ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΕΜΜΑΝΟΥΗΛ ΔΗΜΗΤΣΑΣ ΔΗΜΟΠΟΥΛΟΣ

«Έλεγχοι υποθέσεων για την κανονική κατανομή»

ΣΤΗΝ ΕΠΙΣΤΗΜΟΝΙΚΗ ΠΕΡΙΟΧΗ:

Στατιστικής

ΕΠΙΒΛΕΠΟΥΣΑ:

Βόντα Φιλία, Καθηγήτρια Ε.Μ.Π.

ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ

Βόντα Φιλία, Καθηγήτρια Ε.Μ.Π.

Παπανικολάου Βασίλειος, Ομότιμος Καθηγητής Ε.Μ.Π

Καρώνη Χρυσηίς, Ομότιμη Καθηγήτρια Ε.Μ.Π

[ΑΘΗΝΑ,2024]

ΕΥΧΑΡΙΣΤΙΕΣ

Αρχικά, θα ήθελα να ευχαριστήσω την επιβλέπουσα της Διπλωματικής Εργασίας κυρία Βόντα Φιλία, Καθηγήτρια Ε.Μ.Π., η οποία με έφερε σε επαφή με το παρόν θέμα και την πλούσια και ιδιαίτερα ενδιαφέρουσα σχετική βιβλιογραφία, ενώ μου προσέφερε επίσης πολύτιμη βοήθεια και καθοδήγηση κατά τη διάρκεια της εκπόνησής της. Ευχαριστώ επίσης τον κύριο Παπανικολάου Βασίλειο, Ομότιμο Καθηγητή Ε.Μ.Π. και την κυρία Καρώνη Χρυσηίδα, Ομότιμη Καθηγήτρια Ε.Μ.Π. για το χρόνο που αφιέρωσαν στην διπλωματική μου και την συμμετοχή τους στην τριμελή επιτροπή.

Επίσης, θα ήθελα να ευχαριστήσω ιδιαίτερα την οικογένειά μου και τους φίλους μου για τη στήριξη που μου παρείχαν καθ’ όλη τη διάρκεια των προπτυχιακών σπουδών μου.

....

ΔΗΜΗΤΣΑΣ-ΔΗΜΟΠΟΥΛΟΣ ΕΜΜΑΝΟΥΗΛ
Διπλωματούχος Σχολής Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών

Περίληψη

Η παρούσα διπλωματική εργασία ασχολείται με τον έλεγχο υποθέσεων για την κανονική κατανομή και σκοπεύει στην κατανόηση και την πρακτική εφαρμογή αυτής της διαδικασίας για τη λήψη αποφάσεων βασισμένων σε στατιστικά δεδομένα. Ο έλεγχος υποθέσεων αποτελεί ένα σημαντικό εργαλείο στον χώρο της στατιστικής και της επιστήμης των δεδομένων, καθώς επιτρέπει την αξιολόγηση της στατιστικής σημαντικότητας και την καθοδήγηση στη λήψη αποφάσεων βασισμένων σε παρατηρήσεις/δεδομένα. Αρχικά, γίνεται μια εισαγωγή σχετικά με την χρησιμότητα της κανονικής κατανομής σε διάφορους τομείς, όπως οικονομία, τεχνολογία και την επιστήμη. Στη συνέχεια, εξηγούμε θεμελιώδεις στατιστικές έννοιες που θα χρησιμοποιήσουμε στα επόμενα κεφάλαια. Έπειτα, παρουσιάζουμε μια εισαγωγή στους ελέγχους υποθέσεων, αναλύοντας τόσο την ιστορία όσο και την μαθηματική βάση τους, προκειμένου να κατανοήσουμε πώς μπορούν να χρησιμοποιηθούν για να ελέγξουμε εάν τα δεδομένα μας ακολουθούν την κανονική κατανομή. Επιπλέον, παρουσιάζουμε πώς μπορούν να εφαρμοστούν οι έλεγχοι υποθέσεων στην γλώσσα προγραμματισμού της R χρησιμοποιώντας μια γεννήτρια για τυχαία δεδομένα τα οποία ακολουθούν την κανονική κατανομή για να δείξουμε τα αντίστοιχα στατιστικά πακέτα που απαιτούνται για την υλοποίησή τους. Επιπλέον, χρησιμοποιώντας ένα πραγματικό σύνολο δεδομένων στο περιβάλλον της R, προβαίνουμε σε έλεγχο για να διαπιστώσουμε εάν τα δεδομένα ακολουθούν την κανονική κατανομή και λαμβάνουμε κάποια συμπεράσματα ανάλογα με τον έλεγχο που χρησιμοποιούμε καθώς και κάποιες διαφορές στα αποτελέσματα που εξάγουμε από τους διάφορους ελέγχους. Τέλος στην R φτιάχνουμε κάποια σημαντικά διαγράμματα για την αναπαράσταση των δεδομένων μας όπως το διάγραμμα QQ-plot, το ιστόγραμμα και το κυτιοδιάγραμμα με σκοπό την καλύτερη κατανόηση.

Λέξεις κλειδιά:

Κανονική κατανομή, ελεγχοσυνάρτηση, έλεγχος υποθέσεων, μηδενική υπόθεση, p-value, μέγεθος στατιστικού ελέγχου, διαγράμματα, στατιστικό πακέτο R

Abstract

This thesis focuses on hypothesis testing for the normal distribution and aims to understand and apply these processes in practice in order to make decisions based on statistical data. Hypothesis testing is a crucial tool in the field of statistics and data science, as it allows for the evaluation of statistical significance and provides guidance for decision-making based on observations/data. Initially, we provide an introduction regarding the utility of the normal distribution in various fields such as economics, technology and science. Subsequently, we explore fundamental statistical concepts that will be used in the following chapters. We then provide an introduction to hypothesis testing, exploring both its historical context and mathematical foundations. This helps us understand how hypothesis tests can be used to determine whether our data follow a normal distribution. Additionally, we showcase how hypothesis tests can be applied in the R programming language, using a random data generator that generates data that follow the normal distribution. We also discuss the relevant statistical packages required for their implementation. Moreover, by employing a real dataset within the R environment, we conduct a test to determine if the data follow a normal distribution. We draw conclusions based on the different types of tests that we employed, as well as discuss the differences in the results extracted from the various tests. Finally, we create essential graphical representations of our data in R, such as the QQ-plot, histogram, and boxplot, with the aim of enhancing our understanding.

Keywords:

Normal distribution, statistical test function, hypothesis testing, null hypothesis, p-value, size of a statistical test, diagrams, statistical package R

Περιεχόμενα

ΠΡΟΛΟΓΟΣ

Η σύγχρονη επιστήμη και επιχειρηματικότητα επικεντρώνονται στην χρήση πολύτιμων δεδομένων για τη λήψη στρατηγικών αποφάσεων. Οι δεδομένες πληροφορίες μπορούν να αποκτήσουν ένα νέο επίπεδο νόησης όταν εξετάζονται μέσω στατιστικών μεθόδων και εργαλείων. Σε αυτό το πλαίσιο, η παρούσα διπλωματική εργασία ασχολείται με ένα σημαντικό θέμα της στατιστικής: τον έλεγχο υποθέσεων για την κανονική κατανομή.

Ο έλεγχος υποθέσεων αποτελεί το εργαλείο που μας επιτρέπει να αντιληφθούμε τη στατιστική σημαντικότητα στα δεδομένα μας και να λάβουμε αποφάσεις με βάση αυτή. Από την ανάλυση μέσω κατάλληλων ελέγχων για την κανονική κατανομή, η οποία αποτελεί τον πυρήνα της παρούσας μελέτης, οι ερευνητές μπορούν να ανακαλύψουν αν οι παρατηρήσεις τους ακολουθούν την κανονική κατανομή, η οποία έχει σημαντική εφαρμογή σε πολλούς τομείς, όπως η οικονομία, η τεχνολογία, και η επιστήμη.

Μέσα από αυτήν τη διπλωματική εργασία, θα εξερευνήσουμε τις θεωρητικές και πρακτικές πτυχές του ελέγχου υποθέσεων βασισμένων στην κανονική κατανομή. Θα προχωρήσουμε από την ανάλυση της θεωρίας και των βασικών στατιστικών εννοιών που αφορούν αυτήν τη μέθοδο, προς την πρακτική εφαρμογή της χρησιμοποιώντας την γλώσσα προγραμματισμού της R. Τέλος, με τη χρήση πραγματικών δεδομένων, θα αξιολογήσουμε αν αυτά τα δεδομένα ακολουθούν την κανονική κατανομή και θα εξάγουμε σημαντικά συμπεράσματα.

Αναμένεται ότι η παρούσα διπλωματική εργασία θα αποτελέσει μια χρήσιμη πηγή γνώσης και κατανόησης για όλους όσους ενδιαφέρονται για την εφαρμογή της στατιστικής και την προετοιμασία των δεδομένων τους με ακρίβεια και αξιοπιστία.

Αυτός ο πρόλογος θέτει το πλαίσιο για τη Διπλωματική εργασία και προετοιμάζει για την ανάλυση της κανονικής κατανομής τόσο σε θεωρητικό κομμάτι όσο και σε πρακτικό, χρησιμοποιώντας ένα πραγματικό σύνολο δεδομένων στην R.

ΚΕΦΑΛΑΙΟ 1

ΧΡΗΣΙΜΟΤΗΤΑ ΚΑΙ ΙΣΤΟΡΙΑ ΚΑΝΟΝΙΚΗΣ ΚΑΤΑΝΟΜΗΣ

1.1 Περιγραφή και χρησιμότητα κανονικής κατανομής

Η κανονική κατανομή (γνωστή και ως γκαουσιανή κατανομή) σχετίζεται και περιγράφει συνεχείς μεταβλητές και αποτελεί μία συνεχή συνάρτηση πυκνότητας πιθανότητας. Χρησιμοποιείται πολλές φορές ως μία πρώτη προσέγγιση για να περιγράφουν πραγματικά δεδομένα, τα οποία τείνουν να συγκεντρώνονται γύρω από μια μέση τιμή. Από πολλούς θεωρείται η σπουδαιότερη κατανομή της Θεωρίας Πιθανοτήτων και Στατιστικής και αυτό διότι πολλές τυχαίες μεταβλητές περιγράφονται ικανοποιητικά από την κανονική κατανομή ή περιγράφονται από κατανομές που μπορούν να προσεγγισθούν από αυτή. Επίσης οι ιδιότητές της αξιοποιούνται στην Στατιστική Συμπερασματολογία, δηλαδή αποτελεί το θεμέλιο της Στατιστικής Συμπερασματολογίας. Η κανονική κατανομή είναι ιδιαίτερα χρήσιμη για την περίπτωση που έχουμε να κάνουμε συγκρίσεις μεταξύ μεγάλου πλήθους δεδομένων και έχει πολλές εφαρμογές στην επιστήμη, την τεχνολογία και τον οικονομικό τομέα.

Επιστήμη

Φυσική: Η κανονική κατανομή χρησιμοποιείται σε πολλούς κλάδους της φυσικής.

- Θερμοδυναμική: Στη θερμοδυναμική, η κανονική κατανομή χρησιμοποιείται συχνά για να μοντελοποιήσει την κατανομή των ενεργειακών καταστάσεων μέσα σε ένα σύστημα.
- Κβαντομηχανική: Στην κβαντομηχανική, χρησιμοποιείται για να μοντελοποιήσει την κατανομή των κυματοσυναρτήσεων, οι οποίες περιγράφουν την πιθανότητα εύρεσης ενός σωματιδίου σε μια δεδομένη θέση.

- Αστροφυσική: Στην αστροφυσική, χρησιμοποιείται για τη μοντελοποίηση της κατανομής ιδιοτήτων όπως η μάζα, η ταχύτητα και η θερμοκρασία σε έναν γαλαξία ή άλλο αστρονομικό σύστημα.
- Ρευστοδυναμική: Στον τομέα της δυναμικής των ρευστών, χρησιμοποιείται συχνά για τη μοντελοποίηση της κατανομής των ταχυτήτων μέσα σε ένα ρευστό, όπως η κατανομή θερμοκρασίας μέσα στο ρευστό.
- Ατομική και μοριακή φυσική: Στο πεδίο της ατομικής και μοριακής φυσικής, χρησιμοποιείται για τη μοντελοποίηση της κατανομής των ενεργειακών επιπέδων, των θέσεων και των ροπών των σωματιδίων στο σύστημα. Μπορεί επίσης να χρησιμοποιηθεί για την πρόβλεψη της έκβασης των συγκρούσεων και για την κατανόηση των ιδιοτήτων του υλικού όπως η αγωγιμότητα, η διαχυτικότητα και η θερμική αγωγιμότητα.

· **Βιολογία**: Η κανονική κατανομή έχει πολλές εφαρμογές στους κλάδους της βιολογίας, κάποια παραδείγματα είναι:

- Γενετική πληθυσμού: Στη γενετική πληθυσμού, η κανονική κατανομή χρησιμοποιείται συχνά για να μοντελοποιήσει την κατανομή μιας γενετικής παραλλαγής μέσα σε έναν πληθυσμό.
- Εξελικτική βιολογία: Στην εξελικτική βιολογία, η κανονική κατανομή χρησιμοποιείται συχνά για να μοντελοποιήσει την κατανομή των τιμών των χαρακτηριστικών μέσα σε έναν πληθυσμό.
- Επιδημιολογία: Στην επιδημιολογία, η κανονική κατανομή χρησιμοποιείται συχνά για τη μοντελοποίηση της κατανομής των ποσοστών μολυσματικών ασθενειών σε έναν πληθυσμό.
- Νευροεπιστήμη: Στη νευροεπιστήμη, η κανονική κατανομή χρησιμοποιείται συχνά για τη μοντελοποίηση της κατανομής της εγκεφαλικής δραστηριότητας ή της νευρικής σηματοδότησης.

· **Χημεία**: Όπως στους κλάδους της φυσικής και της βιολογίας έτσι και στην Χημεία η κανονική κατανομή χρησιμοποιείται :

- Αναλυτική χημεία: Στην αναλυτική χημεία, η κανονική κατανομή χρησιμοποιείται συχνά για να μοντελοποιήσει την κατανομή των σφαλμάτων ή των αβεβαιοτήτων στις μετρήσεις.
- Φυσικοχημεία : Στη φυσικοχημεία , η κανονική κατανομή χρησιμοποιείται συχνά για τη μοντελοποίηση της κατανομής των μοριακών ιδιοτήτων, όπως τα μήκη των δεσμών, οι γωνίες των δεσμών και οι μοριακές ενέργειες.

- Μοριακή μοντελοποίηση: Στη μοριακή μοντελοποίηση, η κανονική κατανομή χρησιμοποιείται συχνά για τη μοντελοποίηση της κατανομής των ατομικών θέσεων μέσα σε ένα μόριο.
- Ανακάλυψη φαρμάκων: Στον τομέα της ανακάλυψης φαρμάκων, η κανονική κατανομή χρησιμοποιείται συχνά για τη μοντελοποίηση της κατανομής των φαρμακολογικών ιδιοτήτων των υποψηφίων φαρμάκων, όπως η ισχύς, η αποτελεσματικότητά τους και οι παρενέργειές τους.

Τεχνολογία

Η κανονική κατανομή μπορεί να αξιοποιηθεί με πολλούς τρόπους στον τομέα της τεχνολογίας:

- Ποιοτικός έλεγχος: Στην κατασκευή, η κανονική κατανομή χρησιμοποιείται συχνά για τη μοντελοποίηση της κατανομής των μετρήσεων των διαστάσεων ή άλλων χαρακτηριστικών ενός προϊόντος. Αυτό μπορεί να είναι χρήσιμο για τον εντοπισμό και τη διόρθωση προβλημάτων στη διαδικασία κατασκευής που μπορεί να οδηγήσουν σε ελαττωματικά προϊόντα.
- Αξιολόγηση απόδοσης: Στην επιστήμη των υπολογιστών, η κανονική κατανομή χρησιμοποιείται συχνά για τη μοντελοποίηση της απόδοσης ενός συστήματος ή ενός δικτύου υπολογιστών. Για παράδειγμα, η κατανομή των χρόνων απόκρισης για έναν διακομιστή web θα μπορούσε να μοντελοποιηθεί χρησιμοποιώντας την κανονική κατανομή.
- Επεξεργασία εικόνας: Στον τομέα της επεξεργασίας εικόνας, η κανονική κατανομή χρησιμοποιείται συχνά ως κατανομή αναφοράς για σύγκριση και ταξινόμηση εικόνων. Για παράδειγμα, η κατανομή των τιμών των pixel σε μια εικόνα μπορεί να συγκριθεί με μια κανονική κατανομή για τον εντοπισμό μοτίβων ή ανωμαλιών.
- Sensor data (Δεδομένα αισθητήρα): Η κανονική κατανομή χρησιμοποιείται συχνά για τη μοντελοποίηση της κατανομής των μετρήσεων. Για παράδειγμα, ανάλογα με τη θερμοκρασία μπορεί να παράγει μετρήσεις που κατανέμονται κανονικά, με μέση τιμή που αντιστοιχεί στη θερμοκρασία περιβάλλοντος και τυπική απόκλιση που αντανακλά την ακρίβεια του αισθητήρα.

Οικονομία

Στην οικονομία η κανονική κατανομή χρησιμοποιείται:

- Χρηματοοικονομική μοντελοποίηση: Στα χρηματοοικονομικά, η κανονική κατανομή χρησιμοποιείται συχνά για τη μοντελοποίηση της κατανομής των αποδόσεων των μετοχών, των αποδόσεων των ομολόγων και άλλων χρηματοοικονομικών μεταβλητών. Αυτό μπορεί να είναι χρήσιμο για την πρόβλεψη της πιθανότητας ορισμένων οικονομικών αποτελεσμάτων και για τη δημιουργία χρηματοοικονομικών χαρτοφυλακίων.

- Οικονομικές προβλέψεις: Στα οικονομικά, η κανονική κατανομή χρησιμοποιείται συχνά για τη μοντελοποίηση της κατανομής των οικονομικών μεταβλητών, όπως το ακαθάριστο εγχώριο προϊόν (ΑΕΠ), ο πληθωρισμός και τα ποσοστά απασχόλησης. Έτσι η κανονική κατανομή βοηθάει στην πρόβλεψη της μελλοντικής κατεύθυνσης της οικονομίας και στην αξιολόγηση της αποτελεσματικότητας των οικονομικών πολιτικών.
- Συμπεριφορά καταναλωτή: Στη μελέτη της συμπεριφοράς των καταναλωτών, η κανονική κατανομή χρησιμοποιείται συχνά για τη μοντελοποίηση της κατανομής μεταβλητών όπως το εισόδημα, ο πλούτος και τα πρότυπα κατανάλωσης. Για παράδειγμα αυτό μπορεί να βοηθήσει στην πρόβλεψη της καταναλωτικής ζήτησης και στην κατανόηση των παραγόντων που επηρεάζουν τη λήψη αποφάσεων από τους καταναλωτές.
- Ανάλυση κινδύνου: Στο πεδίο της ανάλυσης κινδύνου, η κανονική κατανομή χρησιμοποιείται συχνά για τη μοντελοποίηση της κατανομής των κινδύνων που σχετίζονται με χρηματοοικονομικές επενδύσεις, επιχειρηματικά εγχειρήματα και άλλες οικονομικές δραστηριότητες. Αυτό μπορεί να είναι χρήσιμο για τον εντοπισμό και τη διαχείριση των κινδύνων και για τη λήψη τεκμηριωμένων αποφάσεων.

1.2 Ιστορική αναδρομή

Η κανονική κατανομή παρουσιάστηκε για πρώτη φορά από το Γάλλο Μαθηματικό Abraham DeMoivre το 1733. Ο DeMoivre ο οποίος χρησιμοποίησε την κατανομή αυτή για να προσεγγίσει πιθανότητες που σχετίζονται με το στρίψιμο νομισμάτων την ονόμασε εκθετική κωδωνοειδή καμπύλη. Ωστόσο η χρησιμότητά της αποκαλύφθηκε πλήρως το 1809 όταν ο διάσημος μαθηματικός Karl Friedrich Gauss την χρησιμοποίησε ως ζωτικό κομμάτι της μεθόδου του για την πρόβλεψη της θέσης ουράνιων σωμάτων. Μετά από αυτό πολλοί την αποκαλούν

Κατανομή Gauss. Όμως στα μέσα με τέλη του 19ου αιώνα οι περισσότεροι στατιστικοί άρχισαν να πιστεύουν ότι η πλειονότητα των συνόλων δεδομένων θα έπρεπε να είχαν ιστογράμματα που συμφωνούν με την κωδωνοειδή καμπύλη της κατανομής Gauss. Πράγματι άρχισε να γίνεται αποδεκτό ως «κανονικό» το γεγονός ότι κάθε «καλώς συμπεριφερόμενο» σύνολο δεδομένων πρέπει να ακολουθεί την καμπύλη αυτή. Για το λόγο αυτό οι άνθρωποι ακολούθησαν το παράδειγμα του Βρετανού στατιστικού Karl Pearson και άρχισαν να αναφέρονται στην Καμπύλη Gauss ως κανονική καμπύλη.

Abraham DeMoivre (1667-1754)

Σήμερα υπάρχει αφθονία συμβούλων στατιστικής, πολλοί από τους οποίους χειρίζονται τις υποθέσεις τους σε πολύ κομψά επιχειρηματικά περιβάλλοντα. Παρόλα αυτά ο πρωτοπόρος του είδους τους εργαζόταν στις αρχές του 18ου αιώνα σε ένα σκοτεινό και βρώμικο πρακτορείο στοιχημάτων στο Long Acres του Λονδίνου γνωστό ως Slaughter's Coffee House (Καφενείο των σφαγέων). Ήταν ο Abraham DeMoivre, ένας προτεστάντης πρόσφυγας από την Καθολική Γαλλία και μπορούσε με το ανάλογο αντίτιμο να υπολογίσει την πιθανότητα των στοιχημάτων για διάφορους τύπους τυχερών παιχνιδιών. Αν και ο DeMoivre, δηλαδή ο άνθρωπος που ανακάλυψε την κανονική καμπύλη έβγαζε το ψωμί του στο καφενείο ήταν ένας μαθηματικός εγνωσμένων ικανοτήτων. Πράγματι , ήταν μέλος της Βασιλικής Ακαδημίας ενώ υπήρχαν αναφορές ότι ήταν στενός φίλος με τον Isaac Newton. Να πως ο Karl Pearson φανταζόταν τον Abraham DeMoivre να εργάζεται στο Slaughter's Coffee House: «Φαντάζομαι το DeMoivre να δουλεύει πάνω σε ένα βρώμικο τραπέζι του καφενείου με κάποιο κατεστραμμένο τζογαδόρο δίπλα του και τον Isaac Newton να περπατάει μέσα στο πλήθος για να βγάλει το φίλο του από αυτό το καταγώγιο».

Karl Friedrich Gauss (1777-1855)

O Karl Friedrich Gauss ο οποίος ήταν ένας από τους πρώτους χρήστες της κανονικής καμπύλης, θεωρείται ένας από τους μεγαλύτερους μαθηματικούς όλων των εποχών. Ακολουθούν τα λόγια του γνωστού ιστορικού των μαθηματικών E.T. Bell όπως εκφράστηκαν στο βιβλίο του Men of Mathematics το 1954. Στο κεφάλαιο με τίτλο «The Prince of Mathematicians» γράφει «ο Αρχιμήδης, ο Newton και ο Gauss , αυτοί οι τρεις αποτελούν μόνοι τους μία ξεχωριστή κλάση ανάμεσα στους μεγάλους Μαθηματικούς και δεν είναι δουλειά των κοινών θνητών να επιχειρήσουν να τους κατατάξουν ως προς την αξία τους». Και οι τρεις προκάλεσαν παλιρροϊκά κύματα τόσο στα θεωρητικά όσο και στα εφαρμοσμένα μαθηματικά. Ο Αρχιμήδης έτρεφε μεγάλη εκτίμηση για τα θεωρητικά μαθηματικά απ' ότι για τις εφαρμογές τους. Για τον Newton, φαίνεται ότι οι επιστημονικές χρήσεις των μαθηματικών του επινοήσεων αποτελούσαν τη βασική καταξίωσή τους, ενώ ο Gauss δήλωνε ότι του ήταν αδιάφορο αν δούλευε στα θεωρητικά ή στα εφαρμοσμένα μαθηματικά.

Ιστορική σημείωση : Κεντρικό οριακό θεώρημα

Pierre – Simon , Marquis de Laplace

Το κεντρικό οριακό θεώρημα διατυπώθηκε αρχικά και αποδείχθηκε από το Γάλλο μαθηματικό Pierre – Simon, Marquis de Laplace , ο οποίος κατέληξε σε αυτό έπειτα από τις παρατηρήσεις του ότι τα σφάλματα των μετρήσεων (τα οποία μπορούν συνήθως να θεωρηθούν ως το άθροισμα ενός μεγάλου πλήθους στοιχειωδών δυνάμεων) τείνουν να έχουν κανονική κατανομή. Ο Laplace που ήταν και διάσημος αστρονόμος (τον αποκαλούσαν Newton της Γαλλίας), ήταν ένας από τους πρώτους μεγάλους θεμελιωτές των πιθανοτήτων και της στατιστικής. Υποστήριζε τη χρήση των πιθανοτήτων στην καθημερινή ζωή. Πίστευε με πάθος στη σπουδαιότητά τους όπως αποδεικνύεται από τα ακόλουθα χωρία που προέρχονται από το δημοσιευμένο βιβλίο του με τίτλο Analytical Theory of Probability: «Βλέπουμε ότι η θεωρία των πιθανοτήτων είναι κατά βάθος η κοινή λογική που έχει αναχθεί σε αριθμητικούς υπολογισμούς, μας βοηθάει να εκτιμήσουμε με ακρίβεια αυτά που ένας κοινός νους αισθάνεται ενστικτωδώς χωρίς συχνά να μπορεί να τα αιτιολογήσει. Είναι αξιοσημείωτο ότι αυτή η επιστήμη, που γεννήθηκε από τη μελέτη των τυχερών παιχνιδιών θα γίνει ενδεχομένως το πιο σημαντικό πεδίο της ανθρώπινης γνώσης. Τα σημαντικότερα ερωτήματα της ζωής είναι ουσιαστικά σε μεγάλο βαθμό προβλήματα πιθανοτήτων.»
Η εφαρμογή του Κεντρικού Οριακού Θεωρήματος στην απόδειξη ότι τα σφάλματα των μετρήσεων έχουν κατά προσέγγιση κανονική κατανομή θεωρείται μία από τις σημαντικότερες συνεισφορές στην επιστήμη. Πράγματι, τον $17^{ο}$ και τον $18^{ο}$ αιώνα, το Κεντρικό Οριακό Θεώρημα συχνά ονομαζόταν Νόμος της συχνότητας των σφαλμάτων (Law of frequency of errors). Ακολουθούν τα λόγια του Francis Galton από το βιβλίο με τίτλο Natural Inheritance που εκδόθηκε το 1889. «Δεν γνωρίζω τίποτα που να εξάπτει την φαντασία τόσο όσο η θαυμάσια μορφή της κοσμικής τάξης, όπως αυτή εκφράζεται μέσα από το νόμο της συχνότητας των σφαλμάτων. Ο Νόμος θα είχε προσωποποιηθεί από τους Έλληνες και θα είχε θεοποιηθεί αν τον ήξεραν. Βασιλεύει με ηρεμία και πλήρη ταπεινοφροσύνη στη μέση της πιο άγριας σύγχυσης. Όσο πιο τεράστιος είναι ο όχλος κι όσο μεγαλύτερη είναι η προφανής αναρχία, τόσο τελειότερη είναι η κυριαρχία του. Είναι ο ανώτατος νόμος του παραλόγου».

1.3 Βασικοί ορισμοί στη Στατιστική

Πληθυσμός

Ορίζεται ένα βασικό σύνολο στοιχείων που πρόκειται να μελετηθεί ως προς μία ή περισσότερες χαρακτηριστικές ιδιότητες. Είναι το σύνολο δηλαδή των απλών στοιχείων για τα οποία μας ενδιαφέρει να βγάλουμε κάποια συμπεράσματα.

Μεταβλητές

Τα χαρακτηριστικά ως προς τα οποία εξετάζουμε ένα πληθυσμό λέγονται μεταβλητές. Τις μεταβλητές τις διακρίνουμε σε ποσοτικές και σε ποιοτικές ή κατηγορικές. Οι ποσοτικές μεταβλητές διακρίνονται ανάλογα με το πλήθος των τιμών που μπορούν να πάρουν σε διακριτές και συνεχείς.

Διακριτή θεωρείται η μεταβλητή που λαμβάνει ένα σύνολο τιμών στις οποίες μπορούν να αντιστοιχισθούν ένα προς ένα, στοιχείο του συνόλου των φυσικών αριθμών. Αντίθετα, μία μεταβλητή χαρακτηρίζεται σαν συνεχής εάν παίρνει, θεωρητικά τουλάχιστον, κάθε τιμή στο διάστημα (α,β). Ποιοτικές μεταβλητές είναι εκείνες για τις οποίες δεν είναι δυνατόν να ληφθούν πλήρως καθοριστικές τιμές αλλά μόνο ενδεικτικές ή συγκριτικές.

