ΚΕΦΑΛΑΙΟ 9

Η ΚΑΝΟΝΙΚΗ ΚΑΤΑΝΟΜΗ

Η κανονική κατανομή ανακαλύφθηκε γύρω στο 1720 από τον Abraham De Moivre στην προσπάθειά του να διαμορφώσει Μαθηματικά που να εξηγούν την τυχαιότητα. Γύρω στο 1870, ο Βέλγος Μαθηματικός Adolph Queteler είχε την ιδέα να χρησιμοποιήσει την καμπύλη της κατανομής αυτής ως ένα ιδανικό ιστόγραμμα με το οποίο θα μπορούσαν να συγκρίνονται άλλα ιστογράμματα που αντιστοιχούσαν σε δεδομένα.

Ορισμός: Έστω X μια (απόλυτα) συνεχής τυχαία μεταβλητή. Το X ακολουθεί την *κανονική κατανομή (normal distribution) με παραμέτρους* m *και* σ^2 , $\sigma>0$ (συμβολικά $X \sim N(\mu,\sigma^2)$) αν

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}, \quad -\infty<x<+\infty, \quad \sigma>0$$

Η κανονική κατανομή είναι η πιο σημαντική και χρήσιμη κατανομή πιθανότητας. Αυτό γιατί:
i) Πολλά πειράματα μπορούν να εκφρασθούν μέσω τυχαίων μεταβλητών που ακολουθούν την κανονική κατανομή.
ii) Η κανονική κατανομή μπορεί να χρησιμοποιηθεί σαν προσέγγιση πολλών άλλων κατανομών.
iii) κατανομή αυτή αποτελεί την βάση για πολλές τεχνικές που χρησιμοποιούνται στην στατιστική συμπερασματολογία.

Ιδιότητες:
α) Η κανονική κατανομή είναι συμμετρική γύρω από το σημείο $x=\mu$.
β) Έχει σχήμα καμπάνας με επικρατούσα τιμή στο σημείο $x=\mu$. Στο σημείο αυτό

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}$$

και επομένως το σ προσδιορίζει το μέγιστο της συνάρτησης. Είναι φανερό ότι η συνάρτηση f(x) στο σημείο x=μ είναι αντιστρόφως ανάλογη της τιμής του σ.

Σημείωση: Η κανονική κατανομή είναι μια καλά ορισμένη κατανομή. Πράγματι

$$\int_{-\infty}^{+\infty} f(x)dx = 2\int_{\mu}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Θέτοντας $\frac{x-\mu}{\sigma} = y$ έχουμε

$dy = \frac{1}{\sigma}dx$ και έτσι

$$\int_{-\infty}^{+\infty} f(x)dx = 2\int_{0}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy$$

Θέτωντας $y^2/2=u$ έχουμε

$$ydy = du \quad \text{δηλαδή} \quad dy = \frac{1}{\sqrt{2u}}du$$

Έτσι

$$\int_{-\infty}^{+\infty} f(x)dx = 2\int_{0}^{+\infty} \frac{1}{\sqrt{2\pi}}\frac{1}{2u}e^{-u}du = \frac{1}{\sqrt{\pi}}\int_{0}^{+\infty} e^{-u}u^{-1/2}du =$$

$$= \frac{\Gamma(1/2)}{\sqrt{\pi}} = \frac{\sqrt{\pi}}{\sqrt{\pi}} = 1$$

Ιδιότητες:

$E(X) = \mu, \quad \Delta(X) = \sigma^2$

Υπολογισμός Πιθανοτήτων της Κανονικής Κατανομής

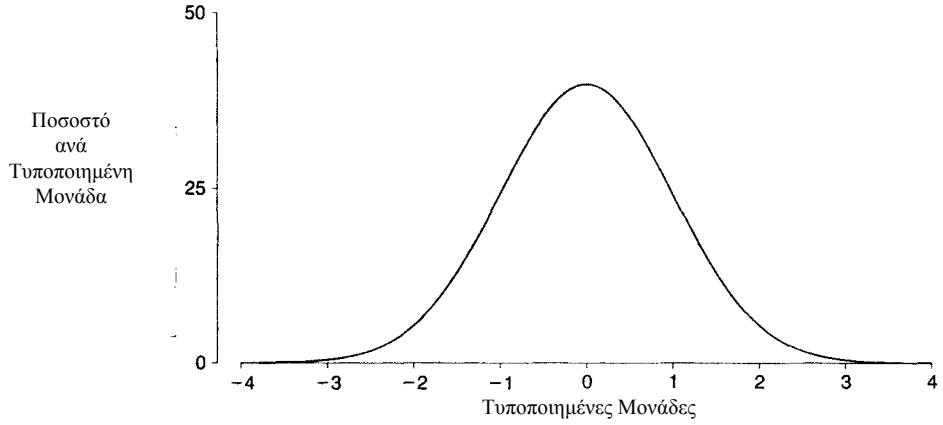
Ο απευθείας υπολογισμός πιθανοτήτων της κανονικής κατανομής είναι δύσκολος λόγω της μορφής της συνάρτησης πυκνότητας πιθανότητας. Ο υπολογισμός αυτός όμως διευκολύνεται με τον ορισμό της τυποποιημένης κανονικής κατανομής.

Η απλούστερη μορφή της κατανομής αυτής, η οποία συνήθως χρησιμοποιείται σε πρακτικές εφαρμογές με μετασχηματισμό, είναι εκείνη που αναφέρεται στην περίπτωση όπου μ=0 και σ^2=1.

Η κατανομή αυτή ονομάζεται *τυποποιημένη κανονική κατανομή* και έχει, προφανώς, την μορφή

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

Σχήμα: Η τυποποιημένη κανονική κατανομή



Όπως είναι προφανές τόσο από το γράφημα της κατανομής, αλλά και από την μαθηματική της μορφή, η κατανομή αυτή είναι συμμετρική γύρω από την μέση τιμή της. Προφανώς, επίσης, όπως σε όλες τις κατανομές, το εμβαδόν της επιφάνειας κάτω από την καμπύλη είναι 1.

Στο γράφημα της τυποποιημένης κανονικής κατανομής, η καμπύλη εμφανίζεται να σταματά σε κάποιο σημείο μεταξύ 3 και 4 και -3 και -4, αντίστοιχα. Στην πραγματικότητα, στα σημεία αυτά, πλησιάζει ασυμπτωτικά τον άξονα των x. Περίπου το 6/10000 μόνο της επιφάνειάς της βρίσκεται έξω από το διάστημα -4 έως 4.

Ορισμός: Αν μ=0, σ=1 (αν δηλαδή X~N(0,1)) λέμε ότι το X ακολουθεί την *τυποποιημένη κανονική κατανομή (standard normal*

distribution). Η συνάρτηση κατανομής επομένως της τυποποιημένης κανονικής κατανομής είναι

$$\Phi(z) = P(-\infty < Z < x) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

Υπάρχουν πίνακες που δίνουν τιμές της συνάρτησης κατανομής της τυποποιημένης κανονικής κατανομής για διάφορες τιμές του z. (Βλέπε παράρτημα).

Παράδειγμα: Έστω $Z \sim N(0,1)$.

α) Να υπολογισθεί η $P(-1 \le Z \le 1)$.

β) Να βρεθεί η τιμή x έτσι ώστε $P(Z<x) = 0.3$.

Λύση:

α) $P(-1 \le Z \le 1) = \Phi(1) - \Phi(-1) = \Phi(1) - (1-\Phi(1)) = 2\Phi(1) - 1$
$= 2(0.8413)-1 = 0.6826$

β) Από τους πίνακες βρίσκουμε ότι η τιμή α με την ιδιότητα $P(Z<\alpha) = 0.7$ είναι $\alpha=0.52$. Επομένως $x = -\alpha = -0.52$

Ο υπολογισμός πιθανοτήτων που αναφέρονται σε μια τυχαία μεταβλητή $X \sim N(\mu,\sigma^2)$ γίνεται με χρήση των πινάκων της τυποποιημένης κανονικής κατανομής και βασίζεται στην εξής πρόταση.

Πρόταση: Αν $X \sim N(\mu,\sigma^2)$, τότε

$$F(x) = P(X \le x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

Απόδειξη: Προφανής.

Σημείωση: Χρησιμοποιώντας την ιδιότητα αυτή, είμαστε σε θέση να καθορίσουμε μια σταθερά c έτσι ώστε αν $X \sim N(\mu,\sigma^2)$ η $P(X \le c)=\alpha$. Αυτό, γιατί

$$P(X \le c) = \alpha \Leftrightarrow \Phi\left(\frac{c-\mu}{\sigma}\right) = \alpha$$

π.χ. αν $\mu=10$, $\sigma^2 =4$ και $\alpha=0.95$,

$$P(X\le c) = 0.95 \Leftrightarrow \Phi\left(\frac{c-10}{2}\right) = 0.95 \Leftrightarrow c=13.29$$

Παραδείγματα:

1. Η γραμμή του οπτικού πεδίου σύμφωνα με την οποία ένα βλήμα αποτυγχάνει τον στόχο του ακολουθεί την κανονική κατανομή με μέσο $\mu = -15$ μέτρα και διασπορά $25(\text{μέτρα})^2$. Να υπολογισθεί η πιθανότητα το βλήμα να αποτύχει τον στόχο σε απόσταση μεγαλύτερη από 20 μέτρα.

Λύση: Έστω X η απόκλιση του βλήματος από τον στόχο.
Τότε $X\sim N(-15, 25)$. Η ζητούμενη πιθανότητα δίδεται από τον τύπο

$$P(|X| > 20) = 1 - P(-20<X<20) = 1 - \{F(20) - F(-20)\} =$$
$$= 1 - \Phi\left(\frac{20+15}{5}\right) + \Phi\left(\frac{-20+15}{5}\right) = 1 - \Phi(7) + \Phi(-1) =$$
$$= 1 - 1 + 1 - \Phi(1) =$$
$$= 1 - 0.8413 = 0.1587$$

2. Σε ένα μετεωρολογικό σταθμό υπάρχουν στοιχεία για την μέγιστη θερμοκρασία της 1ης Ιουνίου για μια σειρά ετών. Περίπου 15% των φορών η μέγιστη αυτή θερμοκρασία ξεπέρασε του 30^oC ενώ περίπου 5% των φορών ήταν μικρότερη από 20^oC. Εάν υποτεθεί ότι η μέγιστη αυτή θερμοκρασία ακολουθεί την κανονική κατανομή να υπολογισθούν το μ και σ^2.

Λύση: Έστω T η μέγιστη θερμοκρασία της 1ης Ιουνίου.
Έχουμε ότι $T\sim N(\mu,\sigma^2)$. Επίσης,

$$P(T<20) = 0.05 \text{ και } P(T>30) = 0.15$$

ή ισοδύναμα

$$\Phi\left(\frac{20-\mu}{\sigma}\right) = 0.05 \quad \text{και} \quad 1-\Phi\left(\frac{30-\mu}{\sigma}\right) = 0.15$$

Από τις σχέσεις αυτές και από τους πίνακες της τυποποίημενης κανονικής κατανομής έχουμε

$$\frac{20-\mu}{\sigma} = -1.645 \quad \text{και} \quad \frac{30-\mu}{\sigma} = 1.036$$

Λύνοντας τις εξισώσεις αυτές, παίρνουμε $\mu=26.12$, $\sigma=3.73$.

