Εκθέτης

Φύλλα Μαθηματικής Παιδείας

Φυλλο 26, 10 Μαϊου 2026

ISSN 2241-3367
Εκδίδεται στην Αθήνα
Διανέμεται & αναπαράγεται ελεύθερα
e-mail: info@ekthetis.gr
Δικτυακός Τόπος:
http://www.ekthetis.gr/
Στοιχειοθετείται με το LaTeX2e
Έκδοση-Επιμέλεια: Ν.Σ. Μαυρογιάννης
Διευθύνεται από Συντακτική Επιτροπή

Ένα πρόβλημα πιθανοτήτων με πολλές αναγνώσεις

Λάμπρος Κατσάπας
Μαθηματικός (MSc)

Εις μνήμην

Ο συγγραφέας αφιερώνει την παρούσα εργασία στην μνήμη του Καθ. Μαθηματικών Δρ Σταύρου Παπαδόπουλου που υπήρξε και μέλος της Σ.Ε. του Εκθέτη.

Περίληψη

Στην παρούσα εργασία μελετάμε ένα στοχαστικό πρόβλημα που προέρχεται από ανάρτηση σε μαθηματική ομάδα κοινωνικού δικτύου. Η προσέγγιση του προβλήματος περιλαμβάνει αναδρομή, ασυμπτωτική ανάλυση για την εξαγωγή της οριακής κατανομής και αξιολόγηση του σφάλματος προσέγγισης εμπειρικά. Τέλος, αναδεικνύουμε τη σύνδεση του προβλήματος με γνωστά θεωρητικά και εφαρμοσμένα προβλήματα.

Λέξεις κλειδιά

Αναδρομή, Ασυμπτωτική ανάλυση, Αναπτύγμα Taylor-Maclaurin, Οριακή κατανομή, Κανονική κατανομή, Κατανομή Poisson, Ομοιόμορφη σύγκλιση, Mathematica software, Monte Carlo προσομοίωση, Πρόβλημα των Γενεθλίων, Συναρτήσεις Hash, Αλγόριθμος Rho, Αλγόριθμος της χελώνας και του λαγού.

Περιεχόμενα

1 Εισαγωγή

Την τελευταία δεκαετία τα μέσα κοινωνικής δικτύωσης έχουν εισχωρήσει δυναμικά στην καθημερινότητά μας. Μας παρέχουν την δυνατότητα να αλληλοεπιδρούμε ως άτομα, να εκφραζόμαστε, να μαθαίνουμε από τους άλλους αλλά και οι άλλοι από εμάς. Η ελληνική μαθηματική κοινότητα έχει αναπτύξει μια ομολογουμένως αξιόλογη δράση μέσα από ιστοσελίδες ή blogs του διαδικτύου αλλά και κοινωνικά δίκτυα όπως για παράδειγμα το Facebook. Στο Facebook, ειδικότερα, μπορεί να συναντήσει κανείς μεγάλο πλήθος ομάδων περί τα μαθηματικά. Δυστυχώς όμως ο μεγαλύτερος όγκος της πληροφορίας χάνεται με την πάροδο του χρόνου με αποτέλεσμα τα ενδιαφέροντα μαθηματικά προβλήματα να περνούν στη λήθη. Ένα τέτοιο ενδιαφέρον πρόβλημα είναι κατά τη γνώμη μας και το παρακάτω το οποίο «κρύβει» αρκετά και ενδιαφέροντα μαθηματικά. Ανέβηκε στην ομάδα Let's solve Math Problems! του Facebook από την Κύπρια συνάδελφο μαθηματικό Καρολίνα Παπαχρυσοστόμου.

2 Το πρόβλημα

Η ετικέτα στο μπουκάλι χαπιών που σας έδωσε ο γιατρός σας, λέει:

«Λήψη μισού χαπιού καθημερινά.»

Ένας τρόπος για να παίρνετε τα χάπια σας είναι ο εξής:

- Καθημερινά αφαιρείτε τυχαία ένα χάπι από το μπουκάλι.
- Αν είναι ολόκληρο χάπι, το σπάτε στη μέση, παίρνετε το μισό και επιστρέφετε το άλλο μισό στο μπουκάλι.
- Αν είναι μισό χάπι επίσης το παίρνετε και συνεχίζετε κανονικά την επόμενη μέρα.

Αν το μπουκάλι έχει αρχικά n χάπια να υπολογίσετε πόσα ολόκληρα χάπια θα επιλεγούν από το μπουκάλι, κατά μέσο όρο, προτού αφαιρεθεί το πρώτο μισό χάπι.

Προτού αναπτύξουμε τη λύση του προβλήματός ας τονίσουμε ότι με την έκφραση «αφαιρείτε τυχαία ένα χάπι από το μπουκάλι» εννοούμε ότι, κάθε δεδομένη στιγμή, όλα τα χάπια που βρίσκονται στο μπουκάλι, ολόκληρα ή μισά, έχουν την ίδια πιθανότητα να επιλεγούν.

3 Θεωρητική Ανάλυση του Προβλήματος

Είναι γνωστό ότι αν X είναι μια τυχαία μεταβλητή (συντομογραφικά τ.μ.) με τιμές στο σύνολο

$$A = \{1, 2, ..., n\}$$

τότε η μέση τιμή της ορίζεται από τη σχέση

$$E(X) = \sum_{x=1}^{n} xP(X = x).$$

Με κατάλληλη αναδιάταξη των όρων του αθροίσματος μπορούμε να καταλήξουμε στην παρακάτω ισότητα:

$$E(X) = \sum_{x=1}^{n} P(X \geq x).$$

Πράγματι, ξεκινώντας από τον ορισμό, αν γράψουμε $x = \sum_{k=1}^{x} 1$ και κάνουμε εναλλαγή των αθροισμάτων τότε έχουμε

$$\begin{aligned} E(X) &= \sum_{x=1}^{n} \left(\sum_{k=1}^{x} 1 \right) P(X = x) \\ &= \sum_{k=1}^{n} \left(\sum_{x=k}^{n} P(X = x) \right) \\ &= \sum_{k=1}^{n} P(X \geq k) = \sum_{x=1}^{n} P(X \geq k) \quad (1) \end{aligned}$$

Μια λιγότερο τυπική αλλά πιο διαισθητική παρουσίαση της παραπάνω σχέσης είναι η εξής:

$$E(X) = \sum_{x=1}^{n} xP(X = x) =$$

$$\underbrace{P(X = 1) + P(X = 2) + P(X = 3) + ... + P(X = n)}_{P(X \geq 1)}$$

$$+ \underbrace{P(X = 2) + P(X = 3) + ... + P(X = n)}_{P(X \geq 2)}$$

$$+ ...$$

$$... + \underbrace{P(X = n)}_{P(X \geq n)}$$

Σε αντίθεση με τον πρώτο τύπο του ορισμού της μέσης τιμής όπου η άθροιση γινόταν στήλη προς στήλη εδώ τώρα αθροίζουμε γραμμή προς γραμμή και παίρνουμε την (1).

Ας ξεκινήσουμε τώρα να επεξεργαζόμαστε το πρόβλημά μας. Θεωρούμε την τ.μ. T που εκφράζει *τον αριθμό των ολόκληρων χαπιών που θα εξάγουμε από το μπουκάλι μέχρι να πάρουμε το πρώτο μισό χάπι.* Η T μπορεί να πάρει τις τιμές $1, 2, 3, ..., n$. Το ενδεχόμενο $\{T \geq k\}, k = 1, 2, 3, ..., n$, είναι το ενδεχόμενο να καταναλώσουμε τουλάχιστον k ολόκληρα χάπια έως ότου να πάρουμε το πρώτο μισό χάπι ή, ισοδύναμα, το ενδεχόμενο όπως οι πρώτες k δοκιμές μας έδωσαν ολόκληρα χάπια.

Για τον υπολογισμό της πιθανότητας $P(T \geq k)$ θα χρησιμοποιήσουμε το:

Θεωρημα. (Πολλαπλασιαστικο Θεωρημα) *Έστω* $A_i, i = 1, 2, ..., n$ *ενδεχόμενα ενός δειγματικού χώρου* Ω *με*

$$P(A_1 \cap A_2 \cap ... \cap A_{n-1}) > 0.$$

Τότε

$$P(A_1 \cap A_2 \cap ... \cap A_n)$$
$$= P(A_1)P(A_2|A_1)...P(A_n|A_1 \cap A_2 \cap ... \cap A_{n-1}).$$

Ας είναι A_i το ενδεχόμενο τη i-οστή ημέρα να εξαχθεί ολόκληρο χάπι. Κάθε φορά που παίρνουμε ένα ολόκληρο χάπι, μέχρι να πετύχουμε το μισό, ο αριθμός των χαπιών στο μπουκάλι είναι σταθερός και ίσος με n (ένα ολόκληρο χάπι βγάζουμε και ένα μισό επιστρέφουμε).

Στην αρχή υπάρχουν μόνο ολόκληρα χάπια στο μπουκάλι, οπότε η πιθανότητα $P(T \geq 1) = P(A_1)$ ισούται με 1. Μετά από την πρώτη επιλογή, επιστρέφουμε ένα μισό χάπι στο μπουκάλι, άρα απομένουν $n-1$ ολόκληρα χάπια και 1 μισό. Η πιθανότητα να επιλέξουμε την πρώτη και τη δεύτερη ημέρα ολόκληρο χάπι είναι:

$$P(T \geq 2) = P(A_1 \cap A_2) = P(A_1)P(A_2|A_1)$$
$$= 1 \cdot \frac{n-1}{n} = \frac{n-1}{n}$$

Αναλόγως, η πιθανότητα να επιλέξουμε την πρώτη, τη δεύτερη και την τρίτη ημέρα ολόκληρο χάπι είναι:

$$P(T \geq 3) = P(A_1 \cap A_2 \cap A_3)$$
$$= P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)$$
$$= 1 \cdot \frac{n-1}{n} \cdot \frac{n-2}{n}$$
$$= \frac{(n-1)(n-2)}{n^2}$$

και γενικά, η πιθανότητα τις πρώτες k ημέρες να επιλέξουμε ολόκληρα χάπια είναι:

$$P(T \geq k) = P(A_1 \cap A_2 \cap ... \cap A_k)$$
$$= P(A_1)P(A_2|A_1)...P(A_k|A_1 \cap A_2 \cap ... \cap A_{k-1})$$
$$= \frac{(n-1)(n-2)...(n-k+1)}{n^{k-1}}$$

Επομένως,

$$P(T \geq k) = \frac{(n-1)(n-2)...(n-k+1)}{n^{k-1}}$$
$$= \frac{n!}{(n-k)!n^k}, \quad k = 1, 2, 3, ..., n.$$

[1]Είναι $\lim_{n\to\infty} \frac{2n^2-n}{2n^2} = 1$ και επομένως, για μεγάλες τιμές του n, το $-n$ είναι αμελητέο.

Από την (1) έχουμε λοιπόν ότι

$$E(T) = \sum_{k=1}^{n} P(T \geq k)$$
$$= \sum_{k=1}^{n} \frac{n!}{(n-k)!n^k}. \quad (2)$$

Το πρόβλημα έχει λυθεί, τουλάχιστον θεωρητικά. Η ζητούμενη μέση τιμή είναι αυτή που δίνεται από τη σχέση (2).

Σε αυτό το σημείο, είναι φυσιολογικό να εξεταστεί η πρακτική εφαρμογή του τύπου. Για παράδειγμα, ας υπολογίσουμε τον αριθμό των πράξεων,πολλαπλασιασμών-διαιρέσεων και προσθέσεων, που απαιτούνται για τον υπολογισμό της μέσης τιμής.Ο αριθμητής απαιτεί $n-1$ πολλαπλασιασμούς όπως και ο παρονομαστής. Έτσι λοιπόν απαιτούνται συνολικά $2n-1$ πράξεις για τον υπολογισμό του κλάσματος. Αθροίζοντας παίρνουμε:

$$\sum_{k=1}^{n} (2n-1) = 2n^2 - n \text{ πράξεις.}$$

Σε γενικές γραμμές μπορούμε να πούμε ότι για μεγάλες τιμές του n χρειαζόμαστε περίπου $2n^2$ πράξεις[1] για τον υπολογισμό της μέσης τιμής. Αυτό εγείρει το ερώτημα:

> *Υπάρχει τρόπος να προσεγγίσουμε αυτήν την ποσότητα με μια απλή και αριθμητικά αποδοτική μέθοδο*;

Η απάντηση στο ερώτημα αυτό είναι καταφατική.

3.1 Προσέγγιση μέσω αναδρομής

Αν και ο τύπος (2) προσφέρει έναν ακριβή τρόπο υπολογισμού του ζητούμενου αθροίσματος, εντούτοις δεν είναι υπολογιστικά βέλτιστος, ειδικά για μεγάλες τιμές του n. Μια πιο αποδοτική προσέγγιση βασίζεται στην παρατήρηση ότι οι όροι του αθροίσματος μπορούν να υπολογιστούν αναδρομικά, με μικρότερο υπολογιστικό κόστος. Συγκεκριμένα, αν δηλώσουμε:

$$S(n,k) = \frac{n!}{(n-k)!n^k}$$

τον k–προσθετέο της (2) τότε

$$S(n,k+1) = \frac{n-k}{n} S(n,k).$$

Αν λοιπόν έχουμε υπολογίσει τον k–προσθετέο, τότε για τον υπολογισμό του $k+1$ προσθετέου απαιτούνται ένας πολλαπλασιασμός και μία διαίρεση. Αν,επιπλέον, ορίσουμε

$$A(n,m) = \sum_{k=1}^{m} S(n,k),$$

τότε είναι

$$A(n,n) = E(T)$$

και ο υπολογισμός του αθροίσματοςτης (2) μπορεί να γίνει αναδρομικά ως εξής:

Ορίζουμε:

- Το αρχικό άθροισμα ως

$$A(n,1) = S(n,1) = 1$$

και

- κάθε επόμενο άθροισμα ως:

$$A(n,k+1) = A(n,k) + S(n,k+1).$$

Πρακτικά:

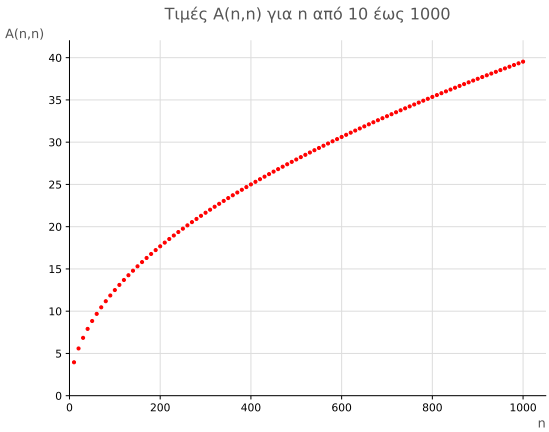
- Ξεκινάμε με $A = 1, S = 1$ και για κάθε $k = 2,3,...,n$ υπολογίζουμε τον νέο όρο: $S := \frac{n-(k-1)}{n} S$.
- Έπειτα προσθέτουμε στο προηγούμενο άθροισμα A Ατον νέο όρο S: $A := A + S$.

