**Α.Τ.Ε.Ι. ΠΑΤΡΑΣ**

**ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ**

**ΤΜΗΜΑ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΥ ΣΧΕΔΙΑΣΜΟΥ**
**ΚΑΙ**
**ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ**

**ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ**

# ΕΚΤΙΜΗΣΗ ΣΗΜΕΙΟΥ ΚΑΙ ΔΙΑΣΤΗΜΑΤΟΣ ΕΜΠΙΣΤΟΣΥΝΗΣ ΠΛΗΘΥΣΜΙΑΚΩΝ ΠΑΡΑΜΕΤΡΩΝ

**ΚΑΠΩΝΗΣ ΝΙΚΟΛΑΟΣ ΑΜ.1211**

**ΠΛΑΤΑΝΙΑΣ ΓΕΩΡΓΙΟΣ ΑΜ.1316**

**ΕΙΣΗΓΗΤΡΙΑ: ΑΘΑΝΑΣΙΑ ΜΠΟΥΜΠΟΥΛΗ**

**ΠΑΤΡΑ 2009**

# Περιεχόμενα

## Περίληψη

Η Επαγωγική Στατιστική είναι το μέρος εκείνο της Στατιστικής που ασχολείται με τις μεθόδους που καθιστούν δυνατή τη γενίκευση των δειγματοληπτικών συμπερασμάτων στον γεννήτορα πληθυσμό, από τον οποίο πήραμε το δείγμα και μας βοηθά στη λήψη σωστών αποφάσεων σε συνθήκες αβεβαιότητας.

Τα σύνολα των αντικειμένων που εξετάζονται εκφράζουν πραγματικές καταστάσεις και η μελέτη τους γίνεται με τη βοήθεια υποδειγμάτων πιθανότητας. Η μεθοδολογική στρατηγική που ακολουθείται στο κρίσιμο θέμα της επαγωγής εξαρτάται από το είδος των πληροφοριών που θεωρούνται σχετικές με το φαινόμενο που διερευνάται. Ανάλογα με τη φύση των πληροφοριών που πρόκειται να επεξεργαστούν, διακρίνουμε δυο βασικές προσεγγίσεις στο θέμα της σταστιστικής επαγωγής: την **κλασική** και την **μπεϋζιανή**.

Στο χώρο της κλασικής στατιστικής επαγωγής ανήκουν δύο συναφή προβλήματα: η εκτίμηση των παραμέτρων του πληθυσμού ( ή του μοντέλου ), και ο έλεγχος υποθέσεων, ο οποίος περιλαμβάνει τόσο τον έλεγχο της αξιοπιστίας των παραμέτρων του πληθυσμού, όσο και τον έλεγχο διαφόρων άλλων υποθέσεων. Στην περίπτωση της εκτίμησης το ενδιαφέρον επικεντρώνεται στην επιλογή (εκτίμηση) της κατάλληλης τιμής για μια ή περισσότερες παραμέτρους του πληθυσμού.

Η εκτίμηση μιας παραμέτρου του πληθυσμού μπορεί να δοθεί με έναν αριθμό. Έχουμε τότε μια σημειακή εκτίμηση της παραμέτρου. Η εκτίμηση μιας παραμέτρου του πληθυσμού μπορεί να δοθεί και με ένα διάστημα, μέσα στο

όποιο πιστεύουμε ότι περιέχεται η παράμετρος του πληθυσμού. Έχουμε τότε μια εκτίμηση διαστήματος της παραμέτρου. Μια πληροφορία για το σφάλμα ή την ακρίβεια της εκτιμήσεως λέμε ότι περιγράφει την αξιοπιστία της εκτιμήσεως.
Στην εργασία αυτή θα παρουσιάσουμε βασικές έννοιες και μεθόδους της εκτιμητικής, θα αναλύσουμε τη σημειακή εκτίμηση, καθώς και τα είδη των διαστημάτων εμπιστοσύνης.

Πιο συγκεκριμένα:

- Στο πρώτο Κεφάλαιο παρουσιάζουμε τις βασικές έννοιες του πεδίου της Στατιστικής, οι οποίες θα μας φανούν πολύ χρήσιμες μιας και αποτελούν λέξεις-κλειδιά για τον προσδιορισμό εννοιών της εκτιμητικής.
- Στο δεύτερο Κεφάλαιο παρουσιάζουμε την Επαγωγική Στατιστική και αναπτύσουμε τις μεθόδους σημειακής εκτίμησης.
- Στο τρίτο Κεφάλαιο αναλύουμε την έννοια του διαστήματος εμπιστοσύνης, παρουσιάζουμε τα είδη που υπάρχουν και κάνουμε ανάλυση του καθενός με τη χρήση παραδειγμάτων.
- Τέλος, το Κεφάλαιο 4 αποτελεί μια πρακτική στατιστική ανάλυση δεδομένων, όπου εφαρμόζουμε με τη βοήθεια του στατιστικού πακέτου SPSS όλα όσα αναφέρουμε στα τρία πρώτα κεφάλαια.

# ΚΕΦΑΛΑΙΟ 1

# Βασικές Έννοιες Στατιστικής

## *1.1. Βασικοί Ορισμοί*

**1. Στατιστική**: είναι ο κλάδος των εφαρμοσμένων μαθηματικών, ο οποίος βασίζεται σε ένα σύνολο αρχών και μεθοδολογιών για:

- Τον σχεδιασμό της διαδικασίας συλλογής δεδομένων (data).
- Τη συνοπτική και αποτελεσματική παρουσίαση των δεδομένων.
- Την ανάλυση και εξαγωγή συμπερασμάτων από τα δεδομένα.

Η στατιστική είναι μια ατελής επαγωγή. Από τις ιδιότητες του μέρους εξάγει συμπεράσματα για το όλον.

**2. Πληθυσμός (Population - Sample Space):** είναι ένα σύνολο, τα στοιχεία του οποίου εξετάζονται ως προς ένα ή περισσότερα χαρακτηριστικά τους.

**3. Μεταβλητή (Variable):** ενός πληθυσμού είναι το χαρακτηριστικό ως προς το οποίο εξετάζεται ο πληθυσμός. Μεταβλητές είναι δύο ειδών, ποιοτικές ή ποσοτικές. Οι ποσοτικές μεταβλητές διακρίνονται σε διακριτές και συνεχείς.

**4. Συχνότητα (Frequency):** μιάς τιμής $\chi_\iota$ της μεταβλητής Χ μεγέθους νΧΝ είναι ο φυσικός αριθμός $\nu_\iota$ΧΝ, $\nu_\iota \leq \nu$, που δείχνει το πλήθος εμφάνισης της τιμής $\chi_\iota$ .

Ισχύει προφανώς: $\nu_1+\nu_2+ . . . +\nu_\mu=\nu$

**5. Σχετική συχνότητα :** μιας τιμής $\chi_\iota$ της μεταβλητής Χ είναι ο αριθμός

$f_i = \frac{\nu_i}{\nu}$ (1) και i = 1,2, . . . , μ όπου μ ≤ ν

Ισχύει προφανώς $f_1 + f_2 + ... + f_\mu = 1$ (2)

Η σχετική συχνότητα εκφράζεται και επί τοις εκατό, οπότε

**$f_i$% = 100. $f_i$** (3)

**6. Αθροιστική συχνότητα $N_i$**: είναι το πλήθος των παρατηρήσεων που είναι μικρότερες ή ίσες της τιμής $\chi_i$.
Ομοίως και σχετική αθροιστική συχνότητα $F_i$.

**7. Ο συμβολισμός $\sum_{i=1}^{\nu} x_i$ :**

$$\sum_{i=1}^{\nu} x_i = x_1 + x_2 + ... + x_\nu \quad (4)$$

Ιδιότητες:

1. $\sum_{i=1}^{\nu} (\lambda .x_i) = \lambda .x_1 + \lambda .x_2 + ... + \lambda .x_\nu = \lambda .\sum_{i=1}^{\nu} x_i$ (5)

2. $\sum_{i=1}^{\nu} (x_i + y_i) = (x_1 + y_1) + (x_2 + y_2) + ... + (x_\nu + y_\nu) = \sum_{i=1}^{\nu} x_i + \sum_{i=1}^{\nu} y_i$ (6)

είναι φανερό ότι ισχύει εν γένει $\sum_{i=1}^{n} x_i^2 \neq (\sum_{i=1}^{n} x_i)^2$ (7)

**8. Μέτρα θέσης:**

**i. Αριθμητικός μέσος ή μέση τιμή (First moment or Mean or Average):**

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_\nu}{\nu} = \frac{1}{\nu} \cdot \sum_{i=1}^{\nu} x_i \quad (8)$$

εάν οι τιμές $x_i$ έχουν συχνότητες $\nu_i$, i=1, 2, . . . , μ τότε η μέση τιμή δίνεται από τον τύπο:

$$\bar{x} = \frac{\nu_1 \cdot x_1 + \nu_2 \cdot x_2 + \ldots + \nu_\mu \cdot x_\mu}{\nu} = \frac{1}{\nu} \cdot \sum_{i=1}^{\mu} \nu_i \cdot x_i = \sum_{i=1}^{\mu} x_i \cdot f_i \quad (9)$$

δοθέντος ότι $f_i = \frac{\nu_i}{\nu}$

Εύκολα αποδεικνύεται ότι εάν στις τιμές της μεταβλητής *X* προσθέσουμε έναν αριθμό *α*X**R** τότε και η μέση τιμή μεταβάλλεται κατά *α*.

Δηλαδή $\overline{X + a} = \bar{X} + a$ (10) με *α*X**R**.

Επίσης εάν οι τιμές της *X* πολλαπλασιασθούν επί *κ*X**R** τότε και η μέση τιμή πολλαπλασιάζεται επί *κ*.

Δηλαδή $\overline{\kappa \cdot X} = \kappa \cdot \bar{X}$ (11) με *κ*X**R.**

**ii. Διάμεσος δ (Median):** ενός δείγματος παρατηρήσεων, οι οποίες έχουν διαταχθεί κατά αύξουσα τάξη, είναι η μεσαία παρατήρηση εάν το πλήθος των παρατηρήσεων είναι περιττό ή ο μέσος όρος των δύο μεσαίων παρατηρήσεων εάν το πλήθος είναι άρτιο.

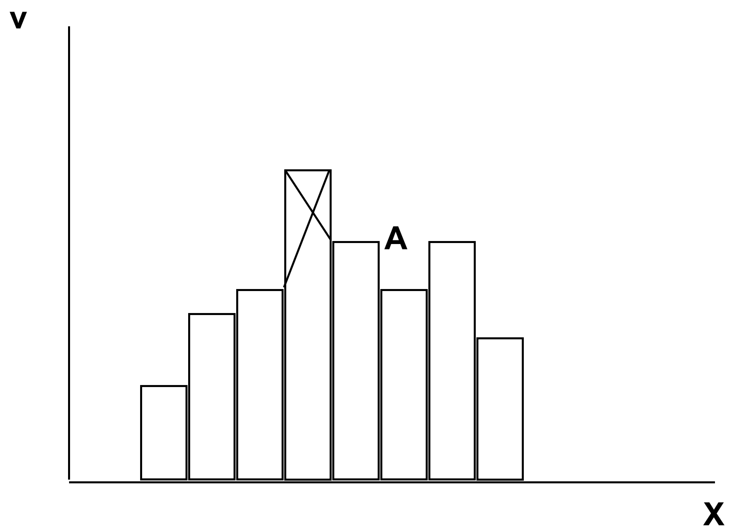
Στην περίπτωση ομαδοποιημένων μεταβλητών η διάμεσος βρίσκεται από το ιστόγραμμα των αθροιστικών συχνοτήτων. Αλγεβρικά η διάμεσος δίνεται από τον τύπο:

$$\delta = l_i + \frac{\frac{\nu}{2} - N_{i-1}}{\nu_i} \cdot c_i \quad (12)$$

όπου $l_i$ το κατώτερο όριο της κλάσης που περιέχει την διάμεσο, $v_i$ η συχνότητα και $c_i$ το πλάτος της κλάσης αντίστοιχα, $N_{i-1}$ η αθροιστική συχνότητα της προηγούμενης κλάσης και ν το μέγεθος του δείγματος.

**iii. Επικρατούσα τιμή $M_0$ (Mode):** ορίζεται ως η παρατήρηση με την μεγαλύτερη συχνότητα.

Εάν έχουμε ομαδοποιημένη μεταβλητή τότε είναι το σημείο τομής των ευθυγράμμων τμημάτων τα οποία ορίζονται από α) ΑΒ, όπου Α το άνω δεξί άκρο της κλάσης με τη μεγαλύτερη συχνότητα και Β το άνω δεξί άκρο της προηγούμενης κλάσης. β) ΓΔ, όπου Γ το άνω αριστερό άκρο της κλάσης με τη μεγαλύτερη συχνότητα και Δ το άνω αριστερό άκρο της επόμενης κλάσης.




**Σχήμα 1**

## 9. Μέτρα Διασποράς.

**i. Εύρος R:** είναι η διαφορά μεταξύ της ελαχίστης παρατήρησης από την μεγίστη. Δηλαδή

**R = $X_{max}$ – $X_{min}$** (13)

**ii. Διακύμανση ή Διασπορά $s^2$ (Second Moment or Variation):**

ορίζεται από την σχέση $s^2 = \frac{1}{\nu}\sum_{i=1}^{\nu}(x_i - \bar{x})^2$ (14)

ο τύπος αυτός μπορεί να μετασχηματισθεί κάνοντας τις πράξεις και ως εξής:

$$s^2 = \frac{1}{\nu}\left[\sum_{i=1}^{\nu} x_i^2 - \frac{1}{\nu}\cdot(\sum_{i=1}^{\nu} x_i)^2\right] \quad (15)$$

$$\text{ή} \quad s^2 = \frac{1}{\nu}\sum_{i=1}^{\nu} x_i^2 - \bar{x}^2 \quad (16)$$

εάν οι παρατηρήσεις έχουν συχνότητες $\nu_i$ οι τύποι γράφονται ως εξής:

$$s^2 = \frac{1}{\nu}\sum_{i=1}^{\mu}(x_i - \bar{x})^2\cdot\nu_i \quad (17)$$

$$s^2 = \frac{1}{\nu}\left[\sum_{i=1}^{\nu} x_i^2\cdot\nu_i - \frac{1}{\nu}\cdot(\sum_{i=1}^{\nu} x_i\cdot\nu_i)^2\right] \quad (18)$$

$$s^2 = \frac{1}{\nu}\sum_{i=1}^{\nu} x_i^2\cdot\nu_i - \bar{x}^2 \quad (19)$$

**iii. Τυπική απόκλιση s (Standard Deviation):** είναι η τετραγωνική ρίζα της διασποράς. Δηλαδή

$$s = \sqrt{s^2} \quad (20)$$

Η τυπική απόκλιση έχει ευρύτερη χρήση διότι εκφράζεται με την ίδια μονάδα που εκφράζονται και οι παρατηρήσεις.

Η τυπική απόκλιση δεν μεταβάλλεται εάν στις τιμές της μεταβλητής X προστεθεί μια σταθερά αX**R**.

Δηλαδή αν **Y = X + α** τότε $\mathbf{s_x = s_y}$ **.**

Εάν οι τιμές μιας μεταβλητής πολλαπλασιασθούν επί μια σταθερά α$\in$**R** τότε η διασπορά πολλαπλασιάζεται επί την απόλυτη τιμή της σταθεράς α.

Δηλαδή αν **Y = X . α** τότε $\mathbf{s_y = s_x.|\alpha|}$

**iv. Συντελεστής Μεταβολής:** είναι το πηλίκο της τυπικής απόκλισης δια της μέσης τιμής. Δηλαδή

$$CV = \frac{s}{\bar{x}} \quad (21)$$

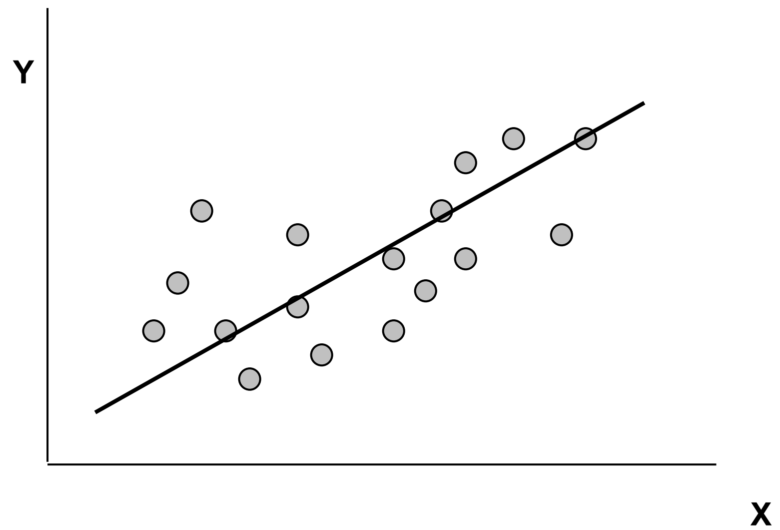
Ο συντελεστής μεταβολής εκφράζεται επί τοις εκατό και είναι ανεξάρτητος από τις μονάδες μέτρησης. Εκφράζει ένα μέτρο σχετικής διασποράς των τιμών της μεταβλητής.

Ένα δείγμα τιμών μιας μεταβλητής είναι **ομοιογενές** όταν ο CV είναι μικρότερος ή ίσος από το 10%.

**10. Ανάλυση Παλινδρόμησης (Regression Analysis):** είναι ο κλάδος της στατιστικής ο οποίος εξετάζει τη σχέση μεταξύ δύο ή περισσοτέρων μεταβλητών, ώστε να είναι δυνατή η πρόβλεψη της μιας από τις υπόλοιπες.

**11. Γραμμική Παλινδρόμηση (Linear Regression):** μιας εξαρτημένης μεταβλητή Y από την ανεξάρτητη μεταβλητή X είναι η σχέση $y = \alpha + \beta.x$ (22) όπου α και β είναι παράμετροι. Ο προσδιορισμός των α και β δίνει μια προσεγγιστική ευθεία, που συνδέει τις τιμές της Y δοθέντων των τιμών της X.

Η ευθεία που προκύπτει λέγεται **ευθεία παλινδρόμησης της Y πάνω στην X.** Η ευθεία αυτή μπορεί να κατασκευασθεί εμπειρικά ή μέσω μαθηματικών μεθόδων, όπως είναι η μέθοδος **ελαχίστων τετραγώνων.** Σκοπός είναι το άθροισμα των τετραγώνων των κατακόρυφων αποστάσεων των σημείων (X,Y) από την ευθεία να είναι ελάχιστο.




**Σχήμα 2**

## *1.2. Διακριτές Κατανομές*

***Η Διωνυμική Κατανομή***

Εφαρμόζεται σε πειράματα όπου πραγματοποιούνται ν επαναλήψεις της ίδιας διαδικασίας και υπάρχουν μόνο τα ενδεχόμενα επιτυχία και αποτυχία.Η πιθανότητα επιτυχίας παραμένει σταθερή και οι επαναλήψεις του πειράματος είναι ανεξάρτητες.Η διωνυμική κατανομή μας δίνει την πιθανότητα να εχουμε ακριβώς χ επιτυχίες στις ανεξάρτητες επαναλήψεις του πειράματος αν p είναι η πιθανότητα επιτυχίας και 1-p=q η πιθανότητα αποτυχίας.

*Χαρακτηριστικά διωνυμικής κατανομής:* αναμενόμενη τιμή

$$E(x) = \frac{\Sigma x_i f(x_i)}{\Sigma f(x_i} = \Sigma x_i f(x_i)$$

***Η Κατανομή Poisson***

Είναι οριακή μορφή της Διωνυμικής Κατανομής όταν η πιθανότητα εμφάνισης του ενδεχομένου είναι μικρή και ο αριθμός των επαναλήψεων μεγάλος. Ο τύπος της κατανομής του Poisson είναι

$$P(x) = P(x = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

όπου x=0,1,2,... ο αριθμός εμφάνισης του ενδεχομένου και λ ο μέσος όρος της κατανομής.

*Χαρακτηριστικά της κατανομής Poisson:* α)αναμενόμενη τιμή Ε(x)=λ,

β)διακύμανση var(x)=λ, γ)συντελεστής ασσυμμετρίας = $B_1 = \frac{1}{\lambda}$,

δ)συντελεστής κύρτωσης = $B_2 = 3 + \frac{1}{\lambda}$

Η κατανομή Poisson έχει εφαρμογή σε καταστάσεις με ενδεχόμενα τα οποία συμβαίνουν ανεξάρτητα στο χώρο και στο χρόνο.

### *Η Υπεργεωμετρική Κατανομή*

Στη διωνυμική κατανομή υποθέσαμε οτι η πιθανότητα εμφάνισης του ενδεχομένου επιτυχίας παραμένει σταθερή σε όλες τις επαναλήψεις του πειράματος. Η υπόθεση αυτή δεν ισχύει στην υπεργεωμετρική κατανομή.

*Χαρακτηριστικά της υπεργεωμετρικής κατανομής:*

α)αναμενόμενη τιμή

$$E(x) = n\frac{m}{N},$$

β)διακύμανση

$$var(x) = npq\frac{(N-n)}{(N-1)}$$

Αν το μέγεθος του πληθυσμού Ν είναι μεγάλο χρησιμοποιούμε τη διωνυμική κατανομή χωρίς μεγάλο σφάλμα.

## *1.3. Συνεχείς Κατανομές*

### *Κανονική Κατανομή*

Η συνεχής τυχαία μεταβλητή χ ακολουθεί την κανονική κατανομή με παραμέτρους μ και $\sigma^2$ όταν έχει την ακόλουθη συνάρτηση πυκνότητας πιθανότητας

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(\frac{(x-\mu)^2}{(2\sigma^2}}$$

, όταν το x ανήκει στο διάστημα(-∞,+∞).

*Χαρακτηριστικά:* α)αναμενόμενη τιμή Ε(x)=μ, β)διακύμανση $var(x) = \sigma^2$, γ)συντελεστής ασυμμετρίας $B_1 = 0$, δ)συντελεστής κύρτωσης $B_2 = 3$

### *Ομοιόμορφη Κατανομή*

Μια συνεχής τυχαία μεταβλητή x ακολουθεί την ομοιόμορφη κατανομή στο διάστημα [α,β] όταν έχει την ακόλουθη συνάρτηση πυκνότητας πιθανότητας

$$f(x) = \frac{1}{\beta} - \alpha, \alpha \leq \chi \leq \beta$$

ή $f(x) = 0$ σε κάθε άλλη περίπτωση,

α)για τη συνεχή περίπτωση όπου $\int_a^b f(x)dx = 1$ και β)για $f(x) \geq 0$

Η πιθανότητα είναι σταθερή κατά μήκος του διαστήματος [α,β].

*Χαρακτηριστικά:* α)αναμενόμενη τιμή $(x) = \frac{(\alpha+\beta)}{2}$, β)διακύμανση

$var(x) = \frac{(\beta-\alpha)^2}{12}$, γ)σταθμικός μέσος= $\sum_{i=1}^{n} P(X = x_i)$

***Η Εκθετική Κατανομή***

Μια συνεχής τυχαία μεταβλητή x ακολουθεί την εκθετική με παράμετρο λ όταν έχει την παρακάτω συνάρτηση πυκνότητας πιθανότητας

$f(x) = \lambda e^{-\lambda x}, x > 0, \lambda > 0$ ή $f(x) = 0$ σε κάθε άλλη περίπτωση.

*Χαρακτηριστικά:*

α)αναμενόμενη τιμή

$$(x) = \frac{1}{\lambda}$$

β)διακύμανση

$$var(x) = \frac{1}{\lambda^2}$$

Η εκθετική κατανομή έχει πολλές εφαρμογές σε καταστάσεις, στις οποίες μελετάμε το χρονικό διάστημα που μεσολαβεί μέχρι να συμβεί ένα γεγονός. Αφορά το χρόνο και είναι συνεχής.

## *1.4. Τυχαίες μεταβλητές και κατανομές πιθανότητας*

***ΤΥΧΑΙΕΣ ΜΕΤΑΒΛΗΤΕΣ***: Έστω S ένας δειγματικός χώρος μιας οποιασδήποτε μορφής και Ai στοιχείο του S. Ο κανόνας που αντιστοιχίζει έναν πραγματικό αριθμό X1 σε κάθε στοιχείο Ai του S, ονομάζεται τυχαία μεταβλητή. Και αυτό διότι η μεταβλητή αυτή παίρνει τιμές κατά τύχη αφού τα Ai είναι αποτελέσματα κάποιου πειράματος τύχης. Σε πιο αυστηρή διατύπωση ως τυχαία μεταβλητή ορίζεται μια μονοσήμαντη συνάρτηση που έχει πεδίο ορισμού το δειγματικό χώρο S και πεδίο τιμών κάποιο υποσύνολο του συνόλου των πραγματικών αριθμών. Με άλλα λόγια, η τυχαία μεταβλητή αποτελεί ένα μετασχηματισμό που απεικονίζει σε κάθε απλό ενδεχόμενο του δειγματικού χώρου ένα μοναδικό σημείο της ευθείας των πραγματικών αριθμών. Οι τυχαίες μεταβλητές διακρίνονται σε συνεχείς και ασυνεχείς ή διακριτές. Μια τυχαία μεταβλητή ονομάζεται ασυνεχής(ή διακριτή ή απαριθμητή) όταν το σύνολο των τιμών της είναι πεπερασμένο ή όταν αυτό είναι απαριθμήσιμο. Όταν ο δειγματικός χώρος δεν αποτελείται από αριθμήσιμα σημεία, αλλά καλύπτει ένα ολόκληρο διάστημα (ή συλλογή διαστημάτων), τότε η μεταβλητή που οι τιμές της εκφράζουν τα αποτελέσματα κάθε πειράματος τύχης ονομάζεται συνεχής τυχαία μεταβλητή. Με άλλα λόγια, μια συνεχής τυχαία μεταβλητή μπορεί να πάρει κάθε τιμή του διαστήματος(πεπερασμένου ή απείρου).

Συνάρτηση κατανομής: Σαν συνάρτηση κατανομής ορίζεται η συνάρτηση F, που για κάθε x, μας δείχνει την πιθανότητα, όπως η τυχαία μεταβλητή να παίρνει την τιμή το πολύ x. Δηλαδή,

$$F(x) = P(X \leq x)$$

***ΑΡΙΘΜΗΤΙΚΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΜΙΑΣ ΤΥΧΑΙΑΣ ΜΕΤΑΒΛΗΤΗΣ***: Έστω X μια τυχαία μεταβλητή και έστω ότι η πυκνότητα πιθανότητας της $f_x(x)$ είναι γνωστή. Τότε, θεωρητικά, μπορούμε να υπολογίσουμε όλες τις πιθανότητες που μας ενδιαφέρουν. Από μαθηματική άποψη υπάρχουν πολλές φορές δυσκολίες που κάνουν αυτούς τους υπολογιαμούς αδύνατους. Για αυτό το λόγο υπάρχουν ορισμένοι αριθμοί που χαρακτηρίζουν την τυχαία μεταβλητή ή την κατανομή της. Η μέση τιμή ή μαθηματική ελπίδα μιας τυχαίας μεταβλητής με πυκνότητα πιθανότητας $f_x(x)$ είναι ο αριθμός:

$$\mu_x = E(X) = \sum_{n=1}^{n\infty} x_n f_x(x_n)$$

εάν X διακριτή και

$$\mu_x = E(X) = \int_{-\infty}^{+\infty} x f_x(x) dx$$

εάν X συνεχής

με την προυπόθεση ότι οι ποσότητες του δεξιού μέλους έχουν νόημα, δηλαδή τόσο το άθροισμα, όσο και το ολοκλήρωμα του δεξιού μέλους είναι πεπερασμένα.

