Ενότητα 1

Παράμετροι και Στατιστικά. Διωνυμική και Κανονική Κατανομή.

Βασικές Έννοιες

Ένας βασικός σκοπός της στατιστικής είναι η *σύνοψη* των δεδομένων που προέρχονται από παρατηρήσεις, ώστε να δούμε τις κυριότερες τάσεις αυτών και να *γενικεύσουμε* τα αποτελέσματα των παρατηρήσεών μας.

Τρεις βασικοί τρόποι για να το κάνουμε αυτό είναι:

- Συγκέντρωση δεδομένων σε πίνακες
- Οπτική αναπαράσταση της συμπεριφοράς των δεδομένων σε ένα **διάγραμμα κατανομής συχνοτήτων**
- Υπολογισμός κάποιων κατάλληλων μέτρων ή χαρακτηριστικών αριθμητικών τιμών που περιγράφουν:
    - τις κύριες τάσεις των δεδομένων μας (μέση τιμή, διάμεσος, κορυφή ή επικρατούσα τιμή)
    - τον τρόπο με τον οποίο και την έκταση στην οποία μεταβάλλονται τα δεδομένα (εύρος, τυπική απόκλιση)

'Εχουμε ήδη πει ότι τα ανεπεξέργαστα δεδομένα, δηλαδή το σύνολο των αποτελεσμάτων κάθε παρατήρησης που έχουμε διενεργήσει για ορισμένο σκοπό, μπορεί να είναι αριθμητικές τιμές (π.χ. ο χρόνος μετατροπής ενός συστατικού κατά ορισμένο ποσοστό σε ένα χημικό αντιδραστήρα) ή μη αριθμητικά (η προτίμηση κάθε ψηφοφόρου όπως αποτυπώνεται στα ψηφοδέλτια που βρίσκουμε με το άνοιγμα της κάλπης). Τα αριθμητικά δεδομένα μπορεί να είναι συνεχή (χρόνος μετατροπής που προαναφέρθηκε) ή διακριτά.

Σε κάθε περίπτωση, τα σύνολο των δεδομένων χαρακτηρίζεται από μία κατανομή δηλαδή έναν τρόπο με τον οποίο αυτά καλύπτουν τα διαστήματα των δυνατών τιμών τους. Ένας άλλος τρόπος που έχουμε ήδη εισάγει για να εκφράσουμε την έννοια της κατανομής έγκειται στην παρουσίαση του τρόπου με τον οποίο μεταβάλλεται η συχνότητα f, εμφάνισης συγκεκριμένων αποτελεσμάτων ή αριθμητικών τιμών x, ως συνάρτηση αυτών ακριβώς των τιμών, f(x). Π.χ. αν τα δεδομένα μας είναι τα βάρη διαφόρων ανθρώπων, τότε η κατανομή των βαρών μπορεί να γίνει αντιληπτή αν περιγράψουμε με κάποιο τρόπο, πόσο συχνά εμφανίζεται το βάρος των 70 κιλών, των 71, 72 κλπ.

Τα δεδομένα δίνονται ως ένα απαριθμήσιμο ή διακριτό σύνολο παρατηρήσεων, αριθμητικών ή μη, δηλαδή ως ένας πεπερασμένος αριθμός στοιχείων. Για να κάνουμε το διάγραμμα συχνοτήτων πρέπει να χωρίσουμε τα δεδομένα σε κατηγορίες. Αν τα δεδομένα είναι μη αριθμητικά, π.χ. το χρώμα ματιών από ένα δείγμα του πληθυσμού, τότε η κατηγοριοποίηση των δεδομένων είναι συνήθως προφανής.

Αν τα δεδομένα είναι αριθμητικά, τότε χωρίζουμε το διάστημα που περιλαμβάνει τις τιμές σε έναν αριθμό από ισομήκη, συνήθως, διαστήματα $(L_i, H_i]$, i=1,, 2, ... n, $b=H_1-L_1=H_2-L_2=...=H_n-L_n$

αρκετά μεγάλα ώστε να μη φαίνονται στο διάγραμμα οι ασήμαντες διακυμάνσεις των δεδομένων που συσκοτίζουν τα ουσιώδη χαρακτηριστικά της κατανομής και αρκετά μικρά ώστε να μη χάνονται αυτά τα χαρακτηριστικά. Μια τυπική εκλογή είναι αυτή των 10 ως 20 διαστημάτων, χωρίς να αποκλείονται περισσότερα ή λιγότερα, ανάλογα με τα ιδιαίτερα χαρακτηριστικά του συνόλου των δεδομένων μας. Συνήθως η χαμηλότερη τιμή, L_1, λαμβάνεται ως λίγο μικρότερη από την ελάχιστη και η υψηλότερη H_n λίγο μεγαλύτερη από τη μέγιστη τιμή των δεδομένων ώστε το μήκος L των **διαστημάτων** (bins) να είναι "στρόγγυλος" αριθμός. Τότε μετράμε τον αριθμό των παρατηρήσεων x_i σε κάθε διάστημα (δηλαδή, έτσι ώστε να ισχύει $L_i < x_i <= H_i$) και κατασκευάζουμε το διάγραμμα κατανομής συχνοτήτων, συνήθως ως ραβδόγραμμα (bar graph). Αυτό ονομάζεται **ιστόγραμμα** (histogram).

Με το ιστόγραμμα παίρνουμε μια ιδέα για το **σχήμα** ή μορφή της κατανομής. Εφόσον τα δεδομένα προέρχονται από συγκεκριμένη πηγή ή αιτία με συγκεκριμένη συμπεριφορά και σε συγκεκριμένες συνθήκες, περιμένουμε ότι αυξάνοντας σταδιακά τις παρατηρήσεις μας θα *αναδύεται* αυτή η συμπεριφορά ως ένα χαρακτηριστικό σχήμα της κατανομής που θα εξομαλύνεται και θα γίνεται πιο καθαρό και σαφές όσο αυξάνεται ο αριθμός των καταγεγραμμένων παρατηρήσεών μας.

Αν έχουμε πολύ μεγάλο αριθμό παρατηρήσεων από δεδομένα με συνεχείς τιμές μέσα από ένα συγκεκριμένο εύρος δυνατών τιμών, μπορούμε να μικρύνουμε το εύρος του κάθε υποδιαστήματος ή κατηγορίας τιμών χωρίς να χαθεί ουσιαστική πληροφορία μέσα στις λεπτομέρειες των τυχαίων διακυμάνσεων. Όσο πιο μεγάλο πλήθος παρατηρήσεων έχουμε, τόσο πιο στενές κατηγορίες μπορούμε να χρησιμοποιήσουμε και τόσο πιο ομαλή περιμένουμε να γίνεται η καμπύλη της κατανομής συχνοτήτων. Για άπειρο αριθμό παρατηρήσεων θα μπορούσαμε να πάρουμε απειροστά διαστήματα και θα αντιστοιχούσαμε μια "πυκνότητα" εμφανίσεων σε κάθε αριθμητική τιμή του συνεχούς (βλ. και πυκνότητα πιθανότητας που ορίσαμε στην εισαγωγή), ενώ η καμπύλη της κατανομής θα ήταν συνεχής, ομαλή και παραγωγίσιμη, εκτός ενδεχομένως από ορισμένα σημεία.

Τότε, η κατανομή είναι δυνατό να περιγραφεί ως ένα βαθμό ικανοποιητικά με τη χρήση κατάλληλων αριθμητικών μέτρων που δίνουν μια ιδέα αφενός για τις κύριες τάσεις που εκφράζονται μέσα από την κατανομή και αφετέρου για το σχήμα της κατανομής γενικότερα. Από μαθηματική άποψη, η καμπύλη της κατανομής θα μπορεί να περιγραφεί από κάποια μαθηματική συνάρτηση και θα υπάρχουν κάποιες παράμετροι της συνάρτησης αυτής που θα περιγράφουν την ακριβή μορφή της καμπύλης για τη δεδομένη συναρτησιακή μορφή. Πρώτα, ας μιλήσουμε για τα σχήματα των κατανομών.

– Παρατηρήσεις φαινομένων που θα μπορούσαμε να χαρακτηρίσουμε "εντελώς τυχαία" τείνουν να απλωθούν ομοιόμορφα στο δυνατό εύρος τιμών, δηλαδή κάθε δυνατό αποτέλεσμα έχει την ίδια πιθανότητα. Για παράδειγμα στο παιγνίδι κορώνα-γράμματα τα ενδεχόμενα είναι κορώνα ή γράμματα και εύκολα επιβεβαιώνουμε ότι οι συχνότητες εμφάνισης όταν στρίψουμε αρκετές φορές ένα νόμισμα, θα είναι πάντα περίπου ίδιες. Για συνεχή αριθμητικά δεδομένα, η καμπύλη θα τείνει σε μια οριζόντια ευθεία που θα αντιστοιχεί στην τιμή των συχνοτήτων για κάθε κατηγορία δεδομένων.