Συνολικά, για κάθε $k = 2,3,...,n$ χρειάζονται 3 πράξεις ανά βήμα. Εφόσον εκτελούμε $n-1$ βήματα για τον υπολογισμό της $E(T)$, ο συνολικός αριθμός πράξεων είναι:

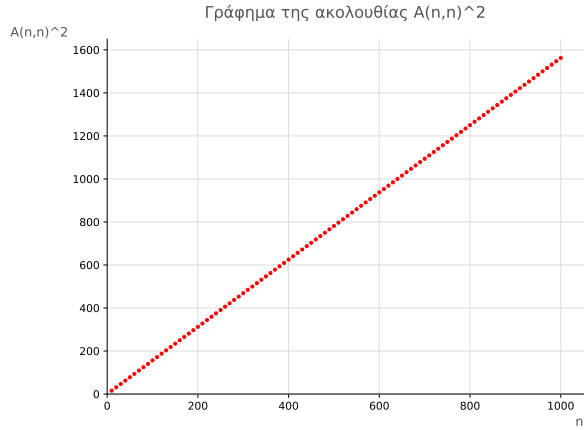
$$3n-3 \text{ πράξεις.}^2$$

Το άθροισμα που χρησιμοποιείται για τον υπολογισμό των τιμών καθιστά δύσκολη την εξαγωγή ενός απλού «κλειστού τύπου». Συμβολικά υπολογιστικά πακέτα, όπως το Maple και το Mathematica, επιστρέφουν λύσεις που περιλαμβάνουν ειδικές συναρτήσεις, όπως η συνάρτηση Kummer.

Μπορούμε όμως να ζητήσουμε από το υπολογιστικό πακέτο Mathematica να σχεδιάσει το γράφημα των τιμών του αθροίσματος $A(n,n)$ για τιμές του από 10 έως 1000 με βήμα 10.



Η προσεκτική παρατήρηση του γραφήματος μας οδηγεί στην εικασία ότι το άθροισμα $A(n,n)$ αυξάνεται με ρυθμό ανάλογο της τετραγωνικής ρίζας του. Για να το διαπιστώσουμε αυτό σχεδιάζουμε την ακολουθία $(A(n,n))^2$. Αν η εικασία μας είναι σωστή, τότε τα σημεία της $(A(n,n))^2$ θα βρίσκονται, περίπου, πάνω σε ευθεία. Παρακάτω φαίνεται το γράφημά της.



Το γράφημα της $(A(n,n))^2$ δείχνει ότι τα σημεία ευθυγραμμίζονται, γεγονός που επιβεβαιώνει την εικασία ότι το άθροισμα $A(n,n)$ αυξάνεται, τουλάχιστον μέχρι το $n = 1000$, με ρυθμό ανάλογο της τετραγωνικής ρίζας του ν, δηλαδή

$$A(n,n) \simeq C\sqrt{n}.$$

Από την κλίση της ευθείας εκτιμάται ότι

$$C^2 \simeq 1,25^2 \Leftrightarrow C \simeq 1,25.$$

Το παραπάνω εμπειρικό αποτέλεσμα ενισχύει την ιδέα ότι η προσέγγιση του αθροίσματος μπορεί να γίνει με έναν ασυμπτωτικό τύπο της μορφής

$$A(n,n) \approx C\sqrt{n},$$

κάτι που είναι ιδιαίτερα χρήσιμο για μεγάλες τιμές του n. Στη συνέχεια, θα προσπαθήσουμε να τεκμηριώσουμε θεωρητικά αυτή την προσέγγιση και να υπολογίσουμε με ακρίβεια τη σταθερά C. Συγκεκριμένα, θα αποδείξουμε ότι

$$C = \sqrt{\pi/2},$$

δηλαδή[3]:

$$E(T) = \sum_{k=1}^{n} \frac{n!}{(n-k)!n^k} \approx \sqrt{\frac{n\pi}{2}} \quad (n \to \infty) \quad (3).$$

[2] Περίπου $3n$ πράξεις, για μεγάλες τιμές του n. Σημαντική βελτίωση σε σχέση με την προηγούμενη μέθοδο.

[3] Σχετικά με το σύμβολο ≈ ας σημειώσουμε ότι οποτεδήποτε χρησιμοποιηθεί παρακάτω θα έχει την εξής έννοια: Αν $(a_n), (b_n)$ είναι δύο ακολουθίες πραγματικών αριθμών με $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$ τότε θα γράφουμε $a_n \approx b_n$.

3.2 Προσέγγιση μέσω ασυμπτωτικού τύπου

Πριν προχωρήσουμε στην απόδειξη της (3), μετασχηματίζουμε τη μέση τιμή σε μια μορφή κατάλληλη για περαιτέρω ανάλυση. Από την (2) έχουμε

$$E(T) = \sum_{k=1}^{n} \frac{n!}{(n-k)!n^k} \overset{j:=n-k}{=}$$

$$= \frac{n!}{n^n} \sum_{j=0}^{n-1} \frac{n^j}{j!}$$

$$= \frac{n!}{n^n} \sum_{\mathrm{j}=0}^{n} \frac{n^j}{j!} - 1. \quad (4)$$

Η δυσκολία στην εύρεση ασυμπτωτικού τύπου για την $E(T)$ έγκειται στην ύπαρξη του αθροίσματος της σχέσης (4). Για την ποσότητα $\frac{n!}{n^n}$ μπορούμε να χρησιμοποιήσουμε τον ασυμπτωτικό τύπο του Stirling για το $n!$.

Πρόταση. Τύπος του Stirling*:*

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

Για την εύρεση της ασυμπτωτικής συμπεριφοράς του αθροίσματος της (4) θα παρουσιάσουμε δύο διαφορετικές αποδείξεις. Στην πρώτη απόδειξη θα χρησιμοποιήσουμε γνωστά αποτελέσματα της Θεωρίας Πιθανοτήτων ενώ στη δεύτερη, θα χρησιμοποιήσουμε εργαλεία της Ανάλυσης. Η πρώτη προσέγγιση ξεχωρίζει για την κομψότητά της όμως η δεύτερη φωτίζει σε βάθος την ασυμπτωτική συμπεριφορά του αθροίσματος της (4), προσφέροντας έτσι καλύτερη κατανόηση του αποτελέσματος. Σε κάθε περίπτωση θα αποδείξουμε την παρακάτω πρόταση:

Πρόταση 1. *Ισχύει:* $\lim_{n\to\infty} \sum_{j=0}^{n} \frac{e^{-n}n^j}{j!} = \frac{1}{2}$ *ή, ισοδύναμα,* $\sum_{j=0}^{n} \frac{n^j}{j!} \approx \frac{e^n}{2}$

Έχοντας αποδείξει την Πρόταση 1 και χρησιμοποιώντας τον τύπο του Stirling θα είμαστε σε θέση να γράψουμε στην (4):

$$E(T) + 1 = \frac{n!}{n^n} \sum_{j=0}^{n} \frac{n^j}{j!}$$

$$\approx \frac{e^n}{2n^n} \cdot \sqrt{2\pi n}\left(\frac{n}{e}\right)^n = \sqrt{\frac{\pi n}{2}}$$

ή

$$E(T) \approx \sqrt{\frac{\pi n}{2}}$$

και έτσι θα έχουμε αποδείξει τον ασυμπτωτικό τύπο της(3). Παρουσιάζουμε παρακάτω τις δύο αποδείξεις της Πρότασης 1 που προαναφέραμε.

Πιθανοθεωρητική Απόδειξη Αρχικά, υπενθυμίζουμε κάποιες βασικές έννοιες και αποτελέσματα της Θεωρίας Πιθανοτήτων.

1. *Διακριτές Μεταβλητές* Μια τυχαία μεταβλητή X λέγεται διακριτή αν παίρνει, με πιθανότητα 1, πεπερασμένο ή αριθμήσιμο σύνολο τιμών

   $$\{x_0, x_1, x_2, \dots\}$$

   δηλαδή

   $$P\left(X \in \{x_0, x_1, x_2, \dots\}\right) = \sum_{k=0}^{\infty} P\left(X = x_k\right) = 1.$$

2. *Συνεχείς Μεταβλητές* Μια τυχαία μεταβλητή X λέγεται συνεχής αν υπάρχει μη αρνητική συνάρτηση $f(x) \geq 0, x \in \mathbb{R}$ με

   $$\int_{-\infty}^{+\infty} f(x)dx = 1,$$

   τέτοια, ώστε για οποιουσδήποτε πραγματικούς αριθμούς α και β με $\alpha \leq \beta$, να ισχύει:

   $$P(\alpha < X \leq \beta) = \int_{\alpha}^{\beta} f(x)dx.$$

3. *Μέση τιμή και διασπορά αθροισμάτων* Αποδεικνύεται ότι αν καθε μιά από τις τυχαίες μεταβλητές $X_1, X_2, ..., X_n$, έχει μέση τιμή μ, τότε

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

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

   Αν, επιπλέον, είναι ανεξάρτητες και κάθε μια έχει διασπορά σ^2 τότε

   $$Var(X_1 + X_2 + ... + X_n)$$

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

   $$= nVar(X_1) = n\sigma^2.$$

4. *Κατανομή Poisson* Μία διακριτή τυχαία μεταβλητή X λέμε ότι ακολουθεί κατανομή Poisson με παράμετρο $\lambda > 0$ (συμβολικά: $X \sim Po(\lambda)$), αν για $k = 0, 1, 2, \dots$, ισχύει:

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

   Η συνάρτηση κατανομής της $X \sim Po(\lambda)$ είναι η

   $$F(x) = P(X \leq x)$$

   $$= \begin{cases} 0 & , \quad -\infty < x < 0 \\ \sum_{k=0}^{[x]} \frac{e^{-\lambda}\lambda^k}{k!} & , \quad 0 \leq x < +\infty \end{cases} \quad (5)$$

   όπου $[x]$ παριστάνει το ακέραιο μέρος του αριθμού x.

5. *Ιδιότητες σχετικές με την Κατανομή Poisson*

   - Για τη μέση τιμή και τη διασπορά της $X \sim Po(\lambda)$ ισχύει:

   $$E(X) = Var(X) = \lambda \quad (6).$$

   - Το άθροισμα ανεξάρτητων τυχαίων μεταβλητών $X_1, X_2, ..., X_n$ με $X_i \sim Po(\lambda_i), i = 1, 2, ..., n$ ακολουθεί κατανομή Poisson με παράμετρο $\lambda_1 + \lambda_2 + ... + \lambda_n$:

   $$X_1 + X_2 + ... + X_n \sim Po(\lambda_1 + \lambda_2 ... + \lambda_n) \quad (7)$$

   - Από τις σχέσεις (6) και (7) έχουμε ότι για τη μέση τιμή και τη διασπορά του αθροίσματος ανεξάρτητων τυχαίων μεταβλητών $Po(\lambda_i), i = 1, 2, ..., n$ ισχύει:

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

   $$= Var(X_1 + X_2 + ... + X_n) = \lambda_1 + \lambda_2 + ... + \lambda_n.$$

6. *Κανονική κατανομή* Μια συνεχής τυχαία μεταβλητή X λέμε ότι ακολουθεί την κανονική κατανομή με παραμέτρους μ και σ^2 (συμβολικά: $X \sim N(\mu, \sigma^2)$), αν έχει συνάρτηση πυκνότητας

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

   όπου $-\infty < \mu < +\infty$ και $0 < \sigma < +\infty$. Η συνάρτηση κατανομής της

   $$X \sim N(\mu, \sigma^2),$$

   είναι η

   $$F(x) = P(X \le x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx.$$

7. *Κεντρικό Οριακό Θεώρημα (Lindeberg-Levy)* Όταν προσθέτουμε ένα μεγάλο πλήθος ανεξάρτητων και ισόνομων τυχαίων μεταβλητών, το άθροισμα «συμπεριφέρεται» όλο και πιο πολύ σαν μια κανονική κατανομή. Αυτό είναι το περιεχόμενο του Κεντρικού Οριακού Θεωρήματος:

   Έστω

   $$X_k, k = 1, 2, ...$$

   μια ακολουθία ανεξάρτητων και ισόνομων τυχαίων μεταβλητών με μέση τιμή

   $$E(X_k) = \mu \in \mathbb{R}$$

   και διασπορά

   $$V(X_k) = \sigma^2 < +\infty, \ k = 1, 2, ...$$

   Συμβολίζουμε με S_n το άθροισμα

   $$\sum_{k=1}^{n} X_k, \ n = 1, 2, ...$$

   και θεωρούμε την ακολουθία των τυχαίων μεταβλητών (Οι τ.μ. Z_n ονομάζονται *τυποποιημένες τ.μ.*).

   $$Z_n = \frac{S_n - E(S_n)}{\sqrt{Var(S_n)}} = \frac{S_n - n\mu}{\sqrt{n\sigma^2}}, \quad n = 1, 2, ..., $$

   Τότε η ακολουθία των συναρτήσεων κατανομής

   $$F_n(z) = P(Z_n \le z), z \in \mathbb{R}, n = 1, 2, ...,$$

   συγκλίνει για κάθε στη συνάρτηση κατανομής

   $$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt$$

   της κανονικής κατανομής με μέση τιμή $\mu = 0$ και διασπορά $\sigma^2 = 1$.(Η κανονική κατανομή με μέση τιμή $\mu = 0$ και διασπορά $\sigma^2 = 1$ ονομάζεται *τυπική κανονική κατανομή*.)

