**TEST- Quantitative Data Analysis (R)**

**(Διάρκεια: 3 ώρες)**

**Απαντήσεις - Νίκος Κουγιανός  
UEL Number: 2018506**

Σε μία περιβαλλοντολογική έρευνα εξετάζεται αν η ανάπτυξη ενός είδους φυτού επηρεάζεται από την γεωγραφική θέση της περιοχής που φύεται. Ο παρακάτω πίνακας περιέχει δεδομένα από 28 περιοχές της Ελλάδος που ευδοκιμεί το φυτό, και περιλαμβάνει το ύψος του φυτού σε μέτρα, και τη γεωγραφική θέση (1: Βόρεια Ελλάδα, 2: Νότια Ελλάδα).



Με βάση τα παραπάνω δεδομένα, εκτελέστε τις ακόλουθες εργασίες με την βοήθεια της R.

1. Να υπολογιστεί ο δειγματικός μέσος, η δειγματική διάμεσος, η δειγματική τυπική απόκλιση, καθώς και το 1ο και 3ο τεταρτημόριο του ύψους των φυτών.
2. Να κατασκευαστεί το ιστόγραμμα συχνοτήτων και σχετικών συχνοτήτων του ύψους των φυτών.

3. Να κατασκευαστούν στο ίδιο γράφημα δύο θηκογραφήματα που να αφορούν το ύψος των φυτών στην βόρεια και νότια Ελλάδα αντίστοιχα. Δώστε ονόματα σε κάθε θηκογράφημα.

4. Με την βοήθεια της εντολής plot να αναπαραστήσετε γραφικά τις τιμές του δείγματος για το ύψος των φυτών και εν συνεχεία με χρήση της παραμέτρου text εμφανίστε τους χαρακτήρες ‘1’ και ‘2’ στο γράφημα ξεχωριστά για τις βόρειες και νότιες περιοχές. Δώστε τίτλο στο γράφημά σας και μια λεζάντα που να εξηγεί την σημασία των χαρακτήρων ‘1’ και ‘2’.

5. Να υπολογιστεί η διάμεσος του ύψους των φυτών ξεχωριστά για τις βόρειες και τις νότιες περιοχές.

**ΑΠΑΝΤΗΣΕΙΣ**

**Σημείωση: Σε όλες τις λύσεις χρησιμοποιούνται οι εντολές**

**# Use these 2 lines to set working directory as the current directory of the**

**# file. Tested on Rstudio @ Windows 10 OS.**

**this.dir <- dirname(parent.frame(2)$ofile)**

**setwd(this.dir);**

**Οι οποίες σετάρουν το working directory της R σε αυτό που βρίσκεται μέσα το υπό εκτέλεση R αρχείο. Έτσι μπορούμε να τρέξουμε τα r scripts ανεξαρτήτως working directory. Σε περίπτωση που δε δουλέψουν αυτές οι εντολές τότε θα πρέπει να φορτωθεί το αρχέιο data.csv που συμπεριλαμβάνεται στα παραδοτέα με absolute path, ή να δοκιμαστεί με relative path χωρίς τις παραπάνω εντολές, με ρίσκο να μην φορτώσει το αρχείο δεδομένων.**

**Επίσης τα δεδομένα φορτώνονται σε κάθε r script ξεχωριστά για διευκόλυνση στην εκτέλεση.**

**Ερώτημα 1**  
Δειγματικός μέσος = 2.890357

Δειγματική διάμεσος = 2.715

Δειγματική τυπική απόκλιση = 1.662966

1ο τεταρτημόριο = 1.87

3ο τεταρτημόριο = 3.75

**Κώδικας R ερωτήματος 1:**

# Author: Nikos Kougianos

# Date: 18/12/2020

# Use these 2 lines to set working directory as the current directory of the

# file. Tested on Rstudio @ Windows 10 OS.

this.dir <- dirname(parent.frame(2)$ofile)

setwd(this.dir);

# Read data

data <- read.table("data.csv", sep=",",header=TRUE)

# Create dataframe with 2 columns instead of 4

df <- data.frame('Plant\_Height'=c(data$Plant.Height,data$Plant.Height.1), 'Geo\_Position'=c(data$Geo.Position,data$Geo.Position.1))

# Calculate mean, median, and standard deviation of plant heights

mean <- mean(df$Plant\_Height)

median <- median(df$Plant\_Height)

sd <- sd(df$Plant\_Height)

