ΜΕΡΟΣ ΠΡΩΤΟ
Στατιστικό υπόβαθρο και βασικός χειρισμός δεδομένων

ΚΕΦΑΛΑΙΟ 1
Βασικές έννοιες

ΠΕΡΙΕΧΟΜΕΝΑ ΚΕΦΑΛΑΙΟΥ

Εισαγωγή

Αυτό το κεφάλαιο περιγράφει ορισμένες από τις βασικές έννοιες που διέπουν μεγάλο μέρος του βιβλίου, συμπεριλαμβανομένων των εννοιών της κατανομής του πληθυσμού και της δειγματικής κατανομής, της σημασίας της τυχαίας δειγματοληψίας, του νόμου των μεγάλων αριθμών και του κεντρικού οριακού θεωρήματος. Στη συνέχεια παρουσιάζει πώς αυτές οι ιδέες υποστηρίζουν την τυπική προσέγγιση για τον έλεγχο υποθέσεων και την κατασκευή διαστημάτων εμπιστοσύνης.

Η οικονομετρία έχει ένα πλήθος από ρόλους όσον αφορά την πρόβλεψη και την ανάλυση πραγματικών δεδομένων και προβλημάτων. Στο επίκεντρο αυτών των ρόλων, ωστόσο, βρίσκεται η επιθυμία να γίνουν αντιληπτά τα μεγέθη των επιδράσεων και να ελεγχθεί η σημαντικότητά τους. Η οικονομική θεωρία συχνά δείχνει προς την κατεύθυνση μιας αιτιώδους σχέσης (εάν το εισόδημα αυξάνεται, μπορούμε να αναμένουμε ότι θα αυξηθεί και η κατανάλωση), αλλά η θεωρία σπάνια υποδεικνύει ένα ακριβές μέγεθος. Ωστόσο, σε ένα πλαίσιο πολιτικής ή επιχειρηματικό, το να έχεις μια ξεκάθαρη εικόνα για το μέγεθος μιας επίδρασης μπορεί να είναι εξαιρετικά σημαντικό, και αυτό είναι η ειδικότητα της οικονομετρίας.

Ο σκοπός αυτού του κεφαλαίου είναι να αποσαφηνίσει ορισμένους βασικούς ορισμούς και ιδέες ώστε να παρέχει στον φοιτητή μια διαισθητική κατανόηση αυτών των βασικών εννοιών. Επομένως, η απόδοσή τους εδώ θα είναι σκόπιμα λιγότερο τυπική απ' ό,τι στο μεγαλύτερο μέρος του βιβλίου.

Ένα απλό παράδειγμα

Ας δούμε ένα πολύ απλό παράδειγμα για να εξηγηθεί η ιδέα που παρατίθεται εδώ. Ο Πίνακας 1.1 δείχνει τη μέση ηλικία θανάτου ανδρών και γυναικών στα 15 κράτη-μέλη της Ευρωπαϊκής Ένωσης (ΕΕ) το 2002.

Από αυτά τα στοιχεία, καθίσταται αρκετά προφανές ότι οι γυναίκες αναμένεται να ζήσουν περισσότερο από τους άνδρες σε καθεμία από τις χώρες αυτές, και εάν λάβουμε τον μέσο όρο όλων των χωρών, μπορούμε να δούμε ότι και πάλι, σε πανευρωπαϊκή βάση, ότι οι γυναίκες τείνουν να ζουν περισσότερο από τους άνδρες. Ωστόσο, υπάρχει μια αρκετά σημαντική διαφοροποίηση μεταξύ των χωρών, και ίσως είναι εύλογο να διερωτηθούμε εάν γενικά, στον παγκόσμιο πληθυσμό, θα αναμέναμε οι γυναίκες να ζουν περισσότερο από τους άνδρες.

Ένας τυπικός τρόπος για να το προσεγγίσουμε θα ήταν να δούμε τη διαφορά στο μέσο προσδόκιμο ζωής για όλη την Ευρώπη και να εξετάσουμε εάν αυτή είναι στατιστικά διαφορετική από το μηδέν. Αυτό προϋποθέτει κάποια βασικά βήματα: αρχικά θα πρέπει να εκτιμηθεί η διαφορά στο μέσο προσδόκιμο ζωής, έπειτα θα πρέπει να κατασκευαστεί ένα μέτρο της αβεβαιότητάς της και, τέλος, θα πρέπει να ελεγχθεί η υπόθεση ότι η διαφορά είναι μηδέν.

Πίνακας 1.1 Μέσος όρος ηλικίας θανάτου για τα 15 κράτη-μέλη της ΕΕ(2002)

	Γυναίκες	**Άνδρες**
Αυστρία	81,2	75,4
Βέλγιο	81,4	75,1
Δανία	79,2	74,5
Φινλανδία	81,5	74,6
Γαλλία	83,0	75,5
Γερμανία	80,8	74,8
Ελλάδα	80,7	75,4
Ιρλανδία	78,5	73,0
Ιταλία	82,9	76,7
Λουξεμβούργο	81,3	74,9
Ολλανδία	80,6	75,5
Πορτογαλία	79,4	72,4
Ισπανία	82,9	75,6
Σουηδία	82,1	77,5
Ηνωμένο Βασίλειο	79,7	75,0
Μέσος όρος	81,0	75,1
Τυπική απόκλιση	1,3886616	1,2391241

Ο Πίνακας 1.1 δίνει τον μέσο όρο (ή μέσο) του προσδόκιμου ζωής των ανδρών και γυναικών για την ΕΕ ως σύνολο ως

$$\bar{Y}_w = \frac{1}{15}\sum_{i=1}^{15} Y_{wi} \qquad \bar{Y}_m = \frac{1}{15}\sum_{i=1}^{15} Y_{mi} \tag{1.1}$$

όπου $\bar{Y}_w$ είναι το μέσο προσδόκιμο ζωής για τις γυναίκες στην ΕΕ και $\bar{Y}_m$ είναι το μέσο προσδόκιμο ζωής για τους άνδρες. Μια τυπική εκτίμηση της διαφοράς μεταξύ των δύο μέσων είναι $(\bar{Y}_w - \bar{Y}_m)$. Ο Πίνακας 1.1 δίνει επίσης της μέση διασπορά για καθέναν από τους μέσους αυτούς, οριζόμενη ως τυπική απόκλιση, η οποία δίνεται από:

$$S.D.j = \sqrt{\sum_{i=1}^{15}(Y_{ji} - \bar{Y}_j)^2} \quad j = w, m \tag{1.2}$$

