Η Κατανομή Poisson

Ας δούμε ένα πρόβλημα: *Σε μια κτηνοτροφική περιοχή υπάρχουν 300000 αιγοπρόβατα. Κάθε χρόνο όλα τα αιγοπρόβατα εμβολιάζονται για προστασία από κάποια ασθένεια. Σύμφωνα με την άδεια χρήσης του εμβολίου, υπάρχει πολύ μικρή πιθανότητα, ίση με* $2\cdot 10^{-5}$*, το εμβόλιο να προκαλέσει μια πολύ σοβαρή παρενέργεια που οδηγεί στο θάνατο του ζώου που εμβολιάζεται. Μας ενδιαφέρει να βρούμε την πιθανότητα, στη συγκεκριμένη περιοχή, να πεθάνουν σε ένα έτος x ζώα από την παρενέργεια που προκαλεί ο εμβολιασμός.*

Είναι προφανές[1] ότι ο αριθμός X των ζώων που πεθαίνουν από τον εμβολιασμό στη συγκεκριμένη περιοχή σε ένα έτος είναι *διωνυμική* τυχαία μεταβλητή με $X \sim B(300000,\, 2\cdot 10^{-5})$. Ενδιαφερόμαστε για τις πιθανότητες $P(X=x)$, όπου $x = 0,1,...,300000$. Έτσι, για την πιθανότητα *σε ένα έτος να μην υπάρξουν θάνατοι ζώων από τον εμβολιασμό* έχουμε,

$$P(X=0)=\binom{300000}{0}\cdot(2\cdot 10^{-5})^0(1-2\cdot 10^{-5})^{300000}=1\cdot 1\cdot(0.99998)^{300000}=0.0024786$$

και για την πιθανότητα *σε ένα έτος να υπάρξουν ακριβώς 2 θάνατοι από τον εμβολιασμό* έχουμε,

$$P(X=2)=\binom{300000}{2}\cdot(2\cdot 10^{-5})^2(1-2\cdot 10^{-5})^{299998}=\frac{300000!}{2!\cdot 299998!}\cdot(0.00002)^2\cdot(0.99998)^{299998}=$$

$$=\frac{299999\cdot 300000}{2}\cdot(0.00002)^2\cdot(0.99998)^{299998}=0.0446165.$$

Ήδη από αυτά τα δύο αριθμητικά παραδείγματα φαίνεται ότι για τιμές των παραμέτρων της *διωνυμικής* κατανομής όπως αυτές του προβλήματός μας (κυρίως για πολύ μεγάλο *n)*, ο τύπος της συνάρτησης πιθανότητάς της δεν είναι πολύ πρακτικός για τον υπολογισμό πιθανοτήτων (ιδιαίτερα χωρίς υπολογιστική μηχανή). Μάλιστα, για τιμές x που δεν είναι κοντά στο 0 ή το n το πρόβλημα γίνεται μεγάλο. Η πιθανότητα, για παράδειγμα, *σε ένα έτος να υπάρξουν ακριβώς 15 θάνατοι από τον εμβολιασμό* είναι…

$$P(X=15)=\binom{300000}{15}\cdot(2\cdot 10^{-5})^{15}(1-2\cdot 10^{-5})^{299985}=\frac{300000!}{15!\cdot 299985!}\cdot(0.00002)^{15}\cdot(0.99998)^{299985}=$$

$$=\frac{299986\cdot 299987\cdot\ldots\cdot 300000}{1\cdot 2\cdot 3\cdot\ldots\cdot 15}\cdot(0.00002)^{15}\cdot(0.99998)^{299985}=\ldots.=0.0008912$$

Οι δυσκολίες αυτές, αναγνωρίσθηκαν από πολύ νωρίς. Ο Γάλλος μαθηματικός *Simeon Denis Poisson (1781-1840)* αναζήτησε τρόπο αντιμετώπισης του προβλήματος και το 1837, σε βιβλίο του με εφαρμογές της θεωρίας πιθανοτήτων σε θέματα δικαστικών υποθέσεων, δημοσίευσε το παρακάτω εντυπωσιακό για την εποχή του οριακό θεώρημα.

