1 ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΜΕΡΟΣ Α

1.1 Εισαγωγή

Στατιστική *(statistics)* είναι η επιστήμη ή «τέχνη» του να μαθαίνουμε από τα δεδομένα. Η στατιστική συνίσταται στη συλλογή δεδομένων που λέγεται **δειγματοληψία** *(sampling)*, στην περιγραφή τους που λέγεται **περιγραφική στατιστική** *(descriptive statistics)* και κυρίως στην ανάλυση τους που οδηγεί και στην απόκτηση συμπερασμάτων και αναφέρεται ως **στατιστική συμπερασματολογία** *(statistical inference)* ή απλά **στατιστική**. Σε αυτό και το επόμενο κεφάλαιο θα ασχοληθούμε με τα δύο τελευταία θέματα. Συγκεκριμένα, σε αυτό το κεφάλαιο θα παρουσιαστούν απλά γραφήματα και πίνακες για την παρουσίαση και σύνοψη των στατιστικών δεδομένων. Θα δοθεί επίσης η βασική μεθοδολογία για την εκτίμηση της μέσης τιμής καθώς και τη σύγκριση δεδομένων ως προς τη μέση τιμή από διαφορετικές ομάδες ή με διαφορετικές ιδιότητες. Γι αυτό θα χρησιμοποιήσουμε τη λεγόμενη **ανάλυση διασποράς** (*Analysis of Variance*, ANOVA).

Η παρουσίαση της στατιστικής μεθοδολογίας γίνεται με παραδείγματα από συγκοινωνιακά θέματα. Δίνεται έμφαση στην πρακτική εφαρμογή με χρήση του στατιστικού λογισμικού SPSS.

1.2 Ορολογία – Βασικές έννοιες

Όλες οι παρατηρήσεις που συλλέγουμε είτε από οργανωμένα πειράματα ή από απλές καταγραφές αποτελούν τα **στατιστικά στοιχεία** ή **δεδομένα** *(data)* που θέλουμε να επεξεργαστούμε με στατιστικές μεθόδους για να καταλήξουμε σε συμπεράσματα. Τα δεδομένα αυτά συλλέγονται από μια καθορισμένη συλλογή στοιχείων που αποτελεί τον **πληθυσμό** *(population)* που μας ενδιαφέρει. Η παρατήρηση όλων του στοιχείων του πληθυσμού είναι πρακτικά πολύ δύσκολη ή αδύνατη, γι αυτό συλλέγουμε ένα μικρό υποσύνολο του πληθυσμού που λέγεται **δείγμα** *(sample)* με κάποιον προκαθορισμένο τρόπο.

Οποιοδήποτε χαρακτηριστικό του οποίου η τιμή αλλάζει από το ένα στοιχείο του πληθυσμού στο άλλο λέγεται **τυχαία μεταβλητή** *(random variable)* και για συντομία θα γράφουμε **τ.μ.**. Στο εξής θα χρησιμοποιούμε κεφαλαία πλάγια λατινικά γράμματα για να δηλώνουμε τις μεταβλητές, όπως

X = μέσο μεταφοράς, Y = χρόνος κάλυψης κάποιας διαδρομής,

και με μικρούς πλάγιους λατινικούς χαρακτήρες θα συμβολίζουμε τις παρατηρήσεις, όπως $x_1, x_2, ..., x_n$ είναι n παρατηρήσεις της μεταβλητής X. Οι τιμές που παίρνει μια τ.μ. μπορεί να είναι κατηγορίες και τότε λέγεται **ποιοτική** *(qualitative)* **τ.μ.**. Η τ.μ. μπορεί επίσης να παίρνει αριθμητικές τιμές σε κάποια μονάδα μέτρησης και τότε λέγεται **ποσοτική** *(quantitative)* **τ.μ.**. Μια ποσοτική τ.μ. που παίρνει τιμές από ένα σύνολο διακεκριμένων τιμών λέγεται **διακριτή** *(discrete)* σε αντίθεση με μια ποσοτική τ.μ. που παίρνει τιμές σ' ένα συνεχές διάστημα και λέγεται **συνεχής** *(continuous)*. Με αντίστοιχο τρόπο διακρίνουμε και τα δεδομένα, π.χ. τα δεδομένα χρόνου κάλυψης μια απόστασης (δρομολόγιο με τρένο) είναι συνεχή ποσοτικά.

1.3 Περιγραφική Στατιστική

Στην αρχή μιας στατιστικής μελέτης συλλέγουμε παρατηρήσεις για μια τ.μ. *X* που μας ενδιαφέρει, που λέγεται και **μεταβλητή ενδιαφέροντος** *(variable of interest)*. Η περιγραφική στατιστική συνίσταται στην παρουσίαση των δεδομένων για κάθε τέτοια μεταβλητή με στατιστικούς πίνακες και διαγράμματα καθώς και στον υπολογισμό συνοπτικών μέτρων. Οι πίνακες, τα διαγράμματα και τα συνοπτικά μέτρα μας βοηθούν να παρατηρήσουμε σημαντικά χαρακτηριστικά των δεδομένων, όπως την κεντρική τάση, το εύρος και τη συμμετρικότητα τους.

1.3.1 Παρουσίαση ποιοτικών ή ποσοτικών διακριτών δεδομένων

Για ποιοτικά η διακριτά ποσοτικά δεδομένα μπορούμε να υπολογίσουμε τη **συχνότητα** εμφάνισης f_i στο δείγμα $x_1, x_2, ..., x_n$, της κάθε διακεκριμένης τιμής a_i για *m* διακεκριμένες τιμές $a_1, a_2, \ldots, a_m$. Η **σχετική συχνότητα** *(relative frequency)* εμφάνισης ή αλλιώς το **ποσοστό** *(percent)* p_i είναι $p_i = \frac{f_i}{n}$. Για διατακτικά κατηγορικά δεδομένα (δηλαδή για κατηγορίες που μπορούν να τεθούν σε διάταξη) ή διακριτά αριθμητικά δεδομένα, υπολογίζεται **η αθροιστική συχνότητα** F_i για την τιμή a_i ως $F_i = \sum_{j=1}^{i} f_j$ όπου οι δείκτες *i* και *j* αντιστοιχούν στις τιμές a_i και a_j και είναι $a_j < a_i$. Με τον ίδιο τρόπο ορίζεται και η **αθροιστική σχετική συχνότητα** P_i.

Ένας **πίνακας συχνοτήτων** *(frequency table)* περιλαμβάνει όλα τα παραπάνω μεγέθη. Κάθε ένα από τα f_i, p_i, F_i και P_i, $i = 1, \ldots, m$, μπορεί να παρασταθεί γραφικά σ' ένα **ραβδόγραμμα** *(bar chart)*, όπου η κάθε ράβδος παρουσιάζει την αντίστοιχη συχνότητα για κάθε τιμή a_i, ή σε ένα **κυκλικό διάγραμμα** ή **διάγραμμα πίτας** *(pie chart)* όπου το κάθε κομμάτι της επιφάνειας του κύκλου («πίτα») παρουσιάζει την αντίστοιχη συχνότητα.

<u>Παράδειγμα:</u> Ας θεωρήσουμε ένα δείγμα για τη χρήση μέσου μεταφοράς προς την εργασία (λεωφορείο, λεωφορείο-μετρό, αυτοκίνητο, άλλο) από 100 εργαζόμενους που επιλέχθηκαν τυχαία. Γι αυτό το δείγμα μπορούμε να σχηματίσουμε πίνακα και γραφήματα συχνοτήτων (και σχετικών συχνοτήτων) αλλά όχι αθροιστικών συχνοτήτων, αφού δεν υπάρχει σχέση διάταξης μεταξύ των τιμών της τ.μ. «μέσο μεταφοράς».

1.3.2 Ομαδοποίηση και παρουσίαση αριθμητικών δεδομένων

Όταν τα δεδομένα είναι αριθμητικά και ο αριθμός των διακεκριμένων τιμών είναι μεγάλος ή οι τιμές ανήκουν σ' ένα διάστημα τιμών, τότε πρέπει πρώτα να τα χωρίσουμε σε **ομάδες** *(groups)*, ή κλάσεις διαστημάτων, και μετά να υπολογίσουμε, όπως πριν, τη συχνότητα για κάθε ομάδα (δηλαδή τον αριθμό των δεδομένων σε κάθε ομάδα). Το εύρος τιμών της κάθε ομάδας είναι συνήθως το ίδιο και το διαλέγουμε ανάλογα με την κλίμακα τιμών για την οποία μας ενδιαφέρει να δούμε διαφορές.

Τα γραφήματα σχηματίζονται με βάση τις ομάδες. Ειδικότερα για το ραβδόγραμμα, η κάθε ράβδος υψώνεται στο κέντρο του διαστήματος της αντίστοιχης ομάδας. Επίσης δεν υπάρχει διάστημα μεταξύ των ράβδων. Το γράφημα αυτό λέγεται **ιστόγραμμα** *(histogram)*. Ένας άλλος χωρισμός σε ομάδες με βάση τα ψηφία των μετρήσεων δίνει το **φυλλογράφημα** *(stem and leaf plot)*.

Από το ιστόγραμμα ή το φυλλογράφημα των αριθμητικών τιμών του δείγματος μπορούμε να αναγνωρίσουμε χαρακτηριστικά της κατανομής της τ.μ. που μελετάμε. Ειδικότερα μας ενδιαφέρει αν η κατανομή είναι συμμετρική (δεν παρουσιάζει λοξότητα) και δεν έχει μεγάλες ουρές (άκρα της κατανομής), αν δηλαδή έχει σχήμα καμπάνας. Τέτοια κατανομή είναι η **κανονική κατανομή** (*normal distribution*). Θεωρώντας κανονική κατανομή για τα δεδομένα διευκολύνεται η στατιστική ανάλυση καθώς υπάρχει συγκεκριμένη μεθοδολογία σε αυτήν την περίπτωση.

Παράδειγμα: Αν έχουμε ένα δείγμα από χρόνους ημι-αστικής διαδρομής συγκεκριμένου μήκους με το ίδιο μέσο μεταφοράς σε 20 διαφορετικές περιοχές σε μια πόλη, το ιστόγραμμα θα μπορούσε να μας δείξει αν οι χρόνοι φαίνεται να μαζεύονται γύρω από μια κεντρική τιμή ή κατανέμονται ομοιόμορφα.

1.3.3 Συνοπτικά μέτρα στατιστικών δεδομένων

Τα συνοπτικά μέτρα μπορεί να αναφέρονται σε χαρακτηριστικές θέσεις (κυρίως μας ενδιαφέρει η θέση της κεντρικής τάσης) ή στη μεταβλητότητα των δεδομένων.

Μέτρα κεντρικής τάσης

Έστω $x_1, x_2, ..., x_n$, οι τιμές των παρατηρήσεων του δείγματος για μια μεταβλητή *X* που μελετάμε. Τα κυριότερα μέτρα κεντρικής τάσης είναι η **δειγματική μέση τιμή** *(sample mean)* ή **μέσος όρος** *(average)* και η **δειγματική διάμεσος** *(sample median)*.

