ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ

ΣΧΟΛΗ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ ΚΑΙ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ

ΤΜΗΜΑ ΔΙΟΙΚΗΣΗΣ ΤΟΥΡΙΣΜΟΥ

(πρώην Τμήμα Λογιστικής & Χρηματοοικονομικής – Μεσολόγγι)

Θέμα «Στατιστική Ανάλυση Δεδομένων από Επιχειρήσεις-Εφαρμογή»

Φοιτητές: Αλεξανδροπούλου Παναγιώτα

Χατζηδημητρίου Μιχαέλα

Περαντάκος Πέτρος

Επιβλέπουσα: κα. Καρυώτη Β.

Μεσολόγγι 2021

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα πτυχιακή εργασία με θέμα «Στατιστική Ανάλυση Δεδομένων από Επιχειρήσεις- Εφαρμογή» πραγματοποιήθηκε στο πλαίσιο της πτυχιακής εργασίας μας στο τμήμα Λογιστικής και Χρηματοοικονομικής του Ανώτατου Εκπαιδευτικού Ιδρύματος Πατρών το έτος 2021.

Θερμές ευχαριστίες απευθύνουμε σε όλους τους καθηγητές μας που είχαμε όλα αυτά τα χρόνια, για τις γνώσεις που μας μετέδωσαν αλλά και γιατί με το μάθημά τους αποτέλεσαν πρότυπα και πηγή έμπνευσης για εμάς.

Τέλος, ένα μεγάλο και εγκάρδιο ευχαριστώ στους γονείς μας που μας στήριξαν ηθικά και οικονομικά όλα αυτά τα χρόνια, δίνοντας μας κουράγιο ώστε να προσπεράσουμε κάθε εμπόδιο και να φτάσουμε στο στόχο μας.

ΠΕΡΙΛΗΨΗ

Η ανάλυση επιχειρησιακών δεδομένων, θεωρείται ο βασικός στόχος για την επίτευξη των απαιτούμενων επιχειρηματικών και αναπτυξιακών ερωτημάτων. Η επεξεργασία των δεδομένων επιτυγχάνεται με τη συλλογή, την επεξεργασία και την επεξήγηση των συμπερασμάτων που προέκυψαν, έτσι ώστε να είναι όσο το δυνατό πιο ορθή η λήψη των αποφάσεων από τις επιχειρήσεις σύμφωνα με τα δεδομένα που έχουν στη βάση τους.

Σκοπός της παρούσας εργασίας είναι αφενός η βιβλιογραφική ανασκόπηση της επιστήμης της στατιστικής, στο πεδίο ανάλυσης των δεδομένων από επιχειρήσεις και αφετέρου η πρακτική εφαρμογή της ανάλυσης ενός σετ επιχειρησιακών δεδομένων με σκοπό την εξαγωγή χρήσιμων συμπερασμάτων. Μέσω της πρακτικής εφαρμογής της στατιστικής ανάλυσης των επιχειρησιακών δεδομένων, επιχειρείται να αναδειχθεί η χρησιμότητά της από όλες τις επιχειρήσεις που διαθέτουν σαφείς αναπτυξιακούς στόχους και να υποδειχθούν οι ενδεχόμενοι περιορισμοί που μπορεί να προκύψουν, όταν η στατιστική ανάλυση εκτελείται με ένα ευρέως διατιθέμενο στατιστικό πακέτο όπως το SPSS και από μη εξειδικευμένους στατιστικούς αναλυτές.

Λέξεις κλειδιά : Επιχειρησιακά δεδομένα, Στατιστική Ανάλυση

ABSTRACT

Business data analysis provides the crucial link between data and the information needed to address business questions. The processing of data is achieved by collecting, processing and explaining the conclusions that have emerged, so that it is as correct as possible for companies to make decisions based on the data they have in their database.

The purpose of the following work is on the one hand the literature review of the science of statistics, in the field of data analysis by companies and on the other hand the practical application of the analysis of a set of business data in order to draw useful conclusions. Through the practical application of statistical analysis of business data, an attempt is made to highlight its usefulness by all companies with clear development objectives and to indicate any limitations that may arise when statistical analysis is performed with a widely available statistic package -such as SPSS- and by non-specialist statistical analysts.

Keywords : business data, statistical analysis

Πίνακας περιεχομένων

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΑΤΑΛΟΓΟΣ ΔΙΑΓΡΑΜΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΣΥΝΤΟΜΟΓΡΑΦΙΩΝ

ΣΥΝΤΟΜ/ΦΙΑ	ΟΛΟΓΡΑΦΩΣ	ΣΕΛΙΔΑ
POS	Point Of Sale/Σημείο Πώλησης	13
TID	Terminal ID/Τερματικό Αναγνώρισης	13
MBA	Market Basket Analysis/Ανάλυση Καλαθιού Αγορών	13
RTF	Rich Text Format/Μορφότυπο εμπλουτισμένου κειμένου	15
HTML	Hyper text Markup Language/Γλώσσα Σήμανσης Υπερκειμένου	15
XML	Extensible Markup language/ Επεκτάσιμη Γλώσσα Σήμανσης	15
PDF	Portable Document Format/Μορφότυπος Φορητού Εγγράφου	15
OLAP	Online Analytical Processing/ Αναλυτική Επιγραμμική Επεξεργασία	15
SQL	Structured Query Language/Δομημένη Γλώσσα Αναζητήσεων	16
BIDS	Business Intelligence Development Studio/Στούντιο Ανάπτυξης Επιχειρηματικής Ευφυΐας	16
SPSS	Statistical Package for the Social Sciences/Στατιστικό Πακέτο για τις Κοινωνικές Επιστήμες	33

ΕΙΣΑΓΩΓΗ

Οι περισσότερες επιστήμες του εικοστού πρώτου αιώνα -από τις βιοϊατρικές μέχρι τις επιστήμες του μάρκετινγκ και της βιομηχανίας- εξαρτώνται σε μεγάλο βαθμό από την ευρεία διαθεσιμότητα και προσβασιμότητα σε δεδομένα μεγάλου όγκου (big data). Μέσα από την εντατική χρήση και αξιοποίηση των δεδομένων με την βοήθεια της στατιστικής επιστήμης επιδιώκεται η πληρέστερη κατανόηση πολύπλοκων ή/και πολυσυζητημένων ζητημάτων που απασχολούν το περιβάλλον, τις αγορές, την ιατρική και τις επενδύσεις (Kolker, Stewart&Ozdemir, 2012).

Τα τεράστια δεδομένα κατακλύζουν κάθε επιχείρηση σε καθημερινή βάση. Ωστόσο, η ποσότητα των δεδομένων δεν είναι σημαντική, σε σύγκριση με τις δυνατότητες αξιοποίησής τους μέσω επιστημών όπως η στατιστική. Τα μεγάλου όγκου δεδομένα μπορούν να αναλυθούν, αποδίδοντας γνώση και πληροφορία που οδηγούν σε καλύτερες αποφάσεις και στρατηγικές επιχειρηματικές κινήσεις.

Η Ανάλυση Δεδομένων χρησιμοποιείται σε πολλές επιχειρήσεις και βιομηχανίες, επιτρέποντας τη λήψη βέλτιστων επιχειρηματικών αποφάσεων σε διάφορες εταιρείες και οργανισμούς. Η Ανάλυση Δεδομένων είναι ένα από τα πολλά βήματα που πρέπει να ολοκληρωθούν κατά τη διεξαγωγή μιας έρευνας, ωστόσο ο ρόλος της χαρακτηρίζεται νευραλγικός. Τα δεδομένα που συλλέγονται από διάφορες πρωτογενείς και δευτερεύουσες πηγές λόγω της ακατέργαστης μορφής τους, είναι συγκεχυμένα, ωστόσο δεν παύουν να είναι εξαιρετικά χρήσιμα. Είναι σχεδόν αδύνατο για τον ερευνητή να διαχειριστεί όλα αυτά τα δεδομένα στην αρχική τους μορφή. Μέσω της ανάλυσης, αυτά τα δεδομένα παρουσιάζονται σε κατάλληλη και συνοπτική μορφή άνευ απώλειας των σημαντικών και αξιοποιήσιμων πληροφοριών, ώστε να μπορούν να χρησιμοποιηθούν αποτελεσματικά για τη λήψη αποφάσεων.

Τα δεδομένα μπορούν να παρουσιαστούν σε πίνακες ή γραφικές παραστάσεις. Η παρουσίαση σε πίνακα συνεπάγεται την παράθεση αριθμητικών δεδομένων. Η γραφική ή σχηματική μορφή περιλαμβάνει την παρουσίαση δεδομένων από άποψη δομής που μπορούν να ερμηνευθούν οπτικά, π.χ., ραβδογράμματα, διαγράμματα πίτας, ιστογράμματα κ.λπ. Όλα τα παραπάνω συνιστούν μορφές ανάλυσης δεδομένων. Αυτές οι μέθοδοι έχουν σχεδιαστεί για να βελτιώσουν και να

ερμηνεύσουν τα δεδομένα, έτσι ώστε οι ενδιαφερόμενοι να αποκτούν φιλτραρισμένη γνώση και πληροφορία εκ των δεδομένων. Στο σημείο αυτό αξίζει να αναφερθεί ότι η περιγραφική στατιστική χρησιμοποιείται για την περιγραφή των δεδομένων, ενώ η επαγωγική στατιστική χρησιμοποιείται για την εξαγωγή συμπερασμάτων και υποθέσεων για τις ίδιες πληροφορίες.

Στο πρώτο κεφάλαιο της παρούσας εργασίας αναπτύσσονται τα τρέχοντα βιβλιογραφικά δεδομένα για τη στατιστική των επιχειρήσεων. Επιχειρείται μια σύντομη ιστορική αναδρομή της εξέλιξης της στατιστικής με το πέρασμα των χρόνων μέχρι τη σύγχρονη εφαρμογή της σε επιχειρησιακά δεδομένα. Αναλύονται οι έννοιες της δειγματοληψίας, της συλλογής ή της εξόρυξης των δεδομένων, της οπτικοποίησης των δεδομένων αλλά και της ανάλυσής τους με την βοήθεια της στατιστικής.

Στο δεύτερο κεφάλαιο, αναπτύσσονται οι έννοιες και τα εργαλεία της περιγραφικής στατιστικής, ενώ στο τρίτο κεφάλαιο αναπτύσσονται οι έννοιες και τα εργαλεία της επαγωγικής στατιστικής.

Στη συνέχεια, περιγράφονται διεξοδικά οι στατιστικές μέθοδοι της απλής και της πολλαπλής παλινδρόμησης.

Στο ειδικό μέρος, επιχειρείται μια πρακτική εφαρμογή στατιστικής ανάλυσης σε δεδομένα από επιχειρήσεις, παραθέτοντας αναλυτικά τα εξαγόμενα συμπεράσματα.

ΓΕΝΙΚΟ ΜΕΡΟΣ

Κεφάλαιο 1 Στατιστική των Επιχειρήσεων

1.1 Γενικά

Δεδομένου ότι η λύση των περισσότερων προβλημάτων στη βιομηχανία και τις επιχειρήσεις περιλαμβάνει τη συλλογή, ανάλυση και ερμηνεία των δεδομένων, η επίλυση τέτοιων προβλημάτων επαφίεται στους επιστήμονες της στατιστικής, οι οποίοι δύνανται να αναλάβουν καθήκοντα σε διάφορα σημεία της εκπόνησης μιας έρευνας (Snee, 2015). Οι στατιστικοί επιστήμονες (ή αλλιώς στατιστικολόγοι) έχουν επινοήσει πολλά εργαλεία που διατίθενται για χρήση για τη γενική βελτίωση της κερδοφορίας των επιχειρήσεων και την επίλυση βιομηχανικών προβλημάτων. Ωστόσο, υπάρχει ένα μεγάλο χάσμα μεταξύ των διαθέσιμων εργαλείων και αυτών που πραγματικά και αποτελεσματικά εφαρμόζονται σε επιχειρήσεις, βιομηχανίες και οργανισμούς. Επομένως, είναι σημαντική η εστίαση στη γεφύρωση αυτού του χάσματος, και ειδικότερα στο κατά πόσο οι στατιστικές μέθοδοι μπορούν να βελτιώσουν τις αποδόσεις των επιχειρήσεων και της βιομηχανίας γενικότερα (Abraham, 2007). Παρακάτω, θα εξεταστούν οι βασικές εφαρμογές στατιστικών μεθόδων στις επιχειρήσεις και τη βιομηχανία.

1.2 Ιστορική Αναδρομή

Ο H.G. Wells είπε κάποτε «Η στατιστική σκέψη θα είναι μια μέρα τόσο απαραίτητη για την αποτελεσματική διαβίωση όσο και η ικανότητα ανάγνωσης και γραφής». Παράλληλα, πρόσφατα, οι Harry & Schroeder (2000) παρατήρησαν ότι «η στατιστική γνώση συσχετίζεται με την εποχή της πληροφορίας και της τεχνολογικής εξέλιξης, με τον ίδιο τρόπο που σχετίζονταν τα ορυκτά καύσιμα με τη βιομηχανική εποχή. Στην πραγματικότητα, το μέλλον της βιομηχανίας εξαρτάται από την κατανόηση της Στατιστικής» (Abraham, 2007). Η στατιστική έχει χρησιμοποιηθεί στη βιομηχανία και τις επιχειρήσεις σε όλο τον κόσμο από τα μέσα του εικοστού αιώνα (Snee, 2015).

Ωστόσο, από πολύ παλαιότερα είχε τεθεί μια τεράστια πρόκληση στη στατιστική ανάλυση μιας χρονοεξαρτώμενης αλληλουχίας δεδομένων. Ειδικότερα, ο άνθρωπος

σχεδόν ανέκαθεν είχε την πρόθεση να κατανοήσει τις διαδικασίες αλλαγής και τα αναλυτικά εργαλεία που χρησιμοποιούσε συχνά σχεδιάζονταν για να τελέσουν συγκρίσεις σε περιπτώσεις όπου η ιστορία αδυνατούσε (Klein & Dave, 1997).

Η λέξη «Στατιστική» προέρχεται από τη Λατινική statisticum collegium (council of state, η Ιταλική λέξη statista σημαίνει πολιτικός) και πιθανολογείται ότι αναφέρθηκε για πρώτη φορά στο γερμανικό βιβλίο Statistik, που δημοσιεύθηκε το 1749, περιγράφοντας μεταξύ άλλων την ανάλυση δημογραφικών και οικονομικών δεδομένων σχετικά με το κράτος (πολιτική αριθμητική) (A Brief History of Statistics, 2017). Η μορφή αυτής της επιστήμης εκείνη την εποχή στόχευε στη συλλογή και την ταξινόμηση πληροφοριών εκ μέρους της δημόσιας διοίκησης σχετικά με α) το μέγεθος και τη σύνθεση του πληθυσμού, β) τη μετανάστευση, γ) τις δημογραφικές αλλαγές, δ) τους πίνακες γέννησης και θνησιμότητας, ε) τα δεδομένα για επιχειρήσεις, ζ) τις καλλιέργειες, στ) την κατανομή του πλούτου, η) την εκπαίδευση και θ) την υγεία (Divisietal., 2017). Το 1800 συμπεριλήφθηκαν στα παραπάνω η συλλογή, περίληψη και ανάλυση δεδομένων οποιουδήποτε τύπου και συνδυάστηκε με τη διερεύνηση πιθανοτήτων για σκοπούς στατιστικής συμπερασματικότητας (A Brief History of Statistics, 2017).

Υπάρχουν διάφοροι τομείς δραστηριότητας στους οποίους οι στατιστικές εφαρμογές έχουν ιστορικά γίνει εμφανείς. Η έρευνα μάρκετινγκ και καταναλωτικών τάσεων, συμπεριλαμβανομένων των δημοσκοπήσεων και των ατομικών και ομαδικών ερευνών έχει χρησιμοποιηθεί ευρέως εδώ και δεκαετίες (Roberts, 1990). Η στατιστική ανάλυση δεδομένων από επιχειρήσεις ιστορικά παρουσιάζεται σταθερά αναπτυσσόμενη, ιδιαιτέρως μέσω της συνεχούς εισαγωγής νέων εφαρμογών και της ανάπτυξης σημαντικών καινοτομιών σε στατιστικές έννοιες και μεθόδους

Η ίδρυση των πρώτων εκπαιδευτικών πανεπιστημίων που ασχολήθηκαν με την επιστήμη της Στατιστικής χρονολογείται από τις αρχές του 20^{ου} αιώνα. Ειδικότερα, το 1911 ιδρύθηκε το University College, στο Λονδίνο, το 1918 το τμήμα Βιομετρίας και Βιοστατιστικής Johns Hopkins, το 1931 το τμήμα Οικονομικής και Κοινωνικής Στατιστικής στο Πανεπιστήμιο της Πενσυλβάνια, το 1933 το Στατιστικό Εργαστήρι της Αιόβα και το 1935 το George Washington Statistics Department που για πρώτη

φορά ανήκε σε κολλέγιο ελευθέρων σπουδών και επιστημών (A Brief History of Statistics, 2017).

Ακολούθησε η ίδρυση του American Society for Quality (ASQ) το 1946 με την επακόλουθη έκδοση περιοδικών με ειδική θεματολογία για βιομηχανικό έλεγχο ποιότητας (ASQ's journals—Industrial Quality Control, Journal of Quality Technology, Quality Engineering and Quality Progress) που περιείχαν πολλές δημοσιεύσεις σχετικά με τις εφαρμογές στατιστικής στις επιχειρήσεις (Snee, 2015).

1.3 Σημασία της Στατιστικής στη διοίκηση Επιχειρήσεων

Μέσω της ανάλυσης δεδομένων, τα δεδομένα παρουσιάζονται σε κατάλληλη και συνοπτική μορφή χωρίς απώλεια σχετικών πληροφοριών, ώστε να μπορούν να χρησιμοποιηθούν αποτελεσματικά για τη λήψη αποφάσεων σε επιχειρήσεις, οργανισμούς και βιομηχανίες (Amit, 2015). Για παράδειγμα, εν όψει της εισαγωγής νέων φαρμακευτικών σκευασμάτων, οι φαρμακευτικές εταιρείες επιδίδονται σε εκτεταμένη χρήση στατιστικής ανάλυσης. Άλλες βιομηχανίες που εξαρτώνται απόλυτα από τα αποτελέσματα της στατιστικής ανάλυσης των επιχειρηματικών δεδομένων είναι οι εταιρείες χημικών και, τα τελευταία χρόνια, οι αυτοκινητοβιομηχανίες όπως και άλλες βιομηχανίες που επηρεάζονται από την αναβίωση του ενδιαφέροντος για τον ποιοτικό έλεγχο (Roberts, 1990). Οι ετήσιες εκθέσεις που εκπονούνται και δημοσιεύονται από επιχειρήσεις ή οργανισμούς (π.χ χρηματοπιστωτικά ιδρύματα) περιέχουν ποικιλία στοιχείων που αφορούν τις πωλήσεις, την παραγωγή, τις δαπάνες, τα αποθέματα, τα κεφάλαια που χρησιμοποιούνται και άλλες δραστηριότητες. Αυτά τα δεδομένα είναι συχνά δεδομένα πεδίου, που συλλέγονται χρησιμοποιώντας τεχνικές επιστημονικής έρευνας. Εάν δεν ενημερώνονται τακτικά, αυτά τα δεδομένα είναι προϊόν μιας μοναδικής προσπάθειας και έχουν περιορισμένη χρήση (Kundu, 2016). Η στατιστική ανάλυση έχει σαφή χρησιμότητα για πολλές εργασίες και οι στατιστικές μέθοδοι διαθέτουν τεράστια δυνητική αξία. Για παράδειγμα, συνηθίζεται αρκετά μέσα από αναλύσεις, η πρόβλεψη της ποιότητας διάφορων γεωργικών προϊόντων όπως το κρασί (αναλύσεις κατά τη συγκομιδή των σταφυλιών), η πρόβλεψη της διακύμανσης της αγοράς κατοικίας και η πρόβλεψη της αγοραστικής τάσης του καταναλωτικού κοινού (Ayres, 2007).

Η αντιμονοπωλιακή νομοθεσία, οι περιβαλλοντικές επιπτώσεις, η ρύθμιση των κινητών αξιών, η επαγγελματική ασφάλεια και άλλοι τομείς έχουν εμπνεύσει εκτεταμένες, παρατεταμένες και μερικές φορές εξελιγμένες στατιστικές μελέτες. Αυτές οι εφαρμογές είναι τόσο εξειδικευμένες που οι εξωτερικοί σύμβουλοι και όχι οι εργαζόμενοι συχνά εκτελούν τη στατιστική εργασία. Τα στατιστικά στοιχεία χρησιμοποιούνται επίσης συνήθως για την επικοινωνία και την παρουσίαση των πληροφοριών της εταιρείας, όπως εκθέσεις πωλήσεων και λογιστικές πληροφορίες, ετήσιες εκθέσεις και εσωτερικές εκθέσεις και παρουσιάσεις. Τα απλά αλλά πολύχρωμα "γραφικά παρουσίασης" παίζουν κεντρικό ρόλο. Η οικονομική πρόβλεψη και ο μακροπρόθεσμος σχεδιασμός είναι διεισδυτικές λειτουργίες, τουλάχιστον σε μεγάλες εταιρείες, και έχουν βασιστεί εδώ και πολύ καιρό σε στατιστικές μεθόδους, ειδικά σε εκείνες που σχετίζονται με την οικονομετρία και την ανάλυση χρονοσειρών. Τέλος, η έρευνα διαχείρισης επιστημονικών λειτουργιών, ως εξειδικευμένη λειτουργία ορισμένων επιχειρήσεων, ασχολείται σε μεγάλο βαθμό με τεχνικές βελτιστοποίησης όπως ο γραμμικός προγραμματισμός και χρησιμοποιεί επίσης σε μεγάλο βαθμό τη στατιστική μεθοδολογία (Roberts, 1990).

Σύμφωνα με τον Kundu (2016), οι σημαντικότερες λειτουργίες της εφαρμογής των στατιστικών μεθόδων σε κάθε επιχείρηση συνοψίζονται ως εξής:

Προγραμματισμός των δραστηριοτήτων: μπορεί να σχετίζεται είτε με ειδικά έργα είτε με επαναλαμβανόμενες δραστηριότητες μιας εταιρείας για μια καθορισμένη περίοδο.

Ο καθορισμός προτύπων: μπορεί να σχετίζεται με το μέγεθος της απασχόλησης, τον όγκο των πωλήσεων, τον καθορισμό προδιαγραφών ποιότητας για το κατασκευασμένο προϊόν, τους κανόνες για την ημερήσια παραγωγή, κ.α.

Η λειτουργία του ελέγχου: συνεπάγεται σύγκριση της πραγματικής παραγωγής που έχει επιτευχθεί σε σχέση με τον στόχο που είχε οριστεί. Σε περίπτωση που η παραγωγή υπολείπεται του στόχου, λαμβάνονται διορθωτικά μέτρα έτσι ώστε μια τέτοια ανεπάρκεια να εξαλειφθεί.

1.4 Έρευνα και Σχεδιασμός Στατιστικής ανάλυσης

1.4.1 Επιστήμη των δεδομένων

Τα τελευταία χρόνια, παρατηρήθηκε μια μεγάλη τάση μετασχηματισμού της κοινωνικής ζωής των ατόμων από «αναλογική» σε «ψηφιακή», προκαλώντας τεράστιο αντίκτυπο σχεδόν σε όλες τις καθημερινές δραστηριότητες και μεθόδους επικοινωνίας (Vander Aalst, 2016).

Η χρήση του όρου «επιστήμη των δεδομένων» γίνεται όλο και ευρύτερα χρησιμοποιούμενη, υπό την έννοια των «μεγάλων δεδομένων» (big data). Με τον όρο «επιστήμη» επισημαίνεται η γνώση που αποκτάται μέσω συστηματικής μελέτης. Η επιστήμη των δεδομένων συνοψίζεται ως μια συστηματική πρωτοβουλία που οικοδομεί και οργανώνει γνώση υπό τη μορφή διερευνώμενων εξηγήσεων και προβλέψεων (Heilbron, 2003).

Η επιστήμη δεδομένων μπορεί συνεπώς να συνεπάγεται με τη συστηματική μελέτη των οργανισμών/επιχειρήσεων και των ιδιοτήτων τους μέσω εστίασης σε δεδομένα (και κατ΄επέκταση σε στατιστικά στοιχεία). Παρά το γεγονός ότι η στατιστική ως επιστήμη υφίσταται εδώ και αιώνες, η επιστήμη των δεδομένων είναι σχετικά νέα, δεδομένου ότι πλέον η διάθεση των δεδομένων είναι ανυπολόγιστη (Dhar, 2013). Ειδικότερα, τα δεδομένα πλέον συλλέγονται με διάφορα κίνητρα, ανά πάσα στιγμή και σε οποιοδήποτε μέρος. Οι προβλέψεις του Γκόρντον Μουρ, συνιδρυτή της Intel, ανέφερε ότι το 1965 ο αριθμός των εξαρτημάτων στα ολοκληρωμένα κυκλώματα θα διπλασιαζόταν ανά χρόνο, τελικά αποδείχθηκαν εύστοχες καθώς τα τελευταία 50 χρόνια σημειώθηκε εκθετική ανάπτυξη των τεχνολογικών καινοτομιών και κατά συνέπεια των σύγχρονων δυνατοτήτων (Vander Aalst, 2016). Πρωτοπόρες τεχνολογικές επεκτάσεις όπως η χωρητικότητα των μέσων αποθήκευσης, η απόδοση των υπολογιστών ανά μονάδα κόστους, ο αριθμός των pixel κ.λπ. συνέβαλλαν στο παραπάνω αποτέλεσμα. Εκτός από αυτές τις τεχνολογικές εξελίξεις, οι άνθρωποι και οι οργανισμοί βρέθηκαν όλο και περισσότερο εξαρτώμενοι από τις ηλεκτρονικές συσκευές και τη διαδικτυακή πληροφόρηση. Επομένως, η επιστήμη των δεδομένων είναι ριζικά διαφορετική την τρέχουσα περίοδο, καθιστώντας τις δυνατότητες της στατιστικής περισσότερο αξιοποιήσιμες και ελκυστικές (Dhar, 2013).

1.4.2 Δειγματοληψία

Σκοπός μιας επιχειρηματικής έρευνας είναι η συλλογή δεδομένων από επιχειρήσεις για διάφορους σκοπούς. Για παράδειγμα, οι Εθνικές Στατιστικές Υπηρεσίες πολλών χωρών χρησιμοποιούν τα δεδομένα για την εκπόνηση τριμηνιαίων ή ετήσιων αναφορών σχετικά με το ακαθάριστο εγχώριο προϊόν, τις επενδύσεις, τις κεφαλαιακές συναλλαγές, τις κρατικές δαπάνες και το εξωτερικό εμπόριο κ.α. Οι επιχειρησιακές έρευνες παράγουν μια σειρά οικονομικών στατιστικών που εστιάζουν στη παραγωγή (έξοδα, εισροές, μεταφορά, κυκλοφορία αγαθών και ρύπανση), στις πωλήσεις (υπηρεσίες χονδρικής και λιανικής), στα εμπορεύματα (εισροές, εκροές, τύποι μεταφερόμενων αγαθών, αποστολές, αποθέματα, και παραγγελίες), στις οικονομικές καταστάσεις (έσοδα, έξοδα, περιουσιακά στοιχεία και υποχρεώσεις), στην εργασία (απασχόληση, μισθοδοσία, ώρες, παροχές και χαρακτηριστικά των εργαζομένων) και στις τιμές (τρέχων δείκτης τιμών και βιομηχανικός δείκτης τιμών) (Hidiroglou & Lavallée, 2009).

Η δειγματοληψία συνεπάγεται με την πραγματοποίηση παρατηρήσεων μόνο σε ορισμένα από τα μέλη ενός πληθυσμού και στη χρήση αυτών για την εξαγωγή συμπερασμάτων σχετικά με τα χαρακτηριστικά του πληθυσμού (Smith, 2013).

Οι επιχειρηματικές έρευνες διαφέρουν με διάφορους τρόπους από τις κοινωνικές έρευνες κυριότερα ως προς το σχεδιασμό στον οποίο περιλαμβάνεται μεταξύ άλλων και η δειγματοληψία. Συνήθως, οι έρευνες που εκπονούνται από τις επιχειρήσεις, λαμβάνουν δείγματα από μητρώα επιχειρήσεων που περιέχουν στοιχεία επικοινωνίας, όπως όνομα, διεύθυνση και σημεία επαφής, από τηρούμενα αρχεία διαχείρισης. Οι κοινωνικές έρευνες, από την άλλη πλευρά, συχνά χρησιμοποιούν πλαίσια περιοχής για την επιλογή νοικοκυριών και, τελικά, ατόμων από αυτά τα νοικοκυριά (Hidiroglou & Lavallée, 2009).

Γενικά, στις επιχειρηματικές έρευνες τα δειγματοληπτικά σχέδια είναι βελτιστοποιημένα για να παρέχουν εκτιμήσεις σε εθνικό συνήθως επίπεδο με ένα εκ των προτέρων καθορισμένο επίπεδο ακρίβειας που μπορεί να οδηγήσει σε ακατάλληλα μικρά μεγέθη δείγματος για συγκεκριμένες υποομάδες ενδιαφέροντος. Εφόσον η ακρίβεια των άμεσων εκτιμήσεων (π.χ των καθοριζόμενων από τη διακύμανση του εκτιμητή) είναι αντιστρόφως ανάλογη με το μέγεθος του δείγματος,

τις περισσότερες φορές τα μικρά μεγέθη δείγματος μπορεί να οδηγήσουν σε αναξιόπιστες άμεσες εκτιμήσεις. Ως εκ τούτου, ενδέχεται να χρειαστούν εναλλακτικοί τρόποι εκτίμησης (Burgard, Münnich & Zimmermann, 2014).

Έτσι, με στόχο να αποφεύγονται τα δειγματοληπτικά σφάλματα, συχνά επιλέγεται η τυχαιοποίηση κατά τη δειγματοληψία, υπό την έννοια ότι τα μέλη του πληθυσμού επιλέγονται με τυχαία διαδικασία, έτσι ώστε οι ιδιότητες των εκτιμήσεων να γίνονται αντιληπτές σε σχέση με επαναλαμβανόμενα δείγματα (Smith, 2013).

Οι σημαντικότερες μέθοδοι δειγματοληψίας περιγράφονται παρακάτω:

Η Απλή τυχαία Δειγματοληψία: πρόκειται για την τεχνική που εξασφαλίζει κάθε ίση πιθανότητα σε κάθε τυχαία επιλογή όπου μπορεί να πραγματοποιηθεί με τη μέθοδο των λαχνών, με πίνακες τυχαίων αριθμών ή μέσω ενός Η/Υ(Δήμας, 2021).

Συστηματική Δειγματοληψία: πρόκειται για την επιλογή του δείγματος που περιέχει το δειγματοληπτικό πλαίσιο σε κανονικά διαστήματα. Υπάρχουν ορισμένα στάδια που ακολουθούνται και είναι τα εξής:

Αρίθμηση στο πλαίσιο με ένα μοναδικό αριθμό με αρχή το 0,1,2,..... Αρχικά επιλέγεται ένας αριθμός τυχαίος. Ακολουθεί ο υπολογισμός του κλάσματος της δειγματοληψίας = πραγματικό μέγεθος του δείγματος /το σύνολο του εξεταζόμενου πληθυσμού. Επιλέγονται οι άλλες περιπτώσεις με την χρήση του κλάσματος(Δήμας, 2021).

Στρωματοποιημένη Δειγματοληψία: σε περίπτωση που το δείγμα του εξεταζόμενου πληθυσμού δεν είναι αρκετά ομοιογενές τότε η μέθοδος της απλής δειγματοληψίας δεν θεωρείται η πλέον αντιπροσωπευτική και εφαρμόζεται η Στρωματοποιημένη δειγματοληψία. Η μέθοδος αυτή χωρίζει τον πληθυσμό σε στρώματα με όσο το δυνατό μεγαλύτερη ομοιογένεια με αντικειμενικό στόχο τη διαμόρφωση της μεγαλύτερης διαφοροποίησης ανάμεσα στις κατηγορίες(Δήμας, 2021).

Δειγματοληψία κατά ομάδες: εμφανίζει παρόμοια χαρακτηριστικά με τη μέθοδο της Στρωματοποιημένης δειγματοληψίας χωρίζοντας τον πληθυσμό σε ομάδες που δημιουργούνται με κάποια διαφοροποίηση ανάλογα με τα δεδομένα όπως ομαδοποίηση ανά γεωγραφική περιοχή, κατά ομάδα αίματος, κατά φύλο κλπ. Στη μέθοδο αυτή περιέχονται τα ακόλουθα βήματα:

Επιλογή της ομαδοποίησης από το δειγματοληπτικό πλαίσιο .
Αρίθμηση όλων των ομάδων που δημιουργήθηκαν
Επιλογή του δείγματος με τη χρήση κάποιας τυχαίας δειγματοληψίας(Δήμας, 2021).

Δειγματοληψία Ποσοστών: δειγματοληψία ποσοστών ονομάζεται το δειγματοληπτικό σχέδιο που εμφανίζει παρόμοια τεχνική με τη μέθοδο της δειγματοληψίας κατά στρώματα, αλλά η επιλογή των μονάδων μέσα σε κάθε στρώμα δεν γίνεται τυχαία αλλά από τους συνεντεύκτες με δικά τους κριτήρια. Τα κριτήρια που καθορίζονται, είναι σχετικά με το θέμα που εξετάζεται και ο υπολογισμός του ποσοστού θα γίνει με βάση τα όσα αναφέρθηκαν, Αν για παράδειγμα, το θέμα εξέτασης σχετίζεται με οικονομική κατάσταση των Ελλήνων που είναι δυσαρεστημένοι, τότε θα δημιουργηθεί ένα στρώμα πληθυσμού με κριτήρια ορισμένες παραμέτρους όπως το φύλλο, την ηλικία, την απασχόληση (δημόσιο, ιδιώτες, άνεργοι, συνταξιούχοι κλπ) τον τομέα (αστικός, αγροτικός), το εισόδημα κλπ.(Δήμας, 2021).

