Μαθηματική Στατιστική

Εργασία 12

Ονομ/νο: Νούλας Δημήτριος
ΑΜ: 1112201800377
email: dimitriosnoulas@gmail.com

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν Πανεπιστήμιον Αθηνών
ΙΔΡΥΘΕΝ ΤΟ 1837

Εκφώνηση:

(1) Αναφέρετε και ερμηνεύστε τα αποτελέσματα του ελέγχου Shapiro-Wilk με την χρήση της p-value (και για τους δύο ελέγχους). Η μηδενική υπόθεση αντιστοιχεί στην υπόθεση ότι τα δεδομένα $(night - day)$ προέρχονται από κανονική κατανομή και η εναλλακτική ότι δεν προέρχονται για κάθε μία από τις 2 ομάδες ξεχωριστά.

(2) Στο μέρος αυτό της εργασίας θα ανακαλύψετε τον έλεγχο Shapiro-Wilk. Ανοίξτε τις σημειώσεις της Μη-παραμετρικής Στατιστικής (σελ. 135) και περιγράψτε σύντομα τα βασικά στοιχεία του ελέγχου Shapiro-Wilk.

(3) Προσομοιώστε 50 δεδομένα από $\mathcal{N}(0,1), t_3$ και t_{10} και πραγματοποιήστε τους ελέγχους κανονικότητας. Σχολιάστε τα αποτελέσματα που πήρατε με την βοήθεια της p-value. Τι θα απαντούσατε αν πρέπει να απαντήσετε σε επίπεδο στατιστικής σημαντικότητας $a = 0.05$;

(4) Πραγματοποιήστε ένα κατάλληλο t-test για να ελέγξετε αν η μέση αύξηση στην περίοδο των καρδιακών χτύπων είναι διαφορετική μεταξύ των 2 ομάδων και ένα t-test για να ελέγξετε αν η μέση αύξηση στην περίοδο των καρδιακών χτύπων είναι μεγαλύτερη για τα άτομα της ομάδας 2 σε σχέση με αυτή της ομάδας 1. Να ελεγχθούν αυτά με ε.σ.σ. $a = 0.05$ με την βοήθεια του λογισμικού της R. Οι διασπορές των αντίστοιχων κατανομών θεωρούνται άγνωστες και όχι κατὰνάγκη ίσες [υπόδειξη: δείτε Welch t-test και πώς εφαρμόζεται στην R].

(5) Ας υποθέσουμε τώρα ότι κάνουμε το μετασχηματισμό δεδομένων $\log(night/day)$. Θέτουμε Z_i και W_i τα δεδομένα που προκύπτουν με αυτό το μετασχηματισμό για τα μέλη της ομάδας 1 και 2 αντίστοιχα. Επαναλάβετε τους ελέγχους κανονικότητας για τα δεδομένα αυτά. Επαναλάβετε επίσης τα t-test.

(6) (προαιρετικό) Ποιό από τα 2 μοντέλα (με ή χωρίς μετασχηματισμό) σας φαίνεται καλύτερο για να μπορέσουμε να φτάσουμε σε πιο ασφαλή συμπεράσματα ως προς την ύπαρξη στατιστικά σημαντικής διαφοράς μεταξύ των δύο ομάδων; Μπορείτε εδώ να κάνετε οποιαδήποτε διερεύνηση σας φαίνεται σχετική.

Λύση:

(1) Αρχικά κάνουμε την επιλογή set.seed(625) για όλες τις εντολές που θα ακολουθήσουν στην εργασία. Με τις εντολές:

```
1    shapiro.test(g1diff)
2    shapiro.test(g2diff)
```

παίρνουμε τα αποτελέσματα:

```
> shapiro.test(g1diff)

        Shapiro-Wilk normality test

data:  g1diff
W = 0.98286, p-value = 0.5057

> shapiro.test(g2diff)

        Shapiro-Wilk normality test

data:  g2diff
W = 0.9783, p-value = 0.7184
```

Βλέπουμε ότι η ελεγχοσυνάρτηση W δίνει τιμή αρκετά κοντά στο 1, δηλαδή οι κατανομές από τις οποίες προέρχονται τα δείγματα "ταιριάζουν" αρκετά με την κανονική. Αυτό συμφωνεί με τα υψηλά p-value που προκύπτουν από τον έλεγχο. Καθώς είναι υψηλά, δεν απορρίπτουμε την μηδενική υπόθεση για κανέναν από τους 2 ελέγχους, δηλαδή δεχόμαστε ότι τα δείγματα προέρχονται από κανονική κατανομή.

