ISBN 978-960-456-511-5

Φωτοστοιχειοθεσία **Π. ΖΗΤΗ & Σια Ι.Κ.Ε.**
Εκτύπωση 18ο χλμ Θεσσαλονίκης - Περαίας
Βιβλιοδεσία Τ.Θ. 4171 • Περαία Θεσσαλονίκης • Τ.Κ. 570 19
Τηλ.: 2392.072.222 - Fax: 2392.072.229 • e-mail: info@ziti.gr

ΕΚΔΟΣΕΙΣ ΖΗΤΗ

www.ziti.gr

ΒΙΒΛΙΟΠΩΛΕΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ:
Αρμενοπούλου 27 - 546 35 Θεσσαλονίκη • Τηλ.: 2310-203.720 • Fax 2310-211.305
e-mail: sales@ziti.gr

ΒΙΒΛΙΟΠΩΛΕΙΟ ΑΘΗΝΩΝ:
Χαριλάου Τρικούπη 22 - Τ.Κ. 106 79, Αθήνα • Τηλ.-Fax: 210-3816.650
e-mail: athina@ziti.gr

ΗΛΕΚΤΡΟΝΙΚΟ ΒΙΒΛΙΟΠΩΛΕΙΟ: www.ziti.gr

Πρόλογος

Τα περισσότερα παραδείγματα είναι αντιπροσωπευτικά πραγματικών προβλημάτων που συναντώνται σε πειραματικές επιστήμες. Η εφαρμογή των τεχνικών της στατιστικής ανάλυσης σε μεγάλα σύνολα δεδομένων προϋποθέτει τη χρήση εξειδικευμένων εφαρμογών λογισμικού.

Η γλώσσα R, ελεύθερο λογισμικό ανοικτού κώδικα για τη στατιστική επεξεργασία θεωρείται σήμερα ένα από τα εργαλεία λογισμικού με την μεγαλύτερη ζήτηση στην αγορά εργασίας. Συνοδεύεται από περισσότερα από 13000 πακέτα επέκτασης και αξιοποιείται σε πολλούς επιστημονικούς τομείς και σε εταιρείες - κολοσσούς. Στην παρούσα έκδοση, υπάρχουν παραδείγματα με πραγματικά δεδομένα, από την Ελληνική Στατιστική Υπηρεσία, κ.α., που επιλύονται χρησιμοποιώντας την γλώσσα ανοικτού κώδικά R.

Με αφορμή τη συγγραφή της τρίτης έκδοσης του βιβλίου, κατασκευάσαμε την βιβλιοθήκη *gginference*, η οποία είναι η πρώτη βιβλιοθήκη στην επίσημη ιστοσελίδα του CRAN της R (Comprehensive R Archive Network- http://cran.r-project.org) που παρουσιάζει γραφικά τα αποτελέσματα των ελέγχων στατιστικών υποθέσεων στατιστικής συμπερασματολογίας. Περιέχει ακόμα σύνολα δεδομένων που χρησιμοποιούνται στις ασκήσεις του βιβλίου.

Το βιβλίο αυτό αποτελείται από εννέα κεφάλαια που το καθένα περιλαμβάνει θεωρία, εφαρμογές, προτεινόμενες ασκήσεις, παραδείγματα στατιστικής ανάλυσης πραγματικών δεδομένων με χρήση της γλώσσας R , και συνοδεύεται από τυπολόγιο.

Για την επίλυση των ασκήσεων χρησιμοποιήθηκε η 3.5.1 έκδοση της R και 1.1.456 του RStudio, IDE, το οποίο είναι ένα ολοκληρωμένο περιβάλλον ανάπτυξης για καλύτερη διαχείριση και εκτέλεση κώδικα.

Στους ιστοτόπους https://www.r-project.org και https://www.rstudio.com, υπάρχουν σχετικά αρχεία και σαφείς οδηγίες εγκατάστασης της R και του RStudio αντίστοιχα.

Στο τέλος του βιβλίου, υπάρχει μια συλλογή με τίτλο "Γενικές Ασκήσεις" για εξοικείωση του αναγνώστη με απλά θέματα ανάλυσης δεδομένων και μια συλλογή

στατιστικών πινάκων που θεωρούνται απαραίτητοι για τη λύση των ασκήσεων, καθώς και δυο σύντομους οδηγούς της γλώσσας R.

Τέλος ευχαριστούμε τον κ. Κλεάνθη Κουπίδη για την πολύτιμη βοήθεια του στη συγγραφή και επιμέλεια του κειμένου και των ασκήσεων στα κεφάλαια της R και τις εκδόσεις Ζήτη για την άψογη συνεργασία και για την προσεγμένη έκδοση του παρόντος συγγράμματος.

Νοέβριος 2018

Φ. Κολυβά - Μαχαίρα
Ε. Μπόρα - Σέντα
Χ. Μπράτσας

Περιεχόμενα

Κεφάλαιο 1: **Στοιχεία Πιθανοτήτων**

Κεφάλαιο 2: **Τυχαίες Μεταβλητές – Κατανομές**

Κεφάλαιο 3: **Περιγραφική Στατιστική**

Κεφάλαιο 4: **Εκτιμητική**

Πίνακας Συντμήσεων

α.δ.	ανάλυση διασποράς
β.ε.	βαθμοί ελευθερίας
δ.ε.	διάστημα εμπιστοσύνης
ε.τ.	ελαχίστων τετραγώνων
Κ.Ο.Θ.	Κεντρικό οριακό θεώρημα
π.σ.	ποσοστιαίο σημείο
σ.α.κ.	συνάρτηση αθροιστικής κατανομής
σ.π.π.	συνάρτηση πυκνότητας πιθανότητας
σ.κ.	συνάρτηση κατανομής
σ.π.	συνάρτηση πιθανότητας
στ.σ.	στατιστική συνάρτηση
σ.σ.	στάθμη σημαντικότητας
τ.δ.	τυχαίο δείγμα
τ.μ.	τυχαία μεταβλητή

Κεφάλαιο 1

ΣΤΟΙΧΕΙΑ ΠΙΘΑΝΟΤΗΤΩΝ

1.1 Εισαγωγή

Η Στατιστική είναι μια εφαρμοσμένη μαθηματική επιστήμη που σκοπό έχει να βοηθήσει στη μελέτη και κατανόηση των φαινομένων ή των ιδιοτήτων των πληθυσμών, χρησιμοποιώντας τις πληροφορίες που δίνει ένα τυχαία επιλεγμένο μέρος μόνο του πληθυσμού ή του φαινομένου (**δείγμα** - sample). Επειδή ούτε η μελέτη του συνόλου του πληθυσμού ούτε η εξ ολοκλήρου παρακολούθηση της εξέλιξης ενός φαινομένου είναι δυνατή, καταφεύγουμε στο **πείραμα** (experiment) αν πρόκειται για μελέτη φαινομένου ή στη **δειγματοληψία** (sampling) αν πρόκειται για πληθυσμό. Για να είναι αξιόπιστα τα συμπεράσματα, θα πρέπει το δείγμα να είναι τυχαίο και αντιπροσωπευτικό του πληθυσμού από τον οποίο προέρχεται.

Ορισμός 1.1

Όλα τα δυνατά αποτελέσματα ενός πειράματος αποτελούν το δειγματοχώρο (sample space) που συμβολίζεται με S ή με Ω. Κάθε δυνατό αποτέλεσμα του πειράματος, δηλαδή κάθε σημείο του δειγματοχώρου, λέγεται **απλό γεγονός** ή **ενδεχόμενο** (simple event). Οι δειγματοχώροι που έχουν πεπερασμένο ή αριθμήσιμο πλήθος σημείων λέγονται **διακριτοί** (discrete), ενώ αυτοί που έχουν μη αριθμήσιμο πλήθος στοιχείων λέγονται **μη διακριτοί** ή **συνεχείς** (continuous).

Π.χ. ο αριθμός των παιδιών σε μια οικογένεια είναι ένα απλό γεγονός ενός διακριτού δειγματοχώρου, ενώ το ύψος των ατόμων δημιουργεί ένα συνεχή δειγματοχώρο.

Ορισμός 1.2

Κάθε διαδικασία που εκτελείται ή παρατηρείται και στην οποία το αποτέλεσμα είναι τυχαίο, ονομάζεται **πείραμα τύχης** (random enperiment).

Ορισμός 1.3

Κάθε δείγμα το οποίο επιλέγεται με τέτοιο τρόπο ώστε οποιοδήποτε άλλο δείγμα του ίδιου μεγέθους να έχει την ίδια πιθανότητα να επιλεγεί, λέγεται **τυχαίο** (random sample).

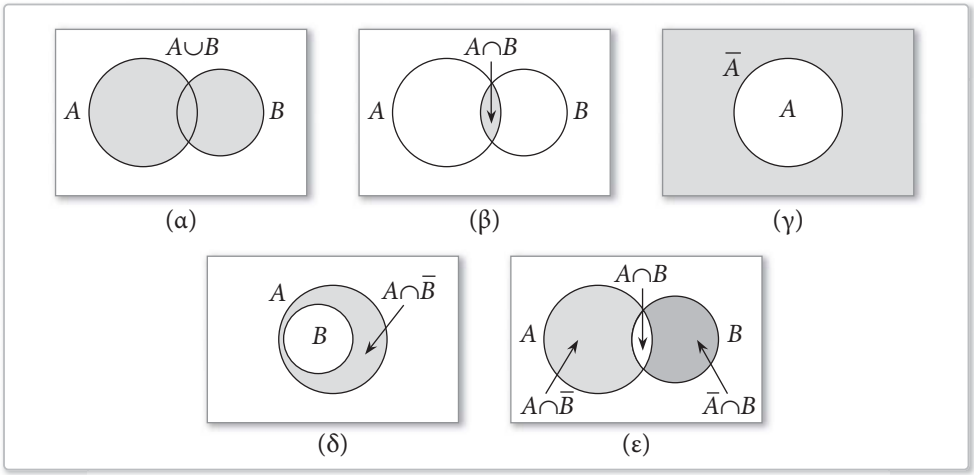
Για να μελετηθούν οι ιδιότητες ή τα φαινόμενα, θα πρέπει να εκφρασθούν μαθηματικά, ώστε να γίνουν μαθηματικά προβλήματα τα οποία θα επιλυθούν και θα δώσουν τα αποτελέσματα. Έτσι λοιπόν θα πρέπει να υπάρχει μια αμφιμονοσήμαντη αντιστοιχία μεταξύ των ιδιοτήτων ή των φαινομένων και κάποιων μαθηματικών εκφράσεων. Η απλούστερη αντιστοιχία επιτυγχάνεται με τη βοήθεια της **θεωρίας συνόλων**.

Η αντιστοιχία μεταξύ των συνόλων, των γεγονότων και των πράξεών τους, δίνεται στον παρακάτω πίνακα.

Γεγονότα	Σύνολα
δειγματοχώρος S ή Ω (βέβαιο γεγονός)	σύνολο αναφοράς S ή Ω
αδύνατο γεγονός	σύνολο $\varnothing$
απλό γεγονός	σύνολο A
δεν συμβαίνει το γεγονός A	σύνολο $\Gamma = \overline{A} = S - A$
τα γεγονότα A και B συμβαίνουν ταυτόχρονα	σύνολο $\Gamma = A \cap B = AB$
τουλάχιστον ένα από τα γεγονότα A, B συμβαίνει	σύνολο $\Gamma = A \cup B = A + B$

(Στη θεωρία των πιθανοτήτων χάριν απλότητας, η ένωση συνόλων $A \cup B$ συμβολίζεται με $A + B$ και η τομή τους $A \cap B$ συμβολίζεται με AB).

Οι πράξεις μεταξύ των συνόλων δίνονται με τα παρακάτω διαγράμματα:



Σχήμα 1.1

Ορισμός 1.4

Δύο γεγονότα A και B ονομάζονται **ασυμβίβαστα** ή **ξένα** όταν η πραγματοποίηση του ενός γεγονότος αποκλείει την πραγματοποίηση του άλλου. Αυτό σημαίνει ότι:

$$A, B \text{ ασυμβίβαστα} \Leftrightarrow A \cap B = AB = \varnothing$$

Π.χ. το να γεννηθεί αγόρι ή κορίτσι είναι δύο γεγονότα ασυμβίβαστα.

Ορισμός 1.5

Δύο γεγονότα A και B λέγονται (**στοχαστικά**) **ανεξάρτητα** (stochastically indepentent) όταν η πραγματοποίηση του γεγονότος A δεν επηρεάζει την πραγματοποίηση του γεγονότος B και αντίστροφα.

Π.χ. το φύλο του πρώτου παιδιού είναι ανεξάρτητο από το φύλο του δεύτερου παιδιού σε μια οικογένεια. Πρέπει να σημειωθεί ότι δύο γεγονότα που είναι ασυμβίβαστα δεν είναι αναγκαστικά και ανεξάρτητα όπως και δύο ανεξάρτητα γεγονότα δεν είναι αναγκαστικά και ασυμβίβαστα.

Ένας από τους στόχους της θεωρίας των πιθανοτήτων είναι ο υπολογισμός της πιθανότητας με την οποία συμβαίνουν τα διάφορα γεγονότα. Έχουν δοθεί διάφοροι ορισμοί της πιθανότητας ενός γεγονότος. Επικρατέστεροι είναι οι δύο παρακάτω:

Ορισμός 1.6: Η πιθανότητα σαν όριο της σχετικής συχνότητας

Αν στις n επαναλήψεις ενός πειράματος ένα γεγονός A εμφανίσθηκε n_A φορές, τότε το πηλίκο $f_A = n_A / n$ ονομάζεται (**σχετική**) **συχνότητα** του γεγονότος A. Όσο το n μεγαλώνει τόσο η σχετική συχνότητα σταθεροποιείται γύρω από έναν αριθμό. Το όριο της σχετικής συχνότητας του $n \to \infty$ ονομάζεται **πιθανότητα του γεγονότος *A*** και συμβολίζεται με $P(A)$.

Π.χ. αν θέλουμε την πιθανότητα να γεννηθεί κορίτσι, τότε το πείραμα που θα μας βοηθήσει να την υπολογίσουμε είναι να καταγράψουμε το φύλο του νεογέννητου σε μία σειρά γεννήσεων. Πρόσφατα παρατηρήθηκε ότι στα 1000 παιδιά που γεννήθηκαν, τα 489 ήταν κορίτσια. Έτσι σύμφωνα με τον ορισμό 1.6 η ζητούμενη πιθανότητα είναι

$$P = \frac{n_A}{n} = \frac{489}{1000} = 0,489 .$$

1.4 Δειγματοληψία

Όταν έχουμε n στοιχεία και θέλουμε να πάρουμε από αυτά ένα δείγμα μεγέθους r, μπορούμε να το πραγματοποιήσουμε με τους εξής τρόπους:

i) Παίρνουμε ένα-ένα στοιχείο, το εξετάζουμε και το επανατοποθετούμε εκεί από όπου το πήραμε πριν πάρουμε το επόμενο στοιχείο.
Συνεχίζουμε αυτήν τη διαδικασία μέχρι να πάρουμε r στοιχεία.
Στην περίπτωση αυτή δείγματα που αποτελούνται από τα ίδια στοιχεία τα οποία όμως πάρθηκαν με διαφορετική σειρά θεωρούνται διαφορετικά. Η δειγματοληψία αυτή ονομάζεται **δειγματοληψία με επανάθεση** (sampling with replacement) και υπάρχουν n^r τέτοια δείγματα.

ii) Παίρνουμε ένα-ένα στοιχείο, το εξετάζουμε και **δεν** το επανατοποθετούμε εκεί απ' όπου το πήραμε. Συνεχίζουμε μέχρι να πάρουμε r στοιχεία. Όπως και προηγουμένως επειδή και εδώ τα στοιχεία λαμβάνονται ένα-ένα, διαφορετική διάταξη ορίζει διαφορετικά δείγματα στα οποία όμως το ίδιο στοιχείο εμφανίζεται μία μόνο φορά. Η δειγματοληψία αυτή ονομάζεται **δειγματοληψία χωρίς επανάθεση** (sampling without replacement) και υπάρχουν $(n)_r = n(n-1)\dots(n-r+1)$ τέτοια δείγματα.

iii) Παίρνουμε r στοιχεία μαζί. Στην περίπτωση αυτή ούτε διάταξη μπορεί να ορισθεί ούτε το ίδιο στοιχείο να εμφανιστεί περισσότερες από μία φορές σε κάθε δείγμα. Το πλήθος τέτοιων δειγμάτων είναι: $\binom{n}{r}$.

iv) Αν έχουμε δειγματοληψία με επανάθεση, το κάθε στοιχείο που παίρνουμε το εξετάζουμε ως προς το είδος του, το επανατοποθετούμε αλλά στο τελικό δείγμα μεγέθους r που φτιάχνουμε δε μας ενδιαφέρει η διάταξη των στοιχείων, τότε υπάρχουν $\mathscr{E}_n^r$ τέτοια δείγματα.

