Εφαρμοσμένη Στατιστική

Δημήτριος Μπάγκαβος

Τμήμα Μαθηματικών και Εφαρμοσμένων Μαθηματικών
Πανεπιστήμιο Κρήτης

6 Μαΐου 2018

Ανάλυση Διακύμανσης.

- Η ανάλυση παλινδρόμησης μελετά τη στατιστική σχέση ανάμεσα στην εξαρτημένη μεταβλητή και σε μία η περισσότερες ανεξάρτητες μεταβλητές.
  - Πρώτα διαπιστώνεται η γραμμικότητα της σχέσης μεταξύ εξαρτημένης και ανεξάρτητων μεταβλητών
  - Μετά εκτιμώνται οι συντελεστές της σχέσης (παράμετροι των μεταβλητών) με σκοπό την ελαχιστοποίηση ενός στατιστικού κριτηρίου.
- Η ανάλυση διακύμανσης είναι πιο γενικό εργαλείο γιατί δεν υποθέτει κάποια συναρτησιακή μορφή μεταξύ εξαρτημένης και ανεξάρτητης τ.μ.
- Είναι σημαντικό να κάνουμε τον εξής διαχωρισμό: στο πλαίσιο της γραμμικής παλινδρόμησης και πάλι λαμβάνει χώρα ανάλυση διακύμανσης για την παραγωγή του F τεστ καλής προσαρμογής.
- Το F τεστ μετράει το κατά πόσο η μεταβλητότητα της εξαρτημένης μεταβλητής εξηγείται από το μοντέλο που υποθέσαμε.
- Η ανάλυση διακύμανσης που μελετάμε εδώ έχει άλλο σκοπό: μπορεί να περιλάβει και ποιοτικές μεταβλητές και αποσκοπεί στο να εξηγήσει κατά πόσο κάθε παράγοντας (δηλ. ποιοτική μεταβλητή) επηρεάζει την εξαρτημένη μεταβλητή.

Ανάλυση Διακύμανσης κατά ένα παράγοντα.

- Στην πιο απλή περίπτωση μας ενδιαφέρει να μελετήσουμε την ανάλυση της διακύμανσης ως προς τα επίπεδα ενός μόνο παράγοντα.
- Θεωρούμε ότι η εξαρτημένη μεταβλητή για το επίπεδο του i παράγοντα ακολουθεί κανονική κατανομή $N(\mu_i, \sigma^2), i = 1, \ldots, m$ όπου m είναι ο αριθμός των επιπέδων του υπό μελέτη παράγοντα.
- Ο σκοπός είναι να διαπιστώσουμε αν οι μέσες τιμές $\mu_i, i = 1, \ldots, m$ διαφέρουν ή αν όντως ο μέσος είναι σταθερός για όλα τα επίπεδα του παράγοντα.
- Η στατιστική μελέτη για την απάντηση του ερωτήματος βασίζεται σε ένα τυχαίο δείγμα τιμών της εξαρτημένης μεταβλητής από τα διάφορα επίπεδα του παράγοντα.
- Ξεκινάμε παίρνοντας δείγμα μεγέθους $n_i, i = 1, \ldots, m$ από το επίπεδο i του παράγοντα.
- Δεν είναι απαραίτητο τα n_i να είναι ίσα για κάθε επίπεδο.
- Πρέπει όμως οι παρατηρήσεις του δείγματος μέσα σε κάθε επίπεδο και μεταξύ των επιπέδων να είναι ανεξάρτητες.

Ανάλυση Διακύμανσης κατά ένα παράγοντα.

- Τα δεδομένα που συγκεντρώσαμε συνοψίζονται ως εξής:

Επίπεδα					Σύνολα	Μέσοι
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}	$Y_{2\cdot}$	$\bar{Y}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
m	Y_{m1}	Y_{m2}	$\dots$	Y_{mn_m}	$Y_{m\cdot}$	$\bar{Y}_{m\cdot}$
					$Y_{\cdot\cdot}$	$\bar{Y}_{\cdot\cdot}$

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

$$Y_{i\cdot} = \sum_{j=1}^{n_i} Y_{ij}, \ \bar{Y}_{i\cdot} = n_i^{-1} Y_{i\cdot}$$

είναι τα μερικά αθροίσματα και μέσοι όροι ανά γραμμή. Επίσης, το ολικό άθροισμα και γενικός μέσος είναι αντίστοιχα

$$Y_{\cdot\cdot} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} Y_{ij}, \ \bar{Y}_{\cdot\cdot} = n^{-1} Y_{\cdot\cdot}$$

Ανάλυση Διακύμανσης κατά ένα παράγοντα.

