ΔΙΑΣΤΗΜΑΤΑ ΕΜΠΙΣΤΟΣΥΝΗΣ ΓΙΑ ΤΗΝ ΔΙΑΦΟΡΑ ΜΕΣΩΝ ΤΙΜΩΝ ΚΑΝΟΝΙΚΩΝ ΠΛΗΘΥΣΜΩΝ

Έστω $X_1, X_2, ..., X_n$ και $Y_1, Y_2, ..., Y_m$ δύο τυχαία δείγματα μεγέθους n και m αντίστοιχα από δύο ανεξάρτητους κανονικούς πληθυσμούς $N(\mu_X, \sigma_X^2)$ και $N(\mu_Y, \sigma_Y^2)$. Ενδιαφερόμαστε να κατασκευάσουμε ένα διάστημα εμπιστοσύνης για τη διαφορά μ_X - μ_Y.

Η προσέγγιση που ακολουθείται για την κατασκευή αυτού του διαστήματος εμπιστοσύνης βασίζεται συνήθως σε ανεξάρτητα δείγματα, δηλαδή σε δείγματα που επιλέγονται με τρόπο ώστε οι παρατηρήσεις τους ενός να είναι ανεξάρτητες των παρατηρήσεων του άλλου.

Πολλές φορές όμως, όπως θα δούμε στην συνέχεια, προκειμένου να ελαχιστοποιήσουμε την επίδραση εξωτερικών παραγόντων, επιλέγουμε τα τυχαία δείγματα με τρόπο που οι παρατηρήσεις του ενός "συνδέονται" με τις παρατηρήσεις του άλλου σχηματίζοντας ζεύγη. Αυτό, για παράδειγμα, συμβαίνει στην περίπτωση που έχουμε δείγματα για εισοδήματα συζύγων ή αδελφών ή στις περιπτώσεις που έχουμε να συγκρίνουμε τις πωλήσεις δύο υποκαταστημάτων μιας αλυσίδας καταστημάτων σε δύο διαφορετικές τοποθεσίες.

Στις επόμενες τρεις ενότητες (Α, Β, Γ) εξετάζουμε την περίπτωση των ανεξαρτήτων δειγμάτων και, στην συνέχεια, στην ενότητα που ακολουθεί (Δ), την περίπτωση δειγμάτων που οι παρατηρήσεις τους έχουν επιλεγεί κατά ζεύγη.

Α. Περίπτωση Γνωστών Διακυμάνσεων
(Ανεξάρτητα Δείγματα)

Όπως είναι γνωστό, μια σημειακή εκτιμήτρια για την διαφορά μ_X-μ_Y είναι το $\overline{X} - \overline{Y}$ όπου $\overline{X}$, $\overline{Y}$ είναι οι δειγματικοί μέσοι των δύο δειγμάτων.

Όπως γνωρίζουμε,

$$\overline{X} \sim N(\mu_X,\ \sigma_X^2/n)$$

και

$$\overline{Y} \sim N(\mu_Y,\ \sigma_Y^2/m)$$

Επομένως,

$$\overline{X} - \overline{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}\right)$$

ή ισοδύναμα,

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}} \sim N(0, 1)$$

Συνεπώς,το 100(1-α)% διάστημα εμπιστοσύνης για το μ_X - μ_Y είναι το

$$\boxed{\overline{X} - \overline{Y} \pm Z_{1-\alpha/2}\sqrt{\sigma_X^2/n + \sigma_Y^2/m}}$$

Αυτό είναι ένα ακριβές διάστημα εμπιστοσύνης.

Παρατήρηση: Αν οι πληθυσμοί δεν είναι ακριβώς κανονικοί αλλά τα μεγέθη των δειγμάτων n και m είναι αρκετά μεγάλα, το διάστημα εμπιστοσύνης που προαναφέρθηκε μπορεί, λόγω του κεντρικού οριακού θεωρήματος, να χρησιμοποιηθεί ως ένα κατά προσέγγιση διάστημα εμπιστοσύνης για τη διαφορά μ_X - μ_Y .

