Τετάρτη 17/02/21

Μη παραμετρικές μέθοδοι:

Συχνά δεν έχουμε προγενέστερη γνώση των κατανομών που περιγράφουν τα δεδομένα μας και σε τέτοιες περιπτώσεις η χρήση της παραμετρικής στατιστικής μπορεί να οδηγήσει σε λανθασμένα αποτελέσματα και συμπεράσματα.

Σε περιπτώσεις όπου δεν είναι επιτρεπτή απλοϊδιτατε υπόθεση για τη μορφή του πληθυσμού ο ερευνητής χρησιμοποιεί το Κ.Ο.Θ ώστε να διεξάγει έλεγχους. Υπάρχουν ωστόσο προβλήματα όταν τα δείγματα που έχουμε συλλέξει από τον υπό μελέτη πληθυσμό είναι μικρά και τα δεδομένα δεν κατανέμονται ομαλά με την κανονική κατανομή.

~~Το βασικό πλεονέκτημα~~

Η βασική ιδέα της μη παραμετρικής στατιστικής συμπερασματολογίας είναι να επιτύχει συμπεράσματα για γενικές ποσότητες, χρησιμοποιώντας τα δεδομένα αλλά παράλληλα θεωρώντας το μικρότερο δυνατό σύνολο υποθέσεων. Το κυριότερο πλεονέκτημα είναι ότι οι μη-παραμετρικές τεχνικές δεν προϋποθέτουν γνώση της κατανομής του πληθυσμού.

- Απαιτούν λιγες υποθέσεις για τους πληθυσμούς από τους οποίους προέρχονται τα δεδομένα
- Επιτρέπουν στον ερευνητή να υπολογίσει σε κάποιες περιπτώσεις την ακριβή τιμή του παρατηρούμενου επιπέδου σημαντικότητας α.
- Πιο απλές τεχνικές σε εφαρμογή και συχνά πιο εύκολες στην κατανόηση
- Δεν είναι ευαίσθητες στις ακραίες τιμές
- Εφαρμόζονται συχνά στις τάξεις των δεδομένων και όχι στα ίδια τα δεδομένα
- Μπορούν να χρησιμοποιηθούν και για δεδομένα που είναι ταξινομημένα σε κατηγορίες (κλίμακα διάταξης ή ονομαστική κλίμακα

- Σε περίπτωση που είναι γνωστό ότι τα δεδομένα προέρχονται από κανονική κατανομή, οι έλεγχοι είναι πιο ~~ισχυροί~~ ισχυροί.
- Αν η κατανομή των δεδομένων δεν είναι η κανονική τότε οι μη-παραμετρικοί έλεγχοι έχουν πλεονέκτημα έναντι των παραμετρικών.

Βασικό Ερώτημα: Όταν δεν έχω / δεν ξέρω πληροφορίες για σ.π.π
Πως μπορώ να την εκτιμήσω?

Η εκτίμηση για να υλοποιηθεί χρειάζεται μόνο ένα δείγμα $X_1, \dots X_n$ από την ζητούμενη σ.π.π $f(x)$.

Η μέθοδος της μη παραμετρικής εκτίμησης έχει διαφορές με την παραμ. εκτίμηση
- Η παραμετρική εκτίμηση της σ.π.π
→ υποθέτει μερική γνώση του υπό εκτίμηση μοντέλου, γνώση συνήθως της οικογένειας που ανήκει η κατανομή
→ δίνει πιο αιχμηρά αποτελέσματα από τη μη παραμετρική προσέγγιση μόνο εφόσον το μοντέλο που υποθέσαμε είναι σωστό.

Σκοπός είναι να μπορούμε να ξεχωρίσουμε με βάση τα δεδομένα πότε είναι η καταλληλη περίπτωση για να εφαρμόσουμε κάθε μια από τις δύο μεθόδους.

Μη παραμετρικοί εκτιμητές των σ.π.π $f(x)$ και α.σ.κ $F(x)$ είναι

$$\hat{f}(x) = (nh)^{-1} \sum_{i=1}^{n} 1\left\{x - \frac{h}{2} \le X_i \le x + \frac{h}{2}\right\}$$ → δείκτρια συνάρτηση

η $\hat{f}(x)$ είναι ο ορισμός του ιστογράμματος και η παράμετρος h είναι το εύρος της μπάρας του.

