συσχετίσεως είναι να μετρήσει το **βαθμό** (την ένταση) της εξαρτήσεως μεταξύ των μεταβλητών Ψ και Χ. Στην Παλινδρόμηση εξετάζουμε αν υπάρχει σχέση εξαρτήσεως μεταξύ των μεταβλητών Ψ και Χ - η οποία εκφράζεται με την εξίσωση παλινδρομήσεως - ώστε να μπορέσουμε να προβλέψουμε τις τιμές της μεταβλητής Ψ με βάση τις τιμές της άλλης μεταβλητής Χ.

Το κεντρικό πρόβλημα της Συσχετίσεως είναι να βρούμε έναν κατάλληλο δείκτη (συντελεστή), ο οποίος να μας δείξει αν η σχέση εξαρτήσεως (συσχέτιση) μεταξύ των μεταβλητών Ψ και Χ είναι έντονη, μέτρια, ασθενής ή ανύπαρκτη. Ο **ποσοτικός προσδιορισμός του βαθμού** (της εντάσεως) της εξαρτήσεως μεταξύ των μεταβλητών Ψ και Χ γίνεται με μιά στατιστική παράμετρο, η οποία ονομάζεται **Συντελεστής Συσχετίσεως** (Correlation Coefficient) και συμβολίζεται με το r.

12.4.2. Συντελεστής γραμμικής συσχετίσεως

Για να μετρήσουμε την **ένταση** (το βαθμό) της εξαρτήσεως μεταξύ των μεταβλητών Ψ και Χ (ή Χ και Ψ), εισάγουμε ένα συντελεστή ανεξάρτητο των μονάδων μετρήσεως των μεταβλητών Ψ και Χ, ο οποίος, όπως είπαμε, ονομάζεται **Συντελεστής Συσχετίσεως** (= r) και ορίζεται ως μέσος γεωμετρικός των δύο συντελεστών παλινδρομήσεως b και b΄. Δηλαδή:

$$r = \sqrt{b \cdot b'} \tag{12.32}$$

Γνωρίζουμε όμως ότι:

$$b = \frac{n\sum X_i\Psi_i - (\sum X_i)(\sum \Psi_i)}{n\sum X_i^2 - (\sum X_i)^2} \quad \text{και} \quad b' = \frac{n\sum X_i\Psi_i - (\sum X_i)(\sum \Psi_i)}{n\sum \Psi_i^2 - (\sum \Psi_i)^2}$$

Αντικαθιστώντας τις τιμές των b και b΄ στη σχέση (12.32) προκύπτει ο παρακάτω τύπος:

$$\mathbf{r} = \frac{\mathbf{n\sum X_i\Psi_i - (\sum X_i)(\sum \Psi_i)}}{\sqrt{\mathbf{n\sum X_i^2 - (\sum X_i)^2}}\cdot\sqrt{\mathbf{n\sum \Psi_i^2 - (\sum \Psi_i)^2}}} \tag{12.33}$$

ο οποίος χρησιμοποιείται όταν διαθέτουμε τα αρχικά δεδομένα των μεταβλητών Χ και Ψ.

Γνωρίζουμε επίσης ότι:

$$b = \frac{\sum (X - \bar{X})(\Psi - \bar{\Psi})}{\sum (X - \bar{X})^2} \quad \text{και} \quad b' = \frac{\sum (X - \bar{X})(\Psi - \bar{\Psi})}{\sum (\Psi - \bar{\Psi})^2}$$

Αντικαθιστώντας τις τιμές των b και b΄ στη σχέση (12.32) προκύπτει ο ακόλουθος τύπος:

$$\mathbf{r} = \frac{\mathbf{\sum (X - \bar{X})(\Psi - \bar{\Psi})}}{\sqrt{\mathbf{\sum (X - \bar{X})^2}}\sqrt{\mathbf{\sum (\Psi - \bar{\Psi})^2}}} = \frac{\mathbf{Cov(X,\Psi)}}{\sqrt{\mathbf{Var(X)}}\cdot\sqrt{\mathbf{Var(\Psi)}}} \tag{12.34}$$

Η έκφραση $\sum (X - \bar{X})(\Psi - \bar{\Psi}) = \mathrm{Cov}(X, \Psi)$ ονομάζεται **Συνδιακύμανση** (Covariance) των μεταβλητών Χ και Ψ, ενώ με τα Var(X) και

Var(Ψ) συμβολίζουμε αντίστοιχα τις διακυμάνσεις των Χ και Ψ. Ο τύπος (12.34) ορίζει το συντελεστή συσχετίσεως του K. Pearson.

