Μεθοδολογία Έρευνας

Αθανάσιος Σαχλάς

asachlas@uth.gr

ΔΗΛΩΣΗ

- Το παρόν διδακτικό υλικό (διαφάνειες) αναρτάται στο e class του μαθήματος και διατίθεται **αποκλειστικά και μόνο** για εκπαιδευτικούς σκοπούς (καλύτερη κατανόηση και αφομοίωση της ύλης) καθώς και για την προετοιμασία για τις εξετάσεις του μαθήματος.

- Κατά συνέπεια, **απαγορεύεται ρητώς** η με οποιοδήποτε μέσο, η διάδοση, η αντιγραφή (πλήρως ή εν μέρει) και η χρήση του υλικού για άλλους σκοπούς, δίχως την άδεια του διδάσκοντα.

- Εάν κατά την ανάγνωση του υλικού υποπέσει στην αντίληψή σας τυχόν οποιοδήποτε λάθος, θα πρόκειται για λάθος εκ παραδρομής και συνεπώς παρακαλείσθε να ενημερώσετε τον διδάσκοντα προκειμένου να το διορθώσει.

Αναλυτική Φάση

Προαπαιτούμενες γνώσεις

- + Πρόσθεση
- – Αφαίρεση (ειδική περίπτωση της πρόσθεσης)
- × Πολλαπλασιασμός
- ÷ Διαίρεση (ειδική περίπτωση του πολλαπλασιασμού)
- Λογική

Στατιστική και Μαθηματικά

- Η μοναδικότητα της Στατιστικής αλλά και η διαφοροποίησή της, από τα Μαθηματικά, αναδεικνύεται από το γεγονός ότι σχετίζεται με έναν τελείως διαφορετικό τρόπο σκέψης και ενδιαφέρεται για διαφορετικού είδους φαινόμενα.
- Τα Μαθηματικά στηρίζονται σε έναν **ντετερμινιστικό** τρόπο σκέψης.
- Η Στατιστική βασίζεται σε αυτό που καλούμε **στοχαστικό** τρόπο σκέψης.
- Σύμφωνα με τον ***ντετερμινιστικό*** (ή συναρτησιακό) τρόπο σκέψης «η εμφάνιση ενός γεγονότος Α οδηγεί ***με απόλυτη βεβαιότητα*** στο αποτέλεσμα Β».
- Σύμφωνα με το ***στοχαστικό*** τρόπο σκέψης «η εμφάνιση του γεγονότος Α οδηγεί ***με κάποια πιθανότητα*** στο αποτέλεσμα Β».

Η Στατιστική

Η **Στατιστική** είναι η επιστήμη, η οποία ασχολείται με

- τη συλλογή,
- την επεξεργασία
- την παρουσίαση

αριθμητικών ή/και μη αριθμητικών στοιχείων (δεδομένων) με απώτερο σκοπό να εξάγει χρήσιμα συμπεράσματα για τους ευρύτερους πληθυσμούς, στους οποίους ανήκουν τα στοιχεία αυτά.

Στατιστικός τρόπος σκέψης

- Ο προπονητής μιας ομάδας ποδοσφαίρου επιλέγει για την εκτέλεση του πέναλτι, τον παίκτη που έχει το μεγαλύτερο ποσοστό ευστοχίας στα πέναλτι.
- Οι ιατρικοί ερευνητές προκειμένου να αποφασίσουν εάν ένα νέο εμβόλιο είναι ασφαλές, δοκιμάζουν το νέο εμβόλιο σε ομάδες ανθρώπων και εάν το ποσοστό των εμφανιζόμενων παρενεργειών είναι μικρό, το χαρακτηρίζουν ως ασφαλές και το προωθούν προς χρήση.
- Εάν δεν καταναλώσω μεγάλη ποσότητα αλκοόλ κατά τη βραδινή μου έξοδο, έχω περισσότερες πιθανότητες να γυρίσω ασφαλής στο σπίτι με το αυτοκίνητο.
- Ο φοιτητής γνωρίζει ότι εάν διαβάσει αυτά που έχει τονίσει περισσότερο ο καθηγητής, έχει περισσότερες πιθανότητες να περάσει το μάθημα.

Συλλογή δεδομένων

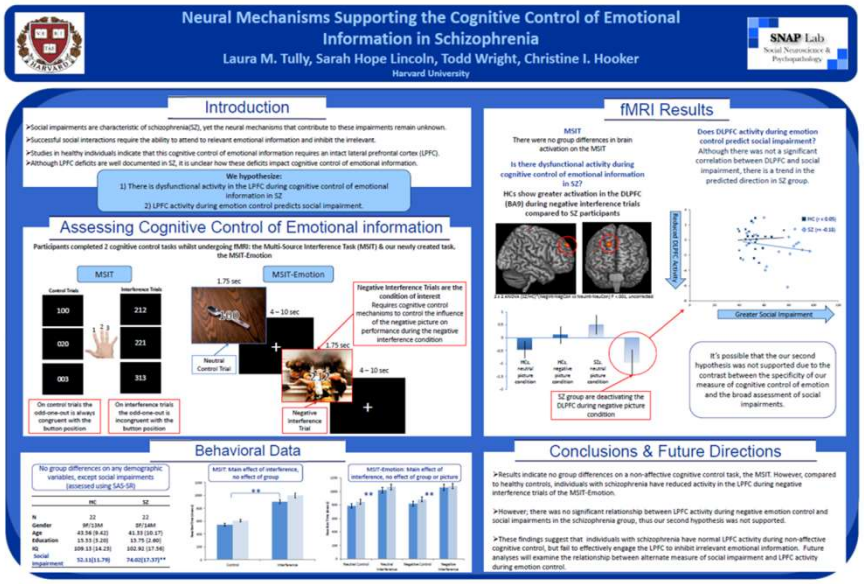
- Απογραφή
- Ιατρικό/Νοσηλευτικό ιστορικό
- Ερωτηματολόγια
- Φυσική Παρατήρηση

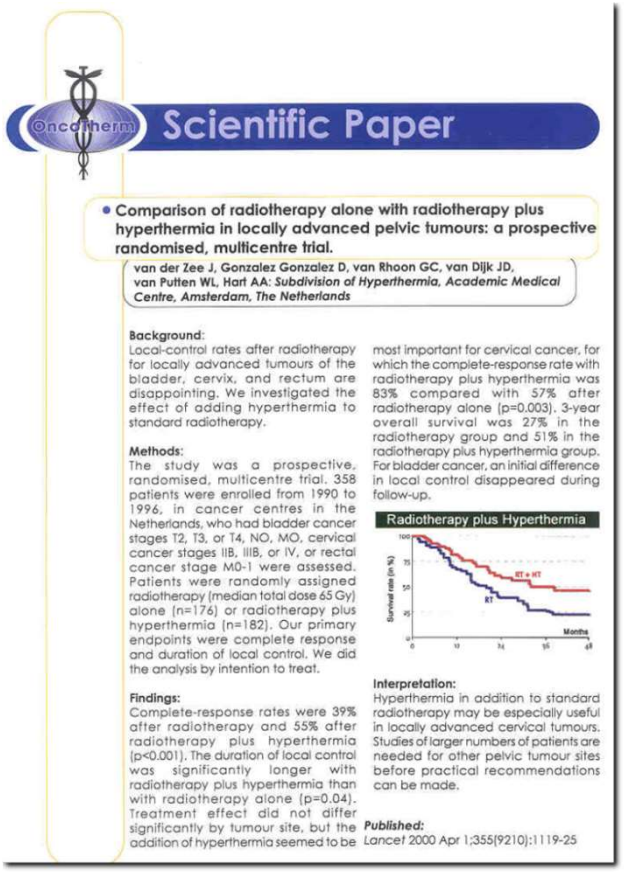
Επεξεργασία δεδομένων

- Με το χέρι
- Με τον υπολογιστή
- Εξειδικευμένα προγράμματα

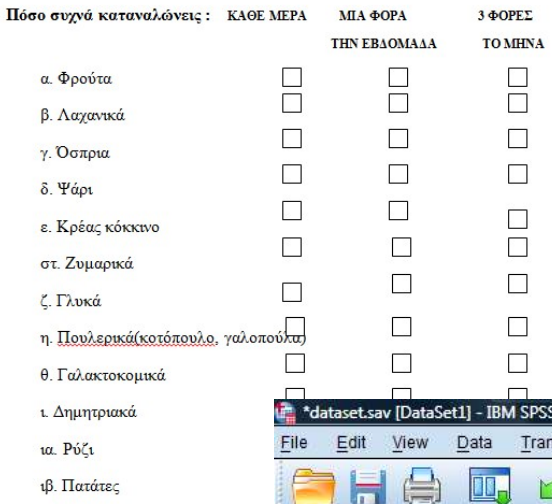
Παρουσίαση δεδομένων

- Συνέδρια
- Συγγραφή άρθρων
- Μεταπτυχιακά/διδακτορικά

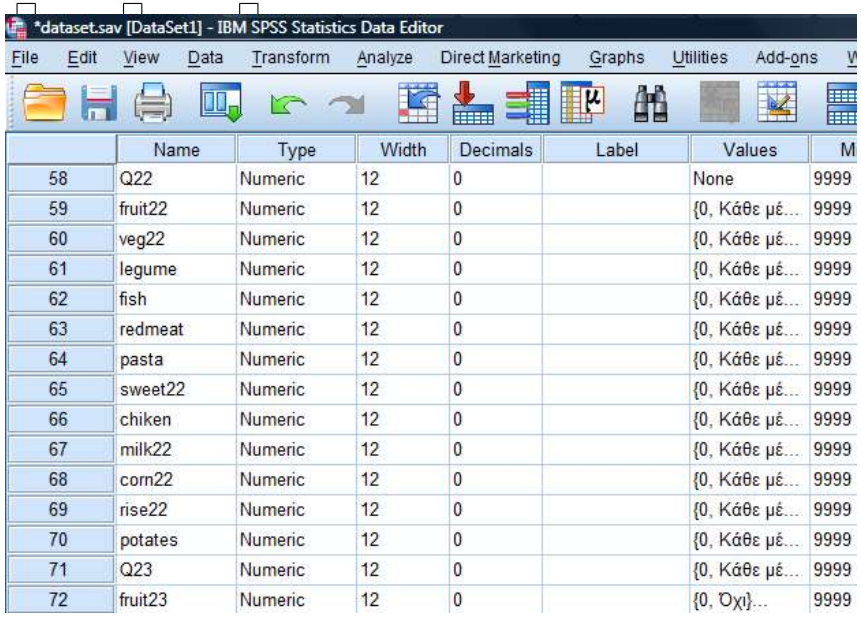




Εισαγωγή δεδομένων στον υπολογιστή



Πόσο συχνά καταναλώνεις :	ΚΑΘΕ ΜΕΡΑ	ΜΙΑ ΦΟΡΑ ΤΗΝ ΕΒΔΟΜΑΔΑ	3 ΦΟΡΕΣ ΤΟ ΜΗΝΑ
α. Φρούτα	☐	☐	☐
β. Λαχανικά	☐	☐	☐
γ. Όσπρια	☐	☐	☐
δ. Ψάρι	☐	☐	☐
ε. Κρέας κόκκινο	☐	☐	☐
στ. Ζυμαρικά	☐	☐	☐
ζ. Γλυκά	☐	☐	☐
η. Πουλερικά(κοτόπουλο, γαλοπούλα)	☐	☐	☐
θ. Γαλακτοκομικά	☐	☐	☐
ι. Δημητριακά	☐	☐	☐
ια. Ρύζι			
ιβ. Πατάτες			



*dataset.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons

	Name	Type	Width	Decimals	Label	Values	M
58	Q22	Numeric	12	0		None	9999
59	fruit22	Numeric	12	0		{0, Κάθε μέ...	9999
60	veg22	Numeric	12	0		{0, Κάθε μέ...	9999
61	legume	Numeric	12	0		{0, Κάθε μέ...	9999
62	fish	Numeric	12	0		{0, Κάθε μέ...	9999
63	redmeat	Numeric	12	0		{0, Κάθε μέ...	9999
64	pasta	Numeric	12	0		{0, Κάθε μέ...	9999
65	sweet22	Numeric	12	0		{0, Κάθε μέ...	9999
66	chiken	Numeric	12	0		{0, Κάθε μέ...	9999
67	milk22	Numeric	12	0		{0, Κάθε μέ...	9999
68	corn22	Numeric	12	0		{0, Κάθε μέ...	9999
69	rise22	Numeric	12	0		{0, Κάθε μέ...	9999
70	potates	Numeric	12	0		{0, Κάθε μέ...	9999
71	Q23	Numeric	12	0		{0, Κάθε μέ...	9999
72	fruit23	Numeric	12	0		{0, Όχι}...	9999

- corn22 ή cereals22???
- rise22???
- potates???
- Τα ονόματα των μεταβλητών εμφανίζονται σε πίνακες και σχήματα
- Χρησιμοποιούμε το Label

Εισαγωγή δεδομένων στον υπολογιστή

smoking

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Όχι	118	72,0	72,4	72,4
	Ναι	10	6,1	6,1	78,5
	2	35	21,3	21,5	100,0
	Total	163	99,4	100,0	
Missing	9999	1	,6		
Total		164	100,0		

alcohol

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Όχι	69	42,1	42,3	42,3
	Ναι	27	16,5	16,6	58,9
	2	67	40,9	41,1	100,0
	Total	163	99,4	100,0	
Missing	9999	1	,6		
Total		164	100,0		

morning

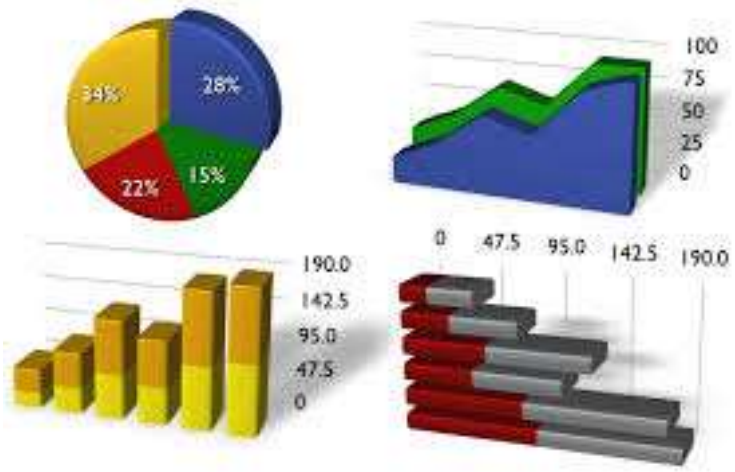
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Όχι	91	55,5	55,8	55,8
	Ναι	13	7,9	8,0	63,8
	Μερικές φορές	19	11,6	11,7	75,5
	3	40	24,4	24,5	100,0
	Total	163	99,4	100,0	
Missing	9999	1	,6		
Total		164	100,0		

- Στη μεταβλητή smoking εμφανίζεται 35 φορές η «άσχετη» τιμή 2
- Μήπως κάποιος περνά τα δεδομένα ως 0-1 και κάποιος ως 1-2???
- Στη μεταβλητή alcohol εμφανίζεται 65 φορές η «άσχετη» τιμή 2
- Στη μεταβλητή morning εμφανίζεται 45 φορές η «άσχετη» τιμή 3
- Προσέχουμε την κωδικοποίηση
- Δημιουργούμε βιβλίο κωδικοποίησης (codebook)

Ανάλυση δεδομένων

- Μια από τις πρώτες ενέργειες που κάνουμε μετά την καταγραφή των δεδομένων και την εισαγωγή τους στον ηλεκτρονικό υπολογιστή, είναι να συνοψίσουμε τα δεδομένα
- Με τον τρόπο αυτό
  - παίρνουμε μια αρχική εικόνα των δεδομένων
  - παίρνουμε χρήσιμες πληροφορίες σχετικά με αυτά (π.χ. περιοδικότητα, τάση, ακραίες τιμές κλπ),
  - αντιλαμβανόμαστε λάθη πληκτρολόγησης

Βασικές έννοιες της Στατιστικής



Πληθυσμός και Δείγμα

- **Πληθυσμός** (**Population**) είναι ένα σύνολο που μπορεί να αποτελείται από ανθρώπους, ζώα ή αντικείμενα, τα χαρακτηριστικά των οποίων θέλουμε να μελετήσουμε.
- **Δείγμα** (**Sample**) είναι μια συλλογή από στοιχειώδεις ή πρωταρχικές μονάδες δειγματοληψίας, τις οποίες επιλέγουμε έτσι ώστε να αποτελούν μια αντιπροσωπευτική εικόνα του πληθυσμού.

Μεταβλητές

- Ένα κατάλληλα επιλεγμένο **δείγμα** μπορεί να μελετηθεί ως προς ένα ή περισσότερα χαρακτηριστικά του.

- Τα χαρακτηριστικά τα οποία μας ενδιαφέρουν τα ονομάζουμε **μεταβλητές** (**variables**).

- Οι **μεταβλητές** διακρίνονται σε **ποιοτικές** και **ποσοτικές**, ανάλογα με τις τιμές που μπορούν να πάρουν και το είδος της μέτρησης που επιδέχονται.

Δεδομένα και Κλίμακες μέτρησης

- Ο όρος **δεδομένα** (**data**) αναφέρεται σε **μετρήσεις** ή **παρατηρήσεις** που προέρχονται από ένα πείραμα ή μια δειγματοληπτική έρευνα.
- Όλες οι μετρήσεις δεν είναι ίδιες. Μερικές μετρήσεις είναι ακριβέστερες από άλλες.
- Η ακρίβεια της μέτρησης μιας μεταβλητής είναι σημαντική στον καθορισμό της στατιστικής μεθόδου που θα χρησιμοποιήσουμε για την στατιστική ανάλυση.

Δεδομένα και Κλίμακες μέτρησης

- Οι κλίμακες που μετριούνται τα δεδομένα μας είναι οι ακόλουθες:
  - **Ονομαστική κλίμακα** (**Nominal scale**), όπου τα δεδομένα είναι ποιοτικά και απλώς οι αριθμοί χρησιμοποιούνται για το διαχωρισμό των δεδομένων.
  - Εάν οι κατηγορίες ποιοτικών δεδομένων υπαινίσσονται και κάποια διάταξη τότε τα δεδομένα μας βρίσκονται σε **κλίμακα διάταξης** (**Ordinal scale**).
  - Τα ποσοτικά δεδομένα χωρίζονται στις εξής δυο κατηγορίες: δεδομένα σε **κλίμακα διαστήματος** (**Interval scale**) και δεδομένα σε **κλίμακα λόγου** (**Ratio scale**).

Ποιοτικές & ποσοτικές μεταβλητές

- Οι μεταβλητές που μετριούνται σε ονομαστική ή τακτική κλίμακα είναι **ποιοτικές μεταβλητές** επειδή η μέτρηση αποτελείται από τις διατεταγμένες (ταξινομημένες) ή μη ιδιαίτερες κατηγορίες, όπως το θρήσκευμα, οι ειδικότητες των ιατρών ενός νοσοκομείου και η κλινική κατάσταση των ασθενών.

- Αντίθετα, οι μεταβλητές που μετριούνται σε κλίμακα διαστήματος ή κλίμακα αναλογίας είναι **ποσοτικές μεταβλητές**. Το ύψος, το βάρος, η θερμοκρασία, ο δείκτης νοημοσύνης είναι παραδείγματα ποσοτικών μεταβλητών.

Ποιοτικές μεταβλητές

Ποιοτικές ονομάζονται οι μεταβλητές που περιγράφουν ποιοτικά χαρακτηριστικά του πληθυσμού όπως φύλο, εθνικότητα, κατάσταση υγείας κλπ.

- Κατηγορικές είναι αυτές που οι τιμές τους εκφράζονται με λέξεις π.χ. θρήσκευμα, φύλο κλπ.
- Διατάξιμες είναι αυτές που επιδέχονται ιεράρχησης, δηλαδή παίρνουν τιμές της μορφής κακό, μέτριο, καλό, πολύ καλό, άριστο

Ποσοτικές μεταβλητές

Ποσοτικές καλούνται οι μεταβλητές που μπορούν να μετρηθούν όπως ύψος, βάρος, αριθμός παιδιών, αριθμός ερυθρών αιμοσφαιρίων, πυρετός κλπ.

- Διακριτές ή ασυνεχείς είναι αυτές που παίρνουν μόνο συγκεκριμένες ακέραιες τιμές που διαφέρουν κατά συγκεκριμένες ποσότητες, π.χ. αριθμός παιδιών, αριθμός ατυχημάτων.
- Συνεχείς είναι αυτές που μπορούν να πάρουν οποιαδήποτε αριθμητική τιμή, π.χ. ύψος, βάρος, θερμοκρασία.

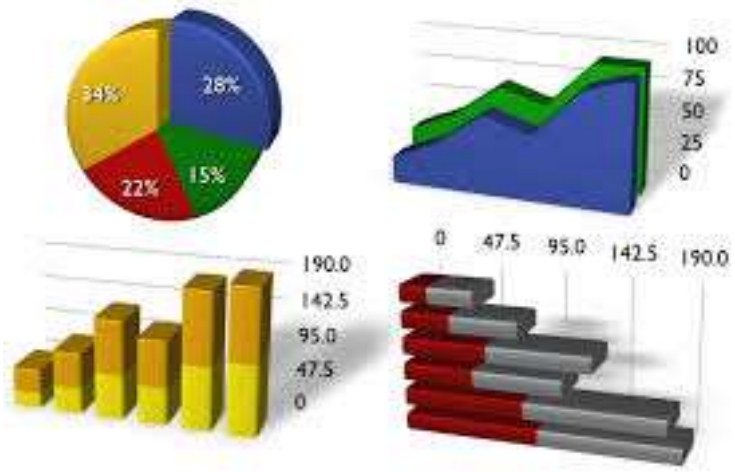
Τομείς Στατιστικής

- Η **Στατιστική** χωρίζεται σε δύο μεγάλους τομείς:
  - την **Περιγραφική Στατιστική**
  - την **Επαγωγική Στατιστική**

- Στόχος της **Περιγραφικής Στατιστικής**: Η περιγραφή είτε του υπό μελέτη πληθυσμού στο σύνολό του, είτε του διαθέσιμου και κατάλληλα επιλεγμένου δείγματος

- Στόχος της **Επαγωγικής Στατιστικής** είναι, η εξαγωγή συμπερασμάτων για τον γεννήτορα πληθυσμό με κατάλληλη χρήση του δείγματος

Περιγραφική Στατιστική



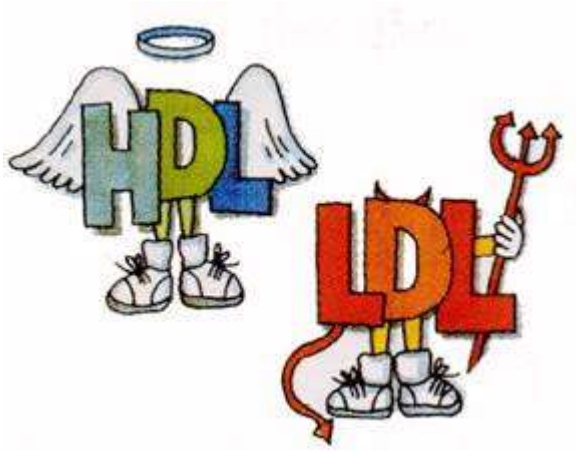
Περιγραφική Στατιστική

- Στόχος της Περιγραφικής Στατιστικής
- Η περιγραφή είτε του υπό μελέτη πληθυσμού-στόχου στο σύνολό του, είτε του διαθέσιμου και κατάλληλα επιλεγμένου δείγματος.
- Απαντά στις ερωτήσεις
  - ποιος?
  - τι?
  - πότε?
  - που?

Αδρά Δεδομένα

- Οι τιμές των μεταβλητών, στην αρχική τους μορφή, ονομάζονται **αδρά δεδομένα** (**raw data**)
- Για παράδειγμα, οι μετρήσεις χοληστερίνης του αίματος με προσέγγιση μονάδας, 60 ατόμων, είναι οι εξής:

239	212	249	227	218	310	281	330	226	233
223	161	195	233	249	284	284	174	170	256
169	299	210	301	199	258	258	195	227	244
355	234	195	196	354	282	282	286	286	176
195	163	297	211	228	309	309	225	223	195
248	284	173	256	169	209	209	200	258	284



Εργαλεία για την περιγραφή δειγμάτων / πληθυσμών:

Πίνακες Συχνοτήτων

Διαγράμματα (Γραφήματα)

Περιγραφικά Μέτρα

Πίνακες Συχνοτήτων

- Συχνότητες:
  - **Απόλυτες συχνότητες n_i :** ο αριθμός των μονάδων του πληθυσμού (ή του δείγματος) που ανήκουν σε μια συγκεκριμένη τιμή (κλάση ή κατηγορία).
  - **Σχετικές συχνότητες f_i:** το ποσοστό των μονάδων του πληθυσμού (ή του δείγματος) που ανήκουν σε μια συγκεκριμένη τιμή (κλάση ή κατηγορία).

  - $f_i = n_i/N$ (πληθυσμός) ή
    $f_i = n_i/n$ (δείγμα)

Πίνακες Συχνοτήτων

Μη διατεταγμένες μεταβλητές

- **Απλός (μη διατεταγμένος) πίνακας**

Φύλο Ασθενούς	***Απόλυτες Συχνότητες***	***Σχετικές Συχνότητες***
Γυναίκες	*187*	*0,48 ή 48%*
Άνδρες	*201*	*0,52 ή 52%*
Σύνολο	*388*	*1,00 ή 100%*

Πίνακες Συχνοτήτων

Διατεταγμένες μεταβλητές

- Διατεταγμένος πίνακας συχνοτήτων

Κατάσταση Υγείας	**Απόλυτες Συχνότητες**	**Απόλυτες Αθροιστικές Συχνότητες**	**Σχετικές Συχνότητες**	**Σχετικές Αθροιστικές Συχνότητες**
Κακή	4	4	20%	20%
Μέτρια	6	10	30%	50%
Άριστη	10	**20**	50%	**100%**
Σύνολο	**20**		**100%**	

Πίνακες Συχνοτήτων

Διακριτές ποσοτικές μεταβλητές

- Πίνακας συχνοτήτων για διακριτές ποσοτικές μεταβλητές

Αριθμός Παιδιών	Απόλυτες Συχνότητες	Απόλυτες Αθροιστικές Συχνότητες	Σχετικές Συχνότητες	Σχετικές Αθροιστικές Συχνότητες
0	8	8	16%	16%
1	10	18	20%	36%
2	15	33	30%	66%
3	14	47	28%	94%
4	3	**50**	6%	**100%**
Σύνολο	**50**		**100%**	

Παράδειγμα

- Καταγράψαμε το ύψος 50 ατόμων:

1,78 1,68 1,88 2,01 1,95 2,04
1,94 1,69 1,64 1,96 1,91 2,03
1,69 1,81 1,87 1,61 1,95 1,85
1,67 2,00 1,88 1,84 1,65 1,74
1,98 1,82 1,68 1,80 1,82 1,67
1,77 1,80 2,05 2,00 1,98 1,89
1,98 1,86 1,83 1,62 2,01 1,67
1,84 1,94 1,91 1,79 2,03 1,88
1,67 1,71

Ύψος	Συχνότητα	Σχετική Συχνότητα	Αθροιστική Σχετική Συχνότητα
1,61	1	2,0	2,0
1,62	1	2,0	4,0
1,64	1	2,0	6,0
1,65	1	2,0	8,0
1,67	4	8,0	16,0
1,68	2	4,0	20,0
1,69	2	4,0	24,0
1,71	1	2,0	26,0
1,74	1	2,0	28,0
1,77	1	2,0	30,0
1,78	1	2,0	32,0
1,79	1	2,0	34,0
1,80	2	4,0	38,0
1,81	1	2,0	40,0
1,82	2	4,0	44,0
1,83	1	2,0	46,0
1,84	2	4,0	50,0
1,85	1	2,0	52,0
1,86	1	2,0	54,0
1,87	1	2,0	56,0
1,88	3	6,0	62,0
1,89	1	2,0	64,0
1,91	2	4,0	68,0
1,94	2	4,0	72,0
1,95	2	4,0	76,0
1,96	1	2,0	78,0
1,98	3	6,0	84,0
2,00	2	4,0	88,0
2,01	2	4,0	92,0
2,03	2	4,0	96,0
2,04	1	2,0	98,0
2,05	1	2,0	100,0
Σύνολο	50	100,0	

Πίνακες Συχνοτήτων

- Πίνακας συχνοτήτων συνεχών ποσοτικών μεταβλητών

Κλάση	Όρια	Τιμή κλάσης ή κεντρική τιμή	Απόλυτη Συχνότητα	Σχετική Συχνότητα	Αθρ. Σχετ. Συχνότητα
1	21,30	25,5	35	9,0%	9,0%
2	31,35	33	63	16,2%	25,2%
3	36,40	38	120	30,9%	56,1%
4	41,45	43	64	16,5%	72,6%
5	46,50	48	46	11,9%	84,5%
6	51,65	58	60	15,5%	100,0%
ΣΥΝΟΛΟ			388	100,0%	

Ποσοτικές μεταβλητές

Πίνακες Συχνοτήτων

- Για την κατασκευή ενός πίνακα συχνοτήτων, για συνεχείς ποσοτικές μεταβλητές, ακολουθούμε τα επόμενα βήματα:
  1. Βρίσκουμε την ελάχιστη και τη μέγιστη τιμή στα δεδομένα μας
  2. Υπολογίζουμε το εύρος των δεδομένων (μέγιστη τιμή – ελάχιστη τιμή)
  3. Υπολογίζουμε τον αριθμό των κλάσεων (τάξεων)
  4. Υπολογίζουμε το εύρος κάθε κλάσης
  5. Καταγράφουμε τον αριθμό των τιμών της μεταβλητής που ανήκουν σε κάθε μια από τις κλάσεις

Περιγραφική Στατιστική

Πίνακες Συχνοτήτων

Διαγράμματα

Περιγραφικά Μέτρα

Γραφικές Μέθοδοι

- Τα διαγράμματα χρησιμοποιούνται ευρύτατα στη Στατιστική για την παρουσίαση των δεδομένων

- Ένα διάγραμμα θα πρέπει να είναι
  - παραστατικό
    - να διευκολύνει την κατανόηση και
    - να παρουσιάζει τα βασικά χαρακτηριστικά της μεταβλητής
  - σαφές
    - να μη δημιουργεί σύγχυση
  - ακριβές
    - να μην παραπλανεί τον αναγνώστη

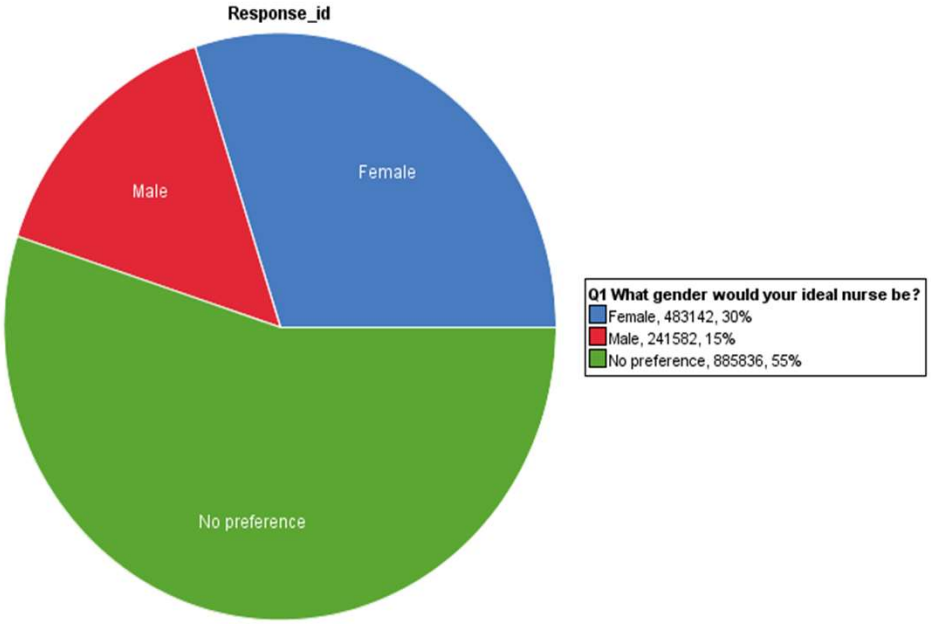
Γραφικές Μέθοδοι

- Ανάλογα με τον τύπο της μεταβλητής έχουμε και τα κατάλληλα διαγράμματα.
  - Για ένα δείγμα έχουμε:
    - Για τις μεταβλητές ονομαστικής κλίμακας (nominal variables) χρησιμοποιούμε το κυκλικό διάγραμμα.
    - Για τις διατάξιμες μεταβλητές (ordinal variables) χρησιμοποιούμε το ραβδόγραμμα.
    - Για τις ποσοτικές μεταβλητές χρησιμοποιούμε το ιστόγραμμα.
  - Για δυο ή περισσότερα δείγματα έχουμε:
    - Τα διαγράμματα διασποράς για τις ποσοτικές μεταβλητές.
    - Τα συνδυασμένα ραβδογράμματα για τις ποιοτικές μεταβλητές.
    - Πολλαπλά θηκογράμματα για μεικτές (ποιοτικές μαζί με ποσοτικές μεταβλητές).

Κυκλικό Γράφημα

- Κυκλικό διάγραμμα (pie chart)
- Για την κατασκευή ενός κυκλικού διαγράμματος, δημιουργούμε έναν κύκλο και τον χωρίζουμε σε τόσους τομείς όσες είναι και οι κατηγορίες της μεταβλητής. Οι μοίρες κάθε κυκλικού τομέα θα πρέπει να αντιστοιχούν στη σχετική συχνότητα της αντίστοιχης κατηγορίας.

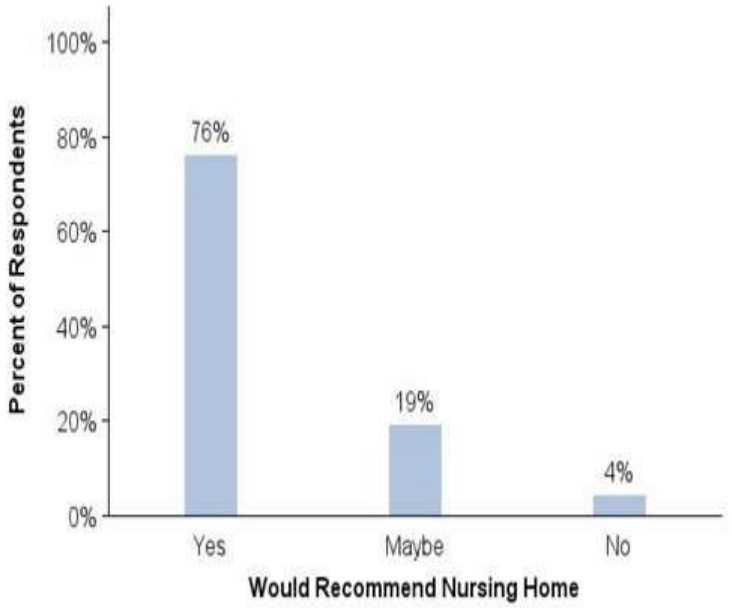
Μεταβλητές ονομαστικής κλίμακας



Ραβδόγραμμα

- Ραβδόγραμμα (bar chart)
- Το ραβδόγραμμα αποτελείται από τόσα ορθογώνια παραλληλόγραμμα όσα οι τιμές της ποιοτικής μεταβλητής, τα οποία έχουν ίσες συνήθως βάσεις και ύψη ίσα με τις απόλυτες ή τις σχετικές συχνότητες των αντίστοιχων τάξεων. Μεταξύ των στηλών υπάρχει απόσταση για να μην υποδηλώνεται συνέχεια.

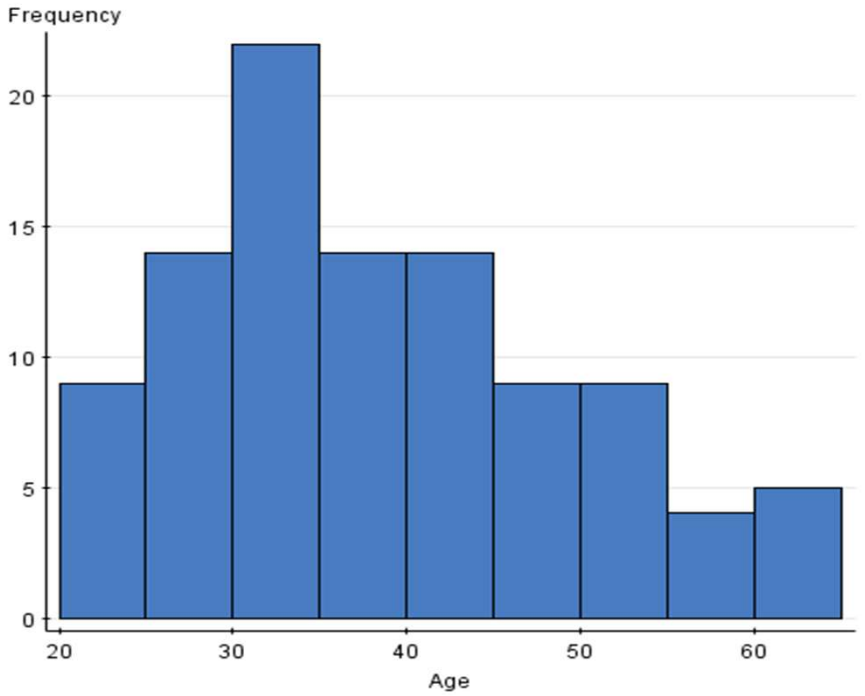
Διατάξιμες ή ονομαστικής κλίμακας μεταβλητές



Ιστόγραμμα

- Ιστόγραμμα (Histogram)
- Το ιστόγραμμα συχνοτήτων αποτελείται από ορθογώνια παραλληλόγραμμα – τόσα όσα οι κλάσεις της μεταβλητής – με ίσες και συνεχόμενες βάσεις και ύψη ανάλογα (όχι ίσα, δεδομένου, ότι το εμβαδόν των παραλληλόγραμμων πρέπει να ισούται με την συνολική πιθανότητα της κλάσης) των συχνοτήτων των αντίστοιχων κλάσεων.

