Έλεγχος Υποθέσεων

Δρ. Αγγελίδης Βασίλης

Περιεχόμενα

- ✓ Η επιστήμη της Στατιστικής
- ✓ Επαγωγική Στατιστική
- ✓ Έλεγχος Υποθέσεων
- ✓ Έλεγχος Κανονικότητας
- ✓ Έλεγχος Ακραίων τιμών

Η Επιστήμη της Στατιστικής

Στατιστική είναι ένα σύνολο αρχών και μεθοδολογιών για

- Το σχεδιασμό της διαδικασίας συλλογής δεδομένων
- Την Συνοπτική και αποτελεσματική παρουσίαση των συλλεχθέντων δεδομένων
- Την ανάλυση και την εξαγωγή χρήσιμων συμπερασμάτων

(Ronald Fisher,1890 – 1962)

Η Επιστήμη της Στατιστικής

Υπάρχουν 2 **βασικές** μορφές

- **Περιγραφική Στατιστική**, η οποία ασχολείται με την περιγραφή και την παρουσίαση των δεδομένων του **δείγματος**
- **Επαγωγική Στατιστική**, η οποία ασχολείται με την εξαγωγή χρήσιμων συμπερασμάτων για τον **πληθυσμό**

Η Επιστήμη της Στατιστικής

Ξέρω στατιστική σημαίνει ότι γνωρίζω ……..

- **ποιο** στατιστικό μέτρο είναι κατάλληλο για κάθε ερευνητική ερώτηση,
- **πώς** να υπολογίσω το στατιστικό αυτό και
- **πώς** να το ερμηνεύσω

Η **επιλογή** του κατάλληλου στατιστικού μέτρου είναι ένα από τα **σημαντικότερα** βήματα στην διαδικασία της εκπαιδευτικής έρευνας και της στατιστικής ανάλυσης

Η Επιστήμη της Στατιστικής

Για την επιλογή του κατάλληλου στατιστικού χρησιμοποιούμε δύο γενικά κριτήρια:

- Το **λόγο** (σκοπό) για τον οποίο χρειαζόμαστε το στατιστικό
  - ✓ η περιγραφή μεταβλητών ή σχέσεων μεταξύ μεταβλητών, με τα οποία ασχολείται η Περιγραφική Στατιστική και
  - ✓ η γενίκευση από το δείγμα στον πληθυσμό, με την οποία ασχολείται η επαγωγική
- Την **κλίμακα** (επίπεδο) μέτρησης των μεταβλητών
  - ✓ Ονομαστική, Τακτική, Ισοδιαστημική, Αναλογική

Επαγωγική Στατιστική

- Ίσως το σπουδαιότερο εργαλείο της Στατιστικής επιστήμης.
- Εξαγωγή συμπερασμάτων για τις τιμές των παραμέτρων του πληθυσμού από το **τυχαίο** δείγμα που έχουμε λάβει από τον πληθυσμό.
- Στατιστική μεθοδολογία με την οποία απορρίπτουμε ή δεν απορρίπτουμε μια στατιστική υπόθεση.

Επαγωγική Στατιστική

Βασικές Έννοιες

- **Ερευνητική Υπόθεση**, μια εικασία που χρειάζεται **μαθηματική** επαλήθευση.
- **Έλεγχος Ερευνητικής Υπόθεσης**, μία στατιστική συμπερασματική – επαγωγική διαδικασία που μας επιτρέπει να αξιοποιήσουμε τα δεδομένα του δείγματος για να εκτιμήσουμε την εγκυρότητα – ορθότητα μιας εικασίας που έγινε για τον πληθυσμό
- **Στατιστική Υπόθεση**, μια οποιαδήποτε στατιστική δήλωση (για κατανομές πληθυσμών, στοχαστικές διαδικασίες, κλπ) που θέτουμε υπό έλεγχο με βάση τις παρατηρήσεις

Έλεγχος Υποθέσεων

- Ο στατιστικός έλεγχος μιας υπόθεσης θα μπορούσε να **προσομοιωθεί** με τη διαδικασία λήψης απόφασης σε μια δικαστική διαδικασία.
- Ο κατηγορούμενος προσάγεται στο δικαστήριο για να δικαστεί με μια συγκεκριμένη διαδικασία. Στην πραγματικότητα, είναι είτε **αθώος** είτε **ένοχος**. Οι ένορκοι όμως δεν το γνωρίζουν και καλούνται να αποφασίσουν.
- Η απόφασή τους θα ληφθεί με βάση τα αποδεικτικά στοιχεία που θα παρουσιαστούν στη διάρκεια της δίκης. Μετά την ολοκλήρωση της ακροαματικής διαδικασίας, οι ένορκοι θα πρέπει να **αποφασίσουν** αν θα δεχθούν την **αθώωση** του κατηγορουμένου ή θα προτείνουν στο δικαστήριο την **ενοχή** του

Έλεγχος Υποθέσεων

ΑΠΟΦΑΣΗ	ΠΡΑΓΜΑΤΙΚΗ ΕΥΘΥΝΗ ΚΑΤΗΓΟΡΟΥΜΕΝΟΥ	
ΕΝΟΡΚΩΝ	Αθώος	Ένοχος
Αθώος	☑ Σωστή απόφαση	☒ Λανθασμένη απόφαση
Ένοχος	☒ Λανθασμένη απόφαση	☑ Σωστή απόφαση

- Καλό θα είναι σε κάθε δίκη οι ένορκοι να παίρνουν τη **σωστή** απόφαση. Αυτό δεν είναι πάντα εφικτό (π.χ. ελλιπή αποδεικτικά στοιχεία, πλάνη, προσωπικοί λόγοι, κ.τ.λ.).
- Γενικά, είναι **αδύνατο** να **μηδενίσουμε** την πιθανότητα της μιας ή της άλλης λανθασμένης απόφασης

Έλεγχος Υποθέσεων

ΑΠΟΦΑΣΗ	ΠΡΑΓΜΑΤΙΚΗ ΕΥΘΥΝΗ ΚΑΤΗΓΟΡΟΥΜΕΝΟΥ	
ΕΝΟΡΚΩΝ	Αθώος	Ένοχος
Αθώος	☑ Σωστή απόφαση	☒ Λανθασμένη απόφαση
Ένοχος	☒ Λανθασμένη απόφαση	☑ Σωστή απόφαση

- οι ένορκοι **προσπαθούν** να φθάσουν κάθε φορά στην ετυμηγορία τους, γνωρίζοντας ότι τόσο στην περίπτωση της αθώωσης όσο και στην περίπτωση της ενοχής υπάρχει κάποια πιθανότητα σφάλματος.
- Η μεθοδολογία που ακολουθείται στο στατιστικό έλεγχο μιας υπόθεσης επιδιώκει ακριβώς την **ελαχιστοποίηση** της πιθανότητας μιας λανθασμένης απόφασης προς τη μια ή την άλλη κατεύθυνση.

Έλεγχος Υποθέσεων

Ο έλεγχος **υποθέσεων** είναι η **διαδικασία** προσδιορισμού αν μια δεδομένη υπόθεση **ισχύει** ή όχι.

- Το πρώτο βήμα στον έλεγχο υποθέσεων είναι να οριστεί η **μηδενική υπόθεση**.
- Η υπόθεση **ελέγχεται** με χρήση της **στατιστικής**.

Η **μηδενική υπόθεση** είναι ένας ισχυρισμός σχετικά με την τιμή μιας πληθυσμιακής παραμέτρου.

Είναι ένας ισχυρισμός ο οποίος θεωρείται σωστός εκτός και εάν υπάρχουν επαρκή στατιστικά στοιχεία για να υποστηριχθεί το αντίθετο συμπέρασμα.

Έλεγχος Υποθέσεων

Στον έλεγχο υποθέσεων υπάρχουν:

- Η **μηδενική** υπόθεση $\mathbf{H_0}$
- Η **ενναλακτική** υπόθεση $\mathbf{H_1}$

Η **εναλλακτική** υπόθεση είναι το αντίθετο της μηδενικής υπόθεσης.

Επειδή υποστηρίζουν αντίθετες υποθέσεις, μόνο 1 από τις 2 θα είναι σωστή. Η απόρριψη της μιας υπόθεσης σημαίνει αποδοχή της άλλης.

Παράδειγμα :

- Μηδενική υπόθεση: $\mathbf{H_0}$: $\mu=100$
- Εναλλακτική υπόθεση: $\mathbf{H_1}$: $\mu\neq100$

Έλεγχος Υποθέσεων

Προφανώς, για να ελεγχθεί μια υπόθεση με απόλυτη ακρίβεια, πρέπει να ελεγχθεί όλος ο πληθυσμός.

Αυτό όμως είναι δύσκολο, οπότε επιλέγεται ένα ικανοποιητικό τυχαίο δείγμα, και εξάγονται συμπεράσματα με βάση αυτό.

Έλεγχος Υποθέσεων

	Αποδοχή υπόθεσης H_0 από το δείγμα	Απόρριψη υπόθεσης από H_0 το δείγμα
Υπόθεση H_0 αληθής στον πληθυσμό	✓	Σφάλμα τύπου I
Υπόθεση H_0 ψευδής στον πληθυσμό	Σφάλμα τύπου II	✓

Σε κάθε στατιστικό έλεγχο υποθέσεων υπάρχει επομένως η δυνατότητα σφάλματος:

- **Σφάλμα τύπου I**: Απόρριψη της H_0 ενώ στην πραγματικότητα είναι αληθής.
- **Σφάλμα τύπου II**: Αποδοχή της H_0 ενώ στην πραγματικότητα είναι ψευδής.

Έλεγχος Υποθέσεων

Βασικές έννοιες

- ✓ **α=P**(σφάλμα τύπου Ι) = **P**(Απόρριψη της H_0 ενώ στην πραγματικότητα είναι αληθής)
- ✓ **β=P**(σφάλμα τύπου ΙΙ) = **P**(Αποδοχή της H_0 ενώ στην πραγματικότητα η H_1 είναι αληθής)
- ✓ Η πιθανότητα **γ=1-β** ονομάζεται **ισχύς** του ελέγχου και εκφράζει το ποσοστό σωστών απορρίψεων της H_0
- ✓ Το **α** ονομάζεται **επίπεδο σημαντικότητας** (π.χ. αν έχουμε επιλέξει α=0,05 και απορρίψουμε την μηδενική υπόθεση H_0 σημαίνει ότι σε 100 όμοιες περιπτώσεις είναι δυνατό να έχουμε κάνει λάθος και να έχουμε απορρίψει την H_0 ενώ είναι αληθής μόνο σε 5).