– Έχουμε επίσης παρατηρήσεις δεδομένων που τείνουν προς μια ορισμένη τιμή αλλά υπάρχουν τυχαίες διακυμάνσεις ή διαταράξεις που αλλοιώνουν τις παρατηρήσεις σε σχέση με αυτή τη χαρακτηριστική τιμή. Οι διαταράξεις ή **αποκλίσεις** από αυτή την τιμή εμφανίζονται πιο συχνά όταν είναι μικρές και πιο σπάνια όταν είναι μεγάλες. Αυτό έχει συχνά ως αποτέλεσμα ένα συμμετρικό "κωδονοειδές" σχήμα κατανομής. Μία πολύ συνηθισμένη μορφή κατανομής αυτού του είδους είναι η λεγόμενη **κανονική** κατανομή (normal distribution) για την οποία είπαμε ήδη μερικά πράγματα στην εισαγωγή. Αυτή χαρακτηρίζεται από μία **κορυφή** ή **επικρατούσα τιμή** (peak, mode) στην τιμή που εμφανίζεται συχνότερα και που αντιστοιχεί σε μηδενικές τυχαίες

διακυμάνσεις. Για παράδειγμα, αν ρίχνουμε βελάκια σε ένα στόχο και μετρήσουμε τις αποστάσεις των σημείων πρόσκρουσης από το κέντρο του στόχου θα βρούμε το μισό μιας καμπύλης κανονικής κατανομής (αφού παίρνουμε θετικές τιμές και η κορυφή είναι στο μηδέν). Επίσης, αν πάρουμε τις x συντεταγμένες και τις y συντεταγμένες των σημείων, θα πάρουμε πάλι κανονική κατανομή με κορυφή που αντιστοιχεί στη συντεταγμένη του κέντρου του στόχου.

– Γενικότερα, μπορεί να έχουμε κατανομές με μία κορυφή, αλλά να μην είναι συμμετρικές. Μπορεί να εμφανίζεται μια "ουρά" από αριστερά ή από δεξιά που αντιστοιχεί σε ακραίες αποκλίσεις από την τιμή κορυφής οι οποίες εμφανίζονται μεν σπάνια αλλά υπάρχουν. Τότε, λέμε ότι η κατανομή είναι **ασύμμετρη** (skewed) από αριστερά ή δεξιά αντίστοιχα. Πάντως, οι κατανομές που έχουν μια σαφώς διακριτή κορυφή λέμε ότι είναι **μονοτροπικές** (monomodal).

– Υπάρχει περίπτωση να έχουμε δύο ή περισσότερες κορυφές σε μια κατανομή. Π.χ. μπορεί να έχουμε δεδομένα που προέρχονται από δυο διαφορετικές διεργασίες που συμπεριφέρονται σύμφωνα με την κανονική κατανομή, αλλά τα διαστήματα των τιμών τους αλληλεπικαλύπτονται. Αυτές λέγονται **διτροπικές** (bimodal) ή **πολυτροπικές** (multimodal) κατανομές.

Τα συνηθέστερα ποσοτικά, αριθμητικά μέτρα που χρησιμοποιούμε για την περιγραφή των κύριων τάσεων μιας κατανομής, είναι τα εξής:

– Η κορυφή ή **τρόπος** (mode) ή επικρατούσα τιμή μιας κατανομής αναφέρθηκε ήδη και είναι ουσιαστικά κάθε τοπικό μέγιστο της καμπύλης κατανομής συχνοτήτων, στο όριο των άπειρων παρατηρήσεων, ώστε να εξασφαλίζεται η ομαλότητα της καμπύλης. Πρακτικά, είναι οι κορυφές που διακρίνονται καθαρά "με το μάτι" μέσα από τις μικροδιακυμάνσεις των δεδομένων γύρω από την καμπύλη κατανομής προς την οποία τείνουν.

– ο **διάμεσος** (median) είναι η "μεσαία" τιμή στο σύνολο των παρατηρήσεών μας. Αφορά αριθμητικά δεδομένα τα οποία έχουμε *ταξινομήσει κατ' αύξουσα σειρά*. Αν έχουμε $2n+1$ δεδομένα x_i, τότε ο διάμεσος είναι η τιμή x_{n+1}. Αν έχουμε $2n$ δεδομένα, τότε παίρνουμε ως διάμεσο την τιμή $(x_n + x_{n+1})/2$

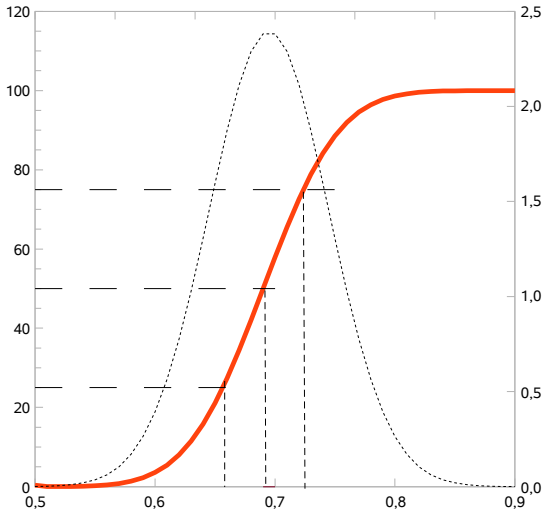
– Η **μέση τιμή** (mean) μ, όπως ξέρουμε, ορίζεται από τη σχέση $\sum_{i=1}^{N} x_i / N$ όπου N το πλήθος των δεδομένων.

Είχαμε αναφερθεί στην εισαγωγή στις παραμέτρους που αναφέρονται στον πληθυσμό και στα στατιστικά που αναφέρονται στα δείγματα, αλλά και τα δύο αποτελούν τρόπους για να περιγράφουμε διάφορα χαρακτηριστικά των κατανομών. Στη συνέχεια αναφερόμαστε πιο αναλυτικά στα κυριώτερα στατιστικά και παραμέτρους.

Διάμεσος (median)

Ο ορισμός δόθηκε πιο πάνω, αλλά υπάρχει και ένας ενδιαφέρων γραφικός τρόπος προσδιορισμού του διάμεσου. Αν από την κατανομή συχνοτήτων (f_1, f_2, f_3, ... f_N) πάρουμε την αθροιστική κατανομή (F_1, F_2, F_3, ... F_N) = (f_1, $f_1 + f_2$, $f_1 + f_2 + f_3$, ... $\Sigma^{N}_{i=1}$ (f_i)) και τη μετατρέψουμε σε ποσοστά διαιρώντας με F_N, μπορούμε να κάνουμε τη γραφική τους παράσταση έναντι των τιμών v_i που έχουν καταγραφεί. Τότε, ο διάμεσος είναι η τιμή που αντιστοιχεί στο ποσοστό 0.5 ή 50%.

Στο επόμενο σχήμα, ο αριστερός άξονας των y αναφέρεται στα % ποσοστά της αθροιστικής κατανομής (παχειά γραμμή) ενώ ο δεξιός άξονας στην αρχική κατανομή (λεπτή μαύρη γραμμή). Η οριζόντια διακεκομμένη ορίζει το ποσοστό 50% και το σημείο τομής της με την καμπύλη της αθροιστικής κατανομής ορίζει τον διάμεσο που, για το συγκεκριμένο παράδειγμα είναι ίσος με 0.69.



Μέση τιμή (mean value) **/ αριθμητικός μέσος** (arithmetic mean).

Αυτή είναι η πιο γνωστή παράμετρος/στατιστικό και, όσον αφορά ένα δείγμα, ουσιαστικά πρόκειται για το άθροισμα των καταγεγραμμένων τιμών x, μιας μεταβλητής X, διά τον αριθμό τους. Έστω ένα δείγμα από n παρατηρήσεις x_i, i=1, 2, ..., n, που παίρνουν m διαφορετικές τιμές v_j, j=1, 2, ..., m ≤ n. Αν μια τιμή v_i στα δεδομένα μας παρατηρείται περισσότερες από μία φορές, τότε η συχνότητα εμφάνισής της, είναι $f_i > 1$. Το άθροισμα των συχνοτήτων είναι ο συνολικός αριθμός δεδομένων. Έτσι, έχουμε τις εξής παραλλαγές του ορισμού του αριθμητικού μέσου:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{r=1}^{m} f_r v_r}{n} = \frac{\sum_{r=1}^{m} f_r v_r}{\sum_{r=1}^{m} f_r}$$

Με $\bar{x}$ παριστάνεται το σταστιστικό της μέσης τιμής του δείγματος. Μία τιμή μπορεί να παρατηρείται περισσότερες από μία φορές, π.χ. μπορεί να βρήκαμε 5 φοιτητές που ζυγίζουν 65 κιλά. Τότε η συχνότητα f της τιμής 65 είναι 5. Οι παρατηρούμενες τιμές θα είναι m τον αριθμό και προφανώς αυτός δε μπορεί να υπερβαίνει τον αριθμό n, των παρατηρήσεων (το πολύ πολύ κάθε αριθμητική τιμή να παρατηρείται μόνο μία φορά). Το άθροισμα των συχνοτήτων δίνει το συνολικό αριθμό παρατηρήσεων n.