Έστω ότι η τυχαία μεταβλητή X ακολουθεί τη διωνυμική κατανομή $B(n,\, p)$ με συνάρτηση πιθανότητας

$$P(X=x)=\binom{n}{x}\,p^x(1-p)^{n-x},\quad x=0,1,2,...,n.$$

Αν, για $n\to\infty$, το $p\to 0$ έτσι ώστε η μέση τιμή της X να συγκλίνει προς μια θετική σταθερά λ, δηλαδή, να ισχύει $np\to\lambda$, τότε

$$\lim_{n\to\infty}\binom{n}{x}\,p^x(1-p)^{n-x}=e^{-\lambda}\,\frac{\lambda^x}{x!},\quad x=0,1,2,...$$

[1] *Με την υπόθεση ότι κάθε ζώο έχει την ίδια πιθανότητα να πεθάνει από τον εμβολιασμό σε ένα έτος.*

Πολύ αργότερα[2] διαπιστώθηκε ότι η συνάρτηση $e^{-\lambda}\frac{\lambda^x}{x!}$, $x = 0,1,2,...$, που εισάγεται με το *οριακό θεώρημα του Poisson,* έχει όλες τις ιδιότητες μιας συνάρτησης πιθανότητας. Έτσι ορίσθηκε η ***κατανομή Poisson***.

<u>Ορισμός</u>

Έστω X μια διακριτή τυχαία μεταβλητή με συνάρτηση πιθανότητας

$$f(x) = P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0,1,2,...$$

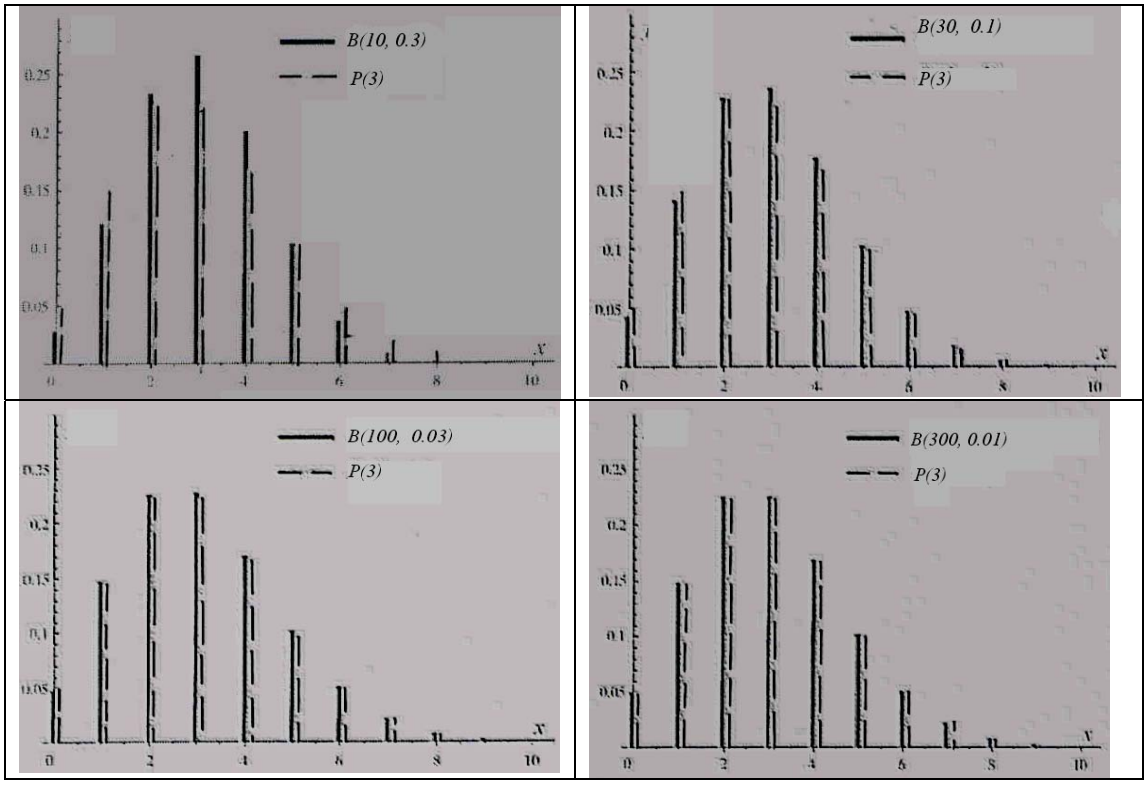
όπου $\lambda > 0$. *Η κατανομή της τυχαίας μεταβλητής X ονομάζεται κατανομή Poisson με παράμετρο* λ *και συμβολίζεται με* $P(\lambda)$.