Συνοπτικά, στις περιπτώσεις **i)** και **iv)** μπορεί να εμφανιστεί στο δείγμα το ίδιο στοιχείο μέχρι r φορές ενώ στις **(ii)** και **(iii)** όλα τα στοιχεία του δείγματος είναι διαφορετικά.

Αν στις περιπτώσεις **(i)** και **(ii)** δεν εξετάζουμε το στοιχείο τη στιγμή που το παίρνουμε αλλά εξετάζουμε τα r στοιχεία στο τέλος της δειγματοληψίας, τότε είναι όπως οι περιπτώσεις **(iv)** και **(iii)** αντίστοιχα.

Οι μεταθέσεις, διατάξεις κ.λπ. βοηθούν πολύ στον υπολογισμό των ευνοϊκών και των δυνατών περιπτώσεων οι οποίες χρειάζονται για να βρεθεί η πιθανότητα ενός γεγονότος θεωρούνται δε από τα πιο δύσκολα μαθηματικά προβλήματα.

1.5 Εφαρμογές – Λυμένες Ασκήσεις

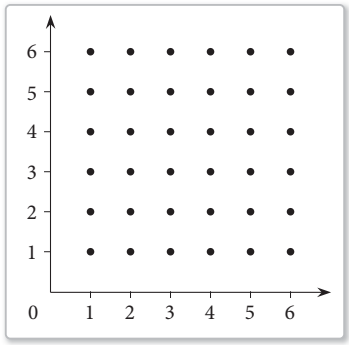
Άσκηση 1.1

Ρίχνονται δύο ζάρια. Να παρασταθεί γραφικά ο δειγματοχώρος των 36 αποτελεσμάτων σ' ένα σύστημα ορθογωνίων καρτεσιανών συντεταγμένων. Με τη βοήθεια αυτού να δοθούν τα αποτελέσματα και το πλήθος για τα παρακάτω ενδεχόμενα:

A ={Το άθροισμα να είναι διαιρετό δια 4}
B = {Και οι δύο αριθμοί να είναι άρτιοι}
C = {Οι αριθμοί να είναι ίσοι}
D = {Οι αριθμοί να διαφέρουν τουλάχιστον κατά 4}
$E = A \cap B$, $C \cup D$, $B - A$, $\overline{A \cup B}$

Λύση

$\Omega = \{1,1),\ (1,2),\ \ldots,(1,6),\ (2,1),\ \ldots,(6,6)\}$



$A = \{(1,3),\ (3,1),\ (2,2),\ (3,5),\ (4,4),\ (5,3),\ (2,6),\ (6,2),\ (6,6)\}$, $n_A = 9$

$B = \{(2,2),\ (2,4),\ (2,6),\ (4,2),\ (4,4),\ (4,6),\ (6,2),\ (6,4),\ (6,6)\}$, $n_B = 9$

$C = \{1,1),\ (2,2),\ (3,3),\ (4,4),\ (5,5),\ (6,6)\}$, $n_C = 6$

$D = \{1,5),\ (1,6),\ (2,6),\ (5,1),\ (6,1),\ (6,2)\}$, $n_D = 6$

$A \cap B = \{(2,2),\ (2,6),\ (4,4),\ (6,2),\ (6,6)\}$, $n_{A\cap B} = 5$

$C \cup D = \{(1,1),\ \ldots,(6,6),\ (1,5),\ (1,6),\ (2,6),\ (5,1),\ (6,1),\ (6,2)\}$ $n_{C\cup D} = 12$

$B - A = \{(2,4),\ (4,2),\ (4,6),\ (6,4)\}$, $n_{B-A} = 4$

$\overline{A \cup B} = \{(1,1),\ (1,2),\ (1,4),\ (1,5),\ (1,6),\ (2,1),\ (2,3),\ (2,5),$
$(3,2),\ (3,3),\ (3,4),\ (3,6),\ (4,1),\ (4,3),\ (4,5),$
$(5,1),\ (5,2),\ (5,4),\ (5,5),\ (5,6),\ (6,1),\ (6,3),\ (6,5)\}$, $n_{\overline{A\cup B}} = 23$.

1.6 Στοιχεία Πιθανοτήτων με χρήση της R

1.6.1 Βασικές εντολές στοιχειών πιθανοτήτων στην R

Για την επίλυση βασικών ασκήσεων πιθανοτήτων στην R χρησιμοποιούνται απλές πράξεις καθώς και συναρτήσεις συνδυαστικής αντικειμένων. Επιπλέον, η βιβλιοθήκη LablacesDemon παρέχει συνάρτηση υπολογισμού του τύπου Bayes.

Πίνακας-R 3.1
Βασικές εντολές συνδυαστικής στην R

Εντολή στην R	Περιγραφή	Βιβλιοθήκη
permutations(n,r) *πλήθος: nrow(permutations)*	Διατάξεις χωρίς επανάληψη	gtools
permutations(n,r,repeats.allowed=T) *πλήθος: nrow(permutations)*	Διατάξεις με επανάληψη	gtools
choose(n,r)	Συνδυασμοί	base R
*factorial(n) / (factorial(r1) * factorial(r2)*…. *factorial(rk))*	Μεταθέσεις με επανάληψη	base R

Πίνακας-R 3.2
Συνάρτηση Bayes στην R

Εντολή στην R	Μέτρο	Βιβλιοθήκη
BayesTheorem(PrA, PrBA)	Bayes	LaplacesDemon

1.6.2 Εφαρμογές - Λυμένες ασκήσεις

Άσκηση-R 1.1

Έστω ότι έχουμε ένα κουτί που περιέχει μια κόκκινη “red”, μια μπλε “blue”, και μια άσπρη “white” μπάλα. Εάν πάρουμε δυο μπάλες, να βρεθούν:

a) Ποιοι και πόσοι είναι οι δυνατοί συνδυασμοί με επανάθεση;

b) Ποιοι και πόσοι είναι οι δυνατοί συνδυασμοί χωρίς επανάθεση;

Έστω ότι προσθέτουμε στο κουτί μπάλες και από τα τρία χρώματα, έτσι ώστε να περιέχει 10 κόκκινες, 8 μπλε και 6 άσπρες. Ποιο είναι το πλήθος των δυάδων που μπορούμε να πάρουμε με το ίδιο χρώμα;

Λύση

a) Οι διατάξεις ή μεταθέσεις με επανάθεση, υπολογίζονται χρησιμοποιώντας την εντολή *permutations*, της βιβλιοθήκης gtool, στην οποία πρέπει να ορίσουμε το διάνυσμα τιμών στην παράμετρο *v*, το μέγεθος του διανύσματος και των συνδυασμών στις παραμέτρους *n*, *r* αντίστοιχα καθώς και την παράμετρο repeats.allowed= TRUE για να προσδιορίσουμε την επανάθεση.

```
#εγκατάσταση της βιβλιοθήκης (εφόσον δεν έχει γίνει ξανά)
#install.packages('gtools')
```

```
#Φόρτωση της βιβλιοθήκης
library(gtools)
```

Δημιουργούμε ένα διάνυσμα x, με τις 3 μπάλες που περιέχονται στο κουτί

```
x <- c('red', 'blue', 'black')
```

Για να βρούμε τους συνδυασμούς ανά 2 μπάλες από το κουτί με επανάθεση, εκτός από το διάνυσμα *x* που ορίσαμε προηγουμένως, θέτουμε την παράμετρο *r* = 2 και την repeats.allowed = TRUE για να καθορίσουμε δύο συνδυασμούς με επανάθεση ως εξής:

```
comb <- permutations(n = 3,
                     r = 2,
                     v = x,
                     repeats.allowed = TRUE)
```

Οι δυνατές μεταθέσεις με επανάθεση είναι:

```
comb
##       [,1]    [,2]
##  [1,] "black" "black"
##  [2,] "black" "blue"
##  [3,] "black" "red"
##  [4,] "blue"  "black"
##  [5,] "blue"  "blue"
##  [6,] "blue"  "red"
##  [7,] "red"   "black"
##  [8,] "red"   "blue"
##  [9,] "red"   "red"
```

Ο αριθμός των μεταθέσεων είναι:

```
nrow(comb)
## [1] 9
```

b) Οι διατάξεις ή μεταθέσεις χωρίς επανάθεση, υπολογίζονται παρόμοια με την ίδια

συνάρτηση (*permutations*), ορίζοντας όμως την παράμετρο repeats.allowed = FALSE.

Για να βρούμε τις μεταθέσεις χωρίς επανάθεση, ορίζουμε τις παραμέτρους τις συνάρτησης *permutations* όπως στο προηγούμενο ερώτημα, αλλά σε αυτή την περίπτωση που δεν έχουμε επανάθεση θα αλλάξουμε την παράμετρο repeats.allowed σε FALSE, δηλαδή:

```
comb2 <- permutations(n = 3,
                      r = 2,
                      v = x,
                      repeats.allowed = FALSE)
```

Οι μεταθέσεις χωρίς επανάθεση είναι οι εξής:

```
comb2
##      [,1]    [,2]
## [1,] "black" "blue"
## [2,] "black" "red"
## [3,] "blue"  "black"
## [4,] "blue"  "red"
## [5,] "red"   "black"
## [6,] "red"   "blue"
```

Ο αριθμός των δυνατών μεταθέσεων είναι:

```
nrow(comb2)
## [1] 6
```

c) Ο υπολογισμός των συνδυασμών n αντικειμένων ανά k γίνεται χρησιμοποιώντας την εντολή *choose*, στην οποία ορίζουμε αντίστοιχα τις παραμέτρους n και k.
Το πλήθος των δυάδων που μπορούμε να εξάγουμε από δέκα κόκκινες μπάλες είναι:

```
choose(n=10,k=2)
## [1] 45
```

Από οκτώ μπλε μπάλες, το πλήθος των δυάδων που μπορούμε να πάρουμε είναι:

```
choose(n=8,k=2)
## [1] 28
```

Ενώ, το πλήθος των δυάδων που μπορούμε να πάρουμε από έξι άσπρες μπάλες είναι:

```
choose(n=6,k=2)
## [1] 15
```

Προτεινόμενες Ασκήσεις

1.51 Ποια είναι η πιθανότητα να κάνει κάποιος φουλ του άσσου, στο πόκερ; (Το πόκερ παίζεται με 52 φύλλα και φουλ του άσσου είναι να έχει κάποιος 3 άσσους και 2 οποιαδήποτε άλλα φύλλα).

1.52 Σ' ένα τμήμα το 70% των φοιτητών είναι αγόρια και το 30% είναι κορίτσια. Από τα αγόρια το 40% παίρνει πτυχίο στα 8 εξάμηνα και το 60% σε περισσότερα από 8 εξάμηνα. Από τα κορίτσια το 45% παίρνει πτυχίο στα 8 εξάμηνα και το 55% παίρνει πτυχίο σε περισσότερο από 8 εξάμηνα.

α) Να βρεθεί η πιθανότητα ένας φοιτητής να πάρει πτυχίο σε 8 εξάμηνα.

β) Αν κάποιος πήρε πτυχίο σε περισσότερα από 8 εξάμηνα, ποια η πιθανότητα να είναι αγόρι και ποια η πιθανότητα να είναι κορίτσι;

1.53 Το ποσοστό αυτών που πάσχουν από AIDS είναι 0,5 % Ένα test δίνει σωστή διάγνωση με πιθανότητα 80% για τους υγιείς και 98% για τους ασθενείς. Να υπολογισθεί η πιθανότητα λανθασμένης διάγνωσης σε ένα άτομο φορέα του AIDS και η πιθανότητα λανθασμένης διάγνωσης γενικά.

1.54 Μια Χριστουγεννιάτικη γιρλάντα αποτελείται από 50 λαμπάκια συνδεδεμένα σε σειρά: αυτό σημαίνει ότι η γιρλάντα ανάβει όταν όλα τα λαμπάκια ανάβουν. Η πιθανότητα να είναι χαλασμένο ένα λαμπάκι είναι 1%. Να υπολογισθεί η πιθανότητα να ανάψει η γιρλάντα.

1.55 Για τα γεγονότα A, B, Γ ισχύουν:

$$P(A)=P(B)=P(\Gamma)=\frac{1}{5}, \quad P(A\cap B\cap\Gamma)=\frac{1}{25},$$

$$P(A\cap B)=P(A\cap\Gamma)=P(B\cap\Gamma)=\frac{1}{25}.$$

Να δειχθεί ότι τα γεγονότα A, B, Γ είναι ανεξάρτητα ανά ζεύγη αλλά δεν είναι ανεξάρτητα μεταξύ τους..

1.56 Σε μια παρέα 15 ατόμων, να βρεθεί η πιθανότητα δύο τουλάχιστον άτομα να ανήκουν στο ίδιο ζώδιο.

1.57 Τα γεγονότα A και B είναι ανεξάρτητα και ισχύει $B\subset A$. Να βρεθεί η πιθανότητα του γεγονότος A.

3.7 Περιγραφική Στατιστική με χρήση της R

3.7.1 Βασικές εντολές περιγραφικής στατιστικής στην R

Η R παρέχει στους χρήστες μια ποικιλία συναρτήσεων για τον υπολογισμό των περιγραφικών στατιστικών μέτρων των μεταβλητών ενός συνόλου δεδομένων.

Πίνακας-R 3.1

Βασικές εντολές περιγραφικής στατιστικής ποιοτικών δεδομένων στην R

Μέτρο	Εντολή στην R	Βιβλιοθήκη
Συχνότητα	table	base R
Σχετικές συχνότητες	prop.table	base R
Επικρατούσα τιμή	Mode	DescTools

Πίνακας-R 3.2

Βασικές εντολές περιγραφικής στατιστικής ποσοτικών δεδομένων στην R

	Μέτρο	Εντολή στην R	Βιβλιοθήκη
Κεντρικής τάσης	Μέση Τιμή	mean	base R
	Διάμεσος	median	stats
	Εκατοστημόρια	quantile	stats
	Μέγιστη τιμή	max	base R
	Ελάχιστη τιμή	min	base R
Διασποράς	Εύρος	range ή Range	base R ή DescTools
	Ενδοτεταρτημοριακό εύρος	IQR	stats
	Διακύμανση/διασπορά	var	stats
	Τυπική απόκλιση	sd	stats
Κατανομής	Συντελεστής μεταβολής	CV	DescriptiveStats.OBeu
	Λοξότητα	ds.skewness	DescriptiveStats.OBeu
	Κύρτωση	ds.kurtosis	DescriptiveStats.OBeu

Ανάλογα με το είδος της μεταβλητής χρησιμοποιούμε τις συναρτήσεις του Πίνακα-R 3.1 (ποιοτικές μεταβλητές) ή αυτές του Πίνακα-R 3.2 (ποσοτικές μεταβλητές).

