Παλινδρόμηση

Αναπλήρωση

9/12/16 ①

Ανάλυση Διακύμανσης κατά ένα παράγοντα

Y ~ Παράγοντα: Ανεξαρτ. ποιοτική μεταβλητή → $1, 2, \dots, I$ → κύρια επίδραση του i-επιπέδου στην Y

Εξαρτη. ποσοτική μεταβλητή

Μοντέλο: $Y = \mu_i + \varepsilon_{ij} \Leftrightarrow Y_{ij} = \mu + \underbrace{a_i}_{\text{ατομική επίδραση επιπέδων}} + \varepsilon_{ij}$ $\quad i = 1, \dots, I \quad j = 1, \dots, I_i$

Στο μοντέλο αυτό βρίσκαμε Ε.Ε.Τ: $\left.\begin{aligned} \hat{\mu} &= \bar{Y}_{oo} \\ \hat{a}_i &= \bar{Y}_{io} - \bar{Y}_{oo} \end{aligned}\right\}$ υπό την πλευρική συνθήκη $\sum_{i=1}^{I} a_i I_i = 0$

Στη συνέχεια αναλύσαμε στα $SS_{tot} = SS_{tr} + SS_{reg}$

→ ΠΙΝΑΚΑΣ ANOVA

Αυτό το μοντέλο θέλει να μελετήσει ποιο από τα επίπεδα του παράγοντα ασκούν την μεγαλύτερη επιρροή στο Y

Υποθέσεις για τα σφάλματα:

1. $E(\varepsilon_{ij}) = 0$
2. $Var(\varepsilon_{ij}) = \sigma^2$
3. $Cov(\varepsilon_{ij}, \varepsilon_{kl}) = 0$ } είναι μεταξύ τους ανεξάρτητα
4. $\varepsilon_{ij} \sim N(0, \sigma^2)$

Οπότε σύμφωνα με το

μοντέλο αφού $E(\varepsilon_{ij})=0$ άρα
$Var(\varepsilon_{ij})\,\sigma^2$

1. $E(Y_{ij})=\mu+\alpha_i$

2. $Var(Y_{ij})=\sigma^2$

3. $Cov(Y_{ij}, Y_{kl})=0$ } τα Y_{ij} ανεξάρτητα

4. $Y_{ij} \sim N(\mu+\alpha_i, \sigma^2)$

<u>Θεώρημα</u>: Υπό τις υποθέσεις για τα σφάλματα

1. $E(MS_{res})=\sigma^2$ (όπως και στα προηγούμενα μοντέλα)

2. $E(MS_{tr})=\sigma^2+\frac{1}{I-1}\sum_{i=1}^{I} J_i a_i^2$

Απόδειξη:

$MS_{res}=\frac{SS_{res}}{n-I}$, όπου $SS_{res}=\sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij}-\bar{Y}_{i.})^2$

Λόγω των υποθέσεων για τα σφάλματα

κάθε γραμμή δεδομένων προέρχεται από πληθυσμό

με διακύμανση σ^2. Δηλαδή $Y_{i1}, Y_{i2}, \ldots, Y_{iJ_i}$

είναι τυχαίο δείγμα από πληθυσμό με διακ. σ^2

Γνωρίζουμε ότι αν $\omega_1,\ldots,\omega_n$ τ.δ. από πληθυσ. με

διακύμανση σ^2 τότε $E(S^2_\omega)=\sigma^2$

όπου S^2_ω δειγματική διακύμανση και $S^2_\omega=\frac{1}{n-1}\sum_{i=1}^{n}(\omega_i-\bar{\omega})^2$

Θεώρημα

$$E(MS_{tr}) = \sigma^2 + \frac{1}{I-1}\sum_{i=1}^{I} J_i \alpha_i^2 .$$

Αποδ. Είναι $MS_{tr} = \dfrac{SS_{tr}}{I-1} = \dfrac{1}{I-1}\sum_{i=1}^{I} J_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$.