Ποσοτικές μεταβλητές

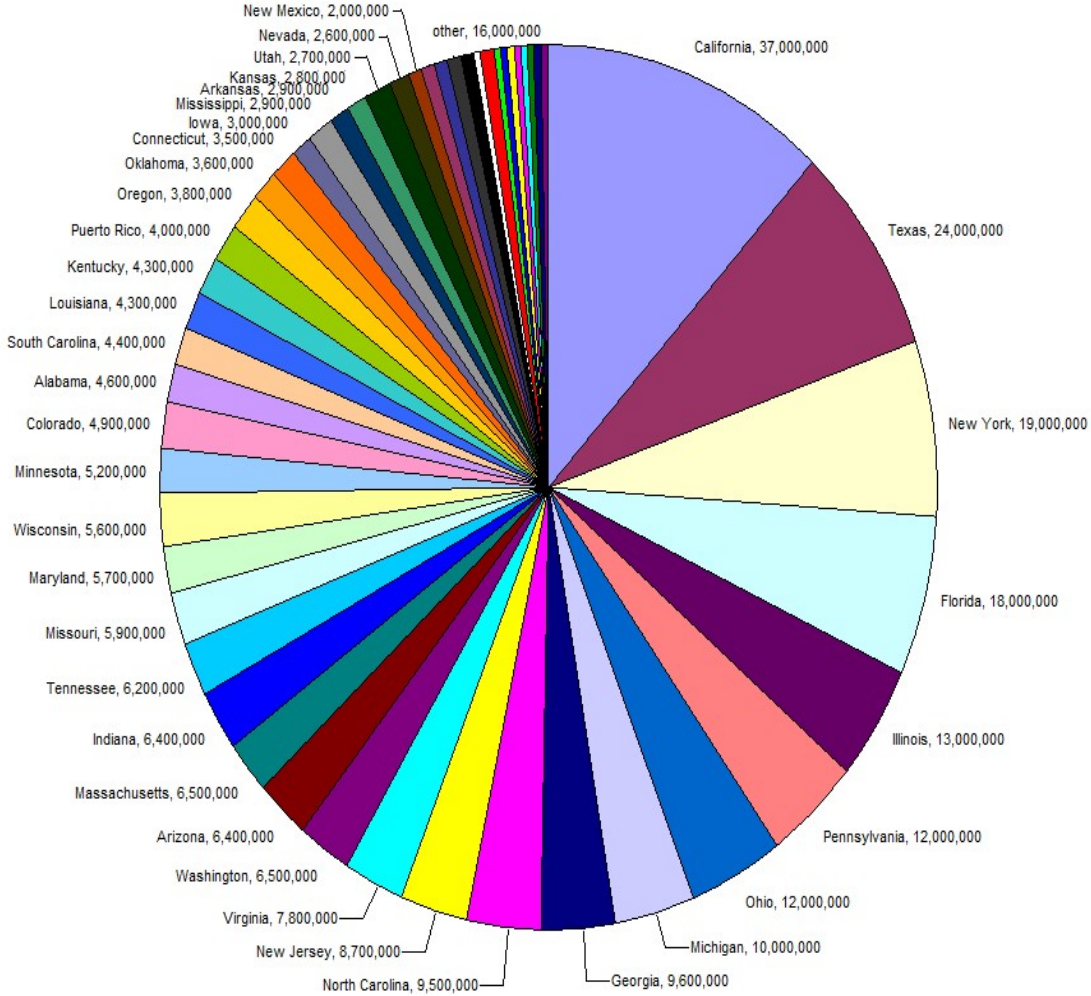


Ερωτήματα Σχετικά με τα Γραφήματα

- Είναι η πληροφορία σωστά παρουσιασμένη;
- Προσπαθεί το γράφημα να σε επηρεάσει;
- Η κλίμακα χρησιμοποιεί κανονικά διαστήματα;
- Τι εντύπωση σου δίνει το γράφημα;

Λάθος ή Παραπλανητικά Γραφήματα

- Το κυκλικό διάγραμμα είναι κατάλληλο για ποιοτικές μεταβλητές με λίγες κατηγορίες



Λάθος ή Παραπλανητικά Γραφήματα

- Το ραβδόγραμμα είναι κατάλληλο για ποιοτικές μεταβλητές με αρκετές ή πολλές κατηγορίες

- Το διπλανό γράφημα όσο δύσκολο να διαβαστεί και εάν είναι, αποτελεί τον σωστό τρόπο γραφικής παρουσίασης της υπό μελέτη μεταβλητής

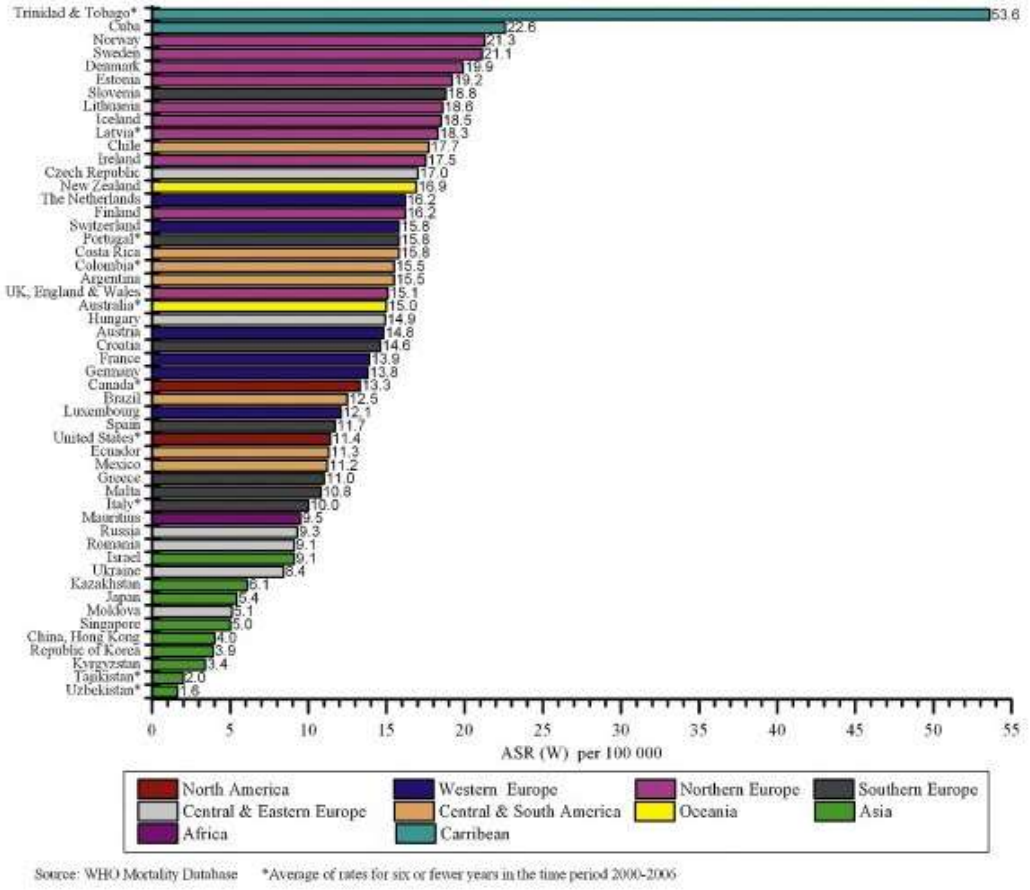
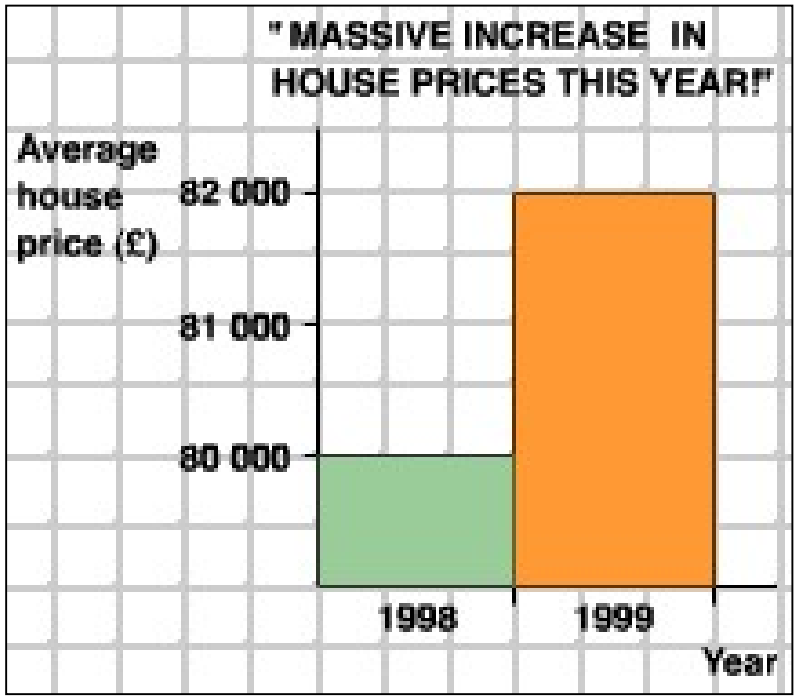


Fig. 4 Prostate cancer mortality rates for select countries, 2000-2006. *Average of rates for ≤6 yr in the time period 2000-2006. Source: World Health Organization mortality database [22].

ASR (W) = age-standardized rate (world).

Λάθος ή Παραπλανητικά Γραφήματα



Ο τίτλος λέει στον αναγνώστη τι να σκεφτεί (ότι υπάρχει τεράστια αύξηση στις τιμές)

Η απόσταση από το 0 έως το 80.000 είναι η ίδια με την απόσταση από το 80.000 στο 81.000

Η πραγματική αύξηση είναι 2.000, η οποία ισοδυναμεί σε λιγότερο από 3% αύξηση

Το γράφημα δείχνει τη δεύτερη ράβδο σαν να είναι τριπλάσια από τη πρώτη, το οποίο ισοδυναμεί με 300% αύξηση

Λάθος ή Παραπλανητικά Γραφήματα

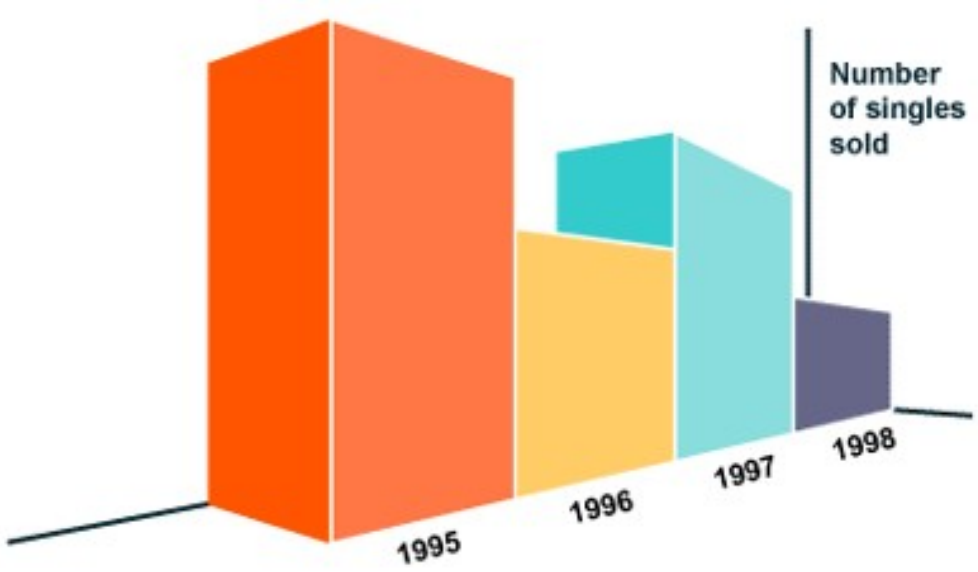
Μήπως αυτό το γράφημα είναι καλύτερο;

Ο τίτλος είναι αμερόληπτος

Η κλίμακα είναι κανονική

Το γράφημα απεικονίζει την σωστή αύξηση

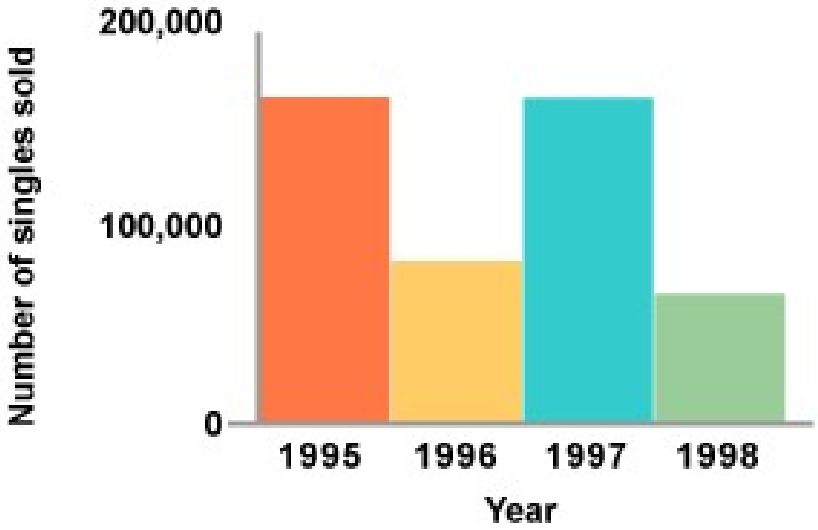
Λάθος ή Παραπλανητικά Γραφήματα



Ο κάθετος άξονας δεν έχει τιμές

Το 1995 πωλήθηκαν τα περισσότερα προϊόντα, ακολούθησε το 1997, στη συνέχεια το 1996 και τέλος το 1998

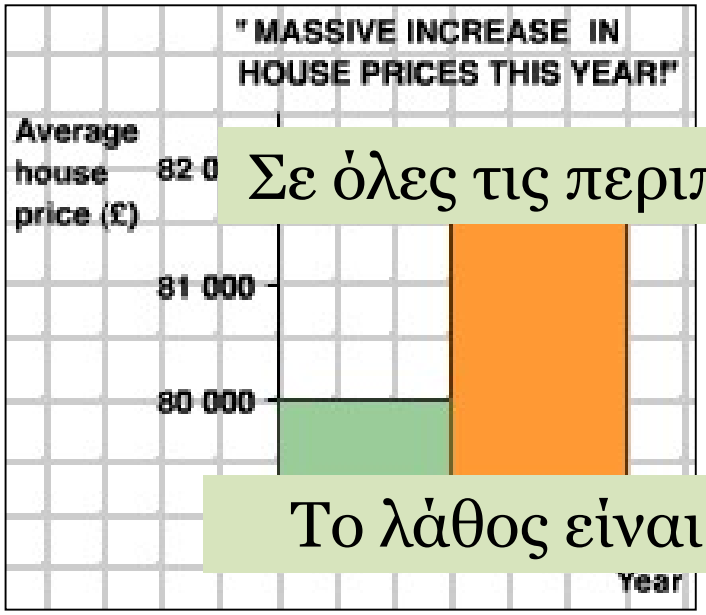
Λάθος ή Παραπλανητικά Γραφήματα

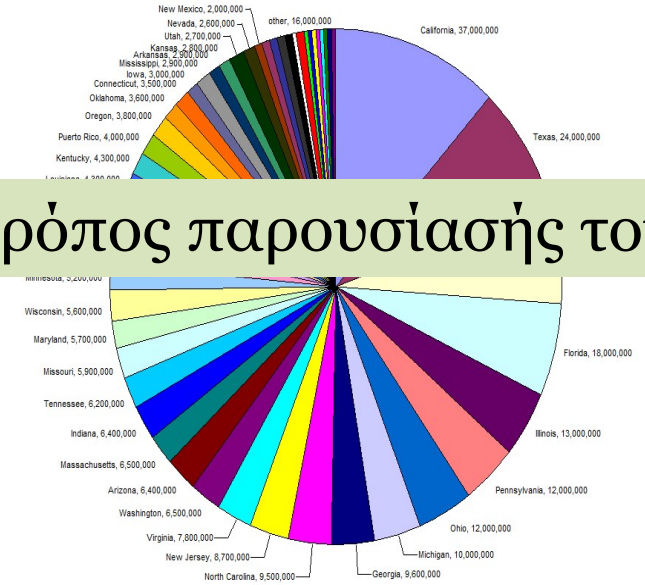


Ο κάθετος άξονας έχει τιμές

Το 1995 και το 1997 έχουμε τις ίδιες πωλήσεις ενώ ακολουθούν το 1996 και το 1998

Λάθος ή Παραπλανητικά Γραφήματα



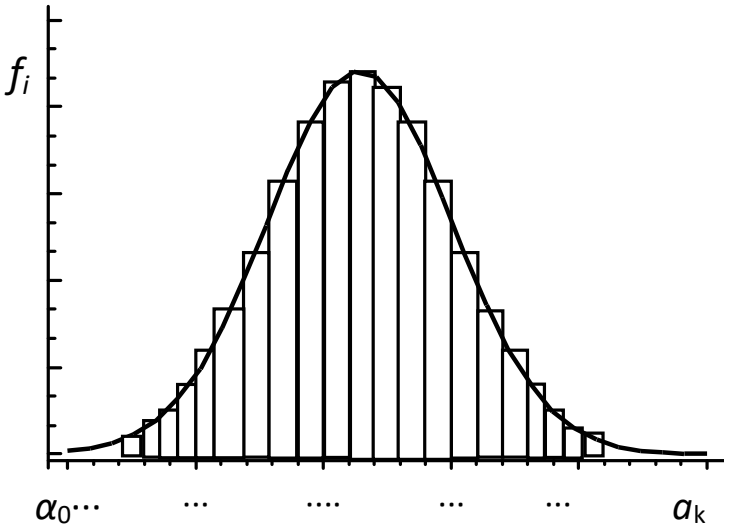


Σε όλες τις περιπτώσεις τα δεδομένα είναι σωστά

Το λάθος είναι ο τρόπος παρουσίασής τους!!!

Καμπύλη Συχνοτήτων

- Περνώντας, από το δείγμα στον πληθυσμό (θεωρώντας ότι ο πληθυσμός έχει πολύ μεγάλο ή θεωρητικά άπειρο μέγεθος), ο αριθμός των κλάσεων του ιστογράμματος για μια συνεχή μεταβλητή μεγαλώνει πάρα πολύ (ή τείνει στο άπειρο) και το πλάτος των κλάσεων γίνεται πολύ μικρό (ή τείνει στο μηδέν).
- Σε αυτή την περίπτωση, η πολυγωνική γραμμή συχνοτήτων τείνει να πάρει τη μορφή μιας ομαλής καμπύλης, η οποία ονομάζεται **καμπύλη συχνοτήτων**.
- Με τον τρόπο αυτό κανείς περνάει από τη δειγματική κατανομή, στην αντίστοιχη κατανομή της μεταβλητής X στον πληθυσμό. Δηλαδή, προσεγγίζει την θεωρητική κατανομή του πληθυσμού από τον οποίο προήλθε το δείγμα.



Περιγραφικά Μέτρα

- Οι ποσότητες που συνοψίζουν πληροφορίες είτε της πληθυσμιακής, είτε της δειγματικής κατανομής μιας μεταβλητής καλούνται **περιγραφικά μέτρα** (descriptive measures).

- Έχουμε τρεις κατηγορίες περιγραφικών μέτρων:
  - τα μέτρα θέσης της κατανομής (μέση τιμή, διάμεσος, κορυφή)
  - τα μέτρα διασποράς της κατανομής (εύρος, διακύμανση, τυπ. απόκλιση)
  - τα μέτρα μορφής της κατανομής (συντ. ασυμμετρίας και κύρτωσης)

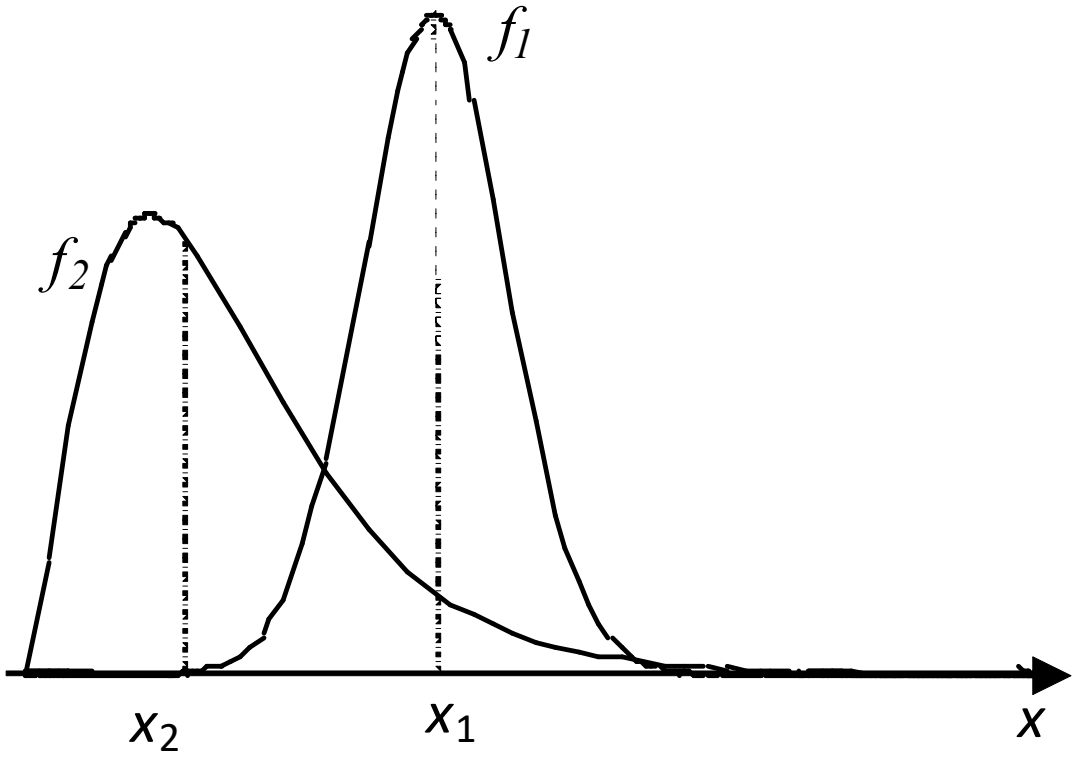
- Εάν οι μετρήσεις προέρχονται από έναν πληθυσμό ή ένα δείγμα, μιλάμε για πληθυσμιακά μέτρα ή δειγματικά μέτρα, αντίστοιχα.

Περιγραφικά Μέτρα

- Οι τιμές κάθε μεταβλητής έχουν την τάση να συγκεντρώνονται γύρω από κάποια τιμή. Η τιμή αυτή ονομάζεται «κεντρική» τιμή και δείχνει τη θέση της μεταβλητής *X* στον οριζόντιο άξονα. Με άλλα λόγια, η **θέση** της κατανομής είναι το σημείο γύρω από το οποίο συγκεντρώνονται οι τιμές της μεταβλητής.
- Με την έννοια **διασπορά**, περιγράφεται το πόσο απομακρυσμένες είναι οι τιμές της μεταβλητής από το σημείο συγκέντρωσης («κεντρική» τιμή) ή με άλλα λόγια το εύρος των τιμών που καταλαμβάνει η μεταβλητή της εμπειρικής ή της πληθυσμιακής κατανομής που μελετάται.
- Με την έννοια της **μορφής** περιγράφεται το σχήμα που έχει η κατανομή. Δηλαδή εάν παρουσιάζει συμμετρία, λοξότητα, αιχμηρότητα, κλπ.

Περιγραφικά Μέτρα

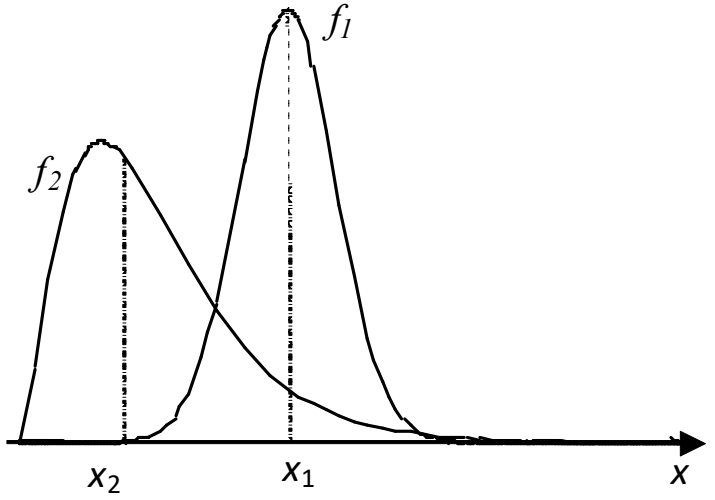
- **θέση**
- **διασπορά**
- **μορφή**



Μέτρα Θέσης

- Τα **μέτρα θέσης** συνοψίζουν πληροφορίες για την κατανομή συχνοτήτων των τιμών μιας μεταβλητής ως προς το σημείο που βρίσκεται πάνω στον οριζόντιο άξονα.

- Μας δείχνουν δηλαδή το σημείο γύρω από το οποίο βρίσκονται οι τιμές της μεταβλητής.



Μέτρα Θέσης

- **Μέση τιμή**
- **Μέση τιμή (μ) – mean, average** τιμών μιας μεταβλητής, καλείται το πηλίκο του αθροίσματος των τιμών αυτών δια του πλήθους τους. Η πληθυσμιακή μέση τιμή συμβολίζεται με το μ, ενώ η δειγματική με το $\overline{X}$ και υπολογίζονται στην περίπτωση των απλών δεδομένων (μη ομαδοποιημένα δεδομένα) με χρήση των ακόλουθων τύπων:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{και} \quad \mu = \frac{\sum_{i=1}^{N} x_i}{N}.$$

Μέτρα Θέσης

- Η μέση τιμή δίνει την τιμή των δεδομένων γύρω από την οποία είναι συγκεντρωμένες οι τιμές της μεταβλητής που έχουμε στη διάθεσή μας.

- Η μέση τιμή αποτελεί αντιπροσωπευτικό μέτρο θέσης της κατανομής των δεδομένων στην περίπτωση συμμετρικών κατανομών. Σε αντίθετη περίπτωση έχει την τάση να ακολουθεί τις ουρές της κατανομής (δεξιά ή αριστερά), επηρεαζόμενη από τυχόν μεγάλες τιμές.

Μέτρα Θέσης

- **Υπολογισμός μέσης τιμής**

- Οι ηλικίες 6 ασθενών σε μια κλινική είναι οι εξής:
  34, 27, 45, 55, 22, 34

- Η μέση τιμή ισούται με
  (34 + 27 + 45 + 55 + 22 + 34)/6 = 217/6 ≈ 36,167

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Εάν ο μεγαλύτερος ασθενής ήταν 95 αντί 55 ετών,
  (34 + 27 + 45 + 95 + 22 + 34)/6 = 257/6 ≈ 42,833

Μέτρα Θέσης

- **Επικρατούσα τιμή ή Κορυφή**
- **Επικρατούσα τιμή ή Κορυφή (τ) – mode** τιμών μιας μεταβλητής, καλείται η τιμή της μεταβλητής που παρουσιάζει τη μεγαλύτερη συχνότητα (δηλαδή η παρατήρηση που εμφανίζεται περισσότερες φορές). Εάν η μεταβλητή έχει μια επικρατούσα τιμή τότε λέγεται μονοκόρυφη ενώ εάν έχει δυο λέγεται δικόρυφη κοκ. Δηλαδή, για μη ομαδοποιημένα δεδομένα, αναζητούμε την παρατήρηση που εμφανίζεται περισσότερες φορές.

Μέτρα Θέσης

- **Υπολογισμός επικρατούσας τιμής**

- Οι ηλικίες 6 ασθενών σε μια κλινική είναι οι εξής:

  34, 27, 45, 55, 22, 34

- Η επικρατούσα τιμή είναι η 34 γιατί εμφανίζεται 2 φορές

- Εάν ο μεγαλύτερος ασθενής ήταν 95 αντί 55 ετών, τότε η επικρατούσα τιμή είναι

  πάλι η 34

Μέτρα Θέσης

- **Διάμεση Τιμή ή Διάμεσος**
- **Διάμεση Τιμή ή Διάμεσος (*m*) - median** τιμών μιας μεταβλητής, καλείται μια τιμή της μεταβλητής, τέτοια, ώστε το μισό (50%) των δεδομένων τιμών να είναι μικρότερες ή ίσες από αυτήν και το άλλο μισό (50%) μεγαλύτερες. Η διάμεσος ή διχοτόμος για μη ομαδοποιημένα δεδομένα δίνεται από τη σχέση:

$$m = \begin{cases} X_{\left[\frac{n+1}{2}\right]}, & \text{για } n \text{ περιττό} \\ \left(\dfrac{X_{\left[\frac{n}{2}\right]} + X_{\left[\frac{n}{2}+1\right]}}{2} \right), & \text{για } n \text{ άρτιο} \end{cases}$$

όπου $X_{[n]}$ είναι η n-οστή παρατήρηση, αφού πρώτα διαταχθούν οι παρατηρήσεις σε αύξουσα σειρά.

Μέτρα Θέσης

- **Υπολογισμός διαμέσου**

- Οι ηλικίες 7 ασθενών σε μια κλινική είναι οι εξής: 36, 34, 27, 45, 55, 22, 34
- Η θέση της διαμέσου είναι η (7 + 1)/2 = 8/2 = 4
- Ταξινομούμε τις ηλικίες: 22, 27, 34, 34, 36, 45, 55
- Εάν ο μεγαλύτερος ασθενής ήταν 95 αντί 55 ετών,
- Τότε η διάμεσος είναι πάλι 22, 27, 34, 34, 36, 45, 95

- Οι ηλικίες 6 ασθενών σε μια κλινική είναι οι εξής: 34, 27, 45, 55, 22, 34
- Η θέση της διαμέσου είναι η (6 + 1)/2 = 7/2 = 3,5
- Ταξινομούμε τις ηλικίες: 22, 27, 34, 34, 45, 55
- Η διάμεσος ισούται με (34 + 34)/2 = 34

Μέτρα Θέσης

- Αντίστοιχα, έχουμε και τα τεταρτημόρια της κατανομής μιας μεταβλητής:
  1. **Πρώτο Τεταρτημόριο (Q_1)** καλείται μια τιμή της μεταβλητής, τέτοια ώστε το ένα τέταρτο ($n/4$ ή 25%) των τιμών της μεταβλητής να είναι μικρότερες ή ίσες με αυτήν
  2. **Τρίτο Τεταρτημόριο (Q_3)** καλείται μια τιμή της μεταβλητής, τέτοια ώστε τα τρία τέταρτα ($3n/4$ ή 75%) των τιμών της μεταβλητής να είναι μικρότερες ή ίσες με αυτήν.

- Το **δεύτερο τεταρτημόριο** (Q_2) συμπίπτει με τη **διάμεσο**.

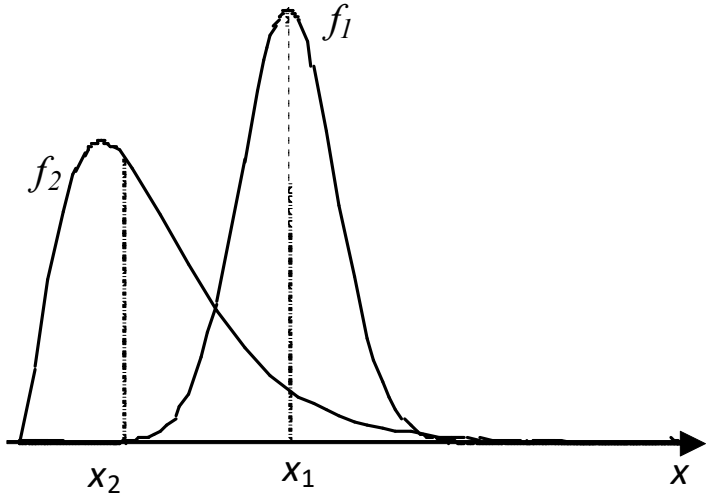
Επιλογή Μέτρου Θέσης

	Μέτρο Θέσης		
Τύπος Μεταβλητής	**Επικρατούσα Τιμή**	**Διάμεσος**	**Μέση Τιμή**
Ονομαστικής Κλίμακας	✓	✗	✗
Τακτικής Κλίμακας	✓	✓	✗
Ποσοτική Διακριτή	✓	✓	✓
Ποσοτική Συνεχής	✗	✓	✓

Μέτρα Διασποράς

- Τα **μέτρα διασποράς** χαρακτηρίζουν την κατανομή των τιμών μιας μεταβλητής ως προς τη μεταβλητότητά της.

- Μας ενημερώνουν, δηλαδή, για το πόσο απέχουν οι τιμές μιας μεταβλητής από συγκεκριμένες τιμές (π.χ. από τη μέση τιμή).

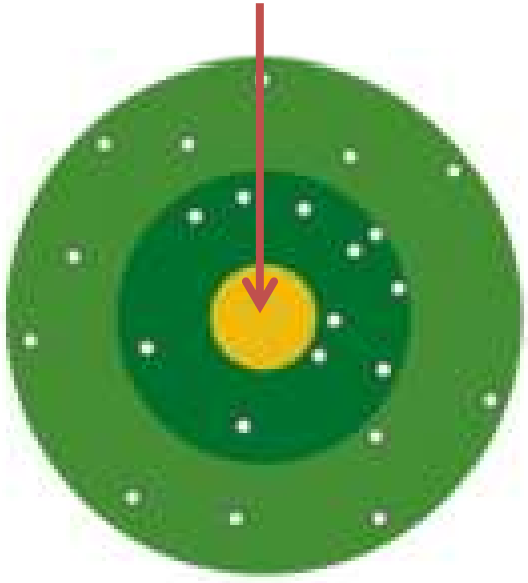


Μέτρα Διασποράς

Μέση τιμή

Μικρή μεταβλητότητα

Μέση τιμή



Μεγάλη μεταβλητότητα

Μέτρα Διασποράς

- **Εύρος μεταβολής**
- **Εύρος μεταβολής (R) – Range** των τιμών μιας μεταβλητής, καλείται η διαφορά μεταξύ της μεγαλύτερης και της μικρότερης τιμής της μεταβλητής, δηλαδή η τιμή

$$R = \text{μέγιστη τιμή} - \text{ελάχιστη τιμή}$$

Μέτρα Διασποράς

- **Ενδοτεταρτημοριακό εύρος**
- **Ενδοτεταρτημοριακό εύρος (*IQR*) – Interquartile Range** των τιμών μιας μεταβλητής, καλείται η διαφορά μεταξύ του 3ου τεταρτημορίου και του 1ου τεταρτημορίου, δηλαδή η τιμή

  IQR = 3ο τεταρτημόριο – 1ο τεταρτημόριο

Μέτρα Διασποράς

- **Διακύμανση**
- **Διακύμανση (σ^2) – variance** των τιμών μιας μεταβλητής, καλείται η μέση τιμή των τετραγώνων των διαφορών (αποκλίσεων) των τιμών της μεταβλητής από τη μέση τιμή της μεταβλητής. Η πληθυσμιακή διακύμανση δίνεται από τον τύπο

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

ενώ η δειγματική διακύμανση συμβολίζεται με s^2 και δίνεται από τον τύπο $\sum_{i=1}^{n}(x_i - \bar{x})^2$

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

Μέτρα Διασποράς

- **Υπολογισμός διακύμανσης**

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

- Οι ηλικίες 6 ασθενών σε μια κλινική είναι οι εξής:

  34, 27, 45, 55, 22, 34
- Υπολογίζουμε το άθροισμα των αποκλίσεων από τη μέση τιμή εις το τετράγωνο

  $(34 – 36{,}167)^2 + (27 – 36{,}167)^2 + (45 – 36{,}167)^2 +$

  $(55 – 36{,}167)^2 + (22 – 36{,}167)^2 + (34 – 36{,}167)^2 \approx 726{,}83$
- Διαιρούμε με την ποσότητα $6 – 1 = 5$
- Άρα η διακύμανση ισούται με $726{,}83/5 \approx 145{,}37$
- Εάν ο μεγαλύτερος ασθενής είναι 95 αντί 55 τότε η διακύμανση θα ισούται με 713,37

Μέτρα Διασποράς

- **Τυπική απόκλιση**
- **Τυπική απόκλιση (σ) – standard deviation** των τιμών μιας μεταβλητής, καλείται η τετραγωνική ρίζα της διακύμανσης.
- Η δειγματική τυπική απόκλιση συμβολίζεται με s και δίνεται από τον τύπο

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

Μέτρα Διασποράς

- **Υπολογισμός τυπικής απόκλισης**

- Οι ηλικίες 6 ασθενών σε μια κλινική είναι οι εξής:
  34, 27, 45, 55, 22, 34

- Η διακύμανση ισούται 145,37

- Άρα η τυπική απόκλιση ισούται με $\sqrt{145,37} = 12,06$

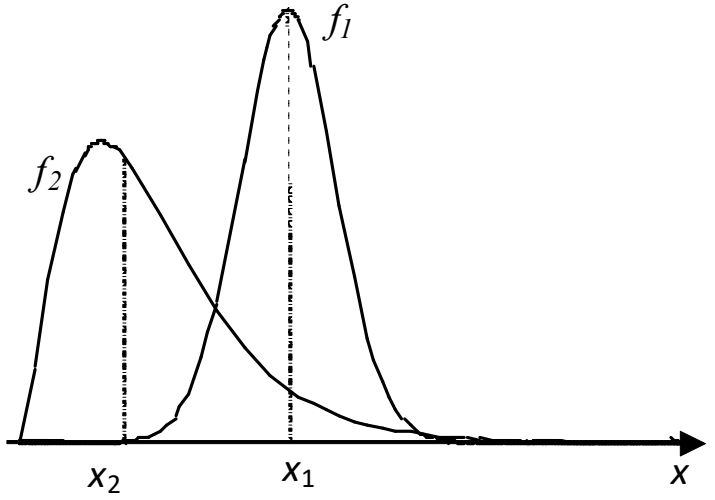
Επιλογή Μέτρου Διασποράς

	Μέτρο Διασποράς		
Τύπος Μεταβλητής	**Εύρος**	**Ενδοτεταρτημοριακό Εύρος**	**Τυπική Απόκλιση**
Ονομαστικής Κλίμακας	✗	✗	✗
Τακτικής Κλίμακας	✓	✓	✗
Ποσοτική	✓	✓	✓

Μέτρα Μορφής

- Τα **μέτρα μορφής** χαρακτηρίζουν το σχήμα της κατανομής.

