ΕΘΝΙΚΟ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ - ΙΑΤΡΙΚΗ ΣΧΟΛΗ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΕΦΑΡΜΟΣΜΕΝΗ ΝΕΥΡΟΑΝΑΤΟΜΙΑ»

«Βιοστατιστική, Μεθοδολογία και Συγγραφή Επιστημονικής Μελέτης»

Ενότητα 3: Ανάλυση Διακύμανσης κατά ένα παράγοντα – One-Way ANOVA

Δρ.Ευσταθία Παπαγεωργίου, Αναπληρώτρια Καθηγήτρια

Τμήμα Ιατρικών Εργαστηρίων

Ανάλυση Διακύμανσης κατά ένα παράγοντα

- Η ανάλυση διακύμανσης χρησιμοποιείται για τον έλεγχο της υπόθεσης της ισότητας των μέσων σε τρεις ή περισσότερους πληθυσμούς.
- Αποτελεί μια επέκταση του κριτηρίου t-test για δύο ανεξάρτητους πληθυσμούς.

Ανάλυση Διακύμανσης κατά ένα παράγοντα

- Η εξαρτημένη μεταβλητή (*response* variable) είναι η μεταβλητή για την οποία συγκρίνουμε τους μέσους
- Παράγοντας (*factor* variable) είναι η κατηγορική μεταβλητή που χρησιμοποιήθηκε για να διαχωριστούν οι πληθυσμοί και να δημιουργηθούν οι ομάδες σύγκρισης
  - Θα θεωρήσουμε k ομάδες (groups)

Ανάλυση Διακύμανσης κατά ένα παράγοντα

Συνθήκες ή Προϋποθέσεις εφαρμογής

- Οι παρατηρήσεις προέρχονται από πληθυσμούς που ακολουθούν την κανονική κατανομή
- Οι παρατηρήσεις αποτελούν τυχαία δείγματα από τους πληθυσμούς
- Οι διακυμάνσεις των πληθυσμών είναι ίσες

Ανάλυση Διακύμανσης κατά ένα παράγοντα

- Η μηδενική υπόθεση είναι ότι όλοι οι μέσοι είναι ίσοι:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$$

- Η εναλλακτική υπόθεση είναι ότι τουλάχιστον ένας εκ των μέσων διαφέρει

Ανάλυση Διακύμανσης κατά ένα παράγοντα

Παράδειγμα:

Η αίθουσα στατιστικής 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

Ανάλυση Διακύμανσης κατά ένα παράγοντα

- Με την ανάλυση διακύμανσης (ANOVA) δεν ελέγχουμε αν κάποιος μέσος είναι μικρότερος από κάποιον άλλο παρά μόνο αν υπάρχει ισότητα ή κάποιος εξ αυτών διαφέρει

$$H_0 : \mu_F = \mu_M = \mu_B$$

Ανάλυση Διακύμανσης κατά ένα παράγοντα

- Εκλέχθηκε ένα τυχαίο δείγμα των φοιτητών κάθε σειράς
- Η επίδοση των φοιτητών στις εξετάσεις καταγράφηκε ως εξής:
  - Front: 82, 83, 97, 93, 55, 67, 53
  - Middle: 83, 78, 68, 61, 77, 54, 69, 51, 63
  - Back: 38, 59, 55, 66, 45, 52, 52, 61

Ανάλυση Διακύμανσης κατά ένα παράγοντα

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

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

Ανάλυση Διακύμανσης κατά ένα παράγοντα

- Διακύμανση (Variation)
  - 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

Ανάλυση Διακύμανσης κατά ένα παράγοντα

- Είναι όλες οι τιμές ταυτόσημες;
  - Όχι, υπάρχει κάποια διασπορά στα δεδομένα
  - Ονομάζεται ολική διασπορά
  - Συμβολίζεται SS(Total) για το ολικό άθροισμα των τετραγώνων
  - Άθροισμα Τετραγώνων (Sum of Squares) είναι εναλλακτική ονομασία

