Ανάλυση Δεδομένων με χρήση του Στατιστικού Πακέτου R

Δημήτρης Φουσκάκης,
Καθηγητής,
Τομέας Μαθηματικών,
Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών,
Εθνικό Μετσόβιο Πολυτεχνείο.

Περιεχόμενα

Τι είναι Στατιστική

"Στατιστική είναι η επιστήμη που ασχολείται με τη συλλογή, παρουσίαση και εν συνεχεία εξαγωγή συμπερασμάτων παρατηρήσεων που υπόκεινται σε τυχαίες μεταβολές."

Στατιστική είναι η επιστήμη της *αβεβαιότητας*, πώς μπορούμε να την ποσοτικοποιήσουμε και να την περιορίσουμε.

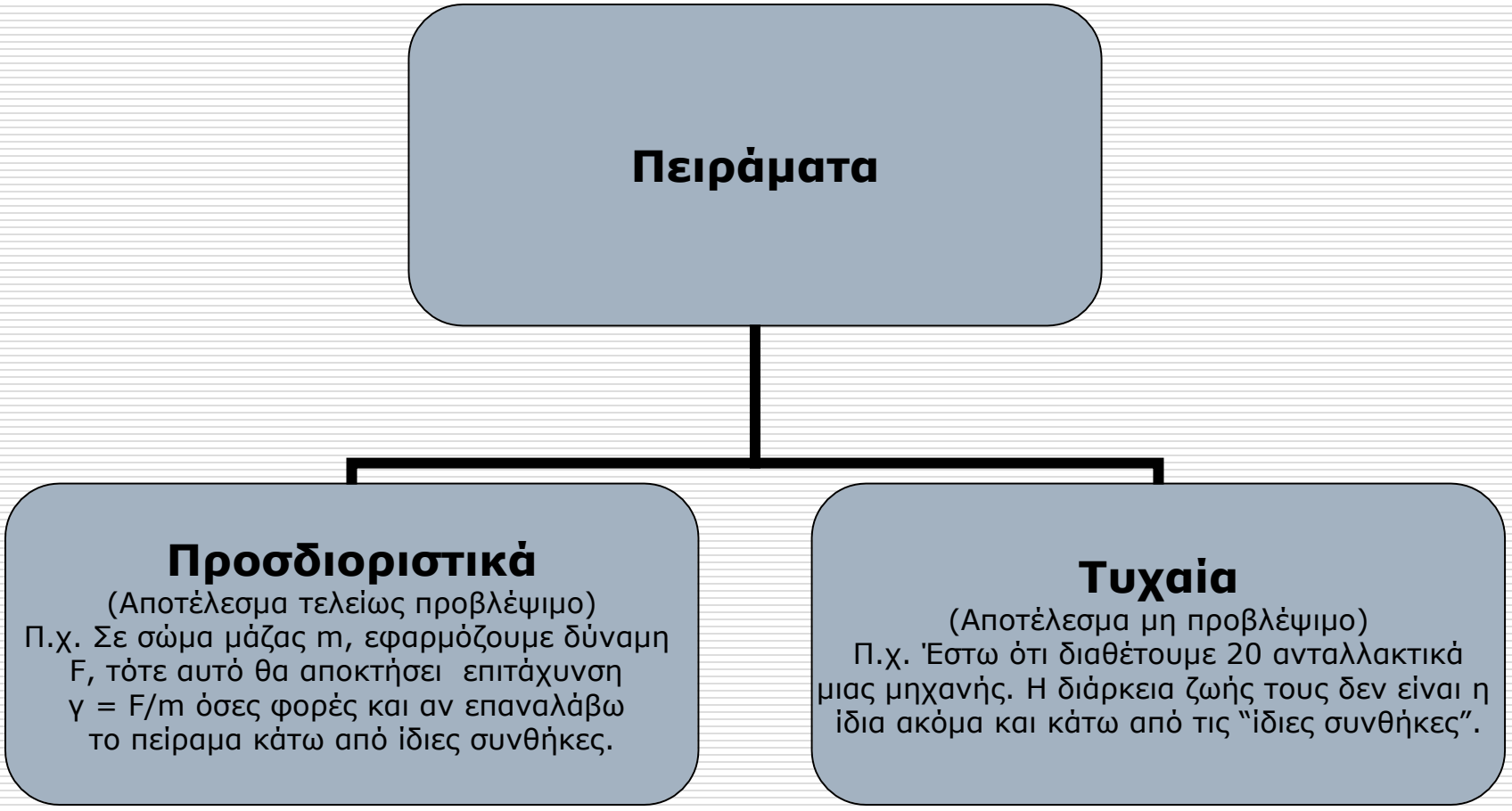
Ο λόγος ύπαρξης αυτής της αβεβαιότητας έχει να κάνει με την ακόλουθη διαπίστωση: πειράματα τα οποία επαναλαμβάνονται κάτω από ίδιες (όπως πιστεύουμε) συνθήκες δεν μας δίνουν πάντα το ίδιο αποτέλεσμα. Τέτοια πειράματα στα οποία δεν μπορούμε να γνωρίζουμε από πριν το αποτέλεσμά τους καλούνται *πειράματα τύχης*.

