ΥΠΟΥΡΓΕΙΟ ΕΘΝΙΚΗΣ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ ΕΠΕΑΕΚ
ΕΥΡΩΠΑΪΚΗ ΕΝΩΣΗ
ΣΥΓΧΡΗΜΑΤΟΔΟΤΗΣΗ
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Η ΠΑΙΔΕΙΑ ΣΤΗΝ ΚΟΡΥΦΗ
Επιχειρησιακό Πρόγραμμα
Εκπαίδευσης και Αρχικής
Επαγγελματικής Κατάρτισης

ΤΟ ΕΡΓΟ ΣΥΓΧΡΗΜΑΤΟΔΟΤΕΙΤΑΙ ΑΠΟ ΤΟ ΕΥΡΩΠΑΙΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ ΚΑΙ ΑΠΟ ΕΘΝΙΚΟΥΣ ΠΟΡΟΥΣ

Αναμόρφωση του Προγράμματος Προπτυχιακών Σπουδών του Τμήματος Μαθηματικών του Πανεπιστημίου Αθηνών με έμφαση στην Πληροφορική, τη Διδακτική και τις Εφαρμογές των Μαθηματικών.

Σημειώσεις για το μάθημα

Στατιστική ΙΙ

Λουκία Μελιγκοτσίδου

Τις σημειώσεις επιμελήθηκαν οι : Σπ. Κάντα
Στ. Καποδίστρια
Τ. Ξιφαρά

Περιεχόμενα

1 Εισαγωγή **7**

2 Γραμμική Παλινδρόμηση **9**

2.1 Ανάλυση Παλινδρόμησης 9

2.2 Εκτίμηση παραμέτρων 10

2.2.1 Εκτίμηση Συντελεστών με τη Μέθοδο Ελαχίστων Τετραγώνων 10

2.2.2 Εκτίμηση της Διασποράς 15

2.3 Γραμμικό Μοντέλο με σφάλματα που ακολουθούν Κανονική Κατανομή 15

2.3.1 Εκτίμηση παραμέτρων με τη Μέθοδο Μέγιστης Πιθανοφάνειας 15

2.4 Το Απλό Γραμμικό Μοντέλο με Πίνακες 20

2.5 Μετασχηματισμοί Συνάρτησης Παλινδρόμησης ώστε να γίνει Γραμμική 22

2.6 Πολλαπλή Παλινδρόμηση 23

2.7 Κατανομές τετραγωνικών μορφών 24

2.8 Κατάλοιπα Παλινδρόμησης 26

2.8.1 Ιδιότητες καταλοίπων 27

2.9 Παλινδρόμηση και Ανάλυση Διασποράς 30

2.9.1 Συντελεστής Προσδιορισμού 34

2.10 Πολυσυγγραμικότητα 34

3 Ανάλυση Διασποράς **39**

3.1 Η σχέση μεταξύ παλινδόμησης και ανάλυσης διακύμανσης 39

3.2 Πειραματικός Σχεδιασμός 39

3.3 Ανάλυση Διασποράς κατά έναν Παράγοντα 40

3.3.1 Έλεγχος Ισότητας Μέσων 41

3.3.2 Υποθέσεις του μοντέλου ANOVA 44

3.3.3 Έλεγχος για την ισότητα των διασπορών 45

3.3.4 Επιμέρους έλεγχοι υποθέσεων για τους μέσους 46

3.3.5 Contrasts 47

3.3.6 Παρατηρούμενο Επίπεδο Σημαντικότητας 49

3.3.7 Η Μέθοδος της Ελάχιστης Σημαντικής Διαφοράς 50

3.3.8 Μέθοδος Scheffé 50

3.4 Ανάλυση Διασποράς με Δυο Παράγοντες (two–factor ANOVA) 51

3.4.1 Έλεγχοι Υποθέσεων 52

3.4.2 ANOVA κατά δυο παράγοντες με αλληλεπίδραση 54

4 Απαραμετρική Στατιστική **59**

4.1 Κριτήριο X^2 59

4.2 Πίνακες Συνάφειας 64

4.2.1 Ο έλεγχος X^2 για την ύπαρξη διαφορών στις πιθανότητες εμφάνισης k κατηγοριών σε l ανεξάρτητα δείγματα 65

4.2.2 Ο έλεγχος ανεξαρτησίας X^2 65

4.2.3 Σύγκριση l ανεξάρτητων πολυωνυμικών κατανομών 68

4.3 Σύγκριση δύο άγνωστων κατανομών 70

4.4 Έλεγχοι Υποθέσεων βασισμένοι στη Διωνυμική Κατανομή 73

4.4.1 Ο Διωνυμικός Έλεγχος 73

4.4.2 Ο Προσημικός Έλεγχος 74

4.5 Έλεγχος Wilcoxon 76

4.5.1 Ο έλεγχος των προσημασμένων τάξεων μεγέθους του Wilcoxon για τη διάμεσο ενός πληθυσμού 76

4.5.2 Έλεγχος Wilcoxon για δείγμα ζευγών παρατηρήσεων 77

Κεφάλαιο 1

Εισαγωγή

Σε ένα πρώτο μάθημα στατιστικής μαθαίνουμε για περιγραφικά μέτρα, αριθμητικά και γραφικά, για συμπερασματολογία για κάποιο πληθυσμό με βάση ένα δείγμα και κάποια στοιχεία θεωρίας κατανομών. Το ουσιαστικό ενδιαφέρον της στατιστικής επιστήμης ωστόσο εστιάζει στην κατανόηση στοχαστικών φαινομένων και την ποσοτικοποίηση της αβεβαιότητας σχετικά με αυτά. Για το σκοπό αυτό αναπτύχθηκαν δύο σημαντικά εργαλεία: η στατιστική μοντελοποίηση και η στατιστική συμπερασματολογία.

Η στατιστική μοντελοποίηση αφορά στην κατασκευή μοντέλων/ υποδειγμάτων με λίγες παραμέτρους για την περιγραφή στοχαστικών φαινομένων-σχέσεων μεταξύ των μεταβλητών. Η στατιστική συμπερασματολογία αφορά στην εκτίμηση των άγνωστων παραμέτρων με βάση παρατηρήσεις/ δεδομένα και στην ποσοτικοποίηση της αβεβαιότητας σχετικά με τις εκτιμήσεις.

Σ' αυτό το μάθημα θα ασχοληθούμε με τα εξής θέματα της στατιστικής επιστήμης:

- Ανάλυση παλινδρόμησης: το γραμμικό μοντέλο
- Ανάλυση διασποράς
- Απαραμετρική συμπερασματολογία

Κεφάλαιο 2

Γραμμική Παλινδρόμηση

2.1 Ανάλυση Παλινδρόμησης

Έστω δύο μεταβλητές X και Y. Στα μαθηματικά έχουμε συναρτησιακή σχέση μεταξύ των μεταβλητών της μορφής

$$Y = f(X)$$

η οποία είναι μια ντετερμινιστική σχέση (*deterministic relationship*), που σημαίνει ότι η τιμή της X καθορίζει πλήρως την τιμή της Y.
π.χ. $Y = \beta_0 + \beta_1 X$, γραμμική σχέση.

Η στατιστική σχέση μεταξύ δυο μεταβλητών είναι της μορφής

$$Y = f(X) + \varepsilon,$$

όπου ε τυχαίος (στοχαστικός) όρος, δηλαδή έχουμε στοχαστική σχέση (stochastic relationship). Η τυχαία μεταβλητή Y εξαρτάται από την μεταβλητή X (η οποία έχει προκαθορισμένες τιμές), αλλά και από κάποιους μη μετρήσιμους παράγοντες που συνοψίζονται στον στοχαστικό όρο ε.
π.χ. $Y = \beta_0 + \beta_1 X + \varepsilon$, απλή γραμμική παλινδρόμηση (regression) ή απλό γραμμικό μοντέλο (simple linear model).

Σκοπός μας είναι να εκτιμήσουμε τις άγνωστες παραμέτρους β_0, β_1 χρησιμοποιώντας δείγμα $(Y_i, X_i), i = 1, \ldots, n$. Έχουμε $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$.

Γραφικά
Σημειόγραμμα(scatter plot)

Η Y είναι η εξαρτημένη μεταβλητή (dependent or response) και η X είναι η ανεξάρτητη μεταβλητή (independent or predictor). Τα ε_i είναι τυχαία σφάλματα.
Προσοχή!: Η Y είναι τ.μ. ενώ η X όχι.

Υποθέσεις για τα τυχαία σφάλματα:

- $E(\varepsilon_i) = 0$ Μηδενική μέση τιμή

- $V(\varepsilon_i) = \sigma^2$ Ομοσκεδαστικότητα (ίση διασπορά)
- $Cov(\varepsilon_i, \varepsilon_j) = 0$ Ασυσχέτιστα σφάλματα (το σφάλμα σε οποιαδήποτε δοκιμή δεν επηρεάζει το σφάλμα άλλων δοκιμών)

Επομένως έχουμε
$E(Y_i) = \beta_0 + \beta_1 X_i$, $V(Y_i) = \sigma^2$, $Cov(Y_i, Y_j) = 0$.

Η γραμμή παλινδρόμησης δίνει την αναμενόμενη τιμή της Y για κάθε τιμή της X.

Απλό γραμμικό μοντέλο

Απλό γιατί υπάρχει μια μόνο ανεξάρτητη μεταβλητή.

Γραμμικό ως προς τις παραμέτρους γιατί καμία παράμετρος δεν παρουσιάζεται σαν εκθέτης ή είναι πολυωνυμική ή διαιρεμένη με άλλη παράμετρο.

Το υπόδειγμα $Y_i = \beta_0 + \beta_1 X_i^2 + \varepsilon_i$ είναι γραμμικό
ενώ το $Y_i = \beta_1^{\beta_1} X_i + \varepsilon_i$ όχι.

Ερμηνεία των Παραμέτρων της Παλινδρόμησης

β_0**:** είναι το σημείο όπου η ευθεία τέμνει τον άξονα των Y, δηλαδή αντιστοιχεί στην αναμενόμενη τιμή του Y για $X = 0$

β_1**:** είναι η κλίση της ευθείας και αντιπροσωπεύει την μεταβολή (αύξηση ή μείωση) στην αναμενόμενη τιμή του Y που αντιστοιχεί σε μεταβολή του X κατά μια μονάδα.

2.2 Εκτίμηση παραμέτρων

2.2.1 Εκτίμηση Συντελεστών με τη Μέθοδο Ελαχίστων Τετραγώνων

Η μέθοδος ελαχίστων τετραγώνων στοχεύει στον προσδιορισμό της γραμμής παλινδρόμησης έτσι ώστε οι αποστάσεις των σημείων από την ευθεία να είναι ελάχιστες (ελαχιστοποίηση των σφαλμάτων).

Έχουμε $\varepsilon_i = Y_i - E(Y_i) = Y_i - (\beta_0 + \beta_1 X_i)$.
Επειδή $E(\varepsilon_i) = 0$ δεν εξετάζουμε την ποσότητα $\sum_{i=1}^{n} \varepsilon_i$ (θα είναι 0), αλλά παίρνουμε το άθροισμα των τετραγώνων

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Οι εκτιμήτριες των β_0, β_1 προκύπτουν από την ελαχιστοποίηση του Q.

$$\begin{cases} \dfrac{dQ}{d\beta_0} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i) = 0 \\ \dfrac{dQ}{d\beta_1} = -2\sum_{i=1}^{n} X_i(Y_i - \beta_0 - \beta_1 X_i) = 0 \end{cases} \Rightarrow \begin{cases} \sum_{i=1}^{n} Y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i Y_i = \beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2 \end{cases}$$ Κανονικές Εξισώσεις

Λύνοντας ως προς β_0 και β_1 έχουμε

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \frac{1}{n}\left[\sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i\right] = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Εναλλακτική μορφή του απλού γραμμικού μοντέλου

$$Y_i = \beta_0^* + \beta_1(X_i - \bar{X}) + \varepsilon_i, \text{ όπου } \beta_0^* = \beta_0 + \beta_1 \bar{X}$$
$$\text{ή } Y_i = \beta_0^* + \beta_1 \widetilde{X}_i + \varepsilon_i \text{ όπου } \widetilde{X}_i = X_i - \bar{X}$$

Η εκτιμήτρια του β_1 είναι η ίδια.
Για το β_0^* είναι: $\hat{\beta}_0 = \hat{\beta}_0 + \hat{\beta}_1 \bar{X} = \bar{Y} - \hat{\beta}_1 \bar{X} + \hat{\beta}_1 \bar{X} = \bar{Y}$

Παράδειγμα 2.1 *Ένα εργοστάσιο ξυλείας κατασκευάζει θρανία μια φορά το μήνα συγκεντρώνοντας τις παραγγελίες που του έγιναν. Τους τελευταίους 10 μήνες οι αριθμοί των θρανίων που κατασκεύασε και οι αντίστοιχες εργατοώρες δίνονται παρακάτω.*

Μήνας	Αριθμός θρανίων X	Εργατοώρες Y
1	30	73
2	20	50
3	60	128
4	80	170
5	40	87
6	50	108
7	60	135
8	30	69
9	70	148
10	60	132

Να εκτιμηθούν οι συντελεστές β_0 και β_1 της γραμμικής παλινδρόμησης $Y = \beta_0 + \beta_1 X + \varepsilon$.

Λύση. Απλή γραμμική παλινδρόμηση: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $i = 1, \ldots, 10$
$n = 10$

X_i	Y_i	X_iY_i	X_i^2
30	73	2190	900
20	50	1000	400
60	128	7680	3600
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\sum X_i = 500$	$\sum Y_i = 1100$	$\sum X_iY_i = 61800$	$\sum X_i^2 = 28400$

οπότε

$$\hat{\beta}_1 = \frac{61800 - \frac{1100 \cdot 500}{10}}{28400 - \frac{500^2}{10}} = 2$$

και

$$\hat{\beta}_0 = \frac{1}{10}[1100 - 2500] = \frac{1}{10}100 = 10$$

Οπότε αφού $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

$$\Rightarrow \hat{Y} = 10 + 2X$$

Δηλαδή εκτιμάται ότι ο μέσος αριθμός ωρών αυξάνεται κατά 2 για κάθε απιπλέον θρανίο.

Θεώρημα 2.1 *Τα $\hat{\beta}_0$ και $\hat{\beta}_1$ είναι γραμμικοί συνδυασμοί των Y_i.*

Απόδειξη. Θα δείξουμε ότι η $\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$ μπορεί να γραφτεί ως $\hat{\beta}_1 = \sum k_i Y_i$,
όπου $k_i = \frac{X_i - \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2}$
Και επειδή τα X_i είναι γνωστές σταθερές και τα k_i θα είναι γνωστές σταθερές και άρα το $\hat{\beta}_1$ είναι γραμμικός συνδυασμός των Y_i.

Έχουμε

$$\begin{aligned}
\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) &= \sum_{i=1}^n (X_i - \bar{X})Y_i - \sum_{i=1}^n (X_i - \bar{X})\bar{Y} \\
&= \sum_{i=1}^n (X_i - \bar{X})Y_i - \bar{Y}\cancel{\sum_{i=1}^n (X_i - \bar{X})} \\
&= \sum_{i=1}^n (X_i - \bar{X})Y_i
\end{aligned}$$

Άρα $\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \sum k_i Y_i$
Προφανώς και $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = \frac{1}{n}\sum Y_i - \bar{X}\sum k_i Y_i$ γραμ. συνδυασμός των Y_i.

Ιδιότητες των ποσοτήτων k_i

- $\sum k_i = 0$, γιατί $\dfrac{\sum_{i=1}^{n}(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \dfrac{0}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = 0$
- $\sum k_i X_i = 1$, γιατί $\sum k_i X_i = \sum \dfrac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} X_i = \dfrac{\sum (X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = 1$

αφού $(X_i - \bar{X})X_i = X_i^2 - X_i\bar{X}) = X_i^2 - X_i\bar{X} + \bar{X}^2 - \bar{X}^2 - X_i\bar{X} + X_i\bar{X} = (X_i - \bar{X})^2 + \bar{X}\cancel{(X_i - \bar{X})} = (X_i - \bar{X})^2$

- $\sum k_i^2 = \dfrac{1}{\sum (X_i - \bar{X})^2}$,

γιατί $\sum k_i^2 = \sum \Big[\dfrac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2}\Big]^2 = \dfrac{\cancel{\sum (X_i - \bar{X})^2}}{(\sum (X_i - \bar{X})^2)\cancel{^2}} = \dfrac{1}{\sum (X_i - \bar{X})^2}$

Θεώρημα 2.2 *(Θεώρημα των Gauss − Markov) Για το απλό γραμμικό μοντέλο οι εκτιμήτριες ελαχίστων τετραγώνων $\hat{\beta}_0$ $\hat{\beta}_1$*

1) είναι αμερόληπτες

2) έχουν ελάχιστη διασπορά μεταξύ των αμερόληπτων εκτιμητριών που είναι γραμμικές συναρτήσεις των Y_i

Απόδειξη

Αμεροληψία της $\hat{\beta}_1$: Θέλουμε να δείξουμε ότι $E(\hat{\beta}_1) = \beta_1$.

$$E(\hat{\beta}_1) = E(\sum k_i Y_i) = \sum k_i E(Y_i) = \sum k_i(\beta_0 + \beta_1 X_i) = \beta_0 \underbrace{\sum k_i}_{0} + \beta_1 \underbrace{\sum k_i X_i}_{1} = \beta_1$$

Έστω ότι όλες οι αμερόληπτες εκτιμήτριες του β_1 που είναι γραμμικές συναρτήσεις των Y_i είναι της μορφής

$$b_1 = \sum c_i Y_i,$$

όπου c_i αυθαίρετες σταθερές. Επειδή έχουμε αμεροληψία:

$$E(b_i) = \beta_i \Rightarrow E(\sum c_i Y_i) = \sum c_i(\beta_0 + \beta_1 X_i) = \beta_0 \sum c_i + \beta_1 \sum c_i X_i = \beta_1.$$

Άρα πρέπει $\sum c_i = 0$ και $\sum c_i X_i = 1$.
Η διασπορά του b_1 είναι

$$V(b_1) = V(\sum c_i Y_i) = \sum c_i^2 V(Y_i) = \sum c_i^2 \sigma^2 = \sigma^2 \sum c_i^2,$$

αφού $Cov(Y_i, Y_j) = 0$.

Έστω ότι τα c_i έχουν τη μορφή $c_i = k_i + d_i$ όπου τα k_i είναι όπως ορίστηκαν στην εκτιμήτρια $\hat{\beta}_1 = \sum k_i Y_i$ και τα d_i είναι αυθαίρετες σταθερές.

Συνεπώς

$$\begin{aligned} V(b_1) &= \sigma^2 \sum c_i^2 = \sigma^2 \sum (k_i + d_i)^2 \\ &= \sigma^2 \Big[\sum k_i^2 + \sum d_i^2 + 2 \sum k_i d_i \Big] \\ &= \underbrace{\sigma^2 \sum k_i^2}_{V(\hat{\beta}_1)} + \sigma^2 \sum d_i^2 + 2\sigma^2 \sum k_i d_i \end{aligned}$$

Έχουμε $\sum k_i = 0$ και $\sum c_i = \sum (k_i + d_i) = 0 \Rightarrow \sum d_i = 0$

$\sum k_i X_i = 1$ και $\sum c_i X_i = \sum (k_i + d_i) X_i = 1 \Rightarrow \sum d_i X_i = 0$.

Είναι $\sum k_i d_i = \dfrac{\sum (X_i - \bar{X}) d_i}{\sum_{i=1}^n (X_i - \bar{X})^2} = \dfrac{\sum X_i d_i}{\sum_{i=1}^n (X_i - \bar{X})^2} - \bar{X} \dfrac{\sum d_i}{\sum_{i=1}^n (X_i - \bar{X})^2} = 0$, οπότε $V(b_1) = V(\hat{\beta}_1) + \sigma^2 \sum d_i^2$.

Η ποσότητα $\sigma^2 \sum d_i^2$ ελαχιστοποιείται για $\sum d_i^2 = 0$. Άρα η διασπορά του b_1 είναι ελάχιστη όταν $\sum d_i^2 = 0. \Leftrightarrow d_i = 0 \ \forall i$, δηλαδή $c_i = k_i, \forall i$.
Συνεπώς η εκτιμήτρια των ελαχίστων τετραγώνων (ε.ε.τ.), $\hat{\beta}_1$, έχει την ελάχιστη διασπορά μεταξύ των αμερόληπτων γραμμικών εκτιμητριών.

Αμεροληψία της $\hat{\beta}_0$:

$$E(\hat{\beta}_0) = E(\bar{Y} - \hat{\beta}_1 \bar{X}) = E\Big[\frac{1}{n}(\sum_{i=1}^n Y_i - \hat{\beta}_1 \sum_{i=1}^n X_i)\Big] =$$

$$= \frac{1}{n}\Big[\sum_{i=1}^n E(Y_i) - (\sum_{i=1}^n X_i) E(\hat{\beta}_1)\Big] =$$

$$= \frac{1}{n}\Big[\sum (\beta_0 + \beta_1 X_i) - \beta_1 \sum X_i\Big] =$$

$$= \frac{1}{n}\Big[n\beta_0 + \cancel{\beta_1 \sum X_i} - \cancel{\beta_1 \sum X_i}\Big] = \beta_0.$$

Τα X_i δεν είναι τυχαίες μεταβλητές.
Τα Y_i είναι τυχαίες μεταβλητές αλλά <u>όχι ισόνομες</u> (έχουν διαφορετικές αναμενόμενες τιμές και κοινή διακύμανση).

2.2.2 Εκτίμηση της Διασποράς

Αν $Y_1, Y_2, \ldots, Y_n$ τυχαίο δείγμα από κατανομή με γνωστό μέσο μ και διασπορά σ^2, τότε η εκτιμήτρια του σ^2 είναι η $\hat{\sigma}^2 = \frac{1}{n}\sum(Y_i - \mu)^2$. Αν ο μέσος μ είναι άγνωστος θα εκτιμηθεί από το $\bar{Y}$ και το σ^2 εκτιμάται από το $S^2 = \frac{1}{n-1}\sum(Y_i - \bar{Y})^2$, όπου $Y_i - \bar{Y}$ είναι οι αποκλίσεις των Y_i από τον κοινό τους μέσο. Έχουμε $E(S^2) = \sigma^2$, δηλαδή η S^2 είναι αμερόληπτη εκτιμήτρια της διασποράς. Διαιρούμε με $n-1$ βαθμούς ελευθερίας γιατί έχουμε εκτιμήσει το $\bar{X}$.

Στο γραμμικό μοντέλο τα Y_i έχουν διαφορετικές κατανομές που εξαρτώνται από τα X_i. Επομένως, η απόκλιση κάθε παρατήρησης πρέπει να υπολογιστεί από το μέσο της, δηλαδή από το $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$. Άρα, αν συμβολίσουμε με $\hat{\varepsilon}_i$ τις εκτιμήσεις των σφαλμάτων, ή αλλιώς τα κατάλοιπα (residuals), υπολογίζουμε το άθροισμα των τετραγώνων των καταλοίπων (error sum of squares or residuals sum of squares)

$$\sum \hat{\varepsilon}_i{}^2 = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2.$$

Μια αμερόληπτη εκτιμήτρια του σ^2 είναι το μέσο τετραγωνικό σφάλμα (mean square error)

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2.$$

Διαιρούμε με $n-2$ βαθμούς ελευθερίας καθώς έχουν εκτιμηθεί δύο παράμετροι.