Επειδή $\sum (X - \bar{X})^2 = (n-1)S_X^2$ και $\sum (\Psi - \bar{\Psi})^2 = (n-1)S_\psi^2$ και $\sum (X - \bar{X})(\Psi - \bar{\Psi}) = \sum X\Psi - n\bar{X}\bar{\Psi}$, αντικαθιστώντας στη σχέση (12.34) βρίσκουμε τον ακόλουθο τύπο:

$$\mathbf{r} = \frac{\mathbf{\sum X\Psi - n\bar{X}\bar{\Psi}}}{\mathbf{(n-1)S_X \cdot S_\psi}} \tag{12.35}$$

Αν πολλαπλασιάσουμε και τα δύο μέλη της (12.35) επί S_ψ/S_X, τότε βρίσκουμε:

$$r \cdot \frac{S_\psi}{S_X} = \frac{\sum X\Psi - n\bar{X}\bar{\Psi}}{(n-1)S_X^2} = \frac{\sum (X - \bar{X})(\Psi - \bar{\Psi})}{\sum (X - \bar{X})^2} = \mathbf{b} \tag{12.36}$$

Αν τώρα πολλαπλασιάσουμε και τα δύο μέλη επί S_X/S_ψ, τότε βρίσκουμε:

$$r \cdot \frac{S_X}{S_\psi} = \frac{\sum X\Psi - n\bar{X}\bar{\Psi}}{(n-1)S_\psi^2} = \frac{\sum (X - \bar{X})(\Psi - \bar{\Psi})}{\sum (\Psi - \bar{\Psi})^2} = \mathbf{b'} \tag{12.37}$$

Οι τιμές των a και a΄ βρίσκονται από τους τύπους:

$$a = \bar{\Psi} - r\,\frac{S_\psi}{S_X}\,\bar{X}, \quad a' = \bar{X} - r\,\frac{S_X}{S_\psi}\,\bar{\Psi} \tag{12.38}$$

Ο συντελεστής συσχετίσεως (r) μετράει την ένταση της εξαρτήσεως μεταξύ των μεταβλητών Χ και Ψ, με την απαραίτητη προϋπόθεση ότι **η σχέση εξαρτήσεως είναι γραμμικής μορφής.** Συνεπώς, ο r δεν είναι κατάλληλο στατιστικό μέτρο συσχετίσεως όταν η σχέση εξαρτήσεως είναι καμπυλόγραμμη.

Ο συντελεστής συσχετίσεως (r) έχει τις ακόλουθες χαρακτηριστικές ιδιότητες:

α) Είναι ένα στατιστικό μέτρο ανεξάρτητο των μονάδων μετρήσεως των μεταβλητών Χ και Ψ. Δηλαδή, αν οι τιμές των μεταβλητών Χ και Ψ πολλαπλασιαστούν ή διαιρεθούν με μιά σταθερή ποσότητα λ, τότε ο r παραμένει αμετάβλητος.

β) Ό,τι πρόσημο (+ ή -) έχει ο συντελεστής παλινδρομήσεως b το ίδιο πρόσημο θα έχει και ο συντελεστής συσχετίσεως (r).

γ) Η τιμή του συντελεστή συσχετίσεως κυμαίνεται ανάμεσα στο -1 και στο +1. Δηλαδή:

$$\mathbf{-1 \leq r \leq +1} \tag{12.39}$$

Αν ο συντελεστής συσχετίσεως είναι **θετικός** (r > 0), τότε η **συσχέτιση** μεταξύ των μεταβλητών Ψ και Χ είναι **θετική**, δηλαδή σε κάθε **αύξηση** (μείωση) της μιάς μεταβλητής αντιστοιχεί **αύξηση** (μείωση) και της άλλης μεταβλητής. Π.χ. αύξηση του εισοδήματος (= Χ) συνεπάγεται αύξηση και της δαπάνης για ψυχαγωγία (= Ψ).

Αν ο συντελεστής συσχετίσεως είναι **αρνητικός** ($r < 0$), τότε η συσχέτιση μεταξύ των μεταβλητών Ψ και Χ είναι **αρνητική**, δηλαδή σε κάθε αύξηση της μιάς μεταβλητής αντιστοιχεί **μείωση** της άλλης μεταβλητής και αντίστροφα. Π.χ. αύξηση της τιμής (= Χ) ενός αγαθού συνεπάγεται μείωση της ζητούμενης ποσότητας του αγαθού και αντίστροφα.

