Μεταπτυχιακός Φοιτητής : **ΜΠΕΓΕΤΗΣ ΝΙΚΟΛΑΟΣ**
Αριθμός Μητρώου : ΠΙΒ0111
Προθεσμία Υποβολής Απαντήσεων : 06/04/2014

ΑΣΚΗΣΗ 1η

1. Επιστήμονας που μελετά ένα συγκεκριμένο προστατευόμενο υπό εξαφάνιση είδος πτηνού στην Πάρνηθα παρατηρεί ότι το πλήθος των πτηνών αυτών έχει μεταβληθεί σε σχέση με πέρυσι. Την ίδια χρονική περίοδο (10/10-25/10) κατά το προηγούμενο έτος, η μη κανονική αλλά συμμετρική κατανομή του ημερήσιου πλήθους των πτηνών του είδους αυτού που παρατηρούσε στην Πάρνηθα είχε διάμεσο ίση με 9. Ο επιστήμονας αποφάσισε να ελέγξει εάν ισχύει όντως η παρατήρησή του ότι το πλήθος των πτηνών αυτών έχει μεταβληθεί σε σχέση με πέρυσι. Για το σκοπό αυτό μέτρησε το ημερήσιο πλήθος του είδους αυτού που παρατηρούσε στην Πάρνηθα. Τα αποτελέσματα των μετρήσεών του για τις 16 ημέρες στο διάστημα 10/10/2013-25/10/2013 καταγράφονται στον ακόλουθο πίνακα :

Ημερομηνία	**10/10/13**	**11/10/13**	**12/10/13**	**13/10/13**	**14/10/13**	**15/10/13**
Πλήθος πτηνών	7	9	12	14	7	7
Ημερομηνία	**16/10/13**	**17/10/13**	**18/10/13**	**19/10/13**	**20/10/13**	**21/10/13**
Πλήθος πτηνών	15	12	10	7	7	12
Ημερομηνία	**22/10/13**	**23/10/13**	**24/10/13**	**25/10/13**		
Πλήθος πτηνών	9	7	15	15		

Υπάρχει ένδειξη σε επίπεδο 5% ότι το πλήθος των πτηνών του συγκεκριμένου προστατευόμενου υπό εξαφάνιση είδους που παρατηρούνται στην Πάρνηθα έχει μεταβληθεί σε σχέση με πέρυσι;

Απάντηση:

Συμμετρική κατανομή όπως γνωρίζουμε είναι η κατανομή της οποίας οι τιμές της μεταβλητής διατάσσονται συμμετρικά γύρω από τη μέση τιμή. Από την εκφώνηση γνωρίζουμε ότι η διάμεσος median2012=9. Επίσης γνωρίζουμε ότι η διάμεσος, η μέση τιμή και η επικρατούσα τιμή συμπίπτουν μόνο όταν η κατανομή είναι συμμετρική και έχει μόνο μία κορυφή. Στην περίπτωσή μας η κατανομή είναι συμμετρική, οπότε μπορούμε να υποθέσουμε ότι αφού γνωρίζουμε τη διάμεσο, και η κατανομή είναι συμμετρική, τότε και η μέση τιμή mean2012=9. Για την επικρατούσα τιμή δεν μπορούμε να είμαστε σίγουροι γιατί δεν γνωρίζουμε τις κορυφές τις κατανομής και αν είναι ζυγές στον αριθμό τότε σίγουρα δεν υπάρχει μία επικρατούσα τιμή, ενώ αν είναι μονές τότε πάλι δεν μπορούμε να εξάγουμε με σιγουριά κάποιο συμπέρασμα.

Ημ/νία	**10**	**14**	**15**	**19**	**20**	**23**	**11**	**22**	**18**	**12**	**17**	**21**	**13**	**16**	**24**	**25**
Πλήθος πτηνών	7	7	7	7	7	7	9	9	10	12	12	12	14	15	15	15

Για να βρούμε τη διάμεσο στις φετινές παρατηρήσεις αρκεί να διατάξουμε τις μετρήσεις σε αύξουσα σειρά και να εκλέξουμε το μέσο. Και επειδή στην περίπτωσή μας έχουμε 16 μετρήσεις, η διάμεσος θα βρεθεί από το μέσο όρο μεταξύ της $8^{ης}$ και $9^{ης}$ θέσης

Από το παραπάνω πινακάκι φαίνεται ότι η διάμεσος για τη φετινή χρονιά είναι ο μέσος όρος του πλήθους των πτηνών που παρατηρήθηκαν την 22/10/2013 και 18/10/2013 και αυτή είναι median2013=9.5. Χρησιμοποιώντας αυτές τις μετρήσεις για επιβεβαίωση σε κώδικα MATLAB και για να βρούμε και τη μέση τιμή για τη φετινή χρονιά έχουμε:

```
function ass1_1()
    x2013=[7 9 12 14 7 7 15 12 10 7 7 12 9 7 15 15];
    mean_x = mean(x2013) % mean_x holds the mean of all x values
    median_x = median(x2013) % median_x holds the meadian of all x
values
end
```

και τα αποτελέσματα ήταν:

```
mean_x = 10.3125
median_x = 9.5000
```

Οπότε συνολικά έχουμε:

Mean2012=9.5

Mean2013=10.3125

Median2012=9.5

Median2013=9.5

Το 2013 δεν είναι συμμετρική η κατανομή αλλά αφού γνωρίζουμε τα δεδομένα της μπορούμε να βρούμε την τυπική της απόκλιση. Ακολούθως, γνωρίζοντας ότι ο συντελεστής μεταβλητότητας εξαρτάται από το μέγεθος της τυπικής απόκλισης και της μέσης τιμής (std(x)/mean(x)), παρατηρούμε ότι ο συντελεστής μεταβλητότητας υπολογίζεται ως εξής σε κώδικα MATLAB:

```
function ass1_1()
    x2013=[7 9 12 14 7 7 15 12 10 7 7 12 9 7 15 15];
    mean_x = mean(x2013) % mean_x holds the mean of all x values
    median_x = median(x2013) % median_x holds the meadian of all x
values
    dispFactor_x = (std(x2013)/mean_x) % dispFactor_x holds the
dispersion factor of all x values
end
```

Τα αποτελέσματα του παραπάνω κώδικα είναι:

```
mean_x = 10.3125
median_x = 9.5000
dispFactor_x = 0.3142
```

2. Ενδοκρινολόγος εφάρμοσε ειδικό πρόγραμμα διατροφής με σκοπό την απώλεια βάρους σε οκτώ εξαιρετικά παχύσαρκους ασθενείς του. Ο πίνακας που ακολουθεί καταγράφει το βάρος των οκτώ ασθενών πριν την εφαρμογή του προγράμματος και μετά από ένα χρόνο εφαρμογής του :

α/α ασθενούς	**Βάρος πριν την εφαρμογή του προγράμματος διατροφής (Kg)**	**Βάρος μετά την εφαρμογή του προγράμματος διατροφής (Kg)**
1	185	170
2	110	114
3	100	90
4	150	130
5	170	177
6	125	122
7	160	140
8	110	110

α) Με τη βοήθεια του διαγράμματος κανονικότητας, να επιβεβαιωθεί η κανονικότητα του βάρους.

β) Να εξεταστεί η αποτελεσματικότητα του προγράμματος διατροφής. Παρουσιάστε όλα τα βήματα του ελέγχου.

Απάντηση:

(α)

Το **διάγραμμα κανονικότητας** είναι ένα χρήσιμο γράφημα για να εκτιμάται εάν τα δεδομένα προέρχονται από κανονική κατανομή.

Όπως έχουμε δει, πολλές στατιστικές διαδικασίες στηρίζονται στην παραδοχή ότι η κατανομή δεδομένων είναι κανονική. Το διάγραμμα κανονικότητας μπορεί να παρέχει κάποια διαβεβαίωση ότι η παραδοχή της κανονικότητας δεν παραβιάζεται ή να παρέχει έγκαιρη προειδοποίηση τυχόν απόκλισης από την παραδοχή της κανονικότητας.

Αυτό το καταλαβαίνουμε εαν όλα τα σημεία των δεδομένων βρίσκονται κοντά στην ευθεία γραμμή και δεν ακολουθούν καμπύλη γραμμή που αποκλίνει από την ευθεία γραμμή.