2.3 Γραμμικό Μοντέλο με σφάλματα που ακολουθούν Κανονική Κατανομή

Στο απλό γραμμικό μοντέλο συχνά υποθέτουμε ότι τα τυχαία σφάλματα ακολουθούν την κανονική κατανομή. Έχουμε τότε το μοντέλο

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, \varepsilon_i \sim N(0, \sigma^2)$$

Παρατήρηση: Η υπόθεση $Cov(\varepsilon_i, \varepsilon_j) = 0$ (ασυσχέτιστα σφάλματα) είναι και υπόθεση ανεξαρτησίας (ανεξάρτητα σφάλματα) κάτω από την κανονική κατανομή.

2.3.1 Εκτίμηση παραμέτρων με τη Μέθοδο Μέγιστης Πιθανοφάνειας

Έχουμε $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^{2)}, \quad i = 1, \ldots, n.$

Συνάρτηση πιθανοφάνειας ($likelihood\ function$):

$$L(\beta_0, \beta_1, \sigma^2) = \prod_1^n \frac{1}{\sqrt{2\pi\sigma^2}} exp\Big\{ -\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2 \Big\}$$

$$= (2\pi\sigma^2)^{-n/2} exp\Big\{ -\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2 \Big\}$$

Συνάρτηση log-πιθανοφάνειας ($log - likelihood\ function$):

$$\log L = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum (Y_i - \beta_0 - \beta_1 X_i)^2$$

Οι εκτιμήτριες μέγιστης πιθανοφάνειας προκύπτουν από τη μεγιστοποίηση της $\log L$ ως προς τις παραμέτρους.

$$\left.\begin{aligned} \frac{\partial \log L}{\partial \beta_0} &= \frac{1}{\sigma^2}\sum (Y_i - \beta_0 - \beta_1 X_i) = 0 \\ \frac{\partial \log L}{\partial \beta_1} &= \frac{1}{\sigma^2}\sum X_i (Y_i - \beta_0 - \beta_1 X_i) = 0 \end{aligned}\right\} \text{αντίστοιχες με τις κανονικές εξισώσεις}$$

$\Rightarrow \hat{\beta}_0, \hat{\beta}_1$ ίδιες με εκτιμήτριες ελαχίστων τετραγώνων.

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum (Y_i - \beta_0 - \beta_1 X_i)^2 = 0$$

$$\Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

Η $\hat{\sigma}^2$ δεν είναι αμερόληπτη εκτιμήτρια του σ^2.

Παρατήρηση
Οι εκτιμήτριες $\hat{\beta}_0$ και $\hat{\beta}_1$ σαν εκτιμήτριες ελαχίστων τετραγώνων
1) είναι αμερόληπτες
2) έχουν ελάχιστη διασπορά μεταξύ των α.ε. που είναι γραμμικοί συνδυασμοί των Y_i.

Επίσης σαν εκτιμήτριες μέγιστης πιθανοφάνειας είναι
(1) συνεπείς
(2) επαρκείς
(3) αμερόληπτες εκτιμήτριες ελάχιστης διασποράς (έχουν ελάχιστη διασπορά μεταξύ <u>όλων</u> των α.ε.)

<u>Κατανομή της $\hat{\beta}_1$</u>

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1, \ldots, \varepsilon_i \sim N(0, \sigma^2), \text{ανεξάρτητα} \Rightarrow Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$

Είναι $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \sum k_i Y_i$

Άρα η $\hat{\beta}_1$ ακολουθεί την κανονική κατανομή ως γραμμικός συνδυασμός κανονικών τ.μ.
Έχουμε δείξει ότι

$$E(\hat{\beta}_1) = \beta_1$$

$$\sigma^2(\hat{\beta}_1) = V(\sum k_i Y_i) = \sum ki_2 V(Y_i) = \sigma^2 \sum k_i^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Άρα η $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2})$

Η διασπορά σ^2 είναι άγνωστη. Μπορεί όμως να εκτιμηθεί από την α.ε. της

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

Οπότε η εκτιμούμενη διασπορά της $\hat{\beta}_1$ είναι

$$S^2(\beta_1) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

.

Κατανομή της $\hat{\beta}_0$

Είναι $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ όπου τα Y_i είναι κανονικές τ.μ. και το $\hat{\beta}_1$ είναι επίσης κανονική τ.μ. Άρα το $\hat{\beta}_0$ ακολουθεί την κανονική κατανομή ως γραμμικός συνδυασμός κανονικών τ.μ.
Εχουμε δείξει ότι

$$E(\hat{\beta}_0) = \beta_0$$

$$\sigma^2(\hat{\beta}_0) = V(\bar{Y} - \hat{\beta}_1 \bar{X}) = V(\bar{Y}) + V(\hat{\beta}_1 \bar{X}) - 2Cov(\bar{Y}, \hat{\beta}_1 \bar{X})$$

$$= \frac{\sigma^2}{n} + \bar{X} V(\hat{\beta}_1) - 2\bar{X} \underbrace{Cov(\bar{Y}, \hat{\beta}_1)}_{0} = \frac{\sigma^2}{n} + \bar{X} \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$= \sigma^2 \Big[\frac{1}{n} + \frac{\bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\Big]$$

Εκτιμούμενη διασπορά: $S^2(\hat{\beta}_0) = \hat{\sigma}\Big[\frac{1}{n} + \frac{\bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\Big]$

Πρόταση 2.1 *Είναι* $Cov(\hat{\beta}_1, \bar{Y}) = 0$

Απόδειξη.
Ιδιότητες συνδιακύμανσης: $Cov(X+Y,Z) = Cov(X,Z) + Cov(X,Y)$
$Cov(X+Y, Z+W) = Cov(X,Z) + Cov(X,W) + Cov(Y,Z) + Cov(Y,W)$

$$Cov(\hat{\beta}_1, \bar{Y}) = Cov(\sum k_i Y_i, \frac{\sum Y_i}{n}) = \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(k_i Y_i, \frac{Y_j}{n}) = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{k_i}{n} Cov(Y_i, Y_j)$$

$$= \sum \frac{k_i}{n} Cov(Y_i, Y_j) = \frac{1}{n}\sum k_i V(Y_i) = \frac{\sigma^2}{n}\underbrace{\sum k_i}_{0} = 0$$

αφού $Cov(Y_i, Y_j) = 0, i \neq j$ ■

Πρόταση 2.2 *Η τ.μ.* $\frac{\hat{\beta}_1 - \beta_1}{s(\beta_1)} \sim t_{(n-2)}$

Απόδειξη.
Ισχύει: Αν $Z \sim N(0,1), U \sim X^2_{(r)}$και Z, U ανεξάρτητες
τότε

$$T = \frac{Z}{\sqrt{\frac{U}{\tau}}} \sim t_{(r)}$$

Έχουμε $\hat{\beta}_1 \sim N(\beta_1, \sigma^2(\hat{\beta}_1)) \Rightarrow \frac{\hat{\beta}_1 - \beta_1}{\sigma(\beta_1)} \sim N(0,1)$

Επίσης ισχύει ότι

$$\frac{\sum \varepsilon_i^2}{\sigma^2} = \frac{\sum (Y_i - \hat{Y}_i)}{\sigma^2} = \frac{\sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{\sigma^2} \sim X^2(n-2)$$

Και επειδή $\hat{\sigma}^2 = \frac{1}{n-2}\sum \hat{\varepsilon}_i^2$, έχουμε $\frac{\hat{\sigma}^2(n-2)}{\sigma^2} \sim X^2(n-2)$.
Τώρα

$$\left.\begin{aligned} S^2(\hat{\beta}_1) &= \frac{\hat{\sigma}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ \sigma^2(\hat{\beta}_1) &= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \end{aligned}\right\} \Rightarrow \frac{S^2(\hat{\beta}_1)}{\sigma^2(\hat{\beta}_1)} = \frac{\hat{\sigma}^2}{\sigma^2}$$

Άρα $\frac{S^2(\hat{\beta}_1)(n-2)}{\sigma^2(\hat{\beta}_1)} \sim X^2(n-2)$
Επομένως

$$\frac{\frac{\hat{\beta}_1 - \beta_1}{\sigma(\hat{\beta}_1)}}{\sqrt{\frac{S^2(\hat{\beta}_1)\cancel{(n-2)}}{\sigma^2(\hat{\beta}_1)\cancel{(n-2)}}}} = \frac{\hat{\beta}_1 - \beta_1}{s(\beta_1)} \sim t_{(n-2)}$$

■

Άσκηση 2.1 *Να βρεθεί διάστημα εμπιστοσύνης με συντελεστή εμπιστοσύνης 1-α για το* β_1

Λύση.

Είναι $P\Big(-t_{\frac{\alpha}{2}}(n-2) \leq \dfrac{\hat{\beta}_1 - \beta_1}{s(\hat{\beta}_1)} \leq t_{\frac{\alpha}{2}}(n-2)\Big) = 1-\alpha$

Άρα το ζητούμενο δ.ε. για το β_1 *είναι*

$$\hat{\beta}_1 \pm t_{\frac{\alpha}{2}}(n-2)s(\hat{\beta}_1)$$

Ομοίως είναι $\dfrac{\hat{\beta}_0 - \beta_0}{s(\beta_0)} \sim t(n-2)$*άρα δ.ε. με συντελεστή 1-α για το* β_0*:*

$$\hat{\beta}_0 \pm t_{\frac{\alpha}{2}}(n-2)s(\hat{\beta}_0)$$

Χρησιμοποιώντας την κατανομή $t(n-2)$ *μπορούμε να κάνουμε και ελέγχους υποθέσεων για τις παραμέτρους* β_0 *και* β_1*:*

$$\begin{array}{ll} H_0: & \beta_0 = 0 \\ H_1: & \beta_0 \neq 0 \end{array} \qquad\qquad \begin{array}{ll} H_0: & \beta_1 = 0 \\ H_0: & \beta_1 \neq 0 \end{array}$$

■

Άσκηση 2.2 *Έστω το εναλλακτικό γραμμικό μοντέλο*

$$Y_i = \beta_0^* + \beta_1(X_i - \bar{X}) + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2), \alpha\nu\varepsilon\xi$$

Να βρεθεί η κατανομή του $\hat{\beta}_0^*$

Λύση

Αρχικό μοντέλο

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \ \delta\eta\lambda\alpha\delta\acute{\eta} \ \ \beta_0^* = \beta_0 + \beta_1\bar{X}$$
$$\text{Άρα} \ \ \hat{\beta}_0^* = \hat{\beta}_0 + \hat{\beta}_1\bar{X} = \bar{Y} - \hat{\beta}_1\bar{X} + \hat{\beta}_1\bar{X} = \bar{Y}$$

Επειδή τα Y_i *είναι κανονικές τ.μ., το* $\hat{\beta}_0^* = \bar{Y}$ *είναι επίσης κανονική τ.μ. ως γραμμικός συνδυασμός ανεξάρτητων κανονικών τ.μ.*

$$E(\hat{\beta}_0^*) = E(\hat{\beta}_0 + \hat{\beta}_1\bar{X}) = E(\hat{\beta}_0) + \bar{X}E(\hat{\beta}_1) = \beta_0 + \beta_1\bar{X} = \beta_0^*$$

Εναλλακτικά

$$E(\hat{\beta}_0^*) = E(\bar{Y}) = E\Big(\frac{\sum Y_i}{n}\Big) = \frac{1}{n}\sum E(Y_i) = \frac{1}{n}\sum(\beta_0^* + \beta_1(X_i - \bar{X}))$$
$$= \frac{1}{n}n\beta_0^* + \frac{1}{n}\underbrace{\sum(X_i - \bar{X})}_{0} = \beta_0^*$$

Δηλαδή $\hat{\beta}_0^*$ *αμερόληπτη εκτιμήτρια.*

$$V(\hat{\beta}_0^*) = \sigma^2(\hat{\beta}_0^*) = V(\bar{Y}) = V\Big(\frac{\sum Y_i}{n}\Big) = \frac{1}{n^2}\sum V(Y_i) = \frac{1}{n^2}\sum\sigma^2 = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

Άρα $\hat{\beta}_0^* \sim N(\beta_0^*, \frac{\sigma^2}{n})$ ■

2.4 Το Απλό Γραμμικό Μοντέλο με Πίνακες

Είναι

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1, \ldots, n$$

Δηλαδή έχουμε

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_1 + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_2 + \varepsilon_2 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 X_n + \varepsilon_n \end{aligned}$$

Ορίζουμε

$$\underset{(n\times 1)}{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \underset{(n\times 2)}{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}, \underset{(2\times 1)}{B} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \underset{(n\times 1)}{\mathcal{E}} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Το μοντέλο μπορεί να γραφτεί σε μορφή πινάκων ως

$$\underset{(n\times 1)}{Y} = \underset{(n\times 1)}{XB} + \underset{(n\times 1)}{\mathcal{E}}$$

Θέλουμε να εκτιμήσουμε το διάνυσμα B.

Θα χρειαστούμε:

- $Y'Y = (Y_1, Y_2, \ldots, Y_n) \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \sum Y_{i^2}$

  $$X'X = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ X_1 & X_2 & \ldots & X_n \end{pmatrix} \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} = \begin{pmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{pmatrix}$$

  $$X'Y = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ X_1 & X_2 & \ldots & X_n \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \sum Y_i \\ \sum X_i Y_i \end{pmatrix}$$

- Αντίστροφος τετραγωνικού πίνακα (2×2)
  $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ και έστω η ορίζουσα $D = ad - cb \neq 0$

Τότε μπορώ να βρώ τον αντίστροφο: $A^{-1} = \begin{pmatrix} d/D & -b/D \\ -c/D & a/D \end{pmatrix}$

$\Rightarrow$ Ο $X'X = \begin{pmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{pmatrix}$ έχει ορίζουσα $D = n\sum X_i^2 - (\sum X_i)^2 =$

$$= n\Big[\sum X_i^2 - \frac{(\sum X_i)^2}{n}\Big] = n\sum (X_i - \bar{X})^2 > 0$$

Άρα $(X'X)^{-1} = \begin{pmatrix} \dfrac{\sum X_i)^2}{n\sum (X_i - \bar{X})^2} & -\dfrac{\bar{X}}{\sum (X_i - \bar{X})^2} \\ -\dfrac{\bar{X}}{\sum (X_i - \bar{X})^2} & \dfrac{1}{\sum (X_i - \bar{X})^2} \end{pmatrix}$

- Έστω Y τυχαίο διάνυσμα $n \times 1$, δηλαδή Y διάνυσμα του οποίου τα στοιχεία είναι τυχαίες μεταβλητές. Είναι

  $E(Y) = \begin{pmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n) \end{pmatrix}$, $\underset{\text{Πίνακας Συνδιακύμανσης}}{\sigma^2(Y) = D(Y)}$ $n \times n$ πίνακας με ij στοιχείο $\sigma(Y_i, Y_j) = Cov(Y_i, Y_j)$

  και στη διαγώνιο οι διασπορές $\sigma^2(Y_i)$

  Άρα, στο γραμμικό μοντέλο $D(Y) = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 \dots & \sigma^2 & \end{pmatrix} = \sigma^2 I_{(n\times n)}$

- Εάν Y, W τυχαία διανύσματα και A πίνακας σταθερών τέτοια ώστε $W = AY$, τότε
  $E(A) = A$
  $E(W) = E(AY) = AE(Y)$
  $D(W) = D(AY) = AD(Y)A'$

Εκτιμήτρια ελαχίστων τετραγώνων του B

Κανονικές εξισώσεις:
$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i$$
$$\hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i$$

Υπό μορφή πινάκων:

$$X'X\hat{B} = X'Y \tag{2.1}$$

γιατί $X'X\hat{B} = \begin{pmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} n\hat{\beta}_0 + \hat{\beta}_1 \sum X_i \\ \hat{\beta}_0 \sum X_i + \hat{\beta}_1 \sum X_i^2 \end{pmatrix}$

και $X'Y = \begin{pmatrix} \sum Y_i \\ \sum X_i Y_i \end{pmatrix}$

(2.1) $\overset{\text{Λύνω ως προς } \hat{B}}{\Longrightarrow}$ $\underbrace{(X'X)^{-1}(X'X)}_{I}\hat{B} = (X'X)^{-1}X'Y \Rightarrow \boxed{\hat{B} = (X'X)^{-1}X'Y}$

$$E(\hat{B}) = E\Big((X'X)^{-1}X'Y\Big) = (X'X)^{-1}X'E(Y)$$
$$= (X'X)^{-1}X'[E(XB+\mathcal{E})]$$
$$= (X'X)^{-1}X'XB = B$$

$$D(\hat{B}) = D\Big((X'X)^{-1}X'Y\Big) = (X'X)^{-1}X'D(Y)\Big((X'X)^{-1}X'\Big)'$$
$$= (X'X)^{-1}X'\sigma^2 I_n X(X'X)^{-1}$$
$$= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} = \sigma^2(X'X)^{-1}$$

2.5 Μετασχηματισμοί Συνάρτησης Παλινδρόμησης ώστε να γίνει Γραμμική

- Πολλαπλασιαστικό Μοντέλο

  $Y_i = \gamma_0\gamma_1^{X_i}\varepsilon_i$, όπου γ_0, γ_1 παράμετροι και ε_i τ.μ. με $E(\varepsilon_i) = 1$.

  Παίρνοντας το μετασχηματισμό $Y_i' = \log_{10} Y_i$ έχουμε

  $$Y_i' = \log_{10} Y_i = \log_{10}\gamma_0 + X_i \log_{10}\gamma_1 + \log_{10}\varepsilon_i$$

  ή $Y_i' = \beta_0 + \beta_1 X_i + \varepsilon_i'$, όπου $\beta_0 = \log_{10}\gamma_0, \beta_1 = \log_{10}\gamma_1$ παράμετροι και ε_i' τυχαία σφάλματα με $E(\varepsilon_i) = 0$.

- Μοντέλο $Y_i = \beta_0 + \dfrac{\beta_1}{X_i} + \varepsilon_i$

  Χρησιμοποιούμε το μετασχηματισμό $X_i' = \dfrac{1}{X_i}$, οπότε έχουμε

  $$Y_i' = \beta_0 + \beta_1 X_i' + \varepsilon_i$$

- Μοντέλο $Y_i = \dfrac{1}{1 + exp\{\beta_0 + \beta_1 X_i + \varepsilon_i\}}$

  Παίρνουμε $Y_i' = \dfrac{1}{Y_i}$, οπότε

  $$Y_i' = 1 + exp\{\beta_0 + \beta_1 X_i + \varepsilon_i\}$$

  Συνεχιζουμε θέτοντας $Y_i'' = \ln(Y_i' - 1)$,οπότε

  $$Y_i' = \beta_0 + \beta_1 X_i + \varepsilon_i$$

2.6 Πολλαπλή Παλινδρόμηση

Γενικό γραμμικό μοντέλο με p ανεξάρτητες μεταβλητές

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i$$

Μετασχηματισμοί

- Μοντέλο $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$
  Θέτουμε $X_{i1} = X_i$ και $X_{i2} = X_i^2$ οπότε παίρνουμε
  $$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$
- Μοντέλο $Y_i = \dfrac{1}{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i}$
  Θέτουμε $Y_i' = \dfrac{1}{Y_i}$ οπότε έχουμε
  $$Y_i' = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$
- Μοντέλο $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i$
  θέτουμε $X_{i3} = X_{i1} X_{i2}$οπότε έχουμε
  $$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + + \beta_3 X_{i3} + \varepsilon_i$$

Γενικά για το μοντέλο πολλαπλής παλινδρόμησης $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i$ οι κανονικές εξισώσεις είναι

$$X'X\hat{B} = X'Y.$$

Για το μοντέλο $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$ έχουμε

$$X = \begin{pmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} \end{pmatrix}, B = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}, Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

$$X'X\hat{B} = \begin{pmatrix} n & \sum X_{i1} & \sum X_{i2} \\ \sum X_{i1} & \sum X_{i1}^2 & \sum X_{i1}X_{i2} \\ \sum X_{i2} & \sum X_{i2}X_{i1} & \sum X_{i2}^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \sum Y_i \\ \sum X_{i1}Y_i \\ \sum X_{i2}Y_i \end{pmatrix} = X'Y$$

Γενικά έχουμε

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} 1 & X_{11} & \ldots & X_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \ldots & X_{np} \end{pmatrix}, B = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \mathcal{E} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{pmatrix}$$

Οι κανονικές εξισώσεις μας δίνουν την εκτιμήτριες ελαχίστων τετραγώνων για το B

$$X'X\hat{B} = X'Y \qquad \Rightarrow \qquad \hat{B} = (X'X)^{-1}X'Y$$

$$E(\hat{B}) = B$$

$$D(\hat{B}) = \sigma^2(X'X)^{-1}$$

2.7 Κατανομές τετραγωνικών μορφών

Ορισμός 2.1 *Ένα ομογενές πολυώνυμο 2ου βαθμού σε n μεταβλητές καλείται τετραγωνική μορφή σε αυτές τις μεταβλητές. Εάν και οι συντελεστές και οι μεταβλητές είναι πραγματικοί αριθμοί έχουμε πραγματική τετραγωνική μορφή.*

Παράδειγματα 2.1 *(1)*$X_1^2 + X_1X_2 + X_2^2$ *τετραγωνική μορφή ως προς* X_1, X_2
(2)$X_1^2 + X_2^2 + X_3^2 - 2X_1X_2$ *τετραγωνική μορφή ως προς* X_1, X_2, X_3
(3)$X_1^2 + X_2^2 - 2X_1 - 4X_2 + 5$ *δεν είναι τετραγωνική μορφή ως προς* X_1, X_2

$$\| \qquad (X_1 - 1)^2 + (X_2 - 2)^2 \qquad \text{αλλά είναι ως προς } (X_1 - 1), (X_2 - 2)$$

Πρόταση 2.3 *Έστω* $X_1, X_2, \ldots, X_n$ *τ.δ. από την τ.μ.* X *και* $\bar{X}, S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$ *ο δειγματικός μέσος και η δειγματική διασπορά αντίστοιχα. Η τ.μ.* $(n-1)S^2$ *είναι μια τετραγωνική μορφή στις* n *τ.μ.* $X_1, X_2, \ldots, X_n$

Απόδειξη

$$\begin{aligned}
(n-1)S^2 &= \sum(X_i - \bar{X})^2 = \sum(X_i - \tfrac{\sum X_i}{n})^2 \\
&= \sum X_i^2 + \frac{n(\sum X_i)^2}{n^2} - 2\sum X_i \frac{\sum X_i}{n} = \\
&= \sum X_i^2 + \frac{(\sum X_i)^2}{n} - \frac{2}{n}(\sum X_i)^2 = \\
&= \sum X_i^2 - \frac{(\sum X_i)^2}{n} = \sum X_i^2 - \frac{1}{n}\Big\{\sum X_i + 2\sum_{\substack{i,j=1 \\ i<j}}^{n} X_iX_j\Big\} = \\
&= \frac{n-1}{n}\sum X_i^2 - \frac{2}{n}\sum_{\substack{i,j=1 \\ i<j}}^{n} X_iX_j \qquad \text{τετραγωνική μορφή}
\end{aligned}$$

■

Παρατήρηση. Γνωρίζουμε ότι εάν $X \sim N(\mu, \sigma^2)$ και $X_1, X_2, \ldots, X_n$ τ.δ. τότε $\frac{(n-1)S^2}{\sigma^2} \sim X^2_{(n-1)}$.

