

1. ΣΤΟ ΓΗΠΕΔΟ ΤΗΣ ΣΤΑΤΙΣΤΙΚΗΣ

Μέσο: ΤΟ ΒΗΜΑ ΚΥΡΙΑΚΗΣ_ΒΗΜΑ SCIENCE

Ημ. Έκδοσης: . . . 24/11/2019 Ημ. Αποδελτίωσης: . . . 24/11/2019

Σελίδα: 1



4-6

ΣΤΟ ΓΗΠΕΔΟ ΤΗΣ... ΣΤΑΤΙΣΤΙΚΗΣ

Ελληνες ερευνητές χρησιμοποιούν στατιστικά μοντέλα για να εξαγάγουν πολύτιμα δεδομένα σχετικά με ομαδικά αθλήματα όπως το ποδόσφαιρο – ήδη **πέρυσι βρήκαν** με σχετική ακρίβεια **την τελική τετράδα του ελληνικού πρωταθλήματος ποδοσφαίρου**



ΜΑΘΗΜΑΤΙΚΑ

ΞΕΡΕΙ ΑΠΟ ΜΠΑΛΑ Η ΣΤΑΤΙΣΤΙΚΗ;

Ερευνητές στο **Οικονομικό Πανεπιστήμιο** Αθηνών χρησιμοποιούν στατιστικά εργαλεία και εξαγουν πολύτιμα δεδομένα για τους αθλητικούς αγώνες - με ένα απλό μοντέλο κατάφεραν να προβλέψουν πέρυσι με αρκετά μεγάλη ακρίβεια την τελική τετράδα της Σούπερ Λίγκας

ΤΟΥ ΑΛΚΗ ΓΑΛΔΑΔΑ
algaldada@yahoo.gr

Η Αθήνα τον Αύγουστο καιγόταν. Και μια από τις ημέρες εκείνου του μήνα μπόρεσαμε να συναντηθούμε για πρώτη φορά στο γραφείο του με τον κ. **Γιάννη Ντζούφρα**, καθηγητή της Στατιστικής στο **Οικονομικό Πανεπιστήμιο** Αθηνών. Όταν μου πρότεινε να πάμε να ρίξουμε μια ματιά και στο «εργαστήριο», στην αρχή σκέφθηκα πως δεν θα υπήρχε ψυχή, παρόμοια με τους απόλυτα ερήμους εκείνη την ημέρα διαδρόμους στο πολυτελές κτίριο της οδού Τροίας 2. Όταν όμως με διαβεβαίωσε πως υπήρχε κόσμος εκεί και δούλευε, προπτυχιακοί και μεταπτυχιακοί φοιτητές μαζί, φαντάστηκα έναν χώρο όπου θα επικρατούσε ένταση, φασαρία, θα έβλεπα ταλαιπωρημένα πρόσωπα, πίνακες γεμάτους με ακατάνοη για τους έξω οχήματα, ίσως απομεινάρια από «ντελβέρσι» φαγητό σε πάγκους.

Και όμως όχι. Περνώντας τη χωρίς τζάμια πόρτα του εργαστηρίου σε υποδεχεται μια ακόμη πιο πυκνή ησυχία σε σχέση με τους διαδρόμους. Υπήρχαν πράγματι κάποια άτομα που δούλευαν Αύγουστο στο μίνα, που σήκωναν για λίγο το βλέμμα τους από τις οθόνες για να μας χαϊρεθούν και που μετά αθόρυβα επέστρεφαν σε ό, τι έκαναν πριν μπούμε.

Στατιστική σιωπηλά εκτελούμενη, και τουλάχιστον εδώ, μια κάθε άλλο παρά «σέξι» δουλειά, όπως την είχε χαρακτηρίσει ακριβώς πριν από δέκα χρόνια ο **Χαλ Βάριαν**, επικεφαλής για τα οικονομικά της Google. «Εκτός και αν με τη λέξη σέξι», είχε απεύσει να γράψει η Harvard Business Review, «εννοείται ότι κάποιος θα έχει σπάνιες ικανότητες που (θα) έχουν μεγάλη ζήτηση».