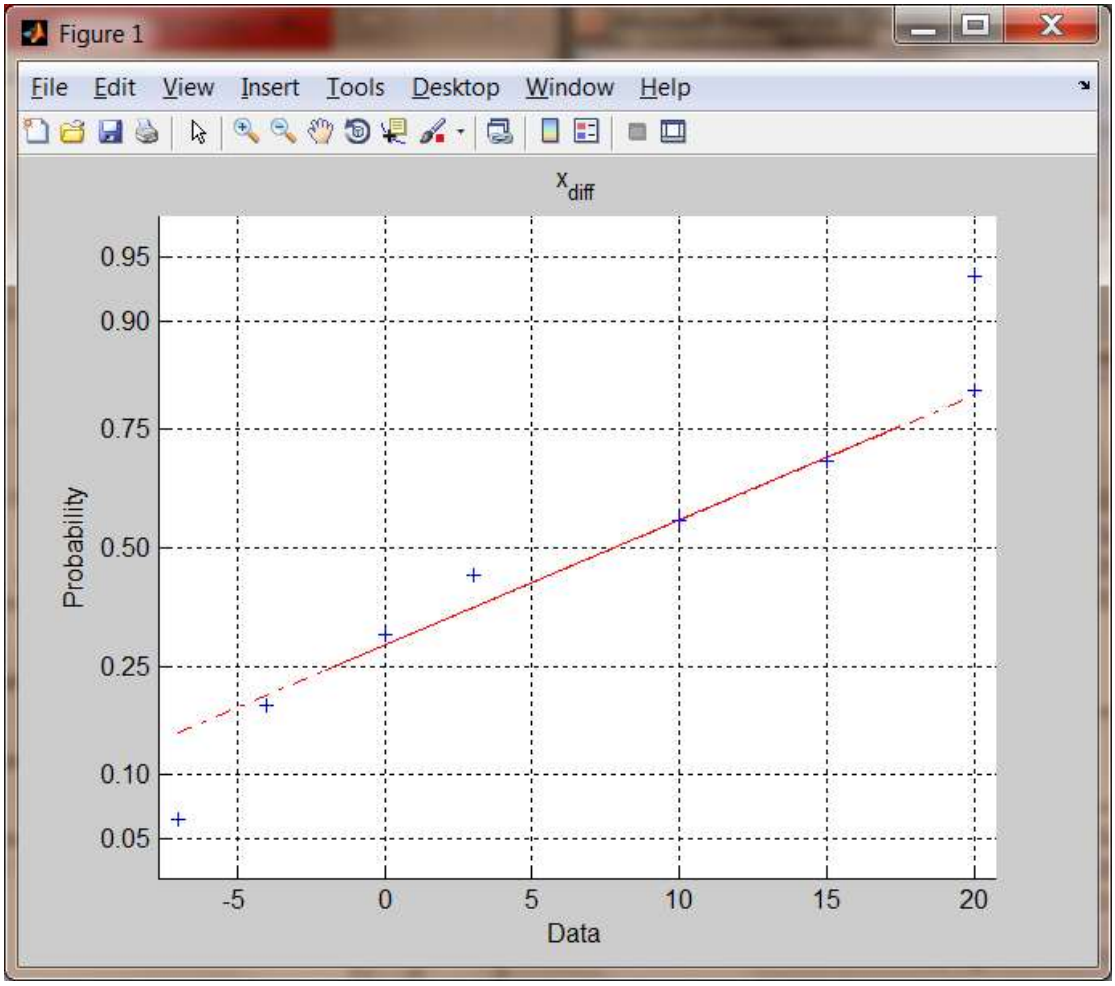
Η εντολή Matlab που παράγει το διάγραμμα κανονικότητας είναι η **normplot(x)** και άρα ο κώδικας MATLAB για τα παραπάνω δεδομένα βάρους των οκτώ ασθενών είναι ο εξής:

```
function ass1_2()

% question 2a
    x_before=[185 110 100 150 170 125 160 110];
%    normplot(x_before)    % uncomment to see the x_before plot
%    title('x_{before}')
    x_after=[170 114 90 130 177 122 140 110];
%    figure
%    normplot(x_after)    % uncomment to see the x_after plot
%    title('x_{after}')

    x_diff=x_before-x_after
%    figure
    normplot(x_diff)
    title('x_{diff}')
end
```

Και το σχήμα που προκύπτει από τον παραπάνω κώδικα είναι το εξής:



Όπως φαίνεται από τα παραπάνω δύο διαγράμματα το διάγραμμα κανονικότητας περιλαμβάνει τρία γραφικά στοιχεία:

- ✓ Τα "+" δείχνουν την εμπειρική πιθανότητα ως προς την τιμή των δεδομένων κάθε σημείου στο δείγμα.
- ✓ Η συμπαγής γραμμή συνδέει το 25ο και το 75ο εκατοστημόριο των δεδομένων και αναπαριστά μία γραμμική παλινδρόμηση (μη ευαίσθητη στα ακραία σημεία του δείγματος).
- ✓ Η διακεκομμένη γραμμή επεκτείνει τη συμπαγή γραμμή στα άκρα του δείγματος

Επίσης, η κλίμακα του y άξονα δεν είναι ομογενή και οι τιμές του y άξονα είναι πιθανότητες και έτσι κυμαίνονται μεταξύ 0 και 1.

Επίσης στο συγκεκριμένο παράδειγμα δεν μπορούμε να διακρίνουμε ότι η απόσταση μεταξύ των σημείων στον άξονα y ταιριάζει με την απόσταση μεταξύ των εκατοστημορίων της κανονικής κατανομής γιατί έχουμε λίγα δεδομένα, αλλά γενικά ισχύει ότι τα εκατοστημόρια πλησιάζουν μεταξύ τους κοντά στη διάμεσο (πιθανότητα = 0.5) και «χαλαρώνουν» συμμετρικά καθώς απομακρυνόμαστε από τη διάμεσο.

Συνεπώς για να απαντήσουμε και στο ερώτημα της άσκησης, μέσα από τα διαγράμματα **η κανονικότητα βάρους μπορεί να επιβεβαιωθεί** επειδή όλα τα σημεία των δεδομένων βρίσκονται κοντά στην ευθεία γραμμή και τα "+" μπορεί ακολουθούν την κατεύθυνση της ευθείας και δεν αποκλίνουν σε καμπύλη σε καμία από τις 2 περιπτώσεις (πριν και μετά την εφαρμογή του προγράμματος διατροφής)

(β) Να εξεταστεί η αποτελεσματικότητα του προγράμματος διατροφής. Παρουσιάστε όλα τα βήματα του ελέγχου

Έχοντας επιβεβαιώσει την κανονικότητα του βάρους των ασθενών βρίσκοντας τη διαφορά μεταξύ των μετρήσεων πριν και μετά την εφαρμογή του προγράμματος διατροφής εφαρμόζουμε το κατά ζεύγη t-τεστ με μηδενική υπόθεση ότι «τα δύο δείγματα x και y προέρχονται από κατανομές με ίσες μέσες τιμές».

Αυτό μπορούμε να το κάνουμε εφαρμόζοντας το Lilliefors τεστ.

Η μηδενική υπόθεση είναι ότι «η διαφορά x-y ακολουθεί κανονική κατανομή με απροσδιόριστη μέση τιμή και διασπορά». Η εναλλακτική υπόθεση εκφράζει το αντίθετο.

Ο έλεγχος υλοποιείται με τις ακόλουθες εντολές MATLAB:

```
function ass1_2()

% question 2a
    x_before=[185 110 100 150 170 125 160 110];
    x_after=[170 114 90 130 177 122 140 110];
    x_diff=x_before-x_after
    normplot(x_diff)
    title('x_{diff}')

% question 2b
    [h,p,lstat,cv] = lillietest(x_diff)
End
```

Τα αποτελέσματα που επεστράφησαν από τον παραπάνω κώδικα είναι:

h = 0

p = 0.5000

lstat = 0.1508

cv = 0.2880

Συνεπώς, η μηδενική υπόθεση κανονικότητας δεν μπορεί να απορριφθεί, γιατί η υπόθεση h=0.

