διάγραμμα

ΣΤΑΤΙΣΤΙΚΗ

Περιγραφική Στατιστική

Εκτιμητική

Διαστήματα Εμπιστοσύνης

Έλεγχοι Στατιστικών Υποθέσεων

Ανάλυση Παλινδρόμησης & Συσχέτισης

Λυμένες Ασκήσεις - Θέματα Εξετάσεων

ΑΘΗΝΑ 1996

Πρόλογος

Οι σημειώσεις αυτές γράφτηκαν για τους φοιτητές του Εθνικού Μετσόβιου Πολυτεχνείου και καλύπτουν πλήρως το μάθημα της Στατιστικής. Το φυλλάδιο αυτό συμπληρώνεται με το φυλλάδιο των Πιθανοτήτων.

Σκοπός των σημειώσεων αυτών είναι να δοθούν με σαφήνεια και απλότητα όλες οι έννοιες και οι εφαρμογές της στατιστικής, διατηρώντας όμως παράλληλα την επιστημονική αυστηρότητα και ευκρίνεια που πρέπει να διέπει τέτοιες προσπάθειες.

Το φυλλάδιο αυτό χωρίστηκε σε πέντε κεφάλαια. Στο πρώτο, την περιγραφική στατιστική, αναφέρονται γενικές έννοιες της στατιστικής. Ακολουθεί το ένα από τα τρία βασικά αντικείμενα της Στατιστικής, η Εκτιμητική, στο οποίο πρωταρχικό ρόλο έχουν οι μέθοδοι εκτίμησης που είναι αυτή των ροπών και η μέθοδος μέγιστης πιθανοφάνειας. Στο τρίτο κεφάλαιο βρίσκονται τα διαστήματα εμπιστοσύνης που αποτελούν μια εκτίμηση κατά διάστημα σε αντιδιαστολή με την εκτίμηση κατά σημείο που αναφέρεται στο δεύτερο κεφάλαιο. Εν συνεχεία εξετάζουμε το τρίτο βασικό κομμάτι της Στατιστικής που είναι ο έλεγχος των στατιστικών υποθέσεων και τέλος παραθέτουμε κάποια στοιχεία από την ανάλυση παλινδρόμησης και συσχέτισης.

Καταβάλαμε μεγάλη προσπάθεια για τη συγκέντρωση πλήθους χαρακτηριστικών ασκήσεων και απλούστευσης της θεωρίας - όχι όμως υπεραπλούστευσης, καθώς το κομμάτι αυτό των Μαθηματικών είναι από τα πλέον δύσκολα. Συνίσταται διεξοδική μελέτη της θεωρίας γιατί είναι απαραίτητη για την πλήρη κατανόηση των ασκήσεων.

Μεγάλο βάρος θα πρέπει να δοθεί στις παρατηρήσεις των ασκήσεων, οι οποίες αποτελούν συμπλήρωμα της θεωρίας. Φροντίστηκε ώστε να υπάρχει ομοιογένεια στην έκφραση καθώς και στην διατύπωση για να μη δημιουργούνται προβλήματα στους αναγνώστες.

Φ. Φωτόπουλος

Α. Χαραλαμπάκης

Περιεχόμενα

Κεφάλαιο 1ο: Περιγραφική Στατιστική

Α. ΘΕΩΡΙΑ

1.1 Εισαγωγή Στο Αντικείμενο

Στατιστική είναι ένας κλάδος των εφαρμοσμένων Μαθηματικών που ασχολείται με τη συλλογή, τη συνοπτική παρουσίαση και την ανάλυση - ερμηνεία παρατηρήσεων που υπόκεινται σε τυχαίες μεταβολές, με τελικό στόχο την εξαγωγή βάσιμων συμπερασμάτων και τη λήψη βέλτιστων αποφάσεων με συνθήκες αβεβαιότητας.

Θα ασχοληθούμε αρχικά με την εκτίμηση μιας ή περισσοτέρων παραμέτρων κατανομής ενός χαρακτηριστικού του πληθυσμού που μελετάμε κάθε φορά. Η εκτίμηση αυτή θα γίνει με δυο τρόπους. Ο πρώτος αναφέρεται στο δεύτερο κεφάλαιο και είναι η εκτίμηση κατά σημείο. Στο τρίτο κεφάλαιο θα ασχοληθούμε με εκτίμηση κατά διάστημα. Κατόπιν θα περάσουμε στον έλεγχο στατιστικών υποθέσεων. Ο έλεγχος αυτός εν γένει δεν είναι ούτε εξαιρετικά ακριβής ούτε και πολύ αξιόπιστος. Παρόλα αυτά, αποτελεί ένα καλό και γρήγορο μέσο για πρόχειρες διαπιστώσεις. Τέλος θα ασχοληθούμε με ορισμένα στοιχεία από την ανάλυση παλινδρόμησης και συσχέτισης.

1.2 Παρουσίαση Αποτελεσμάτων

Για την παρουσίαση των αποτελεσμάτων χρησιμοποιούνται τα ιστογράμματα, τα κυκλικά διαγράμματα, τα φυλλογράμματα κα. Η αναλυτική περιγραφή των μεθόδων αυτών είναι κάτι που ξεφεύγει από τα όρια των σημειώσεων αυτών.

1.3 Χαρακτηριστικές Ποσότητες

Είναι η μέση τιμή του δείγματος που δίνεται από τον τύπο:

$$\boxed{\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i}$$

Παράδειγμα: Πέντε νομίσματα ρίχνονται 50 φορές ταυτόχρονα, και σε κάθε ρίψη καταγράφεται ο αριθμός X των "κεφαλών" που προκύπτει. Έχουμε τον παρακάτω πίνακα που μας δείχνει πόσες φορές εμφανίστηκαν τα εκάστοτε αποτελέσματα του πειράματος. Δηλαδή 15 φορές εμφανίστηκαν τρεις κεφαλές ταυτόχρονα.

$\mathbf{X_i}$	$\mathbf{f_i}$
0	1
1	7
2	19
3	15
4	8
5	0
Σύνολο	50

Η μέση τιμή θα είναι σύμφωνα με τον παραπάνω τύπο:

$$\bar{x} = \frac{1}{50}\sum_{i=1}^{50} x_i = 0*1 + 1*7 + 2*19 + 3*15 + 4*8 + 5*0 = \frac{122}{50} = 2{,}44.$$

Μια άλλη ποσότητα ενδεικτική της μεταβλητότητας του X είναι η <u>διασπορά</u> του δείγματος η οποία και θα δίδεται εκ του τύπου:

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

Για το παραπάνω παράδειγμα θα έχουμε:

$$s^2 = \frac{1}{50-1}\sum_{i=1}^{50}\left(x_i - 2{,}44\right)^2 = \frac{1}{49}[1*(0-2{,}44)^2 + 7*(1-2{,}44)^2 + 19*(2-2{,}44)^2 +$$
$$+15*(3-2{,}44)^2 + 8*(4-2{,}44)^2 + 0*(5-2{,}44)^2] = \frac{48{,}32}{49} = 0{,}986$$

Η τετραγωνική ρίζα της διασποράς, δηλαδή η τιμή s καλείται <u>τυπική απόκλιση</u> του δείγματος. Η δειγματική τυπική απόκλιση όπως άλλωστε και η διασπορά, είναι ένα μέτρο του πόσο μακριά (κατά μέσο όρο) από τη μέση τιμή $\bar{x}$ "διασπείρονται" τα αποτελέσματα $x_1, x_2, \ldots, x_n$. Αν προκύψει (απίθανο στην πραγματικότητα) s=0, τότε όλα τα x_i συμπίπτουν μεταξύ τους.

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

Για το παράδειγμά μας, η τυπική απόκλιση είναι προφανώς $s^2 = 0{,}986 \Rightarrow s = 0{,}993$. Τονίζουμε ότι s>0 πάντα.

<u>Κορυφή</u> είναι η τιμή με τη μεγαλύτερη συχνότητα. Αν όμως έχουμε συνεχή περίπτωση τότε κορυφή είναι η κεντρική τιμή της κλάσης με τη μεγαλύτερη συχνότητα. Προφανώς στο παράδειγμά μας, η ρίψη νομίσματος είναι διακριτή περίπτωση καθώς τα f_i παίρνουν πεπερασμένες τιμές. Θα δώσουμε παρακάτω και ένα παράδειγμα με συνεχή μεταβλητή για την πλήρη κατανόηση του θέματος.

Κορυφή για το παράδειγμα αυτό είναι το x=2 , γιατί εκεί εμφανίστηκε μεγαλύτερη συχνότητα f_2=19.

<u>Διάμεσος</u> είναι η παρατήρηση εκείνη η οποία είναι μεγαλύτερη από το μισό ακριβώς του αριθμού των παρατηρήσεων. Αν έχουμε περιττό αριθμό παρατηρήσεων πχ 15, τότε η διάμεσος είναι το όγδοο στοιχείο που είναι και το "μεσαίο". Αν έχουμε άρτιο πχ 16, τότε διάμεσος θα είναι το ημιάθροισμα του όγδοου και ένατου στοιχείου. Διάμεσο έχουμε προφανώς στο 2.

Τέλος έχουμε τα <u>ρ-ποσοστιαία</u> σημεία. Για ρ=1/4 και ρ=1/3 έχουμε την πρώτη και τρίτη τεταρτητόμο αντίστοιχα. Για ρ=1/2 έχουμε τη διάμεσο.

Παράδειγμα: Σε δείγμα 50 εξαρτημάτων του αυτού τύπου μετρήθηκε η διάρκεια ζωής Χ(ρ) και προέκυψαν τα κάτωθι αποτελέσματα.

46	104	94	114	35	214	15	272	118	193
126	64	5	27	26	57	56	236	72	46
23	85	122	43	159	102	14	73	17	314
120	8	146	117	35	14	263	4	64	113
48	97	73	38	143	9	25	171	37	184

Παρατηρούμε ότι ενώ στο προηγούμενο παράδειγμα είχαμε μόλις 6 κατηγορίες και ήταν εύκολο να βρούμε τις συχνότητες εμφάνισης των αποτελεσμάτων (6 διαφορετικά) εδώ δεν γίνεται κάτι τέτοιο. Αυτό συμβαίνει γιατί κανένα Χ (εν γένει) δεν είναι ίσο με άλλο. Θα πρέπει να χωρίζουμε τα x_i σε κλάσεις όπως λέμε. Δηλαδή μια κλάση μπορεί να είναι από 0 ως 50, η άλλη 50 ως 100, αλλά οπωσδήποτε πρέπει το πλάτος τους να είναι ίδιο (εδώ 50).

Ποιος όμως είναι ο κατάλληλος αριθμός των διαμερίσεων που πρέπει να κάνουμε; Μπορούμε ασφαλώς με πλάτος 50 να κάνουμε 7, αλλά και με πλάτος 25 να κάνουμε 14. Έχει βρεθεί (εμπειρικά) ότι ο αριθμός των διαμερίσεων δίνεται με καλή προσέγγιση από τον παρακάτω τύπο:

$$\boxed{q = 1 + 3{,}32\log_{10} n}$$

όπου n το πλήθος των παρατηρήσεων, εδώ 50. Είναι $q=1+3{,}32\log_{10}50=6{,}64\rightarrow 7$. Μπορούμε να κάνουμε 8 διαμερίσεις με πλάτος 40 για πιο βολικά νούμερα. Παίρνουμε τον παρακάτω πίνακα συχνοτήτων.

Διάρκεια Ζωής	**Συχνότητα**
0-40	16
40-80	11
80-120	10
120-160	5
160-200	3
200-240	2
240-280	2
280-320	1
Σύνολο	*50*

Ο μέσος όρος βρίσκεται από το γνωστό τύπο: $\bar{x} = \frac{1}{50}\sum_{i=1}^{50} x_i$. Όμως το δείγμα αυτό ήδη είναι μεγάλο για να προσθέσουμε όλα τα στοιχεία του και υπάρχουν δείγματα ακόμη μεγαλύτερα. Μια προσέγγιση (αρκετά καλή) του μέσου όρου θα είναι αν προσθέσουμε τα γινόμενα των κέντρων των κλάσεων με τη συχνότητα:

$$\bar{x} = \frac{1}{50}(20*16 + 60*11 + 10*100 + 5*140 + 3*180 + 2*220 + 2*260 + 1*300) = \frac{4480}{50} = 89{,}6.$$

Για να βρούμε τη διασπορά, εφαρμόζουμε τον τύπο:

$s^2 = \frac{1}{50-1}\sum_{i=1}^{50}(x_i - \bar{x})^2$. Όμως μιας και έχουμε κάνει διαμερίσεις και ο προηγούμενος τύπος απαιτεί πολλές πράξεις, μπορούμε αντ' αυτού να κάνομε χρήση του τύπου:

$s^2 = \frac{1}{n-1}(\sum_{i=1}^{n}\left(y_i^2 v_i\right) - n\bar{x}^2)$, οπότε είναι $s^2 = \frac{1}{49}(663200 - 50*89{,}6^2)$ =5342,69 και επομένως η τυπική απόκλιση θα είναι s=73,09 (τετραγωνική ρίζα της διασποράς).

Το επόμενο βήμα είναι η εύρεση της κορυφής. Στην προηγούμενη περίπτωση που ήταν διακριτή μεταβλητή είχαμε την τιμή με τη μεγαλύτερη συχνότητα. Εδώ που η μεταβλητή είναι συνεχής, θα έχουμε ως κορυφή την κεντρική τιμή της κλάσης (διαμέρισης) με τη μεγαλύτερη συχνότητα. Η μεγαλύτερη συχνότητα είναι το 16 στην πρώτη διαμέριση, άρα η κορυφή θα είναι η κεντρική τιμή της πρώτης διαμέρισης, δηλαδή το μέσο του διαστήματος [0,40] που είναι το 20.

Όσον αφορά τη διάμεσο, ένας τρόπος είναι να κατασκευάσουμε το ιστόγραμμα των αθροιστικών συχνοτήτων. Η διάμεσος τότε θα δίνεται από την τετμημένη του σημείου του οποίου η παράλληλος προς τον άξονα των x από το σημείο ν/2 = 25, τέμνει το πολύγωνο. Όμως αυτός ο τρόπος απαιτεί την κατασκευή γραφήματος και ως εκ τούτου δε δύναται να έχει μεγάλη ακρίβεια. Θα χρησιμοποιήσουμε αντί της μεθόδου αυτής, ένα τύπο:

$$\boxed{\delta = L_i + \frac{\frac{a}{100}n - N_{i-1}}{v_i}c}$$

όπου c είναι το πλάτος της διαμέρισης, εδώ 40. L_i θα είναι το κάτω όριο του διαστήματος της επικρατούσας κλάσης. Επειδή έχουμε την [40,80] θα είναι L_i=40. Το α θα είναι 50% για τη διάμεσο, 25% και 75 % για την πρώτη και τρίτη τεταρτητόμο αντίστοιχα. Το δε v_i θα είναι η συχνότητα της επικρατούσας κλάσης. Τέλος το N_{i-1} θα ισούται με το άθροισμα των συχνοτήτων μέχρι τη διαμέριση i-1. Εδώ θα έχουμε N_{i-1} =16. Επομένως έχουμε:

$$\delta = 40 + \frac{\frac{1}{2}50 - 16}{11}40 = 72{,}7$$

Με τον όρο επικρατούσα κλάση εννοούμε την κλάση στην οποία βρίσκεται το μεσαίο στοιχείο (δηλαδή το 25ο.). Προφανώς είναι η δεύτερη, αφού στις δυο πρώτες περιέχονται 16+11=27 στοιχεία.

Αν τώρα αντί για α=50% βάλουμε α=25% και α=75% αντίστοιχα θα έχουμε την πρώτη και τρίτη τεταρτητόμο:

$$Q_1 = 40 + \frac{0{,}25 * 50 - 16}{11} = 27{,}3 \text{ και } Q_3 = 80 + \frac{0{,}75 * 50 - (16+11)}{10} = 122$$

1.4 Δειγματοληπτικές Κατανομές

Ας υποθέσουμε ότι έχουμε δείγματα από κανονικούς πληθυσμούς. Τότε αν $X_1, X_2, \ldots, X_\nu$ τυχαίο δείγμα, επειδή αυτό ακολουθεί κανονική κατανομή θα είναι $N(\mu,\sigma^2)$. Ισχύουν τα παρακάτω θεωρήματα:

ΘΕΩΡΗΜΑ 1ο: Ο μέσος όρος τυχαίου δείγματος θα ακολουθεί επίσης κανονική κατανομή $N(\mu,\sigma^2/\nu)$ όπου ν το πλήθος των δειγμάτων του πληθυσμού.

ΘΕΩΡΗΜΑ 2ο: Αν υποθέσουμε ότι έχουμε δυο δείγματα $X \sim (\mu_1,\sigma_1^2)$ και $Y \sim (\mu_2,\sigma_2^2)$ τότε και η διαφορά τους X-Y θα ακολουθεί κανονική κατανομή $X-Y \sim (\mu_1-\mu_2,\ \sigma_1^2+\sigma_2^2)$. Να προσεχθεί το πρόσημο (+) στις διασπορές.

Παρατήρηση: Θα συμβολίζουμε με $\overline{X}$ το μέσο όρο πληθυσμού.

Β. ΑΣΚΗΣΕΙΣ

Άσκηση 1.1

Δίνονται τα παρακάτω δεδομένα:

2,29	1,95	1,52	1,93	1,59	1,68	2,79	1,15
1,70	1,75	1,65	1,01	1,55	1,35	1,48	2,14

Να ομαδοποιηθούν τα δεδομένα επιλέγοντας αριθμό κλάσεων. Να βρεθούν ο δειγματικός μέσος, η διασπορά, η κορυφή, η διάμεσος, η 1η και 3η τεταρτητόμος.

Λύση:

Για την επιλογή κατάλληλου αριθμού κλάσεων (διαμερίσεων) εφαρμόζουμε τον τύπο: $q = 1 + 3{,}32\log_{10} 16 \approx 5$ κλάσεις. Ελάχιστο στοιχείο το 1,01 και μέγιστο το 2,79. Η διαφορά που πρέπει να καλυφθεί μεταξύ τους με 5 μόνο διαμερίσεις είναι 1,78 που δια του 5 δίνει 0,356 ή 0,36 (το βήμα). Άρα θα αρχίσουμε από το 1,00 και θα προχωρούμε με 0,36 βήμα δημιουργώντας 5 κλάσεις.

i	**Κάτω όριο**	**Άνω όριο**	**y_i**	**v_i**	**N_i**
1	1,00	1,36	1,18	3	3
2	1,36	1,72	1,54	7	10
3	1,72	2,08	1,90	3	13
4	2,08	2,44	2,26	2	15
5	2,44	2,80	2,62	1	16

Η στήλη y_i μας δίνει το ημιάθροισμα κάτω και άνω ορίου. Η στήλη v_i δείχνει πόσα εκ των δεδομένων βρίσκονται εντός του εκάστοτε διαστήματος. Η τελευταία στήλη N_i μας δίνει πόσα στοιχεία έχουν ήδη τοποθετηθεί στις εκάστοτε κλάσεις.

Για το δειγματικό μέσο έχουμε:

$$\bar{x} = \frac{1}{n}\sum x_i = \frac{1}{16}(2{,}29 + 1{,}95 + 1{,}52 + \ldots + 2{,}14) = 1{,}721$$

Μια καλή προσέγγιση του δειγματικού μέσου δίνεται εκ του τύπου:

$$\bar{x} = \frac{1}{n}\sum y_i \nu_{\mathrm{i}} = \frac{1}{16}(1{,}18 * 3 + 1{,}54 * 7 + 1{,}90 * 3 + 2{,}26 * 2 + 2{,}62 * 1) = 1{,}698$$

Για τη δε διασπορά θα έχουμε:

$$s^2 = \frac{1}{(n-1)}\left[\sum_{i=1}^{k} y_i^2 \nu_{\mathrm{i}} - \frac{1}{n}\left(\sum_{i=1}^{k} y_i \nu_{\mathrm{i}}\right)^2\right] = \frac{1}{15}(48{,}688 - \frac{1}{16}(27{,}16)^2) = 0{,}17$$

και επομένως s=0,41 (η τυπική απόκλιση).

Κορυφή έχουμε στο i=2 γιατί εκεί εμφανίζονται τα περισσότερα από τα δεδομένα, την τιμή 1,54 που ισούται με το μέσο της διαμέρισης.

Διάμεσος είναι η παρατήρηση εκείνη η οποία είναι μεγαλύτερη από το ½ ακριβώς του αριθμού των παρατηρήσεων. Αν έχουμε περιττό αριθμό δεδομένων διάμεσος είναι το μεσαίο δεδομένο. Αν έχουμε άρτιο όπως εδώ, διάμεσος είναι το ημιάθροισμα του 8ου και 9ου δεδομένου.