## *1.5. Κατανομές δειγματοληψίας*

**Δειγματικός μέσος και δειγματική μεταβλητότητα**

Έστω δείγμα $X_1, X_2, \ldots, X_n$ προερχόμενο από κάποιο πληθυσμό με μέση τιμή $\mu$ και μεταβλητότητα $\sigma^2$. Έστω $\overline{X}$ και $S^2$ η μέση τιμή και μεταβλητότητα του δείγματος που υπολογίζονται ως εξής:

$$\overline{X} = \frac{\sum X_i}{n}, \quad S^2 = \frac{\sum X_i^2 - n\overline{X}^2}{n-1}.$$

Τότε, ανεξαρτήτως της κατανομής του πληθυσμού, ισχύει

$$\mathrm{E}(\overline{X}) = \mu, \quad \mathrm{var}(\overline{X}) = \frac{\sigma^2}{n}.$$

Αν ο πληθυσμός ακολουθεί κανονική κατανομή ισχύουν οι σχέσεις

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1), \qquad (1)$$

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}, \qquad (2)$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}. \qquad (3)$$

Οι σχέσεις (1),(2) χρησιμοποιούνται για την εξαγωγή συμπερασμάτων για την μέση τιμή του πληθυσμού $\mu$, η (1) όταν η μεταβλητότητα του πληθυσμού υποτίθεται ότι είναι γνωστή, ενώ η (2) όταν είναι άγνωστη. Η σχέση (3) χρησιμοποιείται για την εξαγωγή συμπερασμάτων για την μεταβλητότητα του πληθυσμού $\sigma^2$.

Αν το δείγμα είναι μεγάλο ($n > 30$), τότε, ανεξαρτήτως της κατανομής του πληθυσμού, η σχέση (1) ισχύει και η σχέση (2) αντικαθίσταται από την σχέση

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim N(0,1).$$

**Διαφορά δειγματικών μέσων και λόγος δειγματικών μεταβλητοτήτων**

Έστω δύο δείγματα προερχόμενα από δύο διαφορετικούς πληθυσμούς με μέσες τιμές $\mu_1, \mu_2$ και μεταβλητότητες $\sigma_1^2, \sigma_2^2$. Έστω $\overline{X}_1, \overline{X}_2$ και $S_1^2, S_2^2$ οι μέσες τιμές και μεταβλητότητες των δύο δειγμάτων, και $n_1, n_2$ τα μεγέθη τους. Υπολογίζουμε επίσης τον σταθμισμένο μέσο όρο των $S_1^2, S_2^2$ από τον τύπο

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} .$$

Αν οι δύο πληθυσμοί ακολουθούν κανονικές κατανομές ισχύουν οι σχέσεις

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1) , \qquad (4)$$

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} , \qquad (5)$$

$$\frac{S_1^2 / S_2^2}{\sigma_1^2 / \sigma_2^2} \sim F_{n_1-1, n_2-1} \qquad (6)$$

Η σχέση (4) χρησιμοποιείται για τη σύγκριση των μέσων τιμών των δύο πληθυσμών όταν οι μεταβλητότητές τους υποτίθενται γνωστές, ενώ η σχέση (5) όταν είναι άγνωστες αλλά ίσες. Η σχέση (6) χρησιμοποιείται για την σύγκριση των μεταβλητοτήτων των δύο πληθυσμών.

Αν τα δείγματα είναι μεγάλα, τότε, ανεξαρτήτως της κατανομής των δύο πληθυσμών, η σχέση (4) ισχύει και η σχέση (5) αντικαθίσταται από την σχέση

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1).$$

**Ποσοστά και διαφορά ποσοστών σε δείγματα**

Έστω $p$ το ποσοστό κάποιας κατηγορίας ενός πληθυσμού και $\hat{p}$ το αντίστοιχο ποσοστό σε δείγμα μεγέθους $n$. Αν το δείγμα είναι μεγάλο, ισχύει η σχέση

$$\hat{p} \sim N(p, p(1-p)/n). \quad (7)$$

Κατ' ανάλογο τρόπο, η διαφορά ποσοστών σε δύο μεγάλα δείγματα που προέρχονται από δύο διαφορετικούς πληθυσμούς ακολουθεί επίσης κανονική κατανομή σύμφωνα με την ακόλουθη σχέση

$$\hat{p}_1 - \hat{p}_2 \sim N(p_1 - p_2, p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2). \quad (8)$$

## *1.6. Μέτρηση των μεταβλητών*

- *Κλίμακα λόγου (ratio scale)*: Περιλαμβάνει όλα τα ποσοτικά χαρακτηριστικά αλλά χρησιμοποιείται κυρίως σε δεδομένα όπου υπάρχει νόημα για αναφορά στη μηδενική τιμή.
- *Κλίμακα διαστήματος (interval scale):* Αναφέρεται σε δεδομένα που οι τιμές τους μπορούν να συγκριθούν ως προς τις μεταβολές τους αλλά όχι μεταξύ τους. Χαρακτηριστικό παράδειγμα είναι η θερμοκρασία όπου μπορούμε να συζητήσουμε για αύξηση της θερμοκρασίας αλλά ποτέ δε θα πούμε για ποσοστιαία τιμή.
- *Κλίμακα ιεράρχησης (ordinal scale):* Αφορά τις τιμές των χαρακτηριστικών που αντικείμενό τους είναι να ιεραρχήσουν τις

παρατηρήσεις με αύξουσα ή φθίνουσα σειρά. Η βαθμολογία αποτελεί χαρακτηριστικό παράδειγμα.

- *Ονομαστική κλίμακα (nominal scale):* Όταν χρειαζόμαστε απλή ταξινόμηση των παρατηρήσεων που κατηγοριοποιούν τα δεδομένα (π.χ. Φύλο, οικογενειακή κατάσταση κ.λ.π.).

### 1.6.1. Σχέσεις μεταξύ των μεταβλητών

Ανεξάρτητα από τον τύπο τους, δύο ή περισσότερες μεταβλητές συσχετίζονται εάν σε ένα δείγμα των παρατηρήσεων, οι τιμές εκείνων των μεταβλητών κατανέμονται κατά τρόπο συνεπή. Με άλλα λόγια, οι μεταβλητές συσχετίζονται εάν οι τιμές τους αντιστοιχούν συστηματικά η μια στην άλλη για αυτές τις παρατηρήσεις.

Γενικά, ο απώτερος στόχος κάθε έρευνας ή της επιστημονικής ανάλυσης βρίσκει τις σχέσεις μεταξύ των μεταβλητών. Η φιλοσοφία της επιστήμης μας διδάσκει ότι δεν υπάρχει κανένας άλλος τρόπος για να αναζητήσουμε τη σημασία ενός φαινομένου εκτός από την συσχέτιση ποσοσοτικών και ποιοτικών χαρακτηριστικών, επομένως την αναζήτηση σχέσεων μεταξύ των μεταβλητών. Κατά συνέπεια, η πρόοδος της επιστήμης πρέπει πάντα να περιλάβει την εύρεση των νέων σχέσεων μεταξύ των μεταβλητών. Η συσχετιστική έρευνα περιλαμβάνει τη μέτρηση τέτοιων σχέσεων με τον απλούστερο τρόπο. Εντούτοις, η πειραματική έρευνα δεν είναι καθόλου διαφορετική από αυτή την άποψη. Παραδείγματος χάριν, το προαναφερθέν πείραμα που συγκρίνει χρωμοσώματα στα αρσενικά και τα θηλυκά μπορεί να περιγραφεί ψάχνοντας έναν συσχετισμό μεταξύ δύο μεταβλητών: φύλο και χρωμοσώματα. Οι στατιστικές μόνο αυτό μας βοηθούν, να αξιολογήσουμε τις σχέσεις μεταξύ των μεταβλητών.

### 1.6.2. Δύο βασικά χαρακτηριστικά γνωρίσματα κάθε σχέσης μεταξύ των μεταβλητών

Οι δύο πιό στοιχειώδεις ιδιότητες κάθε σχέσης μεταξύ των μεταβλητών είναι το *μέγεθος* της σχέσης και η *αξιοπιστία* της.

1. *Μέγεθος*. Το μέγεθος είναι πολύ ευκολότερο να το καταλάβουμε και να το μετρήσουμε σε σχέση με την αξιοπιστία. Παραδείγματος χάριν, εάν κάθε αρσενικό στο δείγμα μας βρέθηκε για να είναι ψηλότερο από οποιοδήποτε θηλυκό στο δείγμα, θα μπορούσαμε να πούμε ότι το μέγεθος της σχέσης μεταξύ των δύο μεταβλητών (φύλο και ύψος) είναι πολύ υψηλό στο δείγμα μας. Με άλλα λόγια, θα μπορούσαμε να προβλέψουμε το ένα βασισμένο σε άλλο (τουλάχιστον μεταξύ των μελών του δείγματός μας).
2. *Αξιοπιστία*. Η αξιοπιστία μιας σχέσης είναι μια πολύ λιγότερο διαισθητική έννοια, αλλά ακόμα εξαιρετικά σημαντική. Αναφέρεται στη "αντιπροσωπευτικότητα" του αποτελέσματος που βρίσκεται στο συγκεκριμένο δείγμα μας για τον ολόκληρο πληθυσμό. Με άλλα λόγια, λέει πόσο πιθανό είναι ότι μια παρόμοια σχέση θα βρισκόταν εάν το πείραμα ξαναγινόταν με άλλα δείγματα που προήλθαν από τον ίδιο πληθυσμό. Θυμηθείτε ότι δεν ενδιαφερόμαστε σχεδόν ποτέ "τελικά" μόνο για αυτό που γίνεται στο δείγμα μας αλλά ενδιαφερόμαστε για το δείγμα μόνο στην έκταση που μπορεί να παρέχει τις πληροφορίες για τον πληθυσμό. Εάν η μελέτη μας ικανοποιεί μερικά ειδικά κριτήρια (θα αναφερθούν αργότερα), κατόπιν η αξιοπιστία μιας σχέσης μεταξύ των μεταβλητών που παρατηρούνται στο δείγμα μας μπορεί να υπολογιστεί ποσοτικά και να αντιπροσωπευθεί χρησιμοποιώντας ένα τυποποιημένο μέτρο (τεχνικά αποκαλούμενο ως p-value ή στατιστικό επίπεδο σημαντικότητας.

## *1.7. Τι είναι η p-value*

Η στατιστική σημαντικότητα ενός αποτελέσματος είναι η πιθανότητα ότι η παρατηρηθείσα σχέση (π.χ., μεταξύ των μεταβλητών) ή της διαφοράς (π.χ., μεταξύ των μέσων) σε ένα δείγμα εμφανίστηκε κατά καθαρή τύχη και ότι στον πληθυσμό από τον οποίο το δείγμα προήλθε, καμία τέτοια σχέση ή διαφορά δεν υπάρχει.

Χρησιμοποιώντας λιγότερο τεχνικούς όρους, κάποιος θα μπορούσε να πει ότι η στατιστική σημασία ενός αποτελέσματος μας λέει κάτι για το βαθμό στον οποίο το αποτέλεσμα είναι "αληθινό" (από την άποψη της ύπαρξης "αντιπροσώπευσης του πληθυσμού"). Πιό τεχνικά, η τιμή της p-value αντιπροσωπεύει έναν δείκτη της αξιοπιστίας ενός αποτελέσματος. Όσο υψηλότερη η p-value, λιγότερο μπορούμε να πιστεψουμε ότι η παρατηρηθείσα σχέση μεταξύ των μεταβλητών στο δείγμα είναι ένας αξιόπιστος δείκτης της σχέσης μεταξύ των αντίστοιχων μεταβλητών στον πληθυσμό.

Συγκεκριμένα, η p-value αντιπροσωπεύει την πιθανότητα του λάθους που περιλαμβάνεται στην αποδοχή του παρατηρηθέντος αποτελέσματός μας τόσο έγκυρου, δηλαδή όσο "η αντιπρσώπευση του πληθυσμού". Παραδείγματος χάριν, μια p-value του 0.05 (δηλ., 1/20) δείχνει ότι υπάρχει μια πιθανότητα 5% ότι η σχέση μεταξύ των μεταβλητών που βρίσκονται στο δείγμα μας να είναι "ψευδής". Με άλλα λόγια, υποθέτοντας ότι στον πληθυσμό δεν υπήρξε καμία σχέση μεταξύ εκείνων των μεταβλητών, και επαναλαμβάναμε τα πειράματα μας, θα μπορούσαμε να αναμείνουμε ότι περίπου σε κάθε 20 επαναλήψεις του πειράματος θα υπήρχε ένα στον οποίο η σχέση μεταξύ των εν λόγω μεταβλητών θα ήταν ίση ή ισχυρότερη απ' ό,τι στους δικούς μας υπολογισμούς.

Σε πολλούς τομείς της έρευνας, η p-value του 0.05 είναι συνήθως η διαχωριστική γραμμή ως αποδεκτό "επίπεδο λάθους".

# ΚΕΦΑΛΑΙΟ 2

# Στατιστική Επαγωγή και Σημειακή Εκτίμηση

## *2.1. Επαγωγική Στατιστική*

Η Επαγωγική Στατιστική είναι το μέρος εκείνο της Στατιστικής που ασχολείται με τις μεθόδους που καθιστούν δυνατή τη γενίκευση των δειγματοληπτικών συμπερασμάτων στον γεννήτορα πληθυσμό, από τον οποίο πήραμε το δείγμα και μας βοηθά στη λήψη σωστών αποφάσεων σε συνθήκες αβεβαιότητας. Η διαδικασία της εκτίμησης των αγνώστων παραμέτρων ενός πληθυσμού με τη βοήθεια ενός τυχαίου δείγματος, ο έλεγχος της ισχύος ορισμένων υποθέσεων και κατά πόσο ένα δείγμα προέρχεται από πληθυσμό που ακολουθεί μια ορισμένη θεωρητική κατανομή, αποτελεί βασικό αντικείμενο της Επαγωγικής Στατιστικής. Σπουδαίο εργαλείο στο έργο της Επαγωγικής Στατιστικής είναι η θεωρία των Πιθανοτήτων.

Η Επαγωγική Στατιστική είναι η μεθοδολογία με την οποία εξάγουμε συμπεράσματα για τις παραμέτρους ενός πληθυσμού με βάση τις περιορισμένες πληροφορίες ενός τυχαίου δείγματος.

Σύμφωνα με έναν άλλον ορισμό, με τον όρο Στατιστική Επαγωγή, εννοούμε την επιστημονική προσέγγιση στο θέμα της μελέτης και της εξαγωγής συμπερασμάτων προσδιορίσιμης αξιοπιστίας αναφορικά με τη δομή και τα χαρακτηριστικά ενός συνόλου αντικειμένων, στηριζόμενοι σε πληροφορίες που περιέχονται σε ένα μικρό σχετικά υποσύνολο του πληθυσμού.

Τις προς εξέταση παραμέτρους του πληθυσμού τις παριστάνουμε με $\theta_1, \theta_2, \ldots, \theta_\nu$ και μπορεί να είναι ο μέσος αριθμητικός, η διακύμανση, η διάμεσος, η επικρατούσα τιμή ή μια οποιαδήποτε ροπή του γεννήτορα πληθυσμού.

Ας δούμε ένα παράδειγμα για να κατανοήσουμε την έννοια της Επαγωγικής Στατιστικής. Ας εξετάσουμε λοιπόν τις έρευνες που γίνονται την ημέρα των εκλογών, που δείγμα ψηφοφόρων εξετάζεται την ώρα που βγαίνουν από τα εκλογικά τμήματα για το πολιτικό κόμμα που ψήφισαν, τα γνωστά σαν **exit polls.** Σκοπός του ερευνητή είναι να εκτιμήσει το ποσοστό των ψήφων που τελικά θα πάρει το κόμμα Α. Το σύνολο των ψηφοφόρων που θα ασκήσουν το εκλογικό τους δικαίωμα αποτελεί τον πληθυσμό του εκλογικού σώματος, ενώ ο συγκεκριμένος αριθμός ψηφοφόρων που θα ερωτηθούν από τους ερευνητές αποτελούν το δείγμα.

Μετά την τελική καταμέτρηση των ψήφων θα προκύψει το πραγματικό ποσοστό των ψήφων που πήρε το κόμμα Α. Αυτή η μέτρηση αποτελεί την παράμετρο του πληθυσμού, δηλαδή το πραγματικό ποσοστό του κόμματος Α. Ενώ, το ποσοστό που θα προκύψει, από το δείγμα των ψηφοφόρων που κάλυψε η έρευνα, αποτελεί μια εκτίμηση του άγνωστου ποσοστού που τελικά θα πάρει το κόμμα Α.

Η επαγωγική σταστιστική μας επιτρέπει, χρησιμοποιώντας την εκτίμηση του δείγματος, να διεξάγουμε συμπεράσματα για τον πληθυσμό. Έτσι, με την κατάλληλη μέθοδο, μπορούμε να συμπεράνουμε σε ποιό διάστημα τιμών αναμένεται να διαμορφωθεί το τελικό ποσοστό του κόμματος Α.

Τα σύνολα των αντικειμένων που εξετάζονται εκφράζουν πραγματικές καταστάσεις και η μελέτη τους γίνεται με τη βοήθεια υποδειγμάτων πιθανότητας. Η μεθοδολογική στρατηγική που ακολουθείται στο κρίσιμο θέμα της επαγωγής εξαρτάται από το είδος των πληροφοριών που θεωρούνται σχετικές με το φαινόμενο που διερευνάται. Έτσι, ανάλογα με τη φύση των πληροφοριών που πρόκειται να επεξεργαστούν, διακρίνουμε δυο βασικές προσεγγίσεις στο θέμα της σταστιστικής επαγωγής: την **κλασική** και την **μπεϋζιανή**.

Η κλασική προσέγγιση ακολουθείται στις περιπτώσεις που μοναδική πηγή πληροφοριών είναι τα δεδομένα δείγματος από τον πληθυσμό που αποτελεί αντικείμενο μελέτης. Τα δεδομένα αυτά εκφράζονται με τη συνάρτηση πιθανοφάνειας του δείγματος και τις κατανομές δειγματοληψίας των διαφόρων στατιστικών ( π.χ. του δειγματικού μέσου, της διακύμανσης κ.λπ.). Για τη διαμόρφωση του αναγκαίου αναλυτικού πλαισίου, στις περιπτώσεις αυτές, είναι απαραίτητο η πιθανότητα να ερμηνεύεται ως σχετική συχνότητα, διότι έτσι διευκολύνεται η κατασκευή των σχετικών εμπειρικών κατανομών. Όταν όμως για τη μελέτη του φαινομένου και τη διατύπωση κανόνων λήψης αποφάσεων, κρίνεται αναγκαία η ενσωμάτωση στα δεδομένα του δείγματος και πληροφοριών, οι οποίες αφορούν γνώση που προέρχεται από παρακολούθηση του φαινομένου κατά το παρελθόν ή γενικότερα γνώσεις που οφείλονται σε εξωγενείς πηγές, τότε η κλασική προσέγγιση στο θέμα της στατιστικής επαγωγής δεν είναι επαρκής.

Συχνά οι πληροφορίες που προέρχονται από εξωγενείς πηγές έχουν έντονο το στοιχείο της υποκειμενικής εκτίμησης. Αποτέλεσμα αυτού είναι να απαιτείται άλλη μεθοδολογική προσέγγιση για τη συνδυασμένη επεξεργασία δύο διαφορετικών κατηγοριών, μιας αντικειμενικής, όπως είναι τα δεδομένα του δείγματος, και μιας κατά κύριο λόγο υποκειμενικής, όπως είναι συνήθως οι πληροφορίες από εξωγενείς πηγές. Για την αντιμετώπιση των προβλημάτων που ανακύπτουν στη δεύτερη περίπτωση έχει αναπτυχθεί μια διαφορετική προσέγγιση που είναι γνωστή στη βιβλιογραφία ως μπεϋζιανή προσέγγιση. Σύμφωνα με την προσέγγιση αυτή, η εκ των προτέρων πληροφόρηση, όπως ονομάζονται οι εξωγενείς πληροφορίες, ενσωματώνεται στα δεδομένα του δείγματος μέσα από μια διαδικασία επαναληπτικών χρησιμοποιήσεων του θεωρήματος του Bayes.

Στο χώρο της κλασικής στατιστικής επαγωγής ανήκουν δύο συναφή προβλήματα: η εκτίμηση των παραμέτρων του πληθυσμού ( ή του μοντέλου ), και ο έλεγχος υποθέσεων, ο οποίος περιλαμβάνει τόσο τον έλεγχο της αξιοπιστίας των παραμέτρων του πληθυσμού, όσο και τον έλεγχο διαφόρων άλλων υποθέσεων. Στην περίπτωση της εκτίμησης το ενδιαφέρον επικεντρώνεται στην επιλογή (εκτίμηση) της κατάλληλης τιμής για μια ή περισσότερες παραμέτρους του πληθυσμού. Στον έλεγχο των υποθέσεων, το πρόβλημα γίνεται πιο συγκεκριμένο και αφορά στην αποδοχή ή στην απόρριψη μιας δεδομένης τιμής μιας παραμέτρου σε όρους πιθανότητας.

Το πρόβλημα της εκτίμησης διαχωρίζεται σε δύο επιμέρους θέματα: στην **εκτίμηση σημείου** ή **σημειακή εκτίμηση** και την **εκτίμηση διαστήματος**. Στην πρώτη περίπτωση τα δεδομένα επιτρέπουν τον υπολογισμό μιας μόνο τιμής για την παράμετρο που είναι αντικείμενο μελέτης, ενώ στη δεύτερη περίπτωση προσδιορίζεται ένα διάστημα τιμών μέσα στο οποίο αναμένεται με κάποιο βαθμό εμπιστοσύνης (πιθανότητας) να περιλαμβάνεται η άγνωστη παράμετρος του πληθυσμού.

Η εκτίμηση μιας παραμέτρου του πληθυσμού μπορεί να δοθεί με έναν αριθμό. Έχουμε τότε μια σημειακή εκτίμηση της παραμέτρου. Η εκτίμηση μιας παραμέτρου του πληθυσμού μπορεί να δοθεί και με ένα διάστημα, μέσα στο όποιο πιστεύουμε ότι περιέχεται η παράμετρος του πληθυσμού. Έχουμε τότε μια εκτίμηση διαστήματος της παραμέτρου. Μια πληροφορία για το σφάλμα ή την ακρίβεια της εκτιμήσεως λέμε ότι περιγράφει την αξιοπιστία της εκτιμήσεως.

Οι διαδικασίες επαγωγικής στατιστικής όπως εκτίμηση (estimation), διαστήματα εμπιστοσύνης (confidence intervals), έλεγχος (testing), και πρόβλεψη (prediction) στηρίζονται στις πληροφορίες που συνοψίζονται στη κατανομή του δείγματος (the sampling distribution).

Σε μερικές περιπτώσεις η κατανομή του δείγματος (sampling distribution) είναι εύκολο να την βρεις αλλά γενικά είναι πολύ δύσκολο. Αυτός είναι ο λόγος που συνήθως προσεγγίζουμε την κατανομή του δείγματος ασυμπτωτικά.

## *2.2. Βασικές Έννοιες της Επαγωγικής Στατιστικής*

**Στατιστικά μοντέλα** είναι επαρκείς περιγραφές των παρατηρήσιμων φαινομένων υπό μορφή στοχαστικών μηχανισμών.

**Στοχαστικό φαινόμενο** είναι ένα φαινόμενο του οποίου τα παρατηρήσιμα στοιχεία είναι τυχαία (αβέβαια) - random (uncertain) - σε αντίθεση με ντετερμινιστικά (deterministic). Ή καλύτερα, **στοχαστικό φαινόμενο** είναι το φαινόμενο του οποίου τα παρατηρήσιμα στοιχεία παρουσιάζουν μια τυχαία κανονικότητα - **chance regularity**.

Με τον όρο '**chance'** εννοούμε ότι υπάρχει μια έμφυτη αβεβαιότητα σχετικά με το περιστατικό στις ιδιαίτερες εκβάσεις (απουσία τάξης στο ατομικό επίπεδο).

Με τον όρο '**regularity'** εννοούμε ότι υπάρχει μια μόνιμη κανονικότητα σε σχέση με το συμβάν πολλών τέτοιων εκβάσεων (τάξη στο συνολικό επίπεδο).

***Μορφές τυχαίας κανονικότητας (forms of chance regularity)***

**Κατανομή (Distribution):** οι επαναλαμβανόμενες δοκιμές διαμορφώνουν ένα σταθερό νόμο (stable law).

**Ανεξαρτησία (Independence)**: σε οποιαδήποτε ακολουθία δοκιμών η έκβαση οποιασδήποτε δοκιμής δεν επηρεάζει και δεν επηρεάζεται από αυτήν οποιοδήποτε άλλων.

**Ομοιογένεια (Homogeneity):** οι πιθανότητες που συνδέονται με τις διάφορες εκβάσεις παραμένουν ίδιες για όλες τις δοκιμές.

Οι τρεις διαφορετικές μορφές chance regularity αποτελούν τις ιδιαίτερες περιπτώσεις τριών διαφορετικών και ευρέων κατηγοριών υποθέσεων πιθανότητας:

Κατανομής – Distribution (D)

Εξάρτησης – Dependence (M)

Ετερογένειας – Heterogeneity (H)

Αυτές οι ευρείς κατηγορίες μπορούν να αντιμετωπισθούν ως βασικά συστατικά κάθε στατιστικού μοντέλου. Στην πραγματικότητα, ένα στατιστικό μοντέλο είναι ένα σύνολο συμβατών υποθέσεων πιθανότητας από τις τρεις ευρείς κατηγορίες: (D), (M), και (H).

Στην περίπτωση των οικονομικών φαινομένων πρώτα πρέπει να αποφασίσουμε εάν το υπό εξέταση φαινόμενο μπορεί να αντιμετωπισθεί ως στοχαστικό. Και έπειτα, πρέπει να εκμεταλλευτούμε τη τυχαία κανονικότητα (chance regularity) που διακρίνεται στα διαγράμματα των στοιχείων για να επιλέξουν ένα κατάλληλο στατιστικό μοντέλο.

**<u>Στοχαστικό φαινόμενο - Πληθυσμός (population)</u>**

Συνήθως ο πραγματικός στόχος ενδιαφέροντος ορίζεται ως πληθυσμός .Εμείς όμως θα δουλεύουμε με την έννοια του στοχαστικού φαινομένου όπως ορίστηκε προηγουμένως.

**<u>Δείγμα (sample)</u>**

Συνήθως ορίζεται ως ένα υποσύνολο του πληθυσμού. Τα στοιχεία αντιμετωπίζονται συνήθως ως δείγμα. Εμείς θα το ορίσουμε ως συγκεκριμένες πραγματοποιήσεις του στοχαστικού μηχανισμού που δημιουργούν τα παρατηρήσιμα στοιχεία (δηλ. δείγμα).

**Η περιγραφική στατιστική (descriptive statistics)** ορίζεται ως οι περιληπτικές, γραφικές, και αριθμητικές μέθοδοι που χρησιμοποιούνται για να συνοψίσουν τα στατιστικά στοιχεία.Συμπεράσματα βασισμένα στη περιγραφική στατιστική δεν μπορούν να επεκταθούν πέρα από τα διαθέσιμα στατιστικά στοιχεία.

Αυτός είναι ο λόγος που πρέπει να μελετήσουμε επαγωγική στατιστική (statistical inference).

Η επαγωγική στατιστική είναι η διαδικασία της εκτίμησης και στατιστικού ελέγχου υποθέσεων για τα χαρακτηριστικά του στοχαστικού φαινομένου (πληθυσμού).

**Η Επαγωγική στατιστική είναι ένα σύνολο από διαδικασίες για να** εξάγουμε συμπεράσματα και να θεμελιώνουμε ισχυρισμούς για το στοχαστικό φαινόμενο που προκάλεσε τα στοιχεία χρησιμοποιώντας (α) το υποθετημένο στατιστικό μοντέλο (postulated statistical model) και (β) τη πραγματοποίηση του δείγματος (the sample realization – the observed data).

Η επαγωγική στατιστική παρέχει τις σωστές διαδικασίες που επιτρέπουν την εξαγωγή συμπερασμάτων και τη θεμελίωση ισχυρισμών.Η επαγωγική στατιστική χρησιμοποιεί (α) το προϋποθετημένο (postulated) στατιστικό μοντέλο και (β) την πραγματοποίηση δειγμάτων (παρατηρηθέντα στατιστικά στοιχεία).