- Τα κυριότερα μέτρα είναι ο **συντελεστής ασυμμετρίας** και ο **συντελεστής κύρτωσης**.



Μέτρα Λοξότητας ή Ασυμμετρίας

- Σε έναν θάλαμο του παθολογικού τμήματος ενός νοσοκομείου νοσηλεύονται 7 ασθενείς με ηλικία

  37, 43, 49, 55, 61, 67, 73.

- Μπορούμε να παρατηρήσουμε ότι οι ηλικίες των ασθενών συγκεντρώνονται γύρω από την «κεντρική» τιμή των 55 ετών συμμετρικά.
- Στον διπλανό θάλαμο νοσηλεύονται 7 ασθενείς με ηλικία

  22, 23, 26, 28, 31, 59, 63.

- Παρατηρούμε ότι οι περισσότερες ηλικίες στον θάλαμο είναι συγκεντρωμένες στα αριστερά (οι περισσότεροι ασθενείς είναι μικροί σε ηλικία).
- Στον διπλανό θάλαμο νοσηλεύονται 7 ασθενείς με ηλικία

  32, 35, 68, 72, 73, 76, 80.

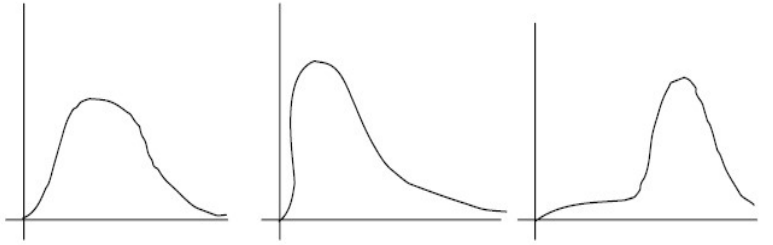
- Παρατηρούμε ότι στον θάλαμο αυτό οι περισσότερες ηλικίες είναι συγκεντρωμένες στα δεξιά (οι περισσότεροι ασθενείς είναι μεγάλοι σε ηλικία).

Μέτρα Λοξότητας ή Ασυμμετρίας

- **συντελεστής ασυμμετρίας β_1:**

$$\beta_1 = \frac{1}{n-1} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{s^3}$$

- Οι κατανομές διακρίνονται γενικά σε
  - συμμετρικές $\beta_1 = 0$
  - θετικά συμμετρικές $\beta_1 > 0$
  - αρνητικά συμμετρικές $\beta_1 < 0$
- Σε μια συμμετρική κατανομή,
  - Οι τιμές κατανέμονται ομοιόμορφα γύρω από τη μέση τους τιμή
  - Η μέση τιμή, η διάμεσος και η επικρατούσα τιμή συμπίπτουν.

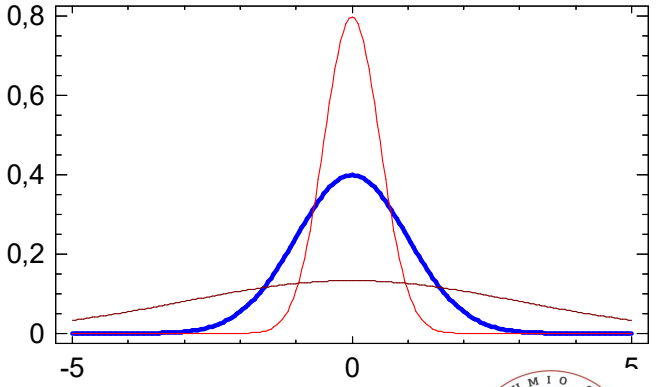


Μέτρα Κύρτωσης

- **συντελεστής κύρτωσης β_2:**

$$\beta_2 = \frac{1}{n-1}\frac{\sum_{i=1}^{n}(x_i-\bar{x})^4}{s^4} - 3$$

- Η κύρτωση χαρακτηρίζει την **αιχμηρότητα** της καμπύλης μιας κατανομής.
- Με βάση την κύρτωση, οι κατανομές διακρίνονται σε:
  - λεπτόκυρτες $\beta_2 > 0$
  - μεσόκυρτες $\beta_2 = 0$
  - πλατύκυρτες $\beta_2 < 0$

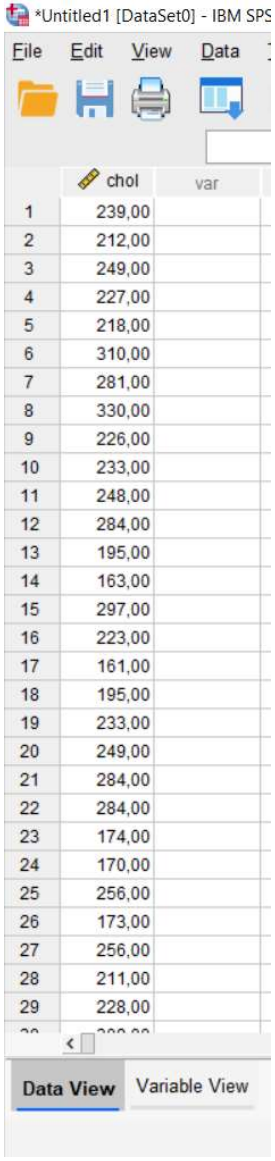


Παράδειγμα

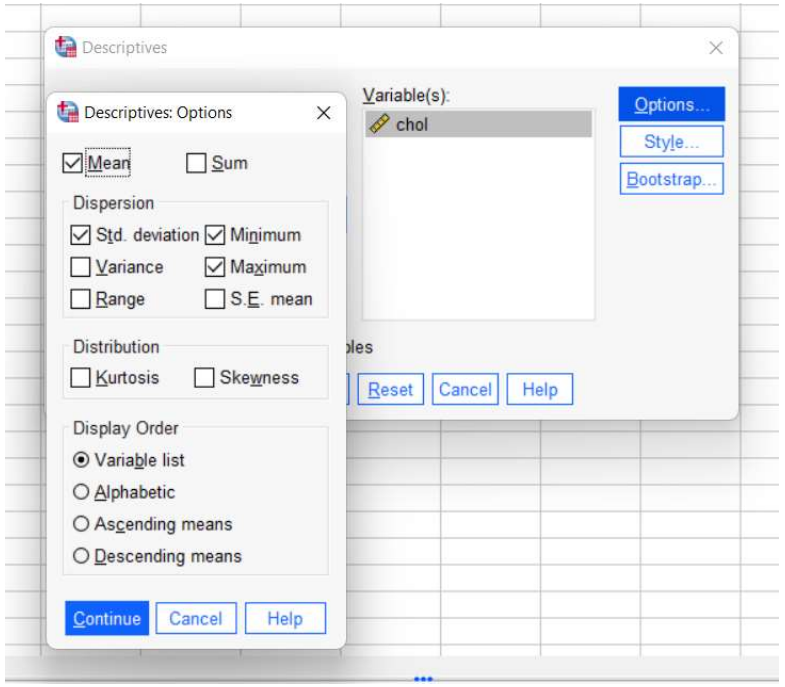
Οι μετρήσεις χοληστερίνης του αίματος με προσέγγιση μονάδας, 60 ατόμων, είναι οι εξής:

239	212	249	227	218	310	281	330	226	233	248	284	195	163	297
223	161	195	233	249	284	284	174	170	256	173	256	211	228	309
169	299	210	301	199	258	258	195	227	244	169	209	309	225	223
355	234	195	196	354	282	282	286	286	176	209	200	195	258	284

Παράδειγμα



Analyze → Descriptive Statistics → Descriptives…

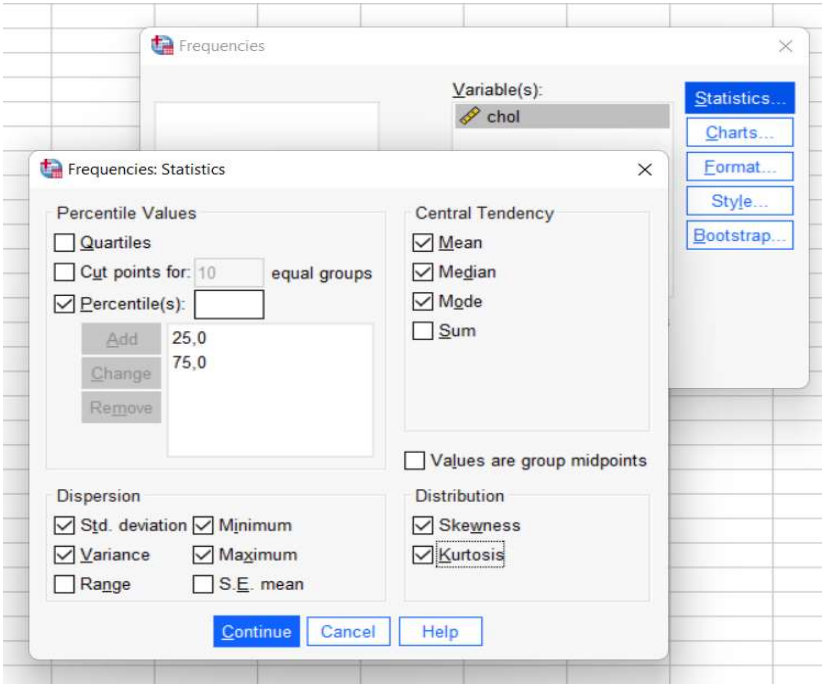


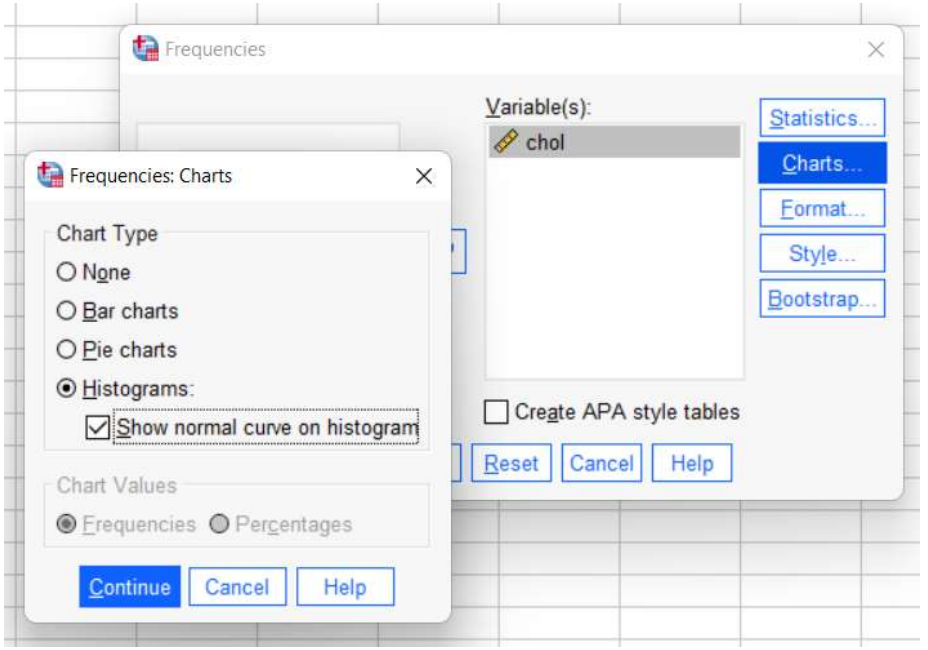
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
chol	60	161,00	355,00	240,0833	48,72729	,363	,309	-,539	,608
Valid N (listwise)	60								

Παράδειγμα

Analyze → Descriptive Statistics → Frequencies… → Statistics και Charts





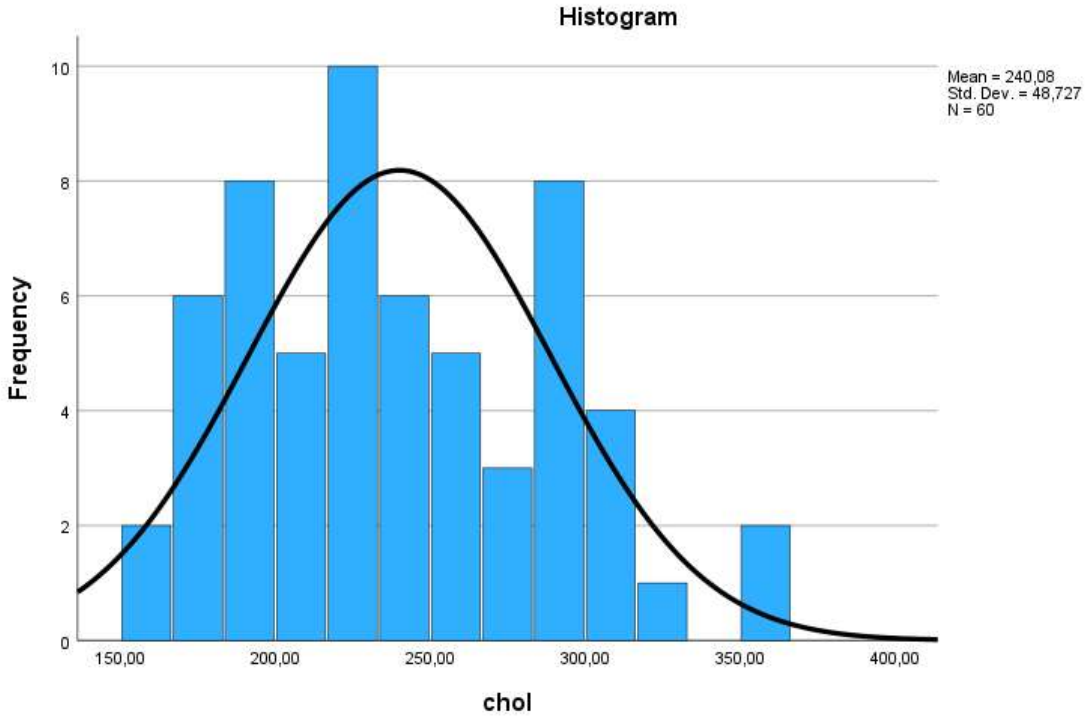
Παράδειγμα

Analyze → Descriptive Statistics → Frequencies… → Statistics και Charts

Statistics

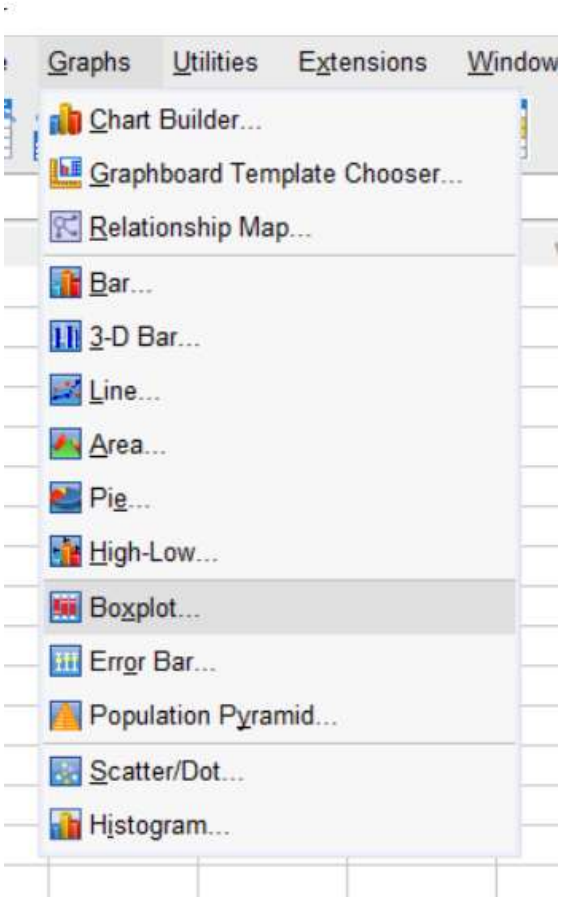
chol

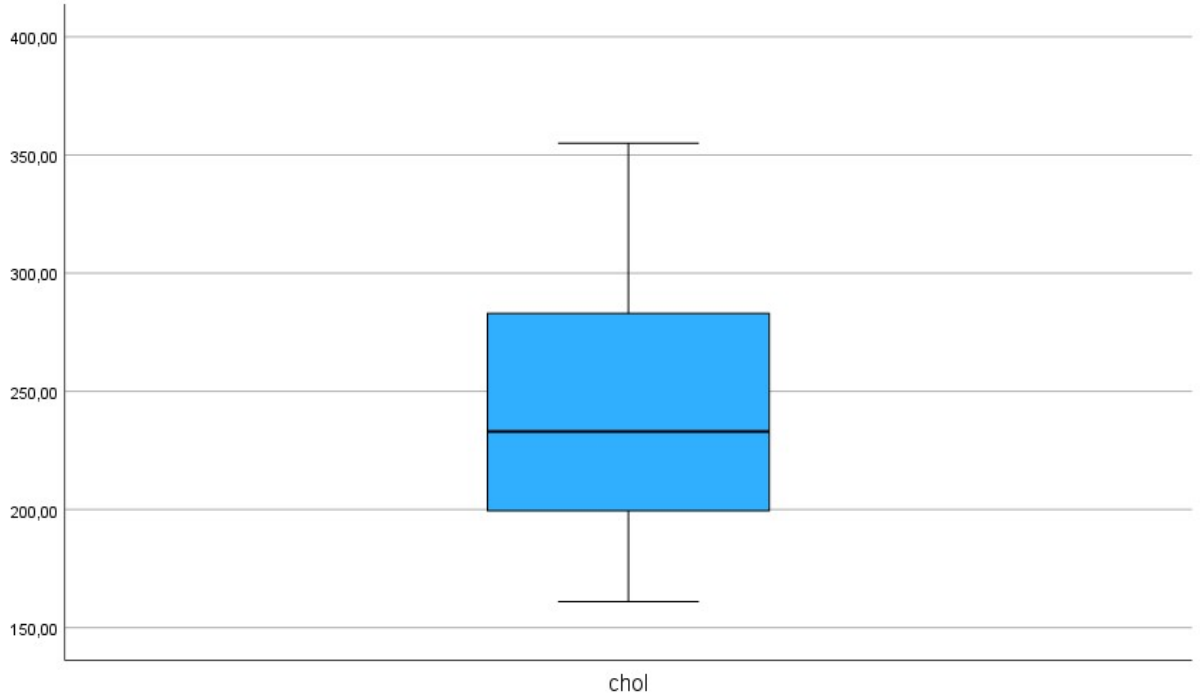
N	Valid	60
	Missing	0
Mean		240,0833
Median		233,0000
Mode		195,00
Std. Deviation		48,72729
Variance		2374,349
Skewness		,363
Std. Error of Skewness		,309
Kurtosis		-,539
Std. Error of Kurtosis		,608
Minimum		161,00
Maximum		355,00
Percentiles	25	199,2500
	75	283,5000



Παράδειγμα

Graphs → Boxplot…





Παράδειγμα

chol

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	161,00	1	1,7	1,7	1,7
	163,00	1	1,7	1,7	3,3
	169,00	2	3,3	3,3	6,7
	170,00	1	1,7	1,7	8,3
	173,00	1	1,7	1,7	10,0
	174,00	1	1,7	1,7	11,7
	176,00	1	1,7	1,7	13,3
	195,00	5	8,3	8,3	21,7
	196,00	1	1,7	1,7	23,3
	199,00	1	1,7	1,7	25,0
	200,00	1	1,7	1,7	26,7
	209,00	2	3,3	3,3	30,0
	210,00	1	1,7	1,7	31,7
	211,00	1	1,7	1,7	33,3
	212,00	1	1,7	1,7	35,0
	218,00	1	1,7	1,7	36,7
	223,00	2	3,3	3,3	40,0
	225,00	1	1,7	1,7	41,7
	226,00	1	1,7	1,7	43,3
	227,00	2	3,3	3,3	46,7
	228,00	1	1,7	1,7	48,3
	233,00	2	3,3	3,3	51,7
	234,00	1	1,7	1,7	53,3
	239,00	1	1,7	1,7	55,0
	244,00	1	1,7	1,7	56,7
	248,00	1	1,7	1,7	58,3
	249,00	2	3,3	3,3	61,7
	256,00	2	3,3	3,3	65,0
	258,00	3	5,0	5,0	70,0
	281,00	1	1,7	1,7	71,7
	282,00	2	3,3	3,3	75,0
	284,00	4	6,7	6,7	81,7
	286,00	2	3,3	3,3	85,0
	297,00	1	1,7	1,7	86,7
	299,00	1	1,7	1,7	88,3
	301,00	1	1,7	1,7	90,0
	309,00	2	3,3	3,3	93,3
	310,00	1	1,7	1,7	95,0
	330,00	1	1,7	1,7	96,7
	354,00	1	1,7	1,7	98,3
	355,00	1	1,7	1,7	100,0
	Total	60	100,0	100,0	

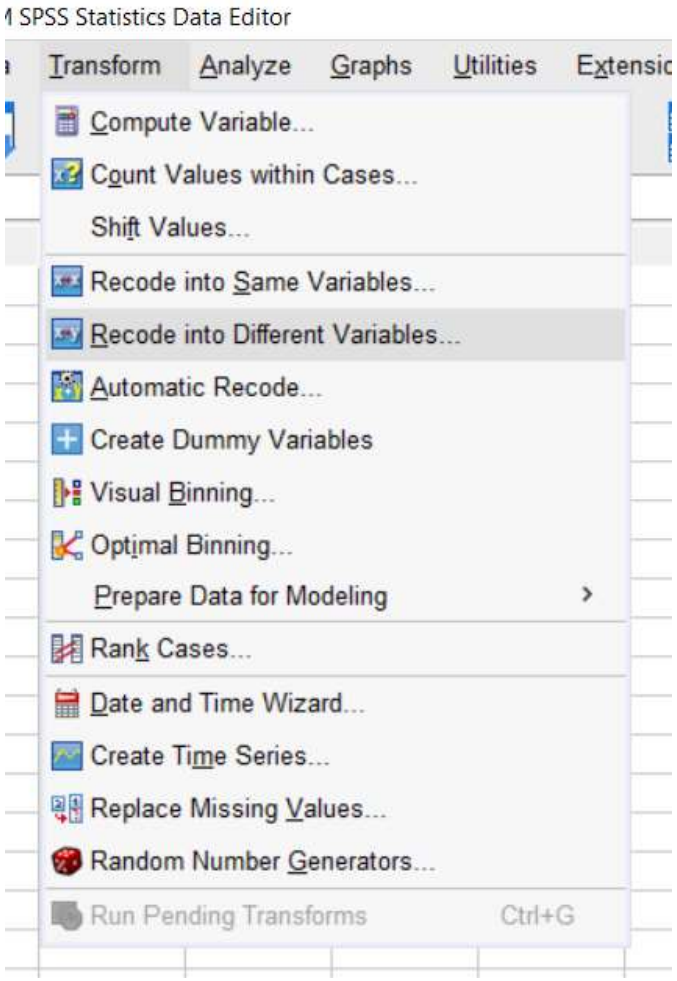
Analyze → Descriptive Statistics → Frequencies…

Ο πίνακας συχνοτήτων δεν είναι πληροφοριακός

Παράδειγμα

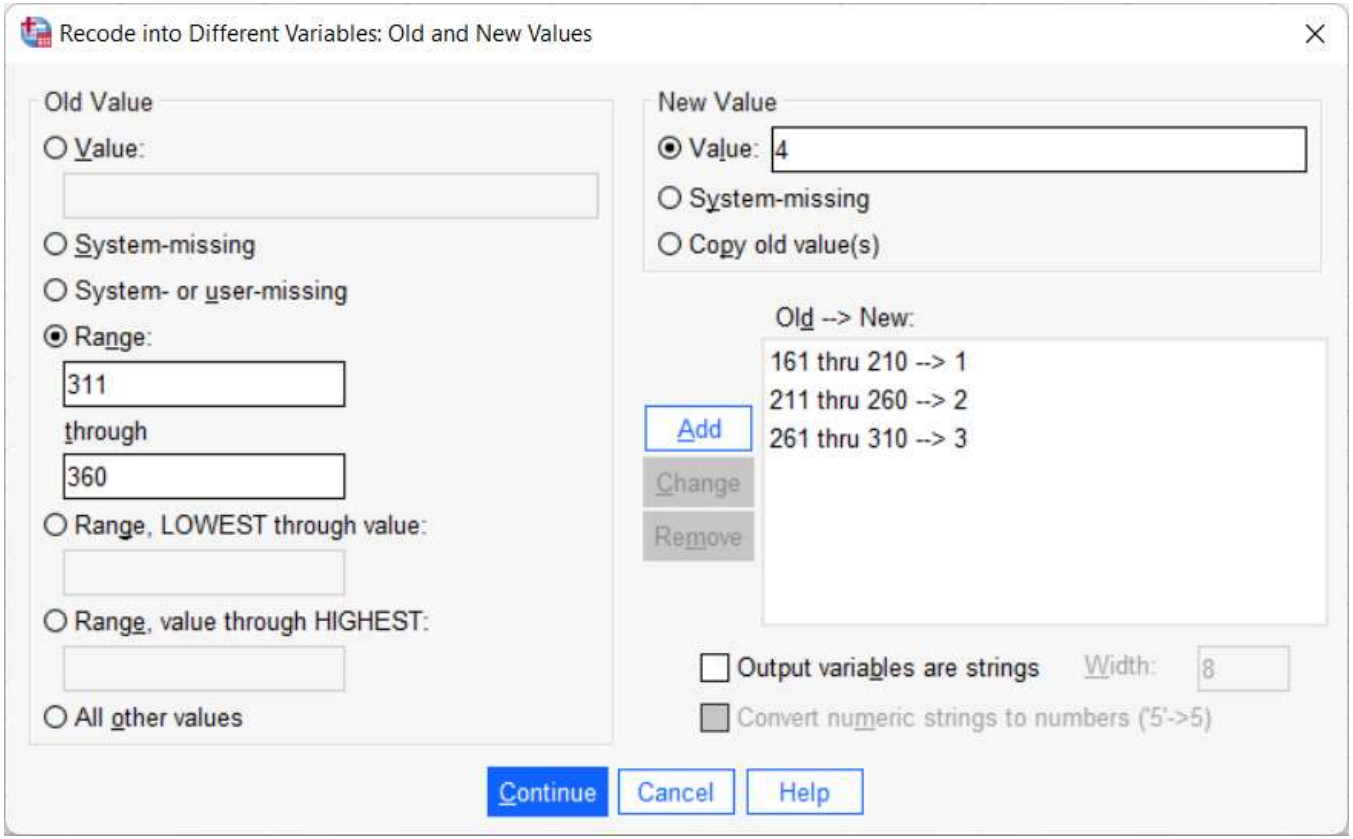
Θα πρέπει να ομαδοποιήσουμε τα δεδομένα σε κατάλληλες ομάδες

Transform → Recode into Different Variables…



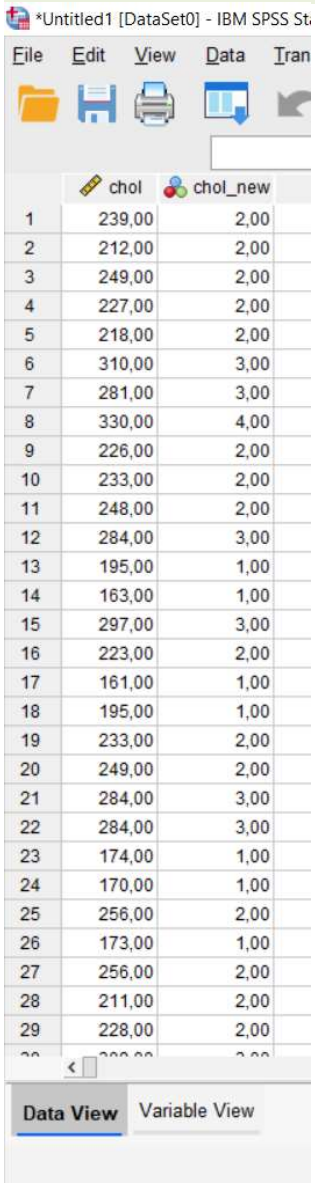
Παράδειγμα

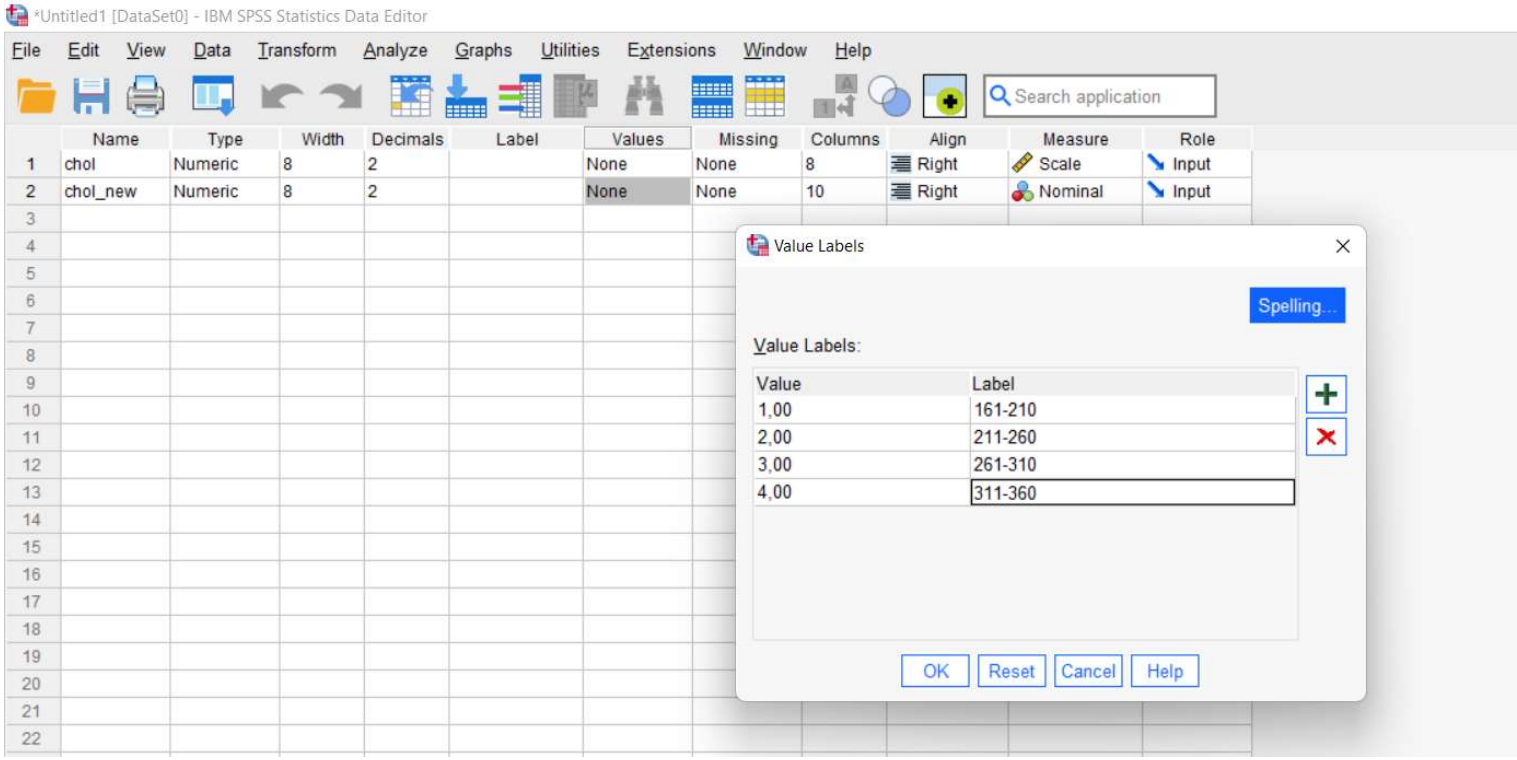
Transform → Recode into Different Variables…



Παράδειγμα

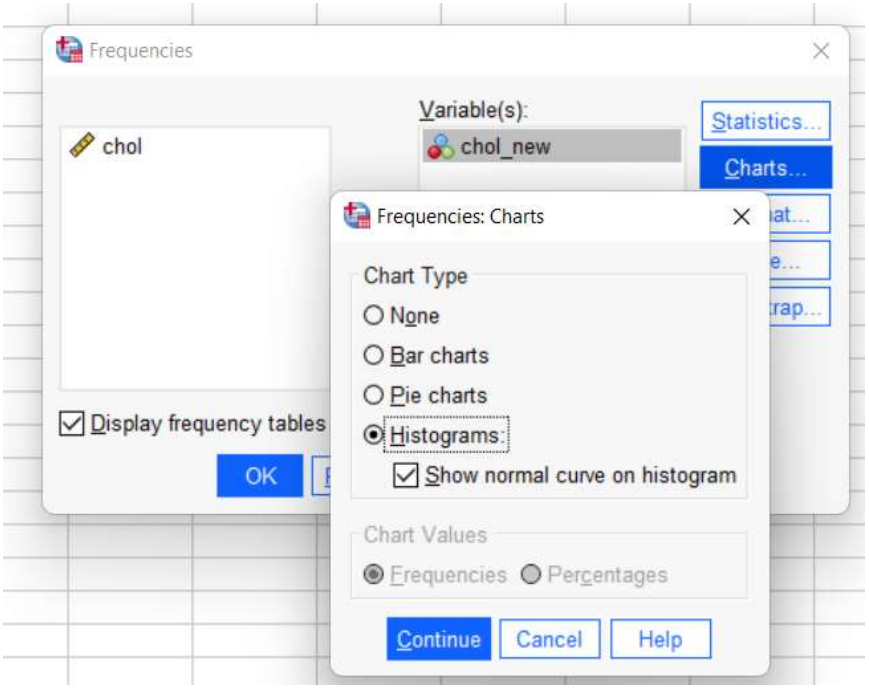
Δίνουμε ετικέτες στις κατηγορίες





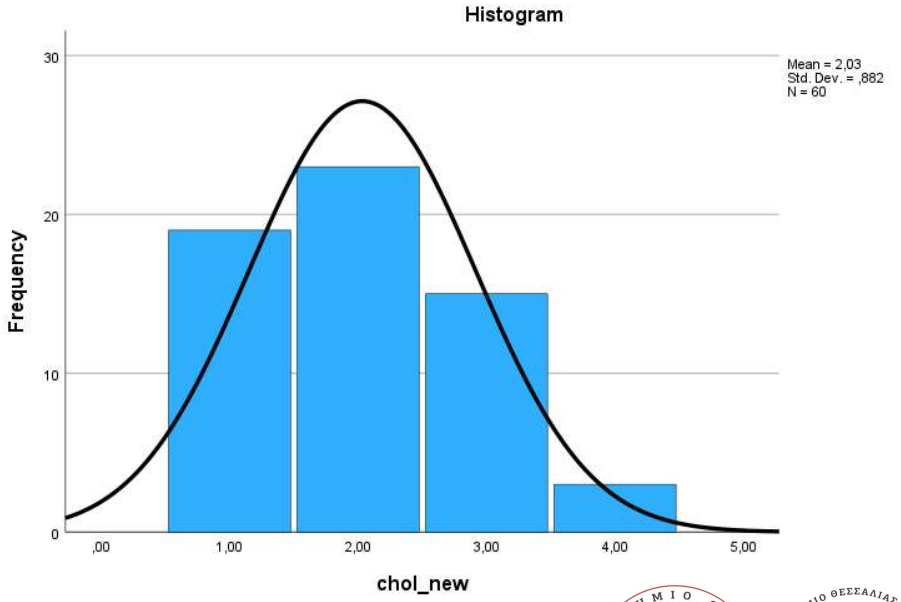
Παράδειγμα

Analyze → Descriptive Statistics → Frequencies… → Charts

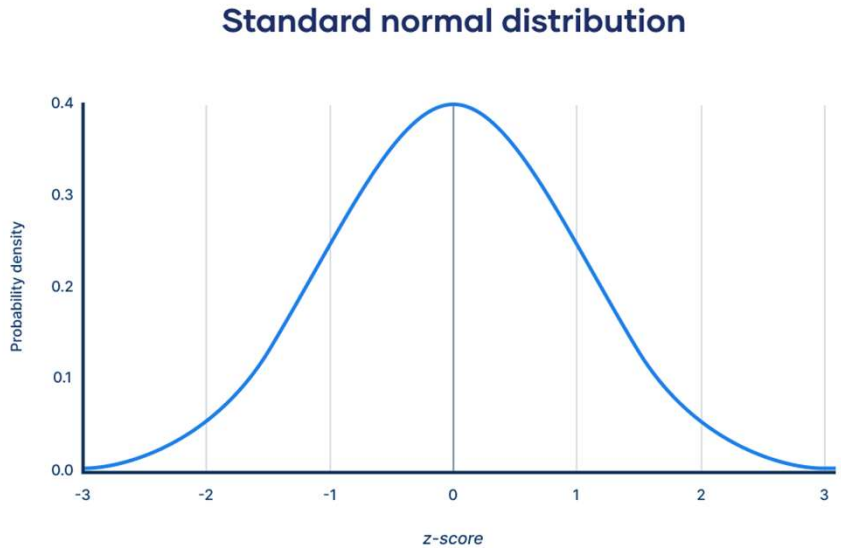


chol_new

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	161-210	19	31,7	31,7	31,7
	211-260	23	38,3	38,3	70,0
	261-310	15	25,0	25,0	95,0
	311-360	3	5,0	5,0	100,0
	Total	60	100,0	100,0	



Στατιστική Συμπερασματολογία



Στατιστική συμπερασματολογία

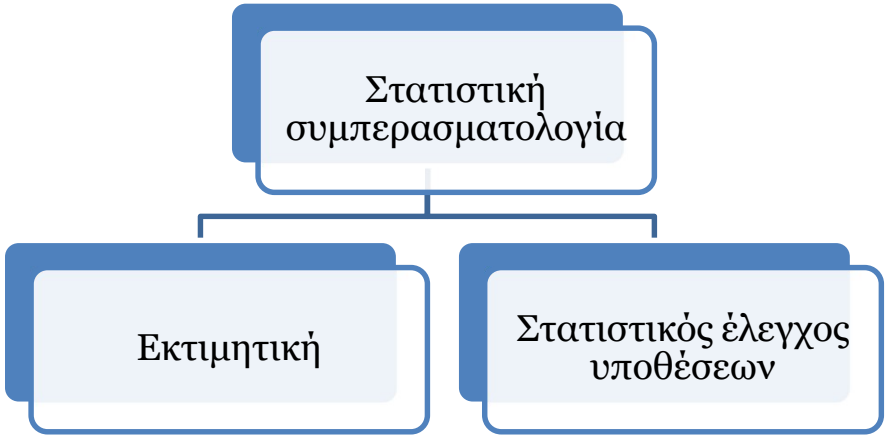
- **Συμπερασματολογία** είναι ο κλάδος της Λογικής, ο οποίος ασχολείται με την εξαγωγή συμπερασμάτων (αποφάσεων ή προβλέψεων για το μέλλον)

- Παραδείγματα:
  - πρόβλεψη του αριθμού εισαγωγών
  - πρόβλεψη της κατάστασης υγείας
  - πρόβλεψη του προσδόκιμου ζωής
  - σύγκριση δυο θεραπειών

- **Στατιστική συμπερασματολογία (Statistical inference)** είναι η επιστήμη που ασχολείται με την εξαγωγή συμπερασμάτων σχετικά με τις παραμέτρους ενός πληθυσμού

- **Στατιστική συμπερασματολογία** είναι η διαδικασία της γενίκευσης των αποτελεσμάτων από το δείγμα στον πληθυσμό

Στατιστική συμπερασματολογία



- Στόχος της **Εκτιμητικής** είναι η εκτίμηση των παραμέτρων της κατανομής κάποιου χαρακτηριστικού ενός πληθυσμού με χρήση ενός δείγματος
- Στόχος του **Στατιστικού Ελέγχου Υποθέσεων** είναι η διατύπωση υποθέσεων, η ανάπτυξη μεθοδολογιών που στηρίζονται σε δείγμα και η εξαγωγή συμπερασμάτων για την στατιστική ορθότητα ή μη των υποθέσεων

Έννοιες Εκτιμητικής

- **Παράμετρος (parameter):** Μια σταθερά θ, η οποία μετρά ένα χαρακτηριστικό μιας κατανομής.
  - Η μέση τιμή μ, για παράδειγμα, μετρά την κεντρική τάση της κατανομής.
  - Η διακύμανση σ^2 , μετρά τη διασπορά της κατανομής.

