ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ

Επικ. Καθ. Στέλιος Ζήμερας

Τμήμα Μαθηματικών
Κατεύθυνση Στατιστικής και Αναλογιστικά – Χρηματοοικονομικά Μαθηματικά

2015

Έλεγχος διακυμάνσεων

Μας ενδιαφέρει να εξετάσουμε 5 δίαιτες που δίνονται σε παιδιά ηλικίας ενός έτους. Το βάρος σε γραμμάρια δινεται από τον παρακάτω πίνακα

Δίαιτα	Βάρος σε γραμμάρια	n_i	$\bar{Y}_{i.}$	$Y_{i.}$
1	28 24.8 27.9 24.7 28.7 34.8 30.9	7	28.54	199.8
2	24.7 30.7 30.5 22 26.7 28.6 28.8 22.6 34.8	9	27.71	249.4
3	30.5 24.8 30.9 22.5 37.6 28.6 28.7 28.8 33.6 24	10	29	290
4	30.4 28.9 27.8 30.7 32.7 30.8 30.5 34.5 32.8 31.7 28.6 38.8	12	31.51	378.2
5	28.5 20.8 22.9 17.8 23.6 18.9 24.7 16.3 25.4 20.9 25.1	11	22.26	244.9

$$Y_{i.} = \sum_{j}^{n_i} Y_{ij} \qquad \bar{Y}_{i.} = \frac{1}{n_i} \sum_{j}^{n_i} Y_{ij}$$

Μοντελοποίηση

Μας ενδιαφέρει να εντοπίσουμε αν υπάρχουν διαφορές μεταξύ των μέσων

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

$$H_0 : \bar{Y}_{1.} = \bar{Y}_{2.} = \bar{Y}_{3.} = \bar{Y}_{4.} = \bar{Y}_{5.}$$

Θα έχουμε ότι για την παρατήρηση Y_{ij} που γράφεται ως

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \; i = 1,2,...,k; \; j = 1,2,...,n_i$$

όπου οι τ.μ. ε_{ij} είναι ανεξάρτητες και $\varepsilon_{ij} \sim N(0,\sigma^2)$ με $\sigma^2>0$

$$n = n_1 + n_2 + \ldots\ldots + n_k$$

Μοντελοποίηση

Η επίδραση της μ_i, i=1,2,..,k δίαιτας ανάλογα με το πώς γίνεται είναι:

- **Σταθερά** (μοντέλο τύπου I – μοντέλο σταθερών επιδράσεων) (όπως στο παράδειγμά μας)
- **Τυχαία μεταβλητή** (μοντέλο τύπου II – μοντέλο τυχαίων επιδράσεων)

Μοντέλο σταθερών επιδράσεων

Το μοντέλο

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \mathrm{i} = 1,2,...,\mathrm{k};\ \mathrm{j} = 1,2,...,\mathrm{n_i}$$

μπορεί να γραφεί με μορφή

$$Y_{ij} = \mu + a_i + \varepsilon_{ij}, \mathrm{i} = 1,2,...,\mathrm{k};\ \mathrm{j} = 1,2,...,\mathrm{n_i}$$

με

$$\sum_{i=1}^{k} n_i a_i = 0$$

Πράγματι αν $\sum_{i=1}^{k} n_i \mu_i = n\mu \Rightarrow \sum_{i=1}^{k} n_i (\mu_i - \mu) = 0$

$$\mu_i = (\mu_i - \mu) + \mu = \alpha_i + \mu$$

όπου $a_i = \mu_i - \mu$ και $\sum_{i=1}^{k} \mathrm{n_i} a_i = 0$

Γενικός μέσος

Μοντέλο σταθερών επιδράσεων

Αν αγνοήσουμε το μέγεθος των δειγμάτων τότε το μοντέλο απλοποιείται σε

$$Y_{ij} = \mu + a_i + \varepsilon_{ij}, \mathrm{i} = 1,2,...,\mathrm{k};\ \mathrm{j} = 1,2,...,\mathrm{n_i}$$

με $\sum_{i=1}^{k} a_i = 0$

Μοντέλο χωρίς βάρη

όπου $a_i = \mu_i - \bar{\mu}$ και $\bar{\mu} = \dfrac{\mu_1 + \mu_2 + ... + \mu_k}{n}$

Μοντέλο σταθερών επιδράσεων

Με την μέθοδο ελαχίστων τετραγώνων εκτιμάμε τους μέσους μ_i και την διασπορά $\sigma^2 \Rightarrow$

$$\min_{\mu_i} \sum_i \sum_j \left(Y_{ij} - \mu_i\right)^2$$

Η ελαχιστοποίηση γίνεται ως προς τους μέσους μ_i και παίρνουμε τις εκτιμήτριες

$$\hat{\mu}_i = \frac{1}{n_i}\left(Y_{i1} + Y_{i2} + \ldots + Y_{in_i}\right) = \frac{Y_{i.}}{n_i} = \bar{Y}_{i.}$$