Δεν θέλει και πολύ για να παρασυρθείς πάντως. Εχοντας παρακολουθήσει παλαιότερα στο HUB την ομιλία «The funny side of statistics», του κ. **Δημήτρη Καρλή**, καθηγητή επίσης στο Τμήμα Στατιστικής του **Οικονομικού Πανεπιστημίου** Αθηνών που ήταν κάπως σαν stand-up comedy, επηρεάζεσαι σε σχέση με αυτή την περιχόη της ανθρώπινης γνώσης καθώς βλέπουν το φως κάθε τόσο και απίθανες στατιστικές που κατά καιρούς έχουν πραγματοποιηθεί (όπως αυτή για το «urinal problem», μια στατιστική μελέτη για το ποια θέση πρέπει να διαλέξεις σε μια δημόσια τουαλέτα ανδρών ώστε να ελαχιστοποιείται η πιθανότητα να έλθει κάποιος στην ακριβώς διπλανή! Carleton University, 2010). Όπως έγραψε ο **Τζον Άλεν Πάουλος**, ο στατιστική παρουσιάζεται με ιμπεριαλιστικές τάσεις απέναντι στις άλλες επιστήμες, θέ-



Ο καθηγητής Στατιστικής στο **Οικονομικό Πανεπιστήμιο** Αθηνών, **Γιάννης Ντζούφρας**

λει να μπει παντού, και εδώ τη βοηθάει πλέον και η πλημμύρα δεδομένων που έφεραν Google και Facebook προς επεξεργασία στις άκρες των δακτύλων του οποιουδήποτε. Γι' αυτό έχουμε γεμίσει και από κάθε είδους ευφυολογήματα για το θέμα Στατιστική του τύπου:

«Στον Θεό έχουμε εμπιστοσύνη, αλλά όλοι οι άλλοι πρέπει να μας δείξουν τα δεδομένα που διαθέτουν»!

Θέμα σωστής χρήσης

Το 1940, πριν από την εμφάνιση του εμβολίου κατά της πολιομυελίτιδας στις Ηνωμένες Πολιτείες οι «ειδικοί» παρατήρησαν αύξηση των κρουσμάτων σε συνάρτηση με την αύξηση στην κατανάλωση... παγωτού και αναψυκτικών. Εγιναν συστάσεις στον κόσμο να ελαττώσει την κατανάλωσή τους. Αλλά αυτό δεν ήταν το σωστό «εμβόλιο». Αλλά κατάλαβαν κάποια στιγμή ότι η ασθένεια παρουσάζει έξαρση τους καλοκαιρινούς μήνες. Ναι, η Στατιστική όταν δεν χρησιμοποιείται σωστά μπορεί να οδηγήσει και σε τρελά συμπεράσματα. Στο συγκεκριμένο παράδειγμα δεν έγινε σωστή ερμηνεία των αποτελεσμάτων καθώς στην ανάλυση δεν ελήφθη υπόψη ο παράγοντας της ζέσης, ο οποίος

είναι ο κοινός συντελεστής που προκαλεί και τα δύο φαινόμενα.

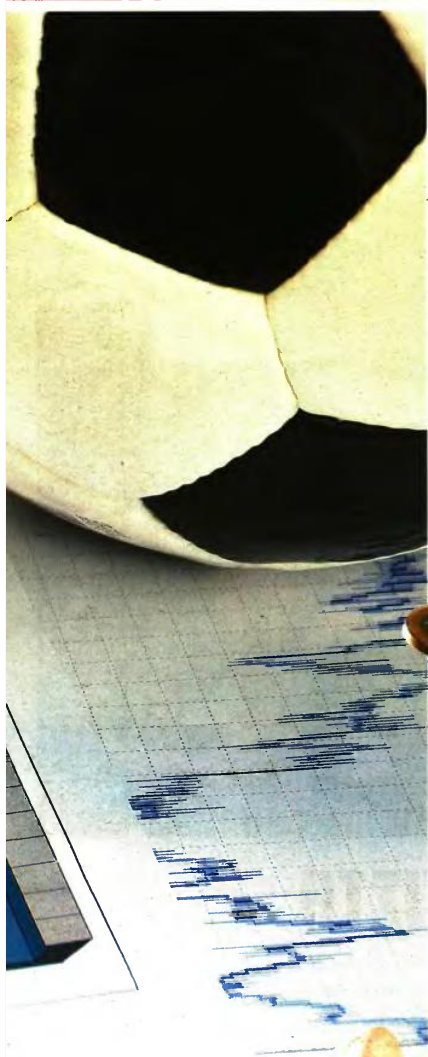
Σήμερα πάντως από την απλή συλλογή δεδομένων και συσχέτιση τα πράγματα έχουν προχωρήσει προς την κατεύθυνση που μας είχε παρουσιάσει από τους πρώτους ο **Γουίλιαμ Γκιμπσον** στα βιβλία του τη δεκαετία του '80. Τεράστιες ποσότητες δεδομένων, κάποια από αυτά πεταμένα στα σκουπίδια, που όταν υποστούν την κατάλληλη επεξεργασία αποφέρουν μεγάλα κέρδη. Data Science για τα Big Data. Πόσο μεγάλο όμως είναι το big; Κάποια petabytes (1,024 terabytes, δηλαδή 1,024 τρισεκατομμύρια bytes) ίσως και exabytes (1,024 petabytes).

