10/1/2023

Αναλυτικές Μέθοδοι στη Γεωπληροφορική

Στατιστικά μέτρα τυχαίων μεταβλητών

Επόμενες ενότητες μαθήματος

- Στατιστικά μέτρα (παράμετροι) κατανομής τυχαίων μεταβλητών
  - Βέλτιστοι και Αμερόληπτοι Εκτιμητές
  - Εκτίμηση παραμέτρων
  - Σημειακές εκτιμήσεις
- Στοιχεία στατιστικής ανάλυσης
  - Διαστήματα εμπιστοσύνης
  - Έλεγχος υποθέσεων
- Υπολογισμοί μέσω του R

Στην έρευνα … δεδομένα συλλέγονται για να εξαχθούν πρότυπα, τάσεις ή άλλα σημαντικά συμπεράσματα

Ο κύκλος χρήσης των δεδομένων

Το πλαίσιο αναλύσεων δεδομένων

Ερευνητικό πρόβλημα ενδιαφέροντος → Ανάγκες / Σχεδιασμός συλλογής πρωτογενών δεδομένων

Μετρητικές διαδικασίες → Εμπεριέχουν σφάλματα και θόρυβο / Ετερογένεια και μεταβλητότητα στα δεδομένα

Οργάνωση δεδομένων / Επιλογή εργαλείων και τεχνικών ανάλυσης

Στατιστική επεξεργασία των δεδομένων → Στατιστικά μέτρα → Αξιολόγηση και συμπεράσματα

Προσήλωση στα δεδομένα

- *Συχνά ο σκοπός μιας ανάλυσης είναι να διερευνηθεί πως και κατά πόσο ένα προγνωστικό μοντέλο ταιριάζει στα δεδομένα.*
  - Η δημιουργία ενός τέτοιου μοντέλου μπορεί να είναι χρήσιμη για την εξαγωγή περαιτέρω συμπερασμάτων, νέου σχεδιασμού, ή δοκιμών με νέα δεδομένα και βελτίωση του μοντέλου

Διαγνωστική ανάλυση: Τι συνέβη μέχρι τώρα;

Προγνωστική ανάλυση: Τι μπορεί να συμβεί;

Προκαθοριστική ανάλυση: Τι πρέπει να γίνει μελλοντικά;

Γνωστική ανάλυση:Μοτίβα, αλληλεπιδράσεις, και συνδέσεις;

Στατιστική := Σύνολο αρχών και μεθοδολογιών

- Σχεδιασμός διαδικασιών συλλογής (ποιοτικών ή ποσοτικών) δεδομένων → *Experimental Design*
- Συνοπτική και αποτελεσματική παρουσίασή τους → Περιγραφική Στατιστική *(Descriptive Statistics)*
- Ανάλυση και εξαγωγή συμπερασμάτων → Επαγωγική Στατιστική *(Inferential Statistics)*

Συλλογή δεδομένων

- Ικανός αριθμός μετρήσεων ή παρατηρήσεων από δειγματικούς χώρους ενός πληθυσμού ενδιαφέροντος που συλλέγονται για αναφορά, πληροφορία και γνώση.
  - Εμπεριέχουν **σφάλματα** και **θόρυβο**

Οργάνωση και παρουσίαση δεδομένων

- Διερεύνηση της ταυτότητας και της μορφής των δεδομένων → Συνεισφέρουν στην επιλογή εργαλείων και τεχνικών ανάλυσης

Επεξεργασία δεδομένων

- Εξαγωγή συμπερασμάτων σχετικά με τα δεδομένα και τις υπό διερεύνηση παραμέτρους του εκάστοτε προβλήματος ενδιαφέροντος

Οι περισσότερες ερευνητικές μελέτες καταλήγουν σε **μεγάλο όγκο ακατέργαστων δεδομένων**

Αυτά είναι συχνά αναγκαίο να περιγραφούν κατάλληλα (συνοπτικά)

Να μελετηθούν με όλες τις λεπτομέρειες και τις προεκτάσεις τους

Αυτό μπορεί καταρχήν να επιτευχθεί εστιάζοντας σε διάφορα **στατιστικά μέτρα**

Αυτά πρέπει να "ταιριάζουν" στις εκάστοτε πληροφοριακές ανάγκες μιας έρευνας

Μέτρηση της ποικιλομορφίας και της ισοδυναμίας των αποτελεσμάτων από μετρήσεις

Παραμετρικά μοντέλα/μέθοδοι

- Τα συμπεράσματά τους βασίζονται σε ισχυρές υποθέσεις/παραδοχές για τα δεδομένα, τα οποία είναι στο επίκεντρο της ανάλυσης → *υπολογίζουμε τις παραμέτρους της κατανομής που υποτίθεται για τα δεδομένα* → παρέχουν εκτιμήσεις και διαστήματα εμπιστοσύνης και γενικεύονται σε πιο σύνθετες αναλύσεις

Μη-παραμετρικά μοντέλα/μέθοδοι

- π.χ. μέθοδοι κατάταξης ή δεδομένα από οπτικές αναλογικές κλίμακες: δεν κάνουν συγκεκριμένες υποθέσεις σχετικά με τη μορφή οποιωνδήποτε υποκείμενων κατανομών πιθανοτήτων

Στα πλαίσιο της Θεωρίας Εκτίμησης

- ***Εκτίμηση*** := Διαδικασία μέσα από την οποία συνάγονται συμπεράσματα σχετικά με ένα ή περισσότερα χαρακτηριστικά κάποιου πληθυσμού, με βάση τις πληροφορίες που λαμβάνονται από ένα τυχαίο δείγμα παρατηρήσεων που εμπεριέχουν σφάλματα και θόρυβο
- ***Εκτιμητής*** := Μια συνάρτηση που αντιστοιχίζει το χώρο του δείγματος, σε ένα σύνολο άγνωστων παραμέτρων μιας συνάρτησης κατανομής πιθανότητας και *παράγει μια εκτίμηση για τις υπό διερεύνηση παραμέτρους ενός προβλήματος* με ακρίβεια αντίστοιχη εκείνης των *θορυβωδών παρατηρήσεων* του δείγματος

'Ετερογένεια' στα δεδομένα;

- Στη στατιστική ο όρος χρησιμοποιείται για να υποδηλώσει ότι οι πληθυσμοί | τα δείγματα | τα αποτελέσματα αναλύσεων είναι **διαφορετικά**.
  - Τέτοιες διαφορές μπορούν να εμφανίζονται τόσο στον χώρο όσο και στον χρόνο
    - **χωρική ετερογένεια** (π.χ., για ένα στοιχείο μεταξύ διαφορετικών θέσεων)
    - **χρονική ετερογένεια** (π.χ., μεταξύ στοιχείων εντός της ίδιας θέσης | πολλαπλά, διαχρονικά δείγματα μετρήσεων)

Χωρική ετερογένεια;

- Αναφέρεται στην άνιση κατανομή (οριζόντια ή κατακόρυφα) ενός χαρακτηριστικού *(π.χ., θερμοκρασία, ρύπανση, ανεργία, εισόδημα, ...)* που μελετάται στο χώρο σε μια περιοχή
  - Η σχέση μεταξύ των μεταβλητών ενός υποδείγματος/μοντέλου δεν παραμένει σταθερή παντού σε μια γεωγραφική περιοχή αλλά μεταβάλλεται από σημείο σε σημείο
  - Χωρική ετερογένεια *(spatial heterogeneity)* + Χωρική εξάρτηση *(spatial dependence)* = **Χωρικές επιδράσεις** ***(spatial effects)***

Ετερογένεια | διαφορετικότητα;

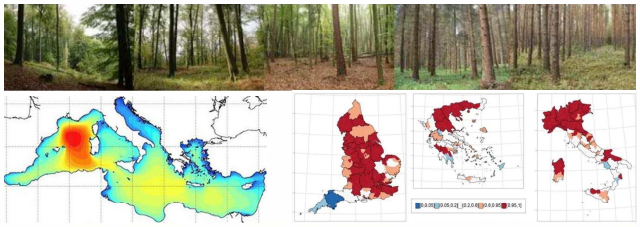
- Αυτό που ξεχωρίζει την έννοια της ετερογένειας από αυτή της διαφορετικότητας είναι η σχετική διάταξη των στοιχείων και όχι απλώς η εμφάνισή τους.
  - Η ετερογένεια μπορεί να εκληφθεί με διαφορετικούς τρόπους για διαφορετικούς σκοπούς, με αποτέλεσμα πολλαπλούς ορισμούς και λειτουργικοποιήσεις.
- Πρέπει να διακρίνουμε τη διακύμανση που είναι επιστημονικά ενδιαφέρουσα και την παραλλαγή που απλώς αντικατοπτρίζει την ετερογένεια του υποβάθρου των παρατηρήσεων μας

Χωρική ετερογένεια μπορεί να είναι

- είτε διακριτή *(discrete spatial heterogeneity)* : οι παράμετροι ενός υποδείγματος σε κάποιες συγκεκριμένες γεωγραφικές περιοχές έχουν διαφορετικές τιμές από ότι σε άλλες περιοχές : *π.χ., Αγροτικές | Αστικές περιοχές*
- είτε *συνεχής (continuous spatial heterogeneity)* : οι συντελεστές ενός υποδείγματος ποικίλλουν σε όλες τις περιοχές → κάθε παράμετρος του υποδείγματος θεωρείται ότι είναι συνάρτηση των γεωγραφικών συντεταγμένων.

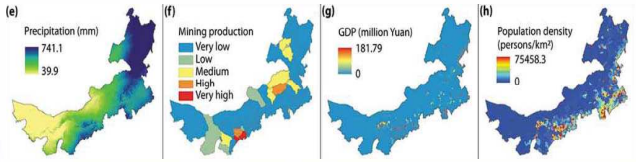
Χωρική ετερογένεια μπορεί να είναι

- **τοπική** *(spatial local heterogeneity)* : αναφέρεται σε φαινόμενα όπου *η τιμή ενός χαρακτηριστικού σε μια τοποθεσία* είναι διαφορετική από τη γύρω περιοχή, π.χ., *hotspots της χερσαίας ή θαλάσσιας επιφάνειας, μετάδοση COVID, …*

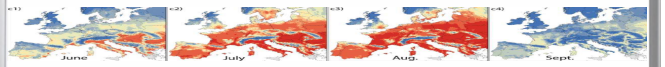


Χωρική ετερογένεια μπορεί να είναι

- *διαστρωματική* ή *στρωματοποιημένη (spatially stratified heterogeneity)*: σε μία γεωγραφική περιοχή, η διακύμανση εντός των στρωμάτων (ομοιογενών υποομάδων) είναι μικρότερη από τη διακύμανση μεταξύ των στρωμάτων, π.χ., οικολογικές ή κλιματικές ζώνες, χρήσεις γης, …→ *χωρική παραλλαγή χαρακτηριστικών*



Χρονική ετερογένεια;



- Αναφέρεται ως η διαφορά στο είδος ή στη διάταξη ενός χαρακτηριστικού ή στο βαθμό αλλαγής, με την πάροδο του χρόνου
  - Η πρόβλεψή της είναι μια σημαντική παράμετρος για τη διαχείριση της δομής και της λειτουργίας ενός συστήματος (π.χ. ενός οικοσυστήματος) → μπορεί να χρησιμοποιηθεί για να περιγράψει σύνθετες αλληλεπιδράσεις μεταξύ διαφόρων παραγόντων σε ένα σύστημα ή περιβάλλον

Εκδηλώσεις El Nino με υψηλό θερμικό στρες

https://doi.org/10.1371/journal.pone.0190957.g005

Γιατί χρειαζόμαστε στατιστικά μέτρα

- Στο φυσικό κόσμο όλα ποικίλλουν: Εάν μετρήσετε το ίδιο μέγεθος δύο φορές, θα λάβετε δύο διαφορετικά αποτελέσματα.
- Η ετερογένεια είναι καθολική: χωρική ετερογένεια & χρονική ετερογένεια → Πρέπει να διακρίνουμε τη διακύμανση που είναι επιστημονικά ενδιαφέρουσα και την παραλλαγή που απλώς αντικατοπτρίζει την ετερογένεια του υποβάθρου των παρατηρήσεων μας