Παρατήρηση. Γνωρίζουμε ότι εάν $Q_1, Q_2, \ldots, Q_n$ είναι ανεξάρτητες X^2 τ.μ. με $r_i, i = 1, \ldots, k$ β.ε. αντίστοιχα, τότε η τ.μ. $Q = \sum_{i=1}^{k} Q_i$ είναι $X^2_{(r_1+r_2+\ldots+r_k)}$.

Θεώρημα 2.3 *Έστω $Q = \sum_{i=1}^{k} Q_i$ όπου $Q_1, Q_2, \ldots, Q_n$ είναι k πραγματικές τετραγωνικές μορφές σε n ανεξάρτητες τ.μ. που ακολουθούν την κανονική κατανομή με μέσους $\mu_1, \mu_2, \ldots, \mu_n$ αντίστοιχα και την ίδια διασπορά σ^2.*
Έστω επίσης ότι (i) $\frac{Q}{\sigma^2}, \frac{Q_1}{\sigma^2}, \frac{Q_2}{\sigma^2}, \ldots, \frac{Q_{k-1}}{\sigma^2}$ έχουν την κατανομή X^2 με $r, r_1, r_2, \ldots, r_{k-1}$ β.ε. αντίστοιχα, και (ii) Q_k μη αρνητική. Τότε
(1) $Q_1, Q_2, \ldots, Q_n$ είναι ανεξάρτητες και
(2) $\frac{Q_k}{\sigma^2} \sim X^2_{(r_k)}$, όπου $r_k = r - (r_1 + r_2 + \ldots + +r_k)$

Πρόταση 2.4 *Έστω το γραμμικό μοντέλο*

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_{i,}$$

ή

$$Y_i = \beta_0^* + \beta_1(X_i - \bar{X}) + \varepsilon_i \text{ με } \beta_0^* = \beta_0 + \beta_1 \bar{X}$$

Εάν $\varepsilon_i \sim N(0, \sigma^2)$, τότε $Y_i \overset{\text{ανεξ}}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$
ή $Y_i \sim N(\beta_0^* + \beta_1(X_i - \bar{X}), \sigma^2)$.
Έχουμε $$\frac{\sum \hat{\varepsilon}_i^2}{\sigma^2} = \frac{\sum (Y_i - \hat{Y}_i)^2}{\sigma^2} = \frac{\sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)}{\sigma^2} \sim X^2_{(n-2)}$$
$$\text{ή } \frac{\sum (Y_i - \hat{\beta}_0^* - \hat{\beta}_1 (X_i - \bar{X})}{\sigma^2} \sim X^2_{(n-2)}$$

Απόδειξη

$$\sum (Y_i - \beta_0 - \beta_1 X_i)^2 = \sum (Y_i - \hat{\beta}_0^* - \hat{\beta}_1 (X_i - \bar{X}))^2 =$$
$$= \sum \left[(\hat{\beta}_0^* - \beta_0^*) + (\hat{\beta}_1 - \beta_1)(X_i - \bar{X}) + [Y_i - \hat{\beta}_0^* - \hat{\beta}_1 (X_i - \bar{X})] \right]^2,$$
όπου έχω προσθαφαιρέσει $\hat{Y}_i = \hat{\beta}_0^* - \hat{\beta}_1 (X_i - \bar{X})$
$$= n(\hat{\beta}_0^* - \beta_0^*)^2 + (\hat{\beta}_1 - \beta_1) \sum (X_i - \bar{X})^2 + \sum \left[(Y_i - \hat{\beta}_0^* - \hat{\beta}_1 (X_i - \bar{X}) \right]^2 + 0$$

Συμβολίζουμε με

$$Q = \sum (Y_i - \beta_0 - \beta_1 X_i)^2 = \sum \left(Y_i - \beta_0^* - \beta_1 (X_i - \bar{X}) \right)^2$$

$$Q_1 = n(\hat{\beta}_0^* - \beta_0^*)^2$$

$$Q_2 = (\hat{\beta}_1 - \beta_1)^2 \sum (X_i - \bar{X})^2$$

$$Q_3 = \sum \left[Y_i - \hat{\beta}_0^* - \hat{\beta}_1 (X_i - \bar{X}) \right]^2$$

Είναι $Q = Q_1 + Q_2 + Q_3$.

Τα Q, Q_1, Q_2, Q_3 είναι τετραγωνικές μορφές ως προς κανονικές τ.μ.

Είναι $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2) \Rightarrow \dfrac{(Y_i - \beta_0 - \beta_1 X_i)^2}{\sigma^2} \sim X^2_{(1)}$

Οι Y_i είναι ανεξάρτητες άρα $\sum \dfrac{(Y_i - \beta_0 - \beta_1 X_i)^2}{\sigma^2} \sim X^2_{(n)}$

$$\text{δηλ. } \frac{Q}{\sigma^2} \sim X^2_{(n)}$$

$$\hat{\beta}_0^* \sim N(\beta_0, \frac{\sigma^2}{n}) \Rightarrow \frac{n(\hat{\beta}_0^* - \beta_0^*)^2}{\sigma^2} \sim X^2_{(1)} \text{ , δηλ. } \frac{Q_1}{\sigma^2} \sim X^2_{(1)}$$

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}) \Rightarrow \frac{\sum_{i=1}^n (X_i - \bar{X})^2 (\hat{\beta}_1 - \beta_1)^2}{\sigma^2} = \frac{Q_2}{\sigma^2} \sim X^2_{(1)}$$

Το Q_3 είναι μη αρνητικό, άρα με βάση το θεώρημα

$$\frac{Q_3}{\sigma^2} = \frac{\sum \left[Y_i - \hat{\beta}_0^* - \hat{\beta}_1 (X_i - \bar{X}) \right]^2}{\sigma^2} = \frac{\sum \hat{\varepsilon}_i^2}{\sigma^2} \sim X^2_{(n-1+1)} \equiv X^2_{(n-2)}$$

Μένει να δείξουμε ότι τα διπλάσια γινόμενα είναι μηδενικά.

(i) $2\sum(\hat{\beta}_0^* - \beta_0^*)(\hat{\beta}_1 - \beta_1)(X_i - \bar{X}) = 2(\hat{\beta}_0^* - \beta_0^*)(\hat{\beta}_1 - \beta_1)\sum(X_i - \bar{X}) = 0$

(ii) $2\sum(\hat{\beta}_0^* - \beta_0^*)\Big(Y_i - \hat{\beta}_0^* - \hat{\beta}_1(X_i - \bar{X})\Big) = 2(\hat{\beta}_0^* - \beta_0^*)\sum(Y_i - \bar{Y}) - 2(\hat{\beta}_0^* - \beta_0^*)\hat{\beta}_1 \sum(X_i - \bar{X}) = 0$

(iii) $2\sum(\hat{\beta}_1 - \beta_1)(X_i - \bar{X})\Big(Y_i - \hat{\beta}_0^* - \hat{\beta}_1(X_i - \bar{X})\Big) =$

$= 2(\hat{\beta}_1 - \beta_1)\sum(X_i - \bar{X})(Y_i - \bar{Y}) - 2(\hat{\beta}_1 - \beta_1)\hat{\beta}_1 \sum(X_i - \bar{X})^2 =$

$= 2(\hat{\beta}_1 - \beta_1)\Big[\sum(X_i - \bar{X})(Y_i - \bar{Y}) - \dfrac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\cancel{\sum(X_i - \bar{X})^2}} \cancel{\sum(X_i - \bar{X})^2}\Big] = 0$

2.8 Κατάλοιπα Παλινδρόμησης

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$E(Y_i) = \beta_0 + \beta_1 X_i,\ \varepsilon_i = Y_i - E(Y_i)$$

Η παλινδρόμηση εκτιμάται από την ευθεία $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ (λέμε ότι προσαρμόζουμε αυτή την ευθεία στα δεδομένα μας). Η τιμή $\hat{Y}_i$ καλείται προσαρμοσμένη τιμή (fitted value) ενώ η Y_i παρατηρούμενη τιμή (observed value). Το i κατάλοιπο, $\hat{\varepsilon}_i$, είναι η διαφορά μεταξύ της παρατηρούμενης και της προσαρμοσμένης τιμής:

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i.$$

2.8.1 Ιδιότητες καταλοίπων

1) $\sum \hat{\varepsilon}_i = 0$ $(*)$

Είναι $\sum \hat{\varepsilon}_i = \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = \sum Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum X_i = 0$ (από την πρώτη κανονική εξίσωση) $(*) \Rightarrow \sum (Y_i - \hat{Y}_i) = 0 \Rightarrow \sum Y_i = \sum \hat{Y}_i$ Οι παρατηρούμενες τιμές και οι προσαρμοσμένες τιμές έχουν ίδιο μέσο.

2) $\sum \hat{\varepsilon}_i^2 = 0$ είναι ελάχιστο (απαίτηση στη μέθοδο ελαχίστων τετραγώνων).

3) $\sum X_i \hat{\varepsilon}_i = 0$

Είναι $\sum X_i \hat{\varepsilon}_i = \sum X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = \sum X_i Y_i - \hat{\beta}_0 \sum X_i - \hat{\beta}_1 \sum X_i^2 = 0$ (από την δεύτερη κανονική εξίσωση)

4) $\sum Y_i \hat{\varepsilon}_i = 0$

Είναι $\sum Y_i \hat{\varepsilon}_i = \sum (\hat{\beta}_0 + \hat{\beta}_1 X_i)(Y_i - \hat{Y}_i) = \hat{\beta}_0 \underbrace{\sum (Y_i - \hat{Y}_i)}_{0} + \hat{\beta}_1 \sum X_i (Y_i - \hat{Y}_i) =$

$$= \hat{\beta}_1 \sum \Big(X_i Y_i - X_i (\hat{\beta}_0 + \hat{\beta}_1 X_i) \Big) = \hat{\beta}_1 \Big[\underbrace{\sum X_i Y_i - \hat{\beta}_0 \sum X_i - \hat{\beta}_1 \sum X_i^2}_{0(\text{από 2η κανονική εξίσωση})} \Big] = 0$$

Ας μην ξεχνάμε ότι τα κατάλοιπα $\hat{\varepsilon}_i$ είναι οι εκτιμήσεις των τυχαίων σφαλμάτων, ε_i. Κάτω από τις υποθέσεις του γραμμικού μοντέλου $E(\varepsilon_i) = 0, V(\varepsilon_i) = \sigma^2$, $Cov(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$. Τα κατάλοιπα λοιπόν πρέπει να είναι τυχαία κατανεμημένα γύεω από το 0, ασυσχέτιστα και ομοσκεδαστικά. Αν αυτά δεν ισχύουν, τότε το μοντέλο που έχουμε προσαρμόσει δεν είναι κατάλληλο για τα δεδομένα μας.

Άσκηση 2.3 *(Παλιό θέμα)*

Δίνεται το γραμμικό μοντέλο

$$Y_i = \beta(X_i - \bar{X}) + \varepsilon_i, i = 1, \ldots, n, \varepsilon_i \sim N(0, \sigma^2) \alpha\sigma\upsilon\sigma\chi\acute{\varepsilon}\tau\iota\sigma\tau\alpha$$

(i) Να βρεθεί η ε.ε.τ. $\hat{\beta}$ της β και η κατανομή της.
(ii) Να βρεθεί η κατανομή της στατιστικής συνάρτησης $\frac{\hat{\beta} - \beta}{s(\hat{\beta})}$ όπου $s(\hat{\beta})$ είναι α.ε. της διασποράς του $\hat{\beta}$.
(iii) Για δεδομένη τιμή X_0 να βρεθεί δ.ε. συντελεστή 1-α για το $\beta(X_0 - \bar{X})$.

Λύση.

(i) Εκτίμηση ελαχίστων τετραγώνων: ελαχιστοποιώ την ποσότητα
$Q = \sum \varepsilon_i^2 \sum (Y_i - \beta(X_i - \bar{X}))^2$ *ως προς* β

$$\frac{\partial Q}{\partial \beta} = -2 \sum (X_i - \bar{X})[Y_i - \beta(X_i - \bar{X})] = 0$$

$$\Rightarrow -\sum (X_i - \bar{X}) Y_i + \beta \sum (X_i - \bar{X})^2 = 0 \Rightarrow \hat{\beta} = \frac{\sum (X_i - \bar{X}) Y_i}{\sum (X_i - \bar{X})^2}$$

Το $\hat{\beta}$ είναι γραμμικός συνδυασμός των Y_i, αφού
$\hat{\beta} = \sum k_i Y_i$, *όπου* $k_i = \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2}$ *σταθερές.*
Επειδή $\varepsilon_i \sim N(0, \sigma^2)$ ασυσχέτιστες άρα και ανεξάρτητες τ.μ., ισχύει ότι

$$Y_i \sim N(\beta(X_i - \bar{X}), \sigma^2)$$

Άρα το $\hat{\beta}$ ακολουθεί κανονική κατανομή σαν γραμμικός συνδυασμός ανεξάρτητων κανονικών τ.μ. Είναι

$$E(\hat{\beta}) = E\Big[\frac{\sum (X_i - \bar{X}) Y_i}{\sum (X_i - \bar{X})^2}\Big] = \frac{\sum (X_i - \bar{X}) E(Y_i)}{\sum (X_i - \bar{X})^2}$$

$$= \frac{\sum (X_i - \bar{X}) \beta (X_i - \bar{X})}{\sum (X_i - \bar{X})^2} = \frac{\beta \sum (X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2} = \beta$$

Δηλαδή είναι α.ε.

$$V(\hat{\beta}) = V\Big[\frac{\sum (X_i - \bar{X}) Y_i}{\sum (X_i - \bar{X})^2}\Big] = \frac{\sum (X_i - \bar{X})^2 V(Y_i)}{\Big(\sum (X_i - \bar{X})^2\Big)^2} = \frac{\sigma^2 \cancel{\sum (X_i - \bar{X})^2}}{\Big(\sum (X_i - \bar{X})^2\Big)^{\cancel{2}}} = \frac{\sigma^2}{\sum (X_i - \bar{X})^2}$$

Άρα

$$\hat{\beta} \sim N(\beta, \frac{\sigma^2}{\sum (X_i - \bar{X})^2})$$

(ii) (Να βρεθεί η κατανομή της $\frac{\hat{\beta}-\beta}{s(\hat{\beta})}$ *όπου* $s(\hat{\beta})$ *είναι α.ε. της* $\hat{\beta}$*.)*

Θα δείξουμε πρώτα ότι το $\frac{\sum(Y_i-\hat{\beta}(X_i-\bar{X}))^2}{\sigma^2} \sim X^2_{(n-1)}$

Οι βαθμοί ελευθερίας είναι αριθμός παρατηρήσεων-αριθμός εκτιμούμενων παραμέτρων

$$\sum(Y_i-\beta(X_i-\bar{X}))^2 = \sum(Y_i-\hat{\beta}(X_i-\bar{X})+\hat{\beta}(X_i-\bar{X})-\beta(X_i-\bar{X}))^2 =$$

$$= \sum\left[(\hat{\beta}-\beta)(X_i-\bar{X})+(Y_i-\hat{\beta}(X_i-\bar{X})\right]^2 =$$

$$= (\hat{\beta}-\beta)^2\sum(X_i-\bar{X})^2+\sum\left(Y_i-\hat{\beta}(X_i-\bar{X})\right)^2\Big]^2+2(\hat{\beta}-\beta)\sum(X_i-\bar{X})[Y_i-\hat{\beta}(X_i-\bar{X})] =$$

$$= (\hat{\beta}-\beta)^2\sum(X_i-\bar{X})^2+\sum\left[Y_i-\hat{\beta}(X_i-\bar{X})\right]^2$$

γιατί $\sum(X_i-\bar{X})[Y_i-\hat{\beta}(X_i-\bar{X}) = \sum(X_i-\bar{X})Y_i-\hat{\beta}\sum(X_i-\bar{X})^2$

$$= \sum(X_i-\bar{X})Y_i-\frac{\sum(X_i-\bar{X})Y_i}{\cancel{\sum(X_i-\bar{X})^2}}\cancel{\sum(X_i-\bar{X})^2} = 0$$

Άρα έχουμε

$Q = Q_1+Q_2$*, όπου*

$Q = \sum\left((Y_i-\beta(X_i-\bar{X})\right)^2$ *πραγματικές*

$Q_1 = (\hat{\beta}-\beta)\sum(X_i-\bar{X})^2$ *τετραγωνικές μορφές*

$Q_2 = \sum(Y_i-\beta(X_i-\bar{X}))^2$ *ως προς τις μεταβλητές* $Y_i, (X_i-\bar{X})$

Επειδή $Y_i \sim N(\beta(X_i-\bar{X}),\sigma^2)$

$$\frac{[Y_i-\beta(X_i-\bar{X})]^2}{\sigma^2} \sim X^2_{(1)}$$

$$\Rightarrow \frac{Q}{\sigma^2} = \frac{\sum[Y_i-\beta(X_i-\bar{X})]^2}{\sigma^2} \sim X^2_{(n)}$$

$$\hat{\beta} \sim N(\beta, \frac{\sigma^2}{\sum(X_i-\bar{X})^2}) \Rightarrow \frac{(\hat{\beta}-\beta)\sum(X_i-\bar{X})^2}{\sigma^2} = \frac{Q_1}{\sigma^2} \sim X^2_{(n)}$$

Επειδή Q_2 *είναι μη αρνητικό, από το θεώρημα τετραγωνικών μορφών έχουμε*

$$\frac{Q_2}{\sigma^2} = \frac{\sum(Y_i-\beta(X_i-\bar{X}))^2}{\sigma^2} \sim X^2_{(n-1)}$$

Ξέρουμε ότι αν $W \sim X^2_{(n-1)}$, *τότε* $E(W) = n-1$.
Άρα $E\Big[\dfrac{\sum(Y_i - \beta(X_i - \bar{X}))^2}{\sigma^2}\Big] = n-1 \Rightarrow E\Big[\dfrac{\sum(Y_i - \beta(X_i - \bar{X}))^2}{n-1}\Big] = \sigma^2$.
Άρα η $\hat{\sigma}^2 = \dfrac{\sum(Y_i - \hat{\beta}(X_i - \bar{X}))^2}{n-1}$ *είναι α.ε. του* σ^2
και η $S^2(\hat{\beta}) = \dfrac{\hat{\sigma}^2}{\sum(X_i - \bar{X})^2}$ *α.ε του* $\sigma^2(\beta) = \dfrac{\sigma^2}{\sum(X_i - \bar{X})^2}$

Έχουμε βρεί ότι $\dfrac{\sum[Y_i - \beta(X_i - \bar{X})]^2}{\sigma^2} = \dfrac{(n-1)\hat{\sigma}^2}{\sigma^2} \sim X^2_{(n-1)}$
Άρα

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

και η κατανομή της είναι ανεξάρτητη του $\hat{\beta}$.

Άρα

$$T = \frac{\dfrac{\hat{\beta} - \beta}{\sigma(\hat{\beta})}}{\sqrt{\dfrac{\cancel{(n-1)}S^2(\hat{\beta})}{\sigma^2(\hat{\beta})\cancel{(n-1)}}}} = \frac{\hat{\beta} - \beta}{s(\hat{\beta})} \sim t(n-1)$$

(*iii*) *Ζητάμε Δ.Ε. για το* $\beta(X_0 - \bar{X})$.

Είναι $P\Big(-t_{\frac{\alpha}{2}}(n-1) < \dfrac{\hat{\beta} - \beta}{s(\hat{\beta})} < t_{\frac{\alpha}{2}}(n-1)\Big) = 1 - \alpha$

Άρα το ζητούμενο δ.ε. για το β *με συντελεστή 1-α είναι*

$$\hat{\beta} \pm t_{\frac{\alpha}{2}}(n-1)s(\hat{\beta})$$

οπότε Δ.Ε. για το $\beta(X_0 - \bar{X})$

$$\hat{\beta}(X_0 - \bar{X}) \pm t_{\frac{\alpha}{2}}(n-1)s(\hat{\beta})(X_0 - \bar{X})$$

■

2.9 Παλινδρόμηση και Ανάλυση Διασποράς

Έστω το μοντέλο $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$. Η διασπορά των Y_i μετριέται από το άθροισμα των τετραγώνων των διαφορών $Y_i - \bar{Y}$, δηλαδή από το συνολικό άθροισμα τετραγώνων

$\sum(Y_i - \bar{Y})^2$ Total sum of squares (SST)

Το άθροισμα των τετραγώνων των καταλοίπων είναι

$$\sum \hat{\varepsilon}_i = \sum(Y_i - \hat{Y})^2 \text{ Error sum of squares } (SSE)$$

Ισχύει ότι

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

Θα δείξουμε ότι

$$\sum(Y_i - \bar{Y})^2 = \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{Y})^2$$

Απόδειξη

$$\sum(Y_i - \bar{Y})^2 = \sum\Big[(Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})\Big]^2 = \sum(\Big[(Y_i - \hat{Y}_i)^2 + (\hat{Y}_i - \bar{Y})^2 + 2(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})\Big]$$

$$= \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{Y})^2$$

Γιατί

$$2\sum(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = 2\sum\hat{Y}_i(Y_i - \hat{Y}_i) - 2\bar{Y}\sum(Y_i - \hat{Y}) =$$

$$= 2\sum\hat{Y}_i\hat{\varepsilon}_i - 2\bar{Y}\sum\hat{\varepsilon}_i = 0 \quad \text{Από ιδιότητες καταλοίπων}$$

$\sum(Y_i - \bar{Y})^2$	$=$	$\sum(Y_i - \hat{Y}_i)^2$	$+$	$\sum(\hat{Y}_i - \bar{Y})^2$
SST		SSE		SSR (Regression sum of squares)
Συνολική μεταβλητότητα των Y_i		Μεταβλητότητα που που αποδίδεται στα σφάλματα		Μεταβλητότητα που εξηγείται από την παλινδρόμηση
$n-1$ β.ε.		$n-2$ β.ε.		1 β.ε.
Τα Y_i είναι ελεύθερα Μια δέσμευση από το $\bar{Y}$ $\sum(Y_i - \bar{Y}_i) = 0$		τα Y_i ελεύθερα. Τα $\hat{Y}_i$ Δυο δεσμεύσεις: β_0 και β_1		Υπάρχουν δυο παράμετροι το β_0 και β_1και μια δέσμευση: $\sum(Y_i - \hat{Y}_i) = 0$

Πίνακας Ανάλυσης Διασποράς (Analysis of Variance Table ANOVA)

Πηγή Μεταβλητότητας *Source of Variation*	Άθροισμα Τετραγώνων *SS*	*B.E.* *d.f.*	*Mean Square* *MS*
Παλινδρόμηση	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$
Σφάλματα	$SSE = \sum(Y_i - \hat{Y})^2$	$n-2$	$MSE = \frac{SSE}{n-2}$
Σύνολο	$SST = \sum(Y_i - \bar{Y})^2$	$n-1$	

Θεώρημα 2.4 *(Θεώρημα Cochran). Έστω n παρατηρήσεις $Y_i \sim N(\mu, \sigma^2)$ και $\sum(Y_i - \bar{Y})^2 = Q_1 + Q_2 + \ldots + Q_k$ όπου τα Q_i είναι τετραγωνικές μορφές ως προς τα $Y_1, Y_2, \ldots, Y_n$ με j, $j = 1, \ldots, k$ β.ε. αντίστοιχα. Τότε οι τ.μ. $Q_1, Q_2, \ldots, Q_k$ είναι ανεξάρτητες και $\frac{Q_j}{\sigma^2} \sim X^2_{(r_j)}$, $j = 1, \ldots, k$ αν και μόνο αν $\sum_{j=1}^{k} r_j = n - 1$*

Το θεώρημα μπορεί να χρησιμοποιηθεί για τον έλεγχο υποθέσεων

$$\begin{aligned} H_0 &: \quad \beta_1 = 0 \quad (\text{δεν υπάρχει γραμμική σχέση}) \\ H_1 &: \quad \beta_1 \neq 0 \end{aligned}$$

με τη χρήση του κριτηρίου F.