- **Εκτιμήτρια (estimator):** Μια συνάρτηση των παρατηρήσεων του δείγματος που περιλαμβάνει άγνωστες παραμέτρους. Είναι τυχαία μεταβλητή και ακολουθεί κάποια κατανομή.

- **Εκτίμηση (estimate):** Η τιμή που παίρνει η εκτιμήτρια για συγκεκριμένα δεδομένα. Συμβολίζεται με $\hat{\theta}$.

Εκτίμηση παραμέτρων

- Η εκτίμηση μιας παραμέτρου θ του πληθυσμού πραγματοποιείται με δυο τρόπους:

  - Λαμβάνοντας μια σημειακή εκτίμηση $\hat{\theta}$, η οποία συνήθως προκύπτει από μια σχετική με την παράμετρο στο δείγμα στατιστική συνάρτηση, π.χ. η πληθυσμιακή διασπορά σ^2 εκτιμάται από τη δειγματική διακύμανση S^2

  - Καθορίζοντας ένα διάστημα τιμών μέσα στο οποίο βρίσκεται η υπό εκτίμηση παράμετρος με έναν σχετικά υψηλό «συντελεστή εμπιστοσύνης»

Σημειακή εκτίμηση

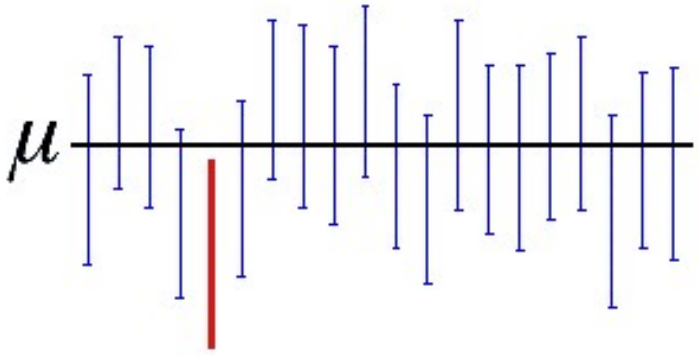
Σημειακή εκτίμηση

- Έστω ένα τ.δ. X_1, X_2, ..., X_n από έναν πληθυσμό με μια άγνωστη παράμετρο, την οποία θέλουμε να εκτιμήσουμε, π.χ. την μέση ηλικία των ασθενών ενός νοσοκομείου

- Χρησιμοποιούμε μια ή περισσότερες αριθμητικές ποσότητες για την εκτίμηση της παραμέτρου (π.χ. για το μ την $\bar{X}$)

- Για παράδειγμα, ο διοικητής ενός νοσοκομείου με βάση παρελθοντικά δεδομένα εκτιμά ότι κατά μέσο όρο η ηλικία των ατόμων που επισκέπτονται τα εξωτερικά ιατρεία του νοσοκομείου είναι τα 63 έτη

Διαστήματα εμπιστοσύνης



Σημειακή εκτίμηση

- Ο διοικητής ενός νοσοκομείου με βάση παρελθοντικά δεδομένα εκτιμά ότι κατά μέσο όρο η ηλικία των ατόμων που επισκέπτονται τα εξωτερικά ιατρεία του νοσοκομείου είναι τα 63 έτη

- Από την άλλη, ο διοικητής ενός νοσοκομείου με βάση παρελθοντικά δεδομένα εκτιμά ότι κατά μέσο όρο η ηλικία των ατόμων που επισκέπτονται τα εξωτερικά ιατρεία του νοσοκομείου είναι μεταξύ των 60 και 65 έτη

Διάστημα εμπιστοσύνης

- Εκτός από τη σημειακή εκτίμηση, μπορούμε επίσης να δημιουργήσουμε ένα διάστημα τιμών μέσα στο οποίο θα βρίσκεται η άγνωστη παράμετρος με έναν υψηλό «βαθμό εμπιστοσύνης»

- Δυο πράγματα πρέπει να ζητάμε από ένα δ.ε.

  - Να περιέχει την αληθινή τιμή της παραμέτρου ένα μεγάλο «ποσοστό φορών»

  - Να έχει όσο το δυνατόν μικρότερο μήκος

Διάστημα εμπιστοσύνης

- Βαθμός εμπιστοσύνης ονομάζεται το «ποσοστό φορών» που ένα δ.ε. (L,U) περιέχει την πραγματική τιμή της παραμέτρου θ. Συμβολίζεται με $100(1-a)\%$ και φανερώνει την πιθανότητα το δ.ε. να περιέχει την πραγματική τιμή της παραμέτρου θ.

$$\text{βαθμός εμπιστοσύνης} = P(L < \theta < U) = 1 - a$$

- a = *συντελεστής εμπιστοσύνης*
- a = *5% ή 10% ή 1%*

Κατασκευή Διαστήματος Εμπιστοσύνης

Τα περισσότερα διαστήματα εμπιστοσύνης έχουν την παρακάτω μορφή:

Δειγματική στατιστική συνάρτηση ± *(τυπικό σφάλμα στατιστικής συνάρτησης)(κρίσιμη τιμή)*

Περιθώριο σφάλματος

Κατασκευή Διαστήματος Εμπιστοσύνης

Δειγματική

στατιστική ± (τυπικό σφάλμα στατιστικής συνάρτησης)(κρίσιμη τιμή)

συνάρτηση

Διάστημα εμπιστοσύνης για τη μέση τιμή με γνωστή διακύμανση:

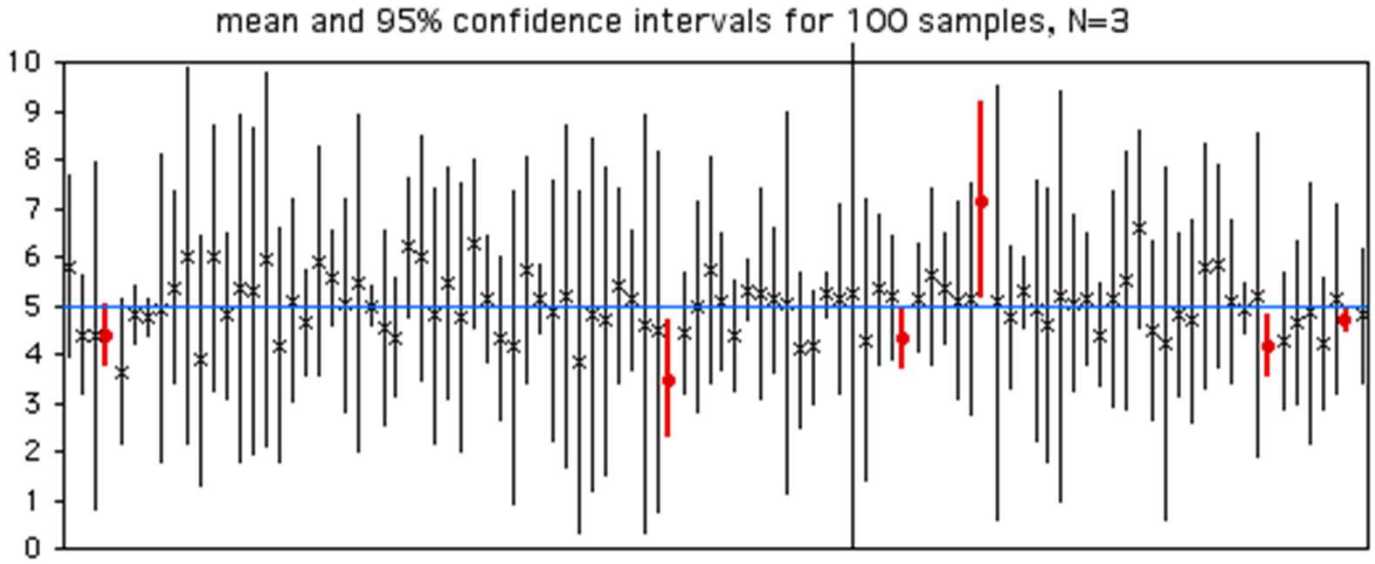
$$\overline{X} \pm \frac{\sigma}{\sqrt{n}} z_{a/2}$$

Διάστημα εμπιστοσύνης για τη μέση τιμή με άγνωστη διακύμανση:

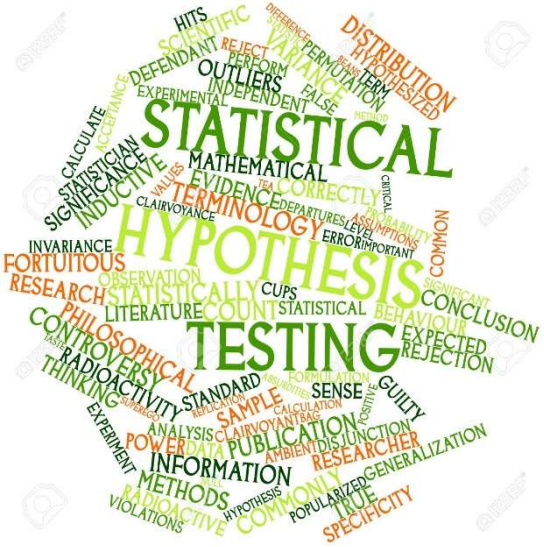
$$\overline{X} \pm \frac{S}{\sqrt{n}} t_{n-1,a/2}$$

Διάστημα εμπιστοσύνης

- **Ερμηνεία δ.ε.**
- *Ένα 95% δ.ε. για μια παράμετρο θ ενός πληθυσμού σημαίνει ότι εάν πάρουμε 100 δείγματα και κατασκευάσουμε 100 δ.ε. τότε τα 95 από αυτά θα περιέχουν την αληθινή τιμή της παραμέτρου θ και τα 5 όχι*



Έλεγχοι Υποθέσεων



Έλεγχοι Υποθέσεων

- Μας απασχολεί μια υπόθεση (ερώτημα) η οποία αφορά μια παράμετρο του πληθυσμού, όπως
  - η πληθυσμιακή μέση τιμή:

Το μέσο βάρος ενός φυσιολογικού μωρού κατά τον τοκετό είναι 3,45 κιλά, δηλαδή, η παράμετρος μ λαμβάνει την τιμή 3,45 (μ=3,45);

  - το ποσοστό στον πληθυσμό:

Το ποσοστό των εφήβων στην Ελλάδα που έχει κάνει χρήση ναρκωτικών είναι 12%, δηλαδή, η παράμετρος p λαμβάνει την τιμή 0,12 (p=0,12);

Βασικές έννοιες του στατιστικού ελέγχου υποθέσεων

- **H_1 ή H_α: εναλλακτική υπόθεση (alternative hypothesis)** – η υπόθεση που δηλώνει το ερευνητικό ερώτημα
  - Το νέο φάρμακο ΕΙΝΑΙ πιο αποτελεσματικό από το παλιό
  - Το κάπνισμα ΠΡΟΚΑΛΕΙ καρκίνο του πνεύμονα
- **H_0: μηδενική υπόθεση (null hypothesis)** – η υπόθεση που εκφράζει ένα «συντηρητικό» ερώτημα (το αντίθετο από την εναλλακτική)
  - Το νέο φάρμακο ΔΕΝ ΕΙΝΑΙ πιο αποτελεσματικό από το παλιό
  - Το κάπνισμα ΔΕΝ ΠΡΟΚΑΛΕΙ καρκίνο του πνεύμονα
- **α: επίπεδο σημαντικότητας** – η πιθανότητα να απορρίψουμε λανθασμένα την μηδενική υπόθεση
- **p-value:** εκφράζει την πιθανότητα να παρατηρήσουμε το αποτέλεσμα που παρατηρήθηκε ή κάποιο πιο ακραίο αποτέλεσμα, όταν ισχύει η μηδενική υπόθεση

Αξιολόγηση των στατιστικών ελέγχων

Πιθανές Εκβάσεις ενός ελέγχου		Υπόθεση	
		Ορθή (H_0)	*Λανθασμένη* (H_1)
Απόφαση	*Αποδοχή* (H_0)	*Σωστό*	*Λάθος*
	Απόρριψη (H_1)	*Λάθος*	*Σωστό*

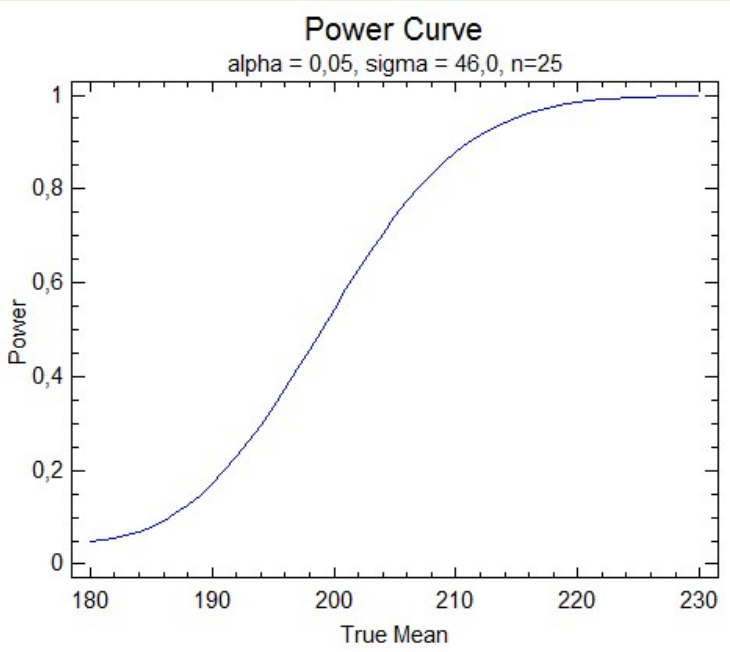
- Η άγνωστη πραγματικότητα:
  - H_0: Η υπόθεση είναι ορθή
  - H_1: Η υπόθεση είναι λανθασμένη
- Η απόφαση:
  - H_0: Η υπόθεση είναι ορθή
  - H_1: Η υπόθεση είναι λανθασμένη
- Υπάρχει περίπτωση να κάνουμε τα εξής σφάλματα:
  - Απορρίπτουμε μια ορθή υπόθεση
  - Αποδεχόμαστε μια λανθασμένη υπόθεση

Αξιολόγηση των στατιστικών ελέγχων

Πιθανές Εκβάσεις ενός ελέγχου		**Υπόθεση**	
		Ορθή (H_o)	*Λανθασμένη* (H_1)
Απόφαση	*Αποδοχή* (H_o)	ο p=1-α	Σφάλμα τύπου ΙΙ p=β
	Απόρριψη (H_1)	Σφάλμα τύπου Ι p=α	ο p=1-β

- **α → Σφάλμα Τύπου Ι: Απορρίπτουμε την H_0 όταν η H_0 είναι ορθή**
- **β → Σφάλμα Τύπου ΙΙ: Δεν απορρίπτουμε την H_0 όταν η H_o είναι λάθος**
- **Ισχύς του τεστ: γ=1-β → Απορρίπτουμε την H_0 όταν η H_0 είναι λάθος**
- **Καθώς το α μικραίνει το β μεγαλώνει**

Ισχύς ελέγχου



- Η **ισχύς** (**power**) του ελέγχου υπόθεσης είναι η πιθανότητα

  γ=P(απόρριψη της H_0|H_0 λανθασμένη)

- Η ισχύς μπορεί να θεωρηθεί ως η πιθανότητα μια συγκεκριμένη μελέτη να διακρίνει μια απόκλιση από την μηδενική υπόθεση δεδομένου ότι αυτή η απόκλιση υπάρχει

- Η ποσότητα γ έχει διαφορετική τιμή για διαφορετική τιμή του μ_1 που ορίζει η εναλλακτική υπόθεση. Εάν σχεδιάσουμε τις τιμές του γ έναντι όλων των εναλλακτικών τιμών του μέσου του πληθυσμού, θα πάρουμε την καμπύλη ισχύος (power curve)

Έλεγχοι Υποθέσεων

Τα βήματα που ακολουθούνται για τον στατιστικό έλεγχο μιας υπόθεσης είναι τα εξής:

1. Διατυπώνουμε το ερευνητικό ερώτημα
2. Διατυπώνουμε την μηδενική υπόθεση H_0 (*null hypothesis*), την οποία, συχνά, θα προσπαθούμε να την απορρίψουμε, και την εναλλακτική υπόθεση H_1 ή H_α (*alternative hypothesis*)
3. Αποφασίζουμε το επίπεδο σημαντικότητας α του ελέγχου
4. Προσδιορίζουμε την στατιστική συνάρτηση στην οποία βασίζεται ο έλεγχος
5. Καθορίζουμε τον κανόνα σύμφωνα με τον οποίο απορρίπτουμε ή «δεχόμαστε» την H_0 (κρίσιμη περιοχή)
6. Υπολογίζουμε την στατιστική συνάρτηση
7. Αποφασίζουμε σχετικά με την απόρριψη της μηδενικής υπόθεσης
8. Ερμηνεύουμε το αποτέλεσμα

Έλεγχοι Υποθέσεων

- Συνήθως είναι προτιμότερο να απορρίπτουμε την H_o, καθώς όταν την απορρίπτουμε, η πιθανότητα σφάλματος είναι γνωστή (μικρότερη ή ίση ενός γνωστού ορίου a)
- Καλό είναι να μην λέμε ότι αποδεχόμαστε την H_o αλλά ότι δεν υπάρχουν αρκετές ενδείξεις ώστε να απορρίψουμε την μηδενική υπόθεση
- Όταν απορρίπτουμε την H_o λέμε ότι ο έλεγχος είναι **στατιστικά σημαντικός**

Έλεγχοι Υποθέσεων

- ***p*-value** ενός στατιστικού ελέγχου είναι η μικρότερη τιμή του a για την οποία απορρίπτουμε την H_0 με βάση τα δεδομένα που έχουμε παρατηρήσει

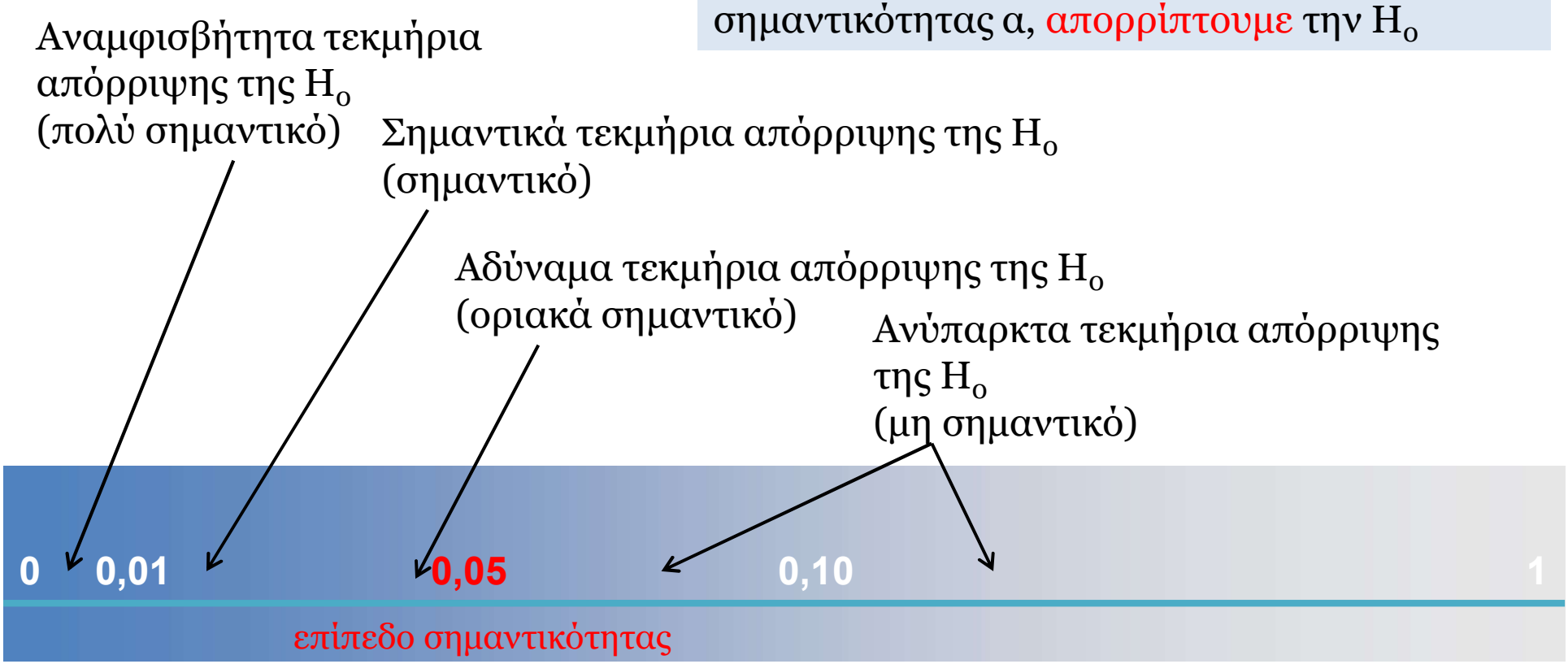
$$p-value = p(x) = P(T(X) \geq T(X_0) \mid H_0),$$

όπου $T(X_0)$ είναι η τιμή της στατιστικής συνάρτησης που καθορίζει την κρίσιμη περιοχή

- Το p-value (**παρατηρούμενο επίπεδο σημαντικότητας**) εκφράζει την πιθανότητα η στατιστική συνάρτηση να λάβει πιο ακραία τιμή από αυτή που πήρε όταν εφαρμόσθηκε στο διαθέσιμο δείγμα
- Διαισθητικά, το p-value είναι ένα μέτρο της σιγουριάς μας για την ισχύ της μηδενικής υπόθεσης
- Το p-value ενός στατιστικού ελέγχου είναι τυχαία μεταβλητή

Κανόνες απόρριψης H_o

Εάν το p-value είναι μικρότερο από το επίπεδο σημαντικότητας α, απορρίπτουμε την H_o

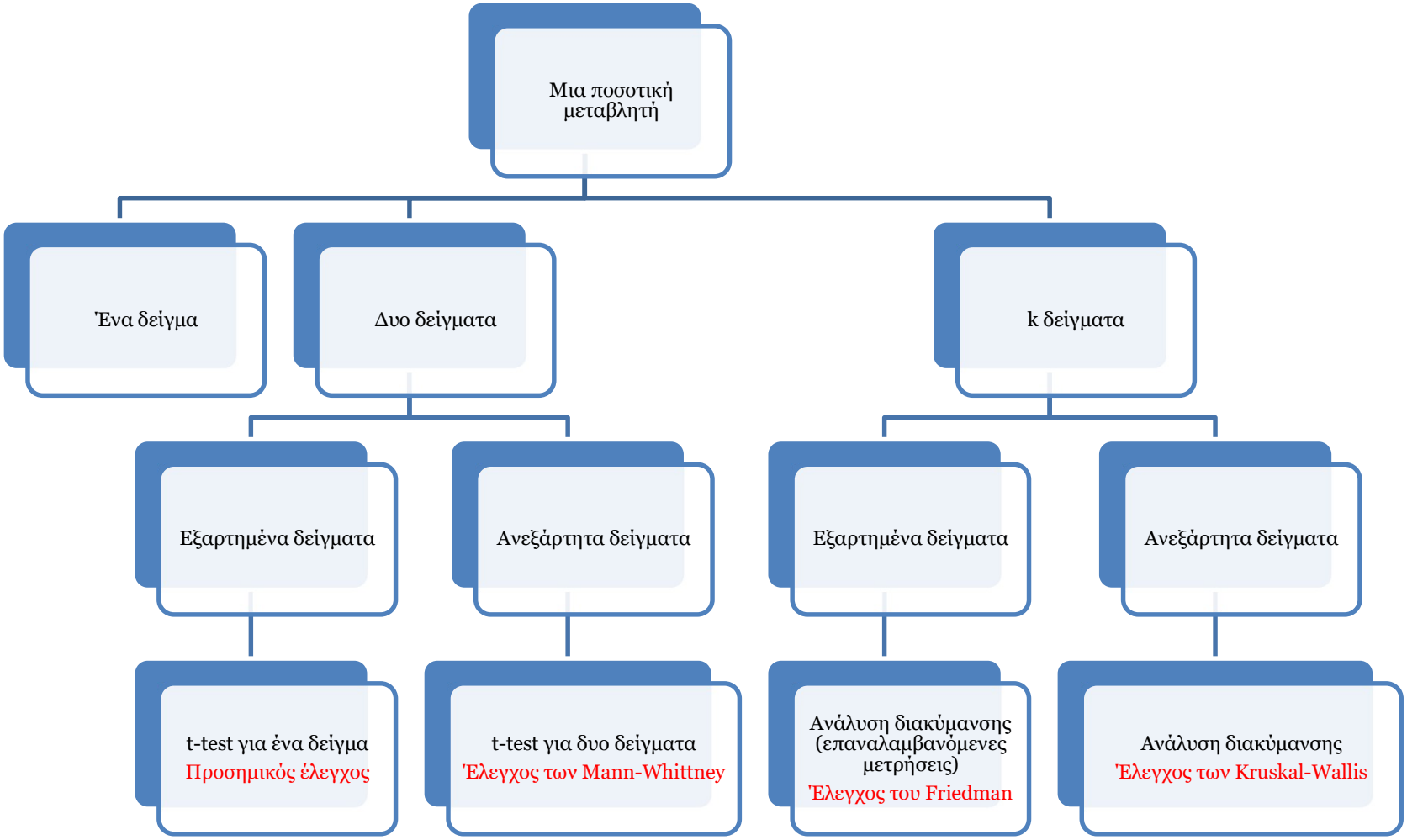


Κατηγορίες ελέγχων υποθέσεων

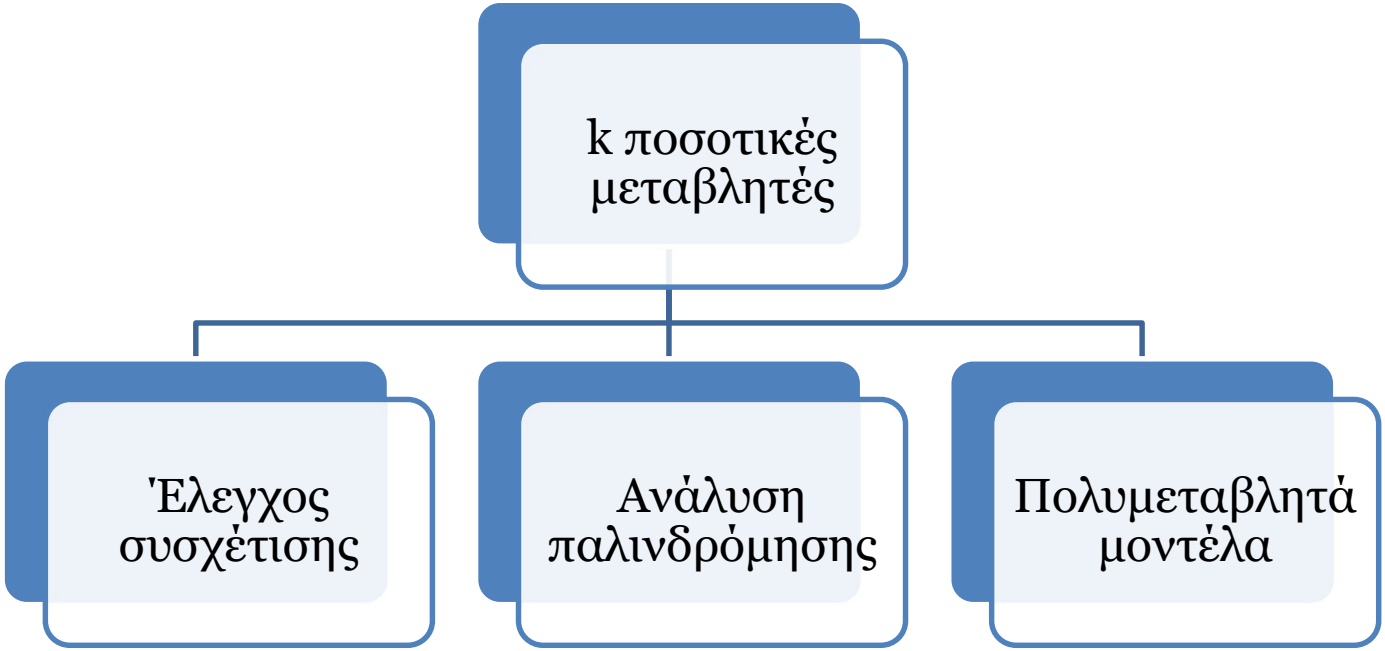
- Οι **παραμετρικοί έλεγχοι** χρησιμοποιούνται όταν η μεταβλητή X που μελετάμε ακολουθεί γνωστή κατανομή (συνήθως την κανονική)

- Οι **μη παραμετρικοί έλεγχοι** χρησιμοποιούνται όταν η μεταβλητή X που μελετάμε ακολουθεί μια άγνωστη κατανομή

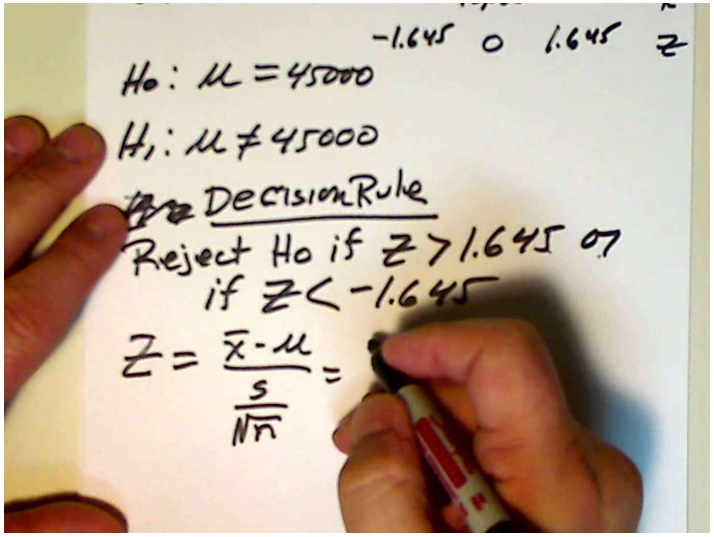
Κατηγορίες ελέγχων υποθέσεων



Κατηγορίες ελέγχων υποθέσεων



Παραμετρικοί έλεγχοι για ένα δείγμα



Έλεγχος για τη μέση τιμή – ένα δείγμα

- Ξεκινάμε με τον ισχυρισμό ότι η μέση τιμή του πληθυσμού ισούται με κάποια προκαθορισμένη τιμή μ_0
- Η δήλωση αυτή καλείται μηδενική υπόθεση ή H_0

$$H_0 : \mu = \mu_0$$

- Η εναλλακτική υπόθεση H_1 ή H_a είναι μια δεύτερη δήλωση (ισχυρισμός) που αντιπαρατίθεται στην H_0

$$H_1 : \mu \neq \mu_0 \longrightarrow \text{Αμφίπλευρος έλεγχος}$$

$$\left.\begin{array}{l} H_1 : \mu < \mu_0 \\ H_1 : \mu > \mu_0 \end{array}\right\} \longrightarrow \text{Μονόπλευροι έλεγχοι}$$

Παράδειγμα

Η μέση διαστολική πίεση στον πληθυσμό των ανδρών που έχουν μια συγκεκριμένη ασθένεια είναι ίση με 77,5mm Hg. Προκειμένου να ελέγξουμε αν η λήψη ενός νέου φαρμάκου **επηρεάζει** τη διαστολική πίεσή τους, ελήφθη τυχαίο δείγμα 30 ανδρών και τους χορηγήθηκε το σκεύασμα. Στη συνέχεια καταγράφηκε η διαστολική τους πίεση.

Υπόθεση προς έλεγχο:

Ερώτημα: Η λήψη του νέου φαρμάκου μεταβάλλει (είτε προς τα επάνω είτε προς τα κάτω) τη διαστολική πίεση;

$$H_0 : \mu = 77.5 - H_1 : \mu \neq 77.5$$

Αμφίπλευρος έλεγχος

Ερώτημα: Η λήψη του νέου φαρμάκου ελαττώνει τη διαστολική πίεση;

$$H_0 : \mu \geq 77.5 - H_1 : \mu < 77.5$$

Μονόπλευρος έλεγχος

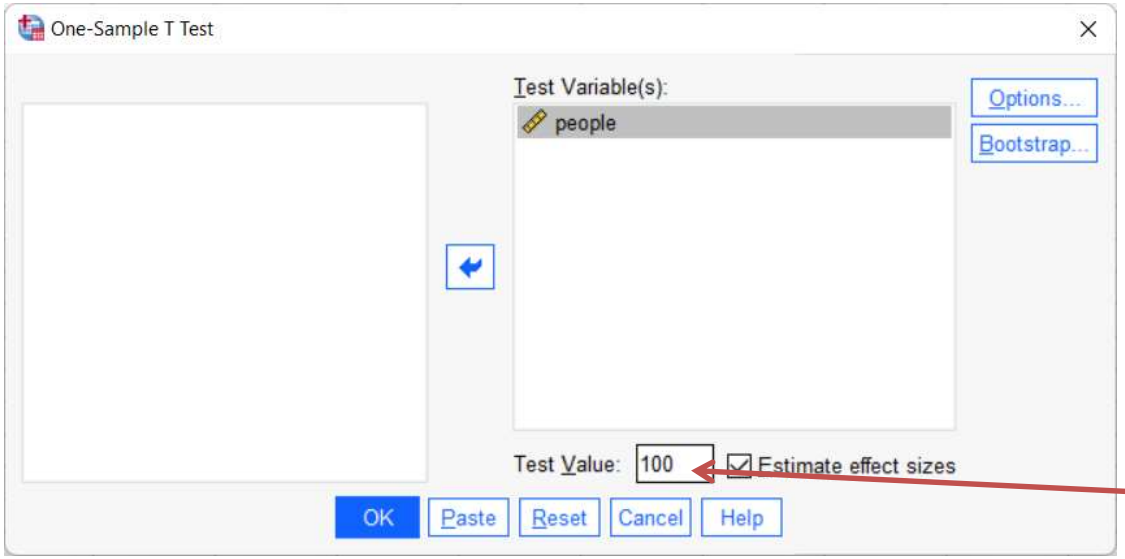
Ερώτημα: Η λήψη του νέου φαρμάκου αυξάνει τη διαστολική πίεση;

$$H_0 : \mu \leq 77.5 - H_1 : \mu > 77.5$$

Μονόπλευρος έλεγχος

Παράδειγμα

Η διοίκηση ενός νοσοκομείου θέλει να μελετήσει την προσέλευση ασθενών στα εξωτερικά ιατρεία του νοσοκομείου. Η διοίκηση πιστεύει ότι καθημερινά επισκέπτονται τα εξωτερικά ιατρεία 100 άτομα. Προκειμένου η διοίκηση να ελέγξει εάν ο αριθμός των ατόμων που επισκέπτονται τα εξωτερικά ιατρεία είναι 100, αποφάσισε να λάβει ένα τυχαίο δείγμα 40 ημερών.



