16ο online μάθημα

20/5/2020

Κεφάλαιο 6

ΜΟΝΤΕΛΑ ΑΝΑΛΥΣΗΣ ΔΙΑΚΥΜΑΝΣΗΣ

Η ανάλυση διακύμανσης μελετά την επίδραση μιας ή περισσοτέρων ποιοτικών ανεξάρτητων μεταβλητών στην εξαρτημένη μεταβλητή. Δεν κατασκευάζεται κάποια σχέση μεταξύ εξαρτημένης και ανεξάρτητων αλλά το ενδιαφέρον επικεντρώνεται στο να αναλυθούν οι επιδράσεις μιας ή περισσοτέρων ανεξάρτητων στην εξαρτημένη

Ορολογία:

Παράγοντας: ο όρος αυτός χρησιμοποιείται για να περιγράψει την ανεξάρτητη μεταβλητή. Κάθε τιμή του αποτελεί ένα επίπεδο του παράγοντα.
Όταν εξετάζεται μόνο ένας παράγοντας τότε λέμε ότι έχουμε ΑΝΑΔΙΑ κατά ένα παράγοντα (ΑΝΑΔΙΑ = ανάλυση διακύμανσης)
Όταν εξετάζονται δύο παράγοντες τότε λέμε ότι έχουμε ΑΝΑΔΙΑ κατά δύο παράγοντες.

• ΑΝΑΔΙΑ ΚΑΤΑ 1 ΠΑΡΑΓΟΝΤΑ: αποτελεί ουσιαστικά γενίκευση του ελέγχου $\mu_1 = \mu_2$ όταν τα επίπεδα του παράγοντα είναι περισσότερα από ένα.
Έστω ένας παράγοντας (π.χ. μόρφωση πατέρα) με I επίπεδα (π.χ. ΔΗΜΟΤΙΚΟ, ΓΥΜΝΑΣΙΟ-ΛΥΚΕΙΟ, ΑΕΙ-ΤΕΙ).
Σε κάθε επίπεδο του παράγοντα υπάρχουν διαθέσιμες $n_1, n_2, \ldots, n_I$ παρατηρήσεις αντίστοιχα για την υπό μελέτη εξαρτημένη μεταβλητή (π.χ. εισόδημα σε ευρώ)
Έτσι

ΕΠΙΠΕΔΟ	ΠΑΡΑΤΗΡΗΣΕΙΣ
1	$Y_{11}, Y_{12}, \ldots, Y_{1n_1}$
2	$Y_{21}, Y_{22}, \ldots, Y_{2n_2}$
$\vdots$	$\vdots$
I	$Y_{I1}, Y_{I2}, \ldots, Y_{In_I}$

Άρα Y_{ij} είναι η j-οστή παρατήρηση από το i-επίπεδο
$i=1,\dots,I$ $\quad j=1,\dots,n_i$

Συμβολίζουμε τότε:
$Y_{i\cdot}$ = άθροισμα παρατηρήσεων στο δείγμα από το i-επίπεδο
δηλαδή

$$Y_{i\cdot} = \sum_{i=1}^{n_i} Y_{ij} \quad \text{και}$$

$$\bar{Y}_{i\cdot} = \frac{1}{n_i}\sum_{i=1}^{n} Y_{ij}$$

$Y_{\cdot\cdot}$ = το άθροισμα όλων των διαθέσιμων παρατηρήσεων
δηλαδή

$$Y_{\cdot\cdot} = \sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij} \quad \text{και}$$

$$\bar{Y}_{\cdot\cdot} = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij} \quad \text{όπου} \quad N=\sum_{i=1}^{I} n_i$$

ΣΧΗΜΑΤΙΚΑ

ΕΠΙΠΕΔΟ	ΠΑΡΑΤΗΡΗΣΕΙΣ	ΣΥΝΟΛΟ	Μ.Ο.
1	$Y_{11}, Y_{12} \dots Y_{1n_1}$	$Y_{1\cdot}$	$\bar{Y}_{1\cdot}$
2	$Y_{21}, Y_{22} \dots Y_{2n_2}$	$Y_{2\cdot}$	$\bar{Y}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	$Y_{I1}, Y_{I2} \dots Y_{In_I}$	$Y_{I\cdot}$	$\bar{Y}_{I\cdot}$

και $Y_{\cdot\cdot} = Y_{1\cdot} + \dots + Y_{I\cdot}$

ενώ $\bar{Y}_{\cdot\cdot} = \dfrac{Y_{\cdot\cdot}}{N}$

$N \rightarrow$ το σύνολο/πλήθος όλων των παρατηρήσεων.

