ΚΕΦΑΛΑΙΟ 6

ΣΥΝΑΡΤΗΣΕΙΣ ΤΥΧΑΙΩΝ ΜΕΤΑΒΛΗΤΩΝ

ΠΟΛΥΜΕΤΑΒΛΗΤΕΣ ΚΑΤΑΝΟΜΕΣ

Σε πολλά προβλήματα, ενδιαφερόμαστε για περισσότερα από ένα χαρακτηριστικά ενός πληθυσμού. Τα χαρακτηριστικά αυτά είναι πιθανό να αλληλοεξαρτώνται (για παράδειγμα εισόδημα με κατανάλωση, τιμή, ποιότητα και αντοχή ενός υλικού κ.λ.π.). Προβλήματα της μορφής αυτής οδηγούν στην μελέτη της σχέσης ανάμεσα σε διαφορετικές τυχαίες μεταβλητές, και τον ορισμό της από κοινού συνάρτησης κατανομής. (Για ευκολία θα δίνουμε τους ορισμούς και τις έννοιες στις δύο διαστάσεις και θα εξηγούμε πως οι έννοιες αυτές γενικεύονται στις n διαστάσεις).

Ορισμός: Έστω X και Y δύο τυχαίες μεταβλητές. Ως *από κοινού συνάρτηση πιθανότητας (joint probability distribution)* (ή αντίστοιχα *από κοινού συνάρτηση πυκνότητας πιθανότητας) (joint probability density function)* των X και Y ορίζουμε την συνάρτηση $P:RxR \to R$ έτσι ώστε

$$p_{xy} \equiv P(x,y) \equiv P(X=x, Y=y) \text{ αν X,Y είναι διακριτές}$$

ή αντίστοιχα τη συνάρτηση $f:RxR \to R$ έτσι ώστε

$f(x,y)dxdy = P(x<X<x+dx,\ y<Y<y+dy)$ αν X,Y είναι συνεχείς.

Οι παραπάνω συναρτήσεις ικανοποιούν τις σχέσεις:

1) $P_{xy} \geq 0$, $f(x,y) \geq 0$, τελικά για όλα τα (x,y).

2) $\sum_x \sum_y P(x,y) = 1$ αν X,Y διακριτές

$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x,y)dxdy = 1$ αν X,Y είναι συνεχείς.

Ορισμός: Ως *από κοινού συνάρτηση κατανομής (joint probability function)* των τυχαίων μεταβλητών X και Y ορίζουμε την συνάρτηση

$$F : RxR \rightarrow R : F(x,y) = P(X \leq x, Y \leq y), -\infty < x,y < \infty$$

Παρατήρηση: Οι προηγούμενοι ορισμοί επεκτείνονται και στην περίπτωση n μεταβλητών $X_1, X_2, ..., X_n$ δηλαδή στην περίπτωση του n-διάστατου τυχαίου διανύσματος $\underset{\sim}{X} \equiv (X_1, X_2, \ldots, X_n)$.

Η κατανομή του X προκύπτει από την από κοινού κατανομή των X και Y ως εξής:

$$F_X(a) = P(X \leq a) = P(X \leq a, Y < \infty) = P(\lim_{b \to \infty}(\{X \leq a, Y \leq b\})$$
$$= \lim_{b \to \infty} P(X \leq a, Y \leq b)$$
$$= \lim_{b \to \infty} F(a,b) = F(a,\infty)$$

Ομοίως

$$F_Y(b) = P(Y \leq b) = \lim_{a \to \infty} F(a,b) = F(\infty,b)$$

Ορισμός: Οι συναρτήσεις κατανομής $F_X(x)$, $F_Y(y)$ λέγονται *περιθώριες συναρτήσεις κατανομής (marginal distribution functions)* των X και Y, αντίστοιχα.

Αν X και Y είναι διακριτές τυχαίες μεταβλητές *οι περιθώριες κατανομές πιθανότητας* των X και Y υπολογίζονται ως εξής:

$$P_X(x) = P(X=x) = \sum_y P(x,y)$$

$$P_Y(y) = P(Y=y) = \sum_x P(x,y)$$

Αν X και Y είναι συνεχείς οι *περιθώριες συναρτήσεις πυκνότητας πιθανότητας* δίνονται από τους εξής τύπους:

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)dy$$

και

$$f_Y(y) = \int_{-\infty}^{+\infty} f(x,y)dx$$

Παράδειγμα: Στρίβουμε ένα αμερόληπτο νόμισμα τρεις φορές. Ορίζουμε τις τυχαίες μεταβλητές X και Y ως εξής:

$$X = \begin{cases} 0 & \text{αν έχουμε Κ στην πρώτη δοκιμή} \\ 1 & \text{αν έχουμε Γ στην πρώτη δοκιμή} \end{cases}$$

Y = ο αριθμός των Κ στις τρεις δοκιμές

x=0,1 , y=0,1,2,3.

Για τις κατανομές πιθανότητας των X και Y έχουμε αντίστοιχα:

x	0	1
$P_X(x)$	1/2	1/2

y	0	1	2	3
$P_Y(y)$	1/8	3/8	3/8	1/8

Επίσης,

{X=0} = {ΚΚΚ, ΚΚΓ, ΚΓΚ, ΚΓΓ}
{X=1} = {ΓΚΚ, ΓΚΓ, ΓΓΚ, ΓΚΚ}

Η από κοινού συνάρτηση πιθανότητας P(x,y) μπορεί να δοθεί με την μορφή του παρακάτω πίνακα.

Y / x	0	1	2	3	$P_X(x)$
0	0	1/8	2/8	1/8	1/2
1	1/8	2/8	1/8	0	1/2
$P_Y(y)$	1/8	3/8	3/8	1/8	1

ΑΝΕΞΑΡΤΗΣΙΑ ΤΥΧΑΙΩΝ ΜΕΤΑΒΛΗΤΩΝ

Ορισμός: Δύο τυχαίες μεταβλητές X και Y θα λέγονται *ανεξάρτητες (independent)* αν για κάθε ζεύγος συνόλων A και B πραγματικών αριθμών (ακριβέστερα συνόλων Borel A και B της πραγματικής ευθείας)

$$P\{X \in A, Y \in B\} = P\{X \in A\}\, P(Y \in B\}$$

Με άλλα λόγια, οι X και Y είναι ανεξάρτητες αν για όλα τα A και B τα ενδεχόμενα $E_A = \{X \in A\}$ και $E_B = \{Y \in B\}$ είναι ανεξάρτητα. (Βλέπε και ορισμό της ανεξαρτησίας ενδεχομένων στο κεφάλαιο 1).

Είναι εύκολο να αποδειχθεί, χρησιμοποιώντας τα τρία αξιώματα των πιθανοτήτων, ότι ο παραπάνω ορισμός είναι ισοδύναμος με τη σχέση

$$F(x,y) = F_X(x)F_Y(y) \qquad \text{για κάθε } (x,y)$$

Στην περίπτωση της συνεχούς διμεταβλητής κατανομής η σχέση ανεξαρτησίας είναι ισοδύναμη με τη σχέση:

$$f(x,y) = f_X(x)f_Y(y) \qquad \text{για κάθε } (x,y)$$

Αν οι X και Y δεν είναι ανεξάρτητες, λέγονται *εξαρτημένες (dependent).*

Παράδειγμα: Στο παράδειγμα του στριψίματος αμερόληπτου νομίσματος τρεις φορές, οι τυχαίες μεταβλητές X και Y είναι εξαρτημένες. Αυτό γιατί

$$P(1,1) = \frac{2}{8} \neq P_X(1)P_Y(1) = \frac{1}{2}\frac{3}{8} = \frac{3}{16}$$

Παράδειγμα: Έστω ότι

$$P(x,y) = \frac{xy^2}{30}, \quad x=1, 2, 3 \quad y=1,2$$

Είναι

$$P_X(x) = \frac{x}{30}\sum_{y=1}^{2} y^2 = \frac{x}{y}, \quad x=1, 2, 3$$

$$P_Y(y) = \frac{y^2}{30}\sum_{x=1}^{3} = \frac{y^2}{5}, \quad y=1, 2$$

Επομένως,

$$P(x, y)=P_X(x)P_Y(y) \quad \text{για κάθε } (x, y)$$

δηλαδή οι X, Y είναι ανεξάρτητες

Παράδειγμα: Έστω ότι

$$f(x,y)=\begin{cases}2(x+4y)/5 & \text{αν } 0\le x,y\le 1\\ 0 & \text{διαφορετικά}\end{cases}$$

Έχουμε

$$f_X(x)=\frac{2}{5}\int_0^1 (x+4y)dy=\frac{2(x+2)}{5},\quad 0\le x\le 1$$

$$f_Y(y)=(1+8y)/5,\qquad 0\le y\le 1$$

Επομένως,

υπάρχει $(x,y)\in[0,1]x[0,1]$ τέτοιο ώστε:

$$f(x,y)\neq f_X(x)f_Y(y)$$

και συνεπώς οι X και Y εξαρτημένες.

ΔΕΣΜΕΥΜΕΝΗ ΚΑΤΑΝΟΜΗ

Στο κεφάλαιο 3 ορίσαμε σαν δεσμευμένη πιθανότητα δύο ενδεχομένων A και B με P(B)>0 την πιθανότητα

$$P(A\mid B)=\frac{P(A\cap B)}{P(B)}$$

Γενικεύοντας την έννοια αυτή σε τυχαίες μεταβλητές μπορούμε να ορίσουμε την δεσμευμένη κατανομή πιθανότητας και την δεσμευμένη πυκνότητα πιθανότητα ως εξής:

Ορισμός: Αν X και Y είναι διακριτές τυχαίες μεταβλητές ορίζουμε ως *δεσμευμένη (conditional) συνάρτηση πιθανότητας* του X δοθέντος ότι Y=y τη συνάρτηση

$$P_{X|Y}(x\mid y)=P(X=x\mid Y=y)=\frac{P(X=x,Y=y)}{P(Y=y)}=\frac{P(x,y)}{P_Y(y)}$$

για κάθε y: P (y)>0.

