

## Περιγραφική στατιστική

### 1.1 ΠΕΡΙΓΡΑΦΙΚΗ ΚΑΙ ΣΥΜΠΕΡΑΣΜΑΤΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

Η στατιστική μάς επιτρέπει να περιγράψουμε δεδομένα, να υπολογίζουμε δείκτες οι οποίοι περιγράφουν με συνοπτικό τρόπο σύνθετα φαινόμενα και να εξάγουμε γενικευμένα συμπεράσματα για πληθυσμούς, χρησιμοποιώντας δείγματα από αυτούς τους πληθυσμούς.

Για να γίνει κατανοητό το παραπάνω, ας αναφερθούμε σε παραδείγματα:

Είμαστε αρκετά εξοικειωμένοι με τις έρευνες πρόθεσης ψήφου που διεξάγονται σε μεγάλα δείγματα ερωτώμενων και ανακοινώνονται στον Τύπο. Συνήθως τέτοιες έρευνες πραγματοποιούνται σε δείγματα μεγέθους έστω 1000 ατόμων. Όταν συλλεχθούν τα ερωτηματολόγια και αναλυθούν, παρουσιάζονται τα ευρήματα με τη μορφή πινάκων. Και ας υποθέσουμε ότι, σύμφωνα με τις απαντήσεις των συγκεκριμένων 1000 ατόμων της έρευνας, το κόμμα Α θα έπαιρνε το 30% των ψήφων αν την ερχόμενη Κυριακή διεξάγονταν εκλογές. Αυτό το εύρημα αφορά το συγκεκριμένο δείγμα, αλλά συνήθως ανακοινώνεται σαν να αφορά ολόκληρο το εκλογικό σώμα. Κάθε εύρημα, δείκτης και ποσότητα που εξάγεται από ένα δείγμα *αφορά μόνο το συγκεκριμένο δείγμα και ονομάζεται στατιστικό* (του δείγματος). Είναι φανερό ότι, παίρνοντας ένα άλλο δείγμα μεγέθους 1000, θα υπολογίζαμε πιθανότατα ένα άλλο ποσοστό για το κόμμα Α, που μπορεί να ήταν κοντά στο 30%, αλλά δεν θα ήταν απαραίτητα ίσο με αυτό. Κάθε δείγμα παράγει τα δικά του στατιστικά που δεν συμπίπτουν απαραίτητα

με τα αντίστοιχα στατιστικά ενός άλλου δείγματος, ίσου μεγέθους, που επιλέχθηκε για τον ίδιο σκοπό. Η εξαγωγή και ο υπολογισμός τέτοιων στατιστικών, και των γραφικών παραστάσεων που τα αφορούν, είναι το αντικείμενο της *περιγραφικής στατιστικής*. Πότε λοιπόν κάνουμε περιγραφική στατιστική; Όταν θέλουμε να περιγράψουμε ένα δείγμα, γνωρίζοντας ότι τα στατιστικά του αφορούν το συγκεκριμένο μόνο δείγμα και όχι όλα τα δείγματα, ούτε και τον πληθυσμό από τον οποίο προέρχονται.

Περιγραφική στατιστική κάνουμε και όταν έχουμε στη διάθεσή μας ολόκληρο τον πληθυσμό. Στις λίγες τέτοιες περιπτώσεις, όπως είναι οι απογραφές πληθυσμού, το σύνολο των ψηφοδελτίων που συλλέγονται την ημέρα των εκλογών μετά την ψηφοφορία ή το σύνολο στοιχείων που μπορούν να συλλέξουν διάφοροι οργανισμοί και φορείς από όλα τα μέλη τους, μπορούμε να κάνουμε περιγραφική στατιστική. Στην περίπτωση που περιγράφουμε στατιστικά έναν πληθυσμό, τα στατιστικά που εξάγουμε ονομάζονται *παραμέτροι* και είναι μοναδικά, γιατί προφανώς ο πληθυσμός είναι ένας και μοναδικός, όχι όπως στην περίπτωση των δειγμάτων, τα οποία είναι πολλά. Αν λοιπόν μελετάμε την εκλογική συμπεριφορά, μπορούμε να πάρουμε ένα δείγμα μεγέθους 1000 ψηφοφόρων και να υπολογίσουμε διάφορα στατιστικά που αφορούν το δείγμα. Και αν είχαμε πάρει ένα άλλο δείγμα, θα υπολογίζαμε πιθανότατα διαφορετικές τιμές για το ίδιο είδος στατιστικών. Αν μπορούσαμε όμως να πάρουμε ολόκληρο τον πληθυσμό των ψηφοφόρων (αυτό πράγματι γίνεται με τα επίσημα εκλογικά αποτελέσματα την ημέρα των εκλογών), τότε θα είχαμε μοναδικές τιμές των παραμέτρων που θα περιέγραφαν ολόκληρο τον πληθυσμό.

