7
Mathematics

ΚΕΦΑΛΑΙΟ 7
ΜΑΘΗΜΑΤΙΚΑ

The R Book Michael J. Crawley

Μπορείτε να κάνετε πολλά μαθηματικά στο R. Εδώ έχουμε συγκεντρώσει σχετικά με τα είδη των μαθηματικών που βρίσκουν πιο συχνή εφαρμογή της επιστημονικής εργασίας και των στατιστικών μοντέλων:

• Συναρτήσεις?

• συνεχείς κατανομές?

• διακριτές κατανομές?

• γραμμική άλγεβρα?

• λογισμός?

• διαφορικές εξισώσεις.

Μαθηματικές Συναρτήσεις

Για τα είδη των συναρτήσεων που θα συναντήσετε στην υπολογιστική στατιστική, υπάρχουν μόνο τρεις μαθηματικοί κανόνες που θα πρέπει να μάθετε: αυτοί που ασχολούνται με δυνάμεις, εκθέτες και λογαρίθμους. Στην X^b έκφραση η επεξηγηματική μεταβλητή αυξάνεται στη **b δύναμη**. Στην e^x επεξηγηματική μεταβλητή εμφανίζεται ως μια δύναμη, σε αυτή την ειδική βάση e, με e = 2. 71828, εκ των οποίων x είναι **ο εκθέτης**. Το αντίστροφο του e^x είναι **ο λογάριθμος** του x, συμβολίζεται με log (x) - σημειώστε ότι όλα τα logs μας είναι σε βάση e και ότι, για εμάς, γράφοντας log (x) είναι το ίδιο με το ln (x).

Είναι επίσης χρήσιμο να θυμηθούμε κάποια στοιχειώδη μαθηματικά ώς γεγονότα που είναι χρήσιμα όταν δουλεύουμε την συμπεριφορά στα όρια. Θα θέλαμε να γνωρίζουμε τι συμβαίνει στο y όταν το x γίνεται πολύ μεγάλο (π.χ. $x \rightarrow \infty$) και τι συμβαίνει στο y όταν x πηγαίνει στο 0 (δηλ. τι το σημείο τομής είναι, αν υπάρχει) Αυτές είναι οι πιο σημαντικοί κανόνες:

• Οτιδήποτε τιμές στη μηδενική δύναμη είναι 1: x^0 = 1.

• Το ένα σε οποιαδήποτε δύναμη εξακολουθεί να είναι 1: 1^x = 1.

• Άπειρο συν 1 είναι άπειρο: ∞ +1 =∞

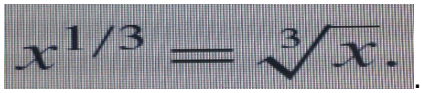
•Ένα πάνω από το άπειρο (Το αντίστροφο του απείρου, ∞^ -1) είναι μηδέν: 1/∞= 0.

• Ένας αριθμός μεγαλύτερος από 1 ανυψωμένος σε δύναμη άπειρο δίνει

άπειρο: 1.2^∞=∞

• Ένα κλάσμα (π.χ. 0.99) ανυψωμένο στη δύναμη άπειρο είναι μηδέν: 0.99^∞= 0.

• Οι αρνητικές δυνάμεις είναι αντίστροφες: x^-b = 1/x^b.

• Οι Κλασματικές δυνάμεις είναι οι ρίζες:  .

• Η βάση των φυσικών λογαρίθμων, e, είναι 2.718 28, έτσι e^∞=∞

• Τελευταίο, αλλά ίσως το πιο χρήσιμο: e^-∞ = 1/e^∞=1/∞=0

Υπάρχουν ενσωματωμένες συναρτήσεις στο R για λογαριθμική, πιθανότητα και τριγωνομετρικές συναρτήσεις (σελ. 11).

Λογαριθμικές Συναρτήσεις

Η λογαριθμική συνάρτηση δίνεται από

$$y = a \ln(bx).$$

Εδώ ο λογάριθμος είναι με βάση το e. Η εκθετική συνάρτηση, στην οποία το y είναι η απόκριση

αντιλογάριθμος της συνεχούς ερμηνευτικής μεταβλητής x, δίνεται από

$$y = ae^{(bx)}$$

Και οι δύο αυτές συναρτήσεις είναι ομαλές συναρτήσεις, και για να σχεδιάσετε ομαλές συναρτήσεις στην R θα πρέπει να δημιουργήσετε μια σειρά από 100 ή περισσότερο τακτικά διαστήματα x τιμών μεταξύ min (x) και max (x):

x <-seq (0,10,0.1)

Στην R η εκθετική συνάρτηση είναι exp και η φυσική log συνάρτηση (ln) είναι log. Έστω a = b = 1.