Δειγματοληψία Χιονόμπαλας: η δειγματοληψία χιονόμπαλας αποτελεί μία τεχνική δημιουργίας ενός δείγματος μέσα από τον πυρήνα γνωστών στοιχείων, που ενισχύουν τη μέθοδο αυτή με την προσθήκη επιπλέον στοιχείων. Η μέθοδος αυτή χρησιμοποιείται όταν δεν υπάρχει διαθέσιμο δειγματοληπτικό πλαίσιο και επομένως είναι δυνατή η χρήση της μεθόδου κυρίως σε πληθυσμούς που χαρακτηρίζονται δύσκολοι προσβάσιμοι(Δήμας, 2021).

Δειγματοληψία Σκοπιμότητας: η μέθοδος αυτή αναφέρεται στην επιλογή δείγματος ορισμένων ομάδων (ή περιπτώσεων) του πληθυσμού που ανταποκρίνονται σε ορισμένες υποθέσεις. Όσο μεγαλύτερο είναι το δείγμα τόσο καλύτερα αντιπροσωπευτικό θεωρείται αλλά δεν μπορεί να εγγυηθεί την αξιοπιστία των αποτελεσμάτων(Δήμας, 2021).

1.4.3 Συλλογή/Εξόρυξη δεδομένων

Με τον όρο «εξόρυξη δεδομένων» αποδίδεται η ανακάλυψη κρυφών και σημαντικών προτύπων και τακτικών καταναλωτικών τάσεων μέσα από μεγάλες ποσότητες δεδομένων (Gan&Dai, 2014).

Η αυξανόμενη παγκοσμιοποίηση του λιανικού εμπορίου, η συνεχής συνδεσιμότητα, η συνάφεια με βάση τα συμφραζόμενα και ένας κόσμος με πολλές οθόνες αλλάζουν τόσο την online όσο και την offline αγοραστική εμπειρία. Καθώς η εμπειρία στο κατάστημα αρχίζει να εκπίπτει, ανοίγονται συναρπαστικές νέες δυνατότητες για τους εμπόρους λιανικής μέσω της εξόρυξης δεδομένων και της τεχνολογίας πληροφοριών. Το σύγχρονο λιανικό εμπόριο εξαρτάται σε μεγάλο βαθμό από τις πληροφορίες και οι επενδύσεις στον τομέα αυτό από τις λιανικές επιχειρήσεις έχουν αυξηθεί σημαντικά. Η γνώση της αγοράς, καθώς και ο έλεγχος των δεδομένων και των πληροφοριών είναι απαραίτητη για την απόκτηση ενός επιχειρηματικού ανταγωνιστικού πλεονεκτήματος. Ουσιαστικά, η τεχνολογία πληροφοριών μπορεί να επιταχύνει τις διαδικασίες και να αυξήσει τις πωλήσεις, να βελτιώσει τα ποσοστά διατήρησης των πελατών και να προσφέρει οφέλη εξοικονόμησης κόστους για μια εταιρία (Shet, 2015).

Η ικανότητα συγκέντρωσης δεδομένων (που είναι εύκολα προσβάσιμα) και αποτελεσματικής παρουσίασης των παραγόμενων πληροφοριών σε πραγματικό χρόνο υπήρξε σημαντικός καταλύτης για την βελτίωση της παραγωγικότητας πολλών οργανισμών και επιχειρήσεων (Lee, 2013). Ωστόσο, υπάρχει μια συντριπτική μάζα ετερογενών δεδομένων που συλλέγονται ως αποτέλεσμα. Η ανάλυση αυτών των δεδομένων γίνεται ένα κρίσιμο και προκλητικό μέρος της επιχειρηματικής διαδικασίας(Roberts & Laramee, 2018).

Ένα περιβόητο παράδειγμα της αξιοποίησης μεγάλου όγκου δεδομένων από επιχειρήσεις ήταν η εξόρυξη δεδομένων από συναλλαγές λιανικού εμπορίου όπως καταγράφονταν από τις συσκευές POS (Point Of Sale /Σημείο Πώλησης) και ειδικούς αλγόριθμους τη δεκαετία του 1970. Τα δεδομένα συναλλαγών λιανικής αρχικά οργανώνονταν σε σχεσιακές βάσεις δεδομένων. Όπως φαίνεται στον παρακάτω πίνακα κάθε εγγραφή στη βάση δεδομένων αντιπροσώπευε μία συναλλαγή με τέσσερα κύρια πεδία: α) TID (Αναγνωριστικό συναλλαγής), β) χρόνος συναλλαγής, γ) λεπτομέρειες των στοιχείων που περιέχονται στη συναλλαγή και δ) Σύνολο. Οι συναλλαγές πελατών που χρησιμοποιούσαν κάρτες επιβράβευσης οργανώνονταν περαιτέρω ως βάσεις δεδομένων ακολουθιών συναλλαγών, όπου οι συναλλαγές ενός πελάτη καταγράφονταν σε χρονική σειρά (Gan &Dai, 2014).

Πίνακας 1. Δείγμα βάσης δεδομένων συναλλαγών λιανικής (Gan&Dai, 2014)

TID	Time	Items' Details				Total ($)
		Item	Unit Price	Quantity	Subtotal ($)	
10000001	09:15, 12/03/2012	Beer	$3.58	2	$7.16	$31.33
		Apple	$2.45	1.8 kg	$4.41	
		Diaper	$9.88	2	$19.76	
10000002	09:20, 12/03/2012	Sugar	$1.26	1.5kg	$1.89	$15.73
		Beef	$8.65	1.6kg	$13.84	
...	...	...	...	...	...	...
99999999	16:34, 09/11/2012	Toothpaste	$2.9	2	$4.80	$8.05
		Toothbrush	$3.25	1	$3.25	

Μέσω της ανάλυσης του καλαθιού αγοράς (Market Basket Analysis-MBA), εφαρμόστηκε και συνεχίζει να εφαρμόζεται μέχρι σήμερα μια άμεση εξόρυξη κανόνων συσχέτισης σε βάσεις δεδομένων συναλλαγών πωλήσεων, όπου κάθε εγγραφή αντιπροσωπεύει μια συναλλαγή πωλήσεων και είναι ένα είδος προϊόντος. Για παράδειγμα, ο κανόνας συσχέτισης «Καφές - Γάλα» που βρίσκεται σε μια βάση δεδομένων συναλλαγών υποδεικνύει ότι υπάρχει συσχέτιση μεταξύ καφέ και γάλακτος και αποκαλύπτει ότι ο πελάτης που αγοράζει καφέ τείνει να αγοράζει γάλα ταυτόχρονα (Gan & Dai, 2014).

1.4.4 Οπτικοποίηση δεδομένων

Ένας ραγδαία αυξανόμενος αριθμός επιχειρήσεων βασίζεται σε λύσεις οπτικοποίησης για τις προκλήσεις της διαχείρισης δεδομένων. Αυτή η ζήτηση πηγάζει από τη συνολικότερη στροφή ολόκληρου του επιχειρηματικού κλάδου προς τις προσεγγίσεις που στηρίζονται στα δεδομένα για λήψη αποφάσεων και επίλυση προβλημάτων (Roberts & Laramee, 2018). Σύμφωνα με τον Friedman, ο κύριος σκοπός της οπτικοποίησης δεδομένων είναι η επικοινωνία πληροφοριών με σαφήνεια και αποτελεσματικότητα μέσω γραφικών μέσων. Η ιδέα βασίζεται στην δημιουργία αισθητικών και λειτουργικών απεικονίσεων δεδομένων προκειμένου να παρέχονται πληροφορίες και διαισθητικοί τρόποι αντίληψης περίπλοκων δεδομένων (Friedman, 2008).

Η χρήση οπτικής ανάλυσης αυξάνει την κατανόηση των δεδομένων, επιτρέποντας έτσι σε ένα ευρύτερο φάσμα χρηστών να ερμηνεύουν την υποκείμενη συμπεριφορά, σε αντίθεση με τους εξειδικευμένους αλλά ακριβούς αναλυτές δεδομένων. Η διεύρυνση της προσέγγισης κοινού με ένα ευρύτερο φάσμα φόντων δημιουργεί νέες ευκαιρίες για λήψη αποφάσεων, επίλυση προβλημάτων, αναγνώριση τάσεων και

δημιουργική σκέψη. Οι συνηθέστερες τάσεις στην οπτικοποίηση των επιχειρησιακών δεδομένων και στην οπτική αναλυτική αφορούν την οπτικοποίηση που χρησιμοποιείται για την αντιμετώπιση προκλήσεων δεδομένων και προσδιορίζει τομείς στους οποίους οι βιομηχανίες χρησιμοποιούν οπτικό σχεδιασμό για να αναπτύξουν την κατανόησή τους για το επιχειρηματικό περιβάλλον (Roberts & Laramee, 2018).

Μέσω της χρήσης των εργαλείων γραφικής απεικόνισης και σχεδίασης, τα δεδομένα μπορούν να οπτικοποιηθούν με γραφικό περιεχόμενο, προβολές κινουμένων σχεδίων, τρισδιάστατα μοντέλα και διαδραστικά εργαλεία οπτικοποίησης, ενώ οι παρουσιάσεις μπορούν να ανατεθούν για φιλοξενία σε ιστοσελίδες ή εκδηλώσεις. Σχέδια επιχειρηματικών διαδικασιών, τοποθεσιών ή τάσεων μπορούν επίσης να παραχθούν για να επεξηγήσουν έννοιες και να βελτιώσουν την παρουσίαση των πληροφοριών (Homocianu, 2010).

Τα δεδομένα μπορούν να παρουσιαστούν σε μορφή πίνακα ή γραφικής αναπαράστασης. Η μορφή πίνακα αφορά περισσότερο την παρουσίαση αριθμητικών δεδομένων. Η γραφική αναπαράσταση περιλαμβάνει την παρουσίαση δεδομένων που μπορούν να ερμηνευθούν οπτικά. Από την άποψη της δομής τους διακρίνονται σε ραβδογράμματα, γραφήματα σε μορφή πίτας, ιστογράμματα, διαγράμματα διασποράς κλπ. Όλα τα παραπάνω, συνιστούν μορφές ανάλυσης δεδομένων, με σκοπό την βελτίωση και το φιλτράρισμα των δεδομένων, έτσι ώστε οι αναγνώστες να μπορούν να λαμβάνουν καθαρή πληροφόρηση χωρίς να χρειάζεται να ταξινομήσουν όλα τα δεδομένα (Amit, 2015).

Υπάρχει μια συνεχής ανάγκη δημιουργίας χρήσιμων επιχειρηματικών αναφορών οι οποίες συνήθως βασίζονται σε πρότυπα που δημιουργούνται και καθορίζονται από τον τελικό χρήστη, και είναι ευρύτερα γνωστά με τον όρο «point and click». Αυτά τα πρότυπα καθορίζουν ποια πεδία περιλαμβάνονται στην αναφορά, τους τύπους μηνυμάτων που λαμβάνουν οι χρήστες και την εμφάνιση της αναφοράς και δημιουργούνται με όργανα που επιτρέπουν την εξαγωγή αναφορών σε διαφορετικές μορφές (rtf, html, xml, pdf).

Επιπλέον, η οπτικοποίηση των επιχειρησιακών δεδομένων περιλαμβάνει εκπόνηση αναφορών βασισμένων στο OLAP (Online Analytical Processing) - μια κατηγορία

τεχνολογίας λογισμικού που δίνει τη δυνατότητα σε αναλυτές, διευθυντές και ανώτατα στελέχη να έχουν διορατικότητα στα δεδομένα μέσω γρήγορης, δομημένης και διαλογικής πρόσβασης σε μία μεγάλη ποικιλία πιθανών απεικονίσεων της πληροφορίας που έχει μετατραπεί από σκέτα δεδομένα, με σκοπό να αντικατοπτρίσει την πραγματική διαστατοποίηση της επιχείρησης, όπως την αντιλαμβάνονται οι χρήστες. Το OLAP περιλαμβάνει κυρίως τη συγκέντρωση μεγάλων ποσοτήτων διαφορετικών δεδομένων με πολύπλοκες σχέσεις. Στόχος είναι η ανάλυση αυτών των σχέσεων και η αναζήτηση μοτίβων, τάσεων και εξαιρέσεων. Ένα τυπικό παράδειγμα εργαλείου OLAP είναι ο υπερκύβος (hypercube) που ορίζεται από τα μέτρα και τις διαστάσεις του (Homocianu, 2010).

Μια άλλη περίπτωση οπτικοποίησης επιχειρησιακών δεδομένων είναι αυτή των αναφορών που χρησιμοποιούν επιχειρηματικές μετρήσεις. Αυτές είναι μετρήσεις που χρησιμοποιούνται για τη μέτρηση ορισμένων ποσοτικοποιήσιμων στοιχείων της απόδοσης μιας εταιρείας, όπως η επενδυτική απόδοση, οι μετρήσεις υπαλλήλων και φοιτητών, έσοδα και έξοδα κ.α (Homocianu, 2010).

Σε διαδικασίες εξόρυξης δεδομένων, με στόχο τις προβλέψεις και την αναγνώριση προτύπων δεδομένων, προβλέπεται η χρήση δέντρων αποφάσεων, συσταδοποίησης κανόνων συσχέτισης κ.λπ., με τις αντίστοιχες αναπαραστάσεις τους.

Πέρα από αυτές τις φόρμες, σήμερα πολλές γραφικές αναπαραστάσεις επιχειρησιακών δεδομένων έχουν τρισδιάστατο προσανατολισμό, με στόχο την διευκόλυνση της σύγκρισης όχι μόνο των αριθμητικών τιμών μεταξύ τους αλλά και ολόκληρης της αλληλουχίας δεδομένων.

Οι περισσότερες από τις παραπάνω μεθόδους οπτικοποίησης δεδομένων μπορούν να γίνουν με προϊόντα λογισμικού υπολογιστικών φύλλων. Για παράδειγμα, το Excel 2003 και 2007 προσφέρει αυτή τη δυνατότητα, ενώ η έκδοση του 2007 μπορεί επιπλέον να κάνει προεπισκόπηση του αποτελέσματος σε τρισδιάστατη (3D) μορφή.

Οι περισσότερες από τις δισδιάστατες γραφικές παραστάσεις που χρησιμοποιούνται σήμερα σε εφαρμογές επιχειρηματικής ευφυΐας έχουν γεωγραφική διάσταση. Σε αυτήν την περίπτωση ασχολούμαστε με Χαρτογράμματα και Χωροπληθείς Χάρτες.

Τέλος, υπάρχουν και τρισδιάστατες μορφές των παραπάνω χαρτών με προορισμό την απεικόνιση της χωρικής κατανομής αναφορικά με οποιοδήποτε αγοραστικό ή καταναλωτικό φαινόμενο και για οποιονδήποτε σχετικό δείκτη.

Κάθε παράδειγμα δεδομένων που αναφέρεται παραπάνω χρησιμοποιείται επί του παρόντος από τους περισσότερους προγραμματιστές λύσεων επιχειρηματικής ευφυΐας. Το κύριο πρόβλημα εξακολουθεί να είναι η περιορισμένη εφαρμογή τους ή το γεγονός ότι εξακολουθεί να παρατηρείται ανεπαρκής εκμετάλλευση αυτών των διαθέσιμων προϊόντων (Homocianu, 2010).

1.4.5 Ανάλυση δεδομένων

Οι στατιστικές ερευνητικές μέθοδοι λειτουργούν συνήθως μέσα από ένα δείγμα δεδομένων και χρησιμοποιούν αυτά τα δεδομένα για να εξαχθούν συμπεράσματα για οτιδήποτε ανησυχεί τους ερευνητές(Wood, 2010).Ζητείται συχνά από τον στατιστικολόγο να αναλύσει δεδομένα που έχουν ήδη συλλεχθεί. Ωστόσο, ο στατιστικός είναι πιο αποτελεσματικός όταν συμμετέχει σε μια μελέτη από τα αρχικά στάδια του σχεδιασμού της. Σε συνεργασία με έναν ερευνητή ή μια διεπιστημονική ομάδα, ο στατιστικολόγος βοηθά στον προσδιορισμό των ερευνητικών προβλημάτων και στην ανάπτυξη μιας κατάλληλης στρατηγικής για τη συλλογή των απαραίτητων δεδομένων (Snee, 2015). Κατά τη διάρκεια συλλογής των δεδομένων, ο στατιστικός επιβλέπει το χρονοδιάγραμμα και την διαδοχή των βημάτων βάση του αρχικού σχεδιασμού και επιμελείται της ομαλής και σωστά κατευθυνόμενης διεξαγωγής. Συχνά και εν όψει δυσκολιών διεξαγωγής, μπορεί να χρειαστεί η αναζήτηση ενδιάμεσων λύσεων ή η αναθεώρηση του προγραμματισμού από τον στατιστικολόγο. Μετά τη συλλογή των δεδομένων, ο ερευνητής ή οι στατιστικοί συνδυάζουν τις δεξιότητές τους στην ανάλυση και ερμηνεία των αποτελεσμάτων. Στο τέλος μιας μελέτης, ο στατιστικός συντάσσει συνήθως μια γραπτή έκθεση που περιγράφει τη διαδικασία λύσης (υπόβαθρο του προβλήματος, σχεδιασμός μελέτης, συλλογή δεδομένων και στατιστική ανάλυση) και συζητά τα αποτελέσματα. Λόγω του συνεργατικού χαρακτήρα της επίλυσης προβλημάτων, είναι πολύ σημαντικό οι στατιστικολόγοι να έχουν καλές δεξιότητες προφορικής και γραπτής επικοινωνίας (Snee, 2015).

Η επιτυχία οποιουδήποτε προγράμματος εφαρμογής της στατιστικής σε επιχειρησιακά δεδομένα προϋποθέτει την πλήρη και ορατή δέσμευση της ανώτερης διοίκησης. Αυτή, οφείλει να αφιερώνει χρόνο και πόρους στην αρχική αξιολόγηση της κατάστασης και στη διανομή καθηκόντων και ρόλων. Τα στατιστικά εργαλεία μπορούν να βοηθήσουν στην αύξηση της ικανοποίησης των πελατών και στη μέτρηση της απόδοσης του οργανισμού. Η εφαρμογή των στατιστικών εργαλείων είναι μια συνεχής διαδικασία και βοηθά τον οργανισμό να είναι ένας οργανισμός μάθησης και μια επιχείρηση βασισμένη στη γνώση. Οι οργανισμοί που βασίζονται στη γνώση τείνουν να είναι περισσότερο επιτυχημένοι μακροπρόθεσμα (Abraham, 2007). Επιπλέον, οι εταιρείες εκμεταλλεύονται ολοένα και περισσότερο τα βασικότερα και απλούστερα στατιστικά εργαλεία για τη βελτίωση της ποιότητας και της παραγωγικότητας και αναπτύσσουν «εταιρικές κουλτούρες», κατάλληλες για την αποτελεσματική χρήση της στατιστικής. Αξίζει να σημειωθεί ότι η στατιστική ανάλυση μπορεί να χρησιμοποιηθεί πιο αποτελεσματικά στην επιχείρηση όταν πολλοί εργαζόμενοι – έχουν στοιχειωδώς κατανοήσει τα στατιστικά εργαλεία και την στατιστική σκέψη. Ευτυχώς, υπάρχουν ενδείξεις ότι τα πολύ στοιχειώδη εργαλεία αρκούν για να ενισχύσουν διάφορους τύπους μελετών που μπορούν να φωτίσουν τα περισσότερα επιχειρηματικά προβλήματα και να διευκολύνουν τις περισσότερες επιχειρηματικές αποφάσεις (Roberts, 1990).

Όπως προαναφέρθηκε, η ανάλυση ερευνητικών δεδομένων παρέχει τον κρίσιμο σύνδεσμο μεταξύ ερευνητικών δεδομένων και πληροφοριών που απαιτούνται για την αντιμετώπιση ερευνητικών ερωτημάτων (Amit, 2015).

Η Ανάλυση Δεδομένων έχει πολλαπλές πτυχές και προσεγγίσεις, που περιλαμβάνουν ποικίλες στατιστικές τεχνικές, με μια ποικιλία ονομάτων σε διαφορετικές επιχειρήσεις, και επιχειρηματικούς τομείς από τη βιομηχανία μέχρι την κοινωνική επιστήμη (Amit, 2015). Η πιθανολογική συλλογιστική, οι αποκλίσεις από μια μέση κατανομή συχνότητας, η παλινδρόμηση των ελαχίστων τετραγώνων, η συσχέτιση και ο έλεγχος υποθέσεων είναι μέθοδοι με τις οποίες εκτελούνται ως επί το πλείστο οι στατιστικές συγκρίσεις.

Η σύγκριση των διαφορών μεταξύ χαμηλότερης και υψηλότερης τιμής, η εξομάλυνση με κυμαινόμενους μέσους όρους και η αναδιάταξη μιας χρονοσειράς σε

ένα πλέγμα διατομών είναι εργαλεία για την αποσύνθεση των σειρών έως ότου η διαδικασία με την οποία μένει είναι ισοδύναμη με τις παρατηρήσεις σχετικά με μια κατάσταση ηρεμίας (Klein & Dave, 1997).

Ακόμη κι αν τα συλλεχθέντα δεδομένα είναι επαρκή, έγκυρα και αξιόπιστα σε οποιοδήποτε βαθμό, δεν μπορούν να εξυπηρετήσουν κανένα χρήσιμο σκοπό, χωρίς την προσεκτική ανάλυσή τους. Οι διάφορες τεχνικές που μπορούν να χρησιμοποιηθούν κατά την ανάλυση των δεδομένων εμπίπτουν σε δύο κατηγορίες, την α) περιγραφική και την β) συμπερασματική ανάλυση. Οι παραπάνω τεχνικές μπορούν να εξυπηρετήσουν πολλούς σκοπούς όπως α) τη σύνοψη δεδομένων με απλό τρόπο, β) την οργάνωση δεδομένων ώστε να καθίστανται πιο ευνόητα και γ) χρήση των δεδομένων για δοκιμή των θεωριών σε έναν μεγαλύτερο πληθυσμό. Η σύγχρονη μεγάλη ποικιλία και διαθεσιμότητα λογισμικού υπολογιστών, μπορούν να οδηγήσουν στην αποφυγή των κουραστικών τύπων και υπολογισμών (Amit, 2015).

Καθώς η περιγραφική στατιστική ασχολείται σε μεγάλο βαθμό με τη μελέτη της κατανομής μιας μεταβλητής, αυτό το είδος ανάλυσης μπορεί να διακριθεί σε τρεις διαφορετικούς τύπους (Amit, 2015) :

- Μονομεταβλητή (univariate) ανάλυση: Όταν μια μεμονωμένη μεταβλητή αναλύεται μόνη της. Ένα τυπικό παράδειγμα μονομεταβλητής ανάλυσης μπορεί να είναι η εύρεση του μέσου όρου της ηλικίας ενός δείγματος ή των ακαδημαϊκών επιδόσεων ενός δείγματος αποτελούμενου από μαθητές ή φοιτητές

- Διμεταβλητή (Bivariate) ανάλυση: Όταν κάποια συσχέτιση μετριέται μεταξύ δύο μεταβλητών ταυτόχρονα. Πιο αναλυτικά, η διμεταβλητή ανάλυση χρησιμοποιείται όταν διερευνάται η ύπαρξη συσχέτισης μεταξύ ενός προσδιοριστή και μιας έκβασης (Γαλάνης, 2014). Ένα τυπικό παράδειγμα τέτοιας ανάλυσης είναι η διερεύνηση καταναλωτικών τάσεων ανά ηλικία καταναλωτών.

- Πολυμεταβλητή (multivariate) ανάλυση: Στην πολυμεταβλητή ανάλυση, τρεις ή περισσότερες μεταβλητές διερευνώνται ταυτόχρονα, επιτρέποντάς μας να εξετάσουμε τα αποτελέσματα περισσότερων από μία μεταβλητών ταυτόχρονα. Ένα τυπικό παράδειγμα τέτοιας ανάλυσης είναι ο προσδιορισμός της ικανοποίησης από την εργασία ανάλογα με όρους όπως η ηλικία, το φύλο, ο μισθό κ.ο.κ. Η

πολυμεταβλητή ανάλυση περιλαμβάνει τεχνικές όπως ανάλυση πολλαπλής παλινδρόμησης, ανάλυση πολλαπλών διακρίσεων, πολυπαραγοντική ανάλυση διακύμανσης (MANOVA), ανάλυση παραγόντων.

Συνοψίζοντας, οι στατιστικές μέθοδοι που χρησιμοποιούνται στη δημοσιευμένη έρευνα επιχειρησιακών δεδομένων ποικίλλουν, ωστόσο υπάρχουν τρεις ευρέως χρησιμοποιούμενες κατηγορίες μεθόδων που ξεχωρίζουν από τις ανεπίσημες έρευνες (Wood, 2010) :

• Περιγραφικά στατιστικά στοιχεία όπως ο μέσος όρος, οι τυπικές αποκλίσεις, οι συσχετίσεις και οι αναλογίες αναφέρονται ευρέως.

• Οι αυτοέλεγχοι μηδενικής υπόθεσης είναι πολύ συχνές και αποτελούν τον πυρήνα της λογικής δομής πολλών ερευνητικών μελετών και των εργασιών που τις αναφέρουν. Τα διαστήματα εμπιστοσύνης και οι μέθοδοι Bayesian, που μπορούν να χρησιμοποιηθούν ως εναλλακτικές, αναφέρονται σπάνια.

• Η παλινδρόμηση, σε πολλές μορφές της, χρησιμοποιείται ευρέως.

Μέσω της ανάλυσης παλινδρόμησης είναι δυνατή η α) περιγραφή των σχέσεων μεταξύ των εξαρτημένων και ανεξάρτητων μεταβλητών, η β) εκτίμηση των τιμών της εξαρτώμενης μεταβλητής, μέσω εκτίμησης των παρατηρούμενων τιμών των ανεξάρτητων μεταβλητών και η γ) πρόγνωση των παραγόντων κινδύνου που επηρεάζουν τον προσδιορισμό του αποτελέσματος (Schneider, Hommel & Blettner, 2010).

Ο όρος «ανάλυση» αναφέρεται στον υπολογισμό ορισμένων μέτρων (όπως μέτρα κεντρικής τάσης, διακύμανση κ.λπ.) μαζί με την αναζήτηση μοτίβων σχέσης (όπως συσχέτιση, παλινδρόμηση) που υπάρχουν μεταξύ των ομάδων δεδομένων. Εκτός από αυτό, στη διαδικασία ανάλυσης, σχέσεων ή διαφορών, η υποστήριξη ή η σύγκρουση με την αρχική ή νέα υπόθεση θα πρέπει να υποβληθεί σε στατιστικά τεστ σημασίας για να καθοριστεί υπό ποια εγκυρότητα τα δεδομένα υποδηλώνουν τυχόντα συμπεράσματα. Η ανάλυση, επομένως, πρέπει να κατηγοριοποιηθεί ως περιγραφική και συμπερασματική ανάλυση (η συμπερασματική ανάλυση είναι επίσης γνωστή ως στατιστική ανάλυση). Στην περιγραφική ανάλυση κυριαρχεί ο υπολογισμός ορισμένων δεικτών από ανεπεξέργαστα δεδομένα και καθιέρωση σχέσης μεταξύ δύο

ή περισσότερων μεταβλητών. Από την άλλη μεριά, η συμπερασματική ανάλυση αφορά την α) εκτίμηση των παραμέτρων ενός πληθυσμού και τον (β) έλεγχο της στατιστικής υπόθεσης ή του επιπέδου στατιστικής σημαντικότητας (Amit, 2015).

1.4.6 Ανάλυση του όρου της μεταβλητής

Η μεταβλητή στην πιο απλή μορφή της αποτελεί μία ποσότητα που μπορεί να μεταβάλλεται (Petrie & Sabin, 2008). Στην στατιστική, ο όρος «μεταβλητή» περιγράφει ένα στοιχείο που είναι διαφορετικό από ένα μεμονωμένο μέλος του πληθυσμού σε ένα άλλο (Kaur, 2013). Η ηλικία και το ανάστημα ενός ατόμου ή η προϋπηρεσία ενός εργαζομένου μπορούν να θεωρηθούν μεταβλητές.

Τα χαρακτηριστικά ή οι ιδιότητες των στατιστικών μονάδων ως προς τα οποία εξετάζεται ένας πληθυσμός ονομάζονται μεταβλητές (variables). Ειδικότερα, ο όρος μεταβλητή χρησιμοποιείται όταν χρειάζεται να αποδοθεί ένα χαρακτηριστικό ή μια ιδιότητα σε ένα πρόσωπο, ένα αντικείμενο, μια κατάσταση, κ.λπ., το οποίο πρόκειται να καταμετρηθεί, ούτως ώστε να εξυπηρετήσει τους σκοπούς της εκάστοτε έρευνας. Για παράδειγμα, σε μια μελέτη του ύψους των ενήλικων παιδιών μιας πόλης, σε κάθε ξεχωριστό άτομο αντιστοιχίζεται ένας αριθμός που δηλώνει το ύψος του. Με αυτό τον τρόπο ορίζεται η μεταβλητή Χ= «ύψος ατόμου»(Χαλικιάς, και συν, 2015).

Οι μεταβλητές μιας έρευνας μπορούν να ταξινομηθούν σε δύο μεγάλες κατηγορίες:

α) Ποσοτικές

β) Ποιοτικές (ή κατηγορικές)

Οι μεταβλητές όπως το ύψος και το βάρος μετρούνται με κάποιο είδος κλίμακας, μεταφέρουν ποσοτικές πληροφορίες που ονομάζονται ποσοτικές μεταβλητές. Το φύλο, το χρώμα του δέρματος και των ματιών δίνουν ποιοτικές πληροφορίες και ονομάζονται ποιοτικές μεταβλητές (Kaur, 2013).

Οι ποσοτικές μεταβλητές παίρνουν αριθμητικές τιμές και μπορούν να διακριθούν σε:

> Διακριτές (discrete) όταν παίρνουν μόνο ακέραιες τιμές.
> Συνεχείς (continuous) όταν παίρνουν τιμές από το σύνολο των πραγματικών αριθμών.

Οι ποιοτικές (ή κατηγορικές) μεταβλητές δεν εκφράζουν κάτι μετρήσιμο. Για παράδειγμα, η προέλευση και ο τύπος ενός τροφίμου, η ένταση ενός αρώματος και η οικογενειακή κατάσταση (έγγαμος/η, άγαμος, χήρος) είναι ποιοτικά ή κατηγορικά δεδομένα. Οι κατηγορικές μεταβλητές είναι μεταβλητές που δεν μπορούν να ταξινομηθούν σε συγκεκριμένη σειρά αλλά σε κατηγορίες (Ali & Bhaskar, 2016).

1.4.7 Κλίμακες Μέτρησης

Ευρέως χρησιμοποιούνται οι εξής τέσσερις κλίμακες: κατηγορίας, διάταξης, διαστήματος και αναλογίας. Χαρακτηριστική ομάδα στις κλίμακες αποτελούν οι δίτιμες/δυαδικές κλίμακες που έχουν μόνο δύο τιμές (0, 1).Οι κλίμακες μέτρησης είναι οι ακόλουθες:

Οι κλίμακες Αναλογίας (ratio): Στις κλίμακες αυτές οι τιμές που λαμβάνουν χώρα βρίσκονται σε αναλογική αντιστοίχιση με την ποσότητα του χαρακτηριστικού που μετρούν (π..χ. ταχύτητα, χρόνος, μήκος, βάρος, εισόδημα). Το μηδέν είναι καλά ορισμένο(Δαφέρμος, 2005).

Διαστήματος (interval): Στις κλίμακες αυτές οι διαφορές μεταξύ των τιμών τους έχουν και ίσες διαφορές για το αντίστοιχο χαρακτηριστικό που μετρά η κλίμακα (π.χ. Θερμοκρασία, ηλικία). Το μηδέν δεν είναι καλά ορισμένο(Δαφέρμος, 2005).

Διάταξης (ordinal): Στις κλίμακες αυτές συνολικά οι τιμές ορίζουν μία σχέση διάταξης(Δαφέρμος, 2005).

Στις διατεταγμένες κλίμακες, περιέχεται μία ορισμένη διάταξη των τιμών της μεταβλητής με ανώτερες και κατώτερες τιμές. Ένα χαρακτηριστικό παράδειγμα τακτικής κλίμακας στη μέτρηση της απόδοσης ενός μαθητή είναι η εξής:

➢ πολύ καλή

➢ καλή

> μέτρια

> κακή

> πολύ κακή (Ρόντος, & Παπαπάνης, 2006)

′Ονομαστική (nominal): Συνολικά οι τιμές, διαφοροποιούνται μόνο σε ένα στοιχείο (π.χ. χρώμα ματιών, τόπος γέννησης, φύλο)(Δαφέρμος, 2005), (π.χ. στη μεταβλητή «φύλο», η τιμή «άρρεν» με την τιμή «θήλυ»). Εδώ οι τιμές δεν αντιστοιχίζονται σε κάποια διάταξη και δεν μετρούν στάσεις. Αξίζει να τονιστεί ότι οι ονομαστικές κλίμακες αποτελούν την απλούστερη μορφή κλίμακας μέτρησης(Ρόντος, & Παπαπάνης, 2006).

Οι δυο πρώτες κλίμακες μέτρησης αφορούν τις ποιοτικές μεταβλητές ενώ οι δυο τελευταίες τις ποσοτικές.

Διατεταγμένη είναι μια ποιοτική μεταβλητή όταν οι κωδικοποιημένες τιμές που δίνουμε στη μεταβλητή αυτή καθορίζονται από μια διάταξη. Παράδειγμα, για την ένταση ενός αρώματος δημιουργούμε μια διατεταγμένη μεταβλητή που παίρνει τις εξής κωδικοποιημένες τιμές: 1 = άοσμο, 2 = ασθενές άρωμα, 3 = έντονο άρωμα. Ονομαστική είναι μια ποιοτική μεταβλητή όταν οι κωδικοποιημένες τιμές που δίνουμε κατηγοριοποιούν τα στοιχεία ενός συνόλου σε ομάδες. Παράδειγμα, 1 = άνδρας, 2 = γυναίκα ή M = άνδρας, F = γυναίκα (Νικήτας, 2013). Οι ονομαστικές ή ποιοτικές μεταβλητές, εκφράζουν μια ατομική ποιότητα (π.χ χρώμα και σχήμα φύλλων και φρούτων). Η ποιοτική μεταβλητή δεν μετράται, αλλά ταξινομείται σε κατηγορίες, βάση του τρόπου παρουσίασής της (ώριμα ή άγουρα φρούτα μπιζέλια, πράσινο ή κίτρινο χρώμα) (Divisi et al., 2017).

