

Homework 1 (12-Νοεμβρίου 2008),

Μάθημα: **Στατιστικά Προγράμματα**, τμ. Στατιστικής και Ασφαλιστικής Επιστήμης

Στα πλαίσια του προγράμματος SHARE (*Survey of Health, Ageing and Retirement in Europe*, www.share-project.org) το έτος 2004, καταγράφηκε ένα σύνολο μεταβλητών σε περισσότερα από 30000 άτομα ηλικίας 50 ετών και άνω, από αρκετές χώρες της Ευρώπης. Θα χρησιμοποιήσουμε ένα μικρό υποσύνολο των μεταβλητών αυτών, οι οποίες βρίσκονται σε ένα αρχείο του SPSS στην ιστοσελίδα του μαθήματος (αρχείο **Share1.sav**).

Κάθε φοιτητής – τρία θα ξεκινήσει με το ίδιο σύνολο δεδομένων. Στη συνέχεια όμως (από την ερώτηση E2 και μετά) θα χρησιμοποιήσει τα δεδομένα που αφορούν μόνο 4 χώρες.

Συγκεκριμένα, οι φοιτητές-τριες με τελευταίο ψηφίο του ΑΜ τους

το 0 θα επιλέξουν τις χώρες 1,2,8,9	το 1 θα επιλέξουν τις χώρες 5,4,8,9
το 2 θα επιλέξουν τις χώρες 6,11,8,9	το 3 θα επιλέξουν τις χώρες 1,7,10,9
το 4 θα επιλέξουν τις χώρες 5,2,10,9	το 5 θα επιλέξουν τις χώρες 6,4,10,9
το 6 θα επιλέξουν τις χώρες 1,11,3,9	το 7 θα επιλέξουν τις χώρες 5,7,3,9
το 8 θα επιλέξουν τις χώρες 6,2,3,9	το 9 θα επιλέξουν τις χώρες 5,4,8,9

Όπου : 1. Αυστρία, 2. Γερμανία, 3. Σουηδία, 4. Ολλανδία, 5. Ισπανία, 6. Ιταλία, 7. Γαλλία, 8. Δανία, 9. Ελλάδα, 10. Ελβετία, 11. Βέλγιο

E1. (α) Να κατασκευάσετε έναν πίνακα συχνοτήτων που να δείχνει πόσα άτομα του δείγματος προέρχονται από κάθε μια από τις 11 χώρες. (β) Κατασκευάστε και το αντίστοιχο ραβδόγραμμα (να φαίνονται και τα αντίστοιχα ποσοστά) (γ) Δημιουργήστε με κατάλληλο τρόπο μια νέα μεταβλητή (age) που να δείχνει την ηλικία του κάθε ατόμου την ημέρα της έρευνας (έγινε το 2004). Κατασκευάστε μια πληθυσμιακή πυραμίδα (population pyramid) που να δείχνει τις συχνότητες των ηλικιών των ατόμων του δείγματος ανά φύλο. (δ) Ερμηνεύστε σύντομα το γράφημα.

E2. (α) Από τα συνολικά δεδομένα, επιλέξτε μόνο τα άτομα που προέρχονται από τις 4 χώρες που αντιστοιχούν στο τελευταίο ψηφίο του ΑΜ σας (βλ. παραπάνω). Τα υπόλοιπα άτομα διαγράψτε τα από το αρχείο των δεδομένων σας. **Σε όλα τα επόμενα ερωτήματα συνεχίσουμε με τα δεδομένα που αφορούν αυτές τις 4 χώρες.** (β) Κατασκευάστε τις πληθυσμιακές πυραμίδες (ηλικίες ανά φύλο) αυτή τη φορά για κάθε μια από τις 4 χώρες. Ερμηνεύστε σύντομα τα γραφήματα. (υποδ. εάν θέλετε να επαναλάβετε την ίδια ανάλυση πάνω σε υποσύνολα του δείγματος, χρησιμοποιείτε την επιλογή *panel by* (όπου υπάρχει) ή την επιλογή *split file*). (γ) Κατασκευάστε και πάλι τις πληθυσμιακές πυραμίδες για κάθε μια από τις 4 χώρες αλλά αυτή τη φορά για τις συχνότητες των ηλικιών ανά είδος απασχόλησης (er005) (δ) Ερμηνεύστε σύντομα το γράφημα.

