**ΚΕΦΑΛΑΙΟ 3 – ΚΑΝΟΝΙΚΗ ΚΑΤΑΝΟΜΗ**

Η κατανομή πολλών συνεχών ποσοτικών χαρακτηριστικών σε ένα πληθυσμό παρουσιάζει συχνά την παρακάτω συμμετρική μορφή



Στο παρακάτω σχήμα βλέπουμε πως ένα ιστόγραμμα που αφορά ένα ποσοτικό χαρακτηριστικό, μπορεί να προσεγγιστεί με μια καμπύλη αυτής της μορφής.



**Παρατήρηση**

Ακόμα και αν η μορφή της κατανομής του χαρακτηριστικού που μελετάμε σε ένα πληθυσμό, δεν προσεγγίζεται πολύ καλά από την κανονική καμπύλη, υπάρχουν τρόποι (μετασχηματισμοί δεδομένων), να γίνει εφικτή αυτή η προσέγγιση.

Το ιστόγραμμα αναφέρεται στο ΔΕΙΓΜΑ, που μελετήσαμε ενώ η καμπύλη θεωρητικά αναφέρεται στον ΠΛΗΘΥΣΜΟ μας ολόκληρο.

Η παραπάνω συμμετρική καμπύλη, που μοιάζει με «καμπάνα» ονομάζεται **Κανονική Καμπύλη** ή **Κανονική Κατανομή** ή **Καμπύλη του Gauss** (Normal Distribution). Θεωρητικά αυτή η καμπάνα επεκτείνεται από το -∞ μέχρι το +∞, όμως πρακτικά θα δούμε ότι σύντομα «αγγίζει» τον άξονα των χ, φτάνει δηλαδή πολύ κοντά σε αυτόν.

Χρειαζόμαστε ΔΥΟ στοιχεία για να περιγράψουμε την κανονική κατανομή:

α) τη μέση τιμή (*μ ή* $\overbar{Χ}$)

β) την τυπική απόκλιση (*σ ή S*)

Όταν θέλουμε να περιγράψουμε ένα χαρακτηριστικό με την κανονική καμπύλη, είναι φανερό ότι δεν γνωρίζουμε εκ των προτέρων τη μέση τιμή και την τυπική απόκλιση. Παίρνουμε λοιπόν ένα δείγμα από τον πληθυσμό, εκτιμούμε τις ποσότητες αυτές και καθορίζουμε την καμπύλη με βάση τις εκτιμήσεις των *μ* και *σ*.

Δίνουμε επιπλέον τον μαθηματικό τύπο με τον οποίο κατασκευάζεται η κανονική καμπύλη

Ο τύπος περιλαμβάνει την μέση τιμή *μ*, την τυπική απόκλιση *σ*, και δύο πολύ γνωστές μαθηματικές σταθερές, το *π* (≈3,14) και το *e* (≈2.71).

H Ειδική περίπτωση όπου το *μ*=0 και το *σ*=1, ονομάζεται **Τυπική Κανονική Κατανομή**.

Στο παρακάτω σχήμα, βλέπουμε κανονικές κατανομές με διαφορετικές μέσες τιμές και ίδιες τυπικές αποκλίσεις.

Στο παρακάτω σχήμα, βλέπουμε κανονικές κατανομές με διαφορετικές τυπικές αποκλίσεις και ίδιες μέσες τιμές.

**Παράδειγμα:**

Η κατανομή του ζαχάρου σε ένα πληθυσμό ανδρών πάνω από 50 ετών, ακολουθεί την κανονική κατανομή με μέση τιμή 155 mg/dL και τυπική απόκλιση 27 mg/dL. H γραφική παράσταση δίνεται στο σχήμα

**Η σημασία της μέσης τιμής (μ ή** $\overbar{Χ}$**).**

Η μέση τιμή στην συμμετρική αυτή καμπύλη είναι επιπλέον και Διάμεσος, δηλαδή στο παραπάνω παράδειγμα μπορούμε να πούμε ότι 50% των ατόμων έχουν ζάχαρο λιγότερο από 155 mg/dL και 50% περισσότερο από το όριο αυτό.

