3. Στατιστική Συμπερασματολογία

Συμπερασματολογία είναι ο κλάδος εκείνος της Λογικής που ασχολείταο με την εξαγωγή συμπερασμάτων. Τα συμπεράσματα εκείνα μπορει να αναφέρονται στο πλαίσιο της απόφασης ή πρόβλεψης. Μερικά παραδείγματα: Ο ερευνητής θέλει να αποδείξει την θεωρία του, ο γιατρός ενδιαφέρεται να δείξει ποιες απο τις δύο θεραπείες είναι η καλύτερη. Για τα παραπάνω χρειάζεται η βοήθεια των σχετικών πληροφοριών για την εξαγωγή συμπερασμάτων που λέγεται δεδομένα ή παρατηρήσεις. Η στατιστική συμπερασματολογία έχει επαγωγικό χαρακτήρα, δηλαδή γενικεύει απο ένα δείγμα για ολόκληρο τον πληθυσμό. Εξετείας αυτού εμπεριέχει και μια σχετική αβεβαιότητα η οποία μετριέται υπο την μορφή της πιθανότητας. Σκοπός της στατιστικής ειναι η μελετη μεθόδων Συμπερασματολογία και τρόπων μέτρησης της αβεβαιότητας αυτής. Επομένως για κάθε εφαρμογή της στατιστικής συμπερασματολογίας εμπεριέχονται δύο στοιχεία, το ένα εκ των οποίων ειναι το συμπέρασμα και το δεύτερο είναι το μέτρο ορθότητας ή το μέτρο καταλληλότητας του.

3.1 Εκτίμηση Παραμέτρων

3.1.1 Γενικά

Έστω $X_1, X_2, \dots, X_n$ τυχαίο δείγμα απο κάποιον πληθυσμό με μία άγωστη παράμετρο. Το πρόβλημα μας είναι να βρούμε μια ή περισσότερες ποσότητες που θα χρησιμοποιηθουν για την εκτίμηση της άγνωστης παραμέτρου. Προφανές είναι ότι οι ποσότητες αυτές θα πρέπει να προέρχονται από το δείγμα. Ένα παράδειγμα είναι όταν θέλουμε να εκτιμήσουμε την μέση τιμή μ ενός πληθυσμού X, μπορούμε να χρησιμοποιήσουμε την ποσότητα $\overline{X}$. Η ποσότητα αυτή λέγεται εκτιμήτρια συνάρτηση και η αριθμητική τιμή της προκύπτει από το δείγμα εκτιμητής της παραμέτρου. ***Οι εκτιμήτριες συναρτήσεις είναι στατιστικές συναρτήσεις και εξαρτώνται μόνο από το δείγμα και όχι από την παράμετρο που ζητάμε να εκτιμήσουμε. Στην περίπτωση αυτή εκτιμούμε ένα άγνωστο σημείο με ένα γνωστό σημείο****. Η σχετική μεθοδολογία λέγεται* ***εκτιμητική*** *και εδώ συγκεκριμένα* ***εκτίμηση σε σημείο ή σημειοεκτιμητική****.*

Την άγνωστη παράμετρο μπορούμε να την εκτιμήσουμε με διάστημα αντί για σημείο καθώς θα έχει και περισσότερη ακρίβεια στην εκτίμηση της. Θα χρειαστούμε καταρχήν δύο ποσότητες που θα προέρχονται από το δείγμα τα όρια του διαστήματος. Η συγκεκριμένη μεθοδολογία ονομάζεται ***εκτιμητική*** *ή* ***εκτίμηση με διάστημα*** *ή* ***εκτίμηση με διάστημα εμπιστοσύνης****.*

Σε προβλήματα της στατιστικής η κατανομή του πληθυσμού είναι γνωστής μορφής εκτός από το γεγονός ότι περιέχει μια ή περισσότερες άγνωστες παραμέτρους. Στην παραπάνω περίπτωση βρισκόμαστε στην σφαίρα των παραμετρικών στατιστικών μοντέλων, όπου γνωρίζουμε μια ή περισσότερες παραμέτρους εκ των προτέρων (a priori). Στην αντίθετη περίπτωση, όπου δεν γνωρίζουμε την κατανομή του πληθυσμού

