Τ. Ε. Ι. Αθήνας

Τμήμα Ιατρικών Εργαστηρίων

ΒΙΟΣΤΑΤΙΣΤΙΚΗ

Δρ Ε. ΠΑΠΑΓΕΩΡΓΙΟΥ

Επίκουρος Καθηγήτρια

ΤΜΗΜΑ ΙΑΤΡΙΚΩΝ ΕΡΓΑΣΤΗΡΙΩΝ

Τ.Ε.Ι. ΑΘΗΝΑΣ

ΣΚΟΠΟΣ ΚΑΙ ΣΤΟΧΟΣ ΔΙΑΛΕΞΕΩΝ

Να γίνουν προσιτοί οι τρόποι:

- Οργάνωσης και παρουσίασης τύπων δεδομένων και
- ανάλυσης των διαφόρων τύπων δεδομένων

Να γίνουν κατανοητές οι έννοιες:

- της στατιστικής συμπερασματολογίας

Σύντομη Ιστορική Αναδρομή

Η επιδημιολογία διαμορφώθηκε ως επιστήμη τον μεσαίωνα για την μελέτη μεγάλων επιδημιών (χολέρα, ευλογιά, πανώλη). Ο πρώτος που ασχολήθηκε με τα αντικείμενα της «Επιδημιολογίας» είναι ο Ιπποκράτης (460-357 π.Χ) στο «Περί αέρος, ύδατος και τόπων». Εκεί κατέγραψε τις εμπειρικές σχέσεις μεταξύ συγκεκριμένων ασθενειών και του τόπου εκδήλωσης ή εμφάνισης, των συνθηκών διαβίωσης, της διατροφής, κατοικίας, κλίματος και άλλων αιτιών. Όλα αυτά αποτελούν το αντικείμενο της σύγχρονης περιγραφικής Επιδημιολογίας. Πολύ αργότερα ακολούθησαν ο ο Graunt (1620-1674) και ο Farr (1807-1883). Την ίδια εποχή, ο Snow (1813-1858) απέδειξε ότι η χολέρα προερχόταν από μικρόβιο που μεταδιδόταν από το νερό και την κοπριά. Στο αποτέλεσμα αυτό κατέληξε αφού μελέτησε το νερό από την κεντρική ύδρευση δύο περιοχών ίδιων χαρακτηριστικών (ηλικία, φύλο, κοινωνικοοικονομική κατάσταση):

το Lambeth (με καθαρό νερό) και το Soutwork (όπου το νερό περιείχε απόβλητα). Το Lambeth είχε 8 φορές μικρότερη θνησιμότητα από χολέρα. Άρα το νερό ήταν το κλειδί και όχι ο αέρας όπως πίστευαν εκείνη την εποχή.

Σύντομη Ιστορική Αναδρομή(συνέχεια)

Στον 20ο αιώνα είχαμε σημαντικές εξελίξεις στην Επιδημιολογία. Μια από τις πιο σημαντικές είναι η σύνδεση του καπνίσματος με τον καρκίνο του πνεύμονα από τον Doll, (για λεπτομέρειες βλ. Doll and Peto,1976).

Πολύ σημαντικό γεγονός είναι και η δημιουργία Σχολή Δημόσιας Υγιεινής το 1922 στο Πανεπιστήμιο του Harvard η οποία περιλαμβάνει Τμήματα Επιδημιολογίας και Βιοστατιστικής. Η σχολή αυτή συνέβαλε στον καθορισμό και στην προώθηση της επιστήμης της Επιδημιολογίας και καθιερώθηκε ως ένα από τα καλύτερα Σχολεία στον τομέα αυτό.

ΒΙΒΛΙΟΓΡΑΦΙΑ :

Ελληνική :

- Σταυρινός Βασίλης Γ., Παναγιωτάκος Δημοσθένης Β. Βιοστατιστική, Εκδόσεις Γ. Δαρδάνος - Κ. Δαρδάνος Ο.Ε.
- Τριχόπουλου Δ., Τζώνου Α., Κατσουγιάννη Κ., Βιοστατιστική, Εκδόσεις Παρισιάνου, 1993
- Τσίμπου Κ., Γεωργιακώδη Φ., Περιγραφική και Διερευνητική Στατιστική Ανάλυση Δεδομένων, Τόμος Α. Εκδόσεις Σταμούλη, 1999
- Τσίμπου Κ., Γεωργιακώδη Φ., Περιγραφική και Διερευνητική Στατιστική Ανάλυση Δεδομένων, Τόμος Β. Εκδόσεις Σταμούλη, 1999.
- Petrie Aviva, Sabin Caroline, Ιατρική Στατιστική με μια ματιά, Εκδόσεις Παρισιάνου, 2008

Ακολουθεί Ξενόγλωσση Βιβλιογραφία.

Με έντονα γράμματα (Bold) επισημαίνονται τα συγγράμματα τα οποία συνάδουν με την παρούσα παρουσίαση και βοηθούν σε μια εισαγωγική μελέτη ενώ τα υπόλοιπα παρατίθενται είτε για όσουν ενδιαφέρονται για περαιτέρω μελέτη ή εμβάνθυνση είτε ως εξειδικευμένα στατιστικά βιβλία.

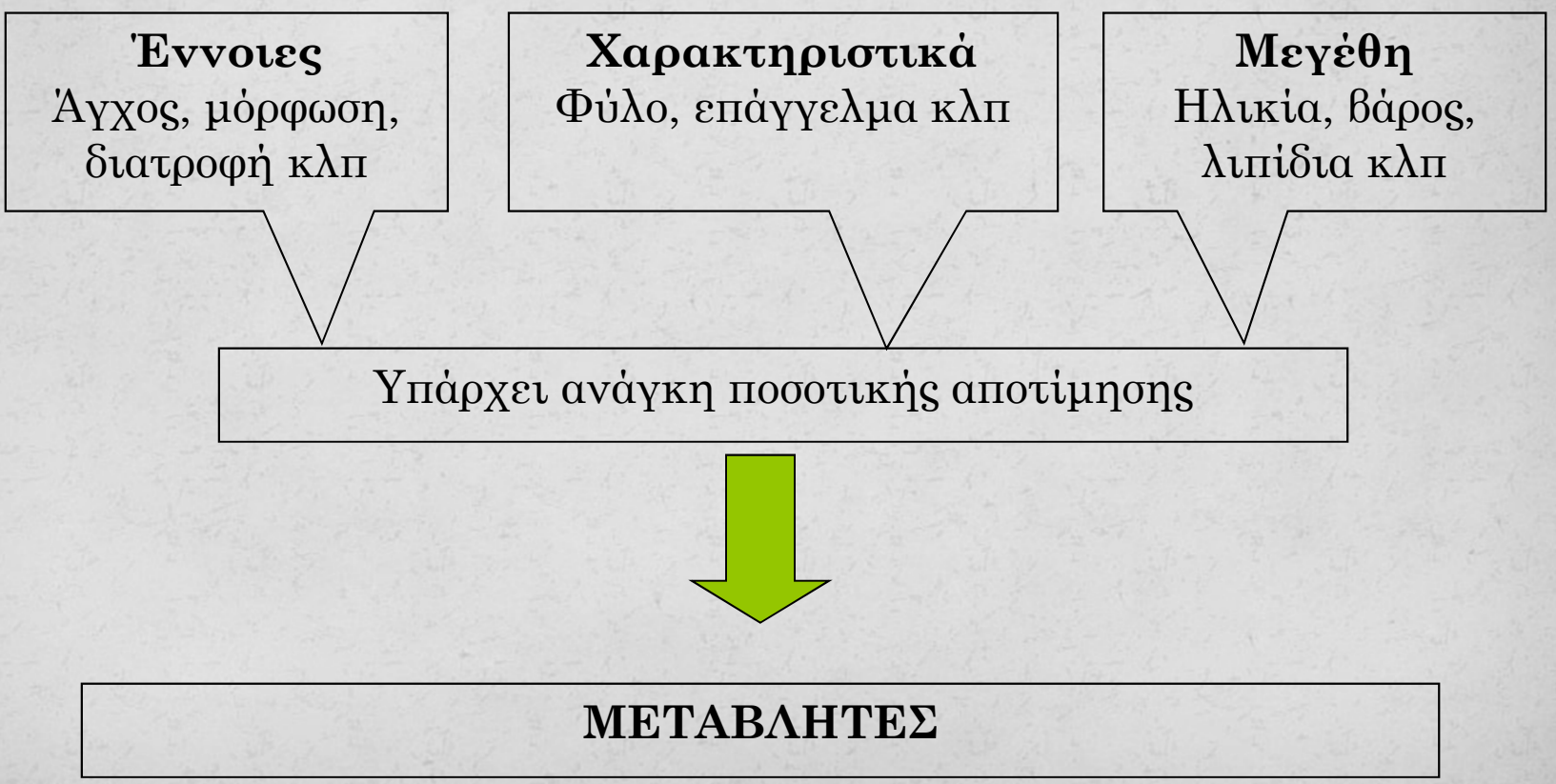
- **D.G. Altman (1992): Practical statistics for medical research. Chapman and Hall.**
- D. F. Andrews and A. M. Herzberg (1985): Data - A Collection of Problems from many Fields for the Student and Research Worker. Wiley, New York.
- Ralf Bender, Stefan Lange (2001): Adjusting for multiple testing— when and how? Journal of Clinical Epidemiology 54(4), 343–349.
- **M. Bland (1995): An Introduction to Medical Statistics. Second Edition. Oxford University Press.**
- J.M. Bland and D.G. Altman (1986): Statistical methods for assessing agreement between two methods of clinical measurement. Lancet, 1:307-310.
- M J. Campbell and D. Machin (1993): Medical Statistics – A Commonsense Approach. John Wiley & Sons, New York.
- B. Dawson and R.G. Trapp (2004): Basic & Clinical Biostatistics. Fourth Edition. McGraw-Hill.
- A.R. Feinstein, D.M. Sosin and C.K. Wells (1985): The Will Rogers phenomenon. Stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer. The New England Journal of Medicine, 312(25), 1604-1608.
- L.D. Fisher and G. van Belle (1993): Biostatistics - Methodology for the Health Sciences. Wiley, New York.
- S. Holm (1979): A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics, 6, 65-70.
- J.C. Hsu (1996): Multiple Comparisons. Theory and methods.Chapman and Hall.
- **M.H. Katz (1999): Multivariable Analysis. A Practical Guide for Clinicians. Cambridge University Press.**
- D. Kendrick, K. Fielding, E. Bentley, R. Kerslake, P. Miller, and M. Pringle. Radiography of the lumbar spine in primary care patients with low back pain: randomised controlled trial. British Medical Journal, 322:400-405, 2001.
- S. Landau and B.S. Everitt (2004): A Handbook of Statistical Analyses
- using SPSS. Chapman & Hall/CRC.

- **H. Motulsky (1995): Intuitive Biostatistics. Oxford University Press.**
- J. Pallant (2005): SPSS survival manual. 2nd edition. Open University Press.
- M.F. Schilling, A.E. Watkins, and W. Watkins. Is human height bimodal? The American Statistician, 56:223-229, 2002.
- **M. Schumacher und G. Schulgen (2002): Methodik klinischer Studien. Methodische Grundlagen der Planung, Durchführung und Auswertung. Springer-Verlag (*German*).**
- J.P. Shaffer (1986): Modified Sequentially Rejective Multiple Test Procedures. Journal of the American Statistical Association, 81(395), 826-831.
- G.W. Snedecor and W.G. Cochran (1989): Statistical methods. 8th edition. Iowa State University Press.
- Y.-K. Tu, Z.L. Nelson-Moon, and M.S. Gilthorpe. Misuses of correlation and regression analyses in orthodontic research: The problem of mathematical coupling. American Journal of Orthodontics & Dentofacial Orthopedics, 130:62-68, 2006.
- Y.-K. Tu and M.S. Gilthorpe. Revisiting the relation between change and initial value: A review and evaluation. Statistics in Medicine, 26:443-457, 2007.

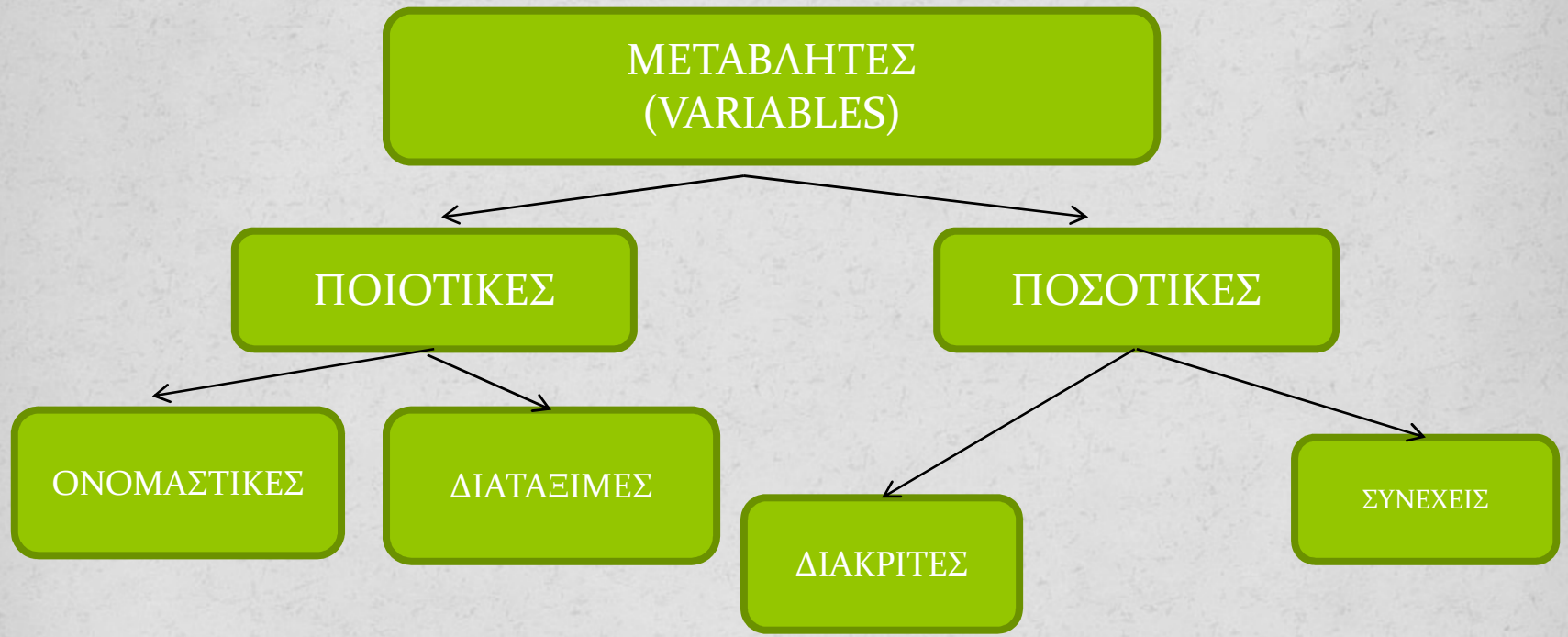
Σύνοψη της Παρουσίασης

- **Εισαγωγικές Έννοιες-Περιγραφική Στατιστική**
- **Ανάλυση Συνεχών Μεταβλητών**
- **Ανάλυση Κατηγορικών Δεδομένων**
- **Έλεγχοι Υποθέσεων – Διαστήματα Εμπιστοσύνης**
- **Συσχέτιση**
- **Γραμμική Παλινδρόμηση**

Η ανάγκη χρήσης μεταβλητών



1. Περιγραφική Στατιστική



1. Περιγραφική Στατιστική

Παράδειγμα: Σε ένα δείγμα 20 οικογενειών από μια περιοχή της Αθήνας, το επάγγελμα του πατέρα, το ημερομίσθιο του πατέρα και ο αριθμός παιδιών της οικογένειας ήταν:

επάγγελμα	ημερομίσθιο	αρ_παιδιών
εργάτης	70	0
οδηγός	75	1
εργάτης	80	0
δημ_υπάλληλος	70	2
δημ_υπάλληλος	80	2
δημ_υπάλληλος	50	2
δάσκαλος	90	3
ιερέας	100	2
οδηγός	60	4
εργάτης	60	1
δάσκαλος	70	1
εργάτης	60	2
εργάτης	80	3
δημ_υπάλληλος	70	4
ιερέας	90	1
δάσκαλος	100	2
εργάτης	90	2
δημ_υπάλληλος	65	2
δάσκαλος	75	2
δημ_υπάλληλος	80	2

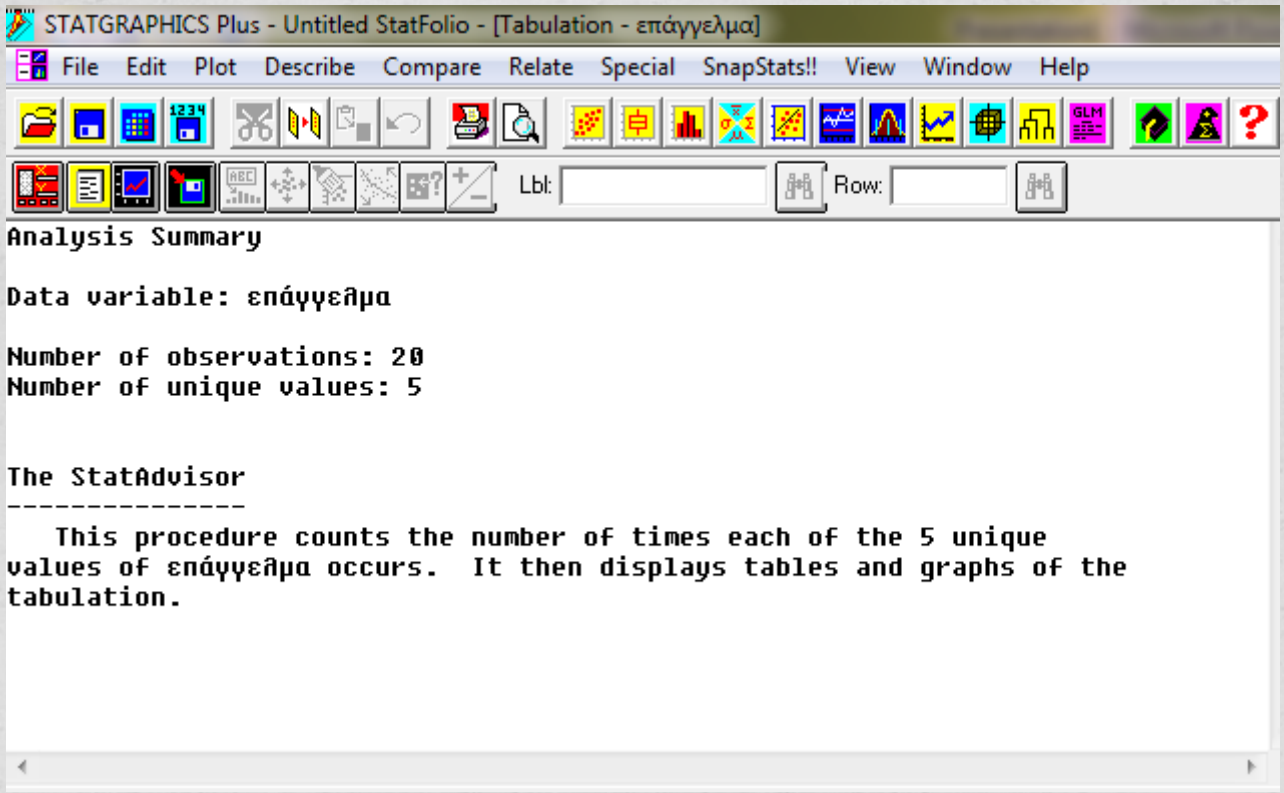
1. Περιγραφική Στατιστική

Στο παράδειγμα 1.1. για να εισάγουμε τη μεταβλητή «επάγγελμα», επιλέγουμε μεταβλητή τύπου character αφού πρόκειται για ποιοτική μεταβλητή, την ονομάζουμε «επάγγελμα» και στη συνέχεια πληκτρολογούμε τα δεδομένα μας:

	επάγγελμα	Col_2	Col_3	Col_4	Col_5
1	εργάτης				
2	οδηγός				
3	εργάτης				
4	δημ_υπάλληλος				
5	δημ_υπάλληλος				
6	δημ_υπάλληλος				
7	δάσκαλος				
8	ιερέας				
9	οδηγός				
10	εργάτης				
11	δάσκαλος				
12	εργάτης				
13	εργάτης				
14	δημ_υπάλληλος				
15	ιερέας				
16	δάσκαλος				
17	εργάτης				
18	δημ_υπάλληλος				
19	δάσκαλος				
20	δημ_υπάλληλος				

1. Περιγραφική Στατιστική

Η ανάλυση που προκύπτει άμεσα είναι η ακόλουθη:



1. Περιγραφική Στατιστική

Και ο πίνακας συχνοτήτων:

Frequency Table for επάγγελμα

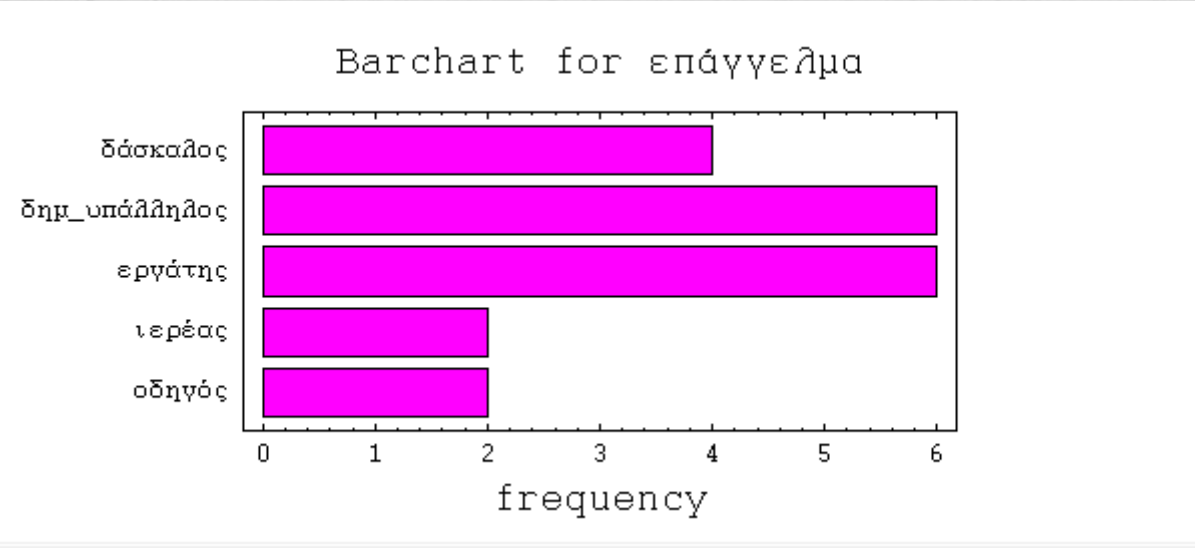
Class	Value	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
1	δάσκαλος	4	0,2000	4	0,2000
2	δημ_υπάλληλος	6	0,3000	10	0,5000
3	εργάτης	6	0,3000	16	0,8000
4	ιερέας	2	0,1000	18	0,9000
5	οδηγός	2	0,1000	20	1,0000

The StatAdvisor

This table shows the number of times each value of επάγγελμα occurred, as well as percentages and cumulative statistics. For example, in 4 rows of the data file επάγγελμα equaled δάσκαλος. This represents 20,0% of the 20 values in the file. The rightmost two columns give cumulative counts and percentages from the top of the

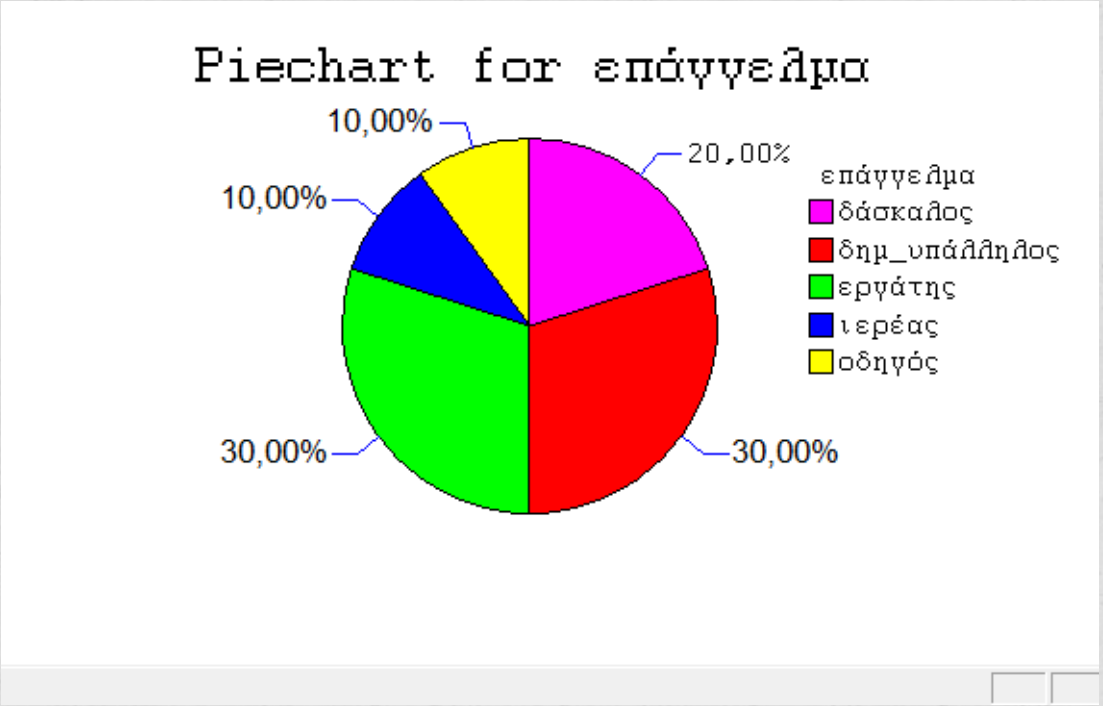
1. Περιγραφική Στατιστική

Και το ραβδόγραμμα συχνοτήτων:



1. Περιγραφική Στατιστική

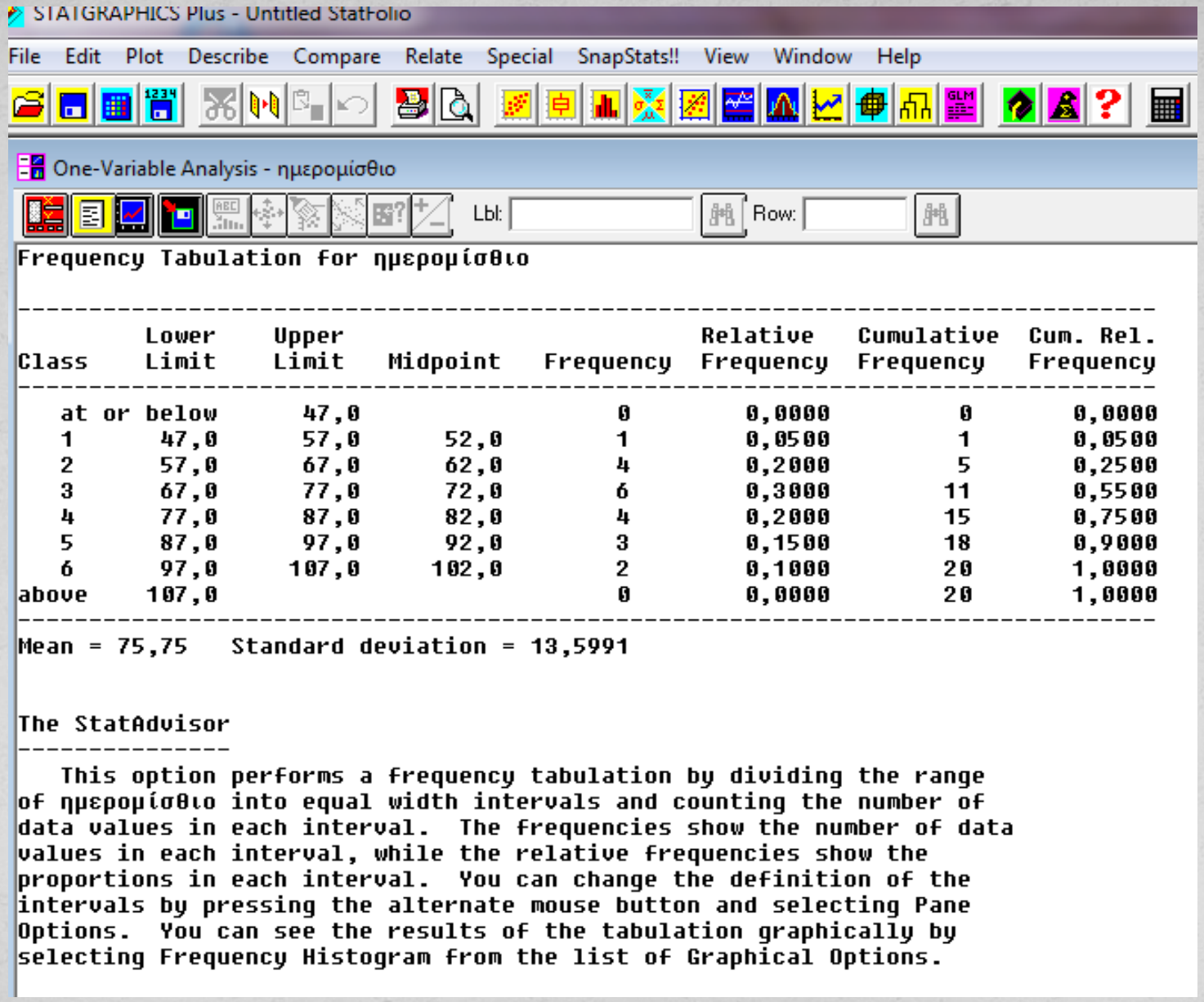
Και το κυκλόγραμμα (πίτα) συχνοτήτων:



1. Περιγραφική Στατιστική

Συνεπώς έχουμε εισάγει τις δύο μεταβλητές και έχουμε αυτή την εικόνα:

	επάγγελμα	ημερομίσθιο
1	εργάτης	70
2	οδηγός	75
3	εργάτης	80
4	δημ_υπάλληλος	70
5	δημ_υπάλληλος	80
6	δημ_υπάλληλος	50
7	δάσκαλος	90
8	ιερέας	100
9	οδηγός	60
10	εργάτης	60
11	δάσκαλος	70
12	εργάτης	60
13	εργάτης	80
14	δημ_υπάλληλος	70
15	ιερέας	90
16	δάσκαλος	100
17	εργάτης	90
18	δημ_υπάλληλος	65
19	δάσκαλος	75
20	δημ_υπάλληλος	80



Frequency Tabulation for ημερομίσθιο

Class	Lower Limit	Upper Limit	Midpoint	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
at or below		47,0		0	0,0000	0	0,0000
1	47,0	57,0	52,0	1	0,0500	1	0,0500
2	57,0	67,0	62,0	4	0,2000	5	0,2500
3	67,0	77,0	72,0	6	0,3000	11	0,5500
4	77,0	87,0	82,0	4	0,2000	15	0,7500
5	87,0	97,0	92,0	3	0,1500	18	0,9000
6	97,0	107,0	102,0	2	0,1000	20	1,0000
above	107,0			0	0,0000	20	1,0000

Mean = 75,75 Standard deviation = 13,5991

The StatAdvisor

This option performs a frequency tabulation by dividing the range of ημερομίσθιο into equal width intervals and counting the number of data values in each interval. The frequencies show the number of data values in each interval, while the relative frequencies show the proportions in each interval. You can change the definition of the intervals by pressing the alternate mouse button and selecting Pane Options. You can see the results of the tabulation graphically by selecting Frequency Histogram from the list of Graphical Options.

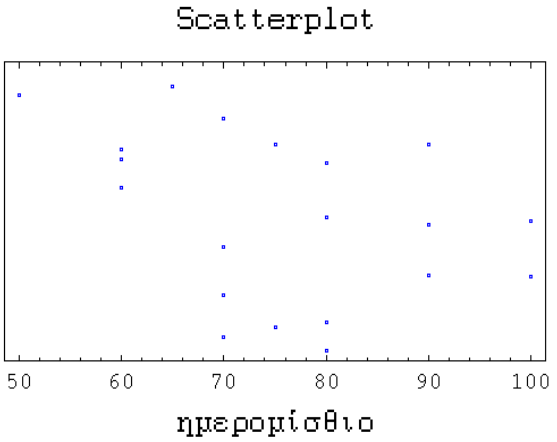
1. Περιγραφική Στατιστική

By default εμφανίζεται η παρακάτω ανάλυση:

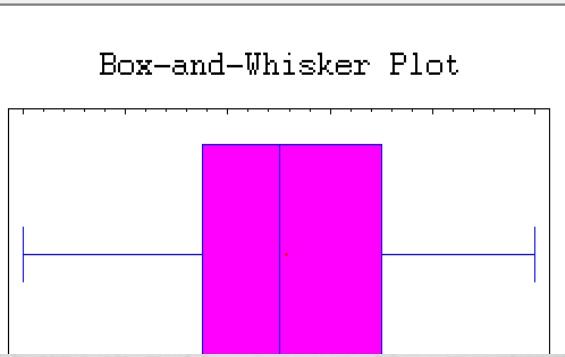
One-Variable Analysis - ημερομίσθιο

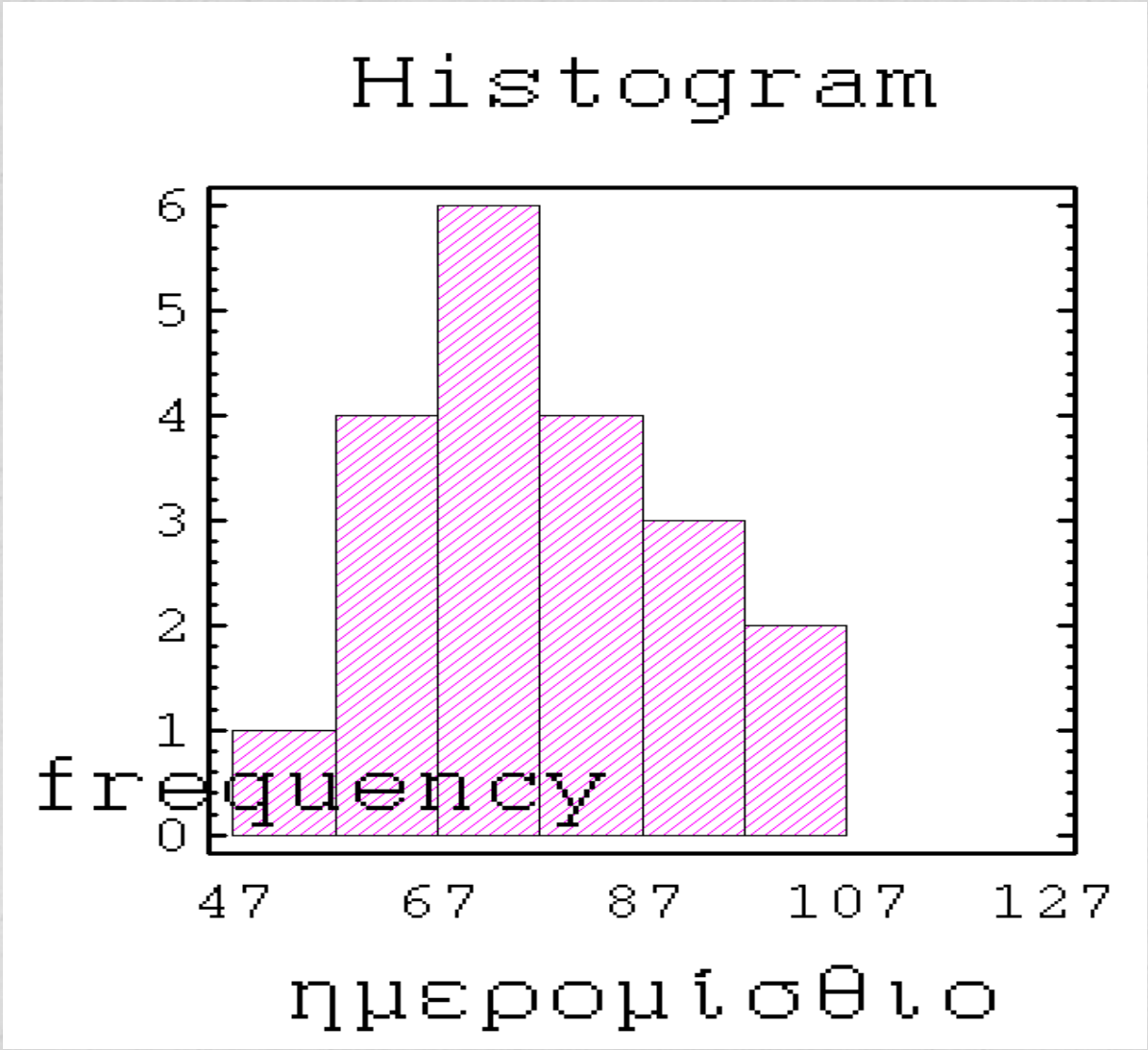
Lbl: Row:

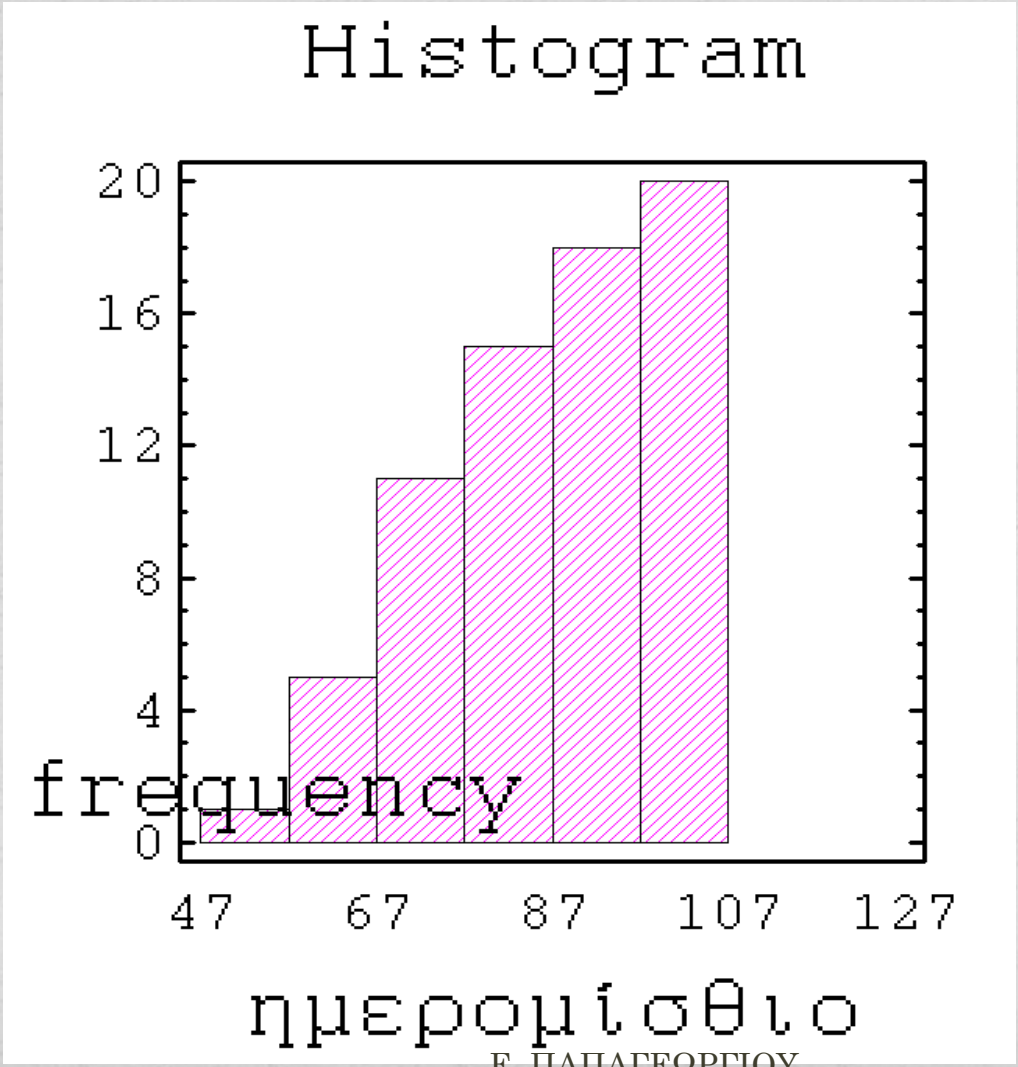
```
Analysis Summary

Data variable: ημερομίσθιο

20 values ranging from 50,0 to 100,0


The StatAdvisor
---------------
   This procedure is designed to summarize a single sample of data.
It will calculate various statistics and graphs.  Also included in the
procedure are confidence intervals and hypothesis tests.  Use the
Tabular Options and Graphical Options buttons on the analysis toolbar
to access these different procedures.
```

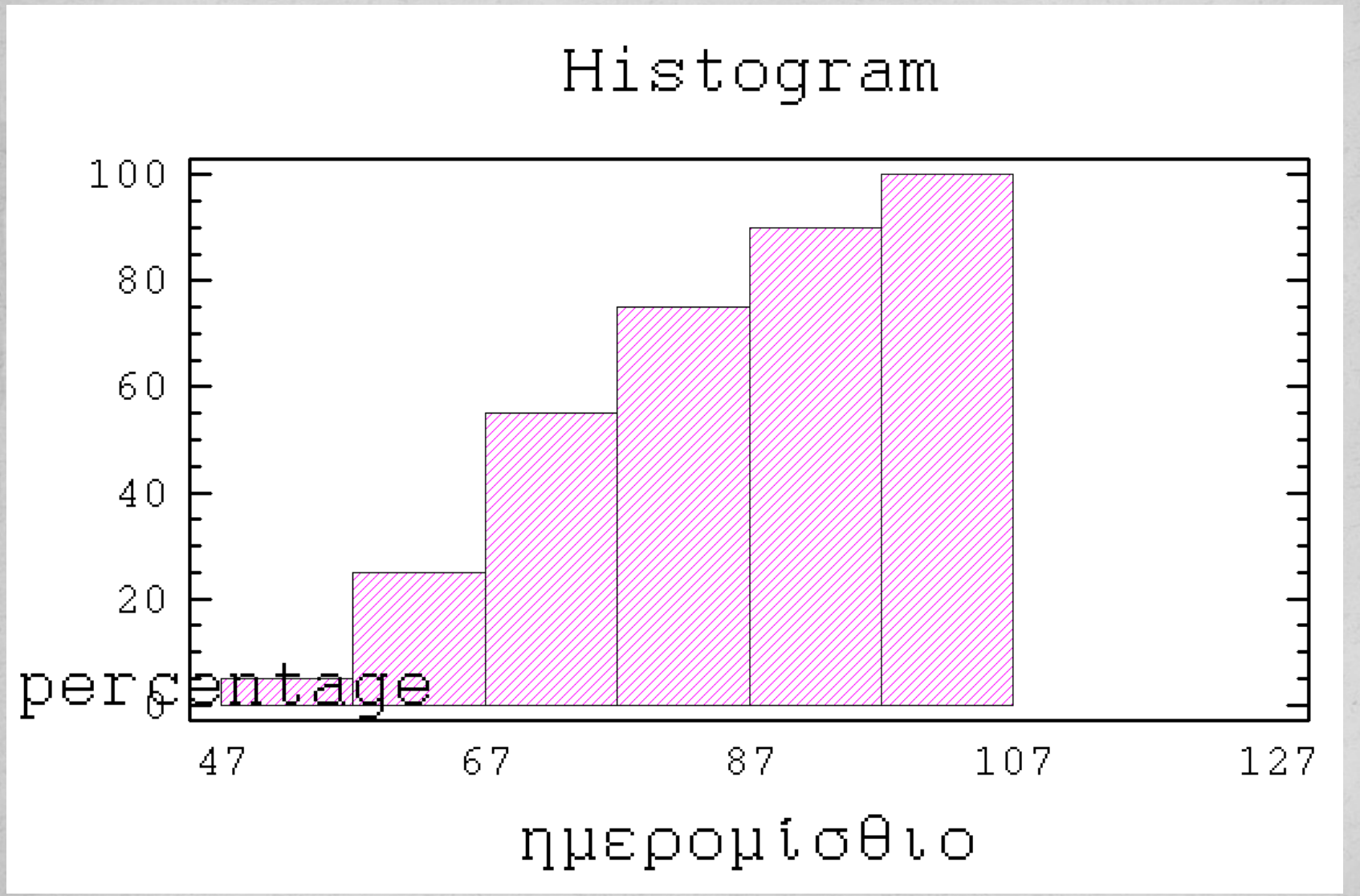


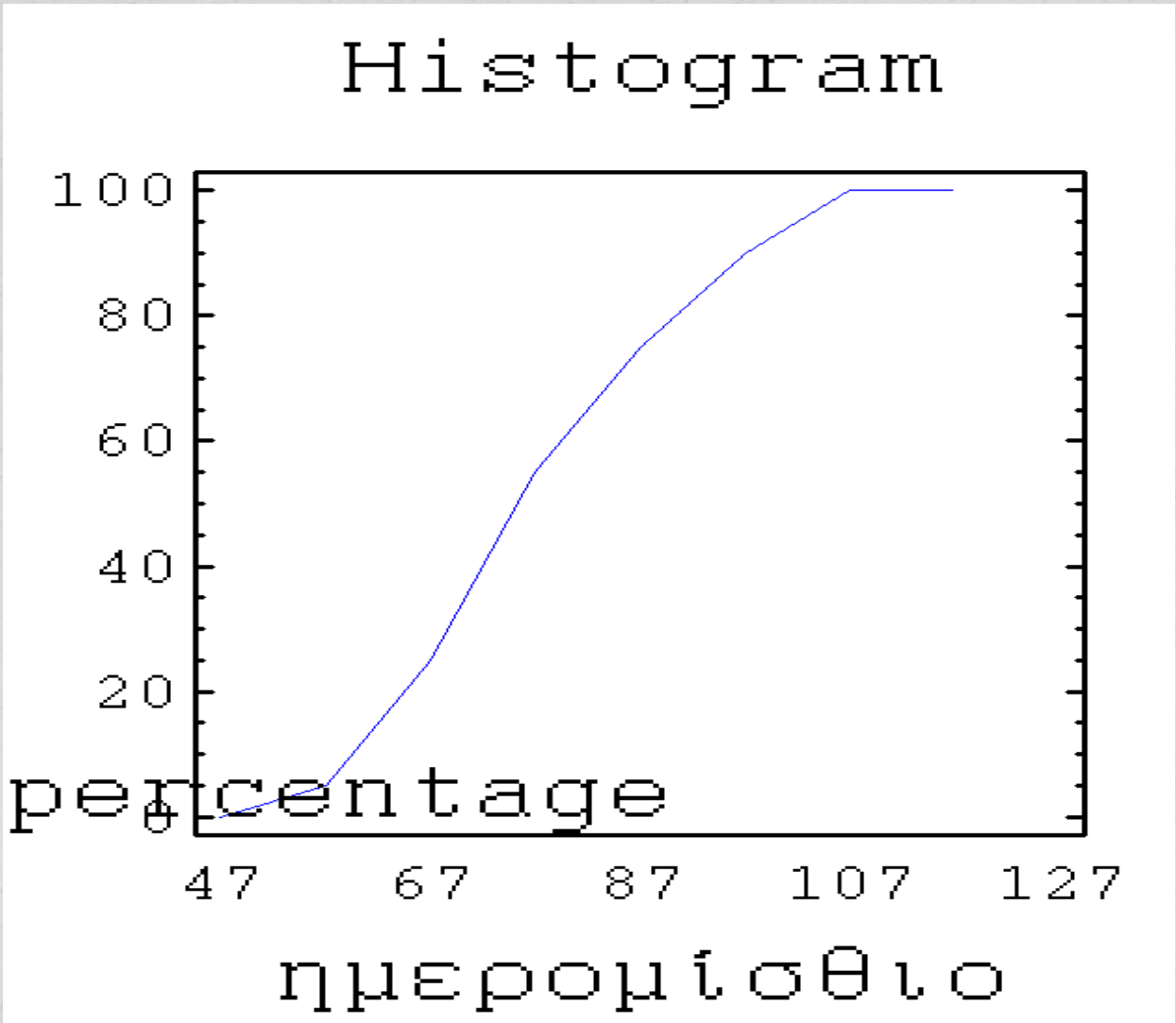

```
Summary Statistics for ημερομίσθιο

Count = 20
Average = 75,75
Variance = 184,934
Standard deviation = 13,5991
Minimum = 50,0
Maximum = 100,0
Range = 50,0
Stnd. skewness = 0,284892
Stnd. kurtosis = -0,41834


The StatAdvisor
---------------
   This table shows summary statistics for ημερομίσθιο.  It includes
measures of central tendency, measures of variability, and measures of
shape.  Of particular interest here are the standardized skewness and
```

