Η 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 Συμπερασμασματολογία. Ουσιαστικά, η κανονική κατανομή, αποτελεί το θεμέλιο της Στατιστικής Συμπερασμασματολογίας.*

Στο δεύτερο μέρος του μαθήματος (Στατιστική), 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 μεγάλο πλήθος φαινομένων και πειραμάτων.

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 (Central Limit Theorem)*** τις βάσεις του οποίου έθεσαν δύο μεγάλοι Μαθηματικοί. Ο *Abraham De Moivre* το 1733 και, έναν αιώνα περίπου αργότερα, το 1812, ο *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.

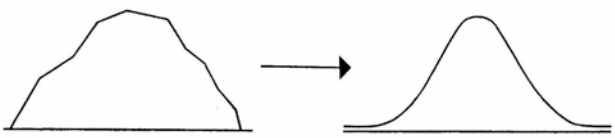
Σύμφωνα με το *Κεντρικό Οριακό Θεώρημα*, το άθροισμα και -επομένως- η μέση τιμή, μεγάλου αριθμού ανεξάρτητων παρατηρήσεων, ακολουθεί κατά προσέγγιση *κανονική κατανομή*, <u>ανεξαρτήτως από το ποια κατανομή ακολουθούν οι παρατηρήσεις</u>. Πώς, όμως, αυτό το αποτέλεσμα ερμηνεύει τη μεγάλη εφαρμοσιμότητα 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, κά. 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* λεει ότι αυτά ακριβώς τα χαρακτηριστικά περιγράφονται ικανοποιητικά από την *κανονική κατανομή*. Επιπλέον, το *Κεντρικό Οριακό Θεώρημα* **<u>συνδέει</u>** την *κανονική κατανομή* με **<u>οποιαδήποτε άλλη κατανομή</u>** (αφού δεν προϋποθέτει να ακολουθούν οι παρατηρήσεις την *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

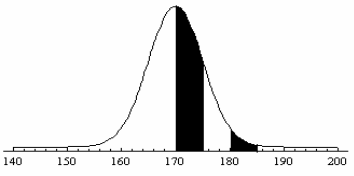
γίνουν μετρήσεις που να επαληθεύουν ένα τέτοιο συμπέρασμα[1]. Μια από τις πρώτες εφαρμογές της *κανονικής κατανομής*, έγινε το 1809 από το μεγάλο Γερμανό Μαθηματικό C*arl F. Gauss* ο οποίος διαπίστωσε ότι τα σφάλματα που γίνονται σε αστρονομικές παρατηρήσεις μπορούν να περιγραφούν ικανοποιητικά από την *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. 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[2], αναφέρουμε ότι στο γερμανικό χαρτονόμισμα των δέκα μάρκων υπήρχαν, φωτογραφία του *Gauss*, η κανονική καμπύλη και ο μαθηματικός τύπος της!!

Ιδιότητες της κανονικής καμπύλης

Στην *κανονική καμπύλη* έχουμε ήδη αναφερθεί. Όπως όλες οι καμπύλες συχνοτήτων, προκύπτει ως προσέγγιση του *πολυγώνου συχνοτήτων* των τιμών μιας συνεχούς μεταβλητής. Αυξάνοντας, δηλαδή, το μέγεθος του δείγματος και κατασκευάζοντας το ιστόγραμμα με ολοένα και μικρότερου πλάτους κλάσεις ($c \to 0$), το αντίστοιχο *πολύγωνο* προσεγγίζει μια ομαλή-λεία καμπύλη.



Η *κανονική καμπύλη* έχει κωδωνοειδή μορφή, 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 <u>πυκνότητα</u>, 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* 170 cm και *τυπική απόκλιση* 5 cm, τότε μεταξύ 170 cm και 175 cm βρίσκονται περισσότερα άτομα από όσα βρίσκονται μεταξύ 180 cm και 185 cm. Επίσης, πολύ λίγα άτομα έχουν ύψος μεγαλύτερο από 185 cm ή μικρότερο από 155 cm.



[1] *Δες και το σχόλιο στη σελίδα 78*

[2] *Ενδεικτική της αναγνώρισης της σημασίας της κανονικής κατανομής και του έργου του Gauss*

Συνήθως, η *ομαλή καμπύλη* μιας συνεχούς μεταβλητής μπορεί να περιγραφεί-προσεγγισθεί από ένα *μαθηματικό μοντέλο* το οποίο ονομάζεται ***συνάρτηση πυκνότητας***. Η *συνάρτηση πυκνότητας* της *κανονικής κατανομής* έχει τύπο:

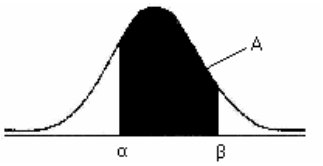
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad -\infty < x < +\infty$$

όπου, $\sigma > 0$ η *τυπική απόκλιση* και μ η *μέση τιμή* της μεταβλητής, με $-\infty < \mu < +\infty$.

