ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ
ΤΜΗΜΑ ΑΓΡΟΤΙΚΗΣ ΑΝΑΠΤΥΞΗΣ

Γ Ε Ω Ρ Γ Ι Κ Ο Σ   Π Ε Ι Ρ Α Μ Α Τ Ι Σ Μ Ο Σ

ΚΟΥΤΡΟΥΜΑΝΙΔΗΣ Θ.
Αν. Καθηγητής Δ.Π.Θ.

ΖΑΦΕΙΡΙΟΥ Ε.
Υπ. Διδάκτορας

Ορεστιάδα 2007

ΠΕΡΙΕΧΟΜΕΝΑ

Κεφάλαιο 1^{o}

Δειγματοληψία

1.1. Βασικές Έννοιες

Δείγμα: Ένα γνήσιο υποσύνολο των στοιχείων του πληθυσμού ή των τιμών της μεταβλητής.

Δειγματοληψία: Η διαδικασία λήψης ενός δείγματος, δηλαδή ενός μέρους του συνόλου των πληροφοριών που μας ενδιαφέρουν.

Η ευρεία χρήση των δειγματοληψιών στην μελέτη των πληθυσμών αποδίδεται στους εξής λόγους:

- Η μεγάλη χρονική διάρκεια λήψης δεδομένων από έναν πληθυσμό καθιστά τα στοιχεία που έχουν ληφθεί να μην ανταποκρίνονται στην πραγματικότητα
- Τα σφάλματα μέτρησης και επεξεργασίας αυξάνουν με τον αριθμό των παρατηρήσεων με αποτέλεσμα οι επιπλέον πληροφορίες να μην δίνουν την πραγματική εικόνα του πληθυσμού.
- Δεν είναι δυνατή η χρήση της απογραφής όταν η παρατήρηση συνεπάγεται την καταστροφή της εξεταζόμενης μονάδας.

- Το κόστος παρατήρησης σε περίπτωση πληθυσμών είναι ιδιαίτερα υψηλό.

Η δειγματοληψία παρέχει πολλές πληροφορίες και σε ιδιαίτερα χαμηλό κόστος υπό τον όρο ότι το δείγμα είναι αντιπροσωπευτικό.

Ένα δείγμα είναι αντιπροσωπευτικό όταν το κάθε στοιχείο του πληθυσμού έχει την ίδια πιθανότητα να επιλεγεί ως στοιχείο του δείγματος.

Η εξαγωγή συμπερασμάτων για τα χαρακτηριστικά του δείγματος με δεδομένη την γνώση των χαρακτηριστικών του πληθυσμού και μέσω μαθηματικής συλλογιστικής αποτελεί την *απαγωγική συμπερασματική*

Η δε αντίστροφή διαδικασία δηλαδή η εξαγωγή συμπερασμάτων για το σύνολο με δεδομένη τη γνώση για το υποσύνολο αποτελεί την *επαγωγική στατιστική*.

Τυχαία μεταβλητή ορίζουμε μία συνάρτηση X έτσι ώστε σε κάθε στοιχείο s του δειγματικού χώρου να αντιστοιχεί ένας πραγματικός αριθμός X(s).

1.2. Κατανομή πιθανότητας Διακριτών τυχαίων μεταβλητών

Αν X είναι μία διακριτή τυχαία μεταβλητή τότε η συνάρτηση μάζας πιθανότητας ή η συνάρτηση πιθανότητας ή κατανομή πιθανοτήτων της X είναι η συνάρτηση η οποία σε κάθε τιμή x_i της X αντιστοιχεί την πιθανότητα εμφάνισης της τιμής x_i δηλαδή την $P(x_i) \equiv P(X = x_i)$ και ικανοποιεί τις συνθήκες:

$$P(x_i) \geq 0 \ \forall x_i$$
$$\sum_i P(x_i) = 1$$

Αθροιστική συνάρτηση πυκνότητας F της τυχαίας μεταβλητής X είναι η συνάρτηση που αντιστοιχεί σε κάθε τιμή x_i της X την αθροιστική πιθανότητα**:** $F(x_i) = \sum_{x_i} P(X \leq x_i)$

Βασικές ιδιότητες αυτής είναι οι εξής:

I) $0 \leq F(x_i) \leq 1 \ \forall x_i$

II) Αν $x_1 < x_2$ τότε $F(x_1) \leq F(x_2)$

1.3. Κατανομή Πιθανότητας Τυχαίων Συνεχών Μεταβλητών

Αν X είναι μία τυχαία συνεχής μεταβλητή τότε η συνάρτηση f(x) ονομάζεται συνάρτηση πυκνότητας πιθανότητας ή απλώς συνάρτηση πυκνότητας της X και ορίζουμε ότι:

$$P(x_1 < x < x_2) = \int_{x_1}^{x_2} f(x)dx$$

ισχύουν δε τα εξής:

I) $f(x) \geq 0 \ \forall x \in (-\infty, \infty)$

II) $\int_{-\infty}^{\infty} f(x)dx = 1$

Η **αθροιστική συνάρτηση πιθανότητας** F(x) ή συνάρτηση κατανομής μιας συνεχούς τυχαίας μεταβλητής χ ορίζεται ως:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

Βασικές ιδιότητες αυτής είναι οι εξής:

$$\lim_{x \to -\infty} F(x) = 0 \quad \kappa\alpha\iota \lim_{x \to +\infty} F(x) = 1$$

$$\forall\, x_1 < x_2,\, F(x_1) \leq F(x_2)$$

$$f(x) = \frac{d}{dx}F(x)\ \alpha\phi o\acute{\upsilon} F(x) = \int_{-\infty}^{x} f(t)$$

$$\kappa\alpha\iota\ P(x_1 \leq x \leq x_2) = P(X \leq x_2) - P(X \leq x_1) = F(x_2) - F(x_1)$$

Η κατανομή πιθανοτήτων μιας τυχαίας μεταβλητής ονομάζεται κατανομή πληθυσμού.

Οι δε παράμετροι της Χ ονομάζονται **παράμετροι του πληθυσμού.**

Αν Χ μία τυχαία μεταβλητή με ορισμένη κατανομή πιθανοτήτων και $X_1, X_2, X_3, \ldots, X_\nu$ ανεξάρτητες τυχαίες μεταβλητές κάθε μία από τις οποίες έχει την ίδια κατανομή πιθανοτήτων με την Χ τότε $X_1, X_2, X_3, \ldots, X_\nu$ ονομάζονται **τυχαίο δείγμα** μεγέθους ν από την μεταβλητή Χ.

Μία τιμή του δείγματος συμβολίζεται με $X_1, X_2, X_3, \ldots, X_\nu$ και ονομάζεται δειγματικό σημείο αφού οι τιμές μπορούν να θεωρηθούν ως οι συντεταγμένες ενός σημείου στον ν-διάστατο Ευκλείδειο χώρο.

Στην περίπτωση συνεχών μεταβλητών με συνάρτηση πυκνότητας πιθανότητας την $f_x(x)$ η κοινή συνάρτηση πυκνότητας πιθανότητας θα δίνεται από την σχέση:

$$h(x_1, x_2, ...x_n) = f_x(x_1) f_x(x_2), ... f_x(x_n)$$

Σε περίπτωση διακριτών μεταβλητών με συνάρτηση μάζας πιθανότητας $P_x(x)$, τότε η κοινή συνάρτηση πιθανότητας θα δίνεται από την σχέση:

$$P(X_1 = x_1, X_2 = x_2, ...X_n = x_n) = P_x(x_1) P_x(x_2) .....P_x(x_n)$$

Η κοινή συνάρτηση πιθανότητας παρέχει την πιθανότητα να προκύψουν οι τιμές του δείγματος με μία ορισμένη σειρά.

Για τον ορισμό ενός τυχαίου δείγματος απαιτείται μία τυχαία μεταβλητή Χ. Αν αυτό γίνει για ένα τυχαίο πείραμα το οποίο επαναλαμβάνεται άπειρες φορές τότε στατιστικό πληθυσμό αποτελούν οι τιμές του Χ σε όλες τις επαναλήψεις (ο αριθμός αυτός είναι άπειρος).

Για πεπερασμένο αριθμό στοιχείων με δεδομένο ότι το χαρακτηριστικό ομαδοποιείται σε ορισμένη κατανομή σχετικών συχνοτήτων τότε με τυχαία λήψη εισάγεται και η τυχαιότητα στην Χ με αποτέλεσμα η κατανομή σχετικών συχνοτήτων να είναι η κατανομή πιθανοτήτων της Χ.

Βασικά χαρακτηριστικά που πρέπει να πληροί μία διαδικασία δειγματοληψίας για να θεωρηθεί τυχαίο δείγμα είναι τα εξής:

- Κάθε δυνατό δείγμα ν στοιχείων θα πρέπει να έχει την ίδια πιθανότητα να ληφθεί
- Οι λήψεις των στοιχείων να είναι ανεξάρτητες
- Η κατανομή του πληθυσμού να είναι η ίδια σε κάθε λήψη

1.4. Η επιλογή ενός τυχαίου δείγματος

1.4.1. Η επιλογή τυχαίου δείγματος από πεπερασμένο πληθυσμό

Δειγματοληψία χωρίς επανατοποθέτηση. Οι λήψεις των στοιχείων είναι ανεξάρτητες.

Η δειγματοληψία αυτή χρησιμοποιείται όταν θέλουμε να πάρουμε μία επιπλέον πληροφορία από ένα στοιχείο που δεν έχει ήδη ληφθεί.

Όταν το μέγεθος του πληθυσμού είναι μεγάλο παίρνουμε πίνακες και επιλέγουμε με τυχαίο τρόπο στοιχεία αυτού.

Όταν είναι μικρό και τα στοιχεία μπορούν να αριθμηθούν βάζουμε π.χ. αριθμημένα χαρτάκια σε ένα καπέλο τα αριθμούμε και επιλέγουμε τυχαία κάθε φορά ένα χαρτάκι.

Δεν αποτελούν τυχαίο δείγμα ενός πληθυσμού όταν για παράδειγμα στεκόμαστε σε ένα πολυσύχναστο σημείο ενός δρόμου και ρωτάμε τους περαστικούς για ένα δεδομένο χρονικό διάστημα.

1.4.2. Η επιλογή ενός τυχαίου δείγματος από άπειρο πληθυσμό

Πρόκειται για μία θεωρητική κατασκευή στην οποία θεωρούμε πως ένα ορισμένο πείραμα τύχης επαναλαμβάνεται απεριόριστα κάτω από τις ίδιες βασικές συνθήκες οπότε έχουμε ένα άπειρο πληθυσμό δυνατών αποτελεσμάτων.

Επομένως αν Χ είναι μία τυχαία μεταβλητή που ορίζεται για το πείραμα αυτό, το τυχαίο δείγμα μεγέθους ν από την Χ αντιστοιχεί σε ν επαναληπτικές μετρήσεις της Χ που γίνονται κάτω από τις ίδιες βασικές συνθήκες και το αποτέλεσμα της μιας δεν επηρεάζεται από τις άλλες.

1.5. Αξιοπιστία του τυχαίου δείγματος

Βασικό χαρακτηριστικό του δείγματος πρέπει να είναι η αντιπροσωπευτικότητα του, δηλαδή να αποτελούν μία μικρογραφία του πληθυσμού.

Στόχος των μεθόδων της δειγματοληψίας είναι η αύξηση της αντιπροσωπευτικότητας του δείγματος.

Στην περίπτωση που από ένα πληθυσμό λαμβάνεται ένα τυχαίο δείγμα η πιθανότητα να προκύψει ένα μη αντιπροσωπευτικό δείγμα είναι μικρή και μετρήσιμη.

Η δε τυχαία δειγματοληψία προστατεύει τους ερευνητές από ένα μη αντιπροσωπευτικό δείγμα αλλά μόνο πιθανοκρατικά.

Σφάλμα δειγματοληψίας: Η διαφορά μεταξύ της εκτίμησης μιας παραμέτρου του πληθυσμού που προκύπτει από τις παρατηρήσεις αυτής και της πραγματικής τιμής της παραμέτρου αποτελεί σφάλμα δειγματοληψίας. Η δε πιθανότητα να εμφανισθεί η διαφορά αυτή είναι θετική. Η ύπαρξη της διαφοράς αυτής οφείλεται στο ότι δεν εξετάζουμε ολόκληρο τον πληθυσμό αλλά μόνο μέρος αυτού.

Όταν η διαφορά μεταξύ της πραγματικής τιμής της παραμέτρου και της εκτίμησης αυτής αποδίδεται σε ατέλειες στην διαδικασία συλλογής των παρατηρήσεων του δείγματος αποτελεί συστηματικό σφάλμα δειγματοληψίας. Ειδικότερα σφάλμα δειγματοληψίας μπορεί να προκληθεί στις ακόλουθες περιπτώσεις:

- Όταν το ερωτηματολόγιο με το οποίο συλλέγονται οι πληροφορίες περιλαμβάνει κακοδιατυπωμένες ερωτήσεις
- Όταν η συλλογή των στοιχείων γίνεται από μη ειδικευμένο προσωπικό
- Όταν αγνοούνται τα άτομα που αρνούνται να απαντήσουν.

Τα άτομα με υψηλότερο επίπεδο εκπαίδευσης έχουν την τάση να απαντούν ενώ τα άτομα με χαμηλό εισόδημα και επίπεδο εκπαίδευσης αρνούνται να απαντήσουν σε ερωτηματολόγια με αποτέλεσμα το δείγμα να μην είναι αντιπροσωπευτικό όλων των κατηγοριών του πληθυσμού.

- Όταν η επιλογή γίνεται από ένα πλαίσιο δηλαδή μία εύχρηστη καταχώρηση στοιχείων του πληθυσμού το οποίο μπορεί να μην περιλαμβάνει ορισμένες κατηγορίες πληθυσμού. Το μέγεθος του σφάλματος είναι συνάρτηση του ποσοστού των ατόμων του πληθυσμού που δεν είναι στον κατάλογο
- Συστηματικό σφάλμα εισάγεται όταν η δειγματοληψία δεν γίνεται με τυχαίο τρόπο αλλά με την κρίση του ερευνητή ή όταν επιχειρείται τυχαία δειγματοληψία χωρίς να ικανοποιούνται οι συνθήκες της.

Κάθε πραγματική συνάρτηση των τιμών του δείγματος αποτελεί **στατιστικό του δείγματος** .

Παράδειγμα στατιστικών αποτελεί ο **δειγματικός μέσος** $\bar{x} = \sum_{i=1}^{n} x_i / n$ και η **δειγματική διακύμανση** $s^2 = \sum^{n} (x_i - \bar{x})^2 / (n-1)$.

Επομένως το στατιστικό ως συνάρτηση τυχαίων μεταβλητών είναι τυχαία μεταβλητή η κατανομή της οποίας ονομάζεται κατανομή δειγματοληψίας .

Το **στατιστικό** που χρησιμοποιείται για την εκτίμηση μιας άγνωστης παραμέτρου του πληθυσμού ονομάζεται **εκτιμητής** και μία ορισμένη τιμή αυτού ονομάζεται **εκτίμηση**.

Κεφάλαιο 2^{ο}

Θεωρητικές και Παράγωγες Κατανομές στην δειγματοληψία

2.1. Γενικά

Μεταβλητή: Κάθε χαρακτηριστικό το οποίο μπορεί να μετρηθεί.

Τιμές της μεταβλητής είναι οι ατομικές μετρήσεις της μεταβλητής.

Παράμετρος: Η τιμή ενός χαρακτηριστικού που πρέπει να εκτιμηθεί με τη χρήση μέρους του πληθυσμού το οποίο καλείται **δείγμα.**

Στατιστικές: Οι ποσότητες που προσδιορίζουν το δείγμα.

Κατανομές Συχνότητας: Αποδίδει τον τρόπο με τον οποίο κατανέμονται οι παρατηρήσεις.

2.2. Κανονική Κατανομή ή κατανομή Gauss

Περιγραφή της Κανονικής Κατανομής

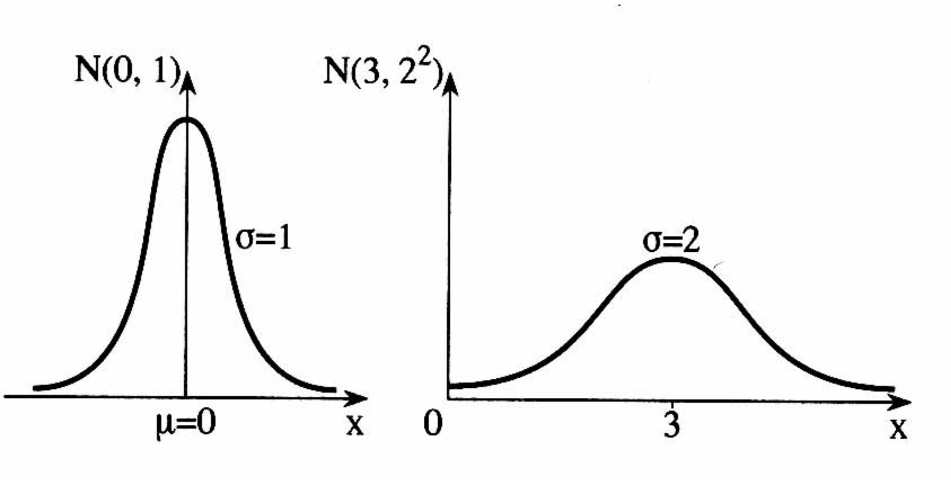
Η Κανονική Κατανομή της πιθανότητας των τιμών της τυχαίας μεταβλητής X απεικονίζεται γραφικά με τη μορφή κωδωνοειδούς καμπύλης η οποία φανερώνει την συχνή εμφάνιση των μεσαίων τιμών της X και την προοδευτική ελάττωση της συχνότητας εμφάνισης των ακραίων τιμών της X.

Ισχύουν οι ακόλουθες σχέσεις στην κανονική κατανομή:

$$P(Z \leq z_o) = \Phi(z_0)$$

$$P(Z \geq z_0) = 1 - P(Z \leq z_o) = 1 - \Phi(z_0)$$

$$P(z_1 \leq Z \leq z_2) = \Phi(z_2) - \Phi(z_1)$$



Σχήμα 1: X ~ N(0,1) .Τυποποιημένη Κανονική Κατανομή και η αντίστοιχή της Κανονική Κατανομή N(μ =3, $\sigma^2 = 2^2$)

Εφαρμογή της κανονικής κατανομής (ενδεικτικές αναφορές)

- Χρήση της κανονικής κατανομής γίνεται για να ελεγχθεί αν **μία τιμή προέρχεται από έναν πληθυσμό.**

Με **σφάλμα α** δίνεται **η απόσταση – θέση** της τιμής x μιας μεταβλητής X από το μέσο όρο μ του πληθυσμού των τιμών της X, που θεωρούμε ότι κατανέμονται κανονικά, εκφρασμένη σε αριθμό τυπικών αποκλίσεων σ του πληθυσμού μέσω της σχέσης:

$$x = \mu - z_{\frac{a}{2}}\sigma_{\bar{x}} \quad ή \quad x = \mu + z_{\frac{a}{2}}\sigma_{\bar{x}}$$

Το $z_{\frac{\alpha}{2}}$ – Κριτήριο. Υπολογίζει την πιθανότητα προς τις δύο πλευρές της καμπύλης ώστε το σφάλμα α να αποτελείται κατά το μισό από τις μεγαλύτερες θετικές τιμές και το άλλο μισό από τις μεγαλύτερες αρνητικές τιμές. Άρα η εκτίμηση του χώρου μεταξύ των δύο σημείων δίνει το ποσοστό των τιμών που περιέχονται ανάμεσα σε δύο σημεία, δηλαδή την πιθανότητα που υπάρχει μία τιμή να περιλαμβάνεται μεταξύ των σημείων αυτών. Πολλές φορές δίνεται η πιθανότητα μεταξύ ενός σημείου και μέχρι το άκρο της κατανομής, δίνεται δηλαδή η πιθανότητα που υπάρχει μία τιμή να είναι μεγαλύτερη ή μικρότερη του σημείου αυτού.

Με την συγκεκριμένη πιθανότητα σφάλματος α λαμβάνουμε τα όρια εμπιστοσύνης ($\bar{x} - z_{\frac{a}{2}}\sigma_{\bar{x}}$, $\bar{x} + z_{\frac{a}{2}}\sigma_{\bar{x}}$) του μέσου μ του πληθυσμού. Τότε το x δεν προέρχεται από τον εν λόγω πληθυσμό με σφάλμα α αν βρίσκεται εκτός του διαστήματος ($\bar{x} - z_{\frac{a}{2}}\sigma_{\bar{x}}$, $\bar{x} + z_{\frac{a}{2}}\sigma_{\bar{x}}$). Όπου $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ **το τυπικό σφάλμα (τυπική απόκλιση)** των μέσων όρων των μεγάλων δειγμάτων από τον πληθυσμό.

Η ποσότητα $z_{\frac{a}{2}}\sigma_{\bar{x}}$ καλείται **ελάχιστη σημαντική διαφορά (ΕΣΔ).**

- Χρήση της κανονικής κατανομής γίνεται για να εκφρασθεί η δειγματική κατανομή του μέσου $\bar{x}$ πολλών ίσων δειγμάτων μεγάλου μεγέθους

Οι δειγματικές μέσες τιμές πολλών μεγάλων δειγμάτων $\bar{x}_i$ από ένα πληθυσμό δημιουργούν την μεταβλητή $\bar{X}$ η οποία ακολουθεί την κανονική κατανομή Ν(μ,σ^2/n) και συνεπώς καταλήγουμε στην **Τυποποιημένη Κανονική Κατανομή** Ζ :

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Όπου $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ το **τυπικό σφάλμα** των μέσων όρων των μεγάλων δειγμάτων:

- Χρήση της κανονικής κατανομής γίνεται για να εκφρασθεί η δειγματική κατανομή της διαφοράς των δύο μέσων όρων, δύο μεγάλου μέγεθος δειγμάτων από δύο πληθυσμούς αντίστοιχα.

Για τη σύγκριση των δύο μέσων όρων δύο μεγάλων δειγμάτων γίνεται χρήση της τυχαίας μεταβλητής Ζ η οποία ακολουθεί την τυπική κανονική κατανομή:

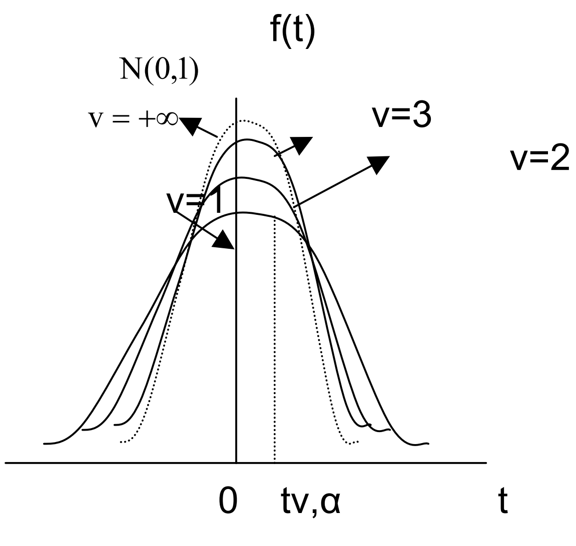
$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Όπου $\sigma_d = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ το **τυπικό σφάλμα** της διαφοράς των δύο μέσων όρων των μεγάλων δειγμάτων.

2.3. Η t - Κατανομή.

Περιγραφή της t – Κατανομής

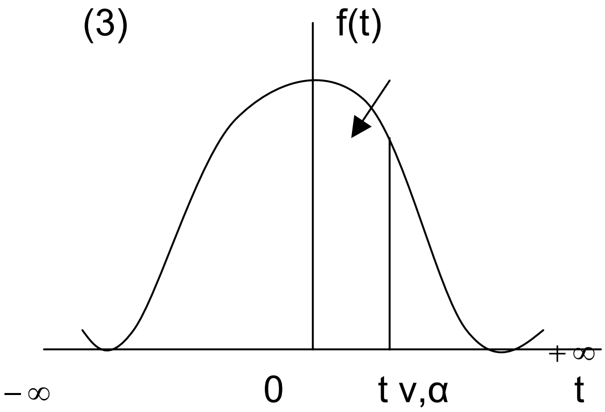
Η t – Κατανομή παρουσιάζει τις μορφές του παρακάτω σχήματος 2 , όπου v (v = n – 1) οι βαθμοί ελευθερίας του δείγματος μεγέθους n, του οποίου μελετάμε την κατανομή πιθανότητας των τιμών του.



Σχήμα 2: Η t - student Κατανομή

Όταν ο ν είναι μεγάλος η κατανομή t τείνει στην τυπική κανονική κατανομή Ν (0,1).

Έχουμε την δυνατότητα να πάρουμε από πίνακες τις τιμές $t_{v,\alpha}$ της t για τις οποίες γνωρίζουμε την πιθανότητα: $P[t \leq t_{v,\alpha}] = \alpha$.

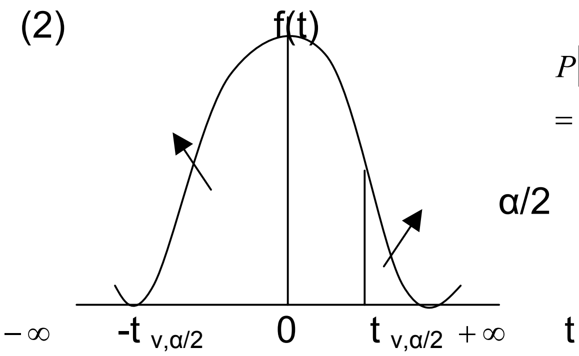


$$P\lfloor T \leq t_{v,\alpha} \rfloor = \\ = 1 - P[T \geq t_{v,\alpha}] = \\ = F(t_{v,\alpha})$$

Σχήμα 3: Η t - student Κατανομή

Επίσης πάλι από πίνακες μπορούμε να πάρουμε τις τιμές $-t_{v,\alpha/2}$ και $+t_{v,\alpha/2}$ για τις οποίες γνωρίζουμε την πιθανότητα:

$$P\lfloor -t_{v,\alpha/2} \leq T \leq t_{v,\alpha/2} \rfloor = 1 - a$$



$$P\lfloor -t_{v,\alpha/2} \leq T \leq t_{v,\alpha/2} \rfloor = \\ = 1 - \alpha$$

Σχήμα 4: Η t - student Κατανομή

Εφαρμογή της t – κατανομής (ενδεικτικές αναφορές)

- Η t – κατανομή εφαρμόζεται στην κατανομή των μέσων $\bar{x}$ δειγμάτων μικρού μεγέθους που προέρχονται από κανονικό πληθυσμό

Η τυχαία μεταβλητή που χρησιμοποιείται είναι η εξής:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$ (με n-1 βαθμούς ελευθερίας).

