Τετάρτη 03/03/21:

Έλεγχοι καλής προσαρμογής:

Έστω δείγμα $X_1, \ldots, X_n$ από κάποια (άγνωστη) κατανομή $F(x)$
Το ερώτημα που θέλουμε να απαντήσουμε είναι αν το δείγμα προέρχεται από την δοθείσα κατανομή $F(x, \theta)$

Έλεγχος της υπόθεσης: $H_0: F(x) = F(x, \theta)$ έναντι $H_a: F(x) \neq F(x, \theta)$.
- μετράμε την απόσταση μεταξύ $F(x)$ και $F(x, \theta)$ και εφαρμόζουμε ένα κριτήριο με το οποίο θα αποφασίσουμε την εγγύτητα της $F(x, \theta)$ στην πραγματική $F(x)$

Η $F(x, \theta)$ θεωρείται δεδομένη, υπό την H_0 ενώ η $F(x)$ άγνωστη, επομένως στην θέση της $F(x)$ χρησιμοποιούμε έναν εκτιμητή της.
[η χρήση διαφορετικού εκτιμητή οδηγεί σε διαφορετικό στατιστικό τεστ.]
[Επίσης μπορούμε να χρησιμοποιήσουμε διαφορετικές μετρικές για την απόσταση μεταξύ $f(x)$ και $F(x, \theta)$].

Οι ιδιότητες του εκτιμητή που χρησιμοποιούμε και ο τύπος της μετρικής που χρησιμοποιούμε καθορίζουν τις στατιστικές ιδιότητες του τεστ:
Το σημαντικότερο είναι η στατιστική συνάρτηση να είναι πάντα τ.μ. αφού πάντα εξαρτάται από τα δεδομένα.
Σαν τ.μ. η στατ. συνάρτηση έχει κατανομή η οποία είναι διαφορετική ανάλογα με τον συνδυασμό μετρικής, εκτιμητή της $f(x)$ και της $F(x, \theta)$
Η κατανομή της χρησιμοποιείται για να θεσπίσουμε το κρ. απορ. ή αποδοχής της υπόθεσης

Ακραίες τιμές της σ.σ σημαίνουν σημαντική απόσταση της $F(x)$ με $F(x, \theta)$
Όταν θα ισχύει η ~~...~~ μηδενική υπόθεση περιμένουμε η τιμή του στατιστικού να είναι κοντά στο 0.

Πιο επίσημα θεωρούμε ότι η τιμή του στατιστικού που πήραμε βάσει του δοθέντος δείγματος προέρχεται όντως από την κατανομή του στατιστικού υπό την H_0 αν είναι μέσα στο $(1-a)\%$ της μάζας της κατανομής των τιμών του στατιστικού.

Το a καθορίζει τα εξής όρια:

- όταν η τιμή του στατιστικού ανήκει στο $(1-a)\%$ του εύρους τιμών της κατανομής του στατιστικού τότε δεν απορ. την H_0.
- αντίθετα αν παίρνει τιμή εκτός του $(1-a)\%$ απορρίπτουμε την H_0

Στην γενική περίπτωση αν η κατανομή της σ.σ $T(x)$ υπό την H_0 έχει ποσοστιαία σημεία $-q_{\frac{a}{2}}, q_{\frac{a}{2}}$, δηλαδή στο διάστημα $(-q_{\frac{a}{2}}, q_{\frac{a}{2}})$ περικλείεται το $100(1-a)\%$ των τιμών της $T(x)$ τότε ισχύει $1-a = P(-q_{\frac{a}{2}} < T(x) < q_{\frac{a}{2}})$. (1)

Το επίπεδο σημαντικότητας: όπως ορίζεται από την (1) είναι η πιθανότητα να παρατηρηθεί τιμή του στατιστικού μεγαλύτερη από αυτή που έδωσε το δείγμα των παρατηρήσεων.
Ή ισοδύναμα το a καθορίζει το μέγεθος του ρίσκου που είμαστε διατεθειμένοι να δεχτούμε για απορ. ή αποδοχή της H_0

Έλεγχος Καλής Προσαρμογής Kolmogorov-Smirnov

Βασίζεται στην μέτρηση της απόστασης μεταξύ $F(x)$ και $F(x,\theta)$ μέσω της μετρικής $D = \sup_{x \in \mathbb{R}} |F(x) - F(x,\theta)|$

Βασικά Χαρακτηριστικά Τεστ:

- Δ $F(x,\theta)$ είναι πλήρως προσδιορισμένη δηλ. η τιμή της παραμέτρου θ δεν έχει προκύψει από εκτίμηση μέσω του δείγματος, δηλ. είναι η ίδια ανεξάρτητα με το δείγμα που έχουμε για να αναλύσουμε τα δεδομένα.