- Το μοντέλο ανάλυσης διασποράς κατά ένα παράγοντα είναι:

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

- Τα $\varepsilon_{ij}, j = 1, \ldots, n_i, i = 1, \ldots, m$ είναι ανεξάρτητα και ισόνομα κατανεμημένα τυχαία σφάλματα
- Οι μέσοι $\mu_i, i = 1, \ldots, m$ είναι οι άγνωστες παράμετροι του μοντέλου
- Δηλ. το μοντέλο υποθέτει ότι τα διάφορα επίπεδα του υπό μελέτη παράγοντα διαφέρουν ως προς το μέσο όρο τους αλλά έχουν κοινή διασπορά.
- Η υπόθεση που μας ενδιαφέρει να ελέγξουμε είναι

$H_0 : \mu_i = \mu, i = 1, \ldots, m$ έναντι $H_1 : \mu_i \neq \mu$ για τουλάχιστον ένα i

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

Ανάλυση Διακύμανσης κατά ένα παράγοντα.

- Η συνολική μεταβλητότητα του δείγματος ορίζεται από το άθροισμα των τετραγώνων των αποκλίσεων των παρατηρήσεων από το συνολικό μέσο του δείγματος:

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

- Παρατήρηση: Η απόκλιση της Y_{ij} από το συνολικό μέσο $\bar{Y}_{..}$ μπορεί να γραφτεί ως

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

- Οπότε μπορούμε να γράψουμε

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

Ανάλυση Διακύμανσης κατά ένα παράγοντα.

- Όμως

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

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

- Άρα

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

- Το άθροισμα των τετραγώνων των αποκλίσεων από τους μέσους των επιπέδων που ανήκουν

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

εκφράζει τη μεταβλητότητα μέσα στα επίπεδα του παράγοντα.

Ανάλυση Διακύμανσης κατά ένα παράγοντα.

- Το άθροισμα των τετραγώνων των αποκλίσεων από το συνολικό μέσο του δείγματος

$$\mathrm{SSF} = (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$$

εκφράζει τη μεταβλητότητα ανάμεσα στα επίπεδα. Συνολικά:

$$\mathrm{SST} = \mathrm{SSE} + \mathrm{SSF}$$

- Η μεταβλητότητα ανάμεσα στα επίπεδα είναι αυτή που οφείλεται στις διαφορές των επιπέδων του παράγοντα και εκφράζεται από το συνολικό άθροισμα τετραγώνων του παράγοντα SSF.
- Η μεταβλητότητα μέσα στα επίπεδα οφείλεται στην τυχαιότητα και εκφράζεται από το συνολικό άθροισμα τετραγώνων των τυχαίων σφαλμάτων.
- Τα m επίπεδα του υπό μελέτη παράγοντα κινούνται σε $m-1$ βαθμούς ελευθερίας
- Μέσα σε κάθε επίπεδο οι βαθμοί ελευθερίας είναι $n_i - 1$ άρα συνολικά στα επίπεδα έχουμε $n = m$ βαθμούς ελευθερίας.
- Οι συνολικοί βαθμοί ελευθερίας στο πείραμα είναι $n-1$.

Ανάλυση Διακύμανσης κατά ένα παράγοντα.

- Για κάθε επίπεδο ορίζουμε τη συνάρτηση

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

- Οι παρατηρήσεις του i επιπέδου αποτελούν δείγμα από την κανονική κατανομή $N(\mu_i, \sigma^2)$.
- Οπότε

$$\sigma^{-2} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2 = \sigma^{-2}(n_i - 1)W_i^2 \sim \chi^2_{n_i - 1}.$$

- Άρα

$$\mathbb{E} = \sigma^{-2}(n_i - 1)W_i^2 = n_i - 1 \Rightarrow \mathbb{E}W_i^2 = \sigma^2.$$

- Άρα η W_i^2 είναι αμερόληπτη εκτιμήτρια της διασποράς σ^2 βασισμένη σε δείγμα από το i επίπεδο.
- Τώρα, το άθροισμα m ανεξάρτητων τ.μ. που ακολουθούν χ^2 κατανομή είναι επίσης τ.μ. με χ^2 κατανομή με βαθμούς ελευθερίας ίσους με το άθροισμα των βαθμών ελευθερίας των αρχικών m κατανομών.

