ΚΕΦΑΛΑΙΟ 1

1. ΕΤΥΜΗΓΟΡΙΑ ΚΑΙ ΡΙΖΕΣ ΤΗΣ ΣΤΑΤΙΣΤΙΚΗΣ

Η λέξη ***στατιστική*** προέρχεται από το αρχαίο ελληνικό ρήμα ***στατίζω*** που σημαίνει τοποθετώ, ταξινομώ. Η αντίστοιχη λατινική λέξη είναι ***status*** που σημαίνει κράτος, πολιτεία.

Η στατιστική έχει τις ρίζες της στις απογραφές των πληθυσμών. Υπάρχουν ενδείξεις απογραφής στην αρχαία Αθήνα την εποχή του βασιλιά Κέκροπα (περί το 3000 π.Χ.). Η πρώτη καταγεγραμμένη απογραφή πληθυσμού έγινε στην Κίνα το 2238 π.Χ., και μια δεύτερη στην αρχαία Ρώμη από τον ιδρυτή της Ρωμύλο (περί το 753 π.Χ). Το 1935 ο Αμερικανός δημοσιογράφος G. Gallup ίδρυσε το ομώνυμο ινστιτούτο σφυγμομέτρησης της κοινής γνώμης. Οι ασχολούμενοι με τη συλλογή και ανάλυση δεδομένων, για τις κρατικές ανάγκες, ονομάζονταν "κρατικοί" ή "στατιστικοί". Οι πληροφορίες αυτές αφορούσαν γεννήσεις - θανάτους, περιουσία - εισοδήματα, και στρατιωτικές δυνάμεις. Το πρώτο βιβλίο στατιστικού περιεχομένου γράφτηκε το 1583 από τον Ιταλό λόγιο Francesco Sansovino.

2. ΟΡΙΣΜΟΣ ΚΑΙ ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ ΤΗΣ ΣΤΑΤΙΣΤΙΚΗΣ. ΚΛΙΜΑΚΕΣ ΜΕΤΡΗΣΗΣ ΤΩΝ ΜΕΤΑΒΛΗΤΩΝ

Στατιστική είναι η επιστήμη που ασχολείται με την ανάπτυξη μεθόδων: ***συλλογής (collecting data), ταξινόμησης (classification), επεξεργασίας (processing), παρουσίασης (presentation), και ανάλυσης (analysis)*** πάνω σε οποιαδήποτε δραστηριότητα του ανθρώπου ή φαινομένου της φύσης με σκοπό την εξαγωγή συμπερασμάτων που χρησιμεύουν στη λήψη αποφάσεων. Ανάλογα με το πεδίο εφαρμογής της η στατιστική διακρίνεται σε κοινωνική, εκπαιδευτική, περιβαλλοντική, βιοστατιστική, χρηματοοικονομική, στατιστική της αγοράς και των επιχειρήσεων, δημογραφική κ.λπ.

Οι πρωταρχικές έννοιες της στατιστικής είναι οι ακόλουθες:

1. Δεδομένα (Data): Είναι τα στοιχεία τα οποία συλλέγονται, αναλύονται και συνοψίζονται για παρουσίαση και ερμηνεία. Αυτά μπορεί να περιλαμβάνουν γράμματα, αριθμούς, και σύμβολα. Το σύνολο των μετρήσεων που συγκεντρώθηκαν

για κάθε χαρακτηριστικό που μας ενδιαφέρει ονομάζονται ***παρατηρήσεις (observations)***. Κατά τη συλλογή των δεδομένων υπάρχει πιθανότητα να πάρουμε λανθασμένα ή ελλιπή στοιχεία. Είναι προτιμότερο να χρησιμοποιούμε, για σωστότερα αποτελέσματα, μόνο αυτά τα στοιχεία για τα οποία έχουμε σιγουριά για την ορθότητά τους.

2. Άτομο-Στοιχείο (Individual-Element): Είναι η μονάδα βάσης την οποία παρατηρεί ο αναλυτής. Π.χ. ένα άτομο, ένας αριθμός, ένα αντικείμενο, μια ιδιότητα, μια γνώμη.

3. Πληθυσμός (Population): Είναι το σύνολο των εξεταζόμενων ατόμων-στοιχείων σε μια έρευνα. Ο πληθυσμός π.χ. μπορεί να είναι το σύνολο των κατοίκων μιας χώρας, το σύνολο των ακινήτων μιας πόλης, το σύνολο των μισθών μιας επιχείρησης κ.λπ. Ο (στατιστικός) πληθυσμός μπορεί να είναι ***άπειρος (infinite)*** δηλ. να περιλαμβάνει άπειρο πλήθος ατόμων-στοιχείων ή ***πεπερασμένος (finite)***. Συνήθως πεπερασμένοι πληθυσμοί με πολύ μεγάλο πλήθος στοιχείων (στην κλίμακα του δισεκατομμυρίου και άνω) θεωρούνται άπειροι.

4. Δείγμα (Sample): Είναι κάθε γνήσιο υποσύνολο του πληθυσμού. Π.χ. το σύνολο των κατοίκων μιας συγκεκριμένης πόλης, το σύνολο των ακινήτων ενός συγκεκριμένου προαστίου ή το σύνολο των μισθών ενός συγκεκριμένου τμήματος μιας επιχείρησης αποτελούν δείγματα των πληθυσμών που αναφέρθηκαν πιο πάνω.

Με άλλα λόγια από τον πληθυσμό επιλέγουμε ένα δείγμα (αποτελούμενο από άτομα-στοιχεία) από το οποίο συλλέγουμε τα δεδομένα προς ταξινόμηση, επεξεργασία, ανάλυση, και παρουσίαση.

5. Πείραμα (Experiment): Είναι η σχεδιασμένη ενέργεια η οποία έχει ως αποτέλεσμα τη δημιουργία σετ δεδομένων. Π.χ. μια έρευνα πρόθεσης αγοράς ενός προϊόντος.