Δείγμα

Είναι ένα σύνολο μονάδων που επιλέγονται από τον πληθυσμό. Το δείγμα επιλέγεται για να μελετηθεί και να εξαχθούν έγκυρα συμπεράσματα για τον πληθυσμό, χωρίς να μελετηθεί όλος ο πληθυσμός γιατί είναι πολύ μεγάλος. Το δείγμα πρέπει να είναι αντιπροσωπευτικό του πληθυσμού και γι' αυτό παίζει ρόλο το μέγεθός του και η μέθοδος δειγματοληψίας που χρησιμοποιείται.

Πείραμα

Είναι οποιαδήποτε διαδικασία ή μελέτη, η οποία έχει ως αποτέλεσμα τη συλλογή των δεδομένων, η έκβαση της οποίας είναι άγνωστη.

Συνάρτηση πυκνότητας πιθανότητας

Έστω ότι X είναι μία συνεχής τυχαία μεταβλητή. Αν υπάρχει μία μη αρνητική συνάρτηση f ορισμένη για κάθε πραγματικό αριθμό $x \in R$ η οποία έχει την εξής ιδιότητα : για κάθε σύνολο B με πραγματικούς αριθμούς ισχύει

$$P[x \in B] = \int_B f(x)dx \quad \text{και} \quad \int_R f(x)dx = 1$$

τότε η συνάρτηση f ονομάζεται συνάρτηση πυκνότητας πιθανότητας της τυχαίας μεταβλητής X.

Λόγω της ενασχόλησής μας με την κανονική κατανομή σε αυτή την εργασία, δε θα ασχοληθούμε με τις διακριτές τυχαίες μεταβλητές για τις οποίες ορίζεται αντίστοιχα η συνάρτηση μάζας πιθανότητας.

1.4 Βασικοί ορισμοί στην Εκτιμητική

Θεωρία εκτίμησης - Σημειοεκτιμητική

Ο κλάδος της στατιστικής που ασχολείται με την σημειακή εκτίμηση των τιμών των παραμέτρων με βάση εμπειρικά δεδομένα. Οι παράμετροι περιγράφουν μια υποκείμενη φυσική ρύθμιση-ιδιότητα του πληθυσμού με τέτοιο τρόπο ώστε οι τιμές τους να επηρεάζουν την κατανομή των δεδομένων. Μία συνάρτηση - εκτιμήτρια επιχειρεί να χρησιμοποιήσει τις μετρήσεις για να προσεγγίσει τις άγνωστες παραμέτρους.

Εκτιμήτρια
Μια συνάρτηση των γνωστών δεδομένων που χρησιμοποιείται για την εκτίμηση μιας άγνωστης παραμέτρου. Μια εκτίμηση είναι το αποτέλεσμα της πραγματικής εφαρμογής της συνάρτησης σε ένα συγκεκριμένο σύνολο δεδομένων. Για παράδειγμα, ο μέσος όρος μπορεί να χρησιμοποιηθεί ως εκτιμήτρια της μέσης τιμής.

Αναμενόμενη τιμή
Το άθροισμα (διακριτό ή συνεχές) των πιθανοτήτων κάθε πιθανού αποτελέσματος ενός πειράματος πολλαπλασιασμένο με την αντίστοιχη τιμή του. Η έννοια είναι, διαισθητικά, μια γενίκευση του σταθμισμένου μέσου όρου όλων των πιθανών αποτελεσμάτων μιας συγκεκριμένης διαδικασίας ή πειράματος και μπορεί να θεωρηθεί ως ο αριθμητικός μέσος όρος ενός μεγάλου αριθμού ανεξάρτητων πραγματοποιήσεων του πειράματος. Η αναμενόμενη τιμή της τυχαίας μεταβλητής X γράφεται συνήθως ως $E(X)$.

1.5 Έλεγχοι υποθέσεων

1.5.1 Ιστορία

Ενώ ο έλεγχος υποθέσεων έγινε ευρέως γνωστός στις αρχές του 20ου αιώνα, οι πρώιμες μορφές του χρησιμοποιήθηκαν το 1700. Η πρώτη χρήση αποδίδεται στον John Arbuthnot (1710) και ακολούθησε ο Pierre-Simon Laplace (δεκαετία 1770), στην ανάλυση της αναλογίας των ανθρώπινων φύλων κατά τη γέννηση.

1.5.2 Βασικοί Ορισμοί

Στατιστική υπόθεση
Ονομάζεται κάθε υπόθεση που αφορά την κατανομή μιας τυχαίας μεταβλητής X. Συγκεκριμένα η υπόθεση (συμβολιζόμενη συνήθως με H (Hypothesis)) είναι δυνατόν να αφορά (α) την άγνωστη παράμετρο έστω θ (ή μια συνάρτηση του θ) που υπεισέρχεται στην κατανομή της τυχαίας μεταβλητής X ή (β) τη συναρτησιακή μορφή της κατανομής της τυχαίας μεταβλητής. Στην πρώτη περίπτωση έχουμε πρόβλημα ελέγχου παραμετρικής υπόθεσης

(parametric test) ενώ στη δεύτερη έχουμε πρόβλημα ελέγχου καταλληλόλητας ή καλής προσαρμογής κατανομής (goodness of fit test).

Εναλλακτική υπόθεση

Λαμβάνεται να είναι η άποψη που έχει διατυπωθεί, ενώ η **μηδενική υπόθεση** λαμβάνεται να είναι η άρνηση της άποψης που έχει διατυπωθεί. Οι δύο υποθέσεις μαζί αποτελούν τη βάση ενός ελέγχου υποθέσεων.

Σφάλμα τύπου Ι

Το Σφάλμα τύπου Ι ενός στατιστικού ελέγχου είναι η πιθανότητα της εσφαλμένης απόρριψης μιας μηδενικής υπόθεσης η οποία συμβολίζεται H_0. Συνήθως ένα σφάλμα τύπου Ι οδηγεί στο συμπέρασμα ότι μια υποτιθέμενη επίδραση ή σχέση υπάρχει όταν στην πραγματικότητα δεν υπάρχει, δηλαδή το διαπράττουμε όταν απορρίπτουμε την H_0 ενώ είναι στην πραγματικότητα αληθής.

Σφάλμα τύπου ΙΙ

Το Σφάλμα τύπου ΙΙ ενός στατιστικού ελέγχου είναι η αποτυχία να απορρίψουμε μια ψευδή μηδενική υπόθεση δηλαδή το διαπράττουμε όταν αποδεχόμαστε την H_0 ενώ στην πραγματικότητα η H_1 είναι αληθής.

Επίπεδο σημαντικότητας

Το επίπεδο σημαντικότητας ενός στατιστικού ελέγχου υποθέσεων είναι μια σταθερή πιθανότητα (η οποία προεπιλέγεται) κακής απόρριψης της μηδενικής υπόθεσης H_0, όταν αυτή είναι όντως αληθινή.
Είναι η πιθανότητα σφάλματος τύπου Ι και καθορίζεται από τον ερευνητή σε σχέση με τις συνέπειες ενός τέτοιου σφάλματος. Δηλαδή, θέλουμε να καταστήσουμε το επίπεδο σημαντικότητας όσο το δυνατόν μικρότερο, προκειμένου να προστατευθεί η μηδενική υπόθεση και να αποτρέψει, στο μέτρο του δυνατού, τον εξεταστή από ακούσιους ψευδείς ισχυρισμούς. Το επίπεδο σημαντικότητας συνήθως συμβολίζεται με α. Επίπεδο σημαντικότητας $P(\sigma\varphi\acute{\alpha}\lambda\mu\alpha\ \tau\acute{\upsilon}\pi o\upsilon\ I) = \alpha$. Συνήθως, το επίπεδο σημαντικότητας επιλέγεται να είναι 0,05 (ή ισοδύναμα, 5%).

Στατιστική ελεγχοσυνάρτηση

Ονομάζεται η στατιστική συνάρτηση η οποία χρησιμοποιείται για τον έλεγχο της υπό μελέτης υπόθεσης. Είναι μία συνάρτηση τυχαίων μεταβλητών και δεν μπορεί να εξαρτάται από άγνωστες παραμέτρους. Οι τιμές που παίρνει η ελεγχοσυνάρτηση καθορίζουν αν θα γίνει δεκτή ή θα απορριφθεί η μηδενική υπόθεση.

Κρίσιμη περιοχή (χωρίο απορρίψεως)

Η Κρίσιμη περιοχή ή αλλιώς χωρίο απορρίψεως είναι ένα σύνολο τιμών της ελεγχοσυνάρτησης, για το οποίο η μηδενική υπόθεση απορρίπτεται σε έναν έλεγχο υποθέσεων. Πιο συγκεκριμένα αν ο δειγματικός χώρος για την ελεγχοσυνάρτηση χωριστεί σε δύο περιοχές τότε η μια θα οδηγεί σε απόρριψη της μηδενικής υπόθεσης ενώ η άλλη σε αποδοχή. Έτσι αν η παρατηρούμενη τιμή της ελεγχοσυνάρτησης ανήκει στην κρίσιμη περιοχή τότε υποδηλώνεται ότι απορρίπτουμε την μηδενική υπόθεση. Αλλιώς δεν απορρίπτουμε την μηδενική υπόθεση (πάντα με δεδομένη πιθανότητα σφάλματος τύπου Ι).

Ισχύς Στατιστικού ελέγχου

Η ισχύς ενός στατιστικού ελέγχου υποθέσεων μετρά την ικανότητα του τεστ να απορρίπτει τη μηδενική υπόθεση όταν είναι στην πραγματικότητα ψευδής - δηλαδή, να πάρει μια σωστή απόφαση. Με άλλα λόγια, η ισχύς ενός ελέγχου υποθέσεων είναι η πιθανότητα να μην συμβαίνει ένα σφάλμα τύπου ΙΙ.

Υπολογίζεται αφαιρώντας την πιθανότητα σφάλματος τύπου ΙΙ από τη μονάδα και συνήθως εκφράζεται ως: Ισχύς = $1 - P(\sigma\varphi\acute{\alpha}\lambda\mu\alpha\ \tau\acute{\upsilon}\pi o\upsilon\ II)$=$1 - \beta$. Η μέγιστη ισχύς που μπορεί να έχει ένας έλεγχος είναι 1, το ελάχιστο είναι 0. Ιδανικά θέλουμε έναν έλεγχο να έχει μεγάλη ισχύ, κοντά στο 1.

P-Value ή πιθανότητα σημαντικότητας

Είναι η μικρότερη τιμή του επιπέδου σημαντικότητας α για την οποία η μηδενική υπόθεση H_0 απορρίπτεται. Αν η p-value<α απορρίπτουμε την H_0.

Έλεγχοι καλής προσαρμογής

Έλεγχοι που προσπαθούν να απαντήσουν σε υποθέσεις με ερώτημα : «ακολουθούν οι παρατηρήσεις μας κάποια συγκεκριμένη κατανομή;». Σχεδιάζονται για να συγκρίνουν το δείγμα με τον τύπο του δείγματος που θα περιμέναμε να έχουμε από την κατανομή που υποθέτουμε, ώστε να μπορούμε να δούμε αν η υποτεθείσα συνάρτηση κατανομής προσαρμόζεται στα δεδομένα του δείγματος. Ο έλεγχος αυτός χρησιμοποιείται σε περιπτώσεις προβλημάτων στα οποία ενδιαφερόμαστε να εξετάσουμε αν τα δεδομένα προέρχονται από μια ορισμένη κατανομή, δηλαδή, για τον έλεγχο υποθέσεων της μορφής

$H_0: F_x(x) = F_0(x)$

$H_1: F_x(x) = F_c(x)$

όπου $F_x(x) = P(X \leq x)$ είναι η συνάρτηση κατανομής της τυχαίας μεταβλητής X και $F_0(x)$ γνωστή συνάρτηση κατανομής και $F_c(x)$ κάποια άλλη γνωστή συνάρτηση κατανομής.

Ο αρχαιότερος και γνωστότερος τέτοιος έλεγχος καλής προσαρμογής είναι ο έλεγχος X^2 ο οποίος προτάθηκε από τον Pearson το 1900. Αργότερα, ο Kolmogorov (1933, 1941) επινόησε έναν εναλλακτικό έλεγχο για τον ίδιο σκοπό και ο Smirnov (1939, 1948) τον επέκτεινε για την περίπτωση του ελέγχου της υπόθεσης ότι δύο ανεξάρτητα δείγματα μπορούν να υποτεθούν ότι προέρχονται από την ίδια κατανομή.

1.5.3 Ιδιότητες Στατιστικών ελέγχων

Για να είναι ένας έλεγχος κατάλληλος, ή τουλάχιστον καταλληλότερος από τους υπόλοιπους , θα πρέπει να ικανοποιεί κάποια κριτήρια σε ένα βαθμό. Τέτοια κριτήρια είναι τα εξής :

1. Ο έλεγχος θα πρέπει να είναι αμερόληπτος.

2. Ο έλεγχος θα πρέπει να είναι συνεπής.

3. Ο έλεγχος θα πρέπει να είναι πιο αποτελεσματικός κατά κάποιο τρόπο από ότι οι υπόλοιποι έλεγχοι.

Από αυτές τις ιδιότητες η αποτελεσματικότητα, η οποία σχετίζεται με την ισχύ, είναι η πιο σημαντική και η πιο ευρέως χρησιμοποιούμενη. Μερικές φορές είμαστε ικανοποιημένοι αν ισχύουν ένα ή δύο από τα τρία κριτήρια. Μόνο σπάνια ισχύουν και οι τρεις προϋποθέσεις .

Σε αυτή την εργασία, θα ασχοληθούμε με τους ελέγχους υποθέσεων της κανονικής κατανομής. Θα αναφέρουμε και θα αναλύσουμε διάφορους ελέγχους για την κανονική κατανομή και θα δοκιμάσουμε κάποιους από αυτούς σε ένα σετ δεδομένων, για να ελεγχθεί η υπόθεση αν τα συγκεκριμένα δεδομένα που θα χρησιμοποιήσουμε ακολουθούν τελικά την κανονική κατανομή.

ΚΕΦΑΛΑΙΟ 2

ΕΛΕΓΧΟΙ ΥΠΟΘΕΣΕΩΝ ΓΙΑ ΤΗΝ ΚΑΝΟΝΙΚΗ ΚΑΤΑΝΟΜΗ - ΙΣΤΟΡΙΚΑ ΣΤΟΙΧΕΙΑ

Στην Στατιστική, ο Έλεγχος Υποθέσεων για την Κανονική Κατανομή ονομάζεται έλεγχος Κανονικότητας [Test of Normality] και είναι ίσως ο σπουδαιότερος στατιστικός έλεγχος υπόθεσης, και χρησιμοποιείται για να εξεταστεί αν η κατανομή μιας μεταβλητής είναι συμβατή με την Κανονική Κατανομή.
Παρακάτω θα αναφέρουμε τις κύριες μεθόδους που έχουν αναπτυχθεί για τον έλεγχο κανονικότητας με ιστορικά στοιχεία. Μπορούμε να διαχωρίσουμε τους ελέγχους υποθέσεων σε τρείς κύριες κατηγορίες :

2.1 Έλεγχοι παλινδρόμησης και έλεγχοι συσχέτισης (regression and correlation tests)

Οι έλεγχοι παλινδρόμησης και συσχέτισης είναι στατιστικές μέθοδοι που χρησιμοποιούνται για την εξέταση της σχέσης μεταξύ δύο ή περισσότερων μεταβλητών.
Η ανάλυση συσχέτισης (correlation test) χρησιμοποιείται για να προσδιοριστεί εάν υπάρχει σχέση μεταξύ δύο συνεχών μεταβλητών και ο βαθμός στον οποίο συσχετίζονται. Ο συντελεστής συσχέτισης είναι ένα μέτρο της ισχύος και της κατεύθυνσης (αρνητική ή θετική) της σχέσης μεταξύ των δύο μεταβλητών.
Η ανάλυση παλινδρόμησης είναι μια πιο ισχυρή στατιστική μέθοδος που μπορεί να χρησιμοποιηθεί για να μοντελοποιήσει τη σχέση μεταξύ δύο ή περισσότερων μεταβλητών και να κάνει προβλέψεις σχετικά με την τιμή μιας μεταβλητής με βάση την τιμή μιας ή περισσότερων άλλων μεταβλητών.

Samuel Sanford Shapiro

Ο Samuel Sanford Shapiro (1911-2001) ήταν Αμερικανός στατιστικός και μαθηματικός που συνέβαλε σημαντικά στους τομείς της στατιστικής και της επιχειρησιακής έρευνας. Γεννήθηκε στη Νέα Υόρκη και πήρε το διδακτορικό του στη στατιστική από το Πανεπιστήμιο της Βόρειας Καρολίνας στο Chapel Hill.

Martin Wilk

Ο Martin Wilk (1922-2013) ήταν ένας Αμερικανός στατιστικός που συνέβαλε σημαντικά στους τομείς της στατιστικής και της ανάλυσης δεδομένων. Γεννήθηκε στο Μπρούκλιν και πήρε το διδακτορικό του στη στατιστική από το Πανεπιστήμιο της Νέας Υόρκης.

2.1.1 Shapiro – Wilk Test

Ο έλεγχος Shapiro-Wilk δημοσιεύθηκε από τους Samuel Shapiro και Martin Wilk το 1965. Ο έλεγχος αναπτύχθηκε ως βελτίωση σε σχέση με προηγούμενους ελέγχους κανονικότητας, όπως τον έλεγχο Kolmogorov-Smirnov και τον έλεγχο X^2, οι οποίοι βρέθηκαν να έχουν περιορισμούς, ιδιαίτερα για μικρά έως μέτρια μεγέθη δειγμάτων. Οι Shapiro και Wilk πρότειναν έναν έλεγχο που θα ήταν πιο ευαίσθητος στις αποκλίσεις από την κανονικότητα στις ουρές της κατανομής και θα είχε καλύτερη ισχύ για την ανίχνευση μη κανονικότητας. Ο έλεγχος Shapiro-Wilk έχει χρησιμοποιηθεί ευρέως σε πολλούς τομείς, όπως η μηχανική, τα οικονομικά, η ψυχολογία και η βιολογία. Χρησιμοποιείται επίσης σε πολλά στατιστικά πακέτα λογισμικού, όπως R, SAS και SPSS.

R. S. Francia

Ο R. S. Francia (1934-2016), ή Ronald S. Francia, ήταν Αμερικανός στατιστικός και μαθηματικός που συνέβαλε σημαντικά στον τομέα της στατιστικής, ιδιαίτερα στον τομέα της μη παραμετρικής στατιστικής. Γεννήθηκε το 1934 στην Νέα Υόρκη και πήρε το πτυχίο του στα μαθηματικά από το City College της Νέας Υόρκης το 1956 και το μεταπτυχιακό του και το διδακτορικό του στη στατιστική από το Πανεπιστήμιο της Βόρειας Καρολίνας στο Chapel Hill το 1961 και το 1963.

2.1.2 Shapiro-Francia Test

Ο έλεγχος Shapiro-Francia είναι ένας στατιστικός έλεγχος που χρησιμοποιείται για να προσδιοριστεί εάν ένα σύνολο δεδομένων ακολουθεί μια κανονική κατανομή. Πήρε το όνομά του από τους δημιουργούς του, Samuel Shapiro και Ronald Francia.

Ο έλεγχος προτάθηκε για πρώτη φορά από τους Shapiro και Francia σε μια εργασία που δημοσιεύτηκε στο Journal of the American Statistical Association το 1972. Οι συγγραφείς προσπάθησαν να δημιουργήσουν ένα πιο αποτελεσματικό έλεγχο για την κανονικότητα από τον έλεγχο Shapiro-Wilk, ο οποίος είχε δημοσιευτεί λίγα χρόνια νωρίτερα. Ο έλεγχος Shapiro-Francia είναι μια τροποποίηση του ελέγχου Shapiro-Wilk, η οποία περιλαμβάνει τον γραφικό έλεγχο της κανονικότητας μέσω του p-p plot (γράφημα πιθανότητας-πιθανότητας) για ένα σύνολο δεδομένων ενώ ο έλεγχος Shapiro-Francia, χρησιμοποιεί μια διαφορετική ελεγχοσυνάρτηση που είναι συνάρτηση των διατεταγμένων τιμών του δείγματος και των αντίστοιχων αναμενόμενων τιμών τους κάτω από την μηδενική υπόθεση της κανονικότητας.

2.2 Έλεγχοι με βάση την Εμπειρική Συνάρτηση Κατανομής (Empirical Distribution Function (EDF))

Η εμπειρική συνάρτηση κατανομής (EDF) είναι μια μη παραμετρική μέθοδος για την εκτίμηση της αθροιστικής συνάρτησης κατανομής (CDF) ενός συνόλου δεδομένων. Η EDF παρέχει έναν τρόπο οπτικής σύγκρισης της κατανομής των δεδομένων με μια υποθετική κατανομή, όπως μια κανονική κατανομή. Η EDF μπορεί να είναι ένα χρήσιμο εργαλείο για την αξιολόγηση της κανονικότητας ενός συνόλου δεδομένων, αλλά είναι σημαντικό να σημειωθεί ότι είναι μια οπτική μέθοδος και μπορεί να είναι υποκειμενική.

Andrey Nikolaevich Kolmogorov

Ο Andrey Nikolaevich Kolmogorov (1903-1987) ήταν ένας Ρώσος μαθηματικός που συνέβαλε σημαντικά σε πολλούς τομείς των μαθηματικών, συμπεριλαμβανομένης της θεωρίας πιθανοτήτων, της τοπολογίας και της ανάλυσης. Το 1933, δημοσίευσε έναν στατιστικό έλεγχο που έγινε γνωστός ως έλεγχος Kolmogorov-Smirnov (KS test). Ο έλεγχος σχεδιάστηκε αρχικά για να προσδιορίσει εάν ένα δεδομένο δείγμα δεδομένων προέρχεται από μια συγκεκριμένη συνεχή κατανομή πιθανότητας, αλλά αργότερα επεκτάθηκε και σε διακριτές κατανομές.

Nikolai Vasilyevich Smirnov

Ο Nikolai Smirnov (1900-1966) ήταν Ρώσος μαθηματικός που εργάστηκε στη θεωρία των πιθανοτήτων και στη στατιστική. Είχε σημαντική συμβολή στη θεωρία της μη παραμετρικής στατιστικής και είναι γνωστός για την εργασία του στον έλεγχο Kolmogorov-Smirnov. Το 1948, δημοσίευσε μια εργασία που παρείχε μια πιο αποτελεσματική μέθοδο για τον υπολογισμό της στατιστικής ελεγχοσυνάρτησης Kolmogorov-Smirnov.

2.2.1 Kolmogorov–Smirnov Test

Ο έλεγχος Kolmogorov-Smirnov (KS) είναι ένας στατιστικός έλεγχος που χρησιμοποιείται για να προσδιορίσει εάν ένα δεδομένο δείγμα δεδομένων προέρχεται από μια συγκεκριμένη κατανομή πιθανοτήτων. Δημοσιεύτηκε για πρώτη φορά από τον Andrey Kolmogorov το 1933 και αργότερα βελτιώθηκε από τον Nikolai Smirnov το 1948. Ο έλεγχος KS έχει γίνει έκτοτε ένας ευρέως χρησιμοποιούμενος στατιστικός έλεγχος σε πολλούς τομείς, όπως η μηχανική, η φυσική, τα οικονομικά και η ψυχολογία. Είναι ιδιαίτερα χρήσιμος όταν η υποκείμενη κατανομή των δεδομένων είναι άγνωστη ή είναι δύσκολο να προσδιοριστεί. Ο έλεγχος μπορεί να χρησιμοποιηθεί για τη σύγκριση ενός δείγματος με μια καθορισμένη κατανομή ή για τη σύγκριση δύο δειγμάτων μεταξύ τους.

Hubert Whitman Lilliefors

Ο Hubert Whitman Lilliefors (1928-2008) ήταν Αμερικανός στατιστικός, γνωστός για την εισαγωγή του ελέγχου Lilliefors.

Ο Lilliefors έλαβε πτυχίο στα μαθηματικά από το Πανεπιστήμιο George Washington το 1952 και το διδακτορικό του στο Πανεπιστήμιο George Washington το 1964 υπό την επίβλεψη του Solomon Kullback. Στο ίδιο πανεπιστήμιο διετέλεσε καθηγητής στατιστικής για 39 χρόνια. Στη δεκαετία του 1970 συμμετείχε στη διαδικασία τυχαιοποίησης για το στρατό των ΗΠΑ.

2.2.2 Lilliefors Test

Στην στατιστική, ο έλεγχος Lilliefors είναι ένας έλεγχος κανονικότητας που βασίζεται στον έλεγχο Kolmogorov–Smirnov. Η διαφορά τους είναι ότι αντί να χρησιμοποιεί τη στατιστική ελεγχοσυνάρτηση Kolmogorov-Smirnov, χρησιμοποιεί μια τροποποιημένη στατιστική ελεγχοσυνάρτηση που είναι πιο κατάλληλη για μικρότερα μεγέθη δειγμάτων. Ο έλεγχος είναι ιδιαίτερα χρήσιμος όταν το μέγεθος του δείγματος είναι μικρό έως μέτριο (λιγότερο από 50), κάτι που είναι κοινό σε πολλούς τομείς, συμπεριλαμβανομένων των κοινωνικών επιστημών και της βιοστατιστικής.

Harald Cramér

Ο Harald Cramér (1893-1985) ήταν Σουηδός μαθηματικός και στατιστικός που συνέβαλε σημαντικά στη θεωρία των πιθανοτήτων, τη στατιστική και την αναλογιστική επιστήμη. Εισήγαγε τον έλεγχο Cramer von Mises σε μια εργασία που δημοσιεύτηκε το 1928, όπου παρουσίασε έναν έλεγχο καλής προσαρμογής βασισμένο στην εμπειρική συνάρτηση κατανομής.

Richard von Mises

Ο Richard von Mises (1883-1953) ήταν Αυστριακός μαθηματικός και στατιστικός που είναι γνωστός για το έργο του στη θεωρία πιθανοτήτων, τη στατιστική και την αεροδυναμική. Το 1931, ανέπτυξε ανεξάρτητα από τον Cramer τον ίδιο έλεγχο, ο οποίος έγινε γνωστός ως έλεγχος Cramer von Mises.

2.2.3 Cramér–von Mises Test

Ο έλεγχος Cramer-von Mises είναι ένας στατιστικός έλεγχος που χρησιμοποιείται για να προσδιοριστεί εάν ένα σύνολο παρατηρήσεων προέρχεται από μια συγκεκριμένη κατανομή. Ο έλεγχος αυτός είναι μια τροποποίηση του προηγούμενου ελέγχου του von Mises, ο οποίος χρησιμοποιήθηκε για να ελέγξει την καλή προσαρμογή των κυκλικών κατανομών. Ο Cramér επέκτεινε τον έλεγχο για να συμπεριλάβει μη κυκλικές κατανομές, όπως οι κανονικές και ομοιόμορφες κατανομές. Ο έλεγχος Cramer-von Mises χρησιμοποιείται στην στατιστική και είναι ιδιαίτερα χρήσιμος όταν η υποκείμενη κατανομή είναι άγνωστη ή όταν τα δεδομένα δεν κατανέμονται κανονικά. Ο έλεγχος έχει επεκταθεί για να χειρίζεται διάφορους τύπους δεδομένων, όπως λογοκριμένα δεδομένα, περικομμένα δεδομένα και ομαδοποιημένα δεδομένα.

Theodore Wilbur Anderson

Ο Theodore Wilbur Anderson (1918 – 2016) ήταν Αμερικανός μαθηματικός και στατιστικός που ειδικεύτηκε στην ανάλυση πολλών μεταβλητών. Γεννήθηκε στη Μινεάπολη της Μινεσότα. Ήταν στη σχολή του Πανεπιστημίου Κολούμπια από το 1946 μέχρι που μετακόμισε στο

Στάνφορντ για το Πανεπιστήμιο του Στάνφορντ το 1967, και έγινε Ομότιμος Καθηγητής το 1988.

Donald Allan Darling

Ο Donald Allan Darling (1915 – 2014) ήταν Αμερικανός στατιστικός, γνωστός για τον έλεγχο Anderson–Darling.

Γεννήθηκε το 1915 στο Λος Άντζελες. Το 1934 ο Darling ξεκίνησε τις προπτυχιακές του σπουδές στο Πανεπιστήμιο της Καλιφόρνια στο Λος Άντζελες, όπου απέκτησε το πτυχίο του στα μαθηματικά το 1939. Το 1940 έγινε μετεωρολόγος στην Pan American Airways και από το 1942 έως το 1946, κατά τη διάρκεια του Β' Παγκοσμίου Πολέμου, ηγήθηκε το τμήμα στατιστικής του Έργου Έρευνας Καιρού της Πολεμικής Αεροπορίας.

2.2.4 Anderson-Darling Test

Ο έλεγχος Anderson-Darling είναι ένα στατιστικός έλεγχος που χρησιμοποιείται για να προσδιορίσει εάν ένα σύνολο παρατηρήσεων προέρχεται από μια συγκεκριμένη κατανομή. Ο έλεγχος αναπτύχθηκε από τον Theodore Anderson και τον Donald Darling, δύο Αμερικανούς στατιστικούς, το 1952.

