

Ενότητα 0

Εισαγωγή – Βασικές Έννοιες.

Παρατηρήσεις και ενδεχόμενα. Πληθυσμοί και δείγματα

Τα δεδομένα με τα οποία ασχολείται η Στατιστική προέρχονται από παρατηρήσεις και πειράματα που εμπίπτουν σε οποιονδήποτε τομέα, της Φυσικής, της Χημείας, των λεγόμενων Ανθρωπιστικών Επιστημών κλπ. Αυτές οι παρατηρήσεις μπορεί να είναι οσοδήποτε απλές ή περίπλοκες, αλλά τελικά καταλήγουν στην καταγραφή ορισμένων αριθμητικών τιμών ή μη αριθμητικών ιδιοτήτων που σχετίζονται με ένα ή περισσότερα συγκεκριμένα μεγέθη ή πειραματικά αποτελέσματα. Στο πιο απλό πείραμα προσφερόμενο για στατιστική επεξεργασία, το γνωστό και χιλιοειπωμένο “κορώνα-γράμματα”, καταγράφουμε τα αποτελέσματα κάθε ρίψης του νομίσματος ως “κορώνα” ή “γράμματα”.

Κάθε δυνατό αποτέλεσμα (αριθμητική τιμή ή μη αριθμητική ιδιότητα) μιας παρατήρησης (αλλά και κάθε σύνολο τέτοιων αποτελεσμάτων) αποτελεί ένα **ενδεχόμενο**. Έτσι, στη ρίψη νομίσματος έχουμε δύο δυνατά αποτελέσματα: κορώνα και γράμματα. Αν μετρούσαμε το ύψος και το βάρος των φοιτητών του ΤΜΕΥ, τα δυνατά ενδεχόμενα είναι θετικοί πραγματικοί αριθμοί που αποτελούν ένα συνεχές σύνολο, καθώς και τα διαστήματα αυτών των αριθμών. Επειδή όμως πάντα κάνουμε κάποια προσέγγιση ή στρογγύλευση (π.χ. μετράμε το βάρος σε kg, το ύψος σε cm) θα πάρουμε ένα διακριτό σύνολο ρητών ή ακεραίων αριθμών ως πιθανά ενδεχόμενα. Τέλος, τόσο στη ρίψη νομίσματος όσο και σε παρατηρήσεις τύπου χρώμα ματιών ή μαλλιών τα ενδεχόμενα σχηματίζουν ένα διακριτό σύνολο *μη αριθμητικών* τιμών.

Συνοψίζοντας, τα στατιστικά δεδομένα κατηγοριοποιούνται ως εξής:

- Μη αριθμητικά
- Αριθμητικά:
 - διακριτά
 - συνεχή

Όπως είπαμε, συνήθως, ακόμη και για συνεχή αριθμητικά δεδομένα, μας ενδιαφέρει η καταγραφή τους ως διακριτά, για πρακτικούς λόγους. Εξ άλλου, τίποτα δε μας εμποδίζει να αντιστοιχίσουμε έναν αριθμό και στα στοιχεία ενός συνόλου μη αριθμητικών δεδομένων, π.χ. Κορώνα = 0, Γράμματα = 1, αν και ορισμένοι τρόποι επεξεργασίας τους δε θα έχουν φυσικό νόημα σε αυτή την περίπτωση. Επομένως, σε κάθε περίπτωση μπορούμε να ορίσουμε μια αντιστοιχία ανάμεσα στα δυνατά αποτελέσματα των παρατηρήσεων ενός μεγέθους ή ιδιότητας και στο σύνολο των πραγματικών, εν γένει, αριθμών. Αυτή, από μαθηματική άποψη, είναι μια συνάρτηση και στα πλαίσια της στατιστικής ορίζει αυτό που είναι γνωστό ως **τυχαία μεταβλητή**. Οι τυχαίες μεταβλητές συνήθως παριστάνονται με κεφαλαία γράμματα X, Y, Z , π.χ. το βάρος B και το ύψος Y των φοιτητών του ΤΜΕΥ.

Οι τιμές μιας τυχαίας μεταβλητής είναι ή κωδικοποιούν το σύνολο των ενδεχομένων. Αν το

μέγεθος ή ιδιότητα που καταγράφουμε διεπόταν από έναν γνωστό νόμο, π.χ. τύπου αιτίας-αποτελέσματος, που μας επέτρεπε να προβλέψουμε με ακρίβεια το αποτέλεσμα κάθε πειράματος ή παρατήρησης, τότε δε θα υπήρχε ανάγκη στατιστικής μελέτης αυτού. Ένας νόμος τέτοιου είδους δεν υπάρχει πάντα και αν υπάρχει δεν είναι πάντα γνωστός. Ένας από τους στόχους της στατιστικής επεξεργασίας είναι η ανακάλυψη τέτοιων “νόμων”. Αν αυτό δε γίνεται, θέλουμε τουλάχιστον να γνωρίζουμε πόσο συχνά εμφανίζεται κάθε ενδεχόμενο και επομένως, ποια ενδεχόμενα θα είναι περισσότερο αναμενόμενα σε κάθε μελλοντική παρατήρηση, θα θέλαμε δε, να εκφράσουμε αυτή την πληροφορία ποσοτικά και με όσο το δυνατό μεγαλύτερη ακρίβεια. Προς αυτή την κατεύθυνση βοηθά η καταγραφή της συχνότητας με την οποία εμφανίζεται κάθε δυνατό αποτέλεσμα και το σύνολο των συχνοτήτων κάθε τιμής μιας τυχαίας μεταβλητής X , βοηθά να ορίσουμε την **κατανομή** αυτής της μεταβλητής. Η έννοια της κατανομής είναι κεντρικής σημασίας και αποτελεί το σημαντικότερο “εργαλείο” με το οποίο επωφελούμαστε από τις μεθόδους στατιστικής επεξεργασίας. Επιδίωξή μας είναι να προσεγγίσουμε την κατανομή της τυχαίας μεταβλητής με μια μαθηματική έκφραση που μπορεί να γραφεί αναλυτικά. Σε πολλές και σημαντικές περιπτώσεις αυτό είναι εφικτό και έτσι μπορούμε να βρούμε τουλάχιστον ένα στατιστικό νόμο αντί για ένα νόμο αιτίας-αποτελέσματος που εκφράζει ποσοτικά τι είδους ενδεχόμενα να περιμένουμε και με ποια συχνότητα εμφάνισης. Πρέπει να τονίσουμε όμως ότι η συσχέτιση μεταξύ δυό μεταβλητών δεν είναι απαραίτητα αιτιακή σχέση! Αυτή μπορεί να είναι βαθύτερα κρυμμένη σε πιο περίπλοκες συνδέσεις των φαινομένων από τις οποίες εμείς βλέπουμε μόνο την επιφάνεια.

Οι παρατηρήσεις μας αφορούν ένα σύνολο στοιχείων συγκεκριμένης κατηγορίας ή είδους που μας ενδιαφέρει να μελετήσουμε. Το σύνολο όλων αυτών των στοιχείων είναι ο **πληθυσμός** που μας ενδιαφέρει. Π.χ. όταν γίνονται δημοσκοπήσεις για τις προτιμήσεις ψήφου των πολιτών, πληθυσμός θεωρείται ολόκληρο το εκλογικό σώμα. Αυτός είναι πολύ μεγάλος αλλά πεπερασμένος. Υπάρχουν και περιπτώσεις όπου ο πληθυσμός έχει άπειρο μέγεθος. Το πείραμα τύπου “κορώνα-γράμματα” αναφέρεται σε πληθυσμό με άπειρο μέγεθος γιατί αυτό περιλαμβάνει όλες τις δυνατές ρίψεις για τις οποίες δεν υπάρχει κανένα όριο (εκτός ίσως από την ηλικία του σύμπαντος...). Δεν πρέπει λοιπόν να συγχέουμε τον πληθυσμό με το σύνολο των δυνατών αποτελεσμάτων (κορώνα, γράμματα) που μπορεί να έχει πεπερασμένο μέγεθος ή εύρος.