Μια κατηγορική μεταβλητή μπορεί επίσης να διακριθεί σε δυαδική (binary) ή διχοτομική (dichotomous), όταν υπάρχουν μόνο δύο πιθανές κατηγορίες. Για παράδειγμα αναφέρονται «Ναι/Όχι», «Άνδρας/Γυναίκα», «Ενήλικος/Ανήλικος» κ.λπ (Petrie & Sabin, 2008).

1.5 Περιορισμοί της στατιστικής Επιστήμης

Υπάρχει μια μεγάλη βιβλιογραφία σχετικά με τα πλεονεκτήματα και τα μειονεκτήματα των διαφορετικών προσεγγίσεων στην εφαρμογή της στατιστικής σε δεδομένα από επιχειρήσεις. Πιο αναλυτικά, διαπιστώνονται πολλές αμφιβολίες σχετικά με την ορθότητα εφαρμογής συγκεκριμένων μεθόδων και τη χρήση τους έναντι άλλων σε συγκεκριμένα προβλήματα, σχετικά με τις δυσκολίες εκπαίδευσης

των χρηστών στατιστικής, της αξιολόγησης των συμπερασμάτων τους από τους αναγνώστες και σχετικά με την παραγωγή νέων μεθόδων (Wood, 2010).

Η στατιστική μεθοδολογία έχει μεγάλες δυνατότητες για χρήσιμη εφαρμογή στις επιχειρήσεις, αλλά αυτό το δυναμικό σπάνια υλοποιείται (Roberts, 1990).

Σύμφωνα με τον Kundu (2016), η εφαρμογή της στατιστικής ανάλυσης σε δεδομένα από επιχειρήσεις διακρίνεται από ορισμένους αλληλοσχετιζόμενους περιορισμούς, οι οποίοι συνοψίζονται ως εξής:

Υπάρχουν ορισμένα φαινόμενα ή έννοιες όπου η στατιστική δεν μπορεί να εφαρμοστεί, δεδομένου ότι κάποια φαινόμενα ή έννοιες δεν επιδέχονται μέτρηση. Για παράδειγμα, έννοιες όπως η ομορφιά, η αντιγήρανση, η καλή ποιότητα ζωής κ.α δεν μπορούν να ποσοτικοποιηθούν. Τα στατιστικά στοιχεία αποκαλύπτουν τη μέση συμπεριφορά, την κανονική ή τη γενική τάση. Η εφαρμογή της έννοιας «μέσος όρος» εάν εφαρμόζεται σε ένα άτομο ή σε μια συγκεκριμένη κατάσταση μπορεί να οδηγήσει σε λάθος συμπέρασμα και μερικές φορές μπορεί να είναι καταστροφική. Για παράδειγμα, η μέτρηση του μέσου βάθους ενός ποταμού μπορεί να παράσχει αναξιόπιστα αποτελέσματα, στα οποία βασιζόμενος κάποιος (σημεία που έχουν μεγαλύτερο βάθος), μπορεί να κινδυνεύσει. Τα στατιστικά στοιχεία, καθώς συλλέγονται για συγκεκριμένο σκοπό, έχουν ειδικότητα για το σκοπό αυτό και δεν είναι ιδιαιτέρως χρήσιμα για κάποιον άλλο σκοπό. Η Στατιστική, όπως τα Μαθηματικά και η Λογιστική δεν είναι 100% ακριβείς.

Αρκετά συχνά, η συσχέτιση μεταξύ δύο ή περισσότερων μεταβλητών δεν υποδεικνύει απαραίτητα μια απαράβατη σχέση αιτίας και αποτελέσματος. Δείχνει απλώς την ομοιότητα ή την διαφοροποίηση στην διακύμανση μεταξύ δύο μεταβλητών. Σε τέτοιες περιπτώσεις, ο χρήστης πρέπει να ερμηνεύσει προσεκτικά τα αποτελέσματα, επισημαίνοντας τον τύπο της σχέσης που αποκτήθηκε.

Ένας σημαντικός περιορισμός των στατιστικών είναι ότι δεν αποκαλύπτει όλες τις επιδράσεις που σχετίζονται με ένα συγκεκριμένο φαινόμενο. Οι στατιστικολόγοι πρέπει να ερμηνεύουν τα αποτελέσματα έχοντας κατά νου όλες τις άλλες πτυχές που έχουν σχέση με το δεδομένο πρόβλημα.

1.6 Προβλέψεις

Παρά το γεγονός ότι τόσο οι επιστημονικές όσο και οι επιχειρηματικές εφαρμογές της στατιστικής ανάλυσης βασίζονται σε ένα κοινό σύνολο στατιστικών εργαλείων, μια από τις βασικότερες διαφορές μεταξύ τους είναι ότι στις στατιστικές εφαρμογές στις επιχειρήσεις η κύρια έμφαση δίδεται στην επίλυση άμεσων προβλημάτων, ενώ στις στατιστικές εφαρμογές στην επιστήμη η κύρια έμφαση δίδεται στις ευρείες εμπειρικές γενικεύσεις (Roberts, 1990).

Η επιστήμη των δεδομένων (Data Science), ασχολείται με την πρόβλεψη πωλήσεων και έχει ως αντικείμενο την εξαγωγή σημαντικών πληροφοριών από αδόμητα ή δομημένα δεδομένα. Ενώ υπάρχει πληθώρα τεχνικών πρόβλεψης πωλήσεων, η παλινδρόμηση (regression) που γίνεται με την βοήθεια της στατιστικής είναι μια από τις κρισιμότερες. Με τον όρο «πρόβλεψη» λογίζεται η διαδικασία εκτίμησης των μελλοντικών γεγονότων με όσο το δυνατόν πιο έγκυρο και αποτελεσματικό τρόπο βασιζόμενη σε ιστορικά και εμπειρικά δεδομένα και γνώση μελλοντικών γεγονότων που θα μπορούσαν να παρεμβληθούν. Η πρόβλεψη χρησιμοποιείται σε διάφορους επιχειρηματικούς, επενδυτικούς και βιομηχανικούς τομείς, λόγω της αυξημένης και ανεπιτήδευτης ανάγκης των επιχειρήσεων για την μείωση του αποφασιστικού ρίσκου σχετικά με μελλοντικές δραστηριότητες. Έτσι, οι επιχειρήσεις δίνουν μεγάλη σημασία στην διαδικασία της πρόβλεψης προσδοκώντας μέσω της ανάλυσης αυτής πιο έγκυρες πληροφορίες για την λήψη αποφάσεων με σκοπό να επιτυγχάνουν τους στόχους τους. Αποτελεί ουσιαστικά το βασικό εργαλείο για κάθε μελλοντική εξέλιξη και απόφαση και η ποιότητα της εξαρτάται από την ορθότητα, τη μεθοδικότητα και την επαλήθευση των μεθόδων με τις οποίες προέκυψαν και αναλύθηκαν τα δεδομένα. Το αποτέλεσμα αυτών των παραγόντων καλείται συχνά πιθανό σφάλμα». Η αξία της ενσωμάτωσης του πλαισίου στην έρευνα για την ενίσχυση της ακρίβειας του μοντέλου αναγνωρίζεται ευρέως και υποστηρίζεται από ολοένα και πιο εξελιγμένες στατιστικές μεθόδους (Bamberger, 2008). Παρόλα αυτά, σπάνια είναι δυνατό να δοθούν ντετερμινιστικές προβλέψεις μέσω της στατιστικής, δεδομένου του πόσο ανεπιτήδευτες είναι εγγενώς οι ανθρώπινες ενέργειες (Wood, 2010).

Κεφάλαιο 2 Περιγραφική Στατιστική

2.1 Αντικείμενο της Περιγραφικής Στατιστικής

Η περιγραφική στατιστική ασχολείται συγκεκριμένα με την ανάλυση και την συστηματική περιγραφή των στατιστικών δεδομένων που παραπέμπει είτε σε ένα μέρος του στατιστικού πληθυσμού ή στο συνολικό πληθυσμό.
Οι ακόλουθες κύριες στατιστικές διαδικασίες και τεχνικές περιλαμβάνονται στην περιγραφική στατιστική

α) τη διαλογή, την οργάνωση, την επεξεργασία και στην παρουσίαση τον στατιστικών δεομένων που θα προκύψουν σε μορφή γραφικών παραστάσεων και κατανομών συχνοτήτων,

β)την ανάλυση των δεδομένων που θα περιέχει και τον υπολογισμό των κύριων στατιστικών παραμέτρων όπως ο μέσος όρος , μέτρα διασποράς ,ασυμμετρίας κυρτότητας , κλπ

γ) την έρευνα που αφορά την αλληλεξάρτηση δύο ή περισσοτέρων μεταβλητών ενός πληθυσμού,
δ)την ανάλυση που αναφέρεται στην μελέτη της συνεχής εξέλιξης και πρόβλεψης κάποιον φαινομένων ενός πληθυσμού.

ε)την εξήγηση των αποτελεσμάτων που προκύπτουν από την στατιστική ανάλυση. Αυτά τα στατιστικά αποτελέσματα αναφέρονται μόνο σε ένα κομμάτι του στατιστικού πληθυσμού χωρίς να αναφέρονται στο σύνολο αυτού. Γενικότερα αν αυτά τα στατιστικά δεδομένα αφορούν το σύνολο των μονάδων ενός πληθυσμού τότε δεν αφορά την περιγραφική στατιστική και εδώ θα τελειώσει η έρευνα που έγινε με τις μεθόδους της περιγραφικής. Εάν από ένα συγκεκριμένο δείγμα ενός πληθυσμού με την συλλογή , την επεξεργασία και την ανάλυση αριθμητικών δεδομένων προήρθαν συμπεράσματα , αυτά τα συμπεράσματα που προέκυψαν από την εφαρμογή της περιγραφικής στατιστικής θα πρέπει να γενικευθούν ως προς το συνολικό πληθυσμό από το οποίο κιόλας πάρθηκε το δείγμα (Κιόχος, 2015).

2.2 Πίνακες συχνοτήτων

Ένα σχετικά μικρό σύνολο κατηγορικών ή αριθμητικών τιμών x1, x2, …, xm μιας διακριτής μεταβλητής X μπορεί εύκολα να παρουσιαστεί σε ένα πίνακα συχνοτήτων(frequency table). Ο πίνακας συχνοτήτων παρουσιάζει για κάθε τιμή xi της X τη συχνότητα εμφάνισής της που συμβολίζεται ως fi, δηλαδή πόσες φορές εμφανίζεται η κάθε διακεκριμένη τιμή στο δείγμα (Κουγιουμτζής, 2014).

Ονομάζουμε συχνότητα (frequency) της τιμής xi τον φυσικό αριθμό vi που δείχνει πόσες φορές επαναλαμβάνεται η τιμή xi στο δείγμα. Αν v = v1 + v2 + … + vm, τότε ο λόγος, fi=vi/v

ονομάζεται σχετική συχνότητα (relative frequency) εμφάνισης ή αλλιώς το ποσοστό (percent) pi της τιμής xi, που ορίζεται από το λόγο της συχνότητας εμφάνισης fi μιας τιμής vi προς το σύνολο των παρατηρήσεων ν του δείγματος (Κουγιουμτζής, 2014).

Αντίστοιχα, η αθροιστική συχνότητα, Fi, ορίζεται από τη σχέση

Fi=f1+f2+……..fi για i=1, 2……..m

Όταν το πλήθος των τιμών του δείγματος είναι μεγάλο αλλά κυρίως όταν η μεταβλητή X είναι συνεχής, οπότε μπορεί να πάρει μια οποιαδήποτε τιμή στο πεδίο ορισμού της, οι συχνότητες δεν ορίζονται σε μια συγκεκριμένη τιμή x_i αλλά σε μια περιοχή τιμών της X, που ονομάζεται κλάση (bin). Συγκεκριμένα αν x_{min} και x_{max} είναι η ελάχιστη και η μέγιστη τιμή της μεταβλητής X στο δείγμα, διαιρούμε το διάστημα $x_{max} - x_{min}$ σε διαστήματα μήκους $\Delta x = (x_{max} - x_{min})/k$, που ονομάζονται κλάσεις και σε κάθε κλάση υπολογίζουμε το σύνολο των τιμών του δείγματος που ανήκουν σε αυτή. Η ποσότητα αυτή, που προφανώς είναι ένας φυσικός αριθμός, είναι η συχνότητα της κλάσης. Αντίστοιχα ορίζονται η σχετική και η αθροιστική συχνότητα μιας κλάσης (Νικήτας, 2013).

2.3 Μέθοδοι Γραφικής Παρουσίασης Δεδομένων

Για την παρουσίαση των στατιστικών δεδομένων χρησιμοποιούνται επίσης διάφοροι τύποι γραφικών παραστάσεων, με κυριότερα τα ραβδογράμματα, τα κυκλικά διαγράμματα, γ) ιστογράμματα και τα δ) θηκογράμματα. Οι δύο πρώτοι τύποι

γραφικών παραστάσεων χρησιμοποιούνται συνήθως όταν η μεταβλητή Χ είναι ποιοτική, ενώ οι δύο τελευταίοι τύποι όταν έχουμε ποσοτικά δεδομένα.

α) Ραβδόγραμμα

Το ραβδόγραμμα σχηματίζεται βάση του πίνακα συχνοτήτων μιας διακριτής μεταβλητής Χ. Στον οριζόντιο άξονα τοποθετούνται ισαπέχοντα τα στοιχεία του δείγματος και σε κάθε στοιχείο αντιστοιχίζεται μια ορθογώνια στήλη με ύψος ίσο με τη συχνότητα του στοιχείου (Νικήτας, 2013). Η κάθε ράβδος μπορεί να παρουσιάζει είτε τη σχετική ή την αθροιστική συχνότητα για κάθε τιμή xi (Κουγιουμτζής, 2014).

β) Κυκλικό διάγραμμα

Το κυκλικό διάγραμμα ή διάγραμμα πίτας παρουσιάζει την ίδια πληροφορία με διαφορετική διάταξη, καθώς συνίσταται από ένα κυκλικό δίσκο χωρισμένο σε κυκλικές τομές. Το κάθε κομμάτι της επιφάνειας του κύκλου (πίτα) παρουσιάζει τη συχνότητα της αντίστοιχης τιμής, δηλαδή το εμβαδόν κάθε κυκλικής τομής είναι ανάλογο προς τη συχνότητα της αντίστοιχής τιμής (Νικήτας, 2013).

γ) Ιστόγραμμα

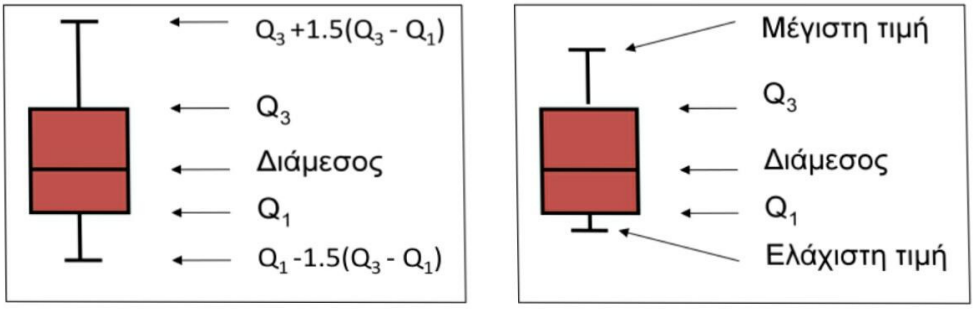
Το ιστόγραμμα (histogram*)* είναι ένας αντίστοιχος τύπος γραφικής αναπαράστασης με το ραβδόγραμμα, μόνο που στον οριζόντιο άξονα τοποθετούνται οι κλάσεις αντί των στοιχείων του δείγματος (Νικήτας, 2013).Επίσης**,** δεν υπάρχει κενό διάστημα μεταξύ των ράβδων του ιστογράμματος. Στον κάθετο άξονα του ιστογράμματος μπορεί να είναι η συχνότητα fi, η σχετική συχνότητα (ποσοστό) pi, η αθροιστική συχνότητα Fi, ή ακόμα η σχετική αθροιστική συχνότητα Pi για την κάθε i ομάδα. Αντίστοιχα με το ραβδόγραμμα σχετικής συχνότητας, το ιστόγραμμα σχετικής συχνότητας δίνει μια εκτίμηση του γραφήματος της συνάρτησης πυκνότητας πιθανότητας fX(x) της συνεχούς μεταβλητής Χ, βασισμένη στο δείγμα και στην επιλεγμένη ομαδοποίηση των παρατηρήσεων (Κουγιουμτζής, 2014).

δ) Θηκόγραμμα (boxplot)

Το θηκόγραμμα (boxplot)απαρτίζεται από ένα ορθογώνιο με δύο κεραίες, η μία στην κάτω βάση του ορθογωνίου με σχήμα αντεστραμμένου Τ και η άλλη στην επάνω βάση του σε σχήμα Τ. Η κάτω βάση του ορθογωνίου βρίσκεται στο Q1 και η επάνω στο Q3. Η διάμεσος παριστάνεται με ένα ευθύγραμμο οριζόντιο τμήμα στο εσωτερικό

του ορθογωνίου. Το μήκος των βάσεων του ορθογωνίου είναι αυθαίρετο. Οι κεραίες, φράκτες(whiskers), εκτείνονται μέχρι τις τιμές Q3 + 1.5(Q3 - Q1) και Q1 – 1.5(Q3 - Q1) (Αν η μέγιστη ή η ελάχιστη τιμή του δείγματος βρίσκονται εντός των περιοχών αυτών, τότε οι φράκτες μετατοπίζονται στη μέγιστη ή στην ελάχιστη τιμή. Αν υπάρχουν ακραίες τιμές αυτές εμφανίζονται ως σημεία εκτός των φρακτών (Νικήτας, 2013).

Σχήμα 1 Θηκόγραμμα



Στην περιγραφική στατιστική, σκοπός του θηκογράμματος (boxplot),αποτελεί το γεγονός ότι θεωρείται ένας βολικός τρόπος γραφικής αποτύπωσης πέντε αριθμητικών δεδομένων μιας σειράς παρατηρήσεων: της μικρότερης παρατήρησης του πρώτου τεταρτημόριου (Q1), της διαμέσου (δ) του τρίτου τεταρτημόριου (Q3), και της μεγαλύτερης παρατήρησης. Επιπλέον το θηκόγραμμα αποτυπώνει τις διαφορές ανάμεσα στους πληθυσμούς. Οι αποστάσεις μεταξύ των διαφόρων τμημάτων του θηκογράμματος βοηθούν να φανεί το μέγεθος της διασποράς και η ασυμμετρία των δεδομένων(https://eclass.uoa.gr/).

2.4 Μέτρα Κεντρικής Τάσης

Τα μέτρα κεντρικής τάσης χρησιμοποιούνται για να περιγράψουν τη θέση που κατέχει το σύνολο τον δεδομένων.

Μέση Τιμή(mean). Η αλλιώς και αριθμητικός μέσος όρος είναι το άθροισμα των τιμών όλων των παρατηρήσεων όπου αυτό διαιρείται με το πλήθος των παρατηρήσεων. Υπολογίζεται και ερμηνεύεται στατιστικά στις ποσοτικές μεταβλητές. Χρειάζεται ιδιαίτερη προσοχή στην περίπτωση ποιοτικών μεταβλητών που οι τιμές αυτών έχουν κωδικοποιηθεί με αριθμούς , όπου η μέση τιμή δεν έχει

νόημα. Μια εξαίρεση αποτελεί στην περίπτωση όπου η ποιοτική μεταβλητή λαμβάνει μόνο δυο τιμές για παράδειγμα "ΝΑΙ" και "ΟΧΙ" τα οποία μπορούν να έχουν κωδικοποιηθεί με τους αριθμούς 0 και 1. Υπάρχει και η αποκλειστική περίπτωση του σταθμισμένου αριθμητικού μέσου όρου (weighted average), όπου, αποδίδεται άνισο βάρος ή σημασία σε όλες τις παρατηρήσεις. Ο υπολογισμός του αριθμητικού μέσου όρου υπολογίζεται αθροίζοντας όλες τις τιμές (θετικές και αρνητικές) του συνόλου δεδομένων και διαιρώντας αυτό το άθροισμα με τον συνολικό αριθμό των παρατηρήσεων (Ali& Bhaskar, 2016).

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Και ο τύπος του πληθυσμού είναι: $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$

Υπάρχουν καταστάσεις στις οποίες το σύνολο δεδομένων έχει ακραίες τιμές στο χαμηλότερο ή υψηλότερο όριο, που στη στατιστική γλώσσα ορίζονται ως ακραίες τιμές(extreme value). Σε τέτοιες περιπτώσεις, η χρήση του αριθμητικού μέσου αντενδείκνυται, καθώς επηρεάζεται εύκολα από αυτές τις ακραίες τιμές (Manikandan, 2011). Για παράδειγμα, ο αριθμητικός μέσος των δεδομένων 2, 3, 5, 2, 22 θα είναι 16,4 που όπως παρατηρείται δεν μπορεί να θεωρηθεί καλός αντιπρόσωπος των δεδομένων. Ως εκ τούτου, σε αυτήν την περίπτωση χρησιμοποιείται ένα άλλο μέτρο που ονομάζεται διάμεσος(Amit, 2015).

Διάμεσος(median)Μ. Είναι αυτή η τιμή της μεταβλητής όπου το 50% των τιμών είναι μικρότερο από εκείνη και το υπόλοιπο 50% μεγαλύτερο. Στην περίπτωση που το πλήθος των παρατηρήσεων είναι περιττό, τότε η διάμεσος εφόσον οι παρατηρήσεις είναι διαταγμένες σε αύξουσα τιμή,είναι η μεσαία παρατήρηση. Διαφορετικά αν το πλήθος είναι άρτιο τότε η διάμεσος υπολογίζεται από την μέση τιμή των δύο μεσαίων παρατηρήσεων. Να σημειωθεί πως χρησιμοποιείται αποκλειστικά για ποσοτικές μεταβλητές(Δαφέρμος, 2011) Δεδομένου ότι η διάμεσος δεν επηρεάζεται από ακραίες τιμές, προτιμάται για την περιγραφή των δεδομένων που εμφανίζουν ακραίες τιμές ως μέτρο θέσης από τη μέση τιμή, η οποία αντιθέτως επηρεάζεται πολύ από τις ακραίες τιμές(Νικήτας, 2013). Αξίζει να σημειωθεί ότι η ύπαρξη ακραίων τιμών(παρατηρήσεων) στο δείγμα δυσκολεύει τη στατιστική περιγραφή ανάλυση. Γι'

αυτό πρέπει εκ των προτέρων να αποφασίζεται η συμπερίληψη των ακραίων τιμών, η αγνόηση ή η ειδική διαχείρισή τους.

Επικρατούσα τιμή (mode) **Μο** χαρακτηρίζεται η τιμή εκείνη των δεδομένων που έχει τη μεγαλύτερη συχνότητα εμφάνισης. Όταν υπάρχουν περισσότερες από μια τιμές οι οποίες έχουν την ίδια συχνότητα εμφάνισης, τότε τα δεδομένα αυτά έχουν περισσότερες επικρατούσες τιμές (Amit, 2015). Η κατανομή των δεδομένων που έχουν μια μόνο επικρατούσα τιμή λέγεται μονοκόρυφη (unimodal) ενώ εάν έχουν δύο επικρατούσες τιμές λέγεται δικόρυφη (bimodal). Χρησιμοποιείται και για τις ποσοτικές μεταβλητές αλλά και για τις ποιοτικές.

Το πρώτο τεταρτημόριο Q1 (Q1 quartile)διαιρεί τα δεδομένα σε δύο μέρη, έτσι ώστε, όταν τα δεδομένα είναι διατεταγμένα κατ' αύξουσα σειρά μεγέθους, το μέρος με τις μικρότερες παρατηρήσεις να αντιστοιχεί στο 25% των δεδομένων. Το τρίτο τεταρτημόριο Q3 (Q3 quartile*)* διαιρεί τα δεδομένα σε δύο μέρη, έτσι ώστε, όταν τα δεδομένα είναι διατεταγμένα κατ' αύξουσα σειρά μεγέθους, το μέρος με τις μεγαλύτερες παρατηρήσεις να αντιστοιχεί στο 25% των δεδομένων (Νικήτας, 2013).

Εύρος Τα σημαντικότερα πλεονεκτήματα αποτελούν ο εύκολος τρόπος υπολογισμού του και το γεγονός ότι περιλαμβάνει και τις ακραίες τιμές της κατανομής. Εκτός από τα πλεονεκτήματα υπάρχουν και τα μειονεκτήματα, όπως και οι αλλοιώσεις από τις ακραίες τιμές, με αποτέλεσμα να μην παρουσιάζει σε πολλές περιπτώσεις, μια αντιπροσωπευτική εικόνα της διασποράς της κατανομής. Επιπλέον, δεν παρέχει καμία πληροφορία αναφορικά με την με τη διασπορά των τιμών μεταξύ των άκρων της κατανομής.

2.5 Μέτρα Διασποράς

Τα μέτρα διασποράς χρησιμοποιούνται κατά κύριο λόγο στις επιχειρηματικές εφαρμογές καθώς αποτελούν σημαντικό συμπλήρωμα στις μελέτες τους. Με τη βοήθεια αυτών, επιτυγχάνεται μια πληρέστερη εικόνα της καταστάσεως που εξετάζεται και κατά συνέπεια οδηγούμαστε σε ασφαλέστερη λήψη ορθών επιχειρηματικών αποφάσεων. Αν για παράδειγμα ένας οικονομολόγος που έχει αναλάβει μια μελέτη για κάποια επιχείρηση και παρατηρήσει ότι η διασπορά των δεδομένων γύρω από το μ είναι μεγάλη, τότε θα πρέπει να εξετάσει πολύ προσεκτικά

τις πληροφορίες που λαμβάνονται από το μ και να προβληματιστεί αρκετά για τη λήψη κάποιας απόφασης.

Τα μέτρα διασποράς θεωρούνται απαραίτητα για τη χρήση των οικονομικών και επιχειρηματικών εφαρμογών, καθώς με τη χρήση τους επιτυγχάνεται ένα καλύτερο αποτέλεσμα που εξετάζεται και κατά συνέπεια επιτυγχάνεται καλύτερη επιχειρηματική απόφαση.

Τα πιο συνήθη μέτρα διασποράς που χρησιμοποιούνται στην στατιστική για την μέτρηση της μεταβλητότητας είναι :

-η μέση απόκλιση,
-η τυπική απόκλιση,
-η διακύμανση/διασπορά,
-το ημιενδοτεταρτημοριακό εύρος,
-ο συντελεστής μεταβλητότητας (CV) (χαρακτηρίζεται και ως μέτρο σχετικής θέσεως)

Ένα μέτρο διασποράς είναι η μέση απόκλιση (M.A.) (meandeviation) που ορίζεται ως ο μέσος αριθμητικός των απόλυτων διαφορών των τιμών της μεταβλητής από το μ. Δίνεται από τον τύπο:

$$M.A. = \frac{\sum |x_i - \mu|}{N}$$

όπου N ο αριθμός των παρατηρήσεων του πληθυσμού.

Όσο πιο μικρό είναι το αποτέλεσμα, τόσο πιο κοντά στο μ βρίσκονται οι παρατηρήσεις, που σημαίνει ότι τόσο αντιπροσωπευτικός και αξιόπιστος είναι ο μ. Λόγω των απόλυτων τιμών, δεν είναι εύκολος ο υπολογισμός του M.A., γι' αυτό χρησιμοποιούνται άλλα μέτρα διασποράς

Η διακύμανση (Variance) είναι ένα μέτρο για το πόσο διεσπαρμένη είναι η κατανομή. Δίνει μια ένδειξη του πόσο κοντά συσσωρεύεται μια μεμονωμένη παρατήρηση σχετικά με τη μέση τιμή (Myles&Gin, 2000).

Ο τύπος της είναι:

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} = \frac{\sum x_i^2}{N} - \mu^2$$

$$s^2 = \frac{1}{n-1}\sum (x - \bar{x})^2 = \frac{1}{n-1}[\sum x_i^2 - \frac{(\sum x_i)^2}{n}]$$

Ο μεν πρώτος αναφέρεται στον πληθυσμό, ενώ ο δεύτερος στο δείγμα.

2.Επειδή η διακύμανση εκφράζεται μέσω του τετραγώνου της μεταβλητής, γι' αυτό παίρνουμε τη θετική τετραγωνική ρίζα της διακύμανσης που ονομάζεται τυπική απόκλιση και η οποία εκφράζεται με τις ίδιες μονάδες μέτρησης με τη μονάδα μέτρησης της μεταβλητής. Η τυπική απόκλιση ορίζεται: $\sigma = \sqrt{\sigma^2}$ και $s = \sqrt{s^2}$.

Όσο μικρότερες είναι οι τιμές της διασποράς και της τυπικής απόκλισης, τόσο πιο συγκεντρωμένες γύρω από το μ βρίσκονται οι τιμές της μεταβλητής. Επίσης, είναι φανερό ότι οι τιμές της διασποράς κυμαίνονται μεταξύ 0 και ∞.

Ένα άλλο μέτρο μεταβλητότητας είναι ο συντελεστής σχετικής μεταβλητότητας Ορίζεται ως εξής:

$$CV = \frac{\sigma}{\mu} 100$$

$$CV = \frac{s}{\bar{x}} 100$$

Για τον πληθυσμό και για το δείγμα αντίστοιχα.

Είναι καθαρός αριθμός, απαλλαγμένος από μονάδες μέτρησης της μεταβλητής. Εκφράζει το 'άπλωμα' των τιμών σε σχέση με το μέσο. Επίσης, χρησιμοποιείται για συγκρίσεις ομάδων μεταξύ τους (είτε οι ομάδες εκφράζονται με ίδιες μονάδες μέτρησης είτε όχι). Επίσης, χρησιμοποιείται για την εξέταση της ομοιογένειας μέσα στην ίδια ομάδα. Επίσης, όταν ο CV δεν ξεπερνά το 10%, θα λέμε ότι το δείγμα είναι ομοιογενές.

Το **ενδοτεταρτημοριακό εύρος** (interquantilerange) είναι η διαφορά του πρώτου από το τρίτο τεταρτημόριο. Επομένως, όσο μικρότερο είναι αυτό το διάστημα, τόσο

μεγαλύτερη θα είναι η συγκέντρωση των τιμών και άρα μικρότερη η διασπορά των τιμών της μεταβλητής. Το μισό του ενδοτεταρτημοριακού εύρους είναι γνωστό ως **ημιενδοτεταρτημοριακό εύρος** (semi-interquantilerange) και συμβολίζεται με Q. Μετριέται με τις ίδιες μονάδες της μεταβλητής και δεν εξαρτάται από όλες τις τιμές αλλά μόνο από εκείνες που περιλαμβάνονται στον υπολογισμό του πρώτου και τρίτου τεταρτημορίου.

2.6 Μέτρα ασυμμετρίας-Μέτρα κύρτωσης

Η κατανομή ενός συνόλου δεδομένων μπορεί να είναι είτε συμμετρική είτε μη συμμετρική. Σαν αριθμητικά μέτρα καθορισμού της ασυμμετρίας έχουν προταθεί διάφοροι παράμετροι εκ' των οποίων σπουδαιότεροι είναι οι εξής:

Συντελεστές ασυμμετρίας κατά Pearson

Ορίζονται από τις σχέσεις:

$$\gamma_1 = \frac{\overline{x} - M_o}{s}, \gamma_2 = \frac{3(\overline{x} - M)}{s}$$

και λέγονται πρώτος και δεύτερος συντελεστής ασυμμετρίας του Pearson αντίστοιχα.

Σε περίπτωση μέτριας ασυμμετρίας ισχύει:

$$\gamma_1 \approx \gamma_2 = \gamma$$

Είναι φανερό ότι:

γ=0, συμμετρία και ισχύει $\overline{x} = M = M_o$

γ<0, αρνητική ασυμμετρία και ισχύει $\overline{x} < M < M_o$

γ>0, θετική ασυμμετρία και ισχύει $\overline{x} > M > M_o$

Συντελεστής ασυμμετρίας του Bowley

Ένα άλλο μέτρο ασυμμετρίας είναι και η ποσότητα:

$$s_A = \frac{Q_1 + Q_3 - 2M}{Q_3 - Q_1} = \frac{(Q_3 - M) - (M - Q_1)}{Q_3 - Q_1}$$

που λέγεται συντελεστής ασυμμετρίας του Bowley ή τεταρτημοριακός συντελεστής. Παίρνει τιμές μεταξύ –1 και 1. Και ισχύουν τα ακόλουθα:

$s_A = 0$, συμμετρία κατά Bowley

$0 < s_A < 1$, θετική ασυμμετρία κατά Bowley

$-1 < s_A < 0$, αρνητική ασυμμετρία κατά Bowley

Επίσης ισχύουν:

Αν $s_A = 0$, συμμετρία και ισχύει $Q_3 - M = M - Q_1 \Leftrightarrow M = \frac{Q_3 + Q_1}{2}$

Αν $s_A = 1$, μεγαλύτερη θετική ασυμμετρία με το πρώτο τεταρτημόριο να προσεγγίζει τη διάμεσο.

Αν $s_A = -1$, μεγαλύτερη αρνητική ασυμμετρία με το τρίτο τεταρτημόριο να τείνει στη διάμεσο.

Συντελεστής ασυμμετρίας με βάση τις ροπές

Γενικεύοντας την έννοια της διασποράς μπορεί κανείς να ορίσει τις λεγόμενες κεντρικές ροπές (centralmoments)t-τάξης από τη σχέση:

$$\mu_t = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^t}{n}, t = 2,3,...$$

$$\mu_t = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^t}{\sum_{i=1}^{n} f_i}, t = 2,3,...$$

αν t=2, τότε η κεντρική ροπή δεύτερης τάξης συμπίπτει με τη διασπορά.