Ο έλεγχος Anderson-Darling είναι μια τροποποίηση του τεστ Kolmogorov-Smirnov. Ο Anderson και ο Darling επέκτειναν τον έλεγχο Kolmogorov-Smirnov εισάγοντας μια συνάρτηση στάθμισης που έδωσε μεγαλύτερη βαρύτητα στις ουρές της κατανομής. Αυτή η τροποποίηση έκανε το τεστ πιο ευαίσθητο σε αποκλίσεις από την υποθετική κατανομή στις ουρές, κάτι που είναι χρήσιμο σε πολλές εφαρμογές όπου οι ακραίες τιμές είναι σημαντικές. Τέλος ο έλεγχος αυτός χρησιμοποιείται στη στατιστική συμπερασματολογία και είναι ιδιαίτερα χρήσιμος όταν η υποκείμενη κατανομή είναι άγνωστη ή όταν τα δεδομένα δεν κατανέμονται κανονικά. Ο έλεγχος έχει επεκταθεί για να χειρίζεται διάφορους τύπους δεδομένων, όπως λογοκριμένα δεδομένα, περικομμένα δεδομένα και ομαδοποιημένα δεδομένα.

2.3 Έλεγχοι βάσει Ροπών (Moment Tests)

Οι έλεγχοι ροπών είναι μια κατηγορία στατιστικών ελέγχων που χρησιμοποιούνται για να προσδιοριστεί εάν ένα δείγμα δεδομένων είναι συνεπές με μια συγκεκριμένη κατανομή πιθανοτήτων. Αυτοί οι έλεγχοι βασίζονται στις ροπές της κατανομής, που αποτελούν μαθηματικές εκφράσεις του σχήματος και των ιδιοτήτων του. Οι έλεγχοι ροπών είναι χρήσιμοι για τον έλεγχο της προσαρμογής μιας συγκεκριμένης κατανομής σε ένα δείγμα δεδομένων, αλλά έχουν περιορισμούς.

Karl Pearson

O Karl Pearson (1857-1936) ήταν Άγγλος μαθηματικός που είχε πολλές σημαντικές συνεισφορές στην ανάπτυξη της σύγχρονης στατιστικής. Θεωρείται ευρέως ένας από τους ιδρυτές της σύγχρονης στατιστικής, μαζί με τον Ronald Fisher και τον Jerzy Neyman .Ο Pearson γεννήθηκε στο Λονδίνο και σπούδασε στο King's College του Cambridge. Εργάστηκε ως καθηγητής πανεπιστημίου και ερευνητής για μεγάλο μέρος της καριέρας του, μεταξύ άλλων στο University College του Λονδίνου και στο Πανεπιστήμιο του Cambridge. Το έργο του Pearson στη στατιστική επικεντρώθηκε στην ανάπτυξη της στατιστικής θεωρίας και μεθοδολογίας, συμπεριλαμβανομένης της ανάπτυξης της ανάλυσης συσχέτισης και παλινδρόμησης. Τέλος ο πρώτος έλεγχος Λοξότητας (Skewness) και Κύρτωσης (Kurtosis) αναπτύχθηκε από αυτόν.

2.3.1 Skewness Test

Η Λοξότητα (Skewness) είναι ένα μέτρο της ασυμμετρίας μιας κατανομής πιθανοτήτων. Ο έλεγχος Λοξότητας είναι ένα στατιστικός έλεγχος που χρησιμοποιείται για να προσδιορίσει εάν η λοξότητα ενός συνόλου παρατηρήσεων είναι σημαντικά διαφορετική από το μηδέν. Τα τελευταία χρόνια έχουν αναπτυχθεί έλεγχοι ασυμμετρίας που δεν απαιτούν υποθέσεις σχετικά με την υποκείμενη κατανομή των δεδομένων. Αυτοί οι έλεγχοι βασίζονται σε μεθόδους μετάθεσης και μπορούν να χρησιμοποιηθούν για τον έλεγχο της υπόθεσης ότι η λοξότητα ενός συνόλου παρατηρήσεων είναι διαφορετική από μια καθορισμένη τιμή, όπως το μηδέν. Τέλος, ο έλεγχος αυτός χρησιμοποιείται ευρέως στη στατιστική ανάλυση και είναι ιδιαίτερα χρήσιμος σε εφαρμογές όπου η συμμετρία των δεδομένων είναι σημαντική, όπως στη χρηματοδότηση και την οικονομία. Χρησιμοποιούνται επίσης σε τομείς όπως η γενετική και η βιολογία, όπου η κατανομή των δεδομένων μπορεί να είναι πολύ λοξή.

2.3.2 Kurtosis Test

Η Κύρτωση (Kurtosis) είναι ένα μέτρο της αιχμής ή της επιπεδότητας μιας κατανομής πιθανότητας. Ο έλεγχος κύρτωσης είναι ένας στατιστικός έλεγχος που χρησιμοποιείται για να προσδιορίσει εάν η κύρτωση ενός συνόλου παρατηρήσεων είναι σημαντικά διαφορετική από την κύρτωση μιας κανονικής κατανομής. Όπως και για τον έλεγχο Λοξότητας έτσι και για τον έλεγχο Κύρτωσης έχουν αναπτυχθεί έλεγχοι κύρτωσης που δεν απαιτούν υποθέσεις σχετικά με την υποκείμενη κατανομή των δεδομένων. Αυτοί οι έλεγχοι βασίζονται σε μεθόδους μετάθεσης και μπορούν να χρησιμοποιηθούν για τον έλεγχο της υπόθεσης ότι η κύρτωση ενός συνόλου παρατηρήσεων είναι διαφορετική από μια καθορισμένη τιμή, όπως η κύρτωση μιας κανονικής κατανομής. Τέλος, ο έλεγχος χρησιμοποιείται σε οικονομικούς τομείς και στην βιολογία όπου το σχήμα της κατανομής είναι πολύ σημαντικό.

Ralph D'Agostino

Ο Ralph B. D'Agostino Sr. είναι ένας Αμερικανός στατιστικός που έχει κάνει πολλές σημαντικές συνεισφορές στον τομέα της βιοστατιστικής. Γεννήθηκε το 1942 στο Somerville της Μασαχουσέτης και έλαβε το προπτυχιακό του πτυχίο από το Πανεπιστήμιο Tufts το 1963 και το διδακτορικό του στην στατιστική από το Πανεπιστήμιο του Χάρβαρντ το 1968. Η έρευνα του D'Agostino έχει επικεντρωθεί στην ανάπτυξη στατιστικών μεθόδων για ιατρική και επιδημιολογική έρευνα, συμπεριλαμβανομένης της ανάπτυξης μοντέλων πρόβλεψης κινδύνου για καρδιαγγειακές παθήσεις και άλλα προβλήματα υγείας. Είναι επίσης γνωστός για το έργο του στην ανάλυση διαχρονικών (longitunidal) δεδομένων και τη μέτρηση της διαγνωστικής ακρίβειας.

Michael A. Stephens

Ο Michael A. Stephens (1938-2015) ήταν ένας Αμερικανός στατιστικός που συνέβαλε σημαντικά στον τομέα της στατιστικής συμπερασματολογίας, ιδιαίτερα στους τομείς της μη παραμετρικής στατιστικής και των ισχυρότατων ελέγχων (most powerful tests). Γεννήθηκε το 1938 και έλαβε το προπτυχιακό του πτυχίο από το Πανεπιστήμιο της Καλιφόρνια στο Μπέρκλεϋ και το διδακτορικό του στην στατιστική από το Πανεπιστήμιο του Στάνφορντ το 1965. Ο Stephens ήταν γνωστός για την εργασία του για την ανάπτυξη ισχυρότατων στατιστικών μεθόδων, οι οποίες έχουν σχεδιαστεί για να είναι ανθεκτικές (robust) σε ακραίες τιμές και άλλες αποκλίσεις από την κανονικότητα. Συνέβαλε επίσης σημαντικά στην ανάπτυξη μη παραμετρικών μεθόδων, οι οποίες δεν απαιτούν υποθέσεις σχετικά με την υποκείμενη κατανομή των δεδομένων.

Ronald B. D'Agostino

Ο Ronald B. D'Agostino είναι ένας Αμερικανός στατιστικός που έχει κάνει σημαντικές συνεισφορές στους τομείς της βιοστατιστικής και της επιδημιολογίας. Γεννήθηκε το 1956 και έλαβε το προπτυχιακό του από το Κολλέγιο του Holy Cross και το διδακτορικό του στη βιοστατιστική από το Πανεπιστήμιο της Βοστώνης το 1984. Η έρευνα του D'Agostino επικεντρώθηκε στην ανάπτυξη στατιστικών μεθόδων για την ιατρική έρευνα, συμπεριλαμβανομένης της ανάλυσης διαχρονικών δεδομένων και της ανάπτυξης μοντέλων πρόβλεψης κινδύνου για καρδιαγγειακές παθήσεις και άλλα προβλήματα υγείας. Έχει επίσης συνεισφέρει σημαντικά στη μέτρηση της διαγνωστικής ακρίβειας και στην αξιολόγηση ιατρικών εξετάσεων και προγραμμάτων προληπτικού ελέγχου.

2.3.3 D'Agostino–Pearson Omnibus Test

Ο Omnibus έλεγχος D'Agostino-Pearson είναι ένας στατιστικός έλεγχος που χρησιμοποιείται για να προσδιοριστεί εάν ένα σύνολο παρατηρήσεων προέρχεται από μια κανονική κατανομή. Ο έλεγχος εισήχθη για πρώτη φορά από τους Ralph D'Agostino και Michael A. Stephens το 1971, και αργότερα βελτιώθηκε από τους D'Agostino και Ronald B. D'Agostino το 1986. Ο έλεγχος βασίζεται στη λοξότητα και την κύρτωση των δεδομένων του δείγματος. Ο πρώτος έλεγχος βασίζεται στη διαφορά μεταξύ της λοξότητας του δείγματος και της αναμενόμενης λοξότητας μιας κανονικής κατανομής και ο δεύτερος στατιστικός έλεγχος βασίζεται στη διαφορά μεταξύ της κύρτωσης του δείγματος και της αναμενόμενης κύρτωσης μιας κανονικής κατανομής. Ο Omnibus έλεγχος D'Agostino-Pearson χρησιμοποιείται για την αξιολόγηση της κανονικότητας στη στατιστική ανάλυση και έχει εφαρμοστεί σε πολλά πακέτα στατιστικών λογισμικού. Ωστόσο, έχει επικριθεί ότι έχει χαμηλή ισχύ στην ανίχνευση αποκλίσεων από την κανονικότητα, ιδιαίτερα για μικρά μεγέθη δειγμάτων.

Carlos Jarque

Ο Carlos Jarque είναι ένας Μεξικανός οικονομολόγος που είναι γνωστός για τη συνεισφορά του στον τομέα της μακροοικονομίας και των διεθνών οικονομικών. Γεννήθηκε το 1949 στην Πόλη του Μεξικού και έλαβε το πτυχίο του στα οικονομικά από το Εθνικό Αυτόνομο Πανεπιστήμιο του Μεξικού (UNAM) το 1971. Συνέχισε με το μεταπτυχιακό του και το Ph.D. στα οικονομικά από το Πανεπιστήμιο της Καλιφόρνια στο Μπέρκλεϋ το 1973 και το 1977, αντίστοιχα. Ο Jarque είναι γνωστός για την εξειδίκευσή του στα διεθνή χρηματοοικονομικά και μακροοικονομικά, και η δουλειά του συνεχίζει να επηρεάζει τις πολιτικές συζητήσεις και την ακαδημαϊκή έρευνα σε αυτούς τους τομείς.

Anil K. Bera

Ο Anil K. Bera είναι ένας ινδοαμερικανός οικονομολόγος και στατιστικός που είναι γνωστός για τη συμβολή του στην ανάλυση χρονοσειρών, στην μη παραμετρική στατιστική και στην οικονομική οικονομετρία. Γεννήθηκε στην Ινδία το 1950 και έλαβε το πτυχίο του στα μαθηματικά από το Πανεπιστήμιο της Καλκούτας το 1971. Συνέχισε με το μεταπτυχιακό του και το Ph.D. στα οικονομικά από το Πανεπιστήμιο της Πενσυλβάνια το 1975 και το 1977, αντίστοιχα. Οι απόψεις του Bera είναι σεβαστές για την εξειδίκευσή στην ανάλυση χρονοσειρών και στην χρηματοοικονομική οικονομετρία, και το έργο του συνεχίζει να επηρεάζει τις συζητήσεις πολιτικής και την ακαδημαϊκή έρευνα σε αυτούς τους τομείς.

2.3.4 Jarque-Bera Test

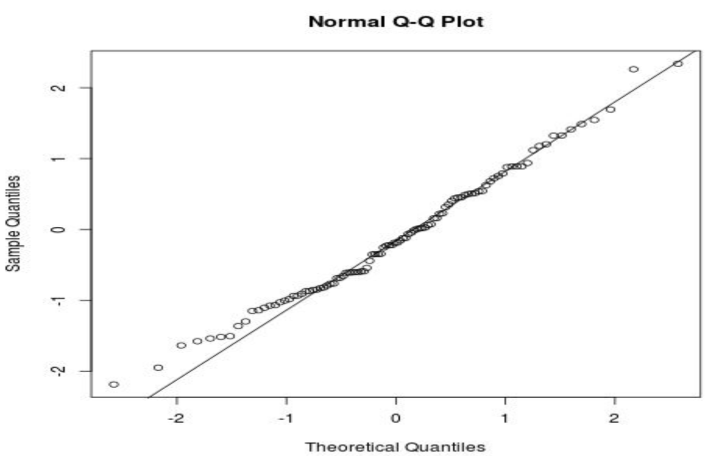
Ο έλεγχος Jarque-Bera είναι ένας στατιστικός έλεγχος που χρησιμοποιείται για τον έλεγχο της κανονικότητας σε ένα σύνολο δεδομένων. Αναπτύχθηκε από τους Carlos Jarque και Anil K. Bera το 1980. Ο έλεγχος Jarque-Bera βασίζεται στη λοξότητα και την κύρτωση των δεδομένων. Η λοξότητα μετρά την ασυμμετρία της κατανομής, ενώ η κύρτωση μετρά τον βαθμό αιχμής ή επιπεδότητας της κατανομής. Ο έλεγχος Jarque-Bera έχει γίνει ένα ευρέως χρησιμοποιούμενο εργαλείο για τον έλεγχο της κανονικότητας σε διάφορους τομείς, συμπεριλαμβανομένων των οικονομικών, της οικονομίας και της μηχανικής. Συχνά χρησιμοποιείται σε συνδυασμό με άλλους ελέγχους κανονικότητας, όπως ο έλεγχος Shapiro-Wilk και ο έλεγχος Kolmogorov-Smirnov, για να παρέχει μια πιο ισχυρή αξιολόγηση της κανονικότητας.

ΚΕΦΑΛΑΙΟ 3

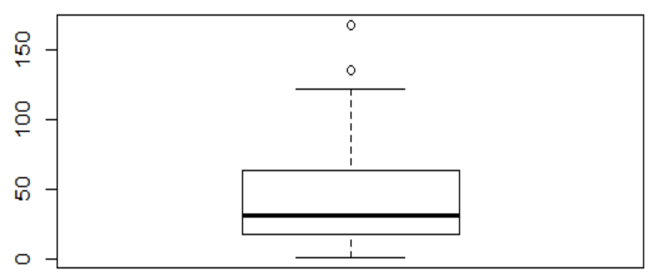
ΕΛΕΓΧΟΙ ΥΠΟΘΕΣΕΩΝ ΓΙΑ ΤΗΝ ΚΑΝΟΝΙΚΗ ΚΑΤΑΝΟΜΗ

Οι περισσότερες στατιστικές διαδικασίες και έλεγχοι (t-test, ANOVA, απλή ή πολλαπλή παλινδρόμηση (regression analysis), διαγράμματα ποιοτικού ελέγχου, κ.λ.π.) βασίζονται στην υπόθεση της κανονικότητας, συχνά μαζί με την ομοσκεδαστικότητα και την ανεξαρτησία των σφαλμάτων. Αν και η one-way ANOVA θεωρείται ένας ισχυρός στατιστικός έλεγχος έναντι της κανονικότητας, αν τα μεγέθη του δείγματος είναι μικρά και η κανονικότητα παραβιάζεται, τα αποτελέσματα μπορεί να είναι εσφαλμένα ή παραπλανητικά. Σε περίπτωση που δεν ικανοποιείται η υπόθεση της κανονικότητας, θα πρέπει να εφαρμοστεί ένας κατάλληλος μετασχηματισμός δεδομένων, ο οποίος συχνά εκ νέου καταλήγει σε αντιφατικά συμπεράσματα. Έτσι, για ερευνητές και επαγγελματίες σε διάφορους τομείς, προκύπτει η ανάγκη να ελεγχθεί η υπόθεση της κανονικότητας προκειμένου να διασφαλιστεί η καταλληλότητα της επιλεγμένης στατιστικής τεχνικής καθώς και η ακρίβεια των αποτελεσμάτων που εξήχθησαν. Αυτός ο έλεγχος μπορεί να επιτευχθεί με διαγνωστικά διαγράμματα όπως διαγράμματα Q-Q (Q-Q plots, εικόνα 1), κυτιοδιαγράμματα (box-plots, εικόνα 2) και ιστογράμματα (histograms, εικόνα 3). Αυτά τα διαγνωστικά διαγράμματα είναι χρήσιμα, όπως επίσης υπάρχουν και τρόποι επιλογής μεταξύ τους όμως χρειάζεται γνώση και εμπειρία για την ερμηνεία των αποτελεσμάτων τους. Τις περισσότερες φορές, οι στατιστικοί έλεγχοι χρησιμοποιούνται για να επιβεβαιώσουν συμπεράσματα που βασίζονται σε οπτική ερμηνεία γραφικών μεθόδων. Θα πρέπει να σημειωθεί ωστόσο, ότι οι περιορισμοί και τα μειονεκτήματα δεν είναι ασυνήθιστα. Πράγματι, για παράδειγμα, αν και οι στατιστικοί έλεγχοι έχουν το πλεονέκτημα της αντικειμενικής κρίσης της κανονικότητας, τις περισσότερες φορές έχουν το μειονέκτημα ότι συχνά εμφανίζουν σφάλματα σε πολύ μικρά μεγέθη ή σε μεγάλα μεγέθη δείγματος. Αποδεικνύεται ότι τουλάχιστον σε ορισμένες από αυτές τις περιπτώσεις μπορεί κανείς να εντοπίσει γραφικά εργαλεία που έχουν το πλεονέκτημα ότι επιτρέπουν σχετικά καλύτερη κρίση για την αξιολόγηση της κανονικότητας από τους στατιστικούς ελέγχους, κάτι που με τη σειρά του συνεπάγεται ότι πρέπει να λαμβάνονται αποφάσεις χωρίς να παραμελείτε η επάρκεια και η αποτελεσματικότητα των γραφημάτων. Ταυτόχρονα, οι στατιστικοί έλεγχοι δεν πρέπει να υπερεκτιμούνται και η στατιστική συλλογιστική πρέπει να ασκείται σε όλες τις περιπτώσεις, όχι μόνο για την επιλογή του καταλληλότερου ελέγχου, αλλά και για την επιλογή μεταξύ μιας ελεγχοσυνάρτησης ή ενός γραφικού εργαλείου για την αξιολόγηση της κανονικότητας. Υπάρχουν πολλοί έλεγχοι κανονικότητας, με διαφορετικές υποθέσεις και εφαρμογές, διαθέσιμες στη βιβλιογραφία και οι ιδιότητές τους έχουν εξεταστεί από πολλούς ερευνητές. Ως εκ τούτου, ο ερευνητής αντιμετωπίζει ένα θεμελιώδες πρόβλημα, δηλαδή πώς να επιλέξει τον καταλληλότερο έλεγχο για το σύνολο των δεδομένων του.

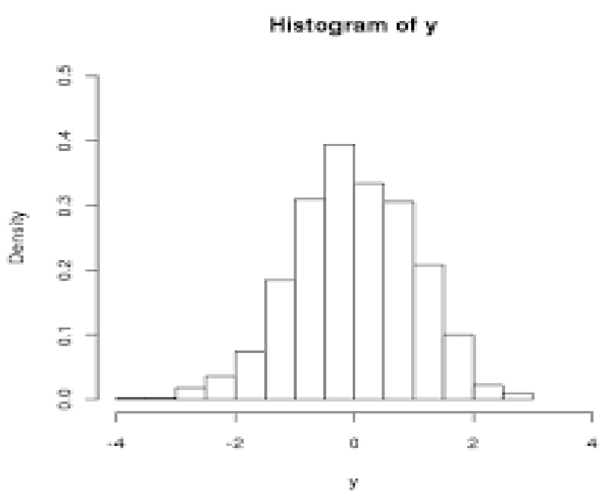
Σε αυτό το κεφάλαιο θα συζητήσουμε τον κάθε στατιστικό έλεγχο αναλυτικά καθώς και την χρησιμότητα του κάθε ελέγχου υποθέσεων ανάλογα με τα δεδομένα που μας δίνονται και τις τρεις βασικές κατηγορίες που χωρίσαμε τους ελέγχους υποθέσεων της κανονικής κατανομής στο προηγούμενο κεφάλαιο δηλαδή 1) στους ελέγχους παλινδρόμησης και ελέγχους συσχέτισης 2) ελέγχους με βάση την εμπειρική συνάρτηση κατανομής και 3) ελέγχους βάσει ροπών.



Εικόνα 1: QQ - Διάγραμμα



Εικόνα 2: Κυτιοδιάγραμμα



Εικόνα 3: Ιστόγραμμα

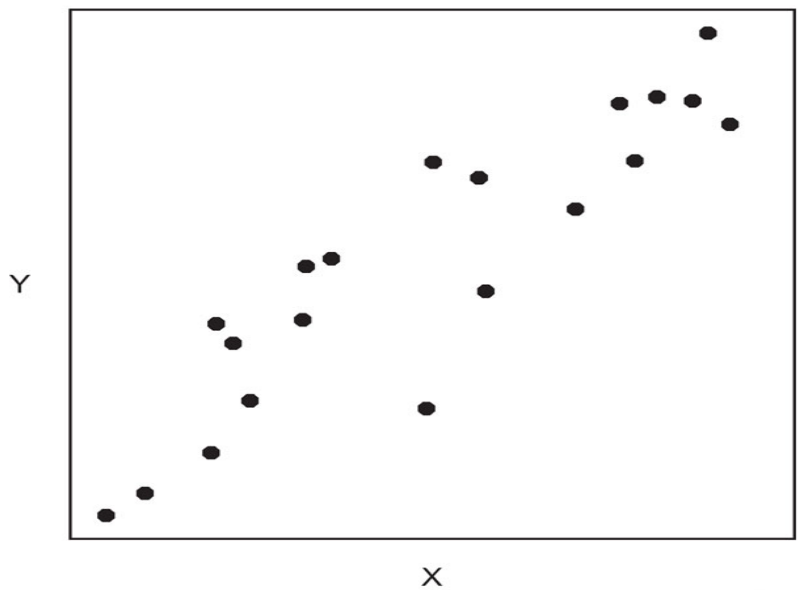
3.1 Έλεγχοι παλινδρόμησης και έλεγχοι συσχέτισης (regression and correlation tests)

Οι πιο συχνά χρησιμοποιούμενες τεχνικές για τη διερεύνηση της σχέσης μεταξύ δύο ποσοτικών μεταβλητών είναι η συσχέτιση και η γραμμική παλινδρόμηση. Η συσχέτιση ποσοτικοποιεί την ισχύ της γραμμικής σχέσης μεταξύ ενός ζεύγους μεταβλητών, ενώ η παλινδρόμηση εκφράζει τη σχέση με τη μορφή εξίσωσης. Κατά τη διερεύνηση μιας σχέσης μεταξύ δύο μεταβλητών, το πρώτο βήμα είναι να εμφανιστούν οι τιμές δεδομένων γραφικά σε ένα διάγραμμα διασποράς. Σε ένα διάγραμμα διασποράς, όσο πιο κοντά βρίσκονται τα σημεία σε μια ευθεία γραμμή, τόσο ισχυρότερη είναι η γραμμική σχέση μεταξύ δύο μεταβλητών. Για να ποσοτικοποιήσουμε την ισχύ της σχέσης, μπορούμε να υπολογίσουμε τον συντελεστή συσχέτισης. Στην αλγεβρική σημειογραφία, αν έχουμε δύο μεταβλητές x και y, και τα δεδομένα έχουν τη μορφή n ζευγών (δηλαδή $[x_1, y_1], [x_2, y_2], \ldots., [x_n, y_n]$) , τότε ο συντελεστής συσχέτισης δίνεται από τον ακόλουθο ορισμό:

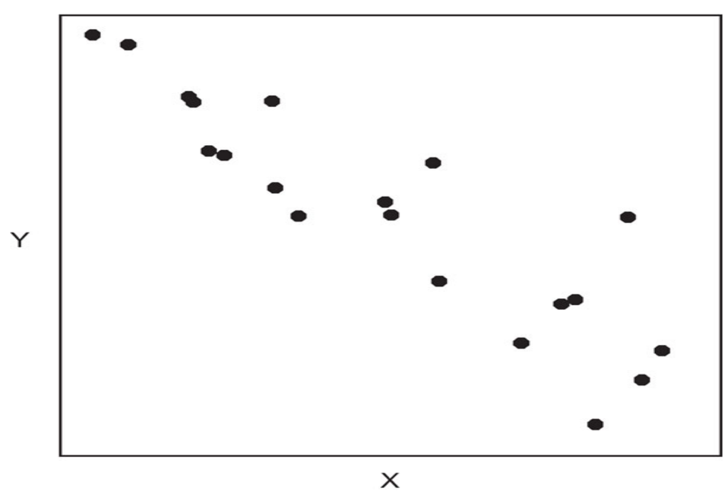
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 * \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

όπου $\bar{x}$ ο μέσος όρος των τιμών x και $\bar{y}$ ο μέσος όρος των τιμών y.

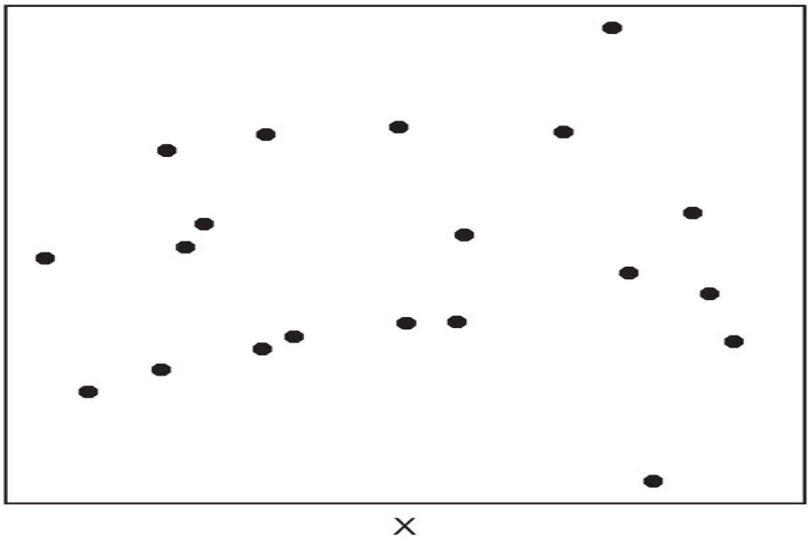
Αυτός είναι ο δειγματικός συντελεστής συσχέτισης (ή συντελεστής συσχέτισης Pearson). Η τιμή του r βρίσκεται πάντα μεταξύ -1 και $+1$. Μια τιμή του συντελεστή συσχέτισης κοντά στο $+1$ υποδηλώνει μια ισχυρή θετική γραμμική σχέση. Μια τιμή κοντά στο -1 υποδηλώνει μια ισχυρή αρνητική γραμμική σχέση. Μια τιμή κοντά στο 0 υποδεικνύει ότι δεν υπάρχει γραμμική σχέση, ωστόσο, θα μπορούσε να υπάρχει μια μη γραμμική σχέση μεταξύ των μεταβλητών.



