ΕΡΓΑΣΤΗΡΙΟ ΓΕΩΣΤΑΤΙΣΤΙΚΗΣ

MATLAB

Πληροφορίες σχετικά με τις στατιστικές συναρτήσεις υπάρχουν στο Statistics Toolbox του documentation του MATLAB, στο Help → MATLAB Help. Εκεί μπορεί να δει κάποιος ποιες συναρτήσεις (εντολές) υπάρχουν, πως συντάσσονται και πως χρησιμοποιούνται σε παραδείγματα.
Επίσης, είναι σχεδόν απαραίτητο να δει και να μελετήσει κάποιος τις βασικές κατανομές πιθανότητας (ιστόγραμμα, pdfs, cdfs) από το Help → Demos → Toolboxes → Statistics → [Probability Distributions , Random Number Generation]. Τέλος, εισαγωγικές σημειώσεις για το MATLAB υπάρχουν στη διεύθυνση *http://www.mred.tuc.gr/home/MatlabIntroV1.pdf*

Παρακάτω θα εξετάσουμε μερικά απλά παραδείγματα.

1.

Το πλήθος των επιτυχιών (π.χ 'γράμματα') στις ρίψεις ενός νομίσματος δίνεται από την δυωνυμική κατανομή. Η συνάρτηση binornd(N,p) δίνει το πλήθος των επιτυχιών σε N επαναλήψεις όταν η πιθανότητα επιτυχίας είναι p και η binornd(N,p,m,n) m x n αποτελέσματα σε ένα πίνακα m x n.
Ας πούμε ότι έχουμε ένα νόμισμα και θέλουμε να δούμε πόσο 'δίκαιο' είναι. Ας υποθέσουμε ότι στην πραγματικότητα είναι κατασκευασμένο έτσι ώστε να δίνει 'γράμματα' με πιθανότητα 55% αντί 50%. Κάνοντας 10 πειράματα με 100 επαναλήψεις τη φορά, παίρνουμε ένα πίνακα 1x10, όπου το κάθε στοιχείο του δίνει το πλήθος των επιτυχιών ('γράμματα')

```
» gramata=binornd(100,0.55,1,10)

gramata =

54   51   56   45   52   55   55   54   55   65
```

Η μέση τιμή και η τυπική απόκλιση του δείγματος είναι

```
» [mean(gramata), std(gramata)]

ans =

54.2000         4.9621
```

Η (άγνωστη) μέση τιμή είναι μ=p=55 και η (άγνωστη) τυπική απόκλιση είναι σ = $\sqrt{Npq}$ = 4.975. Οι δειγματικές τιμές που πήραμε είναι αρκετά κοντά. Μια εκτίμηση λοιπόν της πιθανότητας επιτυχίας ('γράμματα') είναι $\hat{p} = 0.542$, που σημαίνει ότι το νόμισμα *ίσως* ευνοεί τα 'γράμματα'. Το πόσο σίγουροι είμαστε γι' αυτό θα εκφράζεται από ένα επίπεδο εμπιστοσύνης (που το διαλέγουμε εμείς) και ένα αντίστοιχο διάστημα εμπιστοσύνης που θα εξαρτάται από την τυπική απόκλιση και την ίδια την κατανομή. Αυξάνοντας το πλήθος των επαναλήψεων N, πετυχαίνουμε σταδιακά μικρότερη (δειγματική) τυπική απόκλιση, άρα καλύτερη εκτίμηση. Για να δούμε καλύτερα την εξάρτηση από το πλήθος των επαναλήψεων υπολογίζουμε τα αποτελέσματα 5 πειραμάτων με 10, ... ,10^4 επαναλήψεις τη φορά.

```
>> N=logspace(1,4,4)

N =

      10       100      1000     10000

>> gramata=binornd(N,0.55)

gramata =

       6        59       547      5501

>> p=gramata./N

p =

   0.6000    0.5900    0.5470    0.5501
```