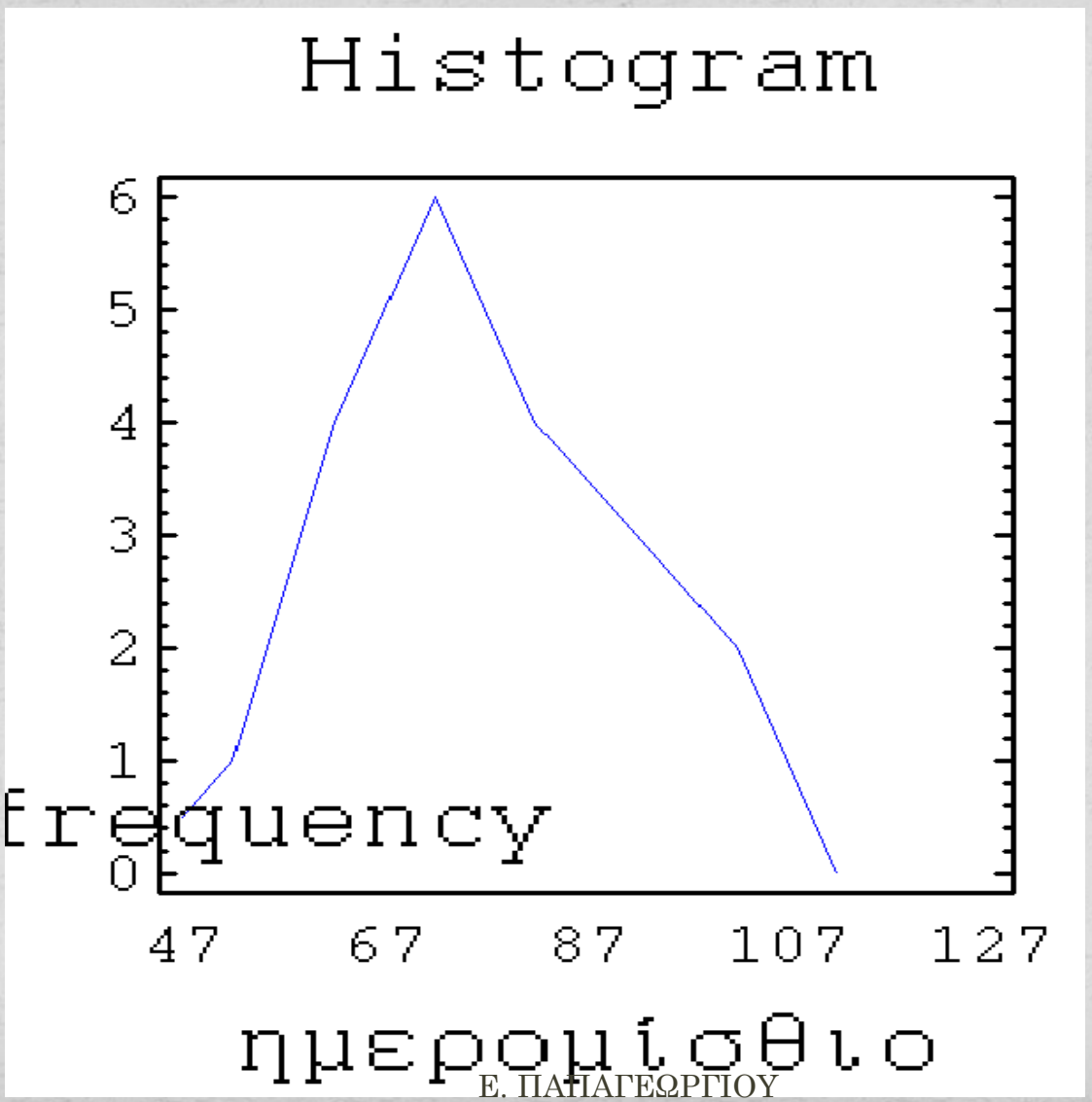



Histogram
frequency
0
1
2
3
4
5
6
47
67
87
107
127
ημερομίσθιο

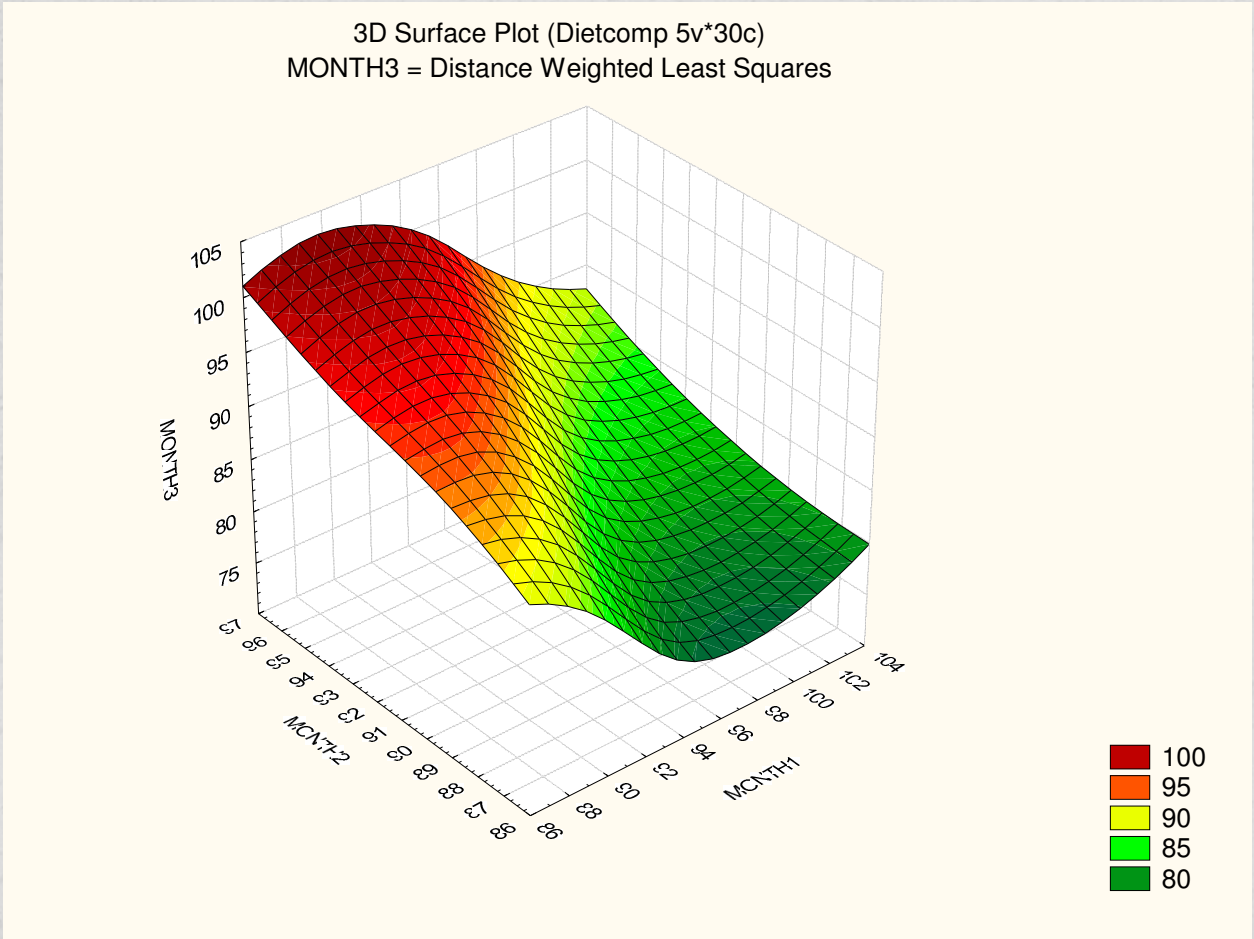

Histogram
20
16
12
8
4
0
frequency
47
67
87
107
127
ημερομίσθιο


Histogram
100
80
60
40
20
0
percentage
47
67
87
107
127
ημερομίσθιο


Histogram
100
80
60
40
20
0
percentage
47
67
87
107
127
ημερομίσθιο


Histogram
6
5
4
3
2
1
0
frequency
47
67
87
107
127
ημερομίσθιο

Διαγράμματα επιφανείας



1. Περιγραφική Στατιστική

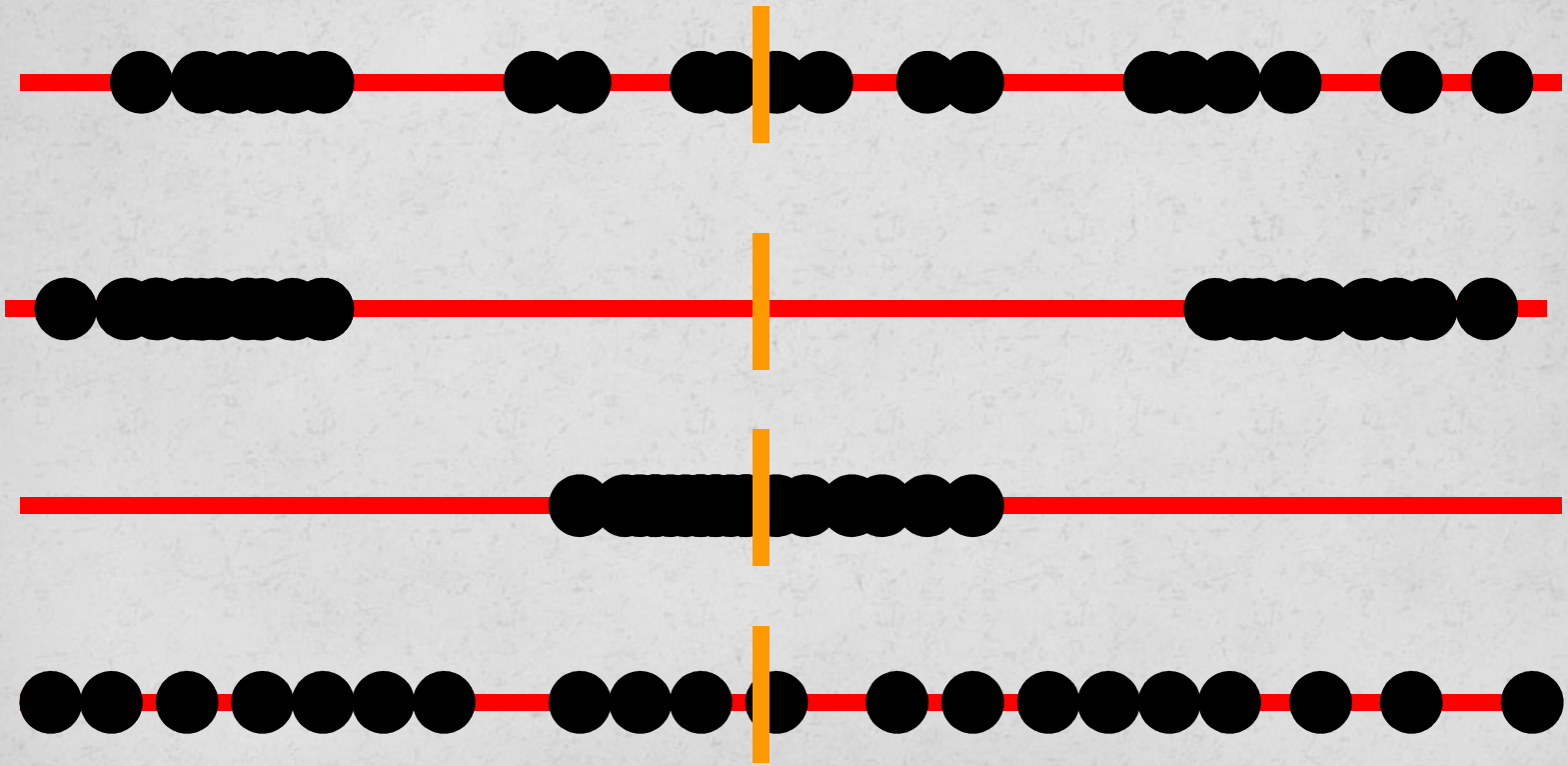
Ως περιγραφικά μέτρα θέσης εννοούμε τα εξής:

- Μέση Τιμή
- Κορυφή ή Επικρατούσα Τιμή
- Διάμεσο
- Ποσοστημόρια

Η ερμηνεία των περιγραφικών μέτρων

- **Αριθμητικός μέσος**
  - Η αναμενόμενη τιμή που θα έχει η ποσοτική μεταβλητή σε ένα τυχαία επιλεγμένο άτομο του δείγματος
    - Πόσο αξιόπιστο μέτρο είναι; (όταν στο δείγμα υπάρχει ανισοκατανομή)

Αριθμητικός μέσος

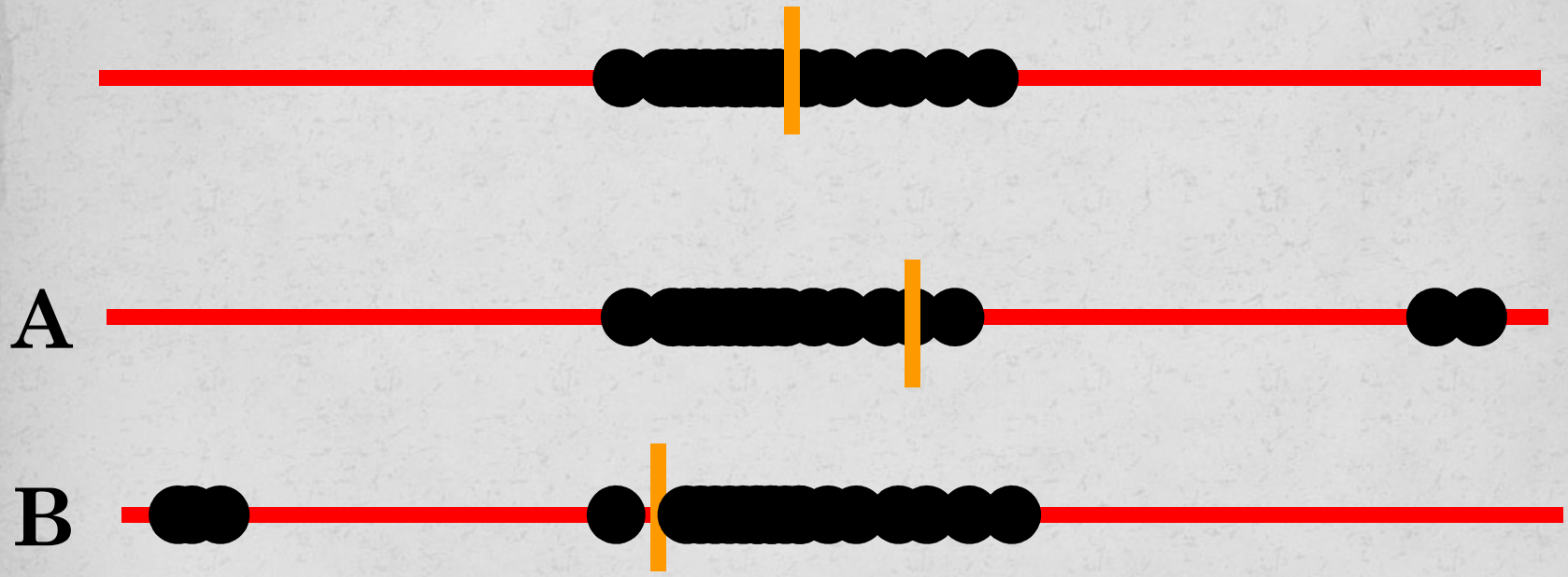


Αριθμητικός μέσος

Η ερμηνεία των περιγραφικών μέτρων

- **Πότε χρειαζόμαστε τη διάμεσο;**
  - **Όταν έχουμε ακραίες τιμές στην κατανομή της ποσοτικής μεταβλητής,**
    - **και ειδικότερα όταν είναι ασύμμετρα κατανεμημένες.**

Αριθμητικός μέσος & ακραίες τιμές



1. Περιγραφική Στατιστική

Ως πιο διαδεδομένα περιγραφικά μέτρα διασποράς εννοούμε τα εξής:

- Εύρος
- Ενδοτεταρτημοριακή απόκλιση
- Μέση απόκλιση
- Διασπορά ή Διακύμανση
- Τυπική απόκλιση

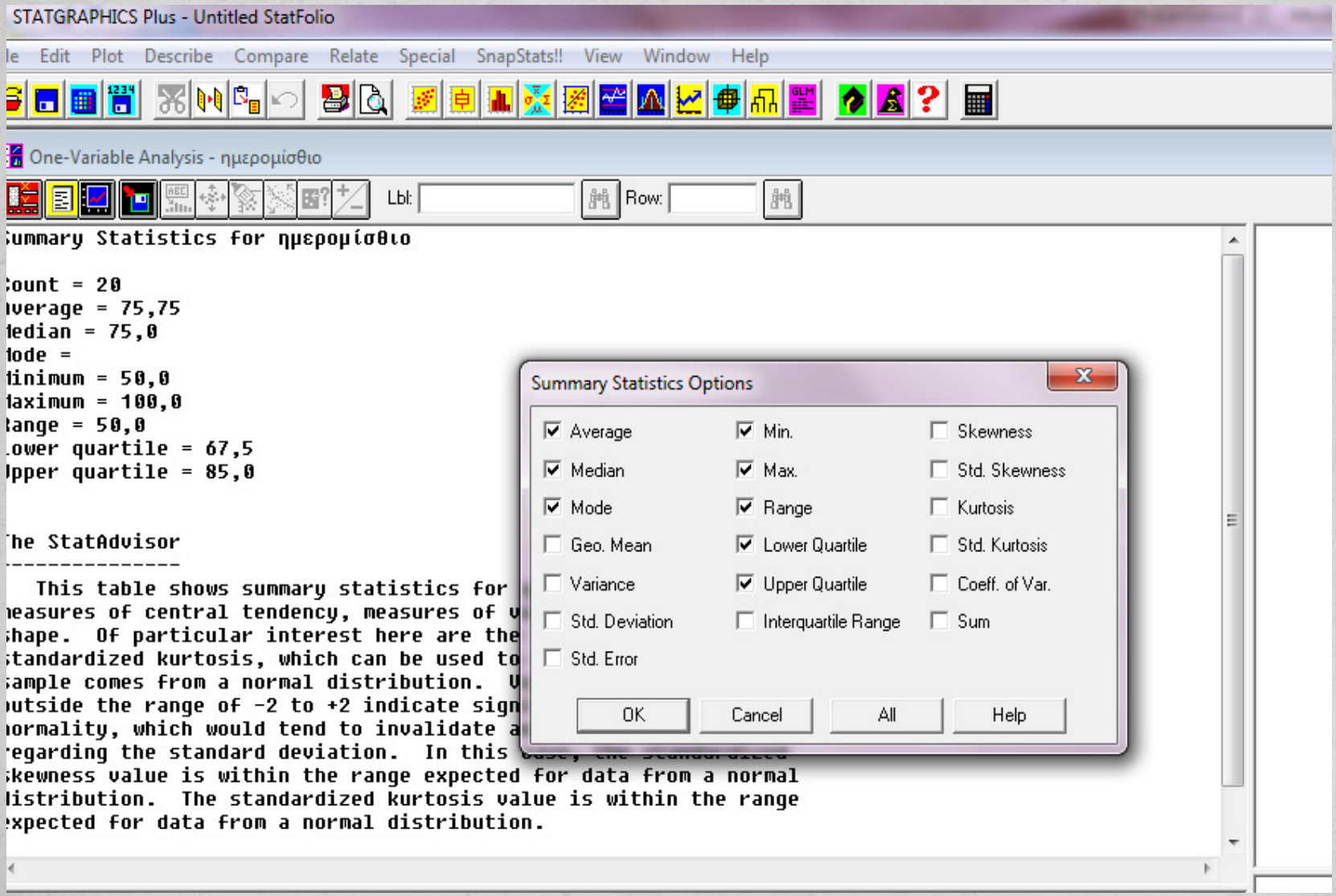
Η ερμηνεία των περιγραφικών μέτρων

- **Τυπική απόκλιση**
  - Ένας δείκτης μεταβλητότητας των τιμών της ποσοτικής μεταβλητής.
    - Όσο μικρότερες τιμές λαμβάνει, τόσο πιο ομοιογενές το δείγμα.
    - **Επηρεάζεται από τις μονάδες μέτρησης.**

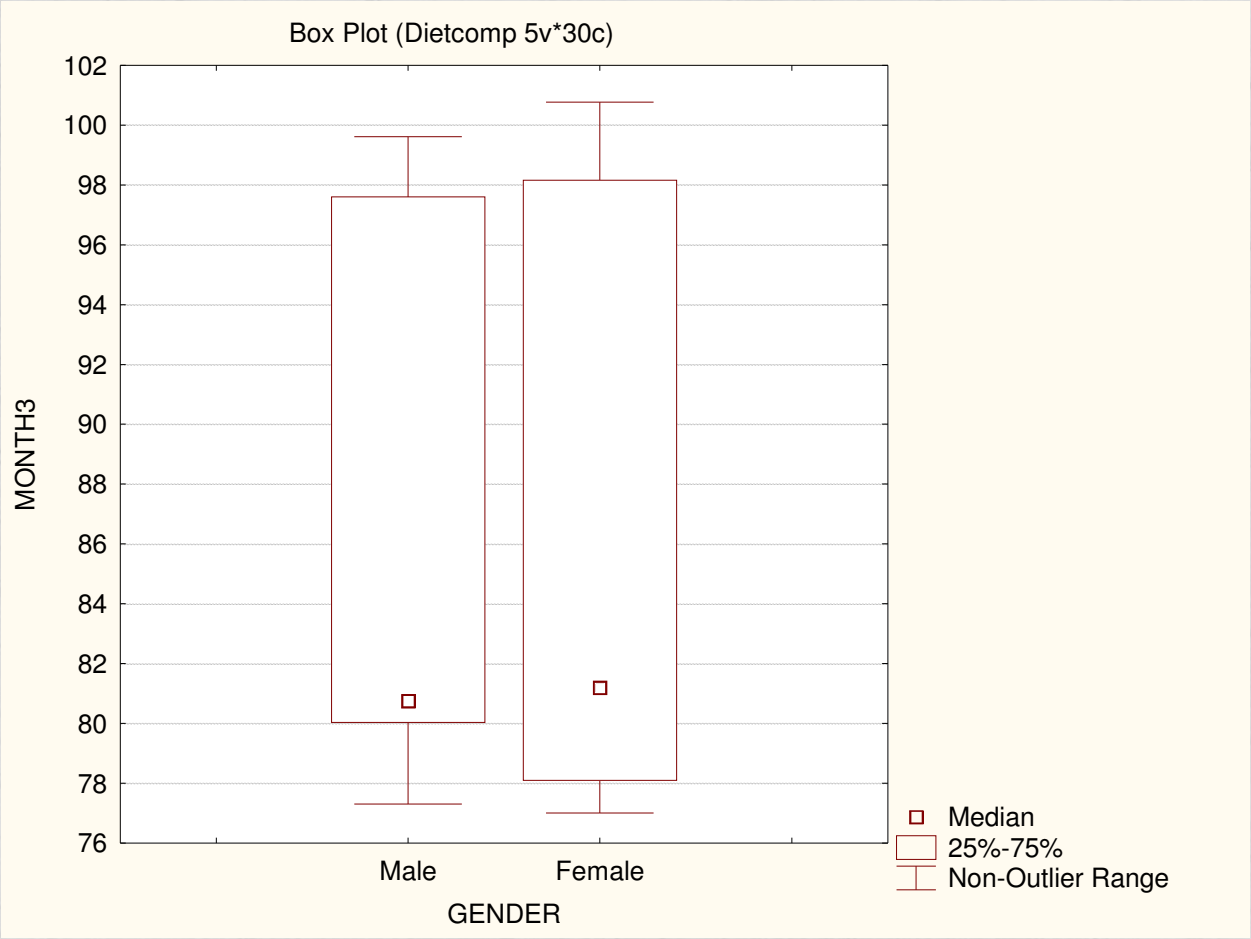
Μέτρηση της μεταβλητότητας

- **Συντελεστής μεταβλητότητας**
  - Ένας δείκτης μεταβλητότητας των τιμών της ποσοτικής μεταβλητής, που λαμβάνει υπόψη την μέση τιμή και δεν επηρεάζεται από τις μονάδες μέτρησης.

