1

Παλινδρόμηση και συσχέτιση

Τεχνικές που επιτρέπουν την μελέτη της αλληλεπίδρασης ανάμεσα σε δύο μεταβλητές

Ποσοτικές τεχνικές (σημείου ή διαστήματος)

ΜΑΘΗΜΑ 5

1) Διερεύνηση σχέσης μεταξύ δύο μεταβλητών

Η σχέση ανάμεσα σε μία ανεξάρτητη μεταβλητή x και σε μία εξαρτημένη μεταβλητή y μπορεί να είναι:

- Γραμμική: $y = b_0 + b_1 x$
- Παραβολική: $y = b_0 + b_1 x + b_2 x^2$
- Εκθετική: $y = b_0 b_1^{x}$ ή
  $\ln y = \ln b_0 + x \ln b_1$
- Λογαριθμική: $y = b_0 + b_1 \ln x$

Γραμμική σχέση με πολλές ανεξάρτητες μεταβλητές

$$y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_n x_n$$

2) Μέτρηση του βαθμού συσχέτισης

Συντελεστής συσχέτισης Pearson.

② Ποιοτικές μέθοδοι

Ιεραρχικής κλίμακας (μέτρηση έντασης
συσχέτισης: συντελεστής συσχέτισης
Spearman, συντελεστής Gamma,
συντελεστές tau-b, tau-c)

Ονομαστικής κλίμακας

χ^2, λόγος πιθανοφάνειας, συντελεστής phi

ΠΑΛΙΝΔΡΟΜΗΣΗ

Απλή (για ανεξάρτητη μεταβλητή)
Πολλαπλή (πολλές ανεξάρτητες μεταβλητές

αναφέρεται σε δύο μετα-
προσδιορίζει τη σχέση
βλητές με σκοπό τη δυνατότητα πρόβλεψης
και ελέγχου.
(π.χ. πρόβλεψη εξόδων με βάση το εισόδημα)

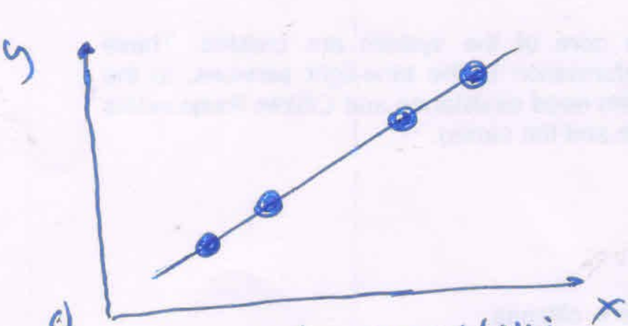
Απλή παλινδρόμηση

μεταβλητές x και y με την
Έστω δύο μεταβλητή y να εξαρτάται
εξαρτημένη μεταβλητή x μέσω μιας
από την ανεξάρτητη μεταβλητή αντιστοιχού
σχέσης 1-προς-1. (εάν σε ένα x αντιστοιχούν πολλά y η εξάρτηση λέγεται στοχαστική ή
στατιστική)

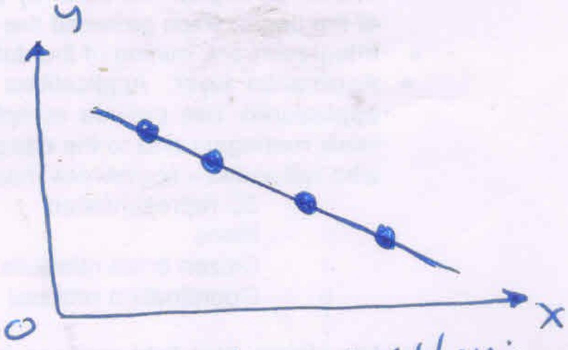
Γραμμική παλινδρόμηση

Έστω ν διατεταγμένα ζεύγη της μορφής (x_i, y_i) $(i = 1, 2, 3, \ldots, \nu)$. Εάν τα ζεύγη αυτά θεωρούμενα ως σημεία απεικονιστούν στο επίπεδο x-y λαμβάνουμε ένα νέφος σημείων ή διάγραμμα διασποράς (scatter plot).

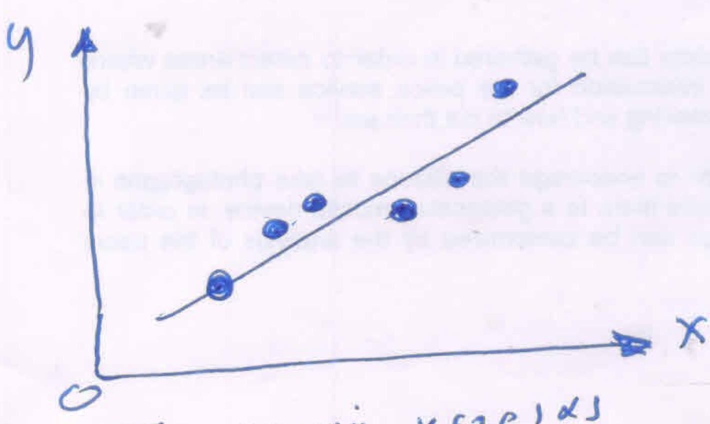
Εάν η εξάρτηση ανάμεσα στα σημεία είναι γραμμικής φύσεως η γραμμή που προκύπτει έχει θετική ή αρνητική κλίση με ισχυρό ή ασθενή βαθμό συσχέτισης. Παραδείγματα