6. Παράμετρος (Parameter): Είναι μια αριθμητική τιμή που δίνει την εικόνα όλων των δεδομένων ενός πληθυσμού ως προς κάποιο χαρακτηριστικό γνώρισμα. Π.χ. το ποσοστό που πήρε ένα κόμμα στις εκλογές. Όταν διενεργούμε σφυγμομέτρηση (γκάλοπ), τότε επιλέγουμε ένα δείγμα του πληθυσμού και κάνουμε ***εκτίμηση*** της ***πληθυσμιακής παραμέτρου*** μέσω της ***δειγματικής παραμέτρου***.

7. Μεταβλητή (Variable) ή Τυχαία Μεταβλητή (Random Variable): Είναι το χαρακτηριστικό γνώρισμα ως προς το οποίο θα μελετηθεί ο πληθυσμός ή το

δείγμα. Οι μεταβλητές διακρίνονται σε:

7.1. Ποσοτικές (Quantitative): Αυτές που επιδέχονται μέτρηση και οι τιμές τους είναι πραγματικοί αριθμοί. Συνοδεύονται από τη μονάδα μέτρησης κάθε φορά (αν υπάρχει). Παραδείγματα αποτελούν το βάρος ενός μαθητή, το ύψος των παικτών μιας ομάδας, η θερμοκρασία, η υγρασία, ο αριθμός των κοριτσιών μιας οικογένειας, η βαθμολογία σ'ένα τεστ κ.λπ.

7.2. Ποιοτικές (Qualitative): Αυτές που δεν επιδέχονται μέτρηση και οι τιμές τους εκφράζονται με αριθμούς ή λέξεις. Παραδείγματα αποτελούν η κατάταξη σ' ένα αγώνισμα, η ένταση της γνώμης, το επάγγελμα, η οικογενειακή κατάσταση, το χρώμα των ματιών, η ένδειξη ενός νομίσματος, το φύλο κ.λπ. Ανάλογα με τη φύση των χαρακτηριστικών που θέλουμε να μετρήσουμε χρησιμοποιούμε και διαφορετική ***κλίμακα μέτρησης (measurement scale)*** των μεταβλητών με αποτέλεσμα οι δύο βασικές κατηγορίες μεταβλητών να διαχωρίζονται περαιτέρω. Έτσι έχουμε τις εξής:

i) ***Κλίμακα τάξης ή ιεράρχησης (Ordinal scale)***: Όταν μια ποιοτική μεταβλητή παίρνει τιμές που μπορούν να ιεραρχηθούν. Κλασικά παραδείγματα είναι η σειρά κατάταξης σ'ένα αγώνισμα (1ος-2ος-3ος), η κατάσταση της υγείας (1 = κακή, 2 = μέτρια,3 = καλή), η απόδοση ενός υπαλλήλου στην εργασία του κ.λπ. Όπως βλέπουμε μπορούμε να αντιστοιχήσουμε αριθμητικές τιμές στις απαντήσεις. Οι τιμές αυτές δηλώνουν σύγκριση/ιεράρχηση/κορύφωση και δε μπορούν ούτε να αφαιρεθούν ούτε να διαιρεθούν μεταξύ τους.

ii) ***Ονομαστική κλίμακα (Nominal scale)***: Όταν μια ποιοτική μεταβλητή παίρνει τιμές που δε μπορούν να ιεραρχηθούν. Και εδώ μπορούν να χρησιμοποιηθούν αριθμητικές τιμές προκειμένου να κατατάξουν/κωδικοποιήσουν τις απαντήσεις σε κατηγορίες. Π.χ. το φύλο των ατόμων (1 = άνδρας, 2 = γυναίκα), φορέας μιας ασθένειας (1 = ναι, 2 = όχι) (αυτές οι μεταβλητές που επιδέχονται μόνο δύο απαντήσεις αμοιβαία αποκλειόμενες ονομάζονται ***διχοτομικές (dichotomical)***), οικογενειακή κατάσταση (1 = ανύπαντρος, 2 = παντρεμένος, 3 = διαζευγμένος), το χρώμα των ματιών (1 = μαύρο, 2 = μπλε, 3 = πράσινο), το θρήσκευμα (1 = χριστιανός ορθόδοξος, 2 = καθολικός, 3 = προτεστάντης, κ.λπ.) κ.λπ. Είναι προφανές ότι οι αριθμητικές τιμές, στην περίπτωση αυτή, όχι μόνο δε μπορούν να αφαιρεθούν ή να διαιρεθούν μεταξύ τους αλλά ούτε και να συγκριθούν.

iii) ***Αναλογική κλίμακα ή κλίμακα λόγου (Ratio scale)***: Όταν μια ποσοτική μεταβλητή παίρνει τιμές οι οποίες είναι αναλογικές. Η ονομασία της προέκυψε από το γεγονός ότι οι τιμές των μεταβλητών αυτών μπορούν να διαιρεθούν μεταξύ τους (π.χ. $\frac{5000€}{2000€} = 2,5$ φορές ακριβότερο, $\frac{12\,\text{κιλά}}{4\,\text{κιλά}} = 3$ φορές βαρύτερο κ.λπ.), αλλά και να αφαιρεθούν (π.χ. 12 κιλά -4 κιλά $= 8$ κιλά βαρύτερο κ.λπ.). Αυτό οφείλεται στο γεγονός ότι οι κλίμακες μέτρησης αυτών των χαρακτηριστικών περιλαμβάνουν την αρχή ή σημείο μηδέν ή σημείο έλλειψης του μετρούμενου χαρακτηριστικού (μηδενική αξία, μηδενικό βάρος, μηδενική παραγωγή κ.λπ.). Αφού λοιπόν υπάρχει συγκεκριμένη αρχή, η διαίρεση μεταξύ δύο τιμών έχει νόημα (όπως βέβαια και η αφαίρεση).

