Θέμα εργασίας (τελευταία ενημέρωση 9/6/2020)

Οδηγίες

*Η εργασία θα αξιολογηθεί με άριστα τη **μία μονάδα**. Θα πρέπει:*

- ***Να είναι ατομική. Η συνεργασία πάνω στα ερωτήματα είναι επιθυμητή αλλά η αντιγραφή απαγορεύεται.***
- *Στην αρχή (πρώτη σελίδα) να αναφέρεται το όνομα και το ΑΕΜ του φοιτητή / της φοιτήτριας.*
- *Να περιέχει τις απαντήσεις και λύσεις με σχόλια, όπως ζητούνται και με τη σειρά που ζητούνται.*
- *Τα γραφήματα και οι πίνακες αποτελεσμάτων από το SPSS θα πρέπει να παρουσιάζονται με τη σειρά που ζητούνται και σε κατάλληλα σημεία μέσα στο κείμενο της εργασίας.*
- ***Το κάθε γράφημα από το SPSS θα πρέπει να έχει στον τίτλο το ΑΕΜ του φοιτητή / της φοιτήτριας και ο κάθε πίνακας από το SPSS θα πρέπει να έχει στην επικεφαλίδα το ΑΕΜ του φοιτητή / της φοιτήτριας. Θα συμπεριλάβετε ΜΟΝΟ σχετικούς πίνακες και σχήματα στο SPSS που υποστηρίζουν τις απαντήσεις σας, διαφορετικά θα μετρήσει αρνητικά στο βαθμό σας.***
- *Η εργασία θα πρέπει να είναι γραμμένη στον υπολογιστή σε πρόγραμμα επεξεργασίας κειμένου, όπως Word. Αν είναι γραμμένη σε άλλο πρόγραμμα επεξεργασίας κειμένου, να σταλεί σε pdf.*

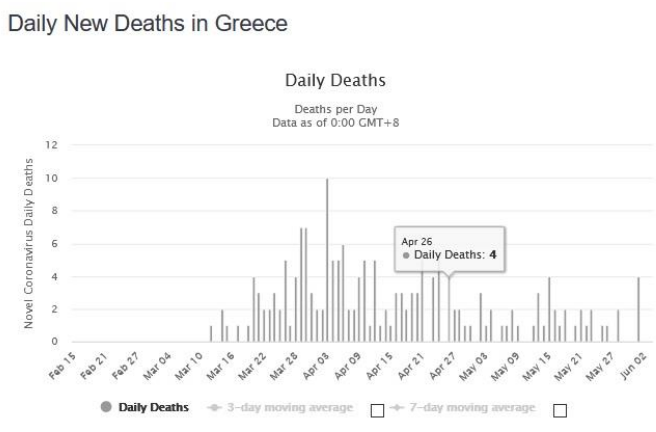
*Η εργασία θα πρέπει να παραδοθεί ηλεκτρονικά μέσω της ιστοσελίδας του μαθήματος στο elearning το αργότερο ως τις **22/7/2020**. Θα πρέπει να παραδοθεί ηλεκτρονικά ΜΟΝΟ ένα αρχείο που περιέχει ολόκληρη την εργασία. Το όνομα του αρχείου θα πρέπει να είναι Stat<ΑΕΜ>.docx όπου <ΑΕΜ> είναι ο Αριθμός Ειδικού Μητρώου του φοιτητή που παραδίδει την εργασία (χωρίς τις ανισότητες) και docx δηλώνει ότι είναι γραμμένο σε Word (άλλες επιτρεπτές μορφές είναι doc και pdf), π.χ. αν το ΑΕΜ είναι 9820 το αρχείο Word θα πρέπει να έχει όνομα Stat9820.docx, Stat9820.doc ή Stat9820.pdf. **Παρακαλώ να εξετάσετε αν το όνομα του αρχείου είναι σωστό πριν το υποβάλλετε. Αρχείο με όνομα που δεν είναι στην προβλεπόμενη μορφή μπορεί να αγνοηθεί και η εργασία να μην αξιολογηθεί!***