Επιπλέον των στατιστικών μέτρων, η R προσφέρει στον χρήστη μια συλλογή από γραφήματα για την καλύτερη κατανόηση των δεδομένων. Στον Πίνακα-R 3.3 παραθέτουμε τα κυριότερα γραφήματα για την παρουσίαση των στατιστικών μέτρων περιγραφικής στατιστικής.

Πίνακας-R 3.3

Εντολές βασικών γραφημάτων περιγραφικής στατιστικής στην R

Γράφημα		Εντολή στην R	Βιβλιοθήκη
Ποιοτικές	Ραβδόγραμμα	barplot	graphics
	Κυκλικό διάγραμμα	Pie	graphics
Ποιοτικές	Ιστόγραμμα	Hist	graphics
	Θηκόγραμμα	boxplot	graphics
	Διασποράς	plot	graphics

Για την ερμηνεία και τον ορισμό των παραμέτρων των εντολών που παραθέτουν οι πίνακες, μπορείτε να συμβουλευτείτε τη βοήθεια της R. Στις ασκήσεις που ακολουθούν γίνεται χρήση όλων των εντολών με στόχο την κατανόηση της χρήσης τους σε πραγματικά δεδομένα της ΕΛΣΤΑΤ[1].

3.7.2 Εφαρμογές – Λυμένες ασκήσεις

Άσκηση-R 3.1

Η Ελληνική Στατιστική Υπηρεσία παρέχει τα δημογραφικά δεδομένα των γεννήσεων και των θανάτων στην Ελλάδα από το 1932 ως το 2016, ως δυο αρχεία σε επεξεργάσιμη μορφή ανοικτών δεδομένων. Ενοποιημένα, τα δεδομένα βρίσκονται στην βιβλιοθήκη gginference, όπως περιγράφονται στον παρακάτω πίνακα:

Μεταβλητή	Περιγραφή	Είδος Μεταβλητής	Τιμές
Year	Χρονιά	Διακριτή	1932,1933,…2016
Births	Γεννήσεις	Διακριτή	117593,111447,…
BeathsRate	Γεννήσεις ανά 1000 κατοίκους	Συνεχής	18 ,16.8 15, ...
Deaths	Θάνατοι	Διακριτή	185523, 189583 ….
DeathsRate	Θάνατοι ανά 1000 κατοίκους	Συνεχής	28.4, 28.6, 31.1, …

Να υπολογισθούν τα μέτρα κεντρικής τάσης, τα μέτρα μεταβλητότητας και να περιγράφουν γραφικά τα δεδομένα.

[1] http://www.statistics.gr/

Λύση

Αρχικά, χρησιμοποιούμε την εντολή της R, *str*, για να δούμε τις βασικές λεπτομέρειες για την δομή των δεδομένων μας. Είναι το πρώτο βήμα που πρέπει να κάνουμε σε κάθε δοθέν σύνολο δεδομένων, έτσι ώστε να αναγνωρίσουμε και να διορθώσουμε πιθανά λάθη, αν υπάρχουν, στον ορισμό μεταβλητών, πριν συνεχίσουμε στο στάδιο της ανάλυσης.

```
# Εγκατάσταση της βιβλιοθήκης (εφόσον δεν έχει γίνει ξανά)
# install.packages('gginference')

# Φόρτωση της βιβλιοθήκης
library(gginference)

# Προβολή των πρώτων παρατηρήσεων
head(BirthDeath)
##   Year Deaths DeathsRate Births BirthRate
## 1 1932 117593      17.97 185523     28.35
## 2 1933 111447      16.82 189583     28.62
## 3 1934 100651      14.96 208929     31.06
## 4 1935 101416      14.83 192511     28.16
## 5 1936 105005      15.14 193343     27.87
## 6 1937 105674      15.04 183878     26.16

# Δομή του BirthDeath
str(BirthDeath)
## 'data.frame':    71 obs. of  5 variables:
##  $ Year      : int  1932 1933 1934 1935 1936 1937 1938 ...
##  $ Deaths    : int  117593 111447 100651 101416 105005 ...
##  $ DeathsRate: num  18 16.8 15 14.8 15.1 ...
##  $ Births    : int  185523 189583 208929 192511 193343 ...
##  $ BirthRate : num  28.4 28.6 31.1 28.2 27.9 ...
```

Από τα αποτελέσματα παρατηρούμε ότι οι μεταβλητές του συνόλου δεδομένων της ΕΛΣΤΑΤ είναι σωστά ορισμένες στο πλαίσιο δεδομένων *BirthBeath* και είναι όλες ποσοτικές. Από την περιγραφή των μεταβλητών γίνεται κατανοητό ότι αρκεί να μελετήσουμε μόνο δυο μεταβλητές. Οι μεταβλητές αυτές είναι η *DeathsRate* και η *BirthsRate*, των οποίων οι τιμές, είναι οι τιμές των μεταβλητών *Deaths* και *Births* αντίστοιχα ανά 1000 κατοίκους. Επομένως, για τις δυο αυτές μεταβλητές θα υπολογισθούν τα μέτρα κεντρικής τάσης, τα μέτρα μεταβλητότητας και θα περιγράφουν γραφικά με ιστογράμματα και θηκογράμματα.

a) Μέτρα Κεντρικής Τάσης

Για τα μέτρα κεντρικής τάσης, πρέπει να υπολογίσουμε την επικρατούσα τιμή, τη διάμεσο και τη δειγματική μέση τιμή.

i) Για τον υπολογισμό της επικρατούσας τιμής, χρησιμοποιούμε την συνάρτηση *Mode* της βιβλιοθήκης DescTools.

```
# Εγκατάσταση της βιβλιοθήκης (εφόσον δεν έχει γίνει ξανά)
# install.packages('DescTools')

library(DescTools)
```

Η επικρατούσα τιμή των γεννήσεων ανά 1000 κατοίκους, είναι:

```
MBirthsR <- Mode(BirthDeath$BirthRate)
MBirthsR
## [1]  9.47  17.95
```

και των θανάτων ανά 1000 κατοίκους, είναι:

```
MDeathsR <- Mode(BirthDeath$DeathsRate)
MDeathsR
## [1] 7.88  9.46
```

Παρατηρούμε ότι και στις δύο μεταβλητές υπάρχουν δυο τιμές που έχουν την μεγαλύτερη συχνότητα εμφάνισης.

ii) Με την εντολή *median* της βιβλιοθήκης stats, υπολογίζουμε τη διάμεσο μιας μεταβλητής. Άρα, η διάμεσος των γεννήσεων ανά 1000 κατοίκους είναι:

```
mBirthsR <- median(BirthDeath$BirthRate)
mBirthsR
## [1] 14.49
```

και των θανάτων ανά 1000 κατοίκους:

```
mDeathsR <- median(BirthDeath$DeathsRate)
mDeathsR
## [1] 9.27
```

iii) Ο υπολογισμός της δειγματικής μέσης τιμής γίνεται με την συνάρτηση της R, *mean*, ως εξής:

Για τις γεννήσεις ανά 1000 κατοίκους:

```
moBirthsR <- mean(BirthDeath$BirthRate)
moBirthsR
## [1] 14.97
```

Και για τους θανάτους ανά 1000 κατοίκους:

```
moDeathsR <- mean(BirthDeath$DeathsRate)
moDeathsR
## [1] 9.727
```

5 Κεφάλαιο

ΕΛΕΓΧΟΙ ΥΠΟΘΕΣΕΩΝ

5.1 Εισαγωγή

Αντικείμενο αυτού του κεφαλαίου είναι οι έλεγχοι υποθέσεων που αναφέρονται στις παραμέτρους του πληθυσμού.

Για καλύτερη κατανόηση έστω το παρακάτω παράδειγμα.

Υπάρχει ένα φάρμακο με το οποίο εάν γίνει θεραπεία, εμφανίζονται τα πρώτα αποτελέσματα βελτίωσης, μέσα σε 10 ημέρες. Ανακαλύπτεται ένα καινούργιο φάρμακο για την ίδια ασθένεια και η φαρμακευτική εταιρεία ισχυρίζεται ότι φέρνει αποτελέσματα σε συντομότερο χρονικό διάστημα. Εάν λοιπόν ο μέσος χρόνος βελτίωσης για το πρώτο φάρμακο είναι $\mu=10$ ημέρες χρειάζεται να ελεγχθεί εάν πράγματι για το δεύτερο φάρμακο, ο μέσος χρόνος είναι μικρότερος, δηλαδή εάν $\mu<10$. Η διαδικασία που ακολουθείται λέγεται **δοκιμασία υποθέσεων** ή **έλεγχος υπόθεσης** ή απλά λέμε ότι κάνουμε ένα **στατιστικό έλεγχο (test)**.

Η υπόθεση ότι $\mu=10$ συμβολίζεται με H_0 και λέγεται **μηδενική υπόθεση**. Η υπόθεση $\mu<10$ η οποία ελέγχεται ως προς την H_0, λέγεται **εναλλακτική υπόθεση** και συμβολίζεται με H_1. Η απόφαση να γίνει δεκτή ή να απορριφθεί η H_0 στηρίζεται σε ένα κριτήριο το οποίο υπολογίζεται από τα δεδομένα του δείγματος. **Απορριπτική περιοχή** ή κρίσιμη περιοχή της H_0 ονομάζεται η περιοχή στα σημεία της οποίας η H_0 απορρίπτεται. Η απορριπτική περιοχή συμβολίζεται με R.

Τα στοιχεία ενός στατιστικού test είναι τα εξής:

1. Ορίζεται η μηδενική υπόθεση H_0.
2. Ορίζεται η εναλλακτική υπόθεση, H_1.
3. Ορίζεται το στατιστικό του ελέγχου από το δείγμα.
4. Ορίζεται η απορριπτική περιοχή R, της υπόθεσης H_0.
5. Εξάγονται τα συμπεράσματα.

Η μηδενική υπόθεση ενός ελέγχου συνήθως είναι η $H_0: \theta=\theta_0$ όπου θ η παρά-

μετρος του πληθυσμού που ελέγχεται και θ_0 μια συγκεκριμένη τιμή της. Στο παράδειγμα $\theta = \mu$ και $\theta_0 = 10$ δηλαδή $H_0 : \mu = 10$.

Η εναλλακτική υπόθεση μπορεί να είναι $\theta > \theta_0$ ή $\theta < \theta_0$ ή $\theta \neq \theta_0$.
Στις δύο πρώτες περιπτώσεις ο έλεγχος είναι μονόπλευρος ενώ στην τρίτη δίπλευρος. Στο παράδειγμα $H_1 : \mu < 10$.

5.2 Σφάλματα – Στάθμη σημαντικότητας

Σ' έναν έλεγχο υπόθεσης υπάρχει περίπτωση να γίνουν δύο ειδών σφάλματα.

Ορισμός 5.1

Ονομάζεται **σφάλμα τύπου I** η απόρριψη της μηδενικής υπόθεσης ενώ είναι σωστή. Η πιθανότητα αυτού του σφάλματος συμβολίζεται με **a** και ονομάζεται **στάθμη** ή **επίπεδο σημαντικότητας** (σ.σ.). Ισχύει a =P(απόρριψης της H_0 / H_0 σωστή).

Ορισμός 5.2

Ονομάζεται **σφάλμα τύπου II** η αποδοχή της H_0 ενώ είναι λάθος. Η πιθανότητα του σφάλματος τύπου II συμβολίζεται με $\boldsymbol{\beta}$. Ισχύει β =P(αποδοχής της H_0 / H_0 λάθος). Η ποσότητα $\gamma = 1 - \beta$ ονομάζεται **ισχύς** του ελέγχου.

Η απάντηση στο ερώτημα ποιο από τα δύο σφάλματα είναι σημαντικότερο είναι σχετική. Στο παράδειγμα, για την εταιρεία που παράγει το νέο φάρμακο και η οποία ενδιαφέρεται κυρίως για αποδοχή της εναλλακτικής υπόθεσης, το σφάλμα τύπου II είναι σημαντικότερο, ενώ για την εταιρεία που παράγει το παλιό φάρμακο, το σφάλμα τύπου I είναι το σημαντικό.

Κανονικά πρέπει να υπολογίζονται και το a και το β· ο υπολογισμός όμως του β είναι αρκετά δύσκολος.

Η σχέση μεταξύ των σφαλμάτων a και β φαίνεται καλύτερα στο παρακάτω παράδειγμα:

❖ **Παράδειγμα 5.1**

Έστω X η τ.μ. που μετρά τη δύναμη θραύσης μιας ατσάλινης ράβδου. Εάν η ατσάλινη ράβδος παραχθεί με τη μέθοδο I, τότε $X \sim N(50, 36)$. Μια καινούργια μέθοδος II που είναι υπό δοκιμή, πιστεύεται ότι δίνει $X \sim N(55, 36)$. Εάν δοθούν 16 ατσάλινες ράβδοι που φτιάχτηκαν με τη μέθοδο II, πώς θα μπορούσε να ελεγχθεί εάν η αύξηση της δύναμης θραύσης είναι πραγματική;

Λύση

Υπάρχουν δύο υποθέσεις: η $\mu = 50$ και η $\mu = 55$. Για να ελεγχθούν οι δύο αυτές υποθέσεις, ορίζεται ένας κανόνας που εξαρτάται από τις 16 τιμές του δείγματος: χωρίζεται ο δειγματοχώρος σε δύο μέρη έστω c και c' και εάν το δείγμα $(x_1, x_2, \ldots, x_{16}) \in c$ η $\mu = 50$ απορρίπτεται, ενώ αν $(x_1, x_2, \ldots, x_{16}) \in c'$ η $\mu = 50$ γίνεται δεκτή (δεν απορρίπτεται). Συνήθως αυτός ο χωρισμός του δειγματοχώρου γίνεται με το στατιστικό του ελέγχου που αναφέρθηκε στην προηγούμενη παράγραφο. Στην περίπτωση του παραδείγματος 5.1 θα μπορούσε να ορισθεί:

$$c = \{(x_1, x_2, \ldots, x_{16}) : \bar{x} > 53\} = \{\underset{\sim}{x} : \bar{x} > 53\} \quad \text{όπου} \quad \bar{X} \sim N\left(50, \frac{36}{16}\right)$$

όταν $\mu = 50$, ενώ $\bar{X} \sim N\left(55, \frac{36}{16}\right)$ όταν $\mu = 55$.

