# *Βασικές συνεχείς κατανομές και το Κεντρικό Οριακό Θεώρημα*

Στο *5ο Κεφάλαιο*, όταν στην *Ενότητα 5.4* μιλήσαμε για τις *συνεχείς* τυχαίες μεταβλητές, είδαμε ότι μια τυχαία μεταβλητή, έστω *X*, με *συνάρτηση πυκνότητας*

$$f(x)=\begin{cases}\dfrac{1}{\beta-\alpha}, & \alpha \le x \le \beta \\ 0, & \text{αλλού}\end{cases}$$

ονομάζεται ***ομοιόμορφη συνεχής τυχαία μεταβλητή*** και δείξαμε *(Παράδειγμα 5.4.4)* ότι έχει *μέση τιμή*

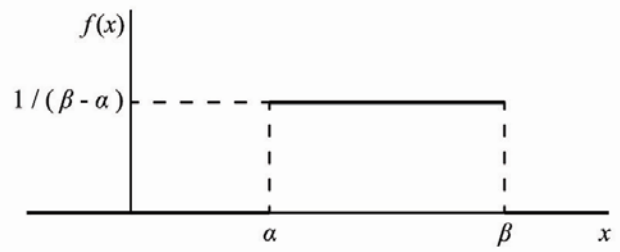
$$\mu = E(X) = \frac{\alpha+\beta}{2}$$

και *διακύμανση*

$$\sigma^2 = Var(X) = \frac{(\beta-\alpha)^2}{12}.$$

Εξηγήσαμε επίσης, ότι πρόκειται για ένα μοντέλο πιθανοτήτων το οποίο σε ίσου πλάτους υποδιαστήματα του $[\alpha,\beta]$ εκχωρεί ίσες πιθανότητες, ή αλλιώς, η πιθανότητα που εκχωρεί σε ένα οποιοδήποτε υποδιάστημα του $[\alpha,\beta]$ είναι ανάλογη του πλάτους του, αφού στο $[\alpha,\beta]$ η *f* είναι σταθερή *(Σχήμα 7.1)*.

Μια τέτοια κατανομή/μοντέλο πιθανοτήτων ονομάζεται ***ομοιόμορφη συνεχής κατανομή***, συμβολίζεται με $U(\alpha,\beta)$ και είναι μια από τις βασικές και με ενδιαφέρουσες εφαρμογές συνεχείς κατανομές[1].




*Σχήμα 7.1*
*Η συνάρτηση πυκνότητας της*
*ομοιόμορφης κατανομής συνεχούς τ.μ.*

Στο κεφάλαιο αυτό, στις *Ενότητες 7.1&7.3* θα γνωρίσουμε τέσσερις ακόμη βασικές συνεχείς κατανομές. Την ***κανονική κατανομή***, την ***κατανομή χ²***, την ***κατανομή t (Student)*** και την ***κατανομή F***. Πρόκειται για μοντέλα πιθανοτήτων τα οποία έχουν μελετηθεί συστηματικά, χρησιμοποιούνται ευρύτατα στη *στατιστική συμπερασματολογία* και καλύπτουν πολύ μεγάλο φάσμα εφαρμογών[2]. Βέβαια, όπως εξηγήσαμε στην *Ενότητα 5.4 (Σχόλιο 5.4.2β)*, αλλά και όπως στη συνέχεια θα διαπιστώσουμε, σε αντίθεση με τα διακριτά μοντέλα πιθανοτήτων, τα συνεχή μοντέλα κατά κανόνα δε «γεννιούνται» από την περιγραφή αντίστοιχων τυχαίων πειραμάτων, αλλά εισάγονται απευθείας μέσω μιας *συνάρτησης πυκνότητας*. Έτσι, το αν μια συνεχής τυχαία μεταβλητή περιγράφεται από συγκεκριμένη *συνάρτηση πυκνότητας*, όπως θα διαπιστώσουμε, είναι «προς απόδειξη».

[1] *Για παράδειγμα, ίσως το πιο χαρακτηριστικό, η ομοιόμορφη κατανομή είναι κατάλληλη για την περιγραφή τυχαίων μεταβλητών που εκφράζουν χρόνους αναμονής σε κυκλικές διαδικασίες.*

[2] *Άλλες συνεχείς κατανομές, στις οποίες όμως δε θα αναφερθούμε, είναι (μεταξύ άλλων) η* ***εκθετική κατανομή****, η* ***κατανομή Γάμμα*** *και η* ***κατανομή Βήτα****.*

## *7.1 Κανονική κατανομή*

Η ***κανονική κατανομή (normal distribution)*** θεωρείται η σπουδαιότερη κατανομή της *Θεωρίας Πιθανοτήτων* και της *Στατιστικής*. Οι λόγοι που εξηγούν την εξέχουσα θέση της είναι βασικά δύο.

*i) Πολλές τυχαίες μεταβλητές περιγράφονται ικανοποιητικά από την κανονική κατανομή ή περιγράφονται από κατανομές που μπορούν να προσεγγισθούν από την κανονική κατανομή.*

*ii) Οι ιδιότητες της κανονικής κατανομής αξιοποιούνται στη στατιστική συμπερασματολογία. Ουσιαστικά, η κανονική κατανομή αποτελεί το θεμέλιο της στατιστικής συμπερασματολογίας.*

Στο Β΄ Μέρος, θα έχουμε την ευκαιρία να διαπιστώσουμε πόσο σημαντική είναι η *κανονική κατανομή* στη *στατιστική συμπερασματολογία*. Προς το παρόν, ας σταθούμε λίγο περισσότερο στον πρώτο από τους παραπάνω λόγους. Ας προσπαθήσουμε δηλαδή, να εξηγήσουμε γιατί η *κανονική κατανομή* βρίσκει εφαρμογή σε πολλά στοχαστικά φαινόμενα και πειράματα.

Το «μυστικό» που εξηγεί το μεγάλο εύρος εφαρμογών της *κανονικής κατανομής,* βρίσκεται σε ένα εκπληκτικά ισχυρό θεωρητικό αποτέλεσμα της *Θεωρίας Πιθανοτήτων* το οποίο επιβεβαιώνεται και πειραματικά. Πρόκειται για το ***Κεντρικό Οριακό Θεώρημα*** τις βάσεις του οποίου έθεσαν δύο μεγάλοι Μαθηματικοί. Ο *Abraham De Moivre* το 1733 και έναν αιώνα περίπου αργότερα, το 1812, ο *Pierre-Simon Laplace*. Σε αυτό το σημείο δε θα διατυπώσουμε αυστηρά ούτε θα αποδείξουμε το *Κεντρικό Οριακό Θεώρημα*. Θα προσπαθήσουμε να εξηγήσουμε μόνο το νόημα και τη σημασία του. Αργότερα, θα δώσουμε μια πληρέστερη διατύπωση.

Σύμφωνα με το *Κεντρικό Οριακό Θεώρημα*, το άθροισμα και – επομένως - η *μέση τιμή*, μεγάλου αριθμού ανεξάρτητων παρατηρήσεων, ακολουθεί κατά προσέγγιση *κανονική κατανομή*, ανεξαρτήτως από το ποια κατανομή ακολουθούν οι παρατηρήσεις. Πώς όμως, αυτό το αποτέλεσμα ερμηνεύει τη μεγάλη εφαρμοσιμότητα της *κανονικής κατανομής*; Είναι απλό. Σε πολλά φαινόμενα και πειράματα, οι τιμές διαφόρων χαρακτηριστικών (μεταβλητών) είναι αποτέλεσμα αθροιστικής επίδρασης πολλών ανεξάρτητων αιτίων-παραγόντων κανένα από τα οποία δεν υπερισχύει των άλλων. Για παράδειγμα, ο χρόνος αναμονής σε μια ουρά είναι αποτέλεσμα πολλών παραγόντων όπως, η ημέρα της εβδομάδας, η ώρα της ημέρας, η αποτελεσματικότητα του υπαλλήλου, το είδος της συναλλαγής που διεκπεραιώνεται, κ.ά. Επίσης, το βάρος των ζώων μιας κτηνοτροφικής μονάδας, οφείλεται σύμφωνα με τους ειδικούς, σε πληθώρα παραγόντων όπως, η ατομικότητα του ζώου, η φυλή, το γένος, οι συνθήκες διατροφής, οι συνθήκες ενσταυλισμού, κ.ά. Καθένας από τους παράγοντες αυτούς επιφέρει ένα θετικό ή αρνητικό αποτέλεσμα και όλοι μαζί αθροιστικά συντελούν στη διαμόρφωση του τελικού αποτελέσματος. Τέτοια χαρακτηριστικά (μεταβλητές), εμφανίζονται σε πολλά φαινόμενα και πειράματα. Το *Κεντρικό Οριακό Θεώρημα* μας διαβεβαιώνει ότι αυτά ακριβώς τα χαρακτηριστικά περιγράφονται ικανοποιητικά από την *κανονική κατανομή*. Επιπλέον, το *Κεντρικό Οριακό Θεώρημα* **συνδέει** την *κανονική κατανομή* με **οποιαδήποτε άλλη κατανομή** (αφού δεν προϋποθέτει να ακολουθούν οι παρατηρήσεις *κανονική κατανομή),* γεγονός το οποίο απαντάει επίσης στο ερώτημα, γιατί η *κανονική κατανομή* βρίσκει εφαρμογή σε μεγάλο πλήθος φαινομένων και πειραμάτων.

Πρέπει να τονίσουμε ότι για να αποδειχθεί ότι ένα συγκεκριμένο χαρακτηριστικό (μεταβλητή) προσεγγίζεται ικανοποιητικά από την *κανονική κατανομή*, πρέπει να γίνουν μετρήσεις που να επαληθεύουν ένα τέτοιο συμπέρασμα[3]. Μια από τις πρώτες εφαρμογές της *κανονικής κατανομής*, έγινε το 1809 από το μεγάλο Γερμανό

[3] *Θυμηθείτε το Σχόλιο 5.4.2 (β) και επίσης δείτε το Σχόλιο 7.1.1 στη συνέχεια.*

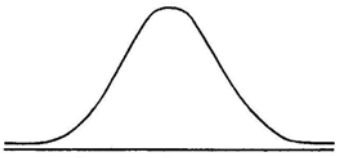
Μαθηματικό C*arl F. Gauss* ο οποίος διαπίστωσε ότι τα σφάλματα που γίνονται σε αστρονομικές παρατηρήσεις μπορούν να περιγραφούν ικανοποιητικά από την *κανονική κατανομή*. Στη συνέχεια, διαπιστώθηκε επίσης, ότι τα τυχαία σφάλματα (όχι τα συστηματικά) που εμφανίζονται σε διάφορες μετρήσεις ακολουθούν με ικανοποιητική προσέγγιση *κανονική κατανομή*. Για το λόγο αυτό, η *κανονική κατανομή* ονομάζεται και ***κατανομή των σφαλμάτων (law of errors)***. Επίσης, είναι γνωστή ως ***κατανομή του Gauss (Gaussian distribution)***, για τη μεγάλη συνεισφορά του *Gauss* στην ανάδειξη των ιδιοτήτων και της σημασίας της. ***Κανονική κατανομή*** ονομάσθηκε στις αρχές του 20ου αιώνα από τον *Pearson*. Όμως, για το πώς και από ποιόν εισήχθη η *κανονική κατανομή*, θα αναφερθούμε αργότερα όταν μιλήσουμε πιο αναλυτικά για το *Κεντρικό Οριακό Θεώρημα*. Τέλος, ως πρόσθετη σχετική πληροφορία[4], αναφέρουμε ότι στο γερμανικό χαρτονόμισμα των δέκα μάρκων υπήρχαν, φωτογραφία του *Gauss*, η κανονική καμπύλη και ο μαθηματικός τύπος της!!


Η ***συνάρτηση πυκνότητας*** της *κανονικής κατανομής* δίνεται από τον τύπο,

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

όπου, $\sigma > 0$ η *τυπική απόκλιση* και $\mu \in (-\infty, +\infty)$ η *μέση τιμή* της κατανομής[5].

Η γραφική της παράσταση *(Σχήμα 7.1.1.)* είναι γνωστή ως ***κανονική καμπύλη*** και έχει κωδωνοειδή μορφή.