Από το μοντέλο $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ και την γνωστή συνθήκη $\sum_{i=1}^{I} J_i \alpha_i = 0$

παίρνουμε: $Y_{i\cdot} = J_i\mu + J_i\alpha_i + \varepsilon_{i\cdot}$ και $Y_{\cdot\cdot} = N\mu + \varepsilon_{\cdot\cdot}$

Έτσι $E(Y_{i\cdot}) = J_i(\mu + \alpha_i)$, $Var(Y_{i\cdot}) = Var(\varepsilon_{i\cdot}) = J_i\sigma^2$

και $E(Y_{\cdot\cdot}) = N\mu$, $Var(Y_{\cdot\cdot}) = N\sigma^2$.

Επομένως

$$E(SS_{tr}) = E\left\{\sum_{i=1}^{I} J_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2\right\} = E\left\{\sum_{i=1}^{I} J_i\left(\frac{Y_{i\cdot}}{J_i} - \frac{Y_{\cdot\cdot}}{N}\right)^2\right\} = E\left\{\sum_{i=1}^{I}\frac{Y_{i\cdot}^2}{J_i} - \frac{Y_{\cdot\cdot}^2}{N}\right\}$$

$$= \sum_{i=1}^{I}\frac{1}{J_i}E(Y_{i\cdot}^2) - \frac{1}{N}E(Y_{\cdot\cdot}^2) = \sum_{i=1}^{I}\frac{1}{J_i}\left\{Var(Y_{i\cdot}) + [E(Y_{i\cdot})]^2\right\} - \frac{1}{N}\left\{Var(Y_{\cdot\cdot}) + [E(Y_{\cdot\cdot})]^2\right\}$$

$$= \sum_{i=1}^{I}\frac{1}{J_i}\left\{J_i\sigma^2 + J_i^2(\mu+\alpha_i)^2\right\} - \frac{1}{N}\left\{N\sigma^2 + N^2\mu^2\right\} = (I-1)\sigma^2 + \sum_{i=1}^{I} J_i\alpha_i^2 .$$

Συνεπώς

$$E(MS_{tr}) = E\left(\frac{SS_{tr}}{I-1}\right) = \sigma^2 + \frac{1}{I-1}\sum_{i=1}^{I} J_i\alpha_i^2$$

Άρα

$$E\left(\frac{1}{J_i-1}\sum_{j=1}^{J_i}(Y_{ij}-\bar{Y}_i)^2\right)=\sigma^2$$

$$\stackrel{\cdot}{=} \quad E\left(\sum_{j=1}^{J_i}(Y_{ij}-\bar{Y}_i)^2\right)=(J_i-1)\sigma^2$$

δηλαδή

$$E(MS_{res})=E\left(\frac{1}{N-I}\sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij}-\bar{Y}_i)^2\right)$$

$$=\frac{1}{N-I}\sum_{i=1}^{I}E\left(\sum_{j=1}^{J_i}(Y_{ij}-\bar{Y}_{i\cdot})^2\right)$$

$$\stackrel{(*)}{=}\frac{1}{N-I}\sum_{i=1}^{I}(J_i-1)\sigma^2$$

$$=\frac{\sigma^2}{N-I}\left(\sum_{i=1}^{I}J_i-I\right)$$

$$=\frac{\sigma^2}{N-I}(N-I)$$

$$=\sigma^2$$

ΕΛΕΓΧΟΣ $\quad H_0: a_1=a_2= \quad =a_I$

Για τεχνικούς λόγους θεωρώ την ισοδύναμη υπόθεση θεωρώ

$H_0: a_1=a_2= \quad =a_I=0$

επομένως αν H_0: αληθής

$E(MS_{tr})=E(MS_{res})$ ή $MS_{res}\simeq MS_{tr}$

όποτε αν δεν είναι περίπου ίσα (διαφέρε

τότε απορρ. την H_0 αφού

αν $A\Rightarrow B$ το $\neg B\Rightarrow\neg A$ (προτασιακός λογισμός)