Αν μπορούσαμε να μελετήσουμε ολόκληρο τον υπό μελέτη πληθυσμό δε θα χρειαζόταν περαιτέρω στατιστική μελέτη αυτού, ακόμη και αν δεν ήταν γνωστός κάποιος νόμος που να διέπει την εμφάνιση των δυνατών ενδεχομένων. Τα αποτελέσματα των παρατηρήσεών μας θα ήταν η μόνη βέβαιη απάντηση σε όλα τα σχετικά ερωτήματά μας, όπως το αποτέλεσμα μιας εκλογικής αναμέτρησης απαντά σε όλες τις εικασίες των δημοσκοπήσεων που προηγήθηκαν. Συνήθως όμως, η εξέταση ολόκληρου του πληθυσμού, δεν είναι εφικτή γιατί αυτός μπορεί να είναι πολύ μεγάλος ή απλά, άπειρος. Αν όμως μπορούμε να μελετήσουμε ένα αρκετά μικρό υποσύνολο του πληθυσμού το οποίο έχουμε λόγους να πιστεύουμε ότι δε θα είχε κάτι το “ξεχωριστό” σε σχέση με το γενικό πληθυσμό, τότε δικαιούμαστε να πούμε ότι η κατανομή των εξεταζόμενων τυχαίων μεταβλητών ως προς αυτό το υποσύνολο θα μοιάζει, ως προς τα ουσιώδη χαρακτηριστικά της, με την κατανομή του πληθυσμού. Ένα τέτοιο υποσύνολο το λέμε **δείγμα** του πληθυσμού.

Έστω ότι ο πληθυσμός μας είναι οι φοιτητές του ΤΜΕΥ και μεταβλητή μας το ύψος τους. Ας πούμε ότι τους μετρήσαμε όλους (εδώ, ο πληθυσμός είναι ευτυχώς αρκετά μικρός!) και βρήκαμε ότι έχουν ύψος από 1.40 μέχρι 2.10 μέτρα. Τότε, αν πάρουμε ένα οποιοδήποτε υποσύνολο των φοιτητών, αναγκαστικά κάθε μέλος του υποσυνόλου θα έχει ύψος ανάμεσα στα παραπάνω όρια. Επίσης, αν το 50% των φοιτητών έχει ύψος μεταξύ 1.60 και 1.80 και πάρουμε ένα δείγμα 10 φοιτητών, δε θα ήταν παράξενο αν οι 5 από αυτούς είχαν ύψος σε αυτό το διάστημα. Δεν είναι απαραίτητο ότι θα είναι ακριβώς 5 και αν βρίσκαμε 6 ή 4 σε αυτά τα όρια θα το περιμέναμε. Καταλαβαίνουμε όμως, ότι αν και οι 10 ήταν μεταξύ αυτών των ορίων, θα μας φαινόταν κάπως παράξενο να απουσιάζουν τελείως οι πιο ακραίες τιμές. Αν, από την άλλη και οι 10 ήταν εκτός των

ορίων, δηλαδή ή κάτω από 1.60 ή πάνω από 1.90, θα ήταν κάτι το απροσδόκητο και θα το θεωρούσαμε πολύ σπάνιο.

Τι είναι αυτό που θα μας επέτρεπε να έχουμε τέτοιες προσδοκίες από το δείγμα μας – και γενικότερα, πού μπορούμε να βασιστούμε για να πούμε ότι το δείγμα μας δεν ξεχωρίζει ιδιαίτερα από τον πληθυσμό ως προς τα βασικά χαρακτηριστικά του (όπως κι αν τα εννοούμε); Στην πραγματικότητα, εφόσον ο πληθυσμός αρχικά μας είναι άγνωστος, δε μπορούμε ποτέ να ξέρουμε από πριν αν το δείγμα μας μοιάζει με τον πληθυσμό ή διαφέρει από αυτόν. Μπορούμε όμως να δικαιολογήσουμε την προσδοκία μας ότι το δείγμα θα μοιάζει με αυτόν, με βάση τον τρόπο που το επιλέξαμε, δηλαδή, με βάση τη διαδικασία **δειγματοληψίας**. Συγκεκριμένα, η σωστή δειγματοληψία είναι αυτή που δίνει *ίσες ευκαιρίες* σε όλα τα δυνατά δείγματα και δεν περιέχει καμία φανερή ή έμμεση προτίμηση προς επιμέρους κατηγορίες του πληθυσμού. Π.χ. αν κατά τη μέτρηση του ύψους των φοιτητών επιλέγουμε μόνο αγόρια ή μόνο κορίτσια, καταλαβαίνουμε ότι αυτό δε θα ήταν η καλύτερη τακτική¹.

Για να δώσουμε ένα εξίσου απλό αλλά πιο γενικό παράδειγμα, ας φανταστούμε ένα πολύ μικρό πληθυσμό $\Pi = \{\alpha, \beta, \gamma, \delta, \epsilon\}$, από τον οποίο θέλουμε να πάρουμε ένα δείγμα των τριών στοιχείων. Τα δυνατά δείγματα είναι $\{\alpha, \beta, \gamma\}$, $\{\alpha, \beta, \delta\}$, $\{\alpha, \beta, \epsilon\}$, $\{\beta, \gamma, \delta\}$, $\{\beta, \gamma, \epsilon\}$, $\{\gamma, \delta, \epsilon\}$. Η δειγματοληψία πρέπει να γίνει έτσι ώστε να μπορεί να επιλεγεί οποιοδήποτε από τα παραπάνω, *εξίσου εύκολα*. Για παράδειγμα, δε θα ήταν επιτρεπτό να επιλέγαμε μόνο δείγματα που περιέχουν το στοιχείο “α” γιατί έτσι θα στερούσαμε από ένα μέρος του πληθυσμού την ευκαιρία να καταγραφεί.

Ως ένα άλλο παράδειγμα, ας φανταστούμε ένα μεγάλο σύνολο από σφαίρες ίσης διαμέτρου φτιαγμένες από δύο υλικά A και B γνωστών διαφορετικών πυκνοτήτων, ρ_A και ρ_B , με διαφορετικές αναλογίες των υλικών σε κάθε σφαίρα, έτσι ώστε τα βάρη τους να διαφέρουν. Θέλουμε να βρούμε την κατανομή βάρους των σφαιρών και από αυτή, την κατανομή των ποσοστών του υλικού A, από ένα δείγμα 50 σφαιρών. Αν το A είναι άσπρο και το B μαύρο, οπότε οι σφαίρες είναι χρωματισμένες σε διαβαθμίσεις του γκριζου, τότε, προφανώς, θα ήταν λάθος να διαλέγουμε π.χ. μόνο ανοιχτόχρωμες ή μόνο σκουρόχρωμες σφαίρες. Το καλύτερο θα ήταν να τις διαλέγαμε με κλειστά μάτια! Ας πούμε ότι η κατανομή του πληθυσμού των σφαιρών είναι τέτοια ώστε το 70% των σφαιρών έχει περιεκτικότητα μεταξύ 40% και 60% σε υλικό A, κατά βάρος, ενώ το υπόλοιπο 30% έχει πιο ακραίες αναλογίες, είτε μικρότερες είτε μεγαλύτερες. Αν κατά τη δειγματοληψία δεν είμαστε μεροληπτικοί με βάση την απόχρωση κάθε σφαίρας, όλα τα δυνατά δείγματα έχουν ίσες ευκαιρίες. Αυτό βέβαια, δεν αποκλείει το ενδεχόμενο να επιλεγεί δείγμα με ακραία κατανομή που να διαφέρει σημαντικά από αυτή του πληθυσμού. Αν έχουμε έναν αρκετά μεγάλο πληθυσμό, τότε ένα ή και περισσότερα από τα δυνατά δείγματα μπορεί να αποτελείται αποκλειστικά από σφαίρες με περιεκτικότητα κάτω από 40% σε A. Τότε, πώς μπορούμε να ισχυριζόμαστε ότι η δειγματοληψία μας είναι σωστή; Το ισχυριζόμαστε γιατί *είναι πολύ σπάνιο* να συμβεί μια τέτοια ακραία περίπτωση.

