ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν
Πανεπιστήμιον Αθηνών

Εισαγωγή στην Ανάλυση Γλωσσικών Δεδομένων

Ενότητα 3: Βασικές αρχές της επαγωγικής στατιστικής

Γεώργιος Κ. Μικρός
Φιλοσοφική Σχολή
Τμήμα Ιταλικής Γλώσσας και Φιλολογίας

Στόχοι της επαγωγικής στατιστικής

- **Εκτίμηση (estimation):**

  χρησιμοποιώντας τις μετρήσεις ενός δείγματος εκτιμούμε τις παραμέτρους του πληθυσμού

- **Έλεγχος Σημαντικότητας (significance testing):**

  αξιολογούμε το κατά πόσο διαφορές που εμφανίζονται σε διαφορετικές ομάδες ή μεταβλητές είναι πραγματικές ή προκύπτουν τυχαία.

Έλεγχος στατιστικής σημαντικότητας

- Όταν ελέγχουμε την στατιστική σημαντικότητα στην ουσία ελέγχουμε την πιθανότητα μια υπόθεση που έχουμε διαμορφώσει για τα δεδομένα μας να βγει αληθινή ή να διαψευσθεί.
  - Η υπόθεση που κάνουμε για τα δεδομένα μας λέγεται «ερευνητική υπόθεση» και συμβολίζεται με το Η1. Στην υπόθεση αυτή θεωρούμε ότι υπάρχει μια διαφορετική συμπεριφορά ομάδων ή μεταβλητών.
  - Η εναλλακτική υπόθεση που θα ισχύσει για τα δεδομένα μας αν η ερευνητική δεν αποδειχθεί λέγεται «μηδενική υπόθεση» και συμβολίζεται με το Η0. Σε αυτήν την περίπτωση θεωρούμε ότι οι ομάδες που εξετάζουμε ή οι μεταβλητές δεν εμφανίζουν διαφορές.

Παράδειγμα

- Διαφορές Ομάδων
  - Η1: Οι γυναίκες χρησιμοποιούν πιο συχνά πληθυντικό ευγενείας από τους άνδρες στον χώρο εργασίας.
  - Η0: Οι γυναίκες και οι άνδρες δεν εμφανίζουν διαφορές στη συχνότητα χρήσης του πληθυντικού ευγενείας στον χώρο εργασίας.
- Σχέση μεταβλητών
  - Η1: Το μέσο μήκος των λέξεων ενός κειμένου σχετίζεται με την δυσκολία κατανόησής του
  - Η0: Το μέσο μήκος των λέξεων ενός κειμένου **δεν** σχετίζεται με την δυσκολία κατανόησής του

Επίπεδο σημαντικότητας

- Η πιθανότητα (p) που ο ερευνητής θέτει ως όριο για να απορρίψει την μηδενική υπόθεση ονομάζεται «επίπεδο σημαντικότητας» (significance level).
- Παραδοσιακά στις κοινωνικές επιστήμες το επίπεδο σημαντικότητας τίθεται στο 0,05 ή αλλιώς θεωρούμε ότι αν επαναλάβουμε το πείραμα ή την έρευνα 100 φορές θα πρέπει να επιβεβαιώσουμε τα αποτελέσματά μας τουλάχιστον 95 φορές.
- Το επίπεδο σημαντικότητας σε άλλες επιστήμες μπορεί να διαφέρει σημαντικά αφού υπάρχουν επιστήμες (Ιατρική, Αστρονομία) όπου αβεβαιότητες της τάξης του 5% μεταφράζονται σε χαμένες ανθρώπινες ζωές. Έτσι τίθενται επίπεδα σημαντικότητας αρκετά μικρότερα. Συνηθισμένα επίπεδα είναι: 0,01 και 0,001.

Η βασική δομή της ερευνητικής διαδικασίας

1. Θέτουμε την ερευνητική υπόθεση και την μηδενική υπόθεση
2. Διεξάγουμε την έρευνα
3. Ελέγχουμε την μηδενική υπόθεση
    1. Θέτουμε το επίπεδο σημαντικότητας
    2. Επιλέγουμε στατιστικό τεστ και υπολογίζουμε την στατιστική τιμή
    3. Συγκρίνουμε την στατιστική τιμή με την κρίσιμη τιμή ενός τεστ

Πίνακας κρίσιμων τιμών για το t test

	LEVEL OF SIGNIFICANCE FOR ONE-TAILED TEST					
	.05	*.025*	*.01*	*.005*	*.001*	*.0005*
	LEVEL OF SIGNIFICANCE FOR TWO-TAILED TEST					
df	*.10*	*.05*	*.02*	*.01*	*.002*	*.001*
1	6.314	12.706	31.820	63.657	318.309	636.619
2	2.920	4.303	6.965	9.925	22.327	31.599
3	2.353	3.182	4.541	5.841	10.215	12.924
4	2.132	2.776	3.747	4.604	7.173	8.610
5	2.015	2.571	3.365	4.032	5.893	6.869
6	1.943	2.447	3.143	3.707	5.208	5.959
7	1.895	2.365	2.998	3.499	4.785	5.408
8	1.860	2.306	2.896	3.355	4.501	5.041
9	1.833	2.262	2.821	3.250	4.297	4.781
10	1.812	2.228	2.764	3.169	4.144	4.587
11	1.796	2.201	2.718	3.106	4.025	4.437
12	1.782	2.179	2.681	3.055	3.930	4.318
13	1.771	2.160	2.650	3.012	3.852	4.221
14	1.761	2.145	2.624	2.977	3.787	4.140
15	1.753	2.131	2.602	2.947	3.733	4.073
16	1.746	2.120	2.583	2.921	3.686	4.015
17	1.740	2.110	2.567	2.898	3.646	3.965
18	1.734	2.101	2.552	2.878	3.610	3.922
19	1.729	2.093	2.539	2.861	3.579	3.883
20	1.725	2.086	2.528	2.845	3.552	3.850
21	1.721	2.080	2.518	2.831	3.527	3.819
22	1.717	2.074	2.508	2.819	3.505	3.792
23	1.714	2.069	2.500	2.807	3.485	3.768
24	1.711	2.064	2.492	2.797	3.467	3.745
25	1.708	2.060	2.485	2.787	3.450	3.725
26	1.706	2.056	2.479	2.779	3.435	3.707
27	1.703	2.052	2.473	2.771	3.421	3.690
28	1.701	2.048	2.467	2.763	3.408	3.674
29	1.699	2.045	2.462	2.756	3.396	3.659
30	1.697	2.042	2.457	2.750	3.385	3.646
50	1.676	2.009	2.403	2.678	3.261	3.496
100	1.660	1.984	2.364	2.626	3.174	3.390
00	1.645	1.960	2.326	2.576	3.090	3.291

Παράδειγμα

- Έστω ότι ένας ερευνητής χρησιμοποιεί το t-test για να διακρίνει την διαφορά στη χρήση παθητικής φωνής μεταξύ ανδρών και γυναικών.
- Η ερευνητική υπόθεση που θα διαμορφώσει μπορεί να είναι μονόδρομη (one-tailed) ή δίδρομη (two-tailed). Δηλ. μπορεί να υποθέσει ότι οι άνδρες χρησιμοποιούν μεγαλύτερο ποσοστό από τις γυναίκες (μονόδρομη υπόθεση) ή να υποθέσει γενικά ότι άνδρες και γυναίκες χρησιμοποιούν διαφορετικά ποσοστά δίχως όμως να έχει συγκεκριμένη ιδέα για το ποιο φύλο χρησιμοποιεί περισσότερο την παθητική φωνή (δίδρομη υπόθεση).
- Έπειτα από τη συλλογή των **δεδομένων** ο ερευνητής επιλέγει το κατάλληλο στατιστικό τεστ για να συγκρίνει τη διαφορά στους μέσους όρους των δύο φύλων (το κατάλληλο τεστ είναι το t-test).
- Επιλέγει το επίπεδο σημαντικότητας το οποίο τις περισσότερες φορές είναι το 0,5
- Υπολογίζει την στατιστική τιμή του t-test η οποία είναι -2,55.
- Υπολογίζει τους βαθμούς ελευθερίας του t-test οι οποίοι τις περισσότερες φορές είναι N-1 για κάθε μεταβλητή (άρα 18).
- Συγκρίνει την στατιστική τιμή του t-test με την κρίσιμη τιμή που εμφανίζεται στον πίνακα κρίσιμων τιμών του t-test. Αν η τιμή είναι μεγαλύτερη τότε ο ερευνητής αποφασίζει να απορρίψει την μηδενική υπόθεση και να δεχθεί την ερευνητική με πιθανότητα λάθους 0,5 ή 5%.