Ανάλυση Διακύμανσης κατά ένα παράγοντα.

- Άρα,

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

- Οπότε,

$$\mathbb{E}\sigma^{-2}\mathsf{SSE}=n-m\Rightarrow\mathbb{E}(n-m)^{-1}\mathsf{SSE}=\sigma^2.$$

- Επομένως η $(n-m)^{-1}\mathsf{SSE}$ είναι μία αμερόληπτη εκτιμήτρια της σ^2.
- Υπό την H_0 όλες οι παρατηρήσεις Y_{ij} είναι δείγμα από την ίδια κατανομή $N(\mu,\sigma^2)$.
- Έστω η συνάρτηση

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

- Έχουμε ότι

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

Ανάλυση Διακύμανσης κατά ένα παράγοντα.

- Άρα
$$\mathbb{E}\left\{\sigma^{-2}(n-1)S^2\right\} = n-1 \Rightarrow (n-1)^{-1}\mathrm{SST} = \sigma^2.$$
- Άρα η $S^2 = (n-1)^{-1}\mathrm{SST}$ υπό την H_0 είναι μια αμερόληπτη εκτιμήτρια της διασποράς σ^2.
- Ισχύει
$$\sigma^{-2}\mathrm{SST} = \sigma^{-2}\mathrm{SSE} + \sigma^{-2}\mathrm{SSF}.$$
- Από το θεώρημα των τετραγωνικών μορφών, υπό την H_0 η συνάρτηση $\mathrm{SSF}/\sigma^2 \sim \chi^2_{m-1}$. Άρα,
$$\mathbb{E}\sigma^{-2}\mathrm{SSF} = m-1 \Rightarrow (m-1)^{-1}\mathrm{SSF} = \sigma^2.$$
- Άρα η $(m-1)^{-1}\mathrm{SSF}$ υπό την H_0 είναι μια αμερόληπτη εκτιμήτρια της διασποράς σ^2.
- Άρα υπό την H_0 έχουμε ότι
$$\frac{\mathrm{SSF}/(m-1)}{\mathrm{SSE}/(n-m)} \sim F_{m-1,n-m}.$$

Ανάλυση Διακύμανσης κατά ένα παράγοντα.

- Η ελεγχοσυνάρτηση
$$\frac{\mathrm{SSF}/(m-1)}{\mathrm{SSE}/(n-m)}$$
μπορεί να χρησιμοποιηθεί για τον έλεγχο της ισότητας μέσων στην ανάλυση διακύμανσης.
- Η H_0 απορρίπτεται σε επίπεδο 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
$$\frac{\mathrm{SSF}/(m-1)}{\mathrm{SSE}/(n-m)} > F_{\alpha,m-1,n-m}.$$
- Παρατήρηση: Η εκτιμήτρια του σ^2 που βασίζεται στο SSE είναι πάντα αμερόληπτη. Όμως η εκτιμήτρια που βασίζεται στο SSF είναι αμερόληπτη κάτω από την H_0. Αν η H_0 δεν είναι αληθής τότε
$$\frac{1}{m-1}\mathbb{E}\{\mathrm{SSF}\} > \sigma^2.$$

Ανάλυση Διακύμανσης κατά ένα παράγοντα.

Απόδειξη:

$$\begin{aligned}
\mathbb{E}\{\mathrm{SSF}\} &= \mathbb{E}\sum_{i=1}^{m}(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot})^2 = \mathbb{E}\sum_{i=1}^{m} n_i\bar{Y}_{i\cdot}^2 - 2\sum_{i=1}^{m} n_i\bar{Y}_{i\cdot}\bar{Y}_{\cdot\cdot} + n\bar{Y}_{\cdot\cdot}^2 \\
&= \mathbb{E}\sum_{i=1}^{m} n_i\bar{Y}_{i\cdot}^2 - 2n\bar{Y}_{\cdot\cdot}^2 + n\bar{Y}_{\cdot\cdot}^2 = \mathbb{E}\sum_{i=1}^{m} n_i\bar{Y}_{i\cdot}^2 - n\bar{Y}_{\cdot\cdot}^2 \\
&= \sum_{i=1}^{m} n_i\mathbb{E}\bar{Y}_{i\cdot}^2 - n\mathbb{E}\bar{Y}_{\cdot\cdot}^2 \\
&= \sum_{i=1}^{m} n_i(\Delta(\bar{Y}_{i\cdot}) + \mathbb{E}^2(\bar{Y}_{i\cdot})) - n(\Delta(\bar{Y}_{\cdot\cdot}) + \mathbb{E}^2(\bar{Y}_{\cdot\cdot})) \\
&= \sum_{i=1}^{m} n_i\left(\frac{\sigma^2}{n_i} + \mu_i^2\right) - n\left(\frac{\sigma^2}{n} + \mu^2\right) \\
&= (m-1)\sigma^2 + \sum_{i=1}^{m} n_i(\mu_i^2 - \mu^2) > \sigma^2.
\end{aligned}$$

