Επίκουρος Καθηγητής **Ιωάννης Παραβάντης**
Τμήμα Διεθνών και Ευρωπαϊκών Σπουδών
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
Ιούλιος 2009

Ανάλυση Διασποράς (Analysis of Variance, ANOVA)

Στην **ANOVA** (Analysis of Variance) συγκρίνουμε τους μέσους όρους (**means**) περισσότερων από δυο πληθυσμών (**populations**).

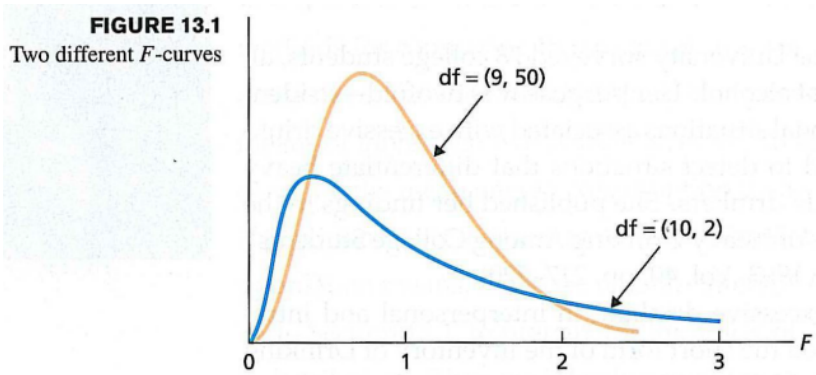
Για **παράδειγμα**, μπορεί να θέλουμε να συγκρίνουμε την μέση ετήσια ενεργειακή κατανάλωση ανά νοικοκυριό (annual mean energy consumption per household) διαφορετικών περιοχών μιας χώρας.

Ο απλούστερος τύπος ANOVA ονομάζεται **one-way ANOVA**.

Προκειμένου να ασχοληθούμε με ANOVA, πρέπει πρώτα να μελετήσουμε την κατανομή F (**F-distribution**), που έχει ονομαστεί έτσι προς τιμήν του **Sir Ronald Fisher** (1890-1962).

Ο Fisher είχε περιγραφεί ως *«slight, bearded, eloquent, reactionary and quirkish; genial to his disciples and hostile to his dissenters»*.

Να δυο καμπύλες που έχουν το σχήμα της κατανομής F:

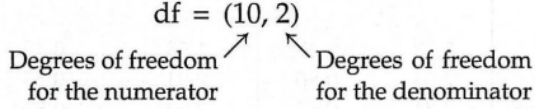


Οι βασικές ιδιότητες (**properties**) των καμπυλών που έχουν το σχήμα της κατανομής F (F curve) είναι:

1. το συνολικό εμβαδόν (**area**) κάτω από την καμπύλη ισούται με ένα
2. η ελάχιστη τιμή (**min**) της καμπύλης είναι μηδέν
3. η μέγιστη τιμή (**max**) της καμπύλης είναι το συν άπειρο (plus infinity) – η καμπύλη τείνει προς τον οριζόντιο άξονα ασυμπτωτικά (asymptotically)
4. η καμπύλη F έχει θετική στρέβλωση (**right skewed**, δηλαδή το «σουβλί» είναι στα δεξιά).

Όπως μπορούμε να δούμε από το προηγούμενο σχήμα, η μορφή της κατανομής F καθορίζεται από δυο αριθμητικές παραμέτρους που ονομάζονται βαθμοί ελευθερίας (**degrees of freedom**):

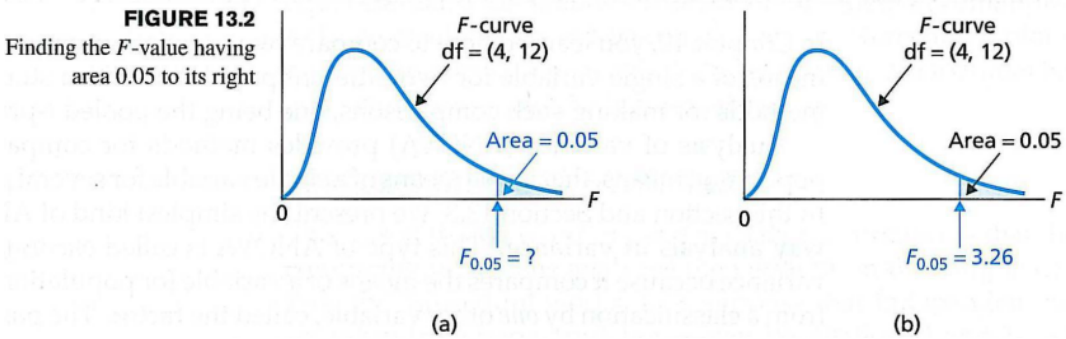
1. ο πρώτος ονομάζεται βαθμοί ελευθερίας του αριθμητή (**numerator**)
2. ο δεύτερος ονομάζεται βαθμοί ελευθερίας του παρονομαστή (**denominator**).