Είδη στατιστικού λάθους

- Τύπου Ι (α λάθος):

  Συμβαίνει όταν ο ερευνητής απορρίπτει τη μηδενική υπόθεση και αποδέχεται την ερευνητική όταν στην ουσία η μηδενική είναι ορθή και θα έπρεπε να γίνει αποδεκτή

- Τύπου ΙΙ (β λάθος) ή λάθος αποδοχής:

  Αποτελεί το αντίθετο του α λάθους και συνίσταται στην αποδοχή της μηδενικής υπόθεσης όταν αυτή στην πραγματικότητα δεν ισχύει.

	Η0 ορθή	**Η0 λανθασμένη**
Αποδοχή Η0	Σωστό	**β λάθος**
Απόρριψη Η0	**α λάθος**	Σωστό

Αναλύοντας διαφορές μεταξύ ομάδων - Στατιστικά τεστ

- Κατηγορικά δεδομένα:
  - χ^2
- Ποιοτικά δεδομένα:
  - Median test
  - Mann-Whitney U test
  - Kruskal-Wallis test
- Αριθμητικά δεδομένα
  - t test
  - ANOVA

Κατηγορικά δεδομένα

- Το χ^2 εξετάζει διαφορές μεταξύ των κατηγοριών μιας ανεξάρτητης μεταβλητής σε σχέση με τις κατηγορίες μιας εξαρτημένης. Υπάρχουν δύο είδη:
  - χ^2 με μια μεταβλητή: εξετάζει διαφορές στις κατηγορίες μιας κατηγορικής μεταβλητής
  - χ^2 με δύο μεταβλητές: εξετάζει διαφορές στις κατηγορίες που εμφανίζονται σε δύο κατηγορικές εξαρτημένες ή ανεξάρτητες μεταβλητές

χ^2 με μια μεταβλητή

- Το τεστ μετράει τη συχνότητα σε κάθε κατηγορία της μεταβλητής (ονομάζονται παρατηρημένες συχνότητες – observed frequencies).
- Στη συνέχεια υπολογίζεται η αναμενόμενη συχνότητα (expected frequency) στις σχετικές κατηγορίες. Αυτή είναι η συχνότητα που θα εμφανιζόταν αν ίσχυε η μηδενική υπόθεση.
- Το χ^2 προκύπτει από τον ακόλουθο τύπο:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Παράδειγμα εφαρμογής του χ^2

- Διερεύνηση της χρήσης επιθέτων σε 5 θεματικά είδη κειμένων

	Επιστήμη	**Ιστορία**	**Πολιτική**	**Οικονομία**	**Σύνολο**
Παρατηρημένες συχνότητες (O)	48	41	27	13	129
Αναμενόμενες συχνότητες (E)	32,25	32,25	32,25	32,25	129

χ^2 με δύο μεταβλητές

- Στο χ^2 με δύο μεταβλητές επεκτείνουμε την χρήση του χ^2 με μια μεταβλητή. Το μόνο που αλλάζει είναι ο υπολογισμός των αναμενόμενων συχνοτήτων. Αυτός προκύπτει από το γινόμενο του συνόλου της στήλης και του συνόλου της σειράς το οποίο διαιρείται από το συνολικό άθροισμα των συχνοτήτων του πίνακα:

$$E = \frac{\sum(r_i) \cdot \sum(c_k)}{\sum(r)(c)}$$

Παράδειγμα εφαρμογής του χ^2 με δύο μεταβλητές

- Διερεύνηση της χρήσης των επιθέτων και των ρημάτων σε διαφορετικά είδη κειμένων.

	Επίθετα	**Ουσιαστικά**	**Άθροισμα σειράς**
Επιστήμη			
Ο	48	7	55
Ε	36,39	18,61	
Οικονομικά			
Ο	27	16	43
Ε	28,45	14,55	
Ιστορία			
Ο	13	22	35
Ε	23,16	11,84	
Άθροισμα στήλης	88	45	133
			Συνολικό άθροισμα

Ποιοτικά δεδομένα

- Πολλές φορές ο ερευνητής ενδιαφέρεται να συγκρίνει δύο ομάδες ανθρώπων σχετικά με το πώς αξιολογούν μια μεταβλητή.
- Π.χ. Μπορούμε να μελετήσουμε κατά πόσο μια ομάδα ομιλητών που έχουν ως μητρική γλώσσα τα ελληνικά αξιολογεί την γραμματικότητα συγκεκριμένων προτάσεων με μια ομάδα ξένων που μαθαίνουν τα ελληνικά ως ξένη γλώσσα.

Ποιοτικά δεδομένα – Median test

- Για το προηγούμενο παράδειγμα μπορούμε να χρησιμοποιήσουμε το median test. Θα πρέπει να καταχωρήσουμε όλες τις τιμές της κάθε ομάδας και να υπολογίσουμε την διάμεσο, την τιμή εκείνη όπου χωρίζει τα δεδομένα μας σε ισοπληθείς ομάδες.
- Στη συνέχεια θα χωρίσουμε τα δεδομένα σε δύο ομάδες. Η πρώτη θα περιλαμβάνει τον αριθμό των περιπτώσεων πάνω από τη διάμεσο και η δεύτερη τον αριθμό των περιπτώσεων κάτω από τη διάμεσο. Έτσι θα συνεχίσουμε και θα υπολογίσουμε το χ^2 με τον τρόπο που δείξαμε πριν.

Ποιοτικά δεδομένα – Άλλα test

- **Mann-Whitney U-test:** χρησιμοποιείται για να αναλύσει διαφορές δύο ομάδων όταν τα δεδομένα είναι σοβαρά σκεβρωμένα.
- **Kruskal-Wallis:** χρησιμοποιείται για να αναλύσει διαφορές περισσότερων των δύο ομάδων.
- **Wilkoxon signed-rank test:** χρησιμοποιείται για να εξετάσει διαφορές μεταξύ κατατάξεων για ένα φαινόμενο που έχουν προκύψει από δύο κριτές.
- **Friedman test:** χρησιμοποιείται για να εξετάσει διαφορές μεταξύ κατατάξεων για ένα φαινόμενο που έχουν προκύψει από πολλούς κριτές

Αριθμητικά δεδομένα – t-test

- Οι ερευνητές χρησιμοποιούν t-test όταν θέλουν να εξετάσουν κατά πόσο η διαφορά μεταξύ **δύο ομάδων** σε κάποιο φαινόμενο που μπορεί να μετρηθεί σε **αριθμητική κλίμακα** είναι πραγματική και όχι τυχαία. Υπάρχουν δύο είδη:
  - t-test ανεξάρτητων δειγμάτων (independent sample t-test)
  - t-test εξαρτημένων δειγμάτων (related measures t-test)

Αριθμητικά δεδομένα – Ανάλυση Διακύμανσης (Analysis of Variance – ANOVA)

- Η ΑΔ είναι η πλέον χρησιμοποιημένη μέθοδος στις κοινωνικές και ψυχολογικές επιστήμες. Όταν θέλουμε να συγκρίνουμε περισσότερες από δύο ομάδες τότε το t-test είναι ακατάλληλο λόγω του προσθετικού λάθους που παράγει. Στην περίπτωση αυτή χρησιμοποιούμε την ΑΔ.
- Προϋπόθεση για να εφαρμόσουμε την ANOVA είναι η ύπαρξη μιας εξαρτημένης αριθμητικής μεταβλητής και η ύπαρξη μιας ή περισσότερων ανεξάρτητων κατηγορικών μεταβλητών.
- Η ΑΔ επιτρέπει την σύγκριση πολλών μέσων όρων και λειτουργεί συγκρίνοντας την διακύμανση εντός της ομάδας και κατά μήκους των ομάδων. Ο λόγος των δύο διακυμάνσεων είναι η τιμή F η οποία έχει συγκεκριμένους βαθμούς ελευθερίας και η στατιστική σημαντικότητάς της ελέγχεται με βάσει σχετικούς πίνακες.
- Η ΑΔ εμφανίζει δύο γενικές μορφές:
  - ΑΔ μιας μεταβλητής
  - ΑΔ πολλών μεταβλητών