Ο συντελεστής ασυμμετρίας με βάση τις ροπές ορίζεται σαν το πηλίκο

$$\beta_1 = \frac{\mu_3^2}{\mu_2^2} \ \eta \ \beta_1 = \frac{\mu_3}{\mu_2^{3/2}}$$

και εκφράζει τη συμμετρία αν είναι ίσο με μηδέν και αν είναι θετικό δηλώνει ασυμμετρία. Το είδος της ασυμμετρίας καθορίζεται από το πρόσημο της κεντρικής ροπής τρίτης τάξης. Αν $\mu_3 > 0$, θετική ασυμμετρία, ενώ $\mu_3 < 0$, αρνητική ασυμμετρία.

Μέτρα Κύρτωσης (Measures of kurtosis)

Τα μέτρα αυτά αφορούν το βαθμό συγκέντρωσης των δεδομένων γύρω από το μέσο και τα άκρα της κατανομής.

Μια κατανομή η οποία έχει σχετικά μεγάλη συχνότητα (κορυφή) και επομένως μεγάλη συγκέντρωση τιμών γύρω από το μέσο λέγεται λεπτόκυρτη (leptokurtic), ενώ αν η μέγιστη συχνότητά της είναι σχετικά μικρή λέγεται πλατύκυρτη (platykurtic). Κατανομές που προσεγγίζονται από την κανονική κατανομή λέγονται μεσόκυρτες (mesokurtic).

Ένα μέτρο που εκφράζει το βαθμό κυρτότητας μιας κατανομής είναι ο συντελεστής κύρτωσης του Pearson ο οποίος ορίζεται από τον τύπο:

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

Επειδή για κανονικές κατανομές έχουμε $\beta_2 = 3$ συνηθίζεται να μετράμε την κυρτότητα με τη διαφορά $\beta_2 - 3$, η οποία για λεπτόκυρτες κατανομές παίρνει θετικές τιμές (θετική κύρτωση), ενώ για πλατύκυρτες κατανομές γίνεται αρνητική (αρνητική κύρτωση).

Κεφάλαιο 3 : Επαγωγική Στατιστική

Η επαγωγική στατιστική αναφέρεται σε ένα τυχαίο δείγμα δεδομένων που περιέχει ένα πληθυσμό με σκοπό την ερμηνεία και την εξαγωγή των συμπερασμάτων. Αποτελεί ένα πολύ χρήσιμο εργαλείο σε περιπτώσεις που είναι αδύνατη η εξέταση μεμονωμένων μελών ενός ολόκληρου πληθυσμού (Satake, 2015).

3.1 Αντιπροσωπευτικότητα του δείγματος

Η ακρίβεια της γενίκευσης των αποτελεσμάτων από το δείγμα στον πληθυσμό από τον οποίο προέρχεται, σχετίζεται άμεσα με την αντιπροσωπευτικότητα του δείγματος. Προκειμένου να υπάρχει μία γενίκευση αποτελεσμάτων στην έρευνα που υλοποιείται στο συνολικό πληθυσμό, είναι αναγκαίο να υπάρχει μία συμφωνία με τις αρχές της δειγματοληψίας (sampling). Η δειγματοληψία, αφορά την υλοποίηση της επιλογής του δείγματος από τον πληθυσμό, με αντικειμενικό σκοπό, να αποτελέσει όσο το δυνατόν αντιπροσωπευτικότερο, δηλαδή, να διακρίνεται από τα στοιχεία εκείνα του πληθυσμού στον οποίο ανήκει. Ως εκ τούτου, ο ερευνητής πρέπει να επιλέξει ένα όσο το δυνατόν πιο αντιπροσωπευτικό δείγμα.

Η αντιπροσωπευτικότητα του δείγματος προσδιορίζεται από δύο παραμέτρους:

α) τον τρόπο επιλογής των στοιχείων που θα αποτελέσουν το δείγμα, και

β) το μέγεθος του δείγματος.

Άρα, τα σημαντικότερα στοιχεία που πρέπει να προσέξει ιδιαίτερα ο ερευνητής σχετικά με τη δειγματοληψία είναι: α) να ταιριάζει το δείγμα όσο πιο πολύ γίνεται με τον πληθυσμό εξασφαλίζοντας έτσι μία αρτιότερη προσέγγιση στις απαιτήσεις για την αληθή τιμή του πληθυσμού β) να κάνει τέτοια επιλογή του μεγέθους του δείγματος που να είναι εφικτή από τη γενικότερη έρευνα που υλοποιεί.

Ο προσδιορισμός του μεγέθους του δείγματος, αποτελεί μία διαδικασία που προσδιορίζεται από ορισμένες παραμέτρους: η επιθυμητή ακρίβεια της έρευνας (το επιθυμητό μέγεθος του δειγματοληπτικού σφάλματος), το ποσό και ο χρόνος που διατίθενται για την υλοποίησή της, η ομοιογένεια ως προς τον πληθυσμό και τέλος

τον αριθμό των μεταβλητών και των υποομάδων που θα συμπεριληφθούν στην έρευνα καθώς και ο συνολικός αριθμός των τιμών κάθε μεταβλητής(Ρόντος & Παπαπάνης, 2006).

Οι Τύποι δειγμάτων που χρησιμοποιούνται ανήκουν σε δύο κατηγορίες:

Δείγματα υπολογιζόμενα με πιθανότητες (probability samples), ώστε κάθε στοιχείο του πληθυσμού, χαρακτηρίζεται σαν να έχει την ίδια πιθανότητα να επιλεγεί και να απορριφθεί και ο πληθυσμός είναι γνωστός πριν την υλοποίηση της έρευνας κατά τη διάρκεια του σχεδιασμού της.

Δείγματα υπολογιζόμενα χωρίς πιθανότητες (non probability samples): στην περίπτωση αυτή κάποια στοιχεία του πληθυσμού, απορρίπτονται εσκεμμένα από τη διαδικασία της επιλογής, ενώ κάποια άλλα επιλέγονται, (δηλαδή δεν έχουν όλα τα στοιχεία του πληθυσμού την ίδια πιθανότητα επιλογής). Το άτομο που ερευνά δεν γνωρίζει το ακριβές μέγεθος του πληθυσμού και παρατηρείται μία μεροληψία σε κάποια στοιχεία.

3.2 Εκτιμητική (Διάστημα Εμπιστοσύνης)

Η μέθοδος η οποία ασχολείται με τον καθορισμό ενός φάσματος πιθανών τιμών μιας υπό εκτίμηση παραμέτρου ονομάζεται μέθοδος των διαστημάτων εμπιστοσύνης(confidence intervals). Η ανάπτυξη της μεθόδου αυτής χρησιμοποιεί μόνο τις αρχές της Θεωρίας Πιθανοτήτων χωρίς να χρειάζεται νέες έννοιες Στατιστικής Συμπερασματολογίας.

Το διάστημα εμπιστοσύνης (confidence interval) χρησιμοποιείται με σκοπό την ασφαλέστερη εκτίμηση μιας παραμέτρου ενός πληθυσμού βάση ενός τυχαίου δείγματος από αυτόν τον πληθυσμό. Το διάστημα αυτό, παρέχει ένα φάσμα εύλογων (πιθανών) τιμών της παραμέτρου, συνοδευόμενο από τον βαθμό εμπιστοσύνης που διατηρείται, ότι περιέχει την πραγματική τιμή της παραμέτρου(Πανάρετος & Ξεκαλάκη, 2003).

Δεδομένου ότι η σημειακή εκτίμηση δίνει μια μόνο τιμή που λαμβάνεται ως η καλύτερη εκτίμηση μιας παραμέτρου, εάν συλλέγονται άλλα δεδομένα από τον ίδιο πληθυσμό, η σημειακή εκτίμηση του σημείου μπορεί να αλλάξει(Ali & Bhaskar,

2016). Το μειονέκτημα της σημειακής εκτίμησης είναι ότι δεν δίνει απαντήσεις της μορφής «η μέση τιμή είναι ίση με…», αλλά της μορφής «η μέση τιμή είναι περίπου ίση με…». Έτσι, τα διαστήματα εμπιστοσύνης είναι ο κλάδος της Στατιστικής Συμπερασματολογίας που ασχολείται με τον προσδιορισμό διαστημάτων, περιοχών γενικότερα, που περιέχουν με μεγάλη πιθανότητα άγνωστες παραμέτρους κάποιας κατανομής.

Μέσω της εκτίμησης σε διάστημα, προκύπτει η βεβαιότητα ότι το υπολογιζόμενο διάστημα περιέχει την πραγματική τιμή της εξεταζόμενης παραμέτρου. Αναμένεται ότι η πραγματική τιμή του πληθυσμού θα εμπίπτει σε αυτό το διάστημα με το επιθυμητό επίπεδο εμπιστοσύνης. Για τον λόγο αυτό, δίνεται το όνομα «διάστημα εμπιστοσύνης» (Nickerson, 2000). Το διάστημα πρέπει να είναι σε λογικά όρια. Αυτά τα όρια (ή τα διαστήματα) υπολογίζονται στατιστικά και αναφέρονται ως διαστήματα εμπιστοσύνης ή όρια εμπιστοσύνης. Δεδομένου ότι εκτιμούμε την παράμετρο του πληθυσμού από τιμές δείγματος, δεν μπορεί ποτέ να υπάρξει καμία εκτίμηση με 100% εμπιστοσύνη. Συνήθως, το επίπεδο εμπιστοσύνης 95% θεωρείται κατάλληλο. Δηλαδή, θεωρείται ότι «με 95% εμπιστοσύνη είναι πολύ πιθανό η παράμετρος του πληθυσμού να συμπέσει σε κάποιο σημείο μεταξύ του διαστήματος εμπιστοσύνης» (Amit, 2015).

Επειδή το προτεινόμενο διάστημα συνοδεύεται με ένα συντελεστή εμπιστοσύνης, το διάστημα αυτό καλείται διάστημα εμπιστοσύνης με βαθμό εμπιστοσύνης γ. Ο αριθμός γ (γ=1-α) εκφράζει την προσδοκώμενη ακρίβεια της εκτίμησης, ενώ ο αριθμός α(επίπεδο σημαντικότητας)(significance level) εκφράζει τον βαθμό ανεκτικότητας ώστε το διάστημα να μην περιέχει την πραγματική τιμή της παραμέτρου. Το διάστημα εμπιστοσύνης έχει μεγαλύτερη έκταση όσο μεγαλύτερος είναι ο συντελεστής εμπιστοσύνης. Δηλαδή, ένα διάστημα εμπιστοσύνης 99% έχει μεγαλύτερη έκταση από ένα διάστημα εμπιστοσύνης 95%, ώστε να είμαστε σίγουροι ότι στο διάστημα αυτό βρίσκεται η παράμετρος που εκτιμάμε.

3.3 Έλεγχος υποθέσεων

Γενικά, σε έναν στατιστικό έλεγχο χρησιμοποιούμε δύο συγκεκριμένες διατυπώσεις στατιστικών υποθέσεων που είναι γνωστές ως μηδενική υπόθεση (H_0) και η

εναλλακτική υπόθεση (H_1). Η θεωρία που αναπτύσσει ένας ερευνητής από μια παρατήρηση συνήθως προβλέπει την εμφάνιση κάποιου αποτελέσματος. Η υπόθεση ότι το αποτέλεσμα αυτό εμφανίζεται ονομάζεται Πειραματική Υπόθεση (Experimental Hypothesis), αλλά κυρίως έχει επικρατήσει ο όρος Εναλλακτική Υπόθεση (Alternative Hypothesis)όπου συμβολίζεται ως H_1. Η αντίθετη της εναλλακτικής υπόθεσης είναι η Μηδενική Υπόθεση (Null Hypothesis)όπου συμβολίζεται με H_0. Με δεδομένο ότι η μηδενική υπόθεση είναι η αντίστροφη της εναλλακτικής, σε περίπτωση που τελικά γίνει αποδεκτή η μηδενική υπόθεση, ο ερευνητής θα αναφέρει ότι το αποτέλεσμα της εναλλακτικής υπόθεσης απορρίπτεται(Λαγουμιντζής, Βλαχόπουλος & Κουτσογιάννης, 2015). Σχεδόν πάντα η απόκτηση γνώσης σχετικά με μια παράμετρο του πληθυσμού είναι ακατόρθωτη, χωρίς την εφαρμογή του ελέγχου υπόθεσης ή του ελέγχου σπουδαιότητας. Η στρατηγική αυτή χρησιμοποιείται για να αποφασιστεί εάν τα αποτελέσματα του εξεταζόμενου δείγματος προσφέρονται για γενικεύσεις ως προς τον γενικό πληθυσμό επί μίας ή περισσότερων υποθέσεων (Amit, 2015).

Ο έλεγχος υπόθεσης ξεκινά συνήθως με κάποιες παραδοχές, υποθέσεις ή ισχυρισμούς σχετικά με μια ή περισσότερες συγκεκριμένες παραμέτρους ενός εξεταζόμενου πληθυσμού (Amit, 2015). Μηδενική υπόθεση καλείται η υπόθεση που κάνουμε για μια συγκεκριμένη τιμή της παραμέτρου και η οποία είναι η σοβαρότερη υπόθεση στον έλεγχο. Η μηδενική υπόθεση συνήθως περιστρέφεται γύρω από τη θέση ότι «δεν υπάρχει σημαντική διαφορά μεταξύ των δειγμάτων». Η εναλλακτική υπόθεση είναι εκείνη για την οποία ελέγχεται η μηδενική υπόθεση(Ali& Bhaskar, 2016).

Στην πλειοψηφία των έγκυρων στατιστικών αναλύσεων, το επίπεδο σημαντικότητας εκφράζεται ως ποσοστό της τάξης του 5% ή του 1% (0,05 ή 0,01), καθώς οι τιμές αυτές έχουν παγιωθεί μέσα από βαρυσήμαντες στατιστικές δοκιμές (Nickerson, 2000). Στη συνέχεια καταρτίζεται ένας πίνακας και μέσω βασικών μαθηματικών υπολογισμών αναδεικνύεται η κρίσιμη τιμή μιας μετρούμενης μεταβλητής. Εάν η υπολογισμένη τιμή (στατιστική) είναι ίση ή μικρότερη από την κρίσιμη τιμή, η διαφορά μεταξύ του αποτελέσματος και της αναμενόμενης τιμής είναι ασήμαντη και αυτή η ασήμαντη διαφορά μπορεί να αποδοθεί σε δειγματοληπτικό σφάλμα. Έτσι, η μηδενική υπόθεση γίνεται αποδεκτή. Αντίθετα, αν η υπολογισμένη τιμή είναι

ψηλότερη από την κρίσιμη τιμή, η διαφορά θεωρείται σημαντική και δεν μπορεί να αποδοθεί σε δειγματοληπτικό σφάλμα, επομένως η μηδενική υπόθεση απορρίπτεται (Nickerson, 2000). Κάθε φορά που λαμβάνουμε μια απόφαση σχετικά με τον πληθυσμό βάσει δείγματος, η απόφαση δεν μπορεί να είναι 100% αξιόπιστη. Οι δυνατότητες του στατιστικού αναλυτή περιορίζονται στην αποδοχή ή απόρριψη της μηδενικής υπόθεσης (Amit, 2015). Ο έλεγχος που εφαρμόζεται στη στατιστική για την παραπάνω εξέταση ακολουθεί τα εξής βήματα:

Ορισμός της μηδενικής ή αρχικής υπόθεσης H_0(null hypotheses).
Ορισμός της εναλλακτικής υπόθεσης H_1(alternative hypotheses).
Ορισμός του ελέγχου που εφαρμόζεται για την αποδοχή ή την απόρριψη της αρχικής υπόθεσης.
Εξαγωγή συμπερασμάτων

Ορισμός της απορριπτικής περιοχής της μηδενική υπόθεσης ή αλλιώς της κρίσιμης περιοχής του ελέγχου (της περιοχής του δείγματος χώρου για την οποία απορρίπτεται η αρχική υπόθεση).

Ακολούθως μετά τον προσδιορισμό της μηδενικής και εναλλακτικής υπόθεσης αποφασίζεται το επίπεδο σημαντικότητας το οποίο χρησιμοποιείται ως κριτήριο απόρριψης της μηδενικής υπόθεσης. Είναι το μικρότερο ε.σ. για το οποίο απορρίπτεται η μηδενική υπόθεση. Η τιμή **p** χρησιμοποιείται για να ερμηνεύσει τον έλεγχο μέσω των στατιστικών προγραμμάτων. Βασικό πλεονέκτημα της χρήσης του p-value, αποτελεί το γεγονός ότι δεν απορρίπτουμε ή δεχόμαστε απλώς την H_0 αλλά, υπάρχει η δυνατότητα να δούμε και πόσο πιθανή ήταν η εμφάνιση του δείγματος x που πήραμε (υπό την H_0) ενώ επίσης είναι δυνατή και η σύγκριση άμεσα με όποιο a και αν γίνει η επιλογή. Και για αυτό p-value , απαιτεί και τη χρήση Η/Υ καθώς δεν μπορεί αλλιώς να υπολογιστεί για κάθε τιμή('Βασικές έννοιες θεωρίας ελέγχων υποθέσεων').

Είναι αριθμητική, κυμαίνεται μεταξύ 0 και 1 και ερμηνεύεται από ερευνητές για να αποφασίσουν εάν θα απορρίψουν ή θα διατηρήσουν την μηδενική υπόθεση. Εάν η τιμή **p** είναι μικρότερη από την αυθαίρετα επιλεγμένη τιμή (γνωστή ως α ή επίπεδο σημαντικότητας), η μηδενική υπόθεση απορρίπτεται. Σε κάθε περίπτωση, εάν η μηδενική υπόθεση απορριφθεί εσφαλμένα, αυτό είναι γνωστό ως σφάλμα τύπου I,

ενώ στην περίπτωση που γίνεται εσφαλμένα αποδοχή της μηδενικής υπόθεσης, ενώ στην πραγματικότητα είναι αληθής η εναλλακτική υπόθεση το είδος του σφάλματος είναι γνωστό ως σφάλμα τύπου II (Bajwa, 2015).

3.4 t- test δύο δειγμάτων

Η δοκιμή t- test ενός δείγματος (one samplet-test) χρησιμοποιείται σε περιπτώσεις προβλημάτων στα οποία θέλουμε να ελέγξουμε αν ένα δείγμα προέρχεται από κάποιο πληθυσμό με γνωστό μέσο όρο ή να ελέγξουμε αν ο μέσος όρος ενός δείγματος είναι ίσος με το μέσο όρο του γενικού πληθυσμού που θεωρούμε ότι είναι γνωστός (Λαγουμιντζής, Βλαχόπουλος & Κουτσογιάννης, 2015).

Η μηδενική υπόθεση H0 : $\theta = \theta_0$ υποδηλώνει ότι η εκτίμηση θ_0 της θ είναι σωστή. Επίσης, τιμές της θ 'κοντά' στη θ_0 υποστηρίζουν την ορθότητα της H_0 ενώ τιμές της θ 'μακριά' από τη θ0 δεν την υποστηρίζουν. Έτσι, χωρίζουμε τις δυνατές τιμές της παραμέτρου θ σε αυτές για τις οποίες αποδεχόμαστε την H_0 που αποτελούν την περιοχή αποδοχής και σ΄ αυτές για τις οποίες την απορρίπτουμε που αποτελούν την περιοχή απόρριψης (rejectionregion) που συμβολίζουμε R.

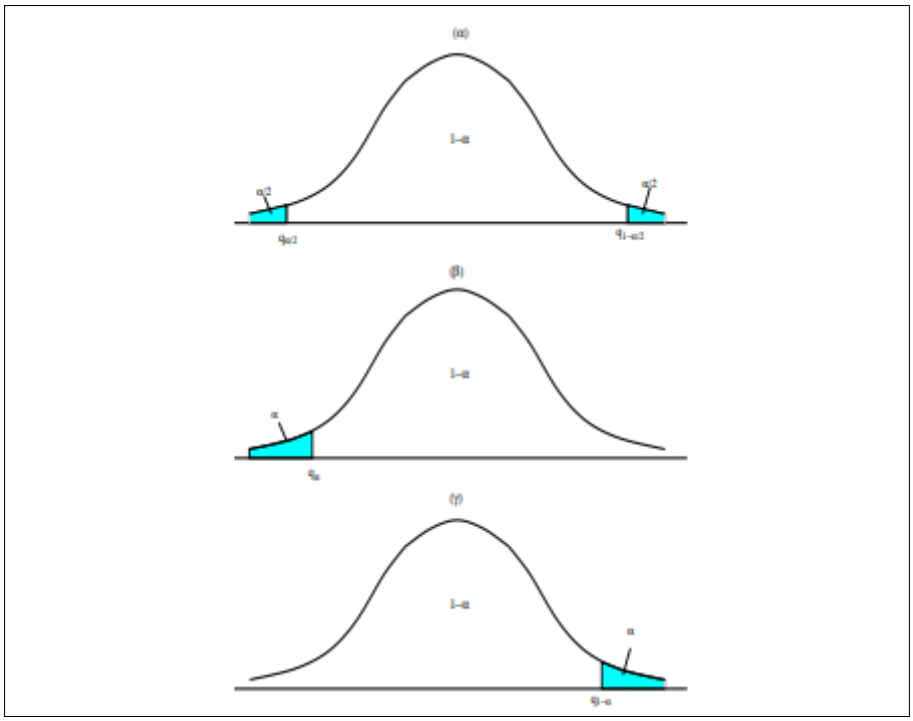
Η απόφαση για την αποδοχή ή απόρριψη της H_0 γίνεται βάση πιθανοτήτων κι όπως για τα διαστήματα εμπιστοσύνης έτσι κι εδώ ορίζουμε επίπεδο εμπιστοσύνης $(1-\alpha)$ για την απόφαση ελέγχου.

Ο έλεγχος ξεκινάει υποθέτοντας ότι η μηδενική υπόθεση H_0 είναι σωστή. Με βάση τη H_0 προσδιορίζουμε τη στατιστική ελέγχου q και την κατανομή της. Όταν ο έλεγχος είναι παραμετρικός υποθέτουμε κάποια κατανομή για την παράμετρο θ κι η q σχετίζεται άμεσα με τη θ (η q προκύπτει από μετασχηματισμό της θ). Όταν ο έλεγχος είναι μη παραμετρικός δε θεωρούμε κάποια κατανομή για τη θ κι η κατανομή της q βασίζεται σε άλλες ιδιότητες της θ. Η κατανομή της στατιστικής ελέγχου q δίνει την πιθανότητα η q να πάρει κάποια τιμή (αν η q είναι διακριτή τ.μ.) ή να βρίσκεται σ΄ ένα διάστημα τιμών (αν η q είναι συνεχής τ.μ.) όταν η H_0 είναι αληθής. Αντίστροφα μπορούμε να πούμε πως με βάση αυτή την κατανομή, αν παρατηρήσουμε τιμές της q που αντιστοιχούν σε μεγάλες πιθανότητες αυτό δείχνει πως η H_0 είναι αληθής, ενώ αν παρατηρήσουμε τιμές της q που αντιστοιχούν σε μικρές πιθανότητες αυτό υποδηλώνει αμφιβολία για την ισχύ της H_0. Άρα μη

πιθανές τιμές της q συνιστούν την απόρριψη της H_0. Η οριακή πιθανότητα για την αποδοχή ή απόρριψη της H_0 είναι το επίπεδο σημαντικότητας α κι αυτό καθορίζει την κρίσιμη τιμή (criticalvalue), ή τις κρίσιμες τιμές, της q για τον προσδιορισμό της περιοχής αποδοχής και της περιοχής απόρριψης R της H_0. Κατά κανόνα η περιοχή απόρριψης σχηματίζεται από τα άκρα της κατανομής της στατιστικής ελέγχου q όπως αυτά ορίζονται από τις κρίσιμες τιμές. Αν ο έλεγχος είναι δίπλευρος τότε οι κρίσιμες τιμές $q_{\alpha/2}$ και $q_{1-\alpha/2}$ ορίζουν την περιοχή απόρριψης της H_0 ως $R = \{q| q<q_{\alpha/2} \vee q>q_{1-\alpha/2}\}$, δηλαδή σχηματίζεται από τις δύο ουρές της κατανομής της q με συνολική πιθανότητα α. Αν ο έλεγχος είναι μμονόπλευρος τότε υπάρχει μόνο μία κρίσιμη τιμή, q_α ή $q_{1-\alpha}$, που ορίζει την περιοχή απόρριψης της H_0, $R = \{q| q<q\alpha\}$ για την αριστερή πλευρά και $R = \{q| q>q_{1-\alpha}\}$ για τη δεξιά πλευρά.

Στο παρακάτω σχήμα δίνονται σχηματικά οι περιοχές αποδοχής κι απόρριψης για δίπλευρο και μμονόπλευρο έλεγχο.

Σχήμα 1 Συνάρτηση πυκνότητας πιθανότητας της στατιστικής ελέγχου q, η περιοχή αποδοχής και η περιοχή απόρριψης (σκιασμένη), για δίπλευρο έλεγχο στο διάγραμμα (α) και μονόπλευρο έλεγχο στο (δ) και (γ).



Ο έλεγχος για ανεξάρτητα δείγματα χρησιμοποιείται σε περιπτώσεις προβλημάτων, στα οποία θέλουμε να συγκρίνουμε τις μέσες τιμές για την ίδια συνεχή μεταβλητή,

δύο δειγμάτων. Συνήθως, η σύγκριση αφορά δυο διαφορετικά δείγματα του ίδιου πληθυσμού που υποβάλλονται σε διαφορετική δοκιμασία και θέλουμε να συγκρίνουμε τις μέσες τιμές μιας μεταβλητής για τα δύο δείγματα(Λαγουμιντζής, Βλαχόπουλος & Κουτσογιάννης, 2015). Για παράδειγμα, αυτή η μέθοδος ανάλυσης μπορεί να χρησιμοποιηθεί για να προσδιοριστεί εάν η μέση τιμή ενός οχήματος τύπου sedan είναι σημαντικά διαφορετική από ένα όχημα τύπου SUV. Στην περίπτωση αυτή, ως μηδενική υπόθεση μπορεί να ορισθεί η υπόθεση ότι οι τύποι αυτοκινήτων SUV και Sedan έχουν ασήμαντες διαφορές ως προς την αξία, ενώ η εναλλακτική υπόθεση θα είναι ότι η μέση τιμή των τύπων SUV και sedan διαφέρει σημαντικά.

Κεφάλαιο 4 Απλή Παλινδρόμηση

4.1 Διάγραμμα διασποράς

Πολλές φορές η ανάλυση και η ερμηνεία απαιτεί την ύπαρξη περισσότερων από μία μεταβλητές λόγω της σχέσης μεταξύ των δύο μεταβλητών. Σε αυτές τις καταστάσεις είναι απαραίτητο να είναι γνωστές οι τιμές και των δύο. Μεγάλο ρόλο σε αυτό παίζει η συσχέτιση που χρησιμοποιείται ώστε να αποδειχτεί αν υπάρχει στατιστική σχέση μεταξύ των δυο μεταβλητών και αν υπάρχει τότε, χρησιμοποιείται η ανάλυση παλινδρόμησης, για την εκτίμηση του μοντέλου που θα αναλύει και τη σχέση των δυο μεταβλητών.

Θεωρώντας ότι υπάρχουν δυο μεταβλητές: X και Y, όπου X καλείται ανεξάρτητη μεταβλητή (independent variable), Y καλείται εξαρτημένη μεταβλητή (dependent variable), οι παρατηρήσεις παρουσιάζονται σε ζεύγη τιμών $(x_1, y_1),...,(x_n, y_n)$.

Στη συνέχεια δημιουργείται η γραφική παράσταση των δεδομένων με τα ζεύγη σε ένα ορθοκανονικό σύστημα συντεταγμένων, όπου στον άξονα X υπάρχουν οι τιμές της ανεξάρτητης μεταβλητής και στον άξονα Y της εξαρτημένης. Έτσι σχηματίζεται ένα πλήθος σημείων που ονομάζεται νέφος σημείων ή διάγραμμα διασποράς.

Η ανεξάρτητη μεταβλητή επιδρά στην εξαρτημένη προκαλώντας αρκετές αλλαγές με αποτέλεσμα να υπάρχει και η ανάλογη μεταβλητότητα της εξαρτημένης μεταβλητής.

Η απλούστερη σχέση που μπορεί να συνδέει δυο μεταβλητές είναι η γραμμική και θα ασχοληθούμε μόνο σε αυτήν την περίπτωση, αφού πολλές άλλες μορφές σχέσεων μπορούν εύκολα με κάποιους κατάλληλους μετασχηματισμούς των μεταβλητών να αναχθούν σε γραμμικές. Οι παρακάτω συναρτήσεις ανάγονται σε γραμμικές

θέτοντας $y = \ln y' \Rightarrow y = a + bx$

$y' = a' x'^{b}$ θέτοντας $y = \ln y', \ln a' = a, \ln x' = x \Rightarrow y = a + bx$

$z = c\exp(bx)$ θέτοντας $y = \ln z, a = \ln c \Rightarrow y = a + bx$

$y = a + b\frac{1}{z}$ θέτοντας $x = \frac{1}{z} \Rightarrow y = a + bx$

$z = \frac{1}{a + bx}$ θέτοντας $y = \frac{1}{z} \Rightarrow y = a + bx$

$z = \frac{1}{(a + bx)^2}$ θέτοντας $y = \frac{1}{z} \Rightarrow y = a + bx$

$\frac{1}{z} = a + b\frac{1}{1 + x'}$ θέτοντας $y = \frac{1}{z}, x = \frac{1}{1 + x'} \Rightarrow y = a + bx$

$y = a + b\sqrt{x'}$ θέτοντας $x = \sqrt{x'} \Rightarrow y = a + bx$

4.2 Γραμμικός συντελεστής συσχέτισης ρ

Σύμφωνα με τα προαναφερόμενα μετά την κατασκευή του διαγράμματος διασποράς, ακολουθεί ο υπολογισμός του συντελεστή συσχέτισης.

Συντελεστής Συσχέτισης ονομάζεται η ποσοτική μέτρηση της έντασης της (γραμμικής) σχέσης μεταξύ δυο μεταβλητών. Ο συντελεστής αυτός καλείται γραμμικός συντελεστής συσχέτισης και συμβολίζεται με ρ ενώ η αντίστοιχη εκτίμησή του με r.

Υπολογίζεται:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{[\sum_{i=1}^{n}(x_i - \overline{x})^2][\sum_{i=1}^{n}(y_i - \overline{y})^2]}} = \frac{n\sum_{i=1}^{n}xy - \sum_{i=1}^{n}x\sum_{i=1}^{n}y}{\sqrt{[n\sum_{i=1}^{n}x^2 - (\sum_{i=1}^{n}x)^2(\sum_{i=1}^{n}y)^2}} = \frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y}$$

Όπου

$$\mathrm{cov}(X,Y) = E(X-\bar{X})(Y-\bar{Y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \frac{1}{n}\sum_{i=1}^{n}xy - \overline{x}\,\overline{y}$$

Και ονομάζεται συνδιακύμανση των δυο μεταβλητών.

Σε περίπτωση που η συνδιακύμανση είναι θετικός αριθμός τότε οι μεταβλητές χαρακτηρίζονται και αυτές θετικά συσχετισμένες και ακολουθεί μία ομόρροπη μεταβολή. Άν η συνδιακύμανση είναι αρνητικός αριθμός τότε οι μεταβλητές χαρακτηρίζονται και αυτές αρνητικά συσχετισμένες και έπεται μία αντίρροπη μεταβολή. Και τέλος, αν η συνδιακύμανση είναι μηδενική τότε δεν παρουσιάζεται καμία γραμμική συμμεταβολή των δυο μεταβλητών και πρόκειται για μεταβλητές που χαρακτηρίζονται ασυσχέτιστες.

Η συνδιακύμανση δεν έχει τη δυνατότητα έκφρασης αντικειμενικού βαθμού της γραμμικής συμμεταβολής και ούτε χρησιμοποιείται για τη σύγκριση μεταξύ του βαθμού γραμμικής συμμεταβολής διαφορετικών κατανομών. Έτσι στη θέση της χρησιμοποιούμε ως μέτρο της γραμμικής συμμεταβολής όλων των μεταβλητών τον γραμμικό συντελεστή συσχέτισης που χαρακτηρίζεται ως καθαρός αριθμός.

Ιδιότητες ρ

Ο γραμμικός συντελεστής συσχέτισης διακρίνεται πάντα από το ίδιο πρόσημο με τη συνδιασπορά. Οι τιμές που παίρνει ανήκουν στο διάστημα -1 και 1 δηλ. $-1 \le \rho \le 1$

Σε περίπτωση που ισχύει $\rho = 1$ ή $\rho = -1$ τότε οι μεταβλητές θεωρούνται ότι έχουν τέλεια θετική ή αρνητική αντίστοιχα, γραμμική σχέση. Ο ρ θεωρείται καθαρός αριθμός

Όταν ισχύει $\rho = 0$ οι μεταβλητές θεωρούνται γραμμικά ασυσχέτιστες και διακρίνονται από άλλου είδους σχέση.

Επίσης Αν $|\rho| \le 0.30$ τότε δεν υφίσταται γραμμική συσχέτιση

Αν $0.30 \le |\rho| \le 0.50$, τότε η συσχέτιση χαρακτηρίζεται ως ασθενής.

Αν $0.50 \le |\rho| \le 0.70$,τότε η συσχέτιση χαρακτηρίζεται μέτρια.

Αν $0.70 \le |\rho| \le 0.80$,τότε η συσχέτιση χαρακτηρίζεται ως ισχυρή.

Αν $|\rho| \ge 0.80$, τότε η συσχέτιση χαρακτηρίζεται ως πολύ ισχυρή.

Αν $|\rho| = 1$, τότε η συσχέτιση χαρακτηρίζεται ως τέλεια.