Συμβολίζουμε τις κρίσιμες τιμές (critical values) της κατανομής F με το σύμβολο F_{α}, που συμβολίζει την τιμή F που έχει εμβαδόν α (άλφα) προς τα δεξιά (δηλαδή από πάνω).

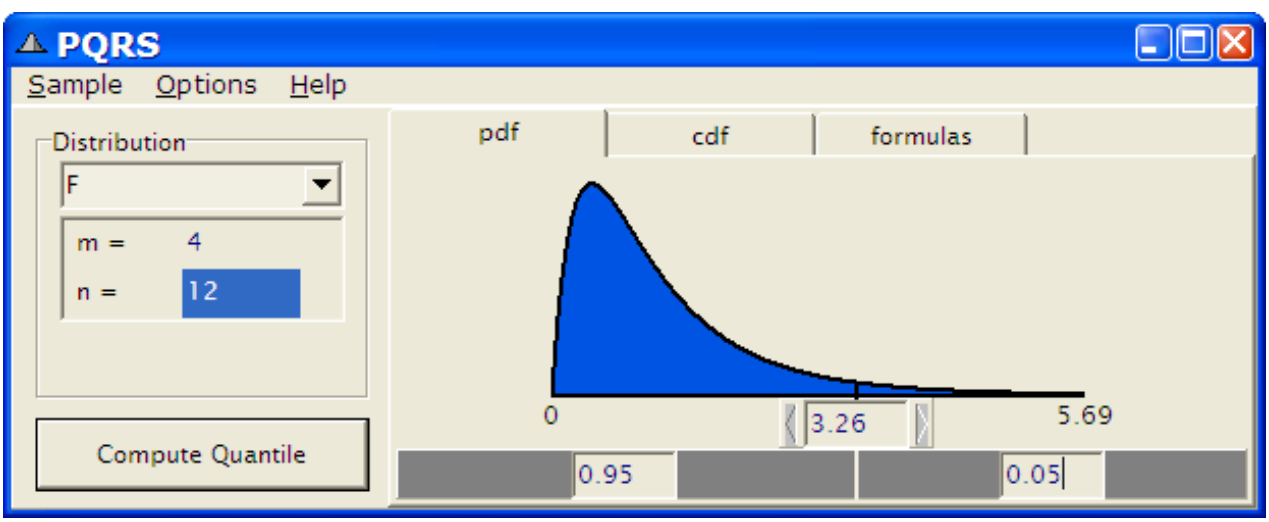
Υπενθυμίζουμε ότι το α ισούται με ένα μείον το επίπεδο εμπιστοσύνης. Συνήθως το επίπεδο εμπιστοσύνης είναι ίσο με **95%** οπότε το α είναι ίσο με **5%** ή **0.05**.

Να ένα παράδειγμα υπολογισμού κρίσιμης τιμής από πίνακες:



Επειδή οι πίνακες της κατανομής F δεν περιέχουν τιμές για όλους τους βαθμούς ελευθερίας, αν χρειαστεί, κάνουμε γραμμική παρεμβολή (**linear interpolation**) ανάμεσα στις τιμές του πίνακα.

Αν έχουμε πρόσβαση σε υπολογιστή, κάνουμε χρήση κατάλληλου λογισμικού όπως το δωρεάν πρόγραμμα **PQRS**, για παράδειγμα στο παρακάτω σχήμα υπολογίζουμε επαληθεύουμε την κρίσιμη τιμή της κατανομής (F=3.26) για βαθμούς ελευθερίας (4,12):



Ας πούμε τώρα δυο λόγια για την **λογική** πίσω από την ανάλυση **one-way ANOVA**.

Καταρχήν υπενθυμίζουμε ότι για να συγκρίνουμε τις μέσες τιμές (means) μιας μεταβλητής ανάμεσα σε **δυο** πληθυσμούς (populations), χρησιμοποιούμε **t-test**.

Η **ANOVA** μας επιτρέπει να συγκρίνουμε τις μέσες τιμές μιας εξαρτημένης (**dependent**) μεταβλητής ανάμεσα σε **περισσότερους από δυο** πληθυσμούς!

Στην **one-way ANOVA**, οι πληθυσμοί καθορίζονται από τις τιμές μιας μόνο άλλης ανεξάρτητης (**independent**) μεταβλητής, που ονομάζεται παράγοντας (**factor**). Οι δυνατές τιμές του παράγοντα (**factor**) ονομάζονται επίπεδα (**levels**).

Για να επιστρέψουμε τώρα στο αρχικό **παράδειγμα** ώστε να αποσαφηνίσουμε τις έννοιες που συζητήσαμε.

Θέλουμε να συγκρίνουμε την μέση ετήσια ενεργειακή κατανάλωση ενός νοικοκυριού (**annual mean energy consumption per household**). Η μεταβλητή αυτή είναι η εξαρτημένη μεταβλητή (**dependent variable**).

