Μεθοδολογία της Έρευνας

Περιγραφική Στατιστική, Εκτιμητική, Έλεγχοι υποθέσεων

Παντελής Μπάγκος
Καθηγητής
Πανεπιστήμιο Θεσσαλίας, 2020

Τι είναι η Στατιστική;

Η Στατιστική είναι ένα σύνολο αρχών και μεθοδολογιών για:

- το σχεδιασμό της διαδικασίας συλλογής δεδομένων (σχεδιασμός πειραμάτων)
- τη συνοπτική και αποτελεσματική παρουσίασή τους (περιγραφική στατιστική)
- την εξαγωγή αντίστοιχων συμπερασμάτων (στατιστική συμπερασματολογία).

Αντικείμενο της Στατιστικής

Η στατιστική ασχολείται με:

- Το σχεδιασμό της διαδικασίας συλλογής πληροφοριών
- Τη συλλογή πληροφοριών από το σύνολο του πληθυσμού (απογραφή) ή από επιλεγμένο δείγμα (ομοιογενές σύνολο ατόμων) του πληθυσμού
- Την οργάνωση των πληροφοριών
- Τη συνοπτική και αποτελεσματική παρουσίασή τους
- Την ανάλυση και εξαγωγή συμπερασμάτων

Οι πληροφορίες για ένα χαρακτηριστικό ονομάζονται **παρατηρήσεις ή δεδομένα**

Βασικοί στατιστικοί όροι

- **Ολική Απογραφή**: Εξέταση όλων των ατόμων του πληθυσμού για το χαρακτηριστικό που μας ενδιαφέρει (σύνολο των βαθμών των μαθητών μιας τάξης, τα εισοδήματα των κατοίκων μιας πόλης και λοιπά)

- **Δείγμα**: Όταν η ολική απογραφή είναι δύσκολη, αδύνατη ή οικονομικά ασύμφορη, ο ερευνητής μαζεύει πληροφορίες από κάποια μικρή ομάδα ή υποσύνολο του πληθυσμού ώστε τα αποτελέσματα που προκύπτουν από την εξέταση της ομάδας αυτής να είναι αντιπροσωπευτικά για τον μελετώμενο πληθυσμό.

Βασικές αρχές δειγματοληψίας

- **Δειγματοληψία**: Οι αρχές και οι μέθοδοι για τη συλλογή και ανάλυση δεδομένων από πεπερασμένους πληθυσμούς.

- **Δειγματοληπτικές μονάδες**: Μη επικαλυπτόμενες συλλογές απλών στοιχείων του πληθυσμού (για παράδειγμα αν ο πληθυσμός μας είναι το σύνολο των κατοίκων μιας περιοχής, οι δειγματοληπτικές μονάδες μπορεί να είναι τα νοικοκυριά ή τα διαμερίσματα).

- **Δειγματοληπτικό πλαίσιο:** Το σύνολο των δειγματοληπτικών μονάδων που αντιστοιχούν σε ένα πληθυσμό. Το δειγματοληπτικό πλαίσιο είναι συνήθως ένας κατάλογος ονομάτων φυστικών προσώπων, αντικειμένων, οικοδομικό σχέδιο πόλεως, μία αεροφωτογραφία ενός νησιού.

Σφάλματα δειγματοληψίας

- Δειγματοληπτικά σφάλματα:

α) Ακατάλληλη μέθοδος δειγματοληψίας

β) Ακατάλληλο δείγμα όσον αφορά τη μεταβλητότητα και το μέγεθος του δείγματος

- Μη δειγματοληπτικά σφάλματα:

α) **Μη συστηματικά** (ή σφάλματα τυχαίου τύπου) των οποίων οι επιδράσεις προσεγγιστικά αλληλοαναιρούνται όταν χρησιμοποιούνται αρκετά μεγάλου μεγέθους δείγματα

β) **Συστηματικά** (ή μεροληψίες) τα οποία οδηγούν τις εκτιμήσεις μας πάντοτε προς τα πάνω ή προς τα κάτω από την πραγματική τιμή και γι αυτό δεν μπορούν να αλληλοεξουδετερωθούν

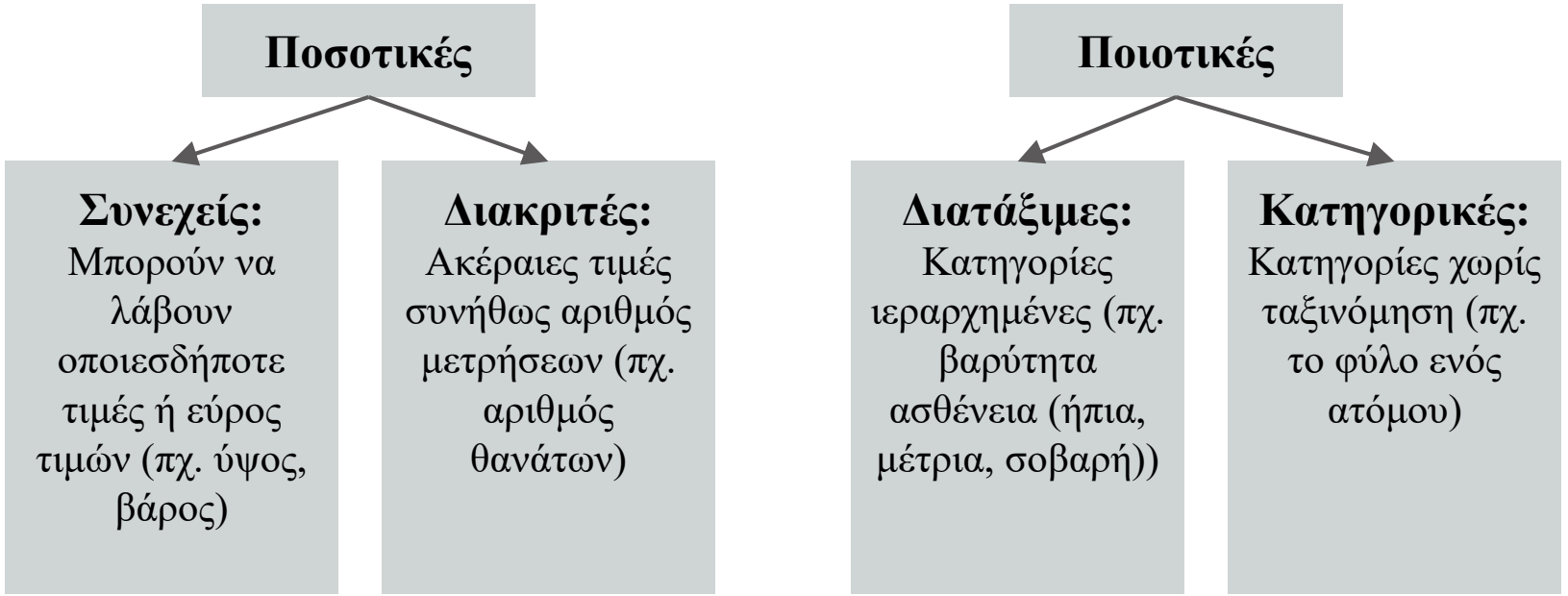
Παραμετρική και μη παραμετρική στατιστική

- **Παραμετρικές μέθοδοι**: Ο αναλυτής υποθέτει ότι η μορφή της κατανομής του πληθυσμού είναι γνωστή και αναζητά μεθόδους για τον προσδιορισμό (εκτίμηση) των αγνώστων παραμέτρων της (για παράδειγμα του μέσου και της διασποράς).

- **Απαραμετρική συμπερασματολογία**: Οι μέθοδοι της περιοχής αυτής μπορούν να εφαρμοστούν ανεξάρτητα από την κατανομή που ακολουθεί ο αρχικός πληθυσμός και για το λόγο αυτό λέγονται ελεύθερες κατανομών.

Τύποι δεδομένων

Έστω λοιπόν ένας πληθυσμός στα άτομα του οποίου καταγράφουμε τις τιμές που παίρνει ένα (ή περισσότερα) συγκεκριμένο χαρακτηριστικό (π.χ. το μηνιαίο εισόδημα, χρώμα ματιών, ύψος, ηλικία κ.λ.π.). Έτσι έχουμε μία τυχαία μεταβλητή X και αν από τον πληθυσμό θεωρήσουμε ένα τυχαίο δείγμα μεγέθους ν θα πάρουμε ν ανεξάρτητες και ισόνομες τυχαίες μεταβλητές X_1, X_2 ,..., X_v . Οι τυχαίες μεταβλητές διακρίνονται ανάλογα με το είδος των τιμών που μπορούν να πάρουν σε **ποσοτικές** και **ποιοτικές**.



Περιγραφική στατιστική

Πίνακες συχνοτήτων

Έστω X μία τυχαία μεταβλητή (χαρακτηριστικό) που αφορά τα άτομα ενός πληθυσμού και X_1, X_2 ,..., X_v ένα τυχαίο δείγμα μεγέθους v. Για ένα συγκεκριμένο δείγμα θα συμβολίζουμε με x_1, x_2 ,..., x_v τις τιμές του χαρακτηριστικού για τα ν άτομα του δείγματος και με y_1, y_2 ,..., y_k ($k \leq v$) τις k διαφορετικές μεταξύ τους τιμές από τα x_1, x_2 ,..., x_v.

Συχνότητα (frequency) v_i της τιμής y_i θα λέγεται το πλήθος των x_1, x_2 ,.., x_v που είναι ίσα με y_i, ενώ **σχετική συχνότητα** (relative frequency) f_i θα λέγεται το αντίστοιχο ποσοστό, δηλαδή:

$$f_i = \frac{v_i}{v} = \frac{v_i}{\sum_{j=1}^{k} v_j}, \quad i = 1, 2, \ldots, k .$$

Συνήθως οι ποσότητες y_i , v_i , f_i [illegible] δείγμα συγκεντρώνονται σε ένα συνοπτικό πίνακα που ονομάζεται **πίνακας συχνοτήτων**.

Παράδειγμα 1: Σε ένα δείγμα 20 οικογενειών από μία περιοχή της Αθήνας, το επάγγελμα του πατέρα, ο μηνιαίος μισθός του πατέρα και ο αριθμός παιδιών της οικογένειας δίνονται στον πιο κάτω Πίνακα.

Δεδομένα ενός δείγματος 20 οικογενειών.

Οικογένεια *i*	Επάγγελμα Πατέρα	Μηνιαίος Μισθός πατέρα	Αριθμ. παιδιών Οικογένειας
1	εργάτης	700	0
2	οδηγός	750	1
3	εργάτης	800	0
4	δημ. υπάλληλος	700	2
5	δημ. υπάλληλος	800	2
6	δημ. υπάλληλος	500	2
7	δάσκαλος	900	3
8	ιερέας	1000	2
9	οδηγός	600	4
10	εργάτης	600	1
11	δάσκαλος	700	1
12	εργάτης	600	2
13	εργάτης	800	3
14	δημ. υπάλληλος	700	4
15	ιερέας	900	1
16	δάσκαλος	1000	2
17	εργάτης	900	2
18	δημ. υπάλληλος	650	2
19	δάσκαλος	750	2
20	δημ. υπάλληλος	800	2

- *Πίνακας συχνοτήτων για το επάγγελμα πατέρα*

1	Εργάτης	IIIII I	6	0.3
2	οδηγός	II	2	0.1
3	δημ. υπάλληλος	IIIII I	6	0.3
4	δάσκαλος	IIII	4	0.2
5	ιερέας	II	2	0.1
		Σύνολο	20	1.0

- *Πίνακας συχνοτήτων για το Μηνιαίο μισθό*

1	50	I	1	0.05
2	60	III	3	0.15
3	65	I	1	0.05
4	70	IIII	4	0.20
5	75	II	2	0.10
6	80	IIII	4	0.20
7	90	III	3	0.15
8	100	II	2	0.10
		Σύνολο	20	1.00

- *Πίνακας συχνοτήτων για τον αριθμό παιδιών*

1	0	II	2	0.1
2	1	IIII	4	0.2
3	2	IIIII II	10	0.5
4	3	III	2	0.1
5	4	II	2	0.1
		II		
		Σύνολο	20	1.0

Αθροιστικές συχνότητες - Αθροιστικές σχετικές συχνότητες

Στην περίπτωση ποσοτικών τυχαίων μεταβλητών εκτός των ποσοτήτων ν_i, f_i χρησιμοποιούνται συνήθως και οι λεγόμενες **αθροιστικές συχνότητες** (cumulative frequencies) N_i , καθώς και οι **αθροιστικές σχετικές συχνότητες** (cumulative relative frequencies) F_i οι οποίες δίνουν το πλήθος και το ποσοστό αντίστοιχα των παρατηρήσεων που είναι μικρότερες ή ίσες του y_i. Αν τα $y_1, y_2,..., y_k$ είναι διατεταγμένα κατά αύξουσα σειρά μεγέθους δηλ. $y_1 \leq y_2 \leq ... \leq y_k$ είναι φανερό ότι:

$$N_i = \nu_1 + \nu_2 + ... + \nu_i \ , \ i = 1, 2,..., k \ ,$$
$$F_i = f_1 + f_2 + ... + f_i \ , \ i = 1, 2,..., k \ ,$$
$$\nu_1 = N_1 \ , \ \nu_i = N_i - N_{i-1}, \ i = 2,3,..., k \ ,$$
$$f_1 = F_1 \ , \ f_i = F_i - F_{i-1} \ , \ i = 2,3,..., k \ .$$

Παράδειγμα 1 (συνέχεια)

- *Πίνακας συχνοτήτων και αθρ. συχνοτήτων για το Μισθό*

i	y_i	v_i	f_i	N_i	F_i
1	50	1	0.05	1	0.05
2	60	3	0.15	4	0.20
3	65	1	0.05	5	0.25
4	70	4	0.20	9	0.45
5	75	2	0.10	11	0.55
6	80	4	0.20	15	0.75
7	90	3	0.15	18	0.90
8	100	2	0.10	20	1.00
		20	1.00		

- *Πίνακας συχνοτήτων και αθρ. συχνοτήτων για αριθμό παιδιών*

1	0	2	0.1	2	0.1
2	1	4	0.2	6	0.3
3	2	10	0.5	16	0.8
4	3	2	0.1	18	0.9
5	4	2	0.1	20	1.0
		20	1.0		

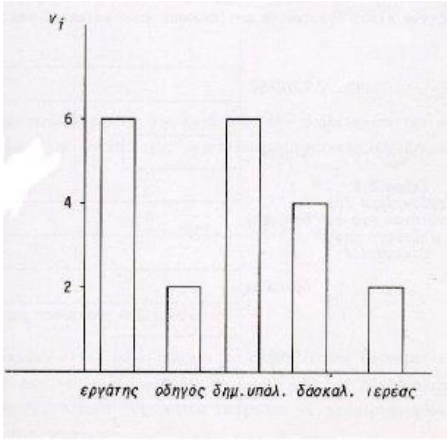
Γραφικές μέθοδοι παρουσίασης στατιστικών δεδομένων (1)

Παρουσίαση ποιοτικών δεδομένων: Για τη γραφική παράσταση ποιοτικών δεδομένων χρησιμοποιούνται κυρίως δύο είδη διαγραμμάτων: το **ραβδόγραμμα** (barchart) και το **κυκλικό διάγραμμα συχνοτήτων** (piechart).

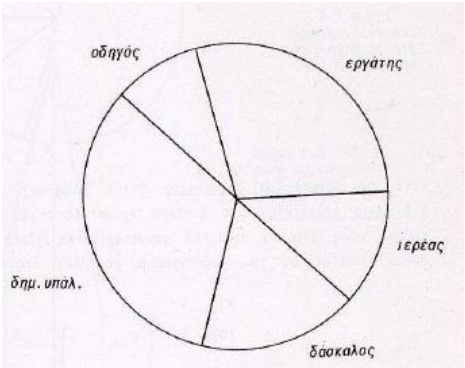
α. Στο **ραβδόγραμμα**, οι κατηγορίες της τυχαίας μεταβλητής παριστάνονται στον οριζόντιο άξονα σαν ισομήκη διαστήματα (με κενά συνήθως μεταξύ τους) ενώ οι αντίστοιχες συχνότητες ή σχετικές συχνότητες στον κατακόρυφο.

β. Τα **κυκλικά διαγράμματα** χρησιμοποιούν για την παράσταση των δεδομένων ένα κύκλο χωρισμένο σε κυκλικά τμήματα. Κάθε κυκλικό τμήμα αναφέρεται σε μία κατηγορία του χαρακτηριστικού και έχει τόξο αi ανάλογο της αντίστοιχης συχνότητας ή σχετικής συχνότητας, δηλαδή:

$$\alpha_i = v_i \frac{360^\circ}{v} = 360 f_i, \qquad i = 1, 2, \ldots, k.$$



Σχήμα 1: Ραβδόγραμμα Συχνοτήτων για τα δεδομένα του Παραδείγματος 1.



Σχήμα 2: Κυκλικό διάγραμμα συχνοτήτων για το επάγγελμα (Παράδειγμα 1).

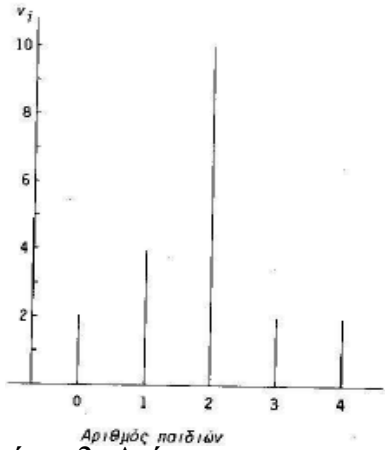
Γραφικές μέθοδοι παρουσίασης στατιστικών δεδομένων (2)

Παρουσίαση ποσοτικών δεδομένων: Όταν τα δεδομένα είναι ποσοτικά και το πλήθος k των διαφορετικών τιμών που πήραμε από το δείγμα είναι μικρό τότε αφού γίνει η πινακοποίηση των δεδομένων σε ένα πίνακα συχνοτήτων μπορούμε να χρησιμοποιήσουμε για την γραφική τους παράσταση είτε ένα **διάγραμμα συχνοτήτων** (line diagram) είτε ένα **κυκλικό διάγραμμα συχνοτήτων**.

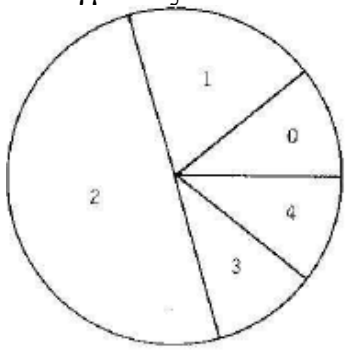
α. Το **διάγραμμα συχνοτήτων** μοιάζει με το ραβδόγραμμα με μόνη διαφορά ότι αντί να χρησιμοποιούμε συμπαγή ορθογώνια, υψώνουμε σε κάθε y_i μία κάθετη γραμμή με μήκος ίσο προς την αντίστοιχη συχνότητα ή σχετική συχνότητα.

β. Το **κυκλικό διάγραμμα συχνοτήτων** σχηματίζεται με τον ίδιο ακριβώς τρόπο, όπως για τα ποιοτικά χαρακτηριστικά.

c. Για μικρά σύνολα δεδομένων, μπορεί κανείς να χρησιμοποιήσει και το λεγόμενο **σημειόγραμμα** (dot diagram) στο οποίο οι παρατηρήσεις παριστάνονται με τελείες στις αντίστοιχες θέσεις ενός οριζόντιου άξονα.



Σχήμα 3: Διάγραμμα συχνοτήτων για τον αριθμό παιδιών του Παραδείγματος 1.

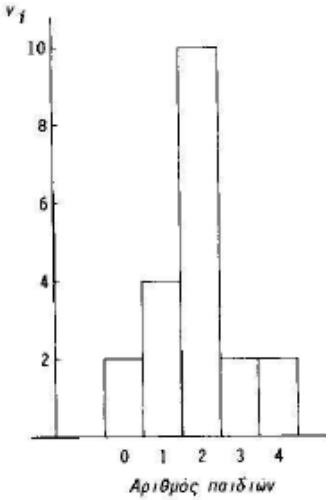


Σχήμα 4: Κυκλικό διάγραμμα συχνοτήτων για τον αριθμό παιδιών του Παραδείγματος 1.

