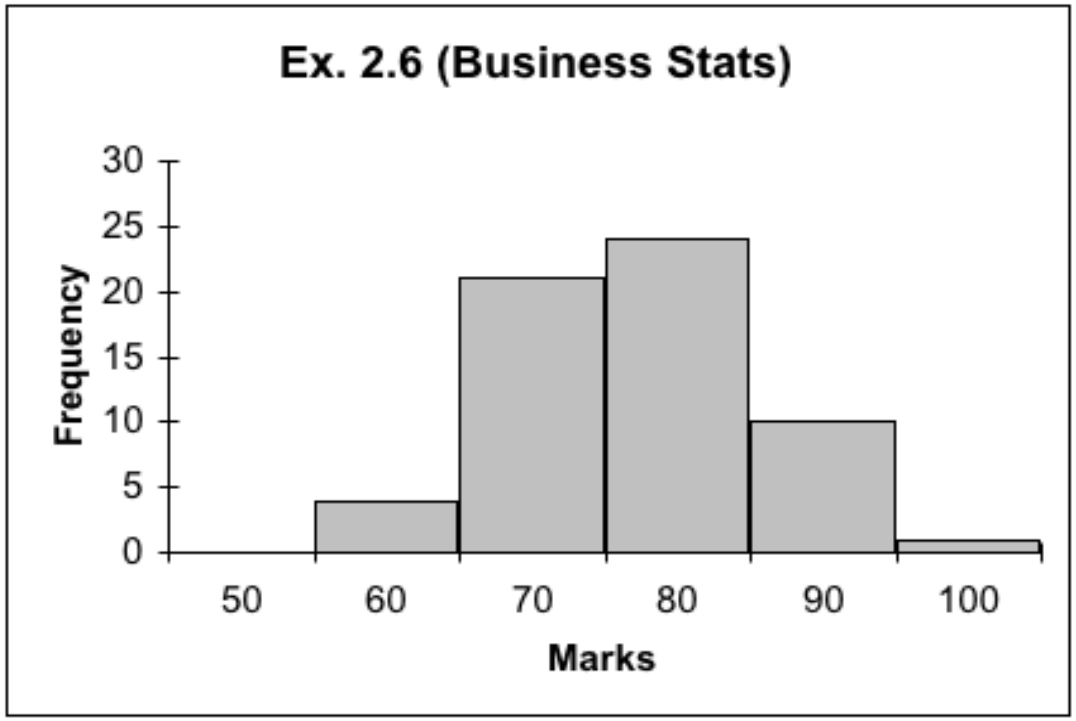
Αριθμητικές Περιγραφικές Στατιστικές

Κλωνάρης Στάθης

Περισσότερα σε Περιγραφική Στατιστική

- Θυμηθείτε, που χρησιμοποιήσαμε γραφικές τεχνικές για να περιγράψουμε δεδομένα:



Ενώ αυτό το ιστόγραμμα δίνει κάποια νέα πληροφόρηση, άλλα ενδιαφέροντα ερωτήματα (π.χ. ποιος είναι ο μέσος όρος της τάξης;) δεν απαντιέται.

Αριθμητικές Περιγραφικές Τεχνικές

- Μέτρα Κεντρικής Θέσης
  - Μέση Τιμή (Mean), Διάμεσος (Median), Επικρατούσα τιμή ή Κορυφή (Mode)
- Μέτρα Μεταβλητότητας
  - Εύρος (Range), Τυπική Απόκλιση (Standard Deviation), Διακύμανση (Variance), Συντελεστής Μεταβλητότητας (Coefficient of Variation)
- Μέτρα Σχετικής Τοποθεσίας
  - Ποσοστημόρια (Percentile), Τεταρτημόρια (Quartile)

Μέτρα Κεντρικής Θέσης …

- Η ***αριθμητική μέση τιμή***, *μέσος όρος*, η απλά *μέση τιμή*, είναι το πιο δημοφιλή και χρήσιμο μέτρο κεντρικής θέσης.

- Υπολογίζεται απλά προσθέτοντας όλες τις παρατηρήσεις και διαιρώντας με τον συνολικό αριθμό των παρατηρήσεων:

$$\text{Μέση Τιμή} = \frac{\text{Άθροισμα των Παρατηρήσεων}}{\text{Αριθμός των Παρατηρήσεων}}$$

Συμβολισμός …

- Όταν αναφερόμαστε στον αριθμό των παρατηρήσεων ενός ***πληθυσμού***, χρησιμοποιούμε το κεφαλαίο γράμμα **N**.

- Όταν αναφερόμαστε στον αριθμό των παρατηρήσεων ενός ***δείγματος***, χρησιμοποιούμε το μικρό γράμμα **n**.

- Η μέση τιμή του ***πληθυσμού*** συμβολίζεται με το ελληνικό γράμμα ***μ***:

- Η μέση τιμή του ***δείγματος*** συμβολίζεται με: $\overline{x}$

Συμβολισμός …

	Πληθυσμός	Δείγμα
Μέγεθος	N	n
Μέση Τιμή	μ	$\bar{x}$

Αριθμητική Μέση Τιμή …

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Μέση Τιμή Πληθυσμού

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

Μέση Τιμή Δείγματος

Συμβολισμός …

	Πληθυσμός	Δείγμα
Μέγεθος	N	n
Μέση Τιμή	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Αριθμητική Μέση Τιμή …

- …είναι κατάλληλη για να περιγράψουμε δεδομένα μετρήσεων, π.χ. ύψη ανθρώπων, βαθμοί από εξετάσεις, κλπ.

- …επηρεάζεται σοβαρά από «ακραίες τιμές». Π.χ. εφόσον ένας εκατομμυριούχος μετακομίζει σε μία γειτονιά, ο μέσο οικογενειακό εισόδημα αυξάνει πολύ από πριν και δίνει λανθασμένη εντύπωση.

Γεωμετρικός Μέσος

Απλός Γεωμετρικός Μέσος

$$G = \sqrt[n]{\prod_{i=1}^{n} X_i}$$

$$\log G = \frac{1}{n}\sum_{i=1}^{n} \log X_i$$

Παράδειγμα

2.0	2.1	2.3	3.2	3.2	2.3
3.3	4.1	3.3	3.5	3.5	3.5
3.7	3.7	3.7	5.7	3.9	3.9
3.9	3.9	4.1	4.3	4.3	4.3
4.4	4.4	4.5	5.7	5.7	6.0

$$G = e^{\left(\frac{39,62051}{30}\right)} = 3,75$$

Γεωμετρικός Μέσος Κατανομής Συχνοτήτων

$$G = \sqrt[\sum_{i=1}^{k} f_i]{\prod_{i=1}^{k} X_i^{f_i}} \qquad \log G = \frac{1}{\sum_{i=1}^{k} f_i} \sum_{i=1}^{k} f_i \log X_i$$

Ιδιότητες Γεωμετρικού Μέσου

α) Για να υπολογιστεί απαιτείται $X_i > 0 \quad \forall i$

β) $G \le \bar{X}$

γ) Επηρεάζεται λιγότερο από τις ακραίες τιμές

δ) Δεν έχει ευρεία εφαρμογή

Μέτρα Κεντρικής Θέσης …

- Η ***διάμεσος*** (median)υπολογίζεται βάζοντας όλες τις παρατηρήσεις στην σειρά. Η *μεσαία* παρατήρηση είναι η διάμεσος.

Δεδομένα: {0, 7, 12, 5, 14, 8, 0, 9, 22} N=9 (μονός αριθμός)
Τα ταξινομούμε από το μικρότερο ως προς το μεγαλύτερο, και βρίσκουμε την κεντρική τιμή:
0 0 5 7 8 9 12 14 22

Δεδομένα: {0, 7, 12, 5, 14, 8, 0, 9, 22, 33} N=10 (ζυγός)
Τα ταξινομούμε από το μικρότερο ως προς το μεγαλύτερο, και υπάρχουν δύο κεντρικές τιμές (8 και 9) από τις οποίες παίρνουμε τον μέσο όρο
0 0 5 7 8 9 12 14 22 33
Διάμεσος = (8+9)÷2 = 8.5

Οι διάμεσοι του δείγματος και του πληθυσμού υπολογίζονται κατά τον ίδιο τρόπο.

Μέτρα Κεντρικής Θέσης …

- Η ***επικρατούσα τιμή ή κορυφή (mode)*** ενός συνόλου παρατηρήσεων είναι η τιμή που εμφανίζεται *πιο συχνά*.

- Ένα σύνολο δεδομένων ενδέχεται να έχει μία κορυφή ή δύο, ή περισσότερες κορυφές.

