Χρήστος Πατρινόπουλος

3200150

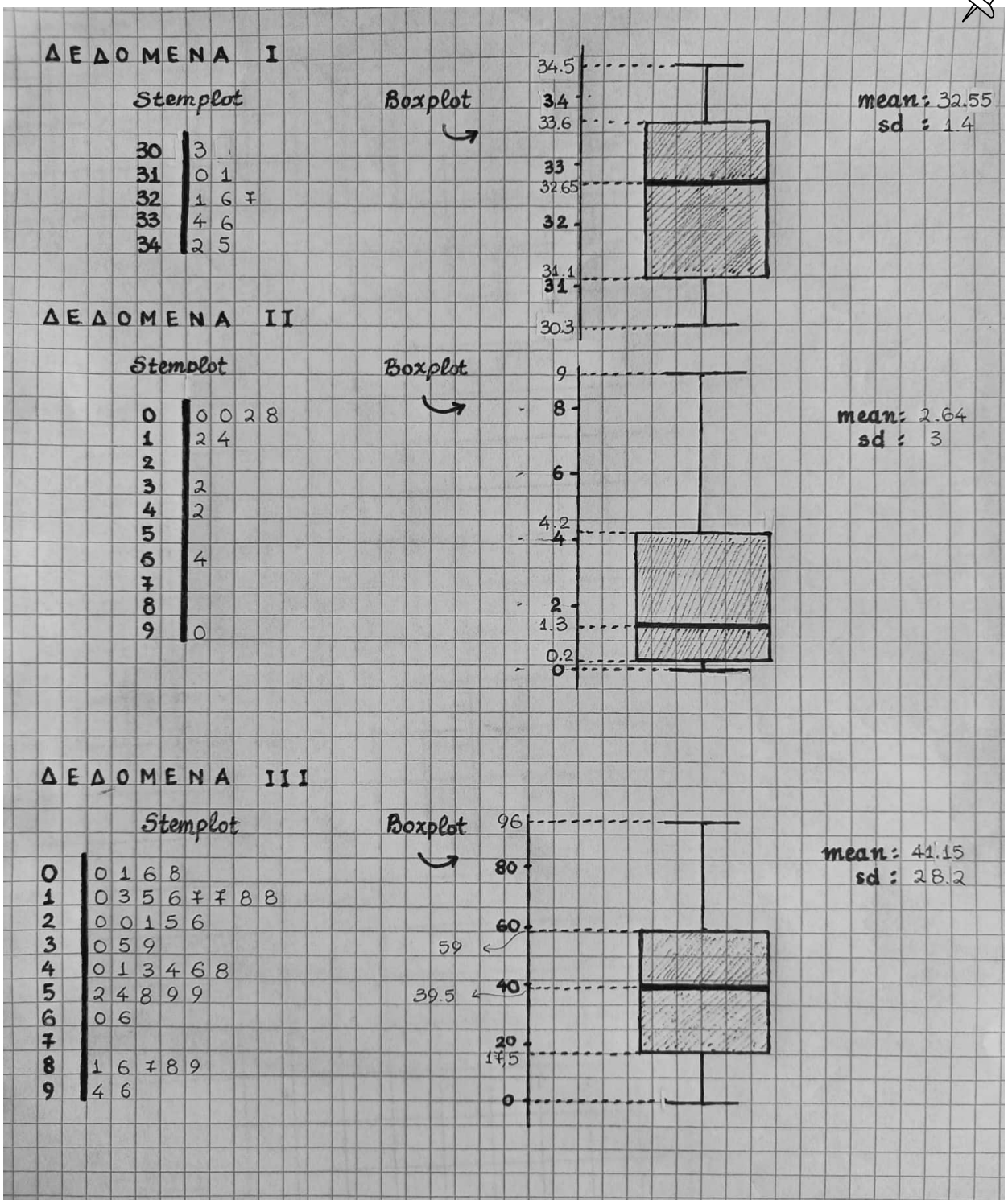
Αλβιόνα Μάντσο

3200098

Στατιστική στην Πληροφορική

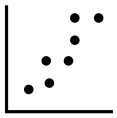
Άσκηση 1

a)



b)

ΔΕΔΟΜΕΝΑ Ι:



Από το stemplot και το boxplot φαίνεται να υπάρχει συσπείρωση δεδομένων κοντά στη μέση τιμή, ενώ επιπλέον mean και median είναι πολύ κοντινά γεγονός που υποδηλώνει μια ισοκατανομή των τιμών:

- mean = 32.55 και median = 32.65 και άρα mean ~ median

Άρα η κατανομή συνοψίζεται καλύτερα από το ζεύγος της μέσης τιμής (mean) και τυπικής απόκλισης (s).