Έλεγχος Υποθέσεων

	Αποδοχή υπόθεσης H_0 από το δείγμα	Απόρριψη υπόθεσης από H_0 το δείγμα
Υπόθεση H_0 **αληθής** στον πληθυσμό	Ορθή Απόφαση Πιθανότητα = 1-α	**Σφάλμα τύπου I** Πιθανότητα = α
Υπόθεση H_0 **ψευδής** στον πληθυσμό	**Σφάλμα τύπου II** Πιθανότητα = β	Ορθή Απόφαση Πιθανότητα = 1-β

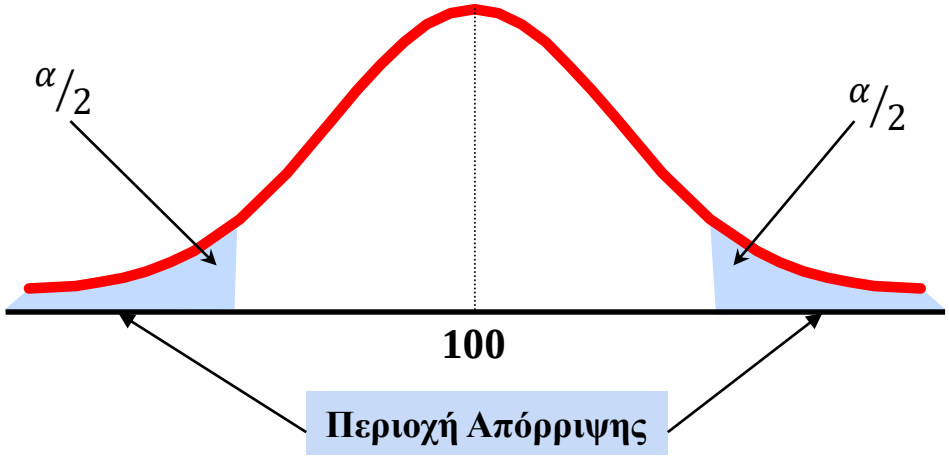
Τα σφάλματα επομένως μπορούν να ταξινομηθούν σε:

- ✓ **Σφάλμα τύπου I**: Απόρριψη της H_0 ενώ στην πραγματικότητα είναι αληθής, με πιθανότητα α
- ✓ **Σφάλμα τύπου II**: Αποδοχή της H_0 ενώ στην πραγματικότητα H_1 αληθής, με πιθανότητα β

Έλεγχος Υποθέσεων

Βασικές έννοιες

- ✓ Το **(1-α)** ονομάζεται και **συντελεστής εμπιστοσύνης** και είναι η πιθανότητα μη απόρριψής της H_0 όταν είναι αληθής
- ✓ Το **(1-α)*100%** ονομάζεται **επίπεδο εμπιστοσύνης** του ελέγχου



Έλεγχος Υποθέσεων

Βασικές έννοιες

- ✓ Η τιμή του α (άλφα) **επηρεάζει**
  - ❖ Τόσο την πιθανότητα σφάλματος τύπου I (όσο αυξάνεται το α τόσο **αυξάνεται** η πιθανότητα σφάλματος τύπου I)
  - ❖ όσο και την πιθανότητα σφάλματος τύπου II (όσο αυξάνεται το α τόσο **μειώνεται** η πιθανότητα σφάλματος τύπου II)
- ✓ Η τιμή του α, επιλέγεται ανάλογα με τις επιπτώσεις/κόστος του κάθε σφάλματος

Παράδειγμα επιλογής **α**

Αν το σφάλμα **τύπου ΙΙ** (δηλαδή αποδοχή λανθασμένης υπόθεσης) είναι πολύ **σημαντικό**, π.χ. γιατί θα προκαλέσει δυσφήμιση στην εταιρία μου να παραχθούν προβληματικά προϊόντα, τότε επιλέγω **μεγαλύτερο α** π.χ. στο 10%, άρα 90% διάστημα εμπιστοσύνης.

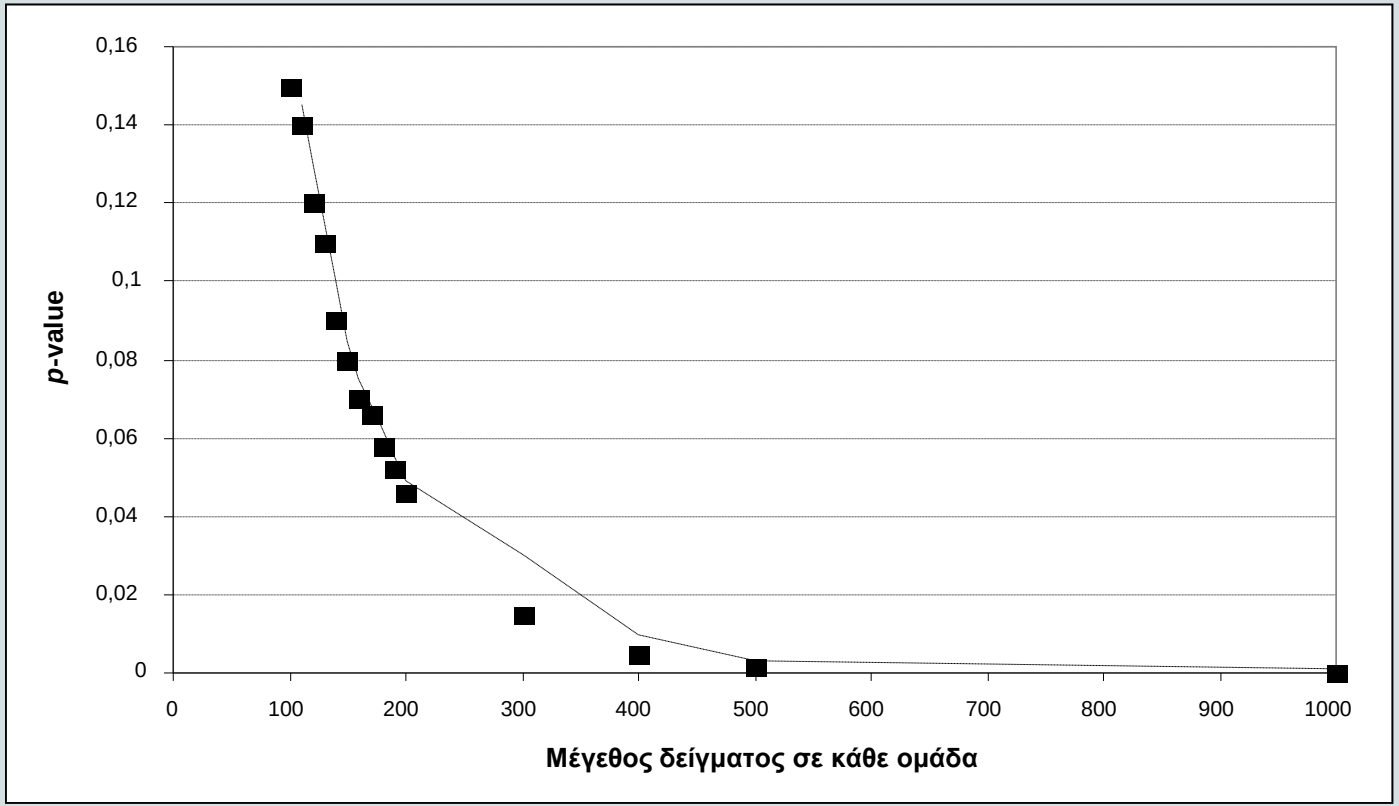
Αν όμως ένα σφάλμα **τύπου ΙΙ** δεν έχει **ιδιαίτερο κόστος**, ενώ ένα σφάλμα τύπου Ι οδηγεί στην απόρριψη μιας καλής παραγωγής προϊόντων, τότε επιλέγω **μικρότερο α** π.χ. στο 1%, άρα 99% διάστημα εμπιστοσύνης.

Έλεγχος Υποθέσεων

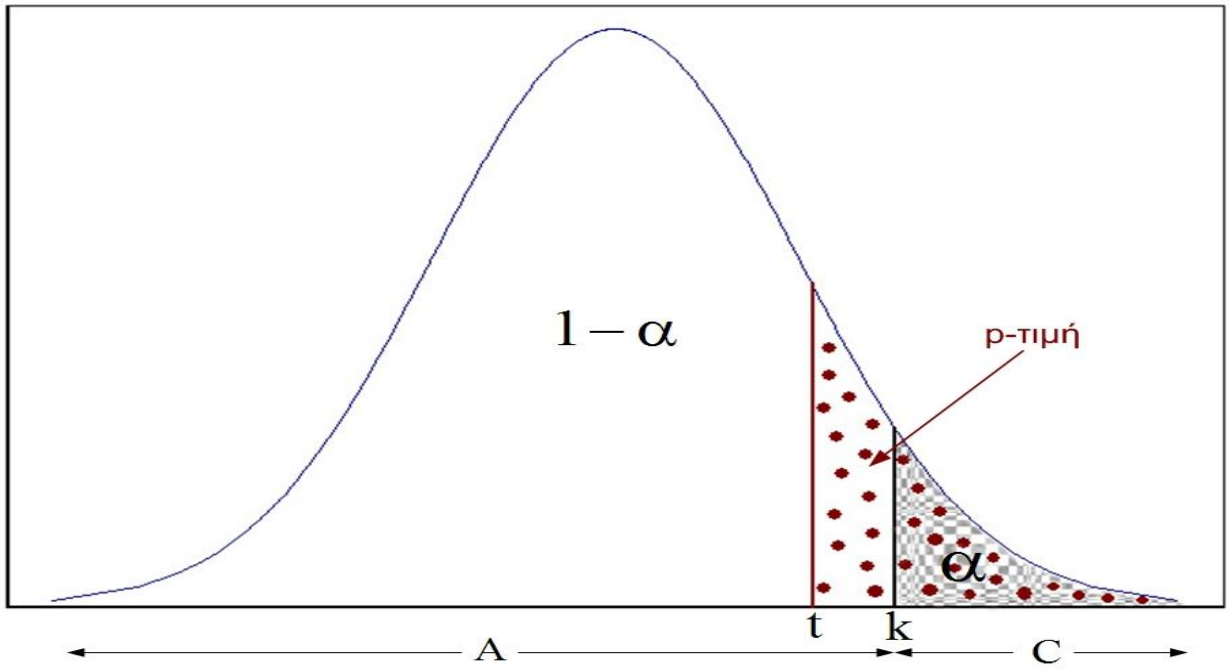
Η τιμή p-value

- ✓ Η τιμή **p-value** είναι το κριτήριο αποδοχής ή όχι της μηδενικής υπόθεσης H_0
- ✓ Πιο συγκεκριμένα απορρίπτουμε την μηδενική υπόθεση H_0 όταν η τιμή **p-value** είναι μικρότερη από το επίπεδο στατιστικής σημαντικότητας **α** (άλφα) που έχουμε δηλώσει.
- ✓ Η τιμή **p-value** δεν είναι η πιθανότητα να επαληθευθεί η μηδενική υπόθεση H_0 και αυτό γιατί οι υποθέσεις δεν εκφράζονται με πιθανότητες στην στατιστική
- ✓ Η τιμή **p-value** επηρεάζεται ισχυρά από το μέγεθος του δείγματος πιο συγκεκριμένα υπάρχει **αντίστροφη** συσχέτιση μεταξύ του μεγέθους του δείγματος και της τιμής **p-value**

P value και **μέγεθος** δείγματος για μια δεδομένη συσχέτιση



P value και διάστημα εμπιστοσύνης



Έλεγχος Υποθέσεων

Υποθέσεις

"*Αν η Γιαγιά μου είχε καρούλια … θα ήταν πατίνι*"

- Κάθε στατιστικό τεστ βασίζεται σε ένα σύνολο υποθέσεων (κριτηρίων)
- Αν οι υποθέσεις δε ισχύουν, το αποτέλεσμα του ελέγχου μπορεί να είναι λανθασμένο
- Πολύ συχνά δεν γίνεται σωστά ο έλεγχος υποθέσεων
- Ένας πολύ σημαντικός έλεγχος στην στατιστική ανάλυση είναι να δούμε αν μπορούμε να χρησιμοποιήσουμε **παραμετρικά** τεστ (αν τα δεδομένα ακολουθούν την κανονική κατανομή)
  - ✓ Τα παραμετρικά τεστ εμφανίζονται πολύ συχνά στην βιβλιογραφία
  - ✓ Είναι πιο ισχυρά και έχουν καλύτερη αντιμετώπιση από τους reviewers

Έλεγχος Υποθέσεων

Βήματα στον έλεγχο Υποθέσεων

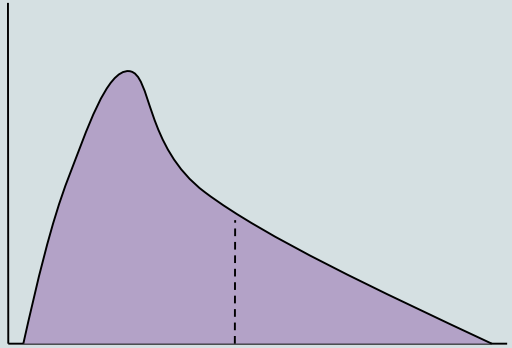
1. Διατυπώστε την **μηδενική** υπόθεση $\mathbf{H_0}$ και την **εναλλακτική** υπόθεση $\mathbf{H_1}$
2. Επιλέξτε το **επίπεδο στατιστικής** σημαντικότητας **α** και το **μέγεθος** του δείγματος n λαμβάνοντας υπόψη την σχετική σημασία των **σφαλμάτων** τύπου I και τύπου II
3. Προσδιορίστε την **κατάλληλή** στατιστική συνάρτηση ελέγχου (ποιο στατιστικό μέτρο θα χρησιμοποιηθεί)
4. Συλλέξτε τα δεδομένα και υπολογίστε την τιμή της στατιστικής συνάρτησης ελέγχου (**p value**)
5. Πάρτε την στατιστική **απόφαση** (αν p value < α απορρίπτεται η μηδενική υπόθεση $\mathbf{H_0}$) και **διατυπώστε** το διοικητικό συμπέρασμα

Δείκτες Κεντρικής Θέσης

Αριθμητικός Μέσος

Ο ***αριθμητικός μέσος***, ή αλλιώς *μέσος όρος*, ή πιο σύντομα απλά *μέσος*, είναι το πιο **γνωστό** και πιο χρήσιμο μέτρο κεντρικής θέσης. **Υπολογίζεται** αθροίζοντας όλες τις τιμές των δεδομένων και διαιρώντας δια το πλήθος τους:

$$\overline{X} = \frac{\chi_1 + \chi_2 + \ldots\ldots + \chi_\nu}{\nu} = \frac{\sum_{i=1}^{\nu} X_i}{\nu}$$

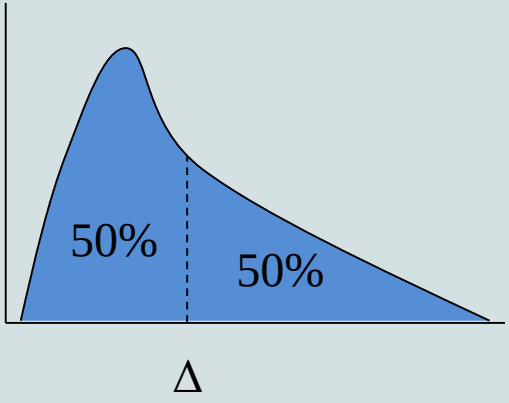


- η μέση τιμή **επηρεάζεται** ιδιαίτερα από τις **ακραίες τιμές** (μεγάλες ή μικρές). Αυτό δημιουργεί προβλήματα σε μη συμμετρικές κατανομές.
- Κάποιες φορές δεν έχει φυσικό νόημα

Δείκτες Κεντρικής Θέσης

Διάμεσος

Δηλαδή η **διάμεσος** είναι μία τιμή η οποία χωρίζει τις παρατηρήσεις του δείγματος σε δύο ισοπληθείς ομάδες, έτσι ώστε οι παρατηρήσεις της πρώτης ομάδας να είναι όλες μεγαλύτερες ή ίσες της διαμέσου και όλες οι παρατηρήσεις της άλλης ομάδας να είναι όλες μικρότερες ή ίσες αυτής. Την συμβολίζουμε με **δ**

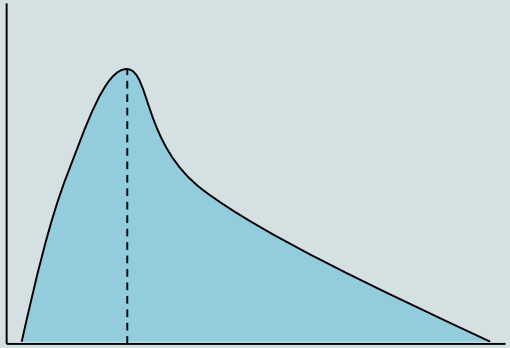


Χωρίζει το εμβαδόν κάτω από την καμπύλη της κατανομής σε δύο ίσα μέρη (50% - 50%)

Δείκτες Κεντρικής Θέσης

Επικρατούσα Τιμή

Η ***επικρατούσα τιμή*** ενός συνόλου δεδομένων είναι η τιμή που εμφανίζεται με τη **μεγαλύτερη** *συχνότητα*.

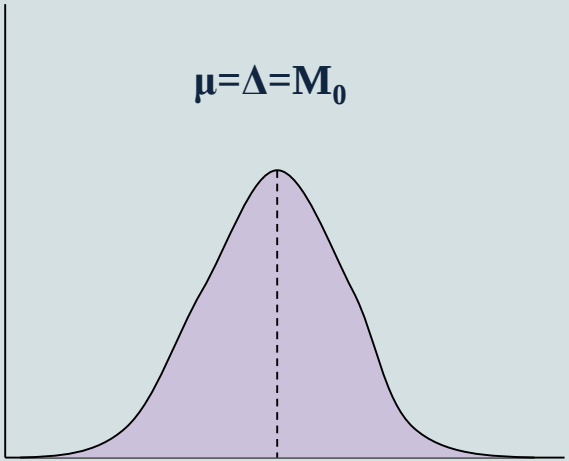


Είναι ο αριθμός x, που εμφανίζεται με τη μεγαλύτερη συχνότητα

Κατανομές

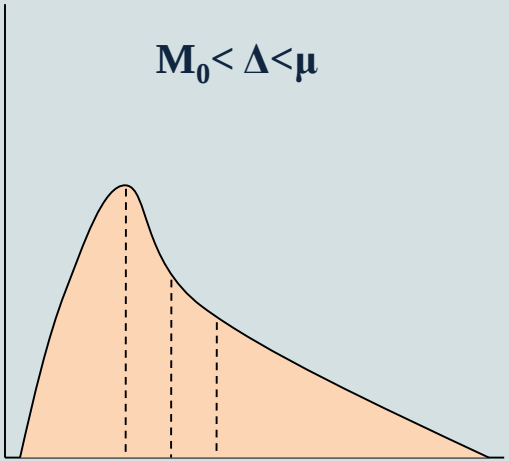
Χαρακτηριστικά Κατανομής - συμμετρικές και μη συμμετρικές κατανομές

$\mu=\Delta=M_0$



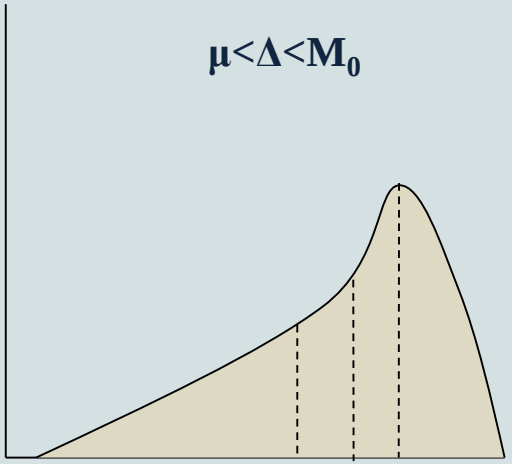
Πολλές τιμές στη μέση, λίγες μεγάλες τιμές και λίγες μικρές τιμές

$M_0<\Delta<\mu$



Πολλές μικρές τιμές, κάποιες τιμές στη μέση και λίγες μεγάλες τιμές

$\mu<\Delta<M_0$

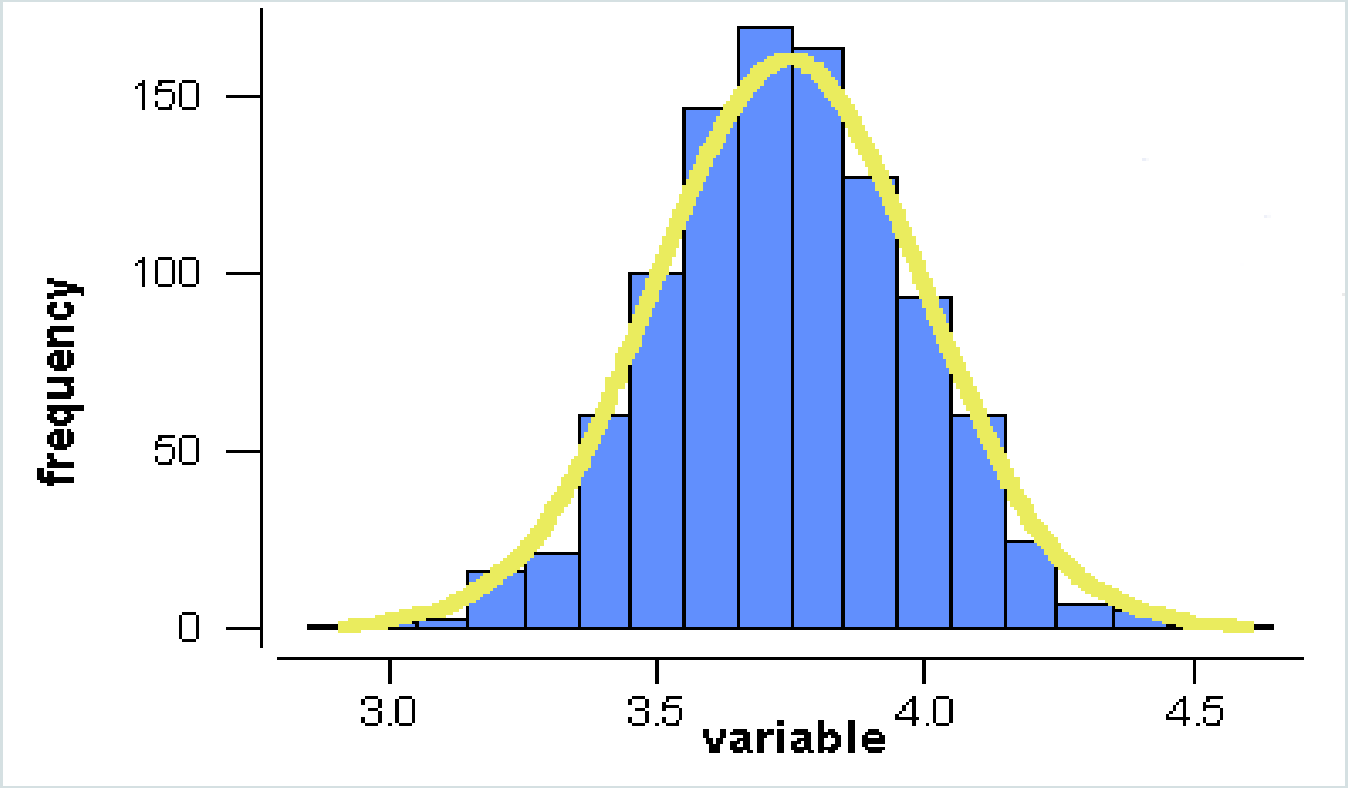


Πολλές μεγάλες τιμές, κάποιες τιμές στη μέση και λίγες μικρές τιμές

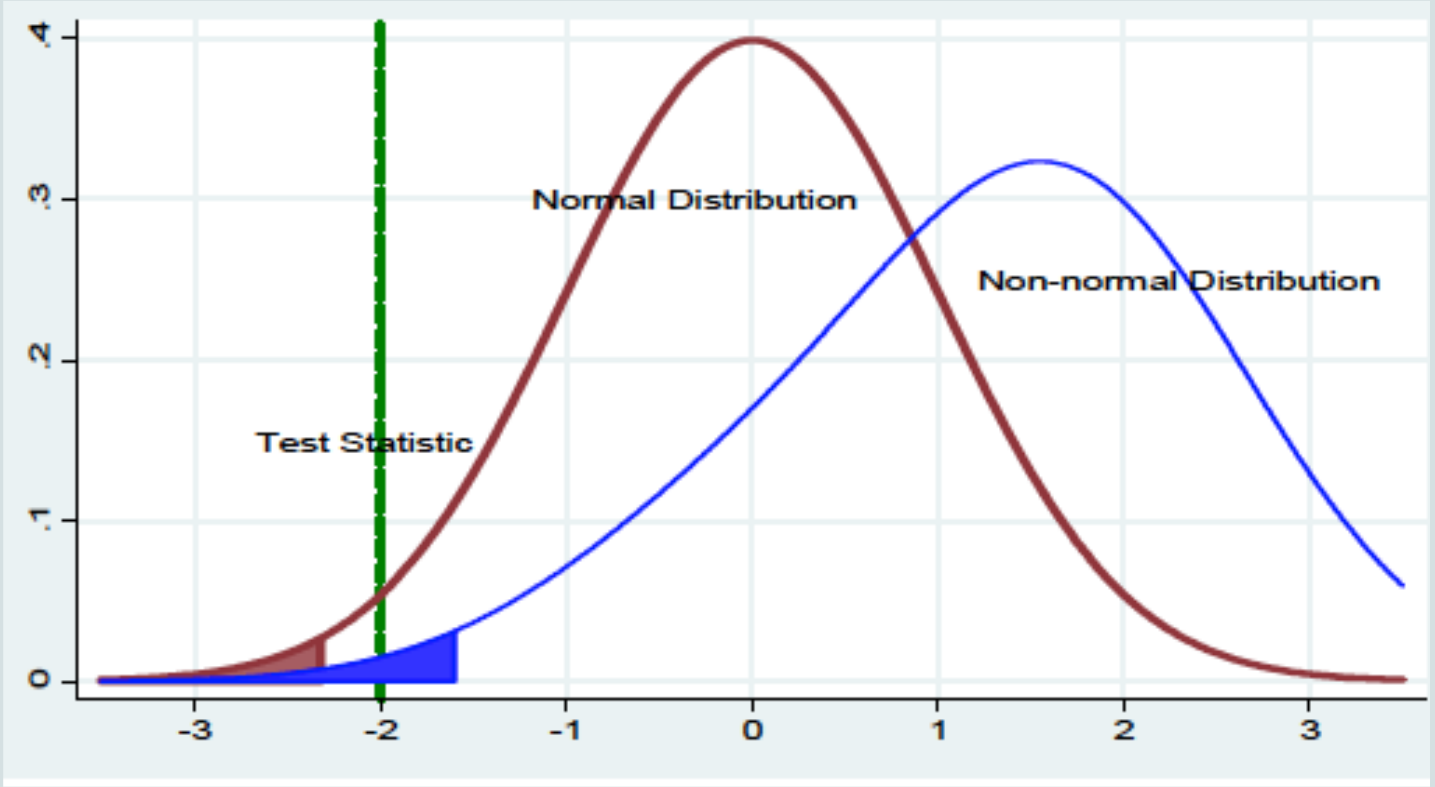
Κανονική Κατανομή

- Η **υπόθεση** της κανονικότητας είναι μία από τις υποθέσεις πάνω στις οποίες έχει θεμελιωθεί η στατιστική συμπερασματολογία.
- Οι περισσότερες από τις μεθοδολογίες της **Παραμετρικής** Στατιστικής υποθέτουν, προϋποθέτουν ότι τα δεδομένα προέρχονται από έναν πληθυσμό, ο οποίος περιγράφεται ικανοποιητικά από την κανονική κατανομή.
- Όταν το ιστόγραμμα συχνοτήτων των ποσοτικών μεταβλητών έχει το σχήμα "**καμπάνας**", τότε λέμε ότι τα δεδομένα ακολουθούν την **κανονική κατανομή** ή κατανέμονται κανονικά.
- Το ιστόγραμμα όμως δεν είναι "**ικανό**" να μας απαντήσει στη ερώτηση αν είναι κανονικά τα δεδομένα ή αν προέρχονται από μία κανονική κατανομή με ένα μέσο και μία διακύμανση.

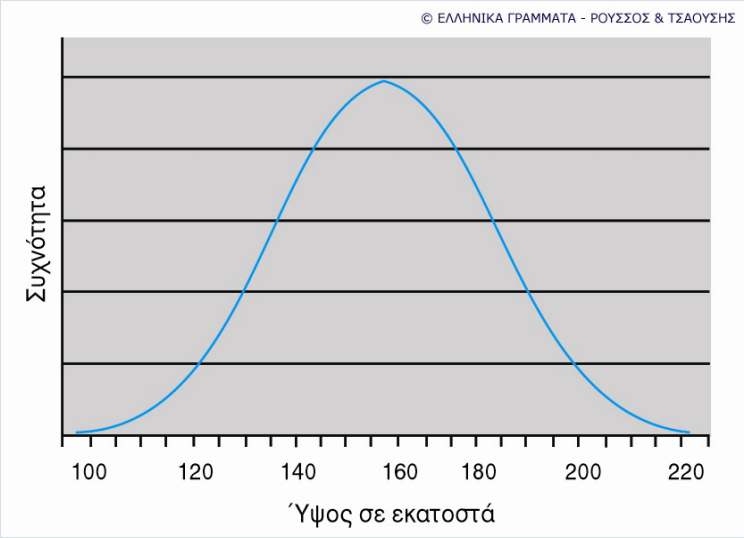
Κανονική Κατανομή



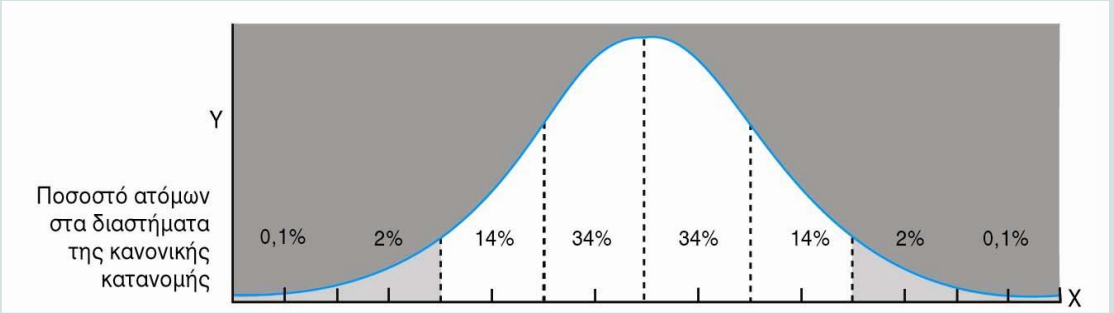
Κανονική Κατανομή



Κανονική Κατανομή



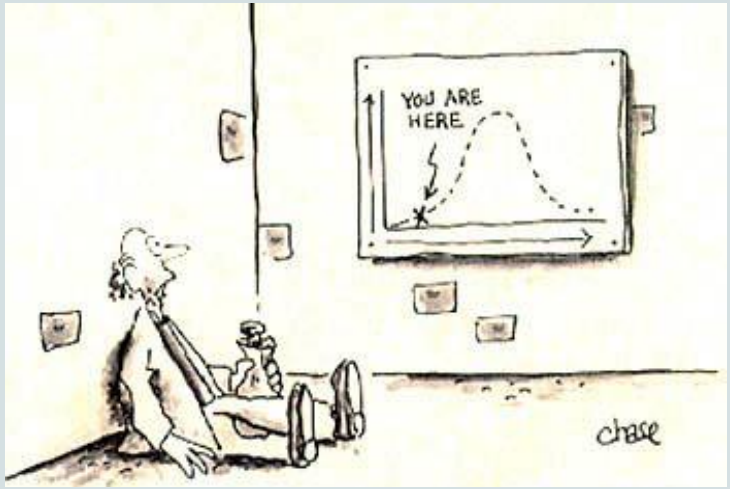
Ιδιότητες Κανονικής Κατανομής



- ✓ Η **επικρατούσα** τιμή, η **διάμεσος** και ο **μέσος όρος** συμπίπτουν
- ✓ **Συμμετρική** ως προς το μέσο μ. (το 50% βρίσκεται δεξιά του μ και το 50% αριστερά του μ)
- ✓ **Συνολικό** εμβαδό κάτω από την καμπύλη f(x) = 1

Κανονική Κατανομή

Ιδιότητες Κανονικής Κατανομής



- ✓ Το σχήμα της κανονικής κατανομής έχει τις εξής ιδιότητες:
  - Το πιο **απότομο** σημείο της καμπύλης βρίσκεται σε απόσταση μιας τυπικής απόκλισης εκατέρωθεν του μέσου όρου
  - Σε **απόσταση** 3 τυπικών αποκλίσεων από το μέσο όρο η κλίση είναι σχεδόν οριζόντια, πολύ κοντά στο μηδέν
  - Παρουσία **ακραίων τιμών** μπορεί να γείρει την καμπάνα δεξιά ή αριστερά **παραβιάζοντας** το κριτήριο της κανονικής κατανομής

Κανονική Κατανομή

- Ο έλεγχος ότι τα τυχαία δεδομένα ακολουθούν μια συγκεκριμένη κατανομή ονομάζεται «***έλεγχος καλής προσαρμογής***».
- Για τον έλεγχο αν τα δεδομένα ακολουθούν την κανονική κατανομή αρχικά μπορούμε να κατασκευάσουμε δύο **γραφήματα** με το SPSS, το P-P Plot και το Q-Q Plot
- Με αυτά τα γραφήματα ελέγχουμε **οπτικά** την ύπαρξη κανονικότητας στα δεδομένα. Όσο πιο κοντά στην ευθεία είναι τα σημεία του σχήματος τόσο πιο πολλές είναι οι ενδείξεις ότι τα δεδομένα ακολουθούν την κανονική κατανομή.
- Το μάτι όμως πάλι μπορεί να "**πέσει έξω**" και να ξεγελαστούμε. Για αυτό το λόγο καταφεύγουμε σε **τεστ κανονικότητας** για να απαντήσουμε στην προηγούμενη ερώτηση.

Κανονική Κατανομή

Για τον έλεγχο της **Κανονικής** κατανομής έχουμε τις υποθέσεις :

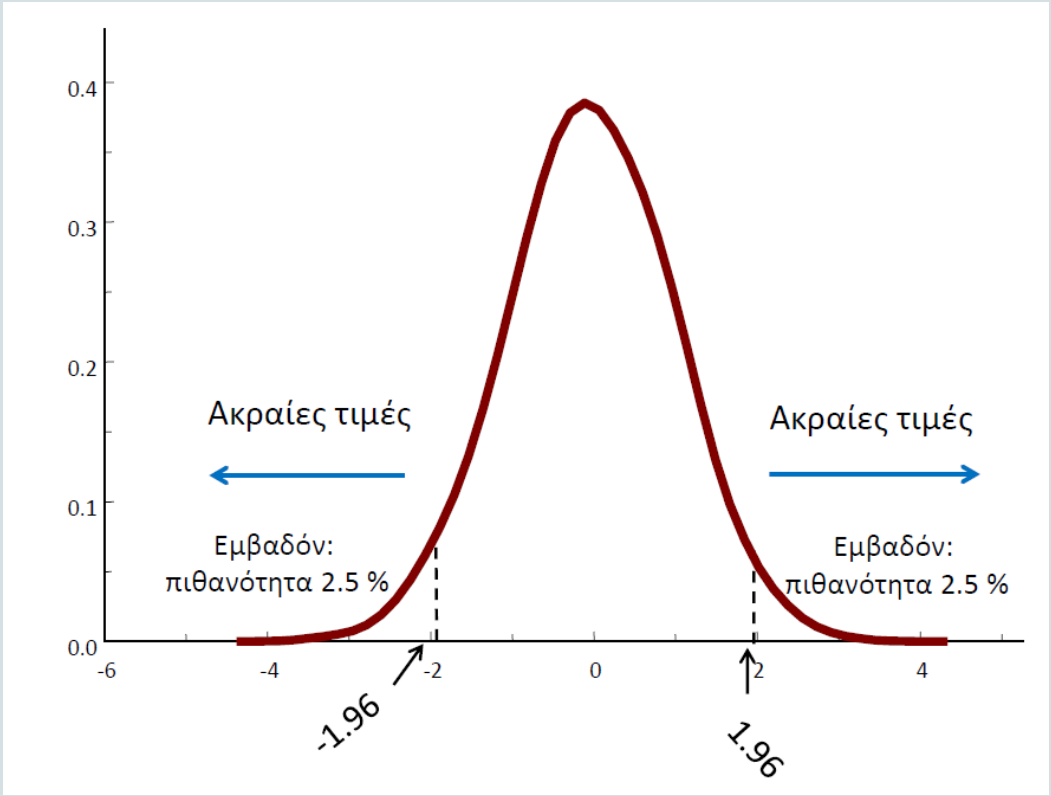
H_0: Η κατανομή των δεδομένων δε διαφέρει από την κανονική κατανομή

H_1: Η κατανομή των δεδομένων διαφέρει από την κανονική κατανομή

Για τον έλεγχο της υπόθεσης συγκρίνουμε την τιμή **p-value** με το επίπεδο στατιστικής σημαντικότητας **α** (άλφα) που έχουμε ορίσει (π.χ α=0,05)

- Αν η **p-value** είναι μικρότερη του 0,05, τότε λέμε ότι η μηδενική υπόθεση απορρίπτεται.
- Αν η **p-value** είναι μεγαλύτερη ή ίση του 0,05, τότε λέμε ότι η μηδενική υπόθεση δεν απορρίπτεται.

Κανονική Κατανομή



Αν p-value > 0.05 **ΔΕΝ** συμπεραίνουμε ότι τα δεδομένα του δείγματος ακολουθούν την κανονική κατανομή, αλλά ότι σε επίπεδο σημαντικότητας α = 0.05 δεν διαπιστώνονται στατιστικά σημαντικές αποκλίσεις από την κανονικότητα.

Έλεγχος Κανονικής Κατανομής

Εντοπισμός Ακραίων Τιμών

Όπως αναφέρθηκε οι ακραίες μπορούν να παραβιάσουν την κανονικότητα επομένως πριν τον έλεγχο κανονικότητας πρέπει να **εντοπίσουμε** και να **απομακρύνουμε** τυχόν ακραίες τιμές.

Για τον εντοπισμό των ακραίων τιμών μπορούμε να χρησιμοποιήσουμε:

- ✓ τα **z–scores** (πόσες τυπικές αποκλίσεις απέχει η τιμή από την μέση τιμή του δείγματος) και
- ✓ τα **θηκογράμματα**

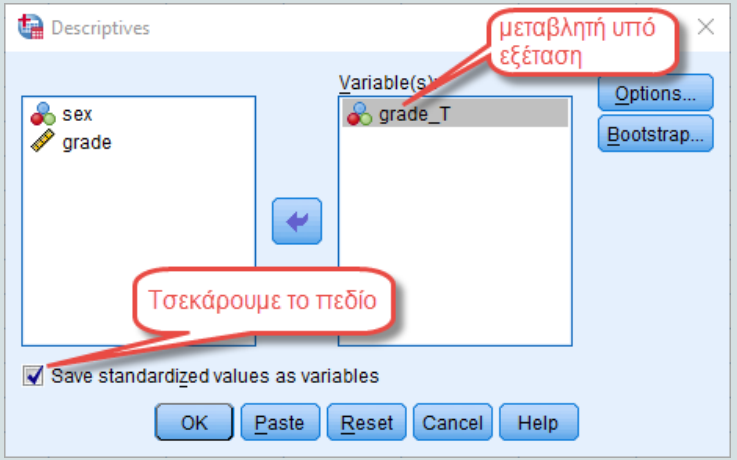
Παράδειγμα Ι

Έλεγχος Ακραίων Τιμών

Χρησιμοποιούμε το αρχείο Lecture_1_1.sav το οποίο περιέχει τους βαθμούς των φοιτητών στο μάθημα της Ανάλυσης Δεδομένων.

Μέθοδος 1η :

- Από το μενού επιλέγουμε Analyze → Descriptive Statistics → Descriptive
- Στο παράθυρο διαλόγου που εμφανίζεται βάζουμε την μεταβλητή που θέλουμε να εξετάσουμε στο πλαίσιο "Variables" και τσεκάρουμε το πεδίο "save standardized values as variables"



Παράδειγμα Ι

Έλεγχος Ακραίων Τιμών

- Η μεταβλητή που προστίθεται στο φύλλο δεδομένων είναι τα **z-scores**
- Τα z-scores σε απόλυτη τιμή παρουσιάζουν τον αριθμό των τυπικών αποκλίσεων μεταξύ των δεδομένων και της μέσης τιμής
- Τιμές z-scores μεγαλύτερες σε απόλυτη τιμή του **τρία** (3) υποδηλώνουν ακραίες τιμές αν και ορισμένοι ερευνητές προτείνουν το **1,96** (α=5%)
- Υπάρχουν ερευνητές που εκτιμούν ότι τα z-scores εξαρτούνται από το μέγεθος του δείγματος και η τιμή για τον εντοπισμό των ακραίων τιμών πρέπει να υπολογιστεί από τον τύπο $\frac{(n-1)}{\sqrt{n}}$ όπου n ο αριθμός των παρατηρήσεων
- Στο παράδειγμα υπό εξέταση η τιμή στην παρατήρηση **12** είναι πιθανόν **ακραία** τιμή

Zgrade_T
,79801
,52331
,24861
-,54117
-1,09057
-,12911
1,00404
-,71286
,38596
-,19779
,74719
3,37336
-,09477
,45463
,72934
-,36948
-,81587
,17993
-,12911
,45463
-,30080
-,54117
,17993
-,30080
-2,36108

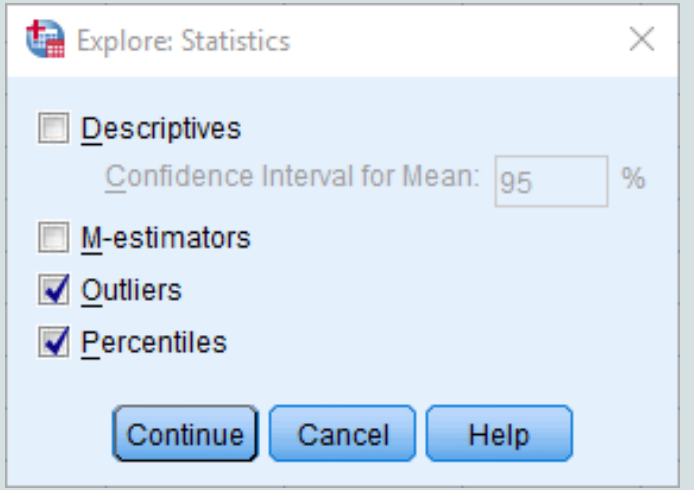
Ακραία

Παράδειγμα Ι

Έλεγχος Ακραίων Τιμών

Μέθοδος 2η :

- Από το μενού επιλέγουμε Analyze → Descriptive Statistics → Explore
- Στο παράθυρο διαλόγου που εμφανίζεται βάζουμε την μεταβλητή που θέλουμε να εξετάσουμε στο πλαίσιο "Dependent List"
- Από την επιλογή "Statistics" επιλέγουμε Outliers και Percentiles



Παράδειγμα Ι

Έλεγχος Ακραίων Τιμών

Από το αποτελέσματα μελετούμε τους Πίνακες Percentiles και Extreme Values

Extreme Values

			Case Number	Value
grade_T	Highest	1	12	168,00
		2	7	99,00
		3	1	93,00
		4	15	91,00
		5	2	85,00
	Lowest	1	25	1,00
		2	5	38,00
		3	17	46,00
		4	11	48,00
		5	8	49,00

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	grade_T	12,1000	42,8000	54,0000	66,0000	83,0000	95,4000	147,3000
Tukey's Hinges	grade_T			54,0000	66,0000	83,0000		

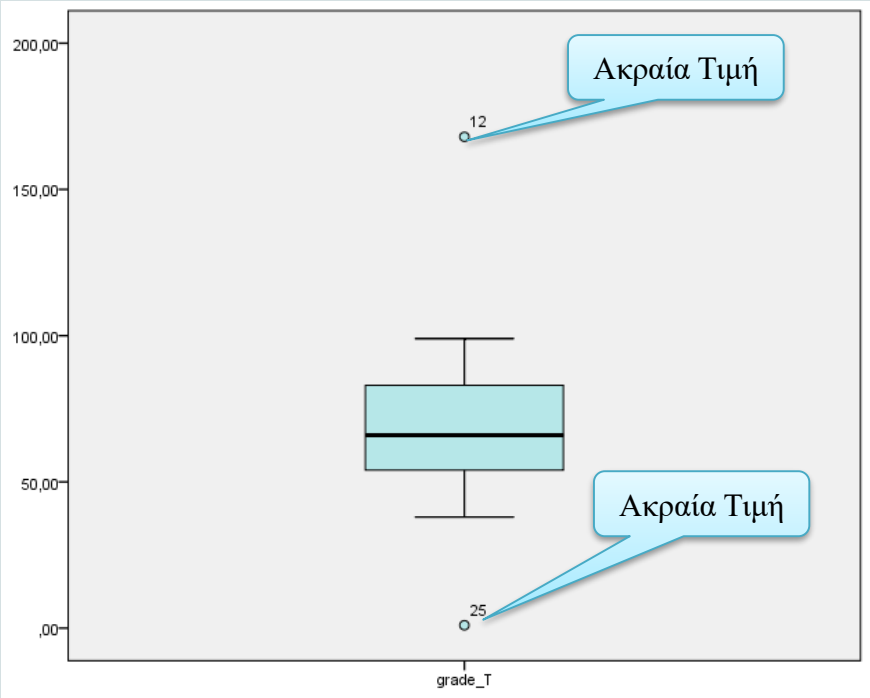
Παρατηρούμε ότι τιμές μικρότερες του **12,1** και μεγαλύτερες του **147,3** είναι πιθανές **ακραίες** τιμές (σειρά 12 και σειρά 25)

Παράδειγμα Ι

Έλεγχος Ακραίων Τιμών

Από το θηκόγραμμα **παρατηρούμε** 2 **ακραίες τιμές** (outliers)

- ✓ τιμές πέρα από τα whiskers, επισημαίνονται με «**o**» και είναι **ακραίες** (outliers), ενώ με * επισημαίνονται οι **έκτροπες** (extreme)
- ✓ πιθανές **αποκλίσεις** από την κανονική κατανομή (αν η διάμεσος είναι πιο κοντά στην κορυφή ή στην αρχή του κουτιού και όχι στο κέντρο).



Παράδειγμα I

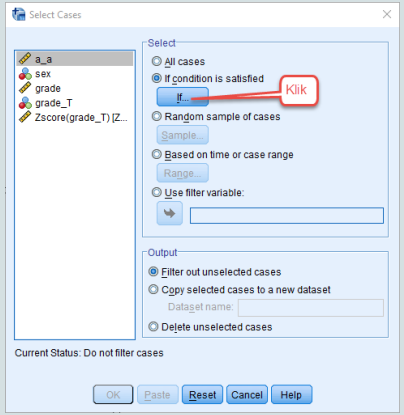
Απομάκρυνση Ακραίων Τιμών

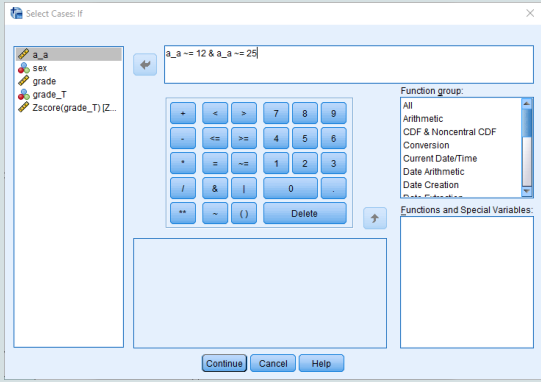
Από το μενού επιλέγουμε Data → Select Cases

Από το παράθυρο διαλόγου επιλέγουμε την επιλογή "If condition is satisfied"

Χρησιμοποιώντας τα πλήκτρα γράφουμε a_a ~= 12 & a_a ~= 25 το οποίο εξαιρεί από τον υπολογισμό την **12** και την **25** παρατήρηση

Στην συνέχεια πατάμε το πλήκτρο continue και το πλήκτρο Ok και εκτελούμε ξανά την ανάλυση χωρίς τις **ακραίες τιμές**

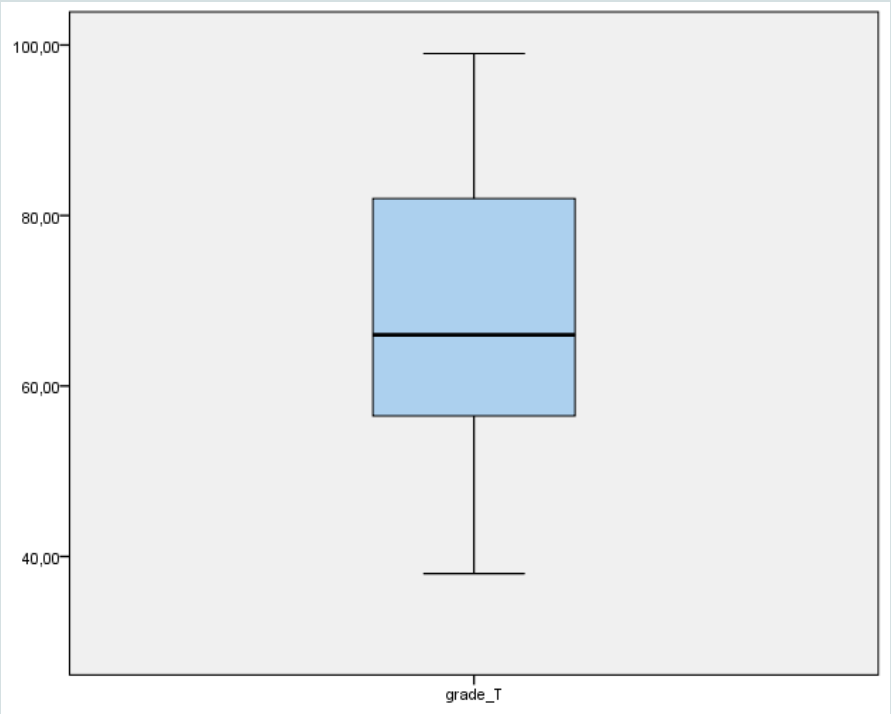




Παράδειγμα I

Εντοπισμός Ακραίων Τιμών

Τα αποτελέσματα χωρίς τις ακραίες τιμές



Έλεγχος Κανονικής Κατανομής

Οι **προϋποθέσεις** για να εξετάσουμε αν τα δεδομένα του δείγματος ακολουθούν την **Κανονική** Κατανομή είναι:

- **Απουσία Ακραίων τιμών** (outliers)
  - Εντοπίζονται είτε με τα z-scores είτε με τα θηκογράμματα
  - Εφόσον εκλεχθούν απομακρύνονται από το δείγμα
- **Ανεξαρτησία** (Independence)
  - Δεν βασίζεται σε κάποιο στατιστικό τεστ, αλλά στη λογική της έρευνας
  - Μία μέτρηση, π.χ. ο βαθμός ενός φοιτητή σε ένα τεστ, θα πρέπει να μην επηρεάζεται από τους βαθμούς άλλων φοιτητών
  - Ανάλογα με το σχεδίαση της έρευνας, μπορεί να δοθεί διαφορετικό νόημα στην ανεξαρτησία

Έλεγχος Κανονικής Κατανομής

Προϋποθέσεις

- **Συνέχεια** (Interval Data)
  - Συνεχείς τιμές σε κλίμακα τιμών (π.χ. 1-10)
  - Αν η βαθμολογία δύο φοιτητών σε ένα τεστ γνώσης είναι 7 και 10 αντίστοιχα, η διαφορά στην κλίμακα θα πρέπει να αντιπροσωπεύει αντίστοιχη πραγματική διαφορά στη γνώση

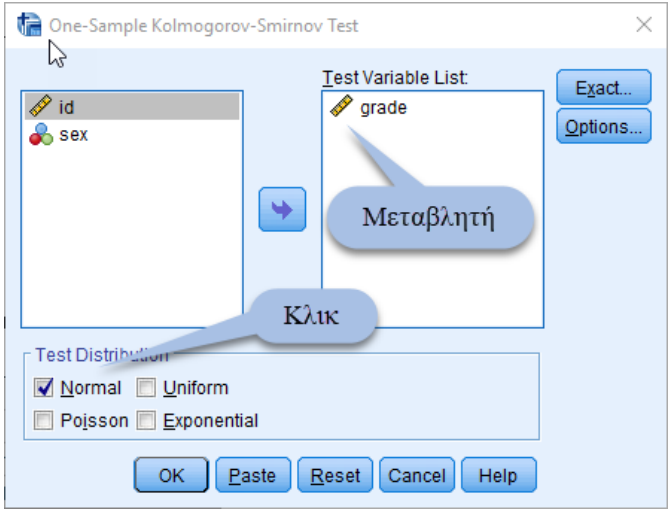
Δεν υπάρχει στατιστικό τεστ για τον έλεγχο – βασίζεται στη λογική

Έλεγχος Κανονικής Κατανομής

Μέθοδος 1η

Για να ελέγξουμε αν η **κατανομή** μιας μεταβλητής είναι συμβατή με την **κανονική** αρχικά θα έπρεπε να εξετάσουμε τις προϋποθέσεις (ακραίες τιμές, ανεξαρτησία, συνέχεια) και στην συνέχεια να εφαρμόσουμε το test **Kolmogorov-Smirnov** (lecture1_3.sav)

- ✓ Analyze → Non parametric tests → One sample K-S
- ✓ Βάζουμε στο test variable list τις μεταβλητές που θέλουμε να ελέγξουμε την κανονικότητα τους,
- ✓ Τσεκάρουμε Normal και OK



Έλεγχος Κανονικής Κατανομής

Παρατηρούμε ότι η τιμή **p-value** είναι 0,001 επομένως **μικρότερη** του 0,05 το οποίο θέσαμε ως επίπεδο στατιστικής σημαντικότητας.

Επομένως απορρίπτουμε την μηδενική υπόθεση H_0 (η κατανομή, δε διαφέρει από την κανονική κατανομή).

One-Sample Kolmogorov-Smirnov Test

		grade
N		50
Normal Parameters[a,b]	Mean	67,4028
	Std. Deviation	6,58471
Most Extreme Differences	Absolute	,174
	Positive	,146
	ative	-,174
Test Statistic		,174
Asymp. Sig. (2-tailed)		,001[c]

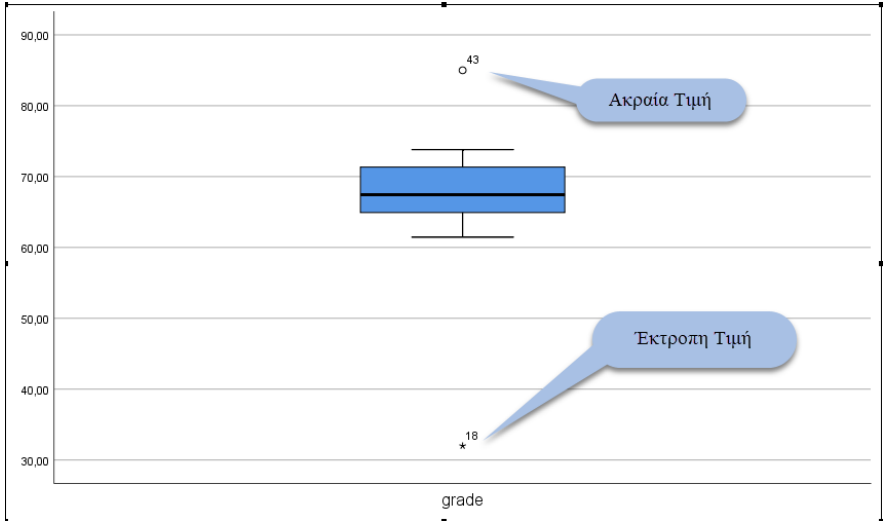
p-value

Έλεγχος Κανονικής Κατανομής

Αν πριν προχωρήσουμε στον έλεγχο κανονικότητας τρέξουμε την διαδικασία **εντοπισμού** και απομάκρυνσης τυχόν ακραίων τιμών ίσως τα αποτελέσματα να ήταν είναι διαφορετικά.

Παρατηρούμε ότι η παρατήρηση 43 είναι **ακραία** και ότι η παρατήρηση 18 είναι **Έκτροπη**.

Τις απομακρύνουμε και τρέχουμε ξανά τον έλεγχο



Έλεγχος Κανονικής Κατανομής

Μετά την απομάκρυνση των δύο παρατηρήσεων παρατηρούμε ότι η τιμή **p-value** είναι 0,200 επομένως **μεγαλύτερη** του 0,05 το οποίο θέσαμε ως επίπεδο στατιστικής σημαντικότητας.

Επομένως δεν μπορούμε να απορρίψουμε την την μηδενική υπόθεση $\mathbf{H_0}$ (η κατανομή, δε διαφέρει από την κανονική κατανομή).

One-Sample Kolmogorov-Smirnov Test

		grade
N		48
Normal Parameters[a,b]	Mean	67,7738
	Std. Deviation	3,43616
Most Extreme Differences	Absolute	,100
	Positive	,068
	Negative	-,100
Test Statistic		,100
Asymp. Sig. (2-tailed)		,200[c,d]

p-value

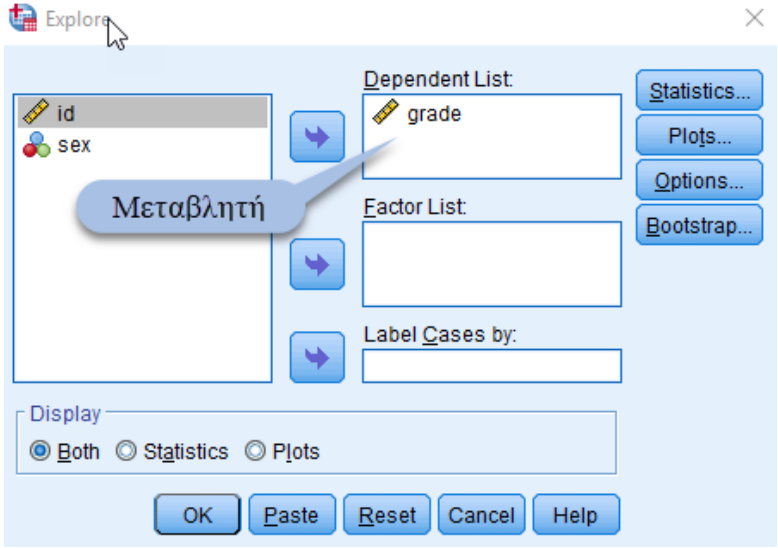
Παρατήρηση: Η ύπαρξη ακραίων τιμών επηρεάζουν την κανονική κατανομή

Έλεγχος Κανονικής Κατανομής

Μέθοδος 2η

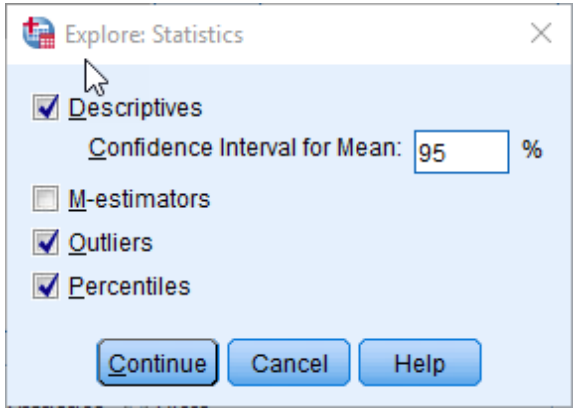
Για να ελέγξουμε αν η **κατανομή** μιας μεταβλητής είναι συμβατή με την **κανονική** μπορούμε να χρησιμοποιήσουμε την παρακάτω διαδικασία η οποία περιέχει και τον εντοπισμό ακραίων τιμών

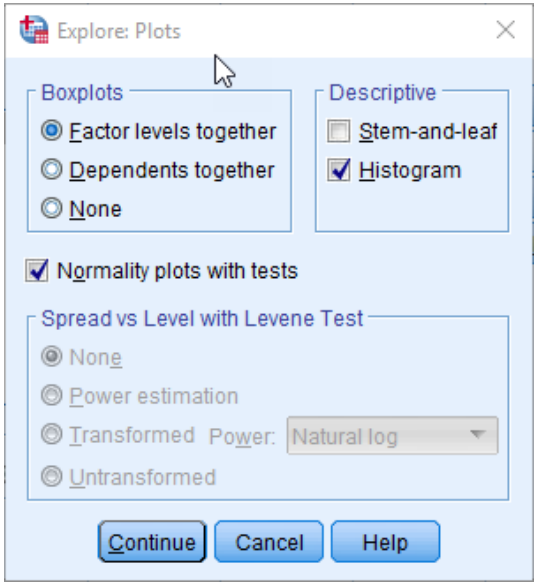
- ✓ Analyze → Descriptive Statistics → Explore
- ✓ Βάζουμε στο Dependent List τις μεταβλητές που θέλουμε να ελέγξουμε την κανονικότητα τους



Έλεγχος Κανονικής Κατανομής

Από την επιλογή **Statistics** τσεκάρουμε τις επιλογές Descriptives, Outlies, Percentiles και ορίζουμε το Διάστημα Εμπιστοσύνης.





Από την επιλογή **Plots** τσεκάρουμε τις επιλογές Histogram, και Normality plots with tests

Έλεγχος Κανονικής Κατανομής

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	grade	61,6800	62,6880	64,9225	67,4400	71,3750	72,6720	73,6625
Tukey's Hinges	grade			64,9300	67,4400	71,3400		

Από το πίνακα Percentiles παρατηρούμε ότι τιμές μικρότερες του 61,68 ή τιμές μεγαλύτερες από 73,66 είναι πιθανές ακραίες τιμές.

Από τον πίνακα Extreme Values παρατηρούμε ότι οι παρατηρήσεις **43,11,18** και **24** είναι πιθανές ακραίες τιμές

Extreme Values

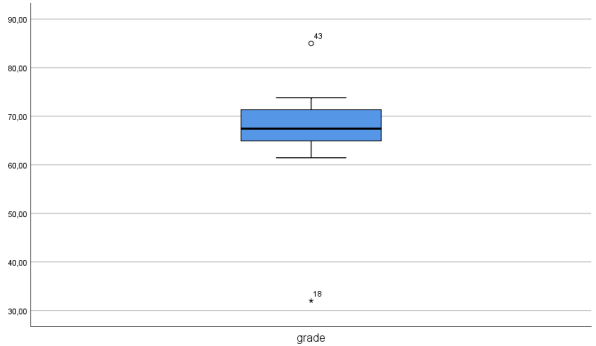
			Case Number	Value
grade	Highest	1	43	85,00
		2	11	73,80
		3	16	73,55
		4	14	73,07
		5	9	72,68
	Lowest	1	18	32,00
		2	24	61,46
		3	44	61,86
		4	40	62,62
		5	4	62,66

Έλεγχος Κανονικής Κατανομής

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
grade	,174	50	,001	,711	50	,000

a. Lilliefors Significance Correction



Από το θηκόγραμμα επιβεβαιώνονται οι παρατηρήσεις που κάναμε στους προηγούμενους πίνακες για ύπαρξη ακραίων τιμών.

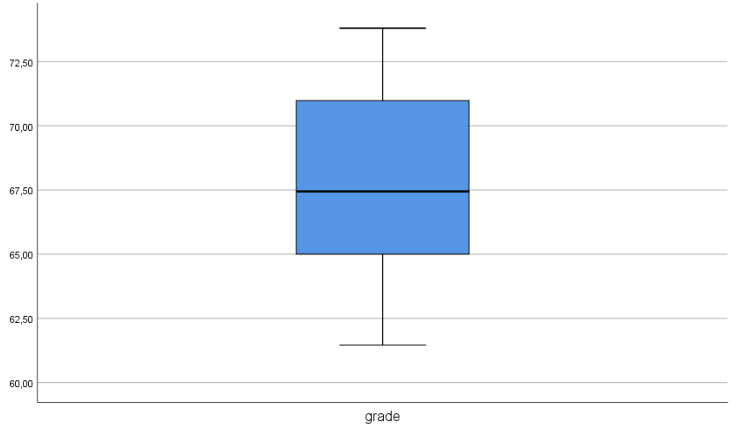
Από τον πίνακα Test of Normality παρατηρούμε ότι η τιμή P-value είναι μικρότερη από το επίπεδο στατιστικής σημαντικότητας (0,05) που θέσαμε και επομένως **απορρίπτουμε** την μηδενική υπόθεση της κανονικής κατανομής.

Απομακρύνουμε τις ακραίες τιμές και ξανατρέχουμε τον έλεγχο

Έλεγχος Κανονικής Κατανομής

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	grade	62,2020	62,9120	64,9650	67,4400	71,1600	72,6080	73,3340
Tukey's Hinges	grade			65,0000	67,4400	70,9800		



Extreme Values

			Case Number	Value
grade	Highest	1	11	73,80
		2	16	73,55
		3	14	73,07
		4	9	72,68
		5	31	72,60
	Lowest	1	24	61,46
		2	44	61,86
		3	40	62,62
		4	4	62,66
		5	23	62,94

Δεν υπάρχουν **ακραίες** τιμές επομένως εξετάζουμε τον πίνακα Test of Normality

Έλεγχος Κανονικής Κατανομής

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
grade	,100	48	,200*	,963	48	,137

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Από τον πίνακα Test of Normality εξετάζουμε :

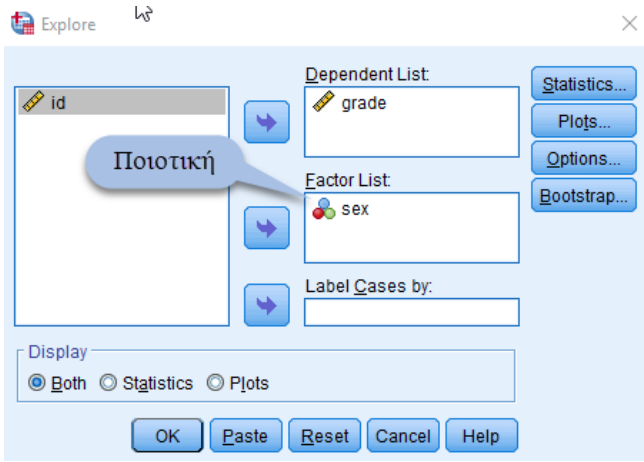
- ✓ Το τεστ του Kolmogorov – Smirnov όταν ο αριθμός των παρατηρήσεων είναι **μεγάλος** (>=50)
- ✓ Το τεστ του Shapiro – Wilk όταν ο αριθμός των παρατηρήσεων είναι **μικρός** (<50)

και στις δύο περιπτώσεις η τιμή **p-value** είναι **μεγαλύτερη** του 0,05 επομένως δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση

Έλεγχος Κανονικής Κατανομής

Αν μία **ποιοτική** μεταβλητή διαχωρίζει το δείγμα σε περισσότερες υποομάδες μπορούμε να εκτελέσουμε το τεστ κανονικότητας για τις επιμέρους ομάδες με την παρακάτω διαδικασία:

- ✓ Analyze → Descriptive Statistics → Explore
- ✓ Βάζουμε στο Dependent List τις μεταβλητές που θέλουμε να ελέγξουμε την κανονικότητα τους
- ✓ Στο Factor List βάζουμε την **ποιοτική** μεταβλητή και εκτελούμε τον έλεγχο



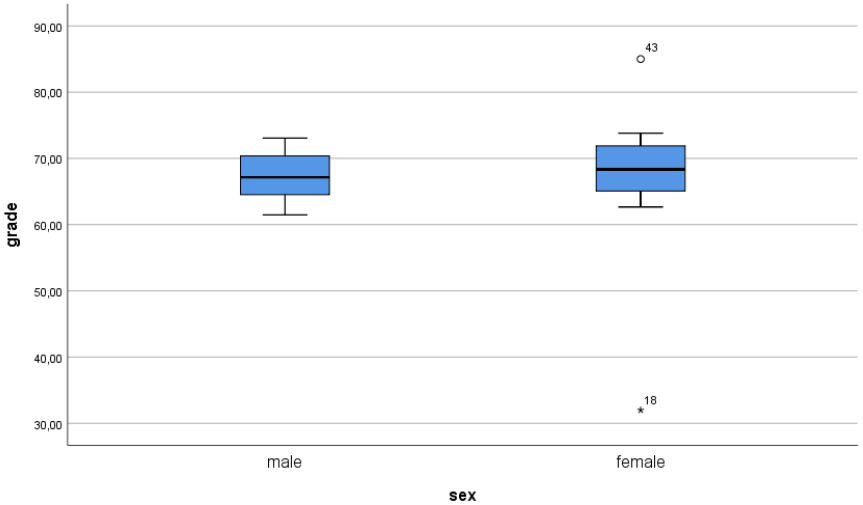
Έλεγχος Κανονικής Κατανομής

Percentiles

		sex	Percentiles 5	10	25	50	75	90	95
Weighted Average (Definition 1)	grade	male	61,6400	62,5440	64,3300	67,1400	70,4975	72,5460	72,8585
		female	36,5990	62,9720	65,0350	68,3400	71,9975	73,7250	83,3200
Tukey's Hinges	grade	male			64,5200	67,1400	70,3750		
		female			65,0700	68,3400	71,8800		

Extreme Values

	sex			Case Number	Value
grade	male	Highest	1	14	73,07
			2	31	72,60
			3	8	72,54
			4	34	71,87
			5	48	71,74
		Lowest	1	24	61,46
			2	44	61,86
			3	40	62,62
			4	23	62,94
			5	20	63,82
	female	Highest	1	43	85,00
			2	11	73,80
			3	16	73,55
			4	9	72,68
			5	36	72,35
		Lowest	1	18	32,00
			2	4	62,66
			3	22	63,70
			4	37	63,73
			5	17	64,93



Απομακρύνουμε τις ακραίες τιμές και εκτελούμε ξανά τον έλεγχο

Έλεγχος Κανονικής Κατανομής

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	sex	Statistic	df	Sig.	Statistic	df	Sig.
grade	male	,096	28	,200*	,958	28	,314
	female	,119	20	,200*	,950	20	,372

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Από τον πίνακα Test of Normality παρατηρούμε ότι και για τις δύο υποομάδες που χωρίζει το δείγμα υπό εξέταση η **ποιοτική** μεταβλητή φύλλο, **δεν** μπορούμε να **απορρίψουμε** την μηδενική υπόθεση

Ερωτήσεις ??????????????????????????

Ασκήσεις

Άσκηση 1η

Στον πίνακα (αρχείο lecture_1_5.sav) βρίσκονται οι βαθμοί 25 φοιτητών στο μάθημα Ανάλυσης Δεδομένων στην εξεταστική και στην πρόοδο

Κάντε έλεγχο κανονικότητας για τις μεταβλητές Βαθμός Εξεταστικής και Βαθμός Προόδου

- για το σύνολο των φοιτητών
- για τα αγόρια και τα κορίτσια ξεχωριστά

Προσοχή στην ύπαρξη Ακραίων Τιμών

a/a	Έτος	Επώνυμο	Φύλλο	Βαθμός Εξεταστικής	Βαθμός Προόδου
1	1	Δρα	man	7	6
2	1	Νίκ	female	4	6
3	1	Παπ	man	9	5
4	1	Κίτ	female	5	6
5	1	Ματ	female	8	5
6	1	Κατ	man	5	5
5	2	Μίτ	man	7	9
7	2	Μίρ	female	6	5
8	2	Μαύ	man	10	6
9	2	Κού	man	2	7,5
10	2	Ταρ	man	0	6,5
11	2	Τσα	female	28	7,5
12	3	Ταβ	female	6	6
13	3	Λίτ	female	8,5	7,5
14	3	Εκε	man	6,5	5,5
15	3	Αυτ	female	5,5	5,5
16	3	Φαρ	female	7	9
17	3	Χαρ	man	4,5	8
18	4	Εκε	female	4,5	7,5
19	4	Ντί	man	7,5	10
20	4	Βασ	man	10	0
21	4	Ζερ	man	9,5	7
22	4	Μπα	man	10	7,5
23	4	Χτα	female	6,5	6,5
24	4	Τρα	female	8,5	7,5
25	4	Στα	female	9	6

Ασκήσεις

Άσκηση 2η

Στον πίνακα (αρχείο lecture_1_4.sav) βρίσκονται οι τιμές της χοληστερίνης ενός δείγματος 60 ατόμων

Κάντε έλεγχο κανονικότητας για τις μεταβλητές Χοληστερίνη και Ηλικία

- για το σύνολο του δείγματος
- για τους άνδρες και τις γυναίκες ξεχωριστά

Προσοχή στην ύπαρξη Ακραίων Τιμών

a/a	Χοληστερίνη	Φύλλο	Ηλικία	Δόση	a/a	Χοληστερίνη	Φύλλο	Ηλικία	Δόση
1	161	man	31	καθόλου	31	212	man	38	μέτρια
2	163	man	19	καθόλου	32	218	woman	34	καθόλου
3	169	man	39	καθόλου	33	223	man	50	μέτρια
4	169	woman	41	ελάχιστη	34	223	woman	51	καθόλου
5	170	woman	35	καθόλου	35	225	woman	49	ελάχιστη
6	173	man	31	μικρή	36	226	woman	54	ελάχιστη
7	174	woman	33	ελάχιστη	37	227	man	39	μικρή
8	176	woman	28	καθόλου	38	227	woman	45	μέτρια
9	195	man	49	μέτρια	39	228	woman	50	καθόλου
10	195	man	41	ελάχιστη	40	233	man	34	ελάχιστη
11	233	woman	54	καθόλου	41	258	woman	53	μικρή
12	234	man	53	ελάχιστη	42	258	woman	54	ελάχιστη
13	239	man	44	μέτρια	43	281	woman	52	μικρή
14	244	woman	50	ελάχιστη	44	282	woman	59	καθόλου
15	248	man	52	μικρή	45	282	woman	60	μικρή
16	249	man	47	μικρή	46	284	man	67	ελάχιστη
17	249	woman	49	μικρή	47	284	woman	62	μέτρια
18	256	man	46	ελάχιστη	48	284	woman	53	καθόλου
19	256	woman	63	μέτρια	49	284	woman	54	ελάχιστη
20	258	woman	64	μέτρια	50	286	woman	62	μέτρια
21	50	man	95	καθόλου	51	286	woman	68	μέτρια
22	195	woman	34	μικρή	52	297	man	64	καθόλου
23	195	woman	38	καθόλου	53	299	man	66	ελάχιστη
24	196	man	36	ελάχιστη	54	301	man	64	μέτρια
25	199	woman	31	μέτρια	55	809	woman	105	καθόλου
26	200	woman	36	καθόλου	56	309	woman	57	ελάχιστη
27	209	woman	36	μικρή	57	310	woman	61	μέτρια
28	209	woman	56	καθόλου	58	330	woman	77	μέτρια
29	210	man	41	ελάχιστη	59	354	woman	63	μικρή
30	211	man	37	ελάχιστη	60	355	man	64	μέτρια