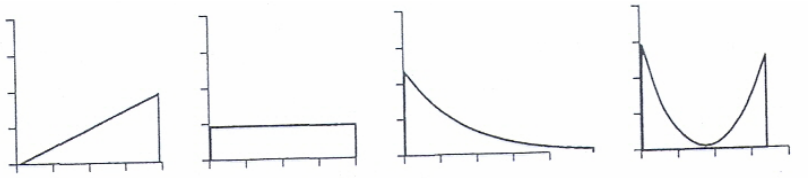
Το Κεντρικό Οριακό Θεώρημα

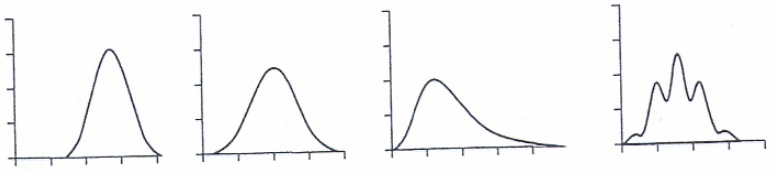
Στα προηγούμενα (σελ. 71), δώσαμε μια πρώτη, γενική, διατύπωση του *Κεντρικού Οριακού Θεωρήματος (Κ.Ο.Θ.)* και τη γενική ιδέα για το πώς το *Κ.Ο.Θ.* εξηγεί το μεγάλο εύρος εφαρμογής της *κανονικής κατανομής* και πώς συνδέει οποιαδήποτε κατανομή με την *κανονική*. Ας δούμε τώρα μια πληρέστερη διατύπωσή του.

Αν από έναν πληθυσμό που ακολουθεί ***οποιαδήποτε κατανομή*** *με μέση τιμή μ και διασπορά σ^2, επιλέξουμε τυχαία δείγματα μεγέθους n και υπολογίσουμε τους μέσους τους, τότε, για μεγάλα n (θεωρητικά $n \to \infty$) η κατανομή αυτών των μέσων (των δειγματικών) είναι κατά προσέγγιση κανονική κατανομή με μέση τιμή επίσης μ και διασπορά σ^2 / n.*

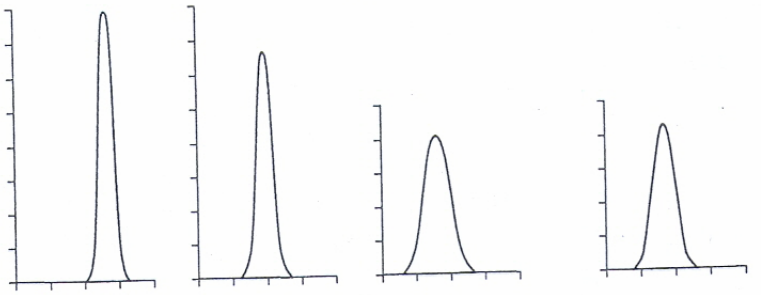
Έτσι, αν για παράδειγμα, θεωρήσουμε τις παρακάτω κατανομές τεσσάρων <u>πληθυσμών</u>



τότε, οι κατανομές των <u>δειγματικών μέσων</u> είναι αντίστοιχα,
για ***n = 4***:



και για ***n = 25***:



Όσο πιο μεγάλο είναι το μέγεθος *n* των δειγμάτων, τόσο καλύτερη (ακριβέστερη) είναι η προσέγγιση της κατανομής των δειγματικών μέσων από την κανονική κατανομή. *(Δες και την Παρατήρηση 2 στη σελίδα 86).*

Ας δούμε όμως, με ένα συγκεκριμένο παράδειγμα, τι σημαίνουν τα παραπάνω στην πράξη.

<u>*Παράδειγμα 1:*</u> *Μας είναι γνωστό, ότι τα μήλα μιας δενδροκαλλιέργειας έχουν μέσο βάρος μ = 50gr με τυπική απόκλιση σ = 10gr. Η παραγωγή της καλλιέργειας συσκευάζεται σε κιβώτια και προωθείται στην αγορά. Σε κάθε κιβώτιο τοποθετούνται 400 μήλα (τυχαία επιλεγμένα). Μπορούμε να υπολογίσουμε ποιο ποσοστό (κατά προσέγγιση) των κιβωτίων περιέχει μήλα με μέσο βάρος: α) μεγαλύτερο των 51.25gr και β) μικρότερο των 49gr;*

<u>*Απάντηση*</u>

Τα μήλα κάθε κιβωτίου είναι ένα τυχαίο δείγμα μεγέθους *n = 400* από τον πληθυσμό των μήλων της παραγωγής. Η μορφή της κατανομής των βαρών των μήλων δεν μας είναι γνωστή. Μπορεί να είναι οποιαδήποτε. Για την άγνωστη αυτή κατανομή γνωρίζουμε μόνο τη *μέση τιμή* της *μ = 50 gr* και την *τυπική απόκλισή της σ = 10 gr.*

Μπορούμε με αυτά τα δεδομένα να απαντήσουμε στα ερωτήματα που θέσαμε; Η απάντηση είναι ναι και ας δούμε πώς. Τα ερωτήματά μας μπορούν να επαναδιατυπωθούν ως εξής:

Ποιο ποσοστό (κατά προσέγγιση) των δειγματικών μέσων α) είναι μεγαλύτερο των 51.25gr και β) είναι μικρότερο των 49gr.