ΔΕΔΟΜΕΝΑ ΙΙ:

s, μ

Όπως φαίνεται από το stemplot και το boxplot υπάρχει ανισοκατανομή των δεδομένων που επιβεβαιώνεται και από τη διαφορά μεταξύ median και mean:

- mean = 2.64 και median = 1.3 και άρα med << mean αν το δούμε σε σχέση και με το συνολικό range των τιμών

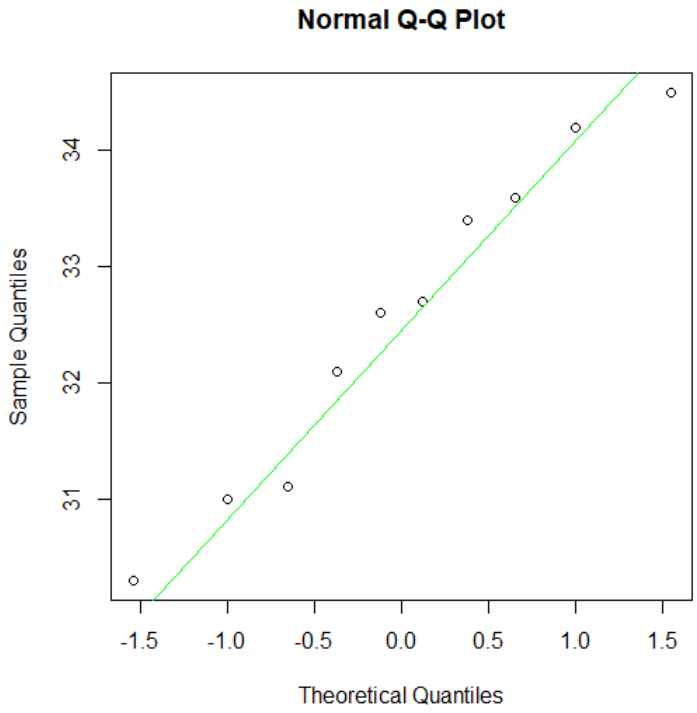
Άρα στην περίπτωση αυτή είναι πιο περιγραφική η σύνοψη των 5 αριθμών.

ΔΕΔΟΜΕΝΑ ΙΙΙ:

Στην περίπτωση αυτή, παρατηρούμε μια κατανομή πιο άνιση από των δεδομένων Ι αλλά και πιο συμμετρική από των δεδομένων 2, όπως φαίνεται από το stemplot και το boxplot. Ο mean=41.15 και ο median 39.5, τιμές που είναι σχετικά κοντινές (δεδομένου και πάλι του μεγάλου range των τιμών).

Άρα για τα δεδομένα ΙΙΙ δεν ενδείκνυται σαφέστατα κάποιος από τους δύο τρόπους, θα μπορούσαν να χρησιμοποιηθούν είτε η μέση τιμή-τυπική απόκλιση είτε η σύνοψη των 5 αριθμών.

c)



mean = 32.55

sd =1.4

```
{ sum( x > mean(x) - sd(x)  &  x < mean(x) + sd(x) ) / length(x) }
```

Το **50%** βρίσκονται μεταξύ των (mean - sd, mean + sd) έναντι **68%** η κανονική κατανομή.

