ΚΕΦΑΛΑΙΟ 1ο

Βασικές Τεχνικές Δειγματοληψίας και Ανάλυση Ερωτηματολογίων με χρήση του ελέγχου (t)

(Basic Sampling Techniques and
Questionnaire Analysis using (t) test)

1. Εισαγωγή

Η Δειγματοληψία είναι ένα σημαντικό τμήμα της Στατιστικής Συμπερασματολογίας

Στο κεφάλαιο αυτό θα υπάρξει μία, περισσότερο λεπτομερής, εξέταση της Δειγματοληψίας.

Ειδικότερα, μετά τη μελέτη του κεφαλαίου αυτού, ο αναγνώστης:

1. Θα μπορεί να αντιλαμβάνεται και να εκτιμά τις ιδέες και διαδικασίες που χρησιμοποιούνται στη δειγματοληψία.
2. Θα μπορεί να εκλέγει τα ακόλουθα είδη δειγμάτων: (α) στρωματοποιημένο δείγμα, (β) δείγμα κατά συστάδες, (γ) συστηματικό δείγμα.
3. Θα μπορεί να εκτιμήσει Μέσους και Σύνολα από τα δεδομένα που προέρχονται από τη στρωματοποιημένη, κατά συστάδες και συστηματική δειγματοληψία..

1.1. Ορισμοί

Η ανάγκη της δειγματοληψίας γεννήθηκε από την δυσκολία, από πλευράς κόστους και μέσων, της εξέτασης όλων των μονάδων ενός πληθυσμών ως προς ένα ιδιαίτερο χαρακτηριστικό τους, π.χ. η εύρεση της μέσης ηλικίας όλων των ενηλίκων που ζουν σε ένα τμήμα μιας πόλης.

Ορισμός 1.1.1: **Πληθυσμός (population)** ή Στατιστικός Πληθυσμός είναι το σύνολο των προσώπων ή πραγμάτων που παρατηρείται ως προς ένα ή περισσότερα χαρακτηριστικά του γνωρίσματα από ένα ερευνητή.

Ορισμός 1.1.2: **Πληθυσμιακή μονάδα (population unit or entity**) είναι η μονάδα του πληθυσμού που παρατηρείται ως προς ένα ή περισσότερο χαρακτηριστικά της γνωρίσματα, από έναν στατιστικό ή έναν ερευνητή.

Ορισμός 1.1.3: **Μεταβλητή (variable**) είναι ένα χαρακτηριστικό που λαμβάνει διαφορετικές τιμές για διαφορετικές πληθυσμιακές μονάδες.

Ορισμός 1.1.4: **Τυχαία μεταβλητή (random variable**) είναι μία μεταβλητή για την οποία μπορεί να προσδιορισθεί μία μαθηματική έκφραση, που **λέγεται συνάρτηση πιθανότητας, (probability function),** η οποία έχει ορισμένες ιδιότητες και δίνει την σχετική συχνότητα με την οποία εμφανίζονται οι τιμές της μεταβλητής. Οι τυχαίες μεταβλητές συμβολίζονται με κεφαλαία γράμματα, π.χ. X, Y, και οι τιμές τους με μικρά γράμματα, π.χ. x, y.

Ορισμός 1.1.5: **Ποσοτική Μεταβλητή (Quantitative Variable**) είναι μία μεταβλητή της οποίας οι τιμές δίνονται ως αριθμητικές ποσότητες, όπως μετρήσεις ή πλήθη. Παραδείγματα ποσοτικών μεταβλητών είναι το ύψος φοιτητών ή ο αριθμός των πελατών ενός καταστήματος.

Ορισμός 1.1.6: **Ποιοτική μεταβλητή (Qualitative Variable)** είναι η μεταβλητή της οποίας οι τιμές δεν είναι μετρήσιμες ποσότητες. Παραδείγματα ποιητικών μεταβλητών είναι οι ιδιότητες προϊόντων, π.χ. ελαττωματικό προϊόν ή μη ελαττωματικό προϊόν, το χρώμα των ματιών π.χ. καστανό, μπλε ή πράσινο ή η συναισθηματική κατάσταση ενός προσώπου, π.χ. χαρά, λύπη, κ.τ.λ.

Ορισμός 1.1.7: **Διακριτή μεταβλητή (Discrete Variable)** είναι μία μεταβλητή που μπορεί να λάβει μόνο ορισμένες τιμές σε ένα διάστημα, π.χ. ο αριθμός των πελατών ενός καταστήματος σε μία δεδομένη ημέρα, ή ο αριθμός των παιδιών μιας οικογένειας.

Ορισμός 1.1.8: **Συνεχής μεταβλητή (Continuous Variable)** είναι μία μεταβλητή που μπορεί να λάβει όλες τις τιμές μέσα σε ένα διάστημα. Το ύψος, το βάρος, το πλάτος, ο όγκος, είναι συνεχείς μεταβλητές γιατί μπορούν να λάβουν όλες τις τιμές στα διαστήματα στα οποία αναφέρονται, π.χ. το ύψος ενός ενήλικα ανθρώπου μπορεί να λάβει όλες τις τιμές μεταξύ 1,40 m και 2,10 m.

Ορισμός 1.1.9: **Δείγμα (sample)** είναι ένα μέρος του στατιστικού πληθυσμού. Αν, π.χ. ο πληθυσμός ορίζεται ως το σύνολο των υψών όλων των

φοιτητών στην Ελλάδα, ένα δείγμα από το πληθυσμό αυτόν είναι τα ύψη των φοιτητών στο Τ.Ε.Ι. του Πειραιά.

Ορισμός 1.1.10: **Τυχαίο Δείγμα (Random Sample)** είναι ένα δείγμα το οποίο έχει εκλεγεί με τέτοιο τρόπο ώστε τα αποτελέσματα της ανάλυσής του να μπορούν να χρησιμοποιηθούν για να εξαγάγουν στατιστικά συμπεράσματα για τον πληθυσμό από τον οποίο το δείγμα έχει εκλεγεί. Ένα τυχαίο δείγμα έχει εκλεγεί με τη χρήση τυχαίων αριθμών ή με έναν άλλο τελείως τυχαίο μηχανισμό. Αν λάβουμε, π.χ. τους 5 πρώτους φοιτητές μιας τάξης στα Μαθηματικά, η εκλογή αυτή δεν αποτελεί ένα τυχαίο δείγμα από το σύνολο των φοιτητών της τάξης αυτής. Ας υποθέσουμε, όμως ότι αριθμούμε τους 50 φοιτητές μιας τάξης από 1 ως 50 και αναγράφουμε τον κάθε αριθμό σε ένα φύλλο χάρτου, τα φύλλα αυτά τα τοποθετούμε σε ομοιόμορφους φακέλους, τους οποίους κλείνουμε και τους τοποθετούμε σε ένα κλειστό κουτί, το οποίο μετακινούμε για να αναμιχθούν οι φάκελοι. Κατόπιν εκλέγουμε τυχαία, χωρίς επανατοποθέτηση, ένα δείγμα 5 φακέλων, τους οποίους ανοίγουμε και αναγράφουμε τους αριθμούς των φύλλων χάρτου που περιέχουν. Οι 5 αριθμοί τους οποίους βρίσκουμε, με τον τρόπο αυτό, αποτελούν ένα απλό τυχαίο δείγμα μεγέθους 5 από τον πληθυσμό των 50 φοιτητών της τάξης.

1.2. Απλή Τυχαία Δειγματοληψία

Δίνουμε τους παρακάτω ορισμούς:

Ορισμός 1.2.1: **Στατιστική Συμπερασματολογία (Statistical Inference)** είναι η διαδικασία με την οποία βρίσκεται ένα συμπέρασμα για ένα πληθυσμό, με βάση πληροφορίες που προέρχονται από ένα δείγμα που εκλέχτηκε τυχαία από τον πληθυσμό αυτό.

Ορισμός 1.2.2: **Δείγμα πιθανότητας (Probability sample)** είναι ένα δείγμα που εκλέχτηκε από ένα στατιστικό πληθυσμό κατά τέτοιο τρόπο ώστε κάθε στοιχείο του πληθυσμού έχει μία γνωστή και διάφορη του μηδενός πιθανότητα να εκλεγεί. Το δείγμα 5 φοιτητών που εκλέγεται στο προηγούμενο υποκεφάλαιο 1.1 (Ορισμός 1.1.10) είναι ένα δείγμα πιθανότητας όπου κάθε στοιχείο έχει την πιθανότητα $1/50$ να εκλεγεί.

Συνήθως γίνεται δειγματοληψία από μεγάλους πληθυσμούς από τους οποίους εκλέγονται δείγματα πιθανοτήτων. Ένα δείγμα πιθανότητας είναι το απλό τυχαίο δείγμα.

Ορισμός 1.2.3: **Απλό Τυχαίο Δείγμα (simple random sample).** Αν ένα δείγμα μεγέθους n εκλέγεται από ένα πληθυσμό μεγέθους N κατά τέτοιο

τρόπο ώστε κάθε πιθανό δείγμα μεγέθους n να έχει την ίδια πιθανότητα να εκλεγεί, τότε το δείγμα λέγεται **απλό τυχαίο δείγμα (simple random sample (srs)).**

Συνήθως εκλέγονται απλά τυχαία δείγματα, χωρίς επανατοποθέτηση (without replacement).

Η εκλογή απλών τυχαίων δειγμάτων, αφού αριθμηθούν τα στοιχεία του πληθυσμού, γίνεται με τη βοήθεια τυχαίων αριθμών (οι οποίοι περιέχονται στο Πίνακα Α, στο τέλος του βιβλίου). **Μία εναλλακτική μέθοδος είναι να χρησιμοποιηθεί ένα από τα ακόλουθα πακέτα λογισμικού, τα οποία θα χρησιμοποιηθούν αρκετά στο βιβλίο αυτό: SPSS, EXCEL, MINITAB, SAS, και XLSTAT.**

Ορισμός 1.2.4: **Παράμετρος (Parameter)** Κάθε περιγραφικό μέτρο ενός πληθυσμού ονομάζεται παράμετρος. Παραδείγματα παραμέτρων είναι ο μέσος μ, διασπορά σ^2, η τυπική απόκλιση σ και η αναλογία (ή ποσοστό) π. Συνήθως οι παράμετροι συμβολίζονται με μικρά ελληνικά γράμματα.

Ορισμός 1.2.5: **Στατιστικό (Statistic)** είναι κάθε περιγραφικό μέτρο το οποίο υπολογίζεται από τα στοιχεία ενός δείγματος και χρησιμοποιείται για να αποτυπώσει περιεκτικά τις ιδιότητες ενός δείγματος.

Ορισμός 1.2.6: **Εκτιμητής σε σημείο (Point Estimate)** είναι μία συνάρτηση (στατιστικό,statistic) των δεδομένων του δείγματος. Η τιμή της συνάρτησης αυτής χρησιμοποιείται σαν η "καλύτερη εικασία" για την τιμή της αντιστοίχου πληθυσμιακής παραμέτρου την οποία ο εκτιμητής εκτιμά. Ο εκτιμητής έχει ορισμένες ιδιότητες όπως η Αμεροληψία (unbiasedness), Συνέπεια (consistency), Αποτελεσματικότητα (efficiency) και Επάρκεια (sufficiency). Περισσότερα για τους εκτιμητές ο αναγνώστης μπορεί να διαβάσει στο κεφάλαιο 12: ΕΚΤΙΜΗΤΙΚΗ, του βιβλίου: Στατιστική Επιχειρήσεων, Χρήστος Κ. Φράγκος, Εκδόσεις Σταμούλη, 1998). Στο ίδιο βιβλίο, στο κεφάλαιο 11: ΔΕΙΓΜΑΤΙΚΕΣ ΚΑΤΑΝΟΜΕΣ, περιλαμβάνεται η θεωρία και οι εφαρμογές της απλής τυχαίας δειγματοληψίας όπως και η περιγραφή των κατανομών (χ^2), (F), (t) και Κανονικής Κατανομής. Ένας βασικός σκοπός της δειγματοληψίας είναι να παράγει εκτιμητές των παραμέτρων του πληθυσμού που εξετάζεται. Η "καλή ποιότητα" ενός δείγματος εξαρτάται από το πόσο καλοί εκτιμητές των παραμέτρων του πληθυσμού παράγονται από αυτό το δείγμα.

Ορισμός 1.2.7: Η διαφορά ενός εκτιμητού ($\hat{\theta}$) και της πραγματικής τιμής της παραμέτρου (θ) που εκτιμά $\hat{\theta}$-θ καλείται **<u>δειγματοληπτικό σφάλμα (sampling error) του εκτιμητή.</u>**

Εάν υπάρχει ένα δείγμα πιθανότητας είναι δυνατόν να παραχθεί ένας εκτιμητής του δείγματοληπτικού σφάλματος.

1.3 Εφαρμογές της Δειγματοληψίας

Οι τεχνικές δειγματοληψίας έχουν σημαντικές εφαρμογές στις ακόλουθες επιστημονικές, κοινωνικές και οικονομικές περιοχές:

α. Δημοσκοπήσεις κοινής γνώμης για το αποτέλεσμα των εκλογών πριν τις εκλογές (exit polls).

β. Έρευνα αγοράς, για να προσδιορισθούν οι προτιμήσεις των καταναλωτών για ορισμένα προϊόντα,

γ. Διαδικασίες ελέγχου ποιότητας για κατασκευαστικές τεχνικές.

δ. Λογιστική, φορολόγηση και έλεγχος επιχειρήσεων.

ε. Προβλέψεις παραγωγής σιταριού και άλλων γεωργικών προϊόντων.

ζ. Προσδιορισμός του ρυθμού εμφάνισης και του ρυθμού επικράτησης ορισμένων ασθενειών σε μία ορισμένη γεωργική περιοχή.

η. Έρευνα σχετική με πολλά κοινωνικά και οικονομικά προβλήματα

θ. Προσδιορισμός πληθυσμιακών χαρακτήρων, όπως η εργασιακή κατάσταση, το οικονομικό εισόδημα και ο βαθμός εκπαίδευσης των πολιτών.

1.4 Επιπλέον Ορισμοί

Στο κεφάλαιο 11: ΔΕΙΓΜΑΤΙΚΕΣ ΚΑΤΑΝΟΜΕΣ, του βιβλίου: ΣΤΑΤΙΣΤΙΚΗ ΕΠΙΧΕΙΡΗΣΕΩΝ, Χρήστος Φράγκος, Εκδόσεις Σταμούλη, 1998, υπάρχει η θεωρία της Απλής Τυχαίας Δειγματοληψίας. Στο υποκεφάλαιο 1.2 του παρόντος βιβλίου επεκτείνονται οι ορισμοί και η ορολογία η οποία χρησιμοποιείται στη βασική θεωρία δειγματοληψίας.

Ειδικότερα, υπάρχουν τα ακόλουθα στοιχεία σε μία δειγματοληψία: α) Πρακτικά, χρησιμοποιούνται τυχαία δείγματα χωρίς επανατοποθέτηση, σχεδόν αποκλειστικά. β) Πρέπει να προσδιορισθούν χωριστά οι μονάδες στις οποίες εκτελείται η δειγματοληψία και οι μονάδες στις οποίες λαμβάνονται οι παρατηρήσεις. γ) Πρέπει να ληφθούν υπ' όψη πολλά δειγματοληπτικά πειράματα.

Δίνονται οι ακόλουθοι ορισμοί εννοιών που χρησιμοποιούνται στις δειγματοληπτικές μεθόδους.

Ορισμός 1.4.1: **Μονάδα (observational unit)** είναι η πληθυσμιακή μονάδα (ορισμός 1.1.2). Στις μονάδες παρατήρησης λαμβάνονται οι μετρήσεις της δειγματοληψίας ή γίνεται κάποια διαδικασία κατάταξης σε κατηγορίες.