Πράγματι, μπορεί ναδειχθεί με συνδυαστικά επιχειρήματα ότι τα πιο πολλά δείγματα θα έχουν κατανομή παρόμοια με αυτή του πληθυσμού και ότι όσο πιο μεγάλα είναι, τόσο πιο μεγάλο είναι και το ποσοστό τους που θα προσομοιάζει τον πληθυσμό. Καθώς μεγαλώνει το μέγεθος δείγματος, το να πετύχουμε μια ακραία περίπτωση που αποκλίνει σημαντικά από τον πληθυσμό, μπορεί να παρομοιαστεί με το να κερδίσουμε το λαχείο!² Αυτό το πετύχαμε δίνοντας ίσες

1 Δεν είναι μικρής σημασίας το ότι γνωρίζουμε εμπειρικά την ύπαρξη σχέσης ύψους-φύλου. Όσο περισσότερα γνωρίζουμε για τον πληθυσμό τόσο πιο καλή (= αμερόληπτη) δειγματοληψία μπορούμε να κάνουμε. Και φυσικά, αν θέλαμε να διερευνήσουμε τη σχέση ύψους-φύλου σε βάθος εννοείται ότι θα κάναμε το αντίθετο: θα παίρναμε δείγματα αποκλειστικά αγοριών και αποκλειστικά κοριτσιών.

2 Μπορούμε να πειστούμε γι' αυτό ως εξής: Αν πάρουμε το μικρότερο δυνατό δείγμα (μία σφαίρα) από ένα πληθυσμό π.χ. 1000 σφαιρών, τότε μόνο 300 σφαίρες έχουν ακραίες τιμές αναλογιών. Αν πάρουμε δείγματα $N=2$ σφαιρών, τότε οι δυνατοί συνδυασμοί θα είναι 1000×999 δηλ. περίπου 10^6 ενώ οι συνδυασμοί όπου και οι δυο σφαίρες έχουν ακραίες αναλογίες είναι 300×299 δηλ. περίπου 3×10^4 . Το

ευκαιρίες σε κάθε δείγμα και για να δώσουμε ίσες ευκαιρίες, *αγνοήσαμε κατά τη δειγματοληψία, ακριβώς τον παράγοντα που θέλουμε να μελετήσουμε* (περιεκτικότητα σε Α) *αποκλείοντας από τα κριτήρια επιλογής κάθε άλλον παράγοντα που σχετίζεται με αυτόν* (εν προκειμένω, απόχρωση). Δείγματα που λαμβάνονται με τέτοια διαδικασία ονομάζονται **αντιπροσωπευτικά** γιατί πιστεύουμε ότι η διαδικασία με την οποία επιλέγονται *συνήθως* οδηγεί σε κατανομές παρόμοιες με αυτές του δειγματοζόμενου πληθυσμού. Κλείνοντας και αυτή την παράγραφο, θα πρέπει να πούμε ότι η *συσχέτιση* μεταξύ διαφόρων παραγόντων (εδώ, απόχρωση και βάρος σφαίρας) είναι καλά ορισμένη έννοια που μπορεί να μελετηθεί στατιστικά και να εκφραστεί με ακριβή ποσοτικά μέτρα, όπως θα δούμε παρακάτω, αλλά, όπως είπαμε και στο παράδειγμα για το ύψος των φοιτητών, για να επιλέξουμε το δείγμα μας, αναγκαστικά θα πρέπει να προβλέψουμε κάποιες δυνατές σχέσεις που δεν έχουμε μετρήσει ακόμη, βασισμένοι σε εύλογα φυσικά επιχειρήματα (εδώ: βάρος $\leftrightarrow$ περιεκτικότητα $\leftrightarrow$ απόχρωση).

Κατανομές και πιθανότητες

Διαβάζοντας τα παραπάνω είναι λογικό να σκεφτεί κάποιος ότι η δειγματοληψία είναι και αυτή ένα είδος παρατήρησης και τα δυνατά δείγματα είναι τα δυνατά αποτελέσματα αυτής. Μέριμνά μας είναι αυτά τα αποτελέσματα να έχουν ίση ευκαιρία να παρουσιαστούν. Επομένως αν παίρναμε πάρα πολλές φορές διάφορα δείγματα και καταγράφαμε τη συχνότητα με την οποία εμφανίστηκε το καθένα, θα βλέπαμε ότι οι μεταξύ τους διαφορές θα ήταν μικρές. Αυτό ισχύει γενικότερα για κάθε πείραμα όπου τα δυνατά αποτελέσματα δεν έχουν λόγο να υπερτερούν το ένα σε σχέση με το άλλο. Κατά τις ρίψεις ενός νομίσματος δεν έχουμε λόγο να περιμένουμε ότι η μια όψη θα εμφανίζεται συχνότερα από την άλλη. Αν κάνουμε ένα μικρό αριθμό πειραμάτων μπορεί να παίρναμε διαφορετικούς αριθμούς για κάθε όψη, π.χ. 7 φορές “κορώνα” και 3 “γράμματα”, αλλά καθώς θα προχωρούσαμε σε περισσότερες δοκιμές θα διαπιστώναμε ότι το *ποσοστό* της συχνότητας εμφάνισης κάθε ενδεχομένου θα πλησίαζε στο 0.5 (ή 50%, αν προτιμούμε να εκφραζόμαστε με εκατοστιαίες αναλογίες).

Το παραπάνω παράδειγμα εισάγει δύο έννοιες ταυτόχρονα. Η μία έννοια είναι αυτή της **ομοιόμορφης κατανομής**. Τέτοια είναι κάθε κατανομή όπου τα ενδεχόμενα ή οι δυνατές τιμές μιας τυχαίας μεταβλητής έχουν ίσες ευκαιρίες εμφάνισης και επομένως οι συχνότητες με τις οποίες παρατηρούνται τείνουν να εξισωθούν, δηλαδή να έχουν όλο και μικρότερες ποσοστιαίες διαφορές μεταξύ τους, για όλο και μεγαλύτερα δείγματα. Πολύ συχνά η κατανομή ενός πληθυσμού υπακούει σε αρκετά απλές μορφές, οι οποίες μπορούν να εκφραστούν με αναλυτικές μαθηματικές εκφράσεις. Η ομοιόμορφη κατανομή είναι η πιο απλή τέτοια περίπτωση. Αν με f_x παριστάνεται η συχνότητα εμφάνισης της τιμής x μιας μεταβλητής X , τότε για τις ρίψεις