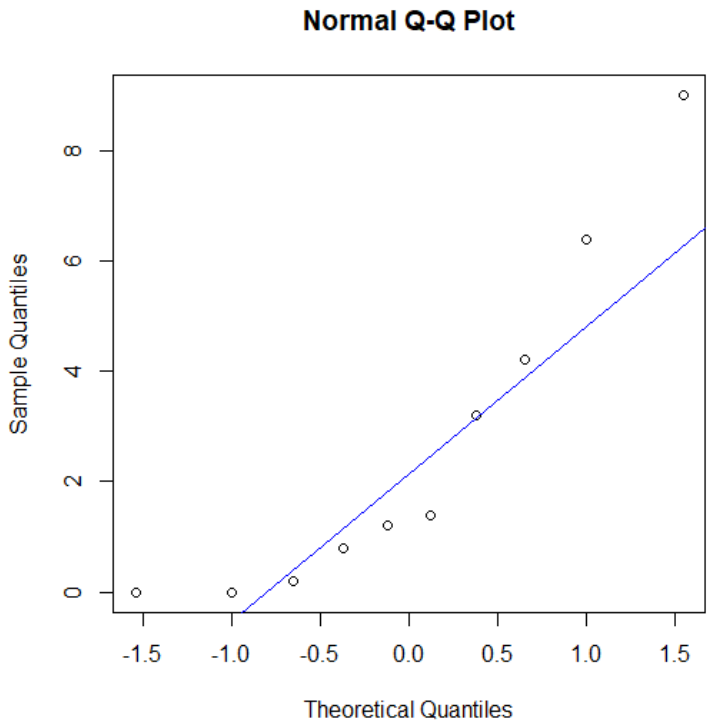
```
{ sum( x > mean(x) - sd(x)  &  x < mean(x) + sd(x) ) / length(x) }
```

Το **100%** βρίσκονται μεταξύ των (mean - 2*sd, mean + 2*sd) έναντι **95%** η κανονική κατανομή.`

```
{ sum( x > mean(x) - sd(x)  &  x < mean(x) + sd(x) ) / length(x) }
```

Το **100%** βρίσκονται μεταξύ των (mean - 3*sd, mean + 3*sd) έναντι **99.7%** η κανονική κατανομή.

Παρατηρείται σημαντική απόκλιση των ποσοστών σε σχέση με την κανονική κατανομή. Από την άλλη το quartile-quartile plot με την κανονική κατανομή προσεγγίζει μεν εν μέρει την ευθεία αλλά δεν αποτελεί ευθεία. Άρα η κανονική κατανομή δεν θα ήταν ιδιαίτερα ακριβής.



mean = 2.64

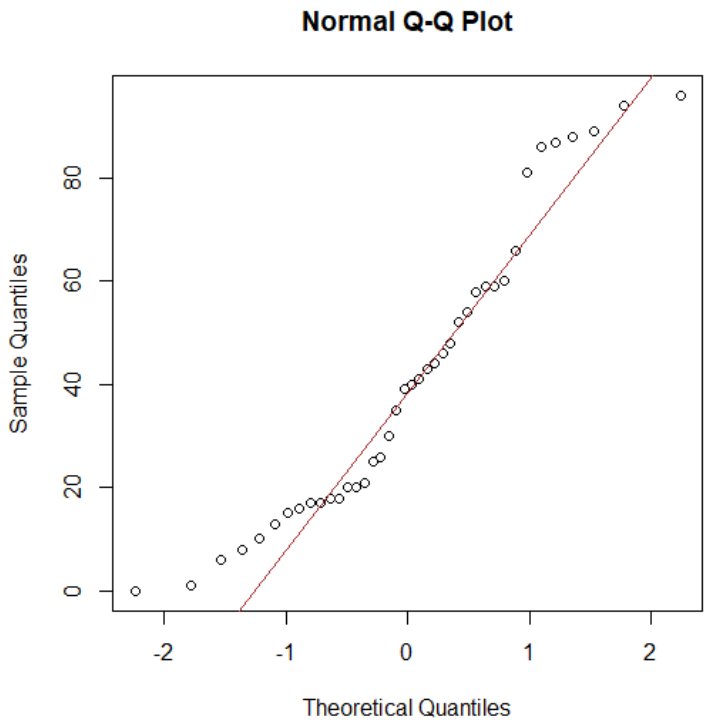
sd = 3

Το **80%** βρίσκονται μεταξύ των (mean - sd, mean + sd) έναντι **68%** η κανονική κατανομή.

Το **90%** βρίσκονται μεταξύ των (mean - 2*sd, mean + 2*sd) έναντι **95%** η κανονική κατανομή.

Το **100%** βρίσκονται μεταξύ των (mean - 3*sd, mean + 3*sd) έναντι **99.7%** η κανονική κατανομή.

Παρατηρείται σημαντική απόκλιση των ποσοστών σε σχέση με την κανονική κατανομή. Μάλιστα το quartile-quartile plot με την κανονική κατανομή δεν προσεγγίζει την ευθεία. Άρα η κανονική κατανομή δεν θα ήταν ακριβής, φαίνεται να πρόκειται για μια εκθετική κατανομή.



mean = 41.15

sd = 28.2

Το **70%** βρίσκονται μεταξύ των (mean - sd, mean + sd) έναντι **68%** η κανονική κατανομή.

