Κεφάλαιο 7

Ανάλυση Διακύμανσης με ένα Παράγοντα (One Way ANOVA)

7.1 Γενικότητες

Η ANOVA περιλαμβάνει μία ομάδα στατιστικών μεθόδων κατάλληλων για την ανάλυση δεδομένων που προκύπτουν από πειραματικούς σχεδιασμούς. Η ανάπτυξη της μεθοδολογίας οφείλεται στον θεμελιωτή της σύγχρονης στατιστικής επιστήμης, Άγγλο στατιστικό Sir Ronald Aylmer Fisher (1890-1962). Η ανάλυση διακύμανσης προτάθηκε από τον ίδιο το 1918. Ευρέως έγινε γνωστή μετά το 1925 όταν εκδόθηκε το βιβλίο του R. A. Fisher, Statistical Methods for Research Workers, στο οποίο είχε συμπεριλάβει και την ανάλυση διακύμανσης. Η ανάλυση διακύμανσης «γεννήθηκε» κατά την ενασχόληση του Fisher με δύσκολα προβλήματα στατιστικής συμπερασματολογίας που εμφανίζονται στο γεωργικό πειραματισμό (πολλές πηγές μεταβλητότητας και συχνά εμφανιζόμενες ετερογένειες, και μάλιστα προς διάφορες κατευθύνσεις του πειραματικού αγρού π.χ. ως προς τη γονιμότητα, την κλίση και την υγρασία των εδαφών, τις προηγούμενες καλλιέργειες, κτλ.). Η προσέγγιση της λύσης τέτοιου είδους προβλημάτων που πρότεινε ο Fisher βασίζεται στην τυχαιοποίηση και στην επανάληψη και ως μαθηματικό εργαλείο για την υποστήριξη αυτής της προσέγγισης πρότεινε την ανάλυση διακύμανσης.

Γι’ αυτό στην ανάλυση διασποράς έχει επικρατήσει να χρησιμοποιείται ορολογία που χρησιμοποιείται στο γεωργικό πειραματισμό και γενικότερα στον πειραματισμό, παρότι δεν εφαρμόζεται μόνο στην ανάλυση πειραματικών δεδομένων.

7.2 Εισαγωγή

Η ανάλυση της διακύμανσης (**An**alysis **O**f **Va**riance - ANOVA) είναι μία στατιστική μέθοδος με την οποία η μεταβλητότητα που υπάρχει σ' ένα σύνολο δεδομένων διασπάται στις επιμέρους συνιστώσες της με στόχο την κατανόηση της σημαντικότητας των διαφορετικών πηγών προέλευσής της.

Ελέγχει τη διαφορά στις μέσες τιμές προκειμένου να αποφασιστεί, αν η διακύμανση μεταξύ δύο ή περισσότερων ομάδων είναι μεγαλύτερη από τη διακύμανση εντός των ομάδων.

Υπάρχουν δύο είδη αναλύσεων:

- μονοδιάστατος σχεδιασμός (one-way design)
- παραγοντικός σχεδιασμός (factorial design), συνήθως δισδιάστατος (two-way design).

Θα ξεκινήσουμε με το επόμενο παράδειγμα.

Παράδειγμα 7.1 (Constantin Yiannoutsos - Principles of Biostatistics)

Ασθενείς από τρία κέντρα, το Johns Hopkins, το Rancho Los Amigos και το St.Louis έλαβαν μέρος σε μια κλινική δοκιμή. Ως μέρος των βασικών αξιολογήσεων, εκτιμήθηκε η πνευμονική απόδοση των ασθενών. Μία καλή δοκιμασία αυτού αποτελεί το «Ποσό Δυναμικής Εκπνευστικής Ροής σε διάρκεια ενός δευτερολέπτου(FEV1)». Τα δεδομένα παρουσιάζονται στον Πίνακα 7.1.