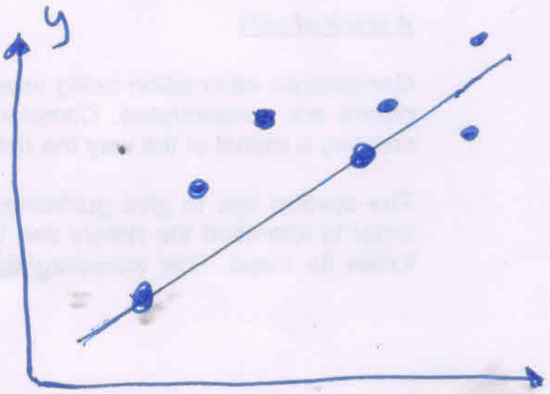
εντελώς γραμμική θετική κλίση



εντελώς γραμμική αρνητική κλίση



Γραμμική μετρίας εντάσης



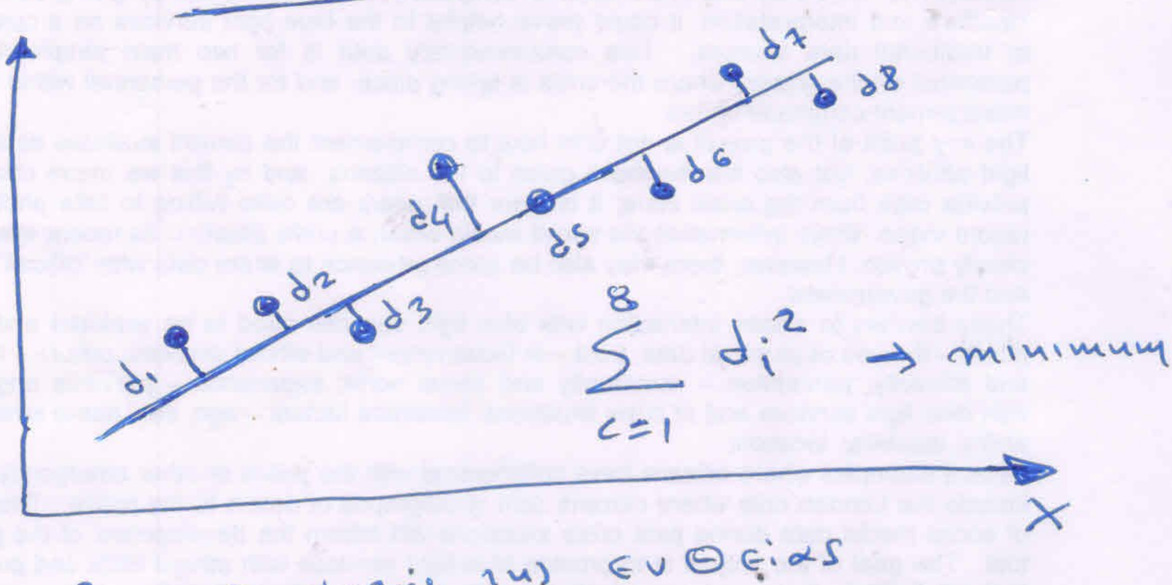
Γραμμική μικρής εντάσης

4 Ευθεία ελαχίστων τετραγώνων

Έστω σημεία (x_i, y_i) $(i=1,2,3,\dots,\nu)$ που φαίνεται πως σχετίζονται μεταξύ τους με γραφικό τρόπο και κατά συνέπεια σχετίζονται με ευθεία. Αναζητούμε την ευθεία εκείνη η οποία χαρακτηρίζεται από την επόμενη ιδιότητα:

Το άθροισμα των τετραγώνων των αποστάσεων δ_i $(i=1,2,3,\dots,\nu)$ των παραπάνω σημείων από αυτή την ευθεία είναι ελάχιστο

Γεωμετρική ερμηνεία



Εάν η εξίσωση αυτής της ευθείας έχει τη μορφή

$$y = b_0 + b_1 x$$

Το πρόβλημα ανάγεται στην εύρεση των τιμών των συντελεστών b_0 και b_1

Απλά δεδομένα ν ζεύγη (x_i, y_i)

Οι τιμές των b_0 και b_1 προκύπτουν από τη λύση του συστήματος $(\nu = N)$

$$\sum_{i=1}^{N} y_i = N b_0 + b_1 \sum_{i=1}^{N} x_i$$

$$\sum_{i=1}^{N} x_i y_i = b_0 \sum_{i=1}^{N} x_i + b_1 \sum_{i=1}^{N} x_i^2$$

από όπου προκύπτει ότι (παραλείπονται οι δείκτες στο άθροισμα για ευκολία)

$$b_1 = \frac{\sum x_i y_i - N\bar{X}\bar{Y}}{\sum x_i^2 - N\bar{X}^2}$$ και στη συνέχεια

$$b_0 = \bar{Y} - b_1 \bar{X}$$ όπου $\bar{X}$ και $\bar{Y}$ οι

μέσοι όροι των x_i και y_i αντίστοιχα.

Στις παραπάνω εκφράσεις το b_1 ονομάζεται **συντελεστής παλινδρόμησης** ή κλίση της ευθείας και εκφράζει τη μεταβολή του y όταν το x μεταβληθεί κατά μία μονάδα.

- $b_1 > 0$ θετική εξάρτηση
- $b_1 < 0$ αρνητική εξάρτηση

Για κάθε τιμή του x_i υφίστανται δύο τιμές y, η τιμή y_i που μαζί με το x_i σχηματίζουν το ζεύγος (x_i, y_i) που χρησιμοποιήθηκε στον παραπάνω υπολογισμό και η τιμή $\hat{y}_i$ που προκύπτει αντικαθιστώντας την τιμή του x_i στην εξίσωση της ευθείας ελαχίστων τετραγώνων – θα είναι δηλαδή

$$\hat{y}_i = b_0 + b_1 x_i$$

Η διαφορά $d_i = y_i - \hat{y}_i$ δεν είναι παρά το σφάλμα που θεωρήσαμε προηγουμένως.

ΠΑΡΑΔΟΧΕΣ

- Το σφάλμα d_i αποτελεί μια τυχαία μεταβλητή με μηδενική μέση τιμή
- Η διακύμανση του σφάλματος είναι η ίδια για όλα τα x
- Οι τιμές του σφάλματος είναι ανεξάρτητες μεταξύ τους
- Το σφάλμα χαρακτηρίζεται από κανονική κατανομή

Η ευθεία ελαχίστων τετραγώνων χρησιμοποιείται για πρόβλεψη μέσω παρεμβολής

Εάν το πλήθος των ζευγών σημείων (7)
(x_i, y_i) είναι πάρα πολύ μεγάλο, προχωρούμε
ως γνωστόν στην ομαδοποίησή τους. Έστω

f_{ij} το πλήθος των ζευγών (x_i, y_j)

f_{x_i}: το πλήθος των ζευγών (x_i, α) που
έχουν ως πρώτο στοιχείο το x_i ανεξάρτητα από
το y_j

f_{y_j} το πλήθος των ζευγών (α, y_j) που έχουν
ως δεύτερο στοιχείο το y_j ανεξάρτητα από το
x_i. Θα είναι τότε

$$b_1 = \frac{N\sum\sum f_{ij}x_i y_j - \left(\sum f_{x_i} x_i\right)\left(\sum f_{y_j} y_j\right)}{N\sum f_{x_i} x_i^2 - \left(\sum f_{x_i} x_i\right)^2}$$

και

$b_0 = \bar{y} - b_1\bar{x}$ όπου $\bar{x}$ και $\bar{y}$ οι μέσοι όροι
των x_i και y_j αντίστοιχα

Η προσαρμογή των σημείων (x_i, y_i) εκτός
από ευθεία μπορεί να γίνει και σε άλλες
καμπύλες όπως <u>παραβολή</u>, <u>υπερβολή</u> ή
<u>εκθετική</u> καμπύλη.

Προσαρμογή παραβολής σε δεδομένα

Έστω σύνολο σημείων $M_i\,(x_i, y_i)$ $(i=1,2,..N)$.
Ζητείται παραβολή με εξίσωση

$$y = b_0 + b_1 x + b_2 x^2$$

η οποία να προσαρμόζεται στα δεδομένα και να οδηγώνται στο μικρότερο δυνατό σφάλμα.

