LABORATORY OF STATISTICS AND DATA ANALYSIS

BE PARSIMONIOUS - CHOOSE WISELY

e-ΠΕΡΙΣΚΟΠΙΟ

No.3/2020

Το e-Περισκόπιο του Εργαστηρίου Στατιστικής και Ανάλυσης Δεδομένων του Πανεπιστημίου αποτελεί μια πρωτοβουλία των φοιτητών-ερευνητών που το πλαισιώνουν και δεν απευθύνεται αποκλειστικά σε άτομα με στατιστικό υπόβαθρο.

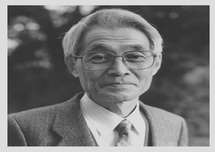
Το κάθε τεύχος είναι ανεξάρτητο των υπολοίπων και ακολουθεί συγκεκριμένη θεματολογία.

Συγκεκριμένα απαρτίζεται από: (1) Βιογραφικό σημείωμα ενός ατόμου που έχει συνδεθεί με το υπό ανάλυση θέμα και η συμβολή του στην στατιστική ήταν καθοριστική και θεμελιώδης; (2) Εισαγωγική συζήτηση του θέματος συνοδευόμενη από κάποια funny corners καθώς και ένα quiz/paradox. Στόχος του e-Περισκοπίου είναι η ενημέρωση, η ψυχαγωγία και ο προβληματισμός των αναγνωστών σε θέματα που έχουν ως κεντρικό άξονα την στατιστική. Αν επιθυμείτε να συμβάλετε στο περισκόπιο θέτοντας κάποιο θέμα προς ανάλυση, επισκεφτείτε την σελίδα του εργαστηρίου ή/και ελάτε σε επικοινωνία μαζί μας μέσω των πληροφοριών που βρίσκονται στην καρτέλα *“ΣΤΟΙΧΕΙΑ ΕΠΙΚΟΙΝΩΝΙΑΣ”*.

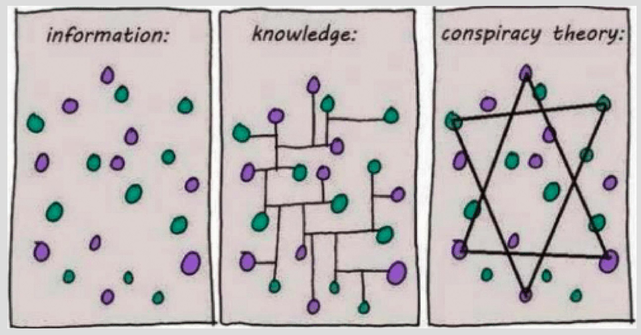
ΣΤΟΙΧΕΙΑ ΕΠΙΚΟΙΝΩΝΙΑΣ

Laboratory of
Statistics and Data Analysis

ΒΙΟΓΡΑΦΙΚΟ ΣΗΜΕΙΩΜΑ



Ο Hirotugu Akaike (*1927-2009*) ήταν Ιαπωνικής καταγωγής στατιστικός με κομβικής σημασίας συνεισφορά στον κλάδο της Στατιστικής. Ήταν ο νεότερος από τα *4* αδέρφια της οικογενείας του. Πραγματοποίησε τις προπτυχιακές του σπουδές στην Σχολή Θετικών Επιστημών του Πανεπιστημίου του Tokyo από όπου και αποφοίτησε το *1952*. Έπειτα εργάστηκε ως ερευνητής στο Institute of Statistical Mathematics και το *1961* πήρε το διδακτορικό του στα Μαθηματικά από το Πανεπιστήμιο του Tokyo. Εννέα χρόνια αργότερα, στις αρχές του *1970*, δημιούργησε ένα πρακτικό και συνάμα ευέλικτο κριτήριο πληροφορίας ονόματι Akaike’s Information Criterion, το οποίο έμελλε να αλλάξει τον ρου της ιστορίας των τεχνικών επιλογής μοντέλου. Συγκεκριμένα, το AIC αποτέλεσε την γέφυρα που ένωσε τον κόσμο των δεδομένων με τον κόσμο της μοντελοποίησης. Σήμερα υπολογίζεται ότι η δημοσίευση όπου παρουσιάζεται για πρώτη φορά το AIC, έχει πάνω από *20000* αναφορές, ενώ το AIC χρησιμοποιείται σε πάνω από *170000* επιστημονικές δημοσιεύσεις και βιβλία. Επιπλέον, ο Akaike συνεισέφερε σημαντικά και στην μελέτη χρονολογικών σειρών καθώς επίσης έπαιξε σημαντικό ρόλο στην εξέλιξη της επιστήμης της Στατιστικής στην Ιαπωνία. Στις *5* Νοεμβρίου *2017*, η Google τίμησε τον Hirotugu Akaike με ένα Doodle για την *90η* επέτειο από την γέννησή του.



