

ΙΕΚ ΞΑΝΘΗΣ

Μάθημα : Στατιστική Ι

Υποενότητα : Κωδικοποίηση

Επαμεινώνδας Διαμαντόπουλος

Ιστοσελίδα : <http://users.sch.gr/epdiaman/>

Email : epdiamantopoulos@yahoo.gr

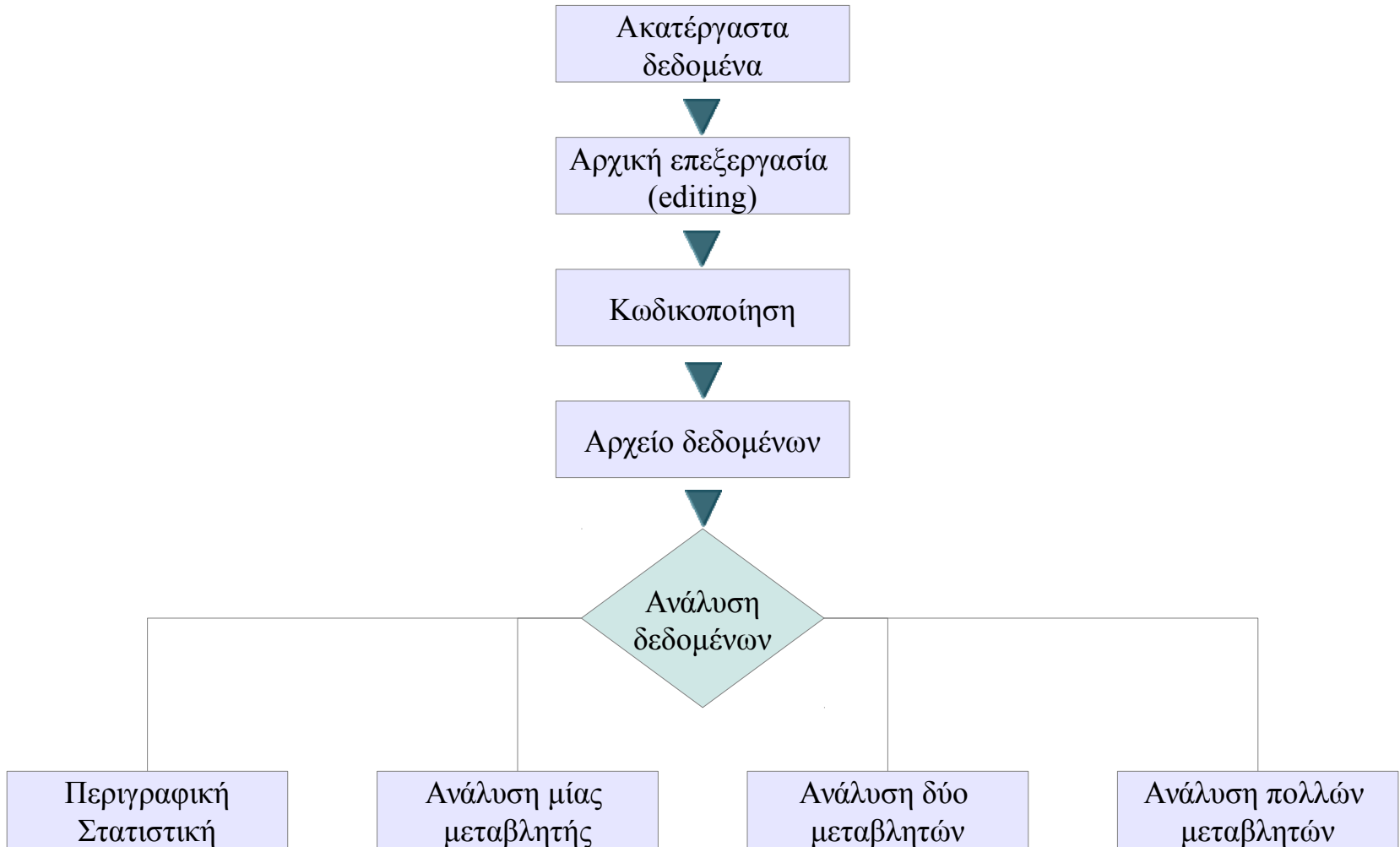
Στόχοι της υποενότητας

Μετά τη μελέτη αυτής της υποενότητας θα...

