ΑΝΑΛΥΣΗ ΤΗΣ ΔΙΑΚΥΜΑΝΣΗΣ

Γενικά: Υπάρχει μια εξαρτημένη μεταβλητή Y (ποσοτική). Επίσης υπάρχουν μία ή περισσότερες ανεξάρτητες μεταβλητές. Η πιο συνηθισμένη περίπτωση είναι οι ανεξάρτητες μεταβλητές να είναι ποιοτικές.

Σε αντίθεση με την παλινδρόμηση, δεν μας ενδιαφέρει η σχέση ανεξάρτητων και εξαρτημένης, ούτε και η πρόβλεψη. Κυρίως τα προβλήματα είναι ελέγχου υποθέσεων.
Οι ανεξάρτητες μεταβλητές αναφέρονται σαν παράγοντες (factors).

Ανάλυση της Διακύμανσης κατά ένα παράγοντα

Μαθηματική Διατύπωση:
Έχουμε παρατηρήσεις $Y_{ij} \sim N(\mu_i, \sigma^2)$ $i = 1, ..., I$, $j = 1, ..., n_i$ και μπορούν να δοθούν υπό μορφή πίνακα:

$$\begin{array}{llll} Y_{11} & Y_{12} & \cdots & Y_{1n_1} \; \} \\ Y_{21} & Y_{22} & \cdots & Y_{2n_2} \; \} \\ \vdots & & & \quad\;\; \} \\ Y_{I1} & Y_{I2} & \cdots & Y_{In_I} \; \} \end{array} \quad \text{Κάθε γραμμή είναι τυχαίο δείγμα από την } N(\mu_i, \sigma^2)$$

Πρόβλημα: Να ελεγχθεί αν οι μέσες τιμές είναι ίσες.

$$H_0 : \mu_1 = ... = \mu_I$$

Εφαρμογή: Έστω ότι υπάρχουν $I = 4$ ποικιλίες κάποιου αγροτικού προϊόντος. Ενδιαφερόμαστε να εξετάσουμε ποια ποικιλία είναι καλύτερη με την έννοια ότι δίνει καλύτερη απόδοση.
Y_{ij} = απόδοση (παραγωγή) της i ποικιλίας στο j αγρόκτημα.
Θεωρώ τον έλεγχο: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ (δεν διαφέρουν οι ποικιλίες)
$H_1 : \mu_\kappa \neq \mu_\lambda$ (υπάρχουν διαφορές)
Εδώ έχουμε ανάλυση της διακύμανσης κατά ένα παράγοντα με 4 επίπεδα (levels).

Έλεγχος: Έστω $n = \sum_{i=1}^{I} n_i$ (ολικός αριθμός παρατηρήσεων).

$$\underline{\mu}_{nx1} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_1 \\ \hline \mu_2 \\ \vdots \\ \mu_2 \\ \vdots \\ \hline \mu_I \\ \vdots \\ \mu_I \end{pmatrix} \begin{matrix} \} \\ \} \; n_1 \; \phi o\rho\varepsilon\varsigma \\ \} \\ \} \\ \} \; n_2 \; \phi o\rho\varepsilon\varsigma \\ \} \\ \\ \} \\ \} \; n_I \; \phi o\rho\varepsilon\varsigma \\ \} \end{matrix} \qquad \underline{Y} = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ \hline Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ \hline Y_{I1} \\ \vdots \\ Y_{In_1} \end{pmatrix} \begin{matrix} \} \\ \} \; n_1 \; \phi o\rho\varepsilon\varsigma \\ \} \\ \} \\ \} \; n_2 \; \phi o\rho\varepsilon\varsigma \\ \} \\ \\ \} \\ \} \; n_I \; \phi o\rho\varepsilon\varsigma \\ \} \end{matrix}$$

$$\underline{\mu_0} = \begin{pmatrix} \mu_. \\ \mu_. \\ \vdots \\ \mu_. \end{pmatrix}_{nx1} \quad ,\text{όπου } \mu_. = \frac{n_1\mu_1 + n_2\mu_2 + \ldots + n_I\mu_I}{n} = \frac{\sum_{i=1}^{I} n_i\mu_i}{n}$$

Όταν ισχύει η H_0 τότε έχω ότι $\underline{\mu_0} = \mu \ . \ \underline{1}$, $\underline{1} = (1,\ldots,1)^T$ και μ η κοινή άγνωστη τιμή των μ_i. Άρα, ο έλεγχος παίρνει τη μορφή:

$$\left.\begin{aligned} H_0&: \underline{\mu} = \underline{\mu_0} \\ H_1&: \underline{\mu} \neq \underline{\mu_0} \end{aligned}\right\} \Leftrightarrow \begin{aligned} H_0&: \left\|\underline{\mu} - \underline{\mu_0}\right\|^2 = 0 \\ H_1&: \left\|\underline{\mu} - \underline{\mu_0}\right\|^2 > 0 \end{aligned}$$

Έστω $\theta = \left\|\underline{\mu} - \underline{\mu_0}\right\|^2 = \sum_{i=1}^{I} n_i(\mu_i - \mu_.)^2$

Άρα, έχω τον ισοδύναμο έλεγχο:
$H_0: \theta = 0$ προς $H_1: \theta > 0$.

