ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΙΘΑΝΟΤΗΤΕΣ ΚΑΙ ΣΤΑΤΙΣΤΙΚΗ

Σημειώσεις Διαλέξεων

ΠΕΙΡΑΙΑΣ 2021

Περιεχόμενα

Μέρος Ι

Πιθανότητες

Κεφάλαιο 1

Εισαγωγή στις πιθανότητες

1.1 Πείραμα τύχης

Η θεωρία πιθανοτήτων μελετά τα τυχαία φαινόμενα τα οποία, στα πλαίσια της θεωρίας, ονομάζονται και **πειράματα τύχης**. Κάθε εκτέλεση ενός πειράματος τύχης ονομάζεται **δοκιμή** (trial) και το αποτέλεσμά της δεν είναι γνωστό εκ των προτέρων.

Κατηγορίες πειραμάτων:

- Πείραμα (πχ. ρίχνω ζάρια, διαλέγω αριθμό στο $\{1, 2, \ldots, n\}$).
- Φαινόμενο (πχ. θερμοκρασία, χρόνος παράδοσης email, χρόνος τηλεφωνικής συνδιάλεξης).

Βασική υπόθεση: Δυνατότητα επανάληψης (δοκιμής) κάτω από τις ίδιες συνθήκες (αν δεν υπάρχει, την υποθέτουμε).

- **Δειγματικός χώρος ή δειγματοχώρος** Ω (sample space): Το σύνολο **όλων** των δυνατών αποτελεσμάτων ενός πειράματος τύχης.

  Ζάρι $\rightarrow \Omega = \{1, 2, \ldots, 6\}$,

  Χρόνος τηλεφωνικής συνδιάλεξης $\rightarrow \Omega = [0, +\infty)$

- Ο δειγματικός χώρος Ω μπορεί να είναι:
  - **πεπερασμένος** (π.χ. $\{1, 2, \ldots, 6\}$),
  - **αριθμήσιμος** (ισοδύναμος του $\mathbb{N}$, π.χ. $\mathbb{Z}$, $\mathbb{N}^2$), ή
  - **υπεραριθμήσιμος** (π.χ $\mathbb{R}$, $[0, 1]$).

- **Ενδεχόμενο** (event): Ένα υποσύνολο $A \subseteq \Omega$ του δειγματικού χώρου Ω.
- **Απλό ή στοιχειώδες ενδεχόμενο**: Ένα ενδεχόμενο της μορφής $\{\omega\}$, όπου $\omega \in \Omega$.
- **Πραγματοποίηση ενός ενδεχομένου**: Όταν το πείραμα οδηγεί σε αποτέλεσμα (απλό ενδεχόμενο) που περιέχεται στο ενδεχόμενο (**ευνοϊκό ενδεχόμενο**).
- Δύο ενδεχόμενα $A, B \subseteq \Omega$ ονομάζονται **ξένα** αν και μόνο αν $A \cap B = \varnothing$.

1.1.1 Τι είναι πιθανότητα;

Υπάρχουν πολλές προσεγγίσεις της έννοιας "πιθανότητα". Ο κλασικός τρόπος προσέγγισής της είναι να υπολογίσουμε τον λόγο του αριθμού των ευνοϊκών περιπτώσεων ενός ενδεχομένου προς τον αριθμό των δυνατών περιπτώσεων.

Ορισμός (Ορισμός κλασικής πιθανότητας (De Moivre 1711, Laplace 1812)).
Αν A είναι ενδεχόμενο του δειγματικού χώρου Ω, τότε

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\textit{πλήθος ευνοϊκών για το } A \textit{ αποτελεσμάτων}}{\textit{πλήθος δυνατών αποτελεσμάτων}}$$

Μειονεκτήματα κλασικού ορισμού:

- Απαιτεί πεπερασμένο πλήθος στοιχείων δειγματοχώρου.
- Απαιτεί ισοπίθανα δυνατά αποτελέσματα.

Πλεονεκτήματα κλασικού ορισμού:

- Χρήσιμος στην πράξη.
- Η πιθανότητα υπολογίζεται εκ των προτέρων.

Η στατιστική προσέγγιση στηρίζεται στην έννοια της συχνότητας πραγματοποίησης ενός ενδεχομένου σε μια μεγάλη σειρά δοκιμών.

Ορισμός (Στατιστικός ορισμός πιθανότητας (όριο σχετικής συχνότητας)).
Αν $N_n(A)$ είναι το πλήθος πραγματοποιήσεων του ενδεχομένου A του δειγματικού χώρου Ω, μετά από n δοκιμές, τότε

$$P(A) = \lim_{n\to\infty} \frac{N_n(A)}{n}$$

Πολλές επαναλήψεις, μεγάλα μεγέθη $\Rightarrow$ σύγκλιση σε μια τιμή (πιθανότητα).

Μειονεκτήματα στατιστικού ορισμού:

- Πρέπει να γίνουν πολλές δοκιμές.
- Η πιθανότητα δεν είναι γνωστή εκ των προτέρων.

Πλεονεκτήματα στατιστικού ορισμού:

- Χρήσιμος στην πράξη.

Παράδειγμα 1.1.1. *Ένα δοχείο περιέχει 3 κόκκινες και 3 μαύρες μπάλες. Επιλέγουμε 3 μπάλες στην τύχη (δηλαδή όλα τα δυνατά αποτελέσματα είναι ισοπίθανα). Ποια είναι η πιθανότητα να διαλέξαμε ακριβώς 2 κόκκινες μπάλες.*

Λύση. Αρχικά, θα πρέπει να προσδιορίσουμε και να αναπαραστήσουμε τον δειγματικό χώρο Ω. Έστω $R = \{1, 2, 3\}$ το σύνολο των κόκκινων μπαλών και έστω $B = \{4, 5, 6\}$ το σύνολο των μαύρων μπαλών. Κάθε στοιχειώδες ενδεχόμενο είναι μια (μη διατεταγμένη) τριάδα από στοιχεία του $R \cup B$, άρα

$$\Omega = \{\{x, y, z\} : x, y, z \in R \cup B\} = \{S \subseteq R \cup B : |S| = 3\}.$$

Το ενδεχόμενο να διαλέξουμε ακριβώς 2 κόκκινες μπάλες είναι το

$$A = \{\{x, y, z\} : x, y \in R, z \in B\} = \{\{x, y\} \cup \{z\} : \{x, y\} \subseteq R, z \in B\}.$$

Με απλές γνώσεις Συνδυαστικής, βρίσκουμε ότι

$$|\Omega| = \binom{6}{3} = 20 \quad \text{και} \quad |A| = \binom{3}{2}\binom{3}{1} = 3 \cdot 3 = 9,$$

επομένως, σύμφωνα με τον κλασικό ορισμό, είναι $P(A) = \dfrac{|A|}{|\Omega|} = \dfrac{9}{20} = 0.45$.

(Παρατήρηση: Μπορούμε να καταλήξουμε στο ίδιο αποτέλεσμα θεωρώντας τις τριάδες διατεταγμένες. Τότε όμως, θα πρέπει να αναπαραστήσουμε τα Ω, A διαφορετικά.)

Σύμφωνα με τον στατιστικό ορισμό, μπορούμε να εκτελέσουμε πολλές φορές το πείραμα της επιλογής μιας τριάδας και να πάρουμε μια προσέγγιση για την πιθανότητα $P(A)$. Ο επόμενος κώδικας υλοποιεί αυτή την εκτέλεση.

```
import random
import time
r, b, k, d = 3, 3, 3, 2 #number of red, black, selected and desired red balls
N = 10000 #number of trials
R = [i for i in range(1,r+1)] #list of red balls
B = [i for i in range(r+1,r+b+1)] #list of black balls
RB = R+B #concatenate
succ = 0 # number of successful trials. Success = exactly d red balls selected
start = time.time()
for i in range(N): #perform N trials
    s = random.sample(RB, k) #s: random sample of k distinct balls
    red = 0
    for j in range(len(s)): #scan sample to count red balls
        if s[j] <= r: red += 1 #red ball found
    if red == d: succ += 1
end = time.time()
print("Success rate:", succ/N) #print success ratio
print("Simulation time (seconds):", end-start)
```

Output:

```
Success rate: 0.4529
Simulation time (seconds): 0.13151121139526367
```

□

1.2 Αξιωματικός ορισμός πιθανότητας

Ορισμός (Αξιωματικός ορισμός πιθανότητας (Kolmogorov 1930)).
Έστω συνάρτηση $P : \mathcal{P}(\Omega) \to \mathbb{R}$. Η P ονομάζεται ***συνάρτηση πιθανότητας*** *ή* ***μέτρο πιθανότητας*** *αν και μόνο αν*

- $P(A) \geq 0$, *για κάθε* $A \subseteq \Omega$.
- $P(\Omega) = 1$.
- $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$, *για κάθε (αριθμήσιμη) οικογένεια* $(A_i)_{i \in \mathbb{N}^*}$ *ξένων ανά δύο υποσυνόλων του* Ω.

Η τριάδα $(\Omega, \mathcal{P}(\Omega), P)$ *ονομάζεται* ***χώρος πιθανότητας****, ή* ***μοντέλο πιθανότητας****.*

Πλεονεκτήματα αξιωματικού ορισμού

- Λειτουργεί και για δειγματικούς χώρους με άπειρο (αριθμήσιμο ή υπεραριθμήσιμο) πλήθος ενδεχομένων (π.χ. ύψος ανθρώπου).
- Λειτουργεί και για μη ισοπίθανα ενδεχόμενα (π.χ. κάλπικο νόμισμα).
- Ταυτίζεται με τον κλασσικό ορισμό, όταν ο δειγματοχώρος είναι πεπερασμένος και τα ενδεχόμενα ισοπίθανα.
- Κάθε ενδεχόμενο προκύπτει ως ένωση στοιχειωδών ενδεχομένων. Στους πεπερασμένους ή αριθμήσιμους χώρους, αν γνωρίζουμε την πιθανότητα για τα στοιχειώδη ενδεχόμενα τότε γνωρίζουμε και την πιθανότητα για κάθε ενδεχόμενο.

Παράδειγμα 1.2.1. *Ρίχνουμε 2 ζάρια. Να βρεθεί η πιθανότητα το άθροισμα των ενδείξεων να είναι* 3 *ή* 7*.*

Λύση. Ο δειγματικός χώρος Ω του πειράματος είναι

$$\Omega = \{(x, y) \in \mathbb{N}^* \times \mathbb{N}^* : 1 \leq x, y \leq 6\},$$

οπότε $|\Omega| = 6 \cdot 6 = 36$.

Αν A είναι το ζητούμενο ενδεχόμενο, τότε $A = A_3 \cup A_7$, όπου A_3 το ενδεχόμενο οι ενδείξεις των δύο ζαριών να αθροίζουν στο 3 και A_7 το ενδεχόμενο οι ενδείξεις των δύο ζαριών να αθροίζουν στο 7. Έχουμε ότι

$$A_3 = \{(1,2),(2,1)\} \qquad \text{και} \qquad A_7 = \{(1,6),(6,1),(2,5),(5,2),(4,3),(3,4)\}.$$

Δεδομένου ότι τα στοιχειώδη ενδεχόμενα είναι ισοπίθανα προκύπτει ότι

$$P(A_3) = \frac{2}{36} \text{ και } P(A_7) = \frac{6}{36}$$

οπότε, αφού τα A_3, A_7 είναι ξένα μεταξύ τους

$$P(A) = P(A_3) + P(A_7) = \frac{2}{36} + \frac{6}{36} = \frac{8}{36} = \frac{2}{9} = 0.222. \qquad \square$$

Παράδειγμα 1.2.2. *Έστω ότι ένα ζάρι δεν είναι αμερόληπτο αλλά έχει τις παρακάτω πιθανότητες εμφάνισης των στοιχείων του:*

$$P(1) = 1/6 \quad P(2) = 1/4 \quad P(3) = 1/3$$
$$P(4) = 1/8 \quad P(5) = 1/16 \quad P(6) = 1/16$$

Να βρεθεί η πιθανότητα το ζάρι να φέρει άρτιο αριθμό.

Λύση. Το ενδεχόμενο A το ζάρι να φέρει άρτιο αριθμό είναι η ένωση των (ξένων) ενδεχομένων:

$$A_2 : \text{το ζάρι φέρνει 2}, \quad A_4 : \text{το ζάρι φέρνει 4}, \quad A_6 : \text{το ζάρι φέρνει 6}$$

Επομένως η ζητούμενη πιθανότητα ισούται με

$$P(A) = P(A_2) + P(A_4) + P(A_6) = 1/4 + 1/8 + 1/16 = 7/16.$$

Παρατήρηση. Με τον κλασικό ορισμό της πιθανότητας δεν θα μπορούσαμε να βρούμε την απάντηση! $\square$

Πρόταση 1.1 (Ιδιότητες αξιωματικού ορισμού πιθανότητας).
Αν $A, B, C \subseteq \Omega$, τότε

i) $P(\overline{A}) = 1 - P(A)$

ii) $P(\varnothing) = 0$

iii) $P(A) \leq 1$

iv) $A \subseteq B \Rightarrow P(A) \leq P(B)$

v) $P(A \setminus B) = P(A \cap \overline{B}) = P(A) - P(A \cap B)$

vi) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

vii) $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$

Κατηγορίες ενδεχομένων

- Δύο ενδεχόμενα A, B ονομάζονται **ασυμβίβαστα** ή **αμοιβαίως αποκλειόμενα** αν η πραγματοποίηση του ενός αποκλείει την πραγματοποίηση του άλλου.

  Ισοδύναμα, $A, B \subseteq \Omega$ **ασυμβίβαστα** $\Rightarrow A \cap B = \emptyset$.

- Το ενδεχόμενο A ονομάζεται **υποενδεχόμενο** (αντ. **υπερενδεχόμενο**) του B αν η πραγματοποίηση (αντ. μη πραγματοποίηση) του A συνεπάγεται την πραγματοποίηση (αντ. μη πραγματοποίηση) του B.

  Ισοδύναμα, A υποενδεχόμενο του $B \Rightarrow A \subseteq B$ (αντ. A υπερενδεχόμενο του $B \Rightarrow B \subseteq A$).

- Τα ενδεχόμενα A, B ονομάζονται **αντίθετα** ή **συμπληρωματικά** αν η πραγματοποίηση του ενός συνεπάγεται την μη πραγματοποίηση του άλλου και αντιστρόφως.

  Ισοδύναμα, $A \cup B = \Omega$, $A \cap B = \emptyset$. Ισοδύναμα, $A = \overline{B}$.

- **Ένωση ενδεχομένων** A, B: Πραγματοποιείται όταν πραγματοποιείται τουλάχιστον ένα από τα δύο.

- **Τομή ενδεχομένων** A, B: Πραγματοποιείται όταν πραγματοποιούνται και τα δύο ενδεχόμενα.

1.2.1 Επιπλέον ιδιότητες της πιθανότητας

Πρόταση 1.2 (Ιδιότητα υποαθροισιμότητας).

i) *Για οποιαδήποτε n-άδα ενδεχομένων $A_1, A_2, \ldots, A_n \subseteq \Omega$ ισχύει ότι*

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) \leq P(A_1) + P(A_2) + \cdots + P(A_n)$$

ii) *Για οποιαδήποτε αριθμήσιμη οικογένεια ενδεχομένων $A_1, A_2, \ldots \subseteq \Omega$ ισχύει ότι*

$$P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i).$$

Απόδειξη. i) (Άσκηση)

ii) Παρατηρήστε ότι

$$\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i, \text{ όπου } B_i = A_i \cap \overline{\left(\bigcup_{k=1}^{i-1} A_k\right)}$$

και επιπλέον ότι τα B_i είναι ξένα ανά δύο με $B_i \subseteq A_i$. Επομένως,

$$P(\bigcup_{i=1}^{\infty} A_i) = P(\bigcup_{i=1}^{\infty} B_i) = \sum_{i=1}^{\infty} P(B_i) \leq \sum_{i=1}^{\infty} P(A_i). \qquad \square$$

Πρόταση 1.3. *Αν για μια ακολουθία ενδεχομένων $A_i \subseteq \Omega$, $i \in \mathbb{N}$, ισχύει ότι*

$$\sum_{i=1}^{\infty} P(A_i) < +\infty,$$

τότε

$$P(\limsup_{n\to\infty} A_n) = 0.$$

Απόδειξη. Θέτουμε $B_n := \bigcup_{i=n}^{\infty} A_i$, $S_n := \bigcap_{j=1}^{n} B_j$ και $S := \limsup_{n\to\infty} A_n$. Εξ ορισμού είναι $S = \lim_{n\to\infty} S_n$. Επιπλέον, θέτουμε $p_n := \sum_{i=1}^{n} P(A_i)$. Επειδή $p := \lim_{n\to\infty} p_n \in [0, +\infty)$, από τον ορισμό του ορίου ακολουθίας αριθμών, έπεται ότι για $\epsilon > 0$ υπάρχει $n_0 > 1$ τέτοιο ώστε για κάθε $n \geq n_0$ να ισχύει

$$\sum_{i=n}^{\infty} P(A_i) = p - \sum_{i=1}^{n-1} P(A_i) \leq |p - p_{n-1}| < \epsilon.$$

Από την άλλη, προφανώς είναι $S \subseteq S_n \subseteq B_n$, οπότε

$$P(S) \leq P(S_n) \leq P(B_n) = P\left(\bigcup_{i=n}^{\infty} A_i\right) \leq \sum_{i=n}^{\infty} P(A_i) < \epsilon,$$

άρα $P(S) < \epsilon$, για κάθε $\epsilon > 0$, δηλαδή $P(S) = 0$. $\square$

Πρόταση 1.4. *Αν μια ακολουθία ενδεχομένων $A_i \subseteq \Omega$, $i \in \mathbb{N}$, είναι αύξουσα δηλαδή ισχύει ότι*

$$A_1 \subseteq A_2 \subseteq \cdots$$

τότε

$$P(\bigcup_{n=1}^{\infty} A_n) = P(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} P(A_n).$$

Απόδειξη. Για κάθε $n \in \mathbb{N}^*$ το σύνολο A_n γράφεται ως ξένη ένωση συνόλων ως εξής:

$$A_n = A_1 \cup (A_2 \setminus A_1) \cup (A_3 \setminus A_2) \cup \cdots \cup (A_n \setminus A_{n-1}).$$

Επίσης, αντίστοιχα

$$\lim_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} A_n = A_1 \cup (A_2 \setminus A_1) \cup \cdots = A_1 \cup \bigcup_{j=2}^{\infty} (A_j \setminus A_{j-1}).$$

Επομένως,

$$P(\lim_{n\to\infty} A_n) = P(\bigcup_{n=1}^{\infty} A_n) = P(A_1) + \sum_{j=2}^{\infty} P(A_j \setminus A_{j-1}) = \lim_{n\to\infty} \left(P(A_1) + \sum_{j=2}^{n} P(A_j \setminus A_{j-1}) \right) = \lim_{n\to\infty} P(A_n). \qquad \square$$

Πόρισμα 1.5. *Αν μια ακολουθία ενδεχομένων $A_i \subseteq \Omega$, $i \in \mathbb{N}$, είναι φθίνουσα δηλαδή ισχύει ότι*

$$A_1 \supseteq A_2 \supseteq \cdots$$

τότε

$$P(\bigcap_{n=1}^{\infty} A_n) = P(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} P(A_n).$$

Απόδειξη. Παρατηρήστε ότι η ακολουθία $(\overline{A_n})$ είναι αύξουσα, επομένως, από την προηγούμενη πρόταση και τον κανόνα του De Morgan ισχύει ότι

$$P(\bigcap_{n=1}^{\infty} A_n) = 1 - P(\overline{\left(\bigcap_{n=1}^{\infty} A_n\right)}) = 1 - P(\bigcup_{n=1}^{\infty} \overline{A_n}) = 1 - \lim_{n\to\infty} P(\overline{A_n}) = \lim_{n\to\infty}(1 - P(\overline{A_n})) = \lim_{n\to\infty} P(A_n). \qquad \square$$

1.2.2 Γεωμετρική πιθανότητα

Ορισμός (Πιθανότητα σε υπεραριθμήσιμους χώρους). *Έστω Ω ένας υπεραριθμήσιμος δειγματικός χώρος οριζόμενος από μια περιοχή του (μοναδιάστατου ή διδιάστατου ή τρισδιάστατου) χώρου στην οποία οποιεσδήποτε στοιχειώδεις περιοχές είναι εξίσου πιθανές. Η πιθανότητα ενός ενδεχομένου A οριζόμενου από μια περιοχή του δειγματικού χώρου Ω δίδεται από την σχέση*

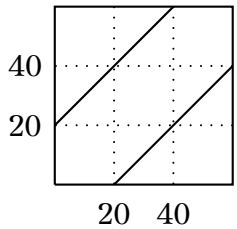
$$P(A) = \frac{\mu(A)}{\mu(\Omega)}$$

όπου $\mu(A)$ και $\mu(\Omega)$ είναι το μέτρο (μήκος, ή εμβαδόν, ή όγκος) των περιοχών A και Ω αντίστοιχα.

Παράδειγμα 1.2.3. *Δύο φίλοι συμφωνούν να συναντηθούν στην πλατεία Κοραή κάποια στιγμή στο διάστημα μεταξύ 5:00 μ.μ. και 6:00 μ.μ. Ο καθένας τους φτάνει κάποια τυχαία στιγμή, περιμένει για 20 λεπτά, και αν δεν έρθει ο άλλος φεύγει. Ποια η πιθανότητα να συναντηθούν;*

Λύση. Έστω F_1, F_2 οι δύο φίλοι. Αν $x \in [0, 60]$ είναι η χρονική στιγμή άφιξης του F_1 στο διάστημα 5:00 μέχρι 6:00 και $y \in [0, 60]$ είναι η χρονική στιγμή άφιξης του F_2, τότε μπορούμε να θεωρήσουμε ως δειγματικό χώρο το σύνολο των σημείων του τετραγώνου $[0, 60] \times [0, 60]$.

Οι δύο φίλοι θα συναντηθούν ανν $|x - y| \leq 20$, επομένως το ενδεχόμενο A οι δύο φίλοι να συναντηθούν αντιστοιχεί στα σημεία του χωρίου με $|x - y| \leq 20$, το οποίο περιγράφεται στο σχήμα:



Επομένως, η ζητούμενη πιθανότητα ισούται με το πηλίκο των εμβαδόν των δύο χωρίων:
$P(A) = \frac{60^2 - 2\frac{1}{2}40^2}{60^2} = 1 - \left(\frac{2}{3}\right)^2 = 1 - \frac{4}{9} = \frac{5}{9} \simeq 0.55.$ $\square$

1.3 Ανεξάρτητα ενδεχόμενα

Ορισμός (Δύο ανεξάρτητα ενδεχόμενα). *Τα ενδεχόμενα $A, B \subseteq \Omega$ ονομάζονται **ανεξάρτητα** αν και μόνο αν*

$$P(A \cap B) = P(A)P(B).$$

Παραδείγματα:

- Οι διαδοχικές ρίψεις ενός ζαριού είναι ανεξάρτητες (το ζάρι δεν έχει μνήμη!)
- Η θερμοκρασία στην Αθήνα σήμερα και η θερμοκρασία στην Αθήνα αύριο δεν είναι ανεξάρτητες.

Πρόταση 1.6. *Αν τα ενδεχόμενα $A, B \subseteq \Omega$ είναι ανεξάρτητα, τότε τα ενδεχόμενα*

i) *$A, \overline{B}$ είναι ανεξάρτητα.*

ii) *$\overline{A}, \overline{B}$ είναι ανεξάρτητα.*

Απόδειξη. i) Ισχύει ότι

$$P(A \cap \overline{B}) = P(A) - P(A \cap B) = P(A) - P(A)P(B) = P(A)(1 - P(B)) = P(A)P(\overline{B}).$$

ii) Ισχύει ότι

$$\begin{aligned} P(\overline{A} \cap \overline{B}) &= P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - P(A) - P(B) - P(A \cap B) \\ &= 1 - P(A) - P(B) - P(A)P(B) = 1 - P(A) - P(B)(1 - P(A)) \\ &= (1 - P(A))(1 - P(B)) = P(\overline{A})P(\overline{B}) \end{aligned}$$

□

Παράδειγμα 1.3.1. *Ρίχνουμε 3 φορές ένα αμερόληπτο νόμισμα.*

i) *Να βρεθεί ο δειγματικός χώρος Ω του πειράματος.*

ii) *Να βρεθεί το ενδεχόμενο A να μην εμφανισθεί κορώνα σε καμία από τις 3 ρίψεις.*

iii) *Να βρεθεί η πιθανότητα $P(A)$.*

Λύση. i) $\Omega = \{ΚΚΚ, ΚΚΓ, ΚΓΚ, ΚΓΓ, ΓΚΚ, ΓΚΓ, ΓΓΚ, ΓΓΓ\}$, $|\Omega| = 2^3 = 8$.

ii) $A = \{ΓΓΓ\}$.

iii) (1ος τρόπος) Αν υποθέσουμε ότι όλα τα ενδεχόμενα του Ω είναι ισοπίθανα, $P(A) = \dfrac{|A|}{|\Omega|} = \dfrac{1}{8}$.

(2ος τρόπος) Το αποτέλεσμα κάθε ρίψης του νομίσματος είναι ανεξάρτητο από τις άλλες. Το A πραγματοποιείται αν δεν έρθει κορώνα ούτε στην 1η, ούτε στην 2η, ούτε στην 3η ρίψη, οπότε

$$P(A) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}.$$

□

Παράδειγμα 1.3.2 (Κορώνα ή γράμματα με μεροληπτικό νόμισμα). *Έστω ένα νόμισμα το οποίο δεν είναι αμερόληπτο αλλά εμφανίζει κορώνα με πιθανότητα p και γράμματα με πιθανότητα $q = 1 - p$, με $p \in (0,1)$. Να βρεθεί ένας τρόπος να χρησιμοποιήσουμε το νόμισμα αυτό με δίκαιο τρόπο (δηλαδή να επιλέξουμε ανάμεσα σε δύο ενδεχόμενα με πιθανότητα $1/2$ το καθένα).*

Λύση του Von Neumann. Ρίχνουμε το νόμισμα 2 φορές:
- Αν φέρει πρώτα κορώνα και μετά γράμματα τότε θεωρούμε ότι έφερε κορώνα.
- Αν φέρει πρώτα γράμματα και μετά κορώνα τότε θεωρούμε ότι έφερε γράμματα.
- Αν φέρει δύο ίδια αποτελέσματα τότε ξαναρίχνουμε από την αρχή.

Επειδή το ζάρι δεν έχει μνήμη:
Η πιθανότητα να φέρει πρώτα κορώνα και μετά γράμματα είναι $p \cdot q$.
Η πιθανότητα να φέρει πρώτα γράμματα και μετά κορώνα είναι $q \cdot p$.
Η πιθανότητα να φέρει 2 φορές γράμματα είναι q^2.
Η πιθανότητα να φέρει 2 φορές κορώνα είναι p^2.
Άρα, ο τρόπος αυτός είναι δίκαιος αφού έχει την ίδια πιθανότητα. □

Παρατήρηση. Η μέθοδος αυτή, που ονομάζεται **τέχνασμα του Von Neumann**, έχει το μειονέκτημα ότι ανάλογα με τα p, q μπορεί να χρειαστεί να επαναλάβουμε τις ρίψεις πολλές φορές. Συγκεκριμένα, η πιθανότητα επιτυχίας είναι $2pq$ και, με βάση την γεωμετρική κατανομή που θα δούμε αργότερα, ο αναμενόμενος αριθμός επαναλήψεων μέχρι την πρώτη επιτυχία είναι $\dfrac{1}{2pq}$.

Όπως φαίνεται και στην επόμενη άσκηση, δεν είναι πάντα προφανές αν δύο ενδεχόμενα είναι ανεξάρτητα.

Παράδειγμα 1.3.3. *Ρίχνουμε ένα αμερόληπτο τετράεδρο ζάρι δύο φορές για το οποίο τα 16 δυνατά αποτελέσματα είναι ισοπίθανα με πιθανότητα $1/16$.*

i) *Να εξετασθεί αν είναι ανεξάρτητα ή όχι τα ενδεχόμενα*

A: Η 1η ρίψη είναι 1, *B: Το άθροισμα των δύο ρίψεων είναι 5.*

ii) *Να εξετασθεί αν είναι ανεξάρτητα ή όχι τα ενδεχόμενα*

A: Το μέγιστο των δύο ρίψεων είναι 2, *B: Το ελάχιστο των δύο ρίψεων είναι 2.*

Λύση. Ο δειγματικός χώρος του πειράματος είναι το σύνολο $\Omega = \{(i,j) : i, j \in [4]\}$.
Σε κάθε περίπτωση θα υπολογίσουμε τις πιθανότητες $P(A)$, $P(B)$ και $P(A \cap B)$.

i) $A = \{(1,j) : j = 1,2,3,4\} = \{11,12,13,14\}$. Άρα, $P(A) = 1/4$.
$B = \{(i,j) : i + j = 5\} = \{14,23,32,41\}$. Άρα, $P(B) = 1/4$.
$A \cap B = \{14\}$. Άρα, $P(A \cap B) = \dfrac{1}{16} = \dfrac{1}{4} \cdot \dfrac{1}{4} = P(A)P(B)$, δηλαδή τα A, B είναι ανεξάρτητα.

ii) $A = \{(i,j) : \max(i,j) = 2\} = \{12,21,22\}$. Άρα, $P(A) = 3/16$.
$B = \{(i,j) : \min(i,j) = 2\} = \{22,23,24,32,42\}$. Άρα, $P(B) = 5/16$.
$A \cap B = \{22\}$. Άρα, $P(A \cap B) = \dfrac{1}{16} \neq \dfrac{3}{16} \cdot \dfrac{5}{16} = P(A)P(B)$, δηλαδή τα A, B δεν είναι ανεξάρτητα. □

Η ανεξαρτησία n ενδεχομένων ορίζεται ως εξής:

Ορισμός (*n* **ανεξάρτητα ενδεχόμενα).** *Τα ενδεχόμενα $A_1, A_2, \ldots, A_n \subseteq \Omega$ ονομάζονται (πλήρως) ανεξάρτητα αν και μόνο αν για κάθε υποσύνολο $\{i_1, i_2, \ldots, i_k\}$ του $[n]$, όπου $2 \leq k \leq n$, ισχύει ότι*

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k}).$$

Παράδειγμα 1.3.4 (Ο νόμος του Murphy). *Έστω $A_1, A_2, \ldots, A_n$ ανεξάρτητα "όχι και τόσο ευχάριστα" ενδεχόμενα που μπορούν να μας συμβούν με μικρές πιθανότητες $p_1, p_2, \ldots, p_n$ το καθένα.*
Να βρεθεί η πιθανότητα να συμβεί τουλάχιστον ένα από αυτά.

Λύση. Είναι ευκολότερο να υπολογίσουμε την συμπληρωματική πιθανότητα να μην συμβεί κανένα από αυτά τα ενδεχόμενα. Επειδή τα A_1, A_2, …, A_n είναι ανεξάρτητα έπεται ότι και τα συμπληρωματικά ενδεχόμενα είναι επίσης ανεξάρτητα.

$$\begin{aligned} P(A_1 \cup A_2 \cup \cdots \cup A_n) &= 1 - P(\overline{A_1 \cup A_2 \cup \cdots \cup A_n}) \\ &= 1 - P(\overline{A_1} \cap \overline{A_2} \cap \cdots \cap \overline{A_n}) \\ &= 1 - P(\overline{A_1})P(\overline{A_2}) \cdots P(\overline{A_n}) \\ &= 1 - (1 - P(A_1))(1 - P(A_2)) \cdots (1 - P(A_n)) \end{aligned}$$

Από την ανισότητα $1 + x \leq e^x$ έπεται ότι $(1 - P(A_i)) \leq e^{-P(A_i)}$ για κάθε $i \in [n]$, άρα

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) \geq 1 - e^{-P(A_1)}e^{-P(A_2)} \cdots e^{-P(A_n)} = 1 - e^{-(P(A_1)+P(A_2)+\cdots+P(A_n))}$$

Άρα, $P(A_1 \cup A_2 \cup \cdots \cup A_n) \geq 1 - e^{-(p_1+p_2+\cdots+p_n)}$.

Αν υποθέσουμε ότι η πιθανότητα να συμβεί κάθε ένα από τα ενδεχόμενα αυτά είναι αρκετά μικρή, για παράδειγμα $p = \dfrac{1}{1000}$ η ανισότητα που βρήκαμε μας λέει πως αν υπάρχουν πολλά τέτοια σπάνια "όχι τόσο ευχάριστα ενδεχόμενα" για παράδειγμα έστω ότι υπάρχουν 100 τέτοια ενδεχόμενα, τότε η πιθανότητα να συμβεί τουλάχιστον ένα είναι μεγαλύτερη ή ίση από

$$1 - e^{-\frac{100}{1000}} = 1 - e^{-\frac{1}{10}} = 0.0951 = 9.5\%,$$

με άλλα λόγια αν υπάρχουν πολλά πράγματα που μπορούν "να πάνε στραβά" (π.χ. 100) ακόμα και με μικρή πιθανότητα το καθένα (π.χ 1 στις 1000) τότε περίπου 1 στις 10 μέρες τουλάχιστον ένα πράγμα θα "πηγαίνει στραβά" (μια εκδοχή του **νόμου του Murphy**.) □

Επίσης, η ανεξαρτησία αριθμήσιμου πλήθους ενδεχομένων ορίζεται ως εξής:

Ορισμός (**Άπειρα ανεξάρτητα ενδεχόμενα).** *Τα ενδεχόμενα $A_i \subseteq \Omega$, $i \in \mathbb{N}$, ονομάζονται (πλήρως) ανεξάρτητα αν και μόνο αν για κάθε $n \in \mathbb{N}$ τα ενδεχόμενα $A_1, A_2, \ldots, A_n$ είναι ανεξάρτητα.*

Πρόταση 1.7. *Αν $(A_i)_{i\in\mathbb{N}}$ είναι μια ακολουθία ανεξάρτητων ενδεχομένων με $\sum_{n=1}^{\infty} P(A_n) = \infty$, τότε $P(\lim_{n\to\infty} \sup A_n) = 1$.*

1.4 Δεσμευμένη πιθανότητα

Μερικές φορές σε κάποιο δειγματικό χώρο Ω υπάρχουν ενδεχόμενα A, B τα οποία δεν είναι ανεξάρτητα. Για παράδειγμα έστω ο δειγματικός χώρος Ω που περιλαμβάνει (όλα) τα γεγονότα που μπορούν να συμβούν στη ζωή μας σήμερα! Θεωρήστε τα ενδεχόμενα:

A: Σήμερα το πρωί, οι δρόμοι θα έχουν μποτιλιάρισμα.

B: Θα αργήσω να φτάσω στη δουλειά μου το πρωί.

Η πραγματοποίηση ή μη του ενδεχόμενου A μπορεί να επηρεάζει την πιθανότητα πραγματοποίησης του ενδεχομένου B. Προκειμένου να χειριστούμε τέτοιες περιπτώσεις θα εισάγουμε την έννοια της δεσμεύμενης πιθανότητας.

Ορισμός (Δεσμευμένη πιθανότητα). *Η **δεσμευμένη πιθανότητα του ενδεχομένου** A **δεδομένου του ενδεχομένου** $B \neq \varnothing$ συμβολίζεται με $P(A|B)$ και ορίζεται ως εξής:*

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

Παρατήρηση. Η συνάρτηση της δεσμευμένης πιθανότητας είναι συνάρτηση πιθανότητας, δηλαδή ικανοποιεί τα αξιώματα του ορισμού της πιθανότητας:

- $P(A|B) \geq 0$, για κάθε $A \subseteq \Omega$. Πράγματι, $P(A \cap B) \geq 0, P(B) > 0 \Rightarrow P(A|B) = \dfrac{P(A \cap B)}{P(B)} \geq 0.$
- $P(\Omega|B) = 1$. Πράγματι, $P(\Omega|B) = \dfrac{P(\Omega \cap B)}{P(B)} = \dfrac{P(B)}{P(B)} = 1.$
- $P(\bigcup_{n=1}^{\infty} A_i|B) = \sum_{n=1}^{\infty} P(A_i|B)$, για κάθε αριθμήσιμη οικογένεια $(A_i)_{i\in\mathbb{N}^*}$ ξένων ανά δύο υποσυνόλων του Ω. Πράγματι, τα σύνολα $A_i \cap B$, $i \in \mathbb{N}^*$, είναι προφανώς ανά δύο ξένα, οπότε

$$P(\bigcup_{n=1}^{\infty} A_i|B) = \frac{P((\bigcup_{n=1}^{\infty} A_i) \cap B)}{P(B)} = \frac{P(\bigcup_{n=1}^{\infty}(A_i \cap B))}{P(B)} = \frac{\sum\limits_{n=1}^{\infty} P(A_i \cap B)}{P(B)} = \sum_{n=1}^{\infty} P(A_i|B).$$

Διαισθητικά, όταν είναι δεδομένο το B, τότε αυτό αποτελεί τον νέο, συρρικνωμένο, δειγματικό χώρο που περιλαμβάνει όλα τα δυνατά αποτελέσματα. Εξάλλου, είναι $P(A) = P(A \cap \Omega) = P(A \cap \Omega)/P(\Omega) = P(A|\Omega)$, δηλαδή και η πιθανότητα $P(A)$ μπορεί να θεωρηθεί ως δεσμευμένη, με δεσμό το Ω. Επομένως, η συνάρτηση της δεσμευμένης πιθανότητας ικανοποιεί και τις ιδιότητες της πιθανότητας.

Πρόταση 1.8 (Ιδιότητες δεσμευμένης πιθανότητας).
Αν $A, B, C, D \subseteq \Omega$, με $D \neq \varnothing$ τότε

i) $P(\overline{A}|D) = 1 - P(A|D)$

ii) $P(\varnothing|D) = 0$

iii) $P(A|D) \leq 1$

iv) $P(A \setminus B|D) = P(A \cap \overline{B}|D) = P(A|D) - P(A \cap B|D)$

v) $A \subseteq B \Rightarrow P(A|D) \leq P(B|D)$

vi) $P(A \cup B|D) = P(A|D) + P(B|D) - P(A \cap B|D)$

vii) $P(A \cup B \cup C|D) = P(A|D) + P(B|D) + P(C|D) - P(A \cap B|D) - P(A \cap C|D) - P(B \cap C|D) + P(A \cap B \cap C|D)$

Παράδειγμα 1.4.1. *Έστω ένα αμερόληπτο ζάρι. Θεωρούμε τα ενδεχόμενα:*
A: Το ζάρι φέρνει 1
B: Το ζάρι φέρνει περιττό αριθμό.
C: Το ζάρι δεν φέρνει 5.
Να βρεθούν οι πιθανότητες $P(A)$, $P(B)$, $P(C)$, $P(A|B)$, $P(B|A)$, $P(B|C)$, $P(C|B)$.

Λύση. Ο δειγματικός χώρος Ω είναι $\Omega = \{1, 2, 3, 4, 5, 6\}$. Επίσης

$$A = \{1\},\ B = \{1, 3, 5\},\ C = \{1, 2, 3, 4, 6\},\ A \cap B = \{1\},\ B \cap C = \{1, 3\}$$

Επειδή τα στοιχειώδη ενδεχόμενα είναι ισοπίθανα, έχουμε ότι

$$P(A) = \frac{1}{6},\ P(B) = \frac{3}{6},\ P(C) = \frac{5}{6},\ P(A \cap B) = \frac{1}{6},\ P(B \cap C) = \frac{2}{6},$$

οπότε

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1}{3},\ P(B|A) = \frac{P(A \cap B)}{P(A)} = 1,\ P(B|C) = \frac{P(B \cap C)}{P(C)} = \frac{2}{5},\ P(C|B) = \frac{P(B \cap C)}{P(B)} = \frac{2}{3}.$$ □

1.4.1 Τύπος ολικής πιθανότητας

Πολλές φορές είναι ευκολότερο να υπολογίσουμε την δεσμευμένη πιθανότητα του ενδεχομένου A δεδομένου του B από ότι την πιθανότητα της τομής $A \cap B$. Από τον ορισμό της δεσμευμένης πιθανότητας $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$ προκύπτει ο χρήσιμος τύπος

$$P(A \cap B) = P(A|B)P(B).$$

Επιπλέον, κάθε σύνολο A εκφράζεται ως ξένη ένωση των συνόλων $A \cap B, A \cap \overline{B}$, για οποιοδήποτε σύνολο $B \subseteq \Omega$, δηλαδή $A = (A \cap B) \cup (A \cap \overline{B})$. Γενικότερα, αν τα $B_1, B_2, \ldots, B_n$ αποτελούν μια διαμέριση του Ω, τότε

$$A = \bigcup_{i=1}^{n} (A \cap B_i).$$

Με βάση τους προηγούμενους τύπους, προκύπτει ο τύπος της ολικής πιθανότητας:

Πρόταση 1.9 (Θεώρημα ολικής πιθανότητας). *Αν τα (μη κενά) σύνολα της οικογένειας $(B_i)_{i\in[n]}$, $n \in \mathbb{N}^*$, αποτελούν διαμέριση του Ω, τότε για κάθε $A \subseteq \Omega$ ισχύει*

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A|B_i)P(B_i).$$

Παράδειγμα 1.4.2. *Σε ένα διήγημα της Αγκάθα Κρίστι ο ντετέκτιβ Πουαρό έχει συγκεντρώσει τους 7 εξίσου υπόπτους (μεταξύ των οποίων και η ανιψιά του δολοφονηθέντος) σε ένα κυκλικό τραπέζι ενώ ο ίδιος στέκεται όρθιος δίπλα τους. Ποια είναι η πιθανότητα η ανιψιά να κάθεται δίπλα στο δολοφόνο;*

Λύση. Έστω A το ενδεχόμενο η ανιψιά να είναι δολοφόνος και B το ενδεχόμενο η ανιψιά να κάθεται δίπλα στον δολοφόνο. Τότε $P(A) = \dfrac{1}{7}$, $P(B|A) = 0$, $P(B|\overline{A}) = \dfrac{2}{6}$ (διότι υπάρχουν 6 θέσεις για να καθίσει ο δολοφόνος και εκ των οποίων δύο είναι δίπλα στην ανιψιά). Από τον τύπο της ολικής πιθανότητας έχουμε ότι

$$P(B) = P(B|A)P(A) + P(B|\overline{A})P(\overline{A}) = 0 \cdot \frac{1}{7} + \frac{2}{6}\left(1 - \frac{1}{7}\right) = \frac{2}{6} \cdot \frac{6}{7} = \frac{2}{7}. \qquad \square$$

Παράδειγμα 1.4.3. *Σε ένα τουρνουά σκάκι υπάρχουν 3 κατηγορίες παιχτών: A, B, Γ.*
Αν παίξουμε με κάποιον στην κατηγορία A η πιθανότητα να κερδίσουμε είναι 0.3.
Αν παίξουμε με κάποιον στην κατηγορία B η πιθανότητα να κερδίσουμε είναι 0.4.
Αν παίξουμε με κάποιον στην κατηγορία Γ η πιθανότητα να κερδίσουμε είναι 0.5.
Η πιθανότητα κάποιος παίχτης να ανήκει στην κατηγορία A είναι 0.5.
Η πιθανότητα κάποιος παίχτης να ανήκει στην κατηγορία B είναι 0.25.
Η πιθανότητα κάποιος παίχτης να ανήκει στην κατηγορία Γ είναι 0.25.
Ποια είναι η πιθανότητα να παίξουμε με ένα τυχαίο παίχτη του τουρνουά και να κερδίσουμε;

Λύση. Έστω W το ενδεχόμενο να κερδίσουμε και A, B, Γ το ενδεχόμενο να παίξουμε με κάποιο παίχτη που ανήκει στην κατηγορία A, B, Γ αντίστοιχα. Από τον τύπο της ολικής πιθανότητας, προκύπτει ότι

$$P(W) = P(W|A)P(A) + P(W|B)P(B) + P(W|\Gamma)P(\Gamma) = 0.3 \cdot 0.5 + 0.4 \cdot 0.25 + 0.5 \cdot 0.25 = 0.375 \qquad \square$$

1.4.2 Τύπος του Bayes

Για τα ενδεχόμενα $A, B \neq \emptyset$, από τον ορισμό της δεσμευμένης πιθανότητας, έχουμε ότι

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad \text{και} \qquad P(B|A) = \frac{P(A \cap B)}{P(A)},$$

οπότε προκύπτουν οι τύποι

$$P(A \cap B) = P(A|B)P(B) \qquad \text{και} \qquad P(A \cap B) = P(B|A)P(A).$$

Εξισώνοντας τα δεύτερα μέλη των δύο παραπάνω τύπων, προκύπτει ο (διάσημος) τύπος του Bayes (Μπέιζ):

Πρόταση 1.10 (Τύπος του Bayes). *Για κάθε $A, B \subseteq \Omega$ με $A, B \neq \emptyset$ ισχύει ότι*

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Επίσης, χρησιμοποιώντας το θεώρημα ολικής πιθανότητας, ο τύπος του Bayes μπορεί να γραφεί ως εξής:

Πρόταση 1.11 (Τύπος του Bayes). *Αν η οικογένεια $(B_i)_{i\in[n]}$, $n \in \mathbb{N}^*$, αποτελεί διαμέριση του Ω, τότε για κάθε μη κενό ενδεχόμενο $A \subseteq \Omega$ και για κάθε $i \in [n]$ ισχύει*

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)}.$$

Παράδειγμα 1.4.4. *Ένας χρήστης έχει παρατηρήσει ότι από τα email που λαμβάνει καθημερινά ότι το 60% είναι γραμμένα στα Αγγλικά και το υπόλοιπο 40% στα Ελληνικά. Επίσης, έχει παρατηρήσει ότι από τα email γραμμένα στα Αγγλικά το 90% είναι ανεπιθύμητα (spam) και από τα email γραμμένα στα Ελληνικά το 30% είναι ανεπιθύμητα.*

Να βρεθεί η πιθανότητα ένα email που διαβάζει ο χρήστης να είναι spam.

Να βρεθεί η πιθανότητα ένα spam email που διαβάζει ο χρήστης να είναι γραμμένο στα Ελληνικά.

Λύση. Έστω A το ενδεχόμενο το email να είναι γραμμένο στα Αγγλικά.
Έστω E το ενδεχόμενο το email να είναι γραμμένο στα Ελληνικά.
Έστω S το ενδεχόμενο το email να είναι spam.
$P(A) = 0.6$, $P(E) = 0.4$
$P(S|A) = 0.9$, $P(S|E) = 0.3$.
$P(S) = P(S|A)P(A) + P(S|E)P(E) = 0.9 \cdot 0.6 + 0.3 \cdot 0.4 = 0.66$.
Άρα, $P(E|S) = \dfrac{P(S|E)P(E)}{P(S)} = \dfrac{0.3 \cdot 0.4}{0.66} = 0.1818$. □

Παράδειγμα 1.4.5. *Έστω ότι μια πάθηση του αίματος εμφανίζεται με συχνότητα 2 φορές στα 1000 άτομα. Υπάρχει ένα τεστ για την πάθηση αυτή το οποίο έχει ποσοστό σφάλματος 5%, τόσο στους υγιείς, όσο και στους ασθενείς.*

Να βρεθεί η πιθανότητα κάποιος άνθρωπος X να πάσχει από αυτήν την πάθηση δεδομένου ότι το τεστ βγήκε θετικό.

Λύση. Έστω A το ενδεχόμενο ο X να έχει την πάθηση αυτή και Θ το ενδεχόμενο το τεστ να βγει θετικό. Ψάχνουμε την πιθανότητα $P(A|\Theta)$. Γνωρίζουμε (από την εκφώνηση) ότι

$$P(A) = \frac{2}{1000} = 0.002, \quad P(\Theta|A) = P(\overline{\Theta}|\overline{A}) = 0.95, \quad P(\overline{\Theta}|A) = P(\Theta|\overline{A}) = 0.05.$$

Από τον τύπο του Bayes, έχουμε ότι $P(A|\Theta) = \dfrac{P(\Theta|A)P(A)}{P(\Theta)}$.

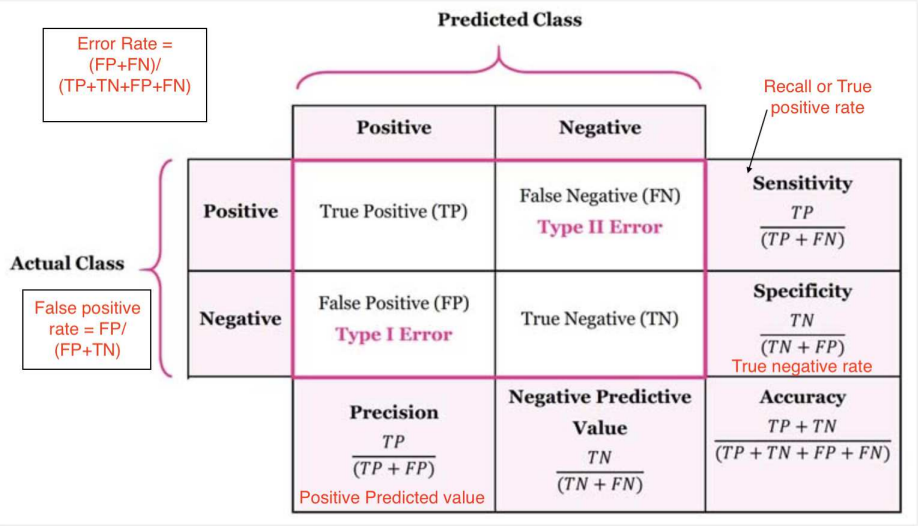
Επιπλέον, επειδή κάθε άνθρωπος ή πάσχει ή δεν πάσχει από την πάθηση, έπεται ότι τα $A, \overline{A}$ αποτελούν διαμέριση του Ω, οπότε, από τον τύπο της ολικής πιθανότητας, έχουμε ότι

$$P(\Theta) = P(\Theta|A)P(A) + P(\Theta|\overline{A})P(\overline{A}) = 0.95 \cdot 0.002 + 0.05 \cdot 0.998 = 0.0518.$$

Άρα

$$P(A|\Theta) = \frac{P(\Theta|A)P(A)}{P(\Theta)} = \frac{0.95 \cdot 0.002}{0.0518} = 0.03667 \approx 3.67\%$$ □

Παρατηρήσεις. Το αποτέλεσμα του τεστ ονομάζεται ψευδώς ή αληθώς αρνητικό (False Negative - FN, True Negative - TN), όταν προκύπτει αρνητικό σε ασθενή ή υγιή άνθρωπο αντίστοιχα και ονομάζεται ψευδώς ή αληθώς θετικό (False Positive - FP, True Positive - TP), όταν προκύπτει θετικό σε υγιή ή ασθενή άνθρωπο αντίστοιχα. Οι δείκτες ευαισθησία (sensitivity - TPR), εξειδίκευση (specificity - TNR), θετική προβλεπτική αξία (positive predictive value - PPV), αρνητική προβλεπτική αξία (negative predictive value - NPV) και ακρίβεια (accuracy - ACC) ορίζονται σύμφωνα με τον ακόλουθο πίνακα (όπου οι ποσότητες TP, TN, FP, FN αναφέρονται σε πλήθος αποτελεσμάτων, κατόπιν διεξαγωγής του τεστ σε όλον τον πληθυσμό):



Ερμηνεύοντας τη συχνότητα εμφάνισης κάθε αποτελέσματος ως πιθανότητα, οι παραπάνω δείκτες σχετίζονται με τις δεσμευμένες πιθανότητες ως εξής:

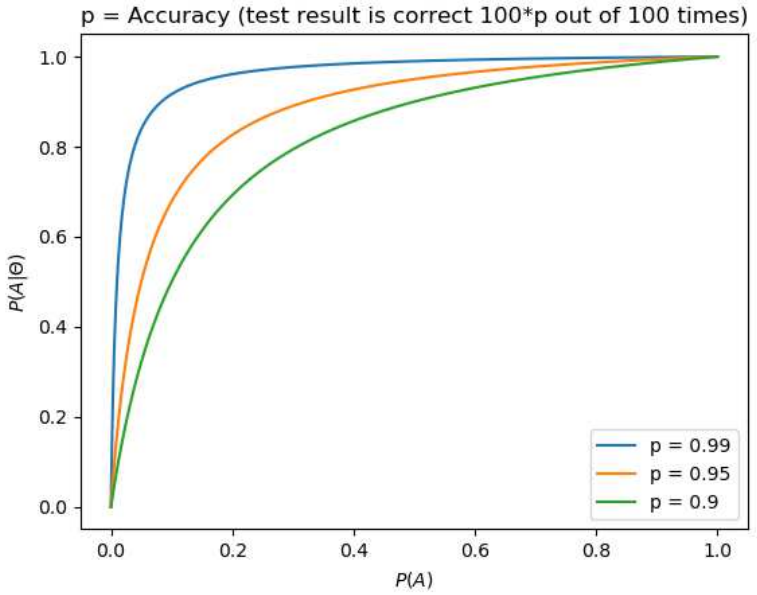
- $P(A|\Theta) = \dfrac{P(A \cap \Theta)}{P(A \cap \Theta) + P(\overline{A} \cap \Theta)} = \dfrac{TP}{TP + FP} = PPV$
- $P(\overline{A}|\overline{\Theta}) = \dfrac{P(\overline{A} \cap \overline{\Theta})}{P(\overline{A} \cap \overline{\Theta}) + P(A \cap \overline{\Theta})} = \dfrac{TN}{TN + FN} = NPV$
- $P(\overline{\Theta}|\overline{A}) = \dfrac{P(\overline{A} \cap \overline{\Theta})}{P(\overline{A} \cap \overline{\Theta}) + P(\overline{A} \cap \Theta)} = \dfrac{TN}{TN + FP} = TNR$ (specificity)
- $P(\Theta|A) = \dfrac{P(A \cap \Theta)}{P(A \cap \Theta) + P(A \cap \overline{\Theta})} = \dfrac{TP}{TP + FN} = TPR$ (sensitivity)
- $P((\Theta \cap A) \cup (\overline{\Theta} \cap \overline{A})) = P(\Theta \cap A) + P(\overline{\Theta} \cap \overline{A}) = P(\Theta|A)P(A) + P(\overline{\Theta}|\overline{A})P(\overline{A}) = ACC$

Παρατηρήστε ότι, όταν $TPR = TNR$ (όπως στο προηγούμενο παράδειγμα), τότε είναι $TPR = TNR = ACC$. Γενικά όμως, είναι $TPR \neq TNR$ και η ακρίβεια ACC είναι ένας σταθμισμένος μέσος των TPR και TNR (όπως προκύπτει από τον τελευταίο τύπο).

Η πιθανότητα $P(A|\Theta)$, που υπολογίσθηκε στο προηγούμενο παράδειγμα, ισούται με την PPV, δηλαδή αν γίνει το τεστ σε όλον τον πληθυσμό, αναμένεται το 3.67% των θετικών να είναι πραγματικά ασθενείς. Ο λόγος που συμβαίνει αυτό, δηλαδή που η πιθανότητα $P(A|\Theta)$ είναι πολύ μικρότερη της πιθανότητας 0.95 ορθού αποτελέσματος του τεστ (ACC = 95%), είναι ότι το ποσοστό $P(A) = 0.002$ των ασθενών στον πληθυσμό είναι πολύ μικρό. Έτσι το πλήθος των FP είναι πολύ μεγάλο (το 5% των υγιών) σε σχέση με τα TP (το 95% των ασθενών).

Αν και η πιθανότητα ασθένειας μετά το τεστ είναι μικρή (3.67%), ωστόσο είναι πολύ μεγαλύτερη από την πιθανότητα πριν το τεστ (0.2%), δηλαδή το τεστ έδωσε ένα σημαντικό ποσό πληροφορίας. Το επόμενο ερώτημα είναι το ποια θα είναι η πιθανότητα ασθένειας μετά από ένα δεύτερο θετικό τεστ. Για να απαντήσουμε σε αυτό το ερώτημα, χρειαζόμαστε την έννοια της ανεξαρτησίας υπό συνθήκη που παρουσιάζεται παρακάτω (βλ. Παράδειγμα 1.4.8).

Στην επόμενη γραφική παράσταση, φαίνεται η εξέλιξη της τιμής της $P(A|\Theta)$, καθώς αυξάνει το ποσοστό $P(A)$ ασθένειας στον πληθυσμό, για διάφορες τιμές ACC.



Το προηγούμενο σχήμα παράγεται με τον ακόλουθο κώδικα:

```
import numpy as np
import matplotlib.pyplot as plt

p = [0.99, 0.95, 0.9] #accuracy: test result is correct 100*p out of 100 times
def f(p,x): #implementation of Bayes' rule
    return p*x/(p*x + (1-p)*(1-x)) #0<=x<=1 is the percentage of sick people

x = np.linspace(0,1,1001) #1001 evenly spaced values from 0 to 1
y = np.zeros((len(p), len(x))) #len(p) rows and len(x) columns filled with zeros
for i in range(len(p)):
    y[i] = f(p[i],x) #1001 results
    plt.plot(x, y[i], label = "p = %s"%(p[i]))

plt.title("p = Accuracy (test result is correct 100*p out of 100 times)")
plt.xlabel("$P(A)$")
plt.ylabel("$P(A|\Theta)$")
plt.legend()
plt.show()
```

Παράδειγμα 1.4.6. *Σε κάποιο μάθημα εξετάστηκαν* 250 *άτομα, τα οποία τοποθετήθηκαν σε 3 αίθουσες* A, B, Γ*:* 70 *στην* A*,* 90 *στην* B *και τα υπόλοιπα στην* Γ*. Προβιβάσιμο βαθμό πήρε το* 80% *των γραπτών της αίθουσας* A*, το* 88% *της* B *και το* 72% *της* Γ*.*

i) *Αν επιλέξουμε τυχαία έναν εξεταζόμενο ποια είναι η πιθανότητα το γραπτό του να βαθμολογήθηκε κάτω από τη βάση;*

ii) *Αν πάρουμε ένα γραπτό που έχει βαθμολογηθεί πάνω από τη βάση, ποια είναι η πιθανότητα να προέρχεται από την αίθουσα* A*;*

Λύση. Θέτουμε F το ενδεχόμενο να επιλεχθεί εξεταζόμενος που πήρε βαθμό κάτω από την βάση και A, B, Γ τα ενδεχόμενα να προέρχεται από την αίθουσα A, B, Γ αντίστοιχα.

i) Από τον τύπο της ολικής πιθανότητας προκύπτει ότι

$$P(F) = P(F|A)P(A) + P(F|B)P(B) + P(F|\Gamma)P(\Gamma) = 0.2 \cdot \frac{70}{250} + 0.12 \cdot \frac{90}{250} + 0.28 \cdot \frac{90}{250} = 0.2$$

ii) Από τον τύπο του Bayes έχουμε ότι

$$P(A|\overline{F}) = \frac{P(\overline{F}|A)P(A)}{P(\overline{F})} = \frac{P(\overline{F}|A)P(A)}{1 - P(F)} = \frac{0.8 \cdot \frac{70}{250}}{1 - 0.2} = 0.28. \qquad \square$$

1.4.3 Κανόνας του πολλαπλασιασμού

Παρατήρηση. Όπως έχει ήδη αναφερθεί, τα ενδεχόμενα $A, B \subseteq \Omega$ ονομάζονται ανεξάρτητα αν και μόνο αν $P(A \cap B) = P(A)P(B)$. Ο τύπος του Bayes δίνει άλλον έναν χαρακτηρισμό της ανεξαρτησίας δύο ενδεχομένων: Επειδή $P(A|B) = P(A \cap B)/P(B)$ και $P(B|A) = P(A \cap B)/P(A)$, προκύπτει ότι τα ενδεχόμενα A, B είναι ανεξάρτητα αν και μόνο αν είναι $P(A|B) = P(A)$ (ή, ισοδύναμα, $P(B|A) = P(B)$).

Πρόταση 1.12. *Δύο ενδεχόμενα A, B είναι ανεξάρτητα αν και μόνο αν*

$$P(A|B) = P(A) \text{ ή } P(B|A) = P(B).$$

Στην περίπτωση που τα ενδεχόμενα ενός δειγματικού χώρου δεν είναι κατ΄ ανάγκη ανεξάρτητα, ο παρακάτω τύπος είναι πολύ χρήσιμος.

Πρόταση 1.13. *Έστω $A_1, A_2, \ldots, A_n \subseteq \Omega$ με $P(A_1 \cap A_2 \cap \cdots \cap A_{n-1}) > 0$. Τότε*

$$P(\bigcap_{i=1}^{n} A_i) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1})$$

Απόδειξη. Ισχύει ότι

$$\begin{aligned} &P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}) \\ &= P(A_1)\frac{P(A_1 \cap A_2)}{P(A_1)}\frac{P(A_1 \cap A_2 \cap A_3)}{P(A_1 \cap A_2)} \cdots \frac{P(A_1 \cap A_2 \cap \cdots \cap A_n)}{P(A_1 \cap A_2 \cap \cdots \cap A_{n-1})} = P(A_1 \cap A_2 \cap \cdots A_n). \end{aligned}$$ □

Παράδειγμα 1.4.7 (Έλεγχος ποιότητας). *Ένας ελεγκτής ποιότητας εξετάζει μια παρτίδα με 100 τεμάχια ενός προϊόντος επιλέγοντας 5 τεμάχια (χωρίς επανατοποθέτηση). Αν κανένα από τα τεμάχια δεν είναι ελαττωματικό, τότε η παρτίδα γίνεται αποδεκτή, αλλιώς υποβάλλεται σε περαιτέρω ελέγχους. Να βρεθεί η πιθανότητα μια παρτίδα που περιέχει 5 ελαττωματικά τεμάχια να γίνει αποδεκτή.*

Λύση. Έστω A_i το ενδεχόμενο το i-στό τεμάχιο που ελέγχεται να μην είναι ελαττωματικό, $i = 1, 2, 3, 4, 5$.

Μια παρτίδα γίνεται δεκτή αν πραγματοποιηθούν και τα 5 ενδεχόμενα, επομένως μας ενδιαφέρει η πιθανότητα $P(A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5)$.

Από τον κανόνα του πολλαπλασιασμού έχουμε ότι

$$P(A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)P(A_4|A_1 \cap A_2 \cap A_3)P(A_5|A_1 \cap A_2 \cap A_3 \cap A_4).$$

Επίσης,

$$P(A_1) = \frac{95}{100},\ P(A_2|A_1) = \frac{94}{99},\ P(A_3|A_1 \cap A_2) = \frac{93}{98},\ P(A_4|A_1 \cap A_2 \cap A_3) = \frac{92}{97},\ P(A_5|A_1 \cap A_2 \cap A_3 \cap A_4) = \frac{91}{96}.$$

Επομένως,

$$P(A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5) = \frac{95 \cdot 94 \cdot 93 \cdot 92 \cdot 91}{100 \cdot 99 \cdot 98 \cdot 97 \cdot 96} = 0.76959.$$

Παρατήρηση. Στην περίπτωση όπου είχαμε επανατοποθέτηση των τεμαχίων το πρόβλημα θα λυνόταν πιο εύκολα:

$$P(A_1) = P(A_2|A_1) = P(A_3|A_1 \cap A_2) = P(A_4|A_1 \cap A_2 \cap A_3) = P(A_5|A_1 \cap A_2 \cap A_3 \cap A_4) = \frac{95}{100}.$$

Επομένως,

$$P(A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5) = \left(\frac{95}{100}\right)^5 = 0.773781.$$ □

1.4.4 Ανεξαρτησία υπό συνθήκη

Επειδή η δεσμευμένη πιθανότητα είναι συνάρτηση πιθανότητας μπορούμε να μιλάμε και για ανεξαρτησία ενδεχομένων υπό συνθήκη (conditional independence).

Ορισμός (Ανεξαρτησία υπό συνθήκη). *Έστω $\varnothing \neq C \subseteq \Omega$. Τα $A, B \subseteq \Omega$ ονομάζονται **υπό συνθήκη** ή **υπό δέσμευση ανεξάρτητα δεδομένου του** C αν και μόνο αν*

$$P(A \cap B|C) = P(A|C)P(B|C).$$

Παρατηρήστε ότι γενικά

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

Για να είναι τα A, B υπο συνθήκη ανεξάρτητα δεδομένου του C πρέπει

$$P(A|B \cap C) = P(A|C)$$

δηλαδή αν είναι γνωστό ότι το C έχει συμβεί, η επιπλέον πληροφορία ότι το B έχει επίσης συμβεί να μην αλλάξει την πιθανότητα του A.

Παράδειγμα 1.4.8. *Για το παράδειγμα 1.4.5, να βρεθεί η πιθανότητα κάποιος άνθρωπος να πάσχει από αυτή την πάθηση, δεδομένου ότι υποβλήθηκε σε δύο διαδοχικά τεστ, τα οποία βγήκαν θετικά.*

Λύση. Υπενθυμίζεται ότι το ποσοστό του πληθυσμού που έχει την πάθηση είναι $x = 0.002$ και ότι το τεστ δίνει σωστό αποτέλεσμα με πιθανότητα $p = P(\Theta|A) = P(\overline{\Theta}|\overline{A}) = 0.95$. Συμβολίζουμε με Θ_1 και Θ_2 τα ενδεχόμενα να βγουν θετικά το πρώτο και το δεύτερο τεστ αντίστοιχα.

Τα ενδεχόμενα Θ_1 και Θ_2 δεν είναι ανεξάρτητα, αλλά είναι ανεξάρτητα δεδομένου του A, δηλαδή $P(\Theta_1 \cap \Theta_2|A) = P(\Theta_1|A)P(\Theta_2|A) = p^2$. Με απλά λόγια, αν κάποιος που είναι ασθενής υποβληθεί σε διαδοχικά τεστ, το κάθε ένα από αυτά θα βγαίνει θετικό με πιθανότητα $p = 0.95$ και ανεξάρτητα από τα υπόλοιπα.

Ζητάμε την πιθανότητα $P(A|\Theta_1 \cap \Theta_2)$ και χρησιμοποιώντας τους τύπους του Bayes και της υπό συνθήκη ανεξαρτησίας, έχουμε ότι

$$\begin{aligned} P(A|\Theta_1 \cap \Theta_2) &= \frac{P(\Theta_1 \cap \Theta_2|A)P(A)}{P(\Theta_1 \cap \Theta_2)} = \frac{P(\Theta_1 \cap \Theta_2|A)P(A)}{P(\Theta_1 \cap \Theta_2|A)P(A) + P(\Theta_1 \cap \Theta_2|\overline{A})P(\overline{A})} \\ &= \frac{P(\Theta_1|A)P(\Theta_2|A)P(A)}{P(\Theta_1|A)P(\Theta_2|A)P(A) + P(\Theta_1|\overline{A})P(\Theta_2|\overline{A})P(\overline{A})} = \frac{p^2 \cdot x}{p^2 \cdot x + (1-p)^2(1-x)} \approx 0.42. \end{aligned}$$

□

Παράδειγμα 1.4.9. *Από το σύνολο των γυναικών που κάνουν τεστ εγκυμοσύνης μόνο το 12% είναι έγκυες. Έστω ότι ένα τεστ εγκυμοσύνης έχει τις εξής πιθανότητες:*

$$P(\Theta ETIKO|OXI\ EΓKYOΣ) = 1\%, \qquad P(ΑΡΝΗΤΙΚΟ|ΕΓΚΥΟΣ) = 3\%.$$

i) *Να βρεθεί η πιθανότητα να είναι έγκυος μια γυναίκα που κάνει το τεστ και βγαίνει θετικό.*

ii) *Ποιά είναι η αντίστοιχη πιθανότητα να είναι έγκυος για μια γυναίκα που κάνει το τεστ 2 ανεξάρτητες φορές και την πρώτη φορά βγαίνει θετικό ενώ την δεύτερη φορά αρνητικό;*

Λύση. Έστω E το ενδεχόμενο να είναι έγκυος και Θ το ενδεχόμενο το τεστ να βγει θετικό.

i) Από τον τύπο της ολικής πιθανότητας, έχουμε ότι

$$\begin{aligned} P(\Theta) &= P(\Theta|E)P(E) + P(\Theta|\overline{E})P(\overline{E}) \\ &= 0.97 \cdot 0.12 + 0.01 \cdot 0.88 = 0.1252 = 12.52\% \end{aligned}$$

Επομένως,

$$P(\mathrm{E}|\Theta) = \frac{P(\Theta|E)P(\mathrm{E})}{P(\Theta)} = \frac{0.97 \cdot 0.12}{0.1252} = 0.9297$$

ii) Έστω Θ_1, Θ_2 τα ενδεχόμενα την πρώτη και δεύτερη φορά το τεστ να βγει θετικό αντίστοιχα. Από τον τύπο του Bayes ισχύει ότι

$$P(E|\Theta_1 \cap \overline{\Theta_2}) = \frac{P(\Theta_1 \cap \overline{\Theta_2}|E)P(E)}{P(\Theta_1 \cap \overline{\Theta_2})}$$

Λόγω ανεξαρτησίας των Θ_1, $\overline{\Theta_2}$ δεδομένου του ενδεχομένου E έπεται ότι

$$P(\Theta_1 \cap \overline{\Theta_2}|E) = P(\Theta_1|E)P(\overline{\Theta_2}|E) = 0.97 \cdot 0.03 = 0.0291.$$

Επίσης, από τον τύπο της ολικής πιθανότητας, προκύπτει ότι

$$\begin{aligned} P(\Theta_1 \cap \overline{\Theta_2}) &= P(\Theta_1 \cap \overline{\Theta_2}|E)P(E) + P(\Theta_1 \cap \overline{\Theta_2}|\overline{E})P(\overline{E}) \\ &= P(\Theta_1|E)P(\overline{\Theta_2}|E)P(E) + P(\Theta_1|\overline{E})P(\overline{\Theta_2}|\overline{E})P(\overline{E}) \\ &= 0.97 \cdot 0.03 \cdot 0.12 + 0.01 \cdot 0.99 \cdot 0.88 = 0.012204. \end{aligned}$$

Άρα,

$$P(E|\Theta_1 \cap \overline{\Theta_2}) = \frac{0.0291 \cdot 0.12}{0.012204} = 0.286136 = 28.6\%$$ □

1.5 Λυμένες ασκήσεις

Άσκηση 1.1. *Ρίχνουμε τρία αμερόληπτα ζάρια. Να βρεθεί τι είναι πιο πιθανό: οι ενδείξεις να αθροίζουν στο* 11 *ή στο* 12*;*

Λύση. Ο δειγματικός χώρος Ω του πειράματος αποτελείται από τις διατεταγμένες τριάδες των ενδείξεων των τριών ζαριών, δηλαδή

$$\Omega = \{(x, y, z) : 1 \leq x, y, z \leq 6\},$$

οπότε $|\Omega| = 6 \cdot 6 \cdot 6 = 6^3$.

Έστω A το ενδεχόμενο τα τρία ζάρια να αθροίζουν στο 11 και B το ενδεχόμενο τα τρία ζάρια να αθροίζουν στο 12.

Το ενδεχόμενο A αποτελείται από τις 27 παρακάτω ισοπίθανες τριάδες:

$$\begin{array}{lllllllll} (6,4,1) & (6,3,2) & (6,2,3) & (6,1,4) & (5,5,1) & (5,4,2) & (5,3,3) & (5,2,4) & (5,1,5) \\ (4,6,1) & (4,5,2) & (4,4,3) & (4,3,4) & (4,2,5) & (4,1,6) & (3,6,2) & (3,5,3) & (3,3,5) \\ (3,4,4) & (3,2,6) & (2,6,3) & (2,5,4) & (2,4,5) & (2,3,6) & (1,6,4) & (1,5,5) & (1,4,6) \end{array}$$

Επομένως, $P(A) = \dfrac{27}{6^3}$.

Το ενδεχόμενο B αποτελείται από τις 25 παρακάτω ισοπίθανες τριάδες:

$$\begin{array}{lllllllll} (6,5,1) & (6,4,2) & (6,3,3) & (6,2,4) & (6,1,5) & (5,6,1) & (5,5,2) & (5,4,3) & (5,3,4) \\ (5,2,5) & (5,1,6) & (4,6,2) & (4,5,3) & (4,4,4) & (4,3,5) & (4,2,6) & (3,6,3) & (3,5,4) \\ (3,4,5) & (3,3,6) & (2,6,4) & (2,5,5) & (2,4,6) & (1,6,5) & (1,5,6) & & \end{array}$$

Επομένως, $P(B) = \dfrac{25}{6^3}$. Επομένως, είναι πιο πιθανό να φέρουμε 11 από ότι 12.

Παρατήρηση. Το ερώτημα αυτό είχε τεθεί στον Pascal από τον Chevalier de Mere, ο οποίος ενώ εμπειρικά γνώριζε ότι το 11 εμφανίζεται συχνότερα από το 12, εν τούτοις προσπαθώντας να υπολογίσει τις αντίστοιχες πιθανότητες θεωρούσε (**εσφαλμένα**) ως δειγματικό χώρο τις μη διατεταγμένες τριάδες, όπου στο ενδεχόμενο A αντιστοιχούν οι 6 τριάδες:

$$\{6,4,1\} \quad \{6,3,2\} \quad \{5,5,1\} \quad \{5,4,2\} \quad \{5,3,3\} \quad \{4,4,3\}$$

ενώ στο ενδεχόμενο B αντιστοιχούν οι 6 τριάδες:

$$\{6,5,1\} \quad \{6,4,2\} \quad \{6,3,3\} \quad \{5,5,2\} \quad \{5,4,3\} \quad \{4,4,4\}$$

οπότε σύμφωνα με τον Chevalier de Mere οι πιθανότητες θα έπρεπε να είναι ίσες. □

Παρατήρηση. Με γνώσεις Συνδυαστικής, ο πληθάριθμος $|A|$ (αντ. $|B|$) μπορεί να υπολογιστεί ως ο συντελεστής του x^{11} (αντ. x^{12}) στη γεννήτρια συνάρτηση (πολυώνυμο)

$$(x + x^2 + x^3 + x^4 + x^5 + x^6)^3 = x^3(1 + x + x^2 + x^3 + x^4 + x^5)^3 = x^3\left(\frac{1 - x^6}{1 - x}\right)^3$$
$$= x^3 + 3x^4 + 6x^5 + 10x^6 + 15x^7 + 21x^8 + 25x^9 + 27x^{10} + 27x^{11} + 25x^{12} + 21x^{13} + 15x^{14} +$$
$$+ 10x^{15} + 6x^{16} + 3x^{17} + x^{18}.$$

Μπορούμε να υπολογίσουμε τον συντελεστή αυτόν χρησιμοποιώντας κάποιο λογισμικό συμβολικού υπολογισμού (π.χ. Mathematica, Maple, Sage). (Δείτε την απάντηση εδώ.)