iv) ***Κλίμακα διαστήματος (Interval scale)***: Όταν μια ποσοτική μεταβλητή παίρνει τιμές των οποίων η διαίρεση δεν έχει καμία αξία (παρά μόνο η αφαίρεσή τους). Αυτό οφείλεται στο γεγονός ότι η κλίμακα μέτρησης αυτών των χαρακτηριστικών δεν περιλαμβάνει "μηδενικό" σημείο. Κλασικό παράδειγμα είναι η θερμοκρασία. Κανείς δεν αμφισβητεί τις μεταβολές της θερμοκρασίας. Π.χ. χθες είχαμε 20°C, σήμερα ανέβηκε στους 25°C και για αύριο προβλέπεται να φθάσει τους 30°C. Όμως δε μπορούμε να ισχυρισθούμε ότι αύριο ο καιρός θα είναι 50% πιο ζεστός από χθες, διότι δεν υπάρχει αρχή (μηδέν) στη θερμοκρασία (αυτό είναι διαφορετικό από τους 0°C). Έτσι οι θερμοκρασίες αφαιρούνται μεταξύ τους αλλά δε διαιρούνται. Άλλο παράδειγμα είναι οι εξετάσεις που περιλαμβάνουν μεγάλο αριθμό ερωτήσεων και ως εκ τούτου εξετάζουν (προσεγγιστικά) όλο το φάσμα των γνώσεων (π.χ. GMAT, GRE, TOEFL κ.λπ.). Έτσι, τα κριτήρια της βαθμολογίας από 400 σε 500, και από 500 σε 600 μπορεί να είναι περίπου τα ίδια και να αντιπροσωπεύουν τις "ίδιες" αποστάσεις σε γνώσεις. Όμως δε μπορούμε να ισχυρισθούμε ότι ο φοιτητής που βαθμολογήθηκε με 500 έχει 25% περισσότερες γνώσεις από εκείνον που βαθμολογήθηκε με 400, διότι δεν υπάρχει αρχή (μηδέν) στην κλίμακα της γνώσης.

Οι ποσοτικές μεταβλητές διακρίνονται επίσης σε:

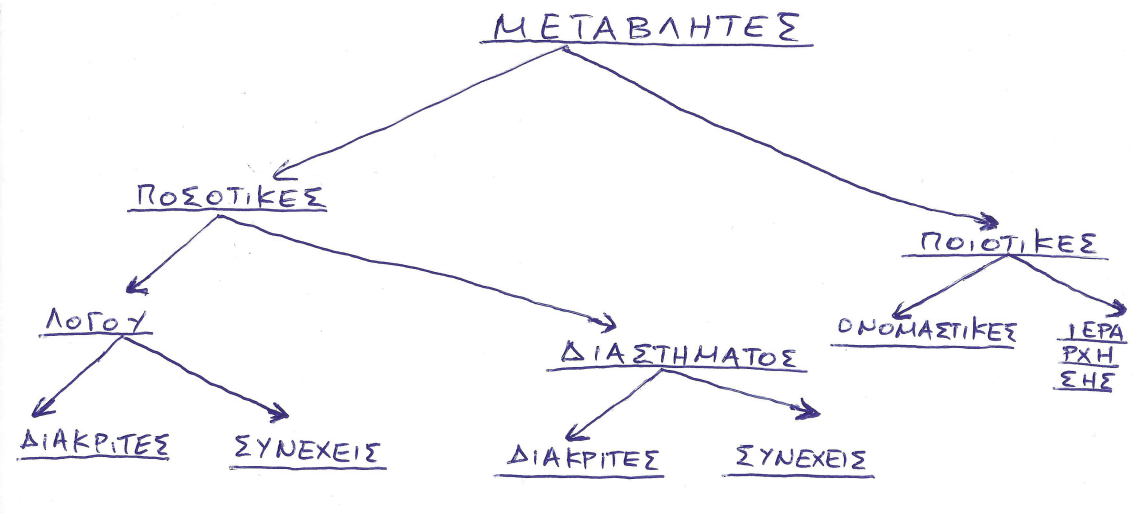
i) ***Συνεχείς (Continuous)***: Όταν παίρνουν τιμές σε όλο το εύρος των τιμών. Π.χ. το βάρος (κιλά, γραμμάρια, δέκατα γραμμαρίων), το μήκος (μέτρα, εκατοστά, χιλιοστά, δέκατα χιλιοστών), ο χρόνος (ώρες, λεπτά, δευτερόλεπτα, δέκατα δευτερολέπτων) κ.λπ.

ii) ***Ασυνεχείς ή Διακριτές (Discrete)***: Όταν παίρνουν μόνο μεμονωμένες

τιμές. Π.χ. αριθμός μελών οικογένειας, αριθμός επισκέψεων στον κινηματογράφο, αριθμός περιοδικών που αγοράζουν τα μέλη μιας οικογένειας, ο αριθμός των μαθημάτων για την απόκτηση πτυχίου κ.λπ.

Έχουμε λοιπόν το ακόλουθο

Σχήμα 1



3. ΒΑΣΙΚΟΙ ΚΛΑΔΟΙ ΤΗΣ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΕΝΔΕΙΚΤΙΚΕΣ ΕΦΑΡΜΟΓΕΣ ΣΤΙΣ ΕΠΙΧΕΙΡΗΣΕΙΣ ΚΑΙ ΤΗΝ ΟΙΚΟΝΟΜΙΑ

Η Στατιστική χωρίζεται σε δύο μεγάλους κλάδους:

1. Περιγραφική Στατιστική (Descriptive Statistics): Με αυτόν τον όρο περιγράφουμε τις μεθόδους που ασχολούνται με τη συλλογή, παρουσίαση και χαρακτηρισμό (ταξινόμηση) των δεδομένων ανάλογα με το είδος των χαρακτηριστικών που περιγράφουν (μετρούν). Οι περισσότερες από τις στατιστικές πληροφορίες τις οποίες βλέπουμε σε εφημερίδες, περιοδικά, τηλεοράσεις ή αλλού, αποτελούνται από δεδομένα τα οποία έχουν συλλεχθεί, επεξεργασθεί και παρουσιάζονται συνήθως με πίνακες και γραφήματα. Στην Περιγραφική Στατιστική διακρίνουμε:

i) ***Μονομεταβλητούς πληθυσμούς (univariate)***: Μελέτη ενός μόνο χαρακτηριστικού

ii) ***Πολυμεταβλητούς πληθυσμούς (multivariate)***: Μελέτη περισσοτέρων του ενός χαρακτηριστικών ταυτόχρονα.