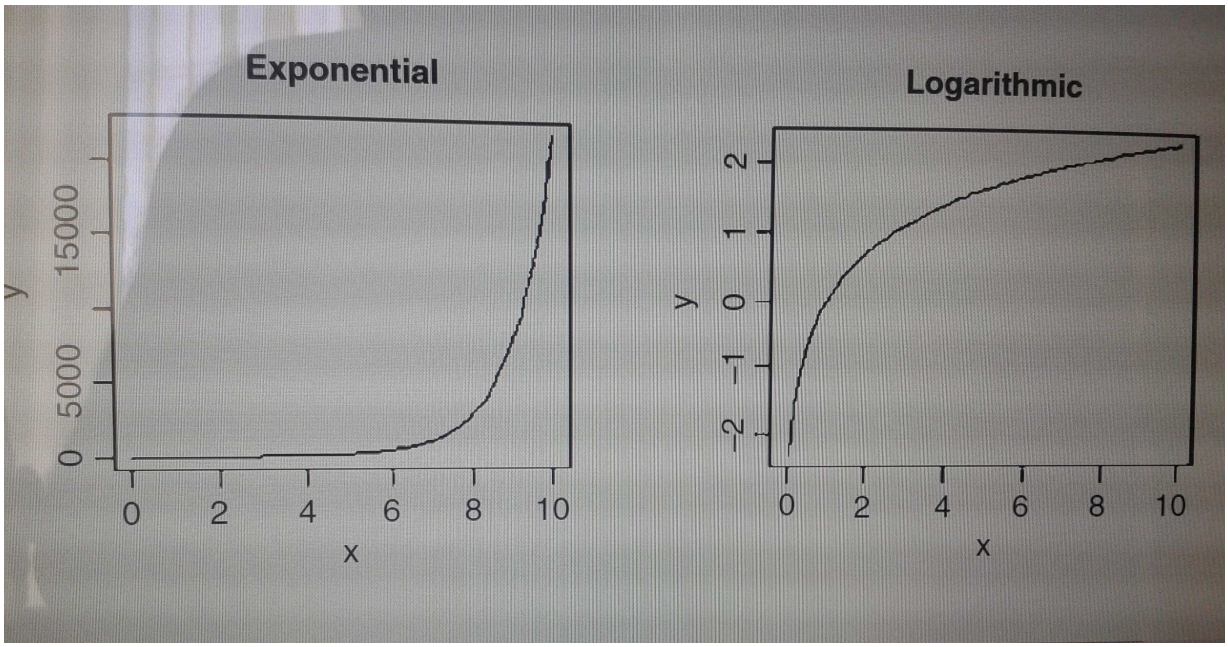
Για να σχεδιάσετε τις εκθετικές και λογαριθμικές συναρτήσεις σε αυτές τις τιμές μαζί σε μια σειρά, γράψτε

y <-exp (x)

plot (y ~ x, type = "l", main = " Exponential ")

y <-log (x)

plot (y ~ x, type = "l", main = " Logarithmic ")



Σημειώστε ότι η γραφική παράσταση της συνάρτησης μπορεί να χρησιμοποιηθεί με ένα εναλλακτικό τρόπο, προσδιορίζοντας τις καρτεσιανές συντεταγμένες της γραμμής χρησιμοποιώντας plot (x, y) μάλλον από το μπούσουλα plot (y ~ x) (βλ. σελ.. 181).

Αυτές οι συναρτήσεις είναι πιο χρήσιμες στη μοντελοποίηση της διαδικασίας της εκθετικής αύξησης και της εξασθένησης.

Τριγωνομετρικές συναρτήσεις

Εδώ είναι το cosine (συνημίτονο=προσκείμενη κάθετη / υποτείνουσα), sine (ημίτονο = απέναντι κάθετη / υποτείνουσα) και tangent (εφαπτομένη=απέναντι κάθετη/ προσκείμενη κάθετη) συναρτήσεις του x (μετρούνται σε ακτίνια) πάνω στην περιοχή 0 έως 2π. Υπενθυμίζουμε ότι ο πλήρης κύκλος είναι 2π ακτίνια, έτσι ώστε 1 ακτίνιο = 360/2π = 57, 295 78 βαθμούς.

x <-seq (0,2 * pi, 2 * pi/100)

y1 <-cos (x)