Άσκηση 1.2. *Έστω A_1, A_2, A_3 τρία ενδεχόμενα ενός δειγματικού χώρου Ω για τα οποία ισχύει ότι $P(A_1) = \frac{1}{3}$, $P(A_2) = \frac{1}{2}$, $P(A_3) = \frac{1}{4}$. Να εξετασθεί αν τα A_1, A_2, A_3 είναι ανά δύο ξένα.*

Λύση. Έστω ότι τα A_1, A_2, A_3 είναι ανά δύο ξένα. Τότε,

$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) = \frac{1}{3} + \frac{1}{2} + \frac{1}{4} = \frac{4+6+3}{12} > 1.$$

Επομένως, τα A_1, A_2, A_3 δεν είναι ανά δύο ξένα. □

Άσκηση 1.3. *Έστω A, B δύο ενδεχόμενα του ίδιου δειγματικού χώρου Ω. Αν $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{3}$ και $P(A \cup B) = \frac{3}{4}$, να υπολογισθούν οι πιθανότητες $P(A \cap B)$, $P(A \cap \overline{B})$ και $P(\overline{A} \cap \overline{B})$.*

Λύση. Ισχύει ότι

$$P(A \cap B) = P(A) + P(B) - P(A \cup B) = \frac{1}{2} + \frac{1}{3} - \frac{3}{4} = \frac{6}{12} + \frac{4}{12} - \frac{9}{12} = \frac{1}{12}.$$

$$P(A \cap \overline{B}) = P(A) - P(A \cap B) = \frac{1}{2} - \frac{1}{12} = \frac{6}{12} - \frac{1}{12} = \frac{5}{12}.$$

$$P(\overline{A} \cap \overline{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - \frac{3}{4} = \frac{1}{4}.$$ □

Άσκηση 1.4. *Αν $P(A \cup B) = P(A \cap B)$, όπου A, B είναι δύο μη κενά υποσύνολα του δειγματικού χώρου Ω, να δειχθεί ότι*

i) $P(A) = P(B)$

ii) $A = B$.

Λύση.

i) Προφανώς, $A \cap B \subseteq A \subseteq A \cup B$ και $A \cap B \subseteq B \subseteq A \cup B$ οπότε

$$P(A \cap B) \leq P(A) \leq P(A \cup B)$$

και

$$P(A \cap B) \leq P(B) \leq P(A \cup B).$$

Επειδή, $P(A \cap B) = P(A \cup B)$, έπεται ότι

$$P(A \cap B) = P(A) = P(B) = P(A \cup B).$$

ii) Γενικά ισχύει ότι

$$P(A) + P(B) - P(A \cap B) = P(A \cup B)$$

επειδή $P(A \cap B) = P(A \cup B)$ έπεται ότι

$$P(A) - P(A \cap B) + P(B) - P(A \cap B) = 0 \Leftrightarrow P(A \setminus B) + P(B \setminus A) = 0$$

Λόγω μη αρνητικότητας της πιθανότητας έπεται ότι $P(A \setminus B) = P(B \setminus A) = 0$, οπότε $A \setminus B = \varnothing$ και $B \setminus A = \varnothing$, δηλαδή $B \subseteq A$ και $A \subseteq B$, άρα $A = B$. □

Άσκηση 1.5. *Αν A, B είναι δύο ενδεχόμενα του ίδιου δειγματικού χώρου Ω με $P(A) = \frac{1}{3}$ και $P(B) = \frac{3}{8}$, να δειχθεί ότι $\frac{3}{8} \leq P(A \cup B) \leq \frac{17}{24}$.*

Λύση. Επειδή $A, B \subseteq A \cup B$ έπεται ότι $P(A \cup B) \geq P(A)$ και $P(A \cup B) \geq P(B)$, οπότε

$$P(A \cup B) \geq \max\{P(A), P(B)\} = \max\{\frac{1}{3}, \frac{3}{8}\} = \frac{3}{8}.$$

Γνωρίζουμε ότι

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \leq P(A) + P(B) = \frac{1}{3} + \frac{3}{8} = \frac{17}{24}. \qquad \square$$

Άσκηση 1.6. *Αν A, B είναι δύο ενδεχόμενα του ίδιου δειγματικού χώρου Ω με $P(A) = \frac{2}{3}$ και $P(B) = \frac{3}{4}$ να δειχθεί ότι $\frac{5}{12} \leq P(A \cap B) \leq \frac{2}{3}$.*

Λύση. Ισχύει ότι $P(A \cap B) \leq P(A)$ και $P(A \cap B) \leq P(B)$ επομένως

$$P(A \cap B) \leq \min\{P(A), P(B)\} = \min\left\{\frac{2}{3}, \frac{3}{4}\right\} = \frac{2}{3}.$$

Επιπλέον,

$$P(A \cap B) = P(A) + P(B) - P(A \cup B) \geq P(A) + P(B) - 1 = \frac{2}{3} + \frac{3}{4} - 1 = \frac{8}{12} + \frac{9}{12} - \frac{12}{12} = \frac{5}{12}. \qquad \square$$

Εδώ χρησιμοποιήσαμε την ανισότητα $P(A \cup B) \leq 1$, η οποία μας έδωσε χρήσιμη πληροφορία διότι είχαμε $P(A) + P(B) > 1$.

Άσκηση 1.7 (Ανισότητα Bonferroni).

i) *Να δειχθεί ότι για οποιαδήποτε δύο ενδεχόμενα A, B ισχύει ότι*

$$P(A \cap B) \geq P(A) + P(B) - 1$$

ii) *Να δειχθεί ότι για οποιαδήποτε ενδεχόμενα $A_1, A_2, \ldots, A_n$ ισχύει ότι*

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) \geq P(A_1) + P(A_2) + \cdots + P(A_n) - (n-1)$$

Λύση.

i) $P(A \cap B) = P(A) + P(B) - P(A \cup B) \geq P(A) + P(B) - 1$, διότι $P(A \cup B) \leq 1$.

ii) Από τους κανόνες του De Morgan έχουμε ότι

$$\begin{aligned} P(A_1 \cap A_2 \cap \cdots \cap A_n) &= P(\overline{(\overline{A_1} \cup \overline{A_2} \cup \cdots \cup \overline{A_n})}) \\ &= 1 - P(\overline{A_1} \cup \overline{A_2} \cup \cdots \cup \overline{A_n}) \\ &\geq 1 - P(\overline{A_1}) - P(\overline{A_2}) - \cdots - P(\overline{A_n}) \\ &= 1 - (1 - P(A_1)) - (1 - P(A_2)) - \cdots - (1 - P(A_n)) \\ &= P(A_1) + P(A_2) + \cdots + P(A_n) - n + 1 \end{aligned}$$

$\square$

Άσκηση 1.8. *Ένας επιστήμονας έστειλε τα αποτελέσματα της έρευνάς του για δημοσίευση σε ένα από τα γνωστά διεθνή περιοδικά. Η εργασία του φτάνει σε 3 ανεξάρτητους κριτές, οι οποίοι αξιολογούν θετικά με πιθανότητες* $6/11$, $3/7$ *και* $4/9$ *αντίστοιχα. Ποια είναι η πιθανότητα η πλειοψηφία των κριτών να αξιολογήσει θετικά την εργασία του επιστήμονα, οπότε αυτή να δημοσιευθεί;*

Λύση. Έστω A_1, A_2 και A_3 το ενδεχόμενο η εργασία να γίνει αποδεκτή από τον 1ο, 2ο και 3ο κριτή αντίστοιχα και A το ενδεχόμενο η εργασία να γίνει δεκτή από το περιοδικό. Τα ενδεχόμενα A_1, A_2 και A_3 είναι ανεξάρτητα.

Για να γίνει δεκτή η εργασία υπάρχουν 4 ξένες ανά δύο περιπτώσεις, που αντιστοιχούν στους παρακάτω όρους του αθροίσματος:

$$\begin{aligned} P(A) &= P(A_1 \cap A_2 \cap \overline{A_3}) + P(A_1 \cap \overline{A_2} \cap A_3) + P(\overline{A_1} \cap A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3) \\ &= P(A_1)P(A_2)P(\overline{A_3}) + P(A_1)P(\overline{A_2})P(A_3) + P(\overline{A_1})P(A_2)P(A_3) + P(A_1)P(A_2)P(A_3) \\ &= \frac{6\cdot 3\cdot 5}{11\cdot 7\cdot 9} + \frac{6\cdot 4\cdot 4}{11\cdot 7\cdot 9} + \frac{5\cdot 3\cdot 4}{11\cdot 7\cdot 9} + \frac{6\cdot 3\cdot 4}{11\cdot 7\cdot 9} = \frac{106}{231} = 0.4588 = 45.88\% \end{aligned}$$

□

Άσκηση 1.9. *Να βρεθεί ο ελάχιστος αριθμός ατόμων που πρέπει να ρωτήσουμε αν έχουν γενέθλια σήμερα, ώστε η πιθανότητα ένα τουλάχιστον άτομο να έχει γενέθλια σήμερα να είναι τουλάχιστον* $\frac{1}{2}$. *(Υποθέτουμε ότι τα γενέθλια κάθε ατόμου είναι ανεξάρτητα.)*

Λύση. Έστω ότι ρωτάμε n άτομα και A το ενδεχόμενο ένα τουλάχιστον από τα άτομα αυτά να έχει γενέθλια σήμερα.

Θα υπολογίσουμε την πιθανότητα του ενδεχόμενου $\overline{A}$ (κανένα από τα άτομα αυτά να μην έχει γενέθλια σήμερα).

$$P(\overline{A}) = \underbrace{\frac{364}{365}\cdot\frac{364}{365}\cdots\frac{364}{365}}_{n \text{ φορές}} = \left(\frac{364}{365}\right)^n$$

Ψάχνουμε το ελάχιστο n ώστε

$$\begin{aligned} P(A) \geq \frac{1}{2} \Leftrightarrow P(\overline{A}) \leq \frac{1}{2} \Leftrightarrow \left(\frac{364}{365}\right)^n \leq \frac{1}{2} \Leftrightarrow 2 \leq \left(\frac{365}{364}\right)^n \\ \Leftrightarrow \ln 2 \leq n \ln\frac{365}{364} \Leftrightarrow n \geq \frac{\ln 2}{\ln 365 - \ln 364} = 252.652. \end{aligned}$$

Άρα, ο ελάχιστος αριθμός ατόμων που πρέπει να ρωτήσουμε είναι $n = 253$. □

Άσκηση 1.10.

i) *Να βρεθεί ο ελάχιστος αριθμός ρίψεων δύο ζαριών ώστε η πιθανότητα να έρθουν "εξάρες" σ΄ αυτές τις ρίψεις να είναι μεγαλύτερη ή ίση του* $\frac{1}{2}$.

ii) *Τι θα συμβεί αν διπλασιάσουμε τον αριθμό των ρίψεων που υπολογίσαμε στο ερώτημα* i*); Θα διπλασιαστεί η πιθανότητα να έρθουν "εξάρες";*

Λύση.

i) Έστω ότι ρίχνουμε τα δύο ζάρια n φορές και A_n το ενδεχόμενο μια τουλάχιστον φορά στις n ρίψεις να φέρουμε "εξάρες". Ψάχνουμε το ελάχιστο n ώστε $P(A_n) \geq \frac{1}{2}$.

Θα υπολογίσουμε το ενδεχόμενο $\overline{A_n}$: Καμία φορά στις n ρίψεις να μην φέρουμε "εξάρες". Ισχύει ότι

$$P(\overline{A_n}) = \underbrace{\frac{35}{36} \cdot \frac{35}{36} \cdots \frac{35}{36}}_{n \text{ φορές}} = \left(\frac{35}{36}\right)^n$$

επομένως

$$P(A_n) \geq \frac{1}{2} \Leftrightarrow P(\overline{A_n}) \leq \frac{1}{2} \Leftrightarrow \left(\frac{35}{36}\right)^n \leq \frac{1}{2} \Leftrightarrow 2 \leq \left(\frac{36}{35}\right)^n$$
$$\Leftrightarrow \ln 2 \leq n \ln \frac{36}{35} \Leftrightarrow n \geq \frac{\ln 2}{\ln 36 - \ln 35} = 24.605.$$

Άρα, απαιτούνται τουλάχιστον 25 ρίψεις των δύο ζαριών για να φέρουμε "εξάρες" με πιθανότητα τουλάχιστον $\frac{1}{2}$.

ii) Όχι, η πιθανότητα να έρθουν "εξάρες" δεν είναι γραμμική συνάρτηση του αριθμού n των ρίψεων. Συγκεκριμένα, με $2 \cdot 25 = 50$ ρίψεις, η αντίστοιχη πιθανότητα είναι

$$P(A_{50}) = 1 - P(\overline{A_{50}}) = 1 - \left(\frac{35}{36}\right)^{50} = 0.755 = 75.5\% \qquad \square$$

Άσκηση 1.11. *Κατά την ρίψη* 3 *ζαριών οι ενδείξεις τους αθροίζουν στο* 7*. Να βρεθεί η πιθανότητα τουλάχιστον μια από τις ενδείξεις να ισούται με* 1*.*

Λύση. Δεδομένου ότι κάθε στοιχειώδες ενδεχόμενο είναι ισοπίθανο και το άθροισμα των ενδείξεων αθροίζει στο 7, μπορούμε να υπολογίσουμε την ζητούμενη πιθανότητα απαριθμώντας όλες τις περιπτώσεις (ευνοϊκές και μη).

Οι τριάδες που αθροίζουν στο 7 είναι οι εξής: 115, 124, 133, 142, 151, 214, 223, 232, 241, 313, 322, 331, 412, 421, 511. Άρα, υπάρχουν 15 τέτοιες τριάδες.

Από τις παραπάνω, οι τριάδες που δεν περιέχουν 1 είναι οι εξής: 223, 232, 322. Άρα, υπάρχουν 12 τριάδες που περιέχουν 1.

Επομένως, η ζητούμενη πιθανότητα ισούται με $\frac{12}{15} = 0.8$. $\square$

Άσκηση 1.12. *Έστω A, B δύο ενδεχόμενα του ίδιου δειγματικού χώρου Ω. Αν $P(A) = P(B) = \frac{2}{3}$ τότε να δειχθεί ότι $P(A|B) \geq \frac{1}{2}$.*

Λύση. Από τον ορισμό έχουμε ότι

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

Όμως,

$$P(A \cap B) = P(A) + P(B) - P(A \cup B) \geq P(A) + P(B) - 1 = \frac{2}{3} + \frac{2}{3} - \frac{3}{3} = \frac{1}{3}$$

οπότε

$$P(A|B) \geq \frac{\frac{1}{3}}{\frac{2}{3}} = \frac{1}{2}.$$ □

Άσκηση 1.13. *Είναι γνωστό ότι το 5% των οχημάτων, σε ένα συγκεκριμένο τμήμα της εθνικής οδού, κινείται με ταχύτητα άνω του επιτρεπτού ορίου. Έστω ότι η τροχαία των εθνικών οδών χρησιμοποιεί ένα αυτοματοποιημένο σύστημα μέτρησης της ταχύτητας των οχημάτων το οποίο ανιχνεύει τα οχήματα που κινούνται στο συγκεκριμένο τμήμα με ταχύτητα πάνω από επιτρεπτό όριο και το ποσοστό σφάλματος του συστήματος είναι 3% (είτε θετικά, είτε αρνητικά).*

i) *Να βρεθεί η πιθανότητα ένα όχημα να πάρει κλήση από το σύστημα.*

ii) *Να βρεθεί η πιθανότητα ένα όχημα να κινούνταν με υπερβολική ταχύτητα δεδομένου ότι πήρε κλήση από το σύστημα της τροχαίας.*

iii) *Να βρεθεί η πιθανότητα ένα όχημα να κινούνταν με υπερβολική ταχύτητα δεδομένου ότι δεν πήρε κλήση από το σύστημα της τροχαίας.*

Λύση. Έστω A το ενδεχόμενο ένα όχημα να υπερβεί το όριο ταχύτητας και έστω K το ενδεχόμενο ένα όχημα να λάβει κλήση από το σύστημα.

Ισχύει ότι $P(A) = 0.05$, $P(\overline{A}) = 0.95$ και $P(K|A) = 0.97$, $P(\overline{K}|A) = 0.03$, $P(K|\overline{A}) = 0.03$, $P(\overline{K}|\overline{A}) = 0.97$.

i) Από τον τύπο της ολικής πιθανότητας έχουμε ότι

$$P(K) = P(K|A)P(A) + P(K|\overline{A})P(\overline{A}) = 0.97 \cdot 0.05 + 0.03 \cdot 0.95 = 0.077 = 7.7\%$$

Ενώ μόνο το 5% των οχημάτων κάνει παράβαση, εν τούτοις ένα επιπλέον 2.7% λαμβάνει κλήση λόγω σφάλματος.

ii) Από τον τύπο του Bayes έχουμε ότι

$$P(A|K) = \frac{P(K|A)P(A)}{P(K)} = \frac{0.97 \cdot 0.05}{0.077} = 0.62987 = 63\%.$$

iii) Από τον τύπο του Bayes έχουμε ότι

$$P(A|\overline{K}) = \frac{P(\overline{K}|A)P(A)}{P(\overline{K})} = \frac{0.03 \cdot 0.05}{1 - 0.077} = 0.00162 = 0.2\% \ll 1\%.$$ □

Άσκηση 1.14. *Η Αφροδίτη έχει κάνει αίτηση για πρόσληψη στην εταιρεία Megasoft και έχοντας τις παρακάτω πληροφορίες θέλει να κάνει μια εκτίμηση της πιθανότητας να προσληφθεί. Η πιθανότητα να γίνει δεκτή μια αίτηση για πρόσληψη στην εταιρεία είναι* 5%*. Από αυτούς που προσλαμβάνονται το* 90% *έχει περάσει από δεύτερη συνέντευξη, ενώ το* 10% *προσλαμβάνεται κατευθείαν από την πρώτη συνέντευξη. Επίσης, από αυτούς που δεν προσλαμβάνονται μόνο το* 2% *περνάει από δεύτερη συνέντευξη. Ένα τηλεφώνημα καλεί την Αφροδίτη για δεύτερη συνέντευξη στην Megasoft. Τι πιθανότητα έχει να προσληφθεί;*

Λύση. Έστω A το ενδεχόμενο να προσληφθεί και B το ενδεχόμενο να περάσει από δεύτερη συνέντευξη. Από τα δεδομένα του προβλήματος, γνωρίζουμε ότι

$$P(A) = 0.05, P(B|A) = 0.9, P(\overline{B}|A) = 0.1, P(B|\overline{A}) = 0.02$$

Το ζητούμενο είναι η πιθανότητα $P(A|B)$. Από τον τύπο του Bayes έχουμε ότι

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\overline{A})P(\overline{A})} = \frac{0.9 \cdot 0.05}{0.9 \cdot 0.05 + 0.02 \cdot 0.95} = 0.703. \qquad \square$$

Άσκηση 1.15. *Ο Κώστας απαντάει σε ένα τεστ με ερωτήσεις πολλαπλής επιλογής, το οποίο περιέχει* 30 *ερωτήσεις με* 5 *επιλογές, μία σωστή και τέσσερεις λάθος. Ο Κώστας γνωρίζει την απάντηση σε κάποιες από αυτές, ενώ τις υπόλοιπες τις απαντάει στην τύχη. Η πιθανότητα να γνωρίζει μια απάντηση δεδομένου ότι απάντησε σωστά στην ερώτηση είναι* 0.9*.*

i) *Να υπολογισθεί πόσες από τις* 30 *ερωτήσεις αναμένεται να γνώριζε ο Κώστας.*

(Υπόδειξη: Να βρεθεί η πιθανότητα p ο Κώστας να γνωρίζει την απάντηση σε μια ερώτηση.)

ii) *Να υπολογισθεί σε πόσες από τις* 30 *ερωτήσεις απάντησε σωστά ο Κώστας.*

Λύση. Έστω A το ενδεχόμενο ο Κώστας να γνωρίζει την απάντηση σε μια ερώτηση και έστω B το ενδεχόμενο να απαντήσει σωστά στην ερώτηση. Από τα δεδομένα του προβλήματος έχουμε ότι

$$P(B|A) = 1, \quad P(B|\overline{A}) = 1/5 = 0.2, \quad P(A|B) = 0.9$$

i) Έστω $P(A) = p$, οπότε $P(\overline{A}) = 1 - p$. Από τον τύπο του Bayes έχουμε ότι

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} = \frac{P(A)P(B|A)}{P(B|A)P(A) + P(B|\overline{A})P(\overline{A})}.$$

Αντικαθιστώντας από τα δεδομένα του προβλήματος, προκύπτει ότι

$$0.9 = \frac{p \cdot 1}{1 \cdot p + 0.2 \cdot (1 - p)} \Leftrightarrow 0.9 \cdot (0.8 \cdot p + 0.2) = p \Leftrightarrow p = 0.64$$

Άρα, ο Κώστας αναμένεται να γνώριζε την απάντηση σε $30 \cdot 0.64 = 19.2$ ερωτήσεις.

ii) Έστω $P(B) = q$. Από τον τύπο της ολικής πιθανότητας, έχουμε ότι

$$P(B) = P(B|A)P(A) + P(B|\overline{A})P(\overline{A}) = 1 \cdot 0.64 + 0.2 \cdot 0.36 = 0.712$$

Άρα, ο Κώστας αναμένεται να απαντήσει σωστά σε $30 \cdot 0.712 = 21.36$ ερωτήσεις. $\square$

Άσκηση 1.16. *Δέκα ελικόπτερα χρησιμοποιούνται για την αναζήτηση ενός αγνοούμενου ορειβάτη, ο οποίος μπορεί να βρίσκεται σε μία από δύο πιθανές περιοχές με πιθανότητες* 0.7 *και* 0.3 *αντίστοιχα. Κάθε ελικόπτερο μπορεί να χρησιμοποιηθεί σε μία μόνο περιοχή και έχει πιθανότητα* 0.2 *να βρει τον αγνοούμενο, αν αυτός βρίσκεται στην περιοχή που ερευνά. Να βρεθεί πόσα ελικόπτερα πρέπει να σταλούν σε κάθε περιοχή, ώστε η πιθανότητα να βρεθεί ο αγνοούμενος να είναι η μέγιστη δυνατή. Ποια είναι η πιθανότητα να βρεθεί ο αγνοούμενος σε αυτή την περίπτωση; (Υπόδειξη: Υποθέστε ότι* k *ελικόπτερα κατανέμονται στην πρώτη περιοχή και* $10-k$ *ελικόπτερα στην δεύτερη περιοχή.)*

Λύση. Θέτουμε A το ενδεχόμενο να βρεθεί ο ορειβάτης από κάποιο ελικόπτερο, B_1 το ενδεχόμενο ο ορειβάτης να βρίσκεται στην πρώτη περιοχή και B_2 το ενδεχόμενο να βρίσκεται στην δεύτερη περιοχή. Γνωρίζουμε ότι $P(B_1) = 0.7$ και $P(B_2) = 0.3$.

Αν κατανέμονται k ελικόπτερα στην πρώτη περιοχή και $10-k$ ελικόπτερα στην δεύτερη, τότε

$$P(A|B_1) = 1 - P(\overline{A}|B_1) = 1 - 0.8^k \qquad \text{και} \qquad P(A|B_2) = 1 - P(\overline{A}|B_2) = 1 - 0.8^{10-k}.$$

Επομένως, από τον τύπο της ολικής πιθανότητας, έχουμε

$$\begin{aligned} P(A) &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) = (1 - 0.8^k)\cdot 0.7 + (1 - 0.8^{10-k})\cdot 0.3 \\ &= 1 - 0.7\cdot 0.8^k - 0.3\cdot 0.8^{10-k} \end{aligned}$$

Οι τιμές της $P(A)$ για $k = 0, 1, 2, \ldots, 10$ δίδονται στον επόμενο πίνακα:

k	0	1	2	3	4	5	6	7	8	9	10
$P(A)$	0.267	0.399	0.501	0.578	0.634	0.672	0.693	0.699	0.690	0.666	0.624

από όπου προκύπτει ότι η $P(A)$ μεγιστοποιείται όταν 7 ελικόπτερα κατανέμονται στην πρώτη περιοχή (και 3 στη δεύτερη) και η αντίστοιχη πιθανότητα να βρεθεί ο ορειβάτης είναι $0.699 \approx 70\%$. □

Άσκηση 1.17. *Ο Πάνος και η Ελένη σχεδιάζουν χωριστά και μυστικά ο ένας από τον άλλον να πάρουν δώρο για την επέτειό τους. Σκοπεύουν να πληρώσουν με κάρτα από τον κοινό τους λογαριασμό για έκτακτες αγορές που έχει υπόλοιπο* 500 *ευρώ. Αν ο καθένας σχεδιάζει να πάρει δώρα οποιασδήποτε αξίας από* 100 *μέχρι* 300 *ευρώ, να υπολογισθεί η πιθανότητα*

i) *να μην φτάσουν τα χρήματα στον ένα από τους δύο,*

ii) *να μείνουν τουλάχιστον* 100 *ευρώ για δείπνο σε εστιατόριο.*

Λύση. Θέτουμε X, Y τα ποσά που θα ξοδεύσει ο Πάνος και η Μαρία αντίστοιχα. Το ζεύγος (X, Y) μπορεί να θεωρηθεί ως τυχαίο σημείο στο τετράγωνο $[100, 300] \times [100, 300]$.

i) Για μην φθάσουν τα χρήματα, θα πρέπει $X + Y > 500$, το οποίο συμβαίνει για όλα τα σημεία $(X, Y) \in [300, 300]$ που βρίσκονται πάνω από την ευθεία $X + Y = 500$, οπότε

$$P(X + Y > 500) = \frac{100\cdot 100/2}{200^2} = \frac{1}{8} = 0.125.$$

ii) Για να μείνουν τουλάχιστον 100 ευρώ για δείπνο, θα πρέπει $X + Y \leq 400$, το οποίο συμβαίνει για όλα τα σημεία $(X, Y) \in [300, 300]$ που βρίσκονται όχι πάνω από την ευθεία $X + Y = 400$.

$$P(X + Y \leq 400) = 1 - P(X + Y > 400) = 1 - \frac{200\cdot 200/2}{200^2} = 1 - \frac{1}{2} = 0.5.$$ □

Άσκηση 1.18. *Έχουμε n δοχεία, καθένα εκ των οποίων περιέχει a λευκές και b μαύρες μπάλες. Μια μπάλα επιλέγεται τυχαία από το πρώτο δοχείο και μεταφέρεται στο δεύτερο, στην συνέχεια μια μπάλα επιλέγεται τυχαία από το δεύτερο δοχείο και μεταφέρεται στο τρίτο, κ.ο.κ. Στο τέλος, μια μπάλα επιλέγεται τυχαία από το τελευταίο δοχείο. Να βρεθεί η πιθανότητα η μπάλα που επιλέχθηκε να είναι λευκή.*

Λύση. Θέτουμε A_1 το ενδεχόμενο να μεταφέρθηκε λευκή μπάλα από το πρώτο στο δεύτερο δοχείο και A_2 το ενδεχόμενο να επιλεγεί μια λευκή μπάλα από το δεύτερο δοχείο μετά την πρώτη μεταφορά. Ισχύει ότι

$$P(A_1) = \frac{a}{a+b} \text{ και } P(\overline{A_1}) = \frac{b}{a+b}$$

Επίσης

$$P(A_2|A_1) = \frac{a+1}{a+b+1} \text{ και } P(A_2|\overline{A_1}) = \frac{a}{a+b+1}$$

Αφού τα A_1, $\overline{A_1}$ διαμερίζουν τον δειγματικό χώρο, από τον τύπο της ολικής πιθανότητας έχουμε ότι

$$\begin{aligned} P(A_2) &= P(A_2|A_1)P(A_1) + P(A_2|\overline{A_1})P(\overline{A_1}) \\ &= \frac{a+1}{a+b+1} \cdot \frac{a}{a+b} + \frac{a}{a+b+1} \cdot \frac{b}{a+b} \\ &= \frac{a(a+1+b)}{(a+b+1)(a+b)} = \frac{a}{a+b} \end{aligned}$$

Άρα, η πιθανότητα να επιλεγεί μια λευκή μπάλα από το δεύτερο δοχείο μετά από την μεταφορά είναι η ίδια όπως και πριν την μεταφορά. Συνεπώς, και η αντίστοιχη πιθανότητα να επιλεγεί μια λευκή μπάλα από το τρίτο δοχείο είναι επίσης ίδια με πριν, ομοίως για το τέταρτο κ.ο.κ., όπως και για το τελευταίο. Άρα

$$P(A_n) = \frac{a}{a+b}$$

όπου A_n το ενδεχόμενο να επιλεγεί μια λευκή μπάλα από το n-οστό δοχείο. □

Άσκηση 1.19. *Δύο διαφορετικοί αριθμοί επιλέγονται τυχαία από το σύνολο $[n]$. Να βρεθεί η πιθανότητα η διαφορά ανάμεσα στον πρώτο και στον δεύτερο αριθμό να είναι μεγαλύτερη ή ίση με m, όπου $m > 0$.*

Λύση. Έστω H_k το ενδεχόμενο ο πρώτος αριθμός να είναι το k. Προφανώς, $P(H_k) = 1/n$. Επιπλέον, θέτουμε A το ενδεχόμενο η διαφορά ανάμεσα στον πρώτο αριθμό k και τον δεύτερο αριθμό (έστω t) να είναι μεγαλύτερη ή ίση με m (δηλαδή $k - t \geq m$). Ισχύει ότι

$$P(A|H_k) = \begin{cases} \dfrac{k-m}{n-1}, & \text{αν } k = m+1, \ldots, n \\ 0, & \text{αν } k = 1, 2, \ldots, m \end{cases}$$

αφού στην πρώτη περίπτωση ο αριθμός t πρέπει να ανήκει στο σύνολο $\{1, 2, \ldots, k-m\}$, ενώ στην δεύτερη περίπτωση δεν υπάρχει τέτοιος αριθμός t. Τα ενδεχόμενα $H_1, H_2, \ldots, H_n$ αποτελούν διαμέριση του δειγματικού χώρου, οπότε από το θεώρημα ολικής πιθανότητας, ισχύει ότι

$$\begin{aligned} P(A) &= \sum_{k=1}^{n} P(A|H_k)P(H_k) = \sum_{k=m+1}^{n} \frac{k-m}{n-1} \cdot \frac{1}{n} = \frac{1}{n(n-1)} \sum_{k=m+1}^{n} (k-m) \overset{\lambda=k-m}{=} \frac{1}{n(n-1)} \sum_{\lambda=1}^{n-m} \lambda \\ &= \frac{1}{n(n-1)} \cdot \frac{(n-m)(n-m+1)}{2}. \end{aligned}$$

□

Άσκηση 1.20. *Δύο φίλοι κάθονται σε μια σειρά ανθρώπων, ο πρώτος στην αρχή της σειράς και ο δεύτερος στο τέλος της. Ανάμεσά τους κάθονται n άτομα $L_1, L_2, \ldots, L_n$, τα οποία έχουν την τάση να λένε ψέμματα. Αν ακούσουν ΝΑΙ τότε με πιθανότητα $p < 1$ μεταφέρουν στον επόμενο ΟΧΙ και με πιθανότητα $1-p$ μεταφέρουν ΝΑΙ, και αν ακούσουν ΟΧΙ τότε με πιθανότητα p μεταφέρουν ΝΑΙ και με πιθανότητα $1-p$ μεταφέρουν ΟΧΙ. Να βρεθεί η πιθανότητα p_n ο L_n να πει στον δεύτερο φίλο ΝΑΙ δεδομένου ότι ο πρώτος φίλος είπε στον L_1 ΝΑΙ.*

Λύση. Θέτουμε p_i την πιθανότητα ο L_i να πει ΝΑΙ (ανεξαρτήτως τι άκουσε), οπότε $p_1 = 1-p$ και

$$p_i = p_{i-1}(1-p) + (1-p_{i-1})p = (1-2p)p_{i-1} + p, \qquad \text{για κάθε } i \in \{2,3,\ldots,n\}.$$

Επομένως, η ακολουθία (p_i) ικανοποιεί την γραμμική αναγωγική σχέση

$$p_i - (1-2p)p_{i-1} = p, \text{ για } i \geq 2,$$

με αρχική συνθήκη $p_1 = 1-p$.

Από την χαρακτηριστική εξίσωση $x-(1-2p)=0$, προκύπτει ότι η λύση της ομογενούς είναι $q_i = c(1-2p)^i$. Επιπλέον, επειδή p είναι σταθερά που δεν εξαρτάται από το i, η μερική λύση είναι μια σταθερά A. Αντικαθιστώντας στην αναγωγική σχέση της p_i, έχουμε ότι

$$A-(1-2p)A = p \Leftrightarrow A = \frac{1}{2}.$$

Επομένως, η γενική λύση είναι $p_i = q_i + A = c(1-2p)^i + \frac{1}{2}$. Για $i=1$, έχουμε ότι $c(1-2p)+\frac{1}{2} = 1-p \Leftrightarrow c = \frac{1}{2}$, άρα τελικά ισχύει ότι $p_i = \frac{1}{2}((1-2p)^i+1)$.

Επομένως, η ζητούμενη πιθανότητα ισούται με

$$p_n = \frac{1}{2}\left((1-2p)^n+1\right).$$

Καθώς $n \to \infty$ η πιθανότητα $p_n \to 1/2$, δηλαδή καθώς αυξάνει το n χάνεται η πληροφορία που μεταφέρουν οι ενδιάμεσοι αφού στο τέλος όλα είναι ισοπίθανα. □

Άσκηση 1.21. *Μια λοτταρία εκδίδει m λαχεία εκ των οποίων n κερδίζουν (όπου $n < m$). Αν κάποιος αγοράσει τυχαία k λαχεία, ποια είναι η πιθανότητα ένα τουλάχιστον από αυτά να κερδίσει;*

Λύση. Έστω $\overline{A}$ το ενδεχόμενο να μην κερδίσει κανένα από τα λαχεία. Υπάρχουν $\binom{m}{k}$ τρόποι να αγοράσουμε k λαχεία από τα m διαθέσιμα. Επίσης, υπάρχουν $\binom{m-m}{k}$ τρόποι να αγοράσουμε k λαχεία από τα $m-n$ που δεν κερδίζουν. Επομένως, η ζητούμενη πιθανότητα ισούται με

$$P(A) = 1 - P(\overline{A}) = 1 - \frac{\binom{m-n}{k}}{\binom{m}{k}} = 1 - \frac{(m-k)!(m-n)!}{m!(m-n-k)!} = 1 - \frac{(m-n)(m-n-1)(m-n-2)\cdots(m-n-k+1)}{m(m-1)(m-2)\cdots(m-k+1)}$$

Για παράδειγμα αν $m = 10^5$ και $n = 500$, τοτε για $k=2$ η πιθανότητα να κερδίσει είναι 0.009975. □

Άσκηση 1.22. *Έστω ότι κάποια άτομα μιας ομάδας συμπληρώνουν επώνυμα ένα ερωτηματολόγιο και έστω ότι θέλουμε να μάθουμε το ποσοστό των ατόμων της ομάδας τα οποία έχουν μια ιδιότητα, αλλά είτε λόγω προσωπικών δεδομένων είτε επειδή η ερώτηση είναι αδιάκριτη, δεν μπορούμε να ρωτήσουμε ευθέως κάθε άτομο. Μια λύση στο πρόβλημα δίνεται από το παρακάτω πρωτόκολλο, το οποίο βασίζεται στην τυχαιότητα:*

Κάθε άτομο ρίχνει μυστικά ένα νόμισμα.

- *Αν έρθει Κορώνα, τότε πρέπει να απαντήσει ειλικρινώς αν έχει ή όχι την ιδιότητα, με ΝΑΙ ή ΟΧΙ.*
- *Αν έρθει Γράμματα, τότε πρέπει να ξαναρίξει το νόμισμα και αν έρθει Κορώνα να απαντήσει ΝΑΙ, ενώ αν έρθει Γράμματα να απαντήσει ΟΧΙ.*

Να δειχθεί ότι, με βάση τον συνολικό αριθμό των απαντήσεων ΝΑΙ και ΟΧΙ, μπορούμε κατά προσέγγιση να βρούμε το ποσοστό των ανθρώπων που έχουν την ιδιότητα, χωρίς να μπορούμε να γνωρίζουμε για κάθε άτομο αν την έχει ή όχι.

Λύση. Θέτουμε A το ενδεχόμενο ένα άτομο να έχει την ιδιότητα, N το ενδεχόμενο ένα άτομο να απάντησε ΝΑΙ, K_1 το ενδεχόμενο το νόμισμα να φέρει την πρώτη φορά Κορώνα, K_2 το ενδεχόμενο το νόμισμα να φέρει την δεύτερη φορά Κορώνα, Γ_1 το ενδεχόμενα το νόμισμα να φέρει την πρώτη φορά Γράμματα, Γ_2 το ενδεχόμενο το νόμισμα να φέρει την δεύτερη φορά Γράμματα.

Το ενδεχόμενο κάποιο άτομο να απαντήσει ΝΑΙ πραγματοποιείται όταν το άτομο ρίξει ΚΟΡΩΝΑ και έχει την ιδιότητα A, ή όταν το άτομο ρίξει ΓΡΑΜΜΑΤΑ και μετά ξαναρίξει ΚΟΡΩΝΑ. Προφανώς, τα ενδεχόμενα A και K_1 είναι ανεξάρτητα, όπως και τα ενδεχόμενα Γ_1 και K_2. Επομένως, η πιθανότητα $P(N)$ να απαντήσει ΝΑΙ ακολουθώντας τους κανόνες του πρωτοκόλλου είναι

$$\begin{aligned} P(N) &= P(K_1 \cap A) + P(\Gamma_1 \cap K_2) = P(A)P(K_1) + P(\Gamma_1)P(\Gamma_1) \\ &= P(A) \cdot \frac{1}{2} + \frac{1}{4} \end{aligned}$$

και προφανώς είναι συνάρτηση της άγνωστης πιθανότητας $P(A)$.

Επίσης, έστω ότι x άτομα απάντησαν ΝΑΙ και y άτομα απάντησαν ΟΧΙ, οπότε $\dfrac{x}{x+y} \approx P(N)$. (Οι αριθμοί x και y θα διαφέρουν σε διαδοχικές επαναλήψεις του πειράματος, αλλά το κλάσμα $x/(x+y)$ περιμένουμε να είναι πάντα σχετικά κοντά στην πραγματική άγνωστη πιθανότητα $P(N)$.)

Επομένως, λύνοντας ως προς το $P(A)$, προκύπτει ότι

$$P(A) = \frac{4P(N) - 1}{2} \approx \frac{3x - y}{2(x+y)} = \frac{4x - n}{2n}, \qquad \text{όπου } n = x + y.$$

Για παράδειγμα, έστω ότι χρησιμοποιούμε το πρωτόκολλο αυτό και ρωτάμε 100 άτομα. Αν απαντήσουν ΝΑΙ $x = 63$ άτομα (και άρα $y = 37$ απαντήσουν ΟΧΙ), τότε,

$$P(A) \approx \frac{3 \cdot 63 - 37}{200} = \frac{152}{200} = \frac{76}{100},$$

δηλαδή περίπου το 76% των ατόμων έχουν την ιδιότητα αυτή. □

Παρατήρηση. Ο αριθμός x που απάντησαν ΝΑΙ σχηματίζεται ως εξής: Τα μισά περίπου άτομα θα φέρουν την πρώτη φορά ΓΡΑΜΜΑΤΑ και μετά τα μισά από αυτά θα φέρουν ΚΟΡΩΝΑ και τα μισά περίπου άτομα θα φέρουν την πρώτη φορά ΚΟΡΩΝΑ και από αυτά ποσοστό $P(A) \cdot 100\%$ θα απαντήσουν ΝΑΙ. Άρα, $x = \dfrac{x+y}{4} + \dfrac{x+y}{2}P(A) \Leftrightarrow P(A) = \dfrac{3x-y}{2(x+y)}$.

Άσκηση 1.23 (Το πρόβλημα της γραμματέως). *Φανταστείτε ότι εξετάζουμε μέσω συνεντεύξεων ένα σύνολο από n υποψήφιες για μια θέση γραμματέως, και στόχος μας είναι να μεγιστοποίησουμε την πιθανότητα να προσλάβουμε την καλύτερη υποψήφια από αυτό το σύνολο.*

Παρότι δεν έχουμε ιδέα πως να βαθμολογήσουμε την κάθε υποψήφια, μπορούμε να αποφανθούμε εύκολα ποια προτιμάμε συγκριτικά. Επίσης, δεν γνωρίζουμε εκ των προτέρων πόσο καλές είναι οι υποψήφιες.

Οι υποψήφιες προσέρχονται για συνέντευξη με τυχαία σειρά, η μία μετά την άλλη. Μπορούμε να αποφασίσουμε να προσφέρουμε τη θέση σε κάποια υποψήφια οποιαδήποτε στιγμή και είναι εγγυημένο ότι θα δεχθεί τη θέση, οπότε η αναζήτηση θα τερματισθεί. Αν όμως εξετάσουμε κάποια υποψήφια και αποφασίσουμε να μην την προσλάβουμε αλλά να δούμε την επόμενη, την έχουμε χάσει για πάντα.

Όταν ψάχνουμε την καλύτερη γραμματέα υπάρχουν δύο τρόποι που μπορούμε να αποτύχουμε: είτε να σταματήσουμε την αναζήτηση νωρίς, είτε να την σταματήσουμε αργά. Η βέλτιστη λύση έχει τη μορφή ενός κανόνα που ονομάζεται "κοιτάμε και μετά ορμάμε": Καθορίζουμε ένα συγκεκριμένο διάστημα κατά το οποίο θα "κοιτάμε" – δηλαδή θα συγκεντρώνουμε δεδομένα για τις υποψήφιες – χωρις να επιλέξουμε κάποια υποψήφια. Μετά από αυτό το σημείο, περνάμε στη φάση όπου "ορμάμε": Επιλέγουμε οποιαδήποτε υποψήφια είναι καλύτερη από όλες τις υποψήφιες που είδαμε στη φάση όπου κοιτούσαμε.

Καθώς το n μεγαλώνει το ακριβές σημείο μετάβασης από τη φάση όπου κοιτάμε στην φάση όπου ορμάμε σταθεροποιείται στο $1/e = 0.367879 \approx 37\%$ του n. Μια από τις περίεργες μαθηματικές συμμετρίες του προβλήματος είναι ότι το ποσοστό που εκφράζει την στρατηγική συμπίπτει ακριβώς με την πιθανότητα επιτυχίας της. Συγκεκριμένα αποδεικνύεται ότι ακολουθώντας αυτή την στρατηγική έχουμε πιθανότητα $1/e$ να προσλάβουμε την καλύτερη υποψήφια.

α) Να αποδειχθεί ότι ακολουθώντας την παραπάνω στρατηγική αν η πρώτη φάση περιλαμβάνει k από τις n υποψήφιες, τότε η πιθανότητα να επιλέξουμε την καλύτερη υποψήφια ισούται με $\frac{k}{n}(H_{n-1} - H_{k-1})$, όπου $H_n = \sum_{i=1}^{n} \frac{1}{i}$ είναι ο n-στός αρμονικός αριθμός.

β) Να αποδειχθεί ο κανόνας του 37% αν ακολουθήσουμε την παραπάνω στρατηγική.

Λύση. α) Έστω $s_1 < s_2 < \ldots < s_n$ οι n υποψήφιες, με βάση τη σειρά σύγκρισής τους (και όχι με βάση τη σειρά που εμφανίζονται). Υπάρχουν $n!$ τρόποι εμφάνισης των n υποψηφίων. Προκειμένου να επιλέξουμε την s_n πρέπει οι πρώτες k υποψήφιες που εξετάσαμε να μην περιέχουν την s_n και οι υπόλοιπες $n-k$ να έρθουν με τέτοια σειρά ώστε η s_n να είναι η πρώτη που είναι καλύτερη από τις πρώτες k που εξετάσαμε (και άρα οι επιπλέον υποψήφιες που εξετάζουμε μέχρι να φτάσουμε στην s_n να μην είναι καλύτερες από όλες τις k πρώτες). Αν s_j είναι η καλύτερη από τις πρώτες k που εμφανίσθηκαν, όπου $j = k, k+1, \ldots, n-1$, τότε υπάρχουν $\binom{j-1}{k-1}k!$ τρόποι να εμφανισθούν οι k πρώτες υποψήφιες. Για να επιλέξουμε την s_n πρέπει αυτή να εμφανισθεί πριν εξαντληθούν οι υπόλοιπες $j-k$ υποψήφιες που είναι χειρότερες απο την s_j και δεν έχουμε ήδη δει. Αν αυτό συμβεί i εμφανίσεις μετά, όπου $i = 0, 1, \ldots, j-k$, τότε υπάρχουν $\binom{j-k}{i}i!$ τρόποι εμφάνισης των υποψηφίων μέχρι και την s_n και $(n-k-i-1)!$ τρόποι να εμφανισθούν οι επόμενες υποψήφιες που δεν μας απασχολούν διότι ήδη θα έχουμε επιλέξει την s_n.

Αθροίζοντας για όλες τις τιμές των j, i προκύπτει ότι η ζητούμενη πιθανότητα $p_{n,k}$ ισούται με

$$p_{n,k} = \sum_{j=k}^{n-1} \sum_{i=0}^{j-k} \frac{\binom{j-1}{k-1}k!\binom{j-k}{i}i!(n-k-i-1)!}{n!} = \frac{1}{\binom{n}{k}} \sum_{j=k}^{n-1} \frac{1}{n-k}\binom{j-1}{k-1} \sum_{i=0}^{j-k} \frac{\binom{j-k}{i}}{\binom{n-1-k}{i}}$$

Χρησιμοποιώντας τον τύπο (4.1) του Gould [14, σελ. 46]:

$$\sum_{i=a}^{b} \frac{\binom{z}{i}}{\binom{x}{i}} = \frac{x+1}{x-z+1}\left(\frac{\binom{z}{a}}{\binom{x+1}{a}} - \frac{\binom{z}{b+1}}{\binom{x+1}{b+1}}\right)$$

προκύπτει ότι

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

οπότε

$$p_{n,k} = \frac{1}{\binom{n}{k}} \sum_{j=k}^{n-1} \frac{1}{n-k}\binom{j-1}{k-1} \cdot \frac{n-k}{n-j} = \frac{1}{\binom{n}{k}} \sum_{j=k}^{n-1} \frac{1}{n-j}\binom{j-1}{k-1}$$

Χρησιμοποιώντας τον τύπο (1.132) του Gould [14, σελ. 17]:

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

άμεσα προκύπτει ότι

$$p_{n,k} = \frac{(H_{n-1} - H_{k-1})\binom{n-1}{k-1}}{\binom{n}{k}} = \frac{k}{n} \cdot (H_{n-1} - H_{k-1}).$$

όπου $H_n = \sum_{i=1}^{n} \frac{1}{i}$ είναι ο n-στός αρμονικός αριθμός.

β) Ο κανόνας του 37% προκύπτει ως εξής: Από την ανισότητα (6.60) των Graham, Knuth, Patashnik [15, σελ. 277] έχουμε ότι

$$\ln n < H_n < \ln n + 1$$

οπότε

$$\begin{aligned} \ln(n-1) < H_{n-1} < \ln(n-1) + 1 &\Leftrightarrow \ln n\left(1 - \frac{1}{n}\right) < H_{n-1} < \ln n\left(1 - \frac{1}{n}\right) + 1 \\ &\Leftrightarrow \ln n + \ln\left(1 - \frac{1}{n}\right) < H_{n-1} < \ln n + \ln\left(1 - \frac{1}{n}\right) + 1 \end{aligned}$$

και αντίστοιχα

$$-\ln k - \ln\left(1 - \frac{1}{k}\right) - 1 < -H_{k-1} < -\ln k - \ln\left(1 - \frac{1}{k}\right)$$

Με πρόσθεση κατά μέλη προκύπτει ότι

$$-\ln\frac{k}{n} - \ln\frac{1-\frac{1}{k}}{1-\frac{1}{n}} - 1 < H_{n-1} - H_{k-1} < -\ln\frac{k}{n} - \ln\frac{1-\frac{1}{k}}{1-\frac{1}{n}} + 1$$

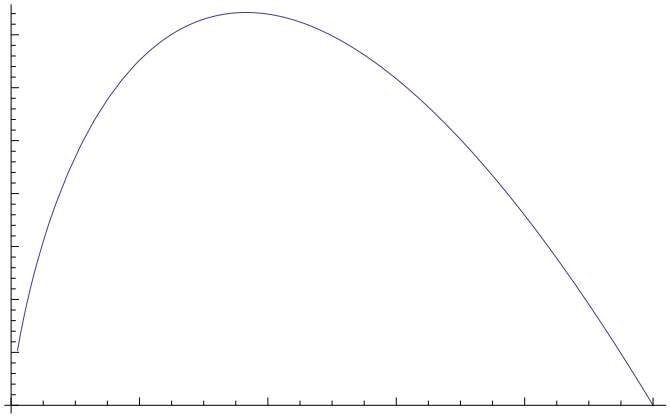
Επομένως,

$$-\frac{k}{n}\ln\frac{k}{n} - \frac{k}{n}\left(\ln\frac{1-\frac{1}{k}}{1-\frac{1}{n}} + 1\right) < p_{n,k} < -\frac{k}{n}\ln\frac{k}{n} - \frac{k}{n}\left(\ln\frac{1-\frac{1}{k}}{1-\frac{1}{n}} - 1\right)$$

Καθώς το n μεγαλώνει η πιθανότητα $p_{n,k} = \frac{k}{n}(H_{n-1} - H_{k-1})$ συγκλίνει στην τιμή της συνάρτησης $-\frac{k}{n}\log\frac{k}{n}$ η οποία έχει μέγιστη τιμή για $\frac{k}{n} = \frac{1}{e}$. Η σύμπτωση ότι η πιθανότητα επιτυχίας είναι $\frac{1}{e}$ προκύπτει επειδή $\log e = 1$. Πράγματι για $\frac{k}{n} = \frac{1}{e}$ έχουμε ότι $-\frac{1}{e}\log\frac{1}{e} = \frac{1}{e}$.

Για παράδειγμα, αν $n = 100$ οι πιθανότητες $p_{100,k}$ δίνονται στον επόμενο πίνακα:

k	$p_{100,k}$	k	$p_{100,k}$	k	$p_{100,k}$	k	$p_{100,k}$	k	$p_{100,k}$
1	0.0517738	21	0.331724	41	0.368522	61	0.303479	81	0.171638
2	0.0835476	22	0.337044	42	0.367267	62	0.29829	82	0.163633
3	0.110321	23	0.34191	43	0.365773	63	0.29294	83	0.155507
4	0.133762	24	0.346341	44	0.364047	64	0.287431	84	0.14726
5	0.154702	25	0.350355	45	0.362093	65	0.281766	85	0.138894
6	0.173643	26	0.353969	46	0.359918	66	0.275947	86	0.13041
7	0.190916	27	0.357199	47	0.357524	67	0.269977	87	0.12181
8	0.206762	28	0.360058	48	0.354919	68	0.263857	88	0.113096
9	0.221357	29	0.36256	49	0.352104	69	0.25759	89	0.104267
10	0.234841	30	0.364717	50	0.349086	70	0.251179	90	0.0953262
11	0.247325	31	0.366541	51	0.345868	71	0.244624	91	0.0862743
12	0.2589	32	0.368042	52	0.342453	72	0.237929	92	0.0771125
13	0.269642	33	0.369231	53	0.338847	73	0.231094	93	0.067842
14	0.279614	34	0.370117	54	0.335051	74	0.224123	94	0.0584639
15	0.288872	35	0.370709	55	0.331071	75	0.217016	95	0.0489795
16	0.297464	36	0.371015	56	0.326909	76	0.209777	96	0.0393898
17	0.30543	37	0.371043	57	0.322568	77	0.202405	97	0.0296959
18	0.312808	38	0.370801	58	0.318051	78	0.194904	98	0.019899
19	0.319631	39	0.370295	59	0.313363	79	0.187275	99	0.01
20	0.325928	40	0.369534	60	0.308504	80	0.179519	100	0.



□

1.6 Ασκήσεις προς επίλυση

1.1) Ποιο είναι το πιο πιθανό αποτέλεσμα για το άθροισμα δύο αμερόληπτων ζαριών; (Απάντηση: 7 με πιθανότητα 1/6.)

1.2) Να δειχθεί ότι αν $A, B, C \subseteq \Omega$ τότε $P(A \cup B \cup C) = P(A) + P(\overline{A} \cap B) + P(\overline{A} \cap \overline{B} \cap C)$.

1.3) Αν $A, B, C \subseteq \Omega$ είναι τρία ανεξάρτητα ανά δύο ενδεχόμενα με πιθανότητες πραγματοποίησης p, q, r αντίστοιχα, να αποδειχθεί ότι $P(A \cup B \cup C) = 1 - (1-p)(1-q)(1-r)$.

1.4) Αν η πιθανότητα να πραγματοποιηθεί ένα ενδεχόμενο είναι τετραπλάσια της πιθανότητας να μην πραγματοποιηθεί, να βρεθεί η πιθανότητα να πραγματοποιηθεί. (Απάντηση: 4/5.)

1.5) Έστω $A, B \subseteq \Omega$ με $P(B) = 1/4$ και $P(A \cup B) = 3/4$. Να βρεθεί η πιθανότητα $P(A)$ ώστε τα A, B να είναι ανεξάρτητα. (Απάντηση: $P(A) = 2/3$.)

1.6) Έστω $A, B \subseteq \Omega$ με $P(A) = 1/3$, $P(B) = 1/2$ και $P(A \cup B) = 2/3$. Να υπολογισθούν οι πιθανότητες $P(A|B)$, $P(B|A)$, $P(\overline{A}|B)$, $P(\overline{B}|A)$, $P(\overline{A}|\overline{B})$, $P(A \cap B|A \cup B)$, $P(A|A \cup B)$ και $P(A|A \cap B)$.

1.7) Έστω ότι θέλουμε να χρησιμοποιήσουμε ένα αμερόληπτο ζάρι αλλά το μόνο που διαθέτουμε είναι ένα αμερόληπτο νόμισμα. Να βρεθεί ένας τρόπος να χρησιμοποιήσουμε το νόμισμα για να προσομοιώσουμε το ζάρι, (δηλαδή να επιλέξουμε με ίση πιθανότητα έναν από έξι αριθμούς).

1.8) Να βρεθεί ο ελάχιστος αριθμός ρίψεων ενός αμερόληπτου ζαριού έτσι ώστε η πιθανότητα να εμφανιστεί 6 τουλάχιστον μια φορά σε αυτές τις ρίψεις να είναι μεγαλύτερη ή ίση από

(α) 2/3, (β) 0.999 (γ) p, $0 \leq p < 1$.

1.9) Ρίχνουμε ένα αμερόληπτο νόμισμα n φορές. Να βρεθεί η πιθανότητα να φέρουμε κορώνα τουλάχιστον σε δύο διαδοχικές ρίψεις.

1.10) Οι διαδοχικές ρίψεις ενός αμερόληπτου νομίσματος είναι ανεξάρτητες, ή όπως λέμε το νόμισμα δεν έχει μνήμη. Αυτό συνήθως δεν συμβαίνει με τον άνθρωπο όταν κατ' επανάληψη κάνει τυχαίες επιλογές. Συγκεκριμένα, έστω ότι ζητάμε από κάποιον να μας δώσει δύο τυχαίες ακολουθίες ρίψεων ενός αμερόληπτου νομίσματος με μήκος 200 η κάθε μια: Την μια να την παράγει χρησιμοποιώντας ένα νόμισμα, και την άλλη να την παράγει επιλέγοντας ο ίδιος με το μυαλό του το αποτέλεσμα κάθε ρίψης. Τότε χωρίς να μας πει ποιά είναι ποιά, μπορούμε με μεγάλη πιθανότητα να διακρίνουμε αυτή που έχει παραχθεί από το νόμισμα, από αυτή που έχει παραχθεί από το μυαλό του.

Ένας απλός τρόπος είναι ο παρακάτω: Θεωρούμε το πείραμα τύχης όπου ρίχνουμε ένα (αμερόληπτο) νόμισμα 200 φορές.

Γράψτε ένα πρόγραμμα που εκτελεί πολλές φορές (π.χ. 100000 φορές) το παραπάνω πείραμα και βρίσκει την σχετική συχνότητα των πειραμάτων που περιέχουν 7 ή περισσότερες διαδοχικές εμφανίσεις της ίδιας όψης.

Με βάση τη συχνότητα που θα βρείτε, θα καταλάβετε ότι η ακολουθία που παράγει το νόμισμα είναι πιθανότερο να είναι αυτή που έχει 7 ή περισσότερες διαδοχικές εμφανίσεις της ίδιας όψης.

Αντίθετα, οι περισσότεροι άνθρωποι δεν σημειώνουν περισσότερες από 4 ή 5 διαδοχικές εμφανίσεις της ίδιας όψης διότι θεωρούν ότι κάτι τέτοιο δεν είναι τυχαίο.

Παραδώστε ηλεκτρονικά τόσο τον κώδικά σας όσο και τα αποτελέσματα των εκτελέσεων του πειράματος.

1.11) Έστω $\varnothing \neq A, B \subseteq \Omega$. Να δειχθεί ότι αν $P(A) > P(B)$ τότε $P(A|B) > P(B|A)$.

1.12) Μπορεί ένα ενδεχόμενο να εξαρτάται από ένα άλλο ενδεχόμενο που έχει πιθανότητα πραγματοποίησης 0; Να αιτιολογηθεί η απάντηση.

1.13) Αν τα A, B, C είναι ανεξάρτητα τότε πρέπει $P(A\cap B\cap C) = P(A)P(B)P(C)$, $P(A\cap B) = P(A)P(B)$, $P(A \cap C) = P(A)P(C)$ και $P(B \cap C) = P(B)P(C)$, δηλαδή υπάρχουν 4 συνθήκες που πρέπει να ελεγχθούν. Πόσες συνθήκες πρέπει να ελεγχθούν για να εξασφαλίσουμε ότι n ενδεχόμενα είναι ανεξάρτητα μεταξύ τους; (Απάντηση: $2^n - n - 1$).

1.14) Από μια κληρωτίδα που περιέχει n λευκά και n μαύρα σφαιρίδια εξάγονται διαδοχικά και χωρίς επανατοποθέτηση δύο σφαιρίδια κάθε φορά μέχρι να εξαχθούν όλα τα σφαιρίδια. Να υπολογισθεί η πιθανότητα όλα τα ζευγάρια που εξάγονται να αποτελούνται από ένα λευκό και ένα μαύρο σφαιρίδιο. (Απάντηση: $2^n/\binom{2n}{n}$.)

1.15) Ένα νόμισμα ρίχνεται $2n$ φορές. Να δειχθεί ότι η πιθανότητα να εμφανισθεί ο ίδιος αριθμός φορών κορώνα και γράμματα ισούται με $\binom{2n}{n}/2^{2n}$. Ποιο είναι το όριο αυτής της πιθανότητας καθώς $n \to \infty$;

1.16) Δύο παίχτες ρίχνουν από n φορές ο καθένας ένα αμερόληπτο νόμισμα. Να βρεθεί η πιθανότητα να εμφανισθεί ο ίδιος αριθμός φορών γράμματα και στους δύο. (Απάντηση: $\binom{2n}{n}/2^{2n}$.)

1.17) Έστω ο δειγματικός χώρος

$$\Omega = \{000, 012, 021, 102, 111, 120, 201, 210, 222\}$$

του οποίου τα στοιχεία θεωρούμε ότι είναι ισοπίθανα. Έστω A_1, A_2 και A_3 τα ενδεχόμενα στην 1η, 2η και 3η θέση αντίστοιχα εμφανίζεται το ψηφίο 0. Να δειχθεί ότι τα A_1, A_2, A_3 είναι ανά δύο ανεξάρτητα, αλλά όχι ανεξάρτητα και τα τρία.

1.18) Επιλέγουμε τυχαία δύο φυσικούς αριθμούς. Να δειχθεί ότι η πιθανότητα p να είναι πρώτοι προς αλλήλους ισούται με $\dfrac{6}{\pi^2}$. (Απάντηση: $\dfrac{1}{p} = \sum\limits_{n=1}^{\infty} \dfrac{1}{n^2} = \dfrac{\pi^2}{6}$.)

1.19) (Το παράδοξο του Galton) Έστω ότι ρίχνουμε 3 αμερόληπτα νομίσματα. Προφανώς, τουλάχιστον 2 θα έχουν την ίδια ένδειξη και το τρίτο νόμισμα θα έχει ίση πιθανότητα να είναι κορώνα ή γράμματα, άρα η πιθανότητα και τα 3 νομίσματα να έχουν την ίδια ένδειξη ισούται με 1/2. Που είναι το σφάλμα στον συλλογισμό αυτό;

1.20) (Polya's Urn) Μια κάλπη περιέχει b μπλε μπάλες και r κόκκινες μπάλες. Μια μπάλα επιλέγεται τυχαία και ξαναρίχνεται πάλι στην κάλπη μαζί με d επιπλέον μπάλες με του ιδιου χρώματος με αυτή. Η διαδικασία αυτή επαναλαμβάνεται συνεχώς.

i) Ποια είναι η πιθανότητα η δεύτερη μπάλα να είναι κόκκινη; (Απάντηση: $\dfrac{r}{b+r}$)

ii) Ποια είναι η πιθανότητα η πρώτη μπάλα να είναι κόκκινη δεδομένου ότι η δεύτερη μπάλα είναι κόκκινη; (Απάντηση: $\dfrac{r+d}{b+r+d}$)

iii) Να δειχθεί οτι η πιθανότητα η μπάλα που εξάγεται την n-οστή φορά να είναι κόκκινη; (Απάντηση: $\dfrac{r}{b+r}$)

1.21) Πώς μπορείς να ελαττώσεις την πιθανότητα να υπάρχει μια βόμβα μέσα στο αεροπλάνο με το οποίο πρόκειται να ταξιδέψεις; Απλώς φέρνεις μια δική σου βόμβα. Η πιθανότητα να υπάρχουν δύο βόμβες μέσα στο ίδιο αεροπλάνο είναι p^2 (όπου p είναι η πιθανότητα να υπάρχει μια βόμβα μέσα στο αεροπλάνο). Αφού το p είναι ήδη μικρό, το p^2 γίνεται αμελητέο. Που είναι το λάθος στον συλλογισμό αυτό;

1.22) Αν $P(A) = a$ και $P(B) = b$, να δειχθεί ότι $P(A|B) \geq \dfrac{a+b-1}{b}$. (Βλέπε λυμένη άσκηση 1.12).

1.23) (Νόμος του Murphy) Ένα αμερόληπτο νόμισμα ρίχνεται ξανά και ξανά. Έστω s μια οποιαδήποτε σταθερή ακολουθία από Κορώνες και Γράμματα με μήκος r. Να δειχθεί ότι με πιθανότητα 1 η ακολουθία s θα εμφανισθεί σε r διαδοχικές ρίψεις του νομίσματος. (Η συνήθης διατύπωση του νόμου του Murphy είναι ότι αν κάτι μπορεί να πάει στραβά, τότε σίγουρα θα πάει στραβά.)

1.24) Έστω A, B, C, D οι κορυφές ενός τετραέδρου. Ένα μυρμήγκι βρίσκεται αρχικά στην κορυφή A. Στη συνέχεια επιλέγει τυχαία με ίση πιθανότητα μια από τις ακμές με άκρο την A και μεταβαίνει στην επόμενη κορυφή. Συνεχίζει με αυτό τον τρόπο και στις υπόλοιπες κορυφές. Ποια είναι η πιθανότητα να βρίσκεται πάλι στην κορυφή A αφού έχει διανύσει n ακμές;

1.25) Έστω ότι n αντικείμενα είναι τοποθετημένα σε μια σειρά. Η πράξη S_k ορίζεται ως εξής: Διάλεξε τυχαία ένα από τα πρώτα k αντικείμενα και αντάλλαξε τη θέση του με το αντικείμενο που βρίσκεται στην θέση k. Εκτελούμε τις πράξεις S_n, S_{n-1}, ..., S_1. Να δειχθεί ότι η τελική τοποθέτηση των αντικειμένων που προκύπτει είναι εξίσου πιθανή να είναι οποιαδήποτε από τις $n!$ μεταθέσεις των αντικειμένων.

1.26) i) n σημεία επιλέγονται ομοιόμορφα στην περιφέρεια ενός κύκλου. Να δειχθεί ότι η πιθανότητα να υπάρχει ένα ημικύκλιο της περιφέρειας που δεν περιέχει κανένα από τα n σημεία ισούται με $\dfrac{2n}{2^n}$.

ii) n σημεία επιλέγονται ομοιόμορφα στην επιφάνεια μιας σφαίρας. Να δειχθεί ότι η πιθανότητα να υπάρχει ένα ημισφαίριο της επιφάνειας που δεν περιέχει κανένα από τα n σημεία ισούται με $\dfrac{n^2-n+2}{2^n}$.

1.27) Έστω μια ομάδα n ατόμων, όπου $2n < 365$. Να δειχθεί ότι η πιθανότητα p_n δύο τουλάχιστον άτομα της ομάδας να έχουν την ίδια μέρα γενέθλια ή σε διαδοχικές μέρες ισούται με

$$p_n = 1 - 365^{-n+1}\frac{(365-n-1)!}{(365-2n)!}.$$

Να βρεθεί ποιος είναι ο ελάχιστος αριθμός ατόμων n ώστε $p_n \geq 1/2$. (Απάντηση. $n = 14$, $p_{14} = 0.537493$)

1.28) Ένας μεγάλος αριθμός φοιτητών σε μια αίθουσα ερωτώνται να αναφέρουν την ημερομηνία των γενεθλίων τους. Ο πρώτος φοιτητής για τον οποίο τα γενέθλια του συμπίπτουν με τα γενέθλια κάποιου από τους ήδη ερωτηθέντες φοιτητές κερδίζει ένα δώρο. Να δειχθεί ότι αν βρίσκεστε στην αίθουσα, η πιθανότητα να κερδίσετε γίνεται μέγιστη αν είστε το 20ο άτομο που θα ερωτηθεί. Στην περίπτωση αυτή πόσο μεγάλη είναι αυτή η πιθανότητα;

Λύση. Αν ήμασταν το n-οστό άτομο που ρωτήθηκε, το ενδεχόμενο A_n να κερδίσουμε πραγματοποιείται όταν στα $n-1$ άτομα που προηγήθηκαν δεν υπάρχουν δύο άτομα που να έχουν την ίδια μέρα γενέθλια (πρώτο ενδεχόμενο) και τα γενέθλια μας συμπίπτουν με ένα από τα γενέθλια των προηγούμενων $n-1$ ατόμων (δεύτερο ενδεχόμενο). Λόγω ανεξαρτησίας των δύο ενδεχομένων η πιθανότητα αυτή ισούται με $p_n = \frac{n-1}{365} \prod_{i=0}^{n-2} \left(1 - \frac{i}{365}\right)$

Οι τιμές της p_n για $n = 2$ έως 80 δίδονται στο επόμενο πίνακα:

2.	0.00273973	21.	0.03225	41.	0.0119198	61.	0.000966138
3.	0.00546444	22.	0.032007	42.	0.0108789	62.	0.000820776
4.	0.00815175	23.	0.0316019	43.	0.00989238	63.	0.000694812
5.	0.0107797	24.	0.031047	44.	0.00896251	64.	0.000586092
6.	0.0133269	25.	0.0303554	45.	0.00809053	65.	0.000492628
7.	0.0157732	26.	0.0295411	46.	0.00727694	66.	0.000412597
8.	0.0180996	27.	0.0286185	47.	0.00652156	67.	0.000344338
9.	0.0202885	28.	0.0276022	48.	0.00582357	68.	0.000286348
10.	0.0223243	29.	0.0265071	49.	0.00518164	69.	0.000237275
11.	0.0241932	30.	0.0253477	50.	0.00459397	70.	0.00019591
12.	0.0258834	31.	0.0241384	51.	0.00405841	71.	0.000161177
13.	0.0273855	32.	0.0228929	52.	0.00357252	72.	0.000132127
14.	0.0286922	33.	0.0216243	53.	0.0031336	73.	0.000107925
15.	0.0297988	34.	0.020345	54.	0.00273885	74.	0.0000878389
16.	0.0307027	35.	0.0190664	55.	0.00238533	75.	0.0000712337
17.	0.0314037	36.	0.0177989	56.	0.00207007	76.	0.0000575593
18.	0.0319038	37.	0.0165519	57.	0.0017901	77.	0.0000463418
19.	0.0322071	38.	0.0153338	58.	0.00154252	78.	0.0000371753
20.	0.0323199	39.	0.0141518	59.	0.00132447	79.	0.0000297138
		40.	0.0130121	60.	0.00113321	80.	0.0000236636

Η μέγιστη τιμή προκύπτει όταν $n = 20$ και ισούται με $0.0323199 \approx 3.23\%$ □

1.29) Σε ένα αεροπλάνο n αριθμένων θέσεων ετοιμάζονται να επιβιβαστούν n επιβάτες. Κάθε επιβάτης γνωρίζει τον αριθμό της θέσης του, όμως ο πρώτος επιβάτης επιλέγει να καθίσει τυχαία σε μια θέση. Στη συνέχεια, κάθε ένας από τους επόμενους επιβάτες, κάθεται στη θέση του εκτός αν είναι ήδη πιασμένη άλλον, οπότε κάθεται τυχαία σε μια από τις άδειες θέσεις.

Να βρεθεί η πιθανότητα p_n ο τελευταίος επιβάτης να καθίσει στην θέση του.

Αν δεν μπορείτε να βρείτε τύπο για την πιθανότητα p_n, γράψτε ένα πρόγραμμα που προσομοιώνει το σενάριο αυτό για $n = 100$, και με τη βοήθεια του οποίου προσεγγίστε πειραματικά την πιθανότητα p_{100}.

Παραδώστε ηλεκτρονικά τόσο τον κώδικά σας όσο και τα αποτελέσματα των εκτελέσεων του προγράμματος.

1.30) Μας δίνονται τρία νομίσματα: το ένα έχει κορώνα και στις δύο πλευρές, το δεύτερο έχει γράμματα και στις δύο πλευρές και το τρίτο έχει κορώνα στη μια πλευρά και γράμματα στην άλλη. Επιλέγουμε ένα νόμισμα στην τύχη, το ρίχνουμε και έρχεται κορώνα. Ποια είναι η πιθανότητα ότι η άλλη πλευρά είναι γράμματα;

1.31) Έστω ότι A, B είναι ενδεχόμενα με $P(A), P(B) > 0$. Λέμε ότι το B *υποστηρίζει* το A αν και μόνο αν $P(A|B) > P(A)$ και ότι το B *δεν υποστηρίζει* το A αν και μόνο αν $P(A|B) < P(A)$.

i) Να δειχθεί ότι αν το B υποστηρίζει το A, τότε και το A υποστηρίζει το B.

ii) Έστω $P(\overline{B}) > 0$. Να δειχθεί ότι το B υποστηρίζει το A αν και μόνο αν $P(\overline{B})$ δεν υποστηρίζει το A.

Κεφάλαιο 2

Διακριτές τυχαίες μεταβλητές

2.1 Εισαγωγή

Παράδειγμα. Θεωρούμε το εξής πείραμα: Ρίχνουμε ένα νόμισμα 4 φορές. Συμβολίζουμε με

- X τον αριθμό των εμφανίσεων της όψης ΓΡΑΜΜΑΤΑ.
- Y τον αριθμό της ρίψης που εμφανίσθηκε πρώτη φορά η όψη ΓΡΑΜΜΑΤΑ.
- Z τη διαφορά του αριθμού των εμφανίσεων της όψης ΓΡΑΜΜΑΤΑ και του αριθμού των εμφανίσεων της όψης ΚΟΡΩΝΑ, δηλαδή $Z = X - (4 - X) = 2X - 4$.

Οι τιμές των X, Y, Z δεν είναι γνωστές εκ των προτέρων, αλλά εξαρτώνται από την έκβαση του πειράματος.

Οι X, Y, Z αποτελούν παραδείγματα τυχαίων μεταβλητών. Διαισθητικά, μια τυχαία μεταβλητή είναι μια ποσότητα που εξαρτάται από την έκβαση ενός τυχαίου πειράματος.

Ορισμός. *Έστω Ω ένας δειγματικός χώρος. Αν σε κάθε στοιχειώδες ενδεχόμενο $\{\omega\}$ του Ω αντιστοιχίσουμε ένα πραγματικό αριθμό, τότε ορίζεται μια συνάρτηση $X : \Omega \to \mathbb{R}$, η οποία ονομάζεται **τυχαία μεταβλητή** (ΤΜ). (Συνήθως, οι ΤΜ συμβολίζονται με κεφαλαία γράμματα.)*

*Το **σύνολο τιμών** (range) της X συμβολίζεται με S_X ή απλά με S, δηλαδή $S_X := X(\Omega)$.*

*Αν το S_X είναι το πολύ αριθμήσιμο, τότε η X ονομάζεται **διακριτή** ΤΜ.*

Το ενδεχόμενο μια ΤΜ X να λάβει την τιμή $x \in S_X$ συμβολίζεται με $\{X = x\}$, δηλαδή

$$\{X = x\} := X^{-1}(x) := \{\omega \in \Omega : X(\omega) = x\},$$

και η πιθανότητα του ενδεχομένου αυτού με $P(X = x)$. (Η πιθανότητα $P(X = x)$ ορίζεται βάσει του μέτρου πιθανότητας που έχει ήδη ορισθεί στο Ω.) Πιο γενικά, συμβολίζουμε με $\{X \in T\}$ το ενδεχόμενο $\bigcup_{x \in T}\{X = x\}$, για κάθε $T \subseteq S_X$, δηλαδή είναι

$$\{X \in T\} := X^{-1}(T) := \{\omega \in \Omega : X(\omega) \in T\},$$

και την αντίστοιχη πιθανότητα με $P(X \in T)$.