Ανάλυση Διακύμανσης κατά ένα παράγοντα

- Είναι όλα τα δείγματα ταυτόσημα;
  - Όχι, υπάρχει διασπορά μεταξύ των δειγμάτων
  - Ονομάζεται διασπορά μεταξύ των δειγμάτων (between group variation)
  - Συμβολίζεται SS(B)

Ανάλυση Διακύμανσης κατά ένα παράγοντα

- Είναι όλες οι τιμές μέσα σ΄ένα δείγμα ταυτόσημες;
- Όχι, υπάρχει διασπορά εντός του κάθε δείγματος (within group variation)
- Ονομάζεται διασπορά μέσα στο δείγμα, υπόλοιπο ή σφάλμα
- Συμβολίζεται SS(W)

Ανάλυση Διακύμανσης κατά ένα παράγοντα

- 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

Ανάλυση Διακύμανσης κατά ένα παράγοντα

- Ο βασικός πίνακας one-way ANOVA είναι:

Source	SS	df	MS	F	p
Between					
Within					
Total					

Ανάλυση Διακύμανσης κατά ένα παράγοντα

- Μεγάλος Μέσος (Grand Mean)
  - Ο μεγάλος μέσος είναι ο μέσος όλων των τιμών όταν αγνοείται ο παράγοντας
  - Αποτελεί σταθμισμένο μέσο των μέσων των δειγμάτων

$$\overline{\overline{x}} = \frac{\sum_{i=1}^{k} n_i \overline{x}_i}{\sum_{i=1}^{k} n_i}$$

$$\overline{\overline{x}} = \frac{n_1 \overline{x}_1 + n_2 \overline{x}_2 + \cdots + n_k \overline{x}_k}{n_1 + n_2 + \cdots + n_k}$$

Ανάλυση Διακύμανσης κατά ένα παράγοντα

- Ο μεγάλος μέσος για το παράδειγμά μας είναι 65.08

$$\bar{\bar{x}} = \frac{7(75.71) + 9(67.11) + 8(53.50)}{7 + 9 + 8}$$

$$\bar{\bar{x}} = \frac{1562}{24}$$

$$\bar{\bar{x}} = 65.08$$

Ανάλυση Διακύμανσης κατά ένα παράγοντα

- Διασπορά μεταξύ των δειγμάτων, SS(B)
  - Η μεταξύ διασπορά των δειγμάτων είναι η διασπορά μεταξύ του μέσου του κάθε δείγματος και του μεγάλου μέσου
  - Κάθε διασπορά εξ αυτών σταθμίζεται από το μέγεθος του δείγματος

$$SS(B) = n_1\left(\bar{x}_1 - \bar{\bar{x}}\right)^2 + n_2\left(\bar{x}_2 - \bar{\bar{x}}\right)^2 + \cdots + n_k\left(\bar{x}_k - \bar{\bar{x}}\right)^2$$

$$SS(B) = \sum_{i=1}^{k} n_i\left(\bar{x}_i - \bar{\bar{x}}\right)^2$$

Ανάλυση Διακύμανσης κατά ένα παράγοντα

Η Διασπορά μεταξύ των δειγμάτων για το παράδειγμά μας είναι: SS(B)=1902

$$SS(B) = 1900.8376 \approx 1902$$

$$SS(B) = 7(75.71 - 65.08)^2 + 9(67.11 - 65.08)^2 + 8(53.50 - 65.08)^2$$

Ανάλυση Διακύμανσης κατά ένα παράγοντα

- Διασπορά εντός των δείγματα, SS(W)
  - Η εντός των δειγμάτων διασπορά είναι το σταθμισμένο ολικό των επιμέρους διασπορών
  - Η στάθμιση επιτυγχάνεται με τους βαθμούς ελευθερίας
  - Ο βαθμός ελευθερίας του κάθε δείγματος είναι κατά ένα λιγότερο από το μέγεθός του