Παρατήρηση: Σε προηγούμενη παρατήρηση σχετική με την ανισότητα του Chebyshev είχαμε δει ότι για οποιαδήποτε τυχαία μεταβλητή τουλάχιστον τα 3/4 των τιμών της βρίσκονται στο διάστημα [μ-2σ, μ+2σ] ενώ στο διάστημα [μ-3σ, μ+3σ] βρίσκονται τουλάχιστον τα 8/9 των τιμών της. Για την κανονική κατανομή όμως έχουμε ειδικότερα

$$P(\mu - \sigma < X < \mu + \sigma) = \Phi\left(\frac{\mu + \sigma - \mu}{\sigma}\right) - \Phi\left(\frac{\mu - \sigma - \mu}{\sigma}\right)$$
$$= \Phi(1) - \Phi(-1) = 0.6826$$

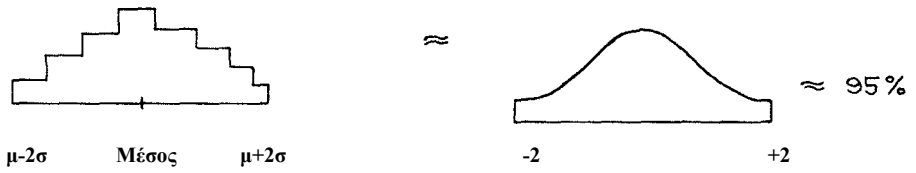
Επίσης,

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = \Phi\left(\frac{\mu + 2\sigma - \mu}{\sigma}\right) - \Phi\left(\frac{\mu - 2\sigma - \mu}{\sigma}\right)$$
$$= \Phi(2) - \Phi(-2) = 0.9544$$

Ενώ,

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = \Phi\left(\frac{\mu + 3\sigma - \mu}{\sigma}\right) - \Phi\left(\frac{\mu - 3\sigma - \mu}{\sigma}\right)$$
$$= \Phi(3) - \Phi(-3) = 0.9974$$

Δηλαδή, το 68% των τιμών μιας κανονικής τυχαίας μεταβλητής βρίσκεται στο διάστημα (μ ± σ), το 95% στο διάστημα (μ ± 2σ), και το 99.7% στο διάστημα (μ ± 3σ).



Ο προσδιορισμός του εμβαδού, επειδή ακριβώς αναφέρεται σε ολοκλήρωση μιας αρκετά πολύπλοκης συνάρτησης, δίνεται σχεδόν σε όλα τα βιβλία Στατιστικής σε πίνακες. Για διευκόλυνση, οι πίνακες αναφέρονται σε τιμές της τυποποιημένης κανονικής κατανομής.

ΤΟ ΚΕΝΤΡΙΚΟ ΟΡΙΑΚΟ ΘΕΩΡΗΜΑ

(Central Limit Theorem)

Όπως είδαμε, ο νόμος των μεγάλων αριθμών προσδιορίζει την μορφή σύγκλισης ακολουθιών τυχαίων μεταβλητών, οι οποίες μπορούν να εκφρασθούν ως μερικά αθροίσματα άλλων ανεξαρτήτων τυχαίων μεταβλητών. Λέει, συγκεκριμένα, ότι τυχαίες μεταβλητές της μορφής

$$S_n = X_1 + X_2 + ... + X_n$$

συγκλίνουν, με κάποια έννοια σύγκλισης, στην μέση τους τιμή, όπως αυτή ορίζεται από την σχέση

$$E(S_n) = \sum_{i=1}^{n} E(X_i)$$

Το επόμενο ερώτημα, που όπως είναι φυσικό ανακύπτει, έχει να κάνει με την κατανομή της τυχαίας μεταβλητής S_n. Απάντηση στο ερώτημα αυτό δίνει το κεντρικό οριακό θεώρημα. Συγκεκριμένα, αποδεικνύει ότι η κατανομή της τυχαίας μεταβλητής S_n είναι, κατά προσέγγιση, η κανονική κατανομή. Το γνωστό αυτό θεώρημα έχει γίνει αντικείμενο έντονης ερευνητικής δραστηριότητας με στόχο την ανακάλυψη των πιο γενικών συνθηκών κάτω από τις οποίες αυτό ισχύει. Από την άλλη μεριά, το θεώρημα αυτό έχει αποτελέσει την βάση μιας εκπληκτικής ποσότητας εφαρμοσμένης έρευνας. Στην θεωρία της μέτρησης των σφαλμάτων, το παρατηρούμενο σφάλμα μπορεί να εκφρασθεί ως άθροισμα ενός μεγάλου αριθμού ανεξάρτητων τυχαίων ποσοτήτων οι οποίες συνεισφέρουν στο αποτελέσμα. Επίσης, στην Στατιστική, ο μέσος n ανεξάρτητων τυχαίων μεταβλητών (που είναι ίσος με μία σταθερά (n^{-1}) επί το άθροισμα των μεταβλητών αυτών) είναι κεντρικής σημασίας. Σε τέτοιες περιπτώσεις, η υπόθεση μιας κανονικής κατανομής μπορεί να είναι κατάλληλη.

Στην συνέχεια, θα εξετάσουμε τις σημαντικότερες μορφές του κεντρικού οριακού θεωρήματος.

Κεντρικό Οριακό Θεώρημα: Έστω $\{X_n\}$ μια ακολουθία ανεξάρτητων τυχαίων μεταβλητών με πεπερασμένες μέσες τιμές και πεπερασμένες διασπορές. Ορίζουμε την ακολουθία $\{S_n\}$των μερικών αθροισμάτων

$$S_n = \sum_{i=1}^{n} X_i \text{ , για κάθε } n \geq 1$$

για την οποία ισχύει ότι

$$E(S_n) = \sum_{i=1}^{n} E(X_i)$$

και

$$V(S_n) = \sum_{i=1}^{n} V(X_i) \text{ για κάθε } n \geq 1$$

Ας θεωρήσουμε την τυχαία μεταβλητή

$$S_n^* = \frac{S_n - E(S_n)}{\sqrt{V(S_n)}}$$

δηλαδή την τυποποιημένη τυχαία μεταβλητή της S_n.
Έστω F_n η συνάρτηση κατανομής της τυχαίας μεταβλητής S_n^*. Τότε,

$$\lim_{n \to \infty} F_n(t) = \Phi(t), \text{ για κάθε } t \in (-\infty, +\infty)$$

όπου $\Phi(t)$ είναι η συνάρτηση κατανομής της τυποποιημένης κανονικής κατανομής.

Δηλαδή σύμφωνα με το κεντρικό οριακό θεώρημα, η κατανομή της τυποποιημένης μεταβλητής του αθροίσματος ανεξάρτητων μεταβλητών τείνει στην κανονική κατανομή όσο αυξάνει ο αριθμός των τυχαίων αυτών μεταβλητών.

Στην βιβλιογραφία συναντάται συχνά μια απλούστερη μορφή με την οποία το θεώρημα αυτό είναι περισσότερο γνωστό. Η μορφή αυτή αντιστοιχεί στην περίπτωση όπου

$$E(X_i) = \mu \text{ και } V(X_i) = \sigma^2 \text{ για κάθε } i$$

όπου δηλαδή όλες οι τυχαίες μεταβλητές του αθροίσματος έχουν την ίδια μέση τιμή και την ίδια διασπορά.
Η ειδική αυτή περίπτωση οδηγεί στην εξής μορφή του κεντρικού οριακού θεωρήματος.

Θεώρημα: Έστω $\{X_n\}$ μια ακολουθία ανεξάρτητων τυχαίων μεταβλητών με ίσες πεπερασμένες μέσες τιμές και ίσες πεπερασμένες διασπορές, δηλαδή

$$E(X_i) = \mu \text{ και } V(X_i) = \sigma^2 \text{ , για κάθε i}$$

Έστω επί πλέον η τυχαία μεταβλητή

$$S_n^* = \frac{S_n - E(S_n)}{\sqrt{V(S_n)}}$$

με συνάρτηση κατανομής F_n. Τότε

$$\lim_{n\to\infty} F_n(t) = \Phi(t), \text{ , για κάθε } t \in (-\infty, +\infty)$$

όπου $\Phi(t)$ είναι η συνάρτηση κατανομής της τυποποιημένης κανονικής κατανομής.

Σημείωση: Μια εναλλακτική παρουσίαση της ερμηνείας του κεντρικού οριακού θεωρήματος είναι ότι, για αρκετά μεγάλες τιμές του n, το άθροισμα S_n n ανεξαρτήτων τυχαίων μεταβλητών $X_1, X_2, \ldots, X_n$ με μέση τιμή μ και διασπορά σ^2 έχει περίπου την κανονική κατανομή με μέση τιμή $\mu_{S_n} = n\mu$ και διασπορά $\sigma^2_{S_n} = n\sigma^2$. Δηλαδή, $P(S_n \le s) \cong \Phi(\frac{s - n\mu}{\sigma\sqrt{n}})$. Η προσέγγιση αφορά μόνο την κατανομή του S_n, επειδή $\mu_{S_n} = n\mu$ και $\sigma^2_{S_n} = n\sigma^2$, ακριβώς.

Παράδειγμα: Η εφαρμογή του κεντρικού οριακού θεωρήματος στο άθροισμα των τιμών n τυχαίων μεταβλητών βοηθά στην ερμηνεία του γεγονότος ότι τόσα πολλά ποσοτικά φαινόμενα έχουν (τουλάχιστον κατά προσέγγιση) κανονική κατανομή. Ας υποθέσουμε, για παράδειγμα, ότι ενδιαφερόμαστε για το ύψος Y ενός ατόμου. Μπορούμε να θεωρήσουμε ότι το ύψος αυτό είναι το αποτέλεσμα άθροισης $Y = \sum_{i=1}^{n} X_i$ ενός μεγάλου αριθμού μικρών (απειροστών) αυξήσεων X_i, όπου κάθε μικρή αύξηση λαβαίνει χώρα μέσα σε κάποιο μικρό χρονικό διάστημα. Εύλογο είναι να υποθέσουμε ότι οι αυξήσεις X_i που αντιστοιχούν σε διαφορετικά μικρά χρονικά

διαστήματα είναι αμοιβαία ανεξάρτητες και έχουν μια κοινή κατανομή (είναι ισόνομες). Στην περίπτωση αυτή, η εφαρμογή του κεντρικού οριακού θεωρήματος μας επιτρέπει να συμπεράνουμε ότι το ύψος $Y = S_n = \sum_{i=1}^{n} X_i$ έχει, τουλάχιστον κατά προσέγγιση, μια κανονική κατανομή.

Σημείωση: Είναι προφανές ότι η ακρίβεια της προσέγγισης της κατανομής της τυχαίας μεταβλητής Z_n από την κανονική κατανομή με την βοήθεια του κεντρικού οριακού θεωρήματος μεταβάλλεται με την τιμή του n. Ας υποθέσουμε, χωρίς βλάβη της γενικότητας, ότι οι μεταβλητές $X_1,X_2,\ldots,X_n$ είναι ισόνομες. Επειδή η κανονική κατανομή είναι συμμετρική γύρω από την μέση τιμή της, είναι φανερό ότι όσο περισσότερο ασύμμετρη (στρεβλή) είναι η κοινή κατανομή των τυχαίων μεταβλητών $X_1,X_2,\ldots,X_n$, τόσο μεγαλύτερο μέγεθος δείγματος απαιτείται για να επιτευχθεί μια καλή προσέγγιση της συνάρτησης κατανομής $F_{Z_n}(z)$ από την συνάρτηση κατανομής της τυποποιημένης κανονικής κατανομής.

Παράδειγμα *(τυχαίος περίπατος):* Έστω $\{Y_n\}$ μια ακολουθία ανεξάρτητων και ισόνομων τυχαίων μεταβλητών με

$$E(Y_n) = 0 \text{ και } V(Y_n) = \sigma^2 < \infty \text{ για κάθε } n \geq 1.$$

Έστω

$$X_n = \sum_{i=1}^{n} Y_i \text{ , } n \geq 1.$$

(Η ακολουθία των τυχαίων μεταβλητών $\{X_n\}$ αποτελεί ένα *τυχαίο περίπατο (random walk).* Η ονομασία οφείλεται στο γεγονός ότι μια τέτοια ακολουθία μπορεί να χρησιμοποιηθεί για την περιγραφή του εξής συστήματος συμπεριφοράς. Σε διακριτό χρόνο t = i, ένα σωματίδιο μετακινείται κατά μία απόσταση Y_i κατά μήκος μιας ευθείας γραμμής. Η καθαρή απόσταση κατά την οποία το σωματίδιο μετακινήθηκε μετά την n κίνησή του, είναι X_n. Οι ατομικές κινήσεις είναι ανεξάρτητες και έχουν την ίδια κατανομή, ενώ η μέση κίνηση είναι μηδέν).