Για να ελέγξουμε την αποτελεσματικότητα του προγράμματος διατροφής εφαρμόζουμε την One-way ANOVA στα δεδομένα, αφού πρώτα επιβεβαιώσαμε την κανονικότητα του βάρους. Έτσι έχουμε:

Μηδενική υπόθεση: Και οι δύο πληθυσμοί έχουν τις ίδιες μέσες τιμές του μετρούμενου μεγέθους

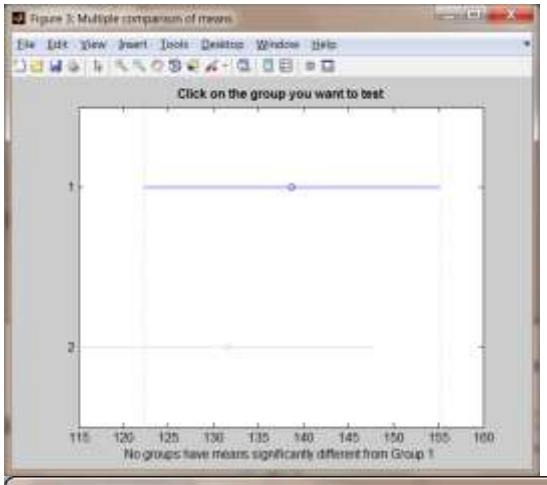
Εναλλακτική υπόθεση: Ο ένας πληθυσμός έχει μέση τιμή του μετρούμενου μεγέθους διαφορετική από τον άλλο.

Ο κώδικας για τον παραπάνω έλεγχο υπόθεση είναι:

```
% question 2b
    [h,p,lstat,cv] = lillietest(x_diff)

    x_comparisons=[
        185 170;
        110 114;
        100 90;
        150 130;
        170 177;
        125 122;
        160 140;
        110 110];
    [p,table,stat]=anova1(x_comparisons);
    c=multcompare(stat)
```

Και τα αποτελέσματα που παράχθηκα είναι τα εξής:



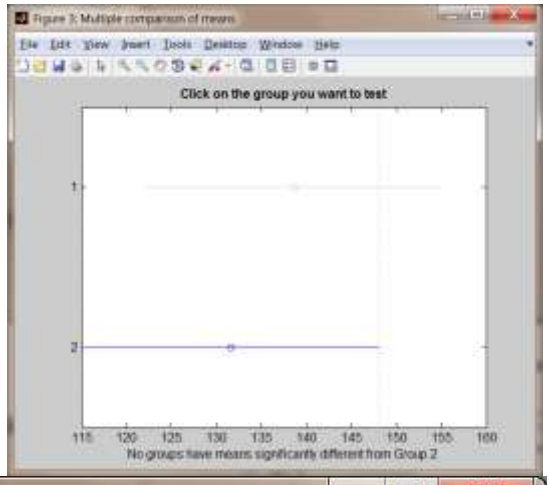


Figure 2: One-way ANOVA

File Edit View Insert Tools Desktop Window Help

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	203.1	1	203.063	0.21	0.6503
Error	13245.4	14	946.098		
Total	13448.4	15			

Επομένως, η μηδενική υπόθεση επιβεβαιώνεται. Η μέση τιμή της διαφοράς των μέσων διαφορετικών βαρών για κάθε ασθενή είναι παρόμοια και στα δύο δείγματα.

Άρα, η αποτελεσματικότητα του προγράμματος διατροφής είναι αρνητική. Το πρόγραμμα διατροφής δεν λειτούργησε όπως θα έπρεπε.

3. Υποθέτοντας ότι οι παρατηρήσεις του ακόλουθου πίνακα είναι ανεξάρτητες και ότι οι γραμμές του αποτελούν δείγματα από πληθυσμούς με ίσες διασπορές, να εξεταστεί εάν όλοι οι πληθυσμοί έχουν την ίδια μέση τιμή :

0.1878	-0.1859	1.1286	1.8057	0.6694	0.8801	1.4694	2.0184
-1.2023	-0.0559	1.6565	1.2316	0.1564	0.9347	0.0964	-0.5804
1.9863	2.4725	-0.1678	0.0102	1.4978	1.4853	1.0359	0.9213
0.4814	1.0557	0.5394	2.3396	2.4885	0.4045	0.3725	0.3183
1.3274	-0.2173	0.7376	1.2895	0.4535	0.8503	1.5354	-0.0246
1.2341	0.9588	-0.2132	2.4789	0.1532	0.5652	1.5529	-0.2344
1.0215	-0.1283	-0.3194	2.1380	0.7537	0.9207	0.7963	1.2888
-0.0039	-0.3493	1.9312	0.3159	1.6630	2.5352	1.0543	0.5707
0.0529	0.7389	1.0112	-0.2919	0.1458	0.3935	1.1326	1.0558

Στην περίπτωση που οι πληθυσμοί δεν έχουν την ίδια μέση τιμή, να βρεθεί ποιοι πληθυσμοί διαφέρουν από τους υπόλοιπους.

Απάντηση:

Group1	0.1878	-0.1859	1.1286	1.8057	0.6694	0.8801	1.4694	2.0184
Group2	-1.2023	-0.0559	1.6565	1.2316	0.1564	0.9347	0.0964	-0.5804
Group3	1.9863	2.4725	-0.1678	0.0102	1.4978	1.4853	1.0359	0.9213
Group4	0.4814	1.0557	0.5394	2.3396	2.4885	0.4045	0.3725	0.3183
Group5	1.3274	-0.2173	0.7376	1.2895	0.4535	0.8503	1.5354	-0.0246
Group6	1.2341	0.9588	-0.2132	2.4789	0.1532	0.5652	1.5529	-0.2344
Group7	1.0215	-0.1283	-0.3194	2.1380	0.7537	0.9207	0.7963	1.2888
Group8	-0.0039	-0.3493	1.9312	0.3159	1.6630	2.5352	1.0543	0.5707
Group9	0.0529	0.7389	1.0112	-0.2919	0.1458	0.3935	1.1326	1.0558

Καταρχήν ελέγχουμε την υπόθεση της ισότητας των διασπορών των εννέα πληθυσμών με τη συνάρτηση vartestn:

```
function ass1_3()
% question 3
    x=[
        0.1878   -0.1859    1.1286  1.8057  0.6694  0.8801  1.4694
2.0184;
        -1.2023 -0.0559 1.6565  1.2316  0.1564  0.9347  0.0964  -0.5804;
        1.9863  2.4725  -0.1678 0.0102  1.4978  1.4853  1.0359  0.9213;
        0.4814  1.0557  0.5394  2.3396  2.4885  0.4045  0.3725  0.3183;
        1.3274  -0.2173 0.7376  1.2895  0.4535  0.8503  1.5354  -0.0246;
        1.2341  0.9588  -0.2132 2.4789  0.1532  0.5652  1.5529  -0.2344;
        1.0215  -0.1283 -0.3194 2.1380  0.7537  0.9207  0.7963  1.2888;
        -0.0039 -0.3493 1.9312  0.3159  1.6630  2.5352  1.0543  0.5707;
        0.0529  0.7389  1.0112  -0.2919 0.1458  0.3935  1.1326  1.0558;
    ];
% we make the computations in the reverted matrix for the 9 groups of
% population
    [p,stats] = vartestn(x') % variance equality
end
```

Τα αποτελέσματα που παρήγαγε η υπόθεση ισότητας με τη συνάρτηση vartestn είναι:

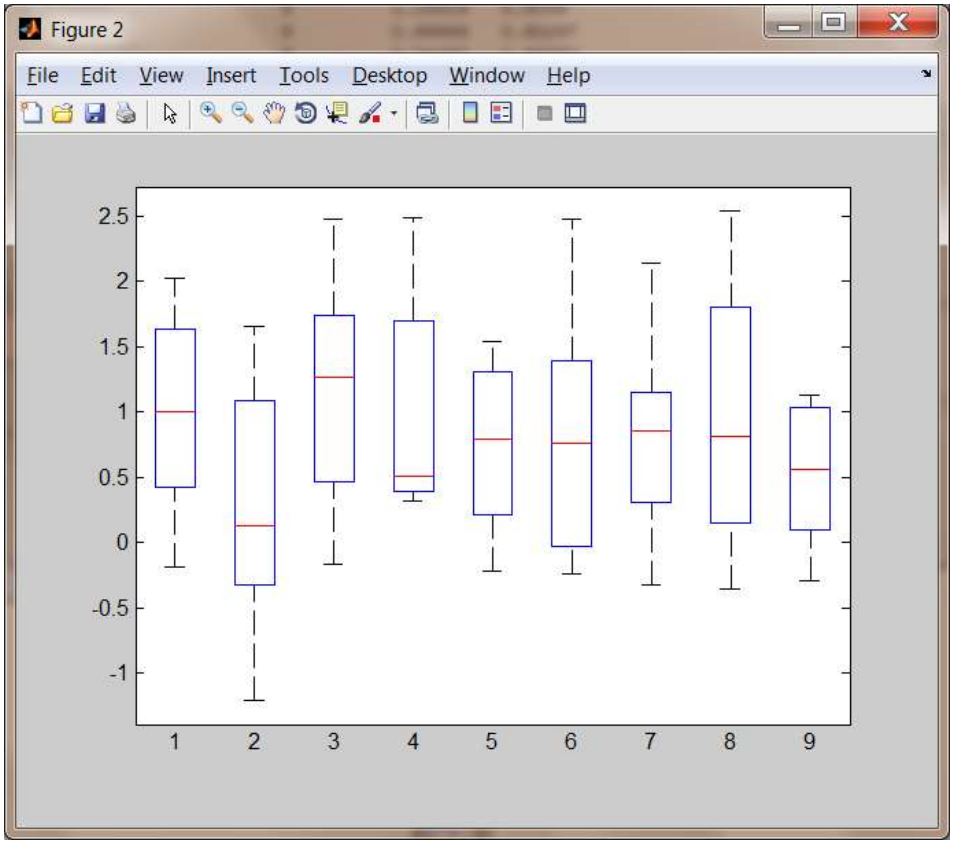


Figure 1: Variance Test

File Edit View Insert Tools Desktop Window Help

Group Summary Table

Group	Count	Mean	Std Dev
1	8	0.99669	0.76651
2	8	0.27963	0.95081
3	8	1.15519	0.9089
4	8	0.99999	0.90297
5	8	0.74397	0.64085
6	8	0.81194	0.9387
7	8	0.80891	0.77544
8	8	0.96464	1.00975
9	8	0.52985	0.5324
Pooled	72	0.81009	0.8386

Bartlett's statistic	4.16081
Degrees of freedom	8
p-value	0.84233

Από τα παραπάνω διαγράμματα p=0.84233 και επειδή p>a=0.05, δεν μπορεί να απορριφθεί η μηδενική υπόθεση. Άρα αποδέχεται την ισότητα διασπορών και επιβεβαιώσαμε ότι οι πληθυσμοί έχουν ίσες διασπορές.

Για την εύρεση αν τα 9 groups πληθυσμών έχουν την ίδια μέση τιμή έχουμε ως εξής:

Εφαρμόζουμε την one-way ANOVA ([p,table,stats] = anova1(X)):

- **Μηδενική Υπόθεση:** Όλα τα δείγματα προέρχονται από πληθυσμό με ίδια μέση τιμή.
- **Εναλλακτική Υπόθεση:** Τουλάχιστον ένα δείγμα προέρχεται από πληθυσμό με διαφορετική μέση τιμή.

Προσθέτουμε στον παραπάνω κώδικα την one-way ANOVA και έχουμε:

```
function ass1_3()
% question 3
    x=[
        0.1878   -0.1859    1.1286  1.8057  0.6694  0.8801  1.4694
2.0184;
        -1.2023 -0.0559 1.6565  1.2316  0.1564  0.9347  0.0964  -0.5804;
        1.9863  2.4725  -0.1678 0.0102  1.4978  1.4853  1.0359  0.9213;
        0.4814  1.0557  0.5394  2.3396  2.4885  0.4045  0.3725  0.3183;
        1.3274  -0.2173 0.7376  1.2895  0.4535  0.8503  1.5354  -0.0246;
        1.2341  0.9588  -0.2132 2.4789  0.1532  0.5652  1.5529  -0.2344;
        1.0215  -0.1283 -0.3194 2.1380  0.7537  0.9207  0.7963  1.2888;
        -0.0039 -0.3493 1.9312  0.3159  1.6630  2.5352  1.0543  0.5707;
        0.0529  0.7389  1.0112  -0.2919 0.1458  0.3935  1.1326  1.0558;
    ];
% we make the computations in the reverted matrix for the 9 groups of
% population
    [p,stats] = vartestn(x') % variance equality
    figure;
    [p,table,stats] = anova1(x')% mean equality
End
```

Τα νέα αποτελέσματα που μας επέστρεψε ο παραπάνω κώδικας είναι τα εξής:

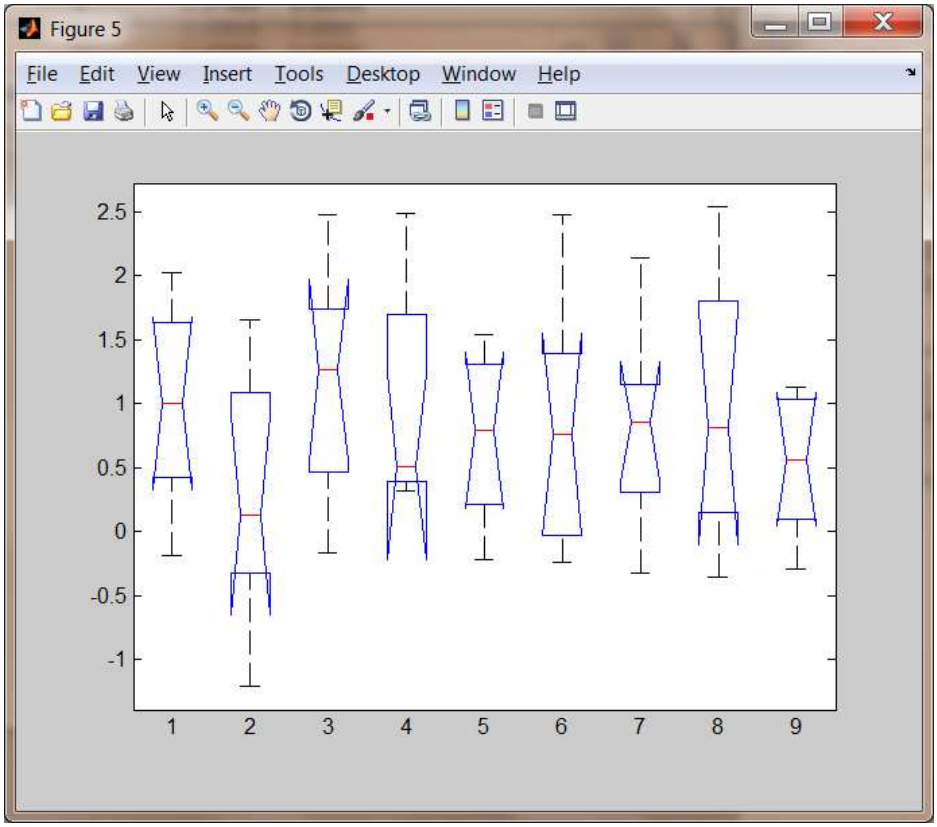


Figure 4: One-way ANOVA

File Edit View Insert Tools Desktop Window Help

ANOVA Table

Source	SS	df	MS	F	Prob>F
Columns	4.6253	8	0.57816	0.82	0.5861
Error	44.3045	63	0.70325		
Total	48.9298	71			