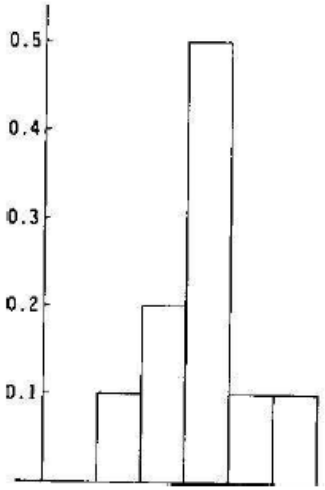
Γραφικές μέθοδοι παρουσίασης στατιστικών δεδομένων (3)

δ. Το πιο συνηθισμένο μέσο περιγραφής ποσοτικών δεδομένων είναι το **ιστόγραμμα** (histogram).

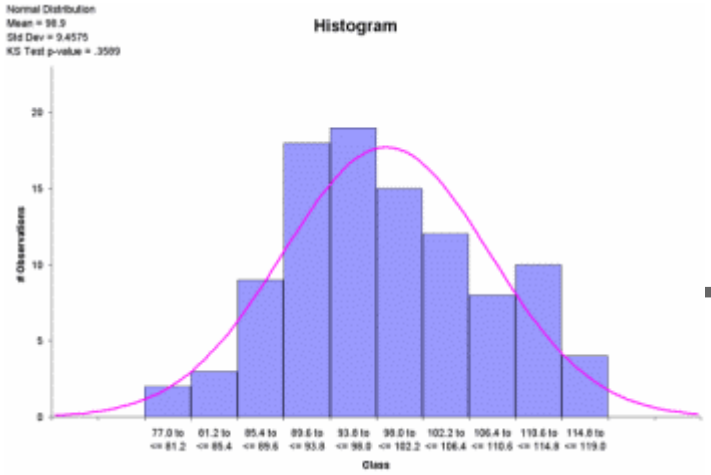
❖Αυτό αποτελείται από διαδοχικά ορθογώνια των οποίων το ύψος διαλέγεται με τέτοιο τρόπο ώστε το εμβαδόν του ορθογωνίου να είναι ίσο με την αντίστοιχη συχνότητα ή σχετική συχνότητα της τιμής στην οποία αναφέρεται.
❖Για διακριτά δεδομένα, ως άκρα των βάσεων των ορθογωνίων διαλέγονται συνήθως τα μεσαία σημεία μεταξύ των διαδοχικών y_i.
❖Αξίζει να σημειωθεί ότι λόγω του τρόπου σχηματισμού του ιστογράμματος συχνοτήτων, το συνολικό εμβαδόν όλων των ορθογωνίων είναι ίσο με το μέγεθος του δείγματος v.
Με παρόμοιο τρόπο σχηματίζεται το **ιστόγραμμα σχετικών συχνοτήτων**, με συνολικό εμβαδόν ίσο με 1.

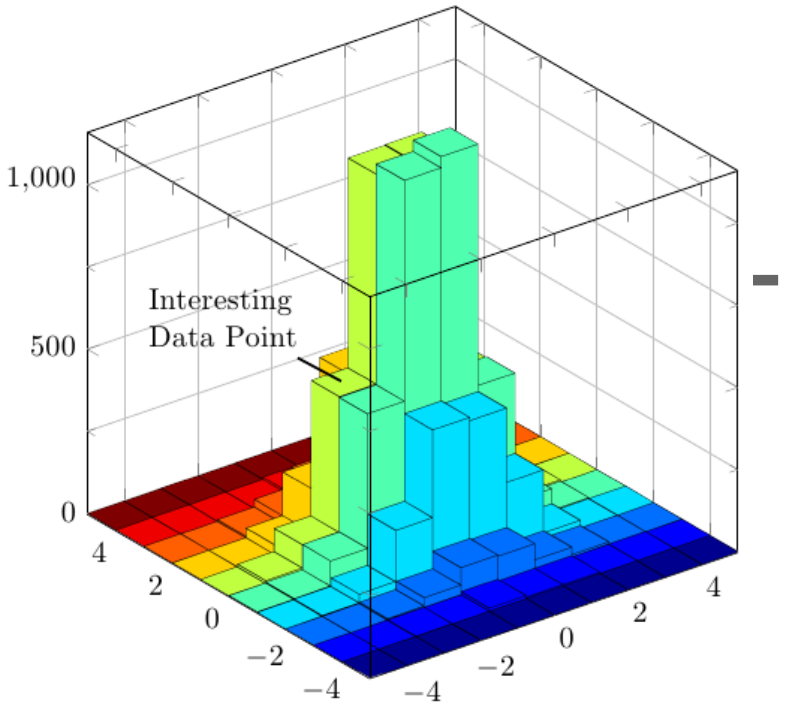


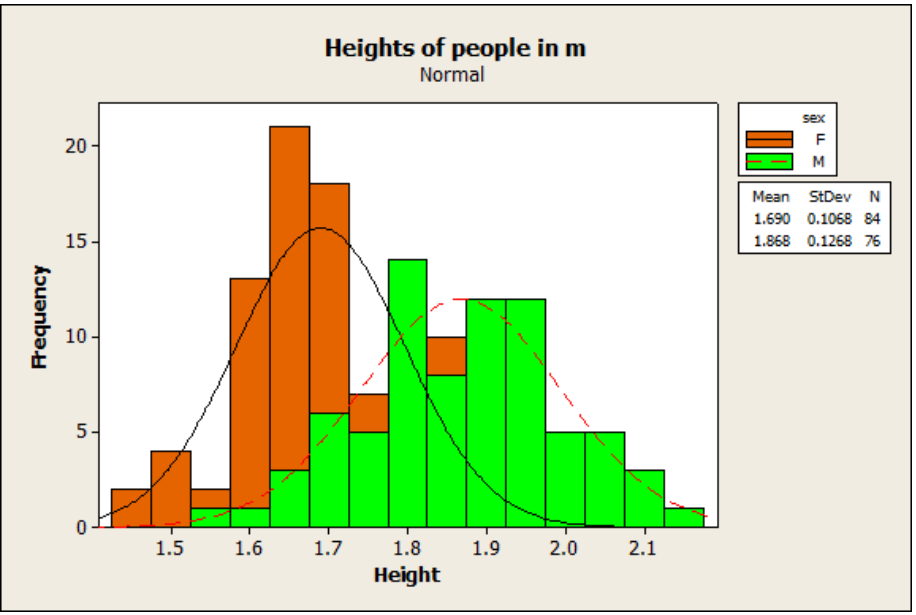
Σχήμα 5: Ιστόγραμμα Συχνοτήτων για τον αριθμό παιδιών (Παράδειγμα 1).

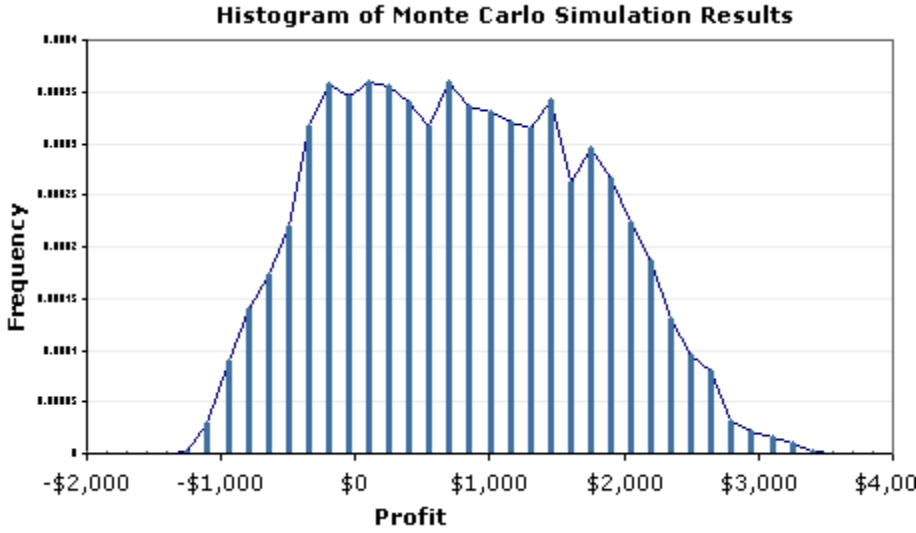


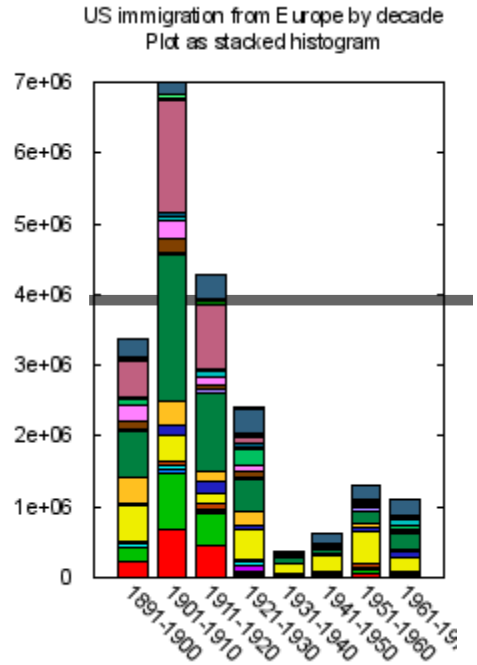
Σχήμα 6: Ιστόγραμμα Σχετικών Συχνοτήτων για τον αριθμό παιδιών (Παράδειγμα 1).

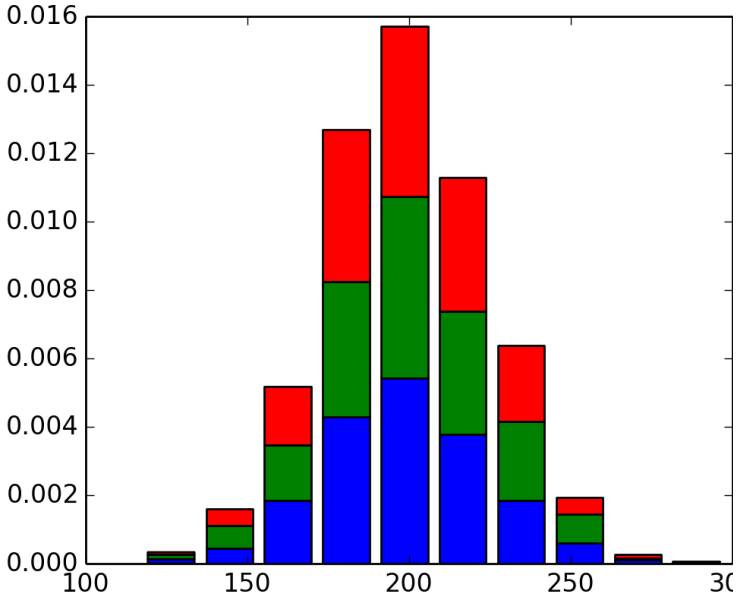

Normal Distribution
Mean = 98.9
Std Dev = 9.4575
KS Test p-value = .3589
Histogram
Observations
Class

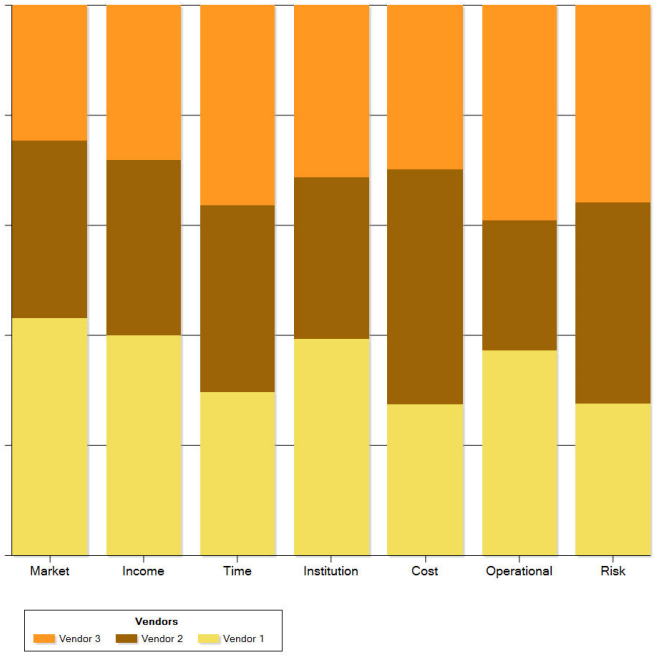

1,000
500
0
Interesting
Data Point
4
2
0
−2
−4

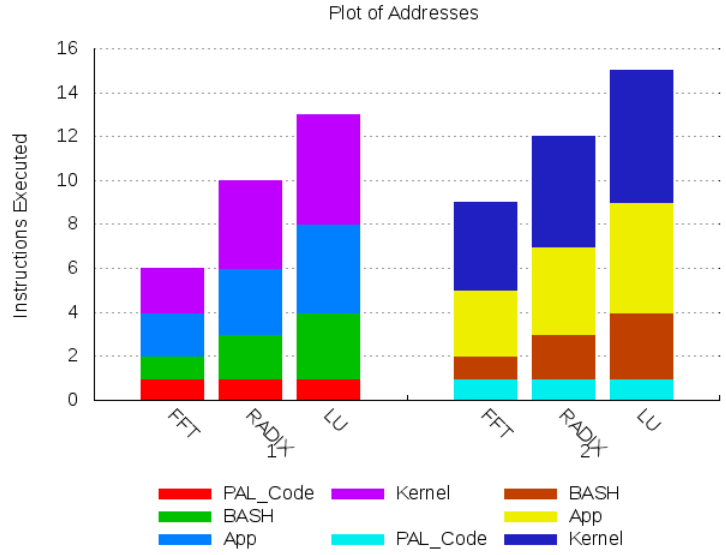

Heights of people in m
Normal
sex
F
M
Mean StDev N
1.690 0.1068 84
1.868 0.1268 76
Frequency
Height

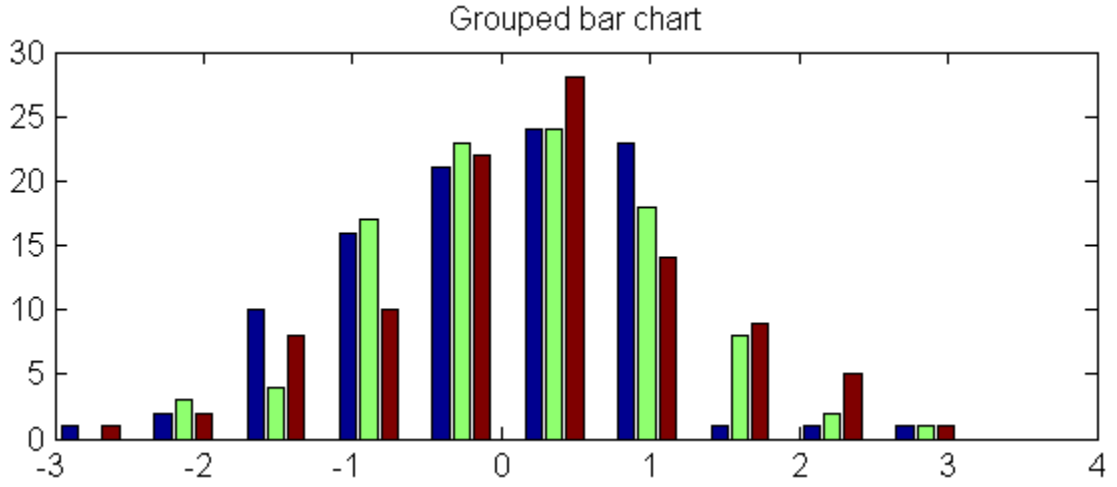

Histogram of Monte Carlo Simulation Results
Frequency
-$2,000
-$1,000
$0
$1,000
$2,000
$3,000
$4,00
Profit

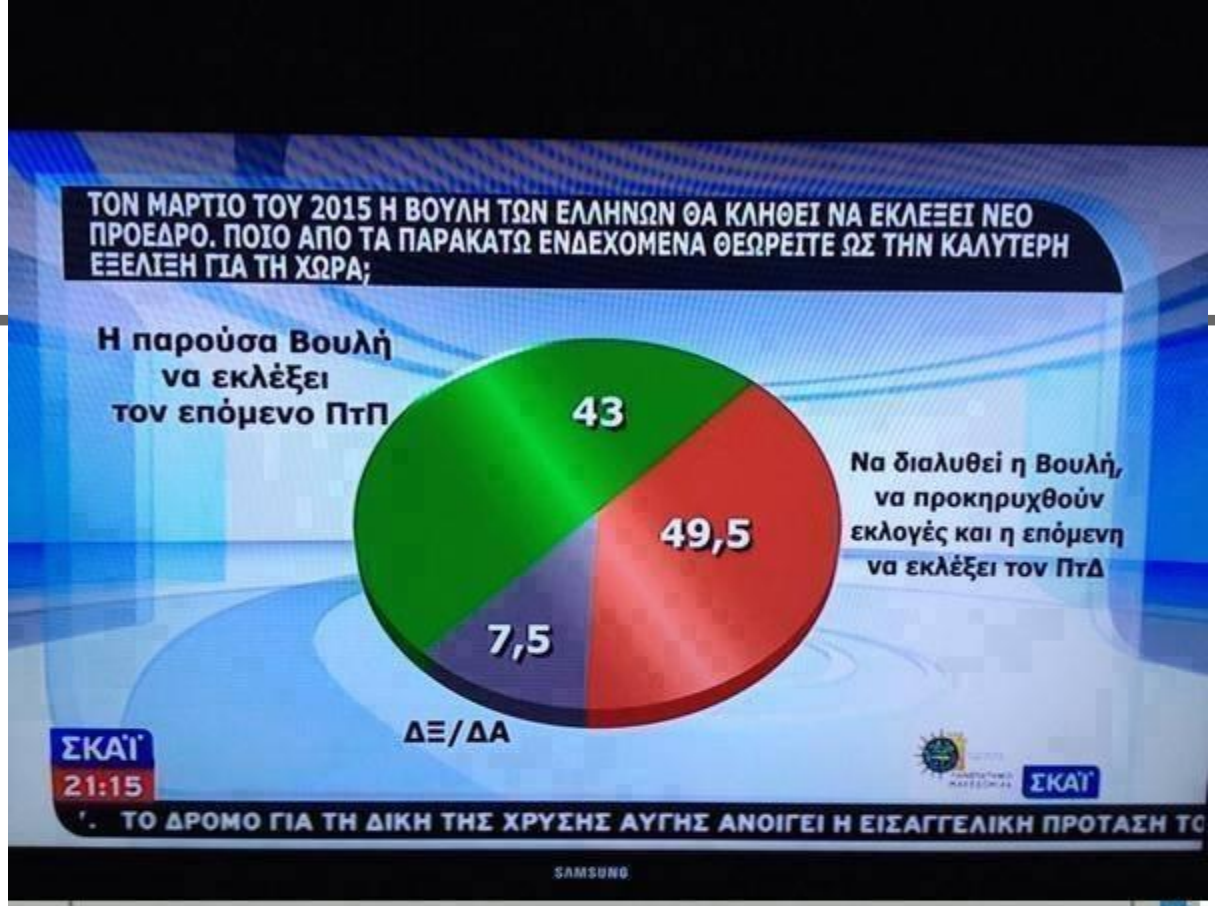

US immigration from Europe by decade
Plot as stacked histogram
7e+06
6e+06
5e+06
4e+06
3e+06
2e+06
1e+06
0
1891-1900
1901-1910
1911-1920
1921-1930
1931-1940
1941-1950
1951-1960


0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
100
150
200
250


Market
Income
Time
Institution
Cost
Operational
Risk
Vendors
Vendor 3
Vendor 2
Vendor 1


Plot of Addresses
Instructions Executed
16
14
12
10
8
6
4
2
0
FFT
RADIX 1
LU
FFT
RADIX 2
LU
PAL_Code
BASH
App
Kernel
PAL_Code
BASH
App
Kernel


Grouped bar chart
30
25
20
15
10
5
0
-3
-2
-1
0
1
2
3
4


ΤΟΝ ΜΑΡΤΙΟ ΤΟΥ 2015 Η ΒΟΥΛΗ ΤΩΝ ΕΛΛΗΝΩΝ ΘΑ ΚΛΗΘΕΙ ΝΑ ΕΚΛΕΞΕΙ ΝΕΟ ΠΡΟΕΔΡΟ. ΠΟΙΟ ΑΠΟ ΤΑ ΠΑΡΑΚΑΤΩ ΕΝΔΕΧΟΜΕΝΑ ΘΕΩΡΕΙΤΕ ΩΣ ΤΗΝ ΚΑΛΥΤΕΡΗ ΕΞΕΛΙΞΗ ΓΙΑ ΤΗ ΧΩΡΑ;
Η παρούσα Βουλή να εκλέξει τον επόμενο ΠτΠ
43
49,5
7,5
ΔΞ/ΔΑ
Να διαλυθεί η Βουλή, να προκηρυχθούν εκλογές και η επόμενη να εκλέξει τον ΠτΔ
ΣΚΑΪ
21:15
ΣΚΑΪ
ΤΟ ΔΡΟΜΟ ΓΙΑ ΤΗ ΔΙΚΗ ΤΗΣ ΧΡΥΣΗΣ ΑΥΓΗΣ ΑΝΟΙΓΕΙ Η ΕΙΣΑΓΓΕΛΙΚΗ ΠΡΟΤΑΣΗ ΤΟ
SAMSUNG


ΣΚΑΪ
SPREAD 10ΕΤΟΥΣ ΚΡΑΤΙΚΟΥ ΟΜΟΛΟΓΟΥ
10%
5%
0
6,60%
6,70%
7,01%
7,85%
8,96%
8,94%
10/10
13/10
14/10
15/10
16/10
17/10


το γράφημα των
ΣΚΑΪ
ξεφτίλες
38,3%
18%
20,4%
14%
6,3%
5,1%
4,5%
4,4%
3,5%
10,6%
ΣΥΡΙΖΑ
ΝΔ
ΤΟ ΠΟΤΑΜΙ
ΚΚΕ
ΑΝ.ΕΛ.
Χ.Α
ΠΑΣΟΚ
ΑΔΙΕΥΚΡΙΝΙΣΤΗ ΨΗΦΟΣ


ΠΡΟΘΕΣΗ ΨΗΦΟΥ
το κανονικό γράφημα
38.3 %
20.4 %
6.3 %
5.1 %
4.5 %
4.4 %
3.5 %
ΝΔ
ΚΚΕ
ΠΑΣΟΚ

Ομαδοποίηση δεδομένων (1)

- Οι μέθοδοι παρουσίασης ποσοτικών δεδομένων που αναφέρθηκαν παραπάνω μπορούν να χρησιμοποιηθούν στην πράξη μόνο όταν ο αριθμός των διαφορετικών παρατηρήσεων είναι σχετικά μικρός.