Χρησιμοποιώντας τον ίδιο συμβολισμό, μπορούμε να γράψουμε το *ποσοστό εμφάνισης* p_r, κάθε αριθμητικής τιμής:

$$p_r = \frac{f_r}{n} = \frac{f_r}{\sum_{r=1}^{n_v} f_r}$$

οπότε, ο αριθμητικός μέσος μπορεί να γραφεί με πιο συμπαγή μορφή και ως εξής:

$$\bar{x} = \sum_{r=1}^{n_v} p_r v_r$$

Από τον ορισμό του ποσοστού προκύπτει ότι το άθροισμα των ποσοστών εμφάνισης κάθε τιμής (και γενικότερα, κάθε ενδεχόμενου) είναι ίσο με τη μονάδα. Η αναλογία με τις πιθανότητες είναι φανερή και τα ποσοστά εμφάνισης μπορούν να χρησιμεύσουν ως **εκτιμήσεις** των αντίστοιχων πιθανοτήτων που υπαγορεύονται από την κατανομή του σχετικού πληθυσμού. Οι συχνότητες εμφάνισης αποτελούν την *κατανομή του δείγματος* και τα ποσοστά ή οι πιθανότητες δίνουν την ίδια κατανομή στην *κανονικοποιημένη* μορφή της.

Αν μπορούσαμε να κάνουμε άπειρες μετρήσεις (ή τόσες ώστε να καλύψουμε το συνολικό πληθυσμό, αν αυτός είναι πεπερασμένος) τότε, με τη βοήθεια των συχνοτήτων εμφάνισης, θα παίρναμε την *κατανομή του πληθυσμού*. Όσο πιο μεγάλο ένα δείγμα τόσο πιο πολύ θα τείνει η κατανομή του να ταυτιστεί με την κατανομή του πληθυσμού. Η παράμετρος που αντιστοιχεί στον αριθμητικό μέσο είναι η **μέση τιμή** που συμβολίζουμε με **μ**. Χρησιμοποιούνται επίσης και οι συμβολισμοί E(X) και <x>, όπου X η θεωρούμενη τυχαία μεταβλητή και x οι τιμές που λαμβάνει. Για διακριτά δεδομένα ο ορισμός μπορεί να γραφεί ανάλογα με αυτόν του αριθμητικού μέσου, αλλά είναι πιο εξυπηρετικό, ιδίως αν πρόκειται για άπειρο πληθυσμό, να βασιστούμε στην κατανομή της πυκνότητας πιθανότητας. Τότε:

$$\mu = E(x) = \sum_x p_x x \quad \text{ή} \quad \mu = E(x) = \int_x x\, p(x)\, dx$$

Αντοχή στατιστικών

Ο διάμεσος συχνά είναι καλύτερο μέτρο από τη μέση τιμή για την περιγραφή της κύριας τάσης μιας κατανομής. Ας δούμε το εξής απλό παράδειγμα. Έχουμε το δείγμα Δ = {1, 3, 3, 4, 5} με διάμεσο 3 και μέσο 3.2. Τώρα, κάνουμε άλλη μία παρατήρηση η οποία έχει την τιμή 40 η οποία είναι πολύ μεγαλύτερη από τις άλλες. Το νέο δείγμα Δ'={1, 3, 3, 4, 5, 40} έχει τις πιο πολλές τιμές του στο διάστημα 1 έως 5 και ο διάμεσος τώρα είναι 3.5 αλλά η μέση τιμή εκτοξεύεται στα 9.333, δίνοντας μια απατηλή εικόνα για το γενικότερο επίπεδο των τιμών. Ένα άλλο "κλασσικό" παράδειγμα είναι η άνοδος του μέσου εισοδήματος εξ αιτίας μιας πολύ πλούσιας μειοψηφίας της οποίας η κατάσταση δεν χαρακτηρίζει αυτή του γενικότερου πληθυσμού.

Ένα στατιστικό λέγεται πιο **ανθεκτικό** από ένα άλλο, όταν οι μεταβολές που υφίσταται με την προσθήκη νέων αποτελεσμάτων στο δείγμα είναι μικρότερες απ' ο,τι για το άλλο. Από τα παραδείγματα φαίνεται ότι ο διάμεσος είναι πιο ανθεκτικό στατιστικό από τον αριθμητικό μέσο.

Παράμετροι και στατιστικά για τη διασπορά των τιμών

Όσο μεγαλύτερο και αντιπροσωπευτικότερο το δείγμα ενός πληθυσμού, τόσο πιο πολύ θα πλησιάζει η μέση τιμή του δείγματος (αριθμητικός μέσος) αυτή του πληθυσμού. Οι παραπάνω τιμές τείνουν να συμπέσουν όχι μόνο μεταξύ τους αλλά και με το διάμεσο, για συμμετρική μονοτροπική κατανομή όπως είναι η κανονική κατανομή. Αυτό όμως, δεν είναι απαραίτητο να συμβαίνει σε όλες τις κατανομές. Όταν το σχήμα της κατανομής είναι πιο πολύπλοκο, αυτές οι τιμές διαφέρουν. Το σχήμα της κατανομής δείχνει ακριβώς πώς κατανέμονται οι παρατηρήσεις μας στις ενδεχόμενες τιμές. Αν και η κατανομή, καθώς και το σχήμα της δίνουν όλη την πληροφορία σχετικά με την πραγματοποίηση των διαφόρων ενδεχομένων, θέλουμε να έχουμε και κάποια ποσοτικά μέτρα που να μας επιτρέπουν την κατάταξη των διαφόρων κατανομών. Θέλουμε π.χ. να περιγράψουμε με έναν αριθμό το πόσο κοντά ή μακριά από τη μέση τιμή βρίσκονται οι περισσότερες άλλες τιμές. Με λίγα λόγια, χρειαζόμαστε ποσοτικά μέτρα της *μεταβλητότητας* (variability) των τιμών.

Εύρος και υποδιαιρέσεις αυτού

Το πιο απλό μέτρο της μεταβλητότητας των δεδομένων είναι το **εύρος** (range) των αριθμητικών τιμών τους δηλαδή η διαφορά της ελάχιστης από τη μέγιστη καταγεγραμμένη παρατήρηση. Τα πολλοστημόρια είναι διαστήματα στα οποία διαιρούμε τις συχνότητες ώστε να έχουμε μία ιδέα για το πού βρίσκονται οι περισσότερες παρατηρήσεις. Τα εκατοστημόρια διαιρούν τις συχνότητες σε 100 ίσα μέρη και ο διάμεσος βρίσκεται στο 50 εκατοστημόριο. Παρόμοια, τα τεταρτημόρια διαιρούν τις συχνότητες σε 4 ίσα μέρη και τα δεκατημόρια σε 10 ίσα μέρη. Ο διάμεσος είναι το δεύτερο τεταρτημόριο.

Στο προηγούμενο σχήμα (γραφική εύρεση διαμέσου) βλέπουμε ότι το δεύτερο και το τρίτο τεταρτημόριο αντιστοιχούν σε πολύ στενά διαστήματα του οριζόντιου άξονα (από 0.65 ως 0.69 και από 0.69 ως 0.73, αντίστοιχα) ενώ το πρώτο και το τελευταίο τεταρτημόριο έχουν πολύ μεγαλύτερο πλάτος. Καταλαβαίνουμε ότι το 50% των καταγεγραμμένων παρατηρήσεων (δεύτερο

και τρίτο τεταρτημόριο μαζί) είναι συσσωρευμένο σε μια στενή περιοχή από 0.65 ως 0.73 ενώ το υπόλοιπο 50% είναι πολύ πιο διεσπαρμένο σε μεγαλύτερες ή μικρότερες περιοχές. Αυτό είναι ένας άλλος τρόπος για να πούμε ότι οι τιμές από 0.65 ως 0.73 είναι οι πιο συχνά απαντώμενες.

Διασπορά και τυπική απόκλιση

Ένα άλλο πολύ διαδεδομένο μέτρο είναι η **τυπική απόκλιση** (standard deviation) που είναι ουσιαστικά ένας τρόπος να μετρήσουμε τις αποκλίσεις από τη μέση τιμή, δηλαδή δείχνει μέχρι που φτάνουν οι πιο συχνά εμφανιζόμενες αποκλίσεις. Η τυπική απόκλιση συνδέεται στενά με την κανονική κατανομή με τρόπο που θα εξηγήσουμε αργότερα.

Σε κάθε κατανομή, ως **απόκλιση** ορίζεται η διαφορά μιας οποιασδήποτε τιμής από τη μέση τιμή, $x - \bar{x}$. Αν πάρουμε το μέσο όρο των αποκλίσεων θα έχουμε μια απατηλή εκτίμηση των αποκλίσεων γιατί κάποιες θετικές αποκλίσεις θα έχουν αντισταθμιστεί από κάποιες αρνητικές οδηγώντας σε αποτέλεσμα ταυτοτικό ίσο με μηδέν, όπως αποδεικνύεται πολύ εύκολα:

$$\frac{1}{n}\sum_i (x_i - \bar{x}) = \frac{1}{n}\sum_i (x_i - \frac{\sum_j x_j}{n}) = \frac{1}{n^2}\sum_i \sum_j (x_i - x_j)$$

και στο παραπάνω διπλό άθροισμα, για κάθε διαφορά $x_i - x_j$ υπάρχει και η διαφορά $x_j - x_i$ οπότε το άθροισμα είναι ταυτοτικά ίσο με μηδέν.

