Χρήση ανιχνευτή Geiger-Mueller για τη στατιστική ανάλυση δεδομένων (Π4)

Φ. Διάκονος, Κ. Θεοφιλάτος, Α. Καπόγιαννης, Ι. Τσοχαντζής

Περίληψη

Στο πείραμα 4 θα μελετήσουμε τον αριθμό διασπάσεων ($k = 0, 1, 2, ...$) ραδιενεργού πηγής που μετρά ανιχνευτής Geiger-Müller σε προκαθορισμένο χρονικό διάστημα καταγραφής ($\Delta t = 2s$). Κρατώντας την απόσταση της πηγής από τον ανιχνευτή σταθερή, θα λάβουμε N γεγονότα καταγραφής ($k_1, k_2, ...k_N$). Στην συνέχεια θα μελετήσουμε συναρτήσει του k το πλήθος των γεγονότων (n_k) για τα οποία ανιχνεύτηκαν k ακριβώς διασπάσεις. Οι πειραματικές μετρήσεις (n_k) θα συγκριθούν με τις θεωρητικά αναμενόμενες (f_k) και θα ποσοτικοποιηθεί η συμφωνία τους. Έλεγχος διαφορετικών θεωρητικών μοντέλων θα γίνει χρησιμοποιώντας το κριτήριο χ^2. Η μετρητική διαδικασία και ο έλεγχος των θεωρητικών μοντέλων θα εφαρμοστεί για 2 διαφορετικές θέσεις της πηγής (κοντά και μακριά από τον ανιχνευτή).

Περιεχόμενα

1 Πριν το εργαστήριο 2

2 Πειραματική διαδικασία 2

3 Επεξεργασία των μετρήσεων και σύνταξη της αναφοράς 3

4 Θεωρία ραδιενεργών διασπάσεων 7

5 Ιστογράμματα και κατανομές 8

6 Θεωρητικά μοντέλα και έλεγχος υποθέσεων 10

7 Περαιτέρω εργασίες 14

8 Ιστογράμματα στην Python 14

1 Πριν το εργαστήριο

Για την προετοιμασία σας πριν το εργαστήριο πρέπει:

- Να διαβάσετε το θεωρητικό μέρος του εργαστηριακού οδηγού συγκεκριμένα των κεφαλαίων Α, Β, Δ, Ε (κοινά για όλες τις ασκήσεις) καθώς και του Γ1 που αφορά τον ανιχνευτή Geiger-Mueller.
- Να είστε σε θέση να σχεδιάσετε ένα ιστόγραμμα, δοσμένου κάποιου δείγματος μετρήσεων καθώς και να υπολογίσετε τον μέσο όρο και την αβεβαιότητά του.
- Να γνωρίζετε το κεντρικό οριακό θεώρημα.
- Να γνωρίζετε τις βασικές κατανομές: διωνυμική, Poisson, κανονική και χ^2.
- Να γνωρίζετε την θεωρία του ελέγχου υποθέσεων.
- Να έχετε μαζί σας ηλεκτρονικό αποθηκευτικό μέσο τύπου usb stick στο οποίο θα αποθηκεύσετε τα δεδομένα που θα επεξεργαστείτε στο σπίτι.

Μπορείτε να βρείτε τον εργαστηριακό οδηγό στον φάκελο έγγραφα στο e-class.

2 Πειραματική διαδικασία

Η άσκηση περιλαμβάνει την λήψη δύο ξεχωριστών σειρών από πειραματικές μετρήσεις, ένα σε υψηλό ρυθμό ($k \sim 30$ διασπάσεις/Δt) και ένα σε χαμηλό ρυθμό ($k \sim 3$ διασπάσεις/Δt). Τα δεδομένα θα ληφθούν κάνοντας χρήση του ανιχνευτή Geiger-Mueller[1].

Για την λήψη της πρώτης, ρυθμίστε το προκαθορισμένο χρονικό διάστημα καταγραφής διασπάσεων στη τιμή $\Delta t = 2s$ και τοποθετήστε την πηγή σε κατάλληλη απόσταση από τον ανιχνευτή ώστε αυτός να καταγράφει $k \sim 30$ διασπάσεις/Δt. Χωρίς να μεταβάλλετε την θέση της πηγής πάρτε $N = 300$ μετρήσεις $(k_1, k_2, ...k_N)$. Γράψτε τις μετρήσεις σας σε αρχείο απλού κειμένου, με κάθε μέτρηση k_i να καταλαμβάνει μια γραμμή, όπως στο παρακάτω υπόδειγμα.

```
33
41
28
17
38
20
31
.
.
.
```

[1] Η λειτουργία του ανιχνευτή βασίζεται στην ανίχνευση του ηλεκτρικού φορτίου που παράγεται λόγω ιονισμού, όταν κάποιο φορτισμένο σωματίδιο διέρχεται μέσα από τον όγκο ενός αδρανούς αερίου (π.χ. He). Βρείτε περισσότερες πληροφορίες για την λειτουργία του ανιχνευτή, εντός του εργαστηριακού οδηγού στο e-class.

Για την λήψη της δεύτερης σειράς δεδομένων, απομακρύνουμε την πηγή σε απόσταση ικανή ώστε ο ρυθμός καταγραφής να πέσει κατά μέσο όρο $k \sim 2$ διασπάσεις $/\Delta t$. Στατιστικώς, θα υπάρχουν αρκετές μετρήσεις με $k = 0$ καθώς και γεγονότα στα οποία ο ανιχνευτής κατέγραψε $k > 2$. Για την εύρεση της κατάλληλης απόστασης ανιχνευτή–πηγής, είναι σκόπιμο να λάβετε ~ 10 μετρήσεις του k και να ελέγξετε αν ο μέσος όρος αυτών είναι ~ 2. (Οι μετρήσεις με $k = 0$ είναι εξίσου σημαντικές και πρέπει να προσμετρηθούν στον μέσο όρο.) Όταν ολοκληρώσετε την ρύθμιση της απόστασης ανιχνευτή–πηγής ώστε ο ανιχνευτής να καταγράφει κατά μέσο όρο $k \sim 2$ διασπάσεις $/\Delta t$, πάρτε τουλάχιστον $N = 300$ μετρήσεις χωρίς να μεταβάλετε περαιτέρω την πειραματική διάταξη. Αποθηκεύεστε τις μετρήσεις σας σε ηλεκτρονικό αρχείο απλού κειμένου (*.txt).