Με την αρχή του νέου αιώνα είχαμε τη διαφοροποίηση μεταξύ της θεωρίας των Πιθανοτήτων και της Μαθηματικής Στατιστικής από τη μια και από την άλλη της προχωρημένης εφαρμογής τους για τη λεγόμενη Ανάλυση Δε-

δομένων ή Data Analytics, όπως είναι ο πιο μοντέρνος όρος. Που δεν μπορεί να υπάρξει χωρίς το πρώτο μέρος γιατί αποτελεί την τεκμηρίωση και την αιτιολόγηση της χρήσης τους.

Μεγάλο κομμάτι της λεγόμενης Μηχανικής Μάθησης (Machine Learning) δεν είναι τίποτα άλλο από ευρέως χρησιμοποιούμενες τεχνικές στατιστικής (αλλά φυσικά δεν είναι μόνο αυτό). Η Μηχανική Μάθηση βασίζεται στην ιδέα ότι μπορούν να κατασκευαστούν μηχανές που επεξεργάζονται δεδομένα και από αυτά μπορούν «να μάθουν» κάποια πράγματα χωρίς να βρίσκονται κάτω από τη συνεχή επίβλεψη του ανθρώπου. Λέγονται έναν κατακλυσμό από στοιχεία και σε αυτά προσπαθούν να αντιστοιχίσουν κάποια λειτουργία. Βελτιώνονται σε ακρίβεια αλλά δεν παρουσιάζουν την εξελικτική ικανότητα του ανθρώπου και σπερδύονται εντελώς της δημιουργικής





Ο ΠΙΘΑΝΟΤΗΤΕΣ

τη ας πούμε) πλευρά. Το αντίστροφο κατά τον Μπέις είναι, αν η μπάλα σταματήσει μέσα σε μια περιοχή με απόσταση X από τη μια πλευρά, χωρίς να γνωρίζεις το μήκος της μεγάλης πλευράς του τραπεζιού του μπιλιάρδου να βρεις ποια είναι η πιθανότητα να έχει ένα συγκεκριμένο μήκος, ας πούμε M . Για παράδειγμα δεν θα λέγαμε ότι είναι πιθανόν η πλευρά αυτή να έχει μήκος 100 μέτρα. Στην ουσία όμως καλούμαστε να δώσουμε τη λύση αφού βάλουμε μέσα και έναν παράγοντα που προκύπτει από εντελώς προσωπική εκτίμηση.

Και αυτό είναι που προκαλέσει κατά καιρούς πολλές συζητήσεις και αμφισβή-
τησεις.

Τα νομίσματα και οι πιθανότητες

Σήμερα υπάρχουν κάποιοι τύποι, βγαλμένοι από τη θεωρία των πιθανοτήτων και ειδικά των λεγόμενων «υπό συνθήκη» ή δεσμευμένων πιθανοτήτων που δίνουν τη δυνατότητα να υπολογιστεί κάτι που είναι κανονικά ένα πρόβλημα της Στατιστικής. Παράδει-

Όταν εκφράζονται αριθμητικά οι πιθανότητες για να συμβεί κάποιο φυσικό φαινόμενο ή να μεταβληθεί ένας πληθυσμός, οι επιστήμονες που ασχολούνται με αυτό ποτέ δεν δίνουν στην πιθανότητα τιμές 0 ή 1, που σημαίνει πως δεν πρόκειται ποτέ να συμβεί κάτι ή αντίθετα ότι θα συμβεί οποιαδήποτε. Στα υποθετικά πειράματα όμως, όπως σε ένα νόμισμα που έχει και στις δύο όψεις «κορόνα», τότε η πιθανότητα για «κορόνα» είναι ακριβώς 1 και για «γράμματα» ακριβώς 0.

