Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
Τμήμα Επικοινωνίας και ΜΜΕ

Γλωσσική μεταβλητότα
Σημειώσεις στατιστικής

(σχεδόν χωρίς μαθηματικά)

Σ. Α. Μοσχονάς

1. Περιγραφική στατιστική (descriptive statistics)

- ταξινομήσεις δεδομένων
  - πίνακες,
  - διαγράμματα κλπ.
- μαθηματική περιγραφή
  - εύρος τιμών (range),
  - μέση τιμή (average, mean value),
  - διάμεσος (median),
  - κορυφή (mode)
  - διακύμανση (variance),
  - τυπική απόκλιση (standard deviation) κλπ.

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

2. Επαγωγική στατιστική (inferential statistics – στατιστική συμπερασματολογία)

- Δειγματοληπτικός έλεγχος ερευνητικών υποθέσεων βάσει μιας *στατιστικής διαδικασίας*
  1. Πληθυσμός
  2. Δείγμα
  3. Χαρακτηριστικά δείγματος
     - Μεταβλητές
  4. Ποσοτικοποίηση
  5. Κατανομή

Στατιστικές διαδικασίες

Βασικές έννοιες – 1. Πληθυσμός, 2. Δείγμα, 3. Χαρακτηριστικά

1. Πληθυσμός
   - Οποιοδήποτε σύνολο προσώπων ή πραγμάτων· π.χ., μία κοινωνική τάξη, οι πραγματώσεις μιας γλωσσικής μεταβλητής, το σύνολο των συμφωνικών συμπλεγμάτων που ένα άτομο προφέρει κατά τη διάρκεια μιας συνέντευξης, οι χώρες της Ευρώπης κλπ.
2. Δείγμα
   - Ένα *αντιπροσωπευτικό* υποσύνολο του πληθυσμού
3. Χαρακτηριστικά του δείγματος (πληθυσμού)
   - Εξαρτημένες μεταβλητές
   - Ανεξάρτητες μεταβλητές
     - Τι ορίζεται ως εξαρτημένη και τι ως ανεξάρτητη μεταβλητή εξαρτάται απολύτως από τον τρόπο που διατυπώνεται η **ερευνητική υπόθεση**.

Βασικές έννοιες – 3. Μεταβλητές

- *Γλωσσική μεταβλητή* είναι μια γλωσσική μονάδα που θα μπορούσε να πραγματωθεί με περισσότερους του ενός τρόπους.
  - (Γλωσσική μεταβλητή = γλ. μονάδα με «εναλλασσόμενους» τύπους)
- Οι γλωσσικές μεταβλητές συνδέουν:

γλωσσική ποικιλία
(= εξαρτημένη μεταβλητή)

με

κοινωνική ποικιλία (στρωμάτωση / ιδεολογία)
(= ανεξάρτητη μεταβλητή).