Περιγραφή προβλήματος και δεδομένα

Όλοι/ες γνωρίζουμε πως η πανδημία του κορονοϊού χτύπησε πολλές χώρες του πλανήτη και όλες τις χώρες της Ευρώπης. Στην εργασία θα χρησιμοποιήσετε στοιχεία κρουσμάτων και θανάτων κορονοϊού σε διάφορες χώρες της Ευρώπης, όπως έχουν καταγραφεί στον ιστοχώρο https://www.worldometers.info/coronavirus/, που θα χρησιμοποιήσετε για να εξάγετε δεδομένα. Δίνεται επίσης το αρχείο σε μορφή excel CoronavirusCountries.xlsx, που περιέχει σε διάταξη τις 37 χώρες τις Ευρώπης με πληθυσμό πάνω από 1 εκ κατοίκους που χτυπήθηκαν από τον κορονοϊό.

Θα χρησιμοποιήσετε δύο από τις 37 χώρες για την στατιστική ανάλυση σας σύμφωνα με τη διάταξη που δίνονται στο αρχείο και με βάση τα 4 τελευταία ψηφία του ΑΕΜ σας ως εξής. Η πρώτη χώρα, συμβολίζεται στο εξής ως Α, έχει αύξοντα αριθμό στο αρχείο CoronavirusCountries.xlsx που είναι το υπόλοιπο της διαίρεσης του διψήφιου αριθμού των δύο τελευταίων ψηφίων του ΑΕΜ σας με το 37. Η δεύτερη χώρα, συμβολίζεται στο εξής ως Β, έχει αντίστοιχα αύξοντα αριθμό στο αρχείο CoronavirusCountries.xlsx που είναι το υπόλοιπο της διαίρεσης του δεύτερου διψήφιου αριθμού από το ΑΕΜ σας με το 37, όπου αυτός ο διψήφιος αριθμός αποτελείται από το τέταρτο από το τέλος ψηφίο του ΑΕΜ (πρώτο ψηφίο του διψήφιου αριθμού) και το τρίτο από το τέλος ψηφίο του ΑΕΜ (δεύτερο ψηφίο του διψήφιου αριθμού). Για παράδειγμα το ΑΕΜ 9820 έχει ως Α τη Τουρκία με αύξοντα αριθμό 20, που είναι το υπόλοιπο της διαίρεσης του 20 με το 37 και ως Β τη Νορβηγία με αύξοντα αριθμό 24, που είναι το υπόλοιπο της διαίρεσης του 98 με το 37. Αν και τα δύο αύξοντες αριθμοί είναι ίδιο, τότε ο δεύτερος αύξων αριθμός θα αυξηθεί κατά 1 (και αν είναι 37 το επόμενο είναι 1). Για παράδειγμα αν το ΑΕΜ είναι 9824 τότε ο πρώτος αύξων αριθμός είναι 24 και ο δεύτερος είναι 25, δηλαδή το υπόλοιπο της διαίρεσης του 98 με το 37 είναι 24 και +1 είναι 25.

Στο αρχείο CoronavirusCountries.xlsx στο όνομα της κάθε χώρας υπάρχει υπερ-σύνδεσμος που σας μεταφέρει στην ιστοσελίδα του ιστοχώρου για αυτήν τη χώρα (μπορείτε φυσικά να μεταβείτε μέσα από την κεντρική σελίδα του ιστοχώρου). Εκεί θα εστιάσετε στα γραφήματα για τα ημερήσια νέα κρούσματα (Daily New Cases) και για τους ημερήσιους νέους θανάτους (Daily New Deaths). Στα παρακάτω σχήματα δίνεται η εικόνα των δύο γραφημάτων (στις 4/6/2020) για την Ελλάδα.



Σε κάθε ένα από τα δύο παραπάνω σχήματα φαίνεται σε ένθετο κάποια πληροφορία και κυρίως η τιμή του αντίστοιχου δείκτη που δηλώνεται με το ύψος της αντίστοιχης μπάρας (για τις 15 Μαΐου στο πρώτο και για τις 25 Απριλίου στο δεύτερο). Μπορείτε εύκολα να δείτε την κάθε τιμή τοποθετώντας το δείκτη του υπολογιστή (cursor) στην κορυφή της αντίστοιχης μπάρας. Με αυτόν τον τρόπου μπορείτε να συλλέξετε τα στοιχεία που θα χρειαστείτε στα ζητήματα της εργασίας που δίνονται παρακάτω.