3 Επεξεργασία των μετρήσεων και σύνταξη της αναφοράς

Στην θεωρητική εισαγωγή της αναφοράς που θα γράψετε, εξηγήστε, με δικά σας λόγια, γιατί αναμένουμε ότι ο αριθμός των διασπάσεων ανά μονάδα χρόνου που καταγράφει ο ανιχνευτής θα ακολουθεί την κατανομή Poisson. Είναι ελέγξιμη πειραματικά η υπόθεση αυτή ή αποτελεί μια παραδοχή που κάνουμε στην παρούσα άσκηση;

Έπειτα, για κάθε μία από τις δύο σειρές μετρήσεων που λάβατε, ακολουθείστε τα παρακάτω βήματα:

- Βρείτε τον μέσο όρο του αριθμού διασπάσεων

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} k_i$$

  και εκτιμήστε την αβεβαιότητά του $\delta\hat{\mu}$. Εξηγήστε, γιατί περιμένουμε το $\hat{\mu}$ να ακολουθεί κατά προσέγγιση κανονική κατανομή.

- Θεωρήστε δυο θεωρητικά μοντέλα για την πρόβλεψη του n_k

$$f_k^{\text{poisson}} = N p_{\text{poisson}}(k; \mu = \hat{\mu}) = N\frac{\hat{\mu}^k}{k!}e^{-\hat{\mu}} \tag{1}$$

$$f_k^{normal} = N p_{\text{normal}}(k; \mu = \hat{\mu}, \sigma = \sqrt{\hat{\mu}}) = N\frac{1}{\sqrt{2\pi\hat{\mu}}}e^{-\frac{(k-\hat{\mu})^2}{2\hat{\mu}}}, \tag{2}$$

  όπου με f_k συμβολίζεται η θεωρητική συχνότητα για το κάθε μοντέλο, ενώ p_{poisson}, p_{normal} είναι οι κατανομές Poisson και Gauss αντίστοιχα.

- Φτιάξτε πίνακα βάσει του υποδείγματος που δίνεται στον πίνακα 2. Περιορίστε το $k_{\max}$ ώστε να συμπεριλάβετε όλα τα πειραματικά δεδομένα αποκόπτοντας τιμές του k για τις οποίες ισχύει ταυτόχρονα ότι $n_k = 0$ και $f_k < 0.01$

- Υπολογίστε το χ^2 για κάθε μοντέλο λαμβάνοντας υπόψιν μόνο τις περιπτώσεις που αναμένουμε θεωρητικά $f_k > 20$. Στα δεδομένα υψηλού ρυθμού χαλαρώστε την απαίτηση σε $f_k > 5$, καθώς έχουμε μεγαλύτερη διασπορά τιμών. Εξηγήστε γιατί η

αποκοπή δεδομένων με χαμηλό f_k είναι καλή ιδέα και πως αυτό θα μπορούσε να αποφευχθεί οργανώνοντας τα δεδομένα διαφορετικά, π.χ., χρησιμοποιώντας διαφορετικό μέγεθος ιστών (bins) στο ιστόγραμμα.

- Υπολογίστε τους βαθμούς ελευθέριας ν για κάθε μοντέλο λαμβάνοντας υπόψιν μόνο τα n_k για τα οποία περιμένουμε $f_k > 20$. Στα δεδομένα υψηλού ρυθμού χαλαρώστε την απαίτηση σε $f_k > 5$, καθώς έχουμε μεγαλύτερη διασπορά τιμών.
- Υπολογίστε την πιθανότητα p να μετρήσουμε χ^2 μεγαλύτερο από αυτό που βρήκαμε στο εργαστήριο, στην περίπτωση που το μοντέλο μας είναι ορθό. Η πιθανότητα αυτή δίνεται από το ολοκλήρωμα

$$p = \int_{\chi^2}^{\infty} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} t^{\nu/2-1} e^{-t/2} dt = 1 - \int_{0}^{\chi^2} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} t^{\nu/2-1} e^{-t/2} dt.$$

Για τον υπολογισμό της τιμής p μπορείτε είτε να χρησιμοποιήσετε την ιστοσελίδα του Matt Bognar είτε (καλύτερα) να χρησιμοποιήσετε τις ακόλουθες γραμμές κώδικα python στο πρόγραμμα της ανάλυσης που θα συντάξετε για τις ανάγκες αυτής της άσκησης.

```
from scipy.stats import chi2
x = 20 # chi2
v = 10 # degrees of freedom
p = 1 - chi2.cdf(x, v)
```

Ακόμη καλύτερο θα ήταν να κατασκευάσετε την κατανομή που περιμένετε για την χ^2 (σχήμα 2), γράφοντας ένα πρόγραμμα Monte–Carlo προσομοίωσης του παρόντος πειράματος θεωρώντας ότι οι μετρήσεις ακολουθούν την κατανομή Poisson με $\mu = \hat{\mu}$.