Τότε, τα σφάλματα τύπου I και II θα είναι:

$$\text{a} = P\ (\text{απόρριψης της } H_0 / \mu = 50) =$$

$$= P(\bar{X} > 53 / \mu = 50) = P\left(\frac{(\bar{X} - 50)4}{6} > \frac{(53 - 50)4}{6}\right) = 0{,}0228$$

ενώ

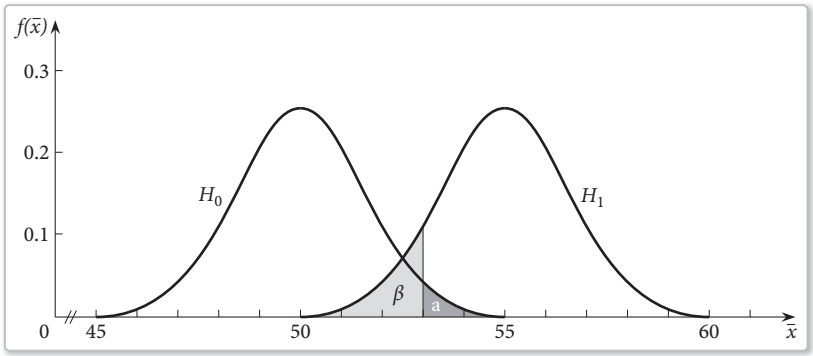
$$\beta = P\ (\text{αποδοχής της } H_0 / \mu = 55) =$$

$$= P(\bar{X} < 53 / \mu = 55) = P\left(\frac{(\bar{X} - 55)4}{6} < \frac{(53 - 55)4}{6}\right) = 0{,}0913$$

Τα σφάλματα a και β φαίνονται στο παρακάτω σχήμα 5.1. Παρατηρούμε ότι:

α) όσο το a μεγαλώνει τόσο το β μικραίνει και αντίθετα.

β) όσο μικρότερη είναι η διαφορά $|50 - 55|$ τόσο μεγαλύτερο είναι το β και



Σχήμα 5.1

γ) τέλος το a και το β μπορούν να ελαττωθούν όταν το μέγεθος του δείγματος αυξηθεί.

Εάν λοιπόν στόχος ήταν ο έλεγχος να έχει πιθανότητα σφάλματος τύπου Ι μέχρι 0,05, τότε επειδή αυτή η πιθανότητα που υπολογίστηκε από το συγκεκριμένο δείγμα είναι μικρότερη, η $\mu = 55$ γίνεται δεκτή.

Εάν ο έλεγχος έπρεπε να έχει πιθανότητα σφάλματος τύπου Ι μέχρι 0,01 τότε επειδή $0,0228 > 0,01$ η $\mu = 50$ δεν απορρίπτεται.

Στην πρώτη περίπτωση το β του ελέγχου θα ήταν μικρότερο του 0,0943 ενώ στη δεύτερη, μεγαλύτερο.

Γενικά, σχέση που να συνδέει τα a και β δεν υπάρχει. ▲

Σ' έναν έλεγχο δεν μπορούν όλες οι παραπάνω πιθανότητες συγχρόνως να ελαττωθούν. Συνήθως επιλέγεται η στάθμη σημαντικότητας a και από όλα τα κριτήρια που έχουν σ.σ. a επιλέγεται αυτό που δίνει μεγαλύτερη ισχύ γ. Το a συνήθως έχει μια μικρή τιμή όπως 0,10, 0,05, 0,01 και σπανίως μικρότερη. Ένας άλλος τρόπος αποδοχής ή απόρριψης της H_0 είναι να οριστεί το παρατηρούμενο επίπεδο σημαντικότητας του ελέγχου.

Ορισμός 5.3

Παρατηρούμενη στάθμη ή **επίπεδο σημαντικότητας** ή απλά **σημαντικότητα** ενός ελέγχου, ονομάζεται η πιθανότητα να παρατηρηθεί μια τιμή του στατιστικού μεγαλύτερη απ' αυτήν που έδωσε το δείγμα.
Δηλαδή είναι η $P(Z > |z'| / H_0$ σωστή) όπου Z η τ.μ. που αντιστοιχεί στο στατιστικό και z' η τιμή του στατιστικού για το συγκεκριμένο δείγμα.

Η παραπάνω πιθανότητα αναφέρεται σε μονόπλευρους ελέγχους, ενώ για δίπλευρους διπλασιάζεται.

Συνήθως η H_0 απορρίπτεται εάν η παρατηρούμενη στάθμη σημαντικότητας είναι **μικρότερη** μιας προκαθορισμένης στάθμης σημαντικότητας a που εκλέγεται από αυτόν που κάνει τη στατιστική ανάλυση.

Αν συμβολίσουμε με L_0 τη συνάρτηση πιθανοφάνειας του τυχαίου δείγματος όταν ισχύει η υπόθεση H_0 και L τη συνάρτηση πιθανοφάνειας του τυχαίου δείγματος για όλες τις τιμές της άγνωστης παραμέτρου, έχει αποδειχθεί ότι ένα καλό κριτήριο είναι ο έλεγχος του γενικευμένου λόγου πιθανοφανειών $\lambda = \dfrac{L_0}{\sup L} < c$.

Προφανώς $0 < \lambda < 1$. Αν ο λόγος λ είναι κοντά στη μονάδα, σημαίνει ότι η πιθανοφάνεια κάτω από την υπόθεση H_0 είναι κοντά στη μέγιστη πιθανοφάνεια άρα η υπόθεση H_0 θα πρέπει να γίνει δεκτή. Αν ο λόγος λ είναι πολύ μικρός, σημαίνει ότι η πιθανοφάνεια κάτω από την υπόθεση H_0 είναι πολύ μικρότερη από τη μέγι-

στη πιθανοφάνεια άρα η H_0 θα πρέπει να απορριφθεί. Η τιμή c που ονομάζεται **κρίσιμο σημείο** μας λέει ποιο είναι το όριο κάτω από το οποίο θα απορρίπτουμε την υπόθεση H_0 και εξαρτάται από τη στάθμη σημαντικότητας a του ελέγχου.

Παράδειγμα 5.2

Ζητάμε να ελέγξουμε αν οι συσκευασίες 100 gr καφέ περιέχουν πράγματι 100 gr. Ο έλεγχος θα γίνει με το παρακάτω τυχαίο δείγμα μεγέθους n=10, το $\underset{\sim}{x}' = (98, 99, 100, 100, 101, 97, 102, 97, 98, 95)$. Υποθέτουμε ότι το βάρος ακολουθεί την κατανομή $N(\mu, 5^2)$.

Οι υποθέσεις είναι:

$$H_0 : \mu = 100$$
$$H_1 : \mu \neq 100$$

Η συνάρτηση πιθανοφάνειας για την κανονική κατανομή είναι:

$$L(\mu / \underset{\sim}{x}) = \prod_{l=1}^{n} \left(\frac{1}{\sigma\sqrt{2\pi}} \right) \exp\left\{ -\frac{1}{2\sigma^2}(x_i - \mu)^2 \right\}$$

Για το παράδειγμά μας:

$$L(\mu / \underset{\sim}{x}) = \left(\frac{1}{5\sqrt{2\pi}} \right)^{10} \exp\left\{ -\frac{1}{2 \cdot 25} \sum_{i=1}^{10} (x_i - \mu)^2 \right\}$$

Κάτω από την υπόθεση H_0 έχουμε:

$$L_0 = \left(\frac{1}{5\sqrt{2\pi}} \right)^{10} \exp\left\{ -\frac{1}{2 \cdot 25} \sum_{i=1}^{10} (x_i - 100)^2 \right\}$$

ενώ για το $\sup L$ πρέπει να βρούμε τον εκτιμητή μέγιστης πιθανοφάνειας. Λογαριθμίζοντας και παραγωγίζοντας την $L(\mu / \underset{\sim}{x})$ ως προς μ βρίσκουμε $\hat{\mu} = \overline{x}$ οπότε:

$$\sup L = \left(\frac{1}{5\sqrt{2\pi}} \right)^{10} \exp\left\{ -\frac{1}{2 \cdot 25} \sum_{i=1}^{10} (x_i - \overline{x})^2 \right\}$$

και ο λόγος λ γίνεται:

$$\lambda = \exp\left\{ -\frac{1}{2 \cdot 25} \sum_{i=1}^{10} (x_i - \mu)^2 + \frac{1}{2 \cdot 25} \sum_{i=1}^{10} (x_i - \overline{x})^2 \right\} = \exp\left\{ -\frac{10(\overline{x} - \mu)^2}{2 \cdot 25} \right\}$$

$$\lambda < c \Rightarrow \ln\lambda < \ln c \Rightarrow \ln\lambda = -\frac{10(\overline{x} - \mu)^2}{2 \cdot 25} < \ln c \Rightarrow \left(\frac{(\overline{x} - \mu)\sqrt{10}}{5} \right)^2 > -2\ln c = C_0$$

Τελικά η περιοχή απόρριψης της H_0 δίνεται από τη σχέση:

$$\left(\frac{(\bar{x}-\mu)\sqrt{n}}{5}\right)^2 > C_0 \Leftrightarrow \left|\frac{(\bar{x}-\mu)\sqrt{n}}{5}\right| > \sqrt{C_0}$$

όπου το κρίσιμο σημείο C_0 θα υπολογιστεί από τη στάθμη σημαντικότητας $\mathrm{a} = 0,05$. Από τον ορισμό της στάθμης σημαντικότητας a έχουμε:

$$\mathrm{a} = P \text{ (απόρριψης της } H_0 / H_0 \text{ ισχύει)} =$$

$$= P\left(\left|\frac{(\bar{X}-\mu)\sqrt{n}}{5}\right| > \sqrt{C_0} / \mu = 100\right) = P\left(\left|\frac{(\bar{X}-100)\sqrt{10}}{5}\right| > \sqrt{C_0}\right) =$$

$$= 1 - P\left(\left|\frac{(\bar{X}-100)\sqrt{10}}{5}\right| \le \sqrt{C_0}\right) = 1 - P\left(-\sqrt{C_0} < \frac{(\bar{X}-100)\sqrt{10}}{5} < \sqrt{C_0}\right) =$$

$$= 1 - (\Phi(\sqrt{C_0}) - \Phi(-\sqrt{C_0})) = 2 - 2\Phi(\sqrt{C_0})$$

$$\Rightarrow \Phi(\sqrt{C_0}) = 1 - \mathrm{a}/2 \quad \text{οπότε} \quad \sqrt{C_0} = z_{a/2} = z_{0,025} = 1,96$$

(Χρησιμοποιήθηκε το γεγονός ότι αν $X_i \sim N(\mu, \sigma^2)$ τότε $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$).

Έτσι
$$R = \left\{ \underset{\sim}{x} : \frac{|\bar{x} - \mu_0|\sqrt{n}}{\sigma} > z_{\mathrm{a}/2} \right\}$$

Για το παράδειγμά μας $\bar{x} = 98,7$ οπότε

$$\frac{|98,7 - 100|\sqrt{10}}{5} = 0,82 < 1,96 = z_{\mathrm{a}/2}$$

δηλαδή αν και κατά μέσο όρο οι συσκευασίες του δείγματος είχαν ποσότητα λιγότερη από 100 gr, εντούτοις δεν μπορούμε να απορρίψουμε την υπόθεση H_0 ▲

5.3 Ορισμός του στατιστικού και της απορριπτικής περιοχής ενός ελέγχου

Από τα στοιχεία ενός στατιστικού ελέγχου, όπως έχουν περιγραφεί στην παράγραφο 5.1, τα κυριότερα είναι ο ορισμός του στατιστικού και ο ορισμός της απορριπτικής περιοχής.

Η γενική ιδέα ενός ελέγχου είναι η εξής: ο πληθυσμός βάσει ενός «κανόνα» ή κριτηρίου χωρίζεται σε δύο περιοχές: στην περιοχή αποδοχής της H_0 και στην περιοχή απόρριψης που είναι και η μικρότερη. Κατόπιν γίνεται ο έλεγχος βάσει του κριτηρίου που ορίστηκε και αποφασίζεται που ανήκει το δείγμα. Εάν ανήκει στην

5.12 Εφαρμογές – Λυμένες Ασκήσεις

Άσκηση 5.1

Πάρθηκαν τυχαία 10 φοιτητές και είχαν βάρη σε kg: 53, 69, 62, 78, 81, 55, 66, 62, 74, 60. Υποθέτοντας ότι το βάρος των φοιτητών, ακολουθεί κανονική κατανομή $N(68, 10^2)$:

i) Μπορούμε να ισχυρισθούμε ότι το μέσο βάρος τους, είναι μικρότερο από 68 κιλά και όχι 68; $(\mathrm{a} = 0{,}05)$.

ii) Να βρεθεί η σημαντικότητα του ελέγχου.

Λύση

i) Ο ισχυρισμός ότι το μέσο βάρος τους είναι μικρότερο από 68 κιλά θα είναι η εναλλακτική υπόθεση και η άρνησή του (δηλαδή βάρος ίσο με 68 κιλά) η μηδενική υπόθεση.

Οι υποθέσεις επομένως που θα ελεγχθούν είναι:

$$H_0 : \mu = 68 = \mu_0$$
$$H_1 : \mu < 68$$

Επειδή σ^2 γνωστό, η απορριπτική περιοχή της H_0 είναι η:

$$R = \{z < -z_{\mathrm{a}}\} \quad \text{όπου} \quad z = \frac{(\overline{x} - \mu_0)\sqrt{n}}{10}.$$

Δίνονται $\overline{x} = 66,\ n = 10,\ \sigma = 10$, οπότε

$$z = \frac{(66-68)\sqrt{10}}{10} = -0{,}6325 \quad \text{και} \quad z_{\mathrm{a}} = z_{0{,}05} = 1{,}64,$$

δε βρισκόμαστε στην απορριπτική περιοχή της H_0 (έχουμε $z > -z_{0{,}05}$) συνεπώς δεχόμαστε την H_0 και συμπεραίνουμε ότι το μέσο βάρος των φοιτητών δεν είναι μικρότερο των 68 κιλών, αλλά 68 κιλά.

ii) Η σημαντικότητα του ελέγχου ο οποίος στην προκειμένη περίπτωση είναι μονόπλευρος, είναι η πιθανότητα

$$P(z > |-0{,}63| \,/\, H_0 \text{ σωστή}) = P(z > 0{,}63 \,/\, \mu = 68) =$$
$$= 1 - \Phi(0{,}63) = 1 - 0{,}7357 = 0{,}2643$$

σύμφωνα με τον πίνακα της κανονικής κατανομής.

Άσκηση 5.2

Οι χρόνοι CPU 725 προγραμμάτων έδωσαν $\bar{x} = 171$ sec, $s = 5$ sec.

i) Να βρεθεί 95% διάστημα εμπιστοσύνης για τον άγνωστο μέσο χρόνο όλων των προγραμμάτων.

ii) Αν γνωρίζουμε ότι ο πραγματικός χρόνος των προγραμμάτων, πριν 15 χρόνια ήταν 170 sec, μπορούμε να δεχθούμε τον ισχυρισμό ότι τώρα ο μέσος χρόνος CPU αυξήθηκε; $(a = 0,01)$.

Λύση

i) Επειδή η διασπορά του πληθυσμού δηλαδή το σ^2 είναι άγνωστη και το δείγμα μεγάλο, το 95% διάστημα εμπιστοσύνης για τη μέση τιμή του πληθυσμού, δίνεται από τον τύπο:

$$\left(\bar{x} - \frac{s}{\sqrt{n}} z_{a/2},\ \bar{x} + \frac{s}{\sqrt{n}} z_{a/2} \right)$$

Αντικαθιστώντας τις γνωστές τιμές, παίρνουμε:

$$\left(171 \pm \frac{5}{\sqrt{725}} 1,96 \right) = (170,64\ ,\ 171,36)$$

δηλαδή ο πραγματικός μέσος χρόνος CPU είναι μεταξύ 170,64 και 171,36 δευτερολέπτων.

ii) Ο ισχυρισμός: «τώρα ο μέσος χρόνος CPU είναι μεγαλύτερος», θα είναι η εναλλακτική υπόθεση. Έτσι οι υποθέσεις που θα ελεγχθούν είναι:

$$H_0 : \mu = 170 = \mu_0$$
$$H_1 : \mu > 170$$

Είμαστε στην περίπτωση όπου σ^2 άγνωστο, και δείγμα μεγάλο.
Η απορριπτική περιοχή της H_0 είναι:

$$P = \{ t > z_a \},$$

όπου:

$$t = \frac{(\bar{x} - \mu_0)\sqrt{n}}{s} = \frac{(171 - 170)\sqrt{725}}{5} = 5,36 \quad \text{και} \quad z_a = z_{0,01} = 2,33$$

Τέλος, επειδή η τιμή του στατιστικού είναι μεγαλύτερη της κρίσιμης τιμής, η H_0 απορρίπτεται, γίνεται αποδεκτή η H_1 και το συμπέρασμά μας είναι ότι τώρα ο μέσος χρόνος CPU αυξήθηκε και είναι μεγαλύτερος των 170 sec.