Analyze → Compare Means and Proportions → One-Sample T Test

Δηλώνουμε ως Test Value την τιμή 100

$$H_0 : \mu = 100 - H_1 : \mu \neq 100$$

Παράδειγμα

Η διοίκηση ενός νοσοκομείου θέλει να μελετήσει την προσέλευση ασθενών στα εξωτερικά ιατρεία του νοσοκομείου. Η διοίκηση πιστεύει ότι καθημερινά επισκέπτονται τα εξωτερικά ιατρεία 100 άτομα. Προκειμένου η διοίκηση να ελέγξει εάν ο αριθμός των ατόμων που επισκέπτονται τα εξωτερικά ιατρεία είναι 100, αποφάσισε να λάβει ένα τυχαίο δείγμα 40 ημερών.

One-Sample Test

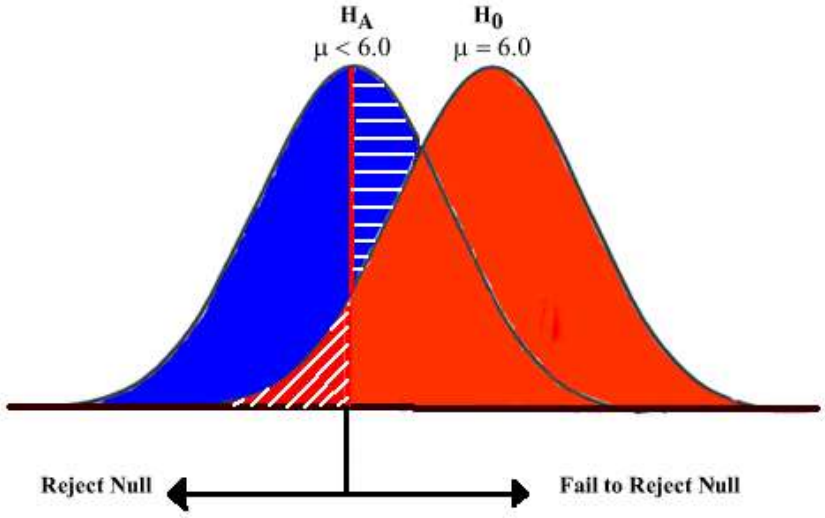
Test Value = 100

	t	df	Significance One-Sided p	Significance Two-Sided p	Mean Difference	95% Confidence Interval of the Difference Lower	95% Confidence Interval of the Difference Upper
people	1,443	39	,078	,157	6,900	-2,77	16,57

Παρατήρηση: Επειδή το p-value = 0,157 > 0,05 συμπεραίνουμε ότι ο αριθμός των ατόμων που επισκέπτονται τα εξωτερικά ιατρεία είναι 100

95% διάστημα εμπιστοσύνης για τον μέσο αριθμό επισκεπτών (100–2,77, 100+16,57)

Παραμετρικοί έλεγχοι για δύο δείγματα



Έλεγχοι για δυο δείγματα

- Παρατηρούμε ένα χαρακτηριστικό (μεταβλητή) σε δυο διαφορετικούς πληθυσμούς και μας ενδιαφέρει να ελέγξουμε εάν υπάρχει διαφορά μεταξύ των δυο πληθυσμών ή όχι (δηλαδή, εάν οι δυο πληθυσμοί συμπεριφέρονται με τον ίδιο τρόπο ως προς το συγκεκριμένο χαρακτηριστικό)

- Εάν συμβολίσουμε με
  - X_1 το χαρακτηριστικό (τυχαία μεταβλητή) για τον πρώτο πληθυσμό και με μ_1 τη μέση του τιμή
  - X_2 το χαρακτηριστικό (τυχαία μεταβλητή) για τον δεύτερο πληθυσμό και με μ_2 τη μέση του τιμή

ένας συνήθης (δίπλευρος) έλεγχος δυο δειγμάτων είναι ο εξής:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Έλεγχοι για δυο δείγματα

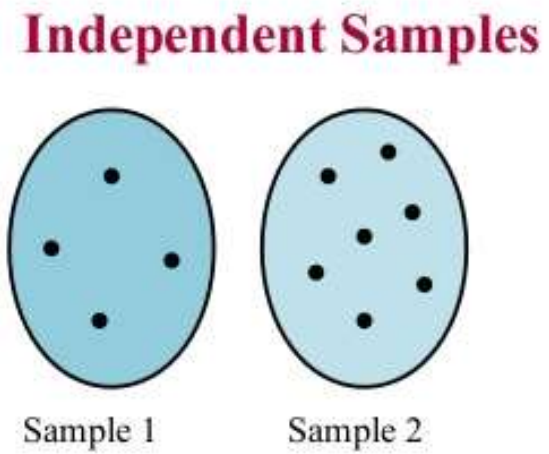
- Όταν τα άτομα του ενός πληθυσμού δεν επηρεάζουν τα άτομα του δεύτερου (ή/και αντίστροφα), τότε λέμε ότι οι δυο πληθυσμοί είναι ανεξάρτητοι μεταξύ τους. Τα δείγματα που προέρχονται από αυτούς τους πληθυσμούς θεωρούνται ανεξάρτητα.

- Όταν οι τυχαίες μεταβλητές X_1 και X_2 αναφέρονται στα ίδια άτομα (αφορούν, συνήθως, παρατηρήσεις σε διαφορετικούς χρόνους) τότε έχουμε ζευγαρωτές παρατηρήσεις (paired samples) και τα δείγματα θεωρούνται εξαρτημένα. Το ίδιο ισχύει επίσης όταν τα άτομα του πρώτου πληθυσμού επηρεάζουν τα άτομα του δεύτερου (ή/και αντίστροφα).

  - Στην περίπτωση αυτή βασίζουμε τα συμπεράσματά μας στις διαφορές $X_1 - X_2$ και χρησιμοποιούμε τις τεχνικές που είναι διαθέσιμες για ένα δείγμα, ελέγχοντας τις υποθέσεις

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0$$

Παραμετρικοί έλεγχοι για δύο **ανεξάρτητα** δείγματα



Έλεγχοι για δυο ανεξάρτητα δείγματα

- Έχουμε δυο **ανεξάρτητα** δείγματα

- Θέλουμε να αξιολογήσουμε κατά πόσον η μέση τιμή του ενός πληθυσμού (από τον οποίο προέρχεται το ένα δείγμα) είναι ίση ή όχι με τη μέση τιμή ενός δεύτερου πληθυσμού (από τον οποίο προέρχεται το άλλο δείγμα)

- Παραδείγματα
  - Σύγκριση της αποτελεσματικότητας δυο φαρμάκων (φάρμακο και placebo)
  - Σύγκριση του BMI μεταξύ ανδρών και γυναικών
  - Σύγκριση χρόνου αντίδρασης οδηγών που μιλούν στο τηλέφωνο και αυτών που δεν μιλούν στο τηλέφωνο

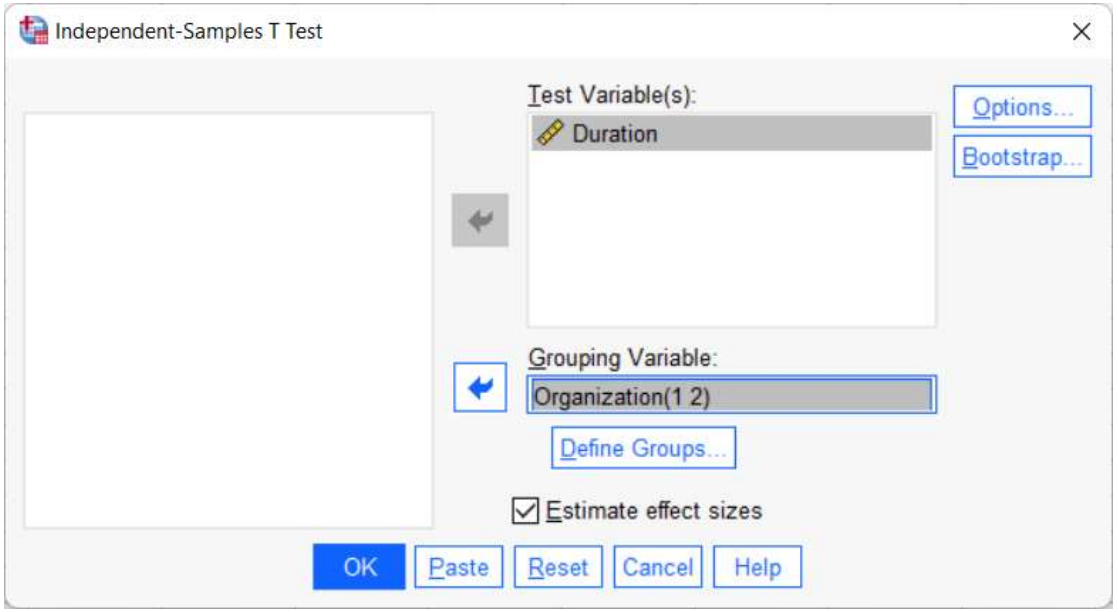
Έλεγχοι για δυο ανεξάρτητα δείγματα

- Στα **ανεξάρτητα** δείγματα υπάρχουν διαφορετικά άτομα
- Στα άτομα αυτά **δεν** υπάρχει κάποια φυσική εξάρτηση μεταξύ τους
- Στα **ανεξάρτητα** δείγματα, οι παρατηρήσεις στο ένα δείγμα **δεν** επηρεάζουν ή **δεν** επηρεάζονται από τις παρατηρήσεις στο άλλο δείγμα
- Έχουμε σχεδιασμό μεταξύ υποκειμένων (between subject design) – μας ενδιαφέρουν οι διαφορές μεταξύ των ατόμων

Παράδειγμα

Η διοίκηση ενός νοσοκομείου θέλει να ελέγξει εάν οι ασφαλιστικοί φορείς Α και Β καλύπτουν στον ίδιο χρόνο τις υποχρεώσεις τους προς το νοσοκομείο. Η διοίκηση του νοσοκομείου επέλεξε τυχαία 32 ασθενείς, εκ των οποίων οι 16 είναι ασφαλισμένοι στον φορέα Α και οι υπόλοιποι 16 στον φορέα Β. Για κάθε ασθενή κατέγραψε το χρόνο σε ημέρες που απαιτήθηκαν, μέχρι ο φορέας να καλύψει τα έξοδα νοσηλείας.

$$H_0 : \mu_A = \mu_B - H_1 : \mu_A \neq \mu_B$$



Analyze → Compare Means and Proportions → Independent-Samples T Test

Παράδειγμα

Η διοίκηση ενός νοσοκομείου θέλει να ελέγξει εάν οι ασφαλιστικοί φορείς Α και Β καλύπτουν στον ίδιο χρόνο τις υποχρεώσεις τους προς το νοσοκομείο. Η διοίκηση του νοσοκομείου επέλεξε τυχαία 32 ασθενείς, εκ των οποίων οι 16 είναι ασφαλισμένοι στον φορέα Α και οι υπόλοιποι 16 στον φορέα Β. Για κάθε ασθενή κατέγραψε το χρόνο σε ημέρες που απαιτήθηκαν, μέχρι ο φορέας να καλύψει τα έξοδα νοσηλείας.

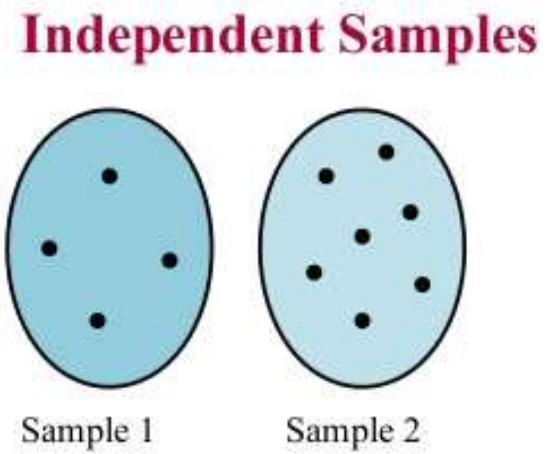
Έλεγχος ισότητας διακυμάνσεων: Επειδή το p-value = 0,359 > 005 συμπεραίνουμε ότι οι δυο πληθυσμοί έχουν ίσες διακυμάνσεις

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means							
						Significance				95% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference	Lower	Upper
Duration	Equal variances assumed	,867	,359	,227	30	,411	,822	4,438	19,522	-35,431	44,306
	Equal variances not assumed			,227	29,083	,411	,822	4,438	19,522	-35,484	44,359

Παρατήρηση: Επειδή το p-value = 0,822 > 0,05 δεν μπορούμε να πούμε ότι ο χρόνος αποπληρωμής μεταξύ των δυο ασφαλιστικών φορέων είναι διαφορετικός

Παραμετρικοί έλεγχοι για δύο **εξαρτημένα** δείγματα



Έλεγχοι για δυο εξαρτημένα δείγματα

- Έχουμε δυο **εξαρτημένα** δείγματα

- Θέλουμε να μελετήσουμε κατά πόσο η μέση τιμή ενός χαρακτηριστικού των ίδιων ατόμων έχει μεταβληθεί μεταξύ δυο χρονικών στιγμών ή να συγκρίνουμε ένα χαρακτηριστικό μεταξύ ατόμων που παρουσιάζουν φυσική εξάρτηση

- Παραδείγματα
  - Σύγκριση της αποτελεσματικότητας μιας δίαιτας στα ίδια άτομα
  - Σύγκριση του IQ δίδυμων αδελφών

Έλεγχοι για δυο εξαρτημένα δείγματα

- Πότε έχουμε **εξαρτημένα** δείγματα;
  - Στον σχεδιασμό «πριν – μετά», επειδή μελετώνται τα ίδια άτομα
  - Στον σχεδιασμό με δυο ομάδες ατόμων που παρουσιάζουν φυσική εξάρτηση (οι παρατηρήσεις στο ένα δείγμα επηρεάζουν ή/και επηρεάζονται από τις παρατηρήσεις στο άλλο δείγμα)

Έλεγχοι για δυο εξαρτημένα δείγματα

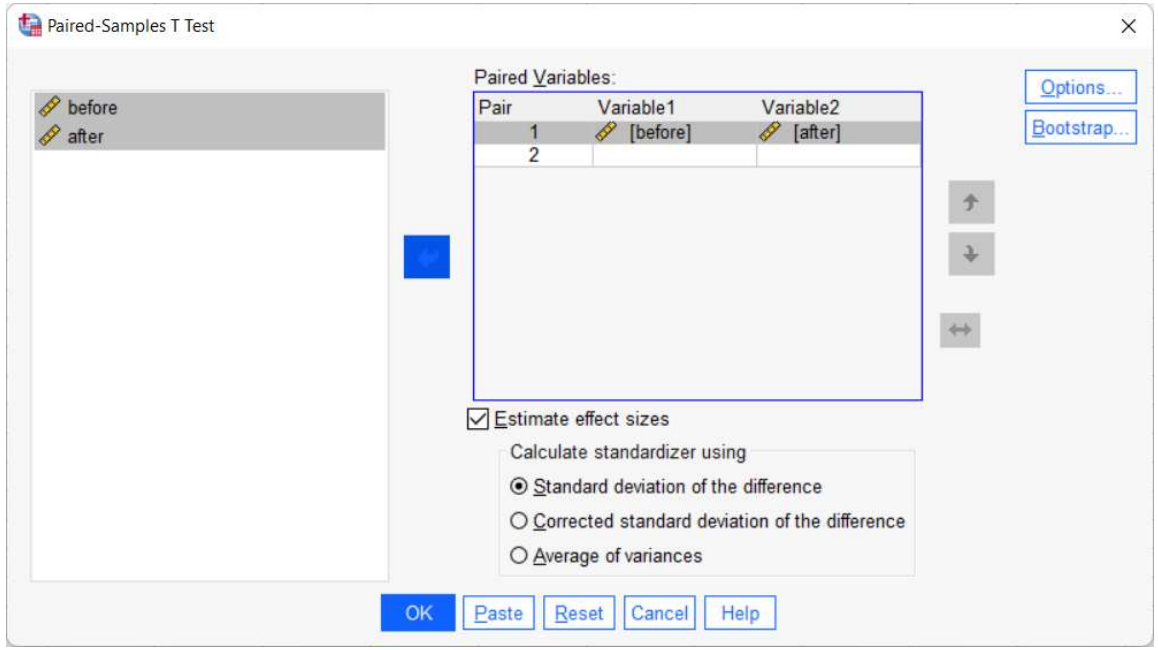
- Υποθέσεις του ελέγχου t για δυο εξαρτημένα δείγματα

- Οι υποθέσεις αφορούν τις διαφορές

    1. Οι διαφορές είναι μεταξύ τους ανεξάρτητες
    2. Οι διαφορές ακολουθούν την κανονική κατανομή
    3. Τα δυο δείγματα έχουν το ίδιο μέγεθος

Παράδειγμα

Ο διοικητής ενός νοσηλευτικού ιδρύματος θέλει να ελέγξει εάν ένα νέο πρόγραμμα εφημεριών του νοσηλευτικού προσωπικού έχει επιτυχία ή όχι. Ο διοικητής του νοσηλευτικού ιδρύματος επέλεξε τυχαία 15 νοσηλευτές και τους έδωσε να συμπληρώσουν ένα ερωτηματολόγιο πριν την εφαρμογή του νέου προγράμματος. Από τις απαντήσεις που έδωσαν, υπολόγισε ένα συνολικό σκορ από το 0 έως το 100. Στη συνέχεια, και αφού οι νοσηλευτές δοκίμασαν το νέο πρόγραμμα, ξανασυμπλήρωσαν το ερωτηματολόγιο και υπολογίστηκε ξανά το σκορ.



$$H_0 : \mu_1 = \mu_2 - H_1 : \mu_1 \neq \mu_2$$

Analyze → Compare Means and Proportions → Paired-Samples T Test

Παράδειγμα

Ο διοικητής ενός νοσηλευτικού ιδρύματος θέλει να ελέγξει εάν ένα νέο πρόγραμμα εφημεριών του νοσηλευτικού προσωπικού έχει επιτυχία ή όχι. Ο διοικητής του νοσηλευτικού ιδρύματος επέλεξε τυχαία 15 νοσηλευτές και τους έδωσε να συμπληρώσουν ένα ερωτηματολόγιο πριν την εφαρμογή του νέου προγράμματος. Από τις απαντήσεις που έδωσαν, υπολόγισε ένα συνολικό σκορ από το 0 έως το 100. Στη συνέχεια, και αφού οι νοσηλευτές δοκίμασαν το νέο πρόγραμμα, ξανασυμπλήρωσαν το ερωτηματολόγιο και υπολογίστηκε ξανά το σκορ.

Paired Samples Test

		Paired Differences					t	df	Significance	
					95% Confidence Interval of the Difference					
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			One-Sided p	Two-Sided p
Pair 1	before - after	-3,467	5,041	1,302	-6,258	-,675	-2,664	14	,009	,019

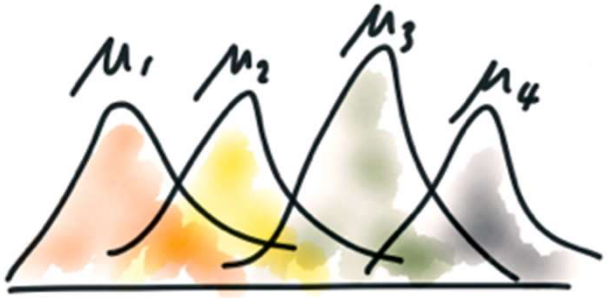
Παρατήρηση: Επειδή το p-value = 0,019 < 0,05 συμπεραίνουμε ότι νοσηλευτικό προσωπικό έχει διαφορετικό σκορ πριν και μετά από το νέο πρόγραμμα

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	before	78,87	15	10,267	2,651
	after	82,33	15	9,209	2,378

Το σκορ μετά από το πρόγραμμα είναι μεγαλύτερο από το σκορ πριν από το πρόγραμμα

Παραμετρικοί έλεγχοι για k **ανεξάρτητα** δείγματα



ANOVA

$\mu_1 = \mu_2 = \mu_3 = \mu_4$?

Έλεγχοι για k>2 ανεξάρτητα δείγματα

Στόχος
Να αξιολογήσουμε κατά πόσον οι μέσες τιμές των πληθυσμών από τους οποίους προέρχονται τα k>2 δείγματα είναι ίσες ή όχι (κάποια μέση τιμή διαφέρει από τις άλλες)

Ο έλεγχος που χρησιμοποιούμε ονομάζεται **Ανάλυση Διασποράς** (Analysis of Variance – ANOVA)

Παραδείγματα ANOVA

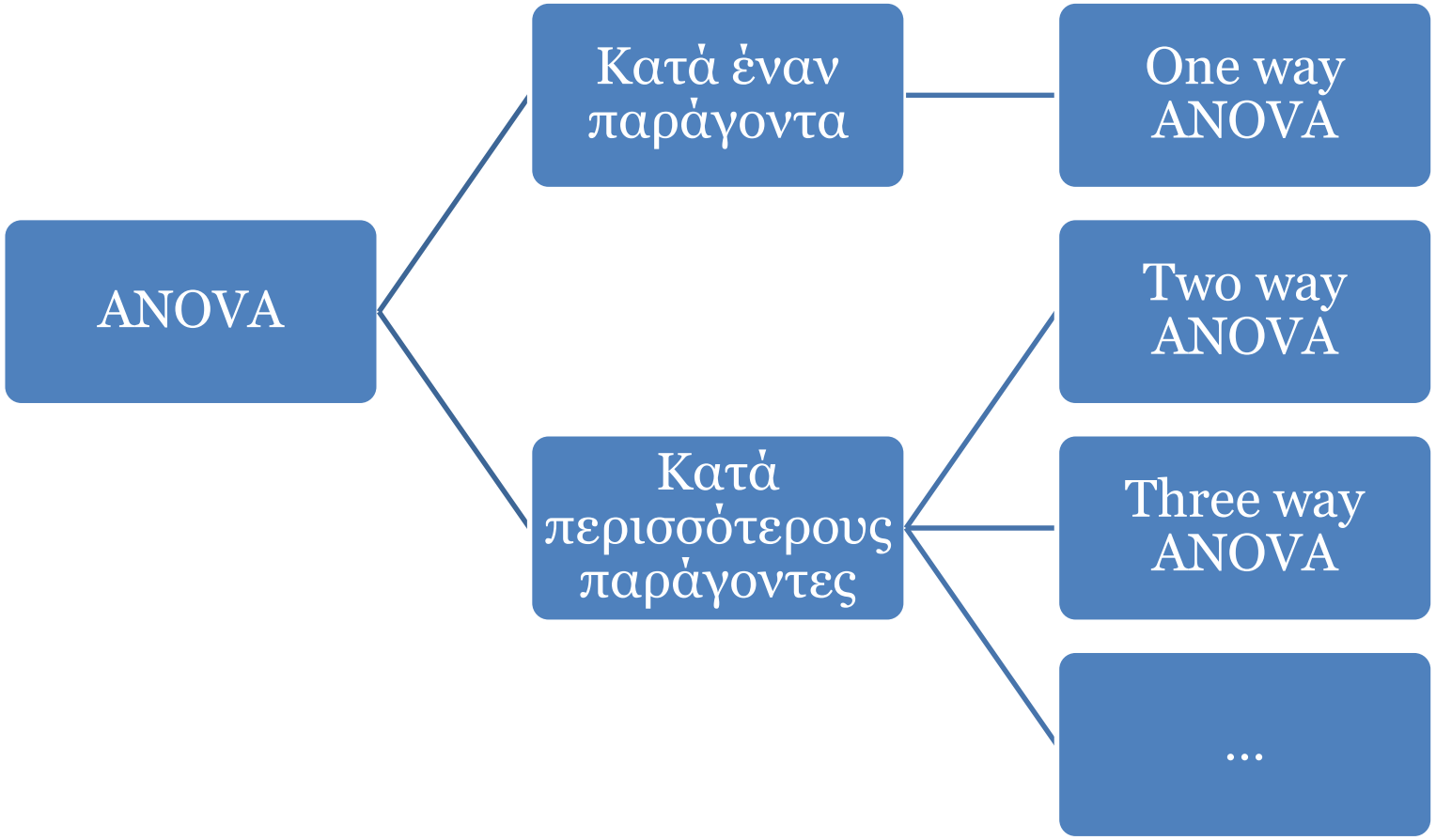
Παράδειγμα 1. Για μια συγκεκριμένη ασθένεια χρησιμοποιείται παραδοσιακά ένα φάρμακο A. Δύο νέα φάρμακα B και Γ εμφανίζονται στην αγορά για την καταπολέμηση της ασθένειας.

Ερώτημα: *Είναι τα τρία φάρμακα ισοδύναμα ως προς την καταπολέμηση της ασθένειας; (βιοϊσοδυναμία). Εάν όχι ποιο είναι το πιο αποτελεσματικό*;

Παράδειγμα 2. Η διοίκηση ενός μεγάλου νοσοκομείου θέλει να μελετήσει κατά πόσο ο χρόνος απουσίας των εργαζομένων (Y) έχει σχέση με το φύλο (X_1), την οικογενειακή κατάσταση (X_1) και το είδος της εργασίας που έχει ανατεθεί στον εργαζόμενο (X_3).

Ερώτημα: *Ποιοι από τους 3 παράγοντες επηρεάζουν το Y, ποιος είναι πιο σημαντικός, κτλ.*

Ανάλυση κατά έναν ή περισσότερους παράγοντες



Ανάλυση κατά έναν ή περισσότερους παράγοντες

Παράγοντας – ποιοτική μεταβλητή (ονομαστικής ή τακτικής κλίμακας)

Παραδείγματα:

1. Ανάλυση κατά έναν παράγοντα
   - X : χρόνος απόκρισης ασθενούς σε ένα φάρμακο
   - Παράγοντας: Φάρμακο (A, B, Γ)

Εξαρτημένη μεταβλητή

Ανεξάρτητη μεταβλητή

2. Ανάλυση κατά τρεις παράγοντες
   - X : χρόνος απόκρισης ασθενούς σε ένα φάρμακο
   - Παράγοντας 1: Φάρμακο (A, B, Γ)
   - Παράγοντας 2: Φύλο (Άνδρας, Γυναίκα)
   - Παράγοντας 3: Ηλικιακή ομάδα (νέος, μεσήλικας, ηλικιωμένος)

Εξαρτημένη μεταβλητή

Ανεξάρτητες μεταβλητές

Ανάλυση Διακύμανσης κατά έναν παράγοντα

- Μια εξαρτημένη μεταβλητή
  - ποσοτική μεταβλητή
- Ένας παράγοντας
  - μια ανεξάρτητη μεταβλητή
  - μια ποιοτική μεταβλητή
  - k επίπεδα

Ανάλυση Διασποράς

- Η **μηδενική υπόθεση** είναι ότι δεν υπάρχει σημαντική διαφορά μεταξύ των αριθμητικών μέσων τιμών (H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$)

- Στην αντίθετη περίπτωση μπορεί να διαφέρουν όλες οι μέσες τιμές
- Στην αντίθετη περίπτωση μπορεί να διαφέρουν ορισμένες από τις μέσες τιμές
- Στην αντίθετη περίπτωση μπορεί να διαφέρουν μόνο δυο από τις μέσες τιμές

- Η **εναλλακτική υπόθεση** είναι ότι υπάρχει τουλάχιστον μία σημαντική διαφορά μεταξύ των μέσων τιμών

Παράδειγμα

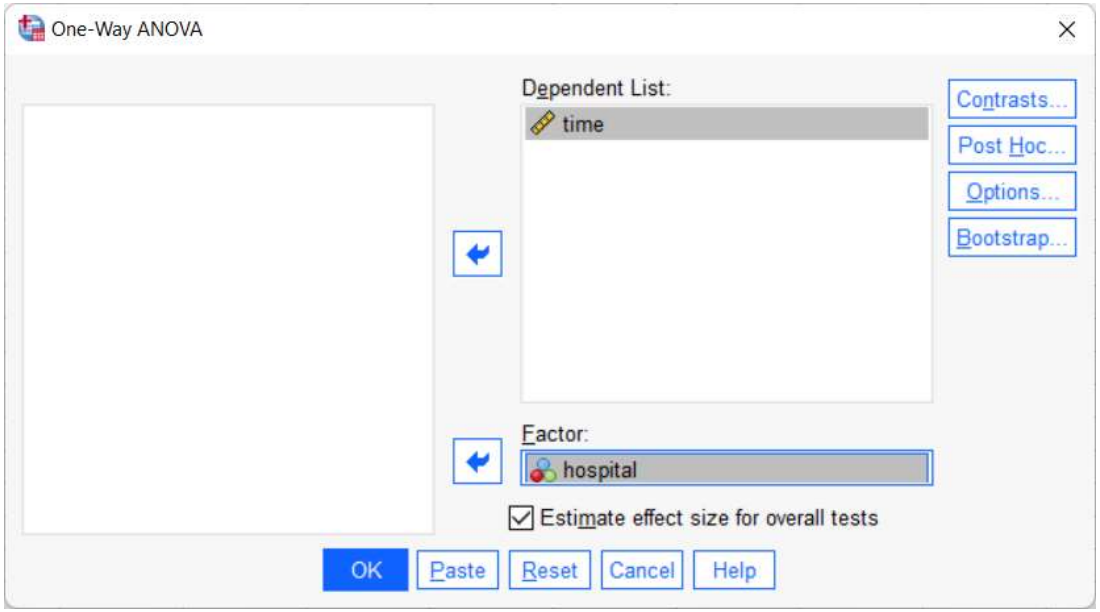
Η διοίκηση ενός ιατρικού ομίλου ενδιαφέρεται να μελετήσει τον χρόνο παραμονής των ασθενών σε τέσσερα νοσηλευτικά ιδρύματα που ανήκουν στον όμιλο. Προκειμένου να πάρει απάντηση στο πρόβλημα, η διοίκηση του ιατρικού ομίλου κατέγραψε τον χρόνο παραμονής σε ημέρες στο νοσηλευτικό ίδρυμα 11 ασθενών από το ίδρυμα Α, 9 ασθενών από το ίδρυμα Β, 10 ασθενών από το ίδρυμα Γ και 8 ασθενών από το ίδρυμα Δ.

Α	Β	Γ	Δ
5	7	12	6
6	7	8	7
6	8	10	8
6	5	13	6
8	6	12	7
8	7	14	7
10	10	11	9
7	6	15	8
7	7	13	
5		9	
9			

$$H_0 : \mu_A = \mu_B = \mu_\Gamma = \mu_\Delta \; - \; H_1 : \text{Διαφορετικά}$$

Παράδειγμα

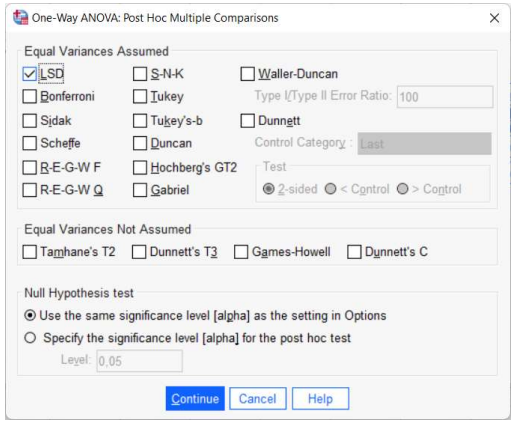
Η διοίκηση ενός ιατρικού ομίλου ενδιαφέρεται να μελετήσει τον χρόνο παραμονής των ασθενών σε τέσσερα νοσηλευτικά ιδρύματα που ανήκουν στον όμιλο. Προκειμένου να πάρει απάντηση στο πρόβλημα, η διοίκηση του ιατρικού ομίλου κατέγραψε τον χρόνο παραμονής σε ημέρες στο νοσηλευτικό ίδρυμα 11 ασθενών από το ίδρυμα Α, 9 ασθενών από το ίδρυμα Β, 10 ασθενών από το ίδρυμα Γ και 8 ασθενών από το ίδρυμα Δ.



Analyze → Compare Means and Proportions → One-Way ANOVA

Παράδειγμα

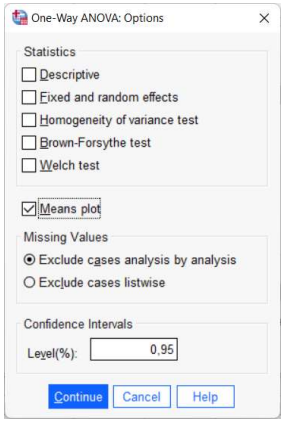
Η διοίκηση ενός ιατρικού ομίλου ενδιαφέρεται να μελετήσει τον χρόνο παραμονής των ασθενών σε τέσσερα νοσηλευτικά ιδρύματα που ανήκουν στον όμιλο. Προκειμένου να πάρει απάντηση στο πρόβλημα, η διοίκηση του ιατρικού ομίλου κατέγραψε τον χρόνο παραμονής σε ημέρες στο νοσηλευτικό ίδρυμα 11 ασθενών από το ίδρυμα Α, 9 ασθενών από το ίδρυμα Β, 10 ασθενών από το ίδρυμα Γ και 8 ασθενών από το ίδρυμα Δ.



Για να κάνουμε πολλαπλές συγκρίσεις:

...One-Way ANOVA → One-Way ANOVA: Post Hoc Multiple Comparisons

Ενεργοποιούμε την επιλογή LSD



Για να διαγραμματοποιήσουμε τους μέσους των υποομάδων:

...One-Way ANOVA → One-Way ANOVA: Options

Ενεργοποιούμε την επιλογή Means Plot

Παράδειγμα

Η διοίκηση ενός ιατρικού ομίλου ενδιαφέρεται να μελετήσει τον χρόνο παραμονής των ασθενών σε τέσσερα νοσηλευτικά ιδρύματα που ανήκουν στον όμιλο. Προκειμένου να πάρει απάντηση στο πρόβλημα, η διοίκηση του ιατρικού ομίλου κατέγραψε τον χρόνο παραμονής σε ημέρες στο νοσηλευτικό ίδρυμα 11 ασθενών από το ίδρυμα Α, 9 ασθενών από το ίδρυμα Β, 10 ασθενών από το ίδρυμα Γ και 8 ασθενών από το ίδρυμα Δ.

ANOVA

time

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	158,216	3	52,739	19,157	<,001
Within Groups	93,600	34	2,753		
Total	251,816	37			

Παρατήρηση: Επειδή το p-value είναι < 0,001 συμπεραίνουμε ότι υπάρχει διαφορά στη διάρκεια παραμονής

Παράδειγμα

Η διοίκηση ενός ιατρικού ομίλου ενδιαφέρεται να μελετήσει τον χρόνο παραμονής των ασθενών σε τέσσερα νοσηλευτικά ιδρύματα που ανήκουν στον όμιλο. Προκειμένου να πάρει απάντηση στο πρόβλημα, η διοίκηση του ιατρικού ομίλου κατέγραψε τον χρόνο παραμονής σε ημέρες στο νοσηλευτικό ίδρυμα 11 ασθενών από το ίδρυμα Α, 9 ασθενών από το ίδρυμα Β, 10 ασθενών από το ίδρυμα Γ και 8 ασθενών από το ίδρυμα Δ.

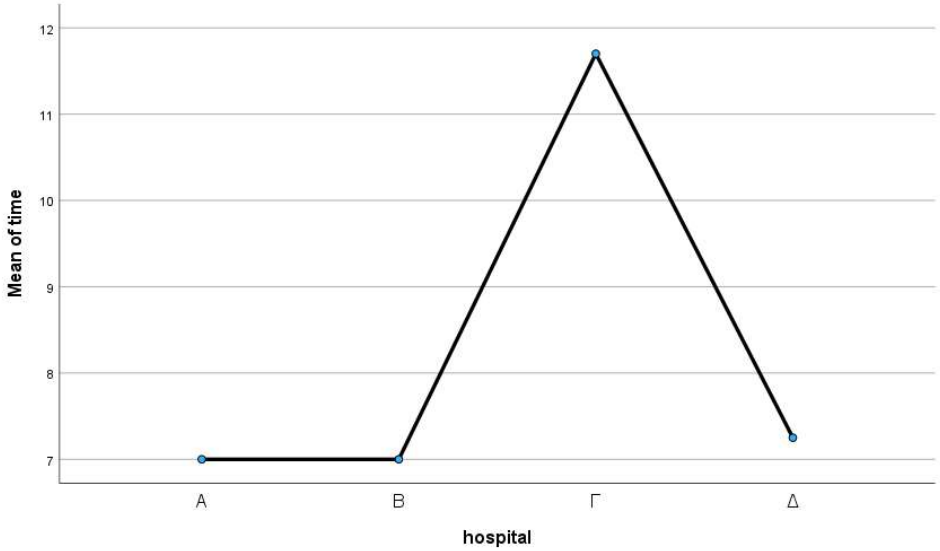
Multiple Comparisons

Dependent Variable: time

LSD

(I) hospital	(J) hospital	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
A	B	,000	,746	1,000	-1,52	1,52
	Γ	-4,700*	,725	<,001	-6,17	-3,23
	Δ	-,250	,771	,748	-1,82	1,32
B	A	,000	,746	1,000	-1,52	1,52
	Γ	-4,700*	,762	<,001	-6,25	-3,15
	Δ	-,250	,806	,758	-1,89	1,39
Γ	A	4,700*	,725	<,001	3,23	6,17
	B	4,700*	,762	<,001	3,15	6,25
	Δ	4,450*	,787	<,001	2,85	6,05
Δ	A	,250	,771	,748	-1,32	1,82
	B	,250	,806	,758	-1,39	1,89
	Γ	-4,450*	,787	<,001	-6,05	-2,85

*. The mean difference is significant at the 0.05 level.



Παρατήρηση: Υπάρχουν στατιστικώς σημαντικές διαφορές μεταξύ Α – Γ, Β – Γ και Γ – Δ

Παράδειγμα

Η διοίκηση ενός ιατρικού ομίλου ενδιαφέρεται να μελετήσει τον χρόνο παραμονής των ασθενών σε τέσσερα νοσηλευτικά ιδρύματα που ανήκουν στον όμιλο. Προκειμένου να πάρει απάντηση στο πρόβλημα, η διοίκηση του ιατρικού ομίλου κατέγραψε τον χρόνο παραμονής σε ημέρες στο νοσηλευτικό ίδρυμα 11 ασθενών από το ίδρυμα Α, 9 ασθενών από το ίδρυμα Β, 10 ασθενών από το ίδρυμα Γ και 8 ασθενών από το ίδρυμα Δ.