Η δειγματική μέση τιμή είναι το «κέντρο ισορροπίας» των δεδομένων, συμβολίζεται $\bar{x}$ κι ορίζεται ως

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \,. \qquad (1.1)$$

Η δειγματική διάμεσος $\tilde{x}$ ορίζεται ως η κεντρική τιμή όταν διατάξουμε τα δεδομένα σε αύξουσα σειρά. Αν ο αριθμός *n* των δεδομένων είναι περιττός τότε $\tilde{x}$ είναι η τιμή στη θέση $(n+1)/2$, ενώ αν το *n* είναι άρτιος τότε $\tilde{x}$ είναι το ημιάθροισμα των τιμών στις θέσεις $n/2$ και $n/2+1$.

Μέτρα μεταβλητότητας

Διαφορετικά δείγματα από τον ίδιο πληθυσμό μπορεί να έχουν το ίδιο μέτρο κεντρικής τάσης αλλά να σκορπίζονται περισσότερο ή λιγότερο γύρω από το κέντρο. Κύρια μέτρα διασποράς των δεδομένων είναι η **δειγματική διακύμανση** ή **δειγματική διασπορά** *(sample variance)*, s^2, η **δειγματική τυπική απόκλιση** *(standard deviation)*, *s*, καθώς και το **ενδοτεταρτομοριακό εύρος** *I (interquartile range)*.

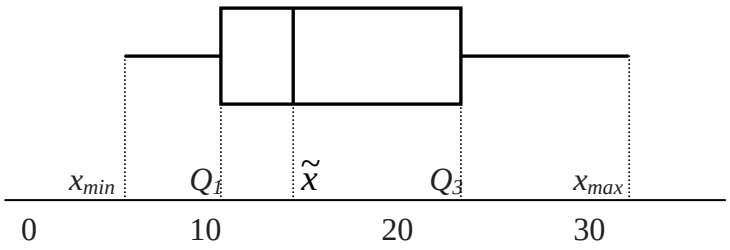
Η δειγματική διασπορά ή διακύμανση μετράει τη μεταβλητότητα των παρατηρήσεων γύρω από τη δειγματική μέση τιμή κι ορίζεται ως

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right). \qquad (1.2)$$

Η διασπορά s^2 προκύπτει από τα τετράγωνα των παρατηρήσεων και γι αυτό είναι δύσκολο να την ερμηνεύσουμε. Γι αυτό συνήθως χρησιμοποιούμε τη δειγματική τυπική απόκλιση *s*, που είναι απλά η θετική ρίζα της δειγματικής διασποράς s^2, μετριέται στην ίδια μονάδα μέτρησης με τα δεδομένα κι εκφράζει πόσο μια τυπική τιμή της μεταβλητής απέχει από τη μέση τιμή.

Μια παρατήρηση ονομάζεται το *p*-εκατοστιαίο σημείο όταν το πολύ *p*% του συνόλου των παρατηρήσεων είναι μικρότερες απ' αυτήν την παρατήρηση. Η διάμεσος είναι το 50-εκατοστιαίο σημείο. Αλλά χαρακτηριστικά εκατοστιαία σημεία είναι το 25-εκατοστιαίο σημείο, δηλαδή το **πρώτο** ή **κατώτερο τεταρτομόριο** *(first or lower quartile)* Q_1 και το 75-εκατοστιαίο σημείο, δηλαδή το **τρίτο** ή **ανώτερο τεταρτομόριο** *(third or upper quartile)* Q_3. Τα Q_1 και Q_3 ορίζονται όπως η διάμεσος αλλά περιορίζοντας το σύνολο των δεδομένων στα αντίστοιχα υποσύνολα (κατώτερο ή ανώτερο μισό, αντίστοιχα). Η διαφορά $I=Q_3 - Q_1$, λέγεται *ενδοτεταρτομοριακό εύρος* και δίνει το εύρος που καλύπτουν τα μισά από τα δεδομένα που είναι πιο κοντά στην κεντρική τιμή (διάμεσο).

Η διάμεσος, το πρώτο και τρίτο τεταρτομόριο και η ελάχιστη και μέγιστη τιμή των δεδομένων αποτελούν τη **σύνοψη των 5 αριθμών** *(five number summary)*. Γραφικά η παρουσίαση της σύνοψης των 5 αριθμών γίνεται με το **θηκόγραμμα** *(box plot)* σε οριζόντια ή κάθετη θέση όπως δείχνει το παρακάτω σχήμα.



Εικόνα 1.1 Σχηματική παρουσίαση οριζόντιου θηκογράμματος.

Η ύπαρξη μεμονωμένων παρατηρήσεων που διαφέρουν σημαντικά από τις υπόλοιπες παρατηρήσεις του δείγματος δυσκολεύει τη στατιστική περιγραφή και ανάλυση. Γι αυτό θα πρέπει πριν προχωρήσουμε να βεβαιωθούμε αν η μακρινή παρατήρηση είναι σωστή και πρέπει να συμπεριληφθεί ή αν υποπτευόμαστε ότι μπορεί να οφείλεται σε λάθος της μέτρησης και να την αγνοήσουμε. Στο σχηματισμό του θηκογράμματος στο SPSS, τιμές που υπερβαίνουν κάποια όρια χαρακτηρίζονται **ύποπτες απόμακρες τιμές** *(outliers)* ή **απόμακρες τιμές** *(extreme values)* και δηλώνονται ως ξεχωριστά σημεία στο θηκόγραμμα (`Graphs > Boxplot` ή `Graphs > Interactive > Boxplot`). Η δειγματική μέση τιμή και η δειγματική διασπορά επηρεάζονται σημαντικά από την ύπαρξη απόμακρων τιμών ενώ η διάμεσος και το ενδοτεταρτομοριακό εύρος όχι.

Γενικά στο SPSS υπάρχουν μια σειρά από γραφήματα στο μενού `Graphs`, ενώ η στατιστική ανάλυση είναι στο μενού `Analyze`.

1.4 Εκτιμητική

Η κατανομή μιας τ.μ. *X* χαρακτηρίζεται από κάποιες παραμέτρους. Για παράδειγμα η κανονική κατανομή ορίζεται από τη μέση τιμή μ και τη διασπορά σ^2. Από το δείγμα των παρατηρήσεων συχνά θέλουμε να εκτιμήσουμε κάποια παράμετρο. Γι αυτό υπολογίζουμε μια τιμή που αντιπροσωπεύει την παράμετρο καλύτερα, κάνουμε δηλαδή **σημειακή εκτίμηση** *(point estimation)*. Επίσης υπολογίζουμε ένα διάστημα τιμών που περιέχει την αληθινή τιμή της παραμέτρου με κάποια μεγάλη πιθανότητα *1-α*, το οποίο λέγεται **(1-α)% διάστημα εμπιστοσύνης** *(confidence interval)*. Το *α* λέγεται και **στάθμη σημαντικότητας** *(significance level)*.

Ιδιαίτερο ενδιαφέρον παρουσιάζει η εκτίμηση της μέση τιμής μ μιας τ.μ. X και της διαφοράς μέσων τιμών $\mu_1 - \mu_2$ δύο τ.μ. X_1 και X_2.

Εκτίμηση μέσης τιμής

Για τη μέση τιμή μ μιας τ.μ. X, η καλύτερη σημειακή εκτίμηση είναι η δειγματική μέση τιμή $\bar{x}$ (δες (1.1)). Για τον ακριβή υπολογισμό του διαστήματος εμπιστοσύνης της μ χρειάζεται να γνωρίζουμε αν η κατανομή της τ.μ. X είναι κανονική, αν η διασπορά σ^2 είναι γνωστή και αν το μέγεθος του δείγματος είναι μεγάλο ($n > 30$) ή μικρό. Στον Πίνακας 1.1 δίνεται το διάστημα εμπιστοσύνης στις διάφορες περιπτώσεις.

Διασπορά	*κατανομή της X*	*N*	*διάστημα εμπιστοσύνης*
γνωστή	κανονική		$\bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$
γνωστή	μη κανονική	$n > 30$	$\bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$
γνωστή	μη κανονική	$n < 30$	μη-παραμετρικό
άγνωστη		$n > 30$	$\bar{x} \pm z_{1-\alpha/2}\frac{s}{\sqrt{n}}$
άγνωστη	κανονική	$n < 30$	$\bar{x} \pm t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}$
άγνωστη	μη κανονική	$n < 30$	μη-παραμετρικό

Πίνακας 1.1 Εκτίμηση του δ.ε. της μ ανάλογα με τη γνώση της διασποράς και κατανομής της τ.μ. X καθώς και του μεγέθους n του δείγματος.

Το διάστημα εμπιστοσύνης ορίζεται με βάση την κατανομή που ακολουθεί ο εκτιμητής $\bar{x}$, ο οποίος θεωρείται επίσης τ.μ.. Το διάστημα εμπιστοσύνης έχει τη γενική μορφή

$$\bar{x} \pm (\text{κρίσιμη τιμή}) \times s_{\bar{x}} \ , \qquad (1.3)$$

όπου $s_{\bar{x}}$ είναι η **εκτίμηση της τυπικής απόκλισης** ή (όπως συνηθίζεται) **του τυπικού σφάλματος** *(estimated standard error)* του εκτιμητή της μέση τιμής $\bar{x}$. Στο γενικό ορισμό του $(1-\alpha)\%$ διαστήματος εμπιστοσύνης στην (1.3) η κρίσιμη τιμή δίνεται από την κατανομή της $\bar{x}$, που κατά περίπτωση είναι η κρίσιμη τιμή της τυπικής κανονικής κατανομής $z_{1-a/2}$ ή η κρίσιμη τιμή της κατανομής student ή t-κατανομής με *n-1* βαθμούς ελευθερίας $t_{n-1,1-a/2}$.

Για μικρά δείγματα μιας τ.μ. X που δε φαίνεται να ακολουθεί κανονική κατανομή (με βάση το ιστόγραμμα ή το θηκόγραμμα) δε γνωρίζουμε την κατανομή της $\bar{x}$ και το διάστημα εμπιστοσύνης δε μπορεί να οριστεί παραμετρικά (δηλαδή με βάση κάποια γνωστή κατανομή). Σε αυτήν την περίπτωση εκτιμούμε διάστημα εμπιστοσύνης για τη διάμεσο χρησιμοποιώντας μη-παραμετρική μέθοδο. Γενικά μη-παραμετρικές μέθοδοι δε θα μας απασχολήσουν εδώ γι αυτό είναι σημαντικό να εξετάζουμε αν η κατανομή της τ.μ. ενδιαφέροντος είναι κανονική.