Ανάλυση Διακύμανσης κατά ένα παράγοντα

Διασπορά εντός των δειγμάτων

$$SS(W) = \sum_{i=1}^{k} df_i s_i^2$$

$$SS(W) = df_1 s_1^2 + df_2 s_2^2 + \cdots + df_k s_k^2$$

Ανάλυση Διακύμανσης κατά ένα παράγοντα

- Η Διασπορά εντός των δειγμάτων για το παράδειγμά μας είναι 3386

$$SS(W) = 6(310.90) + 8(119.86) + 7(80.29)$$

$$SS(W) = 3386.31 \approx 3386$$

Ανάλυση Διακύμανσης κατά ένα παράγοντα

- Μετά τη συμπλήρωση του αθροίσματος των τετραγώνων, προκύπτει

Source	SS	df	MS	F	p
Between	1902				
Within	3386				
Total	5288				

Ανάλυση Διακύμανσης κατά ένα παράγοντα

- Οι μεταξύ των δειγμάτων β.ε. είναι κατά ένα λιγότερο από τον αριθμό των δειγμάτων
  - Έχουμε τρία δείγματα, συνεπώς οι βαθμοί ελευθερίας είναι δύο, df(B) = 2
- Οι εντός των δειγμάτων β.ε. είναι το άθροισμα των επιμέρους β.ε. του κάθε δείγματος
  - Τα μεγέθη των δειγμάτων είναι 7, 9, και 8
  - df(W) = 6 + 8 + 7 = 21
- Οι συνολικοί β.ε. Είναι κατά ένα λιγότερο από το μέγεθος του δείγματος
  - df(Total) = 24 – 1 = 23

Ανάλυση Διακύμανσης κατά ένα παράγοντα

- Η συμπλήρωση των βαθμών ελευθερίας δίνει:

Source	SS	df	MS	F	p
Between	1902	2			
Within	3386	21			
Total	5288	23			

Ανάλυση Διακύμανσης κατά ένα παράγοντα

- Διασπορές
  - Οι διασπορές ονομάζονται επίσης και μέση μεταβλητότητα (Mean of the Squares) και συντομογραφούνται ως MS, συχνά MS(B) ή MS(W)
  - Αποτελούν μία μέση τετραγωνική απόκλιση από το μέσο και υπολογίζονται διαιρώντας την διασπορά με τους βαθμούς ελευθερίας
  - MS = SS / df

$$Variance = \frac{Variation}{df}$$

Ανάλυση Διακύμανσης κατά ένα παράγοντα

- MS(B) = 1902 / 2 = 951.0
- MS(W) = 3386 / 21 = 161.2
- MS(T) = 5288 / 23 = 229.9
  - Σημειώστε ότι η MS(Total) δεν είναι το άθροισμα των MS(Between) και MS(Within).
  - Αυτό ισχύει για το άθροισμα τετραγώνων SS(Total) αλλά όχι για το μέσο τετράγωνο MS(Total)
  - Το MS(Total) συνήθως δεν εμφανίζεται στον πίνακα ANOVA

Ανάλυση Διακύμανσης κατά ένα παράγοντα

- Συμπληρώνοντας και τη στήλη MS παίρνουμε:

Source	SS	df	MS	F	p
Between	1902	2	951.0		
Within	3386	21	161.2		
Total	5288	23	229.9		

Ανάλυση Διακύμανσης κατά ένα παράγοντα

- Κριτήριο F
  - F = MS(B) / MS(W)
  - Το F κριτήριο είναι δεξιάς ουράς test
  - Το p-value είναι η περιοχή στα δεξιά
- Για τα δεδομένα μας, F = 951.0 / 161.2 = 5.9