Αν $r = +1$, τότε έχουμε **τέλεια θετική συσχέτιση**, δηλαδή σε κάθε μεταβολή της μιάς μεταβλητής κατά ορισμένη έννοια ακολουθεί ανάλογη μεταβολή και της άλλης μεταβλητής κατά την ίδια έννοια.

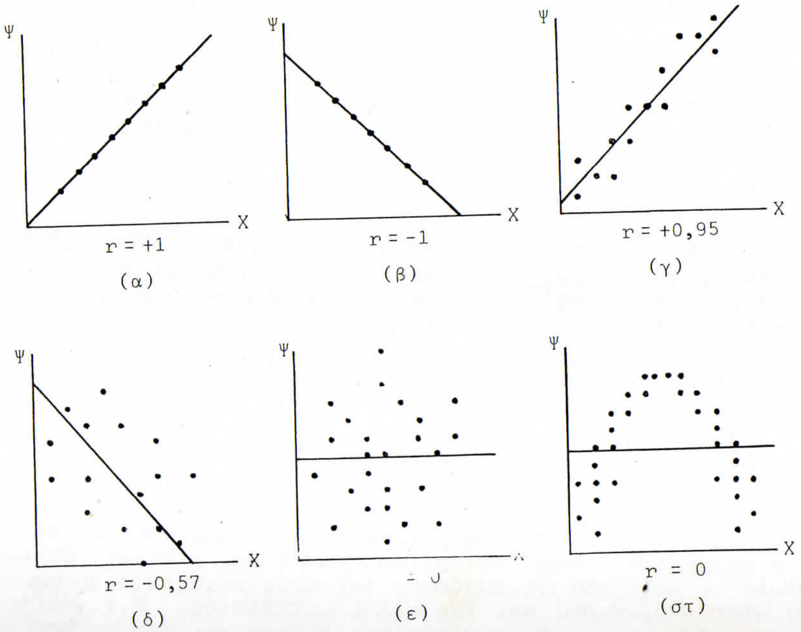
Αν $r = -1$, τότε έχουμε **τέλεια αρνητική συσχέτιση**, δηλαδή σε κάθε μεταβολή της μιάς μεταβλητής κατά ορισμένη έννοια ακολουθεί ανάλογη μεταβολή κατά την αντίθετη έννοια.

Αν $r = 0$, τότε οι μεταβλητές Χ και Ψ είναι **ασυσχέτιστες**, δηλαδή οι τιμές της μεταβλητής Ψ δεν επηρεάζονται από τις μεταβολές της μεταβλητής Χ.

Στο Διάγραμμα 12.7 απεικονίζονται διαγράμματα διασποράς, τα οποία δείχνουν διάφορους βαθμούς συσχετίσεως μεταξύ των μεταβλητών Ψ και Χ.

Διάγραμμα 12.7

Διαγράμματα διασποράς που δείχνουν διάφορους βαθμούς συσχετίσεως



Τα διαγράμματα διασποράς (α) και (β), του Διαγράμματος 12.7, δείχνουν τέλεια θετική και αρνητική συσχέτιση. Τα διαγράμματα (γ) και (δ) δείχνουν αντίστοιχα έντονη θετική και μέτρια αρνητική συσχέτιση. Το διάγραμμα (ε) δείχνει ότι οι μεταβλητές Ψ και Χ είναι ασυσχέτιστες ($r = 0$), ενώ το διάγραμμα (στ) δείχνει ότι δεν υπάρχει γραμμική συσχέτιση ($r = 0$), ενδέχεται όμως να υπάρχει καμπυλόγραμμη συσχέτιση.

Για την καλύτερη κατανόηση και εμπέδωση της συσχετίσεως παραθέτουμε δύο παραδείγματα.

Παράδειγμα 1. Χρησιμοποιούμε τα δεδομένα του παραδείγματος 1 της παραγράφου 12.2.2. Έχουμε:

$n = 10$, $\sum X_i \Psi_i = 4084$, $\sum X_i = 220$, $\sum \Psi_i = 240$, $\sum X_i^2 = 6020$,

$\sum \Psi_i^2 = 7026$, $b = -1{,}014$, $b' = -0{,}945$.