(2) Για ένα τυχαίο δείγμα X μεγέθους n θέλουμε να ελέγξουμε κατά πόσο η κατανομή του ταιριάζει με την κανονική με άγνωστα μ, σ^2. Αφού διατάξουμε τις παρατηρήσεις προσομοιώνουμε ένα δείγμα ίδιου μεγέθους $(Z_1, \dots Z_n)$ από τυπική κανονική κατανομή και επιπλέον το διατάσσουμε. Ορίζουμε:

$$m = (\mu_{(1)}, \dots, \mu_{(n)})^T$$

όπου $\mu_{(i)} = E(Z_{(i)})$ και θεωρούμε τον συμμετρικό πίνακα συνδιακύμανσης V διάστασης $n \times n$ όπου το ij-στοιχείο του είναι το $Cov(Z_{(i)}, Z_{(j)})$.

Θεωρώντας την Ευκλείδεια 2-νόρμα $||\cdot||_2$, θέτουμε το διάνυσμα a με νόρμα 1 ως εξής:

$$a = (a_1, \dots, a_n)^T = \frac{V^{-1} \cdot m}{||V^{-1} \cdot m||}$$

Εδώ ο πίνακας V είναι αντιστρέψιμος καθώς αν είχε ιδιοτιμή 0 και παίρναμε ένα μη μηδενικό διάνυσμα $b \in \mathbb{R}^n$ του ιδιοχώρου που αντιστοιχεί στο 0, έτσι ώστε $Vb = 0$, τότε θα είχαμε:

$$0 = b^T V b = \sum_{ij} b_j Cov(Z_{(i)}, Z_{(j)}) b_i = Var(\sum_i b_i Z_{(i)}) = |b_1| + |b_2| + \dots |b_n|$$

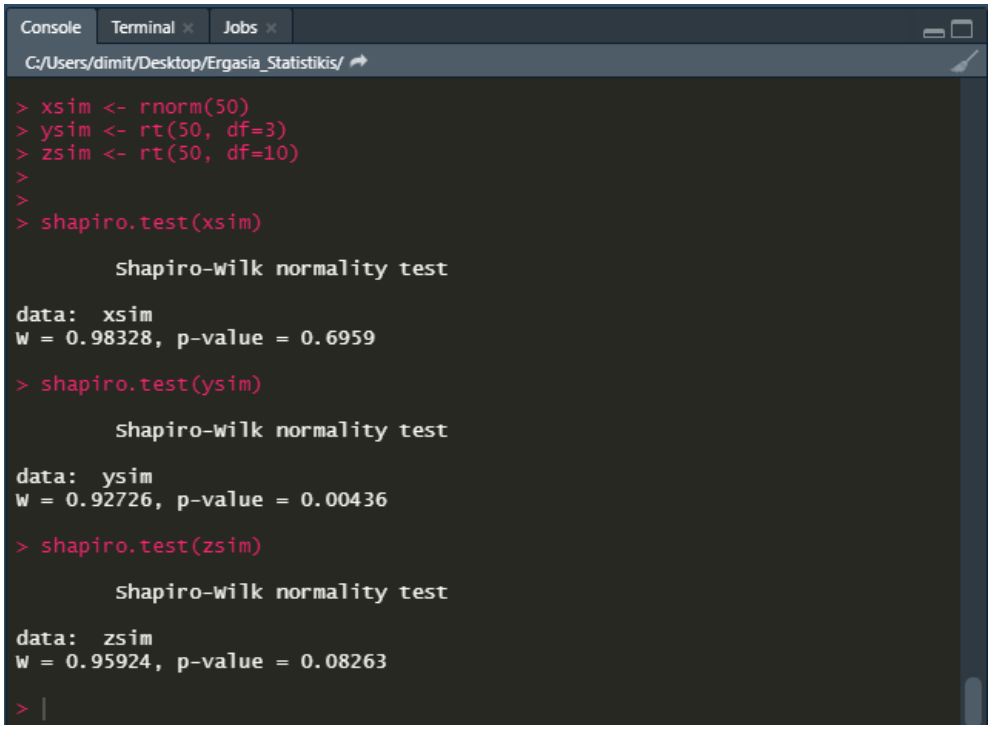
εφόσον $Z_{(i)}$ ανεξάρτητες και ισόνομες. Αυτό έιναι άτοπο, καθώς το $b \neq 0$.