*Σχήμα 7.1.1*
*Η κανονική καμπύλη*

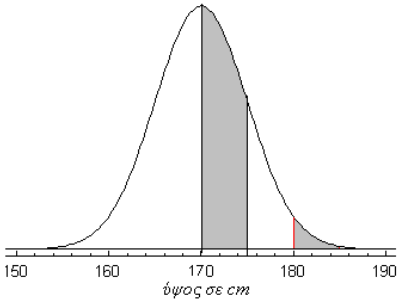
Παρατηρείστε ότι στον τύπο της *συνάρτησης πυκνότητας* της *κανονικής κατανομής*, εμφανίζονται δύο πολύ «διάσημοι» άρρητοι αριθμοί: ο $\pi \cong 3.14$ και ο $e \cong 2.71$.

### *7.1.1 Ιδιότητες της κανονικής καμπύλης*

Η *κανονική καμπύλη* είναι συμμετρική και οι «ουρές» της πλησιάζουν τον οριζόντιο άξονα ομαλά (ασυμπτωτικά). Η *μέση τιμή* και η *διάμεσος* ταυτίζονται. Επίσης, η *κορυφή* ταυτίζεται με τη *μέση τιμή* και τη *διάμεσο*. Έτσι, η περιοχή που παρουσιάζει τη μεγαλύτερη *πυκνότητα*, βρίσκεται και αυτή στο μέσο της κατανομής. Δηλαδή, όταν οι τιμές μιας μεταβλητής είναι κανονικά κατανεμημένες, τότε γύρω από τη *μέση τιμή* τους υπάρχουν σχετικά πολλές τιμές ενώ μακριά από τη *μέση τιμή* βρίσκονται σχετικά λίγες τιμές. Για παράδειγμα, αν το ύψος των ελλήνων ηλικίας 18 έως 25 ετών είναι κανονικά κατανεμημένο, με *μέση τιμή* 170cm και *τυπική απόκλιση* 5cm *(δες Σχήμα 7.1.2)*, τότε μεταξύ 170cm και 175cm βρίσκονται περισσότερα άτομα από όσα

[4] *Ενδεικτική της αναγνώρισης της σημασίας της κανονικής κατανομής και του έργου του Gauss.*
[5] *Αυτός ο τύπος υπήρχε στο χαρτονόμισμα των δέκα μάρκων!!*

βρίσκονται μεταξύ 180cm και 185cm. Επίσης, πολύ λίγα άτομα έχουν ύψος μεγαλύτερο από 185cm ή μικρότερο από 155cm.

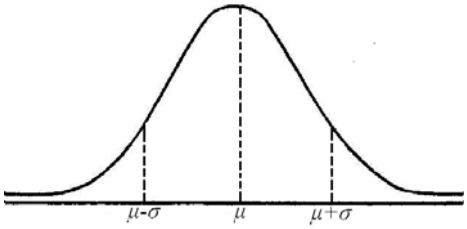



*Σχήμα 7.1.2*
*Το ύψος παίρνει τιμές κοντά στη μέση τιμή με πολύ μεγαλύτερη πιθανότητα απ' ότι μακριά από τη μέση τιμή (προς τις ουρές)*

Παρατηρείστε *(δες Σχήμα 7.1.3)* ότι η καμπύλη της *συνάρτησης πυκνότητας* της *κανονικής κατανομής*, στη θέση $x = \mu$ παρουσιάζει μέγιστη τιμή, ίση με

$$\frac{1}{\sigma\sqrt{2\pi}} = \frac{0.399}{\sigma}$$

και στις θέσεις $x = \mu - \sigma$ και $x = \mu + \sigma$ παρουσιάζει σημεία καμπής.



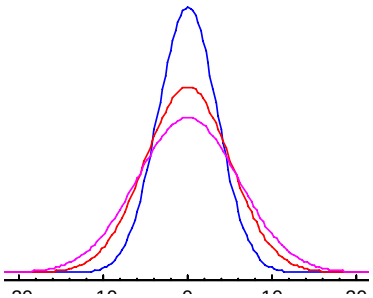

*Σχήμα 7.1.3*
*Οι θέσεις των σημείων καμπής (και της κορυφής) της κανονικής καμπύλης*

***Ερώτηση:*** *Αν μια κανονική κατανομή έχει, για παράδειγμα, τυπική απόκλιση* $\sigma = 0.25$ *τότε η συνάρτηση πυκνότητας της έχει μέγιστη τιμή ίση με*

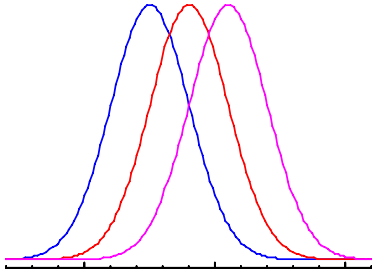
$$f(\mu) = \frac{1}{\sigma\sqrt{2\pi}} = \frac{0.399}{\sigma} = \frac{0.399}{0.25} = 1.596\ .$$

*Είναι άραγε λογικό αυτό; Δηλαδή, μπορεί αυτή η τιμή να είναι μεγαλύτερη του 1;*

Είναι φανερό, ότι η *συνάρτηση πυκνότητας* της *κανονικής κατανομής* δεν ορίζει μια συγκεκριμένη *κανονική καμπύλη* αλλά μια ***οικογένεια κανονικών καμπύλων***. Έτσι, για διαφορετικές τιμές των παραμέτρων $\mu$ και $\sigma$ παίρνουμε διαφορετικές κανονικές καμπύλες. Για παράδειγμα, οι *κατανομές* στα *Σχήματα 7.1.4* είναι όλες *κανονικές κατανομές*. Στο *7.1.4α* έχουν ίδια *μέση τιμή* και διαφορετικές *τυπικές αποκλίσεις* ενώ στο *7.1.4β* έχουν ίδιες *τυπικές αποκλίσεις* και διαφορετικές *μέσες τιμές*.



*(α)*
*Ίδια μέση τιμή και διαφορετική τυπική απόκλιση*



*(β)*
*Ίδια τυπική απόκλιση και διαφορετική μέση τιμή*

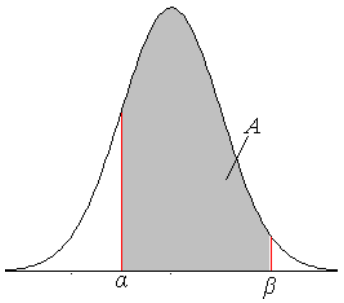
*Σχήμα 7.1.4*
*Κανονικές καμπύλες*

Είναι φανερό, ότι αλλαγή της *μέσης τιμής* προκαλεί μόνο μετατόπιση της *κανονικής καμπύλης* σε μια νέα θέση. Όμως αλλαγή της *τυπικής απόκλισης*, προκαλεί αλλαγή στην *κανονική καμπύλη (χωρίς, φυσικά να αλλάζει η κωδωνοειδής μορφή της)*. **Όσο μικρότερη είναι η *τυπική απόκλιση***, τόσο ψηλότερη και τόσο πιο στενή είναι η *κανονική καμπύλη*, δηλαδή, **τόσο μικρότερο είναι το διάστημα στο οποίο, πρακτικά, εκτείνεται η κατανομή**.

Επισημαίνουμε ότι οι παράμετροι $\mu$ και $\sigma$ χαρακτηρίζουν την *κανονική κατανομή*, δηλαδή, μπορούμε να την προσδιορίσουμε πλήρως αν γνωρίζουμε μόνο τη *μέση τιμή της* $\mu$ και την *τυπική απόκλισή της* $\sigma$. Η *κανονική κατανομή* με *μέση τιμή* $\mu$ και *τυπική απόκλιση* $\sigma$ (δηλαδή *διακύμανση* $\sigma^2$) συμβολίζεται με $N(\mu,\sigma^2)$.

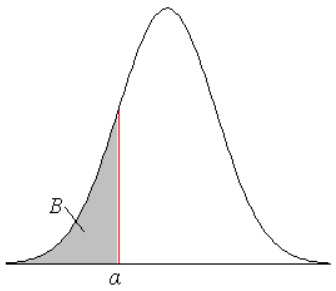
Το εμβαδόν του χωρίου που περικλείεται από την καμπύλη της *συνάρτησης πυκνότητας* και τον άξονα των τιμών μιας συνεχούς τυχαίας μεταβλητής $X$ είναι, όπως γνωρίζουμε. ίσο με 1 και εκφράζει την πιθανότητα η $X$ να πάρει κάποια τιμή μεταξύ $-\infty$ και $+\infty$. Ανάλογα,

- το εμβαδόν του σκιαγραφημένου χωρίου Α στο *Σχήμα 7.1.5*, εκφράζει την πιθανότητα η $X$ να πάρει κάποια τιμή μεταξύ των τιμών $\alpha$ και $\beta$, δηλαδή, $A = P(\alpha \le X \le \beta)$.



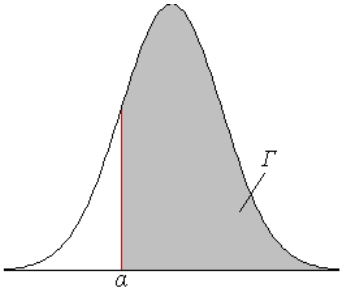

*Σχήμα 7.1.5:* $P(\alpha \le X \le \beta)$

- το εμβαδόν του σκιαγραφημένου χωρίου Β στο *Σχήμα 7.1.6*, εκφράζει την πιθανότητα η $X$ να πάρει κάποια τιμή μικρότερη ή ίση του $\alpha$, δηλαδή, $B = P(X \le \alpha)$.




*Σχήμα 7.1.6:* $P(X \le \alpha)$

- το εμβαδόν του σκιαγραφημένου χωρίου Γ στο *Σχήμα 7.1.7*, εκφράζει την πιθανότητα η $X$ να πάρει κάποια τιμή μεγαλύτερη ή ίση του $\alpha$, δηλαδή, $\Gamma = P(X \ge a)$.



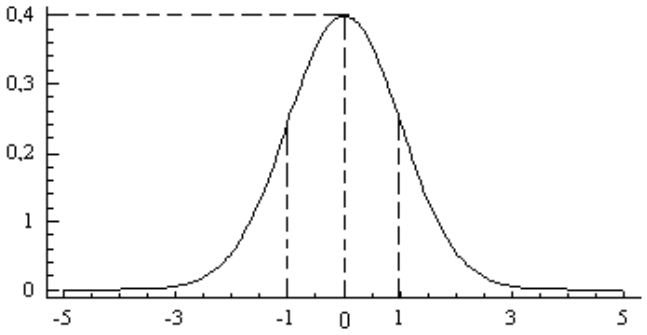

*Σχήμα 7.1.7:* $P(X \ge a)$

### *7.1.2 Η τυποποιημένη κανονική κατανομή*

Στο *5ο Κεφάλαιο* είδαμε ότι μια τυχαία μεταβλητή με μέση τιμή 0 και τυπική απόκλιση 1 ονομάζεται *τυποποιημένη* τυχαία μεταβλητή. Έτσι, η *κανονική κατανομή* που έχει *μέση τιμή* 0 και *τυπική απόκλιση* 1 (άρα και *διακύμανση* 1) ονομάζεται ***τυποποιημένη (ή τυπική) κανονική κατανομή (standard normal distribution)*** και συμβολίζεται με $N(0,1)$. Η τυχαία μεταβλητή που ακολουθεί την *τυποποιημένη κανονική κατανομή,* έχει επικρατήσει να συμβολίζεται με *Z,* η συνάρτηση πυκνότητάς της με $\varphi(z)$ και η συνάρτηση κατανομής της με $\Phi(z)$. Έτσι,

$$\varphi(z) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{z^2}{2}}, \quad -\infty < z < +\infty$$

και η γραφική της παράσταση *(Σχήμα 7.1.8)* παρουσιάζει, σύμφωνα με τις ιδιότητες της κανονικής καμπύλης, *μέγιστη τιμή* (ίση με $1/\sqrt{2\pi} = 0.399$) στη θέση $z = 0$ και *σημεία καμπής* στις θέσεις $z = -1$ και $z = 1$.