τότε έχουμε το Μη-Παραμετρικό στατιστικό μοντέλο. Ένα παράδειγμα είναι ότι η κατανομή ενός πληθυσμού μπορεί να ακολουθεί την Διωνυμική κατανομή $B(n,p)$ με n γνωστό και $p \in (0,1)$ άγνωστο ή κανονική $N(\mu,\sigma^2)$ με $-\infty < \mu < \infty$, $\sigma^2 > 0$ άγνωστα. Όταν έχουμε παρατηρήσεις $x_1, x_2, \dots, x_n$ και δεν γνωρίζω το μοντέλο τους, τότε βρισκόμαστε στο πλαίσιο της μη-παραμετρικής στατιστικής. Γενικά οι άγνωστες παράμετροι θα συμβολίζονται με θ ή θ_1, θ_2, κλπκαι οι κατανομές του πληθυσμού X με $f(x,\theta)$. Το πρόβλημα εστιάζεται στο ότι θέλουμε να εκτιμήσουμε το θ ή συναρτήσεις αυτού $g(\theta)$. Οι εκτιμήτριες συναρτήσεις του θ και θα συμβολίζονται με $\hat{\theta}$ ή πιο γενικά με $T(X_1, X_2, \dots, X_n)$.

3.1.2 Εκτίμηση σε σημείο (Σημειοεκτιμητική)

Ένα διπλό πρόβλημα που έχει να αντιμετωπίσει η σημειοεκτιμητική είναι τα εξής ερωτήματα:

"Ποια είναι τα κριτήρια ή αρχές αξιολόγησης των εκτιμητών;" και "Πώς ορίζουμε τον ¨καλύτερο¨ εκτιμητή μεταξύ των διαφόρων εκτιμητών και ποιες οι μέθοδοι αυτών;". Σε αυτό το σημείο τα κριτήρια αξιολόγησης των εκτιμητών είναι πολλά. Αναφορικά μερικά βασικά κριτήρια αξιολόγησης των εκτιμητών είναι: Η ***Αμεροληψία, Ελάχιστη Διακύμανση, Ακρίβεια, Επάρκεια, Συνέπεια*** *κλπ.*

Αμεροληψία

Ένας εκτιμητής $\hat{\theta}$ μιας παραμέτρου θ λέγεται αμερόληπτος αν η αναμενόμενη τιμή θ για κάθε τιμή της παραμέτρου θ, δηλαδή αν

$$E(\hat{\theta}) = \theta, \qquad \forall \theta \in \Theta$$

Η αμεροληψία ενός εκτιμητή μετρά την ορθότητα της εκτίμησης και εκφράζει την ιδέα ότι σε πολλές επαναλήψεις της δειγματοληψίας μας η μέθοδος εκτίμησης που θα χρησιμοποιούμε θα μας δώσει κατά μέσο όρο την άγνωστη παράμετρο θ. Παράδειγμα αν n=50 οι εκτιμητές

$$\overline{X}_{50} = \frac{1}{50}\sum_{i=1}^{50} X_i \text{ και ο } \overline{X}_{25} = \frac{1}{25}\sum_{i=1}^{25} X_i$$

είναι και οι δυο αμερόληπτοι εκτιμητές του μ, με την διαφορά ότι ο εκτιμητής $\overline{X}_{50}$ είναι καλύτερος από τον $\overline{X}_{25}$ επειδή ο εκτιμητής $\overline{X}_{50}$ χρησιμοποιεί περισσότερη πληροφορία από τον $\overline{X}_{25}$. Εκείνοι λοιπόν που παίζει ρόλο είναι η μεταβλητότητα του εκτιμητή. Αναφορικά η διακύμανση του εκτιμητή $\overline{X}_{50}$ είναι $\sigma^2/50$, ενώ η διακύμανση του εκτιμητή $\overline{X}_{25}$ είναι $\sigma^2/25$. δηλαδή

$$V(\overline{X}_{50}) < V(\overline{X}_{25})$$

αυτά μας οδηγούν στο δεύτερο κριτήριο, εκείνο της Ελάχιστης Διακύμανσης.

Ελάχιστη Διακύμανση

Μεταξύ των αμερόληπτων εκτιμητών προτιμητέος είναι εκείνος που έχει την ***μικρότερη διακύμανση****.*

Η ***ακρίβεια (precision)*** *ενός αμερόληπτου εκτιμητή $\hat{\theta}$ μετριέται συνήθως με την διακύμανση του.*

$$\sigma_{\hat{\theta}}{}^2 = V(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

ή την τυπική απόκλιση του (Τυπικό Σφάλμα)

$$\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})} = \sqrt{E(\hat{\theta} - \theta)^2}$$

Για τον δειγματικό μέσο $\overline{X}$ το τυπικό σφάλμα είναι

$$\sigma_{\overline{X}} = \sqrt{\frac{\sigma^2}{n}}$$

Ένας εκτιμητής λέγεται αμερόληπτος ομοιόμορφα ελάχιστης διακύμανσης (Α.Ο.Ε.Δ) όταν είναι αμερόληπτος και έχει την μικρότερη διακύμανση από κάθε άλλο αμερόληπτο εκτιμητή.

