Μη-παραμετρική Στατιστική (Φροντ. #1) -1-

Εισαγωγή στους Μη-Παραμετρικούς Ελέγχους Υποθέσεων

Αν και σε πολλές περιπτώσεις οι παραμετρικές υποθέσεις μοιάζουν λογικές συχνά δεν έχουμε προηγούμενη γνώση των κατανομών που περιγράφουν τα δεδομένα μας. Σε τέτοιες περιπτώσεις η χρήση της Παραμετρικής Στατιστικής μπορεί να μας οδηγήσει σε λανθασμένα αποτελέσματα και συνεπώς συμπεράσματα.

Σε περιπτώσεις όπου δεν είναι επιτρεπτή οποιαδήποτε υπόθεση για τη μορφή του πληθυσμού ο ερευνητής χρησιμοποιεί το Κ.Ο.Θ. και στηρίζεται στην υπόθεση της κανονικότητας, ώστε να διεξάγει τους ελέγχους που τον ενδιαφέρουν.

Τι γίνεται όμως όταν τα δείγματα που έχουμε συλλέξει είναι μικρά και τα δεδομένα δεν κατανέμονται σύμφωνα με την κανονική κατανομή;

Η βασική ιδέα της μη-Παραμετρικής Στατιστικής **Συμπερασματολογίας** είναι να επάγει συμπεράσματα για άγνωστες ποσότητες ενδιαφέροντος, χρησιμοποιώντας δεδομένα αλλά παράλληλα θεωρώντας το μικρότερο δυνατό σύνολο υποθέσεων.

Προβλήματα υπάρχουν όταν τα δείγματα που έχουμε συλλέξει από τον υπό-μελέτη πληθυσμό είναι μικρά και τα δεδομένα δεν κατανέμονται σύμφωνα με την κανονική κατανομή. Το μικρότερο πλεονέκτημα των μη-παραμετρικών τεχνικών είναι ότι οι έλεγχοι

δεν προϋποθέτουν γνώση της κατανομής του πληθυσμού από τον οποίον έχουν προέλθει τα υπό μελέτη στοιχεία.

Επιπλέον,

- απαιτούν λίγες υποθέσεις για τους πληθυσμούς από τους οποίους προέρχονται τα δεδομένα,
- επιτρέπουν στον ερευνητή να υπολογίσει, σε κάποιες περιπτώσεις ελέγχων την ακριβή τιμή (exact value) του παραπροόμενου επιπέδου σημαντικότητας $\hat{\alpha}$,
- είναι (σχεδόν!) πάντα πιο απλές τεχνικές προς εφαρμογή συγκριτικά με τις αντίστοιχες παραμετρικές τεχνικές,
- συχνά είναι πιο εύκολες στην κατανόηση,
- δεν είναι τόσο ευαίσθητες στις ακραίες τιμές,
- εφαρμόζονται (συχνά) στις τάξεις μεγέθους των δεδομένων και όχι στα ίδια τα δεδομένα,
- μπορούν να χρησιμοποιηθούν και για δεδομένα που είναι ταξινομημένα σε κατηγορίες (κατηγορικά δεδομένα σε κλίμακα διάταξης ή σε ονομαστική κλίμακα).

Όμως:

- σε περίπτωση όπου είναι γνωστό ότι τα δεδομένα προέρχονται από κανονική κατανομή τότε οι παραμετρικοί έλεγχοι είναι πιο <u>ισχυροί</u>.

αλλά:

- αν η κατανομή των δεδομένων δεν είναι κανονική τότε οι μη-παραμετρικοί έλεγχοι έχουν

Μη-παραμετρικοί έλεγχοι για ένα δείγμα

Στο γενικό πρόβλημα για ένα δείγμα, τα διαθέσιμα δεδομένα αποτελούν ένα απλό σύνολο δεδομένων, δηλαδή ένα τυχαίο δείγμα από μια συνάρτηση κατανομής F_X. Η συμπερασματολογία αφορά κάποιο χαρακτηριστικό μέγεθος που σχετίζεται με τη συνάρτηση κατανομής F_X. Στο υποδειγματικό πρόβλημα συμπερασματολογίας για ένα δείγμα, τα δεδομένα χρησιμοποιούνται για να εξαχθούν συμπεράσματα για κάποιο συγκεκριμένο χαρακτηριστικό της κατανομής του πληθυσμού. Οι μη-παραμετρικές τεχνικές είναι χρήσιμες, για παράδειγμα, στην εξαγωγή συμπερασματολογίας ενός μέτρου θέσης για την κατανομή που μας ενδιαφέρει.

Οι υποθέσεις των ελέγχων σχετίζονται με τη διάμεσο που όπως και η μέση τιμή είναι βασικοί δείκτες της κεντρικής τάσης ενός πληθυσμού. Σε κάθε πληθυσμό η διάμεσος πάντοτε υπάρχει και είναι πιο ανθεκτική μετρική από τη μέση τιμή σε σχέση με τη μέση τιμή.