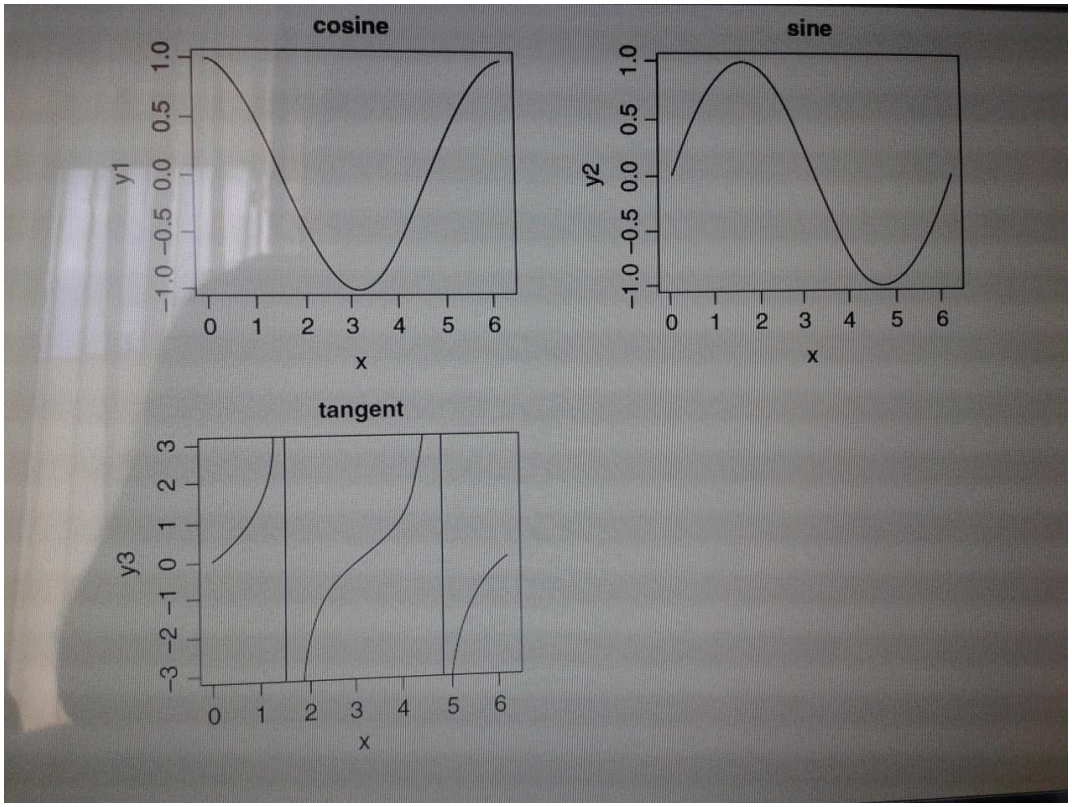
y2 <-sin (x)

Γραφική παράσταση plot(y1 ~ x, type = "l", main = "συνημίτονο")

Γραφική παράσταση plot (y2 ~ x, type = "l", main = "ημίτονο")

y3 <-tan (x)

Γραφική παράσταση plot (y3 ~ x, type = "l", ylim = c (-3,3), main = "εφαπτομένη")



Η εφαπτομένη του x έχει ασυνέχειες, τα γυρίσματα μακριά στο θετικό άπειρο x=π/2 και πάλι στο x =3π/2. Ο περιορισμός του εύρους των τιμών καταγράφεται στον άξονα y (εδώ από -3 σε +3) ως εκ τούτου δίνει μια καλύτερη εικόνα του σχήματος της συνάρτησης tan (εφαπτομένη). Σημειώστε ότι R ενώνει το συν άπειρο και τα «σημεία» μείον άπειρο με μια ευθεία γραμμή x =π / 2 και x = 3π / 2 εντός του πλαίσιου του γραφήματος που ορίζεται από ylim.

Ισχύων νόμος

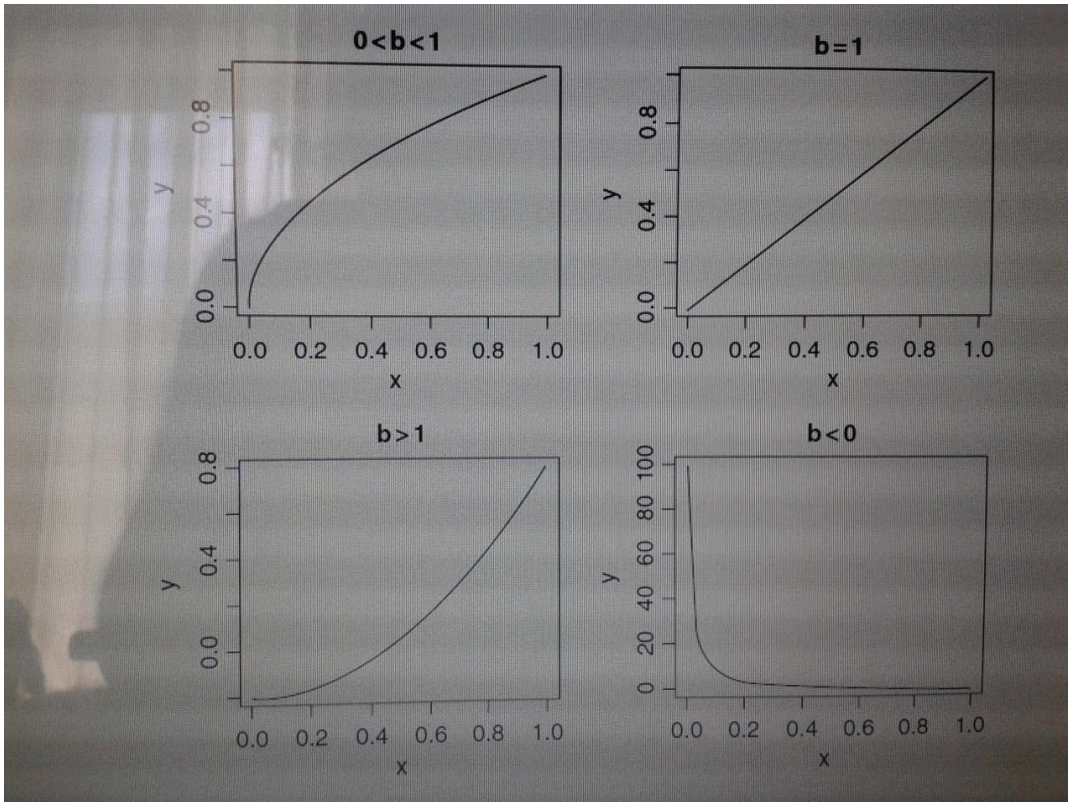
Υπάρχει μια σημαντική οικογένεια των δύο παραμέτρων μαθηματικές συναρτήσεις της μορφής

y = a*(x^b)