<u>Παράδειγμα:</u> Για το παράδειγμα του χρόνου κάλυψης ημι-αστικής διαδρομής συγκεκριμένου μήκους με το ίδιο μέσο μεταφοράς σε 20 διαφορετικές περιοχές, χρησιμοποιώντας τον τύπο (1.1) βρίσκουμε τη σημειακή εκτίμηση του μέσου χρόνου της διαδρομής. Αν η κατανομή των χρόνων φαίνεται να είναι κανονική μπορούμε να υπολογίσουμε το διάστημα εμπιστοσύνης του μέσου χρόνου γι αυτό το μήκος της διαδρομής με κάποια εμπιστοσύνη επιπέδου $(1-\alpha)\%$ *κάνοντας χρήση του τύπου με την κρίσιμη t-τιμή (προτελευταία σειρά στον Πίνακας 1.1).*

<u>Εκτίμηση διαφοράς μέσων τιμών</u>

Συχνά μας ενδιαφέρει να εκτιμήσουμε αν δύο ανεξάρτητες τ.μ. X_1 και X_2 διαφέρουν ως προς τη μέση τιμή τους μ_1 και μ_2 αντίστοιχα. Γι αυτό εκτιμούμε το $(1-\alpha)\%$ διάστημα εμπιστοσύνης για τη διαφορά $\mu_1-\mu_2$ και ελέγχουμε αν αυτό περιέχει το 0. Αν το διάστημα εμπιστοσύνης είναι θετικό, σημαίνει ότι σε στάθμη εμπιστοσύνης $(1-\alpha)\%$ η μέση τιμή μ_1 είναι μεγαλύτερη από τη μ_2 κατά ποσό που δίνεται από το διάστημα εμπιστοσύνης. Αντίστοιχα ερμηνεύεται ένα αρνητικό διάστημα εμπιστοσύνης. Τέλος αν το διάστημα εμπιστοσύνης περιέχει το 0, δεν υπάρχει σημαντική διαφορά.

διασπορές των X_1, X_2	*Κατανομές των X_1, X_2*	*n_1, n_2*	*διάστημα εμπιστοσύνης για $\mu_1-\mu_2$*
γνωστές	κανονική		$\left(\bar{x}_1-\bar{x}_2\right)\pm z_{1-\alpha/2}\cdot\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$
γνωστές	μη κανονική	μεγάλα	$\left(\bar{x}_1-\bar{x}_2\right)\pm z_{1-\alpha/2}\cdot\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$
γνωστές	μη κανονική	μικρά	μη-παραμετρικό
άγνωστες άνισες ή ίσες		μεγάλα	$\left(\bar{x}_1-\bar{x}_2\right)\pm z_{1-\alpha/2}\cdot\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$
άγνωστες και ίσες	κανονική	μικρά	$\left(\bar{x}_1-\bar{x}_2\right)\pm t_{n_1+n_2-2,1-\alpha/2}\cdot s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$
άγνωστες και ίσες	μη κανονική	μικρά	μη-παραμετρικό
άγνωστες και άνισες		μικρά	----

Πίνακας 1.2 Εκτίμηση του δ.ε. της διαφοράς μ_1 - μ_2 ανάλογα με τη γνώση των διασπορών σ_1^2, σ_2^2 και κατανομών των τ.μ. X_1 και X_2 καθώς και των μεγεθών n_1 και n_2 των αντιστοίχων δειγμάτων.

Στον Πίνακας 1.2 δίνεται το διάστημα εμπιστοσύνης της $\mu_1-\mu_2$ στις διάφορες περιπτώσεις, όπως και για τη μέση τιμή. Εδώ εξετάζουμε επίσης αν οι διασπορές των X_1 και X_2 δε διαφέρουν σημαντικά έτσι ώστε να μπορούν να θεωρηθούν ίσες. Στην περίπτωση που οι διασπορές είναι άγνωστες και ίσες, εκτιμούνται από την κοινή διασπορά *(pooled variance)*

$$s_p^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}, \qquad (1.4)$$

όπου s_1^2 και s_2^2 είναι οι δειγματικές διασπορές για τις τ.μ. X_1 και X_2 αντίστοιχα.

Παράδειγμα: Για το παράδειγμα με το χρόνο κάλυψης ημι-αστικής διαδρομής συγκεκριμένου μήκους με το ίδιο μέσο σε 20 διαφορετικές περιοχές σε μια πόλη Α θεωρούμε ότι έχουμε επίσης άλλες 15 μετρήσεις από μια άλλη πόλη Β. Υποθέτοντας κανονικές κατανομές και ίδιες διασπορές για τους χρόνους διαδρομής στις δύο πόλεις, μπορούμε να υπολογίσουμε από τον τύπο (1.4) την κοινή διασπορά από τις δειγματικές διασπορές για τις πόλεις Α και Β. Στη συνέχεια μπορούμε να βρούμε το $(1-\alpha)\%$ *διάστημα εμπιστοσύνης για τη διαφορά των μέσων χρόνων διαδρομής για τις δύο πόλεις κάνοντας χρήση του τύπου με την κρίσιμη t-τιμή (έκτη σειρά στον Πίνακας 1.2).*

1.5 Έλεγχος υπόθεσης

Απάντηση στο ερώτημα αν διαφέρουν οι μέσες τιμές των τ.μ. X_1 και X_2 μπορούμε επίσης να δώσουμε χρησιμοποιώντας έλεγχο στατιστικής υπόθεσης. Γενικά στον έλεγχο στατιστικής υπόθεσης ξεκινάμε με μια **μηδενική υπόθεση** *(null hypothesis)* H_0 που συνήθως θέλουμε να απορρίψουμε για να δεχτούμε την **εναλλακτική υπόθεση** *(alternative hypothesis)* H_1. Για παράδειγμα για τη σύγκριση δύο τ.μ. ως προς τη μέση τιμή τους οι υποθέσεις είναι:

$$H_0: \mu_1 = \mu_2 \qquad H_1: \mu_1 \neq \mu_2 .$$

Στη συνέχεια επιλέγουμε μια κατάλληλη **στατιστική ελέγχου** *(test statistic) q* που ακολουθεί γνωστή κατανομή (π.χ. $q \equiv z \sim \mathrm{N}(0,1)$, $q \equiv t \sim \mathrm{t}_{n_1+n_2-2}$, το σύμβολο $\equiv$ ορίζει ταυτόσημο συμβολισμό). Ορίζεται ένα σύνολο τιμών της *q* που είναι απίθανο (σε στάθμη σημαντικότητας *α*) να πάρει η *q* αν ισχύει η H_0, το οποίο ονομάζεται **απορριπτική περιοχή** *(rejection region), R*. Αυτό το σύνολο αντιστοιχεί στις ουρές της κατανομής της *q*. Για παράδειγμα, για την H_0: $\mu_1 = \mu_2$, χρησιμοποιώντας $q \equiv z$ είναι $R = \left\{ |z| > z_{1-\alpha/2} \right\}$.

Υπολογίζουμε τη δειγματική στατιστική ελέγχου $\tilde{q}$, δηλαδή την τιμή της *q* στο δείγμα, και εξετάζουμε αν αυτή ανήκει στο *R* για να απορρίψουμε την H_0. Στον έλεγχο χρησιμοποιείται συχνά η ***p*-τιμή** *(p-value)*, που δηλώνει τη χαμηλότερη στάθμη σημαντικότητας *α* για την οποία μπορούμε να απορρίψουμε την H_0 με βάση το δείγμα. Στην περίπτωση όπου $q \equiv z$, είναι $p = P\left(|z| > \tilde{z}\right)$, όπου $\tilde{z}$ η τιμή της στατιστικής *q* από το δείγμα που είναι ίδια με την κρίσιμη τιμή $z_{1-p/2}$.

Υπάρχει πλήρης συμφωνία του αποτελέσματος από το διάστημα εμπιστοσύνης και τον στατιστικό έλεγχο στην ίδια στάθμη σημαντικότητας *α*. Για παράδειγμα αν απορρίψουμε την H_0: $\mu = \mu_0$ (η μηδενική υπόθεση για να ελέγξουμε αν η μέση τιμή μ μιας τ.μ. *X* μπορεί να πάρει κάποια τιμή μ_0) για $\alpha = 0.05$ τότε το 95% διάστημα εμπιστοσύνης της μ δεν περιέχει την τιμή μ_0. Το ίδιο ισχύει για τη σύγκριση των μέσων τιμών των τ.μ. X_1 και X_2. Η στατιστική ελέγχου *q* γι αυτούς τους ελέγχους ορίζεται κατά περίπτωση όπως και για τα αντίστοιχα διαστήματα εμπιστοσύνης (δες Πίνακας 1.1 και Πίνακας 1.2).

Ο έλεγχος μπορεί επίσης να είναι **μονόπλευρος** *(one-sided test)*, π.χ. H_0: $\mu_1 \leq \mu_2$, H_1: $\mu_1 \geq \mu_2$ και τότε ορίζεται αντίστοιχα μονόπλευρα η απορριπτική περιοχή, π.χ. $R = \left\{ z > z_{1-\alpha} \right\}$.

Στο SPSS υπάρχει δυνατότητα για πραγματοποίηση ελέγχου και ταυτόχρονα υπολογισμό διαστήματος εμπιστοσύνης για τη μέση τιμή (`Compare Means > One-Sample T test`) και για σύγκριση δύο μέσων τιμών από ανεξάρτητα δείγματα (`Compare Means > Independent Samples T Test`), αλλά μόνο για άγνωστες διασπορές (χρήση κατανομής student, δες Πίνακας 1.1 και Πίνακας 1.2).

1.6 Ανάλυση Διασποράς

Σε πιο σύνθετα προβλήματα, όπου π.χ. δεν έχουμε απλά να εκτιμήσουμε τη μέση τιμή ενός μεγέθους ή να συγκρίνουμε δύο ομάδες ως προς κάποιο χαρακτηριστικό τους, όπως είδαμε πριν, χρειάζεται να οργανώσουμε τη στατιστική μελέτη και να κάνουμε **σχεδιασμό πειράματος** *(experimental design)*. Για παράδειγμα σχεδιασμό πειράματος χρειάζεται να κάνουμε όταν έχουμε να συγκρίνουμε τρεις οι παραπάνω τ.μ. ως προς τη μέση τιμή τους.

Η μέθοδος στατιστικής ανάλυσης που χρησιμοποιούμε λέγεται **ανάλυση διασποράς** (*analysis of variance*, ANOVA). Οι τ.μ. μπορεί να εκφράζουν ομάδες (υπο-πληθυσμούς) από έναν πληθυσμό (π.χ. χρόνος ημι-αστικής διαδρομής ίδιας απόστασης σε τρία αστικά κέντρα) ή διαφορετικά χαρακτηριστικά (π.χ. χρόνος διαδρομής ίδιας απόστασης με αυτοκίνητο, λεωφορείο, συνδυασμένη χρήση λεωφορείου-μετρό, άλλο μεταφορικό μέσο). Άρα οι μετρήσεις των διαφορετικών τ.μ. είναι ουσιαστικά μετρήσεις μιας μεταβλητής ενδιαφέροντος (π.χ. ο χρόνος κάλυψης μιας διαδρομής) άλλα σε διαφορετικές ομάδες ενός **παράγοντα** (*factor*), που αποτελείται από υπο-πληθυσμούς ή χαρακτηριστικά και πιθανόν να επηρεάζει τη μεταβλητή ενδιαφέροντος.

Σε πιο σύνθετους σχεδιασμούς πειράματος μπορεί να εξετάζουμε την επίδραση περισσοτέρων από έναν παράγοντα στη μεταβλητή ενδιαφέροντος ή να θέλουμε να εξαλείψουμε την επίδραση κάποιας άλλης μεταβλητής (*covariate*) που ίσως να καλύπτει την επίδραση του παράγοντα στη μεταβλητή ενδιαφέροντος.