Όπως αναμενόταν η εκτίμηση του p γίνεται σταδιακά καλύτερη, συγκλίνοντας στην θεωρητική τιμή. Αν θέλουμε να ξέρουμε τα (π.χ 90%) διαστήματα εμπιστοσύνης, χρησιμοποιούμε την αντίστοιχη συνάρτηση που κάνει εκτίμηση παραμέτρων για τη διωνυμική κατανομή

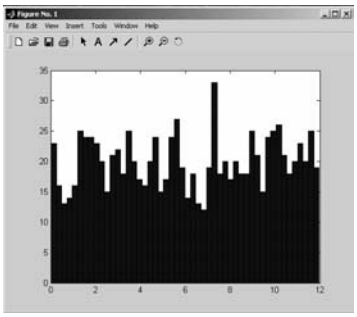
```
>> [p,cip]=binofit(gramata',N',0.1)

p =

    0.6000
    0.5900
    0.5470
    0.5501


cip =

    0.3035    0.8500
    0.5029    0.6730
    0.5205    0.5733
    0.5419    0.5583
```

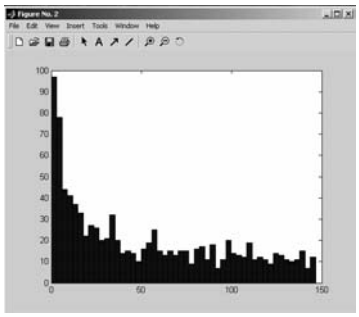
Η συνάρτηση binofit() κάνει μια εκτίμηση από ένα δείγμα (το gramata) με γνωστό πλήθος επαναλήψεων (το N), για την πιθανότητα p (στήλη p) και το αντίστοιχο 90% διάστημα εμπιστοσύνης (στήλες cip). Παρατηρήστε ότι για N=10, το διάστημα εμπιστοσύνης είναι περίπου [0.3, 0.8]. Για ικανοποιητικό πλήθος N=10000, είμαστε 90% σίγουροι (δηλ., με πιθανότητα λάθους 0.01) ότι η πραγματική τιμή είναι στο [0.5419, 0.5583] με εκτίμηση την p = 0.5501

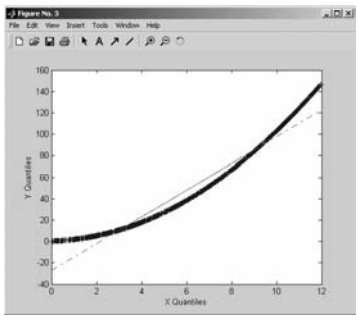
2.

Έστω ότι η τυχαία μεταβλητή X ακολουθεί την ομοιόμορφη κατανομή στο [0,12] και $Y = X^2 + \sqrt{X}$ μια νέα τυχαία μεταβλητή. Θα εξετάσουμε την κατανομή των μεταβλητών και την (γραμμική) εξάρτησή τους.

```
>> x=unifrnd(0,12,1,1000);
>> y=x.^2+sqrt(x);
>> hist(x,50);
>> figure;
>> hist(y,50);
>> figure;
>> qqplot(x,y);


>> cov(x,y)

ans =

  1.0e+003 *

    0.0121    0.1492
    0.1492    1.9509

>> [var(x) var(y)]

ans =

  1.0e+003 *

    0.0121    1.9509
```







Φτιάχνουμε ένα δείγμα για κάθε μεταβλητή, με 1000 παρατηρήσεις το κάθε ένα. Το πρώτο ιστόγραμμα ανήκει στη μεταβλητή X (ομοιόμορφη κατανομή) και το δεύτερο στη μεταβλητή Y, στο οποίο φαίνεται ότι η κατανομή της Y *δεν* είναι ομοιόμορφη. Το ίδιο μπορούμε να δούμε στο διάγραμμα ποσοστιαίων σημείων των δυο μεταβλητών, που δίνει η εντολή qqplot(x,y). Αν ανήκαν στην ίδια κατανομή θα έπρεπε όλα τα ζεύγη τιμών (μπλε) να βρίσκονται πάνω στην ευθεία.

Η κατανομή της Y δεν είναι ομοιόμορφη λόγω των μη-γραμμικών όρων που περιέχει, και μάλιστα περισσότερο λόγω του όρου X^2 αφού είναι σε μεγαλύτερη δύναμη.