Το **100%** βρίσκονται μεταξύ των (mean - 2*sd, mean + 2*sd) έναντι **95%** η κανονική κατανομή.

Το **100%** βρίσκονται μεταξύ των (mean - 3*sd, mean + 3*sd) έναντι **99.7%** η κανονική κατανομή.

Παρατηρείται εγγύτητα των ποσοστών με των αντίστοιχων της κανονικής κατανομής. Επιπλέον το quartile-quartile plot με την κανονική κατανομή

προσεγγίζει την ευθεία (αν και πάλι δεν πρόκειται για ευθεία). Άρα η κανονική κατανομή θα ήταν σε κάποιο βαθμό ακριβής.

Άσκηση 2

a) Τα δεδομένα της άσκησης αντλήθηκαν από τη βάση δεδομένων της Eurostat. Πιο συγκεκριμένα συνδυάστηκαν 2 data sets:

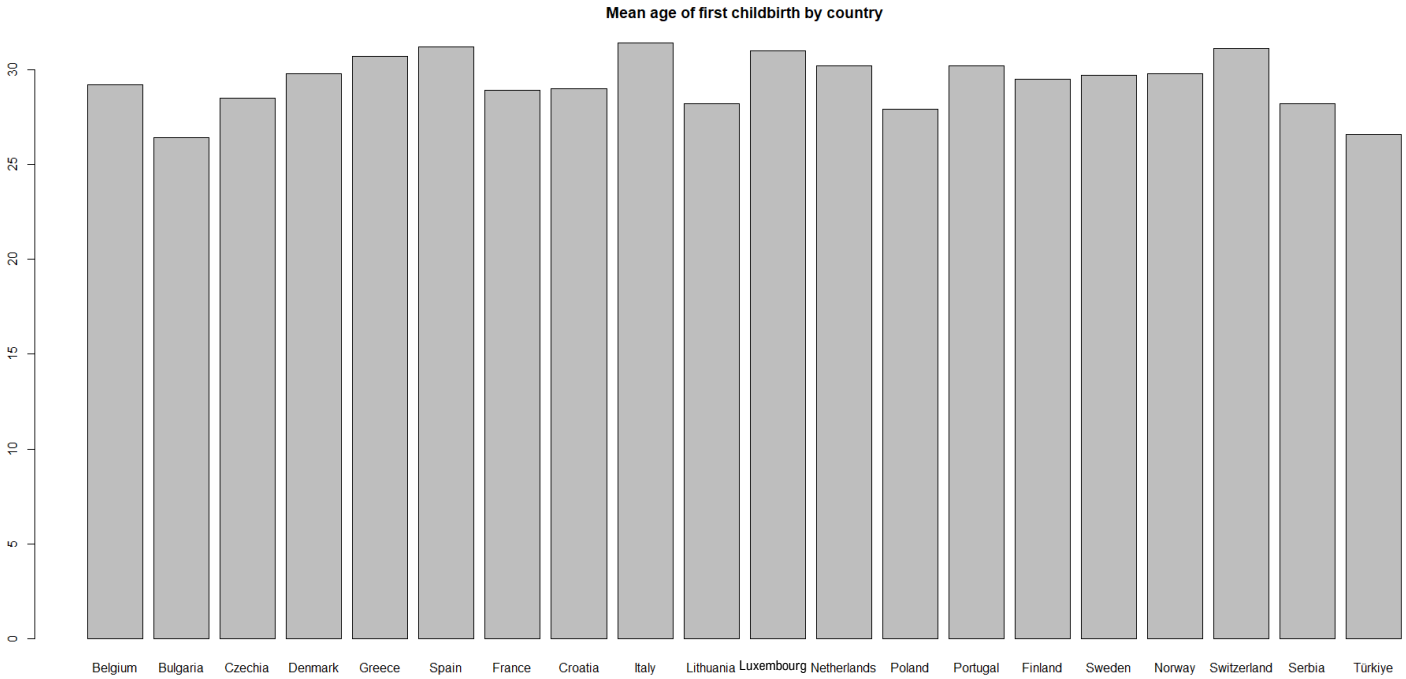
1) ***Mean age of women at childbirth and at birth of first child*** (url: https://ec.europa.eu/eurostat/databrowser/view/TPS00017__custom_3775772/default/table?lang=en)
2) ***Mean age at first marriage by sex*** (url: https://ec.europa.eu/eurostat/databrowser/view/TPS00014__custom_3775767/default/table?lang=en)