- Στην αντίθετη περίπτωση είναι απαραίτητο να ταξινομηθούν τα δεδομένα σε μικρό πλήθος ομάδων και να θεωρούνται όμοιες όλες οι παρατηρήσεις που ανήκουν στην ίδια ομάδα.

- Έτσι μπορούμε να πάρουμε τις συχνότητες (απόλυτες ή σχετικές) και αθροιστικές συχνότητες των διαφόρων ομάδων και να προχωρήσουμε σε πινακοποίηση και γραφική παράσταση των δεδομένων.

Ομαδοποίηση δεδομένων (2)

- Πρώτα επιλέγουμε τον αριθμό q των ομάδων ή διαστημάτων ή κλάσεων.
- Ο αριθμός αυτός συνήθως ορίζεται αυθαίρετα από τον ερευνητή σύμφωνα με την πείρα του.
- Υπάρχει όμως και ένας τύπος που μπορεί να χρησιμοποιηθεί ως οδηγός. Αυτός είναι γνωστός ως τύπος του Sturges και ορίζεται ως εξής:

$$q = 1 + 3.32 \log_{10} v$$

όπου q είναι ο αριθμός των κλάσεων και v το μέγεθος του δείγματος.

- Το δεύτερο βήμα είναι ο προσδιορισμός του πλάτους των κλάσεων (ίδιο για όλες τις κλάσεις). Το πλάτος (c) υπολογίζεται διαιρώντας το εύρος (R) του δείγματος δια του αριθμού των διαστημάτων. Δηλαδή: $c = \frac{R}{q}$

όπου το εύρος $R=\max\{x_i , i=1,2,...,v\}- \min\{x_i , i=1, 2,...,v\}$ ορίζεται ως η διαφορά της μικρότερης παρατήρησης από την μεγαλύτερη.

- Το τρίτο βήμα είναι ο καθορισμός των διαστημάτων. Το πρώτο διάστημα διαλέγεται συνήθως έτσι ώστε να περιέχει τη μικρότερη παρατήρηση και το τελευταίο να περιέχει τη μεγαλύτερη.

Παράδειγμα 2

Η συγκέντρωση (σε μgr / cm^3) ενός συγκεκριμένου ρύπου σε δείγματα αέρος που πάρθηκαν από 57 πόλεις των ΗΠΑ δίνεται από τον πίνακα:

Συγκέντρωση ($\mu gr / cm^3$) ρύπου στον αέρα 57 πόλεων των ΗΠΑ.

68	63	42	27	30	36	28	32	79	27
22	23	24	25	24	65	43	25	74	51
36	42	28	31	28	25	45	12	57	51
12	32	49	38	42	27	31	50	38	21
16	24	69	47	23	22	43	27	49	48
23	12	19	46	30	49	49			

Πηγή: *Statistical Abstract of the United States* 1970, *σελ.* 174.

Από τα δεδομένα βρίσκουμε για τον αριθμό των κλάσεων:

ενώ το εύρος των παρατηρήσεων είναι:

Άρα:

$$q = 1 + 3.32 \log_{10} 57 = 1 + 3.32 \cdot 1.76 = 6.83 \cong 7$$

$$R = 79 - 12 = 67 .$$

$$c = \frac{R}{q} = \frac{67}{7} = 9.6 \cong 10 .$$

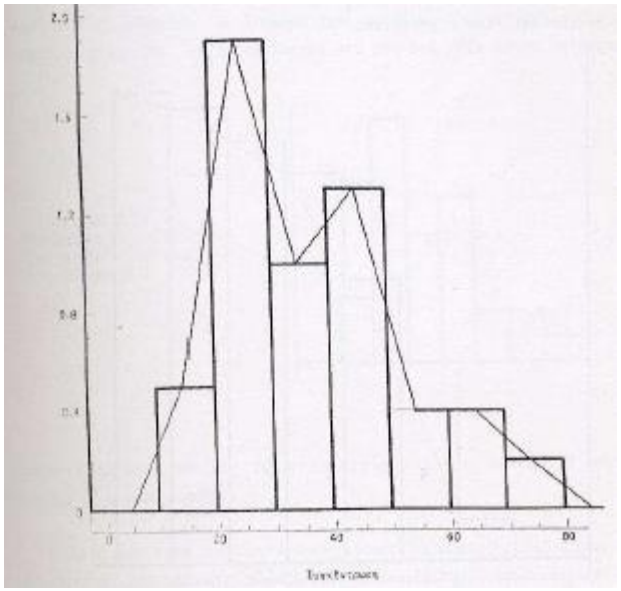
Παράδειγμα 2 (συνέχεια)

Θεωρούμε σαν αρχή του πρώτου διαστήματος το 9.5 (οπότε καμία παρατήρηση δεν πέφτει σε άκρο διαστήματος) θα έχουμε τον επόμενο πίνακα:

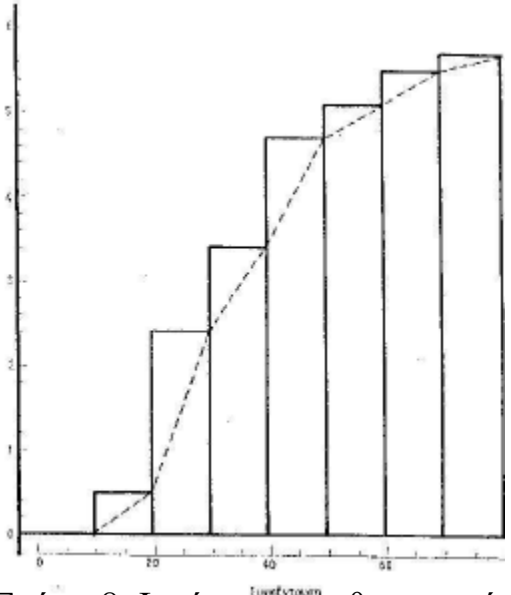
i	Κάτω όριο	Άνω όριο	Κέντρο y_i	v_i	Σχετική Συχνότ.	Αθροιστ Συχνότ	Αρθ. Σχετ. Συχνότητα
1	9.50	19.50	14.50	5	.0877	5	.0877
2	19.50	29.50	24.50	19	.3333	24	.4211
3	29.50	39.50	34.50	10	.1754	34	.5965
4	39.50	49.50	44.50	13	.2281	47	.8246
5	49.50	59.50	54.50	4	.0702	51	.8947
6	59.50	69.50	64.50	4	.0702	55	.9649
7	69.50	79.50	74.50	2	.0351	57	1.0000

Παράδειγμα 2 (συνέχεια)

Για το ιστόγραμμα συχνοτήτων, κατασκευάζουμε ορθογώνια παραλληλόγραμμα που έχουν βάσεις τα διαστήματα των κλάσεων και ύψος τέτοιο, ώστε το εμβαδόν κάθε ορθογωνίου να ισούται με την συχνότητα των παρατηρήσεων στην αντίστοιχη κλάση. Ενώνοντας τα μέσα των άνω βάσεων των ορθογωνίων παραλληλογράμμων (και προσθέτοντας δύο ακόμη υποθετικές κλάσεις με συχνότητα μηδέν δεξιά και αριστερά των πραγματικών κλάσεων) σχηματίζουμε το πολύγωνο συχνοτήτων.

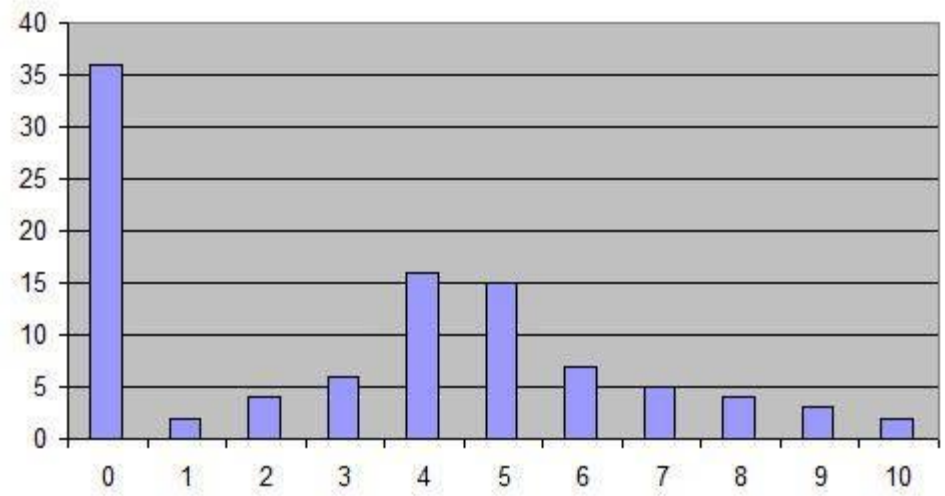


Σχήμα 7: Ιστόγραμμα συχνοτήτων (και πολύγωνο συχνοτήτων)



Σχήμα 8: Ιστόγραμμα αθροιστικών συχνοτήτων και αθροιστικό διάγραμμα (ogive plot)

- Πόσο συμμετρικά είναι τα δεδομένα;
- Πόσο διεσπαρμένα;
- Υπάρχουν διαστήματα με υψηλή συγκέντρωση δεδομένων;
- Υπάρχουν κενά;
- Υπάρχουν παρατηρήσεις μακριά από τις υπόλοιπες (ακραίες παρατηρήσεις - outliers);



http://en.wikipedia.org/wiki/Zero-inflated_model

Αριθμητικά περιγραφικά μέτρα

Διακρίνονται κυρίως σε δύο βασικές κατηγορίες:

• τα μέτρα θέσης ή κεντρικής τάσης (location measures, central tendency measures)

• τα μέτρα διασποράς ή μεταβλητότητας (measures of variability, measures of variance, dispersion measures).

Μέτρα θέσης ή κεντρικής τάσης (1)

Τα μέτρα κεντρικής τάσης είναι χρήσιμα για την περιγραφή της θέσης της κατανομής από την οποία προέρχονται τα δεδομένα. Θα ορίσουμε αρχικά τα μέτρα της κατηγορίας αυτής για την περίπτωση μη ομαδοποιημένων δεδομένων δηλαδή όταν διαθέτουμε τις πρωτογενείς παρατηρήσεις x_1, x_2 ,..., x_v ή ισοδύναμα τις διαφορετικές μεταξύ τους παρατηρήσεις y_1, y_2 ,..., y_k και τις αντίστοιχες συχνότητες.

1. **Μέση Τιμή** (mean, mean value) ή δειγματική μέση τιμή (sample mean) λέγεται το άθροισμα των τιμών των παρατηρήσεων του δείγματος δια του πλήθους των παρατηρήσεων δηλαδή:

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

Όταν χρησιμοποιούμε πίνακα συχνοτήτων, η μέση τιμή προκύπτει από τις ισοδύναμες εκφράσεις:

$$\bar{x} = \frac{\sum_{i=1}^{k} v_i y_i}{\sum_{i=1}^{k} v_i} = \sum_{i=1}^{k} f_i y_i$$

Μέτρα θέσης ή κεντρικής τάσης (2)

Ο δειγματικός μέσος χρησιμοποιείται ευρύτατα ως αριθμητικό περιγραφικό μέτρο αφού είναι πολύ απλός στον υπολογισμό και για ένα σύνολο δεδομένων καθορίζεται μονοσήμαντα. Έχει όμως τα μειονεκτήματα να επηρεάζεται από:

1. πιθανές ακραίες τιμές

2. να μην αντιστοιχεί πάντοτε σε "λογική" τιμή της τυχαίας μεταβλητής που εξετάζουμε

3. δεν μπορεί να χρησιμοποιηθεί για την περιγραφή ποιοτικών χαρακτηριστικών.

Μέτρα θέσης ή κεντρικής τάσης (3)

2. **Κορυφή** (mode) ή επικρατούσα τιμή M_0 ενός συνόλου παρατηρήσεων ορίζεται η παρατήρηση με τη μεγαλύτερη συχνότητα.

3. **Διάμεσος** (median) δ ενός δείγματος είναι η τιμή που χωρίζει το δείγμα σε δύο ίσα μέρη έτσι ώστε ο αριθμός των παρατηρήσεων που είναι μικρότερες ή ίσες από το δ να είναι ίσος με τον αριθμό των παρατηρήσεων που είναι μεγαλύτερες ή ίσες από το δ. Έτσι αν διατάξουμε τις ν παρατηρήσεις x_1, x_2 ,..., x_v και συμβολίσουμε με $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(v)}$ το αντίστοιχο διατεταγμένο δείγμα, τότε η διάμεσος δ ορίζεται από τη σχέση:

$$\delta = \begin{cases} x_{(r)} & \text{αν} \quad v = 2r-1 \\ \dfrac{x_{(r)} + x_{(r+1)}}{2} & \text{αν} \quad v = 2r. \end{cases}$$

4. **Ποσοστημόρια** (quantiles): Το α-στο ποσοστημόριο p_α ($0 < \alpha < 1$) ενός συνόλου παρατηρήσεων είναι η τιμή για την οποία το α100% των παρατηρήσεων είναι μικρότερες ή ίσες του p_α και (1− α)100% μεγαλύτερες ή ίσες του p_α. Ιδιαίτερο ενδιαφέρον παρουσιάζουν επίσης τα τεταρτημόρια (quartiles) που αντιστοιχούν σε α = 0.25, 0.50,0.75. Το $p_{0.25}$ συμβολίζεται με Q_1 και λέγεται πρώτο τεταρτημόριο ενώ το $p_{0.75}$ με Q_3 και λέγεται τρίτο τεταρτημόριο. Το δεύτερο τεταρτημόριο $p_{0.50}$ συμπίπτει με τη διάμεσο δ των παρατηρήσεων.

Παράδειγμα 3

Αν τα βάρη (σε kgr) 10 κοτόπουλων ενός ορνιθοτροφείου ήταν 2, 4, 4, 3, 4, 3, 3, 3, 6, 3 η μέση τιμή του δείγματος θα είναι $\bar{x}$ = 35/10 = 3.5. Στον παρακάτω Πίνακα φαίνεται ο τρόπος υπολογισμού του δειγματικού μέσου με χρήση πίνακα συχνοτήτων:

i	y_i	v_i	$v_i y_i$
1	2	1	2
2	3	5	15
3	4	3	12
4	6	1	6
		10	35

Από τον δεδομένα είναι φανερό ότι M_0 =3 .

Το διατεταγμένο δείγμα είναι 2,3,3,3,3,3,4,4,4,6. Οπότε, αφού ν =10 = 2 × 5 (για r=5), έχουμε:

$$\delta = \frac{x_{(5)} + x_{(6)}}{2} = 3$$

Παράδειγμα 4

Για τις παρατηρήσεις 1,5,3,3,6,4,3,2 (ν=8), το Q_1 θα πρέπει να αφήνει 2 παρατηρήσεις του διατεταγμένου δείγματος αριστερά και 6 δεξιά του.

Επομένως θα πρέπει να πάρουμε Q_1= (2+3) / 2 =2.5 . Όμοια Q_3= (4+ 5) / 2= 4.5.

Μέτρα θέσης ή κεντρικής τάσης (4)

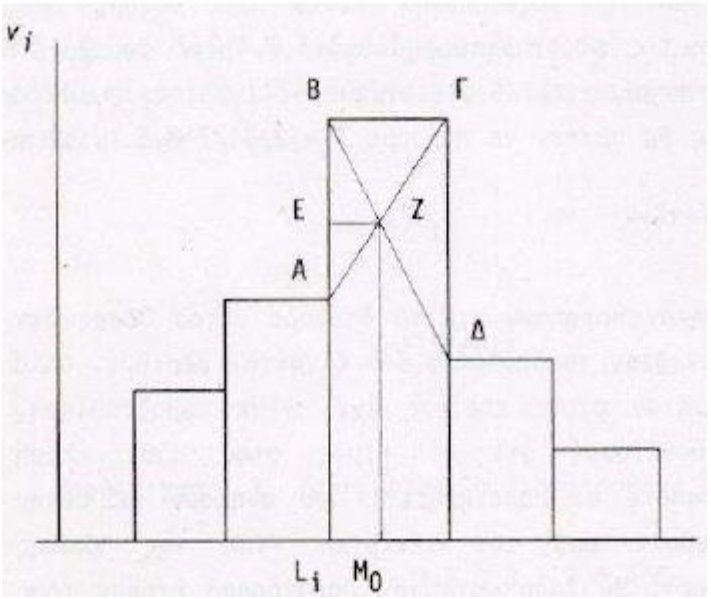
Οι ορισμοί που δόθηκαν παραπάνω για τα διάφορα μέτρα θέσης δεν μπορούν να χρησιμοποιηθούν όταν τα δεδομένα δεν δίνονται ακριβώς, αλλά υπό μορφή πινάκων συχνοτήτων στους οποίους έχει γίνει ομαδοποίηση. Στην περίπτωση αυτή υποθέτουμε ότι οι τιμές στην κάθε κλάση κατανέμονται ομοιόμορφα οπότε οι παρατηρήσεις που ανήκουν σε αυτήν μπορούν να αντιπροσωπευθούν από την κεντρική τιμή της κλάσης (ημιάθροισμα των άκρων της). Με βάση αυτή την παρατήρηση έχουμε τους επόμενους τύπους για τα τέσσερα μέτρα θέσης.

1. **Μέση τιμή**. Αυτή γράφεται στη μορφή:

$$\bar{x} = \frac{1}{v}\sum_{i=1}^{k} v_i y_i = \sum_{i=1}^{k} f_i y_i$$

Μέτρα θέσης ή κεντρικής τάσης (5)

2. **Κορυφή**. Στα ομαδοποιημένα δεδομένα, επειδή οι αρχικές παρατηρήσεις δεν είναι διαθέσιμες δεν μπορούμε να καθορίσουμε την παρατήρηση με τη μεγαλύτερη συχνότητα. Αντί αυτής λοιπόν θεωρούμε την επικρατούσα κλάση, δηλαδή την ομάδα με τη μεγαλύτερη συχνότητα και υπολογίζουμε γραφικά τη κορυφή M_0 από το ιστόγραμμα όπως στο σχήμα:



Σχήμα 9: Γραφικός προσδιορισμός της κορυφής.