νομισμάτων ισχύει ότι $\frac{f_K}{f_\Gamma} \rightarrow 1$ καθώς αυξάνεται το μέγεθος του δείγματος. Ένας άλλος τρόπος για να το εκφράσουμε είναι να γράψουμε το ποσοστό της συχνότητας κάθε ενδεχομένου:

$$\frac{f_\Gamma}{f_K + f_\Gamma}, \frac{f_K}{f_K + f_\Gamma} \rightarrow 0.5$$

Γενικότερα, για ένα πείραμα με N διαφορετικά δυνατά αποτελέσματα μπορούμε να γράψουμε:

$$p_i = \frac{f_i}{f_1 + f_2 + \dots + f_N} \rightarrow \frac{1}{N}, \quad i=1, 2, \dots, N. \text{ Ονομάζουμε τον αριθμό } p_i, \text{ **ποσοστό εμφάνισης**}$$

του αποτελέσματος i και τη διαίρεση κάθε συχνότητας με το άθροισμα όλων των συχνοτήτων για τη λήψη των ποσοστών, ονομάζουμε **κανονικοποίηση** της κατανομής. Ο παράγοντας

$$\frac{1}{f_1 + f_2 + \dots + f_N} \text{ με τον οποίο πολλαπλασιάζεται κάθε συχνότητα για να δώσει το ποσοστό,}$$

ποσοστό τους επί του συνόλου είναι, άρα, 3%. Παρόμοια, για μεγαλύτερα δείγματα βρίσκουμε ότι οι συνδυασμοί με ακραίες αναλογίες γίνονται ακόμη πιο σπάνιοι.

λέγεται **παράγοντας κανονικοποίησης**. Η κανονικοποίηση της κατανομής έχει μεγάλη σημασία γιατί μας επιτρέπει να συγκρίνουμε διαφορετικές κατανομές συχνοτήτων από διαφορετικά, ακόμη και διαφορετικής φύσης πειράματα και να διαπιστώσουμε ότι συμπεριφέρονται με τον ίδιο τρόπο. Έτσι, ομοιόμορφη κατανομή για ένα πείραμα με N ενδεχόμενα είναι αυτή για την οποία τα ποσοστά εμφάνισης υπακούουν στην παρακάτω έκφραση:

$p_i \rightarrow \text{σταθερά}, i=1, 2, \dots, N$ και η κανονικοποιημένη μορφή της είναι:

$$p_i \rightarrow \text{σταθερά} = \frac{1}{N}, i=1, 2, \dots, N$$

Αυτές οι εκφράσεις δεν εξαρτώνται από το μέγεθος του δείγματος ή άλλες λεπτομέρειες του πειράματος και εκφράζουν με γενικό τρόπο αυτό που είναι πιο ουσιώδες: ότι κάθε ενδεχόμενο έχει ίση ευκαιρία εμφάνισης.

Ως εδώ μιλήσαμε για ομοιόμορφες κατανομές διακριτών ενδεχομένων. Όπως έχουμε ήδη πει, τα πειραματικά αποτελέσματά μας μπορεί να είναι συνεχή αριθμητικά δεδομένα. Τότε, το άθροισμα

στον παράγοντα κανονικοποίησης $\frac{1}{\sum_{i=1}^N f_i}$, αντικαθίσταται από ένα ολοκλήρωμα:

$$\frac{1}{\int_a^b f(x) dx}, \text{ όπου } a, b \text{ τα όρια μεταξύ των οποίων κυμαίνονται οι τιμές της μεταβλητής } X. \text{ Από}$$

τους μέχρι τώρα ορισμούς, προκύπτει ότι το άθροισμα ή ολοκλήρωμα των ποσοστών εμφάνισης είναι μονάδα: $\sum_{i=1}^N p_i = 1$ για διακριτά δεδομένα και $\int_a^b p(x) dx = 1$ για συνεχή δεδομένα. Δηλαδή, η κανονικοποίηση μετατρέπει τις συχνότητες σε ποσοστά από το διάστημα $[0, 1]$.

Επειδή αναφερθήκαμε σε δείγματα, χρησιμοποιήσαμε το σύμβολο των ορίων, $\rightarrow$, αντί για την ισότητα στους παραπάνω ορισμούς. Αν έχουμε ένα πληθυσμό για τον οποίο γνωρίζουμε ότι περιγράφεται από την ομοιόμορφη κατανομή, μπορούμε να χρησιμοποιήσουμε το σύμβολο της ισότητας, $=$, στους ορισμούς. Τότε, π.χ. για συνεχή δεδομένα και αντίστοιχη κατανομή, θα ισχύει ότι $f(x) = c$ (κάποια σταθερά) και $p(x) = 1/(b-a)$, $a < x < b$. Για το πέρασμα από το διακριτό στο συνεχές όριο, ας θυμηθούμε τις γενικές μαθηματικές αρχές για τον ορισμό και τη σημασία του ολοκληρώματος. Στο συνεχές όριο δε μιλάμε για ποσοστό εμφάνισης μιας τιμής αλλά για το ποσοστό εμφάνισης τιμών από ένα διάστημα $(x, x+\Delta x]$ το οποίο θα είναι ίσο με το εμβαδό κάτω από την καμπύλη $p(x)$ για αυτό το διάστημα. Για απειροστό πλάτος dx αυτού, το ποσοστό εμφάνισης θα είναι $p(x)dx$.

Η παραπάνω συζήτηση μας επιτρέπει να εισάγουμε και να χρησιμοποιούμε την έννοια της **πιθανότητας**. Η πιθανότητα είναι ουσιαστικά, ένα ποσοτικό μέτρο αυτού που ονομάσαμε, κάπως χαλαρά μέχρι εδώ, "ευκαιρίες εμφάνισης" ενός ενδεχομένου. Έχοντας ορίσει τα ποσοστά εμφάνισης των ενδεχομένων απέχουμε ένα βήμα από τον ορισμό της πιθανότητας που δε θα ήταν άλλος από αυτά τα ποσοστά αν ο πληθυσμός ήταν γνωστός. Η δε κατανομή των ποσοστών του πληθυσμού θα ήταν η κατανομή πιθανότητας (για την τυχαία μεταβλητή που μας ενδιαφέρει) όσον αφορά το συγκεκριμένο πληθυσμό. Για να ξεχωρίζουμε την πιθανότητα από το ποσοστό εμφάνισης που μπορεί να αναφέρεται σε διάφορα δείγματα, υποσύνολα του πληθυσμού, χρησιμοποιούμε κεφαλαίο γράμμα για αυτή και την αντίστοιχη μεταβλητή, $P(X)$ και μικρά γράμματα για το ποσοστό εμφάνισης κάθε τιμής της μεταβλητής: $p(x)$.

Συνήθως, ο πληθυσμός δεν είναι και ούτε καν μπορεί να γίνει γνωστός στο σύνολό του. Επομένως, τόσο η πιθανότητα όσο και η κατανομή αυτής μας είναι κατ' αρχήν άγνωστες. Τι μπορούμε να κάνουμε τότε; Βασικά, υπάρχουν δύο τρόποι για να γνωρίσουμε την κατανομή πιθανότητας που περιγράφει κάποιο πληθυσμό:

- Υπολογισμός με βάση εύλογα επιχειρήματα που αναφέρονται στη φύση των

παρατηρήσεών μας και των δυνατών ενδεχομένων, καθώς και του τρόπου διεξαγωγής του πειράματος. Αυτή είναι η *εκ των προτέρων* ή *a priori* πιθανότητα.

- Εκτίμηση με βάση επαρκή, αντιπροσωπευτικά δείγματα που έχουμε επιλέξει (*εκ των υστέρων* ή *a posteriori* πιθανότητα).