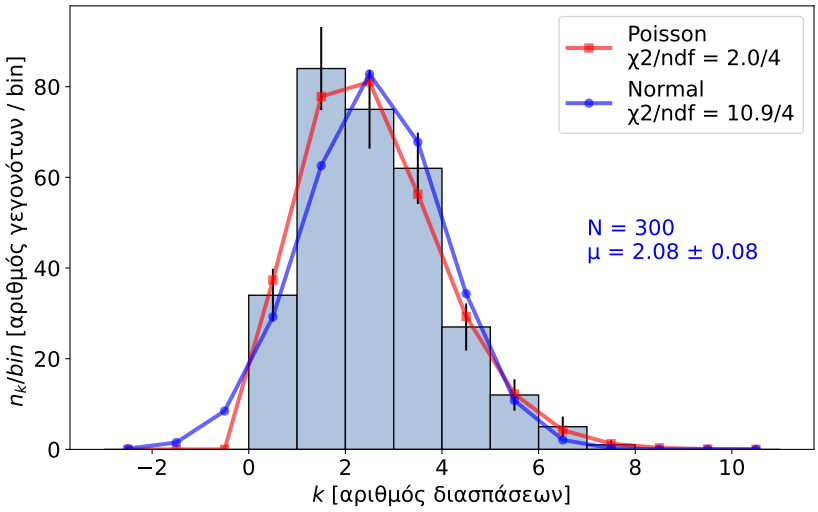
- Φτιάξτε ιστόγραμμα που να συνοψίζει όλα τα αποτελέσματα της εργασίας σας σε δυο διαγράμματα (ένα για κάθε σειρά μετρήσεων), παρόμοια με αυτό του σχήματος 1.
- Συμπληρώστε τον πίνακα αποτελεσμάτων βάσει του υποδείγματος του πίνακα 2, για κάθε μία από τις δύο σειρές μετρήσεων.
- Συντάξτε την αναφορά σας σε ηλεκτρονικό υπολογιστή, χρησιμοποιώντας LaTeX (π.χ. online αν δεν το έχετε στον Η/Υ) ή οποιοδήποτε άλλο πρόγραμμα έχετε ήδη εγκατεστημένο στον Η/Υ σας.
- Επισυνάψτε μαζί με την ηλεκτρονική σας αναφορά, τα δεδομένα που πήρατε στο εργαστήριο σε αρχεία απλού κειμένου (lowmean.txt και highmean.txt).

k	n_k	f_k^{poisson}	f_k^{normal}
0	34	37.4	29.3
1	84	77.8	62.6
2	75	81.1	82.8
3	62	56.3	67.8
4	27	29.3	34.3
5	12	12.2	10.8
6	5	4.2	2.1
7	1	1.3	0.3
8	0	0.3	0.0
9	0	0.1	0.0
10	0	0.0	0.0

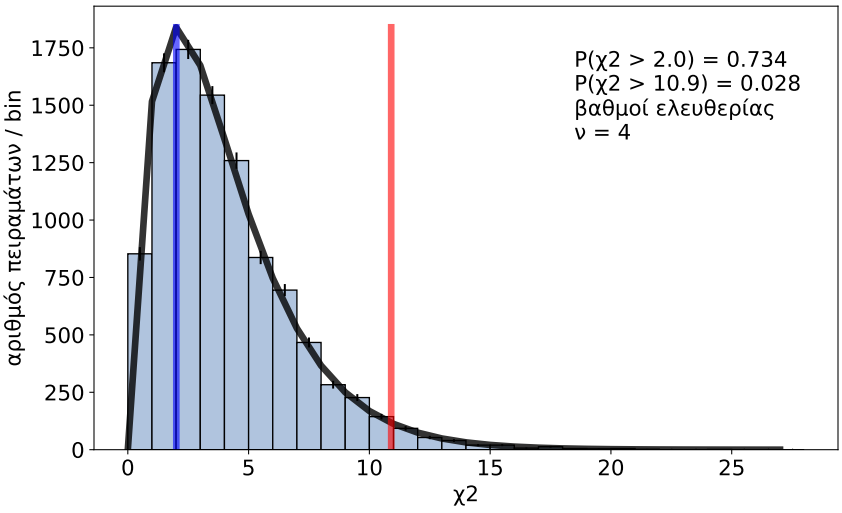
Πίνακας 1: Υπόδειγμα πίνακα μετρήσεων και θεωρητικών προβλέψεων. Ο συμβολισμός διαβάζεται ως έξης, n_k = *πλήθος γεγονότων καταγραφής* που διαπιστώσαμε ότι ο ανιχνευτής μέτρησε k ακριβώς διασπάσεις σε χρόνο Δt. (Δηλαδή k είναι η ένδειξη που βλέπουμε στον ανιχνευτή και n_k πόσες φορές 'είδαμε' την ίδια ένδειξη k στο σύνολο των N γεγονότων καταγραφής που έχουμε στα πειραματικά δεδομένα.) Τα ίδια δεδομένα μπορούν να οπτικοποιηθούν κάνοντας χρήση ενός ιστογράμματος (σχήμα 1).

	Poisson	Normal
$\chi 2$	2.007	10.903
ν	4	4
p	73.4 %	2.8%

Πίνακας 2: Υπόδειγμα πίνακα που συνοψίζει τα αποτελέσματα του έλεγχου των υποθέσεων για το δείγμα δεδομένων χαμηλού ρυθμού με $\hat{\mu} = 2.08 \pm 0.08$.



Σχήμα 1: Υπόδειγμα ιστογράμματος που συνοψίζει τον πίνακα 1.



Σχήμα 2: Ιστόγραμμα που απεικονίζει την κατανομή του χ^2 σε 10^4 ανεξάρτητα ψευδοπειράματα στα οποία έχουμε προσομοιώσει με Monte–Carlo την παρούσα εργαστηριακή άσκηση. Για την παραγωγή των ψευδοδεδομένων χρησιμοποιήθηκε γεννήτρια τυχαίων αριθμών με $P(k) = (2.083^k/k!)e^{-2.083}$ η οποία παρήγαγε ένα δείγμα από $N = 300$ ψευδογεγονότα $(k_1, k_2, ...k_N)$ διασπάσεων ραδιενεργής πηγής για καθένα από τα 10^4 ψευδοπειράματα. Η μπλε γραμμή αντιστοιχεί στο $\chi^2_{\text{Poisson}} = 2.0$ ενώ η κόκκινη στο $\chi^2_{\text{normal}} = 10.9$.

