3. Στατιστική Συμπερασματολογία

Συμπερασματολογία είναι ο κλάδος εκείνος της Λογικής που ασχολείται με την εξαγωγή συμπερασμάτων. Τα συμπεράσματα εκείνα μπορεί να αναφέρονται στο πλαίσιο της απόφασης ή πρόβλεψης. Μερικά παραδείγματα: Ο ερευνητής θέλει να αποδείξει την θεωρία του, ο γιατρός ενδιαφέρεται να δείξει ποιες από τις δύο θεραπείες είναι η καλύτερη. Για τα παραπάνω χρειάζεται η βοήθεια των σχετικών πληροφοριών για την εξαγωγή συμπερασμάτων που λέγεται δεδομένα ή παρατηρήσεις. Η στατιστική συμπερασματολογία έχει επαγωγικό χαρακτήρα, δηλαδή γενικεύει από ένα δείγμα για ολόκληρο τον πληθυσμό. Εξαιτίας αυτού εμπεριέχει και μια σχετική αβεβαιότητα η οποία μετριέται υπό την μορφή της πιθανότητας. Σκοπός της στατιστικής είναι η μελέτη μεθόδων Συμπερασματολογία και τρόπων μέτρησης της αβεβαιότητας αυτής. Επομένως για κάθε εφαρμογή της στατιστικής συμπερασματολογίας εμπεριέχονται δύο στοιχεία, το ένα εκ των οποίων είναι το συμπέρασμα και το δεύτερο είναι το μέτρο ορθότητας ή το μέτρο καταλληλότητας του.

3.1 Εκτίμηση Παραμέτρων

3.1.1 Γενικά

Έστω $X_1, X_2, \dots, X_n$ τυχαίο δείγμα απο κάποιον πληθυσμό με μία άγωστη παράμετρο. Το πρόβλημα μας είναι να βρούμε μια ή περισσότερες ποσότητες που θα χρησιμοποιηθούν για την εκτίμηση της άγνωστης παραμέτρου. Προφανές είναι ότι οι ποσότητες αυτές θα πρέπει να προέρχονται από το δείγμα. Ένα παράδειγμα είναι όταν θέλουμε να εκτιμήσουμε την μέση τιμή μ ενός πληθυσμού X, μπορούμε να χρησιμοποιήσουμε την ποσότητα $\overline{X}$. Η ποσότητα αυτή λέγεται εκτιμήτρια συνάρτηση και η αριθμητική τιμή της προκύπτει από το δείγμα εκτιμητής της παραμέτρου. **Οι εκτιμήτριες συναρτήσεις είναι στατιστικές συναρτήσεις και εξαρτώνται μόνο από το δείγμα και όχι από την παράμετρο που ζητάμε να εκτιμήσουμε. Στην περίπτωση αυτή εκτιμούμε ένα άγνωστο σημείο με ένα γνωστό σημείο**. Η σχετική μεθοδολογία λέγεται **εκτιμητική** και εδώ συγκεκριμένα **εκτίμηση σε σημείο ή σημειοεκτιμητική**.

Την άγνωστη παράμετρο μπορούμε να την εκτιμήσουμε με διάστημα αντί για σημείο καθώς θα έχει και περισσότερη ακρίβεια στην εκτίμηση της. Θα χρειαστούμε καταρχήν δύο ποσότητες που θα προέρχονται από το δείγμα τα όρια του διαστήματος. Η συγκεκριμένη μεθοδολογία ονομάζεται **εκτιμητική** ή **εκτίμηση με διάστημα** ή **εκτίμηση με διάστημα εμπιστοσύνης**.

Σε προβλήματα της στατιστικής η κατανομή του πληθυσμού είναι γνωστής μορφής εκτός από το γεγονός ότι περιέχει μια ή περισσότερες άγνωστες παραμέτρους. Στην παραπάνω περίπτωση βρισκόμαστε στην σφαίρα των παραμετρικών στατιστικών μοντέλων, όπου γνωρίζουμε μια ή περισσότερες παραμέτρους εκ των προτέρων (a priori). Στην αντίθετη περίπτωση, όπου δεν γνωρίζουμε την κατανομή του πληθυσμού τότε έχουμε το Μη-Παραμετρικό στατιστικό μοντέλο. Ένα παράδειγμα είναι ότι η κατανομή ενός πληθυσμού μπορεί να ακολουθεί την Διωνυμική κατανομή $B(n, p)$ με n γνωστό και $p \in (0,1)$ άγνωστο ή κανονική $N(\mu, \sigma^2)$ με $-\infty < \mu < \infty$, $\sigma^2 > 0$ άγνωστα. Όταν έχουμε παρατηρήσεις $x_1, x_2, \dots, x_n$ και δεν γνωρίζω το μοντέλο τους, τότε βρισκόμαστε στο πλαίσιο της μη-παραμετρικής στατιστικής. Γενικά οι άγνωστες παράμετροι θα συμβολίζονται με θ ή θ_1, θ_2, κλπκαι οι κατανομές του πληθυσμού X με $f(x, \theta)$. Το πρόβλημα εστιάζεται στο ότι θέλουμε να εκτιμήσουμε το θ ή συναρτήσεις αυτού $g(\theta)$. Οι εκτιμήτριες συναρτήσεις του θ και θα συμβολίζονται με $\hat{\theta}$ ή πιο γενικά με $T(X_1, X_2, \dots, X_n)$.

3.1.2 Εκτίμηση σε σημείο (Σημειοεκτιμητική)

Ένα διπλό πρόβλημα που έχει να αντιμετωπίσει η σημειοεκτιμητική είναι τα εξής ερωτήματα:

"Ποια είναι τα κριτήρια ή αρχές αξιολόγησης των εκτιμητών;" και "Πώς ορίζουμε τον ¨καλύτερο¨ εκτιμητή μεταξύ των διαφόρων εκτιμητών και ποιες οι μέθοδοι αυτών;". Σε αυτό το σημείο τα κριτήρια αξιολόγησης των εκτιμητών είναι πολλά. Αναφορικά μερικά βασικά κριτήρια αξιολόγησης των εκτιμητών είναι: Η **Αμεροληψία, Ελάχιστη Διακύμανση, Ακρίβεια, Επάρκεια, Συνέπεια** κλπ.

Αμεροληψία

Ένας εκτιμητής $\hat{\theta}$ μιας παραμέτρου θ λέγεται αμερόληπτος αν η αναμενόμενη τιμή θ για κάθε τιμή της παραμέτρου θ, δηλαδή αν

$$E(\hat{\theta}) = \theta, \qquad \forall \theta \in \Theta$$

Η αμεροληψία ενός εκτιμητή μετρά την ορθότητα της εκτίμησης και εκφράζει την ιδέα ότι σε πολλές επαναλήψεις της δειγματοληψίας μας η μέθοδος εκτίμησης που θα χρησιμοποιούμε θα μας δώσει κατά μέσο όρο την άγνωστη παράμετρο θ. Παράδειγμα αν n=50 οι εκτιμητές

$$\overline{X}_{50} = \frac{1}{50}\sum_{i=1}^{50} X_i \text{ και ο } \overline{X}_{25} = \frac{1}{25}\sum_{i=1}^{25} X_i$$

είναι και οι δυο αμερόληπτοι εκτιμητές του μ, με την διαφορά ότι ο εκτιμητής $\overline{X}_{50}$ είναι καλύτερος από τον $\overline{X}_{25}$ επειδή ο εκτιμητής $\overline{X}_{50}$ χρησιμοποιεί περισσότερη πληροφορία από τον $\overline{X}_{25}$. Εκείνοι λοιπόν που παίζει ρόλο είναι η μεταβλητότητα του εκτιμητή. Αναφορικά η διακύμανση του εκτιμητή $\overline{X}_{50}$ είναι $\sigma^2/50$, ενώ η διακύμανση του εκτιμητή $\overline{X}_{25}$ είναι $\sigma^2/25$. δηλαδή

$$V(\overline{X}_{50}) < V(\overline{X}_{25})$$

αυτά μας οδηγούν στο δεύτερο κριτήριο, εκείνο της Ελάχιστης Διακύμανσης.