1.6.1 Μονόδρομη Ανάλυση Διασποράς

Σ' έναν **πλήρως τυχαιοποιημένο σχεδιασμό** *(completely randomized design)* υποθέτουμε πως το πείραμα δίνει τις μετρήσεις της μεταβλητής ενδιαφέροντος *Y* σε *k* διαφορετικές ομάδες (πληθυσμούς ή χαρακτηριστικά), δηλαδή $y_{11}, y_{12}, \ldots, y_{1n_1}$ είναι το δείγμα για την πρώτη ομάδα, $y_{21}, y_{22}, \ldots, y_{2n_2}$ για τη δεύτερη ομάδα κτλ. Θέλουμε να ελέγξουμε την υπόθεση ότι οι μέσες τιμές της *Y* στις *k* ομάδες είναι ίσες

$$\mathrm{H}_0\colon\ \mu_1 = \mu_2 = \cdots = \mu_k .$$

Τυπικά υποθέτουμε ότι η κατανομή της μεταβλητής ενδιαφέροντος στις *k* ομάδες είναι κανονική και η διασπορά είναι η ίδια, άρα οι ομάδες μπορεί να διαφέρουν μόνο ως προς τη μέση τιμή. Η στατιστική ανάλυση που χρησιμοποιείται για να εξετάσουμε αυτήν τη H_0 λέγεται **μονόδρομη ανάλυση διασποράς** *(one way ANOVA)*.

Αν $\bar{y}$ είναι ο μέσος όρος από όλα τα δείγματα *(grand mean)* και $\bar{y}_i$ είναι ο μέσος όρος στο δείγμα *i*, το μοντέλο της μονόδρομης ανάλυσης διασποράς θεωρεί την παρακάτω ανάλυση κάθε μέτρησης y_{ij}

$$\underset{\text{παρατήρηση}}{y_{ij}} = \underset{\text{ολικός μέσος}}{\bar{y}} + \underset{\text{απόκλιση λόγω παράγοντα}}{(\bar{y}_i - \bar{y})} + \underset{\text{τυχαίο σφάλμα}}{(y_{ij} - \bar{y}_i)}$$

Έστω SSA το άθροισμα των τετραγώνων των αποκλίσεων λόγω του παράγοντα (με βάση το δεύτερο όρο στην παραπάνω ανάλυση), δηλαδή το σφάλμα μεταξύ των δειγμάτων *(between-sample error)*. Έστω επίσης SSE το άθροισμα των τετραγώνων των αποκλίσεων μέσα στο κάθε δείγμα *(within-sample error)*, που αντιστοιχεί τον τρίτο όρο στην παραπάνω ανάλυση. Ο λόγος

$$F = \frac{\mathrm{SSA}/(k-1)}{\mathrm{SSE}/(n-k)} \tag{1.5}$$

είναι η στατιστική ελέγχου που, κάτω από την H_0, ακολουθεί κατανομή Fisher, $F_{k-1,n-k}$ με *k-1* βαθμούς ελευθερίας (για το σφάλμα μεταξύ δειγμάτων) και *n-k* βαθμούς ελευθερίας (για το σφάλμα μέσα στο δείγμα). Η απορριπτική περιοχή σε στάθμη σημαντικότητας α είναι $R = \{F > F_{k-1,n-k;a}\}$ και η τιμή $F_{k-1,n-k;a}$ δίνεται σε αντίστοιχο στατιστικό πίνακα. Το SPSS κάνει αυτήν την ανάλυση (`Compare Means > One-Way ANOVA`) και υπολογίζει την τιμή της στατιστικής $\tilde{F}$ από το δείγμα, καθώς και την *p*-τιμή του ελέγχου.

Αν απορρίψουμε την H_0 η ανάλυση συνεχίζεται με τη διερεύνηση διαφορών στις ομάδες ανά ζεύγη κάνοντας **επακόλουθες** ή **πολλαπλές συγκρίσεις** *(follow-up or multiple comparisons)*, υπολογίζοντας δηλαδή διαστήματα εμπιστοσύνης διαφορών μέσων τιμών για κάθε ζευγάρι ομάδων. Επειδή θέλουμε οι $\binom{k}{2}$ συγκρίσεις από κοινού να είναι σωστές, θα πρέπει το κάθε διάστημα εμπιστοσύνης να υπολογισθεί σε μικρότερη στάθμη σημαντικότητας από *α*, έτσι ώστε όλα μαζί να αντιστοιχούν στη στάθμη σημαντικότητας *α*. Υπάρχουν διάφορες τέτοιες διαδικασίες στο SPSS (Bonferroni, Tukey, Dunnett, κτλ.) (επιλογή `Post Hoc…` στο `One-Way ANOVA`).

Για να είναι αξιόπιστα τα αποτελέσματα της (μονόδρομης) ανάλυσης διασποράς είναι σημαντικό να κάνουμε **έλεγχο ορθότητας των υποθέσεων** *(assumption checking)*, οι οποίες είναι ότι η κατανομή στις ομάδες είναι κανονική και οι διασπορές ίσες. Σχηματίζοντας τα θηκογράμματα όλων των ομάδων σ' ένα γράφημα μας επιτρέπει να κάνουμε τον έλεγχο αυτό ποιοτικά. Τα θηκογράμματα πρέπει να μη δείχνουν σημαντικές αποκλίσεις από συμμετρία και να μην έχουν μακριές ουρές ή ακραία σημεία (για να είναι οι κατανομές κανονικές) και το μέγεθος της μεγαλύτερης θήκης (που δηλώνει το ενδοτεταρτομοριακό εύρος) δε θα πρέπει να διαφέρει σημαντικά από το μέγεθος της μικρότερης θήκης (για να είναι ίσες οι διασπορές).

<u>Παράδειγμα:</u> Ας θεωρήσουμε το παράδειγμα με τους χρόνους κάλυψης ημι-αστικής διαδρομής κάποιας απόστασης με το ίδιο μέσο μεταφοράς σε τρεις πόλεις Α, Β, Γ. Αν από την εφαρμογή της μονόδρομής ανάλυσης διασποράς στα δείγματα προκύπτει ότι υπάρχουν διαφορές, δηλαδή απορρίπτεται η μηδενική υπόθεση, τότε η διαδικασία των πολλαπλών συγκρίσεων θα μας καταδείξει σε ποιες πόλεις ο χρόνος διαδρομής διαφέρει.

1.6.2 Ανάλυση διασποράς με δύο παράγοντες

Στο σχεδιασμό πειράματος μπορούμε να εισάγουμε και δεύτερο παράγοντα. Προσπαθούμε με ένα τέτοιο πείραμα να απαντήσουμε στα παρακάτω ερωτήματα:

1. Επηρεάζει ο πρώτος παράγοντας από μόνος του τη μεταβλητή ενδιαφέροντος;
2. Επηρεάζει ο δεύτερος παράγοντας από μόνος του τη μεταβλητή ενδιαφέροντος;
3. Έχουν οι δύο παράγοντες μαζί συνδυασμένη επίδραση στη μεταβλητή ενδιαφέροντος;

Για το κάθε ερώτημα αντιστοιχεί μια μηδενική υπόθεση. Χρησιμοποιώντας την **ανάλυση διασποράς δύο δρόμων** *(two-way ANOVA)* κάνουμε στατιστικό έλεγχο και για τις τρεις υποθέσεις ταυτόχρονα.

Το μοντέλο εδώ είναι πιο πολύπλοκο αλλά η διαδικασία ελέγχου είναι παρόμοια με αυτήν για τη μονόδρομη ανάλυση διασποράς. Ανάλογα με την απόρριψη των μηδενικών υποθέσεων γίνονται και οι επακόλουθες πολλαπλές συγκρίσεις. Οι υποθέσεις για την εφαρμογή της ανάλυση διασποράς δύο δρόμων που πρέπει να ελεγχθούν ως προς την ορθότητα τους είναι και πάλι η κανονική κατανομή σε κάθε ομάδα (και για τους δύο παράγοντες) και η ισότητα των διασπορών.

<u>Παράδειγμα:</u> Θέλουμε να ελέγξουμε αν ο χρόνος κάλυψης ημι-αστικής διαδρομής συγκεκριμένου μήκους επηρεάζεται από την περιοχή στην οποία αναφέρεται (πρώτος παράγοντας) ή από το μέσο μεταφοράς (δεύτερος παράγοντας). Αυτοί οι δύο παράγοντες μπορεί να επηρεάζουν ξεχωριστά το χρόνο διαδρομής. Επίσης μπορεί να έχουν και συνδυασμένη επίδραση στο χρόνο της διαδρομής, π.χ. το λεωφορείο μπορεί να κάνει πιο σύντομα τη διαδρομή σε κάποια συγκεκριμένη περιοχή (αλλά όχι σε όλες). Η ανάλυση διασποράς δύο δρόμων δίνει απαντήσεις σε αυτά τα ερωτήματα.

2 ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΜΕΡΟΣ Β

Σε αυτό το κεφαλαίο θα δούμε μοντέλα που περιγράφουν σχέσεις μεταξύ συνεχών τ.μ.. Συγκεκριμένα θα μελετήσουμε τη γραμμική εξάρτηση μιας τ.μ. από μια ή περισσότερες μεταβλητές. Στη συνέχεια θα ασχοληθούμε με χρονικές σειρές και θα μελετήσουμε μοντέλα που περιγράφουν την εξάρτηση ενός σημείου της χρονοσειράς από τα προηγούμενα. Σε όλες αυτές τις περιπτώσεις η δημιουργία των μοντέλων αποσκοπεί κυρίως στην πρόβλεψη κάποιου μεγέθους, χρησιμοποιώντας τη γνώση μας είτε για άλλα μεγέθη ή για το ίδιο μέγεθος σε προηγούμενους χρόνους.

2.1 Συντελεστής Συσχέτισης

Έστω δύο τ.μ. *X* και *Y* με διασπορά $\sigma_X^2 = \mathrm{Var}[\mathrm{X}]$ και σ_Y^2, αντίστοιχα, και συνδιασπορά $\sigma_{XY} = \mathrm{Cov}[X,Y]$. Η συνδιασπορά εκφράζει τη γραμμική συσχέτιση δύο τ.μ., δηλαδή την αναλογική μεταβολή (αύξηση ή μείωση) της μιας τ.μ. που αντιστοιχεί σε μεταβολή της άλλης μεταβλητής. Η γραμμική αυτή συσχέτιση εκφράζεται καλύτερα με τον **συντελεστή συσχέτισης** (*correlation coefficient*) ρ που ορίζεται ως

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}. \tag{2.1}$$

Το ρ δεν εξαρτάται από τη μονάδα μέτρησης των *X* και *Y* και παίρνει τιμές στο διάστημα [-1,1], όπου τιμές κοντά στο 1 δηλώνουν ισχυρή θετική συσχέτιση (όσο αυξάνει η μια τ.μ. αυξάνει κι άλλη), τιμές κοντά στο -1 δηλώνουν ισχυρή αρνητική συσχέτιση και τιμές κοντά στο 0 δηλώνουν γραμμική ανεξαρτησία των *X* και *Y*.

Ποιοτικά η μορφή της συσχέτισης των *X* και *Y* φαίνεται από την κατανομή των σημείων (x_i, y_i) στο καρτεσιανό σύστημα συντεταγμένων. Αυτό το σχήμα αναφέρεται ως **διάγραμμα διασποράς** *(scatter diagram)*.

