ΘΕΜΑ: «Ανάλυση Διακύμανσης»

Τμήμα: ''Επιχειρηματικού Σχεδιασμού &

Πληροφοριακών Συστημάτων''

Τ.Ε.Ι Πάτρας

2009

<u>Ονοματεπώνυμο: Ανδριάνα Κόρδα</u>

Επιβλέπων Καθηγητής: κος. Αθανάσιος Τσαγγανός

Περίληψη

Ο έλεγχος των μέσων τιμών δύο ανεξάρτητων πληθυσμών είναι η απλούστερη περίπτωση ενός σύνθετου ζητήματος που συναντάται συχνά στην στατιστική συμπερασματολογία και το οποίο, στην γενική του μορφή αφορά τη σύγκριση μέσων τιμών περισσότερων των δύο. Από μεθοδολογικής απόψεως, τέτοια προβλήματα επιλύονται με την βοήθεια των τεχνικών ης ευρύτερης περιοχής της στατιστικής ανάλυσης που φέρει το όνομα ανάλυση διακύμανσης. Ο λόγος για τον οποίο οι τεχνικές αυτές αναφέρονται με αυτήν την ονομασία, οφείλεται στο γεγονός ότι η λογική τους στοχεύει στο διαμερισμό της συνολικής διασποράς ενός συνόλου δεδομένων σε επιμέρους συνιστώσες. Στην εργασία αφού δώσουμε πρώτα κάποιους βασικούς ορισμούς της στατιστικής, θα περιγράψουμε την διαδικασία ανάλυσης διακύμανσης με θεωρεία και παραδείγματα κατά ένα παράγοντα, ανάλυση διακύμανσης κατά δύο παράγοντες και ύστερα το απλό και το γενικό γραμμικό υπόδειγμα. Τέλος θα εφαρμόσουμε την αναπτυχθείσα θεωρεία στην Παναιγιάλειο Ένωση Συνεταιρισμών για να βγάλουμε κάποια συμπεράσματα για την αγροτική παραγωγή.

Πίνακας Περιεχομένων

Εισαγωγή

Σε πολλά προβλήματα συμβαίνει μια ποσοτική εξαρτημένη μεταβλητή να επηρεάζεται από ποιοτικές ανεξάρτητες μεταβλητές. Μπορούμε επίσης να θεωρήσουμε ότι η εξάρτηση είναι γραμμική ως προς τις παραμέτρους και ότι το σφάλμα είναι αθροιστικό και κανονικά κατανεμημένο. Στην περίπτωση αυτή η μέθοδος που εφαρμόζουμε για τη μελέτη τους λέγεται ανάλυση διασποράς (analysis of variance) ή συντομότερα ANOVA. Η μέθοδος ANOVA, την οποία θα δούμε παρακάτω στην ουσία δε διαφέρει ιδιαίτερα από την παλινδρόμηση με ποιοτικές προβλέπουσες μεταβλητές υπάρχουν όμως κάποιοι λόγοι που επιβάλλουν την εκμάθησή της. Τέτοιοι λόγοι είναι, η ευκολία στους υπολογισμούς πράγμα ιδιαίτερα σημαντικό τις παλαιότερες εποχές όπου δεν ήταν δυνατή η χρήση υπολογιστικών συστημάτων. Ο δεύτερος λόγος για τη μελέτη της μεθόδου ANOVA είναι ιστορικός. Είναι αλήθεια ότι η μέθοδος αναπτύχθηκε πρώτη φορά από τους πειραματικούς ερευνητές, οι οποίοι φρόντιζαν οι προβλέπουσες μεταβλητές να έχουν μικρό αριθμό διαφορετικών τιμών οι οποίες μάλιστα να μπορούν να ελεγχθούν. Από εδώ προκύπτει και ένα μεγάλο μέρος της ορολογίας που χρησιμοποιείται στην ανάλυση προβλημάτων.

ΚΕΦΑΛΑΙΟ 1ο

1.0 Η σημασία της ανάλυσης διακύμανσης

Πριν προχωρήσουμε στην ανάλυση της μεθόδου ANOVA, κρίνεται απαραίτητο στο σημείο αυτό να δοθεί ο ορισμός της διακύμανσης. Η διακύμανση λοιπόν ορίζεται ως ο μέσος αριθμητικός των τετραγώνων των αποκλίσεων των τιμών της μεταβλητής από τον μέσο αριθμητικό και συμβολίζεται (διεθνώς) με το s^2, όταν αναφερόμαστε στον πληθυσμό και με το S^2 όταν αυτή αναφέρεται στο δείγμα (Ανδρέας Α. Κίντης, 2002). Έτσι στην περίπτωση που τα δεδομένα είναι αταξινόμητα, έχουμε:

- Για τον πληθυσμό $s^2 = \frac{\sum_i (X_1 - m)^2}{\mathrm{N}}$

- Για το δείγμα $S^2 = \frac{\sum_i (X_i - \bar{X})^2}{n}$

Επειδή η διακύμανση στον πληθυσμό είναι δύσκολο να εκτιμηθεί στην πράξη, διότι ούτε το N ούτε το μ είναι κατά κανόνα γνωστά, η τιμή του s^2 συνάγεται από την τιμή του S^2, δηλαδή από τη διακύμανση που εμφανίζουν οι τιμές στο δείγμα.

Μετά τον σύντομο αυτό ορισμό της διακύμανσης λοιπόν, συνεχίζουμε παρακάτω με την ανάλυση και τη σημασία της.

Η ANOVA χρησιμοποιείται για να ελέγξουμε την ισότητα μεταξύ k πληθυσμιακών μέσων $m_1, m_2 \mathbf{K}, m_k$ (για k>2). Η περίπτωση ελέγχου των μέσων k πληθυσμών παίρνει τη μορφή:

$$H_0 = \mu_1 = m_2 = m_3 = \mathbf{K} = m_k$$

H_1 : όχι όλοι οι πληθυσμιακοί μέσοι ίσοι ($m_1 \neq m_2 \neq m_3 \neq \mathbf{K} \neq m_k$)

Αν η μηδενική υπόθεση είναι σωστή, τότε η πιθανότητα να την απορρίψουμε είναι ισοδύναμη με την πιθανότητα ότι τουλάχιστον σε ένα ζευγάρι από τους k πληθυσμούς να μην έχουμε ίσους μέσους. Ας χρησιμοποιήσουμε ένα απλό παράδειγμα για την κατανόηση του παραπάνω. Αν θεωρήσουμε λοιπόν, την περίπτωση της σύγκρισης τεσσάρων πληθυσμών, τότε υπάρχουν 6 πιθανές συγκρίσεις ζευγών ($1^{ος}$ πληθυσμός με $2^{ο}$, $1^{ος}$ με $3^{ο}$, $1^{ος}$ με $4^{ο}$, $2^{ος}$ με $3^{ο}$, $2^{ος}$ με $4^{ο}$ και $3^{ος}$ με $4^{ο}$) και αν ένα σφάλμα τύπου I για επίπεδο σημαντικότητας 5% καθορίζεται για κάθε σύγκριση, τότε η ολική πιθανότητα μπορεί να πολλαπλασιαστεί ως μια δυωνυμική πιθανότητα του να καταγράψουμε μια ή περισσότερες <<επιτυχίες>> (απορρίψεις) όπου η πιθανότητα κάθε επιτυχίας (απόρριψης) είναι 5%. Με έναν απλό πολλαπλασιασμό ανακαλύπτουμε ότι η πιθανότητα είναι 0,265. Συμπερασματικά, αν κάποιος χρησιμοποιήσει συγκρίσεις ζευγών, η πιθανότητα να απορρίψουμε τη μηδενική υπόθεση όταν αυτή είναι σωστή (η πιθανότητα δηλαδή να υποπέσουμε σε σφάλμα τύπου I) είναι 26,5%.

ΚΕΦΑΛΑΙΟ 2^{ο}

2.0 Στατιστικός έλεγχος της H_0

Εισαγωγή

Ο έλεγχος υποθέσεων αναφέρεται στη διαδικασία ή στον τρόπο με τον οποίο καταλήγουμε στην αποδοχή ή στην απόρριψη μιας στατιστικής υποθέσεως. Με τον όρο στατιστική υπόθεση εννοούμε μια υπόθεση σχετικά με την κατανομή μιας τυχαίας μεταβλητής. Η υπόθεση συνήθως αφορά τις παραμέτρους της κατανομής (πληθυσμού), αλλά μπορεί επίσης να αναφέρεται στην φύση ή τη συναρτησιακή μορφή της κατανομής (Γεώργιος Κ. Χρήστου, 2002). Για παράδειγμα, ας υποθέσουμε ότι ο υπεύθυνος παραγωγής σε κάποιο εργοστάσιο θέλει να ελέγξει, βασισμένος στις πληροφορίες τυχαίου δείγματος, αν το ποσοστό των ελαττωματικών προιόντων που παράγει μια μηχανή είναι το πολύ 5%. Το ποσοστό αναφέρεται στην παράμετρο π της δυωνυμικής κατανομής και η υπόθεση $p \leq 0.05$ είναι μια στατιστική υπόθεση.

Αν η στατιστική υπόθεση καθορίζει ή εξειδικεύει τελείως τη συναρτησιακή μορφή της κατανομής καθώς και τις τιμές των παραμέτρων, τότε ονομάζεται απλή υπόθεση (simple hypothesis). Στην αντίθετη περίπτωση ονομάζεται σύνθετη υπόθεση (composite hypothesis)(Γεώργιος Κ. Χρήστου, 2002). Στο προηγούμενο παράδειγμα η υπόθεση $p \leq 0.05$ είναι σύνθετη υπόθεση γιατί δεν καθορίζει μια τιμή για το p.

Για τον έλεγχο μιας στατιστικής υποθέσεως, όπως η προηγούμενη, θα πρέπει να υπάρχει και μια εναλλακτική υπόθεση (alternative hypothesis). Με άλλα λόγια, εφόσον ο έλεγχος μιας υπόθεσης καταλήγει, βάσει κριτηρίων που θα αναφερθούν παρακάτω, στην αποδοχή ή απόρριψη ή όχι της υπόθεσης, για να έχει νόημα ο έλεγχος θα πρέπει να υπάρχει μια εναλλακτική υπόθεση για την περίπτωση που ελεγχόμενη υπόθεση δεν γίνεται αποδεκτή. Επίσης, η εναλλακτική υπόθεση είναι αναγκαία και για την εξεύρεση των κατάλληλων στατιστικών κριτηρίων με τα οποία θα γίνει ο έλεγχος. Θα παριστάνουμε την

εναλλακτική υπόθεση με H_1 και την ελεγχόμενη υπόθεση, η οποία στη στατιστική αποκαλείται μηδέν υπόθεση (null hypothesis), με H_0.

Όπως η μηδέν υπόθεση, έτσι και η εναλλακτική μπορεί να είναι απλή ή σύνθετη. Να σημειωθεί ότι στην πράξη δύσκολα έχουμε απλή μηδέν και απλή εναλλακτική υπόθεση. Συνήθως η μηδέν ή η εναλλακτική, ή και οι δύο είναι σύνθετες.

Ο καθορισμός της μηδέν και της εναλλακτικής αποτελεί το πρώτο στάδιο στη διαδικασία ελέγχου μιας στατιστικής υποθέσεως. Το επόμενο βήμα είναι η εύρεση ενός κανόνα ή κριτηρίου, σύμφωνα με το οποίο θα γίνει αποδεκτή ή θα απορριφθεί η μηδέν υπόθεση.

2.1 Κριτήριο ελέγχου

Έστω θ μια άγνωστη παράμετρος ενός πληθυσμού και ότι θέλουμε να ελέγξουμε την υπόθεση $q = q_0$ έναντι της υποθέσεως $q < q_0$. Αυτό γράφεται:

$$\mathrm{H}_0 : q = 0$$
$$\mathrm{H}_1 : q < q_0$$

Έστω επίσης ένα τυχαίο δείγμα $(X_1, X_2 \mathbf{K}, X_n)$ και β ένας εκτιμητής της παραμέτρου θ ή απλά μια στατιστική του δείγματος. Ο κανόνας για την αποδοχή ή απόρριψη της μηδέν υποθέσεως, θα μπορούσε να είναι ο εξής:

Αν $b \geq k$, η μηδέν υπόθεση, $q = q_0$, γίνεται δεκτή.