Από το σχήμα είναι φανερό ότι:

$$M_0=L_i+EZ$$

και αν συμβολίσουμε με

c: το πλάτος των κλάσεων

$\Delta_1=\nu_i-\nu_{i-1}$ (διαφορά μεταξύ της μεγαλύτερης συχνότητας και της συχνότητας της προηγούμενης κλάσης)

$\Delta_1=\nu_i-\nu_{i+1}$(διαφορά μεταξύ της μεγαλύτερης συχνότητας και της συχνότητας της επόμενης κλάσης)

θα έχουμε: $AB=\Delta_1,\quad \Gamma\Delta=\Delta_2,\quad B\Gamma=c$.

Επομένως: $EZ=\dfrac{AB}{AB+|\Gamma\Delta}\,B\Gamma=\dfrac{\Delta_1}{\Delta_1+\Delta_2}c$

και η κορυφή M_0 θα δίνεται από τον τύπο:

$$M_0=L_i+\frac{\Delta_1}{\Delta_1+\Delta_2}c$$

Μέτρα θέσης ή κεντρικής τάσης (6)

3. **Διάμεσος**. Αρχικά υπολογίζουμε τη μεσαία κλάση δηλαδή το διάστημα στο οποίο ανήκει η διατεταγμένη παρατήρηση με σειρά $(\nu+1)/2$ (αν το ν είναι άρτιος μας ενδιαφέρουν οι παρατηρήσεις με σειρά $\nu/2$ και $(\nu+1)/2$) και ας συμβολίσουμε με L_i το κάτω όριό της. Ο γραφικός υπολογισμός της διαμέσου δ βασίζεται στο ιστόγραμμα αθροιστικών συχνοτήτων και γίνεται ως εξής: Από το μέσο του τμήματος ΟΗ φέρνουμε παράλληλη με τον άξονα των παρατηρήσεων και από το σημείο όπου αυτή συναντά το αθροιστικό διάγραμμα φέρνουμε παράλληλη με τον άξονα των συχνοτήτων. Το σημείο τομής της τελευταίας με τον οριζόντιο άξονα είναι η διάμεσος δ των παρατηρήσεων. Από το σχήμα είναι φανερό ότι $\delta=L_i+EZ$ και αν συμβολίσουμε

c: το πλάτος των κλάσεων

ν_i: τη συχνότητα της κλάσης με κάτω όριο L_i

$N_{i-1}=\nu_1+\nu_2+\ldots+\nu_{i-1}$ (αθροιστική συχνότητα της κλάσης με άνω όριο το L_i) θα έχουμε:

$$AB=\frac{\nu_i}{c},\quad AE=\frac{\nu}{2c}-\frac{N_{i-1}}{c},\quad B\Gamma=c\,.$$

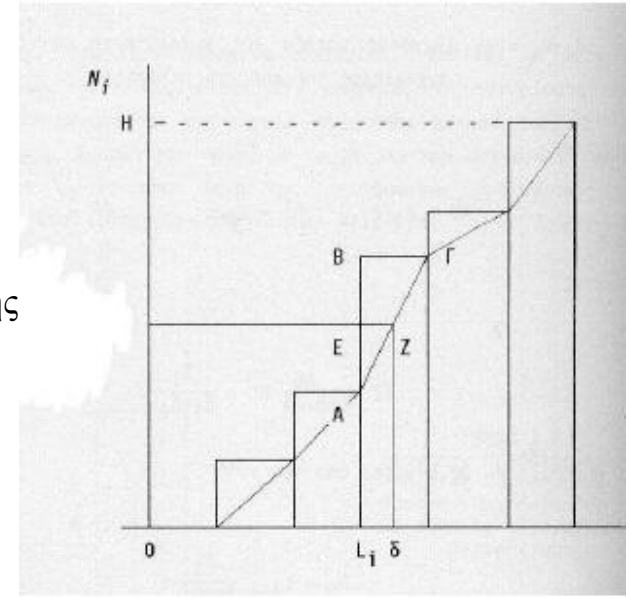
Μέτρα θέσης ή κεντρικής τάσης (7)

Επομένως:

$$EZ = \frac{AE}{AB}\, B\Gamma = \frac{\frac{v}{2} - N_{i-1}}{v_i} \cdot c$$

και η διάμεσος δ θα δίνεται από τον τύπο:

$$\delta = L_i + \frac{\frac{v}{2} - N_{i-1}}{v_i} \cdot c\ .$$



Σχήμα 9: Γραφικός προσδιορισμός της διαμέσου.

Μέτρα θέσης ή κεντρικής τάσης (8)

4. **Ποσοστημόρια**. Δουλεύοντας όπως και στη διάμεσο μπορούμε να δείξουμε ότι το α-στο ποσοστημόριο p_α δίνεται από τον τύπο:

$$p_\alpha = L_i + \frac{\alpha\, v - N_{i-1}}{v_i} \cdot c$$

όπου:

c: το πλάτος των κλάσεων

L_i: το κάτω όριο της κλάσης που περιέχει την διατεταγμένη παρατήρηση με σειρά [α ν]

v_i: η συχνότητα της κλάσης με κάτω όριο το L_i

$N_{i-1} = v_1 + v_2 + \ldots + v_{i-1}$ (αθροιστική συχνότητα της κλάσης με άνω όριο το L_i)

Ειδικά για το πρώτο ($\alpha = 0.25$) και τρίτο ($\alpha = 0.75$) τεταρτημόριο έχουμε τους τύπους:

$$Q_1 = L_i + \frac{\frac{v}{4} - N_{i-1}}{v_i} \cdot c,$$

$$Q_3 = L_i + \frac{\frac{3v}{4} - N_{i-1}}{v_i} \cdot c$$

Παράδειγμα 5

Η βαθμολογία των 28 μαθητών μιας τάξης σε ένα τεστ δίνεται στον επόμενο πίνακα:

Βαθμολογία 28 μαθητών μιας τάξης σε ένα τεστ.

15	22	11	8	10	11	11
11	9	12	11	14	10	10
11	11	12	15	9	6	8
11	7	16	9	10	17	11

από όπου μπορούμε εύκολα να διαπιστώσουμε ότι:

Επίσης

$$M_0 = 11, \quad \delta = 11, \quad Q_1 = (9+10)/2 = 9.5, \quad Q_3 = 12.$$

$$\bar{x} = \frac{\sum_{i=1}^{28} x_i}{28} = \frac{318}{28} = 11.357$$

Παράδειγμα 5 (συνέχεια)

Ομαδοποιώντας τα δεδομένα σε: $q = 1 + 3.32 \log_{10} 28 = 5.8 \cong 6$
ομάδες παίρνουμε τον επόμενο πίνακα:

i	Κάτω όριο	Άνω όριο	Κεντρική Τιμή y_i	Συχνότητα v_i	$v_i y_i$	Αθροιστ. Συχνότητ N_i
1	5.5	8.5	7	4	28	4
2	8.5	11.5	10	16	160	20
3	11.5	14.5	13	3	39	23
4	14.5	17.5	16	4	64	27
5	17.5	20.5	19	0	0	27
6	20.5	23.5	22	1	22	28
				28	313	-

$$\bar{x} = \frac{1}{v}\sum_{i=1}^{\kappa} v_i y_i = \frac{313}{28} = 11.178.$$

για την διάμεσο έχουμε

$$L_2 = 8.5, \quad \Delta_1 = 16 - 4 = 12, \quad \Delta_2 = 16 - 3 = 13$$

$$M_0 = 8.5 + \frac{12}{12+13} \cdot 3 = 9.94$$

$$L_2 = 8.5, \quad v_2 = 16, \quad N_1 = 4 \quad \delta = 8.5 + \frac{14-4}{16} \cdot 3 = 10.375$$

Παράδειγμα 5 (συνέχεια)

Για το πρώτο τεταρτημόριο είναι:

$$L_2 = 8.5, \quad v_2 = 16, \quad M_2 = 4 \qquad Q_1 = 8.5 + \frac{7-4}{16} \cdot 3 = 9.06$$

Για το τρίτο τεταρτημόριο είναι:

$$L_3 = 11.5, \quad v_3 = 3, \quad N_2 = v_1 + v_2 = 20 \qquad Q_3 = 11.5 + \frac{21-20}{3} \cdot 3 = 12.5$$

Μέτρα διασποράς ή μεταβλητότητας (1)

Παράλληλα λοιπόν με τα μέτρα θέσης κρίνεται απαραίτητη και η εξέταση κάποιων μέτρων μεταβλητότητας, δηλαδή μέτρων που εκφράζουν τις αποκλίσεις των τιμών μίας μεταβλητής γύρω από τα μέτρα κεντρικής τάσης. Τέτοια μέτρα λέγονται **μέτρα διασποράς** ή **μεταβλητότητας** (measures of variability, measures of variance, dispersion measures) και τα περισσότερο συνηθισμένα από αυτά είναι τα επόμενα:

1. **Εύρος–Κύμανση**: Το απλούστερο από τα μέτρα διασποράς είναι το εύρος (Range) R που ορίζεται ως η διαφορά της ελάχιστης παρατήρησης από τη μέγιστη παρατήρηση.

2. **Ενδοτεταρτημοριακή** και **Ημιενδοτεταρτημοριακή απόκλιση**: Η ενδοτεταρτημοριακή απόκλιση ή ενδοτεταρτημοριακό εύρος (interquantile deviation, interquantile range) είναι η διαφορά του πρώτου τεταρτημορίου Q_1 από το τρίτο τεταρτημόριο Q_3 .

α. Στο μεταξύ τους διάστημα περιλαμβάνεται το 50% των τιμών του δείγματος.

β. Επομένως όσο μικρότερο θα είναι αυτό το διάστημα, τόσο μεγαλύτερη θα είναι η συγκέντρωση των τιμών και άρα μικρότερη η διασπορά των τιμών.

γ. Το μισό της διαφοράς $Q3 - Q1$ είναι το ημιενδοτεταρτημοριακό εύρος ή απόκλιση (semi-interquantile deviation, semi-interquantile range) και συμβολίζεται με Q

Μέτρα διασποράς ή μεταβλητότητας (2)

3. **Διασπορά** ή **διακύμανση**. Το πιο διαδεδομένο μέτρο διασποράς είναι η δειγματική διασπορά ή διακύμανση (variance) που ορίζεται από τη σχέση:

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

Αυτή ισοδύναμα γράφεται στη μορφή:

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

Στις περιπτώσεις δεδομένων που δίνονται με τη μορφή πινάκων συχνοτήτων η διασπορά μπορεί να υπολογισθεί από τον τύπο:

ή ισοδύναμα,

$$s^2 = \frac{1}{v-1}\sum_{i=1}^{k}v_i(y_i - \bar{x})^2$$

$$s^2 = \frac{1}{v-1}\left[\sum_{i=1}^{k}v_i y_i^2 - \frac{1}{v}\left(\sum_{i=1}^{k}v_i y_i\right)^2\right] = \frac{1}{v-1}\left[\sum_{i=1}^{k}v_i y_i^2 - v(\bar{x})^2\right]$$

Μέτρα διασποράς ή μεταβλητότητας (3)

4. **Τυπική απόκλιση**. Η τετραγωνική ρίζα της διασποράς είναι η τυπική απόκλιση (standard deviation) και συμβολίζεται με s:

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

Όταν τα δεδομένα δίνονται σε μορφή πινάκων συχνοτήτων η τυπική απόκλιση θα δίνεται από τη σχέση:

$$s = \sqrt{\frac{1}{v-1}\left[\sum_{i=1}^{k} v_i y_i^2 - \frac{1}{v}\left(\sum_{i=1}^{k} v_i y_i\right)^2\right]},$$

Παράδειγμα 6

Σε δύο δείγματα 8 οικογενειών είχαμε τον εξής αριθμό παιδιών:

i	1	2	3	4	5	6	7	8
Δείγμα Ι	1	1	3	1	3	10	3	2
Δείγμα ΙΙ	2	1	6	1	6	3	10	3

Δειγμα Ι

$$R = 10 - 1 = 9,$$

$$s^2 = 61/7 = 8.71,$$

$$Q = (Q_3 - Q_1)/2 = (3-1)/2 = 1,$$

$$s = 2.95.$$

Υπολογισμός των μέτρων διασποράς για το δείγμα Ι.

i	y_i	v_i	$v_i y_i$	$\lvert y_i - \bar{x} \rvert$	$v_i \lvert y_i - \bar{x} \rvert$	$(y_i - \bar{x})^2$	$v_i (y_i - \bar{x})^2$
1	1	3	3	2	6	4	12
2	2	1	2	1	1	1	1
3	3	3	9	0	0	0	0
4	10	1	10	7	7	49	49
		8	24		14		61

Δειγμα ΙΙ

$$R = 10 - 1 = 9,$$

$$s^2 = 68/7 = 9.71,$$

$$Q = (Q_3 - Q_1)/2 = (6 - 1 \cdot 5)/2 = 2.$$

$$s = 3 \cdot 12.$$

Υπολογισμός των μέτρων διασποράς για το δείγμα ΙΙ.

i	y_i	v_i	$v_i y_i$	$\lvert y_i - \bar{x} \rvert$	$v_i \lvert y_i - \bar{x} \rvert$	$(y_i - \bar{x})^2$	$v_i (y_i - \bar{x})^2$
1	1	2	2	3	6	9	18
2	2	1	2	2	2	4	4
3	3	2	6	1	2	1	2
4	6	2	12	2	4	4	8
5	10	1	10	6	6	36	36
		8	32		20		68

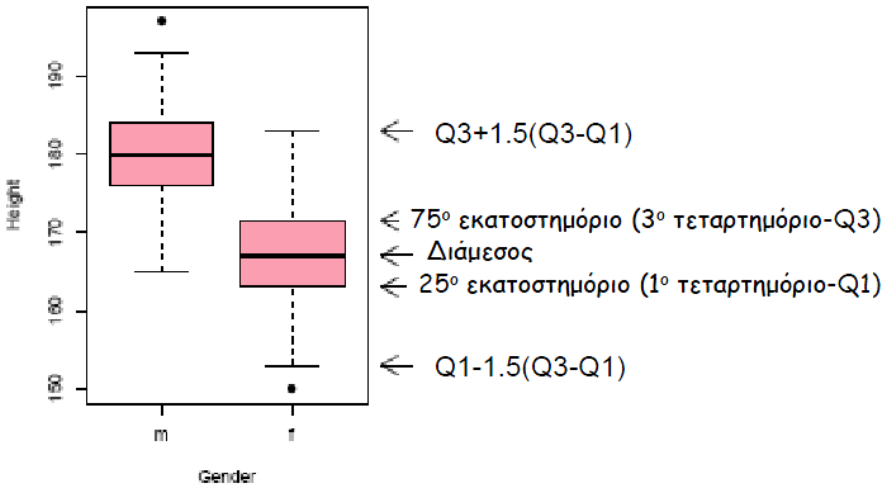
Θηκόγραμμα (box plot)

- Αρχικά υπολογίζουμε τα δύο τεταρτημόρια $Q1$ και $Q3$ και τη διάμεσο δ ($Q2$).
- Μετά κατασκευάζουμε ένα ορθογώνιο με την κάτω βάση στο $Q1$ και την άνω βάση στο $Q3$ Το μήκος των βάσεων του ορθογωνίου λαμβάνεται αυθαίρετα.
- Η διάμεσος παριστάνεται σαν ένα ευθύγραμμο τμήμα μέσα στο ορθογώνιο παράλληλο με τις βάσεις.
- Στη συνέχεια διακεκομμένες γραμμές εκτείνονται από τα μέσα των βάσεων του ορθογωνίου μέχρι τις οριακές (adjacent) τιμές που προκύπτουν ως εξής:

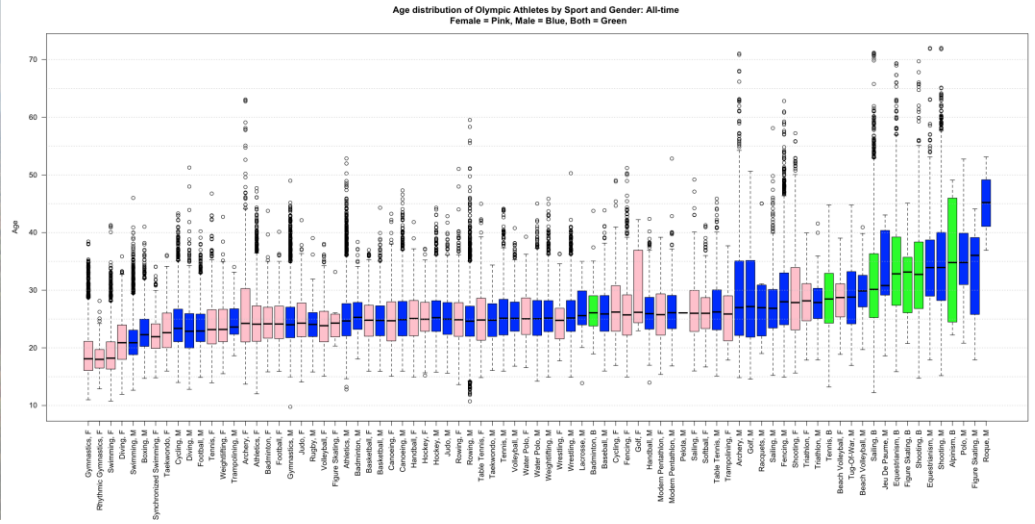
ο Η άνω τιμή ορίζεται ως η μεγαλύτερη παρατήρηση, η οποία είναι μικρότερη ή ίση από το $Q3 + 1.5\ (Q3 - Q1\ ) = Q3 + 3Q$

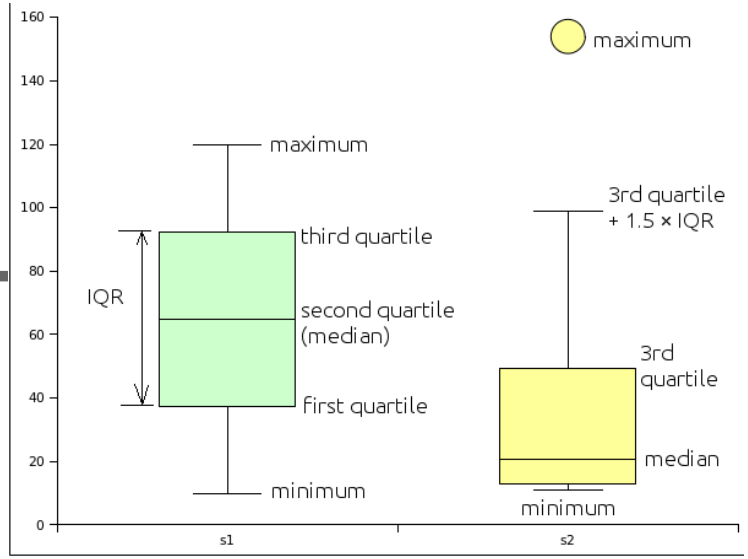
ο Η κατώτερη οριακή τιμή ορίζεται ως η μικρότερη παρατήρηση η οποία είναι μεγαλύτερη ή ίση από το $Q1 - 1.5\ (Q3 - Q1\ ) = Q1 - 3Q$

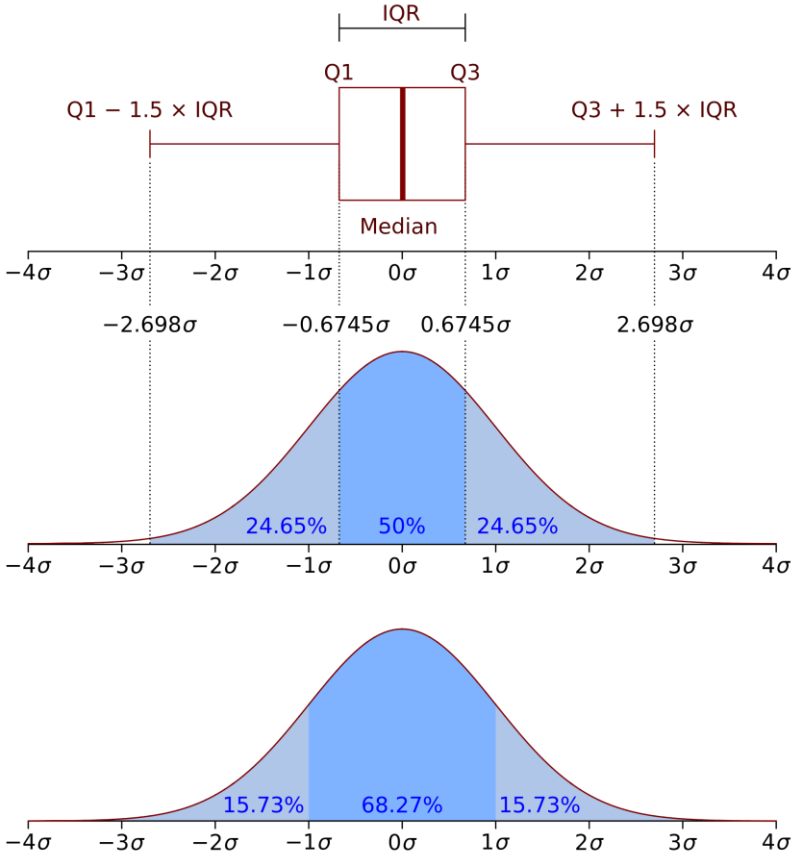
ο Εάν υπάρχουν ακόμη παρατηρήσεις που βρίσκονται έξω από το εύρος των δύο οριακών τιμών, αυτές καλούνται εξωτερικές τιμές και παριστάνονται με κάποιο ιδιαίτερο σύμβολο