   Δηλαδή, για κάθε $z \in \mathbb{R}$ ισχύει:

   $$\lim_{n \to \infty} F_n(z) = \Phi(z).$$

Με τις παραπάνω υπενθυμίσεις ως υπόβαθρο, συνεχίζουμε με την απόδειξη της Πρότασης 1. Για να εξετάσουμε την ασυμπτωτική συμπεριφορά του αθροίσματος

$$\sum_{j=0}^{n} \frac{n^j}{j!}$$

είναι απαραίτητο να αναγνωρίσουμε ότι αυτό σχετίζεται με την κατανομή Poisson. Η μορφή του αθροίσματος είναι παρόμοια με τη συνάρτηση κατανομής της Poisson με παράμετρο n, εκτός από το ότι δεν περιλαμβάνει τον όρο e^{-n}. Θεωρούμε της τυχαίες μεταβλητές

$$X_1, X_2, ..., X_n \sim Po(1).$$

Από την (7) έχουμε ότι

$$X = X_1 + X_2 + ... + X_n \sim$$

$$\sim Po(\underbrace{1 + 1 + ... + 1}_{n \ \text{όροι}}) = Po(n)$$

και από την (5),

$$P(X \le n) = \sum_{j=0}^{n} \frac{e^{-n} n^j}{j!}.$$

Από την (6) έχουμε:

$$E(X) = Var(X) = n.$$

Τυποποιώντας την τ.μ. X (αφαιρούμε από την X τη μέση τιμή της και διαιρούμε το αποτέλεσμα με την τετραγωνική ρίζα της διασποράς της) και χρησιμοποιώντας το Κεντρικό Οριακό Θεώρημα λαμβάνουμε:

$$P(X \le n) = P\left(\frac{X-n}{\sqrt{n}} \le 0\right) \to \Phi(0)$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-t^2/2} dt \quad (n \to \infty)$$

Για τη συνάρτηση κατανομής Φ, όπως και για κάθε συνάρτηση κατανομής, ισχύει

$$\Phi(+\infty) = \lim_{z \to +\infty} \Phi(z) = 1$$

ή, αλλιώς,

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t^2/2} dt = 1.$$

Επειδή η ολοκληρωτέα συνάρτηση $e^{-t^2/2}$ είναι άρτια, άμεσα παίρνουμε ότι

$$\Phi(0) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-t^2/2} dt = \frac{\Phi(+\infty)}{2} = \frac{1}{2}.$$

Απόδειξη βασισμένη σε μεθόδους της Μαθηματικής Ανάλυσης Ο τύπος Taylor με ολοκληρωτικό υπόλοιπο αποτελεί ένα ισχυρό εργαλείο για την προσεγγιστική έκφραση μιας συνάρτησης μέσω πολυωνύμου, ακολουθούμενου από έναν ολοκληρωτικό όρο που εκφράζει το σφάλμα της προσέγγισης:

Πρόταση. *Τύπος Taylor με ολοκληρωτικό υπόλοιπο Έστω $n \in \mathbb{Z}$, $n \ge 0$ και $f : [\alpha, \beta] \to \mathbb{R}$ η οποία είναι $n+1$ φορές παραγωγίσιμη στο $[\alpha, \beta]$ ώστε η $f^{(n+1)}$ να είναι συνεχής στο $[\alpha, \beta]$. Τότε, για κάθε $x \in (\alpha, \beta]$, είναι*

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(\alpha)}{k!}(x-\alpha)^k +$$

$$+ \frac{1}{n!} \int_{\alpha}^{x} (x-t)^n f^{(n+1)}(t) dt$$

Ξεκινάμε με τον τύπο του Taylor με ολοκληρωτικό υπόλοιπο της συνάρτησης e^x:

$$e^x = \sum_{k=0}^{n} \frac{x^j}{j!} + \frac{1}{n!} \int_0^x (x-t)^n e^t dt.$$

Θέτουμε στην παραπάνω σχέση $x := n$ και παίρνουμε:

$$e^n = \sum_{j=0}^{n} \frac{n^j}{j!} + \frac{1}{n!} \int_0^n (n-t)^n e^t dt$$

$$\Leftrightarrow 1 = \sum_{j=0}^{n} \frac{e^{-n} n^j}{j!} + \frac{e^{-n}}{n!} \int_0^n (n-t)^n e^t dt$$

$$\Leftrightarrow 1 = \sum_{j=0}^{n} \frac{e^{-n} n^j}{j!} + \frac{n^n e^{-n}}{n!} \int_0^n \left(1 - \frac{t}{n}\right)^n e^t dt. \quad (8)$$

Χάριν ευκολίας γράφουμε:

$$I_n = \int_0^n \left(1 - \frac{t}{n}\right)^n e^t dt$$

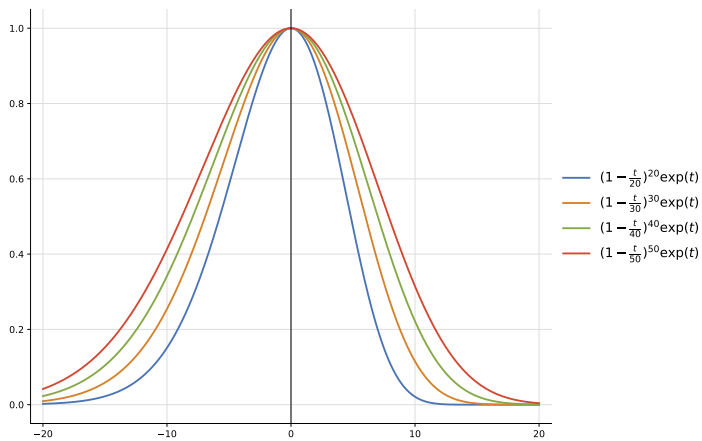
και η (8) οδηγεί στην

$$1 = \sum_{j=0}^{n} \frac{e^{-n} n^j}{j!} + \frac{n^n e^{-n}}{n!} I_n. \quad (9)$$

Παρατηρούμε ότι θα έχουμε αποδείξει την Πρόταση 1 αν αποδείξουμε ότι

$$\lim_{n \to \infty} \frac{n^n e^{-n}}{n!} I_n = \frac{1}{2}.$$

Το επόμενο μέρος της ανάλυσής μας αφιερώνεται στην απόδειξη αυτής της σχέσης. Σταθεροποιούμε προς στιγμήν το n για να μελετήσουμε γραφικά τη συμπεριφορά της ολοκληρωτέας συνάρτησης στην έκφραση του I_n.



Οι παραπάνω γραφικές παραστάσεις δείχνουν ότι το κύριο μέρος του ολοκληρώματος I_n προέρχεται από μια περιοχή κοντά και δεξιά του 0, όπου η ολοκληρωτέα συνάρτηση παίρνει τη μέγιστη τιμή της 1, ενώ στη συνέχεια φθίνει γρήγορα προς το 0.[4]

Για να κατανοήσουμε πότε η συνάρτηση αποκλίνει ουσιαστικά από τη μέγιστη τιμή της, εξετάζουμε τη συμπεριφορά της για διάφορες τιμές του t με στόχο να εντοπίσουμε εκείνες που προκαλούν αισθητή μεταβολή στις τιμές της ολοκληρωτέας συνάρτησης χωρίς όμως αυτές να μπορούν να θεωρηθούν αμελητέες.

Ορίζουμε τη συνάρτηση

$$f(t) = \left(1 - \frac{t}{n}\right)^n e^t, \quad t \in [0, n]$$

και συγκρίνουμε το μέγιστο $f(0) = 1$ με τις τιμές της f όταν η μεταβλητή t παίρνει τιμές κοντά και δεξιά του 0.

[4] Αξίζει, επίσης, να σημειωθεί και το σχήμα «καμπάνας» των γραφικών παραστάσεων.

- Αν $t = x$, όπου x ένας σταθερός αριθμός του διαστήματος $(0, n)$ τότε

$$f(x) = \left(1 - \frac{x}{n}\right)^n e^x \to e^{-x}e^x = 1,$$

όταν $n \to \infty$. Η τιμή της συνάρτησης f παραμένει κοντά στο μέγιστο 1 και δεν παρατηρείται ουσιαστική μεταβολή καθώς το $n \to \infty$.

- Αν $t = cn$, όπου c μια σταθερά στο διάστημα $(0, 1)$ τότε

$$f(cn) = \left(1 - \frac{cn}{n}\right)^n e^{cn} = (1-c)^n e^{cn}.$$

Παίρνοντας λογάριθμο και χρησιμοποιώντας την βασική ανισότητα $\ln x < x - 1$ για $0 < x \neq 1$, έχουμε:

$$n\ln(1-c) + cn = n\left(\ln(1-c) + c\right) \to -\infty,$$

όταν $n \to \infty$. Συνεπάγεται:

$$f(cn) = (1-c)^n e^{cn} \to 0$$

όταν $n \to \infty$. Σε αυτή την περίπτωση η συνάρτηση f «χάνεται» διότι οι τιμές της είναι πρακτικά 0.

Προκύπτει λοιπόν το παρακάτω ερώτημα:

> Πώς πρέπει να επιλέξουμε το t ώστε, καθώς απομακρυνόμαστε από το 0 προς τα δεξιά, οι τιμές της f να αποκλίνουν ουσιαστικά από τη μέγιστη τιμή 1, χωρίς όμως να καταλήγουν να είναι «σχεδόν» 0;

Οι παρατηρήσεις που κάναμε παραπάνω υποδεικνύουν ότι ούτε το σταθερό t αλλά ούτε και το t ανάλογο του n δίνουν το επιθυμητό αποτέλεσμα. Για να προσδιορίσουμε λοιπόν την κατάλληλη επιλογή του t, στρεφόμαστε σε μια πιο λεπτομερή ανάλυση της συμπεριφοράς της συνάρτησης. Ειδικότερα, μελετούμε τον λογάριθμο της συνάρτησης και αναπτύσσουμε τη νέα συνάρτηση σε σειρά Maclaurin. Με τον τρόπο αυτό, μπορούμε να κατανοήσουμε καλύτερα πώς εξαρτάται η συμπεριφορά της από το t και να εντοπίσουμε την κατάλληλη «ζώνη» τιμών του ώστε οι τιμές της αρχικής συνάρτησης να αποκλίνουν με ελεγχόμενο τρόπο από το 1 πλησιάζοντας στο μηδέν, χωρίς όμως να «καταρρέουν» απότομα. Ορίζουμε τη συνάρτηση

$$g(t) := \ln f(t)$$

$$= n\ln\left(1 - \frac{t}{n}\right) + t, t \in [0, n)$$

η οποία έχει σειρά Maclaurin την

$$g(t) = g(0) + g'(0)t + \frac{g''(0)}{2!}t^2 + \frac{g'''(0)}{3!}t^3 + \ldots$$

$$= -\frac{t^2}{2n} - \frac{t^3}{3n^2} - \ldots .$$

Αν επιλέξουμε το t έτσι ώστε αυτό να είναι της τάξης του $\sqrt{n}$,[5] τότε για μεγάλες τιμές του n, οι όροι από τον $-\frac{t^3}{3n^2}$ και πέρα είναι ασυμπτωτικά αμελητέοι[6] και ο επικρατέστερος όρος είναι ο πρώτος όρος $-\frac{t^2}{2n}$.

Βασιζόμενοι σε αυτή την ιδέα κάνουμε την αλλαγή μεταβλητής $t \to \lambda\sqrt{n}$ στο ολοκλήρωμα I_n. Τότε έχουμε:

$$I_n = \int_0^n \left(1 - \frac{t}{n}\right)^n e^t dt$$

$$= \sqrt{n}\int_0^{\sqrt{n}} \left(1 - \frac{\lambda}{\sqrt{n}}\right)^n e^{\lambda\sqrt{n}} d\lambda. \quad (10)$$

Από το ανάπτυγμα Maclaurin της συνάρτησης $\ln(1-x)$:

$$\ln(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \ldots, \ (-1 \le x < 1),$$

για $\lambda \in [0, \sqrt{n})$, λαμβάνουμε:

$$n\ln\left(1 - \frac{\lambda}{\sqrt{n}}\right) + \lambda\sqrt{n} \approx -\frac{\lambda^2}{2} - \frac{\lambda^3}{3\sqrt{n}} - \frac{\lambda^4}{4n} - \ldots$$

Από την τελευταία σχέση βλέπουμε ότι η ολοκληρωτέα συνάρτηση στη (10) γράφεται:

$$\left(1 - \frac{\lambda}{\sqrt{n}}\right)^n e^{\lambda\sqrt{n}} = e^{n\ln\left(1-\frac{\lambda}{\sqrt{n}}\right)+\lambda\sqrt{n}}$$

$$= e^{-\lambda^2/2 - \lambda^3/(3\sqrt{n}) - \lambda^4/(4n) - \ldots}. \quad (11)$$

Από το ανάπτυγμα της (11) παρατηρούμε ότι για λ σταθερό ισχύει:

$$\lim_{n\to\infty}\left(1 - \frac{\lambda}{\sqrt{n}}\right)^n e^{\lambda\sqrt{n}} =$$

$$\lim_{n\to\infty} e^{-\frac{\lambda^2}{2} - \frac{\lambda^3}{3\sqrt{n}} - \ldots} = e^{-\frac{\lambda^2}{2}} \quad (12)$$

και επιπλέον

$$\left(1 - \frac{\lambda}{\sqrt{n}}\right)^n e^{\lambda\sqrt{n}} \le e^{-\frac{\lambda^2}{2}}. \quad (13)$$

Θα θέλαμε, αν είναι δυνατόν, να γράψουμε στη (10):

$$\lim_{n\to\infty} \int_0^{\sqrt{n}} \left(1 - \frac{\lambda}{\sqrt{n}}\right)^n e^{\lambda\sqrt{n}} d\lambda$$

[5] Με την έννοια: $\frac{t}{\sqrt{n}} \to \lambda \in \mathbb{R}^+$, όταν $n \to \infty$.

[6] Με την έννοια: $\frac{t^3}{3n^2} \to 0$, όταν $n \to \infty$,κ.λπ..