Οι εκτιμημένες τιμές των Y_{ij} είναι:

$$\hat{Y}_{ij} = \bar{Y}_{i.}$$

Μοντέλο σταθερών επιδράσεων

Άρα

$$SSE = R_0^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \widehat{Y}_{ij}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{i.}\right)^2 = \sum_{i=1}^{k}\left(\sum_{j=1}^{n_i} Y_{ij}^2 - n_i \overline{Y}_{i.}^2\right)$$

$$\widehat{Y}_{i.} = \overline{Y}_{i.}$$

$$SST_{total} = R_1^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{..}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - n\overline{Y}_{..}^2$$

$$\widehat{\overline{Y}} = \overline{Y}_{..}$$

$$SST_{tr} = R_1^2 - R_0^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\widehat{Y}_{ij} - \overline{Y}_{..}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\overline{Y}_{i.} - \overline{Y}_{..}\right)^2 = \sum_{i=1}^{n_i} n_i\left(\overline{Y}_{i.}\right)^2 - n\left(\overline{Y}_{..}\right)^2$$

Μοντέλο σταθερών επιδράσεων

Αν το αρχικό μοντέλο μας είναι σωστό τότε ισχύει:

$$E\left(Y_{ij}^2\right) = \left(\mu + a_i\right)^2 + \sigma^2$$

$$E\left(\overline{Y}_{i.}^2\right) = \left(\mu + a_i\right)^2 + \frac{\sigma^2}{n_i}$$

$$E\left(\overline{Y}_{..}^2\right) = \mu^2 + \frac{\sigma^2}{n}$$

Μοντέλο σταθερών επιδράσεων

Η εκτίμηση του σ^2 δίνεται από την σχέση

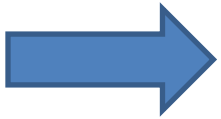
$$s^2 = \sigma^2 = \frac{SSE}{n-k}$$

και τα SST_{tr} και SSE είναι ανεξάρτητα.

Αν ισχύει η H_0 τότε

$$E(MSE) = \frac{SSE}{n-k} = \sigma^2$$

$$E(MST_{tr}) = \frac{SST_{tr}}{k-1} = \sigma^2$$



$$F^* = \frac{MST_{tr}}{MSE} \sim F_{k-1,n-k}$$

Μοντέλο σταθερών επιδράσεων

Τα SST_{tr} και SSE είναι ανεξάρτητα

$$Cov\left(Y_{ij}-\bar{Y}_{i.},\bar{Y}_{i.}-\bar{Y}_{..}\right)=Cov\left(Y_{ij}-\bar{Y}_{i.}\right)-Cov\left(Y_{ij}-\bar{Y}_{..}\right)-Cov\left(\bar{Y}_{i.}-\bar{Y}_{i.}\right)+Cov\left(\bar{Y}_{i.}-\bar{Y}_{..}\right)=$$

$$=\frac{\sigma^2}{n_i}-\frac{\sigma^2}{n}-\frac{\sigma^2}{n_i}+\frac{\sigma^2}{n}=0$$

Αφού τα σ^2 είναι άγνωστα έχουμε:

$$E\left(MSE\right)=\left(n-k\right)\sigma^2=E\left(\frac{SSE}{n-k}\right)=\sigma^2$$

Αμερόληπτος εκτιμητής της σ^2

Μοντέλο σταθερών επιδράσεων

Πρόταση

$$E\left(MST_{tr}\right)=E\left(\sum_{i=1}^{k}n_i\left(\overline{Y}_{i.}-\overline{Y}_{..}\right)^2\right)=\sigma^2+\frac{1}{k-1}\sum_{i=1}^{k}n_i a_i^2$$

Απόδειξη

Έχουμε

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

Αλλά

$$E\left(\overline{Y}_{i.}-\overline{Y}_{..}\right)^2=Var\left(\overline{Y}_{i.}-\overline{Y}_{..}\right)+E^2\left(\overline{Y}_{i.}-\overline{Y}_{..}\right)$$

Μοντέλο σταθερών επιδράσεων

Έχουμε

$$E\left(\bar{Y}_{i.}-\bar{Y}_{..}\right)=\left(\mu+a_i\right)-\left(\mu\right)=a_i$$

$$Var\left(\bar{Y}_{i.}-\bar{Y}_{..}\right)=Var\left(\bar{Y}_{i.}\right)+Var\left(\bar{Y}_{..}\right)-2Cov\left(\bar{Y}_{i.},\bar{Y}_{..}\right)=$$

$$=\frac{\sigma^2}{n_i}+\frac{\sigma^2}{n}-2\frac{\sigma^2}{n}=\frac{\sigma^2}{n_i}-\frac{\sigma^2}{n}$$