Ελάχιστη Διακύμανση

Μεταξύ των αμερόληπτων εκτιμητών προτιμητέος είναι εκείνος που έχει την **μικρότερη διακύμανση**.

Η **ακρίβεια (precision)** ενός αμερόληπτου εκτιμητή $\hat{\theta}$ μετριέται συνήθως με την διακύμανση του.

$$\sigma_{\hat{\theta}}{}^2 = V(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

ή την τυπική απόκλιση του (Τυπικό Σφάλμα)

$$\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})} = \sqrt{E(\hat{\theta} - \theta)^2}$$

Για τον δειγματικό μέσο $\overline{X}$ το τυπικό σφάλμα είναι

$$\sigma_{\overline{X}} = \sqrt{\frac{\sigma^2}{n}}$$

Ένας εκτιμητής λέγεται αμερόληπτος ομοιόμορφα ελάχιστης διακύμανσης (Α.Ο.Ε.Δ) όταν είναι αμερόληπτος και έχει την μικρότερη διακύμανση από κάθε άλλο αμερόληπτο εκτιμητή.

Αν η άγνωστη παράμετρος θ είναι η μέση τιμή του πληθυσμού μ, ο $\overline{X}$ έχει την ιδιότητα της αμεροληψίας, δηλαδή

$$E(\overline{X}) = \mu$$

Και αν ο πληθυσμός είναι κανονικός ο $\overline{X}$ είναι ο αμερόληπτος ομοιόμορφα ελάχιστης διακύμανσης εκτιμητής του μ. Ομοίως αν $X_1, X_2, \dots, X_n$ τυχαίο δείγμα απο κάποιον πληθυσμό με μία άγνωστη παράμετρο σ^2, το στατιστικό S^2είναι αμερόληπτος εκτιμητής του σ^2, δηλαδή

$$E(S^2) = \sigma^2$$

Και ισχύουν πάντα ασχέτως με των τιμών που μπορεί να πάρουν τα μ και σ^2.

3.2 Διαστήματα Εμπιστοσύνης

Έστω $X_1, X_2, \dots, X_n$ τυχαίο δείγμα απο κάποιον πληθυσμό με μία άγνωστη παράμετρο θ. Έστω $L = L(X_1, X_2, \dots, X_n)$ το κάτω όριο (Lower Bound) και $U = U(X_1, X_2, \dots, X_n)$ (Upper Bound) το άνω όριο ενός διαστήματος. Τα όρια αυτά είναι συναρτήσεις των τυχαίων μεταβλητών του δείγματος. Οι τιμές τους είναι συναρτήσεις των τυχαίων μεταβλητών του δείγματος. Οι τιμές τους είναι $l(x_1, x_2, \dots, x_n)$ και $u(x_1, x_2, \dots, x_n)$. Εκτιμούμε το διάστημα (L,U) με τιμή (l,u) για την εκτίμηση του θ. Ζητάμε:

1. Το διάστημα (L, U) να περιέχει την αληθινή τιμή του θ ένα μεγάλο «ποσοστό φορών» και
2. Το διάστημα να έχει όσο το δυνατό μικρότερο μήκος. Το «ποσοστό φορών» που ένα διάστημα (L, U) περιέχει το θ λέγεται **βαθμός εμπιστοσύνης.**

Συμβολίζεται με $100(1-\alpha)\%$ και δείχνει την πιθανότητα το (L, U) να περιέχει το θ δηλαδή,

$$\text{Βαθμός Εμπιστοσύνης} = P(L < \theta < U) = 1 - a$$

Το διάστημα (L, U) λέγεται **Διάστημα Εμπιστοσύνης.**

Κατασκευάζουμε ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για την μέση τιμή ενός κανονικού πληθυσμού με βάση ένα τυχαίο δείγμα με μέγεθος n και με την υπόθεση ότι η διακύμανση σ^2 του πληθυσμού ότι είναι γνωστή. Έστω λοιπόν ότι $\overline{X}$ η μέση τιμή του δείγματος. Γνωρίζουμε ότι η κατανομή του $\overline{X} \sim N(\mu, {}^{\sigma^2}/_{n})$ έτσι για κάθε μ έχουμε ότι

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Έστω $t_{\alpha/2}$ το αντίστροφο εκατοστιαίο σημείο της τυπικής κανονικής κατανομής

$$t_{\alpha/2} = P\left(T \geq t_{\alpha/2}\right) = \alpha/2$$

Τότε έχουμε ότι

$$P(-t_{\alpha/2} \leq Z \leq t_{\alpha/2}) = 1 - a \iff$$

$$P\{-t_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < t_{\alpha/2}\} = 1 - a \iff$$

$$P\{-\frac{\sigma}{\sqrt{n}} t_{\alpha/2} < \overline{X} - \mu < \frac{\sigma}{\sqrt{n}} t_{\alpha/2}\} = 1 - a \iff$$

$$P\{\overline{X} - \frac{\sigma}{\sqrt{n}} t_{\alpha/2} < \mu < \overline{X} + \frac{\sigma}{\sqrt{n}} t_{\alpha/2}\} = 1 - a\,.$$

Επομένως το διάστημα (L, U) με όρια

$$\mathbf{L} = \overline{\mathbf{X}} - \frac{\boldsymbol{\sigma}}{\sqrt{\mathbf{n}}} \mathbf{z}_{\boldsymbol{\alpha}/2} \text{ και } \mathbf{U} = \overline{\mathbf{X}} + \frac{\boldsymbol{\sigma}}{\sqrt{\mathbf{n}}} \mathbf{z}_{\boldsymbol{\alpha}/2}$$

έχουν την ιδιότητα ότι

$$P\{L < \mu < U\} = 1 - a, \qquad \forall \mu$$

Επομένως το Διάστημα εμπιστοσύνης για την μέση τιμή μ από κανονικό πληθυσμό $N(\mu, \sigma^2)$ με σ^2 γνωστό είναι:

$$(\overline{X} - \frac{\sigma}{\sqrt{n}} t_{\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}} t_{\alpha/2})$$

ή διαφορετικά

$$\overline{X} \pm \frac{\sigma}{\sqrt{n}} t_{\alpha/2}$$

με βαθμό εμπιστοσύνης $100(1-\alpha)\%$

Παράδειγμα

Ένα διεγερτικό φάρμακο ελέγχεται για την επίδραση του στην πίεση του αίματος. Οι πιέσεις αίματος n=20 ατόμων μετριούνται πριν από την λήψη και μίση ώρα μετά την λήψη και λαμβάνονται οι ακόλουθες διαφορές

7	6	0	8	-9	-4	0	1	9	1
2	7	0	6	-6	-5	-1	6	-2	4

Είναι γνωστό από προηγούμενες μελέτες ότι η πριν και μετά την λήψη φαρμάκου διαφορά πιέσεων ακολουθεί την κανονική κατανομή με γνωστή διακύμανση $\sigma^2 = 25$. Να κατασκευαστεί ένα 95% διάστημα εμπιστοσύνης για της μέσης διαφοράς μ της πίεσης του αίματος.