2. Επαγωγική Στατιστική ή Στατιστική Συμπερασματολογία (Inferential Statistics): Έτσι ορίζονται οι μέθοδοι που μας βοηθούν να εκτιμήσουμε

τα χαρακτηριστικά (παραμέτρους) ενός πληθυσμού με βάση τα αποτελέσματα που προκύπτουν από τις παρατηρήσεις ενός δείγματος. Στην Επαγωγική Στατιστική διακρίνουμε δύο ομάδες προβλημάτων:

i) ***Την Στατιστική Εκτίμηση (Statistical Estimation)*** με την οποία, με βάση τα δεδομένα ενός δείγματος, εκτιμούμε τις τιμές του πληθυσμού, που περιλαμβάνει:

α. την Εκτίμηση σημείου (Point estimation): Εκτίμηση της τιμής του πληθυσμού βασισμένη σε μία μόνο τιμή του δείγματος. Αν π.χ. η μέση τιμή της αξίας των μετοχών στο Χ.Α.Α. είναι 30€, με βάση ένα δείγμα 50 μετοχών, τότε λέμε ότι και η μέση τιμή όλων των μετοχών στο Χ.Α.Α. είναι περίπου 30€.

β. την Εκτίμηση διαστήματος (Interval estimation): Εκτίμηση της τιμής του πληθυσμού βασισμένη σε ένα διάστημα το οποίο προέκυψε από το δείγμα και καλείται ***διάστημα εμπιστοσύνης (confidence interval)***. Αν π.χ. η μέση τιμή της αξίας των μετοχών του Χ.Α.Α. είναι 30€, με βάση ένα δείγμα 50 μετοχών, μπορούμε να δημιουργήσουμε ένα διάστημα (εμπιστοσύνης), π.χ. το (28,5, 31,5), μέσα στο οποίο θα βρίσκεται η μέση τιμή των αξιών όλων των μετοχών του Χ.Α.Α. Το διάστημα εμπιστοσύνης συνοδεύεται και από έναν συντελεστή που καλείται ***συντελεστής ή επίπεδο εμπιστοσύνης (confidence coefficient)***, παίρνει τιμές στο διάστημα $[0, 1]$ και δίνει την πιθανότητα επιτυχίας της εκτίμησης. Αν λοιπόν στο προηγούμενο παράδειγμα το επίπεδο εμπιστοσύνης είναι 0,95, τότε η εκτίμηση που κάναμε έχει πιθανότητα επιτυχίας 95%.

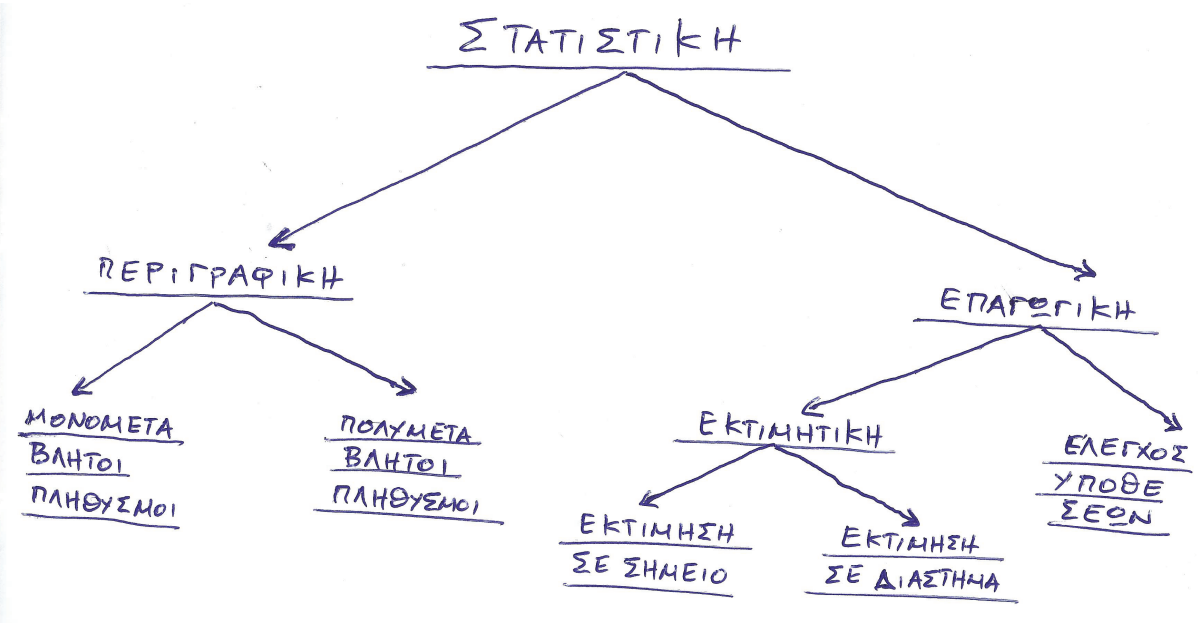
ii) ***τον Έλεγχο (Στατιστικών) υποθέσεων (Testing Hypotheses)*** με τον οποίο ελέγχουμε, με βάση τα δεδομένα ενός δείγματος, την ορθότητα μιας υπόθεσης που έχουμε κάνει για τον πληθυσμό. Στη διαδικασία ελέγχου των υποθέσεων είναι απαραίτητο να κάνουμε δύο βασικές υποθέσεις:

α. τη Μηδενική υπόθεση (Null Hypothesis): Είναι η υπόθεση που κάνουμε για μία συγκεκριμένη τιμή του πληθυσμού και

β. την Εναλλακτική υπόθεση (Alternative Hypothesis): Είναι η αντίθετη της μηδενικής υπόθεσης.

Έχουμε λοιπόν το ακόλουθο

Σχήμα 2



Οι αρχές της στατιστικής συμπερασματολογίας βασίζονται στη ***Θεωρία Πιθανοτήτων***. Επιγραμματικά, και μόνο, αναφέρουμε ότι η εξαγωγή συμπερασμάτων για έναν πληθυσμό με βάση τις παρατηρήσεις ενός δείγματος, προϋποθέτει ότι το δείγμα είναι ***τυχαίο (random sample)*** δηλ. όλα τα μέλη του πληθυσμού έχουν την ίδια πιθανότητα να περιληφθούν σ' αυτό και η επιλογή ενός μέλους στο δείγμα δεν επηρεάζει την πιθανότητα επιλογής οποιουδήποτε άλλου μέλους σ' αυτό.