Πριν προχωρήσουμε σε συγκεκριμένα παραδείγματα, πρέπει να διευκρινίσουμε ορισμένα σημεία:

- Η κατανομή πιθανότητας μπορεί να καθορίζεται από το συνολικό πληθυσμό, αλλά η πιθανότητα $P(X=x)$ ενός συγκεκριμένου ενδεχομένου $X = x$, αναφέρεται σε ένα σύνολο παρατηρήσεων ή και μία και μόνη παρατήρηση. Για να το έχουμε πιο σαφές αυτό, μπορούμε να “μεταφράσουμε” ως εξής: πιθανότητα εμφάνισης του τάδε ενδεχομένου σε μια παρατήρηση (ή N παρατηρήσεις) είναι το ποσοστό των ευκαιριών που δίνονται σε αυτό το ενδεχόμενο να εμφανιστεί κατά την παρατήρηση (ή τις παρατηρήσεις).

- Η πιθανότητα εμφάνισης ενός ενδεχομένου δε συνεπάγεται “δεσμευτική” ή “υποχρεωτική” εμφάνιση αυτού σε προκαθορισμένες αναλογίες. Π.χ. κατά τη ρίψη νομίσματος δύο φορές, δεν είναι υποχρεωτικό να έρθει κορώνα τη μία. Μπορεί να έρθει και τις δύο ή καμία. Είναι μετά από πολλές επαναλήψεις ρίψεων δύο φορές που θα φανεί ότι η αναλογία των εμφανίσεων ακολουθεί το “νόμο” των *a priori* πιθανοτήτων. Αν μας επιτρέπεται η έκφραση, η ευκαιρία που ορίζουν οι πιθανότητες υπάρχει, αλλά είναι δυνατό να χαθεί προσωρινά για ένα ενδεχόμενο και υπέρ ενός άλλου.

- Οι πιθανότητες εκφράζουν το βαθμό της δικής μας γνώσης ή άγνοιας για το υπό μελέτη φαινόμενο, σύνολο, σύστημα κλπ. Π.χ. αν έχουμε ένα σάκκο με 50 λευκά και 50 μαύρα σφαιρίδια, και βγάλουμε ένα, τότε, πριν ανοίξουμε το χέρι μας για να το δούμε, θα πούμε πολύ σωστά, ότι η πιθανότητα να έχουμε επιλέξει μαύρο σφαιρίδιο είναι 50%. Όταν όμως ανοίξουμε το χέρι μας, γνωρίζουμε πλέον το χρώμα του και δεν έχει νόημα η οποιαδήποτε αναφορά στην πιθανότητα με βάση τη γνωστή κατανομή του πληθυσμού. Πριν ανοίξουμε το χέρι μας γνωρίζαμε την κατανομή του πληθυσμού και από αυτή προσπαθούσαμε να συμπεράνουμε το χρώμα της σφαίρας – αυτό είναι το αντίστροφο από την προσπάθεια να γνωρίσουμε την κατανομή του πληθυσμού από ένα μικρότερο δείγμα. Όταν ανοίξαμε το χέρι, δε μας χρειαζόταν πλέον η κατανομή και η αναφορά σε πιθανότητες.

- Όταν η θεωρούμενη τυχαία μεταβλητή X , είναι συνεχής, τότε δεν έχει νόημα να αναφερόμαστε σε πιθανότητα μιας συγκεκριμένης τιμής x . Όπως είπαμε προηγουμένως σε σχέση με τα ποσοστά εμφάνισης, μας ενδιαφέρει το εμβαδό κάτω από την καμπύλη $p(x)$ για ένα διάστημα $(x, x+\Delta x]$. Αν περιοριστούμε σε μια τιμή x , είναι σα να θέτουμε $\Delta x = 0$, οπότε το διάστημα και άρα το αντίστοιχο εμβαδόν, είναι μηδέν. Αυτό που έχει νόημα είναι να αντιστοιχίσουμε σε μια τιμή x το εμβαδόν $p(x)dx$, αναφερόμενοι δηλαδή, στο απειροστό διάστημα $(x, x+dx]$. Σε μια τέτοια περίπτωση (συνεχούς μεταβλητής και αντίστοιχης κατανομής) λέμε ότι η $p(x)$ παριστάνει την **πυκνότητα πιθανότητας** στο σημείο x , ενώ η πιθανότητα μεταξύ δύο τιμών x και $x+\Delta x$, δίνεται από το ολοκλήρωμα της $p(x)$ μεταξύ αυτών των ορίων.

- Όπως και τα ποσοστά στα οποία αναφερθήκαμε πιο πάνω, οι κατανομές πιθανότητας δίνονται σε κανονικοποιημένη μορφή, έτσι ώστε να ισχύει $\sum_x p(x) = 1$ (διακριτά δεδομένα) ή $\int_x p(x)dx = 1$ (συνεχή δεδομένα). Στη δεύτερη περίπτωση, το ολοκλήρωμα κανονικά είναι ορισμένο και τα όρια μπορεί να εκτείνονται μέχρι το $\pm \infty$, ανάλογα και με τη φύση του εκάστοτε προβλήματος.

Παράδειγμα υπολογισμού εκ των προτέρων πιθανότητας, είναι και πάλι η ρίψη νομίσματος: αφού κανένα από τα δύο δυνατά ενδεχόμενα δεν υπάρχει λόγος να ευνοείται έναντι του άλλου, θα ισχύει $P(X=\kappa) = P(X=\gamma) = 0.5$, ή 50%.

Άλλο παράδειγμα με την ίδια λογική είναι το εξής: ρίχνουμε δύο ζάρια και θέλουμε να βρούμε την πιθανότητα εμφάνισης σε κάθε ζαριά, μιας τιμής $x = a + b$, όπου a και b , η τιμή που δίνει κάθε ζάρι χωριστά. Σε αυτό το πρόβλημα θα μας διευκολύνει ο παρακάτω πίνακας. Στην πρώτη στήλη και πρώτη γραμμή του πίνακα έχουμε τις τιμές a και b και στα υπόλοιπα κελιά, τα αντίστοιχα αθροίσματα $x = a + b$.

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Τώρα, η ζαριά είναι το ζεύγος (a, b) και έχουμε 36 τέτοια ζεύγη που έχουν όλα ίσες ευκαιρίες να εμφανιστούν σε κάθε ρίψη. Επομένως, η πιθανότητα εμφάνισης κάθε τέτοιου ζεύγους σε μία και μόνη ζαριά θα είναι $1/36$. Αλλά η τιμή x κυμαίνεται από 2 μέχρι 12, δηλαδή έχει μόνο 11 δυνατές τιμές. Επιπλέον, από τον πίνακα βλέπουμε ότι η κάθε τιμή εμφανίζεται με διαφορετική συχνότητα που μπορούμε να βρούμε μετρώντας κατά μήκος των διαγωνίων από κάτω αριστερά προς τα πάνω δεξιά και να κάνουμε έναν *πίνακα συχνοτήτων* όπου ουσιαστικά, μετράμε τις ευκαιρίες εμφάνισης κάθε τιμής:

Τιμή, x_i	Πλήθος συνδυασμών με $X=x_i, f_i$	A priori πιθανότητα, $p_i = f_i/\Sigma f_i$	Τιμή, x_i	Πλήθος συνδυασμών με $X=x_i, f_i$	A priori πιθανότητα, $p_i = f_i/\Sigma f_i$
2	1	0.028	8	5	0.139
3	2	0.056	9	4	0.111
4	3	0.083	10	3	0.083
5	4	0.111	11	2	0.056
6	5	0.139	12	1	0.028
7	6	0.167	Άθροισμα, Σf_i	36	1.001

Το άθροισμα των συχνοτήτων είναι φυσικά και πάλι 36, αλλά βλέπουμε ότι οι “μεσαίες” τιμές εμφανίζονται συχνότερα από τις ακραίες. Καταλαβαίνουμε ότι αν κάθε συνδυασμός (a, b) έχει 1 στις 36 πιθανότητες να παρατηρηθεί σε μια και μόνη ζαριά, τότε, κάθε τιμή $x = a+b$ έχει $f(x)$ στις 36 πιθανότητες. Έτσι, οι τιμές 2 και 12 (άσσοι και εξάρες) είναι οι πιο απίθανοι με $P(X) = 1/36$, ενώ η τιμή 7 είναι η πιο πιθανή ($P = 6/36 = 1/6$) αφού προκύπτει από τους πιο πολλούς συνδυασμούς. Σημειώνουμε ότι το άθροισμα των πιθανοτήτων είναι εξ ορισμού μονάδα αλλά στον παραπάνω πίνακα υπολογίζεται ίσο με 1.001 λόγω σφαλμάτων στρογγύλευσης στον

υπολογισμό των επιμέρους πιθανοτήτων.