Απάντηση

Επειδή το μοντέλο είναι κανονική κατανομή με γνωστή διακύμανση $\sigma^2 = 25$, το διάστημα εμπιστοσύνης για το μ θα δίνεται από την σχέση

$$\overline{X} \pm \frac{\sigma}{\sqrt{n}} t_{\alpha/2} \ (1)$$

Από τα δεδομένα του προβλήματος μας δίνονται

n=20, $\overline{X}$=0.6 , σ^2=25, σ=5 και 1-α=0.95⇔α=0.05 και από την σχέση $P(T \geq t_{0.025}) = 0.05$ έχουμε από τους πίνακες βρίσκοντας ότι $t_{0.025}$=1.96. Επομένως η (1) θα γίνει:

$$0.6 \pm \frac{5}{\sqrt{20}} t_{0.05/2} = 0.6 \pm \frac{5}{\sqrt{20}} t_{0.025} = 0.6 \pm 2.19$$

Επομένως το διάστημα εμπιστοσύνης της μέσης τιμής της διαφοράς της πίεσης πριν και μετά το φάρμακο με βαθμό εμπιστοσύνης 95% είναι το (-1.59 , 2.79). Αυτό πρακτικά σημαίνει ότι *αν εκτελέσουμε το πείραμα 100 φορές αναμένουμε για την μέση τιμή μ τις 95 η μέση τιμή να βρίσκεται εντός των διαστημάτων αυτών.*

Είπαμε για την έννοια του διαστήματος εμπιστοσύνης και δώσαμε ένα διάστημα τέτοιο με την μέση τιμή μ να είναι άγνωστη και την διακύμανση σ^2 να είναι γνωστή a priori. Ας δούμε μια περίπτωση όπου η μέση τιμή μ και η διακύμανση σ^2 είναι αμφότεροι άγνωστοι a priori. Στην περίπτωση αυτή γνωρίζουμε ότι η τυχαία μεταβλητή

$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

Και είναι και συμμετρική ως προς το μ. Έτσι και σε αυτή την περίπτωση έχουμε τα συμμετρικά σημεία $-t_{n-1,\alpha/2}$ και $t_{n-1,\alpha/2}$ τέτοια ώστε:

$$P\left(-t_{n-1,\alpha/2} \leq t \leq t_{n-1,\alpha/2}\right) = 1 - a \Leftrightarrow$$

$$P\left(-t_{n-1,\alpha/2} \leq \frac{\overline{X} - \mu}{S/\sqrt{n}} \leq t_{n-1,\alpha/2}\right) = 1 - a \Leftrightarrow$$

$$P\left(\overline{X} - \frac{S}{\sqrt{n}} t_{n-1,\alpha/2} \leq \mu \leq \overline{X} + \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}\right) = 1 - a$$

δηλαδή

$$\left(\overline{X} - \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}, \overline{X} + \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}\right)$$

$$\overline{X} \pm \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}$$

Με $P\left(t \geq t_{n-1,\alpha/2}\right) = \alpha/2$

Συνοψίζοντας έχουμε:

$$\overline{X} \pm \frac{\sigma}{\sqrt{n}} t_{\alpha/2} \quad , \quad \sigma^2\text{: γνωστό}$$

$$\overline{X} \pm \frac{S}{\sqrt{n}} t_{n-1,\alpha/2} \quad , \quad \sigma^2\text{: άγνωστο}$$

Στο Παράδειγμα μας αν υποθέταμε πλέον ότι δεν γνωρίζουμε την διακύμανση σ^2, θα είχαμε ότι

n=20, $\overline{X}$= 0.6 , S^2= 25.84 και S = 5.08. Για τον υπολογισμό του S θα είχαμε ότι

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2} = 5.08$$

και το 95% διάστημα εμπιστοσύνης θα ήταν το

$$0.6 \pm \frac{5.08}{\sqrt{20}} \mathrm{t}_{20-1,0.05/2} = 0.6 \pm \frac{5.08}{\sqrt{20}} \mathrm{t}_{19,0.025} = 0.6 \pm 2.38 = (-1.78\,, 2.98)$$

3.3 Έλεγχος Υποθέσεων

Συχνά μας ενδιαφέρει το ερώτημα αν η μέση τιμή ενός πληθυσμού ισούται μια κάποια συγκεκριμένη τιμή. Εξετάζουμε ένα δείγμα από τον πληθυσμό και θέλουμε να ελέγξουμε αν η συγκεκριμένη μέση τιμή που έχουμε ορίσει ισούται όντως με αυτή την τιμή. Για παράδειγμα θέλουμε να εξετάσουμε αν η μέση συστολική πίεση ισούται όντως με 125mmHg. Η συμβολή του ερωτήματος αυτού θα μπορούσε να εδράζεται στην ανάπτυξη ενός φαρμάκου που θα πρέπει να αξιολογηθεί. Αν το φάρμακο μειώνει ή αυξάνει κατά πολύ την συστολική πίεση, τότε αυτό θεωρείται παρενέργεια.

Για το συγκεκριμένο ερώτημα έστω ότι μας ενδιαφέρει η μηδενική υπόθεση του ελέγχου ότι η μέση συστολική πίεση είναι 125mmHg ($\mathrm{H}_0\!: \mu = 125\mathrm{mmHg}$) έναντι της εναλλακτικής υπόθεσης ότι η μέση συστολική πίεση είναι μεγαλύτερη από 125mmHg ($\mathrm{H}_1\!: \mu > 125\mathrm{mmHg}$).

Γενικά μας ενδιαφέρουν οι έλεγχοι της μορφής:

$H_0\!: \mu \leq \mu_0$	$H_0\!: \mu \geq \mu_0$	$H_0\!: \mu = \mu_0$
$H_1\!: \mu > \mu_0$	$H_1\!: \mu < \mu_0$	$H_1\!: \mu \neq \mu_0$

όπου μ_0 είναι μια τιμή ελέγχου.

3.3.1 Βασικοί Έλεγχοι

Έστω ότι έχουμε ένα τυχαίο δείγμα από έναν πληθυσμό με μέση τιμή μ και διακύμανση σ^2 και υπολογίζουμε την δειγματική μέση τιμή και την δειγματική διακύμανση μ και S^2 αντίστοιχα. Για να εκτιμήσουμε την μέση τιμή μ και να ελέγξουμε υποθέσεις για αυτή διακρίνουμε τρείς περιπτώσεις.

1. ο πληθυσμός ακολουθεί την κανονική κατανομή και η διακύμανση είναι γνωστή
2. ο πληθυσμός ακολουθεί την κανονική κατανομή αλλά η διακύμανση είναι άγνωστη
3. ο πληθυσμός ακολουθεί μια άγνωστη κατανομή αλλά το μέγεθος του δείγματος είναι μεγάλο

1. Ο πληθυσμός ακολουθεί την κανονική κατανομή και η διακύμανση είναι γνωστή

Στην περίπτωση που ο πληθυσμός ακολουθεί την κανονική κατανομή με γνωστή διακύμανση, η μέση τιμή μ εκτιμάται από την ποσότητα $\hat{\mu} = \overline{X}$ (η δειγματική μέση τιμή), η οποία έχει τυπικό σφάλμα $\hat{\sigma}_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$. Συνήθως την τιμή της διακύμανσης την γνωρίζουμε από παλαιότερες μελέτες.