Ορίζουμε την ελεγχοσυνάρτηση W, με $0 < W < 1$ ως εξής:

$$W(X) = \frac{\left(\sum_i^n a_i X_{(i)}\right)^2}{\sum_i^n \left(X_i - \overline{X}\right)^2}$$

και ο έλεγχος είναι ο εξής: Όσο πιο κοντά στο 1 είναι το W τόσο πιο πολύ "ταιριάζει" η κατανομή με κανονική και όσο πιο μακριά από το 1 τόσο πιο πολύ "απέχει".

(3) Γνωρίζουμε ότι η κατανομή student με ν βαθμούς ελευθερίας τείνει στην κανονική όσο το ν μεγαλώνει. Οι περισσότερες προσομοιώσεις των εντολών έχουν ένα αποτέλεσμα σαν το ακόλουθο:



Η ελεγχοσυνάρτηση W είναι πολύ κοντά στο 1 σε όλες τις περιπτώσεις, πιο πολύ στην ίδια την κανονική κατανομή και ακολουθεί η t_{10} η οποία "ταιριάζει" περισσότερο με την κανονική από ότι η t_3.

Όπως είναι αναμενόμενο για το δείγμα από την κανονική κατανομή, το p-value είναι και αυτό αρκετά υψηλό. Ωστόσο, στο δείγμα από t_{10} το W παραμένει υψηλό, ενώ το p-value είναι πολύ χαμηλότερο. Βεβαια, στην συγκεκριμένη προσομείωση δεν είναι αρκετά χαμηλό για να απορρίψουμε την μηδενική υπόθεση αν έχουμε θεωρήσει επίπεδο στατιστικής

σημαντικότητας $a = 0.05$. Παρόλα αυτά, το W της t_3 που ταιριάζει λιγότερο με την κανονική παραμένει αρκετά υψηλό. Αυτό έρχεται σε αντίθεση με το χαμηλό p-value με βάση το οποίο στην περίπτωση του t_3 απορρίπτουμε την μηδενική υπόθεση.

(4) Για να απαντήσουμε στο αν η μέση αύξηση των καρδιακών ρυθμών είναι διαφορετική μεταξύ των δύο ομάδων, με ε.σ.σ. $a = 0.05$ χρησιμοποιούμε την εντολή:

```
t.test(g1diff,g2diff,var.equal=FALSE,conf.level = 0.95)
```

Μπορούμε στο τελευταίο όρισμα να αλλάξουμε το $(1-a)$-διάστημα εμπιστοσύνης καθώς και να δώσουμε σαν πληροφορία ότι οι διασπορές των αντίστοιχων κατανομών είναι ίσες.

Το Welch t-test το οποίο χρησιμοποιεί το var.equal=FALSE και χωρίς να το προσδιορίσουμε, κάνει εκτίμηση για τις διασπορές και προσαρμόζει τους βαθμούς ελευθερίας που θα χρησιμοποιηθούν στον έλεγχο. Διαφορετικά, αν γνωρίζουμε ότι οι διασπορές είναι ίσες χρησιμοποιούμε το απλό t-test εφόσον προσδιορίσουμε var.equal=TRUE.

```
> t.test(g1diff,g2diff,var.equal=FALSE,conf.level = 0.95)

        Welch Two Sample t-test

data:  g1diff and g2diff
t = -2.6017, df = 59.153, p-value = 0.0117
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -82.05697 -10.71226
sample estimates:
mean of x mean of y
 169.6154  216.0000
```

Βλέπουμε ότι έχουμε p-value $= 0.0117$ για την μηδενική υπόθεση ότι οι μέσες είναι ίσες και άρα την απορρίπτουμε. Δηλαδή δεχόμαστε την εναλλακτική υπόθεση η οποία εκτυπώνεται και παραπάνω.

Για δεύτερο μέρος του ερωτήματος θεωρούμε ως μηδενική υπόθεση ότι η μέση αύξηση των καρδιακών χτύπων είναι μεγαλύτερη για την ομάδα 2 από την ομάδα 1. Αυτό το προσδιορίζουμε στην R με το να αναφέρουμε την εναλλακτική υπόθεση σαν όρισμα alternative="greater".