Αντικαθιστούμε στον τύπο (12.32) και βρίσκουμε:

$$r = \sqrt{b \cdot b'} = \sqrt{(-1{,}014)(-0{,}945)} = \sqrt{0{,}958} = -0{,}98$$

(Ο r παίρνει αρνητικό πρόσημο, γιατί $b < 0$).

Αν τώρα αντικαταστήσουμε τα παραπάνω δεδομένα στον τύπο (12.33), θα έχουμε:

$$r = \frac{10 \times 4084 - 220 \times 240}{\sqrt{10 \times 6020 - (220)^2}\,\sqrt{10 \times 7026 - (240)^2}} =$$

$$= \frac{-11.960}{108{,}63 \times 112{,}52} = \frac{-11.960}{12.223{,}047} = -0{,}98$$

Από την τιμή $r = -0{,}98$ συμπεραίνουμε ότι η συσχέτιση μεταξύ ζητούμενης ποσότητας (= Ψ) ενός αγαθού και της τιμής (= Χ) αυτού είναι αρνητική. Δηλαδή, όταν αυξάνεται η τιμή του αγαθού μειώνεται η ζητούμενη ποσότητα και αντίστροφα. Από την απόλυτη τιμή $|r| = |-0{,}98|$ συνάγεται ότι η συσχέτιση είναι εντονότατη, γιατί πλησιάζει τη μονάδα.

Παράδειγμα 2. Δεδομένα του παραδείγματος 2 της παραγράφου 12.2.2.

Έχουμε: $n = 10$, $\sum X_i = 290$, $\sum \Psi_i = 80$, $\sum X_i \Psi_i = 2350{,}3$,

$\sum X_i^2 = 8810$, $\sum \Psi_i^2 = 642{,}32$, $b = 0{,}0757$, $b' = 13{,}060$

Αντικαθιστούμε στον τύπο (12.32) και βρίσκουμε:

$$r = \sqrt{b \cdot b'} = \sqrt{0{,}0757 \times 13{,}06} = \sqrt{0{,}988} = 0{,}99$$

Από την τιμή $r = +0{,}99$ συμπεραίνουμε ότι η συσχέτιση μεταξύ κατά κεφαλήν κατανάλωσης (= Ψ) και κατά κεφαλήν εισοδήματος (= Χ) είναι θετική και εντονότατη, γιατί η τιμή του r είναι κοντά στη μονάδα.

12.4.3. Συντελεστής προσδιορισμού

Στην προηγούμενη παράγραφο μιλήσαμε για το συντελεστή συσχετίσεως (= r), ο οποίος μετράει την ένταση (το βαθμό) της εξαρτήσεως μεταξύ των μεταβλητών Χ και Ψ, όταν η σχέση

Έλεγχος της H_0 με το κριτήριο:

$$Z = \frac{p_1 - p_2}{S_{p_1 - p_2}}$$

όπου: $S_{p_1-p_2} = \sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = \sqrt{0{,}14(1-0{,}14)\left(\frac{1}{1000}+\frac{1}{500}\right)} = 0{,}019$

και $\quad p = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2} = \frac{0{,}15 \times 1000 + 0{,}12 \times 500}{1000 + 500} = 0{,}14$

Αντικαθιστώντας στο κριτήριο βρίσκουμε:

$$Z = \frac{0{,}15 - 0{,}12}{0{,}019} = 1{,}58 < 2{,}58$$

Η υπόθεση H_0 γίνεται δεκτή. Οι δύο μέθοδοι βγάζουν το ίδιο ποσοστό ελαττωματικών προϊόντων.

8.6. ΚΑΤΑΝΟΜΗ t-STUDENT

Τα προβλήματα Στατιστικής Επαγωγής (εκτίμηση παραμέτρων του πληθυσμού και έλεγχοι στατιστικών υποθέσεων), τα οποία έχουμε επιλύσει μέχρι τώρα, αναφέρονται σε **πολυπληθή** δείγματα. Είχαμε πει ότι όταν το μέγεθος n του δείγματος είναι αρκετά μεγάλο ($n > 30$ και κυρίως $n > 100$), τότε οι κατανομές δειγματοληψίας των διαφόρων παραμέτρων (μέσων όρων, ποσοστών) ακολουθούν (προσεγγιστικά) την Κανονική Κατανομή, βάσει του κεντρικού οριακού θεωρήματος.