Ομοίως η δεσμευμένη συνάρτηση κατανομής του X δοθέντος ότι Y=y ορίζεται για κάθε y:$P_Y(y)>0$ ως εξής:

$$F_{X|Y}(x|y)=P(X\le x|Y=y)=\sum_{a\le x}PX|\,Y(x|\,y)$$

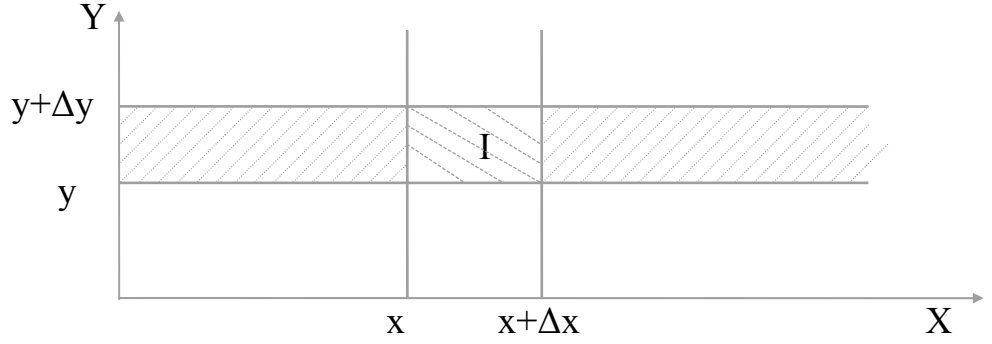
Παράδειγμα: Στο παράδειγμα της σελίδας 88

$$P(Y=2\,|\,X=1)=\frac{P(Y=2,X=1)}{P(X=1)}=\frac{1/8}{1/2}=\frac{1}{4}$$

Ορισμός: Αν οι τυχαίες μεταβλητές X και Y έχουν την από κοινού συνάρτηση πυκνότητας πιθανότητας f(x,y), τότε η *δεσμευμένη πυκνότητα πιθανότητας* του X δοθέντος ότι Y=y, ορισμένη για όλα τα y για τα οποία $f_Y(y)>0$, δίνεται από τη σχέση

$$f_{X|Y}(x\,|\,y)=\frac{f(x,y)}{f_Y(y)}=\lim_{\substack{\Delta x\to 0\\ \Delta y\to 0}}\frac{P\{x\le X\le x+\Delta x,\ y\le Y\le y+\Delta y\}}{\Delta x P\{y\le Y\le y+\Delta y\}}$$

Παρατήρηση: Η $f_{X|Y}(x|y)$ όπως ορίσθηκε παραπάνω, είναι η πιθανότητα να βρισκόμαστε στο ορθογώνιο I δοθέντος ότι είμαστε στην λωρίδα από y έως y + Δy.



Η $f_{X|Y}(x|y)$ είναι πράγματι μία συνάρτηση πυκνότητας πιθανότητας διότι

$$\int_{-\infty}^{+\infty}f_{X|Y}(x\,|\,y)dx=\int_{-\infty}^{+\infty}\frac{f(x,y)dx}{f_Y(y)}=\frac{f_Y(y)}{f_Y(y)}=1$$

Σημείωση: Η χρήση της δεσμευμένης πιθανότητας μας επιτρέπει να ορίσουμε πιθανότητες ενδεχομένων που αναφέρονται σε μία τυχαία μεταβλητή όταν ξέρουμε την τιμή μίας άλλης μεταβλητής. Δηλαδή αν (X,Y) είναι μια συνεχής τυχαία μεταβλητή, τότε για κάθε σύνολο A

$$P(X \in A \mid Y = y) = \int_A f_{X|Y}(x \mid y)dx$$

Ειδικότερα, αν $A = (-\infty, \alpha]$ μπορούμε να ορίσουμε την δεσμευμένη συνάρτηση κατανομής του X δοθέντος ότι $Y = y$ ως

$$F_{X|Y}(a \mid y) = P(X \le a \mid Y = y) = \int_{-\infty}^{a} f_{X|Y}(x \mid y)dx$$

Σημείωση: Είναι ενδιαφέρον να παρατηρηθεί ότι, με τις έννοιες που αναπτύχθηκαν, κατορθώσαμε να καταλήξουμε σε εκφράσεις για δεσμευμένες πιθανότητες της μορφής $P(X \in A \mid Y = y)$, παρά το γεγονός ότι το ενδεχόμενο πάνω στο οποίο γίνεται η δέσμευση (δηλαδή το ενδεχόμενο $\{Y = y\}$), έχει πιθανότητα 0.

Παράδειγμα: Έστω ότι η από κοινού συνάρτηση πυκνότητας πιθανότητας των X και Y δίνεται από τον τύπο:

$$f(x,y) = \begin{cases} \dfrac{e^{-x/y}e^{-y}}{y} & 0 < x < \infty, \quad 0 < y < \infty \\ \\ 0 & \text{διαφορετικά} \end{cases}$$

Να βρεθεί η $P(X > 1 \mid Y=y)$.

Λύση: Βρίσκουμε πρώτα την δεσμευμένη πυκνότητα

$$f_{X|Y}(x \mid y) = \frac{f(x,y)}{f_y(y)} = \frac{e^{-x/y}e^{-y}/y}{e^{-y}\int_0^{\infty}(1/y)\,e^{-x/y}dx} = \frac{1}{y}e^{-x/y}$$

Επομένως,

$$P(X > 1/Y = y) = \int_1^{\infty}\frac{1}{y}e^{-x/y}dx = e^{-x/y}\Big|_1^{\infty} = e^{-1/y}$$

Παρατήρηση: Από τα προηγούμενα γίνεται φανερό ότι γενικά για δύο τυχαίες μεταβλητές X και Y η δεσμευμένη συνάρτηση κατανομής της X δοθέντος ότι $Y \in B$, όπου $P(Y \in B) > 0$, ορίζεται ως

$$F_{X|Y}(x \mid Y \in B) = \frac{P(X \le x, Y \in B)}{P(Y \in B)}$$

Μέση Τιμή και Διασπορά Αθροισμάτων

Τυχαίων Μεταβλητών

Πρόταση: Έστω X_1, X_2, ..., X_n τυχαίες μεταβλητές με $E(X_i)<\infty$, i=1,2,...,n.
Τότε

$$E(X_1 + X_2 + ... + X_n) = E(X_1)+E(X_2)+...+E(X_n)$$

Πρόταση: Έστω X_1, X_2, ..., X_n ανεξάρτητες τυχαίες μεταβλητές με $E(X_i)<\infty$, i=1,2,...,n. Τότε

i) $E(X_1 X_2 ... X_n) = E(X_1)E(X_2)...E(X_n)$
ii) $\Delta(X_1 + X_2 + ...+X_n) = \Delta(X_1)+\Delta(X_2)+...+\Delta(X_n)$

Σημείωση: Για την δεύτερη πρόταση απαιτείται ανεξαρτησία των τυχαίων μεταβλητών, κάτι που δεν είναι απαραίτητο στην πρώτη.

Παράδειγμα: Σε μια δεξίωση έχουν συγκεντρωθεί N ζευγάρια. Οι γυναίκες χωρίζονται από τους άνδρες και κάθε άνδρας διαλέγει στην τύχη μια γυναίκα για να χορέψει. Να βρεθεί ο αναμενόμενος αριθμός των ανδρών που χορεύουν με τις γυναίκες τους.

Λύση: Έστω X ο αριθμός των ταιριασμάτων. Ορίζουμε την τυχαία μεταβλητή X_i , i=1,2,...,N ως εξής:

$$X_i = \begin{cases} 1 & \text{αν ο i άνδρας χορεύει με την γυναίκα του} \\ 0 & \text{διαφορετικά} \end{cases}$$

Τότε $X=X_1 +X_2 +...+ X_n$
Αλλά, για κάθε i

$$E(X_i) = P(X_i=0) \times 0 + P(X_i =1) \times 1 = 1/N$$

και επομένως

$$E(X) = E\left(\sum_{i=1}^{N} X_i\right) = \sum_{i=1}^{N} E(X_i) =(1/N)N = 1$$

Δηλαδή, κατά μέσο όρο, μόνο ένας από τους άνδρες θα χορεύει με την γυναίκα του.

Σημείωση: Οι τυχαίες μεταβλητές $X_1, X_2, ..., X_n$ είναι εξαρτημένες και επομένως η $\Delta(X)$ δεν μπορεί να υπολογισθεί με την μέθοδο της προηγούμενης πρότασης.

Παράδειγμα: Έστω ότι σε ένα κουτί έχουμε 2N κάρτες από τις οποίες 2 είναι σημειωμένες με τον αριθμό 1, δύο με τον αριθμό 2, δύο με τον αριθμό 3 κ.ο.κ. και δύο με τον αριθμό N. Παίρνουμε στην τύχη κ κάρτες από το κουτί. Να βρεθεί ο αναμενόμενος αριθμός ζευγαριών που έχουν μείνει στο κουτί.
(Το πρόβλημα αυτό τέθηκε και λύθηκε τον 18ο αιώνα από τον Daniel Bernoulli ως ένα πιθανό στοχαστικό μοντέλο για τον καθορισμό του αριθμού των γάμων όπου ο άνδρας και η γυναίκα βρίσκονται στη ζωή, όταν έχουν σημειωθεί κ θάνατοι ανάμεσα σε N παντρεμένα ζευγάρια).

Λύση: Ορίζουμε τις τυχαίες μεταβλητές

$$X_i = \begin{cases} 1 & \text{αν το i ζευγάρι παραμένει άθικτο, } i = 1,2,....N \\ 0 & \text{διαφορετικά} \end{cases}$$

Είναι

$$E(X_i) = P(X_i=1) = \frac{\binom{2N-2}{\kappa}}{\binom{2N}{\kappa}}$$

$$= \frac{\frac{(2N-2)!}{\kappa!\,(2N-2-\kappa)!}}{\frac{(2N)!}{\kappa!\,(2N-\kappa)!}} = \frac{(2N-\kappa)(2N-\kappa-1)}{(2N)(2N-1)}$$

Επομένως ο αναμενόμενος αριθμός ζευγαριών στις κάρτες που έχουν μείνει στο κουτί είναι:

$$E(X_1 + X_2 + ... + X_n) = E(X_1) + E(X_2) + ... + E(X_n)$$

$$= \frac{(2N - \kappa)(2N - \kappa - 1)}{2(2N - 1)}$$

Παράδειγμα: Ρίχνουμε ένα αμερόληπτο ζάρι 300 φορές. Να εκτιμηθεί η πιθανότητα του ενδεχομένου να εμφανισθεί 1 ή 2 λιγότερες από 70 φορές.