1. Είσαι σε θέση να επεξεργαστείς τα δεδομένα χωρίς να διακινδυνεύσεις την αξιοπιστία των δεδομένων.
2. Γνωρίζεις τα βασικά της κωδικοποίησης ποιοτικών μεταβλητών.
3. Γνωρίζεις πότε και γιατί χρησιμοποιούμε ψευδομεταβλητές (dummy variables)
4. Έχεις αντιληφθεί τη σημασία της τεχνολογίας στη διαδικασία κωδικοποίησης
5. Καταλαβαίνεις τον τρόπο με τον οποίο αναπαριστώνται τα δεδομένα σε κάποιο αρχείο υπολογιστή.

Χρήσιμες έννοιες

- Ακατέργαστα δεδομένα (Raw data)
 - Οι αποκρίσεις των ερωτώμενων, χωρίς οποιαδήποτε μεταβολή από τον ερευνητή.
- Σφάλμα λανθασμένης καταγραφής.
 - Σφάλμα μη ορθής καταγραφής των αποκρίσεων από τον ερευνητή. (δεν ευθύνεται ο ερωτώμενος)
- Ακεραιότητα δεδομένων (Data Integrity)
 - Έννοια που σημαίνει πως τα δεδομένα περιέχουν τις πληροφορίες που απαιτούνται για την αντιμετώπιση των ερευνητικών υποθέσεων.



Πρωταρχική Επεξεργασία

- Πρωταρχική επεξεργασία (Editing)
 - Έλεγχος για πληρότητα, συνέπεια και ορθότητα των δεδομένων
 - Προετοιμασία των καταγραφών για πέρασμά τους σε ψηφιακό αρχείο.
- Έλεγχος παράλληλα με τη συλλογή (Field Editing)
 - Έλεγχος για παραλείψεις, άκυρες καταγραφές, έλεγχος της αναγνωσιμότητας, διευκρίνηση των αποκρίσεων που δεν μπορούν να γίνουν κατανοητές.
- Έλεγχος μετά τη συλλογή (In-House Editing)
 - Ένας περισσότερο αναλυτικός έλεγχος για το αν τα ερωτηματολόγια που συλλέχθηκαν έχουν λογικά σφάλματα.

Γιατί να γίνει η πρωταρχική επεξεργασία;

Διόρθωση στοιχείων με προφανή σφάλματα

Έλεγχος για πληρότητα

Επεξεργασία “ύποπτων” ερωτηματολογίων

Διευκόλυνση της Διαδικασίας κωδικοποίησης



Έλεγχος συνέπειας

- Έλεγχος συνέπειας
 - Έλεγχος πως το δείγμα ανήκει στον πληθυσμό που θέλαμε να περιγράψουμε.
 - Έλεγχος πως το δείγμα καλύπτει το πλαίσιο δειγματοληψίας που είχαμε αρχικά σχεδιάσει.
- Ενέργειες σε περιπτώσεις προφανούς σφάλματος
 - Διόρθωση των αποκρίσεων μόνο όταν υπάρχουν σοβαρές ενδείξεις για τέτοια ενέργεια.
- Έλεγχος συνέπειας με Η/Υ.
 - Ένας Η/Υ μπορεί εύκολα να εντοπίσει παράλογους συνδυασμούς αποκρίσεων σε επιλεγμένες ερωτήσεις,

Έλεγχος πληρότητας (1)

- Μη απόκριση (Item Nonresponse)
 - Όρος που περιγράφει τις αποκρίσεις που έμειναν μη συμπληρωμένες με αποτέλεσμα το ερωτηματολόγιο να μην είναι πλήρως συμπληρωμένο.
 - Τιμή αντικατάστασης (plug value)
 - ❖ Τιμή που αντικαθιστά τις απύσες τιμές.
 - ❖ Ο τρόπος υπολογισμού πρέπει να είναι προκαθορισμένος.
 - Συμπλήρωση (Impute)
 - ❖ Η συμπλήρωση μίας απύσης τιμής με κάποιο στατιστικό κανόνα/διαδικασία ώστε να μεγιστοποιηθεί το πλήθος των περιπτώσεων από τις οποίες θα προκύψει το αποτέλεσμα (π.χ. συμπλήρωση με τη μέση τιμή ή τη διάμεσο των υπολοίπων κλπ)