E3 (α) Να κατασκευάσετε έναν πίνακα που να δείχνει το πλήθος των ατόμων του δείγματος ανά ηλικία (επιλέξτε 50 έως και 85 ετών) και απασχόληση (*crosstabulation*). (β) Σε κάθε μια από τις 4 χώρες ξεχωριστά: Σε ποια ηλικία οι άντρες συνταξιούχοι γίνονται περισσότεροι από τους άντρες εργαζόμενους; Να απαντήσετε το ίδιο και για τις γυναίκες. (γ) Σε κάθε μια από τις 4 χώρες ξεχωριστά: Να κατασκευάσετε το ραβδόγραμμα των ηλικιών των εργαζομένων ανά φύλο (*bar/stacked*)

E4 (α) Να κατασκευάσετε (για κάθε χώρα) κυκλικά διαγράμματα (*pie charts*) που να δείχνουν τα ποσοστά των ατόμων που ζουν σε μεγάλη πόλη, προάστια, κτλ (περιοχή διαβίωσης). (β) Να κάνετε και τον αντίστοιχο πίνακα (με τα counts και τα ποσοστά).

E5. (α) Να δημιουργήσετε έναν πίνακα με το μέσο, τη διάμεσο, την τυπική απόκλιση και η μέγιστη τιμή του ετήσιου ακαθάριστου εισοδήματος από εργασία ($ydirn$), ανά χώρα (στα άτομα που έχουν $ydirn > 0$). Στις γραμμές να αντιστοιχούν οι 4 χώρες και στις στήλες ο μέσος (με δ.ε. 95%), η διάμεσος, η τυπική απόκλιση και η μέγιστη τιμή (*ανασχηματίστε κατάλληλα τον πίνακα που προκύπτει από τη διαδικασία explore*). (β) Σχολιάστε σύντομα τα αποτελέσματα.

(Στα E6 – E10 θα χρησιμοποιήσετε μόνο τα άτομα που έχουν μη μηδενικό ακαθάριστο εισόδημα από εργασία, δηλ. $ydirn > 0$)

E6. (α) Να κατασκευάσετε ένα σχήμα με 4 θηκογράμματα του ετήσιου ακαθάριστου εισοδήματος από εργασία, ανά χώρα (στα άτομα που έχουν $ydirn > 0$). (β) Να κάνετε το ίδιο ανά χώρα και ανά φύλο (ένα σχήμα με 8 θηκογράμματα). Η κλίμακα στον κάθετο άξονα του δεύτερου σχήματος να είναι από 0 έως 110000 ευρώ. (γ) Σχολιάστε σύντομα τα αποτελέσματα.

E7. (α) Να κατασκευάσετε ένα ραβδόγραμμα με τις μέσες τιμές του ετήσιου ακαθάριστου εισοδήματος (*υποδ. χρησιμ. την επιλογή bars represent: other statistic ...*) από εργασία, ανά χώρα (στα άτομα που έχουν $ydirn > 0$). Στις ράβδους να αναγράφεται και ο αντίστοιχος μέσος. Οι ράβδοι να είναι σε φθίνουσα διάταξη (από αριστερά προς τα δεξιά). (β) Να κάνετε το ίδιο με το (α) αυτή τη φορά ανά χώρα και ανά φύλο (*bar/clustered*). Πάνω στις ράβδους να φαίνονται και τα αντίστοιχα error bars (δ.ε. 95% για το μέσο) (*υποδ. χρησιμ. τα options..*)

E8. (α) Να κατασκευάσετε τα ιστογράμματα του ετήσιου ακαθάριστου εισοδήματος από εργασία, ανά χώρα (στα άτομα που έχουν $ydirn > 0$). Στα ιστογράμματα ο οριζόντιος άξονας να είναι από 0 έως 100000. Επίσης το εύρος των κλάσεων (*interval width*) να είναι 2000 (β) Να κατασκευάσετε τα ιστογράμματα του λογαρίθμου του ετήσιου ακαθάριστου εισοδήματος από εργασία, ανά χώρα (κατασκευάστε πρώτα κατάλληλη μεταβλητή).

E9. (α) Να ελέγξετε για κάθε χώρα ξεχωριστά αν το μέσο ετήσιο ακαθάριστο εισόδημα από εργασία είναι ίσο με 15000 ευρώ (στα άτομα που έχουν $ydirn > 0$) (*υποδ. χρησιμ. split file*). (β) Να κάνετε το ίδιο κρατώντας μόνο τα άτομα ηλικίας 50 ετών.

E10. Για κάθε μια από τις 4 χώρες: (α) Να ελέγξετε αν οι άντρες έχουν το ίδιο (ή διαφορετικό) μέσο ετήσιο ακαθάριστο εισόδημα από εργασία με τις γυναίκες (στα άτομα που έχουν $ydirn > 0$). (β) Να ελέγξετε αν οι άντρες έχουν μεγαλύτερο μέσο ετήσιο ακαθάριστο εισόδημα από εργασία από ό,τι οι γυναίκες. (γ) Έχουν την ίδια διασπορά ακαθάριστου εισοδήματος; (δ) Να κάνετε τα ίδια κρατώντας μόνο τα άτομα ηλικίας 50 ετών.