Δηλαδή, η *κατανομή Poisson* ορίσθηκε ως οριακή κατανομή της *διωνυμικής, έτσι:*

Η ***<u>Διωνυμική κατανομή προσεγγίζεται από την κατανομή Poisson</u>*** *αν για n μεγάλο (θεωρητικά* $n \to \infty$*), η πιθανότητα επιτυχίας p συγκλίνει στο 0 (*$p \to 0$*) έτσι ώστε η μέση τιμή της κατανομής να συγκλίνει σε μια θετική σταθερά* λ *(*$np \to \lambda$*).*

Πρακτικά, όμως, ***πόσο μεγάλο πρέπει να είναι το n*** *και* ***πόσο μικρό το p*** για να είναι ικανοποιητική η προσέγγιση; Έχει παρατηρηθεί ότι *αν* $n \geq 20$ *και* $p \leq \frac{10}{n}$ *ώστε η μέση τιμή* np *να παίρνει μέτριες τιμές (στην πράξη, μικρότερες του 10), η ακρίβεια της προσέγγισης είναι ικανοποιητική.*

Στα σχήματα που ακολουθούν φαίνεται γραφικά η σύγκλιση της *διωνυμικής* κατανομής $B(n, p)$ στην κατανομή *Poisson* με $\lambda = np = 3$ όταν $n = 10,30,100,300$ και $p = 0.3,\ 0.1,\ 0.03,\ 0.01$ αντίστοιχα. Παρατηρείστε ότι για $n = 300$ και $p = 0.01$ η προσέγγιση είναι τέλεια.



[2] *Το 1889, από τον Ρωσο-Γερμανό μαθηματικό L.V. Bortriewicz.*

Ως παράδειγμα εφαρμογής των παραπάνω ας δούμε πάλι το εισαγωγικό πρόβλημα. Επειδή $n = 300000 \geq 20$ και $p = 0.00002 \leq \frac{10}{300000} = 0.00003$ μπορούμε να υπολογίσουμε πολύ πιο εύκολα τις ζητούμενες πιθανότητες αν χρησιμοποιήσουμε την προσέγγιση της διωνυμικής κατανομής $B(300000, 0.00002)$ από την κατανομή *Poisson* με παράμετρο $\lambda = np = 300000 \cdot 0.00002 = 6$. Έτσι έχουμε:

$$P(X = x) \cong e^{-6} \frac{6^x}{x!}, \quad x = 0,1,2,...$$

Άρα:

$$P(X = 0) \cong e^{-6} \frac{6^0}{0!} = 0.0024788$$

$$P(X = 2) \cong e^{-6} \frac{6^2}{2!} = 0.0446176$$

$$P(X = 15) \cong e^{-6} \frac{6^{15}}{15!} = 0.0008913$$

Αποδεικνύεται ότι, αν $X \sim P(\lambda)$ τότε,

$$\mu = E(X) = \lambda \text{ και } \sigma^2 = V(X) = \lambda .$$

Στην πράξη, η παράμετρος λ, συνήθως δεν υπολογίζεται από τα n και p αλλά εκτιμάται εμπειρικά (από στατιστικά στοιχεία).