[7] Εναλλακτικά, η εναλλαγή ορίου και ολοκληρώματος μπορεί να πραγματοποιηθεί στο πλαίσιο της θεωρίας του ολοκληρώματος Lebesgue, με τη βοήθεια του *θεωρήματος της κυριαρχημένης σύγκλισης*. Αξίζει να σημειωθεί ότι το θεώρημα αυτό απαιτεί μόνο σημειακή σύγκλιση (η οποία είναι ασθενέστερη από την ομοιόμορφη), υπό την προϋπόθεση ότι η ακολουθία συναρτήσεων (f_n) είναι φραγμένη από μία ολοκληρώσιμη (κατά Lebesgue,) συνάρτηση.

$$= \int_0^{+\infty} \lim_{n\to\infty} \left(1 - \frac{\lambda}{\sqrt{n}}\right)^n e^{\lambda\sqrt{n}} d\lambda$$
$$\overset{(12)}{=} \int_0^{+\infty} e^{-\frac{\lambda^2}{2}}\, d\lambda.$$

Το τελευταίο είναι το γνωστό (γενικευμένο) ολοκλήρωμα του Gauss το οποίο ισούται με $\sqrt{\pi/2}$. Το «κλειδί» για την εναλλαγή ορίου και ολοκληρώματος βρίσκεται στην ομοιόμορφη σύγκλιση [7].

Ορισμος. *Έστω συνάρτηση $f : A \to \mathbb{R}$ και ακολουθία συναρτήσεων (f_n) με $f_n : A \to \mathbb{R}, n \in \mathbb{N}$. Λέμε ότι η ακολουθία (f_n) συγκλίνει στην f ομοιόμορφα στο A αν για κάθε $\varepsilon > 0$ υπάρχει $n_0 \in \mathbb{N}$ τέτοιος, ώστε*

$$\sup\{|f_n(x) - f(x)| | x \in A\} < \varepsilon,$$

για κάθε $n \geq n_0$. Συμβολικά γράφουμε:

$$\lim_{n\to\infty} f_n \overset{ομ}{=} f$$

στο A.

Αποδεικνύεται ότι:

Προταση. *Αν $\lim_{n\to\infty} f_n \overset{ομ}{=} f$ στο $[\alpha, \beta]$ και κάθε συνάρτηση f_n είναι ολοκληρώσιμη στο $[\alpha, \beta]$, τότε η f είναι ολοκληρώσιμη στο $[\alpha, \beta]$ και*

$$\lim_{n\to\infty} \int_\alpha^\beta f_n(x)\, dx = \int_\alpha^\beta \lim_{n\to\infty} f_n(x)\, dx = \int_\alpha^\beta f(x)\, dx$$

Το αποτέλεσμα αυτό μας επιτρέπει να μεταφέρουμε το όριο μέσα στο ολοκλήρωμα. Ωστόσο, υπάρχει μία κρίσιμη τεχνική λεπτομέρεια που πρέπει να ληφθεί υπόψη. Η εναλλαγή ορίου και ολοκληρώματος δεν είναι εν γένει επιτρεπτή όταν το ολοκλήρωμα είναι γενικευμένο. Για να αντιμετωπίσουμε αυτό το ζήτημα, θα χρησιμοποιήσουμε το παρακάτω:

Προταση. Κριτηριο Συγκλισης *Αν το ολοκλήρωμα*

$$\int_\alpha^{+\infty} f(x)dx$$

συγκλίνει σε κάποιον πραγματικό αριθμό, τότε για κάθε $\varepsilon > 0$ υπάρχει αριθμός $A > 0$ τέτοιος, ώστε

$$\int_A^{+\infty} f(x)dx < \varepsilon.$$

Έστω $\varepsilon > 0$. Επειδή το γενικευμένο ολοκλήρωμα

$$\int_0^{+\infty} e^{-\lambda^2/2} d\lambda$$

συγκλίνει, υπάρχει αριθμός $A > 0$ τέτοιος, ώστε

$$\int_A^{+\infty} e^{-\lambda^2/2} d\lambda < \frac{\varepsilon}{4}.$$

Τότε, από τη σχέση (13), θα ισχύει, επίσης,

$$\int_A^{+\infty} \left(1 - \frac{\lambda}{\sqrt{n}}\right)^n e^{\lambda\sqrt{n}}\, d\lambda < \frac{\varepsilon}{4}.$$

Για $n > A^2$ (ώστε $\lambda/\sqrt{n} < 1$) αποδεικνύουμε ότι η ακολουθία των συναρτήσεων

$$f_n(\lambda) = \left(1 - \frac{\lambda}{\sqrt{n}}\right)^n e^{\lambda\sqrt{n}}, \lambda \in [0, A]$$

συγκλίνει ομοιόμορφα στη συνάρτηση

$$f(\lambda) = e^{-\lambda^2/2}$$

στο $[0, A]$.

Είναι

$$|f_n(\lambda) - f(\lambda)| = \left|\left(1 - \lambda/\sqrt{n}\right)^n e^{\lambda\sqrt{n}} - e^{-\lambda^2/2}\right|$$
$$= e^{-\lambda^2/2} \left|e^{n\ln\left(1-\lambda/\sqrt{n}\right)+\lambda\sqrt{n}+\lambda^2/2} - 1\right|$$
$$\overset{(11)}{=} e^{-\lambda^2/2} \left|e^{-\lambda^3/(3\sqrt{n})-\lambda^4/(4n)-\ldots} - 1\right|$$
$$= e^{-\lambda^2/2} \left(1 - e^{-\lambda^3/(3\sqrt{n})-\lambda^4/(4n)-\ldots}\right)$$
$$\leq 1 - e^{-A^3/(3\sqrt{n})-A^4/(4n)-\cdots} (e^{-\lambda^2/2} \leq 1,\ \lambda \in [0, A])$$
$$\leq \frac{A^3}{3\sqrt{n}} + \frac{A^4}{4n} \ldots . (1 - e^{-x} \leq x,\ x \in \mathbb{R})$$
$$= -\ln\left(1 - \frac{A}{\sqrt{n}}\right) - \frac{A}{\sqrt{n}} - \frac{A^2}{n} \to 0 (n \to \infty)$$
$$|f_n(\lambda) - f(\lambda)| = \left|\left(1 - \lambda/\sqrt{n}\right)^n e^{\lambda\sqrt{n}} - e^{-\lambda^2/2}\right|$$
$$= e^{-\lambda^2/2} \left|e^{n\ln\left(1-\lambda/\sqrt{n}\right)+\lambda\sqrt{n}+\lambda^2/2} - 1\right|$$

Συνεπώς,

$$\sup\{|f_n(\lambda) - f(\lambda)| | \lambda \in [0, A]\} \to 0 (n \to \infty)$$

δηλαδή

$$\lim_{n\to\infty} f_n \overset{ομ}{=} f$$

στο $[0, A]$. Από την ομοιόμορφη σύγκλιση στο $[0, A]$, προκύπτει ότι

$$\lim_{n\to\infty} \int_0^A \left(1 - \frac{\lambda}{\sqrt{n}}\right)^n e^{\lambda\sqrt{n}}\, d\lambda$$
$$= \int_0^A \lim_{n\to\infty} \left(1 - \frac{\lambda}{\sqrt{n}}\right)^n e^{\lambda\sqrt{n}}\, d\lambda$$
$$= \int_0^A e^{-\frac{\lambda^2}{2}}\, d\lambda$$

και συνεπώς, υπάρχει n_0 τέτοιος, ώστε για κάθε $n \geq n_0$ να ισχύει:

$$\left| \int_0^A \left(1 - \frac{\lambda}{\sqrt{n}}\right)^n e^{\lambda\sqrt{n}}\, d\lambda - \int_0^A e^{-\frac{\lambda^2}{2}}\, d\lambda \right| < \frac{\varepsilon}{2}.$$

Τελικά, για κάθε $n \geq n_0$:

$$\left| \int_0^\infty \left(1 - \frac{\lambda}{\sqrt{n}}\right)^n e^{\lambda\sqrt{n}}\, d\lambda - \int_0^\infty e^{-\frac{\lambda^2}{2}}\, d\lambda \right|$$

$$\leq \left| \int_0^A \left(1 - \frac{\lambda}{\sqrt{n}}\right)^n e^{\lambda\sqrt{n}}\, d\lambda - \int_0^A e^{-\frac{\lambda^2}{2}}\, d\lambda \right| +$$

$$+ \int_A^{+\infty} \left(1 - \frac{\lambda}{\sqrt{n}}\right)^n e^{\lambda\sqrt{n}}\, d\lambda + \int_A^{+\infty} e^{-\frac{\lambda^2}{2}}\, d\lambda$$

$$< \frac{\varepsilon}{2} + \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \varepsilon$$

Έτσι λοιπόν αποδείξαμε:

$$\lim_{n\to\infty} \int_0^{\sqrt{n}} \left(1 - \frac{\lambda}{\sqrt{n}}\right)^n e^{\lambda\sqrt{n}}\, d\lambda$$

$$= \int_0^{+\infty} e^{-\frac{\lambda^2}{2}}\, d\lambda = \sqrt{\frac{\pi}{2}} \quad (14)$$

. Οι σχέσεις (10) και (14) μας εξασφαλίζουν το παρακάτω αποτέλεσμα:

$$I_n \approx \sqrt{\frac{\pi n}{2}}$$

απ΄ όπου άμεσα μπορούμε να πάρουμε,χρησιμοποιώντας τον τύπο του Stirling, το προσεγγιστικό αποτέλεσμα:

$$\frac{n^n e^{-n}}{n!} I_n \approx \frac{n^n e^{-n}}{n^n e^{-n}\sqrt{2\pi n}} \cdot \sqrt{\frac{\pi n}{2}} = \frac{1}{2}.$$

Η απόδειξη έχει ολοκληρωθεί.

4 Ταχύτητα πτώσης της ουράς

4.1 Ασυμπτωτική ανάλυση της ουράς

Η γνώση ενός και μόνο χαρακτηριστικού της τ.μ. T, όπως η μέση τιμή της $E(T)$, δεν μας παρέχει αρκετές πληροφορίες για την ίδια την T. Για τον λόγο αυτό θα αναζητήσουμε έναν προσεγγιστικό (οριακό) τύπο για την ουρά[8] της διακριτής κατανομής της τ.μ. T (αριθμός ολόκληρων χαπιών που θα επιλέξουμε μέχρι να πάρουμε το πρώτο μισό χάπι). Είδαμε ότι

$$P(T \geq k) = \frac{n(n-1)(n-2)...(n-k+1)}{n^k}.$$

Η παραπάνω σχέση μπορεί να γραφεί και ως εξής:

$$P(T \geq k) = \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)...\left(1 - \frac{k-1}{n}\right)$$

$$= \prod_{j=1}^{k-1}\left(1 - \frac{j}{n}\right). \quad (15)\,.$$

Επιπλέον, υποθέτοντας ότι το k είναι «πολύ μικρότερο» από το n, εννοώντας ότι

$$\frac{k}{n} \to 0 \quad (n \to \infty)$$

και γνωρίζοντας ότι

$$e^{-x} \approx 1 - x \ \ (x \to 0)$$

είναι εύλογο να υποθέσουμε ότι

$$P(T \geq k) \approx e^{-1/n} e^{-2/n} ... e^{-(k-2)/n}$$

$$= e^{-k(k-1)/(2n)} \approx e^{-k^2/(2n)}$$

Είδαμε παραπάνω ότι η μέση τιμή της T είναι τάξεως $\sqrt{n}$ και πιο συγκεκριμένα

$$E(T) \approx C\sqrt{n}$$

με $C = \sqrt{\pi/2}$. Αν εστιάσουμε λοιπόν στις τιμές του k που είναι «κοντά» στη μέση τιμή της T δηλαδή αν υποθέσουμε:

$$\frac{k}{\sqrt{n}} \to c \in (0, +\infty)\,, \quad (n \to +\infty) \quad (16)$$

τότε μπορούμε να αποδείξουμε ότι η πιθανότητα $P(T \geq k)$ ελαττώνεται εκθετικά. Συγκεκριμένα:

Πρόταση 2. Αν $\frac{k}{\sqrt{n}} \to c \in (0, +\infty)$, όταν $n \to \infty$, τότε

$$\lim_{n\to\infty} e^{k^2/(2n)} P(T \geq k) \ = 1$$

ή, ισοδύναμα,

$$P(T \geq k) \approx e^{-k^2/(2n)}$$

Απόδειξη Παίρνοντας λογάριθμο παρατηρούμε ότι αρκεί να αποδείξουμε:

$$\lim_{n\to\infty} \left(\frac{k^2}{2n} + \ln P(T \geq k)\right) =$$

$$\lim_{n\to\infty} \left(\frac{k^2}{2n} + \ln \prod_{j=1}^{k-1}\left(1 - \frac{j}{n}\right)\right) =$$

$$\lim_{n\to\infty} \left(\frac{k^2}{2n} + \sum_{j=1}^{k-1} \ln\left(1 - \frac{j}{n}\right)\right) = 0.$$

[8] Με τον όρο ουρά διακριτής κατανομής αναφερόμαστε στην πιθανότητα η τιμή μιας τυχαίας μεταβλητής T να υπερβεί μια συγκεκριμένη τιμή k: $P(T > k) \ = \sum_{i=k+1}^{\infty} P(T = i)$.
Η ποσότητα αυτή εκφράζει την πιθανότητα να παρατηρηθεί μια τιμή «μεγάλη-ακραία» σε σχέση με το k.