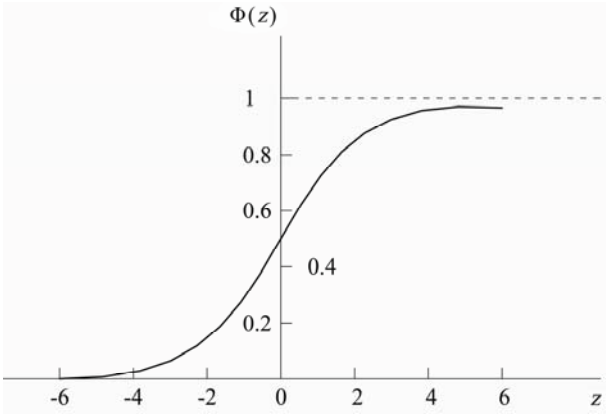



*Σχήμα 7.1.8*
*Η συνάρτηση πυκνότητας* $\varphi(z)$ *της* $Z \sim N(0,1)$

Στο *Σχήμα 7.1.9* φαίνεται η γραφική παράσταση της *συνάρτησης κατανομής*,

$$\Phi(z) = P(Z \le z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}}\, dt, \quad -\infty < z < +\infty,$$

της $Z$. Παρατηρείστε ότι έχει σιγμοειδή μορφή.




*Σχήμα 7.1.9*
*Η συνάρτηση κατανομής* $\Phi(z)$ *της* $Z \sim N(0,1)$

### *7.1.3 Υπολογισμός πιθανοτήτων*

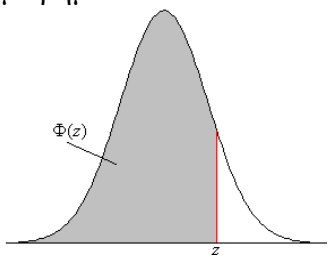
Ο υπολογισμός πιθανοτήτων συνεχούς τυχαίας μεταβλητής ανάγεται, όπως γνωρίζουμε, στον υπολογισμό εμβαδών επίπεδων χωρίων κάτω από τη γραφική παράσταση της συνάρτησης πυκνότητάς της και επομένως στον υπολογισμό κατάλληλων ολοκληρωμάτων. Δυστυχώς, καμία από τις γνωστές τεχνικές ολοκλήρωσης δε μας επιτρέπει τον αναλυτικό υπολογισμό του κατάλληλου, κατά περίπτωση, ορισμένου ολοκληρώματος της συνάρτησης πυκνότητας της κανονικής

κατανομής. Στην πράξη, για να υπολογίσουμε πιθανότητες μιας τυχαίας μεταβλητής που ακολουθεί μια *κανονική κατανομή* $N(\mu,\sigma^2)$, χρησιμοποιούμε τον ***πίνακα της τυποποιημένης κανονικής κατανομής*** $N(0,1)$.

Ο *πίνακας της τυποποιημένης κανονικής κατανομής (Παράρτημα-Α1)* μας δίνει τις τιμές

$$\Phi(z) = P(Z \le z)$$

της *συνάρτησης κατανομής* της *τυποποιημένης κανονικής κατανομής,* δηλαδή, το εμβαδόν του σκιαγραφημένου χωρίου που φαίνεται στο *Σχήμα 7.1.10,* για κάθε *z* από 0 έως 3.59 (ή έως 3.49 ή 3.09) με βήμα 0.01.



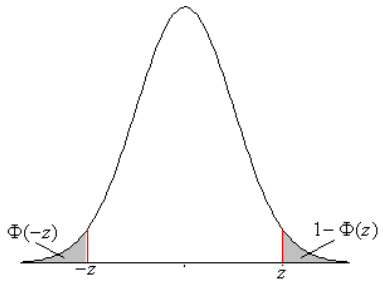

*Σχήμα 7.1.10*
$\Phi(z) = P(Z \le z)$

Από τη συμμετρία της *κανονικής καμπύλης*, εύκολα προκύπτει *(Σχήμα 7.1.11)* ότι

$$\Phi(-z) = 1 - \Phi(z)$$

δηλαδή

$$P(Z \le -z) = \Phi(-z) = 1 - \Phi(z).$$




*Σχήμα 7.1.11*
$\Phi(-z) = 1 - \Phi(z)$

Η ιδιότητα αυτή εξηγεί γιατί ο πίνακας της τυποποιημένης κανονικής κατανομής δίνει τιμές της $\Phi(z)$ μόνο για μη αρνητικά *z*.

Εύκολα επίσης, προκύπτει ότι:

- $P(\alpha \le Z \le \beta) = \Phi(\beta) - \Phi(\alpha)$
- $P(-\alpha \le Z \le \alpha) = \Phi(\alpha) - \Phi(-\alpha) = 2\Phi(\alpha) - 1$
- $P(Z > a) = 1 - P(Z \le \alpha) = 1 - \Phi(\alpha)$.

Είναι φανερό, ότι μπορούμε πλέον να υπολογίσουμε οποιαδήποτε πιθανότητα για τη *Z* με βάση μόνο τις τιμές $\Phi(z)$ του πίνακα *της τυποποιημένης κανονικής κατανομής.* Ας δούμε μερικά παραδείγματα.

$$P(Z \le 0) = \Phi(0) = 0.5$$
$$P(Z \le 1.37) = \Phi(1.37) = 0.9147$$
$$P(Z > 1.37) = 1 - P(Z \le 1.37) = 1 - \Phi(1.37) = 1 - 0.9147 = 0.0853$$
$$P(Z \le -1.55) = \Phi(-1.55) = 1 - \Phi(1.55) = 1 - 0.9394 = 0.0606$$
$$P(-1.55 \le Z \le 2.1) = \Phi(2.1) - \Phi(-1.55) = \Phi(2.1) - [1 - \Phi(1.55)] =$$
$$= \Phi(2.1) - 1 + \Phi(1.55) = 0.9821 - 1 + 0.9394 = 0.9215$$

$$P(-1 \le Z \le 1) = 2\Phi(1) - 1 = 2 \cdot 0.8413 - 1 = 0.6826 \cong 68.3\%$$
$$P(-2 \le Z \le 2) = 2\Phi(2) - 1 = 2 \cdot 0.9772 - 1 = 0.9544 \cong 95.5\%$$
$$P(-3 \le Z \le 3) = 2\Phi(3) - 1 = 2 \cdot 0.9987 - 1 = 0.9974 \cong 99.7\%$$

***Ερώτηση:*** *Μπορείτε να εξηγήσετε γιατί ο πίνακας της τυποποιημένης κανονικής κατανομής δίνει τιμές της* $\Phi(z)$ *μόνο μέχρι* $z = 3.59$ *(ή μόνο μέχρι 3.49 ή 3.09);*

Ας δούμε τώρα πώς μπορούμε με βάση τον πίνακα *της τυποποιημένης κανονικής κατανομής,* να υπολογίζουμε πιθανότητες για οποιαδήποτε *κανονική τυχαία μεταβλητή* $N(\mu, \sigma^2)$.

***Πρόταση 7.1.1:*** *Αν η τυχαία μεταβλητή* $X$ *ακολουθεί μια κανονική κατανομή* $N(\mu, \sigma^2)$ *τότε η τυχαία μεταβλητή*

$$Z = \frac{X - \mu}{\sigma}$$

*ακολουθεί την τυποποιημένη κανονική* $N(0,1)$.

■

Αξιοποιώντας αυτό το αποτέλεσμα, μπορούμε πλέον να υπολογίσουμε πιθανότητες για οποιαδήποτε *κανονική* τ.μ. Ας δούμε για παράδειγμα, πώς για οποιαδήποτε κανονική τυχαία μεταβλητή $X \sim N(\mu, \sigma^2)$ μπορούμε να υπολογίσουμε την πιθανότητα $P(\alpha \le X \le \beta)$.

Προφανώς

$$P(\alpha \le X \le \beta) = P(\frac{\alpha - \mu}{\sigma} \le \frac{X - \mu}{\sigma} \le \frac{\beta - \mu}{\sigma})$$

και επειδή

$$\frac{X - \mu}{\sigma} = Z \sim N(0,1)$$

έχουμε

$$P(\alpha \le X \le \beta) = P(\frac{\alpha - \mu}{\sigma} \le \frac{X - \mu}{\sigma} \le \frac{\beta - \mu}{\sigma}) = P(\frac{\alpha - \mu}{\sigma} \le Z \le \frac{\beta - \mu}{\sigma}) =$$
$$= \Phi(\frac{\beta - \mu}{\sigma}) - \Phi(\frac{\alpha - \mu}{\sigma}).$$

Ασφαλώς, μπορούμε να υπολογίσουμε μέσω της *Z* και οποιαδήποτε άλλη πιθανότητα για την *X*.

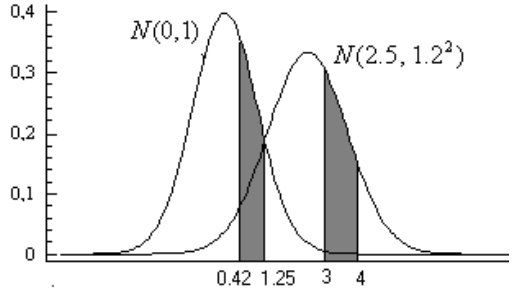
Για παράδειγμα,

$$P(X \le \beta) = P(\frac{X - \mu}{\sigma} \le \frac{\beta - \mu}{\sigma}) = P(Z \le \frac{\beta - \mu}{\sigma}) = \Phi(\frac{\beta - \mu}{\sigma}).$$

Έτσι, αν $X \sim N(2.5, 1.2^2)$, τότε για την πιθανότητα $P(3 \le X \le 4)$ έχουμε

$$P(3 \le X \le 4) = P(\frac{3 - 2.5}{1.2} \le \frac{X - 2.5}{1.2} \le \frac{4 - 2.5}{1.2}) = P(0.42 \le Z \le 1.25) =$$
$$= \Phi(1.25) - \Phi(0.42) = 0.8944 - 0.6628 = 0.2316.$$

Δείτε στο *Σχήμα 7.1.12* το χωρίο κάτω από την $N(2.5, 1.2^2)$ που αντιστοιχεί στην πιθανότητα $P(3 \le X \le 4)$ και το χωρίο κάτω από την $N(0,1)$ που αντιστοιχεί στην πιθανότητα $P(0.42 \le Z \le 1.25)$. Τα δύο χωρία είναι ισεμβαδικά. Αυτό εξάλλου μας διαβεβαιώνει η *Πρόταση 7.1.1*.




*Σχήμα 7.1.12*
*Τα σκιαγραφημένα χωρία έχουν ίδιο εμβαδό*

***Παράδειγμα 7.1.1:*** *Έχει παρατηρηθεί ότι ο χρόνος, έστω X, που χρειάζεται ένα ασθενοφόρο για να φθάσει από ένα κέντρο υγείας στο πλησιέστερο περιφερειακό νοσοκομείο, ακολουθεί κατά προσέγγιση κανονική κατανομή με μέση τιμή* $\mu = 17\,min$ *και τυπική απόκλιση* $\sigma = 3\,min$. *Να βρεθεί η πιθανότητα ο χρόνος που θα χρειασθεί το ασθενοφόρο για να φθάσει στο περιφερειακό νοσοκομείο να είναι α) το πολύ 15 min β) περισσότερο από 22 min και γ) τουλάχιστον 13 min και το πολύ 21 min.*

<u>*Απάντηση:*</u> α) $P(X \le 15) = P(\frac{X-17}{3} \le \frac{15-17}{3}) = P(Z \le -0.67) = \Phi(-0.67) =$

$= 1 - \Phi(0.67) = 1 - 0.7486 = 0.25.$

β) $P(X > 22) = P(\frac{X-17}{3} > \frac{22-17}{3}) = P(Z > 1.67) = 1 - P(Z \le 1.67) =$

$= 1 - \Phi(1.67) = 1 - 0.9525 = 0.0475.$

γ) $P(13 \le X \le 21) = P(\frac{13-17}{3} \le \frac{X-17}{3} \le \frac{21-17}{3}) = P(-1.33 \le Z \le 1.33) =$

$= 2\Phi(1.33) - 1 = 2 \cdot 0.9082 - 1 = 0.8164.$

■

***Παράδειγμα 7.1.2:*** *Στην Περιγραφική Στατιστική, όπως θα δούμε στη συνέχεια, χρησιμοποιείται ένας κανόνας, γνωστός ως* ***εμπειρικός κανόνας (empirical rule)*** *γιατί πολύ συχνά επαληθεύεται εμπειρικά σε διάφορα πειράματα και φαινόμενα, σύμφωνα με τον οποίο, αν η κατανομή ενός δείγματος τιμών μιας τυχαίας μεταβλητής προσομοιάζει με μια κανονική κατανομή (έχει κωδωνοειδή μορφή), τότε το ποσοστό των τιμών του δείγματος που απέχουν από τον μέσο τους (α) λιγότερο από μια τυπική απόκλιση είναι περίπου 68% (β) λιγότερο από δύο τυπικές αποκλίσεις είναι περίπου 95% και (γ) λιγότερο από τρεις τυπικές αποκλίσεις είναι περίπου 99%. Ας αποδείξουμε αυτό τον κανόνα για μια τυχαία μεταβλητή, έστω X, που ακολουθεί μια κανονική κατανομή* $N(\mu, \sigma^2)$.

