Γραμμικά Μοντέλα
Ανάλυση Διασποράς - ANOVA

Διδάσκουσα: Λουκία Μελιγκοτσίδου
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
Τμήμα Μαθηματικών

May 2, 2020

Παλινδρόμηση και Ανάλυση Διασποράς

Η ανάλυση παλινδρόμησης μελετά τη στατιστική σχέση ανάμεσα σε μία ή περισσότερες ανεξάρτητες μεταβλητές και μια εξαρτημένη μεταβλητή. Συγκεκριμένα, η αναμενόμενη τιμή της εξαρτημένης μεταβλητής εκφράζεται ως γραμμική συνάρτηση των ανεξάρτητων μεταβλητών. Η ανάλυση διακύμανσης είναι ένα πιο γενικό στατιστικό εργαλείο. Επίσης μελετά τη στατιστική σχέση ανάμεσα σε μία ή περισσότερες ανεξάρτητες μεταβλητές και μια εξαρτημένη μεταβλητή, χωρίς όμως να υποθέτει απαραίτητα κάποιο συγκεκριμένο μοντέλο για την περιγραφή της σχέσης αυτής.

Είδαμε την ανάλυση διακύμανσης στα πλαίσια του γραμμικού μοντέλου. Γενικά η ανάλυση διακύμανσης στα πλαίσια κάποιου παραμετρικού μοντέλου χρησιμοποιείται σαν ένα μετρο καλής προσαρμογής. Δείχνει κατά πόσο η μεταβλητότητα της εξαρτημένης μεταβλητής εξηγείται από το μοντέλο που υποθέσαμε.

Συχνά όμως στην ανάλυση διακύμανσης οι ανεξάρτητες μεταβλητές είναι ποιοτικές (παράγοντες) και το ενδιαφέρον μας εστιάζει στο κατά πόσο ο κάθε παράγοντας και τα επίπεδά του επηρεάζουν κάποια απαντητική μεταβλητή. Η ανάλυση διακύμανσης κατά παράγοντες χρησιμοποιείται πολύ στο σχεδιασμό πειραμάτων.

Πειραματικός Σχεδιασμός

Πείραμα είναι μια δοκιμή ή ένα σύνολο δοκιμών στις οποίες σκόπιμες αλλαγές γίνονται στα επίπεδα των παραγόντων που επηρεάζουν μια διαδικασία με σκοπό την παρατήρηση και αξιολόγηση των αλλαγών που συνεπάγονται για την απαντητική μεταβλητή. Η απαντητική μεταβλητή είναι μια στατιστική μεταβλητή, η οποία εκφράζει τη λειτουργία της υπό μελέτη διαδικασίας. Τα αποτελέσματα του πειράματος συνοψίζονται σε ένα σύνολο από τιμές για την απαντητική μεταβλητή για κάποιες πειραματικές μονάδες. Επομένως, οι πειραματικές μονάδες, οι οποίες καθορίζονται με τυχαία δειγματοληψία πριν από τη διεξαγωγή του πειράματος, είναι τα στοιχεία του πειράματος για τα οποία έχουμε παρατηρήσεις.

Ο σχεδιασμός πειραμάτων πραγματοποιείται με σκοπό την βελτίωση της υπό μελέτη διαδικασίας. Αυτό περιλαμβάνει την επίτευξη της βέλτιστης αναμενόμενης τιμής για την απαντητική μεταβλητή και την ελαχιστοποίηση της διακύμανσής της. Με συνεχή πειράματα ελέγχονται οι παράγοντες (προσδιορίσιμα ποιοτικά χαρακτηριστικά) που επηρεάζουν την διαδικασία, ώστε να επιτευχθεί το καλυτερο δυνατό αποτέλεσμα. Ένας προσδιορίσιμος παράγοντας είναι δηλαδή μια ποιοτική μεταβλητή η οποία εκ προθέσεως ελέγχεται σε ένα πείραμα ώστε να παρατηρηθεί η επίδρασή της στην απαντητική μεταβλητή. Κάθε παράγοντας εξετάζεται σε επίπεδα, δηλαδή σε ένα σύνολο προκαθορισμένων τιμών.

Παράδειγμα

Έστω η διαδικασία μιας διδασκαλίας (για παράδειγμα η διδασκαλία των μαθητών της Γ' γυμνασίου). Παράγοντες που επηρεάζουν την εκπαιδευτική διαδικασία είναι:

- το εκπαιδευτικό υλικό (βιβλία, διδάσκοντες, κ.λ.π),
- το γνωστικό υπόβαθρο των μαθητών και
- οι εγκαταστάσεις (κτήρια, τεχνολογικός εξοπλισμός).

Οι παραπάνω προσδιορίσιμοι (ελέγξιμοι) παράγοντες εξετάζονται σε επίπεδα. Αξιολογούμε την εκπαιδευτική διαδικασία χρησιμοποιώντας ως πειραματικές μονάδες ένα δείγμα μαθητών. Απαντητική μεταβλητή είναι ο βαθμός των μαθητών στις εξετάσεις.