Είναι φανερό ότι για να μπορέσουμε να απαντήσουμε στα ερωτήματα που θέσαμε (και σε άλλα παρόμοια) **πρέπει να γνωρίζουμε την *κατανομή των δειγματικών μέσων***.

Το *Κ.Ο.Θ.* μας βεβαιώνει ότι, παρότι δε γνωρίζουμε την κατανομή των βαρών των μήλων, εντούτοις, γνωρίζουμε την κατανομή των *δειγματικών μέσων* αφού τα δείγματά μας έχουν μέγεθος αρκετά μεγάλο (400). Δηλαδή, αρκεί μόνο, το ότι γνωρίζουμε τη *μέση τιμή* και την *τυπική απόκλιση* της κατανομής των βαρών των μήλων.

Έτσι, αν συμβολίσουμε με X την τυχαία μεταβλητή που εκφράζει το βάρος ενός τυχαία επιλεγμένου μήλου της καλλιέργειας, και με $\overline{X}$ την τυχαία μεταβλητή που εκφράζει τους *δειγματικούς μέσους*, το *Κ.Ο.Θ.* μας βεβαιώνει ότι η $\overline{X}$ ακολουθεί κατά προσέγγιση *κανονική κατανομή* με *μέση τιμή* $\mu_{\overline{X}} = \mu = 50\,gr$ και *διασπορά* $\sigma^2_{\overline{X}} = \frac{\sigma^2}{n} = \frac{10^2}{400} = 0.25\,gr^2$. Αφού, πλέον, γνωρίζουμε ότι $\overline{X} \sim N(50,\, 0.5^2)$, η απάντηση στα ερωτήματά μας είναι πολύ απλή. Ζητάμε τις πιθανότητες: $P(\overline{X} > 51.25)$ και $P(\overline{X} < 49)$. Έτσι έχουμε:

α) $P(\overline{X} > 51.25) = 1 - P(\overline{X} \le 51.25) = 1 - P(Z \le \frac{51.25 - 50}{\sqrt{0.25}}) = 1 - \Phi(2.5) = 0.0062$. Άρα, το ποσοστό των κιβωτίων που περιέχουν μήλα με μέσο βάρος μεγαλύτερο των 51.25gr είναι κατά προσέγγιση 0.62%.

β) $P(\overline{X} < 49) = P(Z < \frac{49 - 50}{\sqrt{0.25}}) = \Phi(-2) = 1 - \Phi(2) = 0.0228$. Άρα, το ποσοστό των κιβωτίων που περιέχουν μήλα με μέσο βάρος μικρότερο των 49gr είναι κατά προσέγγιση 2.28%.

Μια ισοδύναμη διατύπωση του Κ.Ο.Θ. είναι η εξής:

Αν από έναν πληθυσμό που ακολουθεί οποιαδήποτε κατανομή με μέση τιμή μ και διασπορά σ^2, επιλέξουμε τυχαία δείγματα μεγέθους n και υπολογίσουμε το άθροισμα των παρατηρήσεων κάθε δείγματος, τότε για μεγάλα n (θεωρητικά $n \to \infty$) η κατανομή αυτών των αθροισμάτων είναι κατά προσέγγιση κανονική κατανομή με μέση τιμή nμ και διασπορά $n\sigma^2$, δηλαδή, αν συμβολίσουμε με S_n την τυχαία μεταβλητή που εκφράζει αυτά τα αθροίσματα, τότε κατά προσέγγιση $S_n \sim N(n\mu,\, n\sigma^2)$.

Ας δούμε ένα παράδειγμα.

Παράδειγμα 2: Η ποσότητα ραδιενέργειας που δέχεται κάθε ημέρα κάποιος εργαζόμενος σε ένα εργοστάσιο είναι τυχαία μεταβλητή με μέση τιμή 0.1 και τυπική απόκλιση 0.01. Ποια είναι η πιθανότητα το συνολικό ποσό ραδιενέργειας που θα δεχθεί ο εργαζόμενος σε 100 ημέρες να ξεπερνάει την τιμή 10.02.

Απάντηση: Έστω X_i η ποσότητα ραδιενέργειας που δέχεται ο εργαζόμενος την i ημέρα ($i = 1, 2, ..., 100$). Η κατανομή της $X_i, i = 1, 2, ..., 100$ δεν είναι γνωστή. Είναι γνωστή μόνο η *μέση τιμή* και η *τυπική απόκλισή* της (*0.1* και *0.01* αντίστοιχα). Η συνολική ποσότητα ραδιενέργειας που θα δεχθεί ο εργαζόμενος στις 100 ημέρες είναι

$S_{100} = X_1 + X_2 + ... + X_{100}$ και επειδή το *n* είναι αρκετά μεγάλο, από το *Κ.Ο.Θ.* έχουμε ότι $S_{100} \sim N(100 \cdot 0.1, 100 \cdot 0.01^2)$ ή $S_{100} \sim N(10, 0.1^2)$. Άρα η ζητούμενη πιθανότητα είναι:

$$P(S_{100} > 10.02) = P(Z > \frac{10.02 - 10}{0.1}) = P(Z > 0.2) = 1 - \Phi(0.2) = 0.4207$$