5.13 Έλεγχοι υποθέσεων με χρήση της R

5.13.1 Βασικές εντολές ελέγχων υποθέσεων στην R

Υπάρχουν διάφορες συναρτήσεις για την πραγματοποίηση ελέγχων υποθέσεων και την εύρεση διαστημάτων εμπιστοσύνης, στις διάφορες βιβλιοθήκες της R. Οι βασικές εντολές ελέγχου υποθέσεων, επιστρέφουν στον χρήστη και το διάστημα εμπιστοσύνης όπου η μηδενική υπόθεση είναι αληθής. Τέλος, η εντολή για τον έλεγχο παραμένει η ίδια, ανεξαρτήτως του μεγέθους του δείγματος.

Πίνακας-R 5.1

Βασικές εντολές ελέγχων υποθέσεων και διαστημάτων εμπιστοσύνης για ένα δείγμα

Εντολή στην R	H_0 **Παράμετρος**	H_1 **Παράμετρος**	**Βιβλιοθήκη**
t.test	$\mu = \mu_o$ για τη μηδενική υπόθεση: $mu = \mu_0$ για το διάστημα εμπιστοσύνης: *conf.level=1–a*	$\mu \neq \mu_o$ *alternative="two.sided"*	stats
		$\mu > \mu_o$ *alternative="greater"*	
		$\mu < \mu_o$ *alternative="less"*	
var.test	$\sigma^2 = \sigma_0^2$ για τη μηδενική υπόθεση: $sigma.squared = \sigma_0^2$ για το διάστημα εμπιστοσύνης: *conf.level=1–a*	$\sigma^2 \neq \sigma_0^2$ *alternative="two.sided"*	
		$\sigma^2 > \sigma_0^2$ *alternative="greater"*	
		$\sigma^2 < \sigma_0^2$ *alternative="less"*	
prop.test (n>30) *binom.test (n≤30)*	$p = p_0$ για τη μηδενική υπόθεση: *x=αριθμός επιτυχιών,* *n=αριθμός προσπαθειών,* $p = p_0$ για το διάστημα εμπιστοσύνης: *conf.level=1–a*	$p \neq p_0$ *alternative="two.sided"*	
		$p > p_0$ *alternative="greater"*	
		$p < p_0$ *alternative="less"*	

Σε όλους τους ελέγχους, είναι προεπιλεγμένη η εναλλακτική υπόθεση (alternative=”two.sided”), και το διάστημα εμπιστοσύνης 95% (conf.level=0.95) και μπορούν να παραληφθούν.

Πίνακας-R 5.2

Βασικές εντολές ελέγχων υποθέσεων και διαστημάτων εμπιστοσύνης για δύο δείγματα από κανονική κατανομή.

Εντολή στην R	**H_0 Παράμετρος εντολής**	**H_1 Παράμετρος εντολής**	**Διασπορές Παράμετρος εντολής**	**Βιβλιοθήκη**
t.test	$\mu_1 = \mu_2$ x = διάνυσμα τιμών του 1 δείγματος y = διάνυσμα τιμών του 2 δείγματος Για το διάστημα εμπιστοσύνης conf.level=1–a	$\mu_1 \neq \mu_2$ alternative= “two.sided”	Δείγματα ανεξάρτητα ίσες διασπορές var.equal=TRUE	stats
		$\mu_1 > \mu_2$ alternative= “greater”	Δείγματα ανεξάρτητα, άνισες διασπορές var.equal=FALSE	
		$\mu_1 < \mu_2$ alternative= “less”	Δείγματα εξαρτημένα, ζευγαρωτές paired=TRUE”	
var.test	$\frac{\sigma_1^2}{\sigma_2^2} = 1$ x = διάνυσμα τιμών του 1 δείγματος y = διάνυσμα τιμών του 2 δείγματος Για το διάστημα εμπιστοσύνης conf.level=1–a	$\sigma_1^2 \neq \sigma_2^2$ alternative= “two.sided”	Δείγματα ανεξάρτητα ίσες διασπορές var.equal=TRUE	stats
		$\sigma_1^2 > \sigma_2^2$ alternative= “greater”	Δείγματα ανεξάρτητα, άνισες διασπορές var.equal=FALSE	
		$\sigma_1^2 < \sigma_2^2$ alternative= “less”	Δείγματα εξαρτημένα, ζευγαρωτές paired=TRUE”	

Πίνακας-R 5.3

Βασικές εντολές ελέγχων υποθέσεων και διαστημάτων εμπιστοσύνης για σύγκριση αναλογιών.

Εντολή στην R	H_0 **Παράμετρος**	H_1 **Παράμετρος**	**Βιβλιοθήκη**
prop.test	$p_1 = p_1$ για τη μηδενική υπόθεση: $x = c(x_1, x_2)$, όπου x_1 αριθμός επιτυχιών του πρώτου δείγματος, x_2 του δευτέρου $n = c(n_1, n_2)$, όπου n_1 αριθμός προσπαθειών του πρώτου δείγματος, n_2 αριθμός προσπαθειών του δεύτερου δείγματος, για το διάστημα εμπιστοσύνης conf.level=1-a	$p_1 \neq p_1$ alternative = "two.sided"	stats
		$p_1 > p_1$ alternative = "greater"	
		$p_1 < p_1$ alternaive = "less"	

Τέλος ο χρήστης μπορεί να δει και οπτικά τα αποτελέσματα του ελέγχου, χρησιμοποιώντας τις εντολές του πίνακα-R 5.4 .

Πίνακας-R 5.4

Εντολές οπτικοποίησης ελέγχων υποθέσεων.

Εντολή στην R	**Παράμετροι**	**Βιβλιοθήκη**
ggttest	t=ένα αντικείμενο htest colaccept=χρώμα περιοχής αποδοχής του ελέγχου colreject=χρώμα περιοχής απόρριψης του ελέγχου colstat= χρώμα του στατιστικού	gginference
ggvartest		

5.13.2 Εφαρμογές – Λυμένες ασκήσεις

Άσκηση-R 5.1

Η Ελληνική Στατιστική Υπηρεσία παρέχει τα δεδομένα των μετρήσεων κατανάλωσης σε κυβικούς τόνους διαφόρων ειδών καυσίμου ανά περιφέρεια και νομό. Στη βιβλιοθήκη gginference της R, υπάρχει το σύνολο δεδομένων DieselbioRon95, που αποτελείται από ένα δείγμα 24 νομών της Ελλάδας και τη συνολική κατανάλωση του νομού σε αμόλυβδη βενζίνη και πετρέλαιο κίνησης σε μετρικούς τόνους, για τις χρονιές 2006 και 2016. Υποθέτοντας ότι η συνολική κατανάλωση, ανεξαρτήτως προϊόντος και χρονιάς, προέρχεται από κανονική κατανομή, μπορούμε να ισχυριστούμε ότι η μέση κατανάλωση σε πετρέλαιο κίνησης για τη χρονιά 2016 είναι:

a) μεγαλύτερη από 40000 κυβικούς τόνους και όχι 40000 ($a = 0.05$).

b) 52000 κυβικοί τόνοι ($a = 0.05$ και 0.01).

Λύση

a) Για τη λύση του ερωτήματος, θα πραγματοποιηθεί έλεγχος μέσης τιμής ενός δείγματος (one sample *t*-test). Αυτό μπορεί να γίνει με την εντολή *t.test* στης stats βιβλιοθήκης της R. Στην εντολή αυτή, πρέπει να εισάγουμε τις τιμές τους δείγματος σε μορφή διανύσματος, τη μέση τιμή για την οποία ελέγχουμε για τη μηδενική υπόθεση, την εναλλακτική υπόθεση του ελέγχου καθώς και το επίπεδο εμπιστοσύνης.

Αρχικά, κάνουμε εγκατάσταση την βιβλιοθήκη gginference στην R, που περιλαμβάνει το σύνολο δεδομένων της άσκησης.

```
#Εγκατάσταση της βιβλιοθήκης (εφόσον δεν έχει γίνει ξανά)
#install.packages('gginference')
# Φόρτωση της βιβλιοθήκης
library(gginference)
```

Ορίζουμε τα δεδομένα σε μορφή διανύσματος.

```
diesel<-DieselbioRon95$DieselBio_consumption2016
```

Ο ισχυρισμός ότι η μέση κατανάλωση είναι 40000 κυβικοί τόνοι, είναι η μηδενική υπόθεση, ενώ η υπόθεση η μέση κατανάλωση είναι μεγαλύτερη των 40000 κυβικών τόνων αποτελεί την εναλλακτική υπόθεση του ελέγχου. Επομένως, οι υποθέσεις του ελέγχου είναι:

$$H_0 : \mu = 4000$$
$$H_1 : \mu > 4000$$

```
diesel_t.test = t.test(
  x=diesel,
  mu=40000,
  alternative="greater",
  conf.level=0.95)
```

Τα αποτελέσματα του ελέγχου είναι:

```
diesel_t.test
##
##  One Sample t-test
##
## data:  diesel
## t = 0.41052, df = 23, p-value = 0.3426
## alternative hypothesis: true mean is greater than 40000
## 95 percent confidence interval:
##  34568.52      Inf
## sample estimates:
## mean of x
##  41710.75
```

Η δομή της λίστας της μεταβλητής των αποτελεσμάτων του t-test είναι:

```
str(diesel_t.test)
## List of 9
##  $ statistic  : Named num 0.411
##   ..- attr(*, "names")= chr "t"
##  $ parameter  : Named num 23
##   ..- attr(*, "names")= chr "df"
##  $ p.value    : num 0.343
##  $ conf.int   : num [1:2] 34569 Inf
##   ..- attr(*, "conf.level")= num 0.95
##  $ estimate   : Named num 41711
##   ..- attr(*, "names")= chr "mean of x"
##  $ null.value : Named num 40000
##   ..- attr(*, "names")= chr "mean"
##  $ alternative: chr "greater"
##  $ method     : chr "One Sample t-test"
##  $ data.name  : chr "diesel"
##  - attr(*, "class")= chr "htest"
```

Μπορούμε να ανακτήσουμε τα αποτελέσματα του ελέγχου από τις παραμέτρους τις λίστας *diesel_t.test* .

➲ Η τιμή του t στατιστικού:

```
diesel_t.test$statistic
##         t
## 0.4105172
```

- Οι βαθμοί ελευθερίας για το t στατιστικό:

```
diesel_t.test$parameter
## df
## 23
```

- Η τιμή σημαντικότητας:

```
diesel_t.test$p.value
## [1] 0.3426129
```

- Το 95% διάστημα εμπιστοσύνης:

```
diesel_t.test$conf.int
## [1] 34568.52      Inf
## attr(,"conf.level")
## [1] 0.95
```

- Η δειγματική μέση τιμή:

```
diesel_t.test$estimate
## mean of x
##  41710.75
```

- Η τιμή μ_0 που ελέγχουμε:

```
diesel_t.test$null.value
##  mean
## 40000
```

- Η εναλλακτική υπόθεση:

```
diesel_t.test$alternative
## [1] "greater"
```

- Το όνομα του ελέγχου:

```
diesel_t.test$method
## [1] "One Sample t-test"
```

Το p-value που προκύπτει, είναι η τιμή σημαντικότητας του ελέγχου κι ελέγχουμε αν είναι μεγαλύτερο ή μικρότερο της στάθμης σημαντικότητας a. Παρατηρούμε ότι το p-value = 0.3426 > 0.05, επομένως, ο έλεγχος δεν είναι στατιστικά σημαντικός και η μηδενική υπόθεση δε μπορεί να απορριφθεί. Το στατιστικό σφάλμα που θα

κάνουμε αν απορρίψουμε την H_0 είναι μεγαλύτερο από το 0.05, όποτε δε μπορούμε να την απορρίψουμε. Συνεπώς, η μέση κατανάλωση στον πληθυσμό της Ελλάδας είναι ίση με 40000 κυβικούς τόνους πετρελαίου κίνησης. Επιπλέον, κοιτώντας το διάστημα εμπιστοσύνης, παρατηρούμε ότι σε αυτό περιέχεται η τιμή της μηδενικής υπόθεσης, επομένως ο έλεγχος γίνεται δεκτός.

b) Ο ισχυρισμός ότι η μέση κατανάλωση είναι 52000 κυβικοί τόνοι αποτελεί την μηδενική υπόθεση, ενώ η εναλλακτική υπόθεση του ελέγχου είναι ότι η μέση κατανάλωση είναι διαφορετική από 52000 κυβικούς τόνους. Οι υποθέσεις του ελέγχου διατυπώνονται ως εξής:

$$H_0 : \mu = 52000$$
$$H_1 : \mu \neq 52000$$

Ελέγχοντας την τιμή του *p*-value που προκύπτει από τον έλεγχο *t*-test, είμαστε σε θέση να γνωρίζουμε για ποιες τιμές του a η μηδενική υπόθεση απορρίπτεται η όχι.

Αν, όμως, θέλουμε να δεχτούμε ή να απορρίψουμε τον έλεγχο εξετάζοντας τα διαστήματα εμπιστοσύνης για τη μέση τιμή, τότε πρέπει να γίνει έλεγχος με την εντολή t.test δύο φορές, ώστε να έχουμε τα διαφορετικά διαστήματα εμπιστοσύνης με σιγουριά 95% και 99%.

Ο δίπλευρος έλεγχος *t*-test με $H_0 : \mu = 52000$, και επίπεδο εμπιστοσύνης 95% είναι:

```
t1= t.test(x=diesel,
           mu=52000,
           alternative="two.sided",
           conf.level=0.95)
```

Τα αποτελέσματα του *t*-test ελέγχου είναι:

```
t1
##
##  One Sample t-test
##
## data:  diesel
## t = -2.469, df = 23, p-value = 0.0214
## alternative hypothesis: true mean is not equal to 52000
## 95 percent confidence interval:
##  33090.02 50331.49
## sample estimates:
## mean of x
##  41710.75
```

Ο δίπλευρος έλεγχος *t*-test με $H_0 : \mu = 52000$, και επίπεδο εμπιστοσύνης 99% είναι:

```
t2= t.test(x=diesel,
           mu=52000,
           alternative="two.sided",
           conf.level=0.99)
```

Τα αποτελέσματα του ελέγχου είναι:

```
t2
##
##  One Sample t-test
##
## data:  diesel
## t = -2.469, df = 23, p-value = 0.0214
## alternative hypothesis: true mean is not equal to 52000
## 99 percent confidence interval:
##  30011.72 53409.79
## sample estimates:
## mean of x
##  41710.75
```

Η σημαντικότητα του ελέγχου είναι 0.0214, άρα η H_0 γίνεται αποδεκτή σε στάθμη σημαντικότητας a=0.01, ενώ απορρίπτεται για a=0.05, αφού p-value<0.05.