Ορισμός 1.4.2: **Δειγματοληπτική μονάδα (sampling unit)** είναι η μονάδα στην οποία βασίζεται η διαδικασία της δειγματοληψίας. Η δειγματοληπτική μονάδα μπορεί να είναι μια μονάδα παρατήρησης ή μία ομάδα (cluster) μονάδων παρατήρησης.

Ορισμός 1.4.3: **Γενικό σύνολο διερεύνησης (universe of investigation),** είναι το σύνολο των στοιχείων του οποίου τα χαρακτηριστικά πρέπει να εξετασθούν (π.χ. σύνολο φοιτητών μιας Ανώτατης Σχολής).

Ορισμός 1.4.4: **Πληθυσμός (Population)** είναι το σύνολο των χαρακτηριστικών του Γενικού συνόλου διερεύνησης. Αν το Γενικό σύνολο διερεύνησης είναι το σύνολο των φοιτητών μιας Ανώτατης σχολής, τα σύνολα των ηλικιών και των βαρών των φοιτητών θα είναι δύο διαφορετικοί πληθυσμοί, αντίστοιχα, για το Γενικό σύνολο διερεύνησης.

Ορισμός 1.4.5: **Δειγματοληπτικό πλαίσιο (sampling frame)**. Το δειγματοληπτικό πλαίσιο είναι ένας κατάλογος ή άλλη παρουσίαση των δειγματοληπτικών μονάδων (μονάδων παρατηρήσεις ή τμημάτων). Ένα Γενικό σύνολο διερεύνησης μπορεί να περιέχει μερικά δειγματοληπτικά πλαίσια.

Ορισμός 1.4.6: **Sampling fraction (δειγματοληπτικό κλάσμα)** είναι το ποσοστό τού δειγματοληπτικού πλαισίου που περιέχεται στο δείγμα. Ομάδα (cluster) είναι ένα σύνολο διερεύνησης που μπορεί να θεωρηθεί σαν μία μοναδική δειγματοληπτική μονάδα.

Ορισμός 1.4.7: **Δειγματοληψία κατά ομάδες (cluster sampling)**, είναι μία τεχνική δειγματοληψίας που χρησιμοποιεί σύνολα ή τμήματα στοιχείων του Γενικού συνόλου διερεύνησης σαν δειγματοληπτικές μονάδες.

Ορισμός 1.4.8: **Στρωματοποιημένη δειγματοληψία (stratified sampling)** είναι η τεχνική δειγματοληψίας κατά την οποία διαιρείται όλο το Γενικό σύνολο διερεύνησης σε στρώματα (strata) και εκλέγεται ένα ανεξάρτητο δείγμα από κάθε στρώμα.

Ορισμός 1.4.9: **Μέγεθος δείγματος (size of sample)** είναι ο αριθμός των δειγματοληπτικών μονάδων που περιλαμβάνεται στη μελέτη.

Ορισμός 1.4.10: **Κενό (gap)** είναι η διαφορά μεταξύ των στοιχείων που προέρχονται στο Γενικό σύνολο διερεύνησης και των στοιχείων που προέρχονται στο δειγματοληπτικό πλαίσιο.

1.5 Τα Στάδια της Δειγματοληψίας

Τα κύρια βήματα μιας δειγματοληψίας είναι τα ακόλουθα:

α. **Δήλωση των σκοπών της δειγματοληψίας.** Στο στάδιο αυτό δηλώνονται οι ακριβείς ερευνητικές ερωτήσεις που πρέπει να απαντηθούν και οι ακριβείς Υποθέσεις Έρευνας που πρέπει να ελεγχθούν με βάση τά δεδομένα της δειγματοληψίας.

β. **Γενικές σκέψεις πάνω στο σχέδιο της δειγματοληψίας.** Στο στάδιο αυτό πρέπει να απαντηθούν οι εξής ερωτήσεις: Οι πληροφορίες που ζητούνται υπάρχουν ήδη εντός ή εκτός της Εταιρείας δειγματοληψιών; Πως θα γίνουν οι μετρήσεις; Είναι εφικτή η λήψη παρατηρήσεων; Ποιά είναι η ακρίβεια που απαιτείται; Ποιά είναι η χρηματική ενίσχυση που δίνεται για την διενέργεια της δειγματοληψίας;

γ. **Σχέδιο δειγματοληψίας.** Στο στάδιο αυτό θεωρούνται τα εξής: Εκλογή δειγματοληπτικής μονάδας, δυνατότητα στρωματοποίησης, μέγεθος δείγματος, μέθοδοι αντιμετώπισης του φαινομένου της δυσκολίας λήψης απαντήσεων και το κόστος κάθε εργασίας.

δ. **Εκπόνηση των ερωτηματολογίων.** Η ακρίβεια και χρησιμότητα των αποτελεσμάτων εξαρτάται από την ακρίβεια των δεδομένων. Συνεπώς, είναι εξαιρετικά σπουδαία η εκπόνηση των ερωτηματολογίων και οι μέθοδοι εκπαίδευσης και ελέγχου των ερευνητών που λαμβάνουν τις συνεντεύξεις.

ε. **Σύνοψη και Ανάλυση.** Πρέπει να προσδιορίζεται το είδος και να εκτιμάται το κόστος της ανάλυσης όταν σχεδιάζεται η δειγματοληψία, όχι μετά την συλλογή των δεδομένων.

Στο κεφάλαιο αυτό παρουσιάζεται τρείς μέθοδοι δειγματοληψίας: Στρωματοποιημένη τυχαία δειγματοληψίας, δειγματοληψία κατά Ομάδες και Συστηματική δειγματοληψία. Το κεφάλαιο της απλής τυχαίας δειγματοληψίας παρουσιάζεται στο κεφάλαιο 11 με τίτλο: ΔΕΙΓΜΑΤΙΚΕΣ ΚΑΤΑΝΟΜΕΣ, του βιβλίου "Στατιστική Επιχειρήσεων", Χρήστος Φράγκος, Εκδόσεις Σταμούλης, 1998.

1.6 Στρωματοποιημένη Τυχαία Δειγματοληψία

Στην απλή τυχαία δειγματοληψία η μονάδα παρατήρησης είναι η δειγματοληπτική μονάδα και το δειγματικό πλαίσιο είναι ο πληθυσμός.

Ορισμός 1.6.1: **Στρωματοποιημένη δειγματοληψία (stratified sampling)** είναι εκείνη η δειγματοληψία στην οποία η μονάδα παρατήρησης είναι η δειγματοληπτική μονάδα. Ο πληθυσμός που εξετάζεται υποδιαιρείται σε υποπληθυσμούς (στρώματα) τα οποία βασίζονται σε μια γνωστή μεταβλητή που είναι συνδεδεμένη με την μέτρηση που λαμβάνεται από κάθε μονάδα παρατήρησης.

Η στρωματοποίηση είναι παρόμοια διαδικασία με την τμηματοποίηση (blocking) στη σχεδίαση πειραμάτων.

Ας υποτεθεί ότι κάποιος ενδιαφέρεται να προσδιορίσει το μέσο εισόδημα ενός πληθυσμού των κατοίκων μίας πόλης. Επειδή το εισόδημα είναι συνδεδεμένο με την εργασία, λαμβάνεται ένα δείγμα που περιλαμβάνει μερικούς ανειδίκευτους εργάτες, μερικούς εμπόρους και μερικούς επαγγελματίες επιστήμονες, βρίσκεται το μέσο εισόδημα της κάθε επαγγελματικής ομάδας και μετά ο μέσος των τριών μέσων εισοδημάτων. Η αποτελεσματική στρωματοποίηση παράγει εκτιμητές παραμέτρων που έχουν μικρότερες διασπορές από τους εκτιμητές που απορρέουν από την απλή τυχαία δειγματοληψία.

Δύο άλλα πλεονεκτήματα της στρωματοποιημένης δειγματοληψίας είναι η μείωση του κόστους και η διοικητική ευκολία ελέγχου και αποπεράτωσης της δειγματοληψίας.

Η θεμελιώδης ιδέα που υπάρχει στην στρωματοποιημένη δειγματοληψία είναι η ικανότητα να βελτιώνεται η αποτελεσματικότητα της δειγματοληψίας με τη χρήση μιας γνωστής μεταβλητής που συνδέεται με τη μεταβλητή, οποία μετρείται σε κάθε μονάδα παρατήρησης.

1.7 Εκτίμηση του Συνόλου του Πληθυσμού στη Στρωματοποιημένη Δειγματοληψία

Εκτιμούμε το σύνολο του πληθυσμού T με την εύρεση ενός χωριστού εκτιμητή T_h για κάθε στρώμα

Ισχύει:

$$\hat{T}_{st} = \hat{T}_1 + \hat{T}_2 + \hat{T}_3 + \ldots \hat{T}_h + \ldots \hat{T}_L \,, \qquad \textbf{(1.1)}$$

όπου

$\hat{T}_{st}$ = εκτιμητής συνόλου πληθυσμού με στρωματοποιημένη δειγματοληψία,

L = αριθμός στρωμάτων,

$\hat{T}_1 = N_1 \cdot \overline{x_1} =$ μέσος του δείγματος που λαμβάνεται από το στρώμα 1 επί τον αριθμό των μονάδων του στρώματος 1.

Ισχύει:

$\overline{x_1} = \frac{\sum_{i=1}^{n} x_i}{n_1}$, όπου n_1 είναι το μέγεθος του δείγματος που λαμβάνονται από το στρώμα 1,

$\hat{T}_2 = N_2 \cdot \overline{x_2}$, όπου N_2: αριθμός μονάδων στο στρώμα 2,

$\overline{x_2}$: μέσος δείγματος που λαμβάνονται από το στρώμα 2,

$\hat{T}_h = N_h \cdot \overline{x_h}$, όπου N_h: αριθμός μονάδων στο στρώμα *h*.

$\overline{x_h}$: μέσος δείγματος που εκλέγεται από το στρώμα *h*,

$\hat{T}_L = N_L \cdot \overline{x_L}$, όπου N_L: αριθμός μονάδων στο στρώμα *L*

$\overline{x_L}$: μέσος δείγματος που εκλέγεται από το στρώμα *L*.

Άρα ισχύει:

$$\hat{T}_{st} = N_1 \cdot \overline{x_1} + N_2 \cdot \overline{x_2} + ... + N_h \cdot \overline{x_h} + ... + N_L \cdot \overline{x_L} = \sum_{h=1}^{L} N_h \cdot \overline{x_h} \qquad \textbf{(1.2)}$$

Η διασπορά (διακύμανση) του εκτιμητή $\hat{T}_{st}$ είναι:

$$V(\hat{T}_{st}) = \sum_{h=1}^{L} N_h^2 \cdot \frac{(N_h - n_h)}{N_h} \cdot \frac{S_h^2}{n_h} = \sum_{h=1}^{L} N_h (N_h - n_h) \cdot \frac{S_h^2}{n_h} \qquad \textbf{(1.3)}$$

όπου N_h = αριθμός μονάδων στο στρώμα h,

n_h = αριθμός μονάδων στο δείγμα που εκλέγεται από το στρώμα h.

Ισχύει:

$$S_h^2 = \frac{1}{(N_h - 1)} \cdot \sum_{i=1}^{N_h} (X_{hi} - \mu_h)^2 \qquad \textbf{(1.4)}$$

όπου x_{hi} είναι *η (i)* παρατήρηση του δείγματος που εκλέγεται από το *(h)* στρώμα και

μ_h είναι ο μέσος των παρατηρήσεων του δείγματος που εκλέγεται από το *(h)* στρώμα.

Ο όρος $\frac{(N_h - n_n)}{N_h} = 1 - \frac{n_h}{N_h}$ λέγεται διόρθωση πεπερασμένου πληθυσμού (fiuite population correction, f.p.c.).

Αν ισχύει: $\frac{n_h}{N_h} \leq 0.05$, μπορεί να παραληφθεί η f.p.c. **(1.5)**

Ισχύει ότι: η διασπορά του εκτιμητή $\hat{T}_h$ είναι: $V(\hat{T}_h) = N_h(N_h - n_h) \cdot \frac{S_h^2}{n_h}$ **(1.6)**

$$V(\hat{T}_{st}) = \sum_{h=1}^{L} N_h(N_n - n_h) \cdot \frac{S_h^2}{n_h} \quad \textbf{(1.7)}$$

1.8 Εκτίμηση του Πληθυσμιακού Μέσου στη Στρωματοποιημένη Δειγματοληψία

Ισχύει:

$$\overline{x_{st}} = \frac{\hat{T}_{st}}{M} \quad \textbf{(1.8)}$$

$$V(\overline{x_{st}}) = \frac{1}{N^2} V(\hat{T}_{st}) = \frac{1}{N^2} \cdot \sum_{h=1}^{L} N_h(N_h - n_h) \cdot \frac{S_h^2}{n_h} \quad \textbf{(1.9)}$$

Όπου $N = \sum_{h=1}^{L} N_h$

Επειδή είναι δύσκολο να βρεθεί η S_h^2 για κάθε στρώμα *h*, γιατί δεν είναι γνωστός ο μέσος μ_h, βρίσκονται εκτιμητές των διασπορών $V(\hat{T}_{st})$ και $V(\overline{x_{st}})$ (τύποι (1.7) και (1.9)) με την αντικατάσταση της S_h^2 με την s_h^2, όπου

$$s_h^2 = \frac{1}{(n_h - 1)} \cdot \sum_{i=1}^{n_h} (x_{hi} - \overline{x_h})^2 \quad \textbf{(1.10)}$$

ως εξής:

$$\hat{V}(\hat{T}_{st}) = \sum_{h-1}^{L} N_n - n_h) \cdot \frac{s_n^2}{n_h} \quad \textbf{(1.11)}$$

$$\hat{V}(\overline{x_{st}}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\cdot\frac{s_h^2}{n_h} \qquad \textbf{(1.12)}$$

1.9 Διαστήματα Εμπιστοσύνης για τις παραμέτρους (Τ) και (μ) στη Στρωματοποιημένη Δειγματοληψία

Μπορούμε να κατασκευάσουμε διαστήματα εμπιστοσύνης για το σύνολο του πληθυσμού Τ και το μέσο τον πληθυσμού μ, με την εφαρμογή του γενικού τύπου:

Εκτιμητής ± (παράγοντας αξιοπιστίας) x (τυπικό σφάλμα), **(1.13)**

με την υπόθεση ότι η θεωρία της Κανονικής Κατανομής ισχύει (για μεγάλα δείγματα, ≥ 30).

Τα διαστήματα εμπιστοσύνης με συντελεστή 100(1-α)% για το πληθυσμιακό μέσο και τον πληθυσμιακό σύνολο είναι, αντίστοιχα, τα εξής:

$$\overline{x_{st}} - Z_{\alpha/2}\cdot\sqrt{V(\overline{X_{st}})} \leq \mu \leq \overline{x_{st}} + Z_{\alpha/2}\cdot\sqrt{V(\overline{x_{st}})} \qquad \textbf{(1.14)}$$

$$N\overline{x_{st}} - Z_{\alpha/2}\cdot\sqrt{V(\hat{T}_{st})} \leq T \leq N\overline{x_{st}} + Z_{\alpha/2}\cdot\sqrt{V(\hat{T}_{st})} \qquad \textbf{(1.15)}$$

Όπου $Z_{\alpha/2}$ είναι το σημείο της Κανονικής Κατανομής όπου ισχύει:

$P(X > Z_{\alpha/2}) = \alpha / 2$.