Αν $b < k$, η μηδέν υπόθεση απορρίπτεται υπέρ της εναλλακτικής,

$q < q_0$, όπου k είναι μια σταθερά.

Είναι φανερό ότι η εκλογή ενός κανόνα αυτόματα χωρίζει ολόκληρο το δειγματικό χώρο σε δύο μέρη ή περιοχές: την περιοχή απορρίψεως (region of rejection) ή κρίσιμη περιοχή (critical region) και την περιοχή αποδοχής (region of acceptance). Στο παράδειγμά μας, η κρίσιμη περιοχή περιλαμβάνει όλες τις τιμές της β, για τις οποίες *β<k*, και η περιοχή αποδοχής όλες τις τιμές της β, για τις οποίες $b \geq k$.

Αν επομένως, η τιμή της β από το δείγμα είναι μικρότερη από k, η μηδέν υπόθεση απορρίπτεται και γίνεται αποδεκτή η εναλλακτική. Αν $b \geq k$, η μηδέν υπόθεση γίνεται αποδεκτή.

Έστω ότι θέλουμε να ελέγξουμε την υπόθεση:

$$H_0 : q = q_0$$
$$H_1 : q \neq q_0$$

Ένα κριτήριο ελέγχου ανάλογο με το προηγούμενο θα μπορούσε να είναι το εξής:

Αν $k_1 \leq b \leq k_2$, η μηδέν υπόθεση γίνεται δεκτή.

Αν $b < k_1$ ή $b > k_2$, η μηδέν υπόθεση απορρίπτεται και γίνεται δεκτή η εναλλακτική $q \neq q_0$, όπου k_1, k_2 είναι σταθερές και $k_1 < k_2$.

Αν η τιμή της β από το δείγμα είναι μικρότερη από k_1 ή μεγαλύτερη από k_2, η μηδέν υπόθεση απορρίπτεται. Διαφορετικά γίνεται δεκτή.

Τα προηγούμενα κριτήρια είναι εντελώς αυθαίρετα και λογικά ενδιαφερόμαστε να έχουμε κριτήρια τα οποία έχουν ορισμένες επιθυμητές στατιστικές ιδιότητες.

ΚΕΦΑΛΑΙΟ 3ο

3.0 Ανάλυση διακύμανσης κατά ένα παράγοντα

Με τον όρο αυτό εννοούμε την ανάλυση προβλημάτων όπου μια εξαρτημένη, παρατηρούμενη, ποσοτική μεταβλητή Χ, υποτίθεται ότι επηρεάζεται από έναν παράγοντα (factor). Ο παράγοντας είναι μια ποιοτική, ελέγξιμη μεταβλητή, η οποία μπορεί να παίρνει ένα πεπερασμένο πλήθος τιμές που λέγονται στάθμες (levels). Αν είχαμε δύο ή περισσότερους παράγοντες, τότε οι παρατηρήσεις της Χ θα γίνονταν για συγκεκριμένους συνδυασμούς από στάθμες των παραγόντων. Κάθε τέτοιος συνδυασμός λέγεται αγωγή (treatment). Στην περίπτωση του ενός παράγοντα οι αγωγές ταυτίζονται με τις στάθμες.

Η ανάλυση διακύμανσης ενός παράγοντα υποθέτει ότι οι υπό εξέταση πληθυσμοί είναι όλοι κανονικά κατανεμόμενοι, οι παρατηρήσεις για κάθε πληθυσμό είναι ανεξάρτητες και ότι οι διακυμάνσεις κάθε πληθυσμού είναι ίσες.

Ας υποθέσουμε λοιπόν ότι έχουμε ένα παράγοντα που εμφανίζεται σε k στάθμες ή μια ασθένεια που αντιμετωπίζεται με k διαφορετικές θεραπείες κ.τ.λ. Κάνουμε τότε μια σειρά από n παρατηρήσεις, από τις οποίες r_i, $i = 1, 2, \mathbf{K}, k$ σε κάθε στάθμη ή θεραπεία, όπου $r_1 + r_2 + ... + r_k = n$. Συμβολίζουμε με X_{ij} την τιμή μιας εξαρτημένης μεταβλητής στην j-στη παρατήρηση της i-στης ομάδας. Είναι φανερό ότι η εξαρτημένη μεταβλητή Χ μπορεί να επηρεάζεται ή όχι από τις διαφορετικές στάθμες ή ομάδες. Ένας τρόπος να το ελέγξουμε αυτό είναι να προσαρμόσουμε στα δεδομένα μας το μοντέλο:

$$X_{ij} = m_i + e_{ij} = m + d_i + e_{ij} \ ,$$

όπου μ είναι σταθερά,

d_i είναι αυτό που θα ονομάζεται κύρια επίδραση (main effect) του παράγοντα D στη στάθμη i ($i = 1, 2, \mathbf{K}, k$),

e_{ij} το σφάλμα της j παρατήρησης ($j = 1, 2, \mathbf{K}, r_i$) στην i στάθμη ($i = 1, 2, \mathbf{K}, k$). Το e_{ij} υποτίθεται ότι είναι η τιμή της τυχαίας μεταβλητής ε που έχει κατανομή N ($0, s^2$).

Σχηματικά το όλο πείραμα (ή φαινόμενο) με τη βοήθεια του παραπάνω μοντέλου μπορεί να παρασταθεί όπως στον πίνακα:

Στάθμη 1	Στάθμη2	Στάθμη k
$X_{11} = m + d_1 + e_{11}$ $X_{12} = m + d_1 + e_{12}$ ..... $X_{1r_1} = m + d_1 + e_{1r_1}$	$X_{21} = m + d_2 + e_{21}$ $X_{22} = m + d_2 + e_{22}$ ..... $X_{2r_2} = m + d_2 + e_{2r_2}$	
$X_{k1} = m + d_k + e_{k1}$ $X_{k2} = m + d_k + e_{k2}$ ..... $X_{kr_k} = m + d_k + e_{kr_k}$		

Έστω ότι κάθε πληθυσμός έχει μέσο m_j και διακύμανση s^2 (η οποία από τις υποθέσεις, είναι κοινή μεταξύ των πληθυσμών). Κάθε πληθυσμός είναι κανονικά κατανεμημένος με ανεξάρτητες παρατηρήσεις, έτσι ώστε μια τυχαία παρατήρηση X_{ij} από τον j πληθυσμό να έχει την παρακάτω κατανομή:

$$X_{ij} \square N(m_j, s^2)$$

Η δειγματική διακύμανση για κάθε δειγματικό μέσο $\bar{X}_j$ θα κατανέμεται κανονικά ως:

$$\bar{X}_j \square N(m_j, s^2 / n_j)$$

όπου $\bar{X}_j$ είναι ο μέσος δείγματος εξαγόμενου από τον j πληθυσμό. Τέλος, και μόνο κάτω από τη μηδενική υπόθεση, ο μέσος των μέσων ή γενικός μέσος (grand mean) $\bar{X}$ θα κατανέμεται σύμφωνα με την ακόλουθη κανονική κατανομή:

$$\bar{X} \sim N(m, s^2 / \sum_{j=1}^{k} n_j)$$

Η ουσία ενός τεστ ANOVA βασίζεται στην ανάλυση της συνολικής μεταβλητότητας των δεδομένων. Ένα μέτρο σύγκρισης για τη συνολική μεταβλητότητα στα k δείγματα είναι το συνολικό άθροισμα τετραγώνων (Total Sum of Squares, TSS) το οποίο δίνεται από τον τύπο:

$$TSS = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2$$

Η ANOVA στηρίζεται στη σύγκριση της μεταβλητότητας μεταξύ των δειγματικών μέσων και της μεταβλητότητας των τιμών μέσα σε κάθε δείγμα. Αυτό σημαίνει ότι υπάρχουν δύο πιθανές επεξηγήσεις για κάθε παρατηρημένη μεταβλητότητα στα δεδομένα:

1. Η διακύμανση μεταξύ των δειγμάτων: που περιγράφει τη συστηματική διακύμανση κατά μήκος των δειγμάτων που έχει προκληθεί από διαφορές στους ίδιους τους πληθυσμιακούς μέσους.
2. Η διακύμανση μέσα στα δείγματα: η οποία περιγράφει τις τυχαίες διακυμάνσεις κάθε δείγματος δεδομένων γύρω από κάθε πληθυσμιακό μέσο.

Για να αναλύσουμε τη σχετική σημαντικότητα αυτών των δύο πηγών μεταβλητότητας, μπορούμε να διασπάσουμε τη συνολική μεταβλητότητα στα δεδομένα (TSS) σε αυτή που οφείλεται καθαρά στην τυχαία μεταβλητότητα εντός των δειγμάτων, και σε αυτή που περιλαμβάνει κάποια συστηματική μεταβλητότητα μεταξύ των δειγμάτων. Η ανάλυση αυτή παρουσιάζεται ως εξής:

$$\sum_{j=1}^{k} \sum_{i=1}^{nj} (X_{ij} - \bar{X})^2 = \sum_{j=1}^{k} n_j (\bar{X}_j - \bar{X})^2 + \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2$$

Για να δώσουμε κάποια διαίσθηση στην ανάλυση, μπορούμε να γράψουμε την παραπάνω έκφραση ως εξής:

TSS = SSA + ESS

όπου ο πρώτος όρος SSA συμβολίζει την άθροιση τετραγώνων των αποκλίσεων των δειγματικών μέσων του παράγοντα A από το γενικό μέσο (Sum of Squares of Sample Averages) και ο δεύτερος όρος ESS παριστάνει το άθροισμα τετραγώνων των σφαλμάτων (Error Sum of Squares, ESS). Ο όρος SSA είναι απλός στην επεξήγησή του αφού παριστάνει τη μεταβλητότητα των διαφορών που έχουν μεταξύ τους τα δείγματα. Ο όρος ESS μετρά τη συνολική μεταβλητότητα εντός των δειγμάτων. Ειδικότερα:

$$SSA = \sum_{j=1}^{k} n_j (X_j - \bar{X})^2$$

$$ESS = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2$$

Όπως βλέπουμε, ο όρος ESS είναι ένα μέτρο της μεταβλητότητας των τυχαίων παρατηρήσεων σε κάθε δείγμα γύρω από τους αντίστοιχους μέσους. Έτσι, ESS μετρά τη διακύμανση στα k δείγματα ως σύνολο. Αν εξετάσουμε τη δομή του πρώτου όρου SSA, βλέπουμε ότι μετράει τη διακύμανση στους πληθυσμιακούς μέσους γύρω από το γενικό μέσο, σταθμισμένο από το μέγεθος του κάθε δείγματος. Με άλλα λόγια, SSA, μπορεί να χρησιμοποιηθεί ως δειγματικό στατιστικό μέγεθος για να εκτιμήσει τη διακύμανση μεταξύ των k πληθυσμιακών μέσων. Αν όλοι οι πληθυσμιακοί μέσοι είναι ίσοι, τότε ο όρος SSA παρουσιάζει κάποια μεταβλητότητα στα δεδομένα, που οφείλεται μόνο σε τυχαία διακύμανση (και συμπερασματικά θα είναι της ίδιας τάξης με την ESS). Αν, από την άλλη πλευρά, οι πληθυσμιακοί μέσοι είναι διαφορετικοί, τότε οι διαφορές στην SSA θα είναι μεγαλύτερες.

Αν η μηδενική υπόθεση είναι αληθής ($H_0 : m_1 = m_2 = m_3 = \mathbf{K} = m_k$), τότε δεν θα πρέπει να υπάρχει συστηματική διακύμανση μεταξύ των k δειγματικών μέσων. Με άλλα λόγια, SSA και ESS θα πρέπει να είναι της ίδιας τάξης. Υπό μια έννοια, χρειάζεται να συγκρίνουμε τη σημαντικότητα της διακύμανσης εντός και μεταξύ των ομάδων. Για να γίνει αυτό, χρησιμοποιούμε τη στατιστική F που σχηματίζεται από το λόγο των ως X^2 κατανεμόμενων στατιστικών SSA και ESS. Επειδή μας ενδιαφέρει η μέση μεταβολή, δημιουργούμε το μέσο άθροισμα τετραγώνων των μέσων των δειγμάτων (Mean Squares of Averages, MSA), το οποίο βρίσκεται από τη διαίρεση της άθροισης

τετραγώνων των αποκλίσεων των δειγματικών μέσων προς τους βαθμούς ελευθερίας, δηλαδή $MSA = SSA/(k-1)$. Ομοίως, για τον όρο ESS παίρνουμε το μέσο άθροισμα τετραγώνων των σφαλμάτων (Mean Square Error, MSE) όπου $MSE = ESS/\Sigma(n_j - 1)$. Η στατιστική F της ανάλυσης της διακύμανσης προκύπτει από το λόγο του MSA προς MSE.