ANOVA μιας μεταβλητής

- Έστω ότι θέλουμε να εξετάσουμε την επίδραση του κειμενικού θέματος στην χρήση της γενικής – εως σε 20 κείμενα.
- Στην περίπτωση αυτή θα χρησιμοποιήσουμε μια ANOVA για να δούμε ποιες κατηγορίες εμφανίζουν διαφορετικούς μέσους όρους.
- Η τιμή F ωστόσο δεν μας λέει ποιες κατηγορίες διαφέρουν από ποιες. Ειδικότερα χρειαζόμαστε να ξέρουμε ποιοι μέσοι όροι διαφέρουν στατιστικά σημαντικά από ποιους. Για να λυθεί αυτό το θέμα χρησιμοποιούμε τα τεστ πολλαπλής σύγκρισης (multiple comparison test).
- Υπάρχουν πολλά τέτοια τεστ μερικά από τα οποία είναι:
  - Scheffe test
  - Tukey HSD test
  - Least Significant Difference (LSD) test

ANOVA πολλών μεταβλητών

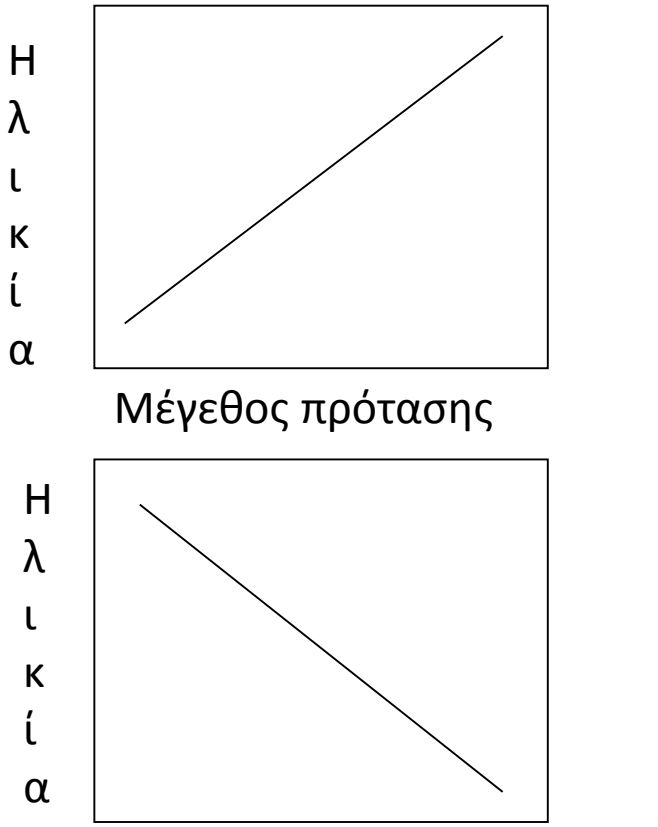
- Η ANOVA μπορεί να περιλαμβάνει περισσότερες της μιας ανεξάρτητων μεταβλητών. Για παράδειγμα θα μπορούσε κάποιος να εξετάσει την επίδραση του κειμενικού θέματος και του κειμενικού γένους στην διαμόρφωση του μέσου μήκους πρότασης του κειμένου.
- Στην περίπτωση αυτή έχουμε την διερεύνηση της επίδρασης δύο ανεξάρτητων κατηγορικών μεταβλητών (κειμενικό θέμα, κειμενικό γένος) στην εξαρτημένη αριθμητική μεταβλητή (μέσο μήκος πρότασης).
- Υπολογίζοντας ANOVA 2 μεταβλητών παίρνουμε δύο τύπους F τιμών. Ο πρώτος αναφέρεται στη γενική επίδραση της ανεξάρτητης μεταβλητής (main effects) και ο δεύτερος αναφέρεται στην επίδραση της αλληλεπίδρασης των δύο μεταβλητών (interaction effects).

Σχέσεις μεταβλητών

- Εκτός από διαφορές ομάδων μπορούμε να μελετήσουμε τις σχέσεις διαφόρων μεταβλητών. Μπορούμε δηλ.να μελετήσουμε τη συμπεριφορά μιας μεταβλητής όταν μια άλλη μεταβλητή αλλάζει.
- Οι πιθανές σχέσεις δύο μεταβλητών μπορεί να είναι οι ακόλουθες:
  - Σχέση
    - Γραμμική
    - Μη γραμμική
  - Μη σχέση

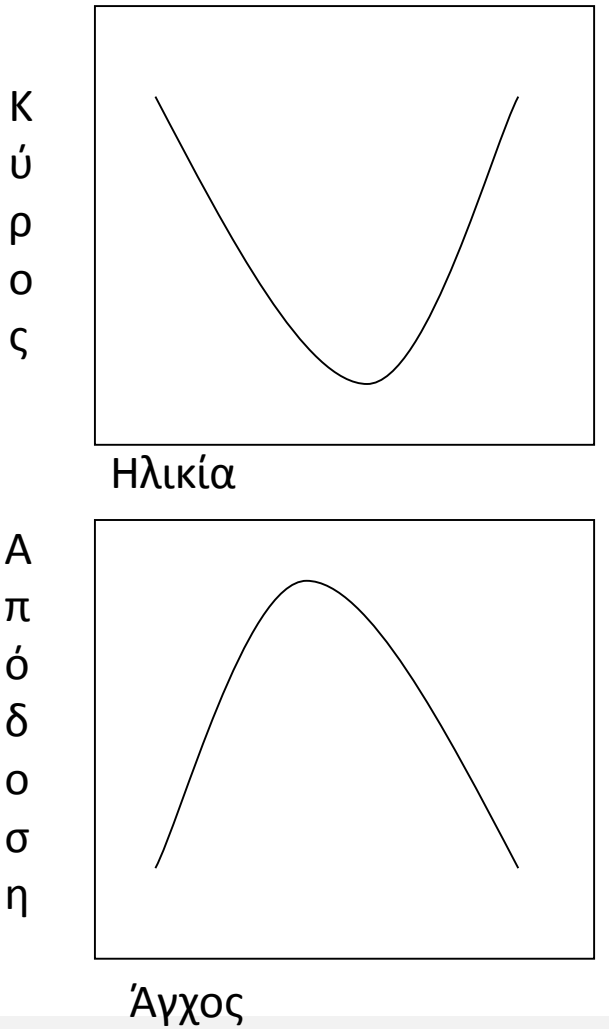
Γραμμική σχέση δύο μεταβλητών

- Η γραμμική σχέση εμφανίζεται γραφηματικά ως μια ευθεία γραμμή. Υπάρχουν δύο είδη γραμμικής συσχέτισης:
  - Θετική γραμμική σχέση: Όταν μεγαλώνει η μία μεταβλητή μεγαλώνει και η άλλη.
  - Αρνητική γραμμική σχέση: Όταν μεγαλώνει η μία μεταβλητή μικραίνει η άλλη.



Μη γραμμική σχέση δύο μεταβλητών

- Όταν η σχέση δύο μεταβλητών δεν μπορεί να αναπαρασταθεί με μια ευθεία γραμμή τότε ονομάζεται μη γραμμική σχέση. Τα σημαντικότερα είδη μη γραμμικής σχέσης είναι:
  - Καμπυλόγραμμη σχέση μορφής U
  - Ανάστροφη καμπυλόγραμμη σχέση μορφής U



Συσχέτιση (correlation)

- Η στατιστική έκφραση της σχέσης δύο μεταβλητών ονομάζεται συσχέτιση.
- Η συσχέτιση δύο μεταβλητών προσεγγίζεται από δύο μετρήσεις:
  - Συντελεστής συσχέτισης (correlation coefficient): Μας δίνει τον τύπο και την ισχύ της συσχέτισης
  - Συντελεστής προσδιορισμού (coefficient of determination): Μας προσδιορίζει το ποσοστό της ποικιλίας της μιας μεταβλητής που εξαρτάται από την ποικιλία της άλλης.