# Calculate quantiles

quantiles <- quantile(df$Plant\_Height)

# Calculate first quantile

quantile1 <- quantiles[2]

# Calculate third quantile

quantile3 <- quantiles[4]

**Ερώτημα 2.**

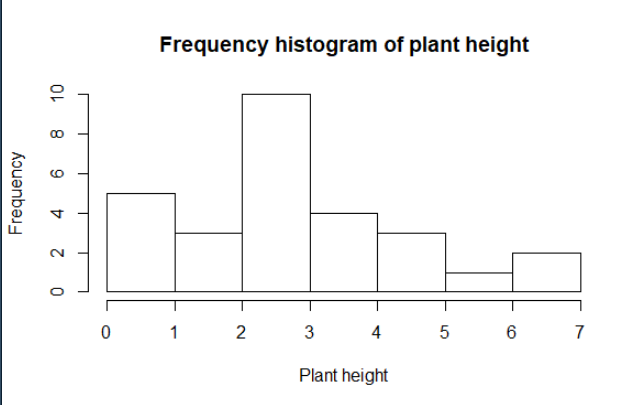
****

Figure 1- ιστόγραμμα συχνοτήτων

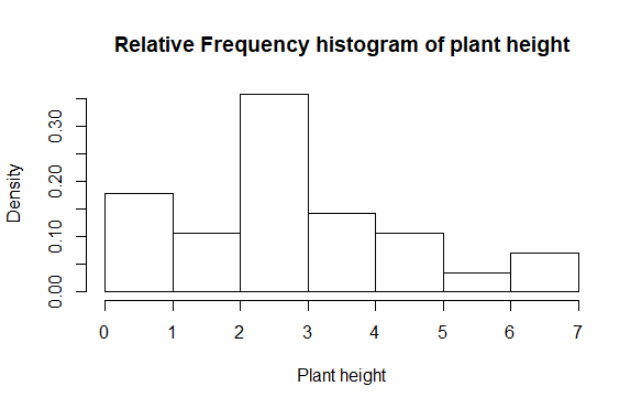


Figure 2- ιστόγραμμα σχετικών συχνοτήτων

**Κώδικας R ερωτήματος 2:**

# Author: Nikos Kougianos

# Date: 18/12/2020

# Use these 2 lines to set working directory as the current directory of the

# file. Tested on Rstudio @ Windows 10 OS.

this.dir <- dirname(parent.frame(2)$ofile)

setwd(this.dir);

# Read data

data <- read.table("data.csv", sep=",",header=TRUE)