Προσεγγίζουμε τη διάμεσο από τον εξής τύπο:

$d = L_i + \dfrac{\frac{n}{2} - N_{i-1}}{N_i} c$, όπου c το πλάτος (βήμα) των κλάσεων, n το πλήθος, L_i το κάτω όριο και i η κορυφή. Εδώ d=1,36+(8-3)*0,36/7=1,617.

Την πρώτη και τρίτη τεταρτητόμο τις προσεγγίζουμε με τον ίδιο τύπο αλλά αντί α=1/2 θα βάλουμε α=1/4 και α=3/4 διαδοχικά :

$d = L_i + \frac{an - N_{i-1}}{N_i} c$, όπου α το ποσοστό (0,25 - 0,5 - 0,75) . Παρατηρούμε ότι για ποσοστό 0,50 (δεύτερη τεταρτητόμο) έχουμε τη διάμεσο. Άρα η διάμεσος είναι ουσιαστικά η δεύτερη τεταρτητόμος. Εδώ

$$Q_1 = 1{,}36 + \frac{4-3}{7} 0{,}36 = 1{,}41 \quad \& \quad Q_3 = 1{,}72 + \frac{12-10}{3} 0{,}36 = 1{,}95.$$

Άσκηση 1.2

Δίδονται $x_1, x_2, \ldots, x_{25}$ και $y_1, y_2, \ldots, y_{25}$ δυο ανεξάρτητα τυχαία δείγματα με Y~N(1,9) και X~N(0,16) . Να βρείτε την πιθανότητα $P(\bar{x} > \bar{y})$.

Λύση:

Αφού Y~N(1,9) $\Rightarrow$ $\bar{Y} \sim (1, \frac{9}{25})$ και $\bar{X} \sim (0, \frac{16}{25})$. Με αφαίρεση κατά μέλη προκύπτει η σχέση $\bar{X} - \bar{Y}$~(-1,1) γιατί ισχύει ως γνωστόν $\bar{X} - \bar{Y} \sim (\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$. Τώρα έχουμε διαδοχικά: $P(\bar{X} > \bar{Y}) = P(\bar{X} - \bar{Y} > 0) = P\left(\frac{\bar{X} - \bar{Y} - (-1)}{1} > \frac{0 - (-1)}{1} \right) = P(Z > 1) =$

$= 1 - P(Z \leq 1) = 1 - F(1) = 1 - 0{,}84134 = 0{,}15866.$

Άσκηση 1.3

Έστω X~N(μ,100). Ποιο το n ώστε P(μ-5< $\bar{x}$ <μ+5)=0,954;

Λύση:

Κατά τα γνωστά έχουμε ένα δείγμα $x_1, x_2, \ldots, x_n$ οπότε X~ $(\mu, \frac{100}{n})$. Είναι:

$$P(\mu - 5 < \bar{x} < \mu + 5) = P(-5 < x - \mu < 5) = P(\frac{-5}{\sqrt{100/n}} < \frac{x-\mu}{\sqrt{100/n}} < \frac{5}{\sqrt{100/n}}) =$$

$$= P(\frac{-5}{\sqrt{100/n}} < Z < \frac{5}{\sqrt{100/n}}) = F(\frac{5}{\sqrt{100/n}}) - F(\frac{-5}{\sqrt{100/n}}) = F(\frac{5}{\sqrt{100/n}}) - 1 +$$

$$F(\frac{5}{\sqrt{100/n}}) = 2F(\frac{5}{\sqrt{100/n}}) - 1 = 0{,}954 \Rightarrow F(\frac{5}{\sqrt{100/n}}) = 0{,}977 \Rightarrow F(0{,}5\sqrt{n}) = 0{,}977 \Rightarrow$$

$$\Rightarrow 0{,}5\sqrt{n} = 2{,}0 \Rightarrow \sqrt{n} = 4 \Rightarrow n = 16.$$

Σημειώνουμε ότι για τις τιμές των F χρησιμοποιήσαμε το γνωστό πίνακα που μας δίνει τα ολοκληρώματα για την κανονική κατανομή.

Άσκηση 1.4

Να βρεθεί η πιθανότητα η μέση διάρκεια ζωής των λαμπτήρων του εργοστασίου Α να είναι μεγαλύτερη από αυτή του Β κατά 160 ώρες.

	Μ.Ο	**Διασπορά**	**πλήθος**
Α	1400	200	125
Β	1200	100	125

Λύση:

Έχουμε από τα δεδομένα $\overline{T}_A \sim (1400, \frac{200^2}{125})$ και $\overline{T}_B \sim (1200, \frac{100^2}{125})$. Είναι σαφές ότι με αφαίρεση κατά μέλη παίρνουμε: $\overline{T}_A - \overline{T}_B \sim (200,400)$. Θέλουμε σύμφωνα με τα δεδομένα της εκφώνησης να είναι:

$$P(\overline{T}_A - \overline{T}_B > 160) = P\left(\frac{\overline{T}_A - \overline{T}_B - 200}{\sqrt{400}} > \frac{160-200}{\sqrt{400}}\right) = P(Z > -2) = 1 - P(Z \le -2).$$

Δηλαδή θα έχουμε P=1-P(Z≤-2)=1-F(-2)=F(2)=0,9772.

Κεφάλαιο 2ο: Εκτιμητική

Α. ΘΕΩΡΙΑ

2.1 Γενικά περί της εκτιμητικής

Στο κεφάλαιο αυτό θα ασχοληθούμε με μια σειρά στατιστικών προβλημάτων που αφορούν την εκτίμηση μιας ή περισσοτέρων παραμέτρων της κατανομής ενός ποσοτικού χαρακτηριστικού Χ των στοιχείων του πληθυσμού που μελετάμε.

Οι μέθοδοι εκτίμησης θα είναι δυο. Ο πρώτος, εξεταζόμενος στο κεφάλαιο αυτό, είναι η εκτίμηση κατά σημείο. Θα γίνεται με δυο μεθόδους, με τη μέθοδο των ροπών και με τη μέθοδο της μέγιστης πιθανοφάνειας. Ο δεύτερος τρόπος εκτίμησης που εξετάζεται στο επόμενο κεφάλαιο είναι η εκτίμηση κατά διάστημα.

2.2 Κριτήρια επιλογής εκτιμήτριας

Θα ονομάζουμε εκτιμήτρια συνάρτηση ή απλά εκτιμήτρια, τη συνάρτηση που μας δίνει εκτίμηση κάποιας παραμέτρου θ ενός τυχαίου δείγματος. Τα κριτήρια επιλογής της πρέπει να είναι τα παρακάτω:

- Αμεροληψία
- Ελάχιστο μέσο τετραγωνικό σφάλμα
- Αποτελεσματικότητα
- Επάρκεια
- Συνέπεια ή Σύγκλιση
- Βέλτιστη ασυμπτωτική κανονικότητα

Αυτό που εμείς πρακτικά θα χρησιμοποιούμε για να χαρακτηρίζουμε αμερόληπτη μια εκτιμήτρια θα είναι το παρακάτω κριτήριο:

$$\boxed{E[T(X)] = \theta}$$

όπου Τ(Χ) η εκτιμήτρια και θ το χαρακτηριστικό. Δηλαδή η μέση τιμή πρέπει να ισούται με το χαρακτηριστικό. Αν για παράδειγμα ελέγχουμε για την κατανομή Poisson την παράμετρο λ με μια συνάρτηση Τ(Χ), θα πρέπει Ε(Τ(Χ))=λ.

2.3 Μέθοδος Των Ροπών

Κατά τη μέθοδο αυτή ενεργούμε ως εξής. Έστω $X_1, X_2, \ldots, X_\nu$ τυχαίο δείγμα. Είναι πρόδηλο ότι ισχύουν οι παρακάτω σχέσεις:

$m_1=E(X)$, $m_2=E(X^2)$, ….., $m_\kappa=E(X^K)$ (*)

Επίσης θα ισχύουν οι σχέσεις:

$$m_1' = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \quad m_2' = \frac{1}{n}\sum_{i=1}^{n} x_i^2\,, \ \ldots . \quad , m_k' = \frac{1}{n}\sum_{i=1}^{n} x_i^k \quad (**)$$

Εξισώνοντας τώρα τις σχέσεις τις κλάσης (*) με της (**) παίρνουμε ένα σύστημα $m_1=m_1'$, $m_2=m_2'$,… το οποίο επιλύουμε για να βρούμε τις εκάστοτε εκτιμήτριες. Το πλήθος των εξισώσεων που θα χρειαστούμε εξαρτάται από το πρόβλημα.

2.4 Μέθοδος Μέγιστης Πιθανοφάνειας

Ας υποθέσουμε ότι έχουμε $X_1, X_2, \ldots, X_\nu$ τυχαίο δείγμα. Με τη βοήθεια της συναρτήσεως που αντιστοιχεί στην κατανομή που ακολουθεί το δείγμα, έστω f(x) προσπαθούμε να σχηματίζουμε την πιθανοφάνεια L(P) ως το γινόμενο $f(x_1)$, $f(x_2)$, … δηλαδή θα είναι $L(P)=f(x_1)\, f(x_2) \ldots f(x_n)$. Στη συνέχεια λογαριθμίζουμε (νεπέριος λογάριθμος) ως προς x την παραπάνω έκφραση της L(P) και έστω Ln[L(P)] αυτό που προκύπτει.

Η μέθοδος της μέγιστης πιθανοφάνειας όπως αποκαλύπτει και ο τίτλος της, συνίσταται στην εύρεση του μέγιστου (τοπικό ακρότατο) της παραπάνω έκφρασης Ln[L(P)]. Αυτό μπορεί να γίνει παραγωγίζοντας και θέτοντας ότι προκύψει ίσο με το μηδέν. Εξάλλου για να βεβαιωθούμε περί του τοπικού μέγιστου (μπορεί να έχουμε τοπικό ελάχιστο), ελέγχουμε αν η δεύτερη παράγωγος είναι αρνητική. Αν δεν είναι τότε έχουμε τοπικό ελάχιστο και άρα είναι άκυρα τα όσα κάναμε.

Υπάρχει βέβαια και η περίπτωση να μην μπορούμε να βρούμε το μέγιστο μιας συνάρτησης αν τη θέσομε με το μηδέν. Ας υποθέσουμε ότι η παράγωγος είναι η $f'(x)=a/x$. Προφανώς όταν $x \to \infty$ έχουμε $f(x)=0$ και μόνο τότε. Παρόλα αυτά η συνάρτηση δύναται να έχει τοπικό μέγιστο, αν είναι ορισμένη σε ένα διάστημα. Έτσι αν οι παράγωγοι δεν μας δίνουν το μέγιστο πρέπει να καταφεύγουμε σε γραφική παράσταση. Αν και τότε δεν προκύψει κάτι, σταματάμε.

Β. ΑΣΚΗΣΕΙΣ

Άσκηση 2.1

Αν X_1, X_2, X_3 τυχαίο δείγμα P(λ) (Poisson) να δείξετε ότι οι $\hat{\lambda}_1 = \dfrac{X_1 + 2X_2 + 3X_3}{6}$ και $\hat{\lambda}_2 = \dfrac{X_1 + X_2 + X_3}{3}$ είναι αμερόληπτες και να συγκρίνετε τις διασπορές.

Λύση:

Θα έχουμε:

$$E(\hat{\lambda}_1) = \frac{1}{6}E(X_1) + \frac{2}{6}E(X_2) + \frac{3}{6}E(X_3) = \frac{1}{6}\lambda + \frac{2}{6}\lambda + \frac{3}{6}\lambda = \frac{6\lambda}{6} = \lambda$$

$$E(\hat{\lambda}_2) = \frac{1}{3}E(X_1) + \frac{1}{3}E(X_2) + \frac{1}{3}E(X_3) = \frac{1}{3}\lambda + \frac{1}{3}\lambda + \frac{1}{3}\lambda = \frac{3\lambda}{3} = \lambda$$

που σημαίνει ότι είναι αμερόληπτες. Θυμίζουμε ότι στην κατανομή Poisson E(X)=λ. Όσον αφορά τις διασπορές θα είναι:

$$Var(\hat{\lambda}_1) = \left(\frac{1}{6}\right)^2 E(X_1) + \left(\frac{2}{6}\right)^2 E(X_2) + \left(\frac{3}{6}\right)^2 E(X_3) = \frac{1+4+9}{36}\lambda = \frac{14}{36}\lambda$$

$$Var(\hat{\lambda}_2) = \left(\frac{1}{3}\right)^2 E(X_1) + \left(\frac{1}{3}\right)^2 E(X_2) + \left(\frac{1}{3}\right)^2 E(X_3) = \frac{3}{9}\lambda = \frac{\lambda}{3}$$

Είναι σαφές ότι $Var(\hat{\lambda}_1) > Var(\hat{\lambda}_2)$. Πρέπει να κάνομε μια σημαντική παρατήρηση. Όταν μια εκτιμήτρια είναι αμερόληπτη, το μέσο τετραγωνικό της σφάλμα συμπίπτει με τη διασπορά της. Είναι ευνόητο ότι μεταξύ δυο εκτιμητριών που έχουν ίδιες μεροληψίες (όπως και οι παραπάνω) θα πρέπει να επιλέγεται εκείνη με το μικρότερο μέσο τετραγωνικό σφάλμα ή ισοδύναμα με τη μικρότερη διασπορά. Έτσι θα πρέπει να επιλέξουμε τη <u>δεύτερη</u> εκτιμήτρια καθόσον έχει τη μικρότερη διασπορά.

Άσκηση 2.2

Έστω $X_1,X_2,\ldots,X_n$ τυχαίο δείγμα με $\delta(\theta)=\begin{cases}\dfrac{1}{\theta}, & 0\le x\le\theta\\ 0, & \alpha\lambda\lambda o\upsilon\end{cases}$. Να βρεθεί η εκτιμήτρια με τη μέθοδο των ροπών.

Λύση:

Θα κάνουμε μια απλή εφαρμογή του τύπου που μας παρέχει την εκτιμήτρια με τη μέθοδο των ροπών:

$$m_1=E(X)=\int_0^{\theta} x\frac{1}{\theta}dx=\frac{\theta^2}{2}\frac{1}{\theta}=\frac{\theta}{2} \quad \kappa\alpha\iota \quad \mathrm{m}_1'=\overline{x}\left(=\frac{1}{n}\sum_{i=1}^{n}x_i\right)\Rightarrow \overline{x}=\frac{\hat{\theta}}{2}\Rightarrow\hat{\theta}=2\overline{x}\,.$$

Για την εύρεση της εκτιμήτριας πήραμε κατά τα γνωστά την εξίσωση m_1 και την m_1'. Τις εξισώσαμε και βρήκαμε το θ. Πρέπει να παρατηρήσουμε ότι $m_1'=\overline{x}$.

Άσκηση 2.3

Έστω τυχαίο δείγμα $X_1,X_2,\ldots,X_3$ ομοιόμορφης κατανομής με U(a,b). Να βρείτε τα $\hat{a},\hat{b}=$;

Λύση:

Αφού έχουμε ομοιόμορφη κατανομή είναι γνωστό ότι: $f(x)=\begin{cases}\dfrac{1}{b-a}, & a\le x\le b\\ 0, & \alpha\lambda\lambda o\upsilon\end{cases}$.

Στην ομοιόμορφη κατανομή ισχύουν για τη μέση τιμή και τη διασπορά αντίστοιχα:

$E(X)=\dfrac{a+b}{2},Var(X)=\dfrac{(b-a)^2}{12}$. Από την πρώτη εξίσωση έχουμε $m_1=m_1'\Rightarrow E(X)=\overline{x}\Rightarrow\Rightarrow$ $a+b=2\overline{x}$ (1) . Επειδή όμως έχουμε δυο αγνώστους a και b και μια εξίσωση, θα χρειαστούμε άλλη μια σχέση. Αυτή θα είναι η δεύτερη εξίσωση $m_2=m_2'\Rightarrow$ $E(X^2)=\sum_{i=1}^{n}x_i^2$. Συνεπώς θα είναι $Var(X)=E(X^2)-E(X)^2 \Rightarrow \frac{(b-a)^2}{12}=\frac{1}{n}\sum_{i=1}^{n}x_i^2-\overline{x}^2$ (2) Η εξίσωση (2) μπορεί όμως να γραφεί: $b-a=\sqrt{12\frac{1}{n}\sum_{i=1}^{n}x_i^2-12\overline{x}^2}$ (3) Οι (1) και (3) με προσθαφαίρεση δίνουν:

$\hat{a}=\overline{x}-\sqrt{\frac{3}{n}\sum_{i=1}^{n}x_i{}^2-3\overline{x}^2}$ και συμμετρικά $b=\overline{x}+\sqrt{\frac{3}{n}\sum_{i=1}^{n}x_i^2-3\overline{x}^2}$

Στην άσκηση αυτή χρειαστήκαμε τις δυο πρώτες εξισώσεις. Πρέπει να παρατηρήσουμε ότι από τη μετάβαση από την (2) στην σχέση (3) δεν βάλαμε απόλυτο. Αυτό έγινε επειδή γνωρίζαμε εκ των προτέρων ότι b>α λόγω του διαστήματος (α,b).

Άσκηση 2.4

Έστω $X_1,X_2,\ldots,X_n$ τυχαίο δείγμα που ακολουθεί κανονική κατανομή $N(\mu,\sigma^2)$. Να βρεθούν με τη μέθοδο των ροπών τα $\hat{\mu}=$; και $\hat{\sigma}=$;

Λύση:

Είναι προφανώς $m_1'=E(X)=\mu=\frac{1}{n}\sum_{i=1}^{n}x_i=\overline{x}$ και αντίστοιχα $m_2'=\frac{1}{n}\sum_{i=1}^{n}x_i^2$, οι δυο πρώτες δειγματικές ροπές περί την αρχή. Αυτά ισχύουν γενικά για όλες τις κατανομές. Ιδιαίτερα για την κανονική όμως κατανομή, γνωρίζουμε πως $m_1=E(X)=\mu$ και $m_2=E(X^2)=\sigma^2+\mu^2$. Θα εξισώσουμε στη συνέχεια κατά τα γνωστά τα $m_1'=m_1$ και $m_2'=m_2$. Με τον τρόπο αυτό λαμβάνουμε: $\mu=\overline{x},\quad \sigma^2+\mu^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2$. Αν τώρα λύσουμε το σύστημα αυτό ως προς μ και σ παίρνουμε τα ζητούμενα: $\hat{\mu}=\overline{x},\quad \hat{\sigma}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2-\overline{x}^2}$.(Το $\overline{x}$ δεν βρίσκεται εντός του αθροίσματος).

Άσκηση 2.5

Έστω $X_1,X_2,\ldots,X_n$ τυχαίο δείγμα που ακολουθεί ομοιόμορφη κατανομή με $U(-\theta,\theta)$. Να βρείτε το $\hat{\theta}=$;

Λύση:

Η μέση τιμή θα είναι E(X)=0 προφανώς. Θα υπολογίσω τη διασπορά:

$$Var(X)=\frac{\left(\theta-(-\theta)\right)^2}{12}=\frac{4\theta^2}{12}=\frac{\theta^2}{3}\Rightarrow Var(X)=E(X^2)-E(X)=E(X^2)-0=\frac{1}{n}\sum_{i=1}^{n}x_i^2\Rightarrow$$

$$\Rightarrow\hat{\theta}=\sqrt{\frac{3}{n}\sum_{i=1}^{n}x_i^2}\ .$$

<u>Παρατήρηση:</u> Δε θα μπορούσαμε να είχαμε πάρει τη μέση τιμή καθόσον αυτή είναι ίση με το μηδέν.

Άσκηση 2.6

Γεγονός Α με πιθανότητα ρ για ν=100 επαναλήψεις, εμφανίστηκε κ=63 φορές. Να βρεθεί η εκτιμήτρια μέγιστης πιθανοφάνειας (ΕΜΠ) $\hat{p}=$;

Λύση:

Η μεθοδολογία των ασκήσεων αυτών είναι στάνταρ. Αρχικά και σύμφωνα με την εκφώνηση θα βρίσκουμε τη συνάρτηση f(x) που θα χρησιμοποιήσουμε και θα προκύπτει από το είδος της κατανομής. Ύστερα θα εκφράζουμε την L(xx) την οποία θα λογαριθμίζουμε, θα παραγωγίζουμε προσπαθώντας να βρούμε το ακρότατο της, το οποίο και θα μας δίνει την εκτιμήτρια.