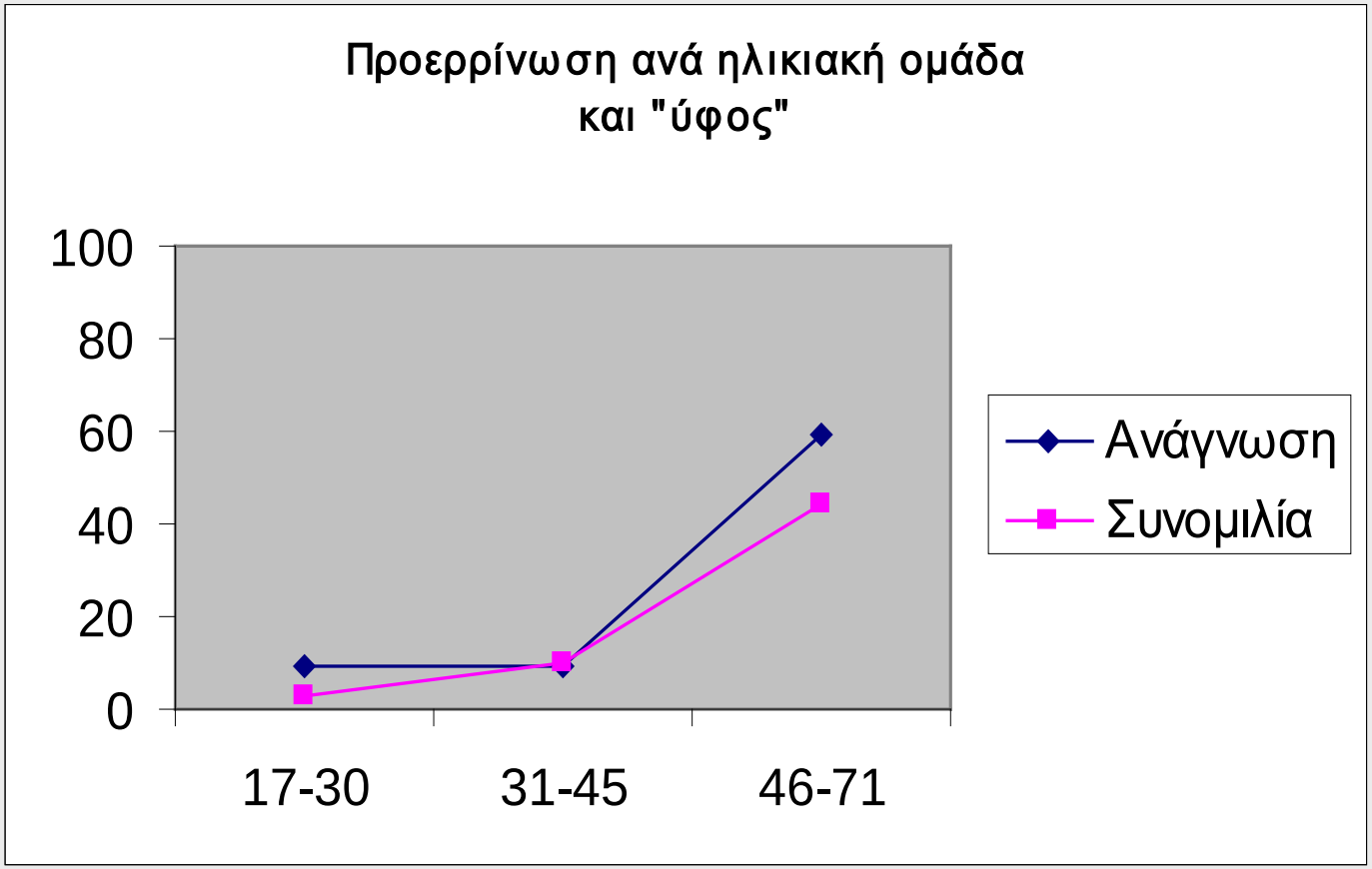
» Παράδειγμα (προερρίνωση κλειστών):

(b, d, g)

[b, d, g] [ᵐb,ⁿd, ᵑg] [mp, nt, ŋk]

Άλλα παραδείγματα: αγαπάω / ώ, αγαπούνε / άνε, αγαπούσα / αγάπαγα, αγαπιόσαν(τ)ε / αγαπιού(ν)ταν / αγαπιό(ν)ταν(ε), γλήγορα / γρήγορα, δανεική / δανεικιά, πατέρες / πατεράδες, έφυγαν / φύγανε, βγέστε / βγείτε, άφησέ τον / άσ' τον [εναλλασσόμενοι τύποι], κτήμα / χτήμα, κτυπώ / χτυπώ, οίνος / κρασί, εφτά / επτά [διαφορετικές πραγματώσεις που οφείλονται στην ύπαρξη διμορφίας], την οδός / την οδό [«λάθη»], …· τα πήρα στο κρανίο, δίνε του, ελλογιμότατε, ρε … [μη εναλλασσόμενων τύπων / ύφους, επιπέδου], ντιμπέιτ / τηλεμαχία [δάνεια] κτλ.