<u>*Σημείωση*</u>
Παρατηρείστε ότι στον τύπο της συνάρτησης πυκνότητας της κανονικής κατανομής, εμφανίζονται δύο πολύ «διάσημοι» άρρητοι αριθμοί: ο $\pi \cong 3{,}14$ *και ο* $e \cong 2{,}71$.

Το εμβαδόν του χωρίου που περικλείεται από την καμπύλη της *συνάρτησης πυκνότητας* και τον άξονα των τιμών της X είναι 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 επόμενο σχήμα, εκφράζει την πιθανότητα η X να πάρει κάποια τιμή μεταξύ των τιμών α και β, δηλαδή, $A = P(\alpha \le X \le \beta)$.



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

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

Επισήμανση
Πρέπει να επισημάνουμε ότι η τιμή $f(x)$ της *συνάρτησης πυκνότητας* για συγκεκριμένη τιμή x της μεταβλητής X, <u>δεν αντιστοιχεί σε πιθανότητα</u>, δηλαδή, <u>δεν ισχύει</u> $f(x) = P(X = x)$. Εξάλλου, στις <u>συνεχείς</u> μεταβλητές, η πιθανότητα $P(X = x)$ είναι μηδέν[3]. Τι εκφράζει επομένως η $f(x)$; Η $f(x)$ εκφράζει <u>*πυκνότητα*</u>,

---

[3] *Αυτός είναι και ο λόγος που στις συνεχείς μεταβλητές έχουμε:* $P(X \le \alpha) = P(X < \alpha)$, $P(X \ge \alpha) = P(X > \alpha)$ *και* $P(\alpha \le X \le \beta) = P(\alpha < X < \beta) = P(\alpha \le X < \beta) =$ $= P(\alpha < X \le \beta)$

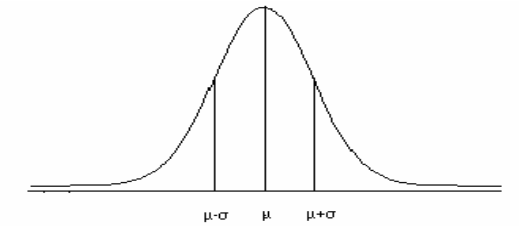
δηλαδή, όσο μεγαλύτερη είναι η τιμή $f(x)$ τόσο περισσότερο πιθανό είναι να πάρει η μεταβλητή X τιμές κοντά στο x.

<u>*Ερώτηση*</u>

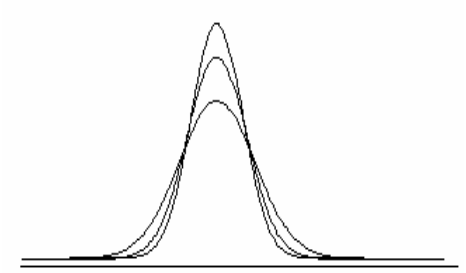
Η $f(x)$ μπορεί να πάρει τιμές μεγαλύτερες του 1;

Παρατήρηση

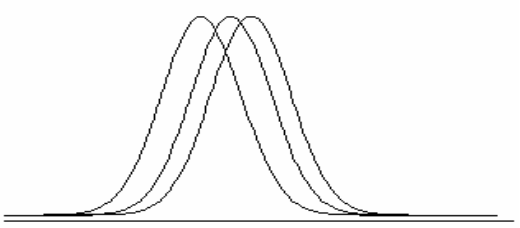
Παρατηρείστε ότι η καμπύλη της συνάρτησης πυκνότητας της κανονικής κατανομής, στη θέση $x = \mu$ παρουσιάζει μέγιστη τιμή (ίση με $\frac{1}{\sigma\sqrt{2\pi}}$) *και στις θέσεις $x = \mu - \sigma$ και $x = \mu + \sigma$ παρουσιάζει σημεία καμπής.*



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 μ και σ παίρνουμε διαφορετικές κανονικές καμπύλες. Για παράδειγμα, οι *κατανομές,*



είναι όλες *κανονικές κατανομές*, με ίδια *μέση τιμή* και διαφορετικές *τυπικές αποκλίσεις*. Επίσης, οι *κατανομές*,



είναι όλες *κανονικές κατανομές* με ίδιες *τυπικές αποκλίσεις* και διαφορετικές *μέσες τιμές*.