Λύση: Έστω X ο αριθμός των 1 και 2 στις 300 δοκιμές. Θέλουμε μια εκτίμηση για την πιθανότητα του ενδεχομένου {X<70}. Ορίζουμε τις τυχαίες μεταβλητές

$$X_i = \begin{cases} 1 & \text{αν στην i δοκιμή το αποτέλεσμα είναι 1 ή 2, } i = 1,2,....n \\ 0 & \text{διαφορετικά} \end{cases}$$

Για την κατανομή πιθανότητας του X_i έχουμε

x_i	0	1
$P(x_i)$	2/3	1/3

Επομένως

$$E(X_i) = 1/3 \quad \text{και} \quad \Delta(X_i) = 2/9$$

και

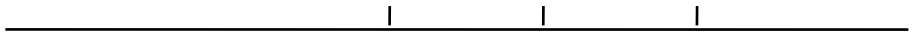
$$E(X) = E\left(\sum_{i=1}^{300} X_i\right) = \sum_{i=1}^{300} E(X_i) = 100$$

Επίσης, επειδή οι τυχαίες μεταβλητές X_i είναι ανεξάρτητες

$$\Delta(X) = \Delta\left(\sum_{i=1}^{300} (X_i)\right) = \sum_{i=1}^{300} \Delta(X_i) = 600/9$$

Προφανώς

$$\{X<70\} \cup \{X>130\} = \{|X-100| > 30\}$$



Συνεπώς,

$$P\{X<70\} \le P(|X-100| > 30)$$

Από το θεώρημα του Chebyshev όμως

$$P(|X-100| > 30) \leq \frac{\Delta(X)}{30^2}$$

Συνεπώς,

$$P(X<70) \leq \frac{\Delta(X)}{30^2}$$

Δηλαδή, τελικά

$$P(X<70) \leq \frac{600/9}{30^2} = \frac{2}{27}$$

Επομένως, σύμφωνα με το θεώρημα του Chebyshev, η πιθανότητα να έχουμε λιγότερα από 70 1 ή 2 σε 300 ριξίματα ενός αμερόληπτου ζαριού είναι το πολύ 2/27. Θα δούμε αργότερα πως η πιθανότητα αυτή μπορεί να καθοριστεί ακριβώς ή με προσέγγιση.

ΣΥΝΔΙΑΚΥΜΑΝΣΗ ΚΑΙ ΣΥΣΧΕΤΙΣΗ

Από την μέχρι τώρα μελέτη των πολυμεταβλητών κατανομών, γίνεται φανερό ότι οι τυχαίες μεταβλητές που υπεισέρχονται σ' αυτές είναι ή εξαρτημένες ή ανεξάρτητες. Στην ενότητα αυτή, σκοπός μας είναι να ορίσουμε και να μελετήσουμε κάποιο μέτρο που να εκφράζει το βαθμό κάποιας μορφής εξάρτησης δύο τυχαίων μεταβλητών. Θα ασχοληθούμε με ένα ειδικό τρόπο εξάρτησης, την γραμμική εξάρτηση.

Ορισμός: Έστω X και Y δύο τυχαίες μεταβλητές με μέσες τιμές μ_X και μ_Y αντίστοιχα. Ορίζουμε ως *συνδιακύμανση* ή *συνδιασπορά* (*covariance*) των X και Y, και συμβολίζουμε με Cov(X,Y) την ποσότητα

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

Όπως φαίνεται και από την ονομασία της, η συνδιακύμανση είναι ένα μέτρο του τρόπου με τον οποίο οι δύο τυχαίες μεταβλητές μεταβάλλονται από κοινού.

Παράδειγμα: (Συνέχεια του παραδείγματος του στριψίματος δύο νομισμάτων). Για το παράδειγμα αυτό βρίσκουμε εύκολα ότι

$\mu_X = E(X) = \Sigma x P_X(x) = 1/2$
$\mu_Y = E(Y) = \Sigma y P_Y(y) = 3/2$

Επομένως, η συνδιακύμανση των X και Y είναι

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = \sum_x \sum_y P(x,y)(x-1/2)(y-3/2) = -1/4$$

Είναι ενδιαφέρον να παρατηρήσει κανείς ότι το αποτέλεσμα αυτό αναμενόταν μια και, από τον ορισμό των X και Y μια "μεγάλη" τιμή της X συνεπάγεται μια "μικρή" τιμή για το Y και επομένως υπάρχει αρνητική εξάρτηση.

Πρόταση: $Cov(X,Y) = E(XY)-E(X)E(Y)$.

Σημείωση: Η παραπάνω ιδιότητα χρησιμοποιείται τις περισσότερες φορές αντί του ορισμού για τον υπολογισμό της συνδιακύμανσης.

Πόρισμα: Αν X,Y είναι ανεξάρτητες, τότε $Cov(X,Y)=0$.

Πρόταση: $Cov(X+Y,Z) = Cov(X,Z)+Cov(Y,Z)$.

Πρόταση: Για τις τυχαίες μεταβλητές X και Y, ισχύει ότι

$$\Delta(X \pm Y) = \Delta(X) + \Delta(Y) \pm 2Cov(X,Y)$$

Η προηγούμενη προτάση επεκτείνεται στην περίπτωση n τυχαίων μεταβλητών, ως εξής:

Πρόταση: Για τις τυχαίες μεταβλητές $X_1, X_2, \ldots, X_n$, ισχύει ότι

$$\Delta\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \Delta(X_i) + 2\sum_{i<j} Cov(X_i, X_j)$$

Ως άμεση συνέπεια των δύο προηγούμενων προτάσεων, προκύπτει το εξής πόρισμα:

Πόρισμα: Αν οι τυχαίες μεταβλητές X και Y είναι ανεξάρτητες τότε

$$\Delta(X+Y) = \Delta(X) + \Delta(Y)$$

Γενικότερα, αν οι τυχαίες μεταβλητές X_1, X_2, …, X_n είναι ανεξάρτητες κατά ζεύγη, τότε

$$\Delta\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \Delta(X_i)$$

Παρατήρηση: Οι δύο προηγούμενες προτάσεις χρησιμεύουν για τον υπολογισμό της διασποράς του αθροίσματος δύο ή περισσοτέρων τυχαίων μεταβλητών, όταν αυτές είναι εξαρτημένες.

Παράδειγμα: Να υπολογισθεί η διασπορά του αριθμού των ταιριασμάτων στο σχετικό παράδειγμα.

Λύση: Είχαμε δει ότι

$$X = X_1 + X_2 + \ldots X_n$$

όπου $X = \begin{cases} 1 & \text{αν ο i άνδρας χορεύει με την γυναίκα του, } i = 1,2,,....N \\ 0 & \text{διαφορετικά} \end{cases}$

Επομένως,

$$\Delta(X) = \sum_{i=1}^{N} \Delta(X_i) + 2\sum_{i<j} \mathrm{Cov}(X_i X_j)$$

Επειδή $P(X_i = 1) = 1/N.$

Έχουμε ότι

$$E(X_i) = 1/N \quad \text{και} \quad E(X_i^2) = 1/N^2$$

και

$$\Delta(X_i) = E(X_i^2) - \{E(X_i)\}^2 = \frac{1}{N^2} - \frac{1}{N} = \frac{N-1}{N^2}$$

Επίσης,

$$\mathrm{Cov}(X_i X_j) = E(X_i X_j) - E(X_i)E(X_j)$$

Αλλά

$$X_i X_j = \begin{cases} 1 & \text{αν και ο i και ο j άνδρας χορεύουν με τις γυναίκες τους} \\ 0 & \text{διαφορετικά} \end{cases}$$

Και έτσι

$$E(X_i X_j) = \sum_{i=0}^{1}\sum_{j=0}^{1} ijP(i,j) = P(X_i = 1, X_j = 1)$$

$$= P(X_i = 1 \mid X_j = 1)P(X_j = 1) = \frac{1}{N-1}\frac{1}{N}$$

Επομένως

$$Cov(X_i, X_j) = \frac{1}{N(N-1)} - \frac{1}{N^2} = \frac{1}{N^2(N-1)}$$

και τελικά

$$\Delta(X) = \frac{N-1}{N} + 2\binom{N}{2}\frac{1}{N^2(N-1)}$$

$$= \frac{N-1}{N} + \frac{1}{N} = 1$$

Παρατηρούμε, δηλαδή, ότι ο μέσος και η διασπορά της κατανομής του X συμπίπτουν. Θα έχουμε την ευκαιρία να δώσουμε μία θεωρητική εξήγηση του αποτελέσματος αυτού σε ένα από τα επόμενα κεφάλαια.

Η συνδιακύμανση δύο τυχαίων μεταβλητών, ως μέτρο εξάρτησης, έχει ένα μεγάλο μειονέκτημα. Μεταβάλλεται αν μεταβληθούν οι μονάδες μέτρησης των τυχαίων μεταβλητών. Έστω για παράδειγμα, X_1 και Y_1 δύο τυχαίες μεταβλητές. Θεωρούμε τις τυχαίες μεταβλητές $X_2=2X_1$ και $Y_2=2Y_1$. Είναι

$$E(X_2)=2E(X_1),\ E(Y_2)=2E(Y_1) \text{ και } E(X_2Y_2)=4E(X_1Y_1)$$

Επομένως

$$Cov(X_2, Y_2) = E(X_2, Y_2) - E(X_2)E(Y_2)$$

$$= 4E(X_1Y_1) - 4E(X_1)E(Y_1)$$

$$= 4Cov(X_1, Y_1)$$

Ενώ, δηλαδή τα ζευγάρια X_1, Y_1 και X_2, Y_2 συμπεριφέρονται με τον ίδιο ακριβώς τρόπο, το δεύτερο ζευγάρι έχει συνδιακύμανση τέσσερις φορές μεγαλύτερη από το πρώτο.