Εδώ οι συντελεστές b_1 και b_2 ονομάζονται συντελεστές παλινδρόμησης ενώ b_0 είναι ένας σταθερός όρος.

Εδώ οι παραπάνω συντελεστές προκύπτουν από τη λύση του συστήματος

$$\sum y_i = N b_0 + b_1 \sum x_i + b_2 \sum x_i^2$$

$$\sum x_i y_i = b_0 \sum x_i + b_1 \sum x_i^2 + b_2 \sum x_i^3$$

$$\sum x_i^2 y_i = b_0 \sum x_i^2 + b_1 \sum x_i^3 + b_2 \sum x_i^4$$

Ποιο είναι το καλύτερο σχήμα καμπύλης? Υπάρχουν κάποια κριτήρια αλλά γενικά αρκεί μια απλή παρατήρηση του διαγράμματος διασποράς. Εάν πχ διαπιστώσουμε πως τα σημεία σχηματίζουν ευθεία ή παραβολή εργαζόμαστε αναλόγως.

Μέσο τετραγωνικό σφάλμα

Αποτελεί τον αριθμητικό μέσο των τετραγώνων των σφαλμάτων d_i.

Για ευθεία με αταξινόμητα δεδομένα έχουμε

$$\sigma^2 = \frac{1}{N} \sum d_i^2 = \frac{1}{N} \sum (y_i - \hat{y}_i)^2 =$$

$$= \frac{1}{\nu} \left(\sum y_i^2 - b_0 \sum y_i - b_1 \sum x_i y_i \right)$$

Για ταξινομημένα δεδομένα η εξίσωση είναι

$$\sigma^2 = \frac{1}{N} \left(\sum_j f_{y_j} y_j^2 - b_0 \sum_j f_{y_j} y_j - \right.$$

$$\left. - b_1 \sum_i \sum_j f_{ij} x_i y_j \right)$$

Για την παραβολή έχουμε

$$\sigma^2 = \frac{1}{\nu} \left(\sum y_i^2 - b_0 \sum y_i - b_1 \sum x_i y_i - \right.$$

$$\left. - b_2 \sum x_i^2 y_i \right)$$

Η τετραγωνική ρίζα σ του μέσου τετραγωνικού σφάλματος ονομάζεται <u>τυπικό σφάλμα εκτίμησης</u>.

Συσχέτιση Μέτρα την ένταση του βαθμού εξάρτησης ανάμεσα σε μεταβλητές.

Συνδιακύμανση (covariance)

Μέτρο της γραμμικής συνεργασίας μεταξύ δύο ποσοτικών μεταβλητών X και y.

Έστω μεταβλητή X με δείγματα $x_1, x_2, \ldots, x_N$ και μεταβλητή y με δείγματα $y_1, y_2, \ldots, y_N$

Ορίζουμε (για αταξινόμητα δεδομένα)

$$Cov(x, y) = \frac{1}{N} \sum (x_i - \bar{x})(y_i - \bar{y}) =$$

$$\frac{1}{N} \sum x_i y_i - \bar{x}\bar{y}$$

όπου $\bar{x}$ και $\bar{y}$ οι μέσοι όροι των X και y.

Εάν τα δεδομένα είναι ταξινομημένα θα έχουμε

$$cov(x, y) = \frac{1}{N} \sum_i \sum_j f_{ij} (x_i - \bar{x})(y_j - \bar{y}) =$$

$$\frac{1}{N} \sum_i \sum_j f_{ij} x_i y_i - \bar{x}\bar{y}$$

Εάν τα δεδομένα σχετίζονται με γραμμικό τρόπο, ο συντελεστής b_1 της ευθείας ελαχίστων τετραγώνων γράφεται ως $b_1 = \frac{cov(x,y)}{\sigma_x^2}$

όπου $\sigma_x^2 = cov(x, x)$ η διακύμανση ή μεταβλητότητα του x. Από τον ορισμό είναι

$$\sigma_x^2 = \frac{1}{N} \sum x_i^2 - (\bar{x})^2$$

Επειδή είναι $\sigma_x^2 > 0$, το πρόσημο του συντελεστή b_1 εξαρτάται από το πρόσημο του $cov(x,y)$. Λαμβάνοντας υπόψη ότι το b_1 είναι η κλίση της ευθείας ελαχίστων τετραγώνων, έχουμε:

- Εάν είναι $cov(x,y) > 0$ οι μεταβλητές μεταβάλλονται ανάλογα (η αύξηση της μιας, οδηγεί σε αύξηση της άλλης).
- Εάν είναι $cov(x,y) < 0$ οι μεταβλητές μεταβάλλονται αντιστρόφως ανάλογα (η αύξηση της μιας οδηγεί σε μείωση της άλλης).
- Εάν είναι $cov(x,y) = 0$ οι μεταβλητές είναι ασυσχέτιστες μεταξύ τους.

ΙΔΙΟΤΗΤΕΣ

$cov(X,Y) = cov(Y,X)$, $cov(X,-Y) = -cov(X,Y)$

$cov(X+Y,Z) = cov(X,Z) + cov(Y,Z)$

$Z = \alpha + bX,\ W = c + dY \Rightarrow cov(Z,W) = bd\, cov(X,Y)$

Η χρησιμότητα της συνδιακύμανσης

cov(X,Y) είναι περιορισμένη, επειδή συνοδεύεται από τις μονάδες των X και Y, εκτός εάν είναι cov(X,Y) = 0 οπότε οδηγούμαστε στο συμπέρασμα πως οι μεταβλητές είναι ασυσχέτιστες

Συντελεστής συσχέτισης

$$r = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{cov(X,Y)}{\sqrt{\sigma_X^2 \sigma_Y^2}} =$$

$$= \frac{cov(X,Y)}{\sqrt{cov(X,X)\,cov(Y,Y)}}$$

Για αταξινόμητα δεδομένα έχουμε

$$r = \frac{(1/N)\sum x_i y_i - \bar{X}\bar{Y}}{\sqrt{\frac{\sum x_i^2}{N} - \bar{X}^2}\sqrt{\frac{1}{N}\sum y_i^2 - \bar{Y}^2}} =$$

$$= \frac{N\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{N\sum x_i^2 - (\sum x_i)^2}\sqrt{N\sum y_i^2 - (\sum y_i)^2}} =$$

$$= \frac{\sum x_i y_i - N\bar{X}\bar{Y}}{N\sigma_X \sigma_Y}$$ τύπος του pearson

Για ταξινομημένα δεδομένα έχουμε

$$r = \frac{N \sum \sum f_{ij} x_i y_j - \sum f_{x_i} x_i \sum f_{y_j} y_j}{\sqrt{N \sum f_{x_i} x_i^2 - \left(\sum f_{x_i} x_i\right)^2} \sqrt{N \sum f_{y_j} y_j^2 - \left(\sum f_{y_j} y_j\right)^2}}$$