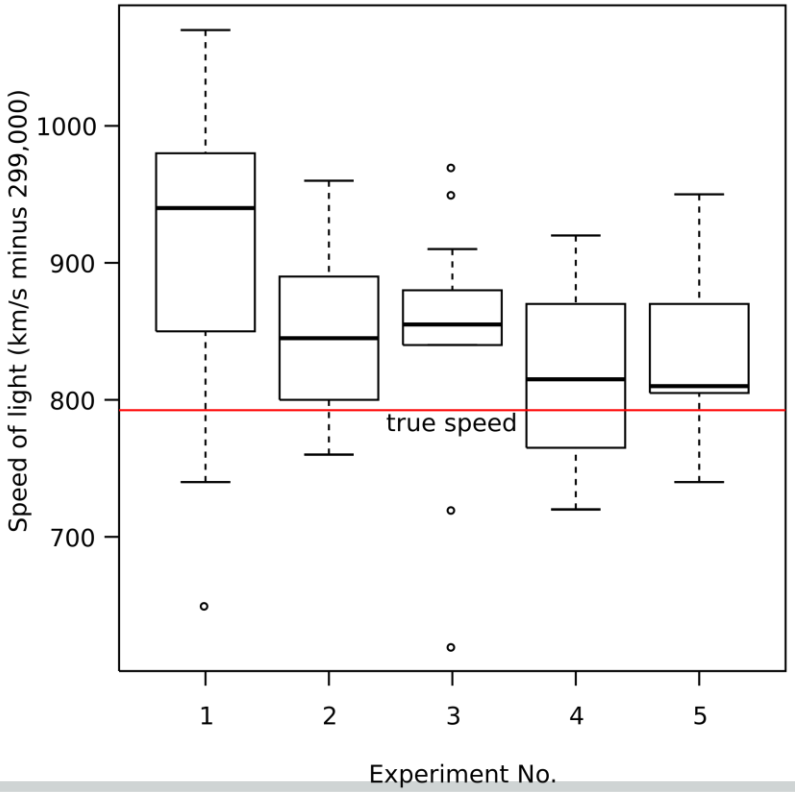


Σχήμα 10: Θηκόγραμμα της κατανομής του ύψους ανάλογα με το φύλο.


Age distribution of Olympic Athletes by Sport and Gender: All-time
Female = Pink, Male = Blue, Both = Green
Age


maximum
maximum
3rd quartile
+ 1.5 × IQR
third quartile
IQR
second quartile
(median)
3rd
quartile
first quartile
median
minimum
minimum
s1
s2


IQR
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%
15.73%
68.27%
15.73%


Speed of light (km/s minus 299,000)
1000
900
800
700
true speed
1
2
3
4
5
Experiment No.

Συντελεστής μεταβλητότητας

Για ένα σύνολο (συνήθως θετικών) παρατηρήσεων, ο λόγος της δειγματικής τυπικής απόκλισης προς τη δειγματική μέση τιμή, δηλαδή το πηλίκο:

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

λέγεται **συντελεστής μεταβλητότητας** (coefficient of variation).

• Ο συντελεστής μεταβλητότητας μπορεί να χρησιμοποιηθεί για συγκρίσεις ομάδων τιμών, είτε σε διαφορετικές μονάδες μέτρησης, είτε εκφράζονται στην ίδια μονάδα μέτρησης αλλά έχουν εντελώς διαφορετικές μέσες τιμές.
• Είναι δηλαδή ένα μέτρο της σχετικής μεταβλητότητας των τιμών και όχι της απόλυτης μεταβλητότητας όπως είναι τα άλλα μέτρα διασποράς που έχουμε αναφέρει.
• Γενικά θα δεχόμαστε ότι ένα δείγμα τιμών μιας μεταβλητής θα είναι ομοιογενές εάν ο συντελεστής μεταβλητότητας δεν ξεπερνά το 10%.
• Προφανώς ο συντελεστής μεταβλητότητας είναι ανεξάρτητος από τις χρησιμοποιούμενες μονάδες μέτρησης των τιμών των διαφόρων μεταβλητών.

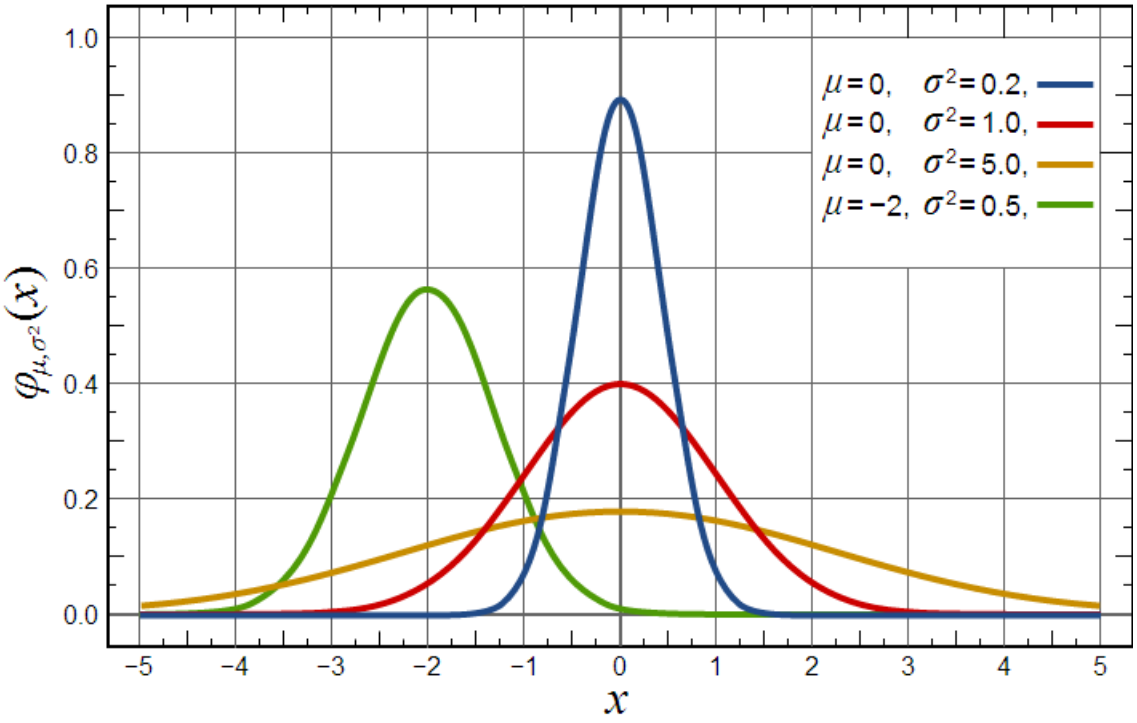
Κατανομές

Η κανονική κατανομή

Η συνάρτηση πυκνότητας πιθανότητας της κανονικής κατανομής είναι:

$$f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

όπου μ ο μέσος της κατανομής και σ η τυπική απόκλιση. Για μ=0 και σ=1 προκύπτει η τυποποιημένη κανονική κατανομή $N(0,1)$.



Σχήμα 11: Η συνάρτηση πυκνότητας πιθανότητας της κανονικής κατανομής κάτω από διάφορες τιμές για τις παραμέτρους μ και σ^2. Με κόκκινο απεικονίζεται η τυποποιημένη κανονική κατανομή.

Η κανονική κατανομή (2)

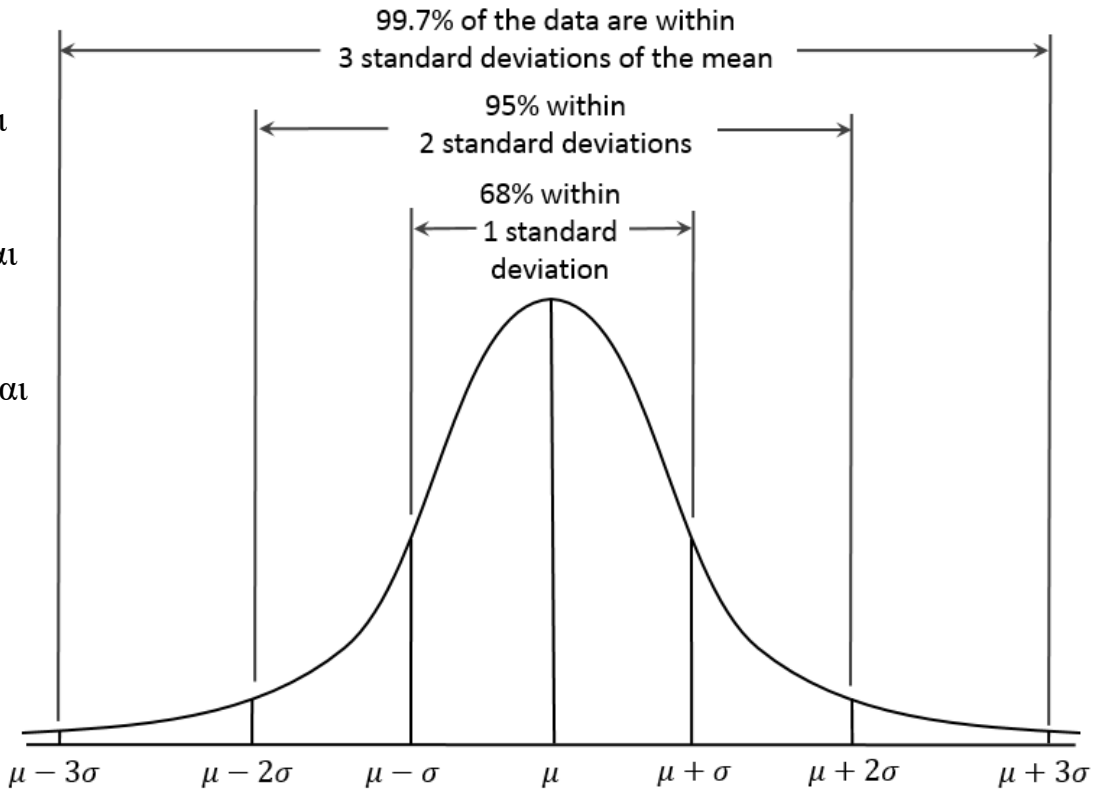
Αν το ιστόγραμμα των δεδομένων μοιάζει με το σχήμα της κανονικής κατανομής (καμπάνα του Gauss) τότε:

i) το 68% περίπου των παρατηρήσεων βρίσκεται στο διάστημα με άκρα τα σημεία $\bar{x} \pm s$,

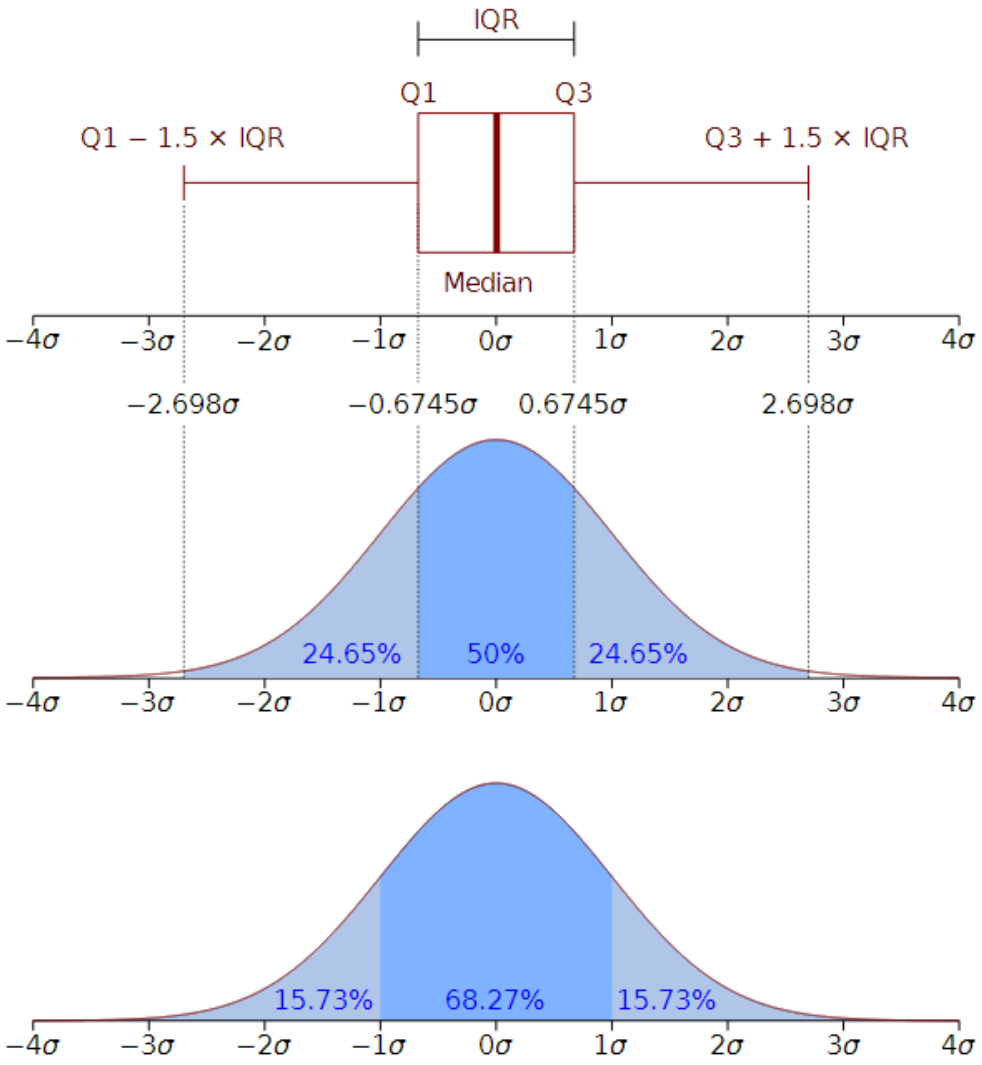
ii) το 95% περίπου των παρατηρήσεων βρίσκεται στο διάστημα με άκρα τα σημεία $\bar{x} \pm 2s$,

iii) το 99% περίπου των παρατηρήσεων βρίσκεται στο διάστημα με άκρα τα σημεία $\bar{x} \pm 3s$,

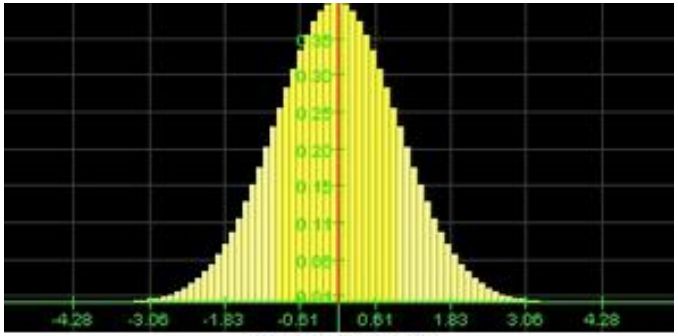
iv) ισχύει προσεγγιστικά η σχέση $R \cong 4s$.



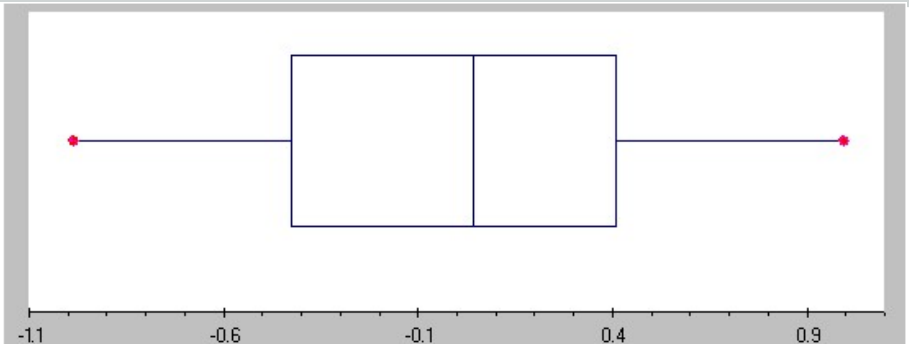
Θηκόγραμμα - συνάρτηση πυκνότητας πιθανότητας μιας $N(0,1)$



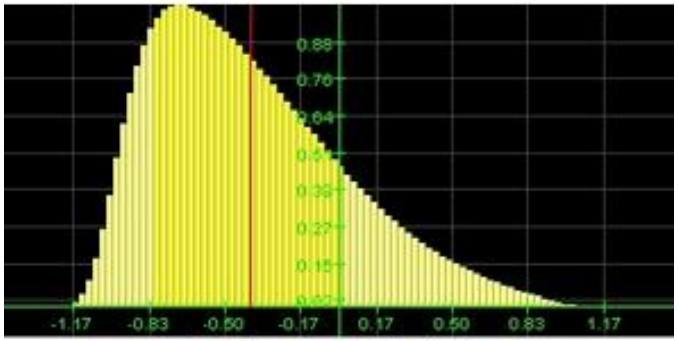
Συμμετρία (ιστόγραμμα-θηκόγραμμα)



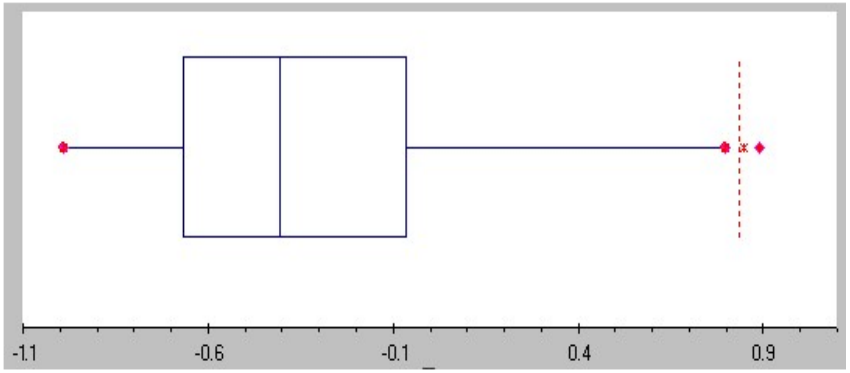
Normal Distribution



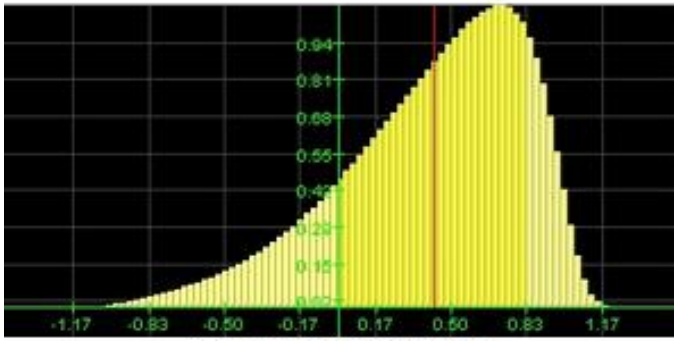
Distribution is (approximately) normal, mean and median should be similar *(the exact numbers are: mean = 0.013 median = 0.041)*