Μερικές ενδεικτικές εφαρμογές της Στατιστικής στις επιχειρήσεις και στην οικονομία είναι οι ακόλουθες:

i) Η αποτύπωση ποιοτικών και ποσοτικών χαρακτηριστικών της επιχείρησης. Μια επιχείρηση θέλει να έχει κάθε στιγμή διαθέσιμα στοιχεία σχετικά με τον όγκο των παραγόμενων προϊόντων, τις πωλήσεις, τις προμήθειες, τις ηλικίες και το φύλο των εργαζομένων, τις αμοιβές κ.λπ. (απλά περιγραφικά στατιστικά μέτρα).

ii) Η προσπάθεια προσδιορισμού των παραγόντων οι οποίοι επιδρούν στις πωλήσεις ή στην τιμή ενός προϊόντος ή στην απόδοση των εργαζομένων. Θέλουμε να μάθουμε π.χ. αν η απόδοση των υπαλλήλων εξαρτάται από το μισθό, την ηλικία, τη μόρφωση κ.λπ. (απλή ή πολλαπλή παλινδρόμηση - συσχέτιση).

iii) Η διαχρονική εξέλιξη των κερδών ή των πωλήσεων ή της τιμής μιας μετοχής. Θέλουμε να διαπιστώσουμε αν τα κέρδη μιας επιχείρησης την τελευταία δεκαετία

μπορούν να μας δώσουν μια ικανοποιητική εκτίμηση για τα επόμενα χρόνια (χρονοσειρές).

iv) Η μεταβολή στην παραγωγικότητα, την απόδοση, τα κέρδη, την τιμή, την ποσότητα. Θέλουμε να γνωρίζουμε π.χ. την ποσοστιαία μεταβολή στην παραγωγικότητα ή τα κέρδη της επιχείρησής μας (αριθμοδείκτες).

4. ΜΕΘΟΔΟΙ ΣΤΑΤΙΣΤΙΚΗΣ ΕΡΕΥΝΑΣ

Για τη διενέργεια μιας στατιστικής έρευνας υπάρχουν διαθέσιμες οι παρακάτω μέθοδοι:

4.1. ***Απογραφή (Census)***: Είναι η συλλογή στοιχείων από όλα τα άτομα του πληθυσμού και προφανώς μπορεί να χρησιμοποιηθεί όταν ο πληθυσμός είναι πεπερασμένος. Το βασικό πλεονέκτημα της απογραφής είναι η (απόλυτη) εγκυρότητα των αποτελεσμάτων μιας και δεν κάνουμε εκτιμήσεις αλλά προσδιορισμούς. Ωστόσο η απογραφή παρουσιάζει και τα ακόλουθα μειονεκτήματα:
i) Τη μεγάλη χρονική διάρκεια που απαιτείται για τη συλλογή και την επεξεργασία των δεδομένων (ιδίως σε μεγάλους πληθυσμούς).
ii) Τα σφάλματα συλλογής και επεξεργασίας των δεδομένων.
iii) Την αδυναμία διεξαγωγής όταν η παρατήρηση συνεπάγεται την καταστροφή των μονάδων (π.χ. σε ελέγχους ποιότητας).
iv) Το κόστος της παρατήρησης είναι "αρκετά" μεγάλο σε σχέση με τα διαθέσιμα μέσα ή τα αναμενόμενα αποτελέσματα. Τα δεδομένα που συλλέγονται με την απογραφή ονομάζονται ***πρωτογενή (primitive)*** και οι πηγές τους επίσης ***πρωτογενείς***.

4.2. ***Δειγματοληψία (Sampling)***: Δειγματοληψία είναι η συλλογή παρατηρήσεων από ένα ***δείγμα (sample)*** δηλ. ένα μέρος των ατόμων του πληθυσμού. Είναι η συνηθέστερη πηγή πρωτογενών δεδομένων, και οι πιο συνηθισμένες είναι οι πολιτικές έρευνες και οι έρευνες αγοράς. Η δειγματοληπτική μέθοδος την οποία θα ακολουθήσουμε σε μία έρευνα (δηλ. το μέγεθος και ο τρόπος επιλογής του δείγματος) αποτελεί τη βάση της ***αξιοπιστίας (reliability)*** την οποία θα παρουσιάσουν τα τελικά αποτελέσματα. Η σύγχρονη δειγματοληψία μας επιτρέπει τη συλλογή πληροφοριών με:

i) Μεγαλύτερη ταχύτητα

ii) Μικρότερο κόστος

iii) Μεγάλη ακρίβεια εκτίμησης

iv) Μεγάλο εύρος εφαρμογών.

Βασικοί κανόνες επιλογής του δείγματος είναι:

i) Αντιπροσωπευτικότητα

ii) Αξιοπιστία

iii) Αντικειμενικότητα

iv) Συγκρισιμότητα.

δηλ. να αποτελεί μία κατά το δυνατό μικρογραφία του πληθυσμού.

Το σημαντικότερο μειονέκτημα της δειγματοληψίας είναι τα δημιουργούμενα ***δειγματοληπτικά σφάλματα (sampling errors)*** τα οποία είναι αριθμητικά σφάλματα που οφείλονται στις τυχαίες κυμάνσεις της δειγματοληψίας. Τα σφάλματα αυτά τείνουν να μηδενιστούν όσο το μέγεθος του δείγματος μεγαλώνει. Με τον όρο δειγματοληπτικό σφάλμα εννοούμε τη διαφορά μεταξύ μιας στατιστικής παραμέτρου που προκύπτει από ένα δείγμα και της αντίστοιχης παραμέτρου που προκύπτει από απογραφή. Έστω π.χ. ότι το μέσο ύψος 3.000 ατόμων, που προέκυψε από απογραφή, είναι 185 cm. Αν από τα 3.000 άτομα επιλέξουμε δείγμα 100 ατόμων και προκύψει μέσο ύψος 180 cm, τότε η διαφορά των 5 cm είναι ένα δειγματοληπτικό σφάλμα. Ένας άλλος παράγοντας σφαλμάτων είναι η διαφορά στους ***ορισμούς (definitions)***. Ορισμός είναι ο καθορισμός της στατιστικής μονάδας. Αν π.χ. ένας ερευνητής χρησιμοποιεί ως στατιστική μονάδα το άτομο και άλλος την οικογένεια, τότε τα αποτελέσματα διαφέρουν και δε μπορούν να ερμηνευθούν με τον ίδιο τρόπο. Τέτοια σφάλματα ονομάζονται ***μη δειγματοληπτικά (non sampling errors)***.