Η εντολή cov(x,y) δίνει τον πίνακα συνδιακύμανσης (covariance matrix) για τις δυο μεταβλητές. Ο πίνακας αυτός είναι συμμετρικός. Τα διαγώνια στοιχεία του είναι οι αντίστοιχες (δειγματικές) διασπορές (var(x), var(y)) των δυο μεταβλητών και τα μη-διαγώνια στοιχεία του η (δειγματική) συνδιακύμανση (συντελεστής συνδιακύμανσης) των δυο μεταβλητών, δηλ. το σ_{XY}^2 ή καλύτερα το s_{XY}^2

Ο πίνακας συσχέτισης (correlation matrix) δίνεται από την corrcoef(x,y) και είναι επίσης συμμετρικός με διαγώνια στοιχεία πάντα ίσα με τη μονάδα. Οι μεταβλητές *X* και *Y* είναι προφανώς εξαρτημένες και μάλιστα μη-γραμμικά. Ο συντελεστής (γραμμικής) συσχέτισης ισούται με το μη-διαγώνιο στοιχείο του πίνακα αυτού.

```
>> corrcoef(x,y)

ans =

   1.0000    0.9721
   0.9721    1.0000
```

3.

Αν ένα σύνολο παρατηρήσεων ακολουθεί την κανονική κατανομή είναι γνωστό ότι η δειγματική μέση τιμή ακολουθεί επίσης την κανονική κατανομή και η δειγματική τυπική διασπορά ακολουθεί την κατανομή χ^2.

Συγκεκριμένα, αν ένα σύνολο *n* παρατηρήσεων (το πλήθος του δείγματος) είναι κανονικά κατανεμημένο με διασπορά σ^2 και αν s^2 είναι η δειγματική διασπορά, τότε ισχύει ότι $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$. Δηλαδή η μεταβλητή $\frac{(n-1)s^2}{\sigma^2}$ ακολουθεί την κατανομή χ^2 με *n*-1 βαθμούς ελευθερίας.

Για να επαληθεύσουμε την ισχύ της παραπάνω πρότασης θα χρησιμοποιήσουμε διάγραμμα ποσοστιαίων σημείων και την εντολή qqplot( ).

Έστω ότι το δείγμα ακολουθεί την κανονική κατανομή *N(0,2)*, δηλ. με μ=0 και σ=2. Διαλέγουμε σαν πλήθος του δείγματος ένα ικανοποιητικά μεγάλο αριθμό, π.χ *n*=5000. Θα πρέπει όμως να εξετάσουμε πολλά τέτοια δείγματα από την κανονική κατανομή, π.χ 500 δείγματα. Φτιάχνουμε έτσι ένα πίνακα 5000 x 500, κάθε *στήλη* του οποίου είναι ένα δείγμα πλήθους *n*=5000

```
>> n=5000;
>> x=normrnd(0,2,n,500);
```

Έπειτα, για κάθε δείγμα παίρνουμε τη (δειγματική) διασπορά του, φτιάχνοντας ένα πίνακα 1 x 500 που περιέχει τις (δειγματικές) διασπορές των 500 δειγμάτων

```
y=var(x);
```

Δημιουργούμε τη μεταβλητή $\frac{(n-1)s^2}{\sigma^2}$

```
y=((n-1)*var(x))/4;
```

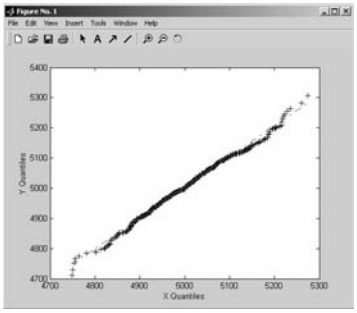
Η μεταβλητή *y* θα ακολουθεί την κατανομή χ^2 με *n*-1 βαθμούς ελευθερίας. Για να το διαπιστώσουμε, φτιάχνουμε μια νέα μεταβλητή *z* από την κατανομή χ^2 με *4999* βαθμούς ελευθερίας, η οποία περιέχει ένα δείγμα πλήθους, π.χ 600 παρατηρήσεων σε ένα πίνακα 1 x 600

```
>> z=chi2rnd(n-1,1,600);
```