**Επαγωγικό επιχείρημα (inductive argument):** χρησιμοποιώντας ένα data set εξάγουμε συμπεράσματα και θεμελιώνουμε ισχυρισμούς για το στοχαστικό φαινόμενο που προκάλεσε τα στοιχεία.

Εντούτοις, αυτό το επαγωγικό επιχείρημα ενσωματώνεται στο αναγωγικό επιχείρημα (deductive argument)

**Αναγωγικό επιχείρημα**: εάν οι υποθέσεις του προυποθετημένου στατιστικού μοντέλου ισχύουν τότε ορισμένα συμπεράσματα ακολουθούν απαραίτητα.

Ως εκ τούτου, μια κρίσιμη πτυχή του στατιστικού συμπεράσματος είναι να εξασφαλιστεί ότι οι υποθέσεις ισχύουν γιατί αλλιώς τα συμπεράσματα δεν ακολουθούν απαραιτήτως.

Ας εξετάσουμε το πιο κάτω απλό γενικό στατιστικό μοντέλο, το οποίο δεν υποθέτει συγκεκριμένη κατανομή.

$$
\begin{aligned}
&\text{(i)}\ \mathrm{X}_{\mathrm{t}} = E(X) + u_t \\
&\text{(ii)}\ \Phi = \{f_X(x;\theta),\theta \in \Theta, \mathrm{x} \in \Re\} \\
&\text{(ii)}\ (\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_{\mathrm{n}})\ \text{είναι τυχαίο δείγμα}
\end{aligned}
$$

Τα στατιστικά στοιχεία $\boldsymbol{x}_n = (x_1, x_2, \ldots, x_n)$ ερμηνεύονται σαν μια αληθινή και τυπική πραγματοποίηση του στοχαστικού μηχανισμού όπως προσδιορίζεται από το στατιστικό μοντέλο. Συγκεκριμένα, οι παρατηρήσεις, ερμηνεύονται σαν συγκεκριμένες τιμές των τυχαίων μεταβλητών του δείγματος $\boldsymbol{X}_n = (X_1, X_2, \ldots, X_n)$. Αυτό είναι πολύ σημαντικό γιατί προσφέρει τη βάση για το επαγωγικό επιχείρημα της επαγωγικής στατιστικής.

***Πως συνδέουμε τη πραγματοποίηση του δείγματος με το στατιστικό μοντέλο;***

Το πρώτο στάδιο της σύνδεσης χρησιμοποιεί τη κατανομή του δείγματος (sampling distribution or distribution of the sample), η οποία μας δίνει τη σχέση μεταξύ των μοντέλων δείγματος και πιθανότητας.

Οι διαδικασίες επαγωγικής στατιστικής όπως εκτίμηση (estimation), διαστήματα εμπιστοσύνης (confidence intervals), έλεγχος (testing), και πρόβλεψη (prediction) στηρίζονται στις πληροφορίες που συνοψίζονται στη κατανομή του δείγματος (the sampling distribution).

Σε μερικές περιπτώσεις η κατανομή του δείγματος (sampling distribution) είναι εύκολο να την βρεις αλλά γενικά είναι πολύ δύσκολο. Αυτός είναι ο λόγος που συνήθως προσεγγίζουμε την κατανομή του δείγματος ασυμπτωτικά.

Η εκτίμηση είναι βασικά η χρησιμοποίηση πληροφοριών στα στατιστικά στοιχεία για επιλέξουμε μια συγκεκριμένη τιμή για την άγνωστη παράμετρο.

Ο εκτιμητής μια άγνωστης παραμέτρου θ ορίζεται ως

$$\hat{\theta} = h(X_1, X_2, \ldots, X_n)$$

Η συγκεκριμένη τιμή του εκτιμητή για δεδομένη πραγματοποίηση του δείγματος ***X***=***x*** αναφέρεται ως *εκτιμημένη τιμή* (*estimate).*

`

$$\hat{\theta} = h(x_1, x_2, ..., x_n)$$

Ο εκτιμητής είναι μια συνάρτηση από τυχαίες μεταβλητές του δείγματος. Άρα και ο εκτιμητής είναι τυχαία μεταβλητή και παίρνει διαφορετικές τιμές για διαφορετικές πραγματοποιήσεις του δείγματος!

Παράδειγμα :

Απλό Στατιστικό Μοντέλο Bernoulli.

$$\begin{aligned}
&\text{(i) } \mathrm{X_i} = E(X_i) + u_i \\
&\text{(ii) } \{f_X(x;\theta) = \theta^x (1-\theta)^{1-x}, 0 < \theta < 1, x = \{0,1\}\} \\
&\text{(ii) } (\mathrm{X_1}, \mathrm{X_2}, ..., \mathrm{X_n})\ \varepsilon\acute{\iota}\nu\alpha\iota\ \tau\upsilon\chi\alpha\acute{\iota}o\ \delta\varepsilon\acute{\iota}\gamma\mu\alpha
\end{aligned}$$

Ας υποθέσουμε ότι έχουμε data με μέγεθος n=10

**x** = {1, 1, 0, 0, 1, 1, 1, 1, 0, 1}

Ξέρουμε ότι θ=*E(X)*. Αυτό εισηγείται ότι ένας φυσικός εκτιμητής του θ είναι ο δειγματικός μέσος (sample mean)

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Οι ακόλουθοι είναι πιθανοί εκτιμητές του θ

$$\text{(a)}\,\hat{\theta}_1 = X_1, \quad \text{(b)}\,\hat{\theta}_2 = \frac{1}{2}(X_1 + X_2), \ \text{(c)}\,\hat{\theta}_3 = \frac{1}{3}(X_1 + X_2 + X_3),$$
$$\text{(d)}\,\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \text{(e)}\,\hat{\theta}_{n+1} = \frac{1}{n+1}\sum_{i=1}^{n} X_i, \ \text{(f)}\,\hat{\theta}_{n+2} = \frac{1}{n+2}\sum_{i=1}^{n} X_i$$

Με τις αντίστοιχες εκτιμημένες τιμές (estimates)

$$\text{(a)}\,\hat{\theta}_1 = 1, \quad \text{(b)}\,\hat{\theta}_2 = \frac{1}{2}(1+1) = 1, \ \text{(c)}\,\hat{\theta}_3 = \frac{1}{3}(1+1+0) = \frac{2}{3},$$
$$\text{(d)}\,\hat{\theta}_n = \frac{7}{10}, \text{(e)}\,\hat{\theta}_{n+1} = \frac{7}{11}, \ \text{(f)}\,\hat{\theta}_{n+2} = \frac{7}{12}$$

Γενικά, δύσκολα μπορεί να βρεθεί η κατανομή του δείγματος!

Ορισμένα λήμματα θα μας βοηθήσουν να βρούμε τη κατανομή του δείγματος.

**Λήμμα 1:**

Αν X1 , X2, X3 ,…, Xn είναι ανεξάρτητες Κανονικές τυχαίες μεταβλητές (independent Normally distributed random variables) με μέσο E(Xi)=μi και Var(Xi)=σi2 for i=1,2,…,n τότε η συνάρτηση

$$Y_n = \sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

Απλό Κανονικό Στατιστικό Μοντέλο

$$\text{(i) } \mathrm{X_t} = \mu + u_t,\ \mathrm{u_t} \equiv \mathrm{X_t} - E(X_t)$$
$$\text{(ii) } \Phi = \{f_X(x;\theta) = \frac{1}{\sigma\sqrt{\pi}} e^{\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]}, \mu \in \Re, \sigma^2 \in \Re_+, x \in \Re\}$$
$$\text{(iii) } (\mathrm{X_1, X_2, ..., X_n}) \text{ is a random sample}$$

Ας υποθέσουμε ότι οι εκτιμητές του μέσου μ και της διακύμανσης $\sigma^2$ είναι

$$\hat{\mu}_n = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ and } \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

Χρησιμοποιώντας το Λήμμα 1 παίρνουμε ότι η κατανομή δείγματος (sampling distribution) του δειγματικού μέσου (sample mean) για το Απλό Κανονικό Στατιστικό Μοντέλο είναι

$$\hat{\mu}_n = \overline{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$$

**Λήμμα 2:** Σχέση μεταξύ Normal and Chi-square distribution

Αν $Z_1, Z_2, \ldots, Z_n$ είναι IID τυπικές Κανονικές τυχαίες μεταβλητές (standard Normal variables) $Z_k \overset{IID}{\sim} N(0,1)$ για $k = 1,2,\ldots,n$

τότε $V_n = \sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)$ , Όπου n συμβολίζει τους βαθμούς ελευθερίας (degrees of freedom) της Κατανομής $X^2(n)$.

E(Vn)=n

Var(Vn)=2n

**Λήμμα 3:**

Αν Z και V είναι δυο ανεξάρτητες τυχαίες μεταβλητές έτσι ώστε Z~N(0,1) και V~ $X^2(m)$ τότε ο λόγος

$$\left(\frac{Z}{\sqrt{V/m}}\right) \text{έχει Student's t κατανομή με m βαθμούς ελευθερίας}$$

$$\frac{Z}{\sqrt{V/m}} \sim St(m)$$

**Λήμμα 4:**

Αν V1 και V2 είναι δυο ανεξάρτητες τυχαίες μεταβλητές έτσι ώστε V1~ $X^2(m1)$ και V2~ $X^2(m2)$ τότε

$$\text{(a) } U = \frac{V_1/m_1}{V_2/m_2} \sim F(m_1, m_2)$$

$$(b)\ \left(V_1 + V_2\right) \sim \chi^2(m_1 + m_2)$$

**Νόμος των Μεγάλων Αριθμών (Law of Large Numbers) και Κεντρικό Οριακό Θεώρημα (Central Limit Theorem)**

Όταν είναι δύσκολο να βρούμε την κατανομή του δείγματος για πεπερασμένο δείγμα χρησιμοποιούμε ασυμπτωτικά αποτελέσματα.

Έστω ότι έχουμε ένα τυχαίο δείγμα (IID) μεγέθους n,

{X1, X2, …., Xn} με μέσο E(Xi)=μ και Var(Xi)=$\sigma^2$ για i=1,2,….,n.

Ας εξετάσουμε το δειγματικό μέσο

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$E(\overline{X}) = \frac{1}{n}E(\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} EX_i \overset{ID}{=} \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

$$\mathrm{Var}(\overline{X}) = \frac{1}{n^2}Var(\sum_{i=1}^{n} X_i) \overset{I}{=} \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \overset{ID}{=} \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$$

Επειδή η ανεμενόμενη τιμή του δειγματικού μέσου είναι μ ανεξάρτητα από το μέγεθος του δείγματος και η διακύμανση του δειγματικού μέσου $\sigma^2$/n πηγαίνει στο μηδέν όταν το n τείνει στο άπειρο, η κατανομή του δειγματικού μέσου συγκεντρώνεται εξολοκλήρου στο σημείο μ.

Αυτό το αποτέλεσμα είναι γνωστό ως Νόμος των Μεγάλων Αριθμών και συμβολίζεται με

$$\overline{X} \xrightarrow{p} \mu$$

και διαβάζεται σαν ο μέσος του δείγματος συγκλίνει σε όρους πιθανότητας στο μέσο της κατανομής (mean or expected value).

Ας εξετάσουμε την τυποποιημένη μορφή του δειγματικού μέσου.

$$\frac{\left(\overline{X} - E(\overline{X})\right)}{\sqrt{Var(\overline{X})}} = \frac{\sqrt{n}\left(\overline{X} - \mu\right)}{\sigma}$$

Τότε όταν όταν το n τείνει στο άπειρο,  η κατανομή του τυποποιημένου δειγματικού μέσου πηγαίνει στην Κανονική κατανομή.

Αυτό το αποτέλεσμα είναι γνωστό ως Κεντρικό Οριακό Θεώρημα Central Limit Theorem (CLT) και γράφεται σαν

$$\sqrt{n}\,\frac{\overline{X} - \mu}{\sigma} \xrightarrow{d} N(0,1)$$

και διαβάζεται σαν

$$\sqrt{n}\left(\overline{X} - \mu\right)/\sigma$$

Τότε, η ασυμπτωτική κατανομή του δειγματικού μέσου δίδεται από

$$\overline{X} \overset{a}{\sim} N(\mu, \frac{\sigma^2}{n})$$

Το LLN και το CLT πάντα αναφέρονται στη συμπεριφορά των "*scaled summations*" όπως ο δειγματικός μέσος . Η δήλωση ότι " όλες οι κατανομές συγκλίνουν στη Κανονική κατανομή όταν το μέγεθος του δείγματος αυξάνει " είναι λάθος !

**Θεώρημα Δειγματικού Μέσου (Sample Mean Theorem)**

Έστω $X_j$, $j = 1,2,...,n$ είναι τυχαίο δείγμα (IID) $E(X_i) = \mu, Var(X_i) = \sigma^2$.

Τότε

a. $E(\overline{X}_n) = \mu$

b. $Var(\overline{X}_n) = \dfrac{\sigma^2}{n}$

c. $(\overline{X}_n - \mu) \overset{a}{\sim} N\left(0, \dfrac{\hat{\sigma}_n^2}{n}\right)$ όταν $n \to \infty$,

όπου $\hat{\sigma}_n^2$ είναι η δειγματική διακύμανση (sample variance)

## *2.3. Σημειακή Εκτίμηση – Η Έννοια του Εκτιμητή*

Το πρόβλημα της σημειακής εκτίμησης συνίσταται στην εύρεση της καλύτερης τιμής μιας ή περισσότερων παραμέτρων του πληθυσμού με τη βοήθεια στοιχείων δείγματος που λαμβάνεται (το δείγμα) κατά τρόπο τυχαίο από τον πληθυσμό. Επειδή στην περίπτωση αυτή αναζητούμε μια μόνο τιμή για κάθε άγνωστη παράμετρο, ομιλούμε περί εκτίμησης σημείου, σε αντιδιαστολή με την εκτίμηση διαστήματος που αναφέρεται σε ένα σύνολο υποψήφιων τιμών για κάθε άγνωστη παράμετρο του πληθυσμού. Η προσφυγή στην εκτίμηση διαστήματος εμπιστοσύνης αντί σημείου, συνήθως οφείλεται σε δύο λόγους. Κατά πρώτον, ορισμένες φορές τα δεδομένα δεν επιτρέπουν εκτίμηση σημείου για όλες τις παραμέτρους και κατά δεύτερο λόγο, η εκτίμηση διαστήματος προτιμάται όταν υπάρχει μεγάλη αβεβαιότητα για την εκτίμηση μόνο μιας τιμής.

Έστω ότι έχουμε έναν πληθυσμό που έχει συνάρτηση πυκνότητας πιθανότητας $f(\chi,\theta)$ και θέλουμε να εκτιμήσουμε με έναν αριθμό την άγνωστη παράμετρο $\theta$ του πληθυσμού αυτού με τη βοήθεια ενός τυχαίου δείγματος $X_1,X_2,...,X_i,...,X_n$.

Στην περίπτωση αυτή η εκτίμηση της παραμέτρου $\theta$ επιτυγχάνεται με την εκτέλεση μιας σειράς αριθμητικών πράξεων (μιας διαδικασίας) επί των τιμών των παρατηρήσεων $\chi_1,\chi_2,...,\chi_i,...,\chi_n$ ενός τυχαίου δείγματος και εκφράζεται με έναν τύπο.

Ο τύπος αυτός που δηλώνει τη συγκεκριμένη διαδικασία πράξεων την οποία πρέπει να ακολουθήσουμε, προκειμένου να εκτιμήσουμε την παράμετρο $\theta$, καλείται ***εκτιμητής ή εκτιμήτρια συνάρτηση***. Η συγκεκριμένη τιμή που παίρνει ο εκτιμητής, όταν εφαρμόζεται σε συγκεκριμένο δείγμα παρατηρήσεων λέγεται εκτίμηση της αντίστοιχης παραμέτρου θ του πληθυσμού. Είναι σαφές ότι ο εκτιμητής είναι μια συνάρτηση, ενώ η εκτίμηση αναφέρεται σε συγκεριμένη τιμή της παραμέτρου.

Κατά κανόνα, για την ίδια παράμετρο του πληθυσμού υπάρχουν περισσότεροι από ένας εκτιμητές, δηλαδή υπαρχουν διάφοροι μαθηματικοί τύποι από τους οποίους είναι δυνατός ο υπολογισμός της παραμέτρου που ενδιαφερόμαστε. Αν, για παράδειγμα, αντικείμενο μελέτης είναι ο προσδιορισμός της κεντρικής τάσης της κατανομής, η ζητούμενη παράμετρος μπορεί να υπολογιστεί από τους τύπους του μέσου αριθμητικού, του μέσου γεωμετρικού, της διαμέσου, του τύπου κ.λπ. Δηλαδή, κάθε ένας από τους τύπους αυτούς αποτελεί εκτιμητή για την παράμετρο κεντρικής τάσης. Προφανώς, εμείς αναζητούμε τον καλύτερο μεταξύ των υποψηφίων εκτιμητών μιας παραμέτρου. Για την επιλογή αυτή είναι απαραίτητη η υιοθέτηση κανόνων αξιολόγησης των εκτιμητών. Οι κανόνες αυτοί είναι γνωστοί ως **επιθυμητές ιδιότητες των εκτιμητών.**

Από τα προηγούμενα είναι φανερό ότι κατά την εκτίμηση των τιμών των παραμέτρων του πληθυσμού από στοιχεία τυχαίου δείγματος αντιμετωπίζονται δυο προβλήματα: α) η ανεύρεση των μαθηματικών τύπων από τους οποίους θα υπολογίζονται οι τιμές των άγνωστων παραμέτρων και β) η διατύπωση κανόνων αξιολόγησης των εκτιμητών.

## *2.4. Η Συνάρτηση Ζημιών*

Στην πράξη η επιλογή μεταξύ εκτιμητών δεν είναι εύκολη υπόθεση. Η απαίτηση του να έχει ένας εκτιμητής ορισμένες ιδιότητες προκειμένου να χαρακτηριστεί ως καλός ή ως ο καλύτερος είναι μια λύση, η οποία δεν είναι πάντα εύκολη κατά την εφαρμογή της. Πρώτον, σε πολλές περιπτώσεις είναι αδύνατη η κατασκευή εκτιμητή που να έχει όλες τις επιθυμητές ιδιότητες και δεύτερον, η επιμονή στην επιλογή του καλύτερου εκτιμητή με βάση τις ιδιότητές του ενδεχομένως να συνεπάγεται ιδιαίτερα υψηλό κόστος σε χρόνο και χρήμα.

Με σκοπό να καταστήσουμε τα προηγούμενα περισσότερο κατανοητά θα εισαγάγουμε την έννοια της **συνάρτησης ζημιών (loss function)** , η οποία παίρνει την εξής μορφή:

L= F ( $\left|\hat{\theta}-\theta\right|$ , c ( $\hat{\theta}$ ))    2.1

όπου $|\hat{\theta}-\theta|$ = απόκλιση (λάθος) σε απόλυτους όρους μεταξύ εκτίμησης και αληθινής τιμής της παραμέτρου θ, και c( $\hat{\theta}$ ) = κόστος σε χρόνο και χρήμα που συνεπάγεται η επιλογή του $\hat{\theta}$. Αυτά σημαίνουν ότι οι ζημιές προσδιορίζονται από δύο συνιστώσες, το λάθος $|\hat{\theta}-\theta|$ και το κόστος απόκτησης του $\hat{\theta}$. Έτσι, όταν έχουμε να επιλέξουμε μεταξύ εναλλακτικών εκτιμητών του θ, θα επιλέξουμε εκείνον που συνεπάγεται τη μικρότερη ζημιά. Παρατηρούμε ότι όσο μικρότερο είναι το λάθος $|\hat{\theta}-\theta|$ , τόσο μικρότερη είναι η τιμή της L. Το ίδιο ισχύει και σε σχέση με το κόστος. Σε μαθηματική γλώσσα αυτά δείχνουν ότι και οι δύο μερικές παράγωγοι της (2.1) πρέπει να είναι θετικές.

Όταν $|\hat{\theta}-\theta|$ = 0, έπεται ότι L=0. Η συνιστώσα του κόστους αποτελούσε σημαντικό κριτήριο επιλογής πριν από 20 και πλέον χρόνια που οι υπολογιστικές δυνατότητες ήταν περιορισμένες. Δε συμβαίνει το ίδιο και σήμερα. Μολονότι το κόστος αναμένεται να είναι, δεν μπορεί ωστόσο να αγνοείται. Όπως είναι η (2.1) δεν προσφέρεται για υπολογισμούς, η δε εξειδίκευσή της δεν είναι καθόλου εύκολο έργο. Μια υπόθεση που υιοθετείται συχνά δέχεται ότι τι κόστος c( $\hat{\theta}$ ) είναι ανάλογο του τετραγώνου του στατιστικού κόστους, το οποίο μετριέται από το λάθος $|\hat{\theta}-\theta|$ . Στο βαθμό που η υπόθεση αυτή απέχει από την πραγματικότητα, δεν αναμένεται να επηρεάσει αισθητά το συμπέρασμα αν πράγματι το κόστος δεν είναι υψηλό, όπως αναφέρθηκε ήδη.

Η παραπάνω υπόθεση επιτρέπει να επαναγράψουμε την (2.1) ως εξής:

L= α( $|\hat{\theta}-\theta|$ )$^2$ α>0 (2.2)

όπου α=συντελεστής κόστους σε χρόνο και χρήμα. Η (2.2), η οποία αποτελεί δευτεροβάθμιο προσέγγιση της συνάρτησης ζημιών, έχει πολύ καλό ιστορικό όσον αφορά τις πρακτικές εφαρμογές. Από τη (2.2) είναι σαφές ότι στις περιπτώσεις που δεν υπάρχει απόκλιση (λάθος) μεταξύ εκτίμησης και πραγματικής τιμής του θ, η ζημιά είναι μηδέν. Ακόμη, η ζημιά είναι συνάρτηση του απόλυτου μεγέθους του λάθους, δηλαδή ένα λάθος -5 προκαλεί την ίδια ζημιά με ένα λάθος +5.

## *2.5. Κριτήρια επιλογής μεταξύ εναλλακτικών εκτιμητών*

Το παραπάνω πρόβλημα ξεπερνιέται αν αναζητήσουμε την προσδοκώμενη (μέση) τιμή του $\hat{\theta}$ που ελαχιστοποιεί τη συνάρτηση ζημιών. Λαμβάνοντας την προσδοκώμενη τιμή και των δύο μελών της σχέσης (2.2)

έχουμε

E(L)= αE $[(\hat{\theta}-\theta)^2]$ (2.3)

Στο σημείο αυτό είναι χρήσιμο να εισαγάγουμε τις ακόλουθες έννοιες:

α) **Δειγματοληπτικό σφάλμα (sampling error).**Συμβολίζεται με $\hat{\theta}=\theta$ και αναφέρεται στις αποκλίσεις (λάθη) μεταξύ επιμέρους εκτιμήσεων του $\hat{\theta}$ και της αληθινής τιμής της παραμέτρου θ στον πληθυσμό.

β) **Μεροληπτικό σφάλμα (bias).** Συμβολίζεται με $E(\hat{\theta})-\theta$ και αναφέρεται στην απόκλιση μεταξύ της προσδοκώμενης τιμής E( $\hat{\theta}$ ) του $\hat{\theta}$ και της αληθινής τιμής του θ στον πληθυσμό.

γ) **Μέσο σφάλμα τετραγώνου (mean square error=MSE).** Συμβολίζεται με E $(\hat{\theta}-\theta)^2$ και μετρά τη διασπορά των τιμών του θ γύρω από την αληθινή τιμή της παραμέτρου θ.

Σημειώνεται ότι το μέσο σφάλμα τετραγώνου διαφέρει από τη διακύμανση και συμπίπτουν μόνο όταν $E(\hat{\theta})=\theta$, δηλαδή μόνο όταν δεν υπάρχει μεροληπτικό σφάλμα. Σε όλες τις άλλες περιπτώσεις $MSE(\hat{\theta}) \neq \sigma^2$.

## *2.6. Ιδιότητες Εκτιμητών*

Οι κανόνες αξιολόγησης των εκτιμητών πρέπει να σχετίζονται με τις παραμέτρους της κατανομής δειγματοληψίας αυτών. Οι κανόνες αυτοί αναφέρονται ως **επιθυμητές ιδιότητες των εκτιμητών**. Επειδή το μέγεθος του δείγματος διαδραματίζει πρωτεύοντα ρόλο στη συμπεριφορά της κατανομής δειγματοληψίας, είναι απαραίτητο να γίνει διαφοροποίηση των ιδιοτήτων των εκτιμητών ανάλογα με το μέγεθος του δείγματος. Έτσι διακρίνουμε τις **ιδιότητες δειγμάτων μικρού μεγέθους** από τις **ιδιότητες δειγμάτων μεγάλου μεγέθους**. Οι πρώτες αφορούν σε δείγματα μικρού και σταθερού μεγέθους, ενώ οι δεύτερες χαρακτηρίζονται ως **ασυμπτωτικές ιδιότητες** και αναφέρονται σε δείγματα μεγάλου και μεταβαλλόμενου μεγέθους. Είναι φυσικό να αναμένει κανείς ότι οι δυσκολίες για την ικανοποίηση της απαίτησης να έχουμε πολύ καλούς εκτιμητές είναι μεγαλύτερες στις περιπτώσεις των δειγμάτων μικρού μεγέθους.

α) **Αμεροληψία (unbiasedness)**

Η αμεροληψία αναφέρεται στην απόκλιση που κατά μέσον όρο υπάρχει μεταξύ των δειγματικών τιμών του $\hat{\theta}$ και της πραγματικής τιμής της παραμέτρου στον πληθυσμό. Έτσι θα λέμε ότι ένας εκτιμητής $\hat{\theta}$ της παραμέτρου θ του πληθυσμού είναι αμερόληπτος εάν και μόνο εάν η προσδοκώμενη τιμή του $\hat{\theta}$ ισούται με την αληθινή τιμή του θ, δηλαδή:

$E(\hat{\theta})=\theta$ ή $E(\hat{\theta})-\theta=0$ (2.4)

Η ικανοποίηση της παραπάνω σχέσης διασφαλίζει ότι το **μεροληπτικό σφάλμα** είναι μηδέν. Η τιμή του $\hat{\theta}$ υπολογίζεται ως σταθμικός μέσος όρος των επιμέρους τιμών που αυτός παίρνει χρησιμοποιώντας διάφορα δείγματα από τον ίδιο πληθυσμό. Ως συντελεστές στάθμισης χρησιμοποιούνται οι πιθανότητες να εμφανιστούν οι τιμές του δείγματος. Η ικανοποίηση της ιδιότητας της αμεροληψίας σημαίνει ότι αν λάβουμε από τον ίδιο πληθυσμό διαφορετικά δείγματα και κάθε φορά παίρνουμε μια εκτίμηση $\hat{\theta}$ για τη θ, τότε ο μέσος των εκτιμήσεων αυτών θα πλησιάζει την αληθινή τιμή της παραμέτρου, καθώς ο αριθμός των δειγμάτων αυξάνεται.

β) **Αποτελεσματικότητα (efficiency)**

Η ιδιότητα αυτή αναφέρεται σε αμερόληπτους εκτιμητές και έχει σχέση με τη διασπορά των τιμών του εκτιμητή γύρω από τη μέση τιμή της παραμέτρου. Ειδικότερα θα λέμε ότι **ο $\hat{\theta}$ είναι αποτελεσματικός εκτιμητής του θ αν είναι αμερόληπτος και συγχρόνως έχει τη μικρότερη διακύμανση μεταξύ όλων των αμερόληπτων εκτιμητών του θ.**

γ) **Ελάχιστο μέσο σφάλμα τετραγώνου**

Πολλές φορές είναι αδύνατον να βρούμε αμερόληπτους εκτιμητές, οπότε δεν μπορούμε να ομιλούμε για την αποτελεσματικότητα του εκτιμητή. Σε άλλες περιπτώσεις η εμμονή στη αμεροληψία ενδέχεται να οδηγήσει σε αποτελέσματα, τα οποία είναι θεωρητικά και πρακτικά απαράδεκτα. Για παράδειγμα, οποιαδήποτε παρατήρηση του δείγματος από μόνη της είναι αμερόληπτος εκτιμητής του μέσου του πληθυσμού. Όμως, είναι σωστό να λαβουμε ως εκτιμητή του μέσου την τιμή μιας παρατήρησης του δείγματος; Ασφαλώς όχι. Ακόμη, στην περίπτωση αυτή δεν έχει νόημα να αναζητούμε μεταξύ των επιμέρους παρατηρήσεων του δείγματος εκείνη που έχει τη μικρότερη διακύμανση για τον απλούστατο λόγο ότι όλες έχουν την ίδια διακύμανση.