Άρα

$$E\left(MST_{tr}\right)=\sum_{i=1}^{k}n_i\frac{\left\{\left(\frac{\sigma^2}{n_i}-\frac{\sigma^2}{n}\right)+a_i^2\right\}}{k-1}=\sum_{i=1}^{k}\frac{\left(\sigma^2-\frac{n_i}{n}\sigma^2+n_ia_i^2\right)}{k-1}=$$

$$=\frac{k\sigma^2-\sigma^2}{k-1}+\frac{1}{k-1}\sum_{i=1}^{k}n_ia_i^2=\sigma^2+\frac{1}{k-1}\sum_{i=1}^{k}n_ia_i^2$$

Μοντέλο σταθερών επιδράσεων

Κάτω από την μηδενική υπόθεση $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_k = 0$ ισχύει

$$E(MST_{tr}) = \sigma^2$$

$$E(MST_{tr}) > \sigma^2$$

Αμερόληπτος εκτιμητής της σ^2 κάτω από την H_0

Αν δεν ισχύει η H_0

Πηγή μεταβλητότητας	Άθροισμα τετραγώνων	Βαθμοί ελευθερίας	Μέσο άθροισμα τετραγώνων	F
Απόκλιση από την H_0 (μεταξύ δειγμάτων)	SST_{tr}	k-1	MST_{tr}	$F^* = \frac{MST_{tr}}{MSE}$
Κατάλοιπα (εντός δειγμάτων)	SSE	n-k	$MSE = s^2$	
Σύνολο	SST	n-1		

$$F^* \to \infty$$

Μεγάλες τιμές αν δεν ισχύει η H_0

Παράδειγμα

Πίνακας Ανάλυσης Διακύμανσης

Πηγή μεταβλητότητας	Άθροισμα τετραγώνων	Βαθμοί ελευθερίας	Μέσο άθροισμα τετραγώνων	F
Απόκλιση από την H_0 (μεταξύ δειγμάτων)	521.262	4	130.315	9.0801 ($p<0.0001$)
Κατάλοιπα (εντός δειγμάτων)	631.508	44	14.352	
Σύνολο	11552.763	48		

Επομένως απορρίπτουμε ότι όλοι οι μέσοι είναι ίσοι.

Η εκτίμηση για την διασπορά των σφαλμάτων είναι:

$$\sigma^2 = 14.352$$

Έλεγχος διακύμανσης

Η σωστή διαδικασία είναι πρώτα ελέγχουμε αν το μοντέλο Y_{ij} είναι σωστό δηλ:

- αν τα σφάλματα είναι ανεξάρτητα με σταθερή διασπορά
- αν ακολουθούν κανονική κατανομή
- αν δεν υπάρχουν άλλοι παράγοντες που επηρεάζουν το αποτέλεσμα

Έλεγχος διακύμανσης

Έλεγχος σφαλμάτων αν έχουν τις ίδιες διακυμάνσεις σε όλες τις ομάδες, θεωρώντας ότι μέσα σε κάθε ομάδα η διασπορά είναι σταθερή.

Κριτήριο Cochran, Barlett , Hartley

Η εκτίμηση της διασποράς σ_i^2 κάθε ομάδας διαίτης είναι:

$$\sigma_i^2 = \frac{SSE_i}{n_i - 1}, \text{με } \mathrm{SSE_i} = \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{i.}\right)^2$$

Το SSE αναλύεται σε k αθροίσματα τετραγώνων

$$\mathrm{SSE} = \mathrm{SSE}_1 + \mathrm{SSE}_2 + ... + \mathrm{SSE}_k, SSE_i \sim \sigma^2 x^2_{n_i - 1}$$

Έλεγχος διακύμανσης

Τα διαστήματα εμπιστοσύνης 100(1-α)% μπορούμε να τα υπολογίσουμε με δύο τρόπους:

- ✓ Χρησιμοποιώντας την εκτίμηση της διασποράς μέσα σε κάθε ομάδα

$$\bar{Y}_{i.} \pm t_{n_1-1,1-\alpha/2}\sqrt{\frac{\hat{\sigma}_i^2}{n_i}}$$

- ✓ Χρησιμοποιώντας την εκτίμηση της διασποράς από όλες τις παρατηρήσεις

$$\bar{Y}_{i.} \pm t_{n-k,1-\alpha/2}\sqrt{\frac{\hat{\sigma}_i^2}{n_i}}$$

- ✓ Η διαφορά δίνεται από την σχέση

$$\bar{Y}_{i.} - \bar{Y}_{.j} \pm t_{n-k,1-\alpha/2}\hat{\sigma}^2\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}$$