Ιδιότητες συντελεστή συσχέτισης

- Είναι αδιάστατο μέγεθος
- Παίρνει τιμές στο διάστημα $[-1, 1]$
- Για $r = +1$ ή $r = -1$ έχουμε τέλεια συσχέτιση (γραμμική μορφή θετική ή αρνητική)
- Εάν $r = 0 \Rightarrow$ οι μεταβλητές είναι ασυσχέτιστες

ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ ΠΙΘΑΝΟΤΗΤΩΝ (1)

Σύνολα: ορισμός, υποσύνολο, καθολικό, ένωση, τομή, διαφορά, καρτεσιανό γινόμενο

κενό σύνολο

ιδιότητες ένωσης και τομής

προσεταιριστική

επιμεριστική

μεταθετική, ουδέτερο στοιχείο

ξένα σύνολα

$\boxed{\text{ΜΑΘΗΜΑ } 6}$

κανόνες de Morgan

$$\overline{A\cup B}=\bar{A}\cap\bar{B}$$

$$\overline{A\cap B}=\bar{A}\cup\bar{B}$$

Στατιστική τυχαιότητα, χάος

Πείραμα τύχης

ορίζεται ως μια διαδικασία η οποία αν και επαναλαμβάνεται με τον ίδιο τρόπο και κάτω από τις ίδιες ακριβώς συνθήκες δεν μπορούμε να προβλέψουμε το αποτέλεσμα που προκύπτει από μία επανάληψη

πχ ρίψη ζαριού

(διαμάχη Bohr - Einstein)

Δειγματοχώρος Ω ενος πειράματος (2)

τυχης ονομαζεται το συνολο ολων των δυνατων αποτελεσματων του π.χ.

$$\text{Ζαρι} = \{1, 2, 3, 4, 5, 6\}$$

$$\text{Κερμα} = \{Κ, Γ\}$$ Κορωνα Γραμματα

Καθε δυνατο αποτελεσμα ενος πειραματος λεγεται απλο γεγονος ή ενδεχομενο

Ενώ

Γεγονός ή ενδεχομενο ονομαζεται καθε υποσυνολο Α του δειγματοχωρου Ω

πχ ζαρι:

Απλα ενδεχομενα: 1, 2, 3, 4, 5, 6

Γεγονοτα: πχ

$$A = \{1, 4\} \quad B = \{3, 5, 1\}$$

Δειγματοχωρος 2 νομισμάτων

$$\Omega = \{ΚΚ, ΚΓ, ΓΚ, ΓΓ\}$$ διακριση κερματων

ή

$$\Omega = \{ΚΚ, ΚΓ, ΓΓ\}$$ χωρις διακριση κερματων

οποτε $ΚΓ = ΓΚ$

ΟΡΙΣΜΟΙ

Αν σε μια επανάληψη ενός πειράματος προκύψει το αποτέλεσμα $\omega \in \Omega$

ΤΟΤΕ

- Αν είναι $\omega \in A$ λέμε πως πραγματοποιείται το γεγονός A
- Αν είναι $\omega \notin A$ λέμε πως δεν πραγματοποιείται το γεγονός A
- Αν είναι $\omega \in A$ και $\omega \in B$ ή ισοδύναμα εάν $\omega \in (A \cap B)$ λέμε πως πραγματοποιούνται ταυτόχρονα τα γεγονότα A και B.
- Αν είναι $\omega \in A$ ή $\omega \in B$ ή ισοδύναμα εάν $\omega \in (A \cup B)$ λέμε πως πραγματοποιείται τουλάχιστον ένα από τα A και B
- Αν είναι $\omega \in A$ και $\omega \notin B$ ή ισοδύναμα $\omega \in (A-B)$ λέμε πως πραγματοποιείται το A αλλά όχι το B
- Αν είναι $\omega \notin A$ και $\omega \notin B$ ή ισοδύναμα $\omega \in \overline{A} \cap \overline{B}$ ή $\omega \in \overline{A \cup B}$ λέμε πως δεν πραγματοποιείται κανένα από τα A και B

(4) Αναφορά στην έννοια της συνάρτησης

$f: A \longmapsto B$ πεδίο ορισμού
σύνολο τιμών

$y = f(x)$, εικόνα, 1 προς 1, επί

ΠΙΘΑΝΟΤΗΤΑ

Έστω ότι ρίχνουμε ένα κέρμα N φορές και καταγράφουμε το πλήθος K του αποτελέσματος "Γράμματα"

Ο λόγος $\boxed{f = \frac{K}{N}}$ ονομάζεται σχετική συχνότητα

Η επί τοις 100 σχετική συχνότητα ορίζεται ως

$$f_i \% \Rightarrow f_i \times 100$$

για το γεγονός f_i.

Παράδειγμα Έστω ότι ρίχνουμε ένα κέρμα 100 φορές και τις 63 φορές έχουμε "γράμματα".

Άρα $f_\gamma = \frac{63}{100} = 0.63 \Rightarrow f_\gamma \% = 63$

Επειδή το άλλο αποτέλεσμα θα είναι μόνο "κορώνα" $\Rightarrow f_x = \frac{37}{100} = 0.37$.

Παράδειγμα

Έστω ότι ρίχνουμε ένα ζάρι 50 φορές και έχουμε

1 17 φορές $\Rightarrow f_1 = \frac{17}{50} = 0.34$

2 8 φορές $\Rightarrow f_2 = 8/50 = 0.16$

3 7 φορες $\Rightarrow f_3 = 7/50 = 0.14$

4 10 φορες $\Rightarrow f_4 = 10/50 = 0.20$

5 3 φορες $\Rightarrow f_5 = 3/50 = 0.06$

6 5 φορες $\Rightarrow f_6 = 6/50 = 0.12$

Παρατηρούμε ότι

$$f_1 + f_2 + f_3 + f_4 + f_5 + f_6 = \frac{17+8+7+10+3+5}{50} = \frac{50}{50} = 1$$

Άρα $F = \sum_{i \in \Omega} f_i = 100\%$ (βεβαιότητα)

Είναι βέβαιο πως το αποτέλεσμα θα είναι ή 1 ή 2 ή 3 ή 4 ή 5 ή 6.

Ας επιστρέψουμε στο πείραμα του κέρματος και έστω πως η σχετική συχνότητα για μεγάλα είναι

Ριψεις	Σχετική συχνότητα
20	0.652
40	0.574
60	0.510
80	0.489
100	0.401
120	0.509 .....

