

Κεφάλαιο 1

Περιγραφική Στατιστική

1.1 Περιγραφική και Συμπερασματική Στατιστική

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

Ας μιλήσουμε με παραδείγματα για να γίνει κατανοητό το παραπάνω:

Είμαστε αρκετά εξοικειωμένοι με τις έρευνες πρόθεσης ψήφου που γίνονται σε μεγάλα δείγματα ερωτώμενων και ανακοινώνονται στον Τύπο. Συνήθως τέτοιες έρευνες γίνονται σε δείγματα μεγέθους, ας πούμε, 1.000 ατόμων. Όταν συλλεχθούν τα ερωτηματολόγια και αναλυθούν, παρουσιάζονται τα ευρήματα με τη μορφή πινάκων. Ας υποθέσουμε ότι σύμφωνα με τις απαντήσεις των συγκεκριμένων 1.000 ατόμων της έρευνας, το κόμμα Α θα έπαιρνε το 30% των ψήφων αν γίνονταν αύριο εκλογές. Αυτό το εύρημα αφορά το συγκεκριμένο δείγμα, αλλά συνήθως ανακοινώνεται σαν να αφορά ολόκληρο το εκλογικό σώμα. Κάθε εύρημα, δείκτης και ποσότητα από τα οποία εξάγεται ένα δείγμα *αφορά μόνο το συγκεκριμένο δείγμα και ονομάζεται Στατιστικό του δείγματος*. Είναι φανερό ότι παίρνοντας ένα άλλο δείγμα μεγέθους 1.000, 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 επιλέχτηκε 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 *Παράμετροι* 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 1.000 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 υπολογίζουμε την τιμή του ποσοστού που θα έπαιρνε το κόμμα Α στις εκλογές (όπου υποτίθεται ότι ψηφίζει ολόκληρο το εκλογικό σώμα και άρα πρόκειται για απογραφή), δηλαδή να υπολογίσουμε την τιμή μιας παραμέτρου βασιζόμενοι στην τιμή του στατιστικού; Το στατιστικό είναι στην περίπτωσή μας το ποσοστό του κόμματος Α που υπολογίζεται από τις απαντήσεις των 1.000 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 1.000. Αν το μέγεθος ήταν μεγαλύτερο, τότε η γενίκευση θα ήταν πιο ακριβής, ενώ αν είναι μικρότερο από 1.000 τότε η γενίκευση θα ήταν λιγότερη 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 βασισμένοι 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 το 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 που χωρίς να μας το έχουν διδάξει, έχουμε ήδη εμπειρία για το τι σημαίνει. Πρόκειται για το ποσοστό $\pm 2\%$ ή 3% ή 4% που μας πληροφορούν οι έρευνες γνώμης και που 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% .

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

Δείγμα (στατιστικά)	Πληθυσμός (παράμετροι)
p, δειγματικό ποσοστό	→ ρ, αναλογία στον πληθυσμό
$\bar{x}$, μέσος όρος	→ μ, μέση τιμή
s^2 , δειγματική διασπορά	→ σ^2 , διασπορά του πληθυσμού
s, δειγματική τυπική απόκλιση	→ σ, 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 μένει από την αρχή στον αναγνώστη είναι ότι το βασικό μας μέλημα είναι να εκτιμήσουμε παραμέτρους του πληθυσμού βασιζόμενοι στα στατιστικά του δείγματος. Γι' αυτό τις πιο πολλές φορές ορίζουμε έτσι τους τύπους υπολογισμού των στατιστικών ώστε να *αποτελούν «καλύτερες» εκτιμήσεις των παραμέτρων του πληθυσμού*. Θυμηθείτε το αυτό όταν θα μιλήσουμε παρακάτω για τη διασπορά και την τυπική απόκλιση.

1.2 Μεταβλητές - Τιμές - Παρατηρήσεις

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 δεύτερο για αργότερα. Ας περιοριστούμε προς το παρόν στο πρώτο, στο ότι οι τιμές τους μεταβάλλονται. Αυτό δεν συνιστά αυστηρό ορισμό βεβαίως. 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 μια ερώτηση (ας πούμε σε ένα ερωτηματολόγιο), η οποία επιδέχεται μία μοναδική απάντηση από κάθε ερωτώμενο.*

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

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

Βαθμός στη Στατιστική (Ερώτηση: τι βαθμό πήρες στη Στατιστική;) Κάθε ερωτώμενος απαντάει και λέει το βαθμό που πήρε.

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

Αν όμως ρωτήσουμε ποια κόμματα συμπαθείτε και ο ερωτώμενος 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 είναι οι τιμές των μεταβλητών που περιγράφηκαν αντίστοιχα παραπάνω.

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

ερωτώμενοι 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 Είδη μεταβλητών

Το είδος μιας μεταβλητής καθορίζεται από τις δυνατές 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).

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 και Γυναίκα. 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 κ.λπ.

Μια διάκριση μεταξύ ποσοτικών μεταβλητών είναι η εξής:

A) Διακριτές: Είναι όσες έχουν τιμές σε ένα σύνολο διακεκριμένων και συνήθως πεπερασμένων τιμών, όπως είναι η μεταβλητή που αναφέραμε.

B) Συνεχείς: Είναι όσες έχουν τιμές μέσα σ' ένα συνεχές διάστημα ή σε μια συλλογή διαστημάτων. Για παράδειγμα, η μεταβλητή Βάρος είναι μια συνεχής μεταβλητή κι αυτό επειδή για οποιεσδήποτε 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 ερωτωμένου που να βρίσκεται ανάμεσα στα 80 και τα 82 κιλά. Επίσης οι τιμές κυμαίνονται συνήθως σε ένα διάστημα, για παράδειγμα από 50-110 κιλά.