Είναι προφανές ότι πρέπει να αναζητηθεί πιο ασφαλές κριτήριο επιλογής και ως τέτοιο χρησιμοποιείται το μέσο σφάλμα τετραγώνου (mean square error=MSE), το οποίο αποτελεί από κοινού έκφραση του μεροληπτικού σφάλματος και της διακύμανσης. Έτσι, δοθέντων έστω τριών εκτιμητών $\hat{\theta}1$, $\hat{\theta}2$ και $\hat{\theta}3$ της παραμέτρου θ και ότι MSE( $\hat{\theta}2$ )<MSE( $\hat{\theta}3$ )<MSE( $\hat{\theta}1$ ). Μεταξύ των τριών θα προτιμηθεί ο $\hat{\theta}2$. Η επιχειρηματολογία υπέρ της επιλογής εκτιμητών με κριτήριο το μέσο σφάλμα τετραγώνου είναι ότι κατά μέσο όρο επιλέγουμε εκτιμητές που οι τιμές τους βρίσκονται πλησιέστερα στην αληθινή τιμή των παραμέτρων του πληθυσμού. Σημειώνεται ότι ενδέχεται το MSE του $\hat{\theta}3$ να είναι μικρότερο από τη διακύμανση του $\hat{\theta}1$, ήτοι MSE( $\hat{\theta}3$ ) < var ( $\hat{\theta}1$ ). Βεβαίως δημιουργείται πρόβλημα κατά τον υπολογισμό του MSE, διότι η τιμή του εξαρτάται από την αληθινή τιμή της αντίστοιχης παραμέτρου, η οποία πολλές φορές είναι άγνωστη.

δ) **Επάρκεια (sufficiency)**

Ο εκτιμητής $\hat{\theta}$ θα ονομάζεται επαρκής εκτιμητής της παραμέτρου θ όταν στον $\hat{\theta}$ έχουν συμπυκνωθεί όλες (χωρίς καμιά απώλεια) οι πληροφορίες για την παράμετρο θ που περιέχονται στο δείγμα. Με άλλα λόγια, ο $\hat{\theta}$ εκφράζει όλες τις πληροφορίες για τη θ που υπάρχουν στο δείγμα. Αυτό σημαίνει ότι η από κοινού κατανομή των τιμών ενός δείγματος από τον πληθυσμό f(x;θ), όταν δίνεται η τιμή του $\hat{\theta}$, είναι ανεξάρτητη της παραμέτρου θ. Αξίζει ωστόσο να τονιστεί ότι η επάρκεια από μόνη της δεν εκφράζει επιθυμητή ιδιότητα. Η σημασία της βρίσκεται στο γεγονός ότι αποτελεί αναγκαία συνθήκη για την εξασφάλιση της αποτελεσματικότητας ενός εκτιμητή, καθώς επίσης και στη γενικότερη προσπάθεια που κάνουμε για την εύρεση άριστων εκτιμητών. Ως παράδειγμα επαρκούς εκτιμητή αναφέρεται ο μέσος του δείγματος. Μετά τον υπολογισμό του δειγματικού μέσου δεν απομένουν στο δείγμα πρόσθετες πληροφορίες που η αξιοποίησή τους θα μπορούσε να βοηθήσει στην επίτευξη καλύτερης εκτίμησης για το μέσο του πληθυσμού. Αντίθετα, η διάμεσος δεν αποτελεί επαρκή εκτιμητή της κεντρικής τάσης μιας κατανομής, διότι κατά τον υπολογισμό της δε χρησιμοποιούνται όλες οι πληροφορίες που υπάρχουν στο δείγμα.

## *2.7. Ασυμπτωτικές Ιδιότητες Εκτιμητών*

Οι ασυμπτωτικές ιδιότητες αναφέρονται σε μεγάλα δείγματα, το μέγεθος των οποίων αυξάνεται με κατεύθυνση το άπειρο (τουλάχιστον δυνητικά). Όμως, καθώς αυξάνεται το μέγεθος αλλάζει και η κατανομή δειγματοληψίας των εκτιμητών. Η μορφή που παίρνει τελικά η κατανομή ενός εκτιμητή όταν το μέγεθος του δείγματος τείνει στο άπειρο ονομάζεται **ασυμπτωτική κατανομή**.

Δύο πολύ χρήσιμοι όροι είναι **ο ασυμπτωτικός μέσος** και **η ασυμπτωτική διακύμανση**. Ο ασυμπτωτικός δειγματικός μέσος είναι το όριο της προσδοκώμενης τιμής του, όταν το μέγεθος του δείγματος τείνει στο άπειρο. Η ασυμπτωτική διακύμανση του δειγματικού μέσου ορίζεται ως το όριο της προσδοκώμενης τιμής αυτής, όταν το μέγεθος του δείγματος τείνει στο άπειρο.

Ένας άλλος χρήσιμος όρος είναι η έννοια του ορίου πιθανότητας (probability limit). Ως **όριο πιθανότητας** ενός εκτιμητή ορίζεται η τιμή γύρω από την οποία συγκεντρώνεται η κατανομή του, καθώς το μέγεθος του δείγματος τείνει στο άπειρο. Συμβολικά το όριο πιθανότητας του $\hat{\theta}$ παριστάνεται ως εξής:

$\text{plim}\, \hat{\theta} = \theta_0$ (2.5) και διαβάζεται πλιμ. Το $\theta_0$ είναι η τιμή (σταθερά) γύρω από την οποία συγκεντρώνεται η κατανομή του $\hat{\theta}_n$.

Οι σημαντικότερες ασυμπτωτικές ιδιότητες ενός εκτιμητή είναι: η ασυμπτωτική αμεροληψία, η συνέπεια και η ασυμπτωτική αποτελεσματικότητα.

**α) Ασυμπτωτική αμεροληψία**

Ο εκτιμητής $\hat{\theta}$ της παραμέτρου θ θα ονομάζεται ασυμπτωτικά αμερόληπτος αν η προσδοκώμενη τιμή του προσεγγίζει διαρκώς και περισσότερο την αληθινή της θ, καθώς το μέγεθος του δείγματος n τείνει στο άπειρο.

**β) Συνέπεια**

Θα λέμε ότι ο $\hat{\theta}$ είναι συνεπής εκτιμητής του θ εάν και μόνο εάν η τιμή του θ συγκλίνει στην αληθινή τιμή της παραμέτρου θ με πιθανότητα που πλησιάζει τη μονάδα, καθώς το μέγεθος του δείγματος αυξάνεται χωρίς όριο.

Αξίζει να σημειωθεί ότι για να είναι ένας εκτιμητής συνεπής δεν αρκεί να έχει μεροληπτικό σφάλμα μηδέν. Πρέπει και η διακύμανσή του να τείνει στο μηδέν όταν το μέγεθος του δείγματος αυξάνεται, δηλαδή η κατανομή του εκτιμητή πρέπει να συγκεντρώνεται γύρω από την αληθινή τιμή της παραμέτρου. **Έτσι, ένας αμερόληπτος εκτιμητής δεν είναι αναγκαστικά και συνεπής, ενώ ένας συνεπής εκτιμητής είναι οπωσδήποτε ασυμπτωτικά αμερόληπτος.**

Ας δούμε ένα παράδειγμα:

Έστω ότι η δειγματοληψία γίνεται από πληθυσμό με μέσο μ και διακύμανση $\sigma^2$. Γνωρίζουμε ότι ο $\bar{X} = 1|n \sum_{i=1}^{n} X_1$ είναι αμερόληπτος εκτιμητής του μ και έχει διακύμανση $var(\bar{X}) = \sigma^2 / n$. Είναι προφανές ότι ο $\bar{X}$ είναι και συνεπής εκτιμητής του μ διότι και αμερόληπτος είναι και η διακύμανσή του τείνει στο μηδέν, όταν το n τείνει στο άπειρο, αφού αυτή δίνεται ως λόγος δύο μεγεθών, όπου το ένα στον αριθμητή είναι σταθερό ($\sigma^2$=σταθερό) και το άλλο στον παρονομαστή αυξάνει καθώς αυξάνεται το n.

Όμως αμερόληπτος εκτιμητής του μ είναι και η τιμή μιας παρατήρησης του δείγματος. Πράγματι, $E(X_1)=\mu$, διότι αν πάρουμε πολλά δείγματα, ο μέσος των όλων των τιμών της πρώτης παρατήρησης όλων των δειγμάτων ισούται με μ. Εξάλλου, $var(X_1) = \sigma^2$, ήτοι η διακύμανση της $X_1$ είναι ανεξάρτητη από το μέγεθος του δείγματος. Αφού η διακύμανση της $X_1$ δεν επηρεάζεται από το μέγεθος του δείγματος, δεν υπάρχει ελπίδα να βελτιωθεί ο εκτιμητής αυτός με την αύξηση του μεγέθους του δείγματος.

Ακόμη σημειώνεται ότι η ιδιότητα της συνέπειας διαδραματίζει σπουδαίο ρόλο στην εφαρμοσμένη έρευνα, κυρίως στην οικονομετρική, όπου συχνά είμαστε αναγκασμένοι να χρησιμοποιούμε εκτιμητές που δεν είναι αποτελεσματικοί αλλά είναι συνεπείς.

Τέλος αξίζει να αναφερθεί η εξής ιδιότητα των συνεπών εκτιμητών: έστω $\hat{\theta}$ ένας εκτιμητής του θ και g(θ) μια συνεχής συνάρτηση του θ. Αν $\hat{\theta}$ είναι συνεπής εκτιμητής του θ, τότε η συνάρτηση $g(\hat{\theta})$ είναι συνεπής εκτιμητής της g(θ). Για παράδειγμα, αν $\hat{\theta}$ είναι συνεπής εκτιμητής του θ, τότε και ο $\log\hat{\theta}$ είναι συνεπής εκτιμητής του logθ.

**γ) Ασυμπτωτική αποτελεσματικότητα**

Ο εκτιμητής $\hat{\theta}$ θα ονομάζεται ασυμπτωτικά αποτελεσματικός εκτιμητής της παραμέτρου θ όταν ισχύουν οι εξής συνθήκες:

i. ο $\hat{\theta}$ έχει ασυμπτωτική κατανομή με πεπερασμένο μέσο και πεπερασμένη διακύμανση
ii. ο $\hat{\theta}$ είναι συνεπής εκτιμητής του θ
iii. δεν υπάρχει άλλος εκτιμητής που να ικανοποιεί τις παραπάνω συνθήκες και να έχει μικρότερη ασυμπτωτική διακύμανση από τον θ.

Εδώ πρέπει να σημειωθεί ότι η επιλογή μεταξύ του εκτιμητή με τη μικρότερη ασυμπτωτική διακύμανση μεταξύ όλων των συνεπών εκτιμητών του θ δεν είναι εύκολη υπόθεση. Για το σκοπό αυτό είναι απαραίτητο να γνωρίζουμε το κατώτατο όριο της ασυμπτωτικής διακύμανσης. Το όριο αυτό προσδιορίζεται από την ανισότητα των Gramer-Rao. Όμως αυτό δεν επιτυγχάνεται πάντοτε, ειδικότερα στα μικρά δείγματα.

## *2.8. Μέθοδοι Εύρεσης Εκτιμητών*

### 2.8.1. Εισαγωγή

Με τη διατύπωση των επιθυμητών ιδιοτήτων των εκτιμητών τέθηκαν τα πλαίσια μέσα στα οποία θα αναζητηθεί η εύρεση μαθηματικών τύπων (δηλαδή εκτιμητών) από τους οποίους θα εκτιμούνται οι άγνωστες παράμετροι του πληθυσμού με τη βοήθεια στοιχείων δείγματος. Οι τύποι αυτοί πρέπει να είναι τέτοιοι ώστε να έχουν κατά το δυνατόν όλες ή τις περισσότερες από τις επιθυμητές ιδιότητες ή και όλες που αναφέρθηκαν πιο πάνω. Έτσι, θα υπάρχει κάποιος βαθμός εμπιστοσύνης ότι οι εκτιμήσεις που λαμβάνονται για τις παραμέτρους του πληθυσμού θα βρίσκονται κοντά στις αληθινές τιμές τους. Εκτιμητές είναι δυνατόν να βρεθούν με τη βοήθεια διάφορων θεωρημάτων ή με ad hoc επιλογές. Όμως, στη στατιστική επιστήμη έχουν αναπτυχθεί μέθοδοι εύρεσης εκτιμητών με γενικότερη εφαρμογή. Κατά κανόνα οι μέθοδοι αυτές υιοθετούν κάποια αρχή βελτιστοποίησης και έτσι οδηγούνται σε εκτιμητές που έχουν αρκετές ή και όλες (ανάλογα με τη φύση του προβλήματος που ερευνάται) τις επιθυμητές ιδιότητες.

Οι μέθοδοι εύρεσης εκτιμητών που χρησιμοποιούνται περισσότερο στην πράξη είναι:

α) Η μέθοδος των ροπών (method of moments)

β) Η μέθοδος των ελαχίστων τετραγώνων (method of least squares)

γ) Η μέθοδος της μέγιστης πιθανοφάνειας ( maximun likelihood method)

δ) Η μέθοδος της άριστης γραμμικής αμερόληπτης εκτίμησης (best linear unbiased estimation = BLUE)

ε) Η μέθοδος του Bayes (Bayesian approach)

Στη συνέχεια γίνεται αναφορά στις τέσσερις πρώτες μεθόδους.

## 2.8.2. Η μέθοδος των ροπών

Η μέθοδος των ροπών είναι η αρχαιότερη μέθοδος εκτίμησης και συνίσταται στην εύρεση με βάση τα στοιχεία του δείγματος των ροπών που αντιστοιχούν στις ροπές του ερευνώμενου πληθυσμού και στην εν συνεχεία εξίσωσή τους. Έτσι προκύπτει ένα σύστημα με τόσες εξισώσεις όσες είναι οι παράμετροι του πληθυσμού που πρόκειται να εκτιμηθούν. Η λύση του συστήματος δίνει τις τιμές των άγνωστων παραμέτρων του πληθυσμού ως συναρτήσεις των στοιχείων του δείγματος.

Έστω, για παράδειγμα, ότι $X_1,X_2,...,X_n$ είναι τυχαίο δείγμα από διακριτό πληθυσμό με κατανομή $f(x;\theta_1,\theta_2,...,\theta_\lambda)$, όπου $\theta_i(i=1,2,...,\lambda)$ οι άγνωστοι παράμετροι του πληθυσμού. Η απλή ροπή r τάξης του πληθυσμού είναι:

$$v'_r = E(X^r) = \sum_x X^r f(X;\theta_1,\theta_2,\ldots,\theta_\lambda) \quad (2.6)$$

ενώ η ροπή r τάξης του δείγματος είναι:

$$m'_r = 1/n \sum_{i=1}^{n} X_i^r \quad (2.7)$$

Στη συνέχεια διαμορφώνεται το σύστημα:

$$m'_r = v'_r \; r = 1,2,\ldots,\lambda$$

το οποίο λύνεται ως προς $\theta_1, \theta_2, \ldots, \theta_\lambda$.

*ΠΑΡΑΔΕΙΓΜΑ 2.1 :* Έστω τυχαίο δείγμα $X_1, X_2, \ldots, X_n$ από πληθυσμό με κατανομή $f(x; \mu, \sigma^2)$, όπου $\mu$=μέσος και $\sigma^2$=διακύμανση.

*ΛΥΣΗ:*

Έχουμε :

$$v'_1 = E(X) = \mu$$

$$v'_2 = E(X^2)$$

$$m'_1 = 1/n \sum_{i=1}^{n} X_i = \bar{X}$$

$$m'_2 = 1/n \sum_{i=1}^{n} X_i^2$$

Επίσης γνωρίζουμε ότι $\sigma^2 = E(X^2) - [E(X)]^2 = E(X^2) - \mu^2$ .

Κατά συνέπεια

$$\nu'_2=\sigma^2+\mu^2$$

Έτσι έχουμε το σύστημα:

$\nu'_1=m'_1$ ή $\mu=\bar{X}$

$\nu'_2=m'_2$ ή $\sigma^2+\mu^2=1/n\sum_{i=1}^{n}X_1^2$

Από τη λύση του συστήματος παίρνουμε:

$\hat{\mu}=\bar{X}=1/n\sum_{i=1}^{n}X_1$ (2.8)

$\sigma^2=1/n\sum_{i=1}^{n}X_1^2-\mu^2=1/n\sum_{i=1}^{n}X_1^2-(\bar{X}^2)=1/n\sum_{i=1}^{n}(X_1-\bar{X})^2$ (2.9)

Οι μαθηματικοί τύποι (2.8) και (2.9) είναι εκτιμητές των μ και $\sigma^2$ αντίστοιχα. Πρέπει ωστόσο να τονιστεί ότι οι εκτιμητές που προκύπτουν με τη μέθοδο των ροπών συχνά είναι είναι μεροληπτικοί, όπως συμβαίνει στην περίπτωση της (2.9) που είναι μεροληπτικός εκτιμητής της $\sigma^2$. Ακόμη, είναι δύσκολο να διατυπωθεί συμπέρασμα για την αποτελεσματικότητα των εκτιμητών αυτών.Όμως, αυτοί είναι συνεπείς κάτω από γενικές συνθήκες.Βεβαίως, η μέθοδος των ροπών δεν μπορεί να εφαρμοστεί στις περιπτώσεις που δεν υπάρχουν οι ροπές του πληθυσμού. Επίσης, πολύ δύσκολα εφαρμόζεται όταν τα προβλήματα εκτίμησης εμφανίζουν πολυπλοκότητα.Ειδικότερα στην οικονομετρία σπάνια χρησιμοποιείται η μέθοδος των ροπών για την εξαγωγή εκτιμητών. Πλεονέκτημα της μεθόδου των ροπών είναι η απλότητα που εμφανίζει, γεγονός που επιτρέπει τη γρήγορη και μη δαπανηρή εύρεση εκτιμητών. Όμως, στην εποχή μας, η οποία κυριαρχείται από τη ραγδαία ανάπτυξη της πληροφορικής, το πλεονέκτημα αυτό έχει χάσει μεγάλο μέρος από τη σπουδαιότητά του.

### 2.8.3. Μέθοδος ελαχίστων τετραγώνων

Στο όνομα της μεθόδου των ελαχίστων τετραγώνων, η οποία πρωτοεμφανίστηκε στις εργασίες των Laplace και Gauss, συναντά κανείς μία από τις περισσότερο χρησιμοποιημένες αρχές εκτίμησης σε πρακτικά προβλήματα που καλύπτουν το μεγαλύτερο φάσμα εφαρμογών της Στατιστικής και της Οικονομετρίας. Οι μεγάλες εφαρμογές της μεθόδου αυτής αφορούν την εκτίμηση υποδειγμάτων παλινδρόμησης και ανάλυσης της διακύμανσης.

Η μέθοδος ελαχίστων τετραγώνων αποβλέπει στην κατασκευή εκτιμητών που να έχουν όσο το δυνατόν περισσότερες από τις επιθημητές ιδιότητες. Ως κριτήριο επιλογής υιοθετείται η ελαχιστοποίηση του αθροίσματος των τετραγώνων των αποκλίσεων μεταξύ των δειγματοληπτικών και των θεωρητικών τιμών μιας τυχαίας μεταβλητής. Έτσι, το πρόβλημα ανάγεται στην εύρεση μαθηματικών τύπων που θα δίνουν τιμές για τις άγνωστες παραμέτρους του πληθυσμού, οι οποίες θα ελαχιστοποιούν το άθροισμα των τετραγώνων των λαθών (αποκλίσεων).

Ας θεωρήσουμε τυχαίο δείγμα $X_1,X_2,...,X_n$ από πληθυσμό με μέσο μ και διακύμανση $\sigma^2$ και ας αναζητήσουμε την εύρεση των ροπών της κατανομής του πληθυσμού ως προς την αρχή μηδέν.

Γνωρίζουμε ότι η ροπή r τάξης ως προς την αρχή μηδέν είναι:

$E(X^r)=v'_r, r=0,1,2,...,$ (2.10)

Για την εύρεση εκτιμητών με τη βοήθεια των οποίων θα υπολογίζονται οι $v_r$΄σχηματίζουμε την παράσταση:

$$S = \sum_{i=1}^{n} (X_i^r - v'_r)^2 \qquad (2.11)$$

η οποία εκφράζει το άθροισμα των τετραγώνων των αποκλίσεων ανάμεσα στις δειγματοληπτικές και τις προς εκτίμηση τιμές των άγνωστων παραμέτρων, δηλαδή των ροπών $v_r$. Σύμφωνα με τη μέθοδο ελαχίστων τετραγώνων, η λύση στο πρόβλημά μας βρίσκεται στην ελαχιστοποίηση της (2.11). Αυτό επιτυγχάνεται με την εύρεση της πρώτης παραγώγου της (2.11) ως προς $v'_r$ και την εξίσωσή της με το μηδέν. Στη συνέχεια εξετάζονται οι συνθήκες β΄ τάξης που εδώ απαιτούν η δεύτερη παράγωγος να είναι θετική. Αφού διαπιστωθεί ότι οι συνθήκες β΄τάξης ικανοποιούνται, τότε λύνεται η εξίσωση ή το σύστημα των εξισώσεων που εκφράζουν οι συνθήκες α΄τάξης και λαμβάνονται οι ζητούμενοι εκτιμητές.

Κατά τα ανωτέρω έχουμε:

$$dS/dv'_r = 2\Sigma(X_i^r - v'_r)(-1) = 0$$

Έχουμε

$$2\Sigma(X_i^r - v'_r) = 0 \Rightarrow [\Sigma X_i^r - nv'_r] = 0$$

Άρα

$v'_r = 1/n\, \Sigma X_1^r$

Έτσι, για r=1,2,3,4, παίρνουμε τους εξής εκτιμητές:

$\hat{v}'_1 = 1/n\, \Sigma X_i$ i=1,2,...,n

$\hat{v}'_2 = 1/n\, \Sigma X_i^2$ i=1,2,...,n

$\hat{v}'_3 = 1/n\, \Sigma X_i^3$ i=1,2,...,n

$\hat{v}'_4 = 1/n\, \Sigma X_i^4$ i=1,2,...,n

Το σύμβολο ^ τίθεται για να δηλώσει ότι πρόκειται για εκτιμήσεις των $v_r$, οι οποίες είναι πολύ πιθανό να διαφοροποιούνται από τις αληθινές τους τιμές.

Μπορεί κάποιος να παρατηρήσει ότι παίρνουμε τους ίδιους εκτιμητές για τις απλές ροπές όπως και με τη μέθοδο των ροπών. Αυτό είναι σωστό. Όμως η μέθοδος ελαχίστων τετραγώνων περιέχει μηχανισμό βελτιστοποίησης, η υπεροχή του οποίου αποδεικνύεται κατά την αντιμετώπιση περισσότερο πολύπλοκων προβλημάτων. Στο σημείο αυτό αναφέρουμε απλώς ότι οι εκτιμητές ελαχίστων τετραγώνων έχουν τη μικρότερη διακύμανση μεταξύ όλων των γραμμικών και αμερόληπτων εκτιμητών μιας παραμέτρου. Με άλλα λόγια, στην κλάση αυτή, οι εκτιμητές ελαχίστων τετραγώνων είναι BLUE. Το αποτέλεσμα αυτό είναι πού σημαντικό. Επίσης, κάτω από αρκετά γενικές συνθήκες, αποδεικνύεται ότι αυτοί είναι συνεπείς εκτιμητές. Εδώ πρέπει να σημειωθεί ότι για μεγάλο βαθμό περιπτώσεων εκτίμησης οικονομετρικών υποδειγμάτων, η μέθοδος ελαχίστων τετραγώνων στην αρχική της μορφή δεν κρίνεται επαρκής.

## 2.8.4. Μέθοδος της μέγιστης πιθανοφάνειας

Η μέθοδος της μέγιστης πιθανοφάνειας είναι μία από τις σημαντικότερες μεθόδους σημειακής εκτίμησης. Η αρχή πάνω στην οποία στηρίζεται συνίσταται στην εύρεση εκτιμητών για την εκτίμηση των άγνωστων παραμέτρων του πληθυσμού τέτοιων ώστε να μεγιστοποιείται η πιθανότητα εμφάνισης των τιμών του χρησιμοποιούμενου δείγματος. Δηλαδή, η μέθοδος στοχεύει στη μεγιστοποίηση της συχνότητας εμφάνισης του συγκεκριμένου δείγματος σε μια διαδφικασία επανάληψης της δειγματοληψίας από τον ίδιο πληθυσμό ή από πληθυσμούς που οι κατανομές τους διαφέρουν μόνο ως προς τις παραμέτρους που ενδιαφερόμαστε να εκτιμήσουμε. Έτσι, το χρησιμοποιούμενο δείγμα θα είναι το πλέον αντιπροσωπευτικό του πληθυσμού, γεγονός που διασφαλίζει ότι οι τιμές των άγνωστων παραμέτρων, οι οποίες υπολογίζονται με τα στοιχεία του δείγματος από εκτιμητές μέγιστης πιθανοφάνειας, είναι οι τιμές που έχουν την υψηλότερη πιθανότητα να συμπέσουν με τις αληθινές τιμές των παραμέτρων.

Με σκοπό να καταστήσουμε τα ανωτέρω σαφέστερα, θεωρούμε ότι ο ερευνώμενος πληθυσμός είναι κανονικός και ότι η διακύμανσή του $σ^2$ είναι γνωστή. Άρα, η μόνη παράμετρος του πληθυσμού που παραμένει άγνωστη είναι ο μέσος της κατανομής. Η εκτίμηση του μέσου θα γίνει με στοιχεία δείγματος. Η κατανομή του δείγματος, ως γνωστόν, κυριαρχείται από τα χαρακτηριστικά της κατανομής του πληθυσμού. Η αρχή της μέγιστης πιθανοφάνειας λέγει ότο μεταξύ όλων των κανονικών κατανομών που έχουν γνωστή διακύμανση, η κατανομή από την οποία προέρχεται το συγκεκριμένο δείγμα είναι αυτή που ο μέσος της συμπίπτει με το μέσο του δείγματος, ο οποίος υπολογίζεται από εκτιμητή μέγιστης πιθανοφάνειας.

Η αρχή του νήματος για την εύρεση εκτιμητών μέγιστης πιθανοφάνειας είναι η αλγεβρική διατύπωση της σχετικής κατανομής δειγματοληψίας. Πρόκειται για την από κοινού κατανομή πυκνότητας πιθανότητας όλων των δυνατών δειγμάτων, η οποία εκφράζεται ως συνάρτηση των παραμέτρων της κατανομής του πληθυσμού. Αρχικά, στην κατανομή δειγματοληψίας τα λαμβανόμενα δείγματα είναι τυχαίες μεταβλητές και οι παράμετροι του πληθυσμού είναι (άγνωστες) σταθερές. Στη συνέχεια θεωρούμε τις τιμές ενός συγκεκριμένου δείγματος ως δεδομένες και τις άγνωστες παραμέτρους του πληθυσμού ως μεταβλητές. Έτσι παίρνουμε την κατανομή πυκνότητας πιθανότητας των τιμών του δεδομένου δείγματος ως συνάρτηση όλων των δυνατών τιμών των άγνωστων παραμέτρων. Για την αποφυγή σύγχυσης, μεταξύ της πρώτης και της ερμηνείας χρησιμοποιείται διαφορετική ορολογία. Για την απόδοση της δεύτερης ερμηνείας χρησιμοποιείται διεθνώς ο όρος **πιθανοφάνεια** (likelihood) αντί του όρου **πιθανότητα** (probability). Έτσι, στη συνάρτηση πιθανοφάνειας οι τιμές συγκεκριμένου δείγματος εκφράζονται ως συνάρτηση των πιθανών τιμών των παραμέτρων του πληθυσμού, ενώ στη συνάρτηση πιθανότητας οι άγνωστες παράμετροι του πληθυσμού εκφράζονται ως συνάρτηση των δειγματικών τιμών.