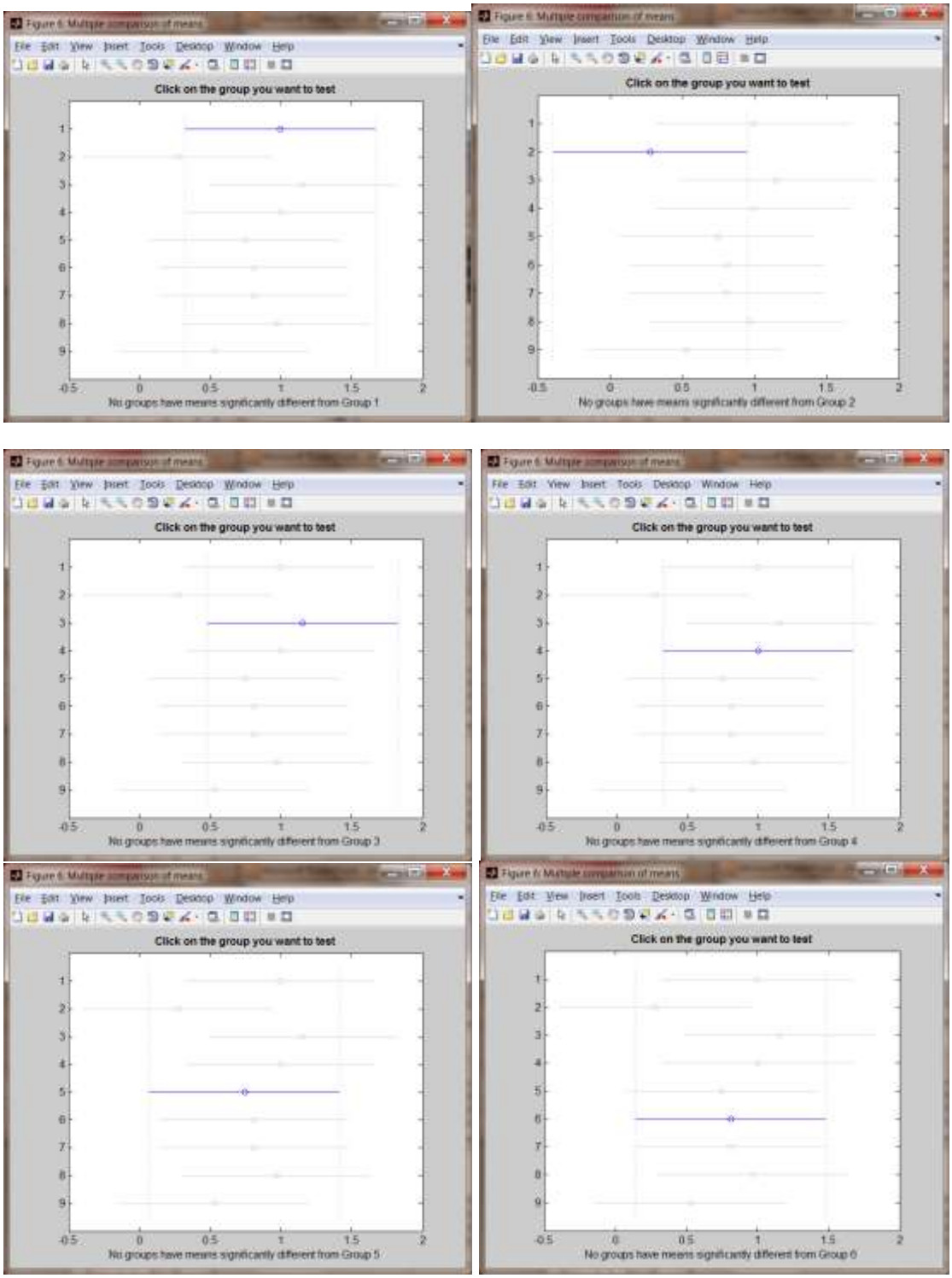
Επειδή p=0.5861>a=0.05, **δεν μπορεί να απορριφθεί η μηδενική υπόθεση. Άρα αποδέχεται την ισότητα στις μέσες τιμές και των 9 πληθυσμών**.

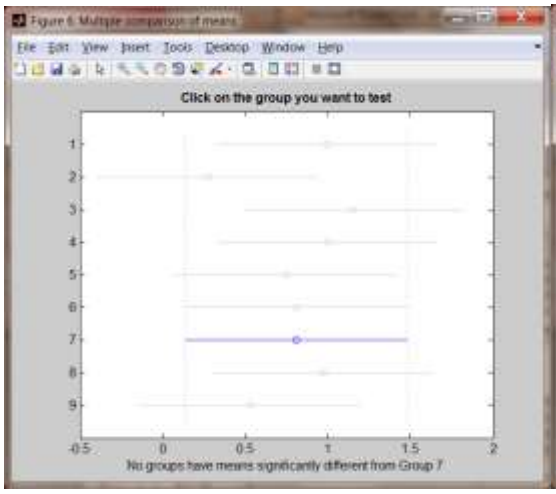
Τέλος για να **επιβεβαιώσουμε** ότι όλα τα δείγματα έχουν τις **ίδιες μέσες τιμές**, χρησιμοποιώντας την εντολή MATLAB **c=multcompare(stats)** λαμβάνουμε τα ακόλουθα αποτελέσματα:

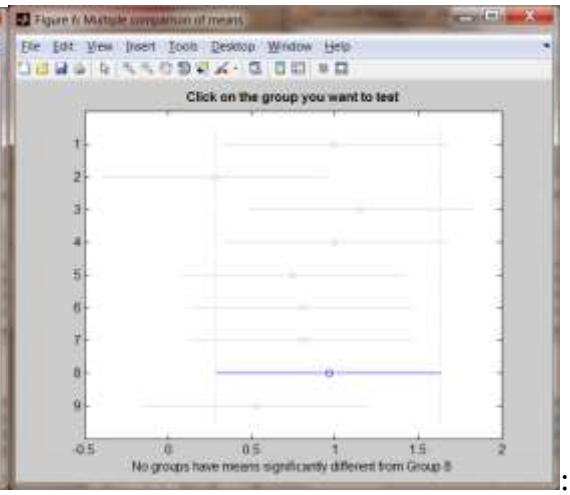
Προσθέτουμε στον παραπάνω κώδικα την **c=multcompare(stats)** και έχουμε:

```
function ass1_3()
% question 3
    x=[
        0.1878   -0.1859    1.1286  1.8057  0.6694  0.8801  1.4694
2.0184;
        -1.2023 -0.0559 1.6565  1.2316  0.1564  0.9347  0.0964  -0.5804;
        1.9863  2.4725  -0.1678 0.0102  1.4978  1.4853  1.0359  0.9213;
        0.4814  1.0557  0.5394  2.3396  2.4885  0.4045  0.3725  0.3183;
        1.3274  -0.2173 0.7376  1.2895  0.4535  0.8503  1.5354  -0.0246;
        1.2341  0.9588  -0.2132 2.4789  0.1532  0.5652  1.5529  -0.2344;
        1.0215  -0.1283 -0.3194 2.1380  0.7537  0.9207  0.7963  1.2888;
        -0.0039 -0.3493 1.9312  0.3159  1.6630  2.5352  1.0543  0.5707;
        0.0529  0.7389  1.0112  -0.2919 0.1458  0.3935  1.1326  1.0558;
    ];
% we make the computations in the reverted matrix for the 9 groups of
% population
    [p,stats] = vartestn(x') % variance equality
    figure;
    [p,table,stats] = anova1(x')% mean equality
    figure;
    c=multcompare(stats)
end
```

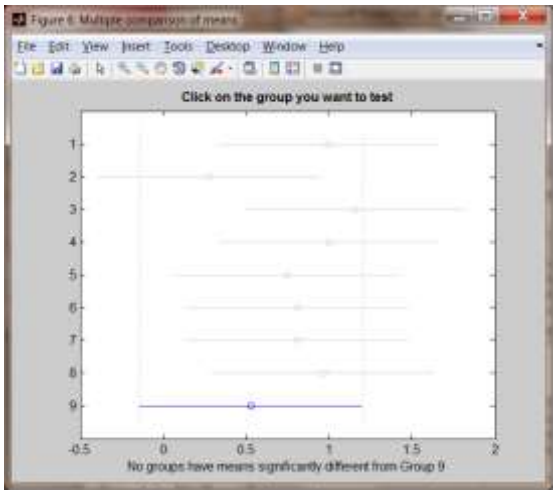
Τα νέα αποτελέσματα που μας επέστρεψε ο παραπάνω κώδικας είναι τα εξής εννέα:







:



Παρατηρούμε ότι όλες οι μέσες τιμές μοιάζουν μεταξύ τους και έχουν όλες χρώμα μπλε (η επιλεγμένη) και γκρι (όλες οι υπόλοιπες) για κάθε μία από τις 9 ομάδες πληθυσμών, ενώ δεν έχει καμία κόκκινο χρώμα**, το οποίο μας επαληθεύει την υπόθεση** και με οπτικό τρόπο.

4. Στον πίνακα που ακολουθεί καταγράφεται το ύψος ενός φυτού σε μέτρα και η γεωγραφική του θέση στον Ελλαδικό χώρο όπου παρατηρήθηκε.