- Η κορυφή είναι χρήσιμη για όλους τους τύπους δεδομένων, και βασικά για <u>ονομαστικά δεδομένα.</u>

Οι κορυφές του δείγματος και του πληθυσμού υπολογίζονται κατά τον ίδιο τρόπο.

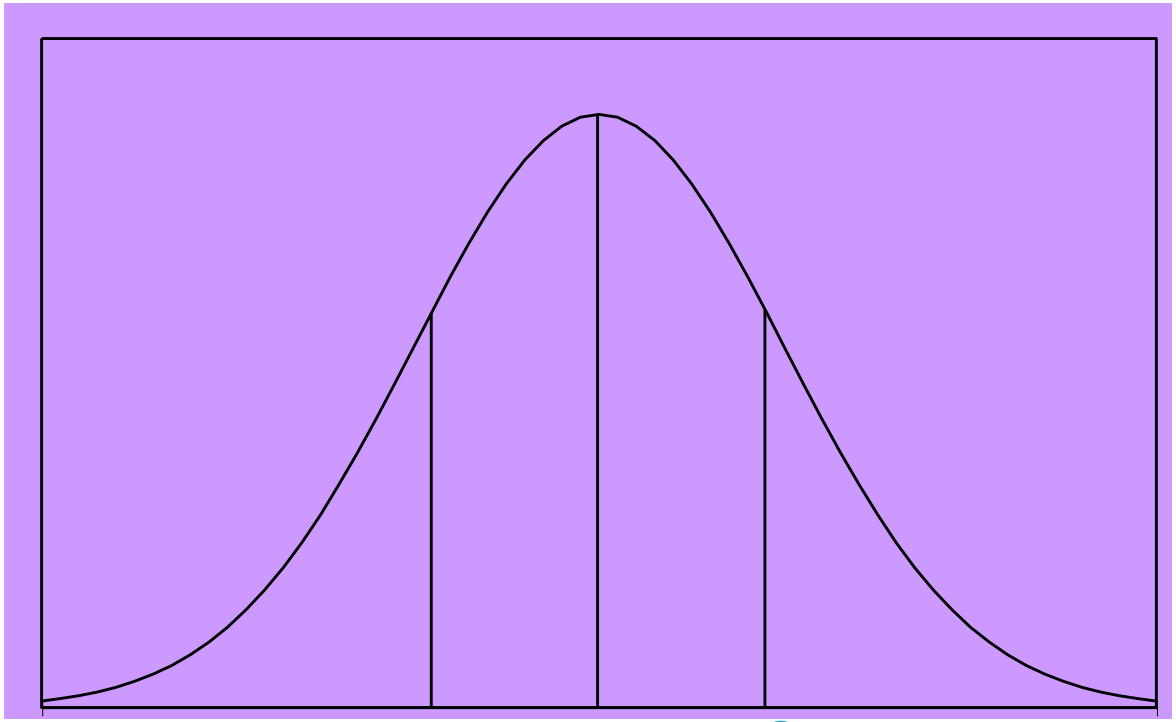
Κορυφή …

- Π.χ. Δεδομένα: {0, 7, 12, 5, 14, 8, 0, 9, 22, 33} N=10

- Ποια παρατήρηση εμφανίζεται ποιο συχνά;
- Η κορυφή για αυτό το σύνολο δεδομένων είναι 0. Πως αυτό είναι ένα μέτρο κεντρικής θέσης;

Μέση τιμή, Διάμεσος …

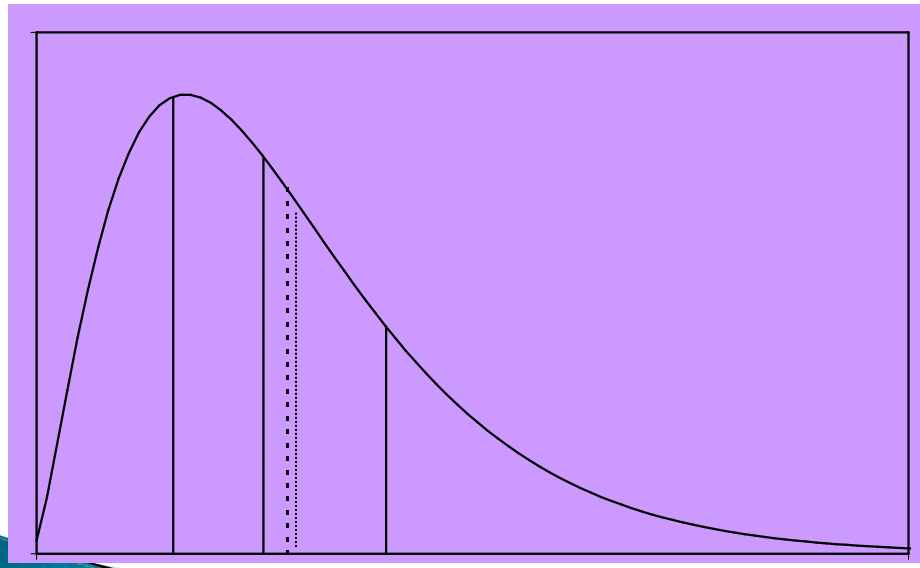
- Εάν η κατανομή είναι συμμετρική, η μέση τιμή, και η διάμεσος συμπίπτουν …



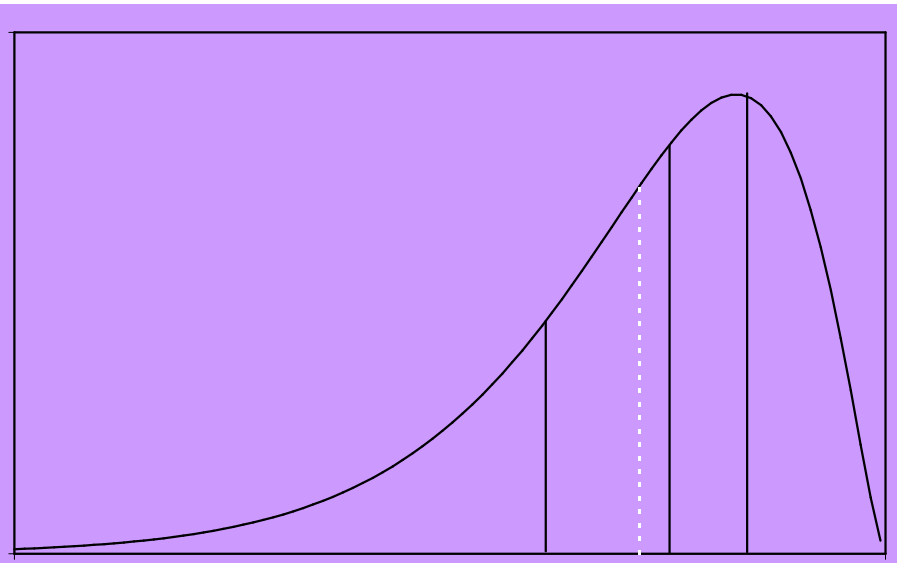
Μέση τιμή, Διάμεσος, …

- Εάν η κατανομή είναι ασύμμετρη, ας πούμε λοξή προς τα αριστερά ή προς τα δεξιά, τα τρία μέτρα θα διαφέρουν, π.χ.

Θετική Ασυμμετρία



Αρνητική Ασυμμετρία



Μέση τιμή, Διάμεσος, Κορυφές για Διατακτικά & Ονομαστικά Δεδομένα…

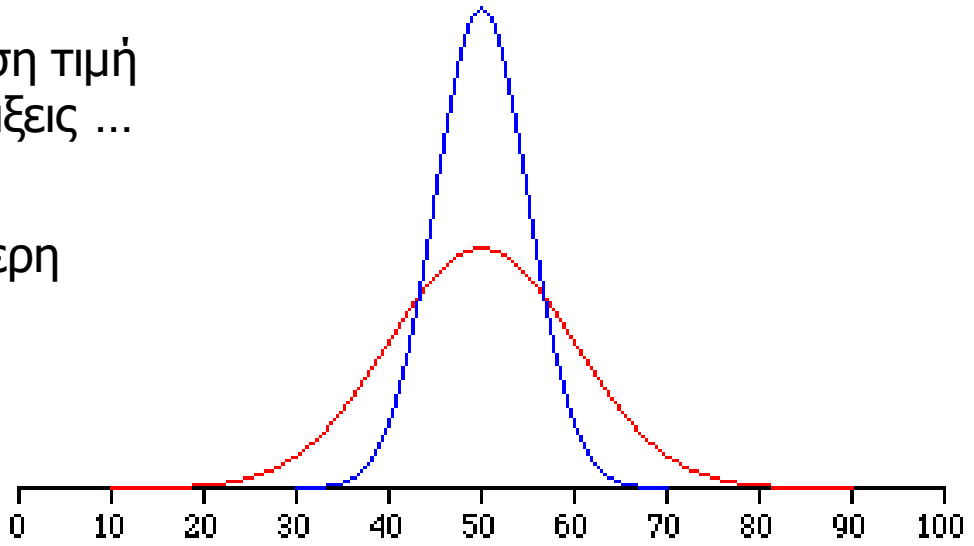
- Για διατακτικά και ονομαστικά δεδομένα ο υπολογισμός της μέσης τιμής δεν είναι έγκυρος.
- Η Διάμεσος είναι κατάλληλη για διατακτικά δεδομένα.
- Για ονομαστικά δεδομένα, ο υπολογισμός της κορυφής είναι χρήσιμος για να καθορίσει την μεγαλύτερη συχνότητα αλλά όχι την «κεντρική θέση».