Οι υποθέσεις που μπορούμε να ελέγξουμε είναι οι εξής:

$H_0\!: \mu \leq \mu_0$	$H_0\!: \mu \geq \mu_0$	$H_0\!: \mu = \mu_0$
$H_1\!: \mu > \mu_0$	$H_1\!: \mu < \mu_0$	$H_1\!: \mu \neq \mu_0$

Η στατιστική συνάρτηση που χρησιμοποιούμε σε αυτές τις περιπτώσεις είναι η

$$Z = \frac{\overline{\mathrm{X}} - \mu_0}{\sigma/\sqrt{\mathrm{n}}} \sim \mathrm{N}(0,1), \qquad \acute{o}\tau\alpha\nu\ \mu = \mu_0$$

Οι αντίστοιχες κρίσιμες περιοχές θα είναι οι:

$$Z \geq z_a\,, \qquad Z \leq -z_a, \qquad |Z| \geq z_{a/2}$$

Αυτό σημαίνει ότι απορρίπτουμε την μηδενική υπόθεση $H_0\!: \mu \leq \mu_0$ έναντι της εναλλακτικής $H_1\!: \mu < \mu_0$ εάν η τιμή της στατιστικής συνάρτησης Z είναι μεγαλύτερη ή ίση από την τιμή z_a. Αντίστοιχα, απορρίπτουμε την μηδενική υπόθεση $H_0\!: \mu \geq \mu_0$ έναντι της εναλλακτικής $H_1\!: \mu < \mu_0$ εάν η τιμή της στατιστικής συνάρτησης Z είναι μικρότερη ή ίση από την τιμή $-z_a$.Τέλος απορρίπτουμε την μηδενική υπόθεση $H_0\!: \mu = \mu_0$ έναντι της εναλλακτικής $H_1\!: \mu \neq \mu_0$ εάν η στατιστική συνάρτηση *Z έχει* τιμή που είναι μικρότερη από την τιμή $-z_{a/2}$ ή μεγαλύτερη από την τιμή $z_{a/2}$.

Παράδειγμα

Έστω ότι τα δεδομένα 71, 79, 80, 76, 103, 121, 114, 100, 85, 99 αποτελούν μετρήσεις του δείκτη νοημοσύνης 10 παιδιών. Υποθέτουμε ότι ο πληθυσμός από τον οποίο πήραμε το δείγμα αυτό ακολουθεί την κανονική κατανομή με $\sigma^2 = 15^2$. *Να ελεγχθεί η υπόθεση ότι ο μέσος δείκτης νοημοσύνης ισούται με 89 έναντι της υπόθεσης ότι είναι διάφορος του 89, σε επίπεδο σημαντικότητας α=0.05.*

Λύση

Η υπόθεση που θέλουμε να ελέγξουμε είναι η

$$H_0\!: \mu = 90$$

$$H_1\!: \mu \neq 90$$

Επειδή γνωρίζουμε τη διακύμανση, ο έλεγχος που θα χρησιμοποιήσουμε είναι το t τεστ με στατιστική συνάρτηση την

$$Z = \frac{\overline{\mathrm{X}} - \mu_0}{\sigma/\sqrt{\mathrm{n}}} \sim \mathrm{N}(0,1)\,, \acute{o}\tau\alpha\nu\ \iota\sigma\chi\acute{\upsilon}\varepsilon\iota\ \eta\ H_0$$

Από τα δεδομένα, έχουμε ότι:

$$\overline{\mathrm{X}} = \frac{1}{10}\sum_{i=1}^{10} X_i = 92.2\,, \mu = 90\,, \sigma^2 = 15^2\ \kappa\alpha\iota\ n = 10$$

άρα η στατιστική συνάρτηση $Z = \frac{\overline{\mathrm{X}} - \mu_0}{\sigma/\sqrt{\mathrm{n}}}$ *ισούται με*

$$Z = \frac{92.2\ - 90}{15/\sqrt{10}} = 0.4638$$

Η κρίσιμη περιοχή του ελέγχου είναι η $|Z| \geq z_{a/2}$*. Απορρίπτουμε δηλαδή την μηδενική υπόθεση όταν η απόλυτη τιμή της στατιστικής συνάρτησης Z είναι μεγαλύτερη ή ίση από την τιμή* $z_{a/2}$*. Από τον πίνακα της τυπικής κανονικής κατανομής βρίσκουμε ότι* $z_{a/2} = z_{0.05/2} = z_{0.025} = 1.96$ *δηλαδή,*

$$Z < z_{a/2}$$

δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση. Συνεπώς, ο μέσος δείκτης νοημοσύνης ισούται με 90. Ο παραπάνω έλεγχος είναι γνωστός και ως ***t-test****.*

2. Ο πληθυσμός ακολουθεί την κανονική κατανομή και η διακύμανση είναι άγνωστη

Στην περίπτωση που ο πληθυσμός ακολουθεί την κανονική κατανομή αλλά δεν γνωρίζουμε την τιμή της διακύμανσης, ο καλύτερος εκτιμητής είναι πάλι ο $\hat{\mu} = \overline{X}$ (η δειγματική μέση τιμή), η οποία έχει τυπικό σφάλμα $\hat{\sigma}_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$. Αυτή είναι η περίπτωση που συναντάται πιο συχνά στην πράξη καθώς τις περισσότερες φορές δεν γνωρίζουμε την διακύμανση του πληθυσμού. Για να ελέγξουμε τις υποθέσεις που είναι ίδιες με την προηγούμενη περίπτωση χρησιμοποιούμε την στατιστική συνάρτηση:

$$\mathrm{t} = \frac{\overline{\mathrm{X}} - \mu_0}{\mathrm{S}/\sqrt{\mathrm{n}}} \sim \mathrm{t}_{\mathrm{n}-1}\, , \acute{o}\tau\alpha\nu\ \iota\sigma\chi\acute{\upsilon}\varepsilon\iota\ \eta\ H_0$$

Απλά εδώ έχουμε αντικαταστήσει την τυπική απόκλιση σ με την δειγματική τυπική απόκλιση S. Οι αντίστοιχες κρίσιμες περιοχές είναι

$$t \geq t_{n-1;a}\ , \qquad t \leq -t_{n-1;a}, \qquad |t| \geq t_{n-1;a/2}$$

Αυτό σημαίνει ότι απορρίπτουμε την μηδενική υπόθεση $H_0\colon \mu \leq \mu_0$ έναντι της εναλλακτικής $H_1\colon \mu < \mu_0$ εάν η τιμή της στατιστικής συνάρτησης t είναι μεγαλύτερη ή ίση από την τιμή $t_{n-1;a}$. Αντίστοιχα, απορρίπτουμε την μηδενική υπόθεση $H_0\colon \mu \geq \mu_0$ έναντι της εναλλακτικής $H_1\colon \mu < \mu_0$ εάν η τιμή της στατιστικής συνάρτησης t είναι μικρότερη ή ίση από την τιμή $-t_{n-1;a}$.Τέλος απορρίπτουμε την μηδενική υπόθεση $H_0\colon \mu = \mu_0$ έναντι της εναλλακτικής $H_1\colon \mu \neq \mu_0$ εάν η στατιστική συνάρτηση *t έχει* τιμή που είναι μικρότερη από την τιμή – $t_{n-1;a/2}$ ή μεγαλύτερη από την τιμή $t_{n-1;a/2}$.*Ο παραπάνω έλεγχος είναι γνωστός και ως* ***t-test****.*

Παράδειγμα

Μια ιατρική ομάδα θέλει να μελετήσει το λόγο βάρους των νεογέννητων μωρών που γεννιούνται από μητέρες με διαβήτη. Ο λόγος βάρους προσδιορίζεται διαιρώντας το βάρος κατά τη γέννηση με το μέσο βάρος για την εκτιμημένη ηλικία κυοφορίας, η οποία είναι διαθέσιμη σε πίνακες που έχουν στηριχθεί σε προηγούμενα δεδομένα. Εξ ορισμού, ο μέσος λόγος βάρους των νεογέννητων του πληθυσμού που χρησιμοποιήθηκε για την κατασκευή των πινάκων ισούται με 1. Η ιατρική ομάδα θέλει να ελέγξει εάν ο λόγος βάρους των παιδιών που θα γεννηθούν κατά τη διάρκεια μιας χρονιάς είναι διαφορετικός του 1. Με λίγα λόγια θέλει να διαπιστώσει εάν υπάρχει μια διαχρονική αλλαγή.
Οι μετρήσεις των 40 νεογέννητων έχουν ως εξής:

1,09	1,92	1,42	1,28	1,03	0,91	1,19	1,29	0,89	0,74
1,16	1,45	0,90	0,73	0,92	1,59	1,44	0,78	1,24	1,44
0,73	1,06	1,01	0,65	1,15	1,02	1,29	1,12	1,23	1,46
0,82	0,67	0,88	1,13	0,59	0,64	1,02	1,32	0,67	0,89

Λύση

Επειδή η ιατρική ομάδα ενδιαφέρεται να ελέγξει εάν ο λόγος βάρους των παιδιών που θα γεννηθούν κατά τη διάρκεια μιας χρονιάς είναι διαφορετικός του 1, ο έλεγχος υπόθεσης θα είναι της μορφής:

$$H_0\colon \mu = \mu_0$$

$$H_1\colon \mu \neq \mu_0$$

όπου $\mu_0 = 1$*. Υποθέτοντας ότι οι 40 μετρήσεις προέρχονται από την κανονική κατανομή και επειδή δεν γνωρίζουμε την διακύμανση του πληθυσμού, πρέπει να χρησιμοποιήσουμε τον παραμετρικό έλεγχο t-test. Η ελεγχοσυνάρτηση που θα χρησιμοποιήσουμε είναι συνεπώς η*

$$\mathrm{t} = \frac{\overline{\mathrm{X}} - \mu_0}{\mathrm{S}/\sqrt{\mathrm{n}}}$$

Για να υπολογίσουμε την τιμή της ελεγχοσυνάρτησης θα πρέπει να υπολογίσουμε τη μέση τιμή μ και την δειγματική τυπική απόκλιση S. Το μέγεθος του δείγματος ισούται με n=40.

$$\overline{\mathrm{X}} = \frac{1}{40}\sum_{i=1}^{40} X_i = 1{,}069$$

$$\mathrm{S} = \sqrt{\frac{1}{40-1}\sum_{\mathrm{i}=1}^{40}(\mathrm{x_i} - \overline{\mathrm{X}})^2} = \sqrt{\frac{1}{39}\sum_{\mathrm{i}=1}^{40}(\mathrm{x_i} - 1{,}069)^2} \approx \sqrt{0{,}091425} = 0{,}30237$$

$$\mathrm{t} = \frac{\overline{\mathrm{X}} - \mu_0}{\mathrm{S}/\sqrt{\mathrm{n}}} = \frac{1{,}069 - 1}{0{,}30237/\sqrt{40}} \approx 1{,}443$$

Η κρίσιμη περιοχή του ελέγχου είναι η $|t| \geq t_{n-1;a/2}$. *Απορρίπτουμε δηλαδή την μηδενική υπόθεση όταν η απόλυτη τιμή της στατιστικής συνάρτησης t είναι μεγαλύτερη ή ίση από την τιμή* $t_{n-1;a/2}$. *Επειδή ισχύει* $\mathrm{t} = 1{,}443 < t_{n-1;a/2} = t_{40-1;0{,}05/2} = t_{39;0{,}025} = 2{,}023$ *δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση. Δηλαδή τα δεδομένα μας συνηγορούν στο γεγονός ότι ο λόγος βάρους των παιδιών που θα γεννηθούν κατά τη διάρκεια της χρονιάς ισούται με 1.*

3. Ο πληθυσμός ακολουθεί μια άγνωστη κατανομή, αλλά το μέγεθος του δείγματος είναι μεγάλο

Σε περίπτωση που ο πληθυσμός δεν ακολουθεί την κανονική κατανομή αλλά κάποια κατανομή την οποία δεν γνωρίζουμε και το μέγεθος του δείγματος είναι «μεγάλο» (είναι μεγαλύτερο από το 30), μπορούμε να χρησιμοποιήσουμε τη μεθοδολογία του z-test, κάνοντας χρήση του Κεντρικού Οριακού Θεωρήματος. Βάσει του θεωρήματος αυτού, η στατιστική συνάρτηση

$$\mathrm{t} = \frac{\overline{\mathrm{X}} - \mu_0}{\mathrm{S}/\sqrt{\mathrm{n}}} \xrightarrow{approx.} \mathrm{N}(0,1)$$

3.3.2 Έλεγχοι Ποσοστών

Στην ενότητα αυτή θα ασχοληθούμε με τη στατιστική συμπερασματολογία στην περίπτωση που έχουμε στη διάθεσή μας δίτιμες κατηγορικές μεταβλητές. Αρχικά θα ασχοληθούμε με την περίπτωση που έχουμε ένα δείγμα, οπότε και επιθυμούμε να ελέγξουμε εάν το δείγμα μας προέρχεται από έναν πληθυσμό, για τον οποίο γνωρίζουμε ότι ένα μέρος του έχει κάποιο χαρακτηριστικό. Για παράδειγμα θέλουμε να ελέγξουμε το ποσοστό των ατόμων ενός πληθυσμού που πάσχει από πνευμονιοκοκκική πνευμονία. Στη συνέχεια θα ασχοληθούμε με την περίπτωση που έχουμε στη διάθεσή μας δυο δείγματα τα μέλη των οποίων μπορούν να καταταχθούν σε δυο κατηγορίες. Για παράδειγμα, θέλουμε να ελέγξουμε εάν το ποσοστό των φοιτητών του τμήματος Νοσηλευτικής που πάσχουν από δυσλεξία είναι το ίδιο με το ποσοστό των φοιτητών του ίδιου τμήματος που δεν πάσχουν από δυσλεξία.

Στην ενότητα αυτή, θα θεωρήσουμε ότι έχουμε ένα δείγμα και επιθυμούμε να ελέγξουμε εάν το δείγμα μας προέρχεται από έναν πληθυσμό, για τον οποίο γνωρίζουμε ότι ένα μέρος του έχει κάποιο χαρακτηριστικό. Πιο συγκεκριμένα, στόχος μας είναι η στατιστική συμπερασματολογία για το διωνυμικό ποσοστό. Για παράδειγμα θέλουμε να ελέγξουμε εάν το ποσοστό των εφήβων στην Ελλάδα που έχει κάνει χρήση ναρκωτικών για τουλάχιστον μια φορά είναι 12%.

Θεωρούμε ότι έχουμε n ανεξάρτητες δοκιμές Bernoulli με πιθανότητα επιτυχίας p και έστω ότι x συμβολίζει τον αριθμό των επιτυχιών. Δοκιμή Bernoulli (Bernoulli trial) είναι ένα πείραμα τύχης με δυο δυνατά αποτελέσματα, για παράδειγμα το στρίψιμο ενός νομίσματος, όπου τα δυο δυνατά αποτελέσματα είναι η «κορώνα» και τα «γράμματα».

Στην περίπτωση αυτή η πιθανότητα επιτυχίας (διωνυμικό ποσοστό) p εκτιμάται από την ποσότητα $\hat{p} = \frac{X}{n}$ η οποία έχει σφάλμα $\hat{\sigma}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ Διαιρούμε δηλαδή τον αριθμό των επιτυχιών προς το μέγεθος του δείγματος. Οι υποθέσεις που θέλουμε να ελέγξουμε είναι οι

$H_0: p \leq p_0$	$H_0: p \geq p_0$	$H_0: p = p_0$
$H_1: p > p_0$	$H_1: p < p_0$	$H_1: p \neq p_0$

Η στατιστική συνάρτηση χρησιμοποιούμε σε αυτή την περίπτωση είναι η

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

η οποία όταν ισχύει η Μηδενική Υπόθεση ακολουθεί την Τυπική κανονική κατανομή $N(0,1)$. Εναλλακτικά ή στατιστική συνάρτηση μπορεί να γραφεί ως

$$Z = \frac{X - np_0}{\sqrt{np_0(1-np_0)}}$$

όπου X συμβολίζει τον αριθμό επιτυχιών. Οι αντίστοιχες κρίσιμες περιοχές είναι

$$Z \geq z_a\,, \qquad Z \leq -z_a, \qquad |Z| \geq z_{a/2}$$

Αυτό σημαίνει ότι απορρίπτουμε την μηδενική υπόθεση H_0: $p \leq p_0$ έναντι της εναλλακτικής H_1: $p>p_0$ εάν η τιμή της στατιστικής συνάρτησης Z είναι μεγαλύτερη ή ίση από την τιμή z_a. Αντίστοιχα, απορρίπτουμε την μηδενική υπόθεση H_0: $p \geq p_0$ έναντι της εναλλακτικής H_1: $p<p_0$ εάν η τιμή της στατιστικής συνάρτησης Z είναι μικρότερη ή ίση από την τιμή $-z_a$. Τέλος, απορρίπτουμε την μηδενική υπόθεση H_0: $p = p_0$ έναντι της εναλλακτικής H_1: $p \neq p_0$ εάν η στατιστική συνάρτηση Z έχει τιμή που είναι μικρότερη από την τιμή $-z_{a/2}$ ή μεγαλύτερη από την τιμή $z_{a/2}$.