Γιατί χρειαζόμαστε στατιστικά μέτρα

Στατιστική σημαντικότητα: Σε μια ανάλυση δεδομένων, θέλουμε να γνωρίζουμε εάν τα ευρήματά μας είναι "σημαντικά"

- Η πρακτική σημασία δεν είναι πάντα το ίδιο πράγμα με την εμπιστοσύνη ότι ένα αποτέλεσμα δεν οφείλεται αποκλειστικά στην τύχη → στατιστική σημασία

Η στατιστική σημασία είναι' μια 'γκρίζα έννοια που συχνά παρεξηγείται

- Απαραίτητη η χρήση στατιστικών μέτρων, αλλά πάντοτε με καλή κρίση

Στατιστικά μέτρα
- Εννοιολογικός ορισμός -

Συνήθως ο όρος υποδηλώνει συνοπτικές πληροφορίες για …

- το σχεδιασμό, τη σύνοψη των συλλεγόμενων/ταξινομημένων δεδομένων
- την ερμηνεία αβέβαιων παρατηρήσεων μιας έρευνας ή μιας μετρητικής διαδικασίας

✓Τυπικά αποτελούν **υποπροϊόν της στατιστικής επεξεργασίας** των διαθέσιμων δεδομένων

✓Δεν μετρούν άμεσα την ποιότητα, αλλά μπορούν να παρέχουν αρκετές **περιγραφικές ή αριθμητικές πληροφορίες** για διάφορες πτυχές της ποιότητας (στα δεδομένα, στα μοντέλα, στις προγνώσεις, …)

Παράμετροι ή μέτρα κατανομής

- Περιγράφουν μια υποκείμενη φυσική διεργασία ή υπόβαθρο με τρόπο που η τιμή τους επηρεάζει την κατανομή των μετρημένων παρατηρήσεων ενός ή περισσότερων χαρακτηριστικών που έχουν μια τυχαία συνιστώσα
- Εξυπηρετούν τις ανάγκες περιγραφής της συμπεριφοράς μιας τ.μ. που ενδιαφέρει να μελετήσουμε, συμπληρώνοντας επίσης την εικόνα της κατανομής πιθανότητας που ακολουθεί η εν λόγω τ.μ.

Μαθηματικός ορισμός στατ. μέτρων

Εάν ***X*** είναι μια τ.μ. που αντιπροσωπεύει ένα ποσοτικό χαρακτηριστικό ή μια ιδιότητα ενός πληθυσμού που μελετάμε που …

↓

ακολουθεί μια κατανομή με σ.π.π. ***f(X,u)*** ή αθροιστική σ.κ.π. ***F(X,u)***, γνωστής συναρτησιακής μορφής εξαρτώμενες από …

↓

μια άγνωστη (k-διάστατη) παράμετρο $\boldsymbol{u} = (u_1, u_2, \ldots, u_k)$ με τιμές από ένα σύνολο επιτρεπτών παραμέτρων U (: = *παραμετρικός χώρος*)

→ **Πως μπορούμε να εκτιμήσουμε την παράμετρο *u* ?**

Εκτιμητική Στατιστική
(μερικές σχετικές έννοιες και ορισμοί)

- *Στατιστική (ή δειγματική) συνάρτηση (στατιστικό/statistic)* := αποκαλείται κάθε συνάρτηση $T(\cdot) = T(X_1, X_2, \ldots, X_n)$ των τιμών ενός δείγματος $X_1, X_2, \ldots, X_n$, η οποία δεν εξαρτάται από τις προς εκτίμηση συνιστώσες $u_1, u_2, .., u_k$ της άγνωστης παραμέτρου u → Μια αριθμητική μέτρηση ενός χαρακτηριστικού ενός δείγματος
  - Προφανώς, κάθε στατιστική συνάρτηση $T(\cdot)$ είναι και η ίδια μια τ.μ. (αφού κάθε φορά που παίρνουμε ένα άλλο τυχαίο δείγμα Χ, η $T(\cdot)$ δίνει διαφορετική τιμή) που εξαρτάται μόνο από τις παρατηρήσεις του δείγματος

- ***Εκτιμήτρια συνάρτηση*** ή απλά ***εκτιμήτρια*** ή ***εκτιμητής*** *της παραμέτρου* ***u*** *(estimator)* := κάθε στατιστική συνάρτηση $g(\mathbf{u}) = g(X_1, X_2, \ldots, X_n)$ των τιμών ενός δείγματος, η οποία χρησιμοποιείται για την εκτίμηση $\hat{u}$ *(estimate)* της παραμέτρου ***u***
  - Σε κάθε δυνατή περίπτωση και κάθε τυχαίο δείγμα, είναι επιθυμητό αυτή να εγγυάται την πλησιέστερη εκτίμηση $\hat{u}$ της πραγματικής τιμής της ***u***
  - Για διαφορετικές τιμές $\{X_1, X_2, \ldots, X_n\}$ (δείγματα), η εκτιμήτρια συνάρτηση $g(\hat{u})$ της παραμέτρου $\hat{u}$, παίρνει διαφορετικές τιμές → η $\hat{u}$ είναι επίσης μια τ.μ. με κάποια κατανομή.

Επιμέρους πρακτικά προβλήματα

- Με βάση τα δεδομένα ενός δείγματος
  - Να προσδιοριστεί η πιο "πιθανή" τιμή της εκτιμήτριας ***u*** → ***σημειακή εκτίμηση***
  - Να βρεθεί διάστημα τιμών, εντός του οποίου να περιέχεται η ***u***, με δεδομένη πιθανότητα → ***διάστημα εμπιστοσύνης***
  - Εάν υπάρχουν ≥2 εναλλακτικές υποθέσεις για την παράμετρο ***u***, ποια από αυτές είναι αποδεκτή → ***έλεγχος υποθέσεων***
  - Εάν υπάρχει μια υπόθεση για την παράμετρο ***u***, σε ποιο βαθμό είναι αποδεκτή ή όχι → ***έλεγχος σημαντικότητας***

Σημειακοί εκτιμητές *(point estimators)*:

- Στατιστικά μεγέθη ενός δείγματος που χρησιμοποιούνται απευθείας ως προσέγγιση αντίστοιχων παραμέτρων ενός πληθυσμού
- Σημειακή εκτίμηση (σ.ε.) μιας παραμέτρου ενός πληθυσμού είναι μια τιμή που παίρνει μια εκτιμήτρια συνάρτηση → αντιπροσωπεύει την πραγματική τιμή της σχετικής παραμέτρου του πληθυσμού από τον οποίο προέρχεται το δείγμα

$$\hat{u} = g(x_1, x_2, \ldots, x_n)$$

  - *π.χ., η μέση τιμή ενός δείγματος είναι μια σ.ε. της μέσης τιμής μ του πληθυσμού. Ομοίως, το ποσοστό p του δείγματος με κάποια ιδιότητα είναι μια σ.ε. του ποσοστού του πληθυσμού με την ίδια ιδιότητα.*

Διαστήματα εκτίμησης *(interval estimators)*:

- Μια ευρύτερη μορφή εκτίμησης από τη σημειακή
- Υποδηλώνουν την εμπιστοσύνη (με όρους πιθανότητας) στην τιμή μιας παραμέτρου με την μορφή ενός διαστήματος εύλογων (πιθανών) τιμών *(confidence interval)* στο οποίο αναμένεται να περιέχεται η πραγματική τιμής της παραμέτρου ενδιαφέροντος με μια προκαθορισμένη πιθανότητα
  - Ένα διάστημα εκτίμησης ορίζεται από δύο τιμές, μεταξύ των οποίων θεωρείται ότι βρίσκεται μια παράμετρος του πληθυσμού, π.χ. $a < \bar{x} < b$ είναι ένα διάστημα εκτίμησης του μέσου **μ** του πληθυσμού

Έλεγχοι Υποθέσεων και Επίπεδα Σημαντικότητας

- Επιτρέπουν να αξιολογηθούν οι ισχυρισμοί/παραδοχές μιας έρευνας σχετικά με τις τιμές μιας μεμονωμένης ή πολλών παραμέτρων ή της μορφής μιας κατανομής πιθανοτήτων ενός πληθυσμού που συνάγονται από τη λήψη ενός δείγματος δεδομένων
- Συνιστούν το κλειδί για την διενέργεια αυστηρών μετρητικών διαδικασιών, όπου συχνά αυτό που ενδιαφέρει είναι η μελλοντική απόδοση ενός συστήματος κάτω από παρόμοιες (ή διαφορετικές) συνθήκες παρατήρησης.

Έλεγχος υποθέσεων (Hypothesis testing)

- Διαφέρει από την εκτίμηση, καθώς ένας έλεγχος υποθέσεων ξεκινάει από κάποια δήλωση ή ιδέα *(που πιστεύεται ότι μπορεί να είναι αληθινή, αλλά δεν μπορεί να αποδειχθεί ότι είναι)* για τον υπό εξέταση πληθυσμό και στη συνέχεια δοκιμάζεται εάν κάποιο δείγμα παρατηρήσεων του πληθυσμού υποστηρίζει την εν λόγω ιδέα.
- Παρόμοια μαθηματική προσέγγιση με εκείνη που ακολουθείται στην εκτίμηση διαστημάτων, αλλά το συμπέρασμα που βγαίνει είναι αρκετά διαφορετικό.
  - " Πώς είναι ο πληθυσμός; " | " Είναι ο πληθυσμός έτσι ή όχι; "

Επίπεδα Σημαντικότητας

- Στον έλεγχο στατιστικών υποθέσεων, ένα αποτέλεσμα έχει *στατιστική σημασία* σε ένα καθορισμένο *επίπεδο σημαντικότητας* μιας μελέτης, που υποδηλώνεται με την πιθανότητα να ληφθεί ένα αποτέλεσμα συμβατό με τα πρότυπα μιας διερεύνησης.
- Τυπικά, το επίπεδο σημαντικότητας για μια μελέτη επιλέγεται πριν από τη συλλογή δεδομένων και τυπικά ορίζεται στο 5% ή πολύ χαμηλότερο → *Όταν αντλούνται δεδομένα από ένα δείγμα, αυτό σημαίνει ότι η περιοχή απόρριψης περιλαμβάνει το 5% της κατανομής δειγματοληψίας*

Μέθοδοι εύρεσης εκτιμητριών

- Μέθοδος μέγιστης πιθανοφάνειας
  - Method of maximum likelihood
  - Προτάθηκε από τον Ronald A. Fisher (1922)
- Μέθοδος των ροπών
  - Προτάθηκε από τον Karl Pearson (1894).
- Μέθοδος των ελαχίστων τετραγώνων
  - Πρωτοεμφανίστηκε σε μια εργασία του Legendre (1805), αλλά περιγράφηκε το 1796 και δημοσιεύθηκε αργότερα από τον Gauss (1809)

Όλες βασίζονται στην υιοθέτηση κάποιας αρχής βελτιστοποίησης που επιτρέπει τον υπολογισμό εκτιμητριών με επιθυμητές ιδιότητες

Μέθοδοι μέγιστης πιθανοφάνειας

- Βασίζονται στη χρήση κάποιου υποκείμενου πιθανοθεωρητικού μοντέλου (που υποθέτουμε ότι ισχύει, *π.χ. μια κανονική κατανομή με άγνωστη μέση τιμή και διασπορά*), του οποίου οι τιμές των παραμέτρων επιλέγονται ώστε να μεγιστοποιείται η πιθανοφάνεια συμφωνίας ενός τυχαίου δείγματος $\{X_1, X_2, \ldots, X_n\}$ που έχουμε παρατηρήσει από τον πληθυσμό, με τις διάφορες τιμές της παραμέτρου $u(\mu, \sigma^2)$