Μέτρα Μεταβλητότητας …

- Τα μέτρα της κεντρικής θέσης δεν μας δίνουν όλα τα χαρακτηριστικά για μία κατανομή. Δηλαδή, πόσο πολύ είναι οι παρατηρήσεις απλωμένες γύρω από κέντρο;

Για παράδειγμα, βαθμοί από δύο διαφορετικές τάξεις δίνονται. Η μέση τιμή (=50) είναι η ίδια και για τις δύο τάξεις …

Αλλά, η κόκκινη τάξη έχει μεγαλύτερη μεταβλητότητα από την μπλε τάξη.



Εύρος …

- Το ***εύρος*** είναι το απλούστερο μέτρο μεταβλητότητας, υπολογίζεται ως:

- Εύρος = Μεγαλύτερη παρατήρηση – Μικρότερη παρατήρηση

Π.χ.
Δεδομένα: {4, 4, 4, 4, 50} Εύρος = 46
Δεδομένα : {4, 8, 15, 24, 39, 50} Εύρος = 46

Το εύρος είναι το ίδιο και στις δύο περιπτώσεις, αλλά τα σύνολα των δεδομένων έχουν πολύ διαφορετικές κατανομές …

Εύρος …

- Τα πλεονέκτημα του είναι η ευκολία με την οποία μπορεί να υπολογιστεί.
- Το βασικό μειονέκτημα είναι ότι δεν δίνει καμία πληροφορία για την διασπορά των παρατηρήσεων ανάμεσα στα δύο ακραία σημεία (min και max).
- Επομένως χρειαζόμαστε ένα μέτρο που να ενσωματώνει **όλα τα δεδομένα** και όχι μόνο δύο παρατηρήσεις. Επομένως …

Διακύμανση …

- Η διακύμανση και το παρεμφερή της μέτρο, τυπική απόκλιση είναι από τις πιο σημαντικές στατιστικές ποσότητες. Χρησιμοποιούνται για να μετρήσουν μεταβλητότητα, και επίσης παίζουν ένα κρίσιμο ρόλο σε όλες σχεδόν τις στατιστικές διαδικασίες για συμπερασματολογία (επαγωγή).

- Η διακύμανση του πληθυσμού συμβολίζεται με σ^2. (μικρό Ελληνικό γράμμα)

- Η διακύμανση του δείγματος συμβολίζεται με s^2. (μικρό "S" στο τετράγωνο)

Συμβολισμός …

	Πληθυσμός	Δείγμα
Μέγεθος	N	n
Μέση Τιμή	μ	$\bar{x}$
Διακύμανση	σ^2	s^2

Διακύμανση…

- Η διακύμανση του **πληθυσμού** είναι:

Μέση τιμή του πληθυσμού

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

Μέγεθος πληθυσμού

- Η διακύμανση του **δείγματος** είναι:

Μέση τιμή του δείγματος

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

Σημειώστε ότι ο παρανομαστής είναι το μέγεθος του δείγματος (n) μείον 1

Διακύμανση…

- Όπως μπορούμε να δούμε, έχουμε να υπολογίσουμε την μέση τιμή (x-παύλα) για να υπολογίσουμε την διακύμανση του δείγματος.

- Εναλλακτικά, υπάρχει ένας πιο σύντομος τύπος για να υπολογίσουμε την διακύμανση του δείγματος άμεσα από τα δεδομένα χωρίς να έχουμε το ενδιάμεσο βήμα του υπολογισμού της μέσης τιμής. Αυτός ο τύπος δίνεται από:

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

Εφαρμογή …

- Παράδειγμα: Από ένα δείγμα 6 γεωργικών εκμεταλλεύσεων το γεωργικό εισόδημα σε χιλ. € υπολογίστηκε αντίστοιχα σε 17, 15, 23, 7, 9, 13.
- Βρείτε το έσο γεωργικό εισόδημα και την διακύμανση του .

Μέση τιμή

$$\bar{x} = \frac{17+15+23+7+9+13}{6} = \frac{84}{6} = 14$$

Διακύμανση του δείγματος

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{1}{6-1}\left[(17-14)^2 + (15-14)^2 + \ldots(13-14)^2\right] = 33.2$$

Διακύμανση του δείγματος (σύντομη μέθοδο)

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right] = \frac{1}{6-1}\left[\left(17^2 + 15^2 + \ldots + 13^2\right) - \frac{(17+15+\ldots+13)^2}{6}\right] = 33.2$$

Τυπική Απόκλιση …

- Η τυπική απόκλιση είναι απλά η τετραγωνική ρίζα της διακύμανσης, έτσι:
- Η τυπική απόκλιση του πληθυσμού: $\sigma = \sqrt{\sigma^2}$
- Η τυπική απόκλιση του δείγματος: $s = \sqrt{s^2}$

Συμβολισμός …

	Πληθυσμός	Δείγμα
Μέγεθος	N	n
Μέση Τιμή	μ	$\bar{x}$
Διακύμανση	σ^2	s^2
Τυπική Απόκλιση	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$

