Μη-Παραμετρική Στατιστική στην **R**

Εργασία 1, 4/5 Απριλίου 2015

Κάθε ένας από εσάς είναι υποψήφιος να αναλάβει μία σημαντική θέση ντετέκτιβ Στατιστικού με ρόλο να εξιχνιάζει μυστήρια που εμπλέκουν στατιστικές αναλύσεις με δεδομένα υψίστης ασφαλείας. Θα πρέπει λοιπόν να περάσετε διάφορα στάδια στα οποία θα πρέπει να αποδείξετε τις ικανότητές σας. Ξεκινάμε λοιπόν με την ανάθεση ένας πρόχειρου μυστικού αριθμού που θα διευκολύνει την επικοινωνία μας.

Ανάθεση κωδικού

(1) Σχηματίστε ένα πλαίσιο δεδομένων που να περιλαμβάνει 2 στήλες. Στην πρώτη στήλη θα πρέπει να περιλαμβάνονται τα γράμματα του αγγλικού αλφαβήτου (τα μικρά) και στη δεύτερη η αριθμητική τους αξία σε αύξουσα σειρά από το 1 μέχρι το 26. Στη συνέχεια δημιουργείστε ένα διάνυσμα χαρακτήρων ως κωδική λέξη που να περιλαμβάνει το μικρό σας όνομά και το επίθετο διαδοχικά το ένα μετά το άλλο με λατινικούς χαρακτήρες και χωρισμένα με το γράμμα w. Αν για παράδειγμα το όνομά σας είναι Μάρθα Κλάψα, τότε μετατρέπεται σε marthawklapsa. Αν λέγεστε Μπάμπης Σφίχτης, τότε babiswsfichtis, ενώ ο αντίστοιχος Σουγιάς γίνεται babiswsouyias. Σε κάθε περίπτωση πρέπει να γράψετε στον κώδικά σας τους αντίστοιχους λατινικούς χαρακτήρες που χρησιμοποιήσατε στην απόδοση του ονόματός σας. Ο ψηφιακός κωδικός τελικά αποδίδεται με εντολή που υπολογίζει το άθροισμα των αριθμών των γραμμάτων που αντιστοιχούν στην κωδική σας λέξη.

(2) Αφού υπολογίσετε τον κωδικό σας αριθμό, χρησιμοποιήστε εντολή που καθορίζει αυτόν τον αριθμό ως σπόρο στο γεννήτορα των ψευδοτυχαίων αριθμών που θα χρειαστούμε.

Ερωτήματα

(1) Προσομοιώστε 500 αριθμούς από την εκθετική κατανομή με παράμετρο 1. Κάντε ένα ιστόγραμμα των δεδομένων συμπεριλαμβάνοντας την καμπύλη της αντίστοιχης συνάρτησης πυκνότητας.

(2) Πραγματοποιήστε ελέγχους X^2 και Kolmogorov-Smirnov με μηδενική υπόθεση την $Exp(1)$ έναντι οποιασδήποτε άλλης κατανομής διαδοχικά για $n = 20, 100, 500$ (με τα ίδια δεδομένα). Σχολιάστε τα αποτελέσματα χρησιμοποιώντας τις p-value.

(3) Φτιάξτε μία συνάρτηση testExp που να παίρνει ως όρισμα N το πλήθος των ανεξάρτητων ελέγχων που θέλουμε να κάνουμε, n το μέγεθος του δείγματος σε κάθε έλεγχο και a το επίπεδο στατιστικής σημαντικότητας που θέτουμε. Η συνάρτηση αυτή θέλουμε να μας επιστρέφει ένα ζεύγος αριθμών που καταγράφει το ποσοστό των απορρίψεων των παραπάνω ελέγχων X^2 και Kolmogorov-Smirnov για τους N ελέγχους πάνω σε αντίστοιχα προσομοιωμένα δείγματα από την $Exp(1)$ (κάθε δείγμα περνάει και τους δύο ελέγχους). Τρέξτε μετά τη συνάρτηση δύο φορές, μία για $(N, n, a) = (10^5, 20, 0.05)$ και μία για $(N, n, a) = (10^5, 500, 0.05)$. Σχολιάστε τα αποτελέσματα.

(4) Στο Ερώτημα 1 έχετε ήδη προσομοιώσει 500 αριθμούς από την εκθετική κατανομή με παράμετρο 1. Με τη βοήθεια αυτών των αριθμών φτιάξτε ένα προσομοιωμένο δείγμα από την ομοιόμορφη κατανομή στο $(0, 1)$. Στη συνέχεια προσομοιώστε ένα καινούριο δείγμα μεγέθους 100 από την ομοιόμορφη κατανομή στο (0,1) με απευθείας χρήση συνάρτησης της R. Ελέγξτε με 3 διαφορετικούς τρόπους αν τα δείγματα προέρχονται από την ίδια κατανομή.

(5) Μετατρέψτε το καινούριο δείγμα των 100 τιμών σε τιμές $X_{1:100}$ που μπορούμε να υποθέσουμε ότι προέρχονται πάλι από την $Exp(1)$. Πραγματοποιήστε ελέγχους X^2 και Kolmogorov-Smirnov για να ελέγξετε τον παραπάνω ισχυρισμό σε ε.σ.σ. $a = 0.05$, για το δείγμα $X_{1:20}$, αλλά και για όλο το δείγμα $X_{1:100}$.

