

# Κατάλοιπα

Σε ένα πίνακα συνάφειας όταν ελέγχουμε την ανεξαρτησία χρησιμοποιούμε τα κριτήρια είτε  $\chi^2$  είτε  $G^2$  με  $(I-1)(J-1)$  βαθμούς ελευθερίας.

Από τον ορισμό του  $\chi^2$  είναι εμφανές ότι μεγάλες αποκλίσεις των αναμενόμενων τιμών από τις παρατηρήσεις προκαλούν μεγαλύτερη τιμή του  $\chi^2$ . Η διαφορά

$$e_{ij} = n_{ij} - \hat{\mu}_{ij}$$

είναι ο ορισμός των καταλοίπων. Ορίζονται ως

$$r_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}}$$

# Κατάλοιπα

Τα κατάλοιπα αντιστοιχούν με τα απλά τυποποιημένα κατάλοιπα των γραμμικών μοντέλων.

Το άθροισμα των τετραγώνων τους είναι ίσο με την τιμή του κριτηρίου  $\chi^2$ .

Τα τυποποιημένα κατάλοιπα προέρχονται από τα κατάλοιπα αλλά λαμβάνουν υπόψη τους την πραγματική διακύμανση της διαφοράς  $e_{ij}$

Επομένως ορίζονται ως

$$d_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{u_{ij}}}, u_{ij} = \left(1 - \frac{n_{i.}}{n_{..}}\right) \left(1 - \frac{n_{.j}}{n_{..}}\right) \hat{\mu}_{ij}$$

# Κατάλοιπα

Κάτω από την μηδενική υπόθεση της ανεξαρτησίας τα  $r_{ij}$  και  $d_{ij}$  ακολουθούν την κανονική κατανομή με μέση τιμή 0.

Η ασυμπτωτική διακύμανση του  $r_{ij}$  είναι μεγαλύτερη του 1, ενώ αυτή του  $d_{ij}$  είναι ασυμπτωτικώς 1.

Τιμές μεγαλύτερες του 2 ή 3 δίνουν πιθανότατα υψηλές τιμές του  $\chi^2$ .

# Κατάλοιπα

Σε 3 κατηγορίες ψυχασθενών ελέγχθηκε αν μετά από μια θεραπεία υπήρξε ίδια βελτίωση της δραστηριότητας.

Τύπος ψυχασθένειας, $X$	Αποτέλεσμα θεραπείας στην καθημερινή δραστηριότητα, $Y$		Σύνολο
	Βελτίωση	Χωρίς βελτίωση	
Σχιζοφρενείς	13 (10)	17 (20)	30
Νευρωτικοί	5 (10)	25 (20)	30
Άλλοι	12 (10)	18 (20)	30
Σύνολο	30	60	90

Ποσοστά βελτίωσης

$$\frac{13}{30} = 0.43$$

$$\frac{5}{30} = 0.17$$

$$\frac{12}{30} = 0.40$$

Αναμενόμενες τιμές  $e_{11} = \frac{30 * 30}{90} = 10$

$$\chi^2 = 5.7$$

$$G^2 = 6.104$$

$$\chi^2_2$$

$$p = 0.058$$

$$p = 0.047$$

# Κατάλοιπα

Λόγο των οριακών τιμών των  $p$ -τιμών δεν μπορούμε να φτάσουμε σε ασφαλές συμπέρασμα σχετικά με την απόρριψη ή την αποδοχή των τιμών.

		Αποτέλεσμα θεραπείας στην καθημερινή δραστηριότητα, $Y$	
Τύπος ψυχασθένειας, $X$	Κατάλοιπα	Βελτίωση	Χωρίς βελτίωση
Σχιζοφρενείς	$e$	3.0	-3.0
	$r$	0.95	-0.67
	$d$	1.42	-1.42
Νευρωτικοί	$e$	-5.0	5.0
	$r$	-1.58	1.12
	$d$	-2.37	2.37
Νευρωτικοί	$e$	2.0	-2.0
	$r$	0.63	-0.45
	$d$	0.95	-0.95