Ανάλυση Διακύμανσης κατά ένα παράγοντα

- Το κριτήριο F είναι ο λόγος δύο εκτιμητών της διασποράς του πληθυσμού
- Αν ισχύει η μηδενική υπόθεση το κριτήριο F αναμένεται κοντά στη μονάδα
- Αν η η μηδενική υπόθεση δεν ισχύει, ο λόγος F αναμένεται μεγαλύτερος από τη μονάδα και όσο πιο μεγάλη είναι η διαφορά μεταξύ των δειγμάτων τόσο μεγαλύτερος είναι και ο λόγος F
- $P(F_{2,21} > 5.9) = 0.009$

Ανάλυση Διακύμανσης κατά ένα παράγοντα

- Προσθέτοντας και το κριτήριο F στον πίνακα:

Source	SS	df	MS	F	p
Between	1902	2	951.0	5.9	
Within	3386	21	161.2		
Total	5288	23	229.9		

Η κατανομή F

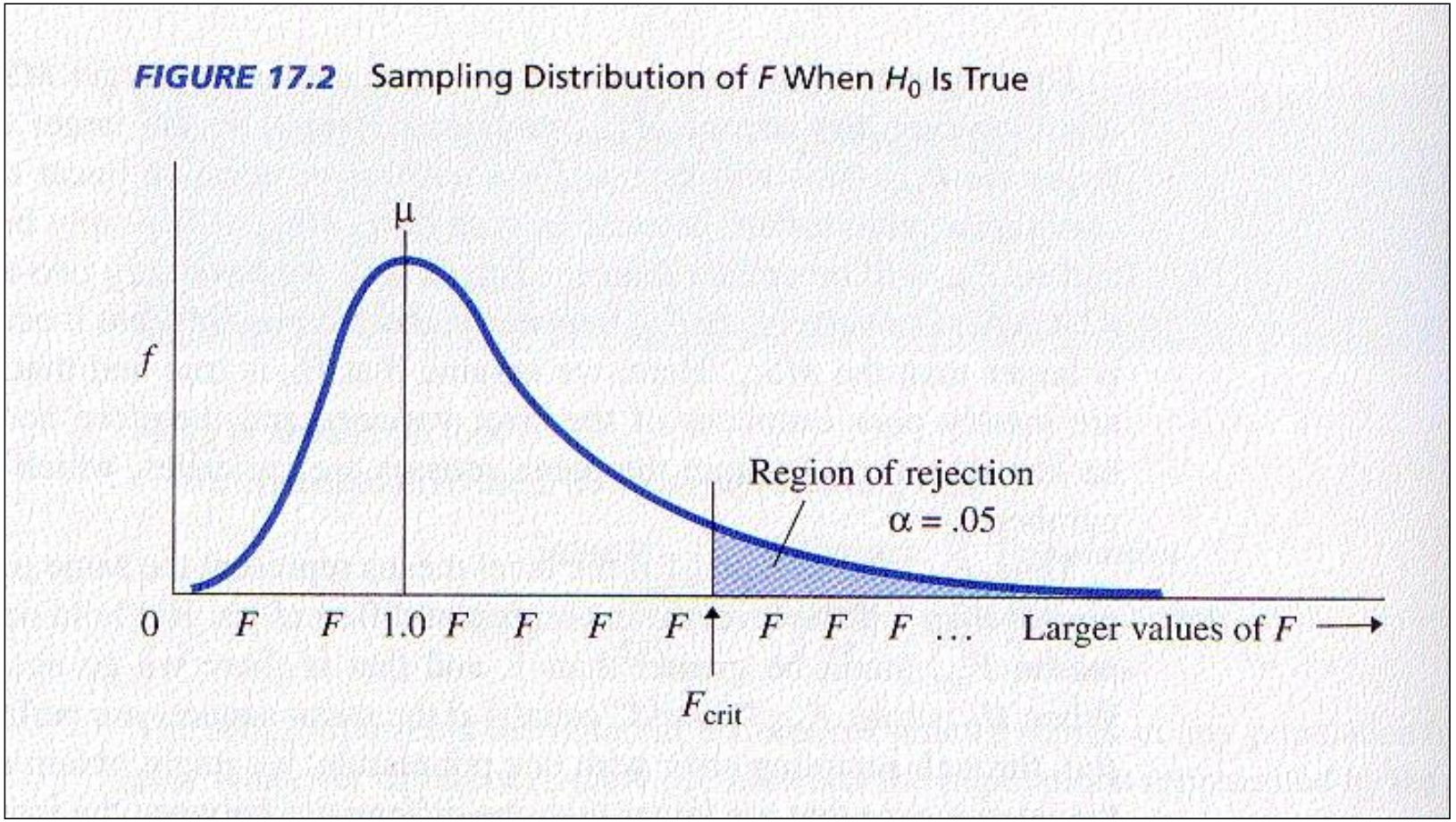
FIGURE 17.2 Sampling Distribution of F When H_0 Is True
μ
f
Region of rejection
α = .05
0
F
F
1.0
F
F
F
F
F
F
F
...
Larger values of F
F_{crit}