Το επόμενο στάδιο είναι να προσπαθήσουμε να «προβλέψουμε» ποια θα είναι η τιμή μιας παραμέτρου του πληθυσμού βασιζόμενοι στην τιμή ενός στατιστικού. Με άλλα λόγια, θα μπορούσαμε να υπολογίσουμε την τιμή του ποσοστού που θα έπαιρνε το κόμμα Α στις εκλογές (όπου υποτίθεται ότι ψηφίζει ολόκληρο το εκλογικό σώμα, και άρα πρόκειται για μια, κατά κάποιον τρόπο, απογραφή), δηλαδή να υπολογίσουμε την τιμή μιας παραμέτρου βασιζόμενοι στην τιμή του στατιστικού. Στην περίπτωση μας, το στατιστικό είναι το ποσοστό του κόμματος Α που υπολογίζεται από τις απαντήσεις των 1000 ψηφοφόρων του δείγματος που χρησιμοποιούμε.

Αν το δείγμα είναι τυχαίο, τότε η γενίκευση αυτή που προσπαθούμε να κάνουμε από το δείγμα προς τον πληθυσμό εξαρτάται από το μέγεθος του δείγματος, εφόσον ο πληθυσμός είναι πολύ μεγάλος ή άπειρος. Εξαρτάται δηλαδή από τον αριθμό 1000. Αν το μέγεθος ήταν μεγαλύτερο, τότε η γενίκευση θα ήταν πιο ακριβής, ενώ αν είναι μικρότερο από 1000, τότε η γενίκευση θα ήταν λιγότερο ακριβής. Υπάρχει συγκεκριμένη διαδικασία και φόρμουλα που υπολογίζει πόσο ακριβής είναι αυτή η γενίκευση.

Μέχρι τώρα αναφερθήκαμε σε πρόβλεψη ή γενίκευση εννοώντας το ίδιο πράγμα, δηλαδή τη διαδικασία με την οποία συμπεραίνουμε για την παράμετρο του πληθυσμού βασιζόμενοι στο στατιστικό του δείγματος. Από τώρα και στο εξής θα αναφέρουμε τη διαδικασία αυτή, ορθά, ως εκτίμηση της παραμέτρου από το στατιστικό.

Οι εκτιμήσεις αποτελούν αντικείμενο μελέτης ενός κλάδου της στατιστικής που ονομάζεται *εκτιμητική ή συμπερασματική στατιστική ή επαγωγική στατιστική*. Εργαζόμενοι μέσω της εκτιμητικής, ξεκινάμε συνήθως από την τιμή του στατιστικού και υπολογίζουμε ένα διάστημα μέσα στο οποίο ανήκει η παράμετρος του πληθυσμού. Το διάστημα ονομάζεται *διάστημα εμπιστοσύνης* και το εύρος του επηρεάζεται από το μέγεθος του δείγματος που χρησιμοποιήθηκε για τον υπολογισμό του στατιστικού, αλλά και από τον βαθμό βεβαιότητας-εμπιστοσύνης που θέλουμε να έχουμε κατά την εκτίμησή μας, δεδομένου ότι στη στατιστική δεν είμαστε 100% σίγουροι για την εκτίμησή μας. Τελικά, αυτό που παράγεται είναι, όπως είπαμε, το διάστημα εμπιστοσύνης, για το οποίο, χωρίς να μας το έχουν διδάξει, έχουμε ήδη εμπειρία για το τι σημαίνει. Πρόκειται για το ποσοστό  $\pm 2\%$  ή  $3\%$  ή  $4\%$  κ.ο.κ., για το οποίο μας πληροφορούν οι έρευνες γνώμης και μας το αναφέρουν ως σφάλμα εκτίμησης. Είναι σαν να μας λένε ότι στο δείγμα μας βρήκαμε το ποσοστό 30% και στον πληθυσμό η τιμή του ποσοστού-παραμέτρου είναι  $30\% +$  ή  $- 3\%$ , δηλαδή η παράμετρος κυμαίνεται από 27% έως 33%, και αυτό με μια σχετική βεβαιότητα που συνήθως ισούται με 95%.

