

Στατιστικά Προγράμματα, 15-11-06, Homework #1

Μετεωρολογικοί σταθμοί σε διάφορες περιοχές της Ελλάδας έχουν καταγράψει στοιχεία που αφορούν τις κλιματολογικές συνθήκες που επικρατούν σε αυτές από το 1994 έως σήμερα. Συγκεκριμένα κάθε ημέρα καταγράφονται 4 μεταβλητές: (1) **Date** ημερομηνία (ημέρα/μήνας/έτος) (2) **Wind** (mean wind speed) Μέση ημερήσια ταχύτητα ανέμου σε m/s. (3) **Temperat** (mean daily air temperature) Μέση ημερήσια θερμοκρασία αέρα σε βαθμούς C. (4) **Precipit** (daily precipitation) Μέση ημερήσια βροχόπτωση σε mm.

Τα δεδομένα προέρχονται από την ιστοσελίδα <http://dapper.pmel.noaa.gov/dchart/> που διατηρεί η National Oceanic and Atmospheric Administration (US Department of Commerce, <http://www.noaa.gov>). Ο κάθε φοιτητής-τρια με τελευταίο ψηφίο στον Αριθμό Μητρώου του το x θα βασιστεί στα ακόλουθα στοιχεία των ετών 1994-2006 για την περιοχή (τα αρχεία δεδομένων txt βρίσκονται στην ιστοσελίδα του μαθήματος)

Av x = 0 : Station: ALEXANDROUPOLI , 25.89E 40.8N 3m	Av x = 5 : Station: KERKYRA , 19.89E 39.6N 4m
Av x = 1 : Station: ATHINAI , 23.69E 37.9N 15m	Av x = 6 : Station: LARISSA , 22.39E 39.6N 74m
Av x = 2 : Station: HERAKLION , 25.1E 35.3N 39m	Av x = 7 : Station: TRIPOLIS , 22.3E 37.5N 652m
Av x = 3 : Station: KASTORIA , 21.19E 40.4N 604m	Av x = 8 : Station: THESSALONIKI , 22.89E 40.5N 4mm
Av x = 4 : Station: KYTHIRA , 23E 36.2N 167m	Av x = 9 : Station: METHONI , 21.69E 36.8N 34m

Περάστε τα δεδομένα στο SPSS. Η μεταβλητή date να ορισθεί ως date στο SPSS με κατάλληλο format (mm/dd/yyyy) και η τιμή -99 όπου εμφανίζεται να θεωρείται ως missing value. Στη συνέχεια, από την μεταβλητή Date κατασκευάστε τρεις νέες μεταβλητές: year, month, dayyear οι οποίες να δείχνουν το έτος (1994-2006), το μήνα (1-12) και την ημέρα του έτους (1-366) αντίστοιχα (υποδ. χρησιμοποιείστε κάποιες από τις συναρτήσεις XDATE: κάντε δεξί κλικ επάνω σε αυτές για να δείτε ποιες είναι κατάλληλες).

ΕΡΩΤΗΜΑΤΑ

(1) Εκτιμήστε τη μέση θερμοκρασία, τη μέση βροχόπτωση και μέση ταχύτητα του ανέμου στο σταθμό αυτό για την περίοδο 94-06. Εκτιμήστε τις αντίστοιχες τυπικές αποκλίσεις. Ποιες είναι οι μεγαλύτερες τιμές που σημειώθηκαν όλα αυτά τα χρόνια στο σταθμό για κάθε μία από τις μεταβλητές; (για την θερμοκρασία να δοθεί και η μικρότερη τιμή). Πότε σημειώθηκε κάθε μία από αυτές τις ακραίες τιμές;

(2) (α) Από όλες τις παρατηρήσεις κατασκευάστε τρία ιστογράμματα, ένα για την θερμοκρασία, ένα για την ταχύτητα του ανέμου και ένα για την βροχόπτωση. (β) Να κατασκευάσετε 12 ιστογράμματα για την θερμοκρασία, ένα για κάθε μήνα του έτους (μπορείτε να χρησιμοποιήσετε split file).