Τι είναι Στατιστική

Για παράδειγμα για να πάω οδικώς κάθε μέρα στη δουλειά από το σπίτι, *ποσότητες* όπως ο αριθμός των κόκκινων φαναριών στα οποία πρέπει να περιμένω ή ο συνολικός χρόνος που θα κάνω για να φτάσω στη δουλειά ποικίλουν από μέρα σε μέρα, ακόμα και αν φεύγω από το σπίτι την ίδια πάντα ώρα και χρησιμοποιώ πάντα την ίδια διαδρομή.

Στο παραπάνω παράδειγμα θα ήταν πολύ χρήσιμο να μπορούσα να καταλάβω ποιοι παράγοντες είναι η *αιτία* της *μεταβλητότητας* που παρατηρώ στον χρόνο που χρειάζεται για να πάω στη δουλειά έτσι ώστε στο μέλλον να μπορώ με *ακρίβεια* να *προβλέψω* τον χρόνο που θα χρειασθώ έως ότου φτάσω στη δουλειά.

Πειράματα


Πειράματα
Προσδιοριστικά
(Αποτέλεσμα τελείως προβλέψιμο)
Π.χ. Σε σώμα μάζας m, εφαρμόζουμε δύναμη F, τότε αυτό θα αποκτήσει επιτάχυνση γ = F/m όσες φορές και αν επαναλάβω το πείραμα κάτω από ίδιες συνθήκες.
Τυχαία
(Αποτέλεσμα μη προβλέψιμο)
Π.χ. Έστω ότι διαθέτουμε 20 ανταλλακτικά μιας μηχανής. Η διάρκεια ζωής τους δεν είναι η ίδια ακόμα και κάτω από τις "ίδιες συνθήκες".

Δεδομένα

- **Δεδομένα**: Πώς μπορούμε λοιπόν να ποσοτικοποιήσουμε την αβεβαιότητα για το αποτέλεσμα ενός τυχαίου πειράματος; Ένας τρόπος είναι να επαναλάβουμε το πείραμα αυτό αρκετές φορές και να καταγράψουμε τις αριθμητικές τιμές που θα προκύψουν για την ποσότητα ή τις ποσότητες του πειράματος που μας ενδιαφέρουν. Έτσι θα έχουμε μια ιδέα για τη μεταβλητότητα των τιμών αυτών των ποσοτήτων και ίσως κατανοήσουμε την αιτία αυτής. Αυτές οι τιμές αποτελούν τα *δεδομένα* ή αλλιώς τις *παρατηρήσεις* μας. Συνήθως τα δεδομένα τα καταγράφουμε σε έναν πίνακα με στήλες όσες οι ποσότητες του πειράματος που μας ενδιαφέρουν και γραμμές όσες οι επαναλαμβανόμενες μετρήσεις που καταγράψαμε.

Δεδομένα

Για παράδειγμα ο επόμενος πίνακας παρουσιάζει τα δεδομένα από το παράδειγμα της μετακίνησής μου από το σπίτι στη δουλειά. Το πείραμα έχει επαναληφθεί 10 φορές και η πρώτη στήλη του πίνακα εκφράζει τον χρόνο (σε λεπτά) της διαδρομής και η δεύτερη τον αριθμό των κόκκινων φαναριών που συνάντησα.

$$\begin{bmatrix} 22 & 2 \\ 31 & 4 \\ 18 & 2 \\ 25 & 1 \\ 22 & 0 \\ 21 & 1 \\ 19 & 1 \\ 35 & 5 \\ 18 & 2 \\ 17 & 1 \end{bmatrix}$$