Μερικοί ορισμοί και κανόνες

Στο σημείο αυτό μπορούμε να δώσουμε μερικούς πιο αυστηρούς ορισμούς μερικών βασικών εννοιών που χρησιμοποιήσαμε “χαλαρά” μέχρι εδώ. Κάναμε λόγο για παρατηρήσεις και πειράματα. Ορίζουμε ως **πείραμα τύχης** εκείνο το πείραμα του οποίου το αποτέλεσμα δεν είναι δυνατό να προβλέψουμε. Θα αναφερόμαστε σε αυτό για συντομία, απλά ως πείραμα. Ο,τι προκύπτει από μια μοναδική εκτέλεση του πειράματος είναι το **αποτέλεσμα** ή **παρατήρηση** και το σύνολο των δυνατών αποτελεσμάτων είναι ο **δειγματικός χώρος** (συνήθως συμβολίζεται με S) του πειράματος. **Ενδεχόμενο** είναι κάθε υποσύνολο του δειγματικού χώρου. Π.χ. αν ρίχνουμε ένα ζάρι, μπορούμε να θεωρήσουμε το υποσύνολο των αποτελεσμάτων $\{1, 2, 5\}$ ως ένα δυνατό ενδεχόμενο, δεν είναι δηλαδή μόνο τα μεμονωμένα αποτελέσματα που θεωρούμε ως ενδεχόμενα αλλά και σύνολα περισσότερων του ενός δυνατών αποτελεσμάτων. Ακόμη και το κενό σύνολο ορίζεται ως το **αδύνατο ενδεχόμενο**.

Ένα ενδεχόμενο πραγματοποιείται όταν τουλάχιστον ένα από τα αποτελέσματα που περιλαμβάνει, έχει συμβεί.

Ορίζοντας τα ενδεχόμενα ως σύνολα, μπορούμε να κάνουμε πράξεις συνόλων μεταξύ δύο ενδεχομένων, όπως η ένωση $A \cup B$ (όλα τα αποτελέσματα που ανήκουν είτε στο A είτε στο B), η τομή $A \cap B$ (μόνο τα αποτελέσματα που ανήκουν και στο A και στο B) και η διαφορά $A - B$ (τα αποτελέσματα που ανήκουν στο A αλλά όχι στο B). **Ασυμβίβαστα** είναι τα ενδεχόμενα που η τομή τους είναι το κενό σύνολο. Αυτό πρακτικά σημαίνει ότι θα πραγματοποιηθεί ή το ένα ή το άλλο (ή κανένα), αλλά ποτέ και τα δύο. Τα *μεμονωμένα* αποτελέσματα ενός πειράματος είναι, φυσικά, ασυμβίβαστα (π.χ. κορώνα ή γράμματα, το αποτέλεσμα ρίψης ενός ζαριού κλπ).

Στη συνέχεια μπορούμε να ορίσουμε την πιθανότητα και να διατυπώσουμε ορισμένους κανόνες για το μαθηματικό χειρισμό των πιθανοτήτων διαφόρων ενδεχομένων. Έστω ο δειγματικός χώρος S . Για κάθε ενδεχόμενο $A \subset S$, ορίζουμε τον αριθμό $P(A)$, για τον οποίο ισχύουν τα εξής:

- $0 \leq P(A) \leq 1$
- $P(S) = 1$
- Για ασυμβίβαστα ενδεχόμενα A, B : $P(A \cup B) = P(A) + P(B)$

Τότε, μπορούμε να δώσουμε δύο ορισμούς για την πιθανότητα:

Ορισμός a priori πιθανότητας (κατά Laplace). Αν για τα αποτελέσματα ενός πειράματος δεν έχουμε λόγο να υποθέσουμε ότι το ένα θα υπερτερεί, δηλ. θα έχει περισσότερες ευκαιρίες να εμφανιστεί, από τα άλλα, τότε για κάθε ενδεχόμενο A , η πιθανότητα $P(A)$ δίνεται διαιρώντας τον αριθμό των αποτελεσμάτων του δια τον αριθμό όλων των αποτελεσμάτων του S : $P(A) = N(A)/N(S)$

Ορισμός a posteriori πιθανότητας (κατά von Mises): αν δε γνωρίζουμε ούτε μπορούμε να υποθέσουμε ότι τα αποτελέσματα έχουν ίσες ευκαιρίες εμφάνισης, εκτελούμε το πείραμα n φορές όπου n πολύ μεγάλος αριθμός, καταγράφουμε τις συχνότητες εμφάνισης των αποτελεσμάτων και ορίζουμε βάσει αυτών την πιθανότητα κάθε ενδεχομένου, με τρόπο ανάλογο με αυτό που χρησιμοποιήσαμε τον αριθμό των δυνατών αποτελεσμάτων στον προηγούμενο ορισμό.

Με τα παραπάνω αξιώματα και ορισμούς, μπορούμε να διατυπώσουμε και αποδείξουμε ορισμένους σημαντικούς κανόνες. Όταν δύο ενδεχόμενα είναι **ανεξάρτητα**, δηλαδή η πραγματοποίηση του ενός δεν εξαρτάται από την πραγματοποίηση του άλλου, ισχύει ο **ειδικός νόμος του πολλαπλασιασμού**:

$$P(A \cap B) = P(A) P(B).$$

Αυτή η σχέση χρησιμεύει και ως ορισμός της ανεξαρτησίας των ενδεχομένων: δύο ενδεχόμενα A, B , είναι ανεξάρτητα το ένα από το άλλο αν και μόνο αν για αυτά ισχύει ο ειδικός νόμος του πολλαπλασιασμού. Στην πραγματικότητα, αυτή η σχέση αποτελεί ειδική περίπτωση του αντίστοιχου γενικού νόμου του πολλαπλασιασμού. Για να γίνει κατανοητός αυτός, πρέπει να εισάγουμε την έννοια της **πιθανότητας υπό συνθήκη** ή **δεσμευμένης πιθανότητας**. Αυτή είναι η πιθανότητα για ένα ενδεχόμενο όταν θεωρούμε ότι πραγματοποιείται ένα άλλο "ευρύτερο" ενδεχόμενο, π.χ. όταν ένα ζάρι φέρει άρτιο αριθμό, ποια η πιθανότητα αυτός να είναι ίσος με 2; Εδώ, το ενδεχόμενο B άρτιου αριθμού $\{2, 4, 6\}$ παίζει το ρόλο ενός νέου δειγματικού χώρου, υποσυνόλου του αρχικού $S = \{1, 2, 3, 4, 5, 6\}$ και η πιθανότητα για το ενδεχόμενο $A = \{2\}$ πρέπει να υπολογιστεί με βάση αυτόν. Θα τη συμβολίσουμε με $P(A | B)$ και θα είναι ίση με $P(A | B) = 1/3$, προφανώς. Επειδή ένα ενδεχόμενο πραγματοποιείται όταν παρατηρείται ένα οποιοδήποτε από τα αποτελέσματά του, μπορούμε να επεκτείνουμε αυτή την έννοια και στην περίπτωση όπου το B δεν είναι υπερσύνολο του A αλλά απλώς ένα άλλο ενδεχόμενο. Π.χ. ποια είναι η πιθανότητα του $A = \{1, 2\}$ όταν πραγματοποιείται το $B = \{2, 4, 6\}$.