Έστω ότι έχουμε στη διάθεσή μας ένα απλό τυχαίο δείγμα, το οποίο προέρχεται από μονομεταβλητό πληθυσμό με κατανομή $f(x;\ \theta_1,\theta_2)$. Αν θεωρήσουμε τα $\theta_1$ και $\theta_2$ ως σταθερές, παίρνουμε την ακόλουθη συνάρτηση κατανομής του δείγματος (πρώτη ερμηνεία):

$f(x_1,x_2,...,x_n;\ \theta_1,\theta_2) = f(x_1;\ \theta_1,\theta_2)\ f(x_2;\ \theta_1,\theta_2)\ ...\ f(x_n;\ \theta_1,\theta_2)$ (2.12)

Αν τώρα μεταχειριστούμε τις τιμές του δείγματος ως σταθερές και τις $\theta_1,\theta_2$ ως μεταβλητές θα έχουμε τη συνάρτηση πιθανοφάνειας του δείγματος ( δεύτερη ερμηνεία ):

$L(x_1,x_2,...,x_n)=f(x_1;\theta_1,\theta_2)\ f(x_2;\ \theta_1,\theta_2)\ ...\ f(x_n;\theta_1,\theta_2)$ (2.13)

Ο μέσος αριθμητικός του δείγματος είναι αμερόληπτος εκτιμητής του μέσου του πληθυσμού σύμφωνα και με τις τρεις μεθόδους εκτίμησης που αναπτύξαμε στο τμήμα αυτό. Δηλαδή τη μέθοδο των ροπών, τη μέθοδο των ελαχίστων τετραγώνων και τη μέθοδο της μέγιστης πιθανοφάνειας. Βεβαίως, η σύμπτωση αυτή δε συμβαίνει πάντοτε και ειδικότερα όταν αντιμετωπίζουμε πιο πολύπλοκα προβλήματα. Ως προς τον $\theta_2$ ως εκτιμητή της διακύμανσης, είναι εύκολο να αποδειχτεί ότι είναι μεροληπτικός, όπως και στην περίπτωση της μεθόδου των ροπών. Όμως είναι ασυμπτωτικά αμερόληπτος, αφού εξαρτάται από το μέγεθος του δείγματος. Γενικά, μπορεί να δειχθεί ότι οι εκτιμητές μέγιστης πιθανοφάνειας είναι συνεπείς και ασυμπτωτικά αποτελεσματικοί.

## 2.8.5. Η μέθοδος BLUE

Στη μέθοδο της άριστης γραμμικής αμερόληπτης εκτίμησης (best linear unbiased estimation-BLUE) υιοθετούμε τις ιδιότητες που επιθυμούμε να έχουν οι εκτιμητές και στη συνέχεια αναζητούμε μαθηματικούς τύπους από τους οποίους θα παίρνουμε, με τη βοήθεια στοιχείων δείγματος, εκτιμήσεις για τις άγνωστες παραμέτρους του πληθυσμού. Με άλλα λόγια, στην περίπτωση αυτή οι εκτιμητές έχουν τις επιθυμητές ιδιότητες από κατασκευής.

Ειδικότερα, αναζητούμε εκτιμητή $\hat{\theta}$ της παραμέτρου θ τέτοιον ώστε:

α) ο $\hat{\theta}$ εκφράζεται ως γραμμική συνάρτηση των παρατηρήσεων του δείγματος

β) ο $\hat{\theta}$ είναι αμερόληπτος εκτιμητής του θ

γ) ο $\hat{\theta}$ έχει τη μικρότερη διακύμανση μεταξύ όλων των γραμμικών και αμερόληπτων εκτιμητών του θ δ) ο τύπος από τον οποίο υπολογίζεται ο $\hat{\theta}$ δεν περιέχει καμία από τις άγνωστες παραμέτρους του πληθυσμού.

# ΚΕΦΑΛΑΙΟ 3

# Διαστήματα Εμπιστοσύνης

### *3.1. Εκτίμηση παραμέτρου σε διάστημα*

Μέχρι τώρα ασχοληθήκαμε με τις ιδιότητες και τις μεθόδους εύρεσης εκτιμητών σημείου. Θεωρήσαμε ότι για την άγνωστη παράμετρο θ του πληθυσμού υπάρχουν περισσότεροι από ένας εκτιμητές και αναζητήσαμε κανόνες επιλογής έτσι ώστε κάθε φορά,ανάλογα με τα δεδομένα του προβλήματος, να είναι δυνατή η επιλογή του καλύτερου. Στη συνέχεια αναφερθήκαμε σε μερικές από τις πλέον γνωστές μεθόδους κατασκευής εκτιμητών. Ωστόσο, ακόμη και στην περίπτωση που ένας εκτιμητής έχει όλες τις επιθυμητές ιδιότητες, παραμένει το θέμα της ακρίβειας μιας συγκεκριμένης εκτίμησης για δύο τουλάχιστον λόγους. Πρώτον, το αποτέλεσμα της εκτίμησης δεν εξαρτάται μόνο από το πόσο καλός είναι ο τύπος από τον οποίο λαμβάνεται η εκτίμηση, αλλά επηρεάζεται και από την ποιότητα των στοιχείων που χρησιμοποιούνται. Έτσι, ένας καλός τύπος δεν εξουδετερώνει τυχόν αδυναμίες των στοιχείων. Δεύτερον, από τον τρόπο που ορίστηκαν οι επιθυμητές ιδιότητες των εκτιμητών είναι σαφές ότι αυτές λειτουργούν στο πλαίσιο της λογικής ότι υπάρχει δυνατότητα λήψης από τον ίδιο πληθυσμό πολλών δειγμάτων, κάτι που δε συμβαίνει ή που δε γίνεται συνήθως. Κατά κανόνα στις εκτιμήσεις χρησιμοποιούμε στοιχεία ενός δείγματος.

Όμως, όταν λέμε ότι ο $\hat{\theta}$ είναι αμερόληπτος εκτιμητής του θ εννοούμε ότι αν πάρουμε πολλά δείγματα από τον ίδιο πληθυσμό και εκτιμήσουμε π.χ. το μέσο αριθμητικό από κάθε δείγμα, τότε ο μέσος όλων των μέσων αναμένεται να συμπέσει με την αληθινή (άγνωστη) τιμή του μέσου του πληθυσμού. Αυτό φυσικά δε διασφαλίζει ότι η εκτίμηση του μέσου από ένα συγκεκριμένο δείγμα θα ισούται με το μέσο του πληθυσμού. Ακόμη και στην περίπτωση που ο $\hat{\theta}$

είναι εκτιμητής μέγιστης πιθανοφάνειας, που σημαίνει ότι το χρησιμοποιούμενο δείγμα είναι το πλέον αντιπροσωπευτικό, τίποτε δε μας εγγυάται ότι κατά τη δειγματοληψία δεν έγιναν λάθη. Επιπλέον, η γνώση που έχουμε για την κατανομή του πληθυσμού ποτέ δεν είναι πλήρης. Κοντολογίς, μεταξύ μιας εκτίμησης του θ και της αληθινής τιμής του θ αναμένεται να υπάρχει κάποια απόκλιση. Έτσι, το ερώτημα δεν αφορά στην ύπαρξη ή όχι αποκλίσεων, αλλά στο μέγεθός τους. Έστω ότι η απόκλιση (σφάλμα) είναι ε. Επειδή, συνήθως, δεν υπάρχει τρόπος να γνωρίζουμε αν η συγκεκριμένη τιμή του $\hat{\theta}$ υπερεκτιμά ή υποεκτιμά την τιμή του θ, είναι λογικό να δεχθούμε ότι η αληθινή τιμή του θ βρίσκεται στο διάστημα | $\hat{\theta}$ -ε, $\hat{\theta}$ +ε|.

Έτσι, μεταφερόμαστε από την εκτίμηση σημείου στην εκτίμηση διαστήματος, όπου αντί μιας έχουμε ένα ολόκληρο διάστημα τιμών μέσα στο οποίο αναμένεται να βρίσκεται η αληθινή τιμή της παραμέτρου που μας ενδιαφέρει. Το διάστημα αυτό εξαρτάται από το ε. Ασφαλώς το ε σχετίζεται με τη διασπορά τιμών (τη διακύμανση) που εμφανίζει η κατανομή δειγματοληψίας του εκτιμητή. Όσο μικρότερη είναι η διασπορά τιμών γύρω από το μέσο, τόσο μεγαλύτερη είναι η πιθανότητα να πλησιάσει η δειγματική εκτίμηση την αληθινή τιμή της παραμέτρου. Αυτός είναι ο λόγος που κατά τη διατύπωση των ιδιοτήτων των εκτιμητών ζητήσαμε αυτοί να είναι αποτελεσματικοί, δηλαδή να έχουν μικρή διακύμανση. Αν η διασπορά (μέση τυπική απόκλιση) είναι μικρή,το ποσοστό των τιμών του $\hat{\theta}$ που θα βρίσκονται σε μικρή απόσταση από την τιμή του θ θα είναι μεγάλο. Μια από τις τιμές αυτές αναμένεται να είναι και η αληθινή τιμή του θ. Αν όμως η διασπορά είναι μεγάλη, τότε η ίδια αναλογία τιμών του $\hat{\theta}$ θα κατανέμεται σε πολύ μεγαλύτερο διάστημα και, κατά συνέπεια, η πιθανότητα να βρίσκεται η συγκεκριμένη τιμή κοντά στην αληθινή μικραίνει. Είναι σαφές ότι ο βαθμός ακρίβειας ενός εκτιμητή είναι θετική συνάρτηση της τυπικής απόκλισης της κατανομής του, δηλαδή του τυπικού σφάλματος.

Η μελέτη της ακρίβειας των εκτιμητών θα διευκολυνόταν πολύ αν γνωρίζουμε την κατανομή δειγματοληψίας αυτών. Ως γνωστόν, η κατανομή δειγματοληψίας ενός εκτιμητή κυριαρχείται από την κατανομή του πληθυσμού από τον οποίο προέρχεται το δείγμα. Η κατανομή που εμφανίζει πολλά πλεονεκτήματα είναι η κανονική. Βεβαίως, όλοι οι πληθυσμοί που αποτελούν αντικείμενα στατιστικής διερεύνησης δεν ακολουθούν την κανονική κατανομή. Γνωρίζουμε ότι η κατανομή δειγματοληψίας του μέσου του δείγματος $\bar{X}$ ακολουθεί ασυμπτωτικά την κανονική κατανομή (κεντρικό οριακό θεώρημα). Αυτό σημαίνει ότι, καθώς το μέγεθος του δείγματος αυξάνεται, η κατανομή δειγματοληψίας πλησιάζει διαρκώς και περισσότερο την κανονική, ανεξάρτητα από την κατανομή του πληθυσμού. Στις περισσότερες περιπτώσεις, στην πράξη έχουμε ικανοποιητική προσέγγιση για σχετικά μικρές τιμές του n, δηλαδή για μεγέθη δείγματος μόλις μεγαλύτερα του 25.

Με άλλα λόγια για την αξιολόγηση της ακρίβειας των εκτιμητών μπορεί να γίνεται προσφυγή στην κανονική κατανομή, η οποία προσδιορίζεται πλήρως από το μέσο και τη διακύμανση. Όμως, έχουμε αποδείξει ότι $E(\bar{X}_n)=\mu$ για όλα τα είδη πληθυσμών και για όλα τα μεγέθη δείγματος. Αυτό σημαίνει ότι αν υπάρχουν διαφορές μεταξύ των δειγματοληπτικών κατανομών των μέσων, αυτές θα αφορούν μόνο τη διακύμανση $\sigma^2_{(x_n)}$. Οι διαφορές στη $\sigma^2_{(x_n)}$ εξαρτώνται από δύο παράγοντες: α) από τη γνώση ή όχι της διακύμανσης $\sigma^2$ του ερευνώμενου πληθυσμού και β) από το γεγονός αν τα στοιχεία του πληθυσμού είναι πεπερασμένα ή όχι.

Μπορούμε να διακρίνουμε τέσσερις περιπτώσεις:

1. πληθυσμός άπειρος με γνωστή διακύμανση
2. πληθυσμός πεπερασμένος με γνωστή διακύμανση
3. πληθυσμός άπειρος με άγνωστη διακύμανση
4. πληθυσμός πεπερασμένος με άγνωστη διακύμανση

Αξίζει ωστόσο να σημειωθεί ότι η γνώση της διακύμανσης του πληθυσμού αποτελεί θεωρητική υπόθεση. Στην πράξη είναι σπάνιες οι περιπτώσεις όπου το $\sigma^2$ είναι γνωστό.

Η ανάπτυξη της λογικής της εκτίμησης παραμέτρου σε διάστημα δεν επηρεάζεται από τη γνώση ή μη της τιμής του $\sigma^2$, αλλά ούτε και από το γεγονός αν ο πληθυσμός είναι ή όχι πεπερασμένος. Γι' αυτό θα επιχειρήσουμε πρώτα μια γενική διατύπωση του προβλήματος και στη συνέχεια θα εξετάσουμε χωριστά τις παραπάνω τέσσερις περιπτώσεις.

Με σκοπό την απλούστερη δυνατή παρουσίαση του θέματος θεωρούμε ότι το δείγμα προέρχεται από κανονικό πληθυσμό με γνωστή διακύμανση. Στις περιπτώσεις που ο πληθυσμός δεν είναι κανονικός θεωρούμε ότι το δείγμα είναι αρκετά μεγάλο ώστε να δικαιολογείται η προσφυγή στην κανονική κατανομή. Έστω $\hat{\theta}$ ο εκτιμητής της παραμέτρου θ (εδώ το θ είναι ο μέσος). Στη συνέχεια σχηματίζουμε την τυχαία μεταβλητή Z τέτοια ώστε:

$$Z=(\hat{\theta}-\theta)/\sqrt{var(\hat{\theta})} \quad (3.1)$$

όπου $\sqrt{var(\hat{\theta})}$ η τυπική απόκλιση του $\hat{\theta}$. Από τα προηγούμενα γνωρίζουμε ότι η Z ακολουθεί την τυπική κανονική κατανομή με μέσο μηδέν και διακύμανση τη μονάδα, ήτοι Z-> N(0,1). Ακόμη, η κατανομή f(z) της Z είναι συμμετρική και ανεξάρτητη από το θ.

Το πρόβλημά μας συνίσταται στην εύρεση ενός διαστήματος τιμών μέσα στο οποίο αναμένεται να βρίσκεται η τιμή του θ με ορισμένη πιθανότητα. Προς τούτο θεωρούμε δυο τυχαίες τιμές $z_1$ και $z_2$ της Z, τέτοιες ώστε $z_1<z_2$, μεταξύ

των οποίων αναμένεται να βρίσκεται η θ με πιθανότητα 1-α. Οι πιθανότητες των $z_1$,$z_2$ είναι $f(z_1)$ και $f(z_2)$ αντίστοιχα. Συμβολικά γράφουμε:

$$P(z_1<Z<z_2)= 1-\alpha \quad (3.2)$$

Με τις κατάλληλες πράξεις και αντικαταστάσεις από τις δύο παραπάνω σχέσεις παίρνουμε την παρακάτω σχέση:

$$\hat{\theta}-z_2\sqrt{var(\hat{\theta})}<\theta<\hat{\theta}-z_1\sqrt{var(\hat{\theta})} \quad (3.3)$$

**Το εύρος των τιμών του θ που προσδιορίζεται από τη διπλή ανισότητα ονομάζεται εκτίμηση του θ σε διάστημα ή διάστημα εμπιστοσύνης του θ με πιθανότηατα 1-α,** ενώ τα δύο όρια ονομάζονται **όρια εμιστοσύνης (confidence limits).** Η έννοια της σχέσης (3.3) είναι ότι αν ληφθούν πολλά δείγματα του ίδιου μεγέθους από κανονικό πληθυσμό και κάθε φορά υπολογίζεται το διάστημα, τότε τα 100(1-α) τοις εκατό των διαστημάτων αυτών θα περιέχουν την αληθινή τιμή του θ. Έτσι, αν λάβουμε 1-α=0.95, τότε στις 95 από τις 100 φορές που θα εκτιμηθεί το διάστημα από την (3.3), η αληθινή τιμή του θ θα περιλαμβάνεται μέσα σε αυτό. Ορισμένες φορές διαβάζουμε την εξής διατύπωση: η αληθινή τιμή της άγνωστης παραμέτρου θ περιλαμβάνεται μεταξύ των δύο ορίων με πιθανότητα 95%. Η διατύπωση αυτή αποτελεί δήλωση πιθανότητας και έχει νόημα μόνο όταν αναφέρεται σε μεταβλητές και όχι σε σταθερές, δηλαδή όταν γίνεται πριν υπολογιστούν τα όρια του διαστήματος εμπιστοσύνης. Από τη στιγμή που με τη βοήθεια συγκεκριμένου δείγματος εκτιμηθούν τα όρια του διαστήματος, τότε η πιθανότητα να βρίσκεται μέσα στο υπολογισμένο διάστημα η τιμή του θ είναι 1 ή 0, διότι η τιμή του θ είναι δεδομένη και κατά συνέπεια είτε βρίσκεται είτε δε βρίσκεται στο διάστημα αυτό.

Το πλεονέκτημα του υπολογισμού διαστημάτων εμπιστοσύνης για την τιμή μιας άγνωστης παραμέτρου βρίσκεται στο γεγονός ότι έτσι έχουμε μια αίσθηση του πόσο κοντά στην αληθινή τιμή της παραμέτρου αναμένεται να είναι η εκτίμηση που παίρνουμε με τη βοήθεια στοιχείων δείγματος. Τέτοια πληροφορία δεν έχουμε στην περίπτωση της σημειακής εκτίμησης. Ακόμη, επειδή το διάστημα περιλαμβάνει ένα σύνολο τιμών, μας παρέχει αυξημένη εμπιστοσύνη ότι η αληθινή τιμή της παραμέτρου που αναζητάμε θα είναι μία από τις τιμές του διαστήματος.

### *3.2. Διάστημα εμπιστοσύνης του μέσου*

Σύμφωνα με όσα προαναφέραμε, διακρίνουμε τις παρακάτω τέσσερις περιπτώσεις:

α)**Γνωστή διακύμανση και άπειρος πληθυσμός:** Για τον προσδιορισμό του διαστήματος εμπιστοσύνης του μέσου μ, όταν είναι γνωστή η διακύμανση $\sigma^2$ και ο πληθυσμός από τον οποίο λαμβάνεται το δείγμα είναι άπειρος, ακολουθούμε τα εξής βήματα:

1.Βρίσκουμε το μέσο $\bar{X}_n$ με τη βοήθεια στοιχείων δείγματος.

2.Εκτιμούμε την τυπική απόκλιση του $\bar{X}_n$ από τον τύπο $\sigma_{x_n} = \sigma / \sqrt{n}$

3.Επιλέγουμε το συντελεστή εμπιστοσύνης (confidence coefficient) 1-α και προσδιορίζουμε την κριτική τιμή της Z που αντιστοιχεί σε εμβαδόν (ή πιθανότητα) (1-α)/2 από τους πίνακες της τυπικής κανονικής κατανομής, η οποία, ως γνωστόν, είναι συμμετρική γύρω από το μηδέν.

4.Διαμορφώνουμε το πλαίσιο του ζητούμενου διαστήματος εμπιστοσύνης. Για το σκοπό αυτό αντικαθιστούμε στην (3.2) την τιμή

$Z = (\bar{X}_n - \mu) / (\sigma - \sqrt{n})$

και κάνουμε τους αναγκαίους μετασχηματισμούς. Τελικά το ζητούμενο διάστημα είναι:

$$\bar{X}_n - Z_\alpha \sigma / \sqrt{n} < \mu < \bar{X}_n + Z_\alpha \sigma / \sqrt{n} \quad (3.4)$$

όπου $Z_\alpha$ η κριτική τιμή της Ζ που αντιστοιχεί στην πιθανότητα ή το εμβαδόν (1-α)/2.

*<u>ΠΑΡΑΔΕΙΓΜΑ 3.1:</u>*

Από δείγμα 225 εργαζομένων στην κλωστοϋφαντουργία προέκυψε ότι το μέσο ημερομίσθιο είναι 4250 δρχ. Ακόμη είναι γνωστό (από προηγούμενες έρευνες ή από τη νομοθεσία) ότι η μέση απόκλιση τετραγώνου των ημερομισθίων στον πληθυσμό των εργαζομένων στον κλάδο αυτό είναι σ=210 δρχ. Να προσδιοριστεί 95% διάστημα εμπιστοσύνης μέσα στο οποίο αναμένεται να βρίσκεται το μέσο ημερομίσθιο μ στον πληθυσμό.

*<u>ΛΥΣΗ:</u>*

Σύμφωνα με τα δεδομένα του προβλήματος έχουμε:

$$\bar{X}_n = 4250 \; \kappa\alpha\iota \; \sigma_x = \sigma / \sqrt{n} = 210 / \sqrt{225} = 14$$

Εξάλλου έχουμε 1-α=0.95 και (1-α)/2=0.95/2=0.4750

Η κριτική τιμή του Ζ που αντιστοιχεί σε εμβαδόν (πιθανότητα) 0.4750 είναι $Z_\alpha$= ± 1.96 . Άρα το ζητούμενο διάστημα είναι:

4250-1.96x14<μ<4250+1.96x14

ή 4250-27.44<μ<4250+2744

ή $4222.56<\mu<4277.44$

Έτσι, το διάστημα [4222.56 , 4277.44] αναμένεται να περιέχει την αληθινή τιμή του μέσου ημερομισθίου στον πληθυσμό με πιθανότητα 95%, ενώ η πιθανότητα να βρίσκεται το μ έξω από το διάστημα είναι α=1-0.95=0.05. Και λόγω της συμμετρικότητας, η πιθανότητα να είναι το μ μικρότερο του 4222.56 είναι 2.5%, ενώ η πιθανότητα να είναι μεγαλύτερο του 4277.49 είναι επίσης 2.5%.

Με δοσμένο το μέγεθος του δείγματος n και την τιμή του $\bar{X}_n$, το διάστημα εμπιστοσύνης μεταβάλλεται ανάλογα με την τιμή του συντελεστή εμπιστοσύνης 1-α. Αν, για παράδειγμα, αντί του 1-α=0.95 χρησιμοποιήσουμε το 1-α=0.90, τότε η τιμή $Z_0$ που αντιστοιχεί στην πιθανότητα 0.90/2=0.45 είναι $z_\alpha=\pm1.65$ περίπου. Κατά συνέπεια το νέο διάστημα εμπιστοσύνης θα είναι:

$4250-1.65x14<\mu<4250+1.65x14$

ή $4226.9<\mu<4273.1$

Άρα βλέπουμε ότι μείωση του συντελεστή εμπιστοσύνης 1-α από 0.95 σε 0.90, είχε ως αποτέλεσμα τη μείωση του διαστήματος εμπιστοσύνης από

[4222.56 , 4277.44] σε [4226.9 , 4273.1]

β)**Γνωστή διακύμανση και πεπερασμένος πληθυσμός:** Εφόσον η διακύμανση είναι γνωστή, το γεγονός ότι ο πληθυσμός είναι πεπερασμένος επηρεάζει μόνο την τυπική απόκλιση του $\bar{X}_n$. Πράγματι, από τα προηγούμενα γνωρίζουμε ότι, στις περιπτώσεις που το δείγμα προέρχεται από πεπερασμένο πληθυσμό χωρίς επανατοποθέτηση, η τυπική απόκλιση του $\bar{X}_n$ τείνει να είναι μικρότερη από την ποσότητα $\sigma/\sqrt{n}$. Η διαφορά εξαρτάται από το μέγεθος του πληθυσμού σε σχέση με το μέγεθος του δείγματος. Ειδικότερα, η $\sigma_X$ υπολογίζεται από τον τύπο:

$\sigma_X = \sigma / \sqrt{n} \sqrt{(N-n)/(N-1)}$ , όπου N=μέγεθος πληθυσμού

Ο όρος $\sqrt{(N-n)/(N-1)} < 1$ και κατά συνέπεια $\sigma_x < \sigma/\sqrt{n}$. Αυτό σημαίνει ότι η αγνόηση της διόρθωσης με $\sqrt{(N-n)/(N-1)}$ στις περιπτώσεις που αναφερόμαστε εδώ, οδηγεί στη διαμόρφωση μη σωστών διαστημάτων εμπιστοσύνης. Βεβαίως, αν το N είναι πολύ μεγάλο σε σχέση με το μέγεθος του δείγματος, η διόρθωση δεν παίζει σημαντικό ρόλο, ενώ όταν το n τείνει στο N, ο όρος $\sqrt{(N-n)/(N-1)}$ τείνει στο μηδέν με αποτέλεσμα και το $\sigma_X$ να τείνει στο μηδέν. Αυτό είναι αναμενόμενο, αφού όταν το μέγεθος του δείγματος πλησιάζει το μέγεθος του πληθυσμού, η εκτίμηση $\bar{X}_n$ αντιστοιχεί στην τιμή του μ.

Με βάση τα δεδομένα του προηγούμενου παραδείγματος (παράδειγμα 3.1) και με την πρόσθετη πληροφορία ότι το μέγεθος του πληθυσμού είναι 1000, έχουμε:

$$\sigma_{x_n} = \sigma/\sqrt{n}\sqrt{(N-n)/(N-1)} = 210/\sqrt{225}\sqrt{(1000-225)/(1000-1)} = 12.33$$

Έτσι, το διάστημα εμπιστοσύνης του μέσου για συντελεστή εμπιστοσύνης 0.95 είναι:

4250-1.96x12.33<μ<4250+1.96x12.33

4225.83<μ<4274.17

Και για συντελεστή εμπιστοσύνης 0.90 είναι:

4250-1.65x12.33<μ<4250+1.65x12.33

4224.66<μ<4270.34

γ) **Άγνωστη διακύμανση και άπειρος πληθυσμός:** Με άγνωστες τις παραμέτρους μ και $\sigma^2$, το πρόβλημα της εκτίμησης περιπλέκεται από το γεγονός ότι δεν μπορούμε να εκτιμήσουμε το $\sigma_X$ από τον τύπο $\sigma/\sqrt{n}$. Η γνώση του $\sigma^2$ επιτρέπει την κατασκευή της τυχαίας μεταβλητής $Z=(\bar{X}-\mu)/(\sigma/\sqrt{n})$ και την προσφυγή στους πίνακες της τυπικής κανονικής κατανομής. Όταν το $\sigma^2$ είναι άγνωστο, ο μόνος τρόπος να το εκτιμήσουμε είναι από τον τύπο $S^2=1/(N-1)\Sigma(X_i-\bar{X})^2$. Στη συνέχεια υπολογίζεται η τυπική απόκλιση του $\bar{X}_n$ από τη σχέση $S_{x_n}=S/\sqrt{n}$, η οποία μας επιτρέπει να σχηματίζουμε την τυχαία μεταβλητή:

$t=(\bar{X}_n-\mu)/(S/\sqrt{n})$ (3.5)

Ως γνωστόν, η μεταβλητή αυτή ακολουθεί την κατανομή t του student με n-1 βαθμούς ελευθερίας, η οποία αποτελεί ικανοποιητική προσέγγιση της κανονικής κατανομής για μεγέθη δείγματος μεγαλύτερα του 25 (δηλαδή n>25). Έτσι, για την κατασκευή διαστήματος εμπιστοσύνης, αντί της κατανομής Ζ προσφεύγουμε στην κατανομή t του student , οι τιμές της οποίας είναι επίσης πινακοποιημένες. Αυτό επιτυγχάνεται με την κατασκευή του διαστήματος:

$\bar{X}_n-t_\alpha S/\sqrt{n}<\mu<\bar{X}_n+t_\alpha S/\sqrt{n}$ (3.6)

όπου $t_\alpha$ είναι η κριτική τιμή της t, η οποία προσδιορίζεται από το συντελεστή εμπιστοσύνης 1-α ή το α και από τους βαθμούς ελευθερίας που είναι n-1.