<u>*Απάντηση:*</u> Αν $X \sim N(\mu, \sigma^2)$, θα δείξουμε ότι

$$P(\mu - k\sigma \le X \le \mu + k\sigma) = 2\Phi(k) - 1.$$

Πράγματι

$$P(\mu - k\sigma \le X \le \mu + k\sigma) = P\big(-k\sigma \le X - \mu \le +k\sigma\big) = P(-k \le \frac{X-\mu}{\sigma} \le +k) =$$

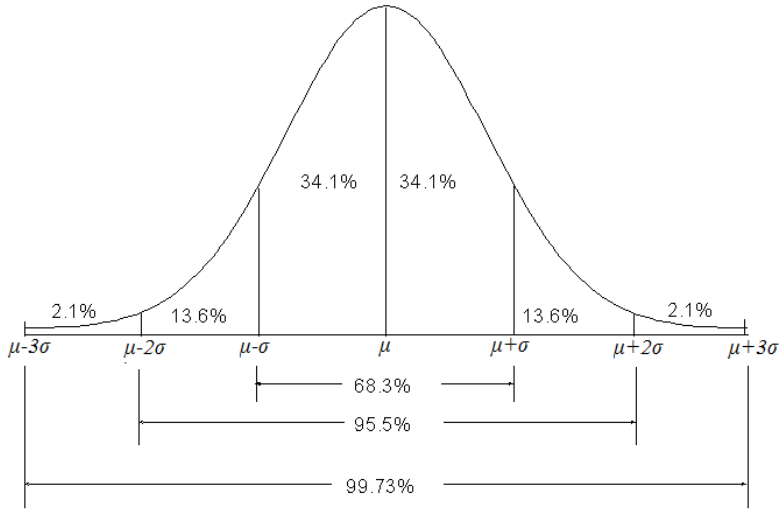
$$= P(-k \le Z \le +k) = 2\Phi(k) - 1.$$

Έτσι *(δες και Σχήμα 7.1.13)*, για $k = 1, 2, 3$ έχουμε

$P(\mu - \sigma \le X \le \mu + \sigma) = 2\Phi(1) - 1 = 0.6826 \cong 68.3\%$

$P(\mu - 2\sigma \le X \le \mu + 2\sigma) = 2\Phi(2) - 1 = 0.9544 \cong 95.5\%$

$P(\mu - 3\sigma \le X \le \mu + 3\sigma) = 2\Phi(3) - 1 = 0.9974 \cong 99.7\%$




*Σχήμα 7.1.13*
*Επιβεβαίωση του εμπειρικού κανόνα*

■

***Σχόλιο 7.1.1:*** *Ίσως σας έχει δημιουργηθεί το εξής ερώτημα: πώς είναι δυνατόν τυχαίες μεταβλητές που παίρνουν μόνο θετικές τιμές ή πεπερασμένου πλήθους τιμές, όπως μεταβλητές που εκφράζουν μήκη, χρόνους ζωής, χρονική διάρκεια φαινομένων κτλ., να περιγράφονται από την κανονική κατανομή η οποία θεωρητικά παίρνει άπειρου πλήθους τιμές και μάλιστα από το* $-\infty$ *μέχρι το* $+\infty$*; Για παράδειγμα, η πιθανότητα* $P(X > \alpha)$ *έχει κάποια τιμή όσο μεγάλο και αν είναι το* $\alpha$*. Αν όμως* $X$ *είναι το ύψος του ανθρώπου και έχει διαπιστωθεί ότι προσεγγίζεται από την κανονική κατανομή, τότε αυτό σημαίνει ότι με βάση το μοντέλο μας (την κανονική κατανομή) θα υπήρχε ένα ποσοστό ανθρώπων, έστω πολύ μικρό, με ύψος X>10 μέτρα! Επίσης, η πιθανότητα* $P(X < 0)$ *έχει κάποια τιμή. Δηλαδή, θα υπήρχε ένα ποσοστό ανθρώπων, έστω πολύ μικρό, με αρνητικό ύψος! Τι μπορεί να συμβαίνει; Μια πρώτη απάντηση είναι η εξής. Οι πιθανότητες αυτές είναι πολύ μικρές και στην πράξη θεωρούνται μηδέν. Για παράδειγμα, η πιθανότητα να είναι αρνητικός ο χρόνος που θα χρειασθεί το ασθενοφόρο για να φθάσει στο περιφερειακό* νοσοκομείο *(Παράδειγμα 7.1.1) είναι*

$$P(X<0) = P(\frac{X-17}{3} < \frac{0-17}{3}) = P(Z<-5.7) = \Phi(-5.7) = 1-\Phi(5.7)$$

*η οποία πρακτικά είναι μηδέν. Όμως, αυτή η απάντηση/εξήγηση δε φαίνεται ικανοποιητική, αφού μπορεί οι πιθανότητες αυτές πρακτικά να είναι μηδέν, αλλά θεωρητικά δεν είναι μηδέν και επομένως το θεωρητικό μοντέλο φαίνεται «προβληματικό». Η απάντηση είναι η εξής: πρέπει να διακρίνουμε την κανονική κατανομή αυτή καθαυτή, από τα τυχαία φαινόμενα που προσεγγίζονται ικανοποιητικά από την κανονική κατανομή. Η κανονική κατανομή δεν είναι «νόμος της φύσης». Είναι, απλά, ένα μοντέλο το οποίο ορίζεται με μια μαθηματική συνάρτηση. Τίποτε περισσότερο και τίποτε λιγότερο. Η κανονική κατανομή δηλαδή, δεν εκφράζει-περιγράφει απολύτως και εξ ορισμού το τυχαίο φαινόμενο που μας ενδιαφέρει. Το πόσο «καλά» το εκφράζει, δηλαδή το πόσο μας βοηθάει να το κατανοήσουμε, είναι πρόβλημα δικό μας και της Στατιστικής (όπως θα δούμε στο Β΄ Μέρος) και όχι της κανονικής κατανομής!*

■

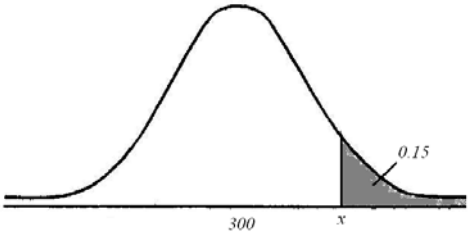
Ας δούμε ένα διαφορετικό παράδειγμα.

***Παράδειγμα 7.1.3:*** *Οι υποψήφιοι για εγγραφή σε ένα Μεταπτυχιακό Τμήμα Πανεπιστημίου, υποβάλλονται σε ένα τεστ. Το τεστ έχει σχεδιασθεί έτσι ώστε οι βαθμοί των υποψηφίων στο τεστ να κατανέμονται κανονικά με μέση τιμή 300 και τυπική απόκλιση 60. α) Αν η πολιτική του Πανεπιστημίου είναι να δέχεται ως φοιτητές το 15% των υποψηφίων με το μεγαλύτερο βαθμό στο τεστ, ποιος είναι ο μικρότερος βαθμός που επιτρέπει την εισαγωγή στο Μεταπτυχιακό Τμήμα; β) Τι βαθμό πρέπει να έχει γράψει*

*ένας υποψήφιος στο τεστ για να κατατάσσεται στο 10% των υποψηφίων με το μικρότερο βαθμό στο τεστ;*

<u>*Απάντηση:*</u> Έστω $X$ η βαθμολογία ενός υποψηφίου στο τεστ. Δίνεται ότι $X \sim N(300,\, 60^2)$.

α) Ζητάμε εκείνη την τιμή $x$ της $X$ για την οποία $P(X \geq x) = 0.15$ *(Σχήμα 7.1.14).*




*Σχήμα 7.1.14*
$P(X \geq x) = 0.15$

Έτσι, έχουμε

$$P(X \geq x) = 0.15 \Leftrightarrow P(\frac{X-300}{60} \geq \frac{x-300}{60}) = 0.15 \Leftrightarrow P(Z \geq \frac{x-300}{60}) = 0.15 \Leftrightarrow$$

$$1 - P(Z < \frac{x-300}{60}) = 0.15 \Leftrightarrow P(Z < \frac{x-300}{60}) = 1 - 0.15 \Leftrightarrow P(Z < \frac{x-300}{60}) = 0.85 \Leftrightarrow$$

$$\Phi(\frac{x-300}{60}) = 0.85\,.$$

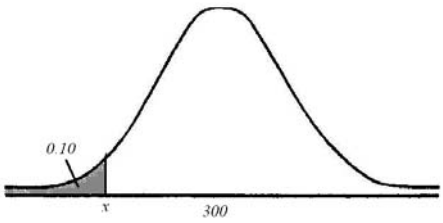
Κάνοντας «αντίστροφη αναζήτηση» στον πίνακα της *τυποποιημένης κανονικής κατανομής*, παρατηρούμε ότι η πιθανότητα 0.85 δεν υπάρχει στον πίνακα όμως οι πλησιέστερες σε αυτήν (που υπάρχουν στον πίνακα) είναι οι 0.8485 και 0.8508 οι οποίες αντιστοιχούν στις τιμές $z = 1.03$ και $z = 1.04$ οπότε ή επιλέγουμε μια από αυτές τις δύο τιμές ή κάνουμε παρεμβολή και βρίσκουμε $z = (1.03 + 1.04)/2 = 1.035$.

Έτσι, έχουμε

$$\frac{x-300}{60} = 1.035 \Rightarrow x = 362.1\,.$$

Άρα, η ζητούμενη βαθμολογία είναι 362.1. Δηλαδή, για να ανήκει ένας υποψήφιος στο 15% των υποψηφίων με το μεγαλύτερο βαθμό στο τεστ, πρέπει να πάρει βαθμό τουλάχιστον ίσο με 362.1.

β) Έστω $x$ εκείνη η τιμή της $X$ για την οποία ισχύει $P(X \leq x) = 0.10$ *(Σχήμα 7.1.15).*




*Σχήμα 7.1.15*
$P(X \leq x) = 0.10$

Έχουμε

$$P(X \leq x) = 0.10 \Leftrightarrow P(\frac{X-300}{60} \leq \frac{x-300}{60}) = 0.10 \Leftrightarrow P(Z \leq \frac{x-300}{60}) = 0.10 \Leftrightarrow$$

$$\Phi(\frac{x-300}{60}) = 0.10 \Leftrightarrow 1 - \Phi(-\frac{x-300}{60}) = 0.10 \Leftrightarrow \Phi(\frac{300-x}{60}) = 0.90\,.$$

Έτσι, με «αντίστροφη αναζήτηση» στον πίνακα της *τυποποιημένης κανονικής κατανομής*, βρίσκουμε $z = (1.28 + 1.29)/2 = 1.285$ και επομένως

$$\frac{300-x}{60} = 1.285 \Rightarrow x = 222.9\,.$$

Άρα, ένας υποψήφιος κατατάσσεται στο 10% των υποψηφίων με το μικρότερο βαθμό στο τεστ, αν έχει πάρει βαθμό το πολύ ίσο με 222.9.