Αυτό σημαίνει, διατηρώντας τη μηδενική υπόθεση ως αληθινή, ότι μπορούμε να ορίσουμε τα παρακάτω:

$$F = \frac{SSA/(k-1)}{ESS/\Sigma(n_j - 1)} = \frac{MSA}{MSE} \square F_{k-1,N-k,a}$$

Παρατηρούμε ότι $\Sigma(n_j - 1) = N - k,$ όπου N ο συνολικός αριθμός των παρατηρήσεων και k ο αριθμός των δειγμάτων. Αν η αρχική υπόθεση είναι σωστή, τότε (επειδή οι προσδοκίες για τον αριθμητή και τον παρονομαστή είναι 1) η στατιστική F θα έπρεπε να κυμαίνεται γύρω στο 1. Μπορούμε να απορρίψουμε τη μηδενική υπόθεση αν η τιμή της στατιστικής F υπερβαίνει μια κριτική τιμή από την F κατανομή σύμφωνα με ένα δεδομένο επίπεδο σημαντικότητας (5%, για παράδειγμα).

Προϋποθέσεις εφαρμογής ανάλυσης διακύμανσης

Η ανάλυση διακύμανσης προς ένα παράγοντα είναι μια κατ' εξοχήν παραμετρική διαδικασία η οποία απαιτεί την πλήρωση συγκεκριμένων προϋποθέσεων:

Οι κατανομές της ποσοτικής μεταβλητής στους κ πληθυσμούς από τους οποίους προέρχονται οι ομάδες, θα πρέπει να είναι κανονικές. Η ανάλυση διακύμανσης είναι αρκετά ανθεκτική σε αποκλίσεις από την κανονικότητα, ώστε αυτό που ουσιαστικά απαιτεί για την πλήρωση της προϋπόθεσης, είναι η στοιχειώδης συμμετρία των κατανομών στις ομάδες και η απουσία πολλών ακραίων τιμών από αυτές. Ο έλεγχος κανονικότητας γίνεται διαγραμματικά με την χρήση των θηκογραμμάτων και των ιστογραμμάτων. Όταν ο συνολικός αριθμός των παρατηρήσεων είναι μικρός $n \le 30$, είναι απαραίτητη η χρήση μη παραμετρικών μεθόδων, όπως η ανάλυση διακύμανσης των Kruskal- Wallis

(διότι η διασφάλιση της κανονικότητας σε αυτή την περίπτωση είναι δύσκολό να επιτευχθεί).

Οι διακυμάνσεις της ποσοτικής μεταβλητής στους k πληθυσμούς θα πρέπει να είναι ίσες. Η προϋπόθεση αυτή είναι απαραίτητο να διασφαλιστεί κατά την χρήση της μεθόδου. Ο έλεγχος της ισότητας των διακυμάνσεων γίνεται με την βοήθεια του test Levene. Η μοναδική περίπτωση να παρακαμφθεί αυτή η προϋπόθεση είναι όταν ο αριθμός των παρατηρήσεων ανά ομάδα είναι περίπου ίδιος. Για τις περιπτώσεις που δεν διασφαλίζεται η ισότητα των διακυμάνσεων, η διαδικασία ONE-WAY ANOVA διαθέτει τα test των Brown-Forsythe και Welch. Τα κριτήρια των δύο αυτών ελέγχων, όταν η μηδενική υπόθεση της ισότητας των μέσων τιμών ισχύει, ακολουθούν προσεγγιστικά την κατανομή F χωρίς να απαιτείται η ισότητα των πληθυσμιακών διακυμάνσεων.

3.1 Παράδειγμα Ανάλυσης Διακύμανσης κατά ένα παράγοντα

Για να πραγματοποιήσουμε μια *One- Way Anova από το SPSS*, ακολουθούμε την εξής διαδικασία

-Από το μενού επιλέγουμε:

Analyse- Compare Means – One-Way Anova

Για το παράδειγμα μας θα εξετάσουμε πόσος είναι ο απαιτούμενος χρόνος εκπαίδευσης ενός ιατρικού επισκέπτη. Μετά το τέλος της εκπαίδευσης κάθε υπάλληλος γράφει ένα τεστ αξιολόγησης. Έχουμε τρεις ομάδες εκπαίδευσης.

(ένας μήνας, δύο μήνες τρεις μήνες)

Descriptives

Score on training exam

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	20	63,5798	13,50858	3,02061	57,2576	69,9020	32,68	86,66
2	20	73,5677	10,60901	2,37225	68,6025	78,5328	47,56	89,65
3	20	79,2792	4,40754	,98556	77,2165	81,3420	71,77	89,69
Total	60	72,1422	12,00312	1,54960	69,0415	75,2430	32,68	89,69

Στον πρώτο πίνακα εμφανίζονται τα αποτελέσματα της επιλογής Descriptives. Στο εσωτερικό του πίνακα δίνεται για κάθε ομάδα ο αριθμός των έγκυρων τιμών, η μέση τιμή, η τυπική απόκλιση, το τυπικό σφάλμα, το 95% διάστημα εμπιστοσύνης και οι ακραίες τιμές. Παρατηρούμε ότι καθώς αυξάνεται η εκπαίδευση, η τυπική απόκλιση μειώνεται.

Test of Homogeneity of Variances

Score on training exam

Levene Statistic	df1	df2	Sig.
4,637	2	57	,014

Το τεστ Levene απορρίπτει την μηδενική υπόθεση ότι απορρίπτει την Ho ότι οι διακυμάνσεις των ομάδων είναι ίσες, όμως το τεστ της ανάλυσης διακύμανσης κατά ένα παράγοντα είναι αρκετά ανθεκτικό όταν το πλήθος των τιμών των ομάδων είναι σχεδόν ίσο.

ANOVA

Score on training exam

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2525,691	2	1262,846	12,048	,000
Within Groups	5974,724	57	104,820		
Total	8500,415	59			

Το Sig ισούται με μηδέν άρα απορρίπτεται η υπόθεση ότι οι μέσες τιμές των ομάδων είναι ίσες, άρα υπάρχει διαφορά στην επίδοση των ιατρικών επισκεπτών ανάλογα με την εκπαίδευση.

3.1 Η ANOVA σε μορφή πίνακα

Ίσως ο πιο εύκολος τρόπος να διεξάγουμε μια ανάλυση διακύμανσης κατά ένα παράγοντα είναι το να χρησιμοποιήσουμε ένα πίνακα ανάλυσης διακύμανσης. Ο πίνακας παρουσιάζει άμεσα όλα όσα έχουμε συζητήσει μέχρι τώρα.

Πηγή της Διακύμανσης	***Αθροίσματα τετραγώνων***	***Βαθμοί ελευθερίας***	***Μέσα αθροίσματα τετραγώνων***	***Λόγος F***
Μεταξύ των Ομάδων (λόγω του Παράγοντα)	$SSA = \sum_{j=1}^{k} n_j (X_j - \bar{X})^2$	k-1	MSA=SSA/ k-1	F=MSA/MSE
Εντός των Ομάδων (σφάλμα)	$ESS = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2$	N-k	MSE=ESS/ N-k	
Σύνολο	$TSS = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2$	N-1		

<u>Παράδειγμα Anova σε μορφή πίνακα</u>

Έστω ότι θέλουμε να συγκρίνουμε την αποτελεσματικότητα τριών φροντιστηρίων για τις εξετάσεις των Αγγλικών. Ένας στατιστικός παίρνει τρία δείγματα, ένα από κάθε φροντιστήριο, και συγκρίνει την απόδοση 21 σπουδαστών στο ίδιο τεστ (σε ποσοστιαία κλίμακα) στα τρία διαφορετικά φροντιστήρια.

$1^{ο}$ Φροντιστήριο: 69, 71, 79, 68, 75, 80, 79

$2^{ο}$ Φροντιστήριο: 62, 65, 49, 72, 61, 67, 66

3^{ο} Φροντιστήριο: 67, 68, 83, 79, 79, 75, 76

Οι μέσοι των τριών δειγμάτων είναι, αντίστοιχα, 74.43, 63.14 και 75.29. Αν θέλουμε να ελέγξουμε, με βάση αυτό το δείγμα, την αποτελεσματικότητα των τριών φροντιστηρίων σχετικά με την απόδοση των σπουδαστών, τότε οι υποθέσεις ελέγχου μπορούν να διατυπωθούν ως εξής:

$$H_0 : m_1 = m_2 = m_3$$
$$H_1 : m_1 \neq m_2 \neq m_3$$

Βασιζόμενοι στα τρία δείγματα από τα δεδομένα, μπορούμε να συνοψίσουμε τους απαιτούμενους υπολογισμούς ως εξής:

Πηγή της διακύμανσης	**Αθροίσματα τετραγώνων**	**Βαθμοί ελευθερίας**	**Μέσα αθροίσματα τετραγώνων**	**Λόγος F**
Παράγοντας Σφάλμα Σύνολο	643 676 1319	2 18 20	643/2=321,5 676/18=37,6	321,5/37,6=8,28

Σε επίπεδο σημαντικότητας α=1%, η κριτική τιμή για την F κατανομή είναι $F_{0.01,2,18} = 6.01$ και, αφού η F από το τεστ υπερβαίνει την κριτική τιμή της κατανομής F, πρέπει να απορρίψουμε τη μηδενική υπόθεση, και καταλήγουμε ότι τα τρία φροντιστήρια δεν έχουν την ίδια αποτελεσματικότητα στην προετοιμασία των σπουδαστών.

3.2 Διαστήματα εμπιστοσύνης για τους k πληθυσμιακούς μέσους στην ANOVA

Αν απορρίψουμε τη μηδενική υπόθεση των ίσων πληθυσμιακών μέσων χρησιμοποιώντας την τεχνική της ανάλυσης διακύμανσης ενός παράγοντα, ίσως είναι χρήσιμο να προχωρήσουμε ακόμα πιο πολύ με το να κατασκευάσουμε διαστήματα εμπιστοσύνης για κάθε πληθυσμιακό μέσο

βασιζόμενοι στο κάθε δείγμα. Ας υποθέσουμε ότι πρέπει να κατασκευάσουμε k διαστήματα εμπιστοσύνης για κάθε πληθυσμιακό μέσο βασιζόμενοι στα ατομικά δείγματα και για 95% επίπεδο πιθανότητας. Όπως γνωρίζουμε, δε μπορούμε να είμαστε κατά 95% σίγουροι ότι οι αληθινοί πληθυσμιακοί μέσοι θα βρίσκονται ταυτόχρονα μέσα σε όλα αυτά τα k διαστήματα εμπιστοσύνης. Στην πραγματικότητα, αφού έχουμε υποθέσει ότι οι k πληθυσμοί είναι ανεξάρτητοι, τότε το συνολικό επίπεδο εμπιστοσύνης υπολογίζεται ως η πιθανότητα της αλληλεπίδρασης των k ανεξάρτητων δειγμάτων, το καθένα με πιθανότητα 0,95, η οποία είναι ίση με $(0.95)^k$. Γενικά, αν κατασκευάσουμε k διαστήματα εμπιστοσύνης, καθένα με επίπεδο πιθανότητας (1-α)%, τότε το κοινό επίπεδο πιθανότητας για τα k διαστήματα εμπιστοσύνης είναι $(1-a)^k$ %. Με άλλα λόγια, μπορούμε να είμαστε μόνο $(1-a)^k$ % σίγουροι ότι οι πληθυσμιακοί μέσοι θα βρίσκονται μέσα σε κάθε ένα από τα k διαστήματα εμπιστοσύνης.

Για την εκτίμηση των διαστημάτων εμπιστοσύνης των k πληθυσμιακών μέσων χρησιμοποιούμε τον παρακάτω τύπο:

$$\Pr\left(\bar{X}_j - t_{a/2} \frac{s_{pooled}}{\sqrt{n_j}} \le m_j \le \bar{X}_j + t_{a/2} \frac{s_{pooled}}{\sqrt{n_j}} \right) = 100(1-a)\,\%$$

όπου s_{pooled} είναι η κοινή τυπική απόκλιση που δίνεται από τον τύπο:

$$s_{pooled} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \ldots + (n_k-1)s_k^2}{N-k}} = \sqrt{MSE}$$

Η κατανομή t κατανέμεται με βαθμούς ελευθερίας ίσους με αυτούς του εκτιμητή της κοινής διακύμανσης, δηλαδή με N-k.