Βασικές έννοιες - 3. Παραδείγματα γλωσσικών μεταβλητών

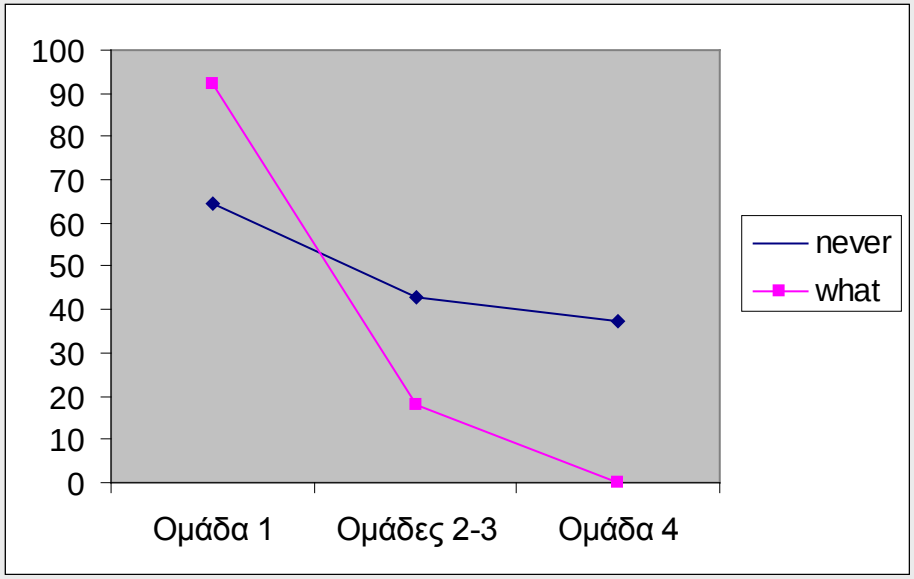


Α. Αρβανίτη, “Sociolinguistic Patterns of Prenasalisation in Greek”

Βασικές έννοιες - 3. Παραδείγματα γλωσσικών μεταβλητών

- Ομάδες:
  - mainstream culture ←→ vernacular culture
- Γλωσσικά χαρακτηριστικά:
  - "We *goes* [αντί για: we go] shopping on Saturdays",
  - "We *has* [αντί για: we have] a little fire",
  - "I'm not going *nowhere*" [αντί για: I'm not going anywhere],
  - "I never *done* it" [αντί για: I did not do it],
    "There's a knob *what* [αντί για: that/which] you turn" κ.ά.

J. Cheshire, "Linguistic Variation and Social Function"



Βασικές έννοιες - 3. Παραδείγματα γλωσσικών μεταβλητών

4. Ποσοτικοποίηση

- Ονομαστικά (nameable) χαρακτηριστικά
  - *Απαρίθμηση*: αρίθμηση των μελών του πληθυσμού ή του δείγματος
- Μετρήσιμα (measurable) χαρακτηριστικά
  - *Ποσόστωση*: μέτρηση του κατά πόσο ένα μέλος του δείγματος εκδηλώνει ένα χαρακτηριστικό

- Εργαλείο μέτρησης
- Μονάδα μέτρησης
- Κλίμακα μέτρησης

» Κλίμακες μέτρησης

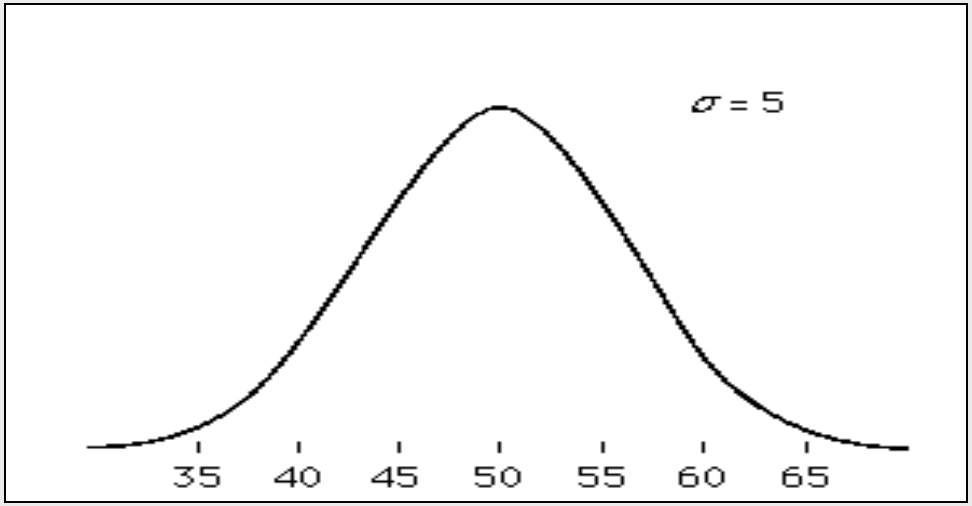
- Ονομαστική /κατηγορική κλίμακα (nominal scale)
- Τακτική / διατεταγμένη (ordinal scale)
- Διαστημική / διαστημάτων (interval scale)
- Αναλογική (ratio scale)