Όταν όμως τα δείγματα είναι **ολιγοπληθή** ($n < 30$), αποδεικνύεται ότι οι κατανομές δειγματοληψίας των πιό πάνω παραμέτρων δεν είναι πλέον ούτε κατά προσέγγιση "κανονικές". Επομένως, δεν επιτρέπεται να εφαρμόζονται οι μέθοδοι των προηγούμενων παραγράφων, οι οποίες ισχύουν μόνον αν το δείγμα είναι πολυπληθές. Επιβάλλεται λοιπόν η μελέτη της κατανομής δειγματοληψίας των πιό πάνω παραμέτρων όταν $n < 30$ και η εξαγωγή χρήσιμων συμπερασμάτων.

Αν από έναν **κανονικό πληθυσμό** με ορισμένο μέσο **μ** και **άγνωστη διακύμανση** σ^2 ληφθούν ισοπληθή δείγματα μεγέθους n και από κάθε δείγμα υπολογίσουμε το μέσο $\bar{x}$ και τη διακύμανση $s^2 = \sum(X_i - \bar{x})^2/n - 1$, τότε η κατανομή δειγματοληψίας της παραμέτρου:

$$t_\nu = \frac{\bar{x} - \mu}{s/\sqrt{n}} \qquad -\infty < t < +\infty \tag{8.55}$$

είναι γνωστή σαν Κατανομή **t** ή Κατανομή **t - Student**.

Η Κατανομή **t** ανακαλύφτηκε στα 1909 από τον Άγγλο Στατιστικό (Χημικό) William C. Gosset, ο οποίος έγραφε με το ψευδώνυμο Student, γιατί τότε ήταν υπάλληλος ενός Ιρλανδού ζυθοποιού, ο οποίος απαγόρευε στους υπαλλήλους του να κάνουν δημοσιεύσεις.

Η συνάρτηση πιθανότητας της Κατανομής **t** έχει την ακόλουθη μορφή:

$$f(t_\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\cdot\pi}\cdot\Gamma\left(\frac{\nu}{2}\right)}\cdot\left(1+\frac{t^2}{\nu}\right)^{-\frac{(\nu+1)}{2}} \quad -\infty < t < +\infty \quad (8.56)$$

όπου: $\nu = n - 1$ το πλήθος των βαθμών ελευθερίας και

$$\Gamma\left(\frac{\nu}{2}\right) = \left(\frac{\nu}{2}-1\right)! = \left(\frac{\nu}{2}-1\right)\left(\frac{\nu}{2}-2\right)\ldots 3\cdot 2\cdot 1, \text{ αν } \nu = \text{άρτιος}$$

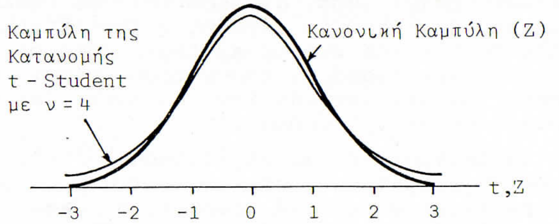
$$= \left(\frac{\nu}{2}-1\right)\left(\frac{\nu}{2}-2\right)\ldots\frac{3}{2}\cdot\frac{1}{2}\cdot\sqrt{\pi}, \text{ αν } \nu = \text{περιττός}$$

παριστάνει τη λεγόμενη Γάμμα Συνάρτηση.

Από τη συνάρτηση (8.56) προκύπτει ότι πρόκειται για ολόκληρη οικογένεια καμπυλών, οι οποίες εξαρτώνται από τα ν, δηλ. από το μέγεθος των δειγμάτων. Για μικρά δείγματα η κατανομή t είναι πλατύκυρτη, αλλά όσο αυξάνει το μέγεθος του δείγματος n, η Κατανομή t τείνει προς την Κανονική Κατανομή (Βλ. Διάγραμμα 8.8).

Διάγραμμα 8.8

Καμπύλες Κανονικής Κατανομής και Κατανομής t-Student



Ο μέσος της Κατανομής t ισούται με το μηδέν, δηλ. $E(t) = 0$ και η διακύμανση $\sigma_t^2 = \nu/(\nu-2)$. Η Κατανομή t είναι συμμετρική και η κύρτωση είναι μικρότερη του 3 (πλατύκυρτη), αλλά προσεγγίζει το 3 όσο αυξάνει το ν. Για $\nu > 30$, η Κατανομή t προσεγγίζει τη μορφή της Κανονικής Καμπύλης.