Εκτός από το μειονέκτημα αυτό, η συνδιακύμανση είναι μια μη φραγμένη συνάρτηση. Για τους λόγους αυτούς, είναι επιθυμητό να ορισθεί ένα άλλο μέτρο εξάρτησης δύο τυχαίων μεταβλητών που σαν συνάρτηση των τυχαίων μεταβλητών να είναι φραγμένη και να μην επηρεάζεται από πιθανές μεταβολές των μονάδων μέτρησης των μεταβλητών. Οι σκέψεις αυτές οδηγούν στον παρακάτω ορισμό:

Ορισμός: Έστω X,Y δύο τυχαίες μεταβλητές με μέσες τιμές μ_X, μ_Y και τυπικές αποκλίσεις σ_X και σ_Y αντίστοιχα. Έστω X΄, Y΄ οι αντίστοιχες τυποποιημένες τυχαίες μεταβλητές. Αν σ_X, $\sigma_Y \neq 0$ ορίζουμε ως *συντελεστή συσχέτισης (correlation coefficient)* των X, Y και συμβολίζουμε με $\rho(X,Y)$ ή με $\rho_{X,Y}$ τη συνδιακύμανση των X΄ και Y΄. Δηλαδή

$$\rho_{X,Y} \equiv \rho(X,Y) = \mathrm{Cov}(X',Y')$$

Αν είτε $\sigma_X = 0$ είτε $\sigma_Y = 0$ ορίζουμε $\rho(X,Y)=0$.
Αν $\rho(X,Y)=0$ οι τυχαίες μεταβλητές λέγονται *ασυσχέτιστες (uncorrelated)*.

Πρόταση: $$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

Ο παραπάνω τύπος χρησιμοποιείται συνήθως, αντί του ορισμού, για τον υπολογισμό του συντελεστή συσχέτισης δύο τυχαίων μεταβλητών.

Παράδειγμα: (συνέχεια του παραδείγματος του στριψίματος δύο νομισμάτων). Για το παράδειγμα αυτό έχουμε ότι

$$E(X^2) = 1/2 \quad \text{και} \quad E(Y^2) = 3$$

και

$$\Delta(X) = 1/4\ ,\ \Delta(Y) = 3/4$$

Έχουμε ήδη υπολογίσει ότι

$$\mathrm{Cov}(X,Y) = -(1/4)$$

και επομένως, χρησιμοποιώντας την προηγούμενη πρόταση

$$\rho(X,Y) = -0.58$$

Οι επόμενες δύο προτάσεις επιβεβαιώνουν ότι τα μειονεκτήματα της συνδιακύμανσης παύουν να υφίστανται με την χρησιμοποίηση του συντελεστή συσχέτισης.

Πρόταση: Αν X_1 και X_2 είναι τυχαίες μεταβλητές και
$Y_1 = \alpha_1 X_1 + \beta_1$ και $Y_2 = \alpha_2 X_2 + \beta_2$ όπου $\alpha_1, \alpha_2, \beta_1, \beta_2$ σταθερές, τότε

$$\rho(Y_1, Y_2) = \rho(X_1, X_2)$$

(δηλαδή η αλλαγή συντεταγμένων δεν επηρεάζει τον συντελεστή συσχέτισης).

Πρόταση: $|\rho(X,Y)| \leq 1$.
(Δηλαδή το $\rho(X,Y)$ είναι μια φραγμένη συνάρτηση).

Τα αποτελέσματα που ακολουθούν σχετίζονται με το γεγονός ότι ο συντελεστής συσχέτισης μπορεί να χρησιμοποιηθεί σαν μέτρο του βαθμού της γραμμικής εξάρτησης δύο τυχαίων μεταβλητών.

Πρόταση: Έστω X και Y δύο τυχαίες μεταβλητές γραμμικά εξαρτημένες, δηλαδή $Y=\alpha X+\beta$ με α και β σταθερές και $\alpha \neq 0$. Τότε

$$\rho(X,Y) = \begin{cases} 1 & \text{αν } \alpha > 0 \\ -1 & \text{αν } \alpha < 0. \end{cases}$$

Η πρόταση που ακολουθεί είναι το αντίστροφο της προηγούμενης.

Πρόταση: Αν $|\rho(X, Y)| = 1$ υπάρχουν σταθερές $\alpha \neq 0$ και β τέτοιες ώστε να $Y = \alpha X+\beta$ εκτός, πιθανόν, για κάποιο σύνολο τιμών με πιθανότητα μηδέν.

Οι προηγούμενες προτάσεις οδηγούν στο συμπέρασμα ότι οι τιμές του $\rho(X,Y)$ που πλησιάζουν το 1 δίνουν μια ένδειξη για θετική γραμμική εξάρτηση ($\alpha>0$) των X και Y ενώ τιμές του $\rho(X,Y)$ που πλησιάζουν το -1 αποτελούν ένδειξη για αρνητική γραμμική εξάρτηση μεταξύ των X και Y ($\alpha<0$). Αντίθετα, τιμές του $\rho(X,Y)$ που διαφέρουν πολύ από το μηδέν οδηγούν στο

συμπέρασμα ότι δεν υπάρχει γραμμική σχέση των X και Y (χωρίς αυτό να αποκλείει εξάρτηση κάποιας άλλης μορφής).

Παρατήρηση: Στο σημείο αυτό, θα πρέπει να τονισθεί ότι ο συντελεστής συσχέτισης θα πρέπει να χρησιμοποιείται ως ένδειξη γραμμικής εξάρτησης μόνο και όχι ως ένδειξη οποιασδήποτε μορφής εξάρτησης. Αυτό, γιατί από την προηγούμενη συζήτηση προκύπτει ότι η έννοια της ανεξαρτησίας δύο τυχαίων μεταβλητών, συνεπάγεται και την έλλειψη συσχέτισης των μεταβλητών αυτών. Το αντίστροφο όμως δεν ισχύει πάντα. Αν δηλαδή, δύο τυχαίες μεταβλητές είναι ασυσχέτιστες, αυτό δεν συνεπάγεται ότι είναι ανεξάρτητες. Το παράδειγμα που ακολουθεί επιβεβαιώνει το γεγονός αυτό.

Παράδειγμα: Θεωρούμε τις τυχαίες μεταβλητές X και Y που ορίζονται ως εξής:

$$P(X = 0) = P(X = 1) = P(X = -1) = 1/3$$

$$Y = \begin{cases} 0 & \text{αν } X \neq 0 \\ 1 & \text{αν } X = 0 \end{cases}$$

Προφανώς οι X και Y είναι εξαρτημένες.
Από τον ορισμό των X και Y έχουμε ότι

$$XY=0 \text{ και επομένως } E(XY)=0$$

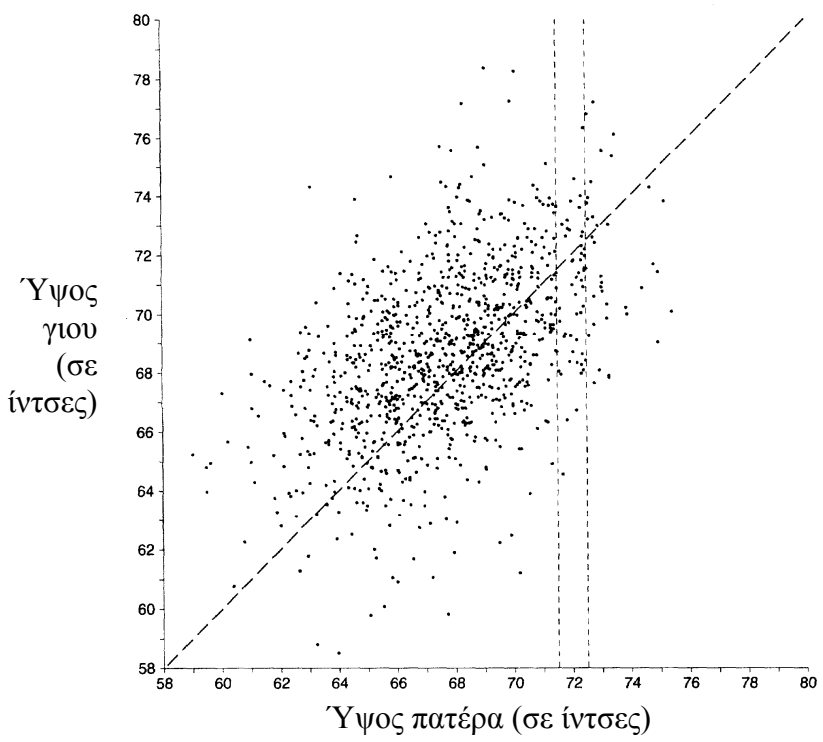
Ομοίως $E(X)=0$.
Συνεπώς $Cov(X,Y)=0$. Άρα οι X και Y είναι ασυσχέτιστες.

Δειγματική Συνδιακύμανση και Συσχέτιση

Η πρώτη προσπάθεια για την μελέτη της σχέσης μεταξύ δύο μεταβλητών έγινε από τον Sir Francies Galton (Αγγλία, 1822-1911). Τον Galton απασχόλησε το πρόβλημα του βαθμού στον οποίο τα παιδιά μοιάζουν στους γονείς τους. Οι Στατιστικοί της Βικτωριανής Αγγλίας είχαν γοητευθεί από την ιδέα να ποσοτικοποιήσουν την επίδραση της κληρονομικότητας και συνέλεξαν μεγάλους όγκους δεδομένων για την διερεύνηση αυτή. Ο μαθητής του Galton, Karl Pearson (1857-1936) στην προσπάθεια αυτή μέτρησε το ύψος 1078 πατέρων και το ύψος των γιών τους μετά την ενηλικίωση.

Η σχέση των δύο μεταβλητών, ύψος πατέρα και ύψος γιού, απεικονίζεται στο διάγραμμα σημείων που ακολουθεί.

Διάγραμμα σημείων του ύψους 1078 πατέρων και υιών



Η διχοτόμος της γωνίας (η ευθεία y = x) δείχνει οικογένειες, στις οποίες το ύψος πατέρα και γιού είναι το ίδιο. Οικογένειες, στις οποίες το ύψος του πατέρα είναι 72 ίντσες, εμφανίζονται στην κατακόρυφη λωρίδα.