Ήταν σημαντικό για τους ερευνητές να εξακριβωθεί αν οι ασθενείς από τα τρία κέντρα είχαν κατά μέσο όρο παρόμοιες πνευμονικές λειτουργίες πριν την διεξαγωγή της έρευνας.

Πίνακας 7.1 *Δεδομένα Παραδείγματος 7.1*

.Λίστα	**FEV1**	**Κέντρο**
1.	3.23	Johns Hopkins
2.	3.47	Johns Hopkins
3.	1.86	Johns Hopkins
4.	2.47	Johns Hopkins
5.	3.01	Johns Hopkins
…		Johns Hopkins
10.	3.36	Johns Hopkins

11.	2.61	Johns Hopkins
12.	2.91	Johns Hopkins
…		Johns Hopkins
57.	2.85	St.Louis
58.	2.43	St.Louis
59.	3.2	St.Louis
60.	3.53	St.Louis

Johns Hopkins : (κέντρο=1)
Rancho Los Amigos : (κέντρο=2)
St.Louis : (κέντρο=3)

Συνοπτικά Στατιστικά Στοιχεία Δεδομένων

Κέντρο:Johns Hopkins					
Μεταβλητή	N	Μέση Τιμή	Τυπική Απόκλιση	Ελάχιστο	Μέγιστο
FEV1	21	2.62619	0.4961701	1.69	3.47

Κέντρο:Rancho Los Amigos					
Μεταβλητή	N	Μέση Τιμή	Τυπική Απόκλιση	Ελάχιστο	Μέγιστο
FEV1	16	3.0325	0.5232399	1.71	3.86

Κέντρο:St.Louis					
Μεταβλητή	N	Μέση Τιμή	Τυπική Απόκλιση	Ελάχιστο	Μέγιστο
FEV1	23	2.878696	0.4977157	1.98	4.06

Με σκοπό να αντιμετωπιστούν οι προβληματισμοί των ερευνητών πρέπει να συγκρίνουμε το μέσο όρο πνευμονικής λειτουργίας των ασθενών στις τρεις περιοχές. Απο τη στιγμή που οι μέσοι όροι και οι τυπικές αποκλίσεις των πνευμονικών λειτουργιών από κάθε περιοχή δεν είναι γνωστές, πρέπει να τις εκτιμήσουμε από τα δεδομένα.

Γενικά, όταν έχουμε να κάνουμε με k ομάδες έχουμε τα ακόλουθα:

	Ομάδα 1	**Ομάδα 2**	**...**	**Ομάδα k**
Πληθυσμός				
Μέση Τιμή	μ_1	μ_2	...	μ_k
Τυπική Από-κλιση	σ_1	σ_2	...	σ_k
Δείγμα				
Μέση Τιμή	$\bar{x}_1$	$\bar{x}_2$	...	$\bar{x}_k$
Τυπική Από-κλιση	s_1	s_2	...	s_k
Μέγεθος Δείγματος	n_1	n_2	...	n_k

Πρέπει να χρησιμοποιήσουμε τις πληροφορίες του δείγματος, έτσι ώστε να εξάγουμε συμπεράσματα (τεστ υπόθεσης, διαστήματα εμπιστοσύνης κλπ) για τις παραμέτρους του πληθυσμού.

Ένα στατιστικό τεστ για επίπεδο σημαντικότητας *a* για την αντιμετώπιση τέτοιου είδους ερευνών, δομείται ως εξής:

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$,

H_1 :Τουλάχιστον μία εκ των μέσων τιμών είναι διαφορετική (ένα ζευγάρι διαφέρει).