Συγκεκριμένα, θα χρειαστεί πρώτα να ορίσετε την κορύφωση της “καμπύλης” για τα ημερήσια νέα κρούσματα και για τους ημερήσιους νέους θανάτους. Η ημέρα της κορύφωσης θα οριστεί ελεύθερα αλλά αιτιολογημένα από εσάς Ένας απλός τρόπος είναι να ορίσετε την κορύφωση από τη μέγιστη τιμή. Εδώ θέλει προσοχή, καθώς κάποια μεγάλη τιμή μπορεί να μην αντιστοιχεί σε κορύφωση, αλλά αποτελεί περισσότερο μια απόμακρη τιμή που μπορεί να προέκυψε από τον τρόπο καταγραφής ή κάποιο μεμονωμένο φαινόμενο. Για παράδειγμα, τα 156 νέα κρούσματα στις 21/4 στην Ελλάδα (αριστερό σχήμα) μπορούν να χαρακτηριστούν ως ειδική καταγραφή, όπου 150 κρούσματα ήταν σε δομή φιλοξενίας στο Κρανίδι. Σε μια τέτοια περίπτωση, μπορείτε να διορθώσετε τη τιμή σε 6 ή με προσέγγιση κάποιας άλλης τιμής ή να την αγνοήσετε αν παρουσιάζεται στο δείγμα σας (δες παρακάτω για το σχηματισμό των δειγμάτων). Υπάρχει υπερ-σύνδεσμος στην πηγή πληροφορίας για την κάθε ημερήσια μέτρηση στο τέλος της ιστοσελίδας για κάθε χώρα, στην παράγραφο Latest News. Άρα ως κορύφωση του φαινομένου στην Ελλάδα θα ήταν πιο σωστό να επιλέξετε την 2/4 με 129 νέα κρούσματα. Η αντίστοιχη κορύφωση για τους ημερήσιους νέους θανάτους είναι 3/4 με 10 θανάτους.

Αφού έχετε ορίσει τη μέγιστη τιμή για τα ημερήσια νέα κρούσματα και τους ημερήσιους νέους θανάτους, θα ορίσετε τα δείγματα σας για την ανάλυση στα ζητήματα του μέρους Α και του μέρους Β.

Για το μέρος Α, θα ορίσετε τις 20 τιμές για τα ημερήσια νέα κρούσματα από τη μέρα μετά την κορύφωση ως και 20 μέρες μετά την κορύφωση. Για την περίπτωση της Ελλάδας αυτό σημαίνει ότι το σύνολο των 20 τιμών είναι από τις 3/4 ως και τις 22/4, όπου θα ήταν καλό να διορθώσετε την ακραία τιμή της 21/4 ή να την αγνοήσετε και να συμπεριλάβετε και την τιμή της 23/4. Αντίστοιχα θα ορίσετε τις 20 συνεχόμενες τιμές για τους ημερήσιους νέους θανάτους από τη μέρα μετά την κορύφωση ως και 20 μέρες μετά την κορύφωση (αν αγνοήσετε την τιμή τις 21/4 για τα ημερήσια νέα κρούσματα που είναι η $19^η$ θα αγνοήσετε επίσης τη $19^η$ τιμή, δηλαδή της 22/4 για τους ημερήσιους νέους θανάτους και θα συμπεριλάβετε και την τιμή της 24/4). Για την περίπτωση της Ελλάδας αυτό σημαίνει ότι το σύνολο των 20 τιμών είναι από τις 4/4 ως και τις 23/4 (ή της 24/4 αν αγνοηθεί η τιμής της 22/4). Στο συγκεκριμένο παράδειγμα τυχαίνει το χρονικό παράθυρο των 20 τιμών του συνόλου των ημερήσιων νέων κρουσμάτων να είναι πολύ κοντά στο χρονικό παράθυρο των 20 τιμών του συνόλου των ημερήσιων νέων θανάτων (διαφορά μιας μέρας), αλλά για άλλες χώρες μπορεί τα δύο αυτά χρονικά παράθυρα να διαφέρουν αρκετά. Στην περίπτωση που δεν υπάρχουν 20 τιμές μετά την κορύφωση (στους ημερήσιους νέους θανάτους που έπεται της κορύφωσης των ημερήσιων νέων κρουσμάτων) θα πάρετε τον αριθμό των διαθέσιμων τιμών, που σε κάθε περίπτωση πρέπει να είναι πάνω από 8.