ΤΟ ΠΑΡΑΔΟΞΟ ΤΟΥ SIMPSON

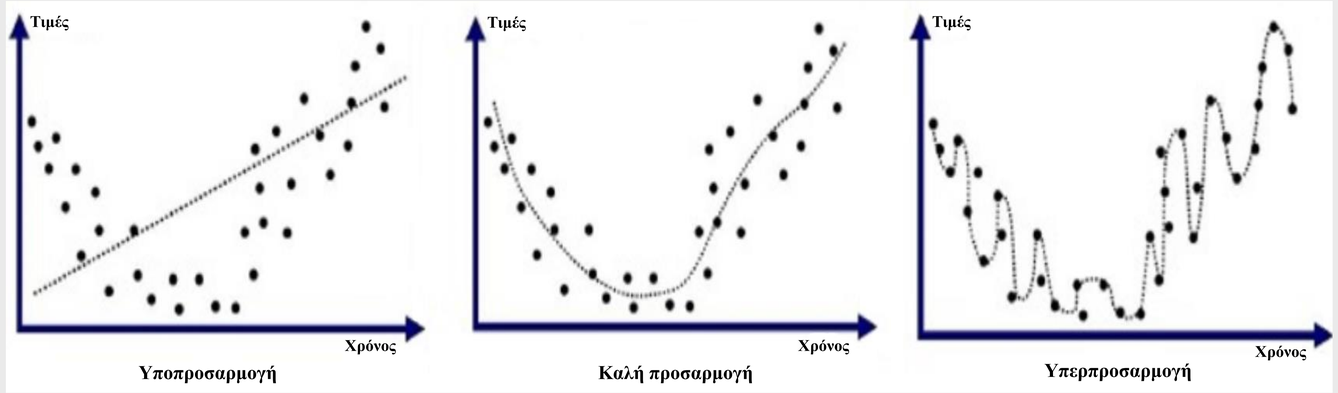
Όταν η ύπαρξη τάσης μπορεί να οδηγήσει στην…μη ύπαρξή της!

Click to learn more

ΤΑ ΚΡΙΤΗΡΙΑ ΠΛΗΡΟΦΟΡΙΑΣ ΩΣ ΤΕΧΝΙΚΗ ΕΠΙΛΟΓΗΣ ΜΟΝΤΕΛΟΥ

Όταν πρόκειται για “μεγάλα δεδομένα” οι αναλυτές συνήθως χρησιμοποιούν διάφορα στατιστικά μοντέλα για σκοπούς εξερεύνησης των ιδιοτήτων των δεδομένων αλλά και πρόβλεψης. Οποιαδήποτε και αν είναι τα δεδομένα υπό ανάλυση, ένα κρίσιμο βήμα είναι η επιλογή του καταλληλότερου μοντέλου από μια πληθώρα μοντέλων. Η επιλογή μοντέλου είναι ένα πολύ σημαντικό “συστατικό” στην στατιστική ανάλυση καθώς μέσω αυτού αποσκοπούμε σε αξιόπιστη και αναπαράξιμη στατιστική συμπερασματολογιά. Η ιστορία των τεχνικών επιλογής μοντέλου είναι μακρά και εγείρεται από έρευνες κυρίως από τους κλάδους της Στατιστικής, της Θεωρίας Πληροφορίας και της Επεξεργασίας Εικόνας. Ανά τα χρόνια, ένας μεγάλος αριθμός μεθόδων έχει αναπτυχθεί ακολουθώντας η κάθε μια την δική της φιλοσοφία και αποδίδοντας καλύτερα κάτω από διαφορετικές συνθήκες. Από τα παραπάνω γίνεται κατανοητό ότι δεν υπάρχει εκείνο το “μαγικό” μοντέλο που αποδίδει καλύτερα σε όλες τις περιπτώσεις υπό εξέταση. Η λανθασμένη επιλογή μοντέλου, μπορεί να οδηγήσει σε παραπλανητικές ευρέσεις ή σε αποθαρρυντική προβλεπτική απόδοση. Μερικές από τις περιπτώσεις μου μπορεί να μας ενδιαφέρει η έννοια της επιλογής είναι οι ακόλουθες:

- για το πλήθος των μεταβλητών ενός μοντέλου,
- τον αριθμό των συνιστωσών σε ένα μικτό μοντέλο,
- τον αριθμό των σημείων αλλαγής συμπεριφοράς μιας χρονολογικής σειράς.



Γράφημα: Προσαρμογή τριών διαφορετικών μοντέλων σε ένα set χρονολογικών δεδομένων.

Τα κριτήρια πληροφορίας αποτελούν ως επί το πλείστον τεχνική επιλογής μοντέλου. Βασίζονται στην συνάρτηση πιθανοφάνειας και εφαρμόζονται σε προβλήματα που μπορούν να περιγραφούν από παραμετρικά μοντέλα. Υπάρχει μια πληθώρα κριτηρίων που έχουν αναπτυχθεί ανά τα χρόνια και υπολογίζεται ότι το πλήθος τους είναι περίπου όσα και τα γράμματα του αγγλικού αλφάβητου. Από τα πιο μνημονευμένα κριτήρια πληροφορίας είναι το Akaike’s Information Criterion (AIC) και το Bayesian ή Schwartz’s Information Criterion (BIC ή SIC).