<u>*Ερώτηση:*</u> Είναι άραγε αναμενόμενο (λογικό) η μέση τιμή της κατανομής *Poisson* να είναι ίση με τη διασπορά της[3];

Η κατανομή *Poisson* ως οριακή κατανομή της *διωνυμικής* κατανομής έχει, όπως και η διωνυμική, μεγάλο εύρος εφαρμογών σε διάφορες επιστημονικές περιοχές. Πιο συγκεκριμένα, χρησιμοποιείται για τη μοντελοποίηση *«διωνυμικών καταστάσεων»* όπου ενδιαφέρει **<u>ο αριθμός εμφανίσεων σπάνιων ενδεχομένων σε μεγάλους πληθυσμούς</u>** (δηλ. όταν σε κάθε επανάληψη, η πιθανότητα επιτυχίας p είναι πολύ μικρή και ο αριθμός επαναλήψεων n πολύ μεγάλος). Γι' αυτό το λόγο, στη βιβλιογραφία αναφέρεται και ως ***κατανομή των σπάνιων ενδεχομένων (distribution of rare events[4])***.

Ας δούμε μερικά χαρακτηριστικά παραδείγματα τυχαίων μεταβλητών που ακολουθούν την κατανομή *Poisson*.

1. Ο αριθμός X των τροχαίων ατυχημάτων σε ένα τμήμα (με μεγάλη κυκλοφορία) του οδικού δικτύου μιας χώρας στη διάρκεια ενός Σαββατοκύριακου (ή μιας ημέρας, ή μιας εβδομάδας, ή ενός μήνα κτλ.). Με την υπόθεση ότι κάθε αυτοκίνητο που περνάει από το συγκεκριμένο σημείο έχει την ίδια πιθανότητα p να εμπλακεί σε τροχαίο ατύχημα, η X ακολουθεί την $B(n, p)$. Επειδή ο αριθμός των αυτοκινήτων n που διέρχονται από το συγκεκριμένο σημείο τη συγκεκριμένη χρονική περίοδο είναι μεγάλος και η πιθανότητα ατυχήματος (επιτυχίας!) p είναι πολύ μικρή[5], η κατανομή της X προσεγγίζεται ικανοποιητικά από την κατανομή *Poisson* με $\lambda = np$. Η παράμετρος λ εκφράζει τον μέσο αριθμό ατυχημάτων στο

[3] *Είναι. Θυμηθείτε τη μέση τιμή και τη διασπορά της διωνυμικής και παρατηρείστε ότι όταν συγκλίνει στην Poisson το 1-p είναι περίπου 1.*

[4] *Αναφέρεται επίσης ως* ***νόμος των μικρών αριθμών.***

[5] *Φοβάμαι ότι θλιβερή εξαίρεση αποτελεί η Ελλάδα (με τους οδηγούς της και τους δρόμους της!!).*

συγκεκριμένο σημείο τη συγκεκριμένη χρονική περίοδο και όπως αναφέραμε, στην πράξη, συνήθως εκτιμάται εμπειρικά (από στατιστικά στοιχεία).