2.1.1 Εκτίμηση του συντελεστή συσχέτισης

Η σημειακή εκτίμηση του συντελεστή συσχέτισης ρ του πληθυσμού από το δείγμα των *n* ζευγαρωτών παρατηρήσεων $\{(x_1, y_1),(x_2, y_2),\ldots,(x_n, y_n)\}$ των *X* και *Y* γίνεται με την αντικατάσταση της συνδιασποράς σ_{XY} και των τυπικών αποκλίσεων σ_X και σ_Y από τις αντίστοιχες εκτιμήσεις από το δείγμα

$$\hat{\rho} \equiv r = \frac{s_{XY}}{s_X s_Y}, \tag{2.2}$$

όπου s_X^2 είναι η δειγματική διασπορά του *X* (δες (1.2)) και s_{XY} είναι η εκτίμηση της συνδιασποράς των *X* και *Y*, που ορίζεται ως

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}\right). \tag{2.3}$$

Η εκτίμηση του συντελεστή συσχέτισης της (2.2) λέγεται και **συντελεστής συσχέτισης Pearson** *(Pearson correlation coefficient)* για να διαφοροποιηθεί από άλλες εκτιμήσεις του συντελεστή συσχέτισης.

Αν οι τ.μ. *X* και *Y* ακολουθούν τη δι-μεταβλητή **κανονική κατανομή** *(bivariate normal distribution)* ορίζεται το $(1-\alpha)\%$ διάστημα εμπιστοσύνης του ρ. Επίσης μπορεί να γίνει παραμετρικός έλεγχος για κάποια τιμή του ρ. Ιδιαίτερο ενδιαφέρον παρουσιάζει ο έλεγχος της H_0: $\rho = 0$ γιατί ουσιαστικά εξετάζει αν τα *X* και *Y* είναι γραμμικά ανεξάρτητα. Στο SPSS υπάρχει η δυνατότητα υπολογισμού του συντελεστή συσχέτισης Pearson *r* καθώς και πραγματοποίησης ελέγχου σημαντικότητας του *r* (`Correlate > Bivariate`).

Παράδειγμα: Θέλουμε να ελέγξουμε αν το αντίτιμο του εισιτηρίου αστικού λεωφορείου σε μια πόλη συσχετίζεται με τη χρήση χώρων παρκαρίσματος στο κέντρο της πόλης. Έχοντας συλλέξει κάποιο δείγμα (από ιστορικά στοιχεία), υπολογίζουμε τον συντελεστή συσχέτισης Pearson r. Αν το r βρεθεί να είναι μικρό (όπως ίσως αναμένεται), θα πρέπει να εξετάσουμε αν είναι διαφορετικό του 0 με στατιστική σημαντικότητα, κάνοντας τον κατάλληλο στατιστικό έλεγχο.

2.2 Απλή γραμμική παλινδρόμηση

Η **ανάλυση παλινδρόμησης** (*regression analysis*) περιγράφει τη μεταβλητότητα μιας τ.μ. *Y* χρησιμοποιώντας την πληροφορία που έχουμε για μια ή περισσότερες μεταβλητές X_1, X_2, κτλ.. Το πρόβλημα της παλινδρόμησης είναι η εύρεση ενός μοντέλου που περιγράφει την εξάρτηση της τ.μ. *Y*, που ονομάζεται **εξαρτημένη μεταβλητή** (*dependent or response variable*), από μια μεταβλητή *X* που ονομάζεται **ανεξάρτητη μεταβλητή** (*independent or explanatory variable*) ή κι από περισσότερες από μια μεταβλητές. Στην περίπτωση της **απλής γραμμικής παλινδρόμησης** (*simple linear regression*) η ανεξάρτητη μεταβλητή είναι μία και η εξάρτηση θεωρείται γραμμική.

Παράδειγμα: Θέλουμε να διερευνήσουμε την επίδραση του αριθμού των σηματοδοτημένων διασταυρώσεων στο χρόνο κάλυψης μιας διαδρομής. Η εξαρτημένη τ.μ. Y είναι ο χρόνος διαδρομής, η ανεξάρτητη μεταβλητή X είναι ο αριθμός σηματοδοτημένων διασταυρώσεων και υποθέτουμε πως η εξάρτηση είναι γραμμική.

2.2.1 Μοντέλο απλής γραμμικής παλινδρόμησης

Για τη δημιουργία του μοντέλου της απλής γραμμικής παλινδρόμησης κάνουμε τις εξής υποθέσεις:

1. Η μέση τιμή της τ.μ. *Y* για κάθε τιμή *x* της *X*, $\mathrm{E}[Y \mid X = x]$, είναι γραμμική συνάρτηση της *x*

$$\mathrm{E}[Y \mid X = x] = \alpha + \beta x . \tag{2.4}$$

2. Η διασπορά της τ.μ. *Y* για κάθε τιμή *x* της *X* είναι σταθερή

$$\mathrm{Var}[Y \mid X = x] \equiv \sigma^2_{Y|X} = \sigma^2 . \tag{2.5}$$

Συνήθως υποθέτουμε επίσης ότι η δεσμευμένη κατανομή της *Y* ως προς τη *X* είναι κανονική

$$Y \mid X = x \ \sim \ \mathrm{N}\left(\alpha + \beta x, \sigma^2\right). \tag{2.6}$$

Η παραπάνω συνθήκη επιτρέπει τον παραμετρικό έλεγχο και την παραμετρική εκτίμηση διαστημάτων για τις παραμέτρους α και β καθώς και για την πρόβλεψη $\hat{y}$ για κάποια τιμή της X.

Η τ.μ. y_i για κάποια τιμή x_i της X δίνεται κάτω από την υπόθεση της γραμμικής παλινδρόμησης ως

$$y_i = \alpha + \beta x_i + \varepsilon_i , \tag{2.7}$$

όπου ε_i είναι το **σφάλμα παλινδρόμησης** (*regression error*). Η διασπορά του σφάλματος είναι $\mathrm{Var}[\varepsilon_i] \equiv \sigma_\varepsilon^2 = \sigma^2$.

2.2.2 Εκτίμηση παραμέτρων απλής γραμμικής παλινδρόμησης

Το πρόβλημα της παλινδρόμησης είναι η εκτίμηση των παραμέτρων α και β καθώς και της διασποράς σ^2 των σφαλμάτων από το δείγμα των ζευγαρωτών παρατηρήσεων $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. Ο σταθερός όρος α είναι η τιμή του y για $x = 0$ και λέγεται **διαφορά ύψους** (*intercept*). Ο συντελεστής του x, β, είναι η **κλίση** (*slope*) της ευθείας ή αλλιώς ο **συντελεστής παλινδρόμησης** (*regression coefficient*).

Η εκτίμηση των α και β γίνεται συνήθως με τη μέθοδο των **ελαχίστων τετραγώνων** (*method of least squares*) και είναι

$$b \equiv \hat{\beta} = \frac{s_{XY}}{s_X^2} \qquad \text{και} \qquad a \equiv \hat{\alpha} = \bar{y} - b\,\bar{x} , \tag{2.8}$$

όπου s_X^2 είναι η δειγματική διασπορά του X (δες (1.2)) και s_{XY} είναι η εκτίμηση της συνδιασποράς των X και Y (δες (2.3)). Η ευθεία των ελαχίστων τετραγώνων είναι

$$\hat{y} = a + b\,x . \tag{2.9}$$

Για κάθε δοθείσα τιμή x_i, με τη βοήθεια της ευθείας ελαχίστων τετραγώνων, εκτιμούμε την τιμή $\hat{y}_i$ που γενικά είναι διαφορετική από την πραγματική τιμή y_i. Η κατακόρυφη απόσταση της πραγματικής τιμής y_i από την ευθεία ελαχίστων τετραγώνων, $e_i = y_i - \hat{y}_i$, είναι το σφάλμα ελαχίστων τετραγώνων ή απλά **υπόλοιπο** (*residual*). Το υπόλοιπο e_i είναι η εκτίμηση του σφάλματος παλινδρόμησης ε_i. Άρα η εκτίμηση της διασποράς του σφάλματος σ^2 δίνεται από τη δειγματική διασπορά s^2 των υπολοίπων e_i

$$s^2 \equiv s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 , \tag{2.10}$$

όπου διαιρούμε με *n-2* γιατί από τους βαθμούς ελευθερίας *n* του μεγέθους του δείγματος αφαιρούμε δύο για τις δύο παραμέτρους που έχουν ήδη εκτιμηθεί.

Διαστήματα εμπιστοσύνης για τις παραμέτρους α και β, δίνονται ως

$$a \pm t_{n-2,1-\alpha/2} \times s_a \qquad \text{και} \qquad b \pm t_{n-2,1-\alpha/2} \times s_b ,$$

ακολουθώντας το γενικό τύπο, όπως και για τη μέση τιμή (δες (1.3)), όπου οι εκτιμήσεις των τυπικών σφαλμάτων των a και b, s_a και s_b, δίνονται από σχετικούς

τύπους. Με τον ίδιο τρόπο μπορεί κάποιος να κάνει έλεγχο υπόθεσης για τα α και β. Συνήθως μας ενδιαφέρει να ελέγξουμε την υπόθεση H_0: $\beta = 0$, γιατί αν η κλίση β της ευθείας παλινδρόμησης βρεθεί να είναι στατιστικά ασήμαντη, τότε οδηγούμαστε στο συμπέρασμα ότι η τ.μ. *Y* δεν εξαρτάται από τη μεταβλητή *X*.

2.2.3 Εκτίμηση και πρόβλεψη της εξαρτημένης μεταβλητής

Το μοντέλο της ευθείας ελαχίστων τετραγώνων της (2.9) δίνει τη σημειακή εκτίμηση $\hat{y}_0$ της μέσης τιμής της *Y* για κάθε τιμή x_0 της *X*. Το αντίστοιχο $(1-\alpha)\%$ διάστημα εμπιστοσύνης είναι

$$(a+bx_0) \pm t_{n-2,1-\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_X^2}} \,. \tag{2.11}$$

Το παραπάνω διάστημα αναφέρεται στη μέση ή αναμενόμενη τιμή της *Y* για μια δεδομένη τιμή x_0 της *X*. Για να ορίσουμε τα **όρια της πρόβλεψης** (limits of prediction) της *Y* για μια τιμή x_0, υπολογίζουμε ένα λίγο ευρύτερο διάστημα με κέντρο επίσης το $\hat{y}_0$, που λέγεται **διάστημα πρόβλεψης** *(prediction interval)* και είναι

$$(a+bx_0) \pm t_{n-2,1-\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_X^2}} \,. \tag{2.12}$$

2.2.4 Επάρκεια του μοντέλου

Ο σχεδιασμός της ευθείας ελαχίστων τετραγώνων $\hat{y} = a + b\,x$ στο διάγραμμα διασποράς δίνει μια πρώτη εικόνα της καταλληλότητας του μοντέλου, δηλαδή πόσο καλά προσαρμόζεται στα σημεία (x_i, y_i). Επίσης φανερώνει, αν υπάρχουν, συστηματικές (όχι τυχαίες) αποκλίσεις των σημείων από την προσαρμοσμένη ευθεία. Αν για παράδειγμα, το γραμμικό μοντέλο δεν είναι σωστό (π.χ. αν μια παραβολική καμπύλη ταιριάζει περισσότερο στα δεδομένα) περιμένουμε περισσότερα σημεία να είναι κάτω από την ευθεία για κάποιες τιμές του *X* και πάνω από την ευθεία για κάποιες άλλες τιμές του *X*. Άλλα χαρακτηριστικά που θα πρέπει ακόμα να ελέγξουμε είναι αν η κατανομή των σημείων γύρω από την ευθεία είναι κανονική και η διασπορά τους για τις διαφορετικές τιμές του *X* σταθερές.