Βασικές έννοιες – 4. Ποσοτικοποίηση

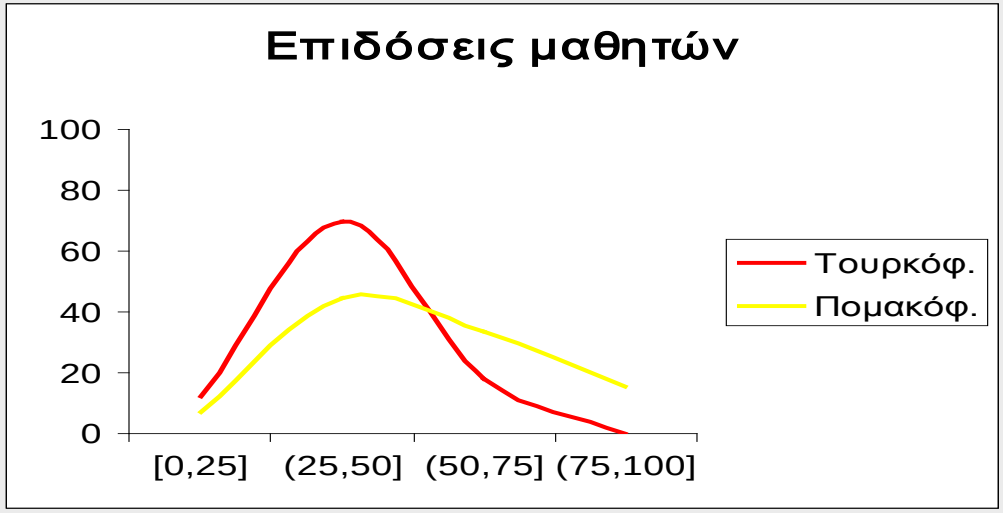
Ταυτότητα	Ονομαστική	Τακτική	Διαστημάτων	Αναλογική
	φύλο			
Μέγεθος	*οι 3 πρώτοι σε αγώνα*			
Ίσα διαστήματα	*C, F*			
Min. μηδέν	*βάρος, χρόνος*			

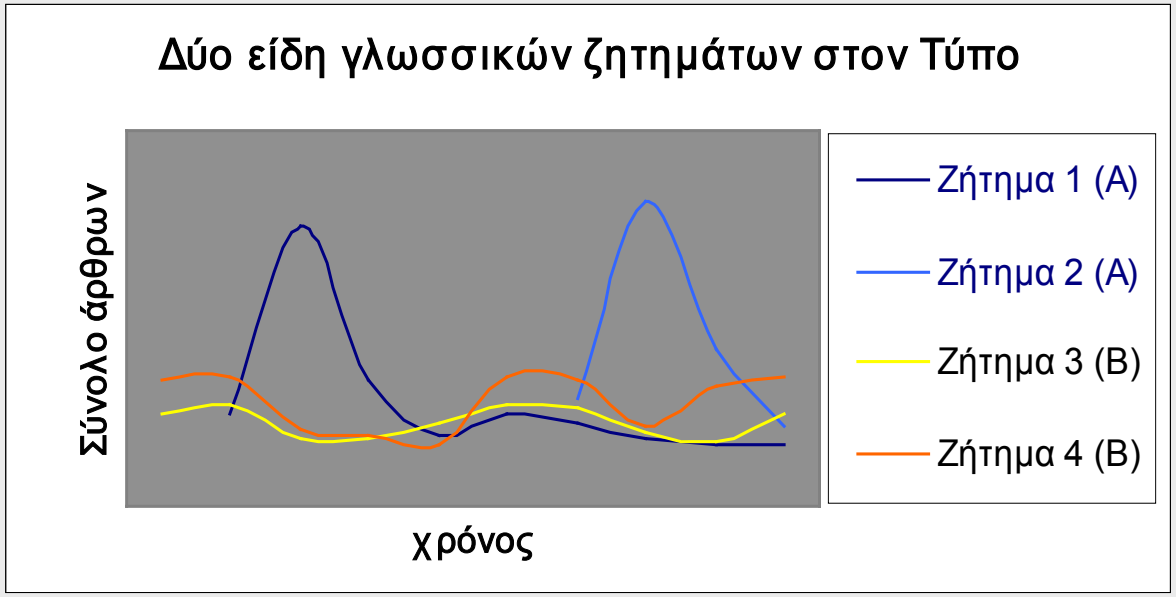
5. Κατανομή

- κανονική κατανομή (normal distribution)



- ασύμμετρη κατανομή (skewed distribution)


Επιδόσεις μαθητών
100
80
60
40
20
0
[0,25]
(25,50]
(50,75]
(75,100]
Τουρκόφ.
Πομακόφ.