2. Ο αριθμός X των τυπογραφικών λαθών σε μια δακτυλογραφημένη σελίδα (ή σε ένα σύνολο σελίδων). Όπως και στο προηγούμενο παράδειγμα πρόκειται για *διωνυμική* κατανομή $B(n, p)$. Επειδή το n είναι μεγάλο (πολλά γράμματα στο κείμενο) και η πιθανότητα τυπογραφικού λάθους (επιτυχίας!) p είναι μικρή[6] η κατανομή του αριθμού των τυπογραφικών λαθών/σελίδα προσεγγίζεται ικανοποιητικά από την κατανομή *Poisson* με $\lambda = np$. Η παράμετρος λ εκφράζει τον μέσο αριθμό τυπογραφικών λαθών/σελίδα και όπως αναφέραμε και στο προηγούμενο παράδειγμα, στην πράξη, συνήθως εκτιμάται εμπειρικά (από στατιστικά στοιχεία).
3. Ο αριθμός X των κλήσεων στο help desk ενός μεγάλου Internet provider σε μια ημέρα (ή σε μια ώρα, ή σε μια εβδομάδα κτλ.).
4. Ο αριθμός X των βλαβών μιας μηχανής σε μια ημέρα (ή σε μια εβδομάδα κτλ.).
5. Ο αριθμός X των ατόμων ενός πληθυσμού που ζουν περισσότερα από 100 χρόνια.
6. Ο αριθμός X των παιδιών ενός πληθυσμού που θα γίνουν ψηλότερα από 1.95μέτρα.
7. Ο αριθμός X των βακτηριδίων σε 1cm^2 μιας πλάκας Petri.
8. Ο αριθμός X των πελατών ενός super market σε μια ημέρα, που θα αγοράσουν σοκολατάκια για σκύλους.
9. Ο αριθμός X των ελαττωματικών προϊόντων που παράγονται από μια συγκεκριμένη γραμμή παραγωγής σε ένα ορισμένο χρονικό διάστημα.
10. Ο αριθμός X των λανθασμένων τηλεφωνικών κλήσεων (άλλος αριθμός πληκτρολογείται και άλλος καλείται) σε μια ημέρα. Επίσης, ο αριθμός X των τηλεφωνικών κλήσεων που φθάνουν σε ένα τηλεφωνικό κέντρο σε μια συγκεκριμένη χρονική περίοδο.
11. Ο αριθμός X των φυσαλίδων σε υαλοπίνακα συγκεκριμένης επιφάνειας.
12. Ο αριθμός X των θανάτων σε μια πόλη από μια σπάνια ασθένεια σε ένα μήνα.
13. Ο αριθμός X των πελατών που φθάνουν σε ένα κέντρο εξυπηρέτησης (τράπεζα, ταχυδρομικό γραφείο, κατάστημα κτλ.) σε μια ημέρα (ή σε μια ώρα, ή σε μια εβδομάδα, κτλ.).
14. Ο αριθμός X των επιβατών μιας αεροπορικής πτήσης που ενώ έχουν κάνει κράτηση θέσης δεν εμφανίζονται την ώρα αναχώρησης.
15. Ο αριθμός X των α-σωματίων που εκπέμπονται από ραδιενεργό υλικό σε συγκεκριμένο χρονικό διάστημα.
16. Ο αριθμός X των σεισμών μεγέθους περισσότερο των 5 βαθμών της κλίμακας *Richter* που πλήττουν μια σεισμογόνο περιοχή σε ένα έτος.
17. Ο αριθμός X των ελαττωματικών σημείων που υπάρχουν σε συγκεκριμένο μήκος καλωδίου.
18. Ο αριθμός X των βακτηριδίων σε διάλυμα συγκεκριμένου όγκου.
19. Ο αριθμός X των αστεριών σε μια γαλαξιακή περιοχή συγκεκριμένου όγκου.
20. Ο αριθμός X των προβλημάτων (ρωγμές και λακκούβες) στο οδόστρωμα ενός εθνικού δρόμου ανά Km.
21. Ως τελευταίο παράδειγμα αναφέρουμε εφαρμογές στη χωροδιάταξη (spatial pattern) φυτών, ζώων κτλ που είναι τυχαία διασκορπισμένα σε μια μεγάλη έκταση ώστε κάθε δειγματοληπτική μονάδα (τετράγωνο «μικρού» εμβαδού) να έχει πολύ μικρή πιθανότητα να «φιλοξενήσει» ένα φυτό ή ζώο.