# Create dataframe with 2 columns instead of 4

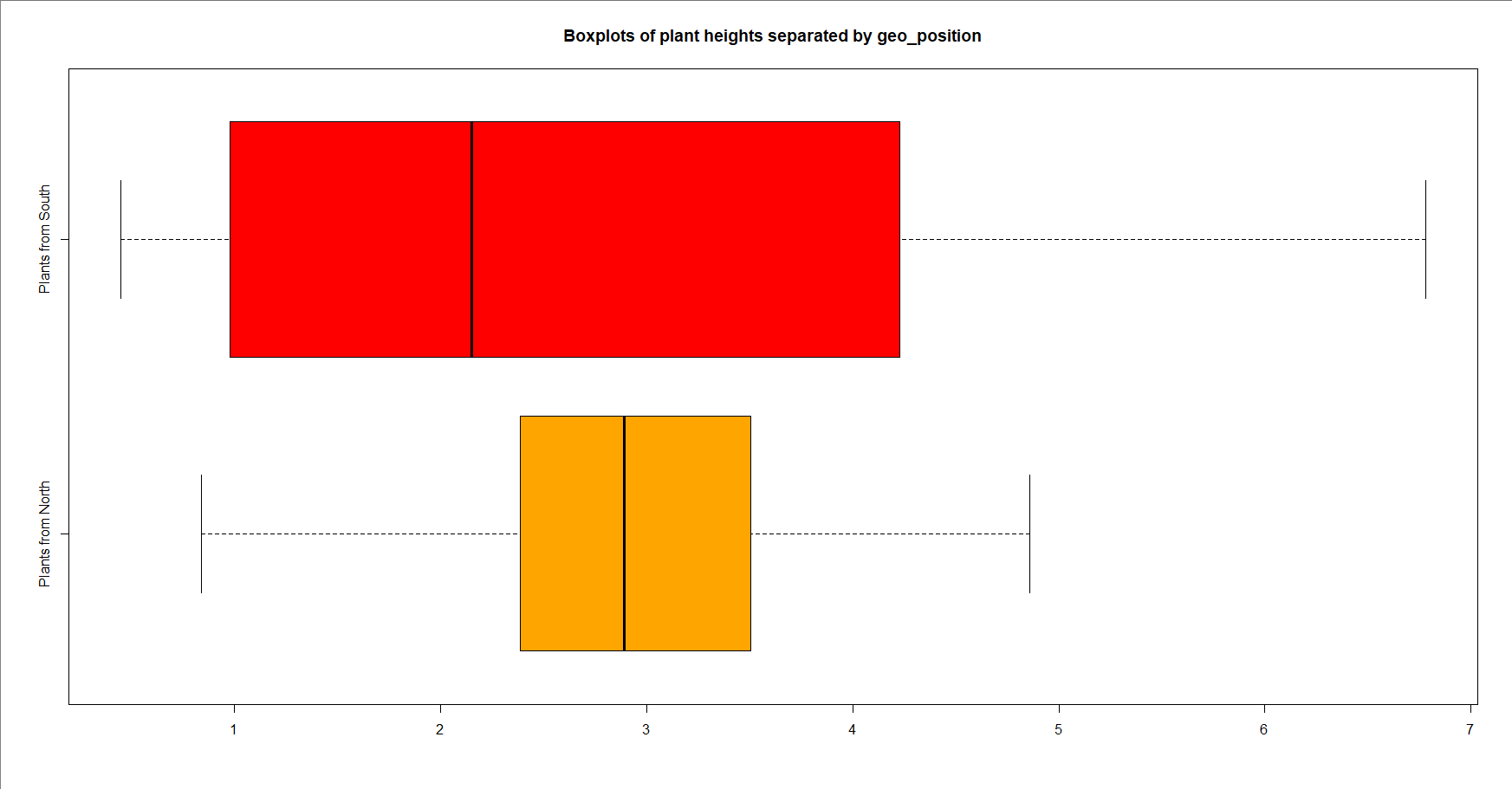
df <- data.frame('Plant\_Height'=c(data$Plant.Height,data$Plant.Height.1), 'Geo\_Position'=c(data$Geo.Position,data$Geo.Position.1))

# Create frequency histogram

frequency\_histogram <- hist(df$Plant\_Height, xlab="Plant height", main = "Frequency histogram of plant height")

# Create relative frequency histogram

relative\_frequency\_histogram <- hist(df$Plant\_Height, freq=FALSE, xlab="Plant height", main = "Relative Frequency histogram of plant height")

**Ερώτημα 3**

**Κώδικας R ερωτήματος 3:**

# Author: Nikos Kougianos

# Date: 18/12/2020

# Use these 2 lines to set working directory as the current directory of the

# file. Tested on Rstudio @ Windows 10 OS.

this.dir <- dirname(parent.frame(2)$ofile)

setwd(this.dir);

# Read data

data <- read.table("data.csv", sep=",",header=TRUE)

# Create dataframe with 2 columns instead of 4

df <- data.frame('Plant\_Height'=c(data$Plant.Height,data$Plant.Height.1), 'Geo\_Position'=c(data$Geo.Position,data$Geo.Position.1))

# Subset dataframe, extract height plants where plant is from North

plantsNorth <- df$Plant\_Height[df$Geo\_Position==1]

# Subset dataframe, extract height plants where plant is from South

plantsSouth <- df$Plant\_Height[df$Geo\_Position==2]

# Create a boxplot that contains both plants from North and plants from South in the same graph

# You can set horizontal=FALSE to swap x and y axis

boxplot(plantsNorth, plantsSouth,

col = c("orange","red"),

names = c("Plants from North", "Plants from South"),

horizontal=TRUE,

main="Boxplots of plant heights separated by geo\_position")

**Ερώτημα 4**

**Στο συγκεκριμένο ερώτημα αντί για την εντολή text χρησιμοποιήθηκε η εντολή legend η οποία βάζει την λεζάντα που ζητείται εκτός του plot για να υπάρχει πιο καθαρή εικόνα. Στο legend επεξηγείται ότι στον άξονα Χ (Plant Geo Position) το 1.0 σημαίνει ότι το φυτό είναι από την Βόρεια Ελλάδα και το 2.0 σημαίνει ότι το φυτό είναι από την Νότια Ελλάδα.**