ΚΕΦΑΛΑΙΟ 4ο

4.0 Ανάλυση διακύμανσης κατά δύο παράγοντες

Σε προηγούμενο κεφάλαιο μελετήσαμε την ανάλυση διακύμανσης κατά ένα παράγοντα. Σε αυτό το κεφάλαιο θα μελετήσουμε την ανάλυση διακύμανσης με δύο παράγοντες.

Ας υποθέσουμε λοιπόν ότι στο πρόβλημα που μας ενδιαφέρει υπεισέρχονται δύο παράγοντες Α και Β, που εμφανίζονται με p και q στάθμες αντίστοιχα. Ας υποθέσουμε επιπλέον ότι σε κάθε στάθμη του παράγοντα Α έγιναν r*q παρατηρήσεις, από r δηλαδή για κάθε στάθμη του Β. Σχηματικά το πρόβλημα μπορεί να περιγραφεί με τον παρακάτω πίνακα:

	B_1	B_2	…	B_q
A_1	$X_{111}, \mathbf{K}, X_{11r}$	$X_{121}, \mathbf{K}, X_{12r}$	…	$X_{1q1}, \mathbf{K}, X_{1qr}$
A_2	$X_{211}, \mathbf{K}, X_{21r}$	$X_{221}, \mathbf{K}, X_{22r}$	…	$X_{2q1}, \mathbf{K}, X_{2qr}$
….	…….	…….	…	…….
A_p	$X_{p11}, \mathbf{K}, X_{p1r}$	$X_{p21}, \mathbf{K}, X_{p2r}$	…	$X_{pq1}, \mathbf{K}, X_{pqr}$

Δύο παράγοντες με r παρατηρήσεις ανά κελί

Σε ένα τέτοιο πρόβλημα η τιμή της εξαρτημένης μεταβλητής Χ μπορεί να εξαρτάται ή όχι από τους παράγοντες Α και Β. Πράγματι, αν αγνοήσουμε τις στάθμες του παράγοντα Β, το πρόβλημα γίνεται ισοδύναμο με αυτό της σελίδας 4, και επομένως μπορούμε να εκτιμήσουμε τις κύριες επιδράσεις του παράγοντα Α. Ομοίως, αν αγνοήσουμε τις στάθμες του Α, εκτιμούμε τις κύριες επιδράσεις του Β. Ας θεωρήσουμε τώρα μόνο τις pr παρατηρήσεις που έχουν γίνει στη στάθμη B_1, και να αγνοήσουμε όλες τις υπόλοιπες. Μελετώντας αυτά τα δεδομένα με ανάλογο, όπως και πριν, τρόπο βρίσκουμε τις κύριες επιδράσεις του παράγοντα Α, όταν όμως ο παράγοντας Β είναι στη στάθμη B_1. Αυτή η διαδικασία μπορεί να γίνει για όλες τις στάθμες του παράγοντα Β. Ενδέχεται τώρα σε ορισμένες από τις στάθμες του Β, οι κύριες επιδράσεις του

Α να είναι πολύ διαφορετικές μεταξύ τους όπως και με τις συνολικές επιδράσεις που εκτιμήθηκαν από όλα τα δεδομένα. Στην περίπτωση αυτή θα λέμε ότι οι παράγοντες Α και Β αλληλεπιδρούν και θα συμβολίζουμε την αλληλεπίδραση (interaction) των Α και Β στο κελί (i,j) με $(ab)_{ij}$.

Αν τώρα συμβολίσουμε με X_{ijk} την τιμή της μεταβλητής Χ στην k-οστή παρατήρηση που έγινε στη στάθμη i του παράγοντα Α και στην j του παράγοντα Β, τότε μπορούμε να θεωρήσουμε ότι ένα κατάλληλο μοντέλο για την περίπτωση είναι το:

$$X_{ijk} = m + a_i + b_j + (ab)_{ij} + e_{ijk},$$

όπου: μ είναι ο μέσος όρος

a_i είναι η επίδραση του παράγοντα Α στη στάθμη i, i=1, 2, ….p

b_j είναι η επίδραση του παράγοντα Β στη στάθμη j, j=1, 2, …, q

$(ab)_{ij}$ είναι η αλληλεπίδραση των παραγόντων Α, Β στο κελί (i , j)

e_{ijk} το σφάλμα της k-οστής παρατήρησης στο κελί (i , j), με k=1, 2, …, r.

Για τις παραμέτρους αυτές υποθέτουμε ότι ισχύουν οι παρακάτω p+q+2 ισότητες:

$$\sum_{i=1}^{p} a_i = \sum_{j=1}^{q} b_j = \sum_{i=1}^{p} (ab)_{ij} = \sum_{j=1}^{q} (ab)_{ij} = 0$$ **(σχέση 10)**

Χρησιμοποιώντας για τα αθροίσματα συμβολισμούς ανάλογους με τη σελίδα 5 και εργαζόμενοι όπως στη σελίδα 4, βρίσκουμε ότι το συνολικό άθροισμα τετραγώνων των αποκλίσεων

$$TSS = \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{k=1}^{r} (X_{ijk} - \bar{X})^2$$

μπορεί να αναλυθεί σε επιμέρους αθροίσματα, δηλαδή

$$TSS = SSA + SSB + SSAB + SSE,$$

όπου:

$$SSA = q * r * \sum_{i=1}^{p} (X_{i..} - \bar{X})^2$$

$$SSB = p * r * \sum_{j=1}^{q} (\bar{X}_{.j.} - \bar{\bar{X}})^2$$

$$SSAB = r * \sum_{i=1}^{p} \sum_{j=1}^{q} (\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{\bar{X}})^2$$

$$SSE = \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{k=1}^{r} (X_{ijk} - \bar{X}_{ij.})^2$$

Απλουστεύουμε τα αθροίσματα και τα συγκεντρώνουμε στον πίνακα ανάλυσης της διασποράς, όπου φαίνονται επίσης και οι βαθμοί ελευθερίας των αθροισμάτων όπως και τα στατιστικά F που ελέγχουν τη σημαντικότητα των επιδράσεων κάθε πηγής.

Οι υποθέσεις που μπορούμε να ελέγξουμε, το αντίστοιχο στατιστικό και η αντίστοιχη κρίσιμη τιμή της κατανομής F_{mn} είναι:

Πηγή	**Άθροισμα τετραγώνων**	**Β.Ε**	**Μ.Τ**	**F**
Παράγοντας Α (γραμμές)	$SSA = \frac{1}{qr}\sum_{i=1}^{p} x_{i..}^2 - \frac{1}{n}x_{...}^2$	p-1	$MSA = \frac{SSA}{p-1}$	$\frac{MSA}{MSE}$
Παράγοντας Β (στήλες)	$SSB = \frac{1}{pr}\sum_{j=1}^{q} x_{.j.}^2 - \frac{1}{n}x_{...}^2$	q-1	$MSB = \frac{SSB}{q-1}$	$\frac{MSB}{MSE}$
Αλληλεπίδραση ΑxΒ (κελιά)	$SSAB = \frac{1}{r}\sum_{i=1}^{p}\sum_{j=1}^{q} x_{ij.}^2 - SSA - SSB - \frac{1}{n}x_{...}^2$	$(p-1)(q-1$	$MSAB = \frac{SSAB}{(p-1)(q-1)}$	$\frac{MSA}{MSE}$
Υπόλοιπα	SSE (με αφαίρεση)	Pq (r-	$MSE = s^2 = SSE / pq(r-1)$	

		1)		
Σύνολο	$TSS = \sum_{i=1}^{p} \sum_{j=1}^{q} \sum_{k=1}^{r} x_{ijk}^2 - \frac{1}{n} X_{...}^2$	pqr-1=n-1		

Ανάλυση Διασποράς με δύο παράγοντες και r παρατηρήσεις σε κάθε κελί **(Πίνακας 5)**

1. Ότι οι κύριες επιδράσεις του παράγοντα A είναι ίσες (οπότε λόγω και της ισότητας 4.28 θα είναι και ίσες με 0), δηλ. $H_0 : a_1 = a_2 = ... = a_p = 0.$ Υπολογίζουμε το στατιστικό F=MSA/MSE και το συγκρίνουμε με την κρίσιμη τιμή $F_{p-1,\, pq(r-1),\, a}$.

2. Ότι οι κύριες επιδράσεις του παράγοντα B είναι ίσες δηλ. $H_0 : b_1 = b_2 = ... = b_q = 0,$ οπότε το στατιστικό είναι F=MSB/MSE και η κρίσιμη τιμή $F_{q-1,\, pq(r-1),\, a}$.

3. Ότι οι αλληλεπιδράσεις των παραγόντων A, B είναι ίσες δηλ. $H_0 : (ab)_{ij} = 0,$ i=1, 2, …,p, j=1, 2, …,q. Το στατιστικό εδώ είναι F=MSAB/MSE και η κρίσιμη τιμή $F_{(p-1)(q-1),\, pq(r-1),\, a}$.

Από τον παραπάνω πίνακα είναι φανερό ότι το s^2 μπορεί να εκτιμηθεί μόνο αν το r>1, όταν δηλ. υπάρχουν επαναλαμβανόμενες μετρήσεις.

Αν το r=1, αν δηλ. έχουμε μόνο μια παρατήρηση σε κάθε κελί, τότε ο πίνακας ανάλυσης της διασποράς περιορίζεται, αφού δε μπορεί τότε να ελεγχθεί η ύπαρξη αλληλεπίδρασης. Στην περίπτωση αυτή το TSS αναλύεται στο SSA, SSB και SSE.

4.1 Παράδειγμα στο SPSS ανάλυσης διακύμανσης κατά δύο παράγοντες

Στο SPSS η ανάλυση διακύμανσης κατά δύο παράγοντες γίνεται μέσω της γενικής διαδικασίας GLM Univariate.

Analyse…..

General Linear Model…..

Univariate..

Θέλουμε να ελέγξουμε αν ο μισθός είναι συνάρτηση του φύλου και των χρόνων εμπειρίας.

Between-Subjects Factors

		Value Label	N
job experience	1,0		9
	2,0		9
	3,0		20
	4,0		20
	5,0		9
	6,0		9
sex	0	female	33
	1	male	43

Ο πρώτος πίνακας περιέχει κάποια περιγραφικά στατιστικά για το πόσοι είναι οι άνδρες και οι γυναίκες και την κατανομή των ετών εμπειρίας. Παρατηρούμε ότι οι άνδρες στην αγορά εργασίας είναι περισσότεροι και ότι οι περισσότεροι υπάλληλοι έχουν 3-4 χρόνια εμπειρίας

Descriptive Statistics

Dependent Variable:salary (x 1000) / year

job experience	sex	Mean	Std. Deviation	N
1,0	female	20,876	4,0428	5
	male	23,025	3,5732	4
	Total	21,831	3,7739	9
2,0	female	29,142	3,6471	6
	male	30,656	8,3539	3
	Total	29,646	5,1316	9
3,0	female	39,321	4,2133	8
	male	40,322	3,5137	12
	Total	39,922	3,7338	20
4,0	female	49,847	4,8632	8
	male	53,824	4,3890	12
	Total	52,233	4,8848	20
5,0	female	54,267	2,2431	3
	male	60,427	3,1407	6
	Total	58,374	4,1123	9
6,0	female	66,229	3,8566	3
	male	72,166	4,7250	6
	Total	70,187	5,1465	9
Total	female	41,032	14,1599	33
	male	49,055	15,1037	43
	Total	45,572	15,1434	76

Στον δεύτερο πίνακα βλέπουμε ότι οι άνδρες αμείβονται καλύτερα από τις γυναίκες κατά μέσο όρο ανεξαρτήτως από τα χρόνια εμπειρίας.