Καθώς έχουμε μια εκτίμηση της διαφοράς και μια εκτίμηση της αβεβαιότητας των μετρήσεών μας, μπορούμε τώρα να διατυπώσουμε έναν τυπικό έλεγχο υπόθεσης. Ο έλεγχος για τη διαφορά μεταξύ των δύο μέσων είναι:

$$t = \frac{\bar{Y}_w - \bar{Y}_m}{\sqrt{\frac{s_w^2}{15} + \frac{s_m^2}{15}}} = \frac{81 - 75.1}{\sqrt{\frac{1.389^2}{15} + \frac{1.24^2}{15}}} = 12.27 \qquad (1.3)$$

Η στατιστική συνάρτηση *t* (*t*-statistic) του 12.27 > 1.96, το οποίο σημαίνει ότι υπάρχει λιγότερο από 5% πιθανότητα εύρεσης μιας στατιστικής συνάρτησης ίσης με 12.27 καθαρά από τύχη όταν η πραγματική διαφορά είναι μηδέν. Επομένως, μπορούμε να συμπεράνουμε ότι υπάρχει σημαντική διαφορά μεταξύ του προσδόκιμου ζωής των ανδρών και γυναικών.

Αν και αυτό φαίνεται αρκετά κατανοητό και απλό, υπάρχουν ορισμένα υποκείμενα λεπτά σημεία, και αυτά αποτελούν το αντικείμενο αυτού του κεφαλαίου. Οι ερωτήσεις που θα διερευνηθούν είναι: Ποιο θεωρητικό πλαίσιο δικαιολογεί όλο αυτό; Γιατί η διαφορά των μέσων αποτελεί μια καλή εκτίμηση της επιπλέον διάρκειας ζωής για τις γυναίκες; Είναι αυτή μια καλή εκτίμηση για τον κόσμο ως σύνολο; Ποιο είναι το μέτρο αβεβαιότητας που αποδίδεται από την τυπική απόκλιση και τι σημαίνει πραγματικά; Στην ουσία, ποιο είναι το υποκείμενο θεωρητικό πλαίσιο που δικαιολογεί αυτό που συνέβη;

Στατιστικό πλαίσιο

Το στατιστικό πλαίσιο που διέπει την παραπάνω προσέγγιση στηρίζεται σε μια σειρά βασικών εννοιών, πρώτη εκ των οποίων είναι ο πληθυσμός. Υποθέτουμε ότι υπάρχει ένας πληθυσμός από γεγονότα ή οντότητες για τον οποίο ενδιαφερόμαστε. Ο πληθυσμός αυτός υποτίθεται ότι είναι απείρως μεγάλος και περιλαμβάνει όλα τα αποτελέσματα που μας αφορούν. Τα δεδομένα στον Πίνακα 1.1 αφορούν τις χώρες της ΕΕ 15 για το έτος 2002. Εάν ενδιαφερόμασταν μόνο για το συγκεκριμένο έτος και για το συγκεκριμένο σύνολο χωρών, τότε δεν θα υπήρχε στατιστικό ερώτημα που θα έπρεπε να τεθεί. Σύμφωνα με τα δεδομένα, οι γυναίκες ζούσαν περισσότερο από τους άνδρες για το συγκεκριμένο έτος και στη συγκεκριμένη περιοχή. Αυτό είναι απλώς ένα γεγονός. Αλλά ο πληθυσμός είναι πολύ μεγαλύτερος· περιλαμβάνει όλους τους άνδρες και γυναίκες σε όλες τις περιόδους, και για να εξαχθεί ένα συμπέρασμα για αυτό τον πληθυσμό χρειαζόμαστε ένα στατιστικό πλαίσιο. Θα μπορούσε, για παράδειγμα, να είναι απλώς τυχαίο γεγονός ότι οι γυναίκες ζουν περισσότερο από τους άνδρες σε αυτό το συγκεκριμένο έτος. Πώς μπορούμε να το εξακριβώσουμε αυτό;

Οι επόμενες σημαντικές έννοιες είναι οι τυχαίες μεταβλητές και η κατανομή του πληθυσμού. Μια τυχαία μεταβλητή είναι απλά ένα μέτρο ενός οποιουδήποτε γεγονότος που συμβαίνει με αβέβαιο τρόπο. Επομένως, η ηλικία, για παράδειγμα, στην οποία ένα άτομο πεθαίνει είναι αβέβαιη, και επομένως η ηλικία θανάτου ενός ατόμου είναι μια τυχαία μεταβλητή. Μόλις ένα άτομο πεθάνει, η ηλικία θανάτου παύει να είναι μια τυχαία μεταβλητή και μετατρέπεται σε μια παρατήρηση ή έναν αριθμό. Η κατανομή του πληθυσμού ορίζει την πιθανότητα να συμβεί ένα συγκεκριμένο γεγονός· για παράδειγ-

μα, είναι η κατανομή του πληθυσμού η οποία θα ορίσει την πιθανότητα ένας άνδρας να πεθάνει πριν γίνει 60 ($\Pr(Y_m < 60)$). Η κατανομή το πληθυσμού έχει διάφορες ροπές που καθορίζουν το σχήμα της. Οι πρώτες δύο ροπές είναι ο μέσος (ορισμένες φορές ονομάζεται αναμενόμενη τιμή, $E(Y_m) = \mu_{Ym}$, ή μέσος όρος) και η διακύμανση ($E(Y_m - \mu_{Ym})$, η οποία είναι το τετράγωνο της τυπικής απόκλισης και συχνά ορίζεται ως $\sigma 2$).