- Δειγματική κατανομή της διαφοράς δύο μέσων όρων που αντιστοιχούν σε μικρού μεγέθους δείγματα τα οποία προέρχονται αντίστοιχα από δύο κανονικούς πληθυσμούς

Η τυχαία μεταβλητή που χρησιμοποιείται είναι η εξής:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Όπου $s_d = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ το **τυπικό σφάλμα** της διαφοράς των δύο μέσων όρων των μικρών δειγμάτων.

2.4. Η X^2 – Κατανομή

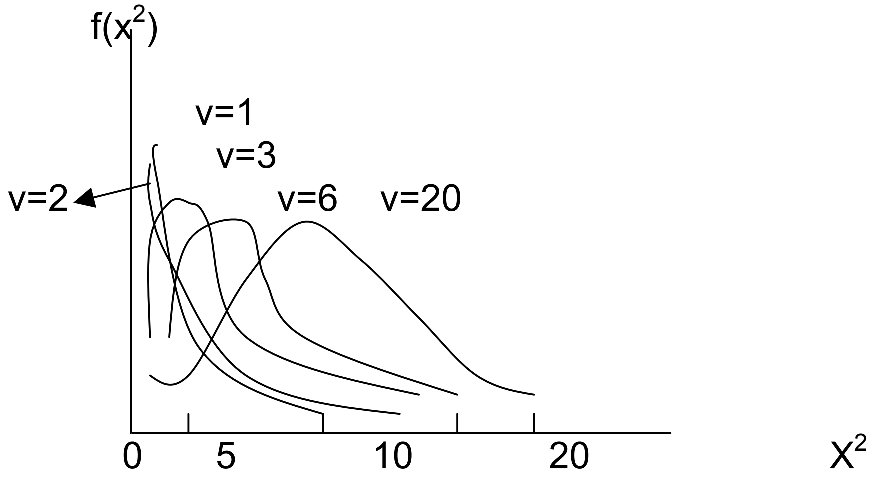
Περιγραφή της X^2 – Κατανομής

Θεωρούμε τις τυχαίες μεταβλητές $X_1, X_2, \ldots, X_v$ που είναι ανεξάρτητες μεταξύ τους και όπου η κάθε μια μεταβλητή έχει τυπική κανονική κατανομή N(0,1).

Δημιουργούμε την τυχαία μεταβλητή $X^2 = X_1^2 + X_2^2 + \ldots + X_v^2$.

Η τυχαία αυτή μεταβλητή ακολουθεί μια κατανομή που ονομάζεται **Χ - τετράγωνο κατανομή ή X^2 – κατανομή.**

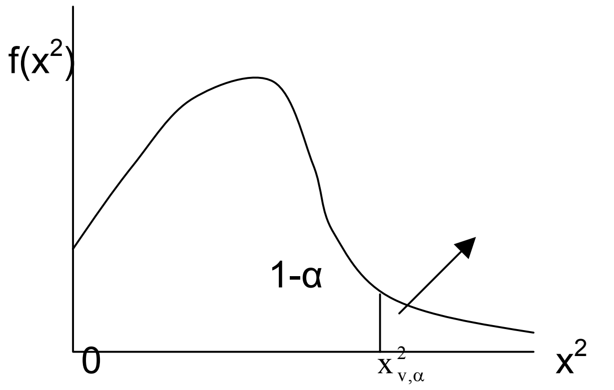
Η μορφή της κατανομής των πιθανοτήτων των τιμών της τυχαίας μεταβλητής $\mathbf{X^2}$ που ακολουθεί **X^2 - κατανομή** παρουσιάζεται στο σχήμα 6 για τους διάφορους βαθμούς ελευθερίας v.



Σχήμα 6: Η X^2 - Κατανομή

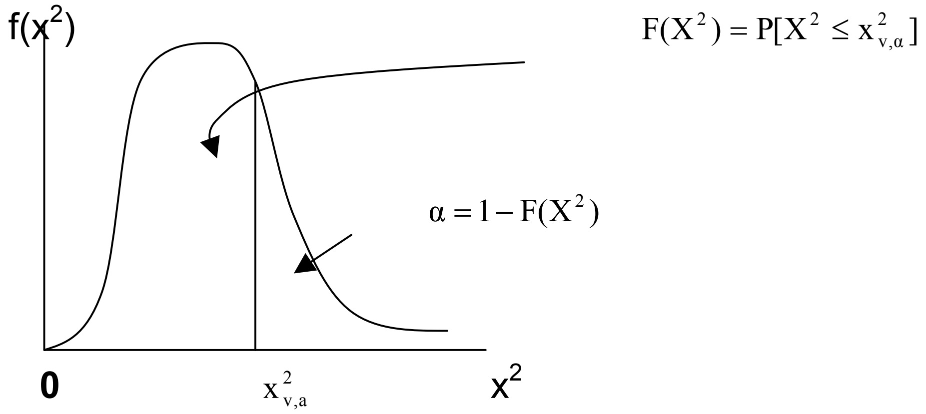
Όταν v > 30 η κατανομή της X^2 είναι σχεδόν συμμετρική και λέμε ότι τείνει προς την κανονική.

Δίνονται πίνακες όπου παρουσιάζονται οι τιμές $x^2_{v,\alpha}$ της X^2 για τις οποίες έχουμε την πιθανότητα $P[X^2 \geq x^2_{v,\alpha}] = \alpha,$ με α δοσμένη πιθανότητα και v - βαθμοί ελευθερίας γνωστά.



Σχήμα 8: Η x^2 κατανομή

Η αθροιστική συνάρτηση $F(x^2)$ εκφράζει το εμβαδόν του τμήματος αριστερά της $x^2_{v,a}$.



Σχήμα 7: Η X^2 κατανομή

Επίσης δίνονται από πίνακες οι τιμές $x^2_{v,\frac{\alpha}{2}}$, $x^2_{v,1-\frac{\alpha}{2}}$ για τις οποίες

με γνωστή πιθανότητα σφάλματος α έχουμε ότι:

$$P[x^2_{v,1-\frac{a}{2}} \leq X^2 \leq x^2_{v,\frac{a}{2}}] = 1\text{-}\alpha$$

Εφαρμογή της X^2 – κατανομής (ενδεικτικές αναφορές)

Η X^2 – κατανομή εφαρμόζεται στην:

- Σύγκριση διακυμάνσεως με γνωστό αριθμό
- Σύγκριση Συχνοτήτων
- Ομοιογένεια διωνυμικής σειράς

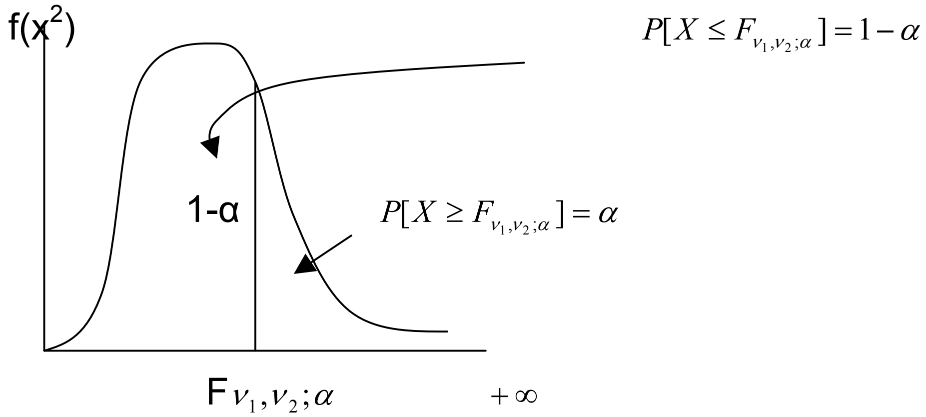
2.5. Η F - Κατανομή

Περιγραφή της F – Κατανομής

Έστω οι τυχαίες μεταβλητές X_1, X_2 που ακολουθούν κατανομές $X^2_{V_1}$, $X^2_{V_2}$, δημιουργούμε την τυχαία μεταβλητή **X =** $\frac{X_1/v_1}{X_2/v_2}$ η οποία ακολουθεί την F – Κατανομή με δύο βαθμούς ελευθερίας το v_1, που είναι ο βαθμός ελευθερίας του αριθμητή και το v_2, που είναι ο βαθμός ελευθερίας του παρανομαστή.

Υπάρχουν πίνακες που δίνουν τις κριτήριες τιμές F v_1,v_2 ; α , για τις διάφορες τιμές των α, v_1, v_2, έτσι ώστε αν η τυχαία μεταβλητή X ακολουθεί F κατανομή τότε:

$$P(X \geq F\ v_1, v_2; \alpha) = \alpha$$



Σχήμα 10: Η F κατανομή

Επίσης δίνονται από πίνακες οι τιμές F v_1, v_2;$\frac{a}{2}$ και F v_1, v_2;1-$\frac{a}{2}$ για τις οποίες έχουμε ότι αν η τυχαία μεταβλητή X που ακολουθεί F κατανομή τότε:

$$P\left(F_{v_1, v_2; 1-\frac{a}{2}} \le X \le F_{v_1, v_2; \frac{a}{2}}\right) = 1- \alpha$$

Εφαρμογή της F κατανομής (ενδεικτική αναφορά)

Εφαρμόζεται σε τυχαίες μεταβλητές που παράγονται από τον λόγο των διασπορών δύο κανονικών πληθυσμών.

Κατανομές στατιστικών δείγματος

3.1. Κατανομή μέσης τιμής δείγματος

Αν πάρουμε πολλά τυχαία δείγματα μεγέθους n από ένα πληθυσμό τότε προκύπτει η τυχαία μεταβλητή $\bar{X}$, από τις τιμές των μέσων τιμών των δειγμάτων. Όταν το δείγμα προέρχεται από **κανονικό πληθυσμό** N(μ, σ^2) ανεξάρτητα από το μέγεθός του έχουμε ότι η $\bar{X}$ ακολουθεί κανονική κατανομή N(μ, $\frac{\sigma^2}{n}$), το ίδιο ισχύει και στην περίπτωση που το δείγμα δεν προέρχεται από κανονικό πληθυσμό αλλά το μέγεθός του είναι μεγάλο **n**$\geq$**30.**

Τότε η $Z = \dfrac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ ή $Z = \dfrac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$

ακολουθεί **τυπική κανονική κατανομή**.

Όταν το μέγεθος του δείγματος είναι **μικρό**, η **διακύμανση του πληθυσμού άγνωστη** και ο πληθυσμός, από όπου προέρχεται το δείγμα, **κανονικός** τότε η τυχαία μεταβλητή:

$$t = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$$

ακολουθεί την **t-κατανομή με n-1 βαθμούς ελευθερίας.**

3.2. Κατανομή διαφοράς μέσων τιμών δύο δειγμάτων

Αν πάρουμε **δύο δείγματα μεγάλα** μεγέθους ν και κ με μέσες τιμές $\bar{x}$ και $\bar{y}$ και διακυμάνσεις s_1^2, s_2^2 και οι πληθυσμιακές διακυμάνσεις είναι σ_1^2 , σ_2^2 αντίστοιχα τότε έχουμε ότι η μεταβλητή:

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{\nu} + \frac{\sigma_2^2}{\kappa}}}$$ ακολουθεί N(0,1).

Εφόσον οι πληθυσμιακές διακυμάνσεις είναι **άγνωστες** αντικαθίστανται από τις δειγματικές διακυμάνσεις s_1^2, s_2^2 και έχουμε ότι η μεταβλητή:

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{\nu} + \frac{s_2^2}{\kappa}}}$$ ακολουθεί N(0,1).

Αν πάρουμε δύο δείγματα **μικρά ανεξάρτητα** μεγέθους ν και κ με μέσες τιμές $\bar{x}$ και $\bar{y}$ και διακυμάνσεις s_1^2, s_2^2 αντίστοιχα προερχόμενα από **κανονικούς πληθυσμούς** με **άγνωστες αλλά ίσες διακυμάνσεις** $\sigma_1^2 = \sigma_2^2 = \sigma^2$ τότε έχουμε ότι η μεταβλητή:

$$t = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{(\nu-1)s_1^2 + (\kappa-1)s_2^2}{\nu+\kappa-2}}\sqrt{\frac{1}{\nu} + \frac{1}{\kappa}}}$$

ακολουθεί **t-κατανομή με ν+κ-2 βαθμούς ελευθερίας**

Αν πάρουμε δύο δείγματα **μικρά ανεξάρτητα** μεγέθους ν και κ με μέσες τιμές $\bar{x}$ και $\bar{y}$ και διακυμάνσεις s_1^2, s_2^2 αντίστοιχα προερχόμενα από **κανονικούς πληθυσμούς** με **άγνωστες και διαφορετικές διακυμάνσεις** $\sigma_1^2 \neq \sigma_2^2$ τότε έχουμε ότι η μεταβλητή:

$$t = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{\nu} + \frac{s_2^2}{\kappa}}}$$

ακολουθεί **t – κατανομή** με **βαθμούς ελευθερίας λ = 2(ν-1)** όταν ν = κ

και $\lambda = \dfrac{\left(\dfrac{s_1^2}{\nu}+\dfrac{s_2^2}{\kappa}\right)^2}{\dfrac{\left(\dfrac{s_1^2}{\nu}\right)^2}{\nu-1}+\dfrac{\left(\dfrac{s_2^2}{\kappa}\right)^2}{\kappa-1}}$ (στρογγυλεμένο στον πλησιέστερο ακέραιο)

όταν $\nu \neq \kappa$

3.3. Κατανομή διακύμανσης δείγματος

Σύμφωνα με την θεωρία η τυχαία μεταβλητή:

$$X^2 = \frac{(n-1)s^2}{\sigma^2}$$

ακολουθεί την **X^2 κατανομή** με **n-1 βαθμούς ελευθερίας.**

Ο πληθυσμός από όπου πάρθηκε το δείγμα μεγέθους n θεωρείται **κανονικός** με σ^2 διασπορά και s^2 είναι η δειγματική διακύμανση.

3.4. Κατανομή λόγου διακυμάνσεων δύο δειγμάτων

Σύμφωνα με την θεωρία αν πάρουμε δύο δείγματα ανεξάρτητα μεγέθους n και m αντίστοιχα με s_1^2, s_2^2 τις δειγματικές διακυμάνσεις τους από δύο **κανονικούς πληθυσμούς** με σ_1^2, σ_2^2 τις πληθυσμιακές διακυμάνσεις, τότε ισχύει ότι η τυχαία μεταβλητή:

$$F = \frac{s_1^2}{s_2^2}\frac{\sigma_2^2}{\sigma_1^2}$$

ακολουθεί την F κατανομή με n-1 και m-1 βαθμούς ελευθερίας

3.5. Κατανομή ποσοστού (αναλογίας) δείγματος

Αν εκτιμήσουμε το ποσοστό (αναλογία) p = $\frac{x}{n}$ των στοιχείων στο **μεγάλο** δείγμα μεγέθους n που έχουν κάποιο συγκεκριμένο χαρακτηριστικό, τότε σύμφωνα με την θεωρία το δειγματικό ποσοστό κατανέμεται κανονικά:

Δηλαδή η δειγματική μεταβλητή P_{Δ}, που παράγεται από τις τιμές p, ακολουθεί την κανονική κατανομή:

$$N(P_{\pi\lambda\eta\theta}, \frac{p_{\pi\lambda\eta\theta}(1-P_{\pi\lambda\eta\theta})}{n})$$

Όπου $P_{\pi\lambda\eta\theta}$ το πραγματικό ποσοστό στον πληθυσμό.

Συνεπώς έχουμε ότι η μεταβλητή:

$$Z = \frac{P_{\Delta} - P_{\pi\lambda\eta\theta}}{\sqrt{\frac{P_{\pi\lambda\eta\theta}(1-P_{\pi\lambda\eta\theta})}{n}}} \quad \text{ακολουθεί την } \mathbf{N(0,1)}.$$

3.6. Κατανομή διαφοράς ποσοστών (αναλογιών) δύο δειγμάτων

Έστω $p_1 = \frac{x}{n}$ και $p_2 = \frac{y}{l}$ είναι οι δειγματικές αναλογίες των στοιχείων, που έχουν κάποιο συγκεκριμένο χαρακτηριστικό, δύο δειγμάτων μεγέθους n, l από δύο πληθυσμούς, όπου οι πραγματικές αναλογίες είναι αντίστοιχα $P_{1\,\pi\lambda\eta\theta}$, $P_{2\,\pi\lambda\eta\theta}$.

Τότε έχουμε ότι οι διαφορές των δειγματικών μεταβλητών $P_{1\,\Delta}$ - $P_{2\,\Delta}$ ακολουθούν κανονική κατανομή:

$$N(P_{1\,\pi\lambda\eta\theta} - P_{2\,\pi\lambda\eta\theta}, \frac{P_{1\pi\lambda\eta\theta}(1-P_{1\pi\lambda\eta\theta})}{n} + \frac{P_{2\pi\lambda\eta\theta}(1-P_{2\pi\lambda\eta\theta})}{l})$$

Κεφάλαιο 4ο

Εκτιμητική

4.1. Σημειακή εκτίμηση μιας άγνωστης παραμέτρου ενός πληθυσμού ως προς μια μεταβλητή.

Όπως έχει αναφερθεί μέχρι τώρα για να μελετήσουμε ένα μεγάλο πληθυσμό ως προς κάποια μεταβλητή X καταφεύγουμε στην δειγματοληψία.

Λαμβάνουμε ένα δείγμα n - στοιχείων του πληθυσμού και μελετάμε αυτό ως προς την μεταβλητή X. Τα συμπεράσματα που προκύπτουν χαρακτηρίζουν, μέσω της επαγωγικής σκέψης, και το πληθυσμό.

Αυτό άλλωστε αποτελεί και τη βασική σκέψη της **Επαγωγικής Στατιστικής** της οποίας ένα από τα σπουδαιότερα κεφάλαια είναι αυτό της **Εκτιμητικής.**

Σε κάθε δείγμα προσδιορίζονται οι στατιστικές παράμετροι αυτού, όπως είναι η μέση τιμή $\bar{x}$, η διακύμανση s^2, η τυπική απόκλιση s κλπ.

Η τιμή μιας από αυτές τις παραμέτρους ονομάζεται **σημειακή εκτιμήτρια** της αντίστοιχης παραμέτρου του πληθυσμού.

Γενικά αν ονομάσουμε $\widehat{\Theta}$ την σημειακή εκτιμήτρια της αντίστοιχης παραμέτρου Θ του πληθυσμού αυτή θα πρέπει να ικανοποιεί κάποιες βασικές ιδιότητες.

Καταρχήν επειδή μπορούμε να πάρουμε πολλά δείγματα ίσου μεγέθους από τον πληθυσμό θα έχουμε και τις αντίστοιχες σημειακές εκτιμήτριες των δειγμάτων αυτών για την άγνωστη παράμετρο του πληθυσμού.

Δηλαδή αν πάρουμε ν δείγματα ίσου μεγέθους από τον πληθυσμό θα έχουμε και τις αντίστοιχες $\widehat{\Theta}_\nu$ εκτιμήτριες της Θ.

Γίνεται λοιπόν φανερό ότι η εκτιμήτρια $\widehat{\Theta}$ είναι και αυτή τυχαία μεταβλητή. Η τιμή της τυχαίας μεταβλητής $\widehat{\Theta}$ σε ένα συγκεκριμένο i-δείγμα $\widehat{\Theta}_i$ i =1,2,3,4,…,ν λέγεται σημειακή εκτίμηση της παραμέτρου Θ.

Οι βασικές ιδιότητες που πρέπει να πληρούνται από την σημειακή εκτιμήτρια $\widehat{\Theta}$ είναι:

i. **Η Αμεροληψία:** θα πρέπει δηλαδή η μέση τιμή της τυχαίας μεταβλητής $\widehat{\Theta}$ να είναι η άγνωστη παράμετρος Θ :

$$E(\widehat{\Theta}) = \Theta.$$

Αν $E(\widehat{\Theta}) \neq \Theta$ τότε η εκτιμήτρια είναι **μεροληπτική** και έχουμε το **σφάλμα μεροληψίας ή σφάλμα εκτίμησης:** $E(\widehat{\Theta}) - \Theta$.

ii. Η Αποτελεσματικότητα:

Θα πρέπει η διακύμανση της αμερόληπτης εκτιμήτριας $\widehat{\Theta}$, η Var$(\widehat{\Theta})$, και είναι μικρότερη ή ίση από την διακύμανση οποιαδήποτε άλλης αμερόληπτης εκτιμήτριας $(\widehat{\Theta})^*$:

$$Var(\widehat{\Theta}) \leq Var(\widehat{\Theta}^*).$$

iii. Η συνέπεια:

Μια σημειακή εκτιμήτρια $\widehat{\Theta}$ είναι συνεπής όταν το σφάλμα μεροληψίας και η διακύμανσή της τείνουν στο μηδέν καθώς το μέγεθος του δείγματος n τείνει στο άπειρο:

$$\lim_{n \to +\infty} E(\widehat{\Theta}) = \Theta \qquad \kappa\alpha\iota \qquad \lim_{n \to +\infty} Var(\widehat{\Theta}) = 0.$$

4.2. Σημειακές εκτιμήτριες της μέσης τιμής και της διακύμανσης ενός πληθυσμού ως προς μια μεταβλητή.

i. Σημειακή εκτίμηση της μέσης τιμής ενός πληθυσμού ως προς μια μεταβλητή.

Έστω $x_1, x_2, \ldots, x_n$ οι παρατηρήσεις της μεταβλητής X και μ η άγνωστη μέση τιμή του πληθυσμού.

Έχουμε προσδιορίσει την δειγματική μέση τιμή $\bar{x}$, την σημειακή εκτιμήτρια της μ του πληθυσμού, από τον τύπο:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{v} xi$$

Αν πάρουμε πολλά δείγματα ίσου μεγέθους μπορούμε να ορίσουμε την τυχαία μεταβλητή $\bar{X}$:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{v} Xi$$

όπου X_1,X_2,…..X_n τυχαίες μεταβλητές με κατανομή την ίδια με αυτή της X.

Η μέση τιμή $\mathrm{E}(\bar{\mathrm{X}})$ είναι μια αμερόληπτη σημειακή εκτιμήτρια της πραγματικής μέσης τιμής μ του πληθυσμού και ισχύει:

$E(\bar{X}) = \mu, Var(\bar{X}) = \frac{\sigma^2}{n}$ και τυπική απόκλιση $\frac{\sigma}{\sqrt{n}}$.

Στις συμμετρικές κατανομές ως σημειακή εκτιμήτρια μπορεί να ληφθεί και η διάμεσος M, διότι έχουμε:

$$\mathrm{E(M)} = \mu.$$

ii. Σημειακή εκτίμηση της διακύμανσης ενός πληθυσμού ως προς μια μεταβλητή.

Ομοίως η άγνωστη διακύμανση ενός πληθυσμού μεγέθους N

$$\sigma^2 = \frac{\sum_{i=1}^{v} (xi - \bar{x})^2}{N}$$

εκτιμάται από την διακύμανση ενός δείγματος x_1, x_2,……..x_n τιμών ως προς την μεταβλητή X του πληθυσμού.

Η δειγματική διακύμανση συμβολίζεται με s^2 και είναι σύμφωνα με όσα αναφέρθησαν:

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

Η δειγματική αυτή διακύμανση είναι τυχαία μεταβλητή S^2, επειδή μπορούμε να λάβουμε πολλά δείγματα ίσου μεγέθους του πληθυσμού και να ορίσουμε έτσι τις αντίστοιχες διακυμάνσεις.

$$S^2 = \frac{\sum_{i=1}^{\nu} (Xi - \overline{X})^2}{n-1}$$

όπου Xi οι προαναφερόμενες τυχαίες μεταβλητές και $\overline{X}$ η μέση τιμή αυτών.

Αποδεικνύεται ότι $E(S^2) = \sigma^2$, δηλαδή η διακύμανση του δείγματος s^2 είναι μια τιμή της τυχαίας μεταβλητής S^2 που έχει μέση τιμή $E(S^2)$ τη διακύμανση σ^2 του πληθυσμού.

4.3. Διάστημα εμπιστοσύνης της παραμέτρου Θ ενός πληθυσμού.

Επειδή η δειγματική σημειακή εκτιμήτρια $\hat{\Theta}$ της πραγματικής τιμής Θ του πληθυσμού δεν μας δίνει πληροφορίες περί του βαθμού ακρίβειάς της, δηλαδή πόσο κοντά στην τιμή Θ βρίσκεται, καταφεύγουμε στο να υπολογίσουμε ένα διάστημα που με κάποια **προκαθορισμένη πιθανότητα** θα περιέχει την άγνωστη τιμή του πληθυσμού.

Το διάστημα αυτό το ονομάζουμε **«διάστημα εμπιστοσύνης»** της τιμής Θ.

Το διάστημα αυτό (β, γ) είναι το διάστημα στο οποίο εκτιμούμε ότι θα βρίσκεται η τιμή Θ του πληθυσμού με ορισμένη **πιθανότητα ή επίπεδο εμπιστοσύνης 1- α.**

$$P(\beta \leq \Theta \leq \gamma) = 1 - \alpha \qquad \mu\varepsilon \qquad 0 \leq \alpha \leq 1.$$

Τα β, γ ονομάζονται **όρια εμπιστοσύνης** και η πιθανότητα α καλείται **επίπεδο σημαντικότητας.**

4.4. Εκτίμηση διαστήματος εμπιστοσύνης της μέσης τιμής μ πληθυσμού.

4.4.1. Μεγάλο ή μικρό δείγμα. Κανονικός πληθυσμός. Γνωστή η διασπορά σ^2 του πληθυσμού.

Είναι δεκτό ότι εφόσον ο πληθυσμός που αντιστοιχεί στην τυχαία μεταβλητή X είναι κανονικής κατανομής τότε και η δειγματική εκτιμήτρια $\bar{x}$ είναι κανονικής κατανομής.

Δηλαδή ισχύει ότι η τυχαία μεταβλητή $\bar{X}$, που παίρνει τιμές τις $\bar{x}$, ακολουθεί την κανονική κατανομή $N(\mu, \frac{\sigma}{\sqrt{n}})$.