γνωστή ως νόμοι δύναμης. Ανάλογα με την τιμή της δύναμης, b, η σχέση μπορεί να λάβει μία από τις πέντε μορφές. Στην οριακή περίπτωση b = 0 η

συνάρτηση y = a (οριζόντια ευθεία). Τα τεσσάρα πιο ενδιαφέροντα σχήματα έχουν ως εξής:

```
x<-seq(0,1,0.01)

y<-x^0.5

plot(x,y,type="l",main="0<b<1")

y<-x

plot(x,y,type="l",main="b=1")

y<-x^2

plot(x,y,type="l",main="b>1")

y<-1/x

plot(x,y,type="l",main="b<0")
```



Οι συναρτήσεις αυτές είναι χρήσιμες σε ένα ευρύ φάσμα των κλάδων. Οι παράμετροι α και b είναι εύκολο να εκτιμηθούν από τα δεδομένα, επειδή η συνάρτηση είναι γραμμική με ένα log-log μετασχηματισμό, log(y)=log(a*x^b)=log(a)+b log(x)

έτσι ώστε σε log-log άξονες το σημείο τομής είναι log (α) και η κλίση είναι b. Αυτά συχνά αποκαλούνται **αλλομετρικές σχέσεις**, επειδή όταν το b≠1, το ποσοστό των x που γίνεται y ποικίλλει ανάλογα με το x. Μια σημαντική εμπειρική σχέση από οικολογική εντομολογία που έχει εφαρμογές σε ένα ευρύ φάσμα της στατιστικής ανάλυσης είναι γνωστή ως **νόμος δύναμης του Taylor**. Έχει να κάνει με η σχέση μεταξύ της διακύμανσης και του μέσου όρου ενός δείγματος. Στα στοιχειώδη στατιστικά μοντέλα, η διακύμανση θεωρείται ότι είναι σταθερή (δηλαδή η διακύμανση δεν εξαρτάται από τη μέση τιμή). Στο πεδίο δεδομένων, ωστόσο, ο Taylor διαπίστωσε ότι η διακύμανση αυξήθηκε με τη μέση τιμή ανάλογα σε ένα νόμο δύναμης, έτσι ώστε σε log-log άξονες τα δεδομένα από περισσότερα συστήματα μειώθηκαν πάνω από τη γραμμή μέσω της προέλευσης με κλίση = 1 (το μοτίβο φαίνεται από τα δεδομένα που διανέμονται στην κατανομή Poisson, όπου η διακύμανση είναι ίση με την μέση τιμή) και κάτω από μια γραμμή μέσω της προέλευσης με κλίση 2. Ο νόμος της δύναμης του Taylor ορίζει ότι, για ένα συγκεκριμένο σύστημα:

• log (διακύμανση) είναι μια γραμμική συνάρτηση του log (μέση)?

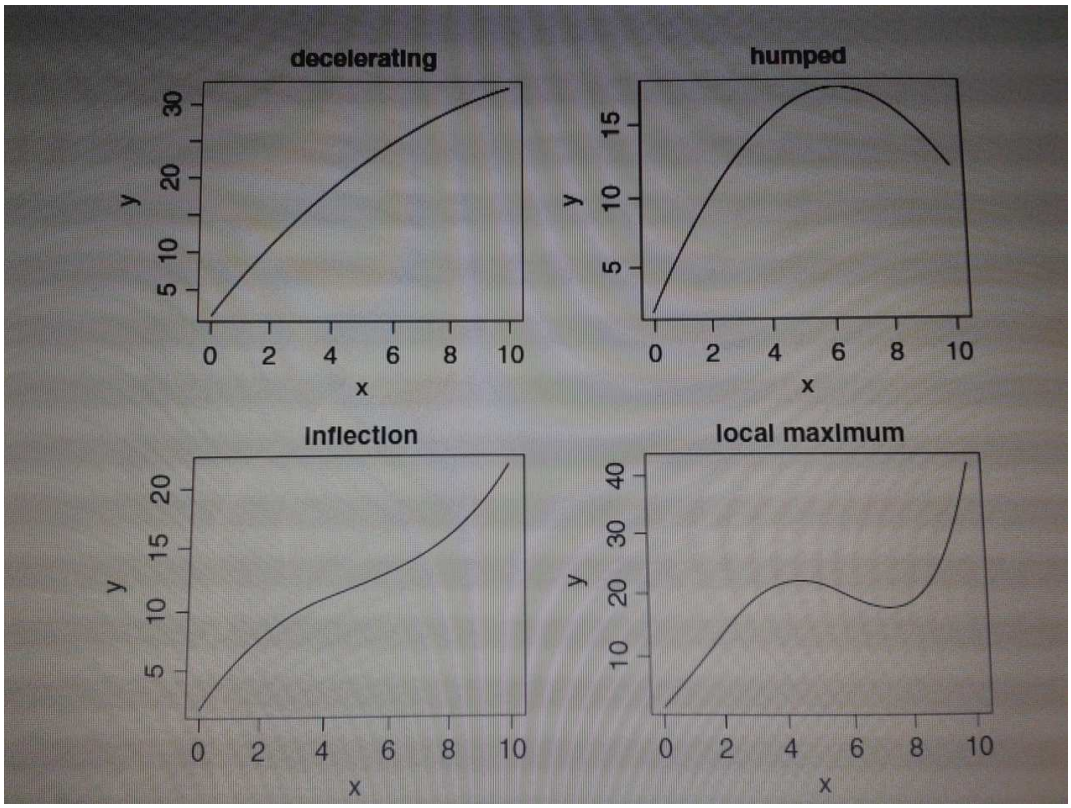
• η διασπορά για αυτό ευθεία γραμμή είναι μικρή?