Μεταβλητές - Παρατηρήσεις

Οι γραμμές του προηγούμενου λοιπόν πίνακα παριστάνουν τις τιμές που προέκυψαν για κάθε ποσότητα του πειράματος που μας ενδιαφέρει έπειτα από κάθε επανάληψη και **κάτω από τις ίδιες φαινομενικά συνθήκες** (ίδια διαδρομή, ίδια ώρα αναχώρηση). Οι ποσότητες αυτές επειδή μεταβάλλονται κατά τυχαίο τρόπο καλούνται *μεταβλητές* και συμβολίζονται συνήθως με κεφαλαία γράμματα του λατινικού αλφάβητου. Ας καλέσουμε λοιπόν Υ τη μεταβλητή που εκφράζει τον χρόνο διαδρομής και Χ τη μεταβλητή που εκφράζει τον αριθμό των κόκκινων φαναριών που συναντάμε.

Τα δεδομένα λοιπόν δεν είναι τίποτα άλλο από *παρατηρήσεις* των αριθμητικών τιμών που πήραν οι εν λόγω μεταβλητές επαναλαμβάνοντας το πείραμα 10 φορές. Οι επαναλήψεις είναι *ισόνομες* (ίδιες φαινομενικά συνθήκες), στον βαθμό που μπορούμε τουλάχιστον να πούμε, και *ανεξάρτητες* μεταξύ τους, δηλαδή το αποτέλεσμα μίας επανάληψης δεν επηρεάζεται από αυτό κάποιας άλλης επανάληψης.

Τυχαίο Δείγμα

- **Απλή Τυχαία Δειγματοληψία**:

  Οι μεταβλητές όπως τις ορίσαμε πριν είναι *χαρακτηριστικά* των μονάδων ενός συνόλου το οποίο το καλούμε *πληθυσμό*. Στο παράδειγμά μας ο πληθυσμός είναι οι διαδρομές που κάνω κάθε πρωί από το σπίτι στην δουλειά. Για να ποσοτικοποιήσουμε την αβεβαιότητά μας για την μεταβλητότητα των ποσοτήτων που μας ενδιαφέρουν, όπως είπαμε και πριν, επαναλαμβάνουμε το πείραμα αρκετές φορές, επιλέγουμε δηλαδή στοιχεία από τον πληθυσμό (δηλαδή διαδρομές) και παρατηρούμε τις τιμές που προκύπτουν για τις μεταβλητές που μας ενδιαφέρουν. Αυτή η ομάδα του πληθυσμού από την οποία συλλέγουμε την απαραίτητη πληροφορία καλείται *δείγμα*. Έχει ιδιαίτερη σημασία το παραγόμενο δείγμα να είναι *αντιπροσωπευτικό* του πληθυσμού, δηλαδή οι μεταβλητές που μας ενδιαφέρουν, αλλά και άλλα χαρακτηριστικά αυτού του πληθυσμού που επηρεάζουν τις τιμές που λαμβάνουν αυτές οι μεταβλητές, να συμπεριφέρονται με τον ίδιο τρόπο στο δείγμα και στον πληθυσμό.

Τυχαίο Δείγμα

Ένα τέτοιο δείγμα λέγεται τότε *τυχαίο*, και μπορεί να προκύψει με την *απλή τυχαία δειγματοληψία*, κατά την οποία κάθε μονάδα του πληθυσμού έχει την ίδια πιθανότητα να συμπεριληφθεί στο δείγμα.

Στο παράδειγμά μας θα ήταν π.χ. καλό στο δείγμα μας να είχαμε συμπεριλάβει διαδρομές από διαφορετικές μέρες της εβδομάδας (ίσως κάποιες μέρες έχει περισσότερη κίνηση από κάποιες άλλες μέρες), με διαφορετικές καιρικές συνθήκες, διαδρομές από μέρες με απεργία των μέσων μεταφοράς, κ.λ.π.