Τελικά, συνοψίζοντας τα παραπάνω, μπορούμε να δώσουμε τον παρακάτω πίνακα, ο οποίος περιγράφει συμβολικά και συνοπτικά όσα αναφέρθηκαν:

Δείγμα (στατιστικά)		Πληθυσμός (παράμετροι)
$p$ , δειγματικό ποσοστό	→	$p$ , αναλογία στον πληθυσμό
$\bar{x}$ , μέσος όρος	→	$\mu$ , μέση τιμή
$s^2$ , δειγματική διασπορά	→	$\sigma^2$ , διασπορά του πληθυσμού
$s$ , δειγματική τυπική απόκλιση	→	$\sigma$ , τυπική απόκλιση του πληθυσμού

Παρατηρήστε ότι ο δειγματικός μέσος αναφέρεται ως *μέσος όρος*, ενώ για τον πληθυσμό αναφερόμαστε στη *μέση τιμή*. Επίσης χρησιμοποιούνται λατινικά γράμματα για τα στατιστικά, ενώ για τις παραμέτρους χρησιμοποιούνται ελληνικά γράμματα, επειδή οι παράμετροι αντιπροσωπεύουν μαθηματικές οντότητες και συμβολίζονται με ελληνικά γράμματα, όπως στα μαθηματικά. Τέλος, η περιγραφική στατιστική ασχολείται μόνο με την περιγραφή των δειγμάτων, δηλαδή την εξαγωγή των στατιστικών, και κάποιες φορές με την εξαγωγή των τιμών των παραμέτρων, όταν μελετά ολόκληρους πληθυσμούς. Για κάθε δείγμα υπολογίζουμε μια διαφορετική ομάδα στατιστικών. Παίρνοντας ένα άλλο δείγμα ίδιου μεγέθους με το προηγούμενο, υπολογίζουμε διαφορετικά στατιστικά. Για τον πληθυσμό, όμως, οι παράμετροι έχουν μοναδικές τιμές.

Η εκτιμητική βασίζεται σε δείγματα για να εκτιμήσει τον πληθυσμό. Προσεγγίζει την κάθε παράμετρο βάσει του αντίστοιχου στατιστικού που υπολογίζεται από το δείγμα. Εκτιμά τη μέση τιμή από τον μέσο όρο, την αναλογία από το ποσοστό στο δείγμα (θυμηθείτε το ποσοστό του κόμματος 30% – αναλογία στον πληθυσμό θα είναι

η αναλογία ψηφοφόρων που ψήφισαν το κόμμα Α), τη διασπορά και την τυπική απόκλιση του πληθυσμού από τη διασπορά και την τυπική απόκλιση του δείγματος.

Αυτό που θα πρέπει να έχει κατά νου από την αρχή το αναγνωστικό κοινό είναι ότι το βασικό μας μέλημα είναι να εκτιμήσουμε παραμέτρους του πληθυσμού βασιζόμενοι στα στατιστικά του δείγματος. Γι' αυτό τις πιο πολλές φορές ορίζουμε με τέτοιον τρόπο τους τύπους υπολογισμού των στατιστικών, ώστε να *αποτελούν «καλύτερες» εκτιμήσεις των παραμέτρων του πληθυσμού*. Θυμηθείτε το αυτό όταν θα συζητήσουμε παρακάτω για τη διασπορά, την τυπική απόκλιση και τους βαθμούς ελευθερίας.