1. Περιγραφική Στατιστική



1. Περιγραφική Στατιστική

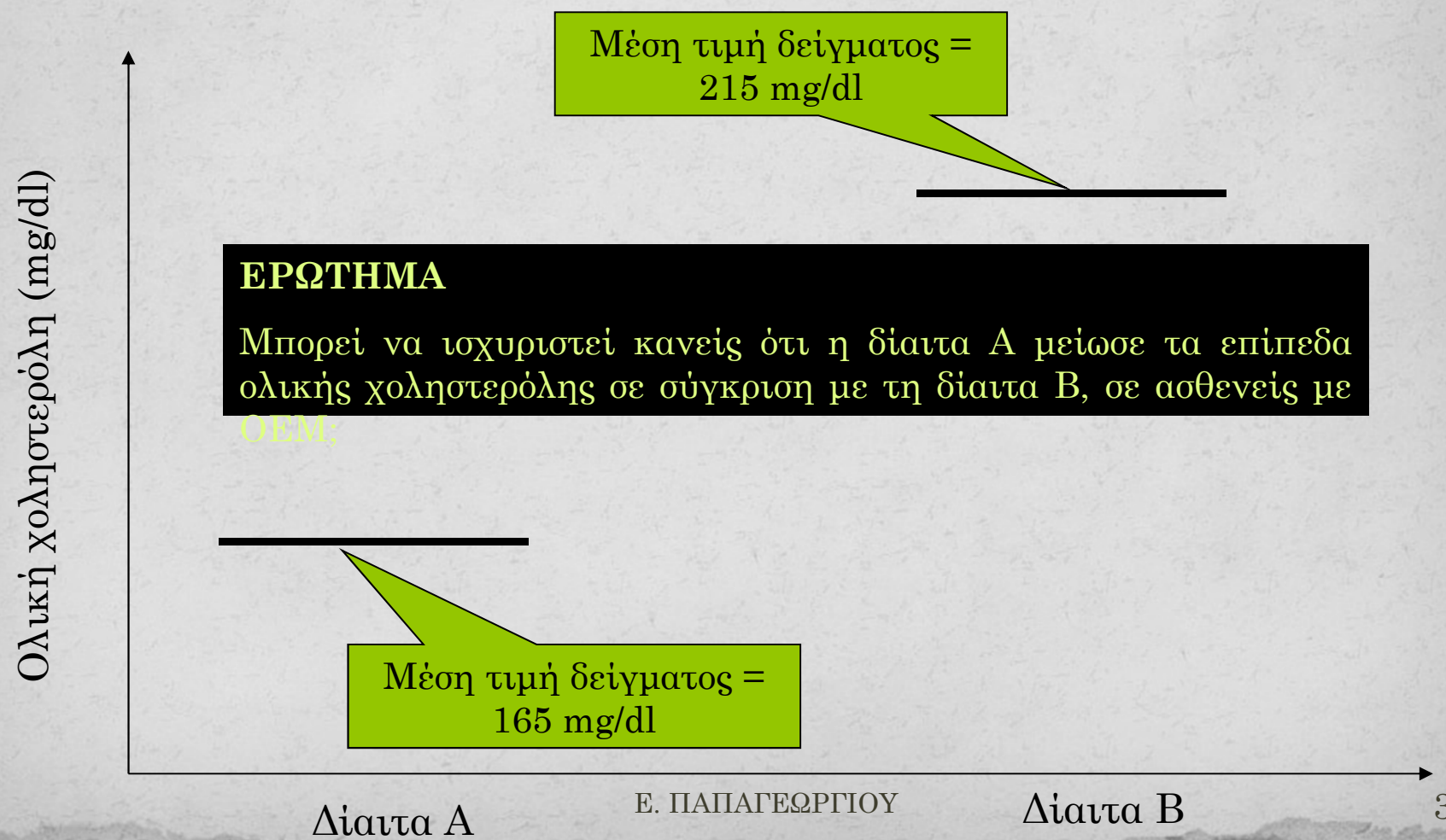


Α. ΠΑΡΑΜΕΤΡΙΚΕΣ ΔΟΚΙΜΑΣΙΕΣ

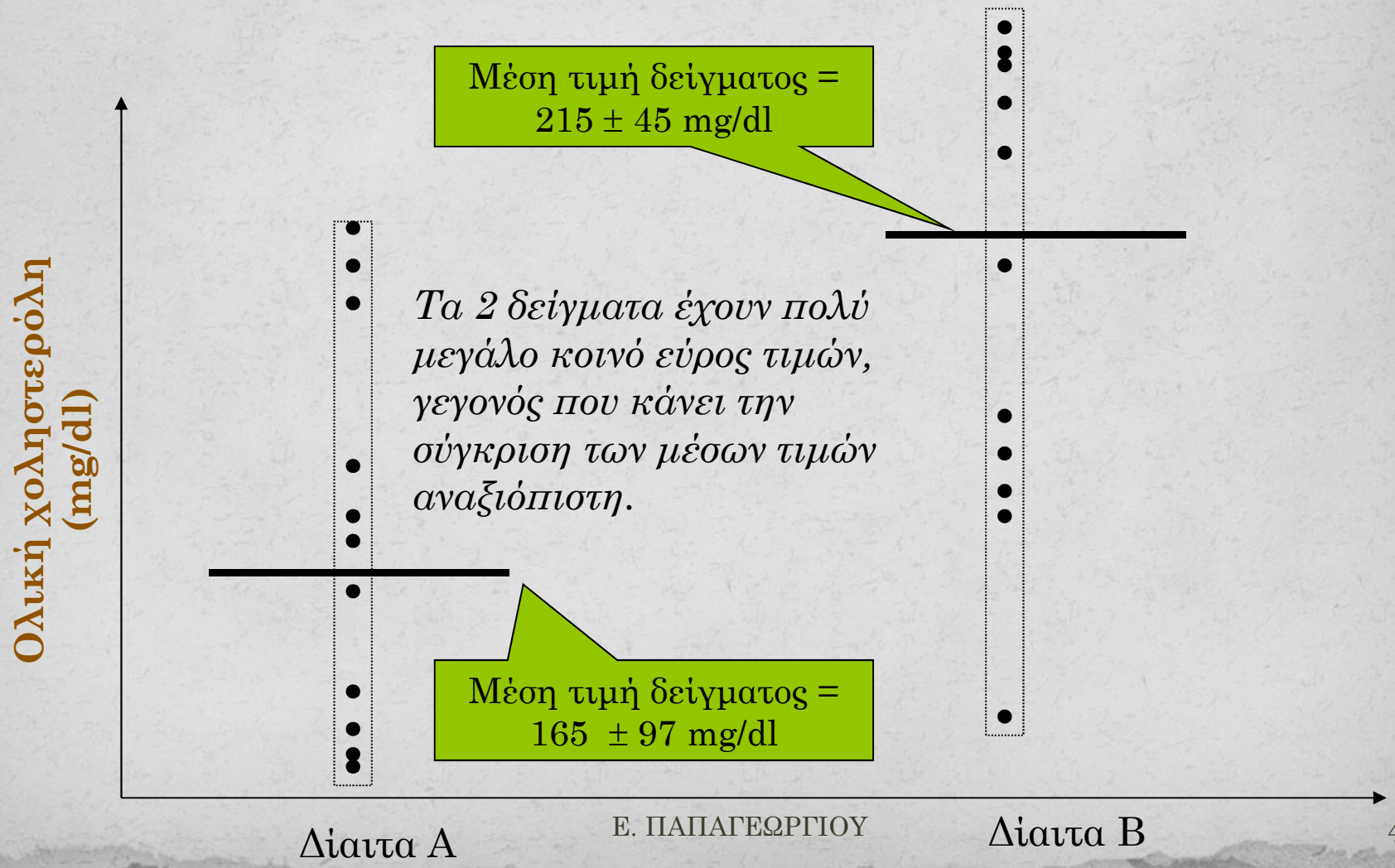
Οι ερευνητικές υποθέσεις

- Στην έρευνα ελέγχουμε υποθέσεις, με βάση τα πραγματικά δεδομένα μας.
  - π.χ. ο μεγάλος χρόνος εισόδου στο Νοσοκομείο από την έναρξη των συμπτωμάτων, συσχετίζεται με αυξημένο κίνδυνο θανάτου;
  - Μια διατροφή πλούσια σε υδατάνθρακες συσχετίζεται με μειωμένο σωματικό βάρος;

Γιατί είναι απαραίτητες οι στατιστικές «συγκρίσεις»;



Γιατί είναι απαραίτητες οι στατιστικές «συγκρίσεις»;

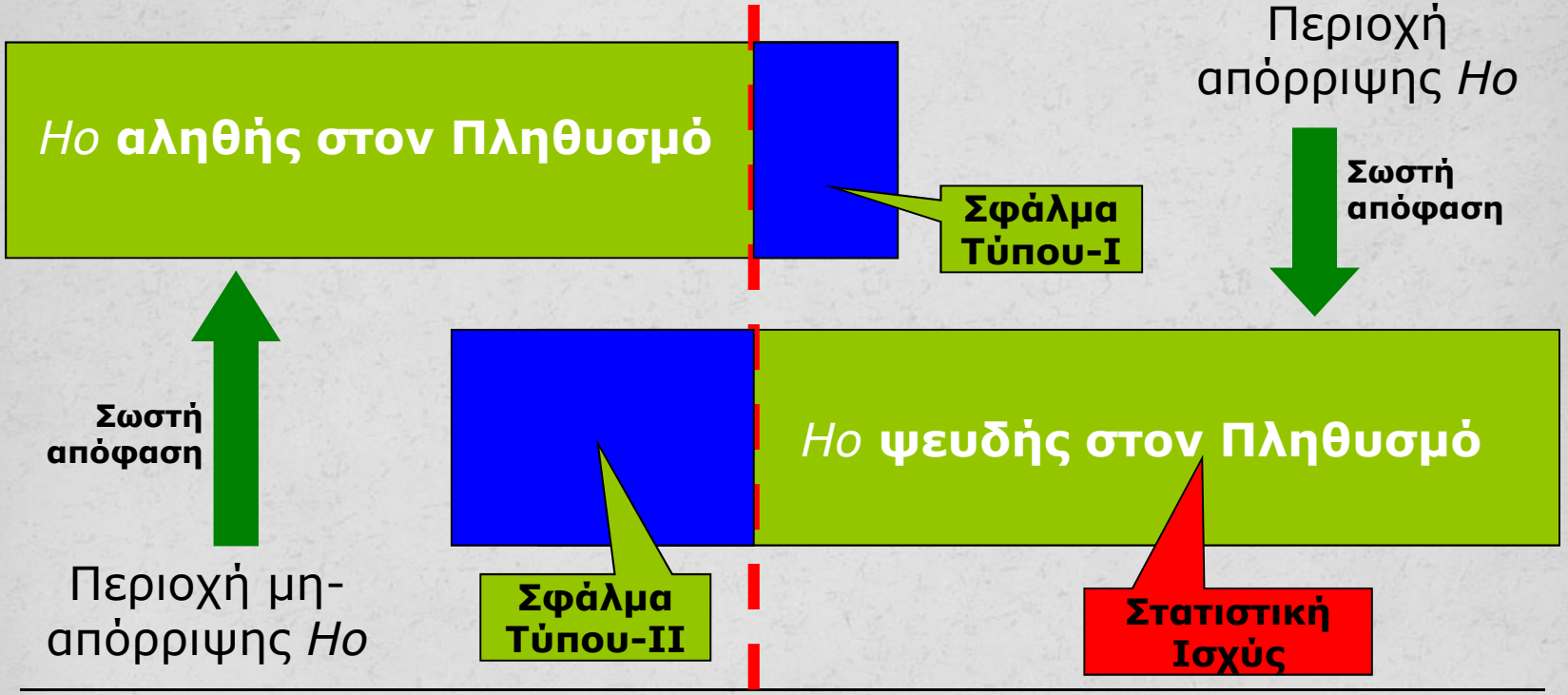


- Η διαδικασία που ακολουθείται για την λήψη τέτοιου είδους αποφάσεων ονομάζεται **έλεγχος υποθέσεων**
- Η υπόθεση που θέλουμε να ελέγξουμε συμβολίζεται με Ηο και ονομάζεται μηδενική υπόθεση ενώ η εναλλακτική της υπόθεση συμβολίζεται με H_1
- Σε κάθε έλεγχο είναι δυνατόν να πραγματοποιηθούν δύο ειδών σφάλματα:
- **Σφάλμα τύπου Ι:** Απόρριψη της Ηο ενώ στην πραγματικότητα είναι αληθής
- **Σφάλμα τύπου ΙΙ:** Απόρριψη της H_1 (Αποδοχή της Ηο) ενώ στην πραγματικότητα η H_1 είναι αληθής

Σφάλματα στη λήψη απόφασης

	Αποδοχή υπόθεσης H_o από το δείγμα	Απόρριψη υπόθεσης H_o από το δείγμα
Υπόθεση Ho αληθής στον πληθυσμό	✓	Σφάλμα τύπου I
Υπόθεση Ho ψευδής στον πληθυσμό	Σφάλμα τύπου II	✓ Στατιστική ισχύς

Έλεγχοι Υποθέσεων



- **α=Ρ(σφάλμα τύπου Ι)=**Ρ(Απόρριψη της Ηο ενώ στην πραγματικότητα είναι αληθής)
- **β=Ρ(σφάλμα τύπου ΙΙ)=**Αποδοχή της Ηο ενώ στην πραγματικότητα η H_1 είναι αληθής)
- **Η πιθανότητα γ= 1-β ονομάζεται ισχύς του ελέγχου και εκφράζει το ποσοστό των «σωστών» απορρίψεων της Ηο.**

Το α ονομάζεται επίπεδο σημαντικότητας.

Ερμηνεία του α:

Εάν για παράδειγμα σε έναν έλεγχο επιλέξουμε επίπεδο σημαντικότητας α=0.05 και απορρίψουμε την υπόθεση, αυτό σημαίνει ότι σε 100 όμοιες περιπτώσεις, είναι δυνατό έχουμε κάνει λάθος και να απορρίψουμε την Ηο ενώ είναι αληθής, μόνο σε 5. Σε μια τέτοια περίπτωση λέμε ότι η υπόθεση απορρίπτεται σε επίπεδο σημαντικότητας 0.05

Κριτήριο για την αποδοχή ή όχι της Ho είναι το p-value.

Το μικρότερο επίπεδο σημαντικότητας για το οποίο απορρίπτεται η Ho ονομάζεται p-value.

Απορρίπτεται η Ho αν η τιμή του p-value είναι μικρή.

Συγκεκριμένα, απορρίπτεται η Ho αν η τιμή του p-value είναι μικρότερη του α για αυτό το επίπεδο σημαντικότητας.

Όσο μειώνεται το α τόσο δυσκολεύει η απόφαση της απόρριψης.

Τι δεν είναι το *p*-value

- Το *p*-value **δεν** είναι η πιθανότητα να επαληθευθεί η μηδενική υπόθεση
  - και αυτό γιατί οι υποθέσεις δεν εκφράζονται με πιθανότητες στην στατιστική.

Τι δεν είναι το *p*-value

- Το *p*-value **δεν** είναι η πιθανότητα να απορριφθεί λανθασμένα η μηδενική υπόθεση.
  - Το να απορριφθεί λανθασμένα η μηδενική υπόθεση είναι το σφάλμα Τύπου I.
    - Αυτό το σφάλμα είναι μια εκδοχή της καλούμενης «σφάλμα του εισαγγελέα» (“prosecutor's fallacy”) όπου κρίνει αθώο τον κατηγορούμενο ενώ έχει διαπράξει το έγκλημα.
      - Το σφάλμα Τύπου I είναι στενά συνυφασμένο με το *p*-value, αφού απορρίπτουμε τη μηδενική υπόθεση όταν το *p*-value είναι μικρότερο από κάποιο προκαθορισμένο όριο α (επίπεδο σημαντικότητας) του σφάλματος τύπου-I.

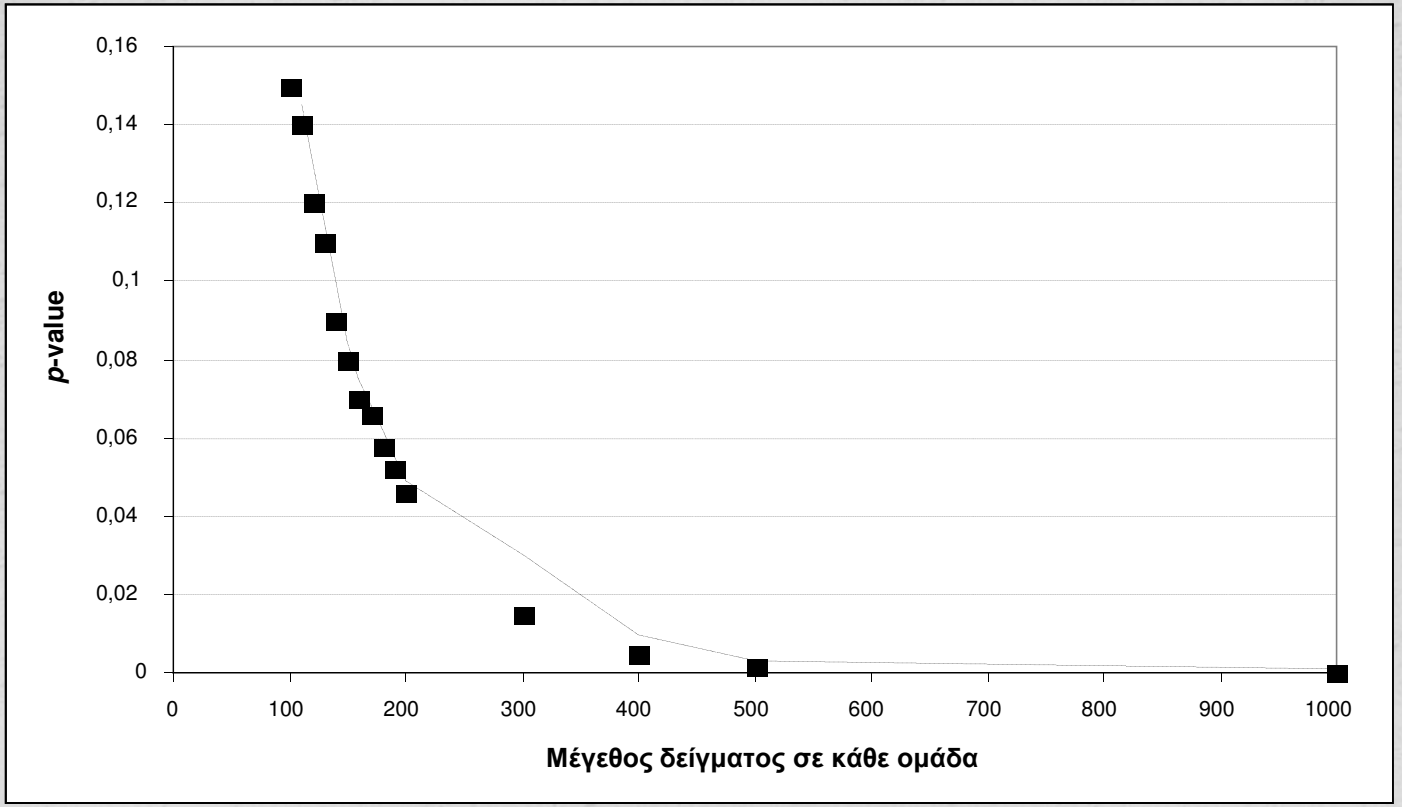
p-value και μέγεθος του δείγματος

- Το *p*-value επηρεάζεται ισχυρά από το μέγεθος του δείγματος.

Συγκεκριμένα

- Υπάρχει **αντίστροφη συσχέτιση** μεταξύ του μεγέθους δείγματος και του *p*-value.

p-value και μέγεθος του δείγματος για μια δεδομένη συσχέτιση



Το μέγεθος του δείγματος

- Το **επαρκές** μέγεθος του δείγματος είναι μεγίστης σημασίας για την αξιοπιστία της έρευνας.

Οι «αρχές» της δειγματοληψίας

- Πρέπει όμως να ληφθεί υπόψη ότι **σχετικά μεγάλο δείγμα** συνεπάγεται και μεγάλο κόστος
  - **χωρίς αυτό να σημαίνει και απαραίτητα αξιόπιστα αποτελέσματα,**
- ενώ **πολύ μικρό δείγμα** μπορεί να οδηγήσει σε **συστηματικό σφάλμα** και **μεροληπτικές** αποφάσεις για τον πληθυσμό.

Το μέγεθος του δείγματος καθορίζεται από:

- το επίπεδο στατιστικής σημαντικότητας των ελέγχων, το οποίο συμβολίζεται με α και στο χώρο των επιστημών έχει καθοριστεί να είναι < 0,01 ή < 0,05
- το μέγεθος της αναζητούμενης σχέσης, π.χ. πόσο μεγάλη θα πρέπει να είναι η διαφορά στα επίπεδα ολικής χοληστερόλης μεταξύ της θεραπευτικής προσέγγισης Α και της θεραπευτικής προσέγγισης Β έτσι ώστε να θεωρείται κλινικά αξιόλογη
- τη στατιστική ισχύ των ελέγχων, η οποία συμβολίζεται με γ και στο χώρο των επιστημών της Υγείας έχει καθοριστεί να είναι > 0,80 ή > 0,90
- το επίπεδο ακρίβειας στις μετρήσεις, το οποίο εξαρτάται και από την συνείδηση των ερευνητών που διεξάγουν την έρευνα
- το μέγεθος του πληθυσμού αναφοράς
- τη μεταβλητότητα στα χαρακτηριστικά του πληθυσμού, η οποία αν είναι μεγάλη συνεπάγεται και ανάλογη αύξηση του μεγέθους του δείγματος
- το διαθέσιμο χρηματικό ποσό για την έρευνα

Τα **διαστήματα εμπιστοσύνης** αποτελούν έναν εναλλακτικό τρόπο εκτίμησης παραμέτρων.
Εκτιμάμε μία παράμετρο, με ένα διάστημα που έχει άκρα τυχαίες μεταβλητές.
Το διάστημα θα έχει την μορφή:

$$P[L \leq \theta \leq U] = \gamma$$

Ένα τέτοιο διάστημα ονομάζεται διάστημα εμπιστοσύνης με βαθμό εμπιστοσύνης γ. Ο αριθμός γ=1-α εκφράζει την ακρίβεια με την οποία θέλουμε να γίνει η εκτίμηση, ενώ ο α εκφράζει τον βαθμό ανεκτικότητας ώστε το διάστημα να μην περιέχει την πραγματική τιμή της παραμέτρου.
Για παράδειγμα αν γ=0.95 αναμένεται σε 100 δείγματα της μορφής [L,U] τα 95 να περιλαμβάνουν την σωστή τιμή.