Στη συνέχεια θα διαιρέσετε την κάθε τιμή ημερήσιων νέων θανάτων με την αντίστοιχη τιμή ημερήσιων νέων κρουσμάτων (η πρώτη τιμή του δεύτερου συνόλου ως προς την πρώτη τιμή του πρώτου συνόλου και όμοια για τις δεύτερες τιμές, τρίτες τιμές κτλ). Θα μετατρέψετε τις τιμές που προκύπτουν σε ποσοστά, πολλαπλασιάζοντας με εκατό. Αν κάποια τιμή ημερήσιων νέων θανάτων ή ημερήσιων νέων κρουσμάτων είναι 0, η τιμή του ποσοστού θα είναι επίσης 0. Αυτές οι τιμές εκφράζουν το ημερήσιο ποσοστό θνητότητας στη χώρα την περίοδο μετά την κορύφωση ως προς τα ημερήσια νέα κρούσματα, δηλαδή προσεγγιστικά μας λέει για τα νέα κρούσματα που έχουμε σε μια μέρα στη χώρα τι ποσοστό από αυτά θα πεθάνει. Το σύνολο αυτών των τιμών θα αποτελέσει το δείγμα για κάθε μια από τις δύο χώρες Α και Β για τα ζητήματα του μέρους Α.

Για το μέρος Β, θα ορίσετε *n* τιμές για τα ημερήσια νέα κρούσματα (δική σας επιλογή και με βάση τα δεδομένα σας, μια καλή επιλογή του *n* είναι μεταξύ 5 και 10, αλλά μπορείτε να επιλέξετε και μεγαλύτερο *n*) από τη μέρα πριν τη κορύφωση πηγαίνοντας προς τα πίσω *n*-1 μέρες, αντίστοιχα, ώστε συνολικά να έχετε *n* τιμές. Για την κάθε τιμή, δε χρειάζεται η πληροφορία της ακριβούς ημερομηνίας παρά μόνο ο αύξων αριθμός της ημέρας, που θα είναι από 1 για την ημέρα της χρονικά πρώτης τιμής ημερήσιων νέων κρουσμάτων ως τον αύξοντα αριθμό *n* της ημέρας πριν την κορύφωση. Το σύνολο αυτών των τιμών ημερήσιων νέων κρουσμάτων και των αντίστοιχων ημερών θα

αποτελέσει το δείγμα ζευγαρωτών παρατηρήσεων για κάθε μια από τις δύο χώρες Α και Β για τα ζητήματα του μέρους Β, δηλαδή για κάθε χώρα το δείγμα θα έχει *n* ζευγάρια τιμών όπου το πρώτο στοιχείο κάθε ζευγαριού είναι ο αύξων αριθμός της ημέρας και το δεύτερο η αντίστοιχη τιμή ημερήσιων νέων κρουσμάτων. Στην επιλογή των τιμών του δείγματος ζευγαρωτών παρατηρήσεων για τη Μελέτη Β, μπορείτε, όπως και για το δείγμα στη Μελέτη Α που αναφέρεται στις τιμές ημερήσιων νέων κρουσμάτων και θανάτων μετά την κορύφωση, να διορθώσετε ή απορρίψετε τιμές υψηλές (ή και αρνητικές) τιμές που αιτιολογημένα μπορούν αν χαρακτηριστούν ως ειδικές καταγραφές που δεν αποτυπώνουν την εξέλιξη του φαινόμενου (δες πηγές των καταγραφών αυτών)

Ζητήματα της εργασίας

Πριν να περάσετε και απαντήσετε τα ζητήματα για τη μελέτη Α και Β θα πρέπει να παρουσιάσετε συνοπτικά πως δημιουργήσατε τα δείγματα που χρησιμοποιούνται στις δύο μελέτες, π.χ. με κάποιο γράφημα, περιγραφή της επιλογής της ημέρας κορύφωσης για τα ημερήσια νέα κρούσματα και ημερήσιους νέους θανάτους, καθώς και προβλήματα που συναντήσατε στην επιλογή των δεδομένων