Άρα το test θα στηριχθεί στην σύγκριση των

MS_{res} και MS_{tr}

Προτιμώ $\frac{MS_{tr}}{M_{res}}$ για την σύγκριση και όχι την διαφορά διότι ειπάξω

ότι μπορεί να έχει και F-κατανομή

Θεώρημα

Υπό τις υποθέσεις για τα σφάλματα και την μηδενική υπόθεση $H_0: a_1 = a_2 = \dots = a_I = 0$ ισχύουν:

1. $\frac{SS_{res}}{\sigma^2} \sim \chi^2_{n-I}$

2. $\frac{SS_{tr}}{\sigma^2} \sim \chi^2_{I-1}$

3. Τα SS_{res}, SS_{tr} είναι μεταξύ τους ανεξάρτητα

Απόδειξη: α. Ξέρω $SS_{res} = \sum_{i=1}^{I} \sum_{j=1}^{J_i} (Y_{ij} - \bar{Y}_{i.})^2$

Από την (Η) υπόθεση για τα σφάλματα
τα $Y_{i1}, Y_{i2}, \dots, Y_{iJ_i}$ είναι τ.δ. από $N(\mu + a_i, \sigma^2)$
υπό H_0: $Y_{i1}, Y_{i2}, \dots$ −"−, −"−, $N(\mu, \sigma^2)$

Γνωρίζω επίσης αν $W_1, \dots, W_n$ τ.δ. από $N(\mu, \sigma^2)$ τότε

$$\frac{(n-1)S_W^2}{\sigma^2} \sim \chi^2_{n-1}, \quad S_W^2 = \frac{1}{n-1}\sum_{i=1}^{n}(W_i - \bar{W})^2$$

Και άρα $\frac{\sum_{j=1}^{J_i}(Y_{ij} - \bar{Y}_{i.})^2}{\sigma^2} \sim \chi^2_{J_i - 1} \Rightarrow \frac{\sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \bar{Y}_{i.})^2}{\sigma^2} \sim \chi^2_{\sum_{i=1}^{I}(J_i - 1)}$

$$\Rightarrow \frac{SS_{res}}{\sigma^2} \sim \chi^2_{n-I}$$

③

B.

$$SS_{tr} = \sum_{i=1}^{I} J_i (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$$

Γνωρίζω ότι $w_1, \dots, w_n$ τ.δ από $N(\mu, \sigma^2)$ τότε $\bar{w} \sim N(\mu, \frac{\sigma^2}{n})$

Τότε $\bar{Y}_{i\cdot} \sim N(\mu, \frac{\sigma^2}{I})$

Αρκεί να δημιουργήσω και τον δεύτερο δειγματικό μέσο οπότε έχω

$$\frac{\bar{Y}_{i\cdot} - \mu}{\sigma / \sqrt{J_i}} \sim N(0,1) \Rightarrow \frac{J_i(\bar{Y}_{i\cdot} - \mu)^2}{\sigma^2} \sim \chi^2_1$$

και λόγω ανεξαρτησίας

$$\sum_{i=1}^{I} \frac{J_i(\bar{Y}_{i\cdot} - \mu)^2}{\sigma^2} \sim \chi^2_{\sum_{i=1}^{I} 1} \equiv \chi^2_I$$

$$\Rightarrow \sum_{i=1}^{I} \frac{J_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2}{\sigma^2} \sim \chi^2_{I-1}$$

$$\Rightarrow \frac{SS_{tr}}{\sigma^2} \sim \chi^2_{I-1}$$

γ. Δεν το αποδυυ.

Οπότε τελικά έχω την Σ.Σ.Τ $F = \frac{MS_{tr}}{MS_{res}}$ όπου γνωρίζω ότι

$$F = \frac{SS_{tr}/(I-1)}{SS_{res}/(N-I)} = \frac{SS_{tr}/\sigma^2(I-1)}{SS_{res}/\sigma^2(N-I)} \sim \frac{\chi^2_{I-1}/I-1}{\chi^2_{N-I}/N-I} \underset{\text{ανεξ}}{\sim} F_{I-1,N-I},$$

υπό την μηδενική υπόθεση !