πιθανότητα να έχω στρίψει το παραποιημένο νόμισμα; Παίρνοντας ένα από τα νομίσματα στο χέρι μου έχω πιθανότητα ίση με $1/3$ να είναι το παραποιημένο. Αυτή είναι η εύκολη και από τα πριν πιθανότητα. Ποια είναι όμως η πιθανότητα να είναι το παραποιημένο αφού έχει προηγηθεί το στρίψιμο 3 φορές στο οποίο εμφανίστηκαν γράμματα και τις 3 φορές; Δηλαδή αφού γνωρίζω το τι έγινε μετά; Χρειάζονται κάποιοι συλλογισμοί και υπολογισμοί γι' αυτό. Τους έχω κάνει για λογαριασμό του αναγνώστη και τους δίνω εδώ περιληπτικά για να μη γεμίσει η σελίδα.

Όλα ξεκινούν φτιάχνοντας παραστατικά τρεις κλάδους, όσους και τα νομίσματα. Η πιθανότητα να πήρα στο χέρι μου το καθένα από αυτά είναι $1/3$. Μετά, αν ήταν το παραποιημένο, η πιθανότητα να είναι «γράμματα» είναι και στα τρία στρίψιμα 1, αφού πάντα αυτό θα βγαίνει. Αντίθετα, για τα άλλα δύο η πιθανότητα είναι κάθε φορά $1/2$. Σύμφωνα με τη θεωρία του Μπέις και έναν τύπο που προκύπτει από αυτήν, η πιθανότητα για τα δύο σωστά νομίσματα βγαίνει $1/24$ (από το γινόμενο: $1/3$ επί $1/2$ επί $1/2$) ενώ για το παραποιημένο $1/3$. Τελικά, αν όλα αυτά μπουν στον τύπο του Bayes βγαίνει ότι η πιθανότητα να είναι ένα από τα σωστά

ισούται με 20% και να είναι το παραποιημένο φυσικά είναι 80%. Εφαρμόζοντας δηλαδή τον συλλογισμό του Μπέις αντί να θεωρείς πως έχεις πιθανότητα $1/3$ να πήρες το παραποιημένο νόμισμα, έπεται από ένα γεγονός όπως η εμφάνιση του «γράμματα» 3 φορές επανέρχεται και βγάζεις από αυτό το αποτέλεσμα μια νέα ενημερωμένη πιθανότητα πολύ πιο κοντά στην πραγματικότητα. Ε, και τι έγινε, θα ρωτήσει κάποιος.

Οι κρίσιμες αποφάσεις

Οι άνθρωποι της Στατιστικής δεν περνούν τον καιρό τους στρίβοντας νομίσματα και συλλέγοντας απλά δεδομένα. Ένα πιο ενδιαφέρον παράδειγμα μπορεί να έρθει από τον χώρο της Βιοστατιστικής, τον κλάδο της Στατιστικής που ασχολείται με την ιατρική έρευνα. Εστω ότι δίνω αίμα για να εξεταστεί κατά πόσο έχω προσβληθεί από κάποια ασθένεια. Η εξέταση αυτή είναι γνωστό πως μπορεί να είναι σε ποσοστό έως και 5% (που επι-

τρέπεται) λάθος, δηλαδή να δώσει ότι έχω προσβληθεί ενώ αυτό δεν συμβαίνει. Λαμβάνω υπόψη μου ότι ο 1 στους 100 στην πόλη όπου ζω έχει την ασθένεια. Υπάρχει όμως και σε ποσοστό 10% η πιθανότητα να βγει ότι είμαι υγιής αλλά στην πραγματικότητα να έχω μολυνθεί. Ποια είναι τελικά η πιθανότητα να έχω την ασθένεια αυτή; Αντε βρες άκρη με όλα αυτά όταν μάλιστα κάτι τέτοιο συμβαίνει σε ένα κάποιο σημαντικό ποσοστό ενός πληθυσμού. Εδώ επίσης μπορεί να βοηθήσει το θεώρημα του Μπέις και κάνοντας τις κατάλληλες πράξεις να μας δώσει πιθανότητα μόλις 15,4% προσβολής από την ασθένεια όταν το τεστ βγει θετικό δείχνοντας ότι έχω την ασθένεια.

Χωρίς τη σωστή στατιστική ανάλυση λοιπόν, ακολουθώντας ένα προκαθορισμένο ερευνητικό πρωτόκολλο, δεν μπορεί να υπάρξει ιατρική έρευνα ή να εγκριθεί η χρήση οποιουδήποτε νέου φαρμάκου. Αλλά δεν είναι μόνον αυτό. Οι κλινικές δοκιμές για την αποτελεσματικότητα των νέων φαρμάκων και η τελική έγκριση για να κυκλοφορήσουν εξαρτώνται από στατιστικά πιστοποιημένα αποτελέσματα. Και ειδικά από ένα μέγεθος που είναι γνωστό ως «η τιμή-p» (p-value) ή (το παρατηρούμενο) «επίπεδο σημαντικότητας».