Είναι φανερό, ότι αλλαγή της *μέσης τιμής* προκαλεί μόνο μετατόπιση της *κανονικής καμπύλης* σε μια νέα θέση. Αλλαγή, της *τυπικής απόκλισης*, όμως, προκαλεί αλλαγή στην *κανονική καμπύλη (χωρίς, φυσικά να αλλάζει η κωδωνοειδής μορφή της)*. Για παράδειγμα, 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.

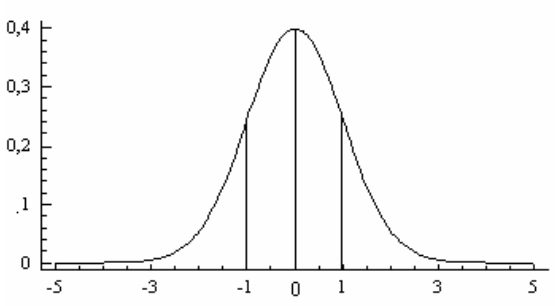
Επισημαίνουμε ότι οι παράμετροι μ και σ χαρακτηρίζουν την *κανονική κατανομή*, δηλαδή, μπορούμε να την προσδιορίσουμε πλήρως αν γνωρίζουμε μόνο τη *μέση τιμή της,* μ και την *τυπική απόκλισή της,* σ. Η *κανονική κατανομή* με *μέση τιμή* μ και *διασπορά* σ^2 (δηλαδή *τυπική απόκλιση* σ) συμβολίζεται με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

Η Τυποποιημένη κανονική Κατανομή

Η *κανονική κατανομή* που έχει *μέση τιμή* **0** και *τυπική απόκλιση* **1** (άρα και *διασπορά* **1**), συμβολίζεται με $N(0,1)$ και ονομάζεται ***τυποποιημένη κανονική κατανομή (standard normal distribution)***. Μια τυχαία μεταβλητή που ακολουθεί την *τυποποιημένη κανονική κατανομή,* έχει επικρατήσει να συμβολίζεται με Z και η συνάρτηση πυκνότητάς της με $\varphi(z)$. Προφανώς είναι:

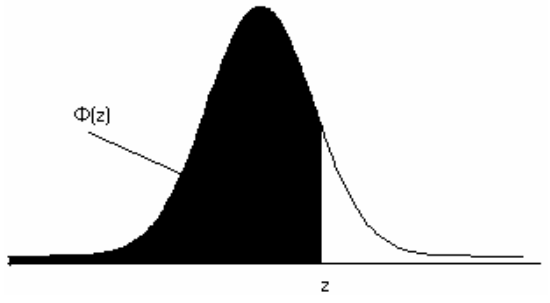
$$\varphi(z)=\frac{1}{\sqrt{2\pi}}\,e^{-\frac{z^2}{2}},\ -\infty<z<+\infty.$$

Σύμφωνα με τα προηγούμενα, η καμπύλη της *τυποποιημένης κανονικής κατανομής* στη θέση $x=0$ παρουσιάζει *μέγιστη τιμή* (ίση με $\frac{1}{\sqrt{2\pi}}=0.4$) και στις θέσεις $x=-1$ και $x=1$ παρουσιάζει *σημεία καμπής.*



Υπολογισμός πιθανοτήτων

Σύμφωνα με όσα ήδη έχουμε αναφέρει, ο υπολογισμός πιθανοτήτων, ανάγεται στον υπολογισμό εμβαδών επιπέδων χωρίων. Δυστυχώς, καμία από τις γνωστές τεχνικές ολοκλήρωσης δε μας επιτρέπει τον αναλυτικό υπολογισμό του κατάλληλου, κατά περίπτωση, ορισμένου ολοκληρώματος της $f(x)$. Στην πράξη, για να υπολογίσουμε τις πιθανότητες που αφορούν τις τιμές τυχαίας μεταβλητής που ακολουθεί *κανονική κατανομή* $N(\mu,\sigma^2)$, χρησιμοποιούμε τον ***πίνακα της τυποποιημένης κανονικής κατανομής*** $N(0,1)$. Ο *πίνακας της τυποποιημένης κανονικής κατανομής*[4], μας δίνει την πιθανότητα $P(Z\le z)$ για όλα τα z από 0 έως 3,59 με βήμα 0,01. Ας συμβολίσουμε αυτή την πιθανότητα με $\Phi(z)$. Δηλαδή, $\Phi(z)=P(Z\le z)$. Ο πίνακας, επομένως, της *τυποποιημένης κανονικής κατανομής* μας δίνει το εμβαδόν του σκιαγραφημένου χωρίου το οποίο συμβολίζεται με $\Phi(z)$.