***Παρατηρήσεις*:**

1. *Η τυπική απόκλιση της κατανομής των δειγματικών μέσων, $\sigma_{\bar{X}} = \sigma / \sqrt{n}$, ονομάζεται **τυπικό σφάλμα (standard error)**. Παρατηρείστε ότι το τυπικό σφάλμα είναι μικρότερο από την τυπική απόκλιση σ της Χ. Δηλαδή, η μεταβλητότητα των δειγματικών μέσων είναι μικρότερη από τη μεταβλητότητα του πληθυσμού από τον οποίο προέρχονται. Μάλιστα, όσο αυξάνει το μέγεθος του δείγματος, η δειγματική μεταβλητότητα (η μεταβλητότητα από δείγμα σε δείγμα των $\bar{x}$, δηλαδή, το $\sigma_{\bar{X}}$) ελαττώνεται.*
2. *Όπως ήδη αναφέραμε, όσο πιο μεγάλο είναι το μέγεθος των δειγμάτων n, τόσο καλύτερη (ακριβέστερη) είναι η προσέγγιση της κατανομής των δειγματικών μέσων από την κανονική κατανομή. Πρακτικά, όμως, **πόσο μεγάλο πρέπει να είναι το n**; Σαφής απάντηση στο ερώτημα αυτό δεν υπάρχει. Γενικά, το πόσο μεγάλο πρέπει να είναι το n, **εξαρτάται από την κατανομή του πληθυσμού**. Για παράδειγμα, αν η κατανομή του πληθυσμού είναι λοξή (ασύμμετρη) απαιτείται μεγαλύτερο μέγεθος δείγματος από αυτό που απαιτείται αν είναι περίπου συμμετρική. Γενικά, το μέγεθος του δείγματος πρέπει να είναι τουλάχιστον 30, δηλαδή, $n \geq 30$. Όμως, όπως φαίνεται και στα σχήματα της σελίδας 84, υπάρχουν περιπτώσεις όπου καλές προσεγγίσεις παίρνουμε και για μικρότερα n. Για διερεύνηση των παραπάνω, μπορείτε να πειραματιστείτε με το λογισμικό προσομοίωσης του Κ.Ο.Θ. που υπάρχει στη διεύθυνση www.aua.gr/gpapadopoulos στην επιλογή «Σημειώσεις παραδόσεων». Τέλος, επισημαίνουμε την περίπτωση όπου ο πληθυσμός από τον οποίο γίνεται η δειγματοληψία είναι κανονικός. Στην περίπτωση αυτή, όπως ήδη έχουμε αναφέρει (σελ. 80-81), αποδεικνύεται θεωρητικά ότι η κατανομή των δειγματικών μέσων (και της S_n), **είναι** κανονική κατανομή ανεξάρτητα από την τιμή του n. Δηλαδή, αποδεικνύεται ότι:*

> *Αν $X_1, X_2, ..., X_n$ ανεξάρτητες τυχαίες μεταβλητές με $X_i \sim N(\mu, \sigma^2)$ τότε,*
> $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \sim N(\mu, \frac{\sigma^2}{n})$ *και* $S_n = \sum_{i=1}^{n} X_i \sim N(n \cdot \mu, n \cdot \sigma^2)$.

Ας δούμε ένα παράδειγμα.

Παράδειγμα 3: Οι ακαθάριστες εβδομαδιαίες εισπράξεις μιας κτηνοτροφικής μονάδας από την πώληση του γάλακτος που παράγει είναι κανονική τυχαία μεταβλητή με μέση τιμή 2.200 € και τυπική απόκλιση 230 €. Ποια είναι η πιθανότητα τις επόμενες δύο εβδομάδες οι συνολικές ακαθάριστες εισπράξεις της μονάδας από την πώληση του γάλακτος που παράγει να ξεπερνούν τις 5.000 €;

Απάντηση: Έστω X_1 οι ακαθάριστες εισπράξεις από την πώληση του γάλακτος την πρώτη εβδομάδα και X_2 οι ακαθάριστες εισπράξεις από την πώληση του γάλακτος τη δεύτερη εβδομάδα. Δίνεται ότι $X_1 \sim N(2.200, 230^2)$ και $X_2 \sim N(2.200, 230^2)$. Οι συνολικές εισπράξεις στις δύο εβδομάδες είναι $S_2 = X_1 + X_2$. Με την υπόθεση ότι οι εισπράξεις από εβδομάδα σε εβδομάδα είναι ανεξάρτητες μεταξύ τους έχουμε ότι $S_2 \sim N(4.400, 2 \cdot 230^2)$. Η ζητούμενη πιθανότητα είναι:

$$P(S_2 > 5.000) = P(Z > \frac{5.000 - 4.400}{\sqrt{2 \cdot 230^2}}) = P(Z > 1.84) = 1 - \Phi(1.84) = 0.0329 .$$

3. *Στα προηγούμενα δε δώσαμε έναν αυστηρό ορισμό για την έννοια του τυχαίου δείγματος μεγέθους n από έναν πληθυσμό. Θα το κάνουμε στη Στατιστική Συμπερασματολογία. Παρόλα αυτά, από όσα στα προηγούμενα έχουμε αναφέρει (βασικά, υπαινιχθεί), είναι (μάλλον) φανερό ότι λέγοντας, «από έναν πληθυσμό παίρνουμε ένα τυχαίο δείγμα μεγέθους n» εννοούμε n ανεξάρτητες τυχαίες μεταβλητές* $X_1, X_2, ..., X_n$ *που ακολουθούν την ίδια κατανομή (αυτήν του πληθυσμού από τον οποίο παίρνουμε το δείγμα). Οι τιμές,* $x_1, x_2, ..., x_n$*, που έχουμε διαθέσιμες για επεξεργασία <u>μετά</u> τη λήψη του δείγματος αποτελούν μια πραγματοποίηση του δείγματος, δηλαδή μια πραγματοποίηση των n ανεξάρτητων και ισόνομων τυχαίων μεταβλητών* $X_1, X_2, ..., X_n$.