Πρέπει να εκτιμήσω την παράμετρο θ. Για να εκτιμήσω την θ έχω να εκτιμήσω τα μ_i και $\mu_.$.

Εκτιμήτρια του μ_i: $\hat{\mu}_i = \sum_{j=1}^{n_i} \frac{Y_{ij}}{n_i} = \frac{Y_{i.}}{n_i} = \overline{Y}_{i.}$

Εκτιμήτρια του $\mu_.$: $\hat{\mu}_. = \frac{n_1\overline{Y}_{1.} + \ldots + n_I\overline{Y}_{I.}}{n} = \overline{Y}_{..}$ ($= \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij}$).

Άρα, η εκτιμήτρια του θ είναι $\hat{\theta} = \sum_{i=1}^{I} n_i(\overline{Y}_{i.} - \overline{Y}_{..})^2$.

Άρα, απορρίπτουμε αν $\hat{\theta} > c$, όπου c σταθερά που υπολογίζεται από το επίπεδο σημαντικότητας.

Ισχύει η βασική σχέση:

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

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

$$= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2 + \sum_{i=1}^{I} n_i(\overline{Y}_{i.} - \overline{Y}_{..})^2 \text{ , αφού } \sum_{i=1}^{I}(\overline{Y}_{i.} - \overline{Y}_{..})\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.}) = 0 \text{ .}$$

Άρα,

$$\underbrace{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{..})^2}_{\substack{n-1 \text{ β.ε.} \\ SS_{total}}} = \underbrace{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i.})^2}_{\substack{n-I \text{ β.ε.} \\ SS_{within}}} + \underbrace{\sum_{i=1}^{I} n_i(\overline{Y}_{i.} - \overline{Y}_{..})^2}_{\substack{I-1 \text{ β.ε.} \\ SS_{between}}} .$$

♬

(άθροισμα τετραγώνων
των υπολοίπων)

Όμως, $\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{..})^2/\sigma^2 \sim X^2_{n-1}$,

γιατί στην περίπτωση που ισχύει η H_0 τα $Y_{ij} \sim N(\mu,\sigma^2)$.

Για τον ίδιο λόγο έχω ότι $\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{i.})^2/\sigma^2 \sim X^2_{n_i-1}$

και άρα $\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{i.})^2/\sigma^2 \sim X^2_{\sum(n_i-1)} = X^2_{n-I}$, λόγω ανεξαρτησίας.

Άρα, αν αποδείξω ότι $\hat{\theta}$ και $\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}-\overline{Y}_{i.})^2$ είναι ανεξάρτητα, η βασική σχέση δίνει ότι $\hat{\theta}/\sigma^2 \sim X^2_{I-1}$.

Αλλά Cov($Y_{ij}-\overline{Y}_{i.},\overline{Y}_{i.}-\overline{Y}..$) = Cov($Y_{ij},\overline{Y}_{i.}$) - Cov($Y_{ij},\overline{Y}_{..}$) - Cov($\overline{Y}_{i.},\overline{Y}_{i.}$) + Cov($\overline{Y}_{i.},\overline{Y}_{..}$)

$=\frac{\sigma^2}{n_i}-\frac{\sigma^2}{n}-\frac{\sigma^2}{n_i}+\frac{\sigma^2}{n}=0$ και λόγω κανονικότητας έχω ανεξαρτησία.

Άρα, $\hat{\theta}/\sigma^2 \sim X^2_{I-1}$.

Επειδή όμως το σ^2 είναι άγνωστο, από τη σχέση

$E(SS_{with})=(n-I)\sigma^2 \Rightarrow E\left(\frac{SS_{with}}{n-I}\right)=\sigma^2$.

Οπότε, ο έλεγχος γίνεται

$F=\frac{\hat{\theta}/(I-1)}{SS_{with}/(n-I)}=\frac{MS_{bet}}{MS_{with}} \sim F_{I-1,n-I}$.

Απορρίπτω αν $F > F_{I-1,n-I,a}$.