Κάτω από την $H_0 : Y_i \sim N(\beta_0, \sigma^2)$, δηλαδή τα Y_i είναι ταυτοτικά κατανεμημένα.

Έχουμε $\sum(Y_i - \bar{Y})^2 = \underbrace{\sum(Y_i - \hat{Y}_i)^2}_{Q_1} + \underbrace{\sum(\hat{Y}_i - \bar{Y})^2}_{Q_2}$

Οι υποθέσεις του θεωρήματος Cochran πληρούνται. Άρα οι $\frac{\sum(Y_i - \hat{Y}_i)^2}{\sigma^2}$και $\frac{\sum(\hat{Y}_i - \bar{Y}_i)^2}{\sigma^2}$είναι ανεξάρτητες με $n-2$ και 1 β.ε. αντίστοιχα.

Παρατήρηση

Αν $Y \sim X^2_{(\nu_1)}$, $V \sim X^2_{(\nu_2)}$ και U, V ανεξάρτητες τότε $F = \frac{U/\nu_1}{V/\nu_2} \sim F(\nu_1, \nu_2)$

Άρα η τ.μ.

$$\frac{\frac{\sum(\hat{Y}_i - \bar{Y}_i)^2}{\cancel{\sigma^2}}}{\frac{\sum(Y_i - \hat{Y}_i)^2}{\cancel{\sigma^2}(n-2)}} = \frac{\sum(\hat{Y}_i - \bar{Y}_i)^2}{\hat{\sigma}^2} \sim F_{(1,n-2)}$$

Αλλά

$$\sum(\hat{Y}_i - \bar{Y})^2 = \sum(\hat{\beta}_0 - \hat{\beta}_1 X_i - \bar{Y})^2 = \sum(\bar{Y} - \hat{\beta}_1\bar{X} + \hat{\beta}_1 X_i - \bar{Y})^2$$
$$= \hat{\beta}_1^2 \sum(X_i - \bar{X})^2$$

Δηλαδή

$$F = \frac{\hat{\beta}_1^2 \sum(X_i - \bar{X})^2}{\hat{\sigma}^2} \sim F_{(1,n-2)}$$

Η $H_0 : \beta_1 = 0$ απορρίπτεται για μεγάλες τιμές της στατιστικής συνάρτησης
$F = \frac{\hat{\beta}_1^2 \sum(X_i - \bar{X})^2}{\hat{\sigma}^2}$, δηλ. η $H_0 : \beta_1 = 0$ απορρίπτεται σε επίπεδο στατιστικής σημαντικότητας α
για $\frac{\hat{\beta}_1^2 \sum(X_i - \bar{X})^2}{\hat{\sigma}^2} > F_{\alpha(1,n-2)}$

Σχέση ανάμεσα στον έλεγχο t και στον έλεγχο F

Ο έλεγχος υποθέσεων

$$H_0 : \quad \beta_1 = 0$$
$$H_1 : \quad \beta_1 \neq 0$$

μπορεί να βασιστεί στη στατιστική συνάρτηση $T = \frac{\hat{\beta}_1 - \beta_1}{s(\hat{\beta}_1)} \sim t_{(n-2)}$, όπου
$s(\hat{\beta}_1)^2 = \frac{\hat{\sigma}^2}{\sum(X_i - \bar{X})^2}$.

Κάτω από την H_0 είναι

$$T^2 = \frac{\hat{\beta}_1^2}{s(\hat{\beta}_1)^2} = \frac{\hat{\beta}_1^2}{\frac{\hat{\sigma}^2}{\sum(X_i - \bar{X})^2}} = \frac{\hat{\beta}_1^2 \sum(X_i - \bar{X})^2}{\hat{\sigma}^2} = F$$

Επίσης κάτω από την $H_0 : \frac{\hat{\beta}_1}{s(\hat{\beta}_1)} \sim t_{(n-2)}$. Απορρίπτουμε την $H_0 :$ σε ε.σ.σ. αν $\left|\frac{\hat{\beta}_1}{s(\hat{\beta}_1)}\right| > t_{a/2(n-2)}$

Πίνακας ANOVA για την Πολλαπλή Παλινδρόμηση

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$, $p+1$ παράμετροι.

Πηγή Μεταβλητότητας	SS	$d.f.$	MS	F	$p-value$
Παλινδρόμηση	SSR	p	SSR/p	MSR/MSE	
Σφάλματα	SSE	$n-p-1$	$SSE/n-p-1$		
Σύνολο	SST	$n-1$			

$p-value$: η πιθανότητα μια τ.μ. που ακολουθεί κατανομή $F_{(p,n-p-1)}$ να πάρει τιμή τόσο ακραία ή περισσότερο ακραία από την τιμή της Fπου παρατηρήσαμε.

Έλεγχος με $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ (δεν υπάρχει γραμμική σχέση των μεταβλητων X και Y).

$F = \dfrac{MSR}{MSE} \sim F(p, n-p-1)$. Απορρίπτουμε την H_0 σε επίπεδο στατιστικής σημαντικότητας α αν $F > F_\alpha(p, n-p-1)$ (ή αν $p-value < \alpha$).

Το $F-test$ είναι πιο γενικό από το $t-test$. Ελέγχει την ύπαρξη γραμμικής σχέσης και στην πολλαπλή παλινδρόμηση.

2.9.1 Συντελεστής Προσδιορισμού

Ο συντελεστής προσδιορισμού ορίζεται ως

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Το R^2 εκφράζει το ποσοστό της συνολικής μεταβλητότητας των Y_i που οφείλεται στην παλινδρόμηση. Είναι ένα μέτρο καλής προσαρμογής του μοντέλου (model fit).

Ο προσαρμοσμένος συντελεστής προσδιορισμού ορίζεται ως

$$R^2_{adj} = 1 - \Big(\frac{n-1}{n-p-1}\Big)\frac{SSE}{SST} < R^2$$

Ο R^2_{adj} λαμβάνει υπόψη του και το πλήθος των παρατηρήσεων σε συνδυασμό με το πλήθος των ανεξάρτητων μεταβλήτων (διόρθωση).

Αν προσθέσουμε ανεξάρτητες μεταβλητές στο μοντέλο το R^2 πάντα αυξάνει ενώ το R^2_{adj} όχι απαραίτητα.

2.10 Πολυσυγγραμικότητα

Το πρόβλημα της πολυσυγγραμμικότητας παρουσιάζεται όταν οι ερμηνευτικές μεταβλητές δεν είναι γραμμμικώς ανεξάρτητες.

Έστω το γραμμικό μοντέλο

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \tag{2.2}$$

όπου για κάποια $\lambda_1 \neq \lambda_2 \neq 0$ ισχύει

$$\lambda_1 X_{i1} + \lambda_2 X_{i2} = 0 \Rightarrow X_{i1} = cX_{i2}, c = -\frac{\lambda_2}{\lambda_1}$$

Τότε το μοντέλο (2.2) μπορεί να γραφτεί ως

$$Y_i = \beta_0 + (\beta_1 c + \beta_2) X_{i2} + \varepsilon_i$$

και δεν μπορούμε να ξεχωρίσουμε τη συμβολή κάθε μεταβλητής X_{ij} στην ερμηνεία του Y_i. Πρακτικά, αν εκτιμήσουμε το μοντέλο (2.2), οι εκτιμήσεις δεν είναι στατιστικά σημαντικές και έχουν μεγάλη διασπορά.

Τι κάνουμε?

1. Αφαιρούμε μία από τις συσχετισμένες μεταβλητές από την ανάλυση

2. Αν έχουμε πάρα πολλές ανεξάρτητες μεταβλητές αρκετές από τις οποίες συσχετισμένες εφαρμόζουμε ανάλυση σε κύριες συνιστώσες

Άσκηση 2.4 *(παλιό θέμα)*

Δίνεται το γραμμικό μοντέλο $Y_i = 1 + \beta X_i^2 + \varepsilon_i$, $i = 1, \ldots, n$, $\varepsilon_i \sim N(0, \sigma^2)$, σ^2 *γνωστό.*
(α) Να βρεθεί η εκτιμήτρια ελαχίστων τετραγώνων της β *και η κατανομή της.*
(β) Να κατασκευασθεί 95% διάστημα εμπιστοσύνης για το $\hat{Y}_0 = 1 + \beta X_0^2$ *όταν* X_0 *γνωστή σταθερά,* σ^2 *γνωστό.*
(γ) Πως θα ελέγχατε την

$$H_0 : \quad \beta = \beta_0$$
$$H_1 : \quad \beta \neq \beta_0$$

Λύση

(α) Ελαχιστοποιώ το

$$Q = \sum_{i=1}^{n} (Y_i - 1 - \beta X_i)^2 = \sum_{i=1}^{n} \varepsilon_i^2$$

$$\frac{dQ}{d\beta} = -2 \sum_{i=1}^{n} (Y_i - 1 - \beta X_i^2) X_i^2 = 0 \Rightarrow$$

$$\sum_{i=1}^{n} X_i^2 Y_i - \sum_{i=1}^{n} X_i^2 - \beta \sum_{i=1}^{n} X_i^4 = 0 \Rightarrow \hat{\beta} = \frac{\sum_{i=1}^{n} X_i^2 Y_i - \sum_{i=1}^{n} X_i^2}{\sum_{i=1}^{n} X_i^4}$$

Επειδή $\varepsilon_i \sim N(0, \sigma^2)$ *έχουμε* $Y_i \sim N(1 + \beta X_i^2, \sigma^2)$

Άρα το $\hat{\beta}$ ακολουθεί κανονική κατανομή ως γραμμικός συνδυασμός κανονικών. Έχουμε

$$E(\hat{\beta}) = E\Big[\frac{\sum X_i^2 Y_i - \sum X_i^2}{\sum X_i^4}\Big] = \frac{1}{\sum X_i^4}\Big[\sum X_i^2 E(Y_i) - \sum X_i^2\Big]$$

$$= \frac{1}{\sum X_i^4}\Big[\sum X_i^2(1+\beta X_i^2) - \sum X_i^2\Big] = \frac{1}{\sum X_i^4}\Big[\cancel{\sum X_i^2} + \beta\sum X_i^4 - \cancel{\sum X_i^2}\Big] = \beta$$

$$Var(\hat{\beta}) = Var\Big[\frac{\sum X_i^2 Y_i - \sum X_i^2}{\sum X_i^4}\Big] = \frac{1}{(\sum X_i^4)^2}\Big[\sum X_i^4 Var(Y_i)\Big]$$

$$= \frac{1}{(\sum X_i^4)^2}\sigma^2 \sum X_i^4 = \frac{\sigma^2}{\sum X_i^4}$$

Άρα $\hat{\beta} \sim N(\beta, \frac{\sigma^2}{\sum X_i^4})$

(β) $\hat{Y}_0 = 1 + \beta X_0^2$

Το $\hat{Y}_0$ ακολουθεί κανονική κατανομή ως γραμμική συνάρτηση κανονικής, με

$$E(\hat{Y}_0) = 1 + \beta X_0^2$$

$$Var(\hat{Y}_0) = X_0^4 Var(\beta) = X_0^4 \frac{\sigma^2}{\sum X_i^4}$$

Άρα $\hat{Y}_0 \sim N(1 + \beta X_0^2, \frac{\sigma^2 X_0^4}{\sum X_i^4})$

$$\Rightarrow Z = \frac{\hat{Y}_0 - 1 - \beta X_0^2}{\sqrt{X_0^4\sigma^2/\sum X_i^4}} \sim N(0,1)$$

Δ.Ε. : $Pr(-Z_{a/2} < Z < -Z_{a/2}) = 1 - \alpha$

$$\Rightarrow P\Big(-Z_{a/2} < \frac{\hat{Y}_0 - 1 - \beta X_0^2}{\sqrt{X_0^4\sigma^2}} < -Z_{a/2}\Big)$$

$$\Rightarrow \hat{Y}_0 - Z_{a/2}\sqrt{X_0^4\sigma^2/\sum X_i^4} < \hat{Y}_0 = 1 + \beta X_0^2 < \hat{Y}_0 - Z_{a/2}\sqrt{X_0^4\sigma^2/\sum X_i^4}$$

(γ) Έχουμε δείξει ότι

$$\hat{\beta} \sim N(\beta, \frac{\sigma^2}{\sum X_i^4})$$

$$\frac{\hat{\beta} - \beta}{\sigma^2 / \sum X_i^4} \sim N(0,1)$$

Κάτω από την H_0 $\frac{\hat{\beta} - \beta_0}{\sigma^2 / \sum X_i^4} \sim N(0,1)$

Άρα απορρίπτουμαι την H_0 *αν* $\left| \frac{\hat{\beta} - \beta_0}{\sigma^2 / \sum X_i^4} \right| > Z_{a/2}$ ■

Κεφάλαιο 3

Ανάλυση Διασποράς

3.1 Η σχέση μεταξύ παλινδόμησης και ανάλυσης διακύμανσης

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. Από την άλλη πλευρά, 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.

3.2 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 και μη ελέγξιμοι παράγοντες.

3.3 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}	$\dots$	Y_{1n_1}	$Y_{1.}$	$\bar{Y}_{1.}$
2	Y_{21}	Y_{22}	$\dots$	Y_{2n_2}	$Y_{2.}$	$\bar{Y}_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
m	Y_{m1}	Y_{m2}	$\dots$	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

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

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

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.

3.3.1 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

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

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).

Τα m επίπεδα του υπό μελέτη παράγοντα κινούνται σε $m-1$ βαθμούς ελευθερίας (degrees of freedom). Μέσα σε κάθε επίπεδο οι βαθμοί ελευθερίας είναι n_i-1, άρα συνολικά μέσα στα επίπεδα έχουμε $n=m$ βαθμούς ελευθερίας. Οι συνολικοί βαθμοί ελευθερίας στο πείραμα είναι $n-1$.

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 Χι τετράγωνο 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 Χι τετράγωνο είναι τυχαία μεταβλητή που επίσης ακολουθεί κατανομή Χι τετράγωνο, 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 Χι τετράγωνο 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 Χι τετράγωνο 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

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

Παρατήρηση 3.1 *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 :

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

■

3.3.2 Υποθέσεις του μοντέλου ANOVA

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. Ισοδύναμα : Οι τυχαίοι όροι ϵ_{ij} είναι ταυτοτικά κατανεμημένοι $\epsilon_{ij} \sim N(0, \sigma^2)$ για κάθε i, j.

3. Οι παρατηρήσεις σε κάθε επίπεδο του παράγοντα είναι ανεξάρτητες και ταυτοτικά κατανεμημένες και είναι ανεξάρτητε από τις παρατηρήσεις στα άλλα επίπεδα.

Στόχοι:

1. Έλεγχος αν οι μέσοι των επιπέδων είναι ίσοι

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

2. 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.

3.3.3 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, απορρίπτουμε την H_0 για μεγάλες τιμές του X_0^2, δηλαδή για $X_0^2 > X_{1-\alpha,m-1}^2$.

Παρατήρηση. Οι υποθέσεις του μοντέλου μπορούν να ελεγχθούν γραφικά.
Κατάλοιπα : $\epsilon_{ij} = Y_{ij} - \hat{\mu}_i = Y_{ij} - \overline{Y}_{\cdot i}$.
Γραφικοί έλεγχοι καταλοίπων $\begin{cases} \text{P-P plot για την κανονικότητα} \\ \text{Scatter plot για ομοσκεδαστικότητα και τυχαιότητα–ανεξαρτησία} \end{cases}$

3.3.4 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 ανάλυση μας.

Μια αμερόληπτη εκτιμήτρια του μέσου του i επίπεδου, μ_i είναι

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

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

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

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

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

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

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

και η εκτιμήτρια της είναι

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

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

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

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 $\left| \frac{\overline{D}}{\sqrt{S^2_{\overline{D}}}} \right| > t_{1-\alpha/2,n-m}$.

3.3.5 Contrasts

Ορισμός 3.1 *Contrast λέγεται μια 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 3.1 *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 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 $\hat{L} = \sum_{i=1}^{m} c_i \overline{Y}_{\cdot i}$. Αφού τα $\overline{Y}_{\cdot i}$ είναι ανεξάρτητα η διασπορά του $\hat{L}$ είναι $\sigma^2_{\hat{L}} = \sum_{i=1}^{m} \frac{c_i^2}{n_i}$, 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

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

Επίσης μπορούν να γίνουν οι έλεγχοι

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

απορρίπτουμε την H_0 αν $\left| \frac{\hat{L}}{\sqrt{S^2_{\hat{L}}}} \right| > t_{1-\alpha/2,n-m}$.

Παράδειγμα 3.2 *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, ενώ οι συσκευασίες 2 και 3 δόθηκαν σε τρία καταστήματα. Άλλες παράμετροι που θα μπορούσαν να επηρρεάσουν 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)*	$Y_{.i}$	$\overline{Y}_{.i}$
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*	$Y_{..} = 180$	$\overline{Y}_{..} = 18$

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 (σφάλμα)	$SSE = 46$	$n - m = 6$	$\frac{46}{6} = 7.67 = \frac{SSE}{n-m}$	
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 $F = 11.2 > F_{0.95,3,6} = 4.76$ *απορρίπτουμε την* H_0 *σε επίπεδο εμπιστοσύνης* $\alpha = 5\%$. *Άρα οι μέσοι διαφέρουν για τα διάφορα είδη συσκευασίας. Για το πρώτο είδος συσκευασίας, δηλαδή για το* μ_1, *έχουμε:*

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

Το αντίστοιχο διάστημα εμπιστοσύνης είναι

$$\begin{aligned} \overline{Y}_{.i} - \sqrt{MSE/n_i}\, t_{1-\alpha/2,n-m} \leq \quad & \mu_1 \quad \leq \overline{Y}_{.i} + \sqrt{MSE/n_i}\, t_{1-\alpha/2,n-m} \\ 15 - \sqrt{3.835}\, 2.447 \leq \quad & \mu_1 \quad \leq 15 + \sqrt{3.835}\, 2.447 \\ 10.2 \leq \quad & \mu_1 \quad \leq 19.8 \end{aligned}$$

σε επίπεδο σημαντικότητας $\alpha = 5\%$. *Ομοίως για* μ_2, μ_3, μ_4.

Αν 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*

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

Άρα το διάστημα εμπιστοσύνης δίνεται

$$\begin{aligned}
\hat{L} - \sqrt{S^2_{\hat{L}}}\, t_{0.975,6} \leq &\quad L \quad \leq \hat{L} + \sqrt{S^2_{\hat{L}}}\, t_{0.975,6} \\
-9 - \sqrt{3.196}\, 2.447 \leq &\quad L \quad \leq -9 + \sqrt{3.196}\, 2.447 \\
-13.4 \leq &\quad L \quad \leq -4.6.
\end{aligned}$$

Εφόσον το 0 *δεν περιέχεται στο διάστημα εμπιστοσύνης η διαφορά των πωλήσεων για τα 2 design είναι στατιστικά σημαντική με συντελεστή εμπιστοσύνης* $\alpha = 5\%$.

3.3.6 Παρατηρούμενο Επίπεδο Σημαντικότητας

Observed level of significance ή p–value

Θεωρούμε τον έλεγχο

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

Στατιστική συνάρτηση : $Z_0 = \frac{\overline{X}}{\sqrt{\sigma^2/n}}$
Παρατηρούμενη τιμή : Z_0^*
Κριτική τιμή : $C_{1-\alpha/2}$ σε επίπεδο σημαντικότητας α
Απορρίπτουμε την H_0 αν $|Z_0^*| \geq C_{1-\alpha/2}$.

Ορισμός 3.2 *p–value καλείται η πιθανότητα να πάρει η στατιστική συνάρτηση ελέγχου τιμή τόσο ακραία ή περισσότερο ακραία από αυτήν που παρατηρήσαμε στο δείγμα μας κάτω από την μηδενική υπόθεση.*
Δηλαδή p–value=$Pr[|Z_0| \geq Z_0^*]$.

Για οποιαδήποτε α >p–value η τιμή της στατιστικής συνάρτησης ελέγχου που πήραμε για το συγκεκριμένο δείγμα θα βρίσκεται στην περιοχή απόρριψης του ελέγχου.
Απορρίπτουμε την H_0 αν α >p–value.
Δεν απορρίπτουμε την H_0 αν α <p–value.

3.3.7 Η Μέθοδος της Ελάχιστης Σημαντικής Διαφοράς

The Least Significant Difference (LSD) method

Το LSD test είναι ο πιο δημοφιλής πολλαπλός έλεγχος. 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 του στατιστικού $t = \frac{\overline{Y}_{i.} - \overline{Y}_{j.}}{\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}} \sim t_{n-m}$. Οι μέσοι μ_i, μ_j διαφέρουν στατιστικά σημαντικά σε επίπεδο σημαντικότητας α, αν $|\overline{Y}_{i.} - \overline{Y}_{j.}| > LSD = t_{\alpha/2,n-m}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$ [Ισοδύναμο με το t–test].
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.

3.3.8 Μέθοδος Scheffé

Ο Scheffé πρότεινε μια μέθοδο για τη σύγκριση κάποιων ή όλων των δυνατών contrasts ανάμεσα σε μέσους επιπέδων. Έστω ένα σύνολο από contrasts $L_j = \sum_{i=1}^m c_{ij}\mu_i$, $j = 1, \ldots, r$. Έχουμε $\hat{L}_j = \sum_{i=1}^m c_{ij}\hat{\mu}_i = \sum_{i=1}^m c_{ij}\hat{Y}_{.i} =$ 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 $C = \sqrt{(m-1)F_{1-\alpha,m-1,n-1}}$, τέτοια ώστε με πιαθνότητα 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 τους πολλαπλούς ελέγχους είναι το πολύ α.

1. 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).
2. Η πιθανότητα σφάλματος τύπου I στη μέθοδο Scheffé είναι α αν πάρουμε όλα τα contrasts. Διαφορετικά είναι μικρότερη από α.