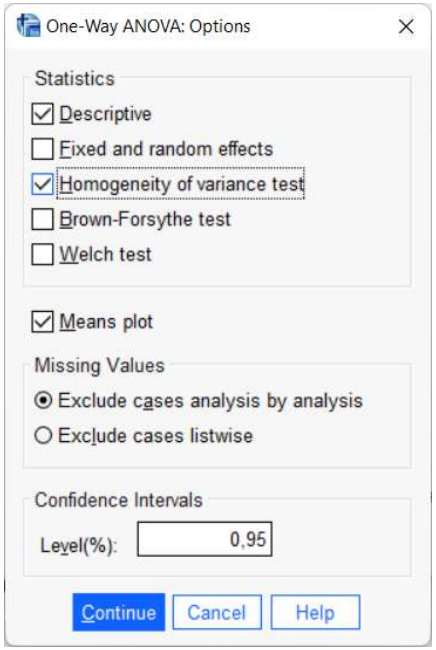
Descriptives

time

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Α	11	7,00	1,612	,486	5,92	8,08	5	10
Β	9	7,00	1,414	,471	5,91	8,09	5	10
Γ	10	11,70	2,214	,700	10,12	13,28	8	15
Δ	8	7,25	1,035	,366	6,38	8,12	6	9
Total	38	8,29	2,609	,423	7,43	9,15	5	15

Tests of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
time	Based on Mean	1,839	3	34	,159
	Based on Median	1,558	3	34	,218
	Based on Median and with adjusted df	1,558	3	28,538	,221
	Based on trimmed mean	1,810	3	34	,164



$$H_0 : \sigma_{\text{Α}}^2 = \sigma_{\text{Β}}^2 = \sigma_{\Gamma}^2 = \sigma_{\Delta}^2$$

$$H_1 : \text{Διαφορετικά}$$

Όλα εύκολα φαίνονται με τους ελέγχους

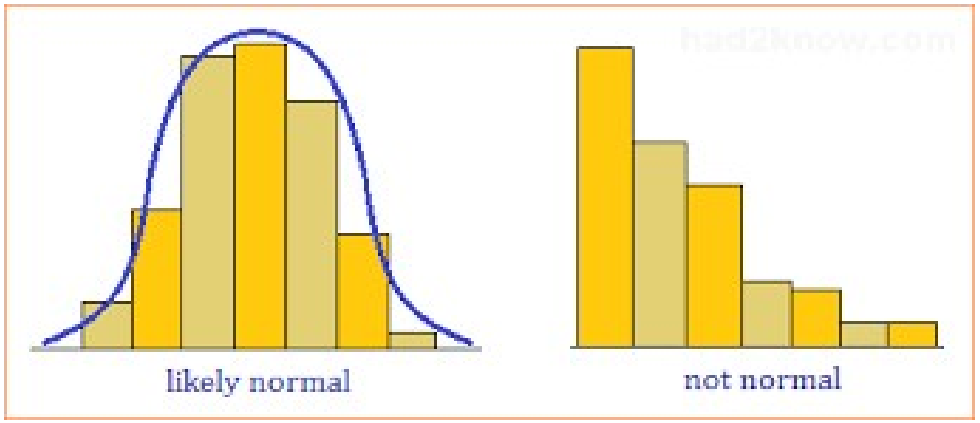
- Έως τώρα έχουμε δει ελέγχους υποθέσεων που αφορούν ένα δείγμα, δυο δείγματα (ανεξάρτητα και εξαρτημένα), περισσότερα από δυο ανεξάρτητα και εξαρτημένα δείγματα

- Μήπως όμως έχουμε ξεχάσει κάτι;

- Όλοι οι έλεγχοι που έχουμε δει απαιτούν κανονική κατανομή!!!

- Πως εξασφαλίζουμε ότι όντως έχουμε κανονική κατανομή;

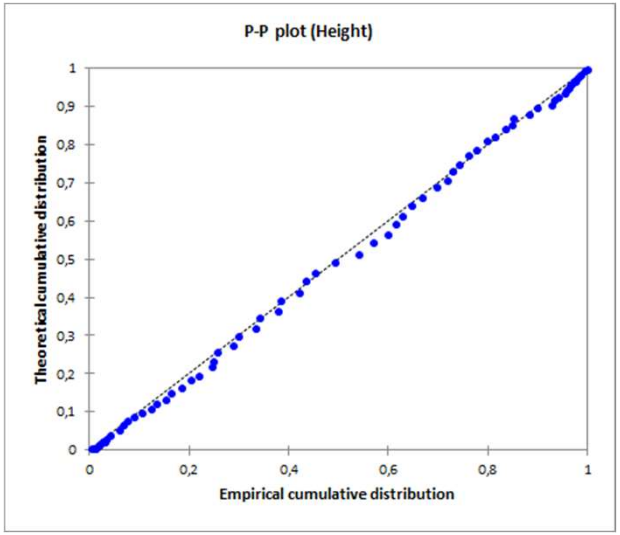
Έλεγχοι κανονικότητας



Έλεγχος κανονικότητας

- Έλεγχος μέσω γραφημάτων
  - P-P plot
  - Q-Q plot
- Έλεγχος των Kolmogorov-Smirnov
- Έλεγχος των Shapiro-Wilk

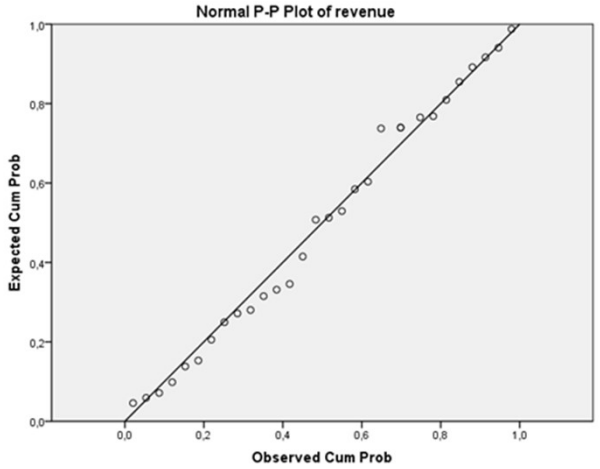
Έλεγχοι κανονικότητας μέσω γραφημάτων

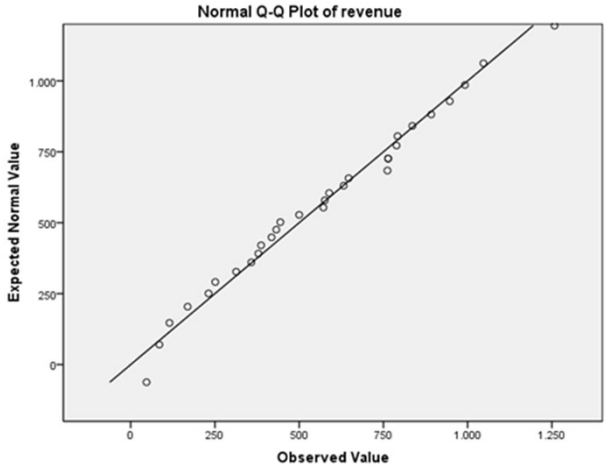


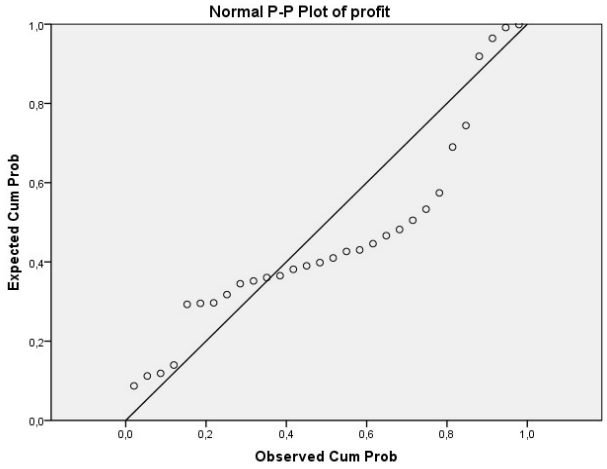
Έλεγχος κατανομής μέσω γραφημάτων

- Έλεγχος μέσω γραφημάτων
  - P-P plot
  - Q-Q plot

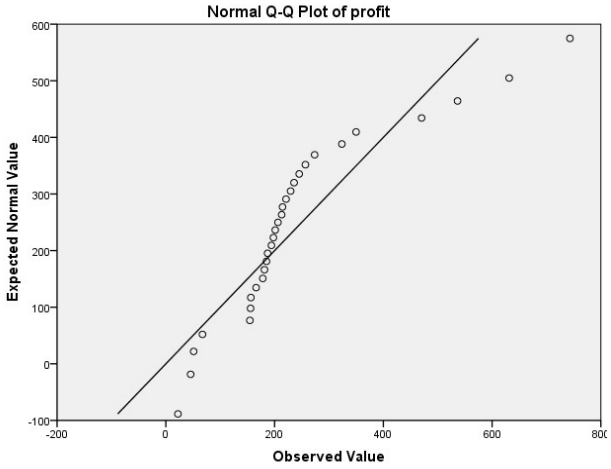
Έλεγχος κατανομής μέσω γραφημάτων



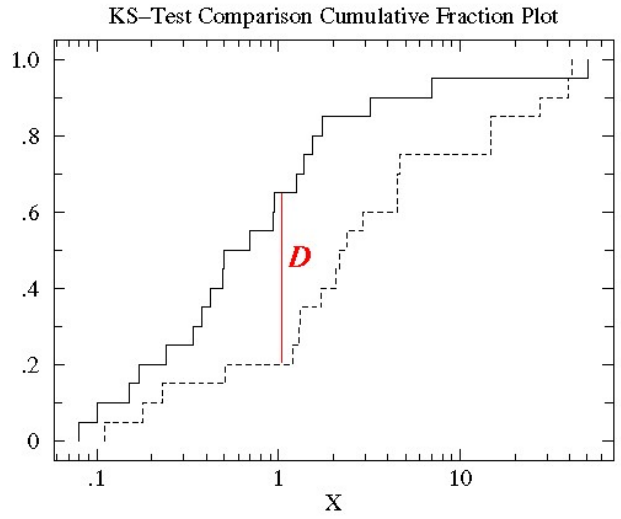




Εάν τα σημεία βρίσκονται αρκετά κοντά στη διχοτόμο ευθεία, τότε συμπεραίνουμε ότι τα δεδομένα ακολουθούν την κανονική κατανομή



Έλεγχος καλής προσαρμογής Kolmogorov-Smirnov



Έλεγχος Kolmogorov-Smirnov

- Ο έλεγχος των Kolmogorov-Smirnov για ένα δείγμα, ελέγχει εάν ένα σύνολο δεδομένων προέρχεται από μια συγκεκριμένη κατανομή (συνήθως μας ενδιαφέρει εάν προέρχεται από την κανονική κατανομή)

Έλεγχος Kolmogorov-Smirnov

- Έστω $X_1, X_2, \ldots, X_n$ ένα τυχαίο δείγμα από έναν πληθυσμό με συνάρτηση κατανομής $F(x)$

- Η μηδενική υπόθεση είναι H_0: $F(x) = F_0(x)$
- Η συνάρτηση κατανομής είναι μια συγκεκριμένη συνάρτηση κατανομής $F_0(x)$

- Η εναλλακτική υπόθεση είναι H_1: $F(x) \neq F_0(x)$
- Η συνάρτηση κατανομής δεν είναι η συγκεκριμένη συνάρτηση κατανομής $F_0(x)$

- Στην περίπτωσή μας, η συνάρτηση κατανομής $F_0(x)$ είναι η συνάρτηση κατανομής της κανονικής κατανομής

Παράδειγμα

Να ελεγχθεί ότι οι παρατηρήσεις του πίνακα προέρχονται από την κανονική κατανομή με μέση τιμή 32,00 και διακύμανση 3,24.

Δεδομένα:

31,00
31,40
33,30
33,40
33,50
33,70
34,40
34,90
36,20
37,00

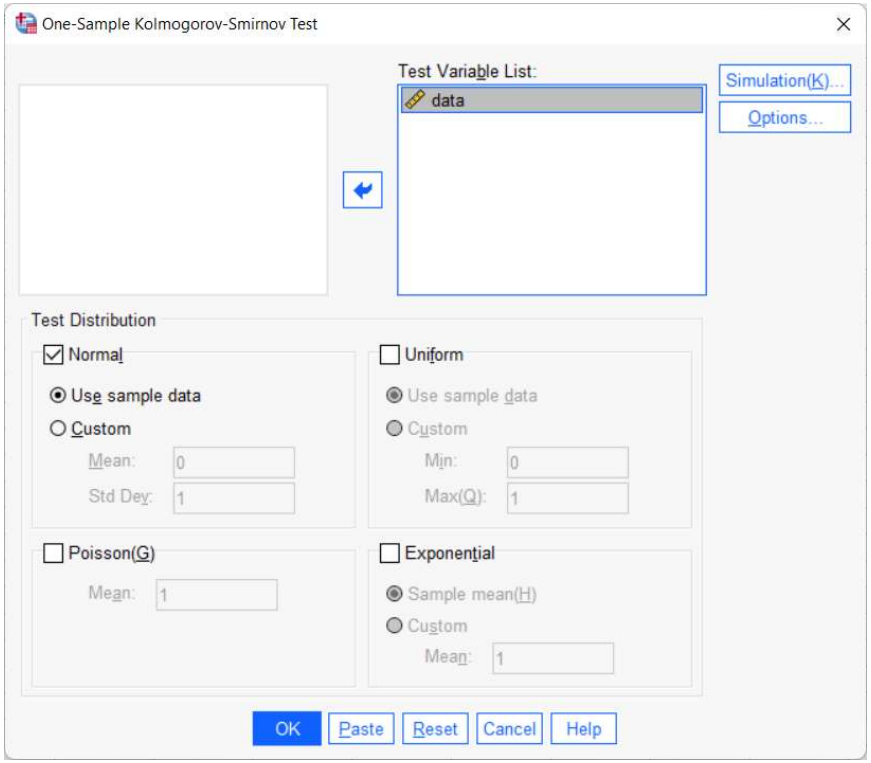
$$H_0 : F(x) = F_0(x)$$
$$H_1 : F(x) \neq F_0(x)$$

F_0 = Συνάρτηση κατανομής της $N(\mu,\sigma^2)$

$$F_0 = N(32,3.24)$$

Παράδειγμα

Να ελεγχθεί ότι οι παρατηρήσεις του πίνακα προέρχονται από την κανονική κατανομή με μέση τιμή 32,00 και διακύμανση 3,24.



Analyze → Nonparametric Tests → Legacy Dialogs → 1-Sample K-S

One-Sample Kolmogorov-Smirnov Test

			data
N			10
Normal Parameters[a,b]	Mean		33,8800
	Std. Deviation		1,87427
Most Extreme Differences	Absolute		,178
	Positive		,138
	Negative		-,178
Test Statistic			,178
Asymp. Sig. (2-tailed)[c]			,200[d]
Monte Carlo Sig. (2-tailed)[e]	Sig.		,487
	99% Confidence Interval	Lower Bound	,474
		Upper Bound	,499

a. Test distribution is Normal.

b. Calculated from data.

c. Lilliefors Significance Correction.

d. This is a lower bound of the true significance.

e. Lilliefors' method based on 10000 Monte Carlo samples with starting seed 2000000.

Παρατήρηση: Επειδή το p-value > 0,200 (> 0,05) συμπεραίνουμε ότι οι παρατηρήσεις προέρχονται από την κανονική κατανομή $N(33,88 , 1,87^2)$

Μετά την απόρριψη της κανονικότητας, τί κάνουμε;

- Εφαρμόσαμε τον έλεγχο της κανονικότητας και απορρίψαμε την μηδενική υπόθεση
- Αυτό σημαίνει ότι τα δεδομένα μας δεν ακολουθούν την κανονική κατανομή
- Αυτό σημαίνει ότι δεν μπορούμε να εφαρμόσουμε τους ελέγχους που ήδη γνωρίζουμε
- Τί κάνουμε;;;

Μη Παραμετρικοί Έλεγχοι

Μη παραμετρικοί έλεγχοι για ένα δείγμα

Μη παραμετρικοί έλεγχοι για ένα δείγμα

- Εφαρμόζονται σε υποθέσεις που αφορούν συνήθως τη διάμεσο (αντί του μέσου) και είναι κατασκευασμένοι για όλα τα είδη των δεδομένων (κανονικά ή μη κανονικά).

- Ο έλεγχος που εφαρμόζουμε όταν έχουμε ένα δείγμα ονομάζεται προσημικός έλεγχος (sign test)

- Η στατιστική συνάρτηση του ελέγχου βασίζεται στον αριθμό των παρατηρήσεων που είναι μεγαλύτερες από την τιμή του ελέγχου (την τιμή που αναφέρει η H_0)

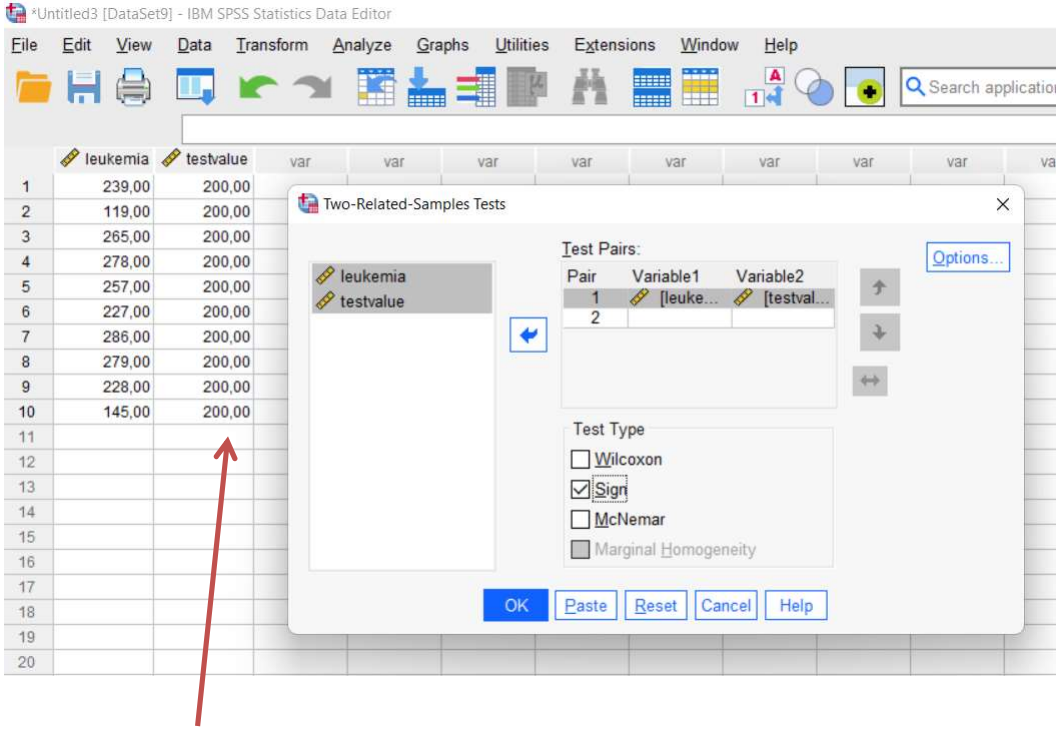
Παράδειγμα

Ο χρόνος διάγνωσης σε ημέρες της λευχαιμίας σε δέκα ζώα μολυσμένα με λευχαιμικά κύτταρα δίνεται στον επόμενο πίνακα. Να ελεγχθεί εάν ο διάμεσος χρόνος διάγνωσης είναι 200 ημέρες.

239		239	+
119		119	-
265	$H_0 : \delta = 200 - H_1 : \delta \neq 200$	265	+
278		278	+
257		257	+
227		227	+
286		286	+
279	X=Αριθμός +=8	279	+
228		228	+
145		145	-

Παράδειγμα

Ο χρόνος διάγνωσης σε ημέρες της λευχαιμίας σε δέκα ζώα μολυσμένα με λευχαιμικά κύτταρα δίνεται στον επόμενο πίνακα. Να ελεγχθεί εάν ο διάμεσος χρόνος διάγνωσης είναι 200 ημέρες.



Αν και έχουμε ένα δείγμα ακολουθούμε διαδικασία για δυο εξαρτημένα δείγματα

Analyze → Nonparametric Tests → Legacy Dialogs → 2 Related Samples

Χρησιμοποιούμε μια βοηθητική στήλη, π.χ. την «testvalue», που ως τιμές έχει την τιμή ελέγχου 200.

Παράδειγμα

Ο χρόνος διάγνωσης σε ημέρες της λευχαιμίας σε δέκα ζώα μολυσμένα με λευχαιμικά κύτταρα δίνεται στον επόμενο πίνακα. Να ελεγχθεί εάν ο διάμεσος χρόνος διάγνωσης είναι 200 ημέρες.

Test Statistics[a]

	testvalue - leukemia
Exact Sig. (2-tailed)	,109[b]

a. Sign Test

b. Binomial distribution used.

Παρατήρηση: Επειδή το p-value = 0,109 > 0,05 συμπεραίνουμε ότι ο διάμεσος χρόνος διάγνωσης ισούται με 200

Έλεγχοι για δυο ανεξάρτητα δείγματα

Μη παραμετρικοί έλεγχοι για δυο δείγματα

- Εφαρμόζονται σε υποθέσεις που αφορούν συνήθως τη διάμεσο (αντί του μέσου) και είναι κατασκευασμένοι για όλα τα είδη των δεδομένων (κανονικά ή μη κανονικά).
- Βασίζονται στις δειγματικές διαμέσους ή στις διατεταγμένες παρατηρήσεις των δυο δειγμάτων.
- Οι μη παραμετρικοί έλεγχοι είναι συνήθως λιγότερο ισχυροί από τους αντίστοιχους παραμετρικούς, είναι όμως πιο ασφαλείς όταν δεν έχει εξασφαλισθεί η κανονικότητα των δεδομένων.

Έλεγχοι για δυο ανεξάρτητα δείγματα
Έλεγχος Mann-Whitney

- Ο έλεγχος Mann-Whitney είναι ο μη παραμετρικός ισοδύναμος του t test για δυο ανεξάρτητα δείγματα.
- Η απαίτηση που θέτει είναι οι δυο πληθυσμιακές κατανομές να είναι περίπου της ίδιας σχηματικής μορφής.
- Η μηδενική υπόθεση του ελέγχου είναι η ισότητα των διαμέσων των δυο κατανομών.

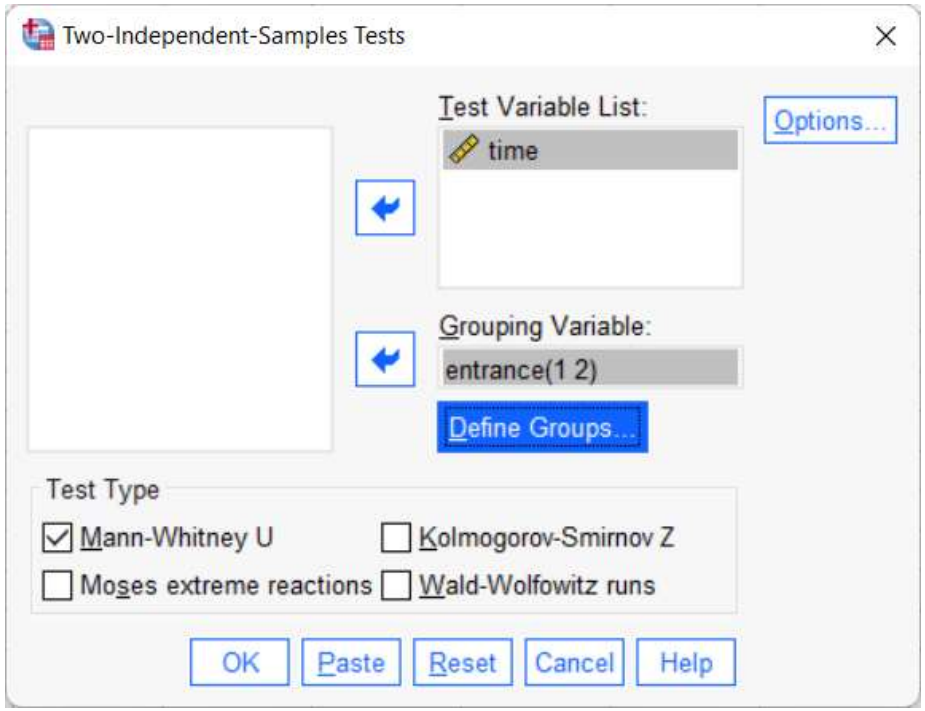
Έλεγχοι για δυο ανεξάρτητα δείγματα

Έλεγχος Wilcoxon-Mann-Whitney

- Επιλέγουμε δυο ανεξάρτητα δείγματα
- Οι τιμές των παρατηρήσεων των δυο δειγμάτων συνδυάζονται σε ένα ενιαίο σύνολο τιμών και διατάσσονται από τη μικρότερη στη μεγαλύτερη.
- Προσδιορίζεται για κάθε μια από τις τιμές η σχετική κατάταξη (rank) που καταλαμβάνει στην ενιαία διάταξη.
- Αθροίζονται οι σχετικές κατατάξεις των παρατηρήσεων χωριστά για κάθε δείγμα.
- Αποδεχόμενοι την μηδενική υπόθεση, αναμένουμε η κατανομή των σχετικών κατατάξεων στα δυο δείγματα να είναι τυχαία και οι τιμές των σχετικών κατατάξεων στα δυο δείγματα να είναι ίσες.

Παράδειγμα

Ο διοικητής ενός νοσοκομείου θεωρεί ότι ο χρόνος που απαιτείται από τη στιγμή που θα περάσει το ασθενοφόρο τη βόρεια είσοδο του νοσοκομείου μέχρι τη στιγμή που θα εξεταστεί ο ασθενής, διαφέρει από το χρόνο που απαιτείται εάν το ασθενοφόρο μπει από τη νότια είσοδο του νοσοκομείου. Ο διοικητής επέλεξε τυχαία 5 περιπτώσεις, που το ασθενοφόρο πέρασε τη βόρεια είσοδο του νοσοκομείου και 5 περιπτώσεις, που το ασθενοφόρο πέρασε τη νότια είσοδο. Στη συνέχεια, κατέγραψε το χρόνο (σε λεπτά) που απαιτήθηκε μέχρι τη στιγμή που εξετάστηκε ο ασθενής.



$$H_0 : \delta_1 = \delta_2 \; - H_1 : \delta_1 \neq \delta_2$$

Analyze →
Nonparametric Tests →
Legacy Dialogs → 2 Independent Samples

Παράδειγμα

Ο διοικητής ενός νοσοκομείου θεωρεί ότι ο χρόνος που απαιτείται από τη στιγμή που θα περάσει το ασθενοφόρο τη βόρεια είσοδο του νοσοκομείου μέχρι τη στιγμή που θα εξεταστεί ο ασθενής, διαφέρει από το χρόνο που απαιτείται εάν το ασθενοφόρο μπει από τη νότια είσοδο του νοσοκομείου. Ο διοικητής επέλεξε τυχαία 5 περιπτώσεις, που το ασθενοφόρο πέρασε τη βόρεια είσοδο του νοσοκομείου και 5 περιπτώσεις, που το ασθενοφόρο πέρασε τη νότια είσοδο. Στη συνέχεια, κατέγραψε το χρόνο (σε λεπτά) που απαιτήθηκε μέχρι τη στιγμή που εξετάστηκε ο ασθενής.

Ranks

	entrance	N	Mean Rank	Sum of Ranks
time	Βόρεια είσοδος	5	6,50	32,50
	Νότια είσοδος	5	4,50	22,50
	Total	10		

Μέση κατάταξη των παρατηρήσεων (Ranks)

Test Statistics[a]

	time
Mann-Whitney U	7,500
Wilcoxon W	22,500
Z	-1,064
Asymp. Sig. (2-tailed)	,287
Exact Sig. [2*(1-tailed Sig.)]	,310[b]

a. Grouping Variable: entrance

b. Not corrected for ties.

Παρατήρηση: Επειδή το p-value = 0,287 > 0,05 συμπεραίνουμε ότι ο χρόνος που απαιτείται από τη στιγμή που θα περάσει το ασθενοφόρο τη βόρεια είσοδο του νοσοκομείου μέχρι τη στιγμή που θα εξεταστεί ο ασθενής, δεν διαφέρει από το χρόνο που απαιτείται εάν το ασθενοφόρο μπει από τη νότια είσοδο του νοσοκομείου

Έλεγχοι για δυο εξαρτημένα δείγματα

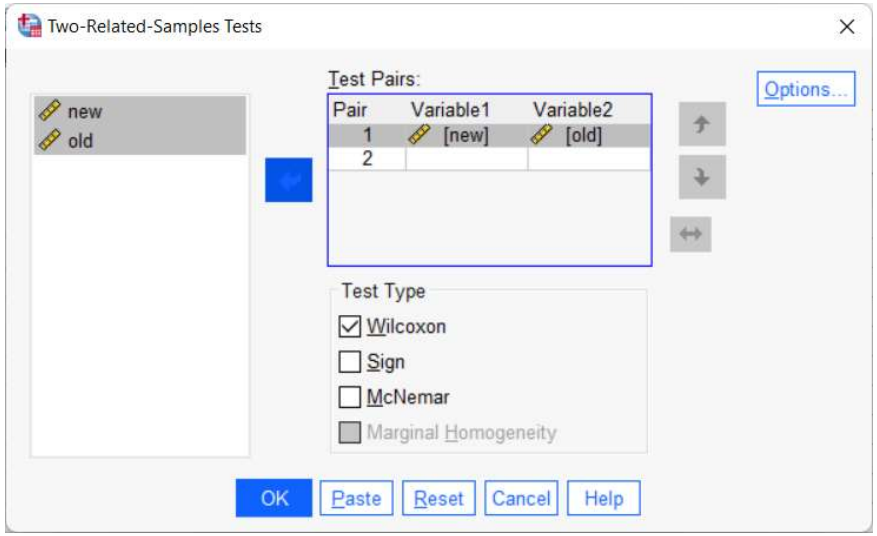
Έλεγχοι για δυο εξαρτημένα δείγματα

- Οι μετρήσεις αναφέρονται στην **ίδια** πειραματική μονάδα και συνεπώς δεν είναι ανεξάρτητες
- Για παράδειγμα, έχουμε μετρήσεις πριν και μετά την εφαρμογή μιας θεραπείας ή μετρήσεις ενός χαρακτηριστικού δίδυμων αδελφών
- Χρησιμοποιούμε τις διαφορές των μετρήσεων
- Κατάλληλοι έλεγχοι: Προσημικός Έλεγχος ή Έλεγχος Προσημασμένης Διάταξης του Wilcoxon

Παράδειγμα

Μια εταιρεία ιατρικών οργάνων, θέλει να ελέγξει την αποτελεσματικότητα ενός νέου σφυγμομανομέτρου (πιεσόμετρου) πριν αυτό κυκλοφορήσει στην αγορά. Για τον έλεγχο αυτό, η εταιρία επέλεξε τυχαία 8 άτομα – υπαλλήλους της εταιρίας. Αρχικά μέτρησε την πίεσή τους με το νέο σφυγμομανόμετρο και στη συνέχεια με ένα σφυγμομανόμετρο που ήδη κυκλοφορεί στην αγορά.

$$H_0 : \delta_1 = \delta_2 - H_1 : \delta_1 \neq \delta_2$$



Test Statistics[a]

	old - new
Z	-,420[b]
Asymp. Sig. (2-tailed)	,674

a. Wilcoxon Signed Ranks Test

b. Based on positive ranks.

Παρατήρηση: Επειδή το p-value = 0,674 > 0,05 συμπεραίνουμε ότι τα δυο σφυγμομανόμετρα είναι ισοδύναμα

Analyze → Nonparametric Tests → Legacy Dialogs → 2 Related Samples

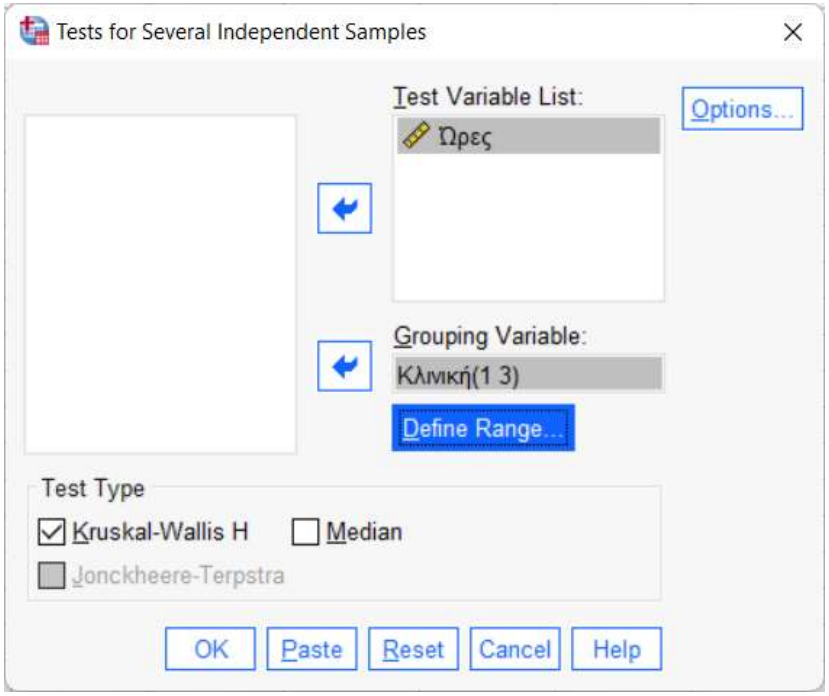
Έλεγχος για k>2 ανεξάρτητα δείγματα (έλεγχος Kruskal-Wallis)

Έλεγχος για k ανεξάρτητα δείγματα

- Επιλέγουμε k > 2 ανεξάρτητα δείγματα
- Οι τιμές των παρατηρήσεων των k δειγμάτων συνδυάζονται σε ένα ενιαίο σύνολο τιμών και διατάσσονται από τη μικρότερη στη μεγαλύτερη.
- Προσδιορίζεται για κάθε μια από τις τιμές η σχετική κατάταξη (rank) που καταλαμβάνει στην ενιαία διάταξη.
- Αθροίζονται οι σχετικές κατατάξεις των παρατηρήσεων χωριστά για κάθε δείγμα.
- Αποδεχόμενοι την μηδενική υπόθεση, αναμένουμε η κατανομή των σχετικών κατατάξεων στα k δείγματα να είναι τυχαία και οι τιμές των σχετικών κατατάξεων στα k δείγματα να είναι ίσες.

Παράδειγμα

Ο διοικητής ενός νοσοκομείου θέλει να ελέγξει κατά πόσο διαφέρουν οι μηνιαίες ώρες απουσίας του νοσηλευτικού προσωπικού 3 τμημάτων του νοσοκομείου. Το παθολογικό τμήμα του νοσοκομείου έχει 24 νοσηλευτές, το καρδιολογικό τμήμα έχει 35 νοσηλευτές ενώ το ορθοπεδικό τμήμα έχει 21 νοσηλευτές.



$$H_0 : \delta_A = \delta_B = \delta_\Gamma \; - \; H_1 : \text{Διαφορετικά}$$

Analyze → Nonparametric Tests → Legacy Dialogs → K Independent Samples

Παράδειγμα

Ο διοικητής ενός νοσοκομείου θέλει να ελέγξει κατά πόσο διαφέρουν οι μηνιαίες ώρες απουσίας του νοσηλευτικού προσωπικού 3 τμημάτων του νοσοκομείου. Το παθολογικό τμήμα του νοσοκομείου έχει 24 νοσηλευτές, το καρδιολογικό τμήμα έχει 35 νοσηλευτές ενώ το ορθοπεδικό τμήμα έχει 21 νοσηλευτές.

Ranks

	Κλινική	N	Mean Rank
Ώρες	Παθολογική	24	52,50
	Καρδιολογική	35	27,73
	Ορθοπεδική	21	48,07
	Total	80	

Test Statistics[a,b]

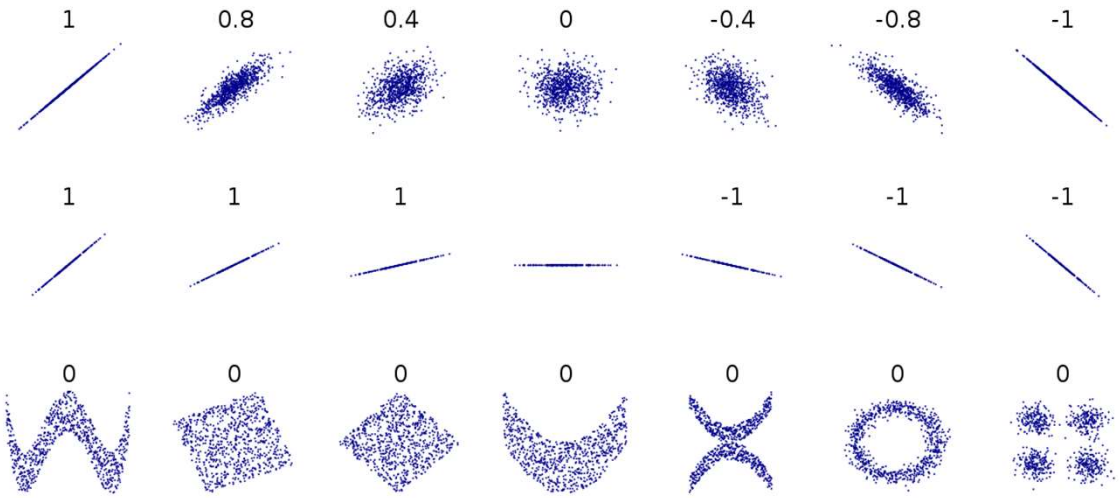
	Ώρες
Kruskal-Wallis H	21,584
df	2
Asymp. Sig.	<,001

a. Kruskal Wallis Test

b. Grouping Variable: Κλινική

Παρατήρηση: Επειδή το p-value είναι < 0,001 συμπεραίνουμε ότι ο μέσος μηνιαίος χρόνος απουσίας των νοσηλευτών διαφέρει μεταξύ των 3 τμημάτων του νοσοκομείου. Με το IBM SPSS Statistics 29 δεν μπορούμε αυτόματα να κάνουμε πολλαπλές συγκρίσεις.