<u>Κανονική προσέγγιση της Διωνυμικής Κατανομής</u>

Ένα σημαντικό αποτέλεσμα της θεωρίας πιθανοτήτων, γνωστό ως *οριακό θεώρημα των De Moivre-Laplace*, μας λέει ότι:

Για μεγάλα n (θεωρητικά $n \to \infty$*), η διωνυμική κατανομή μπορεί να προσεγγισθεί από μια κανονική κατανομή με ίδια μέση τιμή και ίδια διασπορά. Δηλαδή, αν* $X \sim B(n, p)$ *τότε η κατανομή της X προσεγγίζεται, για μεγάλες τιμές του n, από την* $N(\mu, \sigma^2)$ *με* $\mu = np$ *και* $\sigma^2 = np(1-p)$.

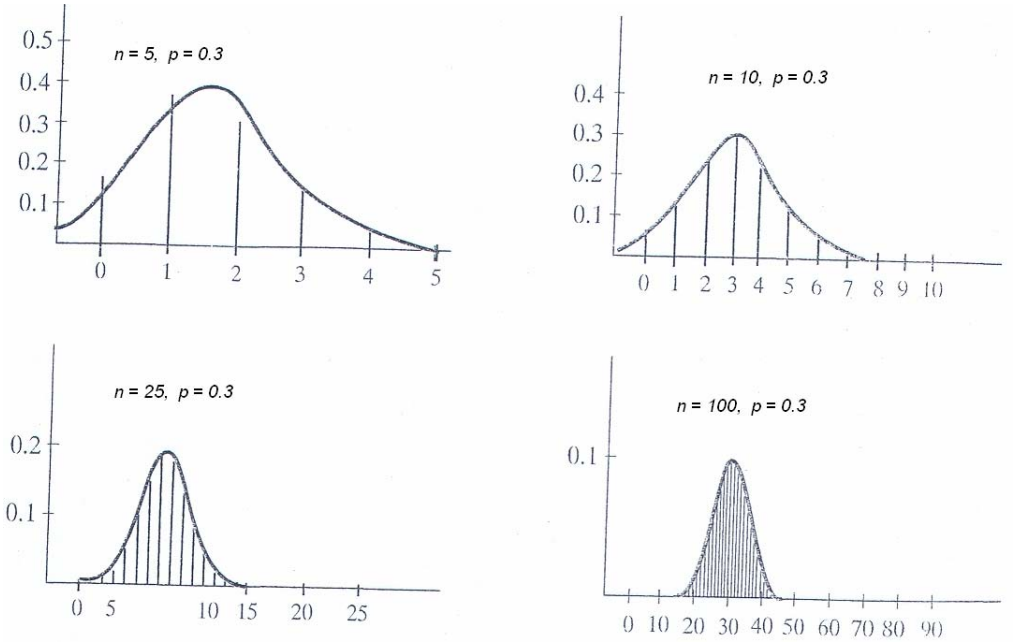
Το αποτέλεσμα αυτό, μπορεί εύκολα να προκύψει ως ειδική περίπτωση του *Κ.Ο.Θ*[1]. Όμως, δεν προέκυψε έτσι. Αντιθέτως, αποδεικνύοντάς το, ο *Abraham De Moivre* το 1733 για $p = \frac{1}{2}$ και, εκατό περίπου χρόνια αργότερα, το 1812, ο *Laplace* για κάθε $p \in (0, 1)$, έθεσαν τις βάσεις για τη διατύπωση και απόδειξη του *Κ.Ο.Θ*. Δηλαδή, η πορεία ήταν αντίστροφη. Πρώτα αποδείχθηκε το *οριακό θεώρημα* των *De Moivre-Laplace,* και πολύ αργότερα διατυπώθηκε και αποδείχθηκε το *Κ.Ο.Θ*. από τον Ρώσο μαθηματικό *Lyapunov* την περίοδο 1901-1902 (είχαν προηγηθεί και άλλες γενικεύσεις του *οριακού θεωρήματος* των *De Moivre-Laplace* από τους *Chebysev* και *Markov*). Μάλιστα, το *οριακό θεώρημα* των *De Moivre-Laplace* προέκυψε –όπως και το *οριακό θεώρημα Poisson* - από την ανάγκη αντιμετώπισης των δυσκολιών που παρουσιάζονται στον υπολογισμό πιθανοτήτων της *διωνυμικής κατανομής*. Έτσι «γεννήθηκε» και η *κανονική κατανομή* (όπως και η κατανομή *Poisson* από το *οριακό θεώρημα Poisson*). Δηλαδή, οι δυσκολίες της *διωνυμικής* «γέννησαν» δύο διάσημες κατανομές!