Έλεγχος Στατιστικής Σημαντικότητας του Συντελεστή Συσχέτισης

Αυτό που ενδιαφέρει περισσότερο είναι, εάν η γραμμική σχέση μεταξύ των μεταβλητών X και Y , είναι στατιστικά σημαντική, δηλαδή αν ο πληθυσμιακός συντελεστής συσχέτισης είναι διάφορος του μηδενός.

Αν X και Y κατανέμονται κανονικά και r μια εκτίμηση του ρ, τότε μας δίνεται η δυνατότητα να ελέγξουμε:

$H_o : \rho = 0 \quad \text{vs} \quad H_1 : \rho \neq 0$

$H_o : \rho = 0 \quad \text{vs} \quad H_1 : \rho > 0$

$H_o : \rho = 0 \quad \text{vs} \quad H_1 : \rho < 0$

Αν ισχύει η μηδενική υπόθεση, τότε οι δυο μεταβλητές είναι ανεξάρτητες. Ο στατιστικός έλεγχος πραγματοποιείται με τη χρήση της ακόλουθης στατιστικής συνάρτησης:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \sim t_{n-2}$$

Ως γνωστό, χρειαζόμαστε έναν κανόνα για τη λήψη της απόφασης, ανάλογα βέβαια με τη μορφή της εναλλακτικής υπόθεσης. Οπότε έχουμε:

Απορρίπτουμε τη μηδενική υπόθεση αν

$$|t| \ge t_{n-2,\frac{a}{2}}$$

$$t > t_{n-2,a}$$

$$t < t_{n-2,a}$$

Αντίστοιχα για τις τρεις μορφές εναλλακτικής. Είναι πολύ σημαντικό να τονιστεί ότι ο συντελεστής συσχέτισης εκφράζει τη γραμμική σχέση μεταξύ των δυο μεταβλητών. Μια χαμηλή τιμή του δε δηλώνει απαραίτητα ότι η σχέση είναι ασθενής, οι μεταβλητές ενδέχεται να έχουν μια άλλη έντονη σχέση για παράδειγμα καμπυλόγραμμη, γι' αυτό και είναι έντονη η χρήση του διαγράμματος.

Ένα άλλο ιδιαίτερο σημείο που καλό είναι να τονιστεί, είναι η σχέση αιτίου – αποτελέσματος, μεταξύ των μεταβλητών, όταν ερμηνεύεται ο γραμμικός συντελεστής συσχέτισης. Μια υψηλή τιμή του δεν δηλώνει απαραίτητα σχέση αιτίου – αποτελέσματος μεταξύ των μεταβλητών.

Για παράδειγμα, υψηλή σχέση μεταξύ πωλήσεις αυτοκινήτων και κατανάλωσης πίτσας δεν σημαίνει ότι καθώς αυξάνει η κατανάλωση πίτσας, θα αυξάνουν και οι πωλήσεις των αυτοκινήτων. Η παρατηρούμενη σχέση είναι συμπωματική και οφείλεται καθαρά στον σύγχρονο τρόπο ζωής.

4.3 Απλή Γραμμική Παλινδρόμηση- Προϋποθέσεις

Η γραμμική παλινδρόμηση θεωρείται ένα από τα πιο σημαντικά εργαλεία που χρησιμοποιούνται στις στατιστικές αναλύσεις. Η απλή παλινδρόμηση χαρακτηρίζεται μόνο από μία α ανεξάρτητη μεταβλητή, σε αντίθεση με την πολλαπλή που περιέχει περισσότερες από μια ερμηνευτικές μεταβλητές.

Τα μοντέλα που χρησιμοποιούνται στην παλινδρόμηση χαρακτηρίζονται ως αιτιοκρατικά ή ντετερμινιστικά μοντέλα (deterministic model), επειδή η σχέση είναι αιτιώδης και οι τιμές των ανεξάρτητων μεταβλητών αναλύουν την εξαρτημένη. Αυτό συμβαίνει επειδή όσες φορές και να επαναληφθεί αυτή η διαδικασία η τιμή της εξαρτημένης μεταβλητής είναι πάντοτε η ίδια και εξαρτάται αποκλειστικά από τις τιμές των ανεξάρτητων μεταβλητών. Κάθε απόκλιση αποδίδεται σε σφάλματα μέτρησης και όχι σε διαφορές στην τιμή της μεταβλητής.

Εκτός από τα παραπάνω μοντέλα υπάρχει και το πιθανοθεωρητικό μοντέλο. Μια απλή σχέση παλινδρόμησης είναι η ακόλουθη:

$y = f(x)$ με την Y να αποτελεί την εξαρτημένη μεταβλητή, και την X την ανεξάρτητη, η f() αντιπροσωπεύει τη συναρτησιακή μορφή της σχέσης της παλινδρόμησης.

Σε περίπτωση που η συναρτησιακή σχέση μεταξύ των μεταβλητών είναι γραμμική τότε το μοντέλο γίνεται ως εξής: $y = \alpha + \beta x + \varepsilon$

Όπου:

y η τιμή της εξαρτημένης μεταβλητής

x η τιμή της ανεξάρτητης μεταβλητής

α ο σταθερός όρος, δηλαδή το σημείο τομής του άξονα y με την ευθεία παλινδρόμησης να έχει την τιμή της εξαρτημένης μεταβλητής όταν η ανεξάρτητη τιμή είναι ίση με μηδέν.

β η κλίση της ευθείας, που ονομάζεται και «γωνιακός συντελεστής» και προσδιορίζει τη μεταβολή της εξαρτημένης μεταβλητής σε μια μοναδική αλλαγή της ανεξάρτητης μεταβλητής.

ε σφάλμα που εμφανίζει τη διαφορά ανάμεσα στην τιμή της εξαρτημένης και της τιμής που δημιουργείται από το προβλεπόμενο μοντέλο.

Το μοντέλο της απλής γραμμικής παλινδρόμησης υποστηρίζει τα παρακάτω:

Οι τιμές της εξαρτημένης μεταβλητής χαρακτηρίζονται μεταξύ τους ανεξάρτητες.

Σε κάθε ορισμένη τιμή της ανεξάρτητης μεταβλητής αναλογούν αρκετές τιμές της εξαρτημένης και αποτελούν κανονική κατανομή.

Όταν το δείγμα μας είναι μεγέθους n, τότε υπάρχουν n κανονικές κατανομές της εξαρτημένης μεταβλητής με την ίδια μεταξύ τους διασπορά σ_{ε}^2.

Ο μέσος της κάθε $y_i(y)$ που έχει κανονική κατανομή περιγράφεται από τον παρακάτω τύπο : $E(y) = \alpha + \beta x$.

Όλοι οι μέσοι προσδιορίζονται σε μία ευθεία γραμμή που ονομάζεται γραμμή παλινδρόμησης του πληθυσμού και αποτελεί τον σύνδεσμο των μέσων της

εξαρτημένης μεταβλητής που αναλογούν στις τιμές της ανεξάρτητης. Τα σημεία α και β ονομάζονται συντελεστές παλινδρόμησης.

Σε περίπτωση μίας θετικής κλίσης ανάμεσα στις μεταβλητές δηλώνει και τη θετική σχέση μεταξύ τους, ενώ μία αρνητική κλίση προσδιορίζει μια αρνητική σχέση μεταξύ των μεταβλητών. Η γραμμικότητα δεν θεωρείται τόσο περιοριστική. Η απλή γραμμική παλινδρόμηση χαρακτηρίζεται ως γραμμική σχετικά με τις παραμέτρους και όχι και ως προς τις μεταβλητές, οπότε υπάρχει μία ευελιξία για την εκτίμηση του μοντέλου, και βέβαια προηγείται η συλλογή ενός τ.δ. (x_i, y_i) μεγέθους n. Με την ανάλυση της παλινδρόμησης υπολογίζονται οι εκτιμήσεις για τις άγνωστες πληθυσμιακές παραμέτρους. Επειδή όμως είναι αρκετά δύσκολο να καθοριστούν όλοι οι τυχαίοι παράγοντες προσδιορίζονται ορισμένες υποθέσεις για τον διαταρακτικό όρο.

4.3.1 Υποθέσεις για τον διαταρακτικό όρο

Για κάθε τιμή της ανεξάρτητης μεταβλητής ο διαταρακτικός όρος είναι μια κανονική τυχαία μεταβλητή με μέσο 0 και διασπορά σ^2. Δηλαδή, $\varepsilon_i \sim N(0, \sigma^2)$.

Η υπόθεση του μηδενικού μέσου εξασφαλίζει ότι η ευθεία θα περνάει από τους μέσους των κατανομών της εξαρτημένης μεταβλητής για δοθείσες τιμές της ανεξάρτητης μεταβλητής. Οι υποθέσεις της κανονικότητας και της σταθερής διασποράς μας εξασφαλίζουν κανονικές κατανομές της εξαρτημένης μεταβλητής με ίδιες διασπορές για διαφορετικές τιμές της ανεξάρτητης. Τα σφάλματα δεν σχετίζονται μεταξύ τους δηλ. ε_i είναι ανεξάρτητα μεταξύ τους. Δηλαδή,

$$\forall\ i, j \quad \mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$$

Αν δεν ισχύει αυτή η υπόθεση τότε εμφανίζεται το πρόβλημα της αυτοσυσχέτισης του διαταρακτικού όρου. Η αυτοσυσχέτιση είναι συνηθισμένη σε δεδομένα χρονοσειρών.

$$Var(\varepsilon_i) = \sigma^2 .$$

Δηλαδή, το τυχαίο σφάλμα έχει σταθερή διασπορά για όλες τις τιμές της ανεξάρτητης μεταβλητής. Αν η διασπορά δεν παρουσιάζει σταθερότητα τότε

εμφανίζεται το πρόβλημα της ετεροσκεδαστικότητας, με εμφανείς συνέπειες στις ιδιότητες των εκτιμητριών. Το πρόβλημα της ετεροσκεδαστικότητας εξετάζεται κατασκευάζοντας τη γραφική παράσταση των υπολοίπων έναντι της ανεξάρτητης μεταβλητής ή τις εκτιμηθείσες της εξαρτημένης. Αν το νέφος σημείων διασκορπίζεται με μη συστηματικό τρόπο τότε υπάρχει και διασπορά σταθερή και ικανοποιείται η προϋπόθεση αυτή. Αντιθέτως, αν το νέφος σημείων αυξάνεται ή μειώνεται τότε η διασπορά δεν είναι σταθερή και δεν ισχύει η υπόθεση. Οι τιμές της ανεξάρτητης μεταβλητής θεωρούνται ως σταθερές και για μια συγκεκριμένη τιμή της ανεξάρτητης αντιστοιχεί μια ολόκληρη κατανομή της εξαρτημένης. Έτσι, κάθε διαφοροποίηση της εξαρτημένης μεταβλητής οφείλεται στους παράγοντες που συμπεριλαμβάνονται στο τυχαίο σφάλμα. Το τυχαίο σφάλμα πρέπει να κατανέμεται κανονικά με μέση τιμή μηδέν, σταθερή διασπορά και οι τιμές να είναι ανεξάρτητες.

4.4 Μέθοδος των ελαχίστων τετραγώνων

Στόχος της ανάλυσης παλινδρόμησης είναι να εκτιμηθούν οι παράμετροι του μοντέλου παλινδρόμησης, κατά τέτοιο τρόπο ώστε η ευθεία που θα προκύψει να περιγράφει όσο το δυνατόν καλύτερο τρόπο τη σχέση μεταξύ των δυο μεταβλητών. Η γραμμή παλινδρόμησης πρέπει να περνάει κοντά από τα σημεία που αντιστοιχούν στα ζεύγη των παρατηρήσεων (x_i, y_i), ώστε να ελαχιστοποιούνται τα σφάλματα της πρόβλεψης. Οι τιμές του διαταρακτικού όρου αντιπροσωπεύουν την κατακόρυφη απόσταση μεταξύ της γραμμής της παλινδρόμησης και της εκάστοτε παρατήρησης. Για την καταλληλότητα του δείγματος ως προς τα συγκεκριμένα δεδομένα, απαιτούμε τα συγκεκριμένα υπόλοιπα να έχουν το μικρότερο δυνατό μέγεθος.

Το κατάλληλο μέτρο είναι το άθροισμα των τετραγώνων των υπολοίπων (SSE, sum of Square of error), δηλαδή:

$$SSE = \sum e_i^2 = \sum (y_i - \hat{a} - \hat{\beta}x)^2$$

Η εκτίμηση με τη μέθοδο των ελαχίστων τετραγώνων επιλέγει ως κατάλληλες τιμές των παραμέτρων αυτές που ελαχιστοποιούν το SSE για δεδομένο δείγμα. Οι εκτιμητές που προκύπτουν καλούνται εκτιμητές ελαχίστων τετραγώνων και συμβολίζονται με $\hat{\alpha}$ και $\hat{\beta}$.

Με βάση αυτούς τους εκτιμητές δύναται να υπολογιστούν οι προβλεπόμενες τιμές για την εξαρτημένη μεταβλητή, δηλαδή:

$$\hat{y} = \hat{a} + \hat{\beta}x$$

Όπου $\hat{y}$ η εκτιμηθείσα τιμή της y.

Επειδή τα υπόλοιπα έχουν και θετικό και αρνητικό πρόσημο, προσπαθούμε να ελαχιστοποιήσουμε τα τετράγωνά τους και μάλιστα το άθροισμά τους.

Συμβολίζουμε με Q το άθροισμα των τετραγώνων των αποκλίσεων και αναζητούμε τις τιμές των παραμέτρων που ελαχιστοποιούν αυτό το άθροισμα. Η ελαχιστοποίηση αυτή προκύπτει παραγωγίζοντας την Q ως προς α και β και εξισώνοντας με το μηδέν και εν συνεχεία λύνουμε ως προς τις άγνωστες παραμέτρους. Δηλαδή,

$$\frac{\partial Q}{\partial a} = -2\sum (y_i - \hat{a} - \hat{\beta}x) = 0$$

$$\frac{\partial Q}{\partial \beta} = -2\sum x_i (y_i - \hat{a} - \hat{\beta}x) = 0$$

Με βάση τις παραπάνω εξισώσεις, που καλούνται κανονικές εξισώσεις, προκύπτουν οι εκτιμήσεις των παραμέτρων α και β.

$$\hat{\beta} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - x)^2} = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2}$$

$$\hat{a} - \overline{y} - \hat{\beta}x$$

Ο γωνιακός συντελεστής έχει το ίδιο πάντοτε πρόσημο με τον γραμμικό συντελεστή συσχέτισης. Οι εκτιμητές εξαρτώνται από το δείγμα.. Άρα οι εκτιμητές ελαχίστων

τετραγώνων είναι τυχαίες μεταβλητές, εφόσον υπολογίζονται από ένα τ.δ. Έτσι, λαμβάνοντας διαφορετικά τ.δ. θα λαμβάνουμε και διαφορετικές εκτιμήσεις των παραμέτρων. Άρα, υπάρχει μια κατανομή πιθανότητας για τις παραμέτρους που εξαρτάται από τις υποθέσεις του διαταρακτικού όρου.

4.5 Συντελεστής Προσδιορισμού. Εκτίμηση της καλής προσαρμογής της γραμμικής παλινδρόμησης

Η καλή προσαρμογή του μοντέλου εκφράζεται ποσοτικά με τον συντελεστή προσδιορισμού R^2, που μετρά την προσαρμοστικότητα του μοντέλου. Αν το μοντέλο έχει τέλεια προσαρμοστικότητα, τα υπόλοιπα είναι όλα μηδέν, ο συντελεστής είναι ίσος με τη μονάδα. Αν το μοντέλο δεν εξηγεί καμία από τις διακυμάνσεις στα δεδομένα, δηλαδή δεν υπάρχει καμία σχέση μεταξύ εξαρτημένης και ανεξάρτητης μεταβλητής, τότε ο συντελεστής προσδιορισμού ισούται με μηδέν.

Δηλαδή, $0 \le R^2 \le 1$.

Στην απλή γραμμική παλινδρόμηση ισχύει $R^2 = \rho^2 \Leftrightarrow \rho = \pm\sqrt{R^2}$

Ο συντελεστής προσδιορισμού χρησιμοποιείται ως μέτρο καλής προσαρμογής του μοντέλου και υπολογίζει την αναλογία της ολικής διακύμανσης στα δεδομένα που εξηγείται από την παλινδρόμηση.

Αποδεικνύεται ότι:

$$\sum e_i = 0$$

$$\sum e_i \hat{y}_i = 0$$

Η πρώτη σχέση δηλώνει ότι το άθροισμα των υπολοίπων πάνω και κάτω από τη γραμμή είναι πάντα μηδέν. Η ευθεία ελαχίστων τετραγώνων περνάει από το μέσο των δεδομένων. Η ελαχιστοποίηση του αθροίσματος του τετραγώνου των υπολοίπων μπορεί επίσης να ερμηνευτεί από τη σχέση αυτή. Γενικά, το άθροισμα του τετραγώνου των υπολοίπων δεν είναι μηδέν αν και το άθροισμα των υπολοίπων είναι.

Η δεύτερη σχέση δηλώνει ότι οι εκτιμηθείσες τιμές και τα υπόλοιπα είναι ασυσχέτιστα. Δηλαδή, η γραμμική σχέση μεταξύ των δυο μεταβλητών έχει ληφθεί υπόψη από την παλινδρόμηση. Καθώς, ισχύουν οι δυο παραπάνω σχέσεις, έχουμε:

$$\sum (y_i - \overline{y})^2 = \sum (\hat{y}_i - \overline{y})^2 + \sum e_i^2$$

$$SST = SSR + SSE$$

Όπου

$SST = \sum (y_i - \overline{y})^2$ είναι η ολική μεταβλητότητα στα δεδομένα ή στις μεταβολές της εξαρτημένης μεταβλητής. $SSR = \sum (\hat{y}_i - \overline{y})^2$ είναι η διακύμανση που εξηγείται από την παλινδρόμηση. $SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$ είναι η μεταβολή της εξαρτημένης μεταβολής που εκφράζεται μέσω των υπολοίπων, δηλαδή η ανεξήγητη διακύμανση των δεδομένων. Διαιρώντας τη τελευταία σχέση με *SST* προκύπτει:

$$\frac{SST}{SST} = \frac{SSR}{SST} + \frac{SSE}{SST} \Leftrightarrow 1 = \frac{SSR}{SST} + \frac{SSE}{SST} \Leftrightarrow \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

και

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \overline{y})^2} = \frac{\sum (\hat{y}_i - \overline{y})^2}{\sum (y_i - \overline{y})^2}$$

Αν *SSR* είναι μικρό τότε η σχέση μεταξύ των δυο μεταβλητών είναι ασθενής. Αν *SSR* είναι μεγάλο τότε η σχέση μεταξύ των δυο μεταβλητών είναι ισχυρή. Μια μεγάλη τιμή του συντελεστή προσδιορισμού, δεν αντιπροσωπεύει απαραίτητα ένα ικανοποιητικό μοντέλο, γιατί το συγκεκριμένο στατιστικό μέτρο απλώς αποκαλύπτει μια ισχυρή σχέση μεταξύ των μεταβλητών, και η εξίσωση παλινδρόμησης είναι πιθανόν να είναι ακατάλληλη.

Οπότε ο συντελεστής προσδιορισμού είναι ανεπαρκής ως κριτήριο για την επιλογή του κατάλληλου μοντέλου. Ο συντελεστής προσδιορισμού ακόμα αυξάνεται αν στο

μοντέλο προστεθούν και άλλες ανεξάρτητες μεταβλητές ακόμα και στην περίπτωση που αυτές είναι ασυσχέτιστες με την Υ. Γι' αυτό τον λόγο χρησιμοποιείται ο διορθωμένος συντελεστής προσδιορισμού R^2adj.

Ένα άλλο μέτρο προβλεπτικότητας του μοντέλου είναι το τυπικό σφάλμα εκτίμησης, που δείχνει το πόσο οι πραγματικές τιμές της εξαρτημένης μεταβλητής αποκλίνουν από τις εκτιμηθείσες. Δηλαδή, εκφράζει την τυπική απόκλιση της εξαρτημένης σε σχέση με την ευθεία παλινδρόμησης.

$$s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum e_i^2}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

Παρατήρηση: Όταν όλα τα σημεία, εκτός από ένα, ακολουθούν μια γραμμή, τότε το σημείο αυτό χαρακτηρίζεται ως ένα μεγάλο τυποποιημένο υπόλοιπο (standard izedresidual). Και σε αυτήν την περίπτωση πρέπει να ελεγχθεί αν υπάρχει κάποιο πρόβλημα με τα δεδομένα και εν συνεχεία να ερμηνευτούν τα αποτελέσματα με και χωρίς αυτή την τιμή.

4.6 Έλεγχος Σημαντικότητας

Έλεγχος Σημαντικότητας για το συντελεστή Προσδιορισμού

Ο έλεγχος αυτός ελέγχει αυτό που μετρά ο συντελεστής προσδιορισμού, δηλαδή ελέγχει εάν το ποσοστό των μεταβολών της Υ οφείλεται στις επιδράσεις της Χ και εξηγείται από την εξίσωση παλινδρόμησης, είναι διάφορο του μηδενός.

Έτσι έχουμε την ακόλουθη υπόθεση:

H_o: η εξίσωση παλινδρόμησης δεν εξηγεί καθόλου τις μεταβολές της εξαρτημένης μεταβλητής δηλαδή το ποσοστό τη Υ που ερμηνεύεται από την Χ είναι μηδέν

Κατά

H_1: η εξίσωση παλινδρόμησης εξηγεί ένα μέρος των μεταβολών της εξαρτημένης μεταβλητής δηλαδή το ποσοστό τη Υ που ερμηνεύεται από την Χ είναι μεγαλύτερο του μηδενός. Άρα, συγκρίνονται οι δυο συνιστώσες της *SST*, η εξηγημένη *SSR* και ανεξήγητη *SSE.* Εάν η πρώτη είναι σημαντικά μεγαλύτερη της δεύτερης τότε σημαίνει ότι η επίδραση της εξίσωσης παλινδρόμησης είναι σημαντική. Ενώ, αν η

δεύτερη είναι σημαντικά μεγαλύτερη της πρώτης τότε η επίδραση της εξίσωσης παλινδρόμησης δεν είναι σημαντική και το ποσοστό της εξαρτημένης που ερμηνεύεται από την εξίσωση είναι αμελητέο. Τα *SSR* και *SSE* είναι αθροίσματα τετραγώνων των αποκλίσεων, που βασίζονται σε διαφορετικό αριθμό β.ε. και γι' αυτό θα πρέπει να διαιρεθούν με τους αντίστοιχους β.ε. για να συγκριθούν. Οι λόγοι που προκύπτουν καλούνται μέσα τετράγωνα (MSE) και ο έλεγχος στηρίζεται στην F κατανομή. Η όλη διαδικασία περιγράφεται από τον ακόλουθο πίνακα που καλείται έλεγχος ανάλυσης διακύμανσης.

Πηγή Μεταβλητότητας	Αθροίσματα Τετραγώνων	Βαθμοί ελευθερίας	Μέσα Τετράγωνα	Λόγος $F_{1,n-2}$
Παλινδρόμηση	SSR	1	SSR/1	(SSR/1)/(SSE/n-2)=SSR/MSE
Υπόλοιπα	SSE	n-2	SSE/ n-2	
Σύνολο	SST	n-1		

Ο λόγος SSE/n-2 καλείται μέσο τετραγωνικό σφάλμα MSE.

Αν η τιμή της στατιστικής συνάρτησης είναι μεγαλύτερη από το ε.σ. τότε απορρίπτεται η μηδενική υπόθεση.

Έλεγχος Σημαντικότητας για τις παραμέτρους α και β

Σε κάθε στατιστική ανάλυση είναι πολύ σημαντικός ο έλεγχος της στατιστικής σημαντικότητας των παραμέτρων του μοντέλου.

Οι πιο συνηθισμένες μορφές είναι:

1.H_0:a=0 vs H_1:a¹0

2.H_0:a=0 vs H_1:a>0

3.H_0:a=0 vs H_1:a<0

Και

1.H_0:β=0 vs H_1:β¹0

2.H_0:β=0 vs H_1:β>0

3.H_0:β=0 vs H_1:β<0

Με τους πρώτους ελέγχουμε κατά πόσο η ευθεία διέρχεται από την αρχή των αξόνων.

Με τους δεύτερους, ελέγχουμε με βάση τα δεδομένα, κατά πόσο η ευθεία έχει μηδενική κλίση και συνεπώς αν η ανεξάρτητη μεταβλητή ερμηνεύει την εξαρτημένη, δηλαδή αν υπάρχει ή όχι συσχέτιση μεταξύ των μεταβλητών. Απορρίπτοντας τη μηδενική υπόθεση σημαίνει ότι υπάρχει στατιστικά σημαντική σχέση μεταξύ των δυο μεταβλητών.

Επίσης, πολλές φορές θέλουμε να ελεγχθεί ένα τμήμα της εκτιμηθείσας εξίσωσης. Για παράδειγμα έστω ότι θέλουμε να διαπιστώσουμε αν η κλίση της ευθείας ενός συνόλου δεδομένων είναι διαφορετική σε σχέση με αυτήν ενός άλλου συνόλου δεδομένων. Ή ακόμα να ελεγχθεί αν κάποιο τμήμα της εξίσωσης έχει μεταβληθεί στη διάρκεια του χρόνου. Οι έλεγχοι γίνονται με χρήση της t στατιστικής συνάρτησης. Έτσι, έχουμε:

$$t = \frac{\hat{a} - a_o}{s_{\hat{a}}} \sim t_{n-2}$$

$$t = \frac{\hat{\beta} - \beta_o}{s_{\hat{\beta}}} \sim t_{n-2}$$

Όπου

$$s^2_{\hat{a}} = s^2_e \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum x^2 - n\bar{x}^2}\right)$$

$$s^2_{\hat{\beta}} = \frac{s^2_e}{\sum (x-\bar{x})^2} = \frac{s^2_e}{\sum x^2 - n\bar{x}^2}$$

Τα τυπικά σφάλματα είναι οι τετραγωνικές ρίζες των παραπάνω σχέσεων. Οι κατανομές θεωρούνται κανονικές. Έτσι, απορρίπτουμε τη μηδενική υπόθεση αν:

$|t| > t_{n-2,a/2}$

$t > t_{n-2,a}$

$t < -t_{n-2,a}$

Αντίστοιχα.

Επίσης, έχουμε και τα αντίστοιχα ΔΕ

$\hat{a} \pm t_{n-2,a/2} s_{\hat{a}}$

$\hat{\beta} \pm t_{n-2,a/2} s_{\hat{\beta}}$

Επίσης ο έλεγχος μπορεί να πραγματοποιηθεί με χρήση του p-value δηλαδή p<α απορρίπτουμε τη μηδενική υπόθεση και η υπό εξέταση μεταβλητή είναι στατιστικά σημαντική.

4.7 Προβλέψεις

Αφού έχει εκτιμηθεί η εξίσωση της ευθείας μπορούμε να βρούμε το $\hat{y}_O$ για τη συγκεκριμένη τιμή x_O και αποτελεί την πρόβλεψη που προκύπτει από την εξίσωση παλινδρόμησης, δηλαδή $\hat{y}_O = \hat{a} + \hat{\beta} x_O$. Οι παράμετροι $\hat{\alpha}$ και $\hat{\beta}$ είναι εκτιμήσεις των παραμέτρων και υπόκεινται σε σφάλματα της δειγματοληψίας, άρα και το $\hat{y}_O$ υπόκεινται σε σφάλματα της δειγματοληψίας, άρα το τυπικό σφάλμα του θα δίνεται από τον τύπο:

$$s_{\hat{y}_o} = s_e \sqrt{\frac{1}{n} + \frac{(x_o - \overline{x})^2}{\sum (x - \overline{x})^2}} = \sqrt{\frac{SSE}{n-2}} \sqrt{\frac{1}{n} + \frac{(x_o - \overline{x})^2}{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

Έτσι μπορεί να εκτιμηθεί το ΔΕ της $\hat{y}_O$ που αντιστοιχεί σε πιθανότητα 1-α:

$$\hat{y}_O \pm t_{n-2,a/2}\sqrt{\frac{SSE}{n-2}}\sqrt{\frac{1}{n}+\frac{(x_O-\overline{x})^2}{\sum x^2-\frac{(\sum x)^2}{n}}}$$

Παρατηρείται ότι το μικρότερο σφάλμα στην πρόβλεψη αντιστοιχεί στην τιμή $x=\overline{x}$. Ενώ το δειγματοληπτικό σφάλμα της πρόβλεψης $\hat{y}$ αυξάνεται όσο η τιμή της ανεξάρτητης μεταβλητής απομακρύνεται από το μέσο του δείγματος.

Το ΔΕ της πρόβλεψης αναφέρεται στη μέση τιμή της Υ που αντιστοιχεί στην τιμή x_O. Η πραγματική τιμή της εξαρτημένης μεταβλητής εξαρτάται και από την τιμή της τυχαίας συνιστώσας e_o. Το δειγματοληπτικό σφάλμα της y_O είναι μεγαλύτερο από το σφάλμα της $\hat{y}_O$ καθώς επηρεάζεται και από τη διασπορά της τυχαίας συνιστώσας. Το τυπικό σφάλμα της πρόβλεψης της y_O δίνεται από τον τύπο:

$$s_{y_o}=s_e\sqrt{1+\frac{1}{n}+\frac{(x_o-\overline{x})^2}{\sum(x-\overline{x})^2}}=\sqrt{\frac{SSE}{n-2}}\sqrt{1+\frac{1}{n}+\frac{(x_o-\overline{x})^2}{\sum x^2-\frac{(\sum x)^2}{n}}}$$

Και το αντίστοιχο ΔΕ:

$$\hat{y}_O \pm t_{n-2,a/2}\sqrt{\frac{SSE}{n-2}}\sqrt{1+\frac{1}{n}+\frac{(x_O-\overline{x})^2}{\sum x^2-\frac{(\sum x)^2}{n}}}$$

Για τιμές της ανεξάρτητης μεταβλητής κοντά στο δειγματικό μέσο το σφάλμα έχει τη μικρότερη τιμή. Όσο απομακρυνόμαστε από τη τιμή του μέσου τόσο το σφάλμα της πρόβλεψης αυξάνεται και μάλιστα εκθετικά. Γι' αυτό και οι προβλέψεις της ανεξάρτητης μεταβλητής που βρίσκονται στα όρια του εύρους τιμών του δείγματος πρέπει να εξετάζονται με επιφύλαξη. Στην πράξη, οι προβλέψεις αφορούν την αναμενόμενη τιμή της εξαρτημένης μεταβλητής και όχι την πραγματική της.

4.8 Ανάλυση Υπολοίπων

Γραφική εξέταση της κανονικής κατανομής → χρησιμοποιείται και για τον εντοπισμό απομονωμένων σημείων (άτυπων, απόμακρων) (outliers), που ενδέχεται να εμφανίζονται και να θέτουν ερωτήματα για την καταλληλότητα του μοντέλου.

Αν η εικόνα δεν είναι ικανοποιητική, τότε ίσως να χρειάζεται να γίνει κάποιος μετασχηματισμός, για παράδειγμα της εξαρτημένης ή να πρέπει να προστεθούν στο μοντέλο και άλλες ανεξάρτητες μεταβλητές.

Ακόμα μπορεί να είναι απαραίτητη και η αφαίρεση κάποιων παρατηρήσεων που μπορεί να προήλθαν από άλλο πληθυσμό ή λάθος καταγραφή.

Γραφική παράσταση των σφαλμάτων επί των εκτιμημένων (προσαρμοσμένων) τιμών → χρησιμοποιείται κυρίως για έλεγχο ομοσκεδαστικότητας. Καθώς επίσης και για σημεία που ενδέχεται να ξεφεύγουν. Αν η εικόνα δεν είναι ικανοποιητική, ίσως να είναι πάλι απαραίτητος κάποιος μετασχηματισμός.

Γραφικές παραστάσεις για τον έλεγχο της συσχέτισης → σε περιπτώσεις που τα δεδομένα έχουν μια ορισμένη χρονική σειρά. Εδώ έχουμε δυο ειδών γραφικές παραστάσεις:

> Γραφική παράσταση των υπολοίπων ως προς τη σειρά, το χρόνο, ή τη θέση των παρατηρήσεων και
> Γραφική παράσταση των e_i κατά e_{i-1} για την εξέταση της συσχέτισης μεταξύ των υπολοίπων.

Τα σημεία που κατανέμονται με τυχαιότητα σε αυτές τις γραφικές παραστάσεις δηλώνουν ανεξαρτησία μεταξύ των σφαλμάτων, οπότε δεν παραβιάζεται η σχετική υπόθεση του μοντέλου. Διαφορετικά θα πρέπει η αυτοσυσχέτιση να ληφθεί υπόψη στο μοντέλο.

Γραφική παράσταση των e_i έναντι των x_i(μιας ανεξάρτητης μεταβλητής). Τυχαία διάσπαρτα σημεία δηλώνουν ανεξαρτησία μεταξύ των σφαλμάτων και της ανεξάρτητης μεταβλητής. Σε αντίθετη περίπτωση ίσως η ανεξάρτητη χρειαστεί κάποιο μετασχηματισμό.

Παρατήρηση: Αν κάποια ή κάποιες γραφικές παραστάσεις δεν ικανοποιούν τις απαραίτητες προϋποθέσεις, τότε το μοντέλο δεν περιγράφει ικανοποιητικά τα δεδομένα.

Κεφάλαιο 5 Πολλαπλή Παλινδρόμηση

Εκτός από την απλή γραμμική παλινδρόμηση και γραμμική συσχέτιση για την ανάλυση και την ερμηνεία των μεταβλητών, υπάρχει και η ανάγκη για πιο σύνθετες αναλύσεις μεταξύ περισσότερων των δυο μεταβλητών. Αυτή εκτίμηση που στηρίζεται σε περισσότερες μεταβλητές της ονομάζεται πολλαπλή παλινδρόμηση.