Συσχέτιση Ποσοτικών Μεταβλητών



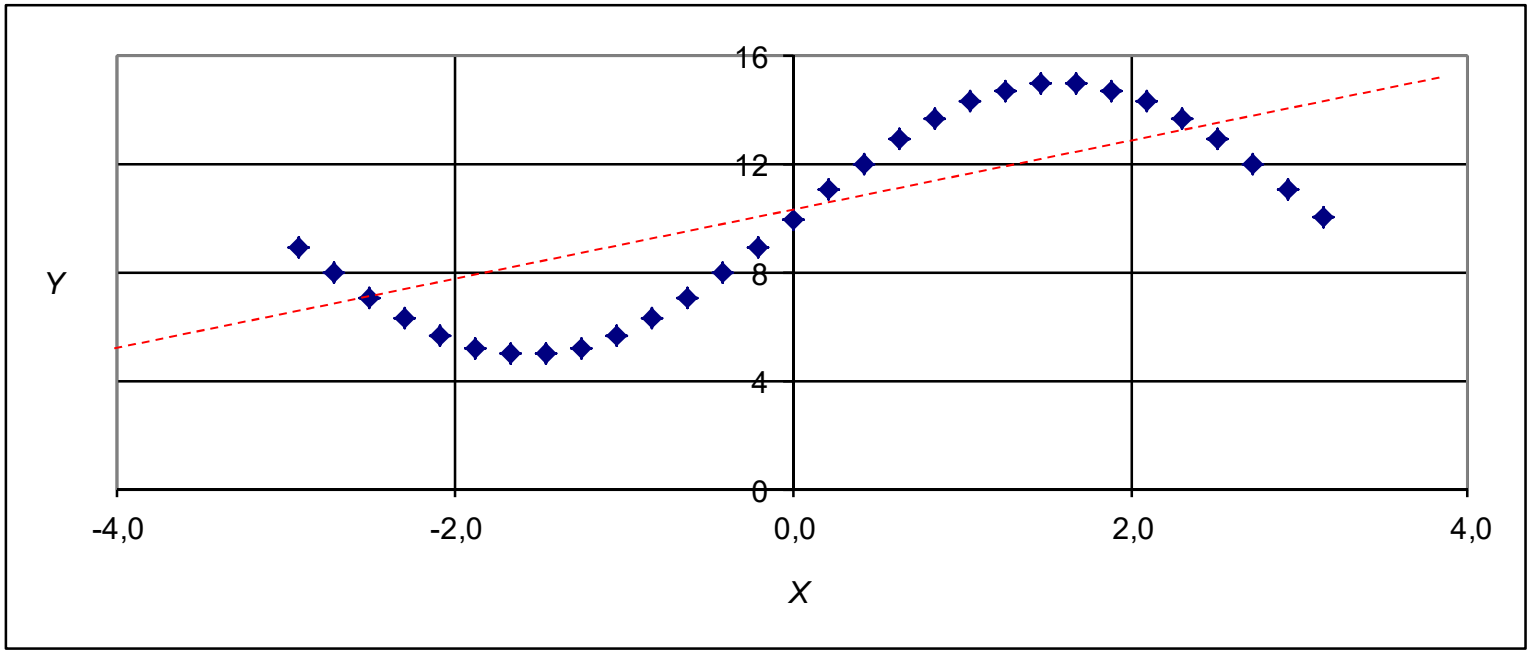
Η έννοια της συσχέτισης

- Όταν οι τιμές δυο ποσοτικών μεταβλητών σχετίζονται με τέτοιον τρόπο ώστε η μια, η οποία καλείται **εξαρτημένη** (Y), να μπορεί να προβλεφθεί όταν η άλλη, η οποία καλείται **ανεξάρτητη** (X), είναι γνωστή, τότε λέμε ότι οι μεταβλητές αυτές παρουσιάζουν συσχέτιση.
  - BMI = f (Βάρος) + σφάλμα
  - BMI = f (*Ώρες Παρακολούθησης Τηλεόρασης*) + σφάλμα

- Τα κύρια χαρακτηριστικά της συσχέτισης είναι:
  - Το είδος της συσχέτισης
  - Η κατεύθυνση της συσχέτισης
  - Η ένταση της συσχέτισης

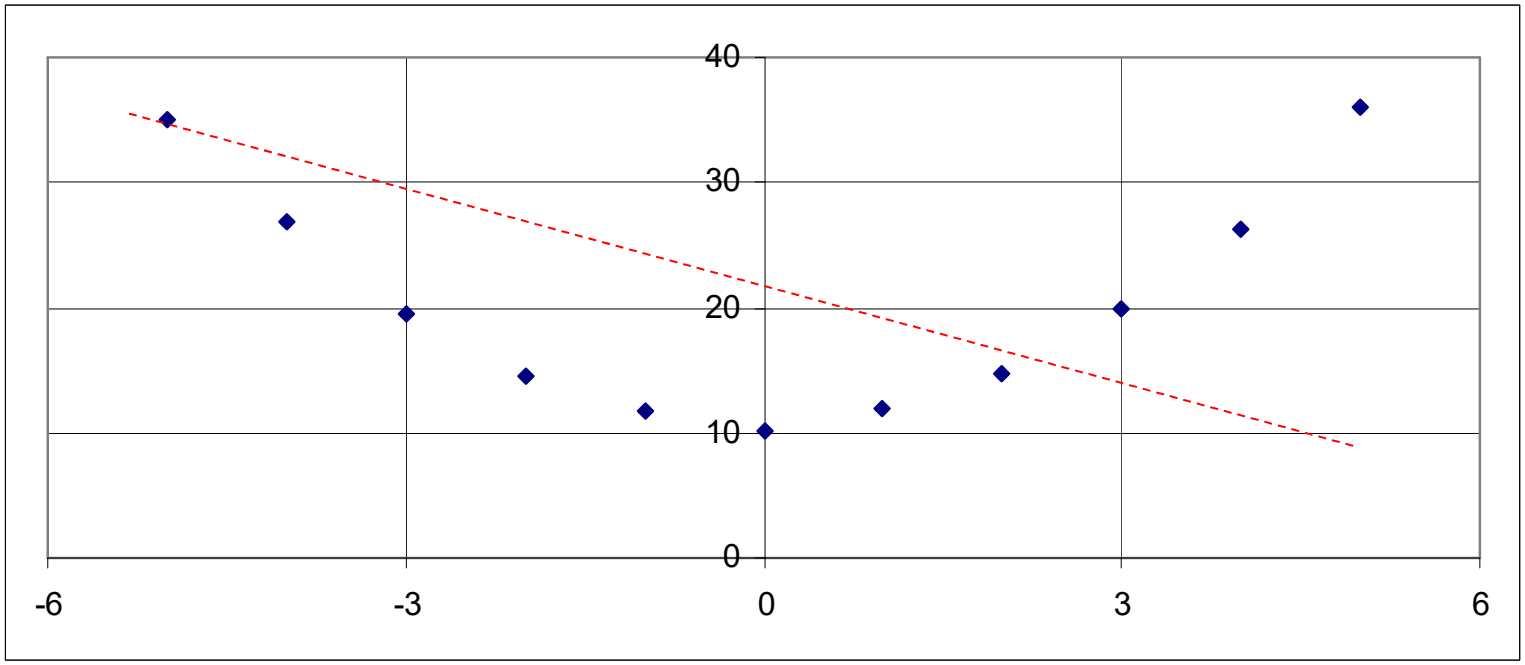
Το είδος της συσχέτισης

- Μη Γραμμική (κυματοειδής)



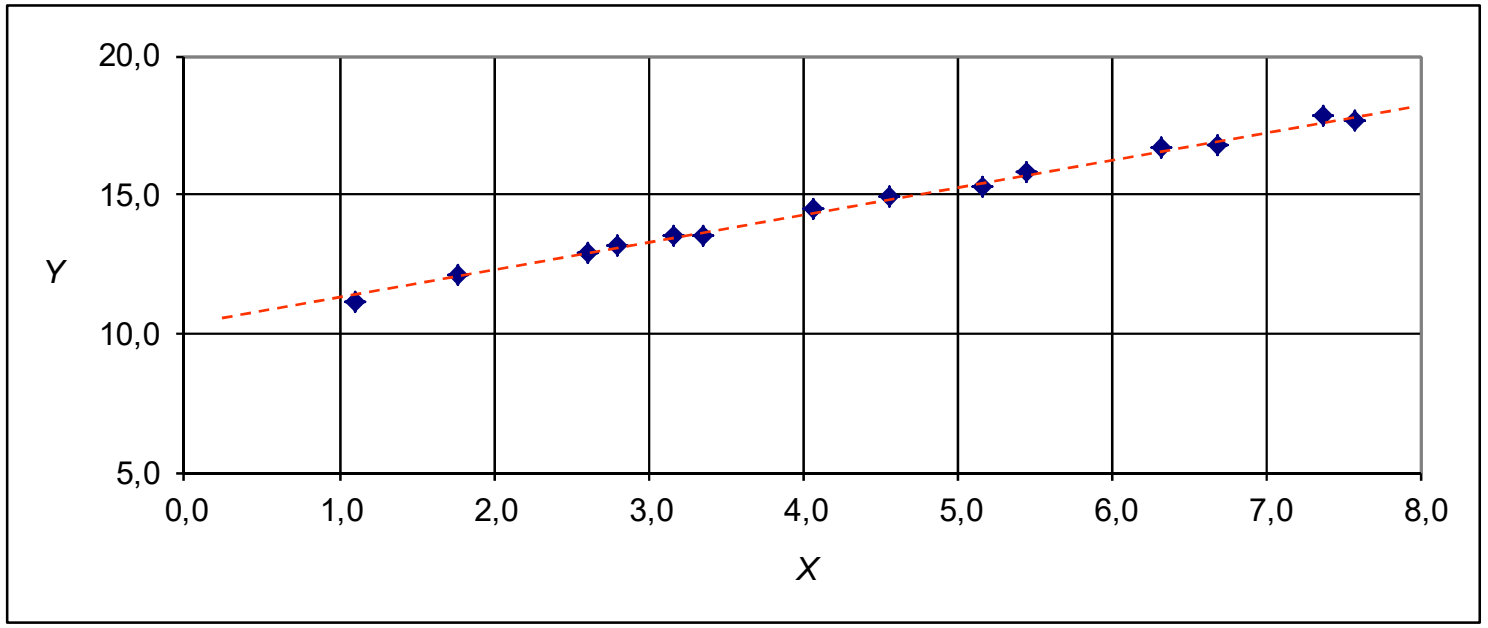
Το είδος της συσχέτισης

- Μη Γραμμική (παραβολή)



Το είδος της συσχέτισης

- Γραμμική



Η ένταση και η κατεύθυνση της συσχέτισης

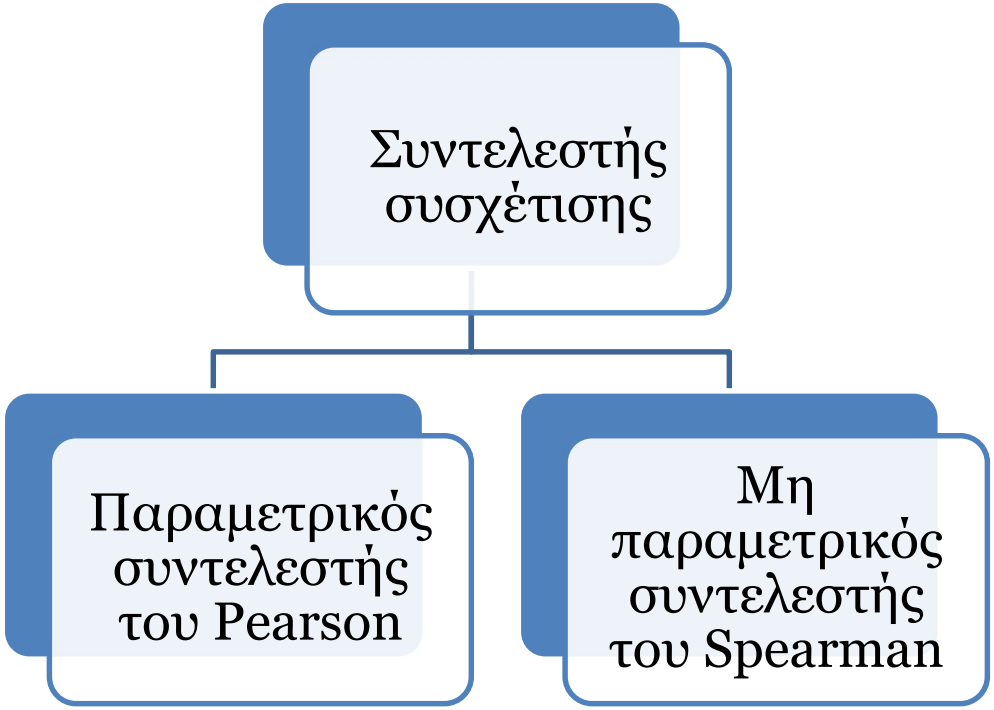
- Το μέτρο συσχέτισης καλείται συντελεστής γραμμικής συσχέτισης
- Λαμβάνει τιμές στο διάστημα [-1,+1]
  - Εάν ο συντελεστής παίρνει τιμές κοντά στο 1 τότε υπάρχει ισχυρή **θετική γραμμική συσχέτιση** μεταξύ των δυο ποσοτικών μεταβλητών.
  - Εάν ο συντελεστής παίρνει τιμές κοντά στο -1 τότε υπάρχει ισχυρή **αρνητική γραμμική συσχέτιση** μεταξύ των δυο ποσοτικών μεταβλητών.
  - Εάν ο συντελεστής παίρνει τιμές κοντά στο 0 τότε **δεν υπάρχει γραμμική συσχέτιση** μεταξύ των δυο ποσοτικών μεταβλητών, δεν αποκλείεται όμως η ύπαρξη μιας άλλου είδους συσχέτισης.

Η ένταση της συσχέτισης

Εύρος τιμών	Ένταση σχέσης
0,00 – 0,20	Πολύ ασθενής
0,20 – 0,40	Ασθενής
0,40 – 0,60	Μέτρια
0,60 – 0,80	Ισχυρή
0,80 – 1,00	Πολύ ισχυρή

Το ίδιο ισχύει και για την αρνητική συσχέτιση

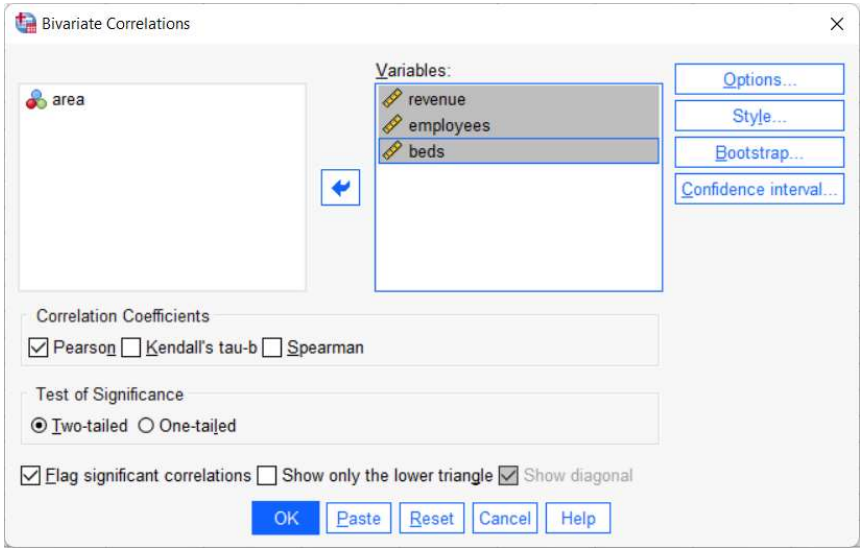
Συντελεστής συσχέτισης



Παράδειγμα

- Ένας ιατρικός όμιλος έχει 30 ιατρικά κέντρα σε όλη την Ελλάδα.
- Η διεύθυνση του ιατρικού ομίλου κατέγραψε
  - τα έσοδα (revenue),
  - την περιοχή (area),
    - 1=Νότια Ελλάδα
    - 2=Βόρεια Ελλάδα
    - 3=Νησιωτική Ελλάδα
  - τον αριθμό εργαζομένων (employees),
  - τον αριθμό κλινών (beds) για κάθε ιατρικό κέντρο
- Η διεύθυνση του ιατρικού ομίλου θέλει να μελετήσει τη σχέση των εσόδων με τις υπόλοιπες μεταβλητές.

Παράδειγμα



Correlations

		revenue	employees	beds
revenue	Pearson Correlation	1	,820**	,894**
	Sig. (2-tailed)		<,001	<,001
	N	30	30	30
employees	Pearson Correlation	,820**	1	,808**
	Sig. (2-tailed)	<,001		<,001
	N	30	30	30
beds	Pearson Correlation	,894**	,808**	1
	Sig. (2-tailed)	<,001	<,001	
	N	30	30	30

**. Correlation is significant at the 0.01 level (2-tailed).

Analyze → Correlate → Bivariate

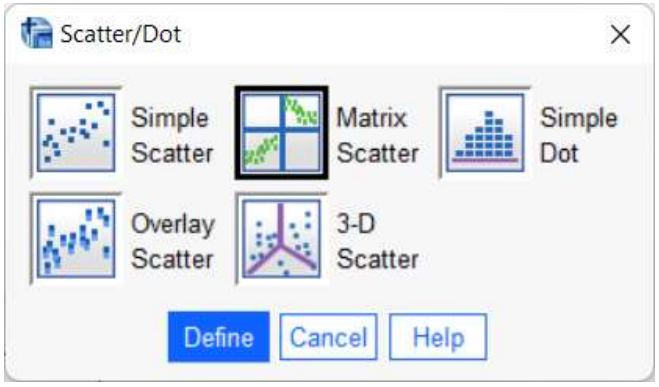
Correlation Coefficients: Pearson

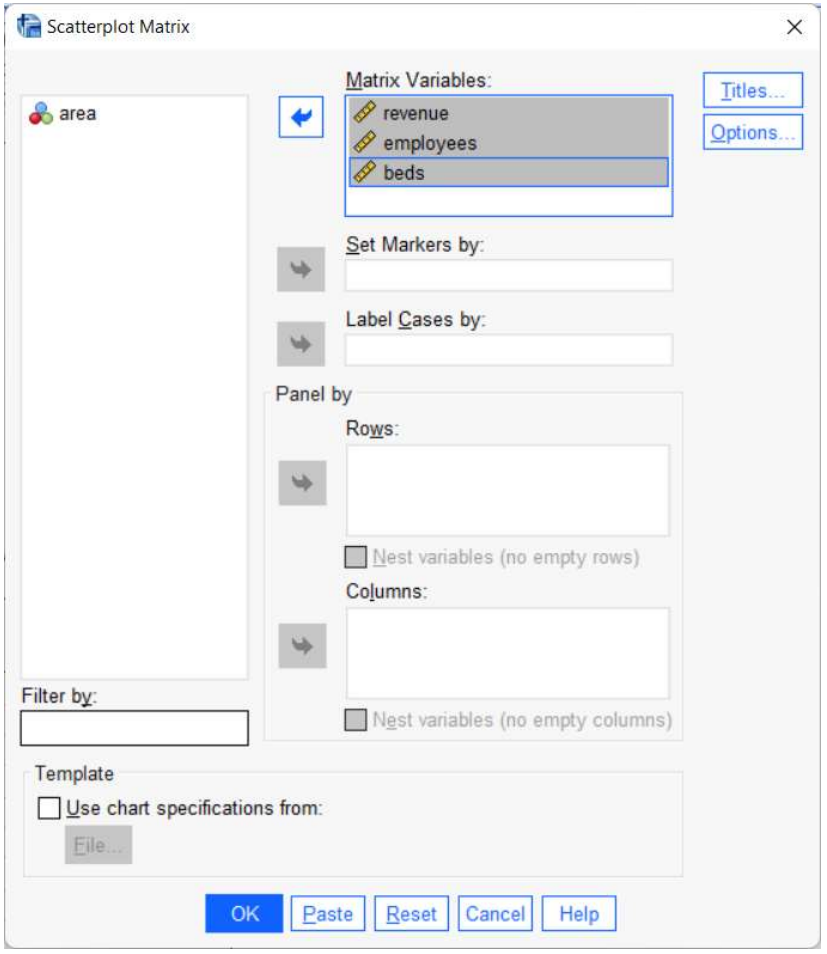
Flag significant correlations

Παράδειγμα

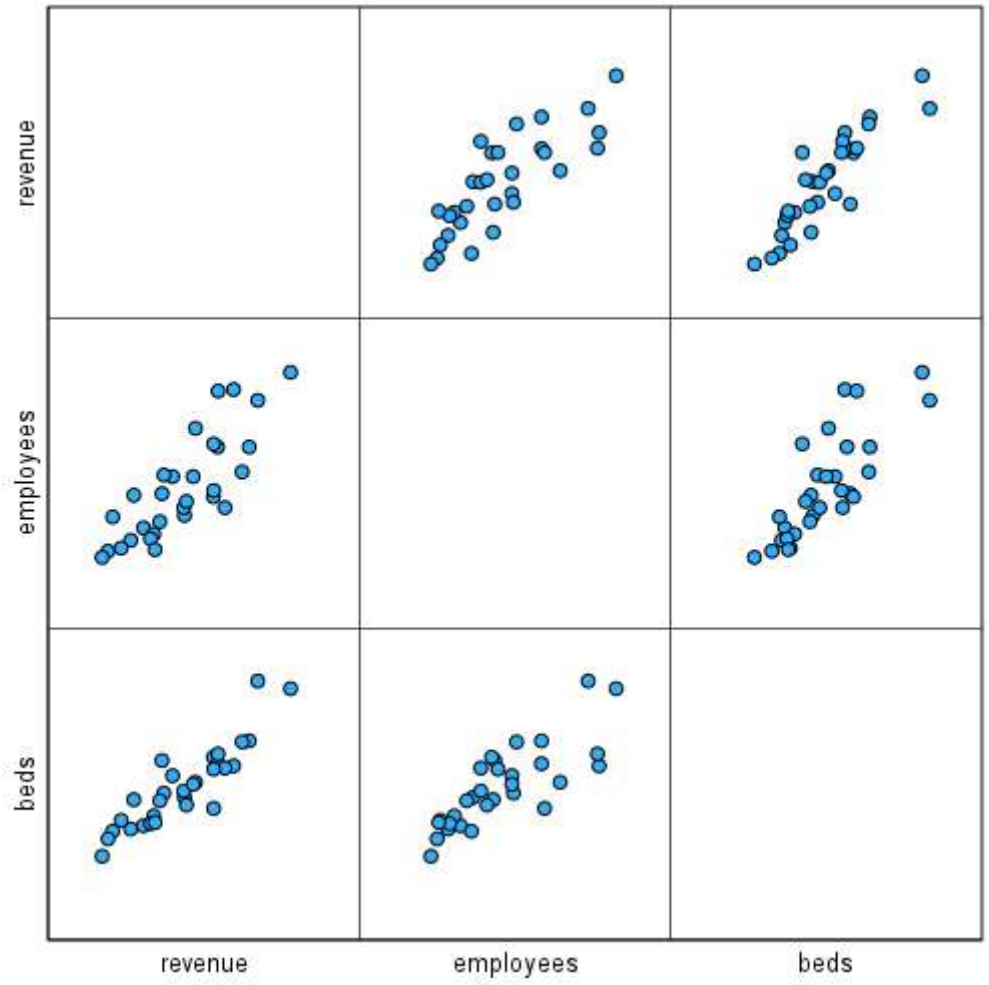
Graphs → Scatter/Dot

Matrix Scatter





Παράδειγμα

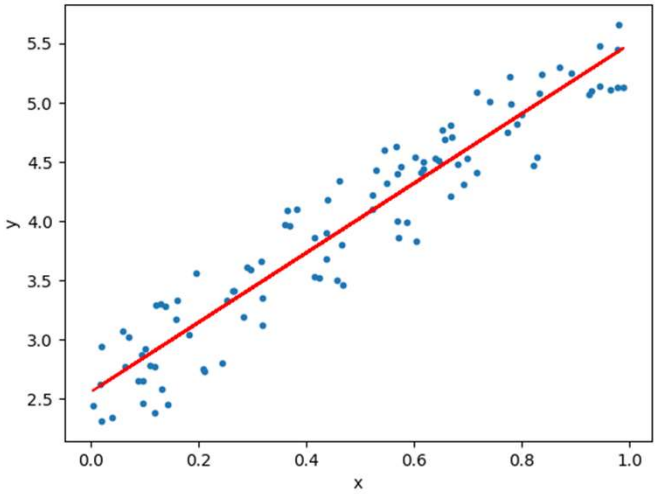


Correlations

		revenue	employees	beds
revenue	Pearson Correlation	1	,820**	,894**
	Sig. (2-tailed)		<,001	<,001
	N	30	30	30
employees	Pearson Correlation	,820**	1	,808**
	Sig. (2-tailed)	<,001		<,001
	N	30	30	30
beds	Pearson Correlation	,894**	,808**	1
	Sig. (2-tailed)	<,001	<,001	
	N	30	30	30

**. Correlation is significant at the 0.01 level (2-tailed).

Απλή Γραμμική Παλινδρόμηση



Απλή Γραμμική Παλινδρόμηση

- Ο απλούστερος τύπος μοντέλου που συνδέει δυο γραμμικά συσχετισμένες μεταβλητές (*X* και *Y) είναι η* ευθεία $Y=\alpha+\beta X$, η οποία λέγεται ευθεία γραμμικής παλινδρόμησης.

- Η μεταβλητή *Y* ονομάζεται εξαρτημένη μεταβλητή (dependent variable) ή μεταβλητή απόκρισης (response), ενώ η *X* ονομάζεται ανεξάρτητη μεταβλητή (independent variable) ή επεξηγηματική μεταβλητή (exploratory variable).

- Ο συντελεστής *α* ονομάζεται σταθερά (intercept), ενώ το *β* είναι η κλίση (slope) της ευθείας. Για μια δεδομένη ευθεία, τα *α* και *β* είναι σταθερές ποσότητες.

- Η σταθερά *α* είναι η τιμή της *Y* για $X=0$, ενώ η κλίση *β* παριστάνει τη μεταβολή που επιφέρει στην *Y* η αλλαγή της *X* κατά μια μονάδα.

Απλή Γραμμική Παλινδρόμηση

- Εάν η παρατήρηση Y δεν υπόκειται σε σφάλματα, δηλαδή αν για κάποια τιμή της ανεξάρτητης μεταβλητής X, μπορούμε να προβλέψουμε ακριβώς την Y, τότε το μοντέλο καλείται προσδιοριστικό (deterministic).

- Στην πραγματικότητα όμως σπάνια μπορούμε να προβλέψουμε ακριβώς την τιμή της Y.

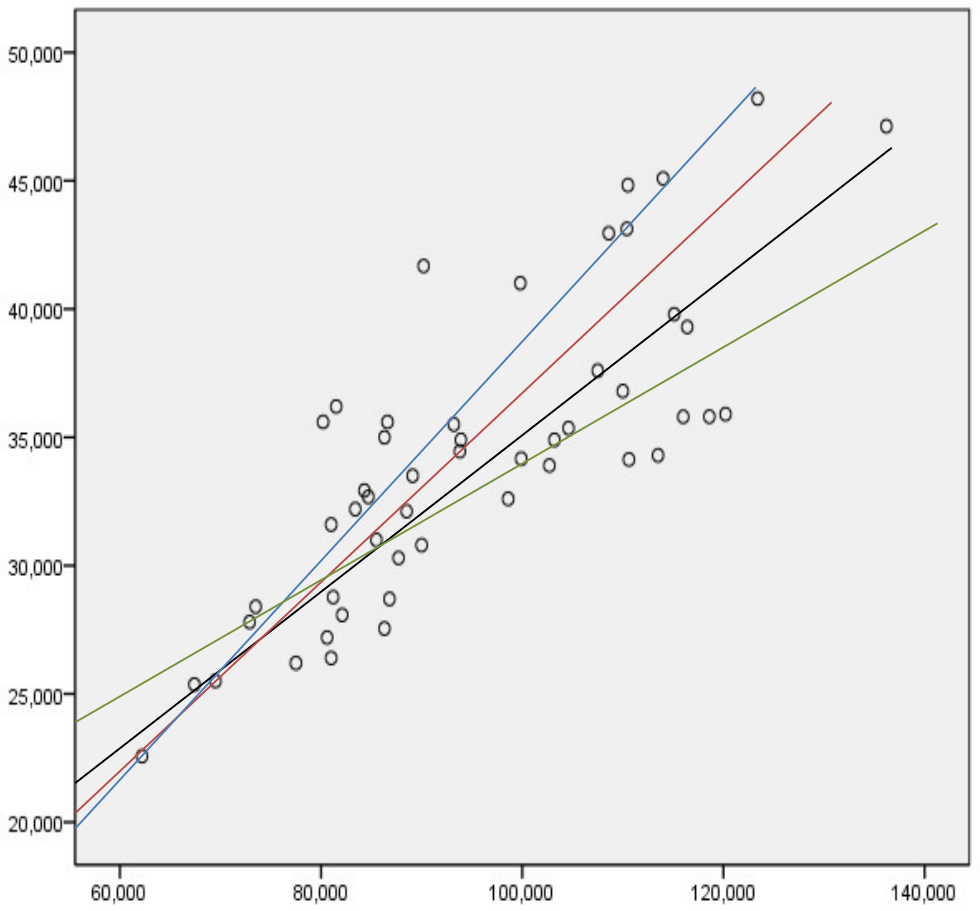
- Σε αυτή την περίπτωση ισχύει $Y=a+\beta X+e$, όπου e είναι ένα τυχαίο σφάλμα (error) και παριστάνει τη διαφορά της παρατηρημένης τιμής Y, για δοθέν X, από τη θεωρητική τιμή $a+\beta X$. Αυτού του είδους τα μοντέλα καλούνται στοχαστικά (stochastic models, probabilistic models).

- Όσον αφορά τα σφάλματα e υποθέτουμε ότι είναι τυχαία, με μέση τιμή $E(e)=0$. Έτσι, αφού τα a και β είναι άγνωστες σταθερές ισχύει $E(Y)=a+\beta X$.

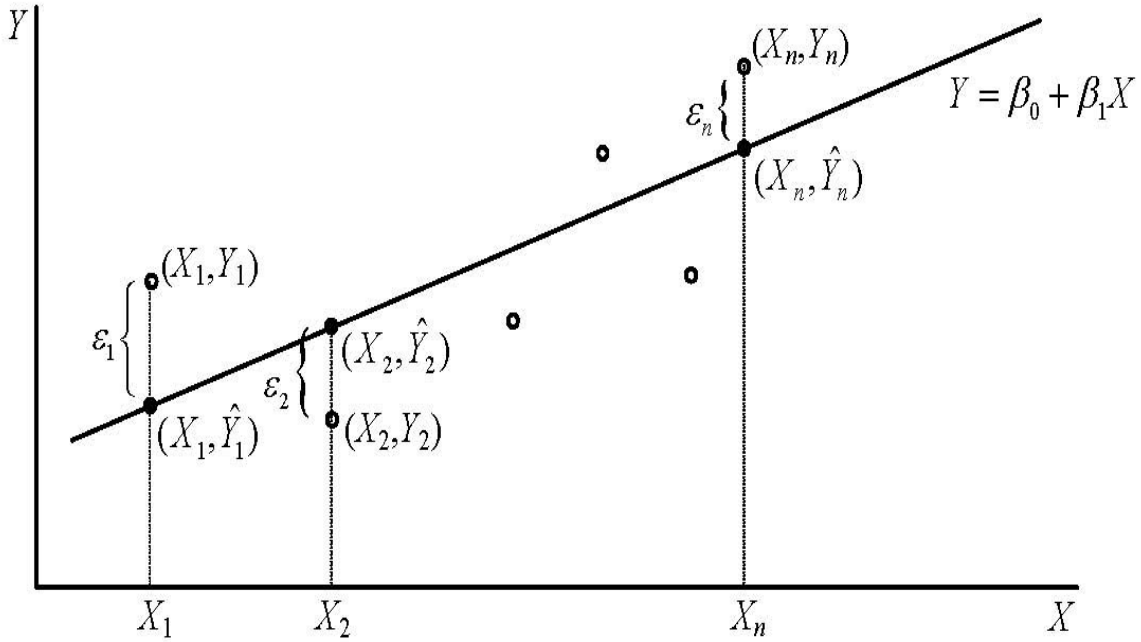
- Υπάρχουν πολλοί τρόποι να προσδιορίσουμε μια εκτιμήτρια της ευθείας $E(Y)=a+\beta X$ που παριστάνεται από την εξίσωση $\hat{Y} = \hat{a} + \hat{\beta}X$

Απλή Γραμμική Παλινδρόμηση

- Ας θεωρήσουμε το διπλανό διάγραμμα διασποράς κάποιων διδιάστατων παρατηρήσεων (X_1,Y_1), ..., (X_n,Y_n).
- Αν και είναι φανερό ότι υπάρχει μια τάση αύξησης των Y_i όταν αυξάνονται τα X_i, η σχέση που έχουν δεν φαίνεται να καθορίζει μια τέλεια ευθεία.
- Αν θελήσουμε να φέρουμε μια ευθεία επάνω από τα δεδομένα, η οποία να δείχνει την «τάση» μεταξύ των Y_i και X_i, τότε μπορεί να βρούμε αρκετές υποψήφιες «καλές» ευθείες πολύ διαφορετικές μεταξύ τους.



Το κριτήριο επιλογής της βέλτιστης ευθείας



- Θεωρούμε το διπλανό διάγραμμα διασποράς κάποιων διδιάστατων παρατηρήσεων (X_1,Y_1), ..., (X_n,Y_n).
- Η ευθεία γραμμικής παλινδρόμησης, με όποιο κριτήριο και αν καθορίζεται αυτή, δεν μπορεί να περάσει και από τα n σημεία.