Συντελεστής συσχέτισης - ΣΣ

- Ο ΣΣ αποτελεί την αριθμητική έκφραση του τύπου και της ισχύος της σχέσης δύο μεταβλητών.
- Ισχύς: Ο ΣΣ παίρνει τιμές από 0 έως 1:
  - Το 0 υποδηλώνει ότι δεν υπάρχει καμία σχέση μεταξύ των δύο μεταβλητών
  - Το 1 υποδηλώνει ότι υπάρχει τέλεια σχέση μεταξύ των δύο μεταβλητών
- Τύπος: Ο τύπος εμφανίζεται ως πρόσημο στον συντελεστή:
  - Το + υποδηλώνει θετική σχέση μεταξύ των μεταβλητών
  - Το – υποδηλώνει αρνητική σχέση μεταξύ των μεταβλητών

Ερμηνεία των ΣΣ

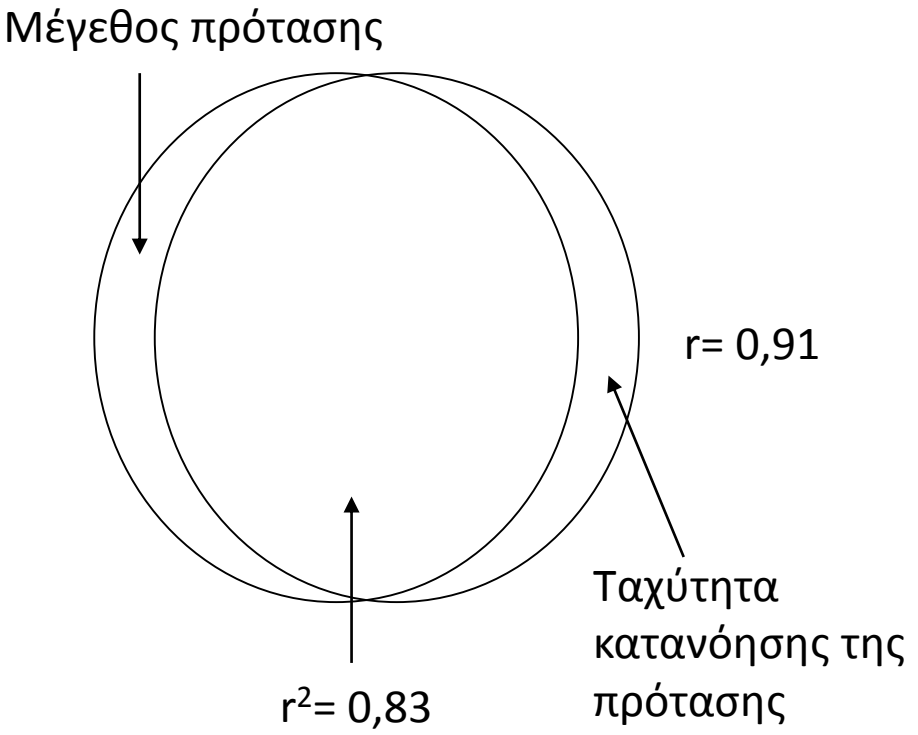
- Δεν υπάρχει αντικειμενικός προσδιορισμός της ισχύος ενός ΣΣ. Ως γενικό οδηγό ωστόσο μπορούμε να ακολουθήσουμε τον παρακάτω πίνακα:
  - < 0,20: Μικρή, σχεδόν ασήμαντη σχέση
  - 0,20 – 0,40: Χαμηλή συσχέτιση, σίγουρη, αλλά μικρή σχέση
  - 0,40 – 0,70: Μέτρια συσχέτιση, σημαντική σχέση
  - 0,70 – 0,90: Υψηλή συσχέτιση, έντονη σχέση
  - > 0,90: Πολύ υψηλή συσχέτιση, άμεσα εξαρτώμενη σχέση

Είδη ΣΣ

- Για αριθμητικές μεταβλητές:
  - Pearson product moment correlation ή Pearson's r
- Για μεταβλητές ποιοτικών δεδομένων
  - Spearman rho (r_s): Διαδεδομένος ΣΣ όταν οι μεταβλητές που συγκρίνονται είναι κατατάξεις ως προς κάποιο χαρακτηριστικό. Συνήθως προτιμάται όταν οι κατατάξεις αυτές έχουν προκύψει από αριθμητικά δεδομένα.
  - Kendall's tau (τ): Όταν δύο κριτές έχουν κατατάξει την ίδια σειρά αντικειμένων.
- Για κατηγορικές μεταβλητές
  - Phi (Φ) & Cramer's V : Για πίνακες με κατηγορικά δεδομένα

Συντελεστής Καθορισμού - ΣΚ

- Ο ΣΣ μας λέει αν και πόσο δυνατά σχετίζονται δύο μεταβλητές. Ωστόσο, δεν μας προσδιορίζει το ποσοστό της διακύμανσης μιας μεταβλητής που οφείλεται στην ύπαρξη της άλλης.
- Ο ΣΚ (ή r^2) παίρνει τιμές από 0 – 1 και προκύπτει από το τετράγωνο του ΣΣ.
- Π.χ. αν ο ΣΣ της σχέσης του μεγέθους μιας πρότασης και της ταχύτητας με την οποία αυτή κατανοείται είναι 0,91, τότε ο ΣΚ είναι 0,83



Ανάλυση Παλινδρόμησης – ΑΠ (Regression Analysis) (1/2)

- Η ΑΠ χρησιμοποιείται για να εξηγήσει ή να προβλέψει τις τιμές μιας αριθμητικής μεταβλητής στηριζόμενη σε μία ή περισσότερες μεταβλητές μεικτής φύσης.
- Η μεταβλητή που ερευνάται ονομάζεται **εξαρτημένη μεταβλητή**, ενώ η μεταβλητή ή οι μεταβλητές που χρησιμοποιούνται για να την προβλέψουν ή να την εξηγήσουν ονομάζονται **ανεξάρτητες μεταβλητές**.
- Π.χ Ας υποθέσουμε ότι ένας ερευνητής εξετάζει τη σχέση του μέσου μήκους λέξης ενός κειμένου και του μέσου μήκους πρότασης. Στην ΑΠ ο ερευνητής σχεδιάζει το ζεύγος τιμών των δύο μεταβλητών σε ένα διάγραμμα και στη συνέχεια προσαρμόζει μια ευθεία γραμμή μεταξύ των τιμών που έχουν παρατηρηθεί έτσι ώστε η γραμμή να απέχει το λιγότερο δυνατό από τα παρατηρημένα σημεία.

Ανάλυση Παλινδρόμησης – ΑΠ (Regression Analysis) (2/2)

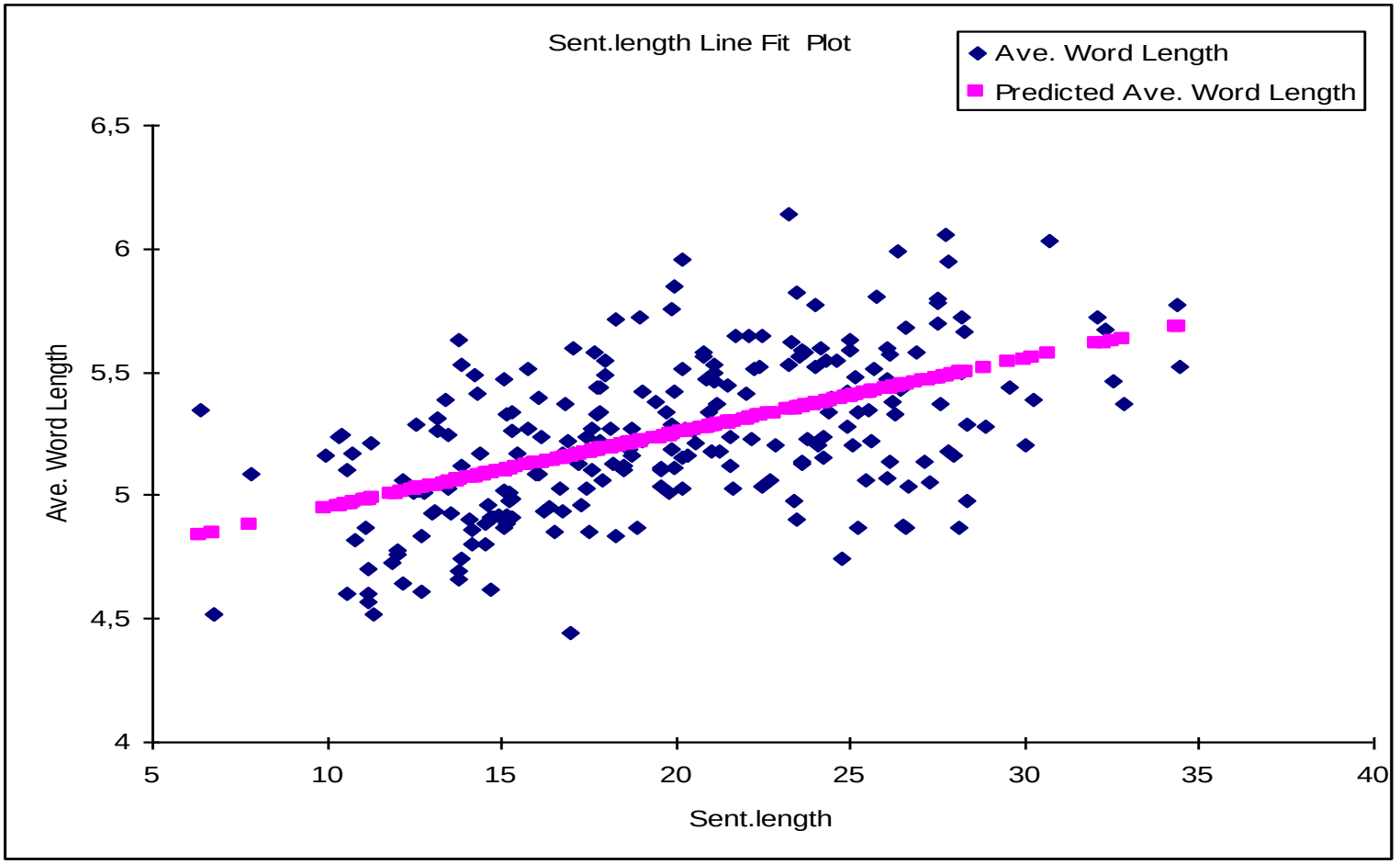
- Η ΑΠ στην ουσία είναι η δημιουργία μιας εξίσωσης που εκφράζει τη σχέση των μεταβλητών της ανάλυσης. Για την περίπτωση των δύο μεταβλητών η εξίσωση έχει τη γενική μορφή: y= a + bx όπου:

  y= Εξαρτημένη μεταβλητή (Μέσο μήκος λέξης)

  a= Τομή (πόσο ψηλά στον y άξονα τέμνει η γραμμή

  b= Κλίση (το μέγεθος σχέσης των δύο μεταβλητών ή πόσες μονάδες αυξάνεται το y σε κάθε μονάδα αύξησης του x.

Διάγραμμα Παλινδρόμησης



Αποτελέσματα ΑΠ

- Μήκος λέξης = 4,64 + 0,03 Χ (μήκος πρότασης)
- Πχ. Μήκος λέξης = 4,64 + 0,03 . 15 => Μήκος λέξης = 4,64 + 0,45 = **5,09**
- R^2 = 0,29
- Κάθε 1 λέξη που προσθέτουμε σε μια πρόταση αυξάνουμε το μήκος της λέξης κατά 0,03 ή 3%.

Πολλαπλή Παλινδρόμηση – ΠΠ (Multiple regression)

- Η ΑΠ μπορεί να επεκταθεί ώστε να συμπεριλάβει πολλές ανεξάρτητες μεταβλητές οι οποίες θα χρησιμοποιούνται ως όργανα πρόβλεψης (predictors) της εξαρτημένης μεταβλητής.
- Δεδομένου ότι τα περισσότερα φαινόμενα (φυσικά και κοινωνικά) έχουν πολυπαραγοντική φύση, η ΠΠ είναι η κατάλληλη ανάλυση για να διερευνήσει **ποιοί** παράγοντες επηρεάζουν ένα φαινόμενο και **πόσο** ο καθένας από αυτούς.
- Γενική μορφή: $y= a + b_1X_1 + b_2x_2 + \dots b_nx_n$
- Προϋποθέσεις:
  - Εξαρτημένη μεταβλητή: Αριθμητική
  - Ανεξάρτητες μεταβλητές: Αριθμητικές, Ποιοτικές, Κατηγορικές
  - Αριθμός ανεξάρτητων μεταβλητών προς μέγεθος δείγματος
  - Μη πολυσυγγραμμικότητα (multicollinearity)

Προϋποθέσεις ΠΠ (1/2)

- Κωδικοποίηση ποιοτικών και κατηγορικών μεταβλητών: Για δεδομένα ποιοτικής ή κατηγορικής μεταβλητής κωδικοποιούμε τις διάφορες κατηγορίες με αριθμούς, π.χ. Άνδρες=1, Γυναίκες=2, Πολύ=3, Αρκετά=2, Λίγο=1 κ.ά. Οι αριθμητές μεταβλητές που σχηματίζονται ονομάζονται dummy variables αφού στην ουσία είναι κατηγορικές με αριθμητική κωδικοποίηση.
- Δείγμα προς αν.μεταβλητές:
  - Ελάχιστο: 5/1
  - Επιθυμητό: 20/1
  - Σε ΠΠ με μέθοδο stepwise: 50/1

Προϋποθέσεις ΠΠ (2/2)

- Πολυσυγγραμικότητα: Το φαινόμενο της συσχέτισης των ανεξάρτητων μεταβλητών μεταξύ τους
  - Διάγνωστικά: Tolerance > 0.10
  - Μειώνει τη δυνατότητα ερμηνείας της συμβολής των ανεξ. μεταβλητών στην εξαρτημένη.
  - Κάνει τον υπολογισμό των συντελεστών μη αξιόπιστο.
  - Αντιμετώπιση: διαγραφή μιας από τις αλληλοσχετιζόμενες μεταβλητές ή χρησιμοποίηση της ΠΠ για προβλέψεις μόνο.

Λογιστική Παλινδρόμηση – ΛΠ (Logistic Regression)

- Η ΠΠ δεν μπορεί να χρησιμοποιηθεί όταν η εξαρτημένη μεταβλητή είναι κατηγορική. Στην περίπτωση αυτή χρησιμοποιούμε ένα ειδικό είδος παλινδρόμησης, την Λογιστική Παλινδρόμηση (ΛΠ)
- Η ΛΠ περιλαμβάνει μια δίτιμη εξαρτημένη κατηγορική μεταβλητή και μια σειρά από ανεξάρτητες μεταβλητής μεικτής φύσης. Το σημαντικότερο πλεονέκτημά της είναι ότι είναι ανθεκτική σε παραβιάσεις κανονικότητας των δεδομένων γεγονός που την καθιστά πολύ σημαντική για την ανάλυση γλωσσικών δεδομένων.
- Ο λόγος δείγματος προς μεταβλητές θα πρέπει τουλάχιστον να φτάνει το 50/1

Αξιολόγηση μοντέλου στη ΛΠ

- Η εξίσωση της ΛΠ είναι ο φυσικός λογάριθμος της πιθανότητας μια περίπτωση να ανήκει σε μια κατηγορία προς την πιθανότητα να ανήκει στην άλλη

- Το μοντέλο που δημιουργείται με τη χρήση ανεξ.μετβλ. συγκρίνεται ως προς το βασικό μοντέλο δίχως μεταβλητές χρησιμοποιώντας την log likelihood (-2LL). Η στατιστική σημαντικότητα κρίνεται με το χ^2.

- Η ερμηνεία των αποτελεσμάτων γίνεται με τρόπο παρόμοιο με την ΠΠ. Ωστόσο δεν υπάρχει ακριβής αντιστοίχιση για τις τιμές Β. Αν και το πρόσημό τους αποκαλύπτει το είδος της σχέσης τους με την εξαρτημένη μεταβλητή. Για μια περισσότερο απλή ερμηνεία των αποτελεσμάτων της ΛΠ χρησιμοποιούμε το Odds ratio (λόγο πιθανοτήτων).