Θα μπορούσαμε να πάρουμε τη μέση τιμή των απόλυτων τιμών των αποκλίσεων και αυτό θα ήταν σωστό αλλά δεν προτιμάται γιατί η απόλυτη τιμή είναι ασυνεχής στο μηδέν και επομένως, όχι και τόσο βολική για μαθηματικούς χειρισμούς. Το επόμενο βήμα για μια εκτίμηση όπου κάθε απόκλιση θα έχει θετική συνεισφορά, είναι να πάρουμε τα τετράγωνα αυτών. Το μέσο τετράγωνο των αποκλίσεων το ονομάζουμε **διασπορά** (variance) της κατανομής ενώ η τετραγωνική ρίζα της διασποράς είναι η **τυπική απόκλιση** (standard deviation). Το σύμβολο της τυπικής απόκλισης είναι s όταν αναφέρεται σε δείγμα (στατιστικό) και σ όταν αναφέρεται στον πληθυσμό (παράμετρος). Έτσι έχουμε τους παρακάτω ορισμούς:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} = \frac{\sum_{r=1}^{m} f_r (v_r - \bar{x})^2}{n} \quad \text{και} \quad s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}} = \sqrt{\frac{\sum_{r=1}^{m} f_r (v_r - \bar{x})^2}{n}}$$

(και ανάλογα για την παράμετρο σ, αρκεί να αντικαταστήσουμε το s με **σ** και το $\bar{x}$ με **μ**).

Χρησιμοποιώντας το συμβολισμό $E(\cdot)$ ή $\langle \cdot \rangle$ για τη μέση τιμή, μπορούμε να γράψουμε τα παραπάνω και ως εξής: $\sigma^2 = E((X-E(X))^2)$ και $s^2 = \langle (x - \langle x \rangle)^2 \rangle$.

Οι παραπάνω σχέσεις μπορούν να μετασχηματιστούν με μερικούς εύκολους αλγεβρικούς χειρισμούς και να δώσουν μια εναλλακτική, χρήσιμη σχέση:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^{m} x_i^2}{n} - 2\bar{x}\frac{\sum_{i=1}^{m} x_i}{n} + \frac{\sum_{i=1}^{m} \bar{x}^2}{n} = \frac{\sum_{i=1}^{m} x_i^2}{n} - 2\bar{x}^2 + \frac{n\bar{x}^2}{n} = \frac{\sum_{i=1}^{m} x_i^2}{n} - \bar{x}^2$$

(και ανάλογα για την παράμετρο σ)

Αυτό μπορεί να γραφεί και έτσι: $\sigma^2 = E(X^2) - (E(X))^2$, $s^2 = \langle x^2 \rangle - \langle x \rangle^2$

Μετασχηματισμός μετατόπισης

Μερικές φορές, εξυπηρετεί να μετατοπίσουμε μια κατανομή κατά σταθερά ποσότητα K, πράγμα που ισοδυναμεί με ένα πολύ απλό μετασχηματισμό μεταξύ αρχικών και μετατοπισμένων τιμών γενικά όσο και ειδικότερα μεταξύ των μέσων τιμών. Αν η μεταβλητή μας ειναι x και ορίσουμε μια νέα μεταβλητή $x' = x + d$ όπου d κάποια σταθερή ποσότητα, τότε είναι πολύ εύκολο να δείξουμε ότι $\langle x' \rangle = \langle x \rangle + d$. Αν θέσουμε $d = -\langle x \rangle$, συνεπάγεται ότι η μετασχηματισμένη μεταβλητή

x' = x - <x> εκφράζει τις αποκλίσεις από τη μέση τιμή και <x'> = 0.

Ανάλογο μετασχηματισμό μπορούμε να βρούμε και για τη διασπορά. Αν θεωρήσουμε μια μέση τιμή x_0 για τη μετασχηματισμένη κατανομή, η διασπορά γύρω από αυτή γράφεται:

$$S^2=\frac{\sum_{i=1}^{n}(x_i-x_0)^2}{n}=\frac{\sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-x_0)^2}{n}=$$

$$=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}+2\frac{\sum_{i=1}^{n}(x_i-\bar{x})(\bar{x}-x_0)}{n}+\frac{\sum_{i=1}^{n}(\bar{x}-x_0)^2}{n}$$

Αλλά,

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}=s^2 \quad ,$$

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})(\bar{x}-x_0)}{n}=(\bar{x}-x_0)\left(\frac{\sum_{i=1}^{n}x_i}{n}-\bar{x}\frac{\sum_{i=1}^{n}1}{n}\right)=0$$

$$\frac{\sum_{i=1}^{n}(\bar{x}-x_0)^2}{n}=(\bar{x}-x_0)^2\frac{\sum_{i=1}^{n}1}{n}=(\bar{x}-x_0)^2=d^2$$

όπου με d παραστήσαμε τη διαφορά μεταξύ αρχικού και μετατοπισμένου μέσου, οπότε καταλήγουμε στον εξής απλό μετασχηματισμό: $S^2=s^2+d^2$, που s η τυπική απόκλιση της αρχικής, μη μετατοπισμένης κατανομής.

Πληθυσμοί και Δείγματα

Η στοιχειώδης "μονάδα" στην οποία βασίζεται η στατιστική είναι η **παρατήρηση**. Τα μεγέθη και οι συναρτήσεις κατανομών που ορίσαμε, καθώς και άλλα που θα ορίσουμε αργότερα, προέρχονται από την επεξεργασία ενός συνόλου παρατηρήσεων ή μάλλον, δεδομένων που προέρχονται από αυτές τις παρατηρήσεις. Συχνά, δεν είναι δυνατό να παρατηρήσουμε όλες τις δυνατές εκφάνσεις ενός φαινομένου, διεργασίας, κατάστασης κλπ, λόγω του μεγάλου πλήθους παρατηρήσεων που πρέπει να γίνουν, οπότε και περιοριζόμαστε σε ένα όσο γίνεται πιο **αντιπροσωπευτικό** δείγμα. Λεπτομερής συζήτηση για τα αντιπροσωπευτικά δείγματα έχει γίνει στην Εισαγωγή.

Η σχέση του δείγματος με τον πληθυσμό έχει επίσης συζητηθεί και είναι αυτή του υποσυνόλου με το υπερσύνολο. **Πληθυσμός** είναι το σύνολο των δυνατών παρατηρήσεων ενός συγκεκριμένου τύπου και **δείγμα** είναι ένα υποσύνολο αυτών, αλλά έτσι επιλεγμένων ώστε να δίνεται *ίση ευκαιρία σε κάθε δυνατή παρατήρηση να εμφανιστεί.* Η προφανής χρησιμότητα του δείγματος είναι ότι μας επιτρέπει να βγάλουμε συμπεράσματα για το σύνολο του πληθυσμού. Ωστόσο, είναι δυνατό και συχνά επιθυμητό να κάνουμε και το αντίστροφο: από τη γνώση της κατάστασης ή των χαρακτηριστικών του πληθυσμού να βγάλουμε συμπέρασμα για το δείγμα. Π.χ. έχουμε μια βιομηχανική διεργασία η οποία παράγει ένα συγκεκριμένο προϊόν σε τεμάχια ή μονάδες και έστω ότι γνωρίζουμε το πόσο συχνά και υπό ποιες προϋποθέσεις προκύπτουν ελαττωματικά τεμάχια. Αυτό μπορεί να το ξέρουμε είτε από την ανάλυση προηγούμενων επαρκών και αντιπροσωπευτικών δειγμάτων είτε από γνώση των φυσικών νόμων και αρχών που διέπουν τη λειτουργία της συγκεκριμένης διαδικασίας. Τότε, θέλουμε να γνωρίζουμε πόσο πιθανό είναι ένα τεμάχιο σε μια παρτίδα των δέκα να είναι ελαττωματικό. Η παρτίδα είναι το δείγμα και η περιγραφή της θα προκύψει από τη γνώση που πιστεύουμε ότι έχουμε για τον πληθυσμό, δηλαδή για όλα τα τεμάχια που παρήχθησαν, παράγονται ή θα παραχθούν.

Σε αντιστοιχία με τις έννοιες του πληθυσμού και του δείγματος έχουμε ορίσει τις έννοιες της

παραμέτρου και του στατιστικού. Παράμετρος είναι ένας αριθμός που συνοψίζει την κατανομή ενός ολόκληρου πληθυσμού, ενώ στατιστικό είναι ένας αριθμός που συνοψίζει την κατανομή ενός συγκεκριμένου δείγματος. Αυτό που μας επιτρέπει να γενικεύουμε τις παρατηρήσεις ενός δείγματος και να τις ανάγουμε στο επίπεδο πληθυσμού, ή το αντίστροφο, είναι η μαθηματική έννοια που λέγεται **κατανομή δειγματοληψίας**. Η χρησιμότητα της έννοιας αυτής θα φανεί καθαρά όταν μιλήσουμε για το Κεντρικό Οριακό Θεώρημα. Αντί να την ορίσουμε, θα τη σκιαγραφήσουμε με ένα παράδειγμα.