Αν η άγνωστη παράμετρος θ είναι η μέση τιμή του πληθυσμού μ, ο $\overline{X}$ έχει την ιδιότητα της αμεροληψίας, δηλαδή

$$E(\overline{X}) = \mu$$

Και αν ο πληθυσμός είναι κανονικός ο $\overline{X}$ είναι ο αμερόληπτος ομοιόμορφα ελάχιστης διακύμανσης εκτιμητής του μ. Ομοίως αν $X_1, X_2, \dots, X_n$ τυχαίο δείγμα απο κάποιον πληθυσμό με μία άγνωστη παράμετρο σ^2, το στατιστικό S^2είναι αμερόληπτος εκτιμητής του σ^2, δηλαδή

$$E(S^2) = \sigma^2$$

Και ισχύουν πάντα ασχέτως με των τιμών που μπορεί να πάρουν τα μ και σ^2.

3.2 Διαστήματα Εμπιστοσύνης

Έστω $X_1, X_2, \dots, X_n$ τυχαίο δείγμα απο κάποιον πληθυσμό με μία άγνωστη παράμετρο θ. Έστω $L = L(X_1, X_2, \dots, X_n)$ το κάτω όριο (Lower Bound) και $U = U(X_1, X_2, \dots, X_n)$ (Upper Bound) το άνω όριο ενός διαστήματος. Τα όρια αυτά είναι συναρτήσεις των τυχαίων μεταβλητών του δείγματος. Οι τιμές τους είναι συναρτήσεις των τυχαίων μεταβλητών του δείγματος. Οι τιμές τους είναι $l(x_1, x_2, \dots, x_n)$ και $u(x_1, x_2, \dots, x_n)$. Εκτιμούμε το διάστημα (L,U) με τιμή (l,u) για την εκτίμηση του θ. Ζητάμε:

1. *Το διάστημα (L, U) να περιέχει την αληθινή τιμή του θ ένα μεγάλο «ποσοστό φορών» και*
2. *Το διάστημα να έχει όσο το δυνατό μικρότερο μήκος. Το «ποσοστό φορών» που ένα διάστημα (L, U) περιέχει το θ λέγεται* ***βαθμός εμπιστοσύνης.***

Συμβολίζεται με $100(1-\alpha)\%$ *και δείχνει την πιθανότητα το* (L, U) *να περιέχει το θ δηλαδή,*

$$Βαθμός\ Εμπιστοσύνης\ =\ P(L<\theta<U)\ =\ 1-a$$

Το διάστημα (L, U) *λέγεται* ***Διάστημα Εμπιστοσύνης.***

Κατασκευάζουμε ένα $100(1-\alpha)\%$ *διάστημα εμπιστοσύνης για την μέση τιμή ενός κανονικού πληθυσμού με βάση ένα τυχαίο δείγμα με μέγεθος n και με την υπόθεση ότι η διακύμανση* σ^2 *του πληθυσμού ότι είναι γνωστή. Έστω λοιπόν ότι* $\overline{X}$ *η μέση τιμή του δείγματος. Γνωρίζουμε ότι η κατανομή του* $\overline{X} \sim N(\mu, {}^{\sigma^2}/_{n})$ *έτσι για κάθε μ έχουμε ότι*

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Έστω $z_{\alpha/2}$ *το αντίστροφο εκατοστιαίο σημείο της τυπικής κανονικής κατανομής*

$$z_{\alpha/2} = P\big(Z \geq z_{\alpha/2}\big) = \alpha/2$$

Τότε έχουμε ότι

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1-a \ \Leftrightarrow$$

$$P\{\ -z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\ \} = 1-a \ \Leftrightarrow$$

$$P\{\ -\frac{\sigma}{\sqrt{n}} z_{\alpha/2} < \overline{X}-\mu\ < \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\ \} = 1-a \ \Leftrightarrow$$

$$P\{\ \overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} < \mu\ < \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\ \} = 1-a\ .$$

Επομένως το διάστημα (L, U) *με όρια*

$$\boldsymbol{L = \overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}} \text{ και } \boldsymbol{U = \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}}$$

έχουν την ιδιότητα ότι

$$P\{L < \mu\ < U\ \} = 1-a, \qquad \forall \mu$$

Επομένως το Διάστημα εμπιστοσύνης για την μέση τιμή μ από κανονικό πληθυσμό $N(\mu, \sigma^2)$ *με* σ^2 *γνωστό είναι:*

$$\left(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right)$$

ή διαφορετικά

$$\overline{X} \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$$

με βαθμό εμπιστοσύνης $100(1-\alpha)\%$

Παράδειγμα

Ένα διεγερτικό φάρμακο ελέγχεται για την επίδραση του στην πίεση του αίματος. ΟΙ πιέσεις αίματος n=20 ατόμων μετριούνται πριν από την λήψη και μίση ώρα μετά την λήψη και λαμβάνονται οι ακόλουθες διαφορές