Οι ΤΜ X, Y, Z του παραδείγματος λαμβάνουν τις παρακάτω τιμές, για κάθε στοιχειώδες ενδεχόμενο του δειγματικού χώρου Ω (πρώτη στήλη του πίνακα):

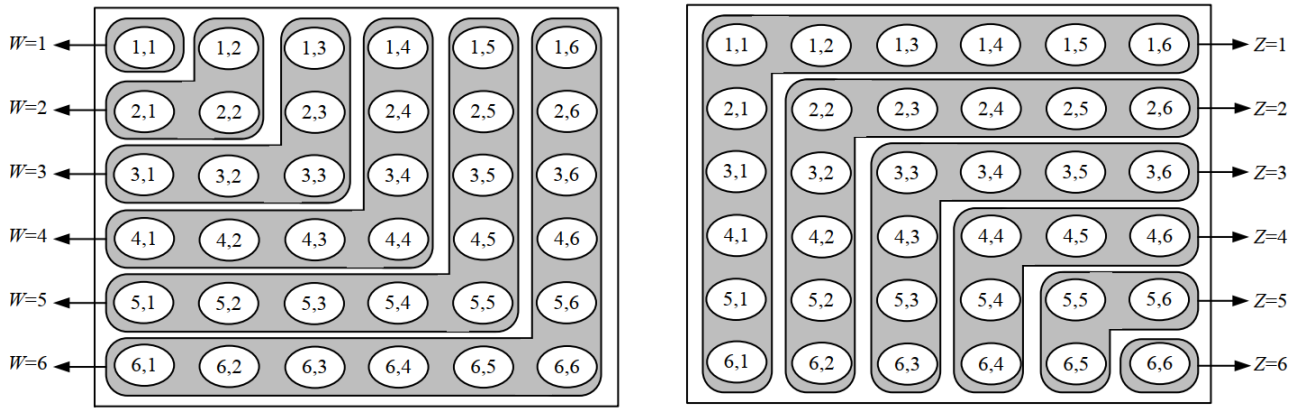
$\omega \in \Omega$	X	Y	Z
ΚΚΚΚ	0	0	-4
ΚΚΚΓ	1	4	-2
ΚΚΓΚ	1	3	-2
ΚΚΓΓ	2	3	0
ΚΓΚΚ	1	2	-2
ΚΓΚΓ	2	2	0
ΚΓΓΚ	2	2	0
ΚΓΓΓ	3	2	2

$\omega \in \Omega$	X	Y	Z
ΓΚΚΚ	1	1	-2
ΓΚΚΓ	2	1	0
ΓΚΓΚ	2	1	0
ΓΚΓΓ	3	1	2
ΓΓΚΚ	2	1	0
ΓΓΚΓ	3	1	2
ΓΓΓΚ	3	1	2
ΓΓΓΓ	4	1	4

Το σύνολο τιμών (δυνατές τιμές) της X είναι $S_X = \{0, 1, 2, 3, 4\}$.
Το σύνολο τιμών της Y είναι $S_Y = \{0, 1, 2, 3, 4\}$.
Το σύνολο τιμών της Z είναι $S_Z = \{-4, -2, 0, 2, 4\}$.

Η X λαμβάνει την τιμή 0 αν δεν εμφανίσθηκε καθόλου η όψη ΓΡΑΜΜΑΤΑ, δηλαδή μόνο για το ενδεχόμενο ΚΚΚΚ, το οποίο εμφανίζεται με πιθανότητα $\frac{1}{2^4}$. Η X λαμβάνει την τιμή 1 αν εμφανίσθηκε ακριβώς μια όψη ΓΡΑΜΜΑΤΑ, δηλαδή για τα ενδεχόμενα ΓΚΚΚ, ΚΓΚΚ, ΚΚΓΚ, ΚΚΚΓ, καθένα εκ των οποίων εμφανίζεται με πιθανότητα $\frac{1}{16}$, άρα η X λαμβάνει την τιμή 1 με πιθανότητα $\frac{4}{16} = \frac{1}{4}$.

Παράδειγμα. Ρίχνουμε δύο ζάρια και συμβολίζουμε με X, Y τα αντίστοιχα αποτελέσματα των 2 ρίψεων. Ο δειγματικός χώρος είναι ο $\Omega = \{(x, y) : x, y \in [6]\}$. Προφανώς οι X, Y είναι ΤΜ, με $S_X = S_Y = [6]$. Επιπλέον ορίζουμε τις ΤΜ $W = \max\{X, Y\}$ και $Z = \min\{X, Y\}$. Προφανώς, είναι $S_W = S_Z = [6]$. Τα ενδεχόμενα $\{W = k\}$ και $\{Z = k\}$ αναπαριστάνονται στο επόμενο σχήμα (ως υποσύνολα του Ω) για κάθε $k \in [6]$.



Παρατηρήστε ότι η οικογένεια $(\{W = k\})_{k\in[6]}$ διαμερίζει το Ω, όπως και η οικογένεια $(\{Z = k\})_{k\in[6]}$. Με συμβολισμό συναρτήσεων, οι διαμερίσεις αυτές είναι οι $(W^{-1}(k))_{k\in[6]}$ και $(Z^{-1}(k))_{k\in[6]}$ αντίστοιχα.

Ορισμός. *Έστω μια διακριτή ΤΜ X με σύνολο τιμών S. Η συνάρτηση $f_X(x) = P(X = x)$ με πεδίο ορισμού το S, ονομάζεται* ***συνάρτηση μάζας πιθανότητας*** *(probability mass function - PMF) ή* ***μάζα*** *της ΤΜ X. (Συνήθως γράφουμε $f(x)$ αντί για $f_X(x)$ όταν δεν υπάρχει κίνδυνος σύγχυσης.)*

Παράδειγμα 2.1.1. *Να βρεθεί η PMF των ΤΜ X, Y, Z του πρώτου παραδείγματος.*

Λύση. Για την X, είναι

$$f_X(0) = P(X = 0) = \frac{1}{16}, \quad f_X(1) = P(X = 1) = \frac{4}{16}, \quad f_X(2) = P(X = 2) = \frac{6}{16},$$

$$f_X(3) = P(X = 3) = \frac{4}{16}, \quad f_X(4) = P(X = 4) = \frac{1}{16}.$$

Για την Y, είναι

$$f_Y(0) = P(Y = 0) = \frac{1}{16} \qquad \text{και} \qquad f_Y(x) = P(Y = x) = \frac{1}{2^{x-1}}\frac{1}{2} = 2^{-x}, \qquad x \in \{1, 2, 3, 4\}$$

Για την Z, είναι

$$f_Z(x) = P(Z = x) = P(2X - 4 = x) = P(X = 2 + x/2) = f_X(2 + x/2), \quad x \in \{-4, -2, 0, 2, 4\}.$$

□

Πρόταση 2.1. *Η PMF f_X μιας διακριτής ΤΜ X με σύνολο τιμών S_X έχει τις εξής ιδιότητες:*

- $0 \le f_X(x) \le 1$, *για κάθε* $x \in S_X$.
- $\sum_{x \in S_X} f_X(x) = 1$.
- $P(X \in T) = \sum_{x \in T} f_X(x)$, *για κάθε* $T \subseteq S_X$.

Απόδειξη. Η πρώτη ανισότητα είναι προφανής, αφού εξ ορισμού είναι $f_X(x) = P(X = x)$ και οποιαδήποτε πιθανότητα παίρνει εξ ορισμού τιμές από 0 έως 1.

Για τη δεύτερη ιδιότητα, παρατηρούμε ότι η οικογένεια $(X^{-1}(x))_{x \in S_X}$ αποτελεί διαμέριση του Ω, επομένως

$$\sum_{x \in S_X} f_X(x) = \sum_{x \in S_X} P(X = x) = \sum_{x \in S_X} P(X^{-1}(x)) = P(\Omega) = 1.$$

Η απόδειξη της τρίτης ιδιότητας είναι ανάλογη και αφήνεται ως άσκηση. □

Το επόμενο αποτέλεσμα αποτελεί τον τύπο ολικής πιθανότητας για διακριτές τυχαίες μεταβλητές.

Λήμμα 2.2. *Αν X, Y διακριτές ΤΜ, τότε*

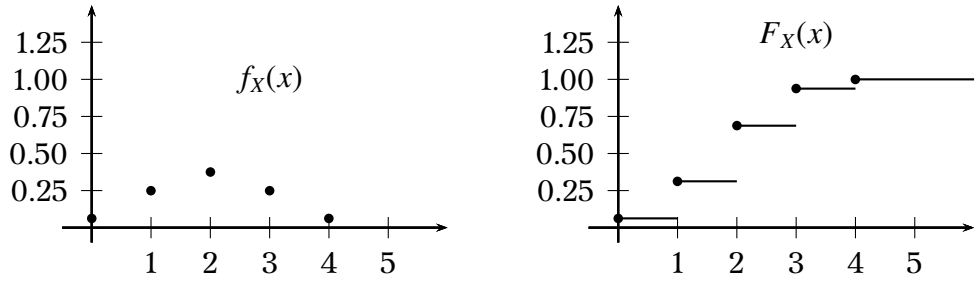
$$P(X = x) = \sum_{y \in S_Y} P(X = x, Y = y), \qquad \text{για κάθε } x \in S_X.$$

Απόδειξη. Το ενδεχόμενο $A = \{X = x\}$ διαμερίζεται στα $A_y = A \cap B_y$, $y \in S_Y$, όπου $B_y = \{Y = y\}$. Επομένως,

$$P(X = x) = P(A) = \sum_{y \in S_Y} P(A \cap B_y) = \sum_{y \in S_Y} P(X = x, Y = y)$$

□

Ορισμός. *Η (αθροιστική) συνάρτηση κατανομής (cumulative distribution function - CDF) της ΤΜ X είναι η συνάρτηση $F_X : \mathbb{R} \to [0,1]$, με $F_X(x) = P(X \leq x)$ για κάθε $x \in \mathbb{R}$,*



Σχήμα 2.1: Η συναρτήσεις f_X και F_X για την ΤΜ X του προηγούμενου παραδείγματος.

Για την (αθροιστική) συνάρτηση κατανομής πιθανότητας $F(x)$ ισχύουν οι παρακάτω ιδιότητες

i) Η $F(x)$ είναι αύξουσα.

ii) Η $F(x)$ είναι συνεχής από τα δεξιά, δηλαδή $\lim_{x \to x_0^+} F(x) = F(x_0)$ για κάθε $x_0 \in \mathbb{R}$.

iii) $\lim_{x \to \infty} F(x) = 1$ και $\lim_{x \to -\infty} F(x) = 0$.

iv) $P(X < x_0) = \lim_{x \to x_0^-} F(x)$.

v) $P(X > x) = 1 - P(X \leq x) = 1 - F(x)$.

vi) $P(x_1 < X \leq x_2) = P(X \leq x_2) - P(X \leq x_1) = F(x_2) - F(x_1)$.

2.2 Αναμενόμενη τιμή και διακύμανση

Ορισμός. *Η **μέση** (ή προσδοκώμενη, ή αναμενόμενη) **τιμή** (expected value) μιας διακριτής ΤΜ X με PMF f_X ορίζεται ως ο αριθμός $E(X)$ με*

$$E(X) := \sum_{x \in S_X} xP(X = x) = \sum_{x \in S_X} x f_X(x).$$

Γενικότερα, για οποιαδήποτε συνάρτηση $g : S_X \to \mathbb{R}$, η μέση τιμή της νέας ΤΜ $g(X)$ δίνεται από τη σχέση

$$E(g(X)) = \sum_{x \in S_X} g(x)P(X = x) = \sum_{x \in S_X} g(x) f_X(x).$$

Παρατηρήσεις. Η μέση τιμή μιας ΤΜ μπορεί και να απειρίζεται (ή και να μην ορίζεται, όταν το άθροισμα του ορισμού δεν συγκλίνει). Στο εξής, θα θεωρούμε πάντα ότι η μέση τιμή είναι πεπερασμένη, εκτός αν αναφέρεται ρητά το αντίθετο.

Ο δεύτερος τύπος είναι άμεση συνέπεια του πρώτου. Πράγματι, η ΤΜ $Y = g(X)$ ορίζεται στον (το πολύ αριθμήσιμο) δειγματικό χώρο S_X, ο οποίος διαμερίζεται από την οικογένεια $(g^{-1}(y))_{y \in S_y}$ και έχει μέτρο πιθανότητας την f_X, επομένως

$$\begin{aligned} E(Y) &= \sum_{y \in S_Y} yP(Y = y) = \sum_{y \in S_Y} y \sum_{x \in g^{-1}(y)} f_X(x) \\ &= \sum_{y \in S_Y} \sum_{x \in g^{-1}(y)} g(x) f_X(x) = \sum_{x \in S_X} g(x) f_X(x). \end{aligned}$$

Ορισμός. *Η **διακύμανση** (ή διασπορά) (variance) μιας διακριτής ΤΜ X, με $E(X) = \mu \in \mathbb{R}$, ορίζεται ως ο αριθμός $\sigma^2 = V(X)$ με*

$$V(X) = E[(X - \mu)^2].$$

*Ο μη αρνητικός αριθμός $\sigma = \sqrt{V(X)}$ ονομάζεται **τυπική απόκλιση** (standard deviation) της X.*

Παρατηρήσεις. Παρατηρήστε ότι η διακύμανση της X είναι ουσιαστικά η μέση τιμή της ΤΜ

$$Y = (X - E(X))^2.$$

Επομένως, η διακύμανση είναι ένα μέτρο που μας δείχνει πόσο συγκεντρωμένες βρίσκονται οι τιμές μιας ΤΜ γύρω από την μέση τιμή της. Αφού η Y παίρνει μη αρνητικές τιμές, η μέση τιμή της θα είναι προφανώς μη αρνητική, οπότε η τυπική απόκλιση σ ορίζεται πάντα.

Η διακύμανση μιας ΤΜ μπορεί επίσης να απειρίζεται (ή και να μην ορίζεται). Στο εξής, θα θεωρούμε πάντα ότι είναι πεπερασμένη, εκτός αν αναφέρεται ρητά το αντίθετο.

Πρόταση 2.3. *Για οποιεσδήποτε διακριτές ΤΜ X, Y και σταθερές $a, b \in \mathbb{R}$, ισχύουν οι παρακάτω ιδιότητες:*

1. $E(aX + bY) = aE(X) + bE(Y)$
2. $V(aX + b) = a^2V(X)$.
3. $V(X) = E(X^2) - (E(X))^2$.

Απόδειξη.

1. Έστω η ΤΜ $Z = aX + bY$. Τότε είναι $S_Z = \{ax + by : x \in S_X, y \in S_Y\}$ και $P(Z = z) = P(X = x, Y = y)$, για κάθε $z = ax + by \in S_Z$. Επομένως, βάσει του Λήμματος 2.2, προκύπτει ότι
$$\begin{aligned} E[Z] &= \sum_{z \in S_Z} zP(Z = z) = \sum_{x \in S_X} \sum_{y \in S_Y} (ax + by)P(X = x, Y = y) \\ &= a \sum_{x \in S_X} x \sum_{y \in S_Y} P(X = x, Y = y) + b \sum_{y \in S_Y} y \sum_{x \in S_X} P(X = x, Y = y) \\ &= a \sum_{x \in S_X} xP(X = x) + b \sum_{y \in S_Y} yP(Y = y) = aE(X) + bE(Y). \end{aligned}$$
2. Βάσει του προηγούμενου αποτελέσματος, ισχύει ότι $E(aX + b) = aE(X) + b = a\mu + b$, οπότε
$$\mathrm{V}(aX + b) = E((aX + b - (a\mu + b))^2) = E(a^2(X - \mu)^2) = a^2E((X - \mu)^2) = a^2\mathrm{V}(X).$$
3. $\mathrm{V}(X) = E((X - \mu)^2) = E(X^2 - 2\mu X + \mu^2) = E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - 2\mu^2 + \mu^2 = E(X^2) - \mu^2$.

□

Παράδειγμα. Έστω ενδεχόμενο $A \subseteq \Omega$ και έστω η διακριτή ΤΜ I_A, με
$$I_A = \begin{cases} 1, & \text{αν πραγματοποιείται το } A, \\ 0, & \text{αλλιώς.} \end{cases}$$
Η ΤΜ ονομάζεται **δείκτρια** (indicator) του ενδεχομένου A και αποτελεί μια **δυαδική** ΤΜ, αφού παίρνει τιμές 0 και 1. Παρατηρήστε ότι αφού η ΤΜ I_A είναι μια συνάρτηση $I_A : \Omega \to \mathbb{R}$, μπορούμε ισοδύναμα να την ορίσουμε ως
$$I_A(\omega) = \begin{cases} 1, & \omega \in A, \\ 0, & \omega \notin A, \end{cases}$$
δηλαδή πρόκειται για την χαρακτηριστική συνάρτηση του συνόλου A.

Αν θέσουμε $p = P(A)$, τότε, βάσει του ορισμού, η μάζα της I_A είναι μια συνάρτηση f με $f(0) = 1-p$ και $f(1) = p$. Η μέση τιμή της ισούται με
$$E(I_A) = \sum_{x=0}^{1} xf(x) = 0 \cdot (1 - p) + 1 \cdot p = p.$$
Διαισθητικά, αν επαναλάβουμε το πείραμα πολλές φορές (υπό τις ίδιες συνθήκες), το ποσοστό πραγματοποιήσεων του Α θα είναι περίπου $p \cdot 100\%$.

Η διακύμανσή της ισούται με
$$V(I_A) = E[(I_A - p)^2] = \sum_{x=0}^{1} (x - p)^2 f(x) = p^2(1 - p) + (1 - p)^2 p = p(1 - p) = p - p^2.$$
Η παράσταση αυτή ελαχιστοποιείται όταν $p \in \{0, 1\}$ και μεγιστοποιείται όταν $p = 1/2$. Διαισθητικά, το αποτέλεσμα γίνεται λιγότερο προβλέψιμο, όταν το p πλησιάζει το $1/2$.

Πόρισμα 2.4. *Αν $X_1, X_2, \ldots, X_n$ διακριτές τυχαίες μεταβλητές και $g_1, g_2, \ldots, g_n$ οποιεσδήποτε συναρτήσεις, όπου $g_i : S_{X_i} \to \mathbb{R}$, τότε:*

$$E(g_1(X_1) + g_2(X_2) + \cdots + g_n(X_n)) = E(g_1(X_1)) + E(g_2(X_2)) + \cdots + E(g_n(X_n)).$$

Ο τύπος του προηγούμενου πορίσματος είναι ιδιαίτερα χρήσιμος για την απλούστευση του υπολογισμού μέσης τιμής σε πολλές εφαρμογές. Αντίστοιχος τύπος ισχύει και για την διακύμανση, μόνο στην περίπτωση που οι τυχαίες μεταβλητές είναι ανεξάρτητες.

Ορισμός. *Οι τυχαίες μεταβλητές $X_1, X_2, \ldots, X_n$ καλούνται ανεξάρτητες όταν*

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = P(X_1 = x_1)P(X_2 = x_2)\cdots P(X_n = x_n),$$

για όλες τις δυνατές τιμές των $x_1, x_2, \ldots, x_n$.

Πρόταση 2.5. *Αν οι τυχαίες μεταβλητές $X_1, X_2, \ldots, X_n$ είναι ανεξάρτητες και $g_1, g_2, \ldots, g_n$ οποιεσδήποτε συναρτήσεις, όπου $g_i : S_{X_i} \to \mathbb{R}$, τότε:*

1. *Οι $g_1(X_1), g_2(X_2), \ldots, g_n(X_n)$ είναι ανεξάρτητες*
2. *$E(g_1(X_1)g_2(X_2)\cdots g_n(X_n)) = E(g_1(X_1))E(g_2(X_2))\cdots E(g_n(X_n))$*
3. *$V(g_1(X_1) + g_2(X_2) + \cdots + g_n(X_n)) = V(g_1(X_1)) + V(g_2(X_2)) + \cdots + V(g_n(X_n))$*

Απόδειξη.

1. Αρχικά, θα δειχθεί ότι $P(X_1 \in A_1, X_2 \in A_2) = P(X_1 \in A_1)P(X_2 \in A_2)$, για οποιαδήποτε $A_1 \subseteq S_{X_1}$, $A_2 \subseteq S_{X_2}$. Πράγματι, είναι

$$\begin{aligned} P(X_1 \in A_1, X_2 \in A_2) &= \sum_{(x,y)\in A_1\times A_2} P(X_1 = x, X_2 = y) = \sum_{(x,y)\in A_1\times A_2} P(X_1 = x)P(X_2 = y) \\ &= \sum_{x\in A_1} P(X_1 = x) \sum_{y\in A_2} P(X_2 = y) = P(X \in A_1)P(X \in A_2) \end{aligned}$$

   Επομένως, για κάθε $(x, y) \in \mathbb{R}^2$ είναι

$$\begin{aligned} P(g_1(X_1) = x, g_2(X_2) = y) &= P(X_1 \in g_1^{-1}(x), X_2 \in g_2^{-1}(y)) = P(X_1 \in g_1^{-1}(x))P(X_2 \in g_2^{-1}(y)) \\ &= P(g_1(X_1) = x)P(g_2(X_2) = y) \end{aligned}$$

   δηλαδή οι $g_1(X_1)$, $g_2(X_2)$ είναι ανεξάρτητες. Το αποτέλεσμα αυτό γενικεύεται επαγωγικά για n όρους.

2. Ομοίως με την Πρόταση 2.3 (1). (Άσκηση.)

3. Έστω $E(X_1) = \mu_1$ και $E(X_2) = \mu_2$, τότε $E(X_1X_2) = \mu_1\mu_2$ και $E(X_1 + X_2) = \mu_1 + \mu_2$, οπότε

$$\begin{aligned} \mathrm{V}(X_1 + X_2) &= E((X_1 + X_2 - \mu_1 - \mu_2)^2) = E((X_1 - \mu_1)^2 + (X_2 - \mu_2)^2 - 2(X_1 - \mu_1)(X_2 - \mu_2)) \\ &= E((X_1 - \mu_1)^2) + E((X_2 - \mu_2)^2) - 2E((X_1 - \mu_1)(X_2 - \mu_2)) \\ &= \mathrm{V}(X_1) + \mathrm{V}(X_2) - 2E(X_1 - \mu_1)E(X_2 - \mu_2) \\ &= \mathrm{V}(X_1) + \mathrm{V}(X_2) - 2(E(X_1) - \mu_1)(E(X_2) - \mu_2) \\ &= \mathrm{V}(X_1) + \mathrm{V}(X_2). \end{aligned}$$

   Κατόπιν, το αποτέλεσμα αυτό γενικεύεται επαγωγικά για n όρους.

□

Παράδειγμα. Θεωρούμε το πείραμα κατά το οποίο ρίχνουμε ένα κέρμα n φορές και την δυαδική ΤΜ X_i, $i \in [n]$, με $X_i = 1$ αν εμφανίσθηκε κορώνα κατά την i-οστή ρίψη, και $X_i = 0$ αλλιώς. Με άλλα λόγια, η X_i είναι η δείκτρια του ενδεχομένου A_i να έρθει κορώνα κατά την i-οστή ρίψη.

Θεωρούμε ότι $P(A_i) = p$, για κάθε i, οπότε, βάσει του προηγούμενου παραδείγματος, έχουμε ότι $E(X_i) = p$ και $V(X_i) = p(1-p)$.

Αν X είναι το πλήθος των φορών που εμφανίσθηκε κορώνα, τότε προφανώς είναι

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

Η X είναι μια νέα ΤΜ με μέση τιμή

$$E(X) = E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n) = np.$$

Επιπλέον, επειδή οι X_i είναι ανεξάρτητες (αφού οι ρίψεις είναι ανεξάρτητες), έχουμε ότι

$$V(X) = V(X_1 + X_2 + \cdots + X_n) = V(X_1) + V(X_2) + \cdots + V(X_n) = np(1-p).$$

Παράδειγμα. Κωδικοποιώντας στο προηγούμενο παράδειγμα την κορώνα με 1 και τα γράμματα με 0, το προηγούμενο πείραμα ισοδυναμεί με την τυχαία επιλογή μιας δυαδικής λέξης με n γράμματα, όταν κάθε γράμμα είναι 1 με πιθανότητα p. Η ΤΜ X αντιστοιχεί στο πλήθος των 1 στη λέξη.

Έστω Y το πλήθος των εμφανίσεων του μοτίβου 11 στη δυαδική λέξη (εμφάνιση κορώνας 2 συνεχόμενες φορές). Ποιά είναι η μέση τιμή και η διακύμανση της Y;

Για την απάντηση, θεωρούμε τη δείκτρια ΤΜ Y_i, $i \in [n-1]$, με $Y_i = 1$ ανν εμφανίζεται το μοτίβο 11 στις θέσεις i και $i+1$. Προφανώς, είναι $E(Y_i) = p^2$ και $V(Y_i) = p^2(1-p^2)$ (γιατί;). Επομένως,

$$E(Y) = E(Y_1 + Y_2 + \cdots + Y_{n-1}) = E(Y_1) + E(Y_2) + \cdots + E(Y_{n-1}) = (n-1)p^2.$$

Όμως, οι Y_i δεν είναι ανεξάρτητες, οπότε δεν μπορεί να χρησιμοποιηθεί ο αντίστοιχος τύπος για τη διακύμανση $V(Y)$.

Πρόταση 2.6. *Αν η X είναι ΤΜ με τιμές στο $\mathbb{N}$, τότε* $E(X) = \sum_{n=1}^{\infty} P(X \geq n)$.

Απόδειξη. Ισχύει ότι $E(X) = \sum_{n=1}^{\infty} nP(X=n) = \sum_{n=1}^{\infty}\sum_{i=1}^{n} P(X=n) = \sum_{i=1}^{\infty}\sum_{n=i}^{\infty} P(X=n) = \sum_{i=1}^{\infty} P(X \geq i)$ □

2.3 Λυμένες ασκήσεις

Άσκηση 2.1. *Σε ένα πρόγραμμα επιλογής τυχαίων φυσικών αριθμών από το 1 έως το n οι αριθμοί έχουν πιθανότητα εμφάνισης ανάλογη του μεγέθους τους, δηλαδή η πιθανότητα να εμφανισθεί ο αριθμός k ισούται με $C \cdot k$, για κάθε $k \in [n]$, όπου C είναι μια σταθερά.*

i) *Να βρεθεί η πιθανότητα εμφάνισης κάθε αριθμού k, όπου $k = 1, \ldots, n$, δηλαδή να προσδιορισθεί η σταθερά C.*

(Υπόδειξη: Το άθροισμα όλων των πιθανοτήτων ισούται με 1.)

ii) *Προτείνετε έναν τρόπο υλοποίησης ενός προγράμματος παραγωγής τυχαίων αριθμών στο διάστημα $[1, n]$ με τις παραπάνω πιθανότητες.*

iii) *Να υπολογισθεί η πιθανότητα εμφάνισης του αριθμού 1 δεδομένου ότι επιλέχθηκε αριθμός από το 1 έως το ν, όπου $\nu \leq n$.*

iv) *Να βρεθεί η αναμενόμενη τιμή των αριθμών που εμφανίζονται.*

v) *Να βρεθεί η διακύμανση των αριθμών που εμφανίζονται.*

Λύση. Έστω X ο αριθμός που θα εμφανισθεί.

i) Η διακριτή ΤΜ X λαμβάνει τιμές στο σύνολο $S = \{1, 2, \ldots, n\}$ και η PMF της X ικανοποιεί τις σχέσεις

$$f(k) = P(X = k) = Ck, \text{ για κάθε } k \in [n].$$

Επιπλέον, πρέπει

$$\sum_{k=1}^{n} f(k) = 1$$

οπότε

$$C(1 + 2 + \cdots + n) = 1 \Leftrightarrow C = \frac{2}{n(n+1)}$$

Άρα,

$$f(k) = P(X = k) = \frac{2k}{n(n+1)}.$$

ii) Ένας τρόπος υλοποίησης είναι να παράγουμε ομοιόμορφα φυσικούς αριθμούς στο σύνολο

$$A = \left\{1, \ldots, \frac{n(n+1)}{2}\right\}.$$

Αν ο αριθμός x που παράγεται είναι στο σύνολο

$$A_k = \left\{1 + \frac{(k-1)k}{2}, \ldots, \frac{k(k+1)}{2}\right\},$$

για κάποιο $k \in [1, n]$ τότε εξάγουμε τον αριθμό k.

Τα σύνολα A_k, $k \in \{1, 2, \ldots, n\}$ αποτελούν μια διαμέριση του A και το καθένα περιέχει k φυσικούς, οπότε η πιθανότητα ο x να ανήκει στο A_k είναι $\frac{2k}{n(n+1)}$, επομένως ο αριθμός k παράγεται με τη ζητούμενη πιθανότητα.

iii) Η ζητούμενη πιθανότητα ισούται με

$$P(X=1|X\le \nu)=\frac{P(X=1)}{P(X=1)+P(X=2)+\cdots+P(X=\nu)}=\frac{C\cdot 1}{C(1+2+\cdots+\nu)}=\frac{2}{\nu(\nu+1)}$$

όπου το C είναι η σταθερά που υπολογίσαμε σε προηγούμενο ερώτημα (η οποία απλοποιήθηκε εδώ).

iv) Η μέση τιμή μ της X ισούται με

$$\mu=E(X)=\sum_{k=1}^{n}kP(X=k)=\sum_{k=1}^{n}\frac{2k^2}{n(n+1)}=\frac{2}{n(n+1)}\sum_{k=1}^{n}k^2=\frac{2}{n(n+1)}\frac{n(n+1)(2n+1)}{6}=\frac{2n+1}{3}.$$

v) Η διακύμανση σ^2 της X ισούται με $\sigma^2=\mathrm{Var}(X)=E(X^2)-E(X)^2$.

Έχουμε ότι

$$E(X^2)=\sum_{k=1}^{n}k^2P(X=k)=\frac{2}{n(n+1)}\sum_{k=1}^{n}k^3=\frac{2}{n(n+1)}\left(\frac{n(n+1)}{2}\right)^2=\frac{n(n+1)}{2}.$$

οπότε

$$\begin{aligned}\sigma^2=E(X^2)-E(X)^2&=\frac{n(n+1)}{2}-\frac{(2n+1)^2}{9}\\&=\frac{9n^2+9n-8n^2+8n-2}{18}=\frac{n^2+n-2}{18}=\frac{(n-1)(n+2)}{18}.\end{aligned}$$

□

Άσκηση 2.2 (Τυχαία γραφήματα). *(Το μοντέλο $G(n,p)$ των Erdos και Rényi)*

Κατασκευάζουμε ένα γράφημα δεσμών $G(n,p)$ με n κορυφές, έτσι ώστε για κάθε ζεύγος κορυφών u,v, ο δεσμός $\{u,v\}$ ανήκει στο γράφημα με πιθανότητα p (ανεξάρτητα από τους άλλους δεσμούς). Συνήθως το p είναι συνάρτηση του n.

i) *Να βρεθεί ο μέσος βαθμός των κορυφών του $G(n,p)$.*

ii) *Να βρεθεί ο μέσος αριθμός δεσμών του $G(n,p)$.*

iii) *Να βρεθεί ο μέσος αριθμός των τριγώνων του $G(n,p)$.*

Λύση.

i) Για κάθε κορυφή ν συμβολίζουμε με X_ν τον βαθμό της.

Για κάθε κορυφή $u\neq\nu$ ορίζουμε την ΤΜ

$$Y_{u\nu}=\begin{cases}1, & \text{αν υπάρχει ο δεσμός } \{u,\nu\}\\ 0, & \text{αν δεν υπάρχει ο δεσμός } \{u,\nu\}\end{cases}$$

Ισχύει ότι

$$X_\nu=\sum_u Y_{u\nu}$$

Για κάθε $u \neq v$ ισχύει ότι $E(Y_{uv}) = 1P(Y_{uv} = 1) + 0P(Y_{uv} = 0) = 1p + 0(1-p) = p$.

Άρα, από την γραμμικότητα της μέσης τιμής ισχύει ότι

$$E(X_v) = E(\sum_u Y_{uv}) = \sum_u E(Y_{uv}) = (n-1)p.$$

Επομένως, για να κατασκευάσουμε ένα τυχαίο γράφημα με n κορυφές όπου ο μέσος βαθμός κορυφής είναι c, επιλέγουμε κάθε δεσμό με πιθανότητα $p = \frac{c}{n-1}$.

ii) Έστω Z το πλήθος δεσμών. Γνωρίζουμε ότι $2Z = \sum_v X_v$. Επομένως,

$$E(Z) = E(\frac{1}{2}\sum_v X_v) = \frac{1}{2}\sum_v E(X_v) = \frac{1}{2}\sum_v (n-1)p = \frac{n(n-1)}{2}p = \binom{n}{2}p.$$

Επομένως, αν θέλουμε μέσο πλήθος δεσμών $E(Z) = c$, τότε θέτουμε $p = \frac{2c}{n(n-1)}$.

iii) Παρατηρούμε ότι υπάρχουν $\binom{n}{3}$ τριάδες κορυφών, κάποιες από τις οποίες αποτελούν τρίγωνο.

Έστω u, v, w 3 κορυφές του γραφήματος. Η πιθανότητα να αποτελούν τρίγωνο ισούται με $p \cdot p \cdot p = p^3$. Παρατηρήστε ότι η ύπαρξη ή μη ενός τριγώνου επηρεάζει την πιθανότητα ύπαρξης των άλλων τριγώνων που έχουν κοινή μια ή δύο πλευρές με αυτό.

Έστω X ο αριθμός των τριγώνων του γραφήματος.

Για κάθε τριάδα κορυφών u, v, w θεωρούμε την ΤΜ

$$Y_{uvw} = \begin{cases} 1, & \text{αν υπάρχει το τρίγωνο } \{u, v, w\} \\ 0, & \text{αν δεν υπάρχει το τρίγωνο } \{u, v, w\} \end{cases}$$

Ισχύει ότι

$$X = \sum_{u,v,w} Y_{uvw}$$

Για κάθε u, v, w ισχύει ότι $E(Y_{uvw}) = 1 \cdot P(Y_{uvw} = 1) + 0 \cdot P(Y_{uvw} = 0) = p^3$.

Άρα,

$$E(X) = E(\sum_{u,v,w} Y_{uvw}) = \sum_{u,v,w} E(Y_{uvw}) = \sum_{u,v,w} p^3 = \binom{n}{3}p^3 = \frac{n(n-1)(n-2)}{6}p^3.$$

Αν επιλέξουμε $p = \frac{c}{n}$, τότε

$$E(X) = \frac{n(n-1)(n-2)}{6} \cdot \frac{c^3}{n^3} \approx \frac{c^3}{6}$$

Επομένως, για να κατασκευάσουμε ένα τυχαίο γράφημα με n κορυφές και k τρίγωνα, θέλουμε $\frac{c^3}{6} = k$, ή ισοδύναμα $c = \sqrt[3]{6k}$, οπότε επιλέγουμε κάθε δεσμό με πιθανότητα $p = \frac{\sqrt[3]{6k}}{n}$.

□

2.4 Ασκήσεις προς επίλυση

1) Να βρεθεί (αν υπάρχει) σταθερά C ώστε η συνάρτηση $f(x) = P(X = k)$ να είναι συνάρτηση μάζας πιθανότητας μιας διακριτής τυχαίας μεταβλητής X με σύνολο τιμών S όταν

i) $P(X = k) = \dfrac{C}{k!}$ και $S = \mathbb{N}$.

ii) $P(X = k) = \dfrac{C}{k}$ και $S = \mathbb{N}^*$.

iii) $P(X = k) = C \cdot \dfrac{(-1)^{n+1}}{n}$ και $S = \mathbb{N}^*$.

(Είναι γνωστό ότι $\sum_{n=1}^{\infty} \dfrac{(-1)^{n+1}}{n} = \ln 2$.)

iv) $P(X = k) = \dfrac{C}{k(k+1)}$ και $S = \mathbb{N}^*$.

v) $P(X = k) = \dfrac{C}{k(k+1)(k+2)}$ και $S = \mathbb{N}^*$.

2) Έστω ότι ρίχνουμε 10 αμερόληπτα νομίσματα και ξαναρίχνουμε τα νομίσματα που έφεραν κορώνα. Να βρεθεί η κατανομή πιθανότητας του συνολικού αριθμού εμφανίσεων της όψης κορώνα.

3) Ένας πελάτης μπαίνει σε ένα κατάστημα με παγωτά και αγοράζει 1, 2, 3, 4, ή 5 μπάλες παγωτού με την εξής συνάρτηση μάζας πιθανότητας:

$X = x$	1	2	3	4	5
$P(X = x)$	0.41	0.37	0.16	0.05	0.01

i) Να υπολογισθεί η συνάρτηση κατανομής $F(x)$ της ΤΜ X: αριθμός από μπάλες παγωτού που αγοράζει ένας πελάτης.

ii) Να υπολογισθεί η αναμενόμενη τιμή και η διακύμανση της ΤΜ X.

iii) Αν κάθε μπάλα παγωτού κοστίζει 2 ευρώ, ποιό είναι το αναμενόμενο κέρδος από κάθε πελάτη;

Κεφάλαιο 3

Σημαντικές διακριτές κατανομές

Όπως είδαμε στα προηγούμενα, η συμπεριφορά μιας ΤΜ X καθορίζεται από το σύνολο τιμών της S_X και από την συνάρτηση $f_X(x)$, $x \in S_X$. Στο κεφάλαιο αυτό παρουσιάζονται ορισμένες σημαντικές κατηγορίες διακριτών τυχαίων μεταβλητών, οι οποίες συναντώνται συχνά στις εφαρμογές.

3.1 Κατανομή Bernoulli

Ορισμός. *Μια διακριτή ΤΜ X λέμε ότι ακολουθεί την* κατανομή Bernoulli με παράμετρο p *και γράφουμε* $X \sim \text{Bern}(p)$, *όπου* $p \in (0,1)$, *ανν έχει σύνολο τιμών το* $S_X = \{0,1\}$ *και PMF* f_X, *με*

$$f_X(1) = P(X = 1) = p, \text{ και } f_X(0) = P(X = 0) = 1 - p.$$

Άμεσα προκύπτουν η CDF (συνάρτηση κατανομής), η μέση τιμή και η διακύμανση:

$$F_X(x) = P(X \leq x) = \begin{cases} 0, & x < 0, \\ 1 - p, & 0 \leq x < 1 \\ 1, & x \geq 1 \end{cases}$$

$$\mu = E(X) = 0 \cdot P(X = 0) + 1 \cdot P(X = 1) = p.$$

και

$$\sigma^2 = V(X) = E((X - p)^2) = (1 - p)^2 p + (0 - p)^2(1 - p) = p(1 - p).$$

Η κατανομή Bernoulli εμφανίζεται σε τυχαία φαινόμενα, όπου υπάρχουν δύο δυνατά αποτελέσματα (επιτυχία - αποτυχία). Κλασικό παράδειγμα ΤΜ που ακολουθεί την κατανομή Bernoulli είναι η έκβαση (επιτυχία: 1, αποτυχία: 0) της ρίψης ενός νομίσματος.

Η κατανομή Bernoulli είναι η βάση για να ορίσουμε πολλές κατανομές. Σημειώνεται ότι η δείκτρια συνάρτηση I_A ενός ενδεχομένου A είναι μια ΤΜ Bernoulli με παράμετρο $p = P(A)$ (την πιθανότητα πραγματοποίησης του A).

3.2 Διωνυμική κατανομή

Παράδειγμα 3.2.1 (Εισαγωγικό παράδειγμα). *Έστω ότι ρίχνουμε ένα νόμισμα N φορές για το οποίο γνωρίζουμε ότι $P(K) = p$ και $P(\Gamma) = 1 - p$ σε κάθε μια από τις ρίψεις. Να υπολογισθεί η PMF της TM*

$$X = \text{αριθμός των ρίψεων που φέραμε Κορώνα.}$$

Λύση. Η X έχει σύνολο τιμών $S = \{0, 1, 2, \ldots, N\}$. Θεωρούμε τις TM X_i, $i \in [N]$, με

$$X_i = \begin{cases} 1, & \text{αν έρθει κορώνα στην } i\text{-οστή ρίψη,} \\ 0, & \text{αν έρθει γράμματα στην } i\text{-οστή ρίψη.} \end{cases}$$

Οι (ανεξάρτητες) TM $X_1, X_2, \ldots, X_N$ ακολουθούν την κατανομή Bernoulli με παράμετρο p και

$$X = X_1 + X_2 + \cdots + X_N.$$

Κωδικοποιούμε την Κορώνα με 1 και τα Γράμματα με 0, οπότε, αν x_i είναι το αποτέλεσμα της i-οστής ρίψης, τότε η ακολουθία $x = x_1 x_2 \cdots x_N$ είναι μια δυαδική λέξη με N γράμματα και το ενδεχόμενο $\{X = k\}$, όπου $k \in S$, πραγματοποιείται όταν η x περιέχει ακριβώς k μονάδες (και $N - k$ μηδενικά). Υπάρχουν $\binom{N}{k}$ τέτοιες δυαδικές λέξεις x και κάθε μία εμφανίζεται με πιθανότητα $p^k(1-p)^{N-k}$. Επομένως, η PMF της TM X είναι η

$$f_X(k) = P(X = k) = \binom{N}{k} p^k (1-p)^{N-k}, \ k = 0, 1, 2, \ldots, N \qquad \square$$

Ορισμός. *Μια διακριτή TM X λέμε ότι ακολουθεί την* διωνυμική κατανομή με παραμέτρους N και p, *για κάποιο $N \in \mathbb{N}^*$ και $p \in (0, 1)$, ανν έχει σύνολο τιμών $S_X = \{0, 1, 2, \ldots, N\}$ και PMF*

$$f_X(k) = P(X = k) = \binom{N}{k} p^k (1-p)^{N-k}.$$

(Συμβολισμός: $X \sim \text{Binom}(N, p)$, ή $X \sim \Delta\iota\omega\nu(N, p)$.)

Πρακτικά, η TM $X \sim \text{Binom}(N, p)$ αντιπροσωπεύει το πλήθος επιτυχιών μετά από N ανεξάρτητες επαναλήψεις, όταν η πιθανότητα επιτυχίας είναι p.

Η μέση τιμή και η διακύμανση της X δίνονται από τους ακόλουθους τύπους:

$$\mu = E(X) = Np \text{ και } \sigma^2 = \text{V}(X) = Np(1-p).$$

Πράγματι, επειδή $X = X_1 + X_2 + \cdots + X_N$ όπου X_1, X_2, …, $X_N \sim \text{Bern}(p)$ και οι X_1, X_2, …, X_N είναι ανεξάρτητες, έπεται ότι

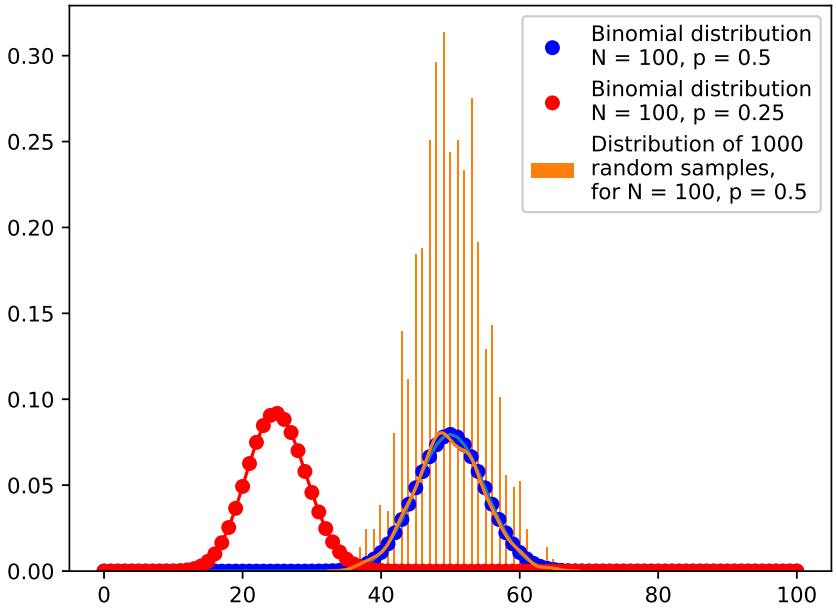
$$E(X) = E(X_1 + X_2 + \cdots + X_N) = E(X_1) + E(X_2) + \cdots + E(X_N) = Np.$$

και

$$\text{V}(X) = \text{V}(X_1 + X_2 + \cdots + X_N) = \text{V}(X_1) + \text{V}(X_2) + \cdots + \text{V}(X_N) = Np(1-p).$$

Επιπλέον, η συνάρτηση κατανομής δίνεται από τον τύπο

$$F_X(x) = P(X \le x) = \sum_{k=0}^{\lfloor x \rfloor} \binom{N}{k} p^k (1-p)^{N-k}.$$



Το παραπάνω σχήμα απεικονίζει γραφικά 2 διωνυμικές κατανομές (μπλε - κόκκινο) με διαφορετικές παραμέτρους και την κατανομή ενός τυχαίου δείγματος (πορτοκαλί) που παράχθηκε με βάση τη διωνυμική κατανομή. Το σχήμα παράγεται από τον ακόλουθο κώδικα:

```
from scipy import stats
import matplotlib.pyplot as plt
import seaborn as sns

trials = 100 #N
P, P2 = 0.5, 0.25 #success probability
S = 1000 #sample size = #experiments
rs = np.random.binomial(n=trials, p=P, size=S) #draw a random sample of size S using
    numpy.random
#rs = stats.binom.rvs(n=trials, p=P, size=S) #draw a random sample of size S using
    scipy.stats

vals = np.arange(trials+1) #list of values 0,1,2,..., trials
f1 = stats.binom.pmf(vals, trials, P)
f2 = stats.binom.pmf(vals, trials, P2)
label1 = "Binomial distribution\n" + "N = " + str(trials) + ", p = " + str(P)
plt.plot(vals,f1, 'bo',label = label1)
plt.plot(vals,f1)
label2 = "Binomial distribution\n" + "N = " + str(trials) + ", p = " + str(P2)
plt.plot(vals,f2, 'ro', label = label2)
plt.plot(vals,f2, color='red')
label3 = "Distribution of " + str(S) + "\nrandom samples,\n" + "for N = " + str(
    trials) + ", p = " + str(P)
sns.distplot(rs, hist=True, kde=True, bins = trials+1, label = label3)
plt.legend()
plt.show()
```

Παράδειγμα 3.2.2. *Για ένα αεροπλάνο* 50 *θέσεων έχουν γίνει* 55 *κρατήσεις (overbooking). Η πιθανότητα ο καθένας από αυτούς να εμφανισθεί στο αεροδρόμιο είναι* 90% *(ανεξάρτητα από τους υπολοίπους).*

i) *Ποιος είναι ο αναμενόμενος αριθμός επιβατών που θα εμφανισθούν για check-in;*

ii) *Ποια είναι η πιθανότητα κάποιοι επιβάτες να μείνουν εκτός πτήσης;*

Λύση. Έστω X ο αριθμός των επιβατών που θα εμφανισθούν και έστω

$$X_i = \begin{cases} 1, & \text{αν εμφανισθεί ο } i\text{-οστός επιβάτης,} \\ 0, & \text{αλλιώς} \end{cases} \qquad i \in [55].$$

Τότε, $X = X_1 + X_2 + \cdots + X_{55}$, όπου $X_1, X_2, \ldots, X_{55} \sim \text{Bern}(0.9)$. Άρα, $X \sim \text{Binom}(55, 0.9)$.

i) Κατά μέσο όρο αναμένονται $E(X) = 55 \cdot 0.9 = 49.5$ επιβάτες.

ii) Η πιθανότητα κάποιοι επιβάτες να μείνουν εκτός πτήσης ισούται με

$$\begin{aligned} P(X > 50) &= P(X = 51) + P(X = 52) + P(X = 53) + P(X = 54) + P(X = 55) \\ &= \binom{55}{51} 0.9^{51} 0.1^4 + \binom{55}{52} 0.9^{52} 0.1^3 + \binom{55}{53} 0.9^{53} 0.1^2 + \binom{55}{54} 0.9^{54} 0.1^1 + \binom{55}{55} 0.9^{55} 0.1^0 \\ &= 0.3451. \end{aligned}$$

□

Παράδειγμα 3.2.3. *Σε ένα τετρακινητήριο αεροπλάνο, το ενδεχόμενο ο* i*-οστός κινητήρας να πάθει βλάβη εν ώρα πτήσης είναι ανεξάρτητο από την πιθανότητα βλάβης των υπολοίπων. Μια πτήση είναι ασφαλής αν λειτουργούν τουλάχιστον οι δύο κινητήρες. Να βρεθεί η πιθανότητα μια πτήση να γίνει με ασφάλεια.*

Λύση. Αν X το πλήθος των κινητήρων που έπαθαν βλάβη, τότε

$$F_X(2) = P(X \le 2) = \sum_{k=0}^{2} \binom{4}{k} p^k (1-p)^{4-k} = (1-p)^4 + 4p(1-p)^3 + 6p^2(1-p)^2$$

όπου p η πιθανότητα βλάβης. Για παράδειγμα, αν $p = 0.1$, τότε $F_X(2) = 0.9963$. □

Παράδειγμα 3.2.4. *Ένας αδιάβαστος φοιτητής, ο οποίος απαντά στην τύχη, συμμετέχει σε μια εξέταση με* 20 *ερωτήσεις πολλαπλής επιλογής με* 4 *απαντήσεις η κάθεμια, και μόνο μια σωστή από αυτές.*

i) *Να βρεθεί η μέση τιμή των σωστών απαντήσεων.*

ii) *Να βρεθεί η πιθανότητα να δώσει ακριβώς* 10 *σωστές απαντήσεις.*

Λύση. Το πλήθος σωστών απαντήσεων X ακολουθεί διωνυμική κατανομή με παραμέτρους $N = 20$ και $p = 1/4$. Επομένως, $E(X) = Np = 5$ και $P(X = 10) = \binom{N}{10} p^{10}(1-p)^{N-10} \simeq 0.00992$. □

3.3 Γεωμετρική κατανομή

Ορισμός. *Μια ΤΜ X λέμε ότι ακολουθεί την γεωμετρική κατανομή με παράμετρο p, όπου $p \in (0,1)$, ανν έχει σύνολο τιμών $S_X = \{1,2,3,\ldots\} = \mathbb{N}^*$ και PMF*

$$f_X(k) = P(X = k) = (1-p)^{k-1}p, \quad k \in \mathbb{N}^*.$$

(Συμβολισμός $X \sim Γεωμ(p)$ ή $X \sim Geom(p)$.)

Η X αντιπροσωπεύει τον αριθμό επαναλήψεων ενός πειράματος μέχρι την πρώτη επιτυχία, όταν η πιθανότητα επιτυχίας είναι ίση με p.

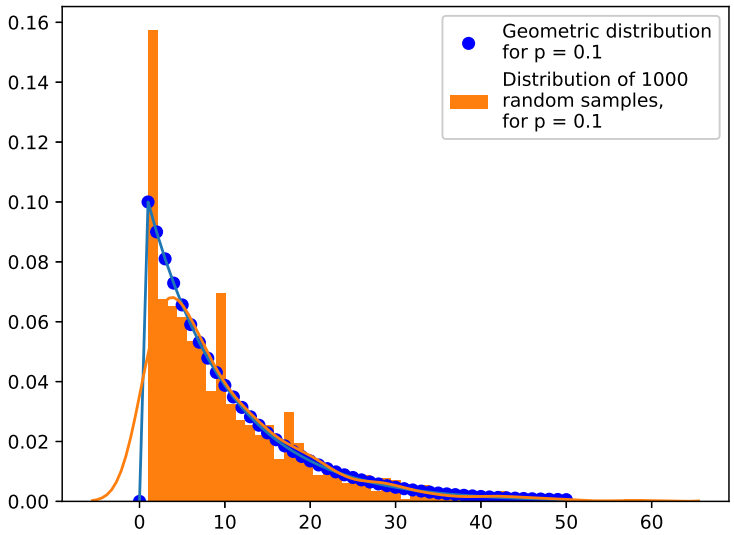
Η μέση τιμή και η διακύμανση δίνονται από τους τύπους

$$E(X) = \frac{1}{p}, \qquad \mathrm{V}(X) = \frac{1-p}{p^2}$$

Η απόδειξη των παραπάνω τύπων δίδεται στην άσκηση 3.2.

Η συνάρτηση κατανομής είναι η

$$F_X(x) = P(X \le x) = \sum_{i=1}^{\lfloor x \rfloor}(1-p)^{i-1}p = p\sum_{i=0}^{\lfloor x \rfloor - 1}(1-p)^i = p\frac{1-(1-p)^{\lfloor x \rfloor}}{1-(1-p)} = 1-(1-p)^{\lfloor x \rfloor}$$



Παρατήρηση: Η γεωμετρική κατανομή ικανοποιεί την ιδιότητα

$$P(X \ge m+n | X > n) = P(X \ge m),$$

όπου το "≥" μπορεί να αντικατασταθεί και από "=" ή ">". Η ιδιότητα αυτή καλείται **έλλειψη μνήμης**. Η απόδειξη ζητείται στην άλυτη άσκηση 3.11.

Παράδειγμα 3.3.1. *Ρίχνουμε ένα νόμισμα συνεχώς. Το νόμισμα έχει πιθανότητα p να φέρει Κορώνα και πιθανότητα $1-p$ να φέρει Γράμματα. Έστω X η ρίψη στην οποία θα εμφανισθεί για πρώτη φορά Κορώνα. Να βρεθεί η πιθανότητα $P(X=k)$.*

Λύση. Αν θεωρήσουμε τη δείκτρια ΤΜ X_i για το ενδεχόμενο εμφάνισης κορώνας στην i-οστή ρίψη, $i \in \mathbb{N}^*$, τότε, λόγω ανεξαρτησίας, ισχύει ότι

$$\begin{aligned} P(X=k) &= P(X_1=0, X_2=0, \ldots, X_{k-1}=0, X_k=1) \\ &= P(X_1=0)P(X_2=0)\cdots P(X_{k-1}=0)P(X_k=1) \\ &\underbrace{(1-p)(1-p)\cdots(1-p)}_{k-1 \text{ φορές}}\, p = (1-p)^{k-1}p. \end{aligned}$$

Άρα, η ΤΜ X ακολουθεί την γεωμετρική κατανομή με παράμετρο p. □

Παράδειγμα 3.3.2. *Κάποιος επιβάτης έχει υπολογίσει ότι η πιθανότητα να συναντήσει ελεγκτές εισιτηρίων σε μια διαδρομή με λεωφορείο είναι $p = \dfrac{1}{100}$.*

i) *Να βρεθεί η πιθανότητα να κάνει 60 διαδρομές χωρίς να συναντήσει καθόλου ελεγκτές εισιτηρίων.*

ii) *Να βρεθεί ο μέσος αριθμός ελέγχων που θα γίνουν στον επιβάτη μέσα σε 60 διαδρομές.*

iii) *Αν το πρόστιμο για μια διαδρομή χωρίς εισιτήριο είναι 84 ευρώ, ποιο είναι το αναμενόμενο πρόστιμο για 60 διαδρομές χωρίς εισιτήριο;*

Λύση. i) Έστω X η διαδρομή στην οποία ο επιβάτης θα συναντήσει για πρώτη φορά ελεγκτές. Η ΤΜ X ακολουθεί την γεωμετρική κατανομή με παράμετρο $p = \frac{1}{100}$. Η πιθανότητα να μην συναντήσει ελεγκτές μέσα σε 60 διαδρομές ισούται με

$$P(X > 60) = (1-p)^{60} = (1 - \frac{1}{100})^{60} = 0.552$$

ii) Έστω Y ο αριθμός των ελέγχων που θα έχει ο επιβάτης μέσα σε 60 διαδρομές και έστω

$$Y_i = \begin{cases} 1, & \text{αν γίνει έλεγχος στην } i\text{-οστή διαδρομή} \\ 0, & \text{αλλιώς.} \end{cases}, i = 1, 2, \ldots, 60.$$

Ισχύει ότι $Y_i \sim \text{Bern}(\frac{1}{100})$, για κάθε $i \in [60]$. Επειδή $Y = Y_1 + Y_2 + \cdots + Y_{60}$ και οι Y_i, $i \in [60]$, είναι ανεξάρτητες, έπεται ότι η Y ακολουθεί την διωνυμική κατανομή με παραμέτρους $N = 60$ και $p = \frac{1}{100}$. Επομένως, η μέση τιμή της ισούται με $\mu = E(Y) = 60 \cdot \frac{1}{100} = 0.6$, δηλαδή ο επιβάτης θα συναντήσει ελεγκτή κατά μέσο 6 φορές ανά 600 διαδρομές.

iii) Το αναμενόμενο πρόστιμο για 60 διαδρομές ισούται με $E(Y) \cdot 84 = 0.6 \cdot 84 = 50.4$ ευρώ. □

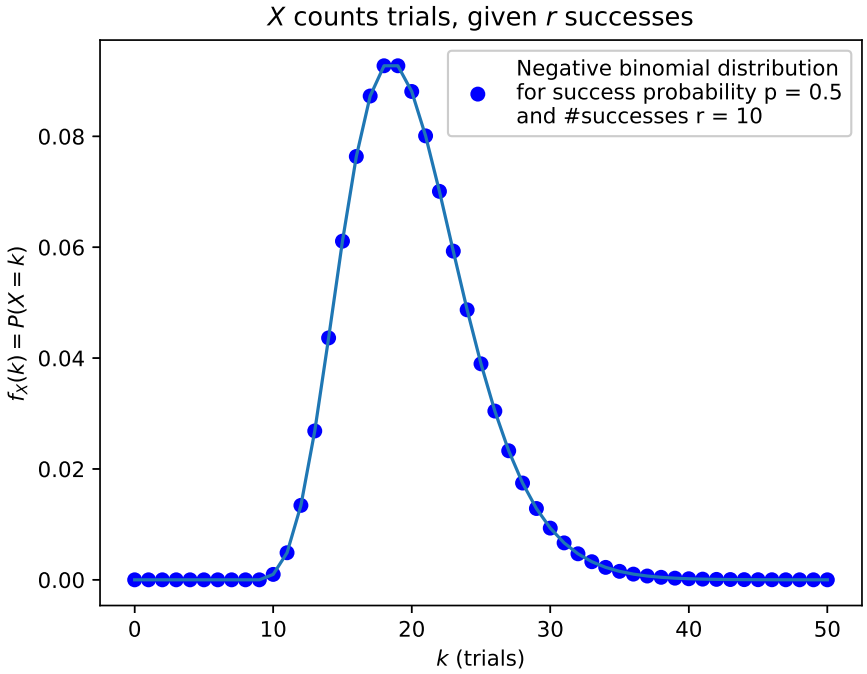
3.4 Αρνητική διωνυμική κατανομή

Ορισμός. *Μια διακριτή ΤΜ X λέμε ότι ακολουθεί την αρνητική διωνυμική κατανομή με παραμέτρους r, p και γράφουμε $X \sim \mathrm{NB}(r,p)$, ανν έχει σύνολο τιμών $S_X = \{r, r+1, r+2, \ldots\}$ και PMF*

$$f_X(k) = P(X=k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}, \qquad 0 < p < 1, r \in \mathbb{N}^*.$$

Η X αντιπροσωπεύει τον αριθμό επαναλήψεων ενός πειράματος μέχρι την r-οστή επιτυχία, όταν η πιθανότητα επιτυχίας είναι ίση με p.

Η μέση τιμή και η διακύμανση είναι αντίστοιχα ίσες με r/p και $r(1-p)/p^2$. Η απόδειξη των παραπάνω τύπων δίνεται στην άσκηση 3.3.



Παρατήρηση: Έστω Y το πλήθος των αποτυχιών μέχρι την r-οστή επιτυχία. Προφανώς, είναι $S_Y = \mathbb{N}$ και $X = Y + r$. Η PMF της Y είναι η

$$f_Y(y) = \binom{y+r-1}{r-1} p^r (1-p)^y \overset{k=y+r}{=} \binom{k-1}{r-1} p^r (1-p)^{k-r} = f_X(k),$$

άρα $f_X(k) = f_Y(k-r)$. Ορισμένες φορές στη βιβλιογραφία, όπως και στη βιβλιοθήκη scipy.stats της Python, θεωρείται η Y (και όχι η X) ως η ΤΜ που ακολουθεί την $\mathrm{NB}(r,p)$, οπότε το τελευταίο σχήμα παράγεται καλώντας την PMF της βιβλιοθήκης ως εξής:

```
stats.nbinom.pmf(k-r, r, p)
```

3.5 Υπεργεωμετρική κατανομή

Ορισμός. *Η διακριτή ΤΜ X λέμε ότι ακολουθεί τη υπεργεωμετρική κατανομή με παραμέτρους M, n, N και γράφουμε $X \sim \mathrm{HGeom(M, n, N)}$, αν έχει σύνολο τιμών $S_X = [\max\{0, N-b\}, \min\{N, a\}]$, $a, b, N \in \mathbb{N}$ και PMF*

$$f_X(k) = P(X = k) = \frac{\binom{a}{k}\binom{b}{N-k}}{\binom{a+b}{N}} = \frac{\binom{n}{k}\binom{M-n}{N-k}}{\binom{M}{N}}, \qquad k \in S_X, n = a, M = a + b.$$

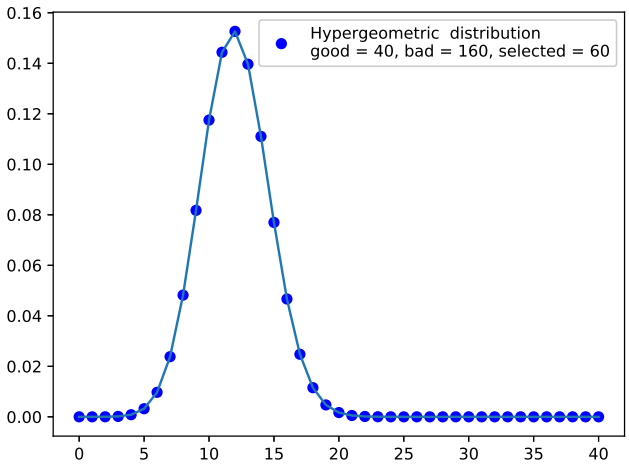
Η X αντιπροσωπεύει τον αριθμό άσπρων (καλών) αντικειμένων που επιλέγονται μετά από N επιλογές χωρίς επανατοποθέτηση μεταξύ $a = n$ άσπρων και $b = M-n$ μαύρων (κακών) αντικειμένων. Η μέση τιμή και η διακύμανση είναι αντίστοιχα ίσες με

$$E(X) = \frac{Na}{a+b} \qquad \text{και} \qquad V(X) = \frac{Nab(a+b-N)}{(a+b)^2(a+b-1)}.$$

Παρατήρηση: Αθροίζοντας για όλες τις δυνατές τιμές του k, προκύπτει ότι

$$\binom{a+b}{N} = \binom{a+b}{N}\sum_k P(X=k) = \sum_k \binom{a}{k}\binom{b}{N-k}.$$

Η παραπάνω σχέση είναι γνωστή ως ταυτότητα Vandermonde ή τύπος του Cauchy και με τη βοήθειά της αποδεικνύονται οι παραπάνω εκφράσεις για τη μέση τιμή και τη διακύμανση.



Προσέγγιση υπεργεωμετρικής κατανομής από τη διωνυμική κατανομή: Αν $a/M = n/M \to p$ καθώς $M = a + b \to +\infty$, τότε

$$\mathrm{HGeom(M, n, N)} \to \mathrm{Binom}(N, p), \qquad \text{καθώς } M \to \infty.$$

Η απόδειξη αφήνεται ως άσκηση. Πρακτικά, η υπεργεωμετρική κατανομή αντιστοιχεί στο παραπάνω πείραμα χωρίς επανατοποθέτηση, ενώ η διωνυμική αντιστοιχεί στο ίδιο πείραμα αλλά με επανατοποθέτηση. Όταν ο συνολικός πληθυσμός M είναι πολύ μεγάλος, η επανατοποθέτηση επηρεάζει ελάχιστα την πιθανότητα επιλογής του κάθε ατόμου.

3.6 Κατανομή Poisson

Ορισμός. *Μια ΤΜ X λέμε ότι ακολουθεί την κατανομή Poisson με παράμετρο $\lambda > 0$, αν $S_X = \mathbb{N}$ και έχει PMF*

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

(Συμβολισμός: $X \sim Poisson(\lambda)$.)

Η μέση τιμή και η διακύμανση της X είναι αντίστοιχα ίσες με $E(X) = \lambda$ και $V(X) = \lambda$.

Πράγματι, είναι

$$E(X) = \sum_{k=0}^{\infty} kP(X = k) = \sum_{k=0}^{\infty} ke^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\lambda\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} = e^{-\lambda}\lambda e^{\lambda} = \lambda$$

και

$$\begin{aligned} V(X) = E(X^2) - (E(X))^2 &= \sum_{k=0}^{\infty} k^2P(X = k) - \lambda^2 = -\lambda^2 + e^{-\lambda}\sum_{k=0}^{\infty} k^2\frac{\lambda^k}{k!} = -\lambda^2 + \lambda e^{-\lambda}\sum_{k=1}^{\infty} k\frac{\lambda^{k-1}}{(k-1)!} \\ &= -\lambda^2 + \lambda e^{-\lambda}\sum_{k=0}^{\infty}(k+1)\frac{\lambda^k}{k!} = -\lambda^2 + \lambda e^{-\lambda}\left(\sum_{k=0}^{\infty} k\frac{\lambda^k}{k!} + \sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\right) \\ &= -\lambda^2 + \lambda e^{-\lambda}\left(\lambda e^{\lambda} + e^{\lambda}\right) = \lambda. \end{aligned}$$

Η κατανομή Poisson ονομάζεται και νόμος των μικρών αριθμών. Χρησιμοποιείται ως μοντέλο για τυχαία φαινόμενα όπως

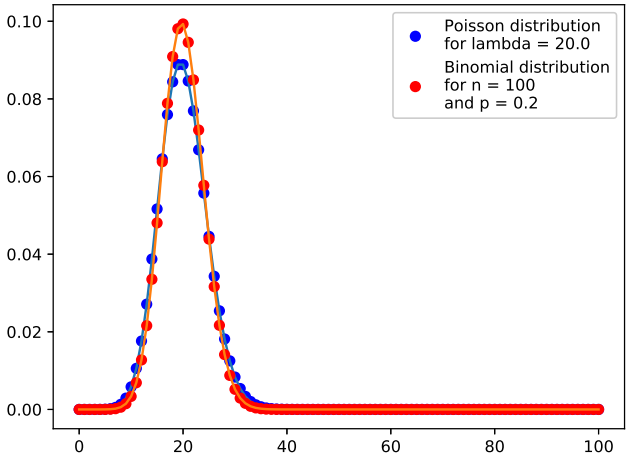
- ο αριθμός των τροχαίων που συμβαίνουν σε μια περιοχή σε κάποιο χρονικό διάστημα
- ο αριθμός των πελατών που φτάνουν σε ένα κατάστημα σε κάποιο διάστημα χρόνου
- ο αριθμός των πλοίων που φτάνουν σε ένα λιμάνι σε κάποιο διάστημα χρόνου
- ο αριθμός των τυπογραφικών λαθών που εμφανίζονται σε κάποια σελίδα ενός βιβλίου
- ο αριθμός των ψαριών (poisson στα Γαλλικά) που πιάνονται στα δίκτυα μας σε κάποιο χρονικό διάστημα.

Προσέγγιση της διωνυμικής κατανομής από την κατανομή Poisson Αν μια ΤΜ X ακολουθεί την $\text{Binom}(n, p)$, τότε για μεγάλες τιμές του n και μικρές τιμές του p, η X προσεγγίζεται από μια ΤΜ η οποία ακολουθεί την κατανομή $\text{Poisson}(np)$, δηλαδή

$$\text{Αν } \lim_{n\to\infty} np = \lambda \in \mathbb{R}, \text{ τότε } \lim_{n\to\infty} \text{Binom}(n, p) = \text{Poisson}(\lambda).$$

Η απόδειξη δίνεται στην άσκηση 3.5.

Στις εφαρμογές, χρησιμοποιείται η προσέγγιση αυτή για την απλοποίηση των υπολογισμών, δηλαδή αν $X \sim \text{Binom}(n, p)$, τότε $P(X = k) = f_X(k) \approx f(k)$, όπου f η PMF μιας ΤΜ Poisson με $\lambda = np$. Η προσέγγιση αυτή είναι πολύ καλή, όταν το n είναι μεγάλο και το p μικρό.



Παράδειγμα 3.6.1. *Ένας server δέχεται κατά μέσο όρο* 5 *αιτήσεις από clients ανά δευτερόλεπτο.*

1. *Ποια είναι η πιθανότητα να δεχθεί* 5 *αιτήσεις στο επόμενο δευτερόλεπτο;*
2. *Ποια είναι η πιθανότητα να δεχθεί* 1000 *αιτήσεις στο επόμενο λεπτό;*

Λύση. Αν X το πλήθος των αιτήσεων σε ένα δευτερόλεπτο, τότε $X \sim \text{Poisson}(\lambda)$, όπου $\lambda = 5$. Επομένως, η ζητούμενη πιθανότητα είναι

$$P(X = 5) = e^{-5}\frac{5^5}{5!} \simeq 0.17546737.$$

Αν X_i το πλήθος των αιτήσεων στο i-οστό δευτερόλεπτο του λεπτού, $i \in [60]$, τότε $X_i \sim \text{Poisson}(\lambda)$, όπου $\lambda = 5$. Οι ΤΜ X_i είναι ανεξάρτητες, οπότε το πλήθος αιτήσεων στο λεπτό είναι $X = \sum_{i=1}^{t} X_i \sim \text{Poisson}(\lambda t)$, όπου $t = 60$. (Η απόδειξη ζητείται στην άσκηση 3.12.)

Κατόπιν τούτου, είναι

$$P(X = 1000) = e^{-300}\frac{300^{1000}}{1000!}$$

Ο αριθμός αυτός προσεγγίζεται πιο εύκολα με τη βοήθεια της κανονικής κατανομής που θα δούμε στα επόμενα. □

Παράδειγμα 3.6.2. *Η πιθανότητα ένας άνθρωπος να έχει αλλεργία σε ένα εμβόλιο είναι* $p = \frac{1}{1000}$. *Κάνουμε ένεση σε* 2000 *άτομα. Να βρεθεί η πιθανότητα να παρουσιάσουν αλλεργία πάνω από* 2 *άτομα.*

Λύση. Έστω X ο αριθμός των ατόμων που θα πάθουν αλλεργία. Η X ακολουθεί την διωνυμική κατανομή $\text{Binom}(2000, 0.001)$. Ψάχνουμε την πιθανότητα $P(X > 2)$. Ισχύει ότι

$$\begin{aligned} P(X > 2) &= 1 - P(X \leq 2) \\ &= 1 - P(X = 2) - P(X = 1) - P(X = 0) \\ &= 1 - \binom{2000}{2} 0.001^2 \cdot 0.999^{1998} - \binom{2000}{1} 0.001^1 \cdot 0.999^{1999} - \binom{2000}{0} 0.001^0 \cdot 0.999^{2000} \\ &= 0.3233. \end{aligned}$$

Μπορούμε να προσεγγίσουμε την ΤΜ X από την ΤΜ Y που ακολουθεί την κατανομή Poisson με παράμετρο $\lambda = 2000 \cdot \frac{1}{1000} = 2$. Ισχύει ότι

$$\begin{aligned} P(X > 2) \simeq P(Y > 2) &= 1 - P(Y \leq 2) \\ &= 1 - P(Y = 2) - P(Y = 1) - P(Y = 0) \\ &= 1 - e^{-2}\frac{2^2}{2!} - e^{-2}\frac{2^1}{1!} - e^{-2}\frac{2^0}{0!} \\ &= 1 - 5e^{-2} = 0.3233. \end{aligned}$$

Παρατήρηση. Στο παράδειγμα αυτό, η προσέγγιση της ΤΜ X από την Y είναι πολύ καλή. Παρόλα αυτά οι πιθανότητες που υπολογίσαμε δεν ταυτίζονται απόλυτα. Συγκεκριμένα, οι αντίστοιχες πιθανότητες με περισσότερη ακρίβεια είναι 0.3233235612 και 0.3233235838. □

Παράδειγμα 3.6.3. *Έστω ότι για κάποιο λαχείο εκδίδονται* $2 \cdot 10^6$ *δελτία από τα οποία* 100 *έχουν σημαντικό κέρδος.*

Πόσα δελτία πρέπει να αγοράσουμε ώστε με πιθανότητα τουλάχιστον 95% *να εξασφαλίσουμε κάποιο σημαντικό κέρδος;*

Λύση. Η πιθανότητα ένα δέλτιο να κερδίσει είναι $p = \dfrac{10^2}{2 \cdot 10^6} = \dfrac{1}{2 \cdot 10^4}$.

Έστω ότι αγοράζουμε N δελτία και έστω X ο αριθμός των δελτίων που κερδίζουν ανάμεσα στα N δελτία που αγοράσαμε.

Η X ακολουθεί την διωνυμική κατανομή με παραμέτρους N και p. Μπορούμε να την προσεγγίσουμε χρησιμοποιώντας κατανομή Poisson με παράμετρο $\lambda = Np = \dfrac{N}{2 \cdot 10^4}$.