Αυτό μπορούμε να το διαπιστώσουμε και από τα διαστήματα εμπιστοσύνης όπου για 95% είναι $(33090.02,50331.49)$ και δεν περιλαμβάνεται σε αυτό η τιμή 52000 ενώ στο 99% $(30011.72,53409.79)$ η τιμή του ελέγχου 52000 περιλαμβάνεται.

Το 95% διαστήματα εμπιστοσύνης είναι:

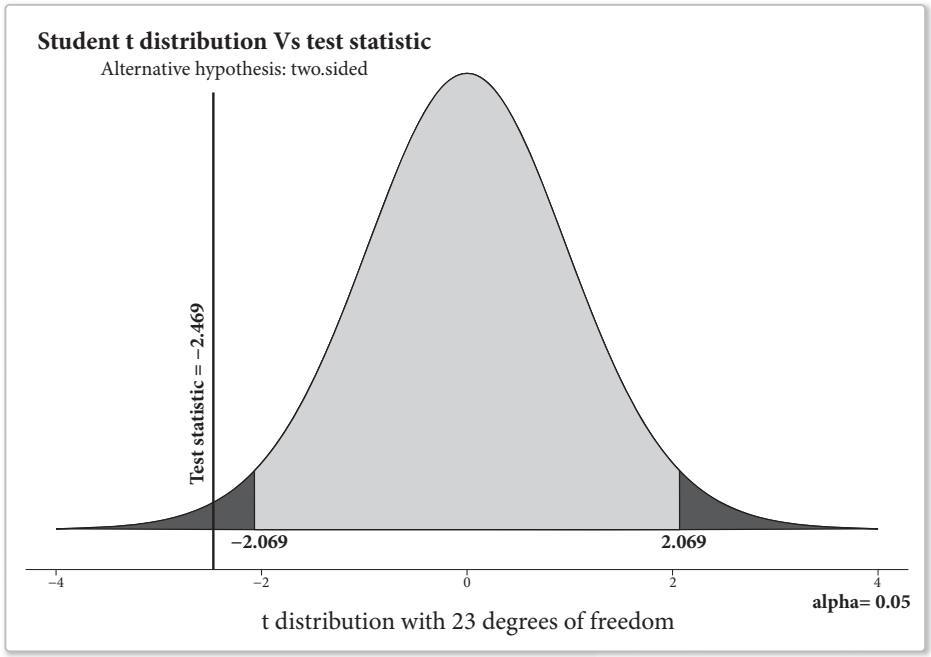
```
t1$conf.int #95%
## [1] 33090.02 50331.49
## attr(,"conf.level")
## [1] 0.95
```

Το 99% διαστήματα εμπιστοσύνης είναι:

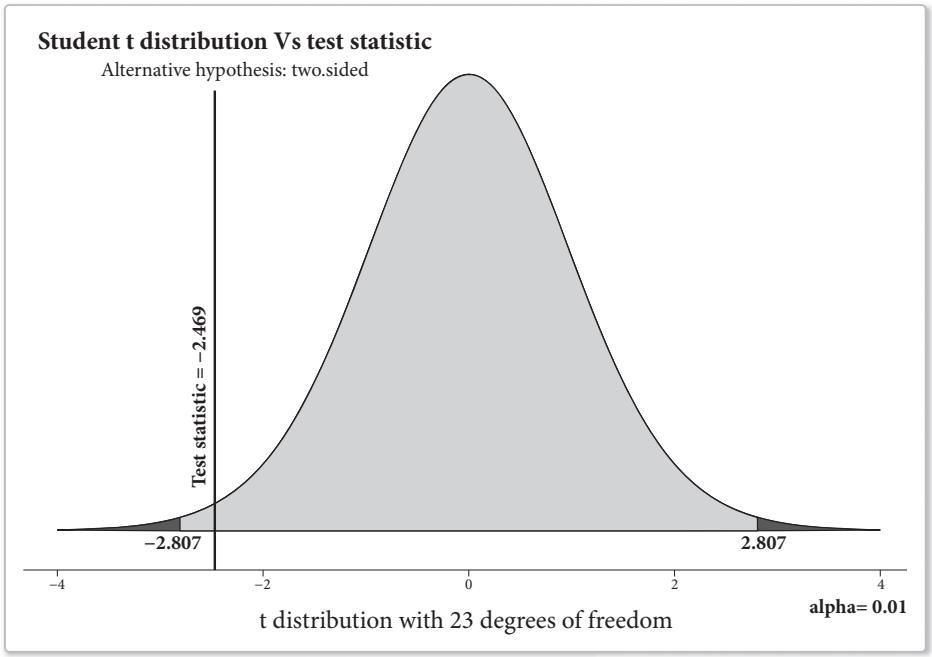
```
t2$conf.int #99%
## [1] 30011.72 53409.79
## attr(,"conf.level")
## [1] 0.99
```

Τέλος, για να δούμε και γραφικά τα αποτελέσματα μπορεί να χρησιμοποιηθεί η εντολή *ggttest* της βιβλιοθήκης *gginference* η οποία δίνει γραφικά την κατανομή που ακολουθείται από το στατιστικό, το στατιστικό που προκύπτει από τον έλεγχο και με διαφορετική απόχρωση την περιοχή αποδοχής και απόρριψης του ελέγχου.

```
library(gginference)
ggttest(t1,
        colaccept = "grey89", #χρώμα αποδεκτής περιοχής
        colreject = "black") #χρώμα απορριπτικής περιοχής
```




```
ggttest(t2,
        colaccept = "grey89", #χρώμα αποδεκτής περιοχής
        colreject = "black") #χρώμα απορριπτικής περιοχής
```



Προτεινόμενες Ασκήσεις

5.23 Ένα φωτοτυπικό μηχάνημα είναι γνωστό ότι τυπώνει 45 αντίγραφα το λεπτό. Σε μια προσπάθεια βελτίωσης της απόδοσής του έγιναν κάποιες αλλαγές και ο έλεγχος που ακολούθησε έδωσε σε τρία διαφορετικά λεπτά 46, 47 και 48 αντίγραφα. Η βελτίωση είναι στατιστικά σημαντική ή απλώς είναι τυχαία;

5.24 Από ένα πληθυσμό διαλέγονται τυχαία 12 στοιχεία. Για το δείγμα $\sum x_i = 1038$ και $\sum x_i^2 = 107888$. Ελέγξτε την υπόθεση $\mu = 100$, για a =0,05 χρησιμοποιώντας ένα δίπλευρο έλεγχο.

5.25 Ένας δημοσιογράφος σε πρόσφατη ανακοίνωση, ισχυρίζεται ότι αυτοί που δεν σπουδάζουν παντρεύονται σε μικρότερη ηλικία απ' αυτούς που σπουδάζουν. Πάρθηκαν δύο δείγματα μεγέθους 100 από κάθε περίπτωση. Η μέση ηλικία γάμου και η τυπική απόκλιση από αυτούς που δε σπούδασαν ήταν 22,5 και 1,4 χρόνια αντίστοιχα, ενώ γι' αυτούς που σπούδασαν η μέση ηλικία και η τυπική απόκλιση ήταν 23 και 1,8 χρόνια. Ελέγξτε τον ισχυρισμό του δημοσιογράφου.

5.26 Ως «χρόνος απόκρισης» ενός υπολογιστή ορίζεται ο χρόνος που πρέπει να περιμένει ο χρήστης μέχρι να πάρει μία πληροφορία από το δίσκο. Υποθέτουμε ότι ένα κέντρο δεδομένων θέλει να συγκρίνει τους μέσους χρόνους απόκρισης δύο μονάδων δίσκων του υπολογιστή. Εάν μ_1 είναι ο μέσος χρόνος απόκρισης για το δίσκο 1 και μ_2 για το δίσκο 2, το κέντρο θέλει να ελέγξει αν $\mu_1 = \mu_2$.

Για την παραπάνω σύγκριση συγκεντρώθηκαν τα εξής δεδομένα:

Χρόνος απόκρισης των δύο δίσκων

Δίσκος 1 ($n_1 = 16$)			
59	73	74	61
92	60	84	58
54	73	47	70
108	975	33	49

Δίσκος 2 ($n_2 = 17$)				
71	63	40	34	49
38	47	60	71	
40	56	53	68	
39	80	72	50	

Υπάρχει σημαντική διαφορά ανάμεσα στους χρόνους απόκρισης;

5.27 Για να δεχτούμε ότι η περιεκτικότητα σε σίδηρο δύο περιοχών A_1 και A_2 είναι η ίδια, πήραμε 6 δείγματα από κάθε περιοχή και μετρήσαμε την περιεκτικότητα σε gr σιδήρου ανά κυβικό μέτρο.

8 Κεφάλαιο

ΑΝΑΛΥΣΗ ΔΙΑΣΠΟΡΑΣ

8.1 Εισαγωγή

Σε προηγούμενο κεφάλαιο παρουσιάσθηκαν οι έλεγχοι που αφορούν στη σύγκριση των μέσων τιμών δύο πληθυσμών, που ακολουθούν κανονική κατανομή βασισμένοι σε τυχαία (ανεξάρτητα) δείγματα. Στην πράξη όμως, αντιμετωπίζονται πολλές φορές προβλήματα στα οποία πρέπει να συγκριθούν οι μέσες τιμές για περισσότερους από δύο πληθυσμούς. Τα προβλήματα αυτά, καθώς και πολυπλοκότερα, λύνονται με μια μέθοδο που καλείται ανάλυση διασποράς ή ανάλυση διακύμανσης (α.δ.) (analysis of variance).

Η μέθοδος της ανάλυσης διασποράς δε διαφέρει ουσιαστικά από τη μέθοδο της παλινδρόμησης, είναι όμως απλούστερη γιατί χρησιμοποιεί αθροίσματα τετραγώνων αντί της αντιστροφής πινάκων πράγμα πολύ θετικό ιδίως για παλαιότερες εποχές που δεν ήταν δυνατή η χρήση του υπολογιστή.

Και εδώ υπάρχει μια εξαρτημένη ποσοτική μεταβλητή Y τις τιμές της οποίας μπορούμε να παρατηρήσουμε και η οποία μπορεί να εξαρτάται από έναν **παράγοντα *A*** (factor *A*) ή από **δύο ή περισσότερους παράγοντες *A*, *B***, (που να επηρεάζουν ή όχι ο ένας τον άλλον κ.λπ.). Οι παράγοντες αυτοί, ισοδυναμούν με τις ανεξάρτητες μεταβλητές της παλινδρόμησης. Οι τιμές που παίρνει ο παράγοντας λέγονται στάθμες (levels) και είναι προφανώς πεπερασμένου πλήθους. Χωρίς περιορισμό της γενικότητας μπορούμε να θεωρήσουμε ότι είναι οι πρώτοι ν φυσικοί αριθμοί. Στα δύο παρακάτω παραδείγματα, επισημαίνονται όλα τα προηγούμενα και εξηγείται ο όρος «ανάλυση διασποράς».

Παράδειγμα 8.1

Θέλουμε να συγκρίνουμε τη ζωή των μπαταριών τριών διαφορετικών τύπων, παίρνοντας δείγματα μεγέθους 5 από κάθε τύπο. (Τα δείγματα μεγέθους 5 είναι πολύ μικρά για ασφαλή συμπεράσματα, μπορούμε όμως να δείξουμε τη βασική ιδέα στην οποία στηρίζεται η μέθοδος).

Με την υπόθεση ότι η διάρκεια ζωής ακολουθεί κανονική κατανομή μπορούν αυτά τα δείγματα να μας δώσουν πληροφορίες για το αν υπάρχουν διαφορές στη διάρκεια ζωής των μπαταριών των τριών τύπων;

Μια πρώτη ματιά στον πίνακα 8.1 όπου δίδονται τα δεδομένα και κάποια στατιστικά, παρατηρείται μικρή μεταβλητότητα μέσα στα δείγματα, ενώ η μεταβλητότητα μεταξύ των δειγμάτων είναι μεγάλη.

Πίνακας 8.1

Διάρκεια ζωής σε ώρες

Τύπος Α	*Τύπος Β*	*Τύπος Γ*
58,0	50,2	40,2
58,4	50,0	40,0
58,2	50,0	39,8
57,8	49,8	39,6
57,6	50,0	40,4
$\overline{x}_1 = 58,0$	$\overline{x}_2 = 50,0$	$\overline{x}_3 = 40,0$
$s_1 = 0,32$	$s_2 = 0,14$	$s_3 = 0,32$

▲

Παράδειγμα 8.2

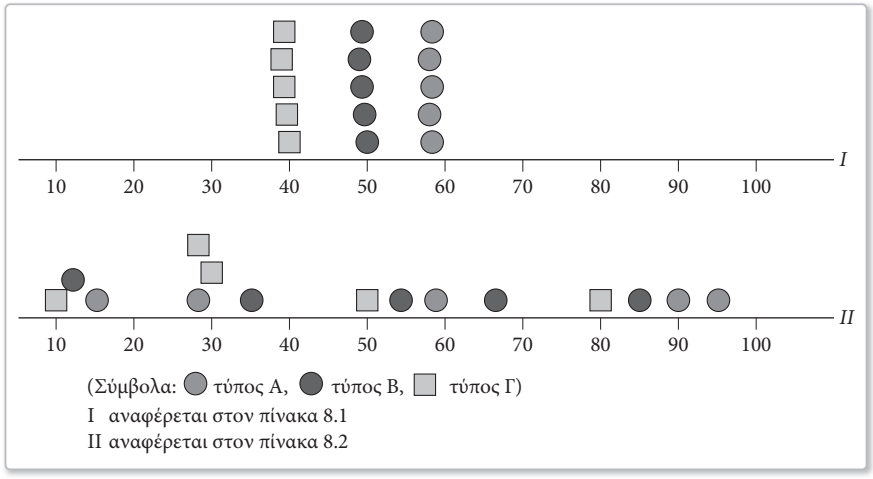
Στα παρακάτω δείγματα, που αναφέρονται στο προηγούμενο παράδειγμα, η μεταβλητότητα μεταξύ των δειγμάτων είναι όπως στο παράδειγμα 8.1, η μεταβλητότητα όμως μέσα στα δείγματα είναι μεγαλύτερη (δειγματικές διασπορές μεγαλύτερες).

Πίνακας 8.2

Διάρκεια ζωής σε ώρες

Τύπος Α	*Τύπος Β*	*Τύπος Γ*
58,0	67,7	30,4
28,6	11,8	78,6
90,0	32,6	29,6
97,8	83,0	51,0
15,6	55,4	10,4
$\overline{y}_1 = 58,0$	$\overline{y}_2 = 50,0$	$\overline{y}_3 = 40,0$
$s_1 = 36,30$	$s_2 = 28,18$	$s_3 = 25,92$

Αν κάνουμε μια γραφική παράσταση των δύο παραπάνω παραδειγμάτων τότε έχουμε:



Σχήμα 8.1

Παρατηρούμε ότι και στους δύο πίνακες 8.1 και 8.2 οι μπαταρίες τύπου A έχουν την ίδια μέση διάρκεια ζωής αλλά διαφορετική διασπορά· το ίδιο και για τις μπαταρίες τύπων B και Γ.