Πως λειτουργεί στην πράξη το ιστόγραμμα:
- Ξεκινώντας από ένα αρχικό σημείο x_0, δημιουργούμε τα διαστήματα $I_j = [x_0 + jh, x_0 + (j+1)h]$, $j = -1, 0, 1, \dots$
Έτσι έχουμε χωρίσει το Π.Ο της f σε υποδιαστήματα

- Η τιμή της $f(x)$ ~~της~~ εκτιμάται μετρώντας τον αριθμό των παρατηρήσεων $X_1, \dots, X_n$ που πέφτουν σε κάθε I_j.
Αυτή η εκτίμηση είναι η εμπειρική πιθανότητα να πέσει μια οποιαδήποτε παρατήρηση στο δίαστ. I_j και προσομοιάζει την αριθμό της σ.π.π.

• στον τύπο της $\hat{f}(x)$ όλα είναι γνωστά εκτός ίσως από το h, το οποίο δίνεται από τύπους που εξαρτώνται μόνο από τα δεδομένα. Η τιμή του h επηρεάζει την ακρίβεια του εκτιμητή.

Παράδειγμα από τα δεδομένα geyser της R από το πακέτο MASS

Κανόνας Sturges $h = \lceil \log_2(299) \rceil + 1 \simeq 10$

Ιδιότητες ιστογράμματος:

Μέση τιμή και διακύμανση της $\hat{f}(x)$

Αν η $f(x)$ είναι συνεχής συνάρτηση πυκνότητας πιθανότητας.

$$E(\hat{f}(x)) = f(x)(1 + o(1))$$

$$Var(\hat{f}(x)) = (nh)^{-1} f(x) + o(n^{-2}).$$

Απόδειξη:

Γνωρίζουμε ότι για συνεχή τ.μ. $X \sim f(x)$ ισχύει:

$$E(g(X)) = \int g(x) \cdot f(x)\, dx$$

$X_1, \dots, X_n$ ισόνομες επιμέρους

$$E\{\hat{f}(x)\} = E\left[(nh)^{-1} \sum_{i=1}^{n} 1_{\{x-\frac{h}{2} \le x_i \le x+\frac{h}{2}\}}\right] = n(nh)^{-1} E\left(1_{\{x-\frac{h}{2} \le x_i \le x+\frac{h}{2}\}}\right)$$

$$= h^{-1} \int_{x-\frac{h}{2}}^{x+h/2} 1 f(t)\, dt + h^{-1} \int_{x-h/2}^{x+h/2} 0 \cdot f(y)\, dy =$$

$$= h^{-1} \int_{x-h/2}^{x+h/2} f(t)\, dt = f(x)(1 + o(1)).$$

Προκύπτει αν εφαρμόσουμε τον ισχυρό νόμο των μεγάλων αριθμών και τις ιδιότητες του ολοκλ. Riemann

Βάση του αριστερού $(nh)^{-1} \xrightarrow{a.s.} h^{-1} \int_{x-h/2}^{x+h/2} f(y)dy \simeq f(x)$

Ισχυρός νόμος των μεγάλων αριθμών:

Έστω $X_1, X_2, \dots$ μια ακολουθία ανεξάρτητων και ισόνομων τ.μ με $E|X_i| = \mu$ και πεπερασμένη διακύμανση $V(X_i) = \sigma^2, i = 1, 2, \dots$ Τότε η ακολουθία δειγματικών μέσων $\bar{X}_v = \frac{1}{v} \sum_{i=1}^{v} X_i$ συγκλίνει με πιθανότητα 1 στη θεωρητική μέση τιμή μ:

δηλ. $P(\lim_{v \to \infty} \bar{X}_v = \mu) = 1.$

Για την διακύμανση:

$Var\{g(x)\} = Eg^2(x) - [E(g(x))]^2.$

$Var\{\hat{f}(x)\} = Var\{(nh)^{-1} \sum_{i=1}^{n} 1_{\{x_0 - \frac{h}{2} \le x_i \le x_0 + \frac{h}{2}\}}\} =$