Όταν οι πληθυσμιακές διασπορές είναι άγνωστες, χρησιμοποιούμε τους εκτιμητές που δίνονται από τους τύπους (1.11) και (1.12)

Παράδειγμα 1.9.1. Ο Διευθυντής προσωπικού της Δημόσιας Επιχείρησης Ενέργειας (Δ.Ε.Ε.) μιάς χώρας θέλει να εκτιμήσει τον μέσο όρο και το συνολικό αριθμό ημερών που οι υπάλληλοι ήταν απόντες κατά τη διάρκεια του προηγούμενου χρόνου.

Το δειγματοληπτικό πλαίσιο αποτελείται από ένα ντοσιέ με τις κάρτες των στοιχείων όλων των υπαλλήλων σε αλφαβητική σειρά. Επειδή ο Διευθυντής Προσωπικού παρατήρησε ότι κατά το παρελθόν οι ημέρες απουσίας ενός υπαλλήλου είναι συνδεδεμένες με τη διάρκεια του χρόνου κατά τον οποίο είναι στο υπαλληλικό προσωπικό της Δ.Ε.Ε., αποφασίζει να εφαρμοστεί μια στρωματοποιημένη δειγματοληψία και να εκλέξει ένα δείγμα υπαλλήλων στρωματοποιημένο στη βάση αυτή.

Τα τρία στρώματα είναι τα εξής: (1): υπηρεσία μικρότερη των 3 χρόνων, (2): υπηρεσία από 3 ως 9 χρόνια, και (3): υπηρεσία μεγαλύτερη των 9 χρόνων. Τα στρώματα αυτά κατασκευάζονται με την διαρρύθμιση των καρτών εργασίας σύμφωνα με

τον χρόνο υπηρεσίας των υπαλλήλων. Ο ακόλουθος πίνακας δείχνει τα αποτελέσματα:

Πίνακας 1.9.1 *Αποτελέσματα στρωματοποιημένης δειγματοληψίας παραδείγματος 1.9.1.*

Στρώμα	N_h	n_h	$\overline{x_h}$	s_h^2
Κάτω από 3 χρόνια	500	50	2	4
3 – 9 χρόνια	700	70	4	5
10 ή περισσότερα χρόνια	1000	100	5	6
Σύνολο	2200	220		

Από την (1.2) έχουμε

$$\hat{T}_{st} = 500(2) + 700(4) + 1000(5) = 8800$$

Από την (1.11) έχουμε

$$\hat{V}(\hat{T}_{st}) = \frac{(50)(500-50)(4)}{50} + \frac{(700)(700-70)(5)}{70} + (1000)\frac{(1000-100)(6)}{100} = 103500$$

Ο μέσος και ο εκτιμητής της διασποράς του είναι από τις εξισώσεις (1.8), (1.12), αντίστοιχα:

$$\overline{x_{st}} = \frac{8800}{2200} = 4 \text{ και } \hat{V}(\overline{X_{st}}) = \frac{103500}{(2200)^2} = 0{,}02$$

Από τις (1.14) και (1.15) έχουμε, για τα διαστήματα εμπιστοσύνης: α=0,05, 100(1-α)%, =0,5%, $Z_{\alpha/2} = 1{,}96$

$$4 - 1{,}96\sqrt{0{,}02} \le \mu \le 4 + 1{,}96 \cdot \sqrt{0{,}02} \text{ ή } 3{,}7 \le \mu \le 4{,}3$$

$$8800 - 1{,}96\sqrt{103500} \le T \le 8800 + 1{,}96 \cdot \sqrt{103500} \text{ ή } 8{,}169 \le T \le 9{,}431.$$

1.10 Δειγματοληψία κατά Ομάδες

Δίνουμε τον παρακάτω ορισμό:

Ορισμός 1.10.1 **Δειγματοληψία κατά Ομάδες (cluster sampling)** λέγεται η δειγματοληπτική μέθοδος κατά την οποία εκλέγεται ένα τυχαίο δείγμα ομάδων ή τμημάτων από το δειγματοληπτικό πλαίσιο και λαμβάνονται πληροφορίες για όλες τις στοιχειώδεις μονάδες παρατήρησης της κάθε ομάδας.

Τα ακόλουθα είναι παραδείγματα ομάδων και στοιχειωδών μονάδων παρατήρησης.

Ομάδα	**Στοιχειώδης μονάδα παρατήρησης**
Νοικοκυριό	Μέλος νοικοκυριού
Τάξη	Σπουδαστής
Ντουλάπα γραφείου	Ντοσιέ, φάκελος
Τιμολόγιο	Προϊόν

Αν, για κάθε ομάδα που εκλέγεται στο δείγμα, λαμβάνουμε πληροφορίες για όλα τα στοιχεία της ομάδας, η διαδικασία αυτή λέγεται απλή δειγματοληψία κατά ομάδες.

Αν, για κάθε ομάδα που εκλέγεται στο δείγμα, εκλέγουμε ένα δείγμα μονάδων παρατήρησης για να συμπεριληφθεί στο δείγμα μας, η διαδικασία αυτή λέγεται **δισταδιακή δειγματοληψία κατά ομάδες**

Η δειγματοληψία κατά ομάδες χρησιμοποιείται στις εξής περιπτώσεις:

α. Όταν δεν υπάρχει δειγματοληπτικό πλαίσιο για τις μονάδες παρατήρησης και δεν είναι δυνατόν να κατασκευασθεί ένα τέτοιο πλαίσιο.

β. Όταν χρειάζεται διοικητική ευκολία ελέγχου της δειγματοληψίας.

γ. Όταν χρειάζεται μείωση κόστους της δειγματοληψίας.

Στην δειγματοληψία κατά ομάδες, ο πληθυσμός αποτελείται από Μ ομάδες.

Η (i) ομάδα έχει N_i μονάδες παρατήρησης (άγνωστες)

Το μέγεθος του πληθυσμού είναι:

$$N = \sum_{i=1}^{M} N_i \text{ μονάδες παρατήρησης} \qquad \textbf{(1.16)}$$

$$T = \sum_{i=1}^{M} \sum_{j=1}^{N_i} x_{ij} \qquad \textbf{(1.17)}$$

$$\mu = \frac{\mathrm{T}}{\mathrm{N}} = \frac{1}{N} \cdot \sum_{i=1}^{M} \sum_{j=1}^{N_i} x_{ij} \qquad \textbf{(1.18)}$$

$$T_{i..} = \sum_{i=1}^{N_i} x_{ij} \qquad \textbf{(1.19)}$$

Το $T_{i.}$ είναι το σύνολο για την (i) ομάδα. Βρίσκουμε την τιμή του $\mathrm{T_{i.}}$, προσθέτοντας τις ομάδες όλων των μονάδων παρατήρησης στην (i) ομάδα.

Στην δειγματοληψία κατά ομάδες εκλέγουμε τυχαία m από τις Μ ομάδες από τις οποίες αποτελείται ο πληθυσμός. Οι m ομάδες είναι το τυχαίο δείγμα. Για κάθε ομάδα πού θα εκλεγεί, χρησιμοποιούνται όλες οι μονάδες παρατήρησης.

Αν εκλεγούν επαναλαμβανόμενα δείγματα ομάδων, η διασπορά στα δειγματοληπτικά αποτελέσματα θα εξαρτάται μόνο από το είδος των ομάδων που εκλέγονται.

Η διασπορά των συνόλων των ομάδων $T_{i.}$ ορίζεται ως εξής:

$$s_b^2 = \frac{1}{(M-1)} \cdot \sum_{i=1}^{M} (T_{i.} - \overline{T}..)^2 \qquad \textbf{(1.20)}$$

όπου $\overline{T_{..}}$ είναι ο μέσος όρος των συνόλων των ομάδων:

$$\overline{T..} = \frac{1}{N} \sum_{i-1}^{M} T_i. \qquad \textbf{(1.21)}$$

Ο παράγοντας s_b^2 θεωρείται ως η διασπορά μεταξύ των ομάδων.

1.11 Εκτίμηση του Συνόλου του Πληθυσμού στη Δειγματοληψία κατά Ομάδες

Παράδειγμα 1.11.1

Υποθέτουμε ότι υπάρχουν 2 μεγάλα πολυκαταστήματα γενικών οικιακών επιπλώσεων των εταιρειών ΙΚΕΑL και ΕΝΤΟΣΜ. Ο Διευθυντής της εταιρείας ΙΚΕΑL θέλει να προσδιορίσει με δειγματοληψία κατά ομάδες την συνολική κατανάλωση Τ της ανταγωνιστικής εταιρείας ΕΝΤΟΣΜ. Παίρνει ένα τυχαίο δείγμα από m τμήματα της πόλης που βρίσκονται τα πολυκαταστήματα της ΕΝΤΟΣΜ και η οποία έχει διαιρεθεί σε Μ τμήματα (clusters). Η εταιρεία ΙΚΕΑL ερωτά όλα τα πρόσωπα (ή τους αρχηγούς νοικοκυριών) που διαμένουν στα m τμήματα για το ποια ήταν η αξία των επίπλων που αγόρασαν από την εταιρεία ΕΝΤΟΣΜ κατά τη διάρκεια των 2 τελευταίων χρόνων.

Το σύνολο Τ της αξίας των αγορών από την εταιρεία ΕΝΤΟΣΜ εκτιμάται ως εξής:

α. Υπολογίζεται το σύνολο του (i) τμήματος (ομάδας), μεγέθους N_i από την εξίσωση:

$$T_i. = \sum_{j=1}^{N_i} x_{ij} \qquad \textbf{(1.22)}$$

β. Υπολογίζεται το σύνολο της αξίας των αγορών από τη σχέση:

$$T_{..} = \sum_{i=1}^{m} T_i. \qquad \textbf{(1.23)}$$

για το δείγμα των m τμημάτων (ομάδων).

γ. Υπολογίζεται ο μέσος του συνόλου των τμημάτων (ομάδων) δηλαδή η μέση συνολική αξία αγορών των Μ τμημάτων, από την εξίσωση:

$$\overline{\hat{T}_{..}} = \frac{1}{m} \cdot \sum_{i-1}^{m} Ti. = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{N_i} x_{ij} \qquad \textbf{(1.24)}$$

δ. Εκτιμάται η συνολική αξία των αγορών από τα Μ τμήματα του πληθυσμού από τη εξίσωση:

$$\hat{T}_{CL} = M.(\frac{1}{m}.\sum_{i=1}^{m} T_{i..}) = \frac{M}{m}.\sum_{i=1}^{m} \sum_{j=1}^{Ni} x_{ij} \qquad \textbf{(1.25)}$$

ε. Η διασπορά του εκτιμητή $\hat{T}_{CL}$ είναι:

$$V(\hat{T}_{CL}) = M^2 \left(\frac{M-m}{M} \right) .. \frac{S_b^2}{m} \qquad \textbf{(1.26)}$$

Ο εκτιμητής της S_b^2 είναι:

$$S_b^2 = \frac{1}{(m-1)} \sum_{i=1}^{m} (T_{i\bullet} - \overline{T_{\bullet\bullet}})^2 \qquad \textbf{(1.27)}$$

Συνεπώς ο εκτιμητής της διασποράς $V(\check{T}_{cl})$ είναι:

$$\hat{V}(\hat{T}_{cl}) = M^2 \left(\frac{M-m}{M} \right) \cdot \frac{s_b^2}{m} \qquad \textbf{(1.28)}$$

1.12 Εκτίμηση του Μέσου του Πληθυσμού στη Δειγματοληψία κατά Ομάδες

Υποθέτουμε ότι όλες οι ομάδες (clusters) έχουν τον ίδιο αριθμό μονάδων (προσεγγιστικά). Τότε έχουμε:

$$N_i = \overline{N} = \frac{N}{M} \text{ i=1,2,…,M, } N = M \cdot \overline{N} \qquad \textbf{(1.29)}$$

$$\overline{x_{cl}} = \frac{\hat{T}_{cl}}{N} = \frac{\hat{T}_{cl}}{M\overline{N}} = \frac{1}{m\overline{N}} . \sum_{i=1}^{m} \sum_{j=1}^{N_I} x_{ij} \qquad \textbf{(1,30)}$$

$$V(\overline{x_{cl}}) = (\frac{1}{M\overline{N}})^2 . V(\hat{T}_{cl}) = \frac{M^2}{M^2 . \overline{N}^2} \cdot \left(\frac{M-m}{M} \right) \cdot \frac{s_b^2}{m} = \frac{1}{\overline{N}^2} \left(\frac{M-m}{M} \right) \cdot \left(\frac{s_b^2}{m} \right) \textbf{(1.31)}$$

1.13 Διαστήματα Εμπιστοσύνης για τις παραμέτρους (Τ) και (μ) στη Δειγματοληψία κατά Ομάδες

Τα διαστήματα εμπιστοσύνης για τις παραμέτρους (μ) και (Τ) με συντελεστή εμπιστοσύνης 100(1-α)%, δίνονται, αντίστοιχα από τις σχέσεις:

$$\overline{x}_{cl} - Z_{a/2}\sqrt{V(\overline{x_{cl}})} \le \mu \le \overline{x_{cl}} + Z_{a/2}\sqrt{V(\overline{x_{cl}})} \qquad \textbf{(1.32)}$$

$$\hat{T}_{cl} - Z_{a/2}\sqrt{V(\hat{T}_{cl})} \le T \le \hat{T}_{cl} + Z_{a/2}\cdot\sqrt{V(\hat{T}_{cl})} \qquad \textbf{(1.33)}$$

Παράδειγμα 1.13.1 Ο Διευθυντής προσωπικού μίας αλυσίδας 300 καταστημάτων κινητής τηλεφωνίας θέλει να εκτιμήσει την μέση ηλικία του προσωπικού. Κάθε κατάστημα απασχολεί περίπου 6 ανθρώπους. Ο Διευθυντής προσωπικού εκλέγει ένα τυχαίο δείγμα 10 καταστημάτων (ομάδων) και προσδιορίζει τις ηλικίες των εργαζομένων στα 10 αυτά καταστήματα. Ο ακόλουθος πίνακας περιέχει τα αποτελέσματα:

Πίνακας 1.13.1 *Ηλικίες εργαζομένων σε 10 καταστήματα*

Κατάστημα	1	2	3	4	5	6	7	8	9	10
	16	21	18	15	16	19	19	16	19	21
	18	20	18	17	21	15	15	15	20	16
	17	18	18	20	17	18	20	16	16	20
	19	16	18	20	19	18	17	21	22	21
	16	19	17	21	19	16	19	24	16	21
	15	15	15	18	17	18	15	17	20	16
Σύνολο	101	109	104	111	109	104	105	109	113	115

Από την (1.2.9.) έχουμε: $N_i = \overline{N} = 6 = \frac{N}{M}$, $N=N_i.M$

N=6, M=10, $N = N_i$. M=6.10=60.

Από την (1.30) έχουμε

$$\overline{\overline{x_{cl}}} = \frac{101+109+\ldots+115}{10(6)} = \frac{1080}{60} = 18$$

Υπολογίζουμε ένα εκτιμητή της διασποράς μεταξύ ομάδων με τη βοήθεια της (1.31).