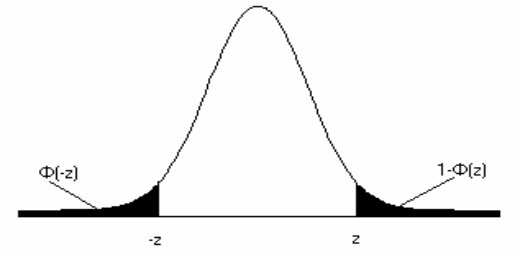


[4] *Υπάρχει σε κάθε βιβλίο Πιθανοτήτων και Στατιστικής (δες σελ. 83).*

Εύκολα μπορεί να αποδειχθεί ότι:

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

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



<u>*Σημείωση*</u>

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

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

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:

$P(Z \le 0) = \Phi(0) = 0.5$

$P(Z \le 1{,}37) = \Phi(1.37) = 0.9147$

$P(Z > 1.37) = 1 - P(Z \le 1.37) = 1 - \Phi(1.37) = 1 - 0.9147 = 0.0853$

$P(Z \le -1.55) = \Phi(-1.55) = 1 - \Phi(1.55) = 1 - 0.9394 = 0.606$

$P(-1.55 \le Z \le 2.1) = \Phi(2.1) - \Phi(-1.55) = \Phi(2.1) - [1 - \Phi(1.55)] =$

$= \Phi(2.1) - 1 + \Phi(1.55) = 0.9821 - 1 + 0.9394 = 0.9215$

$P(-1 \le Z \le 1) = 2 \cdot \Phi(1) - 1 = 2 \cdot 0.8413 - 1 = 0.6826 \cong 68.3\%$

$P(-2 \le Z \le 2) = 2 \cdot \Phi(2) - 1 = 2 \cdot 0.9772 - 1 = 0.9544 \cong 95.5\%$

$P(-3 \le Z \le 3) = 2 \cdot \Phi(3) - 1 = 2 \cdot 0.9987 - 1 = 0.9974 \cong 99.7\%$

<u>*Ερώτηση*</u>

Μπορείτε να εξηγήσετε γιατί ο πίνακας της τυποποιημένης κανονικής κατανομής δίνει τις τιμές της $\Phi(z)$ *μέχρι* $z = 3.59$*;*

Όπως, ήδη, έχουμε αναφέρει, μέσω του πίνακα *της τυποποιημένης κανονικής κατανομής,* μπορούμε να υπολογίσουμε πιθανότητες για οποιαδήποτε *κανονική κατανομή* $N(\mu, \sigma^2)$. Αυτό μπορεί να γίνει διότι έχει αποδειχθεί ότι:

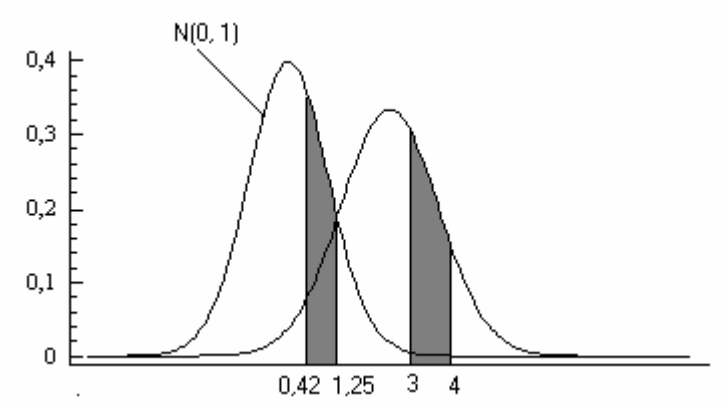
Αν η τυχαία μεταβλητή X ακολουθεί την κανονική κατανομή $N(\mu, \sigma^2)$ *τότε η τυχαία μεταβλητή* $Z = \dfrac{X - \mu}{\sigma}$*, ακολουθεί την τυποποιημένη κανονική* $N(0,1)$.

