ΚΕΦΑΛΑΙΟ ΙΙ

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ

ΕΝΟΤΗΤΕΣ

1. ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ ΚΑΤΑ ΕΝΑ ΚΡΙΤΗΡΙΟ

2. ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ ΚΑΤΑ ΔΥΟ ΚΡΙΤΗΡΙΑ

1. ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ ΚΑΤΑ ΕΝΑ ΚΡΙΤΗΡΙΟ
(One-Way Analysis of Variance)

Η *ανάλυση διακύμανσης* ή όπως αλλιώς λέγεται, *ανάλυση διασποράς* (*analysis of variance)* είναι μια από τις μεθόδους πειραματικών σχεδιασμών (*experimental design*). Η μεθοδολογία αυτή αποσκοπεί στο να ανιχνεύσει διαφορές μεταξύ των μέσων ορισμένων πληθυσμών.

Η ανάλυση διασποράς είναι η μεθοδολογία εκείνη η οποία ασχολείται με την εξέταση και τον προσδιορισμό των πηγών των αποκλίσεων που παρατηρούνται σε δειγματικά δεδομένα.

Εναλλακτικά, μπορούμε να θεωρήσουμε την ανάλυση διακύμανσης *ως τον διαχωρισμό της επιρροής των διαφορετικών υποσυνόλων των παραμέτρων πάνω στις παρατηρήσεις.*

Σε πειράματα οι παράμετροι είναι συνήθως το αποτέλεσμα κάποιων "επιδράσεων" ("*treatments*") πάνω σε μιά μεταβλητή Υ. Σε αγροτικά πειράματα, για παράδειγμα, (από που έχει προέλθει και ο όρος) Υ μπορεί να είναι η παραγωγή σταριού από κάποιο συγκεκριμένο κομμάτι χωραφιού και η "επίδραση" που μελετάμε να είναι η πρόσθεση κάποιου λιπάσματος στο κομμάτι αυτό του χωραφιού κατά την περίοδο της σποράς. Φυσικά ο ερευνητής στο πείραμα του θα χρησιμοποιήσει κομμάτια του χωραφιού που έχουν υποστεί την "επίδραση" και άλλα που δεν την έχουν υποστεί.

Το σημαντικό είναι ότι, από στατιστικής πλευράς, ένα τέτοιο πείραμα μπορεί να παρουσιασθεί με την μορφή του γενικού γραμμικού μοντέλου με τον ορισμό μιάς "εικονικής" μεταβλητής (*dummy variable*) η οποία να είναι 1 αν η επίδραση έχει εφαρμοσθεί και 0 αν δεν έχει εφαρμοσθεί.

Το πρόβλημα, φυσικά, βρίσκεται στο να αποφασισθεί αν η "επίδραση" (χρήση του λιπάσματος), επιφέρει στατιστικά σημαντική βελτίωση της παραγωγής ή όχι.

Παράδειγμα: Τα στοιχεία που ακολουθούν αναφέρονται στη μέση κατανάλωση βενζίνης (σε μίλια/γαλόνι) τριών ειδών μικρών φορτηγών αυτοκινήτων (TOYOTA, DATSUN, MAZDA). Για να καθορισθεί αν υπάρχει στατιστικά σημαντική διαφορά στη μέση κατανάλωση βενζίνης μεταξύ των τριών αυτών διαφορετικών

αυτοκινήτων σχεδιάστηκε το εξής πείραμα: Χρησιμοποιήθηκαν έξι αυτοκινήτα τις κατηγορίας αυτής μαρκας Toyota, πέντε Datsun , και τέσσερα Mazda. Το κάθε ένα από αυτά τα αυτοκίνητα οδηγήθηκε σε μιά διαδρομή (την ιδια διαδρομη για όλα τα αυτοκίνητα) 300 mil με σταθερή ταχύτητα 55 mph (μιλ/ώρα). Για τις διαδρομές αυτές σημειώθηκε η κατανάλωση βενζίνης σε μίλια/γαλόνι. Τα αποτελέσματα δίνονται στον πίνακα που ακολουθεί:

Toyota	Datsun	Mazda
27.1	25.3	23.1
25.5	26.5	24.3
27.0	26.4	23.4
26.9	26.8	24.2
27.7	26.5	
27.3		

Ενα πρώτο βήμα στην εξέταση των δειγματικών αυτών δεδομένων είναι να αναρωτηθούμε ποιά είναι η αιτία που προκαλεί τη διακύμανση στα αποτελέσματα αυτά. Μια προφανής εξήγηση στο ερώτημα αυτό είναι ότι μέρος της διακύμανσης οφείλεται στο γεγονός ότι αυτοκίνητα διαφορετικών κατασκευαστών έχουν διαφορετική απόδοση εξαιτίας της διαφορετικής κατασκευής. Παρατηρούμε, για παράδειγμα, ότι στο συγκεκριμένο πείραμα όλα τα αυτοκίνητα της Mazda έχουν μικρότερη κατανάλωση από τα αντίστοιχα της Toyota και της Datsun. Παρατηρούμε όμως ότι και μεταξύ αυτοκινήτων του ίδιου κατασκευαστή τα αποτελέσματα διαφέρουν. Και για τη διαφορά αυτή, φυσικά, υπάρχουν πολλές εξηγήσεις. Για τα αυτοκίνητα της Toyota, για παράδειγμα, μπορεί κανείς να πει ότι οι έξι διαφορετικές παρατηρήσεις προήλθαν από έξι διαφορετικά αυτοκίνητα τα οποία ίσως οδηγήθηκαν από διαφορετικούς οδηγούς. Είναι ενδεχόμενο τα αυτοκίνητα αυτά να είχαν διαφορετικά λάστιχα με διαφορετικη πίεση αέρα ή ότι, ακόμη τα αυτοκίνητα αυτά είχαν το τελευταίο σέρβις σε διαφορετικές χρονικές στιγμές ή ότι οι καιρικές συνθηκές όταν έγινε το τεστ για κάθε ένα από τα αυτοκίνητα αυτά ήταν ίσως διαφορετικές κλπ. Είναι, επομένως, προφανές ότι τα πειραματικά δεδομένα περιέχουν ένα μεγάλο αριθμό πηγών διακύμανσης. Ενας καλός πειραματικός σχεδιασμός αποσκοπεί στο να καθορίσει την κύρια πηγή

διακύμανσης όπως επίσης και την ποσότητα της διακύμανσης που οφείλεται σε κάθε ένα από τους διαφορετικούς λόγους που μας ενδιαφέρει να εξετάσουμε. Η υπόλοιπη διακύμανση των δεδομένων θεωρείται ότι οφείλεται σε τυχαίους παράγοντες και για το λόγο αυτό ονομάζεται λάθος (error). Ενας καλός σχεδιασμός ελαττώνει την διακύμανση του λάθους όσο το δυνατόν περισσότερο έτσι ώστε οι διαφορές της διακύμανσης που οφείλονται στους λόγους που μας ενδιαφέρουν να καθορισθούν όσο το δυνατόν ακριβέστερα.

Η ανάλυση διακύμανσης αποσκοπεί ακριβώς στο να καθορίσει όλες τις πηγές που συνεισφέρουν στην διακύμανση και το ποσοστό της διακύμανσης που μπορεί να αποδοθεί σε κάθε μία από τις πηγές αυτές.

Η απλούστερη περίπτωση ανάλυσης διακύμανσης είναι αυτή που ονομάζεται *πλήρως τυχαιοποιημένος σχεδιασμός* (*completely randomized design*).

1.1 Πλήρως Τυχαιοποιημένος Σχεδιασμός (Completely Randomized Design)

Το γενικό πρόβλημα μπορεί να τοποθετηθεί ως εξής: Εχουμε k ανεξάρτητους πληθυσμούς και θέλουμε να ελέξουμε την υπόθεση

$H_0 : \mu_1 = \mu_2,..., = \mu_k$

H_1 : τουλάχιστον δύο μέσοι διαφέρουν.

Υποθέτουμε ότι

$$\sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2 \equiv \sigma^2$$

Παίρνουμε k ανεξάρτητα τυχαία δείγματα από τους k πληθυσμούς. Εάν κατατάξουμε τους πληθυσμούς ανάλογα με την "επίδραση" (κάθε "επίδραση" αντιστοιχει σε ένα πληθυσμό) θα μπορούμε να λέμε ότι οι k "επιδράσεις" μπορεί να αναφέρονται σε k διαφορετικά λιπάσματα, k διαφορετικές περιοχέw μιάς χώρας, k σύνολα βαθμών (από k διαφορετικούς καθηγητές) κ.λ.π.

Εστω ότι παίρνουμε k τυχαία δείγματα (ανεξάρτητα) μεγέθους n_1 ,..., n_κ αντίστοιχα, ένα από κάθε επίδραση.

Τότε θα έχουμε την εξής κατάσταση.

Επίδραση Μεταχείρισης (Treatment)

	1	2	…	i	…	k	
	Y_{11}	Y_{21}	…	Y_{i1}	…	Y_{k1}	
	Y_{12}	Y_{22}	…	Y_{i2}	…	Y_{k2}	
	.		.		.		
	.		.		.		
	.		.		.		
	Y_{ln_1}	Y_{2n_2}	…	Y_{in_i}	…	Y_{kn_k}	
Σύνολα	$Y_{1.}$	$Y_{2.}$	…	$Y_{i.}$	…	$Y_{k.}$	$Y_{..}$
Μέσοι	$\overline{Y}_{1.}$	$\overline{Y}_{2.}$	…	$\overline{Y}_{i.}$	…	$\overline{Y}_{k.}$	$\overline{Y}_{..}$

$$Y_{i.} = \sum_{j=1}^{n_i} Y_{ij} \qquad Y_{..} = \sum_{i=1}^{k} Y_{i.}$$

$$\overline{Y}_i = Y_{i.}/n_i \qquad \overline{Y}_{..} = Y_{..}/N$$

(Δηλαδή Y.. είναι το σύνολο $n_1 + n_2 + ... + n_k = N$ παρατηρήσεων και $\overline{Y}$ .. είναι ο μέσος των παρατηρήσεων αυτών.
Θα έχουμε

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

Το ε_{ij} μετρά την απόκλιση της j-παρατήρησης στο i-δείγμα από τον αντίστοιχο μέσο μ_i της i-επίδρασης. (Δηλαδή το ε_{ij} είναι το τυχαίο λάθος). Αν $\grave{\imath} = (1/k)\sum_i \grave{\imath}_i$, τότε $\mu_i = \mu + \alpha_i$ όπου α_i είναι ένας όρος που εκφράζει το αποτέλεσμα (*effect*) της i επίδρασης.
Δηλαδή, τελικά

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad \sum_{i=1}^{k} \alpha_i = 0 \qquad \varepsilon_{ij} \sim N(0, \sigma^2).$$

Αρα, η αρχική υπόθεση είναι ισοδύναμη με την υπόθεση

$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_k = 0$
$H_1 : \alpha_i \neq 0$ για τουλάχιστον ένα i = 1,2,...k

Η συνολική απόκλιση μετριέται από το άθροισμα

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{..}\right)^2 = \sum_i \sum_j \left[\left(Y_{ij} - \overline{Y}_{i.}\right) + \left(\overline{Y}_{i.} - \overline{Y}_{..}\right)\right]^2$$

$$= \sum_i \sum_j \left(Y_{ij} - \overline{Y}_{i.}\right)^2 + 2\sum_i \sum_j \left(Y_{ij} - \overline{Y}_{i.}\right)\left(\overline{Y}_{i.} - \overline{Y}_{..}\right) + \sum_i \sum_j \left(\overline{Y}_{i.} - \overline{Y}_{..}\right)^2$$

Αυτό μπορεί να αποδειχθεί ότι αναλύεται σε δύο προσθετικές συνιστώσες, ως εξής:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{..}\right)^2 = \sum_{i}^{k} n_i\left(\overline{Y}_{i.} - \overline{Y}_{..}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{i.}\right)^2$$

Είναι:

$$\sum\sum\left(\overline{Y}_{i.} - \overline{Y}_{..}\right)^2 = \sum_i n_i\left(\overline{Y}_{i.} - \overline{Y}_{..}\right)^2 = \frac{\sum Y_{i.}^2}{n_i} - \frac{\left(\overline{Y}_{..}\right)^2}{N}$$

Αν συμβολίσουμε με

$$\text{SST } (\equiv \Sigma\text{AT}) = \sum_{i=1}^{k}\sum_{k=1}^{n_i}\left(Y_{ij} - \overline{Y}_{..}\right)^2$$

την συνολική τετραγωνική απόκλιση, με

$$\text{SST}_r\ (\equiv \Pi\text{AT}) = \sum_i \sum_j \left(\overline{Y}_{i.} - \overline{Y}_{..}\right)^2$$

την συνολική τετραγωνική απόκλιση που οφείλεται στις επιδράσεις (*treatments*) και με

$$\text{SSE } (\equiv \Lambda\text{AT}) = \sum_i\sum_j\left(Y_{ij} - \overline{Y}_i\right)^2$$

την τετραγωνική απόκλιση του λάθους θα έχουμε

$$\text{SST} = \text{SSTr} + \text{SSE}.$$

Η συνολική διακύμανση δηλαδή, αναλύθηκε σε δύο προσθετικές συνιστώσες. Σε αυτήν που μπορεί να αποδοθεί σε διαφορές "μεταξύ" των επιδράσεων (*between treatments*) και σε εκείνη που αποδίδεται σε διάφορες "μέσα" σε κάθε επίδραση (*"within" treatments*), δηλαδή τα τυχαία λάθη.

Η μέθοδος της ανάλυσης διασποράς στηρίζεται στο να βρεί κανείς αν οι παρατηρηθείσες διαφορές ανάμεσα στους μέσους των

διαφορών "επιδράσεων" οφείλονται σε τυχαίους λόγους ή σε συστηματική διαφορά ανάμεσα στις διαφορετικές επιδράσεις.
Αποδεικνύεται ότι

$$E\left(\frac{SSTr}{k-1}\right) = \sigma^2 + \frac{\sum_i n_i \alpha_i^2}{k-1}$$

και επομένως, κάτω από την H_0,

$$E\left(\frac{SSTr}{k-1}\right) = \sigma^2$$

Θέτοντας

$$s_0^2 = \frac{SSTr}{k-1}$$

έχουμε

$$E(S_0^2) = \sigma^2$$

Από το άλλο μέρος κάθε ένα από τα k δείγματα δίνουν μιά εκτίμηση του σ^2 .

$$S_i^{*2} = \frac{1}{n_j - 1}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2$$

Αν θεωρήσουμε την σταθμισμένη "εκτιμήτρια" S_p^2 (*pooled variance*) το σ^2 από τα k δείγματα θα έχουμε:

$$S_p^2 = \frac{(n_1 - 1)S_1^{*2} + (n_2 - 1)S_2^{*2} + ... + (n_k - 1)S_k^{*2}}{N - k}$$

$$= \frac{\sum_{i=1}^{k}(n_i - 1)S_i^{*2}}{N - k}$$

$$= \frac{\sum_{i=1}^{k}\sum_{j=1}(Y_{ij} - \overline{Y}_{i.})^2}{N - k}$$

$$= \frac{SSE}{N - k}$$

(ανεξάρτητα από το αν ισχύει η H_0 ή H_1).
Τότε, αν η H_0 είναι σωστή, έχουμε ότι:

$$F = \frac{S_0^2}{S_p^2} \sim F_{k-1,N-k}$$

Οταν η H_0 δεν είναι σωστή θα πρέπει ο παράγοντας SST_r να είναι μεγάλος (δηλαδή το S_0^2 να είναι μεγάλο σε σχέση με το S_p^2).
Τότε, απορρίπτουμε την H_0 αν

$$F > F_{k-1,\ N-k,\ 1-\alpha}$$

Τα παραπάνω αποτυπώνονται στον πίνακα που ακολουθεί. Ο πίνακας αυτός ονομάζεται *πίνακας ανάλυσης διακύμανσης (διασποράς)* (Analysis of Variance ή ANOVA table).

Πίνακας Ανάλυσης Διακύμανσης (Διασποράς)
ANOVA

Αιτία Διασποράς	Αθροισμα Τετραγώνων SS	Βαθμοί Ελευθερίας DF	Μέσα Τετραγωνικά Λάθη MS	F (κάτω από την H_0)
Μεταξύ Επιδράσεων (between treatments)	SSTr	k - 1	$S_0^2 = \frac{SSTr}{k-1}$	$F = \frac{S_0^2}{S_p^2}$
Μέσα στις Επιδράσεις (λάθος) (within treatments)	SSE	N - k	$S_p^2 = \frac{SSE}{N-k}$	
Σύνολο	SST	N - 1		

Συνήθως χρησιμοποιούμε τους ισοδύναμους τύπους:

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - \frac{Y_{..}^2}{N}$$

$$SSTr = \sum_{i=1}^{k} \frac{Y_{i.}^2}{n_i} - \frac{Y_{..}^2}{N}$$

$$SSE = \sum_{i}\sum_{j} Y_{ij}^2 - \sum_{i} \frac{Y_{i..}^2}{n_i}$$

Παράδειγμα: Μια τάξη 20 μαθητών χωρίστηκε, με τυχαίο τρόπο σε 5 τμήματα με τον σκοπό να μελετηθεί η αποτελεσματικότητα 5 διαφορετικών μεθόδων διδασκαλίας της στατιστικής. Μετά από 6 εβδομάδες οι μαθητές έδωσαν ένα διαγώνισμα. Τα αποτελέσματα (οι βαθμοί) δίνονται παρακάτω. Να εξεταστεί αν οι βαθμοί αυτοί δίνουν κάποια ένδειξη στατιστικά σημαντικής διαφοράς των μεθόδων διδασκαλίας.

Πίνακας Ανάλυσης Διακύμανσης για τη Σύγκριση των Πέντε Μεθόδων Διδασκαλίας

	Μέθοδοι				
	1	2	3	4	5
Βαθμοί	93	73	75	89	59
	97	77	84	81	64
	92	67	80	76	55
	85	76	70	75	67
$y_{i.}$	$y_{1.}$=367	$y_{2.}$=293	$y_{3.}$=309	$y_{4.}$=321	$y_{5.}$=245
$\overline{y}_{i.}$	91.75	73.25	77.25	80.25	61.25
$\sum_{i}\sum_{j}^{n_i} y_{ij}^2$	33,747	22,523	23,981	25,883	15,091

$$SSE = 120225 - \frac{479085}{4} = 453.75$$

$$SSTr = \frac{479085}{4} - \frac{2356225}{20} = 1960$$

Πίνακας ANOVA
(Ανάλυση Διακύμανσης)

Αιτία Διασποράς	SS	Βαθμοί Ελευθερίας	MS	F
Μεταξύ Επιδράσεων	1960	4	4.90	16.20
Μέσα στις Επιδράσεις	453.75	15	30.25	
Σύνολο	2413.75	19		

Στο $\alpha = .10$

$$F_{4,\ 15,\ 0.90} = 2.36$$

Επειδή $F > F_{4,\ 15,\ 0.90}$ απορρίπτουμε την H_0.

Σημείωση: Τα συμπεράσματα ισχύουν με την προϋπόθεση ότι:

$$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$$

και ότι $\varepsilon_{ij} \sim N\ (0,\ \sigma^2)$.

Η υπόθεση που ελέγξαμε ήταν ότι:
$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$
$H_1 : \mu_i \neq \mu_j$ για τουλάχιστον ένα ζευγάρι (i, j).

Ας επανέλθουμε στο πρόβλημα της σύγκρισης της απόδοσης αυτοκινήτων που προέρχονται από τρεις διαφορετικές κατασκευαστικές εταιρίες. Η υπόθεση που θέλουμε να ελέγξουμε είναι η:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

(Οτι δηλαδή τα τρία είδη αυτοκινήτων έχουν την ίδια μέση κατανάλωση).
Ο δειγματικός μέσος των 15 παρατηρήσεων που έχουμε διαθέσιμες είναι

$$\bar{\bar{y}}_{..} = 26.0.$$

Η συνολική τετραγωνική απόκλιση των παρατηρήσεων είναι:

$$SST = \sum_{i=1}^{3}\sum_{j=1}^{6}(y_{ij} - \bar{y}_{..})^2 = 32.940$$

Αν δεν υπήρχε καμιά διαφορά στην απόδοση των αυτοκινήτων τότε θα έπρεπε SST = 0.

Αν η μηδενική υπόθεση H_0 δεν απορριφθεί τότε τα στοιχεία του δείγματος αποτελούν ένδειξη ότι τα τρία είδη αυτοκινήτων θα έχουν, περίπου, την ίδια μέση κατανάλωση και επομένως η μόνη αιτία αποκλίσεων στα δεδομένα θα είναι η φυσιολογική διακύμανση στη κατανάλωση που παρατηρείται σε διαφορετικά αυτοκίνητα της ίδιας μάρκας. Οι τιμές των υπολοίπων στατιστικών συναρτήσεων που χρησιμοποιούνται στην ανάλυση διακύμανσης για το συγκεκριμένο πρόβλημα είναι:

$\bar{y}_{1.} = 26.91,\ \bar{y}_{2.} = 26.30,\ \bar{y}_{3.} = 23.75,\ \bar{y}_{..} = 25.86$

$$\sum_{j=1}^{6}(y_{1j} - \bar{y}_{1..})^2 = 2.806$$

$$\sum_{j=1}^{5}(y_{2j} - \bar{y}_{2..})^2 = 1.340$$

$$\sum_{j=1}^{4}(y_{3j} - \bar{y}_{3..})^2 = 1.050$$

Τα τρία τελευταία αθροίσματα αποτελούν μέτρα απόκλισης της απόδοσης για κάθε μοντέλο (*within variation*).

Το άθροισμα των τριών αυτών αιτιών διακύμανσης είναι:

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 = 2.804 + 1.340 + 1.050 = 5.1986$$

Το γεγονός ότι το SSE είναι σημαντικά μικρότερο από το SST αποτελεί ενδειξη ότι η μηδενική υπόθεση δεν είναι αληθινή και ότι στην πραγματικότητα υπάρχει στατιστική διαφορά στους μέσους των υπό εξέταση πληθυσμών.

Εαν απορριφθεί η μηδενική υπόθεση αυτό θα συνεπάγεται ότι μέρος της παρατηρούμενης διασποράς θα πρέπει να αποδοθεί σε διαφορές απόδοσης των διαφορετικών μοντέλων. Αυτή η απόκλιση μετριέται με τον προσδιορισμό του μεγέθους απόκλισης κάθε ενός από τους

τρείς δειγματικούς μέσους (26.91, 26.30, 23.75) από το συνολικό μέσο (25.86) όταν κάθε μια από τις αποκλίσεις αυτές έχει ως συντελεστή βαρύτητας το αντίστοιχο μέγεθος του δείγματος (*between variation*).

$$SSTr = \sum_{i=1}^{3} n_i (\bar{y}_{i.} - \bar{y}_{..})^2$$

$$= \sum_{i} \sum_{j} (\bar{y}_{i.} - \bar{y}_{..})^2$$

$$= 6(26.91 - 25.86)^2 + 5(26.30 - 25.86)^2 + 4(23.75 - 25.86)^2$$

$$= 25.391.$$

(Διασπορά μεταξύ (*between*) μοντέλων).

Αφού οι αποκλίσεις μεταξύ των μοντέλων είναι σχεδόν ίσες (αθροιστικά) με την συνολική απόκλιση του δείγματος έχουμε άλλη μια ένδειξη ότι η H_0 δεν ισχύει.
Είναι προφανές ότι η συνολική απόκλιση μεταξύ (*between*) των μοντέλων και εντός των μοντέλων (within) δίνει την γενική συνολική αποκλιση. Δηλαδή,

$$SST = SST_r + SSE$$

Πράγματι, στη περιπτωσή μας 25.391 + 5.1986 = 30.59.
Ο πίνακας της ανάλυσης δασποράς για το πρόβλημα αυτό έχει ως εξής:

Πίνακας ANOVA
(Ανάλυση Διακύμασης)

Αιτία Διασποράς	SS	Βαθμοί Ελευθερίας	MS	F
Μεταξύ Επιδράσεων	25.391	2	12.695	29.318
Μέσα στις Επιδράσεις	5.1986	12	0.433	
Σύνολο	30.59	14		

Πρόβλημα: Μια εταιρία καταναλωτών ενδιαφέρεται να συγκρίνει την μέση διάρκεια ζωής (σε λεπτά) τεσσάρων ειδών μπαταριών που χρησιμοποιούνται σε παιδικά παιχνίδια. Για το σκοπό αυτό επιλέγεται

ένα τυχαίο δείγμα από κάθε ένα από τα τέσσερα είδη μπαταριών. Στη συνέχεια μετριέται ο χρόνος ζωής για την κάθε μία από τις επιλεγείσες μπαταρίες. Τα αποτελέσματα δίνονται στον πίνακα που ακολουθεί.

Χρόνοι Ζωής Μπαταριών

Είδος 1	Είδος 2	Είδος 3	Είδος 4
43	45	45	45
47	48	43	48
48	49	41	55
45	46	41	47
46	52	38	58
42	45	46	50
46	44	45	46
45	47	41	53
49		43	56
		41	

Να ελεγχθεί κατά πόσο οι χρόνοι μέσης διάρκεια ζωής των τεσσάρων διαφορετικών ειδών διαφέρουν μεταξύ τους.

Λύση: Για το πρόβλημα αυτό τα δεδομένα δίνουν τις παρακάτω τιμές στις αντιστοιχες στατιστικές συναρτήσεις:

$$y.. = 1669,\ \overline{y}_{..} = 1669/36 = 46.361$$

$$\sum_i \sum_j y_{ij}^2 = 78049,\ \overline{y}_{1.} = 45.667,\ \overline{y}_{2.} = 47.000,$$

$$\overline{y}_{3.} = 42.400,\ \overline{y}_{4.} = 50.889$$

$$SSTr = 78049 - 36\,(43.361)^2 = 672.306$$

$$SSTr = \sum_{i=1}^{k} n_i (\overline{y}_{i.} - \overline{y}_{..})^2 = \sum_{i=1}^{k} n_i \overline{y}_{i.}^2 - n\overline{y}_{..}^2$$

$$= 9(45.667)^2+8(47.000)^2+10(42.400)^2+9(50.889)^2-36\,(46.361)^2$$

$$= 349.017.$$

$$SSE = 672.306 - 349.017 = 323.289$$

$$MSTr = 349.017 / (4-1) = 116.339$$

$$MSE = 323.289 / (36-4) = 10.103$$
$$F = 116.339 / 10.103 = 11.52$$

Επειδή $F > F_{3,\ 32,\ 0.95} = 2.922$ η H_0 απορρίπτεται στο $\alpha = 0.05$ επίπεδο σημαντικότητας.

Σημείωση: Η συμπερασματολογία που προηγήθηκε ισχύει με την προϋπόθεση ότι οι διακυμάνσεις για όλα τα είδη μπαταριών, όσον αφορά την μέση διάρκεια ζωής, είναι ίσες και ότι οι αποκλίσεις των παρατηρήσεων από τη μέση διάρκεια ζωής είναι $N(0, \sigma^2)$.
Ο πίνακας της ανάλυσης διασποράς είναι ο εξής:

Πίνακας ANOVA
(Ανάλυση Διακύμασης)

Αιτία Διασποράς	SS	Βαθμοί Ελευθερίας	MS	F
Μεταξύ Επιδράσεων	349.017	3	116.339	11.52
Μέσα στις Επιδράσεις	323.289	32	10.103	
Σύνολο	672.306	35		

Πρόβλημα: Σε μία πόλη όπου η μόλυνση της ατμόσφαιρας έχει φτάσει σε υψηλά επίπεδα το Υπουργείο Περιβάλλοντος προβληματίζεται για το κατά πόσον κάποια από τις τρείς μεγάλες βιομηχανίες που λειτουργούν στην περιοχή της πόλης αυτής ρυπαίνει την ατμόσφαιρα περισσότερο από ότι οι άλλες. Σε τυχαίες χρονικές στιγμές το Υπουργείο Περιβάλλοντος παίρνει μετρήσεις για την ποσότητα των ρύπων που διαφεύγει στην ατμόσφαιρα από τις βιομηχανίες αυτές. Τα αποτελέσματα των μετρήσεων, σε μονάδες ρύπων, για τις τρείς βιομηχανίες καταγράφονται στον πίνακα που ακολουθεί. Υπάρχει στατιστικά σημαντική διαφορά στην ποσότητα των ρύπων που εκλύονται από τις τρείς βιομηχανίες με βάση τα στοιχεία αυτά;

Ποσότητες Ρύπων

Βιομ. Α	Βιομ. Β	Βιομ. Γ
46.3	48.6	45.1
43.7	52.3	46.7
51.2	50.9	41.8
49.6	53.6	40.4
48.8	55.7	42.6

Λύση: $H_0 : \mu_A = \mu_B = \mu_\Gamma$

H_1 : τουλάχιστον δύο από τους μέσους διαφέρουν.

Από τα στοιχεία μας έχουμε

$$y.. = 717.3,\ \overline{y}_{..} = 717.3/15 = 47.82$$

$$\sum_i \sum_j y_{ij}^2 = 34588.99,\ \overline{y}_{1.} = 52.22,\ \overline{y}_{2.} = 47.92,\ \overline{y}_{3.} = 43.32,$$

$$\begin{aligned}
SST &= 34588.99 - 15(47.82)^2 = 287.704 \\
SST &= 5(52.22)^2 + 5(47.92)^2 + 5(43.32)^2 - 15(47.82)^2 \\
&= 34499.386 - 34301.286 = 198.100 \\
SSE &= SST - SSTr \\
&= 287.704 - 198.100 \\
&= 89.604 \\
MSTr &= 198.100/(3-1) = 99.05 \\
MSE &= 89.604/(15-3) = 7.467 \\
F &= 99.05/7.467 = 13.27
\end{aligned}$$

Επειδή $F > F_{2,12,0.95} = 3.885$ η μηδενική υπόθεση θα πρέπει να απορριφθεί.

Σημείωση: Είναι και εδώ απαραίτητες οι υποθέσεις της ισότητας των διασπορών και της κανονικότητας των λαθών προκειμένου να ισχύουν τα συμπεράσματα.

Ο πίνακας της ανάλυσης διασποράς για το πρόβλημα αυτό ειναι ο εξής:

Πίνακας ANOVA
(Ανάλυση Διακύμασης)

Αιτία Διασποράς	SS	Βαθμοί Ελευθερίας	MS	F
Μεταξύ Επιδράσεων	198.100	2	99.05	13.27
Μέσα στις Επιδράσεις	89.604	12	13.27	
Σύνολο	287.704	14		