Ερώτηση: Μπορούμε να χρησιμοποιήσουμε το *t*-τεστ των δυο δειγμάτων για να εκτελέσουμε αυτές τις συγκρίσεις;

Απάντηση:Τα *t*-τεστ των δυο δειγμάτων δεν μπορούν επακριβώς να ικανοποιήσουν την υπόθεση, διότι αυτές οι συγκρίσεις περιλαμβάνουν περισσότερες από δυο ομάδες. Όμως,μπορούμε να χρησιμοποιήσουμε το *t*-τεστ των δυο δειγμάτων για να επιλύσουμε κάθε πιθανή ζευγαρωτή συγκριση μεταξύ των k ομάδων. Σε περιπτωση που k=3, ο αριθμός έστω g των συγκρίσεων ανά 2 που προκύπτει ισούται με 3, (δηλ. Ομαδα1 vs Ομάδα2 , Ομάδα 1 vs Ομάδα3 , και Ομάδα2 vs Ομάδα3).

t-τεστ για ισότητα μέσων τιμών k ομάδων

Ένα τεστ της συνολικής υπόθεσης της ισότητας μεταξυ των k μέσων τιμών, βασισμένο σε t-τεστ δυο δειγμάτων από όλες τις g ζευγαρωτές ομάδες συγκρίσεων, δομείται ως εξης:

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$,

H_1 :Τουλάχιστον μία εκ των μέσων τιμών είναι διαφορετική (ένα ζευγάρι διαφέρει).

Εκτελούμε g το πλήθος *t*-τεστ δυο δειγμάτων (όπου g είναι ο αριθμός όλων των δυνατών ανά 2 συγκρίσεων,δηλαδή συνδυασμός *k* ανα 2) :

$H_{0l} : \mu_i = \mu_j$ όπου i, j είναι δυο εκ των k ομαδων και l=1,…,g,

$H_{1l} : \mu_i \neq \mu_j$.

Το επίπεδο σημαντικότητας είναι *a* .

Η στατιστική συνάρτηση και η απορριπτική περιοχή, δίνονται από τον Πίνακα 6.14 και λαμβάνουν την εξής μορφή:

Κρίσιμη περιοχή	$R = \left\{ \dfrac{\lvert\bar{x}_i - \bar{x}_j\rvert}{s_{ij}\sqrt{\dfrac{1}{n_i}+\dfrac{1}{n_j}}} > t_{n_i+n_j-2,a/2} \right\}$,όπου $s_{ij}^2 = \dfrac{(n_i-1)s_i^2 + (n_j-1)s_j^2}{n_i+n_j-2}$

Εφαρμόζοντας κατά τα γνωστά, τα αντίστοιχα τρία *t*-test στα προκύπτοντα ζεύγη, διαπιστώνουμε ότι η μόνη διαφορά που βρέθηκε να είναι σημαντική σε α=0.05 επίπεδο, ήταν η σύγκριση των Rancho Los Amigos και Johns Hopkins.

Η απορριπτική περιοχή (του συνολικού τεστ):

Απορρίπτουμε την H_0 αν οποιοδήποτε από τα g τεστ απορρίπτει την μηδενική του υπόθεση H_{0l} . Διαφορετικά, δεν απορρίπτουμε την H_0 .

Παρατηρήσεις

Υπάρχουν διάφορα θέματα προς σημείωση:

- Αν ο αριθμός των ομάδων είναι ο ελάχιστος δυνατός (*k*=3), οι μεταξύ τους συγκρίσεις είναι σχετικά λίγες (g=3). Αν όμως για παράδειγμα *k*=10, τότε οι δυνατοί συνδυασμοί 10 ανά 2 είναι 45. Έτσι, το g τείνει να αυξηθεί ραγδαία όταν αυξάνεται το k.
- Αν το επίπεδο σημαντικότητας του συνολικού τεστ είναι α και κάθε ένα από τα g υπο-τεστ είναι επίσης α, τότε το επίπεδο της ση-

μαντικότητας του συνολικού τεστ είναι μεγαλύτερο. Παράδειγμα: Θεωρούμε την περίπτωση όπου α=0.05, τότε (1-α)% είναι 95%, και k=3 (άρα g=3).