3.4 Ανάλυση Διασποράς με Δυο Παράγοντες (two–factor ANOVA)

Έστω ότι το αποτέλεσμα ενός πειράματος εξαρτάται από 2 παράγοντες, A και B. Αρχικά υποθέτουμε ότι δεν υπάρχει αλληλεπίδραση (interaction) μεταξύ των παραγόντων. Θα μελετήσουμε τον παράγοντα A σε m επίπεδα και τον παράγοντα B σε l επίπεδα. Επομένως υπάρχουν $k = m \cdot l$ συνδυασμοί επιπέδων (treatments) για τους οποίους λαμβάνουμε παρατηρήσεις, έστω μια παρατήρηση σε κάθε treatment. Δηλαδή

	B					
A	Y_{11}	Y_{12}	$\cdots$	Y_{1l}	$Y_{1\cdot}$	$\overline{Y}_{1\cdot}$
	Y_{21}	Y_{22}	$\cdots$	Y_{2l}	$Y_{2\cdot}$	$\overline{Y}_{2\cdot}$
			$\vdots$		$\vdots$	$\vdots$
	Y_{m1}	Y_{m2}	$\cdots$	Y_{ml}	$Y_{m\cdot}$	$\overline{Y}_{m\cdot}$
	$Y_{\cdot 1}$	$Y_{\cdot 2}$	$\cdots$	$Y_{\cdot l}$		
	$\overline{Y}_{\cdot 1}$	$\overline{Y}_{\cdot 2}$	$\cdots$	$\overline{Y}_{\cdot l}$		

$$Y_{i\cdot} = \sum_{j=1}^{l} Y_{ij} \qquad Y_{\cdot j} = \sum_{i=1}^{m} Y_{ij} \qquad \overline{Y}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{l} Y_{ij} = \frac{1}{m}\sum_{i=1}^{m}\overline{Y}_{i\cdot} = \frac{1}{l}\sum_{i=1}^{l}\overline{Y}_{\cdot j}\,.$$

$$\overline{Y}_{i\cdot} = \frac{1}{l}Y_{i\cdot} \qquad \overline{Y}_{\cdot j} = \frac{1}{m}Y_{\cdot j}$$

Μοντέλο ANOVA

$$\begin{aligned}
Y_{ij} &= \mu_i + \epsilon_{ij},\ i = 1,\ldots,m,\ j = 1,\ldots,l \\
\epsilon_{ij} &: \text{τυχαία σφάλματα, } \epsilon_{ij} \overset{\text{ανεξ.}}{\sim} N(0,\sigma^2) \\
\mu_{ij} &= \mu_{\cdot\cdot} + \alpha_i + \beta_j \\
\alpha_i &: \text{επίδραση του } i \text{ επιπέδου του παράγοντα A} \\
\beta_j &: \text{επίδραση του } j \text{ επιπέδου του παράγοντα B} \\
&\quad \text{έτσι ώστε } \sum_{i=1}^{m}\alpha_i = 0 \text{ και } \sum_{j=1}^{l}\beta_j = 0
\end{aligned}$$

Δηλαδή, $Y_{ij} \sim N(\mu_{ij},\sigma^2)$.

Υποθέσεις :

1. Κανονικότητα
2. Ομοσκεδαστικότητα
3. Τυχαιότητα/Ανεξαρτησία

Έχουμε $\sum_{i=1}^{m}\alpha_i = 0$ και $\sum_{j=1}^{l}\beta_j = 0$. Αυτό φαίνεται αν θεωρήσουμε $\mu_{ij} = \mu'_{\cdot\cdot} + \alpha'_i + \beta'_j$ και $\overline{\alpha}' = \frac{1}{m}\sum_{i=1}^{m}\alpha'_i = 0$ και $\overline{\beta}' = \frac{1}{l}\sum_{j=1}^{l}\beta'_j = 0$. Τότε

$$\mu_{ij} \;=\; \underbrace{\mu'_{..} + \overline{\alpha}' + \overline{\beta}'}_{\mu_{..}} \;+\; \underbrace{\alpha'_i - \overline{\alpha}'}_{\alpha_i} \;+\; \underbrace{\beta'_i - \overline{\beta}'}_{\beta_j}$$

Προφανώς $\sum_{i=1}^{m} \alpha_i = \sum_{i=1}^{m} (\alpha'_i - \overline{\alpha}') = 0$ και $\sum_{i=j}^{l} \beta_j = \sum_{j=1}^{l} (\beta'_j - \overline{\beta}') = 0$.

3.4.1 Έλεγχοι Υποθέσεων

Για να ελέγξουμε αν υπάρχει επίδραση του παράγοντα A κάνουμε τον έλεγχο

$$\begin{aligned} H_0 &: \mu_{11} = \mu_{21} = \cdots = \mu_{m1} \\ H_1 &: \text{2 τουλάχιστον διαφέρουν} \end{aligned}$$

Ισοδύναμα

$$\begin{aligned} H_0 &: \alpha_1 = \alpha_2 = \cdots = \alpha_m \\ H_1 &: \alpha_i \neq \alpha_j \text{ για τουλάχιστον ένα ζεύγος } i, j \end{aligned}$$

Αντίστοιχα για να ελέγξουμε αν υπάρχει επίδραση του παράγοντα B

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \cdots = \beta_m \\ H_1 &: \beta_i \neq \beta_j \text{ για τουλάχιστον ένα ζεύγος } i, j \end{aligned}$$

Οι παραπάνω έλεγχοι μπορούν να πραγματοποιηθούν με ανάλυση διασποράς και έλεγχο F σε αναλογία με την ANOVA κατά ένα παράγοντα.
Συνολικό άθροισμα :

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

μπορεί να αποδειχθεί ότι όλα τα αθροίσματα διπλασίων γινομένων που σχηματίζονται από $(\overline{Y}_{i.} - \overline{Y}_{..})$, $(\overline{Y}_{.j} - \overline{Y}_{..})$ και $(Y_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y}_{..})$ είναι 0. Επομένως

$$SST \;=\; \underbrace{\sum_{i=1}^{m}\sum_{j=1}^{l} (\overline{Y}_{i.} - \overline{Y}_{..})^2}_{SSA} \;+\; \underbrace{\sum_{i=1}^{m}\sum_{j=1}^{l} (\overline{Y}_{.j} - \overline{Y}_{..})^2}_{SSB} \;+\; \underbrace{\sum_{i=1}^{m}\sum_{j=1}^{l} (Y_{ij} - \overline{Y}_{i.} - \overline{Y}_{.j} + \overline{Y}_{..})^2}_{SSE}$$

SSA **:** άθροισμα τετραγώνων μεταξύ επιπέδων του παράγοντα A

SSB **:** άθροισμα τετραγώνων μεταξύ επιπέδων του παράγοντα B

SSE **:** άθροισμα τετραγώνων των τυχαίων σφαλμάτων.

Απόδειξη για $\sum_{i=1}^{m}\sum_{j=1}^{l}(\overline{Y}_{\cdot j}-\overline{Y}_{\cdot\cdot})(Y_{ij}-\overline{Y}_{i\cdot}-\overline{Y}_{\cdot j}+\overline{Y}_{\cdot\cdot})$

$$\begin{aligned}\sum_{i=1}^{m}\sum_{j=1}^{l}(\overline{Y}_{\cdot j}-\overline{Y}_{\cdot\cdot})(Y_{ij}-\overline{Y}_{i\cdot}-\overline{Y}_{\cdot j}+\overline{Y}_{\cdot\cdot}) &= \sum_{j=1}^{l}(\overline{Y}_{\cdot j}-\overline{Y}_{\cdot\cdot})\sum_{i=1}^{m}(Y_{ij}-\overline{Y}_{i\cdot}-\overline{Y}_{\cdot j}+\overline{Y}_{\cdot\cdot})\\ &= \sum_{j=1}^{l}(\overline{Y}_{\cdot j}-\overline{Y}_{\cdot\cdot})\big(\underbrace{\sum_{i=1}^{m}(Y_{ij}-\overline{Y}_{\cdot j})}_{0}-\underbrace{(\sum_{i=1}^{m}\overline{Y}_{i\cdot}-m\overline{Y}_{\cdot\cdot})}_{0}\big)\\ &= 0\end{aligned}$$

■

Σε πλήρη αντιστοιχία με την ANOVA κατά έναν παράγοντα οι συνολικοί βαθμοί ελευθερίας είναι $ml-1$, οι βαθμοί ελευθερίας του παράγοντα A είναι $m-1$ και του παράγοντα B $l-1$. Τέλος οι βαθμοί ελευθερίας των σφαλμάτων είναι $ml-\cancel{1}-m+\cancel{1}-l+1=m(l-1)-(l-1)=(m-1)(l-1)$. Το $\frac{SSE}{\sigma^2}\sim X^2_{(m-1)(l-1)}$.

Κάτω από την H_0 **:** $\alpha_1=\alpha_2=\cdots=\alpha_m=0$ έχουμε $\frac{SSA}{\sigma^2}\sim X^2_{m-1}$. Άρα απορρίπτουμε την H_0 αν $F_A=\frac{SSA/m-1}{SSE/(m-1)(l-1)}>F_{1-\alpha,m-1,(m-1)(l-1)}$.

Κάτω από την H_0 **:** $\beta_1=\beta_2=\cdots=\beta_m=0$ έχουμε $\frac{SSB}{\sigma^2}\sim X^2_{l-1}$. Άρα απορρίπτουμε την H_0 αν $F_B=\frac{SSB/l-1}{SSE/(m-1)(l-1)}>F_{1-\alpha,l-1,(m-1)(l-1)}$.

Πίνακας ANOVA

Πηγή μεταβλητότητας	SS	df	MS	$F=\frac{MSF}{MSE}$	p–value
Παράγοντας A	SSA	$m-1$	$MSA=\frac{SSA}{m-1}$	$F_A=\frac{MSA}{MSE}$	$\star$
Παράγοντας B	SSB	$l-1$	$MSB=\frac{SSB}{l-1}$	$F_B=\frac{MSB}{MSE}$	
Σφάλμα(Error)	SSE	$(m-1)(l-1)$	$MSE=\frac{SSE}{(m-1)(l-1)}$		
Σύνολο(Total)	SST	$ml-1$			

Παράδειγμα 3.3 *Μας ενδιαφέρει να μελετήσουμε την κατανάλωση βενζίνης από αυτοκίνητα μιας συγκεκριμένης κατηγορίας. Έχουμε 3 μάρκες αυτοκινήτων και 4 διαφορετικούς τύπους βενζίνης. Ο αριθμός των χιλιομέτρων (km) ανά γαλόνι βενζίνης για καθέναν από τους συνδυασμούς είναι*

	τύπος βενζίνης				
μάρκα αυτοκινήτου	*1*	*2*	*3*	*4*	$\overline{Y}_{i\cdot}$
1	*16*	*18*	*21*	*21*	*19*
2	*14*	*15*	*18*	*17*	*16*
3	*15*	*15*	*18*	*16*	*16*
$\overline{Y}_{\cdot j}$	*15*	*16*	*19*	*18*	$\overline{Y}_{\cdot\cdot}$=*17*

Μοντέλο $Y_{ij} = \mu + \alpha_i + \beta_j$, $i = 1,2,3$, $j = 1,2,3,4$. Μας ενδιαφέρει να ελέγξουμε κατά πόσο η μέση κατανάλωση βενζίνης διαφέρει για τους διάφορους τύπους βενζίνης σε επίπεδο στατιστικής σημαντικότητας $\alpha = 1\%$. Δηλαδή, θέλουμε να ελέγξουμε

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \cdots = \beta_m \\ vs & \\ H_1 &: \beta \neq \beta_j \text{ για τουλάχιστον ένα ζεύγος } i,j \end{aligned}$$

Πίνακας ANOVA

Πηγή μεταβλητότητας	SS	df	MS	$F = \frac{MSF}{MSE}$	
Παράγοντας A	$SSA = 24$	$m-1 = 2$	$MSA = \frac{24}{2} = 12$	$F_A = \frac{12}{2/3} = 18$	
Παράγοντας B	$SSB = 30$	$l-1 = 3$	$MSB = \frac{30}{3} = 10$	$F_B = \frac{10}{2/3} = 15$	
Σφάλμα(Error)	$SSE = 4$	$(m-1)(l-1) = 6$	$MSE = \frac{4}{6} = \frac{2}{3}$		
Σύνολο(Total)	$SST = 58$	$ml - 1 = 11$			

Σε $\alpha = 1\%$ απορίπτουμε την H_0 εαν $F_B > F_{0.99,3,6} \Rightarrow F_B = 15 > F_{0.99,3,6} = 9.78$, άρα απορρίπτουμε την H_0. Δηλαδή η μέση κατανάλωση βενζίνης διαφέρει για τους διάφορους τύπους βενζίνης, στις 3 μάρκες αυτοκινήτων που εξετάστηκαν.

3.4.2 ANOVA κατά δυο παράγοντες με αλληλεπίδραση

Οι δυο παράγοντες που εξετάζουμε μπορεί να αλληλεπιδρούν μεταξύ τους. Για να ελέγξουμε για πιθανή αλληλεπίδραση (interaction) εξετάζουμε δείγμα μεγάθους c για κάθε συνδυασμό των δυο παραγόντων (treatment). Έχουμε δηλαδή παρατηρήσεις $Y_{ijr} \sim N(\mu_{ij}, \sigma^2)$, $i = 1,\ldots,m$, $j = 1,\ldots,l$ και $r = 1,\ldots,c$.

Μοντέλο ANOVA

$$Y_{ij} = \mu_{..} + \alpha_i + \beta_j + \gamma_{ij}$$

έτσι ώστε $\sum_{i=1}^{m} \alpha_i = 0$ και $\sum_{j=1}^{l} \beta_j = 0$ και $\sum_{i=1}^{m} \gamma_{ij} = 0$, $j = 1,\ldots,l$ και $\sum_{j=1}^{l} \gamma_{ij} = 0$, $i = 1,\ldots,m$.

$$\overline{Y}_{ij\cdot} = \frac{1}{c}\sum_{r=1}^{c} Y_{ijr} \qquad \overline{Y}_{i\cdot r} = \frac{1}{l}\sum_{j=l}^{c} Y_{ijr} \qquad \overline{Y}_{\cdot jr} = \frac{1}{m}\sum_{i=l}^{m} Y_{ijr}$$

$$\overline{Y}_{i\cdot\cdot} = \frac{1}{lc}\sum_{j=1}^{l}\sum_{r=1}^{c} Y_{ijr} \qquad \overline{Y}_{\cdot j\cdot} = \frac{1}{mc}\sum_{i=1}^{m}\sum_{r=1}^{c} Y_{ijr} \qquad \overline{Y}_{\cdot\cdot r} = \frac{1}{ml}\sum_{i=1}^{m}\sum_{j=1}^{l} Y_{ijr}$$

$$\overline{Y}_{\cdots} = \frac{1}{mlc}\sum_{i=1}^{m}\sum_{j=1}^{l}\sum_{r=1}^{c} Y_{ijr}$$

Συνολική μεταβλητότητα :

$$\begin{aligned} SST &= \sum_{i=1}^{m}\sum_{j=1}^{l}\sum_{r=1}^{c}(Y_{ijr}-\overline{Y}_{...})^2 \\ &= \sum_{i=1}^{m}\sum_{j=1}^{l}\sum_{r=1}^{c}\left((\overline{Y}_{i..}-\overline{Y}_{...})+(\overline{Y}_{.j.}-\overline{Y}_{...})+(Y_{ij.}-\overline{Y}_{i..}-\overline{Y}_{.j.}+\overline{Y}_{...})+(\overline{Y}_{ijr}-\overline{Y}_{ij.})\right)^2 \\ &= \sum_{i=1}^{m} lc(\overline{Y}_{i..}-\overline{Y}_{...})^2+\sum_{j=1}^{l} mc(\overline{Y}_{.j.}-\overline{Y}_{...})^2+\sum_{i=1}^{m}\sum_{j=1}^{l} c(Y_{ij.}-\overline{Y}_{i..}-\overline{Y}_{.j.}+\overline{Y}_{...})^2 \\ &\quad +\sum_{i=1}^{m}\sum_{j=1}^{l}\sum_{r=1}^{c}(\overline{Y}_{ijr}-\overline{Y}_{ij.})^2\,, \end{aligned}$$

Ορίζουμε

$SSA = \sum_{i=1}^{m} lc(\overline{Y}_{i..}-\overline{Y}_{...})^2$

SSA : άθροισμα τετραγώνων μεταξύ επιπέδων του παράγοντα Α

$SSB = \sum_{j=1}^{l} mc(\overline{Y}_{.j.}-\overline{Y}_{...})^2$

SSB : άθροισμα τετραγώνων μεταξύ επιπέδων του παράγοντα Β

$SS(AB) = \sum_{i=1}^{m}\sum_{j=1}^{l} c(Y_{ij.}-\overline{Y}_{i..}-\overline{Y}_{.j.}+\overline{Y}_{...})^2$

$SS(AB)$: άθροισμα τετραγώνων της αλληλεπίδρασης

$SSE = \sum_{i=1}^{m}\sum_{j=1}^{l}\sum_{r=1}^{c}(\overline{Y}_{ijr}-\overline{Y}_{ij.})^2$

SSE : άθροισμα τετραγώνων των τυχαίων σφαλμάτων.

Επομένως

$$SST = SSA + SSB + SS(AB) + SSE$$

Πίνακας ANOVA

Πηγή μεταβλητότητας	SS	df	MS	$F=\frac{MSF}{MSE}$	
Παράγοντας Α	SSA	$m-1$	$MSA=\frac{SSA}{m-1}$	$F_A=\frac{MSA}{MSE}$	
Παράγοντας Β	SSB	$l-1$	$MSB=\frac{SSB}{l-1}$	$F_B=\frac{MSB}{MSE}$	
Παράγοντας ΑΒ	$SS(AB)$	$(m-1)(l-1)$	$MS(AB)=\frac{SS(AB)}{(m-1)(l-1)}$	$F_{AB}=\frac{MS(AB)}{MSE}$	
Σφάλμα(Error)	SSE	$ml(c-1)$	$MSE=\frac{SSE}{ml(c-1)}$		
Σύνολο(Total)	SST	$mlc-1$			

F–test για interaction

Θεωρούμε τον έλεγχο

$$\begin{aligned} H_0 &: \gamma_{ij}=0 \; i=1,\ldots,m,\; j=1,\ldots,l \\ H_1 &: \gamma_{ij}\neq 0 \text{ για ένα τουλαχιστον } i,j \end{aligned}$$

Απορρίπτουμε την H_0 αν $F_{AB} = \frac{MS(AB)}{MSE} > F_{1-\alpha,(m-1)(l-1),ml(c-1)}$. Ελέγχουμε πρώτα την ύπαρξη στατιστικά σημαντικής αλληλεπίδρασης και έπειτα την στατιστική σημαντικότητα των επιμέρους παραγόντων.

Άσκηση 3.1 *(παλιό θέμα). Ένας κατασκευαστής τηλεοράσεων ενδιαφέρεται να ελέγξει αν η αγωγιμότητα του περιβλήματος τεσσάρων διαφορετικών τύπων καλωδιών τηλεόρασης διαφέρει. Μετρήσεις για την αγωγιμότητα των περιβλημάτων καλωδιών δίνονται στον παρακάτω πίνακα.*

Τύπος καλωδίου i	*Αγωγιμότητα Y_{ij}*	$Y_{i\cdot}$	$\overline{Y}_{i\cdot}$
1	*143 141 150 146*	*580*	*145*
2	*152 149 137 143*	*581*	*145.25*
3	*144 146 142 137*	*569*	*142.25*
4	*129 127 132 129*	*517*	*129.25*
		$Y_{\cdot\cdot} = 2247$	$\overline{Y}_{\cdot\cdot} = 140.43$

1. *Να ελεγχθεί σε $\alpha = 5\%$ αν υπάρχει διαφορά στην αγωγιμότητα του περιβλήματος των τεσσάρων τύπων καλωδιών.*

2. *Να κατασκευαστεί διάστημα εμπιστοσύνης σ.ε. 95% για τον μέσο του τέταρτου τύπου καλωδίου.*

3. *Ο κατασκευαστής αποφασίζει να χρησιμοποιήσει τον τέταρτο τύπο καλωδίου γιατί υποστηρίζει ότι δίνει τη μικρότερη αγωγιμότητα περιβλήματος σε σύγκριση με τους άλλους τρεις τύπους. Να ελεγχθεί σε $\alpha = 5\%$ αν ισχύει ο ισχυρισμός του κατασκευαστή.*

   Δίνονται $F_{0.95,3,12} = 3.49$, $t_{0.975,12} = 2.179$ και $t_{0.95,12} = 1.782$.

Λύση:

1.

$$\begin{aligned} SST &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{\cdot\cdot})^2 = 925.9375 \\ SSF &= \sum_{i=1}^{m} n_i(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2 = 689.6875 \end{aligned}$$

Άρα $SSE = SST - SSF = 236.25$.

Πίνακας ANOVA

Πηγή διασποράς	SS	df	MS	$F = \frac{MSF}{MSE}$
F	$SSF = 689.6875$	$m - 1 = 3$	$229.8958 = \frac{SSF}{m-1}$	11.6772
E	$SSE = 236.25$	$n - m = 12$	$19.6875 = \frac{SSE}{n-m}$	
T	$SST = 925.9375$	$n - 1 = 15$		

Θεωρούμε τον έλεγχο

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

Εφόσον $F = 11.6772 > F_{0.95,3,12} = 3.49$ *απορρίπτουμε την* H_0 *σε επίπεδο εμπιστοσύνης* $\alpha = 5\%$. *Άρα οι μέσοι διαφέρουν.*

2. *Για το τέρτατο είδος, δηλαδή για το* μ_4, *έχουμε:*

$$\begin{aligned} \overline{Y}_{4\cdot} &= 129.25 \\ S^2_{\overline{Y}_{4\cdot}} &= \frac{MSE}{n_4} = 4.92219 \end{aligned}$$

$$Pr[-t_{0.975,12} \leq \frac{\overline{Y}_{4\cdot} - \mu_4}{\sqrt{S^2_{\overline{Y}_{4\cdot}}}} \leq t_{0.975,12}] = 0.95$$

Το αντίστοιχο διάστημα εμπιστοσύνης είναι

$$\begin{aligned} \overline{Y}_{4\cdot} - \sqrt{S^2_{\overline{Y}_{4\cdot}}}\, t_{0.975,12} \leq &\quad \mu_4 \quad \leq \overline{Y}_{4\cdot} + \sqrt{S^2_{\overline{Y}_{4\cdot}}}\, t_{0.975,12} \\ 129.25 - \sqrt{4.9219}\, 2.179 \leq &\quad \mu_4 \quad \leq 129.25 + \sqrt{4.9219}\, 2.179 \\ 124.4159 \leq &\quad \mu_4 \quad \leq 134.0841 \end{aligned}$$

σε επίπεδο εμπιστοσύνης 95%.