Επίσης είναι και επικρατούσα τιμή, που σημαίνει ότι τιμές γύρω στα 155 mg/dL, είναι αυτές που εμφανίζονται με μεγαλύτερες συχνότητες στον πληθυσμό.

**Η σημασία της τυπικής απόκλισης (σ ή *S*).**

Η τυπική απόκλιση μας δείχνει γενικά πόσο κοντά στη μέση τιμή, είναι οι υπόλοιπες τιμές που εμφανίζονται στον πληθυσμό μας. Όμως η τυπική απόκλιση, μπορεί να μας δώσει επιπλέον πληροφορίες αρκετά σημαντικές, όπως οι παρακάτω.

α) Σε απόσταση ΜΙΑ τυπική απόκλιση από τη μέση τιμή, υπάρχει περίπου το 68% των περιπτώσεων (δηλαδή μεταξύ του μ-σ και του μ+σ)

β) Σε απόσταση ΔΥΟ τυπικών αποκλίσεων από τη μέση τιμή, υπάρχει περίπου το 95% των περιπτώσεων (δηλαδή μεταξύ του μ-2σ και του μ+2σ)

και

γ) Σε απόσταση ΤΡΙΩΝ τυπικών αποκλίσεων από τη μέση τιμή, υπάρχει περίπου το 99,7% των περιπτώσεων (δηλαδή μεταξύ του μ-3σ και του μ+3σ)

Σχηματικά





**Παράδειγμα:**

Στο παράδειγμα της κατανομής τους ζαχάρου στον πληθυσμό έχουμε:

α) Μία τυπική απόκλιση: 155-27 = 128 / 155+27 = 182. Συνεπώς μεταξύ 128 mg/dl και 182 mg/dl υπάρχει το 68% των ανδρών του παραπάνω πληθυσμού

β) Δύο τυπικές αποκλίσεις: 155-2\*27 = 101 / 155+2\*27 = 209. Συνεπώς μεταξύ 101 mg/dl και 209 mg/dl υπάρχει το 95% των ανδρών του παραπάνω πληθυσμού

γ) Τρείς τυπικές αποκλίσεις: 155-3\*27 = 74 / 155+3\*27 = 236. Συνεπώς μεταξύ 74 mg/dl και 236 mg/dl υπάρχει το 99,7% των ανδρών του παραπάνω πληθυσμού

**ΚΕΦΑΛΑΙΟ 4 – ΔΙΑΣΤΗΜΑΤΑ ΕΜΠΙΣΤΟΣΥΝΗΣ**

Στην Στατιστική ξεκινάμε πάντα με ένα δείγμα από κάποιον πληθυσμό και στη συνέχεια προσπαθούμε να εκτιμήσουμε (προβλέψουμε) διάφορες παραμέτρους του πληθυσμού, από τις αντίστοιχες παραμέτρους του δείγματος.

**4.1 Διαστήματα Εμπιστοσύνης για τη μέση τιμή ενός ποσοτικού χαρακτηριστικού (συνεχούς)**

Έτσι για να εκτιμήσουμε την μέση τιμή (μ) του πληθυσμού, την οποία δεν γνωρίζουμε, παίρνουμε ένα δείγμα n μετρήσεων x1, x2,…xn και υπολογίζουμε τη μέση τιμή του δείγματος ($\overbar{Χ}$ )

Όμως στην περίπτωση αυτή, τι «εμπιστοσύνη» έχουμε στα αποτελέσματα μας τα οποία βασίζονται στο ΔΕΙΓΜΑ; Τι «εμπιστοσύνη» έχουμε, ότι η μέση τιμή που υπολογίσαμε στο Δείγμα, βρίσκεται κοντά στη μέση τιμή του πληθυσμού; Πόσο κοντά είναι το $\overbar{Χ}$ από το δείγμα στο μ του πληθυσμού;

Δυστυχώς ΔΕΝ ΜΠΟΡΟΥΜΕ να είμαστε «σίγουροι» 100% αν δεν μετρήσουμε ΟΛΟΚΛΗΡΟ τον πληθυσμό, πράγμα που είναι φυσικά αδύνατον.