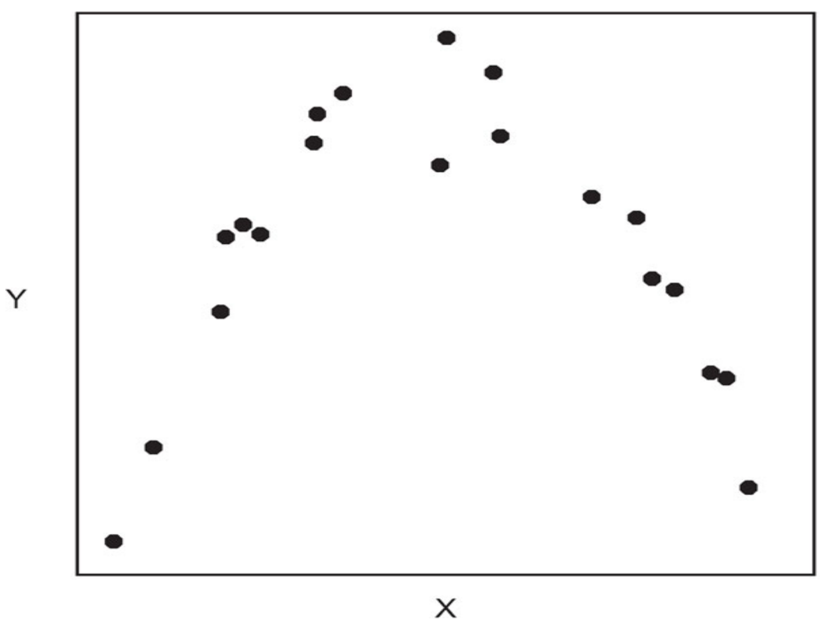
Εικόνα 4: Συντελεστής συσχέτισης r= +0.9 θετικός κοντά στο 1



Εικόνα 5 : Συντελεστής συσχέτισης r= -0.9 αρνητικός κοντά στο -1



Εικόνα 6 : Συντελεστής συσχέτισης r= +0.04 θετικός κοντά στο 0



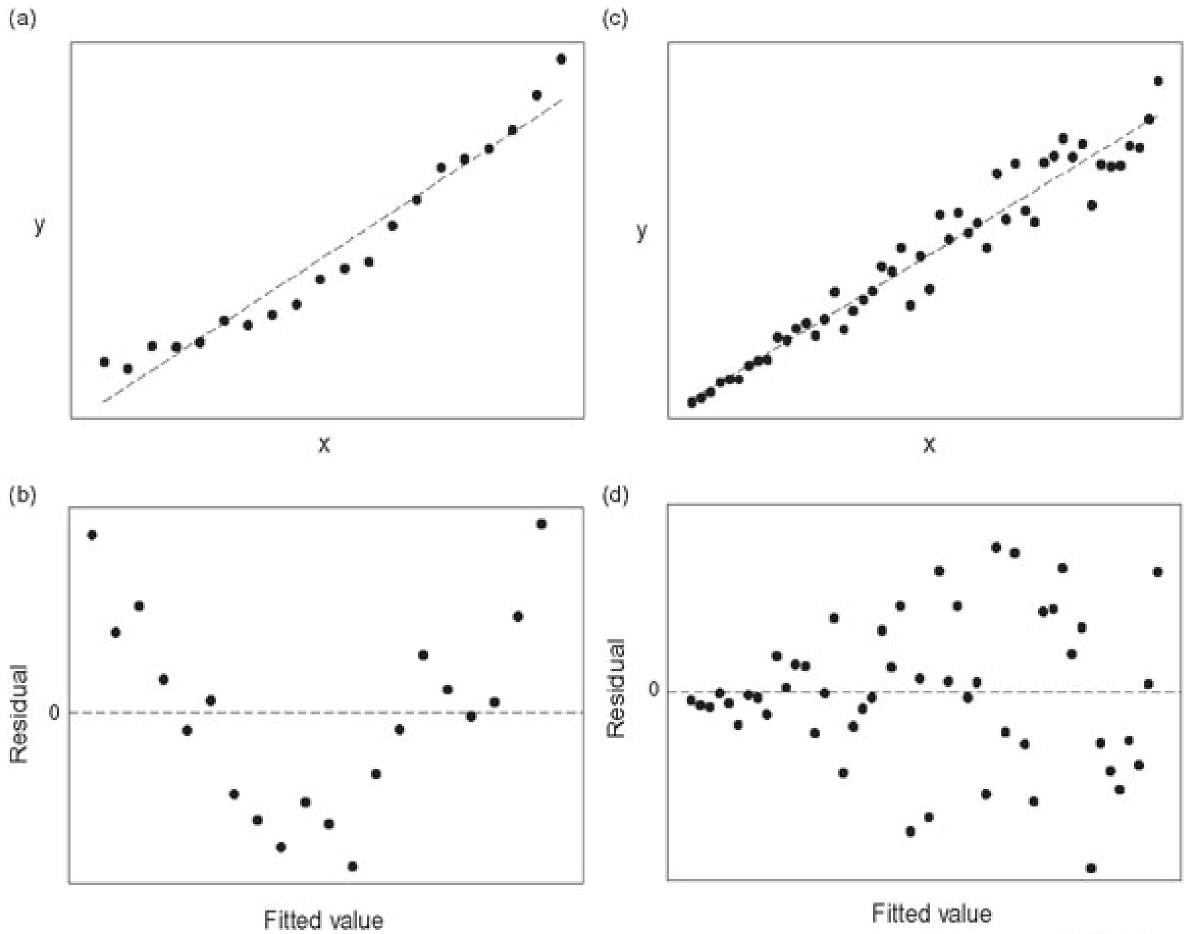
Εικόνα 7 : Συντελεστής συσχέτισης r= -0.03 αρνητικός κοντά στο 0

Η γραμμή παλινδρόμησης προκύπτει με τη μέθοδο των ελαχίστων τετραγώνων. Οποιαδήποτε ευθεία $y = a + bx$ που χαράσσουμε μέσα από τα σημεία δίνει μια προβλεπόμενη ή προσαρμοσμένη τιμή yγια κάθε τιμή του x στο σύνολο δεδομένων. Για μια συγκεκριμένη τιμή τουx η κατακόρυφη διαφορά μεταξύ της παρατηρούμενης και της προσαρμοσμένης τιμής του y είναι γνωστή ως απόκλιση ή υπόλοιπο (residual). Η μέθοδος των ελαχίστων τετραγώνων βρίσκει τις τιμές των a και b που ελαχιστοποιούν το άθροισμα των τετραγώνων όλων των αποκλίσεων. Αυτό δίνει τους ακόλουθους τύπους για τον υπολογισμό των a και b όπου α το σημείο που η ευθεία τέμνει τον άξονα των x και b η κλίση της ευθείας παλινδρόμησης :

$a = \bar{y} - b\bar{x}$ $\qquad$ $b = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

Συνήθως, αυτές οι τιμές υπολογίζονται χρησιμοποιώντας ένα στατιστικό πακέτο ή τις στατιστικές συναρτήσεις σε μια αριθμομηχανή.

Η χρήση της συσχέτισης και της παλινδρόμησης εξαρτάται από ορισμένες υποκείμενες υποθέσεις. Οι παρατηρήσεις υποθέτουμε ότι είναι ανεξάρτητες. Για τον συντελεστή συσχετίσεως και οι δύο μεταβλητές πρέπει να είναι τυχαίες μεταβλητές, αλλά για την παλινδρόμηση μόνο η μεταβλητή απόκρισης y πρέπει να είναι τυχαία. Κατά τη διεξαγωγή στατιστικών ελέγχων ή τον υπολογισμό των διαστημάτων εμπιστοσύνης για τις παραμέτρους παλινδρόμησης, η μεταβλητή απόκρισης πρέπει να ακολουθεί μία Κανονική κατανομή και η μεταβλητότητα του y πρέπει να είναι η ίδια για κάθε τιμή της μεταβλητής πρόβλεψης. Οι ίδιες υποθέσεις απαιτούνται για τον έλεγχο της μηδενικής υπόθεσης ότι η συσχέτιση είναι 0, αλλά για να ερμηνευτούν τα διαστήματα εμπιστοσύνης για τον συντελεστή συσχέτισης και οι δύο μεταβλητές πρέπει να είναι Κανονικά κατανεμημένες. Τόσο η συσχέτιση όσο και η παλινδρόμηση υποθέτουν ότι η σχέση μεταξύ των δύο μεταβλητών είναι γραμμική. Ένα διάγραμμα διασποράς των δεδομένων παρέχει έναν αρχικό έλεγχο των υποθέσεων για παλινδρόμηση. Οι παραδοχές μπορούν να αξιολογηθούν λεπτομερέστερα εξετάζοντας τα γραφήματα των υπολοίπων (residuals). Συνήθως, τα υπόλοιπα σχεδιάζονται με βάση τις προσαρμοσμένες τιμές. Εάν η σχέση είναι γραμμική και η μεταβλητότητα σταθερή, τότε τα υπολείμματα θα πρέπει να είναι ομοιόμορφα διασκορπισμένα γύρω από το 0 κατά μήκος του εύρους των προσαρμοσμένων τιμών. Επιπλέον, μπορεί να παραχθεί ένα κανονικό γράφημα υπολοίπων. Αυτή είναι μια γραφική παράσταση των υπολοίπων σε σχέση με τις τιμές που θα αναμενόταν να λάβουν εάν προέρχονταν από μια τυπική Κανονική κατανομή. Εάν τα υπόλοιπα είναι κανονικά κατανεμημένα, τότε αυτή η γραφική παράσταση θα δείχνει μια ευθεία γραμμή. (Μια τυπική κανονική κατανομή είναι μια κανονική κατανομή με μέση τιμή $\mu = 0$ και τυπική απόκλιση $\sigma = 1$.) Τα κανονικά διαγράμματα είναι συνήθως διαθέσιμα σε στατιστικά πακέτα.



Από τα πιο πάνω γραφήματα βλέπουμε ότι:
a) Το διάγραμμα διασποράς του y έναντι του x υποδηλώνει ότι η σχέση είναι μη γραμμική
b) Γραφικός έλεγχος υπολοίπων έναντι προσαρμοσμένων τιμών για την περίπτωση (a). Η καμπυλότητα της σχέσης φαίνεται πιο καθαρά
c) Το διάγραμμα διασποράς του y έναντι του x υποδηλώνει ότι η μεταβλητότητα στο y αυξάνεται με το x(καλύτερη γραμμική σχέση)
d) Γράφημα υπολοίπων σε σχέση με τις προσαρμοσμένες τιμές για την περίπτωση (c). Η αυξανόμενη μεταβλητότητα στο y με το xφαίνεται πιο καθαρά.

3.1.1 Shapiro – Wilk Test

Ο έλεγχος Shapiro-Wilk ελέγχει τη μηδενική υπόθεση ότι ένα τυχαίο δείγμα $x_1, x_2 \dots \dots, x_n$ προέρχεται από έναν κανονικά κατανεμημένο πληθυσμό. Η στατιστική ελεγχοσυνάρτηση του ελέγχου είναι:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

όπου

1)$x_{(i)}$ είναι η (i) διατεταγμένη παρατήρηση στο δείγμα (δεν πρέπει να συγχέεται με το x_i).
2) $\bar{x} = (x_1 + x_2 + x_3 + \dots . x_n)/n$ ο μέσος του δείγματος
3) Οι συντελεστές a_iδίνονται ως εξής:

$(\alpha_1, \dots, \alpha_n) = \frac{m^T V^{-1}}{C}$ όπου το c είναι η διανυσματική νόρμα
$C = \|V^{-1}m\| = (m^T V^{-1} V^{-1} m)^{(1/2)}$
$m = (m_1, \dots, m_n)^T$

και το διάνυσμα m αποτελείται από τις αναμενόμενες τιμές των διατεταγμένων τυχαίων μεταβλητών από ένα τυχαίο δείγμα ανεξάρτητων και πανομοιότυπα κατανεμημένων τυχαίων μεταβλητών που προέρχονται από την τυποποιημένη κανονική κατανομή. Τέλος, V είναι ο πίνακας συνδιακύμανσης του εν λόγω διατεταγμένου δείγματος. Δεν υπάρχει ονομασία για την κατανομή του W. Οι κρίσιμες τιμές της ελεγχοσυνάρτησης W υπολογίζονται μέσω προσομοιώσεων Monte Carlo.
Η μηδενική υπόθεση αυτού του στατιστικού ελέγχου είναι ότι ο πληθυσμός είναι κανονικά κατανεμημένος. Έτσι, εάν η $p - value$ είναι μικρότερη από το επίπεδο σημαντικότητας α, τότε η μηδενική υπόθεση απορρίπτεται και υπάρχουν ενδείξεις ότι τα δεδομένα που εξετάστηκαν δεν κατανέμονται κανονικά. Από την άλλη πλευρά, εάν η $p - value$ είναι μεγαλύτερη από το επίπεδο σημαντικότητας α, τότε η μηδενική υπόθεση (ότι τα δεδομένα προέρχονται από κανονικά κατανεμημένο πληθυσμό) δεν μπορεί να απορριφθεί. Π.χ., για επίπεδο σημαντικότητας $\alpha = 0.05$ για ένα σύνολο δεδομένων με $p - value$ μικρότερη από 0.05 απορρίπτεται η μηδενική υπόθεση ότι τα δεδομένα προέρχονται από κανονικά κατανεμημένο πληθυσμό. Κατά συνέπεια, αντίθετα για ένα σύνολο δεδομένων με $p - value$ μεγαλύτερη από την τιμή $\alpha = 0.05$ αποτυγχάνουμε να απορρίψουμε τη μηδενική υπόθεση ότι τα δεδομένα προέρχονται από κανονικά κατανεμημένο πληθυσμό.
Όπως και οι περισσότεροι έλεγχοι στατιστικής σημαντικότητας, εάν το μέγεθος του δείγματος είναι αρκετά μεγάλο, ο έλεγχος αυτός μπορεί να ανιχνεύσει ακόμη και ασήμαντες αποκλίσεις από τη μηδενική υπόθεση (δηλαδή, αν και μπορεί να υπάρχει κάποια στατιστικά σημαντική επίδραση, μπορεί να είναι πολύ μικρή για να έχει πρακτική σημασία) - επομένως, συνήθως συνιστάται πρόσθετη διερεύνηση του μεγέθους της επίδρασης, π.χ. ένα διάγραμμα Q-Q σε αυτή την περίπτωση.

3.1.2. Shapiro-Francia test

Ο έλεγχος Shapiro-Francia είναι ένας στατιστικός έλεγχος για την κανονικότητα ενός πληθυσμού, με βάση δειγματικά δεδομένα.
Έστω ένα τυχαίο δείγμα $x_1, x_{2,} \dots, x_n$ και έστω x_i η i-οστή διατεταγμένη τιμή του δείγματος. Για παράδειγμα, αν το δείγμα αποτελείται από τις τιμές $5.6, -1.2, 7.5, 3.4, x_2 = 3.4$ επειδή αυτή είναι η δεύτερη μικρότερη τιμή. Έστω $m_{(i;n)}$ η μέση τιμή της i-οστής διατεταγμένης τυχαίας μεταβλητής από ένα τυχαίο δείγμα με n ανεξάρτητες παρατηρήσεις από μια κανονική κατανομή. Για παράδειγμα, $m_{(2;4)} \approx -0.297$, που σημαίνει ότι η δεύτερη μικρότερη τιμή σε

ένα δείγμα τεσσάρων παρατηρήσεων από μια κανονκή κατανομή είναι συνήθως περίπου 0.297 τυπικές αποκλίσεις κάτω από τη μέση τιμή.
Ο συντελεστής συσχέτισης Pearson μεταξύ των x και των m είναι:

$$W' = \frac{cov(x,m)}{\sigma_x \sigma_m} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (m_i - \bar{m})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 * \sum_{i=1}^{n}(m_i - \bar{m})^2}}$$

Υπό τη μηδενική υπόθεση ότι τα δεδομένα προέρχονται από κανονική κατανομή, η συσχέτιση αυτή θα είναι ισχυρή, οπότε οι τιμές W' θα συγκεντρώνονται λίγο κάτω από το 1, με την κορυφή να γίνεται στενότερη και πιο κοντά στο 1 καθώς το n αυξάνεται. Εάν τα δεδομένα αποκλίνουν έντονα από μια κανονική κατανομή, το W' θα είναι μικρότερο. Αυτή η ελεγχοσυνάρτηση είναι μια τυποποίηση της παλαιότερης πρακτικής του σχηματισμού ενός διαγράμματος Q-Q για τη σύγκριση δύο κατανομών, με τα x να παίζουν τον ρόλο των εμπειρικών ποσοστημόριων της κατανομής του δείγματος και το m να παίζει το ρόλο των αντίστοιχων θεωρητικών ποσοστημορίων μιας κανονικής κατανομής.
Σε σύγκριση με τον στατιστικό έλεγχο Shapiro-Wilk W, ο στατιστικός έλεγχος Shapiro-Francia W' είναι ευκολότερος στον υπολογισμό, επειδή δεν απαιτεί να σχηματίσουμε και να αντιστρέψουμε τον πίνακα των συνδιακυμάνσεων μεταξύ των διατεταγμένων τυχαίων μεταβλητών του δείγματος.

3.2 Έλεγχοι με βάση την Εμπειρική Συνάρτηση Κατανομής (Empirical Distribution Function (EDF))

Οι εμπειρικές συναρτήσεις κατανομής και οι στατιστικές συναρτήσεις με βάση διατεταγμένες τυχαίες μεταβλητές έχουν σημαντικές εφαρμογές σε μη παραμετρικά μοντέλα παλινδρόμησης, σε μεθόδους δειγματοληψίας όπως το jackknife και το bootstrap, σε ακολουθιακούς ελέγχους (sequential testing), καθώς και στην ανάλυση επιβίωσης και αξιοπιστίας. Ειδικότερα, χρησιμεύουν ως βάση για τις γνωστές στατιστικές συναρτήσεις καλής προσαρμογής Kolmogorov-Smirnov και Cramér-von Mises. Δεδομένων των δειγματικών παρατηρήσεων $Y_1, Y_2, \dots, Y_n$ που ακολουθούν κάποια συνάρτηση κατανομής F, η εμπειρική συνάρτηση κατανομής που υπολογίζεται σε έναν δεδομένο πραγματικό αριθμό y είναι

$$F_n(y) = n^{-1} \sum_{i=1}^{n} I(Y_i \leq y) \quad \forall\, y \in R$$

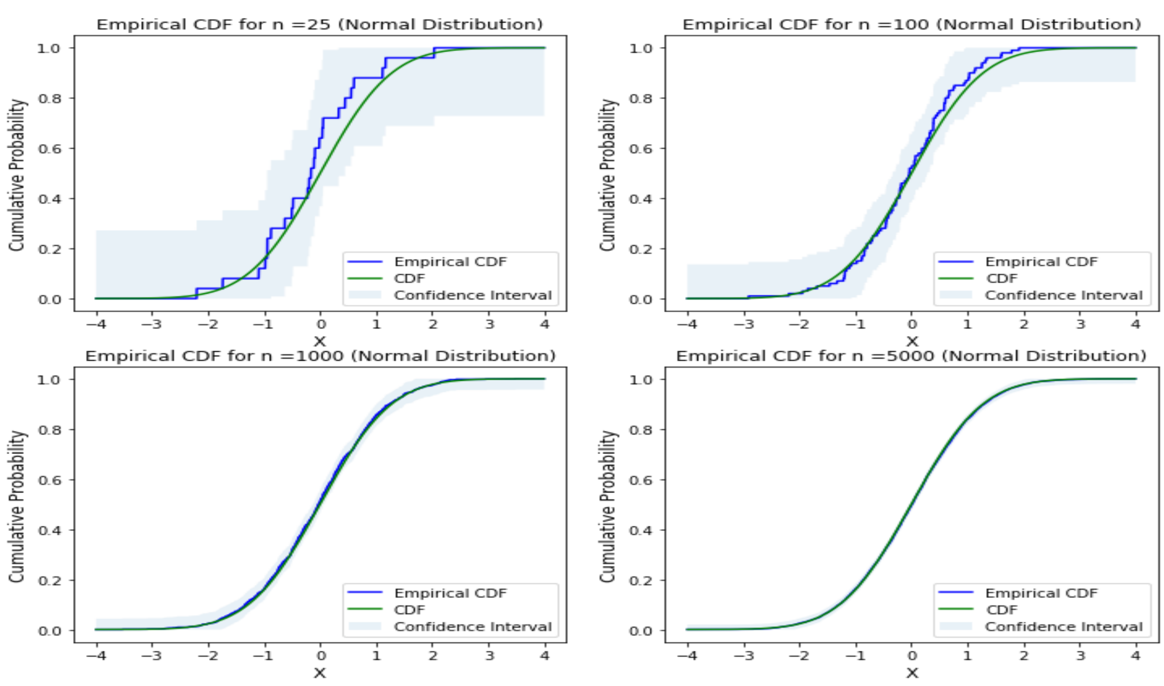
όπου $I(Y_i \leq y)$ είναι μια δείκτρια συνάρτηση που παίρνει την τιμή 1 εάν $Y_i \leq y$ και 0 διαφορετικά. Συνδέεται στενά με τις διατεταγμένες τυχαίες μεταβλητές, $Y_{(n:1)} \leq Y_{(n:2)} \leq \dots \leq Y_{(n:n)}$ όπου $Y_{(n:1)}$ είναι η μικρότερη μεταξύ των $Y_1, \dots, Y_n$ $ενώ$ $Y_{(n:2)}$ είναι η δεύτερη μικρότερη κ.ο.κ. Για δεδομένο δείγμα, η F_n είναι μια συνάρτηση κατανομής όταν θεωρείται ως συνάρτηση του y. Για κάθε δεδομένο y, όταν εξετάζεται ως συνάρτηση των $Y_{1,} \dots, Y_n$, η $F_n(y)$ είναι μια τυχαία μεταβλητή- σε αυτό το πλαίσιο, δεδομένου ότι οι $I(Y_i \leq y)$, $i = 1, \dots, n$, είναι ανεξάρτητες και πανομοιότυπα κατανεμημένες τυχαίες μεταβλητές από τη διωνυμική κατανομή με πιθανότητα επιτυχίας $F(y)$, μπορούμε να εφαρμόσουμε το κεντρικό οριακό θεώρημα για να συμπεράνουμε ότι για κάθε y, η κατανομή της $F_n(y)$ μπορεί να προσεγγιστεί από μια κανονική κατανομή $N(F(y), F(y)\,[1 - F(y)]/n)$, υπό την προϋπόθεση ότι το n είναι αρκετά μεγάλο. Για την επέκταση αυτών των αποτελεσμάτων στη συνάρτηση F_n που υπολογίζεται σε όλες τις πραγματικές τιμές y, απαιτούνται πιο περίπλοκες μέθοδοι, όπως πρότειναν μεταξύ άλλων οι Jurečkova και Sen (1996).

Έστω $(Y_{1,} \dots, Y_n)$ ανεξάρτητες, πανομοιότυπα κατανεμημένες πραγματικές τυχαίες μεταβλητές από την κοινή αθροιστική συνάρτηση κατανομής $F(t)$. Τότε η εμπειρική συνάρτηση κατανομής ορίζεται ως εξής

$$\widehat{F_n}(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{(Y_i \leq t)} = \frac{\text{αριθμός στοιχείων στο δείγμα} \leq t}{n}$$

όπου 1_A είναι η δείκτρια του γεγονότος A. Για ένα σταθερό t η δείκτρια $1_{(Y_i \leq t)}$ είναι τυχαία μεταβλητή διωνυμική με παράμετρο $p = F(t)$ συνεπώς η $nF(t)$ είναι διωνυμική τυχαία μεταβλητή με μέση τιμή $nF(t)$ και διακύμανση $nF(t)(1 - F(t))$. Αυτό σημαίνει ότι είναι ένας αμερόληπτος εκτιμητής για την $F(t)$. Ωστόσο, σε ορισμένα εγχειρίδια, ο ορισμός δίνεται ως εξής

$$\widehat{F_n}(t) = \frac{1}{n+1} 1_{(Y_i \leq t)}$$



Διαγράμματα εμπειρικής CDF (cumulative distribution function), CDF και διαστήματος εμπιστοσύνης για διάφορα μεγέθη δείγματος της Κανονικής Κατανομής.

3.2.1 Kolmogorov–Smirnov Test

Με τον έλεγχο Kolmogorov-Smirnov ελέγχουμε την μηδενική υπόθεση ότι τα δεδομένα ακολουθούν μια συγκεκριμένη κατανομή, με εναλλακτική ότι δεν την ακολουθούν. Με βάση λοιπόν την $p - value$ του εν λόγω ελέγχου φτάνουμε σε τελικά συμπεράσματα σε σχέση με την καταλληλότητα ή μη του μοντέλου δηλαδή την καταλληλότητα της κατανομής που έχουμε επιλέξει. Ο εν λόγω έλεγχος βασίζεται στην εμπειρική συνάρτηση κατανομής. Έστω $x_1, x_2, \dots, x_n$ οι τιμές ενός τυχαίου δείγματος από πληθυσμό με άγνωστη συνάρτηση κατανομής F. Θέλουμε να ελέγξουμε τη μηδενική υπόθεση $H_0: F = F_0$, με εναλλακτική την υπόθεση $H_1: F \neq F_0$,όπου F_0 γνωστή συνάρτηση κατανομής. Θα υπολογίσουμε την ελεγχοσυνάρτηση:

$$D = \sup_{-\infty < x < \infty}(|F_n(x) - F_0(x)|)$$

όπου $F_n(x)$ είναι η εμπειρική συνάρτηση κατανομής που ορίζεται ως ακολούθως:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty \leq x_i)}x_i = \frac{αριθμός\ στοιχείων\ στο\ δείγμα \leq x}{n}$$

Αποδεικνύεται ότι για μεγάλο μέγεθος δείγματος δηλαδή για n>50, για την κατανομή της ελεγχοσυνάρτησης D κάτω από την μηδενική υπόθεση ισχύει ότι:

$$P[\sqrt{n}D \leq x \mid H_0] \rightarrow 1 - 2\sum_{j=1}^{\infty}(-1)^{(j-1)}e^{2j^2x^2}, x \geqslant 0.$$

Με βάση την παραπάνω κατανομή υπολογίζουμε προσεγγιστικά την *p-value* του αμφίπλευρου ελέγχου. Αν θέλουμε να χρησιμοποιήσουμε τον έλεγχο Kolmogorov-Smirnov για την κανονική κατανομή, δηλαδή να δούμε αν τα δεδομένα που διαλέξαμε προέρχονται από κανονικό πληθυσμό, θα θεωρήσουμε πως η F_0 είναι η συνάρτηση κατανομής της Κανονικής κατανομής.

3.2.2. Lilliefors Test

Ο έλεγχος Lilliefors είναι ένας έλεγχος κανονικότητας. Αποτελεί βελτίωση του ελέγχου Kolomogorov-Smirnov (K-S) - διορθώνοντας τον K-S για τις μικρές τιμές στις ουρές των κατανομών πιθανότητας - και ως εκ τούτου μερικές φορές αποκαλείται έλεγχος K-S D. Πολλά στατιστικά πακέτα (όπως το SPSS) συνδυάζουν τους δύο ελέγχους K-S με διόρθωση "Lilliefors". Σε αντίθεση με τον έλεγχο K-S, ο Lilliefors μπορεί να χρησιμοποιηθεί όταν δεν γνωρίζουμε τη μέση τιμή ή την τυπική απόκλιση του πληθυσμού. Ουσιαστικά, ο στατιστικός έλεγχος Lilliefors είναι ένας έλεγχος K-S που μας επιτρέπει να εκτιμήσουμε αυτές τις παραμέτρους από το δείγμα μας.

Έστω $x_1, \ldots, x_n$ οι τιμές ενός τυχαίου δείγματος και ορίζουμε την μηδενική υπόθεση (H_0) ότι τα δεδομένα μας προέρχονται από την κανονική κατανομή με εναλλακτική υποθεσή (H_1) ότι τα δεδομένα μας δεν ακολουθούν την κανονική κατανομή. Οι τύποι που εμπλέκονται στον έλεγχο Lilliefors είναι αρκετά χρονοβόροι (εν μέρει λόγω του γεγονότος ότι πρέπει να υπολογίσουμε τα z-scores για κάθε μέλος του δείγματος). Αρχικά υπολογίζουμε

$Z_i = \frac{x_i - \bar{x}}{S}$, $i = 1,2,\ldots,n$ όπου Z_i= τα ατομικά z-scores για κάθε μέλος του δείγματός μας.

x_i = μεμονωμένο μέλος / σημείο δεδομένων

$\bar{x}$ = μέση τιμή του δείγματος

S = τυπική απόκλιση του δείγματος

με δειγματική διασπορά $S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$.

Στην συνέχεια υπολογίζουμε την ελεγχοσυνάρτηση, η οποία είναι η μέγιστη απόσταση μεταξύ της εμπειρικής συνάρτησης κατανομής (EDF) των Z_i και της συνάρτησης κατανομής της τυποποιημένης κανονικής κατανομής. Ο τύπος που χρησιμοποιούμε είναι:

$$T_1 = \underset{x}{supremum}|F(x) - S(x)|$$

όπου

$F(x)$= η συνάρτηση κατανομής της τυποποιημένης κανονικής κατανομής.

$S(x)$ = η εμπειρική συνάρτηση κατανομής των τιμών Z_i.

Τέλος, βρίσκουμε την κρίσιμη τιμή για τον έλεγχο από τον πίνακα 3 στο παράρτημα και απορρίπτουμε την μηδενική υπόθεση εάν η ελεγχοσυνάρτηση T_1 είναι μεγαλύτερη από την κρίσιμη τιμή.

Οι τύποι που εμπλέκονται στον έλεγχο Lilliefors είναι αρκετά χρονοβόροι (λόγω του γεγονότος ότι πρέπει να υπολογίσουμε τα z-scores για κάθε μέλος του δείγματος).