$= E[(nh)^{-1} \sum_{i=1}^{n} 1_{\{\dots\}}]^2 - [E(\hat{f}(x))]^2$

Γνωρίζουμε ότι $(\sum_{i=1}^{n} x_i)^2 = \sum_{i=1}^{n} x_i^2 + \sum\sum_{i \neq j} x_i x_j$

Έτσι έχουμε $E\{\hat{f}(x)\}^2 = E(nh)^{-1} \sum 1_{\{\dots\}}]^2 =$

$= (nh)^{-2} \{ \sum_{i=1}^{n} E 1^2_{\{\dots\}} + \sum\sum_{i \neq j} E 1_{\{\dots\}} 1_{\{\dots\}} \}$ ομοίως $=$

$= (nh)^{-2} \{ n E 1^2_{\{\dots\}} + n(n-1) E 1_{\{\dots\}} 1_{\{\dots\}} \}$ (1)

$h^{-1} E 1^2_{\{\dots\}} = f(x)(1 + o(1)).$ (2)

$h^{-2} \cdot E 1_{\{\dots\}} 1_{\{\dots\}} = h^{-1} E 1_{\{\dots\}} \cdot h^{-1} E 1_{\{\dots\}} =$

$= \left(h^{-1} \int_{x-h/2}^{x+h/2} f(t)dt \right) \left(h^{-1} \int_{x-h/2}^{x+h/2} f(z)dz \right) = f^2(x)(1 + o(1)).$ (3)

Επομένως $E\{\hat{f}(x)\}^2 = nh^{-1}f(x) + \frac{n(n-1)}{n^2} f^2(x)$

Τελικά $Var.\{\hat{f}(x)\} = nh^{-1}f(x) - n^2 f(x)^2$.

Κεντρικό ρόλο παίζει η ποσότητα $x_i \in I_j$.
Η κατανομή αυτής της ποσότητας ταυτίζεται με τον ορισμό της διωνυμικής κατανομής. Όπου η πιθανότητα να πέσει η i παρατήρηση στο διάστημα I_j είναι $p_j = \int_{I_j} f(x)dx$. , όπου $f(x)$ η πραγματική.

αλλιώς άγνωστη κατανομή

Κατανομή της $\hat{f}(x_i)$
Για $x_j \in I_j = [x_j - h/2, x_j + h/2]$ τότε $f(x_j) \sim binomial(n, p_j)$

Εκτίμηση της α.ο.κ. $F(x)$

$$\hat{F}(x) = n^{-1} \sum_{i=1}^{n} 1\{x_i \le x\}.$$

Θεωρούμε ότι ο $\hat{f}(x)$ προκύπτει από την $\hat{F}(x)$.

$\hat{F}(x) = n^{-1} \sum_{i=1}^{n} 1\{x_i < x\}$ έναν εμπειρικός εκτιμητής της $F(x) = P(X \le x)$.

$$F(x+h) - F(x-h) \simeq n^{-1} \sum_{i=1}^{n} 1\{x_i < x+h\} - n^{-1} \sum_{i=1}^{n} 1\{x-h \le x_i\} =$$

$$= n^{-1} \sum_{i=1}^{n} 1\{x-h < x_i < x+h\}$$ και γράφουμε

$$\hat{f}(x) = \frac{1}{2hn} \sum_{i=1}^{n} 1\{|x-x_i| < h\} = \frac{1}{2hn} \sum_{i=1}^{n} 1\left\{\frac{|x-x_i|}{n} < 1\right\}$$

Παρατηρούμε ότι $\sum_{i=1}^{n} 1\{x_i \le x\} = n\hat{F}(x) \sim B(n, F(x))$.
είναι άθροισμα Bernoulli τ.μ. με παρ. $p = F(x)$.

Έχουμε ότι αν $X \sim B(n,p)$ τότε $EX = n \cdot p$, $Var(X) = np(1-p)$

Άρα $n\hat{F}(x) = n^{-1} \sum_{i=1}^{n} 1\{X_i \le x\}$ έχει Διωνυμική κατανομή με

$$E\hat{F}(x) = n^{-1} n F(x) = F(x)$$
$$Var(\hat{F}(x)) = n^{-2} n F(x)(1-F(x)) = \frac{1}{n} F(x)(1-F(x))$$

Με βάση τις ιδιότητες και με εφαρμογή του Κ.Ο.Θ προκύπτει

Ασυμπτωτική κατανομή της $\hat{F}(x)$.