Στη διαδικασία υπεισέρχονται, όμως, και μη ελέγξιμοι παράγοντες. Στα πλαίσια ενός στατιστικού μοντέλου οι μη ελέγξιμοι/ μη μετρήσιμοι παράγοντες συνοψίζονται σε έναν στοχαστικό όρο.

Ανάλυση Διασποράς κατά έναν Παράγοντα

Έστω ότι μας ενδιαφέρει να μελετήσουμε μια διαδικασία ως προς τα επίπεδα ενός μόνο παράγοντα. Θεωρούμε λοιπόν ότι η απαντητική μεταβλητή για το i επίπεδο του παράγοντα ακολουθεί κανονική κατανομή $N(\mu_i, \sigma^2)$, $i = 1, \ldots, m$, όπου m είναι ο αριθμός των επιπέδων του υπό μελέτη παράγοντα. Σκοπός μας είναι να διαπιστώσουμε αν όντως τα διαφορετικά επίπεδα του παράγοντα επηρεάζουν τη διαδικασία, δηλαδή αν όντως οι μέσοι μ_i των επιπέδων διαφέρουν, ή αν ο μέσος είναι σταθερός για όλα τα επίπεδα του παράγοντα. Στο ερώτημα αυτό θα απαντήσουμε χρησιμοποιώντας στατιστικά εργαλεία, βασιζόμενοι σε ένα τυχαίο δείγμα τιμών της απαντητικής μεταβλητής από τα διάφορα επίπεδα του παράγοντα.

Στο i επίπεδο του παράγοντα λαμβάνουμε δείγμα μεγέθους n_i, $i = 1, \ldots, m$. Τα μεγέθη των δειγμάτων στα διάφορα επίπεδα δεν είναι απαραίτητα ίσα. Οι παρατηρήσεις του δείγματος μέσα σε κάθε επίπεδο και μεταξύ των επιπέδων πρέπει να είναι ανεξάρτητες.

Επίπεδα					Σύνολα	Μέσοι
1	Y_{11}	Y_{12}	$\ldots$	Y_{1n_1}	$Y_{1.}$	$\bar{Y}_{1.}$
2	Y_{21}	Y_{22}	$\ldots$	Y_{2n_2}	$Y_{2.}$	$\bar{Y}_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
m	Y_{m1}	Y_{m2}	$\ldots$	Y_{mn_m}	$Y_{m.}$	$\bar{Y}_{m.}$
					$Y_{..}$	$\bar{Y}_{..}$

όπου Y_{ij} είναι η j παρατήρηση του i επιπέδου, $n = n_1 + n_2 + \ldots + n_m$ είναι ο συνολικός αριθμός των παρατηρήσεων, και

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

$$Y_{..} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} Y_{ij}, \quad \bar{Y}_{..} = \frac{1}{n} \sum_{i=1}^{m} n_i \bar{Y}_{i.}$$

Το μοντέλο ανάλυσης διασποράς κατά έναν παράγοντα είναι

$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2),$$

όπου τα ϵ_{ij}, $j = 1, \ldots, n_i$, $i = 1, \ldots, m$, είναι ανεξάρτητα και ταυτοτικά κατανεμημένα τυχαία σφάλματα και οι μέσοι των επιπέδων μ_i, $i = 1, \ldots, m$, είναι οι άγνωστες παράμετροι του μοντέλου. Δηλαδή, τα διάφορα επίπεδα του ύπό μελέτη παράγοντα διαφέρουν ως προς το μέσο τους αλλά έχουν κοινή διασπορά.

Έλεγχος Ισότητας Μέσων

Η υπόθεση που μας ενδιαφέρει να ελέγξουμε είναι

$$\begin{array}{ll} H_0: & \mu_i = \mu \qquad \text{για } i = 1, \ldots, m \\ H_1: & \mu_i \neq \mu \quad \text{για ένα τουλάχιστον } i \end{array}$$

Για να ελέγξουμε την παραπάνω υπόθεση θα αναλύσουμε τη συνολική διασπορά των δεδομένων σε δύο συνιστώσες: τη διασπορά μέσα στα επίπεδα και τη διασπορά ανάμεσα στα επίπεδα. Διαισθητικά περιμένουμε ότι αν η διασπορά ανάμεσα στα επίπεδα είναι μεγαλύτερη από τη διασπορά μέσα στα επίπεδα, τότε ο μέσος δεν θα είναι σταθερός για όλα τα επίπεδα.

Η συνολική μεταβλητότητα του δείγματος εκφράζεται από το άθροισμα των τετραγώνων των αποκλίσεων των παρατηρήσεων από το συνολικό μέσο του δείγματος, δηλαδή το συνολικό αθροισμα τετραγώνων (total sum of squares)

$$SST = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2$$

Παρατηρώντας ότι η απόκλιση της παρατήρησης Y_{ij} από το συνολικό μέσο του δείγματος $\bar{Y}_{..}$ μπορεί να γραφτεί ως

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

έχουμε ότι

$$\begin{aligned} SST &= \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left[(Y_{ij} - \bar{Y}_{i.}) + (\bar{Y}_{i.} - \bar{Y}_{..}) \right]^2 \\ &= \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 + \sum_{i=1}^{m} \sum_{j=1}^{n_i} (\bar{Y}_{i.} - \bar{Y}_{..})^2 + 2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})(\bar{Y}_{i.} - \bar{Y}_{..}). \end{aligned}$$