Με το λογισμικό R για την εντολή t.test(x,y,alternative="greater") έχουμε ότι η μηδενική υπόθεση είναι ότι $x \leq y$ (για τις μέσες τιμές των δειγμάτων) και η εναλλακτική είναι $x > y$.

Άρα χρησιμοποιούμε την εντολή:

```
t.test(g1diff,g2diff,var.equal=FALSE,conf.level = 0.95,
alternative="greater")
```

και παίρνουμε αρκετά υψηλή p-value. Άρα δεχόμαστε την μηδενική μας υπόθεση.

Μπορούμε φυσικά να θεωρήσουμε τις υποθέσεις αντίστροφα, προσδιορίζοντας alternative="less" και να πάρουμε πάρα πολύ μικρό p-value για να μην απορρίψουμε την νέα μηδενική υπόθεση, δηλαδή την προηγούμενη εναλλακτική.

```
> t.test(g1diff,g2diff,var.equal=FALSE,conf.level = 0.95,alternative="less")

        Welch Two Sample t-test

data:  g1diff and g2diff
t = -2.6017, df = 59.153, p-value = 0.005851
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -16.59314
sample estimates:
mean of x mean of y
 169.6154  216.0000

> t.test(g1diff,g2diff,var.equal=FALSE,conf.level = 0.95,alternative="greater")

        Welch Two Sample t-test

data:  g1diff and g2diff
t = -2.6017, df = 59.153, p-value = 0.9941
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -76.17609       Inf
sample estimates:
mean of x mean of y
 169.6154  216.0000
```

(5) Επαναλαμβάνουμε τους ελέγχους εφόσον έχουμε εφαρμόσει τον μετασχηματισμό:

```
> shapiro.test(logx)

        Shapiro-Wilk normality test

data:  logx
W = 0.98761, p-value = 0.7631

> shapiro.test(logy)

        Shapiro-Wilk normality test

data:  logy
W = 0.97798, p-value = 0.7081

>
>
> t.test(logx,logy,var.equal=FALSE,conf.level = 0.95)

        Welch Two Sample t-test

data:  logx and logy
t = -1.5451, df = 69.416, p-value = 0.1269
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.06169262  0.00783627
sample estimates:
mean of x mean of y
0.1962928 0.2232210

>
> t.test(logx,logy,var.equal=FALSE,conf.level = 0.95,alternative="greater")

        Welch Two Sample t-test

data:  logx and logy
t = -1.5451, df = 69.416, p-value = 0.9366
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.05598265         Inf
sample estimates:
mean of x mean of y
0.1962928 0.2232210
```

Καθώς η εργασία αναφέρεται σε καρδιακούς χτύπους και δεν υπάρχει διπλάσια αύξηση της περιόδου από μέρα σε νύχτα, έχουμε $1 < \frac{night}{day} < 2 < e$ και έτσι ο μετασχηματισμός του λογαρίθμου μας μεταφέρει τις μετρήσεις στο $(0, 1)$. Μάλιστα, σε αυτήν την περίπτωση πιο κοντά στο 0. Υπάρχει και άτομο που η εγγραφή του στο g1diff να έχει αρνητική τιμή και έτσι εφόσον $\frac{night}{day} < 1$, ο μετασχηματισμός θα δώσει αρνητική τιμή. Φυσικά, δεν μας αφορούν τέτοιες μεμονωμένες παρατηρήσεις, καθώς μπορεί να έχει συμβεί λάθος στην καταχώρηση των δεδομένων ή οτιδήποτε άλλο.

Οπότε μπορούμε να θεωρήσουμε ότι έχουμε μετρήσεις στο $(0, 1)$ τις οποίες έτσι μπορούμε να χειριστούμε καλύτερα. Τα test δίνουν όμοια αποτελέσματα εκτός από την περίπτωση του t-test για το αν οι μέσες τιμές ταυτίζονται. Καθώς με αυτόν τον μετασχηματισμό έχουμε τις παρατηρήσεις μας στο $(0, 1)$, οι μέσες τιμές θα είναι αρκετά κοντά. Επειδή αλλάζουμε την "κλίμακα" με αυτόν τον μετασχηματισμό, η μεγαλύτερη διαφορά θα παραμείνει μεγαλύτερη και έτσι θα εξάγουμε το ίδιο αποτέλεσμα από το δεύτερο t-test. Ωστόσο, στο πρώτο test καθώς οι παρατηρήσεις έχουν μεταφερθεί στο $(0, 1)$ η διαφορά των νέων μέσων τιμών θα είναι πολύ μικρή. Έτσι παίρνουμε p-value $= 0.1269$ το οποίο δεν είναι αρκετά μικρό για να απορρίψουμε την μηδενική υπόθεση, πράγμα που δεν συμβαίνει χωρίς τον μετασχηματισμό.