Όπως φαίνεται από τα παραπάνω παραδείγματα, η κατανομή *Poisson* βρίσκει εφαρμογή και σε περιπτώσεις όπου σε ένα τυχαίο πείραμα μας ενδιαφέρει **πόσες φορές εμφανίζεται ένα ενδεχόμενο σε *χρονικό διάστημα t* ή σε *μήκος t* ή σε**

[6] *Εκτός αν η δακτυλογράφος έχει πρόβλημα.*

επιφάνεια t **ή σε** ***όγκο t***. Αυτό συμβαίνει διότι, όταν ικανοποιούνται τρεις συνθήκες[7], τότε και οι περιπτώσεις αυτές είναι *«διωνυμικές καταστάσεις»* με *n* πολύ μεγάλο και *p* πολύ μικρό. Οι συνθήκες αυτές για την περίπτωση χρονικού διαστήματος είναι οι εξής[8]:

Σ1. Η πιθανότητα να εμφανισθεί το ενδεχόμενο σε ένα μικρό χρονικό διάστημα μια φορά είναι ανάλογη του μήκους του.
Σ2. Η πιθανότητα να εμφανισθεί το ενδεχόμενο δύο ή περισσότερες φορές σε ένα μικρό χρονικό διάστημα είναι αμελητέα.
Σ3. Οι εμφανίσεις του ενδεχομένου σε δύο ξένα χρονικά διαστήματα είναι ανεξάρτητα ενδεχόμενα.

Υποθέτουμε επίσης ότι οι συνθήκες του πειράματος παραμένουν αμετάβλητες (στο χρόνο, το χώρο, κτλ.). Η εξήγηση, διαισθητικά, γιατί υπό τις συνθήκες Σ1, Σ2 και Σ3, οι περιπτώσεις αυτές είναι «διωνυμικές καταστάσεις» με *n* πολύ μεγάλο και *p* πολύ μικρό είναι σχετικά απλή[9].

Με βάση τα προηγούμενα, είναι φανερό ότι για κάθε $t \geq 0$ έχουμε μια τυχαία μεταβλητή X_t που εκφράζει πόσες φορές εμφανίσθηκε το ενδεχόμενο σε διάστημα *t*. Πρόκειται δηλαδή για μια οικογένεια τυχαίων μεταβλητών $\{X_t,\ t \geq 0\}$ η οποία ονομάζεται ***στοχαστική διαδικασία (ανέλιξη) Poisson***. Για τη συνάρτηση πιθανότητας της X_t, όπως αναφέραμε προηγουμένως, αποδεικνύεται ότι:

Αν ένα ενδεχόμενο εμφανίζεται σε χρονικό διάστημα t (ή σε μήκος t ή σε επιφάνεια t ή σε όγκο t) έτσι ώστε να ικανοποιούνται οι συνθήκες Σ1, Σ2 και Σ3, τότε υπάρχει ένας θετικός αριθμός λ τέτοιος ώστε η κατανομή του αριθμού X_t των εμφανίσεων του ενδεχομένου σε χρονικό διάστημα t (ή σε μήκος t ή σε επιφάνεια t ή σε όγκο t), να δίνεται από τον τύπο

$$P(X_t = x) = e^{-\lambda t}\frac{(\lambda t)^x}{x!}, \quad x = 0,1,2, \ldots$$

Δηλαδή, η X_t ακολουθεί την κατανομή *Poison* με μέση τιμή λt. Το λ εκφράζει τον μέσο αριθμό εμφανίσεων του ενδεχομένου στη μονάδα του χρόνου (ή μήκους ή επιφάνειας ή όγκου) ή αλλιώς τον ρυθμό εμφάνισης του ενδεχομένου. Στην πράξη, το λ εκτιμάται εμπειρικά (από στατιστικά στοιχεία).

Ας δούμε ένα παράδειγμα.
Στο help desk ενός μεγάλου Internet provider φθάνουν αιτήματα πελατών με ρυθμό 3 αιτήματα ανά λεπτό. Ποια είναι η πιθανότητα ***α)*** *σε ένα λεπτό να φθάσουν το πολύ 2 αιτήματα* ***β)*** *σε μισό λεπτό να φθάσουν το πολύ 2 αιτήματα* ***γ)*** *σε 2 λεπτά να φθάσουν το πολύ 4 αιτήματα και* ***δ)*** *σε 3 διαφορετικά χρονικά διαστήματα του ενός λεπτού να βρεθούν τουλάχιστον δύο τέτοια διαστήματα σε καθένα από τα οποία να έχουν φθάσει το πολύ 2 αιτήματα.*
Απάντηση