Έστω ότι εξετάζουμε τις ακόλουθες 5 περιοχές της Ελλάδας:
1. Πελοπόννησος
2. Δυτική Ελλάδα
3. Ανατολική Ελλάδα
4. Μακεδονία και Θράκη
5. Νησιωτική Ελλάδα

Οι 5 αυτοί πληθυσμοί των οποίων θα συγκρίνουμε την ενεργειακή κατανάλωση, ορίζονται από τον παράγοντα (**factor**) περιοχή (**region**), που λαμβάνει τις ανωτέρω 5 διακριτές τιμές. Η μεταβλητή αυτή είναι η ανεξάρτητη μεταβλητή (**independent variable**).

Δηλαδή, η ANOVA αποτελεί γενίκευση του **pooled t-test** σε περισσότερους από δυο πληθυσμούς.

Ας εξετάσουμε τώρα τις προϋποθέσεις (**assumptions**) για την εφαρμογή της ANOVA:
1. Τα δείγματα (**samples**) που επιλέγονται από τους πληθυσμούς πρέπει να είναι ανεξάρτητα (**independent**).
2. Η μεταβλητή που αναλύεται (εξαρτημένη) πρέπει να είναι κανονικά κατανεμημένη σε κάθε ένα από τους πληθυσμούς. Η προϋπόθεση αυτή λέγεται **normality assumption**.
   - Για να ελέγξουμε την ικανοποίηση αυτής της προϋπόθεσης μπορούμε να ελέγξουμε το σχήμα της κατανομής της εξαρτημένης μεταβλητής σε κάθε πληθυσμό με ιστογράμματα (**histograms**)
   - Ο έλεγχος αυτός γίνεται **πριν** από την ANOVA.
   - Σημειώνεται ότι η ANOVA είναι ανθεκτική (**robust**) σε μέτριες παραβιάσεις (**violations**) αυτής της προϋπόθεσης.
3. Η τυπική απόκλιση (**standard deviations**) της μεταβλητής που εξετάζεται πρέπει να είναι ίδια σε κάθε πληθυσμό.
   - Ως εμπειρικό κανόνα (rule of thumb), θεωρούμε ότι η προϋπόθεση των ίσων τυπικών αποκλίσεων ικανοποιείται εάν ο λόγος της μεγαλύτερης προς τη μικρότερη τυπική απόκλιση είναι μικρότερος από 2.
   - Και εδώ σημειώνεται ότι η ANOVA είναι επίσης ανθεκτική (**robust**) σε μέτριες παραβιάσεις (**violations**) αυτής της προϋπόθεσης.

Οι προϋποθέσεις της **κανονικότητας** και των **ίσων τυπικών αποκλίσεων** μπορούν να ελεγχθούν με γραφική ανάλυση των καταλοίπων (**residuals**).

Το κατάλοιπο (residual) μιας παρατήρησης (observation) ορίζεται ως η διαφορά ανάμεσα στην παρατήρηση και το μέσο όρο του πληθυσμού στον οποίο ανήκει.

Ο έλεγχος αυτός είναι παρόμοιος με αυτόν που γίνεται στην παλινδρόμηση (**regression**) και γίνεται μετά την ANOVA.

Για να εμπεδώσουμε την λογική της ANOVA, ας θεωρήσουμε ότι έχουμε δυο ανεξάρτητα τυχαία δείγματα (**independent random samples**) από δυο πληθυσμούς (populations).

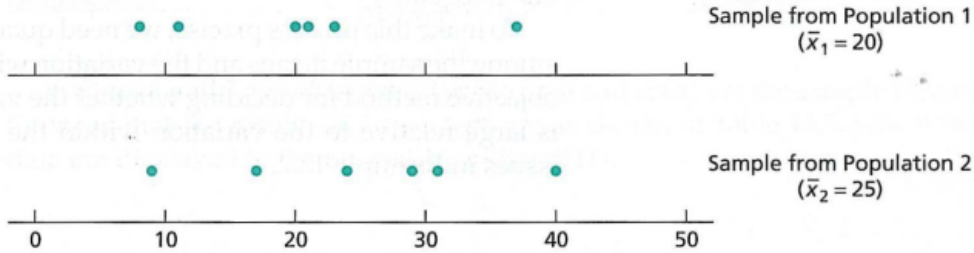
Έστω ότι $\bar{x}_1 = 20$ και $\bar{x}_2 = 25$. Μπορούμε, από τα δεδομένα αυτά, να συμπεράνουμε ότι $\mu_1 \neq \mu_2$;

Στην ακόλουθη περίπτωση (έστω ότι τα δείγματα αποκαλούνται **POP1** και **POP2**), πιθανότατα $\mu_1 = \mu_2$:

TABLE 13.1
Sample data from Populations 1 and 2

Sample from Population 1	21	37	11	20	8	23
Sample from Population 2	24	31	29	40	9	17