Αλλά

$$\begin{aligned} 2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})(\bar{Y}_{i.} - \bar{Y}_{..}) &= 2 \sum_{i=1}^{m} \left[(\bar{Y}_{i.} - \bar{Y}_{..}) \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.}) \right] = \\ 2 \sum_{i=1}^{m} \left[(\bar{Y}_{i.} - \bar{Y}_{..}) \left(\sum_{j=1}^{n_i} Y_{ij} - n_i Y_{i.} \right) \right] &= 2 \sum_{i=1}^{m} \left[(\bar{Y}_{i.} - \bar{Y}_{..})(n_i Y_{i.} - n_i Y_{i.}) \right] = 0. \end{aligned}$$

Άρα

$$SST = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{Y}_{i.} - \bar{Y}_{..})^2,$$

όπου το άθροισμα των τετραγώνων των αποκλίσεων των παρατηρήσεων από τους αντίστοιχους μέσους των επιπέδων στα οποία ανήκουν, $SSE = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2$, εκφράζει τη μεταβλητότητα μέσα στα επίπεδα του παράγοντα, και το άθροισμα των τετραγώνων των αποκλίσεων των μέσων των επιπέδων από το συνολικό μέσο του δείγματος, $SSF = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{Y}_{i.} - \bar{Y}_{..})^2 = \sum_{i=1}^{m} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2$, εκφράζει τη μεταβλητότητα ανάμεσα στα επίπεδα. Δηλαδή,

$$SST = SSE + SSF.$$

Η μεταβλητότητα ανάμεσα στα επίπεδα είναι αυτή που οφείλεται στις διαφορές των επιπέδων του παράγοντα και εκφράζεται από το συνολικό άθροισμα τετραγώνων του παράγοντα (factor sum of squares). Η μεταβλητότητα μέσα στα επίπεδα οφείλεται στην τυχαιότητα και εκφράζεται από το συνολικό άθροισμα τετραγώνων των τυχαίων σφαλμάτων (error sum of squares).

Οι συνολικοί βαθμοί ελευθερίας στο πείραμα είναι $n-1$. Μέσα σε κάθε επίπεδο οι βαθμοί ελευθερίας είναι $n_i - 1$, άρα συνολικά μέσα στα επίπεδα έχουμε $n - m$ βαθμούς ελευθερίας. Οι βαθμοί ελευθερίας ανάμεσα στα επίπεδα είναι $n - 1 - (n - m) = m - 1$.

Για κάθε επίπεδο ορίζουμε τη συνάρτηση

$$W_i^2 = \frac{\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2}{n_i - 1}, \quad i = 1, \ldots, m.$$

Οι παρατηρήσεις του i επιπέδου αποτελούν δείγμα από την κανονική κατανομή $N(\mu_i, \sigma^2)$. Οπότε έχουμε ότι η ποσότητα $\frac{\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2}{\sigma^2} = \frac{(n_i-1)W_i^2}{\sigma^2}$ ακολουθεί κατανομή X^2 με $n_i - 1$ βαθμούς ελευθερίας, δηλαδή

$$\frac{(n_i - 1)W_i^2}{\sigma^2} \sim X^2_{n_i - 1}.$$

Άρα

$$E\left[\frac{(n_i - 1)W_i^2}{\sigma^2}\right] = n_i - 1 \Rightarrow E(W_i^2) = \sigma^2.$$

Επομένως, η συνάρτηση W_i^2 είναι μια αμερόληπτη εκτιμήτρια της διασποράς σ^2, βασισμένη στο δείγμα από το i επίπεδο.

Τώρα, το άθροισμα m ανεξάρτητων τυχαίων μεταβλητών που ακολουθούν κατανομή X^2 είναι τυχαία μεταβλητή που επίσης ακολουθεί κατανομή X^2, με βαθμούς ελευθερίας ίσους με το άθροισμα των βαθμών ελευθερίας των m αρχικών κατανομών, δηλαδή

$$\sum_{i=1}^{m} \frac{(n_i - 1)W_i^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 = \frac{SSE}{\sigma^2} \sim X^2_{n-m}.$$

Άρα

$$E\left[\frac{SSE}{\sigma^2}\right] = n - m \Rightarrow E\left(\frac{SSE}{n-m}\right) = \sigma^2.$$

Επομένως, η συνάρτηση $\frac{SSE}{n-m}$ είναι γενικά μια αμερόληπτη εκτιμήτρια της διασποράς σ^2.

Κάτω από την αρχική υπόθεση H_0 όλες οι παρατηρήσεις Y_{ij} είναι δείγμα από την ίδια κανονική κατανομή $N(\mu, \sigma^2)$. Έστω η συνάρτηση

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 = \frac{SST}{n-1}.$$

Η ποσότητα $\frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2}$ ακολουθεί κατανομή X^2 με $n-1$ βαθμούς ελευθερίας, δηλαδή

$$\frac{(n-1)S^2}{\sigma^2} \sim X^2_{n-1}.$$

Άρα

$$E\left[\frac{(n-1)S^2}{\sigma^2}\right] = n - 1 \Rightarrow E\left(\frac{SST}{n-1}\right) = \sigma^2.$$

Επομένως, η συνάρτηση $S^2 = SST/n-1$, κάτω από την H_0, είναι μια αμερόληπτη εκτιμήτρια της διασποράς σ^2.