Ιδιαίτερη σημασία έχουν τα παραπάνω όταν κάνουμε πειράματα *λήψης χωρίς επαναποτοθέτηση*, για παράδειγμα: Έστω ότι σε ένα κουτί υπάρχουν 8 λευκές και 12 μαύρες σφαίρες. Παίρνουμε ένα σφαιρίδιο και από αυτά που μένουν, παίρνουμε άλλο ένα. Ποια η πιθανότητα το πρώτο να είναι μαύρο και το δεύτερο λευκό;

Για το πρώτο ερώτημα: η πιθανότητα να είναι το πρώτο μαύρο είναι προφανώς $P(1="Μ") = 12/20$. Η πιθανότητα για το δεύτερο να είναι λευκό είναι η *υπό συνθήκη πιθανότητα* το δεύτερο να είναι λευκό *όταν το πρώτο ήταν μαύρο*. Αφού το πρώτο σφαιρίδιο δεν επαναποτοθετήθηκε, έχουμε πλέον 8 λευκές και 11 μαύρες σφαίρες, άρα $P(2="Λ" | 1="Μ") = 8/19$. Τώρα, για να υπολογίσουμε την πιθανότητα $P(1="Μ" \cap 2="Λ")$ πρέπει να σκεφτούμε ως εξής: η παρατήρησή μας είναι το χρώμα και των δύο σφαιριδίων, δηλαδή το ζεύγος $X = (\text{Χρώμα } 1, \text{Χρώμα } 2)$. Επομένως, ο δειγματικός χώρος είναι το σύνολο που περιέχει όλα τα δυνατά ζεύγη πρώτων και δεύτερων σφαιριδίων, *λαμβανομένων χωρίς επαναποτοθέτηση*. Καταλαβαίνουμε ότι ο αριθμός των συνολικών δυνατών ζευγών είναι $20 \times (20 - 1) = 380$, ενώ ο συγκεκριμένος συνδυασμός $(1="Μ", 2="Λ")$ μπορεί να ληφθεί $12 \times 8 = 96$ φορές, γιατί στην αρχή υπάρχουν και τα 12 μαύρα σφαιρίδια ενώ μετά εξακολουθούν να υπάρχουν και τα 8 λευκά. Τότε, βρίσκουμε ότι $P(1="Μ", 2="Λ") = 96/380$. Αλλά αυτό είναι ακριβώς ίσο με $P(1="Μ") P(2="Λ" | 1="Μ")$. Αυτό δεν είναι συμπτωματικό και μπορεί να επαληθευθεί με ανάλογα συνδυαστικά επιχειρήματα και να αποδειχθεί στη γενικότητά του.

Με βάση τα παραπάνω, μπορεί να διατυπωθεί ο **γενικός νόμος του πολλαπλασιασμού** για τα ενδεχόμενα A και B *ισχύοντος του A* , ως εξής:

$$P(A \cap B) = P(A) P(B | A), \text{ αλλά και}$$

$$P(A \cap B) = P(B) P(A | B)$$

Αν ως B πάρουμε το συνολικό δειγματικό χώρο S , τότε προφανώς ισχύει $P(A) = P(A | S)$ και $P(B) = P(B | S)$ (είναι σα να λέμε "η πιθανότητα να συμβεί το A υπό την προϋπόθεση να συμβεί οτιδήποτε"). Επομένως, αν δεν υπάρχει η προϋπόθεση για το B να έχει πραγματοποιηθεί το A , αυτό σημαίνει ότι μπορούμε να γράψουμε $P(A \cap B) = P(B) P(A | S) = P(A \cap B) = P(B) P(A)$ που είναι ο ειδικός νόμος του πολλαπλασιασμού τον οποίο προαναφέραμε.

Εφαρμογή του γενικού νόμου του πολλαπλασιασμού αποτελεί ο **νόμος** ή **κανόνας** ή **θεώρημα του Bayes**. Στο παρακάτω σχήμα παριστάνεται ο δειγματικός χώρος S που υποδιαιρείται στα ασυμβίβαστα ενδεχόμενα $A_1, A_2, \dots, A_n$. Θεωρούμε ένα ενδεχόμενο B (σκιασμένη περιοχή)

		B	
A ₁	A ₂	...	A _n

και από το σχήμα καταλαβαίνουμε ότι $B = (A_1 \cap B) \cup (A_2 \cap B) \dots (A_n \cap B)$ και επειδή τα ενδεχόμενα $(A_i \cap B)$ είναι ασυμβίβαστα, θα ισχύει

$$P(B) = \sum_{i=1}^n P(A_i \cap B) = \sum_{i=1}^n P(A_i) P(B|A_i)$$

όπου η τελευταία ισότητα προκύπτει από το γενικό νόμο του πολλαπλασιασμού και μας δίνει τον **τύπο ολικής πιθανότητας**. Με τη βοήθειά της μπορούμε να βρούμε και την υπό συνθήκη πιθανότητα $P(A_i | B)$ η οποία βρίσκεται αν τη γράψουμε με βάση το γενικό νόμο πολλαπλασιασμού και εκφράσουμε το $P(B)$ με τον τύπο ολικής πιθανότητας:

$$P(A_i \cap B) = P(B) P(A_i | B) \Rightarrow P(A_i | B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i) P(B|A_i)}{P(B)} = \frac{P(A_i) P(B|A_i)}{\sum_{i=1}^n P(A_i) P(B|A_i)}$$

Η τελευταία έκφραση είναι ο κανόνας ή θεώρημα του Bayes. Παρατηρείστε, ως μνημονικό κανόνα, ότι τα $P(B | A_i)$ είναι "συντελεστές βαρύτητας" των $P(A_i)$.

Παράδειγμα: Δυο μηχανές σε εργοστάσιο παράγουν το 10% και 90% της παραγωγής συγκεκριμένων αντικειμένων. Η πιθανότητα να παράγει η πρώτη μηχανή ελαττωματικό αντικείμενο είναι 0.01 και για τη δεύτερη είναι 0.05. Ποια η πιθανότητα ένα ελαττωματικό αντικείμενο να έχει κατασκευαστεί από την πρώτη μηχανή;

Αν B το ενδεχόμενο να είναι το αντικείμενο ελαττωματικό και A_1, A_2 τα ενδεχόμενα να έχει παρασκευαστεί από την πρώτη και τη δεύτερη μηχανή αντίστοιχα, τότε επειδή δίνεται ότι το αντικείμενο είναι ελαττωματικό, εφαρμόζουμε τον τύπο του Bayes και βρίσκουμε:

$$P(A_1|B) = \frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2)} = \frac{0.10 \times 0.01}{0.10 \times 0.01 + 0.90 \times 0.05} = \frac{1}{46}$$

$$P(A_2|B) = \frac{0.90 \times 0.05}{0.10 \times 0.01 + 0.90 \times 0.05} = \frac{45}{46}$$