Αυτή η σχετική συχνότητα ονομάζεται

«πιθανότητα να προκύψουν γράμματα»

Ορισμός ορίζουμε ως πιθανότητα σε έναν δειγματοχώρο Ω για συνάρτηση που αντιστοιχεί σε κάθε γεγονός $A \in \Omega$ έναν θετικό αριθμό $P(A)$ τέτοιον ώστε

1) $P(A) \geq 0$

2) $P(\Omega) = 1$

Για δύο γεγονότα ξένα μεταξύ τους ανά 2 δηλαδή τέτοια ώστε $A_1 \cap A_2 = \emptyset$ είναι

$$P(A_1 \cup A_2) = P(A_1) + P(A_2)$$

Γενικά

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = P(A_1) + P(A_2) + \dots + P(A_n)$$

Δηλαδή Η πιθανότητα εμφάνισης ενός ενδεχομένου A είναι ίση με το άθροισμα των πιθανοτήτων των απλών ενδεχομένων από τα οποία αποτελείται.

Επομένως $P(A) = \sum_{\omega \in A} P(\omega)$

Εξ ορισμού

$$\sum_{\omega \in \Omega} P(\omega) = 1$$

Εάν τα στοιχεία ενός πεπερασμένου δειγματοχώρου Ω με πλήθος στοιχείων $n(\Omega)$ έχουν ίσες πιθανότητες (ισοπίθανος δειγματοχώρος) τότε η πιθανότητα ενός γεγονότος A του Ω με πλήθος $n(A)$ είναι

$$\boxed{P(A) = \frac{n(A)}{n(\Omega)}}$$

Νόμοι πιθανοτήτων

1) $P(\bar{A}) = 1 - P(A)$
2) $P(\emptyset) = 0$
3) $A \subseteq B \Rightarrow P(A) \leq P(B)$
4) $P(A) \leq 1$
5) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
6) $P(A - B) = P(A) - P(A \cap B)$

Κατανομή συχνοτήτων

ΜΑΘΗΜΑ 7

Η κατανομή συχνοτήτων δεν αποτελεί παρά έναν πίνακα στον οποίο αναφέρονται οι τιμές της τυχαίας μεταβλητής και οι αντίστοιχες συχνότητες εμφάνισης.

πχ

Αριθμός παιδιών ανά οικογένεια	Πλήθος οικογενειών
0	50
1	250
2	360
3	120
4	80
5	40
	900

Η σχετική συχνότητα (πχ 250/900 για οικογένειες με 1 παιδί) νοείται ως πιθανότητα. Ο αντίστοιχος πίνακας είναι γνωστός ως πίνακας πιθανοτήτων.

Οι τυχαίες μεταβλητές όπως αναφέραμε σε προηγούμενο μάθημα μπορεί να είναι είτε διακριτές όταν λαμβάνουν τιμές από ένα πεπερασμένο σύνολο ή άπειρες διακριτές μεταξύ τους, είτε συνεχείς όταν λαμβάνουν οποιαδήποτε τιμή σε ένα διάστημα ή σύνολο διαστημάτων.

Ασυνεχείς ή διακριτές μεταβλητές

Μέση (mean) ή **αναμενόμενη (expected) τιμή** ή μαθηματική ελπίδα ή προσδοκία (expectation)

$$E(X) = \mu = \sum x P(x)$$

όπου $P(x)$ η πιθανότητα εμφάνισης της τιμής X.

Διακύμανση (variance)

$$\sigma^2 = \sum \left[x^2 \cdot P(x) \right] - \mu^2$$

Τυπική απόκλιση $\sigma = \sqrt{\sigma^2}$

Κατανομές πιθανότητας

1) **Διωνυμική κατανομή** (Binomial distribution) ή διεργασία Bernoulli

$$X \sim B(n, p)$$

- Χρησιμοποιείται σε σειρά πανομοιότυπων πειραμάτων όπως είναι η ρίψη ενός νομίσματος.
- Για κάθε πείραμα υπάρχουν μόνο δύο αποτελέσματα: επιτυχία (με πιθανότητα p) ή αποτυχία (με πιθανότητα q). Άρα $p+q=1$
- Τα πειράματα είναι ανεξάρτητα μεταξύ τους.
- Η τυχαία μεταβλητή X μετρά το πλήθος των επιτυχιών και μπορεί να πάρει τιμές 0 έως n.

Η πιθανότητα να έχουμε ακριβώς x επιτυχίες σε n επαναλήψεις του πειράματος — δίνεται από την έκφραση

$$P(X=x) = \binom{n}{x} p^{x} q^{n-x} \qquad x = 0,1,2,\ldots,n$$

όπου n το πλήθος των επαναλήψεων και

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!}$$

<u>Παράμετροι</u>

Μέση τιμή $E(x) = np$

Διακύμανση $Var(x) = \sigma^2 = npq$

Τυπική απόκλιση $\sigma = \sqrt{npq}$

<u>ΠΑΡΑΔΕΙΓΜΑ</u>

Το 5% των οδηγών φορτηγών στην Αμερική είναι γυναίκες.
Επιλέγουμε στην τύχη 10 οδηγούς.
Εφόσον το αποτέλεσμα είναι Άνδρας / Γυναίκα η κατανομή είναι διωνυμική

1) Ποια είναι η πιθανότητα ώστε 2 οδηγοί να είναι γυναίκες?

2) Ποια είναι η πιθανότητα κανένας οδηγός να μην είναι γυναίκα?

Εφόσον ενδιαφερόμαστε για γυναίκες οδηγούς και υπάρχει 5% πιθανότητα να συναντήσουμε γυναίκα οδηγό είναι $\underline{p = 0.05}$.

Άρα $\underline{q = 1-p = 0.95}$.

Εφόσον επιλέγουμε 10 οδηγούς είναι $n = 10$

Άρα για $\underline{x=2}$

$$P(X=2) = \binom{10}{2}(0.05)^2(0.95)^8 =$$

$$= \frac{10!}{2!\,(10-2)!}(0.025)(0.66342) = 7.46\%$$

Για $\underline{x=0}$

$$P(X=0) = \binom{10}{0}(0.05)^0(0.95)^{10} = 59.87\%$$

2) **Κατανομή Poisson**

Ασυνεχής κατανομή που χρησιμοποιείται για των εκτίμηση συμβάντων που έχουν λάβει χώρα σε συγκεκριμένο χρονικό διάστημα.

πχ X = αριθμός αυτοκινήτων που επισκέφτηκαν ένα πλυντήριο σε διάστημα μιας ώρας

$X \sim P(\lambda)$ λ ρυθμός άφιξης

ορισμός

$$P(X=x) = \lambda^x \frac{e^{-\lambda}}{x!}$$

- Η πιθανότητα να συμβεί ένα γεγονός είναι η ίδια για δύο διαστήματα ίσου πλάτους.
- Η πραγματοποίηση ή όχι ενός γεγονότος σε ένα διάστημα είναι ανεξάρτητη από την πραγματοποίηση του ή όχι σε κάποιο άλλο διάστημα.