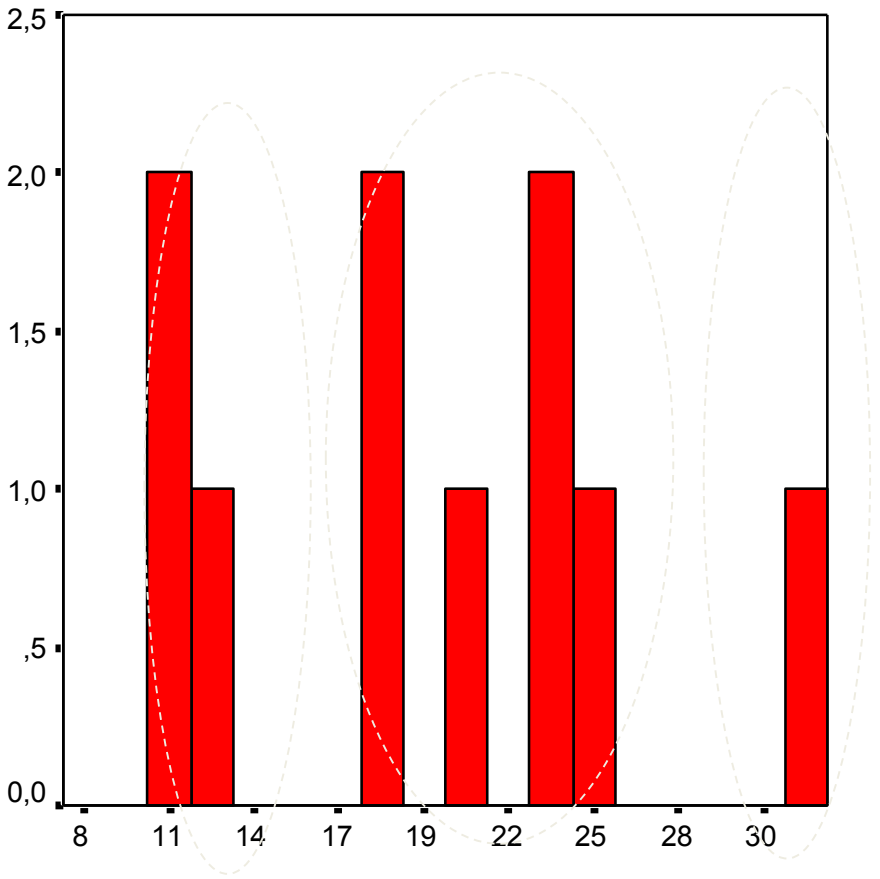
- Π.χ. Έστω ότι ερευνάται η επίδραση της ηλικίας ενός παιδιού στην χρήση προερρινοποίησης. Αν ο λόγος πιθανοτήτων της ανεξ. μεταβλ. βρεθεί να είναι 14,2, τότε για κάθε μονάδα αύξησης της ανεξάρτητης μεταβλητής αυξάνεται κατά 14 φορές η πιθανότητα ο ομιλητής να χρησιμοποιεί προερρινοποίηση.

Ανάλυση Συστάδων (cluster analysis)

- Η ΑΣ κατηγοριοποιεί ένα πλήθος παρατηρήσεων σε δύο ή περισσότερες αμοιβαία αποκλειόμενες ομάδες στηριζόμενη σε συνδυασμούς αριθμητικών μεταβλητών. Ο σκοπός της ΑΣ είναι να εντοπίσει ένα σύστημα που οργανώνει τις παρατηρήσεις σε ομάδες.
- Για παράδειγμα θα μπορούσαμε να διερευνήσουμε το κατά πόσο κάποιοι υφομετρικοί δείκτες (type/token ratio, μέσο μήκος λέξης, μέσο μήκος πρότασης κ.ά.) θα μπορούσαν να διακρίνουν μια σειρά από κείμενα και να τα κατατάξουν θεματικά.
- Μια σημαντική ιδιότητα της ΑΣ είναι ότι κατηγοριοποιεί τις παρατηρήσεις σε άγνωστες ομάδες.

Μια απλή ΑΣ

- Σε περιπτώσεις με μια ή δύο μεταβλητές μια απλή επισκόπηση των δεδομένων χρησιμοποιώντας ιστόγραμμα συχνότητας ή διάγραμμα διασποράς είναι αρκετή για να διαμορφώσουμε μια άποψη για τις δυνατές ομαδοποιήσεις.
- Στην περίπτωση αυτή η διάκριση σε ομάδες των κειμένων βάση της μέτρησης της λεξιλογικής πυκνότητας είναι σχεδόν προφανής.



Ο πίνακας εγγύτητας (proximities matrix)

Η ΑΣ έχει ως αφετηρία με έναν πίνακα δεδομένων όπου τα δείγματα (συνήθως άνθρωποι στις κοινωνικές επιστήμες) είναι σειρές και οι παρατηρήσεις κωδικοποιούνται ως στήλες. Από την αρχή ο πίνακας που δημιουργείται περιλαμβάνει τιμές που είναι μετρήσεις εγγύτητας ή διαφοροποιήσεως μεταξύ δύο παρατηρήσεων.

Κειμενικό θέμα	***Λεξιλογική πυκνότητα***
Οικονομικά	11
Πολιτικά	11
Ανθρωπιστικά	13
Νομικά	18

	Οικο-νομικά	***Πολιτι-κά***	***Ανθρω-πιστικά***	***Νομι-κά***
Οικονομικά				
Πολιτικά				
Ανθρωπιστικά				
Νομικά				

Υπολογίζοντας αποστάσεις

- Τα δεδομένα του πίνακα θα περιγραφούν χρησιμοποιώντας το γράμμα «Α». Η απόσταση γράφεται ως δείκτης στο Α. Έτσι η A_{34} περιγράφει την τομή των Ανθρωπιστικών και των Νομικών κειμένων.
- Η απόσταση υπολογίζεται με την απόλυτη τιμή της διαφοράς των δύο κειμένων. Για παράδειγμα η A_{34}, μεταξύ ανθρωπιστικών και νομικών κειμένων θα είναι |13-18| ή 5. Αν συμπληρώσουμε τον πίνακα εγγύτητας με τον τρόπο αυτό θα έχουμε τον διπλανό πίνακα.
- Ένας δεύτερος τρόπος υπολογισμού του πίνακα εγγύτητας είναι η χρήση των τετραγωνισμένων διαφορών. Π.χ. η απόσταση A_{34} θα γινόταν $(13-18)^2$ ή 25. Η συγκεκριμένη μέτρηση έχει το πλεονέκτημα ότι είναι συναφής με πολλές άλλες στατιστικές μετρήσεις όπως είναι η διακύμανση.

	Οικο-νομικά	***Πολιτι-κά***	***Ανθρω-πιστικά***	***Νομι-κά***
Οικονομικά	0	0	2	7
Πολιτικά	0	0	2	7
Ανθρωπιστικά	2	2	0	5
Νομικά	7	7	5	0

	Οικο-νομικά	***Πολιτι-κά***	***Ανθρω-πιστικά***	***Νομι-κά***
Οικονομικά	0	0	4	49
Πολιτικά	0	0	4	49
Ανθρωπιστικά	4	4	0	25
Νομικά	49	49	25	0

Πολυπαραγοντικές αποστάσεις

Όταν για κάθε δείγμα έχουμε παραπάνω από μια μετρήσεις τότε θα πρέπει να βρεθεί ένας τρόπος για να συνδυαστούν σε έναν πίνακα οι επιμέρους πίνακες εγγύτητας που δημιουργούνται για κάθε μέτρηση. Συνήθως οι επιμέρους πίνακες αθροίζονται σε έναν όπως στο διπλανό παράδειγμα:

Type/token ratio	***Οικο-νομικά***	***Πολιτι-κά***	***Ανθρω-πιστικά***	***Νομι-κά***
Οικονομικά	0	64	81	100
Πολιτικά	64	0	1	4
Ανθρωπιστικά	81	1	0	1
Νομικά	100	4	1	0