Παράδειγμα – Κατανομή Δειγματοληψίας.

Στρίβουμε ένα νόμισμα δύο φορές και καταγράφουμε και τις δυο φορές, αν ήρθε κορώνα ή γράμματα. Αυτό είναι ένα δείγμα. Προφανώς ο πληθυσμός είναι άπειρος γιατί δεν υπάρχει όριο στο πόσες φορές μπορούμε να στρίβουμε νομίσματα.

Μετά υπολογίζουμε το ποσοστό *p* των φορών που ήρθε κορώνα. Προφανώς υπάρχουν τρεις περιπτώσεις:

– Κορώνα , Κορώνα => $p = 1$

– Κορώνα, Γράμματα ή Γράμματα, Κορώνα => $p = 0.5$

– Γράμματα, γράμματα => $p = 0$

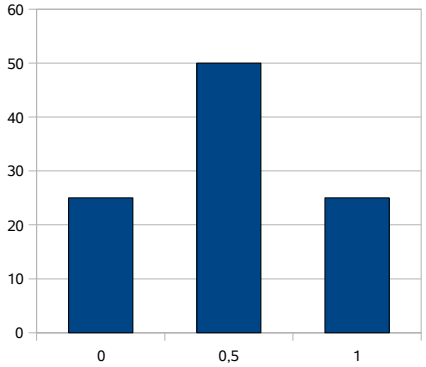
Ο,τι τιμή κι αν υπολογίσαμε, το *p* είναι ένα στατιστικό του δείγματός μας. Από την άλλη, είναι προφανές ότι η μία πλευρά δεν έχει λόγο να ευνοείται έναντι της άλλης, ούτε και υπάρχει περίπτωση κάθε ρίψη να επηρεάζει τις επόμενες, άρα αν ήταν δυνατό να κάνουμε άπειρες παρατηρήσεις αντιλαμβανόμαστε διαισθητικά ότι θα παίρναμε μια μέση τιμή $P = 0.5$. Αυτή θα ήταν μία παράμετρος του πληθυσμού. *Οι παράμετροι συμβολίζονται με κεφαλαία ενώ τα στατιστικά με μικρά γράμματα*.

Αξίζει να συγκρατήσουμε μέχρι εδώ ότι

– για κάθε ρίψη τα δύο αποτελέσματα είναι ισοπίθανα. Τα δυνατά αποτελέσματα τα λέμε **ενδεχόμενα**. Εδώ έχουμε μία περίπτωση **ανεξάρτητων** ενδεχομένων, δηλαδή το αποτέλεσμα κάθε ρίψης δεν εξαρτάται από τις προηγούμενες (ούτε από τις επόμενες!)

– για κάθε δύο ρίψεις, ο συνδυασμός ΚΓ είναι δυο φορές πιθανότερος από τον ΚΚ ή τον ΓΓ. Θεωρώντας ως παρατήρηση τα *ζεύγη* ρίψεων, τα ενδεχόμενα είναι 4 (ΚΚ, ΚΓ, ΓΚ και ΓΓ) και είναι και αυτά ανεξάρτητα. Δεν έχουμε λόγο να υποθέσουμε κάτι άλλο. Αλλά δύο από τα τέσσαρα δίνουν αποτέλεσμα για το στατιστικό που ορίσαμε, την τιμή 0.5. Αν λοιπόν, ως παρατήρηση θεωρήσουμε το στατιστικό *p* του ποσοστού εμφάνισης Κ, τότε έχουμε τρία ενδεχόμενα, 0, 0.5 και 1, από τα οποία το δεύτερο ευνοείται έναντι των άλλων δύο.

Αν, με βάση τις παραπάνω παρατηρήσεις, κάνουμε ένα ιστόγραμμα των % ποσοστών εμφάνισης κάθε τιμής του *p* θα πάρουμε μία συμμετρική κατανομή όπως στο σχήμα:



Στην πραγματικότητα, αυτή θα ήταν η κατανομή κάποιας αντίστοιχης παραμέτρου *P* για άπειρο πληθυσμό ζευγών ρίψεων. Πράγματι, μπορούμε να επαναλάβουμε το πείραμα των δύο διαδοχικών ρίψεων όσες φορές θέλουμε και να καταγράφουμε για κάθε ζεύγος ρίψεων το στατιστικό *p*. Καταγράφοντας τη συχνότητα εμφάνισης των τριών δυνατών τιμών του θα επαληθεύσουμε ότι το αντίστοιχο ιστόγραμμα τείνει να ταυτιστεί με την παραπάνω μορφή.

Το ιστόγραμμα του στατιστικού *p* παριστάνει μία **κατανομή δειγματοληψίας**. Ως γενικό ορισμό μπορούμε να πούμε ότι **κατανομή δειγματοληψίας είναι η κατανομή ενός στατιστικού**. *Δεν πρέπει να συγχέεται με την κατανομή ενός δείγματος.* Για παράδειγμα, αν πάρουμε ένα δείγμα 10 φοιτητών του ΤΜΕΥ και μετρήσουμε το ύψος τους, αυτό θα ήταν μια κατανομή δείγματος. Αν παίρναμε 5 δείγματα των 10 φοιτητών, υπολογίζαμε το μέσο όρο ύψους κάθε δείγματος και κάναμε το ιστόγραμμα αυτών των μέσων όρων, αυτό θα ήταν η κατανομή του στατιστικού "μέσος όρος".

Ας υποθέσουμε ότι ο *διάμεσος* του ύψους των φοιτητών του παν/μίου Ιωαννίνων είναι 1,65. Δηλαδή, οι μισοί είναι πάνω από 1,65 και οι μισοί κάτω από 1,65. Διερωτώμαστε ποια είναι η πιθανότητα τρεις φοιτητές επιλεγμένοι στην τύχη να έχουν ύψος πάνω από 1,65. Είναι προφανές ότι για κάθε ένα φοιτητή ή φοιτήτρια που θα διαλέξουμε, υπάρχουν δύο ενδεχόμενα: κάτω από 1,65, έστω ενδεχόμενο Α και πάνω από 1,65, έστω ενδεχόμενο Β. Επίσης, αφού 1,65 είναι ο διάμεσος, οι μισοί θα είναι πάνω από 1,65 και οι μισοί κάτω από 1,65, άρα δεν έχουμε λόγο να πιστεύουμε ότι για κάθε έναν που επιλέγουμε τυχαία, το ενδεχόμενο Α ευνοείται έναντι του Β, ούτε και το αντίστροφο.

Τώρα, για τρεις φοιτητές, τα ενδεχόμενα προφανώς είναι ΑΑΑ, ΒΑΑ, ΑΒΑ, ΑΑΒ, ΒΒΑ, ΒΑΒ, ΑΒΒ και ΒΒΒ, δηλαδή $2^3 = 8$. Οι αντίστοιχες τιμές του στατιστικού *π* ορισμένου ως συχνότητα επαλήθευσης του Β στην τριάδα φοιτητών (πόσοι από τους τρεις είναι πάνω από 1.65) είναι 0, 0.33..., 0.66... και 1. Με βάση τον αριθμό των δυνατών ενδεχομένων, η κατανομή των διαφόρων τιμών του στατιστικού θα ήταν σύμφωνη με τον παρακάτω πίνακα:

Τιμή	%
0,00	12,5
0,33	37,5
0,66	37,5
1,00	12,5

Το ενδεχόμενο ΒΒΒ αντιστοιχεί σε τιμή στατιστικού 1. Με βάση την κατανομή πληθυσμού που καταστρώσαμε, περιμένουμε 12,5% πιθανότητα να επαληθευθεί αυτό. Όπως και με το παιγνίδι κορώνα-γράμματα, έτσι κι εδώ δεν περιμένουμε η μία δειγματοληψία να επηρεάζει την άλλη, ούτε και κανένα από τα ενδεχόμενα ΑΑΑ, ΒΑΑ κλπ να υπερτερεί έναντι των άλλων, οπότε αν παίρναμε δείγματα πολλές φορές θα περιμέναμε η κατανομή της δειγματοληψίας να προσεγγίζει την προβλεπόμενη κατανομή πληθυσμού.

Τώρα, τα παραπάνω μπορούν να γενικευθούν ως προς τις εξής δύο κατευθύνσεις:

– αύξηση μεγέθους δείγματος

– διαφορετική σχέση μεταξύ ενδεχομένων, δηλαδή η περίπτωση κάποια να ευνοούνται έναντι κάποιων άλλων.