• η κλίση της παλινδρόμησης του log (διακύμανση) έναντι log (μέση τιμή) είναι μεγαλύτερη από 1 και λιγότερη από 2?

• οι τιμές των παραμέτρων της εμπειρικής σχέσης των μεταβλητών log-log είναι θεμελιώδη χαρακτηριστικά του συστήματος.

Πολυωνυμικές συναρήσεις

Πολυωνυμικές συναρτήσεις είναι συναρτήσεις στις οποίες το X εμφανίζεται πολλές φορές, κάθε φορά για να αυξηθεί μια διαφορετική δύναμη. Είναι χρήσιμα για την περιγραφή των καμπυλών με εξογκώματα, κλίσεων ή τοπικά μέγιστα, όπως αυτά:



Το πάνω αριστερό μέρος του πίνακα δείχνει μια επιβράδυνση θετικής συνάρτησης, που διαμορφώθηκε από την τετραγωνική:

x<-seq(0,10,0.1)

y1<-2+5*x-0.2*x^2

Κάνοντας το αρνητικό συντελεστή του x^2 όρο μεγαλύτερο παράγει μια καμπύλη με ένα εξόγκωμα όπως και στο άνω δεξιά πάνελ:

y2<-2+5*x-0.4*x^2

Τα κυβικά πολυώνυμα μπορούν να δείξουν σημεία καμπής, όπως στην κάτω αριστερή πλευρά του πίνακα:

y3<-2+4*x-0.6*x^2+0.04*x^3

Τέλος, τα πολυώνυμα που περιέχουν 4 δυνάμεις είναι ικανές να παράγουν καμπύλες με τα τοπικά μέγιστα, όπως στο κάτω δεξιό πίνακα:

y4<-2+4*x+2*x^2-0.6*x^3+0.04*x^4

par(mfrow=c(2,2)

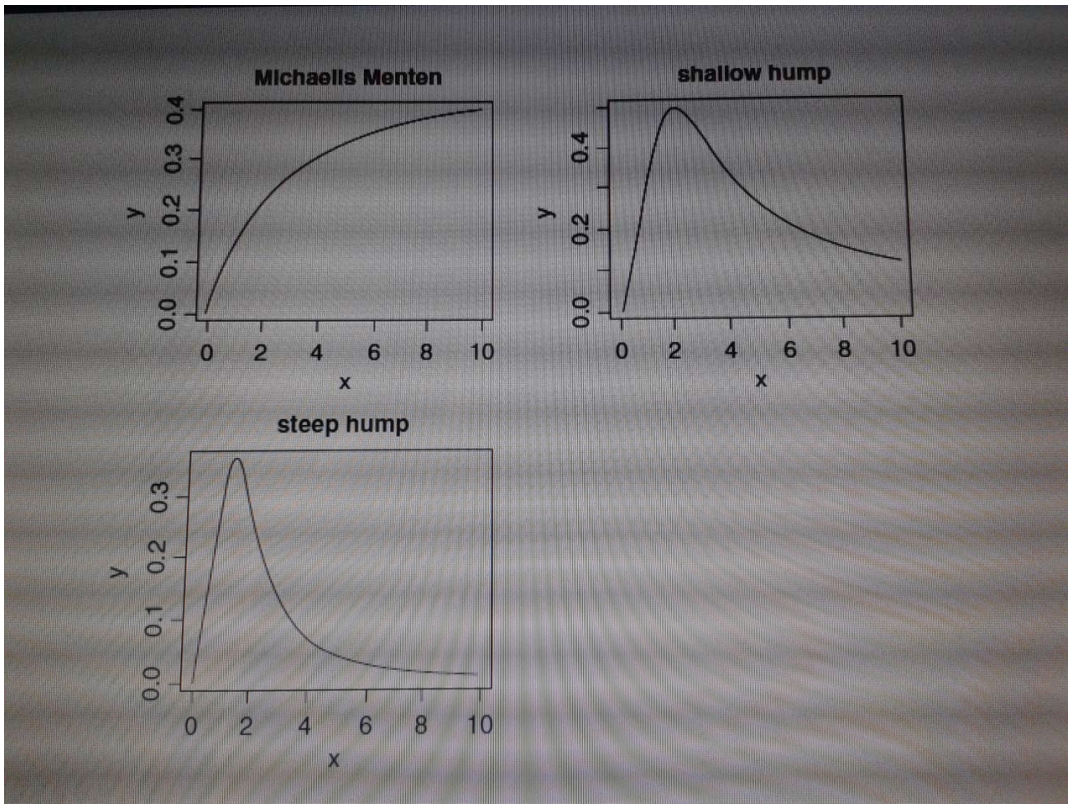
```
plot(x,y1,type="l",ylab="y",main="decelerating")

plot(x,y2,type="l",ylab="y",main="humped")

plot(x,y3,type="l",ylab="y",main="inflection")

plot(x,y4,type="l",ylab="y",main="local maximum")
```