+

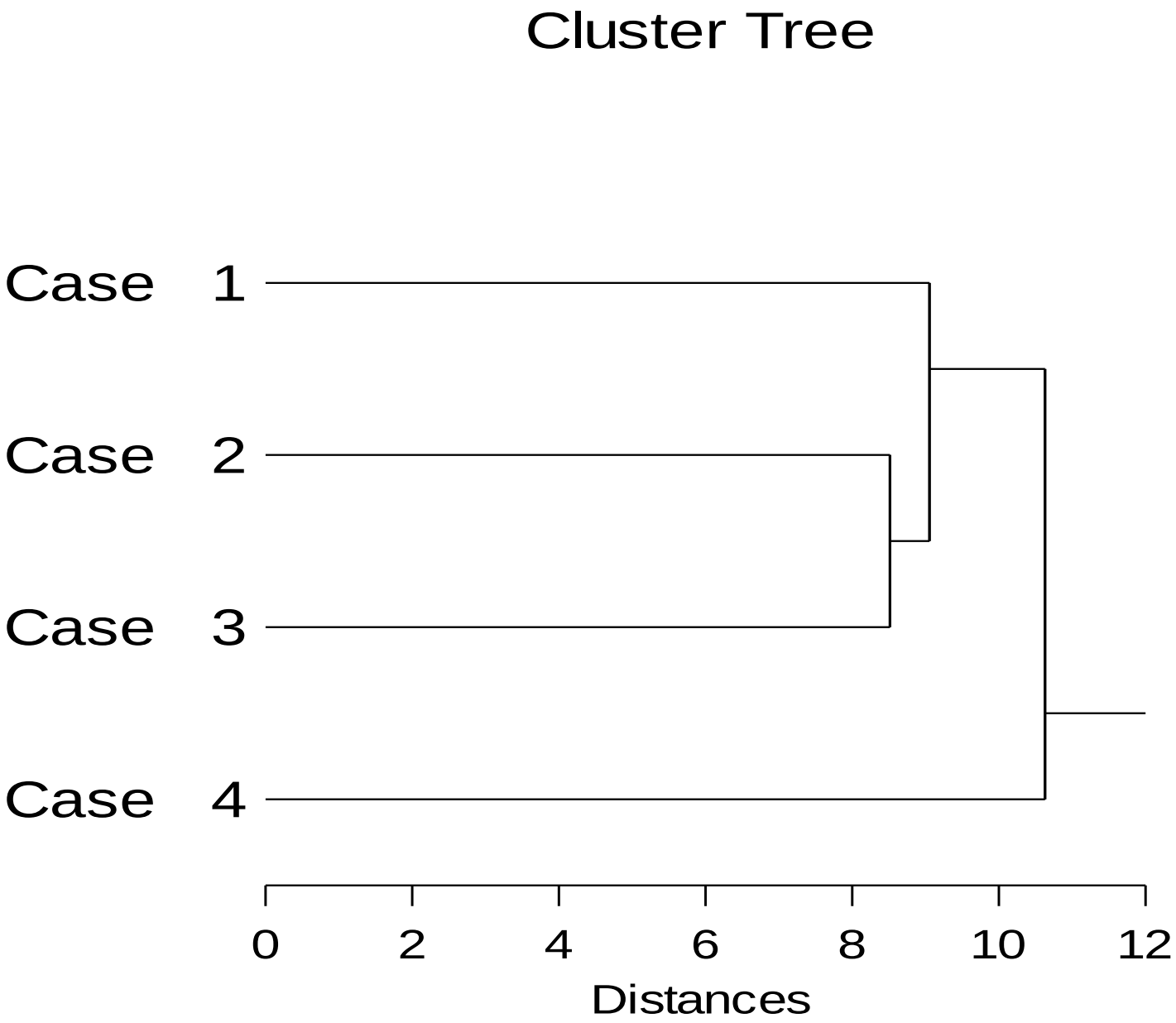
Λεξική πυκνότητα	***Οικο-νομικά***	***Πολιτι-κά***	***Ανθρω-πιστικά***	***Νομι-κά***
Οικονομικά	0	0	4	49
Πολιτικά	0	0	4	49
Ανθρωπιστικά	4	4	0	25
Νομικά	49	49	25	0

=

Σύνολο	***Οικο-νομικά***	***Πολιτι-κά***	***Ανθρω-πιστικά***	***Νομι-κά***
Οικονομικά	0	64	85	149
Πολιτικά	64	0	5	53
Ανθρωπιστικά	85	5	0	26
Νομικά	149	53	26	0

Η χρήση των αποστάσεων για την ομαδοποίηση των δειγμάτων

- Το επόμενο στάδιο μετά την μέτρηση των αποστάσεων είναι η διάκριση των δειγμάτων σε ομάδες βάσει των αποστάσεών τους.
- Αν ο αριθμός των ομάδων είναι γνωστός από πριν χρησιμοποιείται μια «επίπεδη» μέθοδος. Βάσει αυτής τα δείγματα αποδίδονται σε κάποια ομάδα στηριζόμενα σε κάποιο αρχικό κριτήριο. Υπολογίζεται ο μέσος όρος για κάθε ομάδα. Εν συνεχεία ανακατάσσονται τα δείγματα σε ομάδες βάσει τις ομοιότητας του δείγματος στο μέσο όρο της ομάδας. Αυτή η διαδικασία επαναλαμβάνεται αναδρομικά μέχρι όλα τα δείγματα να συμμετάσχουν σε κάποια ομάδα. *Αυτή η μέθοδος ονομάζεται και «k-means cluster analysis».*
- *Οι μέθοδοι ιεραρχικής συσταδοποίησης (hierarchical clustering methods) δεν απαιτούν προηγούμενη γνώση του αριθμού των ομάδων. Οι βασικότερες μέθοδοι είναι η διαιρετική (divisive) και η συσσωρευτική (agglomerative).*
  - *Οι διαιρετικές τεχνικές ξεκινούν προϋποθέτοντας μια ομάδα την οποία την διαιρούν σε υποομάδες συνεχόμενα μέχρι το κάθε δείγμα να αποτελεί το τελικό κλαδί μιας υποομάδες. Οι συσσωρευτικές τεχνικές ξεκινούν από κάθε δείγμα το οποίο περιγράφει μια υποομάδα και με συνεχόμενες συγχωνεύσεις φτάνουμε σε μια ομάδα.*
  - *Και στις δύο περιπτώσεις οι σχετικές τεχνικές περιγράφονται με δενδρόγραμμα ή δίτιμο δένδρο (binary tree). Τα δείγματα εμφανίζονται ως τελικοί κόμβοι στο δενδρόγραμμα, ενώ το μήκος των κλάδων δείχνει την απόσταση μεταξύ των υποομάδων που ενώνονται.*

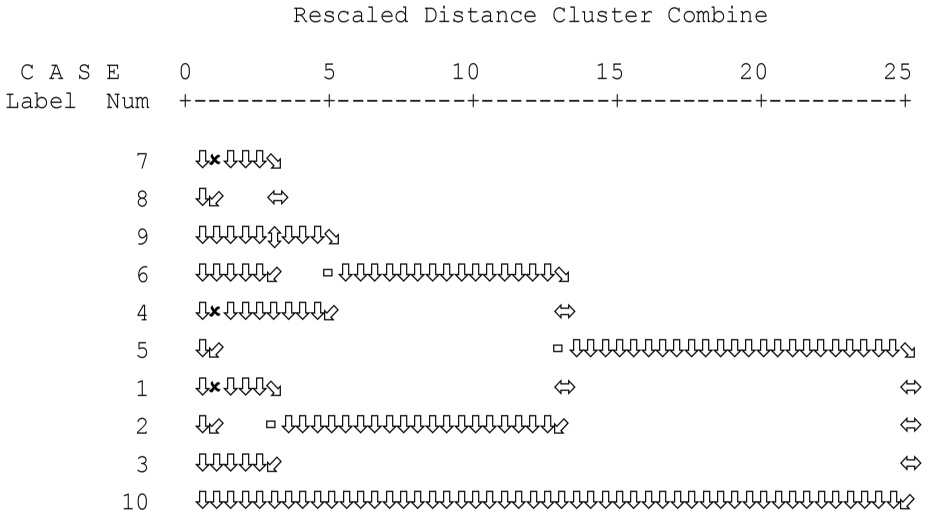

Cluster Tree
Case 1
Case 2
Case 3
Case 4
0
2
4
6
8
10
12
Distances

Μέθοδοι συσταδοποίησης

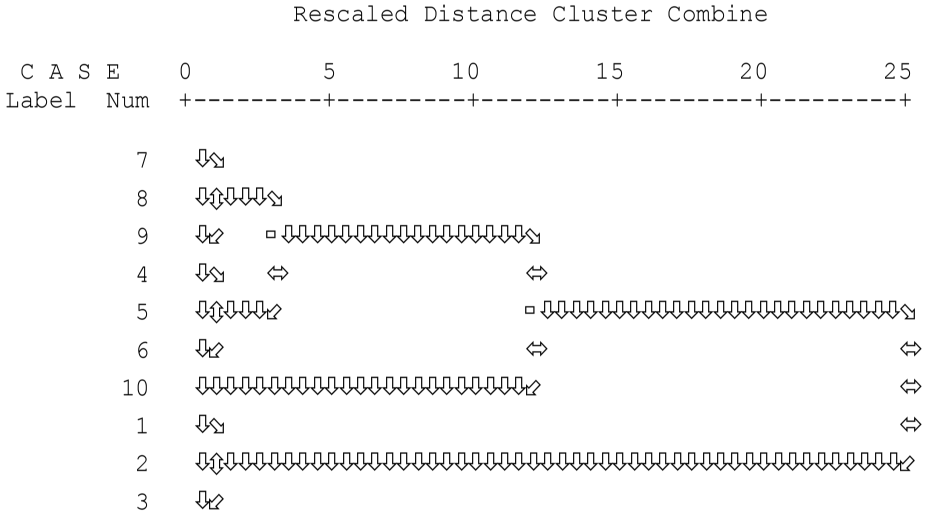
- Απλή διασύνδεση (simple linkage): (Nearest neighbour in SPSS/WIN) υπολογίζει την απόσταση μεταξύ των δύο υποομάδων ως την ελάχιστη απόσταση μεταξύ δύο μελών των αντίθετων ομάδων.
- Πλήρη διασύνδεση (complete linkage): (Furthest neighbour in SPSS/WIN) υπολογίζει την απόσταση ανάμεσα στις δύο υποομάδες ως την μέγιστη απόσταση μεταξύ οποιωνδήποτε μελών στις υποομάδες.
- Μέση διασύνδεση (average linkage): (Centroid Method in SPSS/WIN) υπολογίζει την απόσταση ανάμεσα στις υποομάδες ως τον μέσο όρο μεταξύ των δύο υποομάδων.