Υποθέσεις μοντέλου ανάλυσης διακύμανσης.

1. Η κατανομή των παρατηρήσεων σε κάθε επίπεδο είναι κανονική:

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

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

Στόχοι του μοντέλου ανάλυσης διακύμανσης.

1. Έλεγχος αν οι μέσοι των επιπέδων είναι ίσοι:

$$H_0 : \mu_i = \mu, i = 1, \ldots, m \text{ έναντι } H_1 : \mu_i \neq \mu \text{ για τουλάχιστον ένα } i$$

2. Αν οι μέσοι δεν είναι ίσοι, ποιες είναι οι διαφορές;

Αν οι μέσοι δεν διαφέρουν τότε το συμπέρασμα είναι οτι η εξαρτημένη μεταβλητή δεν εξαρτάται από τα επίπεδα του παράγοντα.

Έλεγχος για την ισότητα των διασπορών.

Το μοντέλο της ανάλυσης διασποράς που αναπτύξαμε υποθέτει ότι οι διακυμάνσεις των κανονικών κατανομών σε όλα τα επίπεδα είναι ίσες. Στην πράξη αυτό ελέγχεται:

Bartlett's test

Έστω S_i^2 η δειγματική διασπορά του i επιπέδου. Για τον έλεγχο της

$$H_0 : \sigma_1^2 = \cdots = \sigma_m^2 = \sigma^2 \text{ έναντι } H_1 : \sigma_i \neq \sigma^2 \text{ για τουλάχιστον ένα } i$$

χρησιμοποιείται η στατιστική συνάρτηση $\chi_0^2 = 2.3026qc^{-1}$ όπου

$$q = (n-m)\log S_p^2 - \sum_{i=1}^{m}(n_i-1)\log S_i^2$$

$$c = 1 + \frac{1}{3(m-1)}\left\{\sum_{i=1}^{m}(n_i-1)^{-1} - (n-m)^{-1}\right\}$$

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

Έλεγχος για την ισότητα των διασπορών.

- 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 απορρίπτουμε την H_0 για μεγάλες τιμές του χ_0^2 δηλαδή για $\chi_0^2 > \chi^2_{1-\alpha,m-1}$.
- Παρατήρηση: οι υποθέσεις του μοντέλου μπορούν να ελεγχθούν και γραφικά με τη βοήθεια των υπολοίπων.
- Τα υπόλοιπα ορίζονται ως

$$\varepsilon_{ij} = Y_{ij} - \hat{\mu}_i = Y_{ij} - \bar{Y}_{i\cdot}$$

- Με ένα pp plot πλοτ ελέγχουμε την κανονικότητα των υπολοίπων
- Με ένα scatter plot ελέγχουμε τυχαιότητα και ανεξαρτησία.

Επιμέρους έλεγχοι υποθέσεων για τους μέσους.

- Αν το F τεστ για την ισότητα των μέσων έδειξε ότι οι μέσοι των επιπέδων διαφέρουν, τότε έχει νόημα προχωρήσουμε σε περαιτέρω ανάλυση.
- Μια αμερόληπτη εκτιμήτρια του μέσου του i επιπέδου μ_i είναι ο δειγματικός μέσος του i επιπέδου $\hat{\mu}_i = \bar{Y}_{\cdot i}$, με διακύμανση $n_i^{-1}\sigma^2$
- Μία αμερόληπτη εκτιμήτρια του σ^2 είναι το MSE/n_i αφού το MSE είναι αμερόληπτη εκτιμήτρια του σ^2.
- Εφόσον $\bar{Y}_{\cdot i} \sim N(\mu_i, n_i^{-1}\sigma^2)$, αυτό συνεπάγεται ότι

$$(\sqrt{\mathsf{MSE}/n_i})^{-1}(\bar{Y}_{\cdot i} - \mu_i) \sim t_{n-m}$$

όπου $n - m$ είναι οι βαθμοί ελευθερίας που συνδέονται με το MSE.