■

***Σημείωση 7.1.1 (άνω α-ποσοστιαίο σημείο):*** *Είναι προφανές ότι με την προηγούμενη μέθοδο υπολογίζουμε ποσοστημόρια της κανονικής κατανομής. Η τιμή* $z$ *της* $Z \sim N(0,1)$ *για την οποία ισχύει*

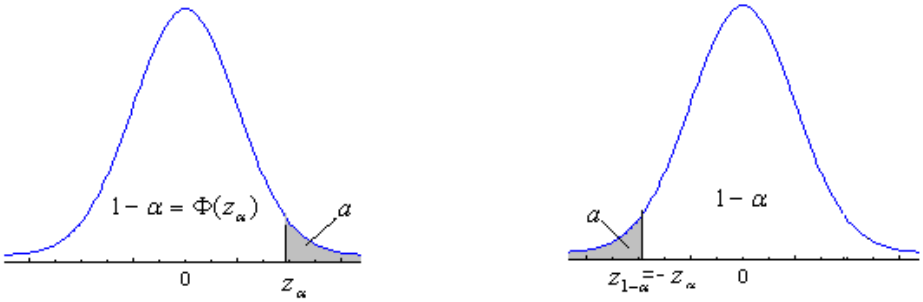
$$P(Z > z) = \alpha, \quad 0 < \alpha < 1$$

*ονομάζεται* ***άνω*** $\alpha$ ***-ποσοστιαίο σημείο*** *της τυποποιημένης κανονικής κατανομής και συμβολίζεται με* $z_\alpha$. *Δηλαδή,*

$$P(Z > z_\alpha) = \alpha.$$

*Προφανώς, λόγω συμμετρίας της κατανομής (δες Σχήμα 7.1.16)*

$$z_{1-\alpha} = -z_\alpha.$$




*Σχήμα 7.1.16*

$P(Z > z_\alpha) = \alpha$ *και* $z_{1-\alpha} = -z_\alpha$

■

***Άσκηση 7.1.1:*** *Δείξτε ότι α)* $z_{0.01} = 2.33$ *και β)* $z_{0.99} = -2.33$.

<u>*Απάντηση:*</u> α) Από τον ορισμό του $z_\alpha$, για $\alpha = 0.01$, έχουμε

$$P(Z > z_{0.01}) = 0.01 \Leftrightarrow 1 - P(Z \le z_{0.01}) = 0.01 \Leftrightarrow 1 - \Phi(z_{0.01}) = 0.01 \Leftrightarrow \Phi(z_{0.01}) = 0.99$$

και επομένως με «αντίστροφη αναζήτηση» στον πίνακα της *τυποποιημένης κανονικής κατανομής* παίρνουμε

$$z_{0.01} = 2.33.$$

β) Από τον ορισμό του *άνω* $\alpha$ *-ποσοστιαίου σημείου* $z_\alpha$, για $\alpha = 0.99$, έχουμε

$$P(Z > z_{0.99}) = 0.99 \Leftrightarrow 1 - P(Z \le z_{0.99}) = 0.99 \Leftrightarrow 1 - \Phi(z_{0.99}) = 0.99 \Leftrightarrow \Phi(-z_{0.99}) = 0.99$$

και επομένως με «αντίστροφη αναζήτηση» στον πίνακα της *τυποποιημένης κανονικής κατανομής* παίρνουμε

$$-z_{0.99} = 2.33 \Leftrightarrow z_{0.99} = -2.33.$$

*(Ασφαλώς, μπορούσαμε να χρησιμοποιήσουμε τη σχέση* $z_{1-\alpha} = -z_\alpha$ *και άμεσα να πάρουμε* $z_{0.99} = z_{1-0.01} = -z_{0.01} = -2.33$*).*

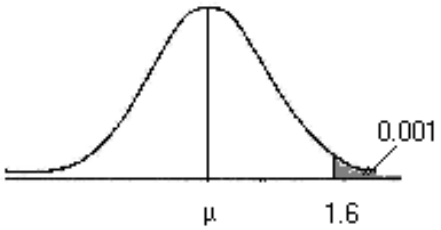
■

***Παράδειγμα 7.1.4:*** *Μια αυτόματη μηχανή συσκευασίας τροφίμων έχει προγραμματισθεί να συσκευάζει δημητριακά σε συσκευασίες των 1.5kg. Έχει παρατηρηθεί ότι η ποσότητα δημητριακών ανά συσκευασία ακολουθεί μια κανονική κατανομή με μέση τιμή* $\mu = 1.5\,kg$ *και τυπική απόκλιση* $\sigma = 0.1\,kg$*. α) Τι ποσοστό των συσκευασιών περιέχει ποσότητα που υπερβαίνει τα 1.6kg; β) Σε τι ποσότητα πρέπει να ρυθμισθεί η μηχανή έτσι ώστε μόνο στο 0.001 των περιπτώσεων η ποσότητα δημητριακών στη συσκευασία να υπερβαίνει τα 1.6kg;*

<u>*Απάντηση:*</u> Έστω $X$ η ποσότητα που περιέχεται ανά συσκευασία.

α) Γνωρίζουμε ότι $X \sim N(1.5,\, 0.1^2)$ και επομένως εύκολα υπολογίζεται το ποσοστό συσκευασιών που υπερβαίνουν τα 1.6kg (για εξάσκηση, επαληθεύστε ότι $P(X > 1.6) = 0.1587$).

β) Έστω $X \sim N(\mu, 0.1^2)$. Πρέπει να προσδιορισθεί η μέση τιμή $\mu$ ώστε $P(X > 1.6) = 0.001$ *(δες και Σχήμα 7.1.17)*..




*Σχήμα 7.1.17*
$P(X > 1.6) = 0.001$

Έχουμε

$$1 - P(X \le 1.6) = 0.001 \Leftrightarrow P(X \le 1.6) = 0.999 \Leftrightarrow P(\frac{X-\mu}{0.1} \le \frac{1.6-\mu}{0.1}) = 0.999 \Leftrightarrow$$

$$P(Z \le \frac{1.6-\mu}{0.1}) = 0.999 \Leftrightarrow \Phi(\frac{1.6-\mu}{0.1}) = 0.999.$$

Άρα

$$\frac{1.6-\mu}{0.1} = 3.09 \Rightarrow \mu = 1.29$$

δηλαδή, η μηχανή πρέπει να ρυθμισθεί στα 1.29kg.

■

Συχνά, σε πρακτικά προβλήματα, ενδιαφέρουν πιθανότητες κάποιας τυχαίας μεταβλητής η οποία εκφράζει το άθροισμα άλλων ανεξάρτητων τυχαίων μεταβλητών που η κάθε μια ακολουθεί *κανονική κατανομή*. Ας δούμε ένα τέτοιο πρόβλημα και πώς αντιμετωπίζεται.

***Παράδειγμα 7.1.5:*** *Στα ζώα μιας κτηνοτροφικής μονάδας δίνεται τροφή τρεις φορές την ημέρα. Η ποσότητα θερμίδων που παίρνουν κάθε φορά ακολουθεί κανονική τυχαία μεταβλητή. Το διαιτολόγιο έχει ρυθμισθεί έτσι, ώστε την πρώτη φορά που δίνεται τροφή η μέση ποσότητα θερμίδων που παίρνουν να είναι* $\mu_1 = 500\,cal$ *με τυπική απόκλιση* $\sigma_1 = 50\,cal$*, τη δεύτερη να είναι* $\mu_2 = 1700\,cal$ *με* $\sigma_2 = 200\,cal$ *και την τρίτη να είναι* $\mu_3 = 800\,cal$ *με* $\sigma_3 = 100\,cal$*. Αν οι ποσότητες θερμίδων που παίρνουν τα ζώα τις τρεις φορές είναι ανεξάρτητες μεταξύ τους, ποια είναι η πιθανότητα η συνολική ημερήσια ποσότητα θερμίδων που παίρνει ένα τυχαία επιλεγμένο ζώο της μονάδας να είναι μεταξύ 2975cal και 3025cal.*

<u>*Απάντηση:*</u> Έστω $X_1$, $X_2$, $X_3$ η ποσότητα θερμίδων που παίρνει το ζώο την $1^η$, τη $2^η$ και την $3^η$ φορά αντίστοιχα (ημερησίως). Γνωρίζουμε ότι το διαιτολόγιο έχει ρυθμισθεί έτσι ώστε

$$X_1 \sim N(500, 50^2),\ X_2 \sim N(1700, 200^2) \text{ και } X_3 \sim N(800, 100^2).$$

Η συνολική ημερήσια ποσότητα θερμίδων $S_3$ που παίρνει το ζώο, προφανώς εκφράζεται από το άθροισμα $X_1 + X_2 + X_3$, δηλαδή

$$S_3 = X_1 + X_2 + X_3.$$

Είναι προφανές ότι για να απαντήσουμε στο ερώτημα που τίθεται (και σε άλλα παρόμοια) πρέπει να γνωρίζουμε την κατανομή της $S_3$. Γι' αυτή την κατανομή, μας πληροφορεί η ακόλουθη πρόταση (δίνεται χωρίς απόδειξη).

***Πρόταση 7.1.2:*** *Αν* $X_1, X_2, \ldots, X_\nu$ *ανεξάρτητες τυχαίες μεταβλητές με* $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2, \ldots, \nu$, *τότε*

$$S_\nu = \sum_{i=1}^{\nu} X_i \sim N(\mu_1 + \mu_2 + \ldots + \mu_\nu,\ \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_\nu^2).$$

*Αν* $X_i \sim N(\mu, \sigma^2)$, $i = 1, 2, \ldots, \nu$, *τότε*

$$S_\nu = \sum_{i=1}^{\nu} X_i \sim N(\nu\mu,\ \nu\sigma^2).$$

*Γενικότερα, αν* $X_1, X_2, \ldots, X_\nu$ *ανεξάρτητες τυχαίες μεταβλητές με* $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2, \ldots, \nu$ *και* $\alpha_1, \alpha_2, \ldots, \alpha_\nu$, $\beta$ *πραγματικοί αριθμοί, τότε*

$$\sum_{i=1}^{\nu} \alpha_i X_i + \beta \sim N(\alpha_1\mu_1 + \alpha_2\mu_2 + \ldots + \alpha_\nu\mu_\nu + \beta,\ \alpha_1^2\sigma_1^2 + \alpha_2^2\sigma_2^2 + \ldots + \alpha_\nu^2\sigma_\nu^2).$$

■

Επειδή οι $X_1$, $X_2$, $X_3$ είναι ανεξάρτητες, από την προηγούμενη πρόταση έχουμε ότι $S_3 \sim N(500 + 1700 + 800,\ 50^2 + 200^2 + 100^2)$ ή $S_3 \sim N(3000,\ 52500)$. Άρα για την ζητούμενη πιθανότητα έχουμε

$$P(2975 < S_3 < 3025) = P\left(\frac{2975 - 3000}{\sqrt{52500}} < \frac{S_3 - 3000}{\sqrt{52500}} < \frac{3025 - 3000}{\sqrt{52500}}\right) =$$
$$= P(-0.11 < Z < 0.11) = 2\Phi(0.11) - 1 = 0.733.$$

■

Από την *Πρόταση 7.1.2* εύκολα προκύπτει η ακόλουθη.