Λίγο υψηλότερες του 2

# Διαμέριση του $\chi^2$

Οι στατιστικές συναρτήσεις  $X^2$  έχουν μια αναπαραγωγική ιδιότητα. Αν έχουμε ανεξάρτητες  $X^2$  με βαθμούς ελευθερίας  $df_1$  και  $df_2$  τότε

$$X_1^2 + X_2^2 \sim \chi_{df_1+df_2}^2$$

Επομένως μπορούν να διαμεριστούν οι τιμές μιας  $X^2$  με  $df_1 > 1$  σε τιμές  $X^2$  με μικρότερους βαθμούς ελευθερίας

Η λογική είναι στο να διαμοιράσει κάποιος τον αρχικό πίνακα σε μικρότερους υποπίνακες έτσι ώστε οι επιμέρους τιμές  $X^2$  να αναπαριστούν συγκεκριμένες μορφές της σχέσης μεταξύ των δύο κατηγορικών μεταβλητών και να αθροίζουν στο  $X^2$  του αρχικού πίνακα.

# Διαμέριση του $\chi^2$

- Κατασκευάζονται  $(I-1)(J-1)$   $2 \times 2$  πίνακες ως

$$\begin{array}{c|c} \sum_{a < i} \sum_{b < j} n_{ab} & \sum_{a < i} n_{aj} \\ \hline \sum n_{ib} & n_{ij} \end{array}$$

- Για κάθε υποπίνακα  $^{b < j}$  υπολογίζεται το  $G^2$ . Το άθροισμα των επιμέρους  $G^2$  ισούται με το  $G^2$  του αρχικού πίνακα.
- Οι βαθμοί ελευθερίας των υποπινάκων πρέπει να αθροίζουν στους βαθμούς ελευθερίας του αρχικού πίνακα
- Κάθε συχνότητα κελιού πρέπει να εμφανίζεται μόνο σε έναν υποπίνακα
- Οι περιθώριες συχνότητες του αρχικού πίνακα πρέπει να εμφανίζονται μόνο σε έναν υποπίνακα

# Διαμέριση του $\chi^2$

- Η στατιστική συνάρτηση  $G^2$  έχει ακριβείς διαμερίσεις. Σε αντίθεση με την στατιστική συνάρτηση  $\chi^2$  της οποίας οι διαμερίσεις δεν αθροίζουν στην τιμή  $\chi^2$  του αρχικού πίνακα.

$$G^2 = 6.104$$

	Αποτέλεσμα θεραπείας στην καθημερινή δραστηριότητα, Y		
Τύπος ψυχασθένειας, X	Βελτίωση	Χωρίς βελτίωση	Σύνολο
Σχιζοφρενείς	13	17	30
Νευρωτικοί	5	25	30
Άλλοι	12	18	30
Σύνολο	30	60	90

	Αποτέλεσμα θεραπείας στην καθημερινή δραστηριότητα, Y		
Τύπος ψυχασθένειας, X	Βελτίωση	Χωρίς βελτίωση	Σύνολο
Σχιζοφρενείς	13	17	30
Νευρωτικοί	5	25	30
Σύνολο	18	42	60

$$G^2 = 5.216 \Rightarrow p = 0.022$$

	Αποτέλεσμα θεραπείας στην καθημερινή δραστηριότητα, Y		
Τύπος ψυχασθένειας, X	Βελτίωση	Χωρίς βελτίωση	Σύνολο
Σχιζοφρενείς-Νευρωτικοί	18	42	60
Άλλοι	12	18	30
Σύνολο	30	60	90

$$G^2 = 0.888 \Rightarrow p = 0.346$$



# Διαμέριση του $\chi^2$

	Αποτέλεσμα θεραπείας στην καθημερινή δραστηριότητα, Y		
Τύπος ψυχασθένειας, X	Βελτίωση	Χωρίς βελτίωση	Σύνολο
Σχιζοφρενείς	13	17	30
Νευρωτικοί	5	25	30
Άλλοι	12	18	30
Σύνολο	30	60	90

Κάνουμε την πρώτη διαμέριση παίρνοντας το υποπίνακα.