Τα παραπάνω ελέγχονται καλύτερα, αλλά πάντα ποιοτικά, αν σχεδιάσουμε τα υπόλοιπα e_i προς τις προσαρμοσμένες τιμές $\hat{y}_i$. Ένα τέτοιο διάγραμμα θα πρέπει να μη δίνει κανενός είδους σχηματισμό των σημείων για να είναι το μοντέλο επαρκές. Τα σημεία απλά θα πρέπει να σκορπίζονται το ίδιο πάνω και κάτω από την οριζόντια γραμμή στο ύψος 0 και αυτό να συμβαίνει κατά μήκος όλης της γραμμής (όπου υπάρχουν σημεία).

2.3 Πολλαπλή Γραμμική Παλινδρόμηση

Στην **πολλαπλή γραμμική παλινδρόμηση** *(multiple linear regression)* θεωρούμε ότι η τ.μ. *Y* εξαρτάται γραμμικά από κάποιες μεταβλητές $X_1, X_2, \ldots, X_k$. Το y_i για κάποια τιμή x_{1i} της X_1, x_{2i} της $X_2, \ldots,$ x_{ki} της X_k, δίνεται ως

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i . \quad (2.13)$$

Οι υποθέσεις της πολλαπλής γραμμικής παλινδρόμησης είναι όπως και για την απλή γραμμική παλινδρόμηση, δηλαδή υποθέτουμε πως τα σφάλματα ε_i της παλινδρόμησης (όπως και η τ.μ. *Y* για κάθε τιμή των $X_1, X_2, \ldots, X_k$) ακολουθούν κανονική κατανομή με σταθερή διασπορά. Γενικά το πρόβλημα και η εκτίμηση της πολλαπλής γραμμικής παλινδρόμησης δε διαφέρει ουσιαστικά από αυτό της απλής γραμμικής παλινδρόμησης. Ένα καινούριο στοιχείο στην πολλαπλή γραμμική παλινδρόμηση είναι ότι, πριν προχωρήσουμε στην εκτίμηση των παραμέτρων, πρέπει να ελέγξουμε αν πράγματι πρέπει να συμπεριλάβουμε όλες τις ανεξάρτητες μεταβλητές στο μοντέλο.

2.3.1 Επιλογή των ανεξάρτητων μεταβλητών

Αν έχουμε στη διάθεση μας ένα μεγάλο αριθμό ανεξάρτητων μεταβλητών είναι πιθανόν κάποιες από αυτές να είναι περιττές, δηλαδή να μην έχουν να προσφέρουν επιπλέον πληροφορία για τη μεταβλητότητα της *Y*, όταν ήδη έχουμε θεωρήσει την εξάρτηση της *Y* από κάποιες άλλες μεταβλητές. Υπάρχουν διάφορες μέθοδοι για την επιλογή των «σημαντικότερων» μεταβλητών και η πιο διαδεδομένη είναι η **τεχνική της βηματικής παλινδρόμησης** *(stepwise regression technique)*.

Η διαδικασία αυτής της τεχνικής αρχίζει με το απλό σταθερό μοντέλο $y_i = \beta_0 + \varepsilon_i$. Σε κάθε βήμα προστίθεται μια ανεξάρτητη μεταβλητή στο μοντέλο μόνο αν αυτή η μεταβλητή προσφέρει σημαντική πληροφορία για τη *Y*, επιπρόσθετα στην πληροφορία που παρέχουν οι ανεξάρτητες μεταβλητές που είναι ήδη στο μοντέλο από το προηγούμενο βήμα. Η διαδικασία αυτή επιτυγχάνεται με τους εξής στατιστικούς ελέγχους:

1. Αν κάποια από τις μεταβλητές που δεν ήταν στο μοντέλο πρέπει να προστεθεί σ' αυτές που ήδη έχουν συμπεριληφθεί.
2. Αν κάποια από τις μεταβλητές που ήταν στο μοντέλο του προηγούμενου βήματος πρέπει να παραμείνει στο νέο μοντέλο, στο οποίο έχει συμπεριληφθεί μια νέα μεταβλητή.

Στο τέλος η διαδικασία της βηματικής παλινδρόμησης δίνει ένα σύνολο από ανεξάρτητα χρήσιμες μεταβλητές για την εξήγηση της *Y*.

<u>Παράδειγμα:</u> Έστω ότι θέλουμε να εξηγήσουμε τη μεταβλητότητα του χρόνου κάλυψης διαδρομής συγκεκριμένου μήκους (με αυτοκίνητο) σε ημι-αστική ζώνη κι εκτός από τον αριθμό σηματοδοτημένων διασταυρώσεων, θεωρούμε κι άλλους πιθανούς παράγοντες, όπως η καταλληλότητα του οδοστρώματος (σε κάποια αριθμητική κλίμακα αξιολόγησης), ο αριθμός βιομηχανικών μονάδων κατά μήκος της διαδρομής (εντός κάποιας προκαθορισμένης ακτίνας), το ποσοστό των φορτηγών στο πλήθος των οχημάτων που κινούνται σε αυτήν τη διαδρομή, κτλ. Η βηματική παλινδρόμηση θα ξεχωρίσει ποιοι από αυτούς τους παράγοντες είναι χρήσιμοι και εξηγούν κάποιο σημαντικό μέρος της μεταβλητότητας του χρόνου διαδρομής που δεν εξηγείται από τους υπόλοιπους παράγοντες. Θα περίμενε ίσως κάποιος το ποσοστό των φορτηγών να μη δίνει κάποια επιπλέον πληροφορία όταν ήδη έχει συμπεριληφθεί ως σημαντική ανεξάρτητη μεταβλητή στο μοντέλο ο αριθμός των βιομηχανικών μονάδων (η κίνηση πολλών φορτηγών αναμένεται να είναι προς και από τις βιομηχανικές μονάδες).

2.3.2 Προσαρμογή μοντέλου

Η εκτίμηση των παραμέτρων του μοντέλου πολλαπλής παλινδρόμησης γίνεται με τη μέθοδο ελαχίστων τετραγώνων όπως και για την απλή γραμμική παλινδρόμηση. Αντίστοιχα διαστήματα εμπιστοσύνης και πρόβλεψης για τη *Y* για κάθε σύνολο τιμών των $X_1, X_2, \ldots, X_k$ υπολογίζονται όπως στις (2.11) και (2.12) (με κάποια διαφοροποίηση στον τύπο έτσι ώστε να συμπεριλαμβάνει όλες τις ανεξάρτητες μεταβλητές του μοντέλου). Τέλος η επάρκεια του μοντέλου εξετάζεται επίσης από το διάγραμμα υπολοίπων προς προσαρμοσμένες τιμές (e_i, y_i).

Στο SPSS υπάρχει η δυνατότητα εκτίμησης μοντέλου απλής και πολλαπλής γραμμικής παλινδρόμησης (`Regression > Linear`).

2.4 Χρονικές Σειρές

Χρονική σειρά ή **χρονοσειρά** *(time series)* είναι ένα σύνολο παρατηρήσεων $x_1, x_2, ..., x_n$ από μια διαδικασία με χρονική διάταξη, $X_1, X_2, \ldots$. Το κάθε στοιχείο X_t της διαδικασίας είναι μια τ.μ. και οι χρονικά διαταγμένες τ.μ. X_t (με βάση το δείκτη *t*) συσχετίζονται με τρόπο που ορίζεται από τη διαδικασία.

Ενώ σε προβλήματα παλινδρόμησης εξετάζουμε την εξάρτηση κάποιας τ.μ. *Y* από άλλες μεταβλητές $X_1, X_2, \ldots, X_k$ (ο δείκτης εδώ δε δηλώνει χρονική διάταξη αλλά απλά διαφορετικές μεταβλητές), σε προβλήματα χρονοσειρών μας ενδιαφέρει κυρίως η εξάρτηση που μπορεί να έχει η τ.μ. X_t από τις προηγούμενες τ.μ. $X_{t-1}, X_{t-2}, \ldots$, για κάθε χρονική στιγμή *t*. Το πρόβλημα αυτό αναφέρεται ως **αυτοπαλινδρόμηση** *(autoregression)*. Πριν περάσουμε όμως σε αυτό το πρόβλημα ας δούμε τη γενικότερη μορφή που μπορεί να έχει μια χρονοσειρά.

2.4.1 Απλό περιγραφικό μοντέλο χρονοσειρών

Κάθε χρονοσειρά $x_1, x_2, ..., x_n$ μπορεί να περιγραφεί από το απλό μοντέλο

$$x_t = \mu_t + s_t + y_t, \tag{2.14}$$

όπου:

- μ_t είναι το **στοιχείο (συνιστώσα) τάσης** *(trend component)* και περιγράφει τη μακροπρόθεσμη συμπεριφορά της χρονοσειράς.
- s_t είναι το **στοιχείο περιοδικότητας** ή **εποχικότητας** *(cyclical or seasonal component)* και περιγράφει κανονικές διακυμάνσεις κάποιας περιόδου (που μπορεί να αντιστοιχούν σε ένα χρόνο, μια εποχή του χρόνου, μια βδομάδα κτλ).
- y_t είναι το **μη ομαλό στοιχείο** *(irregular component)* που απομένει όταν αφαιρεθούν από τη χρονοσειρά τα μ_t και s_t. Η χρονοσειρά των y_t μπορεί να αποτελείται από πλήρως τυχαίες διακυμάνσεις και τότε λέγεται **λευκός θόρυβος** *(white noise)*, ή να παρουσιάζει κάποια δομή και να επιδέχεται κάποιο μοντέλο αυτοπαλινδρόμησης.

Μια χρονοσειρά θεωρείται **στάσιμη** *(stationary)* όταν η κοινή κατανομή των $X_t, X_{t+1}, \ldots$ είναι ανεξάρτητη του χρόνου *t*, ή πιο απλά όταν οι στατιστικές ιδιότητες

της χρονοσειράς δεν αλλάζουν με το χρόνο. Κυρίως μας ενδιαφέρει να μην αλλάζει η μέση τιμή και συνδιασπορά που αναφέρεται και ως **ασθενής στασιμότητα** *(weak stationarity)*. Άρα μια χρονοσειρά με τάση ή περιοδικότητα δεν είναι στάσιμη και η απαλοιφή τους αποσκοπεί πρακτικά να την κάνει στάσιμη.