Ιδιότητες κατανομής Poisson

Μέση τιμή $E(X) = \lambda = np$

Διακύμανση $Var(X) = \sigma^2 = \lambda$

Τυπική απόκλιση $\sigma = \sqrt{\lambda}$.

ΠΑΡΑΔΕΙΓΜΑ

Ταξιδιώτες φτάνουν τυχαία και ανεξάρτητα στο σταθμό ελέγχου ενός αεροδρομίου. Ο μέσος όρος αφίξεων είναι $\lambda = 10 \ \frac{\text{ταξιδιώτες}}{\text{λεπτό}}$.

- Ποια είναι η πιθανότητα να μην φτάσει κανένας ταξιδιώτης κατά τη διάρκεια ενός λεπτού?

$$P(X=0) = 10^0 \frac{e^{-10}}{0!} = 0.000045399 = 0.00454\ \%$$

- Ποια είναι η πιθανότητα να φτάσουν το πολύ τρεις ταξιδιώτες?

$$P(X \le 3) = P(X=0) + P(X=1) + P(X=2) + P(X=3) =$$

$$= 10^0 \frac{e^{-10}}{1!} + 10^1 \frac{e^{-10}}{1!} + 10^2 \frac{e^{-10}}{2!} + 10^3 \frac{e^{-10}}{3!} = 1.03394\ \%$$

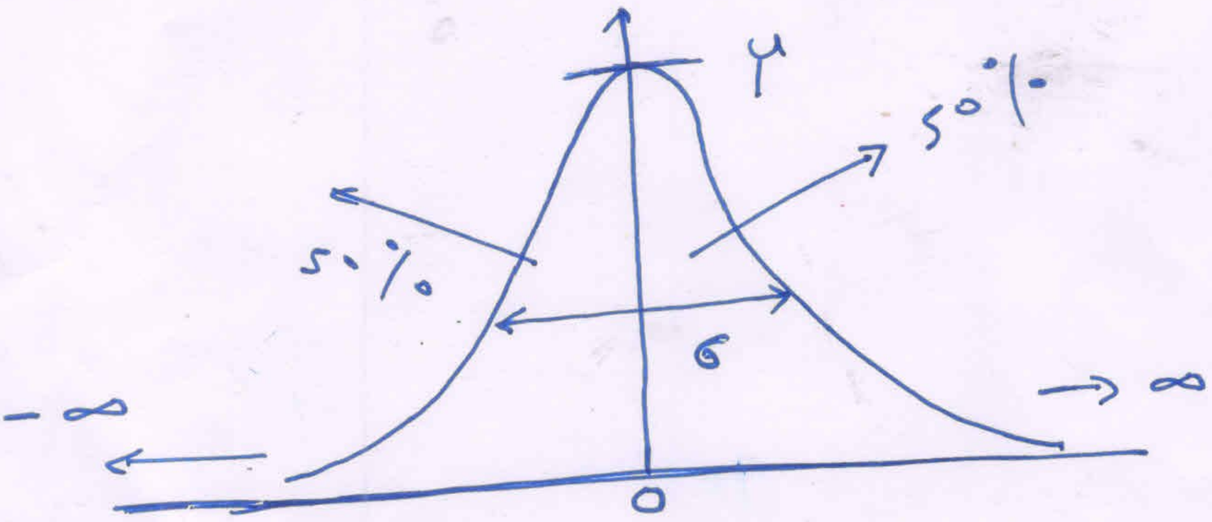
Κανονική κατανομή ή κατανομή Gauss

Αποτελεί μία από τις πιο σημαντικές κατανομές συνεχούς μεταβλητής και ορίζεται ως

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

όπου μ ο αριθμητικός μέσος όρος και σ η τυπική απόκλιση.

Η γραφική της παράσταση έχει τη μορφή



- Ο αριθμητικός μέσος όρος αποτελεί το υψηλότερο σημείο της καμπύλης και αποτελεί επίσης διάμεσο και σημείο μέγιστης συχνότητας. Μπορεί να λάβει θετική, αρνητική ή μηδενική τιμή.
- Η καμπύλη είναι κωδωνοειδής με μία απλή κορυφή που είναι συμμετρική ως προς την ευθεία του αριθμητικού μέσου $x = \mu$.

• Η καμπύλη τείνει ασυμπτωτικά στο μηδέν στα όρια $x \to \infty$ και $x \to -\infty$

• Η τυπική απόκλιση καθορίζει το πλάτος της καμπύλης ενώ το συνολικό εμβαδόν της είναι ίσο με τη μονάδα

• Το 86 % του εμβαδού αντιστοιχεί σε διάστημα $\pm\sigma$, το 95,44 σε διάστημα $\pm 2\sigma$ ενώ το 99,72 σε διάστημα $\pm 3\sigma$.

• Η πιθανότητα να έχει η μεταβλητή x τιμή σε ένα διάστημα είναι ίση με το αντίστοιχο εμβαδόν της καμπύλης.

Γράφουμε $\boxed{X \sim N(\mu, \sigma)}$

Η πιθανότητα για ένα διάστημα $[\alpha, \beta]$ υπολογίζεται ως

$$P(\alpha \le X \le \beta) = \int_{\alpha}^{\beta} f(x)\,dx$$

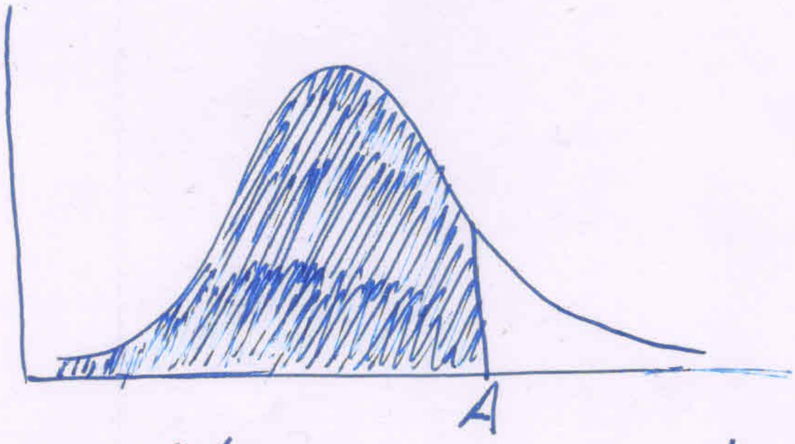
Προκειμένου να αποφύγουμε την διαδικασία υπολογισμού του ολοκληρώματος ορίζουμε τη μεταβλητή κατανομή που προκύπτει

$$Z = \frac{X-\mu}{\sigma}$$

με την αντίστοιχη κατανομή που προκύπτει να χαρακτηρίζεται από όλες τις ιδιότητες της αρχικής κατανομής αλλά να διαθέτει μηδενικό αριθμητικό μέσο όρο και μοναδιαία τυπική απόκλιση $(\sigma = 1)$

Γράφουμε $Z \sim N(0,1)$

Το βασικό πλεονέκτημα αυτής της νέας κατανομής είναι πως οι τιμές της μπορούν να βρεθούν σε πίνακες και δεν απαιτείται ο υπολογισμός του ολοκληρώματος.