Levene's Test of Equality of Error Variances[a]

Dependent Variable:salary (x 1000) / year

F	df1	df2	Sig.
,816	11	64	,625

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Design: Intercept + exp + male + exp * male

Στον πίνακα 3 βλέπουμε ότι απορρίπτεται η Ηο ότι η διακύμανση του σταθερού όρου είναι ίση σε όλες τις ομάδες

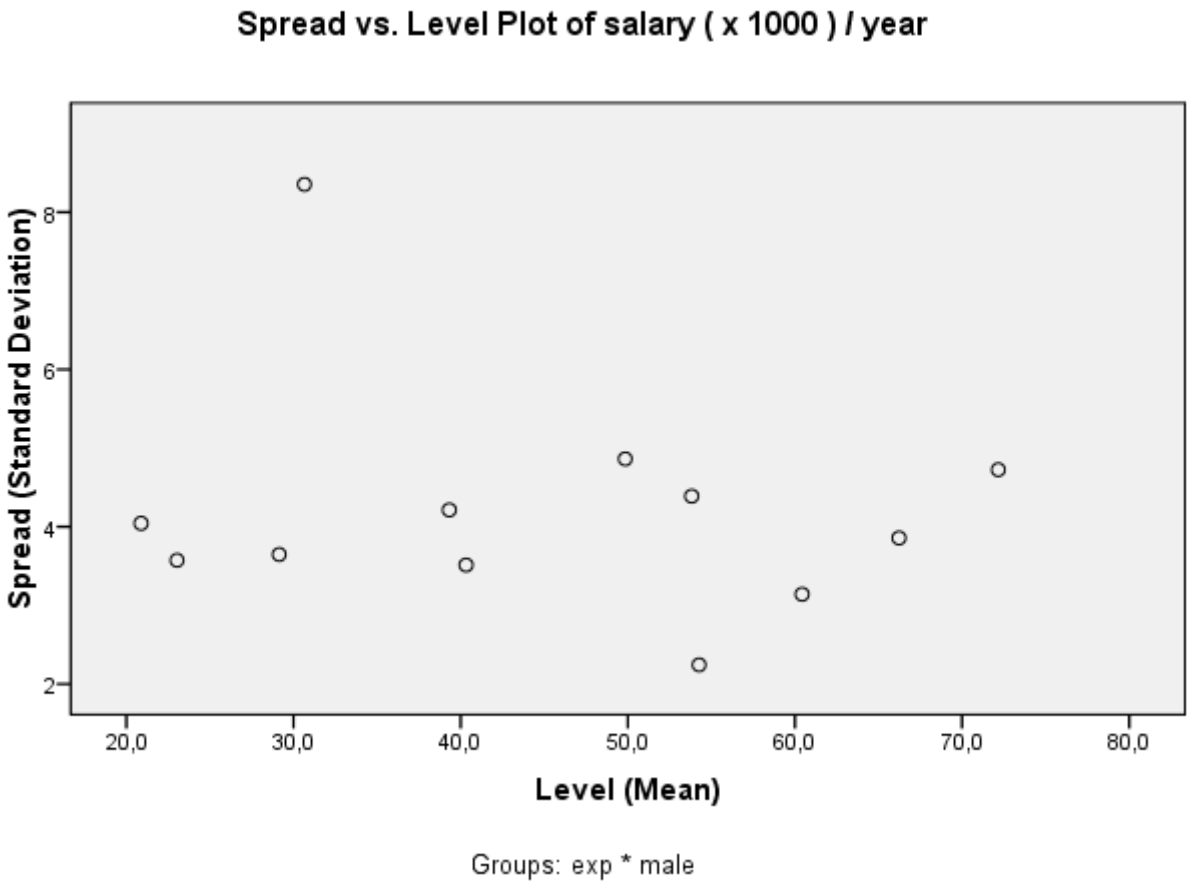
Estimated Marginal Means

job experience * sex

Dependent Variable:salary (x 1000) / year

job experience	sex	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
1,0	female	20,876	1,894	17,093	24,660
	male	23,025	2,118	18,795	27,256
2,0	female	29,142	1,729	25,687	32,596
	male	30,656	2,445	25,771	35,541
3,0	female	39,321	1,497	36,329	42,312
	male	40,322	1,223	37,880	42,765
4,0	female	49,847	1,497	46,855	52,838
	male	53,824	1,223	51,381	56,266
5,0	female	54,267	2,445	49,382	59,152
	male	60,427	1,729	56,973	63,881
6,0	female	66,229	2,445	61,344	71,114
	male	72,166	1,729	68,712	75,621

Σε αυτό το πίνακα μπορούμε να δούμε αν υπάρχει συσχέτιση μεταξύ του φύλου και των χρόνων εμπειρίας σε σχέση με το εισόδημα.



Στο διάγραμμα αυτό μπορούμε να δούμε οπτικά αν οι διακυμάνσεις μεταξύ των ομάδων είναι ίσες

Tests of Between-Subjects Effects

Dependent Variable:salary (x 1000) / year

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	16051,106[a]	11	1459,191	81,346	,000	,933
Intercept	123257,273	1	123257,273	6871,291	,000	,991
exp	13862,309	5	2772,462	154,558	,000	,924
male	181,745	1	181,745	10,132	,002	,137
exp * male	66,691	5	13,338	,744	,594	,055
Error	1148,033	64	17,938			
Total	175033,026	76				
Corrected Total	17199,138	75				

a. R Squared = ,933 (Adjusted R Squared = ,922)

Παρατηρούμε ότι το Significance είναι κάτω από το 0,05 αυτό σημαίνει ότι το φύλο και τα έτη εμπειρίας επηρεάζουν στατιστικά σημαντικά το μισθό.

ΚΕΦΑΛΑΙΟ 5^{ο}

5.0 Κλασσικό γραμμικό υπόδειγμα

5.1 Το απλό γραμμικό υπόδειγμα

Το απλό παλινδρομικό μοντέλο εκφράζει τη σχέση ανάμεσα σε μια μεταβλητή Υ και μια μεταβλητή Χ, σε μορφή ευθείας γραμμής. Η γραμμή παλινδρόμησης στον πληθυσμό της Υ με τη Χ είναι η συσχέτιση των δύο μεταβλητών. Η παλινδρόμηση προϋποθέτει ότι ικανοποιείται το σύνολο των υποθέσεων στον πληθυσμό στον οποίο ανήκουν οι δύο μεταβλητές, προκειμένου να χρησιμοποιηθεί γραμμικό παλινδρομικό μοντέλο.

Η **πρώτη υπόθεση** είναι ότι η σχέση περιγράφεται από μια ευθεία γραμμή, δηλαδή είναι *γραμμική*.

Από μαθηματική άποψη το γραμμικό παλινδρομικό μοντέλο μπορεί να εκφραστεί ως εξής:

$$Y = b_0 + b_1 X + e$$

όπου στην εξίσωση το e είναι το τυχαίο σφάλμα που μετράει την κάθετη απόκλιση της κάθε τιμής της εξαρτημένης μεταβλητής από την πληθυσμιακή παλινδρομική γραμμή για την αντίστοιχη τιμή της ανεξάρτητης μεταβλητής.

Η **δεύτερη υπόθεση** του μοντέλου είναι ότι θα εμφανιστεί η ίδια κατανομή των τιμών της Χ σε οποιαδήποτε επανάληψη του πειράματος. Όμως, η κατανομή των τιμών της Υ για δεδομένες τιμές της Χ μπορεί να διαφέρει από πείραμα σε πείραμα εξαιτίας της επίδρασης του όρου του τυχαίου σφάλματος. Οι τιμές της μεταβλητής Χ, λέμε ότι έχουν προκαθοριστεί, ενώ η εξαρτημένη μεταβλητή Υ, είναι μια τυχαία μεταβλητή.

Η **τρίτη υπόθεση** είναι ότι η αναμενόμενη τιμή του τυχαίου σφάλματος είναι μηδέν και εκφράζεται ως εξής:

$$E(e)=0$$

Δηλαδή κατά μέσο όρο η τιμή του τυχαίου σφάλματος είναι ίση με το μηδέν. Για να κατανοήσουμε την παραδοχή αυτή θεωρούμε την αναμενόμενη τιμή και των δύο μερών της εξίσωσης του υποδείγματος για δεδομένη τιμή της Χ. Η αναμενόμενη τιμή του αθροίσματος:

$$E(Y/X) = E(b_0 + b_1X + e)$$

μπορεί να αναχθεί στο άθροισμα των αναμενόμενων τιμών:

$$E(Y/X) = E(b_0) + E(b_1X) + E(e)$$

b_0 και b_1 είναι σταθερές και η X είναι δεδομένη, έτσι ώστε να μπορούμε να τη μεταχειριστούμε ως σταθερά. Επομένως, η εξίσωση γίνεται:

$$E(Y/X) = b_0 + b_1X$$

εφόσον η αναμενόμενη τιμή μιας σταθεράς είναι μια σταθερά και εφόσον E(e) είναι μηδέν. Αυτή η προσδοκία είναι ο υπό συνθήκη μέσος γιατί μετράει τον μέσο ή τη μέση τιμή της Y που σχετίζεται με μια συγκεκριμένη τιμή της X. Η τρίτη υπόθεση συνεπάγεται ότι για μια δεδομένη τιμή της X, ο μέσος των τιμών Y βρίσκεται πάνω στη γραμμή παλινδρόμησης.

Η **τέταρτη υπόθεση** είναι ότι η διακύμανση του όρου του τυχαίου σφάλματος είναι η ίδια για κάθε τιμή του X. Η υπόθεση αυτή ονομάζεται ομοσκεδαστικότητα:

$$VAR(e) = E\left[e - E(e)\right]^2 = E(e^2) = R$$

όπου R είναι ένας σταθερός όρος. Αλλά από το μοντέλο έχουμε:

$$e = Y - (b_0 + b_1X)$$

Αυτή η υπόθεση, παραπέρα, εισάγει ότι $E\left[Y - (b_0 + b_1X)\right]^2 = R,$ ή ότι η διακύμανση και η σταθερή απόκλιση είναι ίδιες για κάθε τιμή X στον πληθυσμό. Η VAR(e) μετράει τη μεταβλητότητα στην τιμή της εξαρτημένης μεταβλητής, Y, περί τη γραμμή παλινδρόμησης για δεδομένες τιμές της ανεξάρτητης μεταβλητής, X.

Η **πέμπτη υπόθεση** είναι ότι οι τιμές του e είναι ανεξάρτητες μεταξύ τους, δηλαδή $E(e_ie_j) = 0$. Αυτό σημαίνει ότι η τιμή του σφάλματος για οποιαδήποτε δεδομένη τιμή της ανεξάρτητης μεταβλητής δε σχετίζεται με το σφάλμα για οποιαδήποτε άλλη τιμή της X. Όταν τα σφάλματα αυτά σχετίζονται, τότε έχουμε ο πρόβλημα της αυτοσυσχέτισης.

Η **έκτη υπόθεση** είναι ότι ο όρος του σφάλματος για κάθε τιμή της Χ κατανέμεται κανονικά. Από τη στιγμή που οι τιμές της Υ για δεδομένες τιμές της Χ ποικίλουν από πείραμα σε πείραμα μόνον εξαιτίας του όρου του τυχαίου σφάλματος (οι τιμές του Χ μένουν ίδιες από πείραμα σε πείραμα), οι τιμές της Υ πρέπει αναγκαστικά να κατανέμονται κανονικά.