Εδώ αν πούμε $X=\begin{cases}1, p\\ 0, 1-p\end{cases}$ όπου 1=εμφάνιση τότε $f(x)=p^x(1-p)^{(1-x)}, x=0,1.$

α) Θα εκφράσουμε την L(p) με τη βοήθεια της παραπάνω συνάρτησης:

$$L(P)=p^{X_1}(1-p)^{1-X_1}p^{X_2}(1-p)^{X_2}.....p^{X_n}(1-p)^{X_n}=p^{\sum X_i}(1-p)^{n-\sum X_i}$$

β) Θα τη λογαριθμίσουμε (ως προς P μεταβλητή):

$$\ln L(P) = \sum X_i \ln P + (n - \sum X_i)(1-p)$$

γ) Θα την παραγωγίσουμε ως προς P και ό,τι προκύψει, θα το εξισώσουμε με το μηδέν για να προκύψει η ζητούμενη εκτιμήτρια συνάρτηση:

$$\frac{\partial \ln L(P)}{\partial P} = \frac{\sum X_i}{P} - \frac{n - \sum X_i}{1-p} = 0 \Rightarrow \hat{p} = \frac{\sum X_i}{n} = 0{,}63 = 63\%.$$

Άσκηση 2.7

Δίνεται το τυχαίο δείγμα $X_1, X_2, \ldots, X_n$ με πιθανότητα P(λ) που ακολουθεί κατανομή Poisson. Βρείτε την εκτιμήτρια μέγιστης πιθανοφάνειας $\hat{\lambda}$ =;

Λύση:

Θα εφαρμόσομε τη γνωστή μεθοδολογία (βλέπε άσκηση 2.6). Έτσι, η συνάρτηση που θα χρησιμοποιήσουμε αφού πρόκειται για κατανομή Poisson είναι η $f(x) = e^{-\lambda} \frac{\lambda^X}{X!}$. Η μεταβλητή στην προκειμένη περίπτωση θα είναι η X ενώ η εκτιμήτρια η λ.

α) Εκφράζω την L(λ) χρησιμοποιώντας τη συνάρτηση f(x).

$$L(\lambda) = e^{-\lambda} \frac{\lambda^{X_1}}{X_1!} e^{-\lambda} \frac{\lambda^{X_2}}{X_2!} \ldots\ldots e^{-\lambda} \frac{\lambda^{X_n}}{X_n!} = e^{-n\lambda} \frac{\lambda^{\sum X_i}}{X_1! X_2! \ldots X_n!}$$

β) Θα την λογαριθμίσω ως προς λ , χρησιμοποιώντας τις σχέσεις Ln(ab)=Lna+Lnb και την Ln(a/b)=Lna - Lnb.

$$LnL(\lambda) = \ln e^{-n\lambda} + \ln \lambda^{\sum X_i} - \ln(X_1! X_2! \ldots. X_n!) = -n\lambda + \sum X_i \ln \lambda - (\ln X_1! + \ln X_2! + \ldots.$$
$$\ldots \ln X_n!) = -n\lambda + \sum X_i \ln \lambda - \sum \ln X_i! = -n\lambda + n\bar{x} \ln \lambda - \sum \ln X_i!$$

αφού ως γνωστόν $\bar{x} = \frac{1}{n} \sum X_i \Rightarrow \sum X_i = n\bar{x}.$

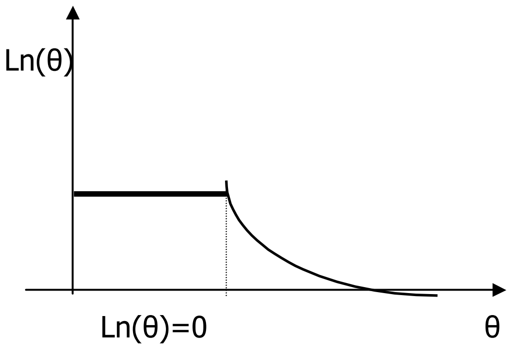
γ) Τέλος θα την παραγωγίσω ως προς λ και ότι προκύψει θα το εξισώσω με το μηδέν για να βρω τη ζητούμενη εκτιμήτρια:

$\frac{\partial LnL(\lambda)}{\partial \lambda} = -n + \frac{n\bar{x}}{\lambda} - 0 = 0 \Rightarrow \hat{\lambda} = \bar{x}$, που είναι η ζητούμενη εκτιμήτρια.

Άσκηση 2.8

Δίδεται τυχαίο δείγμα $X_1, X_2, \ldots, X_n$ που ακολουθεί ομοιόμορφη κατανομή U(0,θ). Να βρείτε την εκτιμήτρια $\hat{\theta}$ =;

Λύση:



Θα ενεργήσουμε κατά τα γνωστά, λαμβάνοντας υπόψη μας, πως λόγω της ομοιόμορφης κατανομής, η συνάρτηση f(x) που θα μας βοηθήσει στη σύνταξη της έκφρασης L(θ) είναι η $f(x) = \begin{cases} \frac{1}{\theta}, 0 \le X_1 \le X_2 \le \ldots \le X_n \le \theta \\ 0, \alpha\lambda\lambda o\upsilon \end{cases}$. Τότε θα έχουμε διαδοχικά:

$L(\theta) = \frac{1}{\theta}\frac{1}{\theta}\ldots\frac{1}{\theta} = \frac{1}{\theta^n}$ γιατί η f(x) είναι ουσιαστικά σταθερή συνάρτηση. Λογαριθμίζοντας παίρνουμε $LnL(\theta) = -Ln\theta^n = -nLn\theta$ και παραγωγίζοντας ως προς θ λαμβάνουμε $\frac{\partial LnL(\theta)}{\partial \theta} = -\frac{n}{\theta} \ne 0, \forall \theta \in R$. Σε τέτοιες περιπτώσεις ανατρέχουμε στη γραφική παράσταση που δίδεται από το σχήμα παραπάνω. Βλέπουμε ότι έχουμε ακρότατο (που είναι ως γνωστόν η ζητούμενη τιμή της εκτιμήτριας) $\hat{\theta} = X_{(n)}$.

<u>Παρατήρηση:</u> Η μέθοδος αυτή καλείται μέθοδος μέγιστης πιθανοφάνειας και η ποσότητα L(q) που υπολογίσαμε στις τρεις προηγούμενες ασκήσεις καλείται πιθανοφάνεια. Η άλλη μέθοδος της εκτιμητικής που εξετάσαμε στις ασκήσεις 2.2, 2.3, 2.4 & 2.5 καλείται μέθοδος των ροπών. Εκτός από τις δυο αυτές μεθόδους, δε θα εξετάσουμε καμία άλλη.

Άσκηση 2.9

Παίρνουμε τυχαίο δείγμα μεγέθους n από πληθυσμό, ο οποίος ακολουθεί την εκθετική κατανομή με παράμετρο θ (θ>0). Να βρεθεί η εκτιμήτρια μέγιστης πιθανοφάνειας της παραμέτρου θ. (*Θέμα 6/1996)*

Λύση:

Κατά τα γνωστά θα βρούμε την έκφραση της πιθανοφάνειας L(θ):

$$L(\theta)=\theta e^{-\theta x_1}\theta e^{-\theta x_2}\ldots\theta e^{-\theta x_n}=\theta^n e^{-\theta\sum x_i}$$

θα λογαριθμίσουμε την παραπάνω σχέση (νεπέριος λογάριθμος - θ η μεταβλητή) οπότε:

$\ln l(\theta)=n\ln\theta-\theta\sum x_i$, γιατί Lnab=Lna + Lnb και θα παραγωγίσουμε ως προς θ:

$\dfrac{\partial \ln l(\theta)}{\partial\theta}=\dfrac{n}{\theta}-\sum x_i=0\Rightarrow\hat{\theta}=\dfrac{n}{\sum x_i}=\dfrac{1}{\overline{x}}$. Είμαστε σίγουροι ότι έχουμε τοπικό μέγιστο γιατί αν πάρουμε δεύτερη παράγωγο, θα βγει $-n/\theta^2<0$.

Κεφάλαιο 3ο: Διαστήματα Εμπιστοσύνης

Α. ΘΕΩΡΙΑ

3.1 Γενικά Περί Δ.Ε.

Εκτός από τον προσδιορισμό της τιμής μιας παραμέτρου θ, πρόβλημα με το οποίο ασχοληθήκαμε στο προηγούμενο κεφάλαιο, συχνά είναι χρήσιμο να προσδιορίσουμε, μέσω των εκτιμητριών, ένα διάστημα το οποίο θα περιέχει την άγνωστη τιμή της παραμέτρου με καθορισμένη πιθανότητα έστω γ. Ένα τέτοιο διάστημα ονομάζεται διάστημα εμπιστοσύνης.

Η πιθανότητα γ θα ονομάζεται βαθμός εμπιστοσύνης και πρακτικά θα είναι της τάξης του 90% ή μεγαλύτερος. Όταν λοιπόν ένα πρόβλημα λέει για 99% σιγουριά, τότε αυτόματα θα είναι ο βαθμός εμπιστοσύνης 99% του διαστήματος που θα δημιουργήσουμε. Αυτό που αρχικά πρέπει να βρίσκουμε σε προβλήματα τέτοιου είδους (αν δε δίνεται) είναι η στάθμη σημαντικότητας α.

Πράγματι αν έχουμε διάστημα εμπιστοσύνης 95% = 0,95 $\Rightarrow$ 1-α = 0,95 $\Rightarrow$ α=0,05. Αλλά και αντίστροφα, αν δίνεται α=0,01 $\Rightarrow$ 1-α = 0,99 $\Rightarrow$ 99% διάστημα εμπιστοσύνης. Η στάθμη σημαντικότητας χρειάζεται στους πίνακες των κατανομών όπως θα δούμε παρακάτω.

3.2 Δ.Ε. της μέσης τιμής μ (σ γνωστό)

Αυτό θα δίνεται εκ του τύπου:

$$\boxed{\bar{x} \pm \frac{\sigma}{\sqrt{n}} Z_{a/2}}$$

όπου n το μέγεθος του δείγματος, α η στάθμη σημαντικότητας και σ η τυπική απόκλιση που είναι γνωστή.

3.3 Δ.Ε. της μέσης τιμής μ (σ άγνωστο)

Αντί του προηγούμενου τύπου θα έχουμε τώρα:

$$\boxed{\bar{x} \pm \frac{s}{\sqrt{n}} t_{n-1,a/2}}$$

όπου τώρα αντί σ=s και αντί $Z_{a/2} = t_{v-1,a/2}$ (t είναι της κατανομής Student).

Παράδειγμα: Για τον προσδιορισμό του συντελεστή θερμικής διαστολής μ του νικελίου έγιναν 25 μετρήσεις και έδωσαν μέσο όρο $\bar{x}$=12,81 και τυπική απόκλιση s=0,04. Να κατασκευαστεί δ.ε. του θερμικού συντελεστή 95%.

Θα είναι ασφαλώς δε 95%=0,95 $\Rightarrow$ 1-a=0,95 $\Rightarrow$ a=0,05 η στάθμη σημαντικότητας. Επειδή δε γνωρίζουμε το σ, θα εφαρμόσουμε το δεύτερο από τους δυο τύπους που μας παρέχουν το δ.ε. του μέσου όρου μ.

Πράγματι το πλήθος n=25 και απομένει να βρούμε το t. Το $t_{25,\ 0.025}$ (a/2=0,025) θα δοθεί από τον πίνακα t των ποσοστιαίων σημείων της κατανομής του Student. Στην οριζόντια στήλη θα έχουμε P=1-0,025=0,975 στη δε κατακόρυφη n=24 και επομένως θα πάρουμε t=2,064.

Πρέπει να τονίσουμε πως στην κατακόρυφη στήλη του πίνακα αυτού, ο παράγοντας v <u>δεν είναι το πλήθος του δείγματος</u> αλλά ο βαθμός ελευθερίας που προκύπτει αν από το μέγεθος του πληθυσμού αφαιρέσουμε τη μονάδα, γι’ αυτό και είπαμε n=25-1=24. Αυτό πρέπει να προσεχθεί ιδιαίτερα, καθώς ισχύει γενικά και για τους άλλους πίνακες.

$$12,81 \pm \frac{0,04}{\sqrt{25}} 2,064 = 12,81 \pm 0,02 = (12,79 \quad , \quad 12,83)$$

<u>Παρατήρηση:</u> Όσον αφορά τα δεκαδικά ψηφία των πράξεων, αυτά θα είναι όσα και του μέσου όρου. Μεγαλύτερη ακρίβεια προφανώς δεν έχει νόημα. Έτσι ενώ στις πράξεις έβγαινε 0,016512 η παράσταση μετά το 12,81, τη στρογγυλοποιήσαμε στο 0,02 για να έχουμε δυο δεκαδικά ψηφία αφού τόσα είναι και ο μέσος όρος. Ο κανόνας αυτός βέβαια δεν

είναι απόλυτος γιατί η έκφραση που ακολουθεί το ± μπορεί να ήταν η 0,0009 που θα στρογγυλευόταν στο 0.

3.4 Δ.Ε. της διαφοράς μ_1-μ_2 (σ_1,σ_2 γνωστά)

Έστω έχουμε δυο πληθυσμούς με μέσους όρους $\bar{x}, \bar{y}$ αντίστοιχα, διασπορές σ_1^2 και σ_2^2 αντίστοιχα και ακολουθούν κανονικές κατανομές $N(\mu_1,\sigma_1^2)$ και $N(\mu_2,\sigma_2^2)$. Τότε το διάστημα εμπιστοσύνης της διαφοράς των μέσων όρων με γνωστές τις τυπικές αποκλίσεις (άρα και τις διασπορές) θα δίνεται από τον τύπο:

$$\boxed{\bar{x} - \bar{y} \pm Z_{a/2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

n_1 και n_2 είναι προφανώς τα μεγέθη των πληθυσμών.

3.5 Δ.Ε. της διαφοράς μ_1-μ_2 ($\sigma_1=\sigma_2=\sigma$ με σ άγνωστο)

Στην ειδική περίπτωση που οι τυπικές αποκλίσεις είναι άγνωστες αλλά ίσες μεταξύ τους ο παραπάνω τύπος αντικαθίσταται από τον παρακάτω, ο οποίος ισχύει για εν γένει μικρά μεγέθη πληθυσμών ήτοι $n_1,n_2<30$:

$$\boxed{\bar{x} - \bar{y} \pm s * t_{n1+n2-2,a/2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

3.6 Δ.Ε. για τη διαφορά μ_1-μ_2 (σ_1^2, σ_2^2 άγνωστα)

Εδώ θα πρέπει τα δείγματα να είναι αρκετά μεγάλα, δηλαδή $n_1,n_2 \geq 30$.

$$\boxed{\bar{x} - \bar{y} \pm Z_{a/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

<u>Παράδειγμα:</u> Δυο αναλυτές Α και Β διεξήγαγαν μικροαναλυτικούς προσδιορισμούς της περιεκτικότητας σε άνθρακα ενός χημικού προϊόντος και πήραν τα κάτωθι αποτελέσματα:

A	47,2	47,5	47,3	47,7	47,6	47,4	47,5	47,8	47,3	47,4			
B	47,7	47,2	47,3	47,9	47,5	47,1	47,4	47,8	47,3	47,9	47	47,6	47,8

Αν οι δυο αναλυτές είναι το ίδιο έμπειροι και ακριβείς στις αναλύσεις τους, αν δηλαδή $\sigma_1=\sigma_2$, ποιο το 0,99 δ.ε. της μέσης διαφοράς στα αποτελέσματά τους;

Τα δείγματα είναι αρκετά μικρά ώστε να μπορεί να εφαρμοσθεί ο τύπος που μας δίνει το διάστημα εμπιστοσύνης με ίσες τυπικές αποκλίσεις (άρα και ίσες διασπορές). Πρέπει αρχικά να βρούμε τους μέσους όρους των μετρήσεων:

$$\bar{x} = \frac{1}{10}(47{,}2 + 47{,}5 + 47{,}3 + 47{,}7 + 47{,}6 + 47{,}4 + 47{,}5 + 47{,}8 + 47{,}3 + 47{,}4) = 47{,}47$$

$$\bar{y} = \frac{1}{13}(47{,}7 + 47{,}2 + 47{,}3 + 47{,}9 + 47{,}5 + 47{,}1 + 47{,}4 + 47{,}8 + 47{,}3 + 47{,}9 + 47 + 47{,}6 + 47{,}8) = 47{,}5$$

Θα βρούμε τις διασπορές s_1^2 και s_2^2 στη συνέχεια:

$$s_1^2 = \frac{1}{10-1}(0{,}0729 + 0{,}0009 + 0{,}0289 + 0{,}0529 + 0{,}0169 + 0{,}0049 + 0{,}0009 + 0{,}1089 +$$
$$+ 0{,}0289 + 0{,}0049) = 0{,}321/9 = 0{,}0356 = 0{,}036$$

$$s_2^2 = \frac{1}{13-1}(0{,}04 + 0{,}09 + 0{,}04 + 0{,}16 + 0 + 0{,}16 + 0{,}01 + 0{,}09 + 0{,}04 + 0{,}16 + 0{,}25 + 0{,}01 +$$
$$+ 0{,}09) = 1{,}14/12 = 0{,}095$$

Για να βρούμε το s έχουμε:

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{9*0{,}036 + 12*0{,}095}{10+13-2} = \frac{1{,}464}{21} = 0{,}0697$$

$\Rightarrow$ s^2=0,0697 $\Rightarrow$ s=0,26 οπότε αντικαθιστούμε στον τύπο και έχουμε:

$$\bar{x} - \bar{y} \pm s * t_{n1+n2-2,a/2}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 47{,}47 - 47{,}50 \pm 0{,}26 * 2{,}831\sqrt{\frac{1}{10} + \frac{1}{13}} = (-0{,}33 \quad , \quad 0{,}27)$$

αφού $t_{21,0.005}$ = 2,831 (P=0,995 και v=21 βαθμοί ελευθερίας).

3.7 Δ.Ε. για τη διασπορά $σ^2$

$$\boxed{\left(\frac{(n-1)s^2}{X^2_{n-1,a/2}},\frac{(n-1)s^2}{X^2_{n-1,1-a/2}}\right)}$$

Σημειώνεται ότι αν ένα πρόβλημα μας ζητά να βρούμε διάστημα εμπιστοσύνης για την τυπική απόκλιση, εμείς θα βρίσκουμε δ.ε. για τη διασπορά και στα άκρα του διαστήματος που θα βρίσκουμε θα βάζουμε τετραγωνικές ρίζες. Για παράδειγμα αν προκύψει διάστημα εμπιστοσύνης (16,25) για τη διασπορά, θα είναι (4,5) για την τυπική απόκλιση.

3.8 Δ.Ε. για το λόγο $σ_1^2/σ_2^2$

Ας υποθέσουμε ότι έχουμε δυο κανονικούς πληθυσμούς $N(μ_1,σ_1^2)$ και $N(μ_2,σ_2^2)$. Τότε το διάστημα εμπιστοσύνης για το λόγο των διασπορών τους θα δίνεται από το τύπο:

$$\boxed{\left(\frac{s_1^2}{s_2^2}\frac{1}{F_{n1-1,n2-1,a/2}},\frac{s_1^2}{s_2^2}\frac{1}{F_{n2-1,n1-1,1-a/2}}\right)}$$

<u>Παρατήρηση:</u> για τα ποσοστιαία σημεία της κατανομής F του Snedecor θα ισχύει η παρακάτω βασική ιδιότητα:

$$\boxed{\frac{1}{F_{n1-1,n2-1,a/2}}=F_{n2-1,n1-1,1-a/2}}$$

3.9 Εξαρτημένα δείγματα

Ας υποθέσουμε ότι έχουμε $X\sim N(μ_1,σ_1^2)$ τυχαίο δείγμα μεγέθους ν και $Y\sim N(μ_2,σ_2^2)$ τυχαίο δείγμα επίσης μεγέθους ν. Οι παρατηρήσεις X και Y συνδέονται μεταξύ τους με κάποια σχέση (δεν είναι δηλαδή ανεξάρτητες).

Τότε θα ισχύει για το διάστημα εμπιστοσύνης:

$$\boxed{\bar{z} \pm \frac{S_Z}{\sqrt{n}} t_{n-1,a/2}}$$

όπου $\bar{z} = \frac{1}{\nu}\sum z_i \quad , z_i = x_i - y_i$. Σημειώνεται επίσης ότι S_Z θα είναι η τυπική απόκλιση του Ζ.

Παράδειγμα: Δέκα μεταλλικοί σωλήνες με αντιδιαβρωτικό επίστρωμα Α τέθηκαν ανά ζεύγη με άλλους δέκα με αντιδιαβρωτικό επίστρωμα Β μέσα στο έδαφος, στο ίδιο βάθος και για το ίδιο χρονικό διάστημα. Ο βαθμός διάβρωσης που υπέστη κάθε σωλήνας εκφράστηκε από το μέγιστο βάθος που προχώρησε η διάβρωση. Με βάση τα παρακάτω δεδομένα να κατασκευαστεί ένα 0,90-δ.ε. της μέσης διαφοράς διαβρώσεων μ_X-μ_Y μεταξύ σωλήνων με αντιδιαβρωτικό επίστρωμα Α και σωλήνων με αντιδιαβρωτικό επίστρωμα Β.