Τα αντίστροφα πολυώνυμα αποτελούν μια σημαντική κατηγορία των συναρτήσεων που είναι κατάλληλες για τη δημιουργία γενικευμένων γραμμικών μοντέλων με τα λάθη γάμα και αντίστροφα συνδεόμενες συναρτήσεις:

1/y=a+bx+cx^2 +dx^3 +…+zx^n

Τα διάφορα σχήματα των συνατήσεων που παράγονται, ανάλογα με τη σειρά του πολυωνύμου (η μέγιστη ισχύς) και οι ενδείξεις των παραμέτρων:

```
par(mfrow=c(2,2))

y1<-x/(2+5*x)

y2<-1/(x-2+4/x)

y3<-1/(x^2-2+4/x)

plot(x,y1,type="l",ylab="y",main="Michaelis-Menten")

plot(x,y2,type="l",ylab="y",main="shallow hump")

plot(x,y3,type="l",ylab="y",main="steep hump")
```



Υπάρχουν δύο τρόποι παραμετροποίησης της εξίσωσης Michaelis-Menten:

y= ax/(1+bx) και y= x/(c+dx)

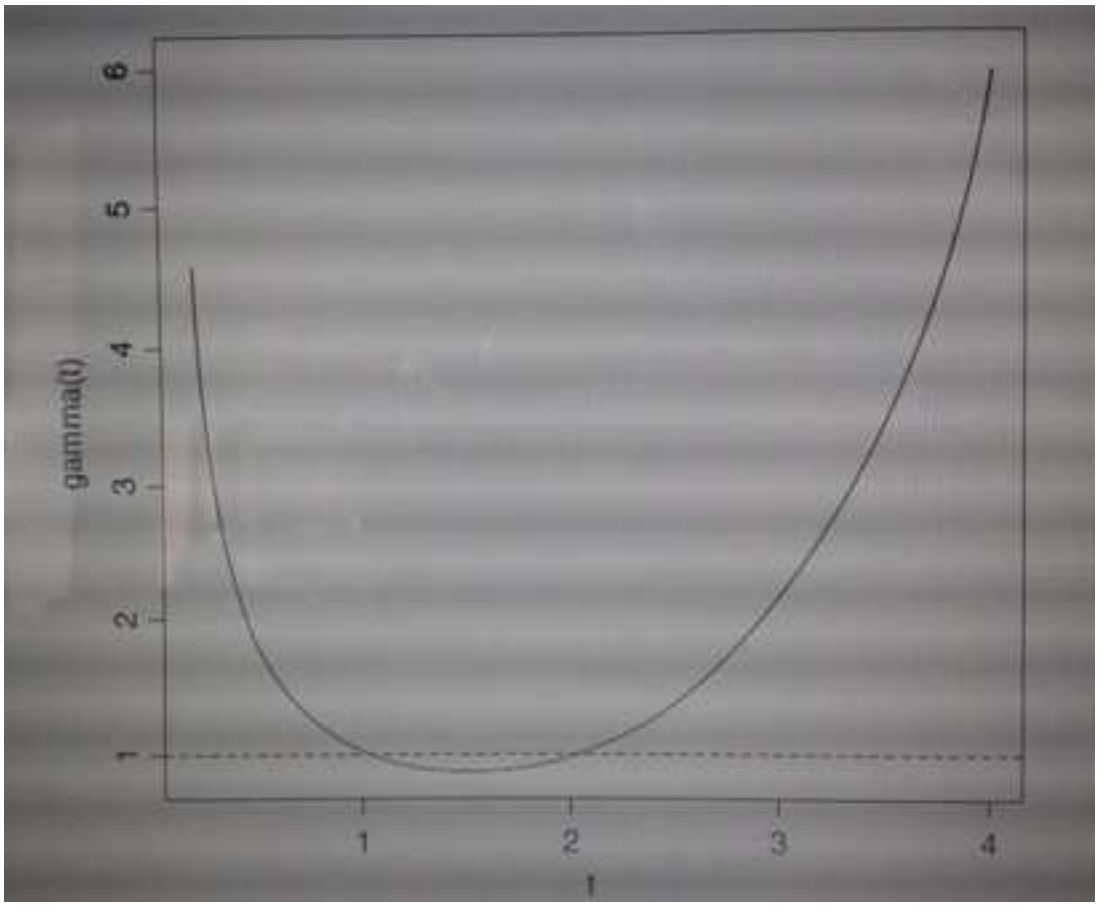
Στην πρώτη περίπτωση, η ασυμπτωτική τιμή του y είναι A / B και στη δεύτερη είναι 1/d.