- Έτσι για κάθε μ_i έχουμε το ακόλουθο διάστημα εμπιστοσύνης

$$\bar{Y}_{\cdot i} - \sqrt{\mathsf{MSE}/n_i}\, t_{1-\alpha/2,n-m} \le \mu_i \le \bar{Y}_{\cdot i} + \sqrt{\mathsf{MSE}/n_i}\, t_{1-\alpha/2,n-m}$$

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

Επιμέρους έλεγχοι υποθέσεων για τους μέσους.

- Για να εκτιμήσουμε τη διαφορά των μέσων 2 επιπέδων:
- Ορίζουμε $\bar{D} = \bar{Y}_{\cdot i} - \bar{Y}_{\cdot j}$.
- Η τ.μ. $\bar{D}$ ακολουθεί κανονική κατανομή ως γραμμικός συνδυασμός ανεξάρτητων κανονικών τ.μ.
- Εφόσον $\bar{Y}_{\cdot i}, \bar{Y}_{\cdot j}$ ανεξάρτητα, η διασπορά της $\bar{D}$ είναι

$$\sigma^2_{\bar{D}} = \sigma^2_{\bar{Y}_{\cdot i}} + \sigma^2_{\bar{Y}_{\cdot j}} = \sigma^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right)$$

και η εκτιμήτρια είναι

$$S^2_{\bar{D}} = \mathrm{MSE} \left(\frac{1}{n_i} + \frac{1}{n_j} \right).$$

- Άρα

$$\frac{\bar{D} - (\mu_i - \mu_j)}{\sqrt{S^2_{\bar{D}}}} \sim t_{n-m}$$

- Ένα δ/μα εμπιστοσύνης για τη διαφορά $\mu_i - \mu_j$ είναι

$$\bar{D} - \sqrt{S^2_{\bar{D}}} t_{1-\alpha/2, n-m} \leq \mu_i - \mu_j \leq \bar{D} + \sqrt{S^2_{\bar{D}}} t_{1-\alpha/2, n-m}.$$

Επιμέρους έλεγχοι υποθέσεων για τους μέσους.

- Για τον έλεγχο

$$H_0 : \mu_i - \mu_j = 0 \text{ έναντι } \mu_i - \mu_j \neq 0$$

απορρίπτουμε την H_0 αν $|\bar{D}(S^2_{\bar{D}})^{-1/2}| > t_{1-\alpha/2,n-m}$.

Ορισμός: Contrasts

Contrast λέγεται μία σύγκριση που εμπλέκει 2 η περισσότερα επίπεδα του παράγοντα

$$L = \sum_{i=1}^{m} c_i \mu_i$$

όπου c_i είναι συντελεστές τέτοιοι ώστε $c_1 + \cdots + c_m = 0$.

Παράδειγμα

Το Contrast $L = \mu_1 - \mu_2$ όπου $c_1 = 1, c_2 = -1$ και $c_i = 3, \ldots, m$ αντιστοιχεί σε διαφορά μέσων. Το Contrast $L = \frac{\mu_1+\mu_2}{2} - \frac{\mu_3+\mu_4}{2}$ με $m = 4, c_1 = c_2 = 1/2$ και $c_3 = c_4 = -1/2$ συγκρίνει μέσες τιμές ζευγών μέσων επιπέδων.

Επιμέρους έλεγχοι υποθέσεων για τους μέσους.