$P(X \leq A)$ = το σκιασμένο εμβαδόν

Δειγματοληπτική κατανομή αριθμητικού μέσου

Έστω πληθυσμός N και δείγμα πλήθους n που χρησιμοποιήθηκε για την διεξαγωγή κάποιας έρευνας. Για να πάρουμε ακριβή αποτελέσματα επαναλαμβάνουμε τη διαδικασία πολλές φορές θεωρώντας πολλά δείγματα από τον ίδιο πληθυσμό με το ίδιο μέγεθος n. Το κάθε δείγμα έχει διαφορετικό αριθμητικό μέσο $\bar{X}$ και η κατανομή πιθανότητας όλων των πιθανών τιμών του δειγματικού μέσου ονομάζεται δειγματική κατανομή.

ΜΑΘΗΜΑ 8

Παράμετροι δειγματοληπτικής κατανομής αριθμητικού μέσου

Αναμενόμενη τιμή του $\bar{X}$: $E(\bar{X}) = \mu$

(η μέση τιμή του πληθυσμού από τον οποίο ελήφθη το δείγμα).

Τυπική απόκλιση του $\bar{x}$

Εάν είναι $n/N > 0.05 \Rightarrow \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

Εάν $N \to \infty$ ή N πεπερασμένο με

$n/N \le 0.05$ τότε $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

όπου : σ η τυπική απόκλιση του πληθυσμού
n το κοινό μέγεθος των επαναλαμβανόμενων δειγμάτων και
N το μέγεθος του πληθυσμού

Όταν το n είναι μεγάλο, η δειγματοληπτική κατανομή του αριθμητικού μέσου $\bar{x}$ μπορεί να προσεγγίζεται από την κανονική κατανομή (central limit theorem).

Μεγάλο δείγμα $\iff n > 30$

Επιθυμητές ιδιότητες μεθόδων για στατιστική εκτίμηση

- **Αμεροληψία**: ο μέσος όρος της δειγματοληπτικής κατανομής των αριθμητικών μέσων που προέρχεται από τον ίδιο πληθυσμό είναι ο ίδιος με τον αριθμητικό μέσο του πληθυσμού.

- **Αποτελεσματικότητα**: Σχετίζεται με το μέγεθος του τυπικού σφάλματος.

- **Συνέπεια**: Ο εκτιμητής θεωρείται συνεπής, εάν καθώς αυξάνει το μέγεθος του δείγματος, τότε τα προς μέτρηση μέγεθος γίνεται ίσο με την τιμή του αντίστοιχου μέτρου του πληθυσμού.

- **Επάρκεια**: Σχετίζεται με τον όγκο των πληροφοριών που εμπλέκονται στη στατιστική εκτίμηση.

Τύποι Στατιστικής Εκτίμησης

Εκτίμηση σε σημείο (point estimation)

Εκτιμούμε την τιμή μιας άγνωστης παραμέτρου του πληθυσμού από μια μοναδική τιμή της αντίστοιχης παραμέτρου της δειγματοληπτικής κατανομής.

Σημειακοί εκτιμητές είναι ο αριθμητικός μέσος, η τυπική απόκλιση και το ποσοστό του δείγματος.

Η τιμή $|\bar{X}-\mu|$ ονομάζεται δειγματοληπτικό σφάλμα και όσο πιο μικρή είναι η τιμή του σφάλματος αυτού, τόσο καλύτερη είναι η εκτίμηση. (3)

Εκτίμηση σε διάστημα (interval estimation)

Η αληθινή τιμή μιας παραμέτρου του πληθυσμού αναμένεται να βρίσκεται μέσα σε ένα διάστημα αναμενόμενων τιμών το οποίο ονομάζεται διάστημα εμπιστοσύνης (confidence interval) $1-\alpha$ το οποίο υποδηλώνει την πιθανότητα να περιλαμβάνεται εντός του διαστήματος η πραγματική τιμή της παραμέτρου του πληθυσμού.

Η πιθανότητα σφάλματος είναι α $(0<\alpha<1)$ και ονομάζεται επίπεδο σημαντικότητας.

Εάν είναι $\alpha=0.05$, ο συντελεστής εμπιστοσύνης θα είναι 0.95 και το διάστημα εμπιστοσύνης ονομάζεται διάστημα εμπιστοσύνης 95 %.

ΔΙΑΣΤΗΜΑ ΕΜΠΙΣΤΟΣΥΝΗΣ ΓΙΑ ΤΗ ΜΕΣΗ ΤΙΜΗ ΠΛΗΘΥΣΜΟΥ

Στον υπολογισμό του λαμβάνονται υπόψη:

- Η μορφή της κατανομής του πληθυσμού (κανονική ή κατά προσέγγιση κανονική)
- Το μέγεθος του δείγματος (μικρό ή μεγάλο)
- Η γνώση ή όχι της τυπικής απόκλισης του πληθυσμού.

- Το εάν ο πληθυσμός N είναι πεπερασμένος ή άπειρος.

- Μεγάλο δείγμα $(n > 30)$, γνωστό σ:

Πεπερασμένος πληθυσμός, $n/N > 0.05$

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

$-$: κατώτατο όριο εμπιστοσύνης

$+$: ανώτατο όριο εμπιστοσύνης

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ τυπικό σφάλμα του μέσου

$E = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ μέγιστο σφάλμα εκτίμησης

Όταν $N \to \infty$ ή N πεπερασμένο με $n/N < 0.05$

Το διάστημα είναι

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- Μεγάλο δείγμα $(n > 30)$, σ άγνωστο

$$\bar{X} \pm Z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}, \quad \hat{\sigma} = S = \sqrt{\frac{\sum(X_i-\bar{X})^2}{n-1}}$$

Για $N \to \infty$ ή N πεπερασμένο που

$n/N < 0.05$:

Το διάστημα είναι

$$\bar{X} \pm Z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}$$

Επειδή <u>ΔΕΝ</u> γνωρίζουμε την τυπική απόκλιση του πληθυσμού, χρησιμοποιούμε την τυπική απόκλιση του δείγματος.

Εδώ μιλούμε για <u>εκτιμώμενο</u> τυπικό σφάλμα του μέσου $\sigma_{\bar{X}}$ και <u>εκτιμώμενο</u> μέγιστο σφάλμα εκτίμησης.