(3) (α) Κατασκευάστε ένα γράφημα με 12 boxplots των παρατηρήσεων της θερμοκρασίας ανά μήνα και ένα γράφημα με 12 boxplots της ταχύτητας του ανέμου ανά μήνα (ανεξαρτήτως έτους). Υπάρχουν σε αυτά έκτροπες παρατηρήσεις το μήνα Αύγουστο και πότε σημειώθηκαν (ημέρα, έτος); (β) Κατασκευάστε τα δύο προηγούμενα γραφήματα αυτή τη φορά ανά μήνα και ανά «δεκαετία» (δεκαετίες 90, 00). Συγκεκριμένα, το κάθε γράφημα να περιέχει 12x2 boxplots, 12 κόκκινα για τους μήνες των ετών 94-99 και 12 πράσινα για τους μήνες των ετών 00-06 (υποδ. χρησιμ. clustered boxplots με κατάλληλη βοηθητική μεταβλητή στα clusters). (γ) Κατασκευάστε ένα γράφημα με 13 boxplots των παρατηρήσεων της θερμοκρασίας ανά έτος και ένα γράφημα με 13 boxplots των παρατηρήσεων της ταχύτητας του ανέμου ανά έτος.

Σχολιάστε περιληπτικά τι παρατηρείτε στα γραφήματα (α), (β), (γ).

(4) Να κατασκευάσετε τα ακόλουθα 4 scatterplots (διαγράμματα διασποράς): (α) ένα που στον κάθετο άξονα να αντιστοιχεί η θερμοκρασία και στον οριζόντιο άξονα ο χρόνος (σε ημέρες από την 1-1-94 με τιμές 1-4696). (β) Ένα που στον κάθετο άξονα να αντιστοιχεί η θερμοκρασία και στον οριζόντιο άξονα ο χρόνος (σε ημέρες από την αρχή κάθε έτους με τιμές 1-366). (γ) Ένα που στον κάθετο άξονα να αντιστοιχεί η βροχόπτωση και στον οριζόντιο άξονα ο χρόνος (σε ημέρες από την αρχή κάθε έτους με τιμές 1-366). (δ) Ένα που στον κάθετο άξονα να αντιστοιχεί η θερμοκρασία και στον οριζόντιο άξονα ο χρόνος (σε ημέρες από την αρχή κάθε έτους με τιμές 1-366) μόνο για τα έτη 1995, 2005 (με διαφορετικό χρώμα οι παρατηρήσεις κάθε έτους). Περιγράψτε σύντομα τι παρατηρείτε (Τα σημεία στα παραπάνω γραφήματα να δίνονται με “•”)

(5) (α) Να κατασκευάσετε ένα ραβδόγραμμα (barchart) με 12 ράβδους: το ύψος κάθε ράβδου να αντιστοιχεί στη μέση θερμοκρασία ανά μήνα (υποδ. χρησιμοποιήστε other summary function: MEAN). Πάνω σε κάθε ράβδο να αναγράφεται και η αντίστοιχη μέση τιμή. Να κάνετε το ίδιο για τη βροχόπτωση και τη ταχύτητα του ανέμου. Τι παρατηρείτε; (β) Να κατασκευάσετε ένα συγκριτικό ραβδόγραμμα (clustered barchart) με 12x2 ράβδους (12 πράσινες και 12 κόκκινες): το ύψος κάθε κόκκινης ράβδου να αντιστοιχεί στη μέση θερμοκρασία ανά μήνα για το έτος 1995 ενώ το ύψος κάθε πράσινης ράβδου να αντιστοιχεί στη μέση θερμοκρασία ανά μήνα για το έτος 2005. Πάνω σε κάθε ράβδο να φαίνεται και η αντίστοιχη μέση τιμή. Να κάνετε το ίδιο για τη βροχόπτωση και τη ταχύτητα του ανέμου. Τι παρατηρείτε; (γ) Να κάνετε την ερώτηση (α) τώρα ανά έτος (αντί ανά μήνα). Να κάνετε την ερώτηση (α) και πάλι ανά έτος αλλά μόνο για το μήνα Αύγουστο (χρησιμοποιείστε select cases). Τι παρατηρείτε;