Ας δοκιμάσουμε να γενικεύσουμε το δείγμα. Αν έχουμε για κάθε μεμονωμένη παρατήρηση δύο ενδεχόμενα Α και Β για τα οποία δεν έχουμε λόγο να πιστεύουμε ότι το ένα ευνοείται έναντι του άλλου, π.χ. κορώνα-γράμματα, τότε λέμε ότι αυτά θα έχουν ίση πιθανότητα $P = 0,5$. Αν πάρουμε ένα δείγμα από *N* παρατηρήσεις, τότε θα έχουμε 2^N διαφορετικούς συνδυασμούς από Α και Β παρμένα συνολικά *N* φορές, όλους ισοπίθανους μεταξύ τους. Αυτοί είναι τα διαφορετικά

ενδεχόμενα για το δείγμα μας. Αφού είναι ισοπίθανα και δεν επηρεάζει το ένα το άλλο, η κατανομή πληθυσμού για καθένα από αυτά θα είναι $1/2^N = (0.5)^N$. Αν ορίσουμε ως στατιστικό τη συχνότητα εμφάνισης του Β, τότε ας παραστήσουμε με $p(X)$ την τιμή του για X εμφανίσεις του Β. Εργαζόμενοι όπως και πριν, βάσει του αριθμού των συνδυασμών, θα προβλέψουμε την κατανομή πληθυσμού ορίζοντας μια αντίστοιχη παράμετρο

$P(X)$ = {αριθμός συνδυασμών με X φορές το Β} / 2^N.

Ξέρουμε όμως από τον κλάδο της συνδυαστικής ότι ο αριθμητής της ανωτέρω παράστασης είναι ο αριθμός των μεταθέσεων με επανάληψη, N αντικειμένων όπου X είναι ίδια και N-X, πάλι ίδια μεταξύ τους, ή "πλήθος συνδυασμών N ανά X" που δίνεται από την παράσταση $N!/X!(N-X)!$

Επομένως, η κατανομή του πληθυσμού που προβλέπουμε θα είναι $P(X) = N!/X!(N-X)!\ (1/2)^N$, δηλαδή το γινόμενο της πιθανότητας κάθε μεμονωμένου συνδυασμού επί τον αριθμό των συνδυασμών που έχουν X φορές το Β, δίνει την πιθανότητα ένα δείγμα N παρατηρήσεων να έχει X "επιτυχίες".

Επειδή η παράσταση $N!/X!(N-X)!$ εμφανίζεται και ως συντελεστής δυνάμεων στο ανάπτυγμα του διωνύμου $(1+\alpha)^N$ για τον X-στό όρο, η παραπάνω κατανομή ονομάζεται **διωνυμική**. Όμως, αυτή δεν είναι η πιο γενική της μορφή. Ας δούμε την περίπτωση όπου τα βασικά ενδεχόμενα Α, Β των παρατηρήσεών μας δεν είναι ισοπίθανα, αλλά το ένα ευνοείται έναντι του άλλου. Π.χ. έχουμε μια μηχανή που παράγει εξαρτήματα και γνωρίζουμε ότι ένα στα δέκα βγαίνει ελαττωματικό. Αν αυτό είναι το ενδεχόμενο Β, ενώ το να βγει "καλό" είναι το ενδεχόμενο Α, τότε θα αποδώσουμε στο Α πιθανότητα 9/10 και στο Β πιθανότητα 1/10. Τότε, για να βγάλουμε την κατανομή για δείγματα N εξαρτημάτων από τα οποία X είναι ελαττωματικά, θα εργαστούμε παρόμοια με πριν. Θα σκιαγραφήσουμε λίγο πιο αναλυτικά, τα σημεία όπου υπάρχουν διαφορές.

Έστω το δείγμα {Β, Β, Α, Α, Α, ... κλπ}. Κάνοντας την πρώτη παρατήρηση υπάρχει 1/10 πιθανότητα να προκύψει το Β. *Με δεδομένο αυτό* υπάρχει 1 στις 10 να προκύψει Β στη δεύτερη και επομένως (1/10) (1/10) να βγούνε δύο Β στη σειρά. Παρόμοια (1/10) (1/10) (9/10) να προκύψει ΒΒΑ κλπ.

Αλλά αν δεχτούμε ότι η μία παρατήρηση δεν επηρεάζει την άλλη (αυτό βέβαια εξαρτάται από το πώς λειτουργεί η μηχανή για το συγκεκριμένο παράδειγμα), τότε για X εμφανίσεις του Β, θα έχουμε $(1/10)^X(9/10)^{N-X}$ πιθανότητα για ένα τέτοιο δείγμα N εξαρτημάτων, με όποια σειρά κι αν εμφανιστούν τα Α και Β. Αλλά το πλήθος των δειγμάτων N εξαρτημάτων με X ελαττωματικά θα δίνεται και πάλι από το διωνυμικό συντελεστή, οπότε θα ισχύει $P(X) = N!/X!(N-X)!\ (1/10)^X(9/10)^{N-X}$

Γενικότερα, αν έχουμε δύο δυνατά ενδεχόμενα για κάθε παρατήρηση, όπου το ένα έχει πιθανότητα p και το άλλο q (με $p+q=1$, αφού δεν υπάρχει άλλο ενδεχόμενο), θα ισχύει

$$P(X)=\frac{N!}{X!(N-X)!}p^X q^{N-X}$$

Αυτή είναι η γενική μορφή της διωνυμικής κατανομής. Προσέξτε ότι τα p, q είναι παράμετροι της κατανομής και όχι στατιστικά δείγματος.

Τώρα, η διωνυμική κατανομή συνηθίζεται να δίνεται σε πίνακες ως συνάρτηση του μεγέθους του δείγματος N, των αριθμών επιτυχίας X και της πιθανότητας "επιτυχίας" ανά παρατήρηση, p (ορίζοντας τα δύο ενδεχόμενα ως "επιτυχία" και "αποτυχία", ο,τι και αν σημαίνει αυτό). Δίνεται επίσης και από συναρτήσεις στα προγράμματα λογιστικών φύλλων όπως Excel και OpenOfficeCalc.

Παράδειγμα:

Πιθανότητα ελαττωματικού εξαρτήματος σε μία παρατήρηση: 0.1

α) Ποια η πιθανότητα να μη βρεθεί κανένα ελαττωματικό σε δείγμα τεσσάρων;

β) Ποια η πιθανότητα να βρεθεί ένα ελαττωματικό σε δείγμα τεσσάρων;

γ) Ποια η πιθανότητα να βρεθεί **το πολύ** ένα σε δείγμα τεσσάρων;

Από πίνακες:

α) $P(0; 4; 0.1) = 0,656 = 65,5\%$

β) $P(1; 4; 0.1) = 0,292 = 29,2\%$

γ) $P(<=1; 4; 0.1) = P(0; 4; 0.1) + P(1; 4; 0.1) = 94,7\%$

Η τελευταία απάντηση μας επαναφέρει και στην έννοια της *αθροιστικής κατανομής* (cumulative distribution) που είναι η πιθανότητα να έχει κάποια μεταβλητή (στατιστικό) τιμή μικρότερη από δεδομένο αριθμό *x*.

Προϋποθέσεις χρήσης της διωνυμικής κατανομής

Πότε **δεν** μπορούμε να χρησιμοποιήσουμε τη διωνυμική κατανομή: όταν τα ενδεχόμενα *δεν είναι* ανεξάρτητα. Πολλές αιτίες μπορεί να υπάρχουν για να συμβεί κάτι τέτοιο αλλά μία από αυτές είναι ο τρόπος με τον οποίο εμείς κάνουμε τη δειγματοληψία: αν πάρουμε ένα αντικείμενο από μία συλλογή *N* άσπρων ή μαύρων αντικειμένων και δεν το επιστρέψουμε, η πιθανότητα να βγει άσπρο ή μαύρο το επόμενο θα έχει αλλάξει. Αν *X* είναι τα μαύρα και *N*-*X* τα άσπρα, έστω ότι το πρώτο ήταν μαύρο. Η πιθανότητα να συμβεί αυτό ήταν *X*/*N*. Αν βγάλουμε το πρώτο εκτός, η πιθανότητα να βγει το δεύτερο μαύρο θα είναι (*X*-1)/(*N*-1) που είναι εν γένει διαφορετικός αριθμός. Αν όμως το επανατοποθετήσουμε, οι πιθανότητες μένουν ίδιες.

Αν ο πληθυσμός είναι αρκετά μεγάλος σε σχέση με το δείγμα τότε μπορούμε να κάνουμε δειγματοληψία χωρίς επανατοποθέτηση και αυτό δε θα αλλοιώσει σημαντικά την πιθανότητα.

Εμπειρικός κανόνας: *αν πληθυσμός > 20 φορές το δείγμα, μπορούμε να πούμε ότι τα ενδεχόμενα είναι προσεγγιστικά ανεξάρτητα και να εφαρμόσουμε τη διωνυμική κατανομή*.