Βρήκα λοιπόν την κατανομή της Σ.Σ.Τ οπότε το τελευταίο που έχω να κάνω είναι να βρω κ.π

Κρίσιμη Περιοχή

Μορφή κρίσιμης περιοχής : $F \geq C$

Προσδιορισμός C: $\alpha = P(\text{απορ } H_0 / H_0 \text{ αληθ})$

$= (F \geq C \,/\, F \sim F_{I-1,N-I})$

$= P(F_{I-1,N-I} \geq C) \Rightarrow$

$C = F_{I-1,N-I,\alpha}$ α-ποσοστιαίο σημείο

Συμπερασματικά: Για τον έλεγχο της $H_0: \alpha_1 = \ldots = \alpha_I$)

(Για τον έλεγχο της ισοεπίδρασης των επιπέδων του παράγοντα στην ανεξ. μετ. Y)

η σ.σ.τ είναι το F-πηλίκο

$$\boxed{F = \frac{MS_{tr}}{MS_{res}} \sim F_{I-1,N-I} \text{ υπό } H_0!!}$$

και κ.π. μεγέθους α,

$$\boxed{F \geq F_{I-1,N-I,\alpha}}$$

Συνήθως η $H_0: \alpha_1 = \ldots = \alpha_I$ απορρίπτεται πως αυτό σημαίνει ότι κάποιο ή κάποια από τα επίπεδα του παράγοντα ασκούν σημαντικότερη επίδραση στη Y

Ερώτημα: Ποιο ή ποια επίπεδα ασκούν σημαντικότερη επίδραση στη Y

Απάντηση: Οι πολλ/λες συγκρίσεις
- ΜΕΘΟΔΟΣ ΕΛΑΧΙΣΤΗΣ ΣΗΜΑΝΤΙΚΗΣ ΔΙΑΦΟΡΑΣ (!)
- Tukey
- Bonferroni
- γραμμικοί συνδυασμοί (!)

Μέθοδος Ελάχιστης Σημαντικής Διαφοράς (ΕΣΔ)

Αν η $H_0: \alpha_1=\alpha_2=\dots=\alpha_I$ απορριφθεί τότε προχωρώ σε συγκρίσεις

ανά δύο Δηλαδή: ελέγχω την

$$H_0: \alpha_i=\alpha_{i'} \text{ έναντι } H_a: \alpha_i \neq \alpha_{i'} \quad i,i'=1,\dots,I$$

Οπότε πρέπει να κατασκευάσω ένα test για τον έλεγχο $H_0\ \alpha_i=\alpha_{i'}$

όπου και στηρίζομαι σε έναν εκτιμητή που εμφανίζεται στην H_0

(κατά Wald θεώρηση)

Ξέρω ότι $\hat{\alpha}_i=\overline{Y}_{i\cdot}-\overline{Y}_{\cdot\cdot}$

ενώ $\hat{\alpha}_{i'}=\overline{Y}_{i'\cdot}-\overline{Y}_{\cdot\cdot}$

$$\left.\begin{array}{l}\hat{\alpha}_i=\overline{Y}_{i\cdot}-\overline{Y}_{\cdot\cdot}\\ \hat{\alpha}_{i'}=\overline{Y}_{i'\cdot}-\overline{Y}_{\cdot\cdot}\end{array}\right\} \hat{\alpha}_i-\hat{\alpha}_{i'}=\overline{Y}_{i\cdot}-\overline{Y}_{i'\cdot}$$

Θα στηριχτώ στην διαφορά $\overline{Y}_{i\cdot}-\overline{Y}_{i'\cdot}$ και

αρκεί να βρώ την κατανομή της υπό την μηδενική υπόθεση!!