(6) (α) Να κατασκευάσετε ένα γράφημα που αποτελείται από 13 γραμμές: μία για κάθε έτος που να δείχνει τη μέση θερμοκρασία ανά μήνα (Multiple line graph). Επίσης, να κατασκευάσετε ένα γράφημα που να αποτελείται από 12 γραμμές: μία για κάθε μήνα που να δείχνει τη μέση θερμοκρασία ανά έτος (Multiple line graph). (β) Να κατασκευάσετε ένα γράφημα το οποίο θα αποτελείται από 13 διαστήματα εμπιστοσύνης για τη μέση ετήσια θερμοκρασία (συντελεστού 95% το καθένα), ένα για κάθε έτος (error bar).
(7) Να δώσετε έναν πίνακα με τις εκτιμήσεις της μέσης θερμοκρασίας ανά μήνα (σημειακά και με δ.ε. 95%) ανεξαρτήτως έτους. Ποια είναι η μέγιστη και ποια η ελάχιστη τιμή της (μέσης ημερήσιας) θερμοκρασίας που σημειώθηκε σε καθένα από τους 12 μήνες. Να γίνει το ίδιο και για τη βροχόπτωση (χρησιμ. explore, κρατήστε από τους πίνακες που παράγονται μόνο ό,τι ζητείται και φτιάξτε έναν συγκεντρωτικό πίνακα).
(8) (α) Να κατασκευάσετε δύο νέες μεταβλητές <code>abs_temp</code> και <code>abs_prec</code> οι οποίες να είναι ίσες με το ακέραιο μέρος της θερμοκρασίας και της βροχόπτωσης αντίστοιχα. Να κατασκευάσετε ένα πίνακα συχνοτήτων για τις τιμές κάθε μιας από αυτές. Τι ποσοστό ημερών είχαμε (μέση ημερήσια) θερμοκρασία μεγαλύτερη από 20 βαθμούς C; Τι ποσοστό ημερών είχαμε βροχόπτωση μεγαλύτερη από 5mm. (β) Να κάνετε τους ίδιους πίνακες με το (α) τώρα μόνο για το μήνα Αύγουστο (select cases).
(9) Να κατασκευάσετε (με recode) μία νέα μεταβλητή <code>c_temp</code> η οποία θα παίρνει τιμές 0, 1, 2, 3, 4 όταν η θερμοκρασία παίρνει τιμές στα διαστήματα $(-\infty, 0]$, $(0, 10]$, $(10, 20]$, $(20, 30]$, $(30, 40]$ αντίστοιχα (προσοχή στα missing values). Να κατασκευάσετε έναν πίνακα που να δείχνει πόσες ημέρες είχαμε τιμές στα παραπάνω διαστήματα ανά μήνα (crosstabulation) που να περιλαμβάνει και τα αντίστοιχα ποσοστά ανά μήνα.
(10) Να ελέγξετε αν η μέση θερμοκρασία στην περιοχή του σταθμού είναι ίση με μ_0 βαθμούς (για όλο το έτος). Να ελέγξετε αν η μέση θερμοκρασία στην περιοχή του σταθμού είναι ίση με $\mu_0 + 10$ βαθμούς για το μήνα Αύγουστο (να γίνει ο αμφίπλευρος και οι αντίστοιχοι μονόπλευροι έλεγχοι). Το μ_0 θα είναι, ανάλογα με τα δεδομένα που έχετε: ALEXANDROUPOLI: 14.5, HERAKLION: 18.5, KERKYRA: 17.5, LARISSA: 15.5, THESSALONIKI: 15.5, ATHINAI: 18.5, KASTORIA: 13, KYTHIRA: 18.5, METHONI: 18, TRIPOLIS: 14
(11) (α) Ελέγξτε αν η μέση τιμή της θερμοκρασίας είναι ίδια τον μήνα Ιανουάριο και τον μήνα Φεβρουάριο. Υπάρχει στατιστικά σημαντική διαφορά στις αντίστοιχες διασπορές; (β) Ελέγξτε αν η μέση τιμή της θερμοκρασίας είναι ίδια τον Αύγουστο του 1995 και τον Αύγουστο του 2005.