Τρόποι Επιλογής Αντιπροσωπευτικού Δείγματος

- **Απλή Τυχαία Δειγματοληψία**. Κάθε μονάδα του πληθυσμού έχει την ίδια πιθανότητα να συμπεριληφθεί στο δείγμα.
- **Συστηματική Δειγματοληψία**. Η επιλογή του δείγματος γίνεται από μια αριθμημένη λίστα των μονάδων του πληθυσμού, π.χ. επιλέγουμε 1 μονάδα ανά 20 μονάδες του πληθυσμού.
- **Στρωματοποιημένη Δειγματοληψία**. Ο πληθυσμός χωρίζεται σε διάφορες μη αλληλοεπικαλυπτόμενες ομάδες (στρώματα) με βάση κάποιο κοινό χαρακτηριστικό (μεταβλητή), το οποίο είναι έντονα διαφοροποιημένο από ομάδα σε ομάδα και σχετίζεται άμεσα με το υπό εξέταση χαρακτηριστικό του πληθυσμού, και προβαίνουμε σε απλή τυχαία δειγματοληψία για κάθε ομάδα χωριστά. Αν θέλουμε π.χ. να μελετήσουμε την επίδοση των φοιτητών στο μάθημα Στατιστικής, μπορούμε να χωρίσουμε τον πληθυσμό με βάση το φύλο έτσι ώστε στο δείγμα μας να συμπεριλάβουμε άντρες και γυναίκες.
- **Κατά συστάδες Δειγματοληψία**. Ο πληθυσμός χωρίζεται σε συστάδες, π.χ. με βάση γεωγραφικά κριτήρια, και απλά τυχαία δείγματα επιλέγονται από κάθε συστάδα. Στο παράδειγμα με την επίδοση τον φοιτητών μπορούμε να χωρίσουμε π.χ. τον πληθυσμό στα διάφορα τμήματα που γίνεται το μάθημα.
- **Πολυεπίπεδη δειγματοληψία**. Επιλογή δειγμάτων από δείγματα, στο παραπάνω παράδειγμα επιλέγουμε τυχαία πανεπιστήμια, μετά τμήματα και μετά φοιτητές.

Μέγεθος δείγματος

- Πόσο μεγάλο πρέπει να είναι το μέγεθος του δείγματος;
  - Το μέγεθος εξαρτάται από τη *μεταβλητότητα* της εξεταζόμενης μεταβλητής, όσο μικρότερη τόσο μικρότερο μέγεθος χρειαζόμαστε.
  - Αν επιθυμούμε *μεγαλύτερη ακρίβεια* (*μικρότερα τυπικά σφάλματα*) στις εκτιμήσεις μας χρειαζόμαστε μεγαλύτερο μέγεθος δείγματος.
  - Το *είδος της Στατιστικής Ανάλυσης*, πιο πολύπλοκη στατιστική ανάλυση απαιτεί πιο μεγάλο δείγμα.
- Απαντώντας στα παραπάνω ερωτήματα υπάρχουν έτοιμοι μαθηματικοί τύποι (για τις "απλές" στατιστικές αναλύσεις) που μας δίνουν το ελάχιστο μέγεθος δείγματος που χρειαζόμαστε για συγκεκριμένη ακρίβεια στις εκτιμήσεις.