Α	47	52	55	76	57	58	39	54	46	71
Β	45	42	46	69	58	54	40	46	42	53
Διαφορά	2	10	9	7	-1	4	-1	11	4	18

Αρχικά θα βρούμε τη γραμμή "διαφορά" $z_i=x_i-y_i$ και κατόπιν το μέσο όρο των Ζ και τη διασπορά. Έχουμε:

$$\bar{z} = \frac{1}{10}(2+10+9+7-1+4-1+11+4+18) = 6{,}3$$

$$s_z^{\ 2} = \frac{1}{10-1}(18{,}49+13{,}69+7{,}29+0{,}49+53{,}29+5{,}29+53{,}29+22{,}09+5{,}29+136{,}89) = 35{,}12$$

άρα s_z=5,93 και $t_{9,0.05}$=1,833 (δ.ε. 90%=0,90 $\Rightarrow$ 1-α=0,90 $\Rightarrow$ α=0,1 $\Rightarrow$ α/2 = 0,05). Άρα με απλή αντικατάσταση παίρνουμε:

$$6{,}3 \pm \frac{5{,}93}{\sqrt{10}} 1{,}833 = 6{,}3 \pm 3{,}4 = (2{,}9 \quad , \quad 9{,}7)$$

<u>Παρατήρηση:</u> Επειδή η τιμή 0 δεν περιέχεται στο παραπάνω δ.ε. μπορούμε να συμπεράνουμε ότι τα δυο επιστρώματα δεν έχουν την ίδια αντιδιαβρωτική ικανότητα. Αν την είχαν τότε $\mu_1=\mu_2 \Rightarrow \mu_1-\mu_2 = 0 \Rightarrow 0$ ανήκει στο δ.ε. άτοπο. Επιπλέον επειδή $\bar{z} = \bar{x} - \bar{y} = 6{,}3 > 0$ είναι φανερό ότι το αντιδιαβρωτικό Β είναι καλύτερο του Α.

3.10 Δ.Ε. για την αναλογία ρ (Διωνυμική Κατανομή)

Η διωνυμική κατανομή θα μας απασχολήσει εκτός από την κανονική που είδαμε μέχρι τώρα, στην εύρεση διαστημάτων εμπιστοσύνης. Απαραίτητη προϋπόθεση είναι η ύπαρξη σχετικά μεγάλων πληθυσμών (ν>30). Η πιθανότητα ρ θα δίνεται από το λόγο x/ν όπου ν το μέγεθος του πληθυσμού. Τέλος θα χρησιμοποιούμε τον παρακάτω τύπο:

$$\boxed{\hat{p} \pm Z_{a/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{\nu}}}$$

3.11 Δ.Ε. για τη διαφορά ρ_1-ρ_2 δυο αναλογιών

Αυτή θα δίδεται εκ του τύπου:

$$\hat{p}_1 - \hat{p}_2 \pm Z_{a/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{\nu_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{\nu_2}}$$

Απαραίτητη προϋπόθεση είναι να έχουμε σχετικά μεγάλα δείγματα, ήτοι $\nu_1,\nu_2 > 30$.

Παράδειγμα: Κατά τον έλεγχο ποιότητας στα προϊόντα μιας μηχανής βρέθηκαν 3 ελαττωματικά σε τυχαίο δείγμα μεγέθους ν=30 αντικειμένων. Να βρεθεί ένα διάστημα εμπιστοσύνης με σιγουριά 90% του ποσοστού ρ των ελαττωματικών προϊόντων της μηχανής.

Είμαστε στο όριο του ν>30 για την εφαρμογή του τύπου. Η πιθανότητα $\hat{p}$ όπως αναφέρθηκε πριν θα είναι ίση με $\hat{p} = \frac{x}{\nu} = \frac{3}{30} = 0,1$. Επίσης σιγουριά 90% $\Rightarrow$ δ.ε. 90%=0,90 άρα 1-α=0,90 $\Rightarrow$ α=0,10 $\Rightarrow$ α/2 = 0,05 και επομένως $Z_{0,05}$ = 1,64 (γιατί πήγαμε στο 0,94950 που αντιστοιχεί στην 1,6 κατακόρυφη στήλη και 0,04 οριζόντια άρα 1,6+0,04=1,64). Με απλή αντικατάσταση παίρνουμε:

$$0,1 \pm 1,64\sqrt{\frac{0,1(1-0,1)}{30}} = 0,1 \pm 0,09 = (0,01\ ,\ \ 0,19)\,.$$

Β. ΑΣΚΗΣΕΙΣ

Άσκηση 3.1

Έχω ένα δείγμα ν=36 χαλύβδινων ράβδων με $\bar{x} = 379{,}2$ και s=124. Να βρεθεί 95% διάστημα εμπιστοσύνης για τη μέση τιμή Ν του πληθυσμού των ράβδων.

Λύση:

Παρατηρούμε ότι ν>30 και σ^2 άγνωστο. Επομένως θα γίνει χρήση του τύπου:

$$\boxed{\bar{x} \pm \frac{s}{\sqrt{n}} t_{n-1,a/2}}$$

Θα έχουμε διαδοχικά: δ.ε.=95%=0,95 $\Rightarrow$ 1-α = 0,95 $\Rightarrow$ α=0,05 $\Rightarrow$ $t_{35,α/2} = t_{35,\ 0,025}$ (1) Αρκεί να βρούμε την τιμή του t. Αυτή την παίρνουμε έτοιμη. Ανατρέχομε στους πίνακες της κατανομής t του Student και επειδή δεν υπάρχει για ν=35 βαθμούς ελευθερίας τιμή του t, θα πάμε κάπου στη μέση (30 - 40) όπου και θα είναι t=2,03 (περίπου). Επομένως με αντικατάσταση στον τύπο του πλαισίου έχουμε:

$$\bar{x} \pm \frac{s}{\sqrt{n}} t_{n,a/2} = 379{,}2 \pm \frac{124}{\sqrt{36}} 2{,}03 = 379{,}2 \pm 41{,}95 = (337{,}25\ ,\ 421{,}15)$$ το ζητούμενο δ.ε.

<u>Παρατήρηση:</u> Αν υποθέσουμε ότι η διασπορά ήταν γνωστή και θέλαμε να έχουμε 99% σιγουριά, τότε θα παίρναμε διαδοχικά: Ζ=99%=0,99 $\Rightarrow$ 1-α=0,99 $\Rightarrow$ α=0,01 $\Rightarrow$ $Z_{α/2} = Z_{0,005}$ (1). Όμως από τους πίνακες της κανονικής κατανομής Φ(Ζ) παίρνουμε Z_1=2,5 και Z_2=0,08 άρα $Z=Z_1+Z_2$ = 2,5+0,08 = 2,58. (Σημειώνεται ότι θέλαμε Φ(Ζ)=1-0,005 = 0,99500 και πήγαμε στο Φ(Ζ)=0,99506). Άρα θα παίρναμε διάστημα εμπιστοσύνης (325,88 , 432,52). Όσο δηλαδή αυξάνεται η σιγουριά, τόσο ευρύτερο διάστημα απαιτείται.

Άσκηση 3.2

Μετρώ την αντοχή χαλύβδινων ράβδων και παίρνω τα παρακάτω:

173	166	168	166	169	166	173	170	170	173
166	161	166	170	168	158	173	166	165	165

Υποθέτω ότι ο πληθυσμός είναι κανονικός Ν(μ,σ^2). Βρείτε διάστημα εμπιστοσύνης 95% για τη μέση αντοχή.

Λύση:

Καταρχήν έχω 20 ράβδους και μέση αντοχή $\bar{x} = \frac{1}{20}\sum X_i = 167{,}60$.

Το δε s^2 θα υπολογιστεί από τον παρακάτω τύπο:

$$s^2 = \frac{1}{20-1}\sum_{i=1}^{20} X_i^2 - \frac{\left(\sum_{i=1}^{20} X_i\right)^2}{20} \Rightarrow s^2 = \frac{1}{19}\sum_{i=1}^{20} X_i^2 - \frac{\bar{x}^2}{20} \Rightarrow s^2 = 15{,}84 \Rightarrow s = 3{,}98$$

Τώρα μπορούμε πλέον να εφαρμόσουμε τον τύπο που θα μας δώσει το διάστημα εμπιστοσύνης:

$$\bar{x} \pm \frac{s}{\sqrt{n}} t_{19,0.025} = 167{,}60 \pm \frac{3{,}98}{\sqrt{20}} 2{,}093 = 167{,}60 \pm 1{,}86 \Rightarrow (165{,}74\ ,\ 169{,}46)$$

<u>Παρατήρηση:</u> Την τιμή του t την υπολογίσαμε από τον πίνακα που μας παρέχει τα ποσοστιαία σημεία της κατανομής t του Student. Πήγαμε στην κατακόρυφη στήλη στο v=20-1= 19 και στην οριζόντια στο P=1-0,025=0,975 και άρα $t_{19,0,025}$ = 2,093.

Άσκηση 3.3

Δίνονται οι τιμές των αντοχών που ελήφθησαν για 8 χαλύβδινες ράβδους.

520	520	530	518	522	525	527	519

Να βρείτε διάστημα εμπιστοσύνης 95% για το σ.

Λύση:

Πρέπει να προσέξουμε ότι ζητάει δ.ε. για το σ και όχι για τη διασπορά σ^2. Άρα αφού βρούμε τα άκρα του διαστήματος για τη διασπορά, θα χρησιμοποιήσουμε την τετραγωνική ρίζα για να βρούμε τα άκρα του διαστήματος για το σ.

Έχουμε ν=20 και προφανώς μπορούμε εύκολα να υπολογίσομε το s.

$$\bar{x} = \frac{1}{8}(520 + 520 + 530 + 518 + 522 + 525 + 527 + 519) = 522{,}625$$

$$s^2 = \frac{1}{8-1}\sum_{i=1}^{8}(x_i - \bar{x})^2 = \frac{1}{7}(13{,}78+54{,}39+21{,}39+0{,}39+5{,}64+4{,}38+13{,}14)=16{,}16 \Rightarrow s=4{,}02$$

Τώρα παίρνουμε: $\left(\frac{(8-1)4{,}02^2}{16{,}01}, \frac{(8-1)4{,}02^2}{1{,}690}\right) = (7{,}07\ ,\ 66{,}94)$ για το σ^2 και άρα το ζητούμενο διάστημα για το σ θα είναι το (2,66 , 8,18).

Παρατήρηση: Τα ζητούμενα $X^2_{7,0.025}$ και $X^2_{7,0.975}$ υπολογίστηκαν από τον πίνακα που μας παρέχει τα ποσοστιαία σημεία της κατανομής X^2. Για παράδειγμα για το πρώτο στην κατακόρυφη στήλη πήγαμε στο ν-1=8-1=7 και στην οριζόντια στο 0,025 άρα πήραμε 1,690. Εντελώς ανάλογα προέκυψε η τιμή 16,01 για το δεύτερο.

Άσκηση 3.4

Χαλύβδινες ράβδοι παράγονται από εργοστάσια Α και Β. Οι πληθυσμοί θεωρούνται κανονικοί. Μετράμε την αντοχή δείγματος ν_1=45 και ν_2=50.

A: $N(\mu_1,\sigma_1^2)$ ν_1=45, $\bar{x}$=76,8 και s_1=7,2

B: $N(\mu_2,\sigma_2^2)$ ν_2=50, $\bar{y}$=80,8 και s_2=7,0.

Να βρείτε διάστημα εμπιστοσύνης 98% για τη διαφορά μ_1-μ_2 των μέσων τιμών των δυο παραπάνω πληθυσμών.

Λύση:

Έχουμε δε 98% $\Rightarrow$ Z=98%=0,98 $\Rightarrow$ 1-α=0,98 $\Rightarrow$ α=0,02 $\Rightarrow$ $Z_{\alpha/2}$ = $Z_{0,01}$ = 2,33 (Z_1=2,3 και Z_2=0,03 πήγαμε δε στο 0,99010 $\approx$ 0,99). Επομένως θα έχουμε:

$$\bar{x}-\bar{y}\pm Z_{0,01}\sqrt{\frac{s_1^2}{\nu_1}+\frac{s_2^2}{\nu_2}}=76{,}8-80{,}8\pm 2{,}33\sqrt{\frac{7{,}2^2}{45}+\frac{7^2}{50}}=-4\pm 3{,}4=(-7{,}4\ ,\ \ -0{,}6)\,,$$

και γράφουμε -7,4 ≤ μ_1-μ_2 ≤ -0,6.

Άσκηση 3.5

Δίνονται τα δυο παρακάτω δείγματα κανονικών πληθυσμών. Ζητείται να βρείτε διάστημα εμπιστοσύνης 98% για τη διαφορά των μέσων τιμών των δυο πληθυσμών.

Δείγμα 1ο	44	44	56	46	47	38	58	53	49	35	46	30	41
Δείγμα 2ο	35	47	55	29	40	39	32	41	42	57	51	39	

Λύση:

Είναι προφανώς ν_1=13, $\bar{x}$ =45,15 και s_1^2=64 για το πρώτο δείγμα. Αντίστοιχα για το δεύτερο δείγμα θα είναι ν_2=12, $\bar{y}$ =42,45 και s_2^2=76,4. Θα βρούμε εν συνεχεία το s^2 από το γνωστό από τη θεωρία τύπο:

$$s^2=\frac{(\nu_1-1)s_1^2+(\nu_2-1)s_2^2}{\nu_1+\nu_2-2}=\frac{12*64+11*76{,}4}{12+13-2}=8{,}36$$

Επίσης από τον πίνακα t του Student παίρνουμε για δε=98% =0,98 $\Rightarrow$ 1-α=0,98 $\Rightarrow$ α=0,02 $\Rightarrow$ α/2=0,01 άρα $t_{23,0.01}$=2,518 (γιατί ν_1+ν_2-2=13+12-2=23).

Τελικά προκύπτει (-5,47 , 11,27). Μια σημαντική παρατήρηση που μπορούμε να κάνουμε εδώ, είναι ότι το μηδέν εμπεριέχεται στο διάστημα εμπιστοσύνης. Άρα δε χρειάζεται να πάρουμε και άλλες μετρήσεις.

Άσκηση 3.6

Μετρήθηκε η θερμοκρασία των δυο μπροστινών τροχών ενός αυτοκινήτου:

71	87	104	107	60	65	75	96	84	107	90	84	93	88	85	92	74
80	84	75	82	69	74	70	60	64	68	71	90	76	68	67	80	98

Να βρεθεί ένα διάστημα εμπιστοσύνης 99% για τη διαφορά μ_1-μ_2 των μέσων τιμών των πληθυσμών. Οι πληθυσμοί θεωρούνται κανονικοί.

Λύση:

Επειδή τα μεγέθη που μετρήθηκαν αφορούν τροχούς του **ιδίου** αυτοκινήτου, έπεται ότι έχουμε να κάνουμε με εξαρτημένα μεγέθη. Άρα τους δυο πληθυσμούς θα τους τρέψω σε έναν, το κάθε στοιχείο του οποίου θα προκύπτει από τη διαφορά του πρώτου από το δεύτερο ήτοι Ζκ=Χκ-Υκ για το τυχόν κ-στοιχείο του πληθυσμού που θα προκύψει.

-9	3	29	25	-9	-9	5	36	20	39	19	-6	17	20	21	12	-24

Οπότε και προκύπτει για τον πληθυσμό αυτό $\bar{z} = 11{,}12$ (ν=17) και s_z=17,17. Επίσης έχουμε διαδοχικά δε 99%=0,99 $\Rightarrow$ 1-α=0,99 $\Rightarrow$ α=0,01 $\Rightarrow$ $t_{\nu-1,\alpha/2}$ = $t_{16,0.005}$ = 2,921. Με απλή τώρα εφαρμογή του τύπου από τη θεωρία λαμβάνουμε:

$$\bar{z} \pm \frac{S_Z}{\sqrt{n}} t_{n-1,a/2} = 11{,}12 \pm \frac{17{,}17}{\sqrt{17}} 2{,}921 = 11{,}12 \pm 12{,}16 = (-1{,}04\ ,\ 23{,}28)$$

Παρατηρούμε ότι το μηδέν περιέχεται στο άνω διάστημα, που σημαίνει ότι δε χρειάζεται να πάρουμε άλλες μετρήσεις. Αν όμως είχαμε για παράδειγμα δ.ε. 95% τότε θα παίρναμε από τους πίνακες t του Student $t_{16,0.025}$=2,120 και το νέο διάστημα εμπιστοσύνης θα ήταν το (2,29 , 19,95). Βλέπουμε ότι το μηδέν δεν περιέχεται στο διάστημα αυτό και άρα θα πρέπει να πάρουμε επιπρόσθετες μετρήσεις.

Παρατήρηση 1η: Η εύρεσις διαστήματος εμπιστοσύνης τέτοιων πληθυσμών ονομάζεται δ.ε. διαφοράς μ_1-μ_2 συσχετισμένων πληθυσμών.

Παρατήρηση 2η: Εν γένει είναι χρήσιμο να κοιτάμε αν στο εκάστοτε δ.ε. της διαφοράς μ_1-μ_2 περιέχεται το μηδέν. Αν περιέχεται μπορούμε να συμπεράνομε ότι κατά μέσο όρο οι δυο αναλυτές (ή οι μετρήσεις για τους δυο πληθυσμούς) δε διαφέρουν στις αναλύσεις των. Έτσι τα δ.ε. δίνουν κατά τρόπο απλό, αλλά και ατελή, απάντηση στο πρόβλημα ελέγχου στατιστικών υποθέσεων (ένα από τα πλέον σύνθετα προβλήματα της στατιστικής) με το οποίο θα ασχοληθούμε στο επόμενο κεφάλαιο.

Άσκηση 3.7

Μετρήθηκε η διάρκεια ζωής ολοκληρωμένου κυκλώματος κάτω από δυο δυναμικά V_1 και V_2. Είναι n_1=10, s_1^2=0,51 και n_2=10, s_2^2=0,20. Οι πληθυσμοί θεωρούνται κανονικοί. Να βρεθεί διάστημα εμπιστοσύνης 90% για το λόγο σ_1^2/σ_2^2 των διασπορών των πληθυσμών. Μπορούμε να θεωρήσουμε ότι οι διασπορές των πληθυσμών είναι ίσες;

Λύση:

Για την εύρεση του δ.ε. θα χρησιμοποιήσουμε τον παρακάτω τύπο:

$$\boxed{\left(\frac{S_1^2}{S_2^2}\frac{1}{F_{n_1-1,n_2-1,a/2}},\frac{S_1^2}{S_2^2}\frac{1}{F_{n_1-1,n_2-1,a/2}}\right)}$$

όπου το $F_{n_1-1,n_2-1,a/2}$ θα το πάρουμε από τον πίνακα F που δίνει τα ποσοστιαία σημεία της κατανομής F του Snedecor. Έχουμε δε 90%=0,90 $\Rightarrow$ 1-α=0,90 $\Rightarrow$ α=0,1 $\Rightarrow$ α/2=0,05 άρα $F_{9,9,0.05}$ = 3,18. Σημειώνεται ότι πήγαμε στον πίνακα P=1-0,05=0,95.

Παρατήρηση: Είναι χρήσιμο να γνωρίζουμε πως επειδή οι μόνοι πίνακες F που δίδονται είναι για P=0,95 και P=0,99 , είναι πρόδηλο πως τα μόνα δ.ε. που μπορούν να ζητηθούν για το λόγο των διασπορών δυο κανονικών πληθυσμών είναι 90% και 98%.

Θα λάβουμε τότε: (0,80 , 8,11). Η απάντηση στο ερώτημα που τίθεται είναι ναι, καθόσον το 1 περιλαμβάνεται στο διάστημα εμπιστοσύνης.

Παρατήρηση: Ενώ πριν είπαμε για τη διαφορά μ_1-μ_2 ότι αν περιέχεται το μηδέν οι μέσες τιμές είναι ίσες, εδώ ζητάμε να περιέχεται η μονάδα για να πούμε ότι οι διασπορές είναι ίσες. Αυτό προκύπτει ως εξής: $\mu_1=\mu_2 \Rightarrow \mu_1-\mu_2=0 \Rightarrow$ το 0 περιλαμβάνεται στο δ.ε. και αντίστοιχα $\sigma_1^2 = \sigma_2^2 \Rightarrow \sigma_1^2/\sigma_2^2=1 \Rightarrow$ το 1 περιλαμβάνεται στο δ.ε.

Άσκηση 3.8

Σε 500 άτομα υπάρχουν 41 άνεργοι. Βρείτε διάστημα εμπιστοσύνης 95% για την αναλογία p της διωνυμικής κατανομής.