Δύο είδη γλωσσικών ζητημάτων στον Τύπο
Σύνολο άρθρων
χρόνος
Ζήτημα 1 (Α)
Ζήτημα 2 (Α)
Ζήτημα 3 (Β)
Ζήτημα 4 (Β)

Έλεγχος υποθέσεων (hypothesis testing)

- ερευνητική /εναλλακτική υπόθεση (research / alternative hypothesis, H_A)
  - υπάρχει ιδιαίτερη, στατιστικά σημαντική σχέση μεταξύ ποσοτικοποιήσιμων χαρακτηριστικών ενός πληθυσμού
- μηδενική υπόθεση (null hypothesis, H_0)
  - δεν υπάρχει ιδιαίτερη σχέση μεταξύ των χαρακτηριστικών

- Στις στατιστικές διαδικασίες δεν αποδεικνύουμε την ερευνητική υπόθεση απευθείας αλλά έμμεσα, έχοντας απορρίψει τη μηδενική υπόθεση. Η ερευνητική υπόθεση συνάγεται ως λογική συνέπεια της άρνησης της μηδενικής υπόθεσης.
  - ακρίβεια του ελέγχου
  - η υπόθεση πρέπει να είναι συμβατή με τα διαθέσιμα δεδομένα

Στατιστικές διαδικασίες

- Από τη στατιστική διαδικασία προκύπτει μια αριθμητική τιμή, που συγκρίνεται με τις τιμές σε μια γνωστή θεωρητική κατανομή.

  Π.χ., το t-test δίνει μια τιμή t που μπορεί να συγκριθεί με τη θεωρητική κατανομή των τιμών t (Student's distribution). Αν η τιμή που προέκυψε από τη στατιστική διαδικασία είναι μικρότερη από την κρίσιμη τιμή t, τότε μάλλον έχει προκύψει κατά τύχη. Αν είναι μεγαλύτερη από την τιμή αυτή, τότε *κατά πάσα πιθανότητα* δεν προέκυψε τυχαία.

- Πόσο πρέπει να απέχει από τη μέση τιμή η αριθμητική τιμή που προκύπτει από μια στατιστική διαδικασία; Ποια είναι η **στατιστική σημαντικότητα** των συμπερασμάτων; Συνήθως θεωρείται ικανοποιητική **p<0,05** (5% πιθανότητα λάθους) ή **p<0,01** (1% πιθανότητα λάθους).

» Η συσχέτιση δύο μεταβλητών (ανεξάρτητης-εξαρτημένης) δεν σημαίνει απαραιτήτως ότι υπάρχει *αιτιακή* σχέση μεταξύ τους.

> «Κι όμως, το Ίντερνετ επικοινωνεί με την ανάγνωση! Σύμφωνα με την έρευνα [VPRC-ΕΚΕΒΙ], βασικός παράγων της αναγνωστικής συμπεριφοράς των πολιτών μέσης και κατώτερης εκπαιδευτικής βαθμίδας είναι η σχέση του με τις νέες τεχνολογίες. "Όσο πιο εξοικειωμένοι είναι με τη χρήση τους (υπολογιστές, Ίντερνετ, ΑΤΜ, κινητό) τόσο περισσότερο διαβάζουν", επισημαίνει η έρευνα.» (Ό. Σελλά, «Δεν διαβάζω γιατί … δεν έχω χρόνο», *Η Καθημερινή/Τέχνες και Γράμματα*, 4/12/2005, σ. 1)

Ποια στατιστική διαδικασία ταιριάζει στα δεδομένα μου;