- Στην περίπτωση ενός τυχαίου δείγματος $x = (X_1, X_2, \ldots, X_n)$ από μια κατανομή με σ.π.π. $f(x; \theta)$, η *συνάρτηση πιθανοφάνειας* έχει τη μορφή

$$L(u) = L(u, x) = \prod_{i=1}^{n} f(X_i; u) \ , \ u \in U$$

Από τη ***μεγιστοποίηση*** της συνάρτησης πιθανοφάνειας

$$L(\hat{u}; x) = \max_{u} L(u; x)$$

Προκύπτουν οι εκτιμήτριες **$\hat{u}$** μέγιστης πιθανοφάνειας της **u**

Μέθοδος των ροπών

- Βασίζεται στην επίλυση ενός συστήματος k εξισώσεων

$$\mu_r = m_r \ , \ r = 1, 2, \ldots, k$$

μερικών από τις πρώτες (r-τάξης) **ροπές** του πληθυσμού περί το μηδέν, $\mu_r(u_1, u_2, \ldots u_k) := E[X^r]$, με τις τιμές των αντιστοίχων ροπών m_r ενός τυχαίου δείγματος $(X_1, X_2, \ldots X_n)$ → απαιτούνται τόσες ροπές σε πλήθος όσος είναι και ο αριθμός των παραμέτρων $u_1, u_2, \ldots u_k$ που θέλουμε να εκτιμήσουμε

- Μέθοδος των ροπών – παράδειγμα
  - π.χ., αν θέλουμε να εκτιμήσουμε τη μέση τιμή και τη διασπορά μιας κανονικής κατανομής

$$\mu_1 = E(X) = \mu \quad \text{και} \quad \mu_2 = E(X^2)$$

Για ένα τυχαίο δείγμα $X_1, X_2, \ldots, X_n$ οι δειγματικές ροπές είναι

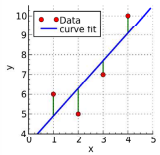
$$m_1 = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X} \quad \text{και} \quad m_2 = \frac{\sum_{i=1}^{n} X_i^2}{n}$$

$$\Rightarrow \hat{\mu} = \overline{X}, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2$$

Μέθοδος των ελαχίστων τετραγώνων

- Βασίζεται στον καθορισμό της εκτιμήτριας $\hat{y}$ μιας παραμέτρου με αντιπροσωπευτική τιμή y (από κάποιο μοντέλο ή των θεωρητικών από την κατανομή τιμών), από την ελαχιστοποίηση του αθροίσματος των τετραγώνων των αποκλίσεων μεταξύ κάθε παρατήρησης y_i και της προβλεπόμενης τιμής y

$$\sum_{i=1}^{n} (y_i - y)^2 \rightarrow min$$



Κριτήρια 'καλών' εκτιμητριών

- Έστω μια τ.μ. ***X*** με κατανομή $F_X(x, u)$
  - ***u***: άγνωστη παράμετρος της κατανομής → π.χ., $u = (\mu, \sigma, p, \ldots)$
- Σε πολλές περιπτώσεις, χρησιμοποιούμε ως εκτιμήτρια της παραμέτρου ενός πληθυσμού την αντίστοιχη συνάρτηση ενός αντιπροσωπευτικού δείγματος $\{ x_1, x_2, \ldots, x_n \}$ του πληθυσμού
  - *$\hat{u} \rightarrow \mu \rightarrow$ εκτίμηση από τη δειγματική μέση τιμή $\overline{x}$*
  - *$\hat{u} \rightarrow \sigma^2 \rightarrow$ εκτίμηση από τη δειγματική διασπορά S^2*

Κριτήρια 'καλών' εκτιμητριών

- ***$\hat{u}$*** : Εκτιμήτρια της *u*
  - συνάρτηση των τ.μ. $\{ x_1, x_2, \ldots, x_n \}$
  - *Η ίδια είναι μια τ.μ. με αναμενόμενη τιμή* $E(\hat{u}) = \mu_{\hat{u}}$ και διασπορά $Var(\hat{u}) = \sigma_{\hat{u}}^2$
- Πόσο σημαντικά μπορεί να διαφέρει η εκτίμηση μιας παραμέτρου ενδιαφέροντος από την πραγματική τιμή της παραμέτρου;
  - Η αξία μιας εκτιμήτριας αξιολογείται από την κατανομή των εκτιμήσεων τις οποίες παράγει, δηλαδή, από τις ιδιότητες της δειγματικής κατανομής της

Στατιστικά μέτρα τυχαίων μεταβλητών

Στατιστικά μέτρα τυχαίων μεταβλητών

ΟΛΑ βασίζονται στον τελεστή ***E[·] → αναμενόμενη τιμή***

Πρακτικά, για την εκτίμηση τους χρειαζόμαστε …

Ένα σύνολο από πραγματώσεις *(δείγμα)* ενός τυχαίου πειράματος, ενός φαινομένου ή μιας μετρητικής διαδικασίας, σε συνδυασμό με …

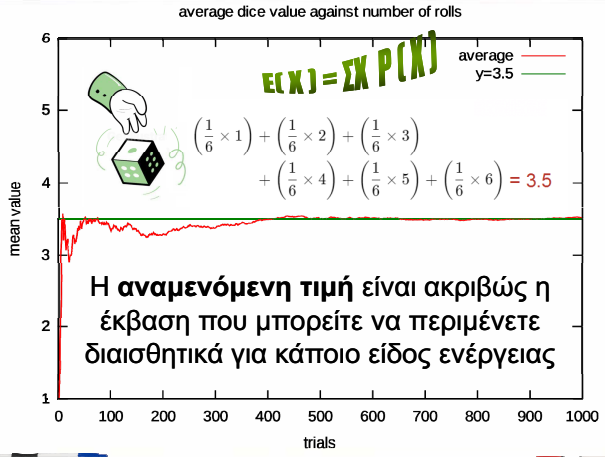
- Πίνακες ή/και γραφικές παραστάσεις που παριστούν περιληπτικά τα δεδομένα, για διερεύνηση τάσεων (μοτίβων) και ιδιαίτερων τιμών στα δεδομένα

Αριθμητικά (στατιστικά) μέτρα τ.μ.

- Μέτρα θέσης ή κεντρικής τάσης
- Μέτρα διασποράς ή μεταβλητότητας
- Μέτρα της μορφής και του σχήματος κατανομής
- Μέτρα αυτοσυσχέτισης
- Μέτρα συμμεταβλητότητας

- Μέτρα κεντρικής τάσης
  - παρέχουν πληροφορίες για ‘συγκεκριμένες θέσεις ή τοποθεσίες ενός συνόλου δεδομένων
    - Αναμενόμενη τιμή / Διάμεσος
    - p-ποσοστημόρια ή p-εκατοστιαία σημεία
    - Επικρατούσα τιμή (mode)
- Μέτρα διασποράς
  - Συνοψίζουν τη μεταβλητότητα των δεδομένων
    - Διακύμανση / Τυπική Απόκλιση
    - Εύρος

- Μέτρα της μορφής και του σχήματος κατανομής
  - Μέτρα συμμετρίας/ασυμμετρίας
    - λοξότητα (skewness)
    - κύρτωση (kurtosis)
- Μέτρα εξάρτησης
  - Μέτρα συσχέτισης | αυτοσυσχέτισης και συμμεταβλητότητας (δύο ή περισσότερων τ.μ.)
    - Συντελεστής (γραμμικής) συσχέτισης (coefficient of correlation)
    - Συνάρτηση αυτοσυσχέτισης



Αναμενόμενη τιμή μιας συνάρτησης τυχαίας μεταβλητής

- Γενικευμένος ορισμός: Αναμενόμενη τιμή μιας πραγματικής συνάρτησης ***g(▪)***, *συμβολικά E[g(.)]* των τιμών μιας τ.μ. ***X*** με συνάρτηση πυκνότητας πιθανότητας ***f(x)***, ορίζεται ως η **σταθμισμένη μέση τιμή** *(weighted average)* της g(▪), με τη συνάρτηση μάζας πιθανότητας ***f(x)*** να παίζει το ρόλο της συνάρτησης βάρους
- Αντιπροσωπεύει κατά κάποιο τρόπο όλες τις δυνατές τιμές μιας τυχαίας μεταβλητής

Αναμενόμενη τιμή μιας διακριτής ή μιας συνεχούς τ.μ.

για συνεχείς τυχαίες μεταβλητές

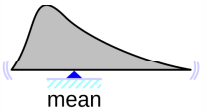
$$E[g(X)] = \int_{-\infty}^{\infty} g(x)\, f(x)\, dx$$

για διακριτές τυχαίες μεταβλητές

$$E[g(X)] = \sum_{i}^{\infty} g(x_i)\, P(X = x_i)$$

- Εάν η αντίστοιχη συνάρτηση μάζας/πυκνότητας πιθανότητας ***f(x)*** παίρνει μη μηδενικές τιμές μόνο στο διάστημα ***[a,b]***

$$\Rightarrow E[x] = \mu = \int_a^b x\, f_x(x)\, dx$$



- Γενικά, η μέση τιμή είναι εύκολη στον υπολογισμό της και εύχρηστη:
  - Δίνει το κέντρο βάρους της κατανομής → μια ένδειξη για τη θέση γύρω από την οποία είναι τοποθετημένες οι τιμές της τ.μ.
  - έχει όμως το μειονέκτημα να επηρεάζεται από τυχόν **ακραίες τιμές** *(outliers)* του πεδίου τιμών της τ.μ

- Εάν η μεταβλητή *X* παίρνει ισοπίθανα όλες τις πιθανές τιμές της… Η μέση τιμή *E*(*X*) ενδέχεται να μην συμπίπτει με καμία από τις πιθανές τιμές της *X*
- Παράδειγμα – ο ημερήσιος αριθμός (X=k, k=0,1,…,5) μικρής διάρκειας διακοπών στη λήψη σημάτων GPS λόγω συνθηκών ορατότητας των δορυφόρων σε έναν σταθμό

Outages per Day (x_i)	$P(X = x_i)$	$x_i P(X = x_i)$
0	0.35	$(0)(0.35) = 0.00$
1	0.25	$(1)(0.25) = 0.25$
2	0.20	$(2)(0.20) = 0.40$
3	0.10	$(3)(0.10) = 0.30$
4	0.05	$(4)(0.05) = 0.20$
5	0.05	$(5)(0.05) = 0.25$
	1.00	$\mu = E(X) = 1.40$

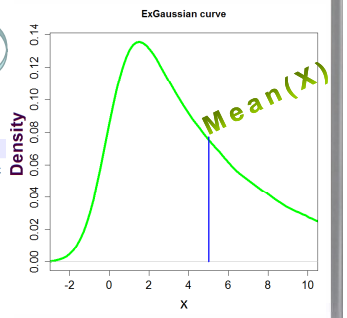
Η μέση τιμή1.4 δεν αποτελεί ένα πιθανό ενδεχόμενο να συμβεί

Η συνάρτηση ***mean()*** της R

```
# Create a vector.
x <- c(12,7,3,4.2,18,2,54,-21,8,-5)
# Find Mean.
mean(x)
[1] 8.22

# Find Mean, trim=0.2
# (2 values from each end will be dropped)
mean(x, trim=0.2)
[1] 6.033333

# Create a vector with NA elements
x <- c(12,7,3,4.2,18,2,54,-21,8,-5,NA) ; x
 [1]  12.0   7.0   3.0   4.2  18.0   2.0
 [7]  54.0 -21.0   8.0  -5.0    NA
# Find mean.
mean(x)
[1] NA
# Find mean dropping NA values.
mean(x,na.rm = TRUE)
[1] 8.22
```

Περικομμένος Μέσος όρος

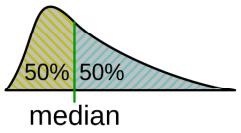
x <- seq(-2.5, 10, length=1000000)



Διάμεσος (median)

- Ιστορικά, για πρώτη φορά η χρήση της διάμεσου τιμής αναφέρεται σε ένα ναυτιλιακό εγχειρίδιο του 1599 *(Edward Wright – Certain errors in Navigation)* … *ως πρόταση για τη βέλτιστη τιμή στο πρόβλημα του προσδιορισμού της κανονικής διακύμανσης των ενδείξεων σε μια πυξίδα*
- Ένα σαφέστερο παράδειγμα χρήσης της διάμεσης τιμής, αναφέρεται στο πλαίσιο των σφαλμάτων μέτρησης, πολύ αργότερα (περί το 1757) όταν ο *Roger Joseph Boscovich* ανέπτυξε μια μέθοδο παλινδρόμησης βασισμένη ουσιαστικά στη διάμεσο

Διάμεσος (median)

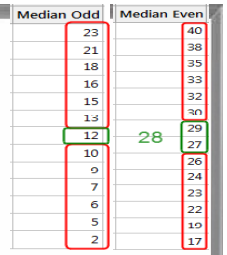


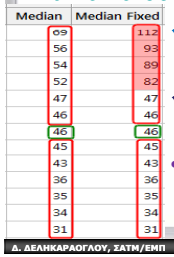
- Η διάμεσος μιας τυχαίας μεταβλητής X με συνάρτηση κατανομής $f(x)$, $\forall\ x \in R$ είναι η τιμή M για την οποία ισχύει:
  $P(X < M) \le 0.5$ και $P(X \le M) \ge 0.5$
  - δηλ., εκείνη η τιμή της τ.μ. που χωρίζει την κατανομή πιθανότητας σε δύο ίσα μέρη
    - *milieu de probabilité* / the middle of the probability (κατά τους Legendre και Laplace, ~18 αι.)