Μπορούμε όμως να υπολογίσουμε το μέγιστο δυνατό λάθος (σφάλμα) που κάνουμε σε μια τέτοια εκτίμηση. Αυτό ονομάζεται Τυπικό Σφάλμα της μέσης τιμής, συμβολίζεται με $SE(\overbar{Χ})$ η απλά με *SE* (Standard Error).

To τυπικό σφάλμα της μέσης τιμής εξαρτάται από δύο παραμέτρους:

1. είναι ανάλογο της τυπικής απόκλισης *S* (του δείγματος ή του πληθυσμού, αν είναι γνωστή βέβαια). Αυτό σημαίνει ότι όταν ένας πληθυσμός παρουσιάζει μεγάλη διασπορά, το σφάλμα της εκτίμησης της μέσης τιμής είναι και αυτό μεγάλο.
2. είναι αντιστρόφως ανάλογο με το μέγεθος του δείγματος *n*. Δηλαδή όσο πιο μεγάλο δείγμα παίρνουμε, τόσο πιο μικρό είναι το τυπικό σφάλμα, συνεπώς και η εκτίμηση της μέσης τιμής του πληθυσμού μας είναι καλύτερη (πιο ακριβής)

**Τυπικό Σφάλμα της Μέσης Τιμής**

Το τυπικό σφάλμα, αποδεικνύεται ότι μπορεί να υπολογιστεί από τον τύπο

$$SE\left(\overbar{X}\right)=\frac{S}{\sqrt{n}}$$

δηλαδή το τυπικό σφάλμα της μέσης τιμής είναι η τυπική απόκλιση αφού διαιρεθεί με την τετραγωνική ρίζα του *n*.

Όταν γίνεται λοιπόν η εκτίμηση της μέσης τιμής ενός πληθυσμού (μ) με έναν απλό αριθμό ($\overbar{Χ}$), η εκτίμηση αυτή εμπεριέχει κάποιο σφάλμα. Για το λόγο αυτό είναι καλύτερα να περικλείουμε την εκτίμηση της μέσης τιμής, μέσα σε ένα διάστημα (Α,Β). Π.χ. αν έχουμε ένα δείγμα από πληθυσμό διαβητικών και θέλουμε να εκτιμήσουμε τη μέση τιμή της γλυκόζης, μπορούμε να αναφέρουμε απλά τη μέση τιμή του δείγματος που βρήκαμε, π.χ. $\overbar{Χ}=173 mg/dL$. Επειδή όμως δεν γνωρίζουμε πόσο καλή είναι η εκτίμηση αυτή, είναι προτιμότερο να αναφέρουμε ότι η μέση τιμή βρίσκεται «πιθανότατα» μέσα σε ένα διάστημα, π.χ. (170 mg/dL , 176 mg/dL) ή αλλιώς μπορούμε να γράψουμε ότι η μέση τιμή είναι 173±3 mg/dL .

Η λέξη «πιθανότατα», που γράψαμε παραπάνω είναι η έννοια της «εμπιστοσύνης» που έχουμε για το αποτέλεσμα μας. Η «εμπιστοσύνη» αυτή, στην Στατιστική είναι αρκετή όταν είναι της τάξης του 95%, για το λόγο αυτό το παραπάνω Διάστημα θα ονομάζεται «95%-Διάστημα Εμπιστοσύνης για τη μέση τιμή» (95%-Confidence Interval). H επιλογή του εύρους του Διαστήματος, ΔΕΝ είναι αυθαίρετη, και βασίζεται σε έναν απλό μαθηματικό τύπο.