Από τα παραπάνω συμπεραίνουμε ότι, για να εφαρμοστεί η Κατανομή t - Student, πρέπει να υπάρχουν οι εξής προϋποθέσεις:

α) Ο πληθυσμός, από τον οποίο έχει ληφθεί το δείγμα, να είναι **κανονικός**, δηλαδή να ακολουθεί την Κανονική Κατανομή.

β) Το δείγμα (ή τα δείγματα) να είναι ολιγοπληθές ($\nu<30$).

γ) Η διακύμανση του πληθυσμού να είναι άγνωστη.

Για τις πρακτικές εφαρμογές υπάρχουν Πίνακες της Κατανομής t - Student (Βλ. Πίνακα IV, Κριτικές Τιμές Κατανομής t - Student στο τέλος του βιβλίου). Στοιχεία εισόδου στον Πίνακα IV είναι: 1) οι βαθμοί ελευθερίας $\nu = 1,2,3,\ldots,120$ που βρίσκονται στην πρώτη στήλη του Πίνακα IV και 2) τα Επίπεδα Σημαντικότητας (α) που είναι στην πρώτη γραμμή του Πίνακα IV.

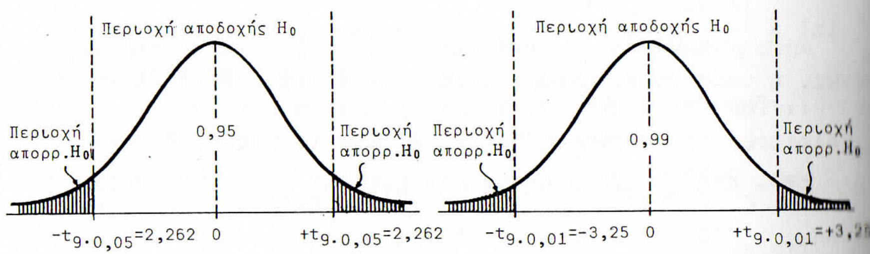
Παράδειγμα: Για δείγμα $n = 10$ έχουμε $\nu = n - 1 = 10 - 1 = 9$ βαθμούς ελευθερίας. Στα επίπεδα σημαντικότητας $\alpha = 0{,}05$ και $\alpha = 0{,}01$ αντιστοιχούν οι κριτικές τιμές:

$$t_{9\cdot 0,05} = \pm 2{,}262 \text{ και } t_{9\cdot 0,01} = \pm 3{,}250$$

Οι περιοχές αποδοχής και απορρίψεως μιάς υποθέσεως H_0, για δίπλευρο κριτήριο ελέγχου, απεικονίζονται στο Διάγραμμα 8.9.

Διάγραμμα 8.9

Περιοχές αποδοχής και απορρίψεως υποθέσεως H_0 σε δίπλευρο κριτήριο ελέγχου



8.7. ΕΛΕΓΧΟΙ ΣΤΑΤΙΣΤΙΚΩΝ ΥΠΟΘΕΣΕΩΝ ΜΕ ΟΛΙΓΟΠΛΗΘΗ ΔΕΙΓΜΑΤΑ

8.7.1. Έλεγχος διαφοράς μέσου δείγματος και μέσου πληθυσμού

Ισχύουν όσα είπαμε για τον έλεγχο στατιστικών υποθέσεων με πολυπληθή (μεγάλα) δείγματα, με τη διαφορά ότι οι τιμές των διαφόρων κριτηρίων ελέγχου συγκρίνονται με τις κριτικές τιμές που παίρνουμε από τον Πίνακα IV (Κριτικές Τιμές Κατανομής t - Student).

Παράδειγμα 1. Δοκιμάζεται ένας νέος τύπος πυραύλων για να διαπιστωθεί αν είναι καλύτερος από έναν παλαιό τύπο. Το μέσο βαληνεκές του παλαιού τύπου είναι 340 μίλια. Παίρνουμε τυχαίο δείγμα 10 πυραύλων νέου τύπου και βρίσκουμε μέσο βεληνεκές 360 μίλια και μέση απόκλιση τετραγώνου 20 μίλια. Αν η κατανομή των βεληνεκών των πυραύλων ακολουθεί την Κανονική Κατανομή, να ελεγχθεί αν ο νέος τύπος πυραύλων είναι καλύτερος από τον παλαιό τύπο. $\alpha = 0{,}05$.

Δεδομένα: $\mu = 340$, $n = 10$, $\bar{x} = 360$, $s = 20$, $\alpha = 0{,}05$