*<u>ΠΑΡΑΔΕΙΓΜΑ 3.2:</u>*

Από τυχαίο δείγμα 24 πρωτοετών φοιτητών του Οικονομικού Πανεπιστημίου Φοιτητών βρέθηκε ότι η μέση βαθμολογία είναι $\bar{X}_{24}=69.8$ μονάδες με άριστα το 100. Επιπλέον, η διακύμανση του πληθυσμού υπολογίστηκε σε 48 μονάδες, ήτοι $S^2$=48. Να κατασκευαστεί 95% διάστημα εμπιστοσύνης μέσα στο οποίο αναμένεται να βρίσκεται η μέση επίδοση του συνόλου των πρωτοετών φοιτητών.

*<u>ΛΥΣΗ:</u>*

Τα βήματα για την εκτίμηση του διαστήματος εμπιστοσύνης είναι τα εξής:

1.Εκτιμούμε τη μέση απόκλιση τετραγώνου του $\bar{X}_{24}$, ήτοι:

$$S_{X_{24}}=\sqrt{S^2/n}=\sqrt{48/24}=1.414$$

2.Αναζητούμε την κριτική τιμή $t_{0.95}$ στους 24-1=23 βαθμούς ελευθερίας. Επειδή, 1-α=0.95, έπεται ότι α=0.05, πράγμα που σημαίνει ότι η πιθανότητα να πάρουμε τιμή μικρότερη του κάτω ορίου ή μεγαλύτερη του άνω ορίου του διαστήματος είναι α/2=0.025. Με 23 βαθμούς ελευθερίας και με πιθανότητα 0.025 βρίσκουμε ότι η κριτική τιμή είναι 2.069.

3.Με βάση τη σχέση (3.6), το διάστημα εμπιστοσύνης είναι:

$69.8-2.069 \times 1.414<\mu<69.8+2.069 \times 1.414$

ή $66.87<\mu<72.73$

δ) **Άγνωστη διακύμανση και πεπερασμένος πληθυσμός:** Στην περίπτωση που η διακύμανση είναι άγνωστη και ο πληθυσμός από τον οποίο λαμβάνεται το δείγμα είναι πεπερασμένος, τότε πρώτα εκτιμούμε την $S^2$ από τα δεδομένα του δείγματος, ήτοι

$$S^2 = 1/(n-1)\Sigma(X_i - \bar{X})^2$$

και στη συνέχεια υπολογίζουμε το $S_x$ από τον τύπο:

$$S_x = S/\sqrt{n}\sqrt{(N-n)/(N-1)}$$

Αν στο προηγούμενο παράδειγμα δεχθούμε ότι το δείγμα των 24 φοιτητών προερχόταν από τους πρωτοετείς του Τμήματος Εφαρμοσμένης Πληροφορικής που είναι 120, τότε θα έχουμε:

$$S_x = \sqrt{48}/\sqrt{24}\sqrt{(120-24)/(120-1)} = \sqrt{2}\sqrt{96/119} = 1.270$$

Έτσι, το μόνο που αλλάζει, σε σχέση με το προηγούμενο παράδειγμα, είναι η τιμή του $S_x$. Κατά συνέπεια, το 95% διάστημα εμπιστοσύνης του μέσου μ των πρωτοετών του Τμήματος Εφαρμοσμένης Πληροφορικής είναι:

$69.8-2.069x1.270<\mu<69.8+2.069x1.270$

ή $67.17<\mu<72.43$ .

## 3.3. Διάστημα εμπιστοσύνης διακύμανσης

Έστω ότι έχουμε έναν κανονικό πληθυσμό, του οποίου επιθυμούμε να εκτιμήσουμε την άγνωστη διακύμανση. Για το σκοπό αυτό παίρνουμε ένα δείγμα n μονάδων, δηλαδή $X_1, X_2, X_3, ..., X_n$. Για την εκτίμηση της διακύμανσης με τη βοήθεια του παραπάνω δείγματος, διακρίνουμε δύο περιπτώσεις:

α) **Εκτίμηση της διακύμανσης ενός κανονικού πληθυσμού με γνωστό το μέσο του πληθυσμού:** Αν η μεταβλητή X του πληθυσμού κατανέμεται κανονικά με άγνωστη διακύμανση $\sigma^2$ και γνωστό μέσο μ, τότε παίρνοντας διαδοχικά τυχαία δείγματα μεγέθους n και εκτιμώντας τις κάθε φορά διακυμάνσεις, έχουμε:

$$\chi^2 = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{\sigma^2} \sim \chi^2_n$$

Δηλαδή η μεταβλητή ακολουθεί την κατανομή $\chi^2$ με βαθμούς ελευθερίας n και όχι n-1, γιατί ο μέσος υποτίθεται γνωστός.

Η ιδιότητα αυτή μας χρησιμεύει στο να προσδιορίσουμε τα όρια του διαστήματος εμπιστοσύνης μέσα στα οποία περιμένουμε ότι θα βρίσκεται η διακύμανση $\sigma^2$ του πληθυσμού, με δοσμένη πιθανότητα.

Από του πίνακες της κατανομής $x^2$ με n βαθμούς ελευθερίας βρίσκουμε δύο τιμές $x^2_{n,1-\alpha/2}$ και $x^2_{n,\alpha/2}$ που αντιστοιχούν σε πιθανότητες 1-α/2 και α/2 και αφήνουν στα δεξιά τους επιφάνειες 1-α/2 και α/2 αντίστοιχα.

Το διάστημα εμπιστοσύνης της διακύμανσης $\sigma^2$ με συντελεστή εμπιστοσύνης (1-α)100% θα είναι:

$$\left( \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{\chi^2_{n,\alpha/2}}, \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{\chi^2_{n,1-\alpha/2}} \right) \quad (3.7)$$

Επίσης ισχύουν και οι παρακάτω σχέσεις για πιθανότητες 1-α και α αντίστοιχα:

$$\left( 0, \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{\chi^2_{n,1-\alpha}} \right) \quad (3.8)$$

$$\left( \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{\chi^2_{n,\alpha}}, +\infty \right) \quad (3.9)$$

*<u>ΠΑΡΑΔΕΙΓΜΑ 3.3:</u>*

Έστω ότι από έναν κανονικό πληθυσμό έχει ληφθεί δείγμα 4 μονάδων και οι συγκεκριμένες παρατηρήσεις είναι: 50, 60, 48 και 74. Αν ο μέσος του πληθυσμού είναι μ=50, ζητείται να εκτιμηθεί το 95% διάστημα εμπιστοσύνης της διακύμανσης.

*<u>ΛΥΣΗ:</u>*

Το διάστημα της διακύμανσης δίνεται από τη σχέση (3.7).Με αντικατάσταση στον αριθμητή, παίρνουμε:

$$\Sigma(x_i-\mu)^2=(50-60)^2+(60-60)^2+(48-60)^2+(74-60)^2=440$$

Από τους πίνακες της κατανομής $x^2$ με n=4 βρίσκουμε δύο τιμές:

$x^2_{4,0.025}=11.14$

και

$x^2_{4,0.975}=0.48$

Επομένως θα έχουμε

$440/11.14 \leq \sigma^2 \leq 440/0.48$ ή $39.5 \leq \sigma^2 \leq 916.7$

Άρα, ενώ η αληθινή τιμή $\sigma^2$ είναι άγνωστη, γνωρίζουμε τώρα ότι 95% των δειγμάτων θα έχουν διακύμανση μεταξύ των τιμών 39.5 και 916.7 .

β)**Εκτίμηση της διακύμανσης ενός κανονικού πληθυσμού με άγνωστο το μέσο του πληθυσμού:** Έστω ότι θέλουμε να εξετάσουμε το διάστημα εμπιστοσύνης της διακύμανσης $\sigma^2$ ενός κανονικού πληθυσμού, του οποίου δε γνωρίζουμε ούτε το μέσο μ, ούτε τη διακύμανση $\sigma^2$.

Γνωρίζουμε ότι η τυχαία μεταβλητή

$$\frac{(n-1)S^2}{\sigma^2}$$

ακολουθεί την κατανομή $X^2$ με n-1 βαθμούς ελευθερίας.

Έχουμε λοιπόν:

$$X^2_{1-\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \Rightarrow \sigma^2 \leq \frac{(n-1)S^2}{X^2_{1-\alpha/2}}$$

και

$$\frac{(n-1)S^2}{\sigma^2} \leq X^2_{\alpha/2} \Rightarrow \sigma^2 \geq \frac{(n-1)S^2}{X^2_{\alpha/2}}$$

Μετά από αντικατάσταση, λαμβάνουμε:

$$P\left( \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} \right) = 1 - a \quad (3.10)$$

Η (3.10) δίνει το διάστημα μέσα στο οποίο αναμένεται να βρίσκεται η αληθινή τιμή του $\sigma^2$ με συντελεστή εμπιστοσύνης 1-α.

*ΠΑΡΑΔΕΙΓΜΑ 3.4:*

Έστω δείγμα μεγέθους n=25 από κανονικό πληθυσμό με μέσο μ και διακύμανση $\sigma^2$. Το δείγμα έχει δειγματική διακύμανση $S^2 = 12$. Να υπολογιστεί διάστημα εμπιστοσύνης 95% για την άγνωστη τιμή του $\sigma^2$.

*ΛΥΣΗ:*

Εδώ έχουμε n-1=25-1=24 βαθμούς ελευθερίας και 1-α=0,95. Άρα, α=0.05, α/2=0.025 και 1-α/2=0.975. Για 24 βαθμούς ελευθερίας, οι κριτικές τιμές που μας χρειάζονται είναι:

$$X^2_{0.975}=12.401$$

και

$$X^2_{0.025}=39.364$$

.

Μετά από αντικατάσταση στην (3.10), παίρνουμε:

$$P[(25-1)(12)/39.364\leq\sigma^2\leq(25-1)(12)/12.401]=0.95$$

ή

$$P(7.32\leq\sigma^2\leq23.22)=0.95 \ .$$

Κοντολογίς, η άγνωστη τιμή του $\sigma^2$ αναμένεται να βρίσκεται στο διάστημα [7.32, 23.22] με συντελεστή εμπιστοσύνης 95%. Για τον προσδιορισμό του διαστήματος έγινε χρήση των πινακοποιημένων τιμών της κατανομής $X^2$ με α/2 = 0.025, 1 – α/2 = 0.975 και n – 1 βαθμούς ελευθερίας. Αυτό προϋποθέτει ότι η τυχαία μεταβλητή

$$\frac{(n-1)\,S^2}{\sigma^2}$$

ακολουθεί την κατανομή $X^2$ με n – 1 βαθμούς ελευθερίας.

Αυτό συμβαίνει μόνο όταν ο πληθυσμός από τον οποίο προέρχεται το δείγμα είναι κανονικός ή στις περιπτώσεις μη κανονικού πληθυσμού όταν έχουμε πολύ μεγάλο δείγμα ( π.χ. n>100). Σε κάθε άλλη περίπτωση η κατανομή $X^2$ δεν αποτελεί κατάλληλο μοντέλο της κατανομής δειγματοληψίας της μεταβλητής

$$\frac{(n-1)\,S^2}{\sigma^2}.$$

Μετά την εκτίμηση του διαστήματος εμπιστοσύνης για την τιμή του $\sigma^2$, μπορούμε να υπολογίσουμε και το διάστημα εμπιστοσύνης της τυπικής απόκλισης αν πάρουμε τις τετραγωνικές ρίζες των ορίων του διαστήματος της $\sigma^2$,

δηλαδή:

$$\sqrt{7.32} \le \sigma \le \sqrt{23.22}$$

ή

$2.71 \le \sigma \le 4.82$.

*ΠΑΡΑΔΕΙΓΜΑ 3.5:*

Από έναν κανονικό πληθυσμό, του οποίου δε γνωρίζουμε τη διακύμανση και το μέσο αριθμητικό, παίρνουμε ένα δείγμα n=21, για να βρούμε ένα διάστημα εμπιστοσύνης της διακύμανσης $\sigma^2$ με συντελεστή εμπιστοσύνης 95%. Αν η δειγματική διακύμανση είναι $S^2=16$, να βρεθεί το διάστημα.

*ΛΥΣΗ:*

Από τους πίνακες της κατανομής $x^2$ για n-1=21-1=20 και α=5% θα έχουμε $\chi^2_{20,0.025}=34.17$ και $\chi^2_{20,0.975}=9.59$

Επομένως θα έχουμε από τη σχέση (3.10) ότι

$(21-1)16/34.17 \leq \sigma^2 \leq (21-1)16/9.59$

ή

$9.36 < \sigma^2 < 33.37$.

## *3.4. Διάστημα εμπιστοσύνης για μια πιθανότητα ή ποσοστό*

Η μέθοδος για τον προσδιορισμό ενός διαστήματος εμπιστοσύνης για την άγνωστη πιθανότητα p=P(A) ενός ενδεχομένου A συνδέεται στενά με τον διωνυμικό έλεγχο.

Ένα δείγμα παρατηρήσεων πάνω σε n ανεξάρτητες δοκιμές εξετάζεται και καταγράφεται η συχνότητα T με την οποία το συγκεκριμένο ενδεχόμενο εμφανίζεται. Οι n δοκιμές είναι αμοιβαία ανεξάρτητες και η πιθανότητα p του εν λόγω ενδεχομένου παραμένει σταθερή από δοκιμή σε δοκιμή. Είναι προφανές, ότι η κατανομή της τυχαίας μεταβλητής T είναι η διωνυμική με παραμέτρους n και p. Ένα διάστημα εμπιστοσύνης για την πιθανότητα p είναι προφανές ότι θα αποτελείται από όλες τις τιμές του $p_0$ οι οποίες είναι τέτοιες ώστε τα δεδομένα του δείγματος θα οδηγούσαν σε μη απόρριψη της μηδενικής υπόθεσης

$$H_0 : p = p_0$$

έναντι της εναλλακτικής

$$H_1 : p \neq p_0 .$$

Ειδικότερα, εάν επιθυμούμε την κατασκευή ενός 100(1–α)% διαστήματος εμπιστοσύνης, παρατηρούμε το δείγμα των αποτελεσμάτων των n δοκιμών, καταγράφουμε την τιμή τ της μεταβλητής T και ρωτάμε:

*«Για την δοθείσα τιμή τ της μεταβλητής T, ποιες είναι οι τιμές εκείνες που θα μπορούσε να πάρει το $p_0$ στην υπόθεση*

$$H_0 : p = p_0$$

ώστε ένας αμφίδρομος διωνυμικός έλεγχος σε επίπεδο σημαντικότητας α να *μην οδηγήσει σε απόρριψη της μηδενικής υπόθεσης; »*

Το σύνολο αυτών των τιμών αποτελούν το 100(1–α)% διάστημα εμπιστοσύνης. Οι τιμές του $p_0$ που οδηγούν σε απόρριψη της μηδενικής υπόθεσης $H_0$ δεν θα ανήκουν στο διάστημα εμπιστοσύνης. Επομένως, αν ($p_1$, $p_2$ ) είναι ένα 100(1–α)% διάστημα εμπιστοσύνης για την παράμετρο p, τότε οι προτάσεις **«** $H_0$ *σε επίπεδο σημαντικότητας* α» και **«** $p_0 \notin (p_1 , p_2 )$ » είναι ισοδύναμες.

Τα άκρα $p_1$, $p_2$ του 100(1–α)% διαστήματος εμπιστοσύνης καθορίζονται έτσι ώστε η παρατηρούμενη τιμή τ της μεταβλητής T να αποτελεί ταυτόχρονα ακραία τιμή των διωνυμικών μεταβλητών $T_1$ ~ διωνυμική (n, $p_1$ ) και $T_2$ ~ διωνυμική (n, $p_2$ ) **,** δηλαδή

$P(T_1 \geq t|n, p = p_1) \cong α/2$ και $P(T_2 \geq t|n, p = p_2) \cong α/2$ .

Επομένως, τα άκρα $p_1$, $p_2$ του διαστήματος εμπιστοσύνης ορίζονται έτσι ώστε

$P(T \leq t\ |n, p = p_2) \cong α/2$

και

$P(T \geq t|n, p = p_1) \cong 1-α/2$

ή, ισοδύναμα,

$P(T \leq t - 1\ |n, p_1) \cong α/2$ .

Επομένως, ο κανόνας απόρριψης του παραπάνω ελέγχου σε επίπεδο σημαντικότητας α μπορεί να διατυπωθεί μέσω του 100(1–α)% διαστήματος εμπιστοσύνης (p1, p2) ως εξής:

Η υπόθεση $H_0$ απορρίπτεται σε επίπεδο σημαντικότητας α αν $p_0 \notin (p_1 , p_2 )$.

**Παρατήρηση:** Στην περίπτωση κατά την οποία το μέγεθος n του δείγματος είναι αρκετά μεγάλο (n >20) και η κατανομή πιθανότητας της στατιστικής συνάρτησης Τ είναι περίπου συμμετρική (np ≥ 5) και (n(1–p)≥ 5), το 100(1–α) % διάστημα εμπιστοσύνης μπορεί να προσδιορισθεί κατά προσέγγιση με βάση την συνάρτηση κατανομής της κανονικής κατανομής με παραμέτρους μ = np και σ2 = np(1–p), η οποία προσεγγίζει τη διωνυμική κατανομή της Τ.

Στην περίπτωση αυτή το διάστημα εμπιστοσύνης με συντελεστή εμπιστοσύνης (1-α) 100% θα είναι:

$$\hat{p}-z_{\alpha}/2\sqrt{P(1-P)/n}<P<\hat{p}+z_{\alpha}/2\sqrt{P(1-P)/n} \quad (3.11)$$

Από την παραπάνω σχέση παρατηρούμε ότι τα όρια εμπιστοσύνης είναι συνάρτηση του άγνωστου ποσοστού P. Επομένως, δεν μπορούμε να ορίσουμε το παραπάνω διάστημα και για το λόγο αυτό αντικαθιστούμε την τιμή P με τη δειγματική εκτίμηση $\hat{p}$. Στην περίπτωση αυτή το διάστημα εμπιστοσύνης για το άγνωστο ποσοστό P δίνεται από τη σχέση:

$$\hat{p}-z_{\alpha}/2\sqrt{\hat{p}\hat{q}/n}<P<\hat{p}+z_{\alpha}/2\sqrt{\hat{p}\hat{q}/n} \quad (3.12)$$

Φυσικά, η παραπάνω διαδικασία εκτίμησης του P είναι περισσότερο ακριβής, όσο μεγαλύτερο είναι το μέγεθος του δείγματος.

*ΠΑΡΑΔΕΙΓΜΑ 3.6:*

Προκειμένου να εκτιμήσουμε το ποσοστό των φοιτητών του πρώτου εξαμήνου που γνωρίζουν ικανοποιητικά μια ξένη γλώσσα, πήραμε ένα δείγμα από 50 φοιτητές. Το δείγμα έδειξε ότι το ποσοστό των φοιτητών που γνωρίζουν ξένη γλώσσα είναι 25%.

Ζητείται με βάση το παραπάνω δείγμα να βρεθεί το 95% διάστημα εμπιστοσύνης για το ποσοστό P.

*ΛΥΣΗ:*

n=50, $\hat{p}=0.25$ , $\hat{q}=1-\hat{p}=0.75, z_{\alpha}/2=1.96$

Το ζητούμενο διάστημα εμπιστοσύνης θα είναι:

$$\hat{p}-z_{\alpha}/2\sqrt{\hat{p}\hat{q}/n}<P<\hat{p}+z_{\alpha}/2\sqrt{\hat{p}\hat{q}/n}\Rightarrow$$

$$0.25-1.96\sqrt{(0.25)(0.75)/50}<P<0.25+1.96\sqrt{(0.25)(0.75)/50}\Rightarrow$$

13%<P<37%.

### *3.5. Διαστήματα εμπιστοσύνης της διαφοράς των μέσων δύο πληθυσμών*

Εστω ένας πληθυσμός Α και ένας άλλος Β και $\mu_1$ και $\mu_2$ οι πραγματικοί μέσοι όροι των δύο πληθυσμών αντίστοιχα. Επιλέγουμε ένα δείγμα $n_1$ από τον πληθυσμό Α και $n_2$ από τον πληθυσμό Β. Έστω $X_1$ ο δειγματικός μέσος του πρώτου και $X_2$ ο δειγματικός μέσος του δεύτερου.

Προκύπτει το πρόβλημα του προσδιορισμού του διαστήματος εμπιστοσύνης της διαφοράς δύο μέσων $\mu_1 - \mu_2$ , με βάση τα 2 δείγματα και σε επίπεδο σημαντικότητας α.

Διακρίνουμε τις παρακάτω περιπτώσεις:

α) **Οι διακυμάνσεις είναι γνωστές:** Αν από δύο κανονικούς πληθυσμούς, που έχουν γνωστές διακυμάνσεις $\sigma_1^2$ και $\sigma_2^2$, ληφθούν δύο τυχαία δείγματα μεγέθους $n_1$ και $n_2$ και οι μέσοι των δειγμάτων αυτών βρεθούν αντίστοιχα $\bar{X}_1$ και $\bar{X}_2$, τότε ο μεν δειγματικός μέσος $\bar{X}_1$ ακολουθεί την κανονική κατανομή

$$\bar{X}_1 \sim N(\mu_{1,}\sigma_1^2/n_1)$$

ο δε δειγματικός μέσος $\bar{X}_2$ ακολουθεί την κατανομή:

$$\bar{X}_2 \sim N(\mu_{2,} \sigma_2^2 / n_2)$$

Άρα η διαφορά των δύο δειγματικών μέσων $\bar{X}_1 - \bar{X}_2$, που είναι μια τυχαία μεταβλητή, θα ακολουθεί την κανονική κατανομή με μέσο τη διαφορά των δύο μέσων μ1-μ2:

$$E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$$

και διακύμανση:

$$Var(\bar{X}_1 - \bar{X}_2) = \sigma^2_{\bar{x}_1 - \bar{x}_2} = Var(\bar{X}_1) - Var(\bar{X}_2) = \sigma_1^2 / n_1 + \sigma_2^2 / n_2$$

Το διάστημα εμπιστοσύνης της διαφοράς δύο μέσων, όταν είναι γνωστές οι διακυμάνσεις σε επίπεδο εμπιστσύνης 1-α, θα είναι:

$$(\bar{X}_1 - \bar{X}_2 - z_{a/2} \sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}, \bar{X}_1 - \bar{X}_2 + z_{a/2} \sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}) \quad (3.13)$$

Για συγκεκριμένο δείγμα:

$$\bar{x}_1 - \bar{x}_2 - z_{a/2} \sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2} < \mu_1 - \mu_2 < \bar{x}_1 - \bar{x}_2 + z_{a/2} \sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2} \quad (3.14)$$

*<u>ΠΑΡΑΔΕΙΓΜΑ 3.7:</u>*

Προκειμένου να βρεθεί η διαφορά του βάρους μεταξύ των σπουδαστών της Ανωτάτης Εμπορικής και Ανωτάτης Βιομηχανικής, έχει ληφθεί από την πρώτη δείγμα $n_1$=1200 σπουδαστών και από τη δεύτερη $n_2$=250 σπουδαστών. Από τα δείγματα αυτά προέκυψε

$\bar{x}_1 = 72\, kgr$ για τους σπουδαστές της Εμπορικής

$\bar{x}_2 = 70\, kgr$ για τους σπουδαστές της Βιομηχανικής.

Αν θεωρηθεί γνωστό ότι η κατανομή των βαρών στις δύο σχολές είναι κανονική και ότι $\sigma_1^2$=100 και $\sigma_2^2$=81, ζητείται το διάστημα εμπιστοσύνης της διαφοράς των δύο μέσων σε επίπεδο σημαντικότητας 5%.

*ΛΥΣΗ:*

Επειδή $\sigma_1^2$ και $\sigma_2^2$ είναι γνωστές, για τον προσδιορισμό του διαστήματος εμπιστοσύνης της διαφοράς των δύο μέσων, εφαρμόζουμε τη σχέση :

$$\bar{x}_1 - \bar{x}_2 - z_{a/2}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2} < \mu_1 - \mu_2 < \bar{x}_1 - \bar{x}_2 + z_{a/2}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$$

Με αντικατάσταση παίρνουμε

$$72 - 70 - 1.96\sqrt{100/1200 + 81/250} \le \mu_1 - \mu_2 \le 72 - 70 + 1.96\sqrt{100/1200 + 81/250}$$

ή

$$1.2 \le \mu_1 - \mu_2 \le 2.8$$

β) **Οι διακυμάνσεις είναι άγνωστες και άνισες και τα δείγματα μεγάλα:** Στην πράξη οι δύο διακυμάνσεις $\sigma_1^2$ και $\sigma_2^2$ δεν είναι γνωστές. Στην περίπτωση αυτή, εφόσον τα δείγματα είναι μεγάλα, μπορούμε να αντικαταστήσουμε τις άγνωστες πληθυσμιακές διακυμάνσεις με τις δειγματικές διακυμάνσεις $S_1^2$ και $S_2^2$.

Το διάστημα εμπιστοσύνης στην περίπτωση αυτή θα είναι:

$$\bar{x}_1-\bar{x}_2-z_{a/2}\sqrt{S_1^2/n_1+S_2^2/n_2}<\mu_1-\mu_2<\bar{x}_1-\bar{x}_2+z_{a/2}\sqrt{S_1^2/n_1+S_2^2/n_2} \quad (3.15)$$

με συντελεστή εμπιστοσύνης (1-α)100%.

γ) **Οι διακυμάνσεις είναι άγνωστες και ίσες και τα δείγματα μικρά:** Αν $S_1^2$ και $S_2^2$ είναι οι δειγματικές διακυμάνσεις που πήραμε από δύο ανεξάρτητα μεταξύ τους τυχαία δείγματα, και επειδή υποτίθεται ότι $\sigma_1^2=\sigma_2^2=\sigma^2$ , τότε η εκτίμηση της $\sigma^2$, δηλαδή η :

$$S^2=[(n_1-1)S_1\, sup2+(n_2-1)S_2^2]/(n_1+n_2-2)$$

αποτελεί εκτιμητή της άγνωστης κοινής διακύμανσης $\sigma^2$ των δύο πληθυσμών και είναι ο σταθμικός μέσος των διακυμάνσεων $S_1^2$ και $S_2^2$.

Το δε διάστημα εμπιστοσύνης είναι :

$$\bar{X}_1-\bar{X}_2-t_{v,\alpha}S\sqrt{1/n_1+1/n_2}<\bar{X}_1-\bar{X}_2+t_{v,\alpha}S\sqrt{1/n_1+1/n_2} \quad (3.16)$$

δ) **Οι διακυμάνσεις είναι άγνωστες και άνισες και τα δείγματα μικρά:** Σε αυτή την περίπτωση ισχύει κατά προσέγγιση η κατανομή t του Student. Αποδεικνύεται ότι το ζητούμενο διάστημα εμπιστοσύνης της διαφοράς των δύο μέσων δίνεται από τη σχέση:

$$\bar{X}_1 - \bar{X}_2 - t_{v,\alpha} S \sqrt{S_1^2/n_1 + S_2^2/n_2} < \bar{X}_1 - \bar{X}_2 + t_{v,\alpha} S \sqrt{S_1^2/n_1 + S_2^2/n_2} \quad (3.17).$$

## *3.6. Διάστημα εμπιστοσύνης του λόγου των διακυμάνσεων δύο κανονικών πληθυσμών*

Εδώ διακρίνουμε δύο περιπτώσεις:

α) **Οι αριθμητικοί μέσοι των δύο κανονικών πληθυσμών είναι γνωστοί:** Στην περίπτωση αυτή το διάστημα εμπιστοσύνης του λόγου $\sigma_1^2/\sigma_2^2$ αποδεικνύεται ότι δίνεται από τη σχέση:

$$\left( \frac{n_2 \sum_{i=1}^{n_1} (x_i - \mu_1)^2}{n_1 \sum_{i=1}^{n_2} (y_i - \mu_2)^2} f^L_{n_2, n_1, \alpha/2} , \frac{n_2 \sum_{i=1}^{n_1} (x_i - \mu_1)^2}{n_1 \sum_{i=1}^{n_2} (y_i - \mu_2)^2} f^U_{n_2, n_1, \alpha/2} \right) \quad (3.18)$$

σε επίπεδο εμπιστοσύνης 1-α. Στη συγκεκριμένη περίπτωση χρησιμοποιούμε τους πίνακες της κατανομής F.