Παράδειγμα: Προκειμένου να ελεγχθεί η ποιότητα δύο ειδών ελαστικών αυτοκινήτων, έγινε ένας έλεγχος σε 100 ελαστικά από κάθε είδος τυχαία επιλεγμένα (n=m=100). Ως στοιχείο ποιότητας χρησιμοποιήθηκε ο αριθμός των χιλιομέτρων που τα ελαστικά αυτά χρησιμοποιήθηκαν μέχρις ότου φθάσουν σε ένα συγκεκριμένο σημείο φθοράς. Τα αποτελέσματα αυτά (σε χιλιόμετρα) ήταν ως εξής:

$$\overline{x} = 26400 \text{ km} \qquad \overline{y} = 25100 \text{ km}$$

Από προηγούμενη πείρα είναι γνωστό ότι

$$\sigma_X^2 = 1440000 \qquad \sigma_Y^2 = 1960000$$

Να κατασκευασθεί ένα 99% διάστημα εμπιστοσύνης για την διαφορά μ_X-μ_Y .

Λύση: Το ζητούμενο διάστημα είναι

$$\overline{x} - \overline{y} \pm Z_{.995}\sqrt{\sigma_X^2/n + \sigma_Y^2/m}$$

Για το δείγμα μας έχουμε,

$$1300 \pm 2.58\,(184) \quad ή$$

$$(825\ ,\ 1775)$$

Επομένως, με πιθανότητα .99 η διαφορά της μέσης διάρκειας ζωής μεταξύ των δύο αυτών ειδών ελαστικών έχει μια τιμή στο διάστημα (825, 1775).

Β. Περίπτωση Αγνώστων Ίσων Διακυμάνσεων

(Ανεξάρτητα Δείγματα)

Υποθέτουμε και πάλι ότι $X \sim N(\mu_X\ ,\ \sigma_X^2)$ και $Y \sim N(\mu_Y\ ,\ \sigma_Y^2)$ και ότι X, Y είναι ανεξάρτητες. Τότε,

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}} \sim N(0,\,1)$$

Υποθέτουμε ότι $\sigma_X^2 = \sigma_Y^2 = \sigma^2$.

Έστω S_X^2, S_Y^2 οι διασπορές των δύο δειγμάτων αντίστοιχα (S_X^{*2}, S_Y^{*2} οι αντίστοιχες αμερόληπτες εκτιμήτριες).
Λόγω της ανεξαρτησίας, θα έχουμε ότι

$$V = \frac{nS_X^2}{\sigma^2} + \frac{mS_Y^2}{\sigma^2} \sim X^2_{n+m-2}$$

(Οι βαθμοί ελευθερίας της V είναι n-1+m-1 = n+m-2).
Δοθέντος επίσης ότι Z και V είναι ανεξάρτητες έχουμε

$$T = \frac{Z}{\sqrt{V/(n+m-2)}} \sim t_{n+m-2}$$

Από τις προηγούμενες σχέσεις προκύπτει ότι,

$$T = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{S_p^2/n + S_p^2/m}} \sim t_{n+m-2}$$

όπου

$$S_p^2 = \frac{nS_X^2 + mS_Y^2}{n+m-2} = \frac{(n-1)S_X^{*2} + (m-1)S_Y^{*2}}{n+m-2}$$

δηλαδή,

$$\boxed{S_p^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 + \sum_{i=1}^{m}\left(Y_i - \overline{Y}\right)^2}{n+m-2}}$$

Επομένως το 100(1-α)% διάστημα εμπιστοσύνης για την διαφορά μ_X-μ_Y το στη περίπτωση αυτή είναι,

$$\boxed{\overline{X} - \overline{Y} \pm t_{n+m-2,\,1-\alpha/2}\sqrt{S_p^2/n + S_p^2/m}}$$

Παρατήρηση: Η διασπορά S_p^2 ονομάζεται *σταθμισμένη διασπορά* (*pooled variance*). Χρησιμοποιείται δε γιατί έχουμε δύο ανεξάρτητες εκτιμήσεις που προέρχονται από δύο διαφορετικά δείγματα για την ίδια ποσότητα (το σ^2). Είναι λοιπόν φυσικό να κάνουμε χρήση και των δύο αυτών εκτιμήσεων λαμβάνοντας όμως υπόψη μας (και δίνοντας την αντίστοιχη βαρύτητα) στην ποιότητα της καθεμιάς από αυτές (δηλαδή στο πόσο ακριβής είναι η καθεμιά από αυτές με βάση το μέγεθος του δείγματος από το οποίο έχει προέλθει).

Παρατήρηση: Είναι δυνατόν να αποδειχθεί, είτε μαθηματικά είτε πειραματικά, ότι η σταθμισμένη εκτιμήτρια S_p^2 της κοινής διασποράς σ^2 είναι μια αμερόληπτη εκτιμήτρια του σ^2.