TABLE 12.4 SELECTING A POSSIBLE STATISTICAL TEST BY NUMBER OF INDEPENDENT VARIABLES AND LEVEL OF MEASUREMENT

Level of Measurement of Dependent Variable	*ONE INDEPENDENT VARIABLE*				*TWO INDEPENDENT VARIABLES*	
	Two Treatments		*More Than Two*		*Factorial Designs*	
	Two Independent Groups	*Two Matched Groups (or Within-Subjects)*	*Multiple Independent Groups*	*Multiple Matched Groups (or Within-Subjects)*	*Independent Groups*	*Matched Groups (or Within-Subjects)*
Interval or ratio	*t* test for independent groups	*t* test for matched groups	One-way ANOVA (randomized)	One-way ANOVA (repeated-measures)	Two-way ANOVA	Two-way ANOVA (repeated-measures)
Ordinal	Mann-Whitney U test	Wilcoxon test	Kruskal–Wallis test	Friedman test	—	
Nominal	Chi square test	—	Chi square test	—	—	—

Μη παραμετρικές διαδικασίες

Ονομαστικές ή τακτικές κλίμακες

στατιστικοί έλεγχοι

χ^2

r (Spearman)

Παραμετρικές διαδικασίες

Διαστημικές ή αναλογικές κλίμακες

t-test

ANOVA

r (Pearson)

χ^2

» *Μη παραμετρική* στατιστική διαδικασία που συνήθως χρησιμοποιείται για να ελεγχθεί η τυχόν αλληλεξάρτηση στις κατανομές δύο *ονομαστικών* χαρακτηριστικών ενός πληθυσμού ή δείγματος.

Πότε χρησιμοποιείται	Υπόθεση
Δεδομένα από έναν *πληθυσμό ή από* ένα *δείγμα* *Δεδομένα για* δύο *τουλάχιστον* ονομαστικά *(μη μετρήσιμα) χαρακτηριστικά* *Δύο* αλληλοαποκλειόμενες *τιμές*	*Ελέγχεται εάν τα δύο χαρακτηριστικά λειτουργούν από κοινού: εάν τα μέλη ενός πληθυσμού που έχουν (ή δεν έχουν) μια συγκεκριμένη τιμή για ένα χαρακτηριστικό, επίσης έχουν (ή δεν έχουν) μια συγκεκριμένη τιμή και για το άλλο.*

χ^2 – Παράδειγμα

Fishman (1966)

	Γλωσσικός παράγοντας	
Ακαθάριστο εθνικό προϊόν	*Ομοιογένεια*	*Ανομοιογένεια*
Πολύ Υψηλό / Μέσο	27	15
Χαμηλό / Πολύ χαμηλό	25	47

42 [=27+15] X 52 [=27+25] = 2.184, 2.184 / 114 [= 27+15+25+47] = 19,2 κ.ο.κ.

	Γλωσσικός παράγοντας	
Ακαθάριστο εθνικό προϊόν	*Ομοιογένεια*	*Ανομοιογένεια*
Πολύ Υψηλό / Μέσο	19,2	22,8
Χαμηλό / Πολύ χαμηλό	32,8	39,2

$\chi^2 = (27-0,5-19,2)^2/19,2 + (25+0,5-32,8)^2/32.8 + \ldots = 8,19$

$\chi^2 = 8,19, p<0,005$

t-test

» *Παραμετρική* στατιστική διαδικασία που ελέγχει εάν οι μέσοι όροι των τιμών που προέρχονται από *δύο* δείγματα διαφέρουν σημαντικά μεταξύ τους.

Πότε χρησιμοποιείται	Υπόθεση
Δεδομένα από δύο *δείγματα ή* δύο *υποσύνολα δείγματος* *Το δείγμα ορίζεται από* ονομαστικό *ή τακτικό χαρακτηριστικό* *Εξετάζονται δεδομένα για το ίδιο* μετρήσιμο *χαρακτηριστικό*	*Ελέγχεται εάν τα δύο δείγματα προέρχονται από τον ίδιο πληθυσμό σε σχέση με το χαρακτηριστικό του οποίου γίνεται μέτρηση (εάν δηλαδή τα δύο δείγματα διαφέρουν σημαντικά σε σχέση με το χαρακτηριστικό αυτό).*

t-test – Παράδειγμα

Milroy (1980)	Μέση συχνότητα % [ɔ̈] (hat)
Άντρες	52,0
Γυναίκες	34,7

Η τιμή t είναι ο λόγος των παρατηρούμενων διαφορών στους μέσους όρους και μιας μέτρησης που ονομάζεται *τυπικό σφάλμα στη διαφορά μεταξύ των μέσων όρων* (standard error of the difference between means), η οποία μετράει τη διαφορά που θα προέκυπτε μεταξύ των μέσων όρων αν υπεισέρχονταν μόνο τυχαίοι παράγοντες. Η τιμή t υπολογίζεται διαιρώντας τη διαφορά μεταξύ των μέσων όρων (52,0 - 34,7 = 17,3) με το τυπικό σφάλμα. Στην ερμηνεία της τιμής t συνυπολογίζονται επίσης οι «βαθμοί ελευθερίας» (degrees of freedom).