4 Θεωρία ραδιενεργών διασπάσεων

Έστω ότι η πιθανότητα διάσπασης ενός ραδιενεργού πυρήνα συγκεκριμένου είδους σε χρονικό διάστημα Δt είναι γνωστή και ίση με $\lambda \Delta t$ όπου λ είναι ο μέσος ρυθμός διάσπασης που χαρακτηρίζει αυτό το είδος των πυρήνων και καλείται **σταθερά διάσπασης**. Αντίστοιχα η πιθανότητα να μην διασπαστεί ένας πυρήνας στο διάστημα Δt θα είναι $1-\lambda\Delta t$. Αν θεωρήσουμε ότι σε μια συλλογή από N πυρήνες κάθε πυρήνας διασπάται ανεξάρτητα από τους άλλους τότε περιμένει κανείς η πιθανότητα να έχουν διασπασθεί n πυρήνες στο διάστημα Δt να δίνεται από την διωνυμική κατανομή:

$$P\left(n, \Delta t\right) = \frac{N!}{(N-n)!n!}(\lambda\Delta t)^n(1-\lambda\Delta t)^{N-n} \tag{3}$$

Όταν $N \gg 1$ και $\lambda\Delta t \ll 1$ με $N\lambda\Delta t =$ σταθερό η κατανομή αυτή τείνει στην κατανομή Poisson:

$$P\left(n, \Delta t\right) = \frac{\mu^n}{n!}e^{-\mu} \tag{4}$$

όπου $\mu = \lambda N \Delta t$ είναι ο μέσος αριθμός διασπάσεων στο διάστημα Δt. Όταν επιπλέον ισχύει $\mu \gg 1$ η κατανομή αυτή τείνει στην κανονική (Gaussian):

$$P\left(n, \Delta t\right) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-(n-\mu)^2}{2\sigma^2}} \tag{5}$$

με διασπορά $\sigma^2 = \mu$.

Ας θεωρήσουμε τώρα τη μεταβολή του πληθυσμού των ραδιενεργών πυρήνων στο χρονικό διάστημα $[t, t+\Delta t]$. Έστω $N(t)$ το πλήθος των αδιάσπαστων πυρήνων τη χρονική στιγμή t. Ο αριθμός των αδιάσπαστων πυρήνων τη χρονική στιγμή $t+\Delta t$ θα δίνεται από τη σχέση: $N(t+\Delta t) = N(t) - \mu$ όπου, όπως προαναφέραμε, $\mu = \lambda N(t)\Delta t$ είναι ο μέσος αριθμός πυρήνων που διασπάστηκαν στο διάστημα $[t, t+\Delta t]$. Θα ισχύει λοιπόν:

$$\frac{N(t+\Delta t) - N(t)}{\Delta t} = -\lambda N(t) \tag{6}$$

και παίρνοντας το όριο $\Delta t \to 0$ καταλήγουμε στη σχέση:

$$-\frac{dN}{dt} = \lambda N \tag{7}$$

όπου λN είναι η **ενεργότητα** της πηγής. Η (7) αποτελεί μια συνήθη διαφορική εξίσωση για το $N(t)$ με λύση την:

$$N(t) = N(0)e^{-\lambda t} \tag{8}$$

Συνήθως αντί της σταθεράς διάσπασης λ χρησιμοποιείται ο χρόνος ημιζωής $T_{1/2}$ ως η φυσική παράμετρος που χαρακτηρίζει την διαδικασία διάσπασης. Ορίζεται σαν το χρόνο υποδιπλασιασμού ενός αρχικού πληθυσμού πυρήνων:

$$T_{1/2} = \frac{\ln 2}{\lambda} \tag{9}$$

Είναι φανερό ότι η μελέτη των διακυμάνσεων του ρυθμού διάσπασης ραδιενεργών πυρήνων μπορεί να υλοποιηθεί με δύο τρόπους:

1. Είτε κρατώντας το διάστημα Δt σταθερό να προσδιορίσει κανείς τις διακυμάνσεις του αριθμού διασπάσεων σε αυτό το διάστημα.
2. Είτε βρίσκοντας τις διακυμάνσεις των χρονικών διαστημάτων στα οποία υλοποιείται προκαθορισμένος αριθμός διασπάσεων.

Στην παρούσα άσκηση θα επιλεγεί ο πρώτος τρόπος δηλ. θα μετρηθεί ο αριθμός διασπάσεων σε προκαθορισμένο σταθερό διάστημα Δt.

5 Ιστογράμματα και κατανομές

Τα ιστογράμματα αποτελούν έναν συνηθισμένο τρόπο πρόχειρης εκτίμησης της υποκείμενης κατανομής που ακολουθεί μια τυχαία μεταβλητή. Το ιστόγραμμα που κατασκευάζεται με τα δεδομένα από επαναλαμβανόμενες μετρήσεις μιας τυχαίας μεταβλητής (π.χ. της ορμής ενός σωματιδίου σε μια διάσπαση) απεικονίζει την μορφή της συνάρτησης πυκνότητας πιθανότητας που διέπει την υπό μελέτη τυχαία μεταβλητή. Στην περίπτωση που η μετρούμενη ποσότητα είναι μια διακριτή τυχαία μεταβλητή (όπως σε αυτήν την άσκηση το k), τότε το ιστόγραμμα απεικονίζει την μορφή της συνάρτησης πιθανότητας μάζας της τυχαίας μεταβλητής.