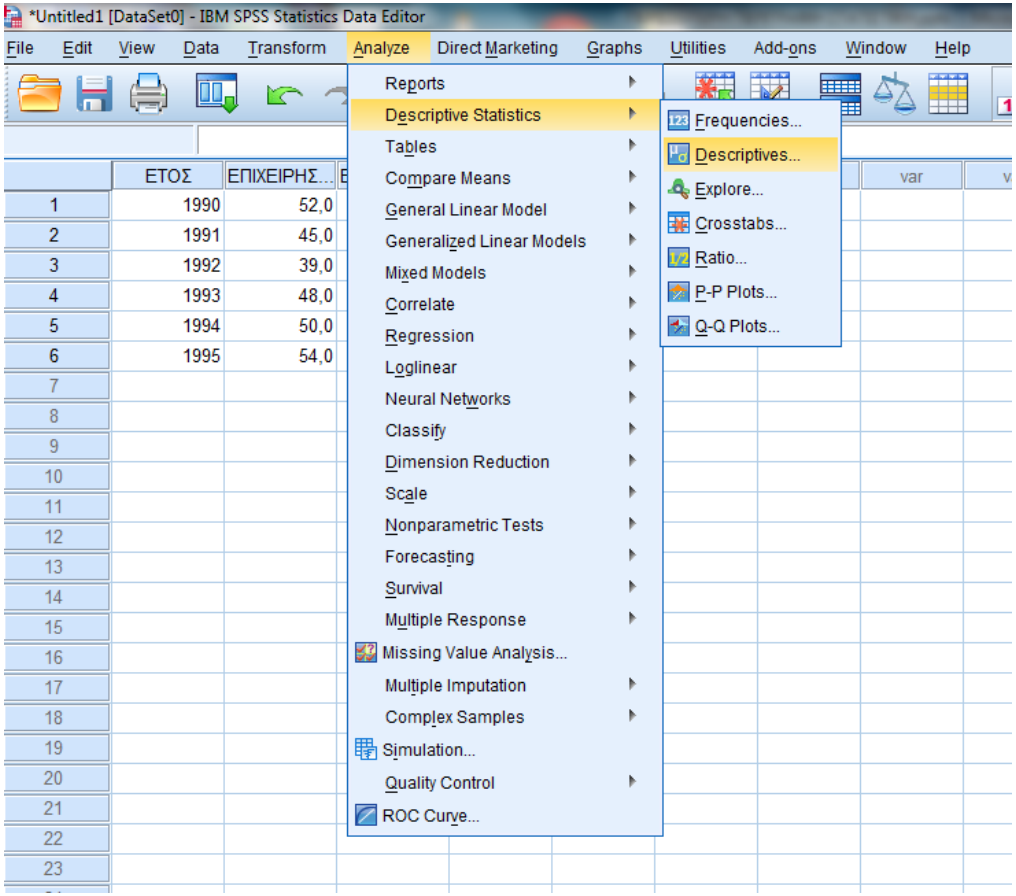
Τυπική Απόκλιση …

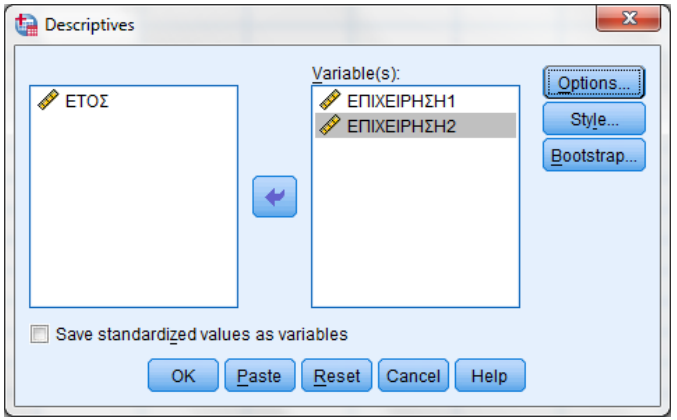
- Παράδειγμά: Δίνεται το ετήσιο ποσοστό κέρδους δύο επιχειρήσεων για 6 χρόνια. Αν έπρεπε να επιλέξετε την μετοχή μιας εκ των 2 με κριτήριο το ποσοστό κέρδους αυτά τα 6 χρόνια. Ποια θα ήταν η επιλογή;

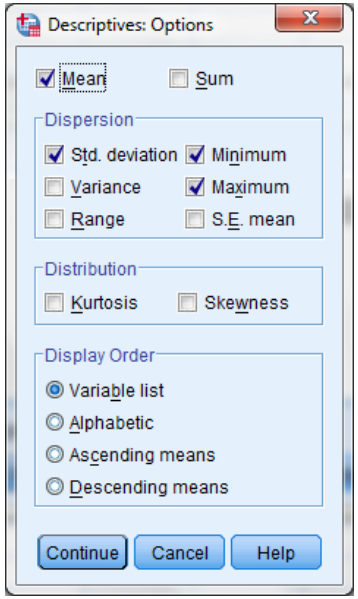
Έτος	1η Επιχείρηση	2η Επιχείρηση
1990	5.2	7.9
1991	4.5	7.0
1992	3.9	-5.3
1993	4.8	14.2
1994	5.0	-11.0
1995	5.4	16.0

Τυπική Απόκλιση …

Χρησιμοποιώντας το SPSS μπορούμε να υπολογίσουμε τις περιγραφικές στατιστικές που θα χρησιμοποιούμε για ερμηνεία …







Τυπική Απόκλιση …

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
ΕΠΙΧΕΙΡΗΣΗ1	6	39,0	54,0	48,000	5,4037	29,200
ΕΠΙΧΕΙΡΗΣΗ2	6	-110,0	160,0	48,000	107,6940	11598,000
Valid N (listwise)	6					

Το μέσο ετήσιο κέρδος και στις δύο επιχειρήσεις είναι το ίδιο (4.8) αλλά η διακύμανση και κατ' επέκταση η τυπική απόκλιση του κέρδους είναι μεγαλύτερη στην 2^{η} επιχείρηση από την πρώτη.

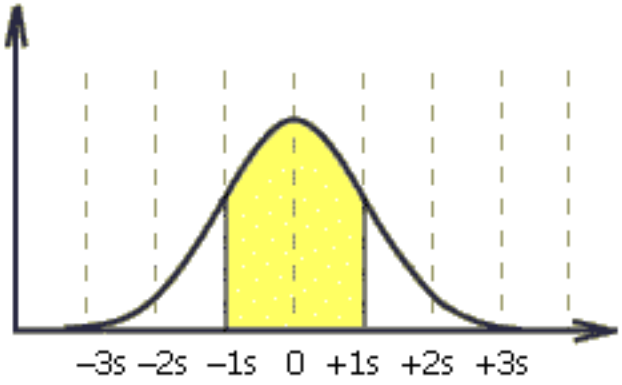
Ερμηνεύοντας την Τυπική Απόκλιση …

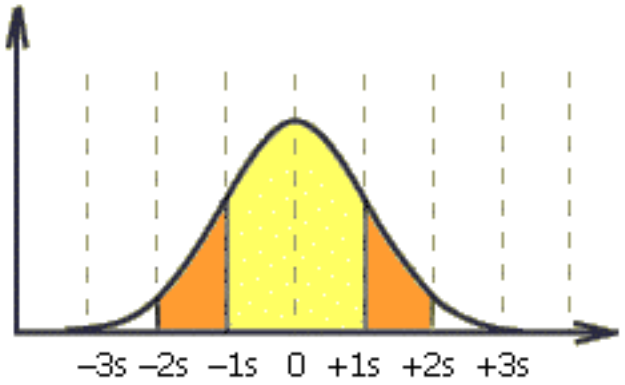
- Η τυπική απόκλιση μπορεί να χρησιμοποιηθεί για την σύγκριση της μεταβλητότητας αρκετών κατανομών και τον χαρακτηρισμό της γενικής μορφής μιας κατανομής. Εάν το ιστόγραμμα έχει **το σχήμα της καμπάνας**, μπορούμε να χρησιμοποιήσουμε τον ***Εμπειρίκο Κανόνα***, ο οποίος λέει:

1) Προσεγγιστικά 68% των παρατηρήσεων βρίσκονται εντός (±) μιας τυπικής απόκλισης από την μέση τιμή.
2) Προσεγγιστικά 95% των παρατηρήσεων βρίσκονται εντός (±) δύο τυπικών αποκλίσεων από την μέση τιμή.
3) Προσεγγιστικά 99.7% των παρατηρήσεων βρίσκονται εντός (±) τρεις τυπικές αποκλίσεις από την μέση τιμή.

Ο Εμπειρικός Κανόνας …

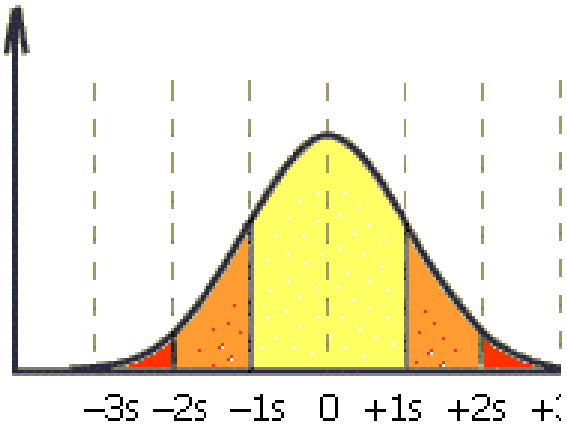
Προσεγγιστικά 68% των παρατηρήσεων βρίσκονται εντός (±) **μιας** τυπικής απόκλισης από την μέση τιμή.





Προσεγγιστικά 95% των παρατηρήσεων βρίσκονται εντός (±) **δύο** τυπικών αποκλίσεων από την μέση τιμή.

Προσεγγιστικά 99.7% των παρατηρήσεων βρίσκονται εντός (±) **τρεις** τυπικές αποκλίσεις από την μέση τιμή.



Το Θεώρημα του Chebysheff's ...

- Μία πιο γενική ερμηνεία της τυπικής απόκλισης εξάγεται από το θεώρημα του ***Chebysheff's***, το ποιο εφαρμόζεται σε όλες τις κατανομές με οποιαδήποτε μορφή.

- Το ποσοστό των παρατηρήσεων στο δείγμα πέφτει εντός (±) **k** τυπικών αποκλίσεων από την μέση τιμή είναι *τουλάχιστον*:

$$1 - \frac{1}{k^2} \quad for \quad k > 1$$

Για k=2 π.χ. το θεώρημα αναφέρει ότι *τουλάχιστον* 3/4 από όλες τις παρατηρήσεις πέφτουν 2 τυπικές αποκλίσεις από την μέση τιμή. Αυτό είναι ένα «κάτω φράγμα» (σε σύγκριση με την προσέγγιση του Εμπειρικό Κανόνα (95%).