Κανονική κατανομή

Τώρα, αν σχεδιάσουμε το ιστόγραμμα της διωνυμικής κατανομής για δεδομένες τιμές παραμέτρων, συχνά θα δούμε ότι προσεγγίζεται πολύ καλά από μια συνεχή καμπύλη που υπακούει στη συνάρτηση

$$P(x)=\frac{1}{\sqrt{(2\pi\sigma^2)}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}=\frac{1}{\sigma}\Phi\left(\frac{x-\mu}{\sigma}\right)$$

Αυτή ορίζει μια συνεχή κατανομή *πυκνότητας πιθανότητας*, τη γνωστή μας από την Εισαγωγή, **κανονική κατανομή**. Για μεγάλους αριθμούς παρατηρήσεων N, αποδεικνύεται ότι η κανονική κατανομή αποτελεί ικανοποιητική προσέγγιση της διωνυμικής κατανομής. Τα στάδια της απόδειξης σκιαγραφούνται στο τρίτο παράρτημα αυτής της ενότητας. Η κανονική κατανομή δίνει το μέσο για ένα μεγάλο δείγμα μετρήσεων. Το σχήμα της έχει τα εξής χαρακτηριστικά:

- μέγιστο στην τιμή *μ*, τη μέση τιμή
- λόγω συμμετρίας, η μέση τιμή είναι επίσης διάμεσος και κορυφή
- το εμβαδόν της είναι μονάδα γιατί τόση είναι η πιθανότητα για όλα τα δυνατά ενδεχόμενα
- η τυπική απόκλιση είναι *σ* που είναι επίσης και χαρακτηριστικό μήκος απόστασης από την κορυφή
- το 64.2% του εμβαδού βρίσκεται στη περιοχή μεταξύ *μ*-*σ* και *μ*+*σ*

Η κανονική κατανομή δίνεται και αυτή σε πίνακες αλλά επειδή είναι συνεχής για να τη χρησιμοποιήσουμε παίρνουμε διαφορές μεταξύ τιμών της αντίστοιχης αθροιστικής κατανομής για

να πάρουμε το εμβαδόν κάτω από την καμπύλη για το συγκεκριμένο διάστημα τιμών Δx που μας ενδιαφέρει κατά περίπτωση. Στους πίνακες δίνεται η **τυπική κανονική κατανομή** που λαμβάνεται αν μετασχηματίσουμε ως εξής: z = (x-μ)/σ, επομένως παίρνουμε μια καμπύλη με κορυφή στο 0 και τυπική απόκλιση σ = 1. Η μαθηματική έκφραση της τυπικής κανονικής κατανομής είναι, επομένως, πιο απλή:

$$P(z)=\frac{1}{\sqrt{(2\pi\sigma^2)}}e^{-z^2}$$

Κεντρικό Οριακό Θεώρημα

Έστω ένας πληθυσμός του οποίου η κατανομή έχει μέση τιμή μ και τυπική απόκλιση σ. Το μέσο όρο από ένα δείγμα *N* μετρήσεων του πληθυσμού τον παριστάνουμε με $\bar{x}$. Αν πάρουμε πολλά δείγματα, τότε ο μέσος των $\bar{x}$ συμβολίζεται με $\mu_{\bar{x}}$ και η τυπική απόκλισή τους με $\sigma_{\bar{x}}$ Αν το *N* είναι αρκετά μεγάλο, τότε η κατανομή δειγματοληψίας τείνει στο σχήμα της κανονικής κατανομής με $\mu_{\bar{x}}=\mu$ και $\sigma_{\bar{x}}=\frac{\sigma}{\sqrt{N}}$

Αυτό είναι το κεντρικό οριακό θεώρημα. Από αυτές τις σχέσεις μπορούμε να βρούμε τα μ και σ του πληθυσμού. Πρακτικά, για μέγεθος δείγματος από 30 και πάνω, ισχύει με πολύ καλή προσέγγιση.

Προσοχή! Το σχήμα της κατανομής του πληθυσμού μπορεί να είναι *οποιοδήποτε*, ακόμη και πολυτροπικό και τελείως ανώμαλο και δεν πρέπει να συγχέεται με το σχήμα της κατανομής δειγματοληψίας. Αλλά ο,τι και αν είναι αυτό, τα μ και σ μπορούμε να τα βρούμε από τις παραπάνω σχέσεις.

Το ΚΟΘ εγγυάται ότι η διωνυμική κατανομή είναι περίπου κανονική για μεγάλο *N* και *p* όχι πολύ κοντά στο 0 ή τη μονάδα.

Το ΚΟΘ εξηγεί επίσης γιατί η κανονική κατανομή απαντάται τόσο συχνά στη φύση και στις παρατηρήσεις που κάνουμε: κάθε μέτρηση είναι αποτέλεσμα πολλών τυχαίων παραγόντων, δηλαδή είναι η μέση τιμή από ένα μεγάλο δείγμα της φυσικής πραγματικότητας και επομένως κατανέμεται σύμφωνα με το ΚΟΘ

Εφαρμογή: διόρθωση στον τύπο της διασποράς για μικρά δείγματα.

Ο,τι μορφή κι αν έχει μια κατανομή, βρίσκουμε σχεδόν πάντα ότι οι περισσότερες τιμές είναι συγκεντρωμένες σε κάποιο διάστημα (α, β) και λίγες τιμές υπάρχουν στις περιοχές x<α και x>β, οι οποίες είναι και οι λεγόμενες "ουρές" της κατανομής. Έτσι, όσο πιο μικρό δείγμα παίρνουμε από ένα πληθυσμό, τόσο πιο λίγες είναι οι ευκαιρίες να κάνουμε δειγματοληψία από τις ουρές της κατανομής με αποτέλεσμα η εκτίμησή μας για τη διασπορά να είναι μικρότερη από την πραγματική.

Αν ο πληθυσμός έχει μέση τιμή μ και διασπορά σ^2, η απόκλιση μιας τιμής x από τη μέση τιμή του πληθυσμού θα είναι:

x – μ = x - $\bar{x}$ + $\bar{x}$ - μ = x - $\bar{x}$ – d, όπου d = μ - $\bar{x}$, η απόκλιση του μέσου του δείγματος από τη μέση τιμή του πληθυμού. Τότε, το άθροισμα από όλο το δείγμα θα δίνει:

$$\sum(x-\mu)^2=\sum(x-\bar{x})^2-\sum 2d(x-\bar{x})+nd^2$$

και επειδή το άθροισμα $\sum 2d(x-\bar{x})$ θα πρέπει να είναι ταυτοτικά ίσο με μηδέν, όπως έχουμε δει στη συζήτηση για την τυπική απόκλιση, βρίσκουμε ότι:

$$\sum(x-\mu)^2=n\sigma^2=\sum(x-\bar{x})^2+nd^2$$

Επίσης,

$nd^2=\sum d^2=\sum(\mu-\bar{x})^2$, άρα η διασπορά των μέσων των δειγμάτων θα είναι $\frac{\sum(\mu-\bar{x})^2}{n}$, το οποίο είναι σ^2/n, βάσει του ΚΟΘ, άρα $\sum(\mu-\bar{x})^2=\sigma^2$ και $\sigma^2=nd^2$, οπότε παίρνουμε:

$n\sigma^2=\sum(x-\bar{x})^2+\sigma^2$ και $\sigma^2(n-1)=\sum(x-\bar{x})^2$ με τελικό αποτέλεσμα:

$$\sigma^2=\frac{\sum(x-\bar{x})^2}{n-1}$$

Βλέπουμε ότι η διόρθωση συνίσταται στην αλλαγή του παρονομαστή από n σε n-1. Η αλλαγή έγινε με βάση το ΚΟΘ που μας επέτρεψε να εκφράσουμε ποσοτικά το σφάλμα εκτίμησης λόγω διασποράς των μέσων τιμών και άρα και των αντίστοιχων ακραίων τιμών των δειγμάτων που δημιουργούν το "πρόβλημα". Για μεγάλα δείγματα, η διαφορά μεταξύ n και n-1 είναι αμελητέα. Στην πράξη, δείγματα με n > 30 θεωρούνται "μεγάλα" και δεν έχει σημασία ποια έκφραση θα χρησιμοποιήσουμε.

Παράρτημα:

Α. Σκαρίφημα απόδειξης της μορφής του διωνυμικού συντελεστή, από συνδυαστικής άποψης.

1) Αν έχουμε *X* διαφορετικά αντικείμενα, τότε οι διαφορετικοί τρόποι διάταξής τους είναι *X*! Πράγματι, το πρώτο μπορεί να είναι οποιοδήποτε από τα *X*, το δεύτερο οποιοδήποτε από *X*-1 αφού το πρώτο τοποθετήθηκε ήδη κ.ο.κ., άρα οι δυνατοί συνδυασμοί είναι *X* (*X*-1) (*X*-2) ... 3 . 2 . 1 = *X*!