Οι ροπές που περιγράφηκαν παραπάνω ορισμένες φορές αναφέρονται και ως αδέσμευτες ροπές· δηλαδή, ισχύουν για το σύνολο της κατανομής του πληθυσμού. Αλλά μπορούμε επίσης να δεσμεύσουμε την κατανομή και τις ροπές για συγκεκριμένες πληροφορίες. Για να γίνει καλύτερα κατανοητό, θεωρήστε το προσδόκιμο ζωής ενός άνδρα που ζει στο Ηνωμένο Βασίλειο. Ο Πίνακας 1.1 μας λέει ότι αυτό είναι τα 75 έτη. Επομένως, ποιο είναι το προσδόκιμο ζωής ενός άνδρα που ζει στο Ηνωμένο Βασίλειο και είναι ήδη 80 ετών; Προφανώς όχι τα 75 έτη! Μια αδέσμευτη ροπή είναι η ροπή για την υπό εξέταση πλήρη κατανομή· μια δεσμευμένη ροπή είναι η ροπή για εκείνα τα μέλη του πληθυσμού που πληρούν ορισμένες συνθήκες, στην προκειμένη περίπτωση να είναι 80 ετών. Μπορούμε να θεωρήσουμε έναν δεσμευμένο μέσο $E(Y_m | Yim = 80)$, σε αυτή την περίπτωση το μέσο των ανδρών που είναι 80, ή δεσμευμένες υψηλότερες ροπές, όπως η δεσμευμένη διακύμανση, η οποία θα αποτελέσει το αντικείμενο άλλου κεφαλαίου. Αυτός είναι άλλος ένας τρόπος σκέψης των υπο-ομάδων του πληθυσμού· θα μπορούσαμε να σκεφτούμε τον πληθυσμό ως αποτελούμενο από όλους τους ανθρώπους ή θα μπορούσαμε να σκεφτούμε την κατανομή του πληθυσμού των ανδρών και γυναικών ξεχωριστά. Αυτό που θα θέλαμε να γνωρίζουμε είναι η κατανομή του πληθυσμού για τον οποίο ενδιαφερόμαστε, δηλαδή το μέσο του προσδόκιμου ζωής όλων των ανδρών και γυναικών. Εάν μπορούσαμε να το μετρήσουμε, και πάλι δεν θα υπήρχε κάποιο στατιστικό θέμα για να αντιμετωπίσουμε· απλώς θα γνωρίζαμε εάν, κατά μέσο όρο, οι γυναίκες ζουν περισσότερο από τους άνδρες. Δυστυχώς, κατά κανόνα μπορούμε να έχουμε άμεσες μετρήσεις μόνο για ένα δείγμα που προέρχεται από τον πληθυσμό. Και θα πρέπει να χρησιμοποιήσουμε αυτό το δείγμα για να εξάγουμε κάποιο συμπέρασμα σχετικά με τον πληθυσμό.

Εάν το δείγμα χαρακτηρίζεται από ορισμένες βασικές ιδιότητες, τότε μπορούμε να προχωρήσουμε στην κατασκευή μιας μεθόδου για την εξαγωγή συμπερασμάτων. Η πρώτη βασική ιδέα είναι αυτή της τυχαίας δειγματοληψίας: τα άτομα που αποτελούν το δείγμα μας θα πρέπει να επιλέγονται τυχαία από τον πληθυσμό. Το προσδόκιμο ζωής ενός άνδρα είναι μια τυχαία μεταβλητή· δηλαδή, η ηλικία θανάτου ενός ατόμου είναι αβέβαιη. Μόλις παρατηρήσουμε την ηλικία θανάτου και η παρατήρηση γίνει μέρος του δείγματός μας παύει να είναι τυχαία μεταβλητή. Τα δεδομένα τότε αποτελούνται από ένα σύνολο μεμονομένων παρατηρήσεων, καθεμία από τις οποίες έχει επιλεχθεί τυχαία από τον πληθυσμό. Επομένως, το δείγμα των ηλικιών θανάτου των ανδρών γίνεται $Y_m = (Y1m, Y2m, \ldots, Y_{nm})$. Η ιδέα της τυχαίας δειγματοληψίας έχει ορισμένες ισχυρές συνέπειες:

Επειδή δύο οποιαδήποτε άτομα επιλέγονται τυχαία από τον πληθυσμό, θα πρέπει να είναι **ανεξάρτητα** το ένα από το άλλο· δηλαδή, το να γνωρίζουμε την ηλικία θανάτου του ενός άνδρα δεν μας λέει τίποτα για την ηλικία θανάτου του άλλου άνδρα. Επί-

σης, καθώς και τα δύο άτομα έχουν επιλεγεί από τον ίδιο πληθυσμό, θα πρέπει να έχουν **πανομοιότυπη κατανομή**. Επομένως, βασιζόμενοι στην υπόθεση της τυχαίας δειγματοληψίας, μπορούμε να ισχυριστούμε ότι καθεμία από τις παρατηρήσεις στο δείγμα μας θα πρέπει να έχει μια ανεξάρτητη και πανομοιότυπη κατανομή· αυτό συχνά εκφράζεται ως IID.

Είμαστε τώρα σε θέση να ξεκινήσουμε να κατασκευάζουμε ένα στατιστικό πλαίσιο. Θέλουμε να εξάγουμε ορισμένα συμπεράσματα για μια κατανομή πληθυσμού από την οποία έχει παρατηρηθεί μόνο ένα δείγμα. Πώς μπορούμε να γνωρίζουμε εάν η μέθοδος που επιλέξαμε για να αναλύσουμε το δείγμα είναι καλή ή όχι; Η απάντηση σε αυτή την ερώτηση βρίσκεται σε μια άλλη έννοια, η οποία ονομάζεται δειγματική κατανομή. Εάν επιλέξουμε ένα δείγμα από τον πληθυσμό μας, ας υποθέσουμε ότι έχουμε μια μέθοδο για να αναλύσουμε αυτό το δείγμα. Θα μπορούσε να είναι οτιδήποτε· για παράδειγμα, επιλέξτε τις παρατηρήσεις με μονή αρίθμηση, αθροίστε τες και διαιρέστε τες με το 20. Αυτό θα μας δώσει μια εκτίμηση. Εάν είχαμε ένα άλλο δείγμα, αυτό θα μας έδινε μια άλλη εκτίμηση και εάν συνεχίζαμε να επιλέγαμε δείγματα, θα είχαμε μια ολόκληρη σειρά από εκτιμήσεις βασισμένες σε αυτή την τεχνική. Θα μπορούσαμε τότε να εξετάσουμε την κατανομή όλων αυτών των εκτιμήσεων και αυτή θα ήταν η δειγματική κατανομή αυτής της συγκεκριμένης τεχνικής. Ας υποθέσουμε ότι η διαδικασία εκτίμησης παράγει μια εκτίμηση του πληθυσμιακού μέσου την οποία ονομάζουμε $\tilde{Y}_m$, τότε η δειγματική κατανομή θα έχει έναν μέσο και μια διακύμανση $E(\tilde{Y}_m)$ και $E(\tilde{Y}_m - E(\tilde{Y}_m))^2$· στην ουσία, η δειγματική κατανομή μιας συγκεκριμένης τεχνικής μάς λέει τα περισσότερα από αυτά που χρειάζεται να γνωρίζουμε σχετικά με την τεχνική. Μια καλή εκτιμήτρια γενικά θα χαρακτηρίζεται από **αμεροληψία**, που υποδηλώνει ότι η μέση τιμή της θα είναι ίση με το πληθυσμιακό χαρακτηριστικό που θέλουμε να εκτιμήσουμε. Δηλαδή, $E(\tilde{Y}_m) = \eta$, όπου η είναι το χαρακτηριστικό του πληθυσμού που θέλουμε να μετρήσουμε. Στην περίπτωση της αμεροληψίας, ακόμη και σε ένα μικρό δείγμα αναμένουμε η εκτιμήτρια να δίνει κατά μέσο όρο τη σωστή απάντηση. Μια ελαφρώς ασθενέστερη απαίτηση είναι η **συνέπεια**· εδώ αναμένουμε η εκτιμήτρια να δώσει τη σωστή απάντηση μόνο εάν έχουμε ένα απείρως μεγάλο δείγμα, $\lim_{n\to\infty} E(\tilde{Y}_m) = \eta$. Μια καλή εκτιμήτρια θα είναι είτε αμερόληπτη είτε συνεπής, αλλά μπορεί να υπάρχει και άλλη διαδικασία η οποία έχει αυτή την ιδιότητα. Στην περίπτωση αυτή μπορούμε να επιλέξουμε μεταξύ ενός αριθμού εκτιμητριών στη βάση της αποτελεσματικότητας· αυτή δίνεται απλά από τη διακύμανση της δειγματικής κατανομής. Ας υποθέσουμε ότι έχουμε μια άλλη τεχνική εκτίμησης, η οποία οδηγεί στην Y, η οποία είναι επίσης αμερόληπτη· τότε θα προτιμήσουμε την $\tilde{Y}$ από αυτή τη διαδικασία εάν $\text{var}(\tilde{Y}) < \text{var}(Y)$. Αυτό απλά σημαίνει ότι, κατά μέσο όρο, και οι δύο τεχνικές δίνουν τη σωστή απάντηση, αλλά τα λάθη που γίνονται με την πρώτη τεχνική είναι, κατά μέσο όρο, μικρότερα.