FIGURE 13.3
Dotplots for sample data in Table 13.1

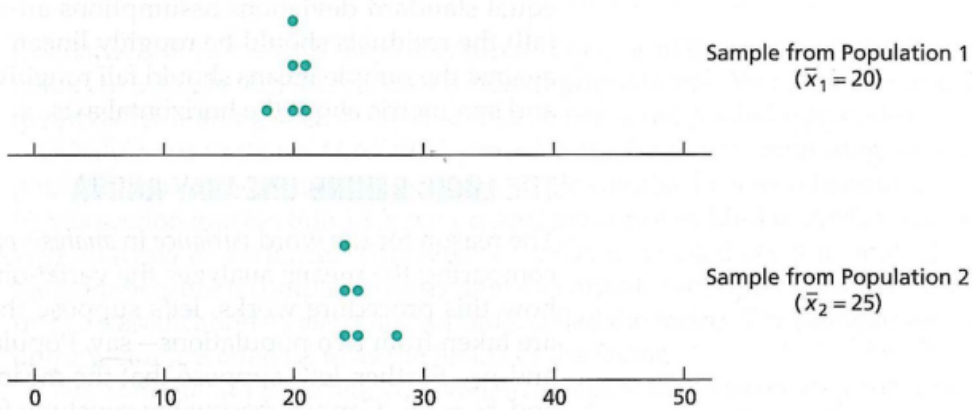


Ενώ στην ακόλουθη περίπτωση (έστω ότι τα δείγματα αποκαλούνται **POP3** και **POP4**), μάλλον $\mu_1 \neq \mu_2$:

TABLE 13.2
Sample data from Populations 1 and 2

Sample from Population 1	21	21	20	18	20	20
Sample from Population 2	25	28	25	24	24	24

FIGURE 13.4
Dotplots for sample data in Table 13.2



Άρα δεν φθάνει να ξέρουμε το μέσο όρο των δειγμάτων – πρέπει να ξέρουμε και τη διασπορά μέσα σε κάθε δείγμα!

Περιγραφικά στατιστικά μεγέθη για τα 4 αυτά δείγματα (**POP1**, **POP2**, **POP3** και **POP4**):

Descriptive Statistics

Variable	N	Mean	SD	Minimum	Maximum
POP1	6	20.000	10.237	8.0000	37.000
POP2	6	25.000	10.936	9.0000	40.000
POP3	6	20.000	1.0954	18.000	21.000
POP4	6	25.000	1.5492	24.000	28.000

Στα δυο πρώτα δείγματα (**POP1** και **POP2**), η διαφορά ανάμεσα στους μέσους όρους ($\bar{x}_1 = 20$ και $\bar{x}_2 = 25$) δεν είναι μεγάλη σε σχέση με την τυπική απόκλιση ανάμεσα στα δείγματα (s_1=10.237 και s_2=10.936)!

Ως εκ τούτου, δεν μπορούμε να είμαστε σίγουροι κατά πόσον η διαφορά ανάμεσα στους μέσους όρους των **δειγμάτων** οφείλεται στην διαφορά των μέσων όρων των **πληθυσμών** ή απλά στη διασπορά των παρατηρήσεων μέσα σε κάθε πληθυσμό!

Η διασπορά αυτή μέσα στα δείγματα αποτελεί «θόρυβο» (noise), που μπερδεύει την εικόνα και δεν μας επιτρέπει να βγάλουμε ασφαλή συμπεράσματα!

Στα δυο δεύτερα δείγματα (**POP3** και **POP4**), η διαφορά ανάμεσα στους μέσους όρους ($\bar{x}_1 = 20$ και $\bar{x}_2 = 25$) είναι μεγάλη σε σχέση με την τυπική απόκλιση ανάμεσα στα δείγματα (s_1=1.0954 και s_2=1.5492)!

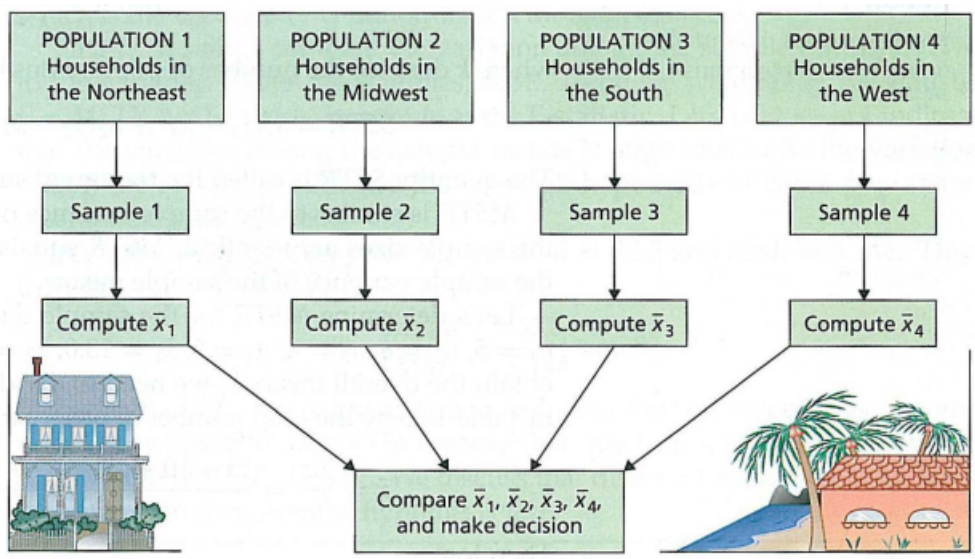
Αυτή τη φορά είναι σαφές ότι η διαφορά ανάμεσα στους μέσους όρους των **δειγμάτων** οφείλεται σε διαφορά ανάμεσα στους μέσους όρους των **πληθυσμών** και όχι στη διασπορά των παρατηρήσεων μέσα σε κάθε πληθυσμό!