Ισχύει: $\overline{\hat{T}}_{..} = \frac{1}{m}\cdot\sum_{i=1}^{m} T_{i.} = \frac{1}{10}\cdot 1080 = 108$ από την (1.24):

Ισχύει: $s_b^2 = \dfrac{(101-108)^2 + (109-108)^2 + ... + (115-108)^2}{(10-1)} = 19{,}56$,

από την (1.27)

Ισχύει, από την (1.31)

$$\hat{V}(\overline{x_{cl}}) = \frac{1}{6^2}\left(\frac{300-10}{300}\right)\cdot\frac{16{,}56}{10} = 0{,}0525$$

Το διάστημα εμπιστοσύνης 95% για το μέσο είναι, από την (1.32):

$$18 - 1{,}96\cdot\sqrt{0{,}0525} \le \mu \le 18 + 1{,}96\cdot\sqrt{0{,}0525} \text{ ή } 17{,}55 \le \mu \le 18{,}45$$

Παράδειγμα 1.13.2 Ο λογιστής μίας αλυσίδας 100 καταστημάτων δώρων θέλει να εκτιμήσει τη συνολική τιμή σε Ευρώ των ακαλύπτων επιταγών που ελήφθησαν από πελάτες σε μία ορισμένη εβδομάδα. Εκλέγει ένα τυχαίο δείγμα από 10 καταστήματα (ομάδες) και υπολογίζει την συνολική αξία των ακαλύπτων επιταγών σε κάθε κατάστημα. Ο παρακάτω πίνακας περιέχει τα αποτελέσματα:

Πίνακας 1.13.2 *Συνολική αξία σε Ευρώ ακαλύπτων επιταγών στα 10 καταστήματα*

Κατάστημα	1	2	3	4	5	6	7	8	9	1
Συνολική Αξία σε €	125	100	130	95	110	105	100	110	115	120

Σύνολο:1110

Έχουμε από την (1.24):

$$\hat{\bar{T}} = \frac{1110}{10} = 111$$

Από την (1.25) έχουμε:

$\hat{T}_{cl}$ =100. (111)=11100

Από την (1.27) έχουμε

$$s_b^2 = \frac{(125-111)^2 + (100-111)^2 + ... + (120-111)^2}{(10-1)} = 132{,}22$$

Από την (1.28) έχουμε:

$$\hat{V}(\hat{T}_{cl}) = 100^2\left(\frac{100-10}{100}\right)\cdot\frac{132{,}22}{10} = 118{,}998$$

Το διάστημα εμπιστοσύνης με συντελεστής 95% για το Τα είναι από την (1.33):

$$11100 - 1{,}96\cdot\sqrt{118998} \le T \le 11100 + 1{,}96\cdot\sqrt{118998} \text{ ή } 10424 \le T \le 11776\,.$$

1.14 Συστηματική Δειγματοληψία

Ορισμός 1.14.1: **Συστηματικό δείγμα** (1 στα Μ) ή $\frac{1}{M}$ λέγεται το δείγμα που λαμβάνεται ως εξής: Εκλέγεται τυχαία 1 μονάδα από τις πρώτες Μ μονάδες π.χ. η μονάδα αριθμός 7 και μετά εκλέγεται η κάθε μονάδα η οποία βρίσκεται Μ θέσεις μετά την 7 (M+7,2M+7,…) έως όταν καλυφθεί όλος ο πληθυσμός.

Παράδειγμα 1.14.1: Ένας Καθηγητής θέλει να υπολογίσει ένα γρήγορο εκτιμητή του βαθμού των φοιτητών σε ένα μάθημα ελέγχοντας τα 2% των φύλλων διαγωνισμάτων στο μάθημα αυτό σε μια εξεταστική περίοδο. Τα φύλλα του διαγωνίσματος είναι ταξινομημένα με αλφαβητική σειρά.

Ισχύει: $2\% = \frac{1}{50}$.

Άρα ο καθηγητής εκλέγει τυχαία ένα διαγώνισμα, π.χ. το διαγώνισμα αριθμός 7 και μετά εκλέγει το 57, 107, ……, διαγώνισμα ως όπου καλύψει όλο τον πληθυσμό των φύλλων διαγωνισμάτων.

Στη συστηματική δειγματοληψία υπάρχουν τα εξής 3 είδη πληθυσμού από τα οποία προέρχεται το τυχαίο δείγμα:

α. Τυχαίος πληθυσμός, στον οποίο οι μονάδες παρατήρησης είναι με τυχαία σειρά. Οι μονάδες παρατηρήσεις βρίσκονται στο δειγματοληπτικό πλαίσιο.

β. Περιοδικός πληθυσμός, στον οποίο οι μονάδες παρατήρησης πλαίσιο είναι σε περιοδικότητα.

γ. Ταξινομημένος κατά τάξη μεγέθους πληθυσμούς, στον οποίο οι μονάδες παρατήρησης είναι ταξινομημένες κατά τάξη μεγέθους.

Όταν λαμβάνουμε δείγματα από έναν ταξινομημένο ή περιοδικό πληθυσμό, μπορούμε να λάβουμε **επαναλαμβανόμενα** συστηματικά δείγματα, όπως δείχνει το παρακάτω παράδειγμα. Στην περίπτωση της μοναδικής συστηματικής δειγματοληψίας, ισχύουν οι τύποι της απλής τυχαίας δειγματοληψίας.

Παράδειγμα 1.14.2 Η Εταιρεία τσιμέντων Ήφαιστος Α.Ε. απασχολεί 200 εργαζομένους στο εργοστάσιό της στη Χαλκίδα. Ο Διευθυντής προσωπικού θέλει να εκτιμήσει την μέση ηλικία των εργαζομένων με επαναλαμβανόμενη συστηματική δειγματοληψία (repeated random sampling). Κάθε εργαζόμενος έχει μία πράσινη κάρτα στην οποία γράφεται η ημερομηνία γεννήσεως του. Οι κάρτες αυτές είναι ταξινομημένες με αλφαβητική σειρά. Η εταιρεία αποφασίζει να εκλέξει ένα δείγμα 10% μεγέθους $200 \cdot 10\% = 20$. Εκλέγει 4 συστηματικά δείγματα ως εξής: Το μέγεθος του κάθε δείγματος είναι $\frac{20}{4} = 5$.

Επειδή 200/5=40, υπάρχουν 40 συστηματικά δείγματα μεγέθους 5 που μπορούν να εκλεγούν από έναν πληθυσμό μεγέθους 200. Η εταιρεία εκλέγει 4 τυχαίους αριθμούς {10,37,8,12}από ένα πίνακα τυχαίων αριθμών. Ο πίνακας 1.14.1 δείχνει τα αποτελέσματα της δειγματοληψίας.

Πίνακας 1.14.1. *τέσσερα συστηματικά δείγματα μεγέθους 5 για το παράδειγμα 1.14.2*

1		**2**		**3**		**4**	
Αριθ-μός	**Ηλικία**	**Αριθ-μός**	**Ηλικία**	**Αριθ-μός**	**Ηλικία**	**Αριθ-μός**	**Ηλικία**
10	61	37	30	8	31	12	64
50	26	77	26	48	55	52	25
90	56	117	37	88	38	92	27
130	37	157	40	128	44	132	44
170	63	197	25	168	41	172	32
Σύνολο	243	**Σύνολο**	158	**Σύνολο**	209	**Σύνολο**	192

Από την (1.30) έχουμε:

$$\overline{\overline{x_{cl}}} = \frac{243+158+209+122}{20} = 40{,}1\,.$$

Ο λόγος για τη παραπάνω εξίσωση είναι ότι τα 4 δείγματα θεωρούνται σαν τέσσερες ομάδες (clusters)

Από την (1.24) έχουμε:

$$\hat{\overline{\overline{T}}}_{..} = \frac{243+158+209+192}{4} = 200{,}5$$

Από την (1,27) έχουμε:

$$s_b^2 = \frac{(243-200{,}5)^2 + (158-200{,}5)^2 + \ldots + (192-200{,}5)^2}{(4-1)} = 1252{,}33$$

Από την (1.31) έχουμε

$$\hat{V}(\overline{\overline{x_{cl}}}) = \frac{1}{(5)^2}\left(\frac{40-4}{40}\right)\cdot\frac{252{,}33}{4} = 11{,}27$$

Το διάστημα εμπιστοσύνης με συντελεστή 95% για τον μέσο (μ) είναι, από την (1.32)

$$40{,}1 - 1{,}96.\sqrt{11{,}27} \le \mu \le 40{,}1 + 1{,}96\cdot\sqrt{11{,}27} \text{ ή } 33{,}5 \le \mu \le 46{,}7$$

Η συστηματική δειγματοληψία καθιστά απαραίτητη την ύπαρξη έτοιμου δειγματοληπτικού πλαισίου. Πρέπει, οι μονάδες παρατήρησης στο δειγματοληπτικό πλαίσιο να είναι τυχαία κατανεμημένες.

1.15 Κόστος Δειγματοληψίας, Αποτελεσματικότητα και Μέγεθος Δείγματος

Το συνολικό κόστος C μίας δειγματοληψίας έχει δύο συνιστώσες: Η μία συνιστώσα είναι η **συνάρτηση του αριθμού των μονάδων (u) παρατήρησης C_u** και η άλλη συνιστώσα είναι το **πάγιο κόστος C_f,** το οποίο είναι συνάρτηση του χρόνου και των εξόδων που πρέπει να διατεθούν για να φθάσει ο δειγματολήπτης στα διάφορα στρώματα ή στις διάφορες μονάδες παρατήρησης.

1.16 Προσδιορισμός Κόστους

Τα έξοδα για τα διαφορετικά είδη δειγματοληψίας είναι τα ακόλουθα:

Απλή τυχαία δειγματοληψία: $C = C_f + n \cdot C_u$ **(1.34)**

Στρωματοποιημένη δειγματοληψία: $C = C_f + \sum_{h=1}^{L} n_h \cdot C_{uh}$ **(1.35)**

Δειγματοληψία κατά ομάδες: $C = C_f + \sum_{i=1}^{m} N_i \cdot C_{u_i}$ **(1.36)**

Συστηματική δειγματοληψία: $C = C_f + N_i \cdot C_{u_i}$ **(1.37)**

Η αποτελεσματικότητα του σχεδιασμού της δειγματοληψίας είναι συνδεδεμένη άμεσα με τη διασπορά του εκτιμητή της παραμέτρου που θα εκτιμηθεί: (Τ) ή (μ) ή (αναλογία= π). Η δειγματοληψία που παράγει εκτιμητές με μικρή διασπορά είναι η καλύτερη, ειδικά αν το κόστος είναι μικρό.

1.16.1 Μέγεθος Δείγματος για ένα Προκαθορισμένο Προϋπολογισμό

Αν ο προϋπολογισμός (C) είναι προκαθορισμένος, το μέγεθος του δείγματος είναι το εξής:

$$n = \frac{C - C_f}{C_u} \text{ (απλή τυχαία δειγματοληψία)} \quad \textbf{(1.38)}$$

$m = \dfrac{C - C_f}{\overline{N}C_{u_i}}$ (δειγματοληψία κατά ομάδες και συστηματική δειγματοληψία – όταν όλα τα έξοδα C_{u_i} είναι σταθερά). **(1.39)**

$n = \dfrac{C - C_f}{C_{u_h}}$ (στρωματοποιημένη δειγματοληψία – όταν όλα τα έξοδα C_{u_h} είναι σταθερά). **(1.40)**

1.17 Μέγεθος Δείγματος για Προκαθορισμένη Ποιότητα Εκτιμητή

1.17.1 Απλή Τυχαία Δειγματοληψία (Εύρεση Δείγματος)

$n = \dfrac{N^2 z^2 S^2}{Nd^2 + z^2 S^2}$, (απλή τυχαία δειγματοληψία) **(1.41)**

όπου:

d= μέγιστο επιθυμητό δειγματοληπτικό σφάλμα

z= τιμή της Κανονικής παρέκκλισης (normit)

S^2= διασπορά πληθυσμού που μπορεί να εκτιμηθεί από τον εκτιμητή S^2 που λαμβάνεται από μία πολιτική δειγματοληψία

N= μέγεθος του πληθυσμού.

Όταν ο πληθυσμός είναι πεπερασμένος, τότε ισχύει

$$n = \frac{N^2 z^2 \sigma^2}{d^2 (N-1) + z^2 \sigma^2} \quad \textbf{(1.42)}$$

Όταν ισχύει $n/N \le 0{,}05$, τότε ισχύει

$$n = \frac{z^2 \sigma^2}{d^2} \quad \textbf{(1.43)}$$

Αν ο εκτιμητής είναι ο μέσος $\bar{x}$ της παραμέτρου μ που ζητείται να εκτιμηθεί, τότε ισχύει:

$$d = \mu - \bar{x} \quad \textbf{(1.44)}$$

1.17.2 Στρωματοποιημένη Δειγματοληψία (Εύρεση Δείγματος)

1.17.2.1. Τυχαία Δείγματα Ισων Μεγεθών από κάθε Δειγματοληπτικό Στρώμα

Ισχύει:

$$n_h = \frac{n}{L} \text{ (τύπος υπολογισμού μεγέθους δειγμάτων)} \quad \textbf{(1.45)}$$

Αν συμβολισθεί με

n: συνολικό δείγμα

L: αριθμός στρωμάτων,

τότε ισχύει:

$$n = \frac{z^2 L \cdot \sum_{i=1}^{L} N_h^2 \cdot S_h^2}{N^2 d^2 + z^2 \cdot \sum_{h=1}^{L} N_h S_h^2} \text{ (τύπος μεγέθους συνολικού δείγματος)} \quad \textbf{(1.46)}$$

1.17.2.2. Τυχαία Δείγματα τα οποία λαμβάνονται από κάθε Στρώμα, ανάλογα με το Μέγεθός του

Ισχύει

$$n_h = \frac{N_h}{N} \cdot (n) \text{ (τύπος υπολογισμών δείγματος στρώματος)} \quad \textbf{(1.47)}$$

$$n = \frac{z^2 N \sum_{h=1}^{L} N_n \cdot S_h^2}{N^2 d^2 + z^2 \sum_{n=1}^{L} N_h \cdot S_h^2} \text{ (τύπος υπολογισμών συνολικού δείγματος)} \quad \textbf{(1.48)}$$

1.17.2.3 Βέλτιστος Επιμερισμός Δείγματος που Επιτρέπει τη Ύπαρξη Μεταβλητότητος στο Κόστος και στη Διασπορά κατά Μήκος των Στρωμάτων

$$n = \frac{N_h \cdot S_h \cdot \sqrt{C_{uh}}}{\sum_{h=1}^{L} \left((N_h S_h) / \sqrt{C_{uh}} \right)} \cdot (n) \text{ (τύπος επιμερισμού δείγματος)} \quad \textbf{(1.49)}$$

$$n = \frac{z^2 (\sum_{h=1}^{L} N_h S_h \cdot \sqrt{C_{uh}}) \cdot \left(\frac{\sum_{h=1}^{L} N_h S_h}{\sqrt{C_{uh}}} \right)}{N^2 d^2 + z^2 \sum_{h=1}^{L} N_h \cdot S_h^2} \text{ (τύπος συνολικού δείγματος)} \quad \textbf{(1.50)}$$

1.17.2.4 Επιμερισμός κατά Neyman που επιτρέπει τη ύπαρξη μεταβλητότητας μόνο στις Διασπορές κατά Μήκος των Στρωμάτων. Υποτίθεται ότι όλα τα έξοδα C_{uh} είναι ίσα κατά μήκος των Στρωμάτων.

Ισχύει

$$n_h = \frac{N_h \cdot S_h}{\sum_{h=1}^{L} N_h \cdot S_h} \cdot (n) \text{ (τύπος επιμερισμού δείγματος)} \quad \textbf{(1.51)}$$

$$n = \frac{z^2 \left(\sum_{h=1}^{L} N_h \cdot S_h \right)^2}{N^2 d^2 + z^2 \sum_{h=1}^{L} N_h \cdot S_h^2} \text{ (τύπος μέγεθος συνολικού δείγματος)} \quad \textbf{(1.52)}$$

Πρακτικά, εκτιμώνται οι διασπορές S_h^2 από δειγματικές διασπορές s_h^2 κατά μήκος των στρωμάτων μετά από πιλοτικές δειγματοληψίες σε όλα τα στρώματα.