2) Τώρα, αν έχουμε *N* αντικείμενα με τα οποία μπορούμε να συμπληρώσουμε *X* < *N* κενές θέσεις, μπορούμε με παρόμοιο συλλογισμό να πούμε: η πρώτη θέση έχει *N* επιλογές, η δεύτερη *N*-1 κλπ κλπ μέχρι την *X*-στή που έχει *N*-*X*+1, άρα οι δυνατοί συνδυασμοί είναι

N (*N*-1) (*N*-2) ... (*N*-*X*+1) = *N* (*N*-1) (*N*-2) ... 3 . 2 . 1 / (*N*-*X*)(*N*-*X*-1)... 3 . 2 . 1 = *N*!/(*N*-*X*)!

3) Τα *X* από τα *N* αντικείμενα που τοποθετήσαμε θεωρούμενα ως ξεχωριστό σύνολο, μπορούν να κάνουν *X*! συνδυασμούς όπως δείξαμε στην αρχή.

4) Αν τα *N* αντικείμενα είναι όλα όμοια μεταξύ τους, τότε οι *X*! συνδυασμοί δε διακρίνονται μεταξύ τους (δεν έχει σημασία ποιο αντικείμενο θα πάει πού) οπότε οι τρόποι συμπλήρωσης *X* θέσεων από *N* όμοια αντικείμενα θα είναι *X*! φορές λιγότεροι από ο,τι αν τα αντικείμενα ήταν εν γένει διαφορετικά, δηλαδή *N*!/*X*!(*N*-*X*)! ό. έ. δ.

Β. Μέση τιμή και διασπορά διωνυμικής κατανομής

Αν ένας πληθυσμός έχει κατανομή πιθανότητας p, τότε η μέση τιμή αυτής είναι $\sum_i p\,x_i$ (για διακριτά δεδομένα). Αυτή τη γενική σχέση θα εφαρμόσουμε για να βρούμε τη μέση τιμή όταν ισχύει η διωνυμική κατανομή. Κατ' αρχήν, η διωνυμική είναι κατανομή πιθανοτήτων γιατί είναι πάντα μή αρνητική και έχει άθροισμα ίσο με τη μονάδα, όπως φαίνεται εύκολα:

$$\sum_x^n P(n,x)=\sum_x^n \binom{n}{x} p^x q^{n-x}=(p+q)^n=1$$

Για τη μέση τιμή, γράφουμε:

$$\mu=\sum_{x=0}^{n} x\,P(x,n)=\sum_{x=0}^{n} x\binom{n}{x}p^x q^{n-x}=\sum_{x=0}^{n} x\frac{n!}{x!(n-x)!}p^x q^{n-x}=$$
$$=np\sum_{x=1}^{n} x\frac{(n-1)!}{(x-1)!(n-x)!}p^x q^{n-x}=np\sum_{x=1}^{n} x\binom{n-1}{x-1}p^{x-1}q^{n-x}$$

Θέτουμε x-1 = r και η παραπάνω σχέση γίνεται

$$\mu=np\sum_{r=0}^{n-1} x\binom{n-1}{r}p^r q^{n-1-r}=np(p+q)^{n-1}=np$$

Για τη διασπορά, που είναι το τετράγωνο της τυπικής απόκλισης, θα χρησιμοποιήσουμε τη σχέση: σ = $\langle X^2\rangle$-$\langle X\rangle^2$, δηλαδή τη διαφορά του τετραγώνου της μέσης τιμής (την οποία υπολογίσαμε) από τη μέση τιμή του τετραγώνου, την οποία πρέπει να βρούμε. Αυτή θα τη βρούμε μέσω της παράστασης X(X-1) για την οποία είναι πιο εύκολο να βρούμε τη μέση τιμή, ως εξής:

$$\langle X(X-1)\rangle=\sum_{x=0}^{n} x(x-1)P(x,n)=\sum_{x=0}^{n} x(x-1)\binom{n}{x}p^x q^{n-x}=\sum_{x=0}^{n} x(x-1)\frac{n!}{x!(n-x)!}p^x q^{n-x}=$$
$$=n(n-1)p^2\sum_{x=0}^{n=2}\frac{(n-2)!}{(x-2)!(n-x)!}p^{x-2}q^{n-x}=n(n-1)p^2\sum_{x=2}^{n}\binom{n-2}{x-2}p^{x-2}q^{n-x}$$

η οποία με το μετασχηματισμό r=x-2 δίνει

$$\langle X(X-1)\rangle=n(n-1)p^2\sum_{r=0}^{n-2}\binom{n-2}{x-2}p^r q^{n-2-r}=n(n-1)p^2(p+q)6n-2=n(n-1)p^2$$

από όπου προκύπτει ότι

$$\langle X^2\rangle=\langle X\rangle+n(n-1)p^2=np+n(n-1)p^2$$

Επομένως,

$$\sigma^2=\langle X^2\rangle-\langle X\rangle^2=np+n(n-1)p^2-n^2p^2=npq$$

Γ. Βασικά στάδια απόδειξης ότι η κανονική κατανομή είναι οριακή μορφή της διωνυμικής.

Η κανονική κατανομή είναι μια συνεχής συνάρτηση ενώ η διωνυμική κατανομή είναι μια ακολουθία η οποία μάλιστα περιέχει τους λεγόμενους διωνυμικούς συντελεστές με διάφορα παραγοντικά. Η πρώτη είναι μια ικανοποιητική προσέγγιση της δεύτερης για μεγάλο αριθμό παρατηρήσεων, n.

Για να βρούμε τη δεύτερη ως προσεγγιστική έκφραση της πρώτης, θα αρχίσουμε αντικαθιστώντας τα παραγοντικά με εκφράσεις που μπορούμε να χειριστούμε πιο εύκολα. Θα χρησιμοποιήσουμε τη σχέση του Stirling η οποία δίνει μια προσεγγιστική έκφραση του n! για μεγάλες τιμές του n. $n!=\sqrt{2\pi n}\,n^n e^{-n}$.

Επίσης, θα χρησιμοποιήσουμε τις εξής ποσότητες:

- μέση τιμή μ = np
- τυπική απόκλιση σ = $(npq)^{1/2}$
- ανηγμένη μεταβλητή z = $(x-\mu)/\sigma$

Εφαρμόζοντας τη σχέση του Stirling, από τη σχέση

$$P(x,n)=\frac{n!}{x!(n-x)!}p^x q^{n-x}$$

περνάμε, μετά από πράξεις, στην έκφραση

$$P(x,n)\approx\frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{x(n-x)}}\left(\frac{np}{x}\right)^{x}\left(\frac{nq}{n-x}\right)^{n-x}$$

Με τη βοήθεια της ανηγμένης μεταβλητής, z, μπορούμε να μετασχηματίσουμε την ποσότητα n/x(n-x) και να ξαναγράψουμε την ανωτέρω έκφραση ως:

$$P(x,n)=\frac{1}{\sqrt{2\pi}\sigma}\left(\frac{np}{x}\right)^{x}\left(\frac{nq}{n-x}\right)^{n-x}$$

Από εδώ μπορούμε να χρησιμοποιήσουμε λογαρίθμους για να πάρουμε την έκφραση

$$\ln\left[\sqrt{2\pi}\sigma P(x,n)\right]=\ln\left[\left(\frac{np}{x}\right)^{x}\left(\frac{nq}{n-x}\right)^{n-x}\right]$$

Είναι καλύτερα όμως τα κλάσματα στο δεξί μέλος να τα αντιστρέψουμε γιατί αυτό θα διευκολύνει τις επόμενες πράξεις. Αυτό γίνεται αλλάζοντας το πρόσημο στους λογαρίθμους:

$$-\ln\left[\sqrt{2\pi}\sigma P(x,n)\right]=\ln\left[\left(\frac{x}{np}\right)^{x}\left(\frac{n-x}{nq}\right)^{n-x}\right]$$

Συνδυάζοντας τον ορισμό της ανηγμένης μεταβλητής με τις τιμές της μέσης τιμής και της τυπικής απόκλισης που δώσαμε αρχικά και εφαρμόζοντας βασικές ιδιότητες των λογαρίθμων, παίρνουμε:

$$-\ln\left[\sqrt{2\pi}\sigma P(x,n)\right]=x\ln\left(1+z\frac{\sigma}{pq}\right)+(n-x)\ln\left(1-z\frac{\sigma}{nq}\right)$$

Σε αυτό το σημείο, κάνουμε την παραδοχή ότι οι τιμές x δεν απέχουν πολύ από τη μέση τιμή, άρα οι ανηγμένες τιμές z δε διαφέρουν πολύ από το μηδέν, οπότε μπορούμε να εφαρμόζουμε στους ανωτέρω λογαρίθμους, την προσεγγιστική σχέση $\ln(1+a)\approx a-\frac{a^2}{2}$ που ισχύει για τιμές του *a* κοντά στο μηδέν. Αντικαθιστώντας, στην τελευταία σχέση, μετά από αρκετές πράξεις όπου χρησιμοποιούμε τους ορισμούς των μ και σ, καθώς και τη σχέση p+q=1, βρίσκουμε:

$$-\ln\left[\sqrt{2\pi}\sigma P(z)\right]=\frac{z^2}{2}$$

από όπου βρίσκουμε αμέσως την έκφραση για την τυπική κανονική κατανομή :

$$P(z)=\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-z^2}{2}}$$