Ιδιότητες της δειγματικής κατανομής του μέσου

Στο παραπάνω παράδειγμα, με βάση τον Πίνακα 1.1, υπολογίσαμε το μέσο προσδόκι-

μο ζωής των ανδρών και των γυναικών. Γιατί αυτό είναι καλή ιδέα; Η απάντηση βρίσκεται στη δειγματική κατανομή του μέσου ως μια εκτίμηση του πληθυσμιακού μέσου. Η μέση τιμή της δειγματικής κατανομής του μέσου δίνεται από:

$$E\left(\frac{1}{n}\sum_{i=1}^{n}Y_i\right)=\frac{1}{n}\sum_{i=1}^{n}E\left(Y_i\right)=\frac{1}{n}\sum_{i=1}^{n}\mu_Y=\mu_Y \tag{1.4}$$

Επομένως, η αναμενόμενη τιμή του μέσου ενός δείγματος είναι ίση με τον πληθυσμιακό μέσο και έτσι η μέση τιμή ενός δείγματος είναι μια αμερόληπτη εκτιμήτρια της μέσης τιμής της κατανομής του πληθυσμού. Η μέση τιμή επομένως πληροί το πρώτο μας κριτήριο να είναι καλή εκτιμήτρια. Αλλά τι γίνεται με τη διακύμανση του μέσου;

$$var\left(\overline{Y}\right)=E\left(\overline{Y}-\mu_Y\right)^2=E\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{i=1}^{n}\left(Y_i-\mu_Y\right)\left(Y_i-\mu_Y\right)\right)$$

$$=\frac{1}{n^2}\left(\sum_{i=1}^{n}var\left(Y_i\right)+\sum_{i=1}^{n}\sum_{i=1,j\neq i}^{n}cov\left(Y_i,Y_j\right)\right)=\frac{\sigma_Y^2}{n} \tag{1.5}$$

Επομένως η διακύμανση του μέσου γύρω από τον πραγματικό μέσο του πληθυσμού σχετίζεται με το μέγεθος του δείγματος που χρησιμοποιείται για να κατασκευαστεί ο μέσος και τη διακύμανση της κατανομής του πληθυσμού. Καθώς το μέγεθος του δείγματος αυξάνεται, η διακύμανση στον πληθυσμό μειώνεται, το οποίο είναι αρκετά εύλογο, καθώς ένα μεγάλο μέγεθος δείγματος οδηγεί σε καλύτερη εκτίμηση του πληθυσμιακού μέσου. Εάν η πραγματική κατανομή του πληθυσμού έχει έναν μικρότερο μέσο, τότε η δειγματική κατανομή θα έχει και αυτή έναν μικρότερο μέσο. Και πάλι, αυτό είναι πολύ λογικό· εάν όλοι πεθαίναν ακριβώς στην ίδια ηλικία, τότε η διακύμανση του πληθυσμού θα ήταν μηδέν, και οποιοδήποτε δείγμα προερχόταν από τον πληθυσμό θα είχε έναν μέσο ακριβώς ίδιο με τον πραγματικό πληθυσμιακό μέσο.

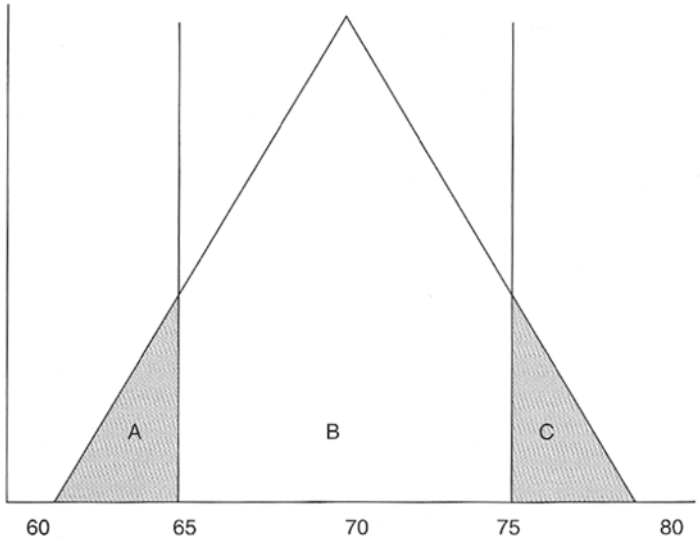
Έλεγχος υποθέσεων και κεντρικό οριακό θεώρημα