1.17.3 Δειγματοληψία κατά ομάδες και Συστηματική Δειγματοληψία (Εύρεση Δείγματος)

Ισχύει ο κάτωθι τύπος για δειγματοληψία κατά ομάδες και συστηματική δειγματοληψία:

$$m = \frac{Mz^2 S_b^2}{Md_{cl}^2 + z^2 S_b^2} \quad \textbf{(1.53)}$$

όπου $d_{cl} = \overline{N}d$ = μέγιστο επιθυμητό δειγματοληπτικό σφάλμα κατά την εκτίμηση του συνόλου Τ στην δειγματοληψία κατά ομάδες.

1.18 Εκτίμηση των Δειγματοληπτικών Διασπορών με τις Μεθόδους "JACKKNIFE" και "BOOTSTRAP".

1.18.1 Η Μέθοδος Εκτίμησης Παραμέτρων Jackknife στη Στατιστική

Για να ορίσουμε τον εκτιμητή **Jackknife** στη θεωρία Δειγματοληψίας, πρέπει να ορίσουμε πρώτα την Πρωτογενή Δειγματοληπτική Μονάδα (PSU).

Ορισμός 1.18.1.1 Πρωτογενής δειγματοληπτική μονάδα (PSU) στην περίπτωση της πολυσταδιακής δειγματοληψίας (multistage sampling) είναι η ομάδα των παρατηρήσεων (cluster) πού εκλέγεται κατά το πρώτο στάδιο της δειγματοληψίας.

Στην περίπτωση της στρωματοποιημένης δειγματοληψίας, θα μπορούσαν να υπάρχουν L στρώματα με (k) PSU's σε κάθε στρώμα και y_{1i}, y_{2i}, …, y_{ni}, μονάδες παρατήρησης σε κάθε σε κάθε $PSU_{(i)}$ (i=1, 2, …, k).

1.18.2 Περιγραφή της Μεθόδου Jackknife

Η ανάπτυξη της μεθόδου Jackknife άρχισε με μια εργασία του Quenouille (1956) με αντικείμενο τη ελάττωση της **μεροληψίας (bias) των εκτιμητών**. Επιπρόσθετες διασκευές της μεθόδου, (**Mosteller and Tukey, 1968),** οδήγησαν στην εφαρμογή της μεθόδου σε μία ποικιλία προβλημάτων στις Κοινωνικές Επιστήμες όπου δεν υπήρχαν έτοιμες μαθηματικές εκφράσεις για τον υπολογισμό των δειγματοληπτικών διασπορών. Ο συγγραφέας του βιβλίου αυτού έχει συγγράψει την διδακτορική του διατριβή με θέμα την θεωρία και τις εφαρμογές των μεθόδων Jackknife και Bootstrap (βλ. Frangos (1980, 1983, 1984, 1987, 1990, 1992, 1994, 1995, 2004b, 2005a, 2005b, 2007, 2007b, 2007c, 2007d, 2007e, 2007f, 2008, 2009a, 2009b, 2009c, 2009d, 2010a, 2010b)

Ας υποθέσουμε ότι $\hat{\theta}$ είναι ένας εκτιμητής μιας παραμέτρου (θ) ενός πληθυσμού, ο οποίος βασίζεται σε όλα τα δεδομένα y_1, y_2, …, y_n ή αν έχουμε (k) ομάδες (g_1, g_2, …, g_k) παρατηρήσεων y_i

$$\hat{\theta} = \theta(g_1, g_2, ..., g_k) \qquad \textbf{(1.54)}$$

Έστω ότι $\hat{\theta}_{(i)}$ είναι ένας εκτιμητής της παραμέτρου θ ο οποίος παράγεται με την ίδια μαθηματική έκφραση όπως ο εκτιμητής θ, αλλά βασίζεται σε όλα τα δεδομένα εκτός των δεδομένων στην (i) ομάδα (group)

$$\hat{\theta}_{(i)} = \theta(g_1, g_2, ..., g_{i-1}, g_{i+1}, ..., g_k), \quad \iota = 1,2,\ldots,k \qquad \textbf{(1.55)}$$

Οι δειγματικοί συντελεστές w_i των δεδομένων που μένουν στο δείγμα αν παραληφθεί η (i) ομάδα (group). πολλαπλασιάζονται επί ($\frac{k}{k-1}$), όπου (k) είναι ο αριθμός των ομάδων (groups), όταν υπολογίζεται ο εκτιμητής $\hat{\theta}_{(i)}$ .. Αυτός ο επαναπροσδιορισμός των βαρών (w_i) είναι απαραίτητος για παραμέτρους που περιλαμβάνουν σύνολα (T), αλλά όχι για παραμέτρους που περιλαμβάνουν μέσους ($\overline{X}$) στην στρωματοποιημένη δειγματοληψία.

Ορισμός 1.18.2.1: Ο εκτιμητής Jackknife της διασποράς του εκτιμητή $\hat{\theta}_{JK}$ αλλά και της διασποράς του αρχικού εκτιμητή $\hat{\theta}$ της παραμέτρου θ δίνεται από τον τύπο:

$$\hat{\text{var}}_{JK}(\hat{\theta}) = \frac{k-1}{k} \cdot \sum_{i=1}^{k} (\hat{\theta}_{(i)} - \hat{\theta})^2 \quad \text{ή} \tag{1.56}$$

$$\hat{\text{var}}_{JK}(\hat{\theta}) = \frac{1}{k(k-1)} \cdot \sum_{i=1}^{k} (\hat{\theta}_i - \bar{\hat{\theta}})^2 \tag{1.57}$$

όπου $\hat{\theta}_i = k\hat{\theta} - (k-1)\hat{\theta}_{(i)}$,

$\hat{\theta}_{(i)} = \theta(g_1, g_2, ..., g_{i-1}, g_{i+1}, ..., g_k)$, ι = 1,2,…,k

$$\bar{\hat{\theta}} = \frac{1}{k} \cdot \sum^{k} \hat{\theta}_i \tag{1.58}$$

Ορισμός 1.18.2.2 : Ο εκτιμητής Jackknife της παραμέτρου θ δίνεται από τον τύπο

$$\bar{\hat{\theta}} = \frac{1}{k} \cdot \sum^{k} \hat{\theta}_i \tag{1.59}$$

1.18.3 Πλεονεκτήματα της Μεθόδου Jackknife

Τα πλεονεκτήματα της μεθόδου εκτιμητικής Jackknife είναι τα εξής:

1 Αν υπάρχει ένας εκτιμητής $\hat{\theta}$, της παραμέτρου θ, ο οποίος έχει αμεροληψία τάξης $\frac{1}{n}$, δηλαδή ισχύει:

$$E(\hat{\theta}) = \theta + \frac{\text{A}}{n} + o\left(\frac{1}{n^2}\right) \tag{1.60}$$

τότε ο εκτιμητής Jackknife της παραμέτρου θ, το οποίο θα συμβολίσουμε με $\bar{\hat{\theta}}$ ή $\hat{\theta}_{JK}$ έχει αμεροληψία μικρότερη από την αμεροληψία του εκτιμητή $\hat{\theta}$, δηλαδή

$$\text{E}(\hat{\theta}_{JK}) = \theta + \frac{B}{n^2} + o\left(\frac{1}{n^3}\right)) \text{(βλ. Quenonille (1956))} \tag{1.61}$$

2 Η έκφραση

$$\hat{\text{var}}_{JK}(\hat{\theta}) = \frac{k-1}{k} \cdot \sum_{i=1}^{k} (\hat{\theta}_{(\iota)} - \hat{\theta})^2 \quad \text{ή} \tag{1.62}$$

$$\hat{\text{var}}_{JK}(\hat{\theta}) = \frac{1}{k(k-1)} \cdot \sum_{i=1}^{k} (\hat{\theta}_{\iota} - \bar{\hat{\theta}})^2 \tag{1.63}$$

$$\text{όπου } \hat{\theta}_{\iota} = k\hat{\theta} - (k-1)\hat{\theta}_{(\iota)} \tag{1.64}$$

$$\bar{\hat{\theta}} = \frac{1}{k} \cdot \sum^{k} \hat{\theta}_i$$

είναι ένας μη παραμετρικός (distribution – free) εκτιμητής της διασποράς του Jackknife εκτιμητή $\bar{\hat{\theta}}$, αλλά και του αρχικού εκτιμητή $\hat{\theta}$ της παραμέτρου θ (Tukey, 1958).

3 Σύμφωνα με την θεμελιώδη εικασία(conjecture) του Tukey (1958) αλλά και με αναλυτική απόδειξη άλλων ερευνητών (Durbin, 1959, Frangos 1980, 1983, 1984, 1986, 1990, 1995), οι εκφράσεις

$$\hat{\theta}_i = \hat{k}\theta - (k-1)\hat{\theta}_{(i)} \tag{1.64a}$$

Όπου $\hat{\theta}_{(i)} = \theta(g_1, g_2, ..., g_{i-1}, g_{i+1}, ..., g_k)$, ι = 1,2,…,k

οι οποίες λέγονται **ψευδοτιμές (pseudovalues)** του εκτιμητή jackknife, είναι προσεγγιστικά ανεξάρτητες μεταξύ τους και έχουν ως Κατανομή Πιθανότητας την Κατανομή Student's (t) (βλ. ΣΤΑΤΙΣΤΙΚΗ ΕΠΙΧΕΙΡΗΣΕΩΝ, Χρ. Φράγκος, Εκδόσεις Σταμούλης,1998).

Το αποτέλεσμα αυτό μπορεί να εφαρμοστεί για τις τιμές $\hat{\theta}_{(i)}$, έτσι ώστε να κατασκευασθεί το ακόλουθο ανθεκτικό (robust) διάστημα εμπιστοσύνης 100(1-α) για τή παράμετρο θ (για $k \geq 30$).

$$\bar{\hat{\theta}} - z_{\alpha/2}\sqrt{\hat{\theta}_{JK}(\hat{\theta})} \leq \theta \leq \bar{\hat{\theta}} + z_{\alpha/2} \cdot \sqrt{\hat{\text{var}}_{JK}(\hat{\theta})} \tag{1.65}$$

1.18.4 Εφαρμογές της Μεθόδου Εκτιμητικής Jackknife στη Δειγματοληψία

Οι ερευνητές Frankel (1971) και Durbin (1959) εφάρμοσαν τη μέθοδο Jackknife στην εκτίμηση της διασποράς των εκτιμητών των παραμέτρων στην θεωρία της

δειγματοληψίας. Στο κεφάλαιο αυτό θα εφαρμοσθεί η μεθόδος Jackknife στην στρωματοποιημένη δειγματοληψία ως εξής:

Υποθέτουμε ότι υπάρχει ένα πολύ-σταδιακό στρωματοποιημένο σχέδιο στο οποίο υπάρχουν L στρώματα. Τυχαία εκλέγονται K_h (PSU) (πρωτογενείς δειγματοληπτικές μονάδες) από το στρώμα h, h=1,2,3,…,L. Υποθέτουμε ότι $\hat{\theta}$ είναι ο εκτιμητής μιας παραμέτρου που βασίζεται σε όλες τις PSU, δηλ. σε K_h παρατηρήσεις. (Η κάθε παρατήρηση,στο παράδειγμα αυτό,είναι μία (PSU)). Στη συνέχεια, θεωρούμε ένα νέο σύνολο δειγματοληπτικών δεδομένων που αποκλείει τις παρατηρήσεις στην (i) PSU του στρώματος (h) πολλαπλασιάζει τους δειγματικούς συντελεστές των παρατηρήσεων που απομένουν στο (h) στρώμα επί τον παράγοντα $(K_h/(K_h-1))$ και δεν αλλάζει τους δειγματικούς συντελεστές των παρατηρήσεων σε άλλα στρώματα. Έστω ότι $\hat{\theta}_{(hi)}$ είναι ο εκτιμητής της παραμέτρου θ που βασίζεται στο νέο αυτό σύνολο δειγματοληπτικών δεδομένων.

Ο εκτιμητής Jackknife της διασποράς του $\hat{\theta}$ δίνεται από τη σχέση:

$$\hat{\text{var}}_{JK}(\hat{\theta}) = \sum_{h=1}^{L}\left(\frac{K_h-1}{K_h}\right)\cdot\sum_{i=1}^{K_h}\left(\hat{\theta}_{(hi)} \hat{-} \hat{\theta}\right)^2 \qquad \textbf{(1.66)}$$

Παράδειγμα 1.18.4.1 (Korn and Granbard, 1999) Εκτίμηση της διασποράς και της τυπικής απόκλισης του εκτιμητή Weighted Mean Hemoglobin Level για γυναίκες που συμμετείχαν σε μία δειγματοληψία (HHANES III) στις Ηνωμένες Πολιτείες, με χρήση της μεθόδου Jackknife.

Το σχέδιο της δειγματοληψίας είναι το εξής:

Ο πληθυσμός των Ισπανόφωνων Αμερικανίδων διαιρείται σε 8 στρωμάτα (strata) h=1,2,…8. (L=8). Λαμβάνεται ένα τυχαίο δείγμα από 2 (PSU) (από κάθε στρώμα. Κάθε (PSU) έχει διαφορετικό αριθμό γυναικών. Εξελέγησαν τυχαία 3603 γυναίκες από τις οποίες 3369 είχαν σοβαρά επίπεδα του χαρακτηριστικών του αίματος. hemoglobin. Ο δειγματικός συντελεστής(βάρος) για κάθε γυναίκα είναι το επίπεδο (σε gm/Dl) της τιμής του δείκτη hemoglobin. Ο σταθμικός μέσος των τιμών hemoglobin ήταν $\hat{\theta} = 13{,}457$ (gm/dL). Ζητείται να εκτιμηθεί η διασπορά του $\hat{\theta}$, με τη μέθοδο Jackknife. Ο παρακάτω πίνακας περιέχει τα αποτελέσματα:

Πίνακας 1.18.4.1 *Μέγεθος δείγματος, Σταθμικός μέσος Hemoglobin και άθροισμα των δειγματικών βαρών (συντελεστών) για τις γυναίκες στη δειγματοληψία HHANES*

Στρώμα	**PSU**	**Μέγεθος δείγματος**	**Σταθμικός Μέσος Hemoglobin**	**Άθροισμα δειγματικών βαρών (συντελεστών)**
1	1	404	13,46	458011
	2	391	13,34	423472
2	1	87	13,04	133529
	2	84	13,08	131240
3	1	181	13,06	183968
	2	239	13,42	261308
4	1	148	13,37	204424
	2	127	13,58	191670
5	1	215	13,57	279680
	2	214	13,28	312737
6	1	238	14,04	264162
	2	193	12,91	238631
7	1	231	13,86	266594
	2	223	13,09	268724
8	1	142	13,62	146645
	2	252	14,05	335698
Σύνολο	3369			4100493

Αν έχουμε ένα σχέδιο στρωματοποιημένης δειγματοληψίας στο οποίο $\overline{y_{hi}}$ είναι ο σταθμικός μέσος των παρατηρήσεων στη (i) (PSU) από το (h) στρώμα, και W_{hi} είναι το άθροισμα των δειγματικών συντελεστών στις παρατηρήσεις που εξελέγησαν (στη (i) (PSU) και στο (h) στρώμα), τότε μπορεί να δειχθεί ότι ο σταθμικός μέσος είναι:

$$\hat{\theta} = \frac{\sum_{h=1}^{L}\sum_{i-1}^{K_h} W_{hi}\cdot \overline{y_{hi}}}{\sum_{h-1}^{L}\sum_{i=1}^{K_h} W_{hi}} \qquad \textbf{(1.67)}$$

Για την εφαρμογή της μεθόδου Jackknife, παραλείπεται η PSU (1) στο στρώμα (1) και πολλαπλασιάζονται οι δειγματικοί συντελεστές (βάρη) της PSU (2) (=423472) επί (2), (διότι ισχύει: $k_1/(k_1-1) = (2/1)=2$)

Τότε ο δειγματικός μέσος του νέου δειγματοληπτικού σύνολο δεδομένων είναι, από τη σχέση (1.67):

$$\hat{\theta}_{((11))} = \frac{13,34 \times 2 \times 423472 + 13,04 \times 133529 + 13,08 \times 131240 + ... + 14,05 \times 335698}{2 \times 423472 + 133529 + 131240 + ... + 335698} = 13,445$$

Στη συνέχεια, από όλο το δείγμα των PSU, παραλείπουμε την PSU (2) του στρώματος (1). Οι δειγματικοί συντελεστές των παρατηρήσεων στην PSU (1) και το στρώμα (1) διπλασιάζονται, γιατί $\frac{k_h}{k_h - 1} = \frac{k_1}{k_1 - 1} = \frac{2}{1} = 2$

Τότε έχουμε:

$$\hat{\theta}_{(12)} = \frac{13,46 \times 2 \times 458011 + 13,04 \times 133529 + 13,08 \times 131240 + ... + 14,05 \times 335698}{2 \times 458011 + 133529 + 131240 + ... + 335698} = 13,470$$

Όμοια, υπολογίζουμε τους εκτιμητές $\hat{\theta}_{(hi)}$ για τα άλλα στρώματα. Οι (16) τιμές των $\hat{\theta}_{hi}$ για h=1,2,3,…8 και i=1,2, είναι οι εξής:

13,445, 13,470, 13,459, 13,456, 13,472, 13,442, 13,467, 13,447, 13,436,
13,479, 13,387, 13526, 13,407, 13,508, 13,498, 13,413

Από τη σχέση (1.66) έχουμε

$\hat{\theta}_{JK}(\hat{\theta}) = 0,0101$.