Συντελεστής Μεταβλητότητας …

- Ο ***συντελεστής μεταβλητότητας*** ενός συνόλου παρατηρήσεων είναι η τυπική απόκλιση των παρατηρήσεων διαιρούμενη με την μέση τιμή,
- δηλαδή:

- Ο συντελεστής μεταβλητότητας του πληθυσμού

$$CV = \frac{\sigma}{\mu}$$

- Συντελεστής Μεταβλητότητας
- του δείγματος

$$CV = \frac{s}{\bar{x}}$$

Συμβολισμός …

	Πληθυσμός	Δείγμα
Μέγεθος	N	n
Μέση Τιμή	μ	$\bar{x}$
Διακύμανση	σ^2	s^2
Τυπική Απόκλιση	σ	s
Συντελεστής Μεταβλητότητας	CV	cv

Συντελεστής Μεταβλητότητας …

- Αυτός ο συντελεστής δείχνει ένα ***αναλογικό*** μέτρο της μεταβλητότητας, π.χ.

- Η τυπική απόκλιση του 10 ενδέχεται να ληφθεί ως μεγάλη όταν η μέση τιμή είναι 100, αλλά μόνο ως μέτρια όταν η μέση τιμή είναι 500.

Μέτρα Μεταβλητότητας …

- Εάν τα δεδομένα είναι συμμετρικά, χωρίς ακραίες τιμές, χρησιμοποιήστε το εύρος και την τυπική απόκλιση.

- Εάν συγκρίνουμε την μεταβλητότητα ανάμεσα σε δύο σύνολα δεδομένων διαφορετικών μονάδων, χρησιμοποιήστε τον συντελεστής μεταβλητότητας.

- Η έννοια της μεταβλητότητας δεν ορίζεται για ονομαστικά δεδομένα.

Μέτρα Σχετικής Τοποθεσίας

- Μέτρα Σχετικής Τοποθεσίας σχεδιάζονται για να προβάλουν πληροφόρηση σχετικά με την ***τοποθεσία*** κάποιων συγκεκριμένων τιμών ***σε σχέση*** με ολόκληρο το σύνολο των δεδομένων.

- ***Ποσοστημόριο:*** το P^{o} ποσοστημόριο είναι η τιμή από την οποία P ποσοστό των τιμών είναι *μικρότερο από* την τιμή αυτή και (100−P)% είναι μεγαλύτερο από την τιμή αυτή.

- Υποθέστε ότι ο βαθμός του πτυχίου σας είναι το 60ο ποσοστημόριο στο έτος σας , το οποίο σημαίνει ότι το 60% των άλλων σκορ ήταν *κάτω* από το δικό σας, ενώ το 40% των άλλων σκορ ήταν *κάτω* από το δικό σας.

Ποσοστημόριο …

- Έχουμε ειδικά ονόματα για το 25ο, 50ο, και 75ο ποσοστημόριο, χαρακτηριστικά ***τεταρτημόρια.***
- Το πρώτο τεταρτημόριο χαρακτηρίζει $Q_1 = 25^{ο}$ ποσοστημόριο.
- Το δεύτερο τεταρτημόριο, $Q_2 = 50^{ο}$ ποσοστημόριο (το οποίο είναι επίσης η διάμεσος).
- Το τρίτο τεταρτημόριο, $Q_3 = 75^{ο}$ ποσοστημόριο.
- Μπορούμε επίσης να αντιστοιχίσουμε ποσοστημόρια σε πεμπτημόρια (quintiles, fifths) και δεκατημόρια (deciles, tenths).

Χρήσιμα Ποσοστημόρια …

- Πρώτο δεκατημόριο = 10ο ποσοστημόριο
- Πρώτο τεταρτημόριο, Q_1, = 25ο ποσοστημόριο
- Διάμεσος, Q_2, = 50ο ποσοστημόριο
- Τρίτο τεταρτημόριο, Q_3, = 75ο ποσοστημόριο
- Ένατο δεκατημόριο = 90ο ποσοστημόριο

- **Σημειώστε:** Εάν ο βαθμός σου σε φέρνει στο 80ο ποσοστημόριο, αυτό δεν σημαίνει ότι απάντησες το 80% των ερωτήσεων σωστά – αυτό σημαίνει ότι το 80% των συμφοιτητών σου είχε σκορ **χαμηλότερο** από το δικό σου. Δείχνει την θέση σου σε σχέση με τους άλλους.

Θέση των Ποσοστημορίων …

- Ο ακόλουθος τύπος μας επιτρέπει να προσεγγίσουμε την θέση του κάθε ποσοστημορίου:

$$L_p = (n+1)\frac{P}{100}$$

Όπου L_p είναι η θέση του Pο ποσοστημόριου

Θέση των Ποσοστημορίων …

- Θεωρήστε τα δεδομένα:
- 0 0 5 7 8 9 12 14 22 33

- Ποια είναι η θέση του 25ο ποσοστημορίου. Δηλαδή, σε ποιο σημείο είναι το 25% των τιμών μικρότερες και 75% των τιμών μεγαλύτερες;

0 0 5 7 8 9 12 14 22 33

- $L_{25} = (10+1)(25/100) = \boxed{2.75}$

Το 25ο ποσοστημόριο είναι τρία-τέταρτα της απόστασης μεταξύ της δεύτερης (που είναι 0) και της τρίτης (που είναι 5) παρατήρησης. Τα τρία-τέταρτα της απόστασης είναι: (.75)(5 – 0) = 3.75
Επειδή η δεύτερη παρατήρηση είναι 0, το 25ο ποσοστημόριο είναι 0 + 3.75 = **3.75**

Θέση των Ποσοστημορίων …

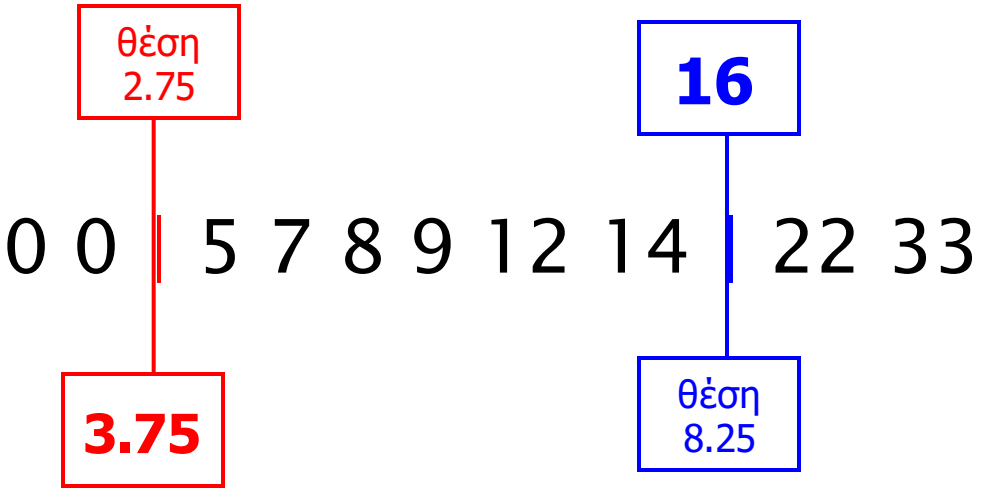
- Ποιο είναι το τρίτο τεταρτημόριο;
- $L_{75} = (10+1)(75/100) = 8.25$

0 0 5 7 8 9 12 14 22 33

Τοποθετείτε ένα-τέταρτο της απόστασης ανάμεσα στην όγδοη και ένατη παρατήρηση, οι οποίες είναι 14 και 22, αντίστοιχα. Το πρώτο τέταρτο της απόστασης είναι: (.25)(22 - 14) = 2, το οποίο σημαίνει ότι το 75ο ποσοστημόριο είναι: 14 + 2 = **16**

Θέση των Ποσοστημορίων …

- Παρακαλώ θυμηθείτε …



L_p καθορίζει την **θέση** στο σύνολο των δεδομένων όπου η τιμή του ποσοστημορίου βρίσκεται, όχι την τιμή του ποσοστημορίου.