Καθώς το $n \to \infty$.

$$\hat{F}(x) \sim N(F(x), n^{-1} F(x)(1-F(x)))$$

$\hat{F}^{-1}(x) = \frac{1}{n} \sum_{i=1}^{n} 1\{X_{(i)} \le x\}$, $X_{(1)}, \dots X_{(n)}$ διατεταγμένο δείγμα από την μικρότερη στην μεγαλύτερη παρατήρηση

Βλέπουμε ότι $\hat{F}(x)$ βαθμωτή συνάρτηση με διαφορά ύψων ίση με $\frac{1}{n}$ σε κάθε σημείο $X_{(i)}$ δηλ δίνει ίδια πιθανότητα $\frac{1}{n}$ σε κάθε $X_{(i)}$, άρα είναι αύξουσα και δεξιά συνεχής.

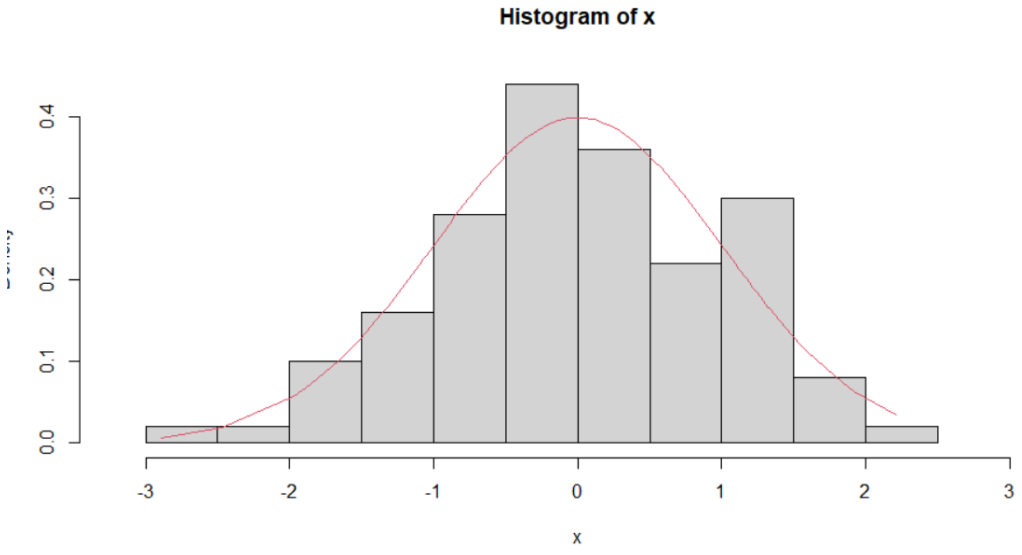
ΜΗ ΠΑΡΑΜΕΤΡΙΚΗ ΣΤΑΤΙΣΤΙΚΗ ΜΑΘΗΜΑ 1 ΕΚΤΙΜΗΣΗ ΣΠΠ ΜΕ ΙΣΤΟΓΡΑΜΜΑ ΣΤΗΝ R

Π.Χ.1

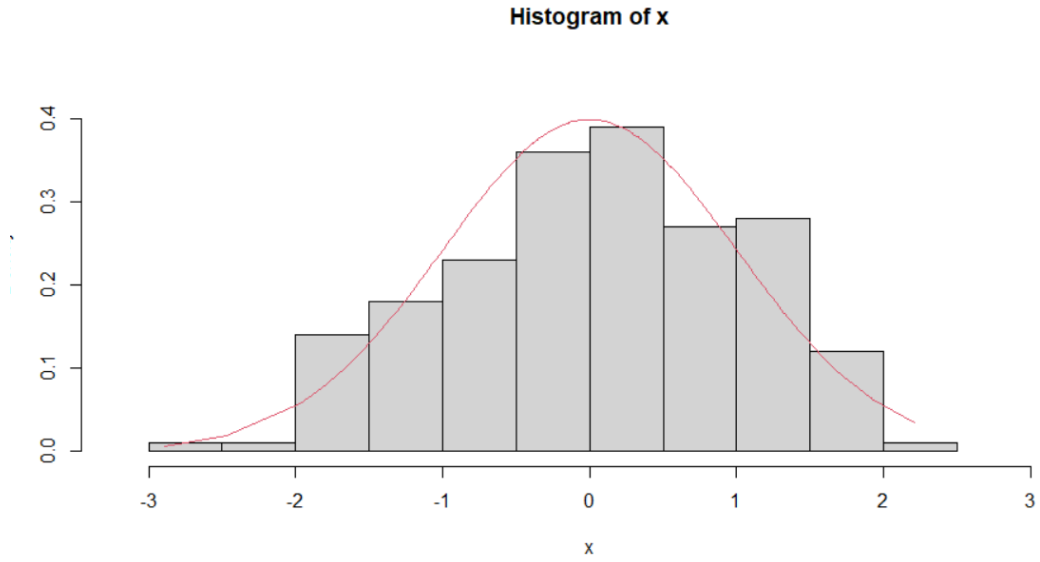
Π.Χ.2

Π.Χ.3

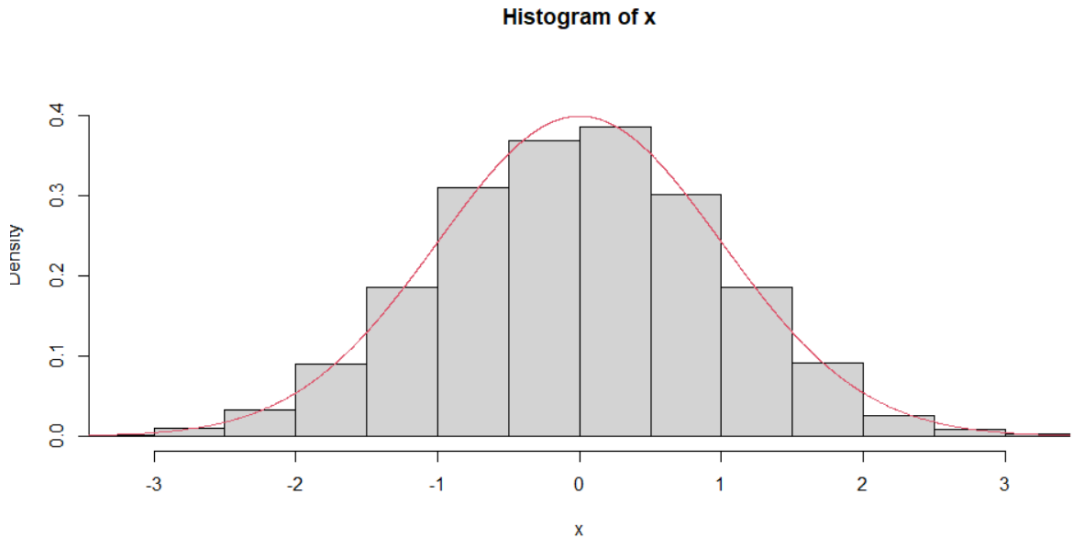
```
#μη παραμετρική στατιστική εκτίμηση της σ.π.π με
#ιστόγραμμα
set.seed(20)
x<-rnorm(100)
x<-sort(x)
hist(x,probability=TRUE,ylim=c(0,0.45),xlim=c(-3.2,3.2))
lines(x,dnorm(x),col=2)

set.seed(20)
x<-rnorm(200)
x<-sort(x)
hist(x,probability=TRUE,ylim=c(0,0.45),xlim=c(-3.2,3.2))
lines(x,dnorm(x),col=2)
#με την αυξηση των παρατηρήσεων  μειωθηκε η διακύμανση όπως μπορούμε να δούμε και από το θεώρημα στις σημειώσεις
#γενικά όσο πιο πολλές παρατηρήσεις έχουμε τόσο πιο καλή θα είναι η εκτίμηση


#δοκιμάζουμε για n=10000
set.seed(20)
x<-rnorm(10000)
x<-sort(x)
hist(x,probability=TRUE,ylim=c(0,0.45),xlim=c(-3.2,3.2))
lines(x,dnorm(x),col=2)
#παρατηρούμε στο ιστόγραμμα οτι έχουμε αισθητά καλύτερη προσέγγιση
```



Π.Χ.1


Histogram of x
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3
x

П.Х.2


Histogram of x
Density
0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3
x

П.Х.3