Για παράδειγμα, μια φαρμακευτική εταιρεία ερευνά το αν ένα καινούργιο σκεύασμα περιορίζει την κατάθλιψη σε μια ομάδα εθελουσών, συγκρίνοντας τη συμπεριφορά τους με τα μέλη μιας άλλης ομάδας στην οποία δεν έχει χορηγηθεί η υπό εξέταση δραστική ουσία. Αυτό που έχει να κάνει δεν είναι να αποδείξει ότι η ουσία αυτή καταπολεμά την κατάθλιψη. Αρκεί να συγκεντρώνει αρκετά στοιχεία ώστε να καταρρίψει την υπόθεση ότι δεν την καταπολεμά. Και εκεί χρησιμοποιείται ως στατιστικό όριο το «επίπεδο σημαντικότητας» ή «τιμή-p», που συνήθως είναι ίσο με 5%. Ετσι η απόκλιση από την υπόθεση αυτή αρκεί να αντιστοιχεί σε τιμή-p μικρότερη αυτού του 5%. Ένα όριο που καθιερώθηκε το 1925 και από τότε παραμένει

φαντασίας που οδηγεί τον άνθρωπο σε λογικές αποφάσεις ακόμα και σε καταστάσεις που συναντά για πρώτη φορά. Πολλοί πιστεύουν πως στη ρίζα αυτής της όλο και πιο δυνατής σε υπολογιστική ικανότητα πορείας βρίσκεται η ιδέα ενός ανθρώπου του 16ου αιώνα.

Ο αιδεσμέτατος Μπέις

Εζησε από το 1702 έως το 1761 και έγινε διάσημος από μια εργασία του που δημοσιεύθηκε αφού είχε φύγει από τη ζωή. Ήταν Άγγλος κληρικός αλλά παρακολούθησε και παραδόσεις Μαθηματικών στη Σκωτία. Αφησε κάποια γραπτά που πολύ αργότερα εκτιμήθηκε η αξία τους από μαθηματικούς όπως ο *Πιερ Σιμόν Λαπλάς*. Από εκεί και πέρα το να ασχολείσαι με τη Στατιστική και να ακολουθείς τις ιδέες του ήταν για δεκαετίες κάτι σαν αίρεση. Από το 1980 περίπου και μετά η θεωρία του Μπέις είναι ένας ακόμη κλάδος της επιστήμης της Στατιστι-

κής με έδρες και καθηγητές σε όλα τα Πανεπιστήμια. Πρώτα βέβαια μαθαίνει κανείς να προσφέρει σωστά το όνομα του Τόμας Μπέις και όσους εργάζονται με εργαλεία που έχουν την καταγωγή τους στην εργασία εκείνη. Όλους αυτούς τους αποκαλούν «Μπεϊσιανούς» ή «Μπεϊζιανούς». Κλεισμένη σε τρεις φράσεις είναι η σκέψη του Μπέις: Είναι βολικό όταν γνωρίζεις το αίτιο να προσδιορίζεις την πιθανότητα ενός αποτελέσματος. Το άβολο είναι το αντίστροφο. Να εκτιμήσεις την πιθανότητα του αιτίου από το αποτέλεσμα. Σε ένα δικό του παράδειγμα ισχυριζόταν ότι είναι εύκολο, χτυπώντας μια μπάλα του μπιλιάρδου, έπειτα από πολλές ανακλάσεις στα τοιχώματα, να εκτιμήσεις την πιθανότητα να σταματήσει μέσα σε μια περιοχή που απέχει το πολύ X εκατοστά από τη μια (τη μικρή) κάθε-

μα: Μου δίνουν τρία νομίσματα. Τα δύο είναι απόλυτα σωστά, οπότε μπορεί να έλθει «κορόνα» ή «γράμματα» με ίσες πιθανότητες. Στο τρίτο έχουν επέμβει και έχει μόνο «γράμματα» και από τις δύο πλευρές. Διαλέγω ένα από τα τρία στην τύχη και το στρίβω 3 φορές. Έρχεται και τις τρεις «γράμματα». Ποια είναι η

0,5%

Στατιστικοί και γιατροί έχουν ζητήσει το περιθώριο λάθους στη δόση ενός φαρμάκου που δοκιμάζεται να κατέβει από το 5% στο 0,5%. Πρόκειται για μια πολύ σημαντική μείωση, που σίγουρα θα συναντήσει τη σθεναρή αντίδραση των φαρμακοβιομηχανιών.