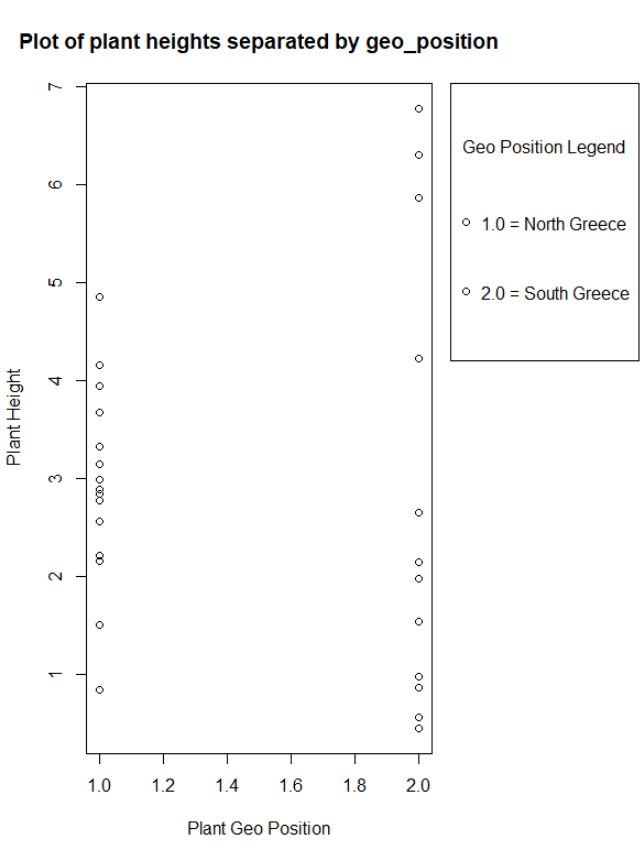


Figure 3- διάγραμμα plot ύψους φυτών σε συνάρτηση με το γεωγραφικό μέρος

**Κώδικας R ερωτήματος 4**

# Author: Nikos Kougianos

# Date: 18/12/2020

# Use these 2 lines to set working directory as the current directory of the

# file. Tested on Rstudio @ Windows 10 OS.

this.dir <- dirname(parent.frame(2)$ofile)

setwd(this.dir);

# Read data

data <- read.table("data.csv", sep=",",header=TRUE)

# Create dataframe with 2 columns instead of 4

df <- data.frame('Plant\_Height'=c(data$Plant.Height,data$Plant.Height.1), 'Geo\_Position'=c(data$Geo.Position,data$Geo.Position.1))

# Create plot parameters

geoPosition <- df[,'Geo\_Position']

plantHeight <- df[,'Plant\_Height']

# Create plot

plot(geoPosition, plantHeight,

main = "Plot of plant heights separated by geo\_position" ,

xlab="Plant Geo Position",

ylab="Plant Height"

)

# Add extra space for legend

par(mar=c(5.1, 4.1, 4.1, 10.1))

# Add legend

legend("topright", inset=c(-0.6,0), legend=c("1.0 = North Greece","2.0 = South Greece"), pch=c(1,1), title="Geo Position Legend")

**Ερώτημα 5**

Διάμεσος ύψους φυτών βόρειας περιοχής = 2.89

Διάμεσος ύψους φυτών νότιας περιοχής = 2.15

**Κώδικας R ερωτήματος 5:**

# Author: Nikos Kougianos

# Date: 18/12/2020

# Use these 2 lines to set working directory as the current directory of the

# file. Tested on Rstudio @ Windows 10 OS.

this.dir <- dirname(parent.frame(2)$ofile)

setwd(this.dir);

# Read data

data <- read.table("data.csv", sep=",",header=TRUE)

# Create dataframe with 2 columns instead of 4

df <- data.frame('Plant\_Height'=c(data$Plant.Height,data$Plant.Height.1), 'Geo\_Position'=c(data$Geo.Position,data$Geo.Position.1))

# Subset dataframe, extract height plants where plant is from North

plantsNorth <- df$Plant\_Height[df$Geo\_Position==1]

# Subset dataframe, extract height plants where plant is from South

plantsSouth <- df$Plant\_Height[df$Geo\_Position==2]

# Calculate median of heights of North plants

medianNorth <- median(plantsNorth)

# Calculate median of heights of South plants

medianSouth <- median(plantsSouth)