Επομένως η τυχαία μεταβλητή Z = $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ με τιμές z = $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ ακολουθεί μια τυπική κανονική κατανομή.

Η πιθανότητα η τιμή $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ να βρίσκεται μέσα σ' ένα δοσμένο διάστημα για παράδειγμα το ± 1,96 δίδεται από τη σχέση:

$$P\left[-1{,}96 \le \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \le 1{,}96\right] = \phi(1{,}96) - \phi(-1{,}96) = 0{,}95.$$

Από την σχέση αυτή αν είναι γνωστή η σ μπορούμε να πούμε ότι:

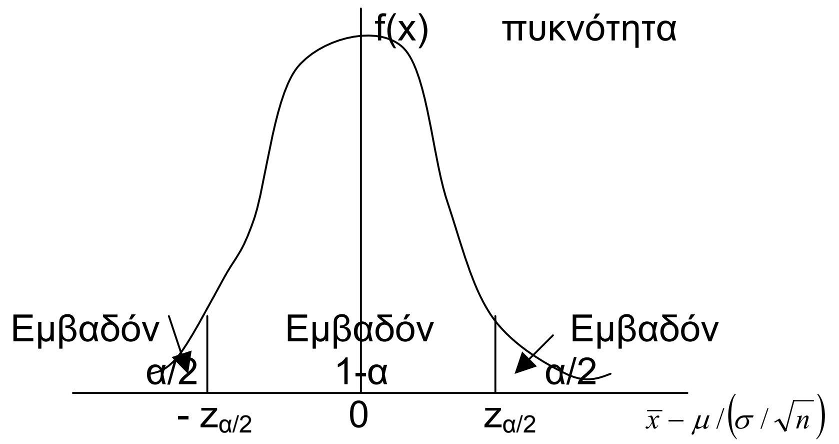
$$\bar{x} - 1{,}96\sigma/\sqrt{n} \le \mu \le \bar{x} + 1{,}96\sigma/\sqrt{n}$$

δηλαδή η μέση τιμή μ του πληθυσμού με πιθανότητα 95% βρίσκεται σε απόσταση $\pm 1{,}96\sigma/\sqrt{n}$ από την δειγματική μέση τιμή $\bar{x}$.

Παρατήρηση:

Μια άλλη ερμηνεία της σχέσεως αυτής είναι ότι «αν χρησιμοποιηθούν διάφορα δείγματα μεγέθους n για τον προσδιορισμό **«διαστημάτων πιθανότητας 95%»** τότε κατά μέσο όρο το 95% από τα διαστήματα αυτά θα περιέχουν την αληθινή τιμή μ.

Σύμφωνα με τα ανωτέρω αν συμβολίσουμε με 1-α το προκαθορισμένο **«επίπεδο εμπιστοσύνης»** και με $\mp z_{\alpha/2}$ τις τιμές της τυπικής κανονικής μεταβλητής με αντίστοιχες τιμές της Αθροιστικής Συνάρτησης κατανομής α/2 και 1-α/2, όπως φαίνεται στο παρακάτω σχήμα:



Σχήμα11. Τυπικής Κανονικής κατανομής

Τότε έχουμε την πιθανότητα:

$$P\left[-z_{\alpha/2} \leq \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right] = 1-\alpha.$$

ή $$P\left[\bar{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = 1-\alpha.$$

Μπορούμε να πούμε ότι υπάρχει εμπιστοσύνη επιπέδου 1-α ότι το διάστημα που εκτιμήθηκε περιέχει την άγνωστη τιμή της μ. Το διάστημα αυτό ονομάζεται **«διάστημα εμπιστοσύνης της μ σε επίπεδο 1-α»** και δίδεται από τη σχέση:

(I) $$<\mu>_{1-\alpha} = \left[(\bar{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}),(\bar{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}})\right], \quad \mu\varepsilon \quad \sigma = \gamma\nu\omega\sigma\tau\acute{\eta}.$$

Παρατήρηση:

Η ανωτέρω σχέση ισχύει για κανονικές τυχαίες μεταβλητές που γνωρίζουμε τη σ.

Για μη κανονικούς πληθυσμούς ισχύει προσεγγιστικά και ο βαθμός προσέγγισης αυξάνει όσο μεγαλώνει το μέγεθος του δείγματος.

Σχόλιο

Η διαδικασία προσδιορισμού διαστημάτων εμπιστοσύνης της μ, όταν η σ είναι γνωστή ακολουθεί τα εξής βήματα:

1° : Επιλέγουμε το επίπεδο εμπιστοσύνης 1-α.

2° : Υπολογίζουμε την τιμή $z_{\alpha/2}$ από τον πίνακα της τυπικής κανονικής κατανομής.

$$z_{\alpha/2} = \phi^{-1}\left(1-\frac{\alpha}{2}\right).$$

3° : Χρησιμοποιούμε τη σχέση (Ι) θέτοντας στη θέση του $\bar{x}$ τη δειγματική μέση τιμή των n - παρατηρήσεων.

Παραδείγματα.

1. Η ημερήσια συγκέντρωση διαλυμένου οξυγόνου σ' ένα σταθμό μέτρησης ενός ποταμού έχει μετρηθεί για 30 ημέρες. Από προηγούμενες εμπειρίες είναι γνωστό ότι η διασπορά της ημερήσιας συγκέντρωσης είναι 4,2 $(mg/l)^2$. Για τις 30 μετρήσεις η μέση δειγματική τιμή $\bar{x} = 2{,}52 mg/l.$

Να υπολογισθεί το διάστημα εμπιστοσύνης επιπέδου 99% για τη μέση ημερήσια συγκέντρωση διαλυμένου οξυγόνου. Θεωρούμε ότι ο πληθυσμός των μετρήσεων της ημερήσιας συγκέντρωσης διαλυμένου οξυγόνου κατανέμεται κανονικά.

$$1-\alpha = 0{,}99 \rightarrow \alpha/2 = 0{,}005 \rightarrow z_{0,005} = \phi^{-1}(0{,}995) = 2{,}58$$

$$\sigma/\sqrt{n}\, z_{\alpha/2} = \frac{\sqrt{4{,}2}}{\sqrt{30}} \cdot 2{,}58 = 0{,}965.$$

$$< \mu >_{0.99} = (\overline{x} - 0{,}965, \overline{x} + 0{,}965) = (1{,}56\ ,\ 3{,}49)\, mg/l.$$

Αυτό είναι το διάστημα εμπιστοσύνης της μ σε επίπεδο 99%.

2. Σε ένα δείγμα 10 κουτιών παστεριωμένου γάλακτος που παράγει μια βιομηχανία γάλακτος το μέσο βάρος των κουτιών είναι 250 γραμμάρια. Από προηγούμενες μετρήσεις είναι γνωστή η διακύμανση που παρατηρείται και ίση, με 60 γραμμάρια^2. Να προσδιοριστεί το διάστημα εμπιστοσύνης στο οποίο θα βρίσκεται το μέσο βάρος του συνόλου των κουτιών γάλακτος που παράγονται με πιθανότητα 99,60%. Θεωρούμε ότι ο πληθυσμός των βαρών των κουτιών κατανέμεται κανονικά.

Έχουμε $1 - \alpha = 0{,}996 \rightarrow \alpha/2 = 0{,}002.$

Άρα $z_{\alpha/2} = z_{0{,}002} = \phi^{-1}\left(1 - \frac{\alpha}{2}\right) = \phi^{-1}(0{,}998) = 2{,}88.$

Επίσης $\sigma/\sqrt{n} \cdot z_{\alpha/2} = \frac{\sqrt{60}}{\sqrt{10}} \cdot 2{,}88 = 7{,}05.$

Το διάστημα εμπιστοσύνης:

$$< \mu >_{1-\alpha} = \left(\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \Rightarrow$$

$$< \mu >_{0.996} = (\overline{x} - 7{,}05, \overline{x} + 7{,}05) = (250 - 7{,}05, 250 + 7{,}05) \Rightarrow$$

$< \mu >_{0.996} = (242{,}95\ \ ,\ \ 257{,}05)$ γραμμάρια.

Το διάστημα εμπιστοσύνης της μ με πιθανότητα ή επίπεδο εμπιστοσύνης 99,6%.

Αν ο πληθυσμός της ημερήσιας συγκέντρωσης είναι κανονικός τότε το παραπάνω διάστημα είναι ακριβές αν όχι τότε το δεχόμαστε προσεγγιστικά.

4.4.2. Δείγμα μεγάλο. Μη κανονικός πληθυσμός. Η Διασπορά του πληθυσμού γνωστή ή άγνωστη.

Όταν το δείγμα είναι μεγάλο σύμφωνα με το κεντρικό οριακό θεώρημα η τυχαία μεταβλητή $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ με τιμές $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ ακολουθεί μια τυπική κανονική κατανομή.

Συνεπώς το **«διάστημα εμπιστοσύνης της μ σε επίπεδο 1-α»** δίδεται επίσης από τη σχέση:

$$\text{(I)}\quad <\mu>_{1-\alpha}=\left[(\bar{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}),(\bar{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}})\right],$$

εφόσον η διακύμανση σ^2 είναι άγνωστη αντικαθίσταται από την s^2 οπότε καταλήγουμε στη σχέση:

$$<\mu>_{1-\alpha}=\left[(\bar{x}-z_{\alpha/2}\frac{s}{\sqrt{n}}),(\bar{x}+z_{\alpha/2}\frac{s}{\sqrt{n}})\right]$$

Επομένως η διαδικασία επιλογής διαστήματος εμπιστοσύνης για την εκτίμηση του μέσου ενός μη κανονικού πληθυσμού του οποίου το δείγμα είναι μεγάλο είναι η εξής:

- Επιλέγεται το διάστημα εμπιστοσύνης 1-α, με $\alpha \in (0,1)$.
- Υπολογίζεται το α/2-ποσοστιαίο σημείο της κανονικής κατανομής το $z_{a/2}$
- Υπολογίζεται ο δειγματικός μέσος $\bar{x}$ καθώς και μία εκτίμηση της διακύμανσης $\hat{\sigma}$ του σ ως εξής:

$$\hat{\sigma} = \begin{cases} \sigma, \alpha\nu\ \sigma\ \gamma\nu\omega\sigma\tau\acute{o} \\ s, \delta\iota\alpha\varphi o\rho\varepsilon\tau\iota\kappa\acute{\alpha} \end{cases}$$

- Υπολογίζεται το $\hat{\sigma}\, z_{a/2}/\sqrt{n}$.
- Έτσι λαμβάνεται το ακόλουθο αριθμητικό διάστημα εμπιστοσύνης:

$$\left[\bar{x} - \frac{z_{a/2}\hat{\sigma}}{\sqrt{n}}, \bar{x} + \frac{z_{a/2}\hat{\sigma}}{\sqrt{n}}\right]$$

Παράδειγμα:

Η ημερήσια συγκέντρωση διαλυμένου οξυγόνου σ' ένα σταθμό μέτρησης ενός ποταμού έχει μετρηθεί για 39 ημέρες και είχαμε $\bar{x}$=2,52 mg/l και s^2=4,2 $(mg/l)^2$ (άγνωστη η σ^2). Να υπολογισθεί το διάστημα εμπιστοσύνης επιπέδου 99% για τη μέση ημερήσια συγκέντρωση διαλυμένου οξυγόνου.

Το διάστημα εμπιστοσύνης για τη μέση ημερήσια συγκέντρωση διαλυμένου οξυγόνου είναι: 1-α = 0,99→ α/2 = 0,005.

Άρα $z_{\frac{a}{2}}$ = 2,58 (παράρτημα).

Οπότε:

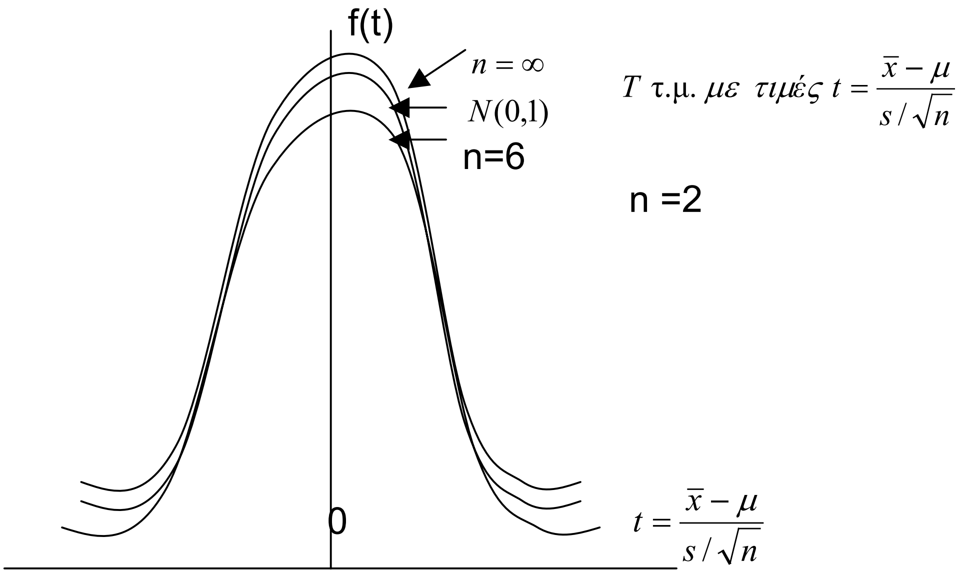
$$<\mu>_{0.99} = \left(2{,}52 - 2{,}58\frac{\sqrt{4{,}2}}{\sqrt{39}}, 2{,}52 + 2{,}58\frac{\sqrt{4{,}2}}{\sqrt{39}}\right) = (1{,}67,\ 3{,}37)\ mg/l.$$

4.4.3. Δείγμα μικρό. Κανονικός πληθυσμός. Η διασπορά του πληθυσμού σ^2 άγνωστη.

Για n μικρό πολύ π.χ. n < 10 τα διαστήματα εμπιστοσύνης της μ από την (I) με s^2 στη θέση του σ^2 είναι αρκετά ανακριβή (όταν δεν είναι γνωστή η διασπορά σ).

Όταν η κατανομή του πληθυσμού της X είναι κανονική τότε και αν ακόμα ή σ^2 δεν είναι γνωστή μπορούν να προσδιοριστούν ακριβή διαστήματα εμπιστοσύνης για τη μ.

Έχει αποδειχθεί ότι η τυχαία μεταβλητή $T\ \mu\varepsilon\ \tau\iota\mu\acute{\varepsilon}\varsigma\ t = \dfrac{\bar{x}-\mu}{s/\sqrt{n}}$ έχει κατανομή t με n-1 βαθμούς ελευθερίας και με συνάρτηση πυκνότητας f(t).



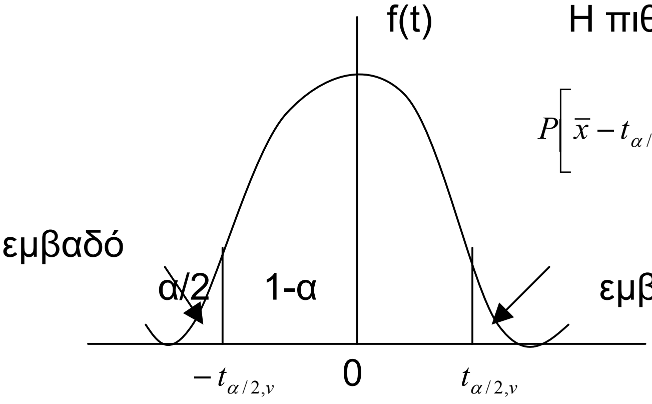
Σχήμα11. t - κατανομής

Η κατανομή t είναι συμμετρική ως προς το μηδέν και έχει σχήμα παρόμοιο της κανονικής κατανομής, όταν το n μεγαλώνει (v = n -1 = βαθμοί ελευθερίας) τότε η κατανομή t προσεγγίζει την κανονική κατανομή.

Έχουμε λοιπόν:

$$P\left[-t_{\alpha/2,v} \le \frac{\bar{x}-\mu}{s/\sqrt{n}} \le t_{\alpha/2,v}\right] = 1-\alpha.$$

Το $t_{\alpha/2,v}$ είναι η τιμή της τυχαίας μεταβλητής T που αντιστοιχεί σε τιμή της Αθροιστικής συνάρτησης κατανομής ίση με 1-α/2, για v = n-1 βαθμούς ελευθερίας. Οι τιμές του $t_{\alpha/2,v}$ δίδονται από πίνακα.



Η πιθανότητα

$$P\left[\bar{x} - t_{\alpha/2,v}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{\alpha/2,v}\frac{s}{\sqrt{n}}\right] = 1-\alpha.$$

Σχήμα12. t - κατανομής

Έτσι οδηγούμαστε στη σχέση

$$<\mu>_{1-\alpha} = \left[(\bar{x} - t_{\alpha/2,v}\frac{s}{\sqrt{n}}),(\bar{x} + t_{\alpha/2,v}\frac{s}{\sqrt{n}})\right]$$

όπου

$\bar{x}$: δειγματική μέση τιμή

s: δειγματική τυπική απόκλιση. Αυτό είναι το διάστημα εμπιστοσύνης της μ, με άγνωστη τη σ, σε επίπεδο 1-α.

Παρατήρηση: Όταν ο αριθμός των βαθμών ελευθερίας είναι πολύ μεγάλος τότε $t_{\alpha/2, v} \to z_{\alpha/2}$ διότι η κατανομή t τείνει τότε στην τυπική κανονική κατανομή.

Για την επιλογή επομένως ενός αριθμητικού διαστήματος εμπιστοσύνης του μέσου μ ενός κανονικού πληθυσμού με άγνωστη διασπορά και που το δείγμα είναι μικρό, ακολουθείται η εξής διαδικασία:

- Επιλέγεται το διάστημα εμπιστοσύνης 1-α, με α $\in (0,1)$.

- Υπολογίζεται το α/2-ποσοστιαίο σημείο της t_{n-1} κατανομής το $t_{n-1;a/2}$

- Υπολογίζεται ο δειγματικός μέσος $\bar{x}$

- Υπολογίζεται το $\frac{st_{n-1;a/2}}{\sqrt{n}}$

- Έτσι λαμβάνεται το ακόλουθο διάστημα εμπιστοσύνης:

$$\left[\bar{x}-\frac{t_{n-1;a/2}S}{\sqrt{n}},\ \bar{x}+\frac{t_{n-1;a/2}S}{\sqrt{n}}\right]$$

Παράδειγμα:

Σε ένα δείγμα 20 κουτιών παστεριωμένου γάλακτος που παράγει μια βιομηχανία γάλακτος το μέσο βάρος των κουτιών είναι 250 γραμμάρια και η δειγματική διακύμανση $s^2 = 56,25$. Να προσδιοριστεί το διάστημα εμπιστοσύνης στο οποίο θα βρίσκεται το μέσο βάρος του συνόλου των κουτιών γάλακτος που παράγονται με πιθανότητα 99,80%.

Γνωρίζουμε ότι:

$$<\mu>_{1-\alpha}=\left[(\bar{x}-t_{\alpha/2,v}\frac{s}{\sqrt{n}},\bar{x}+t_{\alpha/2,v}\frac{s}{\sqrt{n}})\right]$$

$$1-\alpha=0{,}998\rightarrow\frac{\alpha}{2}=0{,}001$$
$$v=20-1=19$$

$$t_{\alpha/2,v}=t_{0{,}001\,,19}=3{,}883.$$

$$t_{\alpha/2,v}\frac{s}{\sqrt{n}}=3{,}883\frac{7{,}5}{\sqrt{20}}=6{,}512.$$

Άρα $<\mu>_{0,998} = (250 - 6{,}512 \quad , \quad 250 + 6{,}512)\text{kgr} \Rightarrow$
$<\mu>_{0,998} = (243{,}488 \quad , \quad 256{,}512)\,\text{kgr}$, το διάστημα εμπιστοσύνης.

Μονόπλευρα όρια εμπιστοσύνης για τη μέση τιμή μ.

Η διακύμανση – διασπορά σ^2 γνωστή, Κανονικός πληθυσμός:

Χρησιμοποιείται η τ.μ. Ζ με τιμές z = $\frac{\overline{x}-\mu}{\sigma\sqrt{n}}$ που ακολουθεί τυπική κανονική κατανομή.

Ορίζουμε το **κατώτατο όριο εμπιστοσύνης της μ σε επίπεδο 1-α.**

$$\left(\overline{x} - z_\alpha \frac{\sigma}{\sqrt{n}}\right), z_\alpha = \phi^{-1}(1-\alpha).$$

* Αυτό προκύπτει από το ότι πρέπει:

$P = \left[\frac{\overline{x}-\mu}{\sigma/\sqrt{n}} \le z_\alpha\right] = 1-\alpha.$, άρα $P\left[\mu \ge \overline{x} - z_\alpha \frac{\sigma}{\sqrt{n}}\right] = 1-\alpha.$

με $z_\alpha = \phi^{-1}(1-\alpha).$

Αυτό δηλώνει ότι η μέση τιμή του πληθυσμού θα είναι μεγαλύτερη από το όριο αυτό με πιθανότητα 1-α.

Ομοίως υπολογίζεται και το **ανώτατο όριο εμπιστοσύνης της μ, επιπέδου 1- α.**

$$\left(\overline{x} + z_\alpha \frac{\sigma}{\sqrt{v}}\right) \quad , \quad z_\alpha = \phi^{-1}(1-\alpha).$$

Σε περίπτωση **μεγάλου δείγματος και άγνωστης πληθυσμιακής διασποράς** ορίζονται αντίστοιχα το **κατώτατο και ανώτατο όριο εμπιστοσύνης της μ, επιπέδου 1-α** όπως τα παραπάνω, εφόσον αντικατασταθεί η **σ** με την **s**.

Η διακύμανση σ^2 άγνωστη (μικρό δείγμα):

Χρησιμοποιείται η t -κατανομή για τον προσδιορισμό **ανώτατου και κατώτατου ορίου εμπιστοσύνης της μ.**

Κατώτατο όριο εμπιστοσύνης της μ, επιπέδου 1-α.

$$\bar{x} - t_{\alpha,v}\frac{s}{\sqrt{n}} \quad , \quad t_{\alpha,v} \text{ από πίνακα.}$$

Ανώτατο όριο εμπιστοσύνης της μ, επιπέδου 1-α.

$$\bar{x} + t_{\alpha,v}\frac{s}{\sqrt{n}} \quad , \quad t_{\alpha,v} \text{ από πίνακα}$$

Παράδειγμα.

Τα εργαστηριακά αποτελέσματα 100 δοκιμίων χάλυβα A36 που διαλέχτηκαν τυχαία δείχνουν για την τάση ροής μια μέση τιμή $\bar{x} = 2200\mathrm{kp/cm^2}$ και μια τυπική απόκλιση $220\mathrm{kp/cm^2}$.

Να προσδιορισθεί το κατώτατο όριο εμπιστοσύνης της μέσης τιμής μ της τάσης ροής του χάλυβα αυτού σε επίπεδο 95%.

Λόγω μεγάλου μεγέθους n = 100 σ ~ s = 220.

$$1-\alpha = 0{,}95 \rightarrow \alpha = 0{,}05 \rightarrow z_{0,05} = \phi^{-1}(1-0{,}05) = \phi^{-1}(0{,}95) = 1{,}65.$$

οπότε το κατώτατο όριο εμπιστοσύνης της μέσης τιμής μ της τάσης ροής του χάλυβα αυτού σε επίπεδο 95% είναι:

$$\bar{x} - z_\alpha\frac{\sigma}{\sqrt{n}} = 2200 - 1{,}65\frac{220}{\sqrt{100}} = 2164\,kgr.$$

Γενική παρατήρηση:

Οι αποφάσεις ή εκτιμήσεις στη Στατιστική έχουν στοχαστικό χαρακτήρα και δεν αποτελούν αποδείξεις.

Έτσι εκτιμούμε ότι ο μέσος ενός πληθυσμού $< \mu >_{1-\alpha} = (\alpha_1, \beta_1)$. Δηλαδή ότι ο μέσος μ βρίσκεται στο διάστημα (α_1, β_1) δίνοντας συγχρόνως την πιθανότητα σφάλματος αυτής της εκτίμησης.

Έτσι λέμε ότι σε επίπεδο εμπιστοσύνης 95% = 1-α η μέση τιμή μ του πληθυσμού ανήκει στο διάστημα: (1,36 , 2,18).

4.5. Εκτίμηση διαστήματος εμπιστοσύνης της διαφοράς των μέσων τιμών δύο πληθυσμών.