Η **Διάμεσος/Median** μπορεί να υπολογιστεί για μεταβλητές που μετρώνται με *κλίμακες διάταξης, ίσου-διαστήματος ή αναλογίας*

- Δεν εφαρμόζεται σε ονομαστικά δεδομένα
- Για διακριτές τ.μ., η διάμεσος είναι η μικρότερη τιμή που δίνει $f(x) \ge ½$.
- Εάν η $f(x)$ είναι συνεχής, η διάμεσος της κατανομής είναι κάθε λύση της εξίσωσης $f(M) = ½$
  - Σε μια συμμετρική κατανομή, η μέση τιμή (=για την οποία μεγιστοποιείται η σ.π. $f(x)$) ταυτίζεται με τη διάμεσο

- Η διάμεσος είναι η τιμή στο μέσο μιας κατανομής δεδομένων οργανωμένων από τη χαμηλότερη στην υψηλότερη τιμή
- Η θέση της διαμέσου δίνεται από το δείκτη *(n+1)/2*





- ✓n: ζυγός → το μέσο στοιχείο του διατεταγμένου συνόλου
- ✓n: μονός → το ημιάθροισμα των δυο μέσων στοιχείων
- Δεν επηρεάζεται από ακραίες τιμές *(outliers)* στα δεδομένα

Η συνάρτηση *median()* της R

- *median(x, na.rm = FALSE, ...)*

```
# Create a vector
  x <- c(-21,-5,-1.2,2,3,4.2,7,8,12,18,54)
# Find Median
   result.median <- median(x)
   print(result.median)
   [1] 5.6
# Create a vector
  x <- c(-21,-5,-1.2,2,3,4.2,7,8,12,18,54)
# Find Median
   result.median <- median(x)
   print(result.mean)
   [1] 4.2
```


Συχνότερη τιμή *(mode)*

- Είναι η τιμή που έχει τον μεγαλύτερο πλήθος εμφανίσεων σε ένα σύνολο δεδομένων. Σε αντίθεση με τη μέση τιμή και τη διάμεσο, μπορεί να αναφέρεται τόσο σε αριθμητικά όσο και σε δεδομένα χαρακτήρων (κατηγορικά δεδομένα).
- **Το R δεν διαθέτει ενσωματωμένη συνάρτηση για τον υπολογισμό της.**
  - Για τον υπολογισμό της απαιτείται μια συνάρτηση χρήστη | ***ΣΗΜΕΙΩΣΤΕ:*** *η βασική συνάρτηση mode() στο R αναφέρεται στην εσωτερική αποθήκευση ενός αντικειμένου*

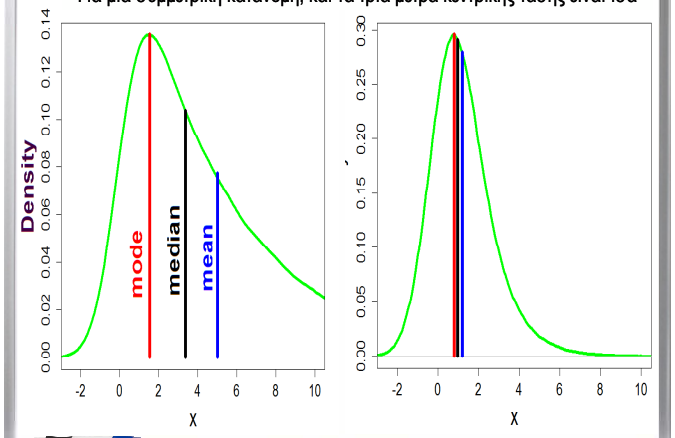

```
mySamples <- c(19, 4, 5, 7, 29, 19, 29, 13, 25, 19, 29)
    user_mode <- function(x) {
          ux <- unique(x)
          ux[which.max(tabulate(match(x, ux)))]
        }
user_mode(mySamples)
[1] 19
```

Επιστρέφει τη "διάμεσο" τιμή που *εμφανίζεται για πρώτη φορά*

```
    multi_modes <- function(x) {
          ux <- unique(x)
          tab <- tabulate(match(x, ux))
          ux[tab == max(tab)]
        }
multi_modes(mySamples)
[1] 19 29
```

Επιστρέφει όλες τις "διαμέσους" τιμές , δηλ. ίδιας μέγιστης συχνότητας εμφάνισης

Για μια συμμετρική κατανομή, και τα τρία μέτρα κεντρικής τάσης είναι ίσα



$$\bar{X}_{geom} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[n]{x_1 . x_1 ..... x_n}$$

Γεωμετρικός μέσος όρος *(geometric mean)*

$x_1, x_2, \ldots, x_n$: Σύνολο αυστηρά θετικών αριθμών

Στο R ο υπολογισμός γίνεται έμμεσα, π.χ.

```
# Defining vector
x <- c(1, 5, 9, 19, 25)
# Print Geometric Mean: prod(x)^(1/length(x))
print(prod(x)^(1 / length(x)))
[1] 7.344821
```


$$\frac{1}{h} = \frac{\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n}}{n} = \frac{\sum_{i=1}^{n} \frac{1}{x_i}}{n} \Rightarrow h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

Αρμονικός μέσος όρος *(harmonic mean)*

Στο R ο υπολογισμός γίνεται έμμεσα, π.χ.

```
# Defining vector
x <- c(1, 5, 8, 10)
# Print Geometric Mean: 1 / mean( 1/x )
print ( 1 / mean( 1/x ) )
[1] 2.807018
```

```
Functions to avoid explicit uses of loop constructs
# apply()    applies a function FUN (mean, median, min, max, etc..)
#            to the rows or columns or both (MARGIN=1,2, or c(1,2))
#            of a data frame or matrix X
apply(X, MARGIN, FUN)

# lapply()   applies a function FUN (mean, median, min, max, etc..)
#            to all the elements of an input list, vector or data frame X
lapply(X, FUN)        # the output is a list

# sapply()   applies a function FUN (mean, median, min, max, etc..)
#            to all the elements of an input list, vector or data frame X
sapply(X, FUN)        # the output is a vector or matrix

# tapply()   computes a measure (mean, median, min, max, etc..) or
#            a function for each factor variable in a vector X
tapply(X, INDEX, FUN)       # the output is group summaries
```


> # Use of apply() - Sum a matrice over all the columns

```
> m1 <- matrix(C<-(1:32),nrow=4, ncol=8)
> m1
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
[1,]    1    5    9   13   17   21   25   29
[2,]    2    6   10   14   18   22   26   30
[3,]    3    7   11   15   19   23   27   31
[4,]    4    8   12   16   20   24   28   32
> a_m1 <- apply(m1, 2, sum)
> a_m1
[1]  10  26  42  58  74  90 106 122
```


Example with dataset 'faithful'

```
> dt <- faithful
> lmn_faithful <- lapply(dt, min)
> smn_faithful <- sapply(dt, min)
> lmn_faithful
$eruptions
[1] 1.6

$waiting
[1] 43

> smn_faithful
eruptions   waiting
      1.6      43.0
```

```
> head(faithful)
  eruptions waiting
1     3.600      79
2     1.800      54
3     3.333      74
4     2.283      62
5     4.533      85
6     2.883      55
> tail(faithful)
     eruptions waiting
267      4.750      75
268      4.117      81
269      2.150      46
270      4.417      90
271      1.817      46
272      4.467      74
```



```
> head(rainfall, 20)
   ReportingPeriodType Year Period Rainfall_mm Temperature_C
1                Month 2010    JAN        69.5          0.80
2                Month 2010    FEB        71.7          1.60
3                Month 2010    MAR        65.1          5.50
4                Month 2010    APR        21.4          8.20
5                Month 2010    MAY        22.1          9.80
6                Month 2010    JUN        42.7         14.30
7                Month 2010    JUL        61.1         16.40
8                Month 2010    AUG        76.6         14.60
9                Month 2010    SEP        83.9         13.40
10               Month 2010    OCT        65.0          9.60
11               Month 2010    NOV       118.2          4.50
12               Month 2010    DEC        40.7         -1.20
13              Season 2010 Winter       228.7          1.52
14              Season 2010 Spring       108.6          7.85
15              Season 2010 Summer       180.5         15.08
16              Season 2010 Autumn       267.2          9.17
17       Calendar Year 2010 Annual       738.3          8.16
> tapply(rainfall$Rainfall_mm, rainfall$Period, mean)
   Annual       APR       AUG    Autumn       DEC       FEB       JAN       JUL       JUN
799.08333  39.94167  77.58333 220.86667  79.26667  63.25833  71.46667  67.91667  68.58333
      MAR       MAY       NOV       OCT       SEP    Spring    Summer    Winter
 53.47500  56.68333  80.57500  80.19167  60.10000 150.13333 214.06667 215.13333
```



```
> str(rainfall)
'data.frame':   204 obs. of  5 variables:
 $ ReportingPeriodType: chr  "Month" "Month" "Month" "Month" ...
 $ Year               : int  2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...
 $ Period             : chr  "JAN" "FEB" "MAR" "APR" ...
 $ Rainfall_mm        : num  69.5 71.7 65.1 21.4 22.1 42.7 61.1 76.6 83.9 65 ...
 $ Temperature_C      : num  0.8 1.6 5.5 8.2 9.8 14.3 16.4 14.6 13.4 9.6 ...
> summary(rainfall)
 ReportingPeriodType      Year         Period          Rainfall_mm      Temperature_C
 Length:204          Min.   :2010   Length:204         Min.   :  9.70   Min.   :-1.200
 Class :character    1st Qu.:2013   Class :character   1st Qu.:  51.08   1st Qu.: 5.500
 Mode  :character    Median :2016   Mode  :character   Median :  80.50   Median : 9.320
                     Mean   :2016                      Mean   : 141.08   Mean   : 9.342
                     3rd Qu.:2018                      3rd Qu.: 148.72   3rd Qu.:13.100
                     Max.   :2021                      Max.   :1043.00   Max.   :17.800
> tapply(rainfall$Rainfall_mm, rainfall$ReportingPeriodType, summary)
$`Calendar Year`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  661.7   734.1   789.0   799.1   831.9  1043.0

$Month
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.70   43.12   64.15   66.59   84.10  148.00

$Season
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   61.0   157.0   194.8   200.1   229.5   344.4
```