Σημείωση: Αν $\sigma_X^2 \neq \sigma_Y^2$ αλλά n, m είναι αρκετά μεγάλα, μπορούμε να αντικαταστήσουμε το σ_X^2 με το $nS_X^2/(n-1)$ και το σ_Y^2 με το

$m S_Y^2/(m-1)$ στην αρχική σχέση (αντίστοιχα το σ_X^2 με το S_X^{*2} και το σ_Y^2 με το S_Y^{*2} δηλαδή, τις αμερόληπτες εκτιμήτριες των σ_X^2 και σ_Y^2 αντίστοιχα). Τότε ένα κατά προσέγγιση 100(1-α)% διάστημα εμπιστοσύνης για την διαφορά $\mu_X - \mu_Y$ είναι το

$$\boxed{\overline{X} - \overline{Y} \pm Z_{1-\alpha/2}\sqrt{S_X^2/(n-1) + S_Y^2/(m-1)}}$$

ή ισοδύναμα, με τις αμερόληπτες εκτιμήτριες των διασπορών,

$$\boxed{\overline{X} - \overline{Y} \pm Z_{1-\alpha/2}\sqrt{S_X^{*2}/n + S_Y^{*2}/m}}$$

Παράδειγμα: Προκειμένου να συγκριθούν οι φοιτητές των τμημάτων Στατιστικής δύο διαφορετικών Πανεπιστημίων, επελέγησαν δύο τυχαία δείγματα από n=50 και m=60 φοιτητές του ιδίου έτους στους οποίους δόθηκε ένα συγκεκριμένο διαγώνισμα. Τα αποτελέσματα στο διαγώνισμα αυτό (σε βαθμολογία με κλίμακα 0-100) ήταν ως εξής:

$$\overline{x} = 77 \qquad \overline{y} = 68$$
$$S_X^2 = 64 \qquad S_Y^2 = 100$$

Να κατασκευασθεί ένα 95% διάστημα εμπιστοσύνης για την διαφορά $\mu_X - \mu_Y$ της πραγματικής μέσης απόδοσης των φοιτητών στα δύο αυτά Πανεπιστήμια.

Λύση: Υποθέτουμε ότι $\sigma_X^2 = \sigma_Y^2$ και ότι η απόδοση στα διαγωνίσματα ακολουθεί την κανονική κατανομή.
Για την σταθμισμένη διασπορά έχουμε,

$$S_p^2 = \frac{50 * 64 + 60 * 100}{108} = 85.18$$

Το 95% διάστημα εμπιστοσύνης για τη διαφορά $\mu_X - \mu_Y$ είναι,

$$9 \pm t_{108,.975}\sqrt{85.18/50 + 85.18/60}$$

ή, ισοδύναμα,

$$(5.60\ ,\ 12.40)$$

(Επειδή στους πίνακες της κατανομής t δεν δίνεται τιμή για 162 βαθμούς ελευθερίας, χρησιμοποιούμε, από τους πίνακες της κανονικής κατανομής, την τιμή $Z_{.975} = 1.96$. Άλλωστε, για τους βαθμούς αυτούς ελευθερίας η κατανομή t προσεγγίζεται σχεδόν πλήρως από την τυποποιημένη κανονική κατανομή).

Σημείωση 1: Μερικές φορές κάποιο από τα άκρα του διαστήματος, ή και τα δύο, μπορεί να είναι αρνητικό πράγμα το οποίο αποτελεί ένδειξη ότι η μέση τιμή του δεύτερου πληθυσμού υπερβαίνει την μέση τιμή του πρώτου πληθυσμού.

Σημείωση 2: Στο προηγούμενο πρόβλημα, δοθέντος ότι τα n και m ήταν μεγάλα, μπορούσαμε να βρούμε ένα κατά προσέγγιση διάστημα εμπιστοσύνης για το $\mu_X - \mu_Y$ χωρίς να υποθέσουμε ότι $\sigma_X^2 = \sigma_Y^2$, χρησιμοποιώντας τον αντίστοιχο τύπο που δώσαμε παραπάνω.

Παρατήρηση: Η χρησιμοποίηση της μεθόδου που αναλύθηκε στην περίπτωση που $\sigma_X^2 = \sigma_Y^2$ απαιτεί την επιβεβαίωση της υπόθεσης αυτής. Όπως θα δούμε αργότερα (στους ελέγχους υποθέσεων), υπάρχει στατιστική μεθοδολογία με την οποία μπορούμε να ελέγξουμε την υπόθεση αυτή.

Παράδειγμα: Σε ένα εργοστάσιο έχει βρεθεί ότι απαιτείται χρόνος περίπου ενός μηνός για να εκπαιδευθεί ένας καινούργιος εργαζόμενος και να φθάσει τη μέγιστη δυνατή απόδοση. Στο εργοστάσιο αυτό έχει προταθεί μια καινούργια μέθοδος εκπαίδευσης για την ελάττωση του χρόνου αυτού. Προκειμένου να ελεγχθεί η αποτελεσματικότητα της μεθόδου αυτής, επιλέγονται δύο ομάδες από 9 εργαζόμενους η καθεμία. Οι εργαζόμενοι αυτοί εκπαιδεύονται για μία περίοδο δύο εβδομάδων, την μιά βδομάδα με χρησιμοποίηση της παλιάς μεθόδου και την άλλη με χρησιμοποίηση της καινούργιας μεθόδου. Στην συνέχεια, ανατίθεται σε καθένα από τους εργαζόμενους που εκπαιδεύτηκαν η συναρμολόγηση μιας συσκευής (εργασία για την οποία εκπαιδεύονται). Οι χρόνοι (σε λεπτά) που απαιτήθηκαν από τον κάθε εργαζόμενο για να συναρμολογήσει το συγκεκριμένο αντικείμενο μετά από το χρόνο της εκπαίδευσης αυτής δίνονται στον πίνακα που ακολουθεί:

ΧΡΟΝΟΣ ΠΟΥ ΑΠΑΙΤΗΘΗΚΕ (σε λεπτά)	
Συνήθης Μέθοδος Εκπαίδευσης X	Νέα Μέθοδος Εκπαίδευσης Y
32	35
37	31
35	29
28	25
41	34
44	40
35	27
31	32
34	31

Να κατασκευασθεί ένα 95% διάστημα εμπιστοσύνης για την διαφορά μεταξύ των μέσων χρόνων που απαιτούνται για την συναρμολόγηση του συγκεκριμένου προϊόντος μετά από περίοδο εκπαίδευσης δύο εβδομάδων για τη συνήθη και την καινούργια μέθοδο εκπαίδευσης.

Λύση: Ας υποθέσουμε ότι μ_X και μ_Y είναι οι μέσοι χρόνοι που απαιτούνται για τη συναρμολόγηση του προϊόντος μετά από εκπαίδευση με τη συνήθη και τη νέα μέθοδο αντίστοιχα.

Ας υποθέσουμε ότι η διακύμανση στους μέσους χρόνους συναρμολόγησης είναι στην πραγματικότητα συνάρτηση των διαφορών που οφείλονται στις προσωπικότητες των εργαζομένων και ότι οι διακυμάνσεις για τις μετρήσεις από τους δύο πληθυσμούς μπορούν να θεωρηθούν κατά προσέγγιση ίσες.

Από τα δύο δείγματα μπορούν να υπολογισθούν τα εξής στοιχεία:

$$\overline{x} = 35.22 \qquad \overline{y} = 31.56 \qquad n = m = 9$$

$$\sum_{i=1}^{n} (x_i - \overline{x})^2 = 195.56 \qquad \sum_{i=1}^{m} (y_i - \overline{y})^2 = 160.22$$

$$S_p^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{m}(y_i - \bar{y})^2}{n + m - 2}$$

$$= \frac{195.56 + 160.22}{9 + 9 - 2}$$

$$= 22.24$$

Επομένως,

$$s_p = \sqrt{22.24} = 4.72$$

Από τους πίνακες της κατανομής t

$$t_{9+9-2, .975} = t_{16, .975} = 2.120$$

Αντικαθιστώντας τις τιμές στον τύπο που δίνει το διάστημα εμπιστοσύνης

$$\overline{X} - \overline{Y} \pm t_{n+m-2, 1-\alpha/2} S_p \sqrt{1/n + 1/m}$$

έχουμε,

$$(35.22 - 31.56) \pm (2.120)\ (4.72)\ \sqrt{2/9}$$

ή ισοδύναμα,

$$3.66 \pm 4.72$$

δηλαδή τελικά,

$$(-1.06\ ,\ 8.38)$$

Επομένως, το εκτιμώμενο 95% διάστημα εμπιστοσύνης για την διαφορά των μέσων χρόνων συναρμολόγησης είναι το (-1.06 , 8.38).