Μελέτη Α

Για τη μελέτη Α, θα θεωρήσετε το δείγμα για τη χώρα Α των 20 τιμών για το ημερήσιο ποσοστό θνητότητας ως προς τα ημερήσια νέα κρούσματα στη χώρα την περίοδο μετά την κορύφωση, που θα ονομάζεται δείγμα Α και το αντίστοιχο δείγμα για τη χώρα Β που θα ονομάζεται δείγμα Β. Στο εξής η τυχαία μεταβλητή που μετρήθηκε στα δείγματα Α και Β θα ονομάζεται απλά *ημερήσιο ποσοστό θνητότητας*.

Θα μελετήσετε την κατανομή του ημερήσιου ποσοστού θνητότητας στη χώρα την περίοδο μετά την κορύφωση, και κυρίως τη μέση τιμή του, για τις δύο χώρες που σας αντιστοιχούν.

1. Σχολιάστε την *κατανομή* του ημερήσιου ποσοστού θνητότητας την περίοδο μετά την κορύφωση στις δύο χώρες Α και Β με βάση τα αντίστοιχα δείγματα των 20 μετρήσεων. Θα πρέπει να συμπεριλάβετε τα παρακάτω, χρησιμοποιώντας το SPSS:
   - Έναν πίνακα με συνοπτικά μέτρα κεντρικής τάσης (μέση τιμή και διάμεσο), μεταβλητότητας (διασπορά, τυπική απόκλιση, εύρος δεδομένων, πρώτο και τρίτο τεταρτομόριο). Ο πίνακας θα πρέπει να περιέχει τα μέτρα και για τα δύο δείγματα.
   - Ιστόγραμμα για το κάθε δείγμα.
   - Ένα σχήμα που να περιέχει τα δύο θηκογράμματα, ένα για κάθε δείγμα.

   Να γίνουν αναλυτικά σχόλια για τον πίνακα και τα σχήματα συγκρίνοντας τα αποτελέσματα για τα δύο δείγματα. Είναι κανονική η κατανομή του ημερήσιου ποσοστού θνητότητας την περίοδο μετά την κορύφωση σε κάθε μια από τις δύο χώρες? Υπάρχουν διαφορές στην κατανομή του ημερήσιου ποσοστού θνητότητας την περίοδο μετά την κορύφωση στις δύο χώρες Α και Β με βάση τα δύο δείγματα;
2. Για κάθε ένα από τα δείγματα Α και Β των 20 παρατηρήσεων, ελέγξετε με κατάλληλο 95% διάστημα εμπιστοσύνης αν η μέση τιμή του ημερήσιου ποσοστού θνητότητας μπορεί να είναι 10%. Στην απάντηση σας συμπεριλάβετε κατάλληλο πίνακα αποτελεσμάτων στο SPSS και για τα δύο δείγματα και σχολιάστε τα αποτελέσματα. Συγκρίνετε επίσης την ακρίβεια της εκτίμησης της μέσης τιμής του ημερήσιου ποσοστού θνητότητας στις δύο χώρες. Σημειώστε τυχόν επιφύλαξη που έχετε για την εγκυρότητα των αποτελεσμάτων.
3. Με βάση τα δείγματα Α και Β, ελέγξετε αν η μέση τιμή του ημερήσιου ποσοστού θνητότητας μπορεί να είναι ίδια στις δύο χώρες χρησιμοποιώντας κατάλληλο 95% διάστημα εμπιστοσύνης. Στην απάντηση σας συμπεριλάβετε κατάλληλο πίνακα αποτελεσμάτων στο SPSS και σχολιάστε τα αποτελέσματα. Σημειώστε τυχόν επιφύλαξη που έχετε για την εγκυρότητα των αποτελεσμάτων.
4. Σχολιάστε αν φαίνεται να υπάρχει χαρακτηριστικό ποσοστό θνητότητας την περίοδο μετά την κορύφωση που μπορεί να εκτιμηθεί και αν αυτό είναι διαφορετικό ανά χώρα με βάση τα αποτελέσματα της στατιστικής ανάλυσης σας.