## 1.2 ΜΕΤΑΒΛΗΤΕΣ – ΤΙΜΕΣ – ΠΑΡΑΤΗΡΗΣΕΙΣ

Στη στατιστική και στις πιθανότητες εργαζόμαστε με μεταβλητές. Φυσικά, λέγονται έτσι γιατί πρόκειται για μαθηματικές έννοιες των οποίων οι τιμές μεταβάλλονται, και μάλιστα μπορούμε να αποδώσουμε πιθανότητες στις τιμές τους. Ας το αφήσουμε όμως το τελευταίο για αργότερα. Ας περιοριστούμε προς το παρόν στο πρώτο, στο ότι οι τιμές τους μεταβάλλονται. Αυτό, βέβαια, δεν συνιστά αυστηρό ορισμό. Στο βιβλίο αυτό θα δώσουμε έναν ορισμό της μεταβλητής που, ενώ είναι αυστηρός και μαθηματικά ορθός, δεν χρησιμοποιεί περίπλοκους όρους και αναφορές. Ορίζουμε λοιπόν ως *μεταβλητή μία ερώτηση (ας πούμε σε ένα ερωτηματολόγιο) η οποία επιδέχεται μία μοναδική απάντηση από κάθε ερωτώμενο και ερωτώμενη*.

Αυτός ο ορισμός στην ουσία ορίζει τη μεταβλητή ως συνάρτηση. Παραδείγματα μεταβλητών αποτελούν τα παρακάτω:

Χρώμα μαλλιών (Ερώτηση: Ποιο είναι το χρώμα των μαλλιών σας;) Κάθε ερωτώμενος δηλώνει το χρώμα των μαλλιών του.

Βαθμός στη στατιστική (Ερώτηση: Τι βαθμό πήρατε στη στατιστική;) Κάθε ερωτώμενος αναφέρει τον βαθμό που πήρε.

Κόμμα που ψηφίσατε (Ερώτηση: Ποιο κόμμα ψηφίσατε;) Δεδομένου ότι κάθε ψηφοφόρος ψηφίζει ένα μόνο κόμμα, αυτή η ερώτηση αποτελεί μεταβλητή.

Αν όμως ρωτήσουμε ποια κόμματα συμπαθείτε και η ερωτώμενη μπορεί να δώσει πολλές απαντήσεις, τότε η ερώτηση δεν είναι μεταβλητή (αν και μπορεί να μετατραπεί σε πολλές μεταβλητές, τόσες όσες είναι οι απαντήσεις που μπορούν να δοθούν).

*Οι μεταβλητές έχουν τιμές.* Ως τιμές ορίζονται οι δυνατές απαντήσεις μιας ερώτησης-μεταβλητής. Όλα τα χρώματα μαλλιών, όλοι οι δυνατοί βαθμοί στη στατιστική, η λίστα των κομμάτων που κατέβηκαν στις εκλογές, όλα αποτελούν τιμές των μεταβλητών που περιγράφηκαν αντίστοιχα παραπάνω.

Από την άλλη, *παρατηρήσεις* είναι οι απαντήσεις των ερωτώμενων. Κάθε ερωτώμενος επιλέγει από το σύνολο των τιμών μίας μεταβλητής, η οποία αποτελεί την απάντησή του, είναι η παρατήρηση για τον συγκεκριμένο ερωτώμενο. Μία τιμή για κάθε

ερωτώμενο. Άρα οι παρατηρήσεις είναι όσες και οι ερωτώμενοι και οι ερωτώμενες της έρευνάς μας και η κάθε παρατήρηση ισούται με κάποια τιμή. Μόνο που οι παρατηρήσεις συνήθως είναι πολλές και οι τιμές που μεταφέρονται στις παρατηρήσεις επαναλαμβάνονται και επανέρχονται. Έτσι, το κόμμα Α, που αποτελεί τιμή για τη μεταβλητή «Ποιο κόμμα ψηφίσατε;», μπορεί να αναφέρεται από πολλούς ερωτώμενους και η τιμή να επαναλαμβάνεται πολλές φορές ως παρατήρηση.