Ο αριθμός των γεγονότων που απαρτίζει έναν ιστό (bin) του ιστογράμματος, ονομάζεται και (απόλυτη) συχνότητα ή και απλώς συχνότητα από πολλούς συγγραφείς, ισούται με την επιφάνεια του ορθογωνίου που καλύπτει ο ιστός. Ο οριζόντιος άξονας του ιστογράμματος είναι διατιμημένος σε ιστούς και έχει απαραιτήτως τις φυσικές μονάδες της υπό μελέτης μεταβλητής (π.χ. ενέργεια, ορμή, απαριθμήσεις ανιχνευτή Geiger-Müller). Το πλάτος του ιστού j είναι το $w_j = x_{\mathrm{j}}^{\max} - x_{\mathrm{j}}^{\min}$. Αν σε κάποιο γεγονός, η μεταβλητή x μετρηθεί να είναι εντός του διαστήματος $[x_{\mathrm{j}}^{\min}, x_{\mathrm{j}}^{\max})$, τότε λέμε ότι το γεγονός αυτό συμπεριλαμβάνεται στον πληθυσμό των γεγονότων που το απαρτίζουν τον ιστό j. Ο κατακόρυφος άξονας, έχει μονάδες πυκνότητας και αναπαριστά τον αριθμό γεγονότων που περιέχονται σε έναν ιστό, ώστε πολλαπλασιαζόμενο με το πλάτος του ιστού w_j να μας δίνει πίσω τον αριθμό γεγονότων που περιέχει ο ιστός (εμβαδόν του ορθογωνίου). Η τεχνική λεπτομέρεια αυτή είναι λιγότερο σημαντική όταν όλοι οι ιστοί έχουν το ίδιο πλάτος $w_j = w_0$. (Για συντομία πολλοί συγγραφείς παραλείπουν την σήμανση που άφορα το πλάτος του ιστού στον τίτλο του κατακόρυφου άξονα ή ακόμη και ολόκληρο τον τίτλο του κατακόρυφου άξονα όταν έχουμε να κάνουμε με ιστογράμματα ομοιόμορφης διαμέρισης.)

Βάσει των ανωτέρω, ο συνολικός αριθμός γεγονότων που συμπεριλήφθηκαν σε ένα ιστόγραμμα ισούται εξ ορισμού με το εμβαδόν του.

Εν γένει, για μια τυχαία μεταβλητή x (διακριτή ή μη), η διαμέριση του οριζόντιου άξονα επιλέγεται έτσι ώστε ο αριθμός των γεγονότων που περιέχονται στους ιστούς ενδιαφέροντος να είναι μεγάλος $n_j > 20$, εφόσον αυτό είναι εφικτό ($N \gg 20$). Η πιθανότητα p_j ένα γεγονός να συνεισφέρει στον ιστό j, ακολουθεί διωνυμική κατανομή καθώς το x είτε θα μετρηθεί εντός του $[x_j^{\min}, x_j^{\max})$ είτε όχι και συνεπώς η διασπορά του n_j περιμένουμε να είναι

$$\sigma_{n_j}^2 = N p_j (1 - p_j). \tag{10}$$

Για παράδειγμα, εάν η τυχαία μεταβλητή x είναι συνεχής και κατανέμεται βάσει της συνάρτησης πυκνότητας πιθανότητας $p(x)$, τότε περιμένουμε

$$n_j = N \int_{x_j^{min}}^{x_j^{max}} p(x)dx = Np_j. \tag{11}$$

Όταν το πλήθος των μετρήσεων είναι μεγάλο $N \gg 1$ και η διαμέριση των ιστών 'λεπτή' ώστε να έχουμε $p_j \ll 1$ η διωνυμική κατανομή προσεγγίζεται καλά από την Poisson και μπορούμε να πούμε ότι η αβεβαιότητα στο πλήθος των γεγονότων που συνεισφέρουν σε έναν ιστό είναι

$$\delta n_j = \sigma_{n_j} \approx \sqrt{Np_j} = \sqrt{n_j}. \tag{12}$$

Στην περίπτωση που η υπό μελέτη μεταβλητή είναι συνεχής μπορούμε πάντα να επιλέγουμε κατάλληλη διμέριση έτσι ώστε $p_j \ll 1$, $n_j > 20$ και συνεπώς να ισχύει ότι $\delta n_j = \sqrt{n_j}$. (Αυτό είναι λιγότερο προφανές όταν έχουμε να κάνουμε με διακριτές τυχαίες μεταβλητές για τις οποίες δεν είμαστε ελεύθεροι να επιλέξουμε οσοδήποτε μικρή διαμέριση επιθυμούμε, όπως για παράδειγμα το αποτέλεσμα της ρίψης ενός ζαριού.)

Η Εξ. 12 μας δίνει ένα απλό κανόνα για τον υπολογισμό της μπάρας σφάλματος (error bar) που βλέπουμε στους ιστούς. Αυτό μας δίνει μια γρήγορη οπτική αίσθηση της αβεβαιότητας που έχουμε στις μετρήσεις. Η αξιοπιστία όμως αυτής, είναι περιορισμένη όταν δεν ικανοποιούνται οι προαναφερθείσες συνθήκες, $N \gg 1$ και $p_j \ll 1$. Σε κάθε περίπτωση, η επεξεργασία των μετρήσεων προς εξαγωγή συμπερασμάτων πρέπει να γίνει προσεκτικά ανεξαρτήτως της μπάρας σφάλματος που χρησιμοποιήθηκε κατά την γραφική αναπαράσταση των δεδομένων σε κάποιο ιστόγραμμα.

Δεν υπάρχει σωστό ή λάθος στην επιλογή της διαμέρισης (εύρος ιστού) ενός ιστογράμματος. Η χρήση πολύ 'λεπτής' διαμέρισης επιφέρει απώλεια στατιστικής ακρίβειας στην εκτίμηση της τιμής στον κατακόρυφο άξονα που αντιστοιχεί σε έναν ιστό. Αντιθέτως, η χρήση πολύ 'χονδρής' διαμέρισης επιφέρει απώλεια στην γνώση που αποκτάμε για το σχήμα της υποκείμενης κατανομής που Θέλουμε να μελετήσουμε. Για συνεχείς τυχαίες μεταβλητές, το ένα όριο είναι να έχουμε ένα πολύ μεγάλο πλήθος 'λεπτών' ιστών με το πολύ ένα γεγονός σε κάθε ιστό, το άλλο όριο είναι να έχουμε ένα και μοναδικό ιστό που συγκεντρώνει όλα τα γεγονότα (και δεν μας δίνει καμία πληροφορία για το σχήμα της υποκείμενης κατανομής). Η βέλτιστη επιλογή διαμέρισης μπορεί να εξαρτάται από το είδος του πειράματος που κάνουμε καθώς και τη διακριτική ικανότητα $\delta x/x$ που έχουμε στην μέτρηση του x.