Θέλουμε

$$\begin{aligned} P(X \geq 1) \geq 0.95 &\Leftrightarrow P(X < 1) \leq 0.05 \Leftrightarrow P(X = 0) \leq 0.05 \\ &\Leftrightarrow e^{-\lambda}\frac{\lambda^0}{0!} \leq 0.05 \Leftrightarrow e^{\lambda} \geq 20 \\ &\Leftrightarrow \frac{N}{2 \cdot 10^4} \geq \ln 20 \\ &\Leftrightarrow N \geq 2 \cdot 10^4 \ln 20 = 59914.6. \end{aligned}$$

Άρα, πρέπει να αγοράσουμε τουλάχιστον $N = 59615$ δελτία. □

Παράδειγμα 3.6.4. *Μια από τις πιο διάσημες εμφανίσεις της κατανομής Poisson δόθηκε από τον Ρώσο μαθηματικό Ladislaus Bortkiewicz (1868 – 1931) ως μοντέλο του αριθμού των θανάτων από κλωτσιά αλόγου των αξιωματικών του ιππικού. Ο Bortkiewicz συνέλεξε στατιστικά στοιχεία για τους θανάτους αυτούς από* 10 *μονάδες ιππικού του πρωσικού στρατού σε μια περίοδο* 20 *ετών (από το* 1875 *μέχρι το* 1894*).*

Στον επόμενο πίνακα δίδονται τα στατιστικά που συνέλεξε ο Bortkiewicz για τις 10 *μονάδες του ιππικού (με λατινικούς αριθμούς) αναλυτικά ανά έτος (εμφανίζονται τα δύο τελευταία ψηφία του).*

	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94
II	-	-	-	2	-	2	-	-	1	1	-	-	2	1	1	-	-	2	-	-
III	-	-	-	1	1	1	2	-	2	-	-	-	1	-	1	2	1	-	-	-
IV	-	1	-	1	1	1	1	-	-	-	-	1	-	-	-	-	1	1	-	-
V	-	-	-	-	2	1	-	-	1	-	-	1	-	1	1	1	1	1	1	-
VII	1	-	1	-	-	-	1	-	1	1	-	-	2	-	-	2	1	-	2	-
VIII	1	-	-	-	1	-	-	1	-	-	-	-	1	-	-	-	1	1	-	1
IX	-	-	-	-	-	2	1	1	1	-	2	1	1	-	1	2	-	1	-	-
X	-	-	1	1	-	1	-	2	-	2	-	-	-	-	2	1	3	-	1	1
XIV	1	1	2	1	1	3	-	4	-	1	-	3	2	1	-	2	1	1	-	-
XV	-	1	-	-	-	-	-	1	-	1	1	-	-	-	2	2	-	-	-	-

Τα παραπάνω δεδομένα, που αφορούν $20 \cdot 10 = 200$ *υπηρεσιακά χρόνια, συνοψίζονται στον επόμενο πίνακα.*

αριθμός θανάτων ανά μονάδα ανά υπηρεσιακό έτος	*0*	*1*	*2*	*3*	*4*	*5 και πάνω*
αριθμός υπηρεσιακών ετών	*109*	*65*	*22*	*3*	*1*	*0*

Ο Bortkiewicz έδειξε ότι ο αριθμός των θανάτων ανά μονάδα ανά υπηρεσιακό έτος ακολουθεί την κατανομή Poisson.

Λύση. Πράγματι, ο συνολικός αριθμός των θανάτων από κλωτσιά αλόγου είναι

$$0 \cdot 109 + 1 \cdot 65 + 2 \cdot 22 + 3 \cdot 3 + 4 \cdot 1 + 5 \cdot 0 = 122,$$

άρα, ο μέσος αριθμός των θανάτων ανά μονάδα ανά υπηρεσιακό έτος είναι

$$\frac{122}{200} = 0.61.$$

Χρησιμοποιώντας την κατανομή Poisson με παράμετρο $\lambda = 0.61$ για την πιθανότητα να έχουμε $X = k$ θανάτους από κλωτσιά αλόγου σε ένα υπηρεσιακό έτος προκύπτει ότι:

Ο αναμενόμενος αριθμός των υπηρεσιακών ετών που θα έχουμε 0 θανάτους μέσα σε 200 υπηρεσιακά έτη ισούται με

$$N(0) = 200 \cdot P(X = 0) = 200 \cdot e^{-0.61} \frac{0.61^0}{0!} = 108.7 \simeq 109$$

Ομοίως, οι αναμενόμενοι αριθμοί των υπηρεσιακών ετών που θα έχουμε 1, 2, 3 και 4 θανάτους μέσα σε 200 υπηρεσιακά έτη ισούνται αντίστοιχα με

$$N(1) = 200 \cdot P(X = 1) = 200 \cdot e^{-0.61} \frac{0.61^1}{1!} = 66.3 \simeq 66$$

$$N(2) = 200 \cdot P(X = 2) = 200 \cdot e^{-0.61} \frac{0.61^2}{2!} = 20.2 \simeq 20$$

$$N(3) = 200 \cdot P(X = 3) = 200 \cdot e^{-0.61} \frac{0.61^3}{3!} = 4.11 \simeq 4$$

$$N(4) = 200 \cdot P(X = 4) = 200 \cdot e^{-0.61} \frac{0.61^4}{4!} = 0.63 \simeq 1$$

Η ομοιότητα των αναμενόμενων τιμών που υπολογίσαμε με τους πραγματικούς στατιστικούς αριθμούς είναι πολύ μεγάλη. Χρησιμοποιώντας το τεστ χ^2, μπορούμε να δείξουμε ότι οι διαφορές δεν είναι στατιστικά σημαντικές.

Το παράδειγμα αυτό (καθώς και ολόκληρο το βιβλίο στα Γερμανικά) του Bortkiewicz είναι διαθέσιμο στο σύνδεσμο

`https://archive.org/stream/dasgesetzderklei00bortrich#page/n65/mode/2up` □

3.7 Λυμένες ασκήσεις

Άσκηση 3.1. *Από τα 10^7 άτομα ενός πληθυσμού, οι 10^5 έχουν οπτική ίνα στο σπίτι. Από αυτόν τον πληθυσμό επιλέγουμε τυχαία (με επανατοποθέτηση) 150 άτομα για να συμμετάσχουν σε μια έρευνα αγοράς. Να δειχθεί ότι το πλήθος Y των ατόμων με οπτική ίνα ανάμεσα στους 150 επιλεχθέντες ακολουθεί την διωνυμική κατανομή με παραμέτρους $N = 150$ και $p = \frac{10^5}{10^7} = 0.01$.*

Λύση. Θεωρούμε ένα τυχαία επιλεγμένο άτομο και θέτουμε A το ενδεχόμενο να έχει οπτική ίνα και B το ενδεχόμενο να επιλεχθεί στους $N = 150$ της έρευνας. Είναι

$$p := P(A|B) = P(A \cap B)/P(B) = P(A)P(B)/P(B) = P(A) = 1/100.$$

Επομένως, κάθε άτομο που συμμετέχει στην έρευνα έχει οπτική ίνα με πιθανότητα p και άρα $Y \sim \text{Binom}(N, p)$. □

Άσκηση 3.2. *Να υπολογισθεί η μέση τιμή και η διακύμανση μια τυχαίας μεταβλητής X που ακολουθεί γεωμετρική κατανομή με παράμετρο p.*

Λύση. Υπενθυμίζονται οι τύποι

$$p \neq -1 \Rightarrow \sum_{k=0}^{n} p^k = \frac{1-p^{n+1}}{1-p}, \qquad |p| < 1 \Rightarrow \sum_{k=0}^{\infty} p^k = \frac{1}{1-p}$$

και

$$|p| < 1 \Rightarrow \frac{1}{(1-p)^2} = \left(\frac{1}{1-p}\right)' = \left(\sum_{k=0}^{\infty} p^k\right)' = \sum_{k=1}^{\infty} k p^{k-1}$$

Βάσει των παραπάνω, είναι $E(X) = \sum_{k=1}^{\infty} k(1-p)^{k-1}p = p\frac{1}{(1-(1-p))^2} = \frac{1}{p}$ και

$$\begin{aligned} E(X^2) &= \sum_{k=1}^{\infty} k^2(1-p)^{k-1}p \overset{q=1-p}{=} (1-q)\sum_{k=1}^{\infty} k^2 q^{k-1} = (1-q)\left(\sum_{k=1}^{\infty} kq^k\right)' = (1-q)\left(\frac{q}{(1-q)^2}\right)' \\ &= \frac{(1-q)^2 + 2(1-q)q}{(1-q)^3} = \frac{1+q}{(1-q)^2} = \frac{2-p}{p^2} \end{aligned}$$

οπότε $\text{V}(X) = E(X^2) - (E(X))^2 = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}$. □

Άσκηση 3.3. *Να αποδειχθεί ότι αν $X \sim \text{NB}(r, p)$, τότε $E(X) = r/p$ και $V(X) = r(1-p)/p^2$.*

Απόδειξη. Θα χρησιμοποιηθεί η γνωστή ιδιότητα

$$\frac{k}{r}\binom{r-1}{k-1} = \binom{r}{k}, \qquad k, r \in \mathbb{N}^*,$$

καθώς και οι γνωστές δυναμοσειρές

$$\frac{1}{(1-x)^{r+1}} = \sum_{n\geq 0}\binom{r+n}{r}x^n, \quad \frac{x^m}{(1-x)^{m+1}} = \sum_{n\geq 0}\binom{n}{m}x^n, \qquad r \in \mathbb{R}, m \in \mathbb{N}, |x| < 1.$$

Για τη μέση τιμή, έχουμε ότι

$$E(X) = \sum_{k\geq r} k\binom{k-1}{r-1}p^r(1-p)^{k-r} = rp^r\sum_{k\geq r}\binom{k}{r}(1-p)^{k-r} \overset{j=k-r}{=} rp^r\sum_{j\geq 0}\binom{j+r}{r}(1-p)^j = rp^r\frac{1}{(1-(1-p))^{r+1}} = \frac{r}{p}.$$

Επιπλέον,

$$E(X^2) = \sum_{k\geq r} k^2\binom{k-1}{r-1}p^r(1-p)^{k-r} = \frac{rp^r}{(1-p)^{r-1}}\sum_{k\geq r} k\binom{k}{r}(1-p)^{k-1} = \frac{rp^r}{(1-p)^{r-1}}\frac{d}{dp}\left(-\sum_{k\geq r}\binom{k}{r}(1-p)^k\right)$$
$$= \frac{rp^r}{(1-p)^{r-1}}\frac{d}{dp}\left(\frac{-(1-p)^r}{(1-(1-p))^{r+1}}\right) = \frac{rp^r}{(1-p)^{r-1}}\frac{r(1-p)^{r-1}p^{r+1}+(r+1)p^r(1-p)^r}{p^{2r+2}}$$
$$= \frac{r}{p^2}(rp+(r+1)(1-p)) = \frac{r}{p^2}(r-p+1).$$

Κατόπιν τούτων, η διακύμανση ισούται με

$$V(X) = E(X^2) - (E(X))^2 = \frac{r}{p^2}(r-p+1) - \frac{r^2}{p^2} = \frac{r^2-rp+r-r^2}{p^2} = \frac{r(1-p)}{p^2}.$$ □

Άσκηση 3.4. *Αν $X_1 \sim \text{Geom}(p_1)$ και $X_2 \sim \text{Geom}(p_2)$ ανεξάρτητες, να ευρεθεί η PMF της $Y = \min\{X_1, X_2\}$.*

Λύση. Για $k \in \mathbb{N}^*$, είναι

$$P(Y=k) = P(X_1=k)P(X_2\geq k) + P(X_1>k)P(X_2=k) = (1-p_1)^{k-1}p_1(1-p_2)^{k-1} + (1-p_2)^{k-1}p_2(1-p_1)^k$$
$$= ((1-p_1)(1-p_2))^{k-1}(p_1+p_2(1-p_1)) = ((1-p_1)(1-p_2))^{k-1}(1-(1-p_1)(1-p_2)) = (1-q)^{k-1}q$$

όπου $q = 1-(1-p_1)(1-p_2)$. Επομένως, $Y \sim \text{Geom}(q)$. □

Άσκηση 3.5. *Να αποδειχθεί ότι αν $\lim_{n\to\infty} np = \lambda \in \mathbb{R}$, τότε $\lim_{n\to\infty} Binom(n,p) = Poisson(\lambda)$.*

Απόδειξη. Θεωρούμε τις ακολουθίες (p_n) και (λ_n), με $\lambda_n = np_n$, και ότι $\lim\limits_{n\to\infty}\lambda_n = \lambda \in \mathbb{R}$, τότε

$$f_X(k) = \binom{n}{k}p^k(1-p)^{n-k} = \frac{n(n-1)\cdots(n-k+1)}{k!}\frac{\lambda_n^k}{n^k}\left(1-\frac{\lambda_n}{n}\right)^{n-k}$$
$$= \frac{\lambda_n^k}{k!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)\left(1-\frac{\lambda_n}{n}\right)^{-k}\left(1-\frac{\lambda_n}{n}\right)^n \overset{n\to\infty}{\to} \frac{\lambda^k}{k!}e^{-\lambda} = f_Y(k)$$

Για το τελευταίο όριο αρκεί να αποδειχθεί ότι αν για μια ακολουθία (a_n), είναι $a_n \to a \in \mathbb{R}$, τότε $(1+a_n/n)^n \to e^a$. Αρχικά αποδεικνύεται για $a=0$. Τότε, υπάρχει $n_0 \in \mathbb{N}^*$, ώστε $n \geq n_0 \Rightarrow |a_n| < 1$, οπότε

$$(1+|a_n|/n)^n < (1+1/n)^n < e, \qquad \text{για κάθε } n \geq n_0,$$

δηλαδή η παράσταση $(1+|a_n|/n)^n$ είναι φραγμένη (για κάθε n), έστω από το $M>0$. Επομένως,

$$|1+a_n/n|^k \leq (1+|a_n|/n)^k \leq (1+|a_n|/n)^n < M, \qquad \text{για κάθε } k \leq n,$$

οπότε,

$$|(1+\frac{a_n}{n})^n - 1| = |\frac{a_n}{n}||1+(1+\frac{a_n}{n})+\cdots+(1+\frac{a_n}{n})^{n-1}| \leq \frac{|a_n|}{n}(1+|1+\frac{a_n}{n}|+\cdots+|1+\frac{a_n}{n}|^{n-1})$$
$$\leq \frac{|a_n|}{n}nM = M|a_n|.$$

Παίρνοντας όρια καθώς $n\to\infty$ προκύπτει το ζητούμενο.

Αν τώρα $a_n \to a$, τότε $a_n - a \to 0$, οπότε

$$\frac{(1+a_n/n)^n}{(1+a/n)^n} = \left(\frac{n+a_n}{n+a}\right)^n = \left(1+\frac{a_n-a}{n+a}\right)^{n+a}\left(1+\frac{a_n-a}{n-a}\right)^{-a} \to 1.$$

Επομένως, $\lim\limits_{n\to\infty}(1+a_n/n)^n = \lim\limits_{n\to\infty}(1+a/n)^n = e^a$. □

Άσκηση 3.6. *Ο Γιώργος και ο Δημήτρης είναι συγκάτοικοι και κάθε φορά που έχουν να πλύνουν πιάτα ρίχνουν ένα νόμισμα. Όποιος φέρει πρώτος κορώνα κερδίζει και ο άλλος πλένει τα πιάτα. Ο Γιώργος παρατήρησε ότι όταν ρίχνει πρώτος κερδίζει συχνότερα, ενώ όταν ξεκινάει δεύτερος καταλήγει συχνότερα στο νεροχύτη. Υπολογίστε τις πιθανότητες και στις δυο περιπτώσεις. Είναι ο ισχυρισμός του Γιώργου σωστός;*

Λύση. Έστω X ο αριθμός ρίψεων μέχρι να έρθει κορώνα για πρώτη φορά. Η X ακολουθεί την γεωμετρική κατανομή με παράμετρο $p = 1/2$, οπότε $P(X = k) = (1-p)^{k-1}p$, για κάθε $k \in \mathbb{N}^*$.

Όταν ο Γιώργος ξεκινάει πρώτος, τότε κερδίζει όταν $X = 1, 3, 5, 7, \ldots$, με πιθανότητα

$$\sum_{k=0}^{\infty} P(X = 2k+1) = \sum_{k=0}^{\infty}(1-p)^{2k}p = p\sum_{k=0}^{\infty}((1-p)^2)^k = \frac{p}{1-(1-p)^2} = \frac{p}{p(2-p)} = \frac{1}{2-p},$$

ενώ, όταν ο Γιώργος ξεκινάει δεύτερος, τότε κερδίζει όταν $X = 2, 4, 6, 8, \ldots$, με πιθανότητα

$$\sum_{k=1} P(X = 2k) = \sum_{k=1}(1-p)^{2k-1}p = \frac{p}{1-p}\sum_{k=1}^{\infty}((1-p)^2)^k = \frac{p}{1-p}\frac{(1-p)^2}{1-(1-p)^2} = \frac{1-p}{2-p}.$$

Επομένως, η πιθανότητα να κερδίσει είναι μεγαλύτερη στην πρώτη περίπτωση. Ειδικά για $p = \dfrac{1}{2}$, οι αντίστοιχες πιθανότητες είναι $\dfrac{1}{2-\frac{1}{2}} = \dfrac{2}{3}$ και $\dfrac{1-\frac{1}{2}}{2-\frac{1}{2}} = \dfrac{1}{3}$, δηλαδή αυτός που αρχίζει να ρίχνει το νόμισμα πρώτος κερδίζει 2 στις 3 φορές. □

Άσκηση 3.7. *Έστω ότι ο αριθμός των γκολ που πετυχαίνει η Εθνική ομάδα ποδοσφαίρου στα εκτός έδρας παιχνίδια ακολουθεί την κατανομή Poisson. Επίσης, είναι γνωστό ότι στα εκτός έδρας παιχνίδια η Εθνική έχει την ίδια πιθανότητα να πετύχει ένα ή δύο γκολ. Να βρεθεί η πιθανότητα στο επόμενο εκτός έδρας παιχνίδι η Εθνική να πετύχει τέσσερα γκολ.*

Λύση. Έστω X ο αριθμός των γκόλ που πετυχαίνει η εθνική. Γνωρίζουμε ότι η X ακολουθεί την κατανομή Poisson με παράμετρο λ και έχει PMF $f_X(k) = P(X = k) = e^{-\lambda}\dfrac{\lambda^k}{k!}$. Επίσης,

$$P(X = 1) = P(X = 2) \Leftrightarrow e^{-\lambda}\frac{\lambda^1}{1!} = e^{-\lambda}\frac{\lambda^2}{2!} \Leftrightarrow \lambda = 2.$$

Επομένως,

$$P(X = 4) = e^{-2}\frac{2^4}{4!} = \frac{2}{3e^2} = 0.09.$$ □

Άσκηση 3.8. *Οι ημερήσιες συχνότητες επιθέσεων σε ένα site τον πρώτο χρόνο λειτουργίας του δίδονται στον επόμενο πίνακα*

Αριθμός επιθέσεων	*0*	*1*	*2*	*3*	*4*
Συχνότητα (ημέρες)	*223*	*110*	*27*	*4*	*1*

Αν υποθέσουμε ότι ο αριθμός X των επιθέσεων που γίνονται στο site μέσα σε μια ημέρα ακολουθεί την κατανομή Poisson να βρεθεί η παράμετρος λ η οποία προσεγγίζει τις παραπάνω παρατηρήσεις.

Στην συνέχεια να βρεθεί με βάση τον τύπο της κατανομής Poisson ο αναμενόμενος αριθμός ημερών στις οποίες θα έχουμε 0, 1, 2, 3, 4 *επιθέσεις αντίστοιχα μέσα στον επόμενο χρόνο.*

(Υπόδειξη: Να υπολογισθεί ο μέσος αριθμός επιθέσεων σε μια μέρα με βάση τα στοιχεία του πίνακα.)

Λύση. Με βάση τα στοιχεία του παραπάνω πίνακα, ο μέσος αριθμός επιθέσεων μέσα σε μια μέρα είναι

$$\mu = \frac{0 \cdot 223 + 1 \cdot 110 + 2 \cdot 27 + 3 \cdot 4 + 4 \cdot 1}{365} = \frac{36}{73} = 0.493.$$

Επομένως, $X \sim \text{Poisson}(\lambda)$, με $\lambda = \mu = 36/73 = 0.493 \approx 0.5$, και

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!} = e^{-0.5}\frac{(0.5)^k}{k!}, k \in \mathbb{N}.$$

Άρα,

$P(X = 0) = e^{-0.5}\dfrac{(0.5)^0}{0!} = 0.606.$

$P(X = 1) = e^{-0.5}\dfrac{(0.5)^1}{1!} = 0.303.$

$P(X = 2) = e^{-0.5}\dfrac{(0.5)^2}{2!} = 0.075.$

$P(X = 3) = e^{-0.5}\dfrac{(0.5)^3}{3!} = 0.012.$

$P(X = 4) = e^{-0.5}\dfrac{(0.5)^3}{4!} = 0.0015.$

Επομένως, μέσα σε ένα χρόνο (365 ημέρες), ο μέσος αριθμός ημερών με 0, 1, 2, 3 και 4 επιθέσεις θα είναι αντίστοιχα $365 \cdot P(X = 0) = 221.38 \approx 221$, $365 \cdot P(X = 1) = 110.69 = 111$, $365 \cdot P(X = 2) = 27.67 \approx 28$, $365 \cdot P(X = 3) = 4.61 \approx 5$ και $365 \cdot P(X = 4) = 0.57 \approx 1$. □

Άσκηση 3.9. *Ένα κουτί περιέχει 6 μπλε και 4 κόκκινες σφαίρες.*

i) *Διαλέγουμε στην τύχη* 5 *σφαίρες χωρίς επανατοποθέτηση. Ποια είναι η πιθανότητα να βγούν* 3 *κόκκινες σφαίρες;*

ii) *Διαλέγουμε στην τύχη μια σφαίρα χωρίς επανατοποθέτηση. Επαναλαμβάνουμε το ίδιο* 5 *φορές. Ποια είναι η πιθανότητα να βγουν* 3 *κόκκινες σφαίρες;*

Λύση.

i) (1ος τρόπος) Ο αριθμός X των κόκκινων σφαιρών που διαλέξαμε ακολουθεί την υπεργεωμετρική κατανομή με παραμέτρους $M = 10$, $n = 4$, $N = 5$. Άρα,

$$P(X = 3) = \frac{\binom{4}{3}\binom{6}{2}}{\binom{10}{5}} = \frac{5}{21} = 0.238.$$

(2ος τρόπος) Όλες οι δυνατές περιπτώσεις είναι $\binom{10}{5}$. Οι ευνοϊκές περιπτώσεις είναι $\binom{4}{3}\binom{6}{2}$. Άρα, η ζητούμενη πιθανότητα είναι

$$\frac{\binom{4}{3}\binom{6}{2}}{\binom{10}{5}}.$$

ii) Η απάντηση είναι ίδια με το πρώτο ερώτημα. □

3.8 Άλυτες ασκήσεις

Άσκηση 3.10. *Να αποδειχθεί ότι αν* $X_i \sim \text{Binom}(N_i, p)$, $i \in [n]$, *ανεξάρτητες, τότε* $X_1 + \cdots + X_n \sim \text{Binom}(N_1 + \cdots + N_n, p)$. *(Υπόδειξη: Να αποδειχθεί πρώτα για* $n = 2$ *και στη συνέχεια να γενικευθεί επαγωγικά).*

Άσκηση 3.11. *Να αποδειχθεί ότι αν* $X \sim \text{Geom}(p)$ *και* $m, n \in \mathbb{N}^*$, *τότε*

$$P(X \geq m + n | X > n) = P(X \geq m).$$

Άσκηση 3.12. *Να αποδειχθεί ότι αν* $X_i \sim \text{Poisson}(\lambda_i)$, $i \in [n]$, *ανεξάρτητες, τότε* $X_1 + \cdots + X_n \sim \text{Poisson}(\lambda_1 + \cdots + \lambda_n)$. *(Υπόδειξη: Να αποδειχθεί πρώτα για* $n = 2$ *και στη συνέχεια να γενικευθεί επαγωγικά).*

Άσκηση 3.13. *Έστω ΤΜ* $X \sim \text{Binom}(n, p)$, $p \in (0, 1)$. *Να αποδειχθεί ότι*

i) $f_X(k+1) = \dfrac{p}{1-p}\dfrac{n-k}{k+1} f_X(k)$, *για* $k \in \{0, 1, 2, \ldots, n-1\}$.

ii) Η $f_X(k)$ *είναι αύξουσα για* $k \leq \lfloor (n+1)p \rfloor$ *και φθίνουσα για* $k \geq \lfloor (n+1)p \rfloor$.

Άσκηση 3.14. *Ένα πείραμα έχει* r *δυνατά αποτελέσματα, τα* $1, 2, \ldots, r$, *με πιθανότητες* $p_1, p_2, \ldots, p_r$ *αντίστοιχα, όπου* $\sum_{i=1}^{r} p_i = 1$. *Αν το πείραμα επαναλαμβάνεται* n *φορές και* X_i *είναι το πλήθος των επαναλήψεων με αποτέλεσμα* $i \in [r]$,

i) Ποια είναι η κατανομή της X_1;

ii) Είναι οι X_1 *και* X_2 *ανεξάρτητες;*

iii) Ποια είναι η κατανομή της $X_1 + X_2$;

iv) Για $k < r$, *ποια είναι η κατανομή της* $X = \sum_{i=1}^{k} X_i$;

Άσκηση 3.15. *Το πλήθος των φορών που ένας άνθρωπος αρρωσταίνει το χρόνο με κρύωμα ακολουθεί κατανομή Poisson με παράμετρο* $\lambda = 3$. *Ένα νέο φάρμακο μειώνει αυτή τη συχνότητα σε* $\lambda = 2$ *για το* 75% *του πληθυσμού, ενώ δεν έχει αποτέλεσμα στο υπόλοιπο* 25%. *Αν ένας άνθρωπος που πήρε το φάρμακο αρρώστησε 0 φορές κατά τη διάρκεια του έτους, ποια η πιθανότητα να τον ωφέλησε το φάρμακο;*

Άσκηση 3.16. *Ένα ηλεκτρονικό κατάστημα παραλαμβάνει ένα φορτίο με 100 ηλεκτρονικές συσκευές. Ελέγχει 10 τυχαία επιλεγμένες από αυτές και αν βρει πάνω από μία ελαττωματική, τότε επιστρέφει το φορτίο στον προμηθευτή. Αν το φορτίο περιέχει 20 ελαττωματικές συσκευές, ποια είναι η πιθανότητα να μην επιστραφεί;*

Άσκηση 3.17. *Σε κάθε χαρακτήρα που πληκτρολογείται σε ένα κείμενο μπορεί υπάρξει τυπογραφικό λάθος ανεξάρτητα με πιθανότητα* p *ή να είναι σωστός με πιθανότητα* $1 - p$. *Έστω* n *ο αριθμός των χαρακτήρων του κειμένου και* X *ο συνολικός αριθμός των τυπογραφικών λαθών.*

i) *Να βρεθεί η πιθανότητα* $P(X = k)$.

ii) *Να δειχθεί ότι* $E(X) = np$.

iii) *Να δειχθεί ότι* $\text{V}(X) = np(1-p)$.

Άσκηση 3.18. *Έστω ότι η ΤΜ* X *ακολουθεί την διωνυμική κατανομή* $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$.

i) *Για σταθερά n, p να βρεθεί η τιμή του k για την οποία η πιθανότητα $P(X = k)$ γίνεται μέγιστη.*

ii) *Για σταθερά k, p να βρεθεί η τιμή του n για την οποία η πιθανότητα $P(X = k)$ γίνεται μέγιστη.*

Άσκηση 3.19. *Η τυχαία μεταβλητή X ακολουθεί την διωνυμική κατανομή* $\mathrm{Binom}(n, p)$ *με αναμενόμενη τιμή* 20 *και διακύμανση* 16.

i) *Να βρεθεί η παράμετρος n.*

ii) *Να υπολογισθεί η $E(X^2)$.*

Άσκηση 3.20. *Ένας μπασκετμπολίστας όταν ρίχνει για τρίποντο πετυχαίνει σε τριπλάσιο αριθμό προσπαθειών απ' ό,τι αποτυχνάνει.*

i) *Να βρεθεί η πιθανότητα να ρίξει 10 φορές και να πετύχει στις 8 από αυτές.*

ii) *Να βρεθεί ο αναμενόμενος αριθμός πόντων που θα συγκεντρώσει αν ρίξει για τρίποντο 10 φορές.*

Άσκηση 3.21. *Να εξηγηθεί γιατί δεν χρησιμοποιούμε την γεωμετρική κατανομή για να μοντελοποιήσουμε τον χρόνο αναμονής για να διαλέξουμε έναν άσσο από μια τράπουλα από την οποία εξάγουμε ένα φύλλο κάθε φορά χωρίς επανατοποθέτηση.*

Άσκηση 3.22. *Έστω ότι η ΤΜ X ακολουθεί την κατανομή Poisson με παράμετρο λ.*

i) *Για σταθερό λ να βρεθεί η τιμή του k για την οποία η πιθανότητα $P(X = k)$ γίνεται μέγιστη.*

ii) *Για σταθερό k να βρεθεί η τιμή του λ για την οποία η πιθανότητα $P(X = k)$ γίνεται μέγιστη.*

Άσκηση 3.23. *Έστω ότι η ΤΜ X ακολουθεί την κατανομή Poisson με παράμετρο λ. Να δειχθεί ότι η πιθανότητα η X να λάβει άρτια τιμή ισούται με $e^{-\lambda}\cosh\lambda$.*

Άσκηση 3.24. *Υποτίθεται ότι ο αριθμός X των ατόμων σε ένα πληθυσμό που τα δακτυλικά τους αποτυπώματα έχουν συγκεκριμένο τύπο ακολουθεί την κατανομή Poisson για κάποια παράμετρο λ. Να εξηγηθεί πότε και γιατί αυτή είναι μια εύλογη υπόθεση.*

Άσκηση 3.25. *Μια αεροπορική εταιρεία πουλάει $m+n$ εισιτήρια για μια πτήση με n καθίσματα. Κάθε πελάτης έχει πιθανότητα p ανεξάρτητα από τους υπόλοιπους να μην εμφανισθεί για επιβίβαση. Οι κενές θέσεις κοστίζουν c ευρώ, ενώ ένας επιβάτης που δεν χωράει να μπει για να πετάξει αποζημιώνεται με b ευρώ. Να βρεθεί η τιμή του m για την οποία η αναμενόμενη ζημία λόγω των κρατήσεων να είναι ελάχιστη.*

Κεφάλαιο 4

Συνεχείς τυχαίες μεταβλητές

Μια ΤΜ X ονομάζεται **συνεχής** όταν το σύνολο τιμών S αυτής είναι διάστημα ή ένωση διαστημάτων.

Η μη αρνητική ολοκληρώσιμη συνάρτηση f/S για την οποία ισχύει

$$\int_S f(x)dx = 1 \qquad \text{και} \qquad P(X \in A) = \int_A f(x)dx, \text{ για κάθε διάστημα } A \subseteq S$$

ονομάζεται **συνάρτηση πυκνότητας πιθανότητας** (PDF) της ΤΜ X.

Συνήθως επεκτείνουμε την f σε όλο το $\mathbb{R}$, θέτοντας $f(x) = 0$ για κάθε $x \in \mathbb{R} \setminus S$. Τότε, η αθροιστική συνάρτηση κατανομής πιθανότητας (CDF) F ισούται με

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt, \qquad \text{για κάθε } x \in \mathbb{R},$$

οπότε, σύμφωνα με το πρώτο Θεμελιώδες Θεώρημα του Ολοκληρωτικού Λογισμού, η F είναι συνεχής και επιπλέον είναι παραγωγίσιμη σε κάθε x όπου η f είναι συνεχής, με

$$f(x) = F'(x), \qquad \text{για κάθε } x \text{ όπου } f \text{ συνεχής.}$$

Βασικές ιδιότητες:

- Για κάθε a, b, με $a \leq b$, είναι

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = \int_a^b f(x)dx.$$

- Για κάθε $a \in \mathbb{R}$, η πιθανότητα η X να ισούται ακριβώς με a είναι 0, δηλαδή

$$P(X = a) = \int_a^a f(x)dx = 0$$

- $P(X \geq a) = P(a \leq X < +\infty) = \int_a^{\infty} f(x)dx$

- $P(a \leq X \leq b) = F(b) - F(a)$

Μέση τιμή και διακύμανση: Η μέση τιμή μ και η διακύμανση σ^2 της συνεχούς ΤΜ X δίδονται από τους τύπους

$$\mu = E(X) = \int_S xf(x)dx = \int_{-\infty}^{\infty} xf(x)dx$$

$$\sigma^2 = \mathrm{V}(X) = E((X - E(X))^2) = E((X-\mu)^2) = \int_S (x-\mu)^2 f(x)dx$$

Γενικότερα, για οποιαδήποτε συνάρτηση $g : S \to \mathbb{R}$, είναι

$$E(g(X)) = \int_S g(x)f(x)dx = \int_{-\infty}^{\infty} g(x)f(x)dx$$

Εύκολα αποδεικνύονται οι ακόλουθες ιδιότητες, όπως και στην περίπτωση διακριτής ΤΜ:

1. $E(aX + bY) = aE(X) + bE(Y)$
2. $\mathrm{V}(aX + b) = a^2\mathrm{V}(X)$.
3. $\mathrm{V}(X) = E(X^2) - (E(X))^2$.

Ανάλογα αποδεικνύεται και η ακόλουθη πρόταση:

Πρόταση 4.1. *Αν οι συνεχείς τ. μ. $X_1, X_2, \ldots, X_n$ είναι ανεξάρτητες και $g_1, g_2, \ldots, g_n$ οποιεσδήποτε συναρτήσεις, όπου $g_i : S_{X_i} \to \mathbb{R}$, τότε:*

1. *Οι $g_1(X_1), g_2(X_2), \ldots, g_n(X_n)$ είναι ανεξάρτητες*
2. *$E(g_1(X_1)g_2(X_2)\cdots g_n(X_n)) = E(g_1(X_1))E(g_2(X_2))\cdots E(g_n(X_n))$*
3. *$V(g_1(X_1) + g_2(X_2) + \cdots + g_n(X_n)) = V(g_1(X_1)) + V(g_2(X_2)) + \cdots + V(g_n(X_n))$*

Παράδειγμα 4.0.1. *Να βρεθεί η σταθερά c ώστε η συνάρτηση $f(x) = \begin{cases} cx^2, & x \in [0,1] \\ 0, & x \notin [0,1] \end{cases}$ να είναι PDF κάποιας ΤΜ X. Στη συνέχεια, να βρεθεί ο τύπος της CDF, η πιθανότητα $P(X \le 2/3)$, η μέση τιμή και η διακύμανση της X.*

Λύση. Πρέπει

$$\int_{-\infty}^{+\infty} f(x)dx = 1 \Leftrightarrow \int_{-\infty}^{+\infty} cx^2dx = 1 \Leftrightarrow c\int_0^1 x^2dx = 1 \Leftrightarrow c\left[\frac{x^3}{3}\right]_0^1 = 1 \Leftrightarrow c\left(\frac{1}{3} - 0\right) = 1 \Leftrightarrow c = 3.$$

Άρα,

$$f(x) = \begin{cases} 3x^2, & x \in [0,1] \\ 0, & x \notin [0,1] \end{cases}$$

Επομένως,

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)dt = \begin{cases} 0, & x \in (-\infty, 0) \\ x^3 & x \in [0,1] \\ 1 & x \in (1, +\infty) \end{cases}$$

Άρα,

$$P(X \leq \frac{2}{3}) = F(\frac{2}{3}) = \left(\frac{2}{3}\right)^3 = \frac{8}{27}.$$

Επίσης,

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx = \int_0^1 x3x^2dx = \left[\frac{3x^4}{4}\right]_0^1 = \frac{3}{4}.$$

Τέλος,

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_0^1 3x^4dx = \left[\frac{3x^5}{5}\right]_0^1 = \frac{3}{5},$$

οπότε

$$\mathrm{V}(X) = E(X^2) - E(X)^2 = \frac{3}{5} - \frac{9}{16} = \frac{3}{80}.$$

□

Παράδειγμα 4.0.2. *Έστω ότι ρίχνουμε με κλειστά μάτια ένα μαγνητικό βελάκι σε ένα στόχο με σχήμα κυκλικού δίσκου ακτίνα 10 εκατοστών με κέντρο την αρχή των αξόνων. Έστω D η ΤΜ που ισούται με την απόσταση του σημείου που στοχεύσαμε με το βελάκι από την αρχή των αξόνων. Θεωρούμε ότι το βελάκι καταλήγει πάντα εντός στόχου.)*

α) Να βρεθεί το σύνολο τιμών της D, η CDF και η PDF της D.

β) Να βρεθεί η μέση τιμή και η διακύμανση της D.

γ) Να υπολογισθούν οι πιθανότητες $P(4 \leq D \leq 6)$, $P(D \leq 1)$.

Λύση. Έστω D η ΤΜ που αντιστοιχεί στην απόσταση από το κέντρο. Το σύνολο τιμών της είναι το $S_D = [0, r]$, όπου $r = 10$. Θεωρώντας όλα τα σημεία ισοπίθανα, θα είναι

$$F_D(x) = P(D \leq x) = \frac{\pi x^2}{\pi r^2} = \frac{x^2}{r^2},$$

οπότε, (αφού η F είναι παραγωγίσιμη)

$$f_D(x) = F'_D(x) = \frac{2x}{r^2}.$$

Κατόπιν τούτου, είναι

$$E(D) = \int_0^r xf_D(x)dx = \int_0^r \frac{2x^2}{r^2}dx = \frac{2}{3r^2}\left[x^3\right]_0^r = \frac{2r}{3}$$

$$E(D^2) = \int_0^r x^2 f_D(x)dx = \int_0^r \frac{2x^3}{r^2}dx = \frac{2}{4r^2}\left[x^4\right]_0^r = \frac{r^2}{2}$$

$$\mathrm{V}(D) = E(D^2) - (E(D))^2 = \frac{r^2}{2} - \frac{4r^2}{9} = \frac{r^2}{18}$$

Τέλος, είναι

$$P(4 \leq D \leq 6) = F_D(6) - F_D(4) = \frac{6^2 - 4^2}{r^2} = \frac{20}{r^2}, \quad P(D \leq 1) = F_D(1) = \frac{1}{r^2}$$

□

4.1 Λυμένες Ασκήσεις

Άσκηση 4.1. *Μια συνεχής τυχαία μεταβλητή X με τιμές στο $\mathbb{R}$ έχει PDF*

$$f(x) = \begin{cases} ax^2e^{-kx} & \text{αν } 0 \le x < +\infty, \\ 0 & \text{αλλιώς,} \end{cases}$$

όπου $k > 0$.

i) *Να βρεθεί η σταθερά a.*

ii) *Να βρεθεί η CDF της X.*

iii) *Να βρεθεί η πιθανότητα $P(0 \le X \le 1/k)$.*

Λύση. *i*) Πρέπει $\int_{-\infty}^{+\infty} f(x)dx = 1$, δηλαδή $\int_0^{+\infty} ax^2e^{-kx}dx = 1$. Θέτουμε

$$I(x) = \int_0^x t^2e^{-kt}dt \qquad \text{και} \quad I = \lim_{x\to+\infty} I(x) = \int_0^{+\infty} t^2e^{-kt}dt$$

και με παραγοντική ολοκλήρωση έχουμε ότι

$$\begin{aligned} I(x) &= \int_0^x t^2\left(\frac{e^{-kt}}{-k}\right)'dt = \left[\frac{t^2e^{-kt}}{-k}\right]_0^x + \frac{2}{k}\int_0^x te^{-kt}dt = -\frac{x^2e^{-kx}}{k} + \frac{2}{k}\int_0^x t\left(\frac{e^{-kt}}{-k}\right)'dt \\ &= -\frac{x^2e^{-kx}}{k} + \frac{2}{k}\left(\left[\frac{te^{-kt}}{-k}\right]_0^x - \int_0^x \frac{e^{-kt}}{-k}\right)dt = -\frac{x^2e^{-kx}}{k} - \frac{2xe^{-kx}}{k^2} + \frac{2}{k^2}\int_0^x e^{-kt}dt \\ &= -\frac{x^2e^{-kx}}{k} - \frac{2xe^{-kx}}{k^2} + \frac{2}{k^2}\left[\frac{e^{-kt}}{-k}\right]_0^x = -\frac{x^2e^{-kx}}{k} - \frac{2xe^{-kx}}{k^2} - \frac{2}{k^3}e^{-kx} + \frac{2}{k^3}. \end{aligned}$$

Παίρνοντας το όριο όταν $x \to +\infty$, βρίσκουμε ότι $I = \lim_{x\to+\infty} I(x) = 2/k^3$. Εναλλακτικά, χρησιμοποιώντας μετασχηματισμό Laplace, βρίσκουμε απευθείας το ίδιο αποτέλεσμα:

$$I = \mathcal{L}[x^2](k) = \frac{\Gamma(3)}{k^3} = \frac{2}{k^3}.$$

Επομένως, $a \cdot I = 1 \Leftrightarrow a = k^3/2$ και άρα $x \in [0, +\infty) \Rightarrow f(x) = \dfrac{k^3}{2}x^2e^{-kx}$.

ii) Για την CDF $F(x)$ της X, έχουμε ότι $x < 0 \Rightarrow F(x) = 0$ και

$$x \ge 0 \Rightarrow F(x) = \int_0^x f(t)dt = \frac{k^3}{2}I(x) = 1 - \left(1 + kx + \frac{k^2x^2}{2}\right)e^{-kx}.$$

iii) Ισχύει ότι

$$P(0 \le X \le 1/k) = F(1/k) - F(0) = 1 - e^{-1}\left(1 + 1 + \frac{1}{2}\right) - 0 = 1 - \frac{5}{2e}. \qquad \square$$

Άσκηση 4.2. *Έστω X μια συνεχής ΤΜ με PDF $f(x)$ και CDF $F(x)$. Να προσδιορισθούν η CDF $F_Y(x)$ και η PDF $f_Y(x)$ της ΤΜ $Y = |X|$. Αν $f(x) = 1/3$, $x \in [-1, 2]$, να βρεθεί η $f_Y(x)$ της Y.*

Λύση. Προφανώς, επειδή $Y \geq 0$, είναι $F_Y(x) = 0$, για κάθε $x \leq 0$, ενώ για $x > 0$ είναι

$$F_Y(x) = P(Y \leq x) = P(|X| \leq x) = P(-x \leq X \leq x) = F(x) - F(-x), \qquad x > 0.$$

Προφανώς, επειδή $Y \geq 0$, είναι $f_Y(x) = 0$, όταν $x < 0$, ενώ σε κάθε $x \geq 0$ όπου f_Y, f συνεχείς, είναι

$$x > 0 \Rightarrow f_Y(x) = F_Y'(x) = F'(x) - (F(-x))' = f(x) + f(-x)$$

Επομένως, μια PDF για την ΤΜ Y είναι η συνάρτηση

$$f_Y(x) = \begin{cases} 0, & x < 0, \\ f(x) + f(-x)), & x \geq 0 \end{cases}$$

Πράγματι, για $x \geq 0$, είναι

$$\begin{aligned} \int_{-\infty}^{x} f_Y(t)dt = \int_{0}^{x} f_Y(t)dt = \int_{0}^{x} f(t)dt + \int_{0}^{x} f(-t)dt = \int_{0}^{x} f(t)dt - \int_{0}^{-x} f(t)dt \\ = \int_{-x}^{x} f(t)dt = F(x) - F(-x) = F_Y(x). \end{aligned}$$

Αν $f(x) = 1/3$, όταν $x \in [-1, 2]$, τότε έχουμε ότι

$$f_Y(x) = \begin{cases} f(x) + f(-x), & 0 \leq x \leq 2, \\ 0, & \text{αλλιώς}, \end{cases} = \begin{cases} f(x) + f(-x), & 0 \leq x \leq 1, \\ f(x), & 1 \leq x \leq 2, \\ 0, & \text{αλλιώς}, \end{cases} = \begin{cases} 2/3, & 0 \leq x \leq 1, \\ 1/3, & 1 \leq x \leq 2, \\ 0, & \text{αλλιώς}. \end{cases}$$

□

4.2 Ασκήσεις προς επίλυση

1. Να βρεθεί (αν υπάρχει) σταθερά C ώστε η συνάρτηση $f(x) = P(X = x)$ να είναι συνάρτηση πυκνότητας πιθανότητας μιας συνεχούς τυχαίας μεταβλητής X με τιμές στο $\mathbb{R}$ όταν

   i) $f(x) = \begin{cases} C(4x^2 + 5x + 2), & \text{αν } 0 \leq x \leq 2 \\ 0, & \text{αλλιώς} \end{cases}$.

   ii) $f(x) = \begin{cases} Cxe^{-x^2}, & \text{αν } x \geq 0 \\ 0, & \text{αλλιώς} \end{cases}$.

   iii) $f(x) = Ce^{-x-e^{-x}}$, για κάθε $x \in \mathbb{R}$.

2. Να εξετασθεί αν η συνάρτηση $F(x)$ είναι (αθροιστική) συνάρτηση κατανομής πιθανότητας μιας συνεχούς τυχαίας μεταβλητής X με τιμές στο $\mathbb{R}$ όταν

   i) $F(x) = e^{-x}$ για κάθε $x \in \mathbb{R}$.

   ii) $F(x) = 1 - \dfrac{x^2}{1 + x^2}$ για κάθε $x \in \mathbb{R}$.

3. Να δειχθεί ότι η συνάρτηση $f(x) = \dfrac{1}{\pi(1+x^2)}$ είναι συνάρτηση πυκνότητας πιθανότητας μιας συνεχούς τυχαίας μεταβλητής X με τιμές στο $\mathbb{R}$, η οποία ονομάζεται *κατανομή Cauchy*.

   Να δειχθεί ότι η ΤΜ X που ακολουθεί την κατανομή Cauchy δεν έχει μέση τιμή και επιπλέον έχει άπειρη διακύμανση.

   Σημείωση: Η κατανομή Cauchy χρησιμοποιείται συχνά σαν τεστ όταν θέλουμε να ελέγξουμε αν μια ιδιότητα ισχύει για όλες τις κατανομές.

4. Η συνάρτηση πυκνότητας πιθανότητας της συνεχούς τυχαίας μεταβλητής X δίδεται από τον τύπο

$$f(x) = \begin{cases} ax, & x \in [0,3) \\ a(6-x), & x \in (3,6] \\ 0, & \text{αλλιώς} \end{cases}$$

   i) Να βρεθεί η σταθερά a. (Απ. $a = 1/9$.)

   ii) Να βρεθεί η συνάρτηση κατανομής πιθανότητας $F(x)$ της ΤΜ X.

   (Απ. $F(x) = \begin{cases} 0, & x < 0 \\ \frac{x^2}{18} & x \in [0,3) \\ \frac{2x}{3} - \frac{x^2}{18} - 1 & x \in [3,6) \\ 1 & x \geq 6 \end{cases}$.)

   iii) Να σχεδιασθούν οι γραφικές παραστάσεις των $f(x)$ και $F(x)$.

   iv) Να βρεθεί η αναμενόμενη τιμή $E(X)$ της X. (Απ. $E(X) = 3$.)

   v) Να βρεθεί η διακύμανση $\mathrm{Var}(X)$ της X. (Απ. $\mathrm{Var}(X) = 3/2$.)

5. Μια συνεχής τυχαία μεταβλητή X με τιμές στο $\mathbb{R}$ έχει συνάρτηση πυκνότητας πιθανότητας (σππ)

$$f(x) = \begin{cases} Cx & x \in [0,2] \\ 0, & \text{αλλιώς} \end{cases}$$

   i) Να βρεθεί η σταθερά C.

   ii) Να βρεθεί η συνάρτηση κατανομής πιθανότητας $F(x)$ της X.

   iii) Να σχεδιασθούν οι γραφικές παραστάσεις των $f(x)$ και $F(x)$.

   iv) Να υπολογισθεί η αναμενόμενη τιμή $E(X)$ της X.

   v) Να υπολογισθεί η διακύμανση $\mathrm{Var}(X)$ της X.

   vi) Να υπολογισθούν οι πιθανότητες $P(X \leq 1)$, $P(X > \frac{3}{2})$, $P(1 \leq X < \frac{3}{2})$,

Κεφάλαιο 5

Σημαντικές συνεχείς κατανομές

5.1 Ομοιόμορφη κατανομή

Μια ΤΜ X η οποία λαμβάνει τιμές στο $[a, b]$, όπου $a < b$, λέμε ότι ακολουθεί την *ομοιόμορφη κατανομή* και γράφουμε $X \sim U(a, b)$, αν η συνάρτηση πυκνότητας πιθανότητας της X δίδεται από τον τύπο

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & \text{αν } x \in [a, b] \\ 0, & \text{αν } x \notin [a, b] \end{cases}$$

οπότε για τη συνάρτηση κατανομής της X ισχύει

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx = \int_{a}^{x} \frac{1}{b-a}dx = \begin{cases} 0, & \text{αν } x < a \\ \dfrac{x-a}{b-a}, & \text{αν } x \in [a, b] \\ 1, & \text{αν } x > b \end{cases}$$

Άμεσα αποδεικνύεται ότι η μέση τιμή και η διακύμανση της X είναι αντίστοιχα ίσες με

$$E(X) = \frac{a+b}{2} \text{ και } \mathrm{V}(X) = \frac{(b-a)^2}{12}.$$

Πράγματι, ισχύει ότι

$$E(X) = \int_{a}^{b} xf(x)dx = \int_{a}^{b} \frac{x}{b-a}dx = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_{a}^{b} = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}.$$

Επίσης,

$$E(X^2) = \int_{a}^{b} x^2 f(x)dx = \int_{a}^{b} \frac{x^2}{b-a}dx = \frac{1}{b-a}\left[\frac{x^3}{3}\right]_{a}^{b} = \frac{b^3-a^3}{3(b-a)} = \frac{a^2+ab+b^2}{3},$$

οπότε

$$\mathrm{V}(X) = E(X^2) - E(X)^2 = \frac{a^2+ab+b^2}{3} - \frac{a^2+2ab+b^2}{4} = \frac{4a^2+4ab+4b^2-3a^2-6ab-3b^2}{12} = \frac{(b-a)^2}{12}.$$

Παράδειγμα 5.1.1. *Τις βραδινές ώρες οι συρμοί του μετρό διέρχονται κάθε 10 λεπτά από κάποιο σταθμό. Αν κάποιος εισέλθει τυχαία στο σταθμό να βρεθεί:*

i) *ο μέσος χρόνος αναμονής,*

ii) *η διακύμανση του χρόνου αναμονής,*

iii) *η πιθανότητα να περιμένει τουλάχιστον 6 λεπτά για να έρθει ο συρμός.*

Λύση. Έστω X η ΤΜ χρόνος αναμονής στο σταθμό (σε λεπτά).

Η X είναι συνεχής, λαμβάνει τιμές στο διάστημα $[0, 10]$ και ακολουθεί την ομοιόμορφη κατανομή.

i) Ο μέσος χρόνος αναμονής είναι $\mu = E(X) = \frac{0+10}{2} = 5$ λεπτά.

ii) Η διακύμανση του χρόνου αναμονής είναι $\sigma^2 = \mathrm{V}(X) = \dfrac{(10-0)^2}{12} = \frac{100}{12} = 8.33$ λεπτά.

iii) Η συνάρτηση κατανομής πιθανότητας της X δίδεται από τον τύπο

$$F(x) = P(X \le x) = \begin{cases} 0, & \text{αν } x < 0 \\ \dfrac{x}{10}, & \text{αν } x \in [0, 10] \\ 1, & \text{αν } x > 10 \end{cases}$$

οπότε

$$P(X \ge 6) = 1 - P(x < 6) = 1 - F(6) = 1 - \frac{6}{10} = 0.4. \qquad \square$$

5.2 Εκθετική κατανομή

Μια ΤΜ X η οποία λαμβάνει τιμές στο $[0, +\infty)$ λέμε ότι ακολουθεί την *εκθετική κατανομή με παράμετρο* $\lambda > 0$ και γράφουμε $X \sim E(\lambda)$, αν η συνάρτηση πυκνότητας πιθανότητας της X δίδεται από τον τύπο

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \ge 0, \\ 0, & x < 0, \end{cases}$$

οπότε η κατανομή πιθανότητας της X δίδεται από τον τύπο

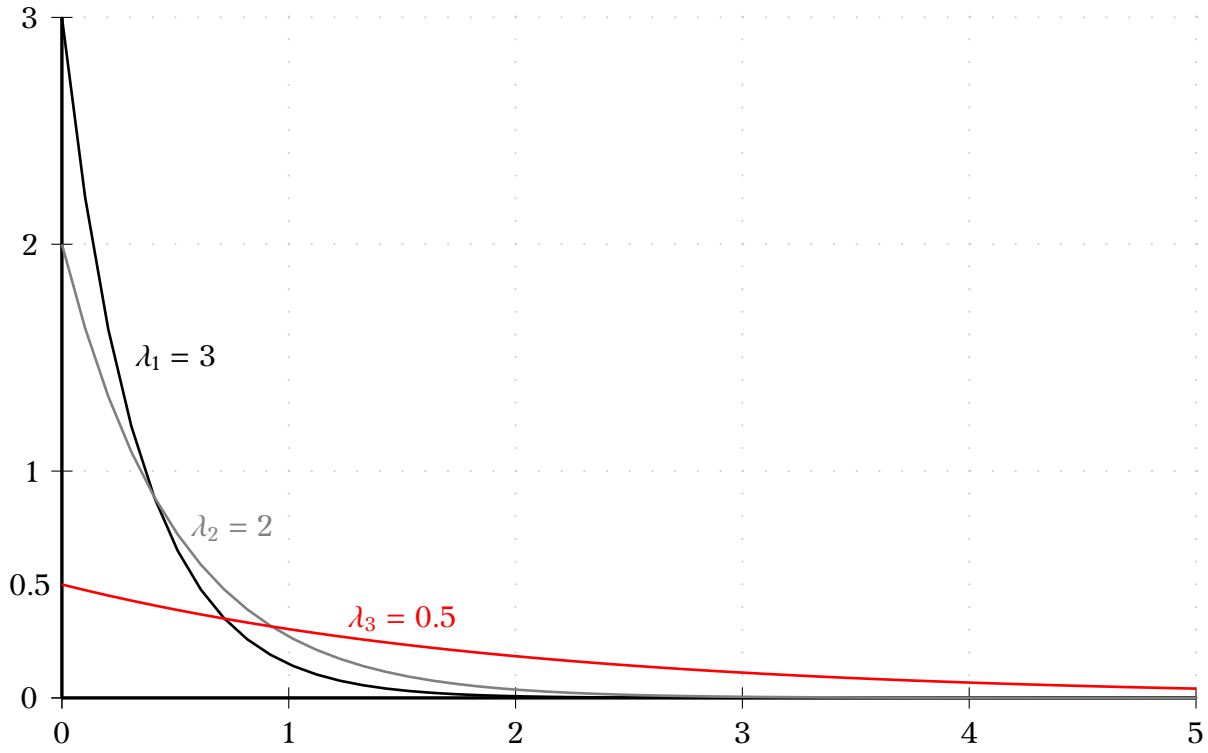
$$F(x) = P(X \le x) = \int_{-\infty}^{x} \lambda e^{-\lambda t} dt = \int_{0}^{x} \lambda e^{-\lambda t} dt = \begin{cases} 1 - e^{-\lambda x}, & x \ge 0, \\ 0, & x < 0. \end{cases}$$

Προφανώς, ισχύει ότι

$$P(X > x) = e^{-\lambda x}.$$

Επιπλέον, αποδεικνύεται άμεσα ότι η μέση τιμή και η διακύμανση της X είναι αντίστοιχα ίσες με

$$\mu = E(X) = \frac{1}{\lambda} \text{ και } \sigma^2 = \mathrm{V}(X) = \frac{1}{\lambda^2}.$$



Η εκθετική κατανομή, όπως και η γεωμετρική κατανομή, έχει την ιδιότητα της έλλειψης μνήμης:

$$P(X > t + s|X > t) = P(X > s) = e^{-\lambda s}, \qquad s, t \geq 0.$$

δηλαδή η πιθανότητα η ΤΜ X να υπερβεί την τιμή $t+s$, όπου $0 < t < t+s$, δοθέντος ότι έχει υπερβεί την τιμή t, είναι ανεξάρτητη του t και ίση προς την αδέσμευτη πιθανότητα να υπερβεί την τιμή s. Πράγματι,

$$P(X > t + s|X > t) = \frac{P(X > t + s)}{P(X > t)} = \frac{1 - P(X \leq t + s)}{1 - P(X \leq t)} = \frac{e^{\lambda(-s-t)}}{e^{-\lambda t}} = e^{-\lambda s} = 1 - (1 - P(X \leq s)) = P(X > s).$$

Η εκθετική κατανομή χρησιμοποιείται για την περιγραφή της συμπεριφοράς μεταβλητών όπως π.χ. η διάρκεια τηλεφωνικών συνδιαλέξεων, η διάρκεια λειτουργίας διάφορων μηχανισμών, οι αφίξεις πελατών, κ.ο.κ.

Παράδειγμα 5.2.1. *Ο χρόνος ζωής μια μπαταρίας ακολουθεί την εκθετική κατανομή με μέση τιμή 32 μήνες.*
α) Ποιά είναι η πιθανότητα η μπαταρία να μην χαλάσει στους επόμενους 24 μήνες;
β) Ποιά είναι η πιθανότητα η μπαταρία να χαλάσει μεταξύ του 10ου και του 20ου μήνα;

Λύση. Έστω X ο χρόνος ζωής της μπαταρίας. Η ΤΜ X ακολουθεί την εκθετική κατανομή. Γνωρίζουμε ότι

$$E(X) = \frac{1}{\lambda} \Leftrightarrow 32 = \frac{1}{\lambda} \Leftrightarrow \lambda = \frac{1}{32}.$$

α)

$$P(X > 24) = 1 - P(X \leq 24) = 1 - (1 - e^{-24\lambda}) = e^{-24/32} = 0.47.$$

β)

$$P(10 \leq X \leq 20) = F(20) - F(10) = (1 - e^{-20/32}) - (1 - e^{10/32}) = 0.19.$$ □

5.3 Η κατανομή Γάμμα

Υπενθυμίζεται ότι η συνάρτηση Γάμμα ορίζεται ως

$$\Gamma(a) = \int_0^{+\infty} x^{a-1}e^{-x}dx, \qquad \text{για } a > 0.$$

Ειδικά, για κάθε $n \in \mathbb{N}$, ισχύει ότι $\Gamma(n+1) = n!$.

Μια ΤΜ X, η οποία παίρνει τιμές στο $[0, +\infty)$, λέμε ότι ακολουθεί την κατανομή Γάμμα με παραμέτρους $a > 0$ και $\theta > 0$ και γράφουμε $X \sim \Gamma(a, \theta)$, αν η PDF της X δίδεται από τον τύπο

$$f(x) = \begin{cases} \dfrac{\theta^a}{\Gamma(a)}x^{a-1}e^{-\theta x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Η μέση τιμή και η διακύμανση είναι αντίστοιχα ίσες με a/θ και a/θ^2.

Η εκθετική κατανομή $E(\theta)$ ταυτίζεται με την $\Gamma(1, \theta)$ και η κατανομή $\Gamma(n, \theta)$ ονομάζεται κατανομή Erlang και συμβολίζεται με $E(n, \theta)$.

5.3.1 Διαδικασία Poisson

Αν θεωρήσουμε ένα γεγονός (π.χ. την άφιξη σε μια ουρά αναμονής) που πραγματοποιείται επανειλημμένα σε χρονικό διάστημα $(0, t)$, διαμερίσουμε το διάστημα σε n υποδιαστήματα διάρκειας $\Delta t = t/n$ και θεωρήσουμε ότι η πραγματοποίηση του γεγονότος σε κάθε υποδιάστημα έχει πιθανότητα $p = \lambda\Delta t$ (ανάλογη του χρόνου), είναι ανεξάρτητη από τα υπόλοιπα υποδιαστήματα και είναι πρακτικά απίθανο να συμβεί περισσότερες από μία φορές επειδή το Δt είναι αρκετά μικρό, τότε η μεταβλητή X: πλήθος πραγματοποιήσεων στο διάστημα $(0, t)$ ακολουθεί τη διωνυμική κατανομή. Αν $n \to \infty$, οπότε $\Delta t \to 0$ και $p \to 0$, τότε η κατανομή της X έχει όριο την κατανομή Poisson με παράμετρο $np = \lambda t$.

Αν $N(t)$ είναι η ΤΜ που αντιστοιχεί στο πλήθος αφίξεων σε ένα σύστημα μέχρι τη χρονική στιγμή $t \geq 0$, τότε η οικογένεια $(N(t))_{t\geq 0}$ είναι μια διαδικασία Poisson με παράμετρο λ, αν $N(0) = 0$ και

$$P(N(t_0 + t) - N(t_0) = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!},$$

αν δηλαδή το πλήθος $N(t_0 + t) - N(t_0)$ των αφίξεων στο χρονικό διάστημα $[t_0, t_0 + t]$ ακολουθεί την κατανομή Poisson.

Η ποσότητα λt εκφράζει τον μέσο αριθμό αφίξεων σε χρονική διάρκεια t. Επομένως, ο μέσος χρόνος μεταξύ δύο διαδοχικών αφίξεων είναι $1/\lambda$, (αφού σε αυτό το χρόνο έχουμε κατά μέσο όρο μία άφιξη).

Αν η τυχαία μεταβλητή $X_n \in [0, +\infty)$ αντιστοιχεί στο χρόνο μέχρι τη n-οστή άφιξη, τότε είναι

$$P(X_n \leq t) = P(N(t) - N(0) \geq n) = \sum_{k\geq n} \frac{e^{-\lambda t}(\lambda t)^k}{k!} = 1 - e^{-\lambda t}\sum_{k=0}^{n-1}\frac{(\lambda t)^k}{k!}$$

Αν $f(t)$ είναι η ΣΠΠ της X_n, τότε $P(X_n \leq t) = \int_0^t f(x)dx$, οπότε παραγωγίζοντας κατά μέλη προκύπτει

$$f(t) = \lambda e^{-\lambda t}\sum_{k=0}^{n-1}\frac{(\lambda t)^k}{k!} - \lambda e^{-\lambda t}\sum_{k=1}^{n-1}\frac{k(\lambda t)^{k-1}}{k!} = \lambda e^{-\lambda t}\left(\sum_{k=0}^{n-1}\frac{(\lambda t)^k}{k!} - \sum_{k=0}^{n-2}\frac{(\lambda t)^k}{k!}\right) = \lambda e^{-\lambda t}\frac{(\lambda t)^{n-1}}{(n-1)!}$$

άρα $X_n \sim E(n, \lambda)$, οπότε και $X_1 \sim E(\lambda)$.

5.4 Η κατανομή Βήτα

Υπενθυμίζεται ότι η συνάρτηση Βήτα ορίζεται ως

$$B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx, \qquad a,b>0$$

και ικανοποιεί τη σχέση $B(a,b) = \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

Μια ΤΜ X, η οποία παίρνει τιμές στο $[0,1]$, λέμε ότι η X ακολουθεί την κατανομή Βήτα με παραμέτρους $a,b>0$ και γράφουμε $X \sim B(a,b)$, αν η συνάρτηση πυκνότητας πιθανότητας της X δίδεται από τον τύπο

$$f(x) = \begin{cases} \dfrac{x^{a-1}(1-x)^{b-1}}{B(a,b)}, & x \in [0,1] \\ 0, & x \notin [0,1] \end{cases}$$

Η μέση τιμή και η διακύμανση είναι αντίστοιχα ίσες με

$$E(X) = \frac{a}{a+b}, \qquad \mathrm{V}(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

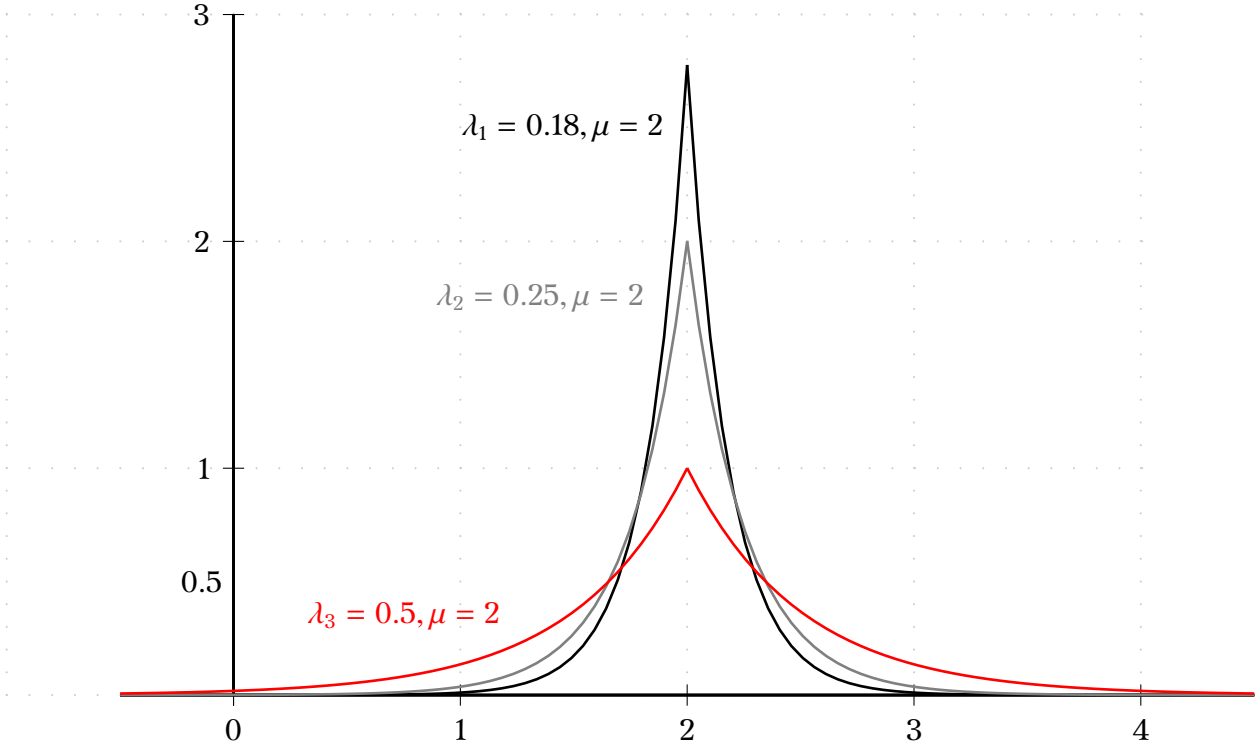
5.5 Κατανομή Laplace ή διπλή εκθετική

Μια ΤΜ X η οποία λαμβάνει τιμές στο $\mathbb{R}$ λέμε ότι ακολουθεί την *κατανομή Laplace ή διπλή εκθετική με παραμέτρους k και λ*, όπου $\lambda > 0$, αν η συνάρτηση πυκνότητας πιθανότητας της X είναι η

$$f(x) = P(X=x) = \frac{1}{2\lambda} e^{-\frac{1}{\lambda}|x-k|}/\mathbb{R}.$$

Η γραφική παράσταση της $f(x)$ είναι συμμετρική ως προς την ευθεία $x=k$. Αποδεικνύεται ότι η μέση τιμή και η διακύμανση της X ισούνται με

$$\mu = E(X) = k \text{ και } \sigma^2 = \mathrm{V}(X) = 2\lambda^2.$$



5.6 Κανονική κατανομή

Μια ΤΜ X η οποία λαμβάνει τιμές στο $\mathbb{R}$ λέμε ότι ακολουθεί την *κανονική κατανομή με παραμέτρους* μ *και* σ^2 (συμβολισμός $X \sim N(\mu, \sigma^2)$ αν η PDF αυτής δίδεται από τον τύπο

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \qquad \text{για κάθε } x \in \mathbb{R},$$

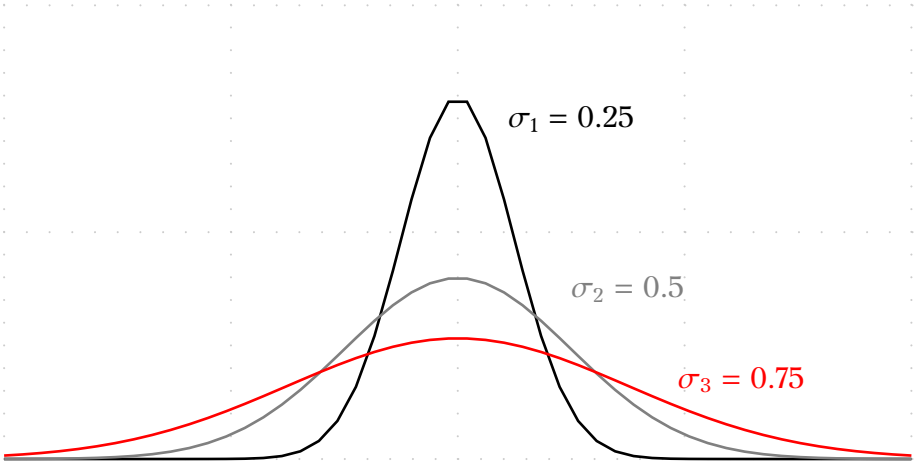
οπότε η CDF δίδεται από τον τύπο

$$F(x) = P(X \le x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt, \qquad \text{για κάθε } x \in \mathbb{R}.$$

Επιπλέον, αποδεικνύεται άμεσα ότι η μέση τιμή και η διακύμανση της X είναι αντίστοιχα ίσες με

$$E(X) = \mu \text{ και } \mathrm{V}(X) = \sigma^2$$

Η γραφική παράσταση της $f(x)$ έχει κωδωνοειδή μορφή συμμετρική ως προς την κατακόρυφη ευθεία $x = \mu$.



Ως γνωστό, το ολοκλήρωμα της $F(x)$ δεν μπορεί να υπολογισθεί χρησιμοποιώντας στοιχειώδεις συναρτήσεις. Οι τιμές της $F(x)$ υπολογίζονται προσεγγιστικά με αναγωγή στις τιμές της λεγόμενης τυπικής κανονικής κατανομής. Συγκεκριμένα, αν η τ.μ X ακολουθεί την κανονική κατανομή $N(\mu, \sigma^2)$ τότε η ΤΜ

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

ακολουθεί την κανονική κατανομή $N(0, 1)$ (δηλαδή έχει μέσο όρο 0 και διακύμανση 1) η οποία ονομάζεται *τυπική ή τυποποιημένη κανονική κατανομή (standard normal distribution).*

Στην περίπτωση αυτή, η CDF της Z συμβολίζεται με $\Phi(z)$, ενώ η αντίστοιχη PDF με $\phi(z)$, δηλαδή

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt \qquad \text{και} \qquad \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad z \in \mathbb{R}.$$

Για την $\Phi(z)$ ισχύει η παρακάτω ιδιότητα

$$\Phi(z) + \Phi(-z) = 1.$$

Επίσης, εύκολα προκύπτουν οι παρακάτω σχέσεις

Αν η $X \sim N(\mu, \sigma^2)$ τότε

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

$$P(a \le X) = 1 - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

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

Παράδειγμα 5.6.1. *Έχει διαπιστωθεί ότι ο δείκτης ευφυΐας (IQ) των ανθρώπων ακολουθεί περίπου την κανονική κατανομή $N(105, 20^2)$.*
α) Να βρεθεί το ποσοστό των ανθρώπων με δείκτη ευφυΐας

i) *τουλάχιστον* 145.

ii) *το πολύ* 90.

iii) *μεταξύ* 80 *και* 125.

β) Τι δείκτη ευφυΐας πρέπει να έχει κάποιος για να ανήκει στο ευφυέστερο 1% των ανθρώπων;

Λύση. α) Έστω $X \sim N(\mu, \sigma^2)$, όπου $\mu = 105$ και $\sigma = 20$. Θέτοντας $Z = (X-\mu)/\sigma$, έχουμε ότι

$$\begin{aligned} P(X > 145) &= P((X-105)/20 > 2) = P(Z > 2) = 1 - \Phi(2) \approx 0.02275, \\ P(X \le 90) &= P((X-105)/20 \le -15/20) = P(Z \le -0.75) = 1 - \Phi(0.75) \approx 0.22662, \\ P(80 \le X \le 125) &= P(-25/20 \le (X-105)/20 \le 20/20) = P(-1.25 \le Z \le 1) \\ &= \Phi(1) - (1 - \Phi(1.25)) \approx 0.7357. \end{aligned}$$

β) Ψάχνουμε μια τιμή x τέτοια ώστε $P(X \ge x) = 0.01$. Είναι

$$\begin{aligned} P(X \ge x) = 1/100 &\Leftrightarrow P(\frac{X-\mu}{\sigma} \ge \frac{x-\mu}{\sigma}) = 0.01 \Leftrightarrow 1 - \Phi(\frac{x-\mu}{\sigma}) = 0.01 \Leftrightarrow \Phi(\frac{x-\mu}{\sigma}) = 0.99 \\ &\Leftrightarrow \frac{x-\mu}{\sigma} = 2.32635 \Leftrightarrow x = 151.527. \end{aligned}$$

□

Παράδειγμα 5.6.2. *Αν μια ΤΜ X ακολουθεί την κανονική κατανομή $N(\mu, \sigma^2)$, να βρεθούν οι πιθανότητες*

i) $P(|X-\mu| \le \sigma)$

ii) $P(|X-\mu| \le 2\sigma)$

Λύση. Θέτουμε $Z = (X-\mu)/\sigma$, οπότε $Z \sim N(0,1)$.

$$P(|X-\mu| \le \sigma) = P(|Z| \le 1) = \Phi(1) - \Phi(-1) = 2\Phi(1) - 1 \approx 0.6827$$

$$P(|X-\mu| \le 2\sigma) = P(|Z| \le 2) = \Phi(2) - \Phi(-2) = 2\Phi(2) - 1 \approx 0.9545$$

i) δηλαδή το 68.27% των τιμών της X απέχουν λιγότερο από σ από την μέση τιμή της μ.

ii) δηλαδή το 95.45% των τιμών της X απέχουν λιγότερο από 2σ από την μέση τιμή της μ. □

Παράδειγμα 5.6.3. *Ένας μαθητής έδωσε γραπτές και προφορικές εξετάσεις σε κάποιο μάθημα. Στις γραπτές εξετάσεις βαθμολογήθηκε με* 71 *στα* 100*, ενώ ο μέσος όρος ήταν* 60 *με διακύμανση* 30*. Στις προφορικές εξετάσεις βαθμολογήθηκε με* 58 *ενώ ο μέσος όρος ήταν* 50 *με διακύμανση* 15*. Έστω ότι και στις δύο εξετάσεις οι βαθμολογίες ακολουθούν την κανονική κατανομή, να βρεθεί σε ποια από τις 2 εξετάσεις πήγε καλύτερα;*

Λύση. Για κάθε εξέταση θα βρούμε το ποσοστό των συνεξεταζόμενων που έλαβαν μικρότερο ή ίσο βαθμό από αυτόν τον μαθητή. Με άλλα λόγια, αν $X \sim N(60, 30)$ είναι ο βαθμός της πρώτης εξέτασης και $Y \sim N(50, 15)$ ο βαθμός της δεύτερης, θα συγκρίνουμε τις πιθανότητες $P(X \leq 71)$ και $P(Y \leq 58)$ και η μεγαλύτερη θα αντιστοιχεί στην καλύτερη εξέταση. Έχουμε λοιπόν ότι

$$P(X \leq 71) = \Phi\left(\frac{71-60}{30}\right) = \Phi\left(\frac{11}{30}\right) \qquad \text{και} \qquad P(Y \leq 58) = \Phi\left(\frac{58-50}{15}\right) = \Phi\left(\frac{16}{30}\right),$$

άρα τα πήγε καλύτερα στην δεύτερη εξέταση. □

Πρόταση 5.1. *Αν οι ΤΜ $X_1, X_2, \ldots, X_n$ είναι ανεξάρτητες και ακολουθούν τις κανονικές κατανομές $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$, …, $N(\mu_n, \sigma_n^2)$ αντίστοιχα, τότε*

$$X_1 + X_2 + \cdots + X_n \sim N(\mu_1 + \mu_2 + \cdots + \mu_n, \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2).$$

Παράδειγμα 5.6.4. *Το βάρος (σε κιλά) των ατόμων ενός πληθυσμού ακολουθεί την κανονική κατανομή $N(75, 8^2)$. Έστω ότι το κεντρικό ασανσέρ ενός κτηρίου μπορεί να μεταφέρει με ασφάλεια βάρος μέχρι* 900 *κιλά. Να βρεθούν:*

i) Η πιθανότητα 15 *άτομα να χρησιμοποιήσουν συγχρόνως το ασανσέρ με ασφάλεια.*

ii) Ο μέγιστος αριθμός ατόμων που με πιθανότητα τουλάχιστον 99% *μπορούν να χρησιμοποιήσουν συγχρόνως το ασανσέρ.*

Λύση. i) Έστω X_i το βάρος του ο i-οστού ατόμου, για κάθε $i \in [15]$ και έστω $X = X_1 + X_2 + \cdots + X_{15}$. Οι ΤΜ X_i ακολουθούν την κανονική κατανομή $N(75, 8^2)$, επομένως η X ακολουθεί την κανονική κατανομή $N(15 \cdot 75, 15 \cdot 8^2) = N(1125, 15 \cdot 8^2)$.