Έχουμε

$$\frac{SST}{\sigma^2} = \frac{SSE}{\sigma^2} + \frac{SSF}{\sigma^2}.$$

Από το θεώρημα τετραγωνικών μορφών, κάτω από την H_0, η συνάρτηση SSF/σ^2 ακολουθεί X^2 κατανομή με $m-1$ βαθμούς ελευθερίας. Άρα

$$E\left[\frac{SSF}{\sigma^2}\right] = m-1 \Rightarrow E\left(\frac{SSF}{m-1}\right) = \sigma^2,$$

δηλαδή η συνάρτηση $SSF/m-1$, κάτω από την H_0, είναι μια αμερόληπτη εκτιμήτρια της διασποράς σ^2.

Επομένως, κάτω από την H_0, έχουμε ότι

$$\frac{SSF/m-1}{SSE/n-m} \sim F_{m-1,n-m}.$$

Η ελεγχοσυνάρτηση $F_0 = \frac{SSF/m-1}{SSE/n-m}$ μπορεί να χρησιμοποιηθεί για τον έλεγχο ισότητας μέσων στην ανάλυση διασποράς. Απορρίπτουμε την αρχική υπόθεση H_0 σε επίπεδο στατιστικής σημαντικότητας α αν η παρατηρούμενη τιμή της ελεγχοσυνάρτησης είναι μεγαλύτερη από το α ποσοστιαίο σημείο της κατανομής F με $m-1$ και $n-m$ βαθμούς ελευθερίας, δηλαδή αν

$$F_0 = \frac{SSF/m-1}{SSE/n-m} > F_{\alpha,m-1,n-m}.$$

(Ισοδύναμα $p-value < \alpha$).

Ακραία (στην ουρά της F κατανομής) παρατηρούμενη τιμή της ελεγχοσυνάρτησης σημαίνει ότι μεγάλο μέρος της συνολικής μεταβλητότητας των δεδομένων οφείλεται στους παράγοντες (μεταβλητότητα ανάμεσα στα επίπεδα - SSF) συγκριτικά με το μέρος που οφείλεται στους τυχαίους όρους (μεταβλητότητα μέσα στα επίπεδα - SSE). Αυτό αποτελεί ένδειξη εναντίον της H_0.

Σημείωση: Η εκτιμήτρια του σ^2 που βασίζεται στο SSE είναι πάντα αμερόληπτη. Όμως η εκτίμητρια που βασίζεται στο SSF είναι αμερόληπτη μόνο κάτω από την H_0. Αν η H_0 δεν είναι αληθής τότε

$$\frac{1}{m-1}E[SSF] > \sigma^2.$$

Απόδειξη:

$$\begin{aligned}
E[SSF] &= E[\sum_{i=1}^{m} n_i(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2] \\
&= E[\sum_{i=1}^{m} n_i\overline{Y}_{i\cdot}^2 - 2\sum_{i=1}^{m} n_i\overline{Y}_{i\cdot}\overline{Y}_{\cdot\cdot} + n\overline{Y}_{\cdot\cdot}^2] \\
&\overset{\sum_{i=1}^{m} n_i\overline{Y}_{i\cdot}=n\overline{Y}_{\cdot\cdot}}{=} E[\sum_{i=1}^{m} n_i\overline{Y}_{i\cdot}^2 - 2n\overline{Y}_{\cdot\cdot}^2 + n\overline{Y}_{\cdot\cdot}^2] \\
&= E[\sum_{i=1}^{m} n_i\overline{Y}_{i\cdot}^2 - n\overline{Y}_{\cdot\cdot}^2] \\
&= \sum_{i=1}^{m} n_i E[\overline{Y}_{i\cdot}^2] - nE[\overline{Y}_{\cdot\cdot}^2] \\
&= \sum_{i=1}^{m} n_i\big[\Delta[\overline{Y}_{i\cdot}] + E^2[\overline{Y}_{i\cdot}]\big] - n\big[\Delta[\overline{Y}_{\cdot\cdot}] + E^2[\overline{Y}_{\cdot\cdot}]\big] \\
&= \sum_{i=1}^{m} n_i\big[\frac{\sigma^2}{n_i} + \mu_i^2\big] - n\big[\frac{\sigma^2}{n} + \mu^2\big] \\
&= (m-1)\sigma^2 + \sum_{i=1}^{m} n_i\big[\mu_i^2 - \mu^2\big]
\end{aligned}$$

Υποθέσεις του μοντέλου ANOVA

- Η κατανομή των παρατηρήσεων σε κάθε επίπεδο είναι κανονική

$$Y_{ij} \sim N(\mu_i, \sigma^2),\ j = 1, \ldots, n_i,\ \text{για το επίπεδο } i.$$

- Η κανονική κατανομή σε κάθε επίπεδο έχει την ίδια διασπορά σ^2. Ισοδύναμα, οι τυχαίοι όροι, ϵ_{ij}, είναι ταυτοτικά κατανεμημένοι $\epsilon_{ij} \sim N(0, \sigma^2)$ για κάθε i, j.
- Οι παρατηρήσεις σε κάθε επίπεδο του παράγοντα είναι ανεξάρτητες και ταυτοτικά κατανεμημένες και είναι ανεξάρτητε από τις παρατηρήσεις στα άλλα επίπεδα.