Στην δική μας περίπτωση, η μεταβλητή ενδιαφέροντος k είναι διακριτή (και αδιάστατη) και το πιο απλό είναι να χρησιμοποιήσουμε ομοιόμορφη διαμέριση του οριζοντίου άξονα σε ιστούς μοναδιαίου μήκους ($w_j = 1$). Το σύνολο των γεγονότων λήψης ικανοποιεί την σχέση

$$N = \sum_{k=0}^{k_{\max}} n_k, \tag{13}$$

όπου το k απαριθμεί τους ιστούς και ταυτοχρόνως συμβολίζει και τον αριθμό διασπάσεων που κατέγραψε ο ανιχνευτής.

6 Θεωρητικά μοντέλα και έλεγχος υποθέσεων

Έστω ότι για N γεγονότα, περιμένουμε θεωρητικώς f_k γεγονότα με k διασπάσεις $/\Delta t$. Πως μπορούμε να αποφανθούμε ποιο από τα διαφορετικά θεωρητικά μοντέλα που έχουμε στην διάθεση ως εναλλακτικές υποθέσεις, περιγράφει καλύτερα τις μετρήσεις που πήραμε στο εργαστήριο;

Στην άσκηση αυτή θα χρησιμοποιήσουμε το κριτήριο χ^2 για την αποδοχή ή απόρριψη κάποιου θεωρητικού μοντέλου. Η μεταβλητή χ^2 ορίζεται ως

$$\chi^2 = \sum_k \left(\frac{n_k - f_k}{\sigma_k} \right)^2 = \sum_k \left(\frac{\text{πείραμα - θεωρία}}{\text{αβεβαιότητα}} \right)^2 \tag{14}$$

και λαμβάνει μια τιμή (για κάθε ανεξάρτητο πείραμα) συνοψίζοντας σε έναν και μόνο αριθμό πόσο 'κοντά' είναι τα πειραματικά δεδομένα σε σχέση με την θεωρητική πρόβλεψη. Πολύ μεγάλο χ^2 σημαίνει κακή συμφωνία θεωρίας–πειράματος, ή υποεκτίμηση του σ_k. Πολύ μικρό χ^2 είναι επίσης απίθανο (και εξίσου προβληματικό) υποδεικνύοντας ότι μάλλον έχουμε υπερεκτιμήσει το σ_k. Σε κάθε περίπτωση, το κριτήριο χ^2 (προ-)υποθέτει ότι η τυχαία μεταβλητή

$$z = \frac{n_k - f_k}{\sigma_k} \tag{15}$$

ακολουθεί την μοναδιαία κανονική κατανομή $z \sim N(\mu = 0, \sigma^2 = 1)$, δηλαδή έχει συνάρτηση πυκνότητας πιθανότητας την

$$p_{\text{normal}}(z) = \frac{1}{\sqrt{2\pi}} \exp^{-z^2/2} . \tag{16}$$

Υπό τον όρο ότι το παραπάνω είναι αληθές, η τυχαία μεταβλητή χ^2 ακολουθεί την κατανομή

$$p_{\chi 2}(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2} \tag{17}$$

με ν βαθμούς ελευθερίας. Οι βαθμοί ελευθερίας είναι ίσοι με το πλήθος των όρων που συνεισφέρουν στο άθροισμα της Εξ. 14 μείον τον αριθμό των παραμέτρων του θεωρητικού μοντέλου που προσδιορίστηκαν κάνοντας χρήση των πειραματικών δεδομένων. Είναι πολύ σύνηθες οι άγνωστες παράμετροι ($\vec{a}$) του θεωρητικού μοντέλου $f(x; \vec{a})$ να προσδιορίζονται από τα δεδομένα, βρίσκοντας την τιμή αυτών για την οποία το μετρούμενο χ^2 ελαχιστοποιείται.

Ένα λεπτό σημείο που προσφέρει συχνή σύγχυση είναι το ποιος είναι ο κατάλληλος ορισμός της αβεβαιότητας σ_k που υπεισέρχεται στον παρανομαστή που βλέπουμε στην Εξ. 14. Στην περίπτωση που έχουμε να κάνουμε με ιστογράμματα, με τον κατακόρυφο άξονα να αναπαριστά την συχνότητα παρατηρήσεων, θέτοντας $\sigma_k = \sqrt{f_k}$ είναι ασφαλές τουλάχιστον για όσους ιστούς έχουν $f_k > 20$. Όταν όμως η μεταβλητή χ^2 χρησιμοποιείται στα πλαίσια της προσαρμογής κάποιου θεωρητικού μοντέλου και τον προσδιορισμό άγνωστων παραμέτρων, η ελαχιστοποίηση της

$$\chi^2 = \sum_k \left(\frac{n_k - f_k}{\sqrt{f_k}} \right)^2 , \tag{18}$$

μπορεί να είναι πολύ δύσκολη καθώς έχουμε τις άγνωστες παραμέτρους του μοντέλου τόσο στον αριθμητή όσο και στον παρανομαστή. Συχνά, απλοποιούμε την Εξ. 18 υποθέτοντας ότι $\sqrt{f_k} \approx \sqrt{n_k}$ και θεωρώντας ότι

$$\chi^2 = \sum_k \left(\frac{n_k - f_k}{\sqrt{f_k}} \right)^2 \approx \sum_k \left(\frac{n_k - f_k}{\sqrt{n_k}} \right)^2 . \tag{19}$$