Ας δουλέψουμε τώρα ένα **πλήρες πρόβλημα** ώστε να κατανοήσουμε πλήρως την εφαρμογή της μεθόδου.

Έστω ότι εξετάζουμε την ετήσια κατανάλωση ενέργειας ανά νοικοκυριό σε 4 περιοχές των ΗΠΑ
1. Northeast
2. Midwest
3. South και
4. West.

Η διαδικασία για τη σύγκριση της ενεργειακής κατανάλωσης στις 4 αυτές περιοχές μέσω ANOVA έχει ως εξής:

FIGURE 13.5
Process for comparing four population means



Έστω ότι τα πραγματικά δεδομένα είναι ως κατωτέρω

TABLE 13.3
Samples and their means of last year's energy consumptions for households in the four U.S. regions

Northeast	Midwest	South	West
15	17	11	10
10	12	7	12
13	18	9	8
14	13	13	7
13	15		9
	12		
13.0	14.5	10.0	9.2 ← *Means*

όπου οι ενεργειακές καταναλώσεις εκφράζονται σε δεκάδες εκατομμύρια BTU (British Thermal Units).

Η μηδενική (**null**) και εναλλακτική (**alternative**) υπόθεση είναι:

H_0: $\mu_1=\mu_2=\mu_3=\mu_4$
H_α: δεν είναι ίσοι όλοι οι μέσοι όροι των 4 πληθυσμών

Σε αντίθεση με απλούστερα τεστ, η διεξαγωγή ανάλυσης ANOVA με το χέρι δεν είναι ούτε εύκολη ούτε σκόπιμη.

Συνεπώς, την εκτελούμε με στατιστικό πακέτο, που μας δίνει τα ακόλουθα αποτελέσματα («έξοδος» του προγράμματος ή **output**):

One-Way AOV for: NORTHEAST MIDWEST SOUTH WEST

Source	DF	SS	MS	F	P
Between	3	97.500	32.5000	6.32	0.0050
Within	16	82.300	5.1438		
Total	19	179.800			

Grand Mean 11.900 CV 19.06

Homogeneity of Variances	F	P
Levene's Test	0.67	0.5856
O'Brien's Test	0.46	0.7127
Brown and Forsythe Test	0.79	0.5189

Welch's Test for Mean Differences

Source	DF	F	P
Between	3.0	5.64	0.0211
Within	8.3		

Component of variance for between groups 5.50797
Effective cell size 5.0

Variable	N	Mean	SE
NORTHEAST	5	13.000	1.0143
MIDWEST	6	14.500	0.9259
SOUTH	4	10.000	1.1340
WEST	5	9.200	1.0143

Ως συνήθως, στο output υπάρχουν πολλές πληροφορίες που μπορούμε, σε πρώτη φάση, να τις αγνοήσουμε.

Κάτω-κάτω, βλέπουμε τους μέσους όρους (**mean**) σε κάθε ένα από τα 4 δείγμα ($\bar{x}_1 = 13$, $\bar{x}_2 = 14.5$, $\bar{x}_3 = 10$ και $\bar{x}_4 = 9.2$).

Ενδιαφέρον ενδιάμεσο μέγεθος στην ANOVA είναι ο συνολικός μέσος όρος (**overall mean**) όλων των δειγμάτων, που μπορεί να βρεθεί και από τους μέσους όρους ($\bar{x}_i$) και τα μεγέθη (n_i) των δειγμάτων:

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i \bar{x}_i}{\sum_{i=1}^{k} n_i} = \frac{5 \times 13 + 6 \times 14.5 + 4 \times 10 + 5 \times 9.2}{5 + 6 + 4 + 5} = 11.9$$

Βλέπουμε ότι ο συνολικός μέσος όρος ($\bar{x} = 11.9$) είναι μικρότερος των μέσων όρων τριών περιοχών ($\bar{x}_1 = 13$, $\bar{x}_2 = 14.5$, $\bar{x}_3 = 10$) και μεγαλύτερος του μέσου όρου της τετάρτης ($\bar{x}_4 = 9.2$).