Φαίνεται ότι η μέση τιμή πληροί τα δύο κριτήριά μας για να είναι μια καλή εκτιμήτρια του πληθυσμού ως σύνολο: είναι αμερόληπτη και η αποτελεσματικότητά της αυξάνεται με το μέγεθος του δείγματος. Ωστόσο, για να μπορέσουμε να ελέγξουμε μια υπόθεση σχετικά με αυτή τη μέση τιμή, χρειαζόμαστε κάποια ιδέα για το σχήμα όλης της δειγματικής κατανομής. Δυστυχώς, ενώ έχουμε καθορίσει μια απλή έκφραση για τη μέση τιμή και τη διακύμανση, δεν είναι γενικά εφικτό να καθορίσουμε το σχήμα όλης της δειγματικής κατανομής. Ένας έλεγχος υποθέσεων κάνει μια υπόθεση σχετικά με την αλήθεια· αυτό το ονομάζουμε μηδενική υπόθεση και συχνά αναφέρεται ως H_0. Στη συνέχεια καθορίζουμε μια συγκεκριμένη εναλλακτική υπόθεση, η οποία συνήθως ονομάζεται H_1. Ο έλεγχος αποτελείται από τον υπολογισμό της πιθανότητας ότι η παρατηρούμενη τιμή του στατιστικού θα μπορούσε να έχει εμφανιστεί τελείως τυχαία, υποθέτοντας ότι

η μηδενική υπόθεση είναι αληθής. Ας υποθέσουμε ότι η μηδενική μας υπόθεση είναι ότι ο πραγματικός πληθυσμιακός μέσος για την ηλικία θανάτου των ανδρών είναι 70, $H_0 : E(\bar{Y}_m) = 70$. Έχοντας παρατηρήσει έναν μέσο ίσο με 75,1, θα μπορούσαμε τότε να ελέγξουμε την εναλλακτική υπόθεση ότι είναι μεγαλύτερος του 70. Θα μπορούσαμε να το κάνουμε αυτό υπολογίζοντας την πιθανότητα ότι το 75,1 θα μπορούσε να εμφανιστεί τυχαία ενώ η πραγματική τιμή του πληθυσμιακού μέσου είναι 70. Με μια συνεχή κατανομή η πιθανότητα να προκύψει ενά ακριβές σημείο είναι μηδέν, επομένως αυτό που υπολογίζουμε ακριβώς είναι η πιθανότητα σχηματισμού μιας οποιασδήποτε τιμής για τον μέσο που είναι μεγαλύτερη από 75,1. Μπορούμε τότε να συγκρίνουμε αυτή την πιθανότητα με μια προκαθορισμένη τιμή, την οποία ονομάζουμε επίπεδο σημαντικότητας του ελέγχου. Εάν η πιθανότητα είναι μικρότερη από το επίπεδο σημαντικότητας, απορρίπτουμε τη μηδενική υπόθεση έναντι της εναλλακτικής. Στην παραδοσιακή στατιστική το επίπεδο σημαντικότητας ορίζεται συνήθως στο 1%, 5% ή 10%. Εάν χρησιμοποιούσαμε ένα 5% επίπεδο σημαντικότητας και βρίσκαμε ότι η πιθανότητα παρατήρησης ενός μέσου μεγαλύτερου από 75,1 ήταν 0,01, καθώς 0,01 < 0,05, θα απορρίπταμε την υπόθεση ότι η πραγματική τιμή του πληθυσμιακού μέσου είναι 70 έναντι της εναλλακτικής ότι είναι μεγαλύτερη από 70.

Η εναλλακτική υπόθεση μπορεί να προσδιοριστεί τυπικά με δύο τρόπους, που οδηγούν είτε σε έναν μονόπλευρο είτε σε έναν δίπλευρο έλεγχο. Το παραπάνω παράδειγμα είναι ένας μονόπλευρος έλεγχος, καθώς η εναλλακτική ήταν ότι η ηλικία θανάτου είναι μεγαλύτερη από 70, αλλά θα μπορούσαμε εξίσου να ελέγξουμε την πιθανότητα ότι η πραγματική μέση τιμή είναι είτε μεγαλύτερη είτε μικρότερη από 70, όπου στην περίπτωση αυτή θα είχαμε διεξαγάγει έναν δίπλευρο έλεγχο. Στη περίπτωση ενός δίπλευρου ελέγχου, θα υπολογίζαμε την πιθανότητα ότι μια τιμή είτε μεγαλύτερη από 75,1 ή μικρότερη από 70 – (75,1 – 70) = 64,9 θα προέκυπτε τυχαία. Προφανώς αυτή η πιθανότητα θα ήταν μεγαλύτερη απ' ό,τι σε έναν μονόπλευρο έλεγχο.

Το Σχήμα 1.1 δείχνει τη βασική ιδέα του ελέγχου υποθέσεων. Απεικονίζει μία πιθανή δειγματική κατανομή για το μέσο προσδόκιμο ζωής των ανδρών υπό τη μηδενική υπόθεση ότι ο μέσος του πληθυσμού είναι 70. Είναι ένα απίθανο σχήμα, ουσιαστικά είναι ένα τρίγωνο, αλλά θα το συζητήσουμε αργότερα· προς το παρόν, ας υποθέσουμε ότι αυτό είναι το σχήμα της κατανομής. Εξ ορισμού, όλη η περιοχή κάτω από το τρίγωνο αθροίζει σε 1. Αυτό σημαίνει ότι με πιθανότητα 1 (βεβαιότητα) η μέση τιμή θα βρίσκεται μεταξύ 62 και 78 και ότι συγκεντρώνεται γύρω από το 70. Στην πραγματικότητα, παρατηρούμε μία μέση τιμή της τάξεως του 75,1, και εάν θέλουμε να ελέγξουμε την υπόθεση ότι η πραγματική μέση τιμή είναι 70 έναντι της εναλλακτικής ότι είναι μεγαλύτερη από 70 (μονόπλευρος έλεγχος), υπολογίζουμε την πιθανότητα να παρατρηρήσουμε μια τιμή ίση με 75,1 ή μεγαλύτερη. Αυτή δίνεται από την περιοχή Γ στο σχήμα. Εάν θέλαμε να διεξάγουμε έναν δίπλευροο έλεγχο, ότι η εναλλακτική είναι είτε για μεγαλύτερη από 75,1 είτε για μικρότερη από 64,9, θα υπολογίζαμε το άθροισμα των περιοχών Α και Γ, το οποίο είναι προφανώς μεγαλύτερο από Γ. Εάν υιοθετούσαμε μια 5% κριτική τιμή και εάν Γ < 0,05, θα απορρίπταμε τη μηδενική υπόθεση σε έναν μονόπλευρο έλεγχο. Εάν Γ + Α < 0,05, θα απορρίπταμε τη μηδενική υπόθεση σε ένα 5% επίπεδο στον δίπλευρο έλεγχο.