Στόχοι:

- Έλεγχος αν οι μέσοι των επιπέδων είναι ίσοι

$$\begin{aligned} H_0 &: \mu_i = \mu_j,\ \text{για όλα τα } i, j \\ H_1 &: \mu_i \neq \mu_j,\ \text{για ένα τουλάχιστον ζεύγος } i, j \end{aligned}$$

- Αν οι μέσοι των επιπέδων δεν είναι ίσοι, έλεγχος για το ποιές είναι οι διαφορές.

Αν οι μέσοι των επιπέδων δεν διαφέρουν (στατιστικά σημαντικά), το συμπέρασμα είναι ότι η απαντητική μεταβλητή δεν εξαρτάται από τα επίπεδα του παράγοντα.

Έλεγχος για την ισότητα των διασπορών

Bartlett's test

Η ανάλυση διασποράς υποθέτει ότι οι διασπορές των κανονικών κατανομών σε όλα τα επίπεδα είναι ίσες. Ο Bartlett πρότεινε τον παρακάτω έλεγχο για την υπόθεση αυτή.

$$\begin{aligned} H_0 &: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_m^2 = \sigma^2 \\ H_1 &: \sigma_i^2 \neq \sigma^2, \text{ για ένα τουλάχιστον } i \end{aligned}$$

Για τον έλεγχο χρησιμοποιείται η στατιστική συνάρτηση

$$X_0^2 = 2.3026\frac{q}{c},$$

όπου

$$\begin{aligned} q &= (n-m)\log S_p^2 - \sum_{i=1}^{m}(n_i-1)\log S_i^2 \\ c &= 1 + \frac{1}{3(m-1)}\Big(\sum_{i=1}^{m}(n_i-1)^{-1} - (n-m)^{-1}\Big) : \text{σταθερά} \end{aligned}$$

S_i^2 η δειγματική διασπορά του i επιπέδου

$$S_p^2 = \frac{1}{n-m}\sum_{i=1}^{m}(n_i-1)S_i^2 : \text{σταθμισμένη διασπορά}$$

Η ποσότητα q είναι ίση με 0 όταν τα S_i^2 είναι ίσα και γίνεται μεγάλη όταν τα S_i^2 διαφέρουν πολύ. Επομένως, μεγάλες παρατηρούμενες τιμές της ελεγχοσυνάρτησης αποτελούν ένδειξη εναντίον της H_0. Απορρίπτουμε την H_0 για ακραίες τιμές του X_0^2, δηλαδή για $X_0^2 > X_\alpha^2(m-1)$ (ή ισοδύναμα αν $p-value < \alpha$).

Παρατήρηση. Οι υποθέσεις του μοντέλου μπορούν να ελεγχθούν γραφικά. Κατάλοιπα : $\hat{\epsilon}_{ij} = Y_{ij} - \hat{\mu}_i = Y_{ij} - \overline{Y}_{i\cdot}$.

Γραφικοί έλεγχοι καταλοίπων $\left\{ \begin{array}{l} \text{P-P plot για την κανονικότητα} \\ \text{Scatter plot για ομοσκεδαστικότητα και τυχαιότητα–ανεξαρτησία} \end{array} \right.$

Επιμέρους έλεγχοι υποθέσεων για τους μέσους

Αν το F–test για την ισότητα των μέσων δείξει ότι οι μέσοι των επιπέδων διαφέρουν, θα πρέπει να προχωρήσουμε με την ανάλυσή μας.

Μια αμερόληπτη εκτιμήτρια του μέσου του i επίπεδου, μ_i, είναι η

$$\hat{\mu}_i = \overline{Y}_{i\cdot} \text{ (δειγματικός μέσος του } i \text{ επιπέδου)}$$

η οποία έχει διασπορά $\sigma^2_{\overline{Y}_{i\cdot}} = \frac{\sigma^2}{n_i}$, με αντίστοιχη αμερόληπτη εκτιμήτρια $S^2_{\overline{Y}_{i\cdot}} = \frac{MSE}{n_i}$, καθώς το MSE είναι αμερόληπτη εκτιμήτρια του σ^2.

Εφόσον $\overline{Y}_{i\cdot} \sim N(\mu_i, \frac{\sigma^2}{n_i})$ έπεται ότι $\frac{\overline{Y}_{i\cdot}-\mu_i}{\sqrt{MSE/n_i}} \sim t(n-m)$, όπου $n-m$ είναι οι βαθμοί ελευθερίας που συνδέονται με το MSE. Το διάστημα εμπιστοσύνης δίνεται ως

$$\overline{Y}_{i\cdot} - \sqrt{MSE/n_i}\, t_{\alpha/2}(n-m) \le \mu_i \le \overline{Y}_{i\cdot} + \sqrt{MSE/n_i}\, t_{\alpha/2}(n-m).$$

Από τέτοια διαστήματα εμπιστοσύνης για όλα τα μ_i παίρνουμε μια πρώτη εικόνα για το πώς διαφέρουν οι μέσοι των επιπέδων.