Έστω Α το ενδεχόμενο «Το συνολικό τεστ απορρίπτει ορθά την H_0 », και

A_l το ενδεχόμενο «Το l τεστ απορρίπτει ορθά την H_{0l} », τότε

$P(A) = P(A_1 \cap A_2 \cap \ldots \cap A_g) = P(A_1) \cdot P(A_2) \cdot \ldots \cdot P(A_g) = (1-a)^g$,

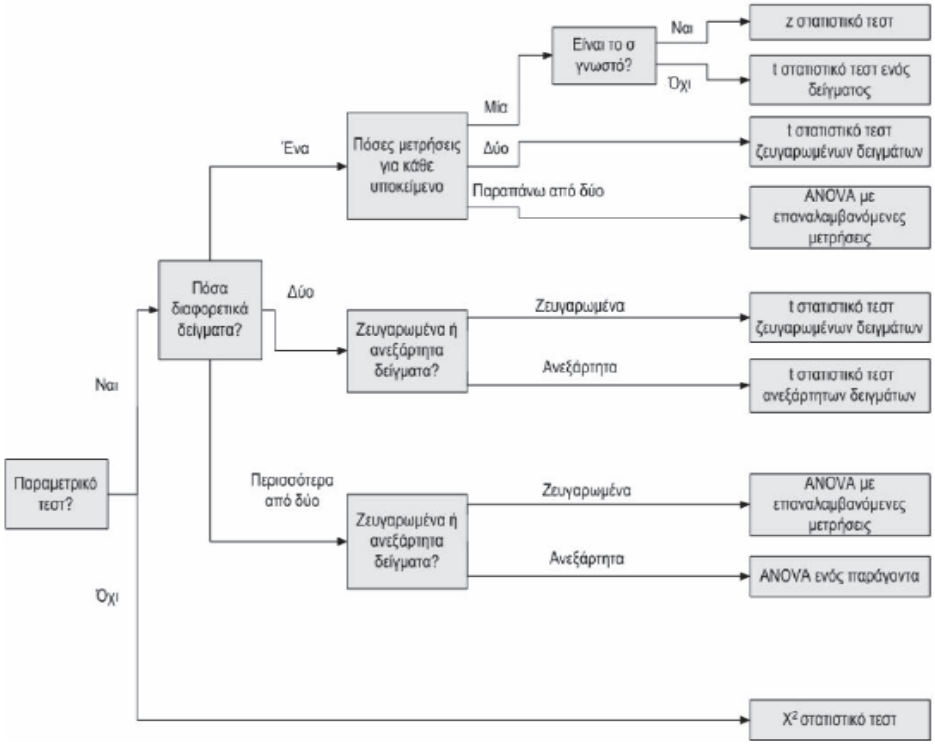
υποθέτοντας ότι τα υπο-τεστ είναι ανεξάρτητα. Αν α=0,05 , $P(A) = (1-a)^g = 0,95^3 = 0.857 < 0.95$. Συνεπώς,η πιθανότητα ενός σφάλματος τυπου Ι είναι α=(1 – 0.857)=0.143 αντί για μόλις 0.05 (με την προϋπόθεση της ανεξαρτησίας). Ακόμη κι αν τα ατομικά ζευγαρωτά τεστ δεν είναι ανεξάρτητα, το επίπεδο σημαντικότητας του συνολικού τεστ μπορεί να είναι πολύ μεγαλύτερο από το αναμενόμενο. Ετσι,το t-τεστ των δυο δειγματων δεν είναι πλήρως ικανοποιητικό.

7.3 Ανάλυση Διασποράς

Μια γενική διαδικασία που αντιμετωπίζει απευθείας την πιο πάνω μηδενική υπόθεση καλείται Ανάλυση Διασποράς. Γενικά, οι πιο συχνά χρησιμοποιούμενες στατιστικές μέθοδοι (π.χ. t-test όπως είδαμε στο προηγούμενο παράδειγμα) χρησιμοποιούν ένα ή δύο δείγματα για να εξετάσουν υποθέσεις για τιμές της ελεγχόμενης μεταβλητής, αλλά υπάρχουν και περιπτώσεις στις οποίες ο ερευνητής θέλει να εξετάσει παραπάνω δείγματα (και παραπάνω μεταβλητές). Η ανάλυση διακύμανσης εξυπηρετεί αυτόν ακριβώς το σκοπό.