ΜΑΘΗΜΑΤΙΚΑ

2,5

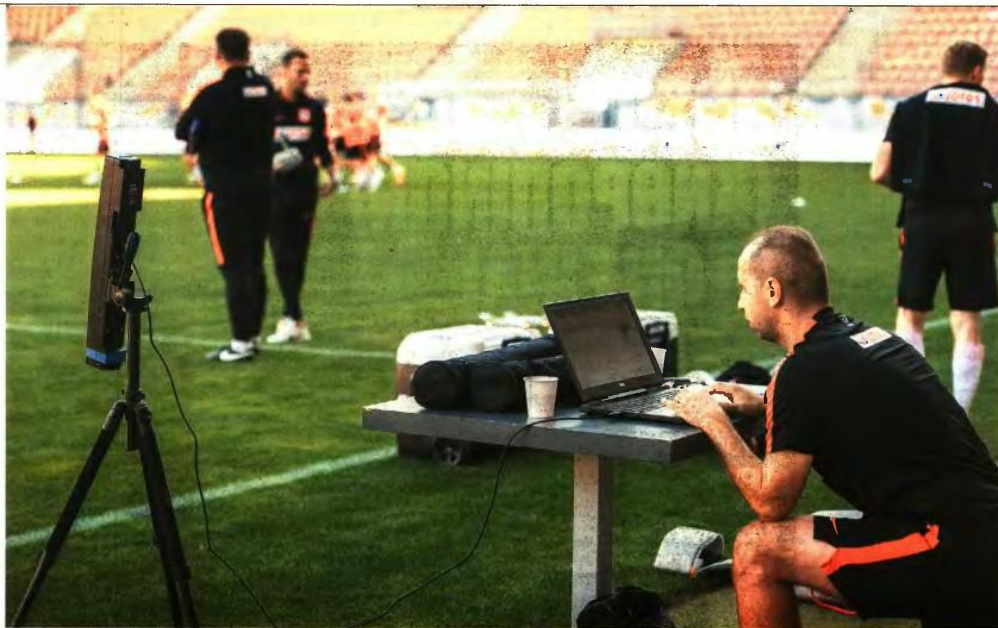
Αλλά πίσω του σέρνει 18 μηδενικά. Διότι πρόκειται για τα 2,5 κουιντίλιον bytes, ή 2,5 δισεκατομμύρια δισεκατομμυρίων που δημιουργούμε όλοι εμείς σε ολόκληρο τον κόσμο με τις δραστηριότητές μας στα διάφορα παγκόσμια δίκτυα. Και από αυτά τα big data δημιουργούνται οι στατιστικές που προδίδουν και τις πιο απόκρυφες συνηθειές μας στους κατασκευαστές των εμπορικών προϊόντων.

ΣΥΝΕΧΕΙΑ ΑΠΟ ΤΗΝ
ΠΡΟΗΓΟΥΜΕΝΗ ΣΕΛΙΔΑ

(βολικά) μαρμαρωμένο. Πολύ πρόσφατα, ο διεθνούς φήμης καθηγητής Στατιστικής Βάλεν Τζόνσον από το Πανεπιστήμιο του Τέξας, άνοιξε μια συζήτηση για τα προβλήματα «αναπαραγωγιμότητας» της έρευνας. Της βασισμένης στους στατιστικούς ελέγχους με επίπεδο σημαντικότητας 5%, οδηγώντας στην πρότασή του για αναθεώρηση (δραστικά προς τα κάτω) αυτού του ορίου. Τον Μάρτιο του 2019, 800 επιστήμονες με αντικείμενο τη Στατιστική υπέγραψαν μια έκκληση δημοσιευμένη και στο περιοδικό Nature [567, 305-307 (2019)], για να πάψει να κρίνεται η εγκυρότητα ενός πειράματος, μιας έρευνας ή μιας δημοσίευσης από αυτό το όριο και να μη χρησιμοποιείται καν ο όρος «στατιστικά σημαντική». Όπως αναφέρει και ο κ. Ντζούφρας «χρειάζεται ακόμα αρκετή έρευνα και συζήτηση ώστε να λήξει αυτή η πρόσφατη διαμάχη και να οδηγηθούμε από τους κλασικούς ελέγχους υποθέσεων (statistical significance tests) προς εναλλακτικούς ελέγχους που να βασίζονται στην Μπεϊζιανή στατιστική (με βάση τον λεγόμενο Bayes Factor)».