β) **Οι αριθμητικοί μέσοι των δύο κανονικών πληθυσμών είναι άγνωστοι:**

Σε αυτή την περίπτωση το διάστημα εμπιστοσύνης του λόγου $\sigma_1^2/\sigma_2^2$ δίνεται από τη σχέση

$$\left( \frac{s_1^2}{s_2^2} f^L_{n_2-1,n_1-1,\alpha/2} , \frac{s_1^2}{s_2^2} f^U_{n_2-1,n_1-1,\alpha/2} \right) \quad (3.19)$$

σε επίπεδο εμπιστοσύνης 1-α.

## *3.7. Διάστημα εμπιστοσύνης διαφοράς δύο ποσοστών*

Έστω ότι επιθυμούμε να εκτιμήσουμε τη διαφορά ποσοστών $P_1$-$P_2$ δύο πληθυσμών που ακολουθούν τη διωνυμική κατανομή, με τη βοήθεια δύο δειγμάτων $n_1$ και $n_2$.

Αν $\hat{p}_1$ και $\hat{p}_2$ είναι τα ποσοστά των δειγμάτων, θα έχουμε:

$$E(\hat{p}_1 - \hat{p}_2) = E(\hat{p}_1) - E(\hat{p}_2) = P_1 - P_2$$

$$Var(\hat{p}_1 - \hat{p}_2) = Var(\hat{p}_1) + Var(\hat{p}_2) = P_1 Q_1 / n_1 + P_2 Q_2 / n_2$$

όπου

$Q_1$=1-$P_1$ και $Q_2$=1-$P_2$

Με τη χρήση του Κεντρικού Οριακού Θεωρήματος και χρησιμοποιώντας τις εκτιμήσεις $\hat{p}_1 \hat{q}_1 / n_{1,}\ \hat{p}_2 \hat{q}_2 / n_2$ θα έχουμε το διάστημα:

$$\hat{p}_1-\hat{p}_2-z_{\alpha/2}\sqrt{\hat{p}_1\hat{q}_1/n_1+\hat{p}_2\hat{q}_2/n_2}<P_1-P_2<\hat{p}_1-\hat{p}_2+z_{\alpha/2}\sqrt{\hat{p}_1\hat{q}_1/n_1+\hat{p}_2\hat{q}_2/n_2}$$

(3.20)

*ΠΑΡΑΔΕΙΓΜΑ 3.8:*

Ρωτήθηκαν 100 άνδρες και 100 γυναίκες μιας πόλης και βρέθηκε ότι 24 άνδρες και 13 γυναίκες προτιμούν μια ορισμένη εκπομπή της τηλεόρασης. Να βρεθεί το διάστημα εμπιστοσύνης, σε επίπεδο σημαντικότητας 95%, της διαφοράς μεταξύ της αναλογίας αντρών και γυναικών που προτιμούν την παραπάνω εκπομπή.

*ΛΥΣΗ:*

Το διάστημα εμπιστοσύνης δίνεται από τη σχέση (3.20).

Έχουμε:

$\hat{p}_1=24/100=0.24,\ \hat{p}_2=13/100=0.13,\ \hat{p}_1-\hat{p}_2=0.11$

Τελικά από την (3.20) παίρνουμε:

$0.11-1.96x0.54<P_1-P_2<0.11+1.96x0.054$

ή $0.004<P_1-P_2<0.21$

## 3.8. Μέγεθος δείγματος

Κατά τον προσδιορισμό διαστημάτων εμπιστοσύνης για τις άγνωστες τιμές των παραμέτρων ενός πληθυσμού, το μέγεθος του δείγματος n θεωρήθηκε δεδομένο. Ωστόσο, η τιμή του n διαδραματίζει σημαντικό ρόλο στην εκτίμηση του μήκους του διαστήματος εμπιστοσύνης, αφού επηρεάζει την τιμή της τυπικής απόκλισης της δειγματικής τιμής της παραμέτρου (π.χ. του μέσου $\bar{X}_n$ ), η οποία αποτελεί μέτρο του λάθους εκτίμησης. Πράγματι, παρατηρούμε ότι η απόκλιση εκατέρωθεν της αληθινής τιμής του μ προσδιορίζεται από τον όρο:

$$z_\alpha \sigma / \sqrt{n} = e \quad (3.21)$$

όπου ως γνωστόν , $z_α$=κριτική τιμή της Z, η οποία προσδιορίζεται από το συντελεστή εμπιστοσύνης, σ=τυπική απόκλιση στον πληθυσμό (γνωστή), n=μέγεθος δείγματος και e=λάθος. Έτσι, το e προσδιορίζεται από την έκταση που η τιμή του $\bar{X}_n$ εκτιμάται ότι ενδέχεται να απέχει (±e) από την αληθινή τιμή του μέσου του πληθυσμού μ με ορισμένο συντελεστή εμπιστοσύνης. Με δοσμένη την τιμή του σ και του συντελεστή εμπιστοσύνης 1-α, το μέγεθος του λάθους, δηλαδή η τιμή του e, εξαρτάται κατά τρόπο αντίστροφο από το μέγεθος του δείγματος.

Λύνοντας τη (3.21) ως προς n, παίρνουμε:

$$n=(\sigma^2 z_\alpha^2/e)^2 \qquad (3.22)$$

Από την (3.22) προκύπτουν τα εξής:

- Για ορισμένο λάθος και συντελεστή εμπιστοσύνης, το μέγεθος του n εξαρτάται από την τιμή του σ. Με άλλα λόγια, το απαιτούμενο μέγεθος του n για ορισμένο αποτέλεσμα είναι μεγαλύτερο όσο μεγαλύτερη είναι η διακύμανση του πληθυσμού.

- Για ορισμένο λάθος και δεδομένη διακύμανση πληθυσμού, το απαιτούμενο μέγεθος δείγματος είναι τόσο μεγαλύτερο όσο μεγαλύτερος είναι ο συντελεστής εμπιστοσύνης και
- Για δεδομένο συντελεστή εμπιστοσύνης και δεδομένη διακύμανση $\sigma^2$, το απαιτούμενο μέγεθος είναι τόσο μεγαλύτερο, όσο μικρότερο είναι το λάθος ανοχής e.

Στην περίπτωση προσδιοριμού διαστημάτων εμπιστοσύνης ποσοστών ισχύουν οι παρακάτω σχέσεις:

$z_a = \sqrt{pq/n} = e,\ n = z_a^2 pq/e^2$ (3.23)

Όταν το p είναι άγνωστο, αυτό εκτιμάται από τα δεδομένα του δείγματος και στη συνέχεια υπολογίζεται το $\hat{q} = 1 - \hat{p}$ .

## 3.9. Διάστημα εμπιστοσύνης και βιβλιογραφία

Στη συγκεκριμένη παράγραφο παρουσιάζουμε τα διάφορα είδη διαστημάτων εμπιστοσύνης με τους τύπους που συναντάμε συχνά στη βιβλιογραφία:

***Διαστήματα εμπιστοσύνης για τη μέση τιμή κανονικού πληθυσμού με γνωστή διακύμανση***

Στατιστική: $z = \dfrac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}} \sim N(0,1)$

Διάστημα εμπιστοσύνης $1 - a$:

$$\left( \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \ , \ \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

$$\left( \bar{x} - z_{\alpha} \frac{\sigma}{\sqrt{n}} \ , \ +\infty \right)$$

$$\left( -\infty \ , \ \bar{x} + z_{\alpha} \frac{\sigma}{\sqrt{n}} \right)$$

***Διαστήματα εμπιστοσύνης για τη μέση τιμή κανονικού πληθυσμού με άγνωστη διακύμανση***

Για την μέση τιμή $\mu$ κανονικού πληθυσμού με άγνωστη διακύμανση $\sigma^2$ έχουμε:

Για μικρά δείγματα $n \leq 30$

Στατιστική: $t = \dfrac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$

Διάστημα εμπιστοσύνης $1 - a$:

$$\left( \bar{x} - t_{n-1,\alpha/2} \frac{s}{\sqrt{n}} \ , \ \bar{x} + t_{n-1,\alpha/2} \frac{s}{\sqrt{n}} \right)$$

$$\left( -\infty \ , \ \bar{x} + t_{n-1,\alpha} \frac{s}{\sqrt{n}} \right)$$

$$\left( \bar{x} - t_{n-1,\alpha} \frac{s}{\sqrt{n}} \ , \ +\infty \right)$$

Για μεγάλα δείγματα $n > 30$

Στατιστική: $t = \dfrac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \underset{\alpha}{\sim} N(0,1)$

Διάστημα εμπιστοσύνης $1 - a$:

$$\left( \overline{x} - z_{\alpha/2} \frac{s}{\sqrt{n}} \ , \ \overline{x} + z_{\alpha/2} \frac{s}{\sqrt{n}} \right)$$

$$\left( -\infty \ , \ \overline{x} + z_{\alpha} \frac{s}{\sqrt{n}} \right)$$

$$\left( \overline{x} - z_{\alpha} \frac{s}{\sqrt{n}} \ , \ +\infty \right)$$

***Διαστήματα εμπιστοσύνης για τη διαφορά μέσων τιμών δύο ανεξάρτητων κανονικών πληθυσμών με γνωστές διακυμάνσεις***

Για την διαφορά $\mu_1 - \mu_2$ των μέσων τιμών δύο ανεξάρτητων κανονικών πληθυσμών με γνωστές διακυμάνσεις $\sigma_1^2, \sigma_2^2$ έχουμε:

Στατιστική: $z = \dfrac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$

Διάστημα εμπιστοσύνης $1 - a$:

$$\left( (\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} , (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

$$\left( -\infty , (\bar{x}_1 - \bar{x}_2) + z_{\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

$$\left( (\bar{x}_1 - \bar{x}_2) - z_{\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} , +\infty \right)$$

***Διαστήματα εμπιστοσύνης για τη διαφορά μέσων τιμών δύο ανεξάρτητων κανονικών πληθυσμών με άγνωστες διακυμάνσεις***

Για την διαφορά $\mu_1 - \mu_2$ των μέσων τιμών δύο ανεξάρτητων κανονικών πληθυσμών με άγνωστες διακυμάνσεις $\sigma_1^2, \sigma_2^2$ έχουμε:

- Για μικρά δείγματα $n_1, n_2 \leq 30$ και $\sigma_1^2 = \sigma_2^2$

Στατιστική: $$t = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1 + n_2 - 2}$$

Διάστημα εμπιστοσύνης $1 - a$:

$$\left( (\bar{x}_1 - \bar{x}_2) - t_{n_n + n_2 - 2; \alpha/2}\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} , (\bar{x}_1 - \bar{x}_2) + t_{n_n + n_2 - 2; \alpha/2}\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \right)$$

$$\left( -\infty \ , (\bar{x}_1 - \bar{x}_2) + t_{n_n + n_2 - 2, \alpha} \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} \right)$$

$$\left( (\bar{x}_1 - \bar{x}_2) - t_{n_n + n_2 - 2, \alpha} \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} , +\infty \right)$$

Μεγάλα δείγματα

$n_1, n_2 > 30$

Στατιστική: $t = \dfrac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \underset{\alpha}{\sim} N(0,1)$

Διάστημα εμπιστοσύνης $1 - a$:

$$\left( (\bar{x}_1 - \bar{x}_2) - z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} , (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right)$$

$$\left( -\infty \ , (\bar{x}_1 - \bar{x}_2) + z_\alpha \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right)$$

$$\left( (\bar{x}_1 - \bar{x}_2) - z_\alpha \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} , +\infty \right)$$

***Διαστήματα εμπιστοσύνης για την αναλογία διωνυμικού πληθυσμού***

Για την αναλογία $p$ διωνυμικού πληθυσμού $b(1, p)$ έχουμε:

Μεγάλα δείγματα $n > 100$

iv.

Στατιστική: $$z = \frac{\frac{X}{n} - p}{\sqrt{\frac{\frac{X}{n}\left(1 - \frac{X}{n}\right)}{n}}} \underset{\alpha}{\sim} N(0,1)$$

Διάστημα εμπιστοσύνης $1 - a$:

$$\left( \frac{x}{n} - z_{\alpha/2} \sqrt{\frac{\frac{x}{n}\left(1 - \frac{x}{n}\right)}{n}} , \frac{x}{n} + z_{\alpha/2} \sqrt{\frac{\frac{x}{n}\left(1 - \frac{x}{n}\right)}{n}} \right)$$

$$\left( -\infty , \frac{x}{n} + z_{\alpha} \sqrt{\frac{\frac{x}{n}\left(1 - \frac{x}{n}\right)}{n}} \right)$$

$$\left( \frac{x}{n} - z_{\alpha} \sqrt{\frac{\frac{x}{n}\left(1 - \frac{x}{n}\right)}{n}} , +\infty \right)$$

***Διαστήματα εμπιστοσύνης για την διαφορά αναλογιών διωνυμικών πληθυσμών***

Για την διαφορά $p_1 - p_2$ των αναλογιών δύο ανεξάρτητων διωνυμικών πληθυσμών $b(1 , p_1)$ , $b(1 , p_2)$ έχουμε:

Μεγάλα δείγματα $n_1, n_2 > 100$

Στατιστική:
$$z = \frac{\frac{X_1}{n_1} - \frac{X_2}{n_2} - (p_1 - p_2)}{\sqrt{\frac{\frac{X_1}{n_1}\left(1 - \frac{X_1}{n_1}\right)}{n_1} + \frac{\frac{X_2}{n_2}\left(1 - \frac{X_2}{n_2}\right)}{n_2}}} \underset{\alpha}{\sim} N(0,1)$$

Διάστημα εμπιστοσύνης $1 - a$:

$$\left( \left(\frac{x_1}{n_1} - \frac{x_2}{n_2}\right) - z_{\alpha/2}\sqrt{\frac{\frac{x_1}{n_1}\left(1 - \frac{x_1}{n_1}\right)}{n_1} + \frac{\frac{x_2}{n_2}\left(1 - \frac{x_2}{n_2}\right)}{n_2}}, \left(\frac{x_1}{n_1} - \frac{x_2}{n_2}\right) + z_{\alpha/2}\sqrt{\frac{\frac{x_1}{n_1}\left(1 - \frac{x_1}{n_1}\right)}{n_1} + \frac{\frac{x_2}{n_2}\left(1 - \frac{x_2}{n_2}\right)}{n_2}} \right)$$

$$\left( -\infty, \left(\frac{x_1}{n_1} - \frac{x_2}{n_2}\right) + z_{\alpha}\sqrt{\frac{\frac{x_1}{n_1}\left(1 - \frac{x_1}{n_1}\right)}{n_1} + \frac{\frac{x_2}{n_2}\left(1 - \frac{x_2}{n_2}\right)}{n_2}} \right)$$

$$\left( \left(\frac{x_1}{n_1} - \frac{x_2}{n_2}\right) - z_{\alpha}\sqrt{\frac{\frac{x_1}{n_1}\left(1 - \frac{x_1}{n_1}\right)}{n_1} + \frac{\frac{x_2}{n_2}\left(1 - \frac{x_2}{n_2}\right)}{n_2}}, +\infty \right)$$

***Διαστήματα εμπιστοσύνης για τη διακύμανση κανονικού πληθυσμού με γνωστό μέσο***

Για την διακύμανση $\sigma^2$ κανονικού πληθυσμού με γνωστό μέσο $\mu$ είναι:

Στατιστική: $\chi^2 = \dfrac{\sum_{i=1}^{n} (X_i - \mu)^2}{\sigma^2} \sim \chi^2_n$

Διάστημα εμπιστοσύνης $1-a$:

$$\left( \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{\chi^2_{n,\alpha/2}}, \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{\chi^2_{n,1-\alpha/2}} \right)$$

$$\left( 0, \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{\chi^2_{n,1-\alpha}} \right)$$

$$\left( \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{\chi^2_{n,\alpha}}, +\infty \right)$$

***Διαστήματα εμπιστοσύνης για τη διακύμανση κανονικού πληθυσμού με άγνωστο μέσο***

Για την διακύμανση $\sigma^2$ κανονικού πληθυσμού με άγνωστο μέσο $\mu$ είναι:

Στατιστική: $\chi^2 = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$

Διάστημα εμπιστοσύνης $1-a$:

$$\left( \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}} \right)$$

$$\left( 0\,, \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha}} \right)$$

$$\left( \frac{(n-1)s^2}{\chi^2_{n-1,\alpha}}, +\infty \right)$$

➧***Διαστήματα εμπιστοσύνης για το λόγο των διακυμάνσεων δύο ανεξάρτητων κανονικών πληθυσμών με γνωστούς μέσους***

Για το λόγο $\sigma_1^2 / \sigma_2^2$ των διακυμάνσεων δύο ανεξάρτητων κανονικών πληθυσμών με γνωστούς μέσους $\mu_1$ και $\mu_2$ είναι:

Στατιστική: $$F = \frac{\left. \dfrac{\sum_{i=1}^{n_1} (X_i - \mu_1)^2}{\sigma_1^2} \right/ n_1}{\left. \dfrac{\sum_{i=1}^{n_2} (Y_i - \mu_2)^2}{\sigma_2^2} \right/ n_2} \sim F_{n_1, n_2}$$

Διάστημα εμπιστοσύνης $1-a$:

$$\left( \frac{n_2 \sum_{i=1}^{n_1} (x_i - \mu_1)^2}{n_1 \sum_{i=1}^{n_2} (y_i - \mu_2)^2} f^L_{n_2, n_1, \alpha/2}, \frac{n_2 \sum_{i=1}^{n_1} (x_i - \mu_1)^2}{n_1 \sum_{i=1}^{n_2} (y_i - \mu_2)^2} f^U_{n_2, n_1, \alpha/2} \right)$$

$$\left( 0\,,\frac{n_2\sum_{i=1}^{n_1}(x_i-\mu_1)^2}{n_1\sum_{i=1}^{n_2}(y_i-\mu_2)^2}f^U_{n_2,n_1,\alpha} \right)$$

$$\left( \frac{n_2\sum_{i=1}^{n_1}(x_i-\mu_1)^2}{n_1\sum_{i=1}^{n_2}(y_i-\mu_2)^2}f^L_{n_2,n_1,\alpha}\;,+\infty \right)$$

**◆*Διαστήματα εμπιστοσύνης για το λόγο των διακυμάνσεων δύο ανεξάρτητων κανονικών πληθυσμών με άγνωστους μέσους***

Για το λόγο $\sigma_1^2/\sigma_2^2$ των διακυμάνσεων δύο ανεξάρτητων κανονικών πληθυσμών με άγνωστους μέσους $\mu_1$ και $\mu_2$ είναι:

Στατιστική: $F=\dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}\sim F_{n_1-1,n_2-1}$

Διάστημα εμπιστοσύνης $1-a$:

$$\left( \frac{s_1^2}{s_2^2}f^L_{n_2-1,n_1-1,\alpha/2}\,,\frac{s_1^2}{s_2^2}f^U_{n_2-1,n_1-1,\alpha/2} \right)$$

$$\left( 0\,,\frac{s_1^2}{s_2^2}f^U_{n_2-1,n_1-1,\alpha} \right)$$

$$\left( \frac{s_1^2}{s_2^2}f^L_{n_2-1,n_1-1,\alpha}\;,+\infty \right)$$

όπου ισχύει: $f^{L}_{n,k,\alpha} = {1}/{f^{U}_{k,n,\alpha}}$

# ΚΕΦΑΛΑΙΟ 4

# Εφαρμογές Διαστημάτων Εμπιστοσύνης

## *4.1. SPSS*

Στην ενότητα αυτή σκοπό έχουμε να παρουσιάσουμε τις βασικές λειτουργίες του στατιστικού πακέτου SPSS . Αυτές θα αναφερθούν στα ακόλουθα σημεία:

1.Εισαγωγή δεδομένων, δημιουργία μεταβλητών, αλλαγές στις τιμές των μεταβλητών, βασικές εργασίες με τα δεδομένα που βοηθούν στον έλεγχο του αρχείου δεδομένων.

2.Μετασχηματισμοί των δεδομένων.

3.Ανάλυση των δεδομένων η οποία θα περιλαμβάνει απλά περιγραφικά μέτρα, περιγραφικά διαγράμματα, παραμετρική σύγκριση μέσων ανεξάρτητων ομάδων, παραμετρική σύγκριση δύο εξαρτημένων ομάδων, παραμετρική σύγκριση περισσότερων από δύο ανεξάρτητες ομάδες, αντίστοιχες προς τα παραπάνω απαραμετρικές συγκρίσεις, γραμμική παλινδρόμηση, ανάλυση κατηγορικών δεδομένων.

***Πεδία δεδομένων-μεταβλητών***

Το SPSS διαθέτει δύο πεδία. Στο πεδίο δεδομένων φαίνονται τα δεδομένα. Κάθε σειρά αναφέρεται σε ένα υποκείμενο, κάθε στήλη σε μια μεταβλητή. Οι μεταβλητές καθορίζονται-φαίνονται στο πεδίο μεταβλητών. Εκεί δίνονται οι ονομασίες και καθορίζονται οι ιδιότητές τους (όνομα, τύπος μεταβλητής, αριθμός ψηφίων, κωδικοποίηση τιμών κ.α)

***Διαδικασίες***

**Α. File**

Τα αρχεία που μπορεί να διαχειρισθεί το SPSS είναι τριών ειδών: αρχεία δεδομένων (προέκταση .sav), αρχεία αποτελεσμάτων (προέκταση .out) και αρχεία syntax στα οποία η ανάλυση δεδομένων γίνεται με τη χρήση εντολών και όχι παραθυρικά (προέκταση .sps). **New:** ανοίγει ένα καινούργιο αρχείο. **Open:** Ανοίγει ένα ήδη δημιουργημένο αρχείο. **Save:** Σώζει στο σκληρό δίσκο ένα αρχείο το οποίο έχει μόλις δημιουργηθεί.

**Β. Edit**

Εντολές με τις οποίες αλλάζουν οι τιμές των μεταβλητών. **Cut:** Σβήνεται η τιμή ή οι τιμές μιας μεταβλητής. **Copy:** Αντιγράφεται η τιμή ή οι τιμές μιας μεταβλητής. **Paste:** Επικολλάται η τιμή ή οι τιμές μιας μεταβλητής σε μιαν άλλη. **Paste Variables:** Επικολλώνται ολόκληρες μεταβλητές σε άλλο σημείο του πεδίου δεδομένων. **Clear:** Σβήνεται μια ή περισσότερες μεταβλητές από το πεδίο δεδομένων. **Find:** Ο cursor βρίσκει και οδηγείται στην τιμή μιας συγκεκριμένης μεταβλητής.

**Γ. Data**

Εντολές με τις οποίες πραγματοποιούνται διάφορες εργασίες στα δεδομένα. **Insert Variable:** Δημιουργείται μια νέα μεταβλητή στα αριστερά του cursor. **Insert Cases:** Προστίθεται μια νέα γραμμή πάνω από το σημείο που βρίσκεται ο cursor. **Go to case:** O cursor οδηγείται σε μια συγκεκριμένη γραμμή. **Sort cases:** Τα δεδομένα ταξινομούνται ανάλογα με τις τιμές μιας ή περισσότερων μεταβλητών. Η ταξινόμηση γίνεται στις γραμμές (στα υποκείμενα). **Transpose:** Οι γραμμές γίνονται στήλες και οι στήλες γραμμές. **Merge Files:** Η διαδικασία ενώνει δύο αρχεία είτε προσθέτοντας μεταβλητές (**Add variables),** είτε προσθέτοντας γραμμές-υποκείμενα (**Add cases). Split file:** Τα υποκείμενα χωρίζονται σε ομάδες ανάλογα με τις τιμές μιας ή περισσότερων μεταβλητών έτσι ώστε οι αναλύσεις να επαναλαμβάνονται για τα υποκείμενα κάθε ομάδας. **Select cases:** Επιλέγονται για την ανάλυση εκείνα τα υποκείμενα που ικανοποιούν μια συνθήκη. **Weight cases:** Οι τιμές των μεταβλητών φαίνονται τόσες φορές όσες δηλώνει η μεταβλητή-βάρος.

### Δ. Transform

Εντολές με τις οποίες τα δεδομένα μετασχηματίζονται και παράγονται νέες μεταβλητές από τις παλαιότερες. **Compute:** Δημιουργείται μια νέα μεταβλητή με τις συνηθισμένες πράξεις της αριθμητικής, είτε χρησιμοποιώντας άλλους πολυπλοκότερους μετασχηματισμούς. Οι μετασχηματισμοί αυτοί γίνονται και κάτω από συνθήκες. **Recode:** Γίνεται επανακωδικοποίηση μεταβλητών και το αποτέλεσμα καταχωρείται είτε στην ίδια μεταβλητή αλλάζοντας τις τιμές της (**Into same variables**), είτε σε μιαν άλλη μεταβλητή (**Into different variables). Count:** Μετρώνται οι εμφανίσεις τιμών σε μια ομάδα μεταβλητών.

### Ε. Analyze

Εντολές με τις οποίες πραγματοποιούνται στατιστικές αναλύσεις στα δεδομένα.

**Descriptive Statistics:** Διαδικασίες για την παραγωγή περιγραφικών αποτελεσμάτων.

**Frequencies:** Οι μεταβλητές για τις οποίες αποτελέσματα είναι επιθυμητά εισάγονται στο πλαίσιο *Variable(s)*. Κουμπί ***Statistics***: Χρήσιμες στατιστικές περιγραφικές συναρτήσεις (μέσος, διάμεσος, εκατοστιαία σημεία, διακύμανση, τυπική απόκλιση, μέγιστο, ελάχιστο). Κουμπί ***Charts***: Χρήσιμα διαγράμματα (ιστόγραμμα).

**Explore:** Διαδικασία που υπολογίζει περιγραφικά αποτελέσματα για κάθε ομάδα υποκειμένων. Οι μεταβλητές για τις οποίες αποτελέσματα είναι επιθυμητά εισάγονται στο πλαίσιο *Dependent List*. Οι μεταβλητές που ορίζουν τις ομάδες εισάγονται στο πλαίσιο *Factor List*. Κουμπί ***Statistics:*** Χρήσιμες στατιστικές συναρτήσεις (Descriptives). Κουμπί ***Plots:*** Χρήσιμα διαγράμματα ανά ομάδα (box plot, ιστόγραμμα).

**Crosstabs:** Διαδικασία που πινακοποιεί τη σχέση μεταξύ δύο κατηγορικών μεταβλητών. Κάθε μεταβλητή πρέπει να παίρνει τουλάχιστον δύο επίπεδα τιμών. Οι μεταβλητές εισάγονται η μία στο πλαίσιο *Row(s)* και η άλλη στο πλαίσιο *Column(s)*. Αν η σχέση αυτή είναι επιθυμητό να διερευνηθεί στην κάθε τιμή μιας τρίτης κατηγορικής μεταβλητής τότε αυτή εισάγεται στο πλαίσιο *Layer 1 of 1.* Κουμπί **Statistics:** Μπορούν να επιλεγούν μέτρα συσχέτισης μεταξύ δύο κατηγορικών μεταβλητών (έλεγχος ανεξαρτησίας $\chi^2$, έλεγχος Cochran and Mantel-Haenszel). Κουμπί ***Cells***: Επιλέγεται αν θα παρουσιαστούν ποσοστά των τιμών της μεταβλητής που βρίσκεται στις γραμμές για κάθε τιμή της μεταβλητής που βρίσκεται στις στήλες (πλαίσιο *Percentages, Column*) ή το αντίστροφο (πλαίσιο *Percentages, Row*).

**Compare means:** Διαδικασίες για την εκτίμηση και τον έλεγχο της ισότητας των μέσων τιμών δύο ή περισσότερων συνεχών μεταβλητών.