Για να μελετήσουμε την οριακή συμπεριφορά του αθροίσματος

$$\sum_{j=1}^{k-1} \ln\left(1-\frac{j}{n}\right)$$

η ιδέα είναι να το αντικαταστήσουμε με το ολοκλήρωμα μιας κατάλληλης συνάρτησης, ώστε να μπορέσουμε να χειριστούμε το πρόβλημα πιο εύκολα[9]. Το άθροισμα μπορεί να ξαναγραφεί ως $n\sum_{j=1}^{k-1}\frac{1}{n}\ln\left(1-\frac{j}{n}\right)$ που θυμίζει άθροισμα Darboux για το ολοκλήρωμα της συνάρτησης $\ln(1-x)$ στο διάστημα $[0,\frac{k}{n}]$. Για $k<n$, το κάτω άθροισμα L και το άνω άθροισμα U:

$$L=\frac{1}{n}\cdot 0+\frac{1}{n}\cdot\ln\left(1-\frac{1}{n}\right)+\ldots+\frac{1}{n}\cdot\ln\left(1-\frac{k-1}{n}\right)$$

$$=\sum_{j=1}^{k-1}\frac{1}{n}\ln\left(1-\frac{j}{n}\right)$$

$$U=\frac{1}{n}\cdot\ln\left(1-\frac{1}{n}\right)+\frac{1}{n}\cdot\ln\left(1-\frac{2}{n}\right)+\ldots+\frac{1}{n}\cdot\ln\left(1-\frac{k}{n}\right)$$

$$=\sum_{j=1}^{k}\frac{1}{n}\ln\left(1-\frac{j}{n}\right)$$

προσεγγίζουν από κάτω και από πάνω αντίστοιχα το ολοκλήρωμα $\int_0^{\frac{k}{n}}\ln(1-x)dx$:

$$L\le\int_0^{\frac{k}{n}}\ln(1-x)dx\le U$$

ή

$$\sum_{j=1}^{k-1}\ln\left(1-\frac{j}{n}\right)\le n\int_0^{\frac{k}{n}}\ln(1-x)dx\le\sum_{j=1}^{k}\ln\left(1-\frac{j}{n}\right).\quad(17)$$

Από την (17) παίρνουμε:

$$\sum_{j=1}^{k-1}\ln\left(1-\frac{j}{n}\right)\le n\int_0^{\frac{k}{n}}\ln(1-x)dx$$

και

$$\sum_{j=1}^{k-1}\ln\left(1-\frac{j}{n}\right)\ge n\int_0^{\frac{k}{n}}\ln(1-x)dx-\ln\left(1-\frac{k}{n}\right)$$

Προσθέτοντας σε κάθε μέλος των παραπάνω ανισοτήτων την ποσότητα $\frac{k^2}{2n}$ προκύπτουν οι παρακάτω σχέσεις:

$$\sum_{j=1}^{k-1}\ln\left(1-\frac{j}{n}\right)+\frac{k^2}{2n}$$

$$\le n\int_0^{\frac{k}{n}}\ln(1-x)dx+\frac{k^2}{2n}\quad(18)$$

και

$$\sum_{j=1}^{k-1}\ln\left(1-\frac{j}{n}\right)+\frac{k^2}{2n}$$

$$\ge\int_0^{\frac{k}{n}}\ln(1-x)dx+\frac{k^2}{2n}-\ln\left(1-\frac{k}{n}\right)\quad(19)$$

Ένας στοιχειώδης υπολογισμός δίνει:

$$n\int_0^{\frac{k}{n}}\ln(1-x)dx$$

$$=-k-(n-k)\ln\left(1-\frac{k}{n}\right)$$

και προσθέτοντας σε κάθε μέλος τον όρο $\frac{k^2}{2n}$, λαμβάνουμε την ακόλουθη έκφραση:

$$n\int_0^{\frac{k}{n}}\ln(1-x)dx+\frac{k^2}{2n}$$

$$=\frac{k^2}{2n}-k-(n-k)\ln\left(1-\frac{k}{n}\right)$$

$$=\frac{k^2}{2n}\left(1-\frac{\frac{k}{n}+\left(1-\frac{k}{n}\right)\ln\left(1-\frac{k}{n}\right)}{\frac{k^2}{2n^2}}\right).\quad(20)$$

Αν στο κλάσμα της παρένθεσης θέσουμε $x:=\frac{k}{n}$ και θεωρήσουμε τη μεταβλητή x συνεχή, τότε, με διπλή εφαρμογή του κανόνα de l'Hôpital, βρίσκουμε:

$$\lim_{x\to 0}\frac{x+(1-x)\ln(1-x)}{x^2}=\frac{1}{2}\quad(21).$$

Από τις (16), (20) και (21) συμπεραίνουμε ότι

$$\lim_{n\to\infty}\left(n\int_0^{\frac{k}{n}}\ln(1-x)dx+\frac{k^2}{2n}\right)$$

$$=\frac{c^2}{2}\left(1-2\cdot\frac{1}{2}\right)=0.\quad(22)$$

Τέλος, παίρνοντας $n\to\infty$ στα δεξιά μέλη των σχέσεων (18) και (19) και κάνοντας χρήση των (16) και (22)

[9] Μπορούμε να κινηθούμε και αλλιώς. Από το ανάπτυγμα Maclaurin: αν θέσουμε $x:=\frac{j}{n}$, παίρνουμε:

$$\ln\left(1-\frac{j}{n}\right)=-\frac{j}{n}-\frac{j^2}{2n^2}-\frac{j^3}{3n^3}-\ldots.$$

Αθροίζοντας προκύπτει ότι:

$$\sum_{j=0}^{k-1}\ln\left(1-\frac{j}{n}\right)$$

$$=-\frac{1}{n}\sum_{j=1}^{k-1}j-\frac{1}{2n^2}\sum_{j=1}^{k-1}j^2-\frac{1}{3n^3}\sum_{j=1}^{k-1}j^3-\ldots.$$

Το άθροισμα

$$\sum_{j=1}^{k-1}j^p,\ p=0,1,2,3,\ldots$$

είναι πολυώνυμο του k με βαθμό $p+1$. Με δεδομένη την (16), οι όροι από τον

$$-\frac{1}{2n^2}\sum_{j=1}^{k-1}j^2$$

και πέρα είναι ασυμπτωτικά αμελητέοι και συνεπώς

$$\sum_{j=0}^{k-1}\ln\left(1-\frac{j}{n}\right)\approx-\sum_{j=1}^{k-1}\frac{j}{n}$$

$$=-\frac{k(k-1)}{2n}\approx-\frac{k^2}{2n}.$$

βρίσκουμε ότι αυτά έχουν όριο το 0. Τελικά, από το κριτήριο παρεμβολής προκύπτει:

$$\lim_{n\to\infty}\left(\frac{k^2}{2n}+\sum_{j=1}^{k-1}\ln\left(1-\frac{j}{n}\right)\right)=0,$$

όπως θέλαμε να αποδείξουμε. ■

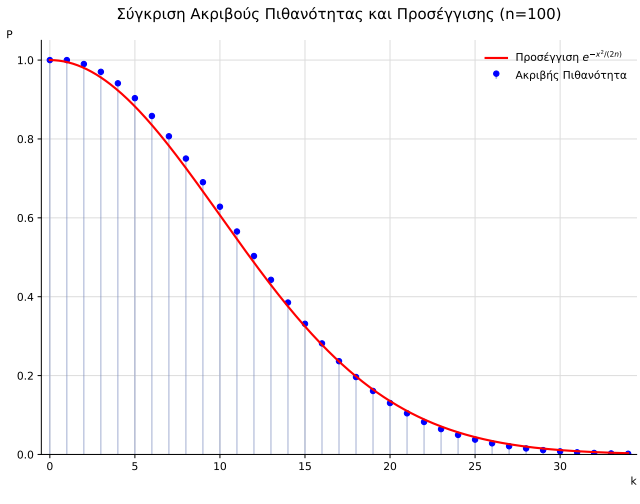
Εύλογα τώρα αν θεωρήσουμε τη μεταβλητή k συνεχή η οποία μπορεί να πάρει τιμές στο διάστημα $(0,+\infty)$ βλέπουμε ότι μια κομψή και συνεχής προσέγγιση είναι η παρακάτω:

$$P(T>x)\approx e^{-\frac{x^2}{2n}}\ \ (n\to\infty).$$

[10].

4.2 Γραφική αξιολόγηση του προσεγγιστικού αποτελέσματος

Ένα οριακό αποτέλεσμα για την ουρά, όπως αυτό της Πρότασης 2, δεν είναι από μόνο του ιδιαίτερα χρήσιμο – ιδίως όταν θέλουμε να το αξιοποιήσουμε σε στατιστικές εφαρμογές – καθώς η Στατιστική βασίζεται σε πεπερασμένα δείγματα. Για τον λόγο αυτό, θα μελετήσουμε τη μέγιστη απόκλιση της προσέγγισης της ουράς από την πραγματική τιμή της. Ακολουθεί μια αρχική οπτική εκτίμηση της συμπεριφοράς του προσεγγιστικού μας μοντέλου για $n = 100$.



[10]Το αποτέλεσμα:

$$P(T>x)\approx e^{-x^2/(2n)}$$

μας παρέχει μία νέα προσέγγιση για την οριακή συμπεριφορά της μέσης τιμής. Η προσεγγιστική μέση τιμή $E(T)$ δίνεται από το ολοκλήρωμα:

$$E(T)=\int_0^{+\infty}P(T>x)dx\approx\int_0^{+\infty}e^{-x^2/(2n)}dx.$$

Με την αλλαγή μεταβλητής $u = x/\sqrt{2n}$ και τη χρήση παραγοντικής ολοκλήρωσης, καταλήγουμε στο αποτέλεσμα:

$$E(T)\approx\sqrt{n\pi/2}.$$

Η προσεγγιστική κατανομή

$$P(T\le x)=1-e^{-x^2/(2n)},$$

Είναι εμφανής η συμφωνία μεταξύ της πραγματικής πιθανότητας:

$$P(T\ge k)=\frac{n!}{(n-k)!n^k}$$

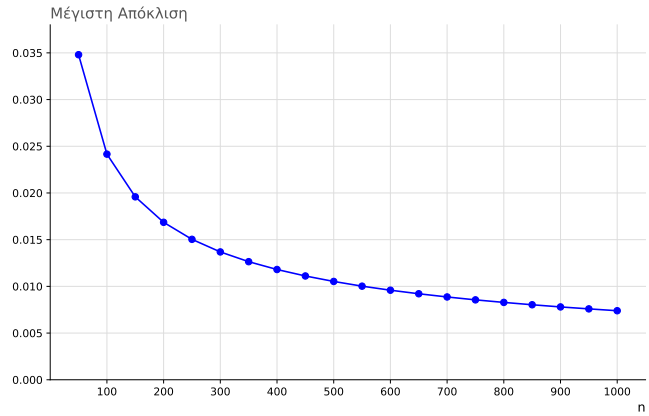
και της προσέγγισής της:

$$P(T>x)\approx e^{-x^2/(2n)}.$$

Στο παρακάτω γράφημα μπορούμε να δούμε και τη μέγιστη απόκλιση

$$A(n):=\max_{1\le k\le n}\left|P(T\ge k)-e^{-k^2/(2n)}\right|$$

μεταξύ της πραγματικής πιθανότητας και της προσέγγισής της για τιμές του n μέχρι το 1000.



Αν και το γράφημα υποδηλώνει φθίνοντα ρυθμό της $A(n)$, παρ' όλα αυτά δεν επιτρέπει με ακρίβεια να διακρίνουμε το ρυθμό αυτής της μείωσης δηλαδή αν η μείωση είναι εκθετική ή ακολουθεί νόμο δύναμης (power law)[11] ή τίποτα από τα δύο. Μετασχηματίζοντας τα δεδομένα του προηγούμενου σχήματος σε λογαριθμική κλίμακα και για τους δύο άξονες παίρνουμε το παρακάτω γράφημα:

δεν είναι καθόλου τυχαία κατανομή αλλά πρόκειται για την κατανομή Rayleigh με μέση τιμή $\sqrt{n\pi/2}$ και παράμετρο κλίμακας $\sigma^2 = n$. Για περισσότερες πληροφορίες, ανατρέξτε στη διεύθυνση: `https://en.wikipedia.org/wiki/Rayleigh_distribution`

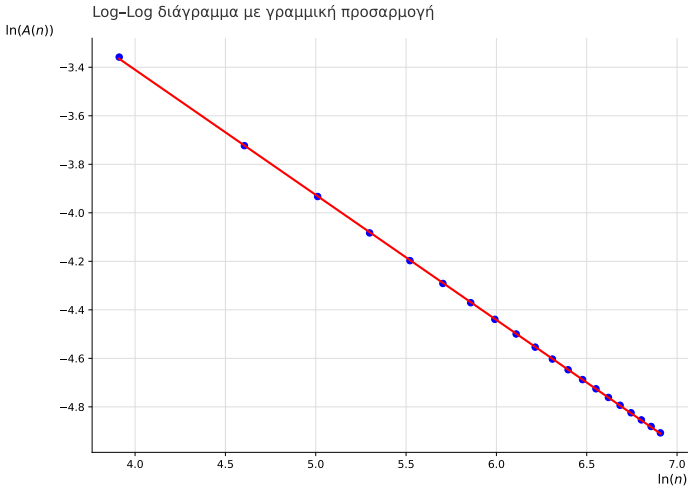
[11]Λέμε ότι η ποσότητα y φθίνει εκθετικά ως προς x, αν

$$y=A\rho^x+\beta$$

με $A > 0, \beta \in \mathbb{R}$ και $0 < \rho < 1$. Λέμε ότι η ποσότητα y φθίνει με νόμο δύναμης ως προς x, αν

$$y=Ax^{-\rho}$$

με $A > 0, \rho > 0$.



Παρατηρούμε ότι τα σημεία στο log – log διάγραμμα[12] σχηματίζουν ευθεία. Αυτό αποτελεί ισχυρή ένδειξη ότι η $A(n)$ φθίνει σύμφωνα με νόμο δύναμης. Ζητώντας από το Mathematica να μας δώσει την εξίσωση της ευθείας, αυτό μας επιστρέφει την παρακάτω:

$$y = -0,51x - 1,34.$$

Συμπεραίνουμε ότι

$$\ln\left(A(n)\right) \simeq -0,51\ln(n) - 1,34$$

$$\Leftrightarrow A(n) \simeq \frac{e^{-1,34}}{n^{0,51}},$$

αποτέλεσμα το οποίο μας δίνει μια ισχυρή ένδειξη ότι η μέγιστη απόκλιση φθίνει σύμφωνα με νόμο δύναμης.[13]

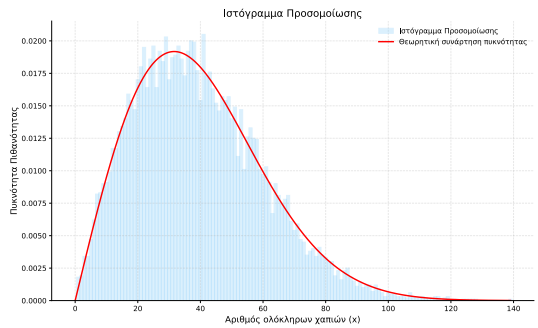
5 Προσομοίωση Monte Carlo

Σε αυτό το μέρος της εργασίας θα ασχοληθούμε με την προσομοίωση του προβλήματος στον υπολογιστή. Η ταχύτατη ανάπτυξη των υπολογιστών τις τελευταίες δεκαετίες και η διάδοση των υπολογιστικών πακέτων που αφορούν τα μαθηματικά έχουν προσφέρει ένα καινούριο εργαλείο στην έρευνα των μαθηματικών.

