Τετάρτη 24/03/21

Τεστ Shapiro - Wilk

Έχουμε διαθέσιμες παρατηρήσεις $X_1, \dots, X_n$ προέρχονται από ανεξάρτητες και ισόνομες τ.μ. με α.σ.κ F

Έστω $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ διατεταγμένες σε σειρά μεγέθους παρατηρ. Το τεστ των Shapiro-Wilk χρησιμοποιείται για να ελέγξουμε αν η κατανομή από την οποία προέρχεται το δείγμα μας είναι κανονική και δίνεται από

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

Συμβολίζουμε με $m^T = (m_1, \dots, m_n)$ το διάνυσμα των αναμενόμενων τιμών διατεταγμένων στατιστικών της τυπικής κανονικής κατανομής και έστω $V = (V_{ij})$ ο αντίστοιχος πίνακας διακυμάνσεων - συνδιακυμάνσεων.

Επειδή αν $X_{(1)} \le X_{(2)} \le \dots \le X_{(n)}$ είναι διατεταγμ. δείγμα από την κανονική κατανομή με μέση τιμή 0 και διακύμανση 1 τότε $E(X_{(i)}) = m_i$ και $cov(X_{(i)}, X_{(j)}) = V_{ij}$, $i, j = 1, 2, \dots n$.

Θέλουμε τώρα ένα διάνυσμα $y^T = (y_1, \dots y_n)$ διατεταγμένων παρατηρ. Θέλουμε να βρούμε ένα τεστ για την υπόθεση ότι το δείγμα είναι από κανονική κατανομή με άγνωστη μέση τιμή μ και διακύμανση σ^2.

Αν το δείγμα y' είναι κανονικό τότε τα y_i μπορούν να γραφούν στη μορφή $y_i = \mu + \sigma X_i$, $i = 1, \dots n$.

Από το θεώρημα των γενικευμένων ελαχίστων τετραγώνων ότι οι καλύτεροι γραμμικοί και αμερόληπτοι εκτιμητές των μ, σ είναι αυτοί που ελαχιστοποιούν την τετραγωνική μορφή.

$(y-\mu\cdot 1-\sigma m)^T V^{-1}(y-\mu 1-\sigma m)$ όπου $1^T=(1,\dots 1)$

Λύνοντας παίρνουμε ότι

$$\boxed{\hat{\mu}=\frac{m^T V^{-1}(m 1^T-1m^T)V^{-1}y}{1^T V^{-1} m^T V^{-1} m-(1^T V^{-1} m)^2}}$$

$$\boxed{\hat{\sigma}=\frac{1^T V^{-1}(1\cdot m-m1^T)V^{-1}y}{1^T V^{-1} 1 m^T V^{-1} m-(1^T V^{-1} m)^2}}$$

Όμως για συμμετρικές κατανομές έχουμε ότι $1^T V^{-1} m=0$ άρα τα παραπάνω γράφονται απλούστερα

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n} y_i=\bar{y}$$

$$\hat{\sigma}=\frac{m^T V^{-1} y}{m^T V^{-1} m}.$$

Αν θέσουμε $S^2=\sum_{i=1}^{n}(y_i-\bar{y})^2$ τον αμερόληπτο εκτιμητή της $(n-1)\sigma^2$ τότε το στατιστικό ορίζεται ως

$$W=\frac{R^4\hat{\sigma}^2}{C^2 S^2}=\frac{b^2}{S^2}=\frac{(a'y)^2}{S^2}=\frac{\left(\sum a_i X_{(i)}\right)^2}{\sum(y_i-\bar{y})^2}.$$

όπου $R^2=m^T V^{-1} m$, $C^2=m^T V^{-1} V^{-1} m$, $b=\frac{R^2\hat{\sigma}}{C}$

$$a'=\frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}=(a_1,\dots a_n)$$

από την μορφή του στατιστικού $W=\frac{b^2}{S^2}$ βλέπουμε ότι το στατιστικό ορίζεται ως ο εκτιμητής της διακύμανσης. (κατάλληλα κανονικοποιημένος) δια την κλασική, εμπειρική διακύμανση)

Με άλλα λόγια αν κάποιος πάρει δείγμα όντως από κανονική κατανομή τότε ο σταθμισμένος εκτιμητής όσο ο παρονομαστής του W εξαιρουμένης μιας σταθεράς θα εκτιμούν την ίδια ποσότητα σ^2.

→ Αυτή η χαρακτηριστική ιδιότητα ισχύει μόνο για κανονικούς πληθυσμούς ενώ για μη κανονικούς πληθυσμούς οι b^2, S^2 δεν εκτιμούν την ίδια ποσότητα.

Παρατηρούμε ότι nb (εξαιρούμενης της σταθεράς C) είναι ο καλύτερος γραμμικός αμερόληπτος εκτιμητής της κλίσης ενός γραμμικού μοντέλου παλινδρόμησης των διατεταγμένων παρατηρήσεων y_i με τις αναμενόμενες τιμές m_i των τυποποιημένων κανονικών διατεταγμένων στατιστικών.

Ισοδύναμα αν κάνουμε το διάγραμμα των $X_{(i)}$, $m_i = \Phi^{-1}(i/(n+1)) = E_{N(0,1)} X_{(i)}$, αναμένουμε υπό την μηδενική τα σημεία του διαγράμματος να είναι κοντά στην ευθεία $X_{(i)} = \mu + \sigma m_i$

Συμπερασματικά βλέπουμε ότι η κατασκευή του στατιστικού και ο τύπος του αντικατοπτρίζει μια χαρακτηριστική ιδιότητα της κανονικής κατανομής και επομένως δικαιολογεί τη χρήση του στατιστικού για έλεγχο της κανονικότητας του δείγματος.

Ένας απλός τρόπος για να εκτιμήσουμε την κατανομή του test είναι να υπολογίσουμε π.χ 1000 τιμές του τεστ και να χρησιμοποιήσουμε το ιστόγραμμα για να εκτιμήσουμε την σ.π.π.

χρησιμοποιούμε δεδομένα από την τυπική κανονική κατανομή και θα ελέγξουμε την ορθότητα της προσαρμογής στην τυπική κανονική α.σ.κ.

```r
require(graphics)
n <- 1000
TestVals <- vector(n, mode="numeric")
set.seed(4)
for (i in 1:n)
{
  x <- rnorm(1000)
  TestVals[i] <- shapiro.test(x)$statistic
}
hist(TestVals, freq=F)
lines(density(TestVals), col=2)
```

Αν ~~[illegible]~~ παράγουμε δεδομένα από άλλη κανονική κατανομή, η κατανομή του στατιστικού δεν αλλάζει.

αν αντί για κανονικό δείγμα έχουμε δείγμα από την ομοιόμορφη κατανομή, παρατηρούμε ότι αλλάζει το ύψος των αναμενόμενων ~~[illegible]~~ άξονα y αλλά και μικραίνει το εύρος των τιμών του test (άξονας x).