Ο έλεγχος Lilliefors έχει χαμηλή ισχύ. Σύμφωνα με την R, είναι "γνωστό ότι αποδίδει χειρότερα" από τους ελέγχους Anderson-Darling ή Cramer-von Mises.

3.2.3 Cramér–von Mises Test

Στη στατιστική ο έλεγχος Cramér-von Mises είναι ένας έλεγχος που χρησιμοποιείται για την αξιολόγηση της καλής προσαρμογής μιας αθροιστικής συνάρτησης κατανομής F σε σύγκριση με μια δεδομένη εμπειρική συνάρτηση κατανομής F_n, ή για τη σύγκριση δύο εμπειρικών κατανομών. Χρησιμοποιείται επίσης ως μέρος άλλων αλγορίθμων, όπως η εκτίμηση ελάχιστης απόστασης. Ορίζεται ως:

$$\omega^2 = \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \mathrm{dF}(x)$$

Σε εφαρμογές ενός δείγματος, F είναι η θεωρητική κατανομή και F_n είναι η εμπειρικά παρατηρούμενη κατανομή. Εναλλακτικά, οι δύο κατανομές μπορεί να είναι και οι δύο εμπειρικά εκτιμώμενες- αυτή ονομάζεται περίπτωση δύο δειγμάτων.
Έστω $(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$ οι παρατηρούμενες τιμές, με αύξουσα σειρά. Υπολογίζουμε την ελεγχοσυνάρτηση

$$T = n\omega^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left(\frac{2i-1}{2n} - F(x_{(i)})\right)^2$$

Εάν η τιμή αυτή είναι μεγαλύτερη από την κρίσιμη τιμή που δίνεται στο Παράρτημα στον πίνακα 4, τότε η υπόθεση ότι τα δεδομένα προέρχονται από την κατανομή F μπορεί να απορριφθεί.

3.2.4 Anderson-Darling Test

Ο έλεγχος Anderson-Darling είναι ένας στατιστικός έλεγχος του κατά πόσον ένα συγκεκριμένο δείγμα δεδομένων προέρχεται από μια συγκεκριμένη κατανομή πιθανότητας. Στη βασική του μορφή, ο έλεγχος υποθέτει ότι δεν υπάρχουν παράμετροι προς εκτίμηση στην ελεγχόμενη κατανομή, οπότε ο έλεγχος και το σύνολο των κρίσιμων τιμών του δεν εξαρτώνται από την κατανομή. Ωστόσο, ο έλεγχος χρησιμοποιείται συχνότερα σε περιπτώσεις όπου ελέγχεται μια οικογένεια κατανομών, οπότε οι παράμετροι αυτής της οικογένειας πρέπει να εκτιμηθούν και αυτό πρέπει να ληφθεί υπόψη στην προσαρμογή είτε του στατιστικού ελέγχου είτε των κρίσιμων τιμών του. Όταν εφαρμόζεται στον έλεγχο του κατά πόσον μια κανονική κατανομή περιγράφει επαρκώς ένα σύνολο δεδομένων, είναι ένα από τα πιο ισχυρά στατιστικά εργαλεία για την ανίχνευση των περισσότερων αποκλίσεων από την κανονικότητα. Οι έλεγχοι K-S και Anderson-Darling είναι διαθέσιμοι για τον έλεγχο του κατά πόσον διάφορες συλλογές παρατηρήσεων μπορούν να μοντελοποιηθούν ως προερχόμενες από έναν ενιαίο πληθυσμό, όπου η συνάρτηση κατανομής δεν χρειάζεται να καθοριστεί.
Εκτός από τη χρήση του ως ελέγχου προσαρμογής για κατανομές, μπορεί να χρησιμοποιηθεί στην εκτίμηση παραμέτρων ως βάση για μια μορφή διαδικασίας εκτίμησης ελάχιστης απόστασης.
Έστω $x_1, x_2, \ldots, x_n$ οι τιμές ενός τυχαίου δείγματος σε αύξουσα σειρά δηλαδή $x_1 < x_2 < \cdots < x_n$ και ορίζουμε την μηδενική υπόθεση (H_0) ότι τα δεδομένα μας προέρχονται από την κανονική κατανομή με εναλλακτική υπόθεση (H_1) ότι τα δεδομένα μας δεν ακολουθούν την κανονική κατανομή. Μπορεί να αποδειχθεί ότι η ελεγχοσυνάρτηση Anderson-Darling για την περίπτωση ελέγχου για την κανονική κατανομή παίρνει τη μορφή:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\ln(F(x_i)) + \ln(1 - F(x_{(n+1-i)}))\right]$$

όπου

n: το μέγεθος του τυχαίου δείγματος
F: είναι η αθροιστική συνάρτηση της κανονικής κατανομής. Εάν οι τιμές x_i έχουν τυποποιηθεί, είναι η συνάρτηση κατανομής της τυποποιημένης κανονικής κατανομής. Οι κρίσιμες τιμές βρίσκονται από την θεωρητική κατανομή.

Μια τροποποιημένη ελεγχοσυνάρτηση για την περίπτωση που θα χρειαστεί εκτίμηση των παραμέτρων μέσης τιμής και διασποράς από το δείγμα, δίνεται από τον D' Agostino (1986):

$$A_c^2 = A^2(1 + \frac{0.75}{n} + \frac{2.25}{n^2})$$

Εάν το A_c^2 είναι μεγαλύτερο από $0{,}752$, τότε η μηδενική υπόθεση ότι τα δεδομένα ανταποκρίνονται σε κανονική κατανομή απορρίπτεται σε επίπεδο σημαντικότητας $\alpha = 5\%$.
Τα δεδομένα μπορούν να ελεγχθούν σε σχέση με άλλες θεωρητικές κατανομές με τη χρήση της κατάλληλης συνάρτησης αθροιστικής κατανομής. Κρίσιμες τιμές και εφαρμογές λογισμικού είναι διαθέσιμες για την κανονική, τη λογαριθμοκανονική, την εκθετική, την Weibull και τη λογιστική κατανομή.

3.3 Έλεγχοι βάσει Ροπών (Moment Tests)

Στη στατιστική, η μέθοδος των ροπών είναι μια μέθοδος εκτίμησης των παραμέτρων του πληθυσμού. Η ίδια αρχή χρησιμοποιείται για τις ανώτερες τάξεις ροπών, όπως η λοξότητα και η κύρτωση.
Ξεκινά με την έκφραση των πληθυσμιακών ροπών (δηλαδή των αναμενόμενων τιμών των δυνάμεων της υπό εξέταση τυχαίας μεταβλητής) ως συναρτήσεις των παραμέτρων ενδιαφέροντος. Στη συνέχεια, οι εκφράσεις αυτές εξισώνονται με τις δειγματικές ροπές. Ο αριθμός αυτών των εξισώσεων είναι ίδιος με τον αριθμό των παραμέτρων που πρέπει να εκτιμηθούν. Οι εξισώσεις αυτές επιλύονται στη συνέχεια για τις παραμέτρους ενδιαφέροντος. Οι λύσεις αποτελούν εκτιμήσεις των εν λόγω παραμέτρων.
Η μέθοδος των ροπών εισήχθη από τον Pafnuty Chebyshev το 1887 στην απόδειξη του κεντρικού οριακού θεωρήματος. Η ιδέα της αντιστοίχισης των εμπειρικών ροπών μιας κατανομής με τις ροπές του πληθυσμού χρονολογείται τουλάχιστον από τον Pearson.

Ας υποθέσουμε ότι το πρόβλημα είναι η εκτίμηση κ άγνωστων παραμέτρων $\theta_1, \theta_2, \dots, \theta_\kappa$ που χαρακτηρίζουν την κατανομή $g_w(w;\theta)$ της τυχαίας μεταβλητής W. Ας υποθέσουμε ότι

κ ροπές της πραγματικής κατανομής (οι "πληθυσμιακές ροπές") μπορούν να εκφραστούν ως συναρτήσεις του θ=$(\theta_1, \theta_2, \dots, \theta_\kappa)$:

$$\mu_1 = E[W] = g_1(\theta_1, \theta_2, \dots, \theta_\kappa)$$
$$\mu_2 = E[W^2] = g_2(\theta_1, \theta_2, \dots, \theta_\kappa)$$

.
.

.
$\mu_\kappa = E[W^\kappa] = g_\kappa(\theta_{1,}\theta_{2,}\ldots, \theta_\kappa)$

Ας υποθέσουμε ότι έχουμε ένα δείγμα μεγέθους *n*, $w_1, \ldots, w_n$. Για $j = 1, \ldots, k$ έχουμε τον εξής ορισμό:

$$\widehat{\mu_J} = \frac{1}{n}\sum_{i=1}^{n} w_i^j$$

είναι η j-οστή δειγματοληπτική ροπή, μια εκτίμηση του μ_j. Ο εκτιμητής της μεθόδου των ροπών για $\theta_1, \theta_2 \ldots, \theta_\kappa$ συμβολίζεται με $\widehat{(\theta_1)}, \ldots, \widehat{(\theta_\kappa)}$ ορίζεται ως η λύση (εάν υπάρχει) των εξισώσεων:

$\widehat{(\mu_1)} = g_1(\widehat{(\theta_1)}, \widehat{(\theta_2)}, \ldots, \widehat{(\theta_\kappa)})$
$\widehat{(\mu_2)} = g_2(\widehat{(\theta_1)}, \widehat{(\theta_2)}, \ldots, \widehat{(\theta_\kappa)})$
.
.
.
$\widehat{(\mu_k)} = g_k(\widehat{(\theta_1)}, \widehat{(\theta_2)}, \ldots, \widehat{(\theta_\kappa)})$

Η μέθοδος των ροπών είναι αρκετά απλή και δίνει συνεπείς εκτιμητές (υπό πολύ αδύναμες υποθέσεις), αν και οι εκτιμητές αυτοί είναι συχνά μεροληπτικοί. Αποτελεί εναλλακτική λύση στη μέθοδο της μέγιστης πιθανοφάνειας.

Ωστόσο, σε ορισμένες περιπτώσεις οι συναρτήσεις πιθανοφάνειας μπορεί να είναι δύσκολες να μεγιστοποιηθούν χωρίς υπολογιστές, ενώ οι εκτιμητές της μεθόδου των ροπών μπορούν να υπολογιστούν πολύ πιο γρήγορα και εύκολα. Λόγω της εύκολης υπολογισιμότητας, οι εκτιμητές της μεθόδου των ροπών μπορούν να χρησιμοποιηθούν ως η πρώτη προσέγγιση της μεγιστοποίησης των συναρτήσεων πιθανοφάνειας και στη συνέχεια μπορούν να βρεθούν διαδοχικές βελτιωμένες προσεγγίσεις με τη μέθοδο Newton-Raphson. Με αυτόν τον τρόπο η μέθοδος των ροπών μπορεί να βοηθήσει στην εύρεση εκτιμήσεων μέγιστης πιθανοφάνειας.
Σε ορισμένες περιπτώσεις, σπάνιες σε μεγάλα δείγματα αλλά λιγότερο σπάνιες με μικρά δείγματα, οι εκτιμήσεις που δίνει η μέθοδος των ροπών βρίσκονται εκτός του χώρου των παραμέτρων. Αυτό το πρόβλημα δεν προκύπτει ποτέ στη μέθοδο της μέγιστης πιθανοφάνειας. Επίσης, οι εκτιμήσεις με τη μέθοδο των ροπών δεν είναι απαραίτητα επαρκείς στατιστικές συναρτήσεις, δηλαδή μερικές φορές δεν λαμβάνουν υπόψη όλες τις σχετικές πληροφορίες του δείγματος.
Κατά την εκτίμηση άλλων δομικών παραμέτρων (π.χ. παραμέτρων μιας συνάρτησης χρησιμότητας, αντί παραμέτρων μιας γνωστής κατανομής πιθανότητας), οι κατάλληλες κατανομές πιθανότητας μπορεί να μην είναι γνωστές και οι εκτιμήσεις με βάση τη μέθοδο των ροπών μπορεί να προτιμηθούν από την εκτίμηση με βάση τη μέγιστη πιθανοφάνεια.
Ένα παράδειγμα εφαρμογής της μεθόδου των ροπών είναι η εκτίμηση πολυωνυμικών κατανομών πυκνότητας πιθανότητας. Στην περίπτωση αυτή, ένα προσεγγιστικό πολυώνυμο τάξης N ορίζεται σε ένα διάστημα [a,b]. Η μέθοδος των ροπών δίνει στη συνέχεια ένα σύστημα εξισώσεων, η λύση του οποίου περιλαμβάνει την αντιστροφή ενός πίνακα Hankel. Στη συνέχεια το παράδειγμα που θα δούμε είναι η απόδειξη του κεντρικού οριακού θεωρήματος με την μέθοδο των ροπών.

Απόδειξη του κεντρικού οριακού θεωρήματος

Έστω $X_1, X_2, \dots, X_n$ ανεξάρτητες τυχαίες μεταβλητές με μέση τιμή 0 και διακύμανση 1, έστω $S_n := \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i$. Μπορούμε να υπολογίσουμε τις ροπές του S_n ως εξής

$E[S_n^0] = 1, E[S_n^1] = 0, E[S_n^2] = 1, E[S_n^3] = 0, \dots.$

Το ανάπτυγμα δείχνει ότι

$$E(S_n^{(2k+1)}) = 0; E(S_n^{2k}) = \frac{n(n-1)\dots\dots(n-k+1)}{n^k}(2k-1)(2\kappa-3)(2\kappa-5)\dots(3)(1)$$

όπου ο αριθμητής είναι ο αριθμός των τρόπων επιλογής k διαφορετικών ζευγών σφαιρών επιλέγοντας από μία από $2k$ κάδους, καθένας από τους οποίους περιέχει σφαίρες αριθμημένες από το 1 έως το n. Στο όριο $n \longrightarrow \infty$, όλες οι ροπές συγκλίνουν σε εκείνες μιας τυπικής κανονικής κατανομής. Περισσότερη θεωρία στη συνέχεια δείχνει ότι αυτή η σύγκλιση ροπών συνεπάγεται σύγκλιση κατά κατανομή. Ουσιαστικά αυτό το επιχείρημα δημοσιεύθηκε από τον Chebyshev το 1887.

3.3.1 Skewness Test

Η λοξότητα περιγράφει την έλλειψη συμμετρίας σε μια κατανομή συχνοτήτων. Μια κατανομή είναι δεξιά (ή θετικά) λοξή αν η ουρά εκτείνεται προς τα δεξιά - προς τους υψηλότερους αριθμούς. Μια κατανομή είναι αριστερά (ή αρνητικά) λοξή αν η ουρά εκτείνεται προς τα αριστερά.
Στη στατιστική, η λοξότητα συνήθως μετράται και ορίζεται με τη χρήση του συντελεστή λοξότητας τον οποίο θα συμβολίσουμε $\widehat{g_1}$. Ο συντελεστής λοξότητας είναι η μέση, τυποποιημένη, κυβική απόκλιση από το δειγματικό μέσο.
Όπως είναι αναμενόμενο, επειδή ο συντελεστής λοξότητας χρησιμοποιεί την κυβική απόκλιση από τον δειγματικό μέσο, η λοξότητα μπορεί να είναι είτε θετική είτε αρνητική. Μια συμμετρική κατανομή έχει μηδενική λοξότητα - παραδόξως όμως, η μηδενική λοξότητα δεν αποδεικνύει ότι η κατανομή είναι συμμετρική!
Για τυχαία δείγματα από κάποια μεταβλητή X, ο συντελεστής λοξότητας $\widehat{g_1}$ μπορεί να εκτιμηθεί χρησιμοποιώντας αυτόν τον τύπο:

$$\widehat{g_1} = \sum_{i=1}^{n} {x_i}'^3/n = \frac{\sum_{i=1}^{n}(X_i-\bar{x})^3/n}{\sqrt{(\sum_{i=1}^{n}(X_i-\bar{x})^2/n)^3}}$$

όπου $x_i' = (X_i - \bar{x})/s_x$: η τυποποιημένη απόκλιση από το δειγματικό μέσο και

s_x: η δειγματική τυπική απόκλιση

Δυστυχώς, ο παραπάνω τύπος παρέχει μεροληπτικές εκτιμήσεις του $\widehat{g_1}$ όταν υπολογίζεται από μικρά δείγματα λοξών πληθυσμών. Ο παρακάτω τύπος χρησιμοποιείται για λιγότερο μεροληπτική εκτίμηση.
Για μεγαλύτερα δείγματα χρησιμοποιείται ο παρακάτω τύπος:

$$\widehat{g_1} = \frac{\sum_{i=1}^{n} {x_i'}^3 n}{(n-1)(n-2)}$$

όπου το x_i' το ίδιο με πριν με την διαφορά ότι εδώ $s_x^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$.

Για έναν κανονικό πληθυσμό και μεγάλα δείγματα (n > 150), το $\widehat{g_1}$ κατανέμεται κατά προσέγγιση κανονικά με μέσο όρο 0 και τυπικό σφάλμα $\sqrt{\frac{6}{n}}$.
Για πολύ μικρά δείγματα πληθυσμών με έντονη στρέβλωση, ακόμη και αυτός ο τύπος αναμένεται να υποεκτιμά την πραγματική του τιμή - με άλλα λόγια, $|E(\widehat{g_1})| < |\widehat{g_1}|$. Σε αυτή την περίπτωση, η προσομοίωση είναι ο μόνος τρόπος για να λάβουμε μια αμερόληπτη εκτίμηση - ή για να εκτιμήσουμε τη μεταβλητότητά του.

3.3.2 Kurtosis Test

Η κύρτωση περιγράφεται συχνά ως ο βαθμός στον οποίο η κορυφή μιας κατανομής πιθανοτήτων αποκλίνει από το σχήμα μιας κανονικής κατανομής (αν είναι πιο οξεία, η κατανομή είναι λεπτόκυρτη, αν είναι πιο επίπεδη, είναι πλατύκυρτη). Δεδομένου ότι οι "ακραίες τιμές" έχουν τη μεγαλύτερη επιρροή, ένας πιο χρήσιμος τρόπος για να θεωρηθεί η κύρτωση είναι από την άποψη του μήκους της ουράς (αν οι ουρές είναι μεγαλύτερες από τις αναμενόμενες είναι πλατύκυρτη, αν είναι μικρότερες είναι λεπτόκυρτη). Ο συντελεστής κύρτωσης ($\widehat{(g_2)}$) είναι ο μέσος όρος της τέταρτης δύναμης των τυποποιημένων αποκλίσεων από τον δειγματικό μέσο.
Για έναν κανονικό πληθυσμό, ο συντελεστής κύρτωσης αναμένεται να ισούται με 3. Μια τιμή μεγαλύτερη από 3 υποδηλώνει μια λεπτόκυρτη κατανομή- μια τιμή μικρότερη από 3 υποδηλώνει μια πλατύκυρτη κατανομή. Για την εκτίμηση μέσω δείγματος ($\widehat{g_2}$), το 3 αφαιρείται έτσι ώστε μια θετική τιμή να υποδηλώνει λεπτοκύρτωση και μια αρνητική τιμή να υποδηλώνει πλατυκύρτωση.
Για τυχαία δείγματα από κάποια μεταβλητή X, ο συντελεστής κύρτωσης ($\widehat{g_2}$) μπορεί να εκτιμηθεί χρησιμοποιώντας αυτόν τον τύπο:

$$\widehat{g_2} = \frac{\sum_{i=1}^{n} {x_i'}^4}{n} - 3 = \frac{\sum_{i=1}^{n}(X_i-\bar{x})^4}{(\sum_{i=1}^{n}(X_i-\bar{x})^2/n)^2} - 3$$

Ο τύπος αυτός παρέχει μεροληπτικές εκτιμήσεις όταν υπολογίζεται από μικρά δείγματα πληθυσμών με κύρτωση. Ο παρακάτω τύπος παρέχει μια λιγότερο μεροληπτική εκτίμηση του $\hat{g}_2$.

Ένας άλλος τρόπος για να υπολογίσουμε τον συντελεστή κύρτωσης είναι :

$$\widehat{g_2} = \left(\frac{\sum_{i=1}^{n} x_i{}'^4 n(n+1)}{(n-1)} - 3(n-1)^2\right)\frac{1}{(n-2)(n-3)}$$

Για ένα μεγάλο δείγμα (n > 150) κανονικού πληθυσμού, το $\widehat{g_2}$ έχει μέση τιμή 0 και τυπικό σφάλμα $\sqrt{\frac{24}{n}}$. Ωστόσο, η κατανομή του δεν είναι κατά προσέγγιση κανονική, εκτός εάν το μέγεθος του δείγματος υπερβαίνει τα 1000.

3.3.3 D'Agostino–Pearson Omnibus Test

Στη στατιστική, ο έλεγχος του D'Agostino, που πήρε το όνομά του από τον Ralph D'Agostino, είναι ένα μέτρο καλής προσαρμογής της απόκλισης από την κανονικότητα, δηλαδή ο έλεγχος αποσκοπεί στο να μετρήσει τη συμβατότητα δεδομένων με τη μηδενική υπόθεση ότι τα δεδομένα είναι μια υλοποίηση ανεξάρτητων, πανομοιότυπα κατανεμημένων τυχαίων μεταβλητών Gauss. Ο έλεγχος βασίζεται σε μετασχηματισμούς της κύρτωσης και της λοξότητας του δείγματος και έχει ισχύ μόνο έναντι των εναλλακτικών υποθέσεων ότι η κατανομή είναι λοξή και/ή έχει κυρτότητα. Η δειγματική λοξότητα $\widehat{g_1}$ και η κύρτωση $\widehat{g_2}$ ακολουθούν αμφότερες ασυμπτωτικά κανονική κατανομή. Ωστόσο, ο ρυθμός σύγκλισής τους στο όριο της κατανομής είναι απογοητευτικά αργός, ειδικά για την $\widehat{g_2}$. Για παράδειγμα, ακόμη και με $n = 5000$ παρατηρήσεις η δειγματική κύρτωση $\widehat{g_2}$ έχει τόσο τη λοξότητα όσο και την κύρτωση περίπου 0.3, που δεν είναι αμελητέα. Για να διορθωθεί αυτή η κατάσταση, έχει προταθεί να μετασχηματιστούν οι ποσότητες $\widehat{g_1}$ και $\widehat{g_2}$ με τρόπο που να καθιστά την κατανομή τους όσο το δυνατόν πιο κοντά στην τυποποιημένη κανονική.
Συγκεκριμένα, οι D'Agostino - Pearson (1973) πρότειναν τον ακόλουθο μετασχηματισμό για τη δειγματική λοξότητα:

$$Z_1(\widehat{g_1}) = \delta asinh\left(\frac{\widehat{g_1}}{\alpha sqrt(\mu_2)}\right) \quad \delta = \frac{1}{\sqrt{(lnW)}}, \alpha^2 = \frac{2}{W^2-1},$$
$$W^2 = \sqrt{(2\gamma_2+4)} - 1$$

όπου $\gamma_2 = \frac{\mu_4(\widehat{g_1})}{\mu_2(\widehat{g_1})^2} - 3 = \frac{36(15n^6-36n^5-628n^4+982n^3+5777n^2-6402n+900)}{n(n-3)(n-2)(n+7)(n+9)(n+11)(n+13)}$ $\mu_2(\widehat{g_1}) = \frac{6(n-2)}{(n+1)(n+3)}$

Ομοίως, οι Anscombe & Glynn (1983) πρότειναν έναν μετασχηματισμό για το $\widehat{g_2}$, ο οποίος λειτουργεί αρκετά καλά για μεγέθη δείγματος 20 ή μεγαλύτερα:

$$Z_2(\widehat{g_2}) = \sqrt{\left(\frac{9A}{2}\right)}\left(1 - \frac{2}{9A} - \left(\frac{1-2/A}{1+(g_2-\mu_1)/\sqrt{(\mu_2)}\sqrt{(2/(A-4))}}\right)^{(1/3)}\right)$$

όπου

$A = 6 + \frac{8}{\gamma_1}(\frac{2}{\gamma_1} + \sqrt{(1 + 4/\gamma_1^2)}),$ $\mu_1 = \frac{-6}{n+1}, \mu_2 = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)},$

$$\gamma_1(\widehat{g_2}) = \frac{6(n^2 - 5n + 2)}{(n+7)(n+9)}\sqrt{(\frac{6(n+3)(n+5)}{n(n-2)(n-3)})}$$

Οι συναρτήσεις Z_1και Z_2 μπορούν να συνδυαστούν για να δημιουργήσουν έναν στατιστικό έλεγχο, ικανό να ανιχνεύσει αποκλίσεις από την κανονικότητα που οφείλονται είτε σε λοξότητα είτε σε κύρτωση (D'Agostino, Belanger & D'Agostino 1990):

$K^2 = Z_1^2(\widehat{g_1}) + Z_2^2(\widehat{g_2})$

Εάν η μηδενική υπόθεση της κανονικότητας είναι αληθής, τότε η K^2 κατανέμεται κατά προσέγγιση σαν X^2 με 2 βαθμούς ελευθερίας.
Σημειώστε ότι οι στατιστικές $\widehat{g_1}, \widehat{g_2}$ δεν είναι ανεξάρτητες, αλλά μόνο ασυσχέτιστες. Επομένως, οι μετασχηματισμοί τους Z_1, Z_2 θα είναι επίσης εξαρτημένοι (Shenton & Bowman 1977), καθιστώντας αμφισβητήσιμη την εγκυρότητα της X^2 προσέγγισης.

3.3.4 Jarque-Bera Test

Ένας έλεγχος κανονικότητας που συνιστάται από ορισμένους συγγραφείς είναι ο έλεγχος Jarque-Bera. Αυτό βασίζεται στην κατανομή ενός συνδυασμένου μέτρου λοξότητας και κύρτωσης. Η στατιστική ελεγχοσυνάρτηση είναι πάντα μη αρνητική. Εάν απέχει πολύ από το μηδέν, αυτό σημαίνει ότι τα δεδομένα δεν ακολουθούν την κανονική κατανομή.
Η στατιστική ελεγχοσυνάρτηση Jarque-Bera ορίζεται ως:

$$J = \frac{n}{6}(\widehat{g_1}^2 + \frac{\widehat{g_2}^2}{4})$$

όπου

n: αριθμός των παρατηρήσεων
$\widehat{g_1}$: η λοξότητα του δείγματος
$\widehat{g_2}$: η κύρτωση του δείγματος

Η στατιστική ελεγχοσυνάρτηση Jarque-Bera έχει ασυμπτωτική κατανομή chi-square με δύο βαθμούς ελευθερίας, όποτε μπορεί να χρησιμοποιηθεί ο πίνακας chi-square για το αν τα δεδομένα προέρχονται από κανονική κατανομή. Ωστόσο, η σύγκλιση σε αυτή την κατανομή είναι αργή και ακανόνιστη. Για μικρά δείγματα η προσέγγιση chi-squared είναι υπερβολικά ευαίσθητη, απορρίπτοντας συχνά τη μηδενική υπόθεση όταν αυτή είναι αληθής. Επιπλέον, η κατανομή των p-τιμών απομακρύνεται από μια ομοιόμορφη κατανομή. Αυτό οδηγεί σε μεγάλο ποσοστό σφάλματος τύπου I. Ο πίνακας 7 στο παράρτημα παρουσιάζει ορισμένες p-τιμές που προσεγγίζονται από μια κατανομή chi-squared για διάφορα επίπεδα σημαντικότητας α για μικρά δείγματα.

ΚΕΦΑΛΑΙΟ 4

ΕΛΕΓΧΟΙ ΥΠΟΘΕΣΕΩΝ ΚΑΝΟΝΙΚΟΤΗΤΑΣ ΓΙΑ ΠΡΟΣΟΜΟΙΩΜΕΝΑ ΚΑΙ ΠΡΑΓΜΑΤΙΚΑ ΔΕΔΟΜΕΝΑ

4.1 Προσομοίωση δεδομένων με την βοήθεια της R

Σε αυτό το κεφάλαιο χρησιμοποιώντας κάποιους από τους ελέγχους υποθέσεων που αναλύσαμε στο προηγούμενο κεφάλαιο θα εξετάσουμε αν ένα σετ δεδομένων που θα προσομοιώσουμε στην R ακολουθεί την κανονική κατανομή. Θα χωρίσουμε το κεφάλαιο σε δύο μέρη. Στο πρώτο μέρος μέσω της R θα προσομοιώσουμε δεδομένα τα οποία θα ακολουθούν την κανονική κατανομή βάσει της εντολής rnorm(). Σκοπός μας είναι να χρησιμοποιήσουμε τους διάφορους ελέγχους για την κανονική κατανομή και να επιβεβαιώσουμε μέσω αυτών ότι το σετ των δεδομένων μας ακολουθεί κανονική κατανομή. Στο δεύτερο μέρος θα χρησιμοποιήσουμε πραγματικά δεδομένα και θα εξετάσουμε αν αυτά ακολουθούν την κανονική κατανομή μέσω στατιστικής συμπερασματολογίας.