Έτσι, αν η τυχαία μεταβλητή *X*, ακολουθεί *κανονική κατανομή* με $\mu = 2.5$ και $\sigma = 1.2$, η πιθανότητα $P(3 \le X \le 4)$ μπορεί να υπολογισθεί ως εξής:

$$P(3 \le X \le 4) = P(\frac{3-2.5}{1,2} \le \frac{X-2.5}{1,2} \le \frac{4-2.5}{1,2}) = P(0.42 \le Z \le 1.25) =$$

$$= \Phi(1.25) - \Phi(0.42) = 0.8944 - 0.6628 = 0.2316$$

Στο παρακάτω σχήμα φαίνεται ο μετασχηματισμός της $N(2.5, 1.2^2)$ στην $N(0,1)$.



<u>*Παράδειγμα 1*</u>

Έχει παρατηρηθεί ότι ο χρόνος 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,

α) να είναι το πολύ 15 *min*

β) να είναι περισσότερο από 22 *min*

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

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

<u>*Παράδειγμα 2*</u>

Στο προηγούμενο κεφάλαιο είχαμε αναφέρει, χωρίς απόδειξη, ότι αν ένα σύνολο παρατηρήσεων προέρχεται από *κανονική κατανομή,* τότε το ποσοστό των παρατηρήσεων που απέχει από τη *μέση τιμή,* λιγότερο

α) από μια *τυπική απόκλιση* είναι περίπου 68%

β) από δύο *τυπικές αποκλίσεις* είναι περίπου 95%

γ) από τρεις *τυπικές αποκλίσεις* είναι περίπου 99.7%.

Μπορούμε τώρα να αποδείξουμε αυτή την πρόταση και μάλιστα, σε γενικότερη μορφή: *Αν ένα σύνολο παρατηρήσεων προέρχεται από την κανονική κατανομή* $N(\mu, \sigma^2)$*, ποιο είναι το ποσοστό των παρατηρήσεων που απέχει από τη μέση τιμή* μ *λιγότερο από k τυπικές αποκλίσεις;*

Ζητάμε την πιθανότητα $P(\mu - k \cdot \sigma \le X \le \mu + k \cdot \sigma)$. Σύμφωνα με τα προηγούμενα, μπορούμε πλέον να υπολογίσουμε αυτήν την πιθανότητα. Πράγματι,

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

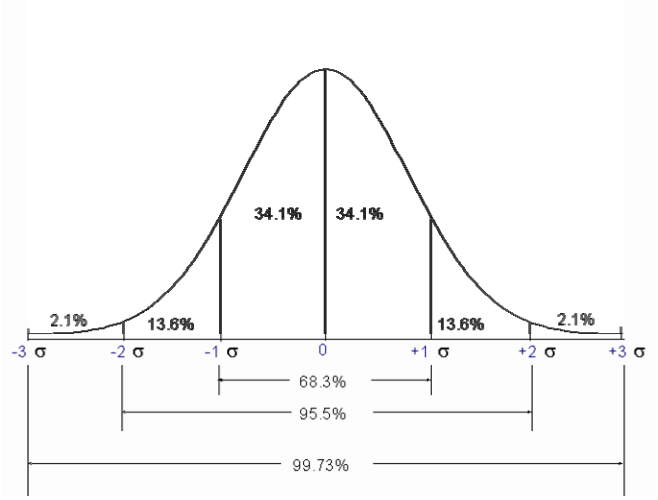
$$= P(-k \le Z \le +k) = 2\cdot\Phi(k) - 1.$$

Έτσι, για $k = 1,\ 2,\ 3$, έχουμε:

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

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

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



<u>*Σχόλιο*</u>

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 κ.λπ., 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 α. 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 εξήγηση 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 *(βλ. παράδειγμα-1) είναι ίση με:*

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

πρακτικά είναι μηδέν. Όμως, αυτή η 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 <u>την κανονική κατανομή αυτή καθαυτή</u>, από τα τυχαία φαινόμενα που <u>προσεγγίζονται ικανοποιητικά από την κανονική κατανομή</u>. 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, όχι της κανονικής κατανομής!

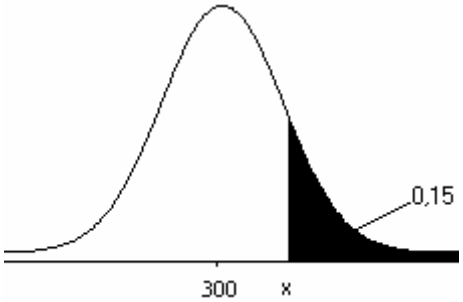
Ας δούμε ένα διαφορετικό παράδειγμα.