Ο Δειγματικός Συντελεστής Συσχέτισης

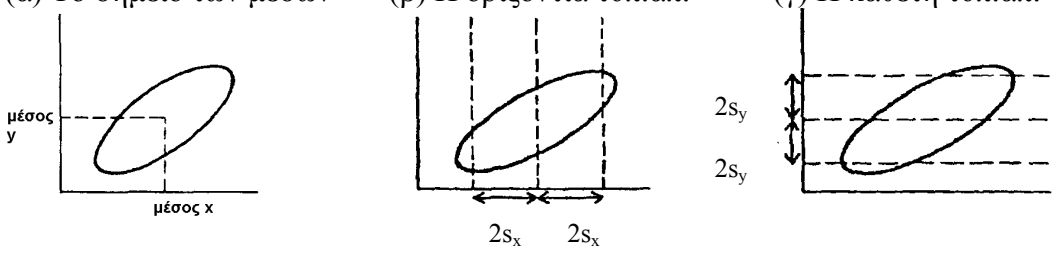
Ας υποθέσουμε ότι ενδιαφερόμαστε να εξετάσουμε την συσχέτιση μεταξύ δύο τυχαίων μεταβλητών. Ο απλούστερος τρόπος είναι να σχεδιάσουμε ένα διάγραμμα σημείων για ένα τυχαίο δείγμα τιμών των δύο μεταβλητών. Το γράφημα είναι ένα διάγραμμα (νέφος) σημείων που έχει την μορφή αμερικάνικης ποδοσφαιρικής μπάλας. Πώς θα μπορούσε να συνοψισθεί αριθμητικά το γράφημα αυτό; (Είναι

προφανές ότι ένα τέτοιο αριθμητικό μέτρο θα αποτελούσε και εκτίμηση του αντίστοιχου μέτρου του πληθυσμού).

Το πρώτο βήμα θα ήταν να σημειώσουμε ένα σημείο που δείχνει τον μέσο των x-τιμών και τον μέσο των y-τιμών (βλέπε σχήμα α). Αυτό είναι *το σημείο των μέσων*, το οποίο προσδιορίζει το κέντρο του διαγράμματος σημείων. Το επόμενο βήμα θα ήταν να μετρήσουμε το "άπλωμα" του διαγράμματος από την μία πλευρά ως την άλλη. Αυτό μπορεί να γίνει χρησιμοποιώντας την τυπική απόκλιση των x-τιμών - την οριζόντια τυπική απόκλιση. Τα περισσότερα σημεία θα βρίσκονται μεταξύ 2 οριζόντιων τυπικών αποκλίσεων προς κάθε μία πλευρά του σημείου που αντιστοιχεί στον μέσο (σχήμα β). Με τον ίδιο τρόπο, η τυπική απόκλιση των y-τιμών - η κάθετη τυπική απόκλιση - θα μπορούσε να χρησιμοποιηθεί για να μετρήσουμε το "άπλωμα" του διαγράμματος (νέφους) από πάνω προς τα κάτω. Τα περισσότερα σημεία θα βρίσκονται μεταξύ των 2 κάθετων τυπικών αποκλίσεων πάνω ή κάτω από το σημείο που αντιστοιχεί στον μέσο (σχήμα γ).

Σύνοψη ενός διαγράμματος σημείων

(α) Το σημείο των μέσων (β) Η οριζόντια τυπ.απ. (γ) Η κάθετη τυπ.απ.



Μέχρι τώρα, η περίληψη των στατιστικών συναρτήσεων που γνωρίζουμε είναι

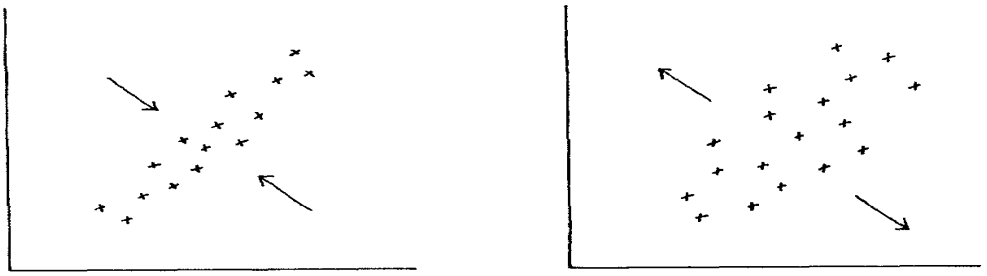
- ο μέσος των x-τιμών και η τυπική απόκλιση των x-τιμών.
- ο μέσος των y-τιμών και η τυπική απόκλιση των y-τιμών.

Αυτές οι στατιστικές συναρτήσεις μας δείχνουν το κέντρο του νέφους και πόσο "απλωμένο" είναι τόσο οριζοντίως όσο και καθέτως. Παρόλα αυτά υπάρχει και κάτι ακόμη που λείπει - η "ένταση" της συσχέτισης μεταξύ των δύο μεταβλητών. Ας δούμε το διάγραμμα σημείων του παρακάτω σχήματος.

Σύνοψη ενός διαγράμματος σημείων. Ο συντελεστής συσχέτισης μετρά την συγκέντρωση γύρω από μία ευθεία γραμμή

(α) Συσχέτιση κοντά στο 1
σημαίνει ισχυρή συγκέντρωση

(β) Συσχέτιση κοντά στο 0
σημαίνει χαλαρή συγκέντρωση



Και τα δύο αυτά νέφη σημείων έχουν το ίδιο "κέντρο" και επιδεικνύουν το ίδιο "άπλωμα" οριζόντια και κατακόρυφα. Παρ' όλα αυτά, τα σημεία στο πρώτο νέφος, συγκεντρώνονται "σφιχτά" γύρω από μία γραμμή. Διαφαίνεται, δηλαδή, μια ισχυρή γραμμική συσχέτιση μεταξύ των δύο μεταβλητών. Στο δεύτερο νέφος, το

"άπλωμα" είναι περισσότερο χαλαρό. Η ισχύς της συσχέτισης είναι διαφορετική στα δύο διαγράμματα.

Προκειμένου να μετρηθεί η δειγματική συσχέτιση, χρειάζεται μία ακόμα στατιστική συνάρτηση αυτή, που ονομάζουμε *δειγματικό συντελεστή συσχέτισης* (sample correlation coefficient). Ο συντελεστής αυτός συνήθως συμβολίζεται με r. Για ένα δείγμα από ζεύγη σημείων που αναφέρονται σε δύο μεταβλητές, έχουμε ότι

Ο δειγματικός συντελεστής συσχέτισης είναι ένα μέτρο γραμμικής σχέσης ή "απλώματος" δειγματικών παρατηρήσεων γύρω από μία γραμμή. Η συσχέτιση μεταξύ δύο μεταβλητών μπορεί να συνοψισθεί με

- τον μέσο των τιμών του δείγματος για την μεταβλητή X και την τυπική απόκλιση των x-τιμών
- τον μέσο των τιμών του δείγματος για την μεταβλητή Y και την τυπική απόκλιση των y-τιμών
- τον δειγματικό συντελεστή συσχέτισης r.

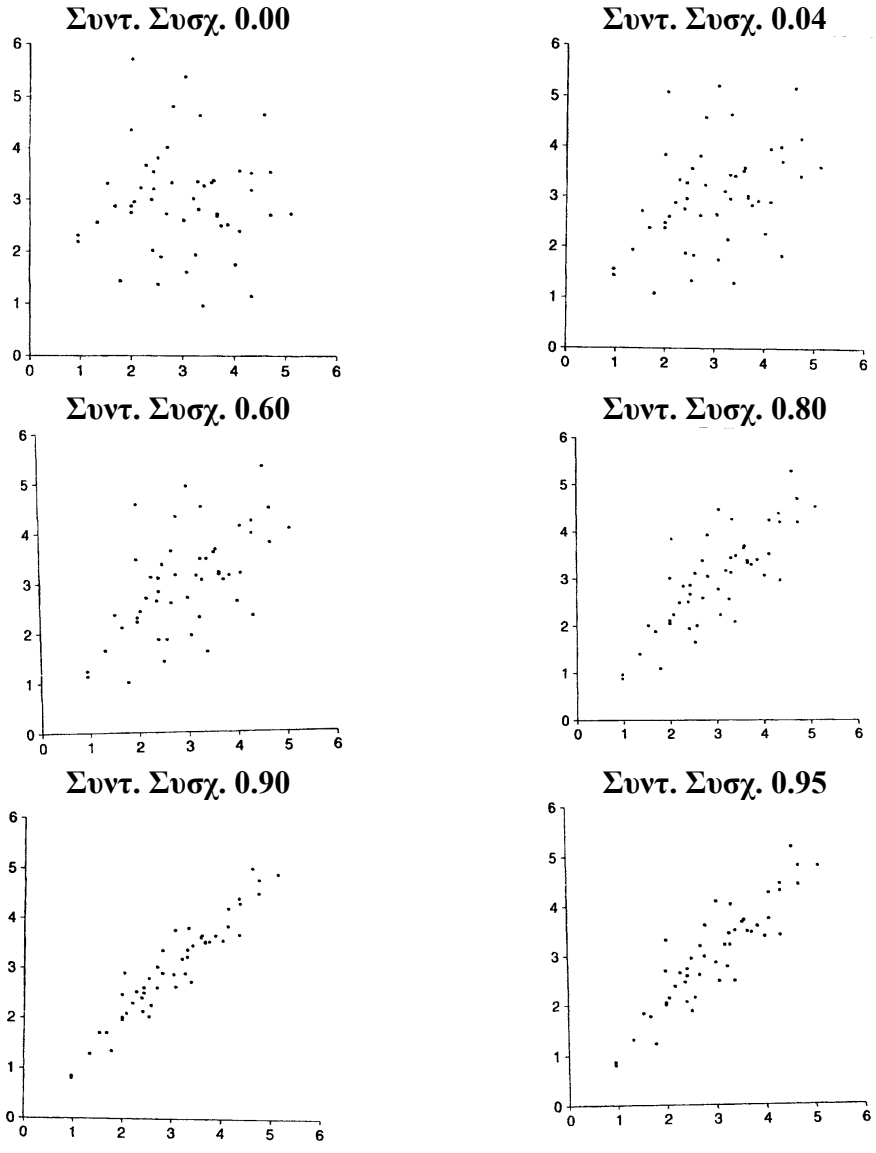
Στην συνέχεια, παρουσιάζονται κάποια γραφήματα που δίνουν μια εικόνα για διάφορες περιπτώσεις δειγματικού συντελεστή συσχέτισης δύο μεταβλητών.