Σχήμα 1.1 Μια πιθανή κατανομή για το προσδόκιμο ζωής

Όπως σημειώθηκε παραπάνω, ενώ έχουμε υπολογίσει τη μέση τιμή και τη διακύμανση για τη δειγματική κατανομή στην περίπτωση του μέσου, δεν είναι γενικά δυνατό να υπολογίσουμε το σχήμα της συνολικής κατανομής. Ωστόσο, υπάρχει ένα αξιοσημείωτο θεώρημα το οποίο γενικά μας επιτρέπει να το κάνουμε αυτό καθώς το μέγεθος του δείγματος αυξάνεται. Αυτό είναι το κεντρικό οριακό θεώρημα.

Κεντρικό οριακό θεώρημα

Εάν ένα σύνολο δεδομένων είναι IID με n παρατηρήσεις, (Y_1, Y_2, ... Y_n), και με πεπερασμένη διακύμανση, τότε καθώς το n τείνει προς το άπειρο, η κατανομή του $\bar{Y}$ γίνεται κανονική. Επομένως, όσο το n είναι αρκετά μεγάλο μπορούμε να σκεφτούμε την κατανομή της μέσης τιμής ότι προσεγγίζει την κανονική.

Αυτό είναι ένα αξιοσημείωτο αποτέλεσμα· αυτό που λέει είναι ότι, ανεξάρτητα από το σχήμα της κατανομής του πληθυσμού, η δειγματική κατανομή θα είναι κανονική εφόσον βασίζεται σε ένα αρκετά μεγάλο δείγμα. Για να πάρουμε ένα ακραίο παράδειγμα, ας υποθέσουμε ότι σκεφτόμαστε μια λαχειοφόρο αγορά η οποία κληρώνει έναν νικηφόρο λαχνό για κάθε 100 λαχνούς που πωλούνται. Εάν το βραβείο για έναν νικηφόρο λαχνό είναι \$100 και το κόστος για κάθε λαχνό είναι \$1, τότε, κατά μέσο όρο, θα αναμένουμε να κερδίσουμε \$1 για κάθε λαχνό που αγοράζουμε. Αλλά η κατανομή του πληθυσμού θα φαινόταν πολύ περίεργη· 99 στους 100 λαχνούς θα είχαν απόδοση μηδέν και ένας λαχνός θα είχε απόδοση \$100. Εάν προσπαθούσαμε να σχεδιάσουμε την κατανομή των αποδόσεων, θα είχε μια τεράστια ακίδα στο μηδέν, μια μικρή ακίδα στα \$100 και καθόλου παρατηρήσεις οπουδήποτε αλλού. Αλλά, όσο επιλέγουμε ένα

αρκετά μεγάλο δείγμα, όταν υπολογίζουμε τη μέση απόδοση πάνω στο δείγμα θα έχει κέντρο στο $1 με μια κανονική κατανομή γύρω από το 1.

Η σημασία του κεντρικού οριακού θεωρήματος είναι ότι μας επιτρέπει να γνωρίσουμε πως θα πρέπει να μοιάζει η δειγματική κατανομή της μέσης τιμής, όσο η μέση τιμή βασίζεται σε ένα αρκετά μεγάλο δείγμα. Επομένως, τώρα μπορούμε να αντικαταστήσουμε την αυθαίρετη τριγωνική κατανομή στο Σχήμα 1.1 με μια πιο λογική, την κανονική κατανομή.

Ένα τελικό μέρος του στατιστικού μας πλαισίου είναι ο **κανόνας των μεγάλων αριθμών**, ο οποίος δηλώνει ότι εάν ένα δείγμα (Y1 , Y2 , ... Y*n*) είναι IID με μια πεπερασμένη διακύμανση, τότε η $\bar{Y}$ είναι μια συνεπής εκτιμήτρια του μ, της πραγματικής μέσης τιμής του πληθυσμού. Αυτό μπορεί να δηλωθεί επίσημα ως $\Pr(|\bar{Y} - \mu| < \varepsilon) \to 1$ καθώς $n \to \infty$, που σημαίνει ότι η πιθανότητα η απόλυτη διαφορά μεταξύ της εκτίμησης του μέσου και της πραγματικής μέσης τιμής του πληθυσμού να είναι μικρότερη από έναν μικρό θετικό αριθμό που τείνει στη μονάδα καθώς το μέγεθος του δείγματος τείνει στο άπειρο. Αυτό μπορεί να αποδειχθεί άμεσα καθώς, όπως έχουμε δει, η διακύμανση της δειγματικής κατανομής του μέσου είναι αντιστρόφως ανάλογη του n· επομένως, καθώς το n τείνει στο άπειρο η διακύμανση της δειγματικής κατανομής τείνει στο μηδέν και η μέση τιμή τείνει προς την πραγματική τιμή του πληθυσμού.

Μπορούμε τώρα να συνοψίζουμε: η $\bar{Y}$ είναι μια αμερόληπτη και συνεπής εκτιμήτρια της πραγματικής μέσης τιμής του πληθυσμού μ· κατανέμεται κατά προσέγγιση ως μια κανονική κατανομή με μια διακύμανση η οποία είναι αντιστρόφως ανάλογη του n· αυτό μπορεί να εκφραστεί ως $N(\mu, \sigma^2/n)$. Επομένως, εάν αφαιρέσουμε τη μέση τιμή του πληθυσμού από το $\bar{Y}$ και διαιρέσουμε με την τυπική απόκλιση, θα δημιουργήσουμε μια μεταβλητή η οποία έχει μέση τιμή μηδέν και διακύμανση ίση με τη μονάδα. Αυτό ονομάζεται τυποποίηση της μεταβλητής.

$$\frac{\bar{Y} - \mu}{\sqrt{\sigma^2 / n}} \sim N(0,1) \tag{1.6}$$

Ένα μικρό πρόβλημα με αυτή την εξίσωση, ωστόσο, είναι ότι περιλαμβάνει το σ^2. Αυτή είναι η διακύμανση του πληθυσμού, η οποία είναι άγνωστη, και χρειάζεται να εξάγουμε μια εκτίμηση από αυτή. Μπορούμε να εκτιμήσουμε τη διακύμανση του πληθυσμού με:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \tag{1.7}$$

Εδώ διαιρούμε με το $n - 1$ διότι χάνουμε στην ουσία μια παρατήρηση όταν εκτιμούμε τη μέση τιμή. Σκεφτείτε τι γίνεται όταν έχουμε ένα δείγμα μεγέθους ένα. Η εκτίμηση της μέσης τιμής θα είναι ίδια με αυτή τη μία παρατήρηση και εάν διαιρούσαμε με το $n = 1$ θα εκτιμούσαμε μια διακύμανση ίση με μηδέν. Διαιρώντας με το $n - 1$ η διακύμανση είναι απροσδιόριστη για ένα δείγμα μεγέθους ένα. Γιατί είναι η $s2$ μια καλή εκτιμήτρια της διακύμανσης του πληθυσμού; Η απάντηση είναι ότι είναι ουσιαστικά απλά ένας άλλος μέσος· επομένως, ο κανόνας των μεγάλων αριθμών εφαρμόζεται και θα είναι μια συνεπής εκτίμηση της πραγματικής διακύμανσης του πληθυσμού.