Να αποδειχθεί ότι, καθώς ο χρόνος αυξάνει, η πιθανότητα το σωματίδιο να βρίσκεται σε απόσταση c από το σημείο εκκίνησής του, τείνει στο μηδέν, ανεξάρτητα από το πόσο μεγάλη είναι η τιμή της σταθεράς c.

Απόδειξη: Σύμφωνα με το κεντρικό οριακό θεώρημα, για μεγάλες τιμές του n, η μεταβλητή

$$X_n^* = \frac{X_n}{\sigma\sqrt{n}}$$

κατανέμεται, κατά προσέγγιση, σύμφωνα με την τυποποιημένη κανονική κατανομή. Επομένως,

$$P\left(|X_n| \le c\right) = P\left(|X_n^*| \le \frac{c}{\sigma\sqrt{n}}\right)$$

$$\cong 2\Phi\left(\frac{c}{\sigma\sqrt{n}}\right) - 1$$

Καθώς $n \to \infty$, έχουμε ότι

$$\frac{c}{\sigma\sqrt{n}} \to 0 \quad \text{και} \quad \Phi\left(\frac{c}{\sigma\sqrt{n}}\right) \to 0.5$$

Κατά συνέπεια,

$$P\left(|X_n| \le c\right) \to 0 \quad \text{καθώς} \quad n \to \infty$$

Σημείωση: Ο αναγνώστης που ενδιαφέρεται για περισσότερες λεπτομέρειες για τους τυχαίους περιπάτους παραπέμπεται σε εγχειρίδιο Στοχαστικών Ανελίξεων (βλέπε π.χ. Ε. Ξεκαλάκη & Ι. Πανάρετος: *Πιθανότητες και Στοιχεία Στοχαστικών Ανελίξεων*).

Η Χρήση του Κεντρικού Οριακού Θεωρήματος στην Στατιστική

Η σημασία του κεντρικού οριακού θεωρήματος στην Στατιστική και τις εφαρμογές της είναι δύσκολο να εκτιμηθεί και οφείλεται στο ότι επιτρέπει την χρήση της κανονικής κατανομής σε πολλά πρακτικά προβλήματα.

Στο πλαίσιο της Στατιστικής Συμπερασματολογίας και ορολογίας και με υποθέσεις που ικανοποιούνται στις περισσότερες

εφαρμογές, το κεντρικό οριακό θεώρημα μπορεί να διατυπωθεί ως εξής:

Στατιστική διατύπωση του κεντρικού οριακού θεωρήματος: Έστω X_1, X_2, ..., X_n n αμοιβαία ανεξάρτητες παρατηρήσεις πάνω σε μια τυχαία μεταβλητή X που περιγράφει τον υπό μελέτη πληθυσμό, με πεπερασμένη μέση τιμή μ και πεπερασμένη διασπορά σ^2. Έστω $\overline{X}$ ο μέσος των παρατηρήσεων αυτών, δηλαδή

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

(ο μέσος αυτός αναφέρεται συχνά ως δειγματικός μέσος).
Έστω ότι η μεταβλητή

$$Z_n = \frac{\overline{X} - \mu}{\sigma\sqrt{n}}$$

είναι η αντίστοιχη τυποποιημένη μεταβλητή και έστω ότι $F_{Z_n}(z)$ είναι η συνάρτηση κατανομής της τυχαίας μεταβλητής Z_n. Τότε, για κάθε z στο διάστημα $(-\infty,+\infty)$, ισχύει ότι

$$\lim_{n\to\infty} F_{Z_n}(z) = \Phi(z)$$

όπου $\Phi(z)$ είναι η συνάρτηση κατανομής της τυποποιημένης κανονικής κατανομής.

Δηλαδή, για αρκετά μεγάλες τιμές του n, ο δειγματικός μέσος n ανεξαρτήτων παρατηρήσεων X_1, X_2, ..., X_n πάνω σε μια τυχαία μεταβλητή X οποιασδήποτε κατανομής με μέση τιμή μ και διασπορά σ^2 ακολουθεί, κατά προσέγγιση, την κανονική κατανομή με μέση τιμή μ=μ και διασπορά $\sigma^2 = \sigma^2$.

Το σημαντικότερο στοιχείο του θεωρήματος είναι ότι το συμπέρασμά του δεν εξαρτάται από την μορφή του πληθυσμού τον οποίο περιγράφει η τυχαία μεταβλητή X (δηλαδή, δεν εξαρτάται από την κατανομή του πληθυσμού αυτού). Η ονομασία του θεωρήματος οφείλεται στο ότι το συμπέρασμά του είναι καθοριστικής σημασίας στην στατιστική θεωρία.

Πόρισμα: Ένα άμεσο συμπέρασμα που μπορεί να εξαχθεί από το κεντρικό οριακό θεώρημα είναι ότι, για οποιοδήποτε δοθέν διάστημα [α,β], ισχύει ότι

$$\lim_{n\to\infty} P(\alpha \le Z_n \le \beta) = \lim_{n\to\infty}\left[F_{Z_n}(\beta) - F_{Z_n}(\alpha)\right]$$

$$= \Phi(\beta) - \Phi(\alpha)$$

Η ΔΙΟΡΘΩΣΗ ΣΥΝΕΧΕΙΑΣ *(Continuity Correction)*

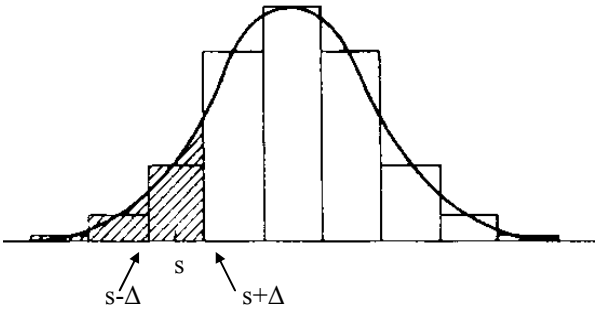
Ας υποθέσουμε ότι οι τιμές των μεταβλητών $S_n = \sum_{i=1}^{n} X_i$ και $\overline{X} = S_n/n$ υπολογίζονται με βάση τις τιμές n τυχαίων μεταβλητών $X_1, X_2, \ldots, X_n$, οι οποίες έχουν διακριτές κατανομές των οποίων οι δυνατές τιμές είναι ισαπέχουσες σε διαστήματα μήκους 2Δ. Για μέτριες τιμές του n μια διόρθωση συνεχείας συχνά βοηθά στην επανόρθωση του σφάλματος το οποίο γίνεται όταν προσεγγίζεται μια διακριτή κατανομή από μια συνεχή κανονική κατανομή. Αν s και x είναι δυνατές τιμές της τυχαίας μεταβλητής S_n και $\overline{X}$, τότε χρησιμοποιούμε τις προσεγγίσεις

$$P(S_n \le s) \cong \Phi\left(\frac{s + \Delta - n\mu}{\sigma\sqrt{n}}\right)$$

$$P(\overline{X} \le x) \cong \Phi\left(\frac{x + \Delta/n - \mu}{\sigma\sqrt{n}}\right)$$

αντίστοιχα. Ας σημειωθεί ότι αν οι διαδοχικές δυνατές τιμές της τυχαίας μεταβλητής X_i (i=1,2,..,n), απέχουν 2Δ, τότε και οι διαδοχικές τιμές της μεταβλητής S_n απέχουν την ίδια απόσταση. Για να αιτιολογηθεί η παραπάνω προσέγγιση για την πιθανότητα που αναφέρεται στη μεταβλητή S_n, θεωρούμε ότι η μάζα πιθανότητας $P(S_n=s)$ που αντιστοιχεί στις δυνατές τιμές s της μεταβλητής S_n εκτείνεται ομοιόμορφα στο διάστημα $[s-\Delta,\ s+\Delta]$, "μετατρέποντας" έτσι την κατανομή της τυχαίας μεταβλητής S_n σε μια συνεχή κατανομή της οποίας η συνάρτηση πυκνότητας πιθανότητας έχει την μορφή ενός ιστογράμματος. Όπως φαίνεται στο σχήμα που ακολουθεί

η πιθανότητα $P(S_n = s)$ είναι η περιοχή κάτω από το ιστόγραμμα που αντιστοιχεί στο διάστημα $[s-\Delta,\ s+\Delta]$.



Έτσι, για να υπολογισθεί η πιθανότητα $P(S_n \le s)$, θα πρέπει να υπολογισθεί το εμβαδόν κάτω από το ιστόγραμμα και υπεράνω του διαστήματος $(-\infty,\ s+\Delta]$. Περιοχές κάτω από την συνάρτηση πυκνότητας πιθανότητας της κανονικής κατανομής με μέση τιμή nμ και διασπορά $n\sigma^2$ χρησιμοποιούνται για να προσεγγισθούν περιοχές κάτω από το ιστόγραμμα, οδηγώντας, έτσι, στην προσέγγιση που παρέχεται από την παραπάνω διόρθωση συνεχείας. Παρόμοιας μορφής επιχειρηματολογία αιτιολογεί την διόρθωση συνεχείας για την πιθανότητα που αναφέρεται στον δειγματικό μέσο $\overline{X}$.

Παράδειγμα: Ένας παίκτης που παίζει Black Jack αποφασίζει να παίξει 100 φορές και να στοιχηματίζει 10000 δρχ. κάθε φορά. Υποθέτουμε ότι οι μοιρασιές (hands) παίζονται ανεξάρτητα και ότι σε κάθε παιχνίδι η πιθανότητα του παίκτη να κερδίσει 10000 δρχ. είναι 0.49. Έστω ότι X_i, i=1,2,..,100, παριστάνει τα κέρδη του παίκτη στο τέλος του i παιχνιδιού.
Τότε,

$$\begin{aligned}\mu_{X_i} &= 10(0.49) + (-10)(0.51) = -0.20,\\ \sigma_{X_i} &= \sqrt{(10+0.20)^2(0.49) + (-10+0.20)^2(0.51)}\\ &= \sqrt{99.96}\\ &= 10\end{aligned}$$

Δηλαδή, το μέσο κέρδος του παίκτη είναι –200 δρχ. (ζημία) και η τυπική απόκλιση είναι 10000 δρχ. Τα συνολικά κέρδη του παίκτη είναι

$$S_{100} = \sum_{i=1}^{100} X_i$$

Ας σημειωθεί ότι η τυχαία μεταβλητή X_i είναι διακριτή της οποίας οι διαδοχικές τιμές απέχουν 20 χιλιάδες δραχμές. Έτσι, για την διόρθωση συνεχείας θέτουμε Δ=10. Η πιθανότητα $P(S_{100} > 0)$, ότι ο παίκτης θα τελειώσει το παιχνίδι με κάποιο κέρδος, είναι $1- P(S_{100}>0)$, η οποία από το κεντρικό οριακό θεώρημα (χρησιμοποιώντας την διόρθωση συνεχείας) είναι περίπου ίση με

$$1 - \Phi\left[\frac{S_{100} + \Delta - n\mu}{\sigma\sqrt{n}}\right] = 1 - \Phi\left[\frac{0 + 10 - 100(-20)}{10\sqrt{100}}\right]$$

$$= 1- \Phi(0.30) = 0.3821$$

Θα ήταν ίσως ενδιαφέρον να υπολογίσουμε την πιθανότητα ότι ο παίκτης θα επιτύχει κάποιο κέρδος εάν παίξει 1.000 φορές αντί 100. Στην περίπτωση αυτή το συνολικό κέρδος S_{1000} του παίκτη έχει πιθανότητα περίπου ίση με $1 - \Phi(0.66) = 0.2546$ να είναι μεγαλύτερο από το μηδέν. Βλέπουμε δηλαδή ότι η πιθανότητα να τελειώσει ο παίκτης το παιχνίδι με κάποιο κέρδος ελαττώνεται όταν ο αριθμός των παιχνιδιών αυξάνει.

Σημείωση: Η κανονική κατανομή, πέραν της χρησιμότητάς της για την περιγραφή φυσικών φαινομένων, έχει μεγάλο πεδίο εφαρμογών ως προσεγγιστική άλλων μορφών κατανομών που αναφύονται ως ακριβείς κατανομές στην μελέτη διαφόρων προβλημάτων ή φαινομένων. Αυτό επιτυγχάνεται με την χρήση του κεντρικού οριακού θεωρήματος και αναπτύσσεται στην συνέχεια.