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

t=3,06, p<0,01

(η διαφορά που παρατηρήθηκε είναι, κατά τρεις φορές περίπου, μεγαλύτερη από την αναμενόμενη τυχαία διαφορά μεταξύ των μέσων όρων)

Ανάλυση διακύμανσης (ANOVA)

» *Παραμετρική* στατιστική διαδικασία που ελέγχει τη διαφορά στις μέσες τιμές τριών ή περισσότερων ομάδων. Δύο είδη αναλύσεων: α) *μονοδιάστατος σχεδιασμός* (one-way design)· β) *παραγοντικός σχεδιασμός* (factorial design), συνήθως *δισδιάστατος* (two-way design).

Πότε χρησιμοποιείται

• *Μονοδιάστατος σχεδιασμός*: παρόμοιες συνθήκες με το t-test, αλλά εφαρμόζεται σε οποιοδήποτε αριθμό δειγμάτων εφόσον αντιπροσωπεύουν διαφορετικά *επίπεδα* του ίδιου γενικού χαρακτηριστικού

• *Παραγοντικός σχεδιασμός*: περισσότερα του ενός ονομαστικά χαρακτηριστικά χρησιμοποιούνται για τον ορισμό των ομάδων

Υπόθεση

Ελέγχουμε τη διαφορά στις μέσες τιμές για να εξακριβώσουμε εάν οι μέσες τιμές των ομάδων διαφέρουν.

Τα ονομαστικά ή τακτικά χαρακτηριστικά προσδιορίζουν τις *κύριες επιδράσεις* (main effects)· ο παραγοντικός σχεδιασμός επιτρέπει επίσης να εξακριβώσουμε τυχόν *συνδυασμένες επιδράσεις* (interaction effects) των χαρακτηριστικών.

Ανάλυση διακύμανσης – Υποθετικό παράδειγμα

Προερρίνωση κλειστών συμφώνων			
Ηλικία	Συνομιλία	Κείμενο	Λέξεις
15-45	84	86	82
46-75	86	94	96

τιμές που υπολογίσαμε

			κρίσιμο F	
	F	p	p<0,05	p<0,01
Ηλικία	19,04	p<0,01	5,99	13,74
Ύφος	8,91	p<0,05	5,14	10,92
Ηλικία X Ύφος	1,00		5,14	10,92

τιμές με τις οποίες συγκρίνουμε

Συσχέτιση (Correlation) – r (Pearson)

» *Παραμετρική* στατιστική διαδικασία που ελέγχει κατά πόσο δύο ή περισσότερα χαρακτηριστικά μεταβάλλονται ταυτοχρόνως. Συντελεστής r (Pearsonian r / Pearson product-moment correlation).

Πότε χρησιμοποιείται	Υπόθεση
Όταν έχουμε μόνο *ένα* δείγμα που *δεν* διαιρείται σε υποσύνολα και κάνουμε μετρήσεις για *δύο* χαρακτηριστικά του δείγματος Τα χαρακτηριστικά είναι μετρήσιμα σε *διαστημική* ή *αναλογική* κλίμακα	Υποθέτουμε ότι οι δύο μετρήσεις μεταβάλλονται ταυτοχρόνως (προς την ίδια κατεύθυνση: *θετική συσχέτιση*, σε αντίθετες κατευθύνσεις: *αρνητική συσχέτιση*). Η σημαντικότητα του r ελέγχεται με t-test.

Συσχέτιση (Correlation) – r (Spearman)

- *Μη παραμετρική* στατιστική διαδικασία που ελέγχει κατά πόσο δύο ή περισσότερα χαρακτηριστικά μεταβάλλονται ταυτοχρόνως. Συντελεστής r (Spearman's r / Spearman rank-order correlation).