Τώρα είμαστε πλέον σε θέση να κατασκευάσουμε έναν τυπικό έλεγχο υποθέσεων. Ο βασικός έλεγχος είναι γνωστός ως έλεγχος «*t*» student και δίνεται από:

$$t = \frac{\bar{Y} - \mu}{\sqrt{S^2 / n}} \tag{1.8}$$

Όταν το δείγμα είναι μικρό, αυτό θα ακολουθεί μια κατανομή *t*-student, η οποία μπορεί να βρεθεί σε ένα οποιοδήποτε τυπικό σύνολο από στατιστικούς πίνακες. Στην πράξη, ωστόσο, μόλις το δείγμα είναι μεγαλύτερο από 30 ή 40, η κατανομή *t* είναι σχεδόν ταυτόσημη με την τυποποιημένη κανονική κατανομή και στην οικονομετρία είναι κοινή πρακτική να χρησιμοποιείται η κανονική κατανομή. Η τιμή της κανονικής κατανομής που συνεπάγεται 0,025 σε κάθε ουρά της κατανομής είναι 1,96. Αυτή είναι η κρίσιμη τιμή η οποία ισχύει για έναν αμφίπλευρο έλεγχο σε ένα 5% επίπεδο σημαντικότητας. Επομένως, εάν θέλουμε να ελέγξουμε την υπόθεση ότι η εκτίμησή μας 75,1 του προσδόκιμου ζωής των ανδρών είναι στην πραγματικότητα μια τυχαία λήψη από έναν πληθυσμό με μέση τιμή 70, τότε ο έλεγχος θα είναι:

$$t = \frac{75.1 - 70}{\sqrt{S^2 / 3.87}} = \frac{5.1}{.355} = 14.37$$

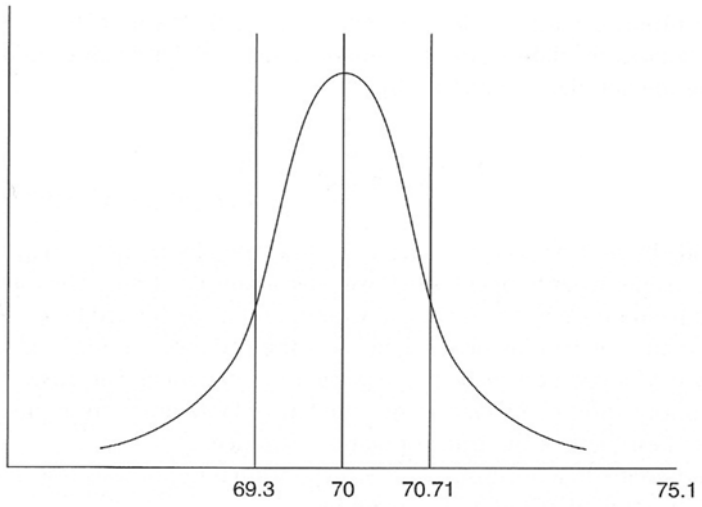
Αυτό είναι μεγαλύτερο από το 1,96 του 5% επιπέδου σημαντικότητας, και επομένως θα απορρίψουμε τη μηδενική υπόθεση ότι η πραγματική μέση τιμή του πληθυσμού είναι 70. Αντίστοιχα, θα μπορούσαμε να εκτιμήσουμε την αναλογία της κατανομής που σχετίζεται με μια απόλυτη τιμή *t* μεγαλύτερη από 4,1, η οποία τότε θα ήταν η πιθανότητα που συζητήθηκε παραπάνω. Τυπικά η πιθανότητα ή *p*-value δίνεται από:

$$p - value = Pr_{Ho}\left(\left|Y - \mu\right| > \left|\bar{Y}^{act} - \mu\right|\right) = Pr_{Ho}\left(\left|t\right| > \left|t^{act}\right|\right)$$

Επομένως, εάν η τιμή *t* είναι ακριβώς 1,96, η *p*-value θα είναι 0,05, και όταν η τιμή *t* είναι μεγαλύτερη από 1,96, τότε η *p*-value θα είναι μικρότερη από 0,05. Περιέχουν ακριβώς την ίδια πληροφορία, αλλά εκφρασμένη με διαφορετικό τρόπο. Η *p*-value είναι, ωστόσο, χρήσιμη σε άλλες περιπτώσεις, καθώς μπορεί να υπολογιστεί για ένα εύρος διαφορετικών κατανομών χωρίς τη βοήθεια από στατιστικούς πινάκες, καθώς η ερμηνεία της είναι πάντα άμεση.

Το Σχήμα 1.2 παρουσιάζει αυτή τη διαδικασία. Δείχνει μια κατά προσέγγιση κανονική κατανομή με κέντρο τη μηδενική υπόθεση και τις δύο ουρές της κατανομή οριζόμενες από 69,3 και 70,71. 95% της περιοχής κάτω από την κατανομή βρίσκεται μεταξύ αυτών των δύο σημείων. Η εκτιμώμενη τιμή 75,1 βρίσκεται αρκετά έξω από αυτή την κεντρική περιοχή και επομένως μπορούμε να απορρίψουμε τη μηδενική υπόθεση ότι η πραγματική τιμή είναι 70 και ότι παρατηρήσαμε το 75,1 καθαρά από τύχη. Η *p*-value είναι δύο φορές η περιοχή κάτω από την καμπύλη η οποία βρίσκεται πέρα από το 75,1, και προφανώς αυτή είναι πράγματι πολύ μικρή.