(12) Με βάση τις παρατηρήσεις του έτους 2005: Εξετάστε αν η μέση θερμοκρασία στο σταθμό που μελετάτε είναι ίδια με τη μέση θερμοκρασία στο σταθμό που αντιστοιχεί στο $x+1 \bmod 10$ (όπου x το τελευταίο ψηφίο του Α.Μ. σας). Υπάρχει θετική συσχέτιση μεταξύ των τιμών των θερμοκρασιών στους δύο σταθμούς;
(13) Σε μια περιοχή έχει βρεθεί ότι το ποσοστό των ημερών με ταχύτητα ανέμου που ανήκει στα διαστήματα $(0, 1.5]$, $(1.5, 2.7]$, $(2.7, 4.5]$, $(4.5, \infty]$ είναι 20%, 30%, 30%, 20% αντίστοιχα. Ελέγξτε αν ισχύει το ίδιο και στο σταθμό που μελετάτε.
(14) Να κατασκευάσετε έναν πίνακα που να δείχνει το πλήθος και το ποσοστό των ημερών ανά μήνα που καταγράφηκε άνεμος $> 7 \text{ m/s}$. Ένας τέτοιος άνεμος καταγράφεται με την ίδια πιθ. σε όλους τους μήνες;
(15) Να ελέγξετε αν οι τιμές της θερμοκρασίας και της ταχύτητας του ανέμου τον μήνα Δεκέμβριο μπορεί να θεωρηθεί ότι προέρχονται από κάποια κανονική κατανομή (εκτιμήστε και τις παραμέτρους της). Να χρησιμοποιήσετε ιστόγραμμα, Q-Q plot, P-P plot, K-S τεστ.
(16) Στο ερώτημα αυτό να κρατήσετε μόνο τις παρατηρήσεις που αντιστοιχούν σε ημέρες που η ταχύτητα του ανέμου καταγράφηκε μεταξύ 2.9 και 3.1 m/s (α) Να δώσετε έναν πίνακα συχνοτήτων αυτών των ημερών ανά μήνα. (β) Αφού διατάξετε στον πίνακα δεδομένων αυτές τις παρατηρήσεις (ως προς την μεταβλητή date) να κατασκευάσετε μία νέα μεταβλητή <code>day0</code> η οποία να δείχνει το πλήθος των ημερών που έχουν περάσει από την 1-1-94. Ποια κατανομή φαίνεται να ακολουθεί αυτή η μεταβλητή (ιστόγραμμα, Q-Q plot, K-S τεστ); (γ) Κατασκευάστε μια νέα μεταβλητή <code>interval</code> η οποία να δείχνει το πλήθος των ημερών που μεσολάβησε από την αμέσως προηγούμενη καταγραφή τέτοιου ανέμου. Η νέα αυτή μεταβλητή καταγράφει τους ενδιάμεσους χρόνους (σε ημέρες) μεταξύ εμφανίσεων ημερών με ταχύτητα ανέμου μεταξύ 2.9 και 3.1 m/s. (υποδ. εκτελέστε create time series με function difference πάνω στην μεταβλητή <code>day0</code>). Ελέγξτε αν αυτή η μεταβλητή μπορεί να ακολουθεί την εκθετική ή την Γάμμα κατανομή (ιστόγραμμα, Q-Q plot, K-S τεστ).
(17) Να ελέγξετε αν η κατανομή των (μέσων ημερήσιων) θερμοκρασιών είναι ίδια τα έτη 1995 και 2005. Να ελέγξετε αν η κατανομή των ημερήσιων ταχυτήτων του ανέμου είναι ίδια τα έτη 1995 και 2005. (χρησιμ. τρία διαφορετικά απαραμετρικά τεστ).
(18) Να ελέγξετε αν οι διαδοχικές ημερήσιες θερμοκρασίες μπορούν να θεωρηθούν ανεξάρτητες τυχαίες μεταβλητές ή αντίθετα μια θερμοκρασία εξαρτάται από την θερμοκρασία των προηγούμενων και επόμενων ημερών (runs test με βάση το mean και το median). Υπάρχει μηδενική, θετική ή αρνητική εξάρτηση των θερμοκρασιών που παρατηρούνται διαδοχικές ημέρες και γιατί;

- όλοι οι έλεγχοι να γίνουν σε ε.σ. $\alpha = 5\%$.