Πίνακας F κατανομής – Critical Values

	Numerator df: df_B				
df_W	1	2	3	4	5
5 5%	6.61	5.79	5.41	5.19	5.05
1%	16.3	13.3	12.1	11.4	11.0
10 5%	4.96	4.10	3.71	3.48	3.33
1%	10.0	7.56	6.55	5.99	5.64
12 5%	4.75	3.89	3.49	3.26	3.11
1%	9.33	6.94	5.95	5.41	5.06
21 5%	4.32	3.47	3.07	2.84	2.68
1%	8.02	5.78	4.87	4.37	4.04

Ανάλυση Διακύμανσης κατά ένα παράγοντα

- Ολοκληρώνουμε τον πίνακα με το p-value

Source	SS	df	MS	F	p
Between	1902	2	951.0	5.9	0.009
Within	3386	21	161.2		
Total	5288	23	229.9		

Ανάλυση Διακύμανσης κατά ένα παράγοντα

ANOVA

Επίδοση

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1901,516	2	950,758	5,896	,009
Within Groups	3386,317	21	161,253		
Total	5287,833	23			

Ανάλυση Διακύμανσης κατά ένα παράγοντα

- Το p-value είναι 0.009, το οποίο είναι μικρότερο από το επίπεδο σημαντικότητας 0.05, έτσι απορρίπτουμε τη μηδενική υπόθεση
- Επίσης η τιμή F=5.9 είναι μεγαλύτερη από τις τιμές της F κατανομής και για ε.σ. α=0.05 (F2,21=3.47) και για α=0.01(F2,21=5.78)
- Η μηδενική υπόθεση είναι ότι οι μέσοι και στις τρείς σειρές της αίθουσας είναι ίδιοι, αλλά την απορρίπτουμε και συνεπώς τουλάχιστον μια σειρά έχει διαφορετικό μέσο

Ανάλυση Διακύμανσης κατά ένα παράγοντα

- Υπάρχει σημαντική ένδειξη για να στηρίξουμε τον ισχυρισμό ότι υπάρχει διαφορά στις μέσες επιδόσεις των μπροστινών μεσαίων και πίσω σειρών στην τάξη
- Ο ANOVA δεν μας πληροφορεί για το ποια σειρά διαφέρει
- Για να διαπιστώσουμε ποια σειρά διαφέρει κάνουμε post hoc tests

Ανάλυση Διακύμανσης κατά ένα παράγοντα

Multiple Comparisons

Dependent Variable: Επίδοση

	(I) Συχνότητες	(J) Συχνότητες	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval
						Lower Bound
Tukey HSD	FRONT	MIDDLE	8,603	6,399	,387	-7,53
		BACK	22,214*	6,572	,008	5,65
	MIDDLE	FRONT	-8,603	6,399	,387	-24,73
		BACK	13,611	6,170	,093	-1,94
	BACK	FRONT	-22,214*	6,572	,008	-38,78
		MIDDLE	-13,611	6,170	,093	-29,16