Ο εκτιμητής jackknife της τυπικής απόκλισης του σταθμικού μέσου είναι:

$0,10 = \sqrt{0,0101}$.

Οι εκτιμητές αυτοί είναι οι ίδιοι που παράγονται με τη βοήθεια άλλων μεθόδων, όπως η μέθοδος της γραμμικοποίησης (linearization method), (βλ. Sampling Techniques, W.G. Cochran, Fourth Edition, 1990, N.Y., Wiley).

1.19 Η Μέθοδος Bootstrap

Η μέθοδος BOOSTRAP δεν έχει χρησιμοποιηθεί αρκετά, ακόμη, στη δειγματοληψία.

Ας υποθέσουμε ότι θέλουμε να εκτιμήσουμε, για μια παράμετρο θ του δείγματος, τον εκτιμητή $\hat{\theta}$ της θ, τον εκτιμητή της διασποράς του $\hat{\theta}$ και ένα διάστημα εμπιστοσύνης για τη παράμετρο θ.

Η ιδέα είναι τα παραχθούν (B) εικονικά δείγματα του ίδιου μεγέθους όπως το αρχικό δείγμα με τη βοήθεια επαναλαμβανομένης, (B) φορές, τυχαίας δειγματοληψίας

ψίας με επανάθεση των αρχικών (PSU) και υπολογισμό του εκτιμητού $\hat{\theta}$ σε κάθε νέο δείγμα. **ΠΡΩΤΟ ΣΤΑΔΙΟ BOOTSTRAP** Αν έχουμε ένα τυχαίο δείγμα από 50 (PSU), τότε, με τη βοήθεια τυχαίων αριθμών, θα μπορούσε να δημιουργηθεί, με τυχαία δειγματοληψία με επανάθεση, ένα άλλο δείγμα 50 (PSU) από το αρχικό δείγμα. Αν, π.χ. το αρχικό δείγμα των 50 (PSU) ήταν:

1, 2, 3, 4, 5, 6, 7, …, 49, 50,
το εικονικό δείγμα θα ήταν

1, 2, 2, 3, 4, 12, 15, 50, 49, …10

1ο δείγμα Bootstrap

Βρίσκουμε αυτό το δείγμα εκλέγοντας 50 τυχαίους αριθμούς με επανάθεση. Αν π.χ. ο πρώτος τυχαίος αριθμός είναι 7, η PSU του νέου δείγματος, θα είναι η PSU του παλαιού δείγματος με αριθμό 7.

Από το νέο δείγμα $i_1,i_2,\ldots,i_{50}$ υπολογίζουμε τον εκτιμητή $\hat{\theta}_1^*$ της παραμέτρου θ της οποίας θέλουμε να εκτιμήσουμε τον εκτιμητή Bootstrap και τη εκτιμήτρια της διασποράς του αρχικού εκτιμητή $\hat{\theta}$.

ΔΕΥΤΕΡΟ ΣΤΑΔΙΟ BOOTSTRAP Αν επαναλάβουμε αυτή τη διαδικασία B=800 φορές, βρίσκουμε 800 εικονικά δείγματα και 800 εκτιμητές $\hat{\theta}_1^*,\hat{\theta}_2^*,\ldots,\hat{\theta}_B^*$.

ΤΡΙΤΟ ΣΤΑΔΙΟ BOOTSTRAP Η εκτιμήτρια Bootstrap της διασποράς τον $\hat{\theta}$ δίνεται από την σχέση:

$$\hat{\text{var}}_{BS}(\hat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \hat{\theta}^*\right)^2 \qquad \textbf{(1.68)}$$

(Efron and Tibshirani, 1993), (Frangos 1990, 1995).

Για τη στρωματοποιημένη δειγματοληψία, κάθε εικονικό δείγμα παράγεται με ανεξάρτητη τυχαία δειγματοληψία των (PSU) με επανάθεση από κάθε στρώμα.

Αν υποτεθεί ότι το αρχικό δείγμα αποτελείται από K_h (PSU) από το στρώμα h, h=1,2,…,L, τότε θα ήταν δυνατό να εκλεγεί **τυχαία με επανάθεση** ένα δείγμα μεγέθους K_h από τις αρχικές K_h (PSU) από κάθε στρώμα h και να ληφθεί το εικονικό δείγμα Bootstrap. Συνήθως εκλέγονται B=1000 ή B=5000 εικονικά δείγματα Bootstrap και εκτελούνται τα στάδια Β και Γ της παραπάνω διαδικασίας.

1.20. Παράδειγμα εισαγωγής δεδομένων ερωτηματολογίου σε αρχείο SPSS

Τα δεδομένα υπάρχουν στο CD ΣΥΝΟΛΑ ΔΕΔΟΜΕΝΩΝ και στο φάκελο ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ 1 (ΚΕΦ. 1, 1. 20)

Ας υποτεθεί ότι δίνεται προς συμπλήρωση το ακόλουθο ερωτηματολόγιο σε φοιτητές:

Ερωτηματολόγιο

Παρακαλούμε συμπληρώστε το παρακάτω ερωτηματολόγιο, βάζοντας αριθμούς στα κουτιά ή κυκλώνοντας τις κατάλληλες εναλλακτικές απαντήσεις

Ηλικία	Χρόνια			
Φύλο	Άνδρας	1	Γυναίκα	2
Σχολή	Ανθρωπιστικές Επιστήμες	1	Φυσικές Επιστήμες	2
	Ιατρική	3	Άλλη	4
Επίπεδο σπουδών	Προπτυχιακό	1	Μεταπτυχιακό για Master	2
	Μεταπτυχιακό για Doctora	3	Άλλο	4
Ύψος (σε cm)			Βάρος (σε Kg)	
Καπνίζετε;	Ναι	1	Όχι	2
Αν καπνίζετε πόσα τσιγάρα καπνίζετε την ημέρα;				
Ομάδα αίματος				

Εκτελούμε τις ακόλουθες εντολές για να εισάγουμε τα δεδομένα 32 απαντήσεων των 32 ερωτωμένων:

SPSS 18,0

Type in data

OK

Οθόνη: Untitled – SPSS Data Editor

Επιλέγουμε: κάτω αριστερά Variable View

Εισάγουμε τις μεταβλητές: case, Name, Age, Sex, Faculty, Bloodtyp, Levelst.

Η Οθόνη: Variable View, έχει τη ακόλουθη μορφή:

Name	***Type***	***Width***	***Decimals***	***Label***	***Values***	***Missing***	***Columns***	***Align***	***Measure***
Case	Numeric	9	0	Case number	none	none	8	Right	Scale
Name	String	15	0	Name of resp.	None	None	8	Left	Nominal
Age	Numeric								
Sex	Numeric				1=Male 2=Female	None	8	Right	Scale
Faculty	Numeric				1=Ανθρωπιστικές 2=Φυσικές 3=Ιατρική 4=Άλλο	None	8	Right	Scale
Bloodtyp	String	20	0		None	None	8	Left	Nominal
Levelst	Numeric				1=Προπτυχ. 2=Msc Μεταπτ. 3=Ph.D Μεταπτ. 4=Άλλο	None	8	Right	Scale
Weight	Numeric	9	2						Scale
Height	Numeric	9	2						Scale
Smoker	Numeric		1=NAI 2=OXI						Scale
Npday	Numeric								Scale

Επιλέγουμε: Data View στο κάτω αριστερό μέρος της οθόνης: Untitled – SPSS Data Editor

Οι απαντήσεις του ερωτηματολογίου έχουν τη μορφή που φαίνεται στον πίνακα 1.20.1

Τα δεδομένα υπάρχουν στο CD ΣΥΝΟΛΑ ΔΕΔΟΜΕΝΩΝ και στο φάκελο ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ 1 (ΚΕΦ. 1, 1. 20)

Πίνακας 1.20.1 *Δεδομένα Ερωτηματολογίου*

Case	Name	Age	Sex	Faculty	Levelst	Height	Weight	Smoker	Npday	Bloodtyp
01	George	18	1	1	1	178	75	1	15	O
02	Mary	17	2	2	2	196	100	2	16	O
03	Chris	19	1	3	3	145	60	1	17	A
04	Jenny	22	2	4	1	170	71	1	20	O
05	Alex	23	1	4	2	180	80	1	21	B
06	John	20	1	3	3	175	69	2	22	O
07	Harry	17	1	2	4	185	78	1	20	AB
08	Charoula	19	2	1	4	190	90	2	19	A
09	Mary	25	2	1	4	183	70	2	18	O
10	Kathy	24	2	1	3	182	85	2	20	B
11	Peter	22	1	3	3	170	72	1	20	A
12	Kathrin	20	2	4	3	160	77	1	20	O
13	Paul	26	1	4	2	170	95	1	26	O
14	Georgia	22	2	4	2	172	68	1	28	AB
15	Bill	21	1	2	1	190	120	1	21	B
16	Jimmy	19	1	1	2	180	75	2	14	O
17	Panos	19	1	2	2	163	60	1	10	O
18	Alexandros	19	2	3	3	142	51	2	9	O
19	Spyros	20	1	4	1	150	55	2	10	A
20	Antony	20	1	1	2	165	64	2	9	O
21	Antonia	20	2	2	3	160	53	1	30	A
22	Andy	21	2	3	2	175	50	1	32	O
23	Peter	22	1	3	3	182	72	1	20	O
24	George	23	1	3	1	169	65	1	20	B
25	Paul	22	1	2	2	162	62	1	20	O
26	Vana	22	2	1	3	182	80	1	19	B
27	Mishel	22	2	4	4	165	67	1	21	O
28	Christine	22	2	4	4	171	50	2	18	A
29	Michael	21	1	1	3	146	55	2	22	O
30	Nafsika	21	2	2	1	151	48	1	21	AB
31	Steven	21	1	2	2	164	59	2	20	O
32	Stephania	20	2	3	2	176	71	1	20	B

Όταν ολοκληρωθεί στην οθόνη Data View η συμπλήρωση των στηλών των μεταβλητών, αποθηκεύουμε το ερωτηματολόγιο ως εξής:

File

Save as…

Οθόνη: Save Data as

Αποθήκευση σε: SPSS

Γράφουμε: Όνομα αρχείου: Exercise questionnaire SPSS

Επιλέγουμε: Αποθήκευση,

File, Exit

1.20.Α Έλεγχος Υποθέσεων – Διαστήματα εμπιστοσύνης με βάση τα δεδομένα ερωτηματολογίου. Εφαρμογή στο SPSS

1.21 Έλεγχος (t) για το μέσο δείγματος που ακολουθεί την Κανονική κατανομή

Με βάση τα δεδομένα του ερωτηματολογίου του υποκεφαλαίου 1.20 για τους καπνίζοντες να γίνει έλεγχος της υπόθεσης

$$H_0: \mu = 160$$

έναντι της

$$H_1: \mu \neq 160,$$

για τη μεταβλητή height of respondent στο επίπεδο σημαντικότητας α = 0,01.

Εκτελούμε τις εντολές

SPSS 18,00 for WINDOWS

C:\Program files\SPSS\Exercise questionnaire SPSS

OK

Analyze

Compare Means

One-sample T Test

Οθόνη: One-Sample T Test

Θέτουμε τη μεταβλητή height of respondent στο κουτί Test Variable(s)

Επιλέγουμε: Test Value: 160

Options

Confidence interval: 95

Exclude cases Analysis by Analysis

Continue

Οθόνη: One-Sample T Test

OK

Το αποτέλεσμα παρουσιάζονται παρακάτω. Η υπόθεση H_0 απορρίπτεται, γιατί η p-value, κάτω από την επικεφαλίδα Sig. (2-tailed), είναι ίση με 0,000<0,01.

T Τεστ

Πίνακας 1.21.1

One-Sample Test				
	N	Mean	Std. deviation	Std. error mean
Height of respondent	32	170,2813	13,6789	2,4181

Πίνακας 1.21.2

One-Sample						
	Test value = 160					
	t	df	Sg. (2-tailed)	Mean difference	95% confidence interval of the difference	
					Lower	Upper
Height of respondent	4,252	31	0,000	10,2813	5,3495	15,2130

1.22. Διάστημα εμπιστοσύνης 95% για το μέσο πληθυσμό. Εφαρμογή στο SPSS

Δίνονται τα δεδομένα του ερωτηματολογίου του κεφαλαίου 1.20 για τους καπνίζοντες φοιτητές. Να βρεθεί ένα διάστημα εμπιστοσύνης 95% για το μέσο μ του πληθυσμού των φοιτητών που καπνίζουν.

Εκτελούμε τις εντολές:

Analyze

Descriptive Statistics

Explore…

Οθόνη: Explore

Μεταφέρουμε τη μεταβλητή age στο κουτί Dependent List

Επιλέγουμε: Statistics (όχι την εντολή Statistics στο κουτί Display)

Οθόνη: Explore: Statistics

Επιλέγουμε: Descriptives

Confidence Interval for Mean: 95

Continue

Οθόνη: Explore

OK

Το αποτέλεσμα φαίνονται στον πίνακα 1.22.1

Πίνακας 1.22.1 *διάστημα εμπιστοσύνης 95% για το μέσο της μεταβλητής age.*

Descriptives				
			Statistic	*Std. error*
Age of respondent	Mean		20,9063	0,36577
	95% confidence interval for mean	Lower bound	20,1603	
		Upper bound	21,6522	
	5% trimmed mean		20,8611	
	Median		21,0000	
	Variance		4,281	
	Std. deviation		2,06912	
	Minimum		17	
	Maximum		26,00	
	Range		9,00	
	Interquartile range		2,7500	
	Skewness		0,297	0,414
	Kurtosis		0,323	0,809

Υποθέτουμε ότι θέλουμε να ελέγξουμε την υπόθεση H_0: μ = 23 στο επίπεδο σημαντικότητας α = 0,05. Το διάστημα εμπιστοσύνης 95% για το μέσο είναι (20,1603, 21,6522). Το διάστημα αυτό δεν περιέχει την τιμή 23. Άρα απορρίπτουμε την υπόθεση H_0: μ = 23 στο επίπεδο σημαντικότητας 0,05.