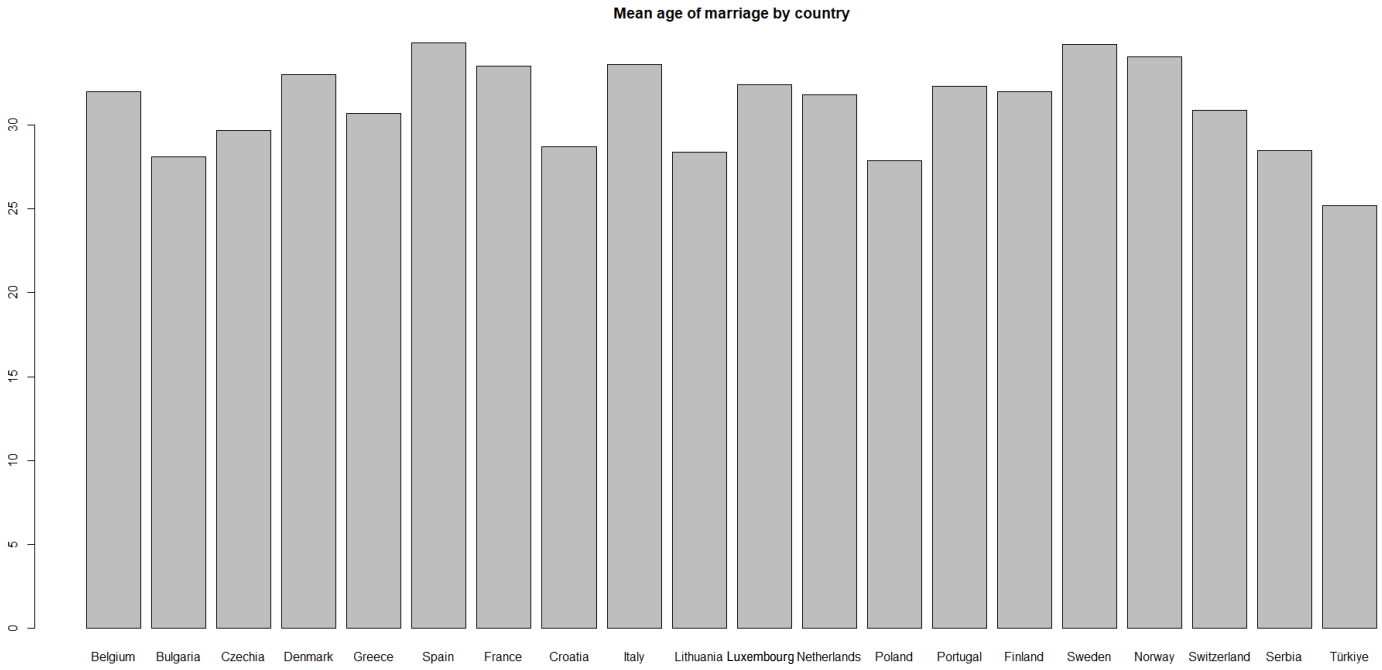
Επιλέχθηκαν 3 στήλες (country, mean age of women at childbirth, mean age at first marriage [women]) όπου το country ήταν κοινό και κρατήθηκαν 20 cases που αφορούν τις εξής ευρωπαϊκές χώρες: Belgium, Bulgaria, Czechia, Denmark, Greece, Spain, France, Croatia, Italy, Lithuania, Luxembourg, Netherlands, Poland, Portugal, Finland, Sweden, Norway, Switzerland, Serbia, Turkey. Τα δεδομένα και των 2 datasets αφορούν μετρήσεις του 2020.

b) **Κατηγορική μεταβλητή**: `country`
Ποσοτικές μεταβλητές: `childbirth` (mean age of women at childbirth), `marriage` (mean age at first marriage)