Η ΠΡΟΣΕΓΓΙΣΤΙΚΗ ΧΡΗΣΗ ΤΗΣ ΚΑΝΟΝΙΚΗΣ ΚΑΤΑΝΟΜΗΣ

Προσέγγιση της Διωνυμικής από την Κανονική Κατανομή

Μια από τις σημαντικές ειδικές χρήσεις του κεντρικού οριακού θεωρήματος είναι η προσέγγιση των πιθανοτήτων μιας διωνυμικής

κατανομής. Αν S_n ακολουθεί την διωνυμική κατανομή με παραμέτρους n και p, τότε, όπως έχει ήδη παρατηρηθεί,

$$S_n = \sum_{i=1}^{n} X_i$$

όπου X_i, i=1,2,…,n είναι ανεξάρτητες μεταβλητές Bernoulli, κάθε μια με παράμετρο p. Για τιμές του n που είναι μέτρια μεγάλες (π.χ. τέτοιες ώστε $(np \geq 5$ και $n(1-p) \geq 5)$, η συνάρτηση κατανομής της τυποποιημένης κανονικής κατανομής δίνει μια καλή προσέγγιση για την συνάρτηση κατανομής της τυχαίας μεταβλητής $(S_n - np)/\sqrt{np(1-p)}$. Η προσέγγιση αυτή συχνά βελτιώνεται με μια διόρθωση συνεχείας όπου Δ=1/2 επειδή κάθε μια από τις μεταβλητές X_i είναι διακριτή με τιμές που αυξάνουν με βήμα 1. Έτσι, όταν η μεταβλητή S_n έχει την διωνυμική κατανομή με παραμέτρους n και p, ισχύει ότι

$$P(S_n \leq s) \cong \Phi\left[\frac{s + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right]$$

Παράδειγμα: Μια γενετική θεωρία υποστηρίζει ότι 25% των απογόνων ενός πειράματος διασταύρωσης ειδών θα έχουν το χαρακτηριστικό Α. Σε ένα σύνολο 50 απογόνων που προήλθαν από ένα τέτοιο πείραμα, μόνο 6 βρέθηκαν να έχουν το χαρακτηριστικό Α. Αν η γενετική θεωρία είναι αληθής, η πιθανότητα να παρατηρηθούν το πολύ 6 απόγονοι με το χαρακτηριστικό Α είναι

$$P(S_{25} \leq 6) \cong \Phi\left[\frac{6 + \frac{1}{2} - (50)(0.25)}{\sqrt{(50)(0.25)\ (0.75)}}\right]$$

$$= \Phi\ (-1.96) = 1\text{-}\Phi\ (1.96)$$

$$= 0.025$$

Επομένως, παρατηρήθηκε ένα αποτέλεσμα που ανήκει σε ένα ενδεχόμενο το οποίο έχει πολύ μικρή πιθανότητα να παρατηρηθεί, αν η γενετική θεωρία είναι σωστή.

Η Κανονική Προσέγγιση για τα Ιστογράμματα Πιθανοτήτων

Για την καλύτερη κατανόηση της σημασίας της προσέγγισης των πιθανοτήτων της διωνυμικής από την κανονική κατανομή, είναι χρήσιμη η μελέτη ορισμένων πρακτικών εφαρμογών.

Σύμφωνα με τον νόμο των μεγάλων αριθμών, αν στρίψουμε ένα νόμισμα ένα μεγάλο αριθμό φορών, το ποσοστό του αποτελέσματος Κ (κεφάλι) θα είναι κοντά στο 50%. Η μαθηματική προσέγγιση του προβλήματος αυτού έγινε γύρω στο 1700 από τον Ελβετό Μαθηματικό James Bernoulli. Είκοσι χρόνια αργότερα, ο De Moivre βελτίωσε σημαντικά την εργασία του Bernoulli αποδεικνύοντας πώς μπορεί να υπολογιστεί η πιθανότητα ότι το ποσοστό του αποτελέσματος "γράμματα" θα βρίσκεται σε οποιοδήποτε διάστημα γύρω από το 50%. Η μέθοδος είναι προσεγγιστική αλλά η προσέγγιση βελτιώνεται όσο αυξάνεται ο αριθμός της επανάληψης του πειράματος.

Τόσο ο Bernoulli όσο και ο De Moivre έκαναν ορισμένες υποθέσεις για το νόμισμα: Τα πειράματα (το στρίψιμο του νομίσματος) είναι ανεξάρτητα και σε κάθε στρίψιμο τα αποτελέσματα Κ και Γ είναι ισοπίθανα. Από τις υποθέσεις αυτές, προκύπτει ότι οποιαδήποτε ακολουθία αποτελεσμάτων έχει την ίδα πιθανότητα με οποιαδήποτε άλλη ακολουθία σε ίσο αριθμό δοκιμών. Αυτό που έκανε ο Bernoulli ήταν να αποδείξει ότι, για τις περισσότερες

ακολουθίες το 50% των αποτελεσμάτων κατά μέσο όρο είναι "γράμματα".

Για να επιβεβαιώσουμε την αλήθεια του ισχυρισμού αυτού, ας ξεκινήσουμε με 5 ρίψεις ενός νομίσματος. Ας υποθέσουμε ότι καταγράφουμε τα αποτελέσματα κάθε φορά. Μπορούμε εύκολα να διαπιστώσουμε ότι υπάρχει μια μόνο δυνατή ακολουθία όπου στις 5 ρίψεις τα αποτελέσματα είναι 5 φορές Γ (η ακολουθία ΓΓΓΓΓ).
Εάν μας ενδιαφέρει ο αριθμός των ακολουθιών όπου το αποτέλεσμα είναι 4 φορές Γ και 1 φορά Κ η απάντηση είναι 5, δηλαδή οι εξής:

ΓΚΚΚΚ ΚΓΚΚΚ ΚΚΓΚΚ ΚΚΚΓΚ ΚΚΚΚΓ

Η ακολουθία ΓΚΚΚΚ, για παράδειγμα, υποδηλώνει ότι το αποτέλεσμα ήταν Γ στην πρώτη δοκιμή και στην συνέχεια σε όλες τις υπόλοιπες 4 φορές ήταν Κ.

Στον πίνακα που ακολουθεί, βλέπουμε τον αριθμό των ακολουθιών που υπάρχουν για οποιοδήποτε αριθμό αποτελεσμάτων Γ στις 5 ρίψεις. Στις 5 ρίψεις υπάρχουν συνολικά 2^5=32 δυνατές ακολουθίες που μπορούν να προκύψουν. Από τις 32 αυτές ακολουθίες στις 20, έχουμε σχεδόν ½ των αποτελεσμάτων να είναι Γ (2 ή 3 από τις 5).

Πίνακας: Αριθμός των ακολουθιών που αντιστοιχούν στον αριθμό των δοκιμών στις οποίες το αποτέλεσμα είναι Γ σε 5 ρίψεις ενός νομίσματος

Αριθμός εμφάνισης του αποτελέσματος "Γράμματα" (Γ)	**Αριθμός Ακολουθιών**
Μηδέν	1
Ένα	5
Δύο	10
Τρία	10
Τέσσερα	5
Πέντε	1

Ο De Moivre κατάφερε να μετρήσει με μικρή απόκλιση λάθους τον αριθμό των ακολουθιών που καταλήγουν σε ένα δεδομένο αριθμό "γραμμάτων" για οποιοδήποτε αριθμό επανάληψης του πειράματος. Με 100 επαναλήψεις, ο αριθμός των δυνατών ακολουθιών είναι 2^{100}.

Είναι σχεδόν αδύνατο να καταγράψει κανείς όλες αυτές τις ακολουθίες.

Παρ' όλα αυτά, χρησιμοποιώντας τον τύπο των συνδυασμών, έχουμε ότι ο αριθμός των ακολουθιών στις οποίες το αποτέλεσμα είναι ακριβώς 50 φορές "γράμματα" είναι

$$\binom{100}{50} = \frac{100!\times 50!}{50!} = \frac{100\times 99\times \ldots \times 51}{50\times 49\times \ldots \times 1} \cong 1.01\times 10^{29}$$

Δεδομένου ότι ο συνολικός αριθμός ακολουθιών που μπορούμε να παρατηρήσουμε είναι $2^{100}\cong 1{,}27\times 10^{30}$, έχουμε ότι η πιθανότητα ακριβώς 50 "γραμμάτων" σε 100 ρίψεις ενός νομίσματος είναι:

$$\frac{\text{Αριθμός ακολουθιών με 50 "γράμματα"}}{\text{Συνολικός αριθμός ακολουθιών}} \cong \frac{1.01\times 10^{29}}{1.27\times 10^{38}}$$

$$\cong 0.08 = 8\%$$

Βέβαια, την εποχή του De Moivre δεν υπήρχαν μηχανές αριθμητικών υπολογισμών. Χρειαζόταν, επομένως, να υπολογισθεί ο διωνυμικός συντελεστής με αριθμητικούς υπολογισμούς. (Από πολλούς, η προσέγγιση αυτή αποδίδεται σε κάποιο άλλο Μαθηματικό τον James Stirling).

Η μέθοδος που χρησιμοποίησε ο De Moivre τον οδήγησε στην χρήση της κανονικής κατανομής. Για παράδειγμα, βρήκε ότι η πιθανότητα να έχει ακριβώς 50 φορές "γράμματα" σε 100 ρίψεις ενός νομίσματος ήταν περίπου ίση με το εμβαδόν κάτω από την κανονική καμπύλη μεταξύ –0,1 και +0,1. Μάλιστα, κατάφερε να αποδείξει ότι όλο το ιστόγραμμα πιθανότητας για τον αριθμό του αποτελέσματος "γράμματα" προσεγγίζεται πολύ καλά από την κανονική καμπύλη, όταν ο αριθμός ρίψεων του νομίσματος είναι μεγάλος. Αργότερα, οι ερευνητές επεξέτειναν το αποτέλεσμα του De Moivre για το άθροισμα επιλογών στην τύχη από μία κληρωτίδα.

Η μαθηματική προσέγγιση που χρησιμοποίησε ο De Moivre είναι αρκετά πολύπλοκη. Παρ' όλα αυτά, στην συνέχεια παρουσιάζεται με γραφικές μεθόδους και με την χρήση γραφημάτων μέσω υπολογιστών η μεθοδολογία αυτή.

Ιστογράμματα Πιθανότητας *(Probability Histograms)*

Όταν ένας τυχαίος μηχανισμός (chance process) παράγει ένα αριθμό, η μέση τιμή και η τυπική απόκλιση είναι ένας οδηγός για την πιθανή τιμή του αριθμού αυτού. Εκείνο όμως που δίνει μια πλήρη εικόνα για τον αριθμό είναι το *ιστόγραμμα πιθανότητας* (probability histogram). (Για ιστογράμματα και τις εφαρμογές τους βλέπε π.χ. το βιβλίο των συγγραφέων *Εισαγωγή στην Στατιστική Σκέψη, τόμος Ι, (Περιγραφική Στατιστική)*).

Το *ιστόγραμμα πιθανότητας* απεικονίζει πιθανότητα (τύχη) και όχι δεδομένα.

Παράδειγμα: Είναι γνωστό το τυχερό παιχνίδι που αναφέρεται στο άθροισμα του αποτελέσματος μιας ζαριάς από δύο ζάρια. Ως γνωστόν, το άθροισμα αυτό παίρνει τιμές από 2 έως 12. Επομένως, η πιθανότητα να κερδίσει κανείς ένα στοίχημα (odds) εξαρτάται από την πιθανότητα που έχει κάθε ένα από τα αθροίσματα αυτά. Ένας τρόπος για να προσδιορισθούν οι πιθανότητες αυτές είναι να το δοκιμάσει κανείς πειραματικά ή να χρησιμοποιήσει ένα υπολογιστή για την προσομοίωση του πειράματος αυτού. Αποτελέσματα μιας τέτοιας προσομοίωσης σε υπολογιστή για 100 ζαριές με δύο ζάρια δίνονται στον πίνακα που ακολουθεί.