Κατανομή

Αφού $Y_{i1},\dots,Y_{iJ_i}$ τ.δ από κανονική $\sim N(\mu+\alpha_i,\sigma^2)$

Το $\overline{Y}_{i\cdot}\sim N\left(\mu+\alpha_i,\frac{\sigma^2}{J_i}\right)$

Επίσης

$Y_{i'1},\dots,Y_{i'J_{i'}}$ τ.δ $\sim N(\mu+\alpha_{i'},\sigma^2)$

Το $\overline{Y}_{i'\cdot}\sim N\left(\mu+\alpha_{i'},\frac{\sigma^2}{J_{i'}}\right)$

ανεξάρτητα

$$\overline{Y}_{i\cdot}-\overline{Y}_{i'\cdot}\sim N\left(\alpha_i-\alpha_{i'},\frac{\sigma^2}{J_i}+\frac{\sigma^2}{J_{i'}}\right) \text{ από μηδενική υπόθεση}$$

$$= N\left(0,\frac{\sigma^2}{J_i}+\frac{\sigma^2}{J_{i'}}\right)$$

$$\Rightarrow \frac{\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}}{\sigma\sqrt{\frac{1}{J_i} + \frac{1}{J_{i'}}}} \sim N(0,1) \text{ υπό } H_0\ a_i = a_{i'}$$

Θυμίζουμε ότι $\frac{SSres}{\sigma^2} \sim \chi^2_{N-I}$ οπότε διαιρώντας

$$\frac{\dfrac{\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}}{\sigma\sqrt{1/J_i + 1/J_{i'}}}}{\sqrt{\dfrac{SSres}{\sigma^2}/(N-I)}} = \frac{\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}}{\sqrt{MSres}\sqrt{1/J_i + 1/J_{i'}}} \sim t_{N-I} \text{ υπό την } H_0$$

Για τον έλεγχο της H_0 $a_i = a_{i'}$

η στατιστική συνάρτηση του Test είναι η: $\frac{\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}}{\sqrt{MSres}\sqrt{\frac{1}{J_i} + \frac{1}{J_{i'}}}} \sim t_{N-I}$ υπό H_0

και κρίσιμη περιοχή μεγάλες τιμές του πηλίκου $|t| > c$

δηλ $\left| \frac{\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}}{\sqrt{MSres}\sqrt{\frac{1}{J_i} + \frac{1}{J_{i'}}}} \right| \geq c = \ldots = \ldots t_{N-I,\alpha/2}$

οπότε $|\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}| \geq t_{N-I,\alpha/2} \sqrt{MSres}\sqrt{\frac{1}{J_i} + \frac{1}{J_{i'}}}$

<u>Πολλαπλές Συγκρίσεις</u>

$$E.S.\Delta = t_{N-I,\alpha/2} \sqrt{MSres\left(\frac{1}{J_i} + \frac{1}{J_{i'}}\right)}$$

Η μέθοδος εφαρμόζεται ως εξής <u>αν</u>: $|\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}| \geq ESD$ τότε H_0 απορρίπτεται

Αν η $H_0: a_i' = a_i$ απορρ. αυτό σημαίνει σε κάποιο από επίπεδα

i και i' ασκεί σημαντικότερη επίδραση στη Y

αν $(\bar{Y}_{i.} - \bar{Y}_{i'.}) > 0$ το i-επίπεδο ασκεί σημαντικότερη επίδραση από το i'

αλλιώς συμβαίνει το αντίθετο. (κρίνω με βάση το πρόσημο)

Παράδειγμα

Μια εταιρεία κατασκευάζει ενδυματικά όργανα για αεροπλάνα με την ιδιότητα τα όργανα αυτά να φωσφορίζουν για κάποιο χρονικό διάστημα μετά το σβήσιμο της λάμπας που τα φωτίζει. Η εταιρεία για να πετύχει την ιδιότητα αυτή μπορεί να διαλέξει μεταξύ 3-μεθόδων παραγωγής.
Τα παρακάτω δεδομένα δίνουν τον χρόνο σε δευτερόλ. που φωσφορίζουν τα εν λόγω όργανα για κάθε μία από τις 3-μεθόδους.
Να αναλυθούν τα δεδομένα αυτά