Μελέτη Β:

Θέλουμε να εξετάσουμε αν κατά την εξάπλωση του κορονοϊού στη χώρα, δηλαδή τη χρονική περίοδο πριν την κορύφωση των ημερήσιων νέων κρουσμάτων, φαίνεται ο αριθμός ημερήσιων νέων κρουσμάτων να ακολουθεί γραμμική εξάρτηση ως προς το χρόνο, δηλαδή αν ο αριθμός των ημερήσιων νέων κρουσμάτων αυξάνεται γραμμικά. Για αυτό θεωρήσετε τα δείγματα *n* ζευγαρωτών παρατηρήσεων για κάθε μια από τις δύο χώρες, όπου το πρώτο στοιχείο κάθε ζευγαριού είναι ο αύξων αριθμός της ημέρας και το δεύτερο η αντίστοιχη τιμή ημερήσιων νέων κρουσμάτων στη χώρα. Η ανάλυση θα γίνει και για τα δύο δείγματα Α και Β των χωρών που σας αντιστοιχούν.

5. Χρησιμοποιώντας το SPSS, κάνετε κατάλληλο διάγραμμα διασποράς και υπολογίστε τον αντίστοιχο συντελεστή συσχέτισης για τον αύξοντα αριθμό της ημέρας και τον αριθμό ημερήσιων νέων κρουσμάτων για κάθε μια από τις δύο χώρες ξεχωριστά χρησιμοποιώντας το αντίστοιχο δείγμα. Σχολιάστε αν εξαρτάται ο αριθμός ημερήσιων νέων κρουσμάτων από την ημέρα για κάθε μια από τις δύο χώρες.

6. Για κάθε ένα από τα δύο ζευγάρια δειγμάτων (για τη χώρα Α και Β) και χρησιμοποιώντας το SPSS, εκτιμήστε το μοντέλο γραμμικής παλινδρόμησης με τη μέθοδο ελαχίστων τετραγώνων. Σχολιάστε το κάθε ένα από τα δύο μοντέλα που εκτιμήσατε με έμφαση στην καταλληλότητα του μοντέλου για προβλέψεις. Στην απάντηση σας θα συμπεριλάβετε τους πίνακες του SPSS που δίνουν την εκτίμηση των παραμέτρων του μοντέλου (σταθερός όρος, κλίση και τυπική απόκλιση σφαλμάτων παλινδρόμησης). Φαίνεται τα δύο μοντέλα παλινδρόμησης που προσεγγίζουν τη γραμμική αύξηση των ημερήσιων νέων κρουσμάτων να συμφωνούν; Αιτιολογείστε την απάντηση σας με αναφορά στις παραμέτρους του κάθε μοντέλου.

7. Προβλέψετε με το κάθε ένα από τα δύο μοντέλα που εκτιμήσατε στο ερώτημα 6 (για το δείγμα Α και Β) τον αριθμό ημερήσιων νέων κρουσμάτων την επόμενη από την τελευταία μέρα που έχετε στο δείγμα, η οποία είναι η ημέρα της κορύφωσης, με βάση τη συλλογή των παρατηρήσεων στο δείγμα. Σημειώνεται ότι σε προβλήματα παλινδρόμησης δεν κάνουμε πρόβλεψη έξω από το πεδίο τιμών της ανεξάρτητης μεταβλητής (εδώ του χρόνου) αλλά το επιχειρούμε στην περίπτωση της χρονικής πρόβλεψης, όπως εδώ που προβλέπουμε για την επόμενη ημέρα. Προσεγγιστικά ποια είναι η ακρίβεια της πρόβλεψης για τη χώρα Α και τη χώρα Β;

8. Μπορείτε να προτείνετε κάποιο άλλο μοντέλο πρόβλεψης που πιστεύετε πως θα απέδιδε καλύτερα στην πρόβλεψη της αύξησης των ημερησίων νέων κρουσμάτων (ως την κορύφωση τους)? Είστε ελεύθεροι/ες (χωρίς να είναι απαραίτητο) να προχωρήσετε και στην υλοποίηση του μοντέλου, δηλαδή την προσαρμογή του στα δεδομένα και πρόβλεψη της επόμενης τιμής για την ημέρα κορύφωσης καθώς και να το συγκρίνετε με το γραμμικό μοντέλο παλινδρόμησης.