Πίνακας: Πείραμα με ζαριά από δύο ζάρια. Πρόκειται για προσομοίωση με υπολογιστή όπου γίνεται μία διαδικασία ανάλογη με το πείραμα και καταγράφεται το άθροισμα του αποτελέσματος σε κάθε ζαριά. Ο υπολογιστής έχει επαναλάβει το πείραμα 10000 φορές. Τα αποτελέσματα για τις πρώτες 100 επαναλήψεις εμφανίζονται στον πίνακα που ακολουθεί.

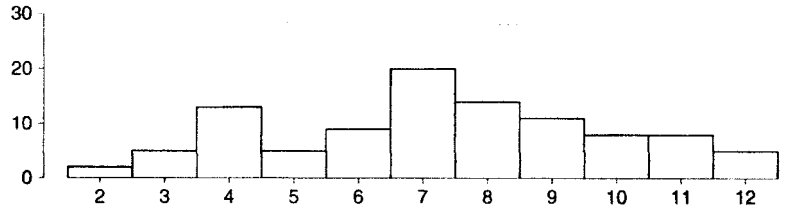
Επανά-ληψη	Αθροι-σμα	Επανά-ληψη	Αθροι-σμα	Επανά-ληψη	Αθροι-σμα	Επανά-ληψη	Αθροι-σμα	Επανά-ληψη	Αθροι-σμα
1	8	21	10	41	8	61	8	81	11
2	9	22	4	42	10	62	5	82	9
3	7	23	8	43	6	63	3	83	7
4	10	24	7	44	3	64	11	84	4
5	9	25	7	45	4	65	9	85	7
6	5	26	3	46	8	66	4	86	4
7	5	27	8	47	4	67	12	87	7
8	4	28	8	48	4	68	7	88	6
9	4	29	12	49	5	69	10	89	7
10	4	30	2	50	4	70	4	90	11
11	10	31	11	51	11	71	7	91	6
12	8	32	12	52	8	72	4	92	11
13	3	33	12	53	10	73	7	93	8
14	11	34	7	54	9	74	9	94	8
15	7	35	7	55	10	75	9	95	7
16	8	36	6	56	12	76	11	96	9
17	9	37	6	57	7	77	6	97	10
18	8	38	2	58	6	78	9	98	5
19	6	39	6	59	7	79	9	99	7
20	8	40	3	60	7	80	7	100	7

Τα τρία πρώτα από τα γραφήματα που ακολουθούν αναφέρονται στο εμπειρικό ιστόγραμμα για 100, 1000 και 10000 επαναλήψεις του πειράματος.

Το πρώτο γράφημα δείχνει ότι σε 100 επαναλήψεις είχαμε 20 φορές άθροισμα 7 και, επομένως, το ορθογώνιο στην θέση 7 έχει εμβαδόν ίσο με το 20% του συνολικού εμβαδού.

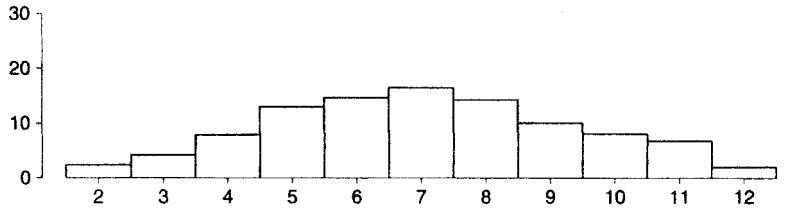
Το τελευταίο γράφημα αναφέρεται στο ιστόγραμμα πιθανότητας. Το ιστόγραμμα αυτό αποτελεί ένα ιδανικό ιστόγραμμα στο οποίο συγκλίνει το εμπειρικό ιστόγραμμα, όταν το πείραμα συνεχίσει να επαναλαμβάνεται επ' άπειρον.

100 Επαναλήψεις



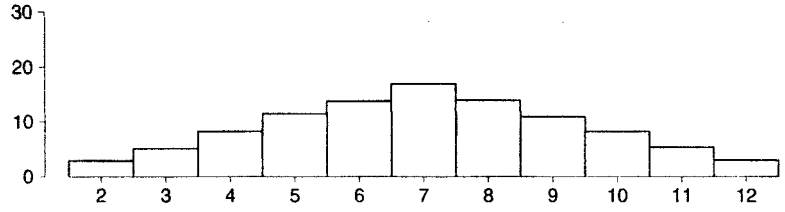
Άθροισμα ζαριάς με δύο ζάρια

1000 Επαναλήψεις



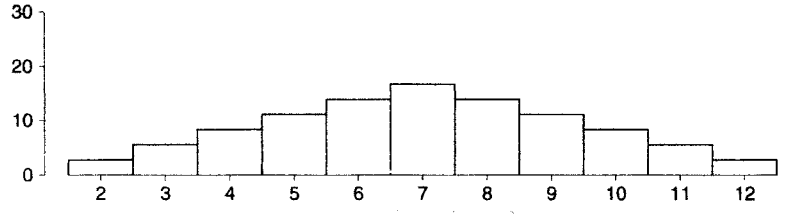
Άθροισμα ζαριάς με δύο ζάρια

10000 Επαναλήψεις



Άθροισμα ζαριάς με δύο ζάρια

Ιστόγραμμα Πιθανότητας



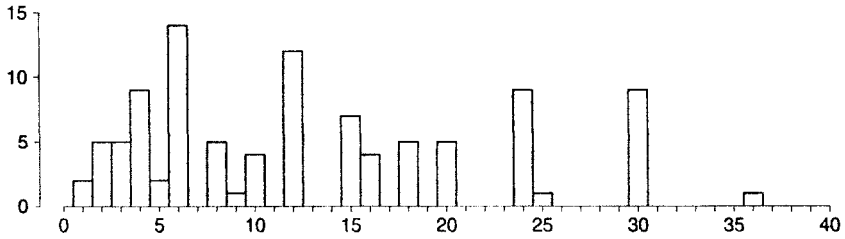
Άθροισμα ζαριάς με δύο ζάρια

Σημείωση: Είναι φανερό ότι το ιστόγραμμα πιθανότητας θα μπορούσε να κατασκευασθεί με θεωρητική προσέγγιση. Όπως είναι γνωστό, οι δυνατοί τρόποι που το άθροισμα σε μια ζαριά από δύο ζάρια είναι 7 είναι 6/36. Δηλαδή, 16 2/3 %. Επομένως, το εμβαδόν του ορθογωνίου πάνω από το σημείο 7 στο ιστόγραμμα πιθανότητας θα είναι 16 2/3 %.

Το ιστόγραμμα πιθανότητας, δηλαδή, εκφράζει την πιθανότητα μέσω εμβαδού και διαμορφώνεται με ορθογώνια όπως φαίνεται στο σχήμα που προηγήθηκε. Η βάση κάθε ορθογωνίου έχει μέσο κάθε μια από τις δυνατές τιμές του αθροίσματος, ενώ το εμβαδόν του ορθογωνίου είναι ίσο με την πιθανότητα να παρατηρηθεί η τιμή αυτή. Το συνολικό εμβαδόν του ιστογράμματος είναι 100%.

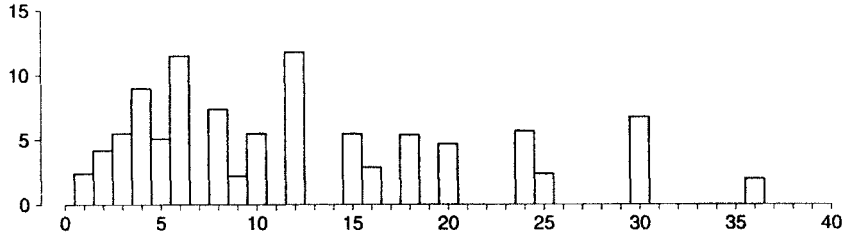
Παράδειγμα: Ας εξετάσουμε τώρα το γινόμενο των αποτελεσμάτων σε μια ζαριά από δύο ζάρια αντί για το άθροισμα. Το πρόγραμμα προσομοίωσης του υπολογιστή είναι να επαναλάβει το τυχαίο πείραμα το οποίο αναφέρεται στο ρίξιμο ενός ζευγαριού από ζάρια και στην συνέχεια να υπολογίσει το γινόμενο των αριθμών που προκύπτουν. Στο σχήμα που ακολουθεί, τα τρία πρώτα γραφήματα αναφέρονται σε 100, 1000 και 10000 επαναλήψεις του πειράματος. Από το πρώτο γράφημα, παρατηρούμε ότι το αποτέλεσμα "γινόμενο 10" εμφανίστηκε 4 φορές, επομένως το εμβαδόν του ορθογωνίου στο σημείο 10 είναι ίσο με 4%. Οι άλλες τιμές προκύπτουν με τον ίδιο τρόπο. Το τελευταίο γράφημα του σχήματος δείχνει το ιστόγραμμα πιθανότητας. Παρατηρούμε ότι το εμπειρικό ιστόγραμμα για 10000 επαναλήψεις είναι σχεδόν το ίδιο με το ιστόγραμμα πιθανότητας.

100 Επαναλήψεις



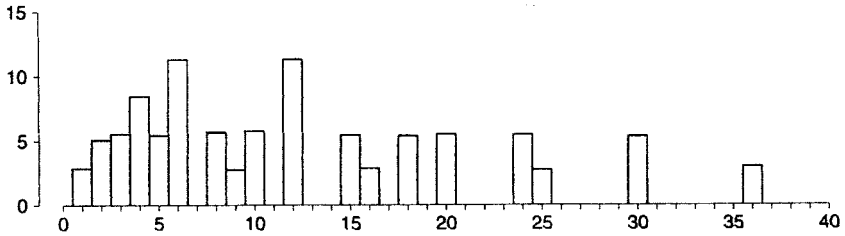
Τιμή του γινομένου

1000 Επαναλήψεις



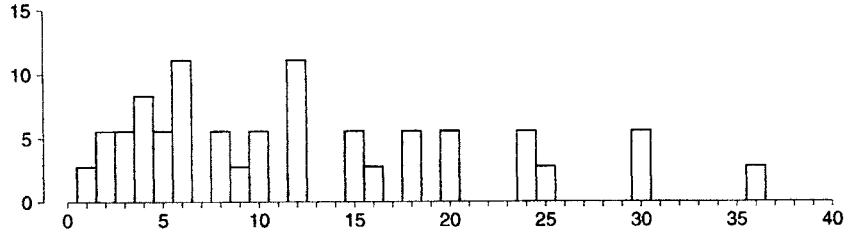
Τιμή του γινομένου

10000 Επαναλήψεις



Τιμή του γινομένου

Ιστόγραμμα Πιθανότητας

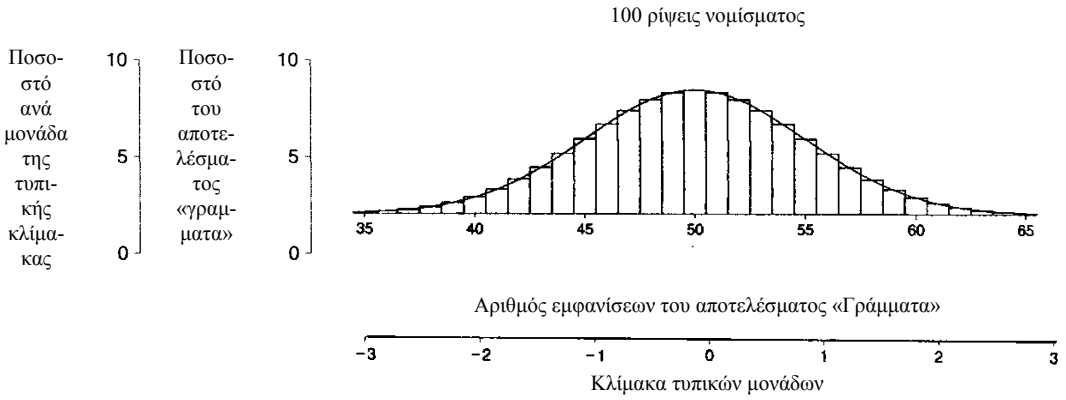


Τιμή του γινομένου

Σημείωση: Το σχήμα 2 διαφέρει από το σχήμα 1 κατά το ότι έχει κενά. Αυτό οφείλεται στις δυνατές τιμές που μπορεί να πάρει το γινόμενο. Προφανώς, η μικρότερη τιμή είναι 1, όταν και τα δύο ζάρια έχουν αποτέλεσμα "1", ενώ η μεγαλύτερη είναι 36, όταν και τα δύο ζάρια είναι "6". Το αποτέλεσμα "7" όμως δεν είναι δυνατόν να παρατηρηθεί. Γι' αυτό τον λόγο, δεν εμφανίζεται ορθογώνιο στην θέση 7 (έχει εμβαδόν 0). Το ίδιο συμβαίνει με την τιμή 11 και τις άλλες τιμές που δεν εμφανίζονται.