4.5.1. Δείγματα ανεξάρτητα μεγάλα, πληθυσμιακές διασπορές σ_1^2 , σ_2^2 γνωστές ή άγνωστες

Αν πάρουμε **δύο δείγματα ανεξάρτητα μεγάλα** μεγέθους ν και κ με μέσες τιμές $\bar{x}$ και $\bar{y}$, διακυμάνσεις s_1^2, s_2^2 και **οι πληθυσμιακές διασπορές είναι γνωστές** σ_1^2, σ_2^2 αντίστοιχα, τότε έχουμε ότι η μεταβλητή:

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{\nu} + \frac{\sigma_2^2}{\kappa}}} \quad \text{ακολουθεί N(0,1).}$$

Τότε ένα **100(1-α)% διάστημα εμπιστοσύνης** για την διαφορά **μ_1 - μ_2** των δύο πληθυσμών όταν οι **διασπορές τους είναι γνωστές** είναι:

$$\bar{x} - \bar{y} - Z_{\frac{a}{2}} \sqrt{\frac{\sigma_1^2}{\nu} + \frac{\sigma_2^2}{\kappa}} \leq \mu_1 - \mu_2 \leq \bar{x} - \bar{y} + Z_{\frac{a}{2}} \sqrt{\frac{\sigma_1^2}{\nu} + \frac{\sigma_2^2}{\kappa}}$$

Εάν **οι πληθυσμιακές διασπορές είναι άγνωστες** αντικαθίστανται από τις δειγματικές διακυμάνσεις s_1^2, s_2^2 και έχουμε ότι η μεταβλητή:

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{\nu} + \frac{s_2^2}{\kappa}}} \quad \text{ακολουθεί N(0,1).}$$

Τότε ένα **100(1-α)% διάστημα εμπιστοσύνης** για την διαφορά **μ_1 - μ_2** των δύο πληθυσμών όταν **οι διασπορές τους είναι άγνωστες** είναι:

$$\bar{x} - \bar{y} - Z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{\nu}+\frac{s_2^2}{\kappa}} \leq \mu_1 - \mu_2 \leq \bar{x} - \bar{y} + Z_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{\nu}+\frac{s_2^2}{\kappa}}$$

4.5.2. Δείγματα **μικρά ανεξάρτητα**, **κανονικοί πληθυσμοί** με **άγνωστες αλλά ίσες διασπορές $\sigma_1^2 = \sigma_2^2 = \sigma^2$**

Αν πάρουμε δύο δείγματα **μικρά ανεξάρτητα** μεγέθους ν και κ με μέσες τιμές $\bar{x}$ και $\bar{y}$ και διακυμάνσεις s_1^2, s_2^2 αντίστοιχα προερχόμενα από **κανονικούς πληθυσμούς** με **άγνωστες αλλά ίσες διασπορές $\sigma_1^2 = \sigma_2^2 = \sigma^2$** έχουμε ότι η μεταβλητή:

$$t = \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{(\nu-1)s_1^2+(\kappa-1)s_2^2}{\nu+\kappa-2}}\sqrt{\frac{1}{\nu}+\frac{1}{\kappa}}}$$

ακολουθεί **t - κατανομή με ν+κ-2 βαθμούς ελευθερίας**

Τότε ένα **100(1-α)% διάστημα εμπιστοσύνης** για την διαφορά **μ_1 - μ_2** των δύο πληθυσμών όταν **οι διασπορές τους είναι άγνωστες αλλά ίσες** είναι:

$$\bar{x} - \bar{y} - t_{\nu+\kappa-2;\frac{\alpha}{2}}\, s\sqrt{\frac{1}{\nu}+\frac{1}{\kappa}} \leq \mu_1 - \mu_2 \leq \bar{x} - \bar{y} + t_{\nu+\kappa-2;\frac{\alpha}{2}}\, s\sqrt{\frac{1}{\nu}+\frac{1}{\kappa}}$$

όπου $s = \sqrt{\frac{(\nu-1)s_1^2+(\kappa-1)s_2^2}{\nu+\kappa-2}}$

4.5.3. Δείγματα μικρά ανεξάρτητα, κανονικοί πληθυσμοί με άγνωστες και διαφορετικές διασπορές $\sigma_1^2 \neq \sigma_2^2$.

Αν πάρουμε δύο δείγματα **μικρά ανεξάρτητα** μεγέθους ν και κ με μέσες τιμές $\bar{x}$ και $\bar{y}$ και διακυμάνσεις s_1^2, s_2^2 αντίστοιχα προερχόμενα από **κανονικούς πληθυσμούς** με **άγνωστες και διαφορετικές διασπορές** $\sigma_1^2 \neq \sigma_2^2$ έχουμε ότι η μεταβλητή:

$$t = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{\nu} + \frac{s_2^2}{\kappa}}}$$

ακολουθεί **t – κατανομή με βαθμούς ελευθερίας λ = 2(ν-1)** όταν **ν = κ**

και $\lambda = \dfrac{\left(\frac{s_1^2}{\nu} + \frac{s_2^2}{\kappa}\right)^2}{\frac{\left(\frac{s_1^2}{\nu}\right)^2}{\nu - 1} + \frac{\left(\frac{s_2^2}{\kappa}\right)^2}{\kappa - 1}}$ (στρογγυλεμένο στον πλησιέστερο ακέραιο)

όταν $\nu \neq \kappa$

Ένα **100(1-α)% διάστημα εμπιστοσύνης** για την διαφορά $\mu_1 - \mu_2$ των δύο πληθυσμών όταν **οι διασπορές τους είναι άγνωστες και διαφορετικές** $\sigma_1^2 \neq \sigma_2^2$ είναι:

$$\bar{x} - \bar{y} - t_{\lambda;\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{\nu} + \frac{s_2^2}{\nu}} \leq \mu_1 - \mu_2 \leq \bar{x} - \bar{y} + t_{\lambda;\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{\nu} + \frac{s_2^2}{\nu}} \quad \text{για } \nu = \kappa$$

ή

$$\bar{x} - \bar{y} - t_{\lambda;\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{\nu} + \frac{s_2^2}{\kappa}} \leq \mu_1 - \mu_2 \leq \bar{x} - \bar{y} + t_{\lambda;\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{\nu} + \frac{s_2^2}{\kappa}} \quad \text{για } \nu \neq \kappa$$

Παραδείγματα:

i. Δύο εργοστάσια κατασκευάζουν το ίδιο εξάρτημα για μια μηχανή.

Παίρνουμε ένα δείγμα 30 εξαρτημάτων από το πρώτο εργοστάσιο και διαπιστώνουμε ότι το μέσο βάρος τους είναι 700 kgr και έχουν δειγματική διακύμανση 400 kgr^2. Λαμβάνουμε επίσης ένα δείγμα 40 εξαρτημάτων από το δεύτερο εργοστάσιο και βρίσκουμε ότι έχει μέσο βάρος 720 kgr με δειγματική διακύμανση 450 kgr^2.

Να βρεθεί το διάστημα εμπιστοσύνης της διαφοράς μ_1 - μ_2 των πραγματικών μέσων των πληθυσμών με πιθανότητα 99%. Ποιου εργοστασίου το εξάρτημα είναι βαρύτερο κατά μέσο όρο;

Έχουμε $\bar{x}_1 = 700 \qquad \bar{x}_2 = 720$

$$s_1^2 = 400, s_2^2 = 450$$
$$n_1 = 30, \qquad n_2 = 40.$$

$$\alpha = 1 - 0{,}99 = 0{,}01 \qquad \to \frac{\alpha}{2} = 0{,}005 \to z_{\frac{\alpha}{2}} = \phi^{-1}\left(1 - \frac{\alpha}{2}\right) = \phi^{-1}(0{,}995) = 2{,}58.$$

$$s^2_{\bar{x}_1 - \bar{x}_2} = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} = \frac{400}{30} + \frac{450}{40} = 24{,}5833$$

$$s_{\bar{x}_1 - \bar{x}_2} = 4{,}9581, \bar{x}_1 - \bar{x}_2 = -20$$

$$-20 - 2{,}58 \cdot 4{,}9581 < \mu_1 - \mu_2 < -20 + 2{,}58 \cdot 4{,}9581$$

$$-32{,}79 < \mu_1 - \mu_2 < -7{,}21 \text{ ή } \quad 32{,}79 > \mu_1 - \mu_2 > 7{,}21.$$

Το εξάρτημα του δεύτερου εργοστασίου είναι βαρύτερο κατά μέσο όρο.

ii. Αν στο πιο πάνω παράδειγμα γνωρίζουμε ότι ύστερα από μετρήσεις οι διακυμάνσεις των πληθυσμών είναι $\sigma_1^2 = 420 \kappa\alpha\iota \sigma_2^2 = 480$ τότε το διάστημα εμπιστοσύνης της διαφοράς μ_1 - μ_2, με πιθανότητα 99%, θα είναι:

$z_{\frac{\alpha}{2}}$ =2,58

$$\sigma^2_{\bar{x}_1-\bar{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{420}{30} + \frac{480}{40} = 26 \rightarrow \sigma_{\bar{x}_1-\bar{x}_2} = 5{,}099$$

$$-20 - 2{,}58 \cdot 5{,}099 < \mu_1 - \mu_2 < -20 + 2{,}58 \cdot 5{,}099$$

$$-33{,}16 < \mu_1 - \mu_2 < -6{,}84 \Rightarrow 33{,}16 > \mu_2 - \mu_1 > 6{,}84.$$

iii. Παίρνουμε ένα δείγμα 10 κονσερβών από ένα εργοστάσιο κονσερβοποιίας και διαπιστώνουμε ότι το μέσο βάρος είναι 700 γρ. με διακύμανση 400 γρ^2. Λαμβάνουμε επίσης ένα δεύτερο δείγμα 18 κονσερβών από ένα δεύτερο εργοστάσιο κονσερβοποιίας και διαπιστώνουμε ότι το μέσο βάρος είναι 720 γρ. με διακύμανση 450 γρ.

Να βρεθεί το διάστημα εμπιστοσύνης της διαφοράς μ_1-μ_2 των πραγματικών μέσων βαρών των κονσερβών που παράγονται στα δύο εργοστάσια με πιθανότητα 99%.

Θεωρούμε ότι οι πληθυσμοί των βαρών των κονσερβών των δύο εργοστασίων είναι κανονικοί και ότι οι διακυμάνσεις των βαρών των κονσερβών στα δύο εργοστάσια είναι άγνωστες αλλά ίσες.

Ένα 100(1-α)% διάστημα εμπιστοσύνης για την διαφορά μ_1- μ_2 των δύο πληθυσμών όταν **οι διακυμάνσεις τους** είναι **άγνωστες αλλά ίσες** είναι:

$$\bar{x} - \bar{y} - t_{\nu+\kappa-2;\frac{\alpha}{2}} s\sqrt{\frac{1}{\nu}+\frac{1}{\kappa}} \leq \boldsymbol{\mu_1 - \mu_2} \leq \bar{x} - \bar{y} + t_{\nu+\kappa-2;\frac{\alpha}{2}} s\sqrt{\frac{1}{\nu}+\frac{1}{\kappa}}$$

όπου $s = \sqrt{\dfrac{(\nu-1)s_1^2 + (\kappa-1)s_2^2}{\nu+\kappa-2}}$

Έχουμε $\bar{x} = 700, \bar{y} = 720$

$s_1^2 = 400, s_2^2 = 450$, $n_1 = 10, n_2 = 18.$, $n_1 + n_2 - 2 = 26$,

$\alpha = 1 - 0{,}99 = 0{,}01, \dfrac{\alpha}{2} = 0{,}005$ και $t_{\nu+\kappa-2;\frac{\alpha}{2}}$ = 2,779 s = $\sqrt{\dfrac{(10-1)400+(18-1)450}{26}}$ = 21,62

s$\sqrt{\dfrac{1}{\nu}+\dfrac{1}{\kappa}}$ = 21,62 * 0,39 = 8,54

$$\bar{x}_1-\bar{x}_2=-20$$

$$-20-2{,}779\cdot 8{,}54<\mu_1-\mu_2<-20+2{,}779\cdot 8{,}54$$

$$-43{,}73<\mu_1-\mu_2<-3{,}73$$

iv. Παίρνουμε ένα δείγμα 10 κονσερβών από ένα εργοστάσιο κονσερβοποιίας και διαπιστώνουμε ότι το μέσο βάρος είναι 700 γρ. με διακύμανση 400 γρ^2. Λαμβάνουμε επίσης ένα δεύτερο δείγμα 10 κονσερβών από ένα δεύτερο εργοστάσιο κονσερβοποιίας και διαπιστώνουμε ότι το μέσο βάρος είναι 720 γρ. με διακύμανση 450 γρ^2.

Να βρεθεί το διάστημα εμπιστοσύνης της διαφοράς μ_1-μ_2 των πραγματικών μέσων βαρών των κονσερβών που παράγονται στα δύο εργοστάσια με πιθανότητα 99%.

Θεωρούμε ότι οι πληθυσμοί των βαρών των κονσερβών των δύο εργοστασίων είναι κανονικοί και ότι οι διακυμάνσεις των βαρών των κονσερβών στα δύο εργοστάσια είναι **άγνωστες και διαφορετικές.**

Ένα 100(1-α)% διάστημα εμπιστοσύνης για την διαφορά μ_1-μ_2 των δύο πληθυσμών, όταν οι διακυμάνσεις τους είναι **άγνωστες και διαφορετικές $\sigma_1^2 \neq \sigma_2^2$** και τα δείγματα είναι ίσα είναι:

$$\bar{x} - \bar{y} - t_{\lambda;\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{\nu}+\frac{s_2^2}{\nu}} \le \mu_1 - \mu_2 \le \bar{x} - \bar{y} + t_{\lambda;\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{\nu}+\frac{s_2^2}{\nu}}$$ για **ν = κ**

Έχουμε $\bar{x}=700, \bar{y}=720$

$s_1^2=400, s_2^2=450$, $n_1=10, n_2=10$,

$\alpha=1-0{,}99=0{,}01, \dfrac{\alpha}{2}=0{,}005$, λ = 2(10-1) = 18 και $t_{\lambda;\frac{\alpha}{2}}$ = 2,878

$\sqrt{\frac{400}{10}+\frac{450}{10}}$ = 9,22

$$-20-2{,}878\cdot 9{,}22<\mu_1-\mu_2<-20+2{,}878\cdot 9{,}22$$

ή

$$-46{,}54 < \mu_1 - \mu_2 < 6{,}54$$

ν. Παίρνουμε ένα δείγμα 10 κονσερβών από ένα εργοστάσιο κονσερβοποιίας και διαπιστώνουμε ότι το μέσο βάρος είναι 700 γρ. με διακύμανση 400 γρ^2. Λαμβάνουμε επίσης ένα δεύτερο δείγμα 16 κονσερβών από ένα δεύτερο εργοστάσιο κονσερβοποιίας και διαπιστώνουμε ότι το μέσο βάρος είναι 720 γρ. με διακύμανση 450 γρ^2.

Να βρεθεί το διάστημα εμπιστοσύνης της διαφοράς μ_1-μ_2 των πραγματικών μέσων βαρών των κονσερβών που παράγονται στα δύο εργοστάσια με πιθανότητα 99%.

Θεωρούμε ότι οι πληθυσμοί των βαρών των κονσερβών των δύο εργοστασίων είναι κανονικοί και ότι οι διακυμάνσεις των βαρών των κονσερβών στα δύο εργοστάσια είναι **άγνωστες και διαφορετικές.**

Ένα 100(1-α)% διάστημα εμπιστοσύνης για την διαφορά μ_1-μ_2 των δύο πληθυσμών όταν οι διακυμάνσεις τους είναι άγνωστες και διαφορετικές $\sigma_1^2 \neq \sigma_2^2$ και τα δείγματα είναι άνισα είναι:

$$\bar{x} - \bar{y} - t_{\lambda;\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{\nu}+\frac{s_2^2}{\nu}} \leq \boldsymbol{\mu_1 - \mu_2} \leq \bar{x} - \bar{y} + t_{\lambda;\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{\nu}+\frac{s_2^2}{\nu}} \quad \text{για } \boldsymbol{\nu \neq \kappa}$$

Έχουμε $\bar{x}=700, \bar{y}=720$

$s_1^2=400, s_2^2=450$, $n_1=10, n_2=16$

$\alpha=1-0{,}99=0{,}01, \frac{\alpha}{2}=0{,}005$,

$$\lambda = \frac{\left(\frac{400}{10}+\frac{450}{16}\right)^2}{\frac{\left(\frac{400}{10}\right)^2}{10-1}+\frac{\left(\frac{450}{16}\right)^2}{16-1}} = 20$$

και $t_{\lambda;\frac{\alpha}{2}} = 2{,}845$

$$\sqrt{\frac{400}{10}+\frac{450}{10}} = 9{,}22$$

$$-20-2{,}845\cdot 9{,}22 < \mu_1 - \mu_2 < -20 + 2{,}845\cdot 9{,}22$$

ή

$$-46{,}23 < \mu_1 - \mu_2 < 6{,}23$$

4.5.4.Δείγματα μικρά εξαρτημένα (ζευγαρωτές παρατηρήσεις). Κανονικοί πληθυσμοί.

Για **δείγματα μικρά εξαρτημένα (ζευγαρωτές παρατηρήσεις)** που προέρχονται από μετρήσεις της ίδιας ομάδας σε δυο διαφορετικές χρονικές στιγμές (ζευγαρωτές παρατηρήσεις) ορίσουμε x_i και y_i, i=1,2,….,n τις παρατηρήσεις στα δύο δείγματα και δημιουργούμε τις αντίστοιχες διαφορές $z_i = x_i - y_i$, i=1,2,….,n που τις θεωρούμε διαφορετικές. Ο πληθυσμός από όπου πήραμε τα ζεύγη θεωρείται **κανονικός.**
Θεωρούμε έναν πρώτο πληθυσμό με μέση τιμή μ_1 (ο οποίος προσδιορίζεται πάντα από τις πρώτες παρατηρήσεις x_i) και έναν δεύτερο πληθυσμό με μέση τιμή μ_2 (ο οποίος προσδιορίζεται πάντα από τις δεύτερες παρατηρήσεις y_i).

Οι παρατηρήσεις z_i ακολουθούν την t – κατανομή με n - 1 βαθμούς ελευθερίας.

Η μεταβλητή $t = \dfrac{\bar{Z}}{\frac{s_Z}{\sqrt{n}}}$ που ακολουθεί t – κατανομή με n -1 βαθμούς ελευθερίας

Ένα 100(1-α)% διάστημα εμπιστοσύνης για τη διαφορά $\mathbf{\mu_1}$- $\mathbf{\mu_2}$ είναι:

$$\bar{Z} - t_{n-1;\frac{a}{2}} \frac{S_z}{\sqrt{n}} \leq \mathbf{\mu_1 - \mu_2 = \mu_Z} \leq \bar{Z} - t_{n-1;\frac{a}{2}} \frac{S_z}{\sqrt{n}}$$

Όταν το δείγμα είναι μεγάλο τότε έχουμε : $t_{n-1;\frac{a}{2}} = Z_{\frac{a}{2}}$

Παράδειγμα:

Έχουμε τις παρακάτω ζευγαρωτές παρατηρήσεις:

X:	4	5	6	4,2	5,2	5,3	6,4	4,8	5,3	5
Y:	4,9	4,8	5,7	5	6	5.2	6,5	5,9	4,8	5,7

Να βρεθεί ένα 90% διάστημα εμπιστοσύνης για την πραγματική διαφορά .

Έχουμε Z: -0,9 0,2 0,3 -0,8 -0,8 0,1 -0,1 -1,1 0,5 -0,7

Και $\bar{Z}$ = -0,33 , S_Z = 0,72 , $t_{9;0,05}$ = 1,833

Επομένως ένα 90% διάστημα εμπιστοσύνης για την πραγματική διαφορά μ_1- μ_2 είναι:

$$\bar{Z} - t_{9;0,05}\ S_Z / \sqrt{n} \leq \mathbf{\mu_1 - \mu_2 = \mu_Z} \leq \bar{Z} + t_{9;0,05}\ S_Z / \sqrt{n}$$

ή

$$-0,33 - 1,833 \times 0,72/ 3,16 \leq \mathbf{\mu_1 - \mu_2 = \mu_Z} \leq -0,33 + 1,833 \times 0,72/ 3,16$$

ή

$$-0,75 \leq \mathbf{\mu_1 - \mu_2 = \mu_Z} \leq 0,09$$

4.6. Εκτίμηση διαστήματος εμπιστοσύνης για την διακύμανση του πληθυσμού

Στην θεωρία διατυπώθηκε ότι η μεταβλητή:

$$X^2 = \frac{(n-1)s^2}{\sigma^2}$$

ακολουθεί την X^2 κατανομή με n-1 βαθμούς.

Ο πληθυσμός από όπου πάρθηκε το δείγμα θεωρείται κανονικός με σ^2 και s^2 η δειγματική διακύμανση.

Ένα 100(1-α)% διάστημα εμπιστοσύνης για την διακύμανση σ^2 του πληθυσμού είναι:

$$\frac{(n-1)s^2}{X^2_{n-1;\frac{a}{2}}} \leq \sigma^2 \leq \frac{(n-1)s^2}{X^2_{n-1;1-\frac{a}{2}}}$$

Ακολουθείται η εξής διαδικασία για τον υπολογισμό του διαστήματος:

- Επιλέγεται το διάστημα εμπιστοσύνης 1-α, με α$\in (0,1)$.
- Υπολογίζεται το $x^2_{n-1;a/2}$ $\kappa\alpha\iota$ $x^2_{n-1;1-a/2}$ ποσοστιαία σημεία της X^2_{n-1} κατανομής
- Υπολογίζονται οι τιμές $\frac{(n-1)S^2}{x^2_{n-1;a/2}}, \frac{(n-1)S^2}{x^2_{n-1;1-a/2}}$
- Έτσι λαμβάνεται το ακόλουθο αριθμητικό διάστημα εμπιστοσύνης:

$$\left[\frac{(n-1)S^2}{x^2_{n-1;a/2}}, \frac{(n-1)S^2}{x^2_{n-1;1-a/2}}\right]$$

Παράδειγμα:

Ένα δείγμα από 51 αγρότες παρουσιάζει διακύμανση των ετήσιων εισοδημάτων τους s^2= 156 ευρώ^2. Να βρεθεί ένα 95% διάστημα εμπιστοσύνης για την διακύμανση σ^2 των ετήσιων εισοδημάτων του πληθυσμού των αγροτών.

Έχουμε ότι:

$$\frac{(n-1)s^2}{X^2_{n-1;\frac{a}{2}}} \leq \sigma^2 \leq \frac{(n-1)s^2}{X^2_{n-1;1-\frac{a}{2}}}$$

ή

$$\frac{(51-1)156}{71,42} \leq \sigma^2 \leq \frac{(51-1)156}{32,36}$$

[$X^2_{50;0,025}$ = 71,42 , $X^2_{50;0,975}$ = 32,36]

ή

$$109,21 \leq \sigma^2 \leq 241,04$$

4.7. Εκτίμηση διαστήματος εμπιστοσύνης για τον λόγο των διακυμάνσεων δύο πληθυσμών

Σύμφωνα με την θεωρία αν πάρουμε δύο δείγματα ανεξάρτητα μεγέθους n και m αντίστοιχα με s_1^2, s_2^2 τις δειγματικές διακυμάνσεις από δύο **κανονικούς πληθυσμούς** αντίστοιχα με σ_1^2 , σ_2^2 τις πληθυσμιακές διακυμάνσεις τότε ισχύει ότι η μεταβλητή:

$$F = \frac{s_1^2}{s_2^2}\frac{\sigma_2^2}{\sigma_1^2}$$

ακολουθεί την F κατανομή με n -1 και m -1 βαθμούς ελευθερίας

Οι πληθυσμοί είναι κανονικοί.

Ένα 100(1-α)% διάστημα εμπιστοσύνης για τον λόγο των διακυμάνσεων σ_1^2,σ_2^2 των δύο πληθυσμών είναι:

$$\frac{s_1^2}{s_2^2}\frac{1}{F_{n-1,m-1;\frac{a}{2}}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2}F_{m-1,n-1;\frac{a}{2}}$$

Παράδειγμα:

Πήραμε δύο δείγματα αγροτών από δύο διοικητικές περιφέρειες της Ελλάδας και εξετάσαμε τα ετήσια εισοδήματά τους. Είχαμε τα ακόλουθα δεδομένα:

n = 31, S_1^2 = 220 ευρώ ^{2} και m = 41, S_2^2 = 200 ευρώ ^{2}

Να βρεθεί ένα 98% διάστημα εμπιστοσύνης για τον λόγο των πραγματικών διακυμάνσεων σ_1^2, σ_2^2 των ετήσιων εισοδημάτων των αγροτικών πληθυσμών στις δύο διοικητικές περιφέρειες της Ελλάδας.

Έχουμε ότι:

$$\frac{s_1^2}{s_2^2}\frac{1}{F_{n-1,m-1;\frac{a}{2}}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2}F_{m-1,n-1;\frac{a}{2}}$$

$F_{30,40;0,01}$ = 2,2 $F_{40,30;0,01}$ = 2,3

$$\frac{220}{200}\frac{1}{2,2} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{220}{200}2,3$$

ή

$$0,50 \le \frac{\sigma_1^2}{\sigma_2^2} \le 2,53$$

4.8. Εκτίμηση διαστήματος εμπιστοσύνης αναλογίας σε πληθυσμό.

Αν εκτιμήσουμε το ποσοστό (αναλογία) $p_\Delta = \frac{x}{n}$ των στοιχείων στο **μεγάλο δείγμα μεγέθους n** που έχουν κάποιο συγκεκριμένο χαρακτηριστικό, τότε σύμφωνα με την θεωρία το δειγματικό ποσοστό κατανέμεται κανονικά:

Δηλαδή η δειγματική μεταβλητή P_{Δ}, που παράγεται από τις τιμές p_{Δ}, ακολουθεί την κανονική κατανομή:

$$N\left(P_{\pi\lambda\eta\theta}, \frac{p_{\pi\lambda\eta\theta}(1-P_{\pi\lambda\eta\theta})}{n}\right)$$

Όπου $P_{\pi\lambda\eta\theta}$ το πραγματικό ποσοστό στον πληθυσμό.

Συνεπώς έχουμε ότι η μεταβλητή:

$$Z = \frac{P_{\Delta} - P_{\pi\lambda\eta\theta}}{\sqrt{\frac{P_{\pi\lambda\eta\theta}(1-P_{\pi\lambda\eta\theta})}{n}}}$$ ακολουθεί την **N(0,1).**

Ένα 100(1-α)% διάστημα εμπιστοσύνης για το ποσοστό (αναλογία) του πληθυσμού είναι:

$$\rho_{\Delta} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\rho_{\Delta}(1-\rho_{\Delta})}{n}} \le P_{\pi\lambda\eta\theta} \le \rho_{\Delta} + z_{\frac{\alpha}{2}}\sqrt{\frac{\rho_{\Delta}(1-\rho_{\Delta})}{n}}$$

Παράδειγμα:

Ο αριθμός των ελαττωματικών προϊόντων μιας μεταποιητικής βιομηχανίας αγροτικών προϊόντων που εντοπίσθηκαν σ' ένα δείγμα 300 προϊόντων είναι 24 προϊόντα.

Να βρεθούν τα όρια μέσα στα οποία θα βρίσκεται το πραγματικό ποσοστό $P_{\pi\lambda\eta\theta}$ των ελαττωματικών προϊόντων της συνολικής παραγωγής με πιθανότητα 98%.

Έχουμε n = 300 , m = 24 $\rightarrow$ $\rho_{\Delta} = \frac{24}{300} = 8\% = 0{,}08.$

Οπότε 1- p_{Δ} = 0,92.

$$\left.\begin{aligned} \alpha &= 1 - 0{,}98 = 0{,}02 \\ \frac{\alpha}{2} &= 0{,}01 \end{aligned}\right\} \rightarrow z_{\frac{\alpha}{2}} = \phi^{-1}\left(1 - \frac{\alpha}{2}\right) = \phi^{-1}(0{,}99) = 2{,}33.$$

$$\frac{\sqrt{\rho_{\Delta} \cdot (1-\rho_{\Delta})}}{\sqrt{n}} = \frac{\sqrt{0{,}08 \cdot 0{,}92}}{\sqrt{300}} = 0{,}0157$$

$$0{,}08 - 0{,}0157 \cdot 2{,}33 < P_{\pi\lambda\eta\theta} < 0{,}08 + 0{,}0157 \cdot 2{,}33$$

$$0{,}0434 < P_{\pi\lambda\eta\theta} < 0{,}1165 \Rightarrow 4{,}34\% < P_{\pi\lambda\eta\theta} < 11{,}6\%.$$

4.9. Εκτίμηση διαστήματος εμπιστοσύνης για τη διαφορά των αναλογιών στοιχείων δύο πληθυσμών.