Προσομοίωση δεδομένων από την R:

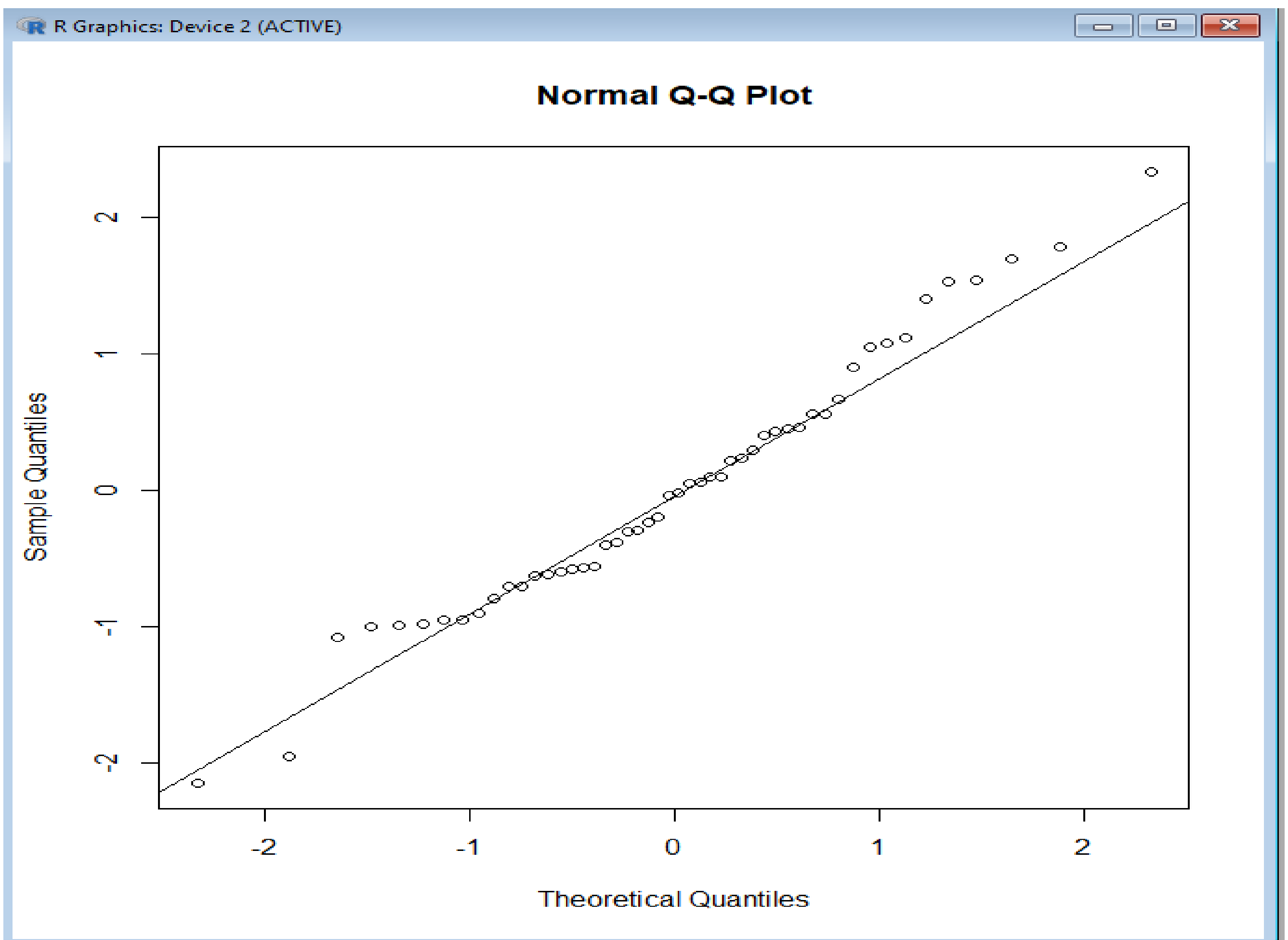
```
data1=rnorm(50)
data1
 [1]  1.53673680 -0.02408032 -0.58041600  0.22944976 -0.55704138  0.04341893
 [7]  0.21425074 -0.98261446  0.28765112  1.52808894 -0.56640129  1.04395435
[13] -2.14985155  0.66585782  0.89630691  0.05863916 -0.63376510  0.44895076
[19] -0.29788424 -0.95007069 -1.07717304 -0.40798884  0.55787516  2.33387676
[25] -0.30313298 -0.94816237  0.40190161 -0.70932739 -0.99363021  1.77758852
[31]  0.10022295 -0.24094516  0.46087230  0.43381900 -0.79425935 -0.62294457
[37] -0.19579926  0.55160707  1.69367756  1.07357667 -0.70409915 -0.04461879
[43]  1.39503966 -0.59677784 -0.38791184 -0.90576979 -0.99762853  1.11914232
[49] -1.95444255  0.09401101
```

Αυτά είναι τα τυχαία δεδομένα ή αλλιώς ένα τυχαίο δείγμα μεγέθους 50 που πήραμε από την R το οποίο ακολουθεί συγκεκριμένα την τυποποιημένη κανονική κατανομή. Γραφικά φτιάχνοντας ένα Q-Q plot διάγραμμα με την βοήθεια της R βλέπουμε ότι:

qqnorm(data1)

> qqline(data1)

Στο διάγραμμα Q-Q plot ο κάθετος άξονας μας είναι οι τιμές του δείγματός μας ενώ ο οριζόντιος άξονας έχει τα θεωρητικά ποσοστημόρια τα οποία είναι οι τιμές που θα περιμέναμε να πάρουμε σε ένα δείγμα αυτού του μεγέθους από την τυποποιημένη κανονική κατανομή. Όπως παρατηρούμε από την εικόνα του διαγράμματος οι παρατηρήσεις βρίσκονται κοντά στην ευθεία $y = x$ οπότε επιβεβαιώνουμε πως το δείγμα μας ακολουθεί την κανονική κατανομή.



Για να βρούμε την μέση τιμή ενός σετ δεδομένων στην R η εντολή είναι:

mean(data1)
[1] 0.006395584

Και η μέση τιμή των προσομοιωμένων δεδομένων μας βλέπουμε ότι είναι περίπου 0 όπως αναμέναμε.
Για να χρησιμοποιήσουμε ελέγχους υποθέσεων για την κανονική κατανομή χρειάζεται στην R να κατεβάσουμε κάποια πακέτα. Το ένα από αυτό είναι το nortsTest. Για να το κατεβάσουμε στην R εισάγουμε την εντολή: **install.packages("nortsTest")**. Το επίπεδο σημαντικότητας που θα χρησιμοποιήσουμε στους παρακάτω ελέγχους θα είναι πάντα το 0.05(5%).

Shapiro-Wilk Test

Για να χρησιμοποιήσουμε στην R τον έλεγχο Shapiro-Wilk γράφουμε την εντολή shapiro.test().
shapiro.test(data1)

Shapiro-Wilk normality test

data: data1
W = 0.97714, p-value = 0.4385

Βλέπουμε ότι η $p-value$ είναι πολύ μεγαλύτερη από το 0.05 άρα η μηδενική υπόθεση δεν απορρίπτεται και συμπεραίνουμε ότι τα δεδομένα μας ακολουθούν την κανονική κατανομή.

Όπως βλέπουμε από τον πίνακα των Shapiro-Wilk που δίνεται στο Παράρτημα στον Πίνακα 1, για $n = 50$ και α=5%, η κρίσιμη τιμή είναι 0.988. Εφόσον η τιμή της ελεγχοσυνάρτησης $W = 0.977$ δεν ξεπερνά το 0.988, η μηδενική υπόθεση της κανονικότητας δεν μπορεί να απορριφθεί. Επομένως έχουμε ένα σωστό συμπέρασμα με τον έλεγχο αυτό.

Shapiro-Francia Test

Για τον έλεγχο Shapiro-Francia στην R χρησιμοποιούμε το πακέτο nortest.
nortest::sf.test(data1)

Shapiro-Francia normality test

data: data1
W = 0.97646, p-value = 0.3497

Παρατηρούμε πάλι ότι η $p-value$ είναι πολύ μεγαλύτερη από το 0.05 άρα δεν έχουμε σοβαρές ενδείξεις να απορρίψουμε την μηδενική υπόθεση δηλαδή ότι τα δεδομένα μας ακολουθούν την κανονική κατανομή.

Kolmogorov-Smirov Test

Ο έλεγχος Kolmogorov-Smirnov στην R είναι:
ks.test(data1,"pnorm")

One-sample Kolmogorov-Smirnov test

data: data1
D = 0.1007, p-value = 0.6539
alternative hypothesis: two-sided

Όπως βλέπουμε από τον πίνακα των Kolmogorov-Smirnov που δίνεται στο Παράρτημα στον Πίνακα 2, για $n = 50$ και α=5%, η κρίσιμη τιμή είναι 0.125. Εφόσον η τιμή της ελεγχοσυνάρτησης $D = 0.1007$ δεν ξεπερνά το 0.125, η μηδενική υπόθεση της κανονικότητας δεν μπορεί να απορριφθεί. Επομένως έχουμε ένα σωστό συμπέρασμα με τον έλεγχο αυτό.

Anderson-Darling Test

Για τον έλεγχο Anderson-Darling χρησιμοποιούμε το πακέτο nortest.
nortest::ad.test(data1)

Anderson-Darling normality test

data: data1
A = 0.45283, p-value = 0.2611
Βλέπουμε ότι η $p-value$ είναι πολύ μεγαλύτερη από το 0.05 άρα τα δεδομένα μας ακολουθούν την κανονική κατανομή.
Όπως βλέπουμε από τον πίνακα Anderson-Darling που δίνεται στο Παράρτημα στον Πίνακα 5, για $n = 50$ και α=5%,η κρίσιμη τιμή είναι 2.492. Η τιμή της ελεγχοσυνάρτησης είναι $A = 0.45283$ η οποία δεν ξεπερνά το 2.492 και συνεπώς η μηδενική υπόθεση της κανονικότητας δεν μπορεί να απορριφθεί. Επομένως έχουμε ένα σωστό συμπέρασμα με τον έλεγχο αυτό.

Cramer-Von Mises Test

Για τον έλεγχο Cramer-Von Mises χρειάζεται να κατεβάσουμε ένα ακόμα πακέτο στην R με την εντολή **install.packages("goftest").** Ο έλεγχος στην R γίνεται:
goftest::cvm.test(data1,'pnorm')

Cramer-von Mises test of goodness-of-fit
Null hypothesis: Normal distribution
Parameters assumed to be fixed

data: data1
omega2 = 0.068626, p-value = 0.7623

Όπως βλέπουμε από τον πίνακα Cramer-Von Mises που δίνεται στο παράρτημα στον Πίνακα 5 για $n = 50$ και $\alpha = 5\%$ η κρίσιμη τιμή είναι 0.461. Η τιμή της ελεγχοσυνάρτησης είναι $\omega^2 = 0.068626$ το οποίο δεν ξεπερνά την κρίσιμη τιμή, οπότε η μηδενική υπόθεση δεν μπορεί να απορριφθεί.

Lilliefors Test

Ο έλεγχος Liliefors στην R γίνεται με την βοήθεια του πακέτου nor.test.
nortest::lillie.test(data1)

Lilliefors (Kolmogorov-Smirnov) normality test

data: data1
D = 0.086298, p-value = 0.4641

Από τον πίνακα Lilliefors που δίνεται στο παράρτημα στον Πίνακα 3 για $n = 50$ και $\alpha = 5\%$ η κρίσιμη τιμή για τον μονόπλευρο έλεγχο είναι 0.125. Η κρίσιμη τιμή για $\alpha = 5\%$ και αμφίπλευρο έλεγχο είναι ίση περίπου με 0.145. Η τιμή της ελεγχοσυνάρτησης είναι $D = 0.086298$ η οποία δεν ξεπερνά αυτές τις κρίσιμες τιμες, άρα η μηδενική υπόθεση δεν μπορεί να απορριφθεί.

Jarque-Bera Test

Για τον έλεγχο Jarque-Bera κατεβάζουμε στην R το πακέτο tseries με την εντολή **install.packages('tseries').** Για τον έλεγχο:
tseries::jarque.bera.test(data1)

Jarque Bera Test

data: data1
X-squared = 0.6247, df = 2, p-value = 0.7317

Από τον πίνακα της X^2 κατανομής (Πίνακας 7) για βαθμούς ελευθερίας 2 και $\alpha = 5\%$ η κρίσιμη τιμή είναι 5.99. Η τιμή της ελεγχοσυνάρτησης μας είναι $x^2 = 0.6247$ η οποία δεν ξεπερνά την κρίσιμη τιμή, άρα η μηδενική υπόθεση δεν μπορεί να απορριφθεί.

D'Agostino–Pearson Omnibus Test

Για τον έλεγχο D'Agostino–Pearson Omnibus στην R θα χρησιμοποιήσουμε το πακέτο **install.packages("Rita").** Μετά γράφουμε την παρακάτω εντολή:

Rita::DPTest(data1)
[[1]]
K
1.187

[[2]]
P-value
0.552

[[3]]
Sig.
FALSE

Από τον πίνακα D'Agostino–Pearson Omnibus που βρίσκεται στο παράρτημα στον Πίνακα 6 για $n = 50$ και $\alpha = 5\%$ η κρίσιμη τιμή είναι 6.339. Η τιμή της ελεγχοσυνάρτησης είναι $K = 1.187$ η οποία δεν ξεπερνά την κρίσιμη τιμή. Επομένως, η μηδενική υπόθεση δεν μπορεί να απορριφθεί.

Kurtosis-Skewness

Από τον υπολογισμό της κύρτωσης και της λοξότητας μπορούμε να καταλάβουμε διάφορα πράγματα. Για να υπολογίσουμε στην R την κύρτωση και την λοξότητα χρειάζεται να κατεβάσουμε σε αυτή το πακέτο moments. Αυτό γίνεται με την παρακάτω εντολή

install.packages('moments')
Για την κυρτότητα μετά γράφουμε:
moments::kurtosis(data1)
[1] 2.895777
Βλέπουμε ότι η τιμή της κυρτότητας είναι πολύ κοντά στο 3 άρα διαπιστώνουμε ότι έχουμε μια συμμετρική κατανομή.
Για την λοξότητα:
moments::skewness(data1)
[1] 0.2687901
Η λοξότητα μπορεί να είναι θετική ή αρνητική. Εδώ όπως βλέπουμε είναι περίπου ίση με 0, επομένως η κατανομή μας είναι συμμετρική όπως αναμέναμε.

4.2 Έλεγχοι υποθέσεων κανονικότητας για πραγματικά δεδομένα

Είδαμε πως χρησιμοποιούμε στην R τους ελέγχους υποθέσεων για να ελέγξουμε αν τα δεδομένα μας ακολουθούν την κανονική κατανομή. Θα εφαρμόσουμε κάποιους από αυτούς τους ελέγχους σε ένα σετ δεδομένων που έχουν να κάνουν με το ύψος και το βάρος αγοριών. Αρχικά θα εισάγουμε τα δεδομένα μας στην R:

data2=read.csv("D:\\Desktop\\δεδομενα κανονικης κατανομης\\student\\student-merge.R") (ύψος των αγοριών)

data3=read.csv("D:\\Desktop\\δεδομενα κανονικης κατανομης\\student\\male-weight.ZIP_OLD")(βάρος των αγοριών)

Τώρα θα τα ενώσουμε σε ένα πίνακα με την μια στήλη να είναι το ύψος (Height) και την άλλη στήλη να είναι το βάρος (Weight). Επειδή τα δεδομένα μας είναι σε inches και σε pounds αντίστοιχα θα κάνουμε τις κατάλληλες πράξεις ώστε να τα μετατρέψουμε σε εκατοστά και κιλά.
Τα δεδομένα μας πλέον είναι σε ένα πίνακα: **data=cbind(data2,data3)**
Μετατροπή από inches σε εκατοστά: **data$Height=2.54*data$Height**
Μετατροπή από pounds σε κιλά: **data$Weight=0.45359237*data$Weight**

Παρακάτω θα δούμε τι τιμές έχουν λάβει τα δεδομένα μας με τις αλλαγές που κάναμε προηγουμένως. Γράφοντας στην R το όνομα που έχουμε ονομάσει τα δεδομένα θα μας δώσει αναλυτικά και τις $n = 192$ τιμές που έχουμε.

data

	Height	**Weight**
1	**173.6052**	**109.72107**
2	**170.2281**	**73.62279**
3	**161.1795**	**96.49763**
4	**180.8363**	**99.80959**
5	**181.9676**	**93.59870**
6	**164.5065**	**69.04227**
7	**175.9790**	**83.42829**
8	**175.8791**	**76.19041**
9	**171.8199**	**79.80025**
10	**183.9425**	**70.94170**
11	**162.4948**	**84.64257**
12	**176.8858**	**96.95136**
13	**172.5575**	**75.80774**
14	**172.5042**	**85.93134**
15	**176.3762**	**84.56512**
16	**168.0188**	**78.10268**

17	191.0232	88.91703
18	172.4487	78.41862
19	173.0858	84.36090
20	175.4877	82.74734
21	184.9141	78.97766
22	171.2500	89.68946
23	173.9809	67.66399
24	174.2900	103.76460
25	188.0459	73.48498
26	181.6817	87.24576
27	175.7176	83.65839
28	176.7261	93.81569
29	178.8184	79.47570
30	175.4535	70.00864
31	170.6715	85.05167
32	167.1506	96.57445
33	163.3326	88.46514
34	172.6548	93.06973
35	183.3611	92.60729
36	165.7867	87.49956
37	167.8691	89.57916
38	171.4762	83.37525
39	178.0662	74.32194
40	173.3597	73.98455
41	183.3187	78.07939
42	175.7168	88.01752
43	185.0907	76.48372
44	164.5478	73.11611
45	178.2662	74.68865
46	173.9683	85.69372
47	171.0203	84.84924
48	170.1570	94.83295
49	168.9089	87.09624
50	173.6168	95.86334
51	179.7671	75.12017
52	180.8884	91.20469
53	177.8339	78.66379
54	181.3641	82.28514
55	176.6621	76.99173
56	187.4991	74.07596
57	170.1721	86.05110
58	181.4029	87.14636
59	165.8094	89.90131
60	173.4164	95.03962
61	184.8240	90.15593

62	172.9724	89.84736
63	174.9976	88.58233
64	175.9954	74.81702
65	179.1290	81.58489
66	176.9459	70.42041
67	171.6920	83.69659
68	184.1706	100.09788
69	180.9826	83.14922
70	182.3953	89.10882
71	183.5027	83.73116
72	168.9421	94.13531
73	167.7137	80.37699
74	171.6460	73.98455
75	173.3463	98.07494
76	187.5184	92.83017
77	177.2983	91.13604
78	167.5373	100.19936
79	180.5407	89.10816
80	169.1543	76.49418
81	175.1369	82.29649
82	173.3409	90.11006
83	183.6847	107.91722
84	182.4112	78.49019
85	165.7021	72.88500
86	179.4269	85.54885
87	164.4191	94.50402
88	170.4430	87.74069
89	165.3984	78.97483
90	182.1211	89.52489
91	169.7555	91.45360
92	168.8370	82.10565
93	176.3053	82.60937
94	177.9325	80.53499
95	169.5288	74.83121
96	168.3368	75.16802
97	173.5543	87.58604
98	177.9928	81.95683
99	174.5818	78.36884
100	171.5924	80.49932
101	168.2842	91.08985
102	175.7228	76.08548
103	171.7271	70.03268
104	176.0036	80.73251
105	173.2043	83.43326
106	181.8820	81.28592

107	175.7600	71.92523
108	170.8454	75.52473
109	187.1604	78.01048
110	176.6245	79.09026
111	173.5114	85.15986
112	172.0570	91.39340
113	182.1326	92.95333
114	161.6260	81.05040
115	174.5518	74.39569
116	170.0513	101.27357
117	159.2758	92.06034
118	183.5364	85.90667
119	180.1080	80.81680
120	171.7290	101.28687
121	179.8360	75.76295
122	176.0341	81.85922
123	168.2570	74.58834
124	171.4302	68.02507
125	167.1479	87.50263
126	181.4670	91.57304
127	173.8996	81.58676
128	162.5293	89.11626
129	180.3448	85.07200
130	173.0749	77.87871
131	173.7243	86.95348
132	172.9644	75.22859
133	173.0771	84.21520
134	174.9046	80.86726
135	168.0181	68.48295
136	168.1633	95.22234
137	171.2776	78.99665
138	176.4566	81.59173
139	179.1103	76.08882
140	181.1995	83.02766
141	180.3449	76.90968
142	168.1540	78.66728
143	184.2600	82.05012
144	171.3976	83.23825
145	165.9900	93.43001
146	179.9439	83.35206
147	177.6437	91.60502
148	164.4391	98.03043
149	176.0433	72.29823
150	174.8497	95.40700
151	157.3208	81.01992

152	**174.2271**	**66.56005**
153	**165.6552**	**88.87553**
154	**163.4075**	**77.19047**
155	**174.6222**	**76.95378**
156	**184.1237**	**89.98306**
157	**170.7742**	**81.60477**
158	**170.8496**	**66.79768**
159	**165.4010**	**80.27871**
160	**168.3076**	**84.12677**
161	**171.9623**	**80.50983**
162	**166.4480**	**77.03245**
163	**177.4672**	**90.07514**
164	**173.9440**	**92.23816**
165	**183.4235**	**87.64206**
166	**173.1760**	**80.80065**
167	**182.8323**	**73.83558**
168	**167.8054**	**72.25115**
169	**169.3067**	**78.06974**
170	**171.7025**	**91.55174**
171	**173.3447**	**77.27052**
172	**164.6139**	**93.16323**
173	**171.4302**	**77.01444**
174	**173.1791**	**90.91683**
175	**176.6656**	**81.00659**
176	**168.6617**	**76.20988**
177	**169.1441**	**80.63349**
178	**182.7172**	**87.66926**
179	**173.4491**	**73.65412**
180	**177.6868**	**83.18150**
181	**182.5019**	**82.12566**
182	**167.0190**	**95.64984**
183	**170.2609**	**75.10061**
184	**194.8430**	**88.97144**
185	**184.3306**	**91.64672**
186	**177.1112**	**77.29831**
187	**184.9100**	**91.60211**
188	**184.2500**	**87.71886**
189	**183.6286**	**81.58117**
190	**170.8234**	**70.98264**
191	**192.8989**	**106.61027**
192	**168.4432**	**104.58050**

Θα εξετάσουμε τα παραπάνω δεδομένα στην R μέσω ελέγχων υποθέσεων με την μηδενική υπόθεση H_0 να αφορά την κανονική κατανομή έναντι της H_1 όπου τα δεδομένα δεν θα

ακολουθούν κανονική κατανομή. Ένα επίπεδο σημαντικότητας που θα θεωρήσουμε στους ελέγχους θα είναι το $\alpha = 0.05$, ενώ φυσικά θα δίνεται και η p-value του ελέγχου.
Θα ξεκινήσουμε τους ελέγχους υποθέσεων για το ύψος των αγοριών και μετά θα εξετάσουμε και το βάρος των αγοριών.

4.3 ‘Ελεγχοι υποθέσεων κανονικότητας για το ύψος των αγοριών

Αρχικά θα βρούμε την μέση τιμή και την τυπική απόκλιση των δεδομένων του ύψους και στην συνέχεια θα εφαρμόσουμε κάποιους από τους στατιστικούς ελέγχους που μελετήθηκαν στα προηγούμενα κεφάλαια.

Για την μέση τιμή: **mean(data$Height)**
[1] 174.5995
Για την τυπική απόκλιση: **sd(data$Height)**
[1] 6.670856

1. Shapiro-Wilk Test

Για τον έλεγχο Shapiro-Wilk γράφουμε στην R:
shapiro.test(data$Height)
Shapiro-Wilk normality test

data: data$Height
W = 0.99206, p-value = 0.3793

Κάνοντας τον έλεγχο Shapiro-Wilk για το ύψος των αγοριών βλέπουμε ότι η $p-value = 0.3793$ που είναι πολύ μεγαλύτερη από το 0.05. Επομένως δεν έχουμε σοβαρές ενδείξεις να απορρίψουμε την μηδενική υπόθεση και τα δεδομένα μας ακολουθούν την κανονική κατανομή.

Η ελεγχοσυνάρτηση παίρνει τιμές μεταξύ του 0 και 1. Γενικά, για τιμές της ελεγχοσυνάρτησης κοντά στο 1, η μηδενική υπόθεση της κανονικότητας δεν απορρίπτεται ενώ για μικρότερες τιμές, απορρίπτεται. Είναι δύσκολο να βρεθούν οι κρίσιμες τιμές αυτής της ελεγχοσυνάρτησης για μεγάλο n.

2. Kolmogorov-Smirnov Test

Για τον έλεγχο Kolmogorov-Smirnov:
ks.test(data$Height,"pnorm",mean=174.5995,sd=6.670856)

One-sample Kolmogorov-Smirnov test

data: data$Height

D = 0.057775, p-value = 0.5432
alternative hypothesis: two-sided

Εδώ όπως βλέπουμε από τον έλεγχο Kolmogorov-Smirnov διαπιστώνουμε ότι η $p - value$ είναι πάρα πολύ μεγαλύτερη από το 0.05. Όπως βλέπουμε από τον πίνακα των Kolmogorov-Smirnov που δίνεται στο Παράρτημα στον Πίνακα 2, για $n = 192$ και $\alpha = 5\%$, η κρίσιμη τιμή είναι περίπου 0.073 (διότι $n = 150$ είναι η τελευταία τιμή που δίνεται στον πίνακα). Εφόσον η τιμή της ελεγχοσυνάρτησης $D = 0.057775$ δεν ξεπερνά το 0.073, η μηδενική υπόθεση της κανονικότητας δεν μπορεί να απορριφθεί.

3.Anderson-Darling Test

Θα χρησιμοποιήσουμε τον έλεγχο Anderson-Darling στην R:

nortest::ad.test(data$Height)

Anderson-Darling normality test

data: data$Height
A = 0.57166, p-value = 0.1363

Από τον έλεγχο Anderson-Darling η τιμή της $p - value$ είναι μεγαλύτερη από το 0.05. Όπως βλέπουμε από τον πίνακα Anderson-Darling που δίνεται στο Παράρτημα στον Πίνακα 5, για $n = 192$ (μεγάλο δηλαδή) και $\alpha = 5\%$, η κρίσιμη τιμή είναι 2.492 . Η τιμή της ελεγχοσυνάρτησης είναι $A = 0.57166$ η οποία δεν ξεπερνά το 2.492 και συνεπώς η μηδενική υπόθεση της κανονικότητας δεν μπορεί να απορριφθεί.

4.D'Agostino–Pearson Omnibus Test

Για τον έλεγχο D'Agostino–Pearson Omnibus έχουμε:

Rita::DPTest(data$Height)

[[1]]
K
1.731

[[2]]
P-value
0.421

[[3]]
Sig.
FALSE

Όπως βλέπουμε και εδώ από αυτόν τον έλεγχο έχουμε ότι η $p - value$ είναι πολύ μεγαλύτερη από το 0.05 . Από τον πίνακα D'Agostino–Pearson Omnibus που βρίσκεται στο παράρτημα στον Πίνακα 6 για $n = 192$ και $\alpha = 5\%$ η κρίσιμη τιμή είναι 6.129. Η τιμή της ελεγχοσυνάρτησης είναι $K = 1.731$ η οποία δεν ξεπερνά την κρίσιμη τιμή. Επομένως, η μηδενική υπόθεση δεν μπορεί να απορριφθεί.

5. Jarque-Bera Test

Για τον έλεγχο Jarque-Bera έχουμε:

tseries::jarque.bera.test(data$Height)

Registered S3 method replaced by 'quantmod':
method from
as.zoo.data.frame zoo

Jarque Bera test

data: data$Height
X-square = 1.6746, df = 2, p-value = 0.4329

Από τον πίνακα της X^2 κατανομής (Πίνακας 7) για βαθμούς ελευθερίας 2 και $\alpha = 5\%$ η κρίσιμη τιμή είναι 5.99. Η τιμή της ελεγχοσυνάρτησης μας είναι $x^2 = 1.6746$ η οποία δεν ξεπερνά την κρίσιμη τιμή, άρα η μηδενική υπόθεση δεν μπορεί να απορριφθεί.

6. Lilliefors Test

Θα εφαρμόσουμε τώρα τον έλεγχο Lilliefors στην R:

nortest::lillie.test(data$Height)
Lilliefors (Kolmogorov-Smirnov) normality test

data: data$Height
D = 0.057778, p-value = 0.1219

Από τον πίνακα Lilliefors που δίνεται στο παράρτημα στον Πίνακα 3 για $n = 192$ και $\alpha = 5\%$ η κρίσιμη τιμή για τον μονόπλευρο έλεγχο είναι 0.063. Για τον αμφίπλευρο έλεγχο και $\alpha = 2\%$ η κρίσιμη τιμή είναι 0.074. Η τιμή της ελεγχοσυνάρτησης είναι $D = 0.057778$ η οποία δεν ξεπερνά καμία από αυτές τις κρίσιμες τίμες, άρα η μηδενική υπόθεση δεν μπορεί να απορριφθεί.

Αυτοί είναι κάποιοι έλεγχοι για να δούμε αν τα δεδομένα του ύψους των αγοριών ακολουθούν την κανονική κατανομή αλλά και να συγκρίνουμε τα αποτελέσματα που δίνουν οι διάφοροι έλεγχοι. Όπως βλέπουμε με βάση τους ελέγχους Shapiro-Wilk, Anderson-Darling, τον D'Agostino–Pearson Omnibus καθώς και τους ελέγχους Kolmogorov-Smirnov ,Lilliefors, Jarque-Bera τα δεδομένα μας ακολουθούν την κανονική κατανομή, δηλαδή δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση.

Kurtosis - Skewness

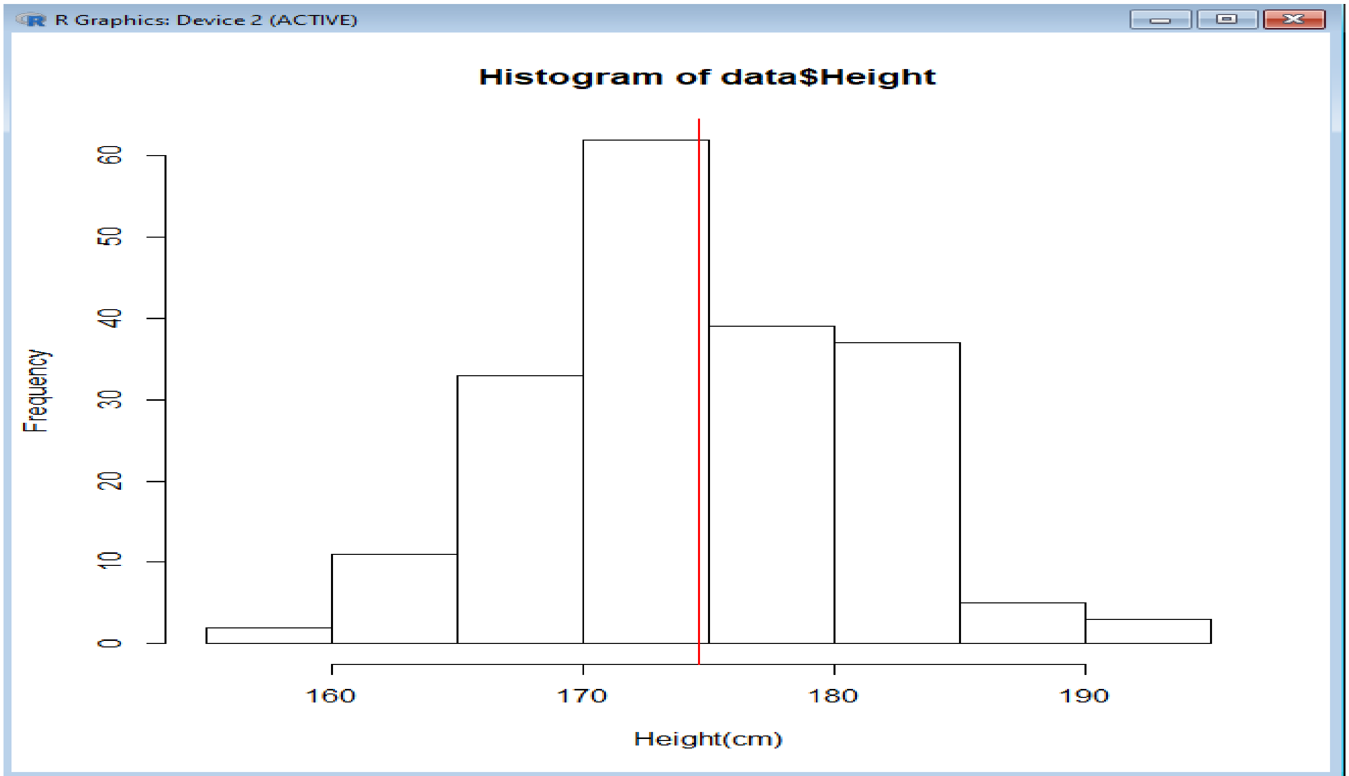
Ας δούμε την τιμή της κύρτωσης για το δείγμα μας:
moments::kurtosis(data$Height)
[1] 2.897794
Βλέποντας το αποτέλεσμα από την εντολή στην R η τιμή αυτή είναι μικρότερη αλλά πολύ κοντά στο 3 άρα έχουμε μια περίπου συμμετρική κατανομή.
Για την λοξότητα έχουμε:
moments::skewness(data$Height)
[1] 0.2229805
Η λοξότητα μπορεί να είναι θετική ή αρνητική. Εδώ όπως βλέπουμε είναι περίπου ίση με 0, επομένως η κατανομή μας είναι συμμετρική όπως αναμέναμε.

Ας εξετάσουμε τώρα γραφικά την κανονικότητα για το ύψος των αγοριών.

4.3.1 Γραφικές παραστάσεις για το ύψος των αγοριών

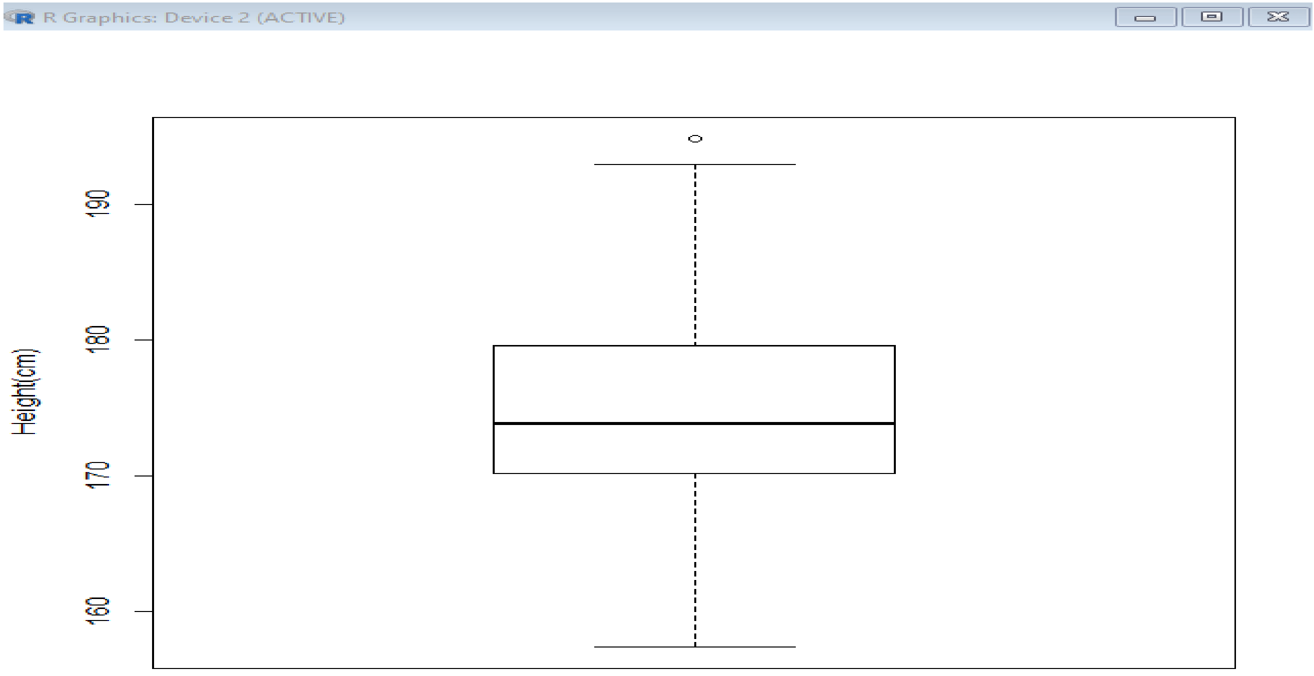
Θα κατασκευάσουμε μερικές χρήσιμες γραφικές παραστάσεις στην R για την κανονική κατανομή. Πρώτα θα δούμε το ιστόγραμμα. Για να φτιάξουμε το ιστόγραμμα στην R:

hist(data$Height,xlab='Height(cm)')
> abline(v=mean(data$Height),col='red')



Αυτό είναι το ιστόγραμμα που προκύπτει για το ύψος των αγοριών, στον οριζόντιο άξονα βλέπουμε το ύψος των αγοριών σε εκατοστά ενώ στον κάθετο άξονα πόσα αγόρια έχουν ύψος που εντάσσεται μέσα σε κάθε κατηγορία . Όπως βλέπουμε το γράφημα παρατηρούμε ότι τα περισσότερα αγόρια έχουν ύψος από $170 - 175$ εκατοστά. Τέλος με την εντολή abline έχουμε βάλει με κόκκινο χρώμα το μέσο ύψος των αγοριών όπως φαίνεται στην παραπάνω εικόνα.
Το ιστόγραμμα φαίνεται συμμετρικό ώστε το δείγμα μας να ακολουθεί την κανονική κατανομή.

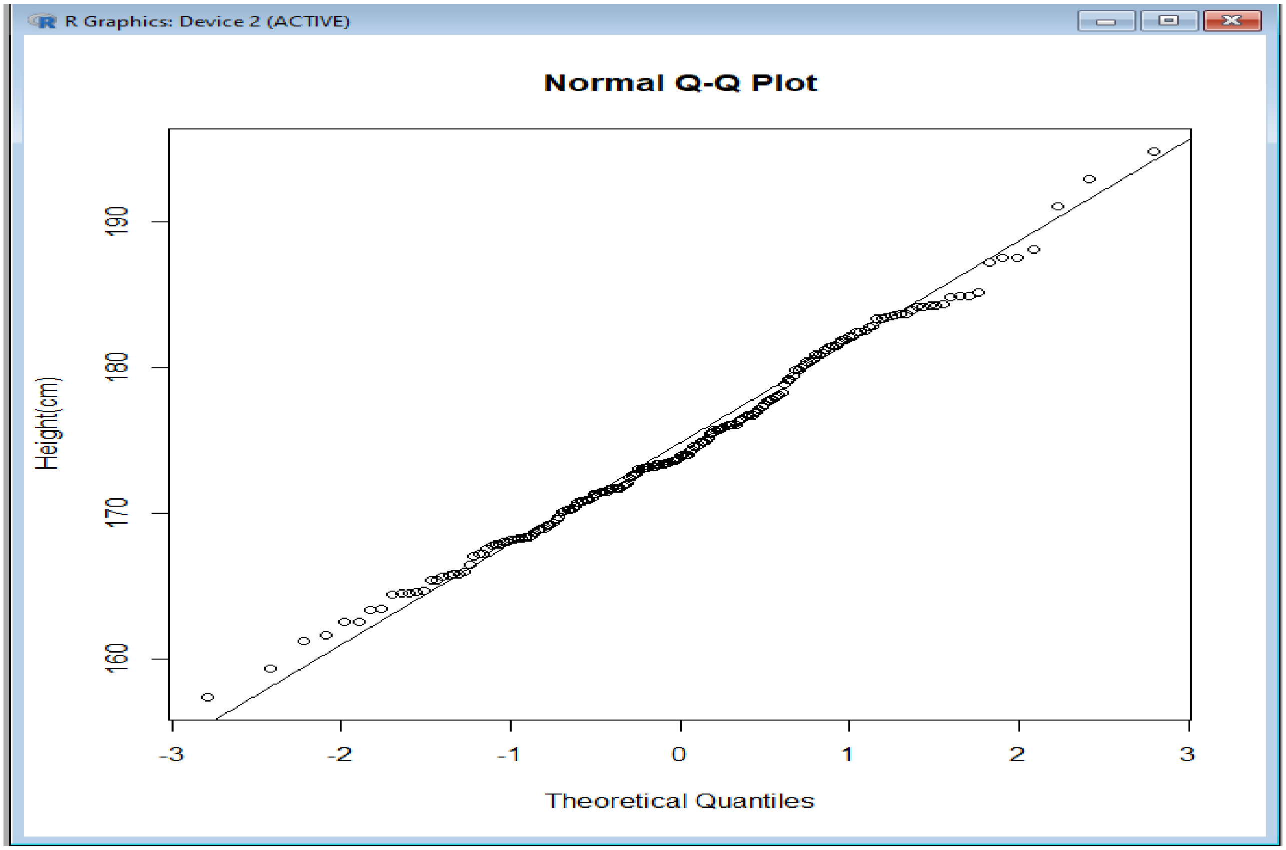
Ας δούμε τώρα το κυτιοδιάγραμμα:
boxplot(data$Height,ylab='Height(cm)')



Αυτό είναι ένα κυτιοδιάγραμμα στο οποίο βλέπουμε στον κάθετο άξονα το ύψος των αγοριών σε εκατοστά, ενώ με την μαύρη οριζόντια γραμμή εμφανίζεται η διάμεσος για το δείγμα μας. Επίσης το πάνω μέρος του ορθογωνίου λέγεται άνω τεταρτημόριο ενώ το κάτω μέρος λέγεται κάτω τεταρτημόριο, ενώ οι γραμμές πάνω και κάτω οδηγούν στο άνω άκρο-μέγιστο των παρατηρήσεων και το κάτω άκρο-ελάχιστο των παρατηρήσεων αντίστοιχα. Το κυτιοδιάγραμμα είναι η απεικόνιση των σημείων δεδομένων σε οριζόντιους και κάθετους άξονες για την παρουσίαση της κατανομής μιας συνεχούς μεταβλητής. Το πλαίσιο αντιπροσωπεύει το μεσαίο 50% περίπου των αριθμητικών τιμών, δηλαδή το 50% των παρατηρήσεων για το ύψος των αγοριών κυμαίνονται μεταξύ των $170-180$ εκατοστών.

Τέλος, ως τελευταίο διάγραμμα για το ύψος των αγοριών θα δούμε το QQ-plot διάγραμμα:

> qqnorm(data$Height,ylab='Height(cm)')
> qqline(data$Height)



Εδώ βλέπουμε το Q-Q διάγραμμα το οποίο στον κάθετο άξονα έχει το ύψος των αγοριών σε εκατοστά ενώ στον οριζόντιο άξονα έχει τα θεωρητικά ποσοστά τα οποία είναι οι τιμές που θα περιμέναμε να πάρουμε σε ένα δείγμα αυτού του μεγέθους από την τυπική κανονική κατανομή. Όπως βλέπουμε το δείγμα μας κινείται αρκετά κοντά σε σχέση με την γραμμή y=x, με λίγες τιμές να είναι μακριά από την ευθεία, έτσι βλέπουμε ότι το δείγμα μας ακολουθεί ικανοποιητικά την κανονική κατανομή.

Ας δούμε τώρα τους ελέγχους υποθέσεων για το βάρος των αγοριών.

4.4 Έλεγχοι υποθέσεων κανονικότητας για το βάρος των αγοριών

Αρχικά θα βρούμε την μέση τιμή και την τυπική απόκλιση των δεδομένων του βάρους των αγοριών και στην συνέχεια θα εφαρμόσουμε κάποιους από τους στατιστικούς ελέγχους που μελετήθηκαν στα προηγούμενα κεφάλαια.

Για την μέση τιμή: **mean(data$Weight)**
[1] 83.95505
Για την τυπική απόκλιση: **sd(data$Weight)**
[1] 8.632141

1. Shapiro-Wilk Test

Αρχικά θα ξεκινήσουμε με τον έλεγχο Shapiro-Wilk για το βάρος των αγοριών:

shapiro.test(data$Weight)
Shapiro-Wilk normality test

data: data$Weight
W = 0.98449, p-value = 0.0325

Όπως βλέπουμε από τον έλεγχο αυτό η $p-value$ είναι μικρότερη από το 0.05 . Αυτό σημαίνει ότι έχουμε ενδείξεις να απορρίψουμε την μηδενική υπόθεση, δηλαδή φαίνεται ότι τα δεδομένα μας δεν ακολουθούν την κανονική κατανομή.

Το αποτέλεσμα αυτό οφείλεται στο γεγονός ότι η ελεγχοσυνάρτηση Shapiro-Wilk δίνει μεγάλο βάρος στις ακραίες παρατηρήσεις και το δείγμα μας έχει έκτροπες παρατηρήσεις προς τα δεξιά.

2. Anderson-Darling Test

Ας δούμε τώρα τον έλεγχο Anderson-Darling:

nortest::ad.test(data$Weight)

Anderson-Darling normality test

data: data$Weight
A = 0.69029, p-value = 0.07038

Από τον έλεγχο αυτό βλέπουμε ότι η $p-value$ είναι μεγαλύτερη από το 0.05. Όπως βλέπουμε από τον πίνακα Anderson-Darling που δίνεται στο Παράρτημα στον Πίνακα 5, για $n = 192$ (μεγάλο δηλαδή) και $\alpha = 5\%$, η κρίσιμη τιμή είναι 2.492 . Η τιμή της ελεγχοσυνάρτησης είναι $A = 0.69029$ η οποία δεν ξεπερνά το 2.492 και συνεπώς η μηδενική υπόθεση της κανονικότητας δεν μπορεί να απορριφθεί.

3. Jarque-Bera Test

Για τον έλεγχο Jarque-Bera έχουμε:

tseries::jarque.bera.test(data$Weight)

Jarque Bera Test

data: data$Weight
X-squared = 5.3509, df = 2, p-value = 0.06887

Από τον πίνακα της X^2 κατανομής (Πίνακας 7) για βαθμούς ελευθερίας 2 και $\alpha = 5\%$ η κρίσιμη τιμή είναι 5.99. Η τιμή της ελεγχοσυνάρτησης μας είναι $x^2 = 5.3509$ η οποία δεν ξεπερνά την κρίσιμη τιμή, άρα η μηδενική υπόθεση δεν μπορεί να απορριφθεί.

4. Lilliefors Test

Θα εφαρμόσουμε τώρα τον έλεγχο Lilliefors στην R:

nortest::lillie.test(data$Weight)

Lilliefors (Kolmogorov-Smirnov) normality test

data: data$Weight
D = 0.057221, p-value = 0.1301

Από τον πίνακα Lilliefors που δίνεται στο παράρτημα στον Πίνακα 3 για $n = 192$ και $\alpha = 5\%$ η κρίσιμη τιμή για τον μονόπλευρο έλεγχο είναι 0.063. Για τον αμφίπλευρο έλεγχο και $\alpha = 2\%$ η κρίσιμη τιμή είναι 0.074. Η τιμή της ελεγχοσυνάρτησης είναι $D = 0.057221$ η οποία δεν ξεπερνά καμία από αυτές τις κρίσιμες τιμες, άρα η μηδενική υπόθεση δεν μπορεί να απορριφθεί.

5. Kolmogorov-Smirnov Test

Τέλος, θα κάνουμε στην R τον έλεγχο Kolmogorov-Smirnov:
ks.test(data$Weight,"pnorm",mean=83.95505,sd=8.632141)

One-sample Kolmogorov-Smirnov test

data: data$Weight
D = 0.057221, p-value = 0.5558
alternative hypothesis: two-sided

Εδώ όπως βλέπουμε από τον έλεγχο Kolmogorov-Smirnov διαπιστώνουμε ότι η $p - value$ είναι πάρα πολύ μεγαλύτερη από το 0.05. Όπως βλέπουμε από τον πίνακα των Kolmogorov-Smirnov που δίνεται στο Παράρτημα στον Πίνακα 2, για $n = 192$ και $\alpha = 5\%$, η κρίσιμη τιμή είναι περίπου 0.073 (διότι για $n = 150$ είναι η τελευταία τιμή που δίνεται στον πίνακα). Εφόσον η τιμή της ελεγχοσυνάρτησης $D = 0.057221$ δεν ξεπερνά το 0.073, η μηδενική υπόθεση της κανονικότητας δεν μπορεί να απορριφθεί.

Οι πιο πάνω έλεγχοι χρησιμοποιήθηκαν για να δούμε αν τα δεδομένα για το βάρος των αγοριών ακολουθούν την κανονική κατανομή αλλά και να συγκρίνουμε τα αποτελέσματα που δίνουν οι διάφοροι έλεγχοι. Όπως βλέπουμε με βάση τους ελέγχους Anderson-Darling, Lilliefors, Kolmogorov-Smirnov και τον Jarque-bera τα δεδομένα μας για το βάρος των αγοριών ακολουθούν την κανονική κατανομή δηλαδή βάσει των ελεγχοσυναρτήσεων δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση. Επίσης, παρατηρούμε για τους ελέγχους Anderson-Darling αλλά και για τον Jarque-bera ότι η $p-value$ είναι λίγο πιο πάνω από το 0.05 οπότε οριακά δεν απορρίπτεται η μηδενική υπόθεση. Σε αντίθεση ο έλεγχος Shapiro-Wilk οδηγεί στο συμπέρασμα ότι τα δεδομένα δεν ακολουθούν την κανονική κατανομή. Αυτός ο έλεγχος είναι πιο ευαίσθητος στις έκτροπες παρατηρήσεις και την ελαφρά λοξότητα των δεδομένων προς τα δεξιά.

Kurtosis -Skewness

Ας δούμε την τιμή της κύρτωσης για το δείγμα μας:
moments::kurtosis(data$Weight)
[1] 2.878316
Βλέποντας το αποτέλεσμα από την εντολή στην R η τιμή αυτή είναι μικρότερη από το $3,$ αλλά πολύ κοντά στο 3 άρα θα μπορούσαμε να συμπεράνουμε ότι η κατανομή των δεδομένων είναι περίπου συμμετρική.

Για την λοξότητα έχουμε:
moments::skewness(data$Weight)
[1] 0.4043696

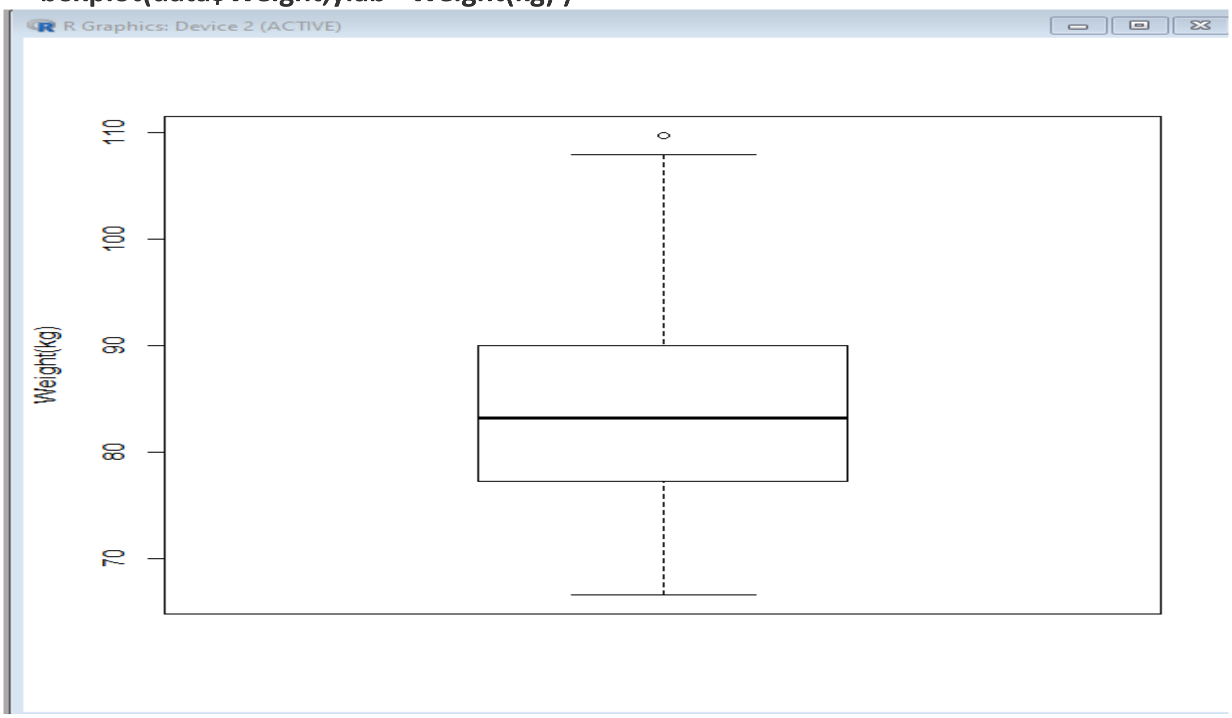
Βλέπουμε ότι η λοξότητα είναι θετική, επομένως η κατανομή είναι δεξιά (ή θετικά) λοξή εφόσον η ουρά εκτείνεται προς τα δεξιά - προς τους υψηλότερους αριθμούς. Δεν είναι τόσο μεγάλη η τιμή της όμως για να απορριφθεί η μηδενική υπόθεση από τον έλεγχο Jarque-Bera ο οποίος βασίζεται στην κύρτωση και τη λοξότητα.

Τέλος θα δούμε μερικές γραφικές παραστάσεις για το βάρος των αγοριών.

4.4.1 Γραφικές παραστάσεις για το βάρος των αγοριών

Αρχικά, θα δούμε το κυτιοδιάγραμμα για το βάρος των αγοριών με την βοήθεια της R:

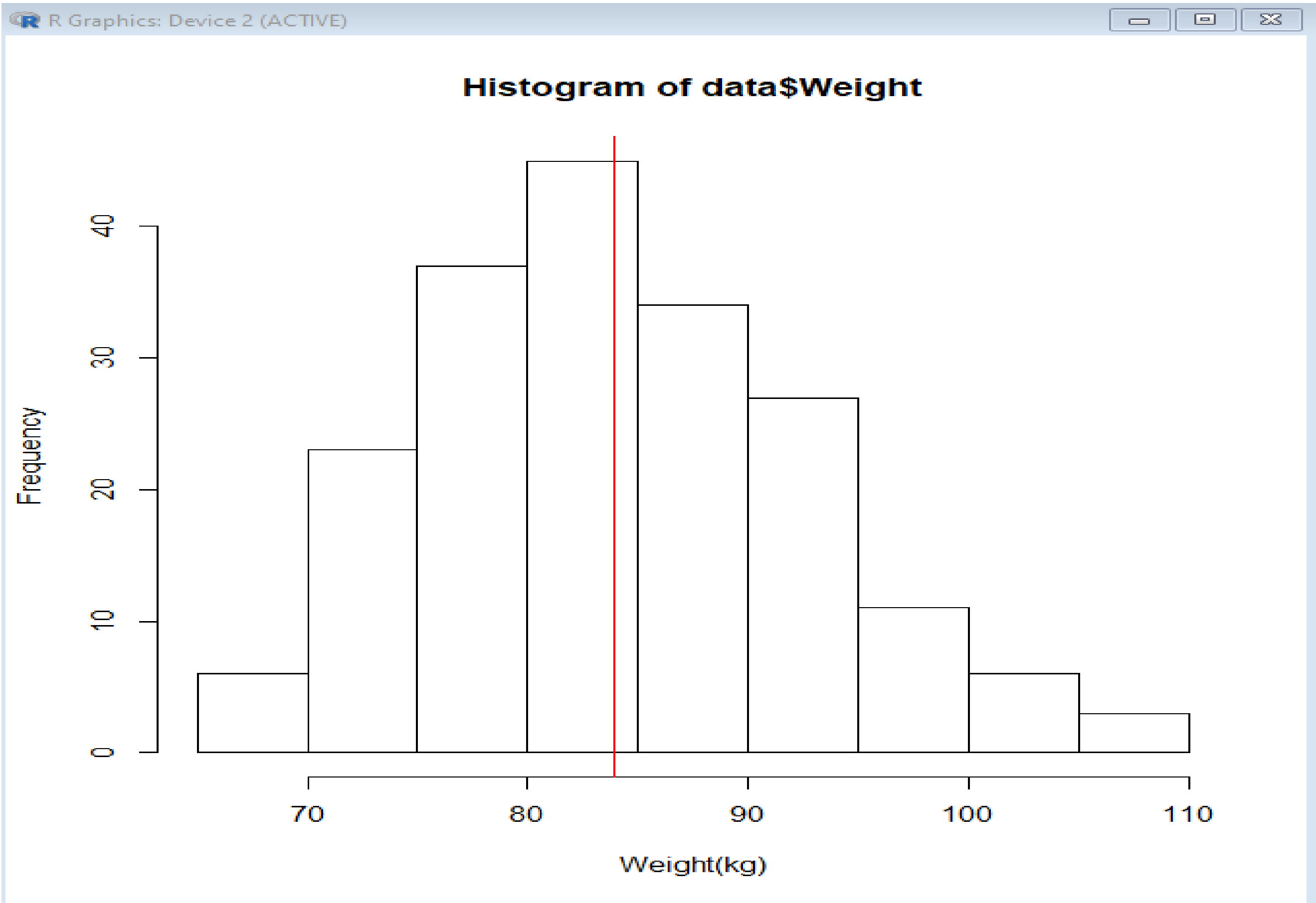
boxplot(data$Weight,ylab='Weight(kg)')



Στον κάθετο άξονα στο κυτιοδιάγραμμα βλέπουμε το βάρος των αγοριών σε κιλά. Η μαύρη οριζόντια γραμμή είναι η διάμεσος του δείγματός μας. Όπως είπαμε και πριν, το πλαίσιο αντιπροσωπεύει περίπου τις το δείγμα μας δηλαδή τα περισσότερα αγόρια έχουν βάρος σε κιλά μεταξύ του $75-90$. Από το κυτιοδιάγραμμα, βλέπουμε κάποιες έκτροπες παρατηρήσεις προς τα μεγαλύτερα βάρη οι οποίες καθιστούν τα δεδομένα λίγο λοξά προς τα δεξιά.

Ας δούμε τώρα το ιστόγραμμα:

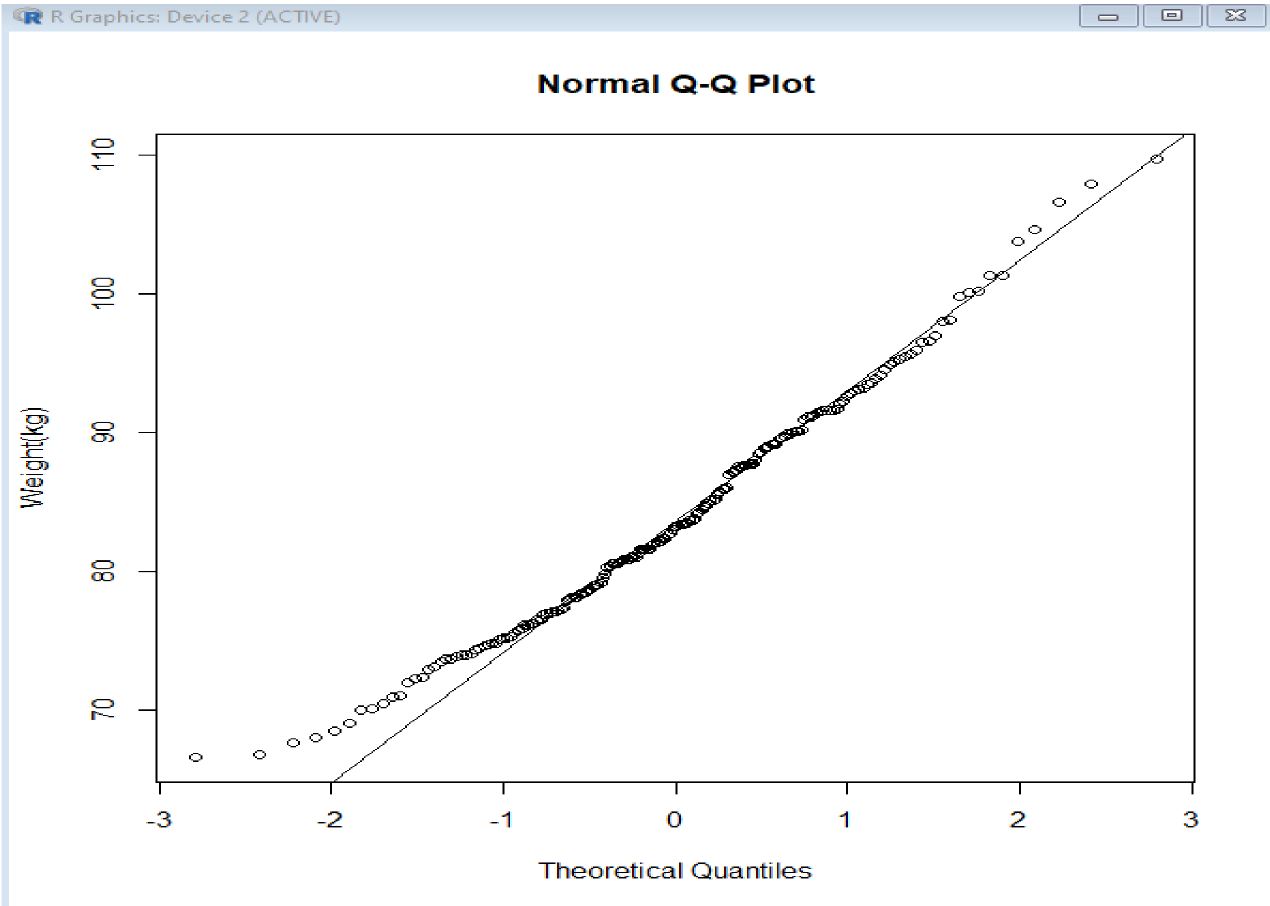
hist(data$Weight,xlab='Weight(kg)')
> abline(v=mean(data$Weight),col='red')



Το πιο πάνω ιστόγραμμα προκύπτει η R για το βάρος των αγοριών. Στον οριζόντιο άξονα βλέπουμε τις τιμές βάρους σε κιλά για τα αγόρια ενώ στον κατακόρυφο άξονα έχουμε πόσα άτομα από το δείγμα μας έχουν βάρος μέσα σε ένα συγκεκριμένο εύρος κιλών (σε κάθε κατηγορία). Τέλος, η κόκκινη γραμμή που έχουμε βάλει στην εικόνα του ιστογράμματος είναι η μέση τιμή για το δείγμα μας. Η μικρή λοξότητα προς τα δεξιά φαίνεται και από το ιστόγραμμα.

Τέλος θα δούμε το διάγραμμα QQ-Plot:

```
qqnorm(data$Weight,ylab='Weight(kg)')
> qqline(data$Weight)
```



Εδώ βλέπουμε το Q-Q διάγραμμα το οποίο στον κάθετο άξονα έχει το βάρος των αγοριών σε κιλά ενώ στον οριζόντιο άξονα έχει τα θεωρητικά ποσοστά τα οποία είναι οι τιμές που θα περιμέναμε να πάρουμε σε ένα δείγμα αυτού του μεγέθους από την τυπική κανονική κατανομή. Όπως βλέπουμε το δείγμα μας είναι αρκετά κοντά σε σχέση με την γραμμή y=x, με λίγες έκτροπες τιμές στα άκρα να είναι μακριά από την ευθεία, έτσι μπορούμε να συμπεράνουμε ότι το δείγμα μας ακολουθεί περίπου την κανονική κατανομή.

ΣΥΜΠΕΡΑΣΜΑΤΑ

Η παρούσα διπλωματική εργασία είχε ως στόχο τον έλεγχο υποθέσεων για το αν τα δεδομένα μας ακολουθούν την κανονική κατανομή, καθώς και την πρακτική εφαρμογή των ελέγχων αυτών σε προσομοιωμένα και πραγματικά δεδομένα με σκοπό την εξαγωγή συμπερασμάτων.

Με την βοήθεια της R χρησιμοποιούμε μια γεννήτρια δεδομένων έτσι ώστε να παράγουμε προσομοιωμένα δεδομένα τα οποία να ακολουθούν την κανονική κατανομή. Σκοπός μας ήταν να δείξουμε πως χρησιμοποιούνται οι έλεγχοι υποθέσεων στην R και ποια στατιστικά πακέτα χρειάζεται να κατεβάσουμε ώστε να μπορούμε να κάνουμε τους ελέγχους. Στην συνέχεια, εφαρμόζουμε τους ελέγχους υποθέσεων σε πραγματικά δεδομένα και συγκεκριμένα σε ένα σετ δεδομένων που αποτελείται από το ύψος και το βάρος 192 αγοριών για να διαπιστώσουμε αν αυτά τα δεδομένα (ύψος και βάρος) ακολουθούν την κανονική κατανομή. Για το ύψος των αγοριών χρησιμοποιήθηκαν οι εξής έλεγχοι: Shapiro-Wilk, Kolmogorov Smirnov, Anderson-Darling, D'Agostino–Pearson Omnibus, Jarque-Bera, και Lilliefors. Όπως είδαμε με αυτούς τους ελέγχους τα δεδομένα μας πάντα ακολουθούσαν την κανονική κατανομή δηλαδή δεν μπορούσαμε να απορρίψουμε την μηδενική υπόθεση της κανονικότητας. Παρατηρήσαμε ότι για τους ελέγχους Shapiro-Wilk , Kolmogorov Smirnov, Anderson-Darling και Jarque Bera η $p-value$ είναι μεγαλύτερη από το 0.3 ενώ για τους ελέγχους Anderson-Darling και Lilliefors η τιμή της $p-value$ είναι κοντά στο 0.10. Αυτό οφείλεται στο ότι οι έλεγχοι Anderson-Darling και Lilliefors χρησιμοποιούνται καλύτερα για μικρότερα δείγματα. Σχετικά με το βάρος των αγοριών, χρησιμοποιήθηκαν οι έλεγχοι Shapiro-Wilk, Kolmogorov Smirnov, Anderson-Darling, Jarque-Bera, και Lilliefors και εδώ είχαμε κάποιες σημαντικές διαφορές. Όπως είδαμε με βάση τους ελέγχους Anderson-Darling, Lilliefors, Kolmogorov-Smirnov και τον Jarque-Bera τα δεδομένα για το βάρος των αγοριών ακολουθούν την κανονική κατανομή δηλαδή βάσει αυτών των ελεγχοσυναρτήσεων δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση της κανονικότητας. Όμως για τον έλεγχο Shapiro-Wilk δεν μπορούμε να πούμε το ίδιο. Κάνοντας τον έλεγχο είδαμε ότι η $p-value$ είναι μικρότερή από το 0.05 και άρα έχουμε σοβαρές ενδείξεις να απορρίψουμε την μηδενική υπόθεση και άρα τα δεδομένα να μην ακολουθούν την κανονική κατανομή. Αυτό οφείλεται στο ότι ο έλεγχος αυτός είναι πιο ευαίσθητος στις έκτροπες παρατηρήσεις και την ελαφρά λοξότητα των δεδομένων προς τα δεξιά. Τέλος, τόσο για το ύψος των αγοριών όσο και για το βάρος χρησιμοποιήθηκαν κάποια διαγράμματα όπως το ιστόγραμμα, το QQ-plot και το κυτιοδιάγραμμα, για την καλύτερη κατανόηση των δεδομένων μας.

Συνοψίζοντας, η εργασία αυτή αποτελεί μια σημαντική πηγή γνώσης για τους ελέγχους υποθέσεων για την κανονική κατανομή και για όσους ασχολούνται με τη στατιστική ανάλυση δεδομένων.

ΠΑΡΑΡΤΗΜΑ

ΣΤΑΤΙΣΤΙΚΟΙ ΠΙΝΑΚΕΣ

	Level								
n	0·01	0·02	0·05	0·10	0·50	0·90	0·95	0·98	0·99
3	0·753	0·756	0·767	0·789	0·959	0·998	0·999	1·000	1·000
4	·687	·707	·748	·792	·935	·987	·992	·996	·997
5	·686	·715	·762	·806	·927	·979	·986	·991	·993
6	0·713	0·743	0·788	0·826	0·927	0·974	0·981	0·986	0·989
7	·730	·760	·803	·838	·928	·972	·979	·985	·988
8	·749	·778	·818	·851	·932	·972	·978	·984	·987
9	·764	·791	·829	·859	·935	·972	·978	·984	·986
10	·781	·806	·842	·869	·938	·972	·978	·983	·986
11	0·792	0·817	0·850	0·876	0·940	0·973	0·979	0·984	0·986
12	·805	·828	·859	·883	·943	·973	·979	·984	·986
13	·814	·837	·866	·889	·945	·974	·979	·984	·986
14	·825	·846	·874	·895	·947	·975	·980	·984	·986
15	·835	·855	·881	·901	·950	·975	·980	·984	·987
16	0·844	0·863	0·887	0·906	0·952	0·976	0·981	0·985	0·987
17	·851	·869	·892	·910	·954	·977	·981	·985	·987
18	·858	·874	·897	·914	·956	·978	·982	·986	·988
19	·863	·879	·901	·917	·957	·978	·982	·986	·988
20	·868	·884	·905	·920	·959	·979	·983	·986	·988
21	0·873	0·888	0·908	0·923	0·960	0·980	0·983	0·987	0·989
22	·878	·892	·911	·926	·961	·980	·984	·987	·989
23	·881	·895	·914	·928	·962	·981	·984	·987	·989
24	·884	·898	·916	·930	·963	·981	·984	·987	·989
25	·888	·901	·918	·931	·964	·981	·985	·988	·989
26	0·891	0·904	0·920	0·933	0·965	0·982	0·985	0·988	0·989
27	·894	·906	·923	·935	·965	·982	·985	·988	·990
28	·896	·908	·924	·936	·966	·982	·985	·988	·990
29	·898	·910	·926	·937	·966	·982	·985	·988	·990
30	·900	·912	·927	·939	·967	·983	·985	·988	·900
31	0·902	0·914	0·929	0·940	0·967	0·983	0·986	0·988	0·990
32	·904	·915	·930	·941	·968	·983	·986	·988	·990
33	·906	·917	·931	·942	·968	·983	·986	·989	·990
34	·908	·919	·933	·943	·969	·983	·986	·989	·990
35	·910	·920	·934	·944	·969	·984	·986	·989	·990
36	0·912	0·922	0·935	0·945	0·970	0·984	0·986	0·989	0·990
37	·914	·924	·936	·946	·970	·984	·987	·989	·990
38	·916	·925	·938	·947	·971	·984	·987	·989	·990
39	·917	·927	·939	·948	·971	·984	·987	·989	·991
40	·919	·928	·940	·949	·972	·985	·987	·989	·991
41	0·920	0·929	0·941	0·950	0·972	0·985	0·987	0·989	0·991
42	·922	·930	·942	·951	·972	·985	·987	·989	·991
43	·923	·932	·943	·951	·973	·985	·987	·990	·991
44	·924	·933	·944	·952	·973	·985	·987	·990	·991
45	·926	·934	·945	·953	·973	·985	·988	·990	·991
46	0·927	0·935	0·945	0·953	0·974	0·985	0·988	0·990	0·991
47	·928	·936	·946	·954	·974	·985	·988	·990	·991
48	·929	·937	·947	·954	·974	·985	·988	·990	·991
49	·929	·937	·947	·955	·974	·985	·988	·990	·991
50	·930	·938	·947	·955	·974	·985	·988	·990	·991

Πίνακας 1 : Ποσοστημόρια για τον έλεγχο Shapiro-Wilk

n	α = 0.15	0.10	0.05	0.025	0.01
1	0.421	0.445	0.486	0.519	0.562
2	0.386	0.408	0.446	0.476	0.516
3	0.350	0.370	0.404	0.432	0.468
4	0.321	0.339	0.371	0.395	0.429
5	0.297	0.314	0.343	0.366	0.397
6	0.278	0.294	0.321	0.343	0.371
7	0.262	0.277	0.303	0.323	0.350
8	0.248	0.263	0.287	0.306	0.332
9	0.237	0.250	0.273	0.292	0.316
10	0.227	0.239	0.262	0.279	0.303
11	0.218	0.230	0.251	0.268	0.290
12	0.209	0.221	0.242	0.258	0.280
13	0.202	0.214	0.234	0.249	0.270
14	0.196	0.207	0.226	0.241	0.261
15	0.190	0.201	0.219	0.234	0.254
16	0.184	0.195	0.213	0.227	0.246
17	0.179	0.190	0.207	0.221	0.240
18	0.175	0.185	0.202	0.215	0.233
19	0.171	0.180	0.197	0.210	0.228
20	0.167	0.176	0.192	0.205	0.222
21	0.163	0.172	0.188	0.201	0.218
22	0.159	0.168	0.184	0.196	0.213
23	0.156	0.165	0.180	0.192	0.205
24	0.153	0.162	0.177	0.189	0.204
25	0.150	0.159	0.173	0.165	0.201
26	0.147	0.156	0.170	0.182	0.197
27	0.145	0.153	0.167	0.179	0.193
28	0.142	0.150	0.164	0.175	0.190
29	0.140	0.148	0.162	0.173	0.187
30	0.138	0.146	0.159	0.170	0.184
31	0.136	0.143	0.157	0.167	0.181
32	0.134	0.141	0.154	0.165	0.179
33	0.132	0.139	0.152	0.162	0.176
34	0.130	0.137	0.150	0.160	0.173
35	0.128	0.135	0.148	0.158	0.171
36	0.126	0.134	0.146	0.156	0.165
37	0.125	0.132	0.144	0.154	0.167
38	0.123	0.130	0.142	0.152	0.164
39	0.122	0.129	0.140	0.150	0.162
40	0.120	0.127	0.139	0.148	0.160
41	0.119	0.126	0.137	0.146	0.159
42	0.117	0.124	0.136	0.145	0.157
43	0.116	0.123	0.134	0.143	0.155
44	0.115	0.121	0.133	0.141	0.153
45	0.114	0.120	0.131	0.140	0.152
46	0.112	0.119	0.130	0.138	0.150
47	0.111	0.118	0.128	0.137	0.149
48	0.110	0.116	0.127	0.136	0.147
49	0.109	0.115	0.126	0.134	0.146
50	0.108	0.114	0.125	0.133	0.144

n	α = 0.15	0.10	0.05	0.025	0.01
52	0.106	0.112	0.122	0.130	0.141
54	0.104	0.110	0.120	0.128	0.129
56	0.102	0.108	0.118	0.126	0.136
58	0.100	0.106	0.116	0.124	0.134
60	0.099	0.104	0.114	0.122	0.132
62	0.097	0.103	0.112	0.120	0.130
64	0.096	0.101	0.111	0.118	0.128
66	0.094	0.100	0.109	0.116	0.126
68	0.093	0.098	0.107	0.115	0.124
70	0.092	0.057	0.106	0.113	0.122
72	0.090	0.056	0.104	0.111	0.121
74	0.089	0.094	0.103	0.110	0.119
76	0.088	0.093	0.102	0.108	0.118
78	0.087	0.092	0.100	0.107	0.116
80	0.086	0.091	0.059	0.106	0.115
82	0.085	0.090	0.098	0.104	0.113
84	0.084	0.089	0.097	0.103	0.112
86	0.083	0.088	0.056	0.102	0.111
88	0.082	0.087	0.095	0.101	0.109
90	0.081	0.086	0.094	0.100	0.108
92	0.080	0.085	0.093	0.099	0.107
94	0.079	0.084	0.092	0.058	0.106
96	0.078	0.083	0.091	0.097	0.105
98	0.078	0.082	0.090	0.056	0.104
100	0.077	0.081	0.089	0.095	0.103
105	0.075	0.079	0.087	0.053	0.100
110	0.073	0.078	0.085	0.050	0.098
115	0.072	0.076	0.083	0.088	0.096
120	0.070	0.074	0.081	0.087	0.094
125	0.069	0.073	0.080	0.085	0.092
130	0.068	0.071	0.078	0.083	0.090
135	0.066	0.070	0.077	0.082	0.089
140	0.065	0.069	0.075	0.080	0.087
145	0.064	0.068	0.074	0.079	0.086
150	0.063	0.067	0.073	0.078	0.084

Πίνακας 2 : Πίνακας για τον έλεγχο Kolmogorov-Smirnov από n=1 έως n=150

Table A22 Table of Critical Values for the Lilliefors Test for Normality

One-tailed Two-tailed	.20 .40	.15 .30	.10 .20	.05 .10	.01 .02
n = 4	.300	.319	.352	.381	.417
5	.285	.299	.315	.337	.405
6	.265	.277	.294	.319	.364
7	.247	.258	.276	.300	.348
8	.233	.244	.261	.285	.331
9	.223	.233	.249	.271	.311
10	.215	.224	.239	.258	.294
11	.206	.217	.230	.249	.284
12	.199	.212	.223	.242	.275
13	.190	.202	.214	.234	.268
14	.183	.194	.207	.227	.261
15	.177	.187	.201	.220	.257
16	.173	.182	.195	.213	.250
17	.169	.177	.189	.206	.245
18	.166	.173	.184	.200	.239
19	.163	.169	.179	.195	.235
20	.160	.166	.174	.190	.231
25	.142	.147	.158	.173	.200
30	.131	.136	.144	.161	.187
n > 30	$.736/\sqrt{n}$	$.768/\sqrt{n}$	$.805/\sqrt{n}$	$.886/\sqrt{n}$	$1.031/\sqrt{n}$

Πίνακας 3 : Πίνακας για τον έλεγχο Lilliefors χρησιμοποιώντας τα Z-scores

k	α = 0.25	0.15	0.10	0.05	0.025	0.01	0.005	0.001
3	0.1980	0.282	0.351	0.472	0.603	0.783	0.922	1.215
4	0.205	0.284	0.351	0.470	0.595	0.767	0.899	1.213
5	0.207	0.284	0.350	0.467	0.590	0.750	0.888	1.204
6	0.208	0.284	0.349	0.465	0.587	0.754	0.883	1.188
8	0.209	0.284	0.348	0.464	0.584	0.749	0.877	1.179
10	0.209	0.284	0.348	0.463	0.583	0.748	0.874	1.175
20	0.209	0.284	0.347	0.462	0.581	0.743	0.871	1.170
40	0.209	0.284	0.347	0.461	0.581	0.744	0.870	1.168
∞	0.209	0.284	0.347	0.461	0.581	0.743	0.869	1.167

Πίνακας 4 : Πίνακας για τον έλεγχο Cramer-von-Mises

Anderson-Darling statistic, A^2								
k	α = 0.25	0.15	0.10	0.05	0.025	0.01	0.005	0.001
3	0.892	1.267	1.580	2.125	2.714	3.52	4.15	5.47
4	0.989	1.363	1.675	2.235	2.821	3.63	4.24	5.71
5	1.043	1.417	1.733	2.289	2.874	3.68	4.28	5.77
6	1.079	1.452	1.763	2.324	2.909	3.72	4.33	5.80
8	1.122	1.495	1.807	2.367	2.952	3.72	4.37	5.84
10	1.147	1.521	1.832	2.392	2.977	3.78	4.40	5.88
∞	1.248	1.621	1.933	2.492	3.077	3.88	4.50	5.97

Πίνακας 5 : Πίνακας για τον έλεγχο Anderson-Darling

	expected value	standard deviation	95% quantile
n = 20	1.971	2.339	6.373
n = 50	2.017	2.308	6.339
n = 100	2.026	2.267	6.271
n = 250	2.012	2.174	6.129
n = 500	2.009	2.113	6.063
n = 1000	2.000	2.062	6.038
$\chi^2(2)$ distribution	2.000	2.000	5.991

Πίνακας 6 : Πίνακας για τον έλεγχο D'Agostino–Pearson Omnibus χρησιμοποιώντας την χ^2 κατανομή

Chi-Square Right-Tail Probability ($\geq \chi^2$)

DF	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Πίνακας 7 : Πίνακας για τον έλεγχο Jarque-Bera(X^2)

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική Βιβλιογραφία

1) Βόντα, Ι. - Καραγρηγορίου, Α. (2017), Εφαρμοσμένη Στατιστική Ανάλυση και Στοιχεία Πιθανοτήτων, Εκδόσεις Παρασκήνιο

2) Δαμιανού, Χ. (2006), Μεθοδολογία Δειγματοληψίας, Τεχνικές και Εφαρμογές, Εκδόσεις Σοφία

3) Φουσκάκης, Δ. , (2013), Ανάλυση Δεδομένων με χρήση της R, Εκδόσεις Τσότρας

Αγγλική Βιβλιογραφία

1) Conover, W.J. (1999), Practical Nonparametric Statistics, 3rd ed., Wiley

2) Ross, S. , (2009), First Course in Probability , 8th Edition, Pearson Prentice Hall

3) Ralph B. D'Agostino (1986). "Tests for the Normal Distribution". In D'Agostino, R.B.; Stephens, M.A. (eds.).Goodness-of-Fit Techniques. New York: Marcel Dekker.

4) Shenton, L.R. Bowman, Kimiko O. (1977). "A bivariate model for the distribution of $\sqrt{b1}$ and $b2$".Journal of the American Statistical Association.72(357): 206–211.

Ιντερνετικές Πηγές

1)https://el.wikipedia.org/wiki/%CE%9A%CE%B1%CE%BD%CE%BF%CE%BD%CE%B9%CE%BA%CE%AE_%CE%BA%CE%B1%CF%84%CE%B1%CE%BD%CE%BF%CE%BC%CE%AE#:~:text=%CE%97%20%CE%BA%CE%B1%CE%BD%CE%BF%CE%BD%CE%B9%CE%BA%CE%AE%20%CE%BA%CE%B1%CF%84%CE%B1%CE%BD%CE%BF%CE%BC%CE%AE%20(%CE%B3%CE%BD%CF%89%CF%83%CF%84%CE%AE%20%CE%BA%CE%B1%CE%B9,%CE%B3%CF%8D%CF%81%CF%89%20%CE%B1%CF%80%CF%8C%20%CE%BC%CE%B9%CE%B1%20%CE%BC%CE%AD%CF%83%CE%B7%20%CF%84%CE%B9%CE%BC%CE%AE,

2) https://en.wikipedia.org/wiki/Statistical_hypothesis_testing,

3)https://en.wikipedia.org/wiki/Glossary_of_probability_and_statistics,
4)https://www.techtarget.com/whatis/definition/normal-distribution,

5)https://www.researchgate.net/publication/348470410_Comparative_evaluation_of_goodness_of_fit_tests_for_normal_distribution_using_simulation_and_empirical_data/~

6) https://www2.ece.ohio-state.edujohnson/shapiro.pdf

7) https://en.wikipedia.org/wiki/Samuel_Sanford_Shapiro

8) https://en.wikipedia.org/wiki/Martin_Wilk

9)https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_testhttps

10) https://en.wikipedia.org/wiki/Shapiro%E2%80%93Francia_test

11) https://en.wikipedia.org/wiki/Andrey_Kolmogorov

12) https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

13) https://en.wikipedia.org/wiki/Hubert_Lilliefors

14) https://en.wikipedia.org/wiki/Lilliefors_test

15) https://en.wikipedia.org/wiki/Harald_Cram%C3%A9r

16) https://en.wikipedia.org/wiki/Richard_von_Mises

17) https://influentialpoints.com/Training/cramer_von_mises_test.htm

18) https://en.wikipedia.org/wiki/Theodore_Wilbur_Anderson

19) https://en.wikipedia.org/wiki/Donald_Allan_Darling

20) https://en.wikipedia.org/wiki/Anderson%E2%80%93Darling_test

21) https://en.wikipedia.org/wiki/Karl_Pearson

22)https://community.gooddata.com/metrics-and-maql-kb-articles-43/normality-testing-skewness-and-kurtosis-241

23) https://en.wikipedia.org/wiki/Ralph_B._D%27Agostino

24) https://www.cardiometabolichealth.org/faculty/ronald-dagostino/

25)https://real-statistics.com/tests-normality-and-symmetry/statistical-tests-normality-symmetry/dagostino-pearson-test/

26) https://en.wikipedia.org/wiki/Carlos_Jarque

27) https://en.wikipedia.org/wiki/Anil_K._Bera

28) https://en.wikipedia.org/wiki/Jarque%E2%80%93Bera_test

29)https://www.researchgate.net/publication/348470410_Comparative_evaluation_of_goodness_of_fit_tests_for_normal_distribution_using_simulation_and_empirical_data/~

30) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC374386/

31) https://en.wikipedia.org/wiki/Empirical_distribution_function

32) https://www.sciencedirect.com/topics/mathematics/empirical-distribution-function

33) https://en.wikipedia.org/wiki/Method_of_moments_(statistics)

34) https://www.sciencedirect.com/topics/mathematics/wilk-test

35) https://en.wikipedia.org/wiki/Shapiro%E2%80%93Francia_test

36) https://influentialpoints.com/Training/anderson_darling_test.htm

37) https://influentialpoints.com/Training/cramer_von_mises_test.htm

38) https://www.statisticshowto.com/lilliefors-test/

39) https://influentialpoints.com/Training/skew-and-kurtosis.htm

40) https://real-statistics.com/tests-normality-and-symmetry/statistical-tests-normality-symmetry/dagostino-pearson-test/

41) https://en.wikipedia.org/wiki/D%27Agostino%27s_K-squared_test

42) https://en.wikipedia.org/wiki/Anderson%E2%80%93Darling_test

43) https://www.researchgate.net/figure/Significant-values-for-Cramer-von-Mises-statistics-for-tests-of-the-discrete-uniform_tbl1_249612820

44)https://www.researchgate.net/figure/Skewness-and-kurtosis-for-the-data-and-fitted-models_tbl1_43033020

45) https://www.statisticshowto.com/wp-content/uploads/2016/03/table-for-Lilliefors.pdf

46) https://luk.staff.ugm.ac.id/stat/ks/Kolmogorov-SmirnovDTable.pdf

47) https://www.epa.gov/sites/default/files/2015-10/documents/monitoring_appendd_1997.pdf

48) https://en.wikipedia.org/wiki/Anderson%E2%80%93Darling_test

49) https://www.researchgate.net/figure/Normality-tests-based-on-kurtosis-and-skewness-and-DAgostino-Pearson-K-2-test-of-return_tbl4_228609219

50) https://en.wikipedia.org/wiki/Jarque%E2%80%93Bera_test

51) https://www.kaggle.com/datasets/majidarif17/weight-and-heightcsv

Η παραπάνω βιβλιογραφία αποτέλεσε το θεμέλιο της εργασίας και βοήθησε σημαντικά στην κατανόηση των ελέγχων υποθέσεων της κανονικής κατανομής και στη διαμόρφωση των συμπερασμάτων. Η ευρεία ποικιλία της βιβλιογραφίας από διάφορες πηγές συνέβαλε στην ποιοτική και ποσοτική ανάλυση του θέματος, ενισχύοντας την αξιοπιστία των αποτελεσμάτων. Η επιλογή της βιβλιογραφίας αποτέλεσε κρίσιμη διαδικασία που διαμόρφωσε το επιστημονικό υπόβαθρο της εργασίας.