Έλεγχος προσήμου - Προσημικός Έλεγχος (Sign Test)

Έστω ένα τ.δ. N παρατηρήσεων $X_1, X_2, \ldots, X_N$ από ένα πληθυσμό με κατανομή F_X και άγνωστη διάμεσο M όπου F_X θεωρείται συνεχής, τουλάχιστον στην "περιοχή" (στη γειτονιά) της διαμέσου M.

Οι N παρατηρήσεις είναι ανεξάρτητες και ταυτοτικά κατανεμημένες και η διάμεσος $M = F_X^{-1}(0.5)$ μοναδική.

Η υπόθεση που μας ενδιαφέρει να ελέγξουμε αφορά την πληθυσμιακή διάμεσο:

$$H_0: M = M_0$$

όπου M_0 είναι μία συγκεκριμένη τιμή έναντι της εναλλακτικής υπόθεσης είτε μονόπλευρου είτε αμφίπλευρου ελέγχου. Ισοδύναμα, μπορούμε να γράψουμε:

$$H_0: \theta = P(X > M_0) = P(X < M_0) = 0.5$$

Έστω $K := \#$ παρατηρήσεων $> M_0$. Επειδή οι παρατηρήσεις που δείγματος έχουν "διχοτομηθεί" (ως προς την τιμή M_0) σε ίσους ένα σύνολο n ανεξάρτητων τ.μ. από έναν πληθυσμό Bernoulli με παράμετρο $\theta = P(X > M_0)$. Άρα, η δειγματική κατανομή της τ.μ. K είναι διωνυμική με παραμέτρους N και θ. Όταν ισχύει η H_0 (μάλιστα από την H_0) τότε $\theta = 0.5$.

Επειδή ισοδύναμα μπορούμε να γράψουμε ότι:

$$K := \#\text{"+"} \text{ ανάμεσα στις } N \text{ διαφορές } X_i - M_0, \quad i = 1, 2, \ldots, N$$

αυτό το μη-παραμετρικό test που βασίζεται στην ποσότητα K καλείται πρόσημο test (sign test).

Συνεπώς, ελέγχουμε $H_0: M = M_0$ έναντι $H_1: M \neq M_0$ ή $H_1: M > M_0$ ή $H_1: M < M_0$

Για την εναλλακτική,

$$H_1: M > M_0 \text{ ή } \theta = P(X > M_0) > P(X < M_0)$$

περιοχή απόρριψης R (ή κρίσιμη περιοχή) της H_0 προσδιορίζεται από την ανισότητα: $K \geq K_\alpha$ όπου K_α είναι ο ελάχιστος ακέραιος για τον οποίον:

$$P(K \geq K_\alpha \mid H_0) = \sum_{i=K_\alpha}^{N} \binom{N}{i} (0.5)^N \leq \alpha$$

Παρόμοια, για την εναλλακτική: $H_1: M < M_0$

ή $\theta = P(X > M_0) < P(X < M_0)$

η κρίσιμη περιοχή επιπέδου α από την ανισότητα: $K \leq K'_\alpha$ όπου K'_α είναι ο μέγιστος ακέραιος για τον οποίον:

$$\sum_{i=0}^{K'_\alpha} \binom{N}{i} (0.5)^N \leq \alpha$$

Για την εναλλακτική, $H_1: M \neq M_0$ ή $\theta = P(X > M_0) \neq P(X < M_0)$ η κρίσιμη περιοχή επιπέδου α προσδιορίζεται από την ανισότητα $K \geq K_{\alpha/2}$ ή $K \leq K'_{\alpha/2}$ όπου $K_{\alpha/2}$ και $K'_{\alpha/2}$ είναι αντίστοιχα, το ελάχιστος και ο μέγιστος ακέραιος τέτοιοι ώστε:

$$\sum_{i=K_{\alpha/2}}^{N} \binom{N}{i} (0.5)^N \leq \alpha/2 \text{ και } \sum_{i=0}^{K'_{\alpha/2}} \binom{N}{i} (0.5)^N \leq \frac{\alpha}{2}.$$

Προφανώς, $K_{\alpha/2} = N - K'_{\alpha/2}$.

Υπολογισμός της p-τιμής: Έστω ότι $H_1: M > M_0$.

Έστω K_0 ότι είναι η παρατηρούμενη τιμή του προσημικού ελέγχου. Τότε η p-τιμή δίνεται από την ακόλουθη πιθανότητα:

$$\sum_{i=K_0}^{N} \binom{N}{i} (0.5)^N$$

Ανάλογα προκύπτουν οι p-τιμές και για τις υπόλοιπες μορφές των εναλλακτικών υποθέσεων.

Προσέγγιση με την Κανονική Κατανομή

Για οποιοδήποτε μέγεθος δείγματος N, μπορούμε να εφαρμόσουμε τον προσημικό έλεγχο χρησιμοποιώντας την διωνυμική κατανομή που είναι η ακριβής κατανομή για τη διενέργεια του ελέγχου. Σημειώνουμε ότι υπάρχουν κατάλληλοι πίνακες που για διάφορες τιμές των N και α μπορούν να μας δώσουν τις κρίσιμες περιοχές. Από τους πίνακες αυτές μπορούμε επίσης να προσδιορίσουμε και τις p-τιμές.