E11. (α) Να δημιουργήσετε μια νέα μεταβλητή (*ageclass*) που να δείχνει την ηλικιακή ομάδα. Να παίρνει τις τιμές 1,2,3 ανάλογα με το αν η ηλικία είναι μεταξύ 50-65, 66-75, και 76 και άνω αντίστοιχα. (β) Να δημιουργήσετε μια νέα μεταβλητή (*gier*) που να δείχνει το σύνολο των ετήσιων ακαθάριστων εισοδημάτων από εργασία και συντάξεις (αθροίστε κατάλληλες μεταβλητές). (γ) Να κατασκευάσετε ένα ραβδόγραμμα με τις μέσες τιμές του ετήσιου ακαθάριστου εισοδήματος από εργασία και συντάξεις ανά ηλικιακή ομάδα και ανά χώρα (στα άτομα που έχουν $gier > 0$). Στις ράβδους να αναγράφεται και ο αντίστοιχος μέσος και να φαίνονται και τα αντίστοιχα error bars (δ.ε. 95% για το μέσο). (δ) Να κατασκευάσετε (για κάθε χώρα) κυκλικά διαγράμματα (*pie charts*) που να δείχνουν τα ποσοστά εργαζομένων, συνταξιούχων κλπ (*job situation*) ανά ηλικιακή ομάδα (*ageclass*) (12 πίττες σε ένα σχήμα, *χρησιμ. panel variables*). Να κάνετε και τον αντίστοιχο πίνακα (με τα counts και τα ποσοστά ανά ηλικιακή ομάδα).

E12. (α) Να κατασκευάσετε ένα πίνακα συχνοτήτων και το αντίστοιχο ραβδόγραμμα με το πλήθος των παιδιών των γυναικών ηλικίας 50-80. Τι ποσοστό των γυναικών έχει το πολύ δύο παιδιά; (β) Να ελέγξετε αν το πλήθος των παιδιών γυναίκας ηλικίας 50-55 ετών είναι μια τ.μ. που λαμβάνει τις τιμές 0,1,2,3,4, ≥ 5 με αντίστοιχες πιθανότητες 0.14, 0.20, 0.45, 0.15, 0.06.

E13. (α) Να κατασκευάσετε ένα ραβδόγραμμα (*clustered*) το οποίο να δείχνει το μέσο πλήθος παιδιών (γυναίκας ηλικίας 50-80) ανά τόπο διαβίωσης και ανά χώρα (μαζί με τα error bars). (β) Περιγράψτε σύντομα τι παρατηρείτε. (γ) Να ελέγξετε αν το μέσο πλήθος παιδιών γυναικών ηλικίας 50-80 που ζουν σε μια μεγάλη πόλη είναι ίσο με αυτών που ζουν σε ένα χωριό (για κάθε χώρα ξεχωριστά με *split file*).

E14. (α) Να δημιουργήσετε μια νέα μεταβλητή (*ageclass2*) που να δείχνει την ηλικιακή ομάδα (τύπου 2). Να παίρνει τις τιμές 50,55,60,65,... ανάλογα με το αν η ηλικία αντίστοιχα είναι μεταξύ 50-54, 55-59, 60-64 κ.ο.κ. (β) Να κατασκευάσετε ένα line chart (*multiple*) με το μέσο πλήθος παιδιών ανά ηλικιακή ομάδα (*ageclass2*) και ανά χώρα (4 γραμμές, κάθε χώρα διαφορετική γραμμή). Να εξομαλύνετε τις γραμμές του διαγράμματος αυτού χρησιμοποιώντας *splines*. (γ) Περιγράψτε σύντομα τι παρατηρείτε.

E15. (α) Να δημιουργήσετε μια μεταβλητή που θα εκφράζει τον δείκτη κατάθλιψης (άθροισμα των μεταβλητών *euro1+euro2+...+euro12*) (β) Να κατασκευάσετε ένα (*clustered*) ραβδόγραμμα με το μέσο δείκτη κατάθλιψης ανά χώρα και ανά φύλο (μαζί με τα error bars) (γ) Να κάνετε το ίδιο, ανά είδος συγκατοίκησης (*living with a spouse/partner*) και ανά φύλο. (δ) Περιγράψτε σύντομα τι παρατηρείτε.