1.23. Έλεγχος ισότητας των μέσων δύο ανεξάρτητων δειγμάτων. Εύρεση διαστήματος εμπιστοσύνης για τη διαφορά των μέσων δύο ανεξάρτητων δειγμάτων. Εφαρμογή στο SPSS

Τα δεδομένα υπάρχουν στο CD ΣΥΝΟΛΑ ΔΕΔΟΜΕΝΩΝ και στο φάκελο ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ 1Α (ΚΕΦ. 1, 1. 23)

(Παράδειγμα από το βιβλίο *Στατιστική Επιχειρήσεων,* Χρ. Φράγκος, Εκδ Σταμούλη, 1998, σελ.429)

Στο μάθημα Οικονομίας του πρώτου έτους ενός Τριτοβάθμιου Εκπαιδευτικού Ιδρύματος μερικοί σπουδαστές υποστηρίζουν ότι οι άνδρες έχουν μεγαλύτερη γνώση του Χρηματιστηρίου από τις γυναίκες. Ο Καθηγητής δίνει μία σειρά ερωτήσε-

ων για τη μέτρηση του βαθμού γνώσης του Χρηματιστηρίου σε δύο ανεξάρτητα τυχαία δείγματα από 15 άνδρες και 15 γυναίκες αντίστοιχα. Τα αποτελέσματα της εξέτασης είναι τα ακόλουθα:

Γυναίκες: 73 96 74 55 91 50 46 82 43 79 79 50 46 81 83

Άνδρες: 57 78 42 44 91 65 63 60 97 85 92 42 86 81 64

Να ελεγχθεί η υπόθεση

$$H_{0:}\ \mu_1 = \mu_2 \qquad \text{έναντι της} \qquad H_1 : \mu_1 \neq \mu_2$$

στο επίπεδο σημαντικότητας α = 0,05, όπου σ_1^2, σ_2^2 είναι οι διακυμάνσεις του πληθυσμού των ανδρών φοιτητών και των γυναικών φοιτητριών αντίστοιχα του παραπάνω Τριτοβάθμιου Εκπαιδευτικού Ιδρύματος.

Στην οθόνη Variable View εισάγουμε τη μεταβλητή sex, με τιμή 1 για τους άνδρες φοιτητές και τιμή 2 για τις γυναίκες φοιτήτριες, και τη μεταβλητή score, με τιμή τους βαθμούς των δοκιμαζόμενων φοιτητών στην εξέταση του Καθηγητή.

Στη οθόνη Data View εισάγουμε τα δεδομένα για τα γυναίκες και τους άνδρες ως ακολούθως:

sex	***score***	***sex***	***score***	***sex***	***score***	***sex***	***score***	***sex***	***score***
2	73	2	79	1	44	1	86	2	96
2	79	1	91	1	81	2	74	2	50
1	65	1	64	2	55	2	46	1	63
2	91	2	81	1	60	2	50	2	83
1	97	2	46	1	57	1	85	2	82
1	78	1	92	2	43	1	42	1	42

Εκτελούμε τις εντολές:

Analyze

Compare Means

Independent – Samples T Test

Οθόνη: Independent – Samples T Test

Μεταφέρουμε τη μεταβλητή score of student στο κουτί Test Variable(s) και τη μεταβλητή sex στο κουτί Grouping Variable

Επιλέγουμε: Define Groups…

Οθόνη: Define Groups

Επιλέγουμε: Group 1 1(= άνδρες)

Group 2 2(= γυναίκες)

Continue

Οθόνη: Independent – Samples T Test

Επιλέγουμε: Options

Οθόνη: Independent – Samples T Test-Options

Επιλέγουμε: Confidence Interval: 95

Continue

Οθόνη: Independent – Samples T Test

Επιλέγουμε OK

Τα αποτελέσματα φαίνονται στους πίνακες 1.23.1 ώς 1.23.4

T Test

Πίνακας 1.23.1

Group Statistics					
	Sex of student	*N*	*Mean*	*Std. deviation*	*Std. Error mean*
Score of student in the test	Άνδρας φοιτητής	15	69,80	18,789	4,851
	Γυναίκα φοιτήτρια	15	68,53	18,150	4,686

Πίνακας 1.23.2. *Έλεγχος Διακυμάνσεων με το κριτήριο Levene*

Independent Samples Test			
		Levene's test for Equality of variances	
		F statistic	Sig.
Score of student in the test	Equal Variances Assumed	0,000	0,993
	Equal Variances not assumed		

Πίνακας 1.23.3. *Έλεγχος ισότητας των μέσων δύο ανεξαρτήτων δειγμάτων*

Independent Samples Test					
		T test for equality of means			
		t	*df*	*Sig. (2-tailed)*	*Mean deference*
Score of student in the test	Equal variances assumed	0,188	28	0,852	1,27
	Equal variances not assumed	0,188	27,967	0,852	1,27

Πίνακας 1.23.4 *Διάστημα εμπιστοσύνης 95% για τη διαφορά των δύο μέσων*

Independent Samples Test				
		T test for equality of means		
		Std. error difference	*95% confidence interval of the difference*	
			Lower	*Upper*
Score of student in test	Equal variances assumed	6,745	-12,550	15,083
	Equal variances not assumed	6,745	-12,551	15,084

Τα συμπεράσματα είναι τα ακόλουθα:

1. Το πλαίσιο Levene's test for the equality of variances δείχνει ότι η p-value του ελέγχου είναι 0,993. Ισχύει α = 0,05. Άρα 0,993 > 0,05. Συνεπώς δεν απορρίπτεται η υπόθεση $H_0 : \sigma_1^2 = \sigma_2^2$.

2. Το διάστημα εμπιστοσύνης 95% για τη διαφορά των δύο μέσων είναι -12,550< $\mu_1 - \mu_2$ < 15,083. Το διάστημα αυτό περιλαμβάνει την τιμή 0. Άρα δεν απορρίπτεται, στο επίπεδο σημαντικότητας α= 0,05, η υπόθεση $H_{0:}$ $\mu_1 - \mu_2 = 0$ ή $\mu_1 = \mu_2$.

3. Το ίδιο με το σημείο (2) συμπέρασμα εξάγεται από την p-value του ελέγχου Τ. Ισχύει p-value = 0,852 > 0,05. Άρα δεν απορρίπτεται η υπόθεση H_0: $\mu_1 = \mu_2$ στο επίπεδο σημαντικότητας α = 0,05.

1.24. Έλεγχος (t) για δύο Ανεξάρτητα Δείγματα. Εφαρμογή στο SPSS

Τα δεδομένα υπάρχουν στο CD ΣΥΝΟΛΑ ΔΕΔΟΜΕΝΩΝ και στο φάκελο ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ 1Γ (ΚΕΦ. 1, 1. 24)

Ένας ερευνητής των ιδιοτήτων των ανθρώπινων ματιών θέλει να διαπιστώσει κατά πόσο αναγνωρίζεται ταχύτερα μια λέξη στο δεξιό μέρος μιας σελίδας παρά το αριστερό μέρος. Ο ερευνητής έγραψε, λοιπόν, μια λέξη στο αριστερό μέρος και μια λέξη στο δεξιό κλειστό πλαίσιο μιας σελίδας. Στη συνέχεια προσκάλεσε 10 άτομα να αναγνωρίσουν τη λέξη στο δεξιό μέρος της σελίδας και 10 διαφορετικά άτομα να αναγνωρίσουν τη λέξη στο αριστερό μέρος της σελίδας. Τα αποτελέσματα φαίνονται στο ακόλουθο πίνακα:

Πίνακας 1.24.1 *Χρόνοι αναγνώρισης λέξης σε milliseconds στο αριστερό και στο δεξιό πεδίο.*

Περίπτωση	***Αριστερό πεδίο***	***Περίπτωση***	***Δεξιό πεδίο***
1	500	11	392
2	513	12	445
3	300	13	271
4	561	14	523
5	483	15	421
6	502	16	489
7	539	17	501
8	467	18	388
9	420	19	411
10	480	20	467

Για να εισάγουμε τα δεδομένα στο SPSS, στην οθόνη SPSS Data editor: Variable View, διαλέγουμε τις μεταβλητές ΠΕΡΙΠΤΩΣ, ΠΕΔΙΟ(1 = αριστερό οπτικό πεδίο, 2 = δεξιό οπτικό πεδίο) και ΧΡΟΝΟΣΑΝ (= χρόνος αναγνώρισης λέξης).

Ο παρακάτω πίνακας είναι μέρος της Οθόνης Variable view:

	Name	Type	Width	Deci mals	Label	Values	Miss-ing	Col-umns	Align
1	ΠΕΡΙΠΤΩΣ	Nu-meric	9	0	Αριθμός περίπτωσης	None	None	8	Right
2	ΠΕΔΙΟ	Nu-meric	9	0	Οπτικό πεδίο	1,Αριστερό οπτικό πεδίο 2,Δεξιό οπτικό πεδίο	None	8	Right
3	ΧΡΟΝΟΣΑΝ	Nu-meric	9	1	Χρόνος αναγνώρισης λέξης	None	None	8	Right

Η μεταβλητή ΠΕΔΙΟ είναι η μεταβλητή ομαδοποίησης (grouping variable), η οποία είναι ανεξάρτητη μεταβλητή. Η μεταβλητή ΧΡΟΝΟΣΑΝ είναι η εξαρτημένη μεταβλητή.

Στην οθόνη Data View γράφουμε για κάθε περίπτωση τον αριθμό της περίπτωσης (ΠΕΡΙΠΤΩΣ), τον αριθμό του πεδίου(ΠΕΔΙΟ) και το χρόνο αναγνώρισης λέξης.(ΧΡΟΝΟΣΑΝ).

Τα δεδομένα έχουν ως εξής:

ΠΕΡΙΠΤΩΣ	**ΠΕΔΙΟ**	**ΧΡΟΝΟ-ΣΑΝ**	**ΠΕΡΙΠΤΩΣ**	**ΠΕΔΙΟ**	**ΧΡΟ-ΝΟΣΑΝ**
1	1	500	11	2	392
2	1	513	12	2	445
3	1	300	13	2	271
4	1	561	14	2	523
5	1	483	15	2	421
6	1	502	16	2	489
7	1	539	17	2	501
8	1	467	18	2	388
9	1	420	19	2	411
10	1	480	20	2	467

Πριν εκτελεστεί ο έλεγχος βρίσκονται οι *απομακρυσμένες τιμές* των δεδομένων (αν υπάρχουν) ή οι *μη συμμετρικές κατανομές* των δεδομένων και γίνεται η ανάλογη διόρθωση.

Εκτελούνται οι εντολές:

Analyze

Descriptive Statistics

Explore

Οθόνη: Explore

Μεταφέρουμε τον μεταβλητή ΧΡΟΝΟΣΑΝ στο κουτί Dependent List και τη μεταβλητή ΠΕΔΙΟ στο κουτί Factor List.

Επιλέγουμε: Plots…

Οθόνη: Explore Plots

Αποεπιλέγουμε: Stem-and-Leaf

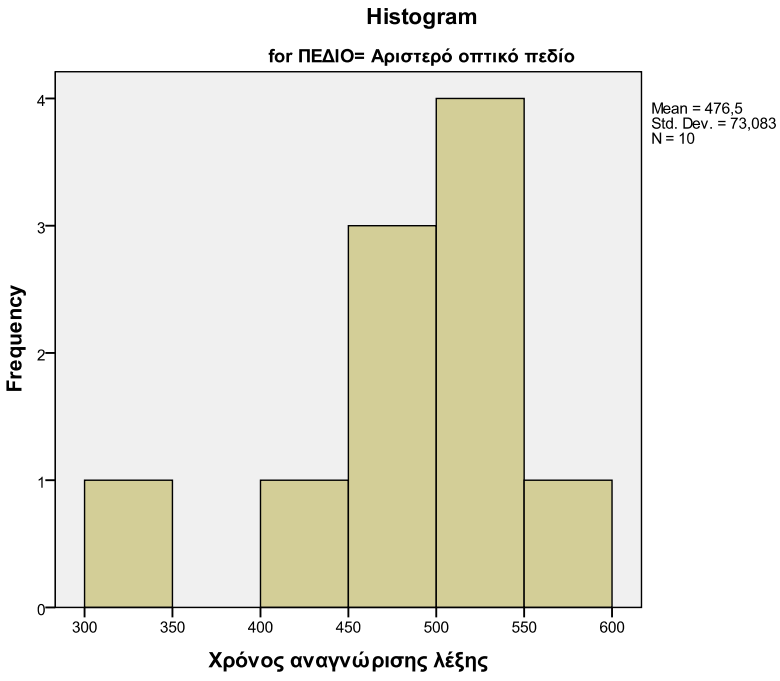
Επιλέγουμε: Histogram

Continue

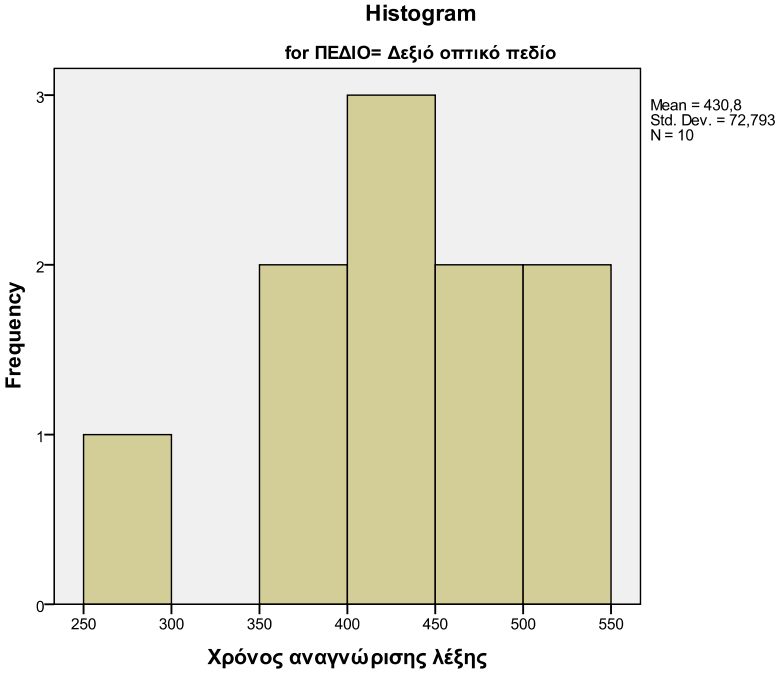
Οθόνη: Explore

OK

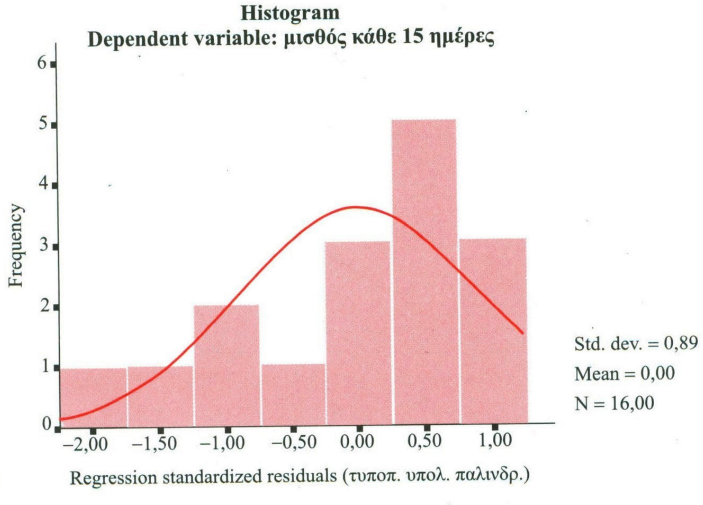
Τα αποτελέσματα είναι προκαταρτικά διερευνητικά (exploratory) και φαίνονται στα γραφήματα 1.24.2 έως 1.24.4



Γράφημα 1.24.2



Γράφημα 1.24.3



Γράφημα 1.24.4

Το **boxplot (κιβωτιοδιάγραμμα)** για το αριστερό πεδίο δείχνει ότι η παρατήρηση (3) (= 300) είναι **απομακρυσμένη (outlier).**

Για να είναι μια παρατήρηση απομακρυσμένη πρέπει να απέχει περισσότερο από 3 μήκη του boxplot από το κάτω μέρος (βάση) του boxplot, προκειμένου για αριστερή απομακρυσμένη παρατήρηση, ή περισσότερο από 3 μήκη του boxplot από το πάνω μέρος (οροφή) του boxplot, όταν πρόκειται για δεξιά απομακρυσμένη παρατήρηση,

Υπενθυμίζεται ότι:

Μήκος boxplot = 75ο εκατοστιαίο σημείο – 25ο εκατοστιαίο σημείο

Γραμμή στο μέσο του boxplot = *διάμεσος* (= 50ό εκατοστιαίο σημείο)

Κάτω βάση του στελέχους του boxplot: ⊥ = μικρότερη τιμή που δεν είναι outlier

Άνω βάση του στελέχους του boxplot: ⊤ = μεγαλύτερη τιμή που δεν είναι outlier.