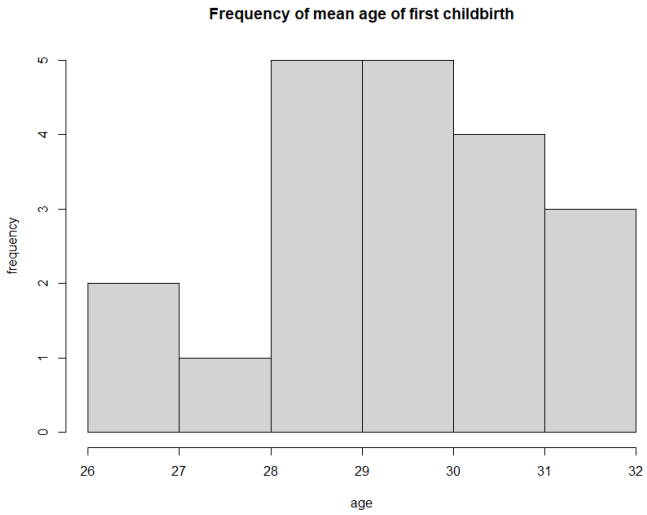
- `country`: ευρωπαϊκές χώρες (string)
- `childbirth`: ηλικία κατά την οποία οι γυναίκες γέννησαν το πρώτο τους παιδί κατά μέσο όρο σε μια χώρα το έτος 2020 (numeric)
- `marriage`: ηλικία κατά την οποία οι γυναίκες παντρεύτηκαν κατά μέσο όρο σε μια χώρα το έτος 2020 (numeric)

c) Κατηγορική (**country**)





Παρατηρούμε ότι εμφανίζονται μικρές αποκλίσεις (πάνω κάτω 5 έτη) μεταξύ των χωρών και στα δύο διαγράμματα, γεγονός που μπορεί να δικαιολογηθεί από την κοινή γεωγραφική θέση και το πολιτισμικό πλαίσιο (θρησκεία, συνήθειες, κλπ). Στις πιο ανεπτυγμένες χώρες (Σουηδία, Νορβηγία, κ.ά.) εμφανίζεται τάση γάμου σε μεγαλύτερη ηλικία από ότι στις λιγότερο ανεπτυγμένες (Τουρκία, Βουλγαρία, κ.ά.), ενώ στα έτη γέννησης του πρώτου παιδιού, οι διαφορές είναι μικρότερες.

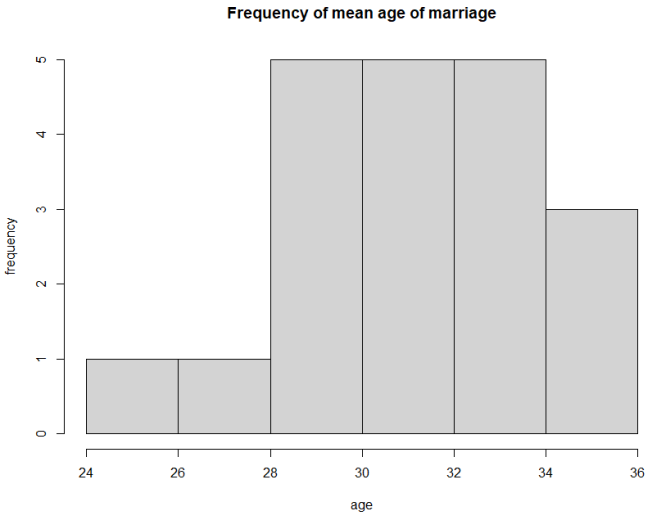


Ποσοτική (childbirth)

Στις ηλικίες 28 έως 30 παρατηρείται η μεγαλύτερη συχνότητα γεννήσεων πρώτου παιδιού. Μικρότερη είναι η συχνότητα σε ηλικίες κάτω των 28 ετών, ενώ δεν παρατηρούνται γενικά ατυπικές τιμές (θα βρίσκονταν μακριά από το κυρίως σώμα τιμών, αριστερά ή δεξιά).

Ποσοτική (marriage)

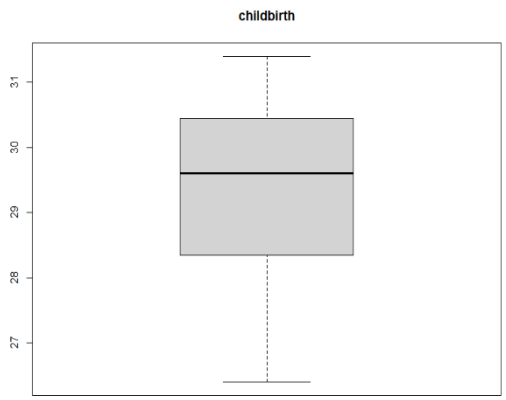
Ομοίως στις ηλικίες 28 με 34 παρατηρείται το μεγαλύτερο σώμα τιμών ηλικίας γάμου στις ευρωπαϊκές χώρες. Χαμηλές είναι πάλι οι τιμές σε ηλικίες κάτω των 28 ετών και δεν εμφανίζονται ατυπικές τιμές.