Οπωσδήποτε, γνωρίζουμε ότι η κανονική προσέγγιση στη διωνυμική είναι "καλή" όταν $\Theta = 0.5$. Έτσι, για μέτριες τιμές του N (τουλάχιστον 12) η κανονική προσέγγιση στη διωνυμική μπορεί να χρησιμοποιηθεί για τον καθορισμό των κρίσιμων περιοχών. Επειδή έχουμε μία συνεχής προσέγγιση σε μία διακριτή κατα-

νομή μπορούμε να συμπεριλάβουμε στους υπολογισμούς και τη διόρθωση συνέχειας.

Όπως είδαμε η σ.σ. ελέγχου δίνεται από τον συνολικό αριθμό των $X_1, X_2, \dots, X_N$ που είναι μεγαλύτεροι από την τιμή M_0 δηλ. $S_N = \sum_{i=1}^{N} I(X_i > M_0)$ όπου I είναι η δείκτρια συνάρτηση. Έτσι, μικρές τιμές της S_N οδηγούν σε απόρριψη της H_0.

$$S_N \overset{H_0}{\sim} Bin(N, 1/2) \text{ δηλ. } P(S_N = k) = \binom{N}{k} \left(\frac{1}{2}\right)^N$$

$$E_{H_0}(S_N) = \sum_{i=1}^{N} \left(\frac{1}{2}\right) = \frac{N}{2}, \quad Var_{H_0}(S_N) = \sum_{i=1}^{N} \left(\frac{1}{4}\right) = \frac{N}{4}$$

Η ασυμπτωτική κανονικότητα της τυποποιημένης μορφής:

$$S_N^* = \frac{S_N - E_{H_0}(S_N)}{\sqrt{Var_{H_0}(S_N)}} = \frac{S_N - \frac{N}{2}}{(N/4)^{1/2}}$$ προκύπτει από το Κ.Ο.Θ.

Για παράδειγμα, για την εναλλακτική $H_1: M > M_0$ η H_0 απορρίπτεται αν $K \geq K_\alpha$ όπου

$K_\alpha = \frac{N}{2} + z_\alpha \frac{\sqrt{N}}{2}$ όπου z_α είναι το α-ποσοστιαίο σημείο της $N(0,1)$.

Επίσης, η p-τιμή δίνεται από την έκφραση

$$1 - \Phi\left(\frac{K_0 - \frac{N}{2}}{\sqrt{0.25N}}\right), \text{ όπου } \Phi \text{ η σ.κ. της } N(0,1).$$

Σχόλια: Οι υποθέσεις που απαιτούνται για τη διενέργεια του Sign test είναι οι ελάχιστες δυνατές.

(1) Ανεξαρτησία παρατηρήσεων

(2) Πληθυσμός που έχει μία συνεχή σ.κ. F_X παντού.

(3) Πειραματικά υπάρχει ένα πρόβλημα για τις διαφορές ευρίσες όπως $X_i - M_0 = 0$. Θα δούμε πώς αντιμετωπίζεται.

(4) Είναι εφαρμόσιμο όταν ποσοτικά δεδομένα δεν είναι εφικτό να βρεθούν

(5) Μπορεί να χρησιμοποιηθεί σε δίχοτομικά δεδομένα με παρατηρήσεις της μορφής ΝΑΙ-ΟΧΙ (yes-no)

(6) Είναι εύκολο στη χρήση και η κανονική προσέγγιση είναι αρκετά ακριβής αυτόρη και για μέτριες τιμές του N.

Αντιμετώπιση του (3). Θεωρητικά, οι διαφορές αυτές δεν δημιουργούν πρόβλημα διότι υποθέτουμε συνέχεια της F_X στην γειτονιά της διαμέσου. Μπορούμε να έχουμε μηδενικές διαφορές είτε λόγω μη-ακριβών μετρήσεων είτε λόγω λανθάσμενης υποτιθέμενης συνέχειας της F_X.

Η πιο κοινή αντιμετώπιση των μηδενικών διαφορών είναι να **αγνοήσουμε** αυτές τις διαφορές και να μειώσουμε αναλογικά το N. Εναλλακτικά, μπορούμε να θεωρήσουμε:

μισές από τις διαφορές αυτές ως "+" και

μισές από τις διαφορές αυτές ως "−".

```
# DATA
x=c(-4.7,3.7,22.4,23.5,14.4,13.6,8.7,9.1,20.2,6.5,-7.8,10.8,15.6,10.1,-6.9)
library("BSDA")
SIGN.test(x, md=0, alternative = "two.sided")
##SIGN.test(x, md=0, alternative= "greater")
##SIGN.test(x, md=0, alternative= "less")
```

-11-

```
One-sample Sign-Test

data:  x
s = 12, p-value = 0.03516
alternative hypothesis: true median is not equal to 0
95 percent confidence interval:
  4.198872 15.386198
sample estimates:
median of x
      10.1

                  Conf.Level L.E.pt  U.E.pt
Lower Achieved CI     0.8815 6.5000 14.4000
Interpolated CI       0.9500 4.1989 15.3862
Upper Achieved CI     0.9648 3.7000 15.6000
```