Με τις συναρτήσεις ***lapply()*** ή ***sapply()*** μπορούμε επίσης να χρησιμοποιήσουμε μια ενσωματωμένη συνάρτηση ή μια συνάρτηση χρήστη

Πόσο αρκετή είναι η πληροφορία της μέσης τιμής;

- ***Όχι πάντα*** → π.χ. περιπτώσεις ετερογένειας
- τ.μ. X και Y με ίδια μέση τιμή (π.χ., 2.75 στο παράδειγμα) → χρειαζόμαστε κάποιο άλλο μέτρο της "απόστασης" της X ή της Y από τη μέση τιμή

X :	1	2	3	4
	1/8	2/8	3/8	2/8

Y :	-20	1	10	30
	2/8	2/8	3/8	1/8

Άλλα στατιστικά μέτρα/παράμετροι

- *Εκτός της μέσης τιμής, ανάλογα με τη μορφή της πραγματικής συνάρτησης g(▪) των τιμών της τ.μ. X, ορίζονται διάφορα άλλα μέτρα της πιθανοθεωρητικής συμπεριφοράς της τ.μ. X*, όπως
  - Η *ν*-οστή (γενική και κεντρική) ροπή
  - Η διασπορά ή διακύμανση
  - Η τυπική απόκλιση
- ΚΑΙ ΑΥΤΑ βασίζονται στον τελεστή *E[·]* της αναμενόμενης τιμής

Στατιστικά μέτρα/παράμετροι

1) $g(X)=X$, η $EX=\mu$ ονομάζεται *μέση τιμή* της τ.μ. X

2) $g(X)=(X-a)^v$, η $E(X-a)^v$ ονομάζεται ν-οστή ροπή της τ.μ. X ως προς α

3) $g(X)=(X-\mu)^v$, η $E(X-\mu)^v$ ονομάζεται ν-οστή κεντρική ροπή της τ.μ. X

4) $g(X)=(X-\mu)^2$, η $E(X-\mu)^2=\sigma^2=VarX$, ονομάζεται διασπορά ή διακύμανση της τ.μ. X και η $\sigma=\sqrt{VarX}=\sqrt{E(X-\mu)^2}$, τυπική απόκλιση της τ.μ. X.

Διακύμανση – το βασικότερο μέτρο διασποράς / μεταβλητότητας γύρω από την αναμενόμενη τιμή

$$E[g(x)] = \mu = \int_{-\infty}^{+\infty} g(x)\, f_x(x)\, dx$$

$$\mathrm{Var}(X) = \sigma^2 = E[(X-\mu)^2] = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)\, dx$$

$$= \int_{-\infty}^{+\infty} x^2 f(x)\, dx - \mu^2 = E[X^2] - \mu^2$$

$$\mathrm{Var}(X) = \sigma^2 = E[\,(X-\mu)^2\,] =$$
$$= \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)\,dx = \underbrace{E[\,X^2\,]}_{\text{Μέση τετραγωνική τιμή}} - \mu^2$$

- Εάν η τ.μ. *X* είναι διακριτή με πεδίο τιμών [$x_1, x_2, \ldots, x_N$]

$$\mathrm{Var}(X) = \sigma^2 = V(X) =$$
$$E[\,(X-\mu)^2\,] = \sum_{1}^{N} (x_i-\mu)^2 f(x_i)$$

- Η διασπορά μιας τ.μ. είναι μέτρο της *"συγκέντρωσης"* ή *"διάχυσης"* των τιμών της μεταβλητής ***γύρω από*** την αντιπροσωπευτική τιμή της, ***τη μέση τιμή***
- Γενικά, όσο περισσότερο αποκλίνει η τ.μ. *X* από τη μέση τιμή της *μ*, τόσο μεγαλύτερη είναι η διαφορά $(X-\mu)^2$ άρα και η διασπορά $Var(X)=\sigma^2$
  - Εάν οι πιθανές τιμές της τ.μ. *X* είναι συγκεντρωμένες κοντά στη μέση τιμή, η διασπορά είναι μικρή.

- *Συνέχεια προηγούμενο παράδειγμα – ο ημερήσιος αριθμός (X=k, k=0,1,…,5) διαλείψεων των σημάτων GPS*

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_{i=1}^{N} [x_i - E(X)]^2 P(X = x_i)}$$

Outages per Day (x_i)	$P(X=x_i)$	$x_i P(X=x_i)$	$[x_i - E(X)]^2 P(X=x_i)$
0	0.35	$(0)(0.35) = 0.00$	$(0-1.4)^2(0.35) = 0.686$
1	0.25	$(1)(0.25) = 0.25$	$(1-1.4)^2(0.25) = 0.040$
2	0.20	$(2)(0.20) = 0.40$	$(2-1.4)^2(0.20) = 0.072$
3	0.10	$(3)(0.10) = 0.30$	$(3-1.4)^2(0.10) = 0.256$
4	0.05	$(4)(0.05) = 0.20$	$(4-1.4)^2(0.05) = 0.338$
5	0.05	$(5)(0.05) = 0.25$	$(5-1.4)^2(0.05) = 0.648$
	1.00	$\mu = E(X) = 1.40$	$\sigma^2 = 2.04$

Ιδιότητες της διακύμανσης μιας τυχαίας μεταβλητής *X*

- $V(a) = 0$, όπου *a*: σταθερά
- $V(X+b) = V(X)$, όπου *b*: σταθερά
- $V(aX) = a^2\,V(X)$, όπου *a*: σταθερά
- $V(aX+b) = a^2\,V(X)$, όπου a, *b*: σταθερές
- $V(X_1+X_2) = V(X_1) + V(X_2)$, εφόσον οι τ.μ. X_1 και X_2 είναι στατιστικά ανεξάρτητες (→ ασυσχέτιστες)
  - Γενικά δεν ισχύει το αντίστροφο

Διακύμανση ή διασπορά πληθυσμού/δείγματος

Διαφορετικός συμβολισμός

… για πληθυσμό

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2 = \frac{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}{N^2}$$

… για δείγμα του πληθυσμού

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

Μερικές χρήσιμες ταυτότητες

$$X_i - \mu = (X_i - \bar{X}) + (\bar{X} - \mu)$$

$$\sum_{1}^{n}(X_i-\mu)^2 = \ldots = \sum_{1}^{n}(X_i-\bar{X})^2 + n(\bar{X}-\mu)^2$$

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

$$Var(X) = \sigma^2 = E[X^2] - \mu^2,$$

$$Var(\bar{X}) = \frac{\sigma^2}{n} = E[\bar{X}^2] - \mu^2$$

Τυπική απόκλιση πληθυσμού/δείγματος

- *Τυπική απόκλιση (standard deviation)* της τ.μ. := η θετική τετραγωνική ρίζα της διασποράς, $\sqrt{\sigma^2}$ | αντίστοιχα ορίζεται η ***δειγματική τυπική απόκλιση*** $S = \sqrt{S^2}$, για την οποία μπορεί να δειχθεί ότι

$$0 < \mathrm{Var}(S) = E[S^2] - (E[S])^2 = \sigma^2 - (E[S])^2 \Rightarrow E[S] < \sigma^2$$

δηλ. είναι ένας προκατειλημμένος εκτιμητής του σ

  - Το πλεονέκτημά της είναι ότι έχει ίδιες 'διαστάσεις' (μονάδες) με την τ.μ.
  - Ωστόσο, σχετικά με τη μεταβλητότητα ενός πληθυσμού/δείγματος, η διακύμανση είναι πιο κατατοπιστική από την τυπική απόκλιση για την παραγωγή στατιστικών συμπερασμάτων

Μια γενικότερη ερμηνεία της τυπικής απόκλισης

- Το όριο που καθορίζεται από τη λεγόμενη ανισότητα Chebyshev

$$\Pr(|X-\mu| \geq k) \leq \frac{\sigma^2}{k^2} \leftrightarrow \Pr(|X-\mu| \leq k) \leq \left(1 - \frac{\sigma^2}{k^2}\right)$$

δίνει ένα άνω φράγμα για την πιθανότητα μια τ.μ. να αποκλίνει από τη μέση τιμή της

  - π.χ., για k=2, η ανισότητα δείχνει ότι τουλάχιστον τα 3/4 των τιμών βρίσκονται σε ακτίνα 2 τυπικών αποκλίσεων από τον μέσο

Οι συναρτήσεις *var()* & *sd()* της R

```
var(x, y=NULL, na.rm=FALSE,use)
sd(x, na.rm=FALSE)
x,y       Complex vector or matrix
na.rm     Boolean with default FALSE meaning to leave
          NA values present and TRUE meaning to remove them
use       Ignored

#  e.g., the dispersion measures of the 'eruption duration'
#        and the 'waiting' in the data set ' faithful '
duration_e = faithful$eruptions      # the eruption durations
var(duration_e)                      # apply the var function
[1] 1.3027
sd(duration_e)                       # apply the sd function
[1] 1.1414
duration_w = faithful$waiting        # the waiting durations
var(duration_w)                      # apply the var function
[1] 184.8233
sd(duration_w)                       # apply the sd function
[1] 13.59497
```

Συντελεστής μεταβλητότητας πληθυσμού/δείγματος

- **Coefficient of variation (CV)**
- Ο συντελεστής μεταβλητότητας $CV_X = c_X = \sigma/\mu$ δίνει το μέτρο της αβεβαιότητας στην εκτίμηση της τ.μ. X
  - $\uparrow CV_X \rightarrow$ μεγάλη διασπορά σε σχέση με τη μέση τιμή → μεγάλη αβεβαιότητα
  - Αν $CV_X = c_X = 0 \rightarrow \sigma = 0$, δηλ. η τ.μ. λαμβάνει μια μόνο τιμή (=μ), για την οποία έχουμε πλήρη βεβαιότητα

$$CV = \frac{\sigma}{\mu}100 \quad \text{για πληθυσμό}$$

$$= \frac{s}{\bar{x}}100 \quad \text{για δείγμα}$$

Αδιάστατη αριθμητική τιμή (ή σε %) …

- Εκφράζει την 'έκταση' των τιμών μιας τ.μ. σε σχέση με την αναμενόμενη τιμή της
- Χρησιμοποιείται για την εξέταση της ομοιογένειας μέσα στην ίδια ομάδα δεδομένων (όσο ↓CV, τόσο ↓μεταβλητότητα των παρατηρήσεων)


```
# The measure of relative variability is the coefficient of variation (CV)
sd(faithful$eruptions, na.rm=TRUE)/
   mean(faithful$eruptions, na.rm=TRUE)*100
[1] 32.72483

CV <- function(mean, sd){
+        (sd/mean)*100
+        }
CV(mean(faithful$eruptions), sd(faithful$eruptions))
[1] 32.72483
# ↑CVX : μεγάλη διασπορά σε σχέση με τη μέση τιμή, μεγάλη αβεβαιότητα
heights <- c(151, 160, 162, 155, 154, 168, 153, 158, 157, 150, 167)
mean(heights)
[1] 157.7273
sd(heights)
[1] 6.034748
CV(mean(heights),sd(heights))
[1] 3.826065
```


Συνάρτηση αυτοσυσχέτισης σε ακολουθία χρονοσειρών

- *Μετρά το διαχρονικό βαθμό συμμεταβολής μιας χρονοσειράς $\{X_t\}$ σε σχέση με την ίδια τη σειρά σε τ-χρόνο (διάστημα) υστέρησης*

$$\{X_t\},\ t=1,\ldots,n$$

$$r(\tau) = \frac{\frac{1}{n-\tau}\sum_{i=1+\tau}^{n}(x_i-\bar{x})(x_{i-\tau}-\bar{x})}{\frac{1}{n-1}\sum_{i=1+\tau}^{n}(x_i-\bar{x})^2}$$

Η r(τ) φανερώνει εάν και κατά πόσο πρότερες τιμές της χρονοσειράς περιλαμβάνουν πληροφόρηση (συσχέτιση) για την πορεία αυτής της χρονοσειράς στο παρόν και στο μέλλον