- **Παράδειγμα**

Μετρήθηκε το κάλιο του ορού σε 9 υγιή άτομα και σε 4 άτομα που έπασχαν από μία νόσο. Στα υγιή άτομα βρέθηκε μέση τιμή 4 m Eq/L και σταθερή απόκλιση 0.9 m Eq/L, ενώ στους ασθενείς βρέθηκε μέση τιμή 5 m Eq/L και σταθερή απόκλιση 0.8 m Eq/L.

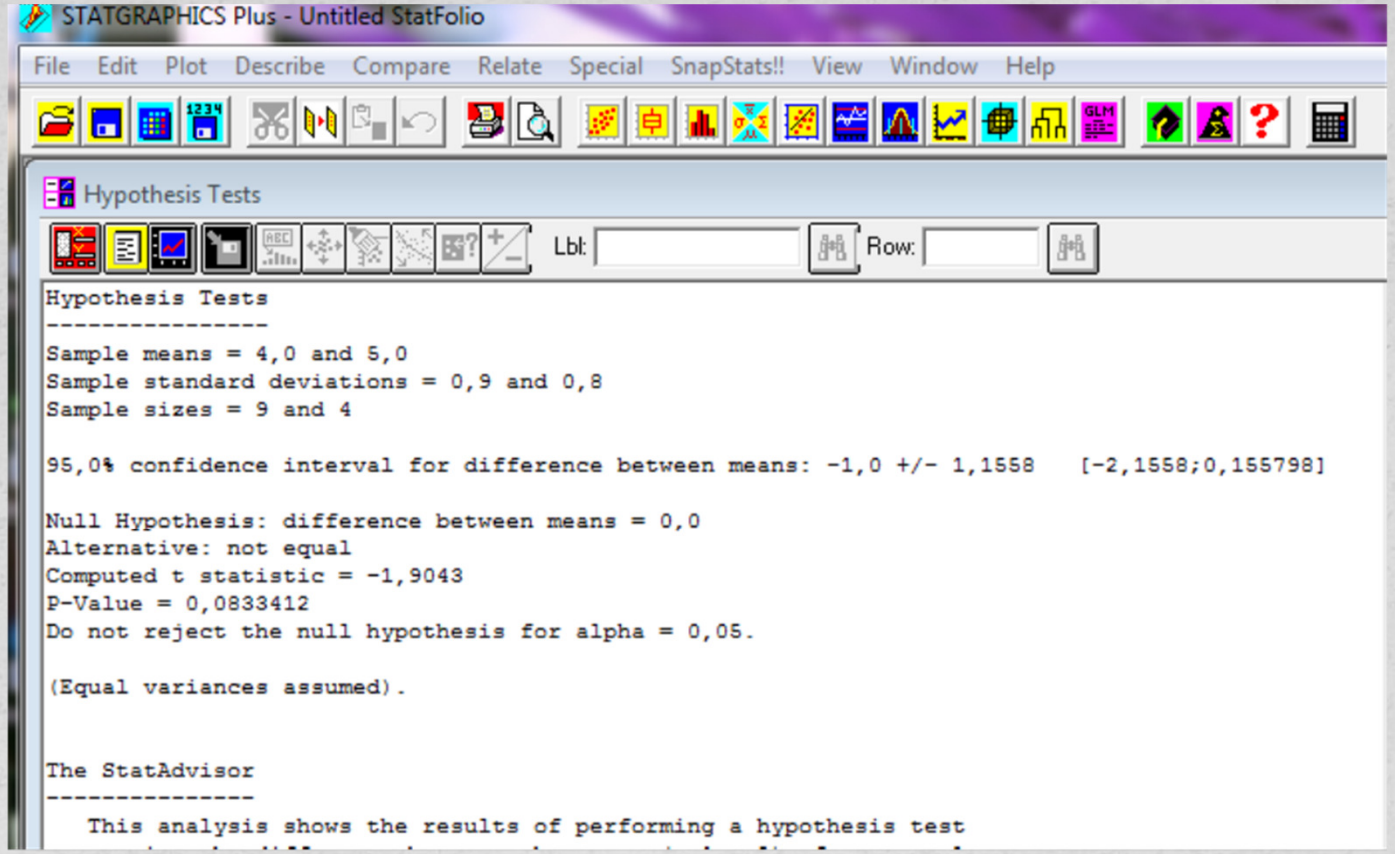
Υπάρχει διαφορά των μέσων τιμών του καλίου του ορού στις δύο αυτές ομάδες;

Έλεγχοι υποθέσεων και δ.ε. για διαφορά μέσων τιμών σε ανεξάρτητους πληθυσμούς σε μικρά δείγματα και με ισότητα διασπορών ($\sigma_1 = \sigma_2 = \sigma$):

H_0: $\mu_1=\mu_2$ H_1: $\mu_1>\mu_2$	H_0: $\mu_1=\mu_2$ H_1: $\mu_1<\mu_2$	H_0: $\mu_1=\mu_2$ H_1: $\mu_1\neq\mu_2$
$R=\{t > t_{n_1+n_2-2;a}\}$	$R=\{t < -t_{n_1+n_2-2;a}\}$	$R=\{\lvert t\rvert > t_{n_1+n_2-2;\frac{a}{2}}\}$

$(\overline{x}_1-\overline{x}_2-s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\ t_{n_1+n_2-2;\frac{a}{2}}\ ,\ \overline{x}_1-\overline{x}_2+s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\ t_{n_1+n_2-2;\frac{a}{2}})$, όπου $s^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$.

Το κριτήριο t δίνεται από τον τύπο: $t=\frac{\overline{x}_1-\overline{x}_2}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$


STATGRAPHICS Plus - Untitled StatFolio
File Edit Plot Describe Compare Relate Special SnapStats!! View Window Help
Hypothesis Tests
Lbl:
Row:
Hypothesis Tests
Sample means = 4,0 and 5,0
Sample standard deviations = 0,9 and 0,8
Sample sizes = 9 and 4
95,0% confidence interval for difference between means: -1,0 +/- 1,1558 [-2,1558;0,155798]
Null Hypothesis: difference between means = 0,0
Alternative: not equal
Computed t statistic = -1,9043
P-Value = 0,0833412
Do not reject the null hypothesis for alpha = 0,05.
(Equal variances assumed).
The StatAdvisor
This analysis shows the results of performing a hypothesis test

Όπως διαπιστώνουμε δεχόμαστε την μηδενική υπόθεση Ho : μ1=μ2 έναντι της εναλλακτικής H_1 : μ1≠μ2, δηλαδή δεχόμαστε ότι δεν υπάρχει διαφορά στις τιμές του καλίου του ορού στις δύο αυτές ομάδες.

Συγκεκριμένα:

Null Hypothesis: difference between means = 0,0

Alternative: not equal

Computed t statistic = -1,9043

P-Value = 0,0833412

Do not reject the null hypothesis for alpha = 0,05.

(Equal variances assumed).

Δεχόμαστε την μηδενική υπόθεση Ho για επίπεδο σημαντικότητας α=0.05, διότι η τιμή του p-value είναι 0.08334 > 0.05. Επίσης το στατιστικό λογισμικό μας υπολογίζει και την τιμή του t κριτηρίου ίση με -1.9043.

Σημειώνεται ότι αναφερόμαστε σε κανονικούς πληθυσμούς με άγνωστες και ίσες διασπορές (σ1=σ2=σ).

Όπως διαπιστώνουμε επίσης το 95% διάστημα εμπιστοσύνης για την διαφορά των μέσων τιμών μ1-μ2 του καλίου του ορού στις δύο αυτές ομάδες είναι:

[-2,1558;0,155798].

Παράδειγμα:

Σε τέσσερα άτομα με αυξημένες τιμές των τριγλυκεριδίων του ορού (mg/dl) χορηγήθηκε για ένα μήνα φάρμακο που πιστεύεται ότι ελαττώνει τα επίπεδα των τριγλυκεριδίων. Οι τιμές των τριγλυκεριδίων στα τέσσερα αυτά άτομα πριν και μετά τη χορήγηση του φαρμάκου ήταν:

Άτομο	Πριν τη χορήγηση	Μετά τη χορήγηση
1ο	180	120
2ο	200	220
3ο	240	130
4ο	230	160

Βρείτε ένα 95% δ.ε. για την διαφορά των μέσων μ1-μ2 στα επίπεδα των τριγλυκεριδίων πριν και μετά την χορήγηση.

Ελαττώνει τα επίπεδα των τριγλυκεριδίων το φάρμακο αυτό;

(Άσκηση 65 σελ. 16 του Βιβλίου Ασκήσεων Βιοστατιστικής Α. Τζώνου & Κ. Κατσουγιάννη)

Έλεγχοι υποθέσεων και δ.ε. για παρατηρήσεις κατά ζεύγη:

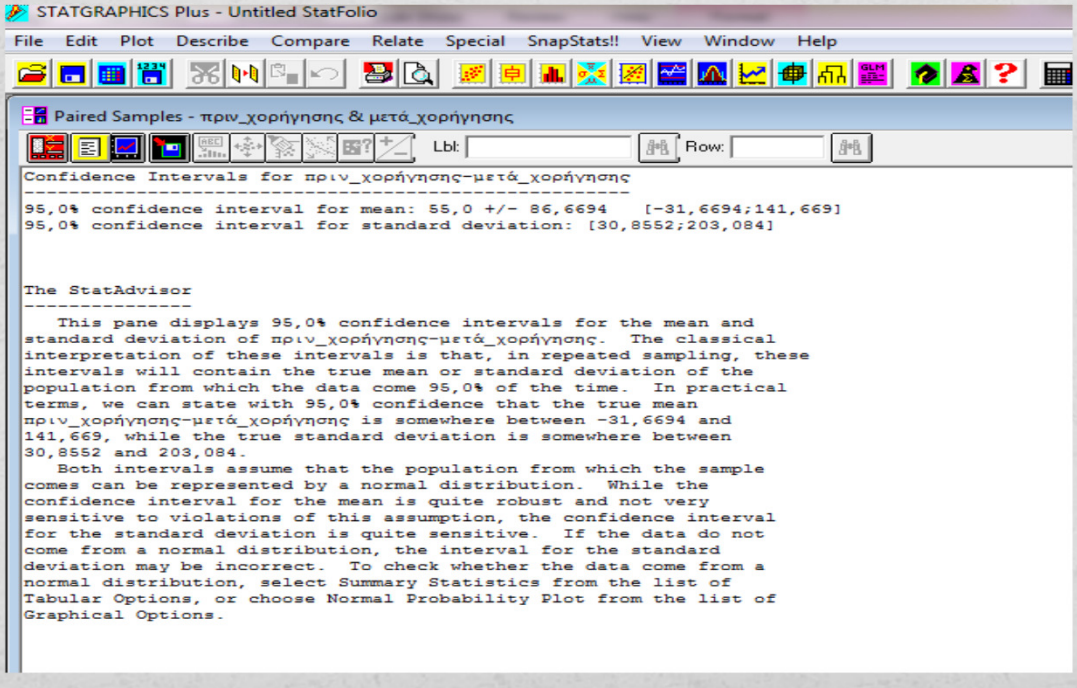
$H_0: \mu_1=\mu_2$ $H_1: \mu1>\mu2$	$H_0: \mu_1=\mu_2$ $H_1: \mu1<\mu2$	$H_0: \mu_1=\mu_2$ $H_1: \mu1\neq\mu2$
$R=\{\frac{\bar{z}}{s_z}\sqrt{n}>t_{n-1;a}\}$	$R=\{\frac{\bar{z}}{s_z}\sqrt{n}<-t_{n-1;a}\}$	$R=\{\left\|\frac{\bar{z}}{s_z}\sqrt{n}\right\|>t_{n-1;\frac{a}{2}}\}$

$(\bar{z}-\frac{s_z}{\sqrt{n}}t_{n-1;\frac{a}{2}},\bar{z}+\frac{s_z}{\sqrt{n}}t_{n-1;\frac{a}{2}})$, όπου $z=x_i-y_i$.

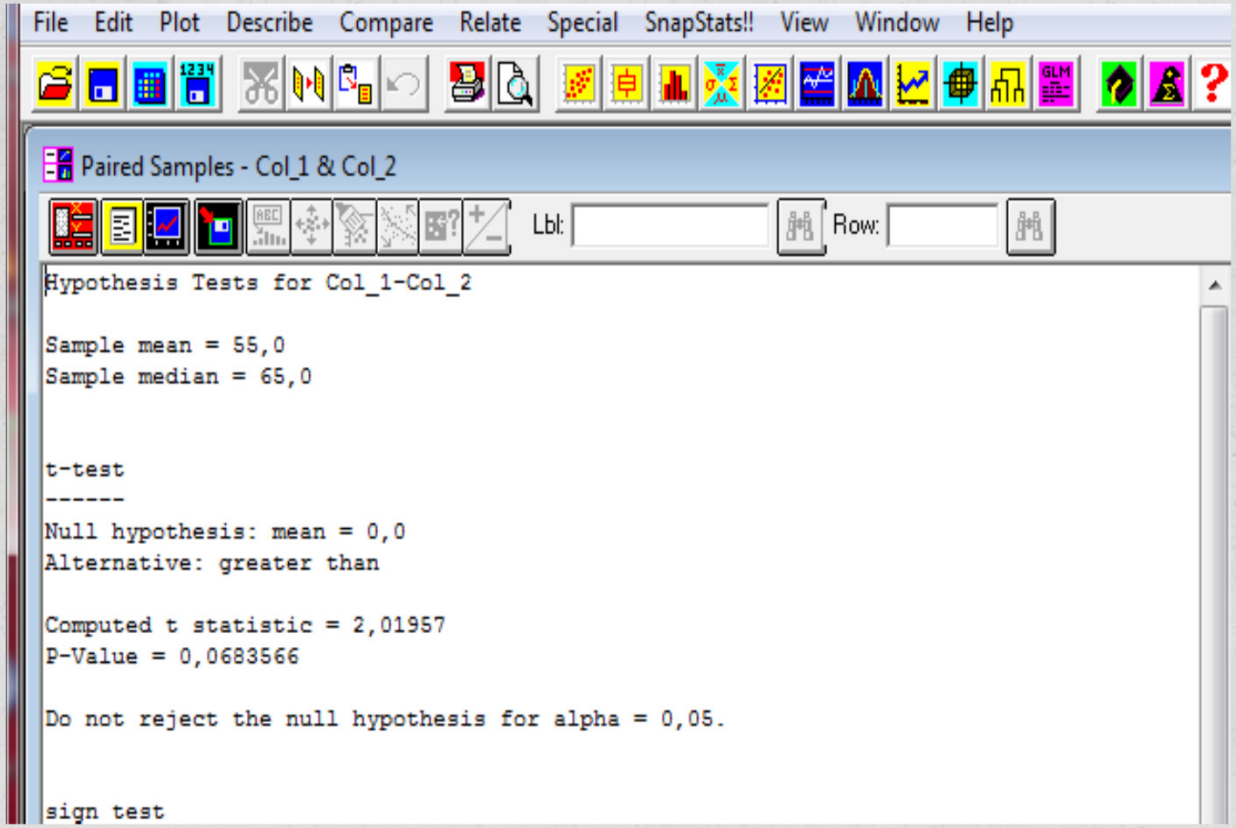
2. Statistical Tests –Confidence Intervals
2.1 Statistical tests I

Όπως φαίνεται και στην παρακάτω εικόνα το 95% διάστημα εμπιστοσύνης για την διαφορά των μέσων στα επίπεδα τριγλικεριδίων πριν και μετά την χορήγηση είναι:

55,0 +/- 86,6694 = [-31,6694;141,669].



2. Statistical Tests –Confidence Intervals
2.1 Statistical tests I



Όπως παρατηρούμε παράγονται τα εξής συμπεράσματα:

Null hypothesis: mean = 0,0

Alternative: greater than

Computed t statistic = 2,01957

P-Value = 0,0683566

Do not reject the null hypothesis for alpha = 0,05.

Δηλ. δεχόμαστε (δεν απορρίπτουμε) την μηδενική υπόθεση σε επίπεδο σημαντικότητας α=5% και συνεπώς το φάρμακο δεν ελαττώνει τα επίπεδα των τριγλυκεριδίων.

Αυτό συμβαίνει διότι η τιμή του P είναι 0,068>0,05 και άρα δέχομαι την Ho: μ1=μ2.

Ταυτόχρονα υπολογίζεται και η τιμή του κριτηρίου t statistic ίση με 2,01957.

2. Απλή Παλινδρόμηση

Η παρούσα παράγραφος αφορά στην απλή παλινδρόμηση.

Στην απλή παλινδρόμηση απαιτούνται δύο ποσοτικές μεταβλητές εκ των οποίων η μία θεωρείται ανεξάρτητη- independent (X) και η άλλη εξαρτημένη - dependent (Y). Η διερεύνηση της μορφής της παλινδρόμησης είναι το βασικό πρόβλημα το οποίο κατ' αρχάς θα πρέπει να επιλυθεί. Είναι δηλαδή απαραίτητο να προσδιορίσουμε αν τα ζεύγη τιμών (X, Y) προσαρμόζονται καλύτερα σε μια ευθεία ή παραβολή ή έλλειψη ή υπερβολή κ.λ.π.

Αν υποθέσουμε ότι η κατάλληλη μορφή παλινδρόμησης, για κάποια συγκεκριμένα ζεύγη τιμών, είναι η γραμμική, τότε για να υπολογίσουμε τους συντελεστές της παλινδρόμησης και τα διάφορα στατιστικά μέτρα τα οποία είναι απαραίτητα, η διαδικασία την οποία πρέπει να ακολουθήσουμε αποτελεί αντικείμενο του συγκεκριμένης παραγράφου.

Παρατίθεται και λυμένο παράδειγμα απλής παλινδρόμησης στο οποίο περιγράφεται η εν λόγω διαδικασία.

2. Απλή Παλινδρόμηση

Στην ανάλυση συνεχών δεδομένων (π.χ. ηλικία, χρόνος, επίδοση κ.α.) χρησιμοποιούμε μοντέλα απλής γραμμικής παλινδρόμησης (με μία μόνο επεξηγηματική μεταβλητή Χ), μοντέλα πολλαπλής γραμμικής παλινδρόμησης (δηλαδή με παραπάνω από μια επεξηγηματική μεταβλητή Χ στο μοντέλο), καθώς και με μοντέλα ανάλυσης διακύμανσης που εκφράζουν την επίδραση κάθε επιπέδου μιας ή περισσοτέρων επεξηγηματικών στην εξαρτημένη Υ. Για παράδειγμα, έστω ότι θέλουμε να ελέγξουμε την επίδραση που θα είχε κάποιο φάρμακο (φάρμακο Α, φάρμακο Β, φάρμακο Γ, δηλαδή 3 επίπεδα στην μεταβλητή Φάρμακο) σε ένα άνθρωπο ανάλογα με την ηλικία του. Εδώ η εξαρτημένη είναι η ηλικία του ανθρώπου (Υ) και η επεξηγηματική είναι του Φάρμακο (Χ).

Και στις τρεις αυτές αναλύσεις, επειδή ακριβώς στηρίζονται στο γραμμικό υπόδειγμα, για να είναι οι εκτιμήσεις συνεπείς, οπότε και σωστές θα πρέπει να πληρούνται κάποιες προϋποθέσεις.