Πότε χρησιμοποιείται	Υπόθεση
Έχουμε πάλι *ένα* μόνο δείγμα που δεν διαιρείται σε υποσύνολα και κάνουμε μετρήσεις για *δύο* χαρακτηριστικά του ίδιου δείγματος Ένα τουλάχιστον χαρακτηριστικό είναι μετρήσιμο σε *τακτική* κλίμακα μόνο	Υποθέτουμε, όπως προηγουμένως, ότι οι δύο μετρήσεις μεταβάλλονται από κοινού (προς την ίδια κατεύθυνση: *θετική συσχέτιση*, σε αντίθετες κατευθύνσεις: *αρνητική συσχέτιση*). Η σημαντικότητα του ρ ελέγχεται με t-test.

Συντελεστής r (Pearson) – Παράδειγμα

# Διαλόγου	Ποσοστό κατανόησης διαλόγου	Ποσοστό αγγλικών λέξεων
9	79,7	93,9
11	79,1	95,5
17	75,6	94,3
12	74,9	97,8
16	68,9	87,3
13	65,6	93,5
15	62,5	92,0
7	60,2	95,3
9	59,1	87,7
10	58,8	93,6
5	56,0	85,9
4	55,5	91,0
3	53,4	84,6
8	50,4	90,5
6	43,6	85,6
2	34,1	90,6
1	32,5	89,5

Flint (1979): μελέτη κατανόησης της ‘διαλέκτου επαφής’ της νήσου Norfolk

$r = +0,544,\ p<0,05$

r	Ερμηνεία
0,01-0,20	μικρή, αμελητέα συσχέτιση
0,21-0,40	χαμηλή, αν και διακριτή
0,41-0,71	μέτρια, ουσιαστική
0,71-0,90	υψηλή, έκδηλη
0,90-0,99	πολύ υψηλή, αξιόπιστη

Συντελεστής r (Spearman) – Παράδειγμα

Ομιλήτρια	(α)	Ισχύς δικτύου
1	2,78	2
2	2,74	5
3	2,70	5
4	2,63	2
5	2,50	3
6	2,48	3
7	2,42	1
8	2,38	1
9	2,35	4
10	2,33	5
11	2,33	3
12	2,25	2
13	2,16	1
14	2,13	1
15	1,75	1
16	1,73	1
17	1,45	0
18	1,05	0

Milroy (1980): μελέτη επικοινωνιακών δικτύων στο Μπέλφαστ

$r = +0,683,\ p<0,01$

Κλίμακες

φωνητικός δείκτης: 1-5 (συχνότητα χρήσης τοπικής διαλεκτικής πραγμάτωσης)

ισχύς δικτύου: 1-5 (Ισχυροί δεσμοί συγγένειας, Εργάζεται στον ίδιο χώρο με τουλάχιστον άλλα δύο άτομα από την ίδια περιοχή κλπ.)

Αναφορές

A. Αρβανίτη (1995). “Sociolinguistic Patterns of Prenasalisation in Greek”. *Μελέτες για την ελληνική γλώσσα* 15, 209-220.

J. Fishman (1966). “Some Contrasts Between Linguistically Homogeneous and Linguistically Heterogeneous Polities”. *Sociological Inquiry* 36.2, 146-158.

E. H. Flint (1979). “Stable Societal Diglossia in Norfolk Island”, στο *Sociolinguistic Studies in Language Contact: Methods and Cases*, επιμ. W. Mackey & J. Ornstein, 295-334. The Hague: Mouton.

W. Labov (1966). *The Social Stratification of English in New York City*, 2η έκδ. Cambridge: CUP, 2006.

L. Milroy (1980). *Language and Social Networks*. Oxford: Blackwell.

Μ. Τριανταφυλλίδης κ.ά. (1941) *Νεοελληνική Γραμματική (της Δημοτικής)*. Αθήνα.

Βιβλία που χρησιμοποιήθηκαν

Αρμενάκης Αντώνης, *Σημειώσεις Στατιστικής*. Αθήνα: ΕΚΠΑ-ΕΜΜΕ, 1991.

Fasold Ralph, *Introduction to Sociolinguistics*, τ. 1: *The Sociolinguistics of Society*, κεφ. 4 (σσ. 85-112): «Statistics". Oxford: Blackwell, 1984.

Woods Anthony, Paul Fletcher & Arthur Hughes, *Statistics in Language Studies*. Cambridge: Cambridge University Press, 1986.

Τέλος