Υπάρχουν διάφορες μέθοδοι επιλογής δείγματος. Το πόσο καλά ένα δείγμα αντιπροσωπεύει τον πληθυσμό εξαρτάται από

i) ***Το δειγματοληπτικό πλαίσιο (sample frame)***

ii) ***Το μέγεθός του (sample size)***

iii) ***Τον σχεδιασμό της διαδικασίας συλλογής (selection procedure)***.

Οι σημαντικότερες τεχνικές δειγματοληψίας είναι οι ακόλουθες:

i) Απλή τυχαία δειγματοληψία (Simple random sampling): Αποτελεί την πιό απλή μορφή επιλογής ενός ***τυχαίου δείγματος (simple random sample)***. Έστω π.χ. ότι θέλουμε να επιλέξουμε ένα δείγμα 200 δικηγόρων από τον

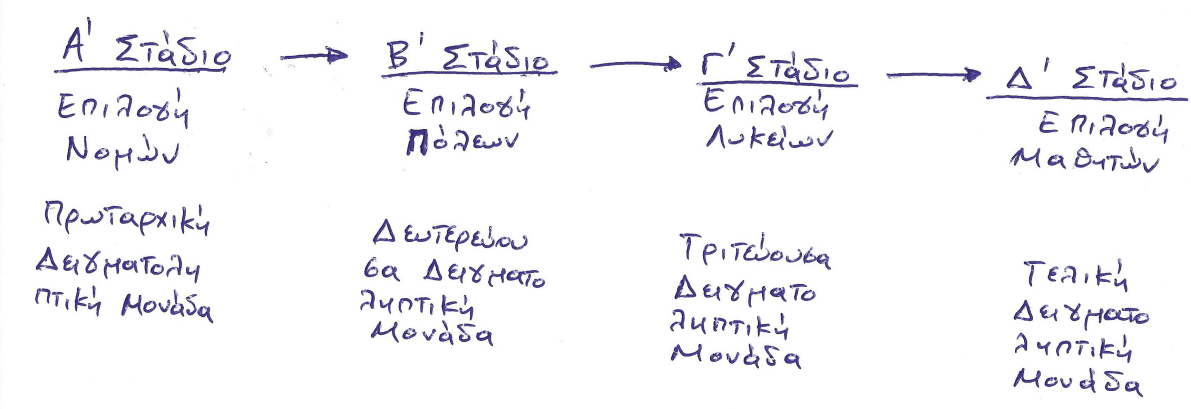
Δικηγορικό Σύλλογο Αθηνών, που αριθμεί 20.000 μέλη, με αντικείμενο τις απόψεις τους για τη νέα κωδικοποίηση του Αστικού Κώδικα. Αριθμούμε τους δικηγόρους από το 1 έως το 20.000 και επιλέγουμε 200 τυχαίους αριθμούς (με την κλασική κληρωτίδα ή με έναν σύγχρονο ηλεκτρονικό τρόπο). Εναλλακτικά, μπορούμε να χρησιμοποιήσουμε τους Αριθμούς Μητρώου των δικηγόρων, αρκεί να είναι μοναδικοί και συνεχόμενοι. Ο κατάλογος (ή μητρώο) των δικηγόρων της Αθήνας είναι το ***δειγματοληπτικό πλαίσιο***, ο κάθε δικηγόρος που είναι εγγεγραμμένος στο μητρώο και αποτελεί αντικείμενο τυχαίας επιλογής είναι η ***δειγματοληπτική μονάδα (sample unit)***, το ***μέγεθος του δείγματος*** είναι 200, και η κλήρωση είναι η ***διαδικασία συλλογής***.

ii) Τυχαία συστηματική δειγματοληψία (Systematic random sampling): Αποτελεί έναν εναλλακτικό τρόπο επιλογής ενός τυχαίου δείγματος, που είναι πιο εύκολος από την απλή τυχαία δειγματοληψία και χρησιμοποιείται ευρέως. Με βάση το προηγούμενο παράδειγμα, η επιλογή 200 δικηγόρων από το σύνολο των 20.000 σημαίνει ότι $\frac{200}{20.000} = \frac{1}{100}$ που ονομάζεται ***δειγματοληπτικό κλάσμα (sampling fraction)***. Δηλ. ανά 100 δικηγόρους επιλέγεται ένας. Έτσι, επιλέγεται ένας τυχαίος αριθμός μεταξύ 1 και 100, π.χ. το 54, και το δείγμα αποτελείται από τους εξής δικηγόρους με βάση την αρίθμησή τους: 54, 154, 254, ..., 19.854, 19.954. Με άλλα λόγια αρχίζουμε με την καταχώρηση 54 και προχωρούμε προσθέτοντας κάθε φορά τον αριθμό 100.

iii) Στρωματοποιημένη τυχαία δειγματοληψία (Stratified random sampling): Σκοπός κάθε έρευνας είναι να καλύψει όσο το δυντόν πιο αντιπροσωπευτικά τον υπό εξέταση πληθυσμό. Η απλή τυχαία δειγματοληψία δε μπορεί να εξασφαλίσει πάντα την αντιπροσωπευτικότητα. Π.χ. μπορεί να επιλεγούν μόνο νέοι δικηγόροι ή μόνο οι μεγαλύτεροι στην ηλικία. Για να εξασφαλίσουμε τη συμμετοχή όλων των κατηγοριών, τους ταξινομούμε πρώτα σε κατηγορίες ή ***στρώματα (strata)*** και στη συνέχεια επιλέγουμε δείγμα από κάθε στρώμα (κατηγορία ή ομάδα). Π.χ. επιλέγουμε ένα τυχαίο δείγμα δικηγόρων από κάθε μία από τις τρεις κατηγορίες δικηγόρων ανάλογα με το βαθμό του δικαστηρίου που μπορούν να παραστούν: Πρωτοδικείο, Εφετείο, Άρειος Πάγος. Αν κάθε κατηγορία αντιπροσωπεύεται στο δείγμα με το ίδιο ποσοστό που συμμετέχει στον πληθυσμό, τότε η δειγματοληψία ονομάζεται ***Αναλογική στρωματοποιημένη δειγματοληψία (Proportional stratified sampling)***.