Για παράδειγμα, μελέτες έχουν δείξει ότι ο συντελεστής συσχέτισης του ύψους ομοίων διδύμων παιδιών είναι περίπου 0.95 (βλέπε H.N. Newman, F.N. Freeman, and K.J. Holzinger (1937), Twins: Study of Heredity and Environment. University of Chicago Press).

Στο τελευταίο από τα γραφήματα που ακολουθούν, μπορεί κανείς να δει πώς εμφανίζονται σημεία που αντιστοιχούν σε συντελεστή συσχέτισης 0.95.

Ένα άλλο παράδειγμα αποτελεί ο συντελεστής συσχέτισης μεταξύ εισοδήματος και επιπέδου εκπαίδευσης. Σε μια μελέτη στις Ηνωμένες Πολιτείες το 1993, βρέθηκε ότι για άντρες ηλικίας 18-24 ετών, ο συντελεστής συσχέτισης εισοδήματος και εκπαίδευσης είναι μόνο 0.15, ενώ αυξάνεται και φθάνει το 0.45 για άντρες ηλικίας 55-64 ετών.

Σχήμα: Ο δειγματικός συντελεστής συσχέτισης - 6 ζευγών θετικά συσχετισμένων τυχαίων μεταβλητών. Τα διαγράμματα διαβαθμίζονται έτσι ώστε ο μέσος να ισούται με 3 και η τυπική απόκλιση με 1, οριζοντίως και καθέτως. Υπάρχουν 50 δειγματικά σημεία σε κάθε διάγραμμα. Η συγκέντρωση έχει μετρηθεί με τον δειγματικό συντελεστή συσχέτισης.

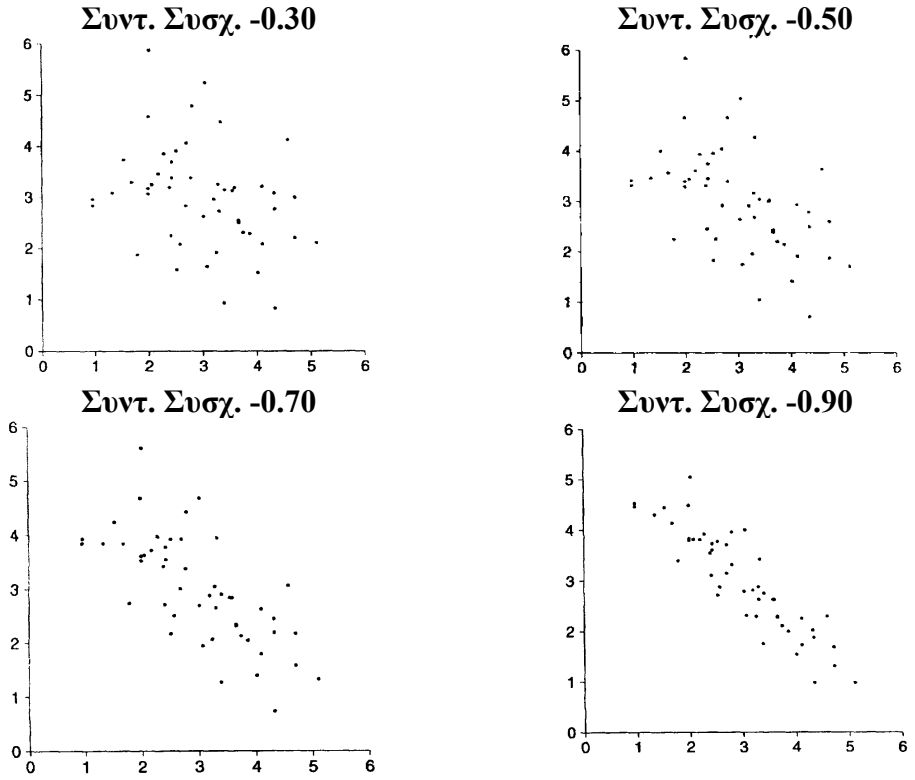


Σημείωση: Θα πρέπει να είναι κανείς προσεκτικός στην ερμηνεία του δειγματικού συντελεστή συσχέτισης. Αν r = 0.80, δεν σημαίνει ότι το

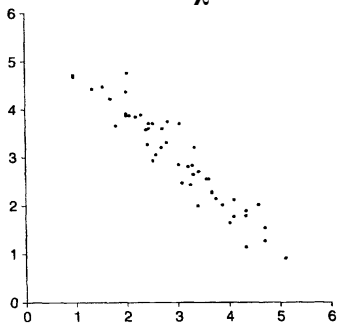
80% των σημείων βρίσκονται πολύ κοντά σε μια γραμμή, ούτε σημαίνει ότι υπάρχει διπλάσια γραμμική συσχέτιση από αυτή της περίπτωσης σημείων που έχουν r = 0.40.

Η αρνητική συσχέτιση δηλώνεται με αρνητικό πρόσημο στον συντελεστή συσχέτισης. Στα διαγράμματα που ακολουθούν εμφανίζονται σημεία που έχουν αρνητική συσχέτιση.

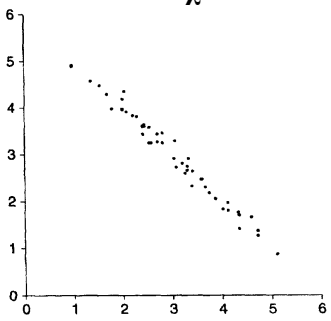
Σχήμα: Ο δειγματικός συντελεστής συσχέτισης - 6 ζευγών αρνητικά συσχετισμένων τυχαίων μεταβλητών. Τα διαγράμματα διαβαθμίζονται έτσι ώστε ο μέσος να ισούται με 3 και η τυπική απόκλιση με 1, οριζοντίως και καθέτως. Υπάρχουν 50 δειγματικά σημεία σε κάθε διάγραμμα. Η συγκέντρωση έχει μετρηθεί με τον δειγματικό συντελεστή συσχέτισης.



Συντ. Συσχ. -0.95



Συντ. Συσχ. -0.99



Ένας συντελεστής συσχέτισης -0.90 παρουσιάζει τον ίδιο βαθμό "απλώματος" όπως ένας συντελεστής 0.90. Το αρνητικό σημείο αναφέρεται σε "άπλωμα" γύρω από μία γραμμή με κλίση προς τα κάτω. Αρνητικός συντελεστής αναφέρεται σε γραμμή με κλίση προς τα πάνω. Για παράδειγμα, στην μελέτη που προαναφέρθηκε, προέκυψε ότι γυναίκες ηλικίας 25-39 ετών στις Ηνωμένες Πολιτείες το 1993 είχαν συντελεστή συσχέτισης μεταξύ εκπαίδευσης και αριθμού παιδιών -0.025 που αντιστοιχεί σε μη ισχυρή αρνητική συσχέτιση.

Ένας τέλειος συντελεστής 1 υποδηλώνει ότι όλα τα σημεία βρίσκονται πάνω σε μια γραμμή με ανοδική κλίση. Ένας τέλειος αρνητικός συντελεστής -1 αποτελεί ένδειξη ότι όλα τα σημεία βρίσκονται σε μια γραμμή με κλίση προς τα κάτω.

Υπολογισμός του Δειγματικού Συντελεστή Συσχέτισης

Ο δειγματικός συντελεστής συσχέτισης είναι ο μέσος των γινομένων των τυποποιημένων δειγματικών τιμών των ζευγών x και y των μεταβλητών X και Y.

Ο μαθηματικός τύπος που δίνει τον δειγματικό συντελεστή συσχέτισης είναι

$$r = \frac{\mathrm{Cov}(x, y)}{s_x s_y}$$

$$\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

Ο τύπος αυτός αποτελεί το δειγματικό ανάλογο του αντίστοιχου τύπου που ήδη είδαμε για τον συντελεστή συσχέτισης του πληθυσμού. Όπως έχει αναφερθεί, ο λόγος που χρησιμοποιούμε τις τυποποιημένες τιμές των X και Y είναι για να αποφύγουμε τις δυσκολίες που προκαλούνται, όταν τα μεγέθη την σχέση των οποίων θέλουμε να προσδιορίσουμε μετρώνται σε διαφορετικές κλίμακες. (Όπως είναι γνωστό, ο τρόπος για να αποφεύγεται το πρόβλημα αυτό είναι η τυποποίηση των τιμών. Η τυποποιημένη τιμή μιας τυχαίας μεταβλητής (που προκύπτει αν από την τιμή αφαιρέσουμε την μέση τιμή των παρατηρήσεων και την διαφορά την διαιρέσουμε με την τυπική απόκλιση των παρατηρήσεων) δείχνει, για κάθε τιμή, πόσες φορές η τιμή αυτή είναι πάνω ή κάτω από την μέση τιμή σε σχέση με την τυπική απόκλιση).

Η διαδικασία, δηλαδή, για τον υπολογισμό του δειγματικού συντελεστή συσχέτισης είναι ο μετασχηματισμός των τιμών κάθε μεταβλητής στις αντίστοιχες τυποποιημένες τιμές τους και στην συνέχεια ο υπολογισμός του μέσου των γινομένων των τυποποιημένων αυτών τιμών.

Αριθμητικό παράδειγμα: Να υπολογισθεί ο δειγματικός συντελεστής συσχέτισης r για τα δεδομένα που ακολουθούν.

x	y
1	5
3	9
4	7
5	1
7	13

Παρατήρηση: Η πρώτη γραμμή του παραπάνω πίνακα αποτυπώνει δύο μετρήσεις του ιδίου αντικειμένου στην συγκεκριμένη μελέτη. Οι δύο αριθμοί είναι οι x και y συντεταγμένες του αντιστοίχου σημείου στο διάγραμμα (νέφος) σημείων. Το ίδιο συμβαίνει και με τις

υπόλοιπες γραμμές. Εχει σημασία να χρησιμοποιηθούν οι αριθμοί του πίνακα ως ζεύγη. Ο συντελεστής συσχέτισης ορίζεται μόνο όταν έχουμε δύο μεταβλητές οι οποίες μετρώνται για κάθε αντικείμενο της μελέτης.