Η τεχνική προσομοίωσης που θα ακολουθήσουμε έχει την ονομασία *Monte Carlo Simulation* και βασίζεται στη παραγωγή (ψευδο)τυχαίων αριθμών από κάποια κατανομή. Το όνομα προέρχεται από το καζίνο του Μόντε Κάρλο στο Μονακό, όπου ο βασικός δημιουργός της μεθόδου, ο μαθηματικός Στάνισλάφ Ούλαμ [14], εμπνεύστηκε από τις συνήθειες του θείου του στον τζόγο. Η επιτυχία της προσομοίωσης ως υπολογιστικό εργαλείο εδράζεται στον *Νόμο των Μεγάλων Αριθμών* [15].

Αρχικά επιλέγουμε τον αριθμό m των επαναλήψεων του πειράματος που επιθυμούμε, για παράδειγμα[16] $m = 10000$ και τον αριθμό n των ολόκληρων χαπιών με τα οποία ξεκινάμε. Δημιουργούμε έπειτα μια κενή λίστα την οποία θα τροφοδοτούμε μετά από κάθε επανάληψη με το αποτέλεσμα του τρέχοντος πειράματος (αριθμός ολόκληρων χαπιών που αφαιρούμε μέχρι να διαλέξουμε το πρώτο μισό χάπι). Μετά το πέρας της διαδικασίας η λίστα θα περιέχει m το πλήθος στοιχεία και θα ζητήσουμε από το Mathematica να μας επιστρέψει το ιστόγραμμα σχετικών συχνοτήτων από κοινού με την προσεγγιστική συνάρτηση πυκνότητας που βρήκαμε. Την τελική λίστα μπορούμε να τη χρησιμοποιήσουμε για να κάνουμε περαιτέρω στατιστική ανάλυση του συγκεκριμένου προβλήματος.

«Τρέχοντας» τον αλγόριθμο για $n = 1000$ το Mathematica μας επιστρέφει στην έξοδο το παρακάτω ιστόγραμμα.



Η συνεχής κόκκινη καμπύλη αναπαριστά την προσεγγιστική συνάρτηση πυκνότητας f της τ.μ. T:

$$f(x) = \frac{d}{dx}P(T \le x)$$

$$= \frac{x}{n}e^{-x^2/(2n)}, \quad x > 0.$$

Παρατηρούμε ότι η προσαρμογή είναι ικανοποιητική, καθώς η μορφή της καμπύλης ταιριάζει καλά με το σχήμα του ιστογράμματος.

[12] Είναι ένα διάγραμμα στο οποίο και οι δύο άξονες, $x'x$ και $y'y$ είναι σε λογαριθμική κλίμακα, συνήθως με λογάριθμους φυσικής βάσης ή δεκαδικής βάσης, ανάλογα με την εφαρμογή. Ως βάση του λογαρίθμου το Mathematica χρησιμοποιεί το e. Είναι δηλαδή ένα διάγραμμα του τύπου $(\ln x, \ln y)$. Αν η απεικόνιση στο log – log διάγραμμα είναι ευθεία γραμμή, τότε

$$\ln y = A\ln x + \beta, (A, \beta \in \mathbb{R})$$

απ' όπου προκύπτει (*νόμος δύναμης*):

$$y = e^{\beta}x^{A}.$$

[13] Σε αυτό το σημείο υποψιαζόμαστε ότι, οριακά, ο εκθέτης του ν είναι $-0,5$, δηλαδή ότι η μέγιστη απόκλιση είναι της μορφής $A(n) \approx \frac{A}{\sqrt{n}}$, όπου A κάποια θετική σταθερά.

[14] `https://en.wikipedia.org/wiki/Stanis%C5%82aw_Ulam`

[15] Υπάρχουν δύο μορφές του Νόμου των Μεγάλων Αριθμών ανάλογα με τον τρόπο σύγκλισης, ο *Ασθενής* και ο *Ισχυρός*. Αμφότερες αναφέρουν ότι ο δειγματικός μέσος όρος συγκλίνει στον θεωρητικό μέσο όρο.

[16] Από εδώ και στο εξής το m θα είναι σταθερό και ίσο με 10000.

5.1 Ανάλυση Περιγραφικών Μέτρων και Σύγκριση με το Προσεγγιστικό Μοντέλο

Μπορούμε να εξαγάγουμε βασικά στατιστικά μέτρα από τα δεδομένα της προσομοίωσης, όπως:

- Μέση τιμή (μ): Ο αναμενόμενος αριθμός ολόκληρων χαπιών πριν το πρώτο μισό.
- Τυπική απόκλιση (sd): Το μέτρο της διασποράς των τιμών γύρω από τη μέση τιμή.
- Εκατοστημόρια P_{25}, P_{50}, P_{75}: Τα σημεία που διαχωρίζουν τα δεδομένα σε 25%, 50% και 75% της κατανομής.

Τα παραπάνω μέτρα μπορούμε να τα συγκρίνουμε με τα μέτρα που προκύπτουν από το προσεγγιστικό μας μοντέλο. Στον παρακάτω πίνακα παρουσιάζονται οι τιμές που προέκυψαν από την προσομοίωση Monte Carlo και οι αντίστοιχες προβλέψεις του προσεγγιστικού μοντέλου για διάφορα n.

ΠΙΝΑΚΑΣ 1

$n = 100$	μ	sd	P_{25}	P_{50}	P_{75}
Προσομοίωση	12,1493	6,2788	7	12	16
Μοντέλο	12,5331	6,5513	7,5852	11,7741	16,6510
$n = 500$	μ	sd	P_{25}	P_{50}	P_{75}
Προσομοίωση	27,6429	14,2791	17	26	37
Μοντέλο	28,0249	14,6461	16,9611	26,3276	37,2329
$n = 1000$	μ	sd	P_{25}	P_{50}	P_{75}
Προσομοίωση	39,0289	19,9669	24	37	51
Μοντέλο	39,6332	20,7172	23,9867	37,2329	52,6553
$n = 2000$	μ	sd	P_{25}	P_{50}	P_{75}
Προσομοίωση	55,7861	28,8569	34	52	74
Μοντέλο	56,0499	29,2985	33,9223	52,6553	74,4659

Παρατηρούμε μια σχετικά καλή εφαρμογή του μοντέλου μας στα δεδομένα της προσομοίωσης. Εντοπίζεται όμως μια μικρή μεροληψία (πρώτη στήλη).Ο μέσος όρος που προκύπτει από το μοντέλο είναι ελαφρώς μεγαλύτερος από τον μέσο όρο της προσομοίωσης. Η διαφορά αυτή θα εξηγηθεί παρακάτω (υποσημείωση 19).

5.2 Quantile-Quantile plot

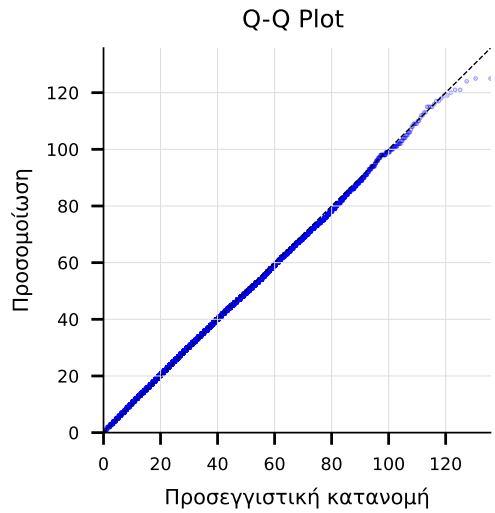
Το *Quantile-Quantile plot* (Q-Q plot) είναι ένα διαγνωστικό εργαλείο που χρησιμοποιείται στη Στατιστική για τη σύγκριση της κατανομής των δεδομένων (εδώ είναι τα δεδομένα της προσομοίωσης) με μια θεωρητική κατανομή (εδώ είναι η προσεγγιστική κατανομή). Ένα Q-Q plot συγκρίνει τα ποσοστημόρια των δεδομένων με τα αντίστοιχα ποσοστημόρια της θεωρητικής κατανομής.

Για την κατασκευή του γραφήματος, χρησιμοποιείται η συνάρτηση ποσοστημορίων $Q(p)$, η οποία για κάθε τιμή της παραμέτρου $p \in [0,1]$ επιστρέφει την τιμή x για την οποία ισχύει[17] $P(T \leq x) = p$. Αρχικά υπολογίζονται τα ποσοστημόρια τόσο για την κατανομή του προσεγγιστικού μοντέλου όσο και για τα δεδομένα της προσομοίωσης, για ένα μεγάλο πλήθος τιμών της παραμέτρου p. Στον οριζόντιο άξονα τοποθετούνται τα ποσοστημόρια της προσεγγιστικής κατανομής, ενώ στον κατακόρυφο άξονα τα αντίστοιχα ποσοστημόρια των δεδομένων που προέκυψαν από την προσομοίωση.

Ως παράδειγμα, ας εξετάσουμε την περίπτωση όπου ο αρχικός αριθμός χαπιών είναι $n = 1000$ και μελετούμε τη διάμεσο ($p = 0,5$). Για την προσεγγιστική κατανομή, η τιμή της συνάρτησης ποσοστημορίων είναι $37,2329$, ενώ για τα δεδομένα της προσομοίωσης η αντίστοιχη τιμή είναι 37 (Πίνακας 1).

Στο Q-Q plot, η σύγκριση αυτή απεικονίζεται ως σημείο με συντεταγμένες $(37,2329,\ 37)$. Αν τα δεδομένα της προσομοίωσης ακολουθούν την προσεγγιστική κατανομή, τότε τα σημεία του διαγράμματος θα βρίσκονται κοντά στη διαγώνιο $y = x$.

Ζητώντας από το Mathematica να μας δημιουργήσει το γράφημα με συνολικό αριθμό χαπιών $n = 1000$ μας επιστρέφει το παρακάτω.



Η συγκέντρωση των σημείων κοντά στην διαγώνιο υποδεικνύει ότι η προσεγγιστική κατανομή αποδίδει πολύ καλά την κατανομή των δεδομένων που παρήχθησαν από την προσομοίωση.

6 Από τη θεωρία στην πράξη!

Τα μαθηματικά δεν είναι μόνο μια μορφή τέχνης, αλλά και ένας θεμελιώδης τομέας της ανθρώπινης δραστηριότητας και του πολιτισμού, με αμέτρητες εφαρμογές στις ε-

[17] Στην περίπτωσή μας η προσεγγιστική συνάρτηση κατανομής είναι συνεχής και γνησίως αύξουσα οπότε η τιμή x είναι μοναδική.

$$P(T \leq x) = 1 - e^{-x^2/(2n)}$$

πιστήμες, την τεχνολογία και την καθημερινή ζωή. Για να αναδείξουμε αυτή τη χρηστική τους πλευρά, θα επιχειρήσουμε στη συνέχεια να προσδώσουμε στο πρόβλημά μας μια πιο πρακτική διάσταση. Όποιος έχει ασχοληθεί με τις πιθανότητες και τη συνδυαστική έχει σίγουρα συναντήσει προβλήματα που αφορούν τράπουλες, μπάλες, κουτιά και άλλα φαινομενικά απλοϊκά αντικείμενα. Κάποιος ίσως αναρωτηθεί αν οι πιθανοθεωρητικοί δεν έχουν άλλα πιο σύνθετα και ενδιαφέροντα θέματα να διερευνήσουν. Ωστόσο, ο βασικός λόγος που τέτοιες ασκήσεις είναι τόσο διαδεδομένες είναι ότι πολλά ρεαλιστικά προβλήματα μπορούν να μοντελοποιηθούν με ανάλογο τρόπο, περιγράφοντας τη λειτουργία τους μέσω αυτών των αντικειμένων. Για παράδειγμα, η διαδικασία τοποθέτησης σφαιρών σε κουτιά μπορεί να αποτυπώσει με ακρίβεια συγκεκριμένους μηχανισμούς κατανομής. Αυτή η προσέγγιση προσφέρει ένα σημαντικό πλεονέκτημα: μας επιτρέπει να εστιάσουμε στον πυρήνα του προβλήματος, αποφεύγοντας άσχετες ή περιττές λεπτομέρειες, διευκολύνοντας έτσι τη βαθύτερη κατανόησή του.

6.1 Η σύνδεση με γνωστό πρόβλημα

Σε αυτό το σημείο θα ακολουθήσουμε μια προσέγγιση αντίστοιχη με την προηγούμενη, επιχειρώντας να αναδιατυπώσουμε το αρχικό πρόβλημα με τη βοήθεια ενός γνωστού μοντέλου: τις σφαίρες και τα κουτιά. Αρχικά, παρατηρούμε ότι κάθε φορά που παίρνουμε ένα χάπι από το μπουκάλι, μέχρι να εμφανιστεί για πρώτη φορά ένα μισό χάπι, ο συνολικός αριθμός των χαπιών παραμένει σταθερός: αφαιρούμε ένα ολόκληρο και επιστρέφουμε ένα μισό. Αυτόν τον σταθερό αριθμό χαπιών τον αντιστοιχίζουμε με έναν σταθερό αριθμό κουτιών. Κάθε φορά που τραβάμε ένα ολόκληρο χάπι, θα θεωρούμε ότι είναι σαν να ρίχνουμε μια σφαίρα σε ένα κουτί, επιλεγμένο τυχαία από τα διαθέσιμα. Όσο εξακολουθούμε να τραβάμε ολόκληρα χάπια, θα υποθέτουμε ότι κάθε σφαίρα καταλήγει σε διαφορετικό κουτί. Η εμφάνιση του πρώτου μισού χαπιού αντιστοιχεί στη στιγμή που για πρώτη φορά μια σφαίρα πέφτει σε κουτί που είναι ήδη κατειλημμένο από σφαίρα. Μπορούμε λοιπόν να δώσουμε στο αρχικό μας πρόβλημα την εξής ισοδύναμη διατύπωση:

Διαθέτουμε αρχικά $n+1$ [18] όμοια σφαιρίδια και n όμοια κουτιά. Αρχίζουμε να τοποθετούμε στην τύχη μια προς μια τις σφαίρες στα κουτιά. Η διαδικασία τερματίζεται τη στιγμή που θα τοποθετηθεί σφαίρα σε κουτί που περιέχει ήδη μια σφαίρα. Ποιος είναι ο μέσος αριθμός σφαιρών που θα έχουμε τοποθετήσει, εξαιρουμένης της τελευταίας σφαίρας, μέχρι να συμβεί το ενδεχόμενο αυτό;

Με μια πρώτη ματιά φαίνεται ότι δεν έχουμε κέρδος από μια τέτοια αναδιατύπωση του προβλήματος. Εντούτοις το κέρδος μας είναι πολύ μεγάλο! Αρκεί να αναρωτηθούμε αν υπάρχει κάποιο γνωστό πρόβλημα που διαθέτει τον ίδιο μηχανισμό. Σε αυτό το σημείο είναι απαραίτητο ο αναγνώστης να έχει μια σχετική τριβή με τη θεωρία πιθανοτήτων και τα προβλήματα που αυτή εξετάζει.