Εκτίμηση της διαφοράς των μέσων δυο επιπέδων, $D = \mu_i - \mu_j$: $\overline{D} = \overline{Y}_{i\cdot} - \overline{Y}_{j\cdot}$. Η τυχαία μεταβλητή $\overline{D}$ ακολουθεί κανονική κατανομή ως γραμμικός συνδυασμός ανεξάρτητων κανονικών τυχαίων μεταβλητών. Αφού τα $\overline{Y}_{i\cdot}$, $\overline{Y}_{j\cdot}$ είναι ανεξάρτητα η διασπορά του $\overline{D}$ είναι

$$\sigma^2_{\overline{D}} = \sigma^2_{\overline{Y}_{i\cdot}} + \sigma^2_{\overline{Y}_{j\cdot}} = \sigma^2\big[\frac{1}{n_i} + \frac{1}{n_j}\big]$$

και η εκτιμήτρια της είναι

$$S^2_{\overline{D}} = MSE\big[\frac{1}{n_i} + \frac{1}{n_j}\big].$$

Άρα $\frac{\overline{D}-(\mu_i-\mu_j)}{\sqrt{S^2_{\overline{D}}}} \sim t(n-m)$. Το διάστημα εμπιστοσύνης δίνεται ως

$$\overline{D} - \sqrt{S^2_{\overline{D}}}\, t_{\alpha/2}(n-m) \le \mu_i - \mu_j \le \overline{D} + \sqrt{S^2_{\overline{D}}}\, t_{\alpha/2}(n-m),$$

ενώ για τον έλεγχο

$$\begin{aligned} H_0 &: \mu_i - \mu_j = 0 \\ H_1 &: \mu_i - \mu_j \neq 0 \end{aligned}$$

απορρίπτουμε την H_0 αν $\left|\frac{\overline{D}}{\sqrt{S^2_{\overline{D}}}}\right| > t_{\alpha/2}(n-m)$.

Contrasts

Ορισμός: Contrast λέγεται μια γενική σύγκριση που εμπλέκει δυο ή περισσότερα επίπεδα του παράγοντα: $L = \sum_{i=1}^{m} c_i \mu_i$, όπου c_i είναι συντελεστές τέτοιοι ώστε $\sum_{i=1}^{m} c_i = 0$.

Παράδειγμα: Το contrast $L = \mu_1 - \mu_2$, όπου $c_1 = 1$, $c_2 = -1$ και $c_i = 0$ για κάθε $i = 3, \ldots, m$, αντιστοιχεί σε διαφορά δύο μέσων. Το contrast $L = \frac{\mu_1+\mu_2}{2} - \frac{\mu_3+\mu_4}{2}$, όπου $m = 4$ και $c_1 = c_2 = \frac{1}{2}$, $c_3 = c_4 = -\frac{1}{2}$ συγκρίνει μέσες τιμές ζευγών μέσων επιπέδων.

Μια αμερόληπτη εκτιμήτρια για το L είναι η $\hat{L} = \sum_{i=1}^{m} c_i \overline{Y}_{i\cdot}$. Αφού τα $\overline{Y}_{i\cdot}$ είναι ανεξάρτητα, η διασπορά του $\hat{L}$ είναι $\sigma^2_{\hat{L}} = \sigma^2 \sum_{i=1}^{m} \frac{c_i^2}{n_i}$, με εκτιμήτρια $S^2_{\hat{L}} = MSE \sum_{i=1}^{m} \frac{c_i^2}{n_i}$. Η εκτιμήτρια $\hat{L}$ ακολουθεί κανονική κατανομή ως γραμμικός συνδιασμός κανονικών τυχαίων μεταβλητών, επομένως

$$\frac{\hat{L} - L}{\sqrt{S^2_{\hat{L}}}} \sim t(n-m).$$

Το διάστημα εμπιστοσύνης δίνεται ως

$$\hat{L} - \sqrt{S^2_{\hat{L}}}\, t_{\alpha/2}(n-m) \leq \mu_i - \mu_j \leq \hat{L} + \sqrt{S^2_{\hat{L}}}\, t_{\alpha/2}(n-m).$$

Επίσης μπορούν να γίνουν οι έλεγχοι στατιστικής σημαντικότητας

$$\begin{aligned} H_0 &: \frac{\mu_1+\mu_2}{2} = \frac{\mu_3+\mu_4}{2} \quad (L = 0) \\ H_1 &: \frac{\mu_1+\mu_2}{2} \neq \frac{\mu_3+\mu_4}{2} \quad (L \neq 0) \end{aligned}$$

Απορρίπτουμε την H_0 αν $\left|\frac{\hat{L}}{\sqrt{S^2_{\hat{L}}}}\right| > t_{\alpha/2}(n-m)$.

Τέλος, μπορούν αντίστοιχα να γίνουν έλεγχοι της γενικής $H_0 : L = L_0$ έναντι αμφίπλευρης ή μονόπλευρης εναλλακτικής.