Τέλος, παίρνουμε το διάγραμμα ποσοστιαίων σημείων των μεταβλητών *y* και *z*

```
>> qqplot(y,z);
```

Από το διπλανό διάγραμμα φαίνεται ότι οι δυο μεταβλητές ακολουθούν την ίδια κατανομή

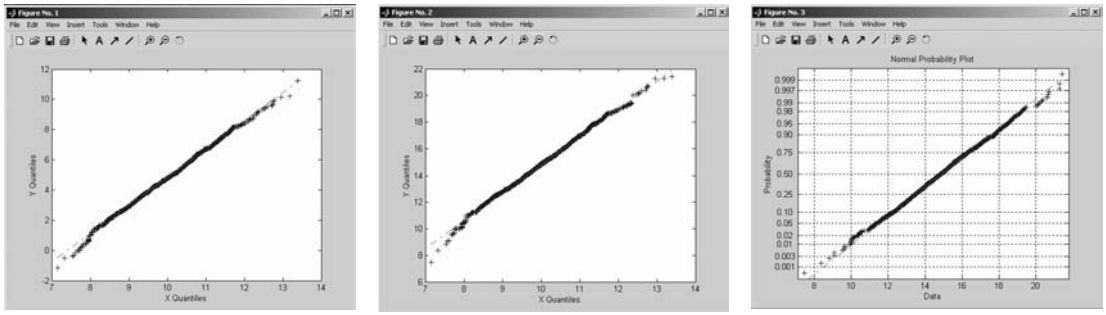


4.

Έστω ότι τα μεγέθη *X* και *Y* είναι τυχαίες μεταβλητές και ακολουθούν την κανονική κατανομή με παραμέτρους μ_X =10 , σ_X^2 =1 και μ_Y =5 , σ_Y^2 =4. Θα εξετάσουμε την κατανομή της τυχαίας μεταβλητής *X* + *Y*, που ως γνωστό είναι επίσης κανονική με μέση τιμή και διασπορά το άθροισμα των αντίστοιχων παραμέτρων των δυο κατανομών, δηλ. μ_{X+Y} = 15 και σ_{X+Y}^2 = 5

Φτιάχνουμε από ένα δείγμα ικανοποιητικού πλήθους, π.χ *n*=1000, για κάθε μία κατανομή και τα προσθέτουμε

```
>> n=1000;
>> x=random('Normal',10,1,1,n);
>> y=random('Normal',5,2,1,n);
>> z=x+y;
>> qqplot(x,y);
>> figure;
>> qqplot(x,z);
>> figure;
>> normplot(z);
```



Στα δύο πρώτα διαγράμματα ποσοστιαίων σημείων (των *X*, *Y* και *X*, *Z*) φαίνεται όπως αναμένεται ότι και οι τρεις μεταβλητές ακολουθούν κοινή κατανομή. Ειδικά για την κανονική κατανομή η εντολή normplot() μπορεί να χρησιμοποιηθεί για να εξεταστεί αν ένα δείγμα είναι κανονικά κατανεμημένο.
Το τρίτο διάγραμμα δηλαδή δείχνει ότι η μεταβλητή *X* + *Y* είναι κανονική. Επιπλέον, δείχνει ότι έχει μέση τιμή κοντά στο 15, διότι η διάμεσος (=μέση τιμή, στην κανονική κατανομή) είναι περίπου 14.9 αν κάνουμε χρήση της εντολής zoom on.