3. *Θεωρούμε το contrast* $L = \frac{\mu_1+\mu_2+\mu_3}{3} - \mu_4$. *Έχουμε*

$$\begin{aligned} \hat{L} &= \frac{\overline{Y}_{1\cdot} + \overline{Y}_{2\cdot} + \overline{Y}_{3\cdot}}{3} - \overline{Y}_{4\cdot} = 14.9167 \\ \sum_{i=1}^{4} \frac{c_i^2}{n_i} &= \frac{1}{4}\big[\frac{1}{9} + \frac{1}{9} + \frac{1}{9} + 1\big] = \frac{1}{3} \\ S^2_{\hat{L}} &= MSE\big[\sum_{i=1}^{4} \frac{c_i^2}{n_i}\big] = 6.5625 \end{aligned}$$

Άρα το διάστημα εμπιστοσύνης δίνεται

$$\begin{aligned} \hat{L} - \sqrt{S^2_{\hat{L}}}\, t_{0.975,12} \leq &\quad L \quad \leq \hat{L} + \sqrt{S^2_{\hat{L}}}\, t_{0.975,12} \\ 14.9167 - 2.5617 \cdot 2.179 \leq &\quad L \quad \leq 14.9167 + 2.5617 \cdot 2.179 \\ 9.3348 \leq &\quad L \quad \leq 20.4986. \end{aligned}$$

Εφόσον το 0 *δεν περιέχεται στο διάστημα εμπιστοσύνης η διαφορά του τέταρτου τύπου καλωδίου από τους άλλους τρείς είναι στατιστικά σημαντική σε* $\alpha = 5\%$. *Συγκεκριμένα είναι* $\mu_4 < \frac{\mu_1+\mu_2+\mu_3}{3}$, *δηλαδή ισχύει ο ισχυρισμός του κατασκευαστή σε* $\alpha = 5\%$. *Εφόσον από τον έλεγχο*

$$\begin{aligned} H_0 &: L = 0 \\ H_1 &: L > 0 \end{aligned}$$

έπεται ότι απορρίπτουμε την H_0 *μιας και* $\frac{\hat{L}}{\sqrt{S^2_{\hat{L}}}} = 5.8230 > 1.782 = t_{0.975,12}$.

Κεφάλαιο 4

Απαραμετρική Στατιστική

4.1 Κριτήριο X^2

Έστω $\underline{Y} = (Y_1, \ldots, Y_k)$ με $\underline{Y} \sim Multinomial(n, \underline{p})$ όπου $\underline{p} = (p_1, \ldots, p_k)$, $0 < p_i < 1,\ i = 1, \ldots, k$ και $\sum_{i=1}^{k} p_i = 1$. Η ποσότητα

$$Q_{k-1} = \sum_{i=1}^{k} \frac{(y_i - np_i)^2}{np_i} \sim X^2_{k-1}.$$

(Βασίζεται στο γεγονός ότι η $\underline{Y}$ ακολουθεί ασυμπτωτικά πολυδιάστατη κανονική κατανομή.)

Έστω ότι σε ένα πείραμα εμφανίζονται k ξένα μεταξύ τους ενδεχόμενα $A_1, \ldots, A_k$. Θέλουμε να ελέγξουμε την

$$H_0 : p_i = p_{i0},\ i = 1, \ldots, k \qquad H_1 : p_i \neq p_{i0},\ \text{για ένα τουλάχιστον } i,$$

όπου p_i η πιθανότητα εμφάνισης του ενδεχομένου A_i. Έστω ότι επαναλαμβάνουμε το πείραμα n φορές και ότι το ενδεχόμενο A_i εμφανίζεται y_i φορές, $i = 1, \ldots, k$. Κάτω από την H_0 τα παρατηρούμενα y_i και ο αναμενόμενος αριθμός εμφανίσεων των A_i, δηλαδή np_{i0}, είναι κοντά. Συνεπώς η τιμή της ελεγχοσυνάρτησης Q_{k-1} πρέπει να είναι μικρή αν ισχύει η H_0, δηλαδή η ποσότητα $q_{k-1} = \sum_{i=1}^{k} \frac{(y_i - np_{i0})^2}{np_{i0}}$ πρέπει να έχει μικρή τιμή. Έχουμε $Q_{k-1} \sim X^2_{k-1}$. Άρα απορρίπτουμε την H_0 σε επίπεδο στατιστικής σημαντικότητας α αν $q_{k-1} > X^2_{k-1,1-\alpha}$.

Παράδειγμα 4.1 *Κανονικοποίηση της βαθμολογίας εξετάσεων*

Υποθέτουμε ότι οι βαθμοί των φοιτητών σε μια εξέταση ακολουθούν $N(\mu, \sigma^2)$ κατανομή. Χωρίζω το διάστημα [0,10] σε 5 υποδιαστήματα ($k = 5$) και καθορίζω εκ των προτέρων πόσοι φοιτητές θα πάρουν 1 ή 2, 3 ή 4, 5 ή 6, 7 ή 8 και 9 ή 10. Για παράδειγμα, κάθε βαθμός στο διάστημα $(\mu + k_2\sigma, \mu + k_1\sigma)$ θα γίνει 7 ή 8. Η επιλογή των $k_1,\ k_2,\ k_3,\ k_4$ είναι αυθαίρετη, έστω $k_1 = \frac{3}{2},\ k_2 = \frac{1}{2},\ k_3 = -\frac{1}{2},\ k_4 = -\frac{3}{2}$.

Κανονικοποίηση: $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$.

$P(\text{βαθμός 9 ή 10}) = P(X > \mu + k_1\sigma) = P(\frac{X-\mu}{\sigma} > k_1) = P(Z > k_1) = P(Z > \frac{3}{2}) = 1 - \Phi(1.5) = 0.0668$

$P(\text{βαθμός 1 ή 2}) = P(X < \mu + k_4\sigma) = P(\frac{X-\mu}{\sigma} < k_4) = P(Z < k_4) = P(Z < -\frac{3}{2}) = \Phi(-1.5) = 1 - \Phi(1.5) = 0.0668$

Άρα το ποσοστό των φοιτητών που θα πάρουν 9 ή 10 είναι 6.68%, καθώς και το ποσοστό των φοιτητών που θα πάρουν 1 ή 2. Ομοίως η αναλογία των 7 ή 8 είναι 0.2417 καθώς και των 3 ή 4. Άρα η αναλογία των 5 ή 6 είναι 1-2·(0.0668)-2·(0.2417)=0.3830.

Άσκηση 4.1 *Έστω ότι σε μία τάξη με* $n = 487$ *φοιτητές δόθηκαν οι βαθμοί*

βαθμοί A_i	*9-10*	*7-8*	*5-6*	*3-4*	*1-2*
παρατηρούμενες συχνότητες	*48*	*110*	*219*	*85*	*25*

Να ελεγχθεί σε επίπεδο σημαντικότητας $\alpha = 5\%$, *με χρήση του κριτηρίου* X^2, *αν έγινε η παραπάνω κανονικοποίηση.*

Λύση:

Θέλουμε να ελέγξουμε

$$\begin{array}{lllllll} H_0: & p_1 & = & P(\text{βαθμός 9 ή 10}) & = & 0.0668 & = & p_{1,0} \\ & p_2 & = & P(\text{βαθμός 7 ή 8}) & = & 0.2417 & = & p_{2,0} \\ & p_3 & = & P(\text{βαθμός 5 ή 6}) & = & 0.3830 & = & p_{3,0} \\ & p_4 & = & P(\text{βαθμός 3 ή 4}) & = & 0.2417 & = & p_{4,0} \\ & p_5 & = & P(\text{βαθμός 1 ή 2}) & = & 0.0668 & = & p_{5,0} \\ H_1: & \text{κάθε δυνατή εναλλακτική} & & & & & & \end{array}$$

Κριτήριο ελέγχου (ελεγχοσυνάρτηση): $Q_{k-1} = \sum_{i=1}^{k} \frac{(y_i - np_{i0})^2}{np_{i0}}$, *όπου* $k = 5$.

βαθμοί A_i	*9 ή 10*	*7 ή 8*	*5 ή 6*	*3 ή 4*	*1 ή 2*
παρατηρούμενες συχνότητες y_i	*48*	*110*	*219*	*85*	*25*
αναμενόμενες συχνότητες np_{i0}	$487 \cdot 0.0668$ *=32.5*	$487 \cdot 0.2417$ *=117.7*	$487 \cdot 0.3830$ *=136.6*	$487 \cdot 0.2417$ *=117.7*	$487 \cdot 0.0668$ *=32.5*

Παρατηρούμενη τιμή κριτηρίου:

$$q_4 = \sum_{i=1}^{5} \frac{(y_i - np_{i0})^2}{np_{i0}} = \left(\frac{(48-32.5)^2}{32.5}\right) + \cdots + \left(\frac{(25-32.5)^2}{32.5}\right) = 24.35$$

Έχουμε $q_4 = 24.35 > 9.488 = X^2_{0.95,4}$. *Άρα απορρίπτουμε την* H_0. ■

Παράδειγμα 4.2 *Έστω* w *το αποτέλεσμα ενός τυχαίου πειράματος,* $F_W(w) = P(W \leq w)$ *η αντίστοιχη αθροιστική συνάρτηση κατανομής και*

$$F_0(w) = \begin{cases} 0, & w < -1 \\ \frac{w^3+1}{2}, & -1 \leq w \leq 1 \\ 1, & w \geq 1. \end{cases}$$

Θέλουμε να ελέγξουμε την

$$H_0 : \ F(w) = F_0(w) \qquad H_1 : \ \text{κάθε δυνατή εναλλακτική.}$$

Λύση:

Χωρίζουμε το διάστημα [0,1] σε 10 υποδιαστήματα με τα σημεία $b_i = \frac{i}{10}$, $i = 0, 1, \ldots, 10$. *Ορίζουμε* $a_i = F_0^{-1}(b_i)$, $i = 0, 1, \ldots, 10$ *οπότε έχουμε* $F_0(a_i) = b_i$. *Αλλά* $F_0(w) = \frac{w^3+1}{2}$, $-1 \leq w \leq 1$. *Έτσι* $F_0(a_i) = \frac{a_i^3+1}{2} = b_i$, *δηλαδή* $a_i = (2b_i - 1)^{\frac{1}{3}}$. *Συνεπώς κάτω από την* H_0 *καθένα από τα σύνολα* $A_1 = [-1, a_1]$, $A_i = (a_{i-1}, a_i]$, $i = 2, \ldots, k-1$, $A_k = (a_9, 1]$ *έχει πιθανότητα* $\frac{1}{10}$. *Δηλαδή* $p_{i0} = \frac{1}{10}$, $i = 1, \ldots, k$. *Από το δείγμα μεγέθους* n *από την κατανομή* $F_W(w)$ *υπολογίζουμε τον αριθμό των παρατηρήσεων σε κάθε διάστημα* A_i, *έστω* y_i. *Έχουμε* $Q_{k-1} = \sum_{i=1}^{k} \frac{(Y_i - np_{i0})^2}{np_{i0}} \sim X^2_{k-1}$ *κάτω από την* H_0. *Απορρίπτουμε την* H_0 *σε επίπεδο στατιστικής σημαντικότητας* α *αν* $q_{k-1} = \sum_{i=1}^{k} \frac{(y_i - np_{i0})^2}{np_{i0}} > X^2_{1-\alpha,k-1}$, *όπου* np_{i0} *ο αναμενομένος παρατηρήσεων στο διάστημα* A_i *κάτω από την* H_0. ■

Παρατήρηση 4.1 *Το κριτήριο* X^2 *μπορεί να χρησιμοποιηθεί για να γίνει ο έλεγχος της υπόθεσης ότι η κατανομή μιας τ.μ. είναι μια συγκεκριμένη κατανομή (π.χ. Διωνυμική*(n,p)*, Εκθετική*(θ)*, Κανονική* (μ, σ^2) *κτλ). Αντικαθιστούμε την αρχική υπόθεση (π.χ.* $X \sim$ *Εκθετικ*(θ)*) με την υπόθεση* $H_0 : P(X \in A_i) = p_{i0}$, $i = 1, \ldots, k$. *Στην επιλογή των* $A_1, A_2, \ldots, A_k$ *υπάρχει υποκειμενικότητα, αλλά η ασυμπτωτική κατανομή του κριτηρίου κάτω από την* H_0 *είναι* X^2_{k-1}. *Δηλαδή η κατανομή του* Q_{k-1} *είναι ελεύθερη από τα* $p_{10}, p_{20}, \ldots, p_{k0}$ *και συνεπώς από την καθορισμένη από την* H_0 *κατανομή του* X. *Δηλαδή έχουμε ένα μη παραμετρικό κριτήριο(non-parametric) ή distribution free κριτήριο.*

Μειονεκτήματα κριτηρίου

i. Ο έλεγχος έχει πολλές εναλλακτικές H_1: κάθε δυνατή εναλλακτική. Ωστόσο μπορεί να χρησιμοποιηθεί σαν αρχικό γενικό κριτήριο και μετά μπορούν να χρησιμοποιηθούν άλλα περισσότερο εξειδικευμένα κριτήρια.

ii. Για μικρό p_{i0}, η αναμενόμενη συχνότητα np_{i0} είναι μικρή και ο όρος $\frac{(y_i - np_{i0})^2}{np_{i0}}$ είναι μεγάλος. Έτσι η τιμή του κριτηρίου $q_{k-1} = \sum_{i=1}^{k} \frac{(y_i - np_{i0})^2}{np_{i0}}$ είναι πολύ μεγάλη. Συνήθως χρησιμοποιούμε τον περιορισμό $np_i > 5$ ή $np_i > 1$. Αν κάποια αναμενόμενη συχνότητα έχει $np_i < 1$, τότε 'ενώνουμε' τα γειτονικά κελιά έτσι ώστε να ισχύει $np_i > 1$ για όλα τα i. Προφανώς μειώνονται και οι βαθμοί ελευθερίας. Η επιλογή των κελιών που ομαδοποιούνται είναι αυθαίρετη και φροντίζουμε να περιορίζεται στο ελάχιστο δυνατό.

Παράδειγμα 4.3 *Τα ακόλουθα δεδομένα παριστάνουν τον αριθμό αφίξεων πελατών σε ένα κατάστημα για ένα χρονικό διάστημα παρατήρησης 121 ωρών.*

Αριθμός αφίξεων ανά ώρα (κελιά)	*Αριθμός ωρών (παρατηρούμενες συχνότητες)*
0	*10*
1	*31*
2	*40*
3	*20*
4	*10*
5	*4*
≥ 6	*6*

Αν X είναι ο αριθμός των αφίξεων ανά ώρα, να εξεταστεί η υπόθεση ότι οι αφίξεις ακολουθούν Poisson(λ=2) σε επίπεδο σημαντικότητας $\alpha = 5\%$.

Λύση:

Θέλουμε να ελέγξουμε

$$H_0: \ X \sim Poisson(\lambda = 2) \qquad H_1: \ \text{κάθε δυνατή εναλλακτική.}$$

Κάτω από την H_0 έχουμε ότι $P(X = x) = e^{-2}\frac{2^x}{x!}$. Έτσι ο έλεγχος μπορεί να γραφτεί

$$H_0 : p_i = p_{i0} = P(X = i) = e^{-2}\,\frac{2^i}{i!}, \ \text{για όλα τα } i \qquad H_1: \ \text{κάθε δυνατή εναλλακτική.}$$

Δηλαδή

$$\begin{aligned} H_0: \quad & p_0 = p_{00} = e^{-2} = 0.135 \\ & p_1 = p_{10} = 2e^{-2} = 0.271 \\ & p_2 = p_{20} = 0.271 \\ & p_3 = p_{30} = 0.180 \\ & p_4 = p_{40} = 0.09 \\ & p_5 = p_{50} = 0.038 \\ & p_{>6} = p_{>60} = 1 - \textstyle\sum_{i=0}^{5} p_{i0} = 0.017 \end{aligned}$$

(Τα παραπάνω υπολογίστηκαν με απλή επαναληπτική μέθοδο υπολογισμού πιθανοτήτων Poisson. Δηλαδή $p_k = e^{-\lambda}\frac{\lambda^k}{k!}$ και $p_{k+1} = e^{-\lambda}\frac{\lambda^{k+1}}{(k+1)!}$ οπότε διαιρώντας κατά μέλη παίρνουμε $\frac{p_{k+1}}{p_k} = \frac{\lambda}{k+1}$ ή $p_{k+1} = \frac{\lambda}{k+1}p_k$ με $p_0 = e^{-\lambda}$.)

i	0	1	2	3	4	5	6
y_i	10	31	40	20	10	4	6
np_{i0}	$121 \cdot 0.135 = 16.376$	32.751	32.751	21.834	10.917	4.367	2.000

$$q_{k-1} = q_5 = \sum_{i=1}^{6} \frac{(y_i - np_{i0})^2}{np_{i0}} = 12.4087 > 12.5916 = X^2_{5,0.95}.$$

Άρα δεν απορρίπτουμε την H_0 *σε* $\alpha = 5\%$. ■

Παράδειγμα 4.4 *Στο προηγούμενο παράδειγμα να εξετασθεί η υπόθεση ότι οι αφίξεις ακολουθούν την κατανομή Poisson.*

Λύση:
Τώρα

$$H_0 : P(X = i) = e^{-\lambda}\frac{\lambda^i}{i!},\ i = 1, \ldots, k\ (\lambda > 0) \qquad H_1 :\ \textit{Κάθε δυνατή εναλλακτική.}$$

Έχουμε τώρα $p_{i0} = p_i(\lambda)$ *και άρα* $Q_{k-1}(\lambda) = \sum_{i=1}^{k} \frac{(Y_i - np_i(\lambda))^2}{np_i(\lambda)}$. *Τα* p_{i0} *εξαρτώνται από την άγνωστη παράμετρο* λ *η οποία πρέπει να εκτιμηθεί. Ποια είναι η κατανομή του* Q_{k-1}*;* ■

Γενικά έστω ότι μας ενδιαφέρει να ελέγξουμε την

$$H_0 :\ X \sim \text{γνωστή κατανομή με παράμετρο } \theta(\tau\text{-διάστατη}) \qquad H_1 : \text{κάθε δυνατή εναλλακτική.}$$

Τα p_{i0} και άρα το $Q_{k-1}(\theta) = \sum_{i=1}^{k} \frac{(Y_i - np_i(\theta))^2}{np_i(\theta)}$ εξαρτώνται από την παράμετρο θ. Θέλουμε να ελέγξουμε αν για κάποια τιμή του θ τα $p_i(\theta),\ i = 1, \ldots, k$ δίνουν μια καλή προσαρμογή στα δεδομένα. Θέλουμε να ελαχιστοποιήσουμε το $Q_{k-1}(\theta)$.

Θεώρημα 4.1 *Έστω* $\hat{\theta}$ *η εκτιμήτρια του* θ *που ελαχιστοποιεί το* $Q_{k-1}(\theta)$. *Κάτω από ορισμένες συνθήκες και για μεγάλο* n*, η ελεγχοσυνάρτηση* $Q(\hat{\theta}) = \sum_{i=1}^{k} \frac{(Y_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})}$ *ακολουθεί ασυμπτωτικά* $X^2_{k-1-\tau}$*, όπου* τ *η διάσταση του* θ.

Πρόταση 4.1 *Ισχύει ότι για μεγάλο* n *το* $\hat{\theta}$ *είναι κατά προσέγγιση η εκτιμήτρια μέγιστης πιθανοφάνειας.*

Για την άσκησή μας 4.4 (κατανομή Poisson) η ε.μ.π. είναι $\hat{\lambda} = \overline{X}$. Οπότε με βάση τα δεδομένα παίρνουμε $\hat{\lambda} = \frac{0\cdot 10 + 1\cdot 31 + \cdots + 6\cdot 6}{121} = 2.207$

$$H_0: \quad p_0 = p_{00} = e^{-2.207}$$
$$p_1 = p_{10} = 2.207e^{-2.207}$$
$$\vdots$$

Συνεχίζουμε τους υπολογισμούς όπως και στην προηγούμενη άσκηση. Η τιμή του κριτηρίου υπολογίζεται μέσω της $q_{k-1-\tau} = q_5 = \sum_{i=1}^{k} \frac{(Y_i - np_i(\hat{\lambda}))^2}{np_i(\hat{\lambda})}$ με $Q_5 \sim X_5^2$ και $X^2_{5,0.95} = 11.0705$. ■

4.2 Πίνακες Συνάφειας

Οι πίνακες συνάφειας είναι πίνακες ταξινόμησης δύο ή περισσότερων διαστάσεων με τα κελιά τους να παριστούν συχνότητες εμφάνισης συγκεκριμένων χαρακτηριστικών.

		Σ	Τ	Η	Λ	Η	
		1	2	$\cdots$	j	$\cdots$	k	Σύνολα στηλών
Γ	1	y_{11}	y_{12}	$\cdots$	y_{1j}	$\cdots$	y_{1k}	$n_{1\cdot}$
Ρ	2	y_{21}	y_{22}	$\cdots$	y_{2j}	$\cdots$	y_{2k}	$n_{2\cdot}$
Α	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
Μ	i	y_{i1}	y_{i2}	$\cdots$	y_{ij}	$\cdots$	y_{ik}	$n_{i\cdot}$
Μ	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
Η	l	y_{l1}	y_{l2}	$\cdots$	y_{lj}	$\cdots$	y_{lk}	$n_{l\cdot}$
Σύνολα γραμμών		$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot k}$	n

y_{ij}: η παρατηρούμενη συχνότητα στο κελί ij.

Χρήσεις των πινάκων συνάφειας (δύο διαστάσεις)

1) Οι γραμμές του πίνακα αντιστοιχούν σε l ανεξάρτητα δείγματα ενώ οι στήλες του πίνακα αντιστοιχούν σε k κατηγορίες. Μας ενδιαφέρει να ελέγξουμε αν οι πιθανότητες εμφάνισης των κατηγοριών διαφέρουν για τα διάφορα δείγματα.

2) Έχουμε ένα δείγμα ταξινομημένο σε k κατηγορίες ως προς ένα χαρακτηριστικό Α και σε l κατηγορίες ως προς ένα χαρακτηριστικό Β. Το κελί ij του πίνακα συνάφειας αντιστοιχεί στο συνδυασμό της i κατηγορίας του χαρακτηριστικού Β και της j κατηγορίας του χαρακτηριστικού Α. Μας ενδιαφέρει να ελέγξομε αν οι πιθανότητες εμφάνισης των κατηγοριών του ενός χαρακτηριστικού επηρεάζονται από τις κατηγορίες του άλλου.

4.2.1 Ο έλεγχος X^2 για την ύπαρξη διαφορών στις πιθανότητες εμφάνισης k κατηγοριών σε l ανεξάρτητα δείγματα

Έστω l αμοιβαία ανεξάρτητα τυχαία δείγματα μεγέθους $n_{1\cdot}$, $n_{2\cdot}$,..., $n_{l\cdot}$ ταξινομημένα σε k κατηγορίες (δηλαδή τα $n_{i\cdot}$ είναι γνωστά εκ των προτέρων). Έστω y_{ij} ο αριθμός των παρατηρήσεων του i δείγματος που ανήκουν στην κατηγορία j, $i = 1, \ldots, l$, $j = 1, \ldots, k$. Έχουμε $n_{i\cdot} = \sum_{j=1}^{k} y_{ij},\ i = 1, \ldots, l$ και $n_{\cdot j} = \sum_{i=1}^{l} y_{ij},\ j = 1, \ldots, k$ καθώς και $n = \sum_{i=1}^{l}\sum_{j=1}^{k} y_{ij} = \sum_{i=1}^{l} n_{i\cdot} = \sum_{j=1}^{k} n_{\cdot j}$. Κάθε παρατήρηση μπορεί να ανήκει σε μία ακριβώς από τις k κατηγορίες. Έστω p_{ij} η πιθανότητα μία τυχαία επιλεγόμενη παρατήρηση από τον i πληθυσμό, $i = 1, \ldots, l$ να ανήκει στην κατηγορία $j,\ j = 1, \ldots, k$. Θέλουμε να ελέγξουμε:

$$H_0 : p_{1j} = p_{2j} = \cdots = p_{lj},\ j = 1, \ldots, k$$

$$H_1 : \text{Κάθε δυνατή εναλλακτική.}$$

Κάτω από την H_0, η αναμενόμενη συχνότητα του κελιού ij είναι

$$E_{ij} = n_{i\cdot} \cdot \left(\frac{n_{\cdot j}}{n}\right),$$

όπου $\frac{n_{\cdot j}}{n}$ είναι το ποσοστό των συνολικών παρατηρήσεων που ανήκουν στην j κατηγορία, δηλαδή η πιθανότητα εμφάνισης της j κατηγορίας κάτω από την H_0.