***Πρόταση 7.1.3:*** *Αν* $X_1, X_2, \ldots, X_\nu$ *ανεξάρτητες τυχαίες μεταβλητές με* $X_i \sim N(\mu, \sigma^2)$ *για κάθε* $i = 1, 2, \ldots, \nu$, *τότε*

$$\overline{X} = \frac{\sum_{i=1}^{\nu} X_i}{\nu} \sim N\left(\mu,\ \frac{\sigma^2}{\nu}\right).$$

■

***Παράδειγμα 7.1.5 (συνέχεια):*** *Ποια είναι η πιθανότητα, η μέση ποσότητα θερμίδων που παίρνει ημερησίως ένα τυχαία επιλεγμένο ζώο σε ένα χρόνο (365 ημέρες) να είναι μεταξύ 2975cal και 3025cal.*

<u>*Απάντηση:*</u> Έστω $S_i$ η συνολική ποσότητα θερμίδων που παίρνει το ζώο την *i* ημέρα, $i = 1, 2, \ldots, 365$. Επειδή, $S_i \sim N(3000,\ 52500)$ θα έχουμε

$$\overline{S} = \frac{\sum_{i=1}^{365} S_i}{365} \sim N\left(3000,\ \frac{52500}{365}\right)$$

και επομένως,

$$P(2975 < \overline{S} < 3025) = P\left(\frac{2975 - 3000}{\sqrt{52500/365}} < \frac{\overline{S} - 3000}{\sqrt{52500/365}} < \frac{3025 - 3000}{\sqrt{52500/365}}\right) =$$
$$= P(-2.08 < Z < 2.08) = 2\Phi(2.08) - 1 = 0.9624.$$

■

***Παράδειγμα 7.1.6:*** *Οι ακαθάριστες εβδομαδιαίες εισπράξεις μιας κτηνοτροφικής μονάδας από την πώληση του γάλακτος που παράγει είναι κανονική τυχαία μεταβλητή με μέση τιμή 2200 € 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 5000 €;*

*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

$$X_1 \sim N(2.200,\ 230^2) \text{ και } X_2 \sim N(2.200,\ 230^2).$$

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

$$S_2 \sim N(4400,\ 2 \cdot 230^2)$$

και επομένως για τη ζητούμενη πιθανότητα έχουμε

$$P(S_2 > 5000) = P(Z > \frac{5000 - 4000}{\sqrt{2 \cdot 230^2}}) = P(Z > 1.84) = 1 - \Phi(1.84) = 0.0329.$$

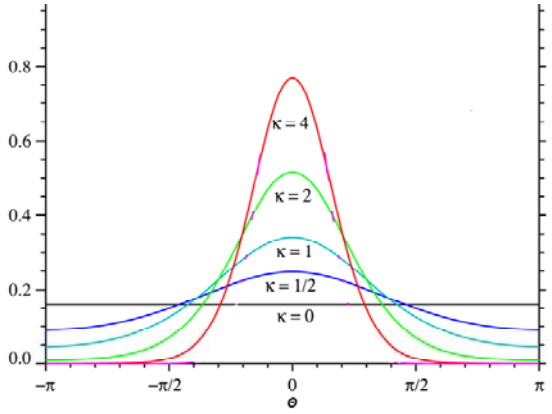
***Σημείωση 7.1.2 (η κατανομή von Mises)***. *Στις κυκλικές μεταβλητές, δηλαδή, στις μεταβλητές που μετρώνται σε κυκλική κλίμακα, η πλέον χρησιμοποιούμενη κατανομή είναι η* ***κατανομή von Mises****. Η κατανομή von Mises, έχει ανάλογα χαρακτηριστικά με την κανονική κατανομή (και αντίστοιχα μεγάλη χρησιμότητα), γι' αυτό στη βιβλιογραφία συναντάται και ως* ***κυκλική κανονική κατανομή (circular normal)****.*

*Αν η κατανομή μιας τυχαίας κυκλικής μεταβλητής, για παράδειγμα, μιας τυχαίας μεταβλητής κατεύθυνσης Θ, περιγράφεται από την κατανομή von Mises, τότε η συνάρτηση πυκνότητας της Θ δίνεται από τον τύπο*

$$f(\vartheta) = \frac{1}{2\pi I_0(k)} e^{k\sigma\upsilon\nu(\vartheta - \mu)}$$

*όπου μ η μέση κατεύθυνση (με τιμές σε διάστημα πλάτους 2π όπως και η Θ), k παράμετρος που παίρνει μη αρνητικές τιμές ($\kappa \geq 0$) και εκφράζει τη συγκέντρωση των τιμών της Θ γύρω από τη μέση κατεύθυνση και*

$$\mathrm{I}_0(k) = \frac{1}{2\pi} \int_0^{2\pi} e^{k\sigma\upsilon\nu\vartheta} d\vartheta.$$




*Σχήμα 7.1.18*
*Η συνάρτηση πυκνότητας της κατανομής von Mises για διάφορες τιμές της παραμέτρου k*

*Για μεγάλα k η κατανομή von Mises προσεγγίζει την* ***κανονική κατανομή*** *με* $\mu = \bar{\theta}$ *και* $\sigma^2 = 1/k$ *(όσο αυξάνεται το k, τόσο αυξάνεται και η πιθανότητα να πάρει η μεταβλητή Θ, τιμή κοντά στη μέση κατεύθυνση).*

*Για μικρά k, δηλαδή όταν το k πλησιάζει στο 0, η κατανομή von Mises προσεγγίζει την* ***ομοιόμορφη κατανομή*** *(σε διάστημα πλάτους 2π), δηλαδή, στην περίπτωση αυτή, για κάθε κατεύθυνση* $\vartheta$*, η πιθανότητα να πάρει η μεταβλητή Θ τιμή κοντά στη* $\vartheta$ *είναι για όλα τα* $\vartheta$ *ίδια.*

## *7.2 Το Κεντρικό Οριακό Θεώρημα*

Στην προηγούμενη ενότητα δώσαμε τη γενική ιδέα για το πώς το ***Κεντρικό Οριακό Θεώρημα (Central Limit Theorem)*** εξηγεί το μεγάλο εύρος εφαρμογής της *κανονικής κατανομής* και πώς συνδέει οποιαδήποτε κατανομή με την *κανονική κατανομή*. Επίσης, στο *5ο Κεφάλαιο (στην ενότητα 5.5)* εξηγήσαμε ότι λέγοντας *«από έναν πληθυσμό παίρνουμε ένα τυχαίο δείγμα μεγέθους* $\nu$ *»* εννοούμε $\nu$ ανεξάρτητες τυχαίες μεταβλητές $X_1, X_2, \ldots, X_\nu$ που ακολουθούν την ίδια κατανομή. Ας δούμε τώρα μια πληρέστερη διατύπωσή του *Κεντρικού Οριακού Θεωρήματος (Κ.Ο.Θ).*

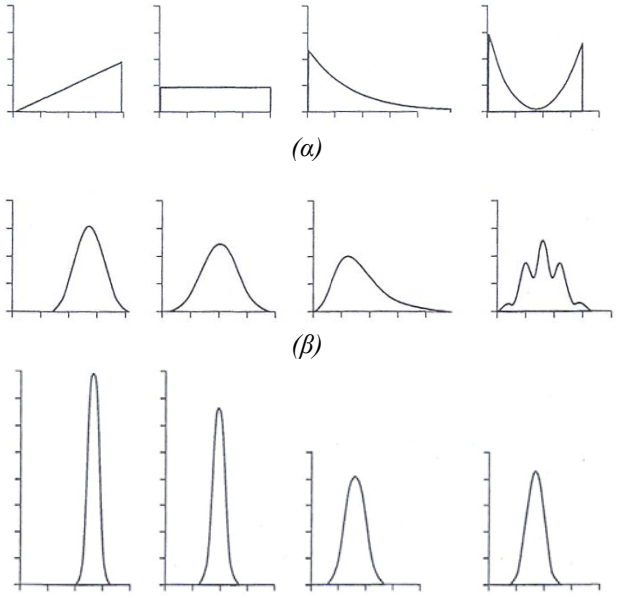
***Θεώρημα 7.2.1 (Το Κεντρικό Οριακό Θεώρημα):*** *Αν* $X_1, X_2, \ldots, X_\nu$ *ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν την ίδια κατανομή με* $E(X_i) = \mu$ *και* $Var(X_i) = \sigma^2$, $i = 1, 2, \ldots, \nu$, *τότε για μεγάλα* $\nu$ *(θεωρητικά* $\nu \to \infty$*), κατά προσέγγιση έχουμε*

$$\bar{X} = \frac{\sum_{i=1}^{\nu} X_i}{\nu} \sim N(\mu, \frac{\sigma^2}{\nu}) \text{ και } S_\nu = \sum_{i=1}^{\nu} X_i \sim N(\nu\mu, \nu\sigma^2).$$

■

Έτσι, αν από έναν πληθυσμό (δηλαδή, από την κατανομή των τιμών μιας τ.μ.) που έχει μέση τιμή $\mu$ και διακύμανση $\sigma^2$, επιλέξουμε τυχαία δείγματα μεγέθους $\nu$ και υπολογίσουμε τους μέσους τους, το *Κ.Ο.Θ.* μας διαβεβαιώνει ότι για μεγάλα $\nu$ (θεωρητικά $\nu \to \infty$), η κατανομή αυτών των μέσων (των δειγματικών) είναι κατά προσέγγιση κανονική με *μέση τιμή* επίσης $\mu$ και *διακύμανση* $\sigma^2/\nu$.

Δείτε, για παράδειγμα[6], στο *Σχήμα 7.2.1α* τέσσερις *πληθυσμούς* (τις κατανομές τεσσάρων τ.μ.) και στο *Σχήμα 7.2.1β* τις αντίστοιχες κατανομές των *δειγματικών μέσων* για $\nu = 4$. Δείτε επίσης στο *Σχήμα 7.2.1γ* τις κατανομές των *δειγματικών μέσων* για $\nu = 25$.



*(α)*

*(β)*

*(γ)*

*Σχήμα 7.2.1*
*Τέσσερις πληθυσμοί (Σχήματα (α)) και οι αντίστοιχες κατανομές των δειγματικών μέσων για* $\nu = 4$ *(Σχήματα (β)) και για* $\nu = 25$ *(Σχήματα (γ))*

[6] *Lapin, L.L., 1982, Statistics for Modern Business Decisions, Harcourt Brace Jovanovich, Inc. New York (από το Davis, J.C., 2002).*

Όσο πιο μεγάλο είναι το μέγεθος $\nu$ των δειγμάτων, τόσο καλύτερη είναι η προσέγγιση της κατανομής των δειγματικών μέσων από την *κανονική κατανομή (δες και την Παρατήρηση 7.2.1 στη συνέχεια).*

Ας δούμε όμως, με συγκεκριμένα παραδείγματα, τι σημαίνουν τα παραπάνω στην πράξη.

***Παράδειγμα 7.2.1:*** *Μας είναι γνωστό ότι τα μήλα στάρκιν που παράγονται στο οροπέδιο της Τεγέας έχουν μέσο βάρος* $\mu = 220gr$ *με τυπική απόκλιση* $\sigma = 80gr$. *Στο συσκευαστήριο του τοπικού συνεταιρισμού τα μήλα συσκευάζονται σε κιβώτια των 60 μήλων και προωθούνται στα ψυγεία και την αγορά. Μπορούμε να υπολογίσουμε ποιο ποσοστό (κατά προσέγγιση) των κιβωτίων περιέχει μήλα με μέσο βάρος μεταξύ 200 και 250gr;*

<u>*Απάντηση:*</u> Τα βάρη των μήλων κάθε κιβωτίου είναι ένα τυχαίο δείγμα μεγέθους $\nu = 60$ από τον πληθυσμό των βαρών των μήλων όλης της παραγωγής. Η μορφή της κατανομής των βαρών των μήλων δε μας είναι γνωστή. Μπορεί να είναι οποιαδήποτε. Για την άγνωστη αυτή κατανομή γνωρίζουμε μόνο τη *μέση τιμή* της $\mu = 220\,gr$ και την *τυπική απόκλισή της* $\sigma = 80gr$. Μπορούμε με αυτά τα δεδομένα να απαντήσουμε στο ερώτημα που θέσαμε; Η απάντηση είναι ναι και ας δούμε πώς.

Το ερώτημά μας μπορεί να επαναδιατυπωθεί ως εξής: *ποιο ποσοστό (κατά προσέγγιση) των δειγματικών μέσων βρίσκεται μεταξύ 200 και 250gr;*.

Είναι φανερό ότι για να μπορέσουμε να απαντήσουμε στο ερώτημα 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. Δηλαδή, αρκεί μόνο 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 στάρκιν που παράγεται στο οροπέδιο της Τεγέας και με $\overline{X}$ την τυχαία μεταβλητή που εκφράζει τους *μέσους των δειγμάτων μεγέθους 60 (από την κατανομή της X)*, το *Κ.Ο.Θ.* μας βεβαιώνει ότι η $\overline{X}$ ακολουθεί κατά προσέγγιση *κανονική κατανομή* με *μέση τιμή* $\mu_{\overline{X}} = \mu = 220gr$ και *διακύμανση*

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{\nu} = \frac{80^2}{60} = 106.67\,gr^2.$$

Αφού γνωρίζουμε ότι

$$\overline{X} \sim N(220, 106.67)$$

η απάντηση πλέον στο ερώτημά μας είναι πολύ απλή. Ζητάμε την πιθανότητα

$$P(200 < \overline{X} < 250)$$

επομένως έχουμε

$$P(200 < \overline{X} < 250) = P\left(\frac{200-220}{\sqrt{106.67}} < Z < \frac{250-220}{\sqrt{106.67}}\right) = P(-1.94 < Z < 2.90) =$$

$$= \Phi(2.90) - \Phi(-1.94) = \Phi(2.90) - [1 - \Phi(1.94)] = \Phi(2.90) - 1 + \Phi(1.94) = 0.9725\,.$$

Δηλαδή, το ποσοστό των κιβωτίων που περιέχουν μήλα με μέσο βάρος μεταξύ 200 και 250gr είναι (κατά προσέγγιση) 97.25%.