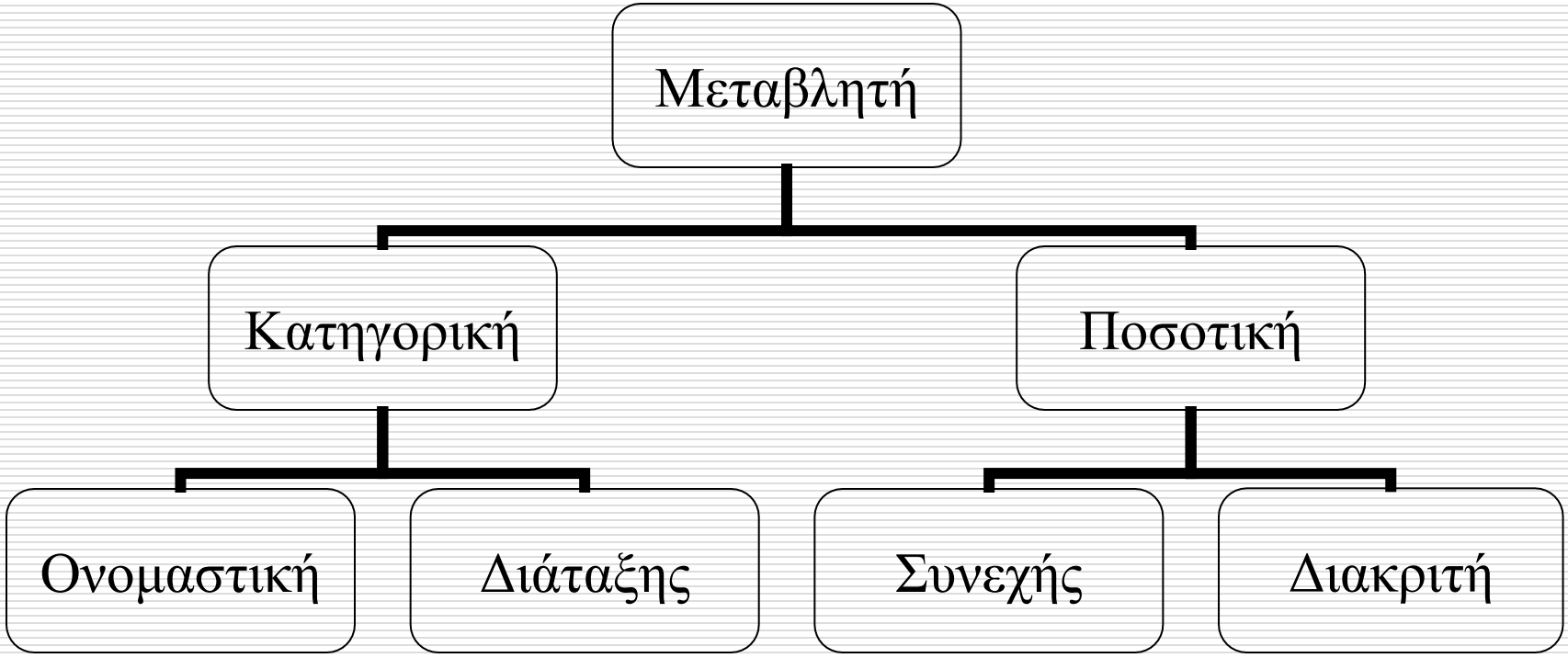
Επαγωγική Στατιστική - Πρόβλεψη

- Κύριος στόχος μιας στατιστικής μελέτης είναι να διερευνήσουμε ένα φαινόμενο με βάση τα δεδομένα του δείγματος (παρατηρήσεις), και από το δείγμα να εξάγουμε συμπεράσματα για τον υπό μελέτη πληθυσμό. Η διερεύνηση αυτή καλείται *επαγωγική στατιστική* ή *στατιστική συμπερασματολογία*. Συνήθως ενδιαφερόμαστε να εκτιμήσουμε ένα άγνωστο μέγεθος που συνοψίζει κατά κάποιον τρόπο τις τιμές της μεταβλητής στον πληθυσμό, π.χ. τον μέσο χρόνο διαδρομής μέχρι την δουλειά. Τέτοια μεγέθη καλούνται *παράμετροι*. Η εκτίμηση τέτοιων παραμέτρων γίνεται με την βοήθεια *εκτιμητριών* οι οποίες είναι κατάλληλα επιλεγμένες συναρτήσεις των παρατηρήσεων που έχουμε, των τιμών δηλαδή του δείγματος. Οι συναρτήσεις αυτές καλούνται *δειγματοσυναρτήσεις* ή *στατιστικές συναρτήσεις*.
- Επίσης αρκετές φορές σε μια στατιστική μελέτη έχουμε το πρόβλημα της *πρόβλεψης* μιας μεταβλητής (*μεταβλητή απόκρισης*) όταν γνωρίζουμε τις τιμές κάποιας ή κάποιων άλλων μεταβλητών (*επεξηγηματικές μεταβλητές*). Ως παράδειγμα μπορεί να ενδιαφερόμαστε για τον βαθμό επίδρασης της επεξηγηματικής μεταβλητής Χ (αριθμός κόκκινων φαναριών που συναντώ στην διαδρομή μου) στην μεταβλητή απόκρισης Υ (χρόνος διαδρομής).

Είδη Μεταβλητών

- **Μεταβλητές:** Ανάλογα με τις τιμές που μια μεταβλητή μπορεί να πάρει μπορεί να ταξινομηθεί ως *κατηγορική* ή ως *ποσοτική*.
  - Ονομάζεται *κατηγορική* μια μεταβλητή η οποία, με κατάλληλη κωδικοποίηση, εκφράζει καταστάσεις, π.χ. το επάγγελμα. Μια κατηγορική μεταβλητή μπορεί να είναι *ονομαστική*, όπου οι κατηγορίες δεν μπορούν να συγκριθούν ή να διαβαθμιστούν (π.χ. χρώμα ματιών) ή *διάταξης* όπου υπάρχει σαφής διαβάθμιση (π.χ. μέτρια, καλή και άριστη φυσική κατάσταση ενός ατόμου).
  - Ονομάζεται *ποσοτική* μια μεταβλητή η οποία εκφράζει ποσότητα, π.χ. βάρος ατόμου. Μια ποσοτική μεταβλητή μπορεί να είναι *διακριτή* όπου το σύνολο τιμών της είναι υποσύνολο των φυσικών αριθμών (π.χ. αριθμός κόκκινων φαναριών που συναντάμε στην διαδρομή μας) ή συνεχής όπου το σύνολο των τιμών της είναι ένα συνεχές διάστημα (π.χ. διάρκεια διαδρομής).