Θα μπορούσαμε να εξετάσουμε την επίδραση όλων αυτών των δυνατών τιμών ανά δύο, χρησιμοποιώντας όσες φορές χρειάζεται μία από τις άλλες στατιστικές μεθόδους. Ωστόσο, όπως εύκολα γίνεται αντιληπτό από το προηγούμενο Παράδειγμα 7.1, η χρήση πολλαπλών στατιστικών τεστ αυξάνει τη συνολική πιθανότητα να οδηγηθούμε σε λανθασμένα συμπεράσματα και φυσικά αυξάνει και το χρόνο της διαδικασίας.

Επιλογή Στατιστικού Ελέγχου



Σχήμα 7.1

ANOVA

Ο έλεγχος αυτός ονομάζεται Ανάλυση Διακύμανσης (Analysis of Variance, ANOVA). Το όνομά του το χρωστάει στο γεγονός ότι η συνολική διακύμανση μπορεί να αναλυθεί σε 2 μέρη, τη διακύμανση μέσα σε κάθε δείγμα (πόσο διαφέρουν δηλαδή οι παρατηρήσεις κάθε ομάδας) και στη διακύμανση ανάμεσα στις ομάδες (πόσο διαφέρουν οι ομάδες μεταξύ τους). Εμπειρικά αν οι μέσες τιμές είναι ίδιες και δεδομένου ότι υποθέτουμε ότι όλες οι ομάδες έχουν ίδια διακύμανση (παρατηρούμε ότι αυτό το πρόβλημα αντιμετωπίσαμε και στον έλεγχο με τους 2 μέσους), θα περίμενε κανείς ότι η διακύμανση ανάμεσα στις ομάδες θα ήταν πολύ μικρή. Αυτό θα αποτελέσει και τη βάση για τον έλεγχο.

Η Ανάλυση Διακύμανσης στηρίζεται σε ένα πλούσιο θεωρητικό υπόβαθρο, και έχει ερευνηθεί με ποικίλους τρόπους. Όταν έχουμε να εξετάσουμε απλά για την ισότητα μέσων τιμών για περισσότερα από 2 δείγματα λέγε-

ται και ανάλυση κατά έναν παράγοντα (one way ANOVA). Στην πραγματικότητα η Ανάλυση Διακύμανσης αποτελεί ένα πολύ δυνατό εργαλείο καθώς μας επιτρέπει να εξετάσουμε πιο πολύπλοκες καταστάσεις, όπως περισσότερους παράγοντες μαζί, αλληλεπιδράσεις παραγόντων κλπ..

Πλεονεκτήματα της ANOVA (συνοπτικά):

- Δεν έχει περιορισμούς στον αριθμό των μέσων τιμών που είναι δυνατόν να συγκριθούν.
- Συντομότερη διαδικασίας ανάλυσης.
- Ακρίβεια της διάγνωσης.
- Μας επιτρέπει να μελετήσουμε ταυτόχρονα την επίδραση δύο ή περισσότερων ανεξάρτητων μεταβλητών. Έτσι, υπολογίζουμε όχι μόνο την επίδραση της κάθε μίας ανεξάρτητης μεταβλητής στην εξαρτημένη αλλά και τις αλληλεπιδραστικές συνέπειες των ανεξάρτητων μεταβλητών στην εξαρτημένη.