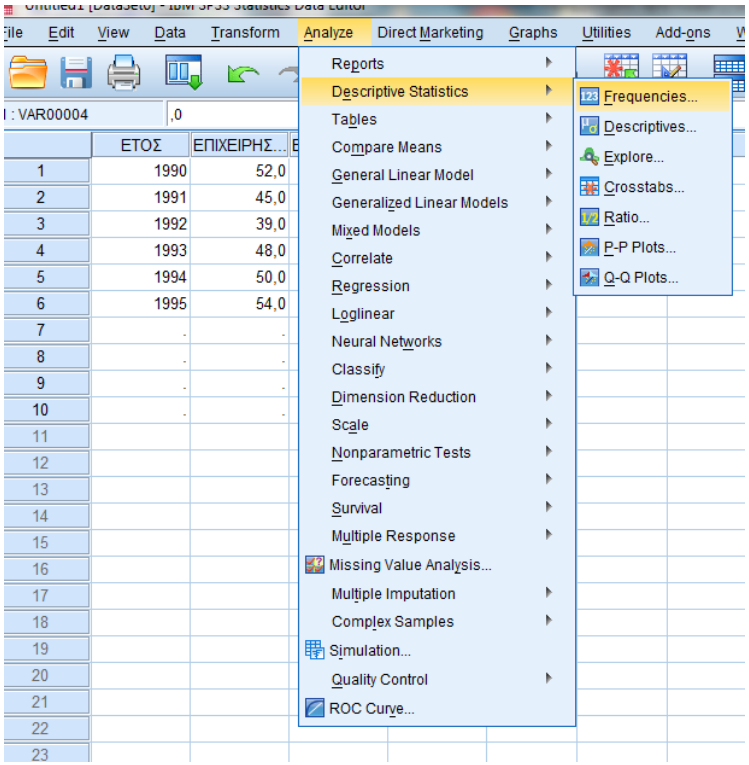
Ενδοτεταρτημοριακό Εύρος …

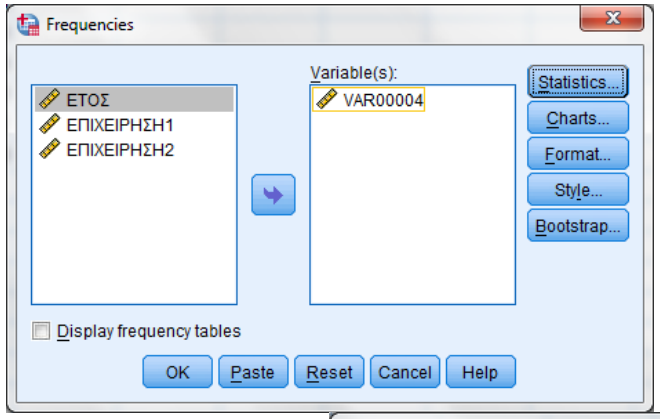
- Τα τεταρτημόρια μπορούν να χρησιμοποιηθούν για να δημιουργήσουν ένα άλλο μέτρο μεταβλητότητας, το ***ενδοτεταρτημοριακό εύρος*** το οποίο ορίζεται ως εξής:

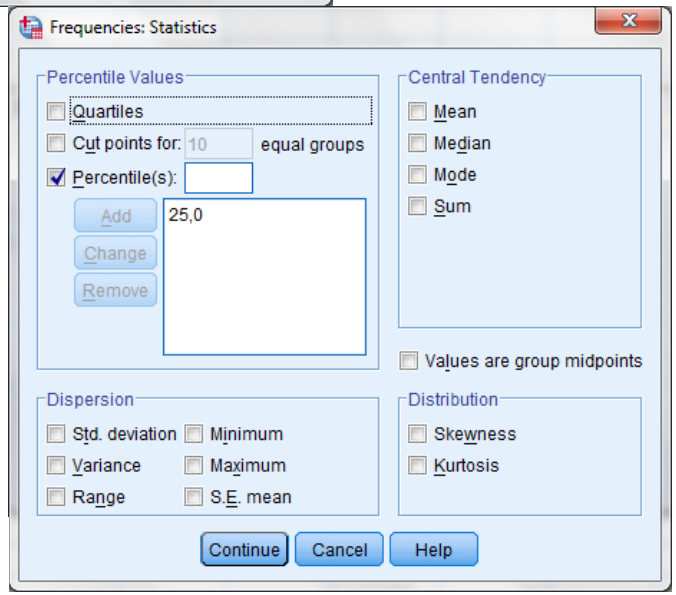
Ενδοτεταρτημοριακό Εύρος = $Q_3 - Q_1$

- Το ενδοτεταρτημοριακό εύρος μετράει το άπλωμα του 50% των μεσαίων παρατηρήσεων.
- Μεγάλες τιμές αυτής της στατιστικής σημαίνουν ότι το 1ο και 3ο τεταρτημόριο απέχουν υποδεικνύοντας υψηλό επίπεδο μεταβλητότητας.

Υπολογισμός ποσοστημορίων στο SPSS







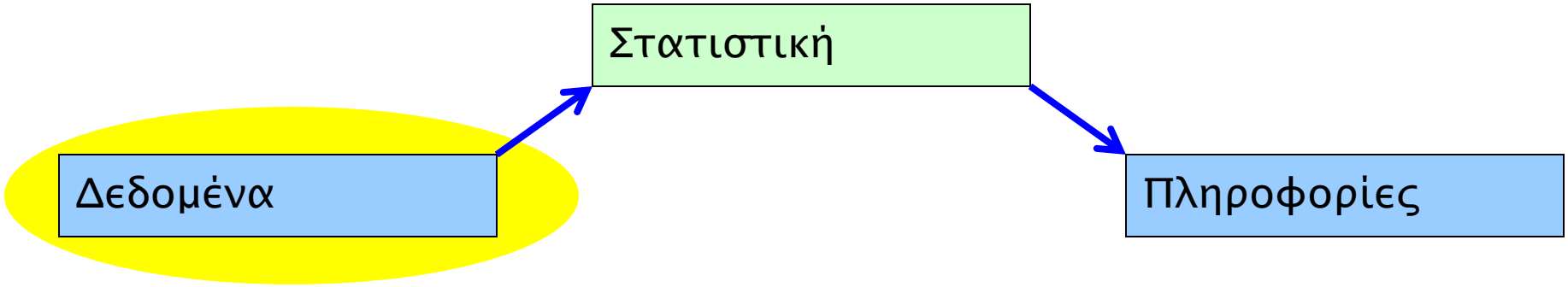
Statistics

VAR00004

N	Valid	10
	Missing	0
Percentiles	25	3,7500

Θυμηθείτε …

- Η Στατιστική είναι ένας μηχανισμός που από τα ***δεδομένα*** παράγει ***πληροφόρηση***:



Αλλά από πού τα ***δεδομένα*** έρχονται; Πως μαζεύονται; Πως εξασφαλίζεται η ορθότητα τους; Αντιπροσωπεύουν τον πληθυσμό από τον οποίο επιλέχθηκαν; Αυτό το κεφάλαιο εξετάζει κάποια από αυτά τα θέματα.

Μέθοδοι για την Συλλογή Δεδομένων …

- Υπάρχουν πολλοί μέθοδοι που χρησιμοποιούνται για την συλλογή ή την επίτευξη δεδομένων για στατιστικές αναλύσεις:
- • Άμεση παρατήρηση
- • Πειράματα, και
- • Δειγματοληψίες.

Δειγματοληψίες …

- Μία ***δειγματοληψία*** αναζητά πληροφόρηση από ανθρώπους, π.χ. σφυγμομέτρηση, προεκλογική σφυγμομέτρηση, έρευνες αγοράς.

- Το ***Ποσοστό Ανταπόκρισης*** (η αναλογία των ανθρώπων που συμμετέχει στην σφυγμομέτρηση) είναι πολύ βασική παράμετρος.

- Οι σφυγμομετρήσεις διαχειρίζονται με ποικίλους τρόπους, π.χ.
- Προσωπική Συνέντευξη,
- Τηλεφωνική Συνέντευξη, και
- Ερωτηματολόγια

Σχεδιασμός Ερωτηματολογίου …

- Με τον καιρό, μεγάλη επιστημονική μελέτη έχει γίνει για τον σχεδιασμό δειγματοληπτικών ερωτήσεων. Βασικές αρχές σχεδιασμού:

1. Κάνουμε το ερωτηματολόγιο όσο πιο **σύντομο** γίνεται.
2. Ρωτάμε σύντομα, απλά, και με **ξεκάθαρες ερωτήσεις**.
3. Αρχίζουμε με **δημογραφικές ερωτήσεις** ώστε να βοηθήσουμε τους ερωτηθέντες να εξοικειωθούν και να αισθάνονται άνετα.
4. Χρησιμοποιήστε **δυαδικές** (ναι/όχι) και ερωτήσεις **πολλαπλών επιλογών**.
5. Χρησιμοποιήστε ελεύθερες (που δεν έχουν ξεκάθαρη απάντηση, **open-ended**) ερωτήσεις με προσοχή.
6. Αποφεύγετε ερωτήσεις που πιέζουν ή καθοδηγούν (**leading**) τον συμμετέχοντα.
7. Δοκιμάστε ένα ερωτηματολόγιο σε ένα μικρό αριθμό ανθρώπων(**pilot**).
8. Σκεφτείτε με πιο τρόπο πρόκειται θα χρησιμοποιήσετε τα δεδομένα που θα επιλέξετε όταν ετοιμάζεται το ερωτηματολόγιο.