Θεωρούμε το μοντέλο:

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

σφάλμα

αληθινή μέση τιμή του i-οστού πληθυσμού

Προσοχή: Μπορεί σε κάποια σύγγραμμα να δείτε το μοντέλο αυτό ως:

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

όμως εκεί Y_{ij} είναι η i-οστή τιμή στο j-επίπεδο, ενώ εμείς έχουμε συμβολίσει

Y_{ij}: την j-οστή παρατήρηση του i-οστού επιπέδου

Δεν διαφοροποιείται κάτι στον τρόπο σκέψης που θα ακολουθήσει. είναι

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

<u>Εκτίμηση των άγνωστων παραμέτρων</u>

Θα χρησιμοποιηθεί η μέθοδος ελαχίστων τετραγώνων.

Θα προσδιορίσουμε έτσι ώστε η

$$S(\mu_1, \dots, \mu_I) = \sum_{i=1}^{I} \sum_{j=1}^{n_i} \varepsilon_{ij}^2$$

ή ισοδύναμα

$$S(\mu_1, \dots, \mu_I) = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2$$

να γίνεται ελαχιστοποίηση ως προς $\mu_1, \mu_2, \dots, \mu_I$

Ουσιαστικά αυτό σημαίνει ότι $\hat{\mu}_I$ είναι τ/ώ να επαληθεύεται το:

$$\frac{\partial S}{\partial \mu_i} = 0 \quad , i = 1, 2, \dots, I$$

Η παραγώγιση αυτή θα μπορούσε να γίνει απευθείας αλλά ... ας την δούμε πιο αναλυτικά

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\mu_i)^2=\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}^2-2y_{ij}\mu_i+\mu_i^2)$$

ή

$$S(\mu_1,\dots\mu_I)=\sum_{i=1}^{I}\sum_{j=1}^{n_i}y_{ij}^2-2\sum_{i=1}^{I}\mu_i\sum_{j=1}^{n_i}y_{ij}+\sum_{i=1}^{I}\mu_i^2\sum_{j=1}^{n_i}1$$

ή

$$S(\mu_1\dots\mu_I)=\sum_{i=1}^{I}\sum_{j=1}^{n_i}y_{ij}^2-2\sum_{i=1}^{I}\mu_i y_{i.}+\sum_{i=1}^{I}\mu_i^2 n_i$$

Άρα

$$S(\mu_1\dots\mu_I)=\sum_{i=1}^{I}\sum_{j=1}^{n_i}y_{ij}^2-2(\mu_1 y_{1.}+\mu_2 y_{2.}+\dots+\mu_I y_{I.})+$$

$$+(n_1\mu_1^2+\dots+n_I\mu_I^2)$$

Επομένως

$$\frac{\partial S}{\partial \mu_i}(\mu_1\dots\mu_I)=-2y_{i.}+2n_i\mu_i \qquad i=1\dots I$$

Άρα $\frac{\partial S}{\partial \mu_i}(\mu_1\dots\mu_I)=0 \Rightarrow \hat{\mu}_i=\frac{y_{i.}}{n_i}=\bar{y}_{i.} \qquad i=1,\dots,I$

και επιπλέον

$$\frac{\partial S^2}{\partial^2 \mu_i}(\mu_1\dots\mu_I)=2n_i>0$$

Επομένως ο εκτιμητής ελαχίστων τετραγώνων

$$\hat{\mu}_i=\bar{y}_{i.} \qquad i=1\dots I$$

ο δειγματικός μέσος του i-οστού πληθυσμού (λογικό αποτέλεσμα)

Παρατήρηση (αποδείξεις): Προφανώς

$$\frac{\partial S}{\partial \mu_i}(\mu_1 \dots \mu_I) = \frac{\partial}{\partial \mu_i} \sum_i \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2 = -2 \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)$$

Άρα πρέπει:

$$\sum_{j=1}^{n_i} (Y_{ij} - \hat{\mu}_i) = 0 \Rightarrow \sum_{i=1}^{n_i} Y_{ij} = n_i \hat{\mu}_i \Rightarrow \hat{\mu}_i = \bar{Y}_{i.}$$

Επίσης

$$\frac{\partial S}{\partial \mu_i}(\mu_1 \dots \mu_I) = 2n_i > 0 \quad \text{(άρα σημείο ελαχίστου)}$$

Επομένως στο

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

οι Ε.Ε.Τ. είναι

$$\hat{\mu}_i = \bar{Y}_{i.} = \frac{1}{n_i} \cdot \sum_{j=1}^{n_i} Y_{ij} \qquad i = 1, \dots, I$$

Ιδιότητες των Ε.Ε.Τ.

Υποθέτουμε ότι $E(\varepsilon_{ij}) = 0$, $Var(\varepsilon_{ij}) = \sigma^2$ και

$Cov(\varepsilon_{ij}, \varepsilon_{kl}) = 0 \quad \forall (i,j) \neq (k,l)$

Τότε:

1. $E\hat{\mu}_i = \mu_i$
2. $Var\,\hat{\mu}_i = \sigma^2 / n_i$

} Απόδειξη εύκολη