5.2 Εκτίμηση του απλού γραμμικού υποδείγματος

Παρατηρήσεις σε ένα δείγμα b_0 τιμών του Υ για ένα καθορισμένο δείγμα τιμών της Χ μας επιτρέπει να υπολογίσουμε την εξίσωση μιας γραμμής της ακόλουθης μορφής

$$\hat{Y} = b_0 + b_1 X$$

όπου b_0, b_1 είναι εκτιμήσεις των παραμέτρων b_0 και b_1 της παλινδρόμησης. $\hat{Y}$ είναι η εκτιμημένη τιμή του m_Y, ο υπό συνθήκη μέσος της Υ για δοσμένο Χ. Για να βρούμε τα $b_0, b_1,$ για ένα δοσμένο σύνολο τιμών του δείγματος, μπορούμε να χρησιμοποιήσουμε τους παρακάτω τύπους:

$$b_1 = \frac{n\Sigma XY - (\Sigma X)(\Sigma X)}{n\Sigma X^2 - (\Sigma X)^2}$$

και

$$b_0 = \frac{(\Sigma Y)(\Sigma X^2) - (\Sigma X)(\Sigma XY)^1}{n(\Sigma X^2) - (\Sigma X)^2}$$

Συνήθως χρησιμοποιείται ένας πιο απλός τύπος για το b_0. Ο τύπος αυτός είναι:

$$b_0 = \frac{\Sigma Y - b_1 \Sigma X}{n}$$

5.3 Το γενικό γραμμικό υπόδειγμα

Το υπόδειγμα της απλής γραμμικής παλινδρόμησης, που αναπτύξαμε στην προηγούμενη παράγραφο, αναφέρεται σε σχέσεις που περιλαμβάνουν μία μόνο ερμηνευτική μεταβλητή. Η συμπεριφορά όμως των περισσότερων

οικονομικών μεταβλητών είναι συνάρτηση όχι μιας αλλά πολλών μεταβλητών. Έστω ότι $Y = f(X_1, X_2, ..., X_k)$, δηλαδή η Υ είναι συνάρτηση των Κ ερμηνευτικών μεταβλητών $X_1, X_2, ..., X_k$. Αν υποθέσουμε ότι η συναρτησιακή σχέση $Y = f(X_1, X_2, ..., X_k)$ είναι γραμμική, για ένα δείγμα από Τ παρατηρήσεις, μπορούμε να γράψουμε:

$$Y_t = b_0 + b_1 X_{t1} + b_2 X_{t2} + ... + b_k X_{tk} + u_t$$

όπου X_{t1} είναι η t παρατήρηση της ερμηνευτικές μεταβλητής X_1, X_{t2} ή t παρατήρηση της ερμηνευτικής μεταβλητής X_2 κ.ο.κ. Ο πρώτος, δηλαδή, δείκτης αναφέρεται στην παρατήρηση και ο δεύτερος στην ερμηνευτική μεταβλητή. Η παραπάνω σχέση αποτελεί το υπόδειγμα της γραμμικής πολυμεταβλητής παλινδρόμησης, που είναι επέκταση της απλής παλινδρόμησης για περισσότερες από μια ερμηνευτικές μεταβλητές. Για k=1, η σχέση γίνεται το υπόδειγμα της απλής γραμμικής παλινδρόμησης.

Οι βασικές υποθέσεις που συνιστούν το κλασσικό γραμμικό υπόδειγμα στη γενική του μορφή, δηλαδή με k ερμηνευτικές μεταβλητές, είναι σχεδόν οι ίδιες με τις υποθέσεις για το διμεταβλητό γραμμικό υπόδειγμα. Οι υποθέσεις αυτές, που πρέπει να ισχύουν για όλες τις παρατηρήσεις, είναι οι ακόλουθες:

$Y_t = b_0 + b_1 X_{t1} + b_2 X_{t2} + ... + b_k X_{tk} + u_t$ (1)
$u_t \sim (0, s^2)$ (2) α) u_t είναι τυχαία μεταβλητή β) $Eu_t = 0$ γ) $Eu_t^2 = s^2$
$Eu_t u_s = 0$ για $t \neq s$ (3)
Οι ερμηνευτικές μεταβλητές δεν είναι στοχαστικές. Οι τιμές τους παραμέ- (4) νουν σταθερές και δεν είναι όλες ίσες μεταξύ τους.

Δεν υπάρχουν ακριβείς γραμμικές σχέσεις ανάμεσα στις ερμηνευτικές μεταβλητές.	(5)
Ο αριθμός των παρατηρήσεων του δείγματος είναι μεγαλύτερος από τον αριθμό των συντελεστών του υποδείγματος που θέλουμε να εκτιμήσουμε.	(6)

Η υπόθεση (1) αναφέρεται στη γραμμική σχέση που συνδέει τις μεταβλητές Y και $X_1, X_2, ..., X_k$. Κάθε τιμή t της εξαρτημένης μεταβλητής είναι γραμμική συνάρτηση των τιμών των ερμηνευτικών μεταβλητών $X_{t1}, X_{t2}, ..., X_{tk}$ και του διαταρακτικού όρου u_t. Οι υποθέσεις (2), (3) και (4) είναι αντίστοιχες με αυτές στο απλό γραμμικό υπόδειγμα. Οι πρόσθετες υποθέσεις (5) και (6) έχουν σχέση με την εκτίμηση και τον έλεγχο του υποδείγματος. Η υπόθεση (5), όπως θα φανεί αργότερα, αποτελεί προϋπόθεση για την εκτίμηση του υποδείγματος και αποκλείει την ύπαρξη *τέλειας πολυσυγραμμικότητας* μεταξύ των ερμηνευτικών μεταβλητών. Αυτό σημαίνει πως καμιά από τις k ερμηνευτικές μεταβλητές δε μπορεί να εκφραστεί ως γραμμικός συνδυασμός των υπολοίπων. Τέλος, η υπόθεση (6) εξασφαλίζει τους απαραίτητους βαθμούς ελευθερίας και για την εκτίμηση αλλά και για τον έλεγχο του υποδείγματος. Ο αριθμός των παρατηρήσεων πρέπει να είναι τουλάχιστον ίσος με τους συντελεστές του υποδείγματος, για να είναι δυνατή η εκτίμησή του. Πρέπει όμως να είναι και μεγαλύτερος, για να είναι δυνατός ο έλεγχος υποθέσεων με τις διάφορες στατιστικές ελέγχου που η κατανομή τους εξαρτάται από τους βαθμούς ελευθερίας, όπως η κατανομή t ή η κατανομή F.

Σύμφωνα με τις προηγούμενες υποθέσεις και γενικεύοντας την ανάλυση της απλής παλινδρόμησης, είναι εύκολο να δειχθεί πως ο μέσος και η διακύμανση της Y_t δίνονται από τις ακόλουθες σχέσεις:

$$E(Y_t) = b_0 + b_1 X_{t1} + b_2 X_{t2} + ... + b_k X_{tk} \quad (7)$$

$$V(Y_t) = s^2 \quad (8)$$

Η σχέση (7), δηλαδή η σχέση που υπάρχει ανάμεσα στους μέσους της Υ και τις τιμές των ερμηνευτικών μεταβλητών, *είναι η παλινδρόμηση στον πληθυσμό.*

Στην απλή παλινδρόμηση, ο συντελεστής b_1 της ερμηνευτικής μεταβλητής Χ παριστάνει τη μεταβολή στη μέση τιμή της Υ, όταν η Χ μεταβάλλεται κατά μία μονάδα, δηλαδή $b_1 = \frac{dE(Y_t)}{dX_t}$. Στην πολυμεταβλητή παλινδρόμηση, ο συντελεστής b_j, για j=1, 2, …, k, παριστάνει τη μεταβολή στη μέση τιμή της Υ, όταν η X_j μεταβάλλεται κατά μία μονάδα και οι υπόλοιπες ερμηνευτικές μεταβλητές παραμένουν σταθερές, δηλαδή $b_j = \frac{\partial E(Y_t)}{\partial X_{tj}}$.

Στην πολυμεταβλητή παλινδρόμηση επομένως, ο συντελεστής b_j είναι η μερική παράγωγος της μέσης τιμής Ε(Υ), ως προς X_j, γι' αυτό και οι συντελεστές $b_1, b_2, ..., b_k$ ονομάζονται και μερικοί συντελεστές παλινδρόμησης (partial regression coefficients), σε αντίθεση προς το συντελεστή b_1 στην απλή παλινδρόμηση, που μπορεί να θεωρηθεί ως η συνολική παράγωγος της μέσης τιμής, Ε(Υ), ως προς Χ.

5.4 Ισοδυναμία ανάλυσης διασποράς και παλινδρόμησης

Ας θεωρήσουμε πάλι ότι έχουμε ένα παράγοντα με k στάθμες, ή k διαφορετικές ομάδες που δέχονται την ίδια μεταχείριση κ.λπ. Χρησιμοποιώντας τότε βωβές μεταβλητές, μπορούμε αντί να κάνουμε ανάλυση διασποράς να κάνουμε παλινδρόμηση, η οποία θα δείξουμε οδηγεί στα ίδια συμπεράσματα.

Πράγματι έστω οι μεταβλητές $X_1, X_2, ..., X_k$, που ορίζονται από τις σχέσεις

$$X_j = \begin{cases} 1, \_an \_h \_paratήrhsh \_gínetai \\ \_\_sth \_stάqmh \_(ή \_omάda) \_i \\ 0, \_allού, \end{cases}$$

όταν i=1, 2, …, k. Προσαρμόζουμε τότε στα δεδομένα μας το μοντέλο

$$Y_{ij} = b_0 + b_1 X_1 + b_2 X_2 + ... + b_k X_k + e_{ij}$$

Οι συντελεστές παλινδρόμησης στο νέο μοντέλο, συνδέονται με τις παραμέτρους a_i και μ του μοντέλου $Y_{ij} = m + a_i + e_{ij}$ και μάλιστα είναι εύκολο να δούμε ότι

$$m + a_i = b_0 + b_i \qquad \text{για} \quad i = 1, 2, ..., k$$

Το παραπάνω μοντέλο μπορεί να γραφτεί και με τη χρήση πίνακα ως:

$$\underset{\%}{Y} = X \underset{\%}{b} + \underset{\%}{e}$$

όπου: $\underset{\%}{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ \text{M} \\ Y_{1r_1} \\ Y_{21} \\ \text{M} \\ Y_{2r_2} \\ \text{M} \\ Y_{k_1} \\ \text{M} \\ Y_{kr_k} \end{bmatrix}$, $X = \begin{bmatrix} 1 & 1 & 0 & 0 & \mathbf{K} & 0 \\ 1 & 1 & 0 & 0 & \mathbf{K} & 0 \\ \text{M} & \text{M} & \text{M} & \text{M} & & \text{M} \\ 1 & 1 & 0 & 0 & \mathbf{K} & 0 \\ 1 & 0 & 1 & 0 & \mathbf{K} & 0 \\ 1 & 0 & 1 & 0 & \mathbf{K} & 0 \\ \text{M} & \text{M} & \text{M} & \text{M} & & \text{M} \\ 1 & 0 & 1 & 0 & \mathbf{K} & 0 \\ 1 & 0 & 0 & 0 & \mathbf{K} & 1 \\ 1 & 0 & 0 & 0 & \mathbf{K} & 1 \\ & & & & & \\ 1 & 0 & 0 & 0 & \mathbf{K} & 1 \end{bmatrix}$, $\underset{\%}{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \text{M} \\ b_{k-1} \end{bmatrix}$, $\underset{\%}{e} = \begin{bmatrix} e_{11} \\ e_{12} \\ e_{1r_1} \\ e_{21} \\ \text{M} \\ e_{2r_2} \\ \text{M} \\ e_{k_1} \\ \text{M} \\ e_{kr_k} \end{bmatrix}$

Έχουμε τότε:

$$\text{X}'\text{X} = \begin{bmatrix} n & \text{r}_1 & \text{r}_2 & \mathbf{K} & r_k \\ r_1 & \text{r}_1 & 0 & \mathbf{K} & 0 \\ r_2 & 0 & \text{r}_2 & \mathbf{K} & 0 \\ \text{M} & \text{M} & \text{M} & & \text{M} \\ r_k & 0 & 0 & \mathbf{K} & \text{r}_k \end{bmatrix} \qquad \text{και} \qquad \text{X}'\underset{\%}{Y} = \begin{bmatrix} Y_{..} \\ Y_{1.} \\ Y_{2.} \\ \text{M} \\ Y_{k.} \end{bmatrix}$$

Ο πίνακας X΄X δεν έχει αντίστροφο γιατί το άθροισμα όλων των γραμμών του εκτός της πρώτης ισούται με την πρώτη γραμμή. Έτσι το σύστημα των κανονικών εξισώσεων που γράφεται

$$(X'X)\underset{\sim}{b} = X'\underset{\sim}{Y}$$

δε μπορεί να λυθεί. Αυτό εξάλλου ήταν αναμενόμενο και από το γεγονός ότι ο πίνακας X έχει ισχυρή πολυσυγραμμικότητα, αφού από τον ορισμό προκύπτει

$$X_1 + X_2 + ... + X_k = 1.$$

Οι δύο παραπάνω σχέσεις γράφονται,

$$\begin{aligned} &nb_0 + r_1 b_1 + r_2 b_2 + ... + r_k b_k = Y_{..} \\ &r_1 b_0 + r_1 b_1 = Y_{1.} \\ &r_2 b_0 + r_2 b_2 = Y_{2.} \qquad (4) \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &r_k b_0 + r_k b_k = Y_{k.} \end{aligned}$$

Ένας τρόπος να ξεπεράσουμε το αδιέξοδο είναι να υποθέσουμε ότι οι παράμετροι b_i, ικανοποιούν και κάποια άλλη σχέση.