■

***Παράδειγμα 7.2.2*****:** *Η ποσότητα ραδιενέργειας που δέχεται κάθε ημέρα ένας ερευνητής είναι τυχαία μεταβλητή με μέση τιμή* $\mu = 0.1$ *μονάδες και τυπική απόκλιση* $\sigma = 0.01$

*μονάδες. Ποια είναι η πιθανότητα το συνολικό ποσό ραδιενέργειας που θα δεχθεί ο ερευνητής σε 100 ημέρες να ξεπερνάει τις 10.02 μονάδες.*

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

$$S_{100} = X_1 + X_2 + \ldots + X_{100}$$

και επειδή το $\nu$ είναι αρκετά μεγάλο, από το *Κ.Ο.Θ.* (κατά προσέγγιση) έχουμε

$$S_{100} \sim N(100 \cdot 0.1, 100 \cdot 0.01^2) \text{ ή } S_{100} \sim N(10, 0.1^2).$$

Άρα η ζητούμενη πιθανότητα είναι

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

■

***Παρατήρηση 7.2.1*:** *Όπως ήδη αναφέραμε, όσο πιο μεγάλο είναι το μέγεθος $\nu$ των δειγμάτων, τόσο καλύτερη (ακριβέστερη) είναι η προσέγγιση της κατανομής των δειγματικών μέσων από την κανονική κατανομή. Πρακτικά, όμως,* ***πόσο μεγάλο πρέπει να είναι το ν****; Απόλυτη απάντηση στο ερώτημα αυτό δεν υπάρχει. Γενικά, το πόσο μεγάλο πρέπει να είναι το $\nu$,* ***εξαρτάται από τον πληθυσμό****. Για παράδειγμα, αν πρόκειται για λοξή (ασύμμετρη) κατανομή απαιτείται μεγαλύτερο μέγεθος δείγματος από αυτό που απαιτείται αν είναι περίπου συμμετρική. Γενικά, το μέγεθος του δείγματος πρέπει να είναι τουλάχιστον 30, δηλαδή, $\nu \geq 30$. Όμως, όπως φαίνεται και στα Σχήματα 7.2.1, υπάρχουν περιπτώσεις όπου καλές προσεγγίσεις παίρνουμε και για μικρότερα $\nu$. Για διερεύνηση των παραπάνω μπορείτε να πειραματιστείτε με λογισμικό προσομοίωσης του Κ.Ο.Θ. που μπορείτε να βρείτε στο Διαδίκτυο. Τέλος, επισημαίνουμε την περίπτωση όπου ο πληθυσμός από τον οποίο γίνεται η δειγματοληψία είναι κανονικός. Στην περίπτωση αυτή, όπως ήδη έχουμε αναφέρει (Πρόταση 7.1.2 και Πρόταση 7.1.3), αποδεικνύεται ότι η κατανομή των δειγματικών μέσων (και της $S_\nu$),* ***είναι*** *κανονική κατανομή ανεξάρτητα από το πόσο μεγάλο είναι το $\nu$.*

■

### *7.2.1 Κανονική προσέγγιση της Διωνυμικής κατανομής*

Ας δούμε ένα ακόμη ενδιαφέρον αποτέλεσμα της *Θεωρίας Πιθανοτήτων*, γνωστό ως *οριακό θεώρημα των De Moivre-Laplace* στο οποίο οφείλεται και η «γέννηση» της *κανονικής κατανομής*.

***Θεώρημα 7.2.2 (Οριακό θεώρημα De Moivre-Laplace):*** *Για μεγάλα* $\nu$ *(θεωρητικά* $\nu \to \infty$*), η διωνυμική κατανομή μπορεί να προσεγγισθεί από μια κανονική κατανομή με ίδια μέση τιμή και ίδια διακύμανση. Δηλαδή, αν* $X \sim B(\nu, p)$ *τότε, για μεγάλες τιμές του* $\nu$*, η κατανομή της* $X$ *προσεγγίζεται από την* $N(\mu, \sigma^2)$ *με* $\mu = \nu p$ *και* $\sigma^2 = \nu p(1-p)$.

■

Το αποτέλεσμα αυτό εύκολα προκύπτει ως ειδική περίπτωση του *Κ.Ο.Θ*[7]. Όμως, δεν προέκυψε έτσι. Αντιθέτως, αποδεικνύοντάς το ο *Abraham De Moivre* το 1733 για $p = 0.5$ και, εκατό περίπου χρόνια αργότερα, το 1812 ο *Pierre-Simon 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 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 De Moivre-Laplace* από τους *Chebysev* και *Markov*). Μάλιστα, το *οριακό θεώρημα 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 παραδείγματα εφαρμογής, σε πρακτικά προβλήματα, της *κανονικής προσέγγισης* της *διωνυμικής κατανομής*, ας δούμε πάλι το ερώτημα: ***πόσο μεγάλο πρέπει να είναι το ν*** για να μπορεί πρακτικά να χρησιμοποιηθεί κανονική προσέγγιση της *διωνυμικής κατανομής*. Το πόσο μεγάλο πρέπει να είναι το $\nu$, ***εξαρτάται από την τιμή της παραμέτρου p***. Αν για παράδειγμα, $p = 0.5$ ή κοντά στο $0.5$, τότε και για όχι πολύ μεγάλες τιμές του $\nu$ παίρνουμε εξαιρετικές προσεγγίσεις της *διωνυμικής*. Αντίθετα, αν το $p$ είναι πολύ μικρό ή πολύ μεγάλο, για καλή προσέγγιση, απαιτούνται πολύ μεγαλύτερες τιμές του $\nu$. Ένας πρακτικός, γενικός, κανόνας είναι ο ακόλουθος.

*Για να πάρουμε μέσω του οριακού θεωρήματος De Moivre-Laplace καλές προσεγγίσεις της διωνυμικής κατανομής αρκεί*

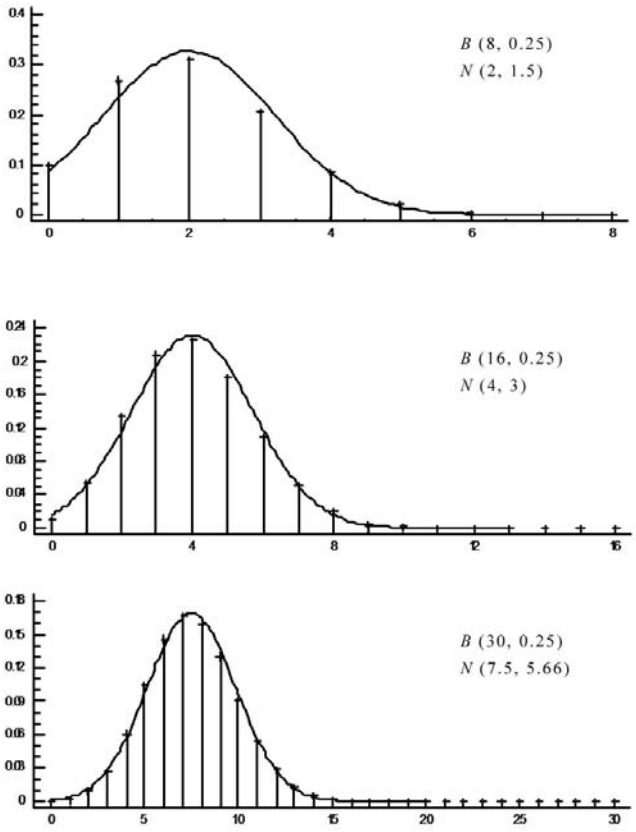
$$\nu p \geq 5 \quad \textbf{\textit{και}} \quad \nu(1-p) \geq 5.$$

*Στη βιβλιογραφία συναντάται επίσης και ο κανόνας ότι αρκεί*

$$\nu p(1-p) \geq 10.$$

Στα *Σχήματα 7.2.2* δίνεται η συνάρτηση πιθανότητας της *διωνυμικής κατανομής* με $p = 0.25$ και $\nu = 8, 16, 30$ καθώς και η συνάρτηση πυκνότητας της *κανονικής κατανομής* με την αντίστοιχη *μέση τιμή* και *διακύμανση*, δηλαδή, με $\mu = \nu p$ και $\sigma^2 = \nu p(1-p)$. Παρατηρείστε ότι όσο μεγαλύτερο γίνεται το $\nu$ τόσο πιο καλή και η προσέγγιση.

[7] *Θυμηθείτε ότι η διωνυμική κατανομή ορίζεται ως άθροισμα ν ανεξάρτητων δίτιμων τυχαίων μεταβλητών.*




*Σχήμα 7.2.2*
*Η διωνυμική κατανομή για* $p = 0.25$ και $\nu = 8, 16, 30$
*και αντίστοιχα η κανονική με* $\mu = \nu p$ *και* $\sigma^2 = \nu pq$

Δείτε επίσης πάλι το *Σχήμα 6.1.2* και παρατηρείστε τη μορφή της *διωνυμικής κατανομής* για $p = 0.5$.

***Παράδειγμα 7.2.4:*** *Ο επιθυμητός/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, κάνει δεκτούς 450 φοιτητές. Ποια είναι η πιθανότητα από τους 450 πρωτοετείς φοιτητές, να παρακολουθούν τελικά τα μαθήματα περισσότεροι από 150.*

<u>*Απάντηση:*</u> Αν $X$ ο αριθμός των πρωτοετών φοιτητών (από τους 450) που παρακολουθούν τα μαθήματα, τότε προφανώς $X \sim B(\nu,\ p)$ με $\nu = 450$ και $p = 0.3$, δηλαδή, $X \sim B(450,\ 3)$. Επειδή, το $\nu$ είναι μεγάλο και $\nu p = 450 \cdot 0.3 = 135 \geq 5$ και $\nu(1-p) = 450 \cdot 0.7 = 315 \geq 5$, η $X$ προσεγγίζεται ικανοποιητικά από την *κανονική* με $\mu = \nu p = 450 \cdot 0.3 = 135$ και $\sigma^2 = \nu p(1-p) = 450 \cdot 0.3 \cdot 0.7 = 94.5$ 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 στο εισαγωγικό κεφάλαιο *(1ο Κεφάλαιο)*, ένα κρίσιμο θέμα στη στατιστική προσέγγιση προβλημάτων είναι το μέγεθος $\nu$ του δείγματος. 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 7.2.5:*** *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 $\nu$. Πόσο πρέπει να είναι το $\nu$ 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%.*

<u>Απάντηση:</u> Ας συμβολίσουμε με $X$ τον αριθμό των ατόμων του δείγματος που έχουν τη συγκεκριμένη ιδιότητα. Η τυχαία μεταβλητή $X$ ακολουθεί τη *διωνυμική κατανομή*[8] με παραμέτρους $\nu$ και $p$, δηλαδή $X \sim B(\nu,\ p)$. Προφανώς, το ποσοστό των ατόμων του δείγματος που έχουν τη συγκεκριμένη ιδιότητα είναι $X/\nu$.

Σύμφωνα με τις απαιτήσεις που θέτει το πρόβλημα πρέπει

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

$$P(-0.01\nu<X-\nu p<0.01\nu)\geq 0.95 \text{ ή } P(\nu p-0.01\nu<X<\nu p+0.01\nu)\geq 0.95.$$

Χρησιμοποιώντας την κανονική προσέγγιση της *διωνυμικής* έχουμε

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

$$=2\Phi\left(\frac{0.01\nu}{\sqrt{\nu p(1-p)}}\right)-1.$$

Άρα, σύμφωνα με τις απαιτήσεις που θέτει το πρόβλημα, πρέπει

$$2\Phi\left(\frac{0.01\nu}{\sqrt{\nu p(1-p)}}\right)-1\geq 0.95$$

ή ισοδύναμα

$$\Phi\left(\frac{0.01\nu}{\sqrt{\nu p(1-p)}}\right)\geq 0.975$$

και επομένως

$$\frac{0.01\nu}{\sqrt{\nu p(1-p)}}\geq 1.96 \text{ ή } \nu\geq 38416\,p(1-p).$$

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 $1/4$ (γιατί;)[9] άρα πρέπει

$$\nu\geq 38416\cdot\frac{1}{4}$$

δηλαδή, το δείγμα πρέπει να έχει μέγεθος τουλάχιστον 9604.