Ο κ. Ντζούφρας, που έχει βραβευτεί με τιμητική μνεία στην εκδήλωση βράβευσης επιστημονικών βιβλίων PROSE Awards στην κατηγορία των Μαθηματικών για το σύγγραμμά του «Bayesian Modeling Using WinBUGS» (J. Wiley), αποφάσισε να επεκτείνει το ερευνητικό του πεδίο και στην επεξεργασία στοιχείων με εφαρμογή στατιστικών μεθόδων για διάφορα αθλήματα, μέσα από την ομάδα Sports Analytics Group (<https://aueb-analytics.wikisite.com/sports/aim-and-scope>).

Το 2000, αν και δεν ασχολείται με το σποήνμα, μαζί με τον Δ. Καρλή, ως διδακτορικοί φοιτητές, γράφουν μια εργασία για προγνωστικά ποδοσφαίρου που βελτιωμένη δημοσιεύεται και στο περιοδικό Statistician της Royal Statistical Society το 2003. Το άρθρο είχε ευρεία απήχηση στη διεθνή επιστημονική κοινότητα, αν κρίνουμε από τις πολυάριθμες αναφορές του, σύμφωνα με το Google Scholar. Ετσι σήμερα υπάρχει μια μικρή ομάδα με πέντε συνεργάτες-ερευνητές, δυο διδακτορικούς φοιτητές και πέντε εξωτερικούς συνεργαζόμενους καθηγητές, που κάνει έρευνα



«ΠΑΣΑ» ΣΤΟΝ ΥΠΟΛΟΓΙΣΤΗ

και διοργανώνει συναντήσεις με βάση το τρίπτυχο: Επιστήμη - Συγκέντρωση στατιστικών στοιχείων - Ανάλυση δεδομένων, με πηγή τα αθλητικά γεγονότα εδώ και στο εξωτερικό. Το ποδόσφαιρο, όπως λέει ο κ. Ντζούφρας, «αποτελεί πρόκληση για έναν επιστήμονα όσον αφορά την ανάπτυξη μοντέλων

και αλγορίθμων πρόβλεψης διότι έχει λίγα «γεγονότα» (εμείς οι άλλοι τα λέμε γκολ, κάρτες, αποβολές) και μια σκόπιμα εμφνευμένη πιθανότητα να μη νικάει πάντα ο καλύτερος (σε αντίθεση π.χ. με την καλαθοσφαίριση όπου η πιθανότητα νίκης του καλύτερου είναι πολύ μεγαλύτερη). Επίσης οι πε-

ρισσότερες επαγγελματικές αθλητικές ομάδες σε όλο τον κόσμο, με εξαίρεση αυτές των Ηνωμένων Πολιτειών, δαπανούν χρήματα όχι για να κερδίσουν περισσότερα όπως θα περίμενε κάποιος από μια αθλητική επένδυση αλλά, οδηγούμενες από το έμφροτο συναίσθημα της νίκης, για να κερδίσουν στο γήπεδο (δηλαδή είναι win-maximizers αντί για profit-maximizers που θα έπρεπε να είναι ως επιχειρήσεις)».

Στην προσπάθεια μαθηματοποίησης των προβλέψεων διακρίνονται δυο περιπτώσεις. Πρόβλεψη του ακριβούς αποτελέσματος ενός αγώνα και πρόβλεψη του τελικού αποτελέσματος με τη μορφή νίκης/ισοπαλίας/ήττας (1X2). Ενοείται πως μπορεί κάποιος να φορτώσει το μοντέλο του με διάφορους παράγοντες επηρεασμού του αποτελέσματος, από τον καιρό μέχρι το ποιος είναι ο προπονητής, πιστεύοντας πως έτσι πηγαίνει πιο κοντά στην πραγματική κατάσταση των ομάδων. Ο άλλος δρόμος είναι να κρατήσεις το μοντέλο πιο απλό και να δεις πώς πάει. Ξεκινώντας από την πιο απλή στατιστική κατανομή πρόβλεψη για το πιο πιθανό επόμενο αποτέλεσμα με βάση όλα τα προηγούμενα, διαλέγεις τη λεγόμενη κα-

τανομή Poisson που μετά μπορεί να πάρει και πιο περίπλοκες ή περίτεχνες μορφές (οι μαθηματικοί τις ονομάζουν sophisticated ή elegant θέλοντας να εκφράσουν την «ομορφιά» τους και την ευελξία τους).