Ιστόγραμμα Πιθανότητας και η Κανονική Κατανομή

Όπως μπορεί να παρατηρήσει κανείς, το ιστόγραμμα πιθανότητας για τον αριθμό των αποτελεσμάτων "γράμματα" στην επανάληψη στριψίματος ενός νομίσματος, πλησιάζει την καμπύλη της κανονικής κατανομής, όταν ο αριθμός των δοκιμών γίνεται μεγάλος. Για παράδειγμα, όπως βλέπουμε στο σχήμα που ακολουθεί, σε 100 ρίψεις ενός νομίσματος το ιστόγραμμα πιθανότητας του αριθμού του αποτελεσμάτος "γράμματα" περιγράφεται αρκετά καλά από την κανονική κατανομή.



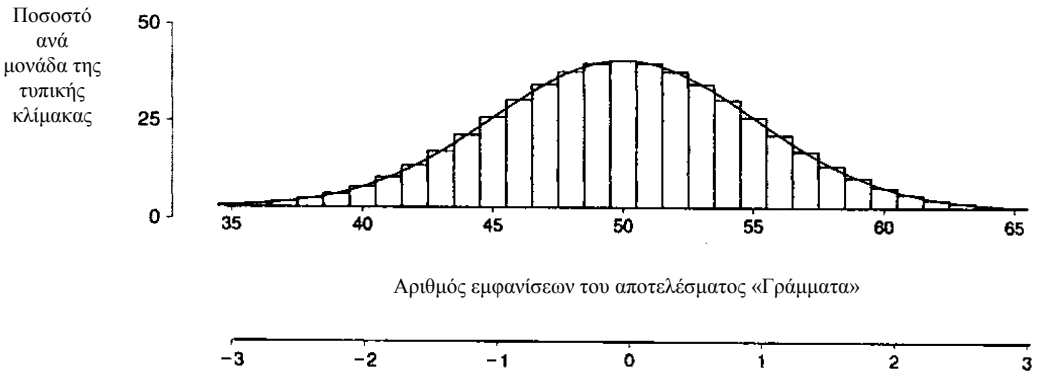
Το σχήμα έχει δύο οριζόντιους άξονες. Το ιστόγραμμα πιθανότητας έχει κατασκευασθεί σε σχέση με τον πρώτο από τους δύο άξονες που δείχνει τον αριθμό του αποτελέσματος "γράμματα". Η κανονική καμπύλη όμως έχει σχεδιαστεί σε σχέση με τον κατώτερο άξονα που είναι εκφρασμένος σε μονάδες της τυπικής κλίμακας (τυποποιημένης κλίμακας). Η μέση τιμή του αποτελέσματος

“γράμματα” είναι 50 και η τυπική απόκλιση είναι 5. Επομένως, ο αριθμός 50 που αναφέρεται στο αποτέλεσμα “γράμματα” σε 100 δοκιμές αντιστοιχεί στο 0 του τυποποιημένου άξονα, το 55 αντιστοιχεί στο +1 κ.ο.κ..

Στο σχήμα, υπάρχουν επίσης δύο κατακόρυφοι άξονες. Το ιστόγραμμα πιθανότητας έχει σχεδιασθεί σε σχέση με τον εσωτερικό άξονα που δείχνει το ποσοστό σε σχέση με το ενδεχόμενο “γράμματα”. Η κανονική καμπύλη έχει σχεδιασθεί σε σχέση με τον εξωτερικό άξονα, ο οποίος αναφέρεται σε ποσοστό σε σχέση με τις μονάδες στην τυποποιημένη κλίμακα. Για παράδειγμα, ας θεωρήσουμε την μεγαλύτερη τιμή σε κάθε ένα από τους δύο άξονες. Το 50% ανά μονάδα στον τυπικό άξονα αντιστοιχεί στο 10% στην κλίμακα ανά ενδεχόμενο “γράμματα”. Η τυπική απόκλιση είναι 5 και επομένως 5 “γράμματα” αντιστοιχούν σε κάθε τυποποιημένη μονάδα. Προφανώς 50:5=10. Οποιοδήποτε άλλο ζευγάρι τιμών διαμορφώνεται με την ίδια λογική.

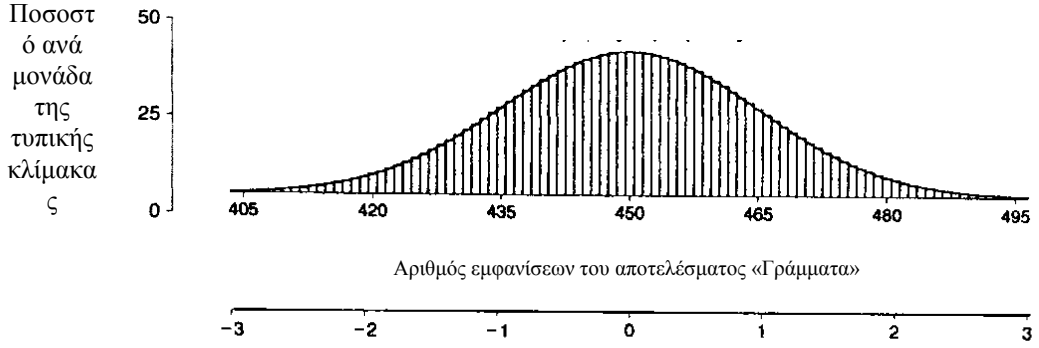
Τα σχήματα που ακολουθούν δίνουν τα ιστογράμματα πιθανότητας για τον αριθμό του αποτελέσματος “γράμματα” σε 100, 400 και 900 ρίψεις ενός νομίσματος. Παρατηρούμε ότι το ιστόγραμμα πλησιάζει καλύτερα την καμπύλη της κανονικής κατανομής όσο αυξάνεται ο αριθμός των φορών που στρίβουμε το νόμισμα. Ο De Moivre απέδειξε την σύγκλιση αυτή στην αρχή του 18ου αιώνα με μαθηματικό τρόπο.

100 ρίψεις νομίσματος



400 ρίψεις νομίσματος

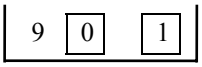
900 ρίψεις νομίσματος



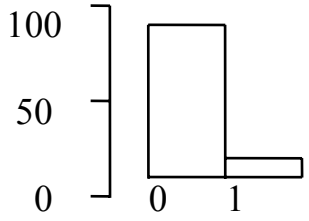
Η Χρήση της Κανονικής Προσέγγισης σε Δειγματοληψία από Υδρία

Στα προηγούμενα, ασχοληθήκαμε με την χρήση της κανονικής κατανομής στο πείραμα που στηρίζεται στο στρίψιμο ενός νομίσματος. Ας δούμε τώρα πώς μπορεί να χρησιμοποιηθεί η κανονική κατανομή στις περιπτώσεις πειραμάτων που αναφέρονται σε επιλογή σφαιρών από υδρία (στην περίπτωση, δηλαδή, που δεν αναφερόμαστε υποχρεωτικά σε ισοπίθανα ενδεχόμενα). Και εδώ, η προσέγγιση μέσω της κανονικής κατανομής εφαρμόζεται πολύ καλά, αρκεί να είναι σαφές ότι, όσο περισσότερο το ιστόγραμμα που αναφέρεται στην συχνότητα των σφαιρών μέσα στην υδρία διαφέρει από την κανονική κατανομή, τόσο μεγαλύτερος αριθμός δειγματοληπτικών δοκιμών απαιτείται προκειμένου η προσέγγιση να είναι εφαρμόσιμη.

Ας θεωρήσουμε, για παράδειγμα, μια υδρία με 9 σφαίρες που έχουν τον αριθμό 0 και μία σφαίρα που έχει τον αριθμό 1. Το ιστόγραμμα της υδρίας αυτής είναι έντονα ασύμμετρο, όπως φαίνεται και στο σχήμα που ακολουθεί.

Σχήμα: Ιστόγραμμα υδρίας 

(εννέα σφαιρών με τον αριθμό 0 και μιας με τον αριθμό 1).

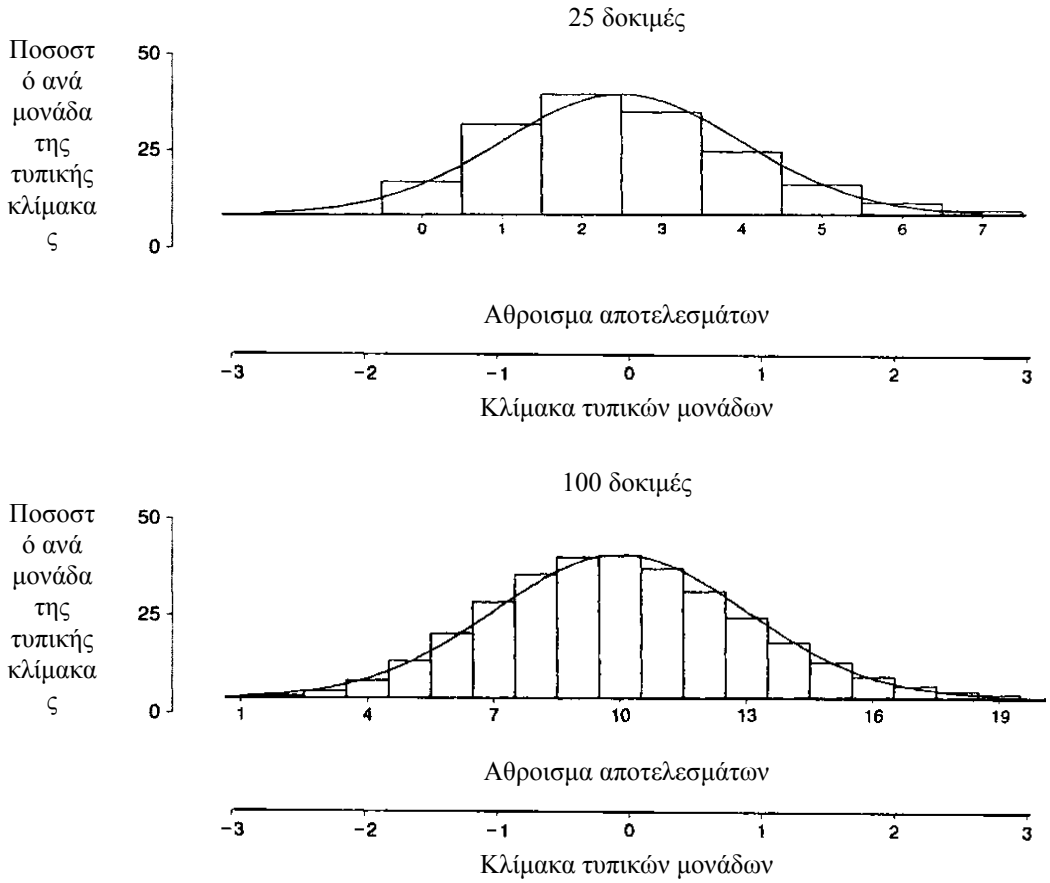


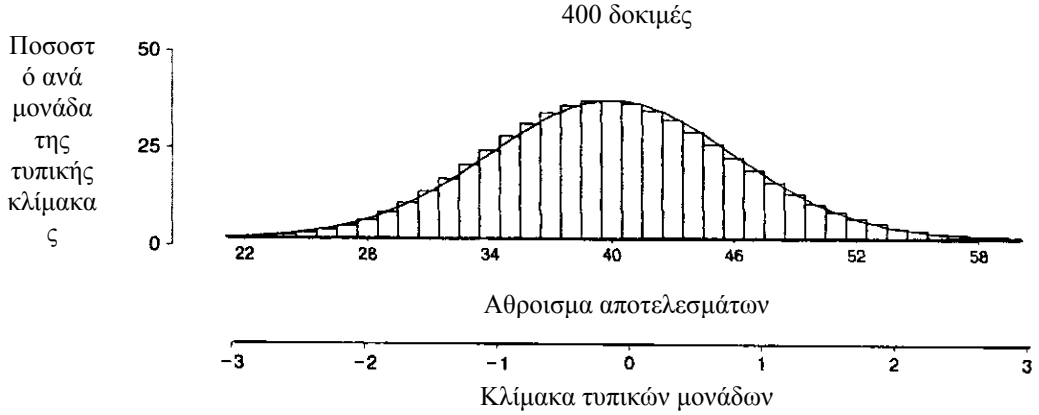
Το ιστόγραμμα πιθανότητας για το άθροισμα των αποτελεσμάτων σε επανάληψη δειγματοληψιών από την υδρία αυτή θα είναι επίσης ασύμμετρο μέχρις ότου ο αριθμός των δειγματοληπτικών επαναλήψεων γίνει αρκετά μεγάλος.