Στοιχείο τάσης

Μια πραγματική χρονοσειρά είναι φυσικό να εμφανίζει κάποια τάση, π.χ. η χρονοσειρά μέτρησης της κυκλοφορίας σε μια οδική αρτηρία μπορεί να παρουσιάζει αυξητική τάση με το χρόνο. Μια τέτοια τάση μπορεί να περιγραφεί από κάποια απλή συνάρτηση του χρόνου $f(t)$, π.χ. ένα πολυώνυμο κάποιας τάξης. Έτσι αν π.χ. η αύξηση της κυκλοφορίας είναι αναλογική ως προς το χρόνο, τότε προσαρμόζουμε στη χρονοσειρά πολυώνυμο πρώτης τάξης $\mu_t = f(t) = a_0 + a_1 t$. Για μια πιο πολύπλοκη τάση θα χρειαστεί να προσαρμόσουμε πιο σύνθετη $f(t)$, π.χ. εκθετική συνάρτηση ή πολυώνυμο μεγάλης τάξης.

Σε πολλές εφαρμογές στην ανάλυση χρονοσειράς το ενδιαφέρον δεν είναι στην μακροπρόθεσμη τάση, που μπορεί να οφείλεται σε εξωγενείς παράγοντες (π.χ. αύξηση της κυκλοφορίας λόγω αύξησης της αγοράς αυτοκινήτων) αλλά σε αλλαγές που γίνονται σε μικρότερη χρονική κλίμακα. Σε τέτοιες περιπτώσεις θέλουμε να απαλείψουμε την τάση. Αν μπορούμε να προσαρμόσουμε ικανοποιητικά μια συνάρτηση $f(t)$ την αφαιρούμε από τη χρονοσειρά και συνεχίζουμε με τη χρονοσειρά των υπολοίπων. Γενικά ένας εύκολος τρόπος να απαλείψουμε την τάση είναι να φτιάξουμε μια άλλη σειρά αποτελούμενη από τις πρώτες διαφορές

$$y_t = \nabla x_t = x_t - x_{t-1}. \qquad (2.15)$$

Στοιχείο περιοδικότητας

Μια χρονοσειρά μπορεί να παρουσιάζει κάποια περιοδικότητα, που είτε οφείλεται σε εξωγενείς παράγοντες που δε μας ενδιαφέρει να μελετήσουμε σε βάθος και γι αυτό θέλουμε να την απαλείψουμε, ή είναι σημαντική για τη μελέτη μας και θέλουμε να τη προσδιορίσουμε. Για παράδειγμα, η κυκλοφορία σε μια οδική αρτηρία παρουσιάζει συστηματικές μεταβολές στις διάφορες ώρες της ημέρας. Αν θέλουμε να μελετήσουμε τη μεταβολή της κυκλοφορίας στις διάφορες περιόδους της ημέρας είναι σημαντικό να προσδιορίσουμε αυτήν τη μεταβολή ως μια συνάρτηση του χρόνου για μια ημέρα. Αν θέλουμε να δούμε αν η κυκλοφορία έχει κάποια δυναμική, δηλαδή αν η κυκλοφορία σε κάποια ώρα της ημέρας εξαρτάται από την κυκλοφορία στις προηγούμενες ώρες, τότε για να μελετήσουμε τέτοιες σχέσεις, πρέπει να απαλείψουμε την περιοδικότητα, δηλαδή την κίνηση που αναλογεί λόγω της συγκεκριμένης ώρας της ημέρας. Και στις δύο περιπτώσεις πρέπει πρώτα να εκτιμήσουμε με κάποιο τρόπο την περιοδικότητα.

Αν υποθέσουμε πως έχουμε μια χρονοσειρά $x_1, x_2, ..., x_n$ χωρίς τάσεις (ή αφού απαλείψαμε ήδη την τάση) αλλά με περιοδικότητα με περίοδο *d*, τότε για κάθε $k = 1, \ldots, d$ υπολογίζουμε το μέσο όρο των παρατηρήσεων x_{k+jd}, όπου $1 \le k + jd \le n$. Αυτοί οι μέσοι όροι αποτελούν την εκτίμηση του στοιχείου περιοδικότητας $\hat{s}_k$, $k = 1, \ldots, d$.

2.4.2 Αυτοσυσχέτιση

Είδαμε πως η γραμμική σχέση δύο τ.μ. *X* και *Y* μετριέται με τη συνδιασπορά και τον συντελεστή συσχέτισης. Στις στάσιμες χρονοσειρές θεωρούμε ως τ.μ. τις X_t και $X_{t-\tau}$, για κάθε χρονική υστέρηση *τ*. Η συνδιασπορά ορίζεται για κάθε υστέρηση *τ* ως $\gamma_X(\tau) = \mathrm{Cov}\left[X_t, X_{t-\tau}\right]$ και λέγεται **συνάρτηση αυτοδιασποράς** *(autocovariance function)*. Ο συντελεστής συσχέτισης για κάθε υστέρηση *τ* είναι

$$\rho_X(\tau) = \frac{\gamma_X(\tau)}{\sigma_X^2} \tag{2.16}$$

και λέγεται **συνάρτηση αυτοσυσχέτισης** *(autocorrelation function)*. Η διασπορά σ_X^2 είναι σταθερή αφού η χρονοσειρά θεωρείται στάσιμη.

Σε αντιστοιχία με τη δειγματική συνδιασπορά, η εκτίμηση της αυτοδιασποράς είναι

$$\hat{\gamma}_X(\tau) = \frac{1}{n-\tau}\sum_{t=\tau+1}^{n}\left(x_t - \bar{x}\right)\left(x_{t-\tau} - \bar{x}\right) = \frac{1}{n-\tau}\left(\sum_{t=\tau+1}^{n} x_t x_{t-\tau} - n\,\bar{x}^2\right) \tag{2.17}$$

και της αυτοσυσχέτισης

$$r_X(\tau) = \hat{\rho}_X(\tau) = \frac{\hat{\gamma}_X(\tau)}{s_X^2}. \tag{2.18}$$

Σημειώνεται ότι $-1 \le r_X(\tau) \le 1$ και $r_X(-\tau) = r_X(\tau)$. Η συνάρτηση αυτοσυσχέτισης ορίζει τη γραμμική συσχέτιση μεταξύ των στοιχείων της χρονικής σειράς. Έτσι αν η χρονοσειρά είναι λευκός θόρυβος, η αυτοσυσχέτιση μηδενίζεται για κάθε υστέρηση $\tau \neq 0$. Πρακτικά αυτό σημαίνει πως η εκτίμηση της αυτοσυσχέτισης $r_X(\tau)$ θα πρέπει να κυμαίνεται στα όρια $\pm z_{1-\alpha/2}\frac{1}{\sqrt{n}}$ για κάποια στάθμη σημαντικότητας *α*. Η αυτοσυσχέτιση χρησιμοποιείται ως στατιστική ελέγχου στον **έλεγχο ανεξαρτησίας** *(test of independence)*, δηλαδή για τη μηδενική υπόθεση ότι η χρονοσειρά είναι λευκός θόρυβος.

2.4.3 Μοντέλα αυτοπαλινδρόμησης

Στη συνέχεια υποθέτουμε ότι η χρονοσειρά δεν έχει τάσεις ή περιοδικότητα, ή αν είχε έχουν απαλειφθεί. Υποθέτουμε επίσης ότι η μέση τιμή για τη χρονοσειρά είναι 0 (αλλιώς στους παρακάτω τύπους, αντικαθιστούμε τη τ.μ. X_t με $X_t - \mu$ και την παρατήρηση x_t με $x_t - \bar{x}$, όπου μ η μέση τιμή της διαδικασίας και $\bar{x}$ η μέση τιμή της παρατηρούμενης χρονοσειράς). Θέλουμε να μελετήσουμε την εξάρτηση της X_t από τις προηγούμενες τ.μ. $X_{t-1}, X_{t-2}, \ldots$, για κάθε χρονική στιγμή *t*.

<u>Το γενικό μοντέλο</u>

Το γενικό μοντέλο αυτοπαλινδρόμησης έχει τη μορφή

$$X_t = \phi_1 X_{t-1} + \cdots \phi_p X_{t-p} + \theta_0 \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \tag{2.19}$$

όπου:

- Οι πρώτοι p όροι του μοντέλου αντιστοιχούν στο **αυτοπαλινδρομούμενο μέρος** *(autoregressive term)* και δίνουν τη γραμμική εξάρτηση της X_t στις $X_{t-1}, X_{t-2}, \ldots, X_{t-p}$.

- Ο όρος $\theta_0 \varepsilon_t$ δηλώνει το σφάλμα ή θόρυβο στη χρονική στιγμή *t*. Συνήθως θεωρούμε $\theta_0 = 1$.

- Οι τελευταίοι q όροι του μοντέλου αντιστοιχούν στο θόρυβο στις προηγούμενες q χρονικές στιγμές και αποτελούν το **μέρος του κινούμενου μέσου** *(moving average term)*.

Το μοντέλο αυτό λέγεται **αυτοπαλινδρομούμενο μοντέλο κινούμενου μέσου** *(autoregressive moving average model)* και συμβολίζεται **ARMA(*p*,*q*)**. Υποθέτουμε πως τα ε_t είναι ανεξάρτητα μεταξύ τους και έχουν πανομοιότυπη κατανομή *(identically and independently distributed, **iid**)*. Επίσης κάθε ε_t είναι ανεξάρτητο των X_s για χρόνο $s \le t$ κι έχει διασπορά σ_ε^2. Συνήθως υποθέτουμε επίσης ότι τα ε_t ακολουθούν κανονική κατανομή για να μπορούμε να υπολογίζουμε παραμετρικά διαστήματα εμπιστοσύνης και πρόβλεψης (αλλά για σημειακές εκτιμήσεις δεν είναι απαραίτητο).

Το αυτοπαλινδρομούμενο μοντέλο

Πιο γνωστό είναι το **αυτοπαλινδρομούμενο μοντέλο** *(autoregressive model)* **AR(*p*)**

$$X_t = \phi_1 X_{t-1} + \cdots \phi_p X_{t-p} + \varepsilon_t . \tag{2.20}$$

Το AR(p) ορίζεται όπως το μοντέλο πολλαπλής παλινδρόμησης (δες (2.13)), με τη διαφορά ότι οι μεταβλητές $X_{t-1}, X_{t-2}, \ldots, X_{t-p}$ είναι από την ίδια διαδικασία, όπως η εξαρτημένη μεταβλητή X_t. Άρα για το μοντέλο AR(p) δε μπορούμε να θεωρήσουμε τις $X_{t-1}, X_{t-2}, \ldots, X_{t-p}$ ως ανεξάρτητες μεταβλητές. Παρ' όλα αυτά χρησιμοποιείται κι εδώ η μέθοδος των ελαχίστων τετραγώνων για την εκτίμηση των παραμέτρων $\phi_1, \ldots, \phi_p, \sigma_\varepsilon^2$. Άλλες μέθοδοι χρησιμοποιούν τη σχέση που υπάρχει μεταξύ της συνάρτησης αυτοσυσχέτισης $\rho_X(\tau)$ για $\tau = 0, 1, \ldots, p$ (όπου $\rho_X(0) = \sigma_X^2$) και των παραμέτρων $\phi_1, \ldots, \phi_p, \sigma_\varepsilon^2$, που αναφέρεται και ως **σύστημα κανονικών ή Yule-Walker εξισώσεων** *(normal or Yule-Walker equations)*:

$$\begin{bmatrix} 1 & \rho_X(1) & \cdots & \rho_X(p-1) \\ \rho_X(1) & 1 & \cdots & \rho_X(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \rho_X(p-1) & \rho_X(p-2) & \cdots & 1 \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{bmatrix} = \begin{bmatrix} \rho_X(1) \\ \rho_X(2) \\ \vdots \\ \rho_X(p) \end{bmatrix} \tag{2.21}$$

και

$$\sigma_X^2 = \sum_{j=1}^{p} \phi_j \rho_X(j) + \sigma_\varepsilon^2 . \tag{2.22}$$

Για την εκτίμηση των $\phi_1,\ldots,\phi_p,\sigma_\varepsilon^2$, αντικαθιστούμε στις εξισώσεις (2.21) και (2.22) τις αυτοσυσχετίσεις $\rho_X(\tau)$ για $\tau=0,1,\ldots,p$ με τις εκτιμήσεις $r_X(\tau)$ από τη χρονοσειρά.