**Means:** Παρουσιάζονται περιγραφικά αποτελέσματα για τις μεταβλητές για τις οποίες υπάρχει ενδιαφέρον σε κάθε ομάδα υποκειμένων. Οι μεταβλητές για τις οποίες αποτελέσματα είναι επιθυμητά εισάγονται στο πλαίσιο *Dependent List*. Οι μεταβλητές που ορίζουν τις ομάδες εισάγονται στο πλαίσιο *Independent List*. Κουμπί ***Options:*** Χρήσιμες στατιστικές περιγραφικές συναρτήσεις μπορούν να επιλεγούν (μέσος, διάμεσος, εκατοστιαία σημεία, διακύμανση, τυπική απόκλιση, μέγιστο, ελάχιστο).

**Independent samples t-test:** Η διαδικασία ελέγχει την ισότητα των μέσων τιμών μιας μεταβλητής όταν αυτή μετράται σε δύο ανεξάρτητες ομάδες υποκειμένων. Η μεταβλητή που ενδιαφέρει εισάγεται στο πλαίσιο *Test Variable(s)*. Η μεταβλητή που ορίζει τις δύο ομάδες υποκειμένων εισάγεται στο πλαίσιο *Grouping Variable*. Κουμπί ***Define groups***: Δίνονται οι τιμές που ορίζουν τις δύο ομάδες.

**Paired Samples t-test:** Η διαδικασία ελέγχει την ισότητα των μέσων τιμών μιας μεταβλητής όταν αυτή μετράται σε δύο εξαρτημένες ομάδες υποκειμένων (τα δεδομένα των δύο ομάδων έχουν εισαχθεί σε δύο διαφορετικές μεταβλητές). Πλαίσιο ***Current Selections***: Επιλέγεται το ζευγάρι των μεταβλητών των οποίων οι μέσες τιμές είναι επιθυμητό να ελεγχθούν ως προ την ισότητά τους.

**One-way ANOVA:** Η διαδικασία ελέγχει την ισότητα των μέσων τιμών μιας μεταβλητής μετρημένης σε τουλάχιστον τρεις ανεξάρτητες ομάδες υποκειμένων. Η μεταβλητή που ενδιαφέρει εισάγεται στο πλαίσιο *Dependent List*. Η μεταβλητή που ορίζει τις δύο ομάδες υποκειμένων εισάγεται στο πλαίσιο *Factor.* Κουμπί ***Post Hoc***: Επιλέγονται τεχνικές πολλαπλών συγκρίσεων για να συγκριθούν οι μέσοι ανά δύο (Bonferroni, Scheffe). Κουμπί ***Options:*** Μπορεί να επιλεγεί παρουσίαση περιγραφικών αποτελεσμάτων (Descriptive), έλεγχος ίσων διακυμάνσεων (Homogeneity of variance test), στατιστικές συναρτήσεις που ελέγχουν την ισότητα των μέσων τιμών σε περίπτωση όπου η ομοσκεδαστικότητα δεν ισχύει (Brown-Forsythe, Welch).

**General Linear Model:** Διαδικασίες που προσαρμόζουν γενικευμένα μοντέλα ANOVA.

**Univariate:** Η διαδικασία ελέγχει την ύπαρξη αλληλεπιδράσεων και κυρίων επιδράσεων παραγόντων στην μεταβλητή που ενδιαφέρει. Αυτό είναι ισοδύναμο με την προσαρμογή ενός γραμμικού μοντέλου στο οποίο οι επεξηγηματικές μεταβλητές είναι δύο ή περισσότερες διακριτές μεταβλητές (παράγοντες). Η μεταβλητή που μελετάται εισάγεται στο πλαίσιο *Dependent Variable*. Οι παράγοντες των οποίων οι κύριες επιδράσεις και αλληλεπιδράσεις είναι επιθυμητό να ελεγχθούν, εισάγονται στο πλαίσιο *Fixed Factor(s)*. Αν είναι επιθυμητό να γίνει ανάλυση χρησιμοποιώντας ANCOVA (Analysis of covariance) τότε η συνεχής επεξηγηματική μεταβλητή ως προς την οποία γίνεται control εισάγεται στο πλαίσιο *Covariate(s)*. Χρήσιμα κουμπιά μπορούν να είναι: Κουμπί ***Plots***: Οι παράγοντες εμφανίζονται στο πλαίσιο *Facors*. Είναι προτιμότερο κάθε φορά να παρουσιάζονται στα διαγράμματα δύο παράγοντες (ακόμα και εάν υπάρχουν περισσότεροι). Ο πρώτος εισάγεται στο πλαίσιο *Horizontal Axis*. Ο δεύτερος εισάγεται στο πλαίσιο *Separate Lines.* Το διάγραμμα θα κατασκευαστεί εφόσον πατηθεί το κουμπί **Add.** Το διάγραμμα που θα παρουσιαστεί είναι ένα διάγραμμα αλληλεπίδρασης των δύο παραγόντων που έχουν χρησιμοποιηθεί στην κατασκευή του διαγράμματος. Κουμπί ***Options***: Χρήσιμο μπορεί να είναι να ζητηθούν περιγραφικά αποτελέσματα (Descriptive statistics).

**Correlate:** Η διαδικασία μετράει τη συσχέτιση ανάμεσα σε ζευγάρια μεταβλητών.

**Bivariate correlations:** Οι μεταβλητές που ενδιαφέρουν εισάγονται στο πλαίσιο *Variables*. Χρήσιμες επιλογές είναι οι **Pearson** ή οι **Spearman** συσχετίσεις.

**Regression:** Η διαδικασία εκτελεί διάφορα είδη αναλύσεων παλινδρόμησης.

**Linear:** Η μεταβλητή που ενδιαφέρει (εξαρτημένη μεταβλητή) εισάγεται στο πλαίσιο *Dependent.* Οι επεξηγηματικές μεταβλητές με τις οποίες θα εξηγηθεί η μεταβλητότητα της εξαρτημένης μεταβλητής, εισάγονται στο πλαίσιο *Independent(s)*. Στο πλαίσιο *Method* μπορεί να επιλεγεί μια μέθοδος για τη βέλτιστη επιλογή επεξηγηματικών μεταβλητών. Αυτή συνήθως αφήνεται *Enter* που σημαίνει ότι στο μοντέλο εισέρχονται όσες μεταβλητές βρίσκονται στο πλαίσιο *Independent(s)* με τη σειρά που γράφονται εκεί. Κουμπί ***Statistics***: Χρήσιμες επιλογές είναι εκτιμήσεις των παραμέτρων του μοντέλου (**Estimates),** διαστήματα εμπιστοσύνης των παραμέτρων **(Confidence intervals)**, πίνακας συνδιακύμανσης των εκτιμώμενων παραμέτρων (**covariance matrix)** καθώς και ένδειξη καλής προσαρμογής (**Model fit)**, αλλαγές στους δείκτες $R^2$ (στην περίπτωση εισαγωγής πολλών επεξηγηματικών μεταβλητών) **(R squared change)**, περιγραφικά στοιχεία **(Descriptives)**, μερικές συσχετίσεις **(part and partial correlations),** διαγνωστικά εργαλεία πολυσυγγραμμικότητας (**collinearity diagnostics).** Κουμπί ***Plots***: Μπορεί να κατασκευαστεί διάγραμμα καταλοίπων για τον οπτικό έλεγχο καλής προσαρμογής του μοντέλου. Στο πλαίσιο *Y* εισάγεται η λέξη *SRESID* και στο πλαίσιο *X* η λέξη *ZPRED*. Το διάγραμμα που προκύπτει είναι ένα διάγραμμα καταλοίπων σε σχέση με τις προβλεπόμενες τιμές που προκύπτουν από το μοντέλο. Αν η προσαρμογή του μοντέλου είναι καλή τότε το διάγραμμα δεν πρέπει να υποδεικνύει καμιά συστηματικότητα και τα σημεία πρέπει να φαίνονται τυχαία κατανεμημένα. Κουμπί ***Save***. Μπορείτε να δημιουργήσετε νέες μεταβλητές οι οποίες να περιέχουν διάφορους τύπους προβλεπόμενων τιμών, καταλοίπων, αποστάσεων ή μέτρων που καθορίζουν την επίδραση ακραίων σημείων στο μοντέλο.

**Nonparametric Tests:** Η διαδικασία εκτελεί απαραμετρικούς ελέγχους ισότητας μέσων τιμών μεταβλητών χωρίς την υπόθεση της κανονικότητας των δεδομένων.

**Two independent samples:** Είναι το απαραμετρικό αντίστοιχο του Independent samples t-test. Η μεταβλητή που ενδιαφέρει εισάγεται στο πλαίσιο *Test Variable List*. Η μεταβλητή που προσδιορίζει τις ομάδες ων υποκειμένων εισάγεται στο πλαίσιο *Grouping Variable*. Αυτό γίνεται με τη βοήθεια του κουμπιού ***Define Groups***. Ο πλέον συνηθισμένος έλεγχος είναι ο ***Mann-Whitney U test.*** Κουμπί ***Options:*** Στο πλαίσιο Statistics μπορούν να επιλεγούν περιγραφικά αποτελέσματα (Descriptive) ή και εκτίμηση των 25%, 50%, 75% εκατοστιαίων σημείων (Quartiles).

**K Independent Samples:** Είναι το απαραμετρικό ισοδύναμο του One-way ANOVA. Η μεταβλητή που ενδιαφέρει εισάγεται στο πλαίσιο *Test Variable List*. Η μεταβλητή που προσδιορίζει τις ομάδες ων υποκειμένων εισάγεται στο πλαίσιο *Grouping Variable*. Αυτό γίνεται με τη βοήθεια του κουμπιού ***Define Range***. Ο πλέον συνηθισμένος έλεγχος είναι ο ***Kruskal-Wallis H test.*** Κουμπί ***Options:*** Στο πλαίσιο Statistics μπορούν να επιλεγούν περιγραφικά αποτελέσματα (Descriptive) ή και εκτίμηση των 25%, 50%, 75% εκατοστιαίων σημείων (Quartiles).

**Two related samples:** Είναι το απαραμετρικό ισοδύναμο του Paired samples t-test. Τα ζεύγη των μεταβλητών των οποίων οι μέσες τιμές είναι επιθυμητό να συγκριθούν επιλέγονται και εισάγονται στο πλαίσιο *Test Pair(s) List.* Ο πλέον συνηθισμένος έλεγχος, εάν οι δύο μεταβλητές είναι σε συνεχή κλίμακα, είναι ο έλεγχος **Wilcoxon**. Εναλλακτικά μπορεί να χρησιμοποιηθεί ο έλεγχος **Sign**. Εάν οι δύο μεταβλητές είναι δίτιμες (επιτυχία-αποτυχία) τότε ο κατάλληλος έλεγχος είναι ο έλεγχος **Mc Nemar**. Εάν οι δύο μεταβλητές είναι σε τακτική κλίμακα τότε μπορούμε να ελέγξουμε την ισότητα των κατανομών των δύο εξαρτημένων δειγμάτων με τον έλεγχο **Marginal Homogeneity**. Κουμπί ***Options:*** Στο πλαίσιο Statistics μπορούν να επιλεγούν περιγραφικά αποτελέσματα (Descriptive) ή και εκτίμηση των 25%, 50%, 75% εκατοστιαίων σημείων (Quartiles).

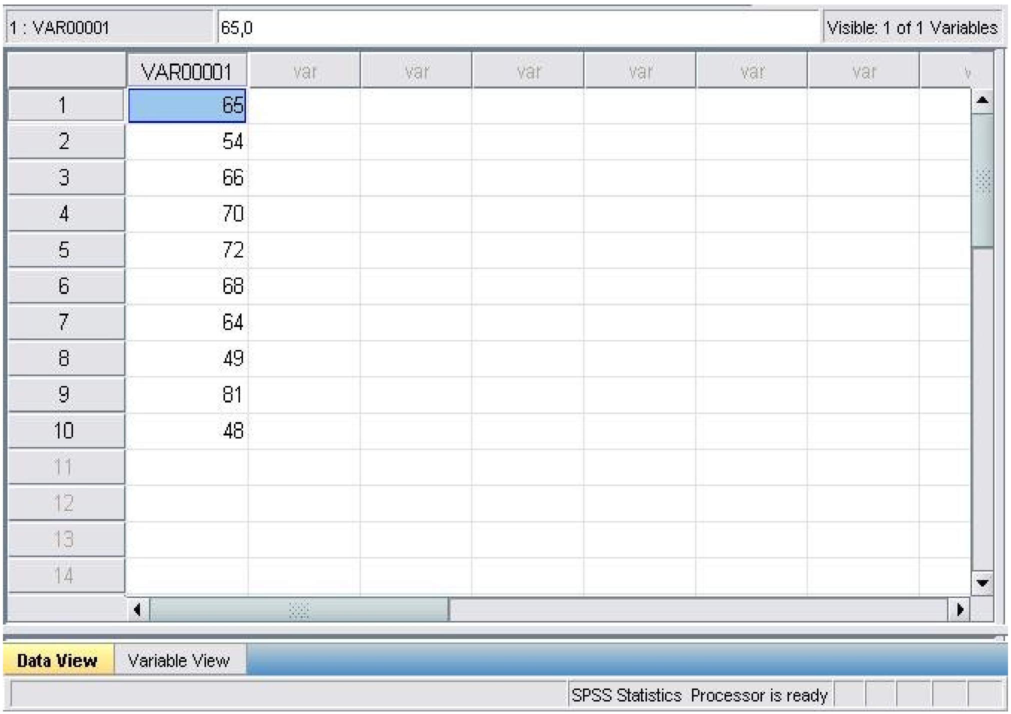
**K related samples:** Είναι η απαραμετρική επέκταση των ελέγχων **Wilcoxon και Mc Nemar** από δύο σε τουλάχιστον τρεις εξαρτημένους πληθυσμούς. Αν οι δύο μεταβλητές είναι σε συνεχή κλίμακα τότε ο κατάλληλος έλεγχος είναι ο **Friedman**. Αν οι μεταβλητές είναι δίτιμες (επιτυχία-αποτυχία) τότε ο κατάλληλος έλεγχος είναι ο **Cochran's Q. Κουμπί *Statistics*:** Μπορούν να επιλεγούν περιγραφικά αποτελέσματα (**Descriptive**) ή και εκτίμηση των 25%, 50%, 75% εκατοστιαίων σημείων (**Quartiles**).

## *4.2. Παραδείγματα σε SPSS*

Για την όσο το δυνατόν καλύτερη ανάλυση καναμε χρήση τριών παραδειγμάτων ώστε να μπορέσουμε να κάνουμε εφαρμογή των διαστημάτων εμπιστοσύνης.

*ΠΑΡΑΔΕΙΓΜΑ 4.1*

Το πρώτο μας παράδειγμα αποτελείται από 10 παρατηρήσεις που αντιπροσωπεύουν τον αριθμό των ατόμων που προσήλθαν για να παρακολουθήσουν ένα σεμινάριο για δέκα διαφορετικές ημέρες.




Στην μπάρα εργασίας επιλέγω Analyze → Descriptive Statistics → Explore. Ορίζουμε σαν εξαρτημένη μεταβλητή τη μοναδική μας μεταβλητή και πατώντας το κουμπί Statistics επιλέγουμε να γίνει το test με διάστημα εμπιστοσύνης 95%.Έτσι παίρνουμε τους παρακάτω πίνακες:

**Case Processing Summary**

| | Cases | | | | | |
|---|---|---|---|---|---|---|
| | Valid | | Missing | | Total | |
| | N | Percent | N | Percent | N | Percent |
| VAR00001 | 10 | 100,0% | 0 | ,0% | 10 | 100,0% |

**Descriptives**

| | | | Statistic | Std. Error |
|---|---|---|---|---|
| VAR00001 | Mean | | 63,70 | 3,317 |
| | 95% Confidence Interval for Mean | Lower Bound | 56,20 | |
| | | Upper Bound | 71,20 | |
| | 5% Trimmed Mean | | 63,61 | |
| | Median | | 65,50 | |
| | Variance | | 110,011 | |
| | Std. Deviation | | 10,489 | |
| | Minimum | | 48 | |
| | Maximum | | 81 | |
| | Range | | 33 | |
| | Interquartile Range | | 18 | |
| | Skewness | | -,225 | ,687 |
| | Kurtosis | | -,506 | 1,334 |

Ο μέσος μας ισούται με 63.70, ενώ το διάστημα εμπιστοσύνης 95% για τον μέσο είναι (56.20,71.20). Επίσης παίρνουμε και άλλες πληροφορίες όπως τη διακύμανση, τη διάμεσο, το εύρος κ.α.

*<u>ΠΑΡΑΔΕΙΓΜΑ 4.2</u>*

Το δεύτερο παράδειγμά μας παρουσιάζει το σύνολο των γάμων, καθώς και τον αριθμό των πολιτικών και θρησκευτικών γάμων που πραγματοποιήθηκαν την περίοδο 1991-2004.

| 7 : YEAR | 1997,0 | | | | | |
|---|---|---|---|---|---|---|
| | YEAR | RELIGIOUS | CIVIL | TOTAL | var | var |
| 1 | 1991 | 59710 | 5858 | 65568 | | |
| 2 | 1992 | 42406 | 6225 | 48631 | | |
| 3 | 1993 | 56204 | 5991 | 62195 | | |
| 4 | 1994 | 50889 | 5924 | 56813 | | |
| 5 | 1995 | 57258 | 6729 | 63987 | | |
| 6 | 1996 | 38670 | 6738 | 45408 | | |
| 7 | 1997 | 53652 | 6883 | 60535 | | |
| 8 | 1998 | 49081 | 6408 | 55489 | | |
| 9 | 1999 | 53495 | 7670 | 61165 | | |
| 10 | 2000 | 40269 | 8611 | 48880 | | |
| 11 | 2001 | 48087 | 10407 | 58491 | | |
| 12 | 2002 | 45578 | 12294 | 57872 | | |
| 13 | 2003 | 47871 | 13210 | 61081 | | |
| 14 | 2004 | 37496 | 13881 | 51377 | | |
| 15 | | | | | | |
| 16 | | | | | | |
| 17 | | | | | | |
| 18 | | | | | | |
| 19 | | | | | | |
| 20 | | | | | | |
| 21 | | | | | | |

Ακολουθώντας παρόμοια διαδικασία με το προηγούμενο παράδειγμα και αφού ορίσουμε σαν εξαρτημένες μεταβλητές το σύνολο των γάμων, τους θρησκευτικούς και τους πολιτικούς, προκύπτουν οι δύο παρακάτω πίνακες.

**Case Processing Summary**

| | Cases | | | | | |
|---|---|---|---|---|---|---|
| | Valid | | Missing | | Total | |
| | N | Percent | N | Percent | N | Percent |
| ΑΡΙΘΜΟΣ ΘΡΗΣΚΕΥΤΙΚΩΝ ΓΑΜΩΝ | 14 | 100,0% | 0 | ,0% | 14 | 100,0% |
| ΑΡΙΘΜΟΣ ΠΟΛΙΤΙΚΩΝ ΓΑΜΩΝ | 14 | 100,0% | 0 | ,0% | 14 | 100,0% |
| ΑΡΙΘΜΟΣ ΣΥΝΟΛΙΚΩΝ ΓΑΜΩΝ | 14 | 100,0% | 0 | ,0% | 14 | 100,0% |

**Descriptives**

| | | | Statistic | Std. Error |
|---|---|---|---|---|
| ΑΡΙΘΜΟΣ ΘΡΗΣΚΕΥΤΙΚΩΝ ΓΑΜΩΝ | Mean | | 48619,00 | 1890,935 |
| | 95% Confidence Interval for Mean | Lower Bound | 44533,88 | |
| | | Upper Bound | 52704,12 | |
| | 5% Trimmed Mean | | 48620,78 | |
| | Median | | 48584,00 | |
| | Variance | | 5,006E7 | |
| | Std. Deviation | | 7075,230 | |
| | Minimum | | 37496 | |
| | Maximum | | 59710 | |
| | Range | | 22214 | |
| | Interquartile Range | | 12418 | |
| | Skewness | | -,129 | ,597 |
| | Kurtosis | | -1,068 | 1,154 |
| ΑΡΙΘΜΟΣ ΠΟΛΙΤΙΚΩΝ ΓΑΜΩΝ | Mean | | 8344,93 | 769,807 |
| | 95% Confidence Interval for Mean | Lower Bound | 6681,86 | |
| | | Upper Bound | 10008,00 | |
| | 5% Trimmed Mean | | 8175,53 | |
| | Median | | 6810,50 | |
| | Variance | | 8296445,456 | |
| | Std. Deviation | | 2880,355 | |
| | Minimum | | 5858 | |
| | Maximum | | 13881 | |
| | Range | | 8023 | |
| | Interquartile Range | | 4712 | |
| | Skewness | | 1,067 | ,597 |
| | Kurtosis | | -,433 | 1,154 |
| ΑΡΙΘΜΟΣ ΣΥΝΟΛΙΚΩΝ ΓΑΜΩΝ | Mean | | 56963,71 | 1662,391 |
| | 95% Confidence Interval for Mean | Lower Bound | 53372,34 | |
| | | Upper Bound | 60555,09 | |
| | 5% Trimmed Mean | | 57127,68 | |
| | Median | | 58181,50 | |
| | Variance | | 3,869E7 | |
| | Std. Deviation | | 6220,098 | |
| | Minimum | | 45408 | |
| | Maximum | | 65568 | |
| | Range | | 20160 | |
| | Interquartile Range | | 10670 | |
| | Skewness | | -,562 | ,597 |
| | Kurtosis | | -,789 | 1,154 |

Για τους θρησκευτικούς γάμους έχουμε ότι ο μέσος ισούται με 48619 και το διάστημα εμπιστοσύνης 95% για το μέσο είναι (44533.88,52704.12).

Ομοίως για τους πολιτικούς έχουμε μέσο 8344 και διάστημα εμπιστοσύνης (6681.16,10008) και για τους γάμους συνολικά ο μέσος είναι 56963.71 και διάστημα εμπιστοσύνης (53372.34,60555.09).

Επιλέγουμε Analyze ➡ Compare Means ➡ One sample t-test, εξετάζουμε τις μεταβλητές των θρησκευτικών και πολιτικών γάμων μεμονωμένα, ορίζουμε απο το κουμπί Options να έχουμε επίπεδο εμπιστοσύνης 95% ενώ ορίζουμε κάθε φορά μια Test value σαν (υποθετική) τιμή του μέσου ενός πληθυσμού. Με αυτό τον τρόπο βρίσκουμε το επίπεδο εμπιστοσύνης της διαφοράς των μέσων δύο πληθυσμών.

Πιο αναλυτικά, στην περίπτωση των θρησκυτικών γάμων ορίζουμε σαν Test value το 50000, δηλαδή ότι έχω έναν πληθυσμό με μέσο 50000.Ακολουθώντας την παραπάνω διαδικασία έχουμε τους εξής πίνακες:

**One-Sample Statistics**

| | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|
| ΑΡΙΘΜΟΣ ΘΡΗΣΚΕΥΤΙΚΩΝ ΓΑΜΩΝ | 14 | 48619,00 | 7075,230 | 1890,935 |

**One-Sample Test**

| | Test Value = 50000 | | | | | |
|---|---|---|---|---|---|---|
| | | | | | 95% Confidence Interval of the Difference | |
| | t | df | Sig. (2-tailed) | Mean Difference | Lower | Upper |
| ΑΡΙΘΜΟΣ ΘΡΗΣΚΕΥΤΙΚΩΝ ΓΑΜΩΝ | -,730 | 13 | ,478 | -1381,000 | -5466,12 | 2704,12 |

Η διαφορά των δύο μέσων ισούται με -1381 και το διάστημα εμπιστοσύνης της διαφοράς τους είναι το (-5466.12,2704.12).

Την ίδια διαδικασία ακολουθούμε και στην περίπτωση των πολιτικών γάμων όπου βάζω τιμή του test το 7000.

**One-Sample Statistics**

| | N | Mean | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|
| ΑΡΙΘΜΟΣ ΠΟΛΙΤΙΚΩΝ ΓΑΜΩΝ | 14 | 8344,93 | 2880,355 | 769,807 |

**One-Sample Test**

| | Test Value = 7000 | | | | | |
|---|---|---|---|---|---|---|
| | | | | | 95% Confidence Interval of the Difference | |
| | t | df | Sig. (2-tailed) | Mean Difference | Lower | Upper |
| ΑΡΙΘΜΟΣ ΠΟΛΙΤΙΚΩΝ ΓΑΜΩΝ | 1,747 | 13 | ,104 | 1344,929 | -318,14 | 3008,00 |

Εδώ, το διάστημα εμπιστοσύνης διαφοράς των μέσων είναι το (-318.14,3008).

*<u>ΠΑΡΑΔΕΙΓΜΑ 4.3</u>*

Το τρίτο παράδειγμά μας παρουσιάζει το βάρος 12 ατόμων πριν και μετά τη χρήση φαρμάκου που διατίθεται για την απώλεια βάρους.

1 : AFTER 55,2

| | AFTER | BEFORE | var | var | var | var |
|---|---|---|---|---|---|---|
| 1 | 55,2 | 55,4 | | | | |
| 2 | 63,6 | 63,9 | | | | |
| 3 | 58,8 | 60,1 | | | | |
| 4 | 77,2 | 78,8 | | | | |
| 5 | 58,5 | 59,2 | | | | |
| 6 | 69,2 | 68,7 | | | | |
| 7 | 59,5 | 59,9 | | | | |
| 8 | 70,0 | 70,0 | | | | |
| 9 | 68,9 | 69,2 | | | | |
| 10 | 74,0 | 73,7 | | | | |
| 11 | 83,9 | 84,9 | | | | |
| 12 | 74,8 | 75,3 | | | | |
| 13 | . | . | | | | |
| 14 | . | . | | | | |

Data View | Variable View

Εδώ για να βρούμε το διάστημα εμπιστοσύνης της διαφοράς των μέσων των δύο πληθυσμών, εφαρμόζουμε το Paired Samples t-test.Αυτό γίνεται επειδή έχουμε πλήρη στοιχεία για τους δύο πληθυσμούς,άρα θεωρούνται γνωστές οι διακυμάνσεις τους, ενώ στο προηγούμενο παράδειγμα αυτό δεν ίσχυε.

**Paired Samples Statistics**

| | | Mean | N | Std. Deviation | Std. Error Mean |
|---|---|---|---|---|---|
| Pair 1 | AFTER | 67,800 | 12 | 8,8204 | 2,5462 |
| | BEFORE | 68,258 | 12 | 8,9147 | 2,5735 |

**Paired Samples Correlations**

| | | N | Correlation | Sig. |
|---|---|---|---|---|
| Pair 1 | AFTER & BEFORE | 12 | ,998 | ,000 |

**Paired Samples Test**

| | | Paired Differences | | | | | t | df | Sig. (2-tailed) |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | 95% Confidence Interval of the Difference | | | | |
| | | Mean | Std. Deviation | Std. Error Mean | Lower | Upper | | | |
| Pair 1 | AFTER - BEFORE | -,4583 | ,6171 | ,1781 | -,8504 | -,0662 | -2,573 | 11 | ,026 |

Βλέπουμε λοιπόν ότι το διάστημα εμπιστοσύνης της διαφοράς των δύο μέσων είναι (-0.8504,-0.662).

# ΒΙΒΛΙΟΓΡΑΦΙΑ

**Γναρδέλλης, Χ.,** *Ανάλυση Δεδομένων με το SPSS 14.0 for Windows.* Εκδ. Παπαζήση, Αθήνα, 2003.

**Γναρδέλλης, Χ.,** *Εφαρμοσμένη Στατιστική.* Εκδ. Παπαζήση, Αθήνα, 2003.

**Κιντής, Α.,** *Σύγχρονη Στατιστική Ανάλυση.* Εκδ. Gutenberg, Αθήνα, 2002.

**Κιόχος, Π.,** *Στατιστική.* Εκδ. INTERBOOKS, Αθήνα, 1993.

**Χαλικιάς, Ι.,**.*Στατιστικη-Μέθοδοι Ανάλυσης για Επιχειρηματικές Αποφάσεις.* Εκδ. Rosili, Αθήνα, 2003.