(6) Καθώς βολεύει να είναι οι παρατηρήσεις μας στο διάστημα $(0, 1)$, αν έχουμε για παράδειγμα δύο τυχαία δείγματα από λογαριθμικές κανονικές κατανομές, δηλαδή:

$$X \sim Lognormal(\mu_1, \sigma_1^2), \quad Y \sim Lognormal(\mu_2, \sigma_2^2)$$

ισοδύναμα

$$ln(X) \sim \mathcal{N}(\mu_1, \sigma_1^2), \quad ln(Y) \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

και χρησιμοποιήσουμε Welch t-test στα X, Y θα έχουμε:

$$H_0 : \quad \exp\left(\mu_1 + \frac{\sigma_1^2}{2}\right) = \exp\left(\mu_2 + \frac{\sigma_2^2}{2}\right)$$

ενάντια στην εναλλακτική H_1: να μην είναι ίσα.

Αν εφαρμόσουμε τον μετασχηματισμό και μετά κάνουμε το ίδιο t-test στα $ln(X), ln(Y)$ θα εξετάζουμε την μηδενική υπόθεση:

$$H_0 : \quad \mu_1 = \mu_2$$

η οποία δεν είναι ισοδύναμη με την προηγούμενη μηδενική υπόθεση. Επιπλέον, αν δεχτούμε την μηδενική υπόθεση για τα δείγματα $ln(X), ln(Y)$, δηλαδή $\mu_1 = \mu_2$ δεν είναι σωστό να συμπεράνουμε ότι και οι μέσες τιμές των X, Y θα είναι ίσες αφού εξαρτώνται από τις άγνωστες διασπορές. Χωρίς να γνωρίζουμε αν τα δεδομένα μας προέρχονται από λογαριθμική κανονική κατανομή, αυτό το φαινόμενο πράγματι το συναντήσαμε στο προηγούμενο ερώτημα. Με βάση αυτά τα συμπεράσματα μπορούμε να θεωρήσουμε ότι είναι δυσκολότερο να κάνουμε ελέγχους υποθέσεων με τον μετασχηματισμό του λογαρίθμου.

Κώδικας R που χρησιμοποιήθηκε:

```
install.packages("readxl")
library("readxl")
setwd("C:/Users/dimit/Desktop/Ergasia_Statistikis")
data1 <- read_excel("ergasia_12/ergasia_12.xls")
head(data1)


g1diff <- data1$night[data1$group==1] - data1$day[data1$group==1]
g2diff <- data1$night[data1$group==2] - data1$day[data1$group==2]

shapiro.test(g1diff)
shapiro.test(g2diff)

num <- 625
set.seed(num)

xsim <- rnorm(50)
ysim <- rt(50, df=3)
zsim <- rt(50, df=10)


shapiro.test(xsim)
shapiro.test(ysim)
shapiro.test(zsim)



t.test(g1diff,g2diff,var.equal=FALSE,conf.level = 0.95)

t.test(g1diff,g2diff,var.equal=FALSE,conf.level = 0.95,
alternative="greater")
t.test(g1diff,g2diff,var.equal=FALSE,conf.level = 0.95,
alternative="less")


logx <- log(data1$night[data1$group==1] /
data1$day[data1$group==1])

logy <- log(data1$night[data1$group==2] /
data1$day[data1$group==2])


shapiro.test(logx)
shapiro.test(logy)


t.test(logx,logy,var.equal=FALSE,conf.level = 0.95)

t.test(logx,logy,var.equal=FALSE,conf.level = 0.95,
alternative="greater")
```

Αναφορές

[1] Σάμης Τρέβεζας. *Μη Παραμετρική Στατιστική*. Σημειώσεις Διαλέξεων, Αθήνα, 2020.

[2] RStudio Team (2020). *RStudio: Integrated Development for R. RStudio*. PBC, Boston, MA, URL: http://www.rstudio.com/

[3] Changyong Feng, Hongyue Wang et al. . *Log-transformation and its implications for data analysis*. Shanghai Arch Psychiatry. (2014) 26:105–9. 10.3969/j.issn.1002-0829.2014.02.009 URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4120293/