Αντικειμενικός στόχος της αποτελεί η περιγραφή της σχέσης μεταξύ της μεταξύ της εξαρτημένης μεταβλητής Y και των k ανεξάρτητων μεταβλητών X_1, X_2,…, X_k. Έτσι το μοντέλο που θα επιλεγεί καθορίζει ότι η εξαρτημένη μεταβλητή χαρακτηρίζεται ως γραμμική συνάρτηση των k ανεξάρτητων μεταβλητών και συναντάται ως εξής:

$$Yi = \beta_o + \beta_1 x_1 i + ... + \beta_k x_{ki} + \varepsilon i$$

Όπου Y η τιμή της εξαρτημένης μεταβλητής και παραπάνω από μία εξαρτημένες μεταβλητές. x_{1i} x_2,…..x_{ki} Δηλαδή, σε ένα υπόδειγμα πολλαπλής παλινδρόμησης η εξαρτημένη μεταβλητή Y από μια σειρά μεταβλητών. Η επίδραση κάθε μεταβλητής μπορεί να καθοριστεί από τους αντίστοιχους συντελεστές παλινδρόμησης β_1, β_2,...,β_κ των k ανεξάρτητων μεταβλητών (regrcssors). Δηλαδή υπάρχουν οι σταθερές β_0, β_1,, β_2, β_κ και (σ) ώστε για κάθε σύνολο, ο αντίστοιχος πληθυσμός των τιμών της Y κατανέμεται κανονικά με μέσο $(\beta_o + \beta_1 x_1 i + ... + \beta_k x_{ki})$ και τυπική απόκλιση (σ).

Μια απλή γραμμική παλινδρόμηση είναι απλώς μία ειδική περίπτωση του υποδείγματος πολλαπλής με μια ερμηνευτική μεταβλητή. Η εκτίμηση του υποδείγματος πολλαπλής παλινδρόμησης γίνεται με τη μέθοδο των ελαχίστων τετραγώνων (OLS) και σύμφωνα με τις ακόλουθες υποθέσεις:

1. Για κάθε παρατήρηση το τυχαίο σφάλμα (ε_i) κατανέμεται κανονικά με μηδενικό μέσο και κοινή διακύμανση σ^2:

$E(\varepsilon i \mid X_{1i}, X_{2i}; ,..., X_{ki}) = 0$

$Var(\varepsilon i \mid X_{1i}, X_{2i}; ,..., X_{ki}) = \sigma^2$

2. Το σφάλμα αυτό δεν συσχετίζεται με τα σφάλματα ταν άλλων παρατηρήσεων:

$Cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i, \varepsilon_j) = 0$

3. Κάθε μία από τις ερμηνευτικές μεταβλητές είναι ανεξάρτητη από το τυχαίο σφάλμα: $\text{Cov}(\varepsilon_i, X_{ij}) = 0$

4. Οι ερμηνευτικές μεταβλητές θεωρούνται σταθερές ποσότητες. Αυτή η υπόθεση συνεπάγεται ότι η τυχαιότητα της εξαρτημένης μεταβλητής Y οφείλεται αποκλειστικά στο τυχαίο σφάλμα (ε).

5. Υπάρχει σωστός προσδιορισμός τον υποδείγματος σχετικά με τη συναρτησιακή μορφή και τις συμπεριλαμβανόμενες μεταβλητές.

6. Οι ανεξάρτητες μεταβλητές δεν συσχετίζονται γραμμικά μεταξύ τους.

Η τελευταία υπόθεση ουσιαστικά απαιτεί ότι οι k ερμηνευτικές μεταβλητές δεν έχουν καμιά γραμμική σχέση μεταξύ τους. Η παραβίαση της υπόθεσης αυτής δημιουργεί πρόβλημα στο διαχωρισμό και αναγνώριση της ατομικής επίδρασης κάθε μεταβλητής πάνω στην εξαρτημένη. Το πρόβλημα αυτό είναι γνωστό ως πρόβλημα πολυσυγγραμμικότητας. Με απλά λόγια, καθώς οι ερμηνευτικές μεταβλητές συσχετίζονται μεταξύ τούς, η ατομική επίδραση κάθε μιας από αυτές τις μεταβλητές πάνω στην εξαρτημένη δεν μπορεί να απομονωθεί.

5.1 Ερμηνεία των συντελεστών πολλαπλής γραμμικής παλινδρόμησης

Η σημασία των παραμέτρων σε μια εξίσωση πολλαπλής γραμμικής παλινδρόμησης εξαρτάται από τον καθορισμό της σχέσης πού υπάρχει στην παλινδρόμηση. Ας πάρουμε την απλή περίπτωση ενός υποδείγματος παλινδρόμησης, το οποίο είναι γραμμικό στα επίπεδα των μεταβλητών. Υποθέτουμε για παράδειγμα, ότι καθορίζουμε μια πολλαπλή παλινδρόμηση ενός υποδείγματος σχετικού με τη ζητούμενη ποσότητα ενός προϊόντος (Q), την τιμή του προϊόντος (P) και το διαθέσιμο εισόδημα (Y).Βασιζόμενοι σε αυτόν τον ορισμό, μπορούμε να υπολογίσουμε την ελαστικότητα τιμής ζήτησης (ETZ) για το προϊόν αυτό. $\beta_1 = \dfrac{\overline{P}}{Q}$

Παρόμοια, για την μεταβλητή του εισοδήματος έχουμε:

$$\beta 2 = \frac{\partial Q}{\partial \mathrm{Y}}$$

Βασιζόμενοι στην κλίση β_2 μπορούμε να βρούμε την εισοδηματική ελαστικότητα ζήτησης από τον αντίστοιχο τύπο της Μικροοικονομικής (και πάλι σε μέσες τιμές). Η ελαστικότητα σε αυτή την περίπτωση εξαρτάται από τα επίπεδα τιμής (P) και κατανάλωσης (Q) για τα οποία η ελαστικότητα εκτιμάται. Συνήθως παίρνουμε τις μέσες τιμές των μεταβλητών και βρίσκουμε την τοξοειδή ελαστικότητα τιμής ζήτησης.

Για τον υπολογισμό των κλίσεων και των ελαστικοτήτων χρήσιμος είναι ο ακόλουθος πίνακας:

Πίνακας 3: Υπολογισμός των κλίσεων και των ελαστικοτήτων

Μοντέλο	Κλίση $\frac{dY}{dX}$	Ελαστικότητες
$\mathrm{Y} = \beta_o + \beta_1 \mathrm{X}$ Γραμμικό	β_1	$\beta_1 \frac{\mathrm{X}}{\mathrm{Y}}$
$\ln \mathrm{Y} = \beta_o + \beta_1 \mathrm{X}$ Λογαριθμικό Γραμμικό	$\beta_{1(\mathrm{Y})}$	$\beta_{1(\mathrm{X})}$
$\ln \mathrm{Y} = \beta_o + \beta_1 \ln \mathrm{X}$ λογαριθμικό-λογαριθμικό	$\beta_1(\frac{\mathrm{Y}}{\mathrm{X}})$	β_1
$\mathrm{Y} = \beta_o + \beta_1 \ln \mathrm{X}$ Γραμμικό-Λογαριθμικό	$\beta_1(\frac{1}{\mathrm{X}})$	$\beta_1(\frac{1}{\mathrm{Y}})$
$\mathrm{Y} = \beta_o + \beta_1(1/\mathrm{X})$ Αντίστροφο μεταβλητής X	$-\beta_1(\frac{1}{\mathrm{X}^2})$	$-\beta_1(\frac{1}{\mathrm{XY}})$

5.2 Εκτιμητές Ελαχίστων Τετραγώνων Πολλαπλής Παλινδρόμησης

Βασιζόμενοι στις υποθέσεις που αναφέρθηκαν παραπάνω μπορούμε, όπως και στην, περίπτωση της απλής παλινδρόμησης, να εξαγάγουμε τους εκτιμητές ελαχίστων τετραγώνων με την ελαχιστοποίηση του αθροίσματος τετραγώνων των υπολοίπων(SSR). Οι τύποι υπολογισμού των εκτιμητών ελαχίστων τετραγώνων μπορούν να υπολογιστούν με παρόμοιο τρόπο με αυτόν τον απλού γραμμικού υποδείγματος.

Αν πάρουμε την πιο απλή μορφή πολλαπλής γραμμικής παλινδρόμησης

$Y_i = \beta_o + \beta_1 X_{1i} + \beta 2\, X_{2i} + \varepsilon_i$

Από τις συνθήκες πρώτης τάξης μπορούμε να πάρουμε τις τρεις κανονικές περιπτώσεις και με τη λύση τους παίρνουμε τα εξής:

$$b_1 = \frac{(\sum x_1 y)(\sum x_2^2) - (\sum x_2 y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$b_2 = \frac{(\sum x_2 y)(\sum x_1^2) - (\sum x_1 y)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$b_0 = \overline{Y} - b_1 \overline{X}_1 - b_2 \overline{X}_2$$

Παρατηρούμε ότι μια πολλαπλή γραμμική παλινδρόμηση ακολουθεί τις ίδιες διαδικασίες με τη διμεταβλητή, αλλά με ολοφάνερα πιο πολύπλοκες παραγώγους. Υπάρχουν κάποιες βασικές ιδιότητες των εκτιμητών ελαχίστων τετραγώνων σε μια πολλαπλή γραμμική παλινδρόμηση, πολλές φορές ανάλογες με αυτές των εκτιμητών της απλής γραμμικής παλινδρόμησης. Ειδικότερα, σε μια πολλαπλή γραμμική παλινδρόμηση με k παραμέτρους ισχύουν τα ακόλουθα:

1. Το άθροισμα των καταλοίπων των ελαχίστων τετραγώνων ισούται με μηδέν.

2. Η εκτιμημένη γραμμή πολλαπλής παλινδρόμησης διέρχεται από τους δειγματικούς μέσους.

3. Η συσχέτιση ανάμεσα στα κατάλοιπα ελαχίστων τετραγώνων και κάθε ερμηνευτικής μεταβλητής είναι μηδέν.

4. Κάτω από τις υποθέσεις του υποδείγματος πολλαπλής παλινδρόμησης, οι εκτιμητές ελαχίστων τετραγώνων $b_1,b_2,...b_k$, είναι οι καλύτεροι γραμμικοί αμερόληπτοι εκτιμητές (Best Linear Unbiased Estimators) σύμφωνα με το θεώρημα Gauss-Markov.. Μπορούμε να συμπεράνουμε ότι:

$E(b_j)=\beta_j$ για $j=0,1,\ldots,k$

Επιπλέον, αν υποθέσουμε ότι ο στοχαστικός διαταρακτικός όρος ε_j κατανέμεται κανονικά με μέσο μηδέν και διακύμανση σ^2, τότε οι εκτιμητές ελαχίστων τετραγώνων b_j για $j=0,1,2,...,k$ θα κατανέμονται και οι ίδιοι κανονικά.

5.3 Έλεγχος στατιστικής σημαντικότητας ενός υποδείγματος πολλαπλής παλινδρόμησης

Σε υποδείγματα πολλαπλής παλινδρόμησης ελέγχουμε τη συνολική στατιστική σημασία από τις ερμηνευτικές μεταβλητές. Αν θέλουμε την προβλεπτικότητα του υποδείγματος, μπορούμε να ελέγξουμε τους συντελεστές($\beta_1,\beta_2,..\ \beta_k$) όλων των ερμηνευτικών μεταβλητών $X_ι,X_2,...,X_k$ σχηματίζοντας υποθέσεις:

H_o:$\beta_1=\beta_2= ...= \beta_k=0$

H_i: τουλάχιστον ένας εκ των συντελεστών β είναι διάφορος του 0

Αν απορρίψουμε την υπόθεση H_0 σημαίνει ότι η παλινδρόμηση είναι χρήσιμη στην ερμηνεία της εξαρτημένης μεταβλητής και κατ΄επέκταση την πραγματικότητα των προβλέψεων. Με άλλα λόγια, η απόρριψη της H_0 σημαίνει ότι τουλάχιστον μία από τις ερμηνευτικές μεταβλητές είναι στατιστικά σημαντική. Στην πολλαπλή παλινδρόμηση χρησιμοποιούμε την κατανομή F.

5.4 Μέτρα καλής εφαρμογής στην πολλαπλή παλινδρόμηση

Όταν σε ένα υπόδειγμα προσθέσουμε περισσότερες ερμηνευτικές μεταβλητές αναμένεται ο απλός συντελεστής προσδιορισμού να αυξηθεί ακόμη και στην περίπτωση που οι προστιθέμενες μεταβλητές δεν έχουν σχέση με τη θεωρητική ερμηνεία της υπό εξέτασης μεταβλητής. Σύμφωνα με τον ορισμό του συντελεστή προσδιορισμού έχουμε:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum e_i^2}{\sum (y - \overline{y})^2} = \frac{\sum (\hat{y} - \overline{y})^2}{\sum (y - \overline{y})^2} = \frac{\hat{\beta}_1 \sum yx_1 + \hat{\beta}_2 \sum yx_2}{\sum y^2}$$

Αν η επιδίωξη μας ήταν ένα υψηλό R^2, στον τύπο του συντελεστή προσδιορισμού τότε θα αρκούσε η προσθήκη μεταβλητών για την επίτευξη του στόχου αυτού. Όμως αυτό που επιδιώκουμε είναι να δούμε αν η προβλεπτικότητα του υποδείγματος βελτιώνεται και να είμαστε σίγουροι ότι η προσθήκη μιας επιπλέον μεταβλητής βελτιώνει το υπόδειγμά μας.

Στη συγκεκριμένη περίπτωση απαιτείται η χρησιμοποίηση του προσαρμοσμένου ή διορθωμένου συντελεστή προσδιορισμού (R^2 ή R^2 — adjusted), ο οποίος δίνεται από τον τύπο:

$$R^2_{adj} = 1 - \frac{\sum e_i^2}{\sum (y - \overline{y})^2} \frac{n-1}{n-k}$$

όπου k ο αριθμός των ανεξάρτητων μεταβλητών συμπεριλαμβανόμενου του σταθερού όρου. Ο τύπος αυτός είναι όμοιος με τον τύπο του απλού συντελεστή προσδιορισμού, αλλά έχει την επιπρόσθετη παρένθεση η οποία λαμβάνει υπόψη τον αριθμό των ερμηνευτικών μεταβλητών στην εξειδίκευση τον υποδείγματος. Όταν το υπόδειγμα περιλαμβάνει μόνο το σταθερό όρο, το προσαρμοσμένο μέτρο καταλήγει να είναι το ίδιο με τον απλό συντελεστή προσδιορισμού και ίσο με μηδέν. Όσο ο αριθμός των ερμηνευτικών μεταβλητών αυξάνεται, δηλαδή όσο k>1, η προσαρμογή δίνει στο διορθωμένο συντελεστή μια μικρότερη τιμή από αυτή του απλού συντελεστή προσδιορισμού. Η σχέση μεταξύ του προσαρμοσμένου και του απλού συντελεστή προσδιορισμού δίνεται ως:

$$R^2_{adj} = 1 - (1 - R^2) \frac{n-1}{n-k}$$

Εδώ πρέπει να αναφερθεί ότι σε ένα πολυμεταβλητό υπόδειγμα μπορούμε να υπολογίσουμε το μερικό συντελεστή συσχέτισης, ο οποίος μετρά τη συσχέτιση μεταξύ της εξαρτημένης μεταβλητής (Y) και μιας από τις ανεξάρτητες μεταβλητές (π.χ. της μεταβλητής X_1), κρατώντας τις υπόλοιπες μεταβλητές σταθερές. Αυτό πραγματοποιείται ως εξής:

$$r_{YX_1} = \frac{\sum x_1 y}{\sqrt{\sum x_1^2}\sqrt{\sum y^2}}$$

$$r_{YX_2} = \frac{\sum x_2 y}{\sqrt{\sum x_2^2}\sqrt{\sum y^2}}$$

$$r_{X_1X_2} = \frac{\sum x_1 x_2}{\sqrt{\sum x_1^2}\sqrt{\sum x_2^2}}$$

Οι ίδιες ερμηνείες των αποτελεσμάτων ισχύουν όπως και στην περίπτωση απλού συντελεστή συσχέτισης. Επιπλέον, μπορούμε να συσχετίσουμε τον συντελεστή συσχέτισης R_{yx1x2} με την τετραγωνική ρίζα του πολλαπλού συντελεστή προσδιορισμού. Υπάρχει όμως μια πρακτική δυσκολία, που έγκειται στο γεγονός ότι ο δείκτης αυτός είναι πάντα θετικός. Αυτή η δυσκολία περιορίζει τη χρήση του συντελεστή στην απλή μέτρηση του βαθμού γραμμικής συσχέτισης ανάμεσα στην εξαρτημένη μεταβλητή Y και τις k ερμηνευτικές μεταβλητές.

5.5 Διαστήματα εμπιστοσύνης των πληθυσμιακών παραμέτρων

Όπως στην απλή παλινδρόμηση, έτσι και στην πολλαπλή μπορούμε να κατασκευάσουμε ένα 100(1-α)% διάστημα εμπιστοσύνης για την πληθυσμιακή παράμετρο β_j, α στις ιδιότητες των εκτιμητών ελαχίστων τετραγώνων b_j, ως: Pr $[b_j - t_{n-k,a/2}\ \tau.\sigma.(b_j,) <\_ \beta_j <\_ b_j, + t_{n-k,a/2}\ \tau.\sigma.(b_j)] = 100(1-\alpha)°/o$ όπου $\tau.\sigma.(b_j)$ είναι το τυπικό σφάλμα του εκτιμητή b_j.

Επίσης ένα 100(1-α)% διάστημα εμπιστοσύνης για τη μέση τιμή των πληθυσμιακών Y τιμών δίνεται ως :

$$\hat{Y}_{p} \pm t_{n-k,a/2} S_{\hat{Y}_{p}}$$

όπου $\hat{Y}$ είναι η προβλεπόμενη τιμή της Y. ενώ το τυπικό σφάλμα της εκτίμησης στην πολλαπλή παλινδρόμηση δίδεται ως εξής:

$$S_e = \sqrt{\frac{\sum (Y_i - \hat{Y})^2}{n-k}} = \sqrt{\frac{\sum e_i^2}{n-k}} = \sqrt{\frac{SSR}{n-k}}$$

Όπου k ο αριθμός των ανεξάρτητων μεταβλητών συμπεριλαμβανόμενού του σταθερού όρου. Στατιστικά προγράμματα μπορούν να χρησιμοποιηθούν για να υπολογίσουν τις πολύπλοκες αυτές εκφράσεις και να παράσχουν ταυτόχρονα την εκτίμηση των αντίστοιχων διαστημάτων εμπιστοσύνης.

5.6 Εκτίμηση της πληθυσμιακής διακύμανσης στην πολλαπλή παλινδρόμηση

Όπως έχουμε αναφέρει η πληθυσμιακή διακύμανση είναι γενικά άγνωστη για τον ερευνητή. Η διακύμανση του εκτιμητή μπορεί να εξαχθεί από το άθροισμα τετραγώνων των καταλοίπων $SSR=\Sigma e_i^2$. Για ένα υπόδειγμα πολλαπλής παλινδρόμησης με δύο ερμηνευτικές μεταβλητές (και τρεις παραμέτρους) μπορούμε να δείξουμε ότι: $S2 =\frac{SSR}{n-3}$, είναι ένας αμερόληπτος εκτιμητής της πληθυσμιακής διακύμανσης, τέτοιος ώστε $E(s^{2)}= \sigma^2$. Πιο γενικά, για μια πολλαπλή γραμμική παλινδρόμηση με k ανεξάρτητες παραμέτρους συμπεριλαμβανόμενου του σταθερού όρου, ο εκτιμητής:

$S^2 =\frac{SSR}{n-\kappa}$ πρέπει να χρησιμοποιείται. Όταν, ως συνήθως, η σ^2 είναι άγνωστη, την αντικαθιστούμε με την s^2 για το σχηματισμό της διακύμανσης των εκτιμητών ελαχίστων τετραγώνων, 'Ετσι, σε μια γενική περίπτωση πολλαπλής γραμμικής παλινδρόμησης, αν η πληθυσμιακή διακύμανση είναι άγνωστη, μπορούμε να γράψουμε την κατανομή των εκτιμητών ελάχιστων τετραγώνων.

5.7 Ατομικοί έλεγχοι στατιστικής σημαντικότητας των συντελεστών παλινδρόμησης

Αν Θέλουμε να εξετάσουμε την ατομική στατιστική σημαντικότητα μιας συγκεκριμένης ανεξάρτητης μεταβλητής χρησιμοποιούμε, όπως και πριν, τη στατιστική t, με τη διαφορά ότι τώρα οι βαθμοί ελευθερίας δίνονται από τον τύπο β.ε.= n-k. Ξέρουμε κάτω από τις υποθέσεις της πολλαπλής παλινδρόμησης και με την υπόθεση ότι ο διαταρακτικός όρος κατανέμεται σύμφωνα με την κανονική

κατανομή, οι δειγματικές κατανομές των εκτιμητών ελαχίστων τετραγώνων είναι κανονικές.

5.8 Σύγκριση συντελεστών προσδιορισμού (P2) διαφορετικών εξισώσεων παλινδρόμησης

Μία σύγκριση των εκτιμητών μιας δειγματικής παλινδρόμησης ίσως συνεπάγεται την άμεση σύγκριση των συντελεστών προσδιορισμού για δύο ή περισσότερες εξισώσεις. Αυτό είναι αποδεκτό, υπό τον όρο ότι συγκρίνουμε υποδείγματα με τις ίδιες εξαρτημένες μεταβλητές. Καθώς ο συντελεστής προσδιορισμού R^2 μετράει την αναλογία της μεταβλητότητας στην εξαρτημένη μεταβλητή που ερμηνεύεται από την παλινδρόμηση δεν μπορούμε άμεσα να συγκρίνουμε τους συντελεστές προσδιορισμού δύο εξισώσεων που οι εξαρτημένες μεταβλητές είναι διαφορετικές στη συναρτησιακή τους μάθηση. Για παράδειγμα, αν θεωρήσουμε δύο εναλλακτικές παλινδρομήσεις στη Y_i, και στις ερμηνευτικές μεταβλητές X_{1i}, , X_{2i} τότε αν:

Παλινδρόμηση 1: $Y_i=\beta_0+\beta_1 X_{1i} +\beta_2 X_{2i} +\varepsilon_i$

Παλινδρόμηση 2: $\text{In}Y_i = \beta_0 + \beta_1 \text{In}X_{1i}+ \beta_2 \text{In}X_{2i} + \varepsilon_i$

Ο πολλαπλός συντελεστής συσχέτισης μπορεί να συσχετιστεί με την τετραγωνική ρίζα του πολλαπλού συντελεστή προσδιορισμού. Όμως υπάρχει ένα πρόβλημα σχετικά με το γεγονός ότι ο δείκτης αυτός είναι πάντα θετικός. Γι' αυτό και συνήθως γίνεται χρήση της απλής συσχέτιση της εξαρτημένης και των k ανεξάρτητων μεταβλητών

5.9 Χρήση Ψευδομεταβλητών στην Παλινδρόμηση

Πολλές φορές, η μελέτη οικονομικών φαινομένων απαιτεί τη μελέτη ποιοτικών φαινομένων. Χρειάζεται λοιπόν η συμπερίληψη στο μοντέλο της παλινδρόμησης μεταβλητών που να εκφράζουν τα φαινόμενα αυτά. Καθώς πρόκειται για μη μετρήσιμες ιδιότητες, αλλά για ποιοτικές, οι μεταβλητές λαμβάνουν μόνο τις τιμές 0 ή 1, για απουσία ή παρουσία του φαινομένου αντίστοιχα. Οι μεταβλητές αυτές λέγονται ψευδομεταβλητές (dummy variables) ή διχοτομικές/δυαδικές μεταβλητές (dichotomous/binary variables). Πέρα από την κύρια χρήση τους για την αξιοποίηση ποιοτικών δεδομένων, χρησιμεύουν επίσης και στο διαχωρισμό των ποσοτικών

δεδομένων σε ομάδες (δηλώνοντας δηλαδή αν μια μεταβλητή βρίσκεται εντός ή εκτός του εύρους μιας ομάδας).

Οι ψευδομεταβλητές χαρακτηρίζονται ως τεχνητές μεταβλητές που εμφανίζονται με τιμές 0 και 1. Όταν έχουμε το 'φύλο΄, ως μία παράμετρο, αυτό μπορούμε να το παραστήσουμε με μία ψευδομεταβλητή που παίρνει την τιμή 1 αν το άτομο είναι άνδρας και την τιμή 0 αν είναι γυναίκα. Οι ψευδομεταβλητές χρησιμοποιούνται και για ποσοτικές μεταβλητές, π.χ. για την μεταβλητή `ηλικία, όταν το πρόβλημα είναι ο διαχωρισμός των δεδομένων σε κάποιο πλήθος ηλικιακών ομάδων. Σε ένα παράδειγμα παλινδρόμησης, και η εξαρτημένη μεταβλητή αναπαρίσταται με ψευδομεταβλητή. Οι ψευδομεταβλητές έχουν ευρεία εφαρμογή στη μελέτη φαινομένων σε βάθος χρόνου, και γενικότερα στο πώς οι μεταβλητές αλλάζουν ανά χρονικές περιόδους (τρίμηνα, εξάμηνα, έτη, δεκαετίες, κ.λπ.). Σε αυτήν την περίπτωση, οι ψευδομεταβλητές μπορούν να εισαχθούν είτε ως υποκατάστατα των σταθερών όρων, είτε των συντελεστών μεταβολής, είτε και των δύο, ανάλογα με το είδος του φαινομένου που μελετάται.

Μία ακόμη χρησιμότητα των ψευδομεταβλητών έγκειται στην ευελιξία της μελέτης της επίδρασης στην εξαρτημένη μεταβλητή Υ όχι μόνο κάθε ψευδομεταβλητής ξεχωριστά, αλλά και σε συνδυασμό με μία άλλη. Σε αυτή την περίπτωση, το μοντέλο θα περιλαμβάνει και το γινόμενο δύο ή περισσοτέρων ψευδομεταβλητών, με τον αντίστοιχο συντελεστή αλληλεπίδρασης (π.χ. $\mathbf{b_3}$)

$$\hat{y} = b_0 + b_1X_1 + b_2X_2 + b_3X_1X_2$$

Μπορούμε να ερμηνεύσουμε την ύπαρξη αλληλεπίδρασης ή μη και μέσω της γραφικής αναπαράστασης των δύο μεταβλητών.

ΕΙΔΙΚΟ ΜΕΡΟΣ

Κεφάλαιο 6 Εφαρμογή στατιστικής ανάλυσης σε επιχειρησιακά δεδομένα

Στα πλαίσια της παρούσας εργασίας, είναι απαραίτητη η πρακτική εφαρμογή όσων αναφέρθηκαν στο γενικό μέρος. Μέσα από αυτήν, προσφέρεται μια βαθιά κατανόηση της απόκτησης νευραλγικών γνώσεων μέσα από τα δεδομένα μιας που μπορεί να προκύψει από αυτά. Επιπλέον, επιδεικνύεται ο τρόπος με τον οποίον η γνώση μπορεί να προκύψει μέσω της χρήσης πρακτικών εφαρμογών σε υπολογιστικά συστήματα.

Όπως τονίστηκε στην βιβλιογραφική ανασκόπηση, η απόφαση για την κατάλληλη μέθοδο στατιστικής ανάλυσης πρέπει να λαμβάνεται πριν από την έναρξη της μελέτης, στο στάδιο του σχεδιασμού, και το επιλεγμένο μέγεθος δείγματος να είναι το βέλτιστο. Αυτά δεν μπορούν να αποφασιστούν αυθαίρετα μετά τη λήξη της μελέτης ή όταν τα δεδομένα έχουν ήδη συλλεχθεί (Nayak & Hazra, 2011). Το τεστ που θα χρησιμοποιηθεί εξαρτάται κυρίως από το είδος του ερευνητικού ερωτήματος που υποβάλλεται. Στην παρούσα εφαρμογή, με σκοπό την ορθή επιλογή της καταλληλότερης στατιστικής μεθόδου, συνεκτιμήθηκαν ο τύπος των δεδομένων προς ανάλυση, ο χρόνος παράδοσης των αποτελεσμάτων καθώς και οι δυνατότητες/περιορισμοί του διαθέσιμου στατιστικού πακέτου.

6.1 Περιγραφική στατιστικού πακέτου (SPSS)

Το SPSS (Statistical Package for the Social Sciences), που αντιπροσωπεύει ένα δημοφιλές στατιστικό πακέτο για τις κοινωνικές επιστήμες, είναι ένα ισχυρό, φιλικό προς τον χρήστη πακέτο λογισμικού για χειρισμό και στατιστική ανάλυση δεδομένων. Το πακέτο είναι ιδιαίτερα χρήσιμο για φοιτητές και ερευνητές κοινωνικών επιστημών, καθώς εκτελεί ένα ευρύ φάσμα ανάλυσης διμεταβλητών και πολυμεταβλητών. Η ανάλυση δεδομένων ξεκινά γενικά με τον υπολογισμό ενός αριθμού συνοπτικών στατιστικών όπως ο μέσος όρος, η μέση τιμή, η τυπική απόκλιση και συνεχίζεται με τη δημιουργία γραφικών απεικονίσεων των δεδομένων, όπως ιστογράμματα, γραφικές αναπαραστάσεις και διαγράμματα στελέχους-φύλλου

(stem-and-leaf plots) (Landau & Everitt, 2004). Μέσω του λογισμικού, μπορούν να εκτελεστούν τύποι στατιστικών αναλύσεων από τον χρήστη, συμπεριλαμβανομένων των α) διασταυρούμενοι πίνακες SPSS Cross tabs (Crosstabulationtables) για τη διερεύνηση της συσχέτισης ή μη μεταξύ δύο μεταβλητών σε μικρό αριθμό κατηγοριών, β) δοκιμή Chi-square για τον προσδιορισμό της ύπαρξης ή μη στατιστικής σημαντικότητας μεταξύ δύο μεταβλητών, γ) δισδιάστατων και τρισδιάστατων διαγραμμάτων (Griffith, 2010).

Η παρακάτω μελέτη πραγματοποιήθηκε χρησιμοποιώντας το στατιστικό πακέτο SPSS έκδοση 25. Το συγκεκριμένο λογισμικό επιλέχθηκε βάση της μεγάλης δημοφιλίας που το χαρακτηρίζει. Εκτός από την δημοτικότητάς του στους ακαδημαϊκούς και επιχειρηματικούς κύκλους, που το καθιστούν το πλέον διαδεδομένο πακέτο στατιστικής ανάλυσης, το SPSS είναι επίσης ένα ευέλικτο πακέτο που επιτρέπει πολλούς διαφορετικούς τύπους αναλύσεων, μετασχηματισμούς δεδομένων, σε συνδυασμό με ευνόητες μορφές παρεχόμενων αποτελεσμάτων. Μάλιστα, ενημερώνεται και βελτιώνεται συνεχώς, και κάθε σημαντική αναθεώρηση ενσωματώνεται σε κάθε νέα έκδοση αυτού του πακέτου.

Εν ολίγοις, το συγκεκριμένο λογισμικό τείνει να εξυπηρετεί περισσότερο τους ερευνητικούς σκοπούς της παρούσας εργασίας και γι’ αυτόν τον λόγο μαζί με τους ανωτέρω επιλέχθηκε.

6.2 Δειγματοληψία

Ο πληθυσμός της έρευνας αφορά άτομα στα οποία χορηγήθηκαν δάνεια από την εταιρία χορήγησης δανείων “Dream Housing Finance”. Η μέθοδος δειγματοληψίας που χρησιμοποιήθηκε ήταν η βολική. Τα δεδομένα ήταν διαθέσιμα μέσω ενός τυπικού ιστότοπου δωρεάν διάθεσης δεδομένων από επιχειρήσεις.. Συνολικά στην έρευνα χρησιμοποιήθηκαν οικονομικά και δημογραφικά στοιχεία 184 δανειοληπτών της εταιρείας.

6.3 Διαδικασία

Τα δεδομένα μας αφορούν δεδομένα της εταιρίας “Dream Housing Finance”, η οποία ασχολείται με την χορήγηση στεγαστικών δανείων. Στα δεδομένα μας έχουμε δάνεια

τα οποία έχουν εγκριθεί μαζί με άλλες πληροφορίες για τα άτομα στα οποία χορηγήθηκαν αυτά τα δάνεια (δανειολήπτες). Χωρίς να μπορούμε να συμπεράνουμε πολλές πληροφορίες από τα διαθέσιμα δεδομένα μας, όπως παρέχονται από το συγκεκριμένο πακέτο δεδομένων (που ανασύρθηκε από έναν τυπικό ιστότοπο δωρεάν διάθεσης δεδομένων από επιχειρήσεις), με μια πρώτη ματιά συμπεραίνουμε ότι πρόκειται για δεδομένα - οικονομικής κατάστασης (π.χ εισόδημα) δανειοληπτών και λοιπών (εγγυητών, συνεγγυητών).

οικογενειακής κατάστασης (π.χ έγγαμος, εξαρτώμενα μέλη) δανειοληπτών
τύπος απασχόλησης (π.χ ιδιωτικός υπάλληλος) δανειολήπτη
ύψος εισοδήματος του υποβάλλοντα και συνυποβάλλοντα την αίτηση δανείου
επίπεδο εκπαίδευσης
ύψος δανείου και άλλα τρέχοντα εξυπηρετούμενα δάνεια, ιστορικό τραπεζικών πιστώσεων πιστώσεων
γεωγραφική περιοχή χορηγούμενου δανείου (π.χ αστικό κέντρο, προαστιακό θέρετρο κλπ)

Τα δεδομένα μεταφορτώθηκαν από τον ιστότοπο φιλοξενίας τους σε μορφή λογιστικού φύλλου και έπειτα πραγματοποιήθηκε περαιτέρω κωδικοποίηση και μορφοποίηση των δεδομένων με σκοπό την ευκολότερη ανάλυσή τους.

Για κάθε μεταβλητή χρησιμοποιήθηκε η περιγραφική μέθοδος ανάλυσης των δεδομένων και υπολογίστηκαν μέτρα όπως η συχνότητα των απαντήσεων και το ποσοστό ενώ δημιουργήθηκαν, όπου κρίθηκε απαραίτητο, τα αντίστοιχα γραφήματα. Στη συνέχεια δημιουργήθηκε ένα προβλεπτικό μοντέλο για τη σύνδεση ανεξάρτητων μεταβλητών με το ποσό του χορηγούμενου δανείου. Οι πίνακες και τα γραφήματα παρουσιάζονται αναλυτικά στα παραρτήματα της μελέτης.