Παράδειγμα

8. Πόσο αλμυρή σας φαίνεται η γεύση της ελιάς;

	καθόλου	Πολύ λίγο	λίγο	μέτρια	Πάρα πολύ
Πικρή	1	2	3	4	5

9. Πόσο ξινή θεωρείτε τη γεύση της ελιάς;

	καθόλου	Πολύ λίγο	λίγο	μέτρια	Πάρα πολύ
Όξινη	1	2	3	4	5

10. Πόσο πικρή πιστεύετε ότι είναι η γεύση της συγκεκριμένης ελιάς;

	καθόλου	Πολύ λίγο	λίγο	μέτρια	Πάρα πολύ
Αλμυρή	1	2	3	4	5

Πως σας φαίνεται η γεύση της ελιάς;

	Καθόλου	Πολύ λίγο	Λίγο	Μέτρια	Πάρα πολύ
Πικρή	1	2	3	4	5
Όξινη	1	2	3	4	5
Αλμυρή	1	2	3	4	5

Δειγματοληψία…

- Θυμηθείτε ότι η επαγωγική στατιστική μας επιτρέπει να βγάλουμε συμπεράσματα σχετικά με τον πληθυσμό βασισμένοι στο δείγμα.

- Η δειγματοληψία (η επιλογή ενός υποσυνόλου του πληθυσμού) γίνεται συχνά για λόγους ***κόστους*** (κοστίζει λιγότερο να κάνεις δειγματοληψία με 1,000 τηλεθεατές από ότι με 100 εκατομμύρια τηλεθεατές) και ***πρακτικούς*** (π.χ. δεν γίνετε να εκτελέσουμε έναν έλεγχο σύγκρουσης για κάθε αυτοκίνητο ή ακόμα για πάρα πολλά αυτοκίνητα).

- Σε κάθε περίπτωση, η προσέγγιση είναι ικανοποιητή όταν ο **πληθυσμός-στόχος** (target population) συμπίπτει με τον **πληθυσμό προέλευσης του δείγματος** (sample population) (περίπτωση Literary Digest)
- **Δείγμα με αυτοεπιλογή** (self-selected sample) SLOP (Self-Selected Listener Opinion Poll).

Δειγματοληπτικό Διάγραμμα …

- Ένα ***δειγματοληπτικό διάγραμμα*** είναι μία μέθοδο ή διαδικασία για να καθορίσουμε πως ένα δείγμα θα επιλεχθεί από έναν πληθυσμό.

- Θα επικεντρώσουμε την προσοχή μας σε αυτές τις τρεις μεθόδους:

- Απλή Τυχαία Δειγματοληψία,
- Στρωματοποιημένη Τυχαία δειγματοληψία, και
- Κατά Συστοιχίες Δειγματοληψία.

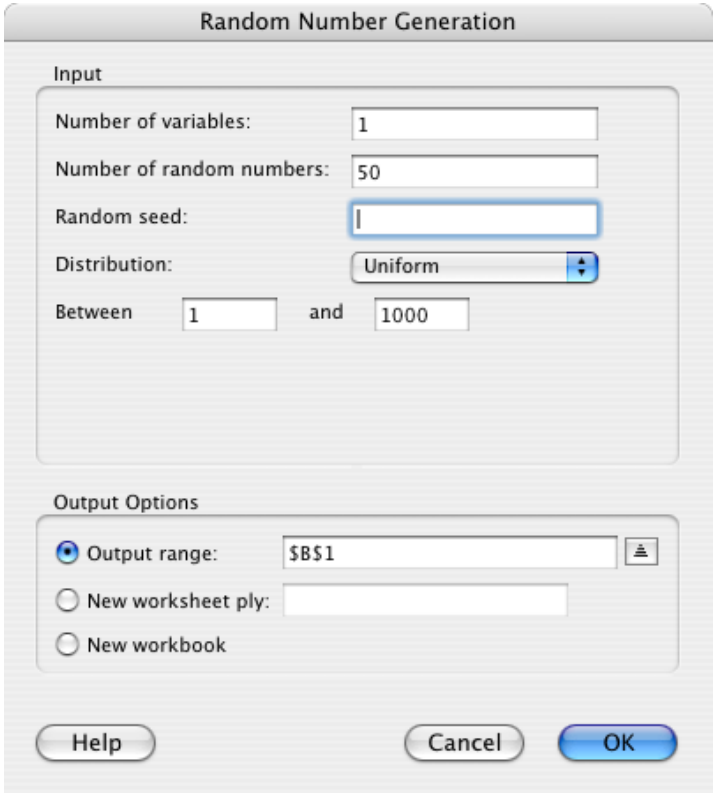
1. Απλή Τυχαία Δειγματοληψία

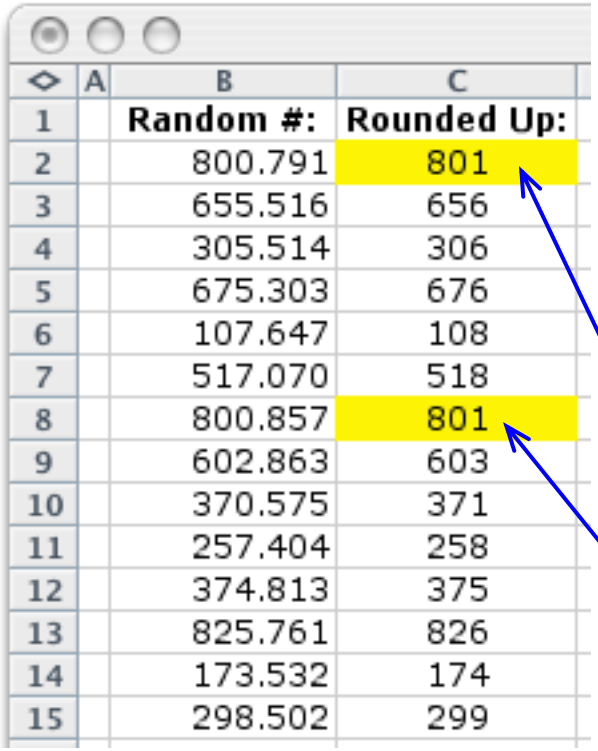
- Ένα ***απλό τυχαίο δείγμα*** είναι ένα δείγμα επιλεγμένο με τέτοιο τρόπο ώστε κάθε δυνατό δείγμα του ιδίου μεγέθους έχει την ίδια πιθανότητα να επιλεγεί.

- Το να επιλέξουμε τρία ονόματα από ένα καπέλο που περιέχει όλα τα ονόματα των φοιτητών της τάξης είναι ένα παράδειγμα απλού τυχαίου δείγματος: κάθε ομάδα τριών ατόμων έχει την ίδια πιθανότητα να επιλεγεί όπως και κάθε ομάδα τριών ατόμων.

1. Απλή Τυχαία Δειγματοληψία

- Παράδειγμα 1: Ένας επιθεωρητής έχει να επιλέξει ένα τυχαίο δείγμα 40 από 1,000 φορολογικές δηλώσεις προσώπων για εξέταση





	A	B	C
1		Random #:	Rounded Up:
2		800.791	801
3		655.516	656
4		305.514	306
5		675.303	676
6		107.647	108
7		517.070	518
8		800.857	801
9		602.863	603
10		370.575	371
11		257.404	258
12		374.813	375
13		825.761	826
14		173.532	174
15		298.502	299

Κάποιοι ακέραιοι μπορούν πολλαπλές φορές να επιλεχθούν.
Απλά επιλέγουμε επιπλέον εφεδρικούς αριθμούς.

2. Στρωματοποιημένη Τυχαία δειγματοληψία …

- Ένα ***στρωματοποιημένο τυχαίο δείγμα*** πετυχαίνεται αν χωρίσουμε τον πληθυσμό σε αμοιβαία αποκλειόμενα σύνολα, ή στρώμα, και μετά επιλέγουμε απλά τυχαία δείγματα από κάθε στρώμα.

Στρώμα 1 : Φύλλο	Στρώμα 2 : Ηλικία	Στρώμα 3 : Επάγγελμα
Αρσενικό	< 20	επαγγελματίας
Θηλυκό	20–30	υπάλληλος
	31–40	εργάτης
	41–50	λοιπά
	51–60	
	> 60	

Μπορούμε να θέλουμε σχετικά με τον συνολικό πληθυσμό, να βγάλουμε συμπεράσματα μέσα σε ένα στρώμα ή να βγάλουμε συμπεράσματα διασταυρώνοντας στρώματα.