$P(X \leq 900) = P(\frac{X-1125}{\sqrt{15}\cdot 8} \leq \frac{900-1125}{\sqrt{15}\cdot 8}) = P(Z \leq -\frac{15\sqrt{15}}{8}) = P(Z \leq -7.26184) \approx 0.$

ii) Έστω $S_n := \sum_{i=1}^n X_i$, οπότε $S_n \sim N(n\mu, n\sigma^2)$ και $Z_n := \dfrac{S_n - n\mu}{\sigma\sqrt{n}} \sim N(0,1)$. Θέτοντας $z_n = \dfrac{900 - n\mu}{\sigma\sqrt{n}}$, είναι

$$P(S_n \leq 900) \geq 0.99 \Leftrightarrow P(Z_n \leq z_n) \geq 0.99 \Leftrightarrow \Phi(z_n) \geq 0.99 \Leftrightarrow z_n \geq \Phi^{-1}(0.99)$$

Η τελευταία ισοδυναμία ισχύει διότι η Φ είναι αύξουσα.

Θέτοντας $x = 900$ και $z = \Phi^{-1}(0.99) \approx 2.32635$, είναι

$$z_n \geq z \Leftrightarrow \frac{x - n\mu}{\sigma\sqrt{n}} \geq z \Leftrightarrow \mu n + z\sigma\sqrt{n} - x \leq 0 \Leftrightarrow \sqrt{n} \leq \frac{-z\sigma + \sqrt{z^2\sigma^2 + 4x\mu}}{2\mu} \approx 3.342251 \Leftrightarrow n \leq 11.17$$

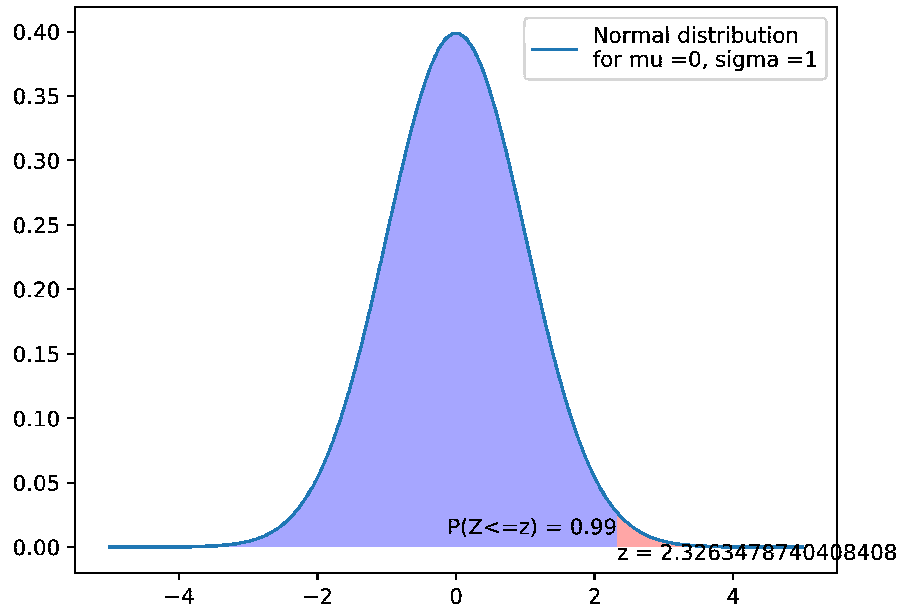
Επομένως, ο μέγιστος αριθμός ατόμων είναι ίσος με 11. □

Ο παρακάτω κώδικας Python δείχνει τη διαδικασία εύρεσης της τιμής z ώστε $P(Z \leq z) = p = 0.99$.

```
import numpy as np
from scipy.stats import norm
import matplotlib.pyplot as plt

mu, sigma = 0, 1    #standard normal distribution
rv = norm(loc=mu, scale=sigma) #a normal random variable rv
x = np.linspace(-5,5,100) #range of values for rv
p= 0.99 #desired accuracy
z = rv.ppf(p) #inverse of cdf
print("P(|Z| <= %s) = %s"%(z, rv.cdf(z)))
print("p-percent point for p = %s is z = %s"%(p,rv.ppf(p)))
#plt.figure(figsize=(12,10))
label1 = "Normal distribution\n"+ "for mu = %s, sigma = %s"%(mu,sigma)
plt.plot(x,rv.pdf(x), label=label1)

plt.fill_between(x=np.arange(z,5,0.01),
                 y1= rv.pdf(np.arange(z,5,0.01)),
                 facecolor='red', alpha=0.35)

plt.fill_between(x=np.arange(-5,z,0.01),
                 y1= rv.pdf(np.arange(-5,z,0.01)),
                 facecolor='blue', alpha=0.35)

plt.text(x=z, y=-0.01, s = "z = %s"%(z))
plt.text(x=z, y=0.01, s = "P(Z<=z) = %s"%(p), horizontalalignment='right')
plt.legend()
plt.show()
```

Output

```
P(|Z| <= 2.3263478740408408) = 0.99
p-percent point for p = 0.99 is z = 2.3263478740408408
```



5.7 Το κεντρικό οριακό θεώρημα

Ορισμός. *Έστω $(X_n)_{n\in\mathbb{N}^*}$ ακολουθία ΤΜ με κοινό σύνολο τιμών S και έστω $F_n(x) = P(X_n \le x)$ η συνάρτηση κατανομής της X_n, $n \in \mathbb{N}$. Λέμε ότι η ακολουθία (X_n) συγκλίνει κατά κατανομή στην ΤΜ X που ακολουθεί μια κατανομή D και έχει συνάρτηση κατανομής $F(x) = P(X \le x)$ και σύνολο τιμών S και γράφουμε $X_n \to D$, αν και μόνο αν*

$$\lim_{n\to\infty} F_n(x) = F(x),$$

για κάθε $x \in S$ όπου F συνεχής.

Πρόταση 5.2 (Κεντρικό οριακό θεώρημα). *Αν οι (διακριτές ή συνεχείς) ΤΜ $X_1, X_2, \ldots, X_n$ είναι ανεξάρτητες και ακολουθούν την ίδια κατανομή με μέση τιμή μ και διακύμανση σ^2, τότε*

i) $S_n := X_1 + X_2 + \cdots + X_n \to N(n\mu, n\sigma^2)$,

ii) $\overline{X}_n := \dfrac{X_1 + X_2 + \cdots + X_n}{n} \to N(\mu, \sigma^2/n)$,

iii) $\overline{S}_n := \dfrac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \dfrac{S_n - n\mu}{\sigma\sqrt{n}} = \dfrac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n}(X_i - \mu) \to N(0,1)$,

iv) $\lim\limits_{n\to\infty} P(\overline{S}_n \le x) = \Phi(x)$, *για κάθε* $x \in \mathbb{R}$.

Το θεώρημα συνήθως εφαρμόζεται για τον προσεγγιστικό υπολογισμό πιθανοτήτων του εμπειρικού μέσου $\overline{X}_n$ που συνήθως προέρχεται από τις παρατηρήσεις σε ένα δείγμα μεγέθους n. Η ακρίβεια της προσέγγισης εξαρτάται κυρίως από το n και πολύ λιγότερο από την κατανομή των ΤΜ X_n, για μικρές τιμές του n. Στις εφαρμογές, συνήθως λαμβάνεται $n \ge 30$.

Προσέγγιση της διωνυμικής κατανομής από την κανονική: Αν $X \sim \text{Binom}(n, p)$, οπότε $\mu = np$ και $\sigma^2 = np(1-p)$, και το n είναι αρκετά μεγάλο, τότε η X ακολουθεί κατά προσέγγιση την $\mathcal{N}(\mu, \sigma^2)$, δηλαδή για κάθε ζεύγος ακεραίων a, b, με $0 \le a \le b \le n$, ισχύει

$$P(a \le X \le b) = P(a - 1/2 \le X \le b + 1/2) \simeq \Phi\left(\frac{b + 1/2 - \mu}{\sigma}\right) - \Phi\left(\frac{a - 1/2 - \mu}{\sigma}\right).$$

Η προσθήκη του 1/2 στα δύο άκρα της ανισότητας ονομάζεται διόρθωση συνέχειας και γίνεται για τη βελτίωση της προσέγγισης, λόγω της μετάβασης από διακριτή σε συνεχή μεταβλητή.

Προσέγγιση της κατανομής Poisson από την κανονική: Ομοίως, αν $X \sim P(\lambda)$, οπότε $\mu = \sigma^2 = \lambda$, και το λ είναι αρκετά μεγάλο ($\lambda > 20$), τότε η X ακολουθεί κατά προσέγγιση την $\mathcal{N}(\mu, \sigma^2)$ και εφαρμόζεται ο προηγούμενος προσεγγιστικός τύπος με διόρθωση συνέχειας.

Παράδειγμα 5.7.1. *Ρίχνουμε ένα αμερόληπτο κέρμα* 1000 *φορές και έστω X η ΤΜ που ισούται με τον αριθμό των εμφανίσεων της όψης ΓΡΑΜΜΑΤΑ.*

i) *Να βρεθούν οι πιθανότητες $P(490 \le X \le 510)$ και $P(485 \le X \le 515)$.*

ii) *Να βρεθεί ο ελάχιστος φυσικός N ώστε $P(X \le N) \ge 0.99$.*

Λύση. Έστω X_i, $i = 1, 2, \ldots, 1000$ οι ΤΜ με

$$X_i = \begin{cases} 1, & \text{αν στην } i\text{-οστή ρίψη έρθει η όψη ΓΡΑΜΜΑΤΑ} \\ 0, & \text{αλλιώς.} \end{cases}.$$

και

$$X = X_1 + X_2 + \cdots X_{1000}.$$

Οι ΤΜ X_i ακολουθούν την κατανομή Bernoulli με παράμετρο $p = \frac{1}{2}$, άρα έχουν μέση τιμή $\mu = \frac{1}{2}$ και διακύμανση $\sigma^2 = p(1-p) = \frac{1}{2}(1-\frac{1}{2}) = \frac{1}{4}$.

Επομένως, η X ακολουθεί την διωνυμική κατανομή με μέση τιμή $1000p = 500$ και διακύμανση $1000\frac{1}{4} = 250$.

Μπορούμε όμως, βάσει του κεντρικού οριακού θεωρήματος, να θεωρήσουμε ότι η X ακολουθεί ασυμπτωτικά την κανονική κατανομή $N(1000 \cdot \frac{1}{2}, 1000 \cdot \frac{1}{4}) = N(500, 250)$.

i) $P(490 \le X \le 510) = P(\frac{490-500}{\sqrt{250}} \le \frac{X-500}{\sqrt{250}} \le \frac{510-500}{\sqrt{250}}) = P(-\sqrt{2/5} \le Z \le \sqrt{2/5}) = P(-0.6324 \le Z \le 0.6324) = \Phi(0.6324) - \Phi(-0.6324) = 2\Phi(0.6324) - 1 = 2 \cdot 0.7357 - 1 = 0.4714.$

$P(485 \le X \le 515) = P(\frac{485-500}{\sqrt{250}} \le \frac{X-500}{\sqrt{250}} \le \frac{515-500}{\sqrt{250}}) = P(-\frac{3}{\sqrt{10}} \le Z \le \frac{3}{\sqrt{10}}) = P(-0.948 \le Z \le 0.948) \approx \Phi(0.95) - \Phi(-0.95) = 2\Phi(0.95) - 1 = 2 \cdot 0.8289 - 1 = 0.6578.$

Εισάγοντας τη διόρθωση συνέχειας, επειδή $\frac{1/2}{\sqrt{250}} = 0.0316$ έχουμε ότι $P(490 \le X \le 510) = P(490 - 1/2 \le X \le 510 + 1/2) = \Phi(0.6324 + 0.0316) - \Phi(-0.6324 - 0.0316) = 2\Phi(0.664) - 1 = 2 \cdot 0.746655 - 1 = 0.49331$
και $P(485 \le X \le 515) = 2\Phi(0.95 + 0.0316) - 1 = 0.6737.$

Για να δούμε πόσο καλή είναι προσέγγιση της X από την κανονική κατανομή μπορούμε να υπολογίσουμε τις αντίστοιχες πιθανότητες με τους τύπους της διωνυμικής κατανομής.

$$P(490 \le X \le 510) = \sum_{i=490}^{510} P(X = i) = \sum_{i=490}^{510} \binom{1000}{i} (\tfrac{1}{2})^i (1-\tfrac{1}{2})^{1000-i} = \frac{1}{2^{1000}} \sum_{i=490}^{510} \binom{1000}{i} = 0.49334.$$

$$P(485 \le X \le 515) = \sum_{i=485}^{515} P(X = i) = \sum_{i=485}^{515} \binom{1000}{i} (\tfrac{1}{2})^i (1-\tfrac{1}{2})^{1000-i} = \frac{1}{2^{1000}} \sum_{i=485}^{515} \binom{1000}{i} = 0.673063.$$

Δηλαδή, η προσέγγιση της X από την κανονική κατανομή είναι πολύ καλή.

ii) $P(X \le N) = P(\frac{X-500}{\sqrt{250}} \le \frac{N-500}{\sqrt{250}}) = P(Z \le \frac{N-500}{\sqrt{250}}) = \Phi(\frac{N-500}{\sqrt{250}}) \ge 0.99.$

Άρα, $\frac{N-500}{\sqrt{250}} \ge 2.33 \Leftrightarrow N \ge 500 + 2.33 \cdot \sqrt{250} = 500 + 2.33 \cdot 15.8114 = 536.841$. Άρα, $N \ge 537$. Επομένως, $N = 537$.

Μπορούμε να υπολογίσουμε με τους τύπους της διωνυμικής κατανομής τις πιθανότητες $P(X \le 536)$, $P(X \le 537)$ για να δούμε αν η τιμή 537 είναι πράγματι η ελάχιστη τιμή N.

$$P(X \le 536) = \sum_{i=0}^{536} P(X = i) = \sum_{i=0}^{536} \binom{1000}{i} (\tfrac{1}{2})^i (1-\tfrac{1}{2})^{1000-i} = \frac{1}{2^{1000}} \sum_{i=0}^{536} \binom{1000}{536} = 0.989536.$$

$P(X \le 537) = P(X \le 336) + P(X = 537) = P(X \le 336) + \binom{1000}{537} (\frac{1}{2})^{537} (1-\frac{1}{2})^{1000-537} = 0.989536 + 0.00163244 = 0.991168.$

Δηλαδή, η τιμή 537 που βρήκαμε χρησιμοποιώντας την προσέγγιση της X από την κανονική κατανομή είναι πράγματι η ελάχιστη τιμή N. □

Παράδειγμα 5.7.2. *Έχει υπολογισθεί ότι το ποσό (σε ευρώ) που ξοδεύουν οι πελάτες ενός καταστήματος μια συγκεκριμένη ημέρα ακολουθεί την κανονική κατανομή* $N(40, 10^2)$.

i) *Να βρεθεί το ποσοστό των ανθρώπων που ξοδεύουν κάτω από* 20 *ευρώ.*

ii) *Αν το κατάστημα δεχθεί* 300 *πελάτες, να βρεθεί η πιθανότητα τα συνολικά έσοδα να είναι τουλάχιστον* 12000 *ευρώ.*

Λύση. Έστω X το πόσο που ξοδεύει ένας πελάτης. Η ΤΜ X ακολουθεί την κανονική κατανομή $N(40, 10^2)$.

i) $P(X \leq 20) = P(\frac{X-40}{10} \leq \frac{20-40}{10}) = P(Z \leq -2) = \Phi(-2) = 0.02275$. Άρα, μόλις το 2.3% περίπου των πελατών του καταστήματος ξοδεύουν το πολύ 20 ευρώ.

ii) Έστω X_i το ποσό που ξοδεύει ο i-οστός πελάτης, για κάθε $i = 1, 2, \ldots, 300$. Τα έσοδα του καταστήματος ισούνται με $Y = X_1 + X_2 + \cdots + X_{300}$. Βάσει του κεντρικού οριακού θεωρήματος η ΤΜ Y ακολουθεί ασυμπτωτικά την κανονική κατανομή $N(300 \cdot 40, 300 \cdot 10^2) = N(12000, 3 \cdot 100^2)$. Επομένως, $P(Y \geq 12000) = P(\frac{Y-12000}{100\sqrt{3}} \geq \frac{12000-12000}{100\sqrt{3}}) = P(Z \geq 0) = 1 - \Phi(0) = 0.5$. □

5.8 Ανισότητες

Σε κάποια προβλήματα ενδέχεται να μην έχουμε πολλές πληροφορίες για μια ΤΜ μεταβλητή εκτός από τη μέση τιμή της, ή και τη διακύμανσή της. Σε αυτές τις περιπτώσεις, μπορούμε να βγάλουμε κάποια συμπεράσματα για την ΤΜ X, χρησιμοποιώντας ανισότητες όπως του Markov ή του Chebyshev, οι οποίες παρουσιάζονται στις επόμενες ενότητες.

5.8.1 Η ανισότητα του Markov

Πρόταση 5.3 (Ανισότητα του Markov). *Αν μια (διακριτή ή συνεχής) ΤΜ X λαμβάνει μόνο μη αρνητικές τιμές και έχει πεπερασμένη μέση τιμή μ, τότε*

$$P(X \geq c) \leq \frac{\mu}{c}, \qquad \text{για κάθε πραγματική σταθερά } c > 0.$$

Απόδειξη. Θα αποδείξουμε την ανισότητα στην περίπτωση όπου η X είναι συνεχής. (Αν η X είναι διακριτή η απόδειξη είναι ανάλογη.) Έστω $f(x)$ η PDF της X. Προφανώς, είναι $x < 0 \Rightarrow f(x) = 0$, οπότε

$$\begin{aligned}\mu = E(X) &= \int_0^{+\infty} xf(x)dx = \int_0^{c} xf(x)dx + \int_c^{+\infty} xf(x)dx \\ &\geq \int_c^{+\infty} xf(x)dx \geq \int_c^{+\infty} cf(x)dx = c\int_c^{+\infty} f(x)dx = cP(X \geq c)\end{aligned}$$

□

Παρατήρηση. Διαισθητικά, αν μια ΤΜ έχει μικρή μέση τιμή, τότε η πιθανότητα να λάβει μεγάλες τιμές είναι μικρή.

Η σημασία της ανισότητας του Markov είναι ότι αν και το μόνο που γνωρίζουμε για μια ΤΜ είναι η μέση τιμή της, εν τούτοις μπορούμε να βγάλουμε κάποιο χρήσιμο συμπέρασμα.

Παράδειγμα 5.8.1. *Μια μη αρνητική ΤΜ X έχει μέση τιμή $\mu = 5$. Να βρεθεί ένα φράγμα για την πιθανότητα $P(X \geq 20)$.*

Λύση. Από την ανισότητα του Markov έχουμε ότι

$$P(X \geq 20) \leq \frac{\mu}{20} = \frac{5}{20} = \frac{1}{4} = 0.25$$

δηλαδή το πολύ 25% των τιμών που λαμβάνει η X είναι μεγαλύτερες ή ίσες του 20, ενώ το 75% των τιμών της ανήκει στο διάστημα 0 έως 20. □

Παράδειγμα 5.8.2. *Ένας βιολόγος ισχυρίζεται ότι το μέσο βάρος του αφρικανικού χελιδονιού είναι 100 γραμμάρια, ενώ το 60% των αφρικανικών χελιδονιών έχει βάρος μεγαλύτερο ή ίσο από 200 γραμμάρια. Είναι δυνατόν να ευσταθεί ο ισχυρισμός αυτός;*

Λύση. Έστω X η ΤΜ βάρος (σε γραμμάρια) του αφρικανικού χελιδονιού. Αν η ΤΜ X έχει μέση τιμή $\mu = 100$, τότε από την ανισότητα του Markov έχουμε ότι

$$P(X \geq 200) \leq \frac{\mu}{200} = \frac{100}{200} = 0.5,$$

το οποίο δεν συμφωνεί με τον ισχυρισμό του βιολόγου ότι η πιθανότητα αυτή είναι 60%.

Μάλιστα, βάσει της παραπάνω ανισότητας, θα πρέπει να είναι $\mu \geq 120$. □

Παράδειγμα 5.8.3. *Έστω ότι ο μέσος βαθμός σε ένα μάθημα είναι 7 (με άριστα το 10). Να βρεθεί ένα κάτω φράγμα για το ποσοστό των φοιτητών που έχουν βαθμολογηθεί με βαθμό μικρότερο του 8.*

Λύση. Αν η ΤΜ X αντιστοιχεί στον βαθμό ενός τυχαία επιλεγμένου φοιτητή, τότε

$$P(X < 8) = 1 - P(X \geq 8) \geq 1 - \frac{\mu}{8} = 1 - \frac{7}{8} = \frac{1}{8} = 0.125,$$

δηλαδή τουλάχιστον το 1/8 των φοιτητών πήραν βαθμό κάτω του 8. □

Παράδειγμα 5.8.4. *Έστω $\sigma = \sigma(1)\sigma(2)\cdots\sigma(n)$ μια μετάθεση του $[n]$. Αν $\sigma(i) = i$ τότε λέμε ότι n σ έχει σταθερό σημείο το i. Να δειχθεί ότι με πιθανότητα τουλάχιστον $1 - \frac{1}{k}$ n σ έχει λιγότερα από k σταθερά σημεία.*

Λύση. Έστω X ο αριθμός των σταθερών σημείων της μετάθεσης σ.

Έστω $X_i = \begin{cases} 1, & \sigma(i) = i \\ 0, & \sigma(i) \neq i \end{cases}$, για κάθε $i \in [n]$.

Προφανώς, $X = X_1 + X_2 + \cdots + X_n$ και $P(X_i = 1) = \frac{1}{n}$, άρα $E(X_i) = \frac{1}{n}$. Επομένως,

$$E(X) = E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots E(X_n) = \frac{n}{n} = 1.$$

Επομένως, $P(X < k) = 1 - P(X \geq k) \geq 1 - \frac{E(X)}{k} = 1 - \frac{1}{k}$. □

Παράδειγμα 5.8.5. *Ο επόμενος αλγόριθμος εντοπίζει το μέγιστο στοιχείο μιας λίστας $a = (a[0], \ldots, a[n-1])$, με n διαφορετικά στοιχεία. Να υπολογισθούν:*

i) Το αναμενόμενο πλήθος αναθέσεων (πόσες φορές αλλάξει η τιμή της μεταβλητής m).

ii) Ένα άνω φράγμα για την πιθανότητα να γίνουν περισσότερες από 15 αναθέσεις.

```
def maxInList(a):
    m=a[0]
    for i in range(len(a)): if m < a[i]: m = a[i]
    return m
```

Λύση. *i)* Έστω X_i η δείκτρια ΤΜ του ενδεχομένου να γίνει ανάθεση στη θέση $i-1 \in [n]$. Επειδή η τιμή $a[i-1]$ είναι η μέγιστη των $a[0], a[1], \ldots, a[i-1]$ με πιθανότητα $1/i$, έπεται ότι $X_i \sim \text{Bernoulli}(1/i)$, οπότε το πλήθος αναθέσεων είναι $X = X_1 + X_2 + \cdots + X_n$, με

$$E(X) = E(X_1 + X_2 + \cdots + X_n) = \sum_{i=1}^{n} \frac{1}{i},$$

όπου το τελευταίο άθροισμα είναι ο n-οστός αρμονικός αριθμός H_n.

ii) Ως γνωστό, ισχύει η σχέση $\ln(1+n) \leq H_n \leq 1 + \ln n$, οπότε, εφαρμόζοντας την ανισότητα Markov για $c = 16$, έχουμε ότι

$$P(X \geq 16) \leq \frac{E(X)}{16} \leq \frac{1 + \ln n}{16}.$$

Για παράδειγμα, για $n = 100$, βρίσκουμε ότι $P(X \geq 16) \leq 0.35$, ενώ για $n = 1000$, βρίσκουμε ότι $P(X \geq 16) \leq 0.494$. □

Πρόταση 5.4. *Έστω ΤΜ X και έστω $c \in \mathbb{R}$.*

i) Αν $E(X) \leq c$, τότε $P(X \leq c) > 0$.

ii) Επιπλέον, αν $S_x \subseteq \mathbb{N}$ και $E(X) < 1$, τότε $P(X = 0) > 0$.

Απόδειξη. *i)* Αν η X είναι διακριτή και $P(X \leq c) = 0$, τότε είναι

$$P(X \leq c) = 0 \Rightarrow P(X > c) = 1 \Rightarrow E(X) = \sum_{k \geq \lfloor c \rfloor + 1} kP(X = k) > \sum_{k \geq \lfloor c \rfloor + 1} cP(X = k) = c,$$

το οποίο είναι άτοπο.

Αν η X είναι συνεχής και $P(X \leq c) = F(c) = 0$, τότε υπάρχει $n \in \mathbb{N}^*$ ώστε $F(c + 1/n) < 1$, οπότε

$$\begin{aligned} E(X) &= \int_c^{+\infty} xf(x)dx = \int_c^{c+1/n} xf(x)dx + \int_{c+1/n}^{+\infty} xf(x)dx \geq c\int_c^{c+1/n} f(x)dx + (c + 1/n)\int_{c+1/n}^{+\infty} f(x)dx \\ &= cF(c + 1/n) + (c + 1/n)(1 - F(c + 1/n)) = c + 1/n(1 - F(c + 1/n)) > c, \end{aligned}$$

το οποίο είναι άτοπο.

ii) Εφαρμόζοντας, την ανισότητα Markov, έχουμε ότι

$$P(X = 0) = 1 - P(X \geq 1) \geq 1 - E(X) > 0. \qquad \square$$

Παράδειγμα 5.8.6. *Αν $v_1, v_2, \ldots, v_n$ είναι μοναδιαία διανύσματα του $\mathbb{R}^n$, να δειχθεί ότι υπάρχουν συντελεστές $x_1, x_2, \ldots, x_n \in \{-1, +1\}$, ώστε $|x_1v_1 + \cdots + x_nv_n| \leq \sqrt{n}$.*

Λύση. Θεωρούμε την ΤΜ $Y = |X_1v_1 + \cdots + X_nv_n|$, όπου οι ανεξάρτητες ΤΜ X_i παίρνουν τιμές ομοιόμορφα στο $\{-1, 1\}$. Τότε, είναι

$$E(X_iX_j) = \begin{cases} 0, & i \neq j, \\ 1, & i = j, \end{cases}$$

και

$$E(Y^2) = E\left(\sum_{i,j} X_iX_j\langle v_i, v_j\rangle\right) = \sum_{i,j}\langle v_i, v_j\rangle E(X_iX_j) = \sum_i |v_i|^2 = n.$$

Επομένως, εφαρμόζοντας την προηγούμενη πρόταση για $c = n$, προκύπτει ότι
$P(Y \leq \sqrt{n}) = P(Y^2 \leq n) > 0$. $\square$

5.8.2 Η ανισότητα Cauchy-Schwarz

Πρόταση 5.5 (Ανισότητα Cauchy-Schwarz). *Για οποιεσδήποτε ΤΜ X, Y ισχύει ότι*

$$(E(XY))^2 \leq E(X^2)E(Y^2)$$

Απόδειξη. Για $a \in \mathbb{R}$, η ΤΜ $(X + aY)^2$ παίρνει μη αρνητικές τιμές, άρα και η μέση τιμή της είναι μη αρνητική. Επομένως,

$$0 \leq E((X + aY)^2) = E(X^2 + 2aXY + a^2Y^2) = a^2E(Y^2) + 2aE(XY) + E(X^2).$$

Αφού η ανισότητα ισχύει για κάθε a, έπεται ότι η διακρίνουσα $(2E(XY))^2 - 4E(Y^2)E(X^2)$ είναι μικρότερη ή ίση του 0, οπότε προκύπτει το ζητούμενο. $\square$

5.8.3 Η ανισότητα του Chernoff

Πρόταση 5.6 (Ανισότητα του Chernoff). *Έστω ΤΜ X και έστω $c \in \mathbb{R}$. Τότε, για κάθε $t > 0$ ισχύει ότι*

$$P(X \geq c) \leq \frac{E(e^{tX})}{e^{tc}}.$$

Απόδειξη. Θεωρούμε την ΤΜ $Y = e^{tX} \geq 0$. Δεδομένου ότι $e^{tc} \geq 0$, εφαρμόζοντας την ανισότητα Markov, έχουμε ότι

$$P(X \geq c) = P(e^{tX} \geq e^{tc}) \leq \frac{E(e^{tX})}{e^{tc}}.$$

□

Η παραπάνω ανισότητα δίνει συχνά καλύτερο άνω φράγμα από την ανισότητα Markov, διότι εφαρμόζεται για οποιοδήποτε $t > 0$ και το φράγμα που δίνει μειώνεται εκθετικά συναρτήσει του t.

Παράδειγμα 5.8.7. *Έστω $X \sim E(\lambda)$, $\lambda > 0$ και $c > 0$. Να ευρεθεί ένα άνω φράγμα για την πιθανότητα $P(X \leq c)$ χρησιμοποιώντας την ανισότητα Markov και την ανισότητα Chernoff. Σε ποιες περιπτώσεις η δεύτερη δίνει μικρότερο άνω φράγμα;*

Λύση. Έστω $Y = e^{tX}$, $t > 0$.

$$E(e^{tX}) = \int_0^{+\infty} e^{tx}\lambda e^{-\lambda x}dx = \lambda \int_0^{+\infty} e^{(t-\lambda)x}dx = \lambda \left[\frac{e^{(t-\lambda)x}}{t-\lambda}\right]_0^{+\infty} dx = \begin{cases} \dfrac{\lambda}{\lambda - t}, & 0 < t < \lambda, \\ +\infty, & t \geq \lambda \end{cases}$$

επομένως, σύμφωνα με την ανισότητα Chernoff, είναι

$$P(X \geq c) \leq \frac{E(e^{tX})}{e^{tc}} = \frac{\lambda e^{-tc}}{\lambda - t}$$

Θεωρώντας την συνάρτηση $g(t) = \dfrac{\lambda e^{-tc}}{\lambda - t}/(0, \lambda)$, βρίσκουμε ότι παρουσιάζει ολικό ελάχιστο όταν $t = \lambda - 1/c$. Επομένως,

$$P(X \geq c) \leq \frac{\lambda e^{(1/c-\lambda)c}}{1/c} = c\lambda e^{1-\lambda c}$$

Από την άλλη, εφαρμόζοντας την ανισότητα Markov, έχουμε ότι

$$P(X \geq c) \leq \frac{E(X)}{c} = \frac{1}{\lambda c}$$

Για να εξετάσουμε πότε είναι $c\lambda e^{1-\lambda c} \leq \dfrac{1}{\lambda c}$, θεωρούμε τη συνάρτηση $h(x) = 1/x - xe^{1-x}/(0, +\infty)$ και βρίσκουμε ότι $h(x) \geq 0$ ανν $x \leq 1$ ή $x \geq 3.513$.

Όταν λοιπόν είναι $c\lambda \leq 1$ ή $c\lambda \geq 3.513$, το πρώτο φράγμα είναι μικρότερο.

Για παράδειγμα, αν $\lambda = 1$, $c = 10$, τότε $\dfrac{1}{c\lambda} = 0.1$ και $c\lambda e^{1-\lambda c} = 0.0012341$. □

5.8.4 Η ανισότητα του Chebyshev

Πρόταση 5.7 (Ανισότητα του Chebyshev). *Αν μια (διακριτή ή συνεχής) ΤΜ X έχει πεπερασμένη μέση τιμή μ και διακύμανση σ^2, τότε*

$$P(|X-\mu| \geq c) \leq \frac{\sigma^2}{c^2}, \qquad \text{για κάθε πραγματική σταθερά } c > 0.$$

ή, ισοδύναμα αν $c = k\sigma$,

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}, \qquad \text{για κάθε πραγματική σταθερά } k > 0.$$

Απόδειξη. Εφαρμόζοντας την ανισότητα του Markov για την ΤΜ $Y = (X-\mu)^2 \geq 0$, προκύπτει ότι

$$P(|X-\mu| \geq c) = P((X-\mu)^2 \geq c^2) \leq \frac{E((X-\mu)^2)}{c^2} = \frac{\sigma^2}{c^2}.$$

□

Παράδειγμα 5.8.8. *Έστω ότι η ΤΜ X έχει πεπερασμένες μέση τιμή μ και διακύμανση σ^2. Να βρεθεί ένα κάτω φράγμα για την πιθανότητα $P(\mu - 2\sigma < X < \mu + 2\sigma)$.*

Λύση. Εφαρμόζοντας την ανισότητα του Chebyshev για $c = 2\sigma > 0$, έχουμε ότι

$$P(\mu-2\sigma < X < \mu+2\sigma) = P(-2\sigma < X-\mu < 2\sigma) = P(|X-\mu| < 2\sigma) = 1 - P(|X-\mu| \geq 2\sigma) \geq 1 - \frac{\sigma^2}{(2\sigma)^2} = \frac{3}{4},$$

δηλαδή τουλάχιστον το 75% των τιμών που λαμβάνει η ΤΜ X είναι στο διάστημα $[\mu - 2\sigma, \mu + 2\sigma]$, δηλαδή απέχει το πολύ 2σ (2 φορές την τυπική απόκλιση) από τη μέση τιμή μ. □

Παράδειγμα 5.8.9. *Έστω ότι η ΤΜ X έχει μέση τιμή $\mu = 0$ και διακύμανση $\sigma^2 = 6^2$. Να βρεθεί μια τιμή $\theta > 0$, ώστε με πιθανότητα τουλάχιστον $a = 0.99$ οι τιμές της X να βρίσκονται στο διάστημα $(-\theta, \theta)$.*

Λύση. Θέλουμε $P(-\theta < X < \theta) \geq a$ και

$$P(-\theta < X < \theta) \geq a \Leftrightarrow P(|X| < \theta) \geq a \Leftrightarrow P(|X| \geq \theta) \leq 1 - a.$$

Από την ανισότητα του Chebyshev έχουμε ότι

$$P(|X| \geq \theta) = P(|X-0| \geq \theta) \leq \frac{\sigma^2}{\theta^2}.$$

Αρκεί λοιπόν

$$\frac{\sigma^2}{\theta^2} \leq 1 - a \Leftrightarrow \frac{\sigma^2}{\theta^2} \leq \sqrt{1-a} \Leftrightarrow \theta \geq \frac{\sigma}{\sqrt{1-a}}.$$

Άρα, για $\theta = \dfrac{\sigma}{\sqrt{1-a}} = \dfrac{6}{0.1} = 60$, με πιθανότητα τουλάχιστον $a = 99\%$, η X λαμβάνει τιμές στο διάστημα $(-60, 60)$. □

Παράδειγμα 5.8.10 (Πόσο ακριβές είναι το φράγμα της ανισότητας του Chebyshev). *Έστω ότι η συνεχής ΤΜ X λαμβάνει τιμές στο διάστημα $[-k, k]$, $k > 0$, και ακολουθεί την ομοιόμορφη κατανομή. Να βρεθεί η πιθανότητα $P(|X - \mu| \geq c)$, $c > 0$ και να συγκριθεί με το φράγμα που δίνει η ανισότητα του Chebyshev.*

Λύση. Αφού η X ακολουθεί την ομοιόμορφη κατανομή στο διάστημα $[-k, k]$ θα είναι $\mu = E(X) = \frac{-k+k}{2} = 0$ και $\sigma^2 = \mathrm{V}(X) = \frac{(k-(-k))^2}{12} = \frac{(2k)^2}{12} = \frac{k^2}{3}$.

Επίσης, η συνάρτηση πυκνότητας πιθανότητας της X είναι

$$f(x) = \begin{cases} \frac{1}{2k}, & x \in [-k, k] \\ 0, & x \notin [-k, k] \end{cases}$$

ενώ, η συνάρτηση κατανομής πιθανότητας της X είναι

$$F(x) = P(X \leq x) = \begin{cases} 0, & x < -k \\ \frac{x+k}{2k}, & x \in [-k, k], \\ 1, & x > k \end{cases}$$

Επομένως,

$$\begin{aligned} P(|X - \mu| \geq c) &= P(|X| \geq c) = 1 - P(|X| \leq c) = 1 - P(-c \leq X \leq c) \\ &= 1 - (F(c) - F(-c)) = \begin{cases} 1 - \frac{c}{k}, & c \in [0, k] \\ 0, & c > k. \end{cases} \end{aligned}$$

Η ανισότητα του Chebyshev δίνει

$$P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2} = \frac{k^2}{3c^2}$$

Αν για παράδειγμα $k = 10$ και $c = 7$, τότε έχουμε ότι

$$P(|X - \mu| \geq 7) = 1 - \frac{7}{10} = \frac{3}{10} = 0.3,$$

ενώ η ανισότητα Chebyshev δίνει

$$P(|X - \mu| \geq 7) \leq \frac{10^2}{3 \cdot 7^2} = 0.68,$$

δηλαδή, ενώ η πραγματική πιθανότητα είναι 0.3, το φράγμα 0.68 του δίνει η ανισότητα Chebyshev είναι υπερεκτίμηση της πραγματικής τιμής. □

5.9 Ο νόμος των μεγάλων αριθμών

Πρόταση 5.8. *Αν $X_1, X_2, \ldots, X_n$ είναι ανεξάρτητες τυχαίες μεταβλητές με μέση τιμή μ και διακύμανση σ^2 και $\epsilon > 0$, τότε*

$$P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}$$

Απόδειξη. Έστω $X = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$. Από την γραμμικότητα της μέσης τιμής ισχύει ότι

$$E(X) = E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{1}{n}\left(E(X_1) + E(X_2) + \cdots + E(X_n)\right) = \frac{1}{n} n\mu = \mu.$$

Επιπλέον, επειδή οι X_i είναι ανεξάρτητες, ισχύει ότι

$$\begin{aligned} \mathrm{V}(X) &= \mathrm{V}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{1}{n^2}\mathrm{V}(X_1 + X_2 + \cdots + X_n) \\ &= \frac{1}{n^2}\left(\mathrm{V}(X_1) + \mathrm{V}(X_2) + \cdots + \mathrm{V}(X_n)\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}. \end{aligned}$$

Επομένως, εφαρμόζοντας την ανισότητα του Chebyshev, προκύπτει ότι

$$P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right) = P\left(|X - E(X)| \geq \epsilon\right) \leq \frac{\mathrm{V}(X)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}.$$ □

Πρόταση 5.9 (Ασθενής νόμος των μεγάλων αριθμών). *Έστω $X_1, X_2, \ldots, X_n$ ανεξάρτητες ΤΜ που ακολουθούν την ίδια κατανομή με μέση τιμή μ. Τότε, για κάθε $\epsilon > 0$ έχουμε ότι*

$$\lim_{n\to\infty} P(|\frac{X_1 + X_2 + \cdots X_n}{n} - \mu| < \epsilon) = 1.$$

Με άλλα λόγια, αν καταγράψουμε τις τιμές X_1, X_2, X_3, …, X_n, …, μιας ΤΜ X, τότε ο εμπειρικός μέσος των τιμών αυτών με πιθανότητα που τείνει στο 1 είναι οσοδήποτε κοντά στην μέση τιμή μ της X.

Μια εφαρμογή του ασθενούς νόμου των μεγάλων αριθμών είναι ότι μπορούμε να εκτιμήσουμε την μέση τιμή μιας ΤΜ X με βάση τον εμπειρικό μέσο όρο των τιμών της.

Πρόταση 5.10 (Ισχυρός νόμος των μεγάλων αριθμών). *Έστω $X_1, X_2, \ldots, X_n$ ανεξάρτητες ΤΜ που ακολουθούν την ίδια κατανομή με μέση τιμή μ. Τότε,*

$$P(\lim_{n\to\infty} \frac{X_1 + X_2 + \cdots + X_n}{n} = \mu) = 1.$$

Με άλλα λόγια, αν καταγράψουμε τις τιμές X_1, X_2, X_3, …, X_n, … μιας ΤΜ X, τότε ο εμπειρικός μέσος των τιμών αυτών τείνει στην μέση τιμή με πιθανότητα 1.

Παράδειγμα 5.9.1 (Πείραμα). *Να καταγραφεί ο μέσος όρος του αριθμού των αδελφιών των φοιτητών της τάξης. Στη συνέχεια να γίνει προσπάθεια προσέγγισης επιλέγοντας ένα τυχαίο δείγμα.*

Παράδειγμα 5.9.2. *Να εκτιμηθεί ο αριθμός των ψαριών μιας λίμνης.*

Λύση. Έστω n ο αριθμός των ψαριών της λίμνης. Ψαρεύουμε k ψάρια, τα μαρκάρουμε με κάποιο τρόπο και τα ρίχνουμε πάλι στη λίμνη.

Η πιθανότητα να ψαρέψουμε ένα μαρκαρισμένο ψάρι είναι $p = \frac{k}{n}$.

Ψαρεύουμε ξανά λ ψάρια. Μέσα σε αυτά θα υπάρχουν μ μαρκαρισμένα.

Από τον νόμο των μεγάλων αριθμών έχουμε ότι

$$\frac{\mu}{\lambda} \approx \frac{k}{n} \Leftrightarrow n \approx \frac{k\lambda}{\mu}$$

Για παράδειγμα, αν $k = 1000$, $\lambda = 200$, $\mu = 17$ τότε $n \approx \dfrac{1000 \cdot 200}{17} = 11764.7$. □

Κατανομή	$f(x)$	Μέσος	Διακύμανση
Ομοιόμορφη	$\frac{1}{n}$, $1 \le x \le n$	$\frac{1}{2}(n+1)$	$\frac{1}{12}(n^2-1)$
Bernoulli	$p^x(1-p)^{1-x}$, $x \in \{0,1\}$	p	$p(1-p)$
Διωνυμική	$\binom{n}{x}p^x(1-p)^{n-x}$, $0 \le x \le n$	np	$np(1-p)$
Γεωμετρική	$(1-p)^{x-1}p$, $x \ge 1$	p^{-1}	$(1-p)p^{-2}$
Poisson	$\frac{e^{-\lambda}\lambda^x}{x!}$, $x \ge 1$	λ	λ
Αρνητική διωνυμική	$\binom{x-1}{k-1}p^k(1-p)^{x-k}$, $x \ge k$	kp^{-1}	$k(1-p)p^{-2}$
Υπεργεωμετρική	$\frac{\binom{n}{x}\binom{M-n}{N-x}}{\binom{M}{N}}$, $0, N-M+n \le x \le N, n$	$\frac{nN}{M}$	$\frac{Nn(M-n)(M-N)}{(M-1)M^2}$

Πίνακας 5.1: Σύνοψη διακριτών κατανομών

Κατανομή	$f(x)$	Μέσος	Διακύμανση
Ομοιόμορφη	$\frac{1}{b-a}$, $a \le x \le b$	$\frac{1}{2}(b+a)$	$\frac{1}{12}(b-a)^2$
Εκθετική	$\lambda e^{-\lambda x}$, $x \ge 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Κανονική	$(2\pi)^{-1}\sigma^{-1}\exp(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2)$, $-\infty \le x \le +\infty$	μ	σ^2
Γάμμα	$x^{r-1}e^{-\lambda x}\frac{\lambda^r}{(r-1)!}$, $x \ge 0$	$r\lambda^{-1}$	$r\lambda^{-2}$
Βήτα	$\frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$, $x \in [0,1]$, $a,b>0$	$\frac{a}{a+b}$	$\frac{ab}{(a+b)^2(a+b+1)}$
Laplace	$\frac{1}{2}\lambda\exp(-\lambda\|x\|)$	0	$2\lambda^{-2}$
Cauchy	$(\pi(1+x^2))^{-1}$	-	-

Πίνακας 5.2: Σύνοψη συνεχών κατανομών

5.10 Λυμένες ασκήσεις

Άσκηση 5.1. *Αν παίζεις τάβλι με έναν ισοδύναμο αντίπαλο, ποιο από τα ενδεχόμενα*
Α: Να κερδίσεις τρεις από τις τέσσερις παρτίδες,
Β: Να κερδίσεις έξι από τις οκτώ παρτίδες,
έχει μεγαλύτερη πιθανότητα;
Παρατηρήστε ότι $\frac{3}{4} = \frac{6}{8}$*, δηλαδή η αναλογία των παιχνιδιών είναι η ίδια. Πώς ερμηνεύετε το αποτέλεσμα που βρήκατε;*

Λύση. Ο αριθμός X των παρτίδων που κερδίζουμε ακολουθεί την διωνυμική κατανομή: Στην πρώτη περίπτωση (ενδεχόμενο Α) με παραμέτρους $N = 4$ και $p = 1/2$ και στη δεύτερη (ενδεχόμενο Β) με παραμέτρους $N = 8$ και $p = 1/2$. Επομένως,

$$P(A) = P(X = 3|N = 4) = \binom{4}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^1 = 0.25 \quad \text{και} \quad P(B) = P(X = 6|N = 8) = \binom{8}{6}\left(\frac{1}{2}\right)^6\left(\frac{1}{2}\right)^2 = 0.11.$$

Παρόλο που τα δύο ενδεχόμενα φαίνονται ίδια: 3 στα 4 στο ένα και 6 στα 8 στο άλλο, παρατηρούμε ότι $P(A) > P(B)$. Η εξήγηση είναι ότι όσο περισσότερα παιχνίδια παίζουν οι παίχτες, τόσο περισσότερο η X θα είναι κοντά στην μέση τιμή της (νόμος των μεγάλων αριθμών), δηλαδή στο $E(X) = Np = N/2$ και αποκλίσεις της τάξης του 75% − 25% θα είναι αρκετά σπάνιες. □

Άσκηση 5.2. *Ο μέσος χρόνος εξυπηρέτησης ενός πελάτη στα ταμεία μιας τράπεζας είναι 3,5 λεπτά. Να βρεθεί η πιθανότητα ο χρόνος που θα περιμένετε για να εξυπηρετηθεί ο προηγούμενος από εσάς πελάτης να είναι:*

i) *μεγαλύτερος από πέντε λεπτά.*

ii) *ανάμεσα σε δύο και τέσσερα λεπτά.*

Λύση. Ο χρόνος εξυπηρέτησης X ενός πελάτη ακολουθεί την εκθετική κατανομή με παράμετρο $\lambda = 1/3.5 = 2/7$.

i) $P(X > 5) = 1 - P(X \le 5) = 1 - (1 - e^{-5\cdot 2/7} = e^{-10/7} = 0.239651.$

ii) $P(2 \le X \le 4) = P(X \le 4) - P(X \le 2) = e^{-4\cdot 2/7} - e^{-2\cdot 2/7} = 0.245812.$ □

Άσκηση 5.3. *Μια εταιρεία τηλεφώνων εκτιμά ότι η μέση διάρκεια ζωής της μπαταρίας τους κάτω από κανονικές συνθήκες ακολουθεί την εκθετική κατανομή με μέση τιμή 4 χρόνια. Αν η εταιρεία έχει την δυνατότητα να αντικαταστήσει δωρεάν (μέσα στην εγγύηση) την μπαταρία μέχρι το 40% των συσκευών που θα πωληθούν στην αγορά, ποιος πρέπει να είναι ο μέγιστος χρόνος εγγύησης που θα δοθεί για την μπαταρία;*

Λύση. Έστω X ο χρόνος ζωής της μπαταρίας της συσκευής και y ο ζητούμενος χρόνος εγγύησης. Αφού η εταιρεία έχει τη δυνατότητα να αντικαταστήσει δωρεάν μέχρι το 40% των συσκευών που θα πωληθούν στην αγορά θα πρέπει

$$P(X \le y) \le 0.4$$

Η X ακολουθεί την εκθετική κατανομή με $\lambda = \dfrac{1}{E(X)} = \dfrac{1}{4}$ οπότε

$$P(X \le y) \le 0.4 \Leftrightarrow 1 - e^{-y/4} \le 0.4 \Leftrightarrow 0.6 \le e^{-y/4} \Leftrightarrow e^{y/4} \le \frac{5}{3} \Leftrightarrow \frac{y}{4} \le \ln\frac{5}{3} \Leftrightarrow y \le 4\ln\frac{5}{3} = 2.043.$$

Άρα, ο μέγιστος χρόνος εγγύησης πρέπει να είναι το πολύ 2 χρόνια. □

Άσκηση 5.4. *Ο χρόνος που χρειάζεται ένας πεζός για να φτάσει από τον σταθμό του Ηλεκτρικού του Πειραιά στο κεντρικό κτήριο του Πα.Πει λόγω φαναριών δεν είναι σταθερός αλλά ακολουθεί την κανονική κατανομή με μέση τιμή 17 λεπτά και διακύμανση 9 λεπτά. Να υπολογισθεί η πιθανότητα ο πεζός να φτάσει:*

i) *σε λιγότερο από 19 λεπτά*

ii) *αργότερα από 22 λεπτά*

iii) *σε λιγότερο από 21 αλλά περισσότερο από 13 λεπτά.*

Λύση. Έστω X ο χρόνος μετάβασης από τον σταθμό του Ηλεκτρικού στο κεντρικό κτήριο του Πα.Πει. Η ΤΜ X ακολουθεί την κανονική κατανομή $N(17, 9)$.

i) $P(X \le 19) = P(\dfrac{X-17}{3} \le \dfrac{19-17}{3}) = P(Z \le \dfrac{2}{3}) = \Phi(\dfrac{2}{3}) = \Phi(0.66) = 0.7454.$

ii) $P(X \ge 22) = 1 - P(X \le 22) = 1 - P(\dfrac{X-17}{3} \le \dfrac{22-17}{3}) = 1 - P(Z \le \dfrac{5}{3}) = 1 - \Phi(\dfrac{5}{3}) = \Phi(-\dfrac{5}{3}) = \Phi(-1.66) = 0.0485.$

iii) $P(13 \le X \le 21) = P(\dfrac{13-17}{3} \le \dfrac{X-17}{3} \le \dfrac{21-17}{3}) = P(\dfrac{-4}{3} \le Z \le \dfrac{4}{3}) = \Phi(\dfrac{4}{3}) - \Phi(-\dfrac{4}{3}) = 2\Phi(\dfrac{4}{3}) - 1 = 2\Phi(1.33) - 1 = 2 \cdot 0.9082 - 1 = 0.8164.$ □

Άσκηση 5.5. *Το 40% των ψηφοφόρων μιας πόλης ευνοούν τον υποψήφιο Α. Αν πάρουμε ένα τυχαίο δείγμα 80 ψηφοφόρων, ποια είναι η πιθανότητα να πλειοψηφούν στο δείγμα οι ευνοούντες τον Α;*

Λύση. Έστω X το πλήθος των ψηφοφόρων μεταξύ των 80 οι οποίοι ευνοούν τον A.

Η X ακολουθεί την διωνυμική κατανομή με παραμέτρους $N = 80$ και $p = 0.4$.

$P(X \ge 41) = 1 - P(X \le 40) = 1 - \sum_{i=0}^{40} \binom{80}{i} 0.4^i 0.6^{80-i} = 0.0271236.$

Από το κεντρικό οριακό θεώρημα μπορούμε επίσης να προσεγγίσουμε την X από την κανονική κατανομή $N(80 \cdot 0.4, 80 \cdot 0.4 \cdot 0.6) = N(32, 19.2)$, οπότε

$P(X \ge 41) = 1 - P(X \le 40) = 1 - P(\frac{X-32}{\sqrt{19.2}} \le \frac{40-32}{\sqrt{19.2}}) = 1 - P(Z \le 1.82574) = 1 - \Phi(1.82574) = \Phi(-1.82574) = 0.0344.$ □

Άσκηση 5.6. *Ο Λεωνίδας έχει κανονίσει συνάντηση στο σταθμό του μετρό Άγιος Δημήτριος με φίλους του και πρέπει να είναι εκεί σε 20 λεπτά αν δεν θέλει να τους καθυστερήσει. Έχει τις εξής επιλογές:*

1η επιλογή: Να περιμένει το λεωφορείο το οποίο αναμένεται να έρθει σε 7 λεπτά και με το οποίο ο χρόνος μετάβασης μέχρι τον Άγιο Δημήτριο ακολουθεί την κανονική κατανομή με μέση

τιμή 10 λεπτά και διακύμανση 9 λεπτά.

2η επιλογή: Να περπατήσει μέχρι το σταθμό του μετρό που απέχει 10 λεπτά. Γνωρίζει ότι οι συρμοί έρχονται στο σταθμό κάθε 10 λεπτά και ο χρόνος μετάβασης διαρκεί 4 λεπτά.

3η επιλογή: Να καλέσει ταξί, το οποίο βρίσκεται 10 λεπτά μακριά και ο χρόνος μετάβασης ακολουθεί την κανονική κατανομή με μέση τιμή 6 λεπτά και διακύμανση 4 λεπτά.

Με βάση τα παραπάνω στοιχεία

i) *Να βρεθεί η πιθανότητα να μην αργήσει στη συνάντηση σε κάθε μια από τις επιλογές.*

ii) *Αν ο Λεωνίδας διαλέγει τυχαία μια από τις 3 επιλογές ποιά είναι η πιθανότητα να μην αργήσει για την συνάντηση;*

iii) *Αν ο Λεωνίδας άργησε να πάει στη συνάντηση, ποια είναι η πιθανότητα να μην πήρε ταξί;*

Λύση. Έστω $B \sim N(10, 9)$ ο χρόνος μετάβασης με λεωφορείο. Η πιθανότητα να μην αργήσει παίρνοντας το λεωφορείο είναι ίση με

$$P(B \leq 13) = \Phi((13-10)/3) = \Phi(1) = 0.84134$$

Έστω $M \sim U(0, 10)$ ο χρόνος αναμονής του συρμού του μετρό. Η πιθανότητα να μην αργήσει παίρνοντας το μετρό είναι ίση με

$$P(M \leq 6) = (6-0)/(10-0) = 0.6$$

Έστω $T \sim N(6, 4)$ ο χρόνος μετάβασης του ταξί. Η πιθανότητα να μην αργήσει παίρνοντας ταξί είναι ίση με

$$P(T \leq 10) = \Phi((10-6)/2) = \Phi(2) = 0.97725$$

Αν A, B, M, T, είναι αντίστοιχα τα ενδεχόμενα να άργησε, να πήρε λεωφορείο, μετρό, ταξί, τότε $P(B) = P(M) = P(T) = 1/3$ και

$$P(\overline{A}) = P(\overline{A}|B)P(B) + P(\overline{A}|M)P(M) + P(\overline{A}|T)P(T) = \frac{1}{3}(0.84134 + 0.6 + 0.97725) = 0.8062$$

και

$$P(T|A) = \frac{P(A|T)P(T)}{P(A)} = \frac{1}{3} \cdot \frac{1-0.97725}{1-0.8062} = 0.04 \Rightarrow P(\overline{T}|A) = 0.96.$$

□

Άσκηση 5.7. *Ένας blogger θέλει να εμφανίζει ότι το site του έχει μεγάλη επισκεψιμότητα και χρησιμοποιεί ένα μετρητή επισκεψιμότητας, ο οποίος αυξάνει από τις πραγματικές επισκέψεις στο site αλλά και αυτόματα χρησιμοποιώντας μια γεννήτρια ψευδοτυχαίων αριθμών η οποία παράγει κάθε λεπτό, με ίση πιθανότητα, ένα φυσικό αριθμό από το 1 εως το 10, ο οποίος προστίθεται στον μετρητή επισκεψιμότητας.*

i) *Να βρεθεί μέσος αριθμός των εικονικών επισκέψεων στο site του blogger μέσα σε 1 μέρα.*

ii) *Αν σε 1 μέρα το site είχε 9000 επισκέψεις (πραγματικές και εικονικές), πόσες από αυτές είναι πραγματικές με πιθανότητα τουλάχιστον 99%;*

Λύση. Έστω X_i ο αριθμός των εικονικών επισκέψεων στο i-στό λεπτό. Η X_i ακολουθεί την (διακριτή) ομοιόμορφη κατανομή, με μέση τιμή $E(X_i) = \frac{10+1}{2} = 5.5$ και διακύμανση $\frac{10^2-1}{12} = 8.25$. Μια μέρα έχει $60 * 24 = 1440$ λεπτά, επομένως ο αριθμός X των εικονικών επισκέψεων στο site είναι ίσος με

$$X = X_1 + X_2 + \cdots + X_{1440}.$$

i) Η μέση τιμή της X είναι

$$E(X) = E(X_1) + E(X_2) + \cdots + E(X_{1440}) = 1440 \cdot 5.5 = 7920.$$

Επιπλέον, η διακύμανση της X είναι, λόγω ανεξαρτησίας,

$$V(X) = V(X_1) + V(X_2) + \cdots + V(X_{1440}) = 1440 \cdot 8.25 = 11880.$$

Επομένως, ο μέσος αριθμός εικονικών επισκέψεων στο site του blogger σε μια μέρα είναι 7920.

ii) Μπορούμε να προσεγγίσουμε την ΤΜ X από την κανονική κατανομή με μέση τιμή $\mu = 7920$ και διακύμανση $\sigma^2 = 11880$. Γνωρίζουμε ότι, για μια ΤΜ $Z \sim N(0,1)$, είναι

$$P(|Z| \le z) = P(-z \le Z \le z) = \Phi(z) - \Phi(-z) = 2\Phi(z) - 1,$$

οπότε

$$P(|Z| \le z) = 0.99 \Leftrightarrow 2\Phi(z) - 1 = 0.99 \Leftrightarrow \Phi(z) = 0.995 \Leftrightarrow z \approx 2.576.$$

Επομένως, βάσει του ΚΟΘ, είναι

$$\begin{aligned} 0.99 &\approx P(-z \le \frac{X-\mu}{\sigma} \le z) = P(\mu - z\sigma \le X \le \mu + z\sigma) = P(\mu - 280.77 \le X \le \mu + 280.77) \\ &= P(7639.23 \le X \le 8200.77), \end{aligned}$$

δηλαδή, με πιθανότητα 99% οι εικονικές επισκέψεις είναι από 7639 έως 8201, οπότε οι πραγματικές είναι από 799 έως 1361. □

Άσκηση 5.8. *Έστω ότι οι ανεξάρτητες ΤΜ $X_1, X_2, \ldots, X_n$ έχουν μέση τιμή $E(X_k) = k$ και διακύμανση $\mathrm{Var}(X_k) = \frac{k}{4}$ για κάθε $k \in [n]$. Θεωρούμε την ΤΜ $S_n = X_1 + X_2 + \cdots + X_n$.*

i) *Να βρεθεί η μέση τιμή μ και η διακύμανση σ^2 της S_n.*

ii) *Με τη βοήθεια της ανισότητας του Chebyshev να βρεθεί ένα διάστημα της μορφής $[\mu - \theta, \mu + \theta]$ στο οποίο η X λαμβάνει τιμές με πιθανότητα τουλάχιστον 60%.*

iii) *Να εξηγηθεί γιατί δεν μπορούμε να θεωρήσουμε ότι η S_n προσεγγίζεται από την κανονική κατανομή.*

Λύση.

i) Από την γραμμικότητα της μέσης τιμής έχουμε ότι

$$E(S_n) = E(X_1) + E(X_2) + \cdots + E(X_n) = \sum_{k=1}^{n} k = \frac{n(n+1)}{2}.$$

Επειδή οι $X_1, X_2, \ldots, X_n$ είναι ανεξάρτητες έπεται ότι

$$\sigma^2 = \mathrm{Var}(S_n) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n) = \sum_{k=1}^{n} \frac{k}{4} = \frac{n(n+1)}{8}.$$

ii) Ψάχνουμε $\theta > 0$ ώστε

$$P(|X-\mu)| \leq \theta) \geq 0.6 \Leftrightarrow P(|X-\mu| \geq \theta)) \leq 0.4$$

Από την ανισότητα Chebyshev έχουμε ότι

$$P(|X-\mu| \geq \theta) \leq \frac{\sigma^2}{\theta^2} = \frac{n(n+1)}{8\theta^2}$$

επομένως αρκεί να διαλέξουμε θ ώστε

$$\frac{n(n+1)}{8\theta^2} \leq 0.4 \Leftrightarrow \theta^2 \geq \frac{5n(n+1)}{16}$$

Άρα, αρκεί να επιλέξουμε $\theta = \frac{5n(n+1)}{16}$. Επομένως, το ζητούμενο διάστημα για το οποίο η X λαμβάνει τιμές με πιθανότητα τουλάχιστον 0.6 είναι

$$[\frac{n(n+1)}{2} - \frac{5n(n+1)}{16}, \frac{n(n+1)}{2} + \frac{5n(n+1)}{16}] = [\frac{3n(n+1)}{16}, \frac{13n(n+1)}{16}]$$

iii) Δεν μπορούμε να θεωρήσουμε ότι η X προσεγγίζεται από την κανονική κατανομή διότι οι $X_1, X_2, \ldots, X_n$ δεν ακολουθούν την ίδια κατανομή αφού έχουν διαφορετικές μέσες τιμές ή/και διακυμάνσεις. □

Άσκηση 5.9. *Έστω (απλό μη κατευθυνόμενο) γράφημα $G = (V, E)$ με n κορυφές και m ακμές. Αν $\frac{n}{2} \leq m \leq \frac{n^2}{4}$, τότε το G περιέχει ένα ανεξάρτητο σύνολο με τουλάχιστον $\frac{n^2}{4m}$ κορυφές.*

Λύση. Θέτουμε $V = \{v_1, v_2, \ldots, v_n\}$ και $E = \{e_1, e_2, \ldots, e_m\}$. Για κάθε $S \subseteq V$, συμβολίζουμε με $G(S) = (S, E(S))$ το υπογράφημα του G με σύνολο κορυφών το S και σύνολο ακμών το $E(S)$, όπου κάθε ακμή $e \in E$ ανήκει στο $E(S)$ ανν τα δύο άκρα της ανήκουν στο S. Αρκεί να δειχθεί ότι υπάρχει σύνολο $S \subseteq V$, ώστε το $G(S)$ να έχει τουλάχιστον $\frac{n^2}{4m}$ περισσότερες κορυφές από ακμές, αφού αν διαγράψουμε μια κορυφή από κάθε ακμή του, θα προκύψει ένα υπογράφημα με τουλάχιστον $\frac{n^2}{4m}$ κορυφές και χωρίς ακμές, δηλαδή το ζητούμενο ανεξάρτητο σύνολο.

Επιλέγουμε τυχαία ένα σύνολο S ως εξής: Κάθε κορυφή $v \in V$ επιλέγεται να ανήκει στο S με πιθανότητα p (η τιμή της θα επιλεχθεί κατάλληλα στη συνέχεια). Το $G(S)$ που προκύπτει είναι τυχαίο από τον τρόπο κατασκευής του, οπότε θεωρούμε τις τυχαίες μεταβλητές

$$X_i \sim \text{Bernoulli}(p), \text{ με } X_i = 1 \Leftrightarrow v_i \in S \qquad \text{και} \qquad Y_j \sim \text{Bernoulli}(p^2), \text{ με } Y_j = 1 \Leftrightarrow e_j \in E(S),$$

όπου $i \in [n]$ και $j \in [m]$. Το πλήθος κορυφών και το πλήθος ακμών του $G(S)$ είναι αντίστοιχα $X = X_1 + X_2 + \cdots + X_n$ και $Y = Y_1 + Y_2 + \cdots + Y_m$, οπότε

$$E(X) = E(X_1 + \cdots + X_n) = np \qquad \text{και} \qquad E(Y) = E(Y_1 + \cdots + Y_m) = mp^2.$$

Θεωρώντας την ΤΜ $Z = X - Y$ (πλήθος κορυφών μείον πλήθος ακμών στο $G(S)$), έχουμε ότι $E(Z) = np - mp^2$. Η συνάρτηση $f(p) = np - mp^2$, με $p \in (0, 1]$, παίρνει μέγιστη τιμή όταν $p = \frac{n}{2m}$, δηλαδή

$$E(Z) \leq f(\frac{n}{2m}) = \frac{n^2}{2m} - \frac{n^2}{4m} = \frac{n^2}{4m} = c$$

οπότε $P(Z \leq c) > 0$ (βλ. Πρόταση 5.4), δηλαδή υπάρχει S ώστε το $G(S)$ να έχει τουλάχιστον c περισσότερες κορυφές από ότι ακμές, όταν επιλέξουμε $p = \frac{n}{2m}$. □

5.11 Ασκήσεις για επίλυση

1) Να εξηγηθεί γιατί δεν μπορούμε να έχουμε μια συνεχή μεταβλητή που ακολουθεί την ομοιόμορφη κατανομή στο διάστημα $(0, +\infty)$;

2) Έστω ότι η ΤΜ X λαμβάνει τιμές στο διάστημα $[-2, 2]$ και ακολουθεί την ομοιόμορφη κατανομή. Να υπολογισθούν οι πιθανότητες α) $P(X \leq 1)$, β) $P(|X - 1| \geq \frac{1}{2})$.

3) Μια ράβδος μήκους 2 μέτρων σπάει τυχαία (ομοιόμορφα) στα δύο. Να βρεθούν:

   α) Η μέση τιμή του μήκους του μικρότερου κομματιού που προκύπτει από το σπάσιμο.

   β) Η μέση τιμή του λόγου των μηκών του μικρότερου κομματιού προς το μεγαλύτερο κομμάτι.

4) Το 45% των καπνιστών μιας πόλης προτιμούν την μάρκα τσιγάρων A. Αν πάρουμε ένα τυχαίο δείγμα 100 καπνιστών, ποια είναι η πιθανότητα να πλειοψηφούν στο δείγμα αυτοί που προτιμούν την A;

5) Μια αλυσίδα αποτελείται από 4 κρίκους. Το όριο αντοχής κάθε κρίκου ακολουθεί την κανονική κατανομή με μέσο $\mu = 60$ κιλά και διακύμανση $\sigma^2 = 25$. Αν η αλυσίδα χρησιμοποιηθεί για να σηκώσει ένα φορτίο βάρους 62 κιλών, ποια είναι η πιθανότητα να σπάσει;

6) Έστω ΤΜ $X \sim N(3, 25)$. Να βρεθεί η πιθανότητα $P(5 < X < 7)$.

7) Έστω ΤΜ $X \sim N(\mu, \sigma^2)$, με $P(2 < X < 3) = \dfrac{1}{5}$ και $P(X < 0) = \dfrac{1}{4}$. Να βρεθούν τα μ και σ^2.

8) Έστω ΤΜ $X \sim N(\mu, \sigma^2)$, με $P(X < 1) = \dfrac{1}{4}$. Ποια είναι η ελάχιστη δυνατή τιμή του μ;

9) Αν η ΤΜ X ακολουθεί την κανονική κατανομή $N(\mu, \sigma^2)$, να δειχθεί ότι $E(|X - \mu|) = \sigma \sqrt{\dfrac{2}{p}}$.

10) Ένα αεροπλάνο έχει 120 επιβάτες. Υπάρχουν διαθέσιμα 60 γεύματα ψαριού και 60 γεύματα ζυμαρικών. Κάθε επιβάτης, ανεξάρτητα από τους υπόλοιπους, προτιμά ζυμαρικά με πιθανότητα 0.55 και ψάρι με πιθανότητα 0.45. Να δειχθεί ότι η πιθανότητα 10 ή περισσότεροι επιβάτες να μην μπορούν να λάβουν την πρώτη τους επιλογή είναι περίπου 0.234. (Δίδεται ότι $\Phi(0.734) = 0.7676$ και $\Phi(2.94) = 0.9984$.)

11) Μια συνεχής ΤΜ X λέμε ότι ακολουθεί την λογιστική κατανομή αν η συνάρτηση κατανομής $F(x)$ ορίζεται από τον τύπο

$$F(x) = \frac{1}{1 + e^{-(ax+b)}}, \text{ όπου } a > 0.$$

   Να βρεθεί η συνάρτηση πυκνότητας πιθανότητας $f(x)$ και να αποδειχθεί η σχέση $f(x) = aF(x)(1 - F(x))$.

12) Για να αποφοιτήσει ένας φοιτητής πρέπει να περάσει 48 μαθήματα. Έστω ότι η βαθμολογία του σε κάθε μάθημα που περνά είναι τυχαία μεταβλητή X που λαμβάνει τις τιμές 5, 6, 7, 8, 9 και 10, κάθε μια με πιθανότητα 1/6.

   i) Να υπολογισθούν τα $E(X)$ και $\mathrm{Var}(X)$.

   ii) Να βρεθεί προσεγγιστικά με χρήση του κεντρικού οριακού θεωρήματος η πιθανότητα ο βαθμός πτυχίου να είναι μεγαλύτερος από 8.

Κεφάλαιο 6

Από κοινού πιθανότητα

6.1 Από κοινού συνάρτηση κατανομής πιθανότητας

Ορισμός. *Η από κοινού συνάρτηση κατανομής πιθανότητας (joint CDF) των ΤΜ X, Y συμβολίζεται με $F_{X,Y}$, έχει πεδίο ορισμού το $S_X \times S_Y$ και ορίζεται από τη σχέση*

$$F_{X,Y}(x, y) := P(X \leq x, Y \leq y), \qquad (x, y) \in S_X \times S_Y.$$

Οι F_X και F_Y μπορούν να προκύψουν από τους τύπους

$$F_X(x) = F_{X,Y}(x, +\infty), \qquad F_Y(y) = F_{X,Y}(+\infty, y).$$

6.2 Από κοινού μάζα και πυκνότητα

Ορισμός. *Η από κοινού συνάρτηση μάζας πιθανότητας (joint PMF) των διακριτών ΤΜ X, Y συμβολίζεται με $f_{X,Y}$, έχει πεδίο ορισμού το $S_X \times S_Y$ και ορίζεται από τη σχέση*

$$f_{X,Y}(x, y) := P(X = x, Y = y), \qquad (x, y) \in S_X \times S_Y.$$

Οι f_X, f_Y ονομάζονται περιθώριες (marginal) PMF των X, Y αντίστοιχα και σύμφωνα με τον τύπο της ολικής πιθανότητας, προκύπτουν αντίστοιχα από τις σχέσεις

$$f_X(x) = \sum_y f_{X,Y}(x, y) \qquad \text{και} \qquad f_Y(y) = \sum_x f_{X,Y}(x, y).$$

Ορισμός. *Η από κοινού συνάρτηση πυκνότητας πιθανότητας (joint PDF) των συνεχών ΤΜ X, Y συμβολίζεται με $f_{X,Y}$, έχει πεδίο ορισμού το $S_X \times S_Y$ και είναι μια μη αρνητική συνάρτηση με την ιδιότητα*

$$P((X, Y) \in A) = \iint_A f_{X,Y}(x, y)dxdy,$$

για κάθε $A \subseteq S_X \times S_Y$. Οι X, Y ονομάζονται από κοινού συνεχείς, όταν υπάρχει τέτοια συνάρτηση $f_{X,Y}$.