Είδη Μεταβλητών



Πιθανότητες

- **Πιθανότητες:** Η Θεωρία Πιθανοτήτων αποτελεί το *Μαθηματικό Εργαλείο* της Στατιστικής. Είναι η *μαθηματική γλώσσα* που ο κόσμος χρησιμοποιεί για να ποσοτικοποιήσει την αβεβαιότητά του για το αποτέλεσμα ενός τυχαίου πειράματος. Αν π.χ. τα φανάρια που συναντώ στην διαδρομή μου για την δουλειά είναι 3, τότε η πιθανότητα να είναι και τα τρία κόκκινα είναι 1/8. Οι μεταβλητές όπως τις ορίσαμε πριν είναι *τυχαίες μεταβλητές (τ.μ.)*, οι οποίες προέρχονται από μια *κατανομή*. Αν γνωρίζουμε την κατανομή τους, τότε η στατιστική μας μελέτη εστιάζεται στην εκτίμηση διαφόρων ποσοτήτων αυτής της κατανομής (παράμετροι) και καλείται *παραμετρική*. Στην αντίθετη περίπτωση η στατιστική μελέτη καλείται *απαραμετρική ή μη-παραμετρική*.

Πιθανότητες / Στατιστική

- **Πιθανότητες / Στατιστική:** Υπάρχει μια ουσιαστική διαφορά μελετώντας προβλήματα πιθανοτήτων και στατιστικής. Η χρήση των πιθανοτήτων αφορούν εφαρμογές *παραγωγικών συλλογισμών*. Στα προβλήματα πιθανοτήτων γνωρίζουμε τις παραμέτρους των κατανομών και μελετάμε την συμπεριφορά των τ.μ., π.χ. αν ένα νόμισμα είναι δίκαιο (δηλαδή η πιθανότητα να φέρουμε κεφαλή ή γράμματα είναι 1/2) ποια είναι η πιθανότητα να φέρουμε 5 κεφαλές μετά από 10 ρίψεις του νομίσματος; Αντίθετα στα προβλήματα στατιστικής χρησιμοποιούμε *επαγωγικές διαδικασίες*, μαθαίνουμε δηλαδή με βάση την υπάρχουσα εμπειρία. Π.χ. αν σε 10 ρίψεις ενός νομίσματος ήρθαν 5 κεφαλές εκτιμήστε την πιθανότητα να φέρουμε κεφαλή;

Πιθανότητες / Στατιστική

Μπορούμε δηλαδή να πούμε ότι στις πιθανότητες γνωρίζουμε τι συμβαίνει στο σύνολο (πληθυσμός) και βγάζουμε συμπεράσματα για ένα τμήμα αυτού του συνόλου (δείγμα), ενώ στη στατιστική με βάση την γνώση που αποκτούμε από ένα τμήμα (δείγμα) βγάζουμε συμπεράσματα για το σύνολο (πληθυσμός).