Παράδειγμα: Μια εταιρεία γαλακτοκομικών ενδιαφέρεται να ελέγξει 4 διαφορετικές συσκευασίες για ένα νέο σοκολατούχο γάλα. 10 καταστήματα με περίπου ίσους όγκους πωλήσεων επιλέχθηκαν ως πειραματικές μονάδες και σε κάθε κατάστημα δόθηκε ένα συγκεκριμένο είδος συσκευασίας. Συγκεκριμένα οι συσκευασίες 1 και 4 δόθηκαν σε δυο καταστήματα η καθεμία, ενώ οι συσκευασίες 2 και 3 δόθηκαν σε τρία καταστήματα η καθεμία. Άλλες παράμετροι που θα μπορούσαν να επηρεάσουν τις πωλήσεις (όπως τιμή, ποσότητα και θέση στο ράφι) διατηρήθηκαν σταθερές για όλα τα καταστήματα. Οι πωλήσεις σε απόλυτους αριθμούς καταγράφηκαν για μια χρονική περίοδο 3 ημερών σύμφωνα με τον ακόλουθο πίνακα.

Επίπεδο i (είδος συσκευασίας)	Παρατηρήσεις (ύψος πωλήσεων στα καταστήματα)	$Y_{i\cdot}$	$\overline{Y}_{i\cdot}$
1	12 18	30	15
2	14 12 13	39	13
3	19 17 21	57	19
4	24 30	54	27
m=4	n=10	$Y_{\cdot\cdot} = 180$	$\overline{Y}_{\cdot\cdot} = 18$

Analysis of Variance (ANOVA table)

Πηγή διασποράς	SS	df	MS	$F = \frac{MSF}{MSE}$
Ανάμεσα στα επίπεδα	$SSF = 258$	$m-1=3$	$\frac{258}{3} = 86 = \frac{SSF}{m-1}$	$\frac{86}{7.67} = 11.2$
Μέσα στα επίπεδα	$SSE = 46$	$n-m=6$	$\frac{46}{6} = 7.67 = \frac{SSE}{n-m}$	
Σύνολο	$SST = 304$	$n-1=9$		

Θεωρούμε τον έλεγχο

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ H_1 &: \mu_i \neq \mu_j \text{ για ένα τουλάχιστον ζεύγος } i, j \end{aligned}$$

Εφόσον $F = 11.2 > F_{0.05}(3,6) = 4.76$ απορρίπτουμε την H_0 σε επίπεδο στατιστικής σημαντικότητας $\alpha = 5\%$. Άρα οι μέσες πωλήσεις διαφέρουν στατιστικά σημαντικά για τα διάφορα είδη συσκευασίας. Για το πρώτο είδος συσκευασίας, δηλαδή για το μ_1, έχουμε:

$$\begin{aligned} & & \overline{Y}_{i\cdot} = 15, \quad n_1 = 2 \\ MSE = 7.67 \quad \Rightarrow \quad & S^2_{\overline{Y}_{i\cdot}} = \frac{7.67}{2} = 3.835 \end{aligned}$$

Το διάστημα εμπιστοσύνης συντελεστή 95% για το μ_1 είναι

$$\begin{aligned} \overline{Y}_{i\cdot} - \sqrt{MSE/n_i}\, t_{\alpha/2}(n-m) \leq &\quad \mu_1 \quad \leq \overline{Y}_{i\cdot} + \sqrt{MSE/n_i}\, t_{\alpha/2}(n-m) \\ 15 - \sqrt{3.835}\, 2.447 \leq &\quad \mu_1 \quad \leq 15 + \sqrt{3.835}\, 2.447 \\ 10.2 \leq &\quad \mu_1 \quad \leq 19.8 \end{aligned}$$

Ομοίως για τα μ_2, μ_3, μ_4.

Έστω ότι επιπλέον διαθέτουμε την πληροφορία ότι τα είδη συσκευασίας διατίθενται σε ποικίλους σχεδιασμούς (design), συγκεκριμένα ότι η συσκευασίες 1 και 2 έχουν design 3 χρωμάτων, ενώ οι συσκευασίες 3 και 4 έχουν design 4 χρωμάτων. Για να συγκρίνουμε τις μέσες πωλήσεις για το design των 3 χρωμάτων με τις μέσες πωλήσεις για το design των 4 χρωμάτων μπορούμε να χρησιμοποιήσουμε το contrast $L = \frac{\mu_1+\mu_2}{2} - \frac{\mu_3+\mu_4}{2}$. Έχουμε

$$\begin{aligned} \hat{L} &= \frac{\overline{Y}_{1\cdot} + \overline{Y}_{2\cdot}}{2} - \frac{\overline{Y}_{3\cdot} + \overline{Y}_{4\cdot}}{2} = \frac{15+13}{2} - \frac{19+27}{2} = -9 \\ \sum_{i=1}^{4} \frac{c_i^2}{n_i} &= \frac{1}{4}\big[\frac{1}{2} + \frac{1}{3} + \frac{1}{2} + \frac{1}{3}\big] = \frac{5}{12} = 0.4167 \\ S^2_{\hat{L}} &= MSE\big[\sum_{i=1}^{4} \frac{c_i^2}{n_i}\big] = 7.76 \cdot 0.4167 = 3.196 \end{aligned}$$