Ανάλυση Διακύμανσης κατά ένα παράγοντα

Multiple Comparisons

Dependent Variable: Επίδοση

	(I) Συχνότητες	(J) Συχνότητες	95% Confidence Interval Upper Bound
Tukey HSD	FRONT	MIDDLE	24,73
		BACK	38,78
	MIDDLE	FRONT	7,53
		BACK	29,16
	BACK	FRONT	-5,65
		MIDDLE	1,94

Ανάλυση Διακύμανσης κατά ένα παράγοντα

- Παρατηρούμε ότι σε επίπεδο σημαντικότητας α=0.05, διαφέρουν η πρώτη και τρίτη σειρά

Σημειώματα

Σημείωμα Αδειοδότησης

Επεξήγηση όρων χρήσης έργων τρίτων

©	Δεν επιτρέπεται η επαναχρησιμοποίηση του έργου, παρά μόνο εάν ζητηθεί εκ νέου άδεια από το δημιουργό.
διαθέσιμο με άδεια CC-BY	Επιτρέπεται η επαναχρησιμοποίηση του έργου και η δημιουργία παραγώγων αυτού με απλή αναφορά του δημιουργού.
διαθέσιμο με άδεια CC-BY-SA	Επιτρέπεται η επαναχρησιμοποίηση του έργου με αναφορά του δημιουργού, και διάθεση του έργου ή του παράγωγου αυτού με την ίδια άδεια.
διαθέσιμο με άδεια CC-BY-ND	Επιτρέπεται η επαναχρησιμοποίηση του έργου με αναφορά του δημιουργού. Δεν επιτρέπεται η δημιουργία παραγώγων του έργου.
διαθέσιμο με άδεια CC-BY-NC	Επιτρέπεται η επαναχρησιμοποίηση του έργου με αναφορά του δημιουργού. Δεν επιτρέπεται η εμπορική χρήση του έργου.
διαθέσιμο με άδεια CC-BY-NC-SA	Επιτρέπεται η επαναχρησιμοποίηση του έργου με αναφορά του δημιουργού. και διάθεση του έργου ή του παράγωγου αυτού με την ίδια άδεια Δεν επιτρέπεται η εμπορική χρήση του έργου.
διαθέσιμο με άδεια CC-BY-NC-ND	Επιτρέπεται η επαναχρησιμοποίηση του έργου με αναφορά του δημιουργού. Δεν επιτρέπεται η εμπορική χρήση του έργου και η δημιουργία παραγώγων του.
διαθέσιμο με άδεια CC0 Public Domain	Επιτρέπεται η επαναχρησιμοποίηση του έργου, η δημιουργία παραγώγων αυτού και η εμπορική του χρήση, χωρίς αναφορά του δημιουργού.
διαθέσιμο ως κοινό κτήμα	Επιτρέπεται η επαναχρησιμοποίηση του έργου, η δημιουργία παραγώγων αυτού και η εμπορική του χρήση, χωρίς αναφορά του δημιουργού.
χωρίς σήμανση	Συνήθως δεν επιτρέπεται η επαναχρησιμοποίηση του έργου.

Διατήρηση Σημειωμάτων

Οποιαδήποτε αναπαραγωγή ή διασκευή του υλικού θα πρέπει να συμπεριλαμβάνει:

- το Σημείωμα Αναφοράς
- το Σημείωμα Αδειοδότησης
- τη δήλωση Διατήρησης Σημειωμάτων
- το Σημείωμα Χρήσης Έργων Τρίτων (εφόσον υπάρχει)

μαζί με τους συνοδευόμενους υπερσυνδέσμους.