Λύση:

Έχουμε δ.ε. 95%=0,95 $\Rightarrow$ 1-α = 0,95 $\Rightarrow$ α=0,05 $\Rightarrow$ α/2=0,025. Άρα $Z_{0,025}$ = 1,96 (Z_1=1,9 και Z_2=0,06 - πήγαμε στο 1-0,025=0,97500 στον πίνακα κανονικής κατανομής Φ(Z)). Θα έχουμε επίσης $\hat{p} = 41/500 = 0{,}082$ και ν=500 άρα θα πάρουμε:

$$\hat{p} \pm Z_{a/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0{,}082 \pm 1{,}96\sqrt{\frac{0{,}082(1-0{,}082)}{500}} = 0{,}082 \pm 0{,}024 = (0{,}058\ ,\ 0{,}106)$$

Παρατήρηση: Για να εφαρμόσουμε τον παραπάνω τύπο πρέπει **απαραίτητα** το μέγεθος ν του δείγματος να είναι μεγαλύτερο του 30 (ν>30), κάτι που ασφαλώς συμβαίνει στην περίπτωση αυτή που εξετάσαμε.

Άσκηση 3.9

Σε πλήθος 400 ανδρών έχομε 280 ανάπηρους, ενώ σε πλήθος 600 γυναικών έχομε 455 ανάπηρες. Βρείτε διάστημα εμπιστοσύνης 95% για τη διαφορά $\rho_1-\rho_2$.

Λύση:

Κατά τα γνωστά θα είναι: $\hat{p}_1 = \frac{280}{400} = 0{,}7$ και $\hat{p}_2 = \frac{455}{600} = 0{,}758$. Επίσης είναι προφανές ότι δ.ε. 95%=0,95 $\Rightarrow$ 1-α=0,95 $\Rightarrow$ α=0,05 $\Rightarrow$ α/2=0,025 $\Rightarrow$ $Z_{0,025}$ = 1,96. Επομένως θα πάρουμε:

$$\hat{p}_1 - \hat{p}_2 \pm Z_{a/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = 0{,}7 - 0{,}758 \pm 1{,}96\sqrt{\frac{0{,}7(1-0{,}7)}{400} + \frac{0{,}758(1-0{,}758)}{600}} =$$

$$= -0{,}058 \pm 0{,}0565 = (-0{,}1145\ ,\ -0{,}0015)$$

Παρατήρηση: Είναι σαφές ότι το μηδέν δεν ανήκει στο παραπάνω διάστημα εμπιστοσύνης. Έτσι, και σύμφωνα με τα όλα ελέχθησαν παραπάνω, δε μπορούμε να θεωρήσουμε ότι $\rho_1=\rho_2 \Rightarrow \rho_1-\rho_2=0$. Δηλαδή δε μπορούμε να πούμε ότι οι αναλογίες ρ των διωνυμικών κατανομών είναι ίσες. Παρόλα αυτά η κατάσταση είναι **οριακή** επειδή είμαστε πολύ κοντά στο μηδέν. Έτσι για δ.ε. 99% θα έπαιρνα $Z_{0,005}=1,98 \Rightarrow$ (-0,1320 , 0,0160) το αντίστοιχο δ.ε. οπότε και θα μπορούσαμε να θεωρήσουμε ότι $\rho_1=\rho_2$.

Κεφάλαιο 4ο: Έλεγχοι Στατιστικών Υποθέσεων

Α. ΘΕΩΡΙΑ

4.1 Εισαγωγή

Ονομάζουμε statistικά υπόθεση κάθε υπόθεση που αφορά την κατανομή μιας τυχαίας μεταβλητής Χ. Συγκεκριμένα η υπόθεση συμβολιζόμενη με Η δύναται να αφορά (α) την παράμετρο θ της συνάρτησης πυκνότητας p(x|θ) με p γνωστή συναρτησιακή μορφή , (β) τη συναρτησιακή μορφή αυτή καθ' αυτή της συνάρτησης πυκνότητας p.

4.2 Έλεγχοι Στατιστικών Υποθέσεων

Θα ορίζουμε την αρχική ή μηδενική υπόθεση H_0 και έπειτα την εναλλακτική υπόθεση H_1. Θα κάνουμε τον έλεγχο (test) αφού προηγούμενα ορίσουμε την απορριπτική περιοχή R της υπόθεσης H_0. Αν ο έλεγχος πέσει εντός της απορριπτικής περιοχής R τότε θα δεχόμαστε την H_1 και θα απορρίπτουμε την H_0. Τα αντίθετα θα συμβούν στην περίπτωση που ο έλεγχος δώσει αποτέλεσμα εκτός της απορριπτικής περιοχής R οπότε και θα δεχόμαστε την αρχική υπόθεση H_0.

Σφάλμα τύπου 1: θα ονομάζεται το σφάλμα να απορρίψουμε την H_0 ενώ αυτή είναι σωστή. (πιθανότητα α)

Σφάλμα τύπου 2: θα ονομάζεται το σφάλμα να δεχτούμε την H_0 ενώ αυτή είναι λανθασμένη. (πιθανότητα β)

Ισχύ του ελέγχου θα ονομάζουμε την ποσότητα π=1-β.

Οι έλεγχοι των στατιστικών υποθέσεων προφανώς θα εξαρτώνται από το τι παράμετρο θέλομε να ελέγξομε. Αφού ορίσουμε τις H_0 και H_1 θα εφαρμόζουμε τον τύπο και ότι προκύψει θα κοιτάμε αν είναι εντός της απορριπτικής περιοχής. Αν είναι, σωστή είναι η υπόθεση H_1 , αν όχι δεχόμαστε την αρχική υπόθεση H_0.

4.3 Έλεγχος για το μ N(μ,σ^2)

Ας υποθέσουμε ότι έχουμε ένα κανονικό πληθυσμό N(μ,σ^2). Θέλουμε να ελέγξουμε αν η αρχική υπόθεση H_0: μ=10 είναι ορθή. Εναλλακτικά λέμε πχ H_1: μ>10. Έχει αποδειχτεί ότι η στατιστική συνάρτηση που μας επιτρέπει κάτι τέτοιο εξαρτάται από το αν γνωρίζουμε τη διασπορά σ^2 ή όχι. Επίσης αν δεν τη γνωρίζουμε, εξαρτάται από το αν το δείγμα (πλήθος n) είναι μεγάλο ή όχι. Επομένως πρέπει να διακρίνομε τρεις ξεχωριστές περιπτώσεις. Αυτές τις περιπτώσεις μας τις δίνει ο παρακάτω πίνακας:

H_0 $\mu_1=\mu_0$ **H_1 $\mu_1>\mu_0$**	**H_0 $\mu_1=\mu_0$** **H_1 $\mu_1<\mu_0$**	**H_0 $\mu=\mu_0$** **H_1 $\mu\neq\mu_0$**	**Στατιστικό Test**
R={ $z > z_a$ }	R={$z < -z_a$)	R={ $\|z\| > z_{a/2}$}	$z=\dfrac{\overline{x}-\mu_0}{\sigma}\sqrt{n}$, σ^2 γνωστό
R={ $z > z_a$ }	R={$z < -z_a$)	R={ $\|z\| > z_{a/2}$}	$z=\dfrac{\overline{x}-\mu_0}{s}\sqrt{n}$, σ^2 άγνωστο, $n\geq30$
R={ $t>t_{n-1,a}$}	R={$t<-t_{n-1,a}$}	R={$\|t\|>t_{n-1,a/2}$}	$t=\dfrac{\overline{x}-\mu_0}{s}\sqrt{n}$, σ^2 άγνωστο, n<30

Λόγω της εκφώνησης είναι σαφές ότι αναφερόμαστε στην πρώτη στήλη του πίνακα. Η διασπορά δίνεται N(μ,σ^2) και επομένως είμαστε και στην πρώτη γραμμή. Θα εφαρμόσουμε τον τύπο $z=\dfrac{\overline{x}-\mu_0}{\sigma}\sqrt{n}$ και αν είμαστε εντός της απορριπτικής περιοχής όπως αυτή δίνεται στην πρώτη στήλη - πρώτη γραμμή του πίνακα, τότε απορρίπτουμε την H_0 αλλιώς τη δεχόμαστε.

<u>Παράδειγμα</u>: Μηχανή ρυθμίζεται έτσι ώστε να παράγει προϊόντα πάχους 15mm. Για τον έλεγχο της λειτουργίας της μηχανής, τυχαίο δείγμα μεγέθους n=9, εξετάστηκε και έδωσε μέσο όρο $\overline{x}$=15,1mm και τυπική απόκλιση s=0,3mm. Να ελεγχθεί αν η μηχανή χρειάζεται ρύθμιση ή όχι σε στάθμη σημαντικότητας 0,01.

Προφανώς έχουμε να ελέγξουμε τη μέση τιμή μ. Η μηχανή δε θα χρειάζεται ρύθμιση αν πράγματι έχουμε μέσο όρο μ=15mm ενώ θα χρειάζεται <u>σε κάθε άλλη περίπτωση.</u> Άρα αν πούμε H_0: μ=15 και εναλλακτικά H_1: μ≠15 προφανώς πάμε στην τρίτη στήλη. Η διασπορά σ^2 δεν είναι γνωστή (δεν πρέπει να συγχέεται με τη s) άρα είμαστε ή στη δεύτερη ή στην τρίτη γραμμή. Επειδή όμως το μέγεθος του πληθυσμού είναι 9, προκύπτει ότι n=9<30 και άρα είμαστε στην τρίτη γραμμή. Εφαρμόζουμε τον τύπο:

$|t| = \left|\frac{\bar{x}-\mu_0}{s}\right|\sqrt{n} = \frac{|15,1-15|}{0,3}\sqrt{9} = 1$ και από τον πίνακα t του Student για 8 βαθμούς ελευθερίας και P=1-0,005=0,995 παίρνουμε $t_{8,0.005}$=3,355. Ελέγχουμε αν ισχύει R={$|t|>t_{n-1,a/2}$}. Αν πράγματι $t=1>3,355=t_{8,0.005}$ τότε απορρίπτεται. Αλλά αυτό δε συμβαίνει, επομένως η μηχανή δε θέλει ρύθμιση. Σημειώνεται ότι μπήκε απόλυτο στον προηγούμενο τύπο, γιατί είχαμε απόλυτο και στην απορριπτική περιοχή R. Θα μπορούσαμε να μην είχαμε βάλει απόλυτο, υπό την προϋπόθεση ότι θα σπάγαμε το απόλυτο της απορριπτικής περιοχής και θα κάναμε δυο συγκρίσεις.

4.4 Έλεγχος για τη διασπορά σ^2 $N(\mu,\sigma^2)$

Και αυτός ο έλεγχος αναφέρεται για κανονικούς πληθυσμούς. Ομοίως με όσα αναφέραμε παραπάνω, έχει αποδειχτεί ότι για να ελέγξουμε τη διασπορά πρέπει να κάνουμε τον έλεγχο όπως αυτός δίνεται από τον παρακάτω πίνακα:

H_0: $\sigma^2=\sigma_0^2$ **H_1: $\sigma^2>\sigma_0^2$**	**H_0: $\sigma^2=\sigma_0^2$** **H_1: $\sigma^2<\sigma_0^2$**	**H_0: $\sigma^2=\sigma_0^2$** **H_1: $\sigma^2\neq\sigma_0^2$**	**Στατιστικό Test**
R={$X^2>X^2_{n-1,a}$ }	R={$X^2<X^2_{n-1,a}$}	R={$X^2 > X^2_{n-1,a/2}$} ή R={$X^2 < X^2_{n-1,1-a/2}$}	$X^2 = \frac{(n-1)s^2}{\sigma_0^2}$

4.5 Έλεγχος για τη διαφορά μ_1-μ_2 κανονικών πληθυσμών

H_0: $\mu_1-\mu_2=\delta$ **H_1: $\mu_1-\mu_2<\delta$**	**H_0: $\mu_1-\mu_2=\delta$** **H_1: $\mu_1-\mu_2\neq\delta$**	**H_0: $\mu_1-\mu_2=\delta$** **H_1: $\mu_1-\mu_2>\delta$**	**Στατιστικό Test**
R={ $z < -z_a$ }	R={ $|z| > z_{a/2}$}	R={ $z > z_a$}	$\frac{\bar{x}-\bar{y}-\delta}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{n}}}$, σ_1^2,σ_2^2 γνωστά

R={ z < -z_α }	R={ |z| > $z_{\alpha/2}$}	R={ z > z_α}	$\frac{\bar{x}-\bar{y}-\delta}{\sqrt{\frac{s_1^2}{n}+\frac{s_2^2}{n}}}$,σ_1^2,σ_2^2 άγνωστα,n,m≥30
R={t<-$t_{n+m-2,\alpha}$}	R={|t|>$t_{n+m-2,\alpha/2}$}	R={t>$t_{n+m-2,\alpha/2}$}	$\frac{\bar{x}-\bar{y}-\delta}{\sqrt{\frac{s_1^2}{n}+\frac{s_2^2}{n}}}$,σ_1^2,σ_2^2 άγνωστα,n,m<30

<u>**Παράδειγμα**</u>: Δυο μηχανές συσκευάζουν προϊόν σε πακέτα των 500gr. Τυχαία δείγματα μεγέθους n_1=40 και n_2=60 πακέτων από τις μηχανές αυτές έδωσαν $\bar{x}_1$=496gr , s_1=6gr και $\bar{x}_2 = 502gr$, s_2=8gr. Να ελεγχθεί αν οι μηχανές διαφέρουν ως προς το μέσο βάρος συσκευασίας του προϊόντος σε στάθμη σημαντικότητας 0,01.

Μια σημαντική παρατήρηση που πρέπει να κάνουμε είναι ότι αν μηδενίσουμε στον παραπάνω πίνακα το δ, θα έχουμε έλεγχο ως προς την ισότητα ή όχι των μ_1,μ_2 γιατί όταν H_0: μ_1-μ_2=δ ⇒ H_0: μ_1-μ_2=0 ⇒ H_0: μ_1=μ_2 κοκ. Επομένως επειδή εμείς θέλουμε να δούμε αν διαφέρουν οι δυο μηχανές ως προς το <u>μέσο</u> βάρος (άρα και θα συγκρίνουμε τις μέσες τιμές μ_1,μ_2) θα πρέπει να θέσουμε στον παραπάνω πίνακα δ=0.

Δε γνωρίζουμε τις διασπορές και τα δείγματα είναι μεγαλύτερα από 30. Επομένως αν εμείς θέσουμε την αρχική υπόθεση H_0: μ_1=μ_2 και την εναλλακτική H_1: $\mu_1 \neq \mu_2$ θα είμαστε στη δεύτερη στήλη, δεύτερη γραμμή. Εφαρμόζουμε το στατιστικό test:

|z| = $\frac{|\bar{x}_1-\bar{x}_2-\delta|}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} = \frac{|496-502-0|}{\sqrt{\frac{36}{40}+\frac{64}{60}}} = \frac{6}{1{,}402} = 4{,}278$. Σημειώνεται ότι στον αριθμητή βάλαμε απόλυτο, καθόσον η απορριπτική περιοχή έχει |z|. Από τον πίνακα της κανονικής κατανομής για α/2=0,005 παίρνουμε $Z_{\alpha/2}$ = 2,58 (κατακόρυφη 2,5 και οριζόντια στήλη 0,08 δίνουν Z=0,99506 περίπου ίσο με 0,995). Επειδή |z| =4,278 > 2,58 = $Z_{\alpha/2}$, η υπόθεση H_0 απορρίπτεται και επομένως οι μηχανές δίνουν διαφορετικό μέσο βάρος συσκευασίας προϊόντος.

<u>**Παράδειγμα**</u>: Οι αντιστάσεις δέκα τμημάτων καλωδίου τύπου Α και δέκα τύπου Β μετρήθηκαν και προέκυψαν τα παρακάτω αποτελέσματα:

A	0,128	0,126	0,130	0,124	0,131	0,127	0,125	0,127	0,125	0,130
B	0,127	0,122	0,126	0,121	0,124	0,120	0,123	0,125	0,124	0,122

Υποθέτοντας ισότητας διασπορών, να ελεγχθεί η υπόθεση ισότητας των μέσων τιμών αντιστάσεως των καλωδίων τύπου A και B.

Εδώ έχουμε τον έλεγχο των υποθέσεων H_0: $\mu_1=\mu_2$ και H_1: $\mu_1 \neq \mu_2$ με $\sigma_1=\sigma_2=\sigma$ άγνωστο. Πρόκειται σαφώς για τη δεύτερη στήλη του παραπάνω πίνακα. Επειδή τα δείγματα είναι μικρά και σαφώς μικρότερα του 30, έχουμε την τρίτη γραμμή. Επειδή υποθέσαμε ισότητα διασπορών, θα είναι και $s_1=s_2=s$ όπου το s δίδεται εκ του τύπου:

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

Όμως από τα δεδομένα του προβλήματος προκύπτουν:

$$\bar{x} = \frac{1}{10}(0{,}128+0{,}126+0{,}130+0{,}124+0{,}131+0{,}127+0{,}125+0{,}127+0{,}125+0{,}130) = 0{,}127$$

$$\bar{y} = \frac{1}{10}(0{,}127+0{,}122+0{,}126+0{,}121+0{,}124+0{,}120+0{,}123+0{,}125+0{,}124+0{,}122) = 0{,}123$$

$$S_1^2 = \frac{1}{9}[(0{,}01)^2 + (-0{,}01)^2 + 2(0{,}03)^2 + (-0{,}03)^2 + (0{,}04)^2 + 0 + (-0{,}02)^2 + 0 + (-0{,}02)^2] = 6*10^{-4}$$

$$S_2^2 = \frac{1}{9}[(0{,}04)^2 + 2(-0{,}01)^2 + (0{,}03)^2 + (-0{,}02)^2 + 2(0{,}01)^2 + (-0{,}03)^2 + 0 + (0{,}02)^2] = 5*10^{-4}$$

άρα προκύπτει $s^2=5{,}5*10^{-4}$ επομένως $s=2{,}35*10^{-2}$ οπότε και παίρνουμε:

$|t| = \dfrac{|\bar{x}-\bar{y}-0|}{2{,}35*10^{-2}\sqrt{\dfrac{1}{10}+\dfrac{1}{10}}} = \dfrac{0{,}04}{0{,}0105} = 3{,}81$. Όμως από τον πίνακα t της κατανομής του Student για 10+10-2=18 βαθμούς ελευθερίας και P=1-α=0,995 παίρνουμε $t_{18,0.995}=2{,}878$ που όμως σημαίνει ότι αφού

$$\boxed{R=\{|t|>t_{n+m-2,\alpha/2}\}}$$

θα είναι $|t|=3,81 > 2,878 = t_{n+m-2,\alpha/2}$ και άρα η H_0 απορρίπτεται που σημαίνει ότι οι μέσες τιμές αντίστασης των καλωδίων τύπου A και B δεν μπορούν να θεωρηθούν ίσες.

4.6 Έλεγχος για την ισότητα διασπορών κανονικών πληθυσμών

H_0: $\sigma_1=\sigma_2$ **H_1: $\sigma_1 \neq \sigma_2$**	**H_0: $\sigma_1=\sigma_2$** **H_1: $\sigma_1>\sigma_0$**	**H_0: $\sigma_1=\sigma_2$** **H_1: $\sigma_1<\sigma_0$**	**Στατιστικό Test**
$R=\{F>F_{\alpha/2,n1-1,n2-1}$ ή $F<F_{1-\alpha/2,n1-1,n2-1}\}$	$R=\{F>F_{\alpha,n1-1,n2-1}\}$	$R=\{F<F_{1-\alpha,n1-1,n2-1}\}$	$F=\frac{S_1^2}{S_2^2}$

<u>Παράδειγμα</u>: Να γίνει έλεγχος ισότητας των διασπορών των πληθυσμών που είχαμε στο προηγούμενο παράδειγμα για τα καλώδια τύπου A και B, σε στάθμη σημαντικότητας 0,02.

Έχουμε να ελέγξουμε τις υποθέσεις H_0: $\sigma_1=\sigma_2$ και την εναλλακτική υπόθεση H_1: $\sigma_1 \neq \sigma_2$. Πρόκειται σαφώς για την πρώτη στήλη του παραπάνω πίνακα. Από τον πίνακα της κατανομής F του Snedecor για $\alpha=0,02 \Rightarrow \alpha/2=0,01$ παίρνουμε $F_{9,9,0.01}=5,35$ αφού $P=1-\alpha=0,99$. Δεν έχουμε όμως τον πίνακα για $1-\alpha/2 = 0,99 \Rightarrow P=0,01$. Ισχύει όμως όπως είδαμε στο προηγούμενο κεφάλαιο ο τύπος:

$$\boxed{\frac{1}{F_{n1-1,n2-1,a/2}} = F_{n2-1,n1-1,1-a/2}}$$

άρα $F_{1-\alpha/2,n1-1,n2-1}=1/5,35=0,187$. Από τον τύπο $F=\frac{S_1^2}{S_2^2}$ παίρνουμε $F=6/5=1,2$. Κοιτάμε να δούμε αν είμαστε εντός ή εκτός της απορριπτικής περιοχής. $F= 1,2 < 5,35 = F_{9,9,0.01}$ (άρα εκτός απορριπτικής περιοχής) και ταυτόχρονα $F= 1,2 > 0,187$ πάλι εκτός της απορριπτικής περιοχής. Δηλαδή η υπόθεση H_0 είναι δεκτή.