- Μια αμερόληπτη εκτιμήτρια για το L είναι η

$$\hat{L} = \sum_{i=1}^{m} c_i \bar{Y}_{.i}.$$

- Αφού τα $\bar{Y}_{.i}$ είναι ανεξάρτητα, η διασπορά του $\hat{L}$ είναι $\sigma^2_{\hat{L}} = \sum_{i=1}^{m} c_i^2 n_i^{-1}$ με εκτιμήτρια

$$S^2_{\hat{L}} = \text{MSE} \sum_{i=1}^{m} c_i^2 n_i^{-1}$$

- Η $\hat{L}$ είναι γραμμικός συνδυασμός κανονικών τ.μ. και άρα ακολουθεί την κανονική κατανομή, επομένως,

$$(\hat{L} - L) S_{\hat{L}}^{-1} \sim t_{n-m}$$

- Το διάστημα εμπιστοσύνης είναι

$$\hat{L} - \sqrt{S^2_{\hat{L}}} t_{n-m,1-\alpha/2} \leq \mu_i - \mu_j \leq \hat{L} + \sqrt{S^2_{\hat{L}}}.t_{n-m,1-\alpha/2}$$

Η μέθοδος της ελάχιστης σημαντικής διαφοράς.

- Η μέθοδος της ελάχιστης σημαντικής διαφοράς είναι ο πιο δημοφιλής πολλαπλός έλεγχος.
- Πρόκειται για σύνολο ελέγχων της μορφής

$$H_0 : \mu_i = \mu_j \text{ για όλα τα } i \neq j \text{ έναντι } H_1 : \mu_i \neq \mu_j$$

- Οι έλεγχοι γίνονται με χρήση του στατιστικού

$$t = \frac{\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}}{\sqrt{\text{MSE}(n_i^{-1} + n_j^{-1})}} \sim t_{n-m}$$

- Οι μέσοι μ_i και μ_j διαφέρουν στατιστικά σημαντικά σε επίπεδο σημαντικότητας α αν

$$|\bar{Y}_{i\cdot} - \bar{Y}_{j\cdot}| > t_{n-m,\alpha/2} \sqrt{\text{MSE}(n_i^{-1} + n_j^{-1})}$$

- Δηλαδή στην ουσία πρόκειται για τεστ ισοδύναμο με το t test.
- Ένα μειονέκτημα βέβαια είναι ότι όσο μεγαλώνει ο αριθμός των τεστ τόσο αυξάνει η πιθανότητα σφάλματος τύπου I.

Η μέθοδος Scheffe.

- Ο Scheffe πρότεινε μια μέθοδο για τη σύγκριση κάποιων ή όλων των δυνατών contrasts ανάμεσα σε μέσους επιπέδων.
- Έστω ένα σύνολο από contrasts

$$L_j = \sum_{i=1}^{m} c_{ij}\mu_i, j = 1, \ldots, r.$$

- Ισχύει,

$$\hat{L}_j = \sum_{i=1}^{m} c_{ij}\hat{\mu}_i = \sum_{i=1}^{m} c_{ij}\hat{Y}_{\cdot i}, S^2_{\hat{L}_j} = \text{MSE} \sum_{i=1}^{m} c_{ij}^2 n_i^{-1}$$

- Ο Scheffe κατασκεύασε ταυτόχρονα διαστήματα εμπιστοσύνης της μορφής

$$\hat{L}_j - \sqrt{S^2_{\hat{L}_j}}C \leq L_j \leq \hat{L}_j + \sqrt{S^2_{\hat{L}_j}}C,\ C = \sqrt{(m-1)F_{m-1,n-1,1-\alpha}}$$

τέτοια ώστε με πιθανότητα τουλάχιστον $1-\alpha$ όλα τα δ/τα να είναι αληθή.

Η μέθοδος Scheffe.

- Για έλεγχο υποθέσεων της μορφής

$$H_0 : L_j = 0 \text{ έναντι } H_1 : L_j \neq 0$$

απορρίπτουμε την H_0 αν

$$\left| \frac{L_j}{\sqrt{S^2_{L_j}}} \right| > C.$$

- Η συνολική πιθανότητα σφάλματος τύπου Ι για τους πολλαπλούς ελέγχους είναι το πολύ α.
  - Τα απλά δ/τα εμπιστοσύνης της κατανομής t είναι πιο στενά από τα αντίστοιχα δ/τα της μεθόδου Scheffe.
  - Αυτό οφείλεται στο ότι κατασκευάζουμε ταυτόχρονα δ/τα για μια οικογένεια contrasts οπότε υπάρχει μεγαλύτερη αβεβαιότητα.
  - Η πιθανότητα σφάλματος τύπου Ι στη μέθοδο Scheffe είναι α αν πάρουμε όλα τα contrasts
  - Στην αντίθετη περίπτωση είναι μικρότερη του α.