Πριν δώσουμε δύο παραδείγματα εφαρμογής, σε πρακτικά προβλήματα, της κανονικής προσέγγισης της *διωνυμικής κατανομής*, ας δούμε πάλι το ερώτημα: ***πόσο μεγάλο πρέπει να είναι το n*** για να μπορεί, πρακτικά, να χρησιμοποιηθεί κανονική προσέγγιση της *διωνυμικής κατανομής*. Όπως φαίνεται και στα σχήματα που υπάρχουν στην επόμενη σελίδα, το πόσο μεγάλο πρέπει να είναι το *n*, ***εξαρτάται από την τιμή της παραμέτρου p***. Αν, για παράδειγμα, $p = \frac{1}{2}$ ή κοντά στο $\frac{1}{2}$, τότε και για όχι πολύ μεγάλες τιμές του *n* παίρνουμε εξαιρετικές προσεγγίσεις της *διωνυμικής*. Αντίθετα, αν το *p* είναι <u>πολύ μικρό ή πολύ μεγάλο</u>, για καλή προσέγγιση, απαιτούνται πολύ μεγαλύτερες τιμές του *n*. Ένας πρακτικός, γενικός, κανόνας είναι ο εξής:

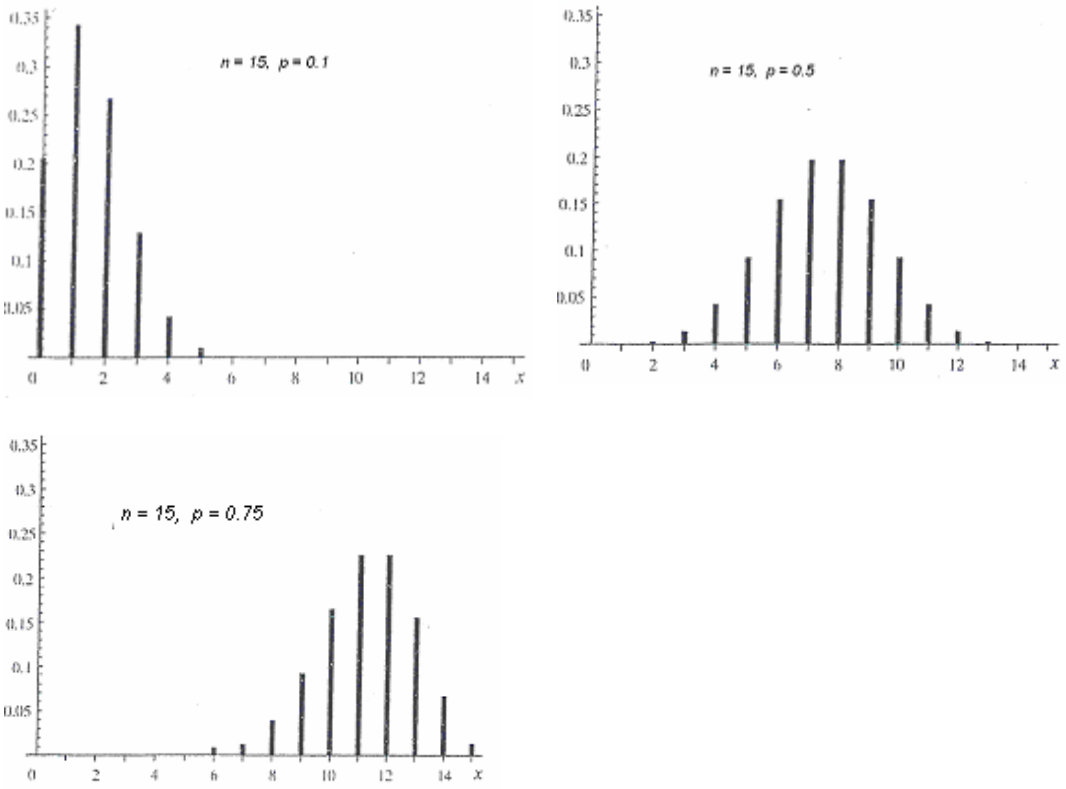
[1] *Παρατηρείστε ότι η διωνυμική κατανομή αναφέρεται στο άθροισμα n ανεξάρτητων δίτιμων τυχαίων μεταβλητών.*

Για να πάρουμε, μέσω του οριακού θεωρήματος De Moivre-Laplace, καλές προσεγγίσεις της διωνυμικής κατανομής αρκεί $n \cdot p \cdot (1-p) \geq 10$. *Επίσης, συχνά, στη βιβλιογραφία συναντάται και ο κανόνας:* $n \cdot p > 5$ ***και*** $n \cdot (1-p) > 5$.

Στα τέσσερα σχήματα που ακολουθούν δίνεται η συνάρτηση πιθανότητας της *διωνυμικής κατανομής* με $p = 0.3$ (σταθερό) και $n = 5, 10, 25, 100$ καθώς και η συνάρτηση πυκνότητας της *κανονικής κατανομής* με την αντίστοιχη *μέση τιμή* και *διασπορά*, δηλαδή, με $\mu = np$ και $\sigma^2 = np(1-p)$. Παρατηρείστε ότι όσο μεγαλύτερο γίνεται το *n* τόσο πιο καλή και η προσέγγιση.