Έτσι υποθέτουμε ότι ικανοποιείται η σχέση

$$r_1 b_1 + r_2 b_2 + ... + r_k b_k = 0, \qquad (4.1)$$

Τότε η πρώτη από τις σχέσεις (4) δίνει

$$\hat{b_0} = \frac{Y_{..}}{n} = \bar{Y}$$

ενώ οι υπόλοιπες δίνουν

$$\hat{b_i} = \frac{Y_{1.}}{r_1} - \bar{Y} = \bar{Y}_{1.} - \bar{Y}$$

ώστε $\hat{b} = (\bar{Y},\ \bar{Y}_{1.} - \bar{Y},\ ...,\ \bar{Y}_{k.} - \bar{Y})'$.

Το άθροισμα των τετραγώνων SSR που θα οφείλεται στην παλινδρόμηση θα είναι τώρα

$$SSR = \underset{\sim}{\hat{b}}'\mathrm{X}'\underset{\sim}{\mathrm{Y}} - \frac{1}{n}(\underset{\sim}{\mathrm{Y}}'\underset{\sim}{1})^2 =$$

$$= \bar{\mathrm{Y}}\mathrm{Y}_{..} + \sum_{i=1}^{k}(\bar{\mathrm{Y}}_{i.} - \bar{\mathrm{Y}})\mathrm{Y}_{i.} - \frac{1}{n}(\mathrm{Y}_{..})^2 \qquad (4.2)$$

ή

$$\mathrm{SSR} = \sum_{i=1}^{k}\frac{\mathrm{Y}_{i.}^2}{r_i} - \frac{\mathrm{Y}_{..}^2}{n}$$

Μπορεί να δειχθεί ότι το άθροισμα των τετραγώνων SSR μένει αμετάβλητο αν αντί της σχέσης 4.1 είχαμε υποθέσει κάποια άλλη γραμμική σχέση μεταξύ των συντελεστών b_i. Ακόμη θα μπορούσαμε να αντιμετωπίσουμε το πρόβλημα χρησιμοποιώντας γενικευμένους αντίστροφους πίνακες. Πράγματι μπορούμε εύκολα να διαπιστώσουμε ότι ο πίνακας

$$\mathrm{A} = \begin{bmatrix} 0 & & & \\ & \mathrm{r}_1^{-1} & (0) & \\ & & \mathrm{r}_2^{-1} & \\ (0) & & \mathbf{O} & \\ & & & \mathrm{r}_k^{-1} \end{bmatrix},$$

ικανοποιεί τη σχέση (X΄X)A(X΄X)=X΄X. Είναι επομένως $\mathrm{A} = (\mathrm{X}'\mathrm{X})^{-}$, είναι δηλαδή ένας γενικευμένος αντίστροφος του X΄X. Τότε θα είναι

$$\underset{\sim}{\hat{b}} = (\mathrm{X}'\mathrm{X})^{-}\mathrm{X}'\underset{\sim}{Y}$$

που δίνει $\underset{\sim}{\hat{b}} = (0,\ \bar{\mathrm{Y}}_{1.},\ ...,\ \bar{\mathrm{Y}}_{k.})$

διαφορετικό από το προηγούμενο $\underset{\sim}{\hat{b}}$. Για τον υπολογισμό όμως του SSR έχουμε

$$SSR = \underset{\sim}{\hat{b}}'\mathrm{X}'\underset{\sim}{Y} - \frac{1}{n}(\underset{\sim}{\mathrm{Y}}'\underset{\sim}{1})^2 =$$

$$= \sum_{i=1}^{k}\frac{\mathrm{Y}_{i.}^2}{r_i} - \frac{\mathrm{Y}_{..}^2}{n}$$

το ίδιο ακριβώς που βρήκαμε στην (4.2).

Επειδή βεβαίως το συνολικό άθροισμα τετραγώνων SST, δεν εξαρτάται από τη μέθοδο, προκύπτει ότι οι δύο μέθοδοι είναι ισοδύναμες.

Μια άλλη προσέγγιση του προβλήματος με παλινδρόμηση, που δεν απαιτεί την εύρεση γενικευμένων αντιστρόφων είναι η παρακάτω:

Έστω ότι έχουμε ένα μοντέλο $Y_{ij} = m + a_i + e_{ij}$. Θέτουμε $b_i = m + a_i$, οπότε το μοντέλο γράφεται:

$$\underset{\sim}{Y} = X\underset{\sim}{b} + \underset{\sim}{e} \qquad (4.3)$$

όπου $\underset{\sim}{Y}$ το διάνυσμα των παρατηρήσεων όπως ορίστηκε στην αρχή της παραγράφου, και X, ο πίνακας που προκύπτει από τον προηγούμενο X χωρίς την πρώτη στήλη του, δηλαδή

$$X = \begin{bmatrix} \underset{\sim}{1}_{r_1} & & & \\ & 0 & & \\ & & \underset{\sim}{1}_{r_2} & \\ 0 & & \mathbf{O} & \\ & & & \underset{\sim}{1}_{r_k} \end{bmatrix} \qquad (4.4)$$

Τότε

$$X'X = diag(r_1,\ r_2,\ ...,\ r_k)$$
$$X'\underset{\sim}{Y} = (Y_{1.},\ Y_{2.},\ ...,\ Y_{k.})$$

και άρα

$$\underset{\sim}{\hat{b}} = (X'X)^{-1}X'\underset{\sim}{Y} = (\bar{Y}_{1.},\ \bar{Y}_{2.},\ ...,\ \bar{Y}_{k.})$$

οπότε και

$$SSR = \underset{\sim}{\hat{b}}'X'\underset{\sim}{Y} - \frac{1}{n}(\underset{\sim}{Y}'\underset{\sim}{1})^2 = \sum_{i=1}^{k}\frac{Y_{i.}^2}{r_i} - \frac{Y_{..}^2}{n}$$

το ίδιο ακριβώς που βρήκαμε και με τις προηγούμενες μεθόδους.

Ένας ακόμη τρόπος προσέγγισης του προβλήματος με παλινδρόμηση είναι να χρησιμοποιήσουμε k-1 μόνο βωβές μεταβλητές.

5.5 Παράδειγμα στο SPSS στην πολλαπλή παλινδρόμηση

Από το μενού

Analyse

Regression

Linear

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,960[a]	,922	,920	4,2806

a. Predictors: (Constant), job experience, sex

b. Dependent Variable: salary (x 1000) / year

Στον παραπάνω πίνακα παρατηρούμε ότι το R^2 είναι πολύ υψηλό, αυτό σημαίνει ότι η ευθεία της παλινδρόμησης ερμηνεύει την μεταβλητή μισθό κατά 96%, δηλαδή οι μεταβλητές που έχουμε επιλέξει ερμηνεύουν σε μεγάλο βαθμό τον μισθό ενός εργαζομένου.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	15861,503	2	7930,752	432,812	,000[a]
	Residual	1337,635	73	18,324		
	Total	17199,138	75			

a. Predictors: (Constant), job experience, sex

b. Dependent Variable: salary (x 1000) / year

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	10,108	1,323		7,640	,000
	sex	3,125	1,006	,103	3,107	,003
	job experience	9,627	,340	,937	28,285	,000

a. Dependent Variable: salary (x 1000) / year

Παρατηρούμε ότι και οι δύο συντελεστές έχουν σ μικρό Significance, αυτό σημαίνει ότι είναι στατιστικά σημαντικοί κατά 95%.

Η ευθεία της παλινδρόμησης που προκύπτει είναι:

salary (x 1000) / year = 10,108 + 3,125*sex + 9,627* job experience

Αυτό που παρατηρούμε είναι ότι ένας άνδρας αμείβεται 3.125 $ παραπάνω από μία γυναίκα και κάθε εργαζόμενος με ένα επιπλέον έτος προϋπηρεσίας 9.627 $ παραπάνω.

ΚΕΦΑΛΑΙΟ 6^{ο}

6.0 Εμπειρική εφαρμογή της θεωρίας της ανάλυσης διακύμανσης

6.1 Λίγα λόγια για τον Συνεταιρισμό

Η εμπειρική εφαρμογή της θεωρίας της ανάλυσης της διακύμανσης θα πραγματοποιηθεί στην Παναιγιάλειο Ένωση Συνεταιρισμών η οποία μετέχει σε τρεις παραγωγικές δραστηριότητες: επεξεργασία και συσκευασία σταφίδας, συσκευασία εσπεριδοειδών και τυποποίηση ελαιολάδου.

Λίγα λόγια για την Παναιγιάλειο Ένωση Συνεταιρισμών (Π.Ε.Σ.) ιδρύθηκε το 1935 στο Αίγιο Αχαΐας, μια όμορφη παραλιακή πόλη του Κορινθιακού κόλπου στα βορειοδυτικά της Πελοποννήσου. Αποτελείται από 59 πρωτοβάθμιους αγροτικούς συνεταιρισμούς, με σύνολο 6.000 περίπου ενεργών μελών.

Η εμπορική δραστηριότητα της Π.Ε.Σ. εξαπλώνεται σε όλον τον κόσμο, με ετήσιο κύκλο εργασιών που υπερβαίνει τα €15 εκ. Η Π.Ε.Σ. είναι ο μεγαλύτερος εξαγωγέας Κορινθιακής σταφίδας τύπου "ΒΟΣΤΙΤΣΑ" ("VOSTIZZA"), η οποία θεωρείται ως η κορυφαία ποιότητα μαύρης Κορινθιακής σταφίδας και έχει καταχωρηθεί (από το 1993) ως Προϊόν Προστατευόμενης Ονομασίας Προέλευσης (ΠΟΠ). Κάθε χρόνο, η Π.Ε.Σ. διαχειρίζεται άνω του 80% της σταφίδας της ευρύτερης περιοχής της Αιγιαλείας.

Οι σπουδαιότερες δραστηριότητες της Π.Ε.Σ. είναι οι παρακάτω:

- Επεξεργασία και εξαγωγή Κορινθιακής σταφίδας "ΒΟΣΤΙΤΣΑ" ("VOSTIZZA")
- Επεξεργασία και εξαγωγή παρθένου ελαιολάδου "ΕΛΙΚΗ"
- Συσκευασία και εξαγωγή εσπεριδοειδών
- Επεξεργασία και εξαγωγή πιστοποιημένων οργανικών προϊόντων

  (προϊόντων βιολογικής γεωργίας)
- Εμπορία λιπασμάτων, φυτοφαρμάκων και ζωοτροφών

Επίσης, ιδιαίτερα σημαντική δραστηριότητα θεωρείται η τεχνική στήριξη των παραγωγών και η παροχή διαφόρων τύπων ασφάλειας. Παράλληλα, η Π.Ε.Σ.

εκπροσωπεί τα μέλη της στην Ευρωπαϊκή Ένωση, στο Υπουργείο Γεωργίας, στη ΠΑΣΕΓΕΣ κτλ

6.2 Εμπειρική εφαρμογή της ανάλυσης διακύμανσης κατά ένα παράγοντα.

Μία από τις εμπορικές δραστηριότητες του Συνεταιρισμού είναι η πώληση ελαιολάδου. Ο Συνεταιρισμός κατέχει ένα σύγχρονο τυποποιητήριο και μπορεί να εμφιαλώνει ελαιόλαδο σε φιάλες των 0,25 l , 0,5 l 0,75 l και 1 λίτρου, καθώς επίσης σε δοχεία λευκοσιδήρου των 3 και 5 λίτρων. Μια από της εμπορικές στρατηγικές του Συνεταιρισμού είναι να προωθεί τις πωλήσεις του τυποποιημένου ελαιόλαδου καθώς με αυτή την κίνηση επιτυγχάνει διαφοροποίηση προϊόντος με αποτέλεσμα να μπορεί να πουλήσει ακριβότερα από τους ανταγωνιστές (π.χ. Ισπανούς). Αντίθετα στο χύμα λάδι ο ανταγωνισμός είναι πολύ μεγάλος και από το εμπόριο δίνεται έμφαση στη χαμηλή τιμή και όχι στην ποιότητα.

Για να εξετάσουμε αν ισχύει αυτή η περίπτωση πήραμε τις μέσες τιμές πώλησης του χύμα ελαιόλαδου και του τυποποιημένου και θα εξετάσουμε αν όντως υπάρχει διαφορά με την βοήθεια της ανάλυσης διακύμανσης κατά ένα παράγοντα. Οι τιμές είναι από το 2002 έως το 2008 και από την τιμή του τυποποιημένου ελαιόλαδου έχει αφαιρεθεί το μέσο κόστος της τυποποίησης.