Αξίζει επίσης να αναφέρουμε από τον πρώτο-πρώτο πίνακα:

- η στήλη **DF** δείχνει τους βαθμούς ελευθερίας (Degrees of Freedom)
- η στήλη **SS** δείχνει το άθροισμα των τετραγώνων (Sum of Squares)
- η στήλη **MS** δείχνει το μέσο άθροισμα των τετραγώνων (Mean Square)

Οι τιμές των MS υπολογίζονται αν διαιρέσουμε το SS με το αντίστοιχο DF:

$$MS_{between} = \frac{SS_{between}}{DF_{between}} = \frac{97.5}{3} = 32.5$$

$$MS_{within} = \frac{SS_{within}}{DF_{within}} = \frac{82.3}{16} = 5.1438$$

Τέλος, η τιμή του F υπολογίζεται ως εξής:

$$F = \frac{MS_{between}}{MS_{within}} = \frac{32.5}{5.1438} = 6.32$$

Από όλα τα ανωτέρω αξίζει να θυμάστε ότι

- το $MS_{between}$ μετράει την διαφορά ανάμεσα στους μέσους όρους των δειγμάτων και για τον υπολογισμό του πρέπει να υπολογίσουμε το συνολικό μέσο όρο!
- το MS_{within} μετράει την διασπορά μέσα σε κάθε δείγμα και για τον υπολογισμό του, ουσιαστικά κάνουμε χρήση της προϋπόθεσης ότι η τυπική απόκλιση της μεταβλητής που εξετάζεται είναι ίδια σε κάθε πληθυσμό
- επομένως, μεγάλες τιμές του **λόγου** τους, δηλαδή του F, δείχνουν ότι η διαφορά ανάμεσα στους μέσους όρους των δειγμάτων είναι μεγάλη σχετικά με τη διασπορά των παρατηρήσεων μέσα σε κάθε δείγμα και ως εκ τούτου θα πρέπει να **απορρίψουμε** την μηδενική υπόθεση!

Να και μια άλλη όψη του πίνακα της ANOVA:

TABLE 13.4
ANOVA table format for a one-way analysis of variance

Source	df	SS	MS = SS/df	F-statistic
Treatment	$k-1$	$SSTR$	$MSTR = \frac{SSTR}{k-1}$	$F = \frac{MSTR}{MSE}$
Error	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$	
Total	$n-1$	SST		

Επικεντρωνόμαστε λοιπόν σε αυτό το **test statistic**, δηλαδή το F=6.32, με αντίστοιχη πιθανότητα p=0.0050. Επειδή 0.0050<0.05, απορρίπτουμε την **μηδενική** υπόθεση

$$H_0: \mu_1=\mu_2=\mu_3=\mu_4$$

και συμπεραίνουμε υπέρ της **εναλλακτικής**

$$H_\alpha: \text{δεν είναι ίσοι όλοι οι μέσοι όροι των 4 πληθυσμών}$$

Αν αγνοούσαμε την πιθανότητα που είναι ήδη υπολογισμένη από το στατιστικό πακέτο (p=0.0050) και θέλαμε να συγκρίνουμε την τιμή του test statistic (F=6.32) με την κρίσιμη τιμή που αντιστοιχεί σε επίπεδο εμπιστοσύνης 95%, θα έπρεπε να πάμε στους πίνακες της κατανομής F με βαθμούς ελευθερίας

$$(k-1,n-k)$$

όπου

k: ο αριθμός των δειγμάτων που εξετάζονται

n: ο συνολικός αριθμός παρατηρήσεων ($n = \sum_{i=1}^{k} n_i$)

Στην περίπτωσή μας:

$$k = 4 \Rightarrow k-1 = 3$$

$$n = n_1 + n_2 + n_3 + n_4 = 5+6+4+5 = 20 \Rightarrow n-k = 20-4 = 16$$

και από τον πίνακα της κατανομής F για βαθμούς ελευθερίας (3,16) βρίσκουμε

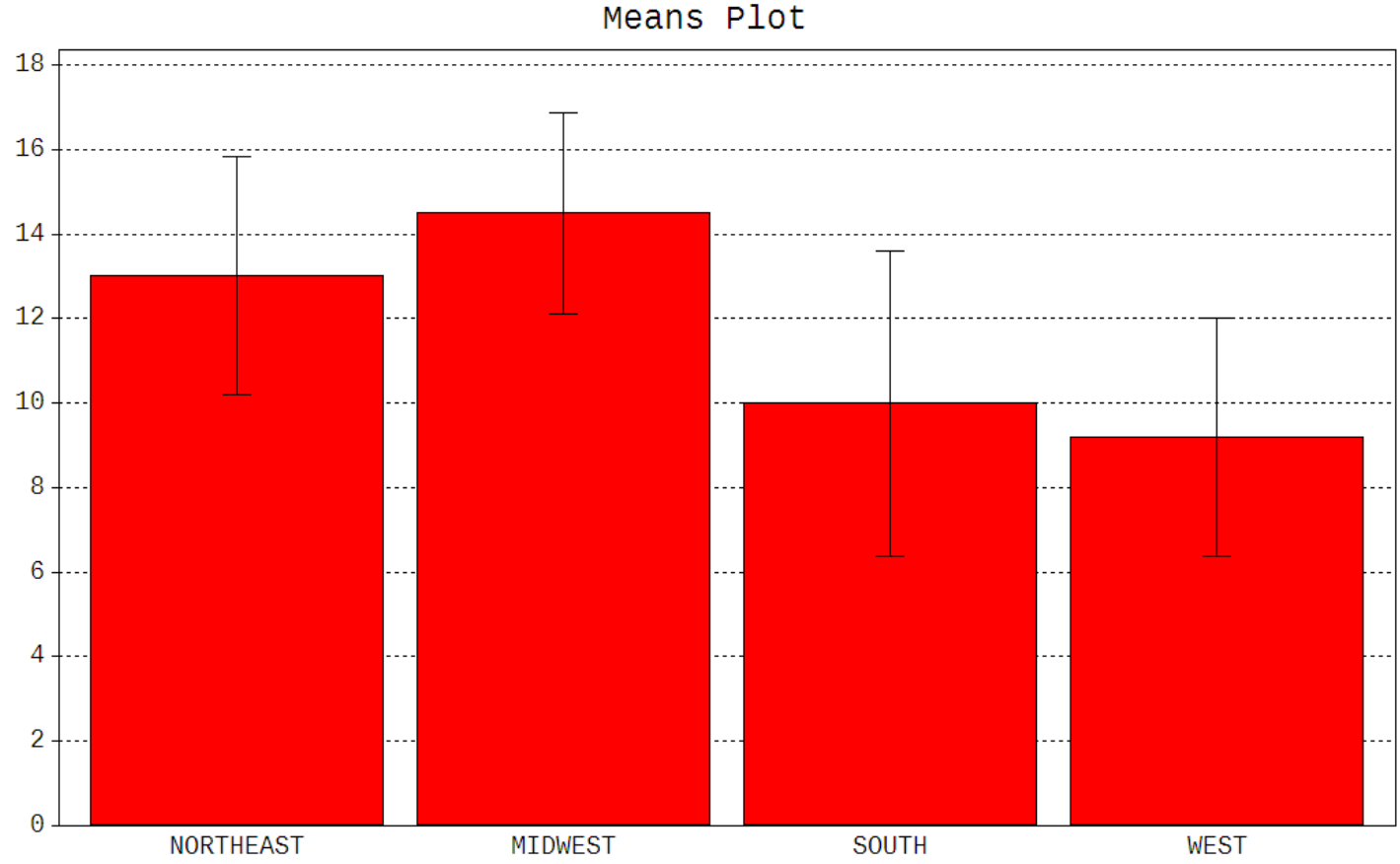
$$F_{critical}=3.24$$

Επαληθεύουμε λοιπόν ότι

$$F=6.32>3.24=F_{critical}$$

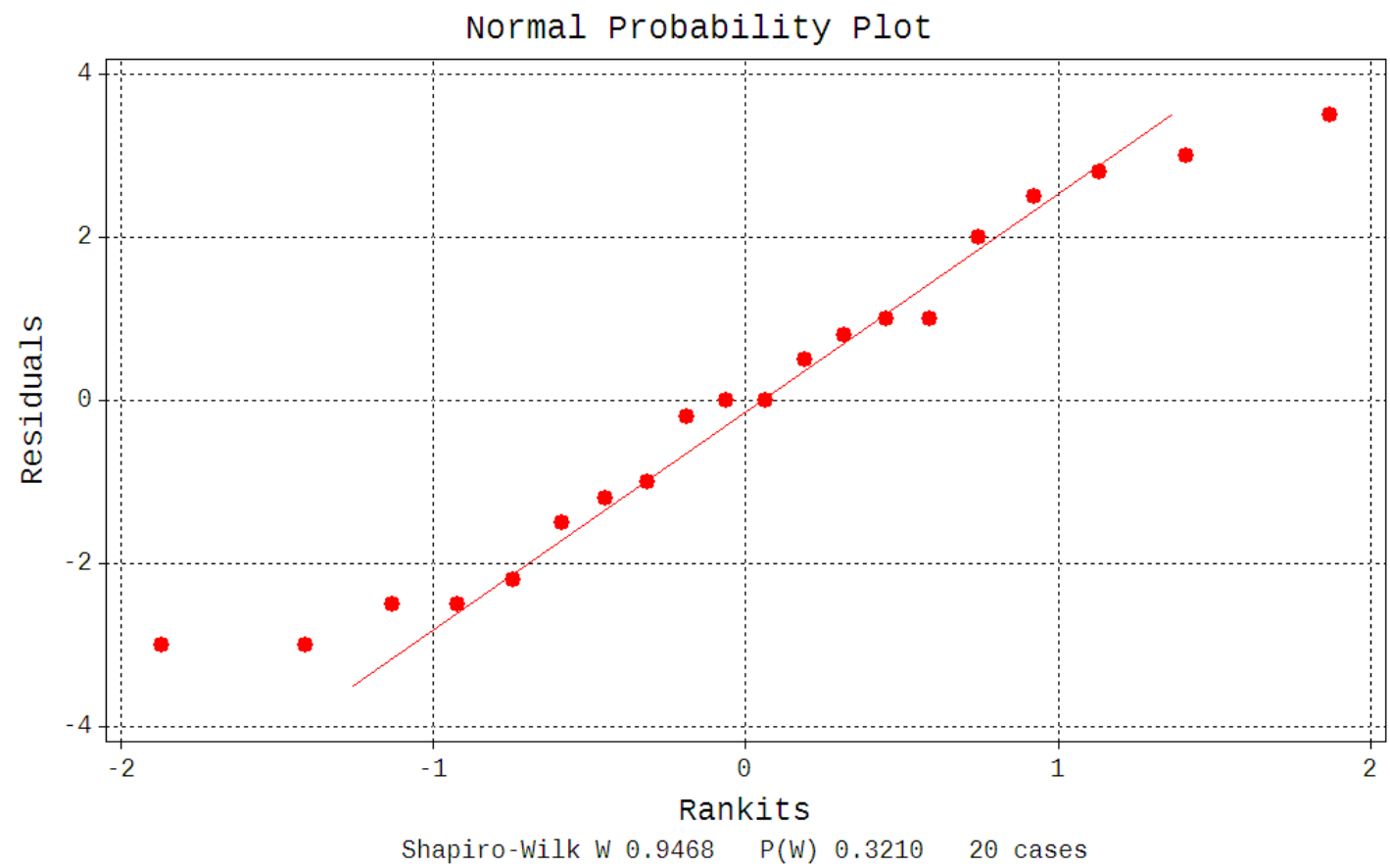
και ως εκ τούτου απορρίπτουμε τη μηδενική υπόθεση υπέρ της εναλλακτικής.