- Στις περισσότερες περιπτώσεις, σε κάθε σημείο X_i η ευθεία θα αντιστοιχήσει ένα εκτιμημένο σημείο Y_i. Η διαφορά των πραγματικών Y_i από τα εκτιμημένα ονομάζεται «κατάλοιπο» ή «υπόλοιπο» (residual) και συμβολίζεται με $\varepsilon_i = Y_i - \hat{Y}_i$.
- Η ευθεία που προσαρμόζεται καλύτερα στα δεδομένα είναι, σύμφωνα με τη **μέθοδο ελαχίστων τετραγώνων**, αυτή που ελαχιστοποιεί το άθροισμα των τετραγώνων των καταλοίπων

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

Μεταβλητότητα – Διακύμανση

- Η ολική διακύμανση των Y_i, δίνεται από τη σχέση

$$SST = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} [(Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y})]^2$$

ή

$$SST = SSE + SSR,$$

όπου $SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ και $SSR = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \overline{Y}) = 0$$

SSR: η μεταβλητότητα που ερμηνεύεται από το μοντέλο παλινδρόμησης που έχουμε εκτιμήσει

SSE: η μεταβλητότητα που δεν ερμηνεύεται από το μοντέλο παλινδρόμησης

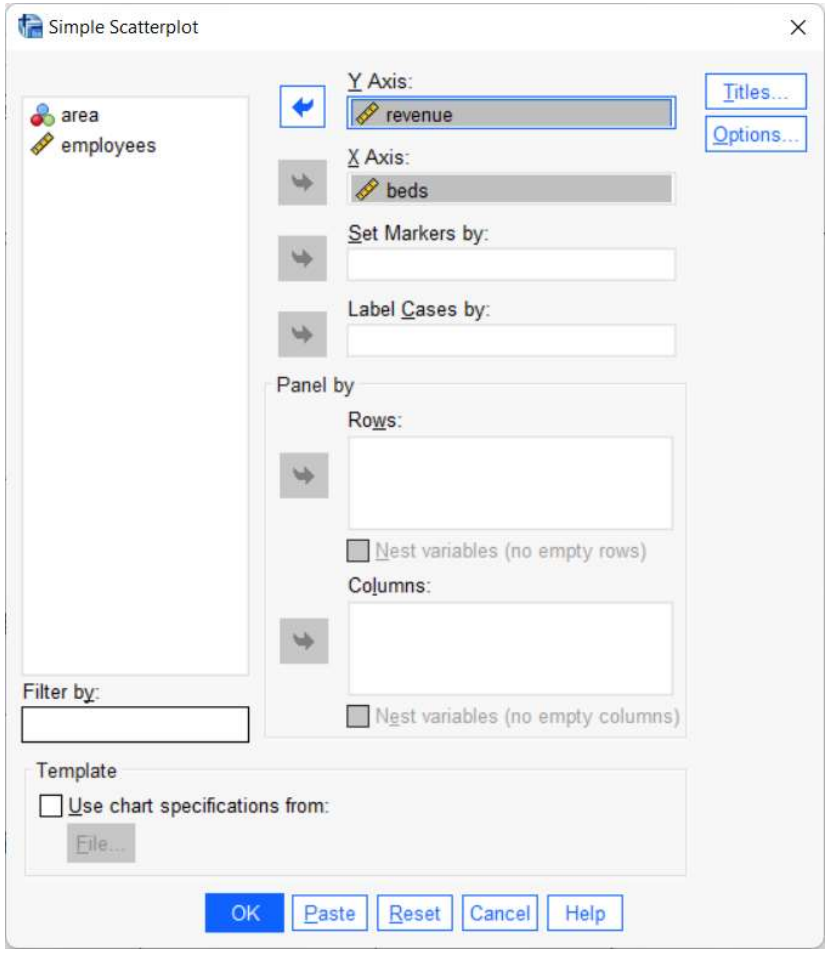
Αξιολόγηση του μοντέλου

- **Συντελεστής προσδιορισμού**
  - Επειδή στόχος μας είναι να μειώσουμε την ανερμήνευτη μεταβλητότητα, το μοντέλο θα είναι πιο χρήσιμο όταν η *SSR* είναι αυξημένη σε σχέση με την *SSE*, δηλαδή όταν το πηλίκο

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = r_{XY}^2$$

    είναι όσο το δυνατόν μεγαλύτερο
  - Παίρνει τιμές μεταξύ του 0 και του 1, με το 1 να σημαίνει πλήρη προσαρμογή του μοντέλου
  - Ο συντελεστής προσδιορισμού δίνει το ποσοστό της μεταβλητότητας μεταξύ των παρατηρημένων τιμών της *Y*, το οποίο ερμηνεύεται από το εκτιμημένο μοντέλο
  - Από τον παραπάνω τύπο προκύπτει ότι το R^2 ισούται με το τετράγωνο του συντελεστή γραμμικής συσχέτισης

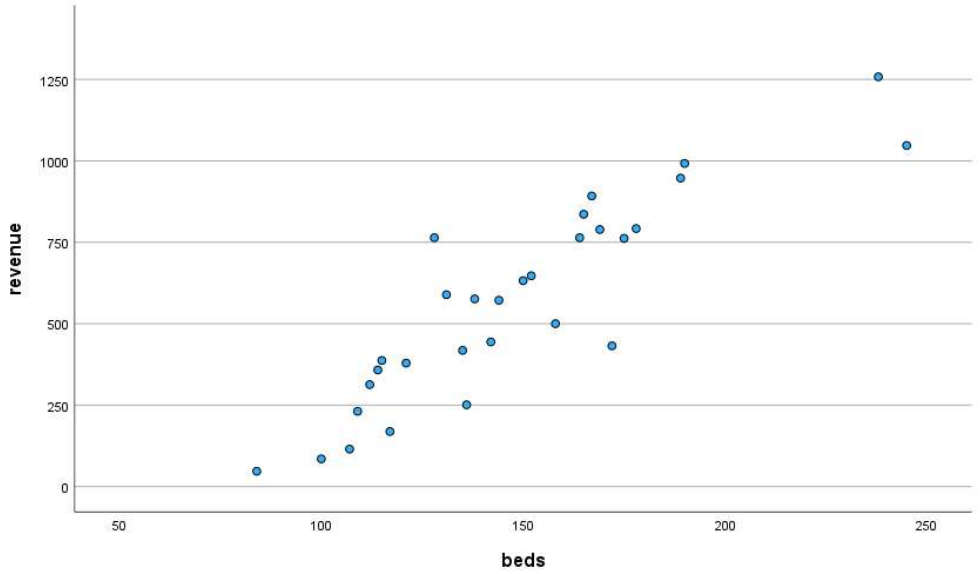
Παράδειγμα



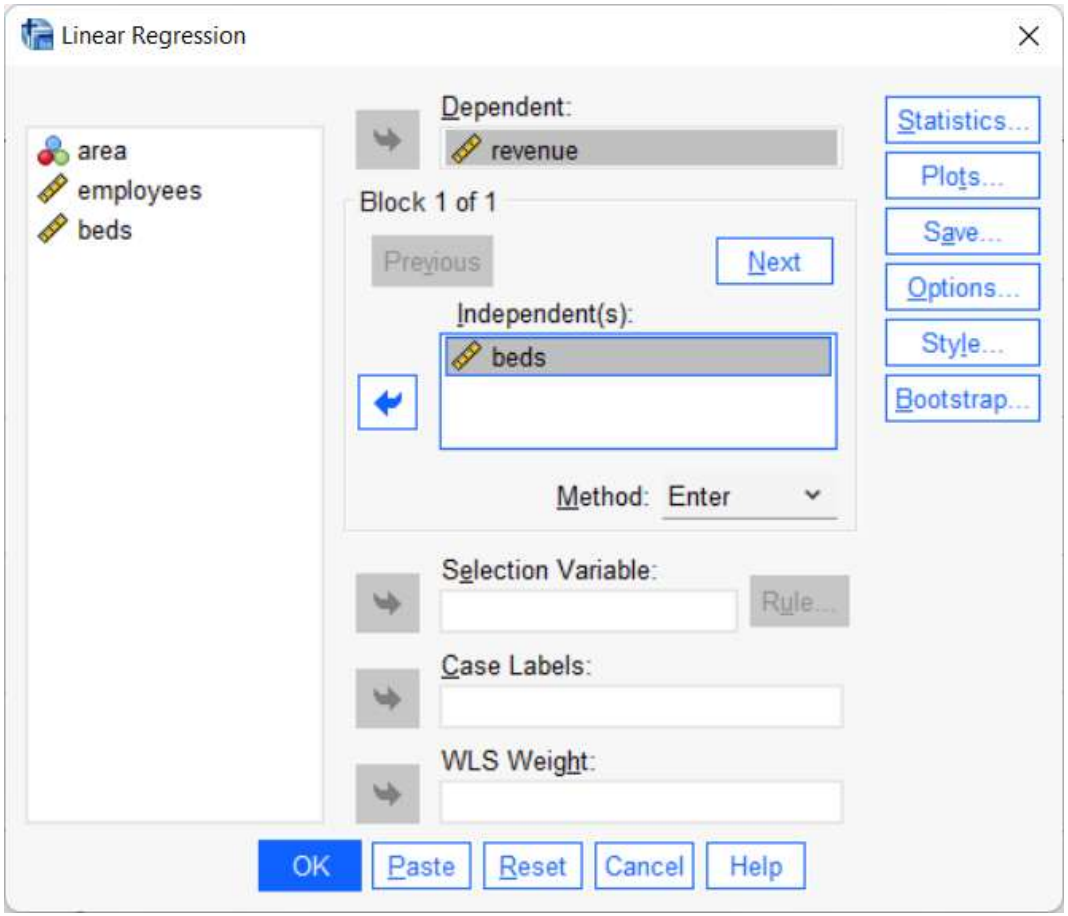
Graphs → Legacy Dialogs → Scatter/Dot → Simple Scatter

Y Axis: revenue

X Axis: beds



Παράδειγμα



Analyze → Regression → Linear

Dependent: revenue

Independent(s): beds

Παράδειγμα

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,894[a]	,800	,793	140,214

a. Predictors: (Constant), beds

Το μοντέλο ερμηνεύει το 80,0% της συνολικής διασποράς

Το p-value (Sig.) είναι μικρότερο από το 0,05, οπότε το μοντέλο είναι στατιστικά σημαντικό (η κλίση της ευθείας β δεν ισούται με το 0)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2201330,488	1	2201330,488	111,970	<,001[b]
	Residual	550481,379	28	19660,049		
	Total	2751811,867	29			

a. Dependent Variable: revenue

b. Predictors: (Constant), beds

Παράδειγμα

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-524,583	106,220		-4,939	<,001
	beds	7,362	,696	,894	10,582	<,001

a. Dependent Variable: revenue

Ούτε η σταθερά (Constant) ούτε ο συντελεστής της beds ισούνται με το 0

$$revenue(predicted) = -524{,}583 + 7{,}362 beds$$

Υποθέσεις της Γραμμικής Παλινδρόμησης

- Ύπαρξη γραμμικής σχέσης
- Κανονικότητα των καταλοίπων
- Μέση Τιμή των καταλοίπων ίση με το 0
- Ομοσκεδαστικότητα των καταλοίπων (σταθερή διασπορά)
- Ανεξαρτησία των καταλοίπων

Πολλαπλή Γραμμική Παλινδρόμηση

- Το μοντέλο της πολλαπλής παλινδρόμησης γράφεται ως

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

- Προσπαθεί να ερμηνεύσει τη σχέση μιας (εξαρτημένης) μεταβλητής με μια σειρά από πολλές ανεξάρτητες μεταξύ τους μεταβλητές.

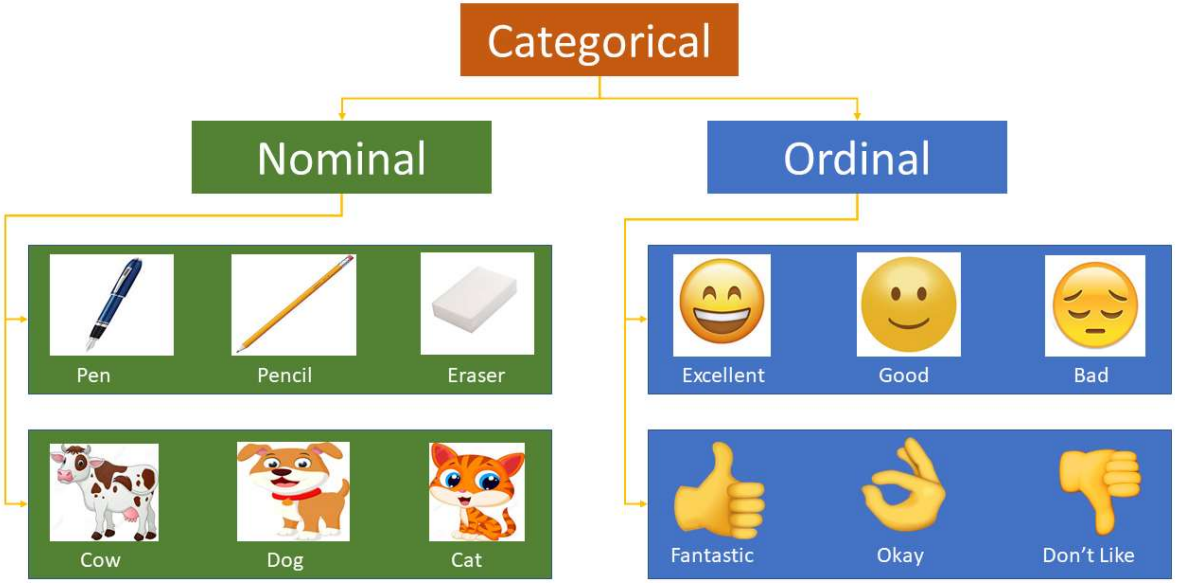
Πολλαπλή Γραμμική Παλινδρόμηση

- Το μοντέλο της πολλαπλής παλινδρόμησης γράφεται ως

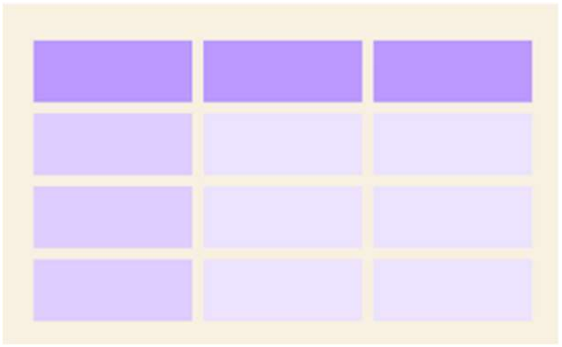
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

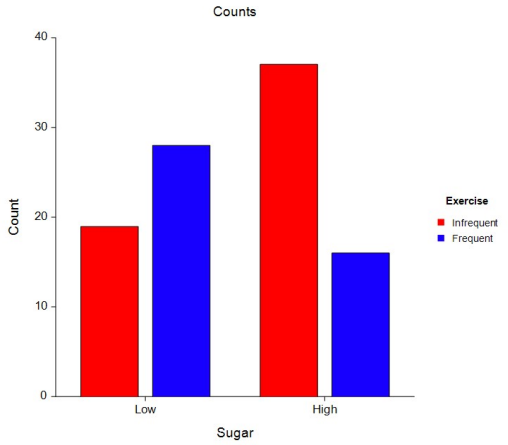
- Η σταθερά β_0 είναι η τιμή της Y όταν όλες οι ανεξάρτητες μεταβλητές πάρουν την τιμή 0
- Η παράμετρος β_1 παριστάνει τη μεταβολή που επιφέρει στην Y η αλλαγή της X_1 κατά μια μονάδα όταν όλες οι άλλες ανεξάρτητες μεταβλητές παραμείνουν σταθερές
- Η παράμετρος β_2 παριστάνει τη μεταβολή που επιφέρει στην Y η αλλαγή της X_2 κατά μια μονάδα όταν όλες οι άλλες ανεξάρτητες μεταβλητές παραμείνουν σταθερές
- ...

Ανάλυση Κατηγορικών Δεδομένων



Παρουσίαση Κατηγορικών Δεδομένων





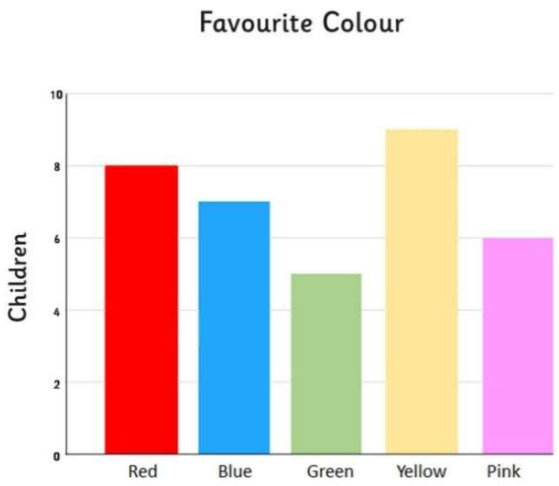
Παρουσίαση Κατηγορικών Δεδομένων

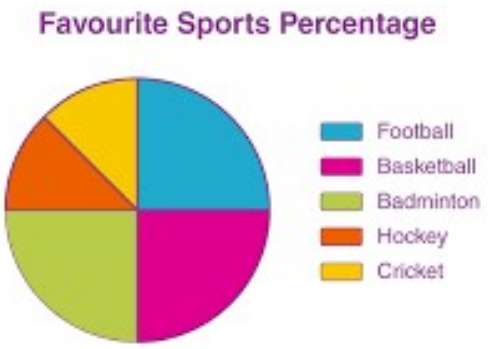
- Πίνακες συχνοτήτων
- Γραφήματα

Μία μεταβλητή

Vehicle	Frequency
Bike	3
Bus	2
Car	12
Lorry	3
Van	5
	Total = 25

Satisfaction	Frequency
Very satisfied	59
Satisfied	42
Neutral	12
Dissatisfied	8
Very dissatisfied	5



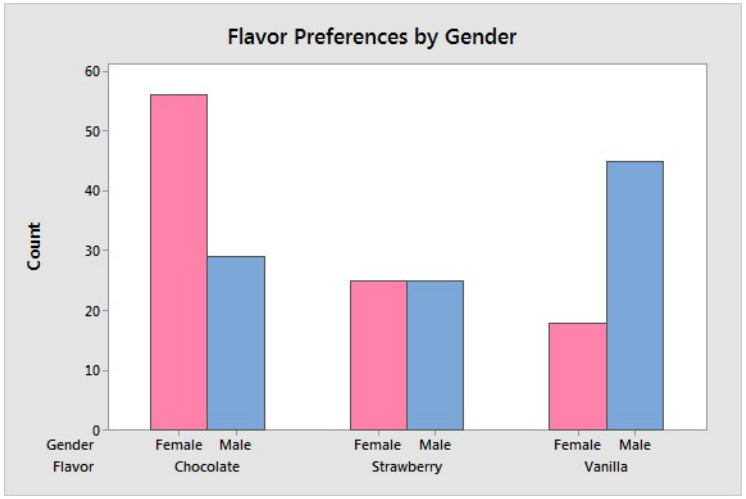


Παρουσίαση Κατηγορικών Δεδομένων

- Πίνακες συνάφειας
- Γραφήματα
  - Συνδυασμένα ραβδογράμματα

Δύο μεταβλητές

		Sport Preference			
		Archery	Boxing	Cycling	
Gender	Female	35	15	50	100
	Male	10	30	60	100
		45	45	110	200



Πίνακες συνάφειας

- **Πίνακας συνάφειας** είναι ένας πίνακας συχνοτήτων που προκύπτει αν ταξινομήσουμε ταυτόχρονα άτομα ή περιπτώσεις σύμφωνα με τις τιμές δυο ή περισσοτέρων ποιοτικών μεταβλητών
- Χρησιμοποιούνται συνήθως για:
  - σύγκριση δυο ή περισσοτέρων θεραπειών
  - αξιολόγηση διαγνωστικών ελέγχων
  - εύρεση συγχυτικών παραγόντων
- Χρήσιμο εργαλείο στην Επιδημιολογία
  - διερεύνηση για την ύπαρξη σχέσης μεταξύ δυο ή περισσοτέρων ποιοτικών μεταβλητών
  - ανίχνευση σχέσης παράγοντα - αποτελέσματος

Πίνακες συνάφειας

- **Παραδείγματα**

		Κάπνισμα		
		Ναι	Όχι	**Σύνολο**
Φύλο	Άνδρας	58	32	**90**
	Γυναίκα	41	67	**112**
	Σύνολο	**99**	**103**	**202**

		Ομάδα αίματος				
		A	B	O	AB	**Σύνολο**
Χρώμα μαλλιών	Μαύρο	63	44	78	56	**241**
	Ξανθό	71	53	69	62	**255**
	Σύνολο	**134**	**97**	**147**	**118**	**496**

Πίνακες συνάφειας

- Η πιο απλή μορφή πίνακα συνάφειας είναι ο 2×2 πίνακας

		Μεταβλητή 2	
		E	***A***
Μεταβλητή 1	***E***	O_{11}	O_{12}
	A	O_{21}	O_{22}

- Με O_{ij} συμβολίζουμε τη συχνότητα εμφάνισης του κελιού (i,j) δηλαδή αυτού που αποτελείται από τη γραμμή i και τη στήλη j
- Τα αθροίσματα των γραμμών και των στηλών ονομάζονται περιθώρια αθροίσματα
  - $O_{11}+O_{12}=O_{1.}$
  - $O_{21}+O_{22}=O_{2.}$
  - $O_{11}+O_{21}=O_{.1}$
  - $O_{12}+O_{22}=O_{.2}$

Πίνακες συνάφειας

- **Παράδειγμα:** Σε μια μελέτη σχετική με μια ασθένεια καταγράψαμε πόσα από τα άτομα που εμφανίζονται στο δείγμα και νοσούν είναι καπνιστές και πόσα μη καπνιστές καθώς και πόσα από τα άτομα που εμφανίζονται στο δείγμα και δεν νοσούν είναι καπνιστές και πόσα μη καπνιστές

- Συγκεκριμένα βρήκαμε:
  - **20 καπνιστές που νοσούν**
  - **40 καπνιστές που δεν νοσούν**
  - **35 μη καπνιστές που νοσούν**
  - **55 μη καπνιστές που δεν νοσούν**
- Αυτές είναι οι παρατηρημένες συχνότητες

	Ασθένεια		
Κάπνισμα	Ασθενής	Υγιής	**Σύνολο**
Καπνιστής	**20**	**40**	**60**
Μη Καπνιστής	**35**	**55**	**90**
Σύνολο	**55**	**95**	**150**

Πίνακες συνάφειας
Αναμενόμενες Συχνότητες

- Οι αναμενόμενες συχνότητες είναι οι συχνότητες που θα περιμέναμε να παρατηρήσουμε εάν η ασθένεια και το κάπνισμα ήταν ανεξάρτητα μεταξύ τους

Συγκεκριμένα έχουμε:

- Για τους καπνιστές που νοσούν: 60×55/150 = 22
- Για τους καπνιστές που δεν νοσούν: 60×95/150 = 38
- Για τους μη καπνιστές που νοσούν: 90×55/150 = 33
- Για τους μη καπνιστές που δεν νοσούν: 90×95/150 = 57

- Για τον υπολογισμό, πολλαπλασιάζουμε το σύνολο της γραμμής επί το σύνολο της στήλης και διαιρούμε με το σύνολο του δείγματος

	Ασθένεια		
Κάπνισμα	Ασθενής	Υγιής	**Σύνολο**
Καπνιστής	**22**	**38**	**60**
Μη Καπνιστής	**33**	**57**	**90**
Σύνολο	**55**	**95**	**150**

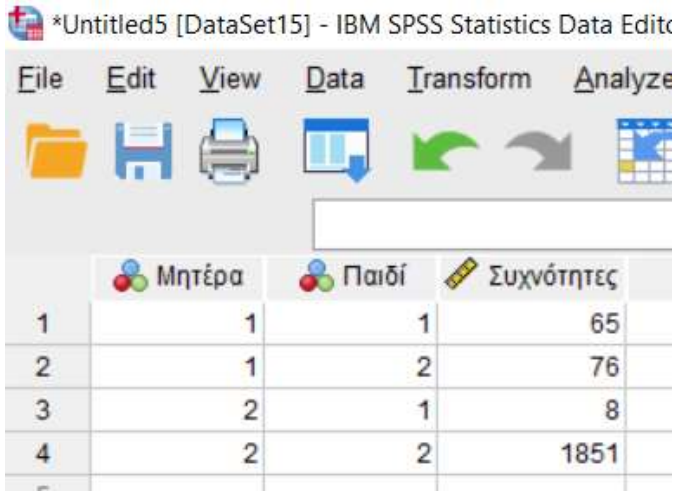
Παράδειγμα

Έστω ότι μας ενδιαφέρει να ελέγξουμε αν υπάρχει σχέση μεταξύ δυο μεταβλητών π.χ. ο μογγολισμός στα παιδιά με την ηπατίτιδα της μητέρας

	Παιδί (πραγματικότητα)		
Μητέρα (τεστ)	μογγολοειδές (+)	μη μογγολοειδές (-)	Σύνολο
ηπατίτιδα (+)	65	76	141
όχι ηπατίτιδα (-)	8	1851	1859
Σύνολο	73	1927	2000

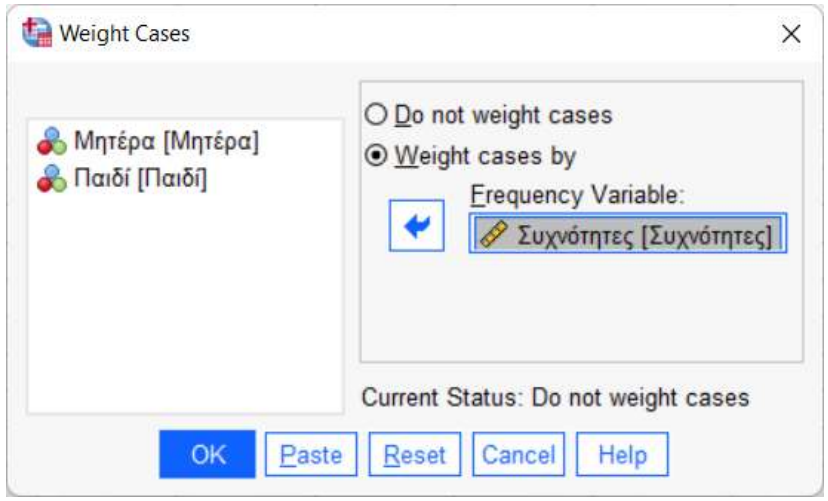
Παράδειγμα

Έστω ότι μας ενδιαφέρει να ελέγξουμε αν υπάρχει σχέση μεταξύ δυο μεταβλητών π.χ. ο μογγολισμός στα παιδιά με την ηπατίτιδα της μητέρας



Data → Weight Cases

Weight cases by: Συχνότητες

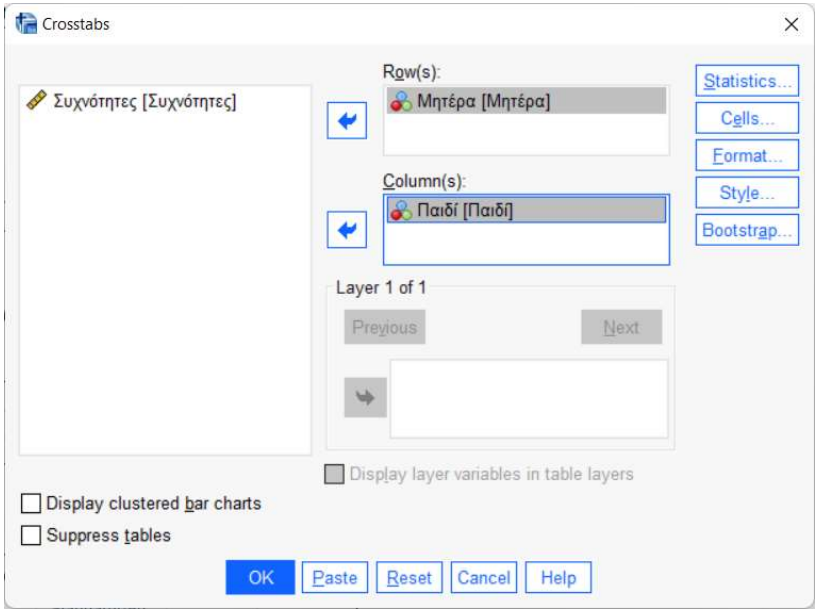


	Παιδί (πραγματικότητα)		
Μητέρα (τεστ)	μογγολοειδές (+)	μη μογγολοειδές (-)	Σύνολο
ηπατίτιδα (+)	65	76	141
όχι ηπατίτιδα (-)	8	1851	1859
Σύνολο	73	1927	2000

Παράδειγμα

Έστω ότι μας ενδιαφέρει να ελέγξουμε αν υπάρχει σχέση μεταξύ δυο μεταβλητών π.χ. ο μογγολισμός στα παιδιά με την ηπατίτιδα της μητέρας

Analyze → Descriptive Statistics → Crosstabs



Row(s): Μητέρα

Column(s): Παιδί

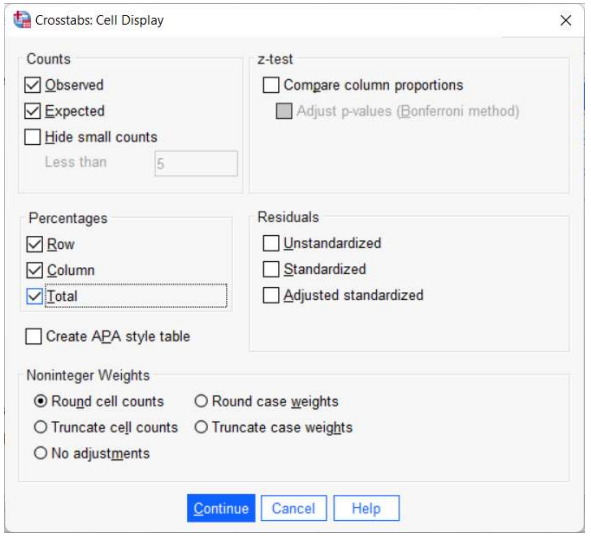
Cells

Counts:

Observed, Expected

Percentages

Row, Column, Total



Παράδειγμα

Μητέρα * Παιδί Crosstabulation

			Παιδί		
			Μογγολοειδές	Όχι μογγολοειδές	Total
Μητέρα	Ηπατίτιδα	Count	65	76	141
		Expected Count	5,1	135,9	141,0
		% within Μητέρα	46,1%	53,9%	100,0%
		% within Παιδί	89,0%	3,9%	7,0%
		% of Total	3,3%	3,8%	7,0%
	Όχι ηπατίτιδα	Count	8	1851	1859
		Expected Count	67,9	1791,1	1859,0
		% within Μητέρα	0,4%	99,6%	100,0%
		% within Παιδί	11,0%	96,1%	93,0%
		% of Total	0,4%	92,6%	93,0%
Total		Count	73	1927	2000
		Expected Count	73,0	1927,0	2000,0
		% within Μητέρα	3,7%	96,4%	100,0%
		% within Παιδί	100,0%	100,0%	100,0%
		% of Total	3,7%	96,4%	100,0%

Παρατηρούμενες τιμές (Count)

Αναμενόμενες τιμές (Expected Count)

Σχετικές συχνότητες (Row) (% within Μητέρα)

Σχετικές συχνότητες (Column) (% within Παιδί)

Σχετικές συχνότητες (Total) (% of Total)

Έλεγχος ανεξαρτησίας χ^2

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

Έλεγχος ανεξαρτησίας χ^2

- Για να εφαρμόσουμε τον έλεγχο χ^2 χρειαζόμαστε:

- Τις παρατηρημένες συχνότητες O_{ij}
  - Τις συχνότητες που έχουμε παρατηρήσει στο δείγμα μας

- Τις αναμενόμενες συχνότητες E_{ij}
  - Τις συχνότητες που αναμένουμε να παρατηρήσουμε εάν υποθέσουμε ότι οι δυο μεταβλητές είναι ανεξάρτητες

Έλεγχος ανεξαρτησίας χ^2

- Έλεγχος ανεξαρτησίας χ^2 σε 2x2 πίνακες συνάφειας
  - Υπάρχει εξάρτηση μεταξύ των δυο μεταβλητών;
  - $H_0: p_{ij}=p_{i.}p_{.j}$ ή $H_0: p_{i|j}=p_{i.}$ (Μηδενική Υπόθεση)

$$p_{ij}=\frac{O_{ij}}{n},\ p_{i.}=\frac{O_{i.}}{n},\ p_{.j}=\frac{O_{.j}}{n}$$

  - Η στατιστική συνάρτηση (ελεγχοσυνάρτηση) είναι η

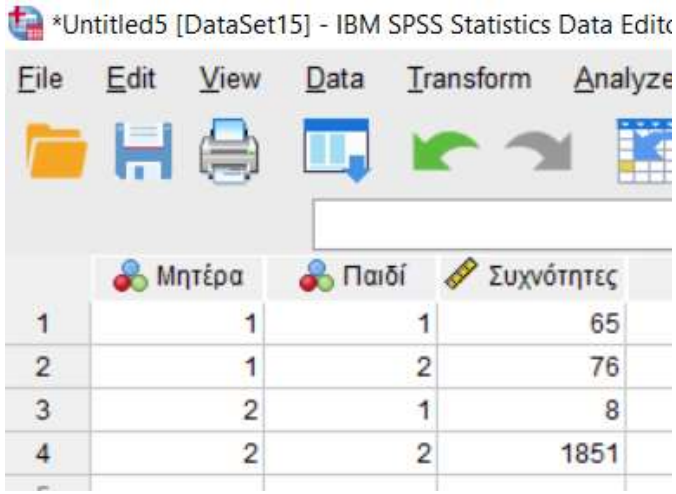
$$\chi^2=\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}}\sim\chi_1^2$$

  - Η κρίσιμη περιοχή (περιοχή απόρριψης) ορίζεται από τη σχέση

$$\chi^2\geq\chi_{a,1}^2$$

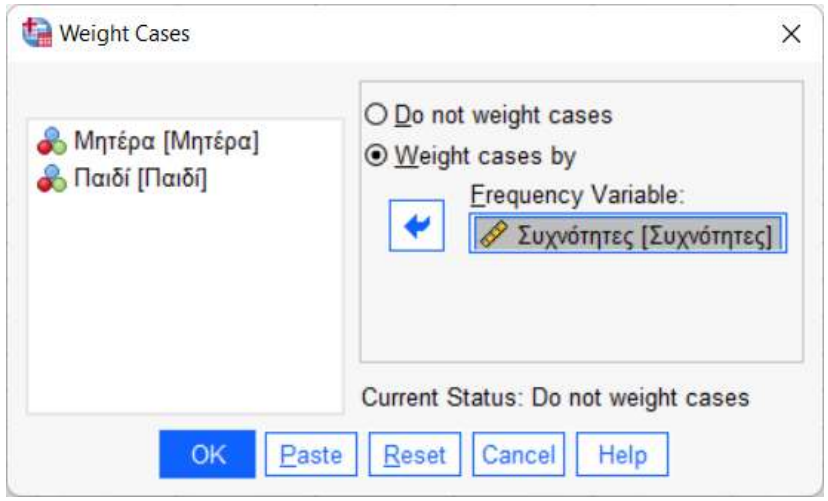
Παράδειγμα

Έστω ότι μας ενδιαφέρει να ελέγξουμε αν υπάρχει σχέση μεταξύ δυο μεταβλητών π.χ. ο μογγολισμός στα παιδιά με την ηπατίτιδα της μητέρας



Data → Weight Cases

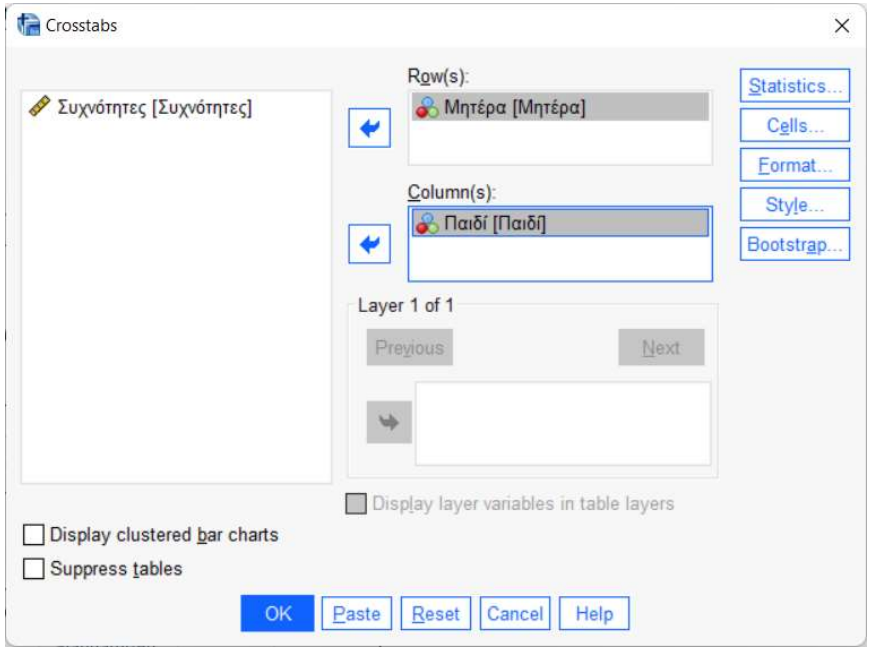
Weight cases by: Συχνότητες



	Παιδί (πραγματικότητα)		
Μητέρα (τεστ)	μογγολοειδές (+)	μη μογγολοειδές (-)	Σύνολο
ηπατίτιδα (+)	65	76	141
όχι ηπατίτιδα (-)	8	1851	1859
Σύνολο	73	1927	2000

Παράδειγμα

Έστω ότι μας ενδιαφέρει να ελέγξουμε αν υπάρχει σχέση μεταξύ δυο μεταβλητών π.χ. ο μογγολισμός στα παιδιά με την ηπατίτιδα της μητέρας



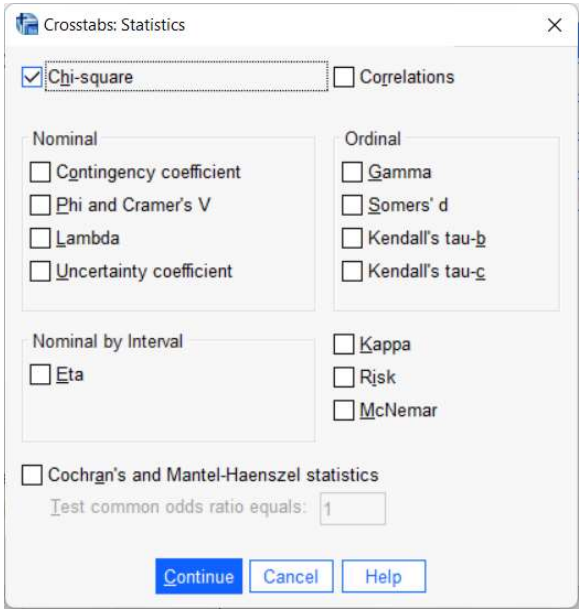
Analyze → Descriptive Statistics → Crosstabs

Row(s): Μητέρα

Column(s): Παιδί

Statistics

Chi-square



Παράδειγμα

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	777,259[a]	1	<,001		
Continuity Correction[b]	764,328	1	<,001		
Likelihood Ratio	328,880	1	<,001		
Fisher's Exact Test				<,001	<,001
Linear-by-Linear Association	776,871	1	<,001		
N of Valid Cases	2000				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 5,15.

b. Computed only for a 2x2 table

Επειδή το p-value είναι < 0,001 απορρίπτουμε την υπόθεση της ανεξαρτησίας. Συνεπώς η ασθένεια του παιδιού εξαρτάται από την ασθένεια της μητέρας.

Παρατήρηση: Το Linear-by-Linear Association χρησιμοποιείται όταν τουλάχιστον μια από τις δυο μεταβλητές είναι διατάξιμη

Κανόνας ορθής εφαρμογής του ελέγχου χ^2

Οι όροι της στατιστικής συνάρτησης χ^2 έχουν στον παρονομαστή τις αναμενόμενες συχνότητες $E_{ij}=np_{ij}$. Σε περίπτωση που τα p_{ij} είναι πολύ μικρά τότε οι E_{ij} είναι πολύ μικρές και έτσι αυξάνεται η τιμή του όρου εξαιτίας των μικρών τιμών των p_{ij} που ελέγχουμε και όχι λόγω απόκλισης των συχνοτήτων. Έτσι η προσέγγιση της κατανομής χ^2 δεν είναι ικανοποιητική και δεν πρέπει να εφαρμοστεί ο έλεγχος χ^2. Οι παρατηρήσεις αυτές οδήγησαν στον λεγόμενο κανόνα ορθής εφαρμογής, ο οποίος έχει ως εξής:

a) Το μέγεθος n του δείγματος δεν πρέπει να είναι μικρότερο του τετραπλάσιου του αριθμού των κελιών του πίνακα

b) Καμία από τις αναμενόμενες συχνότητες E_{ij} δεν πρέπει να είναι μικρότερη του 1

c) Το ποσοστό των αναμενόμενων συχνοτήτων E_{ij}, οι οποίες είναι μικρότερες του 5 δεν πρέπει να είναι μεγαλύτερο του 20% με 25%.

Τέλος Μαθήματος