Έστω $ρ_{1\,Δ} = \frac{x}{n}$ και $ρ_{2\,Δ} = \frac{y}{l}$ είναι οι δειγματικές αναλογίες των στοιχείων, που έχουν κάποιο συγκεκριμένο χαρακτηριστικό, δύο **μεγάλων δειγμάτων μεγέθους n, l** από δύο πληθυσμούς, όπου οι πραγματικές αναλογίες είναι αντίστοιχα $P_{1\,\pi\lambda\eta\theta}$, $P_{2\,\pi\lambda\eta\theta}$.

Τότε έχουμε ότι οι διαφορές των δειγματικών μεταβλητών $P_{1\,Δ}$- $P_{2\,Δ}$ ακολουθούν κανονική κατανομή:

$$N\left(P_{1\,\pi\lambda\eta\theta} - P_{2\,\pi\lambda\eta\theta},\ \frac{P_{1\pi\lambda\eta\theta}(1-P_{1\pi\lambda\eta\theta})}{n} + \frac{P_{2\pi\lambda\eta\theta}(1-P_{2\pi\lambda\eta\theta})}{l}\right)$$

Ένα 100(1-α)% διάστημα εμπιστοσύνης για τη διαφορά των ποσοστών (αναλογιών) των δύο πληθυσμών $P_{1\,\pi\lambda\eta\theta}$, $P_{2\,\pi\lambda\eta\theta}$ είναι:

$$ρ_{1\,Δ} - ρ_{2\,Δ} - Z_{\frac{a}{2}}\sqrt{\frac{\rho_{1Δ}(1-\rho_{1Δ})}{n} + \frac{\rho_{2Δ}(1-\rho_{2Δ})}{l}} \leq \mathbf{P}_{1\,\pi\lambda\eta\theta} - \mathbf{P}_{2\,\pi\lambda\eta\theta} \leq ρ_{1\,Δ} - ρ_{2\,Δ} + Z_{\frac{a}{2}}\sqrt{\frac{\rho_{1Δ}(1-\rho_{1Δ})}{n} + \frac{\rho_{2Δ}(1-\rho_{2Δ})}{l}}$$

Παράδειγμα:

Παίρνουμε δύο μεγάλα δείγματα μιας ποικιλίας ενός φυτού από δύο διαφορετικούς πληθυσμούς και εξετάζουμε πόσα από αυτά ασθένησαν μέσα σε ένα συγκεκριμένο χρονικό διάστημα. Τα δεδομένα που προέκυψαν ήταν: Από τα n=120 του πρώτου δείγματος ασθένησαν τα 12 και από τα m=130 του δεύτερου δείγματος ασθένησαν τα 18. Να βρεθεί ένα 95% διάστημα εμπιστοσύνης για τη διαφορά των πραγματικών ποσοστών (αναλογιών) των ασθενούντων φυτών των δύο πληθυσμών $P_{1\ \pi\lambda\eta\theta}$, $P_{2\ \pi\lambda\eta\theta}$.

Έχουμε ότι:

p_1- p_2- $Z_{\frac{a}{2}}\sqrt{\frac{\rho_{1\Delta}(1-\rho_{1\Delta})}{n}+\frac{\rho_{2\Delta}(1-\rho_{2\Delta})}{l}} \leq$ $\mathbf{P}_{1\ \pi\lambda\eta\theta}$**-** $\mathbf{P}_{2\ \pi\lambda\eta\theta} \leq$ p_1-p_2+ $Z_{\frac{a}{2}}\sqrt{\frac{\rho_{1\Delta}(1-\rho_{1\Delta})}{n}+\frac{\rho_{2\Delta}(1-\rho_{2\Delta})}{l}}$

$\rho_{1\ \Delta}$= 12/120 = 0,1, $\rho_{2\ \Delta}$ = 18/130 = 0,14

Έχουμε: $\rho_{1\ \Delta}$- $\rho_{2\ \Delta}$ = -0,04

$Z_{0,025}$= 1,96

$\sqrt{\frac{0,1x0,9}{120}+\frac{0,14x0,86}{130}}$ = 0,04

Άρα ένα 95% διάστημα εμπιστοσύνης για τη διαφορά των πραγματικών ποσοστών (αναλογιών) των ασθενούντων φυτών των δύο πληθυσμών $P_{1\ \pi\lambda\eta\theta}$, $P_{2\ \pi\lambda\eta\theta}$ είναι:

$$-0,04 - 1,96 \times 0,04 \leq \mathbf{P}_{1\ \pi\lambda\eta\theta} - \mathbf{P}_{2\ \pi\lambda\eta\theta} \leq -0,04 + 1,96 \times 0,04$$

ή

$$-0,118 \leq \mathbf{P}_{1\ \pi\lambda\eta\theta} - \mathbf{P}_{2\ \pi\lambda\eta\theta} \leq 0,038$$

Κεφάλαιο 5^{ο}

Έλεγχοι υποθέσεων

5.1 . Γενικά

Εκτός του προσδιορισμού του διαστήματος εμπιστοσύνης μιας αγνώστου παραμέτρου Θ του πληθυσμού πολλές φορές απαιτείται να κάνουμε υποθέσεις για την τιμή που μπορεί να πάρει η Θ, τις οποίες και ελέγχουμε.

Σημαντικό ρόλο στον έλεγχο της υπόθεσης που κάνουμε για την άγνωστη παράμετρο Θ του πληθυσμού παίζει η **εκτιμήτρια θ** και το **στατιστικό του ελέγχου** από το δείγμα.

Καταρχήν η υπόθεση που διατυπώνουμε για την άγνωστη παράμετρο Θ του πληθυσμού καλείται H_0 και είναι της μορφής:

H_0: $\Theta = \theta^*$ όπου θ^* είναι μια συγκεκριμένη τιμή που υποθέτουμε ότι μπορεί να πάρει η Θ.

Η υπόθεση αυτή ελέγχεται αν ισχύει η όχι και καλείται **μηδενική υπόθεση**.

Οι εναλλακτικές υποθέσεις είναι τις μορφής H_1: $\Theta \neq \theta^*$, H_1: $\Theta > \theta^*$, H_1: $\Theta < \theta^*$

Έτσι διαμορφώνονται οι ακόλουθες υποθέσεις προς έλεγχο:

H_0: $\Theta = \theta^*$
H_1: $\Theta \neq \theta^*$ έχουμε τότε δίπλευρο έλεγχο.

H_0: $\Theta = \theta^*$
H_1: $\Theta > \theta^*$ έχουμε τότε μονόπλευρο έλεγχο (δεξιά)

H_0: $\Theta = \theta^*$
H_1: $\Theta < \theta^*$ έχουμε τότε μονόπλευρο έλεγχο (αριστερά)

Για τον έλεγχο υπολογίζεται η **απορριπτική περιοχή R** της H_0 ,δηλαδή η περιοχή στα σημεία τη οποίας η H_0 απορρίπτεται.

Αυτή προσδιορίζεται από την κατανομή που ακολουθεί **το στατιστικό του ελέγχου,** το **σφάλμα α** που λαμβάνεται υπόψη και **η μορφή του ελέγχου** που γίνεται (δίπλευρος ή μονόπλευρος δεξιά ή αριστερά).

Συνεπώς τα στοιχεία ενός ελέγχου μηδενικής υπόθεσης είναι τα ακόλουθα:

1. Ορισμός της μηδενικής υπόθεσης
2. Ορισμός της εναλλακτικής υπόθεσης
3. Ορισμός του στατιστικού του ελέγχου από το δείγμα
4. Ορισμός της απορριπτικής περιοχής R της H_0
5. Εξαγωγή συμπερασμάτων.

5.2. Σφάλματα – στάθμη σημαντικότητας – περιοχή απόρριψης της H_0

Το α είναι η πιθανότητα να απορρίψουμε την H_0 ενώ είναι σωστή:

$$\alpha = P(\text{απόρριψη της } H_0 \,/\, H_0 \text{ σωστή})$$

Το α καλείται και **σφάλμα τύπου I.**

Το β είναι η πιθανότητα να δεχτούμε την H_0 ενώ είναι λάθος:

$$\beta = P(\text{αποδοχή της } H_0 \,/\, H_0 \text{ λάθος})$$

Το β καλείται και **σφάλμα τύπου II.**

Το γ = 1 – β και εκφράζει **την πιθανότητα απόρριψης της H_0 όταν η H_0 είναι πράγματι λάθος**.
Το γ καλείται και **ισχύς** του στατιστικού του ελέγχου**.**

Η απορριπτική περιοχή της R της H_0 ορίζεται βάσει του σφάλματος α που καλείται **στάθμη σημαντικότητας ή επίπεδο σημαντικότητας (σ. σ).**

Συγκεκριμένα επίπεδο σημαντικότητας α ενός ελέγχου μηδενικής υπόθεσης H_0 ονομάζουμε π.χ την πιθανότητα να παρατηρηθεί μια τιμή του στατιστικού του ελέγχου, που έδωσε το δείγμα, μεγαλύτερη από την κριτήρια τιμή που παίρνουμε από τους πίνακες και η οποία προσδιορίζεται από το μέγεθος του δείγματος και την κατανομή που ακολουθεί το στατιστικό ελέγχου με το οποίο ελέγχουμε τη μηδενική υπόθεση.

Δηλαδή η πιθανότητα P(Y>y / H_0 σωστή), όπου Y η τ.μ που αντιστοιχεί στο στατιστικό και y η κριτήρια τιμή του στατιστικού από τους πίνακες. Η πιθανότητα αυτή αναφέρεται σε μονόπλευρους ελέγχους ενώ σε δίπλευρους ελέγχους η πιθανότητα διπλασιάζεται.

Η υπόθεση H_0 απορρίπτεται εάν η παρατηρούμενη πιθανότητα α = P(απόρριψη της H_0 / H_0 σωστή) που προκύπτει από την τιμή του στατιστικού είναι μικρότερη μιας ορισμένης στάθμης σημαντικότητας που ορίσαμε για να ελέγξουμε την μηδενική υπόθεση.

Πως ορίζεται το στατιστικό και η απορριπτική περιοχή R

Εάν η εκτιμήτρια θ ακολουθεί κανονική κατανομή ή προσεγγιστικά κανονική κατανομή τότε βάσει της θεωρίας η μεταβλητή:

$$Z = \frac{|\theta - \theta^*|}{\tau}$$ ακολουθεί τυπική κανονική κατανομή.

Όπου τ το τυπικό σφάλμα (τυπική απόκλιση) της κατανομής της εκτιμήτριας θ.

Η υπόθεση H_0 απορρίπτεται όταν $Z = \frac{|\theta - \theta^*|}{\tau} > Z_{\frac{\alpha}{2}}$ με α το επίπεδο σημαντικότητας του ελέγχου (δίπλευρος έλεγχος).

Γενικά όταν το δείγμα προέρχεται από **κανονικό πληθυσμό** με την **προϋπόθεση ότι ισχύει η H_0** η μεταβλητή $X = \frac{\theta - \theta^*}{\tau}$ ακολουθεί **γνωστή κατανομή** με περιοχή απόρριψης της H_0 εκεί όπου:

$$X = \frac{\theta - \theta^*}{\tau} > \Phi_\alpha, \quad X = \frac{\theta - \theta^*}{\tau} < -\Phi_\alpha \quad \text{ή} \quad |X| = \left|\frac{\theta - \theta^*}{\tau}\right| > \Phi_{\frac{\alpha}{2}}$$

όταν οι εναλλακτικές υποθέσεις είναι αντίστοιχα:

H_1: Θ > θ^* , H_1: Θ < θ^* ή H_1: Θ $\neq$ θ^*

Οι Φ_{α} , $\Phi_{\frac{\alpha}{2}}$ που παίρνονται από πίνακες είναι τιμές της κατανομής που ακολουθεί η μεταβλητή:

$$X = \frac{\theta - \theta^{*}}{\tau}$$

Για τις οποίες έχουμε :

$P(X = \frac{\theta - \theta^{*}}{\tau} > \Phi_{\alpha}) = \alpha$, $P(X = \frac{\theta - \theta^{*}}{\tau} < -\Phi_{\alpha}) = \alpha$ και

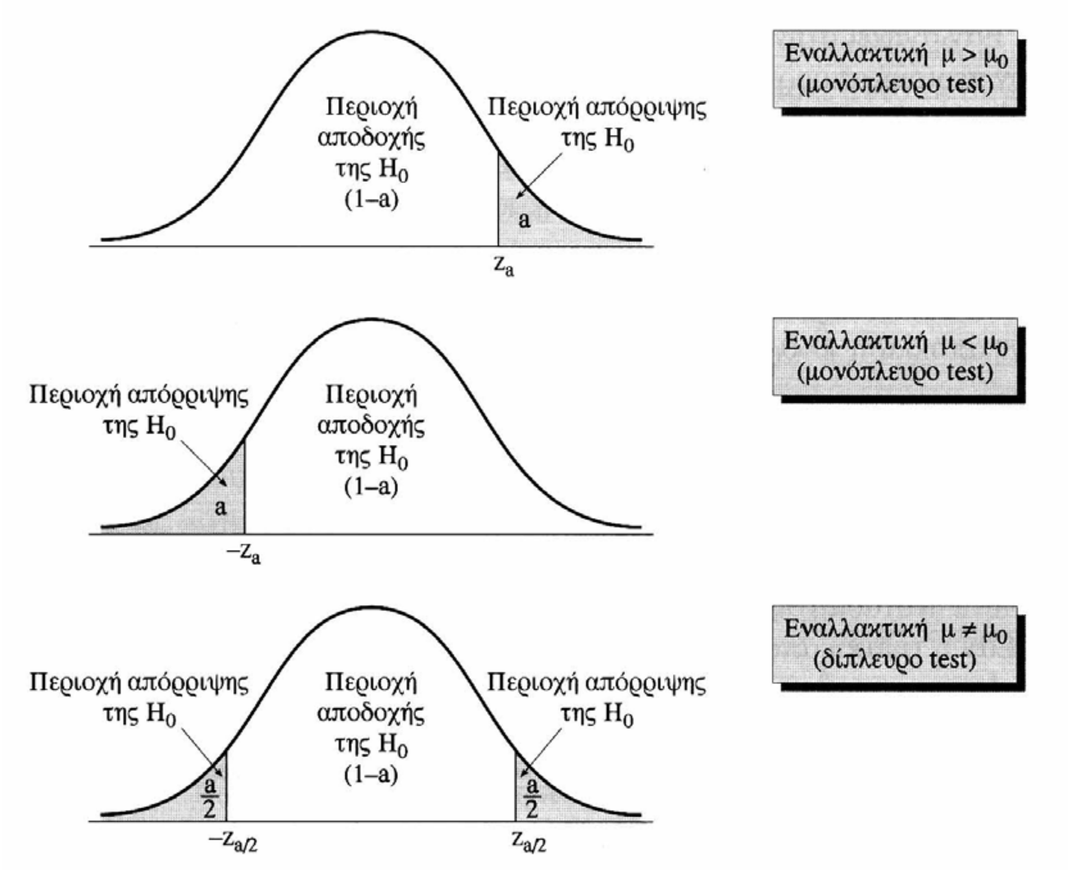
$P(|X| = \left|\frac{\theta - \theta^{*}}{\tau}\right| > \Phi_{\frac{\alpha}{2}}) = \alpha$

<u>Προσδιορισμός της απορριπτικής περιοχής των υποθέσεων:</u>

H_0: μ = μ_0
H_1: μ > μ_0 Μονόπλευρος έλεγχος (δεξιά)

H_0: μ = μ_0
H_1: μ < μ_0 Μονόπλευρος έλεγχος (αριστερά)

H_0: μ = μ_0
H_1: μ > μ_0 Δίπλευρος έλεγχος



5.3. Έλεγχοι υποθέσεων

5.3.1. Έλεγχος υπόθεσης για τη μέση τιμή του πληθυσμού

5.3.1.1. Έλεγχος υπόθεσης για τη μέση τιμή του πληθυσμού (όταν $n \geq 30$ και η διασπορά του πληθυσμού να είναι γνωστή ή άγνωστη).

Οι έλεγχοι υποθέσεων στην προκειμένη περίπτωση παίρνουν τις ακόλουθες μορφές:

$H_0: \mu = \mu^*$
$H_1: \mu \neq \mu^*$ δίπλευρος έλεγχος.

$H_0: \mu = \mu^*$
$H_1: \mu > \mu^*$ μονόπλευρος έλεγχος (δεξιά)

$H_0: \mu = \mu^*$
$H_1: \mu < \mu^*$ μονόπλευρος έλεγχος (αριστερά)

Ο σημειακός εκτιμητής του μ είναι το $\bar{x}$. Όταν το δείγμα προέρχεται από κανονικό πληθυσμό N(μ, σ^2) ανεξάρτητα από το μέγεθός του έχουμε ότι η $\bar{X}$ ακολουθεί κανονική κατανομή N(μ, $\frac{\sigma^2}{n}$), το ίδιο ισχύει και στην περίπτωση που το δείγμα δεν προέρχεται από κανονικό πληθυσμό αλλά το μέγεθός του είναι **n$\geq$30.**

Το στατιστικό για τον έλεγχο είναι το:

$$Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \qquad ή \qquad Z = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$$

που ακολουθεί τυπική κανονική κατανομή.

Στην πράξη παίρνουμε:

$$Z = \frac{\bar{x}-\mu^*}{\frac{\sigma}{\sqrt{n}}} \qquad ή \qquad Z = \frac{\bar{x}-\mu^*}{\frac{s}{\sqrt{n}}}$$

Οι απορριπτικές περιοχές για τους προαναφερόμενους ελέγχους είναι αντίστοιχα:

Για τον δίπλευρο έλεγχο R = $\{|z| > z_{\frac{a}{2}}\}$,

για τον μονόπλευρο έλεγχο (δεξιά) R = $\{z > z_a\}$ και

για τον μονόπλευρο έλεγχο (αριστερά) R = $\{z < -z_a\}$

5.3.1.2. Έλεγχος υπόθεσης για τη μέση τιμή του πληθυσμού (όταν n<30 και η διασπορά του πληθυσμού να είναι άγνωστη).

Όταν το μέγεθος του δείγματος είναι μικρό, η διακύμανση του δείγματος άγνωστη και ο πληθυσμός από όπου προέρχεται το δείγμα κανονικός τότε η μεταβλητή:

$t = \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}}$ ακολουθεί την t κατανομή με n -1 βαθμούς ελευθερίας.

Η μεταβλητή αυτή παίρνεται ως το στατιστικό ελέγχου για τις παρακάτω υποθέσεις και οι τιμές της στην πράξη είναι:

$$t = \frac{\bar{x}-\mu^*}{\frac{s}{\sqrt{n}}}$$

Οι έλεγχοι υποθέσεων:

H_0: μ = μ^*
H_1: μ $\neq$ μ^* δίπλευρος έλεγχος.

H_0: μ = μ^*
H_1: μ > μ^* μονόπλευρος έλεγχος (δεξιά)

H_0: μ = μ^*
H_1: μ < μ^* μονόπλευρος έλεγχος (αριστερά)

(μ^* είναι η τιμή που υποθέτουμε ότι μπορεί να πάρει το μ)

έχουν ως απορριπτικές περιοχές αντίστοιχα:

Για δίπλευρο έλεγχο R = $\{|t| > t_{n-1\,;\,\frac{a}{2}}\}$,
Για μονόπλευρο έλεγχο (δεξιά) R = $\{t > t_{n-1;\,a}\}$ και

Για μονόπλευρο έλεγχο (αριστερά) R = $\{t < -t_{n-1;\,a}\}$

Παραδείγματα:

1. Παίρνουμε ένα δείγμα ημερήσιων αμοιβών n = 40 εργατών με μέση ημερήσια αμοιβή $\bar{x}$ = 28 ευρώ. Θεωρούμε ότι οι ημερήσιες αμοιβές των εργατών κατανέμονται κανονικώς $N(29{,}4^2)$. Μπορούμε να πούμε ότι η μέση ημερήσια αμοιβή στον πληθυσμό είναι μικρότερη του 29.

Έχουμε τον έλεγχο της μηδενικής υπόθεσης: (α = 0,05)

H_0: μ = 29
H_1: μ < 29 μονόπλευρος έλεγχος (αριστερά)

Η απορριπτική περιοχή είναι:

R = {z < - z_a} με $Z = \dfrac{\bar{X}-\mu}{\dfrac{\sigma}{\sqrt{n}}} = \dfrac{28-29}{\dfrac{4}{\sqrt{40}}}$ = -1,58

Έχουμε z_a = $z_{0,05}$ = 1,64 και εφόσον - z_a < Z = -1,58 δεν απορρίπτουμε την H_0

2. Παίρνουμε ένα δείγμα ημερήσιων αμοιβών n = 50 εργατών με μέση ημερήσια αμοιβή $\bar{x}$ = 38 ευρώ και δειγματική διακύμανση 6^2. Μπορούμε να πούμε ότι η μέση ημερήσια αμοιβή στον πληθυσμό είναι μεγαλύτερη του 39.

Έχουμε τον έλεγχο της μηδενικής υπόθεσης: (α=0,05)

H_0: μ = 39
H_1: μ > 39 μονόπλευρος έλεγχος (δεξιά)

Η απορριπτική περιοχή είναι:

R = {z > z_a} με $Z = \dfrac{\bar{X}-\mu}{\dfrac{\sigma}{\sqrt{n}}} = \dfrac{38-39}{\dfrac{6}{\sqrt{50}}}$ = -1,18

Έχουμε z_a = $z_{0,05}$ = 1,64 και εφόσον z_a = $z_{0,05}$ = 1,64 > Z = -1,18

δεν απορρίπτουμε την H_0.

3. Παίρνουμε ένα δείγμα ημερήσιων αμοιβών n = 20 εργατών με μέση ημερήσια αμοιβή $\bar{x}$ = 33 ευρώ και δειγματική διακύμανση 6^2. Μπορούμε να πούμε ότι η μέση ημερήσια αμοιβή στον πληθυσμό διαφέρει του 30.

Έχουμε τον έλεγχο της μηδενικής υπόθεσης: (α=0,05)

H_0: μ = 30
H_1: μ $\neq$ 30   δίπλευρος έλεγχος

Η απορριπτική περιοχή είναι:

R = $\{|t| > t_{n-1\ ;\ \frac{a}{2}}\}$, με $t = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{33-30}{\frac{6}{\sqrt{20}}} = 2,24$

Έχουμε $t_{n-1\ ;\ \frac{a}{2}} = t_{19;0,025} = 2,093$ και εφόσον t = 2,24 > $t_{19;0,025} = 2,093$

απορρίπτουμε την H_0.

5.3.2. Έλεγχος υπόθεσης για τη διαφορά των μέσων τιμών μ_1, μ_2 δύο πληθυσμών

5.3.2.1. Έλεγχος υπόθεσης για τη διαφορά των μέσων τιμών δύο πληθυσμών (δείγματα μεγάλα ανεξάρτητα, διασπορές γνωστές ή άγνωστες)

Από τις κατανομές των στατιστικών ενός δείγματος γνωρίζουμε ότι:
Αν πάρουμε δύο δείγματα μεγάλα μεγέθους ν και κ με μέσες τιμές $\bar{x}$ και $\bar{y}$ και διακυμάνσεις s_1^2, s_2^2 αντίστοιχα τότε έχουμε ότι η μεταβλητή:

$$Z = \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{\nu}+\frac{\sigma_2^2}{\kappa}}} \text{ ακολουθεί } N(0,1).$$

Εφόσον οι διακυμάνσεις είναι άγνωστες αντικαθίστανται από τις δειγματικές διακυμάνσεις s_1^2, s_2^2 και έχουμε ότι η μεταβλητή:

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{\nu} + \frac{s_2^2}{\kappa}}}$$ ακολουθεί N(0,1).

Οι έλεγχοι υπόθεσης για τη <u>διαφορά των μέσων τιμών δύο πληθυσμών</u> **(δείγματα μεγάλα ανεξάρτητα, διασπορές γνωστές ή άγνωστες)** είναι οι ακόλουθοι:

$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 \neq \mu_2$ δίπλευρος έλεγχος.

$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 > \mu_2$ μονόπλευρος έλεγχος (δεξιά)

$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 < \mu_2$ μονόπλευρος έλεγχος (αριστερά)

Χρησιμοποιείται ως στατιστικό ελέγχου το:

$$Z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{\nu} + \frac{\sigma_2^2}{\kappa}}} \quad \text{ή} \quad Z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{\nu} + \frac{s_2^2}{\kappa}}}$$

και έχουμε ως απορριπτικές περιοχές αντίστοιχα:

Για τον δίπλευρο έλεγχο R = $\{|z| > z_{\frac{a}{2}}\}$,

για τον μονόπλευρο έλεγχο (δεξιά) R = $\{z > z_a\}$ και

για τον μονόπλευρο έλεγχο (αριστερά) R = $\{z < -z_a\}$

Παραδείγματα:

1. Από ένα πληθυσμό ημερήσιων αμοιβών εργατών παίρνουμε ένα δείγμα n = 40 ημερήσιων αμοιβών εργατών με μέση ημερήσια αμοιβή $\bar{x}$ = 33 ευρώ και δειγματική διακύμανση 6^2 και από ένα δεύτερο πληθυσμό ημερήσιων αμοιβών εργατών παίρνουμε ένα δεύτερο δείγμα n = 50 ημερήσιων αμοιβών εργατών με μέση ημερήσια αμοιβή $\bar{y}$ = 30 ευρώ και δειγματική διακύμανση 5^2.

Μπορούμε να πούμε ότι η μέση ημερήσια αμοιβή μ_1 στον πρώτο πληθυσμό διαφέρει της μέσης ημερήσιας αμοιβής μ_2 στον δεύτερο πληθυσμό σε επίπεδο σημαντικότητας (α=0,05)

Ο έλεγχος μηδενικής υπόθεσης για τη <u>διαφορά των μέσων τιμών δύο πληθυσμών</u> είναι ο ακόλουθος:

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 \neq \mu_2$ δίπλευρος έλεγχος.