Παράδειγμα

Ο σύλλογος εργαζομένων σε μια μεγάλη νοσηλευτική μονάδα, ισχυρίζεται ότι το ποσοστό των υπαλλήλων του νοσοκομείου που έρχονται σε επαφή με αίμα ή υποπροϊόντα αυτού και είναι θετικοί σε ηπατίτιδα Β είναι μεγαλύτερο από 15%. Μια ερευνητική ομάδα δείχνει ενδιαφέρον για το γεγονός αυτό καθώς εάν αποδειχθεί ο ισχυρισμός του συλλόγου θα πρέπει να ληφθούν δραστικά μέτρα. Η ιατρική ομάδα έλαβε ένα τυχαίο δείγμα 50 εργαζομένων που έρχονται σε συχνή επαφή με αίμα ή υποπροϊόντα αυτού και τους εξέτασε σχετικά με την ηπατίτιδα Β. Τα αποτελέσματα των 50 υπαλλήλων έχουν ως εξής (με Θ συμβολίζουμε το γεγονός ότι ο υπάλληλος βρέθηκε θετικός και με Α ότι βρέθηκε αρνητικός):

Θ	*Α*	*Α*	*Α*	*Α*	*Α*	*Α*	*Α*	*Θ*	*Α*
Α	*Θ*	*Α*	*Α*	*Θ*	*Α*	*Α*	*Θ*	*Α*	*Α*
Θ	*Θ*	*Α*	*Θ*	*Α*	*Α*	*Α*	*Α*	*Α*	*Θ*
Α	*Θ*	*Α*	*Α*	*Α*	*Α*	*Α*	*Θ*	*Α*	*Θ*
Θ	*Α*	*Θ*	*Θ*	*Α*	*Α*	*Θ*	*Α*	*Α*	*Θ*

Να ελεγχθεί σε επίπεδο σημαντικότητας 5% ο ισχυρισμός του συλλόγου των εργαζομένων.

Λύση

Προκειμένου τα μέλη της ιατρικής ομάδας να εξάγουν συμπέρασμα σχετικά με την ορθότητα ή μη του ισχυρισμού του συλλόγου των εργαζομένων, θα πρέπει να εφαρμόσουν τον κατάλληλο έλεγχο υπόθεσης. Η μηδενική και η εναλλακτική υπόθεση του ελέγχου είναι οι

$$H_0 : p \leq 0{,}15$$

$$H_1 : p > 0{,}15$$

Δηλαδή, με βάση το δείγμα θα αποφασίσουμε εάν οι υπάλληλοι της συγκεκριμένης νοσοκομειακής μονάδας προσβάλλονται από ηπατίτιδα Β σε ποσοστό μεγαλύτερο του 15% (p0=0,15).

Η στατιστική συνάρτηση που θα χρησιμοποιήσουμε είναι η

$$Z = \frac{X - np_0}{\sqrt{np_0(1 - np_0)}}$$

όπου Χ είναι ο αριθμός των ατόμων που βρέθηκαν θετικά στην ηπατίτιδα Β. Από τα δεδομένα παρατηρούμε ότι Χ=17 (βρέθηκαν 17 άτομα θετικά στην ηπατίτιδα Β), n=50 και p_0=0,15. Η στατιστική συνάρτηση παίρνει την τιμή

$$Z = \frac{X - np_0}{\sqrt{np_0(1 - np_0)}} = \frac{17 - 50 * 0{,}15}{\sqrt{50 * 0{,}15(1 - 50 * 0{,}15)}} = \frac{17 - 7.5}{\sqrt{6.375}} = \frac{9.5}{\sqrt{6.375}} \approx 3.762$$

Για να εξάγουμε συμπέρασμα ως προς το ποσοστό των υπαλλήλων του νοσοκομείου που έρχονται σε επαφή με αίμα ή υποπροϊόντα αυτού και είναι θετικοί σε ηπατίτιδα Β, θα πρέπει να συγκρίνουμε την τιμή της ελεγχοσυνάρτησης με την τιμή $z_a = z_{0.05} = 1.645$. *Παρατηρούμε ότι η τιμή Ζ είναι μεγαλύτερη από την τιμή 1,645. Αυτό πρακτικά σημαίνει ότι η μηδενική υπόθεση απορρίπτεται, δηλαδή τα δεδομένα μας συνηγορούν στο γεγονός ότι στη συγκεκριμένη νοσηλευτική μονάδα το ποσοστό των υπαλλήλων που είναι θετικοί σε ηπατίτιδα Β είναι μεγαλύτερο από 15%.*

Εφαρμογή με το IBM SPSS Statistics

Για να ελέγξουμε εάν η μέση τιμή ενός πληθυσμού ισούται με μια συγκεκριμένη τιμή, μέσω του IBM SPSS Statistics, από το μενού ***Analyze*** *επιλέγουμε* ***Compare Means*** *και στη συνέχεια το* ***One-Sample T-Test****.*

Για να ελέγξουμε εάν το ποσοστό των ατόμων με ένα συγκεκριμένο χαρακτηριστικό είναι ίσο με μια προκαθορισμένη τιμή, μέσω το IBM SPSS Statistics, από το μενού ***Analyze*** *επιλέγουμε* ***Nonparametric Tests, Legacy Dialogs*** *και στη συνέχεια το* ***Binomial****.*

3.4 Στατιστικοί Έλεγχοι για δύο ανεξάρτητα δείγματα

3.4.1 Εισαγωγή

Στην περίπτωση αυτή κατά την οποία βρισκόμαστε, έχουμε τις τιμές μιας μεταβλητής, σε δυο διαφορετικές και ανεξάρτητες πληθυσμιακές ομάδες. Στόχος είναι να βρεθεί εάν διαφέρουν σημαντικά οι μέσες τιμές της μεταβλητής στους δυο ανεξάρτητους πληθυσμούς.

3.4.2 Θεωρία

Έστω τυχαίο δείγμα μεγέθους n_1 από πληθυσμό με μέση τιμή μ_1 και διακύμανση ${\sigma_1}^2$ και έστω ένα δεύτερο τυχαίο δείγμα μεγέθους n_2 από έναν δεύτερο πληθυσμό με μέση τιμή μ_2 και διακύμανση ${\sigma_2}^2$.Στην συνέχεια, υπολογίζουμε την δειγματική μέση τιμή του πρώτου δείγματος $\overline{X}_1$και την δειγματική μέση τιμή του δεύτερου δείγματος $\overline{X}_2$ καθώς και τις αντίστοιχες διακυμάνσεις τους ${S_1}^2$ και ${S_2}^2$. Για τον έλεγχο της διαφοράς των δυο μέσων τιμών διακρίνουμε τρεις περιπτώσεις:

1. Οι δυο πληθυσμοί ακολουθούν την κανονική κατανομή με γνωστές διακυμάνσεις
2. Οι δυο πληθυσμοί ακολουθούν την κανονική κατανομή με άγνωστες διακυμάνσεις
3. Οι πληθυσμοί ακολουθούν μια άγνωστη κατανομή αλλά τα μεγέθη των δυο δειγμάτων είναι μεγάλα.