Με την ανωτέρω απλοποίηση οι υπό προσδιορισμό παράμετροι του μοντέλου δεν αλλάζουν σημαντικά, αν πράγματι ισχύει ότι $\sqrt{f_k} \approx \sqrt{n_k}$, ενώ παράλληλα γίνεται σημαντικά ευκολότερη η εύρεση του ελάχιστου χ^2. Στην παρούσα άσκηση όμως, αυτό δε θα μας απασχολήσει. Δεν θα προσπαθήσουμε να ελαχιστοποιήσουμε την μεταβλητή χ^2. Αντί αυτού, θα χρησιμοποιήσουμε την αριθμητική τιμή της χ^2 για να ελέγξουμε αν το θεωρητικό μας μοντέλο (που δεν έχει ελεύθερες παραμέτρους προς προσδιορισμό) είναι συμβατό με τα πειραματικά δεδομένα που λάβαμε στο εργαστήριο. Για τις ανάγκες λοιπόν της δικής μας άσκησης, δεν υπάρχει κανένας λόγος να μην χρησιμοποιήσουμε τη Εξ. 18 και να προτιμήσουμε την Εξ. 19.

Τα θεωρητικά μοντέλα που θα μελετήσουμε είναι τα:

$$f_k^{\text{poisson}} = N p_{\text{poisson}}(k;\mu) = N \frac{\mu^k}{k!} e^{-\mu} \tag{20}$$

$$f_k^{normal} = N p_{\text{normal}}(k;\mu,\sigma) = N \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(k-\mu)^2}{2\sigma^2}} . \tag{21}$$

Το πρώτο μοντέλο (Εξ. 20) προβλέπει ότι το k ακολουθεί κατανομή Poisson που έχει μόνο μια ελεύθερη παράμετρο προς προσδιορισμό, την μέση τιμή μ. Το δεύτερο μοντέλο (Εξ. 21) προβλέπει ότι το k ακολουθεί την κατανομή Gauss που έχει δύο ελεύθερες παραμέτρους, την μέση τιμή μ και την διασπορά σ. Προκειμένου να ορίσουμε τα δυο θεωρητικά μοντέλα, πρέπει να επιλέξουμε τις τιμές των ελεύθερων παραμέτρων που έχουν, στην προκειμένη του μ για το μοντέλο poisson και των μ και σ για το μοντέλο normal. Και στα δύο μοντέλα υπεισέρχεται μια πρόσθετη παράμετρος N που αφορά την κανονικοποιησή τους και καθορίζεται από τον αριθμό μετρήσεων που επιλέξαμε να κάνουμε.

Η βέλτιστη τιμή για το μ δίνεται από τον γνωστό μας αριθμητικό μέσο όρο

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} k_i = \frac{1}{N} \sum_{\mathrm{k=0}}^{k_{\max}} k \cdot n_k, \tag{22}$$

με το i στο πρώτο άθροισμα να αριθμεί τα N γεγονότα που λάβαμε στο εργαστήριο και το k να αριθμεί τους ιστούς του ιστογράμματος. Για την απόδειξη του παραπάνω ισχυρισμού, φτιάξτε την συνάρτηση πιθανοφάνειας

$$L(\mu) = \prod_i^N \frac{\mu^{k_i}}{k_i!} e^{-\mu} \tag{23}$$

και βρείτε την τιμή του μ που μεγιστοποιεί τον λογάριθμο της $logL(\mu)$ και συνεπώς μεγιστοποιεί ταυτοχρόνως και την ίδια την $L(\mu)$, υπολογίζοντας το σημείο μηδενισμού $\hat{\mu}$ της παραγώγου

$$\frac{d(logL(\mu))}{d\mu}|_{\mu=\hat{\mu}} = 0. \tag{24}$$

k	$p_{\text{poisson}}(k;\mu=20)$	$p_{\text{normal}}(k;\mu=20,\sigma=\sqrt{20})$
15	0.052	0.048
16	0.065	0.060
17	0.076	0.071
18	0.084	0.081
19	0.089	0.087
20	0.089	0.089
21	0.085	0.087
22	0.077	0.081
23	0.067	0.071
24	0.056	0.060
25	0.045	0.048

Πίνακας 3: Σύγκριση της *συνάρτησης πιθανότητας μάζας* Poisson με την κανονική *συνάρτηση πυκνότητας πιθανότητας* για διάφορες τιμές του k με $\mu = 20$ και $\sigma = \sqrt{20}$.

Η επιλογή του $\mu = \hat{\mu}$ βάσει της Εξ. 22 ορίζει πλήρως το μοντέλο poisson όπως αυτό διατυπώνεται στην Εξ.20. Ταυτοχρόνως, μικραίνει κατά μια μονάδα τον αριθμό των βαθμών ελευθερίας που θα έχουμε στην κατασκευή του χ^2_{Poisson}, αφού χρησιμοποιήσαμε ήδη μια φορά τα δεδομένα για να προσδιορίσουμε το $\hat{\mu}$. Επεκτείνοντας την χρήση του ίδιου μ και για την παράμετρο που εμφανίζεται στο μοντέλο normal της Εξ. 21 και υπό το πρίσμα του ότι η κατανομή Poisson για μεγάλο $\mu > 20$ προσεγγίζεται πολύ καλά από μια κανονική κατανομή που έχει την ίδια μέση τιμή και $\sigma^2 = \mu$, φαίνεται λογική[2] η επιλογή της τυπικής απόκλισης να είναι $\sigma = \sqrt{\hat{\mu}}$, στο μοντέλο normal της Εξ. 21. Παρόλα αυτά, υπεύθυνο του να κρίνει του λόγου το αληθές είναι το πείραμα και αυτό είναι το πνεύμα γραφής της παρούσας άσκησης.