Συνδιακύμανση / Covariance (x,y)

- Συνδιακύμανση και η συσχέτιση είναι δύο βασικές έννοιες στη στατιστική και τη θεωρία πιθανοτήτων που, και οι δύο, μετρούν τη σχέση και την εξάρτηση μεταξύ δύο μεταβλητών.
- Η συνδιακύμανση υποδεικνύει την κατεύθυνση της γραμμικής σχέσης μεταξύ των μεταβλητών.
- Η συσχέτιση μετρά τόσο την ισχύ όσο και την κατεύθυνση της γραμμικής σχέσης μεταξύ δύο μεταβλητών

- *Στη συνδιακύμανση … Θετικές τιμές υποδεικνύουν θετική γραμμική σχέση, και αρνητικές το αντίθετο*

Για πληθυσμό

$$cov(X,Y) = E[(X-\mu_x)(Y-\mu_y)] =$$

$$E[XY] - \mu_x\mu_y = \frac{\sum_1^N (x_i-\mu_x)(y_i-\mu_y)}{N} = \sigma_{XY}$$

Για δείγμα

$$S_{XY} = \frac{\sum_1^n (x_i-\bar{x})(y_i-\bar{y})}{n-1}$$


```
cov(x, y = NULL, use = "all.obs",
    method = c("pearson", "kendall", "spearman"))
x       a numeric vector, matrix or data frame.
y       NULL (default) or a vector, matrix or data frame
        with compatible dimensions to x. The default is y = x.
na.rm   logical. Should missing values be removed?
use     an optional character string giving a method for computing
        covariances in the presence of missing values.
        This must be (an abbreviation of) one of the strings
        "all.obs", "complete.obs" or "pairwise.complete.obs".
method  a character string indicating which correlation coefficient
        (or covariance) is to be computed.
        One of "pearson" (default), "kendall", or "spearman",
        can be abbreviated.

duration = faithful$eruptions   # eruption durations
waiting = faithful$waiting      # the waiting period
cov(duration, waiting)          # apply the cov function
[1] 13.978
```


Άλλα μέτρα/παράμετροι μεταβλητότητας

Για μη ομαδοποιημένα δεδομένα

- Το **εύρος** *(range)* ή **έκταση** := το πλάτος των τιμών της μεταβλητής, *R = Max-Min*
  - Επηρεάζεται από ακραίες τιμές της μεταβλητής
  - Αδυνατεί να δώσει πληροφορίες για τη διασπορά των παρατηρήσεων μεταξύ των δύο ακραίων τιμών

```
R = max(faithful$eruptions) - min(faithful$eruptions)
R
[1] 3.5
range(faithful$eruptions)
[1] 1.6 5.1
```


Άλλα μέτρα/παράμετροι μεταβλητότητας

Για μη ομαδοποιημένα δεδομένα

- η **μέση απόκλιση** *(mean deviation)* **:=** ο αριθμητικός μέσος των απολύτων διαφορών των (*N* σε πλήθος) τιμών της μεταβλητής από την αναμενόμενη μέση τιμή *μ*,

  MD ή $MA = \sum_{i=1,2,\ldots n} | x_i - \mu | / N$

(+) λαμβάνει υπόψη όλες τις παρατηρήσεις

(-) δεν έχει κάποιο άνω φράγμα

```
USER DEFINED Mean Deviation Function
MA <- function(x, n, FUN){
          sum(abs(x-FUN)) / n
          }
# using the mean as centerpoint
FUN <- mean(faithful$eruptions)
MA(faithful$eruptions,length(faithful$eruptions),FUN)
[1] 1.042253

# using the median as centerpoint
FUN <- median(faithful$eruptions)
MA(faithful$eruptions,length(faithful$eruptions),FUN)
[1] 0.9724669
```


Ποσοστημόρια (Percentiles) και ποσοστιαία σημεία (Quantiles)

Εφαρμόζονται σε διατεταγμένα δεδομένα και αφορούν στην τιμή για την οποία το *p%* των παρατηρήσεων είναι μικρότερες ή ίσες από αυτή την τιμή *και το (1-p)% είναι μεγαλύτερες ή ίσες από αυτήν την τιμή*

- Ανάλογα με την τιμή του *p* διακρίνουμε πολλές ειδικές περιπτώσεις ($p_1,p_2,\ldots,p_{99}$ / $p_{10},p_{20},\ldots,p_{90}$

- Σε ένα σύνολο *n* παρατηρήσεων, το *Q-ποσοστιαίο σημείο* εντοπίζεται στην κλάση που περιέχει την q_n παρατήρηση ($p_{25}=Q_1$, $p_{50}=Q_2$, $p_{75}=Q_3$)

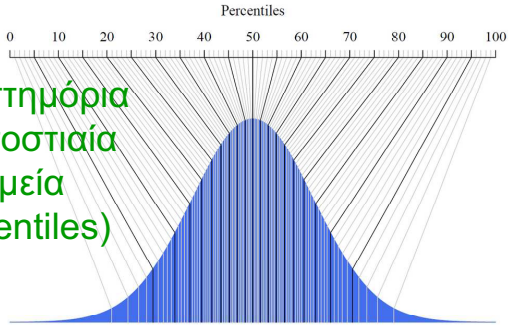
Ποσοστημόρια (Percentiles) και ποσοστιαία σημεία (Quantiles)

Εφαρμόζονται σε διατεταγμένα δεδομένα και αφορούν στην τιμή για την οποία το *p%* των παρατηρήσεων είναι μικρότερες ή ίσες από αυτή την τιμή *και το (1-p)% είναι μεγαλύτερες ή ίσες από αυτήν την τιμή*

- Ανάλογα με την τιμή του *p* διακρίνουμε πολλές ειδικές περιπτώσεις ποσοστιαίων σημείων

- π.χ., τα ποσοστά 25 και 75 τοις εκατό, υποδηλώνουν τα ποσοστιαία σημεία για τα οποία το ένα τέταρτο και τα τρία τέταρτα των δεδομένων είναι μικρότερα. Οι εν λόγω τιμές λαμβάνονται χρησιμοποιώντας τη συνάρτηση ***quantile ()***

Ποσοστημόρια ή ποσοστιαία σημεία (percentiles)

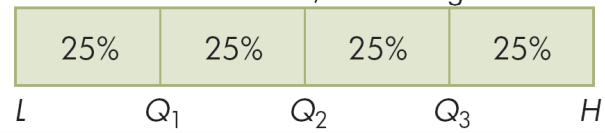


- Οι τιμές μιας τ.μ. που διαχωρίζουν το 100% ενός διατεταγμένου συνόλου δεδομένων σε *n* ίσα υποσύνολα → *n-1* ποσοστιαία σημεία

- Τριτημόρια: 2 θέσεις, 3 ίσα υποσύνολα (→ 33% της συνολικής πληροφορίας)
- Τεταρτημόρια: 3 θέσεις, 4 ίσα υποσύνολα (→ 25% της σ. πληροφορίας)
- Δεκατημόρια: 9 θέσεις, 10 ίσα υποσύνολα (→ 10% της σ. πληροφορίας)
- Εκατοστημόρια: 99 θέσεις, 100 ίσα υποσύνολα (→ 1% της σ. πληροφορίας)
- Χιλιοστημόρια: 999 θέσεις, 1000 ίσα υποσύνολα (→ 0.1% της σ. πληροφορίας)

Τεταρτημόρια (quartiles)

Ranked data, increasing order

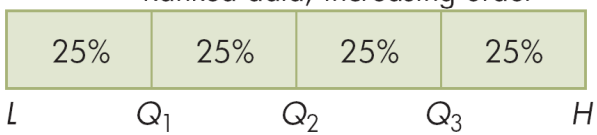


Οι τιμές της τ.μ. που χωρίζουν ένα σύνολο διατεταγμένων δεδομένων σε 4 υποσύνολα (τεταρτημόρια) → μέτρα κεντρικής τάσης

- Αποτελούν γενίκευση της έννοιας της διαμέσου

Τεταρτημόρια (quartiles)

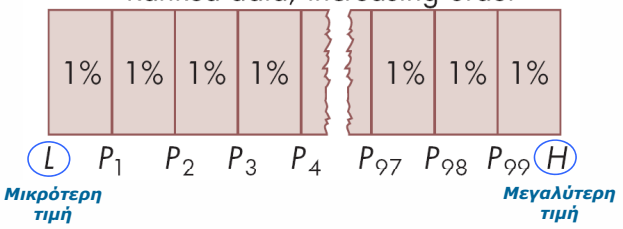
Ranked data, increasing order



- **1ο τεταρτημόριο, Q1**: τουλάχιστον 25% των δεδομένων έχουν τιμές μικρότερες του Q1
- **2ο τεταρτημόριο, Q2**, είναι η διάμεση τιμή
- **3ο τεταρτημόριο, Q3**: τουλάχιστον 25% των δεδομένων έχουν τιμές μεγαλύτερες του Q3

Εκατοστημόρια (percentiles)

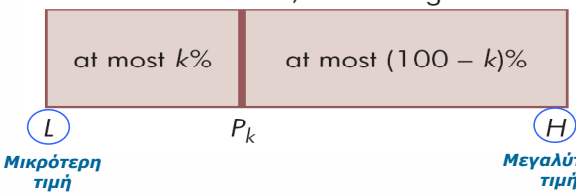
Ranked data, increasing order



- Οι τιμές μιας τ.μ. που διαχωρίζουν ένα διατεταγμένο σύνολο δεδομένων σε 100 ίσα υποσύνολα → 99 ποσοστιαία σημεία.

*k*th Percentile

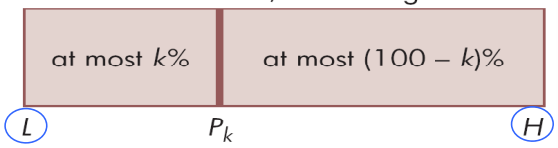
Ranked data, increasing order



- **k-ποσοστιαίο σημείο** της συνάρτησης κατανομής $f(x)$ μιας τυχαίας μεταβλητής X, $\forall x \in R$ είναι η τιμή x_k για την οποία ισχύει: $P(X < x_k) \le k$ και $P(X \le x_k) \ge k$

kth Percentile

Ranked data, increasing order



Μικρότερη τιμή — **Q1=P25, Q2=P50, Q3=P75** — Μεγαλύτερη τιμή

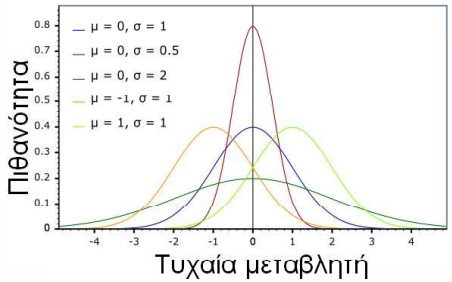
- ή αλλιώς: η τιμή των διατεταγμένων δεδομένων, όπου τουλάχιστον το k% των δεδομένων είναι κάτω από την τιμή P_k, και τουλάχιστον το (1-k)% των δεδομένων είναι πάνω από αυτή την τιμή

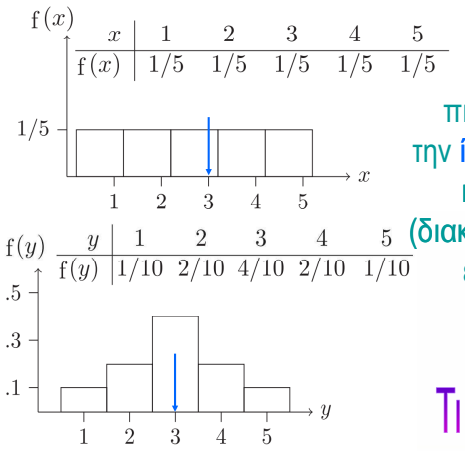

```
quantile(x, probs = seq(0, 1, 0.25), na.rm = FALSE,
         names = TRUE, type = 7, …)

x       numeric vector whose sample quantiles are wanted
        NA and NaN values are not allowed in numeric vectors
        unless na.rm is TRUE.
probs   numeric vector of probabilities with values in [0,1].
        (Values up to 2e-14 outside that range are accepted and
        moved to the nearby endpoint.)
na.rm   logical; if true, any NA and NaN are removed from x before
        the quantiles are computed.
names   logical; if true, the result has a names attribute.
        Set to FALSE for speedup with many probs.
type    an integer between 1 and 9 selecting one of
        nine quantile algorithms (see details in R help).
…       further arguments passed to or from other methods.
```



```
# The quantile() function can provide any quantile we want.
# This is done by using the 'probs' argument.

# The default value for the probs argument is a vector representing
# the minimum (0), the first quartile (0.25), the median (0.5), the third
# quartile (0.75), and the maximum (1).

quantile(faithful$eruptions)
     0%     25%     50%     75%    100%
1.60000 2.16275 4.00000 4.45425 5.10000

quantile(faithful$eruptions, probs = c(0.05, 0.20, 0.95))
     5%     20%     95%
1.8000 2.0034 4.8170
```


- Οι περισσότερες κατανομές πιθανότητας περιγράφονται πλήρως όταν γνωρίζουμε τη μέση τιμή και τη διασπορά (διακύμανση) τους.





Υπάρχουν κατανομές πιθανότητας με την ίδια μέση τιμή και διασπορά (διακύμανση) που έχουν τελείως διαφορετική μορφή

Τι κάνουμε ?

Στατιστικές ροπές

- *Στη στατιστική, είναι ποσότητες που σχετίζονται με το 'σχήμα' ή 'μορφή' ενός συνόλου δεδομένων, π.χ. "πως μοιάζει ένα ιστόγραμμα συχνοτήτων με βάση τις δεδομένες τιμές" – δηλ. πόσο επεκταμένο είναι, πόσο συμμετρικό , ...*
- Πρακτικά, μια ροπή τάξης k είναι ο μέσος όρος όλων των τιμών στο εκάστοτε σύνολο δεδομένων, με κάθε αριθμό στην k δύναμη προτού υπολογιστεί ο μέσος όρος. Για παράδειγμα, η ροπή 1ης τάξης είναι ο αριθμητικός μέσος όρος, η 2ης τάξης είναι ο μέσος όρος των τετραγώνων...

Στατιστικές ροπές

- *Η ροπή k τάξης (k=1,2,3,…) μιας τυχαίας μεταβλητής X, ως προς το κέντρο της κατανομής της*

$$\mu'_k = E[X^k] = \begin{cases} \sum_X x^k P(X=x) & \text{διακριτή τ.μ. X} \\ \int_{-\infty}^{+\infty} x^k P(x)dx & \text{συνεχής τ.μ. X} \end{cases}$$

Στατιστικές ροπές

- *Αντίστοιχα ορίζεται η ροπή k τάξης μιας τυχαίας μεταβλητής X, ως προς τον μέσο μ (central moment, κεντρική ροπή k-τάξης) της κατανομής της*

$$\mu_k = E[(X-\mu)^k] = \begin{cases} \sum_X (X-\mu)^k P(X=x) & \text{διακριτή τ.μ. X} \\ \int_{-\infty}^{+\infty} (X-\mu)^k P(x)dx & \text{συνεχής τ.μ. X} \end{cases}$$

- ***Η κεντρική ροπή 1ης τάξης*** *είναι η* ***μέση τιμή μ***

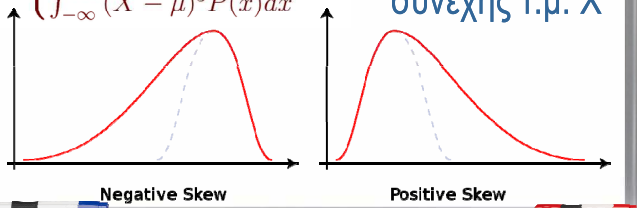
$$\mu'_1 = E[X] = \mu = \begin{cases} \sum_X xP(X=x) & \text{διακριτή τ.μ. X} \\ \int_{-\infty}^{+\infty} xP(x)dx & \text{συνεχής τ.μ. X} \end{cases}$$

- ***Η ροπή 2ης τάξης*** *ως προς τη μέση τιμή είναι* ***η διακύμανση σ²***

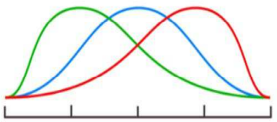
$$\mu_2 = E[(X-\mu)^2] = \sigma^2 = \begin{cases} \sum_X (X-\mu)^2 P(X=x) & \text{διακριτή τ.μ. X} \\ \int_{-\infty}^{+\infty} (X-\mu)^2 P(x)dx & \text{συνεχής τ.μ. X} \end{cases}$$

- Η **κεντρική ροπή *3ης τάξης*** δίνει πληροφορίες για το βαθμό ασυμμετρίας της κατανομής *μιας τ.μ. X* → ***λοξότητα*** *(skewness)*

$$\mu_3 = E[(X-\mu)^3] = \gamma = \begin{cases} \sum_X (X-\mu)^3 P(X=x) & \text{διακριτή τ.μ. X} \\ \int_{-\infty}^{+\infty} (X-\mu)^3 P(x)dx & \text{συνεχής τ.μ. X} \end{cases}$$



- Μια κανονική κατανομή (καμπύλη καμπάνας) εμφανίζει μηδενική λοξότητα.
- Μια θετική λοξότητα υποδεικνύει ότι *η κατανομή είναι δεξιά λοξή ή με δεξιά ουρά ή ότι κλίνει προς τα δεξιά* – σε σχέση με μια κανονική κατανομή
- Ο μέσος όρος θετικά λοξών δεδομένων θα είναι μεγαλύτερος από τον διάμεσο τιμή τους



- Σε μια κατανομή που είναι αρνητικά λοξή, συμβαίνει ακριβώς το αντίθετο
- Υπάρχουν διάφοροι τρόποι μέτρησης της λοξότητας

Λοξότητα ή ασυμμετρία (skewness)

- Συντελεστές ασυμμετρίας κατά Pearson

$$\gamma_1 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{\mathrm{E}[(X-\mu)^3]}{(\mathrm{E}[(X-\mu)^2])^{3/2}} = \dots = \frac{\mathrm{E}[X^3] - 3\mu\sigma^2 - \mu^3}{\sigma^3}$$

1ος

- Συντελεστές ασυμμετρίας κατά Pearson — 2ος

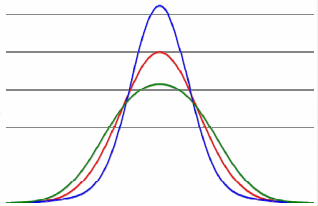
$$\gamma_2 = \frac{3(\bar{x} - M)}{s}$$

Κύρτωση (kurtosis ή peakedness)

- *Κατά παρόμοιο τρόπο με την έννοια της ασυμμετρίας,* η κεντρική ροπή *4ης τάξης* ως προς τη μέση τιμή *περιγράφει το σχήμα της κατανομής πιθανότητας μιας τ.μ. X*

$$\mu_4 = E[(X-\mu)^4] = \beta = \begin{cases} \sum_X (X-\mu)^4 P(X=x) & \text{διακριτή τ.μ. X} \\ \int_{-\infty}^{+\infty} (X-\mu)^4 P(x)dx & \text{συνεχής τ.μ. X} \end{cases}$$

- Η κύρτωση δείχνει τον βαθμό συγκέντρωσης ή αιχμηρότητας (δηλ. οξύτητα της κορυφής) που εμφανίζουν οι τιμές της μεταβλητής γύρω από τη μέση τιμή και τα άκρα της κατανομής



  - Δύο ή περισσότερες κατανομές ακόμη και αν έχουν τον ίδιο συντελεστή ασυμμετρίας είναι δυνατόν να διαφέρουν μεταξύ τους → *αυτό που τις διαφοροποιεί είναι η κύρτωση*
  - Όταν υπάρχει υψηλή κύρτωση, οι ουρές εκτείνονται μακρύτερα από τις τρεις τυπικές αποκλίσεις της κανονικής καμπύλης κατανομής

- Συντελεστής κύρτωσης κατά Pearson

κεντρική ροπή 4ης τάξης

$$\beta_2 = \frac{\mathrm{E}[(X-\mu)^4]}{(\mathrm{E}[(X-\mu)^2])^2} = \frac{\mu_4}{\sigma^4}$$

- Συχνά υπολογίζεται από τη σχέση

$$\beta_2 = \frac{\mu_4}{\sigma^4} - 3$$

που προκύπτει από την τιμή $\beta_2=3$ της κύρτωσης για την κανονική κατανομή ($\beta_2 > 0$ → *λεπτόκυρτες κατανομές*, $\beta_2 < 0$ → *πλατύκυρτες κατανομές*)

- **Η κύρτωση μετρά τις ακραίες τιμές στην 'ουρά', όχι την 'αιχμή'.** Μια κατανομή μπορεί να κορυφωθεί απεριόριστα με χαμηλή κύρτωση ή μπορεί να είναι τέλεια επίπεδη με άπειρη κύρτωση.

Γεννήτριες συναρτήσεις στατιστικών ροπών

- *Σε ορισμένες περιπτώσεις, η μέση τιμή* $\mu = E(X)$ *και η διακύμανση* $\sigma^2 = Var(X) = E(X^2) - \mu^2$ *οι οποίες είναι ροπές χαμηλής τάξης, είναι δύσκολο να υπολογιστούν.*
- *Οι λεγόμενες ειδικές* ***γεννήτριες συναρτήσεις ροπών*** *(ή ροπογεννήτριες) μπορούν μερικές φορές να απλουστεύσουν τον υπολογισμό για τις αναμενόμενες τιμές* $E(X)$, $E(X^2)$, $E(X^3)$, ..., *και* $E(X^r)$ *μιας τυχαίας μεταβλητής*

Moment generating functions (MGF)

$$M(t) = E(e^{tX}) = \sum_x e^{tx} f(x)$$

- Παρέχουν τη βάση για μια εναλλακτική διαδρομή αναλυτικών υπολογισμών σε σχέση με την άμεση χρήση των συναρτήσεων πυκνότητας πιθανότητας ή της αθροιστικής συνάρτησης κατανομής,
- με την προϋπόθεση ότι το παραπάνω άθροισμα υπάρχει και είναι πεπερασμένο για -h<t<h, όπου h>0
- Ωστόσο, ΣΗΜΕΙΩΣΤΕ ότι όλες οι τ.μ. δεν έχουν απαραίτητα γεννήτριες συναρτήσεις ροπών.

Moment generating functions (MGF)

Ένας εναλλακτικός τρόπος περιγραφής της κατανομής πιθανοτήτων

$$M(t) = E(e^{tX}) = \sum_x e^{tx} f(x)$$

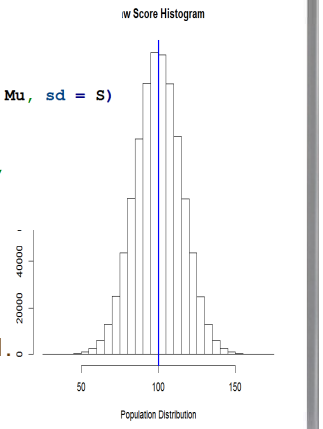
$$\mu = E(X) = M'(0)$$

$$\sigma^2 = E(X^2) - [E(X)]^2 = M''(0) - [M'(0)]^2$$

$$M^{(r)}(0) = E(X^r)$$

- δηλ., εφόσον υπάρχει η κατάλληλη γεννήτρια συνάρτηση ροπών, οι αναμενόμενες τιμές $E(X)$, $E(X^2)$,..., $E(X^r)$ *μιας τυχαίας μεταβλητής μπορούν να υπολογιστούν από τις παραγώγους* $M^{(r)}(t=0)$

```r
set.seed(666)
n=1e6     #1000000 people
Mu=100    # Set pop mean
S = 15    # set pop SD
Normal.Sample.1<-rnorm(n, mean = Mu, sd = S)
hist(Normal.Sample.1,
  main="Raw Score Histogram",
  xlab="Population Distribution",
    ylab="Frequency")
abline(v=100, col="blue")
```



Υπάρχουν 4 ροπές που μπορούν να συλλάβουν σημαντικές πτυχές της κατανομής των δεδομένων (1. κέντρο, 2. εξάπλωση, 3. συμμετρία και 4. αιχμή).


```r
#  Moment 1: Central Tendency. Captures the center of the distribution
#  and be defined three classical ways (Mean, Median, Mode).

mean(Normal.Sample.1)                 # Estimate 'mean'
[1] 99.99483
median(Normal.Sample.1)               # Estimate 'median'
[1] 99.98384

#  Create a function for computing 'mode'
estimate_mode <- function(x) {
  d <- density(x)
  d$x[which.max(d$y)]
}
estimate_mode(Normal.Sample.1)        # Estimate 'mode'
[1] 99.60596
```



```r
#  Moment 2: Spread. Captures the spread of the distribution
#  and can be defined in many classical ways (variance, standard deviation,
#  median absolute difference, interquartile range, ...).
Varience.M<-var(Normal.Sample.1)  ; Varience.M
[1] 224.8388
sd <- sqrt(Varience.mu)           ; sd
[1] 14.99463

#  Moment 3 (Skewness) and Moment 4 (Kurtosis)
#  require use of the R package 'moments'
install.packages(moments)
library(moments)

SK.Calc<-skewness(Normal.Sample.1)
SK.Calc      # print Moment 3: Skewness
[1] -0.0009853765
K.Calc<-kurtosis(Normal.Sample.1)-3
K.Calc       # print Moment 4: Kurtosis
[1] -0.0009629761
```



```r
# install required packages
> install.packages('moments')
# load installed package
> library(moments)
>
> # create a vector of data
> my_data<-c(98,87,96,91,85,89,93,96,99,86)
>
> #Raw Moments: all.moments(x,order.max,central=FALSE)
> print('Raw Moments')
[1] "Raw Moments"
> print(all.moments(my_data))
[1]    1.0   92.0 8487.8
>
> # Central Moments
> print('Central Moments')
[1] "Central Moments"
> print(all.moments(my_data,central=TRUE))
[1]  1.0  0.0 23.8
```


Κριτήρια/Ιδιότητες 'καλών' εκτιμητών

Αμεροληψία *(Unbiasedness)*

- Αμερόληπτος εκτιμητής

Συνέπεια *(Consistency)*

- Συνεπής εκτιμητής

Αποτελεσματικότητα (Efficiency)

- Αποτελεσματικός εκτιμητής

Επάρκεια ή απόδοση (Sufficiency)

- Επαρκής εκτιμητής

Αμεροληψία *(Unbiasedness)*

- Η αναμενόμενη τιμή της εκτίμησης ισούται με την πραγματική τιμή της παραμέτρου ενδιαφέροντος.
- ***E(û) = u*** → ***û***, *για κάθε πιθανή τιμή της u*
- *Το μέσο σφάλμα ή μεροληψία της εκτιμήτριας (bias)*: ***B(û) = E(û) – u***

Αμερόληπτος Εκτιμητής (Unbiased Estimator) όταν, κατά μέσο όρο, δεν υποτιμά ή υπερεκτιμά την πραγματική τιμή της παραμέτρου ενδιαφέροντος, για οποιοδήποτε μέγεθος δείγματος n

Παραδείγματα 'αμερόληπτων' εκτιμητριών

$$E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}E\left[\sum_{i=1}^{n} X_i\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}n\mu = \mu$$

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

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n}E((X_i - \mu)^2) - nE((\bar{X} - \mu)^2)\right) = \ldots = \sigma^2$$

Παράδειγμα 'μη αμερόληπτης' εκτιμήτριας

- Αντίθετα, η στατιστική συνάρτηση $\bar{X}^2$ δεν είναι αμερόληπτη εκτιμήτρια του μ^2

$$E(\bar{X}^2) = V(\bar{X}) + E(\bar{X})^2 = \frac{\sigma^2}{n} + \mu^2 \neq \mu^2$$

αλλά είναι μια ασυμπτωτικά αμερόληπτη εκτιμήτρια του μ^2

$$E(\bar{X}^2) = \frac{\sigma^2}{n} + \mu^2 \xrightarrow[n\to\infty]{} \mu^2$$

- Καθώς $\sigma^2 = E[(X-\mu)^2]$, ο εκτιμητής

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \mu)^2$$

με βάση τη γραμμικότητα του τελεστή προσδοκίας $E[\ ]$, συχνά θεωρείται ως ένας **αμερόληπτος εκτιμητής για τη διακύμανση** σ^2 μιας τ.μ. X

- Επίσης, με βάση το νόμο των μεγάλων αριθμών, είναι επίσης ένας *σταθερός εκτιμητής* για τη διακύμανση.

- Ωστόσο, στην πράξη συχνά δεν γνωρίζουμε την τιμή μ (πραγματική μέση τιμή του πληθυσμού).

- Σε αντίθεση, ο εκτιμητής

$$\overline{S}^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X})^2 = \frac{1}{n}\left(\sum_{k=1}^{n} X_k^2 - n\overline{X}^2\right)$$

είναι ένας **μη αμερόληπτος εκτιμητής** για τη διακύμανση σ^2 καθώς

$$\widehat{\overline{S}}^2 = E\left[\frac{1}{n}\left(\sum_{k=1}^{n} X_k^2 - n\overline{X}^2\right)\right] = \ldots = \frac{(n-1)*\sigma^2}{n} \neq \sigma^2$$

με προκατάληψη (σφάλμα)

$$Bias = B(\overline{S}^2) = E[\overline{S}^2] - \sigma^2 = -\frac{\sigma^2}{n}$$

και τιμή που είναι πολύ μικρή όταν το δείγμα είναι σχετικά μεγάλο (n → ∞)

Συνέπεια *(Consistency)*

- Καθώς αυξάνεται το μέγεθος του δείγματος, η εκτιμήτρια $\hat{u}$ συγκλίνει κατά μέτρο (πιθανότητα) στην πραγματική τιμή της παραμέτρου που εκτιμάται :

$$\lim_{n\to\infty} P(|\hat{u} - u| > \varepsilon) = 0, \forall \varepsilon > 0$$

Συνεπής Εκτιμητής (Consistent Estimator), εκείνος που προσδιορίζει την πραγματική τιμή της παραμέτρου που εκτιμάται όλο και πιο ακριβέστερα καθώς αυξάνεται το μέγεθος του δείγματος

Παραδείγματα 'συνεπών' εκτιμητριών

- Η δειγματική μέση τιμή, εκτός από αμερόληπτη, είναι και μια συνεπής εκτιμήτρια γιατί:

$V(\overline{X}) = \sigma^2/n$ → δηλ., καθώς το *n* αυξάνει, η διακύμανση του δειγματικού μέσου μειώνεται

- Ο αριθμητικός μέσος $x_d = ½(x_{min}+x_{max})$ δεν είναι επαρκής εκτιμήτρια της μέσης τιμής μ

Παραδείγματα 'συνεπών' εκτιμητριών

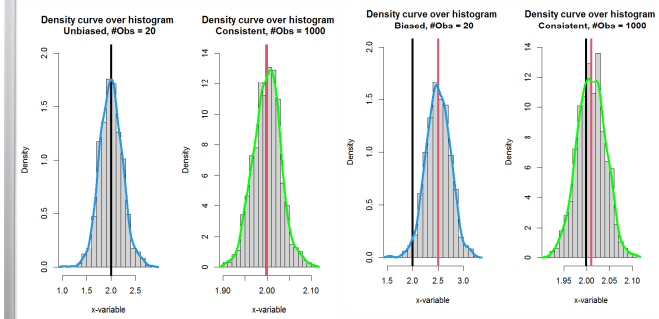
- Η δειγματική ροπή $m_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r$

- Είναι μια συνεπής εκτιμήτρια της ροπής $\mu_r' = E(X_i^r)$ r-τάξης ενός πληθυσμού, γιατί από ένα τυχαίο δείγμα $X_1, X_2, \ldots, X_n$ από κατανομή με σ.κ.π. $F(X, \mu_r')$ προκύπτει

$$E(m_r) = \frac{1}{n}\sum_{i=1}^{n} E(X_i^r) = \frac{1}{n} n\mu_r' = \mu_r'$$

$$V(m_r) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i^r) = \frac{1}{n^2} nV(X_1^r) = \frac{V(X_1^r)}{n} \xrightarrow[n\to\infty]{} 0$$

Παραδείγματα 'συνεπών' εκτιμητριών



Αποτελεσματικότητα (Efficiency)

- Εάν η εκτιμήτρια τιμή $\hat{u}$ έχει την ελάχιστη διασπορά μεταξύ όλων των αμερόληπτων εκτιμητών : $MSE = Var(\hat{u}) - (E[\hat{u}] - u)^2$

Εάν υπάρχει τέτοιος εκτιμητής → μοναδικός

Αποτελεσματικός Εκτιμητής (Efficient Estimator), είναι ο "καλύτερος δυνατόν" ή "βέλτιστος" εκτιμητής μιας παραμέτρου ενδιαφέροντος

Παραδείγματα 'αποτελεσματικών' εκτιμητριών

Για κάποιο δείγμα $\{x_1, x_2, \ldots, x_n\}$, $n=2m+1$ *η δειγματική μέση τιμή είναι μια πιο αποτελεσματική εκτιμήτρια από ότι η διάμεσος για την εκτίμηση του πληθυσμιακού μέσου*

$$var(\text{sample median}) = \frac{\pi\sigma^2}{4m} = \frac{\pi(2m+1)}{4m} var(\text{sample mean})$$

$\sim(\pi/2)=1.571$

- Γενικά, εάν υπάρχουν δύο αμερόληπτοι εκτιμητές $\hat{u}_1$ *και* $\hat{u}_2$ μιας παραμέτρου *u*, αυτός που έχει την μικρότερη διακύμανση (π.χ., $\sigma_{\hat{u}1}^2 < \sigma_{\hat{u}2}^2$) λέγεται ότι είναι σχετικά αποδοτικός ή πιο αποτελεσματικός *(relative efficient)*

Επάρκεια ή **απόδοση** (Sufficiency)

- Εάν η εκτιμήτρια $\hat{u}$ χρησιμοποιεί όλη την πληροφορία από ένα *δείγμα* $\{x_1, x_2, \ldots, x_n\}$ που σχετίζεται με την παράμετρο u

Επαρκής Εκτιμητής (Sufficient Estimator), π.χ. επαρκής εκτιμητής (του μέσου όρου του πληθυσμού) είναι ο μέσος όρος ενός δείγματος, καθότι χρησιμοποιεί όλες τις παρατηρήσεις

Παραδείγματα 'επαρκών' εκτιμητριών

- Η δειγματική μέση τιμή και η δειγματική διασπορά είναι επαρκείς εκτιμητές για τη μορφή μιας γνωστής κανονικής κατανομής *(→ επιτυγχάνουν μια απλοποίηση (ένα περιορισμό ή μία συρρίκνωση) των δεδομένων του δείγματος χωρίς απώλεια πληροφορίας)*, ενώ το εύρος δεν είναι (γιατί χρησιμοποιεί μόνο τις ακραίες τιμές)
- Από την άλλη πλευρά, η διάμεση τιμή δεν είναι επαρκής για τη μέση τιμή, καθώς ακόμη και αν η διάμεσος του δείγματος είναι γνωστή, από το ίδιο το δείγμα μπορούν να εξαχθούν περαιτέρω πληροφορίες σχετικά με τον πληθυσμιακό μέσο

... στη συνέχεια

- Πως εκφράζουμε την εμπιστοσύνη μας στις στατιστικές εκτιμήσεις παραμέτρων ενδιαφέροντος