Descriptives						
Type of ad viewed						
	N	Mean	Std. Deviation	Std. Error	Minimum	Maximum
Strongly appealing	8	2,786169	0,152367	0,05387	2,56	3
Appealing	7	3,781666914		0,155997	3,18	4,17
Total	15	3,250734	0,590665	0,152509	2,56	4,17

Στον παραπάνω πίνακα βλέπουμε κάποια στατιστικά στοιχεία για τις τιμές του ελαιολάδου.

ANOVA

Type of ad viewed

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3,700	1	3,700	40,603	,000
Within Groups	1,185	13	,091		
Total	4,884	14			

Το αποτέλεσμα της ανάλυσης διακύμανσης είναι σαφές το Significance ισούται με μηδέν άρα υπάρχει στατιστική σημαντικότητα άρα υπάρχει διαφορά στην τιμή του τυποποιημένου ελαιόλαδου και του χύμα

6.3 Εμπειρική εφαρμογή της ανάλυσης διακύμανσης κατά δύο παράγοντες.

Ο Συνεταιρισμός εκτός από συσκευασία εσπεριδοειδών παραδίδει πορτοκάλια για χυμοποίηση μέσω του καθεστώς ενίσχυσης από την Ε.Ε. Για κάθε κιλό πορτοκάλι που παραδιδόταν στο χυμοποιητή ο παραγωγός παίρνει 0,09€. Δυστυχώς για να δώσει αυτή την επιδότηση η Ε.Ε. απαιτεί κάποια έγγραφα. Κάθε αρχή της χρονιάς ο Συνεταιρισμός πρέπει να συγκεντρώσει τις ατομικές δηλώσεις από τους παραγωγούς που δηλώνουν την προβλεπόμενη παραγωγή για χυμοποίηση. Έπειτα συντάσσεται το συμφωνητικό και παραδίδεται η κατάστασή στο αρμόδιο όργανο ελέγχου του Υπουργείου Αγροτικής Ανάπτυξης και Τροφίμων. Το πρόβλημα έγκειται στο ότι ο Συνεταιρισμός δικαιούται στατιστικό σφάλμα 30%, αν αποτύχει στην πρόβλεψη του και για παράδειγμα παραδώσει κάτω από 70% της δηλωθείσας ποσότητας τότε οι παραγωγοί δεν θα εισπράξουν την επιδότηση, ενώ αν παραδώσει παραπάνω από το 130% δεν θα εισπράξει για την παραπάνω επιδότηση. Κάποιοι παραγωγοί εκμεταλλευόμενοι το πρόβλημα των εξωτερικών οικονομιών δηλώνουν παραπάνω παραγωγή από αυτή που υπολογίζουν ότι θα φέρουν για να μπορούν να φέρουν παραπάνω κιλά σε περίπτωση που έχουν κάνει λάθους στην πρόβλεψή τους. Έτσι ο

Συνεταιρισμός είναι υποχρεωμένος να κάνει προσαρμόσει την συνολική ποσότητα για να είναι συνεπή με την ποσότητα που θα παραδώσει

Για τις επόμενες δύο εμπειρικές εφαρμογές θα χρησιμοποιήσουμε τα δεδομένα από την χυμοποίηση. Στην ανάλυση διακύμανσης κατά δύο παράγοντες θα χρησιμοποιήσουμε διαχρονικά στοιχεία (5 ετών) από τις ατομικές δηλώσεις 40 παραγωγών σαν εξαρτημένη μεταβλητή την ποσότητα για χυμοποίηση και σαν ανεξάρτητες την ποσότητα για συσκευασία και την περιοχή. Για την περιοχή χρησιμοποιήσαμε την τιμή 0 για τα κτήματα που βρίσκονται σε κάποιο υψόμετρο και την τιμή 1 στα παραθαλάσσια

Tests of Between-Subjects Effects

Dependent Variable:juice

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1,006E10	28	3,592E8	119,730	,000
Intercept	1,053E10	1	1,053E10	3509,939	,000
syskeyasia	9,734E9	23	4,232E8	141,071	,000
place	1600000,000	1	1600000,000	,533	,489
syskeyasia * place	1,964E8	4	4,910E7	16,367	,001
Error	2,100E7	7	3000000,000		
Total	2,396E10	36			
Corrected Total	1,008E10	35			

a. R Squared = ,998 (Adjusted R Squared = ,990)

Στο αποτέλεσμα της ανάλυσης διακύμανσης παρατηρούμε ότι είναι στατιστικά σημαντική η παραδοθείσα ποσότητα για συσκευασία, αλλά όχι η τοποθεσία για στατιστική σημαντικότητα 95%

6.4 Εμπειρική εφαρμογή του μοντέλου της παλινδρόμησης

Συνεχίζοντας την ερευνά μας θα προσθέσουμε και την μεταβλητή, τα στρέμματα,

Στο παρακάτω πίνακα παρατηρούμε ότι το R^2 είναι αρκετά υψηλό 0,949, δηλαδή η ευθεία της παλινδρόμησης ερμηνεύει το 94,9% της παραδοθείσας ποσότητας για χυμοποίησης

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0,949117	0,900823	0,891525	5588,869
a. Predictors: (Constant), place, ektasi, syskeyasia				

Να αποτελέσματα της παλινδρόμησης φαντάζουν μη λογικά καθώς ο συντελεστής της έκτασης εμφανίζεται αρνητικός και στατιστικά μη σημαντικός κάτι που δεν ήταν αναμενόμενο, αυτό μπορεί να εξηγηθεί μόνο από την ιδιομορφία της περιοχής της Αιγιαλείας καθώς το 2004 και το 2007 είχε κτυπηθεί από πάγο με αποτέλεσμα να υπάρχει αναντιστοιχία έκτασης και παραγωγής σε κάποια κτήματα που επηρεάστηκαν από τον πάγο. Τέλος η παραγωγή για χυμοποίηση συσχετίζεται θετικά με την παραδοθείσα ποσότητα για συσκευασία (0,459), κάτι που ήταν και αναμενόμενο. Επίσης παρατηρούμε ότι η τοποθεσία δεν είναι στατιστικά σημαντική μεταβλητή όπως πρόβλεψε και η ανάλυση διακύμανσης.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3952,751	2252,876		1,755	,089
	ektasi	-231,740	321,345	-,094	-,721	,476
	syskeyasia	,459	,058	1,030	7,895	,000
	place	1320,329	2250,282	,033	,587	,561

a. Dependent Variable: juice

Η ευθεία της παλινδρόμησης σύμφωνα με το SPSS είναι:

Juice = 3952,751 -231,740* ektasi + 0,459* syskeyasia + 1320,329 *place

ΚΕΦΑΛΑΙΟ 7^{ο}

7.0 Ανασκόπηση – Συμπεράσματα

Η ανάλυση της διακύμανσης (ANalysis Of VAriance – ANOVA) είναι μία στατιστική μέθοδος με την οποία η μεταβλητότητα που υπάρχει σ' ένα σύνολο δεδομένων διασπάται στις επιμέρους συνιστώσες της με στόχο την κατανόηση της σημαντικότητας των διαφορετικών πηγών προέλευσής της. Η ανάπτυξη της μεθοδολογίας οφείλεται στον θεμελιωτή της σύγχρονης στατιστικής επιστήμης, άγγλο στατιστικό Sir Ronald Aylmer Fisher (1890-1962). Στην πραγματικότητα η ANOVA περιλαμβάνει μία ομάδα στατιστικών μεθόδων καταλλήλων για την ανάλυση δεδομένων που προκύπτουν από πειραματικούς σχεδιασμούς.

Τα δεδομένα ενός δείγματος ανάλογα με την προέλευσή τους διακρίνονται σε παρατηρήσεις σε πειραματικά. Στην πρώτη κατηγορία ο στατιστικός ερευνητής απλά παρατηρεί τις τιμές που εμφανίζονται χωρίς να έχει δυνατότητα επέμβασης στις αντίστοιχες μεταβλητές. Αντίθετα στη δεύτερη κατηγορία ο στατιστικός ερευνητής προσπαθεί να ελέγξει τα επίπεδα μιας η περισσοτέρων ανεξάρτητων μεταβλητών προκειμένου να προσδιορίσει την επίδραση που έχουν πάνω στην υπό μελέτη μεταβλητή που καλείται εξαρτημένη η απόκριση.

Στην εργασία μας αναλύσαμε την θεωρεία της ανάλυσης διακύμανσης κατά ένα παράγοντα και κατά δύο παράγοντες. Επίσης αναπτύχθηκε το γενικό μοντέλο της παλινδρόμησης και εφαρμόστηκαν εμπειρικά και τα τρία στατιστικά μοντέλα στον Παναιγιάλειο αγροτικό συνεταιρισμό.

Βιβλιογραφία

1. ΙΩΑΝΝΙΔΗΣ ΔΗΜΗΤΡΙΟΣ Α. (1999) - Στατιστικές μέθοδοι / Εκδόσεις: ΖΗΤΗ, ΑΘΗΝΑ
2. ΚΑΦΦΕΣ ΔΗΜ. Γ. (1989) - Μαθήματα αναλύσεως διακυμάνσεως / Εκδόσεις: ΣΤΑΜΟΥΛΗΣ, ΠΕΙΡΑΙΑΣ
3. Ε. ΜΠΟΡΑ – ΣΕΝΤΑ, Χ. ΜΩΥΣΙΑΔΗΣ (1997) – Εφαρμοσμένη Στατιστική, Πολλαπλή Παλινδρόμηση-Ανάλυση Διασποράς-Χρονοσειρές / Εκδόσεις: ΖΗΤΗ, ΘΕΣΣΑΛΟΝΙΚΗ
4. JEFFREY JARRETT (2000) - Μέθοδοι Προβλέψεων–Για Οικονομικές και Επιχειρηματικές Αποφάσεις / Εκδόσεις: GUTENBERG, ΑΘΗΝΑ
5. ΜΑΥΡΟΜΑΤΗΣ ΓΕΩΡΓΙΟΣ (1999) - Στατιστικά μοντέλα και μέθοδοι ανάλυσης δεδομένων / University Studio Press, ΘΕΣΣΑΛΟΝΙΚΗ
6. ΧΑΛΙΚΙΑΣ ΙΩΑΝΝΗΣ Γ. (1999) - Στατιστικές μέθοδοι - Ανάλυση παλινδρόμησης - Ανάλυση διακύμανσης / Εκδόσεις: ΜΠΕΝΟΥ, ΑΘΗΝΑ
7. ΓΕΩΡΓΙΟΣ ΕΜΜ. ΧΑΛΚΟΣ (2000) – Στατιστική, Θεωρία-Εφαρμογές & Χρήση Στατιστικών Προγραμμάτων σε Η/Υ / Εκδόσεις: ΤΥΠΩΘΗΤΩ – ΓΙΩΡΓΟΣ ΔΑΡΔΑΝΟΣ, ΑΘΗΝΑ
8. ΤΑΧΥΝΑΚΗΣ ΠΑΝΑΓΙΩΤΗΣ (1999) - Εμπειρική προσέγγιση της ανεξαρτησίας του ορκωτού ελεγκτή στον ελληνικό χώρο
9. ΓΝΑΡΔΕΛΛΗΣ ΧΑΡΑΛΑΜΠΟΣ (2003) - Εφαρμοσμένη στατιστική / Εκδόσεις: ΠΑΠΑΖΗΣΗΣ, ΑΘΗΝΑ
10. ΣΙΩΜΚΟΣ ΓΕΩΡΓΙΟΣ Ι. (2005) - Εφαρμογή μεθόδων ανάλυσης στην έρευνα αγοράς / Εκδόσεις: ΣΤΑΜΟΥΛΗΣ, ΑΘΗΝΑ
11. ΙΩΑΝΝΙΔΗΣ ΔΗΜΗΤΡΙΟΣ Α. (2005) - Στατιστικές μέθοδοι περιγραφική στατιστική, θεωρία πιθανοτήτων, στατιστική συμπερασματολογία, απλή και πολλαπλή γραμμική παλινδρόμηση, χρονολογικές σειρές, ανάλυση διακύμανσης / Εκδόσεις: ΖΗΤΗ, ΘΕΣΣΑΛΟΝΙΚΗ

Ηλεκτρονική Βιβλιογραφία

http://www.guidoluechters.de/IMBIE_SPSS/S_IMBIE.html

http://stat-athens.aueb.gr/~jbn/courses/spps_sem/spss4.pdf