Στα τρία επόμενα σχήματα δίνεται η συνάρτηση πιθανότητας της *διωνυμικής κατανομής* με $n = 15$ (σταθερό) και $p = 0.1, 0.5, 0.7$. Παρατηρείστε τη μορφή της κατανομής για $p = 0.5$.



Παράδειγμα 4: Ο ιδανικός αριθμός πρωτοετών φοιτητών σε ένα πανεπιστήμιο είναι 150. Το πανεπιστήμιο, γνωρίζοντας από προηγούμενη εμπειρία ότι από τους φοιτητές που κάνει δεκτούς για εγγραφή, μόνο το 30% παρακολουθεί τα μαθήματα, εφαρμόζει την εξής πολιτική: κάνει δεκτούς 450 φοιτητές. Ποια είναι η πιθανότητα (κατά προσέγγιση), από τους 450 πρωτοετείς φοιτητές, να παρακολουθούν τελικά τα μαθήματα περισσότεροι από 150.

Απάντηση: Αν X ο αριθμός των πρωτοετών φοιτητών (από τους 450) που παρακολουθούν τα μαθήματα, τότε προφανώς $X \sim B(n,\ p)$ με $n = 450$ και $p = 0.3$, δηλαδή, $X \sim B(450,\ 3)$. Επειδή, το n είναι μεγάλο και $n \cdot p \cdot (1-p) = 450 \cdot 0.3 \cdot 0.7 \geq 10$, η X προσεγγίζεται ικανοποιητικά από την *κανονική* με $\mu = n \cdot p = 450 \cdot 0.3 = 135$ και $\sigma^2 = n \cdot p \cdot (1-p) = 450 \cdot 0.3 \cdot 0.7 = 94.5$ δηλαδή από την $N(135,\ 94.5)$. Άρα, για τη ζητούμενη πιθανότητα έχουμε:

$$P(X \geq 151) = P\left(\frac{X-135}{\sqrt{94.5}} \geq \frac{151-135}{\sqrt{94.5}}\right) = P(Z \geq 1.64) = 1 - \Phi(1.64) = 0.0505\,.$$

Όπως ήδη έχουμε αναφέρει στην *εισαγωγή*, ένα κρίσιμο θέμα στη στατιστική προσέγγιση προβλημάτων είναι το μέγεθος *n* του δείγματος που επιλέγουμε. Τα δύο παραδείγματα που ακολουθούν δίνονται για να πάρουμε μια πρώτη ιδέα για το πώς μπορούμε να εφαρμόσουμε αποτελέσματα της θεωρίας πιθανοτήτων για τον καθορισμό του κατάλληλου μεγέθους δείγματος.

Παράδειγμα 5: Προκειμένου να εκτιμήσουμε το ποσοστό p των ατόμων ενός πληθυσμού που έχουν μια συγκεκριμένη ιδιότητα (π.χ. καπνίζουν, πάσχουν από μια ασθένεια, είναι άνεργοι, ψηφίζουν ένα συγκεκριμένο κόμμα κτλ.) χρησιμοποιούμε ένα δείγμα μεγέθους n. Πόσο πρέπει να είναι το n έτσι ώστε το ποσοστό των ατόμων του δείγματος που έχουν την ιδιότητα, να διαφέρει, κατ' απόλυτη τιμή, από το (άγνωστο) πραγματικό ποσοστό p λιγότερο από 1% με πιθανότητα τουλάχιστον 95%.

Απάντηση: Ας συμβολίσουμε με X τον αριθμό των ατόμων του δείγματος που έχουν τη συγκεκριμένη ιδιότητα. Η τυχαία μεταβλητή X ακολουθεί τη *διωνυμική κατανομή*[2] με παραμέτρους *n* και *p*, δηλαδή, $X \sim B(n,\ p)$. Προφανώς, το ποσοστό των ατόμων του δείγματος που έχουν τη συγκεκριμένη ιδιότητα είναι $\frac{X}{n}$. Σύμφωνα με τις απαιτήσεις που θέτει το πρόβλημα πρέπει:

$$P\left(\left|\frac{X}{n} - p\right| < 0.01\right) \geq 0.95 \text{ ή } P\left(-0.01 < \frac{X}{n} - p < 0.01\right) \geq 0.95 \text{ ή}$$

$$P(-0.01n < X - np < 0.01n) \geq 0.95 \text{ ή } P(np - 0.01n < X < np + 0.01n) \geq 0.95\,.$$

Χρησιμοποιώντας την κανονική προσέγγιση της *διωνυμικής* έχουμε:

$$P(np - 0.01n < X < np + 0.01n) = P\left(\frac{(np-0.01n) - np}{\sqrt{np(1-p)}} < \frac{X - np}{\sqrt{np(1-p)}} < \frac{(np+0.01n) - np}{\sqrt{np(1-p)}}\right) =$$

$$= 2\Phi\left(\frac{0.01n}{\sqrt{np(1-p)}}\right) - 1.$$ Άρα, σύμφωνα με τις απαιτήσεις που θέτει το πρόβλημα,

[2] *Στις περιπτώσεις αυτές, το μέγεθος του δείγματος είναι μικρό σε σχέση με το μέγεθος του πληθυσμού και επομένως μπορούμε να υποθέσουμε ότι η δειγματοληψία γίνεται με επανάθεση.*

πρέπει: $2\Phi\left(\frac{0.01n}{\sqrt{np(1-p)}}\right)-1\geq 0.95$ *ή ισοδύναμα* $\Phi\left(\frac{0.01n}{\sqrt{np(1-p)}}\right)\geq 0.975$ *και επομένως* $\frac{0.01n}{\sqrt{np(1-p)}}\geq 1.96$ *ή* $n\geq 38416\,p(1-p)$.