Πρέπει να επισημανθεί ότι η Εξ. 20 κάνει την υπόθεση ότι η διακριτή τυχαία μεταβλητή k κατανέμεται βάσει της Poisson *συνάρτησης πιθανότητας μάζας* (Poisson PMF) ενώ σε αντιδιαστολή, η Εξ. 21 υποθέτει την κανονική *συνάρτησης πυκνότητας πιθανότητας* (Normal PDF), στην 'συνεχή θεώρηση' της τυχαίας μεταβλητής k. Προφανώς σε ένα πείραμα καταμέτρησης του αριθμού διασπάσεων k που καταγράφει ο ανιχνευτής ανά μονάδα χρόνου, μόνο διακριτές τιμές της μεταβλητής k έχουν νόημα. Η αξία όμως της Εξ.21 γίνεται ορατή, όταν $k \gg 20$, καθώς ο υπολογισμός του f_k^{poisson} γίνεται εξαιρετικά δύσκολος λόγω της εμφάνισης $k!$ στον παρανομαστή της Εξ. 20. Η p_{poisson} της Εξ. 20 προσεγγίζεται ικανοποιητικά για μεγάλα μ από την p_{normal} της Εξ. 21 θέτοντας την παράμετρο $\sigma = \sqrt{\mu}$. Ενδεικτικά δίνεται ο πίνακας 3.

Με τον υπολογισμό του χ^2/ν συναρτήσει των πειραματικών δεδομένων, μπορούμε να ποσοτικοποιήσουμε την ασυμφωνία μεταξύ θεωρίας και πειράματος. Η σημαντικότητα της ασυμφωνίας ορίζεται ως η πιθανότητα που αναμένουμε θεωρητικά να μετρήσουμε μια τιμή για το χ^2, ίδια ή και μεγαλύτερη από αυτή που μετρήσαμε συναρτήσει των πειραματικών

[2]Εναλλακτική υπόθεση για το μοντέλο της κανονικής κατανομής θα ήταν να χρησιμοποιούσαμε ως σ την τετραγωνική ρίζα της διασποράς του k στο δείγμα δεδομένων που πήραμε.

δεδομένων. Η πιθανότητα αυτή είναι

$$p = \int_{\chi^2}^{\infty} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} t^{\nu/2-1} e^{-t/2} dt = 1 - \int_{0}^{\chi 2} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} t^{\nu/2-1} e^{-t/2} dt. \qquad (25)$$

Πρέπει να επισημανθεί ότι από την στιγμή που το μοντέλο που θέλουμε να ελέγξουμε έχει πλήρως προσδιοριστεί και δεν έχει ελεύθερες παραμέτρους, η χ^2 είναι μια τυχαία μεταβλητή που εξαρτάται εξολοκλήρου από τα πειραματικά δεδομένα που συλλέξαμε $\chi^2(k_1, k_2, ...k_N)$ και δεν έχει άγνωστες θεωρητικές παραμέτρους προς προσδιορισμό. Είναι δηλαδή ένας καθαρός αριθμός, που συνοψίζει το πόσο συμβατά είναι τα πειραματικά δεδομένα με την θεωρία που υποθέτουμε στο μοντέλο.

7 Περαιτέρω εργασίες

- Δείξτε ότι η Εξ. 22 αποτελεί λύση της Εξ. 24 στην περίπτωση του μοντέλου poisson.
- Θεωρήστε ως ένα τρίτο εναλλακτικό μοντέλο κανονικής κατανομής

$$k \sim N(\mu = \hat{\mu}, \sigma = \hat{\sigma}) \tag{26}$$

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} k_i \tag{27}$$

$$\hat{\sigma} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(k_i - \hat{\mu})^2}. \tag{28}$$

Πόσοι είναι οι βαθμοί ελευθερίας σε αυτήν την περίπτωση; Συμπληρώστε μια επιπρόσθετη στήλη για το μοντέλο αυτό στον πίνακα των αποτελεσμάτων σας (ακολουθώντας το υπόδειγμα του πίνακα 2).

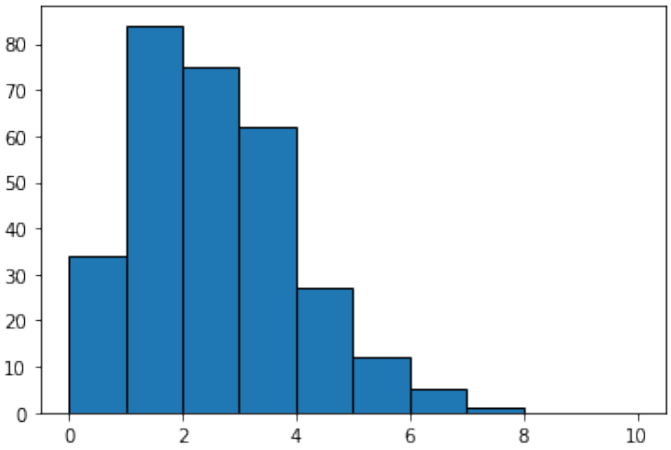
8 Ιστογράμματα στην Python

Το παρακάτω δίνεται ως υπόδειγμα, μπορείτε να χρησιμοποιήσετε **όποιο υπολογιστικό εργαλείο ή οποιαδήποτε άλλη γλώσσα προγραμματισμού** *προτιμάτε.* Το παρακάτω πρόγραμμα είναι επίσης διαθέσιμο διαδραστικά στο νέφος.

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

data = np.loadtxt('data.txt')
for i in range(5):
    print(data[i])

hist, bins = np.histogram(data, bins = range(0,11))

print(hist)
print(bins)

plt.hist(data, bins, ec='black')
```

Ο παραπάνω κώδικας θα φτιάξει το ιστόγραμμα των δεδομένων που βρίσκονται στο αρχείο data.txt και θα εκτυπώσει τα παρακάτω:

```
3.0
4.0
2.0
0.0
2.0
[34 84 75 62 27 12  5  1  0  0]
[ 0  1  2  3  4  5  6  7  8  9 10]
```


80
70
60
50
40
30
20
10
0
0
2
4
6
8
10