iv) Δειγματοληψία σε πολλά στάδια (Multistage sampling): Απαραίτητη προϋπόθεση για την επιλογή ενός τυχαίου δείγματος είναι η καταγραφή όλων των μελών (μονάδων) του πληθυσμού που εξετάζουμε σ' έναν ενιαίο κατάλογο δηλ. το δειγματοληπτικό πλαίσιο. Αυτό όμως δεν είναι πάντα εφικτό. Π.χ. κάθε Λύκειο έχει κατάλογο των μαθητών του, όμως δεν υπάρχει ενιαίος κατάλογος όλων των μαθητών Λυκείων της χώρας. Πώς θα επιλέξουμε ένα δείγμα μαθητών απ' όλα τα γεωγραφικά διαμερίσματα, προκειμένου να καταγράψουμε τις απόψεις τους για το νέο σύστημα εισαγωγής στην τριτοβάθμια εκπαίδευση; Πώς θα επιλεγεί π.χ. ο συγκεκριμένος μαθητής της Α′ Λυκείου της πόλης του Ναυπλίου; Θα χρησιμοποιήσουμε ενδιάμεσα ***στάδια*** για να φθάσουμε στον συγκεκριμένο μαθητή που αποτελεί και την ***τελική δειγματοληπτική μονάδα***. Δηλ. κάθε μαθητής φοιτά σε συγκεκριμένο Λύκειο, κάθε Λύκειο βρίσκεται σε συγκεκριμένη πόλη, και κάθε πόλη ανήκει σε έναν Νομό. Έτσι, σε ***πρώτο στάδιο*** επιλέγουμε ένα τυχαίο δείγμα Νομών, στη συνέχεια σε ***δεύτερο στάδιο*** επιλέγουμε από κάθε Νομό τυχαίο δείγμα πόλεων, στο ***τρίτο στάδιο*** επιλέγουμε από κάθε πόλη τυχαίο δείγμα Λυκείων, και στο ***τέταρτο στάδιο*** (και τελευταίο) επιλέγουμε από κάθε Λύκειο τυχαίο δείγμα μαθητών. Δηλαδή:

Σχήμα 3



Αν ως βασικός κατάλογος (δειγματοληπτικό πλαίσιο) χρησιμοποιείται ο χάρτης, τότε η δειγματοληψία ονομάζεται ***επιφανειακή δειγματοληψία (area sampling)***.

v) Δειγματοληψία ποσοστών (quota sampling): Η μέθοδος αυτή δεν είναι τυχαία και βασίζεται στην υποκειμενική κρίση του ερευνητή. Σκοπός της είναι να συμπεριλάβει στο δείγμα όλες τις κατηγορίες των μελών του πληθυσμού με βάση

διάφορα χαρακτηριστικά, που σε πολλές περιπτώσεις δεν είναι καταγεγραμμένα στον κατάλογο, έτσι ώστε να χρησιμοποιηθεί μετά η τυχαία στρωματοποιημένη δειγματοληψία. Π.χ. στο δείγμα των δικηγόρων θέλουμε ν' αντιπροσωπευθούν όλες οι κατηγορίες (τάσεις) των επαγγελματιών δηλ. φύλο, σπουδές στην Ελλάδα ή στο εξωτερικό, πρώτο πτυχίο ή και μεταπτυχιακές σπουδές, κύρια ειδικότητα (αστικές, ποινικές, οικογενειακές, ναυτιλιακές υποθέσεις), εργασιακό περιβάλλον (ατομικό γραφείο, συνεργασία με άλλους συναδέλφους, νομική εταιρία, νομικός σύμβουλος), χρόνια στο επάγγελμα κ.λπ. Με άλλα λόγια οι ανθρώπινοι πληθυσμοί διακρίνονται με βάση πολλά χαρακτηριστικά που είναι αδύνατο να είναι όλα καταχωρημένα. Εκεί, πρωτεύοντα ρόλο παίζει η εμπειρία και η υποκειμενική κρίση του ερευνητή. Γνωρίζοντας περίπου τα ποσοστά σύνθεσης του πληθυσμού (ποσοστά ανδρών - γυναικών, ποσοστό δικηγόρων με μεταπτυχιακό δίπλωμα κ.λπ.) συνθέτει μόνος του το δείγμα, φροντίζοντας να καλύπτει όλες τις κατηγορίες των δικηγόρων σύμφωνα με τα ***ποσοστά (quotas)*** που συμμετέχουν στον πληθυσμό. Έτσι, δεν βασίζεται στην τύχη, αλλά προσεγγίζει πρώτα τους δικηγόρους (πχ. στους χώρους των δικαστηρίων), τους ρωτά σε ποιά κατηγορία ανήκουν, και στη συνέχεια προχωρεί στη συνέντευξη. Με άλλα λόγια, η έρευνα γίνεται με βάση την "ποσόστωση" που έχει προαποφασίσει ο ερευνητής. Π.χ. το δείγμα θα αποτελείται από 70% άνδρες - 30% γυναίκες, 40% μέχρι 10 χρόνια στο επάγγελμα - 30% από 10 έως 20 χρόνια - 30% με πάνω από 20 χρόνια, 90% απόφοιτοι ελληνικών πανεπιστημίων - 10% απόφοιτοι του εξωτερικού, 40% να ασχολούνται κυρίως με αστικές υποθέσεις - 30% με ποινικά - 20% με οικογενειακά - 10% με ναυτιλιακά κ.ο.κ. Έτσι, ο ερευνητής ψάχνει να βρει δικηγόρους (δειγματοληπτικές μονάδες) που να ικανοποιούν τα διάφορα κριτήρια, έτσι ώστε η σύνθεση του δείγματος να είναι σύμφωνη με τα παραπάνω ποσοστά.