Από το σχήμα 8.1 γίνεται φανερό τι εννοούμε με τον όρο ανάλυση διασποράς: όλες οι διαφορές στους δειγματικούς μέσους κρίνονται στατιστικά σημαντικές ή όχι, σε σχέση με τις διασπορές τους μέσα στα δείγματα. ▲

8.2 Η λογική του κριτηρίου της ανάλυσης διασποράς

Όπως έγινε φανερό στην προηγούμενη παράγραφο, το πρόβλημα που τίθεται είναι αν οι μέσες τιμές των πληθυσμών από τους οποίους προέρχονται τα δείγματα και οι οποίοι στην ανάλυση διασποράς θεωρούνται **κανονικοί και με κοινή διασπορά,** διαφέρουν σημαντικά ή όχι. Η απάντηση που δίνει η μέθοδος της ανάλυσης διασποράς στηρίζεται στη σύγκριση της μεταβλητότητας μεταξύ των δειγματικών μέσων και της μεταβλητότητας των τιμών των y μέσα σε κάθε δείγμα.

Η υπόθεση λοιπόν που ελέγχεται γενικά είναι:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$$

που σημαίνει ότι τα k δείγματα προέρχονται από τον ίδιο πληθυσμό.

Στην παράγραφο 5.5.2 έχει περιγραφεί ο έλεγχος για τη σύγκριση των μέσων τιμών δύο ανεξάρτητων δειγμάτων που προέρχονται από **κανονικές κατανομές** με **κοινή διασπορά** σ^2.

Η χρήση αυτού του κριτηρίου για τη σύγκριση k μέσων τιμών θα σήμαινε $\binom{k}{2}$ ελέγχους υποθέσεων, εργασία κοπιαστική και χρονοβόρα όσο το k μεγαλώνει. Το πιο σημαντικό όμως μειονέκτημα των συγκρίσεων ανά δύο, είναι ότι η πιθανότητα να απορριφθεί ενώ είναι σωστή τουλάχιστον μια μηδενική υπόθεση, αυξάνει με την αύξηση του αριθμού των ελέγχων. Ενδεικτικά αναφέρεται ότι για 5 δείγματα και για σ.σ. a=0,05, το συνολικό σφάλμα ανέρχεται περίπου σε 40%. Χρειάζεται λοιπόν να αντιμετωπισθεί το πρόβλημα της σύγκρισης των μέσων τιμών συνολικά. Γενικεύοντας για k τον τύπο που δίνει την κοινή διασπορά του πληθυσμού, ο εκτιμητής του σ^2, δηλαδή η μεταβλητότητα εντός των δειγμάτων, δίνεται από τη σχέση:

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \ldots + (n_k-1)s_k^2}{n_1 + \ldots + n_k - k} =$$

$$= \frac{\sum_{i=1}^{n_1}(y_{1i}-\overline{y}_1)^2 + \sum_{i=1}^{n_2}(y_{2i}-\overline{y}_2)^2 + \ldots + \sum_{i=1}^{n_k}(y_{ki}-\overline{y}_k)^2}{n_1 + \ldots + n_k - k} =$$

$$= \frac{SSE}{n_1 + n_2 \ \ldots + n_k - k} \tag{8.1}$$

όπου y_{ij} η j-παρατήρηση του i δείγματος $i = 1, 2, \ldots, k$, $\overline{y}_i$ η δειγματική μέση τιμή του i δείγματος και ***SSE*** το **άθροισμα τετραγώνων των σφαλμάτων** το οποίο όπως και στην παλινδρόμηση μετρά τη συνολική μεταβλητότητα **εντός** των δειγμάτων.

Ένα μέτρο της μεταξύ των δειγμάτων μεταβλητότητας, δίνεται από τη σχέση:

$$\frac{1}{k-1}\sum_{i=1}^{k} n_i(\overline{y}_i - \overline{y})^2 = \frac{SSA}{k-1} \tag{8.2}$$

όπου $\overline{y}$ ο γενικός μέσος και με *SSA* συμβολίζεται το άθροισμα τετραγώνων των αποκλίσεων των μέσων τιμών των δειγμάτων του παράγοντα A από το γενικό μέσο. Είναι η εξηγήσιμη μεταβλητότητα, η μεταβλητότητα δηλαδή που οφείλεται στις διαφορές που έχουν μεταξύ τους τα δείγματα.

8.7 Εφαρμογές – Λυμένες Ασκήσεις

Άσκηση 8.1

Η απόδοση σε γάλα (kg/24 h) μιας προβατίνας που έχει γεννήσει, μετρήθηκε ζυγίζοντας το προβατάκι της πριν και μετά το θηλασμό. Χρησιμοποιήθηκαν τρεις διαφορετικές ράτσες *A*, *B*, *Γ* και τα αποτελέσματα ήταν:

Ράτσα							
A	2,4	2,7	1,8	3,2	3,4	2,6	
B	3,2	3,4	4,1	2,8	2,9		
Γ	3,9	4,2	3,6	2,8	3,4	3,7	3,5

Τι συμπεραίνετε; (Δίνεται: δείγματα από κανονικούς πληθυσμούς με κοινή διασπορά).

Λύση

Επειδή τα δεδομένα μας αποτελούνται από τρία διαφορετικά δείγματα ανεξάρτητα μεταξύ τους, που το καθένα αναφέρεται σε μια διαφορετική ράτσα αγελάδων, εκείνο που μπορούμε να ελέγξουμε είναι αν ο παράγοντας «ράτσα» επηρεάζει τις μετρήσεις μας. Οι στάθμες του παράγοντα A είναι τρεις όσα και τα δείγματα έστω οι α_1, α_2 και α_3. Για τη στάθμη α_1 έχουμε δείγμα μεγέθους $n_1 = 6$, για τη στάθμη α_2 έχουμε δείγμα μεγέθους $n_2 = 5$ και για την τρίτη έχουμε μέγεθος δείγματος $n_3 = 7$. Το προτεινόμενο μοντέλο είναι το: $y_{ij} = \mu + \alpha_i + e_{ij}$

Η υπόθεση που ελέγχουμε είναι η $H_0 : \mu_1 = \mu_2 = \mu_3$ ή $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ δηλαδή ο παράγοντας «ράτσα» δεν επηρεάζει την απόδοση σε γάλα, έναντι της H_1: ο παράγοντας «ράτσα» επηρεάζει την απόδοση σε γάλα. Τα δεδομένα μπορούν να τακτοποιηθούν σ' έναν πίνακα, όπου θα αναγραφούν και μερικά από τα μεγέθη που μας χρειάζονται.

Ράτσα			
A_1	A_2	A_3	
2,4	3,2	3,9	
2,7	3,4	4,2	
1,8	4,1	3,6	
3,2	2,8	2,8	
3,4	2,9	3,4	
2,6		3,7	
		3,5	
$\overline{y}_1 = 2,68$	$\overline{y}_2 = 3,28$	$\overline{y}_3 = 8,37$	$\overline{y} = 3,2$

Για το παράδειγμα έχουμε $k=3$, $n_1=6$, $n_2=5$, $n_3=7$

$$n=n_1+n_2+n_3=6+5+7=18$$

Ο πίνακας της ανάλυσης διασποράς συμπληρωμένος, είναι ο παρακάτω:

Πηγή μεταβολής	*Αθροίσματα τετραγώνων*	*β.ε*	*Μέσα τετράγωνα*	*F*
Παράγοντας *A* (μεταξύ των δειγμάτων)	SSA = 2,6751	2	*MSA* =1,3375	$F_A = 5{,}1911$
υπόλοιπο ή σφάλμα (μέσα στα δείγματα)	*SSE* = 3,8649	15	*MSE* = 0,2577	
Ολική	*SST* = *SSA*+*SSE* = 6,5400	17		

Η υπόθεση H_0 απορρίπτεται όταν $F_A > F_{k-1,\, n-k;\, a}$.

Επειδή εδώ $F_A = 5{,}1911 > 3{,}68 = F_{2,15;\, 0{,}05}$ η υπόθεση H_0 δεν γίνεται δεκτή (σε σ.σ. a=0,05) δηλαδή ο παράγοντας «ράτσα» επηρεάζει την απόδοση σε γάλα.

Άσκηση 8.2

Ο παρακάτω πίνακας δίνει το ποσοστό του όγκου που καταλαμβάνουν οι πόροι ενός είδους ελαφρόπετρας που εντοπίστηκε στις περιοχές *A*, *B*, *Γ*:

A :	7,72	8,45	6,90	8,85	9,12	7,63		
B :	6,80	7,30	7,60	7,45	7,20	6,85	6,96	
Γ :	7,85	8,35	8,42	9,17	8,75	8,86	7,85	8,40

Αν υποθέσουμε ότι το ποσοστό του όγκου ακολουθεί κανονική κατανομή με μέση τιμή μ_A, μ_B, μ_Γ για τις περιοχές *A*, *B* και *Γ* αντίστοιχα και διασπορά σ^2, να εξετασθεί αν μπορούμε να ισχυριστούμε ότι:

i) Οι τρεις περιοχές παράγουν το ίδιο είδος ελαφρόπετρας.

ii) Η περιοχή *Γ* παράγει ελαφρόπετρα με πόρους που καταλαμβάνουν ποσοστό όγκου μεγαλύτερο από της περιοχής *B* ; (a=0,05).

Λύση

i) Επειδή πληρούνται οι προϋποθέσεις της ανάλυσης διασποράς, εφαρμόζουμε τη μέθοδο της ανάλυσης διασποράς για έναν παράγοντα.

Εκείνο που διαφοροποιείται από δείγμα σε δείγμα είναι η περιοχή.

Έτσι σαν παράγοντα παίρνουμε την περιοχή με τρεις στάθμες, όσα είναι τα δείγματά μας.

8.8 Ανάλυση Διασποράς με χρήση της R

8.8.1 Βασικές εντολές ανάλυσης διασποράς στην R

Ο έλεγχος της ανάλυσης διασποράς μπορεί να πραγματοποιηθεί με την εντολή aov στην R, αφού πρώτα ελεγχθούν αν ισχύουν οι προϋποθέσεις της ανάλυσης διασποράς. Χρησιμοποιώντας την εντολή aov ορίζουμε το «**μοντέλο**» ανάλυσης διασποράς, και με την εντολή *summary* του «μοντέλου» κατασκευάζουμε τον πίνακα ανάλυσης διασποράς.

Πίνακας-R 8.1
Ορισμός του μοντέλου ανάλυσης διασποράς

Εντολή στην R – Ανάλυση διασποράς	Παράμετροι εισόδου	Τιμές για έναν παράγοντα	Τιμές για δύο παράγοντες	Βιβλιοθήκη
aov	formula	*Y~X*, όπου *Y* η ποσοτική μεταβλητή, ενώ *X* η ποιοτική η οποία διαχωρίζει την ποσοτική μεταβλητή σε διαφορετικούς πληθυσμούς.	*Y~A+B*, όπου *Y* η ποσοτική μεταβλητή, ενώ *A* και *B* ποιοτικές οι οποίες διαχωρίζουν την ποσοτική μεταβλητή σε διαφορετικούς πληθυσμούς.	*Stats*
			Y~A+B+A:B τύπος ανάλυσης διασποράς με αλληλεπίδραση	
	data	Σύνολο δεδομένων		

Πίνακας-R 8.2
Κατασκευή πίνακα Ανάλυσης Διασποράς

Εντολή στην R – Ανάλυση διασποράς	Παράμετροι εισόδου	Περιγραφή	Βιβλιοθήκη
summary	aov	Μοντέλο ανάλυσης διασποράς	base R

Τέλος ο χρήστης μπορεί να δει και γραφικά τα αποτελέσματα του ελέγχου, ζητώντας το διάγραμμα της κατανομής και το στατιστικού.

Πίνακες-R 8.3

Κατασκευή γραφικών των αποτελεσμάτων του *ελέγχου υποθέσεων*.

Εντολή στην R	Παράμετροι	Βιβλιοθήκη
ggaov	*t* =το μοντέλο ανάλυσης διασποράς colaccept = χρώμα περιοχής αποδοχής του ελέγχου colreject = χρώμα περιοχής απόρριψης του ελέγχου	gginference

Για την ερμηνεία των αποτελεσμάτων των εντολών που παραθέτουν οι πίνακες, μπορείτε να συμβουλευτείτε τη βοήθεια της R. Στις ασκήσεις που ακολουθούν γίνεται χρήση όλων των εντολών με στόχο την κατανόηση της χρήσης τους.

8.8.2 Εφαρμογές – Λυμένες ασκήσεις

Άσκηση-R 8.1

Να κατασκευάσετε ένα πλαίσιο δεδομένων που περιέχει τις μετρήσεις του χρόνου πήξης του αίματος για τέσσερις διαφορετικές δίαιτες. Οι χρόνοι πήξης αίματος για τη δίαιτα A είναι: 62, 60, 63, 59, για τη δίαιτα B: 63, 67, 71, 64, 65, 66, για τη δίαιτα: 68, 66, 71, 67, 68, 68 και τέλος για τη δίαιτα D: 56, 62, 60, 61, 63, 64, 63, 59.
Στη συνέχεια να εξετάσετε αν ο παράγοντας δίαιτα έχει οποιαδήποτε επίδραση στο χρόνο πήξης του αίματος, δεδομένου ότι οι μετρήσεις του χρόνου πήξης για κάθε μία από τις τέσσερις δίαιτες προέρχονται από κανονική κατανομή.

Λύση

Αρχικά, ορίζουμε δύο διανύσματα: το πρώτο περιέχει τους χρόνους πήξης του αίματος (ποσοτική μεταβλητή) και το δεύτερο τη δίαιτα (ποιοτική μεταβλητή). Κατασκευάζουμε ένα πλαίσιο δεδομένων με αυτά τα διανύσματα μέσω της εντολής *data.frame*.

Ορίζουμε το διάνυσμα με τιμές τους χρόνους πήξης αίματος

```
coag <- c(62, 60, 63, 59, 63, 67, 71, 64, 65, 66, 68, 66,
          71, 67, 68, 68, 56, 62, 60, 61, 63, 64, 63, 59)
```

Ορίζουμε το διάνυσμα με τιμές το όνομα κάθε δίαιτας, χρησιμοποιώντας την εντολή *rep*. Με την εντολή *rep* επαναλαμβάνουμε τα 4 πρώτα γράμματα του αγγλικού αλφάβητου, τόσες φορές όσες και η συχνότητα εμφάνισης τους στα δεδομένα μας (δηλ. 4 A, 6 B, 6 C και 8 D). Τέλος ορίζουμε με την εντολή *factor* την δίαιτα ως

ποιοτική μεταβλητή.

```
diet <- factor(rep(LETTERS[1:4], times = c(4,6,6,8)))
```

Δημιουργούμε το πλαίσιο δεδομένων και ελέγχουμε την δομή του.

```
coag.df <- data.frame(diet, coag)

str(coag.df)
## 'data.frame':    24 obs. of  2 variables:
##  $ diet: Factor w/4 levels "A","B","C","D": 1 1 2 2 2 ...
##  $ coag: num  62 60 63 59 63 67 71 64 65 66 ...
```

Στο πλαίσιο δεδομένων *coaf.df* που ορίσαμε, έχουμε τέσσερα σύνολα μετρήσεων. Τα σύνολα ορίζονται από τις τέσσερις διαφορετικές δίαιτες και περιλαμβάνουν τις τιμές της πήξης του αίματος. Αν, λοιπόν, πληρούνται οι προϋποθέσεις της ανάλυσης διασποράς, μπορούμε να εφαρμόσουμε ανάλυση διασποράς με έναν παράγοντα, τον παράγοντα «δίαιτα» για να ελέγξουμε αν η μέση τιμή της πήξης του αίματος και στις τέσσερις δίαιτες είναι ίδιες ή διαφορετικές. Εφόσον γνωρίζουμε ότι η τυχαία μεταβλητή «πήξη του αίματος» ακολουθεί κανονική κατανομή, πρέπει να ελέγξουμε, για να είμαστε συνεπείς με τις υποθέσεις της ανάλυσης διασποράς, αν οι διασπορές των πληθυσμών από τους οποίους προέρχονται τα δείγματα είναι ίσες. Επειδή τα δείγματα είναι τέσσερα, αρκεί να συγκρίνουμε τις διασπορές μεταξύ του 1ου και του 2ου δείγματος, μεταξύ του 3ου και του 4ου, και μεταξύ του 1ου και του 3ου.