Στατιστική Συνάρτηση: $Q_{(k-1)(l-1)} = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(y_{ij}-E_{ij})^2}{E_{ij}}$

Ασυμπτωτικά $Q_{(k-1)(l-1)} \sim X^2_{(k-1)(l-1)}$. Επομένως απορρίπτουμε την H_0 σε επίπεδο σημαντικότητας α αν $q_{(k-1)(l-1)} > X^2_{(k-1)(l-1),1-\alpha}$.

4.2.2 Ο έλεγχος ανεξαρτησίας X^2

Έστω ένα τυχαίο δείγμα μεγέθους n ταξινομημένο σε k κατηγορίες ως προς ένα χαρακτηριστικό Α και σε l κατηγορίες ως προς ένα χαρακτηριστικό Β και y_{ij} :ο αριθμός των παρατηρήσεων που ανήκουν στην i κατηγορία του χαρακτηριστικού Β και στη j κατηγορία του χαρακτηριστικού Α, $i = 1, \ldots, l$, $j = 1, \ldots, k$. Κάθε παρατήρηση μπορεί να ανήκει σε μία ακριβώς από τις k κατηγορίες του Α και σε μία ακριβώς από τις l κατηγορίες του Β.

Παρατήρηση 4.2 *Τα σύνολα των γραμμών $n_{i\cdot}$, $i = 1, \ldots, l$ δεν είναι τώρα γνωστά εκ των προτέρων.*

$n_{i\cdot} = \sum_{j=1}^{k} y_{ij},\ i = 1, \ldots, l$ και $n_{\cdot j} = \sum_{i=1}^{l} y_{ij},\ j = 1, \ldots, k$ καθώς και $n = \sum_{i=1}^{l}\sum_{j=1}^{k} y_{ij} = \sum_{i=1}^{l} n_{i\cdot} = \sum_{j=1}^{k} n_{\cdot j}$.

p_{ij} :η πιθανότητα μια τυχαία επιλεγόμενη παρατήρηση από τον πληθυσμό να ανήκει στο κελί ij, $(p_{ij} = P(Y \text{ανήκει στο κελί } ij))$.

$p_{i\cdot}$:η πιθανότητα μια τυχαία επιλεγόμενη παρατήρηση από τον πληθυσμό να ανήκει στη γραμμή i, $(p_{i\cdot} = P(Y \text{ανήκει στη γραμμή } i))$.

$p_{\cdot j}$:η πιθανότητα μια τυχαία επιλεγόμενη παρατήρηση από τον πληθυσμό να ανήκει στη στήλη j, $(p_{\cdot j} = P(Y \text{ανήκει στη στήλη } j))$.

Θέλουμε να ελέγξουμε:

$$H_0 : p_{ij} = p_{i\cdot}p_{\cdot j},\ i = 1, \ldots, l,\ j = 1, \ldots, k(\text{Ανεξαρτησία } A\text{και } B)$$

$$H_1 : \text{Κάθε δυνατή εναλλακτική.}$$

Κάτω από την H_0, η αναμενόμενη συχνότητα του κελιού ij είναι

$$E_{ij} = n\left(\frac{n_{i\cdot}}{n}\right)\left(\frac{n_{\cdot j}}{n}\right) = \frac{n_{i\cdot}n_{\cdot j}}{n},$$

όπου

$\frac{n_{i\cdot}}{n}$: το ποσοστό των παρατηρήσεων που ανήκουν στην i γραμμή, και

$\frac{n_{\cdot j}}{n}$: το ποσοστό των παρατηρήσεων που ανήκουν στη j στήλη.

Στατιστική Συνάρτηση Ελέγχου: $Q_{(k-1)(l-1)} = \sum_{i=1}^{l}\sum_{j=1}^{k}\frac{(y_{ij}-E_{ij})^2}{E_{ij}} \sim X^2_{(k-1)(l-1)}$

Απορρίπτουμε την H_0 σε επίπεδο στατιστικής σημαντικότητας α αν

$$q_{(k-1)(l-1)} > X^2_{(k-1)(l-1),1-\alpha}.$$

Παράδειγμα 4.5 *Ο παρακάτω πίνακας συνάφειας παρουσιάζει δεδομένα που αφορούν 141 ασθενείς με όγκο στον εγκέφαλο οι οποίοι ταξινομήθηκαν με βάση τον τύπο και τη θέση του όγκου στον εγκέφαλό τους.*

		Τύπος όγκου			
		A	*B*	*Γ*	*Σύνολα*
Θέση	*I*	*23*	*9*	*6*	*38*
	II	*21*	*4*	*3*	*28*
	III	*34*	*24*	*17*	*75*
Σύνολα		*78*	*37*	*26*	*141*

Να ελεγχθεί αν ο τύπος και το μέγεθος του όγκου είναι ανεξάρτητα, με χρήση του ελέγχου ανεξαρτησίας X^2 σε επίπεδο σημαντικότητας $\alpha = 5\%$.

Λύση:

Κάτω από την H_0 έχουμε

$$\begin{aligned} E_{11} &= \frac{n_{1\cdot}n_{\cdot 1}}{n} = \frac{38\cdot 78}{141} = 21.02 \\ E_{12} &= \frac{n_{1\cdot}n_{\cdot 2}}{n} = \frac{38\cdot 37}{141} = 9.97 \\ &\vdots \quad\quad \vdots \\ E_{33} &= \frac{n_{3\cdot}n_{\cdot 3}}{n} = \frac{75\cdot 26}{141} = 13.83 \end{aligned}$$

$$\begin{aligned} q_{(3-1)(3-1)} = q_4 &= \frac{(23-21.02)^2}{21.02} + \frac{(9-9.97)^2}{9.97} + \cdots + \frac{(17-13.83)^2}{13.83} \\ &= 0.19 + 0.09 + \cdots + 0.72 = 7.84 < 9.49 = X^2_{4,0.95} \end{aligned}$$

Επομένως δεν απορρίπτουμε την H_0 σε $\alpha = 5\%$ και άρα δεν μπορούμε να ισχυριστούμε ότι η θέση και το μέγεθος του όγκου είναι ανεξάρτητα, με βάση αυτά τα δεδομένα. ■

Παράδειγμα 4.6 *Σε δύο ανεξάρτητα τυχαία δείγματα 20 ατόμων ζητήθηκε να δοκιμάσουν μια νέα μάρκα απορρυπαντικού και να το συγκρίνουν με αυτό που χρησιμοποιούσαν. Οι απαντήσεις τους συνοψίζονται στον παρακάτω πίνακα.*

	καμιά διαφορά	*χειρότερο*	*καλύτερο*	*Σύνολο*
Δείγμα 1	*4*	*7*	*9*	*20*
Δείγμα 2	*1*	*6*	*13*	*20*
Σύνολο	*5*	*13*	*22*	*40*

Δείχνουν τα αποτελέσματα στατιστικά σημαντική διαφορά στις απαντήσεις στα δύο δείγματα; ($\alpha = 5\%$)

Λύση:
Θέλουμε να ελέγξουμε

$$H_0 : p_{1j} = p_{2j},\ j = 1,2,3 \qquad H_1 : p_{1j} \neq p_{2j},\ \text{για ένα τουλάχιστον } j.$$

Κάτω από την H_0, η αναμενόμενη συχνότητα του κελιού ij ($E_{ij} = \frac{n_{i\cdot}n_{\cdot j}}{n}$) είναι

	καμιά διαφορά	*χειρότερο*	*καλύτερο*
Δείγμα 1	*2.5*	*6.5*	*11*
Δείγμα 2	*2.5*	*6.5*	*11*

$$\begin{aligned} q_{(3-1)(2-1)} = q_2 &= \frac{(4-2.5)^2}{2.5} + \cdots + \frac{(13-11)^2}{11} \\ &= 2.604 < 5.991 = X^2_{2,0.95}. \end{aligned}$$

Άρα δεν απορρίπτουμε την H_0 σε επίπεδο σημαντικότητας $\alpha = 5\%$. Επομένως (σε $\alpha = 5\%$) δεν φαίνεται να διαφέρουν τα δύο δείγματα. ■

Παρατήρηση 4.3 *Δύο αναμενόμενες συχνότητες είναι $2.5 < 5$. Μπορούμε να επαναλάβουμε τον έλεγχο ενώνοντας τις κατηγορίες 'καμιά διαφορά' και 'χειρότερο'.*

4.2.3 Σύγκριση l ανεξάρτητων πολυωνυμικών κατανομών

Πολυωνυμική κατανομή: Έστω ένα πείραμα με k δυνατά αποτελέσματα με αντίστοιχες πιθανότητες εμφάνισης $p_1,\ p_2, \ldots, p_k$. Επαναλαμβάνουμε το πείραμα n ανεξάρτητες φορές και έστω

$$\begin{array}{lll} X_1 & \text{ο αριθμός εμφανίσεων του αποτελέσματος} & A_1 \\ X_2 & \text{ο αριθμός εμφανίσεων του αποτελέσματος} & A_2 \\ \vdots & \vdots & \vdots \\ X_k & \text{ο αριθμός εμφανίσεων του αποτελέσματος} & A_k \end{array}$$

με $\sum_{j=1}^{k} X_j = n$. Το διάνυσμα $\underline{X} = (X_1, \ldots, X_k)$ ακολουθεί πολυωνυμική κατανομή

$$P(\underline{X} = \underline{x}) = \frac{n!}{x_1! x_2! \cdots x_k!}\, p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k},$$

όπου $\sum_{j=1}^{k} p_j = 1,\ 0 \le p_j \le 1,\ j = 1, \ldots, k$.

Έστω l ανεξάρτητες πολυωνυμικές κατανομές με πιθανότητες $p_{i1},\ p_{i2}, \ldots,\ p_{ik},\ i = 1, \ldots, l$. Θέλουμε να ελέγξουμε:

$$H_0 :\ p_{1j} = p_{2j} = \cdots = p_{lj} = p_j,\ j = 1, 2, \ldots, k$$

$$H_1 :\ \text{κάθε δυνατή εναλλακτική.}$$

Έστω Y_{ij} ο αριθμός εμφανίσεων του A_j στο δείγμα μεγέθους n_i από την i πολυωνυμική κατανομή, $i = 1, \ldots, l,\ j = 1, \ldots, k$.

Εκτίμηση παραμέτρων: Οι άγνωστες παράμετροι p_j εκτιμώνται ως

$$\hat{p_j} = \frac{\sum_{i=1}^{l} Y_{ij}}{\sum_{i=1}^{l} n_i} = \frac{1}{n} \sum_{i=1}^{l} Y_{ij}.$$

Κριτήριο Ελέγχου: $Q = \sum_{i=1}^{l} \sum_{j=1}^{k} \frac{(Y_{ij} - n_i p_{ij})^2}{n_i p_{ij}}$

Κάτω από την H_0: $Q = \sum_{i=1}^{l} \sum_{j=1}^{k} \frac{(Y_{ij} - n_i \hat{p_j})^2}{n_i \hat{p_j}} \sim X^2_{(l(k-1)-(k-1))} \equiv X^2_{(l-1)(k-1)}$

Απορρίπτω την H_0 αν

$$q_{(l-1)(k-1)} > X^2_{(l-1)(k-1),1-\alpha}$$

Παράδειγμα 4.7 *Θέλουμε να ελέγξουμε δύο διαφορετικές μεθόδους διδασκαλίας. Σχηματίζουμε τυχαία δύο τμήματα από 50 φοιτητές και σε κάθε τμήμα εφαρμόζουμε διαφορετική μέθοδο στη διδασκαλία ενός συγκεκριμένου μαθήματος. Στο τέλος του εξαμήνου οι φοιτητές πήραν τους βαθμούς:*

	A	*B*	*C*	*D*	*F*	*Σύνολο*
Τμήμα 1	*8*	*13*	*16*	*10*	*3*	*50*
Τμήμα 2	*4*	*9*	*14*	*16*	*7*	*50*
Σύνολο	*12*	*22*	*30*	*26*	*10*	*100*

Να βρεθεί αν υπάρχει διαφορά στην απόδοση των φοιτητών στα δύο τμήματα ($\alpha = 5\%$).

Λύση:
($k = 5, l = 2, n_1 = n_2 = 50$) *Θέλουμε να ελέγξουμε*

$$\begin{array}{lccccc} H_0: & p_{1A} & = & p_{2A} & = & p_A \\ & p_{1B} & = & p_{2B} & = & p_B \\ & p_{1C} & = & p_{2C} & = & p_C \\ & p_{1D} & = & p_{2D} & = & p_D \\ & p_{1F} & = & p_{2F} & = & p_F \\ H_1: & \text{κάθε} & \text{δυνατή} & \text{εναλλακτική} & & \end{array}$$

Εκτίμηση πιθανοτήτων:

$$\begin{aligned} \hat{p_A} = \frac{8+4}{50+50} = 0.12 &\rightarrow n_1\hat{p_A} = n_2\hat{p_A} = 6 \\ \hat{p_B} = \frac{13+9}{50+50} = 0.22 &\rightarrow n_1\hat{p_B} = n_2\hat{p_B} = 11 \\ \hat{p_C} = \frac{30}{100} = 0.30 &\rightarrow n_1\hat{p_C} = n_2\hat{p_C} = 15 \\ \hat{p_D} = \frac{26}{100} = 0.26 &\rightarrow n_1\hat{p_D} = n_2\hat{p_D} = 13 \\ \hat{p_F} = \frac{10}{100} = 0.10 &\rightarrow n_1\hat{p_F} = n_2\hat{p_F} = 5 \end{aligned}$$

Παρατηρούμενη τιμή κριτηρίου:

$$\begin{aligned} q_{(l-1)(k-1)} = q_4 &= \frac{(8-6)^2}{6} + \frac{(13-11)^2}{11} + \cdots + \frac{(3-5)^2}{5} + \\ &+ \frac{(4-6)^2}{6} + \frac{(9-11)^2}{11} + \cdots + \frac{(7-5)^2}{5} = 5.18 < 9.49 = X^2_{4,0.95} \end{aligned}$$

Άρα δεν απορρίπτουμε την H_0. *Επομένως δεν έχω αρκετά στοιχεία για να ισχυριστώ ότι οι δύο μέθοδοι διδασκαλίας διαφέρουν σε επίπεδο στατιστικής σημαντικότητας* $\alpha = 5\%$. ■

4.3 Σύγκριση δύο άγνωστων κατανομών

Έστω τυχαίες μεταβλητές U και V με αντίστοιχες αθροιστικές $F(u)$ και $G(v)$. Θέλουμε να ελέγξουμε αν $F(x) = G(x),\ \forall x$.

Χωρίζουμε την πραγματική ευθεία σε k ξένα μεταξύ τους σύνολα $A_1,\ A_2, \ldots,\ A_k$. Έστω ότι

$$p_{1j} = P(U \in A_j),\ j = 1, \ldots, k, \quad \text{και}$$

$$p_{2j} = P(V \in A_j),\ j = 1, \ldots, k.$$

Αν $F(x) = G(x),\ \forall x$, τότε $p_{1j} = p_{2j} = p_j,\ j = 1, \ldots, k$. Δηλαδή, μπορούμε να ανάγουμε τον έλεγχό μας σε έλεγχο δύο ανεξάρτητων πολυωνυμικών.

Έστω δείγμα μεγέθους n_1 από την τ.μ. U και δείγμα μεγέθους n_2 από την τ.μ. V.

Y_{1j} : ο αριθμός των παρατηρήσεων της U στο σύνολο A_j

Y_{2j} : ο αριθμός των παρατηρήσεων της V στο σύνολο A_j.

Εκτιμώ:

$$\hat{p_j} = \frac{Y_{1j} + Y_{2j}}{n_1 + n_2},\ j = 1, \ldots, k$$

$$Q_{(k-1)} = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(Y_{ij} - n_i \hat{p_j})^2}{n_i \hat{p_j}} \sim X^2_{k-1}.$$

-Ερώτημα: Πώς διαλέγουμε τα $A_1, \ldots, A_k$;

Πρακτικά διαιρούμε το εύρος του συνολικού δείγματος σε k ίσα μέρη.

Παράδειγμα 4.8 *Από δύο γνωστές εταιρίες αυτοκινήτων εκλέγονται τυχαία 20 αυτοκίνητα και ελέγχεται η αντοχή τους με διάφορα κριτήρια. Οι χρόνοι μέχρι να παρουσιαστεί πρόβλημα ήταν:*

Μάρκα U										*Μάρκα V*									
25	*31*	*20*	*42*	*39*	*19*	*35*	*36*	*44*	*26*	*28*	*17*	*33*	*25*	*31*	*21*	*16*	*19*	*31*	*27*
38	*31*	*29*	*41*	*43*	*36*	*28*	*31*	*25*	*38*	*23*	*19*	*25*	*22*	*29*	*32*	*24*	*20*	*34*	*28*

Να ελεγχθεί η υπόθεση ότι η κατανομή του χρόνου λειτουργίας μέχρι να παρουσιαστεί πρόβλημα είναι η ίδια και στις δύο μάρκες αυτοκινήτων ($\alpha = 5\%$).

Λύση:
Διαιρώ το συνολικό δείγμα μεγέθους $n = n_1 + n_2 = 40$ *σε* $k = 7$ *ίσα μέρη. Το συνολικό δείγμα παίρνει τιμές στο* $[16, 44]$, *άρα το εύρος είναι* $R = X_{max} - X_{min} = 44 - 16 = 28$. *Άρα, αν έχω* $k = 7$ *μέρη, το πλάτος κάθε διαστήματος θα είναι 28/7=4. Έτσι* $A_1 = (16, 20]$, $A_2 = (20, 24]$, $A_3 = (24, 28]$, $A_4 = (28, 32]$, $A_5 = (32, 36]$, $A_6 = (36, 40]$ *και* $A_7 = (40, 44]$.

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	*Σύνολο*
U	*1*	*1*	*3*	*5*	*1*	*6*	*3*	*20*
V	*4*	*4*	*5*	*4*	*3*	*0*	*0*	*20*
Σύνολο	*5*	*5*	*8*	*9*	*4*	*6*	*3*	*40*

Έστω $p_{1j} = P(U \in A_j)$ *και* $p_{2j} = P(V \in A_j)$, $j = 1, \ldots, 7$. *Τότε*

$$H_0 : \ p_{1j} = p_{2j} = \hat{p_j}, \ j = 1, \ldots, 7$$

όπου $\hat{p_j} = \frac{Y_{1j}+Y_{2j}}{n_1+n_2}, \ j = 1, \ldots, 7$

$$Q_{(7-1)} = Q_6 = \sum_{i=1}^{2} \sum_{j=1}^{7} \frac{(Y_{ij} - n_i \hat{p_j})^2}{n_i \hat{p_j}} \sim X^2_6.$$

$$\begin{aligned} \hat{p_1} = \frac{5}{40} \quad &\rightarrow \quad n_1\hat{p_1} = n_2\hat{p_2} = 20\ \frac{5}{40} = 2.5 \\ \hat{p_2} = \frac{5}{40} \quad &\rightarrow \quad n_1\hat{p_1} = n_2\hat{p_2} = 20\ \frac{5}{40} = 2.5 \\ \vdots \qquad &\qquad \vdots \end{aligned}$$

$$\begin{aligned} q_6 \ &= \ \frac{(1-2.5)^2}{2.5} + \frac{(1-2.5)^2}{2.5} + \cdots + \frac{(3-1.5)^2}{1.5} \\ &+ \ \frac{(4-2.5)^2}{2.5} + \frac{(4-2.5)^2}{2.5} + \cdots + \frac{(0-1.5)^2}{1.5} = 4 < 12.59 = X^2_{6,0.95} \end{aligned}$$

Άρα δεν απορρίπτουμε την H_0 *σε επίπεδο σημαντικότητας* $\alpha = 5\%$. ■

Παράδειγμα 4.9 *Μέλη Δ.Ε.Π. του τμήματος Μαθηματικών εξέφρασαν την άποψη ότι ο αριθμός των φοιτητών οι οποίοι προτιμούν μαθήματα προσφερόμενα από τον τομέα Ανάλυσης, Στατιστικής και Άλγεβρας εκφράζεται από την αναλογία* $p^2 : 2pq : q^2$, *όπου* $p+q = 1$. *Σε ένα τυχαίο δείγμα 116 φοιτητών οι αντίστοιχες συχνότητες ήταν 42, 52 και 22 αντίστοιχα. Ευσταθεί ο ισχυρισμός ότι σε επίπεδο σημαντικότητας* $\alpha = 5\%$ *οι παρατηρήσεις αυτές ταιριάζουν με το θεωρητικό μοντέλο αν i.* $p = 1/2$ *και ii.* p *άγνωστο.*

Λύση:

i.

$$\begin{aligned} H_0 \ : \ p_1 = p_{10} &= \frac{1}{4} \qquad H_1 : \ \textit{κάθε δυνατή εναλλακτική} \\ p_2 = p_{20} &= \frac{1}{2} \\ p_3 = p_{30} &= \frac{1}{4} \end{aligned}$$

Αναμενόμενες συχνότητες κάτω από την H_0:

$$\begin{aligned} np_{10} &= 116 \cdot \frac{1}{4} = 29 \\ np_{20} &= 116 \cdot \frac{1}{2} = 58 \\ np_{30} &= 116 \cdot \frac{1}{4} = 29 \end{aligned}$$

Έλεγχος X^2. Παρατηρούμενη τιμή ελεγχοσυνάρτησης:

$$q_{k-1} = q_2 = \sum_{i=1}^{3} \frac{(y_i - np_{i0})^2}{np_{i0}} = \frac{(42-29)^2}{29} + \frac{(52-58)^2}{58} + \frac{(22-29)^2}{29} = 8.1 > 5.99 = X^2_{2,0.95}.$$

Άρα απορρίπτουμε την H_0 σε επίπεδο σημαντικότητας $\alpha = 5\%$.

ii.

$$\begin{aligned} H_0 \ : \ & p_1 = p^2 \qquad\qquad H_1 : \text{ κάθε δυνατή εναλλακτική} \\ & p_2 = 2p(1-p) \\ & p_3 = (1-p)^2 \end{aligned}$$

Επειδή το p είναι άγνωστο πρέπει να εκτιμηθεί. Θα χρησιμοποιήσουμε τη μέθοδο της μέγιστης πιθανοφάνειας. Έχουμε ότι $\underline{Y} = (Y_1, Y_2, Y_3)$ ακολουθεί πολυωνυμική(τριωνυμική) κατανομή:

$$\begin{aligned} P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3) &= \frac{n!}{y_1!\, y_2!\, y_3!}\, p_1^{y_1} p_2^{y_2} p_3^{y_3} \\ &= \frac{n!}{y_1!\, y_2!\, (n-y_1-y_2)!}\, (p^2)^{y_1} (2p(1-p))^{y_2} ((1-p)^2)^{n-y_1-y_2} \end{aligned}$$

Πιθανοφάνεια: $L(p) \propto p^{2y_1+y_2}(1-p)^{2n-2y_1-y_2}$

$\ln L(p) = (2y_1 + y_2)\ln p + (2n - 2y_1 - y_2)\ln(1-p) + c$

$\frac{\partial \ln L(p)}{\partial p} = \frac{2y_1+y_2}{p} - \frac{2n-2y_1-y_2}{1-p} = 0 \Rightarrow p = \frac{2y_1+y_2}{2n}$.