- Μικρό δείγμα ($n \leq 30$) κανονική ή κατά προσέγγιση κανονική κατανομή, σ άγνωστο

<u>Κατανομή Student's t (W.S. Gosset)</u>

1) Είναι συμμετρική όπως και η κανονική κατανομή
2) Είναι διαφορετική για κάθε διαφορετικό μέγεθος δείγματος επειδή διαφέρουν οι βαθμοί ελευθερίας
3) όσο πιο μεγάλο είναι το μέγεθος του δείγματος τόσο πιο ψηλή είναι η κορυφή της γραφικής παράστασης.
4) Σε σχέση με την κανονική κατανομή είναι πιο χαμηλή αλλά πιο πλατιά
5) όταν οι βαθμοί ελευθερίας αυξάνονται προσεγγίζει την κανονική κατανομή.

Διάστημα εμπιστοσύνης

$\bar{X} \pm t_{\alpha/2} \dfrac{\hat{\sigma}}{\sqrt{n}}$ όπου ξανά $\hat{\sigma} = s = \sqrt{\dfrac{\sum (X_i - \bar{X})^2}{n-1}}$

Το μέγεθος του δείγματος που πρέπει να επιλέξουμε για να έχουμε το μέγιστο επιθυμητό σφάλμα εκτίμησης E με δεδομένο α και γνωστή τυπική απόκλιση είναι

$$n = \frac{(Z_{\alpha/2})^2 \sigma^2}{E^2}$$

Αυτός ο τύπος χρησιμοποιείται και όταν δεν είναι γνωστή η τυπική απόκλιση αλλά έχουμε επιλέξει κάποια τιμή.

Σε όλα τα παραπάνω το εύρος του διαστήματος εμπιστοσύνης εξαρτάται από:

- Το μέγεθος του δείγματος.
- Τη διασπορά του πληθυσμού
- Το συντελεστή εμπιστοσύνης.

Γίνεται τόσο μικρότερο όσο

- μεγαλύτερο είναι το μέγεθος του δείγματος
- μικρότερη είναι η διασπορά
- μικρότερος είναι ο συντελεστής εμπιστοσύνης.

Παράδειγμα Α (Μεγάλο δείγμα, σ γνωστό) 7

Εταιρεία πώλησης καλλυκαθαριστήρων θέλει να εκτιμήσει τη μέση διάρκεια ζωής τους.

Δεδομένα Τυπική απόκλιση $\sigma = 6$ μήνες

$N \to \infty$ (κανονική κατανομή)

Μέση διάρκεια ζωής $\bar{x} = 21$ (μήνες)

$\alpha = 0.05 \Rightarrow 1 - \alpha = 0.95$

(Μεγάλο δείγμα, σ γνωστό).

ΛΥΣΗ $Z_{\alpha/2} = Z_{0.025} = 1.96$ (από πίνακα).

$$E = Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 1.96 \frac{6}{\sqrt{100}} = 1.176.$$

Άρα

$$\left.\begin{aligned} \bar{x}_{min} &= 21 - 1.176 = 19.824 \\ \bar{x}_{max} &= 21 + 1.176 = 22.276 \end{aligned}\right\} \text{ με βεβαιότητα } 95\%.$$

ΠΑΡΑΔΕΙΓΜΑ Β (Μεγάλο δείγμα, σ άγνωστο)

Μέγεθος πληθυσμού $N = 700$ οικογένειες

Μέγεθος δείγματος $n = 50$ οικογένειες (μεγάλο)

καταγραφή ετήσιου εισοδήματος με στόχο οικονομική ενίσχυση.

κανονική κατανομή, $1 - \alpha = 0.9 \Rightarrow \alpha = 0.1$

Από το δείγμα προκύπτει:

Μέσο εισόδημα $\bar{X} = 11800$ €, $\hat{\sigma} = 950$ €

Ποια είναι τα διαστήματα εμπιστοσύνης?

<u>ΛΥΣΗ</u> Είναι $n/N = \frac{50}{700} = 0.071 > 0.05$

<u>η μεγάλο</u>

<u>Άρα</u>

$$\bar{X} \pm Z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}, \quad \hat{\sigma} = s = \sqrt{\frac{\sum (x_i - x)^2}{n-1}}$$

$Z_{\alpha/2} = Z_{0.05} = 1.64$ (από πίνακες)

$$E = Z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} = 1.64 \frac{950}{\sqrt{50}} = 220.3$$

<u>Άρα</u>

$$\bar{x} - Z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = 1180 - 220.3 \sqrt{\frac{700-50}{700-1}}$$

$$= 11800 - 220.3 \times 0.9643 =$$

$$= 11.587.5$$

ομοίως για το (+):

$$\bar{x} + Z_{\alpha/2} = 11800 + 220.3 \times 0.9643 = 12.012.5$$

Άρα με βεβαιότητα 90%, το μέσο ετήσιο εισόδημα των οικογενειών είναι από

<u>11587.5 έως 12.012.5</u>

Διαφορά μέσων τιμών

Δυο δείγματα είναι εξαρτημένα ή εζευγμένα κατά ζεύγη, όταν το ένα επιλέγεται τυχαία και το άλλο έτσι ώστε να είναι ισότιμο με το πρώτο.

Π.χ. για ένα δείγμα 10 τυχαίων μαθητών καταγράφουμε:

- Βαθμό μαθήματος στατιστικής
- IQ
- απουσίες
- μορφωτικό επίπεδο γονέα

Στο δεύτερο δείγμα επιλέγουμε 10 μαθητές για τους οποίους ισχύουν οι ίδιες τιμές IQ, απουσιών και μορφωτικού επιπέδου γονέα και καταγράφουμε το βαθμό της στατιστικής

ΠΙΝΑΚΑΣ ΔΕΔΟΜΕΝΩΝ

Απουσίες	IQ	Μόρφωση	Βαθμ 1	Βαθμ 2
			X_{11}	X_{21}
			X_{12}	X_{22}
			X_{13}	X_{23}

$$d_i = X_{1i} - X_{2i}$$

Ανεξάρτητα δείγματα

Επιλέγονται με τυχαία δειγματοληψία

Ζευγαρωτά δεδομένα

Ο αριθμητικός μέσος των διαφορών είναι

$$\bar{d} = \frac{1}{n}\sum_i d_i$$

Η τυπική απόκλιση είναι

$$S_d = \sqrt{\frac{\sum d_i^2 - \frac{\left(\sum d_i\right)^2}{n}}{n-1}}$$

Διάστημα εμπιστοσύνης

$$\bar{d} \pm \underbrace{t_{\alpha/2}\frac{S_d}{\sqrt{n}}}$$

μέγιστο επιτρεπόμενο τυπικό σφάλμα

Ανεξάρτητα δείγματα

Μέσος όρος $\mu_{\bar{X}_1-\bar{X}_2} = \mu_1-\mu_2$

Τυπικό σφάλμα $\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$

Μεγάλα δείγματα ($n_1, n_2 > 30$) σ_1^2, σ_2^2 γνωστά

$$(\bar{X}_1-\bar{X}_2) \pm \underbrace{Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

E μέγιστο επιτρεπόμενο σφάλμα