Για να εξετάσουμε τώρα τη μέση τιμή και τη διασπορά της μεταβλητής *X* + *Y* μπορούμε να κάνουμε χρήση της normfit(), η οποία εκτιμά τη μέση τιμή και την τυπική απόκλιση ενός δείγματος, με την προϋπόθεση ότι προέρχεται από την κανονική κατανομή. Σε επίπεδο σημαντικότητας 1%, δηλαδή με 0.01 πιθανότητα λάθους, οι εκτιμήσεις των παραμέτρων με τα αντίστοιχα διαστήματα εμπιστοσύνης είναι

```
>> [m,s,mci,sci] = normfit(z,0.01)

m =

   15.0818

s =

    2.1389

mci =

   14.9073
   15.2564

sci =

    2.0220
    2.2691
```

Οι θεωρητικές τιμές είναι μ_{X+Y} = 15 και σ_{X+Y} = $\sqrt{5}$ =2.236. Αν αυξήσουμε το πλήθος του δείγματος οι εκτιμήσεις θα βελτιωθούν και τα διαστήματα εμπιστοσύνης θα ΄στενέψουν΄.

Ένας άλλος τρόπος να εκτιμήσουμε τη μέση τιμή για μια κανονική μεταβλητή είναι να χρησιμοποιήσουμε την ttest() ή την ztest() οι οποίες κάνουν έλεγχο μιας υπόθεσης. Η πρώτη χρησιμοποιείται όταν η διασπορά είναι άγνωστη και χρησιμοποιεί

το t-statistic $t = \frac{\overline{x} - m}{s/\sqrt{n}}$, ενώ η δεύτερη χρησιμοποιείται όταν η διασπορά είναι γνωστή και χρησιμοποιεί το z-statistic $z = \frac{\overline{x} - m}{\sigma/\sqrt{n}}$, όπου $\overline{x}$ είναι η δειγματική μέση τιμή, m είναι η μέση τιμή που θέλουμε να ελέγξουμε αν είναι σωστή, s είναι η δειγματική τυπική απόκλιση και σ είναι η θεωρητική (πραγματική) τυπική απόκλιση.

Έτσι, αν θεωρήσουμε ότι η πραγματική διασπορά για τη μεταβλητή $X + Y$ είναι άγνωστη και θέλουμε να ελέγξουμε την υπόθεση $\mu_{X+Y} = 15$ με εναλλακτική την $\mu_{X+Y} \neq 15$, σε επίπεδο σημαντικότητας 1%

```
>> [h,p,mci] = ttest(z,15,0.01,0)

h =

     0

p =

    0.2267

mci =

   14.9073   15.2564
```

Επειδή h=0 δε μπορούμε να απορρίψουμε την υπόθεση $\mu_{X+Y} = 15$ σε επίπεδο σημαντικότητας 1%. Δηλαδή επειδή η p, που είναι η πιθανότητα (που αντιστοιχεί στο παραπάνω t-statistic) να παρατηρήσουμε την τιμή $\overline{x}$, δεν είναι μικρότερη από το επίπεδο σημαντικότητας (0.01) δε μπορούμε να απορρίψουμε τη μηδενική υπόθεση (null hypothesis)

Αν τώρα θεωρήσουμε ότι η πραγματική διασπορά για τη μεταβλητή $X + Y$ είναι με κάποιο τρόπο γνωστή και θέλουμε να ελέγξουμε την υπόθεση $\mu_{X+Y} = 14$ με εναλλακτική την $\mu_{X+Y} < 14$, σε επίπεδο σημαντικότητας 5%

```
>> [h,p,mci] = ztest(z,14,sqrt(5),0.05,-1)

h =

     0

p =

     1

mci =

     -Inf   15.1981
```

Επειδή h=0 δε μπορούμε να απορρίψουμε την υπόθεση $\mu_{X+Y} = 14$ έναντι της $\mu_{X+Y} < 14$ σε επίπεδο σημαντικότητας 5%. Δηλαδή επειδή η p, που είναι η πιθανότητα (που αντιστοιχεί στο παραπάνω z-statistic) να παρατηρήσουμε την τιμή $\overline{x}$, δεν είναι μικρότερη από το επίπεδο σημαντικότητας (0.01) δε μπορούμε να απορρίψουμε τη μηδενική υπόθεση (null hypothesis). Εδώ όπως είναι λογικό η πιθανότητα να παρατηρήσουμε την τιμή $\overline{x}$ (ή κάποια μεγαλύτερη τιμή) με την προϋπόθεση ότι $\mu_{X+Y} = 14$, είναι μονάδα (βέβαιο ενδεχόμενο)