4.7 Έλεγχοι καλής προσαρμογής

Ας υποθέσουμε ότι έχουμε μια συνάρτηση f(x) μιας κατανομής. Θέλουμε να ελέγξουμε αν αυτή έχει καλή προσαρμογή. Επειδή έχουμε έλεγχο, προφανώς θα απαιτείται κάποια στάθμη σημαντικότητας έστω α. Ο έλεγχος αυτός διαφέρει από τους προηγούμενους. Μέχρι τώρα ασχοληθήκαμε με παραμετρικές υποθέσεις (διασπορά, μέσες τιμές, κτλ.) όμως τώρα θα ασχοληθούμε με την καταλληλότητα μιας κατανομής να περιγράψει κάποιο χαρακτηριστικό X. Το πρόβλημα αυτό είναι γνωστό ως έλεγχος καλής προσαρμογής.

Για τη λύση του προβλήματος αυτού θα μας χρησιμεύσει το κριτήριο X^2:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

όπου Oi είναι οι παρατηρηθείσες συχνότητες και E_i είναι οι αναμενόμενες συχνότητες. Πρέπει να προσέξουμε ιδιαίτερα οι συχνότητες οι αναμενόμενες να είναι μεγαλύτερες ή το πολύ ίσες με το πέντε (5). Αν δεν είναι, θα ενώσουμε γειτονικές κλάσεις έως ότου να είναι.

Τα παραπάνω γίνονται περισσότερο κατανοητά με τη μελέτη του παρακάτω παραδείγματος:

Παράδειγμα: Ο παρακάτω πίνακας δίνει τον αριθμό X των εισερχομένων κλήσεων σε τηλεφωνικό κέντρο. Να εξετασθεί αν ο αριθμός X^2 ακολουθεί κατανομή Poisson σε στάθμη σημαντικότητας α=0,05.

X	0	1	2	3	4	5	
f	15	21	12	5	1	0	*(54)*

Η κατανομή Poisson δίνεται από τη συνάρτηση $f(x) = e^{-\lambda}\frac{\lambda^x}{x!}$. Στην κατανομή όμως Poisson γνωρίζουμε ότι E(X)=λ $\Rightarrow \ \bar{x} = \lambda \Rightarrow \lambda = \frac{1}{54}(0*15 + 1*21 + 2*12 + 3*5 + 4*1) =$

=1,19 άρα η συνάρτηση f(x) γράφεται $f(x) = e^{-1,19}\dfrac{1,19^x}{x!}$. Πρέπει να εφαρμόσουμε τον τύπο X^2. Ξέρουμε τα O_i που είναι οι παρατηρηθείσες συχνότητες. Αυτό που δεν ξέρουμε όμως είναι οι αναμενόμενες συχνότητες E_i. Μπορούμε όμως να τις υπολογίσουμε ολοκληρώνοντας τη δοσμένη συνάρτηση με άκρα [0,1] - [1,2] - …. - [5,∞). Το αόριστο ολοκλήρωμα δίνει $P_i = e^{-1,19}\dfrac{\lambda^i}{i!}$ και αν πολλαπλασιάσουμε με το πλήθος n προκύπτουν τα f_i όπου για i=0,1,2,3,4,5 παίρνουμε τις παρακάτω τιμές:

X	0	1	2	3	4	5
f_i	16,25	19,49	11,72	4,70	1,40	0,44

Επειδή όμως στην 3η , 4η και 5η διαμέριση έχουμε 5,1 και 0 αντίστοιχα παρατηρήσεις πρέπει να τις ενώσουμε. (Αν ενώναμε την 4η και 5η , θα είχαμε μια κλάση με μια παρατήρηση, άρα θα έπρεπε να ενώσουμε υποχρεωτικά και την 3η κλάση). Ο παραπάνω πίνακας γράφεται τώρα:

X	0	1	2	3
f_i	16,25	19,49	11,72	6,54

Άρα το κριτήριο X^2 με αντικατάσταση δίνει:

$$X^2 = \frac{(15-16,25)^2}{16,25} + \frac{(21-19,49)^2}{19,49} + \frac{(12-11,72)^2}{11,72} + \frac{(6-6,54)^2}{6,54} =$$
$$= 0,096 + 0,077 + 0,007 + 0,045 = 0,225$$

Από τον πίνακα της κατανομής X^2 για α=0,05 και 2 βαθμούς ελευθερίας X^2=5,99 και επειδή $X^2_{2,0.05}$ =5,99 > 0,225 = X^2, η υπόθεση ότι ο αριθμός X των εισερχομένων κλήσεων ακολουθεί κατανομή Poisson γίνεται δεκτή.

Παρατήρηση: Οι βαθμοί ελευθερίας δίδονται εκ του τύπου v= n-k-1 όπου n οι διαμερίσεις (εδώ 4) και k οι προσδιοριστέοι συντελεσταί (εδώ 1 - η μέση τιμή λ).

4.8 Πίνακες Συνάφειας

Ας υποθέσουμε ότι έχουμε δυο χαρακτηριστικά Α και Β. Μας ενδιαφέρει να δούμε αν τα χαρακτηριστικά Α,Β είναι ανεξάρτητα. Για το λόγο αυτό θα εκτελέσουμε έλεγχο με αρχική ή μηδενική υπόθεση H_0: Α,Β ανεξάρτητα και εναλλακτική H_1: όχι η H_0. Ο έλεγχος αυτός επειδή εξάγει συμπέρασμα για την ανεξαρτησία των ενδεχομένων, θα καλείται έλεγχος ανεξαρτησίας (X^2-test ανεξαρτησίας).

$$X^2 = \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{(n_{ij} - \theta_{ij})^2}{\theta_{ij}} > X^2_{(s-1)(k-1),a}$$

Αν ισχύει το παραπάνω, η H_0 απορρίπτεται.

Παρατήρηση 1η: Το διπλό άθροισμα προϋποθέτει την ύπαρξη ενός διδιάστατου πίνακα από το σύνολο Π_v των πινάκων. Ένας τέτοιος διδιάστατος πίνακας με στοιχεία αυτά που προκύπτουν από την εφαρμογή της εντός του διπλού αθροίσματος ποσότητας λέγεται πίνακας συνάφειας. Ονομάζεται δε έτσι επειδή εξετάζουμε αν υπάρχει εξάρτηση (συνάφεια) μεταξύ των χαρακτηριστικών Α και Β.

Παρατήρηση 2η: Στον παραπάνω τύπο s θα είναι οι γραμμές του πίνακα συνάφειας και k οι στήλες του.

Παράδειγμα: Μηνιαίο περιοδικό ιστιοπλοΐας ζήτησε από 200 συνδρομητές του, τυχαία επιλεγέντες, να κατατάξουν τους εαυτούς των ανάλογα με τις ικανότητες ιστιοπλοΐας που θεωρούν ότι έχουν. Από τις απαντήσεις που ελήφθησαν συντάχθηκε ο παρακάτω πίνακας κατά φύλο και ικανότητα:

	Άνδρας	Γυναίκα	Σύνολο
Αρχάριος	40	5	45
Ενδιάμεσος	75	60	135
Προχωρημένος	13	7	20

Σύνολο	128	72	200

Να εξετασθεί σε στάθμη σημαντικότητας 0,05 αν τα παραπάνω δεδομένα παρέχουν αρκετές ενδείξεις ότι οι ικανότητες ιστιοπλοΐας έχουν σχέση με το φύλο.

Υποθέτοντας ότι υπάρχει ανεξαρτησία μεταξύ των εν λόγω χαρακτηριστικών και χρησιμοποιώντας τον παρακάτω τύπο προκύπτει ο πίνακας των αναμενόμενων συχνοτήτων:

$$\theta_{ij} = \frac{x_i y_j}{200}$$

	Άνδρας	**Γυναίκα**	*Σύνολο*
Αρχάριος	28,8	16,2	45,0
Ενδιάμεσος	86,4	48,6	135,0
Προχωρημένος	12,8	7,2	20,0
Σύνολο	128	72	200

Επειδή οι αναμενόμενες συχνότητες είναι όλες μεγαλύτερες του 5 δε χρειάζεται να γίνει σύμπτυξη κατηγοριών. Εφαρμόζοντας το κριτήριο X^2 έχουμε:

$$X^2 = \frac{(40-28{,}8)^2}{28{,}8} + \frac{(5-16{,}2)^2}{16{,}2} + \ldots + \frac{(7-7{,}2)^2}{7{,}2} = 16{,}286.$$

Εξάλλου από τον πίνακα X^2 με βαθμούς ελευθερίας (s-1)(k-1)=(3-1)(2-1)=2 και επίπεδο σημαντικότητας α=0,05 $\Rightarrow$ P=0,950 έχουμε $X^2_{2,0.05}$=5,991 < 16,286 = X^2 και επομένως απορρίπτεται η H_0. Συνεπώς υπάρχουν πολύ ισχυρές ενδείξεις ότι οι ικανότητες ιστιοπλοΐας έχουν σχέση με το φύλο και ιδιαίτερα με το ισχυρό.

4.9 Έλεγχος Ομοιογένειας

Ο έλεγχος αυτός απαιτεί την κατασκευή ενός πίνακα συνάφειας όπως αυτός που περιγράψαμε πριν. Η διαφορά εδώ είναι ότι θέλουμε να ελέγξουμε αν τα ποσοστά κάποιου πληθυσμού που έχει ιδιότητα Α και χωρίζεται σε κατηγορίες είναι ίσα μεταξύ τους.

Χρησιμοποιείται η κατανομή X^2 για το λόγο αυτό ο έλεγχος αυτός καλείται και X^2 test ομοιογένειας.

Παράδειγμα: Από την παραγωγή τριών μηχανών ελήφθησαν τρία τυχαία δείγματα 50 προϊόντων. Τα προϊόντα ελέγθησαν ως προς την ποιότητά τους και προέκυψε ο παρακάτω πίνακας συχνοτήτων:

Ποιότητα	M_1	M_2	M_3	*Σύνολον*
1. Χαμηλή	6	5	2	*13*
2. Μέτρια	8	8	4	*20*
3. Καλή	26	30	26	*82*
4. Πολύ καλή	10	7	18	*35*
Σύνολον	*50*	*50*	*50*	*150*

Να γίνει έλεγχος ομοιογένειας της παραγωγής των τριών μηχανών σε στάθμη (επίπεδο) σημαντικότητας α=0,05.

Θα πάρουμε κατά τα γνωστά:

E_{11}=4,33 - E_{12}=4,33 - E_{13}=4,33

$E_{21}=E_{22}=E_{23}$=6,67, $E_{31}=E_{32}=E_{33}$=27,33, $E_{41}=E_{42}=E_{43}$=11,67.

Επειδή η πρώτη κλάση συχνοτήτων έχει τιμή 4,33 μικρότερη του 5 θα την συμπτύξομε με τη δεύτερη, οπότε θα έχουν συχνότητα 4,33+6,67=11. Παίρνουμε τον παρακάτω πίνακα:

Ποιότητα	M_1	M_2	M_3	*Σύνολον*
1. Χαμηλή/ Μέτρια	11,00	11,00	11,00	*53*
2. Καλή	27,33	27,33	27,33	*82*
3. Πολύ καλή	11,67	11,67	11,67	*35*
Σύνολον	*50*	*50*	*50*	*150*

Εφαρμόζοντας το κριτήριο X^2 παίρνουμε:

$$X^2 = \frac{(14-11{,}00)^2}{11{,}00} + \frac{(13-11{,}00)^2}{11{,}00} + \ldots + \frac{(18-11{,}67)^2}{11{,}67} = 9{,}388$$

. Από το δε πίνακα της κατανομής X^2 με στάθμη σημαντικότητας α=0,05 και βαθμούς ελευθερίας (s-1)(k-1)=(3-1)(3-1)=4 παίρνουμε $X^2_{4,0.05}=7{,}779 < 9{,}388 = X^2$ και άρα απορρίπτεται η H_0.

<u>Παρατήρηση:</u> Ενώσαμε τις γραμμές 1 και 2 του αρχικού πίνακα γιατί οι αναμενόμενες συχνότητες της πρώτης γραμμής ήταν 4,33 μικρότερες του 5.

4.10 Μέτρηση Έντασης Της Σχέσης Μεταξύ 2 Μεταβλητών

Για να επιτύχουμε κάτι τέτοιο χρησιμοποιούμε έναν από τους παρακάτω συντελεστές:

$$\Phi = \sqrt{\frac{X^2}{n}}$$

Η παραπάνω σχέση μας δίνει τον συντελεστή Φ που αποτελεί το πρώτο μέτρο μέτρησης της έντασης της σχέσης μεταξύ δυο μεταβλητών.

$$V = \sqrt{\frac{X^2}{n \min(s-1, k-1)}}$$

Ο παραπάνω συντελεστής γνωστός ως συντελεστής V του Crammer προϋποθέτει την εύρεση του ελάχιστου μεταξύ των γραμμών και στηλών του πίνακα συνάφειας. Υπενθυμίζεται ότι με s συμβολίζονται οι γραμμές και με k οι στήλες του πίνακα συνάφειας.

$$C = \sqrt{\frac{X^2}{X^2 + n}}$$

Ο συντελεστής αυτός ονομάζεται συντελεστής συνάρτησης C.

Β. ΑΣΚΗΣΕΙΣ

Άσκηση 4.1

Δίνεται κανονικός πληθυσμός με μέση αντοχή 180. Δείγμα 50 ράβδων με $\bar{x}$=190 και s=35,2. Ελέγξτε τη μέση αντοχή με α=0,01.

Λύση:

Με την έκφραση έλεγχο εννοούμε ουσιαστικά τη σύγκριση (για παράδειγμα είναι ίσα ή μεγαλύτερα;). Αρκεί λοιπόν να πούμε H_0: $\mu=\mu_0$ έναντι H_1: $\mu>\mu_0$. Έχω $n=50\geq30$ και σ^2 άγνωστο. Η αρχική υπόθεση H_0 είναι $\mu_0=180$ και η εναλλακτική $\mu_0\neq180$. Άρα θα χρησιμοποιήσω τον τύπο:

$Z=(\bar{x}-\mu_0)\dfrac{\sqrt{n}}{s}=(190-180)\dfrac{\sqrt{50}}{35{,}2}=2{,}01$. Όμως είναι α=0,01 $\Rightarrow$ $Z_a=Z_{0,01}=2{,}33$.

Επειδή ισχύει $Z_a=2{,}33>2{,}01=Z$, συμπεραίνουμε ότι βρίσκεται εκτός της απορριπτικής περιοχής R και κατά συνέπεια είναι δεκτό. Εδώ πρέπει να τονίσομε ότι η απόφαση απόρριψης ή αποδοχής της εκάστοτε στατιστικής υπόθεσης επηρεάζεται άμεσα από το α. Πράγματι αν κάνουμε τον ίδιο έλεγχο για α=5%=0,05 $\Rightarrow$ $Z_a=Z_{0,05}=1{,}645<2{,}01=Z$, τότε η πρόταση απορρίπτεται. (βρίσκεται εντός της απορριπτικής περιοχής R).

Αν βέβαια απορρίπταμε την υπόθεση αυτή θα δοκιμάζαμε πχ H_0: $\mu=\mu_0$ και εναλλακτικά H_1: $\mu<\mu_0$ και ούτω καθ’ εξής.

Άσκηση 4.2

Για το παρακάτω δείγμα να γίνει έλεγχος της μέσης τιμής (=42) σε στάθμη σημαντικότητας 0,05. Δίνεται το δείγμα: 42,36,46,43,41,35,43,45,40,39.

Λύση:

Ομοίως θα υποθέσουμε H_0: $\mu=\mu_0$ και H_1: $\mu>\mu_0$. Στάθμη σημαντικότητας είναι το α. Άρα α=0,05 $\Rightarrow$ $t_{\alpha,0.05}$=1,833. Πρέπει να βρούμε το t και να το συγκρίνουμε με το tα. Έχουμε για το συγκεκριμένο δείγμα:

$\bar{x}=\frac{1}{10}\sum_{i=1}^{10}X_i=\frac{1}{10}410=41$ και $s^2=\frac{1}{10-1}\sum_{i=1}^{10}(X_i-\bar{X})^2$ =12,89$\Rightarrow$ s=3,59 (s>0).

Είναι $t=(\bar{x}-\mu_0)\frac{\sqrt{n}}{s}=(41-42)\frac{\sqrt{10}}{3{,}59}=-0{,}88.$ Όμως είναι $t_{9,0.05}$=1,833>-0,88=t άρα η υπόθεση H_0 είναι δεκτή. (βρίσκεται εκτός της απορριπτικής περιοχής R).

Άσκηση 4.3

Δίνεται η X με μ=1000gr, πλήθος δειγμάτων n=14. Να ελεγχθεί η διασπορά σ^2 σε στάθμη σημαντικότητας 0,05 αν H_0: σ^2=49 και H_1: σ^2>49. Δίνεται το δείγμα: 997, 999, 1001, 1004, 983, 997, 993, 998, 1010, 1017, 1005, 1002, 1009, 995.

Λύση:

Από τα δοσμένα στοιχεία θα βρούμε το μέσο όρο και το s ώστε να εφαρμόσουμε τον τύπο του ελέγχου. Πράγματι θα είναι $\bar{x}$ =1000,714 και s=8,33.

$X^2=\frac{(n-1)s^2}{\sigma^2}\Rightarrow X^2=\frac{13*8{,}33^2}{49}$ =18,43.

Πάμε στον πίνακα X^2 να βρούμε την τιμή $X^2_{13,0.05}$=22,36. Εδώ πρέπει να τονίσουμε ότι στην οριζόντια στήλη πρέπει να κοιτάξουμε στο P=0,95 καθόσον τόση θα είναι η πιθανότητα και όχι στο 0,05. Συγκρίνουμε, οπότε έχομε:

$X^2_{13,0.05}$ = 22,36 > 18,43 = X^2 (δεκτή η H_0, εκτός R).

Άσκηση 4.4

Για τους δυο παρακάτω κανονικούς πληθυσμούς δίδονται:

A	$\bar{x}$ =79	S_1=12 απόκλιση	n=240 άτομα
B	$\bar{y}$ =74	S_2=14 απόκλιση	m=180 άτομα

Οι πληθυσμοί αυτοί αφορούν τους μέσους όρους σε 1 εξάμηνο και στο ίδιο πλήθος μαθημάτων. Να γίνει έλεγχος H_0 $\mu_1=\mu_2$ με H_1 $\mu_1\neq\mu_2$ για τη διαφορά $\mu_1-\mu_2$ των δυο κατανομών με στάθμη σημαντικότητας 0,05.

Λύση:

Οι έλεγχοι από τη θεωρία ορίζουν $\mu_1-\mu_2=\delta$ για την H_0. Όμως εμείς θέλουμε $\mu_1=\mu_2\Rightarrow \mu_1-\mu_2=0 \Rightarrow \delta=0$. Έτσι θα θέτουμε δ=0 στους σχετικούς με τον έλεγχο αυτό τύπους. Θα βρούμε αρχικά το Z:

$Z=\dfrac{\bar{x}-\bar{y}-0}{\sqrt{\dfrac{s_1^2}{n}+\dfrac{s_2^2}{m}}}=\dfrac{79-74}{\sqrt{\dfrac{12^2}{240}+\dfrac{14^2}{180}}}=3{,}85$. Θα πρέπει να το συγκρίνομε με το ta. Θα είναι όμως

$\alpha=0{,}05 \Rightarrow Z\alpha_{/2}=Z_{0,025}=1{,}960$ ($Z_1=1{,}9$ & $Z_2=0{,}06$ στο 0,97500)

Συγκρίνουμε το $Z_{0,025}$ με το Z που βρήκαμε οπότε και θα έχουμε:

$t_{0,025}=1{,}960 < 3{,}85=t$ που σημαίνει ότι απορρίπτεται καθόσον βρίσκεται εντός της απορριπτικής περιοχής.

Άσκηση 4.5

Ομοίως με την άσκηση 4.4 όμως τώρα (Α): n=64, $\bar{x}$ =24 , $N(\mu_1,7^2)$ και $H_0=\mu_1-\mu_2=3$ - (Β): m=32, $\bar{y}$ =22 και $H_1=\mu_1-\mu_2>3$ με $N(\mu_2,6^2)$.

Λύση:

Είναι σαφές ότι κατά το τεστάκι αυτό θα πρέπει να θέσουμε στους τύπους που αφορούν τον έλεγχο αυτό, δ=3. Έτσι αρχικά θα υπολογίσουμε το Z, ύστερα το $Z_{0.025}$ και θα τα συγκρίνουμε για να δούμε αν απορρίπτεται ή όχι η H_0.

$Z=\dfrac{\bar{x}-\bar{y}-3}{\sqrt{\dfrac{7^2}{64}+\dfrac{6^2}{32}}}=\dfrac{24-22-3}{1{,}375}=-0{,}73$. Από τον πίνακα της κανονικής κατανομής παίρνουμε

$Z_{0,025}=1{,}96 > -0{,}73 = Z$ που σημαίνει ότι η H_0 είναι εκτός της απορριπτικής περιοχής R με άλλα λόγια είναι αποδεκτή.