ΜΕΘΟΔΟΣ 1 : 59.2, 62.1, 57.4, 50, 59.3, 61.2, 60.8, 53.1

62.5
ΜΕΘΟΔΟΣ 2 58.4, 55, 59.8, 64.7, 59.9, 54.7, 58.4

ΜΕΘΟΔΟΣ 3 71.3, 66.6, 63.4, 64.7, 75.8, 65.6, 72.9, 67.3

Λύση

Έστω Y ο χρόνος που φωσφορίζουν

και 3-μεθόδους παραγωγής που πιθανώς επηρεάζουν το χρόνο Y

$Y \rightarrow$ μέθοδος παραγωγής (παράγοντας) < μεθ.1, μεθ.2, μεθ.3

Μοντέλο : $Y_{ij} = \mu + a_i + \varepsilon_{ij}$, $i=1,2,3$

$j = 1, \dots, J_i$ (στις παρατ. που έχω σε καθεμιά)

$J_1 = J_2 = J_3 = 8$

Υποθέτω λοιπόν θα ισχύουν οι υποθέσεις για τα σφάλματα

a_i = παριστά την κύρια επίδραση της i-μεθόδου στο χρόνο Y

ΑΝΟΒΑ

Μεταβλητότητα	Α.Τ Μεταβλητότητα	Β.Ε	Μ-Τ	F-ανάλυση
Μοντέλο	$SS_{tr} = 584,41$	$I-1=2$	$MS_{tr} = 292,205$	$F = 17,044$
Υπόλοιπα	$SS_{res} = 360,015$	$N-I=21$	$MS_{res} = 17,144$	
Ολική	$SS_{tot} = 944,425$	$N-1=23$		

Υπόθεση $H_0: a_1 = a_2 = a_3$ (Υπόθεση ισοεπίδρασης επιπέδων, δηλαδή Υπόθεση της ισοδυναμίας των τριών μεθόδων παραγωγής)

Επειδή $F = 17,044 > F_{I-1,N-I,a} = F_{2,21}$ $\begin{cases} a=0,01: & = 4,32 \\ a=0,05: & = 5,78 \end{cases}$

και επομένως απορρ. την μηδενική υπόθεση!!

Άρα τώρα εφαρμόζω ΕΣΔ (ΕΣΔ)

$\bar{Y}_{1.} = 57,1$, $\bar{Y}_{2.} = 59,175$ $\bar{Y}_{3.} = 68.45$

οπότε

$E.S.D = t_{N-I, a/2} \sqrt{MS_{res}\left(\frac{1}{J_i}+\frac{1}{J_{i'}}\right)} =$

$a=0,05$ $= 2,08 * \sqrt{17,144\left(2 \cdot \frac{1}{8}\right)} = 4,306$

Οπότε παίρνω τις διαφορές των μέσων

1. $|\bar{Y}_{1.} - \bar{Y}_{2.}| = 2,075$
2. $(\bar{Y}_{1.} - \bar{Y}_{3.}) = -11,35$
3. $(\bar{Y}_{2.} - \bar{Y}_{3.}) = -9,275$

Επειδή λοιπόν $|\bar{Y}_{1.}-\bar{Y}_{2.}| \not> 4,306$ άρα η $2=1$

$|\bar{Y}_{2.}-\bar{Y}_{3.}| = 9,275 > 4,306$ και άρα $1 \neq 3$ (απορρ. Ho
και καλύτερη είναι η **3** από το πρόσημο!

$|\bar{Y}_{1.}-\bar{Y}_{3.}| = 11,35 > 4,306$ και άρα $2 \neq 3$
και καλύτερη είναι η **3** από το πρόσημο!

Τελικά **$1=2<3$**