Παράδειγμα 3

Οι υποψήφιοι για εγγραφή σε ένα Μεταπτυχιακό Τμήμα Πανεπιστημίου, υποβάλλονται σε ένα τεστ. Το τεστ έχει σχεδιασθεί έτσι, ώστε η βαθμολογία των υποψηφίων στο τεστ να κατανέμεται κανονικά με *μέση τιμή* 300 και *τυπική απόκλιση* 60. Η πολιτική του Πανεπιστημίου είναι να δέχεται ως φοιτητές, το 15% των υποψηφίων με τη μεγαλύτερη βαθμολογία στο τεστ. Ποια είναι η ελάχιστη βαθμολογία που επιτρέπει την εισαγωγή στο Μεταπτυχιακό Τμήμα;

Απάντηση

Αν X είναι η βαθμολογία των υποψηφίων, ζητάμε την τιμή x της μεταβλητής για την οποία ισχύει: $P(X \geq x) = 0.15$.



Για τον προσδιορισμού του σημείου x της κατανομής, εργαζόμαστε ως εξής:

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

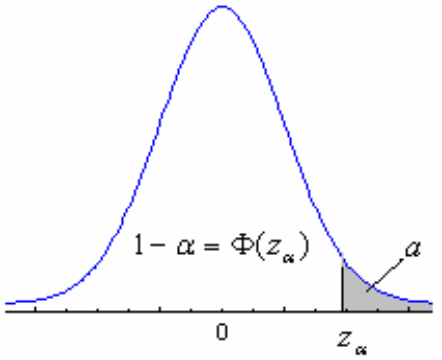
$$1 - P(Z < \frac{x-300}{60}) = 0.15 \Rightarrow P(Z < \frac{x-300}{60}) = 1 - 0.15 \Rightarrow P(Z < \frac{x-300}{60}) = 0.85 \Rightarrow$$

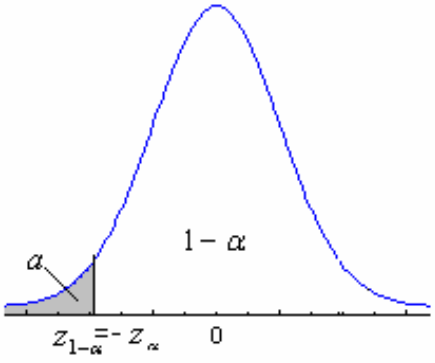
$$\Phi(\frac{x-300}{60}) = 0.85.$$

Στον πίνακα της *τυποποιημένης κανονικής κατανομής*, βλέπουμε ότι το εμβαδόν (η πιθανότητα) 0.85 βρίσκεται μεταξύ των εμβαδών 0.8485 και 0.8508 που αντιστοιχούν στις τιμές 1.03 και 1.04. Κάνοντας παρεμβολή βρίσκουμε $z = \frac{1.03+1.04}{2} = 1.035$.

Επομένως, $\frac{x-300}{60} = 1.035 \Rightarrow x = 362.1$. Άρα, η ζητούμενη βαθμολογία είναι 362.1.

Σημείωση: Είναι προφανές ότι με την προηγούμενη μέθοδο υπολογίζουμε τα ποσοστημόρια της κατανομής. Ειδικότερα, για την τυποποιημένη κανονική κατανομή $Z \sim N(0,1)$*, ο αριθμός* z *για τον οποίο ισχύει* $P(Z \geq z) = \alpha$, $0 < \alpha < 1$, *ονομάζεται* ***άνω*** α ***ποσοστιαίο σημείο*** *της τυποποιημένης κανονικής κατανομής και συμβολίζεται με* z_α. *Δηλαδή,* $P(Z \geq z_\alpha) = \alpha$. *Προφανώς, λόγω συμμετρίας της κατανομής, ισχύει:* $z_{1-\alpha} = -z_\alpha$.





Άσκηση: Επαληθεύστε ότι, $z_{0.01} = 2.33$, $z_{0.05} = 1.645$, $z_{0.10} = 1.28$, $z_{0.99} = -2.33$.

Ας δούμε ένα ακόμη παράδειγμα.

Παράδειγμα 4

Μια αυτόματη μηχανή συσκευασίας τροφίμων έχει προγραμματισθεί να συσκευάζει δημητριακά σε συσκευασίες των 1.5kgr. Έχει παρατηρηθεί ότι η ποσότητα δημητριακών κάθε συσκευασίας ακολουθεί κανονική κατανομή με μέση τιμή $\mu = 1.5\,\text{kgr}$ και τυπική απόκλιση $\sigma = 0.1\,\text{kgr}$.