Τέλος, ας ξαναδούμε την έννοια της τυχαίας μεταβλητής. Μια **τυχαία μεταβλητή X** , όπως έχουμε ήδη αναφέρει, απεικονίζει τα δυνατά αποτελέσματα ενός πειράματος τύχης στο σύνολο των πραγματικών αριθμών. **Κατανομή $f(x)$** της μεταβλητής είναι μια συνάρτηση που απεικονίζει τις τιμές x της μεταβλητής στις πιθανότητες $P(x)$ αυτών. Για συνεχείς μεταβλητές θα ισχύει

$$P(a < x < b) = \int_a^b f(x) dx$$

Αθροιστική ή σωρευτική κατανομή, $F(x)$ ονομάζεται η συνάρτηση που ορίζεται ως

$$F(x) = P(-\infty < X \leq x) = \sum_{u \leq x} f(u) \quad (\text{διακριτή μεταβλητή})$$

$$F(x) = P(-\infty < X \leq x) = \int_{-\infty}^{+\infty} f(u) du \quad (\text{συνεχής μεταβλητή})$$

Από τους παραπάνω ορισμούς, αν X είναι τυχαία μεταβλητή και $a < b$ είναι πραγματικοί αριθμοί, προκύπτουν τα παρακάτω:

$$P(a < x \leq b) = F(b) - F(a)$$

$P(X = a) = 0$ αν η $F(x)$ είναι συνεχής στο σημείο $x = a$ και $P(x=a) =$ μέγεθος άλματος της F στο $x = a$ όταν η F είναι ασυνεχής στο σημείο αυτό. Στη γενική περίπτωση ισχύει:

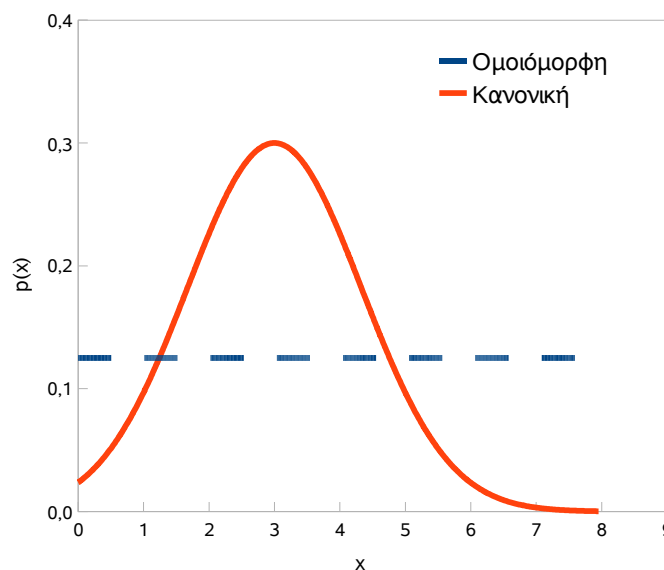
$$P(X < a) = F(a) - P(X=a)$$

Τέλος, αν η X είναι συνεχής μεταβλητή, θα ισχύει:

$$dF(x)/dx = f(x)$$

Εκτός από την ομοιόμορφη, μια άλλη, πολύ σημαντική κατανομή στην οποία θα αναφερθούμε εκτενέστερα στη συνέχεια, είναι η λεγόμενη κανονική κατανομή. Αν παραστήσουμε αυτή και την ομοιόμορφη κατανομή για μια συνεχή μεταβλητή σε ένα διάγραμμα, θα πάρουμε κάτι σαν το σχήμα της επόμενης σελίδας. Δηλαδή, ενώ η ομοιόμορφη κατανομή ισοδυναμεί με ίσες πιθανότητες για όλες τις δυνατές τιμές, η κανονική κατανομή προβλέπει μεγαλύτερη πιθανότητα για τιμές που βρίσκονται κοντά σε μία “μέση τιμή” (εδώ, ίση με 3) όπου και η $p(x)$ μεγιστοποιείται, ενώ όσο απομακρυνόμαστε από αυτή την τιμή, οι πιθανότητες μικραίνουν και τείνουν ασυμπτωτικά στο μηδέν.

Αν και θα αναφερθούμε εκτενέστερα στη συνέχεια, θα πούμε από τώρα ότι τόσο η ομοιόμορφη όσο και η κανονική κατανομή μπορούν να περιγραφούν από σχετικά απλές μαθηματικές εκφράσεις με λίγες χαρακτηριστικές παραμέτρους που σχετίζονται με τα ιδιαίτερα χαρακτηριστικά τους. Για την κανονική κατανομή, μια τέτοια παράμετρος είναι η μέση τιμή, ενώ σημαντική είναι επίσης και η διασπορά, η οποία σχετίζεται με το πόσο πλατειά ή στενή είναι η εικονιζόμενη “κωδονοειδής” καμπύλη. Το ίδιο ισχύει και για άλλες κατανομές που είναι κατάλληλες για την περιγραφή πληθυσμών διαφόρων ειδών.



Η κατανομή ενός πληθυσμού είναι ο άξονας γύρω από τον οποίο περιστρέφεται η Στατιστική. Μερικά από τα κυριώτερα προβλήματα της Στατιστικής είναι:

- Από τη μελέτη αντιπροσωπευτικών δειγμάτων ενός πληθυσμού να συμπεράνουμε το είδος και τη μορφή της κατανομής του,
- Αντίστροφα από το προηγούμενο, αν ξέρουμε ή έχουμε εκτιμήσει την κατανομή ενός πληθυσμού, να προβλέψουμε το πιθανό αποτέλεσμα μιας παρατήρησης ή ενός πειράματος ή τα χαρακτηριστικά ενός δείγματος.
- Παρόμοια, να προβλέψουμε την πιθανότητα εμφάνισης ακραίων τιμών,
- Να εκτιμήσουμε τη βεβαιότητα με την οποία κάποια αποτελέσματα προκύπτουν από τη

δεδομένη κατανομή του πληθυσμού και όχι κατά τύχη (ομοιόμορφη κατανομή) ή από άλλη κατανομή – βλ. έλεγχος υποθέσεων, όπου θα αναφερθούμε σε επόμενες ενότητες.

Σε όλες τις περιπτώσεις γνωρίζουμε ή θέλουμε να μάθουμε την κατανομή του πληθυσμού. Αν γνωρίζουμε τη μορφή της (π.χ. κανονική κατανομή) μπορούμε να την περιγράψουμε επαρκώς με κατάλληλες **παραμέτρους**, μερικές από τις οποίες υπεισέρχονται και στην αναλυτική μαθηματική έκφραση της κατανομής, όπως αναφέραμε. Ποσότητες ανάλογες με τις παραμέτρους του συνολικού πληθυσμού μπορούμε να ορίσουμε και για τα επιμέρους δείγματα. Αυτές οι ποσότητες λέγονται **στατιστικά** και οι τιμές τους εν γένει διαφέρουν από αυτές των πληθυσμιακών παραμέτρων, αλλά στο βαθμό που τα δείγματα είναι αντιπροσωπευτικά και το μέγεθός τους αυξάνει, τα στατιστικά τείνουν να ταυτιστούν με τις παραμέτρους – όπως και οι αντίστοιχες κατανομές. Η μέση τιμή όσον αφορά τον πληθυσμό και ο αριθμητικός μέσος όσον αφορά τα δείγματα είναι η πιο γνωστή περίπτωση παραμέτρου και στατιστικού αντίστοιχα και δίνει μια βασική ιδέα για την κεντρική τάση των δεδομένων μας. Άλλη γνωστή παράμετρος/στατιστικό είναι η τυπική απόκλιση που δίνει μια ιδέα για το πόσο διεσπαρμένα είναι τα δεδομένα γύρω από τη μέση τιμή. Αυτά και άλλα στατιστικά και παράμετροι θα συζητηθούν αναλυτικότερα στην επόμενη ενότητα.