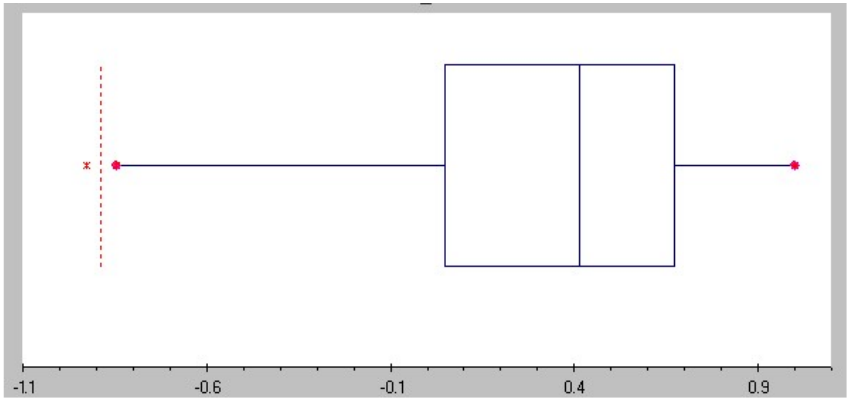
Skewed to the right Distribution



Distribution is shifted to the right, the mean should be greater than the median *(the exact numbers are: mean = -0.3192, median = -0.4061)*



Skewed to the left Distribution



Distribution is shifted to the left, the mean should be less than median *(the exact numbers are: mean = 0.3319, median = 0.4124).*

Δειγματοληπτικές κατανομές (χ^2, t, F)

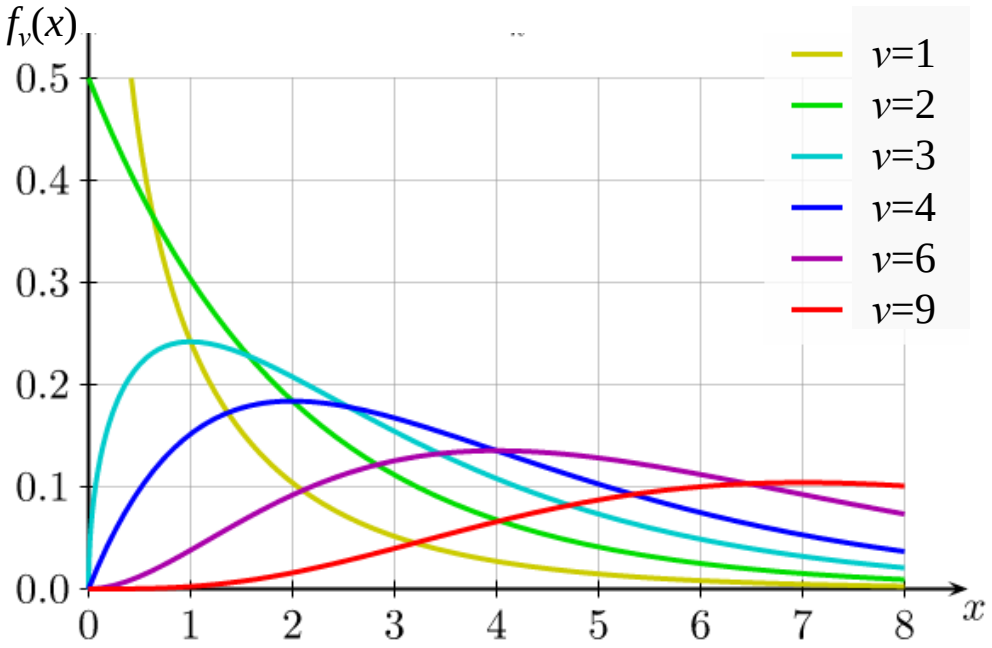
Η κατανομή πιθανότητας που χαρακτηρίζει κάποια στοιχεία της δειγματικής μεταβλητότητας ονομάζεται **δειγματική κατανομή** (sampling distribution). Οι μονοδιάστατες κατανομές που απορρέουν από την κανονική κατανομή διαδραμματίζουν σημαντικό στη στατιστική συμπερασματολογία.

1. **Κατανομή χ^2:** Έστω $X_1, X_2, \ldots X_\nu$ ανεξάρτητες και ισόνομες τυχαίες μεταβλητές που ακολουθούν την τυποποιημένη κανονική κατανομή $N(0,1)$. Η κατανομή της τυχαίας μεταβλητής:

$$U = X_1^2 + X_2^2 + \ldots + X_\nu^2$$

λέγεται χ^2 με ν βαθμούς ελευθερίας (β.ε.). Είναι λοξή προς τα δεξιά και καθώς οι β.ε. αυξάνονται η λοξότητα μειώνεται. Έχει:

$$E(\chi_\nu^2) = \nu \quad Var(\chi_\nu^2) = 2\nu$$



Δειγματοληπτικές κατανομές (χ^2, t, F) (2)

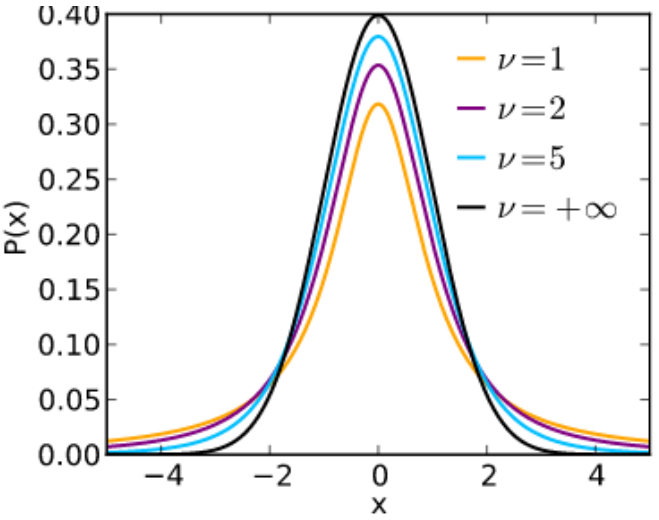
2. **Κατανομή *t*:** Έστω $X \sim N(0,1)$ και $Y \sim \chi^2_\nu$ όπου X και Y ανεξάρτητες τυχαίες μεταβλητές. Η κατανομή της τυχαίας μεταβλητής:

$$T = \frac{X}{\sqrt{Y/\nu}}$$

λέγεται *t* κατανομή νιούς ελευθερίας. Έχει:

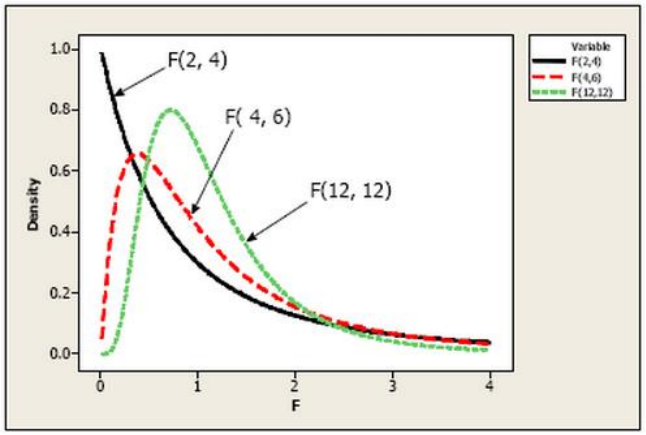
$$E(t_\nu) = 0$$

$$Var(t_\nu) = \frac{\nu}{\nu-2}$$



2. **Κατανομή *F*:** και όπου X_1 και X_2 ανεξάρτητες τυχαίες μεταβλητές. Η κατανομή της τυχαίας μ$X_1 \sim \chi^2_{\nu_1}$ής: $X_2 \sim \chi^2_{\nu_2}$

λέγεται F κατανομή με ν_1 *και* ν_2 βαθμούς ελευθερίας. Έχει: $T = \frac{X_1/\nu_1}{X_2/\nu_2}$



$$E(F_{\nu_1,\nu_2}) = \frac{\nu_2}{\nu_2-2}, \qquad \nu_2 > 2$$

$$Var(F_{\nu_1,\nu_2}) = \frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}, \qquad \nu_2 > 4$$

Εκτιμητική

Σημειακή εκτίμηση (1)

Μια συνάρτηση των παρατηρήσεων ενός δείγματος που εξαρτάται μόνο από τις παρατηρήσεις λέγεται **στατιστική συνάρτηση** (statistic).
Εκτιμήτρια (estimator) είναι μια τυχαία μεταβλητή που χρησιμοποιείται για να εκτιμήσει ένα χαρακτηριστικό του πληθυσμού (π.χ. μια παράμετρο). Η αριθμητική τιμή που η εκτιμήτρια παίρνει για κάποιο συγκεκριμένο δείγμα ονομάζεται **εκτίμηση** (estimate).

Ιδιότητες των σημειακών εκτιμητών:
1.Συνέπεια: Μια εκτιμήτρια $\hat{\theta}_n$ είναι μια συνεπής εκτιμήτρια μιας παραμέτρου θ του πληθυσμού αν για μια οποιαδήποτε πολύ μικρή θετική τιμή ε:

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| < \varepsilon) = 1$$

1.Αμεροληψία: Μια εκτιμήτρια $\hat{\theta}_n$ θα λέγεται αμερόληπτη (unbiased) αν ο μέσος της δειγματικής της κατανομής ισούται με την υπό εκτίμηση παράμετρο θ του πληθυσμού.

Σημειακή εκτίμηση (2)

3. **Αποτελεσματικότητα:** Ορίζουμε ως αποτελεσματικότητα (efficiency) μιας αμερόληπτης εκτιμήτριας $\hat{\theta}_1$ μιας πραγματικής παραμέτρου θ και συμβολίζουμε με eff($\hat{\theta}_1$) τον λόγο:

$$\text{eff}(\hat{\theta}) = \frac{I_\theta^{-1}}{V(\hat{\theta})}$$

4. **Επάρκεια:** Μια στατιστική συνάρτηση T ονομάζεται επαρκής (sufficient) για την παράμετρο θ ενός πληθυσμού αν η στατιστική αυτή συνάρτηση T περιέχει όλες τις πληροφορίες στο δείγμα γύρω από την παράμετρο θ.

Μέθοδοι σημειακής εκτίμησης

Οι κυριότερες από τις μεθόδους που χρησιμοποιούνται για τον καθορισμό σημειακών εκτιμητών είναι:

α) Μέθοδος των ροπών (method of moments)
β) Μέθοδος της μέγιστης πιθανοφάνειας (method of maximum likelihood)

Μέθοδος της μέγιστης πιθανοφάνειας

Η **εκτιμήτρια μέγιστης πιθανοφάνειας** (ΕΜΠ) (maximum likelihood estimator of θ (MLE)) του θ είναι η τιμή εκείνη του θ για την οποία το παρατηρηθέν ενδεχόμενο Ε έχει τη μεγαλύτερη δυνατή πιθανότητα κάτω από το συγκεκριμένο μοντέλο.

Η **συνάρτηση πιθανοφάνειας** (likelihood function) ορίζεται πολλές φορές ως ένα πολλαπλάσιο της P (E; θ), δηλαδή:

$L(\theta)= k\, P(E; \theta)$,

όπου k είναι μια οποιαδήποτε σταθερά. Ο φυσικός λογάριθμος της συνάρτησης πιθανοφάνειας (log likelihood function): $\ell(\theta) = \ln L(\theta)$

έχει την ιδιότητα ότι η τιμή του θ που τη μεγιστοποιεί μεγιστοποιεί και την $L(\theta)$. Ο παραμετρικός χώρος Θ των τιμών μιας παραμέτρου θ, όταν μιλάμε για την εκτίμηση μιας μόνο παραμέτρου, είναι ένα διάστημα πραγματικών τιμών. Επίσης η πρώτη και η δεύτερη παράγωγος $\ell'(\theta) = \frac{\partial}{\partial \theta}\ell(\theta)$ και $\ell''(\theta) = \frac{\partial^2}{\partial \theta^2}\ell(\theta)$

θα υπάρχουν σε εσωτερικά σημεία του Θ. Στην περίπτωση αυτή η ΕΜΠ θα προσδιορίζεται, συνήθως, ως η ρίζα της εξίσωσης μεγίστης πιθανοφάνειας. Μια ρίζα της παραπάνω εξίσωσης για την οποία $\ell''(\theta) < 0$ είναι σημείο τοπικού μέγιστου.

Εκτιμήτριες μέγιστης πιθανοφάνειας των παραμέρων μ και σ της κανονικής κατανομής (1)

Έστω ένα τυχαίο δείγμα $X_1, X_2, ..., X_n$ n ανεξαρτήτων παρατηρήσεων από μιά κανονική κατανομή με άγνωστη μέση τιμή μ και άγνωστη διασπορά σ^2 . Η συνάρτηση πιθανοφάνειας της κανονικής κατανομής είναι:

$$\mathbf{L}(\mu,\sigma^2|\mathbf{x}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\frac{-(x_1-\mu)^2}{2\sigma^2}\right)\cdots\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\frac{-(x_n-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}}\exp-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Ο λογάριθμος της συνάρτησης πιθανοφάνειας είναι:

$$\ln\mathbf{L}(\mu,\sigma^2|\mathbf{x}) = -\frac{n}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Οι μερικές παράγωγοι πρώτης τάξεως της συνάρτησης αυτής ως προς μ και σ^2 δίνουν:

$$\frac{\partial}{\partial\mu}\ln\mathbf{L}(\mu,\sigma^2|\mathbf{x}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = .\frac{1}{\sigma^2}n(\bar{x}-\mu)$$

$$\frac{\partial}{\partial\sigma^2}\ln\mathbf{L}(\mu,\sigma^2|\mathbf{x}) = -\frac{n}{\sigma^2} + \frac{1}{(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2 = \frac{n}{(\sigma^2)^2}\left(\sigma^2 - \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

Εκτιμήτριες μέγιστης πιθανοφάνειας των παραμέρων μ και σ της κανονικής κατανομής (1)

Επομένως, οι εκτιμήτριες μεγίστης πιθανοφάνειας των μ και σ^2 είναι:

$$\hat{\mu}(\mathbf{x}) = \bar{x}$$

$$\hat{\sigma}^2(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x})^2$$

Poisson

Example 1 (Poisson). Let $X_1, \ldots, X_n$ be an i.i.d. collection of Poisson(μ) random variables, where $\mu > 0$. Thus the likelihood function is

$$\begin{aligned} L(\mu; \boldsymbol{x}) &= \prod_{i=1}^{n} \frac{e^{-\mu} \mu^{x_i}}{x_i!} \\ &= e^{-n\mu} \mu^{\sum_{i=1}^n x_i} \frac{1}{\prod_{i=1}^n x_i!} \\ \ell(\mu; \boldsymbol{x}) &= -n\mu + \sum_{i=1}^{n} x_i \log \mu - \log \prod_{i=1}^{n} x_i! \,. \end{aligned}$$

We note that $\ell(\mu; \boldsymbol{x})$ is a differentiable function over the domain $(0, \infty)$, and so we first find the critical points:

$$\begin{aligned} 0 &= \frac{\partial \ell}{\partial \mu}(\mu; \boldsymbol{x}) \\ &= -n + \frac{\sum_{i=1}^n x_i}{\mu} \\ \mu &= \bar{x} \,. \end{aligned}$$

[Here $\bar{x}$ denotes $n^{-1} \sum_{i=1}^n x_i$.]

Bernoulli

Example 2 (Bernoulli trials). *If the experiment consists of n Bernoulli trial with success probability θ, then*

$$\mathbf{L}(\theta|\mathbf{x}) = \theta^{x_1}(1-\theta)^{(1-x_1)}\cdots\theta^{x_n}(1-\theta)^{(1-x_n)} = \theta^{(x_1+\cdots+x_n)}(1-\theta)^{n-(x_1+\cdots+x_n)}.$$

$$\ln \mathbf{L}(\theta|\mathbf{x}) = \ln\theta(\sum_{i=1}^{n} x_i) + \ln(1-\theta)(n-\sum_{i=1}^{n} x_i) = n\bar{x}\ln\theta + n(1-\bar{x})\ln(1-\theta).$$

$$\frac{\partial}{\partial\theta}\ln \mathbf{L}(\theta|\mathbf{x}) = n\left(\frac{\bar{x}}{\theta} - \frac{1-\bar{x}}{1-\theta}\right).$$

This equals zero when $\theta = \bar{x}$. Check that this is a maximum. Thus,

$$\hat{\theta}(\mathbf{x}) = \bar{x}.$$

Έλεγχοι υποθέσεων - διαστήματα εμπιστοσύνης

Έλεγχος υπόθεσης

Ο έλεγχος υπόθεσης είναι μια διαδικασία βάση της οποίας συνάγουμε συμπεράσματα για μια παράμετρο του πληθυσμού (π.χ. τη μέση τιμή), χρησιμοποιώντας πληροφορίες που προέρχονται από το δείγμα μας.

Πάντα ξεκινάμε από τη μηδενική υπόθεση (H_0): υποθέτουμε ότι η πραγματική μέση τιμή του πληθυσμού είναι ίση με δοθείσα τιμή k:
H_0: $\mu=k \rightarrow \Delta=\mu-k=0$
Η εναλλακτική υπόθεση είναι:
H_a: $\mu\neq k \rightarrow \Delta=\mu-k\neq 0$

Αρχίζουμε υποθέτοντας ότι η μηδενική ισχύει και ρωτάμε: «Πόσο πιθανό είναι να έχουμε αυτό το αποτέλεσμα που μας δίνει το δείγμα;»

- Μπορεί το εύρημά μας να οφείλεται στην τύχη.
- Μπορεί η H_0 να ισχύει, και απλά πετύχαμε ένα ασυνήθιστο, μη αντιπροσωπευτικό δείγμα.

Σφάλματα

Τύπου Ι: Είναι το σφάλμα το οπόιο γίνεται όταν απορρίπτουμε την H_o ενώ είναι αληθής

Τύπου ΙΙ: Είναι το σφάλμα το οπόιο γίνεται όταν δεχόμαστε την H_o ενώ είναι ψευδής

	Πραγματικότητα	
Συμπέρασμα	H_0 αληθής	H_0 ψευδής
Απόρριψη H_0	Σφάλμα Τύπου I (Type I error)	Σωστό
'Αποδοχή' H_0	Σωστό	Σφάλμα Τύπου II (Type II error)

P-value

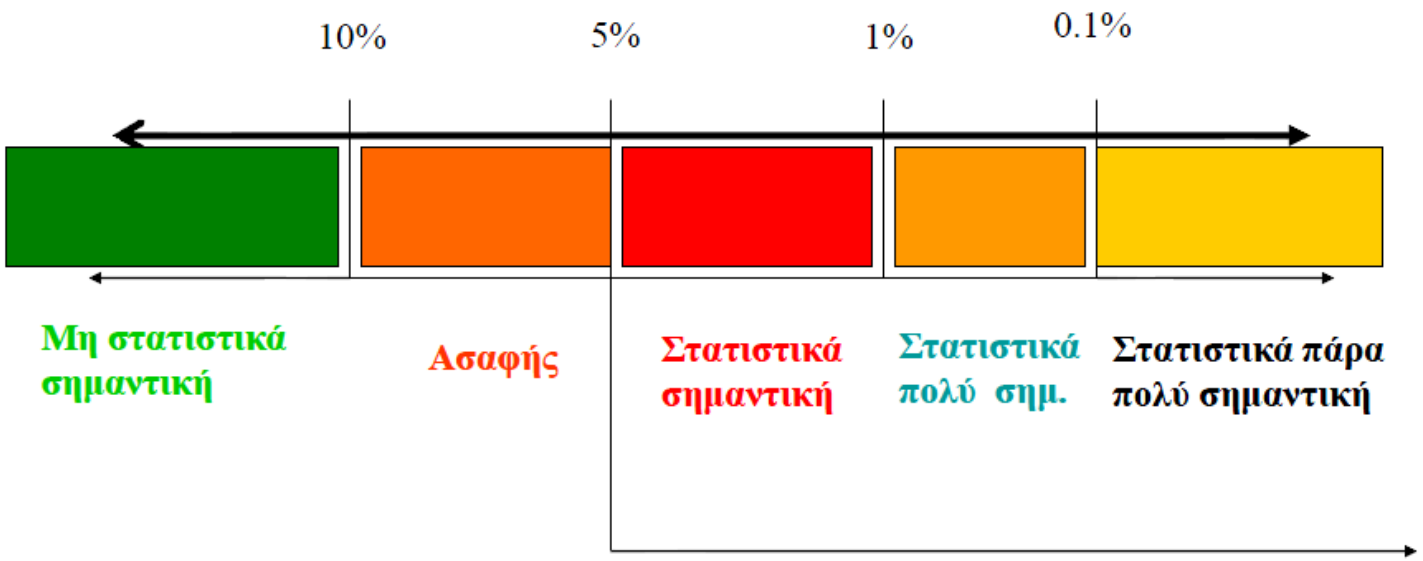
p-value: είναι η πιθανότητα να παρατηρήσουμε ένα αποτέλεσμα τόσο ή περισσότερο ακραίο όσο το αποτέλεσμα ενός συγκεκριμένου δείγματος δεδομένου ότι ισχύει η μηδενική υπόθεση

Μικρό p-value: τα αποτελέσματα από το δείγμα δεν είναι πιθανά δεδομένου της H_0

Χρησιμοποιούμε ως επίπεδο σημαντικότητας a=0,05 (5%) (αυθαίρετο)

Αν p-value < a, τότε απορρίπτουμε την H_0 και αποδεχόμαστε την H_a.