α) Τι ποσοστό των συσκευασιών περιέχει ποσότητα που υπερβαίνει τα 1.6kgr;

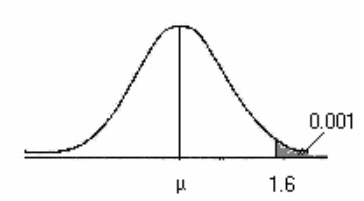
β) Σε τι ποσότητα πρέπει να ρυθμισθεί η μηχανή έτσι ώστε μόνο στο 0.001 των περιπτώσεων η ποσότητα δημητριακών στη συσκευασία να υπερβαίνει τα 1.6kgr;

Απάντηση

Έστω X η ποσότητα που περιέχεται στις συσκευασίες.

α) $X \sim N(1.5,\, 0.1^2)$. Εύκολα υπολογίζεται ότι το ποσοστό συσκευασιών που υπερβαίνουν τα 1.6kgr, δηλαδή, η πιθανότητα $P(X > 1.6)$ είναι 0.1587.

β) $X \sim N(\mu,\, 0.1^2)$. Πρέπει να προσδιορισθεί η μέση τιμή μ ώστε $P(X > 1.6) = 0.001$.



Έχουμε:

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

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

Άρα, $\frac{1.6-\mu}{0.1} = 3.09 \Rightarrow \mu = 1.29$. Δηλαδή, η μηχανή πρέπει να ρυθμισθεί στα 1.29kgr.

Σε πρακτικά προβλήματα, ενδιαφέρουν πολλές φορές 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 5

Στα ζώα μιας κτηνοτροφικής μονάδας δίνεται τροφή τρεις φορές την ημέρα. Η ποσότητα θερμίδων που παίρνουν κάθε φορά είναι 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.

Απάντηση

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 (ημερεσίως). Γνωρίζουμε ότι το διαιτολόγιο έχει ρυθμισθεί έτσι ώστε: $X_1 \sim N(500, 50^2)$, $X_2 \sim N(1700, 200^2)$ 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 (η απόδειξη είναι εκτός των σκοπών του μαθήματος).

Αν $X_1, X_2, ..., X_n$ *ανεξάρτητες τυχαίες μεταβλητές με* $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2, ..., n$, *τότε,* $S_n = \sum_{i=1}^{n} X_i \sim N(\mu_1 + \mu_2 + ... + \mu_n,\ \sigma_1^2 + \sigma_2^2 + ... + \sigma_n^2)$.

Αν για κάθε $i = 1, 2, ..., n$ *είναι* $X_i \sim N(\mu, \sigma^2)$ *δηλαδή αν οι* $X_1, X_2, ..., X_n$ *είναι ανεξάρτητες και ισόνομες κανονικές κατανομές, τότε,* $S_n = \sum_{i=1}^{n} X_i \sim N(n \cdot \mu,\ n \cdot \sigma^2)$

Επειδή οι $X_1,\ X_2,\ X_3$ είναι ανεξάρτητες, από την παραπάνω 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:

$$P(2975 < S_3 < 3025) = P\left(\frac{2975 - 3000}{\sqrt{52500}} < \frac{S_3 - 3000}{\sqrt{52500}} < \frac{3025 - 3000}{\sqrt{52500}}\right) =$$

$$= P(-0.11 < Z < 0.11) = 2\Phi(0.11) - 1 = 0.733.$$

Παρατήρηση: Από την προηγούμενη πρόταση εύκολα προκύπτει η επόμενη (χρήσιμη, επίσης, σε πολλά πρακτικά προβλήματα).

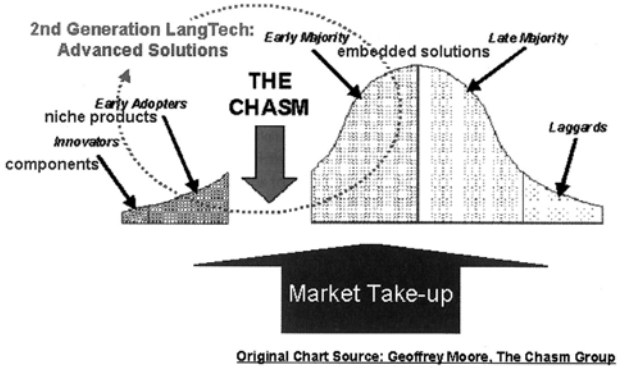
Αν $X_1, X_2, ..., X_n$ *ανεξάρτητες τυχαίες μεταβλητές με* $X_i \sim N(\mu, \sigma^2)$ *για κάθε* $i = 1, 2, ..., n$, *τότε,* $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \sim N(\mu,\ \frac{\sigma^2}{n})$.