Ύψος Φυτού	**Γεωγραφική θέση**	**Ύψος Φυτού**	**Γεωγραφική θέση**
3.33	Βόρεια Ελλάδα	0.84	Βόρεια Ελλάδα
5.87	Νότια Ελλάδα	3.95	Βόρεια Ελλάδα
6.31	Νότια Ελλάδα	4.16	Βόρεια Ελλάδα
1.98	Νότια Ελλάδα	2.99	Βόρεια Ελλάδα
4.86	Βόρεια Ελλάδα	2.15	Βόρεια Ελλάδα
2.65	Νότια Ελλάδα	6.78	Νότια Ελλάδα
2.78	Βόρεια Ελλάδα	2.84	Βόρεια Ελλάδα
2.21	Βόρεια Ελλάδα	4.23	Νότια Ελλάδα
0.45	Νότια Ελλάδα	2.89	Βόρεια Ελλάδα
1.51	Νότια Ελλάδα	2.56	Βόρεια Ελλάδα
0.56	Νότια Ελλάδα	2.65	Νότια Ελλάδα
3.68	Βόρεια Ελλάδα	1.54	Νότια Ελλάδα
2.16	Βόρεια Ελλάδα	0.98	Νότια Ελλάδα
3.15	Βόρεια Ελλάδα	0.87	Νότια Ελλάδα

Ερευνητής υποστηρίζει ότι το μέσο ύψος του φυτού είναι μεγαλύτερο στη Βόρεια Ελλάδα από ό,τι στη Νότια Ελλάδα ενώ η διασπορά του ύψος του φυτού είναι μικρότερη στη Βόρεια Ελλάδα από ό,τι στη Νότια Ελλάδα. Δικαιολογούνται οι ισχυρισμοί του ερευνητή από τα ανωτέρω πειραματικά δεδομένα;

Απάντηση:

(α)

Αρχικά επειδή ο πληθυσμός των δειγμάτων είναι διαφορετικός και επειδή πρέπει να επιβεβαιώσουμε ότι και τα 2 δείγματα έχουν την ίδια κατανομή ώστε να μπορούμε να θεωρήσουμε ότι έχουν αυτό ως κοινό τους στοιχείο και να πατήσουμε σε αυτό για να έχει νόημα ο ισχυρισμός του ερευνητή.

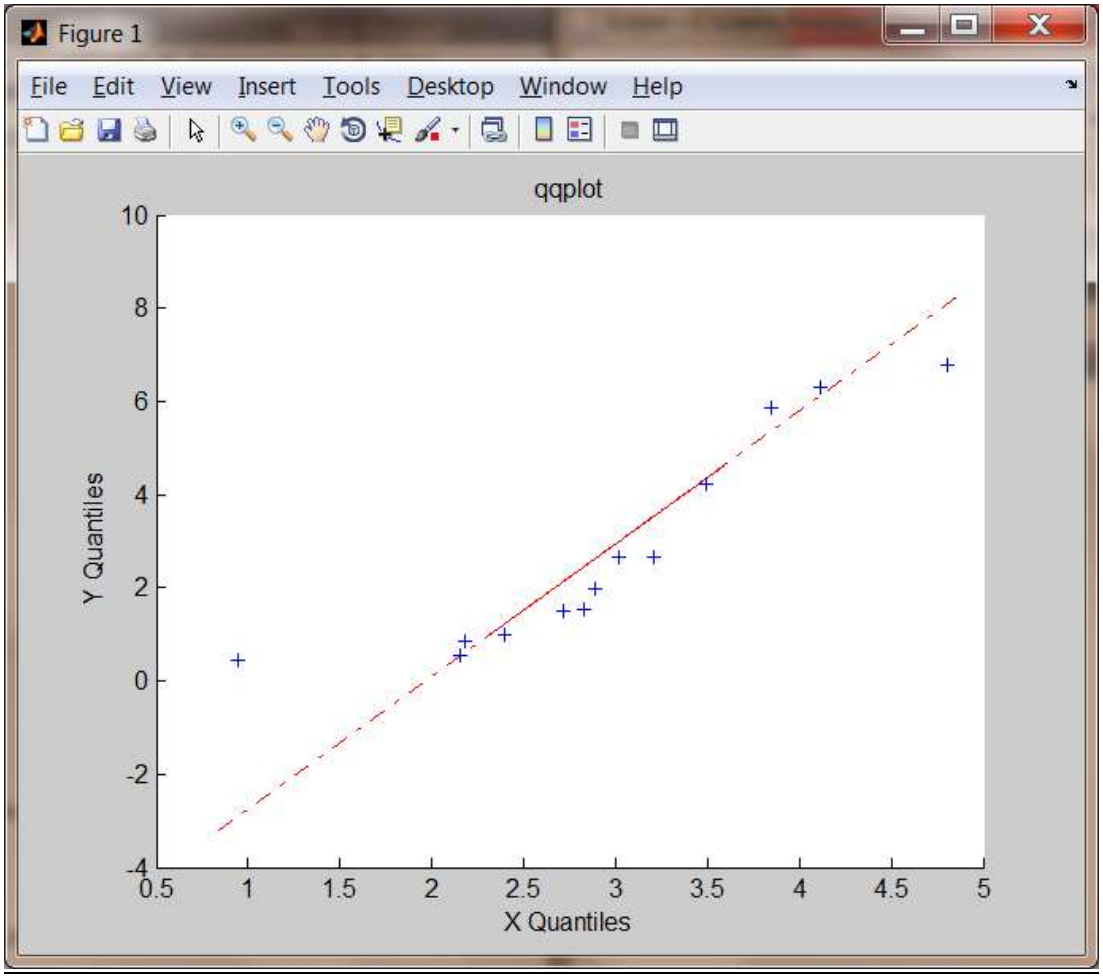
Αυτό θα το κάνουμε με τη συνάρτηση Q-Q plot του MatLab

Άρα ο κώδικας MATLAB είναι ο εξής:

```
function ass1_4()
    x_north=[3.33 0.84 3.95 4.16 2.99 4.86 2.15 2.78 2.84 2.21 2.89
2.56 3.68 2.16 3.15];
    x_south=[5.87 6.31 1.98 2.65 6.78 4.23 0.45 1.51 0.56 2.65 1.54
0.98 0.87];

    qqplot(x_north,x_south)
    title('qqplot')
end
```

Και τα αποτελέσματα του παραπάνω κώδικά είναι τα εξής:



Όπως το διάγραμμα κανονικότητας, έτσι και το Q-Q διάγραμμα έχει τρία γραφικά στοιχεία:

- Τα "+" είναι τα εκατοστημόρια κάθε δείγματος. Εξορισμού το πλήθος των "+" είναι το πλήθος των δεδομένων του μικρότερου δείγματος.
- Η συμπαγής γραμμή συνδέει το 25ο και το 75ο εκατοστημόριο των δεδομένων.
- Η διακεκομμένη γραμμή επεκτείνει τη συμπαγή γραμμή στα άκρα του δείγματος.

Το διάγραμμα παρουσιάζει **γραμμική σχέση** εάν τα δύο δείγματα προέρχονται από την ίδια κατανομή. Πράγματι έτσι είναι. Όπως φαίνεται και από το διάγραμμα που επέστρεψε ο παραπάνω κώδικας. Άρα και τα δύο δείγματα προέρχονται από την ίδια κατανομή.

Για επιβεβαίωση κάνουμε και άλλο ένα τεστ, το Kolmogorov-Smirnov. Το Kolmogorov-Smirnov τεστ για τη σύγκριση των κατανομών δύο δειγμάτων

- Ελέγχει τη **μηδενική υπόθεση** ότι «τα υπό εξέταση δύο δείγματα προέρχονται από την ίδια κατανομή» έναντι της **εναλλακτικής υπόθεσης** ότι «τα δύο υπό εξέταση δείγματα δεν προέρχονται από την ίδια κατανομή».

- Συγκρίνει την αναλογία των τιμών του πρώτου δείγματος που είναι μικρότερα από x με την αναλογία τιμών του δεύτερου δείγματος που είναι μικρότερα από x.

- Η **Kolmogorov-Smirnov στατιστική** είναι η **μέγιστη διαφορά** ως προς όλες τις x τιμές και ακολουθεί την Kolmogorov-Smirnov κατανομή.

