



17/1/2023

## Αναλυτικές Μέθοδοι στη Γεωπληροφορική

### Στατιστικοί Έλεγχοι Υποθέσεων

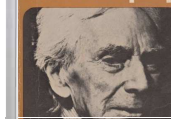
## Ατζέντα σημερινού μαθήματος

### • Έλεγχος υποθέσεων

- Αναγκαιότητα, λογική και σημασία των ελέγχων υποθέσεων στην έρευνα
- Βασικές έννοιες και ορισμοί
  - Πρακτικό και διαδικαστικό πλαίσιο εφαρμογής ελέγχου υποθέσεων
- Στατιστικοί έλεγχοι μέτρων θέσης και διασποράς τ.μ.
- Παραδείγματα με χρήση του R

## Η λογική (και η σημασία) των υποθέσεων στην έρευνα

History of Western Philosophy



BERTRAND RUSSELL

"... μια απλή τακτική διάταξη των δεδομένων θα έκανε προφανή τη σωστή υπόθεση, αλλά αυτό συμβαίνει σπάνια. Κατά κανόνα, η διαμόρφωση των υποθέσεων είναι το πιο δύσκολο μέρος της επιστημονικής έρευνας."

*Bertrand Russell* (1946 - *History of Western Philosophy*), αναφερόμενος στην ανεπαρκή έμφαση της σημασίας των υποθέσεων στην επαγωγική μέθοδο του Francis Bacon

## Διαφορά μεταξύ μιας ερευνητικής και μιας στατιστικής υπόθεσης;

- **Ερευνητική υπόθεση** είναι μια προτεινόμενη απάντηση σε κάποιο ερευνητικό ερώτημα. Συνήθως περιλαμβάνει μια εξήγηση ("...το x επηρεάζει το y επειδή ...").
- **Στατιστική υπόθεση**, από την άλλη πλευρά, είναι μια **μαθηματική δήλωση** σχετικά με μια παράμετρο ενός πληθυσμού
  - Σε μια καλά σχεδιασμένη μελέτη, οι στατιστικές υποθέσεις αντιστοιχούν λογικά στην υπόθεση της έρευνας, π.χ. υπάρχει σχέση μεταξύ μεταβλητών ή διαφορά μεταξύ δειγμάτων, ...

## Στατιστικές δοκιμές / έλεγχοι

- Κάθε διαδικασία μετρήσεων, είναι αναγκαίο να ακολουθείται από μια ή περισσότερες στατιστικές δοκιμές ή ελέγχους για τα δεδομένα που συλλέχθηκαν
  - Τα συμπεράσματα τους επιτρέπουν την αξιολόγηση των αποτελεσμάτων μετρήσεων → **βοηθούν να κριθεί εάν τα δεδομένα είναι σημαντικά** (ακριβή, αξιόπιστα, έγκυρα, ...) ή κατάλληλα για να απαντήσουν σε σύνθετα ερωτήματα ή να εντοπίσουν προβλήματα στα δεδομένα



## Βασικοί ορισμοί

- Ο όρος στατιστική δοκιμή ή υπόθεση ή στατιστικός έλεγχος (*hypothesis testing*) υποδηλώνει **μία υπόθεση ή ισχυρισμό** που γίνεται στα πλαίσια μιας έρευνας ή ανάλυσης για τα δεδομένα ενός πληθυσμού που συλλέχθηκαν από ένα τυχαίο πείραμα, μια μετρητική διαδικασία ή την παρατήρηση ενός φαινομένου.
  - Η εκάστοτε στατιστική υπόθεση που γίνεται δεν είναι υποχρεωτικό να είναι αληθής κάθε φορά.
  - Ο έλεγχος υποθέσεων είναι μια έγκυρη διαδικασία επικύρωσης της εκάστοτε υπόθεσης για μια έρευνα.



## Έλεγχος υποθέσεων - Αναγκαιότητα -

- Θεωρητικά, για να εξαχθούν έγκυρα συμπεράσματα για το ευρύ σύνολο των υποκείμενων μιας έρευνας (π.χ., *ατόμων, αντικειμένων, φαινομένων, ή συμπεριφορών με κάποιο κοινό μετρήσιμο χαρακτηριστικό*), θα πρέπει να ληφθεί υπόψη **το σύνολο του πληθυσμού ενδιαφέροντος** → Συνήθως αυτό δεν είναι **πρακτικά εφικτό** (από πλευράς πόρων, κόστους, χρόνου, προσβασιμότητας, ...)



- Αντί αυτού, χρησιμοποιούμε **τυχαία δείγματα αντιπροσωπευτικά του υπό μελέτη πληθυσμού**,
- για να διατυπωθούν υποθέσεις προκειμένου να ελεγχθούν "ισχυρισμοί" ή "θεωρήσεις" που αφορούν ή εξηγούν τον πληθυσμό, και από
- **ενδείξεις που προκύπτουν από την ανάλυση τους, είτε επιλέγεται είτε απορρίπτεται η εκάστοτε υπόθεση στη βάση ενός συγκεκριμένου (κάθε φορά) στατιστικού κριτηρίου**



- Συνήθως όταν σχεδιάζουμε κάποιο πείραμα ή συλλογή μετρήσεων, έχουμε κάποια θεωρία ή μοντέλο που θέλουμε να επαληθεύσουμε, και επομένως να αποκομίσουμε κάποια ένδειξη για τα τελικά αποτελέσματα.
  - Με στατιστικούς όρους, ουσιαστικά μια στατιστική υπόθεση αποτελεί μια πρόβλεψη για την **επίδραση της αλλαγής της ανεξάρτητης μεταβλητής στην εξαρτημένη μεταβλητή ενός μοντέλου**

- Ελέγχους υποθέσεων (**hypothesis testing**) κάνουμε για να εξακριβώσουμε αν τα δεδομένα φαίνεται να υποστηρίζουν ή όχι κάποια ιδέα που έχουμε για το μηχανισμό που δημιουργεί τα δεδομένα.



- "υπάρχει διαφορά;";
- "υπάρχει σχέση;";
- "υπάρχει επίδραση;"; ...

– Στόχος είναι να συναχθούν συμπεράσματα ως προς το εάν η εκτίμηση των παραμέτρων ενδιαφέροντος από μετρήσεις είναι συμβατή με τις αρχικές παραδοχές της διαδικασίας των μετρήσεων

## Έλεγχοι υποθέσεων

### - Πρακτικό Πλαίσιο -

- Από μια οικογένεια κατανομών πιθανότητας μιας τυχαίας μεταβλητής του ενδιαφέροντος μιας έρευνας, να επιλεγεί η καταλληλότερη που προσαρμόζεται σε ένα δείγμα, όταν
  - υιοθετείται γνώση της συναρτησιακής μορφής (δηλ. της σ.κ.π.  $f(X)$  ή της α.σ.π.  $F(X)$ ) της κατανομής του πληθυσμού από τον οποίο λαμβάνεται το δείγμα, και συνθηθέστερα
  - μερικές ή όλες οι πληθυσμιακές παράμετροι (π.χ., μέση τιμή, διασπορά, ...) που περιγράφουν την κατανομή είναι άγνωστες

- Οι στατιστικές υποθέσεις γίνονται

- είτε επί των αγνώστων παραμέτρων,
- είτε για το σχήμα της κατανομής

- Με βάση ένα τυχαίο δείγμα παρατηρήσεων επιδιώκεται η κατασκευή μιας κατάλληλης ελεγκοσυνάρτησης (**test statistic ή στατιστικό**) για τον έλεγχο της υπόθεσης ότι οι εν λόγω παράμετροι βρίσκονται σε κάποιο ορισμένο υποσύνολο του παραμετρικού χώρου  $\Omega$ .
- Για αυτό το σκοπό, επιλέγεται κάποια κατανομή πιθανότητας για την οποία δεν δεχόμαστε ορισμένη συναρτησιακή μορφή

- Εξετάζεται πόσο είναι πιθανό ότι το υπό εξέταση δείγμα, κάτω από ορισμένες προϋποθέσεις κανονικότητας, προέρχεται από τη συγκεκριμένη κατανομή

- Γενικά, υπάρχουν δύο δυνατές περιπτώσεις

– **Παραμετρικοί στατιστικοί έλεγχοι υποθέσεων** που διεξάγονται κάτω από κάποιες προϋποθέσεις για τις παραμέτρους του πληθυσμού των δεδομένων

– **Μη-παραμετρικοί στατιστικοί έλεγχοι υποθέσεων** στις οποίες λαμβάνονται υπόψη μόνο οι διατάξεις των παρατηρήσεων και όχι οι πραγματικές τιμές των παρατηρήσεων

## Παραμετρικοί έλεγχοι ή δοκιμές

- Προϋποθέτουν ότι τα δεδομένα ακολουθούν κάποιες υποκείμενες στατιστικές κατανομές.
  - Πρέπει να πληρούνται αρκετές προϋποθέσεις εγκυρότητας, ώστε το αποτέλεσμα μιας παραμετρικής δοκιμής να είναι αξιόπιστο.
  - Δεν υπάρχουν ακριβείς κανόνες για τους όρους ισχύος παραμετρικών ελέγχων → **rules of thumb**
    - ✓ Οι μετρήσεις είναι ανεξάρτητες
    - ✓ Τα δείγματα έχουν ίσες διακυμάνσεις
    - ✓ Κανονικότητα πληθυσμού (υποτίθεται ή επαληθεύεται)
    - ✓ Μέγεθος δείγματος

## Παραμετρικοί έλεγχοι συνήθως ...

- Βασίζονται σε δεδομένα που κατανέμονται κανονικά (ή σχεδόν) έτσι ώστε να μπορεί να γίνει εκτίμηση των παραμέτρων του υποκείμενου πληθυσμού.
  - ✓ Οι εκτιμώμενες παράμετροι μπορούν στη συνέχεια να χρησιμοποιηθούν για τον έλεγχο μιας υπόθεσης.
  - ✓ Καθώς μόνο ποσοτικά δεδομένα μπορούν να έχουν κανονική κατανομή, συνεπάγεται ότι **παραμετρικοί έλεγχοι μπορεί να χρησιμοποιηθούν μόνο σε ποσοτικά δεδομένα**.
  - ✓ Οι παραμετρικές δοκιμές έχουν συχνά μη-παραμετρικά ισοδύναμα.

## Μη-παραμετρικοί έλεγχοι ή δοκιμές

- Κάνουν ελάχιστες ή καθόλου υποθέσεις σχετικά με την κατανομή των δεδομένων/του πληθυσμού:

✓ Συνήθως θεωρούνται σχετικά εύκολες διαδικασίες στην εκτέλεση, αλλά μπορεί να προκύψουν ορισμένα προβλήματα, όπως:

- Μπορεί να είναι δύσκολο να πραγματοποιηθούν με μεγάλες ποσότητες δεδομένων.
- Όταν πολλαπλές ίδιες τιμές ή ακραίες τιμές εμφανίζονται αρκετές φορές σε ένα σύνολο δεδομένων, οι δοκιμές με βάση την κατάταξη μπορούν να γίνουν πιο περίπλοκες

## Μη-παραμετρικοί έλεγχοι χαρακτηρίζονται από ορισμένα πλεονεκτήματα όπως:

- ✓ Οι μη-παραμετρικές δοκιμές ισχύουν σε ένα ευρύτερο φάσμα καταστάσεων/ εφαρμογών (λιγότερες συνθήκες εγκυρότητας) σε σχέση με τις παραμετρικές δοκιμές, καθώς ...
- ✓ Δεν απαιτείται η επαλήθευση της καταλληλότητας κάποιας υπόθεσης.
- ✓ Για μικρά δείγματα οι δοκιμασίες οδηγούν σε ακριβή συμπεράσματα και όχι σε προσεγγίσεις.
- ✓ Τα δεδομένα που χρησιμοποιούνται είναι συχνότητες και όχι μετρήσεις.

## Συγκριτικά μεταξύ τους ...

- οι **παραμετρικοί έλεγχοι** εφόσον χρησιμοποιούνται κατάλληλα (π.χ. σε εξάρτηση με το μέγεθος των δειγμάτων), αντλούν περισσότερες πληροφορίες (→ **έχουν μεγαλύτερη στατιστική ισχύ**) για ολόκληρο τον πληθυσμό από ότι αντίστοιχοι μη-παραμετρικοί έλεγχοι.
  - ✓ Τις περισσότερες φορές, η κρίσιμη τιμή πιθανότητας  $p$  που σχετίζεται με μια παραμετρική δοκιμή θα είναι χαμηλότερη από την τιμή  $p$  που σχετίζεται με ένα μη-παραμετρικό ισοδύναμο που εκτελείται στα ίδια δεδομένα.

# Έλεγχος υποθέσεων – Εννοιολογικό Πλαίσιο –

Στους στατιστικούς ελέγχους ενέχονται δύο τύποι υποθέσεων

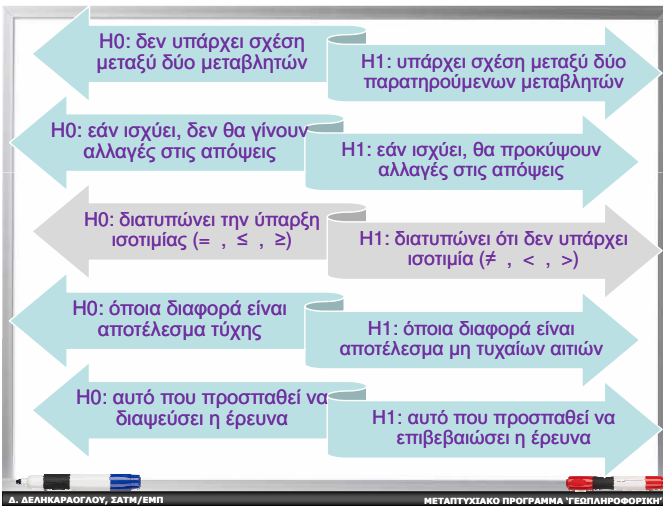
- Η **μηδενική υπόθεση**, *null hypothesis*  $H_0$  (προτάθηκε από τον *R.A. Fisher*).
  - Αφορά το ‘κατεστημένο’ → εκφράζει την ‘αλήθεια’
- Η **εναλλακτική υπόθεση**, *alternative hypothesis*  $H_1$  (προτάθηκε από τους *Jerzy Newman* και *Egon Pearson*. Ο *Fisher* ήταν πολύ αντίθετος στην εισαγωγή της).
  - Θεωρείται έγκυρη αν η  $H_0$  είναι άκαιρη ή παραπλανητική → υπαγορεύει αλλαγή ενεργειών

- Αποσκοπεί στο να ελεγχθεί, με πιθανοθεωρητικό τρόπο, η εγκυρότητα ενός ισχυρισμού, μιας απαίτησης ή μιας παραδοχής που γίνεται για τον πληθυσμό
- Συχνά αντιπροσωπεύει τη συντηρητική (ή την συμβατικά αποδεκτή) παραδοχή ενός ερευνητικού προβλήματος ή μιας άποψης
- Επιλέγεται (διαφορετική κάθε φορά) εκείνη η υπόθεση του εκάστοτε προβλήματος, της οποίας η τυχόν απόρριψη θα συνεπάγεται τη σοβαρότερη επίπτωση για μια έρευνα
- Αφορά στον καθορισμό ενός υποσυνόλου  $\omega$  του παραμετρικού χώρου  $\Omega$ , το οποίο θα περιέχει την αληθινή τιμή  $\theta_0$  μιας παραμέτρου  $\theta$ :  
**Συμβολικά δηλώνεται με  $H_0: \theta_0 \in \omega$**

## Μηδενική υπόθεση

- Είναι η υπόθεση την οποία θα θέλαμε να δείξουμε ότι ισχύει ή θεωρείται έγκυρη, εάν η μηδενική υπόθεση είναι παραπλανητική ή άκαιρη → Αυτή στην οποία στρεφόμεστε εάν δεν ισχύει η μηδενική υπόθεση
- Δηλώνει την ύπαρξη της στατιστικής σχέσης μεταξύ των μεταβλητών → Υποδηλώνει κάποια στατιστικά σημαντική διαφορά ή αποτέλεσμα (μια αλλαγή της ανεξάρτητης μεταβλητής επιδρά στην εξαρτημένη μεταβλητή)
- Αναφέρεται στον καθορισμό ενός υποσυνόλου  $\omega_c$  (του συμπληρώματος του  $\omega$ ), σε σχέση με τον παραμετρικό χώρο  $\Omega$  που περιέχει την αληθινή τιμή  $\theta$  μιας παραμέτρου: Συμβολικά δηλώνεται με  $H_1: \theta \in \omega_c$  ή  $\theta \notin \omega$

## Εναλλακτική υπόθεση



- 1 • Διατύπωση υποθέσεων
  - 2 • Διαμόρφωση των κριτηρίων για την λήψη μιας απόφασης
  - 3 • Συλλογή δεδομένων από τυχαία δείγματα ενός πληθυσμού
  - 4 • Αξιολόγηση της μηδενικής υπόθεσης
- Η λογική του ελέγχου υποθέσεων**
- Εάν τα δεδομένα είναι συνεπή με τη μηδενική υπόθεση  $H_0$  → η  $H_0$  είναι λογική (→ την αποδεχόμαστε)
  - Εάν υπάρχει μεγάλη ασυμφωνία μεταξύ των δεδομένων της μηδενικής υπόθεσης  $H_0$  → η  $H_0$  είναι λανθασμένη (→ την απορρίπτουμε)

## Τα διαδικαστικά βήματα του ελέγχου υποθέσεων

- 1 • Ορίζονται το πρόβλημα και η υπόθεση έρευνας προς έλεγχο
- 2 • Ορίζεται η μηδενική υπόθεση  $H_0$  (Null Hypothesis)
- 3 • Ορίζεται η εναλλακτική υπόθεση  $H_1$  (Alternative Hypothesis)
- 4 • Ορίζεται η στατιστική παράμετρος που θα ελεγχθεί από ένα δείγμα ( $y_1, y_2, \dots, y_k$ ) – π.χ., μέση τιμή, διασπορά, αναλογία (ή διαφορά) παραμέτρων, ...

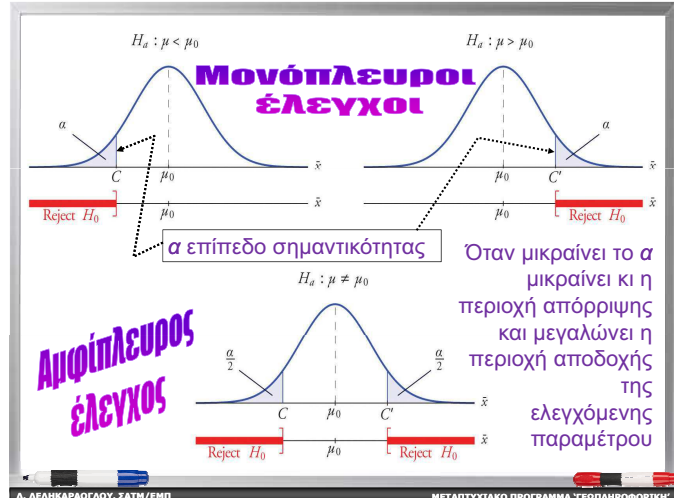
- 5 • Ορίζεται  $R$ , η περιοχή απόρριψης της υπόθεσης για επίπεδο σημαντικότητας  $\alpha$
- 6 • Για την περιοχή απόρριψης της υπόθεσης, επιλέγεται εάν το επίπεδο σημαντικότητας  $\alpha$  εφαρμόζεται για
  - Μονόπλευρη ή δίπλευρη περιοχή απόρριψης
- 7 • Από το δείγμα υπολογίζεται το διάστημα εμπιστοσύνης για την εκτιμήτρια  $\hat{\theta}$  της παραμέτρου ενδιαφέροντος  $\theta$
- 8 • Συνάγεται το συμπέρασμα από τον έλεγχο
  - Η υπόθεση  $H_0$  απορρίπτεται ( $\hat{\theta} \in R$ )
  - Η υπόθεση  $H_0$  δεν απορρίπτεται ( $\hat{\theta} \notin R$ )

## Κάθε διαφορετική μορφή της εναλλακτικής υπόθεσης $H_1$ έχει το δικό της είδος περιοχής απόρριψης

Αν η  $H_1$  έχει τη μορφή  $H_1: \theta < \theta_0$ , απορρίπτουμε την  $H_0$  εάν η εκτιμήτρια από το δείγμα είναι πολύ στα αριστερά του  $\theta_0$ , δηλαδή στα αριστερά κάποιου αριθμού  $C \rightarrow$  η περιοχή απόρριψης έχει τη μορφή διαστήματος  $(-\infty, C]$ .

Αν η  $H_1$  έχει τη μορφή  $H_1: \theta > \theta_0$ , απορρίπτουμε την  $H_0$  εάν η εκτιμήτρια από το δείγμα είναι πολύ δεξιά του  $\theta_0$ , δηλαδή δεξιά του αριθμού  $C \rightarrow$  η περιοχή απόρριψης έχει μορφή διαστήματος  $[C, +\infty)$

Αν η  $H_1$  έχει τη μορφή  $H_1: \mu \neq \mu_0$ , απορρίπτουμε το  $H_0$  αν η εκτιμήτρια της παραμέτρου είναι μακριά του  $\theta_0$  προς οποιαδήποτε κατεύθυνση, δηλαδή προς τα αριστερά κάποιου αριθμού  $C$  ή προς τα δεξιά ενός άλλου αριθμού  $C'$ , οπότε η περιοχή απόρριψης έχει τη μορφή της σύνδεσης δύο διαστημάτων  $(-\infty, C] \cup [C', +\infty)$



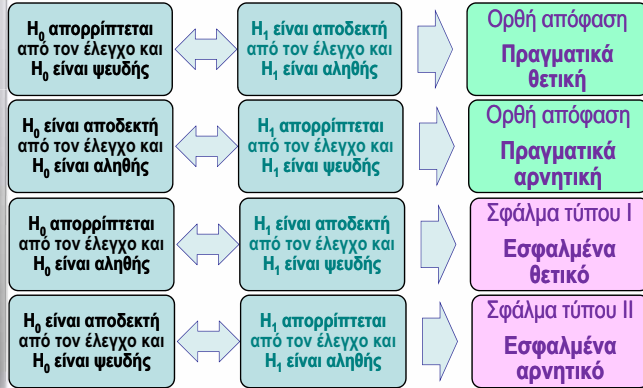
## Καθώς η $H_0$ είναι μια υπόθεση μηδενικής διαφοράς

Εάν απορριφθεί η  $H_0$  τότε ισχύει η  $H_1$  (στατιστικό σημαντικό αποτέλεσμα εφόσον  $p < \alpha$ )  $\rightarrow$  η διαφορά δεν υφίσταται κατά τύχη (ή σύμπτωση ή λάθος)

Εάν δεν απορριφθεί η  $H_0$  τότε την αποδεχόμαστε (χωρίς αυτό να σημαίνει ότι η  $H_0$  είναι αληθής)  $\rightarrow$  διαφορά μπορεί να υφίσταται κατά τύχη (ή σύμπτωση, ή λάθος)

- Στις στατιστικές διαδικασίες δεν αποδεικνύουμε την εκάστοτε ερευνητική υπόθεση απευθείας αλλά έμμεσα, έχοντας απορρίψει τη μηδενική υπόθεση.
- Η εναλλακτική υπόθεση συνάγεται ως λογική συνέπεια της άρνησης της μηδενικής υπόθεσης (σε σχέση πάντα με την ακρίβεια του ελέγχου, και τη συμβατότητά της με τα διαθέσιμα δεδομένα)

## Ενδεχόμενα ελέγχου υποθέσεων



- Το **σφάλμα τύπου I** (συμβολικά  $\alpha$ ) είναι πιο σοβαρό από το **σφάλμα τύπου II** (συμβολικά  $\beta$ ). Το  $1-\beta$  αποκαλείται 'ισχύς του στατιστικού ελέγχου'

		Στον πληθυσμό από όπου εξάγεται το δείγμα, η ...	
		$H_0$ είναι λογική	$H_0$ είναι λανθασμένη
<b>Βασισμένη στο δείγμα, η απόφαση είναι ...</b>	Na μην απορριφθεί η $H_0$	<b>Σωστή απόφαση</b>	<b>Σφάλμα τύπου II (<math>\beta</math>)</b>
	Na απορριφθεί η $H_0$	<b>Σφάλμα τύπου I (<math>\alpha</math>)</b>	<b>Λανθασμένη απόφαση</b>

Για να αποφασίσουμε εάν απορρίπτεται η  $H_0$  πρέπει ...

**1. Να μπορούμε να υπολογίσουμε την πιθανότητα  $p$**  (δηλ. την πιθανότητα το δείγμα να προέρχεται κατά τύχη (ή κατά σύμπτωση, ή λάθος) από έναν πληθυσμό όπου ισχύει η μηδενική υπόθεση)

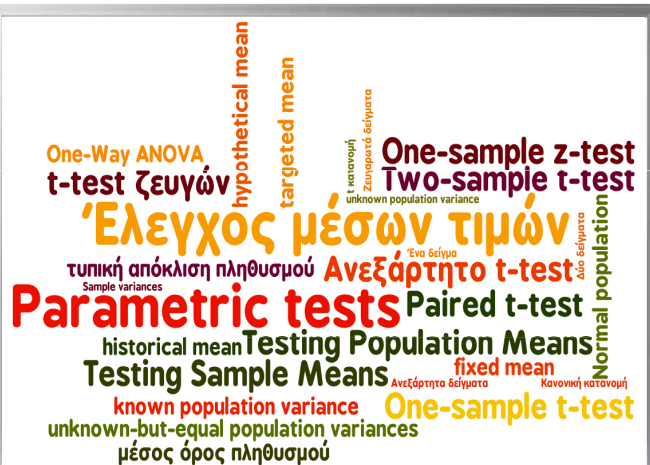
- Απαιτείται να γνωρίζουμε τη δειγματοληπτική κατανομή, και
- τις πιθανότητες που αντιστοιχούν στις διάφορες τιμές αυτής της κατανομής

2. Να αποφασίσουμε εάν η πιθανότητα αυτή, το  $p$ , είναι πολύ μικρή  $\rightarrow$  η διαφορά μηδενικής υπόθεσης που βρήκαμε δεν είναι κατά τύχη (ή κατά σύμπτωση, ή κατά λάθος) και άρα δεν ισχύει η  $H_0$ .

- Αυτό καθορίζεται a priori, μάλλον "αυθαίρετα", με το εκάστοτε επιθυμητό **επίπεδο στατιστικής σημαντικότητας, την παράμετρο  $\alpha$**
- Εάν  $p \leq \alpha \rightarrow$  με βάση τα δεδομένα, η μηδενική υπόθεση φαίνεται ως απίθανη, και απορρίπτεται, δηλ. υπάρχει στατιστικά σημαντική διαφορά
- Εάν  $p > \alpha \rightarrow$  δεν υπάρχει στατιστικά σημαντική διαφορά ( $\rightarrow$  δηλ. με πιθανότητα  $p$ , αποδεχόμαστε ότι το δείγμα προέρχεται από έναν πληθυσμό όπου ισχύει η μηδενική υπόθεση)

## Επιλογή στατιστικών δοκιμών

		Τύποι δεδομένων	
Κατηγορικά δεδομένα		Αριθμητικά δεδομένα	
Μη-παραμετρικές δοκιμές		Τα δεδομένα ακολουθούν την κανονική κατανομή ;	
		OXI	NAI
Μη-παραμετρικές δοκιμές		Είναι οι διασπορές των δύο πληθυσμών των δεδομένων ίσες ;	
		OXI	NAI
	Μη-παραμετρικές δοκιμές		Παραμετρικές δοκιμές



## Έλεγχος μέσων τιμών τυχαίων μεταβλητών

- Αποτελεί μια από τις πιο συχνές επαγωγικές διαδικασίες που αφορά δύο περιπτώσεις:
  - Στη σύγκριση της μέσης τιμής  $\mu$  μιας ποσοτικής μεταβλητής  $X$ , από ένα δείγμα  $\{x_1, x_2, \dots, x_n\}$  μεγέθους  $n$ , με μια προκαθορισμένη τιμή  $\mu_0$ , και
  - Στη σύγκριση των μέσων τιμών  $\mu_1$  και  $\mu_2$  δύο πληθυσμών (π.χ., για τη βροχόπτωση σε δύο περιοχές), που είτε αποτελούνται από ανεξάρτητες είτε από εξαρτημένες παρατηρήσεις

## Τύπος ελέγχου της μέσης τιμής ενός ή περισσότερων δειγμάτων

- Διακρίνουμε διαφορετικές τυπικές περιπτώσεις ελέγχων ...
- Ανάλογα με το αν ο πληθυσμός που ενδιαφέρει ακολουθεί αντίστοιχα
    - κανονική κατανομή με γνωστή διασπορά
    - κανονική κατανομή με άγνωστη διασπορά
    - οποιαδήποτε κατανομή με άγνωστη διασπορά και το μέγεθος δείγματος είναι μεγάλο

- Ανάλογα με το μέγεθος του δείγματος ή/και το πλήθος των δειγμάτων



## Στατιστικοί έλεγχοι ενός δείγματος *One-sample z- και t-tests*

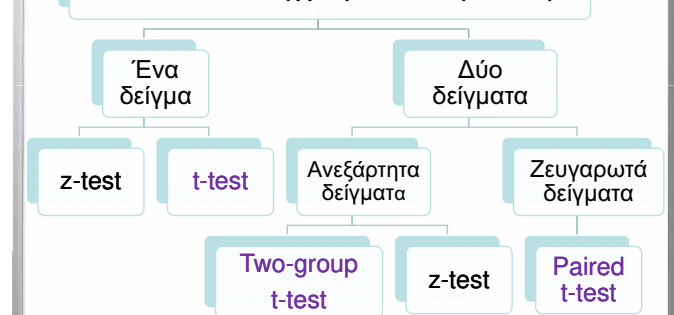
Αφορούν στατιστικούς ελέγχους της αξιοπιστίας

$$\text{της μέσης τιμής } m = \bar{x} = \frac{1}{n} \sum_{i=1}^n (x_1 + x_2 + \dots + x_n)$$

ενός δείγματος που αντιπροσωπεύει μια ποσοτική μεταβλητή  $X$  ενδιαφέροντος, η οποία ακολουθεί μια κανονική κατανομή

- Η δειγματική τιμή λαμβάνεται ως η εκτίμηση της πραγματικής μέσης τιμής  $\mu$  του πληθυσμού
- Ελέγχεται η ισότητα του πληθυσμιακού μέσου  $\mu$ , με μια υποθετική ή δεδομένη μέση τιμή  $\mu_0$

## Στατιστικοί έλεγχοι μέσω τιμών $t$ .



**Διαφορετικοί έλεγχοι ανάλογα με τον αριθμό και το μέγεθος των δειγμάτων**

### Οι έλεγχοι-z (*z-tests*) εφαρμόζονται όταν

1. Το μέγεθος του δείγματος είναι **μεγαλύτερο από 30**.
  - ✓ Επειδή σύμφωνα με το Κεντρικό Οριακό Θεώρημα, όταν ο αριθμός των δειγμάτων αυξάνεται, η μέση κατανομή τους γίνεται παρόμοια με την κανονική κατανομή.
2. Ο γονικός πληθυσμός από όπου αντλούνται τα δείγματα ακολουθεί **κανονική κατανομή με γνωστή διακύμανση  $\sigma^2$** .
  - ✓ Για μέτρια ή μεγάλα δείγματα ( $n \gg 30$ ) μπορούμε να αντικαταστήσουμε τη διακύμανση  $S^2$  του δείγματος με την  $\sigma^2$
3. Οι παρατηρήσεις είναι **ανεξάρτητες μεταξύ τους**.
4. Τα δείγματα είναι **ανεξάρτητα το ένα από το άλλο**.
5. Ομοιογένεια διακυμάνσεων (δηλαδή, διακυμάνσεις περίπου ίσες μεταξύ δειγμάτων, ή  $S^2 \approx \sigma^2$ )

### Οι έλεγχοι-t (*t-tests*) εφαρμόζονται όταν

1. Το μέγεθος  $n$  του δείγματος είναι **μικρότερο από 30**.
2. Ο γονικός πληθυσμός από τον οποίο αντλούνται τα δείγματα ακολουθεί **κανονική ή σχεδόν κανονική κατανομή με άγνωστη διακύμανση  $\sigma^2$** .
3. Οι παρατηρήσεις είναι **ανεξάρτητες μεταξύ τους**.
4. Καθώς το μέγεθος του δείγματος είναι μικρό, τα δεδομένα πρέπει να ελέγχονται ως προς την κανονικότητά τους
  - Μη κανονικές κατανομές πληθυσμού (π.χ., με χοντρή ουρά ή πολύ λοξές), μειώνουν σημαντικά τη δύναμη του ελέγχου
  - Σε μέτρια ή μεγάλα δείγματα, μια παραβίαση της κανονικότητας μπορεί να εξακολουθεί να αποδίδει ακριβείς τιμές  $p$

### Σημαντικές προϋποθέσεις για τα t-test

- **Κανονικότητα**: οι εξαρτημένες μεταβλητές ενδιαφέροντος ακολουθούν κανονική κατανομή.
  - Κρίσιμο για μικρό δείγμα ( $< 30$  παρατηρήσεις)
- **Ανεξαρτησία παρατηρήσεων**: επιλογή τυχαίου δείγματος από τον πληθυσμό και υπό τους ίδιους όρους και συνθήκες
- **Ίσες διακυμάνσεις** των κατανομών των δειγμάτων: Εάν δεν είναι ίσες επιλέγεται το αποτέλεσμα του λεγόμενου **διορθωμένου t-test** ή από μη-παραμετρικές δοκιμές
- **Δείγματα χωρίς ακραίες τιμές** (outliers ή χονδροειδή σφάλματα)

### Έλεγχος μέσω τιμών τυχαίων μεταβλητών

Είναι γνωστή η τυπική απόκλιση του πληθυσμού;

OXI

NAI

Πόσα δείγματα προς έλεγχο;

Πόσα δείγματα προς έλεγχο;

	1	2	1	2
t-test		Παρατηρήσεις κατά ζεύγη;	Z-test	Σύγκριση δύο μέσων τιμών
		OXI	NAI	
		$S_{\text{δείγμα-1}} \approx S_{\text{δείγμα-2}}$	Paired t-test	
		OXI	NAI	
	2-sample t-test, ξεχωριστές διακυμάνσεις	2-sample t-test, συγκεντρωτική διακύμανση		

### Έλεγχος ενός δείγματος (*One-sample tests*)

- Από ένα δείγμα, ελέγχεται η ισότητα του πληθυσμιακού μέσου μιας ποσοτικής μεταβλητής με μια προκαθορισμένη (ή υποθετική) μέση τιμή

### Έλεγχος δύο ανεξαρτήτων δειγμάτων (*Two-independent samples tests*)

- Ελέγχεται η διαφορά των μέσων τιμών δύο διαφορετικών πληθυσμών (π.χ., για τη βροχόπτωση σε δύο διαφορετικές περιοχές). Αποκαλείται και μη συσχετισμένος έλεγχος.

### Έλεγχος ζευγαρωτών δειγμάτων (*Two-paired samples tests*)

- Ελέγχεται η διαφορά των μέσων τιμών δύο διαφορετικών δειγμάτων που προκύπτουν από ζευγαρωτές παρατηρήσεις των στοιχείων ενός πληθυσμού (π.χ., πριν και μετά από την επίδραση ενός εξωτερικού παράγοντα)

### Πλαίσιο στατιστικού ελέγχου

- Ανεξάρτητες παρατηρήσεις
- Ο πληθυσμός κατανέμεται κανονικά;

### Υποθέσεις

- $H_0 : \mu = \mu_0$   $\mu_0$  είναι μια προκαθορισμένη τιμή
- $H_1 : \mu \neq \mu_0$

### Επίπεδο σημαντικότητας

- Συνήθως  $\alpha = 0.05$  ( $p=95\%$ )
- Μονόπλευρος | Αμφίπλευρος έλεγχος

Στατιστική τιμή ελέγχου, για γνωστή διασπορά  $\sigma^2$

$$z = \frac{(\bar{x} - \mu_0)}{(\sigma / \sqrt{n})}$$

Στατιστική τιμή ελέγχου, για άγνωστη διασπορά  $\sigma^2$

$$t_{df} = \frac{(\bar{x} - \mu_0)}{(S / \sqrt{n})}$$

## Έλεγχος της μέσης τιμής δείγματος

εάν ο πληθυσμός ακολουθεί **κανονική κατανομή με γνωστή διασπορά  $\sigma^2$**  → **Έλεγχος z-test**

$$X_i \sim N(\mu_0, \sigma^2), i = 1, 2, \dots, n$$

- Η τ.μ. του δειγματικού μέσου  $\bar{X}$  ακολουθεί την κανονική κατανομή, και επειδή η διασπορά  $\sigma^2$  του μητρικού πληθυσμού είναι γνωστή, ο έλεγχος μπορεί να γίνει με χρήση της τυποποιημένης μεταβλητής Z

$$\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right) \Rightarrow \frac{(\bar{X} - \mu_0)\sqrt{n}}{\sigma} = Z \sim N(0, 1)$$

Απορρίπτουμε την μηδενική υπόθεση  $H_0: \mu = \mu_0$ , έναντι της εκάστοτε εναλλακτικής υπόθεσης  $H_1$

- **Μονόπλευροι έλεγχοι**

$$H_1: \mu > \mu_0, \quad \bar{x} \geq \mu_0 + \frac{\sigma}{\sqrt{n}} z_\alpha$$

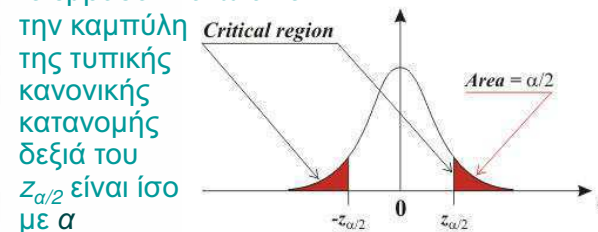
$$H_1: \mu < \mu_0, \quad \bar{x} \leq \mu_0 - \frac{\sigma}{\sqrt{n}} z_\alpha$$

- **Αμφίπλευρος έλεγχος**

$$\bar{x} \leq \mu_0 - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \quad \text{ή} \quad \bar{x} \geq \mu_0 + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$$

## z-test μέσης τιμής - Αμφίπλευρος έλεγχος

- Η μηδενική υπόθεση  $H_0$  απορρίπτεται σε επίπεδο σημαντικότητας  $\alpha$ , εάν ισχύει  $z > z_{\alpha/2}$  ή  $z < -z_{\alpha/2}$ , όπου η τιμή  $z_{\alpha/2}$  είναι τέτοια ώστε το εμβαδόν κάτω από



την καμπύλη της τυπικής κανονικής κατανομής δεξιά του  $z_{\alpha/2}$  είναι ίσο με  $\alpha$

## Έλεγχος της μέσης τιμής

Αν ο πληθυσμός ακολουθεί **άγνωστη κατανομή με άγνωστη διασπορά**

- και το μέγεθος του δείγματος είναι μεγάλο (πρακτικά  $n \gg 30$ , ιδανικά  $n \rightarrow \infty$ ) →  $\sigma^2 \approx S^2$
- ... η στατιστική συνάρτηση ελέγχου  $T$  μπορεί να θεωρηθεί ότι προσεγγίζεται από την τυποποιημένη κατανομή, κάνοντας χρήση της διακύμανσης του δείγματος  $S^2$ :

$$T = \frac{(\bar{X} - \mu_0)\sqrt{n}}{S} \rightarrow Z \sim N(0, 1)$$

Απορρίπτουμε την μηδενική υπόθεση  $H_0: \mu = \mu_0$ , έναντι της εκάστοτε εναλλακτικής υπόθεσης  $H_1$

### Αμφίπλευρος έλεγχος

$$H_1: \mu \neq \mu_0$$

### Μονόπλευροι έλεγχοι

$$H_1: \mu > \mu_0$$

$$H_1: \mu < \mu_0$$

$$H_0: \mu = \mu_0$$

$$Z = \frac{|\bar{X} - \mu_0|}{S/\sqrt{n}} \geq z_{\alpha/2}$$

$$Z = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \geq z_\alpha$$

$$Z = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \leq -z_\alpha$$

### Γνωστή τυπική απόκλιση

$H_1: m \neq \mu_0$ , αμφίπλευρος έλεγχος και περιοχή απόρριψης  $R$

$$R = \{z | z < z_{\alpha/2} \text{ \& } z > z_{1-\alpha/2}\} \quad \text{ή} \quad R = \{z | |z| > z_{1-\alpha/2}\}$$

$H_1: m > \mu_0$ , μονόπλευρος έλεγχος από αριστερά και περιοχή απόρριψης  $R$

$$R = \{z | z < z_\alpha\} = \{z | z < -z_{1-\alpha}\}$$

$H_1: m < \mu_0$ , μονόπλευρος έλεγχος από δεξιά και περιοχή απόρριψης  $R$

$$R = \{z | z > z_{1-\alpha}\}$$

$H_1: m \neq \mu_0$  (αμφίπλευρος έλεγχος), στατιστική τιμή ελέγχου ( $p$ -value) από το δείγμα

$$p = 2 P(z > |z_0|) = 2 (1 - P(z < |z_0|))$$

$H_1: m > \mu_0$ , μονόπλευρος έλεγχος από αριστερά και  $p$ -value

$$p = P(z < z_0) = 1 - P(z < -z_0)$$

$H_1: m < \mu_0$ , μονόπλευρος έλεγχος από δεξιά και  $p$ -value

$$p = P(z > z_0) = 1 - P(z < z_0)$$

$$z_0 = \frac{(\bar{x} - \mu_0)}{(\sigma\sqrt{n})}$$

## Έλεγχος της μέσης τιμής μικρού δείγματος

Αν ο πληθυσμός ακολουθεί **κανονική κατανομή με άγνωστη διασπορά** → **Έλεγχος t-test**

$$X_i \sim N(\mu_0, \sigma^2), i = 1, 2, \dots, n$$

- Χρησιμοποιείται η διακύμανση του δείγματος

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

- ... και η στατιστική συνάρτηση ελέγχου ακολουθεί την κατανομή Student

$$T = \frac{(\bar{X} - \mu_0)\sqrt{n}}{S} \sim t_{n-1}$$

Απορρίπτουμε την μηδενική υπόθεση  $H_0: \mu = \mu_0$ , έναντι της εκάστοτε εναλλακτικής υπόθεσης  $H_1$

- **Μονόπλευροι έλεγχοι**

$$H_1: \mu > \mu_0, \quad \bar{x} \geq \mu_0 + \frac{s}{\sqrt{n}} t_{n-1; \alpha}$$

$$H_1: \mu < \mu_0, \quad \bar{x} \leq \mu_0 - \frac{s}{\sqrt{n}} t_{n-1; \alpha}$$

- **Αμφίπλευρος έλεγχος**

$$\bar{x} \leq \mu_0 - \frac{s}{\sqrt{n}} t_{n-1; \alpha/2} \quad \text{ή} \quad \bar{x} \geq \mu_0 + \frac{s}{\sqrt{n}} t_{n-1; \alpha/2}$$

- Έστω ότι οι ακόλουθες 100 τιμές (σε %), με  $\mu=26.1$ ,  $\sigma^2=17.1$ , αντιπροσωπεύουν τον πληθυσμό των αποκλίσεων σε γωνιακές μετρήσεις ενός θεοδολίου

18.2	26.4	20.1	29.9	29.8	26.6	26.2
25.7	25.2	26.3	26.7	30.6	22.6	22.3
30.0	26.5	28.1	25.6	20.3	35.5	22.9
30.7	32.2	22.2	29.2	26.1	26.8	25.3
24.3	24.4	29.0	25.0	29.9	25.2	20.8
29.0	21.9	25.4	27.3	23.4	38.2	22.6
28.0	24.0	19.4	27.0	32.0	27.3	15.3
26.5	31.5	28.0	22.4	23.4	21.2	27.7
27.1	27.0	25.2	24.0	24.5	23.8	28.2
26.8	27.7	39.8	19.8	29.3	28.5	24.7
22.0	18.4	26.4	24.2	29.9	21.8	36.0
21.3	28.8	22.8	28.5	30.9	19.1	28.1
30.3	26.5	26.9	26.6	28.2	24.2	25.5
30.2	18.9	28.9	27.6	19.6	27.9	24.9
21.3	26.7					

- Πώς μπορούν να οργανωθούν με πιο ουσιαστικό τρόπο αυτά τα δεδομένα, ώστε να είναι περισσότερο κατανοητά;
- Είναι τα δεδομένα αντιπροσωπευτικά των αναγνώσεων που θα έπρεπε λογικά να αναμένονται από το χρησιμοποιούμενο όργανο και έναν έμπειρο παρατηρητή;
- Ποια στατιστικά εργαλεία μπορούν να χρησιμοποιηθούν, τα οποία να αναλύουν και να παρουσιάζουν καλύτερα αυτό το σύνολο δεδομένων;

- Μια γρήγορη ματιά για την ανάλυση των δεδομένων είναι να υπολογιστεί η διασπορά ή το εύρος των τιμών → μια ένδειξη της ακριβείας των δεδομένων.

18.2	26.4	20.1	29.9	29.8	26.6	26.2
25.7	25.2	26.3	26.7	30.6	22.6	22.3
30.0	26.5	28.1	25.6	20.3	35.5	22.9
30.7	32.2	22.2	29.2	26.1	26.8	25.3
24.3	24.4	29.0	25.0	29.9	25.2	20.8
29.0	21.9	25.4	27.3	23.4	38.2	22.6
28.0	24.0	19.4	27.0	32.0	27.3	15.3
26.5	31.5	28.0	22.4	23.4	21.2	27.7
27.1	27.0	25.2	24.0	24.5	23.8	28.2
26.8	27.7	39.8	19.8	29.3	28.5	24.7
22.0	18.4	26.4	24.2	29.9	21.8	36.0
21.3	28.8	22.8	28.5	30.9	19.1	28.1
30.3	26.5	26.9	26.6	28.2	24.2	25.5
30.2	18.9	28.9	27.6	19.6	27.9	24.9
21.3	26.7					

```
> mean(x)
[1] 26.069
> var(x)
[1] 17.46438
> sd(x)
[1] 4.179041
> range(x)
[1] 15.3 39.8
```

- Συλλέγοντας τυχαία δείγματα (ίδιου ή διαφορετικού πλήθους στοιχείων), μπορεί να προσδιοριστεί μια εκτίμηση του μέσου όρου και της διακύμανσης του πληθυσμού.

No.	$\bar{y}$	$S^2$
10	26.9	28.1
20	25.9	21.9
30	25.9	20.0
40	26.5	18.6
50	26.6	20.0
60	26.4	17.6
70	26.3	17.1
80	26.3	18.4
90	26.3	17.8
100	26.1	17.5

- Ωστόσο, δεν πρέπει να αναμένεται ότι αυτές οι εκτιμήσεις (από τα δείγματα) θα ταιριάζουν ακριβώς με τον μέσο όρο και τη διακύμανση του γονικού πληθυσμού.
  - Από τη στιγμή που ο μέσος όρος και η απόκλιση του εκάστοτε δείγματος υπολογίζονται από τυχαίες μεταβλητές (καθόσον οι μετρήσεις εμπεριέχουν τυχαία σφάλματα), είναι επίσης τυχαίες μεταβλητές.

Set 1: 29.9, 18.2, 30.7, 24.4, 36.0, 25.6, 26.5, 29.9, 19.6, 27.9  $\bar{y} = 26.9, S^2 = 28.1$   
 Set 2: 26.9, 28.1, 29.2, 26.2, 30.0, 27.1, 26.5, 30.6, 28.5, 25.5  $\bar{y} = 27.9, S^2 = 2.9$   
 Set 3: 32.2, 22.2, 23.4, 27.9, 27.0, 28.9, 22.6, 27.7, 30.6, 26.9  $\bar{y} = 26.9, S^2 = 10.9$   
 Set 4: 24.2, 36.0, 18.2, 24.3, 24.0, 28.9, 28.8, 30.2, 28.1, 29.0  $\bar{y} = 27.2, S^2 = 23.0$

- Οι διακυμάνσεις των μέσων και των διακυμάνσεων από διαφορετικά δείγματα ίδιου μεγέθους, θέτει ερωτήματα σχετικά με την ικανότητα των τιμών τους να εκτιμήσουν αξιόπιστα τις τιμές του πληθυσμού ενδιαφέροντος, π.χ.
  - Μεγαλύτερη εμπιστοσύνη στο 2<sup>ο</sup> δείγμα, λόγω της μικρότερης διακύμανσης των τιμών του;
  - Μεγαλύτερη εμπιστοσύνη στο 1<sup>ο</sup> ή στο 3<sup>ο</sup> δείγμα, λόγω της εγγύτητας του μέσου όρου τους προς τη μέση τιμή 26.1 του πληθυσμού;

## Πως θα εξυπηρετούσαν κατάλληλοι έλεγχοι υποθέσεων;

- Ποιοι έλεγχοι υποθέσεων μπορούν να χρησιμοποιηθούν για την περιγραφή ή την ποσοτικοποίηση της αξιοπιστίας των εκτιμήσεων μέσου όρου και διακύμανσης από κάποιο δείγμα;
- Με την εφαρμογή τους, θα μπορούσε να διατυπωθούν παραδοχές για την αξιοπιστία σε οποιοδήποτε δεδομένο επίπεδο εμπιστοσύνης των υπολογισμένων εκτιμήσεων των παραμέτρων ενός πληθυσμού.

```
set2 = c(26.9, 28.1, 29.2, 26.2, 30.0, 27.1, 26.5, 30.6, 28.5, 25.5)
summary(set2)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 25.50  26.60   27.60   27.86  29.02   30.60

sd(set2)
[1] 1.700457
alpha = 0.05 ; n = 10
t.half.alpha = qt(1-(alpha/2), df=n-1) ; t.half.alpha
[1] 2.262157
mu0 = 26.069 # Population Mean
# One-sample t-test
res <- t.test(set2, mu = mu0)
# Printing the results
res

One Sample t-test

data:  set2
t = 3.3307, df = 9, p-value = 0.00879
alternative hypothesis: true mean is not equal to 26.069
95 percent confidence interval:
 26.64357 29.07643
sample estimates:
mean of x
 27.86
```

### Παράδειγμα 2<sup>ο</sup> δείγμα

Set 1: 29.9, 18.2, 30.7, 24.4, 36.0, 25.6, 26.5, 29.9, 19.6, 27.9  $\bar{y} = 26.9, S^2 = 28.1$   
 Set 2: 26.9, 28.1, 29.2, 26.2, 30.0, 27.1, 26.5, 30.6, 28.5, 25.5  $\bar{y} = 27.9, S^2 = 2.9$   
 Set 3: 32.2, 22.2, 23.4, 27.9, 27.0, 28.9, 22.6, 27.7, 30.6, 26.9  $\bar{y} = 26.9, S^2 = 10.9$   
 Set 4: 24.2, 36.0, 18.2, 24.3, 24.0, 28.9, 28.8, 30.2, 28.1, 29.0  $\bar{y} = 27.2, S^2 = 23.0$

- Ερμηνεία για το 2<sup>ο</sup> δείγμα (με μέση τιμή δείγματος  $\bar{y} = 27.9$  και τυπική απόκλιση  $S = \sqrt{2.9} = 1.703, n = 10$ ), το επίπεδο εμπιστοσύνης 1- $\alpha$  είναι 0.95 →  $\alpha = 0.05$ 

$$t = \frac{z}{\sqrt{\chi^2/v}} = \frac{(\bar{y} - \mu)/(\sigma/\sqrt{n})}{\sqrt{(vS^2/\sigma^2)/v}} = \frac{(\bar{y} - \mu)/(\sigma/\sqrt{n})}{S/\sigma} = \frac{\bar{y} - \mu}{S/\sqrt{n}}$$

$$t = (27.86 - 26.069) / (1.7004 / \sqrt{10}) = 3.3307 \text{ (από το R)}$$

$$\approx (27.9 - 26.1) / (1.703 / \sqrt{10}) = 3.342$$

```
alpha = .05
n=10
t.half.alpha = qt(1-alpha/2, df=n-1)
c(-t.half.alpha, t.half.alpha)
[1] -2.262157 2.262157
```

- Το στατιστικό στοιχείο ελέγχου  $t = 3.3307$  βρίσκεται ΕΚΤΟΣ των κρίσιμων τιμών  $-2.262$  και  $2.262$  → και από το t.test (... p<a), σε επίπεδο σημαντικότητας 0.05, απορρίπτουμε τη μηδενική υπόθεση ότι η μέση δειγματική τιμή δεν διαφέρει από τη μέση τιμή του πληθυσμού (από τη βαθμονόμηση του οργάνου).
- Το διάστημα εμπιστοσύνης για τη μέση τιμή του πληθυσμού είναι
 
$$\bar{y} - t_{\alpha/2} \frac{S}{\sqrt{n}} < \mu < \bar{y} + t_{\alpha/2} \frac{S}{\sqrt{n}}$$

$$27.86 - 1.216 < \mu < 27.86 + 1.216 \rightarrow 26.644 < \mu < 29.076$$

```
set3 = c(32.2, 22.2, 23.4, 27.9, 27.0, 28.9, 22.6, 27.7, 30.6, 26.9)
summary(set3)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 22.20  24.27   27.35   26.94   28.65   32.20
sd(set3)
[1] 3.337398
alpha = 0.05 ; n = 10
t.half.alpha = qt(1-(alpha/2), df=n-1) ; t.half.alpha
[1] 2.262157
mu0 = 26.069 # Population Mean
# One-sample t-test
res <- t.test(set2, mu = mu0)
# Printing the results
res

One Sample t-test

data: set3
t = 0.8253, df = 9, p-value = 0.4305
alternative hypothesis: true mean is not equal to 26.069
95 percent confidence interval:
 24.55257 29.32743
sample estimates:
mean of x
 26.94
```

### Παράδειγμα 3<sup>ο</sup> δείγμα

Set 1: 29.9, 18.2, 30.7, 24.4, 36.0, 25.6, 26.5, 29.9, 19.6, 27.9  $\bar{y} = 26.9, S^2 = 28.1$   
 Set 2: 26.9, 28.1, 29.2, 26.2, 30.0, 27.1, 26.5, 30.6, 28.5, 25.5  $\bar{y} = 27.9, S^2 = 2.9$   
 Set 3: 32.2, 22.2, 23.4, 27.9, 27.0, 28.9, 22.6, 27.7, 30.6, 26.9  $\bar{y} = 26.9, S^2 = 10.9$   
 Set 4: 24.2, 36.0, 18.2, 24.3, 24.0, 28.9, 28.8, 30.2, 28.1, 29.0  $\bar{y} = 27.2, S^2 = 23.0$

- **Ερμηνεία για το 3<sup>ο</sup> δείγμα** (με μέση τιμή δείγματος  $ybar = 26.9$  και τυπική απόκλιση  $S = \sqrt{10.9} = 3.301, n = 10$ ), το επίπεδο εμπιστοσύνης 1-α είναι 0.95  $\rightarrow a = 0.05$

$$t = \frac{z}{\sqrt{\chi^2/v}} = \frac{(\bar{y} - \mu)/(\sigma/\sqrt{n})}{\sqrt{(vS^2/\sigma^2)/v}} = \frac{(\bar{y} - \mu)/(\sigma/\sqrt{n})}{S/\sigma} = \frac{\bar{y} - \mu}{S/\sqrt{n}}$$

- $t = (26.94 - 26.069) / (3.337 / \sqrt{10}) = 0.8253$  (από το R)  
 $\approx (26.9 - 26.1) / (3.301 / \sqrt{10}) = 0.766$

```
alpha = .05
n=10
t.half.alpha = qt(1-alpha/2, df=n-1)
c(-t.half.alpha, t.half.alpha)
[1] -2.262157 2.262157
```

- Το στατιστικό στοιχείο ελέγχου  $t = 0.8$ . βρίσκεται **ΕΝΤΟΣ** των κρίσιμων τιμών  $-2.262$  και  $2.262 \rightarrow$  και από το  $t.test (... p > a)$ , σε επίπεδο σημαντικότητας 0.05, δεχόμαστε τη μηδενική υπόθεση ότι το μέση δειγματική τιμή δεν διαφέρει από τη μέση τιμή του πληθυσμού (από τη βαθμολόγηση του οργάνου).
- Το διάστημα εμπιστοσύνης για τη μέση τιμή του πληθυσμού είναι

$$\bar{y} - t_{\alpha/2} \frac{S}{\sqrt{n}} < \mu < \bar{y} + t_{\alpha/2} \frac{S}{\sqrt{n}}$$

$$26.9 - 2.387 < \mu < 26.9 + 2.387 \rightarrow 24.552 < \mu < 29.287$$

```
set1 = c(29.9, 18.2, 30.7, 24.4, 36.0, 25.6, 26.5, 29.9, 19.6, 27.9)
summary(set1)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 18.20  24.70   27.20   26.87   29.90   36.00
sd(set1)
[1] 5.305354
alpha = 0.05 ; n = 10
t.half.alpha = qt(1-(alpha/2), df=n-1)
t.half.alpha
[1] 2.262157
# One-sample t-test
res <- t.test(set1, mu = mu0)
# Printing the results
res

One Sample t-test

data: set1
t = 0.47744, df = 9, p-value = 0.6444
alternative hypothesis: true mean is not equal to 26.069
95 percent confidence interval:
 23.07478 30.66522
sample estimates:
mean of x
 26.87
```

### Παράδειγμα 1<sup>ο</sup> δείγμα

Set 1: 29.9, 18.2, 30.7, 24.4, 36.0, 25.6, 26.5, 29.9, 19.6, 27.9  $\bar{y} = 26.9, S^2 = 28.1$   
 Set 2: 26.9, 28.1, 29.2, 26.2, 30.0, 27.1, 26.5, 30.6, 28.5, 25.5  $\bar{y} = 27.9, S^2 = 2.9$   
 Set 3: 32.2, 22.2, 23.4, 27.9, 27.0, 28.9, 22.6, 27.7, 30.6, 26.9  $\bar{y} = 26.9, S^2 = 10.9$   
 Set 4: 24.2, 36.0, 18.2, 24.3, 24.0, 28.9, 28.8, 30.2, 28.1, 29.0  $\bar{y} = 27.2, S^2 = 23.0$

- **Ερμηνεία για το 1<sup>ο</sup> δείγμα** (με μέση τιμή δείγματος  $ybar = 26.9$  και τυπική απόκλιση  $S = \sqrt{28.1} = 5.301, n = 10$ ), το επίπεδο εμπιστοσύνης 1-α είναι 0.95  $\rightarrow a = 0.05$

$$t = \frac{z}{\sqrt{\chi^2/v}} = \frac{(\bar{y} - \mu)/(\sigma/\sqrt{n})}{\sqrt{(vS^2/\sigma^2)/v}} = \frac{(\bar{y} - \mu)/(\sigma/\sqrt{n})}{S/\sigma} = \frac{\bar{y} - \mu}{S/\sqrt{n}}$$

- $t = (26.9 - 26.1) / (5.301 / \sqrt{10}) = 0.477$

```
alpha = .05
n=10
t.half.alpha = qt(1-alpha/2, df=n-1)
c(-t.half.alpha, t.half.alpha)
[1] -2.262157 2.262157
```

- Το στατιστικό στοιχείο ελέγχου  $t = 0.477$  βρίσκεται **ΕΝΤΟΣ** των κρίσιμων τιμών  $-2.262$  και  $2.262 \rightarrow$  και από το  $t.test (... p > a)$ , σε επίπεδο σημαντικότητας 0.05, δεχόμαστε τη μηδενική υπόθεση ότι το μέση δειγματική τιμή δεν διαφέρει από τη μέση τιμή του πληθυσμού (από τη βαθμολόγηση του οργάνου).
- Το διάστημα εμπιστοσύνης για τη μέση τιμή του πληθυσμού είναι

$$\bar{y} - t_{\alpha/2} \frac{S}{\sqrt{n}} < \mu < \bar{y} + t_{\alpha/2} \frac{S}{\sqrt{n}}$$

$$26.87 - 3.795 < \mu < 26.9 + 3.795 \rightarrow 23.075 < \mu < 30.665$$

### Παράδειγμα: Έλεγχος t-test στο R

1.  $H_0 : m = \mu$   $H_a : m \neq \mu$  (different)
2.  $H_0 : m \leq \mu$   $H_a : m > \mu$  (greater)
3.  $H_0 : m \geq \mu$   $H_a : m < \mu$  (less)

- Κάνουμε σύγκριση του μέσου  $m$  ενός μικρού δείγματος ( $n = 10$ ) από έναν πληθυσμό με κάποια καθορισμένη μέση τιμή  $\mu$ : π.χ., εάν τα επίπεδα αλατότητας σε ένα θαλάσσιο ενυδρείο ή μια λιμνοθάλασσα πληρούν ένα επίπεδο κατάστασης π.χ. 37.4 ppt (parts per thousand ή gr άλατος ανά kg θαλάσσιου νερού).

# An example data set containing the salinity observations (in ppt) in a lake  
 # We want to know, if the average salinity of the lake differs from 34.7 ppt?

```
my_data <- data.frame(
  name = paste0(rep("sal_", 10), 1:10),
  salinity = round(rnorm(10, 35, 4), 1)
)

> my_data
> class(my_data)
[1] "data.frame"
> str(my_data)
'data.frame': 10 obs. of 2 variables:
 $ name : Factor w/ 10 levels "sal_1","sal_10",...: 1 3 4 5 6 7 8 9 10 2
 $ salinity: num 30.2 36.1 39.3 25.6 36.7 37 32.7 32.8 32.7 31.4
>
> summary(my_data$salinity)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 25.60  31.73  32.75  33.45  36.55  39.30
```



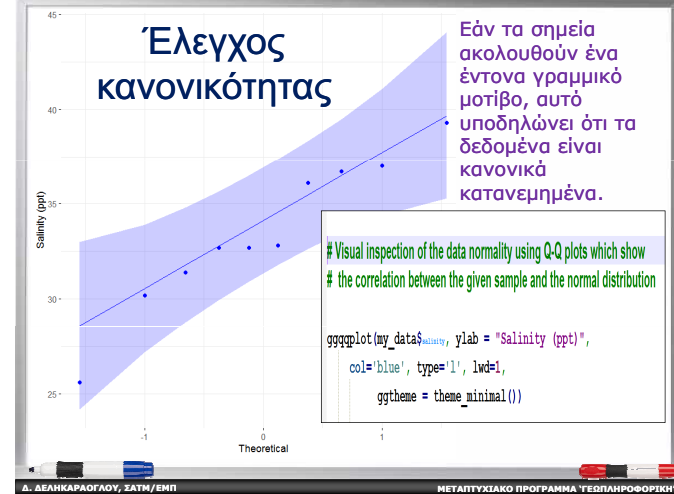
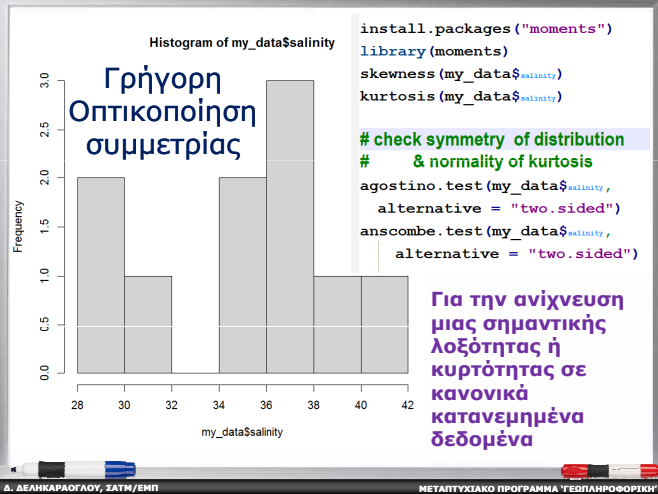
### Γρήγορη Οπτικοποίηση δεδομένων

```
install.packages("ggpubr")
library(ggpubr)

# Visualize the data
> ggboxplot(my_data$salinity,
+ ylab = "Salinity (ppt)", xlab = FALSE,
+ ggtheme = theme_minimal())
```







```
> # Kolmogorov - Smirnov normality test (usually used for n ≥ 50)
> ks.test(my_data$salinity, "pnorm", mean(my_data$salinity), sd(my_data$salinity))

One-sample Kolmogorov-Smirnov test

data: my_data$salinity
D = 0.16039, p-value = 0.9245
alternative hypothesis: two-sided

> # Shapiro - Wilk normality test (more appropriate for small (<50) samples)
> # if p is less than the significance level alpha = 0.05
> # then the sample mean is significantly different from assumed value
> shapiro.test(my_data$salinity)

Shapiro-Wilk normality test

data: my_data$salinity
W = 0.92447, p-value = 0.3958
```

Πιθανότητα εσφαλμένης απόρριψης της  $H_0$   
Εάν  $p\text{-value} < \alpha$ , Reject  $H_0$

```
> res <- t.test(my_data$salinity, mu = 34.7) # One-sample t-test
> res

One Sample t-test

data: my_data$salinity
t = -0.99394, df = 9, p-value = 0.3462
alternative hypothesis: true mean is not equal to 34.7
95 percent confidence interval:
 30.60506 36.29494
sample estimates:
mean of x
 33.45

> # Access to the values returned by t.test() function
> res$p.value # printing the p-value
[1] 0.3462219
> res$estimate # printing the mean
mean of x
 33.45
> res$conf.int # printing the confidence interval
[1] 30.60506 36.29494
attr(,"conf.level")
[1] 0.95
```

$$\bar{x} - t_{\alpha/2} \frac{S}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2} \frac{S}{\sqrt{n}}$$

Outliers? Sample size?

```
> # to test whether the sample mean is less than the assumed value
> t.test(my_data$salinity, mu = 34.7, alternative = "less")

One Sample t-test

data: my_data$salinity
t = -0.99394, df = 9, p-value = 0.1731
alternative hypothesis: true mean is less than 34.7
95 percent confidence interval:
 -Inf 33.75536
sample estimates:
mean of x
 33.45

> # to test whether the sample mean is greater than the assumed value
> t.test(my_data$salinity, mu = 34.7, alternative = "greater")

One Sample t-test

data: my_data$salinity
t = -0.99394, df = 9, p-value = 0.8269
alternative hypothesis: true mean is greater than 34.7
95 percent confidence interval:
 31.14464 Inf
sample estimates:
mean of x
 33.45
```

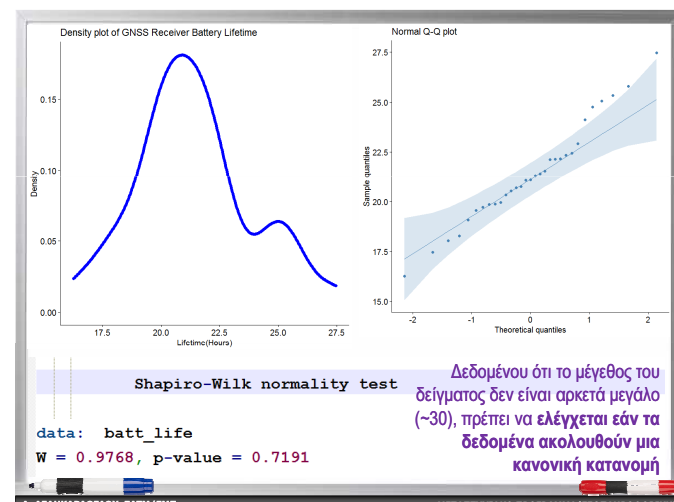
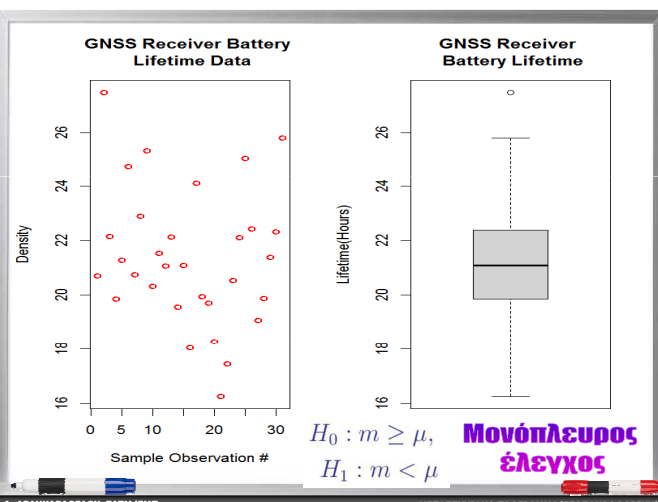
**One-sample t-test** : Δείγμα 31 παρατηρήσεων

• Παράδειγμα:  
Ισχυρισμός ότι οι μπαταρίες κάποιων δεκτών GNSS διαρκούν για τουλάχιστον 25 ώρες συνεχούς λειτουργίας ( $\mu_0 = 25$ );

```
> batt_life # GNSS receiver battery lifetime sample data
[1] 20.70 27.46 22.15 19.85 21.29 24.75 20.75 22.91 25.34 20.33 21.54 21.08
[13] 22.14 19.56 21.10 18.04 24.12 19.95 19.72 18.28 16.26 17.46 20.53 22.12
[25] 25.06 22.44 19.08 19.88 21.39 22.33 25.79
```

$H_0 : m \geq \mu_0$   
 $H_1 : m < \mu_0$

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
batt_life	16.26	19.86	21.10	21.40	22.39	27.46



```
# Assess the normality of the data USING Normality test
shapiro.test(batt_life)

Shapiro-Wilk normality test

data: batt_life
W = 0.9768, p-value = 0.7191

# One-sample t-test, using the R base function
t.test(batt_life, alternative = "less", mu = mu_test)

One Sample t-test

data: batt_life
t = -7.8861, df = 30, p-value = 4.211e-09
alternative hypothesis: true mean is less than 25
95 percent confidence interval:
 -Inf 22.17479
sample estimates:
mean of x
 21.4
```

Καθώς  $p > \alpha$  η κατανομή των δεδομένων δεν διαφέρει σημαντικά από την κανονική κατανομή → Μπορούμε να υποθέσουμε την κανονικότητα

Η τιμή  $p$  της δοκιμής είναι πολύ μικρότερη από το επίπεδο σημαντικότητας  $\alpha = 0.05$ . Μπορούμε να συμπεράνουμε ότι η μέση διάρκεια των μπαταριών δεκτών GNSS διαφέρει σημαντικά από τις 25 h → η  $H_0$  δεν είναι λογική και απορρίπτεται

• **Γενικευμένη περίπτωση** για τους μέσους δύο κανονικών πληθυσμών, π.χ. όταν δύο Τοπογράφοι μετρούν την απόσταση δύο σημείων,  $n_1$  και  $n_2$  φορές αντίστοιχα, και οι υπολογισμένες μέσες τιμές είναι  $\mu_1$  και  $\mu_2$



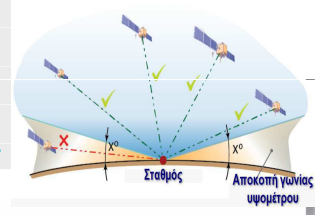
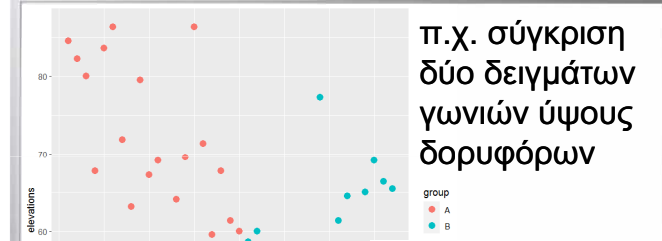
t-test της μέσης τιμής δύο δειγμάτων

- Διακρίνουμε δύο περιπτώσεις: τα δείγματα
  - είναι ανεξάρτητα (*independent*), π.χ. με δύο διαφορετικά αποστασιόμετρα, ή
  - προέρχονται από συνδυασμό (*paired*), π.χ. με το ίδιο αποστασιόμετρο πριν και μετά από βαθμονόμηση

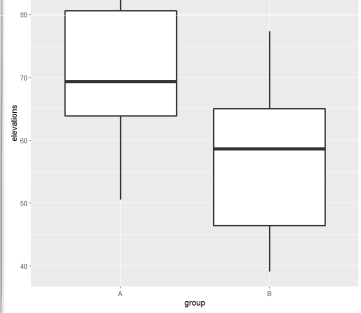
## Δύο δείγματα: Έλεγχος t-test στο R

1.  $H_0 : m_A = m_B$        $H_a : m_A \neq m_B$  (different)
2.  $H_0 : m_A \leq m_B$        $H_a : m_A > m_B$  (greater)
3.  $H_0 : m_A \geq m_B$        $H_a : m_A < m_B$  (less)

- Κάνουμε σύγκριση μεταξύ τους των μέσων δύο ανεξάρτητων δειγμάτων από έναν πληθυσμό : π.χ., εάν οι συνθήκες ορατότητας δορυφόρων GNSS διαφέρουν σε δύο υποψήφιους μόνιμους σταθμούς, π.χ. εξ αιτίας διαφορετικών παρατηρούμενων γωνιών ύψους στον καθένα από τους σταθμούς

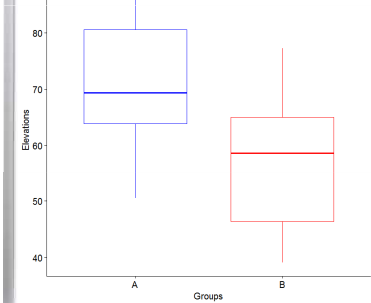


```
# Plot elevations by group and color by group using 'ggplot'
library("ggpubr")
ggboxplot(my_data, x = "group", y = "elevations",
           color = "group", palette = c("#00AFBB", "#E7B800"),
           ylab = "Elevations", xlab = "Groups")
```



Οπτικοποίηση δεδομένων (γωνίες ύψους), με χρήση της ggplot() :: πακέτο ggplot2

```
# Plot elevations by group and color by group using 'ggboxplot'
library("ggpubr")
ggboxplot(my_data, x = "group", y = "elevations",
           color = "group", palette = c("blue", "red"),
           ylab = "Elevations", xlab = "Groups")
```



Οπτικοποίηση δεδομένων (γωνίες ύψους), με χρήση της ggboxplot() :: πακέτο ggpubr

```
# If test reveals a p-value greater than 0.05, it indicates
# that there is no significant difference between the group variances

leveneTest(my_data$elevations, my_data$group, center = mean)
Levene's Test for Homogeneity of Variance (center = mean)
Df F value Pr(>F)
group 1 1.0457 0.3135
      35

# For the Shapiro-Wilk normality test, if the two p-values are greater
# than the significance level 0.05, then the distributions of each
# sample are not significantly different from the normal distribution

# Shapiro-Wilk normality test for site.A's elevations
with(my_data, shapiro.test(elevations[group == "A"]))

      Shapiro-Wilk normality test
data:  elevations[group == "A"]
W = 0.94386, p-value = 0.2833

# Shapiro-Wilk normality test for site.B's elevations
with(my_data, shapiro.test(elevations[group == "B"]))

      Shapiro-Wilk normality test
data:  elevations[group == "B"]
W = 0.93628, p-value = 0.2767
```

```
# Do the two populations have the same variances?
# If the p-value of the F-test is greater than the significance
# level alpha = 0.05, then there is no significant difference
# between the variances of the two sets of data
```

```
res.fstest <- var.test(elevations ~ group, data = my_data)
res.fstest

      F test to compare two variances

data:  elevations by group
F = 0.78943, num df = 19, denom df = 16, p-value = 0.6166
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2925966 2.0451499
sample estimates:
ratio of variances
0.7894339
```

```
# Is there any significant difference between the elevations
# at site.A and site.B?

# Compute t-test
res <- t.test(site.A, site.B, var.equal = TRUE)
res
```

```
Two Sample t-test

data:  site.A and site.B
t = 4.3455, df = 35, p-value = 0.0001137
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 8.326388 22.927453
sample estimates:
mean of x mean of y
71.29545 55.66853
```

```
# Compute independent t-test - Method 2: The data are saved in a data frame.
res <- t.test(elevations ~ group, data = my_data, var.equal = TRUE)
res
```

```
# Test whether the average site.A's elevations is less than
# the average average site.B's elevations:
```

```
t.test(elevations ~ group, data = my_data,
       var.equal = TRUE, alternative = "less")

      Two Sample t-test

data:  elevations by group
t = 4.3455, df = 35, p-value = 0.9999
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
 -Inf 21.70284
sample estimates:
mean in group A mean in group B
71.29545 55.66853

# Similarly, test whether the average site.A's elevations is greater than
# the average average site.B's elevations:
t.test(elevations ~ group, data = my_data,
       var.equal = TRUE, alternative = "greater")
```

## # Welch t-test (applied when variances are not statistically equal)

```
t.test(site.A, site.B, var.equal = FALSE)
```

Welch Two Sample t-test

```
data: site.A and site.B
t = 4.303, df = 32.394, p-value = 0.0001455
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 8.233109 23.020732
sample estimates:
mean of x mean of y
71.29545 55.66853
```

## # Non-parametric Wilcoxon t - test: applied when normality or equal variances are not evident from the data)

```
wilcox.test(site.A, site.B, exact = FALSE)
# or alternative statement
res <- wilcox.test(elevations ~ group, data = my_data, exact = FALSE)
res
```

Wilcoxon rank sum test with continuity correction

```
data: site.A and site.B
W = 285.5, p-value = 0.0004549
alternative hypothesis: true location shift is not equal to 0
```

```
# Print the p-value only
res$p.value
[1] 0.0004548981

# In all t-tests, if the p-value of the test is less than
# the significance level alpha = 0.05. We can conclude that
# the two sample means ARE significantly different (i.e. reject H0)
```

## Ζευγαρωτά δείγματα: t-test στο R

- $H_0 : m = 0$      $H_a : m \neq 0$  (different)
- $H_0 : m \leq 0$      $H_a : m > 0$  (greater)
- $H_0 : m \geq 0$      $H_a : m < 0$  (less)

- Κάνουμε σύγκριση των μέσων δύο σχετισμένων ομάδων δειγμάτων από έναν πληθυσμό : π.χ., εάν οι μετρήσεις αποστάσεων προς γνωστά σημεία με ένα αποστασιόμετρο ΠΡΙΝ και ΜΕΤΑ από μια βαθμονόμηση του οργάνου, έχουν επίδραση στην απόδοση της μετρητικής διαδικασίας

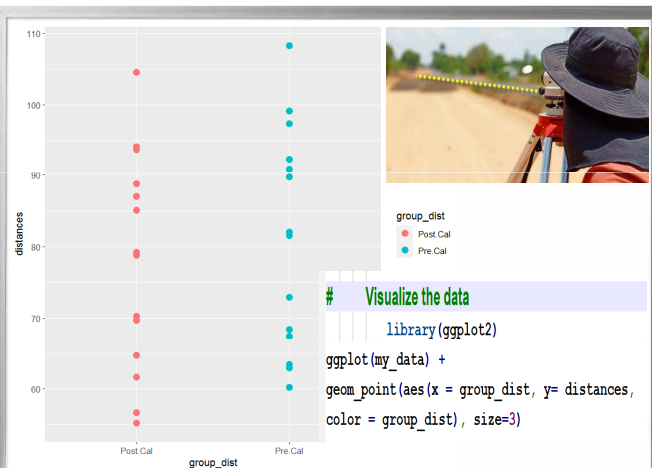
## # Data in two numeric vectors: Measured distances to various targets # Before and After Calibration of EDM instrument

```
pre_cal = c(89.8, 107.5, 105.7, 81.2, 99.3, 76.7, 100.7,
            75.7, 90.3, 105.7, 81.2, 71.7, 71.2, 98, 116.6, 68.5)
post_cal = c(87.1, 102.1, 102.5, 78, 97.1, 73, 95.3, 73,
            87.5, 102.5, 78.5, 69.9, 64.9, 93.4, 112.9, 63.5)
```

```
> summary(pre_cal)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 68.50  76.45   90.05   89.99 101.95  116.60

> summary(post_cal)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 63.50  73.00   87.30   86.33  98.35  112.90
```

```
> summary(my_data)
  group_dist distances
Post.Cal :16  Min.   : 63.50
Pre.Cal  :16  1st Qu.: 75.03
           Median : 88.65
           Mean   : 88.16
           3rd Qu.:101.05
           Max.   :116.60
```



```
# Shapiro-Wilk normality test for the differences.
# If the p-value is greater than the significance level alpha=0.05, then
# the distribution of the differences (diff) are not significantly
# different from normal distribution
diff = pre_cal - post_cal
diff
[1] 2.7 5.4 3.2 3.2 2.2 3.7 5.4 2.7 2.8 3.2 2.7 1.8 6.3 4.6 3.7 5.0
shapiro.test(diff)

Shapiro-Wilk normality test
data: diff
W = 0.92461, p-value = 0.2
```

**Normality is confirmed**

```
# If the data are not normally distributed, can use the non
# parametric paired two-samples Wilcoxon test
res <- wilcox.test(pre_cal, post_cal, paired = TRUE)
res

Wilcoxon signed rank test with continuity correction
data: pre_cal and post_cal
V = 136, p-value = 0.0004772
alternative hypothesis: true location shift is not equal to 0
```

## # Execute the t-test (having confirmed the normality condition)

```
res <- t.test(pre_cal, post_cal, paired = TRUE)
res
```

Paired t-test

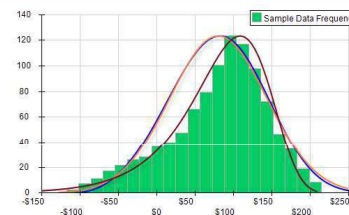
Υπάρχει σημαντική διαφορά στους πληθυσμούς ΠΡΙΝ και ΜΕΤΑ τη βαθμονόμηση?

```
data: pre_cal and post_cal
t = 11.25, df = 15, p-value = 1.039e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.968584 4.356416
sample estimates:
mean of the differences
 3.6625
```

**$p < \alpha \rightarrow$  reject  $H_0$**   
δηλ. υπάρχει σημαντική διαφορά ΠΡΙΝ και ΜΕΤΑ

```
# Compute t-test - 2nd method (must give the same result)
res <- t.test(distances ~ group_dist, data = my_data, paired = TRUE)
res
```

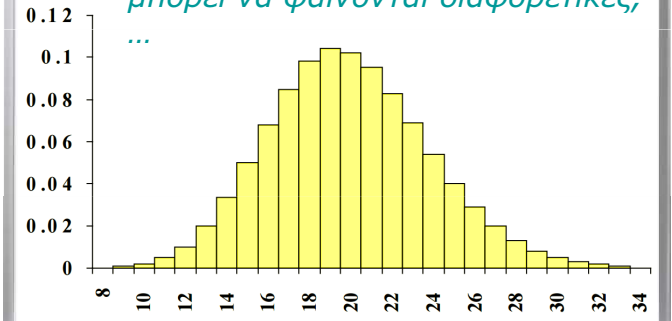
## ΣΤΑΤΙΣΤΙΚΟΙ έλεγχοι καλής προσαρμογής



- Αν έχουμε ένα δείγμα  $\{x_1, x_2, \dots, x_n\}$  που προέρχεται από κάποια (άγνωστη) κατανομή πιθανότητας  $f(x)$   
**Πως μπορεί να προσδιοριστεί αυτή?**  
Όχι διαισθητικά άλλα ποσοτικά από ένα δείγμα που προέρχεται από αυτήν, σύμφωνα με κάποιο κριτήριο

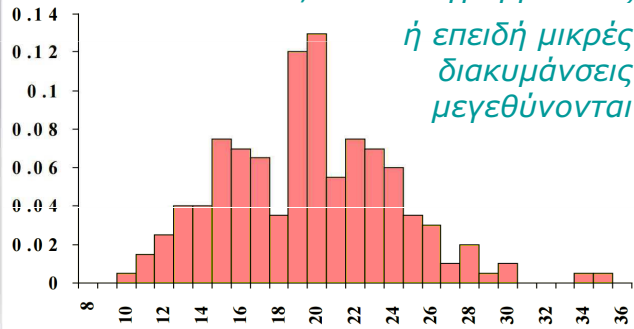
## Ένα σύννηθες πρόβλημα

- Κατανομές που είναι σχεδόν ίδιες μπορεί να φαίνονται διαφορετικές,



... ανάλογα με την επιλογή των κλάσεων, στο ιστόγραμμα τους

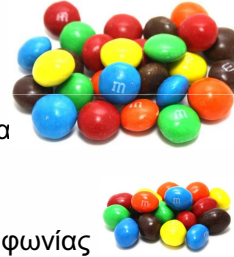
ή επειδή μικρές διακυμάνσεις μεγεθύνονται



## Στατιστικοί έλεγχοι καλής προσαρμογής

### • Goodness of fit tests

υποδεικνύουν πόσο καλά ένα σύνολο δεδομένων περιγράφεται ή προσαρμόζεται από ή σε ένα συγκεκριμένο μοντέλο (ή γονικό πληθυσμό)



- Μας δίνουν το βαθμό ασυμφωνίας ή το βαθμό εγγύτητας των παρατηρούμενων τιμών (*observed values*) από τις τιμές που αναμένονται (*expected values*) αν υιοθετήσουμε το υπό εξέταση μοντέλο

- Ένα μέτρο της απόκλισης μεταξύ των παρατηρούμενων και αναμενόμενων συχνοτήτων δίνεται από τη στατιστική συνάρτηση ελέγχου

$$\chi^2 = \sum_{i=1}^k \frac{(f_i^O - f_i^E)^2}{f_i^E}$$

- Η ποσότητα  $\chi^2$  μπορεί να χρησιμοποιηθεί ως κριτήριο για τη δοκιμή καλής προσαρμογής, με την προϋπόθεση ότι τα σφάλματα των μετρήσεων μπορεί να θεωρηθεί ότι ακολουθούν μια κανονική κατανομή.

$$\chi^2 = \sum_{i=1}^k \frac{(f_i^O - f_i^E)^2}{f_i^E}$$

- Εάν δεν υπάρχουν αποκλίσεις μεταξύ των  $f_i^O$  και  $f_i^E \rightarrow \chi^2 = 0$ , και
- Καθώς  $(f_i^O - f_i^E) \uparrow$ , τότε  $\chi^2 \uparrow$
- Η υπόθεση  $H_0$  δεν αληθεύει όταν  $\chi^2 > \chi^2_{k-1, \alpha}$ , με  $(k-1)$  βαθμούς ελευθερίας και σε επίπεδο σημαντικότητας  $\alpha$ ,  
• δηλ. ισχύει  $P(\chi^2 > \chi^2_{k-1, \alpha}) = \alpha$

## Στατιστικές δοκιμές κατανομής $\chi^2$ Chi-Square Tests

- Αφορούν δοκιμές ελέγχου για την εξέταση διαφορών με κατηγορικές μεταβλητές, σε δύο περιπτώσεις:
  1. Να ελεγχθεί πόσο στενά, μια παρατηρούμενη (από κάποια δεδομένα) κατανομή πιθανότητας ταιριάζει με μια αναμενόμενη κατανομή  $\rightarrow$  Goodness of fit test (δοκιμή 'καλής προσαρμογής') ή
  2. Για την εκτίμηση του κατά πόσον δύο τυχαίες μεταβλητές είναι ανεξάρτητες (Independence testing).

## Δοκιμές κατανομής $\chi^2$ στο R One Variance Chi-Square Tests

- Πρόκειται μια **μη παραμετρική στατιστική δοκιμή** για να ελεγχθεί, με τη χρήση της κατανομής  $\chi^2$ , εάν η διακύμανση μιας μεταβλητής που λαμβάνεται από ένα αντιπροσωπευτικό δείγμα έχει το ίδιο μέγεθος με τη γνωστή ή υποτιθέμενη διακύμανση του πληθυσμού της ίδιας μεταβλητής.
  - Χρησιμοποιεί ένα δείγμα  $n$  ανεξάρτητων κανονικά καταμεμημένων παρατηρήσεων

## Δοκιμές κατανομής $\chi^2$ στο R One Variance Chi-Square Tests

- Αφορούν τρεις διαφορετικούς τύπους δοκιμών ανάλογα με την εναλλακτική

$H_0: \sigma^2 = \sigma_0^2$  υπόθεση που επιλέγεται:  
 $H_a: \sigma^2 < \sigma_0^2$  for a lower one-tailed test  
 $\sigma^2 > \sigma_0^2$  for an upper one-tailed test  
 $\sigma^2 \neq \sigma_0^2$  for a two-tailed test

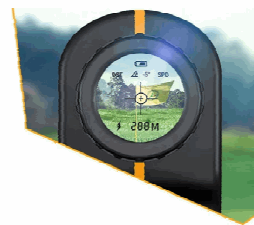
Test Statistic:  $T = \frac{(N-1)(S/\sigma_0)^2}{1}$  **Απόκλιση Δείγματος**  
**Πλήθος Δείγματος**

## Παράδειγμα: Δοκιμή νέου τοπογραφικού οργάνου μέτρησης γωνιών

- Από μια σειρά 30 γωνιομετρικών αναγνώσεων προς ένα στόχο, προέκυψε δείγμα με τυπική απόκλιση  $S = 1.0''$ .

Αναποκρίνεται το συγκεκριμένο όργανο στην προδιαγραφή του κατασκευαστή που δίνεται ως  $\sigma^2 = 1.5''$  σε επίπεδο σημαντικότητας  $\alpha = 0.05$ ;

- $H_0 : S^2 = \sigma^2$
- $H_1 : S^2 \neq \sigma^2$



## # Sample of readings for angular differences to known target

```
x
[1] -0.21859345  0.35722569  0.50158627  0.83737025  1.16426795 -0.03675963
[7] -0.22854185  0.38091994  0.02654751 -0.74957657 -0.58213109  2.22949269
[13]  1.22554667  0.50447188 -1.90382182  0.21540836 -1.88111417 -2.35414604
[19] -1.45410866  0.47009658 -0.08315382  0.80458580 -1.17828793  0.57817767
[25]  0.54666555 -0.27750839  0.27959900  0.23206248 -0.59530343  0.64093022
```

summary(x)

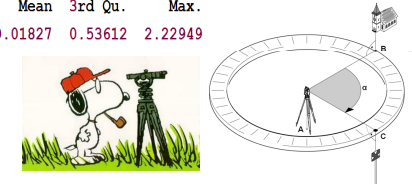
```
Min. 1st Qu.  Median    Mean 3rd Qu.  Max.
-2.35415 -0.50598  0.22374 -0.01827  0.53612  2.22949
```

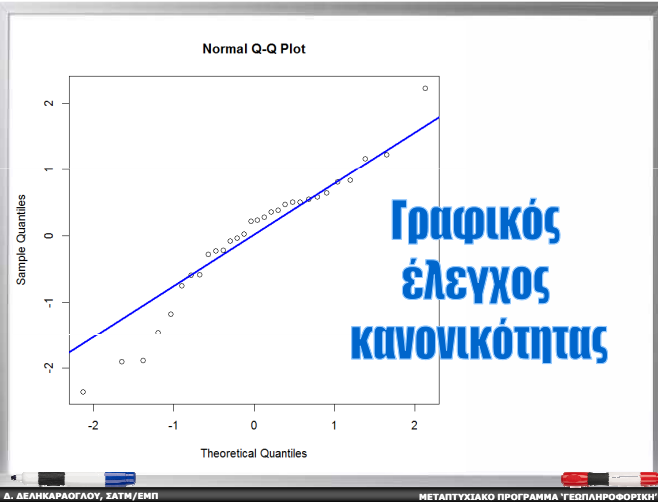
var(x)

```
[1] 1.001305
```

sd(x)

```
[1] 1.000652
```





```
# Perform a one sample chi-squared test to compare the
# variance of a sample with a given value
# If p-value < a, then reject H0 that s^2 (sample) = sigma^2 (population).
# Note the (1-a) confidence interval for the variance.

varTest(x, conf.level = 0.95, sigma.squared = 1.5^2, data.name = NULL)

One Sample Chi-Square test on variance
χ² = (n - 1)s² / σ²

data: x
X-squared = 12.906 df = 29, p-value = 0.008668
alternative hypothesis: true variance is not equal to 2.25
95 percent confidence interval:
 0.6350916 1.8095414
sample estimates:
variance of x
 1.001305
```

## Έλεγχος ανεξαρτησίας

- Ελέγχεται εάν τα δύο χαρακτηριστικά (κατηγορικές τυχαίες μεταβλητές) λειτουργούν από κοινού ή εάν υπάρχει ή όχι στατιστική εξάρτηση μεταξύ τους, δηλ.
  - εάν τα μέλη ενός πληθυσμού που έχουν (ή δεν έχουν) μια συγκεκριμένη τιμή για ένα χαρακτηριστικό, επίσης έχουν (ή δεν έχουν) μια συγκεκριμένη τιμή και για το άλλο
  - Είναι είναι ένα από τα πιο βασικά και κοινά τεστ ελέγχου υποθέσεων σε στατιστικές αναλύσεις κατηγοριοποιημένων δεδομένων.

## Έλεγχος ανεξαρτησίας

Μορφή του πίνακα συνάφειας

- $N_{jk}$  εκφράζει πόσες φορές εμφανίζεται κάθε ένα από τα ζεύγη  $(X_j, Y_k)$  σε ένα (υπό μελέτη) τυχαίο δείγμα

		Y				Σύνολο
		1	2	...	K	
X	1	$N_{11}$	$N_{12}$	...	$N_{1K}$	$N_{1\bullet}$
	2	$N_{21}$	$N_{22}$	...	$N_{2K}$	$N_{2\bullet}$
	...	...	...	...	...	...
	J	$N_{J1}$	$N_{J2}$	...	$N_{JK}$	$N_{J\bullet}$
Σύνολο		$N_{\bullet 1}$	$N_{\bullet 2}$	...	$N_{\bullet K}$	$N = N_{\bullet\bullet}$

- Η αναμενόμενη συχνότητα των παρατηρήσεων στο κελί  $(A_j, B_k)$  είναι το γινόμενο της από κοινού πιθανότητας επί το μέγεθος του δείγματος N:
 
$$f_{jk}^E = P(X_j) P(Y_k) N = \frac{f_{j\bullet}^O}{N} \frac{f_{\bullet k}^O}{N} N = \frac{f_{j\bullet}^O \cdot f_{\bullet k}^O}{N}$$
- Το στατιστικό  $\chi^2_{df, \alpha}$  βάσει του οποίου ενεργείται ο έλεγχος, σε επίπεδο σημαντικότητας  $\alpha$ , με βαθμούς ελευθερίας  $df = (j-1)(k-1)$ :
 
$$\chi^2 = \sum_1^j \sum_1^k \frac{(f_{jk}^O - f_{jk}^E)^2}{f_{jk}^E}$$

## Έλεγχος ανεξαρτησίας – Παράδειγμα

	Π1	Π2	Π3	Σύνολο
GPS II	3	3	2	8
GPS IIR	3	2	4	9
GPS IIR-M	2	4	3	9
Σύνολο	8	9	9	26

Παρατηρούμενος # δορυφόρων από τις περιοχές Π1, Π2, Π3

- Παρατηρούμενοι δορυφόροι GPS-II, -IIR, -IIR-M σε τρεις περιοχές της χώρας, από ένα δείγμα 26 δορυφόρων GPS
- Ο αριθμός των ορατών δορυφόρων GPS είναι ανεξάρτητος από την περιοχή της χώρας?

## Έλεγχος ανεξαρτησίας – Παράδειγμα

	Π1	Π2	Π3	Σύνολο
GPS II	3	3	2	8
GPS IIR	3	2	4	9
GPS IIR-M	2	4	3	9
Σύνολο	8	9	9	26

Παρατηρούμενος # δορυφόρων από τις περιοχές Π1, Π2, Π3

**Υποθέσεις να διερευνηθούν**

- $H_0$ : ο αριθμός των ορατών δορυφόρων είναι ανεξάρτητος της περιοχής των μετρήσεων
- $H_1$ : ο αριθμός των ορατών δορυφόρων εξαρτάται από την περιοχή

## Έλεγχος ανεξαρτησίας – Παράδειγμα

	Π1	Π2	Π3	Σύνολο
GPS II	2.46	2.77	2.77	8
GPS IIR	2.77	3.12	3.12	9
GPS IIR-M	2.77	3.12	3.12	9
Σύνολο	8	9	9	26

Αναμενόμενος # ορατών δορυφόρων από τις περιοχές Π1, Π2, Π3

Πίνακας αναμενόμενων συχνοτήτων:

- π.χ. η αναμενόμενη συχνότητα του κελιού (1,2) είναι

$$f_{12}^E = \frac{f_{1\bullet}^O \cdot f_{\bullet 2}^O}{N} = \frac{8 \times 9}{26} = 2.77$$

## Έλεγχος ανεξαρτησίας – Παράδειγμα

	Π1	Π2	Π3	Σύνολο
GPS II	2.46	2.77	2.77	8
GPS IIR	2.77	3.12	3.12	9
GPS IIR-M	2.77	3.12	3.12	9
Σύνολο	8	9	9	26

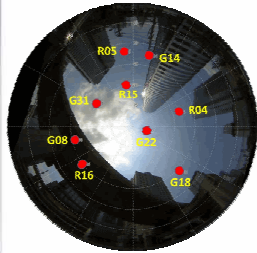
Αναμενόμενος # ορατών δορυφόρων από τις περιοχές Π1, Π2, Π3

$$\chi^2 = \frac{(3 - 2.46)^2}{2.46} + \dots + \frac{(3 - 3.12)^2}{3.12} = 1.49$$

... Με βάση τα δεδομένα του δείγματος, η  $H_0$  αληθεύει

$$\chi^2 = 1.49 < 13.3 = \chi^2_{df=4, \alpha=0.01}$$

# Έρευνα: Απόδοση των παγκόσμιων δορυφορικών συστημάτων πλοήγησης (GNSS) σε αστικά περιβάλλοντα



## Παράδειγμα:

Σύγκριση αποτελεσμάτων προσομοίωσης με δεδομένα παρατήρησης δορυφόρων GPS, BeiDou, Glonass

# # Simulation of GNSS Availability in Urban Environments

	PV	PNV	PVNT	PNVT
GPS	9607	1248	116	9982
BEIDOU	10017	2872	286	3313
GLONASS	2722	130	5086	644

Έρευνα για 4 ΕΦΙΚΤΕΣ ΠΕΡΙΠΤΩΣΕΙΣ:

- PV** – Predict Visible satellite | Receiver tracked satellite
- PNV** – Predict Non Visible satellite | Receiver did not track the satellite
- PVNT** – Predict Visible satellite | Receiver did Not Track the satellite
- PNVT** – Predict Non Visible satellite | Receiver Track the satellite

```

PV  PNV  PVNT  PNVT
GPS  9607 1248  116 9982
BEIDOU 10017 2872  286 3313
GLONASS 2722 130 5086  644

install.packages("gplots")
library(gplots)

# 1. convert the data as a table
dtable <- as.table(as.matrix(gnss.df))
dtable

# 2. Graphical display of contingency table
balloonplot(t(dtable), main="gnss.df", xlab="", ylab="",
            label=FALSE, show.margins=TRUE)
    
```

```

install.packages("graphics")
library(graphics)
# Visualize the contingency table as a mosaic plot
mosaicplot(dtable, shade=TRUE, las=2,
           main="gnss.df")

# Blue color : the observed value
# is higher than the expected value
# Red color : the observed value
# is lower than the expected value
    
```

	PV	PNV	PVNT	PNVT
GPS	9607	1248	116	9982
BEIDOU	10017	2872	286	3313
GLONASS	2722	130	5086	644

```

# Chi-square test whether rows and columns of the contingency table
# are statistically significantly associated.

# (H0): the row and the column variables of the contingency
# table are independent.
# (H1): row and column variables are dependent

chisq <- chisq.test(gnss.df)
chisq

# Pearson's Chi-squared test
data: gnss.df
X-squared = 26987, df = 6, p-value < 2.2e-16

chisq$statistic # the value the chi-squared test statistic.
chisq$parameter # the degrees of freedom
chisq$p.value # the p-value of the test
chisq$observed # the observed count
chisq$expected # the expected count
    
```

$$\chi^2 = \sum \frac{(o - e)^2}{e}$$

p < 0.05

```

chisq$observed # the observed count
PV  PNV  PVNT  PNVT
GPS  9607 1248  116 9982
BEIDOU 10017 2872  286 3313
GLONASS 2722 130 5086  644

chisq$expected # the expected count
PV  PNV  PVNT  PNVT
GPS  10173.516 1934.9075 2498.535 6346.041
BEIDOU 8005.581 1522.5865 1966.107 4993.726
GLONASS 4166.903 792.5059 1023.358 2599.233

round(chisq$residuals, 3) # Look at the residuals
PV  PNV  PVNT  PNVT
GPS  -5.617 -15.616 -47.665 45.642
BEIDOU 22.481 34.582 -37.891 -23.784
GLONASS -22.384 -23.534 126.997 -38.351
    
```

• Θετικά υπόλοιπα: συσχέτιση

• Αρνητικά υπόλοιπα: μη-συσχέτιση

• Τα τυποποιημένα υπόλοιπα των παρατηρήσεων είναι επίσης χρήσιμα για την ερμηνεία της σχέσης μεταξύ σειρών και στηλών (κατηγορικών μεταβλητών)

• Σχετική συνεισφορά κάθε κελιού στη συνολική τιμή του  $\chi^2 \rightarrow$  ένδειξη της εξάρτησης μεταξύ σειρών και στηλών του πίνακα

• Επιβεβαιώνει την προηγούμενη οπτική ερμηνεία των δεδομένων, η οποία μπορεί να είναι περίπλοκη όταν ο πίνακας συνάφειας είναι πολύ μεγάλος.

$$contrib <- 100 * chisq$residuals^2 / chisq$statistic$$

## Παράδειγμα

• Έστω ότι το σφάλμα ενός παρατηρητή στη σκόπευση με ένα θεοδόλιχο που παρέχει μετρήσεις με ακρίβεια 1",

• υπολογίζεται με τη συλλογή 20 αναγνώσεων προς έναν καλά καθορισμένο στόχο σε κάποια απόσταση.

• Η τυπική απόκλιση του δείγματος προσδιορίζεται ως  $S=1.8$ .

• Ποιο είναι το 95% διάστημα εμπιστοσύνης για τη διακύμανση  $\sigma^2$  του πληθυσμού?

## Διαστήματα εμπιστοσύνης για τη διακύμανση πληθυσμού

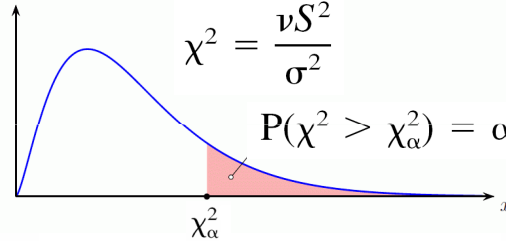
- Βασίζονται στη χρήση της κατανομής  $\chi^2 \rightarrow$

$$P(\chi^2 > \chi_\alpha^2) = \alpha$$

- Σε αντίθεση με την κανονική κατανομή και την κατανομή t (Student), η κατανομή  $\chi^2$  δεν είναι συμμετρική γύρω από το μηδέν  $\rightarrow$  απαιτείται ο υπολογισμός πιθανοτήτων της μορφής

$$P(\chi_{1-\alpha/2}^2 < \chi^2 < \chi_{\alpha/2}^2) = 1 - \alpha$$

pdf  $f(x)$



$$P(\chi_{1-\alpha/2}^2 < \chi^2 < \chi_{\alpha/2}^2) = 1 - \alpha$$

$$P\left(\chi_{1-\alpha/2}^2 < \frac{\nu S^2}{\sigma^2} < \chi_{\alpha/2}^2\right) = P\left(\frac{\chi_{1-\alpha/2}^2}{\nu S^2} < \frac{1}{\sigma^2} < \frac{\chi_{\alpha/2}^2}{\nu S^2}\right)$$

- Χρησιμοποιώντας την ιδιότητα των μαθηματικών ανισοτήτων ... δηλ., με τη λήψη του αντιστρόφου μιας συνάρτησης, η ανισότητα αντιστρέφεται ...

$$P\left(\frac{\chi_{1-\alpha/2}^2}{\nu S^2} < \frac{1}{\sigma^2} < \frac{\chi_{\alpha/2}^2}{\nu S^2}\right) \Leftrightarrow P\left(\frac{\nu S^2}{\chi_{\alpha/2}^2} < \sigma^2 < \frac{\nu S^2}{\chi_{1-\alpha/2}^2}\right) = 1 - \alpha$$

- συνεπώς, το 1-α διάστημα εμπιστοσύνης για τη διακύμανση του πληθυσμού είναι

$$\frac{\nu S^2}{\chi_{\alpha/2}^2} < \sigma^2 < \frac{\nu S^2}{\chi_{1-\alpha/2}^2}$$

```
> qchisq(.975, df=19, lower.tail = FALSE)
[1] 8.906516
> qchisq(.975, df=19)
[1] 32.85233
```

$$P\left(\frac{\nu S^2}{\chi_{\alpha/2}^2} < \sigma^2 < \frac{\nu S^2}{\chi_{1-\alpha/2}^2}\right) = 1 - \alpha$$

$$\frac{(20 - 1)1.8^2}{32.85} < \sigma^2 < \frac{(20 - 1)1.8^2}{8.91}$$

$$1.87 < \sigma^2 < 6.91$$



### Υπενθύμιση:

1. ΠΕΜΠΤΗ 02/02/2023, 11:00-14:00 - Τελική εξέταση, **On-line**
2. ΔΕΥΤΕΡΑ 06/02/2023, 12:00-15:00 - Παρουσίαση Ατομικών Θεμάτων Εξαμήνου, Μικρό Αμφ. Κτ. Λαμπαδάρου