7	*6*	*0*	*8*	*-9*	*-4*	*0*	*1*	*9*	*1*
2	*7*	*0*	*6*	*-6*	*-5*	*-1*	*6*	*-2*	*4*

Είναι γνωστό από προηγούμενες μελέτες ότι η πριν και μετά την λήψη φαρμάκου διαφορά πιέσεων ακολουθεί την κανονική κατανομή με γνωστή διακύμανση $\sigma^2 = 25$. *Να κατασκευαστεί ένα 95% διάστημα εμπιστοσύνης για της μέσης διαφοράς μ της πίεσης του αίματος.*

Απάντηση

Επειδή το μοντέλο είναι κανονική κατανομή με γνωστή διακύμανση $\sigma^2 = 25$, *το διάστημα εμπιστοσύνης για το μ θα δίνεται από την σχέση*

$$\overline{X} \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \ (1)$$

Από τα δεδομένα του προβλήματος μας δίνονται

n=20, $\overline{X}$*=0.6 ,* σ^2*=25, σ=5 και 1-α=0.95⇔α=0.05 και από την σχέση* $P(Z \geq z_{0.025}) = 0.05$ *έχουμε από τους πίνακες βρίσκοντας ότι* $z_{0.025}$*=1.96. Επομένως η (1) θα γίνει:*

$$0.6 \pm \frac{5}{\sqrt{20}} z_{0.05/2} = \ 0.6 \pm \frac{5}{\sqrt{20}} z_{0.025} = 0.6 \pm 2.19$$

Επομένως το διάστημα εμπιστοσύνης της μέσης τιμής της διαφοράς της πίεσης πριν και μετά το φάρμακο με βαθμό εμπιστοσύνης 95% είναι το (-1.59 , 2.79). Αυτό πρακτικά σημαίνει ότι ***αν εκτελέσουμε το πείραμα 100 φορές αναμένουμε για την μέση τιμή μ τις 95 η μέση τιμή να βρίσκεται εντός των διαστημάτων αυτών.***

Είπαμε για την έννοια του διαστήματος εμπιστοσύνης και δώσαμε ένα διάστημα τέτοιο με την μέση τιμή μ να είναι άγνωστη και την διακύμανση σ^2 να είναι γνωστή a priori. Ας δούμε μια περίπτωση όπου η μέση τιμή μ και η διακύμανση σ^2 είναι αμφότεροι άγνωστοι a priori. Στην περίπτωση αυτή γνωρίζουμε ότι η τυχαία μεταβλητή

$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

Και είναι και συμμετρική ως προς το μ. Έτσι και σε αυτή την περίπτωση έχουμε τα συμμετρικά σημεία $-t_{n-1,\alpha/2}$ και $t_{n-1,\alpha/2}$ τέτοια ώστε:

$$P\left(-t_{n-1,\alpha/2} \le t \le t_{n-1,\alpha/2}\right) = 1 - a \Leftrightarrow$$

$$P\left(-t_{n-1,\alpha/2} \le \frac{\overline{X} - \mu}{S/\sqrt{n}} \le t_{n-1,\alpha/2}\right) = 1 - a \Leftrightarrow$$

$$P\left(\overline{X} - \frac{S}{\sqrt{n}} t_{n-1,\alpha/2} \le \mu \le \overline{X} + \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}\right) = 1 - a$$

δηλαδή

$$\left(\overline{X} - \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}\, , \overline{X} + \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}\right)$$

$$\overline{X} \pm \frac{S}{\sqrt{n}} t_{n-1,\alpha/2}$$

Με $P\left(t \ge t_{n-1,\alpha/2}\right) = \alpha/2$

Συνοψίζοντας έχουμε:

$$\overline{X} \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \quad , \quad \sigma^2: \gamma\nu\omega\sigma\tau\acute{o}$$

$$\overline{X} \pm \frac{S}{\sqrt{n}} t_{n-1,\alpha/2} \quad , \quad \sigma^2: \acute{\alpha}\gamma\nu\omega\sigma\tau o$$

Στο Παράδειγμα μας αν υποθέταμε πλέον ότι δεν γνωρίζουμε την διακύμανση σ^2, θα είχαμε ότι n=20, $\overline{X}$= 0.6 , S^2= 25.84 και S = 5.08. Για τον υπολογισμό του S θα είχαμε ότι

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2} = 5.08$$

και το 95% διάστημα εμπιστοσύνης θα ήταν το

$$0.6 \pm \frac{5.08}{\sqrt{20}} t_{20-1,0.05/2} \;=\; 0.6 \pm \frac{5.08}{\sqrt{20}} t_{19,0.025} = 0.6 \pm 2.38 = (-1.78\,, 2.98)$$