Έλεγχος πληρότητας (2)

- Διαγραφή κενής καταχώρησης (απουσίας τιμής)
 - Διαγραφή όλων των αποκρίσεων του ερωτώμενου (η γραμμή στο αρχείο δεδομένων) (list wise)
 - ❖ Όλες οι αποκρίσεις των ερωτώμενου απομακρύνονται από τα δεδομένα και δε λαμβάνονται υπόψη για οποιαδήποτε υπολογισμό.
- Διαγραφή των αποκρίσεων που δεν συμμετέχουν στη στατιστική ανάλυση (pair wise)
 - ❖ Παραμένουν στοιχεία του ερωτώμενου που μπορούν να χρησιμοποιηθούν σε άλλα σημεία όπως π.χ. το φύλο του, η οικογενειακή κατάσταση κλπ.

Έλεγχος πληρότητας (3)

One-Way ANOVA: Options

Statistics

- Descriptive
- Fixed and random effects
- Homogeneity of variance test
- Brown-Forsythe
- Welch

Means plot

Missing Values

- Exclude cases analysis by analysis
- Exclude cases listwise

Continue
Cancel
Help

Factor Analysis: Options

Missing Values

- Exclude cases listwise
- Exclude cases pairwise
- Replace with mean

Coefficient Display Format

- Sorted by size
- Suppress absolute values less than:

Continue
Cancel
Help

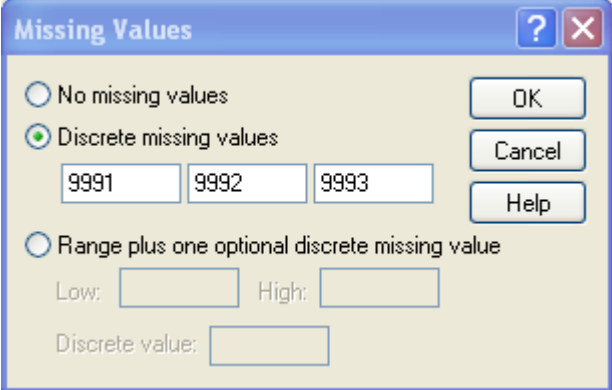
Λεπτομέρειες της διαδικασίας κωδικοποίησης (1)

- Διαχείριση των “ΔΞ/ΔΑ” αποκρίσεων
 - Ο ερωτώμενος δεν έχει άποψη.
 - Ο ερωτώμενος δεν επιθυμεί να απαντήσει.
 - Ο ερωτώμενος δεν καταλαβαίνει την ερώτηση.

Λεπτομέρειες της διαδικασίας κωδικοποίησης (2)

Άποψη από την εξυπηρέτηση

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Πολύ καλή	3	16,7	23,1	23,1
	Καλή	4	22,2	30,8	53,8
	Ούτε καλή ούτε κακή	2	11,1	15,4	69,2
	Κακή	2	11,1	15,4	84,6
	Πολύ κακή	2	11,1	15,4	100,0
	Total		13	72,2	100,0
Missing	Δεν έχει άποψη	2	11,1		
	Δεν επιθυμεί να απαντήσει	2	11,1		
	Δεν καταλαβαίνει την ερώτηση	1	5,6		
	Total	5	27,8		
Total		18	100,0		