d)

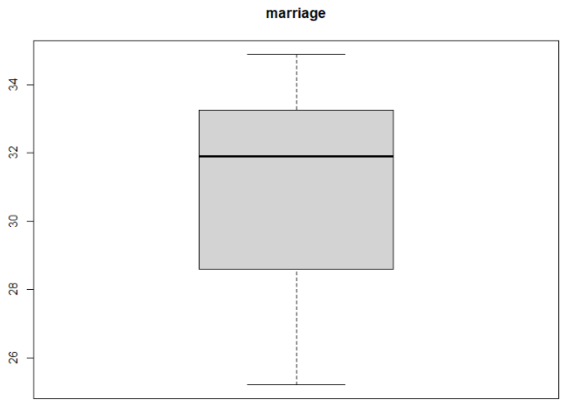
Μεταβλητή: childbirth

mean	29.38
sd	1.4381
min	26.40
Q1	28.35
median	29.60
Q3	30.32
max	31.40



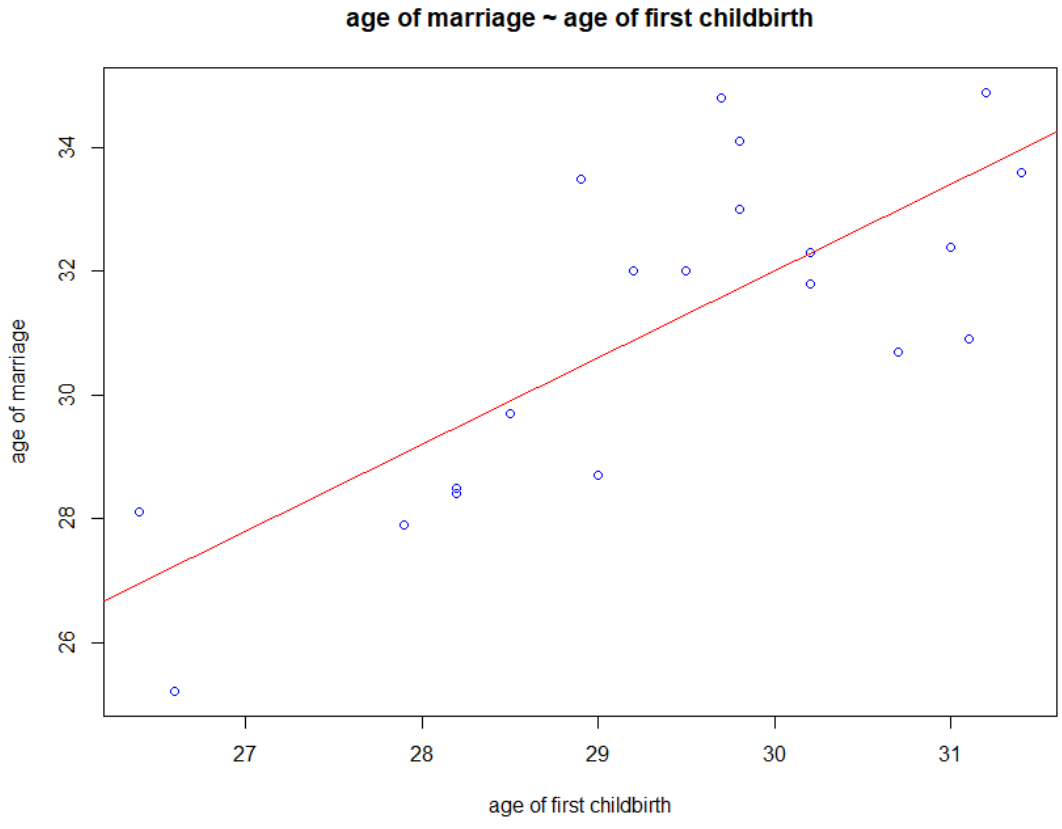
Μεταβλητή: **marriage**

mean	31.12
sd	2.6556
min	25.2
Q1	28.6
median	31.9
Q3	33.25
max	34.9



Όπως φαίνεται από το ιστόγραμμα και το boxplot, οι τιμές δεν είναι συμμετρικά κατανεμημένες γύρω από τη μέση τιμή, αφού πέρα τον πολλών τιμών κοντά στη μέση τιμή παρατηρείται μεγαλύτερη πυκνότητα τιμών στο τρίτο και τέταρτο τεταρτημόριο. Οπότε θεωρούμε πιο περιγραφικές τις τιμές των 5 αριθμών, από ότι η μέση τιμή με την τυπική απόκλιση.