Η απορριπτική περιοχή είναι:

R = $\{|z| > z_{\frac{a}{2}}\}$,

Για τον έλεγχο της μηδενικής υπόθεσης χρησιμοποιούμε το στατιστικό

$$Z = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_1^2}{\nu}+\frac{s_2^2}{\kappa}}} = \frac{33-30}{\sqrt{\frac{36}{40}+\frac{25}{50}}} = 2{,}54$$

Έχουμε ότι: $z_{\frac{a}{2}} = 1{,}96$

Επειδή Z > $z_{\frac{a}{2}}$ απορρίπτεται η H_0. Άρα σε επίπεδο σημαντικότητας α = 5% δεν δεχόμαστε ότι $\mu_1 = \mu_2$.

2. Από ένα πληθυσμό ημερήσιων αμοιβών εργατών παίρνουμε ένα δείγμα n = 38 ημερήσιων αμοιβών εργατών με μέση ημερήσια αμοιβή $\bar{x}$ = 37 ευρώ. Από ένα δεύτερο πληθυσμό ημερήσιων αμοιβών εργατών παίρνουμε ένα δεύτερο δείγμα n = 45 ημερήσιων αμοιβών εργατών με μέση ημερήσια αμοιβή $\bar{y}$ = 34 ευρώ. Γνωρίζουμε τις πληθυσμιακές διασπορές των ημερήσιων αμοιβών ότι είναι 4^2 και 5^2 αντίστοιχα. Μπορούμε να πούμε ότι η μέση ημερήσια αμοιβή μ_1 στον πρώτο πληθυσμό διαφέρει της μέσης ημερήσιας αμοιβής μ_2 στον δεύτερο πληθυσμό σε επίπεδο σημαντικότητας (α=0,05)

Ο έλεγχος μηδενικής υπόθεσης για τη διαφορά των μέσων τιμών δύο πληθυσμών είναι ο ακόλουθος:

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 \neq \mu_2$ δίπλευρος έλεγχος.

Η απορριπτική περιοχή είναι:

R = $\{|z| > z_{\frac{a}{2}}\}$,

Για τον έλεγχο της μηδενικής υπόθεσης χρησιμοποιούμε το στατιστικό

$$Z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{\nu} + \frac{\sigma_2^2}{\kappa}}} = \frac{37-34}{\sqrt{\frac{16}{38} + \frac{25}{45}}} = 3{,}03$$

Έχουμε ότι: $z_{\frac{a}{2}}$ = 1,96

Επειδή Z > $z_{\frac{a}{2}}$ απορρίπτεται η H_0. Άρα σε επίπεδο σημαντικότητας α=5% δεν δεχόμαστε ότι $\mu_1 = \mu_2$.

5.3.2.2. Έλεγχος υπόθεσης για τη διαφορά των μέσων τιμών δύο πληθυσμών (δείγματα μικρά ανεξάρτητα, διασπορές άγνωστες και ίσες)

Από τις κατανομές των στατιστικών ενός δείγματος γνωρίζουμε ότι: Αν πάρουμε δύο δείγματα μικρά ανεξάρτητα μεγέθους ν και κ με μέσες τιμές $\bar{x}$ και $\bar{y}$ και διακυμάνσεις s_1^2, s_2^2 αντίστοιχα προερχόμενα από κανονικούς πληθυσμούς με άγνωστες αλλά ίσες διακυμάνσεις $\sigma_1^2 = \sigma_2^2 = \sigma^2$ τότε έχουμε ότι η μεταβλητή:

$$t = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{(\nu-1)s_1^2 + (\kappa-1)s_2^2}{\nu+\kappa-2}}\sqrt{\frac{1}{\nu}+\frac{1}{\kappa}}}$$

ακολουθεί t-κατανομή με ν+κ-2 βαθμούς ελευθερίας

Οι έλεγχοι υπόθεσης για τη <u>διαφορά των μέσων τιμών δύο πληθυσμών</u> (δείγματα μικρά ανεξάρτητα, διασπορές άγνωστες και ίσες) είναι οι ακόλουθοι:

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 \neq \mu_2$ δίπλευρος έλεγχος.

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 > \mu_2$ μονόπλευρος έλεγχος (δεξιά)

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 < \mu_2$ μονόπλευρος έλεγχος (αριστερά)

Χρησιμοποιείται ως στατιστικό ελέγχου το:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{(\nu-1)s_1^2 + (\kappa-1)s_2^2}{\nu+\kappa-2}}\sqrt{\frac{1}{\nu}+\frac{1}{\kappa}}}$$

και έχουμε ως απορριπτικές περιοχές αντίστοιχα:

Για δίπλευρο έλεγχο R = $\{|t| > t_{\nu+\kappa-2\,;\,\frac{a}{2}}\}$,

Για μονόπλευρο έλεγχο (δεξιά) R = $\{t > t_{\nu+\kappa-2;\,a}\}$ και

Για μονόπλευρος έλεγχος (αριστερά) R = $\{t < -\,t_{\nu+\kappa-2;\,a}\}$

Παράδειγμα:

Από ένα πληθυσμό στρεμματικών αποδόσεων μιας καλλιέργειας σε μια περιοχή παίρνουμε ένα δείγμα ν =16 στρεμματικών αποδόσεων με μέση τιμή $\bar{x}$ = 370 κιλά/στρεμ. Από ένα δεύτερο πληθυσμό στρεμματικών αποδόσεων της ίδιας καλλιέργειας σε μια άλλη περιοχή παίρνουμε ένα δεύτερο δείγμα κ = 15 στρεμματικών αποδόσεων με μέση τιμή $\bar{y}$ = 340 κιλά/στρεμ. Οι δειγματικές διακυμάνσεις είναι αντίστοιχα 41^2 και 52^2. Μπορούμε να πούμε ότι η μέση στρεμματική απόδοση μ_1 στον πρώτο πληθυσμό διαφέρει της μέσης στρεμματική απόδοση μ_2 στον δεύτερο πληθυσμό σε επίπεδο σημαντικότητας (α = 0,05). Θεωρούμε ότι οι πληθυσμιακές διασπορές είναι άγνωστες και ίσες.

Ο έλεγχος μηδενικής υπόθεσης για τη διαφορά των μέσων τιμών δύο πληθυσμών είναι ο ακόλουθος:

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 \neq \mu_2$ δίπλευρος έλεγχος.

Χρησιμοποιείται το στατιστικό:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{(\nu-1)s_1^2 + (\kappa-1)s_2^2}{\nu+\kappa-2}}\sqrt{\frac{1}{\nu}+\frac{1}{\kappa}}} = \frac{370-340}{\sqrt{\frac{(16-1)1681+(15-1)2704}{29}}\sqrt{\frac{1}{16}+\frac{1}{15}}} = 0,34$$

Η απορριπτική περιοχή είναι:

$R = \{|t| > t_{\nu+\kappa-2\ ;\ \frac{a}{2}}\}$, $t_{\nu+\kappa-2\ ;\ \frac{a}{2}} = t_{29\ ;\ 0,025} = 2,045$

Εφόσον $t < t_{\nu+\kappa-2\ ;\ \frac{a}{2}}$ δεν απορρίπτεται η H_0. Άρα σε επίπεδο σημαντικότητας α = 5% δεχόμαστε ότι $\mu_1 = \mu_2$ και δεν μπορούμε να ισχυριστούμε ότι διαφέρουν μεταξύ τους.

5.3.2.3. Έλεγχος υπόθεσης για τη διαφορά των μέσων τιμών δύο πληθυσμών (δείγματα μικρά ανεξάρτητα, διασπορές άγνωστες και διαφορετικές)

Από τις κατανομές των στατιστικών ενός δείγματος γνωρίζουμε ότι:

Αν πάρουμε δύο δείγματα μικρά ανεξάρτητα μεγέθους ν και κ με μέσες τιμές $\bar{x}$ και $\bar{y}$ και διακυμάνσεις s_1^2, s_2^2 αντίστοιχα προερχόμενα από κανονικούς πληθυσμούς με άγνωστες και διαφορετικές διασπορές $\sigma_1^2 \neq \sigma_2^2$ τότε έχουμε ότι η μεταβλητή:

$$t = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{\nu} + \frac{s_2^2}{\kappa}}}$$

ακολουθεί t – κατανομή με βαθμούς ελευθερίας λ = 2(ν-1) όταν ν = κ

και $\lambda = \frac{\left(\frac{s_1^2}{\nu} + \frac{s_2^2}{\kappa}\right)^2}{\frac{\left(\frac{s_1^2}{\nu}\right)^2}{\nu - 1} + \frac{\left(\frac{s_2^2}{\kappa}\right)^2}{\kappa - 1}}$ (στρογγυλεμένο στον πλησιέστερο ακέραιο)

όταν $\nu \neq \kappa$

Οι έλεγχοι υπόθεσης για τη διαφορά των μέσων τιμών δύο πληθυσμών (δείγματα μικρά ανεξάρτητα, διασπορές άγνωστες και διαφορετικές) είναι οι ακόλουθοι:

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 \neq \mu_2$ δίπλευρος έλεγχος.

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 > \mu_2$ μονόπλευρος έλεγχος (δεξιά)

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 < \mu_2$ μονόπλευρος έλεγχος (αριστερά)

Χρησιμοποιείται ως στατιστικό ελέγχου το:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{\nu} + \frac{s_2^2}{\kappa}}}$$

και έχουν ως απορριπτικές περιοχές αντίστοιχα:

Για δίπλευρο έλεγχο R = $\{|t| > t_{\lambda;\, \frac{a}{2}}\}$,

Για μονόπλευρο έλεγχο (δεξιά) R = $\{t > t_{\lambda;\, a}\}$ και

Για μονόπλευρο έλεγχο (αριστερά) R = $\{t < -t_{\lambda;\, a}\}$

Παραδείγματα:

1. Από ένα πληθυσμό στρεμματικών αποδόσεων μιας καλλιέργειας σε μια περιοχή παίρνουμε ένα δείγμα ν=16 στρεμματικών αποδόσεων με μέση τιμή $\bar{x}$ = 370κιλά/στρεμ. Από ένα δεύτερο πληθυσμό στρεμματικών αποδόσεων της ίδιας καλλιέργειας σε μια άλλη περιοχή παίρνουμε ένα δεύτερο δείγμα κ=16 στρεμματικών αποδόσεων με μέση τιμή $\bar{y}$ = 340 κιλά/στρεμ. Οι δειγματικές διακυμάνσεις είναι αντίστοιχα 41^2 και 52^2. Μπορούμε να πούμε ότι η μέση στρεμματική απόδοση μ_1 στον πρώτο πληθυσμό διαφέρει της μέσης στρεμματική απόδοση μ_2 στον δεύτερο πληθυσμό σε επίπεδο σημαντικότητας (α=0,05). Θεωρούμε ότι οι πληθυσμιακές διασπορές είναι άγνωστες και διαφορετικές.

Ο έλεγχος μηδενικής υπόθεσης για τη διαφορά των μέσων τιμών δύο πληθυσμών είναι ο ακόλουθος:

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 \neq \mu_2$ δίπλευρος έλεγχος.

Χρησιμοποιείται το στατιστικό:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{\nu} + \frac{s_2^2}{\kappa}}} = \frac{370 - 340}{\sqrt{\frac{41^2}{16} + \frac{52^2}{16}}} = 1{,}81$$

η απορριπτική περιοχή είναι:

R = $\{|t| > t_{\lambda\,;\,\frac{a}{2}}\}$, $t_{\lambda\,;\,\frac{a}{2}} = t_{30\,;\,0,025}$ = 2,042, λ = 2(ν-1) = 30

Εφόσον t < $t_{\lambda\,;\,\frac{a}{2}}$ δεν απορρίπτεται η H_0. Άρα σε επίπεδο σημαντικότητας α=5% δεχόμαστε ότι $\mu_1 = \mu_2$ και δεν μπορούμε να ισχυριστούμε ότι διαφέρουν μεταξύ τους.

2. Από ένα πληθυσμό στρεμματικών αποδόσεων μιας καλλιέργειας σε μια περιοχή παίρνουμε ένα δείγμα ν=10 στρεμματικών αποδόσεων με μέση τιμή $\bar{x}$ = 370κιλά/στρεμ. Από ένα δεύτερο πληθυσμό στρεμματικών αποδόσεων της ίδιας καλλιέργειας σε μια άλλη περιοχή παίρνουμε ένα δεύτερο δείγμα κ=16 στρεμματικών αποδόσεων με μέση τιμή $\bar{y}$ = 340 κιλά/στρεμ. Οι δειγματικές διακυμάνσεις είναι αντίστοιχα 41^2 και 52^2. Μπορούμε να πούμε ότι η μέση στρεμματική απόδοση μ_1 στον πρώτο πληθυσμό διαφέρει της μέσης στρεμματική απόδοση μ_2 στον δεύτερο πληθυσμό σε επίπεδο σημαντικότητας (α=0,05). Θεωρούμε ότι οι πληθυσμιακές διασπορές είναι άγνωστες και διαφορετικές.

Ο έλεγχος μηδενικής υπόθεσης για τη διαφορά των μέσων τιμών δύο πληθυσμών είναι ο ακόλουθος:

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 \neq \mu_2$ δίπλευρος έλεγχος.

Χρησιμοποιείται το στατιστικό:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{\nu} + \frac{s_2^2}{\kappa}}} = \frac{370 - 340}{\sqrt{\frac{41^2}{10} + \frac{52^2}{16}}} = 1,63$$

η απορριπτική περιοχή είναι:

R = $\{|t| > t_{\lambda\ ;\ \frac{a}{2}}\}$, $t_{\lambda\ ;\ \frac{a}{2}} = t_{23\ ;\ 0,025}$ = 2,069

$$\lambda = \frac{\left(\frac{s_1^2}{\nu} + \frac{s_2^2}{\kappa}\right)^2}{\frac{\left(\frac{s_1^2}{\nu}\right)^2}{\nu - 1} + \frac{\left(\frac{s_2^2}{\kappa}\right)^2}{\kappa - 1}} = 23$$

Εφόσον t < $t_{\lambda\ ;\ \frac{a}{2}}$ δεν απορρίπτεται η H_0.

Άρα σε επίπεδο σημαντικότητας α = 5% δεχόμαστε ότι $\mu_1 = \mu_2$ και δεν μπορούμε να ισχυριστούμε ότι διαφέρουν μεταξύ τους.

5.3.2.4. Έλεγχος υπόθεσης για τη διαφορά των μέσων τιμών δύο πληθυσμών (δείγματα μικρά εξαρτημένα)

Στην περίπτωση που έχουμε **δείγματα μικρά εξαρτημένα** που προέρχονται από μετρήσεις της ίδιας ομάδας σε δυο διαφορετικές χρονικές στιγμές (ζευγαρωτές παρατηρήσεις) ορίσουμε x_i και y_i, i=1,2,….,n τις παρατηρήσεις στα δύο δείγματα και δημιουργούμε τις αντίστοιχες διαφορές $z_i = x_i - y_i$, i=1,2,….,n.

Οι παρατηρήσεις z_i ακολουθούν την t – κατανομή με n - 1 βαθμούς ελευθερίας.

Οι έλεγχοι υπόθεσης για τη διαφορά των μέσων τιμών δύο πληθυσμών **(από όπου προέρχονται τα μικρά δείγματα τα εξαρτημένα)** είναι οι ακόλουθοι:

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 \neq \mu_2$ δίπλευρος έλεγχος.

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 > \mu_2$ μονόπλευρος έλεγχος (δεξιά)

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 < \mu_2$ μονόπλευρος έλεγχος (αριστερά)

χρησιμοποιούν ως στατιστικό ελέγχου το:

$t = \dfrac{\bar{Z}}{\frac{s_Z}{\sqrt{n}}}$ που ακολουθεί t – κατανομή με n -1 βαθμούς ελευθερίας.

και έχουν ως απορριπτικές περιοχές αντίστοιχα:

Για δίπλευρο έλεγχο R = $\{|t| > t_{n-1\,;\,\frac{a}{2}}\}$,

Για μονόπλευρο έλεγχο (δεξιά) R = $\{t > t_{n-1;\,a}\}$ και

Για μονόπλευρο έλεγχο (αριστερά) R = $\{t < -t_{n-1;\,a}\}$

Παράδειγμα:

Έχουμε τις παρακάτω ζευγαρωτές παρατηρήσεις:

X:	4	5	6	4,2	5,2	5,3	6,4	4,8	5,3	5
Y:	4,9	4,8	5,7	5	6	5.2	6,5	5,9	4,8	5,7

Να ελεγχθεί η μηδενική υπόθεση:

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 < \mu_2$ μονόπλευρος έλεγχος (αριστερά),

σε επίπεδο σημαντικότητας α=0,05

Έχουμε Z: - 0,9 0,2 0,3 -0,8 -0,8 0,1 -0,1 -1,1 0,5 -0,7

και $\bar{Z}$ = - 0,33 , S_Z = 0,72 , $t_{9;0,05}$ = 1,833

$$t = \frac{\bar{Z}}{\frac{s_Z}{\sqrt{n}}} = \frac{-0,33}{\frac{0,72}{\sqrt{10}}} = -1,43$$

Η απορριπτική περιοχή είναι: R = $\{t < -t_{n-1;\,a}\}$

Επειδή όμως $t > -t_{n-1;\,a}$ δεν απορρίπτεται η $H_0: \mu_1 = \mu_2$

5.4. Έλεγχος υπόθεσης για τη διασπορά ενός πληθυσμού

Στην θεωρία διατυπώθηκε ότι η μεταβλητή:

$X^2 = \frac{(n-1)s^2}{\sigma^2}$ ακολουθεί την X^2 κατανομή με n-1 βαθμούς.

Ο πληθυσμός από όπου πάρθηκε το δείγμα θεωρείται κανονικός με πληθυσμιακή διασπορά σ^2 και ελέγχεται αν η σ^2 παίρνει την τιμή σ_0^2. Όπου s^2 η δειγματική διακύμανση.

Οι έλεγχοι υποθέσεων στην προκειμένη περίπτωση παίρνουν τις ακόλουθες μορφές:

$H_0: \sigma^2 = \sigma_0^2$
$H_1: \sigma^2 \neq \sigma_0^2$ δίπλευρος έλεγχος.

$H_0: \sigma^2 = \sigma_0^2$
$H_1: \sigma^2 > \sigma_0^2$ μονόπλευρος έλεγχος (δεξιά)

$H_0: \sigma^2 = \sigma_0^2$
$H_1: \sigma^2 < \sigma_0^2$ μονόπλευρος έλεγχος (αριστερά)

Το στατιστικό ελέγχου που χρησιμοποιείται είναι το:

$X^2 = \frac{(n-1)s^2}{\sigma_0^{\,2}}$ που ακολουθεί X^2 κατανομή με n-1 βαθμούς

Οι απορριπτικές περιοχές για τους προαναφερόμενους ελέγχους είναι αντίστοιχα:

Για τον δίπλευρο έλεγχο R = { $X^2 > \chi^2_{n-1;\frac{a}{2}}$ ή $X^2 < \chi^2_{n-1;1-\frac{a}{2}}$ },

για τον μονόπλευρο έλεγχο (δεξιά) R = {$X^2 > \chi^2_{n-1;a}$} και

για τον μονόπλευρο έλεγχο (αριστερά) R = {$X^2 < \chi^2_{n-1;1-a}$}

Παράδειγμα:

Ένα δείγμα από 51 αγρότες παρουσιάζει διακύμανση των ετήσιων εισοδημάτων τους s^2=156 ευρώ^2. Η πληθυσμιακή διασπορά είναι η σ^2.

Να ελεγχθεί η υπόθεση:

H_0: σ^2 = 140
H_1: $\sigma^2 \neq$ 140

Σε επίπεδο σημαντικότητας α=5%

Έχουμε το στατιστικό:

$$X^2 = \frac{(n-1)s^2}{\sigma_0^{\ 2}} = \frac{(51-1)156}{140} = 55{,}71$$

[$\chi^2_{n-1;\frac{a}{2}}$ = $X^2_{50;0,025}$ = 71,42 , $\chi^2_{n-1;1-\frac{a}{2}}$ = $X^2_{50;0,975}$ = 32,36]

Για τον δίπλευρο έλεγχο η απορριπτική περιοχή είναι:

R = {$X^2 > \chi^2_{n-1;\frac{a}{2}}$ ή $X^2 < \chi^2_{n-1;1-\frac{a}{2}}$ },

Εφόσον $\chi^2_{n-1;1-\frac{a}{2}}$ **<** $\boldsymbol{X^2}$ **<** $\chi^2_{n-1;\frac{a}{2}}$ η υπόθεση H_0: σ^2 = 140 δεν απορρίπτεται.

5.5. Έλεγχος υπόθεσης για το λόγο των διασπορών δύο πληθυσμών

Σύμφωνα με την θεωρία αν πάρουμε δύο δείγματα ανεξάρτητα μεγέθους n και m αντίστοιχα με s_1^2, s_2^2 τις δειγματικές διακυμάνσεις από δύο κανονικούς πληθυσμούς αντίστοιχα με σ_1^2 , σ_2^2 τις πληθυσμιακές διασπορές τότε ισχύει ότι η μεταβλητή:

F = $\frac{s_1^2}{s_2^2}\frac{\sigma_2^2}{\sigma_1^2}$ ακολουθεί την F κατανομή με n -1 και m -1 βαθμούς ελευθερίας

Οι έλεγχοι υποθέσεων στην προκειμένη περίπτωση παίρνουν τις ακόλουθες μορφές:

H_0: $\frac{\sigma_1^2}{\sigma_2^2} = 1$

H_1: $\frac{\sigma_1^2}{\sigma_2^2} \neq 1$ δίπλευρος έλεγχος

H_0: $\frac{\sigma_1^2}{\sigma_2^2} = 1$

H_1: $\frac{\sigma_1^2}{\sigma_2^2} > 1$ μονόπλευρος έλεγχος

Το στατιστικό ελέγχου που χρησιμοποιείται είναι το:

$$F = \frac{s_1^2}{s_2^2} \text{ (όταν } s_1^2 > s_2^2 \text{)}$$

που ακολουθεί την F κατανομή με n-1 και m-1 βαθμούς ελευθερίας

ή $$F = \frac{s_2^2}{s_1^2} \text{ (όταν } s_2^2 > s_1^2 \text{)}$$

που ακολουθεί την F κατανομή με m-1 και n-1 βαθμούς ελευθερίας

Οι απορριπτικές περιοχές για τους προαναφερόμενους ελέγχους είναι αντίστοιχα:

Για τον δίπλευρο έλεγχο R = $\{F > F_{n-1,m-1;\frac{a}{2}}\}$,

για τον μονόπλευρο έλεγχο R = $\{F > F_{n-1,m-1;a}\}$

Παραδείγματα:

1. Πήραμε δύο δείγματα αγροτών από δύο διοικητικές περιφέρειες της Ελλάδας και εξετάσαμε τα ετήσια εισοδήματά τους. Είχαμε τα ακόλουθα δεδομένα:

n= 31, S_1^2 = 220 ευρώ^2 και m = 41, S_2^2 = 200 ευρώ^2

Να ελεγχθεί η υπόθεση:

H_0: $\frac{\sigma_1^2}{\sigma_2^2}$ = 1

H_1: $\frac{\sigma_1^2}{\sigma_2^2}$ > 1 σε επίπεδο σημαντικότητας α = 1%

Όπου $σ_1^2$, $σ_2^2$ οι πραγματικές διασπορές των ετήσιων εισοδημάτων των αγροτικών πληθυσμών στις δύο διοικητικές περιφέρειες της Ελλάδας.

Το στατιστικό ελέγχου είναι:

F = $\frac{s_1^2}{s_2^2}$ = 1,1 ($S_1^2 > S_2^2$) έχουμε $F_{30,40;0,01}$ = 2,2

Εφόσον F < $F_{30,40;0,01}$ = 2,2 άρα η H_0: $\frac{\sigma_1^2}{\sigma_2^2}$ = 1 δεν απορρίπτεται.

2. Πήραμε δύο δείγματα αγροτών από δύο διοικητικές περιφέρειες της Ελλάδας και εξετάσαμε τα ετήσια εισοδήματά τους. Είχαμε τα ακόλουθα δεδομένα:

n = 31, S_1^2 = 220 ευρώ^2 και m = 41, S_2^2 = 200 ευρώ^2

Να ελεγχθεί η υπόθεση:

H_0: $\frac{\sigma_1^2}{\sigma_2^2}$ = 1

H_1: $\frac{\sigma_1^2}{\sigma_2^2}$ $\neq$ 1 σε επίπεδο σημαντικότητας α = 2%

Όπου $σ_1^2$, $σ_2^2$ οι πραγματικές διασπορές των ετήσιων εισοδημάτων των αγροτικών πληθυσμών στις δύο διοικητικές περιφέρειες της Ελλάδας.

Το στατιστικό ελέγχου είναι:

F = $\frac{s_1^2}{s_2^2}$ = 1,1 $(S_1^2 > S_2^2)$ έχουμε $F_{30,40;0,01}$ = 2,2 , όπου $\frac{a}{2}$ = 0,01

Εφόσον F < $F_{30,40;0,01}$ = 2,2 άρα η H_0: $\frac{\sigma_1^2}{\sigma_2^2}$ = 1 δεν απορρίπτεται.

5.6. Έλεγχος υπόθεσης για το ποσοστό των στοιχείων ενός πληθυσμού

Αν εκτιμήσουμε το ποσοστό (αναλογία) p = $\frac{x}{n}$ των στοιχείων στο μεγάλο δείγμα μεγέθους n που έχουν κάποιο συγκεκριμένο χαρακτηριστικό, τότε σύμφωνα με την

θεωρία το δειγματικό ποσοστό κατανέμεται κανονικά:

Δηλαδή η δειγματική μεταβλητή P_Δ, που παράγεται από τις τιμές p, ακολουθεί την κανονική κατανομή:

$$N(P_{\pi\lambda\eta\theta}, \frac{p_{\pi\lambda\eta\theta}(1-P_{\pi\lambda\eta\theta})}{n})$$

Όπου $P_{\pi\lambda\eta\theta}$ το πραγματικό ποσοστό στον πληθυσμό.

Συνεπώς έχουμε ότι η μεταβλητή:

$$Z = \frac{P_\Delta - P_{\pi\lambda\eta\theta}}{\sqrt{\frac{P_{\pi\lambda\eta\theta}(1-P_{\pi\lambda\eta\theta})}{n}}} \quad \text{ακολουθεί την N(0,1).}$$

Οι έλεγχοι υποθέσεων στην προκειμένη περίπτωση παίρνουν τις ακόλουθες μορφές:

H_0: $P_{\pi\lambda\eta\theta} = p^*$
H_1: $P_{\pi\lambda\eta\theta} \neq p^*$ δίπλευρος έλεγχος.