Άσκηση 4.6

Παρακάτω δίνεται ο χρόνος ισχύος δυο δειγμάτων ενός φαρμάκου:

Δείγμα 1^o:	10,2	10,5	10,3	10,8	9,8	10,6	10,7	10,2	10	10,6
Δείγμα 2^o:	9,8	9,7	9,6	9,6	10,1	9,8	9,9	10,1	10,2	9,5

Οι πληθυσμοί είναι κανονικοί. Να γίνει έλεγχος για H_0: $\mu_1=\mu_2$ και H_1:$\mu_1 \neq \mu_2$ σε στάθμη σημαντικότητας 0,05.

Λύση:

Έχουμε μικρά δείγματα και άγνωστη διασπορά, άρα αντί του Ζ θα πρέπει να χρησιμοποιήσουμε το t του Student. Βρίσκουμε για το 1ο δείγμα : n=10, $\bar{x}$=10,37 και s_1^2=0,105. Ομοίως για το 2ο δείγμα παίρνουμε: m=10, $\bar{y}$=9,83 και s_2^2=0,058. Επομένως εφαρμόζοντας το σχετικό με τον έλεγχο τύπο έχουμε:

$$t = \frac{\bar{x} - \bar{y} - \delta}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{10{,}37 - 9{,}83 - 0}{0{,}285\sqrt{\frac{1}{10} + \frac{1}{10}}} = \frac{0{,}54}{0{,}1275} = 4{,}235 \text{ όπου } S_p^{\,2} = \frac{9 * 0{,}105 + 9 * 0{,}058}{10 + 10 - 2} = 0{,}0815$$

δηλαδή s_p=0,285. Έχουμε επίσης από τον πίνακα t του Student n=10+10-2=18 και α/2=0,05/2 =0,025 άρα πάμε στη γραμμή που έχει v=18 και στη στήλη με P=0,975, οπότε παίρνουμε $t_{18,0.025}$=2,101. Ελέγχουμε: $t_{18,0.025}$ = 2,101 < 4,235 = t. (απορρίπτεται η υπόθεση που κάναμε).

Άσκηση 4.7

Να γίνει έλεγχος καλής προσαρμογής με στάθμη σημαντικότητας α=0,05 της εκθετικής κατανομής που έχει $\bar{x}$=1 στα διαστήματα (0,1], (1,2], και (2,∞) με 60, 25 και 15 παρατηρήσεις αντίστοιχα.

Λύση:

Καταρχήν έχουμε εκθετική κατανομή, οπότε γνωρίζουμε τη συνάρτηση $f(x) = \lambda e^{-\lambda x}$ και αφού το πρόβλημα λέει $\bar{x}=1 \Rightarrow$ E(X)=1 $\Rightarrow$ λ=1 τότε η f(x) είναι προφανώς η $f(x) = e^{-x}$. Σε ασκήσεις αυτού του είδους θα βρίσκουμε όπως κάναμε πριν τη συνάρτηση. Το επόμενο βήμα είναι να χρησιμοποιήσουμε ολοκληρώματα (έχουμε άπειρες τιμές στα διαστήματα αυτά και όχι πεπερασμένες ώστε να πάρουμε αθροίσματα) με άκρα τα άκρα των διαστημάτων της συνάρτησης που βρήκαμε.

Τα ολοκληρώματα αυτά θα μας δώσουν τις πιθανότητες έστω P_{10}, P_{20} και P_{30}. Το πλήθος των ολοκληρωμάτων που θα χρησιμοποιήσουμε θα είναι όσο το πλήθος των δοσμένων διαστημάτων, στα οποία ζητήθηκε να γίνει έλεγχος καλής προσαρμογής.

Είναι προφανές ότι το άθροισμα των πιθανοτήτων P_{10}, P_{20} και P_{30} θα πρέπει να είναι ίσο με τη μονάδα, οπότε πρακτικά θα υπολογίζουμε τα δυο πρώτα και το τρίτο θα το βρίσκουμε από τη σχέση $P_{10} + P_{20} + P_{30} = 1$. Αυτό είναι πολύ χρήσιμο διότι αφενός είναι γρήγορο και αφετέρου δεν υπολογίζουμε ολοκλήρωμα με άκρο το άπειρο, εν γένει δύσκολο.

$$P_{10} = \int_0^1 e^{-x}dx = -e^{-x}\Big|_0^1 = -e^{-1}+1 = 0{,}632 \Rightarrow \theta_1 = 100 * P_{10} = 100 * 0{,}632 = 63{,}2$$

$$P_{20} = \int_1^2 e^{-x}dx = -e^{-x}\Big|_1^2 = -e^{-2}+e^{-1} = 0{,}233 \Rightarrow \theta_2 = 100 * P_{20} = 100 * 0{,}233 = 23{,}3$$

$$P_{30} = \int_2^{+\infty} e^{-x}dx = -e^{-x}\Big|_2^{\infty} = 0+e^{-2} = 0{,}135 \Rightarrow \theta_3 = 100 * P_{30} = 100 * 0{,}135 = 13{,}5$$

<u>Σημείωση</u>: Το 100 στις παραπάνω σχέσεις είναι το πλήθος n=60+25+15 που έχουμε.

Αν και θα μπορούσαμε να είχαμε βρει το P_{30} δίχως τη χρήση του ολοκληρώματος σύμφωνα με τα όσα αναφέρθησαν παραπάνω, προτιμήθηκε ο υπολογισμός του ώστε να δείξουμε τη σχέση που αναφέραμε πριν. Ταυτόχρονα δε με τα ολοκληρώματα βρήκαμε και τις ποσότητες θ_κ, Τώρα έχουμε το τελευταίο στάδιο που είναι ο υπολογισμός του X^2.

$$X^2 = \frac{(60-63{,}2)^2}{63{,}2} + \frac{(25-23{,}2)^2}{23{,}2} + \frac{(15-13{,}6)^2}{13{,}6} = 0{,}451$$

Πάμε στον πίνακα X^2 και βρίσκουμε ότι $X^2_{1,0.05}$=3,841 (πήγαμε στο ν=1 και στο P=0,950). Έπεται ότι $X^2_{1,0.05}=3,841 < 0,451 = X^2$ άρα δεκτή η H_0.

<u>Παρατήρηση</u>: Αν θέλαμε να βρούμε προσεγγιστικά το μέσο όρο θα είχαμε:

		y_i	v_i	$v_i y_i$
0	1	0,5	60	30
1	2	1,5	25	37,5
2	3	2,5	15	37,5

όπου y_i το ημιάθροισμα άνω και κάτω άκρου διαστήματος. Τα διαστήματα που εκλέξαμε είναι τα (0,1], (1,2] και (2,3]. Προφανώς δε θα μπορούσαμε να υπολογίσουμε το διάστημα (2,∞) της εκφώνησης. Αν αθροίσουμε την τελευταία στήλη παίρνουμε $\bar{x}$=1,05 περίπου ίσο με το δοσμένο μέσο όρο.

Άσκηση 4.8

Τυχαίο δείγμα 200 υπαλλήλων ελέγχεται ως προς τη συμπεριφορά ως προς την έκπτωση (=Β) και ιδιότητα υπαλλήλου (=Α).

	Κάνει	**Δεν Κάνει**	**Δ.Ξ / Δ.Α.**	**Β**
Πωλητής	30	15	15	60
Τμηματάρχης	40	50	10	100
Διαχειριστής	10	25	5	40
Σύνολο Α	80	90	35	ν=200

Να ελεγθεί η υπόθεση ανεξαρτησίας των Α και Β, H_0 τα Α,Β ανεξάρτητα και H_1 τα ΑΒ όχι ανεξάρτητα σε στάθμη σημαντικότητας α=5%.

Λύση:

Θα φτιάξουμε ένα πίνακα 3x3 με στοιχεία $\theta_{\kappa\lambda}$ όπου κ,λ=1,2,3. Τα στοιχεία του θ θα τα φτιάξουμε ως εξής:

$$\theta_{11}=\frac{60*80}{200}=24,\quad \theta_{12}=\frac{60*90}{200}=27,\quad \theta_{13}=\frac{60*35}{200}=9,\quad \theta_{21}=\frac{100*80}{200}=40$$

$$\theta_{22}=\frac{100*90}{200}=45,\quad \theta_{23}=\frac{100*30}{200}=15,\quad \theta_{31}=\frac{40*80}{200}=16,\quad \theta_{32}=\frac{40*90}{200}=18$$

$$\theta_{33}=\frac{40*30}{200}=6$$

Έτσι δημιουργήσαμε τον πίνακα συνάφειας $\begin{bmatrix} 24 & 27 & 9 \\ 40 & 45 & 15 \\ 16 & 18 & 6 \end{bmatrix}$. Στη συνέχεια θα βρούμε το X^2 ώστε να το συγκρίνουμε με αυτό που θα εκλάβουμε από το σχετικό πίνακα των ποσοστιαίων σημείων της κατανομής X^2 με στάθμη σημαντικότητας 5%. Το X^2 θα υπολογιστεί κάνοντας χρήση των στοιχείων του παραπάνω πίνακα θ ως εξής:

$$X^2=\sum_{i=1}^{3}\sum_{j=1}^{3}\frac{(n_{ij}-\theta_{ij})^2}{\theta_{ij}}=\frac{(30-24)^2}{24}+\frac{(15-27)^2}{27}+\ldots+\frac{(5-6)^2}{6}=18{,}2\,.$$

Όπου τα n_{ij} τα παίρνουμε έτοιμα από τα δεδομένα του πίνακα (εκφώνηση) και τα θ_{ij} τα παίρνουμε από τον πίνακα θ που δημιουργήσαμε. Τώρα πάμε στον πίνακα των X^2 όπου για $a=5\%=0{,}05 \Rightarrow P=1-0{,}05=0{,}950$ και $v=(3-1)(3-1)=2*2=4$, παίρνουμε $X^2_{4,0.950}=9{,}488$. Επειδή δε ισχύει $X^2_{4,0.950}=9{,}488 < 18{,}2 = X^2$, απορρίπτεται η H_0 (εντός της απορριπτικής περιοχής).

<u>Παρατήρηση</u>: Για το X^2 το ν δεν είναι 200, αλλά το γινόμενο των γραμμών επί των στηλών του πίνακα που μας δώσαμε (3x3) μείον τη μονάδα και στις γραμμές και στις στήλες (2x2). Αυτό ισχύει πάντα στις ασκήσεις αυτές και χρειάζεται ιδιαίτερη προσοχή.

Άσκηση 4.9

(X^2 - test ομοιογένειας). Σε μελέτη μας ενδιαφέρει να εξετάσουμε τα ποσοστά των αλκοολικών σε ένα πληθυσμό (1200 άτομα). Παίρνουμε τα παρακάτω:

	Αλκοολικοί	**Μη Αλκοολικοί**	
Κλήρος	32	268	*300*
Εκπαιδευτικοί	51	199	*250*
Υπάλληλοι	67	233	*300*
Έμποροι	83	267	*350*
Σύνολο	*233*	*967*	*1200*

Να ελεγθεί η υπόθεση H_0 αν τα ποσοστά P_1,P_2,P_3 και P_4 είναι ίδια $P_1=P_2=P_3=P_4$ (γι' αυτό και το αποκαλούμε test ομοιογένειας) και H_1 όχι ίδια. Η στάθμη σημαντικότητας του ελέγχου είναι 5%.

Λύση:

Θα εργαστούμε ανάλογα με την προηγούμενη άσκηση. Θα δημιουργήσουμε αρχικά τον πίνακα θ:

$$\theta_{11} = \frac{300*233}{1200} = 58{,}25, \quad \theta_{12} = \frac{300*967}{1200} = 241{,}75, \quad \kappa o\kappa .$$

Θα πάρουμε τότε τον πίνακα $\theta = \begin{bmatrix} 58{,}25 & 241{,}75 \\ 48{,}54 & 201{,}96 \\ 58{,}25 & 241{,}75 \\ 67{,}96 & 282{,}04 \end{bmatrix}$. Στη συνέχεια πρέπει να βρούμε το X^2 όπως ακριβώς κάναμε και στην προηγούμενη άσκηση:

$X^2 = \dfrac{(32-58{,}25)^2}{58{,}25} + \dfrac{(51-48{,}54)^2}{48{,}54} + \ldots + \dfrac{(267-282{,}04)^2}{282{,}04} = 20{,}59.$ Για να βρούμε το X^2 από τον πίνακα, αφενός μεν πρέπει να γνωρίζουμε το P που εδώ θα είναι α=5%=0,05$\Rightarrow$ P=0,950 και αφετέρου το ν που καλείται <u>βαθμός ελευθερίας</u> και όπως είπαμε σε προηγούμενη παρατήρηση δεν έχει καμία σχέση με το πλήθος του πληθυσμού, αλλά είναι ίσο με το γινόμενο των γραμμών μείον τη μονάδα του πίνακα που δίδεται επί των στηλών μείον πάλι τη μονάδα. Έτσι ν=(4-1)(2-1)=3*1=3. Πάμε στον πίνακα να βρούμε το $X^2_{3,0.950}$ οπότε και παίρνουμε 7,815. Βλέπουμε ότι:

$X^2_{3,0.950} = 7{,}815 < 20{,}59 = X^2$ και άρα η υπόθεση X^2 απορρίπτεται (έχουμε μάλλον ανομοιογένεια).

Άσκηση 4.10

Να ελεγθεί αν τα Α,Β είναι ανεξάρτητα. Να υπολογισθεί ο συντελεστής Φ, ο συντελεστής V του Crammer και ο συντελεστής συνάρτησης C.

A\B	**<0,5=B_1**	**0,5=B_2**	**>0,5=B_3**	
A_1	16	12	19	*47*
A_2	32	12	9	*53*
	48	*24*	*28*	*100*

Λύση:

Ομοίως με τις προηγούμενες ασκήσεις θα κατασκευάσουμε τον πίνακα θ:

$$\theta_{11} = \frac{47*48}{100} = 22{,}56.....\Rightarrow \theta = \begin{bmatrix} 22{,}56 & 11{,}28 & 13{,}16 \\ 25{,}44 & 12{,}72 & 14{,}84 \end{bmatrix}.$$

Άρα $X^2 = \dfrac{(16-22{,}56)^2}{22{,}56} + \dfrac{(12-11{,}28)^2}{11{,}28} + ... + \dfrac{(9-14{,}84)^2}{14{,}84} = 8{,}57.$

Από τον πίνακα X^2 παίρνουμε (α=0,05 $\Rightarrow$ P=0,950 και βαθμοί ελευθερίας ν=(3-1)(2-1)=2) $X^2_{2,0.950}$ = 5,991 < 8,57 =X^2 και επομένως η υπόθεση H_0 απορρίπτεται.

$$\Phi = \sqrt{\frac{X^2}{n}} = \sqrt{\frac{8{,}57}{100}} = 0{,}293$$

$V = \sqrt{\dfrac{X^2}{n*\min(s-1,k-1)}} = \sqrt{\dfrac{8{,}57}{100*\min(2{,}1)}} = 0{,}293$ (s= γραμμές, k=στήλες)

$$C = \sqrt{\frac{X^2}{X^2+n}} = \sqrt{\frac{8{,}57}{8{,}57+100}} = 0{,}281$$

Κεφάλαιο 5ο: Ανάλυση Παλινδρόμησης & Συσχέτισης

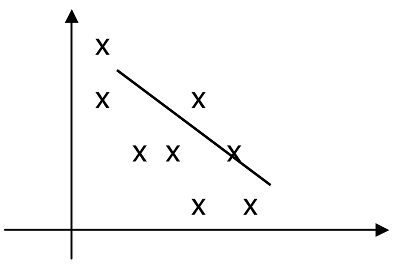
Α. ΘΕΩΡΙΑ

5.1 Εισαγωγή

Ενώ μέχρι τώρα ασχοληθήκαμε κυρίως με το πρόβλημα εκτίμησης των παραμέτρων της κατανομής ενός πληθυσμού, στο κεφάλαιο αυτό θα ασχοληθούμε με το πρόβλημα της πρόβλεψης μιας "εξαρτημένης" μεταβλητής, έστω Y, όταν γνωρίζουμε τις τιμές των "ανεξάρτητων" μεταβλητών $X_1, X_2, \ldots, X_\nu$ συμβολιζόμενες χάριν συντομίας X.

5.2 Απλό Γραμμικό Μοντέλο (Παλινδρόμηση)

Ας υποθέσουμε Y μια εξαρτημένη μεταβλητή από τη X. Στο συνημμένο σχήμα φαίνεται το διάγραμμα διασποράς (scatter diagram). Υποθέτουμε ότι $y=a+bx$ το προσδιοριστικό μοντέλο με α σταθερό όρο και b την κλίση.



Βλέπουμε ότι το απλό γραμμικό μοντέλο είναι στην ουσία μια ευθεία, η οποία τείνει να προσεγγίσει το σύνολο των σημείων που έχουμε, χωρίς όμως να περνάει από όλα τα σημεία. Ας υποθέσουμε ότι e_i είναι το σφάλμα (απόκλιση) της ευθείας από ένα τυχαίο σημείο i. Τότε είναι σαφές ότι το στοχαστικό μοντέλο θα ήταν $y_i=a+bx_i+e_i$.

Θα προσπαθήσουμε να απλοποιήσουμε την απόδειξη των κανονικών εξισώσεων. Έστω $y_i = a + bx_i$ η ευθεία. Τότε $(y_i\text{-}a\text{-}bx_i)^2$ θα είναι οι αποστάσεις των εκάστοτε σημείων. Το δε άθροισμά των αποστάσεων τους παραγωγιζόμενο πρέπει να είναι ίσο με το μηδέν. Παραγωγίζοντας και εξισώνοντας με το μηδέν μια έκφραση, παίρνουμε το τοπικό της ακρότατο.

Επειδή το μοντέλο της ευθείας θα είναι τόσο επιτυχημένο, όσο μικρότερες είναι οι αποστάσεις των σημείων από αυτή και ιδανικό αν είναι μηδέν (όλα τα σημεία πάνω στην ευθεία) θα αναζητήσουμε τοπικό (ή ολικό) ελάχιστο.

$$\left.\begin{aligned} \frac{\partial}{\partial \alpha}\sum_{i=1}^{n}(y_i - \alpha - bx_i) = 0 \\ \frac{\partial}{\partial b}\sum_{i=1}^{n}(y_i - \alpha - bx_i) = 0 \end{aligned}\right\} \Rightarrow \begin{cases} \sum y_i = na + b\sum x_i \\ \sum x_i y_i = a\sum x_i + b\sum x_i^2 \end{cases} \quad (*)$$

-

Οι εξισώσεις (*) ονομάζονται πρώτη και δεύτερη κανονική εξίσωση αντίστοιχα και έχουν ορισμένες εξαιρετικά χρήσιμες εφαρμογές. Ισχύουν οι παρακάτω τύποι που δίνονται χωρίς απόδειξη:

$$\overline{x} = \frac{\sum x_i}{n}, \quad \overline{y} = \frac{\sum y_i}{n}$$

Επίσης η διασπορά γράφεται:

$$s_x^2 = \frac{\sum x_i^2 - n\overline{x}^2}{n-1}, \quad s_{xy}^2 = \frac{\sum x_i y_i - n\overline{x}\overline{y}}{n-1}$$

Άρα μπορούμε άμεσα να προσδιορίσουμε το σταθερό όρο α και την κλίση b του απλού γραμμικού μοντέλου ως εξής:

$$\hat{b} = \frac{S_{xy}}{S_x^2}, \quad \hat{\alpha} = \overline{y} - \hat{b}\overline{x}$$

Η παρακάτω ποσότητα είναι η αμερόληπτη εκτιμήτρια της δεσμευμένης διασποράς $V[Y|x]=\sigma^2_{y|x}$ υπό την προϋπόθεση ότι η τελευταία παραμένει σταθερή για τις διάφορες τιμές του x (ομοσκεδαστικότητα).

$$s^2_{y|x} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Ορίζεται ο συντελεστής προσδιορισμού (προσαρμογής) R^2 από τον τύπο:

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \hat{y}_i)^2}$$

Τέλος ορίζουμε τον συντελεστή συσχέτισης r, που στο απλό γραμμικό μοντέλο το τετράγωνό του συμπίπτει με το τετράγωνο του συντελεστή προσαρμογής R^2. Αυτό συμβαίνει γιατί ο συντελεστής συσχέτισης r χρησιμοποιείται γενικά για τον έλεγχο της ύπαρξης ή όχι γραμμικής εξάρτησης μεταξύ των x_i και y_i.