**95% - Διάστημα Εμπιστοσύνης για τη Μέση Τιμή μ:**

Προσεγγιστικά το Διάστημα υπολογίζεται με τον παρακάτω τύπο

$$\overbar{Χ}\pm 2∙SE(\overbar{X})$$

ή με αν το γράψουμε ως διάστημα, με κάτω και πάνω άκρο (Α,Β)

$$(\overbar{Χ}-2∙SE\left(\overbar{X}\right), \overbar{Χ}+2∙SE\left(\overbar{X}\right))$$

**Ερμηνεία του Διαστήματος:**

Με τον τρόπο αυτό της εκτίμησης μέσω διαστήματος εμπιστοσύνης, έχουμε 95% Εμπιστοσύνη, ότι η πραγματική μέση τιμή του χαρακτηριστικού στον πληθυσμό που μελετήσαμε, βρίσκεται κάπου ανάμεσα στα δύο αυτά άκρα.

Αυτό σημαίνει ότι υπάρχει μια πιθανότητα να έχουμε κάνει λάθος και η μέση τιμή του πληθυσμού, να μην βρίσκεται μέσα εκεί, αλλά η πιθανότητα αυτή είναι τελικά μικρότερη από 5%.

**Παρατηρήσεις:**

1. Ο παραπάνω τύπος του Διαστήματος είναι προσεγγιστικός και δίνει καλύτερα αποτελέσματα, όταν το μέγεθος δείγματος είναι αρκετά μεγάλο (μεγαλύτερο από 30 συνήθως). Σε περιπτώσεις μικρών δειγμάτων, ο τύπος χρειάζεται κάποιας μορφής διόρθωση, που είναι θέμα των σημειώσεων αυτών.
2. Η τυπική απόκλιση (*S*) αναφέρεται στο χαρακτηριστικό που μετρήσαμε (στο δείγμα ή σε ολόκληρο τον πληθυσμό)
3. Το τυπικό σφάλμα (SE) αναφέρεται στη μέση τιμή του χαρακτηριστικού. Μας δείχνει πόσο κοντά στην μέση τιμή του πληθυσμού (*μ*), μπορεί να βρίσκεται η εκτίμησή μας, με άλλα λόγια μας δείχνει πόσο καλή εκτίμηση μπορούμε να έχουμε κάνει.
4. Για την τυπική απόκλιση *S*, γνωρίζουμε ότι στο διάστημα

$$\overbar{Χ}\pm 2∙S$$

μας δείχνει ανάμεσα σε ποιες τιμές βρίσκεται το 95% των παρατηρήσεων μας, ενώ το διάστημα

$$\overbar{Χ}\pm 2∙SE(\overbar{X})$$

μας δείχνει (με 95% εμπιστοσύνη) ανάμεσα σε ποιες τιμές βρίσκεται η ΜΕΣΗ ΤΙΜΗ του πληθυσμού.

**Παράδειγμα**

Αν η μέση τιμή ενός δείγματος με 60 μετρήσεις, βρέθηκε 7,2 και η τυπική απόκλιση βρέθηκε 1,6, τότε το τυπικό σφάλμα της μέσης τιμής είναι



Οπότε το 95% διάστημα εμπιστοσύνης για τη μέση τιμή γίνεται



Εκτελώντας την πράξη, χωριστά για το + και για το -, βρίσκουμε το κάτω και το πάνω άκρο του διαστήματος (6,788 και 7,612), οπότε τελικά έχουμε 95% εμπιστοσύνη ότι η πραγματική μέση τιμή του πληθυσμού που μελετήσαμε είναι μεταξύ

 6,788 και 7,612

**4.2 Διαστήματα Εμπιστοσύνης για ένα ποσοστό (αναφέρονται σε ποιοτικά χαρακτηριστικά συνήθως)**

Όταν έχουμε κάποιο ποιοτικό χαρακτηριστικό, συνήθως υπολογίζουμε σχετικές συχνότητες από τις διάφορες κατηγορίες του χαρακτηριστικού. Π.χ. αν μελετήσαμε τις ομάδες αίματος σε κάποιο πληθυσμό, με ένα δείγμα 550 ατόμων, μπορεί να βρήκαμε ότι στην Ομάδα Α, ανήκαν 175 άτομα. Η σχετική συχνότητα (ποσοστό) της Ομάδας Α, είναι