Προϋποθέσεις εφαρμογής ανάλυσης διασποράς:

- Κανονική κατανομή πληθυσμού.
- Ίσες διασπορές δειγμάτων.
- Η κλίμακα μέτρησης της εξαρτημένης μεταβλητής να είναι τουλάχιστον ίσων διαστημάτων.
- Οι διακυμάνσεις είναι ομοιογενείς.

7.4 Μονοπαραγοντική Ανάλυση Διακύμανσης (one-way ANOVA)

Παραμετρικό στατιστικό κριτήριο για τη μελέτη της επίδρασης μιας ανεξάρτητης μεταβλητής στην εξαρτημένη. Η λογική είναι παρόμοια με το κριτήριο *t* αλλά επιτρέπει στον ερευνητή να συγκρίνει μέσους όρους από περισσότερα από δύο δείγματα. Χρησιμεύει στην ανάλυση δεδομένων που προέρχονται από σχεδιασμούς τόσο ανεξάρτητων δειγμάτων (δηλαδή μετρήσεις που προέρχονται από ομάδες στις οποίες συμμετέχουν διαφορετικά άτομα) όσο και εξαρτημένων δειγμάτων (δηλαδή μετρήσεις που προέρχονται από τη συμμετοχή του κάθε ατόμου σε όλες τις ερευνητικές συνθήκες).

Η διακύμανση των τιμών μεταξύ των πειραματικών συνθηκών προκύπτει από τρεις πηγές:

α) την επίδραση της ανεξάρτητης μεταβλητής,

β) τις ατομικές διαφορές,

γ) το σφάλμα μέτρησης.

Ελέγχουμε τη διαφορά στις μέσες τιμές για να εξακριβώσουμε εάν η διακύμανση είναι μεγαλύτερη μεταξύ των ομάδων απ' ό,τι εντός των ομάδων (εάν δηλαδή η διακύμανση γύρω από το συνολικό μέσο όρο είναι μεγαλύτερη από τη διακύμανση γύρω από το μέσο όρο κάθε ομάδας). Ο ερευνητής επιλέγει ένα δείγμα για κάθε διαφορετική τιμή της ανεξάρτητης μεταβλητής και ελέγχει υποθέσεις που συγκρίνουν τις μέσες τιμές των δειγμάτων αυτών. Η μηδενική υπόθεση δηλώνει ότι δεν υφίσταται καμία διαφορά ανάμεσα σε όλες τις μέσες τιμές των δειγμάτων. Άρα στην περίπτωση που ο ερευνητής απορρίψει τελικά τη μηδενική υπόθεση, το μόνο που μπορεί να ισχυριστεί είναι ότι τα δείγματα διαφέρουν μεταξύ τους, αλλά δεν είναι σε θέση να γνωρίζει ποια συγκεκριμένα δείγματα διαφέρουν. Αν θέλει να εξακριβώσει τέτοιου είδους πληροφορίες μπορεί να χρησιμοποιήσει μετά την ανάλυση διακύμανσης ειδικά στατιστικά τεστ.

Προϋποθέσεις εφαρμογής one-way ANOVA

- Τα δείγματα είναι αντιπροσωπευτικά και οι τιμές που τα απαρτίζουν οφείλονται σε ανεξάρτητες παρατηρήσεις.
- Η κατανομή των τιμών των δειγμάτων είναι κανονική.
- Οι πληθυσμοί από τους οποίους προέρχονται τα δείγματα είναι κανονικοί με κοινή διασπορά.

Ο έλεγχος που πρέπει να πραγματοποιηθεί είναι ο ακόλουθος:

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$,

H_1 :Τουλάχιστον μία εκ των μέσων τιμών είναι διαφορετική.

Η διαδικασία εφαρμογής ANOVA θα δοθεί και με τη βοήθεια παραδείγματος.

Παράδειγμα 7.2

Η αίθουσα Βιοστατιστικής διαχωρίζεται σε τρείς σειρές: Μπροστινή (Front), Μεσαία (Middle) και Πίσω (Back). Ο καθηγητής τους παρατήρησε ότι η επίδοση των φοιτητών είχε κάποια σχέση με την θέση τους. Θέλησε

να ελεγχθεί αν οι φοιτητές που κάθονταν πιο πίσω είχαν και χειρότερη επίδοση.