Αναλυτικά,

1. Οι δυο πληθυσμοί ακολουθούν την κανονική κατανομή με γνωστές διακυμάνσεις

Θα ελέγξουμε τις παρακάτω υποθέσεις:

$H_0: \mu_1 \leq \mu_2$	$H_0: \mu_1 \geq \mu_2$	$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 > \mu_2$	$H_1: \mu_1 < \mu_2$	$H_1: \mu_1 \neq \mu_2$

Χρησιμοποιώντας την στατιστική συνάρτηση

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}} \xrightarrow{H_0} N(0,1)$$

Οι αντίστοιχες κρίσιμες περιοχές θα είναι οι:

$$Z \geq z_a\,, \qquad Z \leq -z_a, \qquad |Z| \geq z_{a/2}$$

Αυτό σημαίνει ότι απορρίπτουμε την μηδενική υπόθεση $H_0: \mu_1 \leq \mu_2$ έναντι της εναλλακτικής $H_1: \mu_1 < \mu_2$ εάν η τιμή της στατιστικής συνάρτησης Z είναι μεγαλύτερη ή ίση από την τιμή z_a. Αντίστοιχα, απορρίπτουμε την μηδενική υπόθεση $H_0: \mu_1 \geq \mu_2$ έναντι της εναλλακτικής $H_1: \mu_1 < \mu_2$ εάν η τιμή της στατιστικής συνάρτησης Z είναι μικρότερη ή ίση από την τιμή $-z_a$.Τέλος απορρίπτουμε την μηδενική υπόθεση $H_0: \mu_1 = \mu_2$ έναντι της εναλλακτικής $H_1: \mu_1 \neq \mu_2$ εάν η στατιστική συνάρτηση Z *έχει* τιμή που είναι μικρότερη από την τιμή $-z_{a/2}$ ή μεγαλύτερη από την τιμή $z_{a/2}$. Ο έλεγχος αυτός είναι γνωστός ως *z-test για δυο ανεξάρτητα δείγματα.*

Παράδειγμα 3.4

Μια ιατρική ομάδα θεωρεί ότι ένα νέο φάρμακο βοηθά στο να ελαττωθεί το σάκχαρο σε άνδρες ηλικίας μεγαλύτερης των 70 ετών. Η ιατρική ομάδα επέλεξε τυχαία 32 άνδρες με τα ίδια γενικά χαρακτηριστικά και χορήγησε στους πρώτους 16 για ένα μήνα το νέο φάρμακο. Αφού πέρασε η περίοδος του ενός μηνός, μετρήθηκε το σάκχαρο. Κατά το ίδιο χρονικό διάστημα, στους υπόλοιπους 16 άνδρες χορηγήθηκε το συνήθως χορηγούμενο φάρμακο και στο τέλος μετρήθηκε και σε αυτούς το σάκχαρο. Από παλαιότερες μελέτες η ιατρική ομάδα γνωρίζει ότι η διακύμανση των δυο πληθυσμών ισούται με 0,55. Οι μετρήσεις των 32 ανδρών δίνονται στον παρακάτω πίνακα. Να ελεγχθεί εάν υπάρχει μεταβολή λόγω του φαρμάκου.

Σύνηθες Φάρμακο		*Νέο Φάρμακο*	
Άτομο	*Τιμή*	*Άτομο*	*Τιμή*
1	0,61	1	1,84
2	1,42	2	1,57
3	1,40	3	1,85
4	0,54	4	0,80
5	1,98	5	1,68
6	0,82	6	0,96
7	2,01	7	1,41
8	1,34	8	2,21
9	2,34	9	0,72
10	1,80	10	1,49
11	1,56	11	0,84
12	2,35	12	2,26
13	1,89	13	1,73
14	2,25	14	1,84
15	1,24	15	2,14
16	2,26	16	1,76

Λύση

Προκειμένου τα μέλη της ιατρικής ομάδας να εξάγουν συμπεράσματα για το συγκεκριμένο. Υ συγκεκριμένο πρόβλημα απαιτείται η χρήση ενός ελέγχου υπόθεσης για δυο πληθυσμούς της μορφής

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

Συνεπώς, με βάση το δείγμα θα αποφασίσουμε εάν ο πληθυσμός των ανδρών στους οποίους χορηγήθηκε το νέο φάρμακο έχει την ίδια μέση τιμή σακχάρου με τον πληθυσμό των ανδρών στους οποίους χορηγήθηκε το συνήθως χορηγούμενο φάρμακο (δηλαδή $\mu_1 = \mu_2$). Υποθέτοντας ότι τα δεδομένα προέρχονται από την κανονική κατανομή με γνωστή διακύμανση, θα χρησιμοποιήσουμε *τον κατάλληλο παραμετρικό έλεγχο, ο οποίος είναι ο λεγόμενος έλεγχος z (z-test) για δυο ανεξάρτητα δείγματα. Η στατιστική συνάρτηση που χρησιμοποιούμε είναι η*

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

Για το πρώτο δείγμα υπολογίζουμε

$$\overline{X}_1 = \frac{1}{16}\sum_{i=1}^{16} X_{1_i} = 1{,}6131$$

Και

$$\overline{X}_2 = \frac{1}{16}\sum_{i=1}^{16} X_{2_i} = 1{,}5688$$

Γνωρίζουμε επίσης ότι $n_1 = n_2 = 16$ *και* ${\sigma_1}^2 = {\sigma_2}^2 = 0{,}55$

Συνεπώς η στατιστική συνάρτηση παίρνει την τιμή

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}} = \frac{1{,}6131 - 1{,}5688}{\sqrt{\frac{0{,}55}{16} + \frac{0{,}55}{16}}} = \frac{1{,}6131 - 1{,}5688}{\sqrt{\frac{0{,}55}{16} + \frac{0{,}55}{16}}} = \frac{0.0443}{\sqrt{2 * \frac{0{,}55}{16}}} = \frac{0.0443}{0.2622} = 0.169$$

Για να εξάγουμε συμπέρασμα εάν το νέο φάρμακο βοηθά στο να ελαττωθεί το σάκχαρο, θα πρέπει να συγκρίνουμε την τιμή της στατιστικής συνάρτησης με την τιμή $z_{a/2} = z_{0.05/2} = z_{0.025} = 1.96$ *. Παρατηρούμε ότι*

$$Z = 0.169 < z_{a/2} = 1.96$$

Αυτό πρακτικά σημαίνει ότι ***η μηδενική υπόθεση δεν μπορεί να απορριφθεί.***

2. *Οι δυο πληθυσμοί ακολουθούν την κανονική κατανομή με άγνωστες διακυμάνσεις*

Έστω ότι έχουμε δύο πληθυσμούς n_1 και n_2 αντίστοιχα. Υποθέτουμε ότι οι δύο πληθυσμοί ακολουθούν την κανονική κατανομή και ότι οι διακυμάνσεις τους ${\sigma_1}^2$ και ${\sigma_2}^2$ είναι άγνωστες. Αυτή είναι η περίπτωση που συναντάται πιο συχνά στην πράξη καθώς τις περισσότερες φορές δεν γνωρίζουμε την διακύμανση των πληθυσμών. Πριν συνεχίσουμε, θα πρέπει να διακρίνουμε δυο περιπτώσεις. Η πρώτη είναι ότι οι διακυμάνσεις είναι μεν άγνωστες αλλά θεωρούμε ότι είναι ίσες, ενώ η δεύτερη είναι ότι οι διακυμάνσεις είναι άνισες μεταξύ τους.