Επειδή το πραγματικό ποσοστό στον πληθυσμό, p, είναι άγνωστο, για να εξασφαλίσουμε για κάθε τιμή του p την ισχύ της τελευταίας ανισότητας πρέπει να βρούμε τη μέγιστη τιμή του $p(1-p)$. Αυτή είναι $\frac{1}{4}$ (γιατί;)[3] άρα πρέπει $n\geq 38416\cdot\frac{1}{4}$, δηλαδή, το δείγμα πρέπει να έχει μέγεθος τουλάχιστον 9604.

<u>*Παράδειγμα 6:*</u> *Ένας αστρονόμος θέλει να μετρήσει (σε έτη φωτός) την απόσταση μεταξύ του αστεροσκοπείου που εργάζεται και ενός άστρου. Παρότι εφαρμόζει μια αναγνωρισμένη μέθοδο μέτρησης, γνωρίζει ότι κάθε φορά που μετράει την απόσταση δεν παίρνει την πραγματική τιμή της αλλά μόνο μια εκτίμησή της (Αυτό συμβαίνει για διάφορους λόγους, όπως αλλαγές στις ατμοσφαιρικές συνθήκες, κ.ά.). Γι' αυτό σχεδιάζει να κάνει έναν αριθμό μετρήσεων n, να υπολογίσει τη μέση τιμή τους και να τη χρησιμοποιήσει ως εκτιμήτρια της άγνωστης πραγματικής απόστασης d. Αν οι n μετρήσεις,* $X_1, X_2, ..., X_n$, *είναι ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν την ίδια (άγνωστη) κατανομή με ίδια μέση τιμή d (την άγνωστη πραγματική απόσταση) και ίδια διασπορά (4 έτη φωτός), πόσες μετρήσεις πρέπει να κάνει ο αστρονόμος ώστε η μέση τιμή τους να διαφέρει, κατ' απόλυτη τιμή, από την άγνωστη πραγματική απόσταση d, λιγότερο από 0.5 έτη φωτός με πιθανότητα 95%.*

<u>*Υπόδειξη:*</u> Σύμφωνα με τις απαιτήσεις που θέτει το πρόβλημα πρέπει:

$P\left(\left|\frac{\sum_{i=1}^{n}X_i}{n}-d\right|<0.5\right)=0.95$. Άρα .....(και βρίσκουμε ότι πρέπει n = 62)

<u>Κανονική προσέγγιση της Κατανομής Poisson</u>

Με εφαρμογή του *Κ.Ο.Θ.,* αποδεικνύεται ότι και η *κατανομή Poisson* μπορεί να προσεγγισθεί ικανοποιητικά από την *κανονική κατανομή*. Πιο συγκεκριμένα, αν X είναι μια τυχαία μεταβλητή που ακολουθεί την *κατανομή Poisson* με παράμετρο λ, τότε η κατανομή της X προσεγγίζεται, για μεγάλες τιμές του λ (στην πράξη για $\lambda>10$), από την $N(\mu,\sigma^2)$ με $\mu=\lambda$ και $\sigma^2=\lambda$.

Ας δούμε ένα παράδειγμα.

<u>*Παράδειγμα 7:*</u> *Σε μια αγροτική καλλιέργεια κηπευτικών, έχει παρατηρηθεί ότι ο αριθμός των φυτών που δεν αναπτύσσονται (ξεραίνονται) είναι τυχαία μεταβλητή X που ακολουθεί κατανομή Poisson με μέση τιμή* $\lambda=100$ *φυτά/καλλιεργητική περίοδο. Ποια είναι η πιθανότητα σε μια καλλιεργητική περίοδο ο αριθμός των φυτών που δεν θα αναπτυχθούν να είναι τουλάχιστον 120.*

<u>*Απάντηση:*</u> Επειδή η τιμή του λ είναι μεγάλη, η κατανομή της X προσεγγίζεται ικανοποιητικά από την *κανονική* με $\mu=\lambda=100$ και $\sigma^2=\lambda=100$, δηλαδή, από την $N(100,100)$. Άρα, για τη ζητούμενη πιθανότητα έχουμε:

$$P(X\geq 120)=P\left(\frac{X-100}{\sqrt{100}}\geq\frac{120-100}{\sqrt{100}}\right)=P(Z\geq 2)=1-\Phi(2)=0.0228.$$