6.2 Το Πρόβλημα των Γενεθλίων

Η παραπάνω αναδιατύπωση του προβλήματος παραπέμπει στον μηχανισμό που κρύβεται πίσω από το γνωστό πρόβλημα πιθανοτήτων, το Πρόβλημα των Γενεθλίων. Το κλασικό πρόβλημα των Γενεθλίων έχει την εξής διατύπωση:

Σε ένα σύνολο n ατόμων ποια είναι η πιθανότητα τουλάχιστον δύο από αυτά να έχουν την ίδια ημερομηνία γενεθλίων, υποθέτοντας ομοιόμορφη κατανομή γενεθλίων και σύνολο 365 ημερών το χρόνο;

Αποδεικνύεται ότι η πιθανότητα αυτή ξεπερνάει το 50% για $n = 23$ ενώ με μόλις 60 άτομα αγγίζει το $99,4\%$!
Εδώ έχουμε να κάνουμε με την παρακάτω τροποποίηση του κλασικού προβλήματος.
Φανταστείτε ότι σε ένα δωμάτιο εισέρχονται, ένα-ένα κάθε φορά, $n+1$ άτομα ($n \geq 365$) και για κάθε άτομο ξεχωριστά καταγράφουμε την ημέρα των γενεθλίων του. Πόσα άτομα, κατά μέσο όρο, με διαφορετικές ανα δύο ημέρες γενεθλίων θα καταγράψουμε, μέχρι να υπάρξουν δύο άτομα με την ίδια ημέρα γενεθλίων;

Η απάντηση έχει ήδη δοθεί από την ανάλυση του προβλήματος με τα χάπια. Κατά μέσο όρο θα καταγράψουμε

$$\sqrt{\frac{\pi n}{2}}$$

άτομα. Το πρόβλημα των γενεθλίων είναι αρκετά μελετημένο στη διεθνή βιβλιογραφία και η αναδιατύπωση του αρχικού προβλήματος που πετύχαμε μας προσφέρει γόνιμο έδαφος για την αναζήτηση εφαρμογών του γνωστού προβλήματος. Το άθροισμα της σχέσης (2) μπορεί να το συναντήσει κάποιος στη βιβλιογραφία με την ονομασία «Ramanujan's Q-Function»[19]. Στην πεντάτομη έκδοση *The Art of Computer Programming* του θεωρητικού της πληροφορικής και παγκοσμίως γνωστού Donald Ervin Knuth και συγκεκριμένα στον πρώτο τόμο *Fundamental Algorithms* (σελ.116-120, βλέπε [6]) γίνεται αναφορά στο άθροισμα της σχέσης (2). Εκεί επιτυγχάνεται μια έξυπνη ομολογουμένως ασυμπτωτική προσέγγιση του αθροίσματος μέσω της σύζευξής του με ένα φαινομενικά όμοιο άθροισμα. Επιπλέον, μια ανάλυση του αθροίσματος

[18] Υπάρχει πάντα το ακραίο ενδεχόμενο να εξαντλήσουμε το σύνολο των ολόκληρων χαπιών μέχρι να πετύχουμε το πρώτο μισό χάπι. Σε αυτή την περίπτωση το πείραμα θα εκτελεστεί $n+1$ φορές . Εκείνη τη στιγμή η $(n+1)$–οστή σφαίρα θα εισέρχεται υποχρεωτικά – από την αρχή της περιστεροφωλιάς– σε κουτί το οποίο περιέχει ήδη μια σφαίρα και η διαδικασία θα ολοκληρωθεί.

[19] Srinivasa Ramanujan, ο γνωστός Ινδός μαθηματικός ο οποίος στις εργασίες του ασχολήθηκε, εκτός των άλλων, και με το άθροισμα της σχέσης (2) του παρόντος άρθρου. Απέδειξε ότι, για μεγάλες τιμές του n, ισχύει $E(T) = \sqrt{n\pi/2} - 1/3+$ (όροι ασυμπτωτικά αμελητέοι). Η σταθερά $1/3$ είναι η μεροληψία που παρατηρήσαμε στον Πίνακα 1.

με προχωρημένα εργαλεία της Μιγαδικής Ανάλυσης συναντάμε στο άρθρο των Philippe Frajoletetal (βλέπε [5]).

6.3 Εφαρμογές στην Πληροφορική και την Κρυπτογραφία

Η σημασία του αθροίσματος της σχέσης (2) στην πληροφορική και συγκεκριμένα στην κρυπτογραφία γίνεται φανερή με τις συναρτήσεις κατακερματισμού (hash functions). Μια συνάρτηση κατακερματισμού μετατρέπει δεδομένα αυθαίρετου μήκους σε δεδομένα συγκεκριμένου και σταθερού πάντα μήκους. Σε αυστηρή μαθηματική γλώσσα θα λέγαμε ότι η h είναι μια συνάρτηση από ένα αυθαίρετο σύνολο στοιχείων A, πεπερασμένο ή ενδεχομένως και άπειρο, σε ένα σύνολο B με πεπερασμένο πλήθος στοιχείων. Η χρησιμότητα τους έγκειται στην ψηφιακή ασφάλεια συναλλαγών όπου η πιστοποίηση αυθεντικότητας ενός ψηφιακού μηνύματος μεταξύ αποστολέα και παραλήπτη πραγματοποιείται μέσω αυτών των συναρτήσεων. Τυχόν αλλαγές στο μήνυμα οι οποίες έχουν συμβεί οδηγούν σε διαφορετικές απεικονίσεις της συνάρτησης και συνεπώς σε μη αυθεντικότητα του κειμένου. Επιπλέον, βρίσκουν εφαρμογές στην σύγχρονη τάση των κρυπτονομισμάτων.

Η χαρακτηριστική ιδιότητα αυτών των συναρτήσεων είναι αυτή της μη *αντιστρεψιμότητας* (irreversibility) και της *ανθεκτικότητας* σε συγκρούσεις (collision-free property). Για την πρώτη ιδιότητα μπορούμε να πούμε ότι αν γνωρίζουμε την τιμή $h(\alpha)$ ($\alpha \in A$) είναι πρακτικά αδύνατο να μπορέσουμε να μάθουμε ποιο είναι το α. Αυτός είναι και ο λόγος για τον οποίο τα προσωπικά στοιχεία μας (κωδικοί, (usernames κ.λπ.) αποθηκεύονται στους servers αφού έχουν περάσει πρώτα από το στάδιο του κατακερματισμού. Έτσι αν κάποιος αποκτήσει πρόσβαση στη βάση δεδομένων θα έχει στη διάθεσή του μόνο τις τιμές της συνάρτησης χωρίς να γνωρίζει όμως από που προήλθαν αυτές οι τιμές. Όσον αφορά τη δεύτερη ιδιότητα, αυτή της ανθεκτικότητας σε συγκρούσεις, θα μπορούσαμε να αναφέρουμε με κάποια χαλαρότητα ότι είναι επιθυμητό μια συνάρτηση κατακερματισμού να είναι «σχεδόν ένα προς ένα» με την εξής έννοια: Η πιθανότητα[20] να μπορούν να βρεθούν δύο διαφορετικά $A, A' \in A$ με $h(A) = h(A')$ (σύγκρουση) είναι αμελητέα. Ας δούμε όμως ένα απλουστευμένο παράδειγμα.

Παράδειγμα Μια συνάρτηση κατακερματισμού h έχει τη δυνατότητα να παράγει

$$2^{128} \simeq 3,4 \cdot 10^{38}$$

διακριτές τιμές[21]. Ας υποθέσουμε ότι τροφοδοτούμε την h με 20 ψήφιους κωδικούς (passwords) τους οποίους μπορούμε να δημιουργήσουμε από ένα σύνολο 94 διαφορετικών χαρακτήρων. Είναι προφανές ότι το πλήθος των διαφορετικών κωδικών που μπορούμε να κατασκευάσουμε είναι

$$94^{20} \simeq 2,9 \cdot 10^{39}$$

δηλαδή μεγαλύτερο από το πλήθος των δυνατών τιμών που μπορεί να αποδώσει η h. Από την αρχή της περιστεροφωλιάς αναμένουμε ότι θα υπάρχουν κωδικοί που θα αντιστοιχίζονται στην ίδια τιμή. Ξεκινώντας τη διαδικασία δημιουργίας τυχαίων κωδικών και κατακερματισμού τους θα χρειαστεί να παράγουμε περίπου

$$\sqrt{\frac{\pi}{2} \cdot 3,4 \cdot 10^{38}} \simeq 2.3 \cdot 10^{19}$$

κωδικούς κατά μέσο όρο μέχρι να πετύχουμε την πρώτη σύγκρουση[22].

6.4 Ο στοχαστικός αλγόριθμος παραγοντοποίησης Rho

Το πρόβλημα των γενεθλίων βρίσκει επιπλέον εφαρμογή στην παραγοντοποίηση μεγάλων ακεραίων. Ο αλγόριθμος Rho αποτελεί μια αποτελεσματική και έξυπνη μέθοδο για την εύρεση μη τετριμμένων διαιρετών ενός μεγάλου και σύνθετου ακεραίου N. Η πιο απλοϊκή προσέγγιση που μπορούμε να ακολουθήσουμε για την παραγοντοποίηση ενός μεγάλου ακέραιου Nείναι να ξεκινήσουμε δοκιμάζοντας τους αριθμούς $2, 3, 4, ...$ μέχρι να βρούμε κάποιον αριθμό ο οποίος διαιρεί τον N. Ας σημειωθεί ότι, από γνωστή πρόταση της Θεωρίας Αριθμών, ο ακέραιος N θα έχει τουλάχιστον έναν πρώτο διαιρέτη $p \leq \sqrt{N}$. Οπότε γνωρίζουμε εκ των προτέρων ότι η διαδικασία απαιτεί το πολύ $\sqrt{N}$ ελέγχους. Ο έλεγχος μπορεί να γίνει υπολογίζοντας τον $\text{ΜΚΔ}(d, N)$ (μέγιστο κοινό διαιρέτη), όπου $d \leq \sqrt{N}$. Όταν εντοπίσουμε κάποιον ακέραιο $d \leq \sqrt{N}$ με $\text{ΜΚΔ}(d, N) > 1$, τότε ο d θα είναι ένας μη τετριμμένος

[20] Πρέπει να σημειωθεί ότι η συνάρτηση κατακερματισμού είναι ντετερμινιστική, επομένως δεν υπάρχει τυχαιότητα στη λειτουργία της. Η αναφορά στην πιθανότητα προκύπτει από τον τρόπο με τον οποίο αναλύουμε θεωρητικά αυτές τις συναρτήσεις, θεωρώντας τη συμπεριφορά της συνάρτησης ως τυχαία ώστε να εκτιμήσουμε πόσο σπάνιο είναι το φαινόμενο μιας σύγκρουσης.

[21] Η συνάρτηση MD5 έχει αυτή την ιδιότητα. Περισσότερες πληροφορίες μπορούν να βρεθούν στον σύνδεσμο `https://en.wikipedia.org/wiki/MD5`. Στην ηλεκτρονική διεύθυνση `https://www.md5hashgenerator.com/` μπορεί να παράγει οποιοσδήποτε θελήσει τις δικές του τιμές με τη συνάρτηση MD5. Το όνομα του συντάκτη του παρόντος άρθρου αντιστοιχεί στην τιμή $89e2839987f6297e79e080231df996f8$!

[22] Η εύρεση συγκρούσεων αποτελεί σημαντικό γεγονός στην επιστήμη της κρυπτογραφίας λόγω της σπανιότητάς τους. Τον Αύγουστο του 2004 Κινέζοι ερευνητές κατάφεραν να κατασκευάσουν αλγόριθμο (χωρίς όμως να παρουσιάσουν σημαντικές λεπτομέρειες της αρχιτεκτονικής του) με τον οποίο παρήγαγαν συγκρούσεις στη συνάρτηση $MD5$. Τον Μάρτιο του 2005 ο Τσέχος κρυπτογράφος Vlastimil Klima παρουσίασε τη δικιά του μέθοδο παραγωγής συγκρούσεων. Η εύρεση αλγορίθμου παραγωγής συγκρούσεων σε μια συνάρτηση κατακερματισμού καθιστά την ίδια τη συνάρτηση επισφαλή και οδηγεί σε αντικατάστασή της από άλλη η οποία θα μετατρέπει τις συγκρούσεις ακόμα πιο απίθανες. Ο ενδιαφερόμενος αναγνώστης μπορεί να αναζητήσει περισσότερες πληροφορίες στη σελίδα του Vlastimil Klima: `https://cryptography.hyperlink.cz/`

διαιρέτης του N και επομένως $N = d \cdot \frac{N}{d}$. Ας σημειωθεί ότι ο πρώτος ακέραιος d που θα συναντήσουμε με την ιδιότητα $\text{ΜΚΔ}(d, N) > 1$, αν ξεκινήσουμε τον έλεγχο από τον αριθμό 2 και αυξάνουμε κάθε φορά το d κατά μια μονάδα, θα είναι αναγκαστικά πρώτος διαιρέτης του N(αν ήταν σύνθετος, θα είχε μικρότερο πρώτο παράγοντα, τον οποίο θα είχαμε ήδη συναντήσει). Μπορούμε όμως να τα καταφέρουμε καλύτερα μέσω του αλγορίθμου Rho[23]. Η κεντρική και ιδιοφυής ιδέα πίσω από τον συγκεκριμένο στοχαστικό αλγόριθμο έχει ως εξής: Έστω ότι ο N έχει πρώτο διαιρέτη τον p με τον pνα είναι «μικρός» σε σύγκριση με τον N. Ας φανταστούμε ότι μπορούμε να επιλέξουμε τυχαία μερικούς αριθμούς $x_1, x_2, x_3, ... x_m$ από το σύνολο $\mathbb{Z}_N = \{0, 1, ..., N-1\}$. Στο εξής εργαζόμαστε υπό τις ακόλουθες παραδοχές:

- Όλοι αυτοί οι αριθμοί είναι διαφορετικοί μεταξύ τους [24], δηλαδή για κάθε $1 \leq i, j \leq m$ με $i \neq j$ ισχύει $x_i \neq x_j \,(\text{mod}N)$.
- Κάποιοι δύο από αυτούς αφήνουν το ίδιο υπόλοιπο όταν διαιρεθούν με το p, δηλαδή υπάρχουν δείκτες $1 \leq i < j \leq m$ τέτοιοι ώστε $x_i = x_j \,(\text{mod}p)$ [25].