Συνάρτηση Gamma
Η συνάρτηση γάμμα Γ(t); είναι μια επέκταση της συνάρτηση παραγοντικό, t!, με θετικούς πραγματικούς αριθμούς:

$$\Gamma(t) = \int_0^\infty x^{t-1} \mathrm{e}^{-x} \mathrm{d}x.$$

Μοιάζει με αυτό:

```
t<-seq(0.2,4,0.01)
plot(t,gamma(t),type="l")
abline(h=1,lty=2)
```



Σημειώστε ότι Γ(t) είναι ίσο με 1, τόσο σε t = 1 και t = 2. Για ακέραιες τιμές του t, Γ(t+1)= t!, και.

Ασυμπτωτικές συναρτήσεις

Πολύ η πιο συχνά χρησιμοποιούμενη ασυμπτωτική συνάρτηση είναι

y=ax/(1+bx),

η οποία έχει διαφορετικό όνομα σε σχεδόν σε κάθε επιστημονικό πεδίο. Για παράδειγμα, στη βιοχημεία λέγεται Michaelis-Menten, και εμφανίζει ταχύτητα αντίδρασης ως συνάρτηση της συγκέντρωσης ενζύμου; στην οικολογία ονομάζεται εξίσωση του δίσκου Holling και δείχνει αρπακτικό ποσοστό σίτισης ως συνάρτηση της πυκνότητας των αρπακτικών. Η γραφική παράσταση που διέρχεται από την αρχή και οι αυξήσεις με φθίνουσες αποδόσεις σε μία ασυμπτωτική τιμή στην οποία αυξάνοντας την τιμή του x δεν θα οδηγήσει σε περαιτέρω αύξηση στο y.
Η άλλη κοινή συνάρτηση είναι η ασυμπτωτική εκθετική
y=a[1−e^(−bx)]

Αυτή, επίσης, είναι ένα μοντέλο δύο παραμέτρων, και σε πολλές περιπτώσεις οι δύο συναρτήσεις θα περιγράφουν δεδομένα εξίσου καλά (βλέπε σελ. 664 για ένα παράδειγμα αυτής της σύγκρισης).
Ας εργαστούμε για τη συμπεριφορά των όριων των δύο ασυμπτωτικών συναρτήσεων, ξεκινώντας με την ασυμπτωτική εκθετική. Για x = 0 έχουμε

y=a[1−e^(−b×0)]=a(1−e^0)=a(1−1)=a×0=0.

έτσι ώστε η γραφική παράσταση περνά μέσα από την αρχή των συντεταγμένων. Στο άλλο άκρο, για x =∞, έχουμε

y=a[1−e^(-bx∞)]=a(1−e^-∞)=a(1−0)=a(1)=a,
πράγμα που αποδεικνύει ότι η σχέση είναι ασυμπτωτική, και ότι η ασυμπτωτική τιμή του y είναι a.
Για την εξίσωση του Michaelis-Menten, ο προσδιορισμός της συμπεριφοράς των ορίων είναι κάπως πιο δύσκολος, διότι για x =∞ καταλήγουμε με y =∞/∞ το οποίο μπορείτε να φανταστείτε είναι πάντα θα είναι 1, δεν έχει σημασία ποιες είναι οι τιμές του a και b. Στην πραγματικότητα, υπάρχει ένας ειδικός μαθηματικός κανόνας για την περίπτωση αυτή, που ονομάζεται l''Hospital κανόνας: όταν μπορείτε να πάρετε μια αναλογία του απείρου σε άπειρο, εργάζεστε από το λόγο των παραγώγων για την απόκτηση της συμπεριφοράς στο όριο. Για x=0 το όριο είναι εύκολο:

y=(a*0)/(1+b*0)=0/1+0=0/1=0.

Για x=∞ παίρνουμε y=∞/(1+∞)=∞/∞. Ο αριθμητής είναι ax που η παράγωγος τoy σε σχέση με το x είναι a. Ο παρανομαστής είναι 1+bx έτσι ο παράγωγος του σε σχέση με το x είναι b είναι 0+b=b. Έτσι ώστε η αναλογία των παραγώγων είναι a/b και αυτή είναι η ασυμπτωτική τιμή της εξίσωσης του Michaelis-Menten
.