Οι f_X, f_Y ονομάζονται περιθώριες (marginal) PDF των X, Y αντίστοιχα και σύμφωνα με τον τύπο της ολικής πιθανότητας, προκύπτουν αντίστοιχα από τις σχέσεις

$$f_X(x) = \int_{S_Y} f_{X,Y}(x,y)dy \qquad \text{και} \qquad f_Y(y) = \int_{S_X} f_{X,Y}(x,y)dx.$$

Όταν και μόνο όταν οι X, Y είναι ανεξάρτητες, ως γνωστό ισχύει ότι $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$, για κάθε $A \subseteq S_X$, $B \subseteq S_Y$, οπότε ισχύουν οι τύποι

$$F_{X,Y}(x,y) = F_X(x)F_Y(y), \qquad f_{X,Y}(x,y) = f_X(x)f_Y(y).$$

Παράδειγμα 6.2.1. *Σε μια κοινότητα, το 15% των οικογενειών δεν έχουν κανένα παιδί, το 20% έχουν 1 παιδί, το 35% έχουν 2 παιδιά και το 30% έχουν 3 παιδιά. Θεωρούμε ότι κάθε παιδί είναι αγόρι ή κορίτσι με πιθανότητα 1/2 και ανεξάρτητα από τα υπόλοιπα. Αν B το πλήθος των αγοριών και G το πλήθος των κοριτσιών σε μια τυχαία επιλεγμένη οικογένεια, να υπολογισθεί η από κοινού PMF των B, G.*

Λύση. Θέτουμε A_0, A_1, A_2, A_3 τα ενδεχόμενα η οικογένεια να έχει αντίστοιχα $0, 1, 2$ ή 3 παιδιά. Έχουμε ότι

$$\begin{aligned}
P(B=0,G=0) &= P(A_0) = 0.15\\
P(B=0,G=1) &= P(A_1,G=1) = P(G=1|A_1)P(A_1) = 0.5\cdot 0.2 = 0.1 = P(B=1,G=0)\\
P(B=0,G=2) &= P(A_2,G=2) = P(G=2|A_2)P(A_2) = 0.25\cdot 0.35 = 0.0875 = P(B=2,G=0)\\
P(B=0,G=3) &= P(A_3,G=3) = P(G=3|A_3)P(A_3) = 0.125\cdot 0.3 = 0.0375 = P(B=3,G=0)\\
P(B=1,G=1) &= P(A_2,G=1) = P(G=1|A_2)P(A_2) = 0.5\cdot 0.35 = 0.175\\
P(B=1,G=2) &= P(A_3,G=2) = P(G=2|A_3)P(A_3) = \frac{3}{8}\cdot 0.3 = 0.1125 = P(B=2,G=1)
\end{aligned}$$

Συνολικά, η από κοινού PMF των B, G (δηλαδή οι τιμές $P(B=i, G=j)$), δίνεται στον ακόλουθο πίνακα:

$G=j$ / $B=i$	0	1	2	3	$P(B=i)$
0	0.15	0.1	0.0875	0.0375	0.375
1	0.1	0.175	0.1125	0	0.3875
2	0.0875	0.1125	0	0	0.2
3	0.0375	0	0	0	0.0375
$P(G=j)$	0.375	0.3875	0.2	0.0375	

□

Παράδειγμα 6.2.2. *Να εξετασθεί αν οι ΤΜ X και Y, με από κοινού PMF όπως στον επόμενο πίνακα, είναι ανεξάρτητες.*

Y \ X	0	1	2
0	1/6	0	1/4
1	0	1/6	0
2	1/4	0	1/6

Λύση. Από τα αθροίσματα της πρώτης γραμμής και της δεύτερης στήλης, βρίσκουμε ότι

$$P(X=0) = 1/6 + 1/4 = 5/12 \qquad \text{και} \qquad P(Y=1) = 1/6,$$

οπότε $P(X=0, Y=1) = 0 \neq P(X=0)P(Y=1)$. Επομένως, οι X, Y δεν είναι ανεξάρτητες. □

Παράδειγμα 6.2.3. *Η ΤΜ X παίρνει ομοιόμορφα τιμές στο $\{2,3,4\}$ και η ΤΜ Y ακολουθεί Γεωμετρική κατανομή με παράμετρο $1/X$. Να υπολογισθεί η $f_{X,Y}$. Στη συνέχεια, να υπολογισθεί η f_Y.*

Λύση. Είναι

$$f(x,y) = P(X=x, Y=y) = P(Y=y|X=x)P(X=x) = \frac{1}{3}\left(1-\frac{1}{x}\right)^{y-1}\frac{1}{x}.$$

Για την περιθώρια PMF της Y, είναι

$$f_Y(y) = \sum_{x=2}^{3} \frac{1}{3}\left(1-\frac{1}{x}\right)^{y-1}\frac{1}{x} = \frac{1}{3}\left(\frac{1}{2^y} + \frac{2^{y-1}}{3^y} + \frac{3^{y-1}}{4^y}\right)$$

□

Παράδειγμα 6.2.4. *Δύο ΤΜ X, Y κατανέμονται ομοιόμορφα στο $S = [0,1]^2$, δηλαδή $f_{X,Y} = \begin{cases} 1, & (x,y) \in S, \\ 0, & (x,y) \notin S \end{cases}$.*

i) Να υπολογισθούν οι πιθανότητες $P(0 \le X \le 1/2, 0 \le Y \le 1/2)$ και $P(0 \le Y \le X \le 3/4)$.

Λύση. i) είναι

$$P(0 \le X \le 1/2, 0 \le Y \le 1/2) \int_0^{1/2}\int_0^{1/2} 1 dx dy = \int_0^{1/2} [y]_0^{1/2} dy = \int_0^{1/2} 1/2 dy = 1/4$$

και

$$P(0 \le Y \le X \le 3/4) = \int_0^{3/4}\left(\int_0^x 1 dy\right) dx = \int_0^{3/4} x dx = [x^2/2]_0^{3/4} = 9/32.$$

□

Παράδειγμα 6.2.5. *Δύο ανεξάρτητες συνεχείς ΤΜ X, Y με θετικές τιμές έχουν την ίδια PDF $f_X(x) = f_Y(x) = \begin{cases} e^{-x}, & x > 0, \\ 0, & x \le 0 \end{cases}$. Να βρεθεί η PDF της ΤΜ $Z = X/Y$.*

Λύση. Για $z > 0$, θέτουμε $A_z = \{(x, y) \in \mathbb{R}^2 : 0 < x \leq zy\}$. Λόγω ανεξαρτησίας, είναι $f_{X,Y}(x, y) = f_X(x)f_Y(y)$, οπότε

$$\begin{aligned} F_Z(z) = P(X/Y \leq z) = P(X \leq zY) &= \iint_{A_z} f_{X,Y}(x, y)dxdy = \int_0^{+\infty} \int_0^{zy} f_{X,Y}(x, y)dxdy \\ &= \int_0^{+\infty} e^{-y} \left(\int_0^{zy} e^{-x}dx \right) dy = \int_0^{+\infty} e^{-y}[1 - e^{-zy}]dy = \left[-e^{-y} + \frac{e^{-(z+1)y}}{z+1} \right]_0^{+\infty} \\ &= 1 - \frac{1}{z+1}. \end{aligned}$$

Επομένως, $f_Z(x) = \dfrac{d}{dz}F_Z(z) = \dfrac{1}{(1+z)^2}$. □

6.3 Μέση τιμή και συνδιακύμανση

Η μέση τιμή μιας συνάρτησης $g(x, y)$ δύο διακριτών ΤΜ X, Y με από κοινού μάζα $f_{X,Y}(x, y)$, δίνεται από τον τύπο

$$E(g(x, y)) = \sum_x \sum_y g(x, y)f_{X,Y}(x, y).$$

Στην περίπτωση που οι X, Y είναι συνεχείς, τότε ισχύει ο τύπος

$$E(g(x, y)) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x, y)f_{X,Y}(x, y)dxdy.$$

Ορισμός. *Η συνδιακύμανση (covariance) δύο ΤΜ X, Y συμβολίζεται ως* $\mathrm{COV}(X, Y)$ *και ορίζεται από τη σχέση*

$$\mathrm{COV}(X, Y) := E((X - E(X))(Y - E(Y))).$$

Η συνδιακύμανση των ΤΜ X, Y ικανοποιεί τις επόμενες ιδιότητες:

- $\mathrm{COV}(X, X) = \mathrm{V}(X)$,
- $\mathrm{COV}(X, -X) = -\mathrm{V}(X)$,
- $\mathrm{COV}(X, Y) = E(XY) - E(X)E(Y)$,
- $\mathrm{V}(X + Y) = \mathrm{V}(X) + \mathrm{V}(Y) + 2\mathrm{COV}(X, Y)$.

Από τις δύο τελευταίες ιδιότητες προκύπτει ότι

$$X, Y \text{ ανεξάρτητες } \Rightarrow \mathrm{COV}(X, Y) = 0.$$

Το αντίστροφο όμως δεν ισχύει πάντα. Γενικά, αν $\mathrm{COV}(X, Y) = 0$, τότε οι X, Y ονομάζονται ασυσχέτιστες (uncorrelated).

6.4 Άλυτες ασκήσεις

1. Ρίχνουμε ένα δίκαιο ζάρι 2 φορές και έστω X_1 και X_2 τα αποτελέσματα των δύο ρίψεων. Να βρεθεί η από κοινού PMF των X_1, X_2. Αν $X = \min\{X_1, X_2\}$ και $Y = \max\{X_1, X_2\}$, να βρεθεί η από κοινού PMF των X, Y.

Κεφάλαιο 7

Τεχνικές δειγματοληψίας

7.1 Εισαγωγή

Μερικές φορές θέλουμε να παράγουμε τυχαίους αριθμούς (ή διανύσματα) οι οποίοι ακολουθούν μια συγκεκριμένη κατανομή πιθανότητας.

Το πρώτο βήμα στην κατεύθυνση αυτή είναι η προσομοίωση μιας ΤΜ $X \sim U((0,1))$ με ομοιόμορφη κατανομή στο $(0,1)$. Για το σκοπό αυτό χρησιμοποιούνται αλγόριθμοι (random number generators) που συνήθως ξεκινούν από μια αρχική τιμή x_0 (seed) και στη συνέχεια υπολογίζουν επαναληπτικά ακέραιες τιμές στο $[0, 1, \ldots m-1]$ βάσει του αναγωγικού τύπου

$$x_{n+1} = (kx_n + \ell) \pmod{m}, \qquad k, \ell, m \in \mathbb{N}^*$$

Αν οι συντελεστές k, ℓ, m επιλεχθούν κατάλληλα, τότε οι τιμές x_n/m κατανέμονται περίπου ομοιόμορφα στο $(0,1)$ και για το λόγο αυτό ονομάζονται ψευδοτυχαίοι αριθμοί.

Σημειώνεται ότι με αυτόν τον τρόπο μπορούμε να προσομοιώσουμε οποιαδήποτε ΤΜ $Y = U((a,b))$, αρκεί να ορίσουμε $Y = a + (b-a)X$.

Παράδειγμα 7.1.1. *Να προσομοιωθεί μια διακριτή ΤΜ X που παίρνει τιμές ομοιόμορφα στο $[n] = \{1, 2, \ldots, n\}$.*

Απάντηση. Έστω $U \sim U((0,1))$. Θέτουμε $X = \lfloor nU \rfloor + 1$. Βάσει της ανισότητας $\lfloor x \rfloor \le x < \lfloor x \rfloor + 1$, $x \in \mathbb{R}$, έχουμε ότι

$$\lfloor nU \rfloor + 1 = k \Leftrightarrow \lfloor nU \rfloor = k-1 \Leftrightarrow k-1 \le nU < k \Leftrightarrow \frac{k-1}{n} \le U < \frac{k}{n}.$$

Άρα

$$P(X=k) = P(\lfloor nU \rfloor + 1 = k) = P(\frac{k-1}{n} \le U < \frac{k}{n}) = \frac{k}{n} - \frac{k-1}{n} = \frac{1}{n},$$

δηλαδή η X κατανέμεται ομοιόμορφα στο $[n]$. □

Βάσει της ΤΜ X του προηγούμενου παραδείγματος, μπορούμε να κατασκευάσουμε μια τυχαία μετάθεση του $[n]$ ομοιόμορφα (δηλαδή όλες οι δυνατές μεταθέσεις) είναι ισοπίθανες. Ξεκινάμε από μια οποιαδήποτε μετάθεση, π.χ. την ταυτοτική $\sigma = (1, 2, 3, \ldots, n)$ και για κάθε j από 1 έως n παράγουμε έναν τυχαίο αριθμό X στο $[j, j+1, \ldots, n]$ και αντιμεταθέτουμε τα στοιχεία σ_j και σ_X.

```
from numpy import random
def randPerm(n):
    a = [i for i in range(n+1)] #identity permutation, a[0] is dummy
    for j in range(1,n+1): #for j = 1 to n
        X = random.randint(j,n+1) #choose random X from j to n
        temp = a[X] #swap A[j], A[X]
        a[X] = a[j]
        a[j] = temp
    return a[1:] #ignore a[0]
```

Η παραπάνω ρουτίνα εκτελείται σε χρόνο $O(n)$ και χρησιμοποιείται συχνά για την τυχαιοποίηση της εισόδου ενός αλγορίθμου, για τον οποίο η διάταξη της εισόδου δεν επηρεάζει την ορθότητα, αλλά επηρεάζει το χρόνο εκτέλεσης. Χρησιμοποιείται σε περιπτώσεις που αν η είσοδος έχει τυχαία διάταξη, τότε η περίπτωση χειρότερης εκτέλεσης έχει πολύ μικρή πιθανότητα.

7.2 Προσομοίωση Monte Carlo

Παράδειγμα 7.2.1. *Δίνεται μια συνάρτηση $g(x)$ και ζητείται να υπολογισθεί προσεγγιστικά το ολοκλήρωμα*

$$I = \int_a^b g(x)dx$$

Απάντηση. Θεωρούμε την ΤΜ $X \sim U((a,b))$, οπότε $E(g(X)) = \int_a^b g(x)\frac{1}{b-a}dx = \frac{I}{b-a}$. Στη συνέχεια, παράγουμε ένα τυχαίο δείγμα $(x_1, x_2, \ldots, x_n)$ τιμών της X και προσεγγίζουμε το I βάσει του δειγματικού μέσου της $g(X)$, δηλαδή

$$I = (b-a)E(g(X)) \approx \frac{b-a}{n}\sum_{i=1}^{n} g(x_i).$$

□

Η παραπάνω προσέγγιση γενικεύεται άμεσα για την προσέγγιση πολλαπλού ολοκληρώματος.

7.3 Προσομοίωση διακριτών τυχαίων μεταβλητών

Αν θέλουμε να προσομοιώσουμε μια διακριτή ΤΜ X με συνάρτηση κατανομής F, μπορούμε να το επιτύχουμε μέσω του παρακάτω αλγορίθμου:

Αλγόριθμος δειγματοληψίας αντίστροφου μετασχηματισμού (inverse transform sampling)

- Παράγουμε ομοιόμορφα ένα τυχαίο $U \in (0,1)$
- Υπολογίζουμε το ελάχιστο $x \in S_X$ ώστε $U \leq F(x)$.
- Επιστρέφουμε το x ως την τιμή της ΤΜ X.

Απόδειξη. Έστω $S_x = \{x_i : i \in I\}$, όπου $I \subseteq \mathbb{N}^*$ (πεπερασμένο ή άπειρο) και $(x_i)_{i\in I}$ γνησίως αύξουσα. Αν τεθεί $p_i = P(X = x_i)$, τότε $F(x_i) = P(X \leq x_i) = \sum_{j=1}^{i} p_i$. Δεδομένης της τυχαίας τιμής της U, επειδή η F είναι αύξουσα με τιμές από 0 έως 1, προφανώς υπάρχει μοναδικός δείκτης $i \geq 1$ τέτοιος ώστε

$F(x_{i-1}) < U \leq F(x_i)$, όπου ορίζουμε $F(x_0) = 0$. Ταυτίζοντας το ενδεχόμενο $X = x_i$ με το ενδεχόμενο $F(x_{i-1}) < U \leq F(x_i)$, έχουμε ότι

$$P(X = x_i) = P(F(x_{i-1}) < U \leq F(x_i)) = P(\sum_{j=1}^{i-1} p_i < U \leq \sum_{j=1}^{i} p_i) = \sum_{j=1}^{i} p_i - \sum_{j=1}^{i-1} p_i = p_i,$$

δηλαδή οι τιμές της X πράγματι ακολουθούν την F. □

Παρατήρηση: Αν το σύνολο τιμών S_x είναι πεπερασμένο, δηλαδή $|S_x| = n$, για κάποιο $n \in \mathbb{N}^*$, τότε η ακολουθία $(F(x_i))_{i \in I}$ είναι αύξουσα και πεπερασμένη, οπότε μπορούμε να εντοπίσουμε τον ελάχιστο δείκτη i ώστε $U \leq F(x_i)$ με δυαδική αναζήτηση σε χρόνο $O(\log n)$.

Παράδειγμα 7.3.1 (Διακριτή κατανομή). *Να παραχθούν* 10 *τυχαίοι αριθμοί που ακολουθούν την εξής κατανομή*

$$P(X = 1) = \frac{1}{6}, \quad P(X = 2) = \frac{1}{7}, \quad P(X = 3) = \frac{1}{3}, \quad P(X = 4) = \frac{15}{42}.$$

Λύση. Υπολογίζουμε την συνάρτηση κατανομής πιθανότητας $F(x)$ για την τυχαία μεταβλητή X.

$$F(X) = \begin{cases} 0, & x \in (-\infty, 1) \\ \frac{1}{6}, & x \in [1, 2) \\ \frac{1}{6} + \frac{1}{7}, & x \in [2, 3) \\ \frac{1}{6} + \frac{1}{7} + \frac{1}{3}, & x \in [3, 4) \\ 1, & x \in [4, +\infty) \end{cases} = \begin{cases} 0, & x \in (-\infty, 1) \\ \frac{1}{6}, & x \in [1, 2) \\ \frac{13}{42}, & x \in [2, 3) \\ \frac{27}{42}, & x \in [3, 4) \\ 1, & x \in [4, +\infty) \end{cases} = \begin{cases} 0, & x \in (-\infty, 1) \\ 0.166, & x \in [1, 2) \\ 0.309, & x \in [2, 3) \\ 0.64, & x \in [3, 4) \\ 1, & x \in [4, +\infty) \end{cases}$$

Διαλέγουμε ομοιόμορφα 10 αριθμούς $U_1, U_2, \ldots, U_{10} \in (0, 1)$.

Αν π.χ. $U_1 = 0.3$, τότε βρίσκουμε το ελάχιστο x ώστε $F(x) \geq 0.3$, δηλαδή $x = 2$. Άρα, παράγουμε το 2.

Αν π.χ. $U_2 = 0.59$, τότε το ελάχιστο x ώστε $F(x) \geq 0.59$ είναι το $x = 3$. Άρα, παράγουμε το 3.

Αν π.χ. $U_3 = 0.48$, τότε το ελάχιστο x ώστε $F(x) \geq 0.48$ είναι το $x = 3$. Άρα, παράγουμε το 3, κ.ο.κ. □

Παράδειγμα 7.3.2. *Να προσομοιωθεί μια ΤΜ X που ακολουθεί κατανομή Bernoulli με παράμετρο $p \in (0, 1)$.*

Λύση. Θεωρούμε ΤΜ $U \sim U((0, 1))$. Επειδή $S_x = \{0, 1\}$,

$$P(X = 1) = p = P(U \leq p) \qquad \text{και} \qquad P(X = 0) = 1 - p = P(U > p),$$

θέτουμε

$$X = \begin{cases} 1, & \text{αν } U \leq p, \\ 0, & \text{αν } U > p \end{cases}$$

□

Άσκηση 7.1. *Να προσομοιωθεί μια ΤΜ X που ακολουθεί:*

i) Διωνυμική κατανομή με παραμέτρους $n \in \mathbb{N}^*$, $p \in (0, 1)$.

ii) Κατανομή Poisson με παράμετρο $\lambda > 0$.

iii) Γεωμετρική κατανομή με παράμετρο $p \in (0, 1)$.

7.4 Δειγματοληψία μέσω αντίστροφου μετασχηματισμού

Πρόταση 7.1. *Έστω συνάρτηση κατανομής $F(x)$ μιας συνεχούς τυχαίας μεταβλητής. Αν η U κατανέμεται ομοιόμορφα στο $(0,1)$, και η F είναι γνησίως αύξουσα (άρα και αντιστρέψιμη) τότε η ΤΜ*

$$X = F^{-1}(U)$$

είναι καλά ορισμένη και ακολουθεί την κατανομή $F(x)$.

Απόδειξη. Υπενθυμίζεται ότι για την ομοιόμορφη κατανομή στο $(0,1)$ ισχύει ότι $P(U \leq x) = x$ για κάθε $x \in [0,1]$.

$$P(X \leq x) = P(F^{-1}(U) \leq x) \overset{1}{=} P(F(F^{-1}(U) \leq F(x)) = P(U \leq F(x)) = F(x).$$

□

Παράδειγμα 7.4.1 (Συνεχής κατανομή). *Να παραχθούν 10 τυχαία δείγματα της εκθετικής κατανομής με παράμετρο $\lambda = 2$.*

Απόδειξη. Η συνάρτηση κατανομής πιθανότητας της εκθετικής κατανομής είναι γνησίως αύξουσα, με τύπο

$$F(x) = P(X \leq x) = 1 - e^{-\lambda x}.$$

Υπολογίζουμε τον τύπο της F^{-1}:

$$y = 1 - e^{-\lambda x} \Leftrightarrow e^{-\lambda x} = 1 - y \Leftrightarrow -\lambda x = \ln(1-y) \Leftrightarrow x = \frac{-1}{\lambda}\ln(1-y),$$

Άρα $F^{-1}(x) = \dfrac{-1}{\lambda}\ln(1-x)$.

Σύμφωνα με τον αλγόριθμο δειγματοληψίας με αντίστροφο μετασχηματισμό, θεωρούμε την ΤΜ $F^{-1}(U) = \dfrac{-1}{\lambda}\ln(1-U)$, όπου $U \in (0,1)$.

Επιπλέον, παρατηρούμε αν το U επιλέγεται ομοιόμορφα στο $(0,1)$, τότε και το $V = 1 - U$ επιλέγεται εξίσου ομοιόμορφα στο $(0,1)$. Επομένως, μπορούμε ισοδύναμα να θεωρήσουμε την ΤΜ

$$X = \frac{-1}{\lambda}\ln U, \qquad U \in (0,1).$$

Ο τελικός τύπος για την παραγωγή των ζητούμενων δειγμάτων είναι ο

$$x = -\frac{\ln U}{\lambda} \text{ όπου } U \text{ παράγεται ομοιόμορφα στο } U \in (0,1).$$

□

Άσκηση 7.2. *Να προσομοιωθεί μια ΤΜ $X \sim N(0,1)$ μέσω του αντίστροφου μετασχηματισμού. (Υπόδειξη: Η F^{-1} μπορεί να προσδιορισθεί έμμεσα μέσω της παραγώγου της.)*

Υπάρχουν περιπτώσεις που είτε η εύρεση της F^{-1} είναι δύσκολη ή και αδύνατη, είτε απαιτούν πολλούς υπολογισμούς (π.χ. λογαρίθμους). Για ορισμένες τυχαίες μεταβλητές έχουν αναπτυχθεί ειδικές μέθοδοι, που αποφεύγουν αυτές τις δυσκολίες.

[1] Η F είναι αύξουσα.

Πρόταση 7.2 (Μέθοδος Box-Muller). *Έστω ζεύγος (U_1, U_2) ανεξάρτητων συνεχών ΤΜ ομοιόμορφα κατανεμημένων στο $(0,1)$. Το ζεύγος ΤΜ (R, Θ), με*

$$R = \sqrt{-2\ln U_1}, \qquad \Theta = 2\pi U_2$$

αποτελεί τις πολικές συντεταγμένες ενός ζεύγους (X, Y) ανεξάρτητων ΤΜ που ακολουθούν την $N(0,1)$.

Απόδειξη. Το ζεύγος (X, Y) 2 ανεξάρτητων ΤΜ που ακολουθούν την $N(0,1)$, έχει από κοινού συνάρτηση πυκνότητας πιθανότητας την

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}\frac{1}{\sqrt{2\pi}}e^{-y^2/2} = \frac{1}{2\pi}e^{-(x^2+y^2)/2}.$$

Έστω (R, Θ) οι πολικές συντεταγμένες του ζεύγους (X, Y).

Άμεσα προκύπτει ότι οι R, Θ είναι ανεξάρτητες και ότι $\Theta \sim U([0, 2\pi])$ και $R^2 \sim E(1/2)$. Πράγματι, αν η αλλαγή σε πολικές συντεταγμένες μετατρέπει το χωρίο A στο $A' = [0, \sqrt{r}] \times [0, t]$, τότε

$$\begin{aligned} F_{R^2,\Theta}(r,t) &= P(R^2 \le r, \Theta \le t) = P((X,Y) \in A) = \iint_A \frac{1}{2\pi}e^{-(x^2+y^2)/2} = \frac{1}{2\pi}\iint_{A'} \rho e^{-(\rho^2)/2}d\rho d\theta \\ &= \frac{1}{2\pi}\int_0^t\int_0^{\sqrt{r}} \rho e^{-\rho^2/2}d\rho d\theta = \frac{1}{2\pi}\int_0^t d\theta \int_0^{\sqrt{r}}(-e^{-\rho^2/2})'d\rho = \frac{t}{2\pi}\left[-e^{-\rho^2/2}\right]_0^{\sqrt{r}} \\ &= \frac{t}{2\pi}(1-e^{-r/2}) = F_\Theta(t)F_{R^2}(r). \end{aligned}$$

(Η τελευταία ισότητα προκύπτει θεωρώντας τις περιθώριες κατανομές.)

Άρα, η R^2, σύμφωνα με προηγούμενο παράδειγμα, μπορεί να προσομοιωθεί βάσει του τύπου $R^2 = -2\ln U_1$, ενώ η Θ προσομοιώνεται βάσει του τύπου $\Theta = 2\pi U_2$.

Χρησιμοποιώντας τον τύπο του μετασχηματισμού σε πολικές συντεταγμένες

$$X = R\cos\Theta, \qquad Y = R\sin\Theta,$$

καταλήγουμε ότι οι

$$X = \sqrt{-2\ln U_1}\cos(2\pi U_2) \qquad \text{και} \qquad Y = \sqrt{-2\ln U_1}\sin(2\pi U_2)$$

είναι ανεξάρτητες και ακολουθούν την $N(0,1)$. □

7.4.1 Δειγματοληψία μέσω απόρριψης

Σε ορισμένες κατανομές πιθανότητας είναι πολύ δύσκολο ή αδύνατο να βρούμε τον αντίστροφο μετασχηματισμό κάποιων συνεχών κατανομών.

Παράδειγμα 7.4.2. *Να παραχθούν ομοιόμορφα n σημεία (x, y) μέσα στον κυκλικό δίσκο που ορίζεται από την ανίσωση*

$$\left(x-\frac{1}{2}\right)^2 + \left(y-\frac{1}{2}\right)^2 \le 1$$

Λύση. Μια πρώτη προσέγγιση είναι παράγουμε ομοιόμορφα και ανεξάρτητα $x, y \in [0,1]$ (δηλαδή διαλέγουμε ομοιόμορφα σημεία (x, y) στο τετράγωνο $[0,1] \times [0,1]$ και κρατάμε μόνο αυτά τα οποία ικανοποιούν την ανισότητα του κυκλικού δίσκου, ενώ απορρίπτουμε (reject) τα υπόλοιπα. □

Παρατήρηση: Μπορούμε να παράγουμε ομοιόμορφα σημεία στον δίσκο χωρίς απόρριψη με το εξής τρόπο: Ας υποθέσουμε ότι ο δίσκος είναι κέντρου $(0,0)$ και ακτίνας r για απλότητα. Συμβολίζουμε έναν τέτοιο δίσκο με $C(r)$. Αν το ζεύγος (X,Y) κατανέμεται ομοιόμορφα στο $C(r)$, τότε

$$P((X,Y) \in C(d)) = \frac{\pi d^2}{\pi r^2} = \frac{d^2}{r^2}, \qquad d \in [0,r].$$

Αν θεωρήσουμε για τις αντίστοιχες πολικές συντεταγμένες (R,Θ) ότι είναι ανεξάρτητες, με $R^2 \sim U([0,r^2])$ και $\Theta \sim U([0,2\pi))$, τότε

$$P((X,Y) \in C(d)) = P((R^2,\Theta) \in [0,d^2] \times [0,2\pi)) = \frac{2\pi d^2}{2\pi r^2} = \frac{d^2}{r^2}.$$

Επομένως, μπορούμε να πάρουμε τυχαία δείγματα $U_1 \sim U([0,1])$ και $U_2 \sim U([0,1))$ και έπειτα να θέσουμε

$$X = \sqrt{r^2 U_1}\cos(2\pi U_2), \qquad Y = \sqrt{r^2 U_1}\sin(2\pi U_2)$$

Παράδειγμα 7.4.3. *Να προσεγγισθεί η τιμή του ολοκληρώματος*

$$I = \int_a^b f(x)dx$$

μιας συνεχούς μη αρνητικής συνάρτησης $f/[a,b]$.

Λύση. Έστω M ένα άνω φράγμα για τη συνάρτηση $f/[a,b]$. Παράγουμε n σημεία ομοιόμορφα στο ορθογώνιο $[a,b]\times[0,M]$, χρησιμοποιώντας δύο ανεξάρτητες μεταβλητές $X \sim U((a,b))$ και $Y \sim U([0,M])$. Έστω m το πλήθος των ζευγών (X,Y) με $Y \le f(X)$. Τότε $I \approx m/n$. □

Αλγόριθμος δειγματοληψίας μέσω απόρριψης (rejection sampling)

Έστω ότι θέλουμε να παράγουμε δείγματα της τυχαίας μεταβλητής X με συνάρτηση πυκνότητας πιθανότητας $f(x)$. Αν υποθέσουμε ότι γνωρίζουμε να παράγουμε δείγματα της ΤΜ Y με συνάρτηση πυκνότητας πιθανότητας $g(y)$, τότε

- Βρίσκουμε μια σταθερά $c > 0$ ώστε $f(x) \le cg(x)$ για κάθε $x \in \mathbb{R}$. (Η σταθερά c μπορεί να υπολογισθεί ως η μέγιστη τιμή της συνάρτησης $\frac{f(x)}{g(x)}$.)
- Παράγουμε ένα τυχαίο δείγμα y που ακολουθεί την κατανομή της Y.
- Παράγουμε ομοιόμορφα ένα τυχαίο $U \in (0,1)$
- Αν $U \le \frac{f(y)}{cg(y)}$, τότε το y είναι δείγμα της κατανομής της X.
- Αν $U > \frac{f(y)}{cg(y)}$, τότε απορρίπτεται το y.
- Επιστρέφουμε το y.

Απόδειξη. Αρκεί να δειχθεί ότι $P(X \le x) = \int_{-\infty}^{x} f(y)dy$. Έστω Y μια τιμή που επιστρέφεται, δηλαδή ισχύει ότι $U \le \frac{f(Y)}{cg(Y)}$. Θέτοντας $K = \frac{f(Y)}{cg(Y)} = P(U \le \frac{f(Y)}{cg(Y)}) > 0$, έχουμε ότι

$$P(X \le x) = P(Y \le x | U \le K) = \frac{1}{K} P(Y \le x, U \le K)$$

και

$$P(Y \le x, U \le K) = \int_{-\infty}^{+\infty} P(Y \le x, U \le K | Y = y) g(y) dy = \int_{-\infty}^{x} P(U \le \frac{f(y)}{cg(y)}) g(y) dy = \frac{1}{c} \int_{-\infty}^{x} f(y) dy.$$

Επομένως,

$$P(X \le x) = \frac{1}{cK} \int_{-\infty}^{x} f(y) dy$$

και παίρνοντας όρια όταν $x \to +\infty$, έχουμε ότι $cK = 1$. □

Παράδειγμα 7.4.4. *Να προσομοιωθεί μια ΤΜ X με σππ $f(x) = x(1-x)^2$, $x \in (0,1)$.*

Λύση. Θεωρούμε $g(x) = 1$ τη σππ μιας $Y \sim U((0,1))$. Η συνάρτηση $h(x) = f(x)/g(x) = x(1-x)^2$ έχει πρώτη παράγωγο

$$h'(x) = (1-x)^2 - 2x(1-x) = (1-x)(1-x-2x) = (1-x)(1-3x) \le 0 \Leftrightarrow x \ge 1/3$$

άρα έχει μέγιστη τιμή $c = h(1/3) = \frac{1}{3}\frac{2^2}{3^2} = \frac{4}{27}$.

Παράγουμε λοιπόν ένα δείγμα από την $U \sim U((0,1))$ και ένα από την Y και αν

$$U \le \frac{f(Y)}{cg(Y)} = \frac{27}{4} Y(1-Y)^2,$$

τότε επιστρέφουμε το Y, αλλιώς το απορρίπτουμε. □

7.4.2 Προσομοίωση διαδικασίας Poisson

Έστω $N(t)$ το πλήθος αφίξεων σε ένα σύστημα μέσα σε χρόνο t. Υπενθυμίζεται ότι η ακολουθία $(N(t))_{t \ge 0}$, όπου $N(0) = 0$, ονομάζεται διαδικασία Poisson με παράμετρο $\lambda > 0$, όταν

$$P(N(t) = k) = e^{-\lambda t} \frac{(\lambda t)^k}{k!}$$

και τότε η ΤΜ X_1 που αντιπροσωπεύει το χρόνο μέχρι την πρώτη άφιξη ακολουθεί εκθετική κατανομή με παράμετρο λ.

Επομένως, μπορούμε να προσομοιώσουμε τη διαδικασία, προσομοιώνοντας τους χρόνους μεταξύ διαδοχικών αφίξεων οι οποίοι ακολουθούν εκθετική κατανομή.

Μπορούμε όμως να την προσομοιώσουμε χρησιμοποιώντας απευθείας ομοιόμορφη ΤΜ Αν γνωρίζουμε ότι συνέβη μια άφιξη μέσα σε χρόνο t, τότε η πιθανότητα να συνέβη σε χρόνο $t_1 < t$ είναι ίση με

$$P(N(t_1) = 1 | N(t) = 1) = \frac{P(N(t_1) = 1, N(t) = 1)}{P(N(t) = 1)} = \frac{P(N(t_1) = 1) P(N(t) - N(t_1) = 0)}{P(N(t) = 1)}$$
$$= \frac{e^{-\lambda t_1} \lambda t_1 e^{-\lambda (t - t_1)}}{e^{-\lambda t} \lambda t} = \frac{t_1}{t}$$

Δηλαδή ο χρόνος t_1 κατανέμεται ομοιόμορφα στο $(0, t)$.

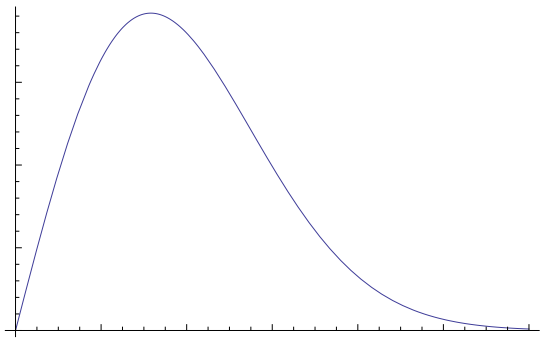
Επομένως, προσομοιώνουμε την ΤΜ $N(t) \sim \text{Poisson}(\lambda t)$ για δεδομένο t και λ, και αν αυτή πάρει την τιμή k, παίρνουμε k δείγματα $U_1, U_2, \ldots, U_k$ από την $U((0,1))$, τα ταξινομούμε και δημιουργούμε την ακολουθία χρονικών στιγμών των αφίξεων $(tU_1, tU_2, \ldots, tU_k)$, η οποία προσομοιώνει τη λειτουργία του συστήματος κατά το χρονικό διάστημα $(0, t)$.

7.5 Λυμένες ασκήσεις

7.6 Ασκήσεις για επίλυση

1. Έστω X μια συνεχής τυχαία μεταβλητή που ακολουθεί την κατανομή Rayleigh με παράμετρο $k > 0$ έχει συνάρτηση πυκνότητας πιθανότητας

$$f(x) = \begin{cases} kxe^{-\frac{kx^2}{2}}, & x \geq 0 \\ 0, & \text{αλλιώς} \end{cases}$$



   i) Να δειχθεί ότι η συνάρτηση κατανομής της X ισούται με $F(x) = 1 - e^{-\frac{kx^2}{2}}$, $x \geq 0$.

   ii) Να βρεθεί με την μέθοδο του αντίστροφου μετασχηματισμού πως μπορούμε να παράγουμε τυχαία δείγματα της μεταβλητής X αν μπορούμε να παράγουμε τυχαία δείγματα της μεταβλητής U που ακολουθεί την ομοιόμορφη κατανομή στο $[0, 1]$.

   iii) Να παραχθούν 30 τυχαία δείγματα της ΤΜ X όταν $k = 1/10$.

2. Να βρεθεί πως μπορούμε να λάβουμε δείγματα από την τυχαία μεταβλητή X με συνάρτηση πυκνότητας πιθανότητας

$$f(x) = \frac{1}{4}\left(x^{-\frac{1}{2}} + (1-x)^{-\frac{1}{2}} \right), 0 < x < 1.$$

3. Ένα σημείο επιλέγεται στο εσωτερικό μιας μοναδιαίας σφαίρας. Να βρεθεί η πιθανότητα το σημείο να απέχει λιγότερο από 1/2 από το κέντρο της σφαίρας.

Μέρος II

Στατιστική

Κεφάλαιο 8

Εκτιμητική και διαστήματα εμπιστοσύνης

8.1 Δείγματα και στατιστικές συναρτήσεις

Συχνά καλούμαστε να βγάλουμε συμπεράσματα για κάποιο χαρακτηριστικό ενός πληθυσμού, βασιζόμενοι στις παρατηρήσεις που λαμβάνουμε από ένα (τυχαίο) δείγμα του πληθυσμού, το οποίο ορίζει κάποια ΤΜ X. Η ΤΜ X ακολουθεί κάποια (ίσως άγνωστη) κατανομή $F(x;\theta)$, όπου θ κάποια άγνωστη παράμετρος που συνήθως θέλουμε να εκτιμήσουμε (π.χ. η μέση τιμή ή η διακύμανση). Τυχαίο δείγμα μεγέθους n από την κατανομή F ονομάζεται μια n-άδα ανεξάρτητων ΤΜ $(X_1, X_2, \ldots, X_n)$ που ακολουθούν την κατανομή F (independent, identically distributed - iid), και γράφουμε $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F$. Στο εξής, όταν θα λέμε τυχαίο δείγμα, θα εννοούμε ότι $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} F$, εκτός αν αναφέρεται το αντίθετο.

Η ΤΜ $T(X_1, X_2, \ldots, X_n)$, όπου $T/\mathbb{R}^n$ συνάρτηση n μεταβλητών, ονομάζεται **στατιστική συνάρτηση** (ή απλά **στατιστική**), εφόσον δεν εξαρτάται από άλλες παραμέτρους, παρά μόνο από τις τιμές των $X_1, \ldots, X_n$, και έχει πεδίο ορισμού το $S_{X_1} \times S_{X_2} \times \cdots \times S_{X_n}$. Μερικές σημαντικές στατιστικές συναρτήσεις για ένα τυχαίο δείγμα $(X_1, X_2, \ldots, X_n)$ είναι:

Δειγματικός μέσος: $\overline{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$.

Δειγματικό ποσοστό: $\overline{P} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$. Αποτελεί ειδική περίπτωση δειγματικού μέσου, όταν οι X_i ακολουθούν κατανομή Bernoulli με κάποια παράμετρο p.

Δειγματική διακύμανση: $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \dfrac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right)$.

Η τελευταία ισότητα προκύπτει από την επόμενη ιδιότητα:

Λήμμα 8.1. *Για οποιουσδήποτε $x_1, x_2, \ldots, x_n \in \mathbb{R}$, $n \in \mathbb{N}^*$, ισχύουν οι σχέσεις*

$$\min_a \sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad \text{και} \qquad \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2.$$

όπου $\overline{x} = (x_1 + \cdots + x_n)/n$.

Απόδειξη.

$$\begin{aligned}\sum_{i=1}^{n}(x_i-a)^2 &= \sum_{i=1}^{n}(x_i-\overline{x}+\overline{x}-a)^2 = \sum_{i=1}^{n}(x_i-\overline{x})^2+\sum_{i=1}^{n}(\overline{x}-a)^2+2\sum_{i=1}^{n}(x_i-\overline{x})(\overline{x}-a)\\ &= \sum_{i=1}^{n}(x_i-\overline{x})^2+\sum_{i=1}^{n}(\overline{x}-a)^2+2\sum_{i=1}^{n}(x_i\overline{x}-ax_i-\overline{x}^2+a\overline{x})\\ &= \sum_{i=1}^{n}(x_i-\overline{x})^2+\sum_{i=1}^{n}(\overline{x}-a)^2+2n(\overline{x}^2-a\overline{x}-\overline{x}^2+a\overline{x})\\ &= \sum_{i=1}^{n}(x_i-\overline{x})^2+\sum_{i=1}^{n}(\overline{x}-a)^2\end{aligned}$$

Η τελευταία παράσταση προφανώς ελαχιστοποιείται όταν $a=\overline{x}$.

Η δεύτερη σχέση προκύπτει θέτοντας $a=0$ στην τελευταία. □

Οι στατιστικές συναρτήσεις είναι τυχαίες μεταβλητές που οι τιμές τους μεταβάλλονται από δείγμα σε δείγμα. Θεωρούμε ότι ακολουθούν κάποια κατανομή η οποία ονομάζεται **κατανομή δειγματοληψίας**. Στη συνέχεια θα δούμε τις κατανομές δειγματοληψίας για ορισμένες χρήσιμες στατιστικές συναρτήσεις.

8.2 Κατανομές δειγματοληψίας

8.2.1 Κατανομή δειγματικού μέσου

1 Αν $(X_1, X_2, \ldots, X_n)$ τυχαίο δείγμα ενός (άπειρου ή πεπερασμένου) πληθυσμού με μέσο μ και διακύμανση σ^2, όπου η δειγματοληψία γίνεται με επανατοποθέτηση ή γενικότερα όταν οι X_i θεωρούνται ανεξάρτητες, τότε

$$E(\overline{X})=\frac{1}{n}E(X_1+\cdots+X_n)=\mu \qquad \text{και} \qquad \mathrm{V}(\overline{X})=\frac{1}{n^2}\mathrm{V}(X_1+\cdots+X_n)\overset{\text{iid}}{=}\frac{\sigma^2}{n}$$

2 Αν $(X_1, X_2, \ldots, X_n)$ τυχαίο δείγμα από πληθυσμό μεγέθους N, με $0<n<N$, με μέσο μ και διακύμανση σ^2, όπου η δειγματοληψία γίνεται χωρίς επανατοποθέτηση ή γενικότερα όταν οι X_i θεωρούνται εξαρτημένες, τότε (βλ. λυμένη άσκηση 8.1)

$$E(\overline{X})=\mu \text{ και } \mathrm{V}(\overline{X})=\frac{\sigma^2}{n}\cdot\frac{N-n}{N-1}.$$

Παρατηρήστε ότι καθώς N τείνει στο $+\infty$ ο τύπος της διακύμανσης ταυτίζεται με αυτόν στην προηγούμενη περίπτωση.

3 Αν ένα τυχαίο δείγμα μεγέθους n προέρχεται από ένα πληθυσμό που ακολουθεί την $N(\mu,\sigma^2)$, τότε ο δειγματικός μέσος $\overline{X}$ ακολουθεί την $N(\mu,\frac{\sigma^2}{n})$, αλλιώς, σύμφωνα με το Κ.Ο.Θ., ισχύει η προσέγγιση $\overline{X}\to N(\mu,\frac{\sigma^2}{n})$.

4 Αν οι μεταβλητές X_i ακολουθούν κατανομή Bern(p), με άγνωστη παράμετρο p, μετρώντας μια ιδιότητα που τα άτομα του πληθυσμού έχουν σε ποσοστό p, τότε $X_1+X_2+\cdots+X_n \sim \text{Binom}(n,p)$, οπότε

$$E(\overline{P}) = \frac{np}{n} = p \qquad \text{και} \qquad \text{V}(\overline{P}) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

5 Αν $(X_1, X_2, \ldots, X_{n_1})$ και $(Y_1, Y_2, \ldots, Y_{n_2})$ δύο ανεξάρτητα τυχαία δείγματα δύο άπειρων πληθυσμών με μέσους μ_1, μ_2 και διακυμάνσεις σ_1, σ_2 αντίστοιχα, τότε

$$E(\overline{X} - \overline{Y}) = \mu_1 - \mu_2 \text{ και } \text{V}(\overline{X} - \overline{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Ειδικά Αν $X_i \sim N(\mu_1, \sigma_1^2)$ και $Y_i \sim N(\mu_2, \sigma_2^2)$, τότε

$$\overline{X} - \overline{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}).$$

όπου $\overline{X}, \overline{Y}$ είναι δειγματικοί μέσοι των δύο δειγμάτων αντίστοιχα.

8.2.2 Κατανομή χ^2

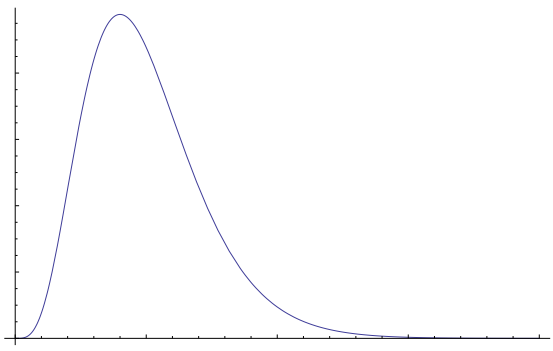
Αν οι ΤΜ X_1, X_2, …, X_ν είναι ανεξάρτητες και ακολουθούν την τυποποιημένη κανονική κατανομή $N(0,1)$ τότε η τυχαία μεταβλητή

$$Y = X_1^2 + X_2^2 + \cdots + X_\nu^2$$

ακολουθεί την λεγόμενη **κατανομή χι τετράγωνο με ν βαθμούς ελευθερίας**, η οποία συμβολίζεται με χ^2_ν και έχει συνάρτηση πυκνότητας πιθανότητας

$$f(x) = \begin{cases} \dfrac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2} & x > 0 \\ 0 & x < 0 \end{cases}$$

όπου $\Gamma(x)$ είναι η συνάρτηση Γάμμα.



Σχήμα 8.1: Η συνάρτηση πυκνότητας πιθανότητας της χ^2_ν για $\nu = 10$

Η κατανομή χ^2_ν είναι ειδική περίπτωση της κατανομής Γάμμα $\Gamma(a,b)$ με παραμέτρους $a = \nu/2$ και $\beta = 1/2$. Αποδεικνύεται εύκολα ότι αν $Y \sim \chi^2_\nu$ τότε

$$E(Y) = \nu \text{ και } \text{V}(Y) = 2\nu.$$

Για την χ^2_ν κατανομή υπάρχουν πίνακες όπου για δοσμένη πιθανότητα p και δοσμένους βαθμούς ελευθερίας $\nu \leq 30$ δίνουν την τιμή $\chi^2_{\nu,p}$, για την οποία $P(Y > \chi^2_{\nu,p}) = p$. Για παράδειγμα, για $p = 0.05$, $\nu = 20$ είναι $\chi^2_{20,0.05} = 31.41$ το οποίο σημαίνει ότι για 20 βαθμούς ελευθερίας

$$P(\chi^2_{20} > \chi^2_{20,0.05}) = P(\chi^2_{20} > 31.41) = 0.05.$$

1 Αν η ΤΜ X ακολουθεί την τυποποιημένη κανονική κατανομή $N(0,1)$ τότε η ΤΜ X^2 ακολουθεί την κατανομή χ^2_1 με ένα βαθμό ελευθερίας.

2 Αν $X_1, X_2, \ldots, X_n$ είναι ανεξάρτητες ΤΜ που ακολουθούν την χ^2 κατανομή με $\nu_1, \nu_2, \ldots, \nu_n$ βαθμούς ελευθερίας αντίστοιχα, τότε η ΤΜ

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

ακολουθεί την κατανομή χ^2_m, όπου $m = \nu_1 + \nu_2 + \cdots + \nu_n$.

3 Αν ο πληθυσμός ακολουθεί κανονική κατανομή $N(\mu, \sigma^2)$ και n το μέγεθος του δείγματος, τότε η στατιστική συνάρτηση

$$\frac{(n-1)S^2}{\sigma^2}$$

ακολουθεί την κατανομή χ^2_{n-1} και επιπλέον οι ΤΜ $\overline{X}$ και $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ είναι ανεξάρτητες.

8.2.3 Κατανομή t του Student

Αν X και Y είναι δύο ανεξάρτητες ΤΜ με $X \sim N(0,1)$ και $Y \sim \chi^2_\nu$ τότε λέμε ότι η μεταβλητή

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

ακολουθεί την **κατανομή t με ν βαθμούς ελευθερίας**, ή συμβολικά $T \sim t_\nu$, η οποία έχει συνάρτηση πυκνότητας πιθανότητας

$$f(x) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}}\left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}, \quad x \in \mathbb{R}.$$

Όταν $\nu \to \infty$ τότε η T προσεγγίζει την $N(0,1)$ (Πρακτικά, αν $n > 30$).

Αποδεικνύεται ότι αν $T \sim t_\nu$ τότε

$$E(T) = 0 \text{ και } \mathrm{V}(T) = \frac{\nu}{\nu - 2}.$$

Υπάρχουν πίνακες που για δοσμένη δοσμένη πιθανότητα p και δοσμένους βαθμούς ελευθερίας ν δίνουν την τιμή $t_{\nu,p}$, όπου $P(t > t_{\nu,p}) = p$.

Για παράδειγμα, για $p = 0.05$, $\nu = 30$ είναι $t_{30,0.05} = 1.645$ το οποίο σημαίνει ότι για 30 βαθμούς ελευθερίας

$$P(t_\nu > t_{\nu,p}) = p \Leftrightarrow P(t_{30} > 1.645) = 0.05.$$

Η t κατανομή είναι συμμετρική και ισχύει ότι $F(t) + F(-t) = 1$ καθώς επίσης $t_{\nu,1-p} = -t_{\nu,p}$.

1 Αν $\overline{X}$ και S^2 είναι ο δειγματικός μέσος και η δειγματική διακύμανση ενός τυχαίου δείγματος μεγέθους n που έχει ληφθεί από κανονικό πληθυσμό με μέσο μ διακύμανση σ^2 τότε η στατιστική συνάρτηση

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

ακολουθεί την κατανομή t_{n-1}.

Απόδειξη. Επειδή η ΤΜ $\overline{X}$ ακολουθεί την κανονική κατανομή $N(\mu, \frac{\sigma^2}{n})$ έπεται ότι η ΤΜ $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ ακολουθεί την τυποποιημένη κανονική κατανομή $N(0,1)$.

Επίσης, ισχύει ότι η ΤΜ $Y = \frac{(n-1)S^2}{\sigma^2}$ ακολουθεί την κατανομή χ^2_{n-1}. Επιπλέον, επειδή οι ΤΜ $\overline{X}$ και S^2 είναι ανεξάρτητες, έπεται ότι και οι ΤΜ Z, Y είναι επίσης ανεξάρτητες.

Επομένως, η ΤΜ

$$T = \frac{Z}{\sqrt{Y/(n-1)}} = \frac{\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

ακολουθεί την κατανομή t με $n-1$ βαθμούς ελευθερίας. □

Παρατήρηση. Η t κατανομή είναι κατάλληλη για να περιγράψει τυχαίες μεταβλητές της μορφής $t = \frac{\theta - \mu_\theta}{\sigma_\theta}$ όπου θ είναι μια ΤΜ που ακολουθεί την κανονική κατανομή $N(\mu_\theta, \sigma_\theta)$, ενώ σ_θ είναι η δειγματική τυπική απόκλιση της θ και οι τ.μ θ και σ_θ είναι στατιστικά ανεξάρτητες.

8.2.4 Κατανομή F του Fisher

Αν X, Y είναι ανεξάρτητες ΤΜ με $X \sim \chi^2_{\nu_1}$ και $Y \sim \chi^2_{\nu_2}$ τότε η ΤΜ

$$F = \frac{X/\nu_1}{Y/\nu_2}$$

ακολουθεί την **κατανομή F με ν_1, ν_2 βαθμούς ελευθερίας**, συμβολικά $F \sim F_{\nu_1,\nu_2}$, η οποία έχει συνάρτηση πυκνότητας πιθανότητας

$$f(x) = \frac{\Gamma((\nu_1+\nu_2)/2)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \nu_1^{\nu_1/2} \nu_2^{\nu_2/2} x^{\nu_1/2-1} (\nu_1 x + \nu_2)^{-(\nu_1+\nu_2)/2}.$$

Αποδεικνύεται ότι αν $F \sim F_{\nu_1,\nu_2}$ τότε

$$E(F) = \frac{\nu_2}{\nu_2 - 2}, \text{ αν } \nu_2 > 2 \qquad \text{και} \qquad \mathrm{V}(F) = \frac{2\nu_2^2(\nu_1+\nu_2-4)}{\nu_1(\nu_2-2)^2(\nu_2-4)}, \text{ αν } \nu_2 > 4$$

Υπάρχουν πίνακες που για δοσμένη δοσμένη πιθανότητα p και δοσμένους βαθμούς ελευθερίας ν_1, ν_2 δίνουν την τιμή $F_{\nu_1,\nu_2,p}$, όπου $P(X > F_{\nu_1,\nu_2,p}) = p$.

Για παράδειγμα, για $p = 0.05$, $\nu_1 = 20$, $\nu_2 = 21$ είναι $F_{20,21,0.05} = 2.09$ το οποίο σημαίνει ότι

$$P(F_{\nu_1,\nu_2} > F_{\nu_1,\nu_2}) = 0.05 \Leftrightarrow P(F_{20,21} > 2.09) = 0.05.$$

Οι τιμές $F_{\nu_1,\nu_2,p}$ και $F_{\nu_2,\nu_1,1-p}$ είναι αντίστροφοι αριθμοί.

Αν $X \sim F_{\nu_1,\nu_2}$ τότε η ΤΜ $Y = \frac{1}{X} \sim F_{\nu_2,\nu_1}$.

1 Αν S_1^2 και S_2^2 είναι οι δειγματικές διακυμάνσεις δύο τυχαίων δειγμάτων μεγέθους ν_1 και ν_2 αντίστοιχα που έχουν ληφθεί από κανονικούς πληθυσμούς με διακυμάνσεις σ_1^2 και σ_2^2 αντίστοιχα, τότε η στατιστική συνάρτηση

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

ακολουθεί την κατανομή F_{ν_1-1,ν_2-1}.

Απόδειξη. Γνωρίζουμε ότι $X = \dfrac{(\nu_1-1)S_1^2}{\sigma_1^2} \sim \chi^2_{\nu_1-1}$ και $Y = \dfrac{(\nu_2-1)S_2^2}{\sigma_2^2} \sim \chi^2_{\nu_2-1}$. Επειδή οι ΤΜ S_1^2 και S_2^2 είναι ανεξάρτητες, έπεται ότι και οι ΤΜ X, Y είναι επίσης ανεξάρτητες. Επομένως, η ΤΜ

$$F = \frac{X/(\nu_1-1)}{Y/(\nu_2-1)} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

ακολουθεί την κατανομή F_{ν_1-1,ν_2-1}. □

Παρατήρηση. Η ιδιότητα αυτή είναι πολύ χρήσιμη όταν $\sigma_1 = \sigma_2$.

Οι επόμενες προτάσεις συνοψίζουν τα προηγούμενα:

Πρόταση 8.2. *Αν $X_1, X_2, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, τότε*

i) $\overline{X} \sim \mathcal{N}(\mu, \dfrac{\sigma^2}{n})$ *και* $Z = \dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ *ii)* $\dfrac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1} = \Gamma(\dfrac{n-1}{2}, \dfrac{1}{2})$

iii) Οι ΤΜ $\overline{X}$ και S^2 είναι ανεξάρτητες. *iv)* $T = \dfrac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$.

Σύμφωνα με το Κεντρικό Οριακό Θεώρημα (Κ.Ο.Θ.), ανεξάρτητα από το ποια είναι η F, όταν το n είναι πολύ μεγάλο, έχουμε ότι

$$\overline{X} \to \mathcal{N}(\mu, \sigma^2/n), \quad Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \to \mathcal{N}(0,1), \quad T = \frac{\overline{X}-\mu}{S/\sqrt{n}} \to \mathcal{N}(0,1), \quad \overline{P} \to \mathcal{N}(p, p(1-p)/n),$$

καθώς $n \to \infty$.

Πρόταση 8.3. *Αν $X_1, X_2, \ldots, X_{n_1} \sim \mathcal{N}(\mu_1, \sigma_1^2)$ και $Y_1, Y_2, \ldots, Y_{n_2} \sim \mathcal{N}(\mu_2, \sigma_2^2)$, τότε*

i) $Z = \dfrac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim \mathcal{N}(0,1)$ *ii)* $\dfrac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$

iii) Αν $\sigma_1 = \sigma_2$, τότε $T_{n_1+n_2-2} = \dfrac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{S_p\sqrt{1/n_1+1/n_2}} \sim t_{n_1+n_2-2}$, *όπου* $S_p^2 = \dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$

8.3 Σημειακές εκτιμήτριες

Στην ενότητα αυτή θα δούμε τεχνικές για την εκτίμηση μιας άγνωστης παραμέτρου θ της κατανομής $F(x;\theta)$ ενός πληθυσμού, με βάση τις τιμές ενός δείγματος $(X_1, X_2, \ldots, X_n)$. Μια στατιστική συνάρτηση $T = T(X_1, X_2, \ldots, X_n)$ ονομάζεται (σημειακή) **εκτιμήτρια της παραμέτρου** θ, αν δεν εξαρτάται από την θ. Η ποσότητα

$$\text{bias}(T) = E(T) - \theta$$

καλείται **μεροληψία** της εκτιμήτριας T. Η ποσότητα

$$\text{mse(T)} = E((T-\theta)^2) = V(T) + \text{bias}^2(T)$$

καλείται **μέσο τετραγωνικό σφάλμα** της εκτιμήτριας T από την θ.

Υπάρχουν 4 βασικά χαρακτηριστικά που επιθυμούμε να έχει μια εκτιμήτρια:

Αμεροληψία: Μια εκτιμήτρια T μιας παραμέτρου θ ονομάζεται **αμερόληπτη** ανν $E(T) = \theta$.

Αποτελεσματικότητα: Η εκτιμήτρια T_1 είναι αποτελεσματικότερη από την T_2, ανν $\text{mse}(\text{T}_1) < \text{mse}(\text{T}_2)$. Ειδικότερα, αν είναι αμερόληπτες, τότε $\text{V}(T_1) < \text{V}(T_2)$.

Συνέπεια: Η ακολουθία εκτιμητριών (T_n) είναι **συνεπής**, αν τείνει κατά πιθανότητα στην θ (γράφουμε $T_n \xrightarrow{p} \theta$), δηλαδή αν $\lim_{n\to\infty} P(|T_n - \theta| \geq \epsilon) = 0$, για κάθε $\epsilon > 0$.

Ένα απλό κριτήριο για τη συνέπεια είναι το ακόλουθο:

Πρόταση 8.4. *Αν για την ακολουθία εκτιμητριών (T_n) ισχύουν $\lim_{n\to\infty} E(T_n) = \theta$ και $\lim_{n\to\infty} \text{V}(T_n) = 0$, τότε η (T_n) είναι συνεπής.*

Απόδειξη. Έστω $\epsilon > 0$. Αφού $\lim_{n\to\infty} E(T_n) = \theta$, έπεται ότι υπάρχει $n_0 \in \mathbb{N}^*$ τέτοιο ώστε $n \geq n_0 \Rightarrow |E(T_n) - \theta| < \epsilon \Rightarrow \epsilon - |E(T_n) - \theta| > 0$. Επομένως, για $n \geq n_0$, βάσει της ανισότητας Chebyshev, είναι

$$\begin{aligned} P(|T_n - \theta| \geq \epsilon) &= P(|T_n - E(T_n) + (E(T_n) - \theta)| \geq \epsilon) \leq P(|T_n - E(T_n)| + |E(T_n) - \theta| \geq \epsilon) \\ &= P(|T_n - E(T_n)| \geq \epsilon - |E(T_n) - \theta|) \leq \frac{\text{V}(T_n)}{(\epsilon - |E(T_n) - \theta|)^2} \end{aligned}$$

και το δεύτερο μέλος τείνει στο 0 καθώς $n \to \infty$. □

Επάρκεια: Μια εκτιμήτρια T της παραμέτρου θ ονομάζεται **επαρκής** αν χρησιμοποιεί όλες τις πληροφορίες του δείγματος (η τιμή της εξαρτάται από όλες τις ΤΜ $X_1, \ldots, X_n$).

Βάσει της επόμενης πρότασης, μπορούμε να κατασκευάσουμε εκτιμήτριες για τη μετασχηματισμένη παράμετρο $g(\theta)$:

Πρόταση 8.5. *Αν για τις ακολουθίες εκτιμητριών $(T_n), (W_n)$ ισχύουν $T_n \xrightarrow{p} \theta_1$ και $W_n \xrightarrow{p} \theta_2$, και g συνεχής συνάρτηση στο θ_1, τότε*

$$g(T_n) \xrightarrow{p} g(\theta_1), \quad T_n + W_n \xrightarrow{p} \theta_1 + \theta_2, \quad T_n - W_n \xrightarrow{p} \theta_1 - \theta_2, \quad T_n W_n \xrightarrow{p} \theta_1\theta_2, \quad T_n/W_n \xrightarrow{p} \theta_1/\theta_2.$$

Παράδειγμα 8.3.1. *Έστω ένας πληθυσμός με μέση τιμή μ και διακύμανση σ^2. Να δειχθεί ότι:*

i) *Ο δειγματικός μέσος $\overline{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$ είναι αμερόληπτη και συνεπής εκτιμήτρια της μέσης τιμής μ.*

ii) *Η στατιστική συνάρτηση $T = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ είναι αμερόληπτη εκτιμήτρια για την διακύμανση σ^2.*

iii) *Η στατιστική συνάρτηση $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ είναι αμερόληπτη εκτιμήτρια για την διακύμανση σ^2.*

iv) *Η στατιστική συνάρτηση $T = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ δεν είναι αμερόληπτη εκτιμήτρια για την διακύμανση σ^2.*

Λύση. i) Έχει ήδη δειχθεί ότι $E(\overline{X}) = \mu$ και $\mathrm{V}(\overline{X}) = \dfrac{\sigma^2}{n}$, άρα η $\overline{X}$ είναι αμερόληπτη και συνεπής.

ii) $$E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\right) = \frac{1}{n}\sum_{i=1}^{n}E((X_i-\mu)^2) = \frac{1}{n}\sum_{i=1}^{n}\sigma^2 = \sigma^2.$$

iii)

$$\begin{aligned} E((n-1)S^2) &= E\left(\sum_{i=1}^{n}X_i^2 - n\overline{X}^2\right) = \sum_{i=1}^{n}E(X_i^2) - nE(\overline{X}^2) = \sum_{i=1}^{n}(\sigma^2+\mu^2) - n(\mathrm{V}(\overline{X}) + \mu^2) \\ &= n\sigma^2 - \frac{n}{n^2}n\sigma^2 = (n-1)\sigma^2 \end{aligned}$$

άρα $E(S^2) = \sigma^2$.

iv) $E(nT/(n-1)) = E(S^2) = \sigma^2$, άρα $E(T) = \dfrac{n-1}{n}\sigma^2 \neq \sigma^2$.

□

8.3.1 Μέθοδος μέγιστης πιθανοφάνειας

Η πιο γνωστή μέθοδος εύρεσης εκτιμητριών είναι η μέθοδος μέγιστης πιθανοφάνειας. Έστω $(X_1, X_2, \ldots, X_n)$ ένα τυχαίο δείγμα ενός πληθυσμού με PDF (ή PMF) $f(x;\theta)$, όπου θ είναι μια άγνωστη παράμετρος που θέλουμε να εκτιμήσουμε. Τότε επειδή οι $X_1, X_2, \ldots, X_n$ είναι ανεξάρτητες τυχαίες μεταβλητές από την ίδια κατανομή, η από κοινού PDF (ή PMF) είναι η

$$f(x_1, x_2, \ldots, x_n; \theta) = f(x_1;\theta)f(x_2;\theta)\cdots f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

Αν γνωρίζουμε τα $x_1, x_2, \ldots, x_n$ (τις τιμές του δείγματος), τότε το γινόμενο αυτό είναι μια συνάρτηση του θ που συμβολίζεται με

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

και ονομάζεται **συνάρτηση πιθανοφάνειας** για το δείγμα $(X_1, X_2, \ldots, X_n)$.

Αν η $L(\theta)$ έχει μέγιστο για κάποιο $\theta = \bar{\theta}$, ή γενικότερα

$$L(\bar{\theta}) = \sup_{\theta} L(\theta),$$

τότε το $\bar{\theta}$ αυτό είναι προφανώς συνάρτηση των $(X_1, X_2, \ldots, X_n)$, άρα είναι μια εκτιμήτρια του θ, και ονομάζεται εκτιμήτρια μέγιστης πιθανοφάνειας (Maximum Likelihood Estimator - MLE).
Παρατήρηση. Επειδή η $L(\theta)$ είναι γινόμενο μη αρνητικών όρων, αντί να μεγιστοποίησουμε την $L(\theta)$ συνήθως μεγιστοποιούμε την $\ell(\theta) := \ln L(\theta)$.

Παράδειγμα 8.3.2. *Έστω $X_1, X_2, \ldots, X_n$ τυχαίο δείγμα από πληθυσμό που ακολουθεί την κατανομή Bernoulli με παράμετρο p. Να βρεθεί εκτιμήτρια για το p με την μέθοδο της μέγιστης πιθανοφάνειας.*

Λύση. Για κάθε $i \in [n]$, η συνάρτηση πιθανότητας της X_i είναι η $f(x_i; p) = p^{x_i}(1-p)^{1-x_i}$, $x_i \in \{0, 1\}$. Επομένως,

$$L(p) = \prod_{i=1}^{n} f(x_i; p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

και, θέτοντας $a = \sum_{i=1}^{n} x_i$,

$$\ell(p) = \ln L(p) = \left(\sum_{i=1}^{n} x_i\right) \ln p + \left(n - \sum_{i=1}^{n} x_i\right) \ln(1-p) = a \ln p + (n-a)\ln(1-p)$$

Παραγωγίζοντας ως προς p,

$$\ell'(p) = \frac{a}{p} - \frac{n-a}{1-p} = \frac{a(1-p) - (n-a)p}{p(1-p)} = \frac{a - np}{p(1-p)} \geq 0 \Leftrightarrow p \leq \frac{a}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Επομένως, η $L(p)$ μεγιστοποιείται όταν $p = \frac{1}{n}\sum_{i=1}^{n} x_i$ και η MLE για την παράμετρο p είναι το δειγματικό ποσοστό $\overline{P} = \frac{1}{n}\sum_{i=1}^{n} X_i$. □

Παράδειγμα 8.3.3. *Έστω $X_1, X_2, \ldots, X_n$ τυχαίο δείγμα από πληθυσμό που ακολουθεί την κατανομή Poisson με παράμετρο λ. Να βρεθεί εκτιμήτρια για το λ με την μέθοδο της μέγιστης πιθανοφάνειας.*

Λύση. Ομοίως με το προηγούμενο παράδειγμα,

$$L(\lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda}\frac{\lambda^{x_1+\cdots+x_n}}{x_1!\cdots x_n!}$$

$$\ell(\lambda) = -n\lambda + (x_1 + \cdots + x_n)\ln\lambda - c, \qquad c = \ln(x_1!\cdots x_n!)$$

Θέτοντας $a = x_1 + \cdots + x_n$ και παραγωγίζοντας ως προς λ,

$$\ell'(\lambda) = -n + \frac{a}{\lambda} \geq 0 \Leftrightarrow \lambda \leq \frac{a}{n} = \frac{x_1 + \cdots + x_n}{n}.$$

Επομένως, η $L(\lambda)$ μεγιστοποιείται όταν $\lambda = \frac{1}{n}\sum_{i=1}^{n} x_i$ και η MLE για την παράμετρο λ είναι ο δειγματικός μέσος $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. □

Παράδειγμα 8.3.4. *Έστω $X_1, X_2, \ldots, X_n$ τυχαίο δείγμα από πληθυσμό που ακολουθεί την εκθετική κατανομή με παράμετρο θ. Να βρεθεί εκτιμήτρια για το θ με την μέθοδο της μέγιστης πιθανοφάνειας.*

Λύση. □

Παράδειγμα 8.3.5. *Έστω $X_1, X_2, \ldots, X_n$ τυχαίο δείγμα από πληθυσμό που ακολουθεί την κανονική κατανομή $N(\mu, \sigma^2)$. Να βρεθούν εκτιμήτριες για τις μ και σ^2 με την μέθοδο της μέγιστης πιθανοφάνειας.*

Λύση.

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(\frac{-\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right)$$

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

Παραγωγίζοντας ως προς μ,

$$\frac{\partial \ell}{\partial \mu} = \frac{-1}{2\sigma^2}\sum_{i=1}^{n}\frac{\partial}{\partial \mu}(x_i-\mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - n\mu\right) = 0 \Leftrightarrow \mu = \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Παραγωγίζοντας ως προς σ^2 και θέτοντας $a = \sum_{i=1}^{n}(x_i-\mu)^2$,

$$\frac{\partial \ell}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{a}{2(\sigma^2)^2} = \frac{a - n\sigma^2}{2\sigma^4} = 0 \Leftrightarrow \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

Άρα, οι δύο μερικές παράγωγοι γίνονται ίσες με 0 στο σημείο $(\hat{\mu}, \hat{\sigma}^2)$, όπου $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2$, στο οποίο η L μεγιστοποιείται και οι εκτιμήτριες μέγιστης πιθανοφάνειας είναι αντίστοιχα η $\overline{X}$ για το μ και η $\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$ για το σ^2. □

8.4 Διαστήματα εμπιστοσύνης

Είναι σχεδόν βέβαιο ότι η τιμή μιας εκτιμήτριας T μιας παραμέτρου θ δεν ταυτίζεται ακριβώς με την πραγματική τιμή της παραμέτρου θ. Επομένως, είναι πιο χρήσιμο, αντί μιας συγκεκριμένης τιμής, να δοθεί ένα διάστημα εντός του οποίου βρίσκεται η πραγματική τιμή της παραμέτρου θ, με κάποια πιθανότητα. Το διάστημα αυτό ονομάζεται **διάστημα εμπιστοσύνης** (δ.ε.) για την παράμετρο θ. Συγκεκριμένα, αν

$$P(\theta_1 \le \theta \le \theta_2) = 1 - a$$

τότε το διάστημα $[\theta_1, \theta_2]$ ονομάζεται $(1-a)100\%$ **διάστημα εμπιστοσύνης για την παράμετρο** θ.

Η πιθανότητα $1-a$ ονομάζεται **επίπεδο εμπιστοσύνης** και η πιθανότητα a ονομάζεται **επίπεδο σημαντικότητας**.

Συνήθως αναζητούμε διαστήματα εμπιστοσύνης τα οποία είναι συμμετρικά ως προς την σημειακή εκτίμηση T της θ, δηλαδή είναι της μορφής $[T-\epsilon, T+\epsilon]$ (διότι αποδεικνύεται ότι τα διαστήματα αυτά έχουν ελάχιστο μήκος).

Στον προσδιορισμό διαστημάτων εμπιστοσύνης κεντρικό ρόλο παίζει η έννοια του ποσοστιαίου σημείου μια κατανομής:

Ορισμός. *Το a-άνω ποσοστιαίο σημείο, $a \in (0,1)$, μιας ΤΜ X που ακολουθεί κάποια κατανομή F συμβολίζεται με x_a και ορίζεται από τις ισοδύναμες ισότητες:*

$$P(X > x_a) = a \Leftrightarrow 1 - P(X \le x_a) = a \Leftrightarrow 1 - F(x_a) = a \Leftrightarrow F(x_a) = 1-a \Leftrightarrow x_a = F^{-1}(1-a)$$

Γενικά, αν η F δεν είναι αντιστρέψιμη, τότε $x_a = \inf\{x \in S_X : F(x) \ge 1-a\}$.

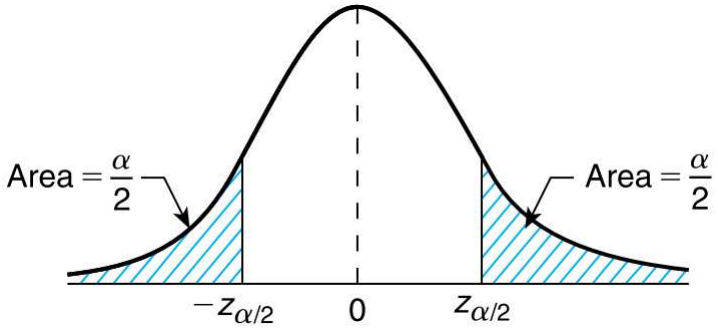
Παρατήρηση: Ορισμένοι συγγραφείς ορίζουν το x_a βάσει των σχέσεων $P(X \ge x_a) = a$ και $P(X \le x_a) = 1-a$. Ο ορισμός αυτός ταυτίζεται με τον προηγούμενο όταν η ΤΜ X είναι συνεχής. Όταν είναι διακριτή, τότε μπορεί το x_a να μην ανήκει στο S_X. Για παράδειγμα, αν η X είναι διακριτή ομοιόμορφη στο $S_X = \{1,2,3,4\}$ και $a = 0.5$, τότε με τον πρώτο ορισμό είναι $x_a = 2$, ενώ με τον δεύτερο μπορεί να είναι οποιοσδήποτε $x_a \in (2,3)$ και συνήθως θεωρείται ο $x_a = 2.5$.

Για την τυπική κανονική κατανομή, το a-άνω ποσοστιαίο σημείο συμβολίζεται με z_a, δηλαδή

$$z_a := \Phi^{-1}(1-a), \quad a \in (0,1).$$

Συνηθισμένες τιμές:

a	0.0005	0.001	0.005	0.01	0.025	0.05	0.10
z_a	3.29	3.09	2.576	2.326	1.960	1.645	1.282



8.4.1 Δ.Ε. για τον πληθυσμιακό μέσο

Πρόταση 8.6. *Αν $\overline{X}$ είναι ο δειγματικός μέσος ενός τυχαίου δείγματος μεγέθους n από ένα πληθυσμό που ακολουθεί την κανονική κατανομή $N(\mu, \sigma^2)$, όπου σ^2 γνωστό, τότε ένα $(1-a)\cdot 100\%$ διάστημα εμπιστοσύνης για τον μέσο όρο μ είναι το*

$$[\overline{X} \pm B] := \left[\overline{X} - B, \overline{x} + B\right], \qquad \text{όπου } B = z_{a/2}\frac{\sigma}{\sqrt{n}}.$$

Απόδειξη. Ως γνωστό, η ΤΜ $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ ακολουθεί τυπική κανονική κατανομή $N(0,1)$, οπότε

$$P(|Z| \le z) \ge 1 - a \Leftrightarrow \Phi(z) - \Phi(-z) \ge 1 - a \Leftrightarrow 2\Phi(z) - 1 \ge 1 - a \Leftrightarrow \Phi(z) \ge \frac{2-a}{2} = 1 - a/2 \Leftrightarrow z \ge z_{a/2}$$

Επομένως,

$$1 - a = P(|Z| \le z_{a/2}) = P(|\overline{X} - \mu| \le B) = P(\mu \in [\overline{X} \pm B]) \qquad \square$$

Παρατήρηση: Η τελευταία πιθανότητα $P(\mu \in [\overline{X} \pm B]) = P(\overline{X} \in [\mu \pm B]) = 1 - a$ που υπολογίσθηκε στην παραπάνω απόδειξη είναι η πιθανότητα το διάστημα $[\overline{X} \pm B]$ που κατασκευάσαμε βάσει της παρατήρησης να περιέχει το μ. Δηλαδή, η ΤΜ είναι το διάστημα και όχι ο μ, ο οποίος θεωρείται σταθερός, αν και άγνωστος.