6.4 Αποτελέσματα στατιστικής ανάλυσης-Περιγραφική Στατιστική

6.4.1 Περιγραφική στατιστική

Βασικός σκοπός της περιγραφικής στατιστικής είναι η παρουσίαση των τιμών του δείγματος με τέτοιο τρόπο ώστε να μπορεί να γίνει μια πρώτη ερμηνεία των αποτελεσμάτων. Στην μελέτη μας, το δείγμα έχει ληφθεί από την εταιρία «Dream

Housing Finance», κατά συνέπεια προχωράμε με τα περιγραφικά μέτρα για το δείγμα μας. Η χρήση περιγραφικών μέτρων είναι σημαντική διότι μπορούν να ανιχνευτούν κάποια ιδιαίτερα χαρακτηριστικά των τιμών του δείγματος (άρα πιθανότατα και του πληθυσμού) τα οποία θα μελετηθούν αναλυτικά αργότερα. Περιγραφικά στοιχεία είναι η μέγιστη τιμή του δείγματος, η ελάχιστη τιμή του δείγματος, το εύρος των τιμών του δείγματος ή η διάμεσος.

Οι μεταβλητές με τις οποίες θα εργαστούμε είναι οι εξής:

Ποσοτικές

Loan Amount (Ποσό δανείου): ποσά σε χρηματικές μονάδες.

Applicant Income (Εισόδημα αιτούντος): ποσά σε χρηματικές μονάδες.

Coapplicant Income (Εισόδημα συναιτούντος): ποσά σε χρηματικές μονάδες.

Ποιοτικές

Gender (Φύλο): με κατηγορίες άντρας (male), γυναίκα (female).

Married (Οικογενειακή κατάσταση): με κατηγορίες έγγαμος (married), άγαμος (not married).

Education (Απόφοιτος τριτοβάθμιας εκπαίδευσης): με κατηγορίες απόφοιτος (graduate), όχι απόφοιτος (notgraduate).

Property_area (Περιοχή ακινήτου): με κατηγορίες αστική περιοχή (urban), ημιαστική περιοχή (semiurban), αγροτική περιοχή (rural).

Dependents (Εξαρτώμενα μέλη): με κατηγορίες 0, 1, 2, 3++

Self_Employed (Αυτοαπασχολούμενος) : με κατηγορίες Ναι, Όχι

6.4.2 Παράθεση αποτελεσμάτων στατιστικής ανάλυσης

Πρώτα θα ασχοληθούμε με τις ποσοτικές μεταβλητές.

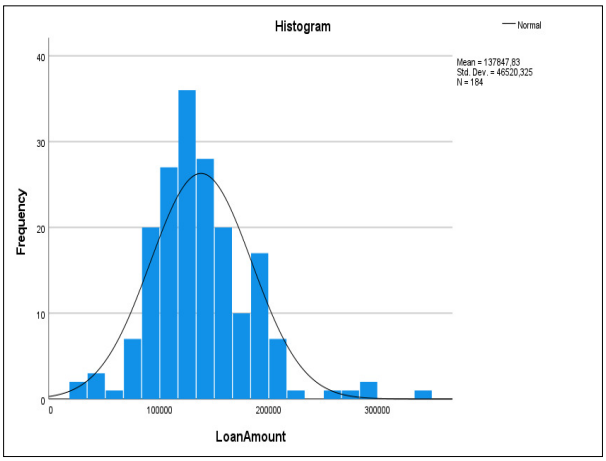
Πίνακας 5 Descriptive Statistics

Πρώτα θα ασχοληθούμε με τις ποσοτικές μεταβλητές.

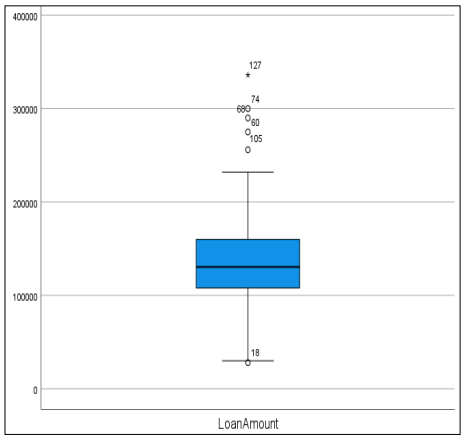
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
ApplicantIncome	184	0	10166	3546,24	1619,162
CoapplicantIncome	184	187	8000	2384,02	1307,976
LoanAmount	184	28000	336000	137847,83	46520,325
Valid N (listwise)	184				

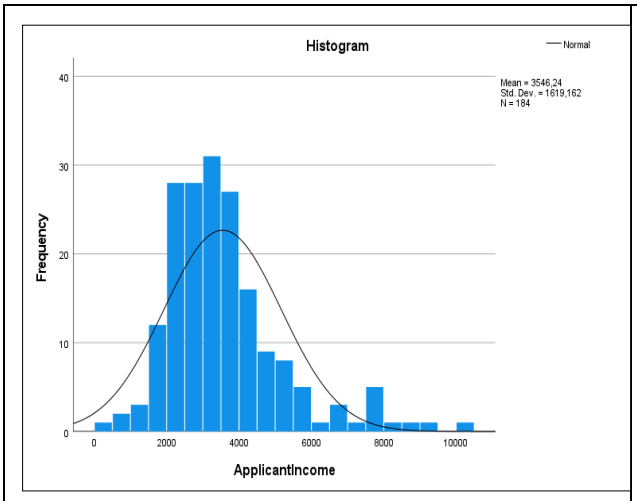
Από τον παραπάνω πίνακα διαπιστώνουμε ότι το μέγεθος του δείγματός μας είναι 184 άτομα. Εξετάζονται οι ποσοτικές μεταβλητές και κάθε στήλη περιέχει τα το μικρότερο, μεγαλύτερο και τον μέσο όρο που αντιστοιχεί στις συγκεκριμένες μεταβλητές.



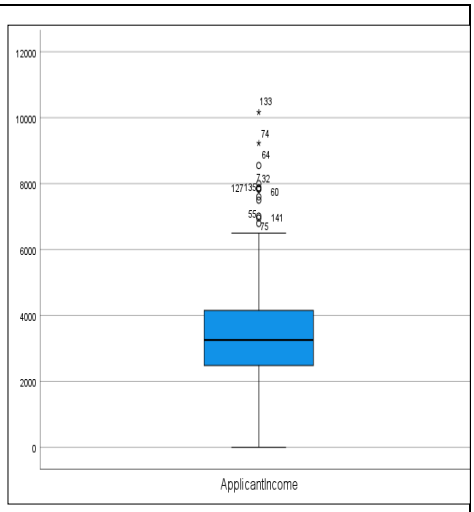
Διάγραμμα 1 Loan Amount



Διάγραμμα 2 Θηκόγραμμα Loan Amount



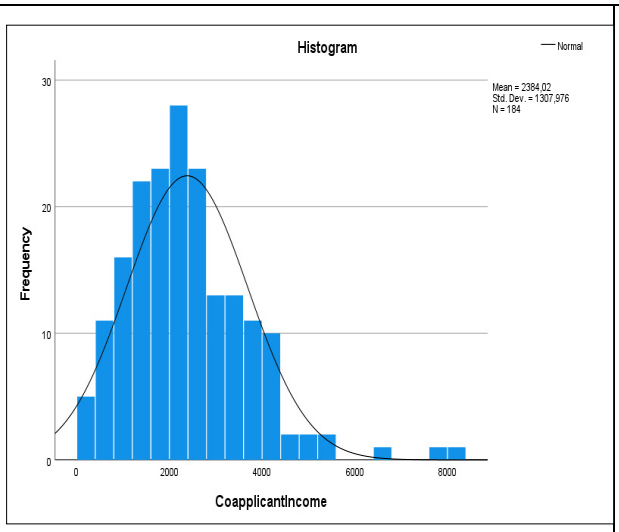
Διάγραμμα 3 Applicant Income



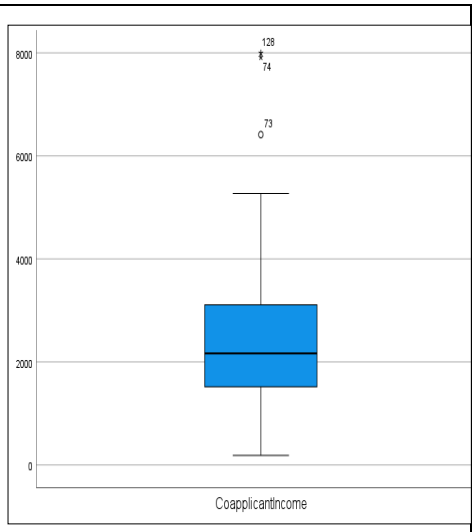
Διάγραμμα 4 ΘηκόγραμμαApplicantIncome

Στο ιστόγραμμα Loan Amount βλέπουμε την προσαρμογή των δεδομένων στην κανονική καμπύλη για τη μεταβλητή καθώς και στο αντίστοιχο θηκόγραμμα όπου θηκόγραμμα παρατηρούμε ότι το ποσό του δανείου έχει μικρή μεταβλητότητα και ότι η μεταβλητή έχει σχεδόν συμμετρική κατανομή με τη διάμεσο στο κέντρο του ορθογωνίου σχήματος. Επίσης, παρατηρούμε κάποιες ακραίες τιμές, με την αντιμετώπιση των οποίων θα ασχοληθούμε αργότερα στην ανάλυση. Παρατηρώντας το ιστόγραμμα και θηκόγραμμα της μεταβλητής Applicant Income ότι η κατανομή είναι σχεδόν συμμετρική ενώ από το θηκόγραμμα μπορούμε να συμπεράνουμε ότι το εισόδημα έχει επίσης μικρή μεταβλητότητα με σχεδόν συμμετρική κατανομή, με τη διάμεσο στο κέντρο περίπου του ορθογωνίου σχήματος. Επίσης, παρατηρούμε κάποιες ακραίες τιμές, με την αντιμετώπιση των οποίων θα ασχοληθούμε αργότερα στην ανάλυση.

Προχωράμε με το ιστόγραμμα και θηκόγραμμα της μεταβλητής Coapplicant Income:



Διάγραμμα 5 Coapplicant Income



Διάγραμμα 6Θηκόγραμμα Coapplicant

Από το ιστόγραμμα φαίνεται ότι η κατανομή είναι σχεδόν συμμετρική. Από το θηκόγραμμα παρατηρούμε ότι η μεταβλητή CoapplicantIncome έχει επίσης μικρή μεταβλητότητα και σχεδόν συμμετρική κατανομή με τη διάμεσο στο κέντρο περίπου του ορθογωνίου σχήματος. Επίσης, παρατηρούμε κάποιες ακραίες τιμές, με την αντιμετώπιση των οποίων θα ασχοληθούμε αργότερα στην ανάλυση.

Συνεχίζουμε με τις ποιοτικές μεταβλητές. Από τον παρακάτω πίνακα διαπιστώνουμε πως καμία από τις ποιοτικές μεταβλητές μας δεν έχει ελλειπούσες τιμές.

Πίνακας 6 Statistics

Statistics

		Gender	Married	Education	PropertyArea	Dependents	Self_Employed
N	Valid	184	184	184	184	184	184
	Missing	0	0	0	0	0	0

Στον παρακάτω πίνακα διαπιστώνουμε το γεγονός πως το 14,1% (26) των αιτούντων ήταν γυναίκες ενώ το 85,9% (158) ήταν άντρες.

Πίνακας 7 Gender

Gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Female	26	14,1	14,1	14,1
	Male	158	85,9	85,9	100,0
	Total	184	100,0	100,0	

Από τον παρακάτω πίνακα διαπιστώνουμε πως το 23,9% (44) των αιτούντων είναι άγαμοι ενώ το 76,1% (140) είναι έγγαμοι.

Πίνακας 8 Οικογενειακή κατάσταση

Married

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Not married	44	23,9	23,9	23,9
	Married	140	76,1	76,1	100,0
	Total	184	100,0	100,0	

Πίνακας 9 Education

Education

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Not graduate	43	23,4	23,4	23,4
	Graduate	141	76,6	76,6	100,0
	Total	184	100,0	100,0	

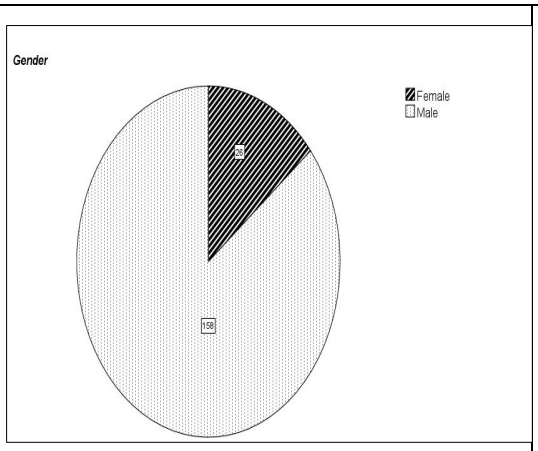
Βλέπουμε από τον πίνακα ότι το 23,4% των αιτούντων δεν είναι απόφοιτοι τριτοβάθμιας εκπαίδευσης ενώ το 76,6% είναι.

Κοιτάζοντας τον παρακάτω πίνακα το 39,7% (73) των ακινήτων βρίσκεται σε αστική περιοχή, το 33,7% (62) βρίσκεται σε ημιαστική περιοχή και το 26,6% (49) βρίσκεται σε αγροτική περιοχή.

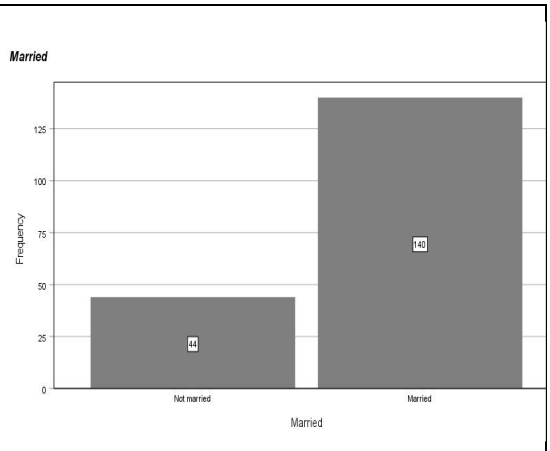
Πίνακας 10 Property Area

Property_Area

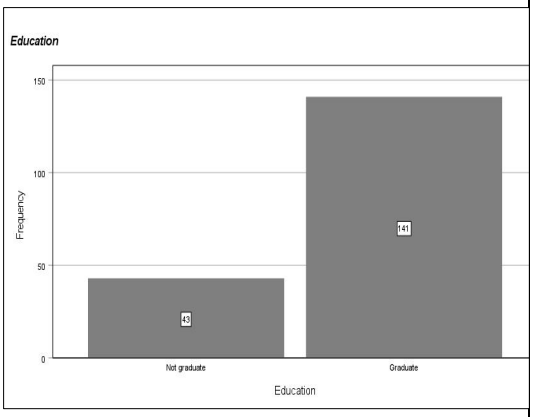
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Rural	49	26,6	26,6	26,6
	Semiurban	62	33,7	33,7	60,3
	Urban	73	39,7	39,7	100,0
	Total	184	100,0	100,0	



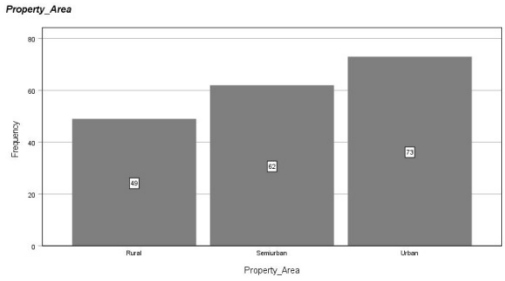
Διάγραμμα 7 Gender



Διάγραμμα 8 Οικογενειακή κατάσταση



Διάγραμμα 9 Education



Διάγραμμα 10 PropertyArea

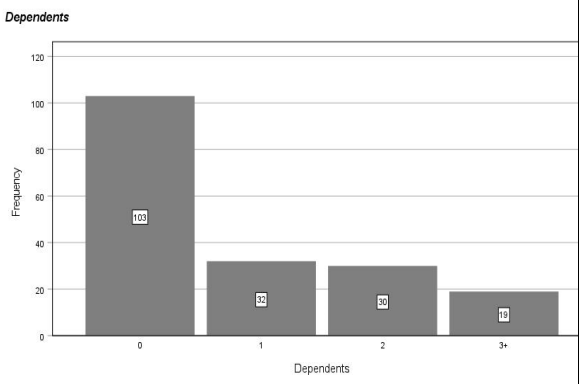
Από το ότι το 23,9% (44) των αιτούντων είναι άγαμοι ενώ το 76,1% (140) είναι έγγαμοι αποτυπώνεται οπτικά στο παραπάνω ραβδόγραμμα καθώς και η πληροφορίας ότι το 23,4% (43) των αιτούντων δεν είναι απόφοιτοι τριτοβάθμιας εκπαίδευσης ενώ το 76,6% (141) είναι. Επίσης το 39,7% (73) των ακινήτων βρίσκεται σε αστική περιοχή, το 33,7% (62) βρίσκεται σε ημιαστική περιοχή και το 26,6% (49) βρίσκεται σε αγροτική περιοχή.

Κοιτάζοντας τον παρακάτω πίνακα το 56% (103) των αιτούντων δεν έχει εξαρτώμενα μέλη, το 17,4% (32) έχει 1 εξαρτώμενο μέλος, το 16,3% (30) έχει 2 εξαρτώμενα μέλη και το 10,3% (19) έχει 3 ή περισσότερα εξαρτώμενα μέλη.

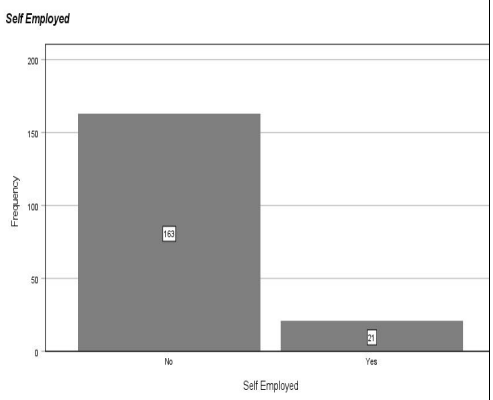
Πίνακας 11 Dependents

Dependents

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	103	56,0	56,0	56,0
	1	32	17,4	17,4	73,4
	2	30	16,3	16,3	89,7
	3+	19	10,3	10,3	100,0
	Total	184	100,0	100,0	



Διάγραμμα 11 Dependents



Διάγραμμα 12 Dependents

Στον παρακάτω πίνακα το 88,6% (163) των αιτούντων δεν είναι αυτοαπασχολούμενοι ενώ το υπόλοιπο 11,4% (21) είναι.

Πίνακας 12 Self Employed

Self Employed

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	163	88,6	88,6	88,6
	Yes	21	11,4	11,4	100,0
	Total	184	100,0	100,0	

6. 5 ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ (REGRESSIONANALYSIS)

Είδαμε τα περιγραφικά μέτρα και προχωράμε με την γραμμική παλινδρόμηση. Σκοπός της μελέτης μας βρίσκοντας ένα μοντέλο γραμμικής παλινδρόμησης είναι να

είμαστε ικανοί μέσω του μοντέλου με τις ανεξάρτητες μεταβλητές να προβλέψουμε το ποσό του δανείου το οποίο μπορεί να χορηγηθεί στον πελάτη.

Η πολλαπλή παλινδρόμηση έχει ως εξαρτημένη μεταβλητή την Loan Amount και ανεξάρτητες τις μεταβλητές Applicant Income, Coapplicant Income, Gender, Married και Property_Area. Για τη μεταβλητή Property_Area θα δημιουργήσουμε ψευδομεταβλητές (dummy variables) για να την προσθέσουμε στο μοντέλο. Οι ψευδομεταβλητές θα είναι οι Urban_rural και Semi urban με κωδικοποίηση 1-0. Όταν και οι δύο είναι μηδέν τότε έχουμε την περίπτωση της αγροτικής περιοχής. Όταν είναι Semi urban=1 τότε Urban_rural=0, οπότε έχουμε την περίπτωση της ημιαστικής περιοχής. Στην αντίθετη περίπτωση, δηλαδή Urban_rural=1 τότε Semi urban=0, μιλάμε για αστική περιοχή. Επίσης θα μετατρέψουμε τη μεταβλητή Dependents σε διχοτομική με τιμές 0 (0 ή 1 εξαρτώμενο μέλος) και 1(2 ή περισσότερα εξαρτώμενα μέλη). Για να πραγματοποιηθεί γραμμική παλινδρόμηση θα πρέπει να ελεγχθεί η γραμμική συσχέτιση ποσοτικών μεταβλητών οι οποίες θέλουμε να βρίσκονται στο μοντέλο της παλινδρόμησης.

Πίνακας 13 Correlations

Correlations

		ApplicantIncome	CoapplicantIncome	LoanAmount
ApplicantIncome	Pearson Correlation	1	-,004	,448**
	Sig. (2-tailed)		,953	<,001
	N	184	184	184
CoapplicantIncome	Pearson Correlation	-,004	1	,350**
	Sig. (2-tailed)	,953		<,001
	N	184	184	184
LoanAmount	Pearson Correlation	,448**	,350**	1
	Sig. (2-tailed)	<,001	<,001	
	N	184	184	184

**. Correlation is significant at the 0.01 level (2-tailed).

Παρατηρούμε πως υπάρχει στατιστικά σημαντική ασθενής συσχέτιση μεταξύ των μεταβλητών Loan Amount και Applicant Income με $r = 0.448$ όπως έχουμε ήδη δει και πολύ ασθενής στατιστικά σημαντική συσχέτιση των μεταβλητών Loan Amount και Coapplicant Income με r=0.350.

Πίνακας 14 Variables Entered/Removed

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	UrbanRural, Gender, Education, Self Employed, Dependents, CoapplicantIncome, Married, ApplicantIncome, SemiUrban[b]	.	Enter

a. Dependent Variable: LoanAmount

b. All requested variables entered.

Πίνακας 15 Coefficients

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	65245,269	13819,591		4,721	<,001	37969,659	92520,878		
	ApplicantIncome	12,727	1,896	,443	6,712	<,001	8,985	16,470	,877	1,140
	CoapplicantIncome	12,392	2,338	,348	5,300	<,001	7,777	17,007	,884	1,132
	Gender	-926,678	8450,946	-,007	-,110	,913	-17606,238	15752,883	,949	1,054
	Married	-7642,695	7130,557	-,070	-1,072	,285	-21716,214	6430,824	,889	1,125
	Education	1798,316	7147,808	,016	,252	,802	-12309,252	15905,884	,899	1,113
	SemiUrban	-1898,611	7857,327	-,019	-,242	,809	-17406,549	13609,327	,596	1,678
	Self Employed	6243,588	9295,182	,043	,672	,503	-12102,234	24589,410	,941	1,063
	Dependents	6093,753	6885,529	,058	,885	,377	-7496,156	19683,662	,887	1,127
	UrbanRural	3687,695	7514,612	,039	,491	,624	-11143,831	18519,220	,608	1,644

a. Dependent Variable: LoanAmount

Διαπιστώνουμε από το μοντέλο ότι η μεταβλητή Gender δεν είναι στατιστικά σημαντική με p-value=0.913>0.05 και έχει το μεγαλύτερο p-value οπότε την αφαιρούμε από το μοντέλο και εφαρμόζουμε πάλι την παλινδρόμηση:

Πίνακας16 Variables Entered/Removed

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	UrbanRural, Education, Self Employed, Married, CoapplicantIncome, Dependents, ApplicantIncome, SemiUrban[b]	.	Enter

a. Dependent Variable: LoanAmount

b. All requested variables entered.

Πίνακας17 Coefficients

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	64607,912	12502,362		5,168	<,001	39933,094	89282,729		
	ApplicantIncome	12,716	1,888	,443	6,735	<,001	8,990	16,442	,880	1,137
	CoapplicantIncome	12,398	2,331	,349	5,319	<,001	7,798	16,999	,884	1,131
	Married	-7760,676	7028,989	-,071	-1,104	,271	-21633,175	6111,824	,909	1,100
	Education	1745,249	7111,247	,016	,245	,806	-12289,598	15780,096	,903	1,108
	SemiUrban	-1893,834	7834,995	-,019	-,242	,809	-17357,078	13569,410	,596	1,678
	Self Employed	6154,638	9233,543	,042	,667	,506	-12068,798	24378,073	,948	1,054
	Dependents	6065,420	6861,229	,058	,884	,378	-7475,987	19606,827	,889	1,125
	UrbanRural	3722,784	7486,573	,039	,497	,620	-11052,809	18498,377	,609	1,641

a. Dependent Variable: LoanAmount

Έχοντας αφαιρέσει την μεταβλητή Gender, βλέπουμε πως από τις εναπομείνασες μεταβλητές η μεταβλητή SemiUrban δεν είναι στατιστικά σημαντική με p-value=0.809>0.05 και έχει το μεγαλύτερο p-value οπότε την αφαιρούμε από το μοντέλο και εφαρμόζουμε πάλι την παλινδρόμηση:

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	UrbanRural, Education, Self Employed, Married, CoapplicantIncome, Dependents, ApplicantIncome[b]	.	Enter

a. Dependent Variable: LoanAmount

b. All requested variables entered.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	63095,744	10795,585		5,845	<,001	41790,285	84401,202		
	ApplicantIncome	12,696	1,881	,442	6,749	<,001	8,983	16,408	,881	1,135
	CoapplicantIncome	12,519	2,270	,352	5,515	<,001	8,039	17,000	,927	1,079
	Married	-7652,278	6995,880	-,070	-1,094	,276	-21458,887	6154,331	,913	1,095
	Education	1870,644	7073,303	,017	,264	,792	-12088,763	15830,050	,908	1,102
	Self Employed	6635,523	8992,517	,045	,738	,462	-11111,519	24382,565	,995	1,006
	Dependents	6011,073	6839,176	,057	,879	,381	-7486,276	19508,422	,890	1,124
	UrbanRural	4824,534	5923,190	,051	,815	,416	-6865,084	16514,152	,968	1,033

a. Dependent Variable: LoanAmount

Έχοντας αφαιρέσει και την μεταβλητή SemiUrban, βλέπουμε πως από τις εναπομείνασες μεταβλητές η μεταβλητή Education δεν είναι στατιστικά σημαντική με p-value=0.792>0.05 και έχει το μεγαλύτερο p-value οπότε την αφαιρούμε από το μοντέλο και εφαρμόζουμε πάλι την παλινδρόμηση:

Επαναλαμβάνουμε την διαδικασία:

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	UrbanRural, Self Employed, Dependents, CoapplicantIncome, ApplicantIncome, Married[b]	.	Enter

a. Dependent Variable: LoanAmount

b. All requested variables entered.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	63996,658	10217,077		6,264	<,001	43833,694	84159,623		
	ApplicantIncome	12,817	1,820	,446	7,044	<,001	9,226	16,408	,937	1,067
	CoapplicantIncome	12,590	2,249	,354	5,599	<,001	8,152	17,027	,940	1,064
	Married	-7658,488	6977,436	-,070	-1,098	,274	-21428,159	6111,184	,913	1,095
	Self Employed	6611,447	8968,401	,045	,737	,462	-11087,307	24310,202	,995	1,005
	Dependents	5722,878	6734,039	,055	,850	,397	-7566,460	19012,215	,913	1,095
	UrbanRural	4873,145	5904,762	,051	,825	,410	-6779,650	16525,940	,969	1,032

a. Dependent Variable: LoanAmount

Σε αυτή την περίπτωση παρατηρούμε ότι η μεταβλητή Self Employed δεν είναι στατιστικά σημαντική με p-value=0.462>0.05 οπότε θα την αφαιρέσουμε από το μοντέλο:

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	UrbanRural, Dependents, CoapplicantIncome, ApplicantIncome, Married[b]	.	Enter

a. Dependent Variable: LoanAmount

b. All requested variables entered.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	64578,801	10173,443		6,348	<,001	44502,724	84654,879		
	ApplicantIncome	12,830	1,817	,447	7,060	<,001	9,244	16,416	,937	1,067
	CoapplicantIncome	12,610	2,246	,355	5,616	<,001	8,179	17,042	,940	1,064
	Married	-7398,695	6959,588	-,068	-1,063	,289	-21132,614	6335,224	,915	1,092
	Dependents	5588,399	6722,929	,053	,831	,407	-7678,501	18855,300	,914	1,095
	UrbanRural	4659,297	5890,064	,049	,791	,430	-6964,043	16282,637	,972	1,029

a. Dependent Variable: LoanAmount

Σε αυτή την περίπτωση παρατηρούμε ότι η μεταβλητή Urban Rural δεν είναι στατιστικά σημαντική με p-value=0.430>0.05 οπότε θα την αφαιρέσουμε από το μοντέλο:

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Dependents, CoapplicantIncome, ApplicantIncome, Married[b]	.	Enter

a. Dependent Variable: LoanAmount

b. All requested variables entered.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	65996,367	10003,895		6,597	<,001	46255,627	85737,108		
	ApplicantIncome	12,933	1,811	,450	7,143	<,001	9,360	16,506	,942	1,062
	CoapplicantIncome	12,474	2,237	,351	5,577	<,001	8,060	16,887	,946	1,057
	Married	-6967,428	6930,945	-,064	-1,005	,316	-20644,299	6709,443	,921	1,086
	Dependents	5826,926	6709,139	,056	,869	,386	-7412,255	19066,107	,915	1,092

a. Dependent Variable: LoanAmount

Σε αυτή την περίπτωση παρατηρούμε ότι η μεταβλητή Dependents δεν είναι στατιστικά σημαντική με p-value=0.386>0.05 οπότε θα την αφαιρέσουμε από το μοντέλο:

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Married, ApplicantIncome, CoapplicantIncome[b]	.	Enter

a. Dependent Variable: LoanAmount

b. All requested variables entered.

Σε αυτή την περίπτωση παρατηρούμε ότι και η μεταβλητή Married δεν είναι στατιστικά σημαντική με p-value=0.373>0.05 οπότε θα την αφαιρέσουμε από το μοντέλο.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	62191,559	8594,787		7,236	<,001	45232,695	79150,423		
	ApplicantIncome	12,914	1,755	,449	7,359	<,001	9,451	16,377	1,000	1,000
	CoapplicantIncome	12,525	2,173	,352	5,765	<,001	8,238	16,812	1,000	1,000

a. Dependent Variable: LoanAmount

Το μοντέλο μας με όλους τους συντελεστές στατιστικά σημαντικούς σε επίπεδο σημαντικότητας 5% είναι το εξής:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3$$

$$\Rightarrow$$

$$LoanAmount = 62191.559 + 12.914 \times ApplicantIncome + 12.525 \times CoapplicantIncome$$

Αύξηση του εισοδήματος του αιτούντος κατά 1 χρηματική μονάδα σημαίνει αύξηση του ποσού του δανείου που μπορεί να χορηγηθεί κατά 12914. Αύξηση του εισοδήματος του συναιτούντος κατά 1 χρηματική μονάδα σημαίνει αύξηση του ποσού του δανείου που μπορεί να χορηγηθεί κατά 12525.

Στον παρακάτω πίνακα ανάλυσης διακύμανσης ελέγχεται αν υπάρχει τουλάχιστον μία προβλεπτική μεταβλητή που να επηρεάζει την μεταβλητή κριτήριο. Σε αυτή την περίπτωση φαίνεται ότι το προβλεπτικό μας μοντέλο έχει καλή προσαρμογή καθώς

F-test value-43.509, p-value=<0.001. Επίσης από τη σύνοψη του μοντέλου μας φαίνεται ότι το μοντέλο μας εξηγεί το 32.5% της συνολικής μεταβλητότητας του ποσού του δανείου.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.286E+11	2	6.429E+10	43,509	<,001[b]
	Residual	2.675E+11	181	1477658540		
	Total	3.960E+11	183			

a. Dependent Variable: LoanAmount

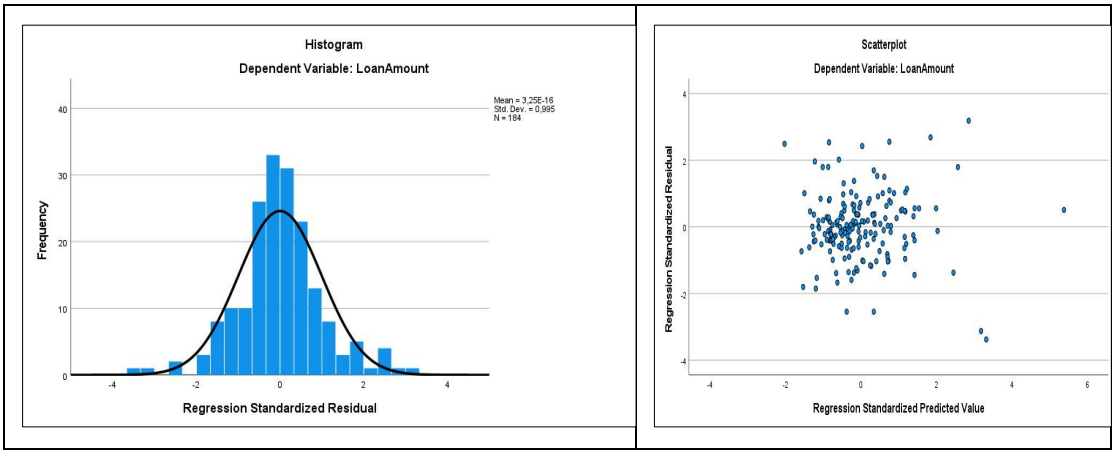
b. Predictors: (Constant), CoapplicantIncome, ApplicantIncome

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,570[a]	,325	,317	38440,324	,325	43,509	2	181	<,001

a. Predictors: (Constant), CoapplicantIncome, ApplicantIncome

b. Dependent Variable: LoanAmount



Από το ιστόγραμμα διαπιστώνουμε πως το κατάλοιπα κατανέμονται σχετικά κανονικά. Στο διάγραμμα σκεδασμού φαίνεται ότι υπάρχει η τυχαιότητα αλλά δεν μπορούμε να μιλήσουμε για ομοσκεδαστικότητα, διότι και στα δύο διαγράμματα βλέπουμε πως οι ακραίες τιμές επηρεάζουν την παλινδρόμηση και κατά συνέπεια τα κατάλοιπα. Προχωράμε με την stepwise μέθοδο του SPSS, να διασταυρώσουμε τα αποτελέσματά μας. (Stepwise: Το πακέτο αποφασίζει βάσει αποτελεσμάτων ποια μεταβλητή θα προστεθεί και ποια θα αφαιρεθεί από το μοντέλο καταλήγοντας σε ένα τελικό μοντέλο.)