Δ Η $F(x)$ εκτιμάται από την εμπειρική της α.σ.κ.

Δ Η μετρική που χρησιμοποιούμε για την απόσταση μεταξύ των $\hat{F}(x)$ και $F(x,\theta)$ είναι η sup norm

δεδομένη, δεν αλλάζει κάθε φορά που μας δίνουν άλλο δείγμα.

$D_n = \sup_{x\in\mathbb{R}} |\hat{F}(x) - F(x,\theta)|$, τ.μ. έχει διαφορετική τιμή για κάθε δείγμα.

Για να μπορέσουμε να χρησιμοποιήσουμε την D_n πρέπει να προσδιορίσουμε την κατανομή της.
αν $\theta \in \mathbb{R}^p, p \geq 1$ τ.μ. η κατανομή του τεστ αλλάζει δραματικά

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

π.χ) δεδομ. από τυπική κανονική κατανομή.

```
require(graphics)
n<-1000
TestVals <-vector(length=n, mode="numeric")
set.seed(4)
for(i in 1:n)
{
 x<-rnorm(100)
 TestVals[i]<-ks.test(x,"pnorm")$statistic
}
hist(TestVals, freq=F)
lines(density(TestVals), col=2)
```

"ΔΕΝ ΠΡΕΠΕΙ ΝΑ ΠΑΡΑΒΙΑΖΟΥΜΕ ΤΟΝ ΚΑΝΟΝΑ ΟΤΙ ΟΙ ΠΑΡΑΜΕΤΡΟΙ ΤΗΣ ΚΑΤΑΝΟΜΗΣ ΠΡΕΠΕΙ ΝΑ ΕΙΝΑΙ ΣΤΑΘΕΡΟΙ ΑΡΙΘΜΟΙ!!

Test Cramer-von Mises:

Αν αντί της μετρικής sup, χρησιμοποιήσουμε την L_2 μετρική η συνάρτηση ελέγχου γίνεται

$$W = \int \{F(x) - F(x,\theta)\}^2 \, dF(x).$$

Με τον όρο $dF(x)$ η διαφορά $F(x) - F(x,\theta)$ ζυγίζεται με βάση την πυκνότητα των παρατηρήσεων ώστε πολύ απομακρυσμένες (ακραίες) παρατηρήσεις να μην επηρεάζουν πολύ το τελικό αποτέλεσμα.

Αν χρησιμοποιήσουμε την $\hat{F}(x)$ στην θέση της $F(x)$ έχουμε

$$\hat{W} = \int \{\hat{F}(x) - F(x,\theta)\}^2 \, dF(x,\theta)$$

Η πρακτική υλοποίηση του $\hat{W}$ εξαρτάται μόνο από το δειγματικό δείγμα και τη δοθείσα κατανομή υπό την H_0.

Δεδομένου ότι ο υπολογισμός του ολοκληρώματος στην πράξη είναι δύσκολος χρησιμοποιούμε την ιδιότητα της $\hat{F}(x)$ ότι για το διατεταγμένο δείγμα $X_{(1)}, X_{(2)}, \dots, X_{(n)}$

$$\hat{F}(X_{(i)}) = \frac{i}{n}, \quad i = 1, \dots, n$$

Έτσι $$\boxed{\hat{W} = \sum_{i=1}^{n} \left\{ \frac{2i-1}{2n} - F(X_{(i)},\theta) \right\}^2 + \frac{1}{12n}}$$

Test Anderson-Darling

βασίζεται στην ελαφρώς τροποποιημένη μετρική L_2.

$$A = n \int \{ F(x) - F(x,\theta) \}^2 w(x) \, dF(x)$$

Η $w(x)$ είναι μια δεύτερη συνάρτηση βάρους διαφορετική από την $f(x)$ αν $w(x) = 1$ έχουμε το τεστ Cramer-von Mises.

Το τεστ των Anderson-Darling προκύπτει από την A θέτοντας $w(x) = (F(x)(1-F(x)))^{-1}$.

$$A = n \int \frac{\{F(x) - F(x,\theta)\}^2}{F(x)(1-F(x))} \, dF(x).$$

Αυτή η επιλογή της $w(x)$ δίνει μεγαλύτερο βάρος στις παρατηρήσεις στις ουρές (άκρα της κατανομής).

Η A δεν μπορεί να χρησιμοποιηθεί στην πράξη διότι περιέχει την άγνωστη $F(x)$ επομένως αν χρησιμοποιήσουμε την $\hat{F}(x)$ προκύπτει η εκτιμήτρια $\hat{A}$, της A

$$\boxed{\hat{A} = \sum_{i=1}^{n} \frac{2i-1}{n} \left[\ln(F(x_i),\theta)) + \ln\{1 - F(x_{(n+1-i)},\theta)\} \right]}$$