---

[7] *που είναι αρκετά απλές και φυσιολογικές-λογικές*
[8] *Ανάλογα διατυπώνονται για μήκος, επιφάνεια ή όγκο.*
[9] *Αν χωρίσουμε το διάστημα [0, t) σε n υποδιαστήματα ίδιου πλάτους $\frac{t}{n}$ (όπου n πολύ μεγάλο ώστε $\frac{t}{n} \to 0$), η συνθήκη Σ2 εξασφαλίζει ότι πρόκειται για δοκιμές Bernoulli και οι Σ1, Σ3 ότι είναι ανεξάρτητες με σταθερή πιθανότητα επιτυχίας $\frac{\lambda \cdot t}{n}$, όπου λ ο αναμενόμενος αριθμός εμφανίσεων στη μονάδα χρόνου, χώρου, κτλ.*

Ο αριθμός X_t των αιτημάτων που φθάνουν στο help desk σε διάστημα t λεπτών ακολουθεί κατανομή Poisson με

$$P(X_t = x) = e^{-3t}\frac{(3t)^x}{x!}, \quad x = 0,1,2,\ldots$$

Επομένως έχουμε:

α) $P(X_1 \le 2) = P(X_1 = 0) + P(X_1 = 1) + P(X_1 = 2) = \sum_{x=0}^{2} e^{-3}\frac{3^x}{x!} = 0.4232$

β) $P(X_{1/2} \le 2) = P(X_{1/2} = 0) + P(X_{1/2} = 1) + P(X_{1/2} = 2) = \sum_{x=0}^{2} e^{-1.5}\frac{1.5^x}{x!} = 0.8088$

γ) $P(X_2 \le 4) = \sum_{x=0}^{4} e^{-6}\frac{6^x}{x!} = 0.2851$

δ) *Τα τρία χρονικά διαστήματα του ενός λεπτού μπορούν να θεωρηθούν ως τρεις ανεξάρτητες δοκιμές Bernoulli στις οποίες επιτυχία σημαίνει: σε ένα λεπτό φθάνουν το πολύ δύο αιτήματα. Έτσι αν συμβολίσουμε με Y τον αριθμό των επιτυχιών στις 3 δοκιμές είναι προφανές ότι* $Y \sim B(3,\ 0.4232)$ δηλαδή

$$P(Y = y) = \binom{3}{y}(0.4232)^y(0.5768)^{3-y}, \ y = 0,1,2,3$$

και επομένως η ζητούμενη πιθανότητα είναι:

$$P(Y \ge 2) = \binom{3}{2}(0.4232)^2(0.5768)^1 + \binom{3}{3}(0.4232)^3(0.5768)^0 = 0.3857.$$

Παρατηρήσεις

1. *Όπως και στη διωνυμική κατανομή, εύκολα αποδεικνύεται ο παρακάτω αναδρομικός τύπος υπολογισμού των πιθανοτήτων* $P(X = x),\ x = 1,2,\ldots$ *μιας Poisson τυχαίας μεταβλητής X.*

> *Για τις πιθανότητες* $P(X = x),\ x = 1,2,\ldots,$ *της τυχαίας μεταβλητής* $X \sim P(\lambda)$ *ισχύει,* $P(X = x) = \frac{\lambda}{x}\cdot P(X = x-1)$ *με αρχική συνθήκη,:* $P(X = 0) = e^{-\lambda}$.

2. *Όπως και στη διωνυμική κατανομή, εύκολα αποδεικνύεται ότι η πιο πιθανή τιμή της* $X \sim P(\lambda)$ *είναι η* $x_0 = [\lambda]$ *όταν ο θετικός αριθμός λ <u>δεν είναι</u> ακέραιος, ενώ όταν <u>είναι</u> ακέραιος οι τιμές της X με τη μεγαλύτερη πιθανότητα είναι δύο: η* $x_0 = \lambda$ *και η* $x_0' = \lambda - 1$.