H_0: $P_{\pi\lambda\eta\theta} = p^*$
H_1: $P_{\pi\lambda\eta\theta} > p^*$ μονόπλευρος έλεγχος (δεξιά)

H_0: $P_{\pi\lambda\eta\theta} = p^*$
H_1: $P_{\pi\lambda\eta\theta} < p^*$ μονόπλευρος έλεγχος (αριστερά)

Το στατιστικό ελέγχου που χρησιμοποιείται είναι το:

$Z = \dfrac{p_\Delta - p^*}{\sqrt{\dfrac{p^*(1-p^*)}{n}}}$ που ακολουθεί την τυπική κανονική κατανομή.

Οι απορριπτικές περιοχές για τους προαναφερόμενους ελέγχους είναι αντίστοιχα:

Για τον δίπλευρο έλεγχο R = $\{|z| > z_{\frac{a}{2}}\}$,

για τον μονόπλευρο έλεγχο (δεξιά) R = $\{z > z_a\}$ και

για τον μονόπλευρο έλεγχο (αριστερά) R = $\{z < - z_a\}$

Παράδειγμα:

Ο αριθμός των ελαττωματικών προϊόντων μιας μεταποιητικής βιομηχανίας αγροτικών προϊόντων που εντοπίσθηκαν σ' ένα δείγμα 300 προϊόντων είναι 24 προϊόντα.

Να ελεγχθεί σε επίπεδο σημαντικότητας α = 2% η μηδενική υπόθεση ότι το πραγματικό ποσοστό $P_{\pi\lambda\eta\theta}$ των ελαττωματικών προϊόντων της συνολικής παραγωγής είναι μεγαλύτερο του 10%.

H_0: $P_{\pi\lambda\eta\theta}$ = 0,10
H_1: $P_{\pi\lambda\eta\theta}$ > 0,10 μονόπλευρος έλεγχος (δεξιά)

Το στατιστικό ελέγχου που χρησιμοποιείται είναι το:

$Z = \dfrac{p_\Delta - p^*}{\sqrt{\dfrac{p^*(1-p^*)}{n}}}$ που ακολουθεί την τυπική κανονική κατανομή.

Η απορριπτική περιοχή είναι: R = $\{z > z_a\}$

Έχουμε n = 300 , m = 24 $\rightarrow$ $\rho_\Delta = \dfrac{24}{300} = 8\% = 0{,}08.$

$$\left.\alpha = 1 - 0{,}98 = 0{,}02 \right\} \rightarrow z_a = \phi^{-1}(1-a) = \phi^{-1}(0{,}98) = 2{,}06.$$

$$Z = \frac{p_\Delta - p^*}{\sqrt{\frac{p^*(1-p^*)}{n}}} = \frac{0{,}08 - 0{,}10}{\sqrt{\frac{0{,}10 * 0{,}90}{300}}} = -1$$

Εφόσον $z < z_a$ δεν απορρίπτεται η H_0: $P_{πληθ}$ = 0,10 , άρα δεν μπορούμε να ισχυριστούμε ότι το πραγματικό ποσοστό $P_{πληθ}$ των ελαττωματικών προϊόντων της συνολικής παραγωγής είναι μεγαλύτερο του10%.

5.7. Έλεγχος υπόθεσης για τη διαφορά των ποσοστών των στοιχείων δύο πληθυσμών

Έστω $p_1 = \frac{x}{n}$ και $p_2 = \frac{y}{l}$ είναι οι δειγματικές αναλογίες των στοιχείων, που έχουν κάποιο συγκεκριμένο χαρακτηριστικό, δύο δειγμάτων μεγέθους n, l από δύο πληθυσμούς, όπου οι πραγματικές αναλογίες είναι αντίστοιχα $P_{1\ πληθ}$, $P_{2\ πληθ}$.

Τότε έχουμε ότι οι διαφορές των δειγματικών μεταβλητών $P_{1\ \Delta}$ - $P_{2\ \Delta}$ ακολουθούν κανονική κατανομή:

$$N(P_{1\ πληθ} - P_{2\ πληθ},\ \frac{P_{1πληθ}(1-P_{1πληθ})}{n} + \frac{P_{2πληθ}(1-P_{2πληθ})}{l})$$

Οι έλεγχοι υποθέσεων στην προκειμένη περίπτωση παίρνουν τις ακόλουθες μορφές:

H_0: $P_{1\ πληθ} = P_{2\ πληθ}$
H_1: $P_{1\ πληθ} \neq P_{2\ πληθ}$ δίπλευρος έλεγχος.

H_0: $P_{1\ πληθ} = P_{2\ πληθ}$
H_1: $P_{1\ πληθ} > P_{2\ πληθ}$ μονόπλευρος έλεγχος (δεξιά)

H_0: $P_{1\ πληθ} = P_{2\ πληθ}$
H_1: $P_{1\ πληθ} < P_{2\ πληθ}$ μονόπλευρος έλεγχος (αριστερά)

Το στατιστικό ελέγχου που χρησιμοποιείται είναι η μεταβλητή:

Z = $\dfrac{p_{1\Delta} - p_{2\Delta}}{\sqrt{pq(\frac{1}{n}+\frac{1}{l})}}$ η οποία ακολουθεί την τυπική κανονική κατανομή

όπου p = $\dfrac{x+y}{n+l}$ εκτιμά την **P = P**$_{1\ \pi\lambda\eta\theta}$**= P**$_{2\ \pi\lambda\eta\theta}$ και q = 1- p

Οι απορριπτικές περιοχές για τους προαναφερόμενους ελέγχους είναι αντίστοιχα:

Για τον δίπλευρο έλεγχο R = {$|z| >$ z$_{\frac{a}{2}}$ },

για τον μονόπλευρο έλεγχο (δεξιά) R = {z > z$_a$} και

για τον μονόπλευρο έλεγχο (αριστερά) R = {z < - z$_a$}

Παράδειγμα:

Παίρνουμε δύο μεγάλα δείγματα μιας ποικιλίας ενός φυτού από δύο διαφορετικούς πληθυσμούς και εξετάζουμε πόσα από αυτά ασθένησαν μέσα σε ένα συγκεκριμένο χρονικό διάστημα. Τα δεδομένα που προέκυψαν ήταν: Από τα n=120 του πρώτου δείγματος ασθένησαν τα 12 και από τα m =130 του δεύτερου δείγματος ασθένησαν τα 18.

Να ελεγχθεί η μηδενική υπόθεση

H$_0$: P$_{1\ \pi\lambda\eta\theta}$= P$_{2\ \pi\lambda\eta\theta}$
H$_1$: P$_{1\ \pi\lambda\eta\theta}$ $\neq$ P$_{2\ \pi\lambda\eta\theta}$

σε επίπεδο σημαντικότητας α = 5%.

Έχουμε ότι:

p$_{1\ \Delta}$= 12/120 = 0,1, p$_{2\ \Delta}$ = 18/130 = 0,14 και p$_{1\ \Delta}$- p$_{2\ \Delta}$ = - 0,04

z$_{\frac{a}{2}}$= Z$_{0,025}$= 1,96

p = $\dfrac{x+y}{n+l}$= 0,12 , q = 0,88

$$Z = \frac{p_{1\Delta} - p_{2\Delta}}{\sqrt{pq(\frac{1}{n} + \frac{1}{l})}} = \frac{0{,}1 - 0{,}14}{\sqrt{0{,}12 * 0{,}88(\frac{1}{120} + \frac{1}{130})}} = -0{,}98$$

Η απορριπτική περιοχή είναι: R = $\{|z| > z_{\frac{a}{2}}\}$

Εφόσον $|z| < z_{\frac{a}{2}}$ η H_0: $P_{1\ \pi\lambda\eta\theta} = P_{2\ \pi\lambda\eta\theta}$ δεν απορρίπτεται.

5.8. Έλεγχος υπόθεσης και διάστημα εμπιστοσύνης

Εάν υπολογίσουμε το 100(1-α) % διάστημα εμπιστοσύνης μιας παραμέτρου Θ (μ, σ, p) ενός πληθυσμού τότε για τον έλεγχο μηδενικής υπόθεσης H_0: Θ = θ (η εκτιμήτριά της από ένα δείγμα) μπορούμε να πούμε ότι:

Αν το θ ανήκει στο 100(1-α)% διάστημα εμπιστοσύνης της παραμέτρου Θ τότε η :

H_0: Θ = θ γίνεται δεκτή, αλλιώς απορρίπτεται

Η εναλλακτική υπόθεση είναι πάντα η H_1: $\Theta \neq \theta$

Κεφάλαιο 6ο

Απλή Παλινδρόμηση

6.1. Παλινδρόμηση

Η απλή παλινδρόμηση αποδίδει τον υπολογισμό του εξαρτημένου παράγοντα από τον ανεξάρτητο Η παλινδρόμηση χαρακτηρίζεται και ως **εξίσωση συμμεταβολής** και μπορεί να πάρει τις εξής μορφές:

- **Γραμμική**: Εκφράζεται από μία εξίσωση α΄ βαθμού και αποδίδεται από μία ευθεία.

- **Μη γραμμική ή καμπυλόγραμμη**: Εκφράζεται από μία εξίσωση μη γραμμική η οποία μπορεί να έχει την μορφή ενός πολυωνύμου μιας λογαριθμικής ή μιας εκθετικής εξίσωσης.

Επιπλέον, η γραμμική παλινδρόμηση διακρίνεται σε **απλή** και **πολλαπλή παλινδρόμηση**.

Απλή παλινδρόμηση: Η εξαρτημένη μεταβλητή είναι συνάρτηση μόνο μιας ερμηνευτικής μεταβλητής $Y = aX + b$.

Πολλαπλή παλινδρόμηση: Η εξαρτημένη μεταβλητή είναι συνάρτηση περισσότερων του ενός ερμηνευτικών μεταβλητών (π.χ. κ - ερμηνευτικών μεταβλητών) $Y = a_1X_1 + a_2X_2 + \ldots\ldots + a_kX_k$

Υπάρχουν μορφές που μετασχηματίζονται σε πολλαπλές ή απλές παλινδρομήσεις όπως:

$Y = a + BX + cX^2$ με $(X^2 = Y)$

$Y = a + BX + cX^2 + dX^3$ με $(X^2 = Y, X^3 = Z)$

$Y = a + BX + cX^2 + dX^3 + eX^4$ με $(X^2 = Y, X^3 = Z, X^4 = F)$

$Y = aX^b$ ή $LogY = Log\alpha + bLogX$ με $0 < a < 1, X > 0$

Παρατήρηση:

Για να προσδιοριστούν οι παραπάνω εξισώσεις, όπως και άλλων μορφών, πρέπει να προσδιοριστούν οι συντελεστές α, b, c, d, e,…, με βασικό μέλημα ότι η κάθε εξίσωση να έχει την καλύτερη δυνατή προσαρμογή με τα δεδομένα του προβλήματος.

6.2. Απλή Παλινδρόμηση

Η μελέτη της ευθύγραμμης παλινδρόμησης $Y = \alpha + \beta x$ εστιάζεται στην εκτίμηση των συντελεστών α και β με τη βοήθεια της χρήσης μιας σειράς δεδομένων. Ειδικότερα στόχος είναι η εξίσωση να εμφανίζει την καλύτερη προσαρμογή στα εν λόγω δεδομένα.

Ο συντελεστής α αποτελεί τον σταθερό όρο ενώ ο συντελεστής β αποτελεί τον συντελεστή κλίσης. Ο σταθερός όρος εκφράζει την τιμή της εξαρτημένης μεταβλητής όταν η τιμή της ανεξάρτητης είναι ίση με το μηδέν, ενώ ο συντελεστής κλίσης εκφράζει την μεταβολή της εξαρτημένης μεταβλητής όταν η ανεξάρτητη μεταβλητή μεταβληθεί κατά μία μονάδα. Οι εκτιμήσεις των συντελεστών με την χρήση της μεθόδου των ελαχίστων τετραγώνων δίνονται από τις σχέσεις:

$$\hat{\beta} = \frac{\sum xy - \left(\sum x \sum y\right)/n}{\sum x^2 - \left(\sum x\right)^2/n}$$

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\,\overline{X} = \frac{\sum Y}{n} - \hat{\beta}\frac{\sum X}{n}$$

Έτσι εκτιμάται η απλή παλινδρόμηση:

$$y = \hat{\alpha} + \hat{\beta}\,x$$

σύμφωνα με την οποία η εξαρτημένη μεταβλητή Υ παλινδρομεί πάνω στην ανεξάρτητη μεταβλητή Χ.

6.3. Καμπυλόγραμμη Παλινδρόμηση

Οι μορφές της καμπυλόγραμμης παλινδρόμησης είναι η υπερβολή, η εκθετική, η λογαριθμική μορφή καθώς και κάποιες για τις οποίες θα γίνει μια απλή αναφορά.

1. **Υπερβολή**

Η εξίσωση της δίνεται από την σχέση:

$$Y = \frac{1}{\alpha + \beta x} \quad \text{ή} \quad \frac{1}{Y} = \alpha + \beta\, x$$

Σύμφωνα με τη δεύτερη εξίσωση υπάρχει γραμμική σχέση μεταξύ της ανεξάρτητης και της αντίστροφης της εξαρτημένης μεταβλητής. Οι εκτιμήσεις του σταθερού όρου και του συντελεστή κλίσης δίνονται από τις σχέσεις:

$$\hat{\beta} = \frac{\sum x \frac{1}{y} - \left(\sum x \sum \frac{1}{y} \right) / n}{\sum x^2 - \left(\sum x \right)^2 / n}$$

$$\hat{\alpha} = \frac{\sum \frac{1}{y}}{n} - \hat{\beta} \frac{\sum x}{n}$$

Με αυτό τον τρόπο προκύπτει η εκτιμημένη εξίσωση της παλινδρόμησης:

$$\frac{1}{Y} = \hat{\alpha} + \hat{\beta}\, x$$

2. **Εκθετική ή λογαριθμική μορφή**
3. Η συναρτησιακή σχέση της εκθετικής ή λογαριθμικής μορφής είναι:

$$Y = \alpha X^{\beta}$$

Με λογαρίθμιση και των δυο μελών της εξίσωσης καθώς και με τη χρήση των ιδιοτήτων των λογαρίθμων είναι δυνατός ο γραμμικός μετασχηματισμός της αρχικής σχέσης. Ειδικότερα, η σχέση που προκύπτει είναι η εξής:

$$\log Y = \log \alpha + \beta \log X$$

Η εξίσωση εκτιμάται με τη βοήθεια της μεθόδου των ελαχίστων τετραγώνων. Οι εκτιμήσεις του γραμμικά μετασχηματισμένου υποδείγματος δίνονται από τις σχέσεις:

$$\hat{\beta} = \frac{\sum \log(x)\log(y) - \left(\sum (\log(x)\sum \log(y)\right)/n}{\sum (\log(x))^2 - \left(\sum (\log(x))\right)^2 / n}$$

$$\log(\hat{\alpha}) = \frac{\sum (\log(Y))}{n} - \hat{\beta}\frac{\sum (\log(X))}{n}$$

Με αυτόν τον τρόπο εκτιμάται η εξίσωση της εκθετικής παλινδρόμησης:

$$\log Y = \log \hat{\alpha} + \hat{\beta} \log X$$

6.4. Άλλες μορφές μη γραμμικών εξισώσεων

Α) Αντίστροφη συνάρτηση:

$$Y = \alpha + \beta \frac{1}{X}$$

Ο γραμμικός μετασχηματισμός είναι δυνατός όταν θέσουμε X΄=1/X. Άρα η τροποποιημένη γραμμική μορφή της παλινδρόμησης είναι: $Y = \alpha + \beta X'$.

Β) Μια άλλη μορφή είναι η εξής:

$$Y = \alpha e^{\beta x}$$

Μέσω του γραμμικού μετασχηματισμού προκύπτει η εξίσωση παλινδρόμησης:

$$\ln Y = \ln a + \beta X$$

Η εν λόγω εξίσωση εκτιμάται με χρήση της μεθόδου ελαχίστων τετραγώνων. Επομένως η εκτιμημένη εξίσωση παλινδρόμησης θα είναι η εξής:

$Y^* = \hat{\alpha}^* + \hat{\beta} X$ με lnY = Y^* και $\hat{a}$ * = με την εκτίμηση του lnα

Από την σχέση $\hat{a}$ * = εκτίμηση του lnα προκύπτει η $\hat{a}$.

6.5. Πολλαπλή Παλινδρόμηση

Η εν λόγω παλινδρόμηση αφορά μία εξαρτημένη μεταβλητή η οποία είναι συνάρτηση πολλών ερμηνευτικών μεταβλητών. Θα εξεταστεί μία παλινδρόμηση τριών μεταβλητών:

$$Y = \alpha + \beta X + \gamma Z$$

Οι εκτιμήσεις των τριών συντελεστών δίνονται από τις σχέσεις:

$$\hat{\beta} = \frac{\left[\sum z^2 - \left(\sum z\right)^2/n\right]\left[\sum xy - \left(\sum x \sum y\right)/n\right] - \left[\sum yz - \left(\sum y \sum z\right)/n\right]\left[\sum xz - \left(\sum x \sum z\right)/n\right]}{\left[\sum x^2 - \left(\sum x\right)^2/n\right]\left[\sum z^2 - \left(\sum z\right)^2/n\right] - \left[\sum xz - \left(\sum x \sum z\right)/n\right]^2}$$

$$\hat{\gamma} = \frac{\left[\sum x^2 - \left(\sum x\right)^2/n\right]\left[\sum yz - \left(\sum y \sum z\right)/n\right] - \left[\sum xy - \left(\sum x \sum y\right)/n\right]\left[\sum xz - \left(\sum x \sum z\right)/n\right]}{\left[\sum x^2 - \left(\sum x\right)^2/n\right]\left[\sum z^2 - \left(\sum z\right)^2/n\right] - \left[\sum xz - \left(\sum x \sum z\right)/n\right]^2}$$

$$\hat{\alpha} = \frac{\sum y - \hat{\beta}\sum x - \hat{\gamma}\sum z}{n}$$

6.6. Πολυμεταβλητή καμπυλόγραμμη παλινδρόμηση

Η εξίσωση αυτής της μορφής παλινδρόμησης δίνεται από την σχέση: $Y = \alpha + \beta X + cX^2$.

Ο γραμμικός μετασχηματισμός της σχέσης αυτής είναι εφικτός με την αντικατάσταση z = X^2 .

6.7. Ο βαθμός συσχέτισης

Κατά την εκτίμηση της εξίσωσης παλινδρόμησης σημαντικά βήματα είναι τα εξής:

- ο έλεγχος της ικανότητας προσαρμογής της εξίσωσης της παλινδρόμησης στα δεδομένα που χρησιμοποιήθηκαν για την εκτίμηση αυτής.
- Η εφαρμογή κριτηρίων σημαντικότητας (έλεγχοι υποθέσεων για τους συντελεστές της εξίσωσης)

Ο έλεγχος της ικανότητας προσαρμογής είναι δυνατός με την γραφική απεικόνιση των σημείων των δεδομένων. Πιο συγκεκριμένα, τα δεδομένα αποδίδονται μέσω του διαγράμματος διασποράς και στην συνέχεια αναζητούμε την μορφή της καμπύλης που προσαρμόζεται στα δεδομένα αυτά.

Επιπλέον, ο έλεγχος της καλής προσαρμογής είναι δυνατός και με την εκτίμηση του μεγέθους της διασποράς των σημείων του δείγματος γύρω από την γραμμή παλινδρόμησης.

Η εκτίμηση του μεγέθους της διασποράς είναι δυνατή με τα εξής μέτρα:

- Το σφάλμα εκτίμησης της εξαρτημένης μεταβλητής y.
- Τον συντελεστή συσχέτισης
- Τον συντελεστή προσδιορισμού

6.8. Το σφάλμα εκτίμησης της εξαρτημένης μεταβλητής y.

Ο τύπος που χρησιμοποιείται είναι ο εξής:

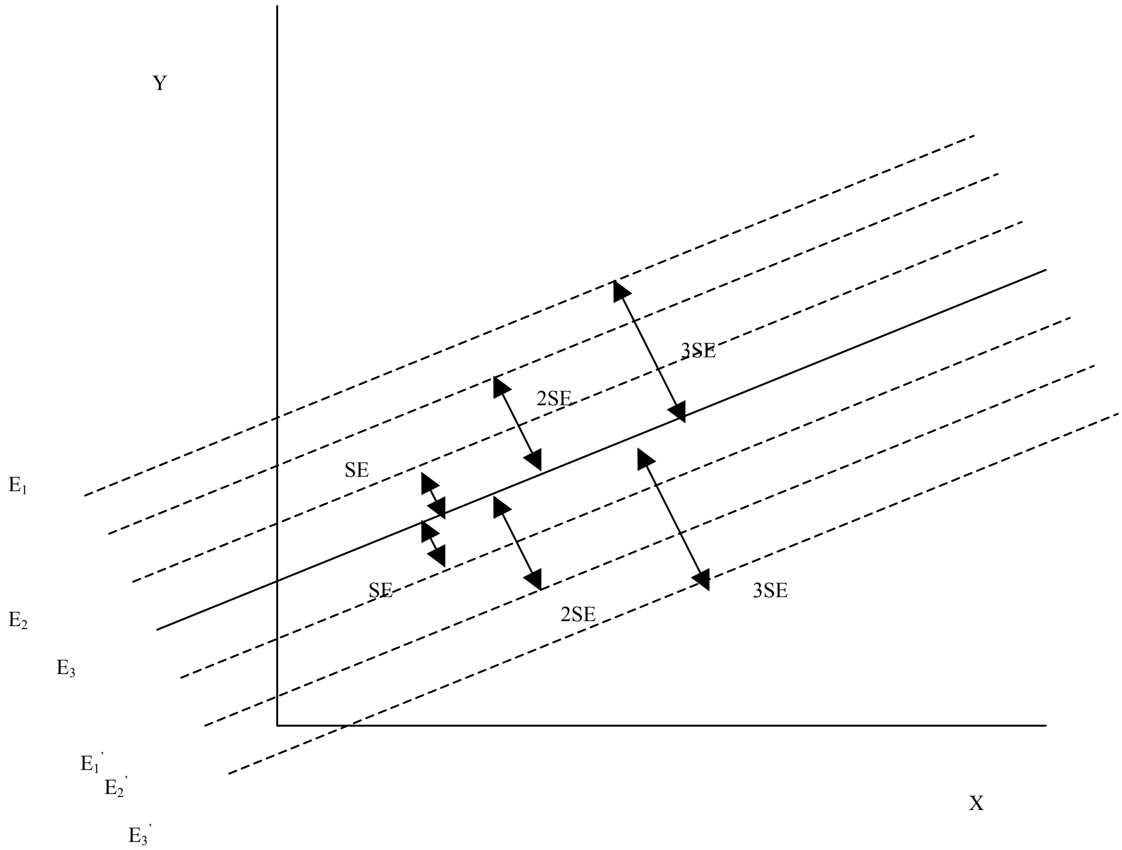
$$S_e = \sqrt{\frac{\sum\left(Y - \hat{Y}\right)^2}{n}}.$$

Το εν λόγω σφάλμα εκτιμά την απόκλιση των εκτιμηθέντων τιμών της εξαρτημένης μεταβλητής από τις τιμές των δεδομένων, παρέχοντας το μέγεθος της διασποράς, των υπό εξέταση σημείων, γύρω από τη γραμμή παλινδρόμησης.

Βασικά χαρακτηριστικά του εν λόγω μέτρου διασποράς είναι τα ακόλουθα:

1. Αναφέρεται στην απλή παλινδρόμηση

2. Η προσαρμογή βελτιώνεται όσο μικρότερο είναι το σφάλμα εκτίμησης.
3. Η χρήση αυτού του μέτρου προαπαιτεί την εκτίμηση της εξίσωσης παλινδρόμησης.



Τα ποσοστά των σημείων που ανήκουν σε κάθε ζώνη δίνονται από τα αντίστοιχα ποσοστά:

Στη ζώνη (-s.e., s.e.) ανήκει το 68% των σημείων

Στην ζώνη (-2s.e., +2s.e.) ανήκει το 95% των σημείων

Στην το ζώνη (-3s.e., +3s.e.) ανήκει 99% των παρατηρήσεων.

6.9. Ο συντελεστής συσχέτισης

Η μαθηματική σχέση που δίνει τον συντελεστή γραμμικής συσχέτισης είναι η ακόλουθη:

$$r = \frac{\sum xy - \left(\sum x \sum y\right)/n}{\left[\sqrt{[\sum x^2 - (\sum x)^2/n][\sum y^2 - \left(\sum y\right)^2/n]}\right]}$$

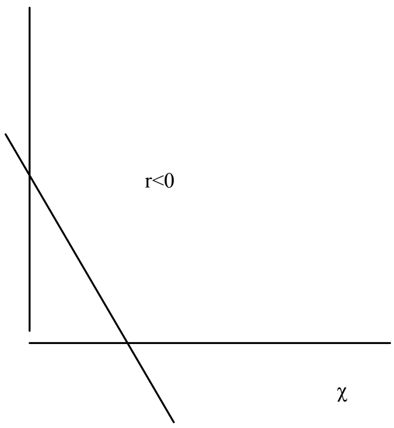
Βασικά χαρακτηριστικά του συντελεστή συσχέτισης είναι τα εξής:

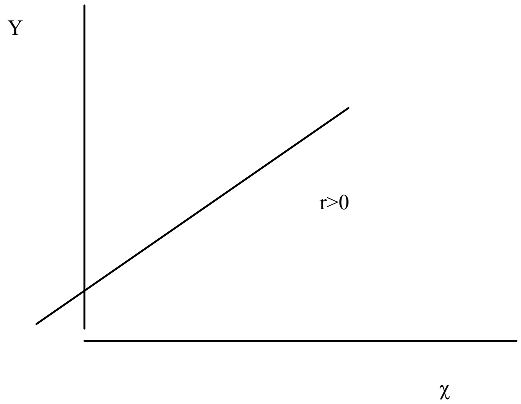
- Χρησιμοποιείται μόνο για την απλή παλινδρόμηση
- Δεν προαπαιτείται η εκτίμηση της εξίσωσης παλινδρόμησης
- Οι τιμές που παίρνει ανήκουν στο διάστημα [-1, 1].
- Αύξηση του συντελεστή συσχέτισης συνεπάγεται βελτίωση της προσαρμογής της εξίσωσης παλινδρόμησης.