Από την παράσταση $B = z_{a/2}\dfrac{\sigma}{\sqrt{n}}$ μπορούμε να υπολογίσουμε ένα εκ των a, B, n, αν δίνονται τα άλλα δύο. Για παράδειγμα, αν δίνονται τα a και B, τότε μπορούμε να υπολογίσουμε το απαιτούμενο μέγεθος δείγματος $n = z_{a/2}^2\dfrac{\sigma^2}{B^2}$ ώστε να επιτύχουμε το απαιτούμενο επίπεδο σημαντικότητας και πλάτος διαστήματος.

Παράδειγμα 8.4.1. *Έστω ένα τυχαίο δείγμα 4 μετρήσεων*

$$1.2, 3.4, 0.6, 5.6$$

από ένα πληθυσμό που ακολουθεί την κανονική κατανομή $N(\mu, 9)$. Να βρεθεί ένα 90% διάστημα εμπιστοσύνης για την μέση τιμή μ.

Λύση. Βάσει της εκφώνησης, είναι $n = 4$ και $a = 0.1$, οπότε $z_{a/2} = 1.645$

Μια σημειακή εκτίμηση για την μέση τιμή μ είναι η

$$\overline{X} = \frac{1.2 + 3.4 + 0.6 + 5.6}{4} = 2.7.$$

Βάσει της προηγούμενης πρότασης, είναι

$$B = z_{a/2}\frac{\sigma}{\sqrt{n}} = 1.645\frac{3}{2} = 2.4675$$

και το ζητούμενο διάστημα είναι το $[\overline{X} \pm B] = [0.2325, 5.1675]$. $\square$

Παράδειγμα 8.4.2. *Έστω ένα δείγμα μεγέθους $n = 100$ έχει δειγματικό μέσο $\overline{x} = 17$ και ο πληθυσμός απ' όπου επιλέξαμε το δείγμα ακολουθεί την κανονική κατανομή με διακύμανση $\sigma^2 = 9$. Να βρεθεί ένα διάστημα εμπιστοσύνης για την μέση τιμή μ του πληθυσμού με επίπεδο εμπιστοσύνης*

i) $1 - a = 0.95$

ii) $1 - a = 0.99$

Λύση. i) Για $a = 0.05$ είναι $z_{a/2} = 1.96$, οπότε $B = z_{a/2}\sigma/\sqrt{n} = 1.96 \cdot 3/10 = 0.588$, οπότε $[\overline{X} \pm B] = [17 \pm 0.588] = [16.412, 17.588]$.

ii) Για $a = 0.01$ είναι $z_{a/2} = 2.576$, οπότε $B = z_{a/2}\sigma/\sqrt{n} = 2.576 \cdot 3/10 = 0.7728$, οπότε $[\overline{X} \pm B] = [17 \pm 0.7728] = [16.2272, 17.7728]$. □

```
import numpy as np
import scipy.stats as st
xbar, a, sigma, N = 17, 0.01, 3, 100
se = sigma/np.sqrt(N) #standard error
c = st.norm.interval(alpha = 1-a, loc = xbar, scale = se)
print("CI (for a = %s):"%a, c)
```

Output:

```
CI (for a = 0.01): (16.22725120893533, 17.77274879106467)
```

Παράδειγμα 8.4.3. *Ένας πληθυσμός ακολουθεί την κανονική κατανομή με μέση τιμή μ και διακύμανση $\sigma^2 = 16$. Να βρεθεί το ελάχιστο μέγεθος n ενός τυχαίου δείγματος που θα μας επιτρέψει να εκτιμήσουμε την μέση τιμή μ με*

i) *επίπεδο εμπιστοσύνης* 0.95 *και σφάλμα* ± 1.

ii) *επίπεδο εμπιστοσύνης* 0.95 *και σφάλμα* ± 0.5.

iii) *επίπεδο εμπιστοσύνης* 0.99 *και σφάλμα* ± 1.

Λύση. i) Για $a = 0.05$ και $B = 1$, είναι $z_{a/2} = 1.96$ και

$$B = z_{a/2}\sigma/\sqrt{n} \leq 1 \Leftrightarrow n \geq z_{a/2}^2\sigma^2 = 1.96^2 \cdot 16 = 61.4656 \Leftrightarrow n \geq 62.$$

ii) Για $a = 0.05$ και $B = 0.5$, είναι $z_{a/2} = 1.96$ και

$$B = z_{a/2}\sigma/\sqrt{n} \leq 0.5 \Leftrightarrow n \geq \frac{z_{a/2}^2\sigma^2}{0.5^2} = \frac{1.96^2 \cdot 16}{0.25} = 4 \cdot 61.4656 = 245.8624 \Leftrightarrow n \geq 246.$$

iii) Για $a = 0.01$ και $B = 1$, είναι $z_{a/2} = 2.576$ και

$$B = z_{a/2}\sigma/\sqrt{n} \leq 1 \Leftrightarrow n \geq \frac{z_{a/2}^2\sigma^2}{B^2} = 2.576^2 \cdot 16 = 106.172416 \Leftrightarrow n \geq 107.$$

□

Παράδειγμα 8.4.4. *Σε ένα εργοστάσιο εμφιαλώσεως νερού παρατηρήθηκε ότι η ποσότητα X του νερού σε κάθε μπουκάλι ακολουθεί την κανονική κατανομή $N(\mu, 1.5^2)$. Ένας έλεγχος σε δείγμα $n = 25$ μπουκαλιών έδωσε $\overline{x} = 499.28$. Να βρεθεί ένα 95% διάστημα εμπιστοσύνης για το μ.*

Λύση. Για $a = 0.05$, είναι $z_{a/2} = 1.96$, οπότε

$$B = z_{a/2}\sigma/\sqrt{n} = 1.96 \cdot 1.5/5 = 0.588$$

και το ζητούμενο δ.ε. είναι το $[499.28 \pm 0.588]$.

```
import numpy as np
import scipy.stats as st
xbar, a, sigma, N = 499.28, 0.95, 1.5, 25
se = sigma/np.sqrt(N) #standard error
c = st.norm.interval(alpha = a, loc = xbar, scale = se)
print("CI (for a = %s):"%a, c)
```

Output:

```
CI (for a = 0.95): (498.69201080463796, 499.867989195362)
```

□

Πρόταση 8.7 (Διάστημα εμπιστοσύνης μέσης τιμής όταν η διακύμανση σ^2 είναι άγνωστη). *Έστω τυχαίο δείγμα $(X_1, X_2, \ldots, X_n)$ από κανονικό πληθυσμό $N(\mu, \sigma^2)$ με άγνωστή αλλά πεπερασμένη διακύμανση σ^2. Η στατιστική συνάρτηση*

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

ακολουθεί την κατανομή t_{n-1} και ένα $(1-a)\cdot 100\%$ διάστημα εμπιστοσύνης για την μέση τιμή μ είναι το

$$[\overline{X} \pm B],$$

όπου $B = t_{n-1,a/2}\dfrac{S}{\sqrt{n}}$ και $t_{n-1,a/2}$ το $a/2$-άνω ποσοστιαίο σημείο της κατανομής t με $n-1$ βαθμούς ελευθερίας.

Απόδειξη. Έχει ήδη λεχθεί ότι $T \sim t_{n-1}$ και ότι η συνάρτηση κατανομής F της T είναι συμμετρική, δηλαδή $F(t) + F(-t) = 1$. Επομένως, όπως και στην προηγούμενη απόδειξη, προκύπτει ότι

$$1 - a = P(|T| \le t_{n-1,a/2}) = P(|\overline{X} - \mu| \le B) = P(\mu \in [\overline{X} \pm B])$$ □

Παρατήρηση: Για μεγάλο n ($n \ge 30$), είναι $T \to N(0,1)$ και ένα προσεγγιστικό $(1-a)\cdot 100\%$ δ.ε. για τη μέση τιμή είναι το

$$[\overline{X} \pm B], \qquad \text{όπου } B = z_{a/2}\frac{S}{\sqrt{n}}.$$

Η προσέγγιση αυτή μπορεί να χρησιμοποιηθεί για οποιαδήποτε κατανομή δείγματος, όταν το n είναι μεγάλο.

Παράδειγμα 8.4.5. *Ένας δείκτης της κυκλοφορίας οχημάτων είναι ο αριθμός χιλιομέτρων που κάνει ένα όχημα το χρόνο. Σε μια περιοχή επιλέχθηκε ένα τυχαίο δείγμα 200 αυτοκινήτων και καταγράφηκε για κάθε αυτοκίνητο ο αριθμός των χιλιομέτρων που διένυσε τον τελευταίο χρόνο και βρέθηκε ότι* $\overline{X} = 14500$ *και* $S = 4000$.

i) *Να βρεθεί ένα 95% διάστημα εμπιστοσύνης για τον μέσο αριθμό χιλιομέτρων που διανύει το χρόνο ένα αυτοκίνητο.*

ii) *Να βρεθεί ένα 99% διάστημα εμπιστοσύνης για τον ίδιο αριθμό και να συγκριθούν τα αποτελέσματα.*

Λύση. Έχουμε ότι $n = 200$, $\overline{X} = 14500$, $S = 4000$ και η διασπορά είναι άγνωστη. Θεωρούμε τη στατιστική $T = \dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$.

i) Για $a = 0.05$ είναι $t_{n-1,a/2} = 1.972$, οπότε

$$B = t_{n-1,a/2} S/\sqrt{n} = \frac{1.972 \cdot 4000}{\sqrt{200}} = \frac{1.972 \cdot 400}{\sqrt{2}} = 557.76583$$

και το ζητούμενο 95% δ.ε. είναι το

$$[14500 \pm 557.76583]$$

Αν χρησιμοποιηθεί η προσέγγιση από κανονική κατανομή, τότε είναι

$$B' = z_{a/2} S/\sqrt{n} = \frac{1.96 \cdot 400}{\sqrt{2}} = 554.3717.$$

```
import numpy as np
import scipy.stats as st
a, n, xbar, s = 0.05, 200, 14500, 4000
df = n-1
q = st.t.ppf(1-a/2,df)
qz = st.norm.ppf(1-a/2)
print("confidence level = %s, sample mean = %s"%(1-a, xbar))
print("t_{n-1,a/2}=%s, P(T>=t_{n-1,a/2}) = %s"%(q, 1-st.t.cdf(q,df)))
print("z_{a/2}=%s, P(Z>=z_{a/2}) = %s"%(qz, 1-st.norm.cdf(qz)))
print("(1-a)100% CI (using t):", st.t.interval(alpha = 1-a, df = n-1, loc = xbar,
    scale = s/np.sqrt(n)))
print("(1-a)100% CI (using normal):", st.norm.interval(alpha = 1-a, loc = xbar, scale
    = s/np.sqrt(n)))
```

Output:

```
confidence level = 0.95, sample mean = 14500
t_{n-1,a/2}=1.971956544249395, P(T>=t_{n-1,a/2}) = 0.02500000000013569
z_{a/2}=1.959963984540054, P(Z>=z_{a/2}) = 0.025000000000000022
(1-a)100% CI (using t): (13942.246462142424, 15057.753537857576)
(1-a)100% CI (using normal): (13945.638470260128, 15054.361529739872)
```

ii) Για $a = 0.01$ είναι $t_{n-1,a/2} = 2.60076$, οπότε

$$B = t_{n-1,a/2} S/\sqrt{n} = \frac{2.60076 \cdot 4000}{\sqrt{200}} = \frac{2.60076 \cdot 400}{\sqrt{2}} = 735.606$$

και το ζητούμενο 99% δ.ε. είναι το

$$[14500 \pm 735.606]$$

□

Παράδειγμα 8.4.6. *Ένα κτηματολογικό γραφείο θέλει να εκτιμήσει την μέση τιμή των σπιτιών μιας περιοχής. Ένα δείγμα 25 σπιτιών έδωσε δειγματικό μέσο $\overline{X} = 50$ και διακύμανση $S^2 = 64$. Να βρεθεί ένα 90% διάστημα εμπιστοσύνης για το μ.*

Λύση. Για $a = 0.1$ και $n = 25$, είναι $t_{n-1,a/2} = 1.711$, οπότε $B = t_{n-1,a/2} S/\sqrt{n} = 1.711 \cdot 8/5 = 2.7376$, και το ζητούμενο δ.ε. είναι το $[\overline{X} \pm B] = [50 \pm 2.7376]$. □

8.4.2 Δ.Ε. για πληθυσμιακό ποσοστό

Πρόταση 8.8 (Διάστημα εμπιστοσύνης ποσοστού όταν το δείγμα είναι μεγάλο). *Έστω τυχαίο δείγμα $(X_1, X_2, \ldots, X_n)$ από τον πληθυσμό* Bernoulli(p). *Η στατιστική συνάρτηση*

$$Z = \frac{\overline{P} - p}{\sqrt{\overline{P}(1-\overline{P})/n}} \to N(0,1)$$

και ένα προσεγγιστικό $(1-a)\cdot 100\%$ διάστημα εμπιστοσύνης για το ποσοστό p είναι το

$$[\overline{P} \pm B], \qquad \text{όπου } B = z_{a/2}\sqrt{\overline{P}(1-\overline{P})/n}.$$

Απόδειξη. Ως γνωστό, $X = X_1 + \cdots + X_n \sim \text{Binom}(\text{np}, \text{np}(1-\text{p}))$, οπότε η $\overline{P} = X/n$ έχει μέση τιμή p και διακύμανση $\sigma^2 = p(1-p)/n$, οι οποίες είναι άγνωστες.

Θέτοντας $\overline{S}^2 = \overline{P}(1-\overline{P})/n$, ως εκτιμήτρια της σ^2, έχουμε ότι

$$\begin{aligned} E(\overline{S}^2) &= \frac{1}{n}E(\overline{P}(1-\overline{P})) = \frac{1}{n}(E(\overline{P}) - E(\overline{P}^2)) = \frac{1}{n}(p - \frac{p(1-p)}{n} - p^2) = \frac{1}{n^2}(np - p(1-p) - np^2) \\ &= \frac{n-1}{n^2}p(1-p) = \frac{n-1}{n}\sigma^2 \approx \sigma^2 \end{aligned}$$

Επειδή, βάσει του ΚΟΘ, είναι $\overline{P} \to N(p, \sigma^2)$ και $Z' = (\overline{P} - p)/\sigma \to N(0,1)$, η ζητούμενη ακτίνα του δ.ε. θα είναι $B \approx z_{a/2}\sigma \approx z_{a/2}\sqrt{\overline{P}(1-\overline{P})/n}$. □

Παράδειγμα 8.4.7. *Από τα προϊόντα μιας μηχανής λαμβάνεται τυχαίο δείγμα μεγέθους $n = 125$. Σε αυτά υπάρχουν 7 ελαττωματικά. Να βρεθεί ένα 98% διάστημα εμπιστοσύνης του ποσοστού των ελαττωματικών προϊόντων που παράγονται από την μηχανή.*

Λύση. Για $a = 0.02$ και $\overline{P} = 7/125 = 0.056$, είναι $z_{a/2} = 2.326$, οπότε $B = z_{a/2}\sqrt{\overline{P}(1-\overline{P})/n} = 0.047833$.

```
import numpy as np
import scipy.stats as st
a, p, n = 0.02, 7/125, 125
s = np.sqrt(p*(1-p)/n)
print("%d"%((1-a)*100)+"% CI (using normal):", st.norm.interval(alpha = 1-a, loc = p,
    scale = s))
```

Output:

```
98% CI (using normal): (0.00815906462425664, 0.10384093537574336)
```

□

8.4.3 Δ.Ε. για την πληθυσμιακή διακύμανση

Πρόταση 8.9 (Διάστημα εμπιστοσύνης για τη διακύμανση σ^2 όταν μ γνωστό). *Έστω τυχαίο δείγμα $X_1, X_2, \ldots, X_n$ από κανονικό πληθυσμό $N(\mu, \sigma^2)$ με γνωστή μέση τιμή μ.*

Η στατιστική συνάρτηση

$$Y = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2$$

ακολουθεί την χ^2_n και ένα $(1-a)\cdot 100\%$ διάστημα εμπιστοσύνης για την διασπορά σ^2 είναι το

$$\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{n,a/2}} \le \sigma^2 \le \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{n,(1-a/2)}}$$

Απόδειξη. Οι ΤΜ $Z_i = (X_i - \mu)/\sigma$ ακολουθούν την $N(0,1)$, άρα $Y = \sum_{i=1}^{n} Z_i^2 \sim \chi^2_n$.

Για το δ.ε., αναζητάμε $y_1 < y_2$, με

$$P(Y > y_2) = a/2 \Leftrightarrow y_2 = \chi^2_{n,a/2}$$

και

$$P(Y < y_1) = a/2 \Leftrightarrow P(Y \ge y_1) = 1 - a/2 \Leftrightarrow y_1 = \chi^2_{n,1-a/2}$$

Επομένως,

$$1 - a = P(y_1 \le Y \le y_2) = P(\frac{1}{y_2} \le \frac{1}{Y} \le \frac{1}{y_1}) = P\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{n,a/2}} \le \sigma^2 \le \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{n,(1-a/2)}} \right).$$ □

Πρόταση 8.10 (Διάστημα εμπιστοσύνης για τη διακύμανση σ^2 όταν μ άγνωστο). *Έστω τυχαίο δείγμα $X_1, X_2, \ldots, X_n$ από κανονικό πληθυσμό $N(\mu, \sigma^2)$ με γνωστή μέση τιμή μ.*

Η στατιστική συνάρτηση

$$Y = \frac{(n-1)S^2}{\sigma^2}$$

ακολουθεί την χ^2_{n-1} και ένα $(1-a)\cdot 100\%$ διάστημα εμπιστοσύνης για την σ^2 είναι το

$$\frac{(n-1)S^2}{\chi^2_{n-1,a/2}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{n-1,(1-a/2)}}$$

Απόδειξη. Ομοίως με την προηγούμενη (άσκηση). □

Παράδειγμα 8.4.8. *Ο υπεύθυνος για τον ποιοτικό έλεγχο των προϊόντων μια εταιρείας ενδιαφέρεται να εκτιμήσει την διακύμανση των μηκών των μεταλλικών ράβδων που παράγει μια νέα μηχανή προκειμένουν να διαπιστώσει ότι είναι σύμφωνα με τις προδιαγραφές. Για το σκοπό αυτό λαμβάνει τυχαίο δείγμα 25 ράβδων για το οποίο υπολογίζει ότι $s = 1.1$ cm. Να βρεθεί ένα 99% διάστημα εμπιστοσύνης για την διακύμανση του μήκους των ράβδων που παράγει η μηχανή.*

Λύση. Ο μέσος είναι άγνωστος, οπότε θεωρούμε την ΤΜ $Y = (n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$. Για $a = 0.01$, είναι $\chi^2_{n-1,a/2} = 45.5585$ και $\chi^2_{n-1,1-a/2} = 9.886$, οπότε το ζητούμενο δ.ε. είναι το

$$[(n-1)S/\chi^2_{n-1,a/2}, (n-1)S/\chi^2_{n-1,1-a/2}] = [0.58, 2.67]$$

```
import numpy as np
import scipy.stats as st
a, n, s  = 0.01, 25, 1.1
df=n-1
u = st.chi2.ppf(1-a/2, df)
l = st.chi2.ppf(a/2, df)
#(l,u) = st.chi2.interval(1-a, df)
print(l,u)
print((n-1)*s*s/u,(n-1)*s*s/l)
```

Output:

```
9.886233502241467 45.558511936530586
0.6374220483859704 2.937417975553165
```

□

Παράδειγμα 8.4.9. *Ένα δείγμα μεγέθους $n = 15$ έδωσε $S^2 = 17.2$. Να βρεθεί ένα 90% διάστημα εμπιστοσύνης για την σ^2.*

Λύση. Η μέση τιμή είναι άγνωστη, οπότε θα χρησιμοποιηθεί η στατιστική: $Y = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$. Το $(1-a)100\%$ δ.ε. είναι το $[(n-1)s^2/\chi^2_{n-1,a/2}, (n-1)s^2/\chi^2_{n-1,1-a/2}]$. Για $n = 15$, $s^2 = 17.2$, $a = 0.1$, είναι $\chi^2_{n-1,a/2} = 23.685$ και $\chi^2_{n-1,1-a/2} = 6.571$ και τελικά το ζητούμενο δ.ε. είναι το

$$[\frac{14 \cdot 17.2}{23.685}, \frac{14 \cdot 17.2}{6.571}] = [10.166, 36.645]$$

```
import numpy as np
import scipy.stats as st
a, n, s2  = 0.1, 15, 17.2
df=n-1
(l,u) = st.chi2.interval(1-a, df)
print(l,u)
print((n-1)*s2/u,(n-1)*s2/l)
```

Output:

```
6.57063138378934 23.684791304840576
10.16686180176671 36.64792406314057
```

□

Παράδειγμα 8.4.10. *Μια μηχανή έχει ρυθμιστεί να παράγει προϊόντα μέσου βάρους* $18kg$. *Δείγμα* $n = 9$ *προϊόντων έδωσε*

$$18.2,\ 17.9,\ 18.3,\ 18,\ 18.3,\ 17.9,\ 18.1,\ 17.9,\ 18.2$$

Να βρεθεί ένα 95% *διάστημα εμπιστοσύνης για την διακύμανση του βάρους των προϊόντων.*

Λύση. Επειδή δίνεται $\mu = 18$, χρησιμοποιούμε τη στατιστική συνάρτηση $Y = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 \sim \chi^2_n$.

```
import numpy as np
import scipy.stats as st
data = np.array([18.2,  17.9,  18.3,  18,  18.3,  17.9,  18.1,  17.9,  18.2])
a, n, mu = 0.05, len(data), 18
d2 = data - 18
sse = np.dot(d2,d2)
(l1,u1) = st.chi2.interval(1-a, n)
print("chi2 points:",(l1,u1))
print("%d%% CI for sigma^2:"%(100*(1-a)), (np.sqrt(sse/u1),np.sqrt(sse/l1)))
```

Output:

```
chi2 points: (2.7003894999803584, 19.02276779864163)
95% CI for sigma^2: (0.1255809528608859, 0.3333092927256988)
```

□

Παράδειγμα 8.4.11. *Οι χωρητικότητες 10 μπαταριών μετρήθηκαν ως εξής:*

$$140, 136, 150, 144, 148, 152, 138, 141, 143, 151.$$

i) Να υπολογισθεί ένα 99% *δ.ε. για την πληθυσμιακή διακύμανση* σ^2.

ii) Να υπολογισθεί μια τιμή v *ώστε να είναι* $\sigma^2 < v$ *με* 90% *εμπιστοσύνη.*

Λύση. Θεωρούμε τη στατιστική συνάρτηση $Y = (n-1)S^2/\sigma^2$. *ii)* Είναι

$$P(\sigma^2 < v) = 0.9 \Leftrightarrow P(Y > (n-1)S^2/v) = 0.9 \Leftrightarrow (n-1)S^2/v = \chi^2_{n-1,0.9} \Leftrightarrow v = (n-1)S^2/\chi^2_{n-1,0.9},$$

$S^2 = 32.23$ και $\chi^2_{n-1,0.9} = 4.168$, οπότε $v = 69.6$.

```
import numpy as np
import scipy.stats as st
data = np.array([140, 136, 150, 144, 148, 152, 138, 141, 143, 151])
a, n = 0.01, len(data)
S2 = st.tvar(data)
(l,u) = st.chi2.interval(1-a, n-1)
print("sample variance S^2 =", S2)
print("chi2 points:",(l,u))
print("%d%% CI for sigma^2:"%(100*(1-a)), ((n-1)*S2/u, (n-1)*S2/l))
p=0.9
q = st.chi2.ppf(1-p, n-1) #q = 0.1-lower percent point
print("P(chi^2 < %s) = %s" %(q, st.chi2.cdf(q, n-1)))
print("P(sigma^2 < %s) = %s" %((n-1)*S2/q,p))
```

Output:

```
sample variance S^2 = 32.23333333333333
chi2 points: (1.7349329049966606, 23.589350781257387)
99% CI for sigma^2: (12.297922172173351, 167.2110772494446)
P(chi^2 < 4.168159008146107) = 0.09999999999999999
P(sigma^2 < 69.59907226020852) = 0.9
```

□

Παράδειγμα 8.4.12. *Ένα μηχάνημα κατασκευάζει κυκλικά μεταλλικά εξαρτήματα διαμέτρου* $\mu = 11.8cm$. *Ένα δείγμα μεγέθους 10 έδωσε τις ακόλουθες μετρήσεις:*

$$11.66, 11.92, 11.75, 11.80, 11.82, 11.71, 11.84, 11.77, 11.81, 11.79$$

Να κατασκευαστεί ένα 95% *δ.ε. για την τυπική απόκλιση* σ *της διαμέτρου*

i) λαμβάνοντας υπόψη τον μέσο μ,

ii) λαμβάνοντας υπόψη μόνο τον δειγματικό μέσο.

Αν η τυπική απόκλιση δεν επιτρέπεται να ξεπερνά το $0.1cm$, *με ποιο επίπεδο εμπιστοσύνης το μηχάνημα τηρεί αυτή την προδιαγραφή;*

8.4.4 Δ.Ε. για την διαφορά μέσων δύο πληθυσμών

Πρόταση 8.11 (Δ.Ε. διαφοράς δύο μέσων τιμών). *Έστω δύο ανεξάρτητα τυχαία δείγματα* $(X_1, X_2, \ldots, X_{n_1})$ *και* $(Y_1, Y_2, \ldots, Y_{n_2})$ *από κανονικούς πληθυσμούς* $N(\mu_1, \sigma_1^2)$ *και* $N(\mu_2, \sigma_2^2)$ *αντίστοιχα, με* σ_1^2 *και* σ_2^2 *γνωστές. Αν* $\overline{X}$ *και* $\overline{Y}$ *οι δειγματικοί μέσοι, τότε*

$$\overline{X} - \overline{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}) \quad \textit{και} \quad Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

και ένα $(1-a)\cdot 100\%$ *διάστημα εμπιστοσύνης για την διαφορά* $\mu_1 - \mu_2$ *είναι το*

$$[(\overline{X} - \overline{Y}) \pm B], \qquad \textit{όπου } B = z_{a/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Πρόταση 8.12 (Δ.Ε. διαφοράς δύο μέσων τιμών (άγνωστη διακύμανση - μεγάλα δείγματα)). *Έστω δύο ανεξάρτητα τυχαία δείγματα* $(X_1, X_2, \ldots, X_{n_1})$ *και* $(Y_1, Y_2, \ldots, Y_{n_2})$ *από κανονικούς πληθυσμούς με άγνωστες μέσες τιμές* μ_1, μ_2 *και άγνωστες διακυμάνσεις* σ_1^2, σ_2^2 *αντίστοιχα. Τότε,*

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \to N(0,1)$$

και ένα προσεγγιστικό $(1-a)\cdot 100\%$ *διάστημα εμπιστοσύνης για την διαφορά* $\mu_1 - \mu_2$ *είναι το*

$$[(\overline{X} - \overline{Y}) \pm B], \qquad \textit{όπου } B = z_{a/2} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}.$$

Πρόταση 8.13 (Δ.Ε. διαφοράς δύο μέσων τιμών (άγνωστες διακυμάνσεις αλλά ίσες - μικρά δείγματα)). *Έστω δύο ανεξάρτητα τυχαία δείγματα* $(X_1, X_2, \ldots, X_{n_1})$ *και* $(Y_1, Y_2, \ldots, Y_{n_2})$ *από κανονικούς πληθυσμούς με άγνωστες μέσες τιμές* μ_1, μ_2 *και άγνωστες διακυμάνσεις* $\sigma_1^2 = \sigma_2^2 = \sigma^2$ *αντίστοιχα. τότε,*

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2} \qquad \textit{όπου } S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

και ένα $(1-a)\cdot 100\%$ *διάστημα εμπιστοσύνης για την διαφορά* $\mu_1 - \mu_2$ *είναι το*

$$[(\overline{X} - \overline{Y}) \pm B], \qquad \textit{όπου } B = t_{n_1+n_2-2,a/2} \cdot \sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

Παράδειγμα 8.4.13. *Μετρήσεις των τελευταίων* 25 *ετών έδειξαν ότι η μέση βροχόπτωση στην περιοχή Α για έναν ορισμένο μήνα είναι* 12.2 *cm. Σε μια άλλη περιοχή Β τα τελευταία* 20 *έτη για τον ίδιο μήνα η μέση βροχόπτωση ήταν* 10.5 *cm. Αν θεωρήσουμε ότι οι κατανομές των δύο βροχοπτώσεων είναι* $N(\mu_1, \sigma_1^2)$ *και* $N(\mu_2, \sigma_2^2)$ *να βρεθεί ένα διάστημα εμπιστοσύνης με επίπεδο σημαντικότητας* 95% *για την διαφορά των μέσων βροχοπτώσεων στις δύο περιοχές όταν*

i) $\sigma_1 = 1.5$ *cm*, $\sigma_2 = 0.5$ *cm*.

ii) $\sigma_1^2 = \sigma_2^2 = \sigma^2$ *άγνωστη, αλλά* $S_1 = 1.2$ *cm*, $S_2 = 0.3$ *cm*.

Λύση. Είναι $\overline{X} = 12.2$, $n_1 = 25$, $\overline{Y} = 10.5$, $n_2 = 20$ και $a = 0.05$, οπότε $z_{a/2} = 1.96$ και $t_{n_1+n_2-2,a/2} = 2.0167$.

```
import numpy as np
import scipy.stats as st

xbar, ybar, n1, n2, a = 12.2, 10.5, 25, 20, 0.05
#known variance
sigma1, sigma2 = 1.5, 0.5
se = np.sqrt(sigma1**2/n1 + sigma2**2/n2)
z = st.norm.ppf(1-a/2)
B = z*se
print("i) %s%% CI for mu1 - mu2:"%(1-a),(xbar-ybar - B, xbar-ybar + B))

#unknown variance
s1, s2 = 1.2, 0.3
t = st.t.ppf(1-a/2, n1+n2-2)
sp2 = ((n1-1)*s1**2+(n2-1)*s2**2)/(n1+n2-2)
B2 = t*np.sqrt(sp2*(1.0/n1 + 1.0/n2))
print("ii) %s%% CI for mu1 - mu2:"%(1-a),(xbar-ybar - B2, xbar-ybar + B2))
#approximate
se2 = np.sqrt(s1**2/n1 + s2**2/n2)
B3 = z*se2
print("ii) approximated %s%% CI:"%(1-a),(xbar-ybar - B3, xbar-ybar + B3))
```

Output:

```
i) 0.95% CI for mu1 - mu2: (1.0725053553048052, 2.3274946446951934)
ii) 0.95% CI for mu1 - mu2: (1.1443511472091799, 2.2556488527908187)
ii) approximated 0.95% CI: (1.211579491866787, 2.1884205081332113)
```

□

8.4.5 Δ.Ε. για την διαφορά ποσοστών δύο πληθυσμών

Πρόταση 8.14 (Διάστημα εμπιστοσύνης διαφοράς ποσοστών). *Αν* $(X_1, X_2, \ldots, X_n)$, $(Y_1, Y_2, \ldots, Y_n)$ *είναι δύο ανεξάρτητα τυχαία δείγματα από πληθυσμούς* Bernoulli(p_1) *και* Bernoulli(p_2), *μεγέθους* n_1 *και* n_2 *αντίστοιχα, τότε*

$$Z = \frac{\overline{P}_1 - \overline{P}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0,1) \qquad \text{όπου } \overline{P}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \quad \overline{P}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$$

και ένα προσεγγιστικό $(1-a)\cdot 100\%$ *διάστημα εμπιστοσύνης για την διαφορά* $p_1 - p_2$ *είναι το*

$$[\overline{P}_1 - \overline{P}_2 \pm B], \qquad \text{όπου } B = z_{a/2}\sqrt{\frac{\overline{P}_1(1-\overline{P}_1)}{n_1} + \frac{\overline{P}_2(1-\overline{P}_2)}{n_2}}.$$

Παράδειγμα 8.4.14. *Σε μια ομάδα* 200 *ανδρών βρέθηκαν* 82 *καπνιστές, ενώ σε μια ομάδα* 300 *γυναικών βρέθηκαν* 87 *καπνίστριες. Να βρεθεί ένα* 95% *δ.ε. για την διαφορά των ποσοστών* $p_1 - p_2$ *όπου* p_1 *το ποσοστό των καπνιστών και* p_2 *το ποσοστό των καπνιστριών.*

Λύση.

```
import numpy as np
import scipy.stats as st
n1, n2, a = 200, 300, 0.05
p1, p2 = (1.0)*82/200, (1.0)*87/300
z = st.norm.ppf(1-a/2)
B = z*np.sqrt(p1*(1-p1)/n1 + p2*(1-p2)/n2)
print("Approximated %s%% CI for p1 - p2:"%(1-a),(p1-p2-B, p1-p2+B))
```

Output:

```
Approximated 0.95% CI for p1 - p2: (0.03466087836812895, 0.20533912163187104)
```

□

Παράδειγμα 8.4.15. *Σε μια σφυγμομέτρηση παρατηρήθηκε ότι στην περιοχή* A *σε τυχαίο δείγμα* 500 *ψηφοφόρων οι* 420 *ψηφίζουν το* X *κόμμα, ενώ στην περιοχή* B *σε τυχαίο δείγμα* 300 *ψηφοφόρων οι* 219 *ψηφίζουν το* X *κόμμα. Να βρεθεί με πιθανότητα* 0.99 *ένα δ.ε. για την αληθινή διαφορά στο ποσοστό των ψηφοφόρων που θα ψηφίσουν το* X *κόμμα στις δύο περιοχές.*

Λύση.

```
import numpy as np
import scipy.stats as st
n1, n2, a = 500, 300, 0.01
p1, p2 = (1.0)*420/500, (1.0)*219/300
z = st.norm.ppf(1-a/2)
B = z*np.sqrt(p1*(1-p1)/n1 + p2*(1-p2)/n2)
print("Approximated %s%% CI for p1 - p2:"%(1-a),(p1-p2-B, p1-p2+B))
```

Output:

```
Approximated 0.99% CI for p1 - p2: (0.031625340362937196, 0.1883746596370628)
```

□

8.4.6 Δ.Ε. για τον λόγο των διακυμάνσεων δύο πληθυσμών

Πρόταση 8.15 (Δ.Ε. λόγου διακυμάνσεων). *Αν $(X_1, X_2, \ldots, X_{n_1})$ και $(Y_1, Y_2, \ldots, Y_{n_2})$ είναι δύο ανεξάρτητα τυχαία δείγματα από δύο κανονικούς πληθυσμούς, με δειγματικούς μέσους $\overline{X}$, $\overline{Y}$ και δειγματικές διακυμάνσεις S_1^2, S_2^2 αντίστοιχα, τότε*

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$$

και ένα $(1-a)\cdot 100\%$ διάστημα εμπιστοσύνης για τον λόγο σ_1^2/σ_2^2 είναι το

$$F_{n_2-1,n_1-1,1-a/2}\frac{S_1^2}{S_2^2} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq F_{n_2-1,n_1-1,a/2}\frac{S_1^2}{S_2^2}$$

Παράδειγμα 8.4.16. *Ένα δείγμα μεγέθους $n_A = 8$ από τον πληθυσμό A έδωσε $\overline{X}_A = 43$ και $S_A^2 = 17$. Άλλο δείγμα μεγέθους $n_B = 13$ από τον πληθυσμό B έδωσε $\overline{X}_B = 31$ και $S_B^2 = 22$. Βρείτε ένα 98% διάστημα εμπιστοσύνης για τον λόγο των διακυμάνσεων των δύο πληθυσμών.*

Λύση.

```
import numpy as np
import scipy.stats as st
n1, n2, svar1, svar2, a = 8, 13, 17, 22, 0.02
(l,u) = st.f.interval(1-a, n2-1, n1-1)
#u1 = st.f.ppf(1-a/2, n2-1, n1-1)
#l1 = st.f.ppf(a/2, n2-1, n1-1)
ci = (l*svar1/svar2, u*svar1/svar2)
print(l,u)
print("%s%% CI for var1/var2:"%(1-a), ci)
```

Output:

```
0.21554035406113947 6.469091278841487
0.98% CI for var1/var2: (0.16655390995633504, 4.998843260922968)
```

□

8.5 Λυμένες ασκήσεις

Άσκηση 8.1. *Αν $(X_1, X_2, \ldots, X_n)$ είναι τυχαίο δείγμα που προέρχεται από δειγματοληψία χωρίς επανατοποθέτηση από έναν πληθυσμό μεγέθους N, όπου $0 < n < N$, με μέσο μ και διακύμανση σ^2, να δειχθεί ότι $V(\overline{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$.*

Λύση. Έστω $S_n = X_1 + \cdots + X_n$, οπότε $\overline{X} = S_n/n$. Γνωρίζουμε ότι $E(S_n) = E(X_1) + \cdots + E(X_n) = n\mu$. Οι X_i δεν είναι ανεξάρτητες, όμως είναι ισοκατανεμημένες, επομένως κάθε ζεύγος (X_i, X_j), $i \neq j$, έχει την ίδια από κοινού κατανομή άρα και την ίδια συνδιακύμανση, έστω την $c = \mathrm{COV}(X_i, X_j)$. Θέτοντας $Y_i = X_i - \mu$, έχουμε ότι $E(Y_i) = 0$, $V(Y_i) = V(X_i)$ και

$$\begin{aligned} V(S_n) &= E\left((X_1 + \cdots + X_n - n\mu)^2\right) = E\left((Y_1 + \cdots + Y_n)^2\right) = E\left(\sum_{(i,j)\in[n]^2} Y_i Y_j\right) = \sum_{(i,j)\in[n]^2} E(Y_i Y_j) \\ &= \sum_{i=1}^{n} E(Y_i^2) + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} E(Y_i Y_j) = \sum_{i=1}^{n} V(Y_i) + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} \mathrm{COV}(Y_i, Y_j) \\ &= \sum_{i=1}^{n} V(X_i) + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} \mathrm{COV}(X_i, X_j) = n\sigma^2 + n(n-1)c. \end{aligned}$$

Η παραπάνω σχέση ισχύει και όταν $n = N$. Τότε όμως η S_N είναι σταθερή (και όχι τυχαία), οπότε έχει διακύμανση 0, δηλαδή $0 = V(S_N) = N\sigma^2 + N(N-1)c$, από όπου προκύπτει ότι

$$c = \frac{-\sigma^2}{N-1}$$

Αντικαθιστώντας στην προηγούμενη, προκύπτει ότι

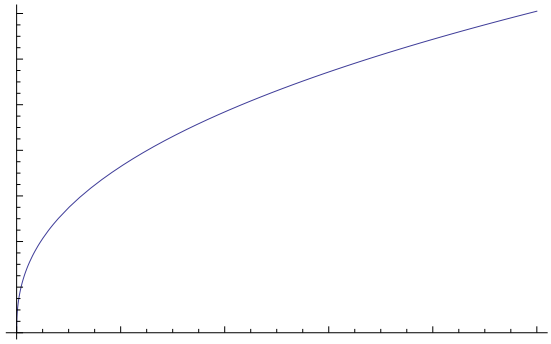
$$V(S_n) = n\sigma^2 - n(n-1)\frac{\sigma^2}{N-1} = n\sigma^2\left(1 - \frac{n-1}{N-1}\right) = n\sigma^2\frac{N-n}{N-1}.$$

Επομένως, $V(\overline{X}) = V\left(\frac{S_n}{n}\right) = \frac{1}{n^2}V(S_n) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$. □

8.6 Ασκήσεις προς επίλυση

1) Έστω $x_1, x_2, \ldots, x_n$ ένα τυχαίο δείγμα που προέρχεται από πληθυσμό με συνάρτηση πυκνότητας πιθανότητας

$$f(x) = \begin{cases} (k+1)x^k, & \text{αν } 0 \le x \le 1 \\ 0, & \text{αλλιώς} \end{cases}, \text{ όπου } k > -1.$$



i) Να βρεθεί με τη μέθοδο της μέγιστης πιθανοφάνειας μια εκτιμήτρια για την παράμετρο k.

ii) Με βάση την παραπάνω εκτιμήτρια, ποιά είναι η εκτίμηση του k αν το δείγμα αποτελούνταν (σε αύξουσα σειρά) από τις επόμενες 10 τιμές 0.108, 0.206, 0.401, 0.417, 0.603, 0.657, 0.675, 0.726, 0.778, 0.895; (Μπορείτε να δείτε την πραγματική τιμή του k αν παρατηρήσετε την γραφική παράσταση της $\sigma.\pi.\pi$ που συνοδεύει την άσκηση.)

2) Χρησιμοποιήστε τυχαίο δείγμα μεγέθους 100 από την ομοιόμορφη κατανομή $U(0,1)$ για να προσεγγίσετε την τιμή του ολοκληρώματος

$$I = \int_0^1 \sqrt{1-x^2}dx = E(\sqrt{1-U^2}), \qquad U \sim U(0,1).$$

Δώστε ένα 95% δ.ε. για την τιμή του I. (Η πραγματική του τιμή είναι $\pi/4$.)

3) Οι παρατηρητές A και B έκαναν τις παρακάτω ανεξάρτητες μετρήσεις:

$$A : 142,\ 150,\ 137,\ 141,\ 139,\ 147,\ 142$$

$$B : 160,\ 164,\ 156,\ 161,\ 162,\ 163,\ 159,\ 166.$$

Να βρεθεί ένα 0.98 διάστημα εμπιστοσύνης για την διαφορά $\mu_A - \mu_B$, αν $\sigma_A^2 = \sigma_B^2$.

4) Έστω $(X_1, X_2), \ldots, X_n, X_{n+1})$ δείγμα από κανονική κατανομή με άγνωστα μ, σ^2. Υποθέστε ότι γνωρίζετε τις n πρώτες τιμές του δείγματος και θέλετε να προβλέψετε την τιμή του X_{n+1}.

Εκφράστε τη διακύμανση της ΤΜ $X_{n+1} - \overline{X}$, όπου $\overline{X} = (X_1 + \cdots + X_n)/n$, συναρτήσει της σ.

Δώστε ένα $(1-a)100\%$ δ.ε. για το X_{n+1}.

5) Έστω $X_1, X_2, \ldots, X_n \sim E(\theta) = \Gamma(1, \theta)$, τυχαίο δείγμα με εκθετική κατανομή με μέσο $1/\theta$. Δείξτε ότι

$$2\theta \sum_{i=1}^{n} X_i \sim \chi^2_{2n} = \Gamma(n, 1/2).$$

Βρείτε ένα $100(1-a)\%$ δ.ε. για τον μέσο $1/\theta$.

(Γενικά, αν $X \sim \Gamma(a, \theta)$, τότε $bX \sim \Gamma(a, \theta/b)$, $b > 0$.)

Κεφάλαιο 9

Έλεγχος υποθέσεων και σημαντικότητας

9.1 Έλεγχος υποθέσεων

Ένα διάστημα εμπιστοσύνης που κατασκευάσαμε με βάση τα προηγούμενα, μπορεί να περιέχει την πραγματική τιμή μιας άγνωστης παραμέτρου με μεγάλη πιθανότητα. Όμως, πολύ συχνά θέλουμε να πάρουμε μια απόφαση, δηλαδή να απαντήσουμε με ναι ή όχι σε μια ερώτηση που εμπεριέχει τυχαιότητα, εξασφαλίζοντας ταυτόχρονα ότι η απόφασή μας είναι η σωστή με μεγάλη πιθανότητα. Τα στατιστικά εργαλεία με τα οποία κατασκευάσαμε διαστήματα εμπιστοσύνης, μπορούν να χρησιμοποιηθούν και για τον σκοπό αυτό με παρόμοιο τρόπο. Η διαφορά είναι ότι τώρα δεν προσπαθούμε να εκτιμήσουμε την άγνωστη παράμετρο, αλλά να ελέγξουμε πόσο στατιστικά ισχυρή είναι η υπόθεση στην οποία βασίζεται η απόφασή μας.

Για παράδειγμα, ας υποθέσουμε ότι μια εταιρεία που κατασκευάζει μπαταρίες, ισχυρίζεται ότι έχουν μέση διάρκεια ζωής 240 ώρες. Για απλότητα, θεωρούμε ότι η διάρκεια ζωής είναι μια ΤΜ $X \sim N(\mu, \sigma^2)$. Έχουμε ένα τυχαίο δείγμα μεγέθους $n = 20$, για το οποίο ο δειγματικός μέσος $\overline{X}$ είναι ίσος με $\overline{x} = 220$. Πόσο σίγουροι είμαστε για τον ισχυρισμό της εταιρείας; Συμβολίζουμε τον ισχυρισμό αυτό ως $H_0 : \mu \geq \mu_0$, όπου $\mu_0 = 240$. Ο ισχυρισμός αυτός ονομάζεται **υπόθεση 0** ή **μηδενική υπόθεση** (null hypothesis). Ο αντίθετος ισχυρισμός ονομάζεται **υπόθεση 1** ή **εναλλακτική υπόθεση** (alternative hypothesis) και συμβολίζεται με H_1. Εδώ είναι $H_1 : \mu < \mu_0$. Προκειμένου να απορρίψουμε την H_0, απαιτούμε η πιθανότητα $p = P(\overline{X} \leq \overline{x}|H_0)$ να παρατηρηθεί μια τιμή της $\overline{X}$ τόσο "μακριά" από την υπόθεση H_0, δεδομένου ότι η H_0 είναι σωστή, να είναι πολύ μικρή, μικρότερη από κάποια τιμή $a \in (0, 1)$, η οποία ονομάζεται **επίπεδο σημαντικότητας** του ελέγχου. Αν συμβεί το αντίθετο, τότε αυτό θεωρείται ότι δεν αποτελεί επαρκή ένδειξη για την απόρριψη της H_0, οπότε την αποδεχόμαστε, χωρίς όμως αυτό να σημαίνει ότι είναι σωστή.

Η αναλογία με τα διαστήματα εμπιστοσύνης είναι άμεση: Ένα (αριστερόπλευρο) $(1 - a)100\%$ δ.ε. για την μ είναι το $(-\infty, \overline{X} + z_a\sigma/\sqrt{n}]$. Πράγματι, είναι

$$P(\mu \in (-\infty, \overline{X} + z_a\sigma/\sqrt{n}]) = P(\overline{X} \geq \mu - z_a\sigma/\sqrt{n}) = P(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \geq -z_a) = 1 - \Phi(-z_a) = 1 - a.$$

Επομένως, αν

$$\mu_0 \notin (-\infty, \overline{X} + z_a\sigma/\sqrt{n}] \Leftrightarrow \mu_0 > \overline{X} + z_a\sigma/\sqrt{n} \Leftrightarrow \overline{X} < \mu_0 - z_a\frac{\sigma}{\sqrt{n}} \Leftrightarrow Z < -z_a,$$

όπου $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$ η στατιστική συνάρτηση του ελέγχου, τότε απορρίπτουμε την H_0, με εμπιστοσύνη $(1 - a)100\%$. Η τιμή $x_c = \mu_0 - z_a\dfrac{\sigma}{\sqrt{n}}$ ονομάζεται **κρίσιμη τιμή** του ελέγχου.

Προφανώς, η απόρριψη εξαρτάται τόσο από την τιμή της στατιστικής συνάρτησης Z, όσο και από την τιμή του a. Καθώς μειώνεται το a, μειώνεται και το $-z_a$, οπότε γίνεται πιο πιθανό να ισχύει $Z \geq -z_a$, οπότε η H_0 δεν απορρίπτεται. Η μέγιστη τιμή του a για την οποία δεν απορρίπτεται η H_0 ονομάζεται **p-τιμή (p-value)** του ελέγχου. Στο συγκεκριμένο παράδειγμα, αυτό συμβαίνει όταν

$$\overline{X} = x_c \Leftrightarrow Z = -z_a \Leftrightarrow \Phi(Z) = a,$$

οπότε είναι p-value $= \Phi(Z)$. Σημειώνεται ότι η p-value είναι μια ΤΜ, αφού εξαρτάται από την Z. Μάλιστα, αποτελεί ένα άνω φράγμα της πιθανότητας p, αφού είναι

$$p = P(\overline{X} \leq \overline{x}|H_0) \overset{*}{\leq} P(\overline{X} \leq \overline{x}|\mu = \mu_0) = P(Z \leq \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}|\mu = \mu_0) \overset{**}{=} \Phi(\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}) = \text{p-value}.$$

Η ισότητα (**) ισχύει διότι όταν είναι $\mu = \mu_0$, τότε $Z \sim N(0,1)$. Η ανισότητα (*) δείχνει ότι η p μεγιστοποιείται όταν $\mu = \mu_0$. Επομένως, κρατώντας συντηρητική στάση, μπορούμε να θεωρούμε ότι σε κάθε περίπτωση η μηδενική υπόθεση είναι η $H_0 : \mu = \mu_0$ και ότι αυτή που αλλάζει είναι η H_1. Το πλεονέκτημα είναι ότι η $H_0 : \mu = \mu_0$ καθορίζει πλήρως την κατανομή της στατιστικής συνάρτησης και για το λόγο αυτό ονομάζεται απλή (αλλιώς ονομάζεται σύνθετη). Στη συγκεκριμένη περίπτωση, υπό την απλή υπόθεση H_0, είναι $Z \sim N(0,1)$. Για τον λόγο αυτόν, στα επόμενα θα θεωρούμε ότι η H_0 είναι απλή (ισότητα) και θα διακρίνουμε περιπτώσεις ανάλογα με την H_1.

Τελικά, απορρίπτουμε την H_0, όταν

$$\text{p-value} < a \Leftrightarrow \Phi(Z) < a \Leftrightarrow Z < -z_a \Leftrightarrow \overline{X} < \mu_0 - z_a\frac{\sigma}{\sqrt{n}},$$

και την αποδεχόμαστε όταν p-value $\geq a$. Το σύνολο $R = (-\infty, -z_a) \subseteq S_Z$ ονομάζεται **χωρίο απόρριψης** (ή κρίσιμη περιοχή), και συμβολίζεται για συντομία ως $R = \{Z < -z_a\}$. Επειδή δεν είναι κάτω φραγμένο, ο έλεγχος αυτός ονομάζεται **αριστερόπλευρος έλεγχος**. Αντίστοιχα προκύπτουν ο δεξιόπλευρος και ο δίπλευρος έλεγχος, ανάλογα με τη μορφή της H_1. Οι τρεις περιπτώσεις συνοψίζονται στον επόμενο πίνακα:

Έλεγχος για την μέση τιμή μ κανονικού πληθυσμού με διακύμανση σ^2 γνωστή

H_1	Χωρίο απόρριψης	Κρίσιμη τιμή	p-value
(αριστερόπλευρος έλεγχος) $\mu < \mu_0$	$R = \{Z < -z_a\}$	$x_c = \mu_0 - z_a\frac{\sigma}{\sqrt{n}}$	$\Phi(Z)$
(δεξιόπλευρος έλεγχος) $\mu > \mu_0$	$R = \{Z > z_a\}$	$x_c = \mu_0 + z_a\frac{\sigma}{\sqrt{n}}$	$\Phi(-Z)$
(δίπλευρος έλεγχος) $\mu \neq \mu_0$	$R = \{\|Z\| > z_{a/2}\}$	$x_c = \mu_0 \pm z_{a/2}\frac{\sigma}{\sqrt{n}}$	$2\Phi(-\|Z\|)$

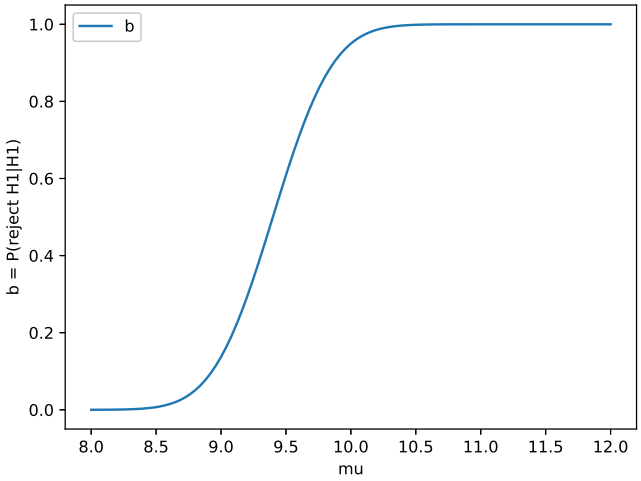
Η απόρριψη μιας σωστής μηδενικής υπόθεσης ονομάζεται **σφάλμα τύπου Ι**. Ο παραπάνω έλεγχος ορίσθηκε με τέτοιο τρόπο ώστε να ελαχιστοποιεί την πιθανότητα σφάλματος τύπου Ι: Η πιθανότητα αυτή με βάση τα παραπάνω είναι το πολύ ίση με a. Πράγματι, είναι

$$P(\overline{X} < \mu_0 - z_a\sigma/\sqrt{n}|\mu \geq \mu_0) \leq P(\overline{X} < \mu_0 - z_a\sigma/\sqrt{n}|\mu = \mu_0) = \Phi(-z_a) = a.$$

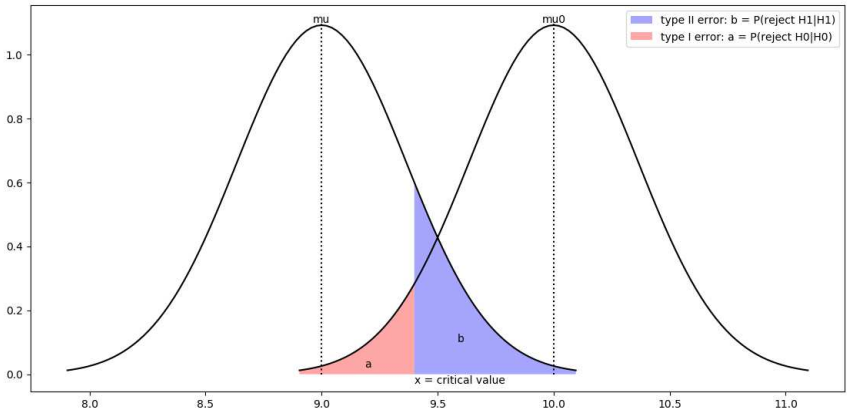
Επομένως, επιλέγοντας μικρότερο a, μειώνουμε αυτή την πιθανότητα. Όμως, υπάρχει και το ενδεχόμενο αποδοχής μιας λανθασμένης μηδενικής υπόθεσης, το οποίο ονομάζεται **σφάλμα τύπου ΙΙ**, με πιθανότητα που συμβολίζεται με β. Στη συγκεκριμένη περίπτωση, είναι

$$\beta = P(\overline{X} \geq \mu_0 - z_a\frac{\sigma}{\sqrt{n}}|H_1) = P(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \geq \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_a|\mu < \mu_0) = \Phi(z_a - \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}) \leq \Phi(z_a) = 1 - a$$

Στην επόμενη εικόνα φαίνεται η γραφική παράσταση του $\beta = \Phi(z_a - \frac{\mu_0 - \mu}{\sigma/\sqrt{n}})$, ως συνάρτηση του μ, για $\mu_0 = 10$, $n = 30$, $a = 0.05$, $\sigma = 2$:



Στην επόμενη εικόνα, για τις ίδες τιμές και για $\mu = 9$, φαίνονται οι τιμές των a και β ως τα εμβαδά του κόκκινου και του μπλε χωρίου αντίστοιχα. Η αριστερή καμπύλη αντιστοιχεί στην πραγματική κατανομή της ΤΜ $\overline{X} \sim N(\mu, \sigma^2/n)$, με $\mu = 9$. Η δεξιά αντιστοιχεί στην κατανομή της $\overline{X}$ όταν ισχύει η $H_0 : \mu = \mu_0 = 10$.



Η πιθανότητα $1 - \beta$ απόρριψης μιας λανθασμένης μηδενικής υπόθεσης ονομάζεται **ισχύς του ελέγχου**. Προφανώς, το β είναι συνάρτηση των μ, a, n. Λύνοντας ως προς n, έχουμε ότι

$$\Phi(z_a - \frac{\mu_0 - \mu}{\sigma/\sqrt{n}}) = \beta = \Phi(-z_\beta) \Rightarrow z_a - \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} = -z_\beta \Rightarrow \frac{\mu_0 - \mu}{\sigma}\sqrt{n} = z_a + z_\beta \Rightarrow \sqrt{n} = \frac{(z_a + z_\beta)\sigma}{\mu_0 - \mu},$$

δηλαδή, αν δίνονται τα μ, a, μ_0, μπορούμε να υπολογίσουμε το κατάλληλο μέγεθος δείγματος n ώστε να εξασφαλίσουμε ένα συγκεκριμένο μικρό άνω φράγμα για το β.

Παράδειγμα 9.1.1. *Κανονικός πληθυσμός έχει διακύμανση $\sigma^2 = 1$. Θέλουμε να ελέγξουμε την υπόθεση ότι η μέση τιμή μ του πληθυσμού είναι $\mu_0 = 9$, έναντι της $H_1 : \mu \neq 9$. Λαμβάνουμε ένα τυχαίο δείγμα μεγέθους $n = 16$, στο οποίο ο δειγματικός μέσος ισούται με $\overline{X} = 9.4$. Να ελεγχθεί η υπόθεσή μας σε επίπεδο σημαντικότητας $a = 0.05$.*

Λύση. Η μηδενική υπόθεση είναι η $H_0 : \mu = \mu_0$, οπότε θα πραγματοποιηθεί δίπλευρος έλεγχος. Για $a = 0.05$, είναι $z_{a/2} = 1.96$.

Η στατιστική συνάρτηση ελέγχου είναι η $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$, με

$$Z = \frac{9.4 - 9}{1/4} = 1.6 < 1.96 = z_{a/2},$$

οπότε η H_0 δεν απορρίπτεται. Στο ίδιο συμπέρασμα καταλήγουμε, υπολογίζοντας

$$\text{p-value} = 2\Phi(-1.6) = 0.1096 > 0.05 = a.$$

Μπορούμε να υπολογίσουμε το μέγεθος n του δείγματος, για το οποίο θα απορρίπταμε την H_0:

$$Z > z_{a/2} \Leftrightarrow \sqrt{n}(\overline{X} - \mu_0) > 1.96 \Leftrightarrow \sqrt{n} > 1.96/0.4 = 4.9 \Leftrightarrow n \geq 25.$$

Για την ισχύ του ελέγχου, έχουμε

$$\begin{aligned} 1 - \beta &= 1 - P(-z_{a/2} \leq \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \leq z_{a/2}|H_1) = 1 - P(-z_{a/2} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} - \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} \leq z_{a/2}|H_1) \\ &= 1 - \Phi(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{a/2}) + \Phi(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_{a/2}) \end{aligned}$$

□

Παράδειγμα 9.1.2. *Εικάζουμε ότι κανονικός πληθυσμός έχει μέση τιμή $\mu_0 = 20$ και γνωστή διακύμανση $\sigma^2 = 7$. Ένα δείγμα μεγέθους $n = 25$ από τον πληθυσμό έδωσε $\overline{x} = 21.5$. Να ελεχθεί, σε επίπεδο σημαντικότητας $a = 0.04$, η υπόθεση $H_0 : \mu = 20$, έναντι της υπόθεσης*

i) $H_1 : \mu \neq 20$.

ii) $H_1 : \mu > 20$.

Λύση. i) Είναι $H_1 : \mu \neq \mu_0$, οπότε θα πραγματοποιηθεί δίπλευρος έλεγχος. Για $a = 0.04$, είναι $z_{a/2} = 2.054$. Η στατιστική συνάρτηση ελέγχου είναι η $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$, με

$$Z = \frac{21.5 - 20}{\sqrt{7/25}} = 2.8347 > 2.054 = z_{a/2},$$

οπότε η H_0 απορρίπτεται. Στο ίδιο συμπέρασμα καταλήγουμε αν υπολογίσουμε την p-value του ελέγχου:

$$\text{p-value} = 2\Phi(-|Z|) = 2\Phi(-2.8347) = 0.0046 < 0.04 = a.$$

ii) Στην περίπτωση αυτή, είναι p-value $= \Phi(-Z) = 0.0023 < a$, οπότε η H_0 απορρίπτεται. □

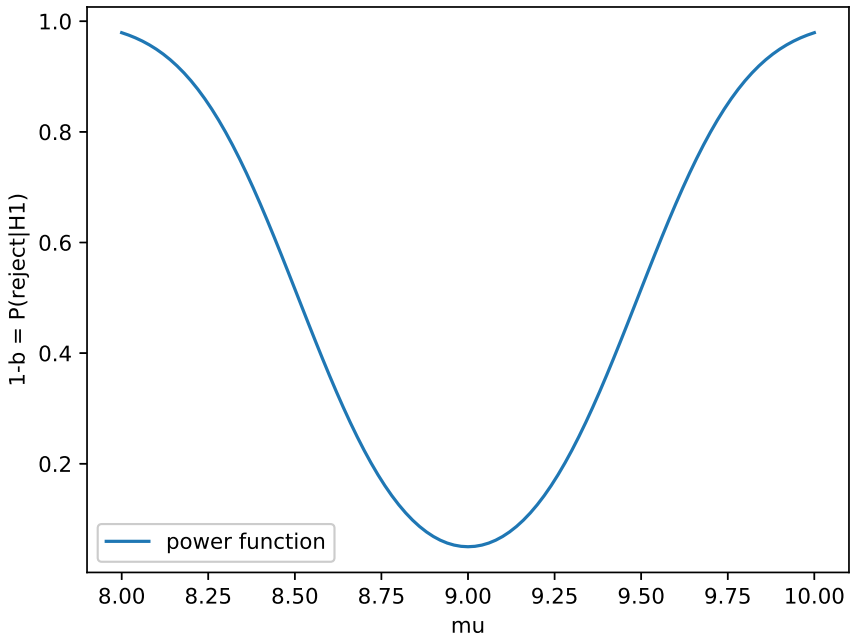
Παρατηρήστε ότι η p-value είναι διπλάσια στην περίπτωση δίπλευρου ελέγχου, διότι στην πιθανότητα μιας τιμής $\overline{x}$ τόσο μεγαλύτερης από την μ_0 προστίθεται και η πιθανότητα μια τιμής εξίσου μικρότερης.

Παρακάτω δίνεται ο κώδικας για τη λύση του παραδείγματος 9.1.1.

```
import numpy as np
import scipy.stats as st
import matplotlib.pyplot as plt

xbar, mu0, n, a, sigma = 9.4, 9, 16, 0.05, 1 #9.1.1
Z = np.sqrt(n)*(xbar-mu0)/sigma
pval = 2*st.norm.cdf(-np.abs(Z))
zpp = st.norm.ppf(1-a/2)
if(pval < a): print("H0 is rejected")
else: print("H0 is accepted")
print("Z=%s, p-value = %s, z_{a/2} = %s"%(Z, pval, zpp))

mu = np.linspace(8,10, 101)
err = (mu0-mu)*np.sqrt(n)
p = 1-st.norm.cdf(err+zpp) + st.norm.cdf(err-zpp)
fig, ax = plt.subplots(1, 1)
ax.plot(mu, p, label = 'power function')
ax.legend()
ax.set_xlabel("mu")
ax.set_ylabel("1-b = P(reject|H1)")
plt.show()
```

Output:

```
H0 is accepted
Z=1.6000000000000014, p-value = 0.10959858339911567, z_{a/2} = 1.959963984540054
```



Έλεγχος για την μέση τιμή μ κανονικού πληθυσμού με σ^2 άγνωστη (one sample t-test)

Η περίπτωση αυτή αντιμετωπίζεται ομοίως με την προηγούμενη, με τη διαφορά ότι τώρα η στατιστική συνάρτηση του ελέγχου είναι η

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

Υπό την υπόθεση $H_0 : \mu = \mu_0$, η T ως γνωστό ακολουθεί την κατανομή t με $n-1$ βαθμούς ελευθερίας, οπότε, αντίστοιχα με πριν, προκύπτει ο ακόλουθος πίνακας:

H_1	Χωρίο απόρριψης	Κρίσιμη τιμή	p-value
(αριστερόπλευρος έλεγχος) $\mu < \mu_0$	$R = \{T < -t_{n-1,a}\}$	$x_c = \mu_0 - t_{n-1,a}\frac{S}{\sqrt{n}}$	$F_{n-1}(T)$
(δεξιόπλευρος έλεγχος) $\mu > \mu_0$	$R = \{T > t_{n-1,a}\}$	$x_c = \mu_0 + t_{n-1,a}\frac{S}{\sqrt{n}}$	$F_{n-1}(-T)$
(δίπλευρος έλεγχος) $\mu \neq \mu_0$	$R = \{\|T\| > t_{n-1,a/2}\}$	$x_c = \mu_0 \pm t_{n-1,a/2}\frac{\sigma}{\sqrt{n}}$	$2F_{n-1}(-\|T\|)$

όπου F_{n-1} η αθροιστική συνάρτηση της κατανομής t με $n-1$ βαθμούς ελευθερίας.

Αν δίνεται η λίστα τιμών A του δείγματος, ο παραπάνω t-έλεγχος εκτελείται στην Python με τη συνάρτηση

```
scipy.stats.ttest_1samp(A, popmean = mu0, alternative = s)
```

όπου η s παίρνει αντίστοιχα μία από τις τιμές "less", "greater", "two-sided". Επιστρέφονται οι τιμές T και p-value.

Παράδειγμα 9.1.3. *Δείγμα μεγέθους $n = 36$ έδωσε δειγματικό μέσο $\overline{X} = 71$ και δειγματική διακύμανση $S^2 = 15$. Να ελεχθεί, σε επίπεδο σημαντικότητας $a = 0.02$, η υπόθεση $H_0 : \mu = 70$, έναντι της υπόθεσης: i) $H_1 : \mu \neq 70$, ii) $H_1 : \mu > 70$.*

Λύση. i) Είναι $H_1 : \mu \neq \mu_0$, όπου $\mu_0 = 70$, οπότε θα πραγματοποιηθεί δίπλευρος έλεγχος. Για $a = 0.02$, είναι $t_{n-1,a/2} = 2.4377$. Η στατιστική συνάρτηση ελέγχου είναι η $T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$, με

$$T = \frac{71 - 70}{\sqrt{15/36}} = 1.55 < 2.4377 = t_{n-1,a/2},$$

οπότε η H_0 δεν απορρίπτεται. Στο ίδιο συμπέρασμα καταλήγουμε αν υπολογίσουμε την p-value του ελέγχου:

$$\text{p-value} = 2F_{n-1}(-1.55) = 0.13 > 0.02 = a.$$

ii) Είναι $H_1 : \mu > \mu_0$, όπου $\mu_0 = 70$, οπότε θα πραγματοποιηθεί δεξιόπλευρος έλεγχος. Είναι $T = 1.55 < 2.133 = t_{n-1,a}$ και p-value $= F_{n-1}(-T) = 0.065 > a$, οπότε πάλι η H_0 δεν απορρίπτεται. □

```
import numpy as np
import scipy.stats as st
xbar, mu0, n, a, s = 71, 70, 36, 0.02, np.sqrt(15)
T = np.sqrt(n)*(xbar-mu0)/s
pval = 2*st.t.cdf(-np.abs(T), n-1)
tpp = st.t.ppf(1-a/2, n-1)
print("i) H1: mu != %s, T=%s, p-value = %s, t_{a/2,n-1} = %s"%(mu0, T, pval, tpp))
```

Output:

```
i) H1: mu != 70, T=1.5491933384829668, p-value = 0.13033124887220834, t_{a/2,n-1} =
    2.437722547143737
```

Έλεγχος για το ποσοστό p πληθυσμού

Έστω τυχαίο δείγμα $(X_1, \ldots, X_n)$, με $X_i \sim \text{Bernoulli}(p)$ και έστω η μηδενική υπόθεση $H_0 : p = p_0$, για το ποσοστό p του πληθυσμού. Χρησιμοποιούμε τον συμβολισμό

$$F(x; n, p) := \sum_{i=0}^{\lfloor x \rfloor} \binom{n}{i} p^i (1-p)^{n-i}.$$

για την αθροιστική συνάρτηση της κατανομής $\text{Binom}(n, p)$.