Τότε θα ισχύει $p|x_i - x_j$ και αφού $p|N$, προκύπτει ότι $p|\text{ΜΚΔ}(x_i - x_j, N)$. Επιπλέον, με δεδομένο ότι $0 < |x_i - x_j| < N$, θα ισχύει:

$$1 < p \leq \text{ΜΚΔ}(x_i - x_j, N) < N. \quad (24)$$

Άρα ο $\text{ΜΚΔ}(x_i - x_j, N)$ είναι ένας μη τετριμμένος διαιρέτης του N και έχουμε πετύχει τον στόχο μας! Ο p μας είναι προφανώς άγνωστος όμως, τα παραπάνω μας λένε ότι, αν υπολογίσουμε[26] τον $\text{ΜΚΔ}(x_\beta - x_A, N)$ για κάθε ζεύγος (α, β), με $1 \leq \alpha < \beta \leq m$, τότε ενδέχεται να βρούμε έναν μη τετριμμένο παράγοντα του N.

Το ερώτημα που προκύπτει σε αυτό το σημείο είναι:

Πόσους, κατά μέσο όρο, τυχαίους αριθμούς πρέπει να πάρουμε από το σύνολο $\mathbb{Z}_N$, ώστε να βρούμε δύο από αυτούς που να είναι ισοϋπόλοιποι mod$p$*;*

Η απάντηση είναι $\sqrt{\frac{\pi p}{2}} \simeq 1,25\sqrt{p}$ τυχαίους αριθμούς. Αυτό είναι ουσιαστικά μια αναδιατύπωση του προβλήματος του παρόντος άρθρου. Δεδομένου ότι $p \leq \sqrt{N}$ θα χρειαστούμε τελικά, κατά μέσο όρο, το πολύ $1,25\sqrt[4]{N}$ τυχαίους αριθμούς.

Για την παραγωγή τυχαίων αριθμών από το σύνολο $\mathbb{Z}_n$ κατασκευάζουμε αναδρομικά μια ακολουθία $x_0, x_1, x_2, ...,$ όπου τα x_i παράγονται με βάση τη σχέση: $x_{i+1} = f(x_i) \,(\text{mod}N)$, με συνήθη επιλογή την $f(x) = x^2 + c$, όπου c ακέραιος με $1 \leq c < N - 2$. Η αρχική τιμή x_0μπορεί να είναι κάποια ακέραια τιμή του διαστήματος $[0, N)$. Φυσικά η ακολουθία αυτή παράγεται με αιτιοκρατικό τρόπο αλλά παρουσιάζει συμπεριφορά ψευδοτυχαίας ακολουθίας και «ξεγελά» τους στατιστικούς ελέγχους τυχαιότητας που διαθέτουμε για να αξιολογήσουμε αν ένα συγκεκριμένο σύνολο αριθμών έχει προέλθει από μια πραγματικά τυχαία διαδικασία [27]. Ωστόσο, δεν είναι πρακτικό να ελέγχουμε όλα τα ζεύγη αριθμών που παράγονται από αυτή την αναδρομική σχέση για το αν ικανοποιούν την (24). Αν έχουμε ήδη δημιουργήσει μια λίστα με m το πλήθος αριθμούς, η σύγκριση απαιτεί το πολύ

$$\binom{m}{2} \approx \frac{m^2}{2}$$

ελέγχους, διόλου ευκαταφρόνητο μέγεθος για μεγάλες τιμές του m. Αντ΄ αυτού ο αλγόριθμος χρησιμοποιεί έναν άλλον πολύ πιο αποδοτικό αλγόριθμο ο οποίος ονομάζεται «Αλγόριθμος του Floyd» και εφευρέθηκε από τον Robert W. Floyd το 1969. Ονομάζεται, επίσης, «Αλγόριθμος της χελώνας και του λαγού» [28] και βρίσκει εφαρμογή στην

[23] Η μέθοδος ανακαλύφθηκε το 1975 από τον John Pollard (βλέπε [8]). Η συγκεκριμένη μέθοδος πέτυχε να παραγοντοποιήσει τον όγδοο αριθμό Fermat:

$F_8 =$	$2^{2^8} + 1 =$	11579	20892	37316	19542
35709	85008	68790	78532	69984	66564
05640	39457	58400	79131	29639	937

Ο αλγόριθμος Rho μας έδωσε την παρακάτω παραγοντοποίηση:

$$F_8 = \alpha\beta$$

όπου:

	$\alpha = 1238926361552897$			
$\beta =$	93461	63971	53579	77769
16355	81996	06896	58405	12375
41638	18858	02803	21	

[24] Αν το N είναι μεγάλο σε σύγκριση με το m, τότε αυτό είναι πολύ πιθανό να συμβεί.

[25] Τα δυνατά υπόλοιπα mod$p$ συγκροτούν το σύνολο $\{0, 1, ..., p-1\}$. Αν ο p είναι πολύ μικρότερος από τον N, τότε είναι αρκετά πιθανό ότι κάποια στιγμή θα προκύψουν δύο τιμές x_i και x_j που είναι ίσες mod$p$ αλλά όχι ίσες mod$N$.

[26] Για δύο θετικούς ακεραίους μ και ν, ο μέγιστος κοινός διαιρέτης τους μπορεί να υπολογιστεί με τον αλγόριθμο του Ευκλείδη ο οποίος βασίζεται στην επαναληπτική εφαρμογή της σχέσης

$$\text{ΜΚΔ}(\mu, \nu) = \text{ΜΚΔ}(\mu \bmod \nu, \nu).$$

Ο αλγόριθμος ολοκληρώνεται σε αριθμό βημάτων που είναι ανάλογος του πλήθους των ψηφίων του μικρότερου από τους δύο αριθμούς. Στην περίπτωσή μας, επειδή $|x_i - x_j| < N$, ο αλγόριθμος ολοκληρώνεται σε αριθμό βημάτων που είναι ανάλογος του πλήθους των ψηφίων του αριθμού $|x_i - x_j|$.

[27] Είναι γεγονός ότι αναδρομικές σχέσεις παρόμοιες με του αλγορίθμου Rho χρησιμοποιούνται για την παραγωγή ψευδοτυχαίων αριθμών του διαστήματος $(0, 1)$. Αυτές έχουν τη μορφή $x_{i+1} = f(x_i) \,(\text{mod}M)$ όπου η συνάρτηση f είναι κάποιο κατάλληλο πολυώνυμο και ο ακέραιος είναι αρκετά μεγάλος ώστε να επιτευχθεί η μέγιστη δυνατή περίοδος της ακολουθίας των τιμών x_i.

[28] Περισσότερες πληροφορίες παρέχονται στον σύνδεσμο: `https://en.wikipedia.org/wiki/Cycle_detection#Floyd's_tortoise_and_hare`

ανίχνευση κύκλων σε επαναληπτικές διαδικασίες. Ας φανταστούμε ότι υπάρχει ένα κλειστό μονοπάτι στο οποίο κινούνται μια χελώνα και ένας λαγός, με τον λαγό να κινείται με διπλάσια ταχύτητα από τη χελώνα. Ο αλγόριθμος του Floyd εγγυάται ότι η χελώνα και ο λαγός θα συναντηθούν κάποια στιγμή στην ίδια θέση. Αυτή την κυκλική δομή που εμφανίζεται στην ακολουθία των τιμών αξιοποιεί ο αλγόριθμος του Floyd, προκειμένου να καταστήσει τον αλγόριθμο Rho υπολογιστικά πιο αποδοτικό.

Η λειτουργία του αλγορίθμου περιγράφεται μέσω δύο ακολουθιών, οι οποίες ορίζονται ως εξής:

- $(x_k)_{k\geq 0}$: Είναι η ακολουθία «Χελώνα». Προχωράει ένα βήμα κάθε φορά και οι όροι της παράγονται από την αναδρομική σχέση

$$x_{k+1} = f(x_k) \,(\mathrm{mod} N)$$

με κάποια αρχική τιμή $0 \leq x_0 < N$.

- $(y_k)_{k\geq 0}$: Είναι η ακολουθία «Λαγός». Προχωράει δύο βήματα κάθε φορά και οι όροι της παράγονται από την αναδρομική σχέση

$$y_{k+1} = f\left(f(y_k)\right) (\mathrm{mod} N)$$

με $y_0 = x_0$.

Είναι φανερό ότι $y_k = x_{2k}$. Ο αλγόριθμος διασφαλίζει ότι θα βρεθούν δύο αριθμοί x_k και y_k με $y_k = x_{2k} = x_k \,(\mathrm{mod} p)$. Άρα αρκεί να ελέγχουμε κάθε φορά αν ισχύει $1 < \mathrm{ΜΚΔ}(x_{2k} - x_k, N) < N$.[29] Η αποδοτικότητα του αλγορίθμου έγκειται στο γεγονός ότι δεν χρειάζεται να κρατήσει ιστορικό τιμών αλλά καταναλώνει σταθερή μνήμη (μόνο για τους δύο δείκτες), γεγονός που τον καθιστά ιδανικό για χρήση σε περιβάλλοντα περιορισμένων πόρων. Ας δούμε όμως ένα αριθμητικό παράδειγμα για το πώς δουλεύει ο αλγόριθμος Rho.

Παράδειγμα Θα παραγοντοποιήσουμε με τη χρήση του Rho τον ακέραιο $N = 1189$ παίρνοντας ως αρχική τιμή τον ακέραιο $x_0 = 2$ και $f(x) = x^2 + 1$. Στον παρακάτω πίνακα φαίνονται οι τιμές που μας επιστρέφει ο αλγόριθμος.

Χελώνα	Λαγός	
x_k	$y_k (= x_{2k})$	$\mathrm{ΜΚΔ}(y_k - x_k, N)$
5	26	1
26	565	1
677	124	1
565	456	1
574	21	1
124	369	1
1109	166	41

Προκύπτει λοιπόν ότι $N = 41 \cdot \frac{1189}{41} = 41 \cdot 29$.

[29] Αν δεν προκύψει μη τετριμμένος διαιρέτης, τότε μπορούμε να επανεκκινήσουμε τη διαδικασία είτε με διαφορετική αρχική τιμή x_0 είτε με διαφορετική ακέραια σταθερά c στο πολυώνυμο $f(x) = x^2 + c$.

7 Βιβλιογραφία

ΞΕΝΟΓΛΩΣΣΗ

1 R.P. Brent, *An improved Monte Carlo Factorization Algorithm*, BIT 20: 176-184 (1980)

2 N.G. De Brujin, *Asymptotic Methods in Analysis*, North Holland Publishing Co., (1958)

3 W. Feller, *An introduction to Probability Theory and It's Applications, Vol.1*, Third Edition, John Wiley & Sons. (1968)

4 W. Feller, *An introduction to Probability Theory and It's Applications, Vol.1*, Second Edition, John Wiley & Sons. (1971)

5 P. Frajolet, P. J. Grabner, P. Kirschenhofer, H. Prodinger, *On Ramanujan's Q-function*, Journal of Computational and Applied Mathematics 58 103-116 (1995).

6 D. E. Knuth, *The Art of Computer Programming, Vol.1* Addison-Wesley, Third Edition (1997).

7 D. P. Kroese, T. Taimre, Z. I. Botev, *Handbook of Monte Carlo Methods*, First Edition (2011).

8 J. M. Pollard, *A Monte Carlo method for factorization*, BIT Numerical Mathematics 15 (3): 331-334 (1975).

9 Joel Spencer with Laura Florescu, *Asymptopia*, American Mathematical Society (Student mathematical library, volume 71) (2014)

ΕΛΛΗΝΙΚΗ

10 P. G. Hoel, S. C. Port, C. J. Stone, *Εισαγωγή στη Θεωρία Πιθανοτήτων*, Πανεπιστημιακές Εκδόσεις Κρήτης (2015)

11 Αριστείδης Δούμας, *Στοιχεία Ασυμπτωτικής Ανάλυσης – Ολοκληρώματα, Αθροίσματα, Ειδικές Συναρτήσεις*, Ανοικτές Ακαδημαϊκές Εκδόσεις Καλλίπος(2022)

12 Ιωάννης Κοντογιάννης, Σταύρος Τουμπής, *Στοιχεία Πιθανοτήτων – Με Εφαρμογές στη Στατιστική και την Πληροφορική*, Ανοικτές Ακαδημαϊκές Εκδόσεις Καλλίπος(2015)

13 Μιχαήλ Μπούτσικας, *Μέθοδοι Προσομοίωσης και Υπολογιστικές Στατιστικές Τεχνικές* (Πανεπιστημιακές σημειώσεις) (2003)
(Στη διεύθυνση: `https://oldsite.unipi.gr/faculty/mbouts/teaching.htm`)

14 Μιχαήλ Παπαδημητράκης, *Πραγματικές Συναρτήσεις μιας Μεταβλητής*, Ανοικτές Ακαδημαϊκές Εκδόσεις Καλλίπος(2015)

15 Χαράλαμπος Α. Χαραλαμπίδης, *Θεωρία πιθανοτήτων και εφαρμογές, Τεύχος 1*, Εκδόσεις Συμμετρία (2000)

16 Χαράλαμπος Α. Χαραλαμπίδης, *Θεωρία πιθανοτήτων και εφαρμογές*, Τεύχος 2, Εκδόσεις Συμμετρία (1999)

Ευχαριστίες

Θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες προς τους κριτές για τα ουσιαστικά σχόλια και τις εποικοδομητικές παρατηρήσεις τους οι οποίες συνέβαλαν καθοριστικά στη βελτίωση του παρόντος άρθρου. Είμαι, επίσης, ιδιαίτερα ευγνώμων προς τη Συντακτική Επιτροπή για την προθυμία της να φιλοξενήσει το άρθρο μου στο περιοδικό Εκθέτης.