Παράδειγμα: Μετρήσεις της ποσότητας οξειδίου y(A) που σχηματίζεται στην επιφάνεια μετάλλου που τοποθετείται για χρόνο x(min) σε κλίβανο σταθερής θερμοκρασίας δίνεται από τον παρακάτω πίνακα. Να εκτιμηθούν με τη μέθοδο ελαχίστων τετραγώνων η σταθερά α και η κλίση b της ευθείας y=αx+b και να προσδιορισθεί ο δειγματικός συντελεστής συσχέτισης r, ο συντελεστής προσαρμογής (προσδιορισμού) R^2 και η διασπορά σφαλμάτων $S^2_{y|x}$.

X	10	20	30	40	50	60	70	80	90
Y	2,0	5,0	6,5	9,5	11,0	13,5	15,0	17,5	19,0

Για την επίλυση της άσκησης, σχηματίζουμε τον παρακάτω πίνακα:

X	Y	$X-\overline{X}$	$Y-\overline{Y}$	$(X-\overline{X})^2$	$(X-\overline{X})(Y-\overline{Y})$	$(Y-\overline{Y})^2$
10	2,0	-40	-9,0	1600	360	81,00
20	5,0	-30	-6,0	900	180	36,00
30	6,5	-20	-4,5	400	90	20,25
40	9,5	-10	-1,5	100	15	2,25
50	11,0	0	0,0	0	0	0,00
60	13,5	10	2,5	100	25	6,25
70	15,0	20	4,0	400	80	16,00
80	17,5	30	6,5	900	195	42,25
90	19,0	40	8,0	1600	320	64,00
450	*99,0*	--	--	*6000*	*1265*	*268,00*

Αφού βρίσκουμε εύκολα ότι $\bar{x}=50, \bar{y}=11{,}0$.

Από τις κανονικές εξισώσεις αν λύσουμε ως προς α και b παίρνουμε: $\hat{b}=\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$

επομένως b=0,21 και από την $\bar{y}=\hat{a}+\hat{b}\bar{x} \Rightarrow \hat{a}=\bar{y}-\hat{b}\bar{x} \Rightarrow \hat{a}=11-(0{,}21)50=0{,}50$.

Βρίσκουμε επίσης ότι $R^2=\dfrac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=\dfrac{\sum_{i=1}^{n}(y_i-\bar{y})^2-\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=0{,}994=r^2$ και

τέλος $S_{y|x}^2=\dfrac{1}{9-1}\sum_{i=1}^{9}(y_i-\hat{y}_i)^2=0{,}217$. Το απλό γραμμικό μοντέλο (η ευθεία δηλαδή) είναι προφανώς η $\hat{y}=0{,}50+0{,}21x$.

5.3 Δ.Ε. για τα α και b.

Οι τύποι που δίνουν τα δ.ε. στάθμης σημαντικότητας α για τα α (σταθερά) και b (κλίση) του απλού γραμμικού μοντέλου είναι οι παρακάτω:

$(\hat{a}\pm Sa_{tn-1,a/2})$ και $(\hat{b}\pm Sb_{tn-2,a/2})$

όπου Sa και Sb δίνονται από τους τύπους:

$$S^2a = s^2(\frac{1}{n} + \frac{\overline{x}^2}{(n-1)S_x^2}) \quad \text{και} \quad S^2b = \frac{s^2}{(n-1)s_x^2}$$

5.4 Έλεγχος για το b

Είναι πολύ σημαντικό να μπορούμε να κάνουμε έλεγχο όχι μόνο για την κλίση b αλλά όπως θα δούμε παρακάτω και για τη μέση και ατομική πρόβλεψη. Οι έλεγχοι αυτοί διαφέρουν μόνο ως προς τους τύπους που εφαρμόζονται, καθόσον η φιλοσοφία τους παραμένει ίδια. Το επίπεδο σημαντικότητας θα είναι το λιγότερο 90%.

H_0: $b=b_0$ **H_1: $b>b_0$**	**H_0: $b=b_0$** **H_1: $b<b_0$**	**H_0: $b=b_0$** **H_1: $b \neq b_1$**	**Στατιστικό Test**
$R=\{t > t_{n-2,\alpha}\}$	$R=\{t<-t_{n-2,\alpha}\}$	$R=\{\lvert t\rvert>t_{n-2,\alpha/2}\}$	$t = \frac{\hat{b} - b_0}{Sb}$

5.5 Δ.Ε. για τη μέση και ατομική πρόβλεψη

Αυτά θα δίδονται από τους τύπους:

$$(\hat{y} \pm st_{n-2,\alpha/2}\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)S_x^2}}) \text{ [μέση πρόβλεψη]}$$

$$(\hat{y} \pm st_{n-2,\alpha/2}\sqrt{1+\frac{1}{n} + \frac{(x-\overline{x})^2}{(n-1)S_x^2}}) \text{ [ατομική πρόβλεψη]}$$

5.6 Έλεγχος για τη μέση πρόβλεψη E(y)

Θα γίνεται βάση του πίνακα:

H_0: $E(y)=\mu_0$ **H_1: $E(y)>\mu_0$**	**H_0: $E(y)=\mu_0$** **H_1: $E(y)<\mu_0$**	**H_0: $E(y)=\mu_0$** **H_1: $E(y)\neq\mu_0$**	**Στατιστικό Test**
$R=\{t>t_{n-2,a}\}$	$R=\{t<t_{n-2,a}\}$	$R=\{\lvert t\rvert>t_{n-2,a/2}\}$	$t=\frac{\hat{y}-\mu_0}{s\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{(n-1)s_x^2}}}$

Β. ΑΣΚΗΣΕΙΣ

Άσκηση 5.1

Αν $\hat{y} = \hat{a} + \hat{b}x$ η εκτιμήτρια ελαχίστων τετραγώνων και $\hat{e}_i = y_i - \hat{y}_i$ τα εκτιμώμενα σφάλματα, τότε να δείξετε ότι:

$$\sum_{i=1}^{n} \hat{e}_i = 0$$

$$\sum_{i=1}^{n} x_i \hat{e}_i = 0$$

$$\sum_{i=1}^{n} \hat{y}_i \hat{e}_i = 0 \qquad \textit{(Θέμα 6/1996)}$$

Λύση:

α) Καταρχήν είναι $\hat{y} = \hat{a} + \hat{b}x$ και $\bar{y} = \hat{a} + \hat{b}\bar{x}$ άρα $\hat{y} = \bar{y} + \hat{b}(x - \bar{x})$. Κατόπιν τούτων έχουμε:

$$\sum_{i=1}^{n} \hat{e}_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}x_i) = \sum y_i - n\hat{a} - \hat{b}\sum x_i = n\bar{y} - n\hat{a} - \hat{b}n\bar{x} = n(\bar{y} - \hat{a} - \hat{b}x) = 0$$

λόγω της πρώτης κανονικής εξίσωσης.

β) $\sum_{i=1}^{n} x_i \hat{e}_i = \sum x_i (y_i - \hat{a} - \hat{b}x_i) = \sum x_i y_i - \hat{a}\sum x_i - \hat{b} \sum x_i^2 = 0$ από τη δεύτερη κανονική εξίσωση.

γ)

$$y_i \hat{e}_i = (\hat{a} + \hat{b}x_i)(y_i - \hat{a} - \hat{b}x_i) = \hat{a}(y_i - \hat{a} - \hat{b}x_i) + \hat{b}(x_i - \hat{a} - \hat{b}x_i) \Rightarrow$$

$$\Rightarrow \sum_{i=1}^{n} y_i \hat{e}_i = \hat{a}(\sum y_i - n\hat{a} - \hat{b}\sum x_i) + \hat{b}(\sum x_i y_i - \hat{a}\sum x_i - \hat{b}\sum x_i^2) = 0 \text{ ,}$$

λόγω της πρώτης και δεύτερης κανονικής εξίσωσης.

Άσκηση 5.2

Να βρεθεί η ευθεία ελαχίστων τετραγώνων, η διασπορά σφαλμάτων και ο συντελεστής προσδιορισμού R^2.

Y	9	5	7	14	10
X	3	1	2	5	4

Λύση:

Θα έχουμε $\sum y_i = n\hat{a} + \hat{b}\sum x_i \Rightarrow 45 = 5\hat{a} + 15\hat{b} \quad (1)$. Από δεύτερη κανονική εξίσωση έχουμε $\sum x_i y_i = \hat{a}\sum x_i + \hat{b}\sum x_i^2 \Rightarrow 156 = 15\hat{a} + 55\hat{b}. \quad (2) \quad \xrightarrow{(1),(2)} \hat{a} = 2{,}7 \ \& \ \hat{b} = 2{,}1.$ Η ζητούμενη ευθεία θα είναι η $\hat{y} = 2{,}7 + 2{,}1x$. Μπορούμε να ελέγξουμε αν η ευθεία είναι αξιόπιστο μοντέλο με το δεδομένα του πίνακα. Για παράδειγμα αν θέσουμε x=3 $\Rightarrow$ y=9 και για x=1 $\Rightarrow$ y=4,8 (αντί 5). Πράγματι η ευθεία είναι ένα πολύ αξιόπιστο μοντέλο.

y	$\bar{y}$	$\hat{y}$	$(y-\bar{y})^2$	$(\hat{y}-\bar{y})^2$	$(y-\hat{y})^2$
9	9	9	0	0	0
5	9	4,8	16	17,64	0,04
7	9	6,9	4	4,41	0,01
14	9	13,2	25	17,64	0,64
10	9	11,1	1	4,41	1
			SST=46	**SSR=44,10**	**SSE=1,90**

Είναι $R^2 = \dfrac{SSR}{SST} = \dfrac{44{,}10}{46} = 0{,}96 \Rightarrow 96\%$ και $s^2 = \dfrac{SSE}{n-2} = \dfrac{1{,}90}{3} = 0{,}633.$

Άσκηση 5.3

Για την άσκηση 5.2, να υπολογισθούν τα διαστήματα εμπιστοσύνης 95% για τα α και b. Επίσης να γίνει έλεγχος για την υπόθεση H_0: b=0 και H_1 b≠0 με στάθμη σημαντικότητας α=0,05.

Λύση:

Θα έχουμε $t_{5-2,0.025}$ = $t_{3,\ 0.025}$ = 3,182 οπότε μπορούμε να υπολογίσουμε τα Sa^2 και Sb^2:

$S_a^2 = s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right) = 0{,}633\left(\frac{1}{5} + \frac{3^2}{(5-1)2{,}5}\right) = 0{,}696 \Rightarrow S_a = 0{,}83$ και ανάλογα έχουμε

$S_b^2 = \frac{s^2}{(n-1)s_x^2} = \frac{0{,}633}{(5-1)2{,}5} = 0{,}0633 \Rightarrow S_b = 0{,}25$ οπότε τα διαστήματα εμπιστοσύνης είναι:

για το μεν α είναι $(\hat{a} \pm S_a) = (2{,}7 \pm 0{,}83) = (1{,}9\ ,\ 3{,}5)$ για το δε b $(\hat{b} \pm S_b) = (1{,}9\ ,\ 2{,}4)$.

Άσκηση 5.4

Να βρεθεί η ευθεία ελαχίστων τετραγώνων του παρακάτω δείγματος:

Y	12	13	13	14	15	15	14	16	17	18
X	2	2	3	3	4	4	5	5	6	6

Ποια θα είναι η διασπορά; Να βρεθεί διάστημα εμπιστοσύνης 90% για τη μέση πρόβλεψη για X=5. Να βρεθούν όλα τα διαστήματα εμπιστοσύνης για όλες τις δυνατές τιμές του X.

Λύση:

Λειτουργούμε κατά τα γνωστά, οπότε παίρνουμε το σύστημα:

$$\sum y_i = n\hat{a} + \hat{b}\sum x_i \Rightarrow 147 = 10\hat{a} + 40\hat{b} \quad (1)$$

$$\sum x_i y_i = \hat{a}\sum x_i + \hat{b}\sum x_i^2 \Rightarrow 611 = 40\hat{a} + 180\hat{b} \quad (2)$$

Επιλύοντας το σύστημα αυτό παίρνουμε: $\hat{a}=10{,}1$, $\hat{b}=1{,}15$. Η ευθεία των ελαχίστων τετραγώνων θα είναι η $\hat{y}=10{,}1+1{,}15x$. Ας πάρουμε το ζεύγος (15,4) για Χ=4 έχουμε Υ=14,7 (αξιόπιστο μοντέλο).

Για τη διασπορά των σφαλμάτων θα έχουμε όπως και προηγούμενα:

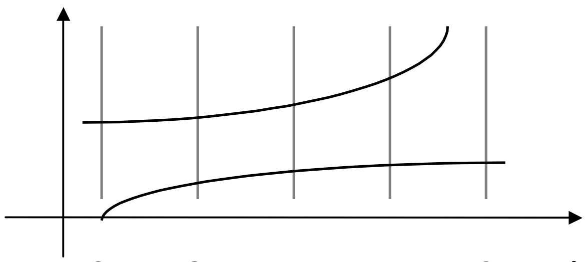
$$s^2=\frac{\sum(y-\hat{y})^2}{n-2}=\frac{(12-12{,}4)^2+(13-12{,}4)^2+(13-13{,}55)^2+(14-13{,}55)^2+(15-14{,}7)^2+}{8}$$

$$\frac{+(15-14{,}7)^2+(14-15{,}85)^2+(16-15{,}85)^2+(17-17)^2+(18-17)^2}{}=\frac{5{,}65}{8}=0{,}706$$

Θέλουμε διάστημα εμπιστοσύνης 90% με Χ=5 και άρα $\hat{y}=15{,}85$. Θα πάρουμε:

$$(\hat{y}\pm s*t_{n-2,0.05}\sqrt{\frac{1}{n}+\frac{(X-\bar{x})^2}{(n-1)S_x^2}})=(15{,}85\pm 0{,}84*1{,}86\sqrt{\frac{1}{10}+\frac{(5-4)^2}{9*2{,}22}})=(15{,}24\ ,\ 16{,}46),$$

αφού $S_x^2=\dfrac{\sum x_i^2-n\bar{x}^2}{n-1}$ =2,22. Αν επαναλάβουμε την ίδια διαδικασία για όλες τις δυνατές τιμές του Χ(2,3,4,6) θα πάρουμε τον παρακάτω πίνακα διαστημάτων εμπιστοσύνης:



2 3 4 5 6 (X)

X		
2	11,54	13,26
3	12,96	14,16
4	14,21	15,19
5	15,24	16,46
6	16,14	17,85

<u>Παρατήρηση</u>: Θα μπορούσαμε να κάνουμε τις γραφικές παραστάσεις που φαίνονται στο σχήμα. Παρατηρούμε ότι το διάστημα εμπιστοσύνης (confidence band) είναι μια ζώνη, η οποία έχει κάποιο ελάχιστο κοντά στο 4, όσο δε απομακρυνόμεθα αυξάνει το πλάτος της (αβεβαιότητα).

Άσκηση 5.5

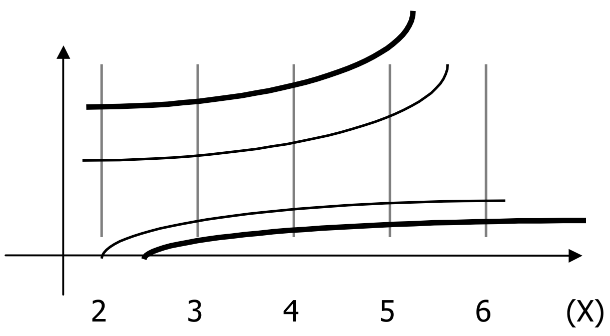
α) Ελέγξτε για την άσκηση 5.4, αν η υπόθεση H_0 E(y)=16 με H_1 E(y)>16 αν X=6 σε στάθμη σημαντικότητας α=0,05, β) να βρείτε διάστημα εμπιστοσύνης 90% για την ατομική πρόβλεψη για X=5, γ) να συγκρίνετε με το δ.ε. που βρήκατε για τη μέση πρόβλεψη, επαναλαμβάνοντας τη διαδικασία για όλες τις δυνατές τιμές του X. Τι συμπέρασμα εξάγετε;

Λύση:

α) Θα είναι $\hat{y} = 10{,}1 + 1{,}15 * 6 = 17$, επομένως και σύμφωνα με τη θεωρία θα έχουμε $t = \dfrac{\hat{y} - \delta}{s\sqrt{\dfrac{1}{n} + \dfrac{(x-\bar{x})^2}{(n-1)S_x^2}}} = \dfrac{17-16}{0{,}84\sqrt{\dfrac{1}{10} + \dfrac{(6-4)^2}{9*2{,}22}}} = \dfrac{1}{0{,}84*0{,}55} = 2{,}17$ > 1,86 = $t_{8,0.05}$ και άρα η H_0 απορρίπτεται.

β) Θα είναι $\hat{y} = 15{,}85$ και προφανώς s=0,84. Το διάστημα εμπιστοσύνης θα δοθεί εκ του τύπου: $15{,}85 \pm 1{,}86 * 0{,}84\sqrt{1 + 0{,}15} \to (14{,}17 \quad , \quad 17{,}53)$.

γ) Επαναλαμβάνοντας την ίδια διαδικασία του ερωτήματος (β) για όλες τις δυνατές τιμές του X (2,3,4,6) παίρνουμε τον παρακάτω πίνακα.



X		
2	10,62	14,18
3	11,87	15,23
4	13,06	16,34
5	14,17	17,53
6	15,22	18,78

Παρατηρούμε ότι η νέα καμπύλη (με σκούρο χρώμα) περιβάλλει την παλαιά. Αυτό σημαίνει ότι η ζώνη εμπιστοσύνης είναι στενότερη από τη ζώνη πρόβλεψης (σκούρο χρώμα). Δηλαδή η ζώνη πρόβλεψης διακατέχεται από μεγαλύτερη αβεβαιότητα και για το λόγο αυτό έχει μεγαλύτερο πλάτος.

Άσκηση 5.6

Δίνεται ο παρακάτω πίνακας:

X	0,10	0,30	0,40	0,55	0,70	0,80	0,95
Y	15,0	18,5	19,3	21,2	22,6	23,8	26,0

Να βρεθούν:

Η ευθεία ελαχίστων τετραγώνων

διάστημα εμπιστοσύνης 95% τα a,b.

i) να γίνει έλεγχος H_0 b=0 αληθής H_1 b≠0 σε στάθμη σημαντικότητας a=0,05

ii) να βρεθεί δ.ε. 95% για την ατομική πρόβλεψη Y όταν X=0,5.

Λύση:

(i) Γράφουμε τις δυο κανονικές εξισώσεις:

$$\sum y_i = na + b\sum x_i \Rightarrow 146{,}4 = 7a + 3{,}8b \quad (1)$$

$$\sum x_i y_i = a\sum x_i + b\sum x_i^2 \Rightarrow 85{,}99 = 3{,}8a + 2{,}595b \quad (2)$$

Οπότε αν λύσουμε τις (1),(2) πχ με τη μέθοδο Crammer κατά α,b θα βρούμε α=14,27 και b=12,24 επομένως η ζητούμενη ευθεία ελαχίστων τετραγώνων με κλίση b και σταθερό όρο α θα είναι η $\hat{y} = 14{,}27 + 12{,}24x$.

(ii) Θα είναι $S_b^2 = \dfrac{s^2}{(n-1)S_x^2}$, όπου η διασπορά θα βρεθεί από τα δεδομένα του προβλήματος και $S_x^2 = \dfrac{\sum x_i^2 - n\overline{x}^2}{n-1} = \dfrac{2{,}595 - 7*0{,}54^2}{7-1} = \dfrac{0{,}5538}{6} = 0{,}09$. Η διασπορά s θα ισούται με $s^2 = \dfrac{n-1}{n-2}\left(S_y^2 - \dfrac{S_{xy}^2}{S_x^2}\right) = \dfrac{6}{5}\left(13{,}63 - \dfrac{1{,}018^2}{0{,}09}\right) = 0{,}15$ αφού το S_x^2 είναι ήδη γνωστό και επίσης $S_y^2 = \dfrac{\sum y_i^2 - n\overline{y}^2}{n-1} = \dfrac{81{,}7833}{6} = 13{,}63$ και $S_{xy}^2 = \dfrac{\sum x_i y_i - n\overline{xy}}{n-1} = 1{,}018$. Οπότε και βρίσκουμε για το Sb=0,53 και Sa=0,32 άρα τα δ.ε. είναι για το μεν α: (13,45 , 15,09) για το δε b: (10,88 , 13,60)

(iii) Η υπόθεση απορρίπτεται καθόσον το μηδέν δεν ανήκει στο παραπάνω διάστημα εμπιστοσύνης για το b.

(iv) $(\hat{y} \pm st_{n-2,\alpha/2}\sqrt{1 + \dfrac{1}{n} + \dfrac{(x-\overline{x})^2}{(n-1)S_x^2}})$, οπότε αν λάβουμε υπόψη μας ότι $t_{5,0.025}$=2,571 θα πάρουμε (19,33 , 21,45) το ζητούμενο δ.ε. για την ατομική πρόβλεψη Y (X=0,5).