■

***Παράδειγμα 7.2.6:*** *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 $\nu$, να υπολογίσει τη μέση τιμή τους και να τη χρησιμοποιήσει για να εκτιμήσει την άγνωστη πραγματική απόσταση d. Αν οι $\nu$ μετρήσεις, $X_1, X_2, \ldots, X_\nu$, είναι ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν την ίδια (άγνωστη) κατανομή με μέση τιμή d (την άγνωστη πραγματική απόσταση) και*

[8] *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.*

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

*διακύμανση 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>*Απάντηση:*</u> Για μεγάλες τιμές του $\nu$, από το *Κ.Ο.Θ.* έχουμε ότι κατά προσέγγιση

$$\overline{X} = \frac{\sum_{i=1}^{\nu} X_i}{\nu} \sim N(d, \frac{4}{\nu}).$$

Σύμφωνα με τις απαιτήσεις που θέτει το πρόβλημα πρέπει

$$P\left(\left|\overline{X} - d\right| < 0.5\right) = 0.95 \Leftrightarrow P(-0.5 < \overline{X} - d < 0.5) = 0.95 \Leftrightarrow$$

$$\Leftrightarrow P(-0.5 + d < \overline{X} < 0.5 + d) = 0.95 \Leftrightarrow$$

$$P(\frac{-0.5}{2/\sqrt{\nu}} < Z < \frac{0.5}{2/\sqrt{\nu}}) = 0.95 \Leftrightarrow 2\Phi(\sqrt{\nu}/4) - 1 = 0.95 \Leftrightarrow \Phi(\sqrt{\nu}/4) = 0.975$$

και επομένως

$$\frac{\sqrt{\nu}}{4} = 1.96 \Rightarrow \nu = 61.46.$$

Άρα ο αστρονόμος πρέπει να πάρει 62 μετρήσεις.

■

### *7.2.2 Κανονική προσέγγιση της κατανομής Poisson*

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 $\lambda$, τότε η κατανομή της $X$ προσεγγίζεται, για μεγάλες τιμές του $\lambda$ (στην πράξη για $\lambda > 10$), από την $N(\mu, \sigma^2)$ με $\mu = \lambda$ και $\sigma^2 = \lambda$.

Ας δούμε ένα παράδειγμα.

***Παράδειγμα 7.2.7:*** *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 δε θα αναπτυχθούν να είναι τουλάχιστον 120.*

<u>*Απάντηση:*</u> Επειδή η τιμή του $\lambda$ 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

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

■

### *7.2.3 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* $X \sim B(\nu, p)$ προσεγγίζεται από την $N\left(\nu p, \nu p(1-p)\right)$ οι πιθανότητες

$$P(a \leq X \leq b), \; P(X \leq b) \text{ και } P(X \geq a)$$

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 7.2.4* τη διόρθωση συνέχειας έχουμε

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

Ομοίως, χρησιμοποιώντας στο *Παράδειγμα 7.2.7* τη διόρθωση συνέχειας έχουμε

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

Τέλος, σημειώνουμε 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 7.2.1:*** *Το πόσο σημαντικό είναι το γεγονός ότι το Κ.Ο.Θ. μας πληροφορεί για την κατανομή της $\overline{X}$ και της $S_\nu$, θα φανεί και στη συνέχεια όταν αναφερθούμε στα διαστήματα εμπιστοσύνης και τους στατιστικούς ελέγχους υποθέσεων.*

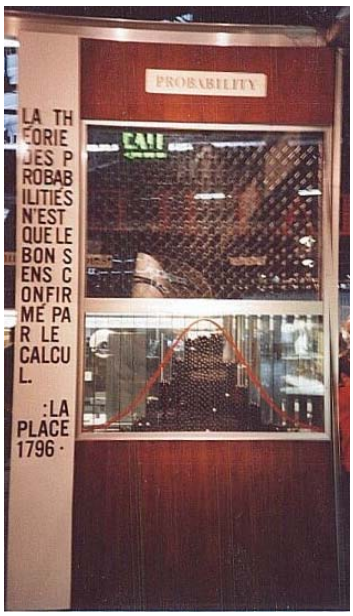
***Για προβληματισμό***

*Σε πολλά μουσεία επιστημών, ένα από τα εκθέματα (μοντέλα/κατασκευές) που εντυπωσιάζουν μικρούς και μεγάλους επισκέπτες και προκαλούν το ενδιαφέρον και την περιέργειά τους, είναι η μηχανή του Galton (Galton's machine)* γνωστή και ως *Galton's board, quincunx* και *bean machine*[10] *(Σχήμα 7.2.3 και Εικόνα 7.2.1). Πρόκειται για έναν κατακόρυφο πίνακα στο πάνω μισό του οποίου υπάρχει μια συστοιχία από καρφιά/πείρους κατανεμημένα σε σειρές και σε ίσες μεταξύ τους αποστάσεις. Το κάτω μισό του πίνακα έχει διαιρεθεί με ορθογώνιες κατακόρυφες υποδοχές τοποθετημένες σε ίσες αποστάσεις επίσης. Στην κορυφή της συστοιχίας καρφιών υπάρχει μια χοάνη από την οποία όταν ο μηχανισμός αρχίσει να λειτουργεί, πέφτουν σφαιρίδια τα οποία προσκρούουν στα καρφιά και κινούμενα στα κενά μεταξύ των καρφιών καταλήγουν στις υποδοχές. Η εμπρός επιφάνεια του πίνακα έχει διαφανές κάλυμμα ώστε η όλη κατασκευή να είναι ορατή (όπως και η λειτουργία της).*

*Ο τρόπος που είναι τοποθετημένα τα καρφιά (αλλά και γενικότερα η όλη κατασκευή) διασφαλίζει ότι κάθε φορά που ένα σφαιρίδιο προσκρούει σε ένα καρφί κινείται με ίδια πιθανότητα, ίση με $1/2$, προς τα δεξιά ή προς τα αριστερά και πέφτει στην επόμενη σειρά καρφιών. Όταν φθάσει στην τελευταία σειρά πέφτει σε μια από τις υποδοχές που υπάρχουν στο κάτω μέρος όπου και στοιβάζεται μαζί με άλλα σφαιρίδια (Σχήμα 7.2.3). Όταν αφεθούν πολλά σφαιρίδια να πέσουν, το μοντέλο που δημιουργείται από τις κατακόρυφες στοίβες σφαιριδίων έχει κωδωνοειδή μορφή που προσομοιάζει με αυτή της κανονικής κατανομής. Δηλαδή, παρότι η διαδρομή κάθε σφαιριδίου είναι τυχαία και*

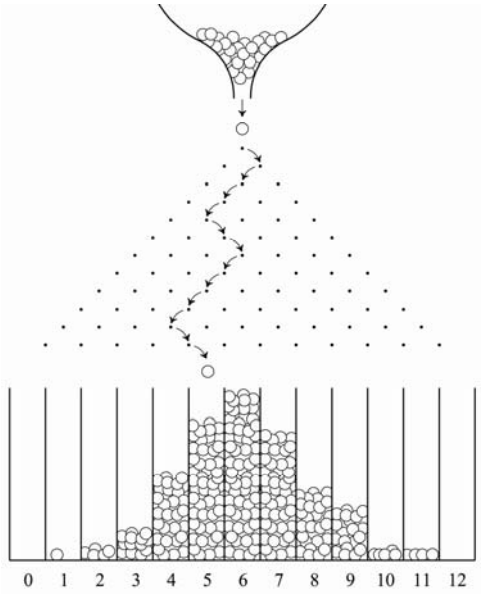
[10] *Ο Sir Francis Galton (1822-1911) ήταν Άγγλος ανθρωπολόγος με πολυσχιδή επιστημονική δραστηριότητα. Συνεργάσθηκε με τον Karl Pearson και χρησιμοποίησε την κανονική κατανομή για τη μελέτη και την ερμηνεία* ανθρωπομετρικών δεδομένων.

*επομένως όχι προβλέψιμη, το μοντέλο που αυτά δημιουργούν είναι πάντοτε (σε κάθε δηλαδή επανάληψη του πειράματος) προβλέψιμο. Μάλιστα όσο περισσότερα είναι τα σφαιρίδια που κάθε φορά αφήνονται να πέσουν, τόσο καλύτερα το μοντέλο που δημιουργείται προσαρμόζεται στην κανονική καμπύλη. Εντυπωσιακό πράγματι αποτέλεσμα.*




*Εικόνα 7.2.1*
*Η μηχανή του Galton στο Μουσείο Επιστημών της Βοστόνης*

*Πρόκειται για ένα ευφυέστατο στη σύλληψη και την υλοποίησή του πείραμα που επιβεβαιώνει και αναδεικνύει με πολύ απλό και παραστατικό τρόπο πώς προκύπτει (τι εκφράζει) η διωνυμική κατανομή και πώς αυτή προσεγγίζεται από την κανονική (οριακό θεώρημα De Moivre-Laplace). Προσομοιώνει επίσης, το νόμο των μεγάλων αριθμών (στατιστική ομαλότητα)! Τι λέτε, μπορείτε να εξηγήσετε πώς τα καταφέρνει όλα αυτά η μηχανή του Galton[11];*

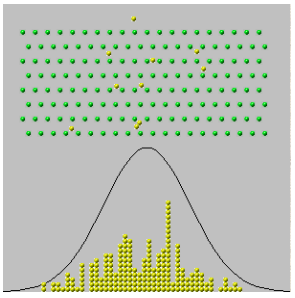



*Σχήμα 7.2.3*
*Η μηχανή του Galton*

[11] *Το ότι προς τις ακραίες υποδοχές στοιβάζονται πολύ λιγότερα σφαιρίδια από ότι στις κεντρικές, είναι βέβαια κάτι προφανώς αναμενόμενο αφού οι διαδρομές που οδηγούν στις ακραίες υποδοχές είναι πολύ λιγότερες.*

*<u>Υπόδειξη-1:</u> Σκεφθείτε ότι η πρόσκρουση ενός σφαιριδίου σε ένα καρφί είναι μια δοκιμή Bernoulli με δύο δυνατά αποτελέσματα: κίνηση προς τα δεξιά, κίνηση προς τα αριστερά. Η διαδρομή επομένως ενός σφαιριδίου ορίζεται από $\nu$ ανεξάρτητες δοκιμές Bernoulli, όσες οι σειρές καρφιών. Σκεφθείτε επίσης ότι ένα σφαιρίδιο θα καταλήξει στην υποδοχή 0 (δες Σχήμα 7.2.3) αν σε κάθε πρόσκρουση κινηθεί προς τα αριστερά, δηλαδή αν σε καμία από τις $\nu$ προσκρούσεις δεν κινηθεί προς τα δεξιά. Αντίστοιχα, θα καταλήξει στην υποδοχή 1 αν κινηθεί προς τα δεξιά σε μία μόνο (οποιαδήποτε) από τις $\nu$ προσκρούσεις, στην υποδοχή 2 αν κινηθεί προς τα δεξιά σε δύο μόνο (οποιεσδήποτε) από τις $\nu$ προσκρούσεις, κ.ο.κ. Έτσι, είναι προφανές ότι η θέση στην οποία καταλήγει ένα σφαιρίδιο καθορίζεται από τον αριθμό των επιτυχιών στις $\nu$ ανεξάρτητες δοκιμές Bernoulli[12]. Πρόκειται επομένως για ένα μοντέλο που περιγράφεται από τη διωνυμική κατανομή $B(\nu,1/2)$ (δείτε επίσης πάλι στο Σχήμα 6.1.2β το διάγραμμα πιθανοτήτων της $B(18,1/2)$). Τέλος, όσο περισσότερες είναι οι σειρές καρφιών και επίσης όσα περισσότερα τα σφαιρίδια που πέφτουν σε μια εκτέλεση του πειράματος τόσο καλύτερη είναι η προσαρμογή του μοντέλου στην κανονική καμπύλη (σκεφθείτε γιατί).*

*<u>Υπόδειξη-2:</u> Μπορείτε μέσω του διαδικτύου να βρείτε Java applets τα οποία προσομοιώνουν τη μηχανή του Galton σε υπολογιστικό περιβάλλον. Προσπαθήστε. Η παρακάτω εικόνα είναι ένα στιγμιότυπο από ένα τέτοιο πρόγραμμα.*



[12] *Θεωρείστε ως επιτυχία την κίνηση προς τα δεξιά και ότι ο αριθμός των υποδοχών είναι* $\nu+1$.