E16. (α) Να κατασκευάσετε ένα line chart (*multiple*) με το μέσο δείκτη κατάθλιψης ανά ηλικιακή ομάδα (*ageclass2*) και ανά χώρα (4 γραμμές, κάθε χώρα διαφορετική γραμμή). Να εξομαλύνετε τις γραμμές του διαγράμματος αυτού χρησιμοποιώντας *splines*. Περιγράψτε σύντομα τι παρατηρείτε. (β) Να ελέγξετε αν ο μέσος δείκτης κατάθλιψης είναι ίσος μεταξύ αυτών που πίνουν και αυτών που δεν πίνουν περισσότερα από 2 ποτήρια αλκοόλ την ημέρα. (ξεχωριστά για γυναίκες - άντρες).

E17. Σε όσα ερωτήματα ακολουθούν (E17-E20) να κρατήσετε μόνο ένα άτομο από κάθε νοικοκυριό (με *select cvid = 1*). (α) Να κατασκευάσετε ένα ραβδόγραμμα με το μέσο ετήσιο ακαθάριστο εισόδημα (*hgtincv*) ενός νοικοκυριού ανά χώρα. (β) Να κατασκευάσετε ένα ραβδόγραμμα για κάθε χώρα με : το μέσο ετήσιο ακαθάριστο εισόδημα (*hgtincv*), μέσο ύψος καταθέσεων στην τράπεζα (*hbaccv*), μέση αξία ομολόγων (*hbondv*), μέση αξία μετοχών (*hstocv*), μέσο ύψος επενδύσεων σε αμοιβαία κεφάλαια (*hmutfv*) ανά περιοχή διαβίωσης του νοικοκυριού (χρησιμ. *Clustered bar/symmaries of separate vars*). (γ) Σχολιάστε σύντομα τα γραφήματα.

E18. Να ελέγξετε αν ο λογάριθμος του ετήσιου ακαθάριστου εισοδήματος ενός νοικοκυριού μπορεί να προέρχεται από την κανονική κατανομή (για κάθε χώρα ξεχωριστά με *split file*, χρησιμοποιήστε Ιστόγραμμα με *normal curve*, P-P, Q-Q plots, K-S τεστ).

E19. Να ελέγξετε αν ο λογάριθμος της αξίας των μετοχών που κατέχει ένα νοικοκυριό μπορεί να προέρχεται από την κανονική κατανομή (για κάθε χώρα ξεχωριστά με *split file*, χρησιμοποιήστε Ιστόγραμμα με *normal curve*, P-P, Q-Q plots, K-S τεστ).

E20. Να ελέγξετε αν η κατανομή του λογάριθμου του ετήσιου ακαθάριστου εισοδήματος ενός νοικοκυριού σε μια μεγάλη πόλη, είναι ίδια με την αντίστοιχη κατανομή ενός νοικοκυριού σε χωριό (ανά χώρα, χρησιμοποιήστε τρία απαραμετρικά κριτήρια και περιγράψτε σύντομα σε ποια στατιστική συνάρτηση βασίζεται το καθένα).

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

Μεταβλητές

	Name	Label
1	sampid2	SAMPLE ID
2	cvid	COVERSCREEN ID
3	country	country
4	nchild	Number of children
5	ydipv	Annual gross income from employment previous year
6	annpen1v	Annual public old age pension previous year
7	annpen2v	Annual public early or pre-retirement pension previous year
8	annpen3v	Annual public disability insurance previous year
9	annpen4v	Annual public unemployment benefit or insurance previous year
10	annpen5v	Annual public survivor pension from partner previous year
11	annpen6v	Annual public invalidity or incapacity pension previous year
12	annpen7v	Annual war pension previous year
13	annrentc	Hhd annual rent paid
14	annocscv	HHd annual other rent-related expenditures
15	hnetwv	Hhd net worth
16	hgtincv	Hhd total gross income
17	hbaccv	Hhd bank accounts
18	hbondv	Hhd government and corporate bonds
19	hstocv	Hhd stocks/shares
20	hmutfv	Hhd mutual funds
21	gender	male or female
22	yrbirth	year of birth
23	mstat	living with spouse/partner
24	euro1	depression
25	euro2	pessimism
26	euro3	suicidality
27	euro4	guilt
28	euro5	sleep
29	euro6	interest
30	euro7	irritability
31	euro8	appetite
32	euro9	fatigue
33	euro10	concentration
34	euro11	enjoyment
35	euro12	tearfulness
36	chronic	number of chronic disease
37	cusmoke	current smoking
38	drinkin2	drinking > 2 glasses of alcohol almost every or 5/6 days a week
39	ep005	current job situation
40	areabldg	area of building