$p=\frac{175}{550}=0,318$ ή 31,8%

Από το ποσοστό αυτό, που βρήκαμε στο δείγμα, μπορούμε να βγάλουμε συμπέρασμα για το ποσοστό της Ομάδας Α, σε ολόκληρο τον πληθυσμό. Στην περίπτωση όμως αυτή, όπως και κατά την εκτίμηση της μέσης τιμής, υπάρχει περίπτωση να έχουμε κάνει κάποιο σφάλμα. Για το λόγο αυτό μπορούμε να κλείσουμε την εκτίμησης του ποσοστού, μέσα σε ένα Διάστημα Εμπιστοσύνης, παρόμοιο με αυτό της μέσης τιμής.

**Τυπικό Σφάλμα ενός Ποσοστού**

Το τυπικό σφάλμα, αποδεικνύεται ότι μπορεί να υπολογιστεί από τον τύπο

$$SE\left(p\right)=\sqrt{\frac{p∙(1-p)}{n}}$$

Το τυπικό σφάλμα του ποσοστού, εξαρτάται πολύ από το μέγεθος δείγματος (n), που σημαίνει ότι όσο πιο μεγάλο δείγμα έχουμε, τόσο μικρότερο γίνεται το σφάλμα και άρα έχουμε καλύτερη εκτίμηση του ποσοστού.

**95% - Διάστημα Εμπιστοσύνης για ένα ποσοστό:**

Δίνεται από τον τύπο (προσεγγιστικά για μεγάλα δείγματα)



ή με αν το γράψουμε ως διάστημα, με κάτω και πάνω άκρο (Α,Β)



**Ερμηνεία του Διαστήματος:**

Με τον τρόπο αυτό της εκτίμησης μέσω διαστήματος εμπιστοσύνης, έχουμε 95% Εμπιστοσύνη, ότι το πραγματικό ποσοστό της κατηγορίας που μας ενδιαφέρει στον πληθυσμό που μελετήσαμε, βρίσκεται κάπου ανάμεσα στα δύο αυτά άκρα.

Αυτό σημαίνει ότι υπάρχει μια πιθανότητα να έχουμε κάνει λάθος και το πραγματικό ποσοστό να μην βρίσκεται μέσα εκεί, αλλά η πιθανότητα αυτή είναι τελικά μικρότερη από 5%.

**Παράδειγμα:**

Σε κάποια δημοσκόπηση πριν τις εκλογές ερωτήθηκαν 585 ψηφοφόροι και δήλωσαν ότι ψηφίζουν το κόμμα Α οι 213.

Έχουμε

 (ή 36,4 %)

Το τυπικό σφάλμα της εκτίμησης του ποσοστού είναι



Οπότε τα όρια εμπιστοσύνης για το ποσοστό που θα πάρει το κόμμα Α στις

εκλογές είναι



Δηλ. 0,324 μέχρι 0,404. (ή 32,4% μέχρι 40,4% αν θέλουμε να το αναφέρουμε επί τοις εκατό)

**Παρατήρηση:**

Το Διάστημα Εμπιστοσύνης για ένα ποσοστό εξαρτάται κατά κανόνα από το μέγεθος δείγματος, αφού ο αριθμητής δεν μπορεί να γίνει μεγαλύτερος από

$$\sqrt{0,50∙0,50}=\sqrt{0,25}=0,50$$

Αυτό συμβαίνει γιατί μπορεί να αποδειχτεί ότι η μέγιστη τιμή του γινομένου $p∙(1-p)$ επιτυγχάνεται όταν το *p* είναι ακριβώς 0,5 δηλαδή 50%.

Συνήθως για να έχουμε μικρό Διάστημα Εμπιστοσύνης για ένα ποσοστό (να μην υπερβαίνει π.χ. το ±2%), πρέπει να πάρουμε πολύ μεγάλο δείγμα, πάνω από 1000 ή 2000 άτομα