Κατασκευάζουμε τα διανύσματα των τιμών του χρόνου πήξης του αίματος ανά δίαιτα.

```
x1 <- coag.df[coag.df$diet == "A","coag"]
x2 <- coag.df[coag.df$diet == "B","coag"]
x3 <- coag.df[coag.df$diet == "C","coag"]
x4 <- coag.df[coag.df$diet == "D","coag"]
```

Πραγματοποιούμε έλεγχο υποθέσεων ισότητας διασπορών για τις τιμές του χρόνου πήξης αίματος για τις δίαιτες *A* και *B*.

```
var_ab <- var.test(x1,
                   x2,
                   ratio = 1,
                   alternative = "two.sided",
                   conf.level = 0.95)
var_ab
##
##  F test to compare two variances
##
## data:  x1 and x2
## F = 0.41667, num df = 3, denom df = 5, p-value = 0.5021
```

```
## alternative hypothesis: true ratio of variances is not
equal to 1
## 95 percent confidence interval:
##  0.05366933 6.20200955
## sample estimates:
## ratio of variances
##          0.4166667
```

Συγκρίνουμε τις διασπορές του χρόνου πήξης αίματος για τις δίαιτες C και D.

```
var_ab <- var.test(x3,
                   x4,
                   ratio = 1,
                   alternative = "two.sided",
                   conf.level = 0.95)
var_ab
##
##  F test to compare two variances
##
## data:  x3 and x4
## F = 0.40833, num df = 5, denom df = 7, p-value = 0.3413
## alternative hypothesis: true ratio of variances is not
equal to 1
## 95 percent confidence interval:
##  0.07725923 2.79833922
## sample estimates:
## ratio of variances
##          0.4083333
```

Τέλος, συγκρίνουμε τις διασπορές του χρόνου πήξης αίματος π.χ. για τις δίαιτες A και C.

```
var_ab <- var.test(x1,
                   x3,
                   ratio = 1,
                   alternative = "two.sided",
                   conf.level = 0.95)
var_ab
##
##  F test to compare two variances
##
## data:  x1 and x3
## F = 1.1905, num df = 3, denom df = 5, p-value = 0.8044
## alternative hypothesis: true ratio of variances is not
equal to 1
## 95 percent confidence interval:
##   0.153341 17.720027
## sample estimates:
## ratio of variances
##           1.190476
```

Για κάθε έναν από τους ελέγχους που πραγματοποιήθηκαν, οι διασπορές είναι ίσες, επομένως μπορούμε να προχωρήσουμε σε ANOVA. Με την ανάλυση διασποράς, ελέγχουμε αν υπάρχει διαφορά στις μέσες τιμές των διαφορετικών δειγμάτων. Δηλαδή η μηδενική υπόθεση του ελέγχου είναι:

H_0: οι μέσες τιμές μ_1, μ_2, μ_3, μ_4 του χρόνου πήξης του αίματος για τις τέσσερις δίαιτες είναι ίσες, με εναλλακτική υπόθεση

H_1: υπάρχει τουλάχιστο μία δίαιτα, της οποίας η μέση τιμή διαφέρει από τις υπόλοιπες

Στην R, για την ανάλυση διασποράς, χρησιμοποιείται η εντολή aov της βιβλιοθήκης *stats*. Ως παραμέτρους, η εντολή δέχεται έναν τύπο (formula), με τον οποίο ορίζουμε τους πληθυσμούς από τους οποίους προέρχονται τα δείγματα, και την παράμετρο data, με την όποια ορίζουμε το σύνολο δεδομένων που επεξεργαζόμαστε. Συγκεκριμένα, για την ανάλυση διασποράς με έναν παράγοντα, ο τύπος formula στην εντολή aov ορίζεται από μια ποσοτική μεταβλητή, το σύμβολο ~, και μια ποιοτική μεταβλητή, η οποία διαχωρίζει την ποσοτική μεταβλητή σε διαφορετικούς πληθυσμούς σύμφωνα με τις κατηγορίες της.

Ορίζουμε το μοντέλο ανάλυσης διασποράς για έναν παράγοντα

```
coag.aov <- aov( formula = coag~diet, data = coag.df)
coag.aov
## Call:
##    aov(formula = coag ~ diet, data = coag.df)
##
## Terms:
##                 diet Residuals
## Sum of Squares   228       112
## Deg. of Freedom    3        20
##
## Residual standard error: 2.366432
## Estimated effects may be unbalanced
```

Χρησιμοποιώντας την εντολή summary του ελέγχου (coag.aov) παίρνουμε τον πίνακα ανάλυσης διασποράς.

```
summary(coag.aov)
##             Df Sum Sq Mean Sq F value   Pr(>F)
## diet         3    228    76.0   13.57 4.66e-05 ***
## Residuals   20    112     5.6
## ---
## Signif. codes:0'***'0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Η εντολή *summary*, με παράμετρο την λίστα των αποτελεσμάτων της aov επιστρέφει τον πίνακα ANOVA για τη σύγκριση των τεσσάρων ειδών δίαιτας. Η πρώτη

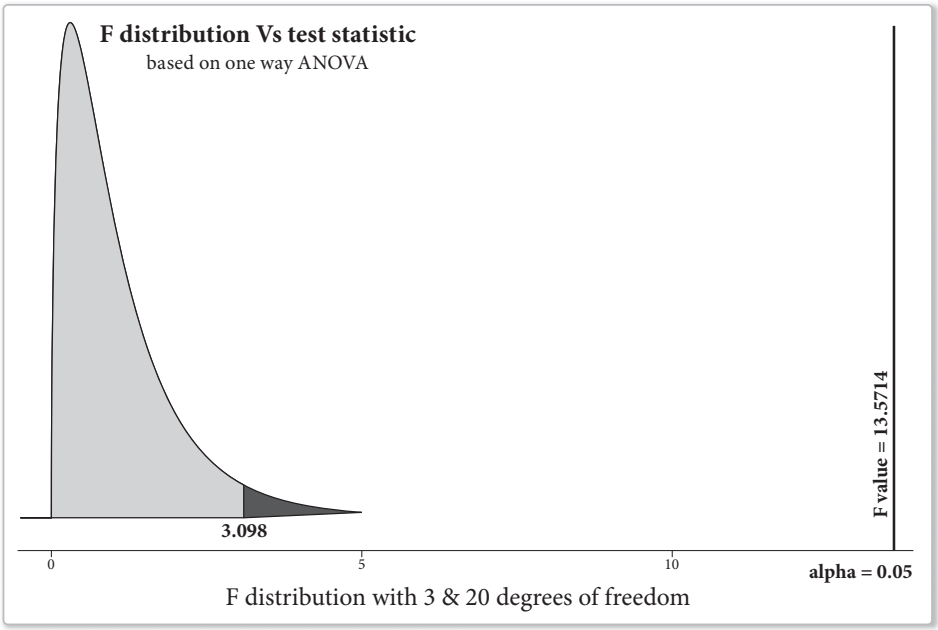
στήλη μας δίνει τους βαθμούς ελευθερίας, η δεύτερη το άθροισμα των τετραγώνων και η τρίτη το μέσο άθροισμα τετραγώνων της ποιοτικής μεταβλητής diet και των υπολοίπων. Η τέταρτη στήλη περιέχει την τιμή *F*-στατιστικό του ελέγχου. Στην πέμπτη στήλη λαμβάνουμε την τιμή της πιθανότητας λάθους αποδοχής της εναλλακτικής υπόθεσης *p*-value. Παρατηρούμε ότι το *p*-value είναι *p*-value=0.00000466 μικρότερο από το 0.05, επομένως ο έλεγχος είναι στατιστικά σημαντικός και η μηδενική υπόθεση μπορεί να απορριφθεί. Από τον έλεγχο διασποράς, φαίνεται ότι υπάρχει τουλάχιστον μία δίαιτα που επηρεάζει διαφορετικά από τις υπόλοιπες το χρόνο πήξης του αίματος.

Εναλλακτικός τρόπος για απόρριψη ή αποδοχή του ελέγχου είναι να ελέγξουμε την τιμή F του ελέγχου με την κρίσιμη τιμή $F_{3,20;0.05}$.

```
Fcrit <- qf(p = 0.05, df1 = 3, df2 = 20, lower.tail = FALSE)
Fcrit
## [1] 3.098391
```

Από τα αποτελέσματα, προκύπτει ότι απορρίπτουμε τη μηδενική υπόθεση αφού $F = 13.57 > 3.0983 = Fcrit = F_{3,20;0.05}$.

Τέλος, μία τρίτη εναλλακτική, είναι να απεικονίσουμε γραφικά την κατανομή και την κρίσιμη τιμή του ελέγχου, καθώς και το στατιστικό του ελέγχου σε αυτή, με την εντολή ggaov της βιβλιοθήκης *gginference*.



```
library(gginference)
ggaov(coag.aov,
        colaccept = "grey89", #Χρώμα αποδεκτής περιοχής
        colreject = "black") #Χρώμα απορριπτικής περιοχής
```

Ο έλεγχος της ανάλυσης διασποράς δίνει απάντηση μόνο στο ερώτημα αν οι μέσες τιμές των δειγμάτων είναι ίσες ή όχι. Στην περίπτωση που δεν είναι ίσες οι μέσες τιμές, όπως στην συγκεκριμένη άσκηση, η ανάλυση διασποράς δε δίνει στοιχεία για το ποια ή ποιες μέσες τιμές διαφοροποιούνται από τις υπόλοιπες. Για να προσδιοριστούν ποιες μέσες τιμές διαφέρουν μεταξύ τους, πρέπει να γίνουν πολλαπλές συγκρίσεις μεταξύ των ομάδων (ανά δύο). Υπάρχουν διάφοροι μέθοδοι, όπως: Tukey, Hochberg's GT2, Gabriel, Scheffe, Bonferroni, LSD-Least Significant Difference, κ.α. Η πιο ευρέως διαδεδομένη είναι ο έλεγχος του Tukey.

Η μέθοδος Tukey στην R, υλοποιείται με την εντολή *TukeyHSD*, η οποία δέχεται ως παραμέτρους το αντικείμενο των αποτελεσμάτων της ANOVA, την μεταβλητή ως προς την οποία θα υπολογιστούν τα διαστήματα εμπιστοσύνης και το επίπεδο εμπιστοσύνης. Η *TukeyHSD*, υπολογίζει τα 95% διαστήματα εμπιστοσύνης για όλα τα ζεύγη διαφορών ανάμεσα στις μέσες τιμές των παραμέτρων. Αν το μηδέν συμπεριλαμβάνεται στο διάστημα μιας διαφοράς, τότε αυτή σχηματίζει μια κατηγορία.

Τα αποτελέσματά της *TukeyHSD* απαντούν στο ερώτημα: Ποιος παράγοντας ή ποιοι παράγοντες προκαλούν τις διαφορές;

Παρακάτω, υπολογίζουμε το 95% διαστήματος εμπιστοσύνης για όλα τα ζεύγη διαφορών ανάμεσα στις μέσες τιμές των κατηγοριών της δίαιτας (diet)

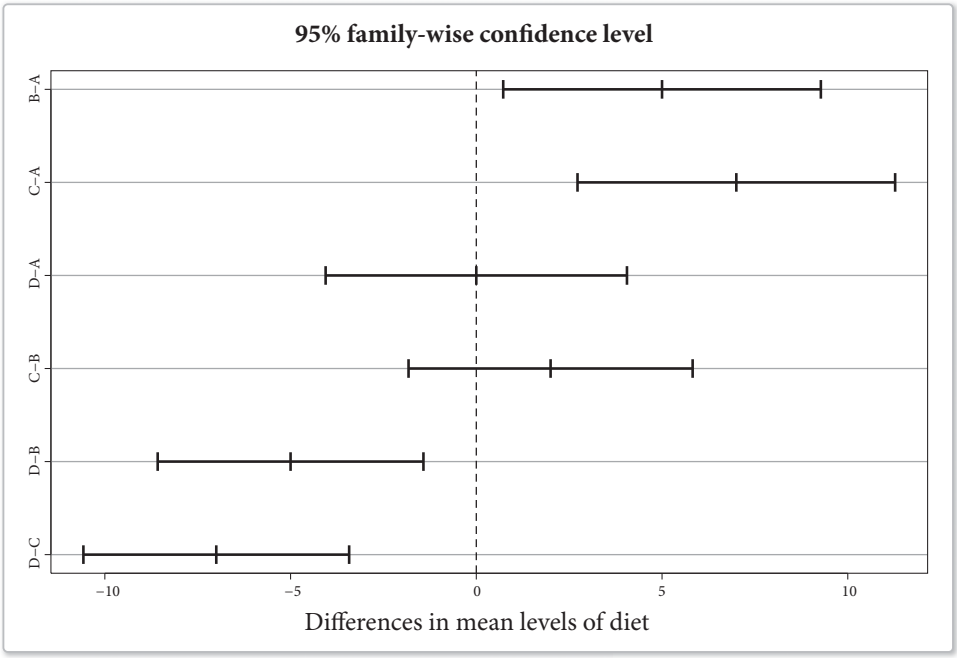
```
mca.coag <- TukeyHSD(
  x = coag.aov,
  which = "diet",
  conf.level = 0.95)

mca.coag
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = coag ~ diet, data = coag.df)
##
## $diet
##     diff         lwr       upr     p adj
## B-A    5   0.7245544  9.275446 0.0183283
## C-A    7   2.7245544 11.275446 0.0009577
## D-A    0  -4.0560438  4.056044 1.0000000
## C-B    2  -1.8240748  5.824075 0.4766005
## D-B   -5  -8.5770944 -1.422906 0.0044114
## D-C   -7 -10.5770944 -3.422906 0.0001268
```

Οι δίαιτες *A* και *D* σχηματίζουν μια κατηγορία, ενώ και οι *B* και *C* μια άλλη κατηγορία, αφού περιέχεται το μηδέν στο 95% διάστημα εμπιστοσύνης της διαφοράς τους.

Μπορούμε να δούμε το αποτέλεσμα γραφικά, μέσω της εντολής *plot* των αποτελεσμάτων του *TukeyHSD*.

```
plot(mca.coag)
```



Άσκηση-R 8.2

Στη βιβλιοθήκη datasets της R, βρίσκεται το σύνολο δεδομένων ChickWeight που περιέχει μετρήσεις για το σωματικό βάρος 50 πουλερικών που ακολούθησαν μία από 4 δίαιτες. Οι μετρήσεις του βάρους τους γίνονταν κάθε δύο μέρες.

a) Μπορούμε να ισχυριστούμε ότι και οι τέσσερις δίαιτες είχαν το ίδιο αποτέλεσμα στο τελικό βάρος των πουλερικών;
b) Αν όχι, μπορούμε να βρούμε ποια δίαιτα είναι πιο αποτελεσματική στην αύξηση του βάρους των πουλερικών;
c) Τέλος, οι δίαιτες που ακολουθούνται, φαίνεται να έχουν διαφορετική αύξηση του βάρους στα πουλερικά ήδη από τις 10 μέρες;

Σημείωση: το βάρος των πουλερικών ακολουθεί κανονική κατανομή, κάθε μέρα που έγινε η μέτρηση.

Ευρετήριο Όρων

Ευρετήριο Εντολών της R