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	ApplicantIncome	.	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
2	CoapplicantIncome	.	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).

a. Dependent Variable: LoanAmount

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	92206,909	7419,329		12,428	<,001	77567,949	106845,869		
	ApplicantIncome	12,870	1,904	,448	6,759	<,001	9,113	16,627	1,000	1,000
2	(Constant)	62191,559	8594,787		7,236	<,001	45232,695	79150,423		
	ApplicantIncome	12,914	1,755	,449	7,359	<,001	9,451	16,377	1,000	1,000
	CoapplicantIncome	12,525	2,173	,352	5,765	<,001	8,238	16,812	1,000	1,000

a. Dependent Variable: LoanAmount

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,448[a]	,201	,196	41705,913	,201	45,689	1	182	<,001
2	,570[b]	,325	,317	38440,324	,124	33,236	1	181	<,001

a. Predictors: (Constant), ApplicantIncome

b. Predictors: (Constant), ApplicantIncome, CoapplicantIncome

c. Dependent Variable: LoanAmount

Το τελικό μοντέλο στο οποίο κατέληξε το υπολογιστικό πακέτο SPSS συμπίπτει με το δικό μας το οποίο κατασκευάσαμε παραπάνω. Συνεπώς, ισχύουν τα ίδια συμπεράσματα και οι ίδιες ερμηνείες.

ΑΚΡΑΙΕΣ ΤΙΜΕΣ (OUTLIERS)

Θα εφαρμόσουμε τα παραπάνω έχοντας αφαιρέσει τις ακραίες τιμές για να ελέγξουμε εάν θα υπάρξει βελτίωση στο τελικό μοντέλο και στην εικόνα των καταλοίπων.

Για την αντιμετώπιση των ακραίων τιμών θα χρησιμοποιήσουμε δύο τεχνικές τις οποίες μας δίνει το υπολογιστικό πακέτο SPSS:

Mahalanobis' Distance:Η απόσταση Mahalanobis είναι ένα μέτρο της απόστασης μεταξύ ενός σημείου P και μιας κατανομής D, που εισήγαγε ο PCMahalanobis το 1936. Είναι μια πολυδιάστατη γενίκευση της ιδέας της μέτρησης του αριθμού των τυπικών αποκλίσεων μακριά από το P από τη μέση τιμή D. Η απόσταση Mahalanobis μίας παρατήρησης από ένα σύνολο παρατηρήσεων με μέση και πίνακα συνδιακύμανσης S ορίζεται ως εξής:

$X^{2\sim(}$απόσταση Mahalanobis, αριθμός ανεξάρτητων μεταβλητών στην παλινδρόμησή μας), υπολογίζουμε πιθανότητες βασισμένες στην απόσταση Mahalanobis. Όσες παρατηρήσεις έχουν πιθανότητα μικρότερη από 0,001, θεωρούνται ακραίες τιμές οι οποίες αλλοιώνουν την ανάλυσή μας οπότε θα πρέπει να αφαιρεθούν από την ανάλυσή μας.

Leveragevalues: Είναι οι τιμές της διαγώνιου (στοιχεία *hii*) του πίνακα προβολής $H = X(X^TX)^{-1}X^T$που ονομάζεται μόχλευση (leverage). Αν ένα σημείο έχει πολύ μεγάλη μόχλευση, τότε το σημείο αυτό επιδρά στην προσαρμογή της παλινδρόμησης ώστε η ευθεία παλινδρόμησης να περνά πολύ κοντά από το σημείο αυτό. Σημεία με μόχλευση $h_{ij}>2p/n$,όπου p, ο αριθμός των μεταβλητών στην εκτίμηση της παλινδρόμησης και n ο αριθμός του δείγματος, θεωρούνται υποψήφια σημεία επιρροής και χρίζουν περαιτέρω εξέτασης.

	UrbanRural	ZRE_1	pMAH_1	LEV_1
1	1,00	1,02354	,00000	,08180
2	,00	-2,57676	,00008	,07524
3	1,00	-,93035	,00011	,01910
4	1,00	-2,67072	,00298	,04547
5	,00	,64980	,00613	,00849
6	,00	,59031	,00711	,03053
7	,00	3,20949	,01332	,04958
8	,00	-,00911	,01505	,00620
9	1,00	-,64379	,02843	,00503
10	1,00	1,91053	,02908	,02655
11	1,00	,68477	,02931	,00721
12	,00	,73794	,02990	,01027

Σύμφωνα με τα παραπάνω προκύπτει ότι τρεις παρατηρήσεις επηρεάζουν την παλινδρόμησή μας οπότε θα τις αφαιρέσουμε για να δούμε εάν υπάρχει διαφορά με την παραπάνω παλινδρόμηση που καταλήξαμε.

Θα εργαστούμε όπως παραπάνω, δηλαδή θα εφαρμόσουμε μία πολλαπλή παλινδρόμηση με εξαρτημένη μεταβλητή Loan Amount και με όλες τις ανεξάρτητες μεταβλητές (Applicant Income, Copplicant Income, Gender, Married, Education, Urban_rural, Semi urban, Dependents,) και θα καταλήξουμε στο βέλτιστο μοντέλο γραμμικής πολλαπλής παλινδρόμησης.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	UrbanRural, Education, Self Employed, Gender, Dependents, Married, CoapplicantIncome, ApplicantIncome, SemiUrban[b]	.	Enter

a. Dependent Variable: LoanAmount

b. All requested variables entered.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	56423,077	14301,474		3,945	<,001	28192,912	84653,243		
	ApplicantIncome	13,588	1,993	,449	6,817	<,001	9,653	17,522	,889	1,125
	CoapplicantIncome	14,727	2,580	,382	5,709	<,001	9,635	19,819	,860	1,163
	Gender	-285,459	8186,732	-,002	-,035	,972	-16445,527	15874,610	,950	1,053
	Married	-6468,037	6941,490	-,062	-,932	,353	-20170,080	7234,007	,883	1,132
	Education	1968,879	6941,468	,019	,284	,777	-11733,121	15670,879	,897	1,114
	SemiUrban	-3811,356	7637,814	-,040	-,499	,618	-18887,896	11265,184	,596	1,677
	Self Employed	4321,502	9017,869	,031	,479	,632	-13479,175	22122,179	,939	1,065
	Dependents	10860,454	6789,339	,106	1,600	,112	-2541,253	24262,161	,884	1,131
	UrbanRural	1954,893	7351,370	,021	,266	,791	-12556,226	16466,012	,608	1,645

a. Dependent Variable: LoanAmount

Διαπιστώνουμε από το μοντέλο ότι η μεταβλητή Gender δεν είναι στατιστικά σημαντική με p-value=0.972>0.05 και έχει το μεγαλύτερο p-value οπότε την αφαιρούμε από το μοντέλο και εφαρμόζουμε πάλι την παλινδρόμηση:

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	UrbanRural, Education, Self Employed, Married, Dependents, CoapplicantIncome, ApplicantIncome, SemiUrban[b]	.	Enter

a. Dependent Variable: LoanAmount

b. All requested variables entered.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	56220,770	13033,666		4,314	<,001	30494,239	81947,301		
	ApplicantIncome	13,585	1,986	,449	6,841	<,001	9,665	17,504	,890	1,123
	CoapplicantIncome	14,731	2,570	,382	5,731	<,001	9,657	19,804	,861	1,161
	Married	-6503,551	6846,394	-,062	-,950	,343	-20017,321	7010,219	,903	1,108
	Education	1952,056	6904,547	,018	,283	,778	-11676,498	15580,611	,902	1,109
	SemiUrban	-3809,751	7615,467	-,040	-,500	,618	-18841,558	11222,055	,596	1,677
	Self Employed	4293,496	8955,912	,030	,479	,632	-13384,150	21971,142	,947	1,056
	Dependents	10852,370	6765,650	,105	1,604	,111	-2502,023	24206,763	,885	1,130
	UrbanRural	1966,192	7322,871	,021	,269	,789	-12488,072	16420,456	,609	1,642

a. Dependent Variable: LoanAmount

Διαπιστώνουμε από το μοντέλο ότι η μεταβλητή Urban Rural δεν είναι στατιστικά σημαντική με p-value=0.789>0.05 και έχει το μεγαλύτερο p-value οπότε την αφαιρούμε από το μοντέλο και εφαρμόζουμε πάλι την παλινδρόμηση:

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Dependents, Self Employed, Education, SemiUrban, Married, CoapplicantIncome, ApplicantIncome[b]	.	Enter

a. Dependent Variable: LoanAmount

b. All requested variables entered.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	57655,771	11855,579		4,863	<,001	34255,568	81055,973		
	ApplicantIncome	13,601	1,979	,449	6,871	<,001	9,694	17,508	,891	1,122
	CoapplicantIncome	14,602	2,519	,379	5,798	<,001	9,631	19,573	,892	1,121
	Married	-6461,939	6826,259	-,061	-,947	,345	-19935,414	7011,535	,903	1,107
	Education	1934,616	6885,700	,018	,281	,779	-11656,182	15525,414	,902	1,109
	SemiUrban	-5058,527	6014,079	-,053	-,841	,401	-16928,943	6811,890	,951	1,052
	Self Employed	3893,823	8807,627	,028	,442	,659	-13490,418	21278,064	,974	1,027
	Dependents	10986,575	6729,043	,107	1,633	,104	-2295,017	24268,166	,890	1,124

a. Dependent Variable: LoanAmount

Σε αυτή την περίπτωση παρατηρούμε ότι η μεταβλητή Education δεν είναι στατιστικά σημαντική με p-value=0.779>0.05 οπότε θα την αφαιρέσουμε από το μοντέλο:

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Dependents, Self Employed, SemiUrban, ApplicantIncome, CoapplicantIncome, Married[b]	.	Enter

a. Dependent Variable: LoanAmount

b. All requested variables entered.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	58599,618	11339,544		5,168	<,001	36218,856	80980,379		
	ApplicantIncome	13,740	1,912	,454	7,187	<,001	9,967	17,513	,950	1,052
	CoapplicantIncome	14,682	2,496	,381	5,882	<,001	9,756	19,608	,903	1,107
	Married	-6469,800	6808,111	-,062	-,950	,343	-19906,909	6967,310	,903	1,107
	SemiUrban	-5190,377	5979,852	-,055	-,868	,387	-16992,759	6612,004	,957	1,045
	Self Employed	3828,875	8781,258	,027	,436	,663	-13502,620	21160,370	,974	1,026
	Dependents	10685,692	6625,675	,104	1,613	,109	-2391,345	23762,730	,913	1,095

a. Dependent Variable: LoanAmount

Σε αυτή την περίπτωση παρατηρούμε ότι η μεταβλητή Self Employed δεν είναι στατιστικά σημαντική με p-value=0.663>0.05 οπότε θα την αφαιρέσουμε από το μοντέλο:

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Dependents, SemiUrban, ApplicantIncome, Married, CoapplicantIncome[b]	.	Enter

a. Dependent Variable: LoanAmount

b. All requested variables entered.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	58978,043	11280,090		5,229	<,001	36715,517	81240,570		
	ApplicantIncome	13,762	1,907	,454	7,218	<,001	9,999	17,525	,951	1,052
	CoapplicantIncome	14,705	2,490	,381	5,907	<,001	9,792	19,619	,904	1,106
	Married	-6362,734	6787,920	-,061	-,937	,350	-19759,457	7033,988	,905	1,105
	SemiUrban	-5568,685	5902,868	-,059	-,943	,347	-17218,659	6081,290	,977	1,023
	Dependents	10626,506	6608,938	,103	1,608	,110	-2416,977	23669,989	,913	1,095

a. Dependent Variable: LoanAmount

Σε αυτή την περίπτωση παρατηρούμε ότι η μεταβλητή Married δεν είναι στατιστικά σημαντική με p-value=0.350>0.05 οπότε θα την αφαιρέσουμε από το μοντέλο:

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Dependents, SemiUrban, ApplicantIncome, CoapplicantIncome[b]	.	Enter

a. Dependent Variable: LoanAmount

b. All requested variables entered.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	53885,391	9882,102		5,453	<,001	34382,723	73388,060		
	ApplicantIncome	13,538	1,891	,447	7,159	<,001	9,806	17,270	,966	1,035
	CoapplicantIncome	15,159	2,441	,393	6,209	<,001	10,341	19,977	,939	1,065
	SemiUrban	-4980,530	5867,401	-,052	-,849	,397	-16560,048	6598,988	,988	1,012
	Dependents	9832,429	6552,160	,096	1,501	,135	-3098,484	22763,341	,929	1,077

a. Dependent Variable: LoanAmount

Σε αυτή την περίπτωση παρατηρούμε ότι η μεταβλητή SemiUrban δεν είναι στατιστικά σημαντική με p-value=0.397>0.05 οπότε θα την αφαιρέσουμε από το μοντέλο:

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Dependents, ApplicantIncome, CoapplicantIncome[b]	.	Enter

a. Dependent Variable: LoanAmount

b. All requested variables entered.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Collinearity Statist	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VI
1	(Constant)	51583,860	9495,362		5,433	<,001	32845,170	70322,551		
	ApplicantIncome	13,566	1,889	,448	7,181	<,001	9,838	17,295	,966	1,
	CoapplicantIncome	15,375	2,426	,399	6,337	<,001	10,587	20,163	,950	1,
	Dependents	9812,059	6546,941	,095	1,499	,136	-3108,049	22732,167	,929	1,

a. Dependent Variable: LoanAmount

Σε αυτή την περίπτωση παρατηρούμε ότι η μεταβλητή Dependents δεν είναι στατιστικά σημαντική με p-value=0.136>0.05 οπότε θα την αφαιρέσουμε από το μοντέλο:

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	CoapplicantIncome, ApplicantIncome[b]	.	Enter

a. Dependent Variable: LoanAmount

b. All requested variables entered.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	54308,446	9352,277		5,807	<,001	35852,841	72764,051		
	ApplicantIncome	14,017	1,872	,463	7,489	<,001	10,323	17,710	,991	1,009
	CoapplicantIncome	14,628	2,383	,379	6,139	<,001	9,926	19,330	,991	1,009

a. Dependent Variable: LoanAmount

Το τελικό μοντέλο το ποίο συμπίπτει με το προηγούμενο με όλους τους συντελεστές στατιστικά σημαντικούς σε επίπεδο σημαντικότητας 5% είναι το εξής:

$$\hat{Y}_i=\hat{\beta}_0+\hat{\beta}_1 X_{1i}+\hat{\beta}_2 X_{2i} \Rightarrow$$

Loan Amount=54308.446+14.017 Applicant Income +14.628 Coapplicant Income

Αύξηση του εισοδήματος αιτούντος κατά 1 χρηματική μονάδα σημαίνει αύξηση του ποσού του δανείου που μπορεί να χορηγηθεί κατά 14.017. Αύξηση του εισοδήματος συναιτούντος κατά 1 χρηματική μονάδα σημαίνει αύξηση του ποσού του δανείου που μπορεί να χορηγηθεί κατά 14.628.

Στον παρακάτω πίνακα ανάλυσης φαίνεται ότι η παλινδρόμηση είναι στατιστικά σημαντική καθώς *F-test value*=43.010,*p-value*< 0.001. Επίσης από τη σύνοψη του μοντέλου μας φαίνεται ότι το μοντέλο μας εξηγεί το 32.5% της συνολικής μεταβλητότητας του ποσού του δανείου.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.200E+11	2	5.998E+10	43,010	<,001[b]
	Residual	2.482E+11	178	1394613199		
	Total	3.682E+11	180			

a. Dependent Variable: LoanAmount

b. Predictors: (Constant), CoapplicantIncome, ApplicantIncome

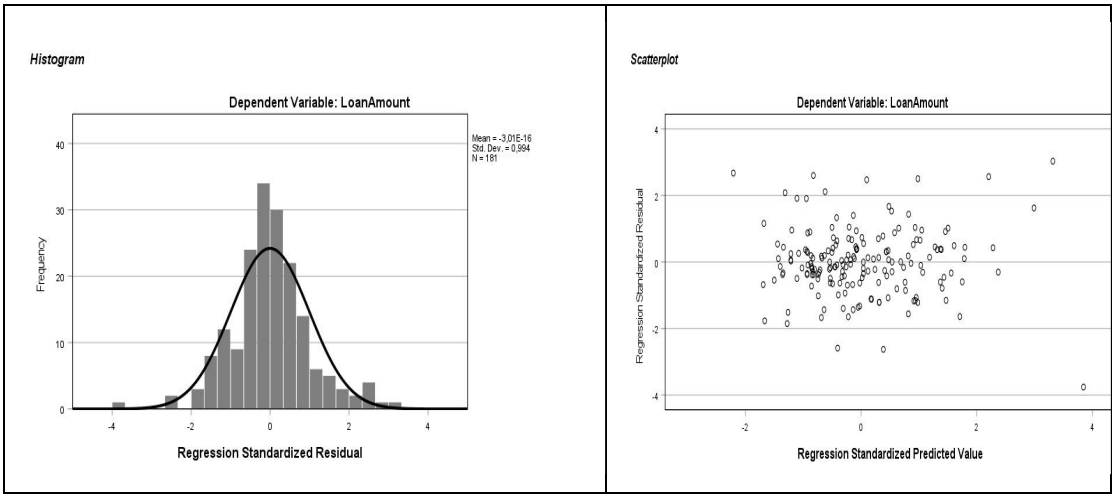
Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,571[a]	,326	,318	37344,520	,326	43,010	2	178	<,001

a. Predictors: (Constant), CoapplicantIncome, ApplicantIncome

b. Dependent Variable: LoanAmount

Από το παρακάτω ιστόγραμμα διαπιστώνουμε πως το κατάλοιπα κατανέμονται σχετικά κανονικά από το ιστόγραμμα. Στο διάγραμμα σκεδασμού διακρίνεται η τυχαιότητα οπότε μπορούμε να πούμε πως ισχύει και η υπόθεση της ομοσκεδαστικότητας.



Προχωράμε με την stepwise μέθοδο του SPSS, να διασταυρώσουμε τα αποτελέσματά μας.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	ApplicantIncome	.	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
2	CoapplicantIncome	.	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).

a. Dependent Variable: LoanAmount

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	92095,273	7729,577		11,915	<,001	76842,456	107348,089		
	ApplicantIncome	12,956	2,046	,428	6,333	<,001	8,919	16,993	1,000	1,000
2	(Constant)	54308,446	9352,277		5,807	<,001	35852,841	72764,051		
	ApplicantIncome	14,017	1,872	,463	7,489	<,001	10,323	17,710	,991	1,009
	CoapplicantIncome	14,628	2,383	,379	6,139	<,001	9,926	19,330	,991	1,009

a. Dependent Variable: LoanAmount

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,428[a]	,183	,178	40993,690	,183	40,107	1	179	<,001
2	,571[b]	,326	,318	37344,520	,143	37,692	1	178	<,001

a. Predictors: (Constant), ApplicantIncome

b. Predictors: (Constant), ApplicantIncome, CoapplicantIncome

c. Dependent Variable: LoanAmount

Το τελικό μοντέλο στο οποίο κατέληξε το υπολογιστικό πακέτο SPSS συμπίπτει με το δικό μας το οποίο κατασκευάσαμε παραπάνω. Συνεπώς, ισχύουν τα ίδια συμπεράσματα και οι ίδιες ερμηνείες.

Το τελικό μοντέλο ήταν ελάχιστα διαφορετικό με και χωρίς τις ακραίες τιμές αλλά η αξιοπιστία του πρώτου μοντέλου είναι αμφίβολη ενώ του δεύτερου μοντέλου είναι μεγαλύτερη.

ΣΥΜΠΕΡΑΣΜΑΤΑ

Οι εκτιμήσεις των συντελεστών του μοντέλου έχοντας αφαιρέσει τις ακραίες τιμές είναι πιο αξιόπιστες από αυτές του πρώτου μοντέλου. Αυτό συμβαίνει διότι τα κατάλοιπα δεν είχαν την επιθυμητή εικόνα βάσει των διαγραμμάτων ενώ στην δεύτερη περίπτωση βελτιώθηκε πολύ.

Ο συντελεστής προσδιορισμού r=0.350 έχει τιμή μικρή κοντά στο +**1**, δεν παρουσιάζει δηλαδή τέλεια προσαρμοστικότητα.

Η παλινδρόμηση ήταν στατιστικά σημαντική με τις ίδιες μεταβλητές μέσα στο μοντέλο αλλά η σημαντική διαφορά ήταν στα κατάλοιπα. Έπεται τότε πως οι προβλέψεις του μοντέλου θα είναι καλύτερες, δηλαδή πιο κοντά στην πραγματικότητα. Παραθέτουμε ένα παράδειγμα πρόβλεψης του μοντέλου:

Έστω πελάτης με εισόδημα αιτούντος=10000, εισόδημα συναιτούντος=5000. Η πρόβλεψη για το ποσό του δανείου θα είναι:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$$

$$\Rightarrow$$

Συντελεστής προσδιορισμού r=0.337

Loan Amount = 54308.446 +14.017 x Applicant Income +14.628 x CoapplicantIncome

= 54308.446 +14.017 x 10000 + 14.628 x 5000

= 54308.446 + 140170 + 73140

= 267618.446

Προκύπτει από το μοντέλο πως με τα παραπάνω στοιχεία η εταιρία μπορεί να χορηγήσει στον συγκεκριμένο πελάτη δάνειο ύψους 267618.446 χρηματικών μονάδων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Van der Aalst, W. (2016) Data Science in Action. In: Process Mining. Springer, Berlin, Heidelberg.

Heilbron, J.L., Ed. The Oxford Companion to the History of Modern Science. OxfordUniversityPress, NewYork, 2003.

Dhar, V. (2013) Data science and prediction. Commun. ACM 56, 12, 64–73. DOI:https://doi.org/10.1145/2500499

Burgard, J., Münnich, R., & Zimmermann, T. (2014). The Impact of Sampling Designs on Small Area Estimates for Business Data, Journal of Official Statistics, 30(4), 749-771. doi: https://doi.org/10.2478/jos-2014-0046.

Hidiroglou, M.A. &Lavallée, P. Sampling and Estimation in Business Surveys, Chapter in Handbook of Statistics, Elsevier, Part A, 29, 441-470, 2009.

Smith, P. (2013). Sampling and Estimation for Business Surveys. In Designing and Conducting Business Surveys (eds G. Snijkers, G. Haraldsen, J. Jones and D.K. Willimack). doi:10.1002/9781118447895.ch05 2013 The Clute Institute Copyright by author(s) ative Commons License CC-BY 53

Lee, P.M. (2013). Use Of Data Mining In Business Analytics To Support Business Competitiveness. Review of Business Information Systems, 17(2), 53-58.

Gan, M., & Dai, H. (2014). Data Mining for Business Analytics in Retail. In Wang, J. (Ed.), Encyclopedia of Business Analytics and Optimization (pp. 618-627). IGI Global. http://doi:10.4018/978-1-4666-5202-6.ch057

Shet, A.R.(2015). Information Technology in Retail Sector International Journal of Scientific Engineering and Research (IJSER), 4(5),43-46. VisualisingBusinessData: A Survey

Roberts, .r.c.&Laramee, R.S. (2018). Visual and Interactive Computing Group, Information, 9(11), 285; https://doi.org/10.3390/info9110285.

Friedman, V. (2008). Data Visualization & Infographics, Graphics, Monday Inspiration.

Homocianu, D. (2010). Data Visualization in Business Intelligence. The (ISI) Proceeding of WSEAS MCBEC2010-Recent Advances in Mathematics and Computers in Business, Economics, Biology & chemistry. [ISSN:1790-2769][ISBN:978-960-474-194-6]. pp.164-167.

Abraham, B. (2007). Implementation of Statistics in Business and Industry. RevistaColombiana de Estadística, 30(1), 1-11.

Snee, R. (2015). Industry and Business, Statistics Chapter In : Wiley. StatisticsReferenceOnline, pp.1-8. 10.1002/9781118445112.stat00076.pub2.

Klein, J.L. & Dave, K. (1997). Statistical Visions in Time: A History of Time Series Analysis, 1662-1938, Cambridge University Press.

A Brief History of Statistics (Selected Topics) ALPHA Seminar, August 29, 2017. Pdf. [online] Available at : https://www.google.com/url?sa=t&source=web&rct=j&url=http://homepage.divms.uiowa.edu/~dzimmer/alphaseminar/Statistics-history.pdf&ved=2ahUKEwiHnZ-Su_rqAhXSC-wKHU0lAlwQFjAbegQIAhAB&usg=AOvVaw1flwYg5Sb1-XqVKA8QgM08 accessed in 01 August, 2020.

Amit, S.A. (2015). Data Analysis in Business Research: Key Concepts. International Journal of Research in Management&BusinessStudies. 2. 50-55.

Γαλάνης (2014). Μονομεταβλητήανάλυσηεπιδημιολογικώνδεδομένων, ΑρχείαΕλληνικήςΙατρικής, 31(2) 221-243.

Ξακελάκη, Η. (2020). Περιγραφή αριθμητικών δεδομένων. Κεφάλαιο στο Εισαγωγή στη στατιστική σκέψη Τόμος Ι - Περιγραφική Στατιστική. ΕκδόσειςΜπένου, Αθήνα, σελ. 149-150.

Swinnen J., Depaire B., Jans M.J., Vanhoof K. (2012) A Process Deviation Analysis – A Case Study. In: Daniel F., Barkaoui K., Dustdar S. (eds) Business Process Management Workshops. BPM 2011. Lecture Notes in Business Information Processing, vol 99. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-28108-2_8

Roberts, H.V. (1990). Applications in Business and Economic Statistics: Some Personal Views, Statistical Science, 4(5), 372—390.

Kundu, S. (2016). Business statistics (course lesson 01). An Introduction to business statistics. Pdf. [Online] Available at : https://www.google.com/url?sa=t&source=web&rct=j&url=http://www.ddegjust.ac.in/studymaterial/mcom/mc-106.pdf&ved=2ahUKEwi-hd_j8_7qAhUGsaQKHeRsDr8QFjAAegQIARAB&usg=AOvVaw0oI_9-OD7egFd4P8yQqPSw accessed in 02 July, 2020.

Wood, M. (2010). The use of statistical methods in management research: a critique and some suggestions based on a case study. Pdf. [Online] Available at : http://userweb.port.ac.uk/~woodm/papers.htm accessed in 02 July, 2020.

Bamberger, P. (2008). From the editors. Beyond contextualization: using context theories to narrow the micro-macro gap in management research. Academy of Management Journal, 51(5), 839-846.

Ayres, I. (2007). Super crunchers: how anything can be predicted. London: JohnMurray.

Nayak, B. K., &Hazra, A. (2011). How to choose the right statistical test?.Indianjournal of ophthalmology, 59(2), 85–86. https://doi.org/10.4103/0301-4738.77005

https://ec.europa.eu › eurostat

European Commission (2017). Processing methods in business statistics (at national level). PDF. [Online] Available at : https://www.google.com/url?sa=t&source=web&rct=j&url=https://ec.europa.eu/eurostat/documents/54610/7779382/Processing-methods-in-business-statistics.pdf&ved=2ahUKEwitueyZ8__qAhUGM-wKHf0aD34QFjAAegQIAhAB&usg=AOvVaw0T7eM2adV_WbcqS7vZNinL accessed in 02 July, 2020.

Kaur SP. (2013). Variables in research. Indian J Res Rep Med Sci. 4:36–8.

Ali, Z., & Bhaskar, S. B. (2016). Basic statistical tools in research and data analysis. Indianjournal of anaesthesia, 60(9), 662–669. https://doi.org/10.4103/0019-5049.190623

Satake EB. (2015). Statistical Methods and Reasoning for the Clinical Sciences Evidence-Based Practice. Isted. SanDiego: PluralPublishing, Inc. pp. 1–19.

Manikandan S. (2011). Measures of central tendency: Median and mode. J PharmacolPharmacother. 2:214–5.
Myles PS, Gin T. (2000). Statistical Methods for Anaesthesia and Intensive Care. Isted. Oxford: ButterworthHeinemann. pp. 8–10.
Nickerson RS. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. PsycholMethods. 5:241–301.
Bajwa SJ. (2015). Basics, common errors and essentials of statistical tools and techniques in anesthesiology research. J AnaesthesiolClinPharmacol. 31:547–53.
Divisi D, Di Leonardo G, Zaccagna G, Crisci R. (2017). Basic statistics with Microsoft Excel: a review. J ThoracDis. 9(6):1734-1740. doi:10.21037/jtd.2017.05.81
Landau, S. &Everitt, S.B. (2004). A handbook of statistical analyses using SPSS. Chapman&Hall/CRC, BocaRaton.
Griffith, A. (2010). SPSS For Dummies®, 2nd Edition, Publisher: For Dummies ISBN: 978-0-470-48764-8
Kolker, E. Stewart, E., Ozdemir, V. (2012). Opportunities and Challenges for the Life Sciences Community. Journal of integrativebiology. 12:140-147.
Aiken, L.S., West, S.G., Pitts, S.C., Baraldi, A.N. and Wurpts, I.C. (2012). Multiple Linear Regression. In Handbook of Psychology, Second Edition (eds I. Weiner, J.A. Schinka and W.F. Velicer). https://doi.org/10.1002/9781118133880.hop202018
Uyanık, K.&Güler, N. A Study on Multiple Linear Regression Analysis, Procedia - Social and Behavioral Sciences, 2013, 106, 234-240.
Marill, KA (2004), Advanced Statistics: Linear Regression, Part II: Multiple Linear Regression. Academic Emergency Medicine, 11: 94-102. https://doi.org/10.1197/j.aem.2003.09.006
Zou, K.H., Tuncali, K. & Silverman, S.G. Correlation and Simple Linear Regression. Radiology 2003 227:3, 617-628.
Altman, N., Krzywinski, M. Simple linear regression. Nat Methods 12, 999–1000 (2015). https://doi.org/10.1038/nmeth.3627

Schneider A, Hommel G, Blettner M. Linear regression analysis: part 14 of a series on evaluation of scientific publications. DtschArzteblInt. 2010;107(44):776-782. doi:10.3238/arztebl.2010.0776

Sawyer, S.F. Analysis of Variance: The Fundamental Concepts. The Journal of Manual & Manipulative Therapy, 2017, 17(2), 28-38.

Bower, K.M. Analysis of Variance (ANOVA) Using Minitab. SemanticScholar 2007, 2(3),1-6.

Sthle, L. &Wold, S. Analysis of variance (ANOVA). Chemometrics and Intelligent Laboratory Systems, 1989, 6(4), 259-272.

Παπαγεωργίου, Ε. Ανάλυση Διακύμανσης με ένα Παράγοντα (OneWayANOVA). Κεφάλαιο στο Βιοστατιστική και Εφαρμογές, Εκδόσεις Πολιτεία Αθήνα, 2017, σελ. 282-284.

Χαλικιάς, Μ., Μανωλέσου, Α. &Λάλου, Π. (2015). Μεθοδολογία Έρευνας και Εισαγωγή στη Στατιστική Ανάλυση Δεδομένων με το IBMSPSSSTATISTICS. Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, Αθήνα, σελ. 30-31.

Νικήτας, Π.Ι. (2013). Εισαγωγή στη στατιστική ανάλυση πειραματικών δεδομένων με χρήση EXCEL και SPSS. Εκδόσεις ΣΙΜΩΝΗ, Θεσσαλονίκη, σελ. 15.

Petrie, A. &Sabin, C. (2008). Ιατρική Στατιστική με μια ματιά. Εκδόσεις Παρισιάνου, Αθήνα, σελ. 16-17.

Κουγιουμτζής, Δ. (2014). Στατιστική για Πολιτικούς Μηχανικούς. Σημειώσεις μέρους Β'. Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, σελ. 14-16.

Πανάρετος, Ι. &Ξεκαλάκη, Ε. (2003). Εισαγωγή στη Στατιστική Σκέψη Τόμος ΙΙ. Εκδόσεις Μπένου Ε., σελ. 301-310.

Παπαδόπουλος, Γ. (2015). Εισαγωγή στις πιθανότητες και τη στατιστική. Εκδόσεις Ιανός, Αθήνα, σελ. 100.

Δριτσάκη, Χ. (2015). Οικονομετρία Ι. Έκδοση: 1.0. ΤΕΙ Δυτικής Μακεδονίας, Κοζάνη Διαθέσιμο από τη δικτυακή διεύθυνση:link

Πετρίδης, Δ. (2015). Πολλαπλή γραμμική παλινδρόμηση & συσχέτιση. Κεφάλαιο Συγγράμματος στο Ανάλυση πολυμεταβλητών τεχνικών. [ηλεκτρ.

βιβλ.] Αθήνα:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. κεφ 2. Διαθέσιμο στο: http://hdl.handle.net/11419/2127

BoutsikasM.V. (2004).Σημειώσεις μαθήματος «Στατιστικά Προγράμματα» Τμήμα Στατ. &Ασφ. Επιστήμης, Πανεπιστήμιο Πειραιώς. Διαθέσιμο από τη δικτυακή διεύθυνση:link

Διαφέρμος Β.(2011) , «Κοινωνική Στατιστική & Μεθοδολογία Έρευνας με το SPSS», Εκδόσεις ΖΗΤΗ,Θεσσαλονίκη

Κιόχος Π.,. (2015). «Στατιστική για τις επιχειρήσεις και την οικονομία». Αθήνα.

Σωσσίδου Ε. , Ψευτογιάννη Δ. «Μεθοδολογία Έρευνας & Στατιστική με την χρήση του SPSS 13.00 forWindows», Θεσσαλονίκη,2007.