Πράγματι όταν το μέγεθος του συντελεστή συσχέτισης τείνει προς τη μονάδα σε απόλυτη τιμή τότε αυξάνει η ικανότητα προσαρμογής, ενώ όταν ο συντελεστής συσχέτισης προσεγγίζει το μηδέν τότε είτε δεν υπάρχει γραμμική συσχέτιση είτε η σχέση των χ και y περιγράφεται από καμπυλόγραμμη σχέση.

Όταν η τιμή του r είναι θετική τότε η συσχέτιση μεταξύ των δύο μεταβλητών είναι θετική. Άρα η σχέση των δύο μεταβλητών περιγράφεται από μία ευθεία γραμμή με θετική κλίση.

Όταν η τιμή του r είναι αρνητική τότε η συσχέτιση μεταξύ των δύο μεταβλητών είναι αρνητική. Άρα η σχέση των δύο μεταβλητών περιγράφεται από μία ευθεία γραμμή με αρνητική κλίση. Οι δύο σχέσεις αποδίδονται γραφικά στα σχήματα που ακολουθούν:





Ο συντελεστής συσχέτισης στην ουσία εκφράζει τη διασπορά κατά γενικό τρόπο.

Πράγματι, μικρή τιμή του συντελεστή συσχέτισης σημαίνει μεγάλη διασπορά των σημείων γύρω από τη γραμμή παλινδρόμησης, ενώ μεγάλη τιμή του συντελεστή συσχέτισης σημαίνει μικρή διασπορά των σημείων γύρω από τη γραμμή παλινδρόμησης και άρα καλή προσαρμογή του υποδείγματος.

6.10. Ο συντελεστής προσδιορισμού γραμμικής παλινδρόμησης

Ο συντελεστής προσδιορισμού εκφράζεται ως το πηλίκο της μεταβλητότητας που οφείλεται στην παλινδρόμηση προς την συνολική μεταβλητότητα.

Ειδικότερα, ο συντελεστής προσδιορισμού εκφράζει το ποσοστό της συνολικής μεταβλητότητας που μπορεί να ερμηνευθεί από την παλινδρόμηση. Για παράδειγμα αν r^2 = 0,88 τότε το 88% της μεταβλητότητας της εξαρτημένης μεταβλητής ερμηνεύεται από την παλινδρόμηση ενώ το 12% ερμηνεύεται από τυχαίους παράγοντες.

Ειδικότερα ο τύπος που χρησιμοποιείται για τον συντελεστή προσδιορισμού είναι ο εξής:

$$r^2 = 1 - \frac{\sum \hat{u}^2}{\sum y^2}$$

6.11. Ο συντελεστής προσδιορισμού μη γραμμικής (καμπυλόγραμμης) παλινδρόμησης

Ο τύπος που χρησιμοποιείται για τη δεδομένη μορφή παλινδρόμησης ($Y = \alpha + \beta X + \gamma x^2$) είναι ο ακόλουθος:

$$r^2 = \frac{\hat{\beta}\left[\sum xy - \left(\sum x \sum y\right)/n\right] + \hat{\gamma}\left[\sum yx^2 - \left(\sum y \sum x^2\right)/n\right]}{\sum y^2 - \left(\sum y\right)^2 / n}$$

6.12. Έλεγχος στατιστικής σημαντικότητας του συντελεστή συσχέτισης

Η στατιστική σημαντικότητα του συντελεστή συσχέτισης δηλώνει την συσχέτιση των μεταβλητών X και Y. Ο έλεγχος αυτός είναι απαραίτητος με δεδομένο ότι η τιμή και μόνο του συντελεστή συσχέτισης δεν αποδίδει την συσχέτιση των δύο μεταβλητών αφού μεγάλη τιμή αυτής μπορεί να είναι αποτέλεσμα της επίδρασης άλλων αιτιών.

Για την πραγματοποίηση του ελέγχου απαιτείται η χρήση του κριτηρίου t.

Ο έλεγχος της στατιστικής σημαντικότητας του συντελεστή συσχέτισης γίνεται ως εξής:

Διατυπώνεται η υπόθεση και γίνεται έλεγχος αυτής:

$$H_o : \rho = 0$$
$$H_1 : \rho \neq 0$$

Η στατιστική που χρησιμοποιείται είναι η εξής:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

ρ: ο συντελεστής συσχέτισης στον πληθυσμό

r: ο συντελεστής συσχέτισης στο δείγμα

n: το μέγεθος του δείγματος

Το επίπεδο σημαντικότητας και συνεπώς το σφάλμα τύπου I είναι α%. Η κριτική τιμή που χρησιμοποιείται είναι η $t_{a/2,\, n-2}$ όπου n-2 είναι οι βαθμοί ελευθερίας του ελέγχου.

- Αν $t > t_{a/2,n-2}$, απορρίπτουμε την H_o και άρα ο συντελεστής συσχέτισης είναι στατιστικά σημαντικός (υπάρχει γραμμική συσχέτιση μεταξύ των δύο μεταβλητών).
- Αν $t < t_{a/2,n-2}$ αποδεχόμαστε την H_o και άρα ο συντελεστής συσχέτισης δεν είναι στατιστικά σημαντικός (δεν υπάρχει γραμμική σχέση μεταξύ των μεταβλητών).

6.13. Έλεγχος στατιστικής σημαντικότητας της διαφοράς μεταξύ των δύο συντελεστών συσχέτισης

Η μηδενική υπόθεση που ελέγχουμε και η εναλλακτική είναι οι εξής:

$$H_o : \rho_1 = \rho_2$$
$$H_1 : \rho_1 \neq \rho_2$$

Για τον έλεγχο αυτό απαιτείται η χρήση του κριτηρίου z ως προς και τις δύο πλευρές. Έστω r_1, r_2 είναι οι συντελεστές συσχέτισης. Οι τιμές των συντελεστών συσχέτισης κανονικοποιούνται με τη χρήση του τύπου:

$$z = \frac{1}{2}\log\frac{1+r}{1-r}$$

Στην συνέχεια υπολογίζεται η τυπική απόκλιση της διαφοράς των δύο z.

Δηλαδή:

$$\sigma_{z_1 - z_2} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}$$

Η στατιστική Z με τιμές z που χρησιμοποιείται είναι η εξής:

$$Z = \frac{z_1 - z_2}{\sigma_{z_1 - z_2}}$$

Λαμβάνουμε α% το επίπεδο σημαντικότητας.

- Αν $z > z_a$, απορρίπτουμε την H_o και άρα οι συντελεστές συσχέτισης διαφέρουν στατιστικά σημαντικά .
- Αν $z < z_a$ αποδεχόμαστε την H_o και άρα οι συντελεστές συσχέτισης δεν διαφέρουν στατιστικά σημαντικά.

<u>Ασκήσεις Γεωργικού Πειραματισμού</u>

1.) Οι αποδόσεις 10 πειραματικών τεμαχίων σε κιλά είναι 8, 12, 10, 14, 11, 13, 12, 9, 9, 7. Να βρεθεί

α) ο μέσος όρος

β) η τυπική απόκλιση

γ) το τυπικό σφάλμα των μέσων όρων

δ) Τα 95% όρια εμπιστοσύνης μεταξύ των οποίων βρίσκεται ο μέσος του πληθυσμού

x_i	v_i	x_i v_i	X_i^2	$V_i x_i^2$
7	1	7	49	49
8	1	8	64	64
9	2	18	81	162
10	1	10	100	100
11	1	11	121	121
12	2	24	144	288
13	1	13	169	169
14	1	14	196	196
	$\sum v_i$ =10	$\sum v_i x_i$ 105		$\sum v_i x_i^2$ 1149

Το μέγεθος του δείγματος είναι μικρό ενώ οι τιμές των παραμέτρων που πρέπει να υπολογιστούν είναι οι εξής:

$$\bar{x} = \frac{\sum v_i x_i}{n} = \frac{105}{10} = 10{,}5$$

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = 2{,}27$$

$$s_x = \frac{s}{\sqrt{n}} = 0{,}72$$

α=5%, β.ε.=10-1=9, $t_{9,0,025} = 2{,}262$

Άρα το αντίστοιχο διάστημα εμπιστοσύνης θα είναι το εξής:

$$\left[\bar{x}-\frac{t_{n-1;a/2}S}{\sqrt{n}},\ \bar{x}+\frac{t_{n-1;a/2}S}{\sqrt{n}}\right]$$ = [8,37, 11,63].

2) Δίνονται τα βάρη σε κιλά ενός τυχαίου δείγματος 15 κολοκυθιών με μέσο όρο $\bar{x}$ = 11,8 και διακύμανση s^2 =2,06.

α) Να βρεθεί η πιθανότητα ένα κολοκύθι στην τύχη να ζυγίζει μεταξύ 9 και 14 κιλά.

β) Να βρεθεί η πιθανότητα ο μέσος όρος του πληθυσμού των κολοκυθιών να βρίσκεται στο διάστημα (11, 13). Ο πληθυσμός θεωρούμε ότι είναι κανονικός με μ=12 και σ=2.

α) Αναζητούμε το $P(9 \le X \le 14) = P\left(\frac{9-\mu}{\sigma} < z < \frac{14-\mu}{\sigma}\right)$ = 0,775

β) Έχουμε ότι: $s_x = \frac{s}{\sqrt{n}} = 0{,}41$.

Με α% επίπεδο σημαντικότητας το διάστημα εμπιστοσύνης θα δίνεται από την σχέση:

$$\left[\bar{x}-\frac{t_{n-1;a/2}S}{\sqrt{n}},\ \bar{x}+\frac{t_{n-1;a/2}S}{\sqrt{n}}\right] = \left[11{,}8 - t_{a/2,n-1}0{,}41,\ 11{,}8 + t_{a/2,n-1}0{,}41\right]$$

Επειδή θέλουμε το κάτω όριο να είναι 11 και το άνω όριο να είναι 13 από τις σχέσεις:

11,8 - $t_{\frac{a}{2},n-1}$ =11 και 11,8 + $t_{\frac{a}{2},n-1}$ = 13 για το $t_{a/2,n-1}$ βρίσκουμε την τιμή 2,492 τιμή που αντιστοιχεί σε 2% επίπεδο σημαντικότητας.

* Αν το **μέγεθος του δείγματος ήταν 64** τότε έπρεπε να το διάστημα εμπιστοσύνης να υπολογισθεί με τη **χρήση της κανονικής κατανομής.**

Πιο συγκεκριμένα θα έχουμε:

$$s_x = \frac{s}{\sqrt{n}} = 0{,}179$$

Ενώ για α% επίπεδο σημαντικότητας το διάστημα εμπιστοσύνης θα δίνεται από την σχέση:

$$\left[\bar{x} - \frac{z_{a/2}S}{\sqrt{n}},\ \bar{x} + \frac{z_{a/2}S}{\sqrt{n}}\right] = \left[11{,}8 - z_{a/2}0{,}179,\ 11{,}8 + z_{a/2}0{,}179\right]$$

Τότε για το κάτω όριο να είναι 11 και το άνω όριο να είναι 13 από τις σχέσεις το z που βρίσκεται είναι 3,09 τιμή αντιστοιχεί σε 0,5% επίπεδο σημαντικότητας.

3) Το μέσο ύψος των φυτών μιας ποικιλίας καλαμποκιού βρέθηκε ίσο με 95,14 εκατοστά. Η τυπική απόκλιση του μέσου όρου είναι $\sigma_{\bar{x}} = 1{,}2$. Να εξεταστεί αν ισχύει η μηδενική υπόθεση:

$$H_0 : \mu = 90,50$$
$$H_1 : \mu \neq 90,50$$

Δίνονται n = 49 και α = 5%.

Η στατιστική που χρησιμοποιείται η Ζ με τιμές z:

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = 3,87$$

Έχουμε ότι $z_{0,05} = 1,96$. Συνεπώς z > $z_{0,05}$, άρα απορρίπτουμε την μηδενική.

* Αν το **δείγμα είναι μικρό** (n=25) τότε χρησιμοποιούμε την Τ μεταβλητή με τιμές t που ακολουθεί την t – student κατανομή.

$$t = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = 2,76$$

Έχουμε $t_{0,025,24} = 2,064$. Άρα και πάλι επειδή t > $t_{0,025,24}$, απορρίπτεται η μηδενική υπόθεση.

4) Να ελεγχθεί αν δύο ποικιλίες κριθαριού έχουν τον ίδιο μέσο όρο για τα δείγματα των οποίων έχουμε:

$\overline{X}_1 = 95,14 \quad \overline{X}_2 = 92.10$
$\sigma_1^2 = 70,44 \quad \sigma_2^2 = 80,50$
$n_1 = 49 \quad n_2 = 64$

Η στατιστική που χρησιμοποιείται είναι η Z με τιμές τις z και ακολουθεί κανονική κατανομή:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{3,04}{1,642} = 1,85$$

Έχουμε $z_{0.05} = 1,96$

Επειδή $z < z_{0.05}$ αποδεχόμαστε την μηδενική υπόθεση και άρα οι μέσοι όροι των δύο δειγμάτων δεν διαφέρουν σημαντικά.

* Έστω ότι τα **δύο δείγματα είναι μικρά** και τα παίρνουμε από δύο κανονικούς πληθυσμούς που θεωρούμε ότι έχουν **άγνωστες αλλά ίσες διασπορές.** Έχουμε ότι:

$$\bar{x}_1 = 95,14 \quad \bar{x}_2 = 92,10$$
$$s_1^2 = 70,44 \quad s_2^2 = 63,45$$
$$n_1 = 25 \quad n_2 = 16$$

Σε αυτήν την περίπτωση η στατιστική που χρησιμοποιούμε είναι η T με τιμές τις t και ακολουθεί t – student κατανομή:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{n_1 + n_2}{n_1 n_2} s^2}} = \frac{3,04}{3,06} = 0,99$$

Όπου $s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = 91,11$

Έχουμε $t_{0.05;39} = 2,020$

Επειδή t < $t_{0.05;39}$ αποδεχόμαστε την μηδενική υπόθεση και άρα οι μέσοι όροι των δύο δειγμάτων δεν διαφέρουν σημαντικά.

5) Στη ζυθοποιία η μορφή και το μέγεθος των κόκκων του κριθαριού παίζουν μεγάλο ρόλο. Για την αγορά του κριθαριού η εταιρία θέτει ως προϋπόθεση η διακύμανση του μεγέθους των κόκκων να μη διαφέρει από το $\sigma^2_0 = 22{,}56$. Το δείγμα για το οποίο πρέπει να αποφασιστεί είναι μεγέθους 28 με δειγματική διακύμανση $s^2=36{,}48$. Να εξεταστεί αν θα αγοραστούν οι κόκκοι από την εταιρία για πιθανότητα 5%.

Ορίζουμε τις υποθέσεις:

Η μηδενική υπόθεση είναι: H_0: $\sigma^2 = \sigma_0^2$

Η εναλλακτική υπόθεση είναι: H_1: $\sigma^2 \neq \sigma_0^2$

Η στατιστική που θα χρησιμοποιηθεί είναι η :

$$X^2 = \frac{1}{\sigma_0^2}(n-1)S^2$$

n -1= 27, $s^2 = 36{,}48$, $\sigma^2_0 = 22{,}56$

Με αντικατάσταση προκύπτει η τιμή της στατιστικής:

$$X^2 = \frac{1}{\sigma_0^2}(n-1)S^2 = 43{,}66$$

Η κριτική τιμή $X^2{}_{0,025;27} = 43{,}19$

Επειδή η τιμή της στατιστικής υπερβαίνει την κριτική τιμή η μηδενική υπόθεση δεν μπορεί να γίνει αποδεκτή και για το λόγο αυτό την απορρίπτουμε και άρα δεν θα επιλεχθεί από την εταιρία η ποικιλία αυτή του κριθαριού.

6) Μία γραπτή εξέταση δίνεται σε μία τάξη του Πανεπιστημίου για να συγκριθεί η μεταβλητότητα της επίδοσης φοιτητών και φοιτητριών. Στο δείγμα συμμετέχουν 20 φοιτητές και 28 φοιτήτριες και έχουμε ότι οι δειγματικές διακυμάνσεις είναι αντίστοιχα:

$$s_1^{\,2} = \frac{\sum (x_i - \bar{x})^2}{n-1} = 2{,}80$$

$$s_2^{\,2} = \frac{\sum (x_i - \bar{x})^2}{n-1} = 1{,}94$$

Έστω σ_1^2, σ_2^2 οι πληθυσμιακές διακυμάνσεις των φοιτητών και των φοιτητριών.

Ελέγχουμε τις υποθέσεις:

H_0: $\sigma_1^2 / \sigma_2^2 = 1$

H_1: $\sigma_1^2 / \sigma_2^2 \neq 1$

Σε επίπεδο σημαντικότητας α = 10%

Η στατιστική που θα χρησιμοποιηθεί είναι η :

$$F = \frac{s_1^2}{s_2^2}$$, με βαθμούς ελευθερίας n_1-1=19, n_2-1=27

Η κριτική τιμή που θα χρησιμοποιηθεί είναι η εξής: $F_{19,27;0,05} = 2{,}63$

Έχουμε F = 1,44 < $F_{19,27;0,05} = 2,63$, άρα αποδεχόμαστε την μηδενική υπόθεση και συνεπώς οι δύο διακυμάνσεις είναι ίδιες για 10% επίπεδο σημαντικότητας.

7) Ένα κέντρο παραγωγής σπόρων συγκεντρώνει σπόρους ενός φυτού όταν αυτοί παρουσιάζουν φυτρωτική ικανότητα τουλάχιστον 50%. Ένας παραγωγός προσέφερε σπόρους από τους οποίους σε μία σχετική δοκιμή από τους 54 φυτρώνουν οι 25. Ερωτάται αν το ποσοστό φυτρώματος πληροί τους όρους ώστε η συγκέντρωση να παραλάβει το σπόρο και να υπάρχει η πιθανότητα να πέσει έξω τουλάχιστον 5%.

Διατυπώνουμε τις υποθέσεις:

Η μηδενική υπόθεση H_0: $p = p_0$

Η εναλλακτική υπόθεση H_1: $p \neq p_0$

Η στατιστική που χρησιμοποιείται είναι η Ζ με τιμές z:

$$z = \frac{\frac{x}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0,463 - 0,5}{0,068} = -0,544$$

Έχουμε p = $\frac{x}{n}$ = 0,463, p_0 = 0,5

Έχουμε ότι $z_{0,025} = 1,96$, επομένως $z > -z_{0,025}$ και άρα το δείγμα πληροί τους όρους της συγκεντρώσεως, εφόσον δεν απορρίπτεται η μηδενική υπόθεση.

8) Δύο ορμόνες δοκιμάστηκαν για την εκτίμηση της ικανότητας τους να αυξήσουν τη ριζοβολία των μοσχευμάτων. Στην ορμόνη Α εμβαπτίστηκαν 55 μοσχεύματα από τα οποία ριζοβόλησαν τα 32 ενώ στην ορμόνη Β εμβαπτίστηκαν 68 από τα οποία ανέπτυξαν ρίζες τα 38. Μπορεί να λεχθεί για πιθανότητα 5% πως οι δύο ορμόνες είναι εξίσου αποτελεσματικές;

Διατυπώνουμε τις υποθέσεις:

Η μηδενική υπόθεση H_0: $p_1 = p_2$

Η εναλλακτική υπόθεση H_1: $p_1 \neq p_2$

Το στατιστικό ελέγχου που χρησιμοποιείται είναι:

$$Z = \frac{p_{1\Delta} - p_{2\Delta}}{\sqrt{pq(\frac{1}{n} + \frac{1}{l})}}$$ ακολουθεί την τυπική κανονική κατανομή

όπου $p = \frac{x_1 + x_2}{n_1 + n_2}$ εκτιμά την **$P_{1\ \pi\lambda\eta\theta}$ = $P_{2\ \pi\lambda\eta\theta}$ = P** και q = 1- p

Η απορριπτική περιοχή είναι: $R = \{|z| > z_{\frac{a}{2}}\}$, Έχουμε ότι:

$p_{1\Delta} = \frac{x_1}{n_1} = \frac{32}{55} = 0{,}58$, $p_{2\Delta} = \frac{x_2}{n_2} = \frac{38}{68} = 0{,}558$ και $p = \frac{x_1 + x_2}{n_1 + n_2} = \frac{32+38}{55+68} = 0{,}57$

με q = 0,43

$$\text{Άρα } z = \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}} = \frac{0{,}58 - 0{,}558}{0{,}008} = \frac{-0{,}022}{0{,}00} = -2{,}75$$

Έχουμε $z_{0,025} = 1{,}96$, επομένως $z \prec -z_{0.025}$ και άρα οι δύο ορμόνες διαφέρουν ως προς την αποτελεσματικότητά τους

9) Από μία ποικιλία σταριού εκτιμήθηκαν δύο χαρακτηριστικά του κόκκου, το βάρος και το μήκος. Να εκτιμηθεί η εξίσωση που εκφράζει το βάρος ως συνάρτηση του μήκους του κόκκου, το σφάλμα εκτίμησης, ο συντελεστής συσχέτισης και ο συντελεστής προσδιορισμού. Τα δεδομένα δίνονται αναλυτικά στις δύο πρώτες στήλες του πίνακα που ακολουθεί:

x	y	x - $\bar{x}$	(x - $\bar{x}$)^2	y - $\bar{y}$	x*y	(y - $\bar{y}$)^2
5	1,6	-7,5	56,25	-0,775	5,8125	0,600625
5	2,2	-7,5	56,25	-0,175	1,3125	0,030625
5	1,4	-7,5	56,25	-0,975	7,3125	0,950625
10	1,9	-2,5	6,25	-0,475	1,1875	0,225625
10	2,4	-2,5	6,25	0,025	-0,0625	0,000625
10	2,6	-2,5	6,25	0,225	-0,5625	0,050625
15	2,3	2,5	6,25	-0,075	-0,1875	0,005625
15	2,7	2,5	6,25	0,325	0,8125	0,105625
15	2,8	2,5	6,25	0,425	1,0625	0,180625
20	2,6	7,5	56,25	0,225	1,6875	0,050625
20	2,9	7,5	56,25	0,525	3,9375	0,275625
20	3,1	7,5	56,25	0,725	5,4375	0,525625
$\sum x_i =$ 150	$\sum y_i =$ 28,5	0	$\sum (x_i - \bar{x})^2 =$ 375	0	$\sum x * y =$ 27,75	$\sum (y_i - \bar{y})^2 =$ 3,0025

Με εφαρμογή των τύπων θα έχουμε:

$$\hat{\beta}_1 = \frac{\sum xy - \left(\sum x \sum y\right)/n}{\sum x^2 - \left(\sum x\right)^2 / n} = 0{,}074$$, συντελεστής συμμεταβολής

$$\hat{\beta}_o = \overline{Y} - \hat{\beta}_1 \overline{X} = \frac{\sum Y}{n} - \hat{\beta}_1 \frac{\sum X}{n}$$ =1,45

$$S_e = \sqrt{\frac{\sum\left(Y - \hat{Y}\right)^2}{n}} = 2{,}41$$

$$r = \left[\frac{\sum xy - \left(\sum x \sum y\right)/n}{\sqrt{[\sum x^2 - (\sum x)^2/n][\sum y^2 - (\sum y)^2/n}}\right] = 0{,}827 \text{ και } r^2 = 0{,}827^2$$

Από τα παραπάνω διαπιστώνουμε την ύπαρξη σημαντικής θετικής γραμμικής συσχέτισης.

10) Έστω ένα δείγμα των X, Y με n = 14 και r = 0,65. Να εξεταστεί αν υπάρχει γραμμική συσχέτιση μεταξύ των δύο μεταβλητών σε επίπεδο σημαντικότητας 5%.

Διατυπώνουμε τις υποθέσεις:

$$H_o : \rho = 0$$
$$H_1 : \rho \neq 0$$

Έχουμε r = 0,65, $r^2 = 0{,}4225$

Για α = 5%, και β.ε = 14 - 2 = 12, ισχύει: $t_{0,05;12} = 2{,}179$

Για τον υπολογισμό της στατιστικής χρησιμοποιούμε τον τύπο:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{2{,}25}{0{,}76} = 2{,}96 > 2{,}179$$

επομένως απορρίπτουμε την μηδενική υπόθεση και άρα η συσχέτιση μεταξύ των δύο μεταβλητών είναι στατιστικά σημαντική.

11) Έστω δύο δείγματα των X, Y και των Z, Y. Τα αντίστοιχα μεγέθη των δειγμάτων είναι n_1= 34 n_2= 28, ενώ οι συντελεστές συσχέτισης είναι r_1= 0,89, r_2= 0,93. Να ελεγχθεί αν οι συντελεστές συσχέτισης διαφέρουν στατιστικά σημαντικά σε επίπεδο σημαντικότητας 5%.

Διατυπώνουμε τις υποθέσεις:

$$H_o : \rho_1 = \rho_2$$
$$H_1 : \rho_1 \neq \rho_2$$

Έχουμε $z_1 = \frac{1}{2}\log\frac{1+r_1}{1-r_1} = 0{,}617$ και $z_2 = \frac{1}{2}\log\frac{1+r_2}{1-r_2} = 0{,}72$

Επίσης υπολογίζουμε το τυπικό σφάλμα:

$$\sigma_{z_1 - z_2} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}} = 0{,}268$$

Η στατιστική ελέγχου Z που παίρνει τιμές z ακολουθεί την κανονική κατανομή:

$$Z = \frac{z_1 - z_2}{\sigma_{z_1 - z_2}} = -0{,}384$$

Έχουμε $|z| = 0{,}384$, $z_{0,025}$ = 1,96.

Εφόσον έχουμε $|z| = 0{,}384$ < $z_{0,025}$ = 1,96 αποδεχόμαστε τη μηδενική υπόθεση και άρα δεν διαφέρουν στατιστικά σημαντικά σε α=5% οι συντελεστές συσχέτισης.

Βιβλιογραφία

Ζαχαροπούλου Χ. Στατιστική Μέθοδοι – Εφαρμογές. Θεσσαλονίκη, 1996

Κουτρουμανίδης Θ. Πανεπιστημιακές Παραδόσεις του μαθήματος Γεωργικού Πειραματισμού. Ορεστιάδα, 2005.

Φασούλας Α., Στοιχεία Πειραματικής Στατιστικής. Θεσσαλονίκη, 2004