Με ένα πρόγραμμα προσομοίωσης σε υπολογιστή κατασκευάστηκαν ιστογράμματα πιθανότητας για το άθροισμα 25, 100 ή 400 δοκιμών επιλογής από την υδρία.

Όπως φαίνεται στα σχήματα που ακολουθούν με 25 δοκιμές το ιστόγραμμα πιθανότητας είναι ασύμμετρο (υψηλότερο από την κανονική καμπύλη στα αριστερά και χαμηλότερο στα δεξιά). Η κανονική προσέγγιση στην περίπτωση αυτή δεν είναι πολύ ικανοποιητική. Αντίθετα, με 100 δοκιμές, το ιστόγραμμα είναι πολύ πλησιέστερο προς την κανονική κατανομή. Στις 400 δοκιμές είναι πολύ δύσκολο να δούμε την διαφορά μεταξύ του ιστογράμματος και της κανονικής κατανομής.

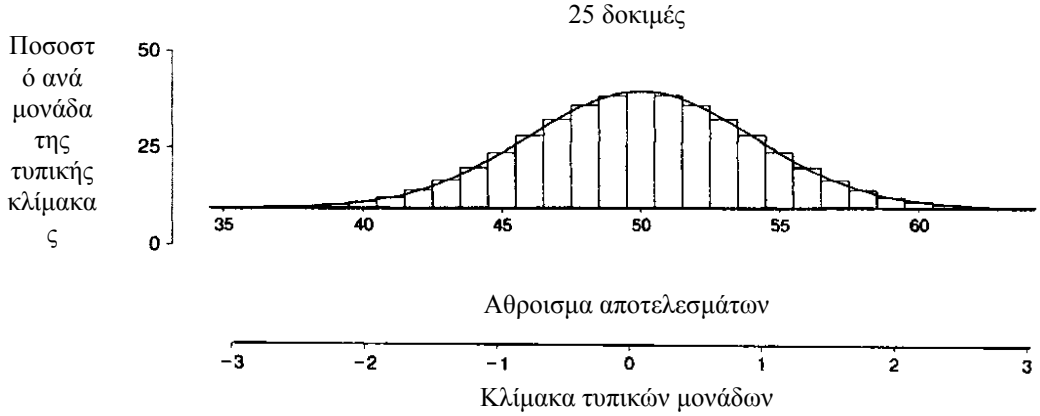
Η κανονική προσέγγιση για το άθροισμα αποτελεσμάτων σε δειγματοληψία από υδρία με 9 σφαίρες που φέρουν τον αριθμό 0 και μια που φέρει τον αριθμό 1

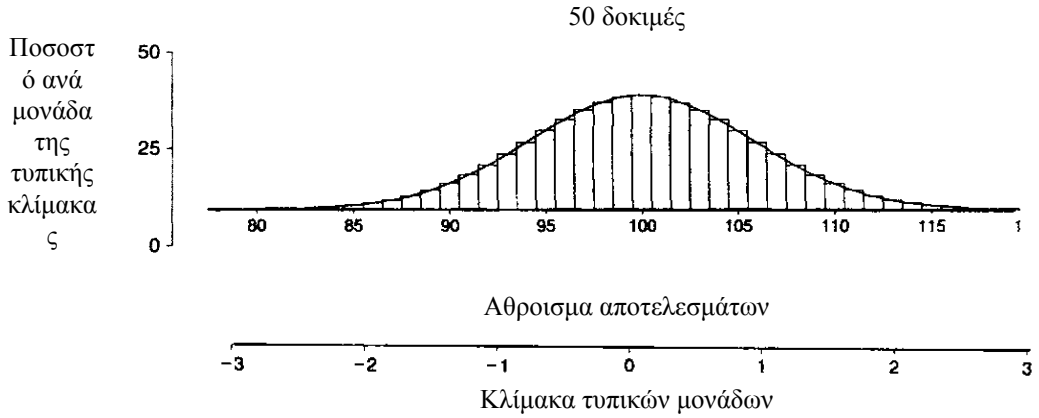




Μέχρι τώρα έχουμε ασχοληθεί με υδρία που έχει σφαίρες που φέρουν τους αριθμούς 0 και 1. Ας δούμε μια περίπτωση που τα πράγματα είναι διαφορετικά. Ας θεωρήσουμε μια υδρία που περιέχει 3 σφαίρες μια με τον αριθμό 1, μια με τον αριθμό 2 και μια με τον αριθμό 3. Το ιστόγραμμα πιθανότητας για το άθροισμα δειγματοληψιών (*με επανάθεση*) με 25 δοκιμές από την υδρία αυτή είναι από την αρχή πολύ κοντά στην κανονική κατανομή. Στις 50 δοκιμές το ιστόγραμμα σχεδόν ταυτίζεται με την κανονική κατανομή.

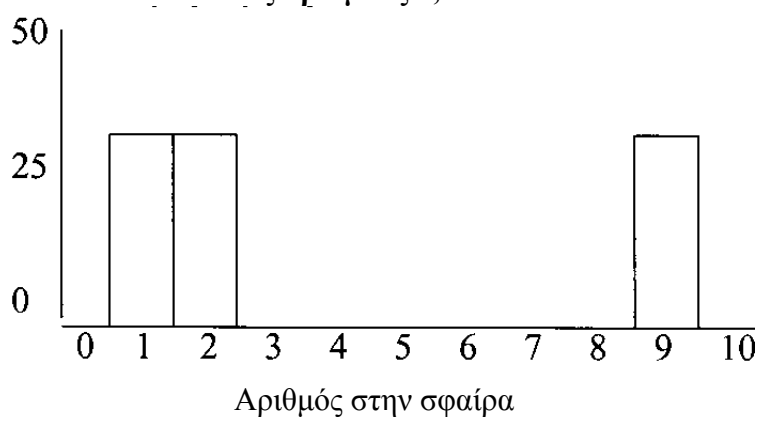
Ιστόγραμμα πιθανότητας για το άθροισμα των αποτελεσμάτων σε 25 και 50 δειγματοληψίες (με επανάθεση) από υδρία με σφαίρες που φέρουν τους αριθμούς 1, 2 και 3





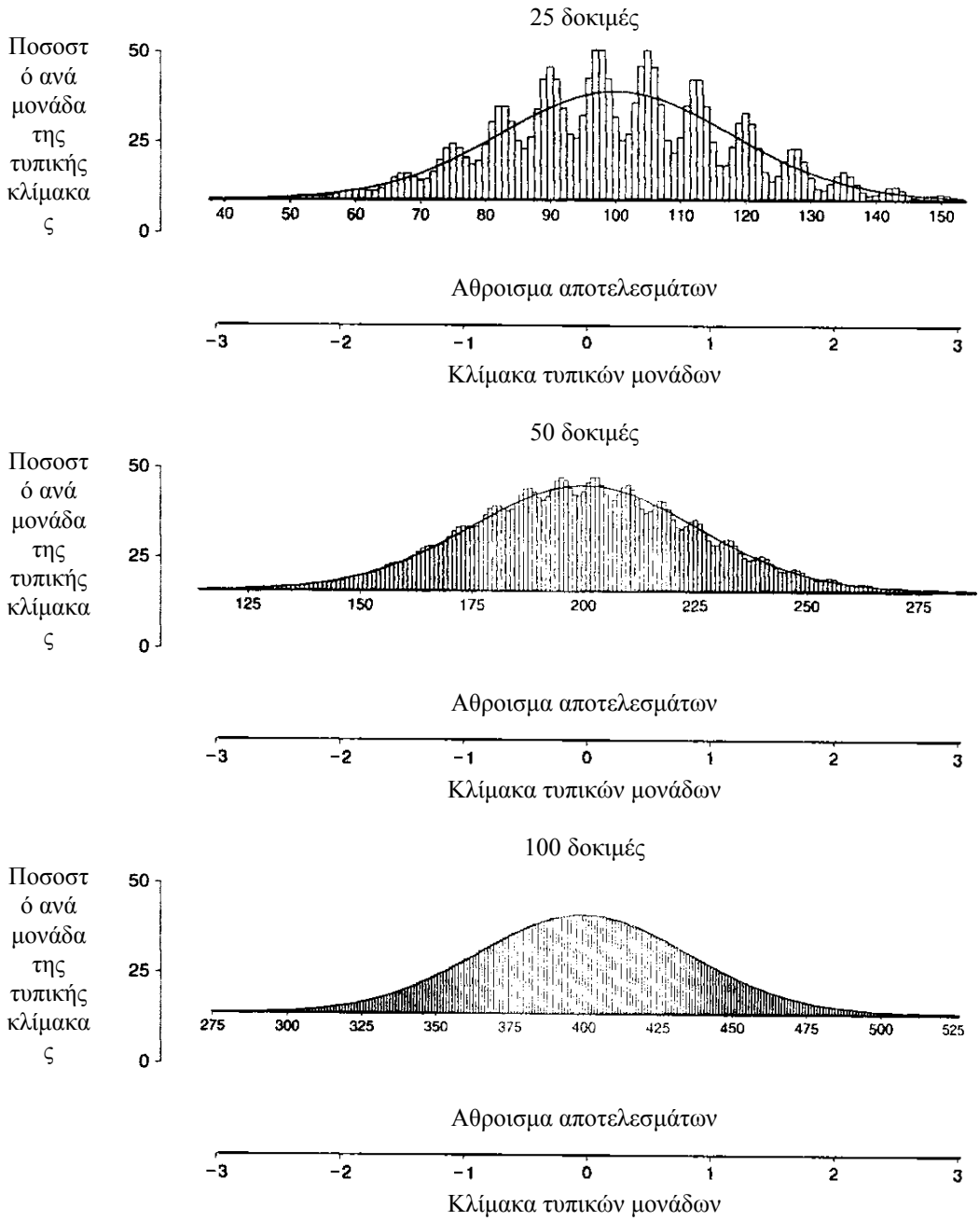
Ας αναφερθούμε τέλος σε μια υδρία με σφαίρες που φέρουν τους αριθμούς 1, 2 και 9. Το ιστόγραμμα πιθανότητας για μια τέτοια υδρία φαίνεται στο σχήμα που ακολουθεί. Είναι προφανές ότι το ιστόγραμμα αυτό δεν έχει καμία σχέση με κανονική κατανομή.

Ιστόγραμμα από υδρία με σφαίρες που φέρουν τους αριθμούς 1, 2 και 9



Με 25 δοκιμές το ιστόγραμμα πιθανότητας του αθροίσματος διαφέρει σημαντικά από την κανονική κατανομή, όπως φαίνεται στα σχήματα που ακολουθούν. Το ιστόγραμμα εμφανίζει μια κυμματοειδή μορφή. Με 50 δοκιμές η κυμματοειδής μορφή εξακολουθεί να υπάρχει αλλά είναι πολύ μικρότερη. Στις 100 δοκιμές το ιστόγραμμα πιθανότητας είναι σχεδόν ίδιο με την κανονική καμπύλη.