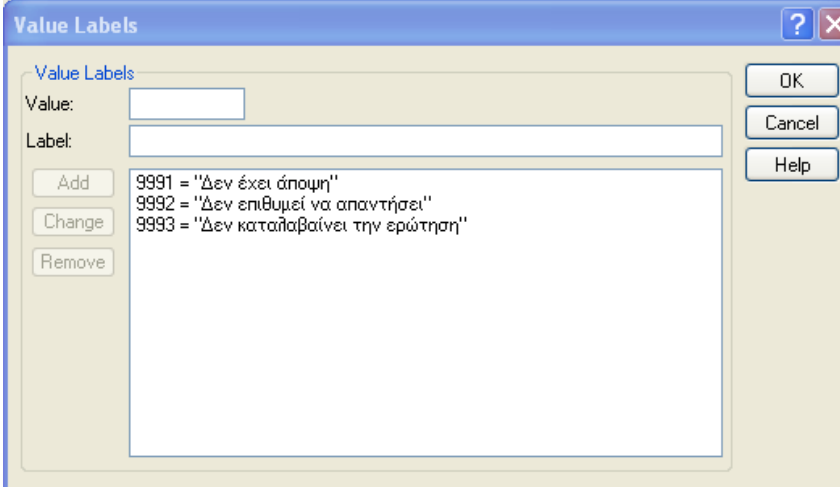
Missing Values

No missing values
 Discrete missing values

Range plus one optional discrete missing value

Low: High:

Discrete value:



Value Labels

Value Labels

Value:

Label:

9991 = "Δεν έχει άποψη"
 9992 = "Δεν επιθυμεί να απαντήσει"
 9993 = "Δεν καταλαβαίνει την ερώτηση"

Αρχική Διαχείριση

- Παγίδες κατά την αρχική επεξεργασία
 - Υποκειμενική διαχείριση των ελλείψεων των αρχικών δεδομένων
 - ❖ Οι διαχειριστές των πρωταρχικών δεδομένων πρέπει να είναι έμπειροι, προσεκτικοί και αντικειμενικοί.
 - Ο υπεύθυνος της αρχικής διαχείρισης πρέπει να έχει στα χέρια του ξεκάθαρες οδηγίες για κάθε μία περίπτωση που θα βρει μπροστά του.
- Επεξεργασία των δοκιμαστικών καταγραφών
 - Η προσεκτική επεξεργασία των δεδομένων που θα συλλεχθούν κατά τη δοκιμαστική εφαρμογή του ερωτηματολογίου μπορούν να αναβαθμίσουν την ποιότητα του κατά την κύρια εφαρμογή, εντοπίζοντας σημεία που χρήζουν βελτίωσης στη διάταξη του ερωτηματολογίου, στον τρόπο έκφρασης των ερωτήσεων ή στις οδηγίες προς τον ερωτώμενο.
- Έλεγχος σφαλμάτων
 - Έλεγχος για την ορθότητα της κωδικοποίησης.

Κωδικοποίηση ποιοτικών μεταβλητών

- Κωδικοποίηση (coding)
 - Αντιστοίχιση διαφορετικών αριθμητικών τιμών σε διαφορετικές τιμές της μεταβλητής.
- Κωδικοί (codes)
 - Κανόνες αντιστοίχισης, ταξινόμησης και καταγραφής των τιμών κατά τη διαδικασία κωδικοποίησης
 - Οι αριθμοί που αντιστοιχούν σε κάθε διαφορετική τιμή των δεδομένων.
- Χρήση ψευδομεταβλητών (dummy coding)
 - Αντιστοίχιση του “1” και του “0” σε διαφορετικές τιμές μίας ποιοτικής μεταβλητής όπως “Ανδρας” και “Γυναίκα”
 - Αν η ποιοτική μεταβλητή έχει k διαφορετικές τιμές τότε απαιτούνται $k - 1$ ψευδομεταβλητές
 - Χρήσιμες στις μεθόδους παλινδρόμησης

Καταχώρηση δεδομένων με χρήση νέων τεχνολογιών

- Καταχώρηση δεδομένων
 - Μεταφορά των δεδομένων από την παρατήρηση στον υπολογιστή (π.χ. με ασύρματους ή wi-fi καταγραφείς)
- Συστήματα οπτικής αναγνώρισης
 - Συσκευή εισόδου με δυνατότητα αναγνώρισης συμπληρωμένων αποκρίσεων σε προκαθορισμένες θέσεις.

Δημιουργία ψηφιακού αρχείου δεδομένων (1)

- Πεδίο (field)
 - Ονομασία που χαρακτηρίζει μία μεταβλητή.
- Πεδίο κειμένου (string characters)
 - Μεταβλητή που φιλοξενεί αλφαριθμητικές τιμές (δηλαδή όχι αριθμούς)
- Φάκελος (record)
 - Συλλογή από σχετιζόμενα πεδία που φιλοξενούν όλες τα δεδομένα από μία δειγματική μονάδα.