Γενικότερα,

αν $X_1, X_2, ..., X_n$ *ανεξάρτητες τυχαίες μεταβλητές με* $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2, ..., n$, *τότε,* $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \sim N\left(\frac{\mu_1 + \mu_2 + ... + \mu_n}{n},\ \frac{\sigma_1^2 + \sigma_2^2 + ... + \sigma_n^2}{n^2}\right)$.

Ως εφαρμογή, απαντήστε στο ακόλουθο ερώτημα-συνέχεια του *Παραδείγματος 5*: Ποια είναι η πιθανότητα, η μέση ποσότητα θερμίδων που θα πάρει το ζώο σε ένα χρόνο (365 ημέρες) να είναι μεταξύ 2975cal και 3025cal.
Απάντηση: Η ζητούμενη πιθανότητα είναι 0.9624 (γιατί;).

Ερώτηση: Τι καταλαβαίνετε από την παρακάτω εικόνα[5]*;*



[5] *Η εικόνα αυτή δημοσιεύθηκε στη σελίδα 13 της έκδοσης Benchmarking Human Language Technologies (HLT) progress in Europe, The EUROMAP study, Andrew Joschelyne and Rose Lockwood, Copenhagen, 2003.*

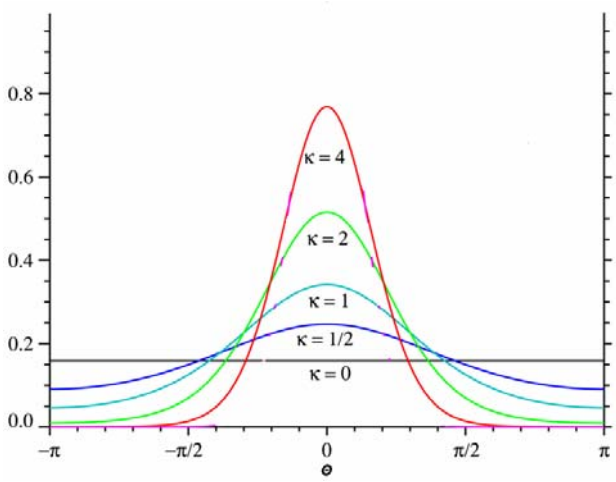
Η Κατανομή von Mises

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:

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

όπου: *μ η μέση κατεύθυνση* (με τιμές σε διάστημα πλάτους 2π όπως και η *Θ)* και *k* παράμετρος που παίρνει μη αρνητικές τιμές ($\kappa \geq 0$) και εκφράζει τη *συγκέντρωση* των τιμών της *Θ* γύρω από τη *μέση κατεύθυνση*. Το $I_0(\kappa)$ είναι σταθερά[6].



Για μεγάλα *k,* η κατανομή *von Mises* προσεγγίζει την ***κανονική κατανομή*** με $\mu = \bar{\theta}$ και $\sigma^2 = \frac{1}{k}$ (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, όλες οι κατευθύνσεις έχουν την ίδια πιθανότητα ή, ακριβέστερα, για κάθε ϑ, δηλαδή, για κάθε κατεύθυνση ϑ, η πιθανότητα να πάρει η μεταβλητή *Θ* τιμή κοντά στη ϑ είναι για όλα τα ϑ ίδια[7].

Σημείωση:

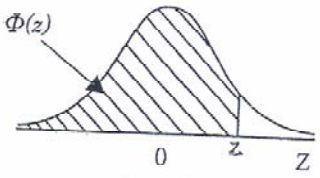
Αν $\vartheta_1, \vartheta_2, \ldots, \vartheta_n$ δείγμα από πληθυσμό που ακολουθεί κατανομή von Mises, τότε, για την εφαρμογή μεθόδων της στατιστικής συμπερασματολογίας (π.χ. στατιστικοί έλεγχοι), η παράμετρος k εκτιμάται μέσω του μέσου μέτρου $\bar{r}$ του διανύσματος $\vec{r}$ (υπάρχουν σχετικοί πίνακες που δίνουν εκτιμήσεις των τιμών του k για διάφορες τιμές του $\bar{r}$).

---

[6] $I_0(x)$ *είναι η συνάρτηση Bessel τάξης 0.*

[7] *ή και αλλιώς, η πιθανότητα να πάρει η Θ τιμή σε ένα διάστημα είναι ανάλογη του πλάτους του διαστήματος.*

Η Συνάρτηση Κατανομής της Τυποποιημένης Κανονικής Κατανομής



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

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

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7703	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7.995	.8023	.8051	.8078	.8106	.8133
.9	8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998