4.3. *Συνεχής καταγραφή (Continuous registration)*: Σε πολλές περιπτώσεις στατιστικά στοιχεία συλλέγονται κατά τη λειτουργία κρατικών και άλλων υπηρεσιών και οργανισμών, με συνεχή καταγραφή. Τέτοια στοιχεία περιλαμβάνουν: εκθέσεις από την Ελληνική Στατιστική Αρχή (ΕΛ.ΣΤΑΤ.), από την Στατιστική Υπηρεσία των Ευρωπαϊκών Κοινοτήτων (Eurostat), το Διεθνές Κέντρο Εμπορίου (International Trade Center), στοιχεία ανεργίας από τον ΟΑΕΔ, μετρήσεις κοινής γνώμης από εταιρείες δημοσκοπήσεων κ.λπ. Σήμερα που η πληροφορική κυριαρχεί σε όλες τις δραστηριότητες υπάρχουν χιλιάδες βάσεις δεδομένων με στοιχεία (σε

ηλεκτρονική μορφή όπως CD's και DVD's, ιστοσελίδες στο διαδίκτυο κ.λπ) που καλύπτουν όλες τις δραστηριότητες (εμπόριο, χρηματιστήρια, συνάλλαγμα, τιμές πρώτων υλών, μακροοικονομικά μεγέθη, κ.λπ.). Αυτό το είδος των δεδομένων που προέρχονται από άλλες πηγές καλούνται ***δευτερογενή δεδομένα*** και οι πηγές προέλευσής τους ***δευτερογενείς πηγές***.

Μετά την συγκέντρωση των στατιστικών στοιχείων ακολουθεί η ***επεξεργασία*** τους. Κατά την επεξεργασία ελέγχουμε αν συμπληρώθηκαν σωστά οι απαντήσεις των ερωτηματολογίων ή αν αντιγράφτηκαν σωστά τα διάφορα στοιχεία που ζητήσαμε να συγκεντρωθούν. Μετά την επεξεργασία των στατιστικών στοιχείων ακολουθεί η ***ταξινόμησή*** τους. Αν τα στοιχεία είναι λίγα (< 30) τότε μπορούμε να τα αναλύσουμε χωρίς ταξινόμηση. Όταν όμως είναι περισσότερα τότε πρέπει να τα ταξινομήσουμε ως προς ένα ή περισσότερα χαρακτηριστικά. Έτσι μελετούμε τις ***κατανομές*** δηλ. κατασκευάζουμε τους πίνακες ***συχνοτήτων*** και ***αθροιστικών συχνοτήτων*** και με βάση αυτούς τα ***ραβδογράμματα*** (για ποιοτικές μεταβλητές), τα ***ιστογράμματα*** (για ποσοτικές μεταβλητές) και τα ***κυκλικά διαγράμματα***, Αν θέλουμε να εξετάσουμε συγχρόνως δύο ή και περισσότερα χαρακτηριστικά του πληθυσμού ή του δείγματος τότε φτιάχνουμε έναν ***πίνακα διπλής εισόδου*** όπως π.χ. ο παρακάτω που αφορά την κατανομή αγοριών-κοριτσιών στις διάφορες βαθμίδες εκπαίδευσης κατά το διδακτικό έτος 2011-2012 σύμφωνα με το ΥΠ.Ε.Π.Θ.

Βαθμίδα εκπαίδευσης	Αγόρια	Κορίτσια	Σύνολο
Προσχολική	55.898	52.459	108.357
Δημοτική	486.354	450.769	937.123
Μέση Γενική	277.453	269.563	547.016
Μέση Τεχνική	113.966	17.362	131.328
Ανώτατη (Α.Ε.Ι. & Τ.Ε.Ι.)	59.684	35.701	95.385
Ανώτερη	13.834	7.660	21.494

Συνήθως, για τις επιχειρήσεις και τους διάφορους οργανισμούς και υπηρεσίες, είναι χρήσιμη η παρακολούθηση της διαχρονικής εξέλιξης διαφόρων μεγεθών με απώτερο σκοπό να προβλεφθεί η μελλοντική συμπεριφορά αυτών. Για τον λόγο αυτό φτιάχνουμε πίνακες όπου αναγράφονται τα έτη και η εξέλιξη των μεγεθών που θέλουμε να μελετήσουμε και τους ονομάζουμε ***πίνακες χρονολογικών σειρών***

όπως π.χ. ο παρακάτω που δείχνει την εξέλιξη του πληθυσμού των αγροτών στις χώρες του ευρωπαϊκού νότου κατά τα έτη 1985, 1990, 1994, 1995 σύμφωνα με έκθεση της Ευρωπαϊκής Επιτροπής για την κατάσταση της γεωργίας.

Χώρα	1985	1990	1994	1995
Ιταλία	2.494.100	2.153.400	1.827.600	1.802.000
Γαλλία	1.564.500	1.288.600	1.081.600	1.043.000
Ισπανία	1.300.400	1.121.700	1.060.200	1.025.000
Ελλάδα	931.000	769.200	683.100	665.400
Πορτογαλία	1.020.700	744.400	582.500	599.000

Τα παραπάνω μπορούν να απεικονισθούν σε κοινό διάγραμμα που ονομάζεται ***χρονόγραμμα***. Πίνακες διπλής εισόδου εμφανίζονται και εδώ όπως ο παρακάτω που δείχνει την αναλογία κατοίκων ανά γιατρό και κατοίκων ανά κρεβάτι νοσοκομείου στην Ελλάδα, κατά τα έτη 1970, 1980, 1985, 1995, 1996, 1997, 1998, σύμφωνα με την Ε.Σ.Δ.Υ.

Έτος	κάτοικοι ανά γιατρό	κάτοικοι ανά κρεβάτι
1970	616	160
1980	413	161
1985	341	182
1995	255	155
1996	252	154
1997	244	156
1998	235	156