Συγκριση μεθόδων ομαδοποίησης

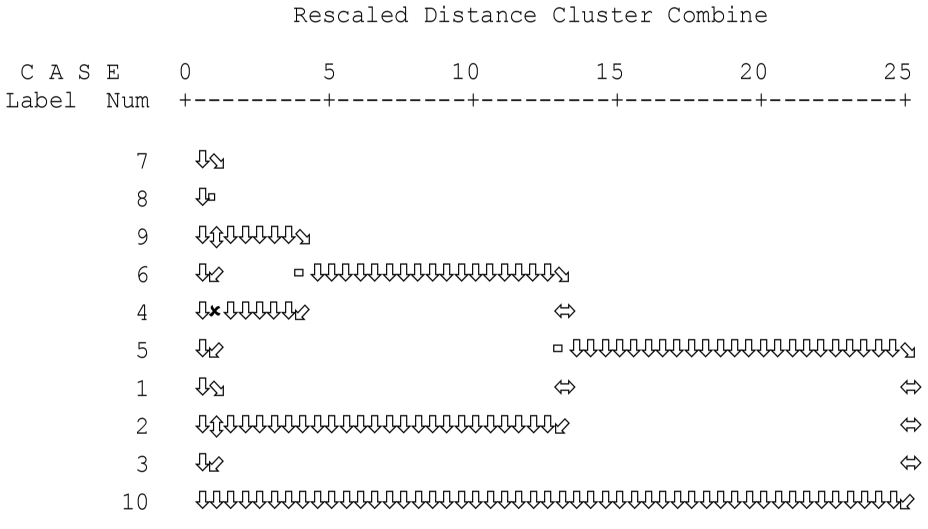
Single Linkage



Complete Linkage



Centroid Method

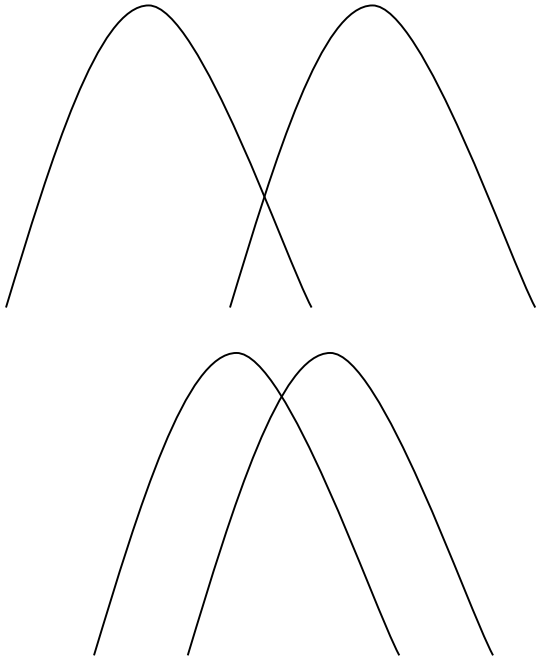


Διακριτική Ανάλυση (Discriminant Function Analysis)

- Η ΔΑ χρησιμοποιείται για να εξηγηθεί ή και προβλεφθεί η κατηγοριοποίηση σε δύο ή περισσότερες κατηγορίες μιας σειράς παρατηρήσεων δεδομένων κάποιων χαρακτηριστικών τους.
- Π.χ. μπορούμε να χρησιμοποιήσουμε ΔΑ για να προβλέψουμε την κατηγοριοποίηση μιας σειράς κειμένων βάσει κάποιων χαρ/κών τους.
- Απαραίτητη προϋπόθεση είναι η εξαρτημένη μεταβλητή να είναι κατηγορική και οι ανεξάρτητες να είναι αριθμητικές. Η ΔΑ είναι γραμμικός συνδυασμός που αντιπροσωπεύει το σταθμισμένο άθροισμα δύο ή περισσότερων ανεξάρτητων μεταβλητών.
- $Z_{jk} = a + W_1X_{1k} + W_2X_{2k} + \ldots + W_nX_{nk}$
  - Z_{jk}= Διακριτική τιμή Z της διακριτικής συνάρτησης j για το αντικείμενο k
  - α= Σταθερά
  - W_i= Διακριτικό βάρος για την ανεξάρτητη μεταβλητή i
  - X_{ik}= Ανεξάρτητη μεταβλητή i για το αντικείμενο k

Υπολογισμός της ΔΑ

Η ΔΑ είναι η κατάλληλη μέθοδος για να ελεγχθεί η υπόθεση ότι οι μέσοι όροι δύο ή περισσότερων κατηγοριών για δύο ή περισσότερες ανεξάρτητες μεταβλητές είναι ίδιοι. Στη ΔΑ κάθε ανεξάρτητη μεταβλητή πολλαπλασιάζεται με το σχετικό «βάρος» της και τα γινόμενα αθροίζονται. Το αποτέλεσμα είναι μια σύνθετη διακριτική Z τιμή (Zδτ) (discriminant Z score). Εν συνεχεία βγάζουμε τον μέσο όρο των Zδτ ανά κατηγορία για όλα τα μέλη. Όταν η ΔΑ γίνεται για δύο ομάδες έχουμε δύο μέσους όρους, για τρεις ομάδες τρεις μέσους όρους κοκ. Όσο μεγαλύτερη απόσταση έχουν οι δύο ή περισσότερες κατανομές των Zδτ τόσο καλύτερη διάκριση επιτυγχάνεται με βάση τις ανεξάρτητες μεταβλητές που χρησιμοποιούνται.



Διακριτικές συναρτήσεις

- Η ΔΑ είναι η μοναδική τεχνική που όταν οι κατηγορίες της εξαρτημένης μεταβλητής ξεπεράσουν τις 2 χρησιμοποιούνται παραπάνω από 1 συναρτήσεις. Ο γενικός κανόνας είναι για Ν κατηγορίες χρησιμοποιούνται Ν-1 συναρτήσεις.
- Κάθε διακριτική συνάρτηση υπολογίζει μια Ζδτ. Στη περίπτωση 3 κατηγοριών κάθε περίπτωση θα έχει δύο Ζδτ από δύο αντίστοιχες διακριτικές συναρτήσεις. Οι δύο αυτές τιμές μπορούν να χρησιμοποιηθούν για να γίνει ένα δισδιάστατο διάγραμμα της κατηγοριοποίησης.
- Η συνάρτηση διάκρισης δεν θα πρέπει να συγχέεται με την συνάρτηση κατάταξης (classification function). Για κάθε κατηγορία μπορεί να υπολογιστεί η συνάρτηση κατάταξης η οποία για κάθε περίπτωση υπολογίζει μια τιμή κατάταξης (classification score). Σε όποια κατηγορία εμφανίζεται η μεγαλύτερη τιμή για την συγκεκριμένη περίπτωση, σε αυτήν κατηγοριοποιείται.

Τέλος Ενότητας

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στο πλαίσιο του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Αθηνών**» έχει χρηματοδοτήσει μόνο την αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Σημειώματα

Σημείωμα Ιστορικού Εκδόσεων Έργου

Το παρόν έργο αποτελεί την έκδοση 1.0.

Σημείωμα Αναφοράς

Σημείωμα Αδειοδότησης

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.

Σημείωμα Χρήσης Έργων Τρίτων

"Η δομή και οργάνωση της παρουσίασης, καθώς και το υπόλοιπο περιεχόμενο, αποτελούν πνευματική ιδιοκτησία του συγγραφέα και του Πανεπιστημίου Αθηνών και διατίθενται με άδεια Creative Commons Αναφορά Μη Εμπορική Χρήση Παρόμοια Διανομή Έκδοση 4.0 ή μεταγενέστερη.

Οι εικόνες/σχήματα/διαγράμματα/φωτογραφίες που περιέχονται στην παρουσίαση αποτελούν πνευματική ιδιοκτησία τρίτων. Απαγορεύεται η αναπαραγωγή, αναδημοσίευση και διάθεσή τους στο κοινό με οποιονδήποτε τρόπο χωρίς τη λήψη άδειας από τους δικαιούχους. "