Ως γνωστό, η ΤΜ $X = \sum_{i=1}^{n} X_i$ ακολουθεί διωνυμική κατανομή με παραμέτρους n, p. Υπό την υπόθεση H_0, έχουμε ότι $X \sim \text{Binom}(n, p_0)$, με μέση τιμή $\mu_0 = np_0$. Έστω ότι η X έχει πάρει την τιμή k στο συγκεκριμένο δείγμα. Προκειμένου να απορρίψουμε την H_0, θα πρέπει η πιθανότητα η X να πάρει μια τιμή τόσο μακριά από την μ_0 όσο η παρατηρηθείσα τιμή k, δεδομένου ότι η H_0 είναι σωστή, να είναι μικρότερη από ένα επίπεδο εμπιστοσύνης a. Στην περίπτωση αριστερόπλευρου ελέγχου, δηλαδή όταν $H_1 : p < p_0$, η πιθανότητα αυτή είναι

$$\text{p-value} = P(X \le k | H_0) = P(X \le k | p = p_0) = F(k; n, p_0),$$

Στην περίπτωση δεξιόπλευρου ελέγχου, δηλαδή όταν $H_1 : p > p_0$, η πιθανότητα αυτή είναι

$$\text{p-value} = P(X \ge k | H_0) = 1 - P(X < k | p = p_0) = 1 - F(k-1; n, p_0).$$

Στην περίπτωση δίπλευρου ελέγχου, δηλαδή όταν $H_1 : p \neq p_0$, η H_0 απορρίπτεται ανν ισχύουν οι ισοδύναμες συνθήκες

$$\begin{aligned} &P(X \le k | H_0) < a/2 \text{ ή } P(X \ge k | H_0) < a/2 \Leftrightarrow 2F(k; n, p_0) < a \text{ ή } 2(1 - F(k-1; n, p_0)) < a \\ \Leftrightarrow &2\min\{F(k; n, p_0), 1 - F(k-1; n, p_0)\} < a. \end{aligned}$$

Εναλλακτικά, όταν το n είναι μεγάλο, βάσει του ΚΟΘ, μπορεί να χρησιμοποιηθεί προσεγγιστικά ο z-έλεγχος που παρουσιάστηκε προηγουμένως. Πράγματι, ως γνωστό, όταν ισχύει η $H_0 : p = p_0$, τότε είναι

$$Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} = \frac{\overline{X} - p_0}{\sqrt{p_0(1-p_0)/n}} \to N(0,1)$$

Ο τρόπος αυτός προτιμάται όταν το n είναι μεγάλο, αφού τότε οι υπολογισμοί μέσω της διωνυμικής κατανομής είναι υπολογιστικά απαιτητικοί. Εισάγοντας και τη διόρθωση συνέχειας, ανάλογα με τη μορφή της H_1, η προσέγγιση μπορεί να βελτιωθεί σημαντικά, όταν το n δεν είναι πολύ μεγάλο. Συγκεκριμένα, όταν έχουμε αριστερόπλευρο (αντ. δεξιόπλευρο) έλεγχο, είναι $P(X \le k) = P(X \le k + 1/2)$ (αντ. $P(X \ge k) = P(X \ge k - 1/2)$), οπότε χρησιμοποιούμε αντίστοιχα τις στατιστικές συναρτήσεις

$$Z_+ = \frac{X - np_0 + 1/2}{\sqrt{np_0(1-p_0)}}, \qquad \text{αντ.} \quad Z_- = \frac{X - np_0 - 1/2}{\sqrt{np_0(1-p_0)}}$$

Κατόπιν τούτων, προκύπτουν οι παρακάτω περιπτώσεις για τον έλεγχο:

H_1	p-value (binomial test)	p-value (z-test)
(αριστερόπλευρος έλεγχος) $p < p_0$	$F(X; n, p_0)$	$\Phi(Z_+)$
(δεξιόπλευρος έλεγχος) $p > p_0$	$1 - F(X-1; n, p_0)$	$\Phi(-Z_-)$
(δίπλευρος έλεγχος) $p \neq p_0$	$2\min\{F(X; n, p_0), 1 - F(X-1; n, p_0)\}$	$2\min\{\Phi(Z_+), \Phi(-Z_-)\}$

Στον κώδικα που ακολουθεί, εκτελείται αρχικά αριστερόπλευρος, δεξιόπλευρος και δίπλευρος διωνυμικός έλεγχος με μηδενική υπόθεση $H_0 : p = p_0 = 0.5$, όπου το X παράγεται τυχαία από την κατανομή Binom(n, p), όπου $n = 30, p = 0.6$. Στη συνέχεια, εκτελείται z-έλεγχος πρώτα χωρίς και έπειτα με διόρθωση συνέχειας. Ο τελευταίος δίνει καλύτερη προσέγγιση, όπως φαίνεται και από το αποτέλεσμα.

```
import numpy as np
import scipy.stats as st

n, p, p0 = 30, 0.6, 0.5
x = np.random.binomial(n, p)
H1 = {'greater':'p>p0', 'less':'p<p0', 'two-sided':'p!=p0'}

print("X = %s, X/n = %s"%(x, x/n))
print("\nexact binomial test:")
theory_pval = {'greater':1-st.binom.cdf(x-1,n,p0),
               'less':st.binom.cdf(x,n,p0),
               'two-sided':2*np.min([1-st.binom.cdf(x-1,n,p0),st.binom.cdf(x,n,p0)])}
for k, v in H1.items():
    result = st.binom_test(x, n, p0, alternative = k)
    print("H1:%s, p-value = %s, theory p-value = %s"%(v, result, theory_pval[k]))

print("\nz-test:")
se = np.sqrt(n*p0*(1 - p0))
z = (x - n*p0)/se
theory_pval = {'greater':st.norm.cdf(-z),
               'less':st.norm.cdf(z),
               'two-sided':2*st.norm.cdf(-np.abs(z))}
for i in H1.keys():
    print("H1:%s, theory p-value = %s"%(H1[i], theory_pval[i]))

print("\nz-test with continuity correction:")
zm = (x - n*p0 - 1/2)/se
zp = (x - n*p0 + 1/2)/se
theory_pval = {'greater':st.norm.cdf(-zm),
               'less':st.norm.cdf(zp),
               'two-sided':2*np.min([st.norm.cdf(-zm),st.norm.cdf(zp)])}
for i in H1.keys():
    print("H1:%s, theory p-value = %s"%(H1[i], theory_pval[i]))
```

Output:

```
X = 19, X/n = 0.6333333333333333

exact binomial test:
H1:p>p0, p-value = 0.10024421103298661, theory p-value = 0.10024421103298664
H1:p<p0, p-value = 0.9506314266473055, theory p-value = 0.9506314266473055
H1:p!=p0, p-value = 0.20048842206597323, theory p-value = 0.20048842206597328

z-test:
H1:p>p0, theory p-value = 0.07206351740800766
H1:p<p0, theory p-value = 0.9279364825919924
H1:p!=p0, theory p-value = 0.14412703481601533

z-test with continuity correction:
H1:p>p0, theory p-value = 0.10062131047886197
H1:p<p0, theory p-value = 0.9498258767688547
H1:p!=p0, theory p-value = 0.20124262095772394
```

Παράδειγμα 9.1.4. *Σε μια πόλη, 1 στους 10 καταναλωτές προτιμά την μάρκα Α. Μετά από μια έντονη διαφημιστική καμπάνια, εξετάστηκε ένα τυχαίο δείγμα 200 ατόμων, προκειμένου να διαπιστωθεί η αποτελεσματικότητα της καμπάνιας. Οι μετρήσεις έδειξαν ότι 26 άτομα προτιμούν την Α. Να ελεχθεί σε επίπεδο σημαντικότητας $a = 0.05$ η αποτελεσματικότητα της καμπάνιας.*

Λύση. Η μηδενική υπόθεση είναι η $H_0 : p \leq p_0$, όπου $p_0 = 0.1$ και $H_1 : p > p_0$. Η τελευταία δηλώνει ότι το ποσοστό αυξήθηκε μετά την καμπάνια. Έχουμε $X = 26 > 20$, οπότε εφαρμόζοντας δεξιόπλευρο διωνυμικό έλεγχο, βρίσκουμε ότι

$$\text{p-value} = 1 - F(X-1; n, p_0) = 1 - F(25; 200, 0.1) = 0.1$$

οπότε η H_0 δεν μπορεί να απορριφθεί. □

Στο ίδιο αποτέλεσμα καταλήγουμε και αν εφαρμόσουμε z-έλεγχο, όπου προκύπτει p-value = 0.097425, όπως φαίνεται στον κώδικα που ακολουθεί.

```
import numpy as np
import scipy.stats as st

n, a, p0 = 200, 0.05, 0.1 #9.1.4
x = 26
H1 = 'greater'
#binomial test
pval = st.binom_test(x, n, p0, alternative = H1)
print("binomial test:\nx =", x, ", p-value =", pval, ", Upper a-percent point:", st.
    binom.ppf(1 - a,n,p0),"\n")
#print("p-value = ", 1-st.binom.cdf(x-1,n,p0))

#z-test
z = (x - n*p0 - 1/2)/np.sqrt(p0*(1 - p0)*n)
print("z-test:\nz = %f, z_{a} = %f, p-value = %f\n"
      %(z, st.norm.ppf(1-a), st.norm.cdf(-z)))
```

Output:

```
binomial test:
x = 26 , p-value = 0.10045728651286649 , Upper a-percent point: 27.0

z-test:
z = 1.296362, z_{a} = 1.644854, p-value = 0.097425
```

Παράδειγμα 9.1.5. *Εικάζεται ότι το ποσοστό των χορτοφάγων σε μια πόλη είναι 22%. Επιλέγουμε τυχαία 260 άτομα και βρίσκουμε ανάμεσά τους 62 χορτοφάγους. Σε επίπεδο σημαντικότητας $a = 0.02$, είναι σωστή αυτή η εικασία;*

Λύση. Θέτουμε $H_0 : p = p_0$ και $H_1 : p \neq p_0$, όπου $p_0 = 0.22$. Εφαρμόζοντας δίπλευρο z-έλεγχο (χωρίς διόρθωση συνέχειας), βρίσκουμε

$$Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} = 0.718614, \quad z_{a/2} = 2.326348, \qquad \text{p-value} = 0.472379$$

οπότε η H_0 δεν απορρίπτεται. Σημειώνεται ότι το δειγματικό ποσοστό $62/260 = 0.23846$ είναι πολύ κοντά στο 0.22. Αν εισάγουμε διόρθωση συνέχειας, τότε προκύπτει p-value = 0.51973. □

Έλεγχος για την διαφορά $\mu_1 - \mu_2$ δύο πληθυσμών $N(\mu_1, \sigma_1^2)$ και $N(\mu_2, \sigma_2^2)$, όταν σ_1, σ_2 γνωστές Αν $(X_1, \ldots, X_{n_1}) \sim N(\mu_1, \sigma_1^2)$ και $(Y_1, \ldots, Y_{n_2}) \sim N(\mu_2, \sigma_2^2)$ είναι ανεξάρτητα τυχαία δείγματα από 2 πληθυσμούς και $\overline{X}$, $\overline{Y}$ οι αντίστοιχοι δειγματικοί μέσοι, τότε ως γνωστό είναι

$$\overline{X} - \overline{Y} \sim N(\mu_1 - \mu_2, \sigma^2), \qquad \text{όπου } \sigma^2 = \sigma_1^2/n_1 + \sigma_2^2/n_2.$$

Αν θέλουμε να ελέγξουμε π.χ. την $H_0 : \mu_1 = \mu_2$, έναντι της $H_1 : \mu_1 \neq \mu_2$, τότε θεωρούμε την στατιστική συνάρτηση

$$Z = \frac{\overline{X} - \overline{Y}}{\sigma} = (\overline{X} - \overline{Y}) / \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$$

η οποία ακολουθεί την $N(0,1)$, όταν ισχύει η H_0. Αυτό σημαίνει ότι η πιθανότητα να παρατηρηθεί μια ακραία τιμή $c > 0$ της $|\overline{X} - \overline{Y}|$, όταν ισχύει η H_0, είναι

$$P(|\overline{X} - \overline{Y}| \geq c | H_0) = P(|Z| \geq c/\sigma | H_0) = 2\Phi(-c/\sigma).$$

Επομένως, η p-value του ελέγχου είναι η $2\Phi(-|Z|)$ και η H_0 απορρίπτεται σε επίπεδο σημαντικότητας a, όταν p-value $< a \Leftrightarrow 2\Phi(-|Z|) < a \Leftrightarrow -|Z| < -z_{a/2} \Leftrightarrow |Z| > z_{a/2} \Leftrightarrow |\overline{X} - \overline{Y}| > z_{a/2}\sigma$.

Ομοίως προκύπτουν και οι υπόλοιπες περιπτώσεις, που συνοψίζονται στον επόμενο πίνακα:

H_1	Χωρίο απόρριψης	p-value
(αριστερόπλευρος έλεγχος) $\mu_1 - \mu_2 < 0$	$R = \{Z < -z_a\}$	$\Phi(Z)$
(δεξιόπλευρος έλεγχος) $\mu_1 - \mu_2 > 0$	$R = \{Z > z_a\}$	$\Phi(-Z)$
(δίπλευρος έλεγχος) $\mu_1 - \mu_2 \neq 0$	$R = \{\|Z\| > z_{a/2}\}$	$2\Phi(-\|Z\|)$

Παράδειγμα 9.1.6. *Δοκιμάζουμε ελαστικά αυτοκινήτου μιας μάρκας σε μια τοποθεσία Α και μιας άλλης μάρκας σε μια άλλη τοποθεσία Β και παίρνουμε τα ακόλουθα δείγματα για τη διάρκεια ζωής τους (σε 100km):*

$$A : 61.1, 58.2, 62.3, 64, 59.7, 66.2, 57.8, 61.4, 62.2, 63.6$$

$$B : 62.2, 56.6, 66.4, 56.2, 56.2, 57.4, 58.4, 57.6, 65.4$$

Από την εμπειρία γνωρίζουμε ότι η τυπική απόκλιση της διάρκειας ζωής εξαρτάται από την τοποθεσία, με $\sigma_A = 3$ και $\sigma_B = 4$, καθώς και ότι η διάρκεια ζωής ακολουθεί κανονική τιμή με μέσο που εξαρτάται μόνο από την κατασκευή. Να ελεχθεί η υπόθεση $H_0 : \mu_A = \mu_B$, έναντι της υπόθεσης $H_1 : \mu_A \neq \mu_B$.

Λύση. Υπολογίζοντας όπως παρακάτω ότι p-value = 0.21, δεν απορρίπτουμε την H_0.

```
import numpy as np
import scipy.stats as st
A = [61.1, 58.2, 62.3, 64, 59.7, 66.2, 57.8, 61.4, 62.2, 63.6]
B = [62.2, 56.6, 66.4, 56.2, 56.2, 57.4, 58.4, 57.6, 65.4]
sA, sB = 3, 4
xbar, ybar, sigma = st.tmean(A), st.tmean(B), np.sqrt(sA**2/len(A) + sB**2/len(B))
z = (xbar-ybar)/sigma
pval = 2*st.norm.cdf(-np.abs(z))
print("Z = %s, sigma = %s, p-value = %s"%(z, sigma, pval))
```

Output:

```
Z = 1.2527562972298287, sigma = 1.636391694484477, p-value = 0.21029441080002742
```

□

Έλεγχος για την διαφορά $\mu_1-\mu_2$ δύο πληθυσμών $N(\mu_1,\sigma_1^2)$ και $N(\mu_2,\sigma_2^2)$, όταν σ_1,σ_2 άγνωστες, αλλά ίσες (two sample t-test)

Θεωρούμε 2 ανεξάρτητα δείγματα $(X_1,\dots,X_{n_1})\sim N(\mu_1,\sigma_1^2)$ και $(Y_1,\dots,Y_{n_2})\sim N(\mu_2,\sigma_2^2)$ όπως πριν, αλλά τώρα είναι $\sigma_1=\sigma_2=\sigma$, αλλά σ άγνωστη. Αν θέλουμε να ελέγξουμε π.χ. την $H_0:\mu_1=\mu_2$, έναντι της $H_1:\mu_1\neq\mu_2$, τότε θεωρούμε την στατιστική συνάρτηση

$$T=\frac{\overline{X}-\overline{Y}}{S_p\sqrt{1/n_1+1/n_2}},\qquad \text{όπου } S_p^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$$

η οποία ως γνωστό ακολουθεί την $t_{n_1+n_2-2}$, όταν ισχύει η H_0. Αυτό σημαίνει ότι η πιθανότητα να παρατηρηθεί μια ακραία τιμή $c>0$ της $|\overline{X}-\overline{Y}|$, όταν ισχύει η H_0, είναι

$$P(|\overline{X}-\overline{Y}|\geq c|H_0)=P(|T|\geq c/D|H_0)=2F_{n_1+n_2-2}(-c/D),$$

όπου $F_{n_1+n_2-2}$ η αθροιστική συνάρτηση της κατανομής $t_{n_1+n_2-2}$ και D ο παρονομαστής της T. Επομένως, η p-value του ελέγχου είναι η $2F_{n_1+n_2-2}(-|T|)$ και η H_0 απορρίπτεται σε επίπεδο σημαντικότητας a, όταν

$$\text{p-value}<a\Leftrightarrow 2_{n_1+n_2-2}(-|T|)<a\Leftrightarrow -|T|<-t_{a/2,n_1+n_2-2}\Leftrightarrow |T|>t_{a/2,n_1+n_2-2}\Leftrightarrow |\overline{X}-\overline{Y}|>t_{a/2,n_1+n_2-2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}.$$

Ομοίως προκύπτουν και οι υπόλοιπες περιπτώσεις, που συνοψίζονται στον επόμενο πίνακα:

H_1	Χωρίο απόρριψης	p-value
(αριστερόπλευρος έλεγχος) $\mu_1-\mu_2<0$	$R=\{T<-t_{a,n_1+n_2-2}\}$	$F_{n_1+n_2-2}(T)$
(δεξιόπλευρος έλεγχος) $\mu_1-\mu_2>0$	$R=\{T>t_{a,n_1+n_2-2}\}$	$F_{n_1+n_2-2}(-T)$
(δίπλευρος έλεγχος) $\mu_1-\mu_2\neq 0$	$R=\{\|T\|>t_{a/2,n_1+n_2-2}\}$	$2F_{n_1+n_2-2}(-\|T\|)$

Έλεγχος για την διαφορά $\mu_1-\mu_2$ δύο πληθυσμών $N(\mu_1,\sigma_1^2)$ και $N(\mu_2,\sigma_2^2)$, όταν σ_1,σ_2 άγνωστες και άνισες: Σε αυτή την περίπτωση, χρησιμοποιείται η στατιστική

$$T=\frac{\overline{X}-\overline{Y}}{\sqrt{S_1^2/n_1+S_2^2/n_2}}\to t_d,\qquad d=\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2\left(\frac{s_1^4}{n_1^2(n_1-1)}+\frac{s_2^4}{n_2^2(n_2-1)}\right)^{-1},$$

της οποίας η κατανομή αποδεικνύεται ότι συγκλίνει στην κατανομή t με d βαθμούς ελευθερίας. Ο έλεγχος αυτός είναι γνωστός ως t-έλεγχος Welch. Αν τα n_1,n_2 είναι αρκετά μεγάλα, τότε για απλότητα χρησιμοποιείται απλός z-έλεγχος, αφού $T\to N(0,1)$. Σημειώνεται ότι και οι δύο έλεγχοι είναι προσεγγιστικοί.

H_1	Χωρίο απόρριψης (z-test)	p-value (z-test)	Χωρίο απόρριψης (t-test)	p-value (t-test)
(αριστερόπλευρος έλεγχος) $\mu_1-\mu_2<0$	$R=\{T<-z_a\}$	$\Phi(T)$	$R=\{T<-t_{a,d}\}$	$F_d(T)$
(δεξιόπλευρος έλεγχος) $\mu_1-\mu_2>0$	$R=\{T>z_a\}$	$\Phi(-T)$	$R=\{T>t_{a,d}\}$	$F_d(-T)$
(δίπλευρος έλεγχος) $\mu_1-\mu_2\neq 0$	$R=\{\|T\|>z_{a/2}\}$	$2\Phi(-\|T\|)$	$R=\{\|T\|>t_{a/2,d}\}$	$2F_d(-\|T\|)$

Αν δίνονται οι λίστες A,B τιμών των δύο δειγμάτων, ο παραπάνω t-έλεγχος εκτελείται στην Python με τη συνάρτηση (equal_var = True, για ίσες διακυμάνσεις):

```
scipy.stats.ttest_ind(A, B, equal_var = False)
```

Παράδειγμα 9.1.7. *Από τυχαίο δείγμα 40 εργαζομένων μιας εταιρείας A, προέκυψε ότι το μέσο ετήσιο εισόδημα είναι $\overline{X} = 54000$, με τυπική απόκλιση $S_1 = 6000$. Επίσης, από τυχαίο δείγμα 50 εργαζομένων μιας εταιρείας B, προέκυψε ότι το μέσο ετήσιο εισόδημα είναι $\overline{Y} = 57000$, με τυπική απόκλιση $S_2 = 8000$. Μπορούμε να δεχθούμε ότι οι εργαζόμενοι στις δύο εταιρείες έχουν τον ίδιο μέσο μισθό σε επίπεδο σημαντικότητας $a = 0.01$; Να γίνει έλεγχος και για τις δύο περιπτώσεις που για τις άγνωστες διακυμάνσεις είναι $\sigma_A = \sigma_B$ και $\sigma_A \neq \sigma_B$.*

Λύση. Θεωρούμε $H_0 : \mu_1 = \mu_2$. Ο κώδικας που ακολουθεί εκτελεί αριστερόπλευρο και δίπλευρο έλεγχο. Η στατιστική συνάρτηση ελέγχου είναι η $T = (\overline{X} - \overline{Y})/s$, όπου

$$s^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right), \text{ όταν } \sigma_A = \sigma_B, \qquad s^2 = \frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}, \text{ όταν } \sigma_A \neq \sigma_B,$$

```
import numpy as np
import scipy.stats as st
n1, n2, xbar, ybar, s1, s2, a = 40, 50, 54000, 57000, 6000, 8000, 0.01 #9.1.7
H1 = ['mu1 < mu2', 'mu1 != mu2']
s = np.sqrt(((n1-1)*s1**2 + (n2-1)*s2**2)/(n1+n2-2)*(1/n1+1/n2))
T = (xbar-ybar)/s #equal variances
tp, pval = st.t.ppf(1-a, n1+n2-2), st.t.cdf(T, n1+n2-2) #less
tp2, pval2 = st.t.ppf(1-a/2, n1+n2-2), 2*st.t.cdf(-np.abs(T), n1+n2-2) #two-sided
print("t-test, equal variances: a =", a)
print("\tH1: %s: T = %.6f, -t_{a} = %.6f, p-value = %.6f"%(H1[0], T, -tp, pval))
print("\tH1: %s: T = %.6f, t_{a/2} = %.6f, p-value = %.6f"%(H1[1], T, tp2, pval2))

T = (xbar-ybar)/(np.sqrt(s1**2/n1 + s2**2/n2)) #unequal variances
zp, zpval = st.norm.ppf(1-a), st.norm.cdf(T) #less
zp2, zpval2 = st.norm.ppf(1-a/2), 2*st.norm.cdf(-np.abs(T)) #two-sided
print("z-test, unequal variances: a =", a)
print("\tH1: %s: Z = %.6f, -z_{a} = %.6f, p-value = %.6f"%(H1[0], T, -zp, zpval))
print("\tH1: %s: Z = %.6f, z_{a/2} = %.6f, p-value = %.6f"%(H1[1], T, zp2, zpval2))

df = np.floor((s1**2/n1 + s2**2/n2)**2/(s1**4/(n1*n1*(n1-1))+ s2**4/(n2*n2*(n2-1))))
tp, tpval = st.t.ppf(1-a, df), st.t.cdf(T, df) #less
tp2, tpval2 = st.t.ppf(1-a/2, df), 2*st.t.cdf(-np.abs(T), df) #two-sided
print("t-test, unequal variances: a =", a)
print("\tH1: %s: T = %.6f, -t_{a} = %.6f, p-value = %.6f"%(H1[0], T, -tp, tpval))
print("\tH1: %s: T = %.6f, t_{a/2} = %.6f, p-value = %.6f"%(H1[1], T, tp2, tpval2))
```

Output:

```
t-test, equal variances: a = 0.01
  H1: mu1 < mu2: T = -1.968922, -t_{a} = -2.369472, p-value = 0.026054
  H1: mu1 != mu2: T = -1.968922, t_{a/2} = 2.632858, p-value = 0.052109
z-test, unequal variances: a = 0.01
  H1: mu1 < mu2: Z = -2.031856, -z_{a} = -2.326348, p-value = 0.021084
  H1: mu1 != mu2: Z = -2.031856, z_{a/2} = 2.575829, p-value = 0.042168
t-test, unequal variances: a = 0.01
  H1: mu1 < mu2: T = -1.968922, -t_{a} = -2.369977, p-value = 0.022609
  H1: mu1 != mu2: T = -1.968922, t_{a/2} = 2.633527, p-value = 0.052145
```

Σε κάθε περίπτωση, δεν απορρίπτουμε την H_0. Η περίπτωση $\sigma_A \neq \sigma_B$ έχει μεγαλύτερη p-value, διότι η παρατηρηθείσα διαφορά των μέσων μπορεί να οφείλεται σε κάποιο βαθμό στη διαφορά των διακυμάνσεων. □

Έλεγχος για την διαφορά $\mu_1 - \mu_2$ δύο πληθυσμών $N(\mu_1, \sigma_1^2)$ και $N(\mu_2, \sigma_2^2)$, όχι ανεξάρτητων (paired t-test)

Θεωρούμε 2 τυχαία δείγματα $(X_1, \dots, X_n) \sim N(\mu_1, \sigma_1^2)$ και $(Y_1, \dots, Y_n) \sim N(\mu_2, \sigma_2^2)$ αλλά τώρα η X_i εξαρτάται από την Y_i. Για παράδειγμα, X_i και Y_i είναι η κατανάλωση βενζίνης του ίδιου αυτοκινήτου i χρησιμοποιώντας την βενζίνη Α και Β αντίστοιχα. Αν θέλουμε να ελέγξουμε π.χ. την $H_0 : \mu_1 = \mu_2$, έναντι της $H_1 : \mu_1 \neq \mu_2$, τότε θεωρούμε την ΤΜ $W_i = X_i - Y_i$, η οποία είναι κανονική $N(\mu_1 - \mu_2, \sigma^2)$, όπου $\sigma^2 = \sigma_1^2 + \sigma_2^2 - 2\mathrm{COV}(X_i, Y_i)$, και, αν ισχύει η H_0, έχει μέση τιμή 0. Το διάνυσμα $(W_1, \dots, W_n)$ είναι ένα τυχαίο δείγμα με δειγματική διακύμανση, έστω S^2. Επομένως, η στατιστική συνάρτηση

$$T = \frac{\overline{X} - \overline{Y}}{S/\sqrt{n}}$$

ακολουθεί την t_{n-1}, όταν ισχύει η H_0. Μπορούμε λοιπόν και σε αυτήν την περίπτωση να εφαρμόσουμε t-έλεγχο με τη στατιστική συνάρτηση T.

Αν δίνονται οι λίστες A, B τιμών των δύο δειγμάτων, ο παραπάνω t-έλεγχος εκτελείται στην Python με τη συνάρτηση:

```
scipy.stats.ttest_rel(A, B, alternative='two-sided')
```

Παράδειγμα 9.1.8. *Σε ένα δείγμα $n = 28$ ανθρώπων, ο μέσος χρόνος απόκρισης σε ένα ερέθισμα είναι $\overline{X} = 1$ sec. Μετά τη λήψη 100ml αλκοόλ, ο μέσος χρόνος απόκρισης αυξήθηκε σε $\overline{Y} = 1.2$ sec, με δειγματική διακύμανση $S^2 = 0.25$. Σε ποιο επίπεδο εμπιστοσύνης μπορούμε να συμπεράνουμε ότι το αλκοόλ αυξάνει το χρόνο απόκρισης;*

Λύση. Ο κώδικας που ακολουθεί εκτελεί αριστερόπλευρο και δίπλευρο έλεγχο.

```
import numpy as np
import scipy.stats as st

n, xbar, ybar, s, a = 28, 1, 1.2, 0.5, 0.05 #9.1.9
H1 = ['mu1 < mu2', 'mu1 != mu2']

T = np.sqrt(n)*(xbar-ybar)/s
tp, pval = st.t.ppf(1-a,n-1), st.t.cdf(T, n-1) #less
tp2, pval2 = st.t.ppf(1-a/2,n-1), 2*st.t.cdf(-np.abs(T), n-1) #two-sided

print("t-test, a =", a)
print("H1: %s: T = %.6f, -t_{a} = %.6f, p-value = %.6f"%(H1[0], T, -tp, pval))
print("H1: %s: T = %.6f, t_{a/2} = %.6f, p-value = %.6f"%(H1[1], T, tp2, pval2))
```

Output:

```
t-test, a = 0.05
H1: mu1 < mu2: T = -2.116601, -t_{a} = -1.703288, p-value = 0.021827
H1: mu1 != mu2: T = -2.116601, t_{a/2} = 2.051831, p-value = 0.043655
```

Μπορούμε λοιπόν, από τον δίπλευρο έλεγχο και για $a = 0.05$, να συμπεράνουμε με εμπιστοσύνη 95% ότι το αλκοόλ επηρεάζει τον χρόνο απόκρισης και από τον αριστερόπλευρο έλεγχο και για $a = 0.03$, να συμπεράνουμε με εμπιστοσύνη 97% ότι το αλκοόλ αυξάνει τον χρόνο απόκρισης. □

Έλεγχος για την διαφορά $p_1 - p_2$ των ποσοστών δύο πληθυσμών Θεωρούμε 2 ανεξάρτητα τυχαία δείγματα $(X_1, \ldots, X_{n_1}) \sim \text{Bernoulli}(p_1)$ και $(Y_1, \ldots, Y_{n_2}) \sim \text{Bernoulli}(p_2)$. Ως γνωστό, είναι $X = \sum_{i=1}^{n_1} X_i \sim \text{Binom}(n_1, p_1)$ και $Y = \sum_{i=1}^{n_2} Y_i \sim \text{Binom}(n_2, p_2)$. Επίσης, ως γνωστό, από το ΚΟΘ, είναι

$$\overline{X} = X/n_1 \to N(p_1, p_1(1-p_1)/n_1), \qquad \overline{Y} = Y/n_2 \to N(p_2, p_2(1-p_2)/n_2),$$

και αφού $\overline{X}, \overline{Y}$ ανεξάρτητες, έχουμε ότι $\overline{X} - \overline{Y} \to N(p_1 - p_2, p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2)$. Επομένως, υπό την μηδενική υπόθεση $H_0 : p_1 = p_2$, έχουμε ότι $\dfrac{\overline{X} - \overline{Y}}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}} \to N(0,1)$.

Επειδή τα p_1, p_2 είναι άγνωστα, ο παρονομαστής αντικαθίσταται από την εκτιμήτρια

$$S_p = \sqrt{\overline{P}(1-\overline{P})(1/n_1 + 1/n_2)}, \qquad \text{όπου } \overline{P} = \frac{X+Y}{n_1+n_2}$$

και μπορούμε να εφαρμόσουμε (προσεγγιστικό) z-έλεγχο με στατιστική συνάρτηση την

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\overline{P}(1-\overline{P})(1/n_1 + 1/n_2)}}, \qquad \overline{P} = \frac{X+Y}{n_1+n_2}.$$

Αν θέλουμε να εφαρμόσουμε ακριβή έλεγχο, τότε, αν παρατηρήσαμε $X = k_1$ και $Y = k_2$, είναι

$$P(X = k_1 | X + Y = k_1 + k_2, H_0) = \frac{\binom{n_1}{k_1}\binom{n_2}{k_2}}{\binom{n_1+n_2}{k_1+k_2}} = f(k_1; M, n, N), \qquad M = n_1 + n_2, n = n_1, N = k_1 + k_2$$

όπου $f(k; M, n, N)$ η συνάρτηση πιθανότητας της υπεργεωμετρικής κατανομής HGeom(M, n, N).

Στην περίπτωση δίπλευρου ελέγχου, η H_0 απορρίπτεται όταν ισχύει μια από τις ισοδύναμες συνθήκες

$$\begin{aligned} &P(X \le k_1 | X + Y = k_1 + k_2, H_0) \le a/2 \text{ ή } P(X \ge k_1 | X + Y = k_1 + k_2, H_0) \le a/2 \\ \Leftrightarrow &2\min\{F(k_1; M, n, N), 1 - F(k_1 - 1; M, n, N)\} < a \end{aligned}$$

Επομένως, είναι p-value $= 2\min\{F(k_1; M, n, N), 1 - F(k_1 - 1; M, n, N)\}$.

Ανάλογα, προκύπτουν και οι υπόλοιπες περιπτώσεις, που συνοψίζονται στον επόμενο πίνακα:

H_1	p-value (exact test)	p-value (z-test)
(αριστερόπλευρος έλεγχος) $p_1 < p_2$	$F(X; M, n, N)$	$\Phi(Z)$
(δεξιόπλευρος έλεγχος) $p_1 > p_2$	$1 - F(X-1; M, n, N)$	$\Phi(-Z)$
(δίπλευρος έλεγχος) $p_1 \ne p_2$	$2\min\{F(X; M, n, N), 1 - F(X-1; M, n, N)\}$	$2\Phi(-\lvert Z\rvert)$

όπου

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\overline{P}(1-\overline{P})(1/n_1 + 1/n_2)}}, \qquad \overline{P} = \frac{X+Y}{n_1+n_2},$$

και $F(x; M, n, N)$ η αθροιστική συνάρτηση της υπεργεωμετρικής κατανομής HGeom(M, n, N).

Παράδειγμα 9.1.9. *Δύο απορρυπαντικά δοκιμάστηκαν για την ικανότητά τους να καθαρίζουν λεκέδες. Το πρώτο ήταν επιτυχές σε* 63 *από* 91 *δοκιμές, ενώ το δεύτερο σε* 49 *από* 79*. Κάποιοι πιστεύουν ότι το πρώτο είναι καλύτερο. Να ελεχθεί η υπόθεση αυτή σε επίπεδο σημαντικότητας* $a = 0.05$*.*

Λύση. Ο κώδικας που ακολουθεί εκτελεί δεξιόπλευρο έλεγχο, με $H_0 : p_1 = p_2$ και $H_1 : p_1 > p_2$.

```
import numpy as np
import scipy.stats as st

x,n1, y, n2, a = 63, 91, 49, 79, 0.05
xbar, ybar = x/n1, y/n2
pbar = (x + y)/(n1 + n2)
sp = np.sqrt(pbar*(1 - pbar)*(1/n1 + 1/n2))
sp2 = np.sqrt(xbar*(1-xbar)/n1 + ybar*(1-ybar)/n2)
Z = (xbar - ybar)/sp

print(xbar, ybar, pbar, sp, sp2)

print("\nz-test: a =", a)
za = st.norm.ppf(1-a)
pval = st.norm.cdf(-Z) #greater
print("Z = %.6f, z_{a} = %.6f, p-value = %.6f"%(Z, za, pval))

print("\nexact test: a =", a)
M,n,N = n1+n2, n1, x+y
pval2 = st.hypergeom.sf(x,M,n,N)
print("p-value = %.6f"%(pval2))
```

Output:

```
0.6923076923076923 0.620253164556962 0.6588235294117647 0.07290617364270878
   0.07295452675511863

z-test: a = 0.05
Z = 0.988319, z_{a} = 1.644854, p-value = 0.161498

exact test: a = 0.05
p-value = 0.125001
```

□

Έλεγχος για την διακύμανση σ^2 κανονικού πληθυσμού: Θεωρούμε τυχαίο δείγμα $(X_1, \ldots, X_{n_1}) \sim N(\mu, \sigma^2)$, με δειγματικό μέσο $\overline{X}$ και δειγματική διακύμανση S^2. Υπό τη μηδενική υπόθεση $H_0 : \sigma = \sigma_0$, η στατιστική συνάρτηση

$$X = \frac{(n-1)S^2}{\sigma_0^2}$$

ακολουθεί κατανομή χ^2_{n-1} με $n-1$ βαθμούς ελευθερίας και για $a \in (0,1)$ είναι

$$P(\chi^2_{1-a/2,n-1} \leq X \leq \chi^2_{a/2,n-1}) = 1 - a$$

Επομένως, για έναν δίπλευρο έλεγχο με επίπεδο εμπιστοσύνης a, αποδεχόμαστε την H_0 όταν $X \in [\chi^2_{1-a/2,n-1}, \chi^2_{a/2,n-1}]$, αλλιώς την απορρίπτουμε. Ανάλογα προκύπτουν και οι υπόλοιπες περιπτώσεις, που συνοψίζονται στον ακόλουθο πίνακα:

H_1	χωρίο απόρριψης	p-value
(αριστερόπλευρος έλεγχος) $\sigma < \sigma_0$	$R = \{X < \chi^2_{1-a,n-1}\}$	$F_{n-1}(X)$
(δεξιόπλευρος έλεγχος) $\sigma > \sigma_0$	$R = \{X > \chi^2_{a,n-1}\}$	$1 - F_{n-1}(X)$
(δίπλευρος έλεγχος) $\sigma \neq \sigma_0$	$R = \{X < \chi^2_{1-a/2,n-1}\} \cup \{X > \chi^2_{a/2,n-1}\}$	$2\min\{F_{n-1}(X), 1 - F_{n-1}(X)\}$

Παράδειγμα 9.1.10. *Αν δείγμα μεγέθους $n = 20$ έχει δειγματική τυπική απόκλιση $S = 0.12$, να ελεγχθεί η υπόθεση $H_0 : \sigma \leq 0.1$ έναντι της $H_1 : \sigma > 0.1$.*

Λύση. Εφαρμόζουμε δεξιόπλευρο έλεγχο. (Ο επόμενος κώδικας εκτελεί και δίπλευρο έλεγχο, ο οποίος όμως δεν είναι απαραίτητος για τη λύση.)

```
import numpy as np
import scipy.stats as st

n, sigma0, S, a = 20, 0.1, 0.12, 0.05
H1 = ['sigma > sigma0', 'sigma != sigma0']

X = (n-1)*S**2/sigma0**2
pval = 1-st.chi2.cdf(X, n-1) #greater
pp = st.chi2.ppf(1-a, n-1)
print("H1:%s, a = %s, X = %f, x_{a} = %f, p-value = %f"%(H1[0], a, X, pp, pval))
pval2 = 2*np.min([st.chi2.cdf(X, n-1), st.chi2.sf(X, n-1)]) #two-sided
pp2 = st.chi2.ppf(1-a/2, n-1)
print("H1:%s, a = %s, X = %f, x_{a/2} = %f, p-value = %f"%(H1[1], a, X, pp2, pval2))
```

Output:

```
H1:sigma > sigma0, a = 0.05, X = 27.360000, x_{a} = 30.143527, p-value = 0.096543
H1:sigma != sigma0, a = 0.05, X = 27.360000, x_{a/2} = 32.852327, p-value = 0.193086
```

Συμπεραίνουμε ότι η $H_0 : \sigma \leq 0.1$ μπορεί να απορριφθεί με επίπεδο εμπιστοσύνης $a = 0.1$. □

Έλεγχος για τον λόγο σ_1/σ_2 των διακυμάνσεων δύο κανονικών πληθυσμών:

Θεωρούμε 2 ανεξάρτητα τυχαία δείγματα $(X_1,\dots,X_{n_1}) \sim N(\mu_1,\sigma_1^2)$ και $(Y_1,\dots,Y_{n_2}) \sim N(\mu_2,\sigma_2^2)$. Αν S_1^2 και S_2^2 είναι οι αντίστοιχες δειγματικές διακυμάνσεις, υπό τη μηδενική υπόθεση $H_0 : \sigma_1 = \sigma_2$, η στατιστική συνάρτηση

$$F = \frac{S_1^2}{S_2^2}$$

ακολουθεί ως γνωστό την κατανομή Fisher F_{n_1-1,n_2-1}, επομένως, για $a \in (0,1)$, είναι

$$P(F_{1-a/2,n_1-1,n_2-1} \le X \le F_{a/2,n_1-1,n_2-1}) = 1-a$$

Επομένως, για έναν δίπλευρο έλεγχο με επίπεδο εμπιστοσύνης a, αποδεχόμαστε την H_0 όταν $X \in [F_{1-a/2,n_1-1,n_2-1}, F_{a/2,n_1-1,n_2-1}]$, αλλιώς την απορρίπτουμε. Ανάλογα προκύπτουν και οι υπόλοιπες περιπτώσεις, που συνοψίζονται στον ακόλουθο πίνακα:

H_1	χωρίο απόρριψης	p-value
(αρ. έλεγχος) $\sigma_1 < \sigma_2$	$R = \{F < F_{1-a,n_1-1,n_2-1}\}$	$F_{n_1-1,n_2-1}(F)$
(δεξ. έλεγχος) $\sigma_1 > \sigma_2$	$R = \{F > F_{a,n_1-1,n_2-1}\}$	$1 - F_{n_1-1,n_2-1}(F)$
(δίπλ. έλεγχος) $\sigma_1 \neq \sigma_2$	$R = \{F < F_{1-a/2,n_1-1,n_2-1}\} \cup \{F > F_{a/2,n_1-1,n_2-1}\}$	$2\min\{F_{n_1-1,n_2-1}(F), 1 - F_{n_1-1,n_2-1}(F)\}$

Παράδειγμα 9.1.11. *Αν δύο ανεξάρτητα δείγματα μεγέθους $n_1 = 10$ και $n_2 = 12$ έχουν δειγματική διακύμανση $S_1^2 = 0.14$ και $S_2^2 = 0.28$ αντίστοιχα, να ελεγχθεί η υπόθεση $H_0 : \sigma_1 = \sigma_2$ έναντι της $H_1 : \sigma_1 \neq \sigma_2$, σε επίπεδο σημαντικότητας $a = 0.05$.*

Λύση. Εφαρμόζουμε δίπλευρο έλεγχο. Για $a = 0.05$, βρίσκουμε $\ell = F_{1-a/2,n_1-1,n_2-1} = 0.2556$ και $u = F_{a/2,n_1-1,n_2-1} = 3.5879$. Επειδή $X = S_1^2/S_2^2 = 0.5 \in [\ell, u]$, η H_0 δεν μπορεί να απορριφθεί. □

Ο κώδικας που ακολουθεί, υπολογίζει τα παραπάνω μεγέθη, καθώς και την p-value του ελέγχου.

```
import numpy as np
import scipy.stats as st

n1, n2, Svar1, Svar2, a = 10, 12, 0.14, 0.28, 0.05
H1 = 'sigma1 != sigma2'

F = Svar1/Svar2
pval = 2*np.min([st.f.cdf(F, n1-1,n2-1), st.f.sf(F, n1-1,n2-1)]) #two-sided
lpp = st.f.ppf(a/2, n1-1,n2-1)
upp = st.f.ppf(1-a/2, n1-1,n2-1)
print("H1:%s, a = %s, F = %f, lpp = %f, upp = %f, p-value = %f"%(H1, a, F, lpp, upp,
    pval))
```

Output:

```
H1:sigma1 != sigma2, a = 0.05, F = 0.500000, lpp = 0.255619, upp = 3.587899, p-value
    = 0.307519
```

Έλεγχος καλής προσαρμογής (goodness of fit test)

Έστω τυχαίο δείγμα $(X_1, \dots, X_n)$ από άγνωστη διακριτή κατανομή, και έστω ότι τα X_i παίρνουν τιμές στο $[k]$. Κάποιος ισχυρίζεται ότι οι τιμές προέρχονται από μια συγκεκριμένη διακριτή κατανομή με $P(X = i) = p_i$, για δεδομένα p_i. Θέλουμε να ελέγξουμε την υπόθεση

$$H_0 : \forall i \in [k], P(X = i) = p_i, \qquad \text{έναντι της } H_1 : \exists i \in [k], P(X = i) \neq p_i,$$

όπου το X αντιπροσωπεύει οποιοδήποτε από τα X_i και $(p_1, \dots, p_k)$ η υποτιθέμενη συνάρτηση πιθανότητας της X. Θέτουμε $N_i = |\{j \in [n] : X_j = i\}|$ το πλήθος των εμφανίσεων της τιμής i στο δείγμα. Υπό την H_0, η ΤΜ N_i ακολουθεί διωνυμική κατανομή με παραμέτρους n και p_i, οπότε $E(N_i) = np_i$. Θεωρούμε τη στατιστική συνάρτηση

$$T = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{N_i^2}{np_i} - n$$

η οποία αντιπροσωπεύει την "απόσταση" των τιμών που παρατηρήθηκαν από τις αναμενόμενες σύμφωνα με την H_0. Δηλαδή, όσο μεγαλύτερη τιμή πάρει η T, τόσο ισχυρότερη ένδειξη έχουμε ότι η H_0 δεν ισχύει. Αποδεικνύεται ότι $T \to \chi^2_{k-1}$ καθώς $n \to \infty$, οπότε ένας έλεγχος με επίπεδο σημαντικότητας a είναι να απορρίψουμε την H_0 αν $T > \chi^2_{a,k-1}$, αλλιώς την αποδεχόμαστε. Η p-value του ελέγχου είναι η

$$\text{p-value} = 1 - F_{k-1}(T)$$

όπου F_{k-1} η αθροιστική συνάρτηση κατανομής της χ^2_{k-1}.

Η προσέγγιση θεωρείται ικανοποιητική όταν $N_i \geq 5$, για κάθε i.

Παράδειγμα 9.1.12. *Ένας κατασκευαστής ισχυρίζεται ότι τα προϊόντα του είναι ποιότητας 1, 2, 3, 4 ή 5 (μεγαλύτερος αριθμός χαμηλότερη ποιότητα) με πιθανότητες*

$$(p_1, p_2, p_3, p_4, p_5) = (0.15, 0.25, 0.35, 0.20, 0.05).$$

Αν σε 30 προϊόντα μετρήσαμε αντίστοιχα πλήθη εμφανίσεων $(N_1, N_2, N_3, N_4, N_5) = (3, 6, 9, 7, 5)$, μπορούμε να απορρίψουμε τον ισχυρισμό σε επίπεδο εμπιστοσύνης 95%;

Λύση. Έχουμε $k = 5$ διαφορετικές ποιότητες - δυνατές τιμές της τυχαίας μεταβλητής.

```
import numpy as np
import scipy.stats as st

a = 0.05
f_obs = np.array([3,6,9,7,5])
n, k = f_obs.sum(), len(f_obs)
f_exp = n*np.array([0.15, 0.25, 0.35, 0.20, 0.05])
T, pval = st.chisquare(f_obs, f_exp) #use built-in function
print("T = %s, p-value = %s"%(T, pval))

T = np.sum((f_obs - f_exp)**2/ f_exp) #use formulas from theory
print("T = %s, chi2_{a,k-1} = %s, p-value = %s"%(T,st.chi2.isf(a,k-1),st.chi2.sf(T,k
    -1)))
```

Output:

```
T = 9.347619047619046, p-value = 0.052974280436963686
T = 9.347619047619046, chi2_{a,k-1} = 9.487729036781158, p-value =
    0.052974280436963686
```

Επομένως, δεν μπορούμε να απορρίψουμε την H_0 (οριακά). □

Στην περίπτωση που η ΤΜ παίρνει άπειρες τιμές, μπορούμε να διαμερίσουμε το σύνολο τιμών σε k υποσύνολα S_i, $i \in [k]$, οπότε N_i είναι το πλήθος των εμφανίσεων τιμών του S_i στο δείγμα. Αν η H_0 δεν καθορίζει πλήρως τις παραμέτρους της κατανομής, τότε χρησιμοποιούμε στη θέση τους τις εκτιμήσεις αυτών από το δείγμα.

Παράδειγμα 9.1.13. *Έστω ότι ο αριθμός ατυχημάτων ανά ημέρα για 30 ημέρες από τον ακόλουθο πίνακα:*

$$8, 0, 0, 1, 3, 4, 0, 2, 12, 5, 1, 8, 0, 2, 0, 1, 9, 3, 4, 5, 3, 3, 4, 7, 4, 0, 1, 2, 1, 2$$

Να ελεγχθεί η υπόθεση ότι η κατανομή των ατυχημάτων είναι Poisson.

Λύση. Εκτιμάμε την παράμετρο της Poisson: $\hat{\lambda} = 3.167$ (πλήθος ατυχημάτων/ πλήθος ημερών). Έστω ΤΜ $X \sim P(\hat{\lambda})$. Διαμερίζουμε το σύνολο τιμών S_X σε 5 υποσύνολα:

$$p_0 = P(X = 0), \quad p_1 = P(X = 1), \quad p_2 = P(X = 2)+P(X = 3), \quad p_3 = P(X = 4)+P(X = 5), \quad p_4 = P(X > 5)$$

Υπολογίζουμε τα N_i, $0 \le i \le 4$. και στη συνέχεια τη στατιστική συνάρτηση $T = \sum_{i=0}^{4} \frac{(N_i - np_i)^2}{np_i}$ και την p-value = 0.000311.

```
import numpy as np
import scipy.stats as st
import matplotlib.pyplot as plt

a = 0.05
obs = np.array([8, 0, 0, 1, 3, 4, 0, 2, 12, 5, 1, 8, 0, 2, 0,
                  1, 9, 3, 4, 5, 3, 3, 4, 7, 4, 0, 1, 2, 1, 2])

lbar = np.mean(obs) #estimate lambda
f_obs = np.bincount(obs) #count occurrences
n, k = np.sum(f_obs), len(f_obs)
f_obs2 = [f_obs[0], f_obs[1], f_obs[2]+f_obs[3], f_obs[4]+f_obs[5], np.sum(f_obs
    [6:-1])]
f_exp = n*st.poisson.pmf(np.arange(0,k), lbar)
f_exp2 = [f_exp[0], f_exp[1], f_exp[2]+f_exp[3], f_exp[4]+f_exp[5], np.sum(f_exp
    [6:-1])]

print("N_i's: %s\nlbar = %f"%(f_obs,lbar))
T, pval = st.chisquare(f_obs2, f_exp2)
print("T = %s, p-value = %s"%(T, pval))

plt.plot(np.arange(k), f_obs/n, label ="observed")
plt.plot(np.arange(k), st.poisson.pmf(np.arange(k),lbar), label="Poisson")
plt.legend()
plt.show()
```

Output:

```
N_i's: [6 5 4 4 4 2 0 1 2 1 0 0 1]
lbar = 3.166667
T = 21.04012686868212, p-value = 0.0003109205102098774
```

Τελικά απορρίπτουμε την H_0 με μεγάλη εμπιστοσύνη. □

Έλεγχος ανεξαρτησίας Θεωρούμε ότι κάθε άτομο ενός πληθυσμού χαρακτηρίζεται από ένα ζεύγος χαρακτηριστικών (ΤΜ) (X, Y) με τιμές στο $[r] \times [c]$. Κάθε άτομο έχει μια τιμή σε κάθε χαρακτηριστικό με κάποια άγνωστη πιθανότητα. Συμβολίζουμε με

$$P_{i,j} = P(X = i, Y = j), \qquad p_i = P(X = i), \qquad q_j = P(Y = j)$$

τις αντίστοιχες πιθανότητες. Η μηδενική υπόθεση είναι ότι τα χαρακτηριστικά X, Y είναι ανεξάρτητα, δηλαδή

$$H_0 : \forall (i, j), P_{i,j} = p_i q_j, \qquad H_1 : \exists (i, j), P_{i,j} \neq p_i q_j.$$

Σε ένα τυχαίο δείγμα μεγέθους n βρέθηκαν $N_{i,j}$ σε πλήθος άτομα με $(X, Y) = (i, j)$. Επομένως $E(N_{i,j}) = nP_{i,j} = np_i q_j$. Η τελευταία ισότητα ισχύει δεδομένης της H_0. Θέτουμε

$$N_i = \sum_{j=1}^{c} N_{i,j}, \qquad M_j = \sum_{i=1}^{r} N_{i,j}$$

οπότε προκύπτουν οι εκτιμήτριες των p_i, q_j

$$\hat{p}_i = \frac{N_i}{n}, \qquad \hat{q}_j = \frac{M_j}{n}.$$

Κατόπιν τούτων, προκύπτει ότι η στατιστική συνάρτηση

$$T = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(N_{i,j} - n\hat{p}_i\hat{q}_j)^2}{n\hat{p}_i\hat{q}_j} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{N_{i,j}^2}{n\hat{p}_i\hat{q}_j} - n$$

για την οποία αποδεικνύεται ότι $T \to \chi^2_{(r-1)(c-1)}$, όταν $n \to \infty$. Επομένως, η p-value του ελέγχου είναι η p-value $= 1 - F(T)$, όπου F η συνάρτηση κατανομής της χ^2 με $(r-1)(c-1)$ βαθμούς ελευθερίας.

Παράδειγμα 9.1.14. *Δίνεται ο επόμενος πίνακας για την προτίμηση πολιτικού κόμματος ενός δείγματος 300 ψηφοφόρων.*

$i \backslash j$	1	2	3	Σύνολο (N_i)
Γυναίκες	68	56	32	156
Άνδρες	52	72	20	144
Σύνολο (M_j)	120	128	52	$n = 300$

Να ελεγχθεί η υπόθεση ότι η επιλογή κόμματος είναι ανεξάρτητη του φύλου.

Λύση. X είναι το φύλο του ατόμου και Y το κόμμα που υποστηρίζει.

```
import numpy as np
import scipy.stats as st
obs = [[68, 56, 32],
       [52, 72, 20]]
T, pval, df, e = st.chi2_contingency(obs)
print("Expected:\n", e)
print("\nT = %s, p-value = %s, df = %s"%(T, pval, df))
```

Output:

```
Expected:
 [[62.4  66.56 27.04]
 [57.6  61.44 24.96]]

T = 6.432856673241291, p-value = 0.04009801943167609, df = 2
```

Απορρίπτουμε την H_0 με επίπεδο εμπιστοσύνης 95%. □

Στα παραπάνω, μπορούμε να θεωρήσουμε ότι μια από τις 2 ΤΜ, π.χ. η Y, αντιπροσωπεύει τον αριθμό δείγματος. Δηλαδή στην περίπτωση που έχουμε c δείγματα από c πληθυσμούς, μπορούμε να ελέγξουμε όπως παραπάνω την υπόθεση ότι ένα χαρακτηριστικό X ακολουθεί την ίδια κατανομή σε κάθε πληθυσμό, δηλαδή είναι ανεξάρτητο επιλογής πληθυσμού:

$$H_0 : \forall j, P_1(X = j) = P_2(X = j) = \cdots = P_c(X = j)$$

Παράδειγμα 9.1.15. *Σε 4 χώρες επιλέχθηκε δείγμα 500 γυναικών και ρωτήθηκαν αν έχουν υποστεί σεξουαλική παρενόχληση στον χώρο εργασίας.*

$i \backslash j$	*AU*	*DE*	*JP*	*US*	*Σύνολο* (N_i)
Ναι	*28*	*30*	*58*	*55*	*171*
Όχι	*472*	*470*	*442*	*445*	*1829*
Σύνολο (M_j)	*500*	*500*	*500*	*500*	$n = 2000$

Να ελεγχθεί η υπόθεση ότι τα ποσοστά παρενόχλησης είναι ίδια σε κάθε χώρα.

Λύση.

```
import numpy as np
import scipy.stats as st
obs = [[28, 30, 58, 55],
       [472, 470, 442, 445]]
T, pval, df, e = st.chi2_contingency(obs)

print("Observed:\n", obs)
print("Expected:\n", e)
print("\nT = %s, p-value = %s, df = %s"%(T, pval, df))
```

Output:

```
Expected:
 [[ 42.75  42.75  42.75  42.75]
 [457.25 457.25 457.25 457.25]]

T = 19.510229921441113, p-value = 0.00021440518558965625, df = 3
```

Απορρίπτουμε την H_0 με επίπεδο εμπιστοσύνης 99%. □

9.2 Ασκήσεις προς επίλυση

1) Το ακόλουθο δείγμα, δίνει τους χρόνους εξυπηρέτησης (σε λεπτά) 28 πελατών σε ένα σύστημα.

$$8.6, 9.4, 5.0, 4.4, 3.7, 11.4, 10.0, 7.6, 14.4, 12.2, 11.0, 14.4, 9.3, 10.5,$$
$$10.3, 7.7, 8.3, 6.4, 9.2, 5.7, 7.9, 9.4, 9.0, 13.3, 11.6, 10.0, 9.5, 6.6$$

Να ελεχθεί η υπόθεση $H_0 : \mu \geq 8$, έναντι της υπόθεσης $H_1 : \mu < 8$.

2)

3) Προκειμένου να συγκριθούν δύο τάξεις μαθητών 2 σχολείων, επιλέχθηκαν τυχαία $n_1 = 50$ και $n_2 = 72$ μαθητές από κάθε σχολείο, για να δώσουν ένα κοινό τεστ, στο οποίο κάθε μαθητής βαθμολογήθηκε από 1 έως 100. Η μέση βαθμολογία και η διακύμανση για κάθε σχολείο ήταν $\overline{X_1} = 70$, $S_1^2 = 25$ και $\overline{X_2} = 75$, $S_2^2 = 36$ αντίστοιχα.

   i) Να βρεθεί ένα 95% διάστημα εμπιστοσύνης για την διαφορά μέσων βαθμολογιών $\mu_1 - \mu_2$.

   ii) Διαφέρουν οι επιδόσεις ανάμεσα στα δύο σχολεία, σε επίπεδο σημαντικότητας $a = 0.05$;

Κεφάλαιο 10

Γραμμική παλινδρόμηση

10.1 Απλή γραμμική παλινδρόμηση

Δίνεται ένα σύνολο n ζευγών (x_i, Y_i) από τιμές των μεταβλητών x (input) και Y (response), που παρατηρήθηκαν σε δείγμα μεγέθους n, και ζητείται να προσδιορισθεί κατά πόσο οι μεταβλητές x, Y έχουν γραμμική σχέση. Υποθέτοντας ότι υπάρχει γραμμική σχέση, δηλαδή ότι

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i \in \{1, 2, \ldots, n\},$$

για κάποιες (άγνωστες) σταθερές β_0, β_1, με τη μεταβλητή ε_i να αντιπροσωπεύει το τυχαίο σφάλμα που εξαρτάται από οποιουσδήποτε άλλους παράγοντες πλην της τιμής του x_i και προκαλεί απόκλιση από την τέλεια γραμμική σχέση, η σχέση αυτή εκφράζεται από την εξίσωση

$$Y = \beta_0 + \beta_1 x + \varepsilon, \tag{10.1}$$

η οποία αποτελεί το **μοντέλο της γραμμικής παλινδρόμησης**. Η μεταβλητή ε θεωρείται ως μια ΤΜ με $E(\varepsilon) = 0$. Μια τιμή της x μπορεί να αντιστοιχεί σε διαφορετικές Y με διαφορετικά σφάλματα, οπότε, η Y θεωρείται ως ΤΜ, ενώ η x δεν θεωρείται ως ΤΜ Επομένως, το μοντέλο εκφράζεται ισοδύναμα από τη σχέση

$$E(Y|x) = \beta_0 + \beta_1 x, \tag{10.2}$$

δηλαδή από μια ευθεία, η οποία ονομάζεται (πληθυσμιακή) ευθεία παλινδρόμησης. Το μοντέλο προσδιορίζεται πλήρως από τις τιμές των συντελεστών παλινδρόμησης β_0 (intercept) και β_1 (slope).

Εκτίμηση των συντελεστών παλινδρόμησης με τη μέθοδο ελαχίστων τετραγώνων: Έστω $\widehat{Y} = B_0 + B_1 x$, όπου τα B_0, B_1 αποτελούν εκτιμήσεις των β_0, β_1 αντίστοιχα. Για τον υπολογισμό των B_0, B_1, χρησιμοποιείται συνήθως η μέθοδος ελαχίστων τετραγώνων, η οποία αποσκοπεί στην ελαχιστοποίηση της συνάρτησης (sum of squared residuals)

$$SSR = SSR(B_0, B_1) = \sum_{i=1}^{n} (Y_i - \widehat{Y_i})^2 = \sum_{i=1}^{n} (Y_i - B_0 + B_1 x_i)^2.$$

Λαμβάνοντας υπόψη τις μερικές παραγώγους της SSR ως προς τις μεταβλητές B_0, B_1, προκύπτει τελικά ότι η SSR λαμβάνει την ελάχιστη τιμή της όταν

$$B_0 = \overline{Y} - B_1 \overline{x}, \qquad B_1 = \frac{\sum_{i=1}^{n} x_i Y_i - n\overline{x}\,\overline{Y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}, \qquad \text{όπου } \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i. \tag{10.3}$$

Θέτοντας για απλότητα

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2, \qquad S_{YY} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2,$$

$$S_{xY} = \sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y}) = \sum_{i=1}^{n} x_i Y_i - n\overline{x}\overline{Y},$$

τότε η σχέση (10.3) γράφεται ως

$$B_0 = \overline{Y} - B_1\overline{x}, \qquad B_1 = \frac{S_{xY}}{S_{xx}}, \qquad SSR = S_{YY} - B_1 S_{xY}. \tag{10.4}$$

Κατόπιν τούτων, η ευθεία $y = B_0 + B_1 x$ αποτελεί μια εκτίμηση της γραμμικής σχέσης των x, Y και ονομάζεται εκτιμώμενη ευθεία παλινδρόμησης. Οι συντελεστές B_0, B_1 υπολογίζονται πλήρως από τη σχέση (10.4). Η τιμή $\widehat{Y} = B_0 + B_1 x$ αποτελεί μια εκτίμηση της υπο συνθήκη αναμενόμενης τιμής $E(Y|x)$.

Κατανομές των εκτιμητριών: Για να εκτιμήσουμε πόσο καλό είναι το μοντέλο, κάνουμε την παραδοχή ότι τα σφάλματα ε_i είναι ανεξάρτητα μεταξύ τους και ακολουθούν κανονική κατανομή $\mathcal{N}(0, \sigma^2)$ με άγνωστη αλλά κοινή διακύμανση. Τότε, άμεσα προκύπτει ότι

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2), \qquad B_1 \sim \mathcal{N}(\beta_1, \sigma^2/S_{xx}), \qquad B_0 \sim \mathcal{N}\left(\beta_0, \frac{\sigma^2 \sum x_i^2}{nS_{xx}}\right), \qquad \frac{SSR}{\sigma^2} \sim \chi^2_{n-2}$$

οπότε οι

$$\frac{SSR}{n-2}, \qquad S^2(B_0) = \frac{SSR \sum x_i^2}{(n-2)nS_{xx}}, \qquad S^2(B_1) = \frac{SSR}{(n-2)S_{xx}}$$

αποτελούν αμερόληπτες εκτιμήτριες των σ^2 και των διακυμάνσεων των B_0 και B_1 αντίστοιχα.

Διαστήματα εμπιστοσύνης: Τα $(1-a)100\%$ δ. ε. για τα β_0, β_1 είναι αντίστοιχα τα

$$B_0 \pm t_{a/2,n-2} S(B_0), \qquad B_1 \pm t_{a/2,n-2} S(B_1),$$

διότι

$$T_0 = \frac{B_0 - \beta_0}{S(B_0)} \sim t_{n-2}, \qquad T_1 = \frac{B_1 - \beta_1}{S(B_1)} \sim t_{n-2}.$$

Έλεγχοι υποθέσεων: Οι παραπάνω στατιστικές συναρτήσεις T_0, T_1 μπορούν να χρησιμοποιηθούν για την κατασκευή στατιστικών ελέγχων σχετικά με τα β_0, β_1. Ένας σημαντικός έλεγχος είναι αυτός της υπόθεσης $H_0 : \beta_1 = 0$, έναντι της $H_1 : \beta_1 \neq 0$, διότι $\beta_1 = 0$ σημαίνει ότι οι τιμές της Y δεν εξαρτώνται από την x, οπότε, όταν η H_0 απορρίπτεται, η γραμμική σχέση των x, Y θεωρείται στατιστικά σημαντική. Η στατιστική συνάρτηση του ελέγχου είναι η $T_1 = \dfrac{B_1}{S(B_1)}$, η οποία υπό την H_0 ακολουθεί την t_{n-2}, οπότε σύμφωνα με τα γνωστά είναι p-value $= 2F_{n-2}(-|T_1|)$, όπου F_{n-2} η συνάρτηση κατανομής της t_{n-2}.

Συντελεστής συσχέτισης: Στην ταυτότητα

$$S_{YY} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 + \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 = SSR + \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$$

η ποσότητα S_{YY} του πρώτου μέλους εκφράζει τη διακύμανση των τιμών της Y. Η διακύμανση αυτή οφείλεται στη διακύμανση των σφαλμάτων ε_i, η οποία αποδίδεται από τον όρο SSR, και στη διακύμανση των τιμών x_i, η οποία αποδίδεται από τον δεύτερο όρο του δεύτερου μέλους. Ο **συντελεστής προσδιορισμού** R^2 ορίζεται ως

$$R^2 = \frac{S_{YY} - SSR}{S_{YY}} = 1 - \frac{SSR}{S_{YY}} = \frac{B_1 S_{xY}}{S_{YY}} = \frac{S_{xY}^2}{S_{xx}S_{YY}}$$

και αντιπροσωπεύει το ποσοστό διακύμανσης που οφείλεται μόνο στην x (και όχι στην ε). Προφανώς, $R^2 \in [0,1]$, όπου μια τιμή της R^2 κοντά στο 1 υποδεικνύει ότι το γραμμικό μοντέλο ταιριάζει πολύ καλά στα δεδομένα, ενώ μια τιμή κοντά στο 0 υποδεικνύει το αντίθετο.

10.2 Πολυμεταβλητή γραμμική παλινδρόμηση

Σε αυτή την περίπτωση η ΤΜ Y_i εξαρτάται από ένα διάνυσμα μεταβλητών $(x_1, x_2, \ldots, x_p)$, δηλαδή το γραμμικό μοντέλο είναι το

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon = \beta^T \mathbf{x} + \varepsilon \tag{10.5}$$

όπου $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ και $\mathbf{x} = (1, x_1, \ldots, x_n)$. Η μεταβλητή $x_0 = 1$ απλώς βοηθά να εκφράσουμε το μοντέλο στην τελευταία διανυσματική μορφή.

Όπως και πριν, για την εκτίμηση του διανύσματος β, χρησιμοποιώντας τη μέθοδο ελαχίστων τετραγώνων, θέτοντας $B = (B_0, B_1, \ldots, B_p)$ το διάνυσμα που εκτιμά το β και $\widehat{Y} = B^T\mathbf{x}$, προκύπτει ότι η συνάρτηση SSR (η οποία τώρα είναι συνάρτηση του B) ελαχιστοποιείται όταν

$$X^T X B = XY \tag{10.6}$$

όπου X $n \times (p+1)$ πίνακας με $j+1$-οστή στήλη το δείγμα της μεταβλητής x_j, $0 \le j \le p$.

Αν ο πίνακας $X^T X$ αντιστρέφεται, τότε υπάρχει μοναδική λύση της παραπάνω εξίσωσης, η οποία είναι η

$$B = (X^T X)^{-1} XY.$$

Στον παρακάτω κώδικα σε R, το διάνυσμα x καταγράφει το ύψος σε ένα δείγμα ανδρών ενώ το y το ύψος των γιών τους.

```
x = c(152.4, 157.5, 162.5, 165, 167.6, 171, 173, 178, 183, 188)
y = c(161.5, 165.6, 167.6, 166.4, 170, 170.5, 171.2, 173.5, 178, 178)
height = lm(y ~ x)
summary(height)
plot(x,y)
abline(height)
```

Output:

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5957 -0.6097 -0.4041  0.7875  1.6257

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 91.19160    5.43816   16.77 1.62e-07 ***
x            0.46548    0.03196   14.56 4.85e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.07 on 8 degrees of freedom
Multiple R-squared:  0.9636,  Adjusted R-squared:  0.9591
F-statistic: 212.1 on 1 and 8 DF,  p-value: 4.847e-07
```

Ο αντίστοιχος κώδικας σε Matlab φαίνεται παρακάτω

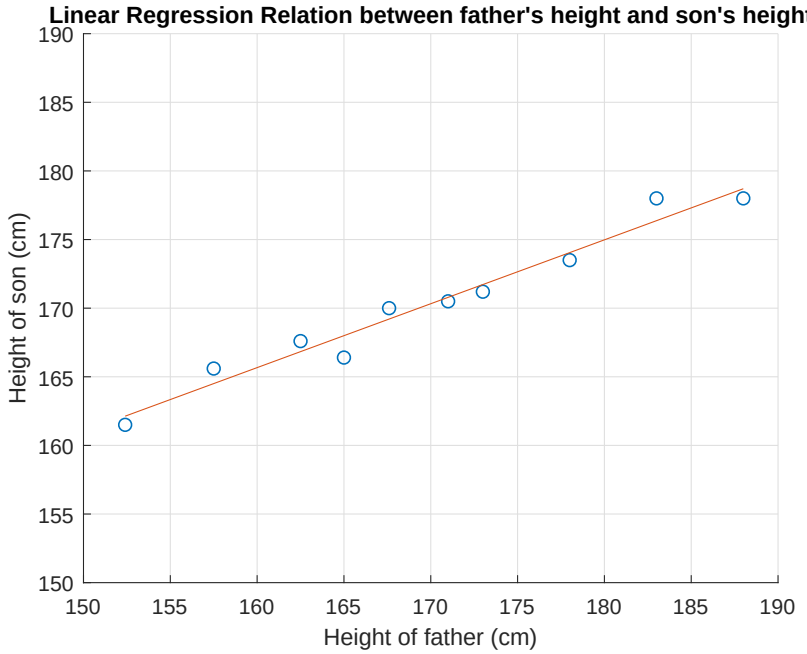
```
%-----------------------create data----------------------
%father = [152.4, 157.5, 162.5, 165, 167.6, 171,  173, 178, 183, 188].'; %transpose
%son = [161.5, 165.6, 167.6, 166.4, 170, 170.5, 171.2, 173.5, 178, 178].';
%save('heights.mat', 'father', 'son')
%--------------------------------------------------------
tbl = table(father,son,'VariableNames',{'Father','Son'});
mdl = fitlm(tbl) %fitlm(x,Y) %the last column is always the response variable
```

Output:

```
Linear regression model:
    Son ~ 1 + Father

Estimated Coefficients:
                   Estimate       SE       tStat       pValue
                   ________    ________    ______    __________

    (Intercept)     91.192       5.4382    16.769    1.6193e-07
    Father         0.46548     0.031965    14.562    4.8471e-07


Number of observations: 10, Error degrees of freedom: 8
Root Mean Squared Error: 1.07
R-squared: 0.964,  Adjusted R-Squared 0.959
F-statistic vs. constant model: 212, p-value = 4.85e-07
```

Οι συντελεστές παλινδρόμησης και το R^2 μπορούν επίσης να υπολογισθούν απλά και ως εξής:

```
load('heights.mat', 'father', 'son');
x = father;
X = [ones(length(x),1) x]; %prepend a column of ones
Y = son;

b = X\Y %calculate the coefficients vector of the linear regression model

Yhat = X*b;
R2 = 1 - sum((Y - Yhat).^2)/sum((Y - mean(Y)).^2)
scatter(X(:,2),Y)
hold on
plot(X(:,2),Yhat)
xlabel('Height of father (cm)')
ylabel('Height of son (cm)')
title('Linear Regression Relation between father''s height and son''s height')
grid on
```



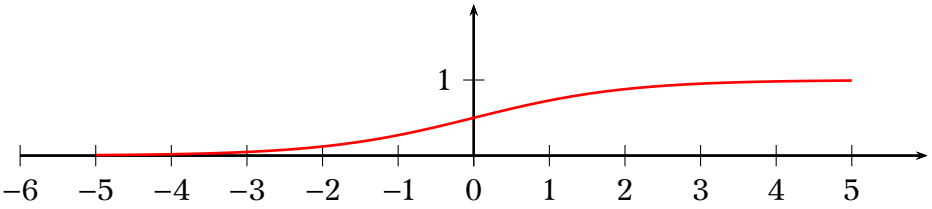
Επίσης, υπάρχει και η συνάρτηση regress(y,x):

```
load('heights.mat', 'father', 'son');
x = father;
X = [ones(length(x),1) x]; %prepend a column of ones
Y = son;
alpha = 0.05;
[b,bint,r,rint,stats] = regress(Y,X, alpha) %default alpha=0.05
%b: coefficients, bint: CI for coefficients
%r: vector of residuals, rint: residuals intervals
%stats: R^2, F, p-value, and MSE
```

10.3 Λογιστική Παλινδρόμηση (Logistic Regression)

Η μέθοδος αυτή χρησιμοποιείται σε προβλήματα ταξινόμησης (classification), μοντελοποιώντας την πιθανότητα $P(C = i|\mathbf{x})$ μια παρατήρηση $\mathbf{x} = (x_0, x_1, x_2, \ldots, x_p)$ να ανήκει στην κλάση i, όπου $i \in [K] = \{1, 2, \ldots, K\}$, από την λογιστική συνάρτηση. Η μεταβλητή x_0 είναι βοηθητική και πάντα ίση με 1. Η λογιστική συνάρτηση είναι η συνάρτηση

$$f : (-\infty, +\infty) \to (0, 1), \qquad \text{με } f(x) = \frac{e^x}{1 + e^x} = \frac{1}{1 + e^{-x}}.$$



Η αντίστροφή της ονομάζεται συνάρτηση logit και έχει τύπο $f^{-1}(x) = \ln \frac{x}{1 - x}$.

Συγκεκριμένα, το μοντέλο αυτό υποθέτει ότι για κάθε $i \in [K-1]$ υπάρχει ένα διάνυσμα σταθερών $B_i = (\beta_{i,0}, \beta_{i,1}, \ldots, \beta_{i,p})$, ώστε να ισχύει η σχέση

$$\ln \frac{P(C = i|\mathbf{x})}{P(C = K|\mathbf{x})} = \beta_{i,0} + \beta_{i,1}x_1 + \cdots + \beta_{i,p}x_p = B_i^T\mathbf{x}, \qquad i \in [K-1],$$

οπότε, προκύπτουν (μετά από πράξεις) οι πιθανότητες

$$P(C = i|\mathbf{x}) = \frac{e^{B_i^T\mathbf{x}}}{1 + \sum_{j=1}^{K-1} e^{B_j^T\mathbf{x}}} = \frac{e^{B_i^T\mathbf{x}}}{\sum_{j=1}^{K} e^{B_j^T\mathbf{x}}} = F_i(B_1^T\mathbf{x}, \ldots, B_K^T\mathbf{x}), \quad i \in [K], \qquad \text{όπου} \qquad B_k = 0$$

και $F_i(x_1, x_2, \ldots, x_K) = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}}$ η συνάρτηση softmax, που γενικεύει τη λογιστική συνάρτηση για πολλές μεταβλητές. Οι πιθανότητες αυτές προφανώς αθροίζουν στο 1 και επιπλέον υπολογίζονται εύκολα από τον τελευταίο τύπο. Ειδικά στην περίπτωση όπου $K = 2$ (binary classification), ο τελευταίος τύπος γίνεται

$$P(C = 1|\mathbf{x}) = f(B_1^T\mathbf{x}), \qquad P(C = 2|\mathbf{x}) = 1 - f(B_1^T\mathbf{x}).$$

Η συνάρτηση του Matlab για την λογιστική παλινδρόμηση είναι η mnrfit().

```
x1 = randn(50,1); %random values from standard normal distribution
x2 = randn(50,1);
y = (2*x1 + x2 + randn(size(x1))) > 1;
X = [x1 x2];
[B,dev,stats] = mnrfit(X,y+1);
logisticfun = @(x) 1./(1+exp(-x));
%compute the output (class assignment) of the model for each data point
modelfit = logisticfun([ones(50,1) X]*B) < 0.5; %if <0.5 assign to class y=1
%percentage of correctly classified points by the model
pctcorrect = sum(modelfit==y) / length(y)
% visualize the data
figure;
hold on;
h1 = scatter(x1(y==0),x2(y==0),50,'k','filled');  % black dots for 0
h2 = scatter(x1(y==1),x2(y==1),50,'w','filled');  % white dots for 1
set([h1 h2],'MarkerEdgeColor',[.5 .5 .5]);        % outline dots in gray
legend([h1 h2],{'y==0' 'y==1'},'Location','NorthEastOutside');
xlabel('x_1');ylabel('x_2');
```

Βιβλιογραφία

[1] Δ. Αθανασόπουλος, *Θεωρία πιθανοτήτων, Μέρος ΙΙ: Κατανομές πιθανότητας τυχαίων μεταβλητών*, Εκδόσεις Σταμούλη, 1991.

[2] Φ. Γεωργιακώδης, Ι. Τριανταφύλλου, *Στοιχεία πιθανοτήτων και στατιστικής στην επιστήμη των υπολογιστών*, Εκδόσεις Σταμούλη, 2011.

[3] Χ. Δαμιανού, Ν. Παπαδάτος, Χ. Α. Χαραλαμπίδης, *Εισαγωγή στις πιθανότητες και τη στατιστική (Διδακτικές σημειώσεις)*, Πανεπιστήμιο Αθηνών, Αθήνα, 2003

[4] Χ. Ευαγγελάρας, *Σημειώσεις πιθανοτήτων και στατιστικής*, Λαμία, 2010.

[5] Ι. Κοντογιάννης, Σ. Τουμπής, *Στοιχεία πιθανοτήτων με εφαρμογές στη στατιστική και την πληροφορική*, ΣΕΑΒ, Αθήνα, 2015.

[6] Δ. Μπερτσεκάς, Γ. Τσιτσικλής, *Εισαγωγή στις πιθανότητες με στοιχεία στατιστικής*, Εκδόσεις Τζιόλα, 2018.

[7] Γ. Πετράκος, *Εισαγωγή στη στατιστική της καθημερινότητάς μας*, Εκδόσεις Σταμούλη, 2011.

[8] M. R. Spiegel, *Πιθανότητες και στατιστική*, ΕΣΠΙ, 1977.

[9] Θ. Ξένος, *Πιθανότητες*, Εκδόσεις Ζήτη, 2012.

[10] Ε. Φούντας, Κ. Πατσάκης, Χ. Φούντας, *Πιθανότητες - Στατιστική & Εφαρμογές*, Εκδόσεις Βαρβαρήγου, 2015.

[11] P. Brémaud, *Discrete probability models and methods*, Springer, 2017.

[12] B. Christian, T. Griffiths, *Algorithms to live by*, Henry Holt and Company, 2016.

[13] V. A. Dobrushkin, *Methods in algorithmic analysis*, CRC Press, 2010.

[14] H. W. Gould, *Combinatorial identities*, Morgantown, 1972.

[15] R. L. Graham, D. E. Knuth, O. Patashnik, *Concrete mathematics*, 2nd edition, Addison-Wesley, 1994.

[16] S. J. Miller, *The probability lifesaver: All the tools you need to understand chance*, Princeton university press, 2017.

[17] F. Mosteller, *Fifty challenging problems in probability with solutions*, Dover publications, 1987.

[18] D. Stirzaker, *Elementary probability*, 2nd edition, Cambridge University Press, 2003.

[19] A. A. Sveshnikov, *Problems in probability theory, mathematical statistics and theory of random functions*, Dover publications, 1978.

[20] M. M. Woolfson, *Everyday probability and statistics: Health, elections, gambling and war*, Imperial College Press, 2008.