Έλεγχος διακύμανσης

Διάιτα	n_i	$\bar{Y}_{i.}$	$\hat{\sigma}_i/\sqrt{n_i}$	$\hat{\sigma}/\sqrt{n_i}$	95% δ.ε των μ_i	
1	7	28.54	1.329	1.643	25.21	31.85
2	9	27.71	1.383	1.449	24.79	30.63
3	10	29	1.438	1.374	26.23	31.77
4	12	31.51	0.86	1.094	29.95	33.07
5	11	22.64	1.111	1.142	20.63	23.89
Σύνολο	49	27.8	0.541	0.541	27.03	28.57

Έλεγχος ομοσκεδαστικότητας των παρατηρήσεων

Για να είναι αξιόπιστα όλα τα παραπάνω θα πρέπει τα σφάλματα να έχουν την ίδια διασπορά σ^2.

- Μπορούμε και πάλι να χρησιμοποιήσουμε το **Levene's Τεστ ομοσκεδαστικότητας**
- Το τεστ αυτό βασίζεται στην ανάλυση διασποράς

Βασίζεται στις τυχαίες μεταβλητές

$$W_{ij} = | Y_{ij} - \bar{Y}_{i\bullet} |, j = 1, 2, ..., n_i, i=1,2,...,k.$$

Αν τα Y_{ij} έχουν την ίδια διασπορά σε όλες τις στάθμες, τότε τα W_{ij} θα έχουν την ίδια μέση τιμή σε όλες τις στάθμες του παράγοντα Χ. Επομένως αρκεί να ελέγξουμε αν οι τ.μ. W_{ij} έχουν την ίδια μέση τιμή σε όλες τις στάθμες του παράγοντα Χ.

Έλεγχος ομοσκεδαστικότητας των παρατηρήσεων

Το Levene τεστ ουσιαστικά είναι το F-ratio τεστ του πίνακα ANOVA που αντιστοιχεί στο μοντέλο ανάλυσης διασποράς της W ως προς τον παράγοντα X.

$$F = \frac{SSTr_W/(k-1)}{SSE_W/(n-k)} = \frac{\sum_{i,j}(\overline{W}_{i\bullet} - \overline{W})^2/(k-1)}{\sum_{i,j}(W_{ij} - \overline{W}_{i\bullet})^2/(n-k)}.$$

Αν, με βάση αυτό το F-ratio τεστ, απορρίπτεται ότι ο μέσος της W είναι σταθερός σε όλες τις στάθμες της X τότε απορρίπτεται ότι και η διασπορά της Y είναι σταθερή σε όλες τις στάθμες της X.

Έλεγχος ομοσκεδαστικότητας των παρατηρήσεων

Σε περίπτωση

- ✓ Μη κανονικότητας των παρατηρήσεων
- ✓ Οι διακυμάνσεις διαφέρουν (τεστ Levene)
- ✓ Ο αριθμός των παρατηρήσεων δεν είναι ίδιος σε κάθε δείγμα

τότε χρησιμοποιούμε το **τεστ των Brown – Forsythe** και το **τεστ του Welch**

Έλεγχος ομοσκεδαστικότητας των παρατηρήσεων

Τεστ των Brown – Forsythe

Ο έλεγχος των Brown-Forsythe βασίζεται στο κριτήριο:

$$F_{BF} = \frac{\sum_{j=1}^{k} n_j \left(\overline{Y}_j - \overline{Y}\right)^2}{\sum_{j=1}^{k} \left(1 - \frac{n_j}{N}\right) s_j^2}, \; j = 1,2,\ldots,k$$

όπου κάτω από την H_0 ακολουθεί την F κατανομή με k, f βαθμούς ελευθερίας

Έλεγχος ομοσκεδαστικότητας των παρατηρήσεων

Τεστ των Brown – Forsythe

όπου

$$\frac{1}{f} = \sum_{j=1}^{k} \frac{c_j^2}{(n_j - 1)} \text{ και } c_j = \frac{\left(1 - \frac{n_j}{N}\right) s_j^2}{\sum_{j=1}^{k} \left(1 - \frac{n_j}{N}\right) s_j^2}$$

Έλεγχος ομοσκεδαστικότητας των παρατηρήσεων

Τεστ των Brown – Forsythe

Εναλλακτικά ισχύει

$$F_{BF} = \frac{(n-k)\sum_{i=1}^{k} n_i \left(Z_i - \bar{Z}\right)^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(Z_{ij} - \bar{Z}_i\right)^2}$$

όπου

$$Z_{ij} = \left|Y_{ij} - \bar{Y}_i\right|$$

$$\bar{Z}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Z_{ij}$$

$$\bar{Z} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Z_{ij}$$

$$F_{BF} \sim F_{k-1,n-k}$$