Χρήσιμο θα ήταν και να βλέπαμε μια γραφική παράσταση των μέσων όρων των πληθυσμών:



Στην κορυφή κάθε κόκκινης μπάρας είναι σχεδιασμένο διάστημα εμπιστοσύνης 95%.

Ο έλεγχος καταλοίπων (residuals) γίνεται με τα επόμενα δυο διαγράμματα:



Το πρώτο διάγραμμα καταλοίπων (**normal probability plot**) δείχνει ότι τα κατάλοιπα είναι κανονικά κατανεμημένα, με εξαίρεση τις ουρές του διαγράμματος (κάτω αριστερά και πάνω δεξιά) όπου τα σημεία ξεφεύγουν λίγο από την ευθεία γραμμή.

Το δεύτερο διάγραμμα καταλοίπων (**residuals by fitted value plot**) δείχνει ότι τα κατάλοιπα είναι σχετικά ομοιόμορφα κατανεμημένα πάνω και κάτω από τον άξονα του μηδενός και στους 4 εξεταζόμενους πληθυσμούς.

Κρίνουμε λοιπόν ότι οι γραφικοί έλεγχοι δεν μας έδειξαν κάτι ανησυχητικό ως προς την τήρηση των προϋποθέσεων.

Συνεπώς, τα δεδομένα των 4 δειγμάτων μας επιτρέπουν να συμπεράνουμε, σε επίπεδο εμπιστοσύνης 95%, ότι η μέση ετήσια ενεργειακή κατανάλωση ~~είναι~~ δεν είναι ίδια στις 4 αυτές περιοχές που εξετάσαμε.

Ας κάνουμε τώρα μερικές χρήσιμες παρατηρήσεις ως προς το τι μπορούμε και τι δεν μπορούμε να κάνουμε με την ANOVA.

Πρώτον, ενώ η ANOVA μας επιτρέπει να εξάγουμε το συμπέρασμα ότι η μέση ετήσια ενεργειακή κατανάλωση δεν είναι ίδια στις 4 περιοχές που εξετάσαμε, δεν μας επιτρέπει να γνωρίζουμε τη σχέση ανάμεσα στις περιοχές, δηλαδή ποια μέση ετήσια ενεργειακή κατανάλωση είναι μεγαλύτερη ή μικρότερη. Τέτοια συμπεράσματα μπορούν να εξαχθούν με τη μέθοδο των πολλαπλών συγκρίσεων (**multiple comparisons**), που δεν θα εξετασθούν και είναι εκτός ύλης.

Δεύτερον, τι θα κάναμε εάν κάποια από τις **προϋποθέσεις** της ANOVA δεν τηρείτο; Τότε, θα καταφεύγαμε σε μη παραμετρική (**nonparametric**) μέθοδο, που στην περίπτωση της ANOVA είναι ο έλεγχος **Kruskal Wallis**. Για την εφαρμογή του ελέγχου **Kruskal Wallis** αρκεί να έχουμε ανεξάρτητα δείγματα και το σχήμα της κατανομής στους διαφορετικούς πληθυσμούς να είναι ίδιο (όχι απαραίτητα κανονικό).