Δημιουργία ψηφιακού αρχείου δεδομένων (2)

- Αρχείο δεδομένων (data file)
 - Αρχείο στο οποίο τα στοιχεία της έρευνας έχουν τοποθετηθεί σε γραμμές και στήλες έτσι ώστε οι αποκρίσεις κάθε ερωτώμενου να βρίσκονται σε κάθε μία σειρά ενώ οι διαφορετικές μεταβλητές βρίσκονται σε διαφορετικές στήλες.
- Ετικέτες των τιμών (value labels)
 - Λέξεις ή φράσεις που επεξηγούν το νόημα κάθε αριθμητικής τιμής που έχει αποδοθεί σε κάθε διαφορετική αρχική τιμής της μεταβλητής.

Δημιουργία κωδικών (code construction)

- Δύο βασικοί κανόνες για την κωδικοποίηση :
 1. Οι κωδικοί πρέπει να καλύπτουν όλες τις πιθανές περιπτώσεις.
 2. Πρέπει να είναι ανεξάρτητοι μεταξύ τους. Κάθε μία διαφορετική κατηγορία πρέπει να αντιστοιχείται σε μία και μόνο μία διαφορετική τιμή.
- Δοκιμαστική καταγραφή (test tabulation)
 - Πρωταρχική ανίχνευση όλων των διαφορετικών αποκρίσεων από το δείγμα και έπειτα αντιστοίχιση με συγκεκριμένους αριθμούς.

Σύνοψη της διαδικασίας κωδικοποίησης

- Η διαδικασία κωδικοποίησης δεν πρέπει να είναι περίπλοκη/δυσνόητη.
 - Στόχος είναι η μορφοποίηση των δεδομένων για την εύκολη διαχείρισή τους από το στατιστικό πρόγραμμα.
 - Οι κατηγορίες πρέπει να είναι ξεκάθαρες ώστε να αποφευχθεί το ενδεχόμενο εσφαλμένης αντιστοίχισης
- Φύλλο κωδικοποίησης
 - Σύντομη αναφορά όλων των στοιχείων που χρησιμοποιήθηκαν στην κωδικοποίηση (ονομασία και περιγραφή της μεταβλητής, αντίστοιχοι κωδικοί)

Πίνακας 2

Δραστηριότητα κωδικοποίησης

ΕΡΩΤΗΜΑΤΟΛΟΓΙΟ

1. **Ηλικία επισκεπτών:** Σημειώστε από το 1 (μεγαλύτερο ποσοστό) έως το 6 (μικρότερο ποσοστό) τις ηλικίες των επισκεπτών:

- Έως 18 ετών
- Από 18 έως 25
- Από 25 έως 35
- Από 35 έως 50
- Από 50 έως 65
- Από 65 και άνω

2. **Για ποιο/οις λόγο/ους κατά τη γνώμη σας οι τουρίστες επέλεξαν να επισκεφθούν την περιοχή σας.** (Σημειώστε με X όποιες απαντήσεις θέλετε).

- Τοποθεσία και φυσικό περιβάλλον της περιοχής
- Επάρρο ή από διαφημιστική εικόνα της περιοχής
- Επίσκεψη σε συγγενείς/ φίλους
- Επίσκεψη σε γειτονικούς νομούς
- Πολιτιστικούς λόγους (αξιοθέατα, γιορτές, φεστιβάλ)
- Καταναλωτικούς λόγους (αγορές ντόπιων προϊόντων)
- Διασκέδαση – Αναψυχή
- Άλλοι λόγοι (αναφέρετε)

3. **Με ποιόν τρόπο κατά τη γνώμη σας, οι τουρίστες πληροφορούνται για την περιοχή σας** (Σημειώστε έως τρεις απαντήσεις με σειρά προτεραιότητας 1,2,3)

- Φίλου/ Συγγενείς
- Διαφημιστικά φυλλάδια και έντυπα
- Διαφήμιση στα ΜΜΕ
- Άρθρα σε περιοδικά, εφημερίδες
- Διαδίκτυο
- Τουριστικές εκθέσεις
- Άλλη πηγή