Αν p-value > a, τότε αποτυγχάνουμε να απορρίψουμε την H_0



Διάστημα εμπιστοσύνης μέσης τιμής

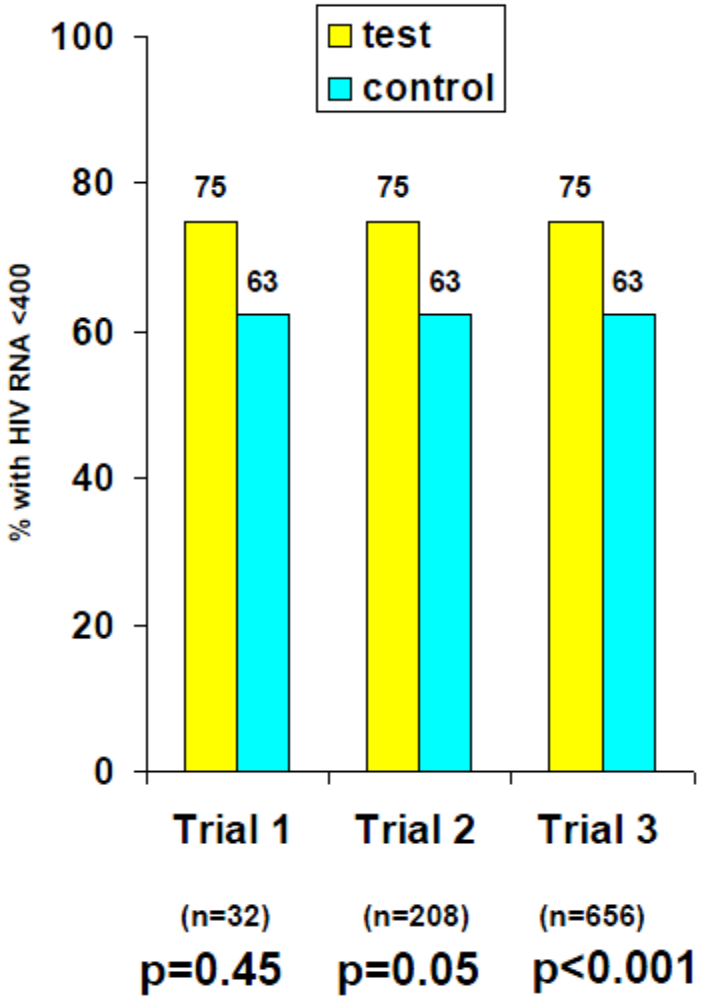
Το διάστημα: μέση τιμή+1,96*SE περιλαμβάνει την πραγματική μέση τιμή με πιθανότητα 95% (95% Διάστημα Εμπιστοσύνης) όπου:

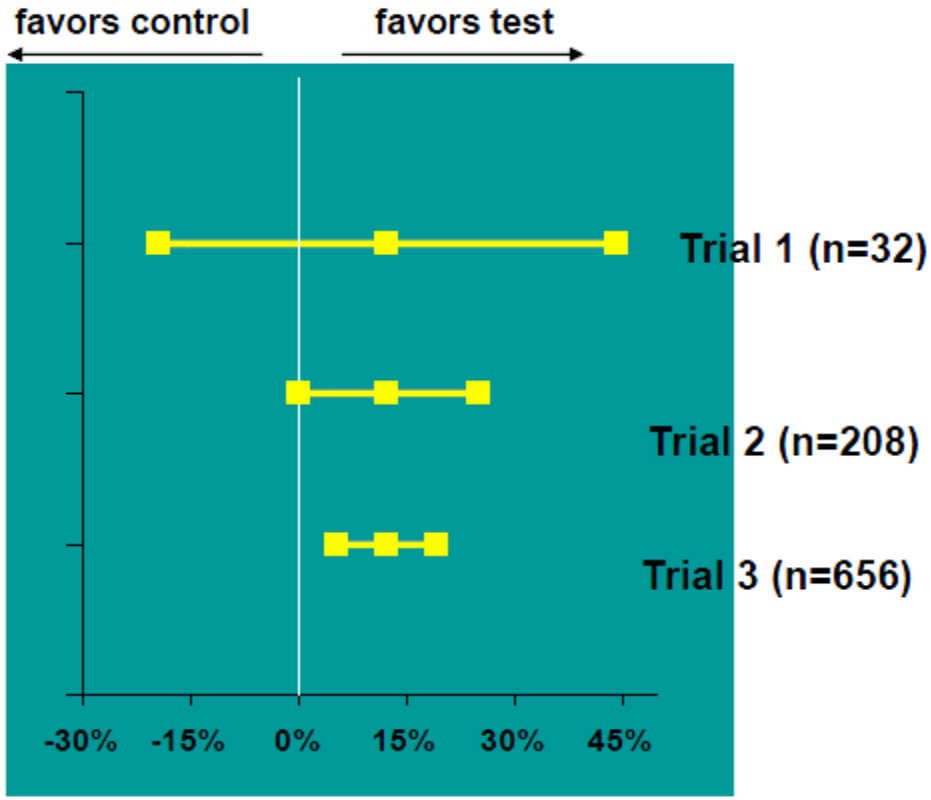
$$\mathrm{SE}_{\bar{x}} = \frac{s}{\sqrt{n}}$$

SE: Πιθανό σφάλμα
s: Τυπική απόκλιση
n: μέγεθος του δείγματος

Αν το δείγμα είναι μικρό τότε: μέση τιμή+t_{n-1}* SE, από την κατανομή t.

P-value και διαστήματα εμπιστοσύνης ανάλογα με το μέγεθος του δείγματος





95% CI for difference in response rates

Elegxos ypotheseon -ttest.pdf - Adobe Reader

Ανάλυση ποσοτικών δεδομένων

t-test για ένα δείγμα (1)

Ας υποθέσουμε ότι έχουμε μια τυχαία μεταβλητή X από πληθυσμό με μέση τιμή μ και διασπορά σ^2 (μ, σ^2 άγνωστα) και ενδιαφερόμαστε να ελέγξουμε την υπόθεση H_0: $\mu=\mu_0$ έναντι της H_a:$\mu\neq\mu_o$ σε επίπεδο σημαντικότητας a. Η στατιστική συνάρτηση ελέγχου είναι:

$$Z = \frac{\bar{X} - \mu_0}{SE(\bar{X})} = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

που ακολουθεί την $N(0,1)$. Επειδή το σ είναι άγνωστο το εκτιμούμε από τη δειγματική τυπική απόκλιση s και έτσι η στατιστική συνάρτηση ελέγχου γίνεται:

$$T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$$

και ακολουθεί την t-κατανομή με n-1 βαθμούς ελευθερίας. Στην περίπτωση που το μέγεθος του δείγματος είναι μεγάλο η t-κατανομή προσεγγίζεται από την κανονική κατανομή. Για αμφίπλευρο έλεγχο το P-value προκύπτει ως η διπλάσια πιθανότητα να παρατηρήσουμε την τιμή του στατιστικού |t| και οτιδήποτε πιο ακραίο από αυτό. Απαραίτητη προϋπόθεση για την εφαρμογή του t-test είναι τα δεδομένα μας να είναι κανονικά κατανεμημένα. Για την κατασκευή ενός 1-α % διαστήματος επιστοσύνης χρησιμοποιούμε την ακόλουθη σχέση:

$$\left(\bar{x} - t_{n-1,\alpha/2} S / \sqrt{n},\ \bar{x} + t_{n-1,\alpha/2} S / \sqrt{n}\right)$$

t-test για ένα δείγμα (2)

Βήμα 1: Διατύπωση μηδενικής υπόθεσης H_0: $\mu=\mu_0$ (ο πραγματικός μέσος του πληθυσμού ισούται με μία τιμή μ_0 που θέλουμε να ελέγξουμε) με εναλλακτικές τις ακόλουθες περιπτώσεις :

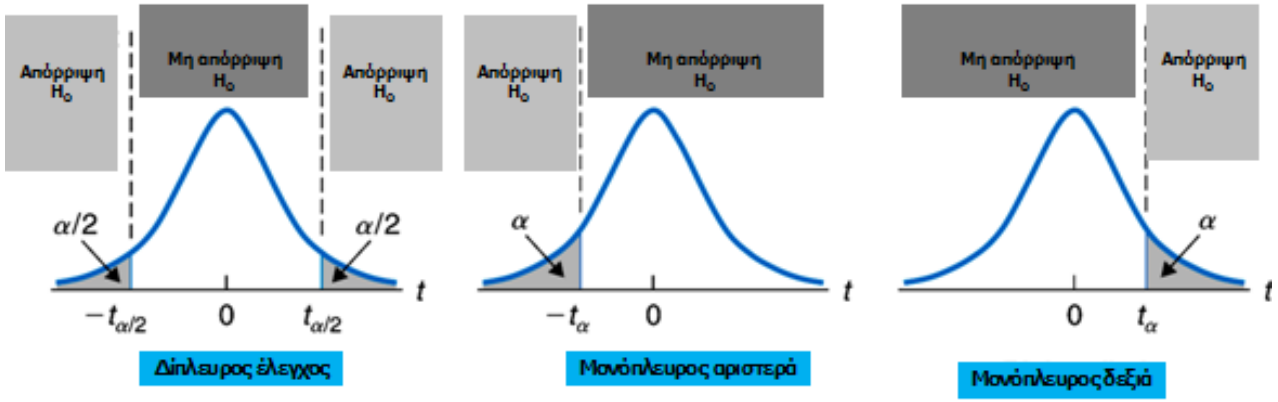
$H_a:\mu\neq\mu_o$	$H_a:\mu<\mu_o$	$H_a:\mu>\mu_o$
(Αμφίπλευρος Έλεγχος)	(Μονόπλευρος αριστερά)	(Μονόπλευρος δεξιά)

Βήμα 2: Επιλογή επιπέδου σημαντικότητας *a*

Βήμα 3: Εύρεση κρίσιμων τιμών για β.ε.=*ν* - 1.

Βήμα 4: Υπολογισμός της στατιστικής συνάρτησης ελέγχου t

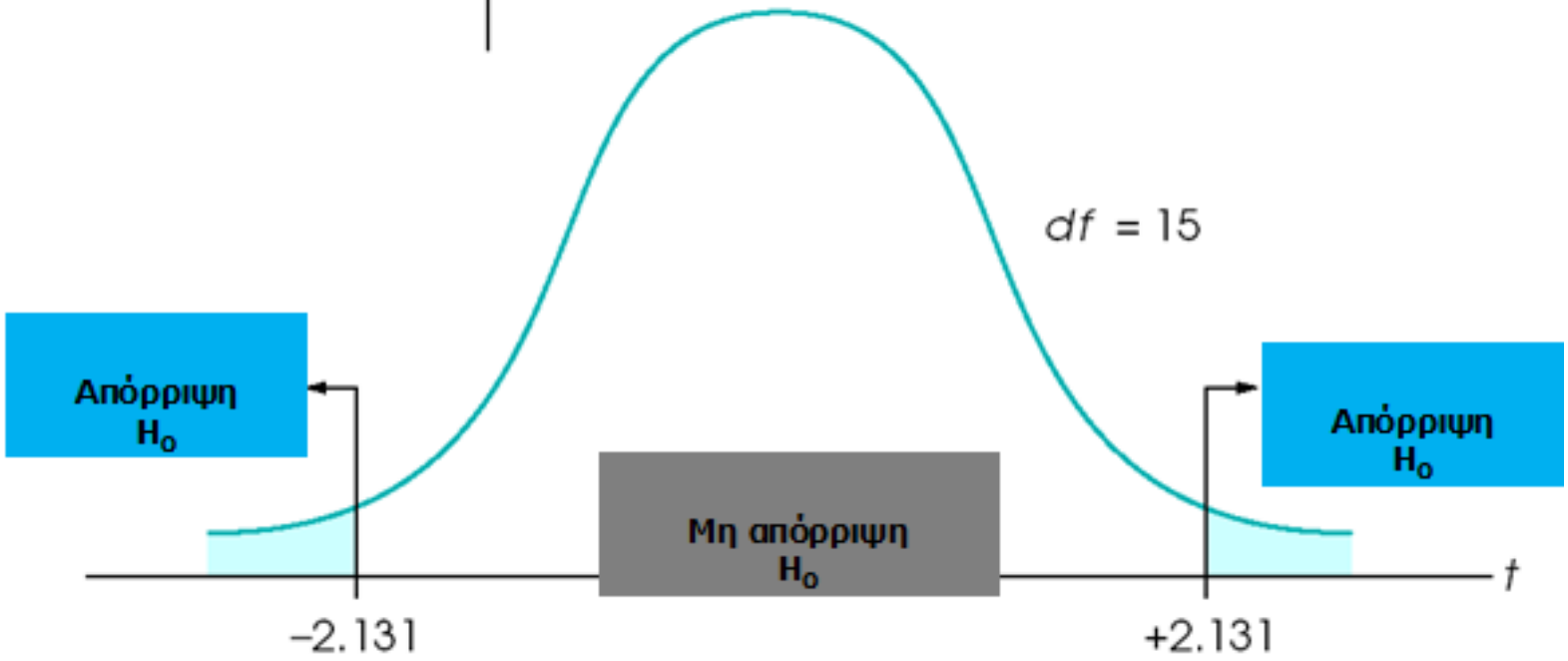
Βήμα 5: Εάν η τιμή της t βρίσκεται εντός της περιοχής απόρριψης, απορρίπτουμε την H_0. Διαφορετικά δεν μπορούμε να απορρίψουμε την H_0



t-test για ένα δείγμα (3)

Εύρεση κρίσιμων τιμών για β.ε.= 15, <u>αμφίπλευρος</u> *a* = 0,05.

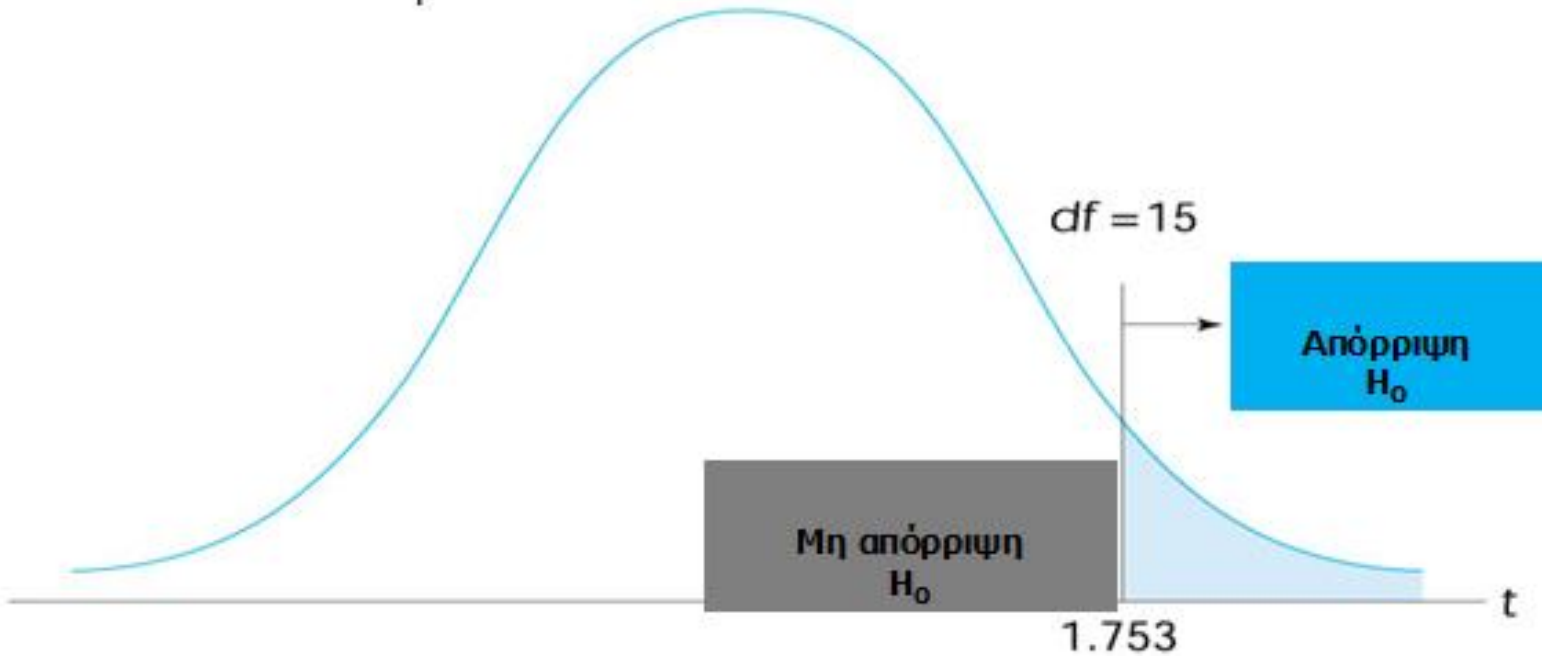
One Sided	75%	80%	85%	90%	95%	97.5%	99%	99.5%	99.75%	99.9%	99.95%
Two Sided	50%	60%	70%	80%	90%	95%	98%	99%	99.5%	99.8%	99.9%
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015



t-test για ένα δείγμα (4)

Εύρεση κρίσιμων τιμών για β.ε.= 15, μονόπλευρος *a* = 0,05.

One Sided	75%	80%	85%	90%	95%	97.5%	99%	99.5%	99.75%	99.9%	99.95%
Two Sided	50%	60%	70%	80%	90%	95%	98%	99%	99.5%	99.8%	99.9%
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015



t-test για δυο ανεξάρτητα δείγματα (1)

Θέλουμε να συγκρίνουμε δύο μέσες τιμές, που προέρχονται από δυο ανεξάρτητους πληθυσμούς. (για παράδειγμα οι τιμές χοληστερόλης σε άνδρες και γυναίκες, ή η επίδοση στη βιοχημεία σε φοιτητές από τη σχολή Α και τη σχολή Β)

Προϋποθέσεις:

❖Η μεταβλητή που μας ενδιαφέρει να ακολουθεί την κανονική κατανομή και στους 2 ανεξάρτητους πληθυσμούς.