- Εκλέχθηκε ένα τυχαίο δείγμα των φοιτητών κάθε σειράς.
- Η επίδοση των φοιτητών στις εξετάσεις καταγράφηκε ως εξής:

o Front: 82, 83, 97, 93, 55, 67, 53

o Middle:83, 78, 68, 61, 77, 54, 69, 51, 63

o Back: 38, 59, 55, 66, 45, 52, 52, 61

Ο επόμενος πίνακας παρουσιάζει συνοπτικά τα περιγραφικά στατιστικά μέτρα των βαθμών κάθε σειράς:

Πίνακας 7.2 *Περιγραφικά στατιστικά μέτρα δεδομένων*

Σειρά	**Front**	**Middle**	**Back**
Μέγεθος δείγματος (Sample size)	7	9	8
Μέσος (Mean)	75.71	67.11	53.50
Τυπική απόκλιση (St. Dev)	17.63	10.95	8.96
Διακύμανση (Variance)	310.90	119.86	80.29

- Η διακύμανση ορίζεται με την βοήθεια του αθροίσματος των τετραγώνων των αποκλίσεων κάθε τιμής από το μέσο.
- Το άθροισμα των τετραγώνων συντομογραφείται με **SS** και συχνά ακολουθείται από μια παρένθεση όπως SS(B) ή SS(W), ώστε να γνωρίζουμε σε ποιο άθροισμα τετραγώνων αναφερόμαστε.

Στη συνέχεια τίθενται τα εξής ερωτήματα:

Είναι όλες οι τιμές ταυτόσημες;

- Όχι, υπάρχει κάποια διασπορά στα δεδομένα που ονομάζεται ολική διασπορά και συμβολίζεται **SS(Total)** για το ολικό άθροισμα των τετραγώνων.
- Άθροισμα Τετραγώνων (Sum of Squares) είναι εναλλακτική ονομασία.

Είναι όλα τα δείγματα ταυτόσημα;

- Όχι, υπάρχει διασπορά μεταξύ των δειγμάτων που ονομάζεται διασπορά μεταξύ των δειγμάτων (between group variation).
- Συμβολίζεται **SS(B).**

Είναι όλες οι τιμές μέσα σ᾽ ένα δείγμα ταυτόσημες;

- Όχι, υπάρχει διασπορά εντός του κάθε δείγματος (within group variation) και ονομάζεται διασπορά μέσα στο δείγμα, υπόλοιπο ή σφάλμα.
- Συμβολίζεται **SS(W).**

Υπάρχουν δύο πηγές διασποράς:

- Η διασπορά μεταξύ των δειγμάτων, **SS(B),** ή διασπορά λόγω του παράγοντα.
- Η διασπορά μέσα στο δείγμα, **SS(W)**, ή η διασπορά που δεν μπορεί να εξηγηθεί από τον παράγοντα και έτσι ονομάζεται και σφάλμα.

Ο βασικός πίνακας one-way ANOVA είναι ως εξής:

Πίνακας 7.3 Βασικός πίνακας one-way ANOVA

Source	SS	df	MS	F	p
Between					
Within					
Total					

Στη συνέχεια, προκειμένου να συμπληρώσουμε τις απαραίτητες ποσότητες στον βασικό πίνακα ANOVA, θα παραθέσουμε τις αντίστοιχες σχέσεις και θα τις εφαρμόσουμε για το Παράδειγμα 7.2.

- **Μεγάλος Μέσος (Grand Mean)**

  Ο μεγάλος μέσος είναι ο μέσος όλων των τιμών όταν αγνοείται ο παράγοντας.

  Αποτελεί σταθμισμένο μέσο των μέσων των δειγμάτων.