Βήμα 1: Μετατρέπουμε τις τιμές της μεταβλητής X σε τυποποιημένες τιμές. Για να γίνει αυτό χρειαζόμαστε τον μέσο $\overline{x}$ και την τυπική απόκλιση s_x των τιμών του x.
Βρίσκουμε ότι $\overline{x} = 4$, $s_x = 2$.
Στην συνέχεια, αφαιρούμε τον μέσο από κάθε τιμή της μεταβλητής X και διαιρούμε με την τυπική απόκλιση.

$$\frac{1-4}{2} = -1.5 \quad \frac{3-4}{2} = -0.5 \quad \frac{4-4}{2} = 0 \quad \frac{5-4}{2} = 0.5 \quad \frac{7-4}{2} = 1.5$$

Οι αριθμοί αυτοί προσδιορίζουν πόσο μακριά, πάνω ή κάτω, βρίσκονται οι τιμές της μεταβλητής X από τον μέσο όρο τους σε σχέση με την τυπική απόκλιση. Στο παράδειγμά, μας η τιμή 1 είναι 1.5 τυπικές αποκλίσεις χαμηλότερη από τον μέσο.

Βήμα 2: Μετατρέπουμε τις τιμές της μεταβλητής Y σε τυπικές τιμές.

Βήμα 3: Για κάθε ένα από τα ζευγάρια των τυπικών τιμών υπολογίζουμε το γινόμενο: (τυπική τιμή του x) * (τυπική τιμή του y).

Βήμα 4: Θεωρούμε τον μέσο των γινομένων

$$r = \text{μέσος (τυπικών τιμών του x) * (τυπικών τιμών του y)} =$$
$$= \frac{0.75 - 0.25 + 0.00 - 0.75 + 2.25}{5} = 0.40$$

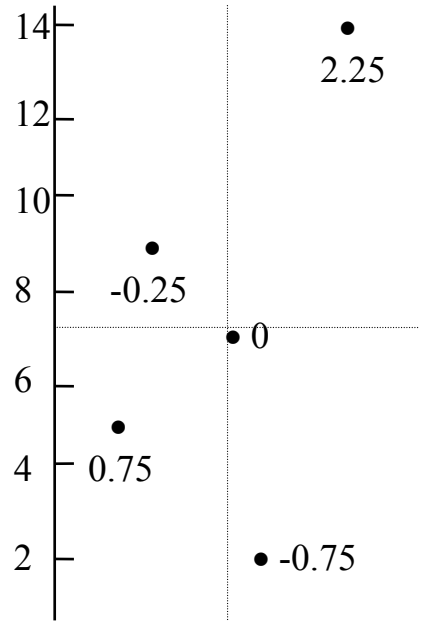
Τα αποτελέσματα αυτά συνήθως τοποθετούνται σε ένα πίνακα με την μορφή του πίνακα που ακολουθεί.

Υπολογισμός του Συντελεστή Συσχέτισης r

x	y	Τυποποιημένες τιμές του x	Τυποποιημένες τιμές του y	Γινόμενο
1	5	-1.5	-0.5	0.75
3	9	-0.5	0.5	-0.25
4	7	0.0	0.0	0.00
5	1	0.5	-1.5	-0.75
7	13	1.5	1.5	2.25

Εάν θεωρήσουμε το διάγραμμα του παραδείγματος αυτού, θα δούμε ότι η κλίση της ευθείας συσχέτισης είναι θετική, αλλά τα σημεία υποδηλώνουν ένα χαλαρό "άπλωμα" γύρω από την ευθεία.

Διάγραμμα σημείων του συντελεστή συσχέτισης (της τυποποιημένης συνδιακύμανσης) του αριθμητικού παραδείγματος

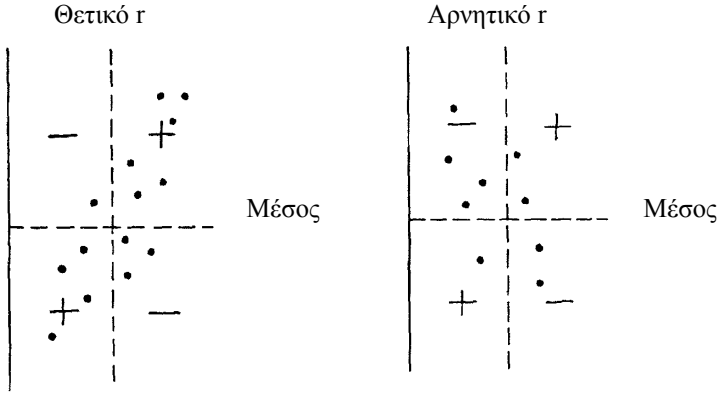


Ποιοί είναι οι λόγοι που κάνουν τον συντελεστή συσχέτισης ένα κατάλληλο μέτρο για γραμμική σχέση;

Στο σχήμα που προηγήθηκε, τα γινόμενα των τυποποιημένων τιμών εμφανίζονται ως σημεία. Εχουν επίσης συρθεί η οριζόντα και η κατακόρυφη γραμμή που περνά από το σημείο των μέσων των x και y

(Προφανώς, το σημείο αυτό είναι το μηδέν αφού μιλάμε γαι τυποποιημένες τιμές). Οι γραμμές αυτές χωρίζουν το διάγραμμα σημείων σε τέσσερα τετράγωνα. Αν ένα σημείο βρίσκεται στο κάτω αριστερό τετράγωνο, το σημείο αυτό αντιστοιχεί σε ζεύγος τιμών των μεταβλητών που είναι μικρότερες από τις αντίστοιχες μέσες τιμές τους και επομένως έχουν αρνητικές τυποποιημένες τιμές. (Το γινόμενο δύο αρνητικών αριθμών είναι θετικός αριθμός). Στο πάνω δεξιό τετράγωνο, το γινόμενο δύο θετικών τιμών είναι θετικό. Στα δύο άλλα τετράγωνα, το γινόμενο ενός θετικού και ενός αρνητικού αριθμού είναι αρνητικός αριθμός. Ο μέσος όλων αυτών των γινομένων είναι ο δειγματικός συντελεστής συσχέτισης. Αν το r είναι θετικό, τότε τα σημεία που βρίσκονται στα δύο "θετικά τετράγωνα" θα υπερισχύουν, όπως αυτό φαίνεται στο πρώτο από τα δύο σχήματα που ακολουθούν. Αν το r είναι αρνητικό, τότε τα σημεία στα δύο αρνητικά τετράγωνα θα υπερισχύουν όπως φαίνεται στο δεύτερο από τα σχήματα που ακολουθούν.

Η Ευθεία Τυπικής Απόκλισης (The Standard Deviation Line)



Ο ΝΟΜΟΣ ΤΩΝ ΜΕΓΑΛΩΝ ΑΡΙΘΜΩΝ

(Law of Large Numbers)

Η μεγάλη εμπειρία που έχει συσσωρευθεί από τον άνθρωπο μας διδάσκει ότι φαινόμενα που έχουν πιθανότητα πολύ κοντά στην μονάδα συμβαίνουν σχεδόν οπωσδήποτε. Αντίστροφα, ενδεχόμενα των οποίων η πιθανότητα να συμβούν είναι πολύ μικρή (πολύ κοντά στο μηδέν) συμβαίνουν πάρα πολύ σπάνια. Η παρατήρηση αυτή παίζει βασικό ρόλο στις πρακτικές εφαρμογές της Θεωρίας

Πιθανοτήτων, γιατί αυτό το *πειραματικό γεγονός* επιτρέπει την εξής θεώρηση στην πράξη:

Ενδεχόμενα που έχουν πολύ μικρή πιθανότητα να συμβούν θεωρούνται *πρακτικά αδύνατα* και ενδεχόμενα τα οποία μπορούν να συμβούν με πιθανότητα πολύ κοντά στην μονάδα θεωρούνται *πρακτικά βέβαια γεγονότα*. Παρ' όλα αυτά δεν είμαστε σε θέση να δώσουμε μια σαφή απάντηση στο πολύ φυσικό ερώτημα: Ποιά πρέπει να είναι η τιμή της πιθανότητας ώστε να μπορούμε να θεωρήσουμε ότι ένα ενδεχόμενο είναι πρακτικά αδύνατο (αντίστοιχα πρακτικά σίγουρο); Το μόνο βέβαιο είναι ότι τα κριτήρια σύμφωνα με τα οποία ενδεχόμενα θεωρούνται πρακτικά αδύνατα ή πρακτικά σίγουρα υπαγορεύονται μόνο από τις απαιτήσεις της πρακτικής. Για παράδειγμα, εάν κατά τη μέτρηση της απόστασης μεταξύ δύο σημείων αυτή βρέθηκε ίση με 5340 μέτρα και το σφάλμα αυτής της μέτρησης είναι μεγαλύτερο ή ίσο από 20 μέτρα με πιθανότητα 0.02, θα μπορούσαμε να αγνοήσουμε την πιθανότητα ενός τέτοιου σφάλματος και να θεωρήσουμε ότι η απόσταση είναι πράγματι ίση με 5340 μέτρα. Επομένως, στην περίπτωση αυτή θεωρούμε ότι το ενδεχόμενο που έχει πιθανότητα 0.02 είναι πρακτικά ασήμαντο και το αγνοούμε στην πράξη. Όμως, κάτω από άλλες περιπτώσεις δεν μπορούμε να αγνοούμε πιθανότητες της τάξης του 0.02 ή και μικρότερες ακόμη. Ας θεωρήσουμε, για παράδειγμα, την περίπτωση της κατασκευής ενός μεγάλου υδροηλεκτρικού σταθμού, η οποία απαιτεί μεγάλες δαπάνες σε υλικά και σε ανθρώπινο δυναμικό. Ας υποθέσουμε ότι υπολογίσθηκε ότι η πιθανότητα που έχει το νερό να φθάσει σε ένα επίπεδο που θα απειλεί καταστροφικές πλημμύρες κάτω από τις παρούσες συνθήκες είναι ίση με 0.02. Τότε η πιθανότητα αυτή θα εθεωρείτο υψηλή και θα έπρεπε να ληφθεί υπόψη στον σχεδιασμό αυτού του σταθμού και επομένως να μην αγνοηθεί όπως στο προηγούμενο παράδειγμα.