❖Οι τυπικές αποκλίσεις να μη διαφέρουν. (F-test)

t-test για δυο ανεξάρτητα δείγματα (2)

Βήμα 1: Διατύπωση μηδενικής υπόθεσης H_0: $\mu_1=\mu_2$ ή $\mu_1-\mu_2=0$ (δεν υπάρχει διαφορά στις μέσες τιμές των δύο δειγμάτων) με εναλλακτικές τις ακόλουθες περιπτώσεις :

$H_a:\mu_1\neq\mu_2$ (Αμφίπλευρος Έλεγχος)

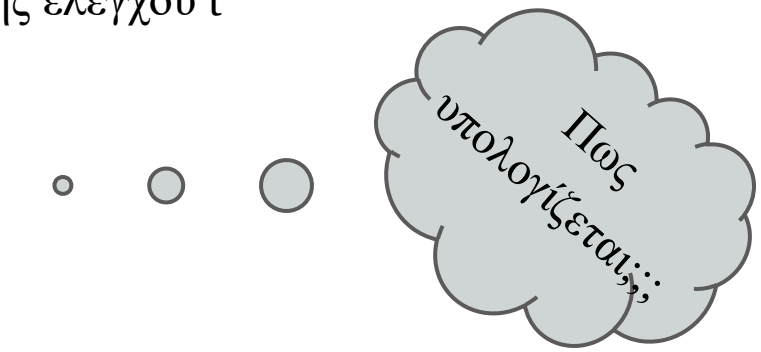
$H_a:\mu_1<\mu_2$ (Μονόπλευρος αριστερά)

$H_a:\mu_1>\mu_2$ (Μονόπλευρος δεξιά)

Βήμα 2: Επιλογή επιπέδου σημαντικότητας *a*

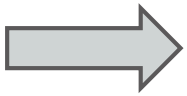
Βήμα 3: Υπολογισμός της στατιστικής συνάρτησης ελέγχου t

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE_{1/2}}$$



t-test για δυο ανεξάρτητα δείγματα (3)

Υπόθεση 1: $\sigma_1=\sigma_2$ άγνωστες αλλά ίσες

Υπολογίζεται μια "μέση απόκλιση"  $$Sw_{1/2} = \sqrt{\frac{(v_1 - 1)s_1^2 + (v_2 - 1)s_2^2}{v_1 + v_2 - 2}}$$

Βήμα 3: Εύρεση κρίσιμων τιμών για β.ε.=$\nu_1+\nu_2$ - 2

Βήμα 4: Υπολογισμός της στατιστικής συνάρτησης ελέγχου t

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE_{1/2}} = \frac{(\bar{x}_1 - \bar{x}_2)}{SE_{1/2}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{(\frac{s_1}{\sqrt{v_1}})^2 + (\frac{s_2}{\sqrt{v_2}})^2}}$$

$$= \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{(\frac{S_{w_{1/2}}}{\sqrt{v_1}})^2 + (\frac{S_{w_{1/2}}}{\sqrt{v_2}})^2}} = \frac{(\bar{x}_1 - \bar{x}_2)}{S_{w_{1/2}}\sqrt{\frac{1}{v_1} + \frac{1}{v_2}}}$$

Βήμα 5: Εάν η τιμή της t βρίσκεται εντός της περιοχής απόρριψης, απορρίπτουμε την H_0. Διαφορετικά δεν μπορούμε να απορρίψουμε την H_0

t-test για δυο ανεξάρτητα δείγματα (4)

Υπόθεση 2: $\sigma_1=\sigma_2$ άγνωστες αλλά άνισες

Βήμα 3: Εύρεση κρίσιμων τιμών για β.ε.:

Satterthwaite

$$v = \frac{\left(\frac{s_1^2}{v_1} + \frac{s_2^2}{v_2}\right)^2}{\frac{(s_1^2/v_1)^2}{v_1 - 1} + \frac{(s_2^2/v_2)^2}{v_2 - 1}}$$

Welch

$$v = -2 + \frac{\left(\frac{s_1^2}{v_1} + \frac{s_2^2}{v_2}\right)^2}{\frac{(s_1^2/v_1)^2}{v_1 + 1} + \frac{(s_2^2/v_2)^2}{v_2 + 1}}$$

Βήμα 4: Υπολογισμός της στατιστικής συνάρτησης ελέγχου t

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{SE_{1/2}} = \frac{(\bar{x}_1 - \bar{x}_2)}{SE_{1/2}} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\left(\frac{s_1}{\sqrt{v_1}}\right)^2 + \left(\frac{s_2}{\sqrt{v_2}}\right)^2}}$$

$$= \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{v_1} + \frac{s_2^2}{v_2}}}$$

Βήμα 5: Εάν η τιμή της t βρίσκεται εντός της περιοχής απόρριψης, απορρίπτουμε την H_0. Διαφορετικά δεν μπορούμε να απορρίψουμε την H_0

Παράδειγμα 7

Για τη μελέτη των πιθανών επιδράσεων των επιπέδων του μολύβδου στην ανάπτυξη των παιδιών μελετήθηκε ένα τυχαίο δείγμα 25 παιδιών σχολικής ηλικίας από περιοχή με χαμηλά επίπεδα μολύβδου στο περιβάλλον (περιοχή Α) και ένα αντίστοιχο δείγμα 20 παιδιών από περιοχή με υψηλά επίπεδα μολύβδου στο περιβάλλον (περιοχή Β). Στον πίνακα που ακολουθεί δίνεται το ανάστημα των παιδιών κάθε περιοχής.

(α) Διαφέρει το ανάστημα των παιδιών στις δύο αυτές περιοχές;
(β) Ποια είναι τα 95% όρια αξιοπιστίας της διαφοράς του μέσου αναστήματος των παιδιών στις περιοχές αυτές;

ΠΕΡΙΟΧΗ Α		ΠΕΡΙΟΧΗ Β	
Α/Α	Ανάστημα παιδιού (cm)	Α/Α	Ανάστημα παιδιού (cm)
1	125	1	122
2	122	2	114
3	130	3	124
4	132	4	122
5	116	5	120
6	122	6	123
7	127	7	115
8	120	8	116
9	128	9	126
10	127	10	129
11	122	11	122
12	131	12	122
13	120	13	120
14	124	14	118
15	128	15	118
16	121	16	122
17	120	17	122
18	122	18	118
19	117	19	125
20	123	20	113
21	116		
22	125		
23	121		
24	134		
25	119		

Παράδειγμα 7 (συνέχεια)

$$\bar{X}_1 = \frac{125+122+130+.....+119}{25} = \frac{3092}{25} = 123{,}68 \text{ cm}$$

$$\bar{X}_2 = \frac{122+114+124+...+113}{20} = \frac{2411}{20} = 120{,}55 cm$$

$$SD_1 = \sqrt{\frac{\sum_i (X_{1i} - \bar{X}_1)^2}{n_1 - 1}} = \sqrt{\frac{\sum_i X_{1i}^2 - \frac{(\sum_i X_{1i})^2}{n_1}}{n_1 - 1}}$$

$$\sum_i X_{1i}^2 = 125^2 + 122^2 + 130^2 + ... + 119^2 = 383002$$

Παράδειγμα 7 (συνέχεια)

$$(\sum_{i} X_{1i})^2 = 3092^2 = 9560464$$

$$SD_1 = \sqrt{\frac{383002 - \frac{9560464}{25}}{24}} = \sqrt{\frac{383002 - 382418,56}{24}}$$

$$= \sqrt{\frac{583,44}{24}} = \sqrt{24,31} = 4,93 \text{ cm}$$

$SD_2 = 4,12$ cm

$$SE_1 = \frac{SD_1}{\sqrt{n_1}} = \frac{4,93}{\sqrt{25}} = 0,986\,cm$$

$$SE_2 = \frac{SD_2}{\sqrt{n_2}} = \frac{4,12}{\sqrt{20}} = 0,922\,cm$$

Παράδειγμα 7 (συνέχεια)

$$t = \frac{\left|\bar{X}_1 - \bar{X}_2\right|}{\sqrt{SE_1^{\,2} + SE_2^{\,2}}} = \frac{\left|123{,}68 - 120{,}55\right|}{\sqrt{0{,}986^2 + 0{,}922^2}} = \frac{3{,}13}{1{,}35} = 2{,}32$$

B.E= n1+n2-2 = 25+20-2 = 43

Αξιολόγηση στους πίνακες του t -test

B.E.	10%	5%	1%	1‰
43	1,68	2,02	2,71	3,55

2,02 < 2.32< 2,71

5% > p > 1%

Συμπέρασμα: Το ανάστημα των παιδιών διαφέρει σε βαθμό στατιστικά σημαντικό στις δύο περιοχές

Ερμηνεία: Τα παιδιά από περιοχές με χαμηλά επίπεδα μολύβδου είναι κατά μέσο όρο υψηλότερα από τα παιδιά από περιοχές με υψηλά επίπεδα μόλυβδου

Παράδειγμα 7 (συνέχεια)

β) 95 % όρια αξιοπιστίας της διαφοράς του μέσου αναστήματος

$$\bar{\delta} = \bar{X}_1 - \bar{X}_2 = 3{,}13 \quad cm$$

$$95\%CI : \bar{\delta} \pm t * SE_{\delta}$$

$$3{,}13 \pm 2{,}02 * 1{,}35 \; cm$$

↓ ↓

0,40 5,86

95% CI: (0,40 - 5,86) cm

Τα παιδιά από περιοχές με χαμηλά επίπεδα μολύβδου είναι υψηλότερα κατά μέσο όρο κατά 3,13 cm από τα παιδιά από περιοχές με υψηλά επίπεδα μολύβδου, με 95% όρια αξιοπιστίας της διαφοράς 0,40 – 5,86 cm

Ανάλυση ποιοτικών δεδομένων

Η δοκιμασία X^2 ως κριτήριο καλής εφαρμογής

Βήματα:

- Διατύπωση υποθέσεων, μηδενικής H_0 και εναλλακτικής H_a
- Επιλογή επιπέδου σημαντικότητας *a*
- Η θεωρητική κατανομή υπό την H_0 είναι η χ^2 με *k* - 1 β.ε.
- Υπολογίζουμε την κρίσιμη τιμή χ^2_c & διατυπώνουμε τον κανόνα λήψης αποφάσεων: Απόρριψη H0 εάν $\chi^2 > \chi^2_c$
- Υπολογισμός χ^2
- Εφαρμογή του κανόνα λήψης αποφάσεων

Η δοκιμασία X^2 ως κριτήριο καλής εφαρμογής

Διασταύρωση υβριδίων με ροζ χρώμα (Κα x Κα)

Απόγονοι: 50% ροζ, 25% κόκκινο, 25% λευκό

❖Διατύπωση μηδενικής και εναλλακτικής υπόθεσης:

H_0 : P(ροζ, άσπρου, κόκκινου) = 0.5, 0.25, 0.25

H_a : P(ροζ, άσπρου, κόκκινου) ≠ 0.5, 0.25, 0.25

❖Η συνάρτηση ελέγχου είναι η ακόλουθη:

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

όπου:

O : Η παρατηρηθείσα συχνότητα (Observed value)

E : Η αναμενόμενη συχνότητα (Expected value)

Η δοκιμασία X^2 ως κριτήριο καλής εφαρμογής

Μελέτη χρώματος άνθεων σε 120 φυτά

	Ροζ	Κόκκινο	Άσπρο
Παρατηρηθείσες συχνότητες (O)	75	25	20
Αναμενόμενες συχνότητες (E)	120 x 0,5 = 60	120 x 0,25 = 30	120 x 0,25 = 30

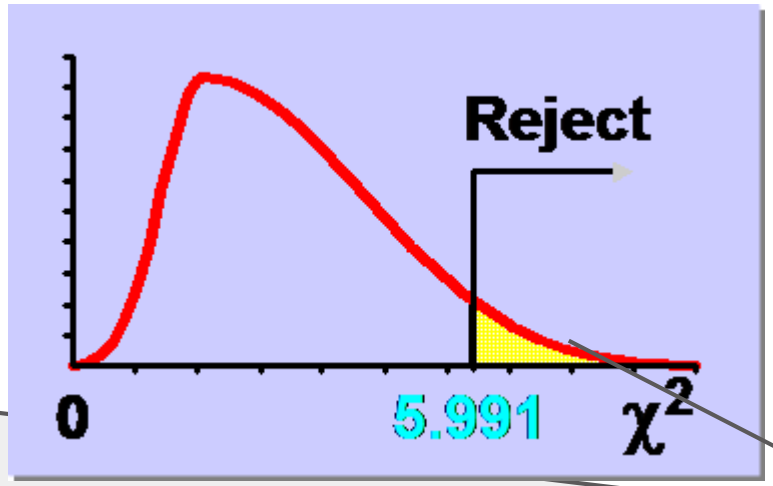
$$X^2 = \sum \frac{(O-E)^2}{E} = \frac{(75-60)^2}{60} + \frac{(25-30)^2}{30} + \frac{(20-30)^2}{30} = 3,75 + 0,83 + 3,33 = 7,91$$

Η δοκιμασία X^2 ως κριτήριο καλής εφαρμογής

Κελιά = 3 & α =0,05
β.ε. = k - 1 = 2

Πίνακας με τις κρίσιμες τιμές της $\chi 2$ κατανομής)

df	.995	...	.95	...	.05
1	...	...	0.004	...	3.841
2	0.010	...	0.103	...	5.991



Απόφαση:
Απόρριψη σε a = 0,05.
Δεν ισχύει το γενετικό μοντέλο κληρονομικότητας που υποθέσαμε

$\chi 2$ = 7,91

Η δοκιμασία X^2 ως κριτήριο συσχέτισης ποιοτικών χαρακτηριστικών

Γενικοί ορισμοί

❖Μεταβλητή με *r* κατηγορίες (γραμμές)

❖Μεταβλητή με *c* κατηγορίες (στήλες)

❖P*i*. η πιθανότητα ένα άτομο να εντάσσεται στην *i* από τις *r* κατηγορίες της μίας μεταβλητής

❖P.*j* η πιθανότητα ένα άτομο να εντάσσεται στην *j* από τις *c* κατηγορίες της άλλης μεταβλητής

❖P*ij* η πιθανότητα ένα άτομο να εντάσσεται ταυτόχρονα στην *i* από τις *r* κατηγορίες της μίας μεταβλητής και στην *j* από τις *c* κατηγορίες της άλλης μεταβλητής

Προϋποθέσεις εφαρμογής ελέγχου χ^2

❖Η αναμενόμενη συχνότητα σε κάθε κελί πρέπει να είναι > 1 και το 80% > 5

❖Ο συνολικός αριθμός παρατηρήσεων πρέπει να είναι > 20

Η δοκιμασία X^2 ως κριτήριο συσχέτισης ποιοτικών χαρακτηριστικών

	Στήλη 1	Στήλη 2	…	Στήλη *j*	…	Στήλη *c*	Σύνολο
Γραμμή 1	n_{11}	n_{12}		n_{1j}		n_{1c}	$N_{1.}$
Γραμμή 2	n_{21}	n_{22}		n_{2j}		n_{2c}	$N_{2.}$
….							
Γραμμή *i*	n_{i1}	n_{i2}		n_{ij}		n_{ic}	$N_{i.}$
….							
Γραμμή *r*	n_{r1}	n_{r2}		n_{rj}		n_{rc}	$N_{r.}$
Σύνολο	$S_{.1}$	$S_{.2}$		$S_{.j}$		$S_{.c}$	T

$$H_0 : P_{ij} = P_{i.} * P_{.j} = \frac{N_{i.}}{T} * \frac{S_{.j}}{T}, \quad O = n_{ij}, \quad E = P_{ij} * T = \left(\frac{N_{i.}}{T} * \frac{S_{.j}}{T}\right) * T = \frac{N_{i.} * S_{.j}}{T}$$

$$H_1 : P_{ij} \neq P_{i.} * P_{.j}$$

Η δοκιμασία X^2 ως κριτήριο συσχέτισης ποιοτικών χαρακτηριστικών

Επιδημιολογικό πρόβλημα

Υπάρχει συσχέτιση μεταξύ επαγγέλματος και ανάπτυξης καρκίνου της ουροδόχου κύστης;

	Επάγγελμα			
	Γεωργοί	*Υπάλληλοι*	*Εργάτες*	*Σύνολο*
Καρκίνος				
Ναι	140 (320*380)/890=136,6	125 (320*390)/890=140,2	55 (320*120)/890=43,2	320
Όχι	240	265	65	570
Σύνολο	380	390	120	890

$$X^2 = \sum \frac{(O-E)^2}{E} = \frac{(140-136,6)^2}{136,6} + \frac{(125-140,2)^2}{140,2} + \frac{(55-43,2)^2}{43,2} + \frac{(240-243,4)^2}{243,4} + \frac{(265-249,8)^2}{249,8} + \frac{(65-76,8)^2}{76,8} =$$

$$0,08 + 1,65 + 3,22 + 0,05 + 0,92 + 1,81 = 7,73$$

Η δοκιμασία X^2 ως κριτήριο συσχέτισης ποιοτικών χαρακτηριστικών

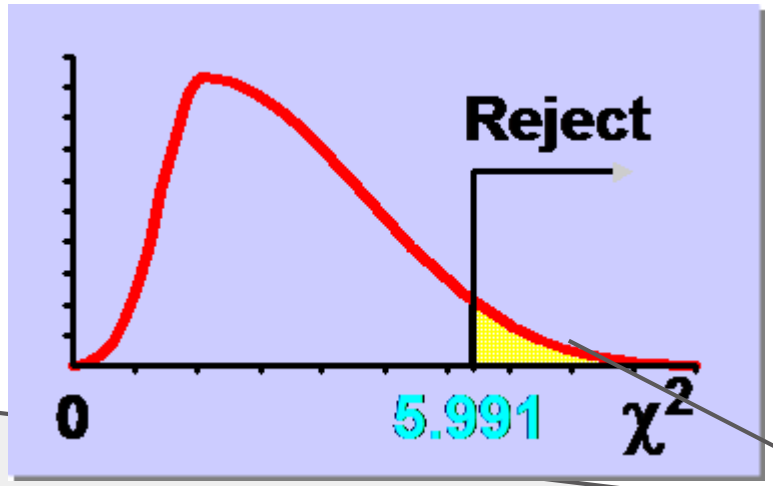
- Διατύπωση υποθέσεων, μηδενικής H_0 και εναλλακτικής H_a

H0: Καρκίνος και επάγγελμα ανεξάρτητες μεταβλητές

H1: Υπάρχει στατιστικά σημαντική σχέση

- Καθορισμός επιπέδου σημαντικότητας α = 0,05
- β.ε. = (*r*-1)*(*c*-1)=(2-1)*(3-1)=2

Απόφαση:
Απόρριψη σε *a* = 0,05.
Πιθανή επίδραση του επαγγέλματος στην ανάπτυξη καρκίνου ουροδόχου κύστης



$\chi 2 = 7,73$

Ο σχετικός λόγος συμπληρωματικών πιθανοτήτων (Odds Ratio-OR)

Σε μια έρευνα ασθενών-μαρτύρων επιλέγουμε ασθενείς και άτομα χωρίς το υπό μελέτη νόσημα (μάρτυρες) και συγκρίνουμε τη συχνότητα του υπό μελέτη παράγοντα στις 2 ομάδες.

Παράγοντας	Νόσημα	
	Ναι	Όχι
Ναι	a	b
Όχι	c	d

$$OR=(a*d)/(b*c)$$

OR≈1: απουσία συσχέτισης
OR>1: παρουσία του παράγοντα είναι επιβαρυντική (δηλαδή τα «εκτεθειμένα» άτομα έχουν μεγαλύτερη πιθανότητα να πάθουν τη νόσο σε σχέση με τα «μη εκτεθειμένα»)
OR<1: η παρουσία του παράγοντα είναι προστατευτική (δηλαδή τα «εκτεθειμένα» άτομα έχουν μικρότερη πιθανότητα να πάθουν τη νόσο σε σχέση με τα «μη εκτεθειμένα»)

Παράδειγμα 8

Σχετίζεται η απόκτηση παιδιών με τον καρκίνο των ωοθηκών;

	Καρκίνος ωοθηκών		
Αριθμός παιδιών	Ναι	Όχι	Σύνολο
0	48	30	78
1+	152	150	302
Σύνολο	200	180	380

$\chi^2(1)=3,13$, p-value=0,077, Οριακά στατιστικά σημαντική σχέση

OR= (48 X 150)/ (30 X 152) = 1,58

Οι γυναίκες χωρίς παιδιά έχουν 1,58 φορές μεγαλύτερο κίνδυνο για εμφάνιση καρκίνου των ωοθηκών σε σχέση με τις με τις γυναίκες με παιδιά ή αλλιώς 58% μεγαλύτερο κίνδυνο

Βιβλιογραφία

1) Δαμιανού Χ. , Κούτρας Μ. (1991-1998) *Εισαγωγή στη Στατιστική, Μέρος Ι και ΙΙ*. Εκδόσεις ΣΥΜΜΕΤΡΙΑ, Αθήνα

2) Δαμιανού, Χαράλαμπος Χ. , Παπαδάτος, Νικόλαος Δ. ,Χαραλαμπίδης, Χαράλαμπος Α. (2010) *Εισαγωγή στις Πιθανότητες και τη Στατιστική*. Εκδόσεις ΣΥΜΜΕΤΡΙΑ, Αθήνα

3) Introduction to Biostatistics, by Larry Winner, University of Florida, 2004 (http://www.stat.ufl.edu/~winner/sta6934/st4170_int.pdf)

4) An introduction to Statistical Inference and its Applications, Michael W. Trosset, College of William & Mary, 2006 (http://www.astrohandbook.com/ch17/statistical_inference.pdf) [νεότερη έκδοση: http://mypage.iu.edu/~mtrosset/StatInfeR.html]

5) A series of tutorials in biostatistics published in *British Medical Journal (BMJ) by* Altman and Bland (http://www.compgen.org/material/biostatistics#TOC-A-series-of-tutorials-in-biostatistics-published-in-British-Medical-Journal-BMJ-)