Ένας τελευταίος τρόπος για να σκεφτούμε όσον αφορά την εμπιστοσύνη που έχουμε στην εκτίμησή μας είναι να κατασκευάσουμε ένα διάστημα εμπιστοσύνης γύρω από



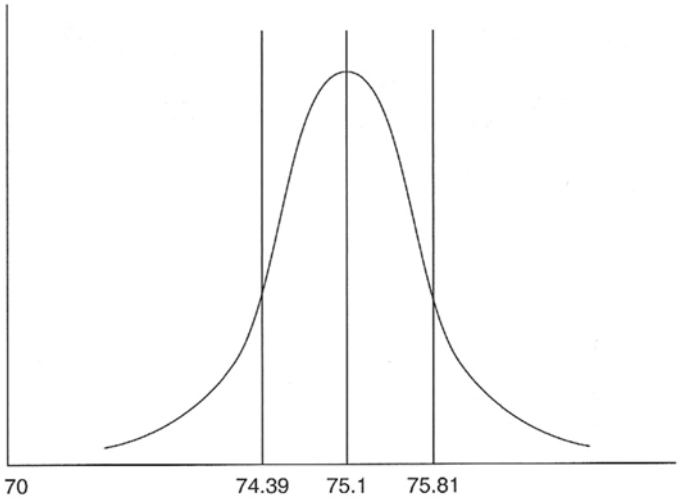
Σχήμα 1.2 Μια κανονική κατανομή για το προσδόκιμο ζωής γύρω από τη μηδενική υπόθεση

την εκτιμώμενη παράμετρο. Έχουμε μια εκτιμώμενη μέση τιμή της τάξεως του 75,1, αλλά γνωρίζουμε ότι υπάρχει ορισμένη αβεβαιότητα ως προς το ποια είναι η πραγματική τιμή. Ο κανόνας των μεγάλων αριθμών μάς λέει ότι αυτή είναι μια συνεπής εκτίμηση της πραγματικής τιμής, επομένως με μόνο αυτή την παρατήρηση η καλύτερη πρόβλεψή μας είναι ότι η πραγματική τιμή είναι 75,1. Το κεντρικό οριακό θεώρημα μας λέει ότι η κατανομή γύρω από την τιμή αυτή είναι κατά προσέγγιση κανονική, και γνωρίζουμε τη διακύμανση αυτής της κατανομής. Επομένως, μπορούμε να κατασκευάσουμε ένα διάστημα γύρω από το 75,1 το οποίο θα περιέχει οποιαδήποτε απαιτούμενη ποσότητα της κατανομής. Το σύνηθες και πάλι είναι να χρησιμοποιήσουμε ένα 95% διάστημα εμπιστοσύνης, και αυτό μπορεί να γίνει ως εξής:

$$CI_{95\%} = \left\{ \bar{Y} + 1.96\frac{s}{\sqrt{n}}, \bar{Y} - 1.96\frac{s}{\sqrt{n}} \right\} = \bar{Y} + 0,71, \bar{Y} - 0.71$$

Επομένως, με 95% βεβαιότητα μπορούμε να πούμε ότι η πραγματική μέση τιμή βρίσκεται μεταξύ 74,39 και 75,81. Αυτό φαίνεται στο Σχήμα 1.3· εδώ η εικόνα έχει μετακινηθεί έτσι ώστε τώρα το κέντρο του να βρίσκεται στην εκτιμώμενη τιμή του 75,1 και το 95% του σχήματος βρίσκεται μέσα στο διάστημα εμπιστοσύνης. Προφανώς, η μηδενική τιμή 70 βρίσκεται αρκετά έξω από αυτή την περιοχή, και επομένως μπορούμε και πάλι να συμπεράνουμε ότι η πραγματική τιμή του μέσου είναι εξαιρετικά απίθανο να είναι 70.

Το ίδιο συμπέρασμα προκύπτει από τον υπολογισμό του τυπικού έλεγχου *t* ή της *p*-value ή λαμβάνοντας υπόψη το διάστημα εμπιστοσύνης, διότι όλα είναι διαφορετικοί τρόποι έκφρασης της ίδιας υποκείμενης κατανομής.



Σχήμα 1.3 Ένα 95% διάστημα εμπιστοσύνης γύρω από τον εκτιμώμενο μέσο

Συμπέρασμα

Σε αυτό το κεφάλαιο έχουμε περιγράψει τα βασικά βήματα για την κατασκευή μιας θεωρίας εκτίμησης και ελέγχου υποθέσεων. Ξεκινήσαμε από την τυχαία δειγματοληψία, η οποία έδωσε αφορμή για την πρόταση ότι τα στοιχεία ενός δείγματος θα έχουν μια κατανομή IID. Από αυτήν ήμασταν σε θέση να ορίσουμε μια κατανομή του πληθυσμού και να εξάγουμε ορισμένα συμπεράσματα σχετικά με την κατανομή αυτή κατασκευάζοντας τον μέσο και μετά ορίζοντας τη δειγματική κατανομή του μέσου. Με τη χρήση του κανόνα των μεγάλων αριθμών και του κεντρικού οριακού θεωρήματος, προσδιορίσαμε το σχήμα της δειγματικής κατανομής και, τέλος, με αυτό το δεδομένο, μπορέσαμε να περιγράψουμε τη βασική διαδικασία ελέγχου που χρησιμοποιείται στην κλασική οικονομετρία.

Ενώ εκ πρώτης όψεως αυτό μπορεί να φαίνεται ότι σχετίζεται συγκεκριμένα με μια απλή διαδικασία εκτίμησης, της μέσης τιμής, τα ίδια βήματα μπορούν να εφαρμοστούν σχεδόν σε οποιαδήποτε διαδικασία εκτίμησης, όπως θα δούμε στα επόμενα κεφάλαια. Επομένως, όταν εκτιμούμε μια παράμετρο σε ένα μοντέλο από ένα σύνολο δεδομένων, ουσιαστικά ακολουθούμε τα ίδια βήματα. Οποιαδήποτε διαδικασία εκτίμησης είναι ουσιαστικά η λήψη ενός δείγματος δεδομένων και ο υπολογισμός του μέσου όρου με κάποιον τρόπο. Έχουμε μια δειγματική κατανομή για την παράμετρο και μπορούμε να διερευνήσουμε την αμεροληψία και τη συνέπεια της διαδικασίας εκτίμησης. Μπορούμε να συνεχίσουμε εφαρμόζοντας το κεντρικό οριακό θεώρημα, το οποίο θα αποδείξει ότι

αυτή η δειγματική κατανομή τείνει σε μια κανονική κατανομή, καθώς το μέγεθος του δείγματος μεγαλώνει. Τέλος, μπορούμε να χρησιμοποιήσουμε αυτό το αποτέλεσμα για να κατασκευάσουμε ελέγχους υποθέσεων σχετικά με τις παραμέτρους που έχουν εκτιμηθεί και να υπολογίσουμε *p*-values και διαστήματα εμπιστοσύνης.