Στι Σούπερ Λίγκες

Στις πρώτες προσπάθειες που έγιναν κρατήθηκαν τα πράγματα όσο γινόνταν πιο απλά. Μπήκαν μόνον παράγοντες σχετικοί με το αν παίζει μια ομάδα εντός ή εκτός έδρας, ότι η επίτευξη ενός γκολ είναι ανεξάρτητη από τα προηγούμενα στον ίδιο αγώνα και περνάς με έναν ειδικό τρόπο τα στοιχεία σε ένα φύλλο του Excel ή σε ένα πιο εξειδικευμένο λογισμικό Στατιστικής όπως η R. Τελικά οι προβλέψεις βγαίνουν μέσα από μια μαθηματική διαδικασία και τύπους που βασίζονται στο μοντέλο πιθανοτήτων που έχει χρησιμοποιηθεί. Σε μια ομιλία του τον Φεβρουάριο του 2019 ο κ. Ντζούφρας, όταν οι ομάδες της Σούπερ Λίγκας είχαν παίξει 19 από τους συνολικά 30 αγώνες, με το εντελώς απλό μοντέλο (ένα μοντέλο βανίλια, δηλαδή ό,τι πιο σκέτο ή απλό, όπως λένε οι άνθρωποι της Στατιστικής), χωρίς τις πολλές παραδοχές που θα το έκαναν πιο ακριβές, παρουσίασε τις προβλέψεις του για την τελική βαθμολογία. Ειδ-

νε κατά σειρά: ΠΑΟΚ>77 βαθμοί, Ολυμπιακός>68, ΑΕΚ>62, Ατρόμητος>51. Η τελική βαθμολογία ως γνωστόν ήταν: 80, 75, 57, 52. Πέφτοντας έτσι έξω 6% κατά μέσον όρο.

Σίγουρα οι άνθρωποι του Sports Analytics Group δεν στοχεύουν να βγάλουν λεφτά από τις προβλέψεις στο σποήνμα. Θέλουν να καταλάβουν πώς η επιστήμη της Στατιστικής θα μπορούσε να βοηθήσει σε διάφορες περιπτώσεις αθλητές, προπονητές και ομάδες. Τα ερωτήματα πάντως για την επίτευξη αυτού του σκοπού είναι αρκετά. Οι ελληνικές ομάδες δεν έχουν δείξει ενδιαφέρον για το κέρδος που μπορεί να τους δώσει μια επιτυχημένη επεξεργασία των δεδομένων τους. Από την άλλη, αν και πολλά δεδομένα είναι διαθέσιμα στο Διαδίκτυο, αρκετά από τα αναλυτικά δεδομένα για διάφορους αγώνες δεν είναι ευρέως διαθέσιμα ή παλινυπόδητοι πολύ ακριβά ως προϊόντα προς εκμετάλλευση σε όποιον ενδιαφέρεται. Ο δρόμος είναι ανθρωπίνος αλλά έχει επιστημονικό ενδιαφέρον. Γι' αυτό και η ερευνητική ομάδα δεν απογοητεύεται, παρά τις δυσκολίες που περιλαμβάνονται στην έλλειψη χρηματοδότησης της έρευνας και τις αναταραχές στις Σχολές.

4ο AUEB Sports Analytics Workshop

Αύριο Δευτέρα 25/11 αρχίζει το 4ο AUEB Sports Analytics Workshop που αποτελεί μια συνάντηση μεταξύ των ανθρώπων που ενδιαφέρονται για το θέμα της Στατιστικής των Αθλημάτων. Στο workshop θα συμμετάσχουν ερευνητές από την Ελλάδα και την Ιταλία και θα συνοδεύεται από ένα μάθημα βραχείας διάρκειας 12 ωρών σε Basketball Data Science από την κυρία Μαρίκα Μανιέρα, αναπληρώτρια καθηγήτρια του Πανεπιστημίου της Μηρέσιας. Επίσης το workshop θα ανοίξει όπως κάθε χρόνο ένας από τους πιο γνωστούς οικονομολόγους του αθλητισμού, ο κ. Στεφάν Κεσόν που διετέλεσε σύμβουλος του Ζαν Μαρκ Μποσάν στην πολυκροτική υπόθεση της αλλαγής για πάντα του ευρωπαϊκού ποδοσφαίρου. Περισσότερες λεπτομέρειες και τον τρόπο συμμετοχής στο workshop μπορείτε να βρείτε στην ιστοσελίδα: <https://aueb-analytics.wikisite.com/saw2019>.