2. Στρωματοποιημένη Τυχαία δειγματοληψία

- Αφού στρωματοποιήσουμε τον πληθυσμό, μπορούμε να χρησιμοποιήσουμε ***απλή τυχαία δειγματοληψία*** για να παράγουμε το πλήρες δείγμα:

Income Category	Population Proportion	Sample Size n = 400	n = 1000
under $25,000	25%	100	250
$25,000 - $39,999	40%	160	400
$40,000 – $60,000	30%	120	300
over $60,000	5%	20	50

Εάν μόνο έχουμε την δυνατότητα να επιλέξουμε 400 ανθρώπους συνολικά, Θα επιλέγαμε 100 από αυτούς από την ομάδα των χαμηλών εισοδημάτων …

…Εάν επιλέξουμε 1000 ανθρώπου, θα επιλέγαμε 50 από αυτούς από την κατηγορία με τα υψηλότερα εισοδήματα.

3. Κατά Συστάδες Δειγματοληψία

- Μία ***κατά συστάδες δειγματοληψία*** είναι ένα απλό τυχαίο δείγμα ομάδων ή συστοιχιών (σε αντίθεση με την απλή τυχαία δειγματοληψία από μεμονωμένα άτομα). Ο πληθυσμός χωρίζεται σε συστάδες, π.χ. με βάση γεωγραφικά κριτήρια, και απλά τυχαία δείγματα επιλέγονται από κάθε συστάδα

- Αυτή η μέθοδος είναι χρήσιμη όταν είναι δύσκολο ή κοστίζει να έχουμε μία πλήρης λίστα των μελών του πληθυσμού ή όταν τα στοιχεία του πληθυσμού είναι ευρέως διάσπαρτα γεωγραφικός.

- Η κατά συστάδες δειγματοληψία ενδέχεται να αυξήσει το λάθος του δείγματος οφειλόμενο στις ομοιότητες μεταξύ των μελών των ομάδων.

4. Πολυεπίπεδη δειγματοληψία

Αφορά την επιλογή δειγμάτων από δείγματα. Π.χ. για την αξιολόγηση του νέου συστήματος Εισαγωγής στην τριτοβάθμια εκπαίδευση, επιλέγουμε πρώτα ένα δείγμα από χαρακτηριστικούς Νομούς (γεωγραφικό κριτήριο) κατόπιν εκπαιδευτικές περιφέρειες δευτεροβάθμιας εκπαίδευσης για κάθε νομό, κατόπιν Λύκεια σε κάθε εκπαιδευτική περιφέρεια και τέλος μαθητές της Β και Γ Λυκείου από κάθε Λύκειο.

Μέγεθος Δείγματος …

- Πόσο μεγάλο πρέπει να είναι το μέγεθος του δείγματος;
  - Το μέγεθος εξαρτάται από την **μεταβλητότητα** της εξεταζόμενης μεταβλητής, όσο μικρότερη τόσο μικρότερο μέγεθος δείγματος χρειαζόμαστε.
  - Αν επιθυμούμε **μεγαλύτερη ακρίβεια (μικρότερα τυπικά σφάλματα)** στις εκτιμήσεις μας χρειαζόμαστε μεγαλύτερο μέγεθος δείγματος.
  - **Το είδος της Στατιστικής Ανάλυσης**, πιο πολύπλοκη στατιστική ανάλυση απαιτεί μεγαλύτερο δείγμα.

- Απαντώντας στα παραπάνω ερωτήματα, υπάρχουν αριθμητικές τεχνικές για να καθορίσουμε τα μεγέθη του δειγμάτων για όλες τις δειγματοληπτικές τεχνικές, π.χ. στρωματοποιημένη ή κατά συστάδες, και διδάσκονται σε μαθήματα για δειγματοληπτικές έρευνες (survey sampling). Γενικά όσο πιο μεγάλο το μέγεθος του δείγματος τόσο πιο ακριβή αναμένονται οι εκτιμητές του δείγματος να είναι.

Δειγματοληπτικά και Μη-Δειγματοληπτικά Λάθη…

- Δύο βασικοί τύποι λαθών μπορούν να εμφανισθούν όταν το δείγμα των παρατηρήσεων παίρνεται από έναν πληθυσμό:
- ***δειγματοληπτικό λάθος*** και ***μη-δειγματοληπτικό λάθος***.

- ***Το δειγματοληπτικό λάθος*** αναφέρεται σε διαφορές μεταξύ του δείγματος και του πληθυσμού οι οποίες υπάρχουν επειδή αυτές οι συγκεκριμένες παρατηρήσεις έτυχε να επιλεχθούν.

- Τα ***μη-δειγματοληπτικά λάθη*** είναι πιο σοβαρά και οφείλονται σε λάθη κατά την απόκτηση των δεδομένων ή οφείλεται στην ακατάλληλη επιλογή των δειγματοληπτικών παρατηρήσεων.

Δειγματοληπτικό λάθος …

- ***Το δειγματοληπτικό λάθος*** αναφέρεται σε διαφορές μεταξύ του δείγματος και του πληθυσμού οι οποίες υπάρχουν επειδή αυτές οι συγκεκριμένες παρατηρήσεις έτυχε να επιλεχθούν.

- Ένας άλλος τρόπος για να δούμε αυτό είναι: οι διαφορές αποτελεσμάτων για διαφορετικά δείγματα (ιδίου μεγέθους) οφειλόμενη καθαρά δειγματοληπτικό λάθος:

- Π.χ. Δύο δείγματα μεγέθους 10 από 1,000 νοικοκυριά. Εάν συνέβη να πάρουμε τα δεδομένα με το υψηλότερα εισοδήματα στο πρώτο μας δείγμα και όλα τα χαμηλότερα στο δεύτερο, αυτή η διαφορά οφείλεται καθαρά σε δειγματοληπτικό λάθος.

Διαφορετικά αποτελέσματα από δείγμα σε δείγμα

- Αύξηση του δείγματος, μειώνει το δειγματοληπτικό λάθος.

Μη-Δειγματοληπτικό λάθος …

- Τα ***μη-δειγματοληπτικά λάθη*** είναι πιο σοβαρά και οφείλονται σε λάθη κατά την απόκτηση των δεδομένων ή οφείλεται στην ακατάλληλη επιλογή των δειγματοληπτικών παρατηρήσεων.
- Τρεις τύποι μη-δειγματοληπτικών λαθών:

- Λάθη κατά την απόκτηση των δεδομένων,
- Λάθη από αναπάντητα ερωτηματολόγια, και
- Μεροληψία στην επιλογή του δείγματος.

Σημειώστε: Αύξηση του δείγματος, **δεν** μειώνει το μη-δειγματοληπτικό λάθος.

Λάθη κατά την Απόκτηση των Δεδομένων …

…εμφανίζονται από την καταγραφή λανθασμένων απαντήσεων, οφειλόμενη σε:

— λανθασμένες μετρήσεις παίρνονται εξαιτίας ελαττωματικού εξοπλισμού,
— λάθη γίνονται κατά την μεταγραφή από βασικές πηγές,
— λανθασμένη καταγραφή των δεδομένων οφειλόμενη σε παρερμηνεία των όρων, ή
— λανθασμένες απαντήσεις σε ερωτήσεις οφειλόμενες σε ευαίσθητα θέματα.

Λάθη από αναπάντητα ερωτηματολόγια …

- … αναφέρεται σε λάθος (ή ***μεροληψία***) που εισάγεται όταν απαντήσεις δεν πετυχαίνονται από κάποια μέλη του δείγματος, π.χ. οι παρατηρήσεις του δείγματος που συλλέγονται ενδεχομένως δεν αντιπροσωπεύουν τον πληθυσμό τον οποίο εξετάζουμε.

- Όπως αναφέρθηκε νωρίτερα, το ***Ποσοστό Ανταπόκρισης*** (η αναλογία των ανθρώπων που συμμετέχει στην σφυγμομέτρηση) είναι πολύ βασική παράμετρος και βοηθάει στην κατανόηση της εγκυρότητας της σφυγμομέτρησης και στην κατανόηση πηγών με λάθη από αναπάντητα ερωτηματολόγια

Μεροληψία στην Επιλογή του Δείγματος …

- … συμβαίνει όταν το δειγματοληπτικό σχέδιο είναι τέτοιο που κάποια μέλη του πληθυσμού δεν μπορούν να επιλεχθούν και δεν συμπεριλαμβάνονται μέσα στο δείγμα.