### 1.3 ΕΙΔΗ ΜΕΤΑΒΛΗΤΩΝ

Το είδος μιας μεταβλητής καθορίζεται από τις τιμές που μπορεί να πάρει. Έτσι, υπάρχουν μεταβλητές που μπορούν να πάρουν διακριτές τιμές (όπως «Ναι» και «Όχι») ή τιμές που παρουσιάζουν συνέχεια, όπως η μέτρηση του ύψους ή του βάρους κ.λπ.

Οι μεταβλητές χωρίζονται γενικά σε δύο μεγάλες κατηγορίες:

- A) Ποιοτικές:** Οι τιμές τους δεν εκφράζουν ποσότητες. Τέτοιες μεταβλητές είναι, για παράδειγμα, το φύλο με τιμές «Άνδρας», «Γυναίκα», «Άλλο». Οι ποιοτικές μεταβλητές είναι κατηγορικές.
- B) Ποσοτικές:** Οι τιμές τους εκφράζουν ποσότητες. Τέτοιες μεταβλητές είναι το ύψος, το βάρος, μεταβλητές που εκφράζουν αριθμό-πλήθος αντικειμένων, ηλικία σε χρόνια κ.λπ.

Μια διάκριση μεταξύ ποσοτικών μεταβλητών είναι η εξής:

- A) Διακριτές:** Είναι όσες έχουν τιμές σε ένα σύνολο διακεκριμένων και συνήθως πεπερασμένων τιμών.
- B) Συνεχείς:** Είναι όσες έχουν τιμές μέσα σε ένα συνεχές διάστημα ή σε μια συλλογή διαστημάτων. Για παράδειγμα, η μεταβλητή βάρος είναι μια συνεχής μεταβλητή, κι αυτό επειδή για οποιοδήποτε δύο τιμές βάρους μπορούμε να βρούμε μια άλλη τιμή η οποία να βρίσκεται ενδιάμεσα, π.χ. για βάρος 80 κιλά, βάρος 82 κιλά, θα μπορούσε να υπάρξει τιμή βάρους κάποιου ερωτώμενου που να βρίσκεται ανάμεσα στα 80 και στα 82 κιλά. Επίσης, οι τιμές κυμαίνονται συνήθως σε ένα διάστημα, για παράδειγμα από 50-110 κιλά.

Συνοψίζοντας, θα αναφέρουμε την πλέον τυπική κατηγοριοποίηση των μεταβλητών όπως αυτή αναφέρεται στη σχετική βιβλιογραφία.

#### *Είδη μεταβλητών*

**Ονομαστικές:** Πρόκειται για μεταβλητές που δέχονται ως τιμές κατηγορίες χωρίς αριθμητική αξία. Επιδέχονται στατιστική ανάλυση που κάνει χρήση της συχνότητας για κάθε τιμή. Τέτοιες μεταβλητές μπορεί να είναι το φύλο, με τιμές Άνδρας-Γυναίκα-Άλλο, το χρώμα των μαλλιών ή το χρώμα των ματιών, η χώρα καταγωγής ή προέλευσης, το κόμμα που προτίθεται να ψηφίσει κάποιος.

**Διάταξης:** Οι τιμές αυτών των μεταβλητών εκφράζουν διάταξη και παριστάνονται με αριθμητικές τιμές, ξεκινώντας συνήθως από το 1 ή το 0. Η μεταβλητή «επίπεδο εκπαίδευσης» μπορεί να δέχεται τρεις τιμές ως απάντηση: «πρωτοβάθμια», «δευτεροβάθμια» και «τριτοβάθμια». Είναι σαφές ότι μπορούμε να μιλάμε για χαμηλότερο και υψηλότερο επίπεδο εκπαίδευσης. Μπορούμε να αντιστοιχίζουμε τις τρεις βαθμίδες εκπαίδευσης με αριθμούς: πρωτοβάθμια = 1, δευτεροβάθμια = 2 και τριτοβάθμια = 3 με την ιδιότητα ότι  $1 < 2 < 3$ . Δεν επιτρέπεται να γίνουν οποιεσδήποτε πράξεις με τις τιμές αλλά επιτρέπεται η σύγκριση τιμών. Επίσης, επιτρέπονται οι στατιστικές αναλύσεις που γίνονται για τις ονομαστικές μεταβλητές. Τέλος, για τις μεταβλητές διάταξης μπορούν να υπολογιστούν ποσοστιαία σημεία (όπως θα δούμε παρακάτω).