- **Κανονικότητα:** τα κατάλοιπα θα πρέπει να ακολουθούν κανονική κατανομή με μέσο 0 και διακύμανση γνωστή.
- **Ομοσκεδαστικότητα:** Ισότητα διακυμάνσεων Στην περίπτωση απλής γραμμικής παλινδρόμησης κάνουμε ένα διάγραμμα σημείων (Scatter Plot) των καταλοίπων με την επεξηγηματική. Αν τα σημεία είναι τυχαία και δεν παρουσιάζουν κάποια τάση τότε υπάρχει ομοσκεδαστικότητα.
- **Ανεξαρτησία καταλοίπων:** Κάνουμε ένα διάγραμμα σημείων μεταξύ των προβλεπόμενων τιμών (Predicted values) και των καταλοίπων (Residuals). Αν είναι τυχαία τα σημεία τότε έχουμε ανεξαρτησία.
- **Γραμμικότητα:** Θα κάνουμε ένα διάγραμμα σημείων (Scatter Plot) προβλεπόμενων τιμών (Unstandardized Predicted Values) έναντι καταλοίπων (Standardized Residuals).

2. Απλή Παλινδρόμηση

- **Παράδειγμα:**

Δίνονται οι τιμές της απορρόφησης πρωτείνης (σε μήκος κύματος 280nm) ανάλογα με την πυκνότητα (συγκέντρωση) της πρωτείνης αυτής (gr/lt). Υπάρχει σχέση απορρόφησης και πυκνότητας πρωτείνης;

Απορρόφηση Πρωτείνης	Πυκνότητα Πρωτείνης
0,10	5
0,21	10
0,25	15
0,32	20
0,40	25
0,48	30
0,55	35
0,64	40
0,75	45
0,80	50

(Άσκηση 167 σελ. 48 του Βιβλίου Ασκήσεων Βιοστατιστικής Α. Τζώνου & Κ. Κατσουγιάννη)

Λύση:

Ακολουθούμε τα παρακάτω βήματα:

Εισάγουμε τα δεδομένα σε στήλες (με τον γνωστό τρόπο) όπως φαίνεται και στην παρακάτω εικόνα:

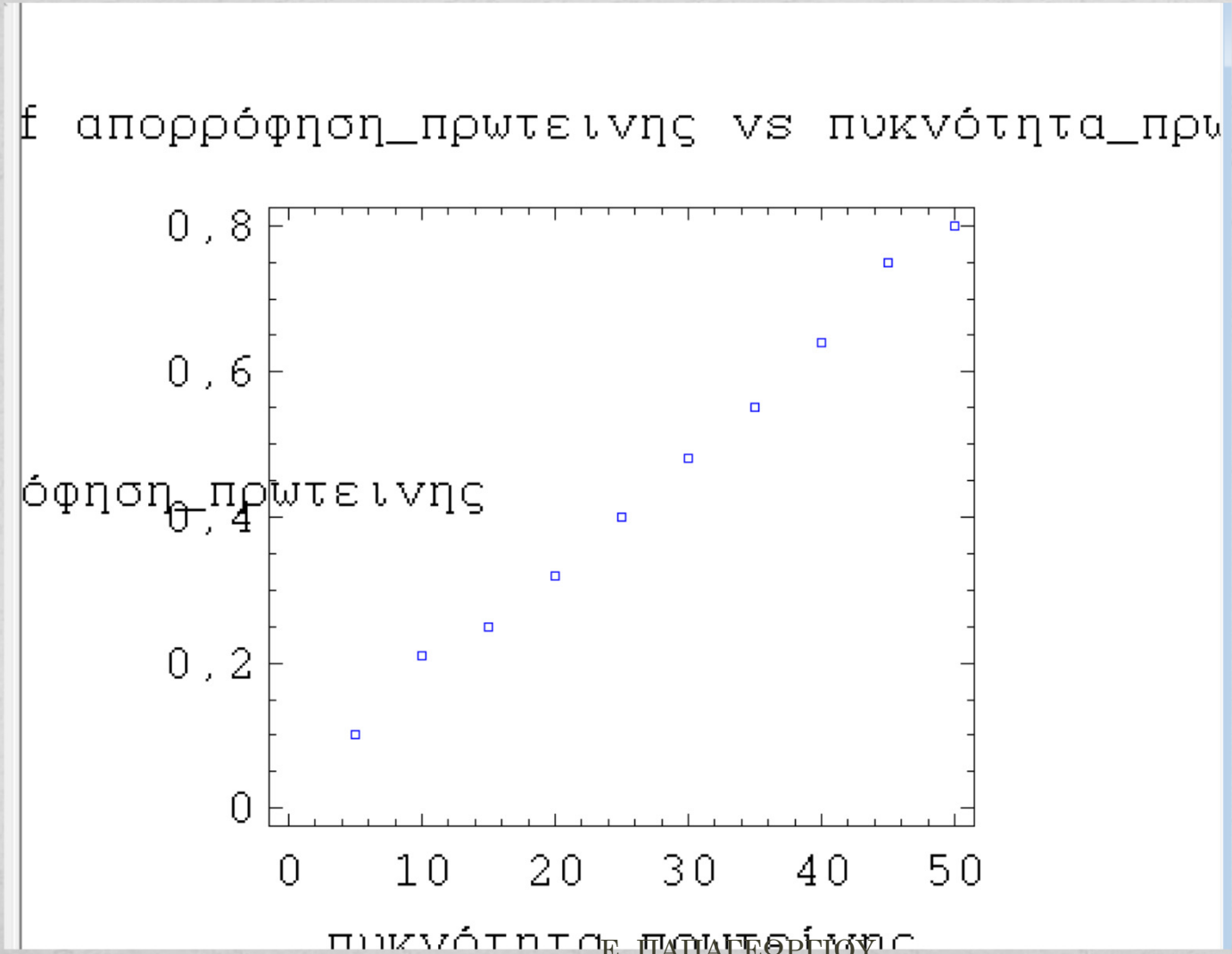
167.sf3

	απορρόφηση_πρ	πυκνότητα_πρω	Col_3	Col_4
1	0,10	5		
2	0,21	10		
3	0,25	15		
4	0,32	20		
5	0,40	25		
6	0,48	30		
7	0,55	35		
8	0,64	40		
9	0,75	45		
10	0,80	50		
11				
12				
13				
14				

2. Απλή Παλινδρόμηση

Στη συνέχεια κατασκευάζουμε ένα διάγραμμα διασποράς (scatter plot) το οποίο κρίνεται απαραίτητο προκειμένου να αναζητήσουμε αν υπάρχει κάποιου είδους σχέση μεταξύ των δύο μεταβλητών ή αν αυτές εμφανίζονται τυχαία κατανεμημένες.

2. Απλή Παλινδρόμηση



2. Απλή Παλινδρόμηση

Από το προηγούμενο διάγραμμα διασποράς (Scatterplot) είναι εμφανές ότι το μοντέλο μας είναι γραμμικό και συνεπώς μπορούμε να προχωρήσουμε στην εφαρμογή της αντίστοιχης θεωρίας για την απλή παλινδρόμηση και να εκτιμήσουμε τους συντελεστές της ευθείας που προσαρμόζεται στα δεδομένα μας.

2. Απλή Παλινδρόμηση

Διαπιστώσαμε:

- Ύπαρξη συσχέτισης μεταξύ X, Y
- Ύπαρξη γραμμικής συσχέτισης μεταξύ X,Y

$$Y = \alpha + \beta x \qquad (Y = \alpha + \beta x + \varepsilon)$$

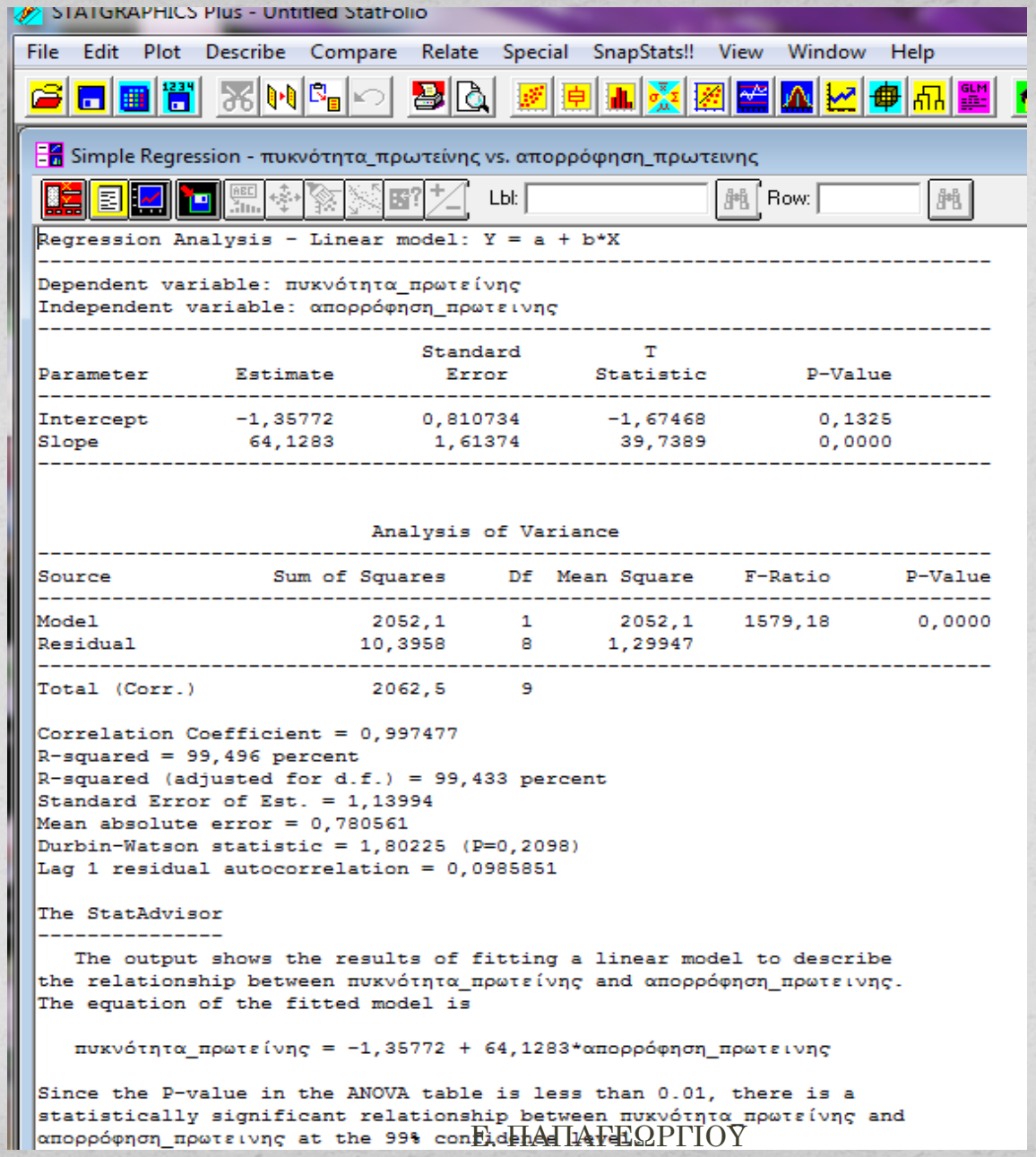
Συνεπώς προχωράμε σε εκτίμηση των παραμέτρων α,β:

$$\hat{\beta} = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - \sum y_i^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_x^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$\hat{Y} = \hat{\alpha} + \hat{\beta} x$$

Προχωράμε στην επίλυση και παρατίθεται το αποτέλεσμα:


STATGRAPHICS Plus - Untitled StatFolio
File Edit Plot Describe Compare Relate Special SnapStats!! View Window Help
Simple Regression - πυκνότητα_πρωτείνης vs. απορρόφηση_πρωτεινης
Lbl:
Row:
Regression Analysis - Linear model: Y = a + b*X
Dependent variable: πυκνότητα_πρωτείνης
Independent variable: απορρόφηση_πρωτεινης
Parameter Estimate Standard Error T Statistic P-Value
Intercept -1,35772 0,810734 -1,67468 0,1325
Slope 64,1283 1,61374 39,7389 0,0000
Analysis of Variance
Source Sum of Squares Df Mean Square F-Ratio P-Value
Model 2052,1 1 2052,1 1579,18 0,0000
Residual 10,3958 8 1,29947
Total (Corr.) 2062,5 9
Correlation Coefficient = 0,997477
R-squared = 99,496 percent
R-squared (adjusted for d.f.) = 99,433 percent
Standard Error of Est. = 1,13994
Mean absolute error = 0,780561
Durbin-Watson statistic = 1,80225 (P=0,2098)
Lag 1 residual autocorrelation = 0,0985851
The StatAdvisor
The output shows the results of fitting a linear model to describe the relationship between πυκνότητα_πρωτείνης and απορρόφηση_πρωτεινης.
The equation of the fitted model is
πυκνότητα_πρωτείνης = -1,35772 + 64,1283*απορρόφηση_πρωτεινης
Since the P-value in the ANOVA table is less than 0.01, there is a statistically significant relationship between πυκνότητα_πρωτείνης and απορρόφηση_πρωτεινης at the 99%

2. Απλή Παλινδρόμηση

- Όπως διαπιστώνουμε από την προηγούμενη διαφάνεια, η ζητούμενη ευθεία είναι:

$$\hat{y} = \hat{\alpha} + \hat{\beta}x$$

πυκνότητα_πρωτείνης = -1,35772 + 4,1283*απορρόφηση_πρωτεινης

$$\hat{y} = -1,35772 + 4,1283x$$

Επίσης η προσαρμογή του μοντέλου είναι πάρα πολύ καλή αφού:

R-squared = 99,496 percent

Ταυτόχρονα εμφανίζεται και το ακόλουθο γράφημα:

2. Απλή Παλινδρόμηση

- Η τιμή ***του Συντελεστή γραμμικής συσχέτισης του Pearson***είναι 0.997477, και ερμηνεύεται όπως αναφέραμε στην αντίστοιχη παράγραφο.
- Η τιμή ***R square (Δείκτης προσδιορισμού)***είναι 99.496 και είναι το τετράγωνο του συντελεστή γραμμικής συσχέτισης του Pearson, (0.997477^2=0.99496). Εκφράζεται σε % και όσο πιο κοντά στο 100 βρίσκεται η τιμή του, τόσο πιο καλή προσαρμογή του μοντέλου έχουμε.

Στο παράδειγμά μας έχουμε σχεδόν τέλεια προσαρμογή του γραμμικού μοντέλου

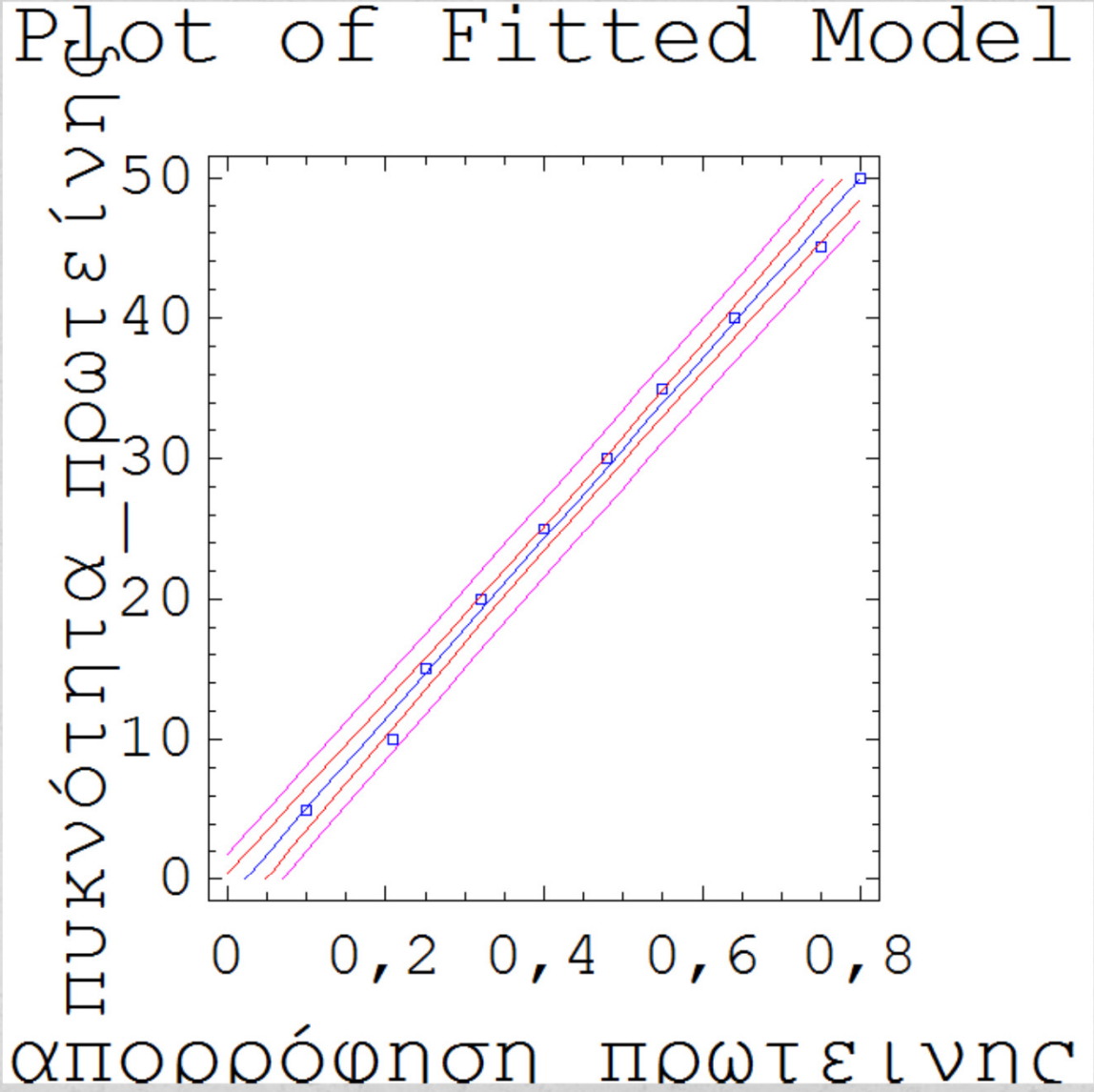
2. Απλή Παλινδρόμηση

Οι αντίστοιχες τιμές του ***p-value*** που εμφανίζονται στον πίνακα υπολογισμού του σταθερού όρου(α) και της κλίσης(β), αφορούν ελέγχους με μηδενικές υποθέσεις αντίστοιχα: το α=0 ή β=0.

Στο παράδειγμά μας έχουμε για το α p-value=0.1325 και για το β το p-value=0.

Συνεπώς αντιστοίχως αποδεχόμαστε την μηδενική υπόθεση για το α, δηλαδή δεχόμαστε α=0, ενώ απορρίπτουμε την μηδενική υπόθεση για το β, δηλαδή το β δεν είναι μηδέν.

2. Απλή Παλινδρόμηση



Ερμηνεία του β:

Ο συντελεστής β εκφράζει την μεταβολή στην εξαρτημένη μεταβλητή Υ όταν η ανεξάρτητη μεταβλητή Χ αυξηθεί κατά μία μονάδα.