Αν εξετασθούν τα δύο κιβωτιοδραγράμματα, θα γίνει φανερό ότι η παρατήρηση 13 (= 271) είναι ακραία παρατήρηση, η οποία βρίσκεται σε μακρυνή απόσταση από το κύριο σώμα των ιστογραμμάτων.

Για να διορθωθούν τα δεδομένα πρέπει να απομακρυνθούν οι παρατηρήσεις 3 (=300) και 13 (=271).Για να γίνει η διόρθωση αυτή, εκτελούνται οι εντολές:

Data

Select Cases…

Οθόνη: Select Cases

Επιλέγουμε: If condition in satisfied

If

Οθόνη: Select Cases: If

Μεταφέρουμε τη μεταβλητή rectime στο πάνω δεξιό κουτί της υποθετικής έκφρασης

Γράφουμε την έκφραση > 300

Επιλέγουμε: Continue

Οθόνη: select Cases

OK

Μια ματιά στην οθόνη Data View αποκαλύπτει ότι οι παρατηρήσεις 3 και 13 έχουν αποεπιλεγεί (filter $ με τιμή 0). Τώρα μπορούμε να προχωρήσουμε στο έλεγχο Τ.

Εκτελούμε τις εντολές:

Analyze

Compare Means

Independent Samples T Test…

Οθόνη: Independent Samples T Test

Μεταφέρουμε τη μεταβλητή word recognition στο κουτί Test Variable(s)

Μεταφέρουμε τη μεταβλητή ΠΕΔΙΟ στο κουτί Grouping Variable

Επιλέγουμε: Define Groups

Οθόνη: Define Groups

Use specified values

Group 1: Γράφουμε 1

Group 2: Γράφουμε 2

Continue

Οθόνη: Independent samples T Test

OK

Τα αποτελέσματα φαίνονται στους πίνακες 1.24.5 έως 1.24.8

Πίνακας 1.24.5 *Group Statistics*

Group Statistics					
	Οπτικό πεδίο	*N*	*Mean*	*Std. deviation*	*Std. error mean*
Χρόνος αναγνώρισης λέξης	Αριστερό πεδίο	9	496,111	41,0135	13,6712
	Δεξιό πεδίο	9	448,556	49,1378	16,3793

Πίνακας 1.24.6 *Independent samples Test*

Independent samples Test			
		Levene's test for equality of variances	
		F	*Sig*
Χρόνος αναγνώρισης λέξης (in milliseconds)	Equal variances assumed	0,997	0,333
	Equal variances not assumed		

Πίνακας 1.24.7 *Independent Samples Test*

Independent Samples Test					
		T test equality of means			
		t	*df*	*Sig. (2-tailed)*	*Mean difference*
Χρόνος αναγνώρισης λέξης (in milliseconds)	Equal variances assumed	2,229	16	0,040	47,556
	Equal variances not assumed	2,229	15,504	0,041	47,556

Πίνακας 1.24.8 *Independent Samples Test*

Independent Samples Test				
		T Test for equality of means		
		Sid. error difference	*95% confidence interval of the difference*	
			Lower	*Upper*
Χρόνος αναγνώρισης λέξης (in milliseconds)	Equal variances assumed	21,3350	2,3274	92,7837
	Equal variances not assumed	21,3350	2,2096	92,9015

Μία περιληπτική στατιστική ερμηνεία των αποτελεσμάτων είναι η ακόλουθη:

1. Το στατιστικό του Levene έχει τιμή F ίση με 0,997 και p-value ίση με 0,333. Υπενθυμίζουμε ότι

p-value = P (t > τιμή στατιστικού του ελέγχου)

Αν η p-value είναι μικρότερη του επιπέδου σημαντικότητας (α), τότε απορρίπτουμε την υπόθεση H_0, όταν πρόκειται για μονόπλευρο έλεγχο.

Αν η p-value είναι μικρότερη του $\frac{\alpha}{2}$, τότε απορρίπτουμε την υπόθεση H_0, όταν πρόκειται για δίπλευρο έλεγχο.

Άρα δεχόμαστε την υπόθεση H_0, ότι οι δύο διακυμάνσεις είναι ίσιες. Συνεπώς μπορούμε να εφαρμόσουμε τον κατάλληλο τύπο για τον έλεγχο Τ όταν οι διακυμάνσεις είναι ίσες αλλά άγνωστες.

2. Η τιμή t είναι 2,23 με 16 βαθμούς ελευθερίας. Η p-value για ένα δίπλευρο έλεγχο είναι 0,04.

υπόθεση H_0: $\mu_1 = \mu_2$
έναντι H_1: $\mu_; \neq \mu_2$

3. Το διάστημα εμπιστοσύνης 95% για τη διαφορά των δύο μέσων είναι [2,33, 2,78]. Το διάστημα αυτό δεν περιέχει την τιμή 0, άρα οι δύο μέσοι δεν είναι ίσοι, στο επίπεδο σημαντικότητας 0.05.

4. Παράλληλα με τον έλεγχο Τ, έγινε και έλεγχος διασπορών

υπόθεση $H_0 : \sigma_1^2 = \sigma_2^2$

έναντι $H_0 : \sigma_1^2 \neq \sigma_2^2$

Με την βοήθεια του στατιστικού του Levene, συμπεραίνεται ότι η υπόθεση $H_0 : \sigma_1^2 = \sigma_2^2$ δεν απορρίπτεται γιατί η p-value = 0,333 > 0,05 = επίπεδο σημαντικότητας.

5. Αν διατηρηθούν οι παρατηρήσεις 3 και 13 στα δεδομένα και γίνει ό έλεγχος Τ, τα αποτελέσματα θα δίνουν p-value > 0,05, πράγμα που σημαίνει ότι δεν απορρίπτεται η υπόθεση H_0: $\mu_1 = \mu_2$ στο επίπεδο σημαντικότητας 0,05 και ότι ισχύει $\mu_1 = \mu_2$.

Για περισσότερα σχόλια βλέπε *Στατιστική Επιχειρήσεων,* Χρ. Φράγκος, Εκδόσεις Σταμούλη 1998.

1.25 Έλεγχος (t) για δύο συσχετισμένα δείγματα. Εφαρμογή στο SPSS

Τα δεδομένα υπάρχουν στο CD ΣΥΝΟΛΑ ΔΕΔΟΜΕΝΩΝ και στο φάκελο ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ 1Δ (ΚΕΦ. 1, 1. 25)

Πίνακας 1.25.1 *Χρόνοι (σε milliseconds) αναγνώρισης λέξης στο αριστερό και στο δεξιό πεδίο μιας οθόνης Η/Υ.*

Περίπτωση	***Αριστερό πεδίο***	***Δεξιό πεδίο***
1	323	304
2	512	493
3	502	491
4	385	365
5	453	426
6	343	320
7	543	523
8	440	442
9	682	580
10	590	564

Ένας ερευνητής μετρά το χρόνο που απαιτείται για να αναγνωρίζουν 10 πρόσωπα μια λέξη που εμφανίζεται αφενός στο δεξιό οπτικό πεδίο μιας οθόνης Η/Υ και α-

φετέρου στο αριστερό οπτικό πεδίο της ίδιας οθόνης του Η/Υ. Σε αναλογία με την άσκηση της παραγράφου 1.24. (ανεξάρτητα δείγματα) ο πίνακας 1.25.1 δίνει τα δεδομένα.

Εισάγουμε τα δεδομένα στο SPSS ως εξής:

Ανοίγουμε την οθόνη Variable View

Δημιουργούμε τις μεταβλητές: ΠΕΡΙΠΤΩΣ, ΑΡΙΣΤΠΕΔ, ΔΕΞΠΕΔ με πλήρη ονομασία « Αριστερό οπτικό πεδίο» και «Δεξιό οπτικό πεδίο».

Εισάγουμε τα δεδομένα στην οθόνη Data View

Διερευνητική ανάλυση δεδομένων (exploratory data analysis)

Για να ελέγχουν τα δεδομένα ως προς την ύπαρξη ανωμαλιών πριν εκτελεσθεί το T-Test, κατασκευάζουνε το *διάγραμμα νέφους* (scatter plot)

Εκτελούμε τις εντολές:

Graphs

Scatter…

Οθόνη: Scatterplot

Επιλέγουμε: Simple

Define

Οθόνη: Simple Scatterplot

Μεταφέρουμε τη μεταβλητή left (ΑΡΙΣΤΠΕΔ) στο κουτί Y Axis και τη μεταβλητή (ΔΕΞΠΕΔ) στο κουτί X Axis

Επιλέγουμε: Titles…

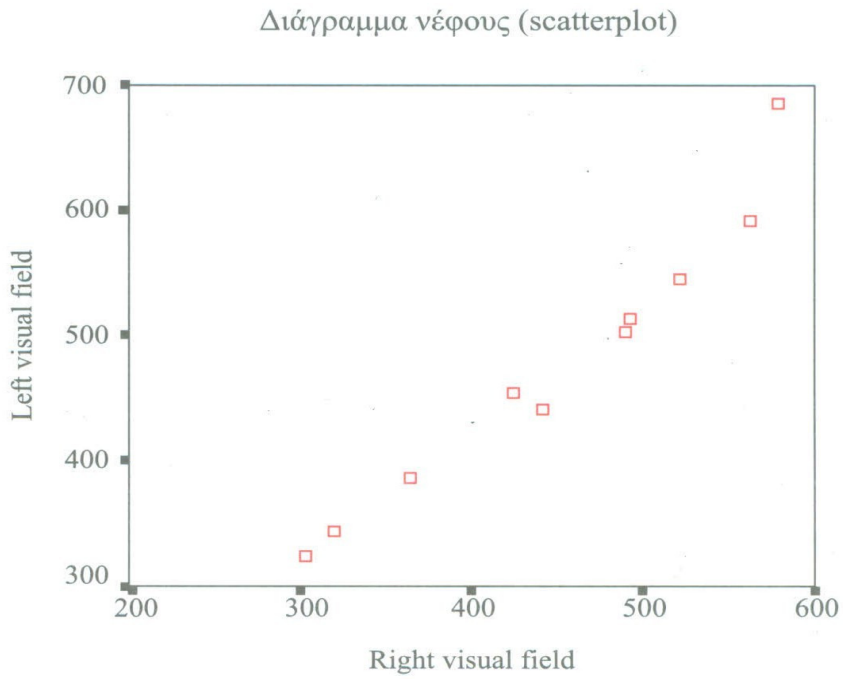
Θέτουμε τίτλο: Διάγραμμα νέφους (scatterplot)

Continue

Οθόνη: Simple Scatterplot

Επιλέγουμε: OK

Τα αποτελέσματα φαίνονται στο γράφημα 1.25.2



Γράφημα 1.25.2

Από τα αποτελέσματα παρατηρούμε ότι δεν υπάρχει απομακρυσμένο ζεύγος τιμών (x, y).

Η παρουσία απομακρύσμένων ζευγών έχει ως συνέπεια τη μείωση της τιμής του στατιστικού (t), γιατί το στατιστικό αυτό είναι ένα κλάσμα με παρανομαστή την τετραγωνική ρίζα της διακύμανσης, η οποία αυξάνεται όταν υπάρχουν απομακρυσμένες παρατηρήσεις.

Επειδή υπάρχουν απομακρυσμένες παρατηρήσεις, διατηρούμε τα δεδομένα όπως είναι και εκτελούμε τον έλεγχο Τ για συσχετισμένα δείγματα, εκτελώντας τις εντολές:

Analyze

Compare Means

Οθόνη: Paired Samples T Test

Επιλέγουμε τις μεταβλητές (ΑΡΙΣΤΠΕΔ) και (ΔΕΞΠΕΔ)

Τα ονόματα των επιλεγμένων μεταβλητών εμφανίζονται στο κουτί Paired Variables

Επιλέγουμε OK

Τα αποτελέσματα φαίνονται στον πίνακα 1.25.3. ώς 1.25.6

T-Test

Πίνακας 1.25.3 *Paired Samples Statistics*

Paired Samples Statistics					
		Mean	***N***	***Std. deviation***	***Std error mean.***
Pair 1	Αριστ. οπτ. πεδίο	447,30	10	112,091	35,446
	Δεξιό οπτ. πεδίο	450,80	10	97,085	30,701

Πίνακας 1.25.4 *Paired Samples Correlations*

Paired Samples Correlations				
		N	***Correlation***	***Sig.***
Pair 1	Αριστ. οπτ. πεδίο Δεξιό οπτ. πεδίο	10	0,975	0,000

Πίνακας 1.25.5 *Paired Samples Test*

Paired Samples Test							
		Paired differences					*t*
		Mean	*Std. deviation*	*Std. error mean*	*95% confidence interval of difference*		
					Lower	*Upper*	
Pair 1	Αριστ. οπτ. πεδίο Δεξιό οπτ. πεδίο	26,50	27,814	8,7960	6,60	46,40	3,013

Πίνακας 1.25.6 *Paired Samples Test*

Paired Samples Test			
		df	*Sig. (2-tailed)*
Pair 1	Αριστ. οπτ. πεδίο Δεξιό οπτ. πεδίο	9	0,015

Τα στατιστικά συμπεράσματα είναι τα εξής:

1. Ο συντελεστής συσχέτισης είναι 0,97.
2. Το διάστημα εμπιστοσύνης 95% είναι (6,60, 40,60). Το διάστημα αυτό δεν περιλαμβάνει την τιμή 0, άρα απορρίπτουμε την υπόθεση H_0 ότι οι μέσοι των δύο δειγμάτων είναι ίσοι.
3. Η τιμή του στατιστικού (t) με 9 βαθμούς ελευθερίας είναι ίση με 3,01 και η p-value (πιθανότητα P) για ένα δίπλευρο έλεγχο είναι 0,015. Η p-value για μονόπλευρο έλεγχο είναι 0,015/2 = 0,008. Επειδή ισχύει: p-value = 0,015 < 0,05 = επίπεδο σημαντικότητας, απορρίπτεται η υπόθεση H_0 ότι οι μέσοι των δύο δειγμάτων είναι ίσοι στο επίπεδο σημαντικότητας 0,05, όπως και στο επίπεδο σημαντικότητας 0,01.