Ταυτόχρονα, θα πρέπει να σημειωθεί ότι κάθε ενδεχόμενο που έχει θετική πιθανότητα, ανεξάρτητα από το πόσο μικρή είναι αυτή, μπορεί να συμβεί. Στην περίπτωση αυτή αν ο αριθμός των δοκιμών είναι πολύ μεγάλος, τότε η πιθανότητα μίας τουλάχιστον εμφάνισης του ενδεχομένου αυτού μπορεί να είναι οσοδήποτε κοντά στη μονάδα.

Από την άλλη μεριά, εάν η πιθανότητα κάποιου ενδεχομένου είναι πολύ μικρή, τότε είναι πολύ δύσκολο να αναμένει κανείς την εμφάνιση του ενδεχομένου αυτού σε κάποια συγκεκριμένη δοκιμή. Έτσι, εάν κάποιος προβλέψει ότι, στην πρώτη μοιρασιά τραπουλόχαρτων μεταξύ 4 παικτών, κάθε παίκτης θα πάρει στα χέρια του κάρτες ενός μόνο είδους (π.χ. κούπες), τότε είναι φυσικό να υποπτευθούμε ότι το πρόσωπο αυτό γνωρίζει ότι η τράπουλα είναι "στημένη". Η υποψία αυτή δικαιολογείται από το γεγονός ότι η πιθανότητα μιας τέτοιας μοιρασιάς, αν η τράπουλα είναι καλά ανακατεμένη, είναι ίση με

$$(9!)^4\ 4!/36! < 1.1 \text{ x } 10^{-18}$$

η οποία είναι πάρα πολύ μικρή. Η περίπτωση αυτή αποτελεί ένα πολύ καλό παράδειγμα για την διαφορά η οποία υπάρχει μεταξύ της έννοιας του πρακτικά αδύνατου ενδεχομένου και του πραγματικά αδύνατου ενδεχομένου.

Από όσα αναφέραμε προηγουμένως, γίνεται σαφές ότι τόσο στις πρακτικές εφαρμογές όσο και στα γενικότερα θεωρητικά προβλήματα, τα ενδεχόμενα με πιθανότητες πολύ κοντά στην μονάδα ή στο μηδέν έχουν μεγάλη σημασία. Επομένως, η εδραίωση κανόνων οι οποίοι αναφέρονται σε πιθανότητες πολύ κοντά στην μονάδα είναι μεγάλης σημασίας και ιδιαίτερα σημαντικός γίνεται ο ρόλος των νόμων, οι οποίοι προκύπτουν ως αποτέλεσμα της παρουσίας ενός μεγάλου αριθμού ανεξάρτητων ή ασθενώς εξαρτημένων τυχαίων παραγόντων. Ο νόμος των μεγάλων αριθμών είναι ένας τέτοιος νόμος της Θεωρίας των Πιθανοτήτων – ο πιο σημαντικός.

Είναι φυσικό να θεωρεί κανείς τον νόμο των μεγάλων αριθμών ως εκφράζοντα την ενοποίηση όλων των προτάσεων (αποτελεσμάτων) που καταλήγουν στο συμπέρασμα ότι με πιθανότητα πολύ κοντά στην μονάδα κάποιο ενδεχόμενο θα συμβεί. Το ενδεχόμενο αυτό εξαρτάται από ένα άπειρα αυξανόμενο αριθμό τυχαίων ενδεχομένων κάθε ένα από τα οποία έχει μια πολύ μικρή συνεισφορά στο συγκεκριμένο ενδεχόμενο.

Αυτή η γενική αντίληψη για τα θεωρήματα που σχετίζονται με το νόμο των μεγάλων αριθμών, μεταξύ των οποίων συγκαταλέγονται

και τα θεωρήματα σύγκλισης που ήδη εξετάσαμε, μπορεί να διατυπωθεί με περισσότερο συγκεκριμένο τρόπο.

Θεώρημα *(νόμος των μεγάλων αριθμών)*: Έστω X_1, X_2, …, X_n μία ακολουθία τυχαίων μεταβλητών. Ας θεωρήσουμε τις μεταβλητές Y_n οι οποίες ορίζονται ως συμμετρικές συναρτήσεις των τυχαίων μεταβλητών X_i, i = 1,2,…,n από την σχέση $Y_n = f_n(X_1, X_2, ..., X_n)$.
Η ακολουθία των τυχαίων μεταβλητών $\{X_n\}$ λέγεται ότι ακολουθεί τον *νόμο των μεγάλων αριθμών (law of large numbers)* για δεδομένες μορφές των συναρτήσεων f_n, τότε και μόνο τότε εάν υπάρχει μία ακολουθία σταθερών α_1, α_2, ..., α_n τέτοια ώστε, για κάθε ε>0

$$\lim_{n\to\infty} P\left(\left|Y_n - \alpha_n\right| < \varepsilon\right) = 1$$

Συνήθως, προσδίδεται μία περισσότερο συγκεκριμένη ερμηνεία στην έννοια του νόμου των μεγάλων αριθμών. Συγκεκριμένα, περιοριζόμαστε στην περίπτωση όπου η συνάρτηση f_n είναι ο μέσος των μεταβλητών X_1, X_2, …, X_n. Στην περίπτωση αυτή ο νόμος των μεγάλων αριθμών μας λέει ότι, για αρκετά μεγάλο n, η πιθανότητα ότι η τιμή του μέσου των τυχαίων μεταβλητών (η τιμή του μέσου του δείγματος) διαφέρει από το μ το πολύ κατά μια προκαθορισμένη σταθερά μπορεί να πλησιάσει την μονάδα (την βεβαιότητα) όσο εμείς θέλουμε. Αυτό ουσιαστικά συνεπάγεται ότι, για μεγάλο n, ο μέσος του τυχαίου δείγματος είναι "κοντά" στην μέση τιμή του πληθυσμού.

Κάτι αντίστοιχο ισχύει για τα ποσοστά. Συγκεκριμένα, αν X~b(x; n,p), γνωρίζουμε ότι X/n είναι το ποσοστό (η σχετική συχνότητα) "επιτυχιών" στις n δοκιμές και είναι μια εκτιμήτρια του p.

Από τον νόμο των μεγάλων αριθμών προκύπτει ότι, η πιθανότητα ότι η σχετική συχνότητα X/n απέχει λιγότερο από ε από το p είναι κοντά στην μονάδα, όταν ο αριθμός των δοκιμών n είναι αρκετά μεγάλος. Η παρατήρηση αυτή ενισχύει τον ορισμό της πιθανότητας ως σχετικής συχνότητας.

Όταν μελετάμε απλά φαινόμενα, τα παρατηρούμε μαζί με όλες τις ιδιαιτερότητες που τα συνοδεύουν. Οι ιδιαιτερότητες αυτές μας εμποδίζουν να αντιληφθούμε τους νόμους που διέπουν την μελέτη ενός μεγάλου αριθμού παρόμοιων φαινομένων. Από πολύ παλιά είχε παρατηρηθεί ότι παράγοντες οι οποίοι δεν συνδέονται με την ουσία

της διεργασίας που προκάλεσε το φαινόμενο ως σύνολο και οι οποίοι εμφανίζονται μόνο σε ορισμένες μεμονωμένες περιπτώσεις, εξουδετερώνονται αμοιβαία, όταν θεωρηθεί ο μέσος ενός μεγάλου αριθμού παρατηρήσεων. Αργότερα, αυτό το εμπειρικό αποτέλεσμα παρατηρήθηκε κατ' επανάληψη χωρίς όμως να έγινε ποτέ κάποια προσπάθεια για να εξηγηθεί θεωρητικά. Επιπλέον, σύμφωνα με πολλούς συγγραφείς δεν χρειαζόταν καμιά ερμηνεία γιατί η παρουσία "τάξης" (κανόνων) τόσο στα φυσικά όσο και στα κοινωνικά φαινόμενα δεν ήταν για αυτούς παρά τα αποτελέσματα θετικής επέμβασης.

Ακόμα και σήμερα, μερικοί συγγραφείς υποβαθμίζουν το περιεχόμενο του νόμου των μεγάλων αριθμών και διαστρεβλώνουν τον πειραματικά παρατηρούμενο κανόνα. Στην πραγματικότητα, η επιστημονική αξία της έρευνας του Chebyshev, του Markov και άλλων ερευνητών στην περιοχή του νόμου των μεγάλων αριθμών δεν βρίσκεται στο γεγονός ότι οι ερευνητές αυτοί ανίχνευσαν την εμπειρική σταθερότητα των μέσων, αλλά στο γεγονός ότι βρήκαν τις γενικές προϋποθέσεις των οποίων η ικανοποίηση αναδεικνύει την στατιστική σταθερότητα των μέσων.

Ως παράδειγμα για τον τρόπο εφαρμογής του νόμου των μεγάλων αριθμών, θεωρούμε την εξής περίπτωση: Σύμφωνα με τους νόμους της Φυσικής, ένα αέριο αποτελείται από έναν μεγάλο αριθμό σωματιδίων που βρίσκονται σε συνεχή και χαοτική κίνηση. Όσο αφορά κάθε ένα από αυτά τα μόρια, δεν είναι δυνατόν να προβλέψουμε την ταχύτητα που θα έχει και το σημείο στο οποίο θα βρίσκεται σε κάποια χρονική στιγμή. Όμως, μπορούμε, εάν δοθούν ορισμένες συνθήκες, να υπολογίσουμε το ποσοστό των μορίων που θα κινούνται με μία δεδομένη ταχύτητα. Αυτό, όμως, είναι ακριβώς ό,τι χρειάζεται να γνωρίζει ο φυσικός, γιατί τα βασικά χαρακτηριστικά ενός αερίου, όπως η πίεση, η θερμοκρασία κ.λ.π. καθορίζονται όχι από την συμπεριφορά ενός μορίου του αλλά από την από κοινού συμπεριφορά όλων των μορίων του.