(6) Σας πληροφορούνε ότι μία καινούρια κατανομή είναι υποψήφια να περιγράφει την παραγωγή αυτών των δεδομένων. Λέγεται διπλωμένη κανονική και θα λέγαμε χαριτολογόντας ότι είναι μία ξεχασιάρα κανονική που δεν θυμάται πρόσημα και τα κάνει όλα θετικά. Ανακαλύψτε ποιά είναι και ρυθμίστε τη μέση τιμή και τη διασπορά της να είναι 1.

(7) Πραγματοποιήστε ελέγχους με μηδενική υπόθεση την $Exp(1)$ (έναντι οποιασδήποτε άλλης) προσομοιώνοντας δείγματα από τη διπλωμένη κανονική με μέση τιμή και διασπορά 1. Οι έλεγχοι πρέπει να πραγματοποιηθούν με τη συνάρτηση testExp που φτιάξατε, αλλά πρέπει να την τροποποιήσετε κατάλληλα ώστε να προσομοιώνει εναλλακτικά από τη συγκεκριμένη διπλωμένη κανονική. Μπορείτε να προσθέσετε ένα επιπλέον όρισμα που δηλώνει από ποιά κατανομή θα προσομοιώνει. Τρέξτε τη συνάρτηση δύο φορές, μία για $(N, n, a) = (10^4, 20, 0.05)$ και μία για $(N, n, a) = (10^4, 100, 0.05)$. Τί στόχο έχουν αυτοί οι επιπλέον έλεγχοι ; Σχολιάστε τα αποτελέσματα.

(8) Έχετε κάνει όλα τα προηγούμενα και έχετε σχολιάσει επαρκώς τα αποτελέσματά σας. Τώρα σας λένε ότι αν επιλεγείτε θα πηγαίνετε συχνά σε μία σεισμογενή περιοχή. Σας πληροφορούνε επίσης ότι η μέγιστη ένταση των σεισμών που συμβαίνουν καθημερινά εκεί έχει συνάρτηση κατανομής που δίνεται από την $F(x) = 0.5+0.5\log(1+x^{4/17})$ για $x \geq 0$ και 0 διαφορετικά. Ως μαθηματικοί συνειδητοποιείτε ότι κάτι λάθος συμβαίνει και η F πρέπει να σταθεροποιείται από κάποιο σημείο και πέρα. Ποιό σημείο x_{max} είναι αυτό ; Υπολογίστε το και στη συνέχεια προσομοιώστε ένα δείγμα 365 τιμών από αυτήν την κατανομή για να έχετε μία ιδέα σε ένα χρόνο πώς θα συμπεριφέρεται στοχαστικά η μέγιστη ένταση των σεισμών. Κάντε ένα ιστόγραμμα των δεδομένων και σε ένα άλλο γράφημα αντιπαραβάλλεται την εμπειρική συνάρτηση κατανομής με τη θεωρητική.

(9) Ότι ακολουθεί κάνει χρήση των πραγματικών δεδομένων που μαζέψαμε στην τάξη. Πραγματοποιήστε ελέγχους κανονικότητας των δεδομένων του ύψους και του βάρους των φοιτητών διακρίνοντας με βάση το φύλο. Συμπεριλάβετε τους ελέγχους Shapiro-Wilk και Lilliefors.

(10) Ελέγξτε με διάφορους τρόπους κατά πόσον υπάρχει στατιστική εξάρτηση μεταξύ των βαθμών των Πιθανοτήτων Ι και Στατιστικής Ι χρησιμοποιώντας τα ζεύγη παρατηρήσεων που προκύπτουν από τα πραγματικά δεδομένα.

(11) Ελέγξτε επίσης με το κριτήριο προσημασμένων βαθμών του Wilcoxon αν υπάρχει στατιστικά σημαντική διαφοροποίηση μεταξύ των βαθμών των Πιθανοτήτων Ι και της Στατιστικής Ι στα συνήθη επίπεδα στατιστικής σημαντικότητας.

(12) Και τώρα η τελευταία **προαιρετική δοκιμασία** για την Εργασία 1. Ελέγξτε με ένα απλό παράδειγμα αν οι μηδενικές διαφορές λαμβάνονται υπόψη στο κριτήριο του Wilcoxon. Μπορείτε να τροποποιήσετε το παραπάνω κριτήριο ώστε να συμπεριλαμβάνει και τις μηδενικές διαφορές ; Για να τα καταφέρετε πρέπει να πείσετε ότι ο τρόπος που χρησιμοποιήσατε αυξάνει την ισχύ του κριτηρίου προσημασμένων βαθμών του Wilcoxon ιδίως εκεί που εμφανίζονται πολλά μηδενικά. Όποιος τα καταφέρει δικαιωματικά θα πάρει και το πρώτο του στατιστικό αστέρι.

Η προσπάθεια πρέπει να είναι ατομική, εκτός του τελευταίου ερωτήματος!

Καλή Επιτυχία!