Επίσης, εάν β=0, το μοντέλο παίρνει την μορφή:

$$Y=\alpha$$

Και συνεπώς η ανεξάρτητη μεταβλητή Χ δεν επηρεάζει καθόλου την Υ.

Σ᾽αυτό το σημείο πρέπει να προσέξουμε γιατί στην ακρίβεια η ανεξάρτητη μεταβλητή Χ δεν έχει καμμία γραμμική σχέση με την Υ. Δεν αποκλείονται όμως άλλου είδους επιδράσεις.

2. Απλή Παλινδρόμηση

Παράλληλα έχουμε την δυνατότητα να κάνουμε πρόβλεψη μέσω του γραμμικού μας μοντέλου για την πυκνότητα πρωτείνης για οποιαδήποτε τιμή της απορρόφησης πρωτείνης που μας ενδιαφέρει.

Παρατίθεται ένα παράδειγμα:

Έστω ότι θέλουμε να κάνουμε πρόβλεψη στις τιμές της απορρόφησης της πρωτείνης $\chi=0.1$ και $\chi=0.8$.

Ταυτόχρονα έχουμε την δυνατότητα να έχουμε ένα 95% Δ.Ε. για την πρόβλεψη.

Παίρνουμε το ακόλουθο αποτέλεσμα:

2. Απλή Παλινδρόμηση

Predicted Values

X	Predicted Y	95,00% Prediction Limits Lower	95,00% Prediction Limits Upper	95,00% Confidence Limits Lower	95,00% Confidence Limits Upper
0,1	5,05511	2,00591	8,10431	3,50999	6,60023
0,8	49,9449	46,8957	52,9941	48,3998	51,49

The StatAdvisor

This table shows the predicted values for πυκνότητα_πρωτεΐνης using the fitted model. In addition to the best predictions, the table shows:

(1) 95,0% prediction intervals for new observations

(2) 95,0% confidence intervals for the mean of many observations

The prediction and confidence intervals correspond to the inner and outer bounds on the graph of the fitted model.

Β. ΜΗ-ΠΑΡΑΜΕΤΡΙΚΕΣ ΔΟΚΙΜΑΣΙΕΣ

Έλεγχος ανεξαρτησίας (συσχέτισης) 2 κατηγορικών μεταβλητών

- Παράδειγμα
  - «εξαρτάται το βρογχικό άσθμα από το κάπνισμα των γονέων; »
  - «επηρεάζει η έντονη φυσική δραστηριότητα την κατηγορία σωματικού βάρους;»
  - «οι υπερτασικοί ασθενείς διαφέρουν ανά φύλο;»

Έλεγχος ανεξαρτησίας 2 ποιοτικών χαρακτηριστικών

Το κριτήριο X^2

Το στατιστικό κριτήριο που χρησιμοποιείται είναι το X^2

- Είναι ένα μέτρο απόστασης δύο «καταστάσεων»

$$X^2 = \sum \frac{(\Pi - A)^2}{A}$$

Π=παρατηρηθείσες συχνότητες, **Α**=αναμενόμενες συχνότητες

Η «φιλοσοφία» του κριτηρίου

Δειγματοληπτικά στοιχεία (**πραγματικά δεδομένα**)

Το κριτήριο χ^2 «μετρά» την απόσταση των δύο πινάκων

Θεωρητικά στοιχεία που θα είχαμε «αν δεν υπάρχει εξάρτηση (Ho)»

X / Y	**A1 (π.χ. ασθενείς)**	**A2 (π.χ. υγιείς)**	Σύνολο
B1 (παράγοντας παρών)	**α**	**β**	R1
B2 (παράγοντας απών)	**γ**	**δ**	R2
Σύνολο	**C1**	**C2**	**n**

X' / Y'	**A1 (π.χ. ασθενείς)**	**A2 (π.χ. υγιείς)**	Σύνολο
B1 (παράγοντας παρών)	**Α'**	**Β'**	R1
B2 (παράγοντας απών)	**Γ'**	**Δ'**	R2
Σύνολο	**C1**	**C2**	**n**

Το κριτήριο X^2

Με βάση τη θεωρία το κριτήριο X^2 είναι το ακόλουθο:

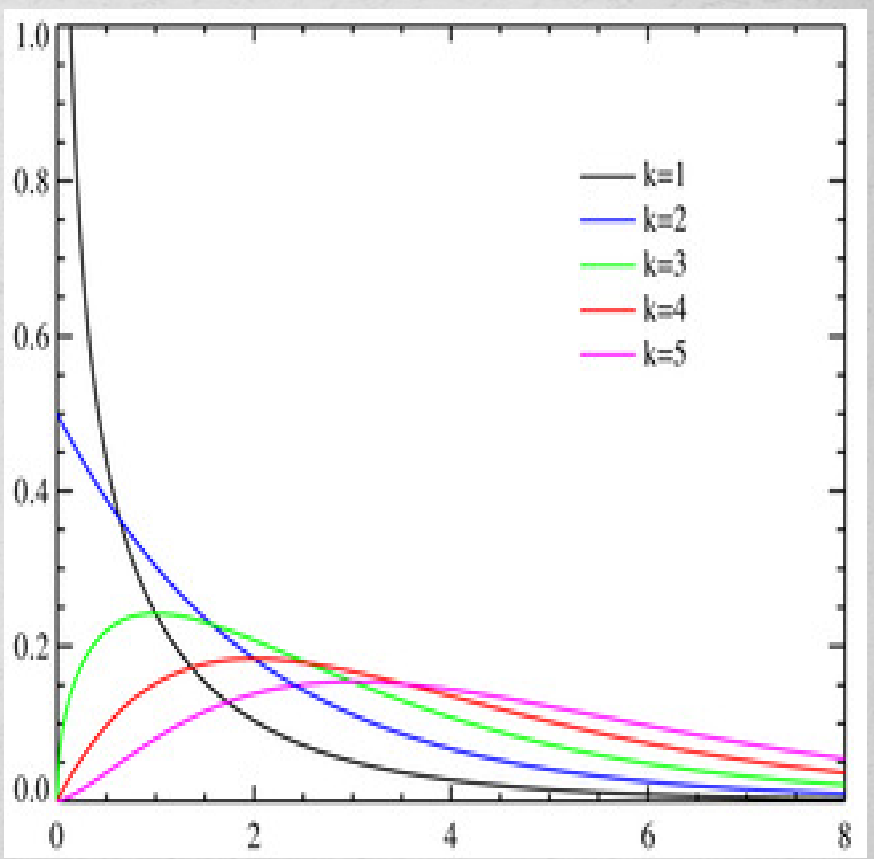
$$X^2 = \frac{(a - \text{A}')^2}{\text{A}'} + \frac{(\beta - \text{B}')^2}{\text{B}'} + \frac{(\gamma - \Gamma')^2}{\Gamma'} + \frac{(\delta - \Delta')^2}{\Delta'}$$

Έλεγχος ανεξαρτησίας 2 ποιοτικών χαρακτηριστικών

- Όσο **πιο μεγάλες τιμές** λαμβάνει το κριτήριο X^2 (άρα p<<) τόσο πιο κοντά είμαστε στο να απορρίψουμε την H_0 , **δηλαδή υπάρχει συσχέτιση**.
- Όσο **πιο μικρές τιμές** (≈0) λαμβάνει το κριτήριο X^2 (άρα p>>) τόσο πιο κοντά είμαστε στο να ΜΗΝ απορρίψουμε την H_0, **δηλαδή δεν υπάρχει συσχέτιση**.

Η κατανομή X^2

- Ασύμμετρη
- Θετικά ορισμένη
- Η μορφή της εξαρτάται από τους βαθμούς ελευθερίας Β.Ε= (κ-1) (λ – 1) όπου κ, λ ο αριθμός των γραμμών και των στηλών του πίνακα
- Με βάση τους βαθμούς ελευθερίας και την χρήση ειδικών πινάκων υπολογίζουμε την κρίσιμη τιμή του ελέγχου ξ



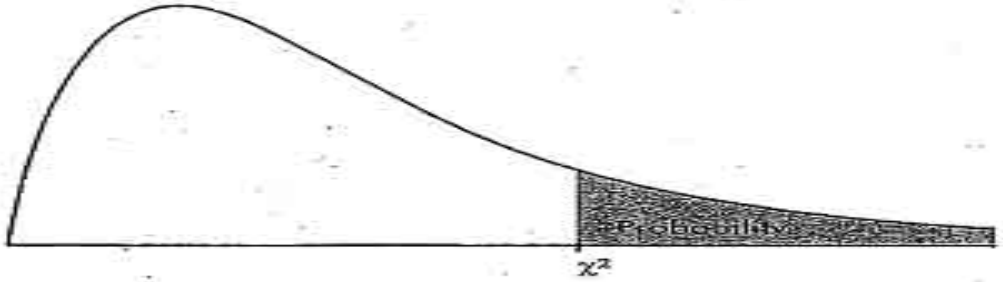


TABLE C: χ^2 CRITICAL VALUES

	Tail probability p										
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59
11	13.70	14.63	15.77	17.28	19.68	21.92	22.62	24.72	26.76	28.73	31.26
12	14.85	15.81	16.99	18.55	21.03	23.34	24.05	26.22	28.30	30.32	32.91
13	15.98	16.98	18.20	19.81	22.36	24.74	25.47	27.69	29.82	31.88	34.53
14	17.12	18.15	19.41	21.06	23.68	26.12	26.87	29.14	31.32	33.43	36.12
15	18.25	19.31	20.60	22.31	25.00	27.49	28.26	30.58	32.80	34.95	37.70
16	19.37	20.47	21.79	23.54	26.30	28.85	29.63	32.00	34.27	36.46	39.25
17	20.49	21.61	22.98	24.77	27.59	30.19	31.00	33.41	35.72	37.95	40.79
18	21.60	22.76	24.16	25.99	28.87	31.53	32.35	34.81	37.16	39.42	42.31
19	22.72	23.90	25.33	27.20	30.14	32.85	33.69	36.19	38.58	40.88	43.82
20	23.83	25.04	26.50	28.41	31.41	34.17	35.02	37.57	40.00	42.34	45.31
21	24.93	26.17	27.66	29.62	32.67	35.48	36.34	38.93	41.40	43.78	46.80
22	26.04	27.30	28.82	30.81	33.92	36.78	37.66	40.29	42.80	45.20	48.27
23	27.14	28.43	29.98	32.01	35.17	38.08	38.97	41.64	44.18	46.62	49.73
24	28.24	29.55	31.13	33.20	36.42	39.36	40.27	42.98	45.56	48.03	51.18
25	29.34	30.68	32.28	34.38	37.65	40.65	41.57	44.31	46.93	49.44	52.62
26	30.43	31.79	33.43	35.56	38.89	41.92	42.86	45.64	48.29	50.83	54.05
27	31.53	32.91	34.57	36.74	40.11	43.19	44.14	46.96	49.64	52.22	55.48
28	32.62	34.03	35.71	37.92	41.34	44.46	45.42	48.28	50.99	53.59	56.89
29	33.71	35.14	36.85	39.09	42.56	45.72	46.69	49.59	52.34	54.97	58.30
30	34.80	36.25	37.99	40.26	43.77	46.98	47.96	50.89	53.67	56.33	59.70
40	45.62	47.27	49.24	51.81	55.76	59.34	60.44	63.69	66.77	69.70	73.40
50	56.33	58.16	60.35	63.17	67.50	71.42	72.61	76.15	79.49	82.66	86.66
60	66.98	68.97	71.34	74.40	79.08	83.30	84.58	88.38	91.95	95.34	99.61
80	88.13	90.41	93.11	96.58	101.9	106.6	108.1	112.3	116.3	120.1	124.8
100	109.1	111.7	114.7	118.5	124.3	129.6	131.1	135.8	140.2	144.3	149.4

Προϋποθέσεις εφαρμογής του κριτηρίου X^2

- Τυχαίο δείγμα και ανεξαρτησία των παρατηρήσεων
- Κανένα κελί με μηδενική τιμή
- Όλες οι αναμενόμενες τιμές των κελιών 2x2 πινάκων συνάφειας >5
- Το 80% των κελιών πινάκων r x c να έχουν αναμενόμενες τιμές > 5

- **Παράδειγμα**

Σε 500 μαθητές δημοτικού σχολείου μελετήθηκε η σχέση της υγείας του στόματος τους με τη χλωρίωση του νερού στην περιοχή διαμονής τους. Η κατανομή των 500 μαθητών ανάλογα με την υγεία του στόματος και τη χλωρίωση του νερού ήταν:

		Υγεία στόματος		
Χλωρίωση νερού		Κακή	Μέτρια	Καλή
Ανεπαρκής		80	120	75
Επαρκής	40	80	105	
Σύνολο		120	200	180

Σχετίζεται η υγεία του στόματος των μαθητών με τη χλωρίωση του νερού;

Η μηδενική υπόθεση στην δοκιμασία X^2 αφορά στην ανεξαρτησία των μεταβλητών.
Αρχικά θα υπολογίσουμε τα θεωρητικά μεγέθη δηλ. τα «Expected », τα οποία συμβολίζονται με Ε στον κάτωθι τύπο. Με Ο συμβολίζονται τα παρατηρούμενα δηλ. τα «Observed».

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

Εν συνεχεία με τον ανωτέρω τύπο υπολογίζουμε την τιμή του κριτηρίου X^2 (χι-τετράγωνο) και την συγκρίνουμε με την τιμή της κατανομής X^2, προκειμένου να αποφανθούμε.

$$R = \{ X^2 > X^2_{(s-1)(k-1);a} \}$$

Όπως φαίνεται στο παρακάτω παράθυρο «Frequency Table» τα θεωρητικά μεγέθη εμφανίζονται κάτω από τα παρατηρούμενα:

Frequency Table

	kaki	kalh	metria	Row Total
Row_1	80	75	120	275
	66,00	99,00	110,00	27,50%
Row_2	40	105	80	225
	54,00	81,00	90,00	22,50%
Row_3	120	180	200	500
	120,00	180,00	200,00	50,00%
Column Total	240	360	400	1000
	24,00%	36,00%	40,00%	100,00%

Chi-Square Test

Chi-Square Test

Chi-Square	Df	P-Value
21,55	4	0,0002

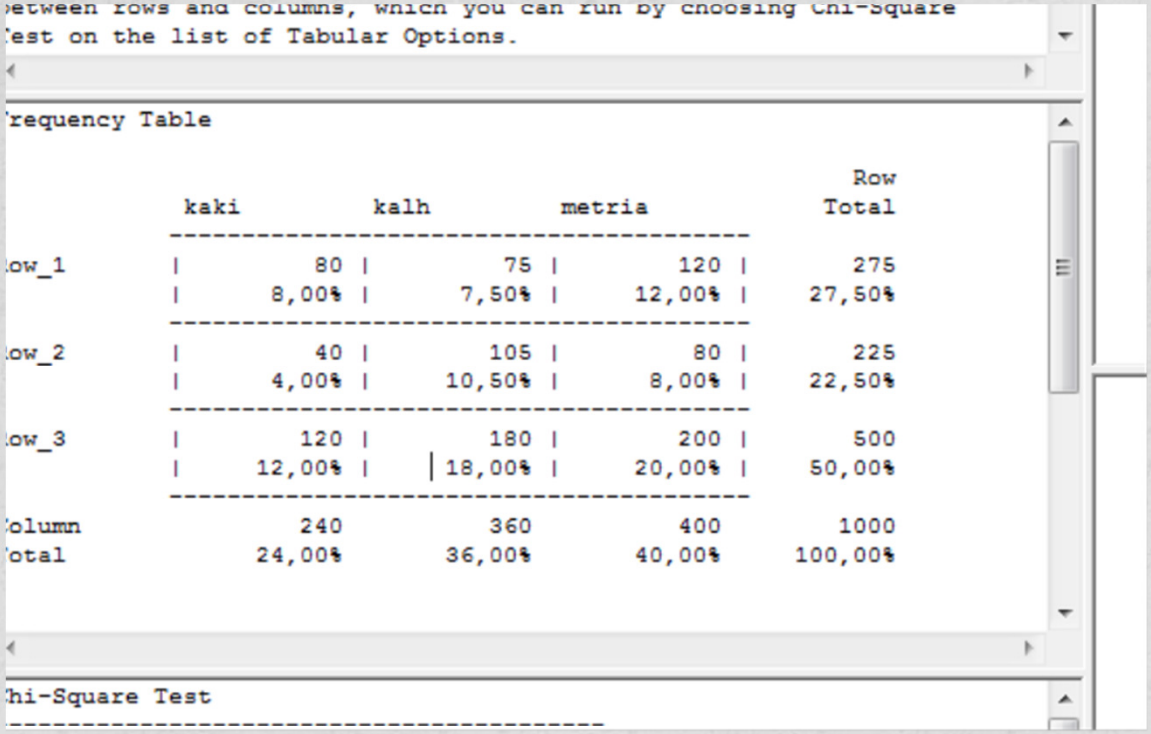
The StatAdvisor

The chi-square test performs a hypothesis test to determine whether or not to reject the idea that the row and column classifications are independent. Since the P-value is less than 0.01, we can reject the hypothesis that rows and columns are independent at the 99% confidence level. Therefore, the observed row for a particular case is related to its column.

Παρατηρούμε ότι η τιμή του X^2 κριτηρίου είναι 21.55, οι βαθμοί ελευθερίας 4 ενώ η τιμή του P-Value 0.0002 το οποίο μας οδηγεί στο συμπέρασμα ότι με βεβαιότητα 99% απορρίπτουμε την Ho αφού η τιμή του είναι μικρότερη από το επίπεδο σημαντικότητας α (0,0002<0,01). Συνεπώς, η υγεία του στόματος των μαθητών δεν είναι ανεξάρτητη της χλωρίωσης του νερού που πίνουν.

Chi-Square	Df	P-Value
21,55	4	0,0002

Σε αυτό το σημείο η άσκηση με την χρήση του στατιστικού λογισμικού λύθηκε. Αν θελήσουμε όμως παραπάνω πληροφορίες για τις ποσοστιαίες αναλογίες των παρατηρούμενων τιμών ή για τον υπολογισμό των θεωρητικών τιμών κάνουμε τα εξής επιπλέον βήματα:



between rows and columns, which you can run by choosing Chi-Square
Test on the list of Tabular Options.

Frequency Table

	kaki	kalh	metria	Row Total
ow_1	80	75	120	275
	8,00%	7,50%	12,00%	27,50%
ow_2	40	105	80	225
	4,00%	10,50%	8,00%	22,50%
ow_3	120	180	200	500
	12,00%	18,00%	20,00%	50,00%
olumn	240	360	400	1000
otal	24,00%	36,00%	40,00%	100,00%

hi-Square Test

Συμπεράσματα

Έλεγχος ανεξαρτησίας 2 ποιοτικών χαρακτηριστικών

- Ο έλεγχος χ^2 **αναδεικνύει** πιθανή εξάρτηση μεταξύ 2 κατηγορικών μεταβλητών.
- Ο έλεγχος χ^2 **ΔΕΝ αναδεικνύει** γραμμική σχέση μεταξύ 2 κατηγορικών μεταβλητών.
- Ο έλεγχος χ^2 **ΔΕΝ αναδεικνύει** επιμέρους διαφορές στις κατηγορίες των κατηγορικών μεταβλητών.

Ευχαριστώ για
το ενδιαφέρον σας