πριν και μετά ακολουθούν κανονική κατανομή.

Αρχή	100	85	110	80	95	90	70	85
Τέλος	120	98	115	120	115	125	105	130