Η προσέγγιση μέσω της κανονικής κατανομής του ιστογράμματος πιθανότητας για δειγματοληψία (με επανάθεση) από υδρία που έχει 3 μπάλλες με τους αριθμούς 1, 2 και 9

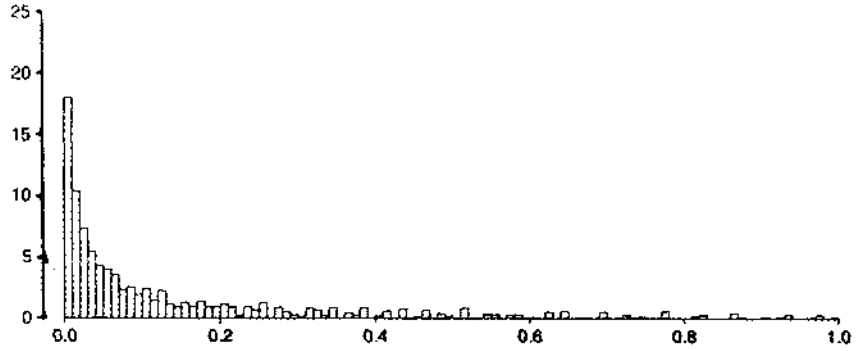


Σημείωση: Πρέπει να τονισθεί ότι η κανονική καμπύλη συνδέεται με το άθροισμα αποτελεσμάτων τυχαίων πειραμάτων. Έτσι, όταν αναφερόμαστε στο γινόμενο αποτελεσμάτων το ιστόγραμμα πιθανότητας συνήθως θα είναι αρκετά διαφορετικό από την καμπύλη της κανονικής κατανομής.

Στο πρώτο από τα σχήματα που ακολουθούν έχουμε το ιστόγραμμα πιθανότητας για το γινόμενο των αποτελεσμάτων σε 10 ρίψεις ενός ζαριού. Είναι προφανές ότι το ιστόγραμμα αυτό δεν έχει καμιά σχέση με την κανονική καμπύλη. Ακόμα και όταν αυξήσουμε τον αριθμό των δοκιμών δεν υπάρχει καμιά βελτίωση όπως φαίνεται στο δεύτερο από τα σχήματα που αναφέρεται στο γινόμενο σε 25 ρίψεις ενός ζαριού. Τα πράγματα είναι ακόμα χειρότερα. Ο πολλαπλασιασμός έχει τελείως διαφορετικά αποτελέσματα από ότι το άθροισμα.

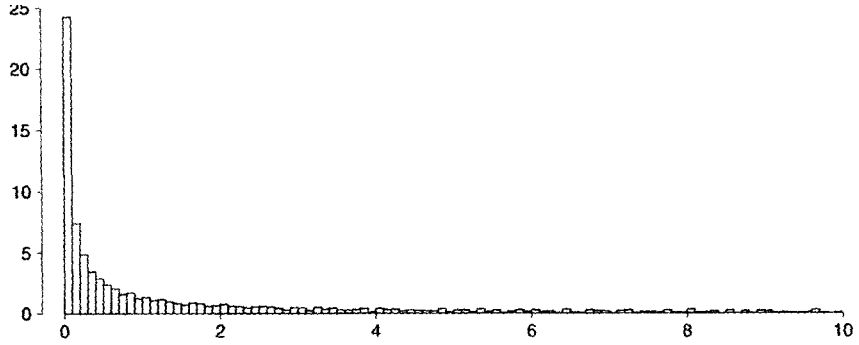
Ιστόγραμμα πιθανότητας για το γινόμενο των αποτελεσμάτων σε 10 και 25 ρίψεις ενός ζαριού

10 ρίψεις ζαριού



Τιμή του γινομένου σε εκατομμύρια

25 ρίψεις ζαριού

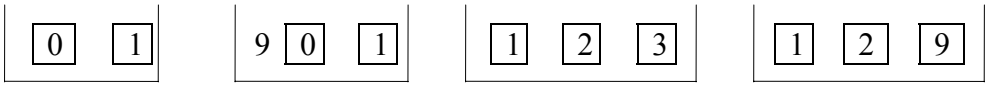


Τιμή του γινομένου σε τρισεκατομμύρια

Σημείωση: Στα ιστογράμματα αυτά η βάση κάθε ορθογωνίου αναφέρεται σε μια έκταση των τιμών του γινομένου και το εμβαδόν του ορθογωνίου ισούται με την πιθανότητα το γινόμενο να πάρει τιμή στην έκταση αυτή. Σε 10 ρίψεις ζαριού η εικόνα αναφέρεται στο 94% περίπου του εμβαδού. Σε 25 ρίψεις ζαριού η εικόνα αναφέρεται στο 80% περίπου του εμβαδού[1]. Στο πρώτο από τα σχήματα, η κατακόρυφη κλίμακα είναι ποσοστό ανά 10000, ενώ στο δεύτερο σχήμα, η κατακόρυφη κλίμακα είναι ποσοστό ανά 10^{11}.

Συμπεράσματα

Εξετάσαμε το άθροισμα των αποτελεσμάτων δειγματοληψίας από υδρία σε τέσσερις διαφορετικές περιπτώσεις



(Στην πρώτη υδρία, υπάρχει μία σφαίρα με τον αριθμό 0 και μία με το 1. Στην δεύτερη, εννέα σφαίρες με 0 και μία με 1. Στην τρίτη, μία σφαίρα με 1, μία με 2 και μία με 3 και στην τέταρτη μία με 1, μία με 2 και μία με 9).

[1] Στις 10 δοκιμές, το ιστόγραμμα για το γινόμενο είναι κατασκευασμένο σε κλίμακα με μονάδα το ένα εκατομμύριο. Το 6% του εμβαδού είναι πέρα από το σημείο που αντιστοιχεί στο 1 εκατομμύριο και δεν εμφανίζεται. Η μεγαλύτερη τιμή για το γινόμενο είναι το 6 πολλαπλασιασμένο με τον εαυτό του 10 φορές, δηλαδή $6^{10} = 60466176$. Σε 25 ρίψεις του ζαριού, το ιστόγραμμα είναι κατασκευασμένο σε κλίμακα με μονάδα το ένα τρισεκατομμύριο. Το 20% του εμβαδού είναι πέρα από το σημείο που αντιστοιχεί στο 1 τρισεκατομμύριο και δεν εμφανίζεται. Το μεγαλύτερο δυνατό αποτέλεσμα του γινομένου είναι ένας πραγματικά μεγάλος αριθμός $6^{25} \cong 3 \times 10^{19}$.

Θα μπορούσε κανείς να εξετάσει πολύ περισσότερες περιπτώσεις. Το συμπέρασμα όμως είναι ότι η ακολουθία είναι περίπου η ίδια. Ότι δηλαδή με αρκετές επαναλήψεις το ιστόγραμμα πιθανότητας για το άθροισμα θα προσεγγίζεται πολύ καλά από την κανονική κατανομή. Αυτό, όπως ήδη είδαμε, είναι συνέπεια του *κεντρικού οριακού θεωρήματος (central limit theorem)*. Στο πλαίσιο του συγκεκριμένου παραδείγματος η διατύπωση του κεντρικού οριακού θεωρήματος μπορεί να απλουστευθεί ως εξής:

Σε δειγματοληψία με επανάθεση αριθμημένων σφαιρών από μια υδρία το ιστόγραμμα πιθανότητας για το άθροισμα των αριθμών που φέρουν οι σφαίρες που θα επιλεγούν ακολουθεί την κανονική κατανομή έστω και αν η συχνότητα των σφαιρών στην υδρία δεν ακολουθεί την κανονική κατανομή. Το ιστόγραμμα θα πρέπει να αναφέρεται σε τυποποιημένες μονάδες και ο αριθμός των δοκιμών θα πρέπει να είναι ικανοποιητικά μεγάλος.

Σημείωση: Το κεντρικό οριακό θεώρημα εφαρμόζεται μόνο στο άθροισμα και όχι σε άλλες πράξεις όπως π.χ. το γινόμενο.

Σημείωση: Ένα φυσιολογικό ερώτημα είναι πόσες δοκιμές χρειάζονται για να είναι η προσέγγιση αυτή ικανοποιητική. Στο ερώτημα αυτό δεν υπάρχει μία προκαθορισμένη απάντηση. Η απάντηση εξαρτάται από το περιεχόμενο της υδρίας. Παρ' όλα αυτά, για πάρα πολλές περιπτώσεις το ιστόγραμμα πιθανότητας για το άθροισμα σε 100 δοκιμές είναι πολύ κοντά στην κανονική κατανομή.

Όταν το ιστόγραμμα πιθανότητας ακολουθεί την κανονική κατανομή μπορεί να συνοψισθεί με την μέση τιμή και το τυπικό σφάλμα (standard error). Για παράδειγμα, αν θέλουμε να κατασκευάσουμε ένα ιστόγραμμα χωρίς παραπέρα πληροφορίες είναι δυνατόν αυτό να γίνει σε μονάδες τυπικής κλίμακας, τουλάχιστον κατά προσέγγιση.

Για να ολοκληρωθεί η εικόνα χρειάζεται να μετασχηματίσουμε τις μονάδες της τυποποιημένης κλίμακας στις αρχικές μονάδες με αντικατάσταση των ερωτηματικών στο σχήμα. Τότε έχουμε στην διαθεσή μας ό,τι πληροφορίες θα μπορούσε να έχει κανείς από το ιστόγραμμα δοθέντος ότι αυτό ακολουθεί την κανονική κατανομή. Η μέση τιμή αναφέρεται στο κέντρο του ιστογράμματος πιθανότητας στον οριζόντιο άξονα και το τυπικό σφάλμα καθορίζει την έκταση της κατανομής.

Η μέση τιμή και το τυπικό σφάλμα για το άθροισμα των αποτελεσμάτων σε επαναλαμβανόμενες δειγματοληπτικές επιλογές μπορούν να υπολογισθούν από τον αριθμό των δειγματοληπτικών δοκιμών, τον μέσο των αριθμών των σφαιριδίων στην υδρία και την τυπική απόκλιση των αριθμών αυτών. Οι τρεις αυτές ποσότητες καθορίζουν την συμπεριφορά του αθροίσματος. Γι' αυτό ακριβώς τον λόγο η τυπική απόκλιση των αριθμών των σφαιρών στην υδρία είναι τόσο σημαντική για την διασπορά.

Σημείωση: Στην ενότητα που προηγήθηκε ασχοληθήκαμε με την χρήση της κανονικής καμπύλης για την ερμηνεία δεδομένων. Σε πολλές περιπτώσεις η ερμηνεία αυτή μπορεί να προκύψει και μαθηματικά με την χρήση δύο ειδών σύγκλισης στις οποίες ουσιαστικά αναφέρονται όσα προηγήθηκαν. Όταν ο αριθμός των επαναλήψεων του πειράματος είναι μεγάλος το εμπειρικό ιστόγραμμα θα πλησιάζει το ιστόγραμμα πιθανότητας. Όταν ο αριθμός των

δοκιμών είναι μεγάλος το ιστόγραμμα πιθανότητας για το άθροισμα θα πλησιάζει την κανονική κατανομή. Συνεπώς, όταν τόσο ο αριθμός των επαναλήψεων όσο και ο αριθμός των επιλογών είναι και οι δύο μεγάλοι το εμπειρικό ιστόγραμμα για το άθροισμα θα πλησιάζει την κανονική καμπύλη. Για να εφαρμόσουμε όμως την προσέγγιση αυτή σε πραγματικά δεδομένα θα πρέπει να επιβεβαιώσουμε ότι μπορούμε να αντιστοιχίσουμε την διαδικασία που παράγει τα δεδομένα με διαδικασία επιλογής αριθμημένων σφαιρών από μία υδρία και στην συνέχεια υπολογισμό του αθροίσματος των αποτελεσμάτων.

Με μαθηματικούς όρους αν n είναι ο αριθμός των δοκιμών και k είναι ο αριθμός των επαναλήψεων η μαθηματική συνθήκη που χρειαζόμαστε είναι ότι

$$k/\sqrt{n}\,\log n \to \infty .$$