Άρα το διάστημα εμπιστοσύνης για το contrast δίνεται ως

$$\begin{aligned} \hat{L} - \sqrt{S^2_{\hat{L}}}\, t_{0.025}(6) \leq &\quad L \quad \leq \hat{L} + \sqrt{S^2_{\hat{L}}}\, t_{0.025}(6) \\ -9 - \sqrt{3.196}\, 2.447 \leq &\quad L \quad \leq -9 + \sqrt{3.196}\, 2.447 \\ -13.4 \leq &\quad L \quad \leq -4.6. \end{aligned}$$

Εφόσον το 0 δεν περιέχεται στο διάστημα εμπιστοσύνης η διαφορά των μέσων πωλήσεων για τα δύο είδη design είναι στατιστικά σημαντική σε επίπεδο σημαντικότητας $\alpha = 5\%$.

Η Μέθοδος της Ελάχιστης Σημαντικής Διαφοράς

The Least Significant Difference (LSD) method

Πολλαπλό έλεγχο υποθέσεων ονομάζουμε μια σειρά από ελέγχους υποθέσεων οι οποίοι διενεργούνται ταυτόχρονα.

Στα πλαίσια της ανάλυσης διασποράς, ο πιο διαδεδομένος πολλαπλός έλεγχος ισότητας ζευγών μέσων επιπέδων είναι το LSD test. Πρόκειται για ένα σύνολο ελέγχων της μορφής

$$\begin{aligned} H_0 &: \mu_i = \mu_j, \text{ για όλα τα } i \neq j \\ H_1 &: \mu_i \neq \mu_j. \end{aligned}$$

Οι έλεγχοι γίνονται με χρήση της στατιστικής συνάρτησης $T = \frac{\overline{Y}_{i\cdot} - \overline{Y}_{j\cdot}}{\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}} \sim t(n-m)$, κάτω από H_0.

Οι μέσοι μ_i, μ_j διαφέρουν στατιστικά σημαντικά σε επίπεδο σημαντικότητας α, αν $|\overline{Y}_{i\cdot} - \overline{Y}_{j\cdot}| > LSD = t_{\alpha/2}(n-m)\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$ [Ισοδύναμο με τα επιμέρους t–tests στατιστικής σημαντικότητας].

Μειονέκτημα : όσο μεγαλώνει ο αριθμός των t–test που κάνουμε τόσο αυξάνει η πιθανότητα σφάλματος τύπου I.

Άλλα κριτήρια : Duncan's, Tukey's, Newman-Keuls.

Μέθοδος Scheffé

Ο Scheffé πρότεινε μια μέθοδο για τη σύγκριση κάποιων ή όλων των δυνατών contrasts ανάμεσα σε μέσους επιπέδων στην οποία προκαθορίζεται η συνολική πιθανότητα σφάλματος τύπου I του πολλαπλού ελέγχου.

Έστω ένα σύνολο από contrasts $L_j = \sum_{i=1}^{m} c_{ij}\mu_i,\ j = 1, \ldots, r.$ Έχουμε $\hat{L}_j = \sum_{i=1}^{m} c_{ij}\hat{\mu}_i = \sum_{i=1}^{m} c_{ij}\hat{Y}_{i\cdot}$ και $S^2_{\hat{L}_j} = MSE \sum_{i=1}^{m} \frac{c_{ij}^2}{n_i}$. Ο Scheffé κατασκεύασε ταυτόχρονα διαστήματα εμπιστοσύνης της μορφής

$$\hat{L}_j - \sqrt{S^2_{\hat{L}_j}}\, C \leq L_j \leq \hat{L}_j + \sqrt{S^2_{\hat{L}_j}}\, C\,,$$

όπου $C = \sqrt{(m-1)F_\alpha(m-1, n-1)}$, τέτοια ώστε με πιθανότητα τουλάχιστον $1-\alpha$ όλα τα διαστήματα εμπιστοσύνης είναι αληθή.

Για έλεγχο υποθέσεων της μορφής

$$\begin{aligned} H_0 &: L_j = 0 \\ H_1 &: L_j \neq 0 \end{aligned}$$

η κριτική τιμή είναι C, δηλαδή, απορρίπτουμε την H_0 αν $\left|\frac{L_j}{\sqrt{S^2_{\hat{L}_j}}}\right| > C$. Η συνολική πιθανότητα σφάλματος τύπου I για τον πολλαπλό έλεγχο είναι το πολύ α.

- Τα απλά διαστήματα εμπιστοσύνης της κατανομής t είναι πιο στενά από τα αντίστοιχα διαστήματα εμπιστοσύνης της μεθόδου Scheffé οφείλεται στο ότι κατασκευάζουμε ταυτόχρονα διαστήματα εμπιστοσύνης για μια οικογένεια contrasts (μεγαλύτερη αβεβαιότητα).
- Η πιθανότητα σφάλματος τύπου I στη μέθοδο Scheffé είναι α αν πάρουμε όλα τα contrasts. Διαφορετικά είναι μικρότερη από α.