Οπότε η συνάρτηση MATLAB είναι η **kstest2** και ο κώδικας ο εξής:

```
function ass1_4()
    x_north=[3.33 0.84 3.95 4.16 2.99 4.86 2.15 2.78 2.84 2.21 2.89
2.56 3.68 2.16 3.15];
    x_south=[5.87 6.31 1.98 2.65 6.78 4.23 0.45 1.51 0.56 2.65 1.54
0.98 0.87];

    qqplot(x_north,x_south)
    title('qqplot')
    [h,p,ksstat] = kstest2(x_north,x_south) % a=0.05
end
```

Και τα αποτελέσματα είναι:

h = 0

p = 0.0609

ksstat = 0.4718

Επομένως, αφού **h=0** η μηδενική υπόθεση ότι τα δύο δείγματα x1 και x2 προέρχονται από την ίδια κατανομή δεν μπορεί να απορριφθεί. Και άρα και τα 2 δείγματα προέρχονται από την ίδια κατανομή! Σκεφτόμαστε τώρα αν είναι η κανονική κατανομή ώστε να μπορούμε μετά να κάνουμε τις συγκρίσεις στις μέσες τιμές και τις διασπορές μιας και τα δείγματα πολύ λίγο ελάχιστο στο πλήθος

Όπως γνωρίζουμε η εφαρμογή του t-test2 θεωρείται ασφαλής μόνο εάν οι **δύο τυπικές αποκλίσεις** δεν διαφέρουν πολύ.

Έτσι ο κώδικας είναι:

```
function ass1_4()
    x_north=[3.33 0.84 3.95 4.16 2.99 4.86 2.15 2.78 2.84 2.21 2.89
2.56 3.68 2.16 3.15];
    x_south=[5.87 6.31 1.98 2.65 6.78 4.23 0.45 1.51 0.56 2.65 1.54
0.98 0.87];

    qqplot(x_north,x_south)
    title('qqplot')
```

```
    [h,p,ksstat] = kstest2(x_north,x_south) % a=0.05
    [h,p,ci,stats] = ttest2(x north,x south)
end
```

Και τα αποτελέσματα είναι:

h = 0

p = 0.7912

ci = -1.1466 1.4897

stats = tstat: 0.2675 df: 26 sd: 1.6923

Επειδή h=0 ισχύει η μηδενική υπόθεση για την ttest2 που είναι τα 2 δείγματα να προέρχονται από κανονική κατανομή.

Συνεπώς μιας και τα δείγματα προέρχονται από κανονικές κατανομές, αν βρούμε τις μέσες τιμές τους και τις διασπορές τους και τις συγκρίνουμε μπορούμε να εξάγουμε αν ο ερευνητής είχε δίκιο ή όχι.

Έχουμε:

```
function ass1_4()
    x_north=[3.33 0.84 3.95 4.16 2.99 4.86 2.15 2.78 2.84 2.21 2.89
2.56 3.68 2.16 3.15];
    x_south=[5.87 6.31 1.98 2.65 6.78 4.23 0.45 1.51 0.56 2.65 1.54
0.98 0.87];

    if(mean(x_north)>mean(x_south))
        disp('The average mean height is taller in North Greece')
    else
        disp('The average mean height is shorter in North Greece')
    end
    if(var(x_north)>var(x_south))
        disp('The variance in heights is bigger in North Greece')
    else
        disp('The variance in heights is smaller in North Greece')
    end
end
```

Και τα αποτελέσματα είναι:

The average mean height is taller in North Greece The variance in heights is smaller in North Greece

Οπότε **επιβεβαιώνεται** ο ισχυρισμός του ερευνητή.

5. Η κατανομή των μετρήσεων του ουρικού οξέος σε mg/100ml σε υγιείς άρρενες ηλικίας 30-39 ετών καταγράφεται στον ακόλουθο πίνακα :

Ουρικό οξύ	**Συχνότητα**
3.0 – 3.4	2
3.5 – 3.9	15
4.0 – 4.4	33
4.5 – 4.9	38
5.0 – 5.4	51
5.5 – 5.9	47
6.0 – 6.4	37
6.5 – 6.9	16
7.0 – 7.4	15
7.5 – 7.9	3
8.0 – 8.4	1
8.5 – 8.9	3

(α) Να υπολογισθούν :

(i) η μέση τιμή, (ii) η διάμεσος, (iii) η επικρατούσα τιμή, (iv) η διασπορά και (v) ο συντελεστής μεταβλητότητας

του ουρικού οξέος των αρρένων του δείγματος.

(β) Να κατασκευασθεί το αντίστοιχο θηκόγραμμα.

(γ) Αφού συγχωνευθούν ανά δύο οι κλάσεις των ανωτέρω δεδομένων, να υπολογιστούν τα μέτρα (i) έως και (v) της ερώτησης (α) και να συγκριθούν με τις αντίστοιχες τιμές του ερωτήματος (α).

Απάντηση:

(α)

Αρχικά αντιπροσωπεύουμε κάθε διάστημα τιμών με τη μέση τιμή το, Χ. Έτσι:

Ουρικό οξύ	X	Συχνότητα
3.0 – 3.4	3.2	2
3.5 – 3.9	3.7	15
4.0 – 4.4	4.2	33
4.5 – 4.9	4.7	38
5.0 – 5.4	5.2	51
5.5 – 5.9	5.7	47
6.0 – 6.4	6.2	37
6.5 – 6.9	6.7	16
7.0 – 7.4	7.2	15
7.5 – 7.9	7.7	3
8.0 – 8.4	8.2	1
8.5 – 8.9	8.7	3

Επειδή σε κάθε διάστημα υπάρχει διαφορετική συχνότητα εμφάνισης του ουρικού οξέος σε mg/100ml σε υγιείς άρρενες ηλικίας 30-39 ετών πρέπει να ληφθεί υπόψη για την εύρεση της Μ.Τ, αφού στο δείγμα μας για παράδειγμα η συχνότητα f=2 για τη μεταβλητή X=3.2 σημαίνει ότι πρέπει να προσμετρηθεί 2 φορές στην κατανομή αυτή η μεταβλητή [3.2, 3.2, …]. Ομοίως, η συχνότητα f=15 για τη μεταβλητή X=3.7 σημαίνει ότι πρέπει να προσμετρηθεί 15 φορές στην κατανομή αυτή η μεταβλητή […, 3.7, 3.7, 3.7, 3.7, 3.7, 3.7, 3.7, 3.7, 3.7, 3.7, 3.7, 3.7, 3.7, 3.7, 3.7, 3.7, …] κ.ο.κ.

Άρα ο κώδικας MATLAB είναι ο εξής:

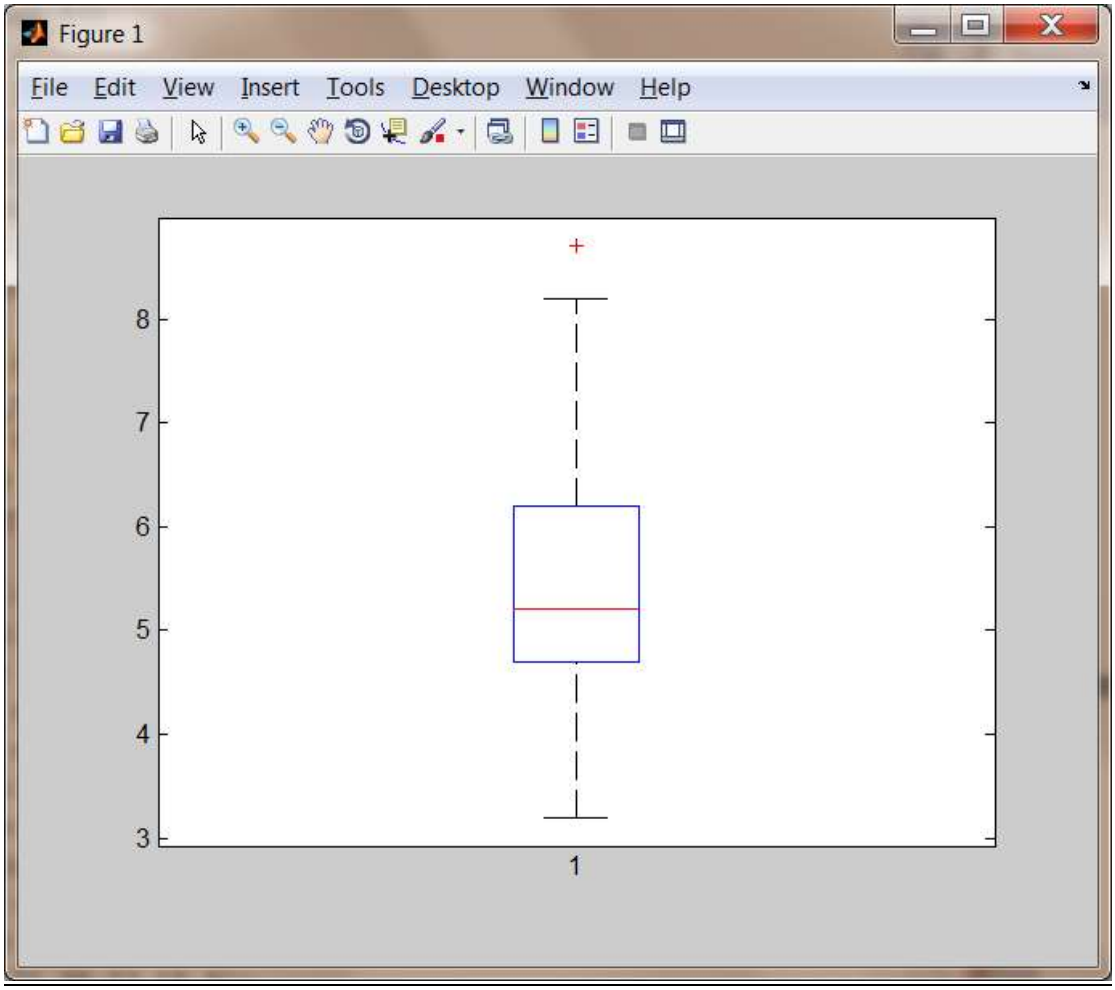
```
% question a
    x=[3.2 3.7 4.2 4.7 5.2 5.7 6.2 6.7 7.2 7.7 8.2 8.7];
    fx=[2 15 33 38 51 47 37 16 15 3 1 3];
    x_new=[];       % including all x values with their frequency
    n=length(x);
    for i=1:n
        x_new=[x_new x(i)*ones(1,fx(i))];
    end;
    mean_x = mean(x_new) % mean_x holds the mean of all x values
    median_x = median(x_new) % median_x holds the meadian of all x
values
    mode_x = mode(x_new) % mode_x holds the prevailing value of all x
values
    var_x = var(x_new) % var_x holds the variance of all x values
    dispFactor_x = (std(x_new)/mean_x) % dispFactor_x holds the
dispersion factor of all x values
```

Και τα αποτελέσματα του παραπάνω κώδικά είναι τα εξής:

```
mean_x = 5.4184
median_x = 5.2000
mode_x = 5.2000
var_x = 1.0810
dispFactor_x = 0.1919
```

(β)

Για να βρούμε το θηκόγραμμα του παραπάνω πίνακα πρέπει αρκεί να καλέσουμε στον πίνακα με όλες τις τιμές x_new την συνάρτηση boxplot



(γ)

Όπως μάθαμε στο 2^{ο} μάθημα, για την ομαδοποίηση των δεδομένων σε κλάσεις και δεδομένου ότι μας ζητείται να συγχωνεύσουμε ανά δύο τις κλάσεις αλλά χωρίς να βρούμε από την αρχή αν είναι αυτή η σωστή ομαδοποίηση των ανωτέρω δεδομένων, το μόνο που έχουμε να κάνουμε είναι τις 12 κλάσεις που μας δίνονται να τις συγχωνεύσουμε σε 6, ανά 2 και να πάρουμε αυτή τη φορά τη μέση τιμή, έστω Y, των νέων διαστημάτων για χρήση στους υπολογισμούς μας. Έτσι, για τον υπολογισμό των μέτρων (i) έως και (v) της ερώτησης (α) έχουμε τον εξής πίνακα:

Ουρικό οξύ	**Y**	**Συχνότητα**
3.0 – 3.9	3.45	17
4.0 – 4.9	4.45	71
5.0 – 5.9	5.45	98
6.0 – 6.9	6.45	53

7.0 – 7.9	7.45	18
8.0 – 8.9	8.45	4

και να συγκριθούν με τις αντίστοιχες τιμές του ερωτήματος (α).

Επειδή σε κάθε διάστημα υπάρχει διαφορετική συχνότητα εμφάνισης του ουρικού οξέος σε mg/100ml σε υγιείς άρρενες ηλικίας 30-39 ετών πρέπει να ληφθεί υπόψη για την εύρεση της Μ.Τ, αφού στο δείγμα μας για παράδειγμα η συχνότητα f=2 για τη μεταβλητή X=3.2 σημαίνει ότι πρέπει να προσμετρηθεί 2 φορές στην κατανομή αυτή η μεταβλητή [3.2, 3.2, …]. Ομοίως, η συχνότητα f=15 για τη μεταβλητή X=3.7 σημαίνει ότι πρέπει να προσμετρηθεί 15 φορές στην κατανομή αυτή η μεταβλητή […, 3.7, 3.7, 3.7, 3.7, 3.7, 3.7, 3.7, 3.7, 3.7, 3.7, 3.7, 3.7, 3.7, 3.7, 3.7, 3.7, …] κ.ο.κ.

Άρα ο κώδικας MATLAB είναι ο εξής:

```
function ass1_5()
% question a
    x=[3.2 3.7 4.2 4.7 5.2 5.7 6.2 6.7 7.2 7.7 8.2 8.7];
    fx=[2 15 33 38 51 47 37 16 15 3 1 3];
    x_new=[];       % including all x values with their frequency
    n=length(x);
    for i=1:n
        x_new=[x_new x(i)*ones(1,fx(i))];
    end;
    mean_x = mean(x_new) % mean_x holds the mean of all x values
    median_x = median(x_new) % median_x holds the meadian of all x
values
    mode_x = mode(x_new) % mode_x holds the prevailing value of all x
values
    var_x = var(x_new) % var_x holds the variance of all x values
    dispFactor x = (std(x new)/mean x) % dispFactor x holds the
dispersion factor of all x values

% question b
%    boxplot(x_new)
    s(1) = subplot(1,2,1); % left subplot
    s(2) = subplot(1,2,2); % right subplot
    boxplot(s(1),x_new)
    title(s(1),'x new')
    ylabel(s(1),'values')
% question c
    y=[3.45 4.45 5.45 6.45 7.45 8.45];
    fy=[17 71 98 53 18 4];
    y new=[];       % including all x values with their frequency
    n=length(y);
    for i=1:n
        y_new=[y_new y(i)*ones(1,fy(i))];
    end;
    mean_y = mean(y_new) % mean_y holds the mean of all y values
    median_y = median(y_new) % median_y holds the meadian of all y
values
    mode_y = mode(y_new) % mode_y holds the prevailing value of all y
values
    var_y = var(y_new) % var_y holds the variance of all y values
    dispFactor_y = (std(y_new)/mean_y) % dispFactor_y holds the
dispersion factor of all y values
```

```
%    figure;
%    boxplot(y_new)
    boxplot(s(2),y_new)
    title(s(2),'y new')
    ylabel(s(2),'values')

% question c comparison
    x_comparisons=[x_new; y_new];
    [p,table,stat]=anova1(x_comparisons);
    c=multcompare(stat)

end
```

Και τα αποτελέσματα του παραπάνω κώδικά είναι τα εξής:

```
mean_x = 5.4184
median_x = 5.2000
mode_x = 5.2000
var_x = 1.0810
dispFactor_x = 0.1919
mean_y = 5.4347
median_y = 5.4500
mode_y = 5.4500
var_y = 1.1536
dispFactor_y = 0.1976
```

Όπως παρατηρούμε από τα παραπάνω αποτελέσματα παρόλης της συγχώνευσης των κλάσεων οι μέσες τιμές και οι συντελεστές μεταβλητότητας δεν μεταβάλλονται πολύ και αυτό γιατί επηρεάζονται από όλο το δείγμα. Αντίθετα, οι διάμεσοι και οι επικρατούσες τιμές όπως είναι λογικό μεταβάλλονται αλλά σε επιτρεπτά πλαίσια μιας και πλέον οι κλάσεις μειώθηκαν στο μισό. Γι αυτό το λόγο, επειδή το εύρος των διαστημάτων κάθε κλάσης μεγάλωσε, η διασπορά των τιμών είναι μεγαλύτερη, οπότε και αυτή η μεταβολή κινήθηκε σε λογικά πλαίσια όπως αναμενόταν.

Τα boxplots των x και y πινάκων που περιγράφτηκαν παραπάνω είναι τα εξής: