Ανάλυση διακύμανσης κατά ένα κριτήριο

One-Way Analysis of Variance

Είναι εκείνη η μεθοδολογία που ασχολείται με την εξέταση και τον προσδιορισμό των πηγών των αποκλίσεων που παρατηρούνται σε δειγματικά δεδομένα. Είναι η μέθοδος διαχωρισμού της επιρροής των διαφορετικών υποσυνόλων των παραμέτρων πάνω στις μετρήσεις. Με τον όρο "παράμετροι" συνήθως εννοούμε το αποτέλεσμα κάποιων επιδράσεων (treatments) σε μία μεταβλητή Y.

<u>Π.χ.</u> Τι είναι η επίδραση και πως επηρεάζει ένα πείραμα;

Έστω $Y :=$ παραγωγή σιταριού από κάποιο συγκεκριμένο κομμάτι χωραφιού

Επίδραση $:=$ προσθήκη λιπάσματος στο κομμάτι του χωραφιού κατά τη σπορά.

Ο ερευνητής εξάγει συμπεράσματα μελετώντας κομμάτια χωραφιού που έχουν υποστεί την επίδραση και άλλα που δεν την έχουν υποστεί.

Πρόβλημα που απασχολεί τον ερευνητή: Η "επίδραση" δηλαδή η χρήση του λιπάσματος επιφέρει στατιστικά σημαντική βελτίωση της παραγωγής; ΝΑΙ ή ΟΧΙ;

<u>Π.χ.</u>

	Toyota ($n_1=6$)	Datsun ($n_2=5$)	Mazda ($n_3=4$)
	27.1	25.3	23.1
	25.5	26.5	24.3
	27	26.4	23.4
	26.9	26.8	24.2
	27.7	26.5	
	27.3		

$Y :=$ μέση κατανάλωση βενζίνης σε Km/lt

Δεδομένα: Μέση κατανάλωση βενζίνης τριών ειδών διαφορετικών αυτοκινήτων. Υπάρχει στατιστικά σημαντική διαφορά στη μέση κατανάλωση βενζίνης μεταξύ των τριών αυτών διαφορετικών αυτοκινήτων;

Βάζουμε οδηγούς να οδηγήσουν τα αυτοκίνητα σε μία διαδρομή (ίδια για όλα τα αυτοκίνητα) μήκους 300 km με σταθερή ταχύτητα 55 km/h. Ποιες είναι οι αιτίες που οδηγούν στη διακύμανση των αποτελεσμάτων;

Διαφορετικών κατασκευαστών αυτοκίνητα έχουν διαφορετική απόδοση λόγω διαφορετικής κατασκευής.

π.χ. Κατανάλωση MAZDA < κατανάλωση TOYOTA DATSUN

Μεταξύ αυτοκινήτων του ίδιου κατασκευαστή παρουσιάζονται αποκλίσεις. 6 αυτοκίνητα οδηγήθηκαν από 6 διαφορετικούς οδηγούς. Αυτοκίνητα με διαφορετικά λάστιχα (διαφορετική πίεση αέρα). Τελευταίο service σε διαφορετική χρονική στιγμή. Ενδεχομένως διαφορετικές καιρικές συνθήκες κάτω από τις οποίες έγινε το πείραμα. Άρα, τα δεδομένα περιέχουν ένα μεγάλο αριθμό πηγών διακύμανσης.

Στόχος: Να καθοριστούν οι κύριες πηγές διακύμανσης και η ποσότητα της διακύμανσης που οφείλεται σε καθένα από τους διαφορετικούς λόγους που μας ενδιαφέρει να εξετάσουμε. Η υπόλοιπη διακύμανση των δεδομένων θεωρείται από τον ερευνητή ότι οφείλεται σε τυχαίους παράγοντες και καλείται σφάλμα (error).

Στόχος: Μείωση στο ελάχιστο του λάθους ώστε η ποσότητα της διακύμανσης που οφείλεται στους λόγους που έχουμε καθορίσει να προσδιοριστούν όσον το δυνατόν ακριβέστερα.

Άρα, η ανάλυση διακύμανσης προσπαθεί να καθορίσει όλες τις πηγές διακύμανσης και το ποσό της διακύμανσης που μπορεί να αποδοθεί σε καθεμία από τις πηγές αυτές.

Η πιο απλή περίπτωση ανάλυσης διακύμανσης είναι η ανάλυση διακύμανσης κατά ένα κριτήριο (μία μόνο επίδραση). Έχουμε κ ανεξάρτητα δείγματα από κ πληθυσμούς. Υποθέτουμε ότι $\sigma_1^2=\sigma_2^2=\dots=\sigma_k^2=\sigma^2$. Ελέγχουμε την υπόθεση:

$H_0: \mu_1=\mu_2=\dots=\mu_k$

H_1: τουλάχιστον δύο μέσοι διαφέρουν

π.χ. κ διαφορετικά λιπάσματα

κ διαφορετικά είδη αυτοκινήτων ⟶ 1 επίδραση (π.χ. μέση κατανάλωση)

κ σύνολα βαθμών από κ διαφορετικούς καθηγητές

1	2	...	i	...	k
y_{11}	y_{21}		y_{i1}		y_{k1}
y_{12}	y_{22}		y_{i2}		y_{k2}
y_{1n_1}	y_{2n_2}		y_{in_i}		y_{kn_k}

κ ανεξάρτητα δείγματα μεγέθους $n_1, n_2, \dots, n_k$.

π.χ. Κάθε μάρκα αυτοκινήτων i περιέχει παρατηρήσεις που θα σκορπούν περισσότερο ή λιγότερο γύρω από το μέσο μ_i (i: μάρκα αυτοκινήτου) καθώς εκτός από το είδος αυτοκινήτου, η μέτρηση επηρεάζεται και από άλλους παράγοντες. Ερώτημα: Είναι όλα τα μ_i ίδια; Μήπως οι διαφορές στις παρατηρήσεις y_{ij}, $j=1,\dots,n_i$, $i=1,\dots,k$, θα μπορούσαν να προέρχονται από τυχαίες αποκλίσεις γύρω από το ίδιο μέσο μ; Δηλαδή, το πρόβλημα που τίθεται είναι αν οι μέσες τιμές των πληθυσμών από τους οποίους προέρχονται τα δείγματα (οι πληθυσμοί θεωρούνται κανονικοί και με κοινή (άγνωστη) διασπορά) διαφέρουν στατιστικά σημαντικά ή όχι.

Έστω οι παρατηρήσεις y_{ij} είναι μετρήσεις ανεξάρτητων τυχαίων μεταβλητών με αναμενόμενη τιμή μ_i. Y_{ij}: τ.μ.

δηλ. $EY_{11} = EY_{12} = \dots = EY_{1n_1} = \mu_1$

$$EY_{k1} = EY_{k2} = \dots = EY_{kn_k} = \mu_k$$

Θεωρούμε ότι όλες οι τμ. έχουν κοινή (άγνωστη) διασπορά σ^2 δηλ. $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 = \sigma^2$.

Το μοντέλο μπορεί να γραφεί:

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, \dots, k, \quad j = 1, \dots, n_i$$

$$E(\varepsilon_{ij}) = 0, \ var(\varepsilon_{ij}) = \sigma^2, \ \varepsilon_{ij} \text{ ανεξάρτητα}$$

ε_{ij}: τυχαίο λάθος (μετρά την απόκλιση της j-παρατήρησης στο i-δείγμα από τον αντίστοιχο μέσο του i-δείγματος).

Έλεγχος υπόθεσης: $H_0: \mu_1 = \mu_2 = \dots = \mu_{k-1} = \mu_k$ (δηλ. τα k-δείγματα προέρχονται από τον ίδιο πληθυσμό)

H_1: τουλάχιστον ένα μ_i διαφορετικό από τα άλλα

$$\mu = \frac{1}{n_+} \sum_{i=1}^{k} n_i \mu_i, \quad n_+ = \sum_{i=1}^{k} n_i$$

$$a_i = \mu_i - \mu \Rightarrow \mu_i = \mu + a_i, \quad \sum_{i=1}^{k} n_i a_i = 0$$

$$H_0: a_1 = a_2 = \dots = a_k = 0$$

$H_1: a_i \neq 0$ για ένα τουλάχιστον i, $i = 1, \dots, k$.

$$y_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}, \quad i=1,\dots,k$$

$$y_{\cdot\cdot} = \frac{1}{n_+}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = \sum_{i=1}^{k}\frac{n_i}{n_+}\, y_{i\cdot}$$ (δειγματικός μέσος όλων των παρατηρήσεων)

$$Y_{ij} = \mu + a_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0,\sigma^2)$$

Ερώτηση: Γιατί δεν λαμβάνουμε **τα δείγματα** ανά δύο για να κάνουμε το t-test?

$H_0: \mu_1 = \mu_2 \qquad n_1, n_2 \leq 30 \qquad \sigma^2 = \sigma_1^2 = \sigma_2^2$

$H_1: \mu_1 \neq \mu_2$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\sqrt{\frac{(n_1-1)S_1^2+(n_2-2)S_2^2}{n_1+n_2-2}}} \sim t_{n_1+n_2-2;\,\alpha/2}$$

Πληθυσμοί από τους οποίους προέρχονται τα δείγματα κανονικοί. Θα χρειαζόταν να κάνουμε $\binom{k}{2}$ ελέγχους. Αν το κάθε test γινόταν σε ε.σ.σ. α τότε το σφάλμα τύπου I

$P\left(\begin{matrix}\text{απόρριψης} \\ \text{της } H_0\end{matrix} \,\middle|\, H_0 \text{ σωστή}\right)$ για τη συνολική διαδικασία

θα ήταν $k\alpha$. Για να ελέγξω το συνολικό λάθος τύπου I σε ε.σ.σ. 5% θα έπρεπε να κάνω κάθε t-test σε ε.σ.σ. $5\%/k$. Έτσι, όμως θα είμασταν αρκετά συντηρητικοί στην απόρριψη της H_0. Ο ταυτόχρονος έλεγχος της σύνθετης υπόθεσης γίνεται με το F-test.

Συνολική απόκλιση: $$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-Y_{\cdot\cdot}\right)^2$$

$$=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[(Y_{ij}-Y_{i\cdot})+(Y_{i\cdot}-Y_{\cdot\cdot})\right]^2$$

$$=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-Y_{i\cdot})^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{i\cdot}-Y_{\cdot\cdot})^2$$

$$+2\sum_{i=1}^{k}(Y_{i\cdot}-Y_{\cdot\cdot})\underbrace{\sum_{j=1}^{n_i}(Y_{ij}-Y_{i\cdot})}_{=\,n_iY_{i\cdot}-n_iY_{i\cdot}=0}$$

$$\Rightarrow \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-Y_{\cdot\cdot})^2}_{\text{(ολική διακύμανση)}}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-Y_{i\cdot})^2+\sum_{i=1}^{k}n_i(Y_{i\cdot}-Y_{\cdot\cdot})^2$$

(διακύμανση των παρατηρήσεων γύρω από το μέσο κάθε δείγματος)

(διακύμανση των μέσων των δειγμάτων γύρω από τον κοινό μέσο)

$$SST = SST_r + SSE$$

διακύμανση που εξηγείται από το μοντέλο

Η ιδέα χρήσης του F-test βασίζεται στις σχέσεις:

(i) $E(SSE)=\sigma^2(n_T-k)$

(ii) $E(SST_r) = \sigma^2(k-1) + \sum_{i=1}^{k} n_i a_i^2$

$$MSSE = \frac{1}{(n_+ - k)} SSE \Rightarrow E(MSSE) = \sigma^2$$

μέσο τετραγωνικό σφάλμα σφαλμάτων

$$MSST_r = \frac{1}{(k-1)} SST_r \Rightarrow E(SST_r) = \sigma^2 + \frac{1}{k-1} \sum_{i=1}^{k} n_i a_i^2$$

μέσο τετράγωνο των δειγμάτων

Ορίζουμε $F = \frac{MSST_r}{MSSE}$ → μεγάλο όταν κάποια $a_i \neq 0$ (αυτό που θέλουμε)

Όταν ισχύει η $H_0: a_1 = a_2 = \dots = a_k = 0$

$F \sim F_{k-1, n_+ - k}$ γιατί $SSE \sim \sigma^2 \chi^2_{n_+ - k}$

$SST_r \sim \sigma^2 \chi^2_{k-1}$

SSE, SST_r ανεξάρτητες τμ.

Σε ε.σ.σ. α απορρίπτουμε την H_0 αν

$F > F_{k-1, n_+ - k; 1-\alpha}$. Τα αθροίσματα τετραγώνων, οι β.ε., τα μέσα τετράγωνα και η σ.σ. F συγκεντρώνονται σε έναν πίνακα που καλλείται πίνακας ανάλυσης διασποράς (ANOVA table).

ANOVA TABLE
ΠΙΝΑΚΑΣ ΑΝΑΛΥΣΗΣ ΔΙΑΣΠΟΡΑΣ

Αιτία διασποράς	Άθροισμα τετραγώνων	Βαθμοί Ελευθερίας	Μέσο Τετραγωνικό Σφάλμα	
Μεταξύ δειγμάτων	SST_r	$k-1$	$MSST_r = \frac{SST_r}{k-1}$	$F = \frac{MSST_r}{MSSE}$
Μέσα στα δείγματα	SSE	n_+-k	$MSSE = \frac{SSE}{n_+-k}$	

Αν ισχύει η H_0: $F = \frac{MSST_r}{MSSE} = \frac{SST_r/k-1}{SSE/n_+-k}$

$$= \frac{SST_r(n_+-k)}{SSE(k-1)} \sim F_{k-1, n_+-k}$$

Αν Ho δεν ισχύει, ο παράγοντας SST_r θα πρέπει να είναι μεγάλος δηλαδή το SST_r πρέπει να είναι μεγάλο σε σχέση με το MSSE κάτι που δικαιολογεί την επιλογή της συγκεκριμένης σ.σ.

Παράδειγμα (με τα αυτοκίνητα)

$H_0: \mu_1=\mu_2=\mu_3$ (τα τρία είδη αυτοκινήτων έχουν την ίδια μέση κατανάλωση)

H_1: Ho δεν ισχύει

$n_+=15$ παρατηρήσεις $(n_1=6, n_2=5, n_3=4)$

$$y_{..} = \frac{1}{n_+}\sum_{i=1}^{3}\sum_{j=1}^{n_i} y_{ij} = 26 \Rightarrow \underline{y_{..} \cong 26}$$

$$SST=\sum_{i=1}^{3}\sum_{j=1}^{n_i}(y_{ij}-y_{..})^2=30,673$$

$y_{1.}=26,9167$, $y_{2.}=26,3$, $y_{3.}=23,75$, $y_{..}\cong 26(\cong 25.8667)$

$$\sum_{j=1}^{6}(y_{1j}-y_{1.})^2 \qquad \sum_{j=1}^{5}(y_{2j}-y_{2.})^2$$

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

Τα τρία αθροίσματα είναι μέτρα απόκλισης της απόδοσης για κάθε είδος αυτοκινήτου.

$$SSE=\sum_{i=1}^{3}\sum_{j=1}^{n_i}(y_{ij}-y_{i.})^2$$
$$=5.198$$

Αν η Ho δεν απορριφθεί τότε τα στοιχεία του δείγματος παρέχουν ενδείξεις ότι τα τρία είδη αυτοκινήτων θα την ίδια περίπου μέση κατανάλωση και επομένως η μόνη αιτία αποκλίσεων θα είναι η φυσιολογική διακύμανση που παρατηρείται σε διαφορετικά αυτοκίνητα της ίδιας μάρκας, και αποδίδεται με το SSE.

Το SSE είναι σημαντικά μικρότερο του SST κάτι που αποτελεί ένδειξη ότι η Ho δεν είναι αληθινή και ότι υπάρχει στατιστικά σημαντική διαφορά στους μέσους των υπό εξέταση πληθυσμών. Αν η Ho απορριφθεί, αυτό θα σημαίνει ότι μέρος της διασποράς οφείλεται σε διαφορές απόδοσης των διαφορετικών πληθυσμών. Το SST_r είναι μέτρο της διασποράς αυτής.

$$SST_r=\sum_{i=1}^{3}n_i(y_{i.}-y_{..})^2=25,475$$

$SST = SST_r + SSE = 25,475 + 5,198 = 30,673$

ΠΙΝΑΚΑΣ ANOVA

Αιτία διασποράς	SS	Β.Ε	MS	F	Sig
μεταξύ δειγμάτων	25,475	2	12,737	29,404	= α = 0
μέσα στα δείγματα	5,198	12	0,433		
Σύνολο	30,673	14			

$f_{2,12} = 3,89$ σε ε.σ.σ. α

$F = 29,404 > f_{2,12,\alpha=0.05} = 3.89$

Απορρίπτουμε την H_0 σε ε.σ.σ. α.

SPSS

Analyze → Compare Means
→ One-Way Anova
→ Options (✓Descriptives
✓Homogeneity of variance test)
• Exclude cases listwise

Levene's test of homogeneity of variance is computed by SPSS to test the ANOVA assumption that each group (category, mark) has the same variance.

If the Levene statistic is significant at e.g. the $\alpha = 0.05$ (or $\alpha = 0.01$) level, the researcher rejects the H_0 hypothesis that the auto marks have equal variances. The Levene statistic is not significant in our case and the researcher concludes that the three marks are homogenous in variance satisfying an important assumption of ANOVA. Note that using the 0.01 level, the H_0 hypothesis of equal variances is not rejected, too.