**Διαστήματος:** Οι τιμές αυτών των ποσοτικών μεταβλητών έχουν αριθμητική αξία. Επιτρέπονται οι πράξεις της πρόσθεσης και της αφαίρεσης αλλά όχι του πολλαπλασιασμού και της διαίρεσης. Παράδειγμα τέτοιας μεταβλητής είναι η θερμοκρασία, για τη μέτρηση της οποίας χρησιμοποιούμε κλίμακες όπως του Κελσίου. Αν σε κάποιο σημείο η θερμοκρασία είναι 15 βαθμοί Κελσίου και αυξηθεί σε 30 βαθμούς, τότε έχουμε αύξηση 15 βαθμών. Η πρόσθεση και η αφαίρεση επιτρέπονται, μπορούμε δηλαδή να μιλάμε για αύξηση 15 βαθμών Κελσίου. Δεν μπορούμε να πούμε όμως ότι διπλασιάστηκε η θερμότητα στο σημείο που μετράμε επειδή διπλασιάστηκε η θερμοκρασία σε βαθμούς Κελσίου. Αν μετρίοταν σε βαθμούς Φαρενάιτ και επιχειρούσαμε να διαιρέσουμε τις δύο αντίστοιχες θερμοκρασίες, δεν θα βρίσκαμε πηλίκο 2. Ας σημειωθεί ότι μηδέν βαθμοί Κελσίου αντιστοιχούν στους 32 βαθμούς Φαρενάιτ. Άρα, *η τιμή μηδέν δεν εκφράζει απουσία θερμότητας*. Για τις μεταβλητές διαστήματος, *η τιμή μηδέν δεν ισοδυναμεί με απουσία του φαινομένου*.

**Αναλογίας:** Σε αυτές τις ποσοτικές μεταβλητές οι τιμές αντιστοιχούν αναλογικά στην ποσότητα του χαρακτηριστικού που μετρούν. Οι τιμές έχουν αριθμητική αξία και επιτρέπονται όλες οι αριθμητικές πράξεις και οι υπολογισμοί όλων των στατιστικών. Το βάρος είναι μια μεταβλητή αναλογίας. Σε οποιαδήποτε σύστημα και αν μετράμε το βάρος, διατηρούνται οι αναλογίες και η τιμή μηδέν δηλώνει την πλήρη απουσία βάρους. Η τιμή μηδέν ανήκει στο σύνολο τιμών, ενώ εκφράζει την απουσία του φαινομένου που μετράμε.

Στο κεφάλαιο αυτό θα ασχοληθούμε κυρίως με ποσοτικές μεταβλητές, αν και η περιγραφική στατιστική αφορά και τις ποιοτικές.

Ας υποθέσουμε ότι έχουμε μια ποσοτική μεταβλητή  $X$ . Τη μετράμε σε ένα δείγμα μεγέθους  $n$ , άρα έχουμε  $n$  παρατηρήσεις. Για να διευκολυνθείτε, μπορείτε να υποθέσετε ότι η  $X$  μετρά τη βαθμολογία των φοιτητών και των φοιτητριών στο μάθημα της στατιστικής. Οι τιμές της μεταβλητής είναι από το 1 έως το 10 και κάθε φοιτητής και φοιτήτρια παίρνει έναν βαθμό που είναι και η παρατήρηση γι' αυτόν τον φοιτητή ή τη φοιτήτρια. Οι παρατηρήσεις συμβολίζονται με  $x_i$ , δηλαδή  $x_1, x_2, x_3, \dots, x_n$ .