Akaike’s Information Criterion

Το AIC αποτελεί ένα score με βάση το οποίο μπορεί να προσδιοριστεί το “καλύτερο” μοντέλο από ένα σύνολο υποψηφίων μοντέλων. Το score αυτό θεωρείται χρήσιμο στην περίπτωση που πρέπει να συγκριθεί με άλλα AIC scores με την προϋπόθεση ότι αφορούν το ίδιο set δεδομένων. Το μοντέλο με την χαμηλότερη τιμή AIC είναι και το “καλύτερο” ανάμεσα στα υποψήφια.

Δίνεται από τον τύπο:

$$AIC = -2\,log(L) + 2p,$$

όπου L είναι η πιθανοφάνεια και p είναι το πλήθος των παραμέτρων του μοντέλου υπό εξέταση.

Παρατηρώντας την παραπάνω σχέση, γίνεται αντιληπτό ότι μοντέλα με υψηλή πιθανοφάνεια οδηγούν σε χαμηλές τιμές AIC. Βέβαια υπάρχει και ένας όρος penalty ($2p$) ο οποίος “τιμωρεί” εκείνα τα μοντέλα που έχουν πολλές παραμέτρους και ως εκ τούτου είναι πολυπλοκότερα.

Το AIC χρησιμοποιεί την (log) πιθανοφάνεια ενός μοντέλου ως μέτρο προσαρμογής. Με τον όρο πιθανοφάνεια εννοούμε το πόσο πιθανό είναι να παρατηρήσουμε τα δεδομένα που έχουμε στην διάθεσή μας δοθέντος ενός μοντέλου. Το μοντέλο με την μέγιστη πιθανοφάνεια είναι και αυτό που προσαρμόζεται καλύτερα στα δεδομένα αλλά ο όρος penalty είναι αυτός που ελέγχει την πολυπλοκότητα και οδηγεί στην επιλογή ευέλικτων και εύχρηστων μοντέλων.

Bayesian Information Criterion

Το BIC προτάθηκε το *1978* από τον Gideon Schwartz και αποσκοπεί στο ίδιο με το AIC. Η διαφορά του έγκειται στον τρόπο με τον οποίον διαχειρίζεται το penalty που δίνεται στην πιθανοφάνεια του μοντέλου.

Δίνεται από τον τύπο:

$$BIC = -2\,log(L) + plog(n),$$

όπου L είναι η πιθανοφάνεια, p είναι το πλήθος των παραμέτρων του μοντέλου υπό εξέταση και n το μέγεθος του δείγματος.

Το BIC δίνει μεγαλύτερο penalty από ότι το AIC όταν υπεισέρχεται μια νέα παράμετρος στο μοντέλο. Έτσι, το BIC θεωρείται “ασφαλέστερο” για την αποφυγή πιθανής υπερπροσαρμογής του μοντέλου στα δεδομένα.

Τα παραπάνω κριτήρια “χτίστηκαν” πάνω στην λογική του μέτρου των Kullback και Leibler για το οποίο και γίνεται μικρή αναφορά ακολούθως.

Kullback-Leibler Divergence

Η βασική ιδέα του Kullback-Leibler (KL) μέτρου, είναι να μετρήσει το κατά πόσο δύο κατανομές πιθανότητας διαφέρουν μεταξύ τους. Εφαρμόζεται κυρίως στον χαρακτηρισμό της εντροπίας κατά Shannon στα πληροφοριακά συστήματα, στην τυχαιότητα συνεχών χρονολογικών σειρών και στην πληροφορία που αποσπάται από την σύγκριση στατιστικών μοντέλων. Όταν το KL πάρει την τιμή 0, τότε οι υπό εξέταση κατανομές πιθανότητας είναι ταυτόσημες.

Έστω $p(x) > 0$ και $q(x) > 0$ δύο κατανομές πιθανότητας που αθροίζουν στο 1, $\forall x \in X$. Τότε το KL μέτρο ορίζεται ως:

$$D_{KL}(p(x)||q(x)) = \begin{cases} \sum_{x \in X} p(x) log \frac{p(x)}{q(x)}, \text{αν x διακριτή} \\ \\ \int_{-\infty}^{+\infty} p(x) log \frac{p(x)}{q(x)} dx, \text{αν x συνεχής} \end{cases}.$$

Αξίζει να σημειωθεί ότι παρόλο που το KL μετρά την απόσταση μεταξύ δύο κατανομών, δεν αποτελεί ένα μέτρο απόστασης υπο την έννοια ότι δεν πληροί τις προϋποθέσεις μιας μετρικής.

Μερικά ακόμη γνωστά κριτήρια πληροφορίας είναι τα ακόλουθα:

- Deviance Information Criterion
- Divergence Information Criterion
- Fisher Information Criterion
- Hannan-Quinn Information Criterion
- Mallows’s C_p
- Modified Divergence Information Criterion
- Rissanen’s Minimum Description Length

All the statistics in the world can't measure the warmth of a smile.

-Chris Hart

And that’s all for the day fellow statistician folks.
Until next time take care and remember…
The probability to be killed by a cow is low
but never zero!