e)



r = 0.7616901

Μεταξύ των δύο μεταβλητών παρατηρείται αύξουσα, μέτρια ισχυρή γραμμική σχέση, οπότε υπάρχει κάποια συσχέτιση μεταξύ των μεταβλητών, η οποία ίσως εμφανίζεται επειδή ζευγάρια τείνουν να παντρεύονται λόγω κάποιας απρόσμενης εγκυμοσύνης, κοντά στην ηλικία των 28-30. Σε αυτή την περίπτωση θεωρούμε ως αίτιο-επεξηγηματική μεταβλητή το birthchild και ως μεταβλητή απόκρισης το marriage.

Άσκηση 3

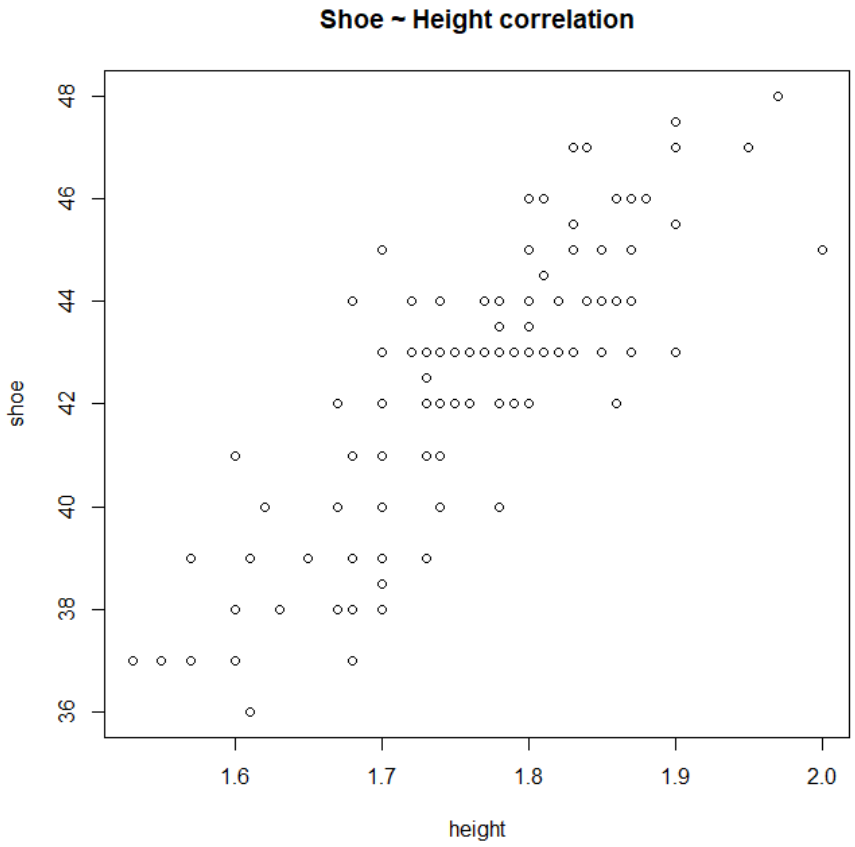
✓ Πηγή: **survey_data_2022.csv**

Ποσοτικές μεταβλητές:

- **height** (ύψος)
- **shoe** (μέγεθος παπουτσιού)

a)

```
{ plot(shoe~height, main="Shoe ~ Height correlation") }
```



Παρατηρούνται τα εξής:

- **Μορφή:** γραμμική
- **Κατεύθυνση:** αύξουσα
- **Δύναμη:** αρκετά ισχυρή

β)

r = 0.832742

Coefficients:

(Intercept)	height
-1.934	25.201

```
{   cor(shoe, height, use="complete.obs")
    m ← lm(shoe~height)
    abline(m, col="hotpink")
    m$residuals
    b1 ← sd(shoe)/sd(height)*cor(shoe, height)
    b0 ← mean(shoe) - b1*mean(height)   }
```

Linear Regression: $shoe_{expected} = 25.201 * height - 1.934$

* $shoe_{expected}$ = τιμή παπουτσιού πάνω στη γραμμή