Βαθμός μοντέλου

Όπως στην πολλαπλή γραμμική παλινδρόμηση επιλέγουμε με κάποιο τρόπο τις σημαντικές ανεξάρτητες μεταβλητές έτσι και στην αυτοπαλινδρόμηση υπάρχουν κριτήρια **επιλογής του βαθμού** του μοντέλου *(order selection)*, δηλαδή του αριθμού των μεταβλητών $X_{t-1},X_{t-2},\ldots,X_{t-p}$ που θα συμπεριληφθούν στο μοντέλο. Τέτοια κριτήρια είναι το **κριτήριο πληροφορίας του Akaike** (Akaike information criterion, AIC) και το **τελικό σφάλμα πρόβλεψης** *(Final Prediction Error, FPE)*. Επίσης ένα χρήσιμο εργαλείο για την επιλογή του βαθμού του μοντέλου είναι η **συνάρτηση μερικής αυτοσυσχέτισης** *(partial autocorrelation function)*.

Επάρκεια μοντέλου

Για να είναι κατάλληλο ένα μοντέλο AR(p) ή ARMA(p,q) θα πρέπει τα υπόλοιπα e_t να είναι λευκός θόρυβος. Αυτό μπορεί να ελεγχθεί με κατάλληλο έλεγχο ανεξαρτησίας στη χρονοσειρά των υπολοίπων.

2.4.4 Αυτοπαλινδρομούμενο μοντέλο για μη-στάσιμη χρονοσειρά

Για τα μοντέλα AR(p) και ARMA(p,q) θεωρήσαμε πως η χρονοσειρά $x_1,x_2,...,x_n$ είναι στάσιμη. Αν δεν είναι στάσιμη μπορούμε να απαλείψουμε πρώτα την τάση μ_t και την περιοδικότητα s_t και μετά να εφαρμόσουμε το μοντέλο AR(p) ή ARMA(p,q) στη σειρά των υπολοίπων $y_1,y_2,...,y_n$. Αυτό ακριβώς κάνουν τα **ολοκληρωμένα αυτοπαλινδρομούμενα μοντέλα** *(autoregressive integrated moving average models)* **ARIMA(*p,d,q*)**, που εκτιμούνται απευθείας από την αρχική χρονοσειρά. Ένα μοντέλο ARIMA(p,d,q) περιγράφεται ως εξής: εφαρμόζουμε d φορές τον τελεστή πρώτων διαφορών $\nabla x_t = x_t - x_{t-1}$ στην αρχική χρονοσειρά και το μοντέλο ARMA(p,q) στη χρονοσειρά $y_t = \nabla^d x_t$ που προκύπτει από τις d πρώτες διαφορές. Ο βαθμός ολοκλήρωσης d είναι κατά κανόνα μικρός και συνήθως είναι 1. Το μοντέλο ARIMA(p,d,q) μπορεί να επεκταθεί ώστε να συμπεριλάβει και την περιοδικότητα ή εποχικότητα που μπορεί να έχει η χρονοσειρά.

2.4.5 Πρόβλεψη χρονοσειρών

Στην πρόβλεψη χρονοσειρών ξεχωρίζουμε δύο περιπτώσεις: όταν θέλουμε να προβλέψουμε την τάση ή την περιοδικότητα και όταν θέλουμε να προβλέψουμε με κάποιο μοντέλο αυτοπαλινδρόμησης.

Πρόβλεψη τάσης και περιοδικότητας

Αν το χαρακτηριστικό της χρονοσειράς που μας ενδιαφέρει να προβλέψουμε είναι η τάση ή η περιοδικότητα, η πρόβλεψη γίνεται με την εκτίμηση της τάσης ή της περιοδικότητας από μια συνάρτηση του χρόνου, $f(t)$. Τότε η **πρόβλεψη για *T* χρονικά βήματα μπροστά** (T time steps ahead prediction) είναι

$$\hat{x}_{n+T} \equiv x_n(T) = f(n+T). \qquad (2.23)$$

Ο συμβολισμός $x_n(T)$ αναφέρεται στην πρόβλεψη T χρονικών στιγμών μπροστά όταν γνωρίζουμε τη χρονοσειρά ως και τη χρονική στιγμή n.

Πρόβλεψη με μοντέλα αυτοπαλινδρόμησης

Ας θεωρήσουμε πρώτα ότι η χρονοσειρά $x_1, x_2, ..., x_n$ είναι στάσιμη και έχουμε εκτιμήσει τις παραμέτρους ενός μοντέλου AR(p). Η πρόβλεψη για ένα χρονικό βήμα μπροστά ($T=1$) με το μοντέλο AR(p) είναι

$$x_n(1) = \phi_1 x_n + \cdots + \phi_p x_{n-p+1}. \tag{2.24}$$

Για πρόβλεψη σε βήμα $T>1$ αντικαθιστούμε κάθε φορά τις άγνωστες τιμές των $x_{n+1}, x_{n+2}, \ldots$, που περιέχονται στον αντίστοιχο τύπο της πρόβλεψης ενός βήματος (2.24) για χρόνο $n+T-1$ αντί n, με τις προβλέψεις $x_n(1), x_n(2), \ldots$ από τις προηγούμενες προβλέψεις.

Με τον ίδιο τρόπο επιτυγχάνεται η πρόβλεψη με ARMA(p,q) μοντέλο, όπου κάθε φορά απαλείφουμε από το μοντέλο το θόρυβο που είναι ανεξάρτητος από τις παρατηρήσεις (για χρόνο $s>n$). Το σφάλμα πρόβλεψης $e_n(T)$ είναι η απόκλιση της πρόβλεψης $x_n(T)$ από την πραγματική τιμή x_{n+T}, $e_n(T) = x_{n+T} - x_n(T)$. Υπολογίζοντας τη διασπορά των σφαλμάτων $\mathrm{Var}[e_n(T)]$ μπορεί κάποιος να υπολογίσει τα $(1-\alpha)\%$ **όρια πρόβλεψης** (*prediction bounds, tolerance intervals*)

$$x_n(T) \pm c_{1-\alpha/2}\sqrt{\mathrm{Var}[e_n(T)]}, \tag{2.25}$$

για κάποια κρίσιμη τιμή $c_{1-\alpha/2}$ (είναι $z_{1-\alpha/2}$ αν τα σφάλματα ακολουθούν κανονική κατανομή).

Αντίστοιχα γίνεται η πρόβλεψη με μοντέλο ARIMA(p,d,q) όταν η χρονοσειρά δεν είναι στάσιμη. Για $d=1$, εφαρμόζοντας το μοντέλο ARMA(p,q) στην $\{y_2, y_3, \ldots, y_n\}$ βρίσκουμε την πρόβλεψη για ένα χρονικό βήμα, $y_n(1)$. Η πρόβλεψη για την αρχική χρονική σειρά είναι

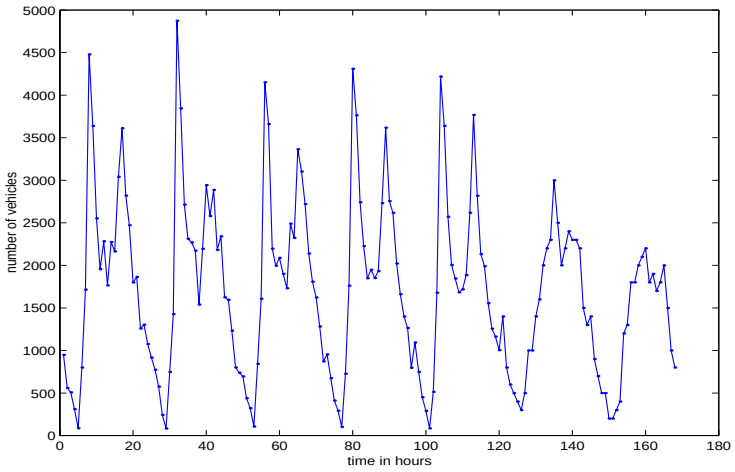
$$x_n(1) = x_n + y_n(1). \tag{2.26}$$

Επαναληπτικά προκύπτει ότι για T χρονικά βήματα η πρόβλεψη είναι

$$x_n(T) = x_n(T-1) + y_n(T), \tag{2.27}$$

όπου $y_n(T)$ είναι η πρόβλεψη του y_{n+T} με το μοντέλο ARMA(p,q) και το $x_n(T-1)$ είναι γνωστό από την πρόβλεψη του x_{n+T-1} στο προηγούμενο βήμα.

Παράδειγμα: Θέλουμε να μελετήσουμε τη μεταβλητότητα του πλήθους των οχημάτων που διέρχονται από μια κύρια αρτηρία. Για το λόγο αυτό μετρήθηκε το πλήθος των οχημάτων ανά ώρα που περνούσε από αυτήν την αρτηρία για μια εβδομάδα, αρχίζοντας τη μέτρηση από τα μεσάνυχτα της Κυριακής.



Εικόνα 2 Χρονοσειρά του πλήθους οχημάτων ανά ώρα σε μια οδική αρτηρία για μια βδομάδα.

Τα δεδομένα του παραπάνω σχήματος αποτελούν μια χρονοσειρά 7·24 = 168 στοιχείων με τη γραμμή να ενώνει τα σημεία της μέτρησης. Η χρονοσειρά αυτή φαίνεται να έχει περιοδικότητα που αντιστοιχεί στη διάρκεια ενός 24ωρου (φαίνεται ότι η μορφή είναι παρόμοια για τις 5 πρώτες μέρες της εβδομάδας). Αν θέλουμε να μελετήσουμε σχέσεις μεταξύ των στοιχείων της χρονοσειράς (αν η κυκλοφορία για κάποια ώρα συσχετίζεται με την κυκλοφορία την προηγούμενη ώρα, την προ-προηγούμενη κτλ) θα πρέπει να απαλείψουμε αυτήν τη περιοδικότητα. Στη συνέχεια θα πρέπει να εκτιμήσουμε ένα μοντέλο αυτοπαλινδρόμησης για να περιγράψουμε αυτές τις συσχετίσεις, αν υπάρχουν. Το μοντέλο αυτό, αν είναι επαρκές, μπορούμε να το χρησιμοποιήσουμε για να κάνουμε προβλέψεις (για τις επόμενες ώρες).