[3] *Πρόκειται για την εύρεση ακρότατης τιμής συνάρτησης του p.*

Διόρθωση συνέχειας

Τόσο στην περίπτωση της κανονικής προσέγγισης της *διωνυμικής κατανομής* όσο και στην περίπτωση της κανονικής προσέγγισης της *κατανομής Poisson*, γίνεται προσέγγιση *διακριτής* κατανομής από *συνεχή*. Στις περιπτώσεις αυτές (που γίνεται προσέγγιση *διακριτής* κατανομής από *συνεχή*), καλό είναι να ενσωματώνεται στον προσεγγιστικό τύπο η λεγόμενη ***διόρθωση συνέχειας***. Έτσι, όταν, για παράδειγμα, η *διωνυμική* $X \sim B(n,\ p)$ προσεγγίζεται από την $N\big(np,\ np(1-p)\big)$ οι πιθανότητες

$$P(a \le X \le b),\ P(X \le b) \text{ και } P(X \ge a)$$

με διόρθωση συνέχειας υπολογίζονται ως εξής:

$$P(a \le X \le b) \cong P(a-0.5 \le X \le b+0.5) =$$

$$= P\left(\frac{a-0.5-np}{\sqrt{np(1-p)}} \le \frac{X-np}{\sqrt{np(1-p)}} \le \frac{b+0.5-np}{\sqrt{np(1-p)}}\right) = \Phi\left(\frac{b+0.5-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a-0.5-np}{\sqrt{np(1-p)}}\right).$$

$$P(X \le b) \cong P(X \le b+0.5) = P\left(\frac{X-np}{\sqrt{np(1-p)}} \le \frac{b+0.5-np}{\sqrt{np(1-p)}}\right) = \Phi\left(\frac{b+0.5-np}{\sqrt{np(1-p)}}\right).$$

$$P(X \ge a) \cong P(X \ge a-0.5) = P\left(\frac{X-np}{\sqrt{np(1-p)}} \ge \frac{a-0.5-np}{\sqrt{np(1-p)}}\right) = 1-\Phi\left(\frac{a-0.5-np}{\sqrt{np(1-p)}}\right).$$

Χρησιμοποιώντας στο *Παράδειγμα 4* τη διόρθωση συνέχειας έχουμε:

$$P(X \ge 151) \cong P(X \ge 151-0.5) = P\left(\frac{X-135}{\sqrt{94.5}} > \frac{150.5-135}{\sqrt{94.5}}\right) = P(Z > 1.59) = 0.0559$$

Ομοίως, χρησιμοποιώντας στο *Παράδειγμα 7* τη διόρθωση συνέχειας έχουμε:

$$P(X \ge 120) \cong P(X \ge 120-0.5) = P\left(\frac{X-100}{\sqrt{100}} \ge \frac{119.5-100}{\sqrt{100}}\right) = P(Z \ge 1.95) = 0.0256.$$

Σημειώνουμε, τέλος, ότι, χρησιμοποιώντας τη διόρθωση συνέχειας, μπορούμε να υπολογίζουμε και τις πιθανότητες $P(X = a),\ a = 0,1,....$ μιας *διακριτής* τυχαίας μεταβλητής X, μέσω της *κανονικής κατανομής*, ως εξής:

$$P(X = a) \cong P(a-0.5 \le X \le a+0.5) = ...$$

Σχόλιο: *Το πόσο σημαντικό είναι το γεγονός ότι το Κ.Ο.Θ. μας πληροφορεί για την κατανομή της $\overline{X}$ και της S_n, θα φανεί και στη συνέχεια όταν αναφερθούμε σε ζητήματα Στατιστικής Συμπερασματολογίας (π.χ. διαστήματα εμπιστοσύνης, στατιστικοί έλεγχοι).*

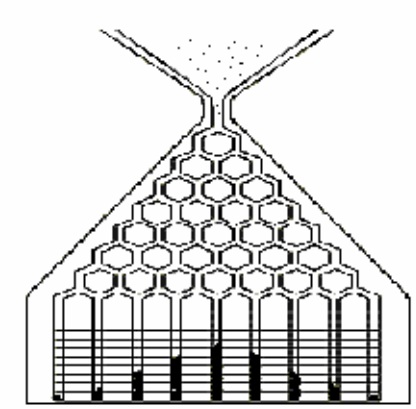
Για προβληματισμό

1. *Τα casino είναι επιχειρήσεις και επομένως μια επένδυση για τη δημιουργία ενός casino, προϋποθέτει, μεταξύ άλλων, προβλεψιμότητα των κερδών. Έχετε σκεφθεί, πώς γίνεται να επαφίεται η βιωσιμότητα μιας τέτοιας επένδυσης στην «τύχη»;*

2. *Σε πολλά μουσεία επιστημών, μεταξύ των διαφόρων εκθεμάτων (κατασκευών-μοντέλων), υπάρχει και η μηχανή του Galton*[4]*.*

[4] *Ο Sir Galton (1822-1911) ήταν Άγγλος ανθρωπολόγος ο οποίος, σε συνεργασία με τον Karl Pearson, χρησιμοποίησε την κανονική κατανομή για τη μελέτη και την ερμηνεία ανθρωπομετρικών δεδομένων.*

Ο τρόπος λειτουργίας της, φαίνεται καλύτερα στο επόμενο σκίτσο.



Μπορείτε να εξηγήσετε τι προσομοιώνει η μηχανή του Galton;
<u>Υπόδειξη:</u> Μπορείτε να βρείτε, μέσω του Internet, Java applets τα οποία προσομοιώνουν τη μηχανή του Galton σε υπολογιστικό περιβάλλον. Η παρακάτω εικόνα είναι ένα στιγμιότυπο από ένα τέτοιο πρόγραμμα.