Για την επόμενη διαμέριση παρατηρούμε ότι οι τελευταίες τιμές 12 και 18 είναι μόνες τους. Επομένως παίρνουμε το άθροισμα των δύο προηγούμενων γραμμών και τις τελευταίες

# Συμπληρωματικές ασκήσεις

## Άσκηση 1

Έστω ότι η διδιάστατη τυχαία μεταβλητή έχει την τριωνυμική κατανομή με συνάρτηση πιθανότητας την

$$f(x, y) = \frac{n!}{x! y! (n - x - y)!} p_1^x p_2^y p_0^{n-x-y}$$

όπου  $x, y = 0, 1, \dots, n$  και  $x + y \leq n$ .

Τότε η περιθώρια κατανομή της τ.μ.  $X$  είναι διωνυμική με σ.π.π. την

$$f(x, y) = \binom{n}{x} p_1^x (1 - p_1)^{n-x}$$

# Συμπληρωματικές ασκήσεις

## Άσκηση 1

και η περιθώρια κατανομή της τ.μ.Υ είναι διωνυμική  
με σ.π.π. την

$$f(x, y) = \binom{n}{y} p_2^y (1 - p_2)^{n-y}$$

# Συμπληρωματικές ασκήσεις

## Λύση

$$\begin{aligned} f_X &= \sum_{y=0}^{n-x} f(x, y) = \sum_{y=0}^{n-x} \frac{n!}{x! y! (n-x-y)!} p_1^x p_2^y (1-p_1-p_2)^{n-x-y} = \\ &= \frac{n!}{x! (n-x)!} p_1^x \sum_{y=0}^{n-x} \frac{(n-x)!}{y! (n-x-y)!} p_2^y (1-p_1-p_2)^{n-x-y} = \\ &= \binom{n}{x} p_1^x \sum_{y=0}^{n-x} \binom{n-x}{y} p_2^y (1-p_1-p_2)^{n-x-y} = \binom{n}{x} p_1^x (p_2 + 1 - p_1 - p_2)^{n-x} = \\ &= \binom{n}{x} p_1^x (1-p_1)^{n-x} \end{aligned}$$

# Συμπληρωματικές ασκήσεις

## Άσκηση 2

**Παράδειγμα. (Νόμος των Hardy-Weinberg):** Στον νόμο των Hardy-Weinberg (παράδειγμα κεφ. 4) είχαμε υπολογίσει ότι οι πιθανότητες (ποσοστά) με τις οποίες τα τρία είδη γονοτύπων AA, Aa και aa συναντώνται σε ένα πληθυσμό είναι

$$P(AA)=p^2, \quad P(Aa)=2p(1-p) \text{ και } P(aa)=(1-p)^2$$

όπου  $p$  είναι η πιθανότητα μεταφοράς του γονιδίου A στον απόγονο. Έστω ότι επιλέγουμε τυχαία οκτώ άτομα από ένα πληθυσμό θέλοντας να καθορίσουμε τα γονότυπά τους. Να υπολογισθούν, συναρτήσει του  $p$ , οι πιθανότητες ότι

- (α) Δεν υπάρχουν γονότυπα της μορφής AA στο δείγμα.
- (β) Υπάρχουν δύο AA, τέσσερα Aa και δύο aa.
- (γ) Ποιά είναι η τιμή του  $p$  που δίνει την μέγιστη τιμή στην πιθανότητα του ερωτήματος (β);

# Συμπληρωματικές ασκήσεις

## Λύση

**Λύση:** Έστω  $X_1$  ο αριθμός των ΑΑ,  $X_2$  των Αα και  $X_3$  των αα στο δείγμα. Τότε το τυχαίο διάνυσμα  $(X_1, X_2, X_3)$  ακολουθεί την πολυωνυμική κατανομή.

$$(\alpha) P(\text{μηδέν ΑΑ στο δείγμα}) = P(X_1=0) = \binom{8}{0} (p^2)^0 (1-p^2)^8 = (1-p^2)^8$$

$$\begin{aligned} (\beta) P(X_1=2, X_2=4, X_3=2) &= \frac{8!}{2!4!2!} (p^2)^2 (2p(1-p))^4 ((1-p)^2)^2 \\ &= 6720p^8(1-p)^8 \end{aligned}$$

$$(\gamma) \frac{\partial P(X_1 = 2, X_2 = 4, X_3 = 2)}{\partial p} = 0 \Leftrightarrow p = \frac{1}{2}.$$