Άρα $\hat{p} = \frac{2y_1+y_2}{2n} = \frac{2\cdot 42+52}{2\cdot 116} = \frac{68}{116}$. *Επιπλέον* $\hat{p_{10}} = \hat{p}^2$, $\hat{p_{20}} = 2\hat{p}(1-\hat{p})$ *και* $\hat{p_{30}} = (1-\hat{p})^2$.

$$q_{k-1-1} = q_1 = \sum_{i=1}^{3} \frac{(y_i - n\hat{p_{i0}})^2}{n\hat{p_{i0}}} = \frac{(42 - 39.86)^2}{68} + \ldots\ldots\ldots\ldots$$

$X^2_{1,0.95} = 3.84$

■

4.4 Έλεγχοι Υποθέσεων βασισμένοι στη Διωνυμική Κατανομή

4.4.1 Ο Διωνυμικός Έλεγχος

Έστω δείγμα μεγέθους n ανεξάρτητων δοκιμών Bernoulli και έστω Y ο αριθμός των 'επιτυχιών' και $n - Y$ ο αριθμός των 'αποτυχιών' στο δείγμα. Υποθέτουμε ότι οι n δοκιμές είναι αμοιβαία ανεξάρτητες και ότι κάθε δοκιμή έχει πιθανότητα επιτυχίας p. Έστω μια συγκεκριμένη τιμή p_0 στο διάστημα $[0, 1]$. Θέλουμε να ελέγξουμε

$$H_0 : p = p_0 \qquad H_1 : p \neq p_0.$$

Η στατιστική συνάρτηση $Y \sim Binomial(n, p)$. Κάτω από την H_0, $Y \sim Binomial(n, p_0)$ και έστω ότι η παρατηρούμενη τιμή της ελεγχοσυνάρτησης είναι y.

Παρατήρηση 4.4 *Επειδή η στατιστική συνάρτηση Y είναι διακριτή, ο έλεγχος σπάνια μπορεί να γίνει σε επίπεδο σημαντικότητας ακριβώς ίσο με το επιθυμητό α.*

$\alpha = \alpha_1 + \alpha_2$ και $\alpha_1 \simeq \alpha_2$. Απορρίπτω την H_0 αν $y \leq y_1$ ή $y \geq y_2$ (σε επίπεδο περίπου ίσο με α). Προσδιορίζουμε τα $y_1,\ y_2$ από πίνακες διωνυμικής κατανομής ώστε

$$\begin{aligned} & P(Y \leq y_1 | n, p_0) &=& \ \alpha_1 \simeq \frac{\alpha}{2} \\ \text{και} \quad & P(Y > y_2 | n, p_0) &=& \ 1 - P(Y \leq y_2 | n, p_0) = \alpha_2 \simeq \frac{\alpha}{2} \Rightarrow \\ \text{δηλ.} \quad & P(Y \leq y_2 | n, p_0) &=& \ 1 - \alpha_2 \simeq 1 - \frac{\alpha}{2} \end{aligned}$$

Μονόπλευροι Έλεγχοι

$$H_0 :\ p = p_0 \quad vs \quad H_1 :\ p > p_0$$

Προφανώς μεγάλες τιμές της στατιστικής συνάρτησης Y θα αποτελούν ένδειξη ότι δεν ισχύει η H_0.

Απορρίπτω την H_0, σε επίπεδο στατιστικής σημαντικότητας$\simeq \alpha$, αν $y > y_2$, όπου y_2 τέτοιο ώστε $P(Y > y_2) = \alpha_2 \simeq \alpha \Rightarrow P(Y \leq y_2) = 1 - \alpha_2 \simeq 1 - \alpha$.

$$H_0 :\ p = p_0 \quad vs \quad H_1 :\ p < p_0$$

Προφανώς μικρές τιμές της στατιστικής συνάρτησης Y θα αποτελούν ένδειξη ότι δεν ισχύει η H_0.

Απορρίπτω την H_0, σε επίπεδο στατιστικής σημαντικότητας περίπου ίσο με α, αν $y \leq y_1$, όπου y_1 τέτοιο ώστε $P(Y \leq y_1) = \alpha_1 \simeq \alpha$.

Παράδειγμα 4.10 *Για να μελετήσει τα αποτελέσματα της κόπωσης, ένας ερευνητής δίδαξε 18 παιδιά δύο διαφορετικούς τρόπους να δένουν κόμπο. Τα μισά παιδιά (που επιλέχθηκαν τυχαία) διδάχθηκαν πρώτα τον τρόπο A και μετά τον τρόπο B, ενώ τα υπόλοιπα διδάχθηκαν πρώτα τον τρόπο B. Στο τέλος μιας κουραστικής μέρας ζητήθηκε από όλα τα παιδιά να δέσουν κόμπο. Η πρόβλεψη του ερευνητή ήταν ότι τα παιδιά θα κατέληγαν να χρησιμοποιήσουν τον πρώτο τρόπο που έμαθαν λόγω κόπωσης. Τελικά, 16 παιδιά χρησιμοποίησαν τον τρόπο που έμαθαν πρώτο. Να ελεγχθεί σε $\alpha = 5\%$ αν ισχύει η πρόβλεψη του ερευνητή.*

Λύση:
Έστω p η πιθανότητα ένα παιδί να χρησιμοποιήσει τον τρόπο που έμαθε πρώτα. Ο ισχυρισμός του ερευνητή μπορεί να διατυπωθεί ως $p > 1 - p \Leftrightarrow p > 0.5$. Επομένως ένα κατάλληλο ζεύγος υποθέσεων για τον έλεγχο του ισχυρισμού του ερευνητή είναι

$H_0 :\ p = 0.5$*(η κόπωση δεν επηρεάζει την επιλογή του παιδιού)*

$H_1 :\ p > 0.5$*(η κόπωση οδηγεί το παιδί να επιλέξει τον τρόπο που έμαθε πρώτο).*

Έστω Y ο αριθμός των παιδιών στο δείγμα μεγέθους $n = 18$ που χρησιμοποίησαν τον τρόπο που έμαθαν πρώτο. Κάτω από την H_0, $Y \sim Binomial(n = 18, p_0 = 0.5)$. Απορρίπτω την H_0 σε επίπεδο στατιστικής σημαντικότητας περίπου 0.05 αν $y = 16 > y_2$, όπου y_2 τέτοιο ώστε

$$P(Y \leq y_2 | n = 18, p_0 = 0.5) \simeq 0.95.$$

Από τον πίνακα της Διωνυμικής κατανομής έχουμε ότι $P(Y \leq 12 | n = 18, p_0 = 0.5) = 0.9519$. Άρα παίρνω $y_2 = 12$. Επομένως, σε επίπεδο σημαντικότητας $\alpha_2 = 0.0481 \simeq 0.05$, απορρίπτω την H_0 καθώς $y = 16 > y_2 = 12$. Δηλαδή, σε επίπεδο α_2 οι ενδείξεις από το δείγμα ενισχύουν τον ισχυρισμό του ερευνητή. ■

Παρατήρηση 4.5 *Ο Διωνυμικός έλεγχος είναι στην ουσία ένας παραμετρικός έλεγχος υποθέσεων καθώς τα δεδομένα ακολουθούν διωνυμική κατανομή και ελέγχουμε αν η πιθανότητα επιτυχίας (παράμετρος της κατανομής) είναι ίση με συγκεκριμένη τιμή. Ωστόσο, διάφοροι μη-παραμετρικοί έλεγχοι ανάγονται δε διωνυμικούς.*

4.4.2 Ο Προσημικός Έλεγχος

Ο προσημικός έλεγχος χρησιμοποιείται κυρίως για τον έλεγχο της υπόθεσης ότι οι τιμές μιας από τις τυχαίες μεταβλητές (X, Z) τείνουν να είναι μεγαλύτερες από τις τιμές της άλλης. Χρησιμοποιείται ακόμα για τον έλεγχο της ύπαρξης τάσης καθώς και για τον έλεγχο ύπαρξης συσχέτισης.

Τα δεδομένα αποτελούνται από παρατηρήσεις πάνω σε δύο εξαρτημένες τ.μ. (X, Z). Κάθε ζεύγος (X_i, Z_i) χαρακτηρίζεται ως "$+$" ζεύγος αν $X_i < Z_i$, ως "$-$" ζεύγος αν $X_i > Z_i$ ή ως "0" ζεύγος αν $X_i = Z_i$. Θέλουμε να ελέγξουμε:

$$H_0 :\ P(+) = P(-) \qquad H_1 :\ P(+) \neq P(-).$$

Αγνοούμε τα "0" ζεύγη και θέτουμε

$$n = (\text{αριθμός "}+\text{" ζευγών}) \ + \ (\text{αριθμός "}-\text{" ζευγών}).$$

Θεωρούμε κάθε ζεύγος (X_i, Z_i), $i = 1, \ldots, n$ ως μια δοκιμή με αποτελέσματα: "$+$", επιτυχία με πιθανότητα $P(+)$, "$-$", αποτυχία με πιθανότητα $P(-)$.

Έτσι, αν Y είναι ο αριθμός των "$+$" ζευγών, κάτω από την H_0, $Y \sim Binomial(n, P(+) = 0.5)$. Δηλαδή, μπορούμε να χρησιμοποιήσουμε το διωνυμικό έλεγχο για να ελέγξουμε την ισχύ της H_0. Αν y είναι ο παρατηρούμενος αριθμός των "$+$" ζευγών, απορρίπτουμε την H_0 σε επίπεδο στατιστικής σημαντικότητας περίπου α αν $y \leq y_1$ ή $y > y_2$, όπου y_1, y_2 τέτοια ώστε

$$\begin{aligned} P(Y \leq y_1 | n, \frac{1}{2}) &= \alpha_1 \simeq \frac{\alpha}{2} \\ P(Y \leq y_2 | n, \frac{1}{2}) &= 1 - \alpha_2 \simeq 1 - \frac{\alpha}{2} \end{aligned}$$

Αντίστοιχα για τους αμφίπλευρους ελέγχους.

Παρατήρηση: Ο προσημικός έλεγχος υπονοεί ουσιαστικά έναν έλεγχο για τις αναμενόμενες τιμές των μεταβλητών X και Z. Δηλαδή,

$$H_0 : \ P(+) = P(-) \qquad H_1 : \ P(+) \neq P(-) \Rightarrow H_0 : \ E(X) = E(Z) \qquad H_1 : \ E(X) \neq E(Z).$$

Χωρίς να κάνουμε συγκεκριμένες υποθέσεις για την κατανομή της X και της Z (πράγμα που θα οδηγούσε σε παραμετρικό έλεγχο), απλά μετράμε τις φορές που $X_i > Z_i$(αποτυχίες) και οδηγούμαστε σε ένα διωνυμικό έλεγχο.

Παράδειγμα 4.11 *Έξι άτομα υποβλήθηκαν σε μία δίαιτα για να χάσουν βάρος. Τα βάρη τους πριν (X_i) και μετά (Z_i) τη δίαιτα είναι τα εξής:*

X_i	*174*	*191*	*188*	*182*	*201*	*188*
Z_i	*165*	*186*	*183*	*178*	*203*	*181*

Παρέχουν τα δεδομένα ενδείξεις ότι η δίαιτα ήταν αποτελεσματική;($\alpha = 0.05\%$)

Λύση:
Ορίζουμε ως "$+$" το ενδεχόμενο $\{X < Z\}$. Θέλουμε να ελέγξουμε

$$H_0 : \ P(+) = P(-) \ \textit{(μη αποτελεσματική)} \qquad H_1 : \ P(+) < P(-) \ \textit{(αποτελεσματική)}$$

δηλαδή

$$H_0 : \ P(+) \geq 0.5 \qquad H_1 : \ P(+) < 0.5.$$

Έστω Y ο αριθμός των "$+$" ζευγών.

X_i	*174*	*191*	*188*	*182*	*201*	*188*
Z_i	*165*	*186*	*183*	*178*	*203*	*181*
πρόσημο:	$-$	$-$	$-$	$-$	$+$	$-$

Παρατηρούμε ότι $n = 6$ και παρατηρούμενη τιμή $y = 1$. Κάτω από την H_0, $Y \sim Binomial(6, 0.5)$. Απορρίπτουμε την H_0 σε επίπεδο σημαντικότητας $\alpha \simeq 0.05$ αν $y = 1 < y_1$, όπου y_1 τέτοιο ώστε $P(Y \leq y_1 | n = 6, p_0 = 0.5) \simeq 0.05$. Από τον πίνακα της διωνυμικής έχουμε ότι $P(Y \leq 2 | n = 6, p_0 = 0.5) = 0.3438$ και $P(Y \leq 3 | n = 6, p_0 = 0.5) = 0.6562$. Άρα απορρίπτουμε την H_0. ■

4.5 Έλεγχος Wilcoxon

4.5.1 Ο έλεγχος των προσημασμένων τάξεων μεγέθους του Wilcoxon για τη διάμεσο ενός πληθυσμού

The Wilcoxon signed rank test for a median

Έστω τυχαίο δείγμα $X_1, \ldots, X_{n'}$ από την τυχαία μεταβλητή X της οποίας η κατανομή είναι συμμετρική. Έστω m η διάμεσος της κατανομής της X. Θέλουμε να ελέγξουμε

$$H_0 : \ m = m_0 \quad H_1 : \ m \neq m_0.$$

Θεωρούμε το δείγμα των διαφορών $D_i = m_0 - X_i, \ i = 1, \ldots, n'$. Η κάθε παρατήρηση D_i θα είναι θετική ή αρνητική με πιθανότητα 0.5 (κάτω από την H_0), καθώς η κατανομή των D_i είναι συμμετρική γύρω από το 0. Εξαιρούμε τις διαφορές που είναι ίσες με 0 και δουλεύουμε με τις απόλυτες τιμές των $n \leq n'$ μη μηδενικών διαφορών

$$|D_i| = |m - X_i|, \ i = 1, \ldots, n.$$

Διατάσσουμε σε αύξουσα τάξη μεγέθους τις τιμές $|D_i|$ και αντιστοιχίζουμε βαθμούς από το 1 ως το n στις παρατηρήσεις του διατεταγμένου δείγματος (ο βαθμός 1 αντιστοιχεί στη μικρότερη απόλυτη διαφορά και ο βαθμός n στη μεγαλύτερη). Σε περίπτωση ισότητας, αντιστοιχίζουμε στις ίσες διαφορές το μέσο βαθμό που αυτές θα είχαν αν διέφεραν.

Έστω $R(|D_i|), \ i = 1, \ldots, n$ η ακολουθία των βαθμών (τάξεων μεγέθους) των απολύτων διαφορών $|D_i|, \ i = 1, \ldots, n$. Ορίζουμε τις μεταβλητές (προσημασμένες τάξεις μεγέθους)

$$R_i = \begin{cases} +R(|D_i|), & \text{αν } D_i = m - X_i > 0 \\ -R(|D_i|), & \text{αν } D_i = m - X_i < 0 \end{cases} \quad i = 1, \ldots, n$$

Στατιστική συνάρτηση ελέγχου:

$$T = \frac{\sum_{i=1}^{n} R_i}{\sqrt{\sum_{i=1}^{n} R_i^2}}$$

Επειδή $R_1, \ldots, R_n$ τυχαίο δείγμα από συμμετρικό πληθυσμό με διάμεσο ίση με 0, η κατανομή της τ.μ. $S = \sum_{i=1}^{n} R_i$ έχει $E[S] = 0$ και $Var(S) = \sum_{i=1}^{n} R_i^2$. Άρα η T είναι η τυποποιημένη μορφή της S.

Η ελεγχοσυνάρτηση T λαμβάνει υπ'όψη:

i. Το μέγεθος των διαφορών $D_i, \ i = 1, \ldots, n$, το οποίο είναι ενδεικτικό της απόστασης των X_i από τη διάμεσό τους.

ii. Το πρόσημο των διαφορών $D_i, \ i = 1, \ldots, n$, το οποίο είναι ενδεικτικό της θέσης των X_i ως προς τη διάμεσο.

Η κατανομή της T προσεγγίζεται από την Κανονική κατανομή.

Παρατήρηση 4.6 *Στην περίπτωση που δεν υπάρχει ισότητα στα* $R_1, \ldots, R_n$, *έχουμε*

$$Var(S) = \sum_{i=1}^{n} R_i^2 = \text{άθροισμα των τετραγώνων των } n \text{ πρώτων φυσικών αριθμών} = \frac{n(n+1)(2n+1)}{6}.$$

Ακραίες τιμές της T *συνεπάγονται ότι είτε οι θετικές είτε οι αρνητικές διαφορές υπερτερούν σε μέγεθος ή/και αριθμό. Επομένως απορρίπτουμε την* H_0 *σε επίπεδο σημαντικότητας* α *αν* $|T| > Z_{1-\frac{\alpha}{2}}$.

Παράδειγμα 4.12 *Οι πληθωρισμένες ετήσιες αποδόσεις των μετοχικών κεφαλαίων σε ένα χρηματιστήριο κατανέμονται συμμετρικά. Έστω τυχαίο δείγμα αποδόσεων*

$$+8.4 \quad -4.3 \quad -0.8 \quad +12.5$$

Να ελεγχθεί ($\alpha = 0.05$) *η*

$$H_0: \ m = 3 \qquad H_1: \ m \neq 3.$$

Λύση:

X_i	$D_i = 3 - X_i$	$\lvert D_i \rvert$	$R(\lvert D_i \rvert)$	R_i
$+8.4$	-5.4	*5.4*	*2*	-2
-4.3	$+7.3$	*7.3*	*3*	$+3$
-0.8	$+3.8$	*3.8*	*1*	$+1$
$+12.5$	-9.5	*9.5*	*4*	-4

$$T = \frac{\sum_{i=1}^{4} R_i}{\sqrt{\sum_{i=1}^{4} R_i^2}} = \frac{1-2+3-4}{\sqrt{4+9+1+16}} = -\frac{2}{\sqrt{30}} = 0.365.$$

Είναι $|T| = 0.365 < 1.96 = Z_{1-\alpha/2}$. *Άρα δεν απορρίπτουμε την* H_0. ■

4.5.2 Έλεγχος Wilcoxon για δείγμα ζευγών παρατηρήσεων

Είναι ένας έλεγχος βασισμένος στις τάξεις μεγέθους των παρατηρήσεων. Έστω ζεύγη παρατηρήσεων (X_i, Z_i), $i = 1, \ldots, n'$. Ορίζουμε τις διαφορές $D_i = Z_i - X_i$ $i = 1, \ldots, n'$ ως ένα δείγμα μονοδιάστατων παρατηρήσεων. Υποθέτουμε ότι η κατανομή των $D_1, \ldots, D_{n'}$ είναι συμμετρική. Η υπόθεση της συμμετρίας δεν είναι τόσο ισχυρή όσο η υπόθεση της κανονικότητας ή γενικά η υπόθεση κάποιας συγκεκριμένης κατανομής. Εάν μία κατανομή είναι συμμετρική, η μέση τιμή της ταυτίζεται με τη διάμεσο.

Θεωρούμε τις απόλυτες τιμές των $n \leq n'$ μη μεδενικών διαφορών $|D_i| = |Z_i - X_i|$ $i = 1, \ldots, n$. Έστω d η διάμεσος της κατανομής των διαφορών D. Η τιμή d παρέχει πληροφορίες για τις

σχετικές θέσεις των δύο πληθυσμών. Αν οι τιμές της Z είναι μεγαλύτερες από τις τιμές της X, τότε $d > 0$. Αν οι τιμές της X είναι μεγαλύτερες από τις τιμές της Z, τότε $d < 0$. Τέλος, αν οι τιμές της D είναι θετικές και αρνητικές με ίσες συχνότητες, τότε $d = 0$. Θέλουμε να ελέγξουμε:

$$H_0 : \ d = 0 \quad H_1 : \ d \neq 0 \qquad \text{(ή μονόπλευρους ελέγχους)}.$$

Παράδειγμα 4.13 *Για να ελεγχθούν τα αποτελέσματα ενός νέου εμβολίου, απαιτείται η γνώση της μεταβολής της θερμοκασίας του σώματος πριν(X_i) και μετά τον εμβολιασμό(Z_i). Οι μετρήσεις σε $n' = 10$ ασθενείς δίνονται στον παρακάτω πίνακα.*

X_i	*37.0*	*37.0*	*36.4*	*36.7*	*37.0*	*36.9*	*37.0*	*37.0*	*36.8*	*37.0*
Z_i	*38.0*	*37.2*	*37.3*	*38.6*	*37.8*	*36.9*	*36.9*	*39.6*	*37.6*	*37.5*

Αποτελούν τα δεδομένα ένδειξη αύξησης της θερμοκρασίας του σώματος μετά τον εμβολιασμό;($\alpha = 0.05$)

Λύση:
Είναι

$$H_0 : \ d = 0 \quad H_1 : \ d > 0.$$

i	X_i	Z_i	$D_i = Z_i - X_i$	$\lvert D_i \rvert$	$R(\lvert D_i \rvert)$	R_i
1	*37.0*	*38.0*	$+1.0$	*1.0*	*7*	$+7$
2	*37.0*	*37.2*	$+0.2$	*0.2*	*2*	$+2$
3	*36.4*	*37.3*	$+0.9$	*0.9*	*6*	$+6$
4	*36.7*	*38.6*	$+1.9$	*1.9*	*8*	$+8$
5	*37.0*	*37.8*	$+0.8$	*0.8*	*4.5*	$+4.5$
6	*36.9*	*36.9*	*0.0*	*0.0*	$-$	$-$
7	*37.0*	*36.9*	-0.1	*0.1*	*1*	-1
8	*37.0*	*39.6*	$+2.6$	*2.6*	*9*	$+9$
9	*36.8*	*37.6*	$+0.8$	*0.8*	*4.5*	$+4.5$
10	*37.0*	*37.5*	$+0.5$	*0.5*	*3*	$+3$

Η τιμή της ελεγχοσυνάρτησης:

$$T = \frac{43}{\sqrt{284.5}} = 2.549 > 1.645 = Z_{0.95}.$$

Άρα σε $\alpha = 5\%$ τα δεδομένα παρέχουν ενδείξεις ότι η θερμοκρασία του σώματος αυξάνει σημαντικά μετά τον εμβολιασμό. ■