Άγνωστες και ίσες διακυμάνσεις

Έστω ότι ${\sigma_1}^2 = {\sigma_2}^2 = \sigma^2$, Η καλύτερη εκτιμήτρια συνάρτηση για την κοινή διακύμανση είναι η

$$S^2 = \frac{(n_1 - 1){S_1}^2 + (n_2 - 1){S_2}^2}{n_1 + n_2 - 2}$$

η οποία έχει $n_1 + n_2 - 2$ βαθμούς ελευθερίας και ονομάζεται **αθροιστικός εκτιμητής της διακύμανσης** *(pooled estimator of variance).* Το εκτιμώμενο τυπικό σφάλμα του εκτιμητή $\overline{X}_1 - \overline{X}_2$ της διαφοράς των μέσων είναι η $S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$. Για να ελέγξουμε τις υποθέσεις που είναι ίδιες με την προηγούμενη περίπτωση, χρησιμοποιούμε την στατιστική συνάρτηση.

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \overset{H_0}{\rightarrow} t_{n_1+n_2-2}$$

Οι αντίστοιχες κρίσιμες περιοχές είναι οι

$$t \geq t_{n_1+n_2-2;a}\ , \qquad t \leq -t_{n_1+n_2-2;a}, \qquad |t| \geq t_{n_1+n_2-2;a/2}$$

Αυτό σημαίνει ότι απορρίπτουμε την μηδενική υπόθεση $H_0: \mu_1 \leq \mu_2$ έναντι της εναλλακτικής $H_1: \mu_1 < \mu_2$ εάν η τιμή της στατιστικής συνάρτησης *t* είναι μεγαλύτερη ή ίση από την τιμή $t_{n_1+n_2-2;a}$. Αντίστοιχα, απορρίπτουμε την μηδενική υπόθεση $H_0: \mu_1 \geq \mu_2$ έναντι της εναλλακτικής $H_1: \mu_1 < \mu_2$ εάν η τιμή της στατιστικής συνάρτησης *t* είναι μικρότερη ή ίση από την τιμή $-t_{n_1+n_2-2;a}$.Τέλος απορρίπτουμε την μηδενική υπόθεση $H_0: \mu_1 = \mu_2$ έναντι της εναλλακτικής $H_1: \mu_1 \neq \mu_2$ εάν η στατιστική συνάρτηση *t έχει* τιμή που είναι μικρότερη από την τιμή $-\ t_{n_1+n_2-2;a/2}$ ή μεγαλύτερη από την τιμή $t_{n_1+n_2-2;a/2}$. *Ο παραπάνω έλεγχος είναι γνωστός και ως* ***t-test για δυο ανεξάρτητα δείγματα***. Θα εξετάσουμε το προηγούμενο παράδειγμα 3.4 με την μόνη διαφορά ότι δεν έχουμε πληροφορία για την διακύμανση των δύο ανεξάρτητων δειγμάτων και θα χρειαστεί να τις *εκτιμήσουμε*. Θα εφαρμόσουμε *t-test για δυο ανεξάρτητα δείγματα*

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

Με βάση το δείγμα θα αποφανθούμε αν οι μέσοι των δυο πληθυσμοί διαφέρουν μεταξύ τους, υποθέτοντας ότι τα δεδομένα ακολουθούν την κανονική κατανομή. Η στατιστική συνάρτηση που θα χρησιμοποιήσουμε θα είναι η

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

με

$$S^2 = \frac{(n_1 - 1){S_1}^2 + (n_2 - 1){S_2}^2}{n_1 + n_2 - 2}, \qquad n_1 + n_2 - 2 > 0\ \vee S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \neq 0]$$

Υπολογίζουμε τις παρακάτω ποσότητες

Για το πρώτο δείγμα

$$\overline{X}_1 = \frac{1}{16}\sum_{i=1}^{16} X_{1_i} = 1{,}6131$$

$$S_1{}^2 = \frac{\sum_{i=1}^{n_1=16}\left(X_i - \overline{X}_1\right)^2}{n_1 - 1} = \frac{(0.61 - 1.6131)^2 + \cdots + (2.26 - 1.6131)^2}{16 - 1} = \frac{5.3853}{15} \approx 0.3590$$

Για το δεύτερο δείγμα

$$\overline{X}_2 = \frac{1}{16}\sum_{i=1}^{16} X_{2_i} = 1{,}5688$$

$$S_2{}^2 = \frac{\sum_{i=1}^{n_2=16}\left(X_i - \overline{X}_1\right)^2}{n_1 - 1} = \frac{(0.61 - 1{,}5688)^2 + \cdots + (2.26 - 1{,}5688)^2}{16 - 1} = \frac{3.7609}{15} \approx 0.2507$$

Κοινή τυπική απόκλιση

$$S^2 = \frac{(n_1 - 1)S_1{}^2 + (n_2 - 1)S_2{}^2}{n_1 + n_2 - 2} = S^2 = \frac{(16-1) * 0.3590 + (16-1) * 0.2507}{16 + 16 - 2} = 0{,}3049$$

$$S = \sqrt{\frac{(n_1 - 1)S_1{}^2 + (n_2 - 1)S_2{}^2}{n_1 + n_2 - 2}} = \sqrt{0{,}349} = 0{,}5522$$

Συνεπώς η στατιστική συνάρτηση θα πάρει την τιμή

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{1{,}6131\ - 1{,}5688}{0{,}5522\sqrt{\frac{1}{16} + \frac{1}{16}}} \approx 0.22731$$

Συγκρίνουμε αυτή την τιμή που μόλις βρήκαμε από την στατιστική συνάρτηση t *με την τιμή της* $t_{n_1+n_2-2;a/2} = t_{16+16-2;0.05/2} = t_{30;0.025} = 2.042$ *και βρίσκουμε ότι*

$$t = 0.22731 < t_{n_1+n_2-2;a/2} = t_{30;0.025} = 2.042$$

Αυτό πρακτικά σημαίνει ότι ***η μηδενική υπόθεση δεν μπορεί να απορριφθεί***.

3. Οι δυο πληθυσμοί ακολουθούν μια άγνωστη κατανομή, με άγνωστες διακυμάνσεις

Στην περίπτωση αυτή μπορούμε να χρησιμοποιήσουμε το Κεντρικό Οριακό Θεώρημα και στην συνέχεια να χρησιμοποιήσουμε την μεθοδολογία του z-test προσεγγιστικά.

Έλεγχος Διασπορών

Από τα παραπάνω παρατηρούμε ότι η στατιστική συνάρτηση που χρησιμοποιούμε στον έλεγχο αυτό, εξαρτάται από το εάν οι διασπορές των δυο πληθυσμών είναι ίσες ή άνισες. Για το λόγο αυτό πριν την εφαρμογή του ελέγχου t για δυο ανεξάρτητα δείγματα, θα πρέπει πρώτα να ελέγχουμε την υπόθεση της ισότητας των δυο διακυμάνσεων. Ένας γνωστός έλεγχος ομογένειας των διακυμάνσεων, είναι ο έλεγχος του **Levene** ίσων Διασπορών. Γενικά, ο έλεγχος του Levene (Levene, 1960) χρησιμοποιείται για να ελέγξει εάν k δείγματα έχουν ίσες διακυμάνσεις. Η ισότητα των διακυμάνσεων ανάμεσα σε δείγματα καλείται ομογένεια ή ομοιογένεια (homogeneity) των διακυμάνσεων. Η μηδενική υπόθεση του ελέγχου είναι η

$$H_0: \sigma_1 = \sigma_2 = \cdots = \sigma_k$$

$$H_1: \sigma_i \neq \sigma_j$$

Για τουλάχιστον ένα ζεύγος (i,j).

Στην εναλλακτική υπόθεση αρκεί και μονό ένα ζεύγος να διαφέρει στατιστικά για να μην ισχύει η H_0.