Στάδια Στατιστικών Μελετών

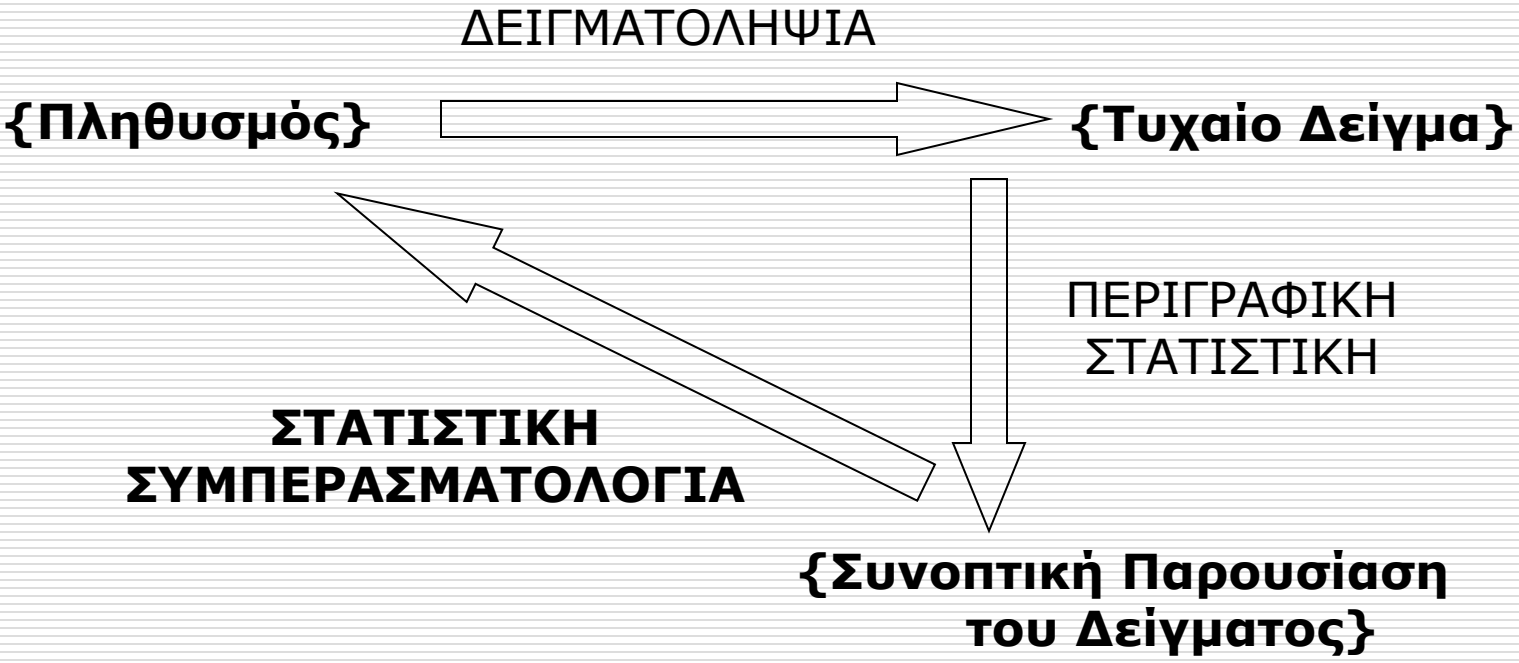
- **Στάδια Στατιστικών Μελετών:**
  1. *Ερευνητικό Ερώτημα.* Πληροφορίες και σχετική βιβλιογραφία. Εδώ θέτουμε το στατιστικό πρόβλημα που έχουμε και τον σκοπό την στατιστικής ανάλυσης που θα επακολουθήσει.
  2. *Δειγματοληψία*. Με βάση το παραπάνω ερευνητικό ερώτημα, προσδιορίζουμε τον πληθυσμό και τις μεταβλητές που μας ενδιαφέρουν και συλλέγουμε "κατάλληλο" δείγμα.
  3. *Κωδικοποίηση και μεταφορά δεδομένων σε Η/Υ.* Για την ευκολότερη στατιστική ανάλυση ίσως χρειασθεί να γίνει κωδικοποίηση μεταβλητών, δηλαδή αντιστοίχηση κωδικών (συνήθως αριθμών) σε κατηγορικές μεταβλητές. Επίσης τα δεδομένα συνήθως μεταφέρονται σε κάποιο στατιστικό πακέτο, με την βοήθεια του οποίου θα γίνει η στατιστική μελέτη, ενώ είναι σημαντικό να γίνει έλεγχος της λογικότητας των τιμών και χειρισμός τυχών ελλιπών τιμών.

Στάδια Στατιστικών Μελετών

4. *Περιγραφική Στατιστική*. Συνοπτική παρουσίαση των δεδομένων που προήλθαν από το δείγμα συνήθως με αριθμητικούς δείκτες και γραφήματα, έτσι ώστε να μπορούν να εξαχθούν διάφορα συμπεράσματα σχετικά με το δείγμα.
5. *Στατιστικό Μοντέλο*. Χρησιμοποιώντας κοινή λογική, προηγούμενες αντίστοιχες μελέτες και τα αποτελέσματα από την περιγραφική στατιστική διατυπώνουμε ένα λογικό στατιστικό μοντέλο για τα δεδομένα. Το στατιστικό μοντέλο αφορά π.χ. την επιλογή της κατανομής της υπό μελέτης μεταβλητής του πληθυσμού, ή τον τρόπο (π.χ. γραμμικά) σύνδεσης των επεξηγηματικών μεταβλητών με την μεταβλητή απόκρισης σε προβλήματα πρόβλεψης. Συνήθως προσαρμόζουμε το μοντέλο στα δεδομένα και προβαίνουμε σε ελέγχους καταλληλότητας του.
6. *Στατιστική Συμπερασματολογία*. Με την βοήθεια του τυχαίου δείγματος και του επιλεγμένου μοντέλου εκτιμούμε τις παραμέτρους του πληθυσμού που μας ενδιαφέρουν.
7. *Παρουσίαση – Ερμηνεία Αποτελεσμάτων.*

Στάδια Στατιστικών Μελετών

- Κύρια Στάδια Στατιστικής Μελέτης



Μεθοδολογία Στατιστικής Συμπερασματολογίας

- **Βασική μεθοδολογία της στατιστικής συμπερασμοτολογίας:** Έστω ότι ενδιαφερόμαστε για το χαρακτηριστικό Υ του πληθυσμού Ω, π.χ. έστω Υ η διάρκεια της διαδρομής μέχρι την δουλειά μου. Όπως αναφέραμε το εν λόγω χαρακτηριστικό είναι μια τυχαία μεταβλητή και έστω f η σ.π.π. του. Έστω ότι ενδιαφερόμαστε να εκτιμήσουμε τη μέση τιμή της διαδρομής, την άγνωστη παράμετρο δηλαδή θ = Ε[Υ]. Για το λόγο αυτόν συλλέγουμε ένα τυχαίο δείγμα μεγέθους n, και έστω $y_1,\ldots,y_n$ οι παρατηρήσεις, και ας υποθέσουμε ότι χρησιμοποιούμε την στατιστική συνάρτηση (δειγματικός μέσος)

$$\bar{y} = n^{-1}\sum_1^n y_i$$ για την εκτίμηση του θ.

Μεθοδολογία Στατιστικής Συμπερασματολογίας

Η τιμή της παραπάνω στατιστικής συνάρτησης αποτελεί και την εκτίμησή μας για το θ. Αλλά τι *σφάλμα* έχουμε σ'αυτή την εκτίμηση. Αν είχαμε πάρει άλλο τυχαίο δείγμα τον ίδιο δειγματικό μέσο θα είχαμε παρατηρήσει; Η *δειγματοληπτική κατανομή* αναφέρεται στην κατανομή της στατιστικής συνάρτησης που προκύπτει από απείρως επαναλαμβανόμενες δειγματοληψίες. Επειδή η τιμή $\bar{y}$ αλλάζει από δείγμα σε δείγμα θεωρούμε ότι απλά έχουμε παρατηρήσει μια τιμή από τις πολλές που μπορεί να πάρει η τ.μ. $\bar{Y}$.

Η επαγωγική στατιστική στηρίζεται στην τυχαιότητα του δείγματος και στην κατανομή αυτού. Το τυχαίο δείγμα δεν είναι τίποτα άλλο από μια συλλογή *ανεξάρτητων και ισόνομων* τυχαίων μεταβλητών $Y_1,\ldots,Y_n$ όπου κάθε μία ξεχωριστά ακολουθεί την κατανομή f του χαρακτηριστικού Υ. Διαλέγοντας ένα συγκεκριμένο τυχαίο δείγμα απλά παρατηρούμε τις τιμές $y_1,\ldots,y_n$ που έτυχε να λάβουν οι εν λόγω τυχαίες μεταβλητές στο συγκεκριμένο δείγμα.

Μεθοδολογία Στατιστικής Συμπερασματολογίας

Όμοια κάθε στατιστική συνάρτηση που χρησιμοποιούμε για να εκτιμήσουμε μια άγνωστη παράμετρο θ του πληθυσμού είναι μια τυχαία μεταβλητή, ως συνάρτηση τυχαίων μεταβλητών, και εμείς μόνο παρατηρούμε μια μόνο τιμή της που προέκυψε από το συγκεκριμένο δείγμα που συλλέξαμε.

Άρα χρησιμοποιούμε μια μεταβλητή ποσότητα (**στατιστική συνάρτηση**), της οποίας έχουμε παρατηρήσει την τιμή που έλαβε στο τυχαίο δείγμα που διαθέτουμε, για να εκτιμήσουμε μια άγνωστη αλλά σταθερή ποσότητα (**παράμετρος**) του πληθυσμού.

Εφαρμογές Στατιστικής

- Ιατρική
- Οικονομετρία
- Μηχανική
- Διοίκηση Επιχειρήσεων
- Αθλητισμός
- Κοινωνικές Επιστήμες
- Βιολογία