Εκτίμηση παραμέτρων σε ασυμπτωτικές συναρτήσεις

Δεν υπάρχει κανένας τρόπος να δίνει γραμμική μορφή στο εκθετικό ασυμπτωτικό μοντέλο, οπότε θα πρέπει να καταφύγουμε σε μη γραμμικά ελαχίστα τετραγώνα (nls), για να υπολογίσουμε τις τιμές των παραμέτρων για αυτό (σελ. 662). Ένα από τα πλεονεκτήματα της Michaelis-Menten συνάρτησης είναι ότι είναι εύκολο να δώσει γραμμική μορφή. Χρησιμοποιούμε τον αμοιβαίο μετασχηματισμό
1/y=(1+bx)/ax,
ο οποίος, εκ πρώτης όψεως, δεν είναι μια μεγάλη βοήθεια. Αλλά μπορούμε να διαχωρίσουμε τους όρους σχετικά με ακρίβεια, διότι έχουν ένα κοινό παρονομαστή. Στη συνέχεια, μπορούμε να ακυρώσουμε το xs, όπως αυτό:
1/y=1/ax+bx/ax=1/ax+b/a.
Έτσι αν εμείς βάλουμε y=1/y, A=1/α, και C=b/α, εμείς βλέπουμε ότι

Y=AX+C

Η οποία είναι γραμμική, C είναι το σημείο τομής και ο A είναι η κλίση. Έτσι για να υπολογίζει τις τιμές των a και b από τα δεδομένα, θα μετατρέψει και τα δύο X και Y σε αντίστροφα κλάσματα, η γραφική παράσταση του 1/y απέναντι στο 1/x, πραγματοποιήσει μια γραμμική εμπειρική σχέση μεταβλητών, στη συνέχεια πίσω-μετασχηματισμό, για να πάρει:

α=1/A,
b=αC,

Ας υποθέσουμε ότι γνωρίζαμε ότι η γραφική παράσταση διέρχεται από τα δύο σημεία (0,2, 44.44) και (0.6,70,59). Πώς εργαζόμαστε για τις τιμές των παραμέτρων α και b;
Πρώτον, έχουμε υπολογίσει τις τέσσερις αντίστροφες. Η κλίση της γραμμικοποιημένης συνάρτησης, A, είναι η αλλαγή σε 1/y διαιρούμενο με μεταβολή σε 1/x:
(1/44.44 - 1/70.59) / (1/0.2 - 1/0.6)
[1] 0.002500781
έτσι ώστε a = 1 / A = 1/0; 0025 = 400. Τώρα μπορούμε να αναδιατάξουμε την εξίσωση και να χρησιμοποιήσουμε ένα από τα σημεία
(δηλαδή x = 0.2, y = 44. 44?) για να πάρουμε την τιμή του b:
b=1/x*(αx/y-1)=1/0.2(400*0.2/44.44-1)=4

Sigmoid (σχήματος S) συαρτήσεις

Η απλούστερη σχήματος S είναι η συνάρτηση δύο παραμέτρων λογιστική όπου, για 0≤y≤1, y=exp(α+bx):[1+exp(α+bx)]
η οποία είναι κεντρικής σημασίας για την τοποθέτηση των γενικευμένων γραμμικών μοντέλων για του ποσοστού δεδομένων (κεφάλαιο 16).
Οι τριών παραμέτρων υλικοτεχνική συνάρτηση επιτρέπει y να ποικίλουν σε κάθε κλίμακα:
y=α:[1+b*exp(-cx)].
Το σημείο τομής είναι α:(1+b), η ασυμπτωτική τιμή είναι μια και η αρχική κλίση μετριέται από το c. Εδώ είναι η καμπύλη με παραμέτρους 100, 90 και 1.0:

```
par(mfrow=c(2,2))
x<-seq(0,10,0.1)
y<-100/(1+90*exp(-1*x))
plot(x,y,type="l",main="τριών παραμέτρων λογιστική ")
```

Η τεσσάρων παραμέτρων συνάρτηση υλικοτεχνική έχει ασύμπτωτες στην αριστερή –(α) και το δεξιά - άκρη (b) του άξονα x και κλίμακες (c) η απόκριση σε x για το μέσο (d) όπου η καμπύλη έχει κλίση του:

y = α +(β-α)/(1 + exp[c*(d-x)]).

Αφήνοντας a=20,b=120,c=0.8 και d=3, η συνάρτηση

y=20+100/(1+exp[0.8*(3-x)])

μοιάζει με αυτό

```
y<-20+100/(1+exp(0.8*(3-x)))

plot(x,y,ylim=c(0,140),type="l",main="τεσσάρων παραμέτρων λογιστική ")
```

Αρνητικές σιγμοειδής καμπύλες έχουν την παράμετρο c <0, όπως για τη συνάρτηση

y=20+(100:[1+exp(-0.8*(3-x)].

Μια ασύμμετρη καμπύλη σχήματος S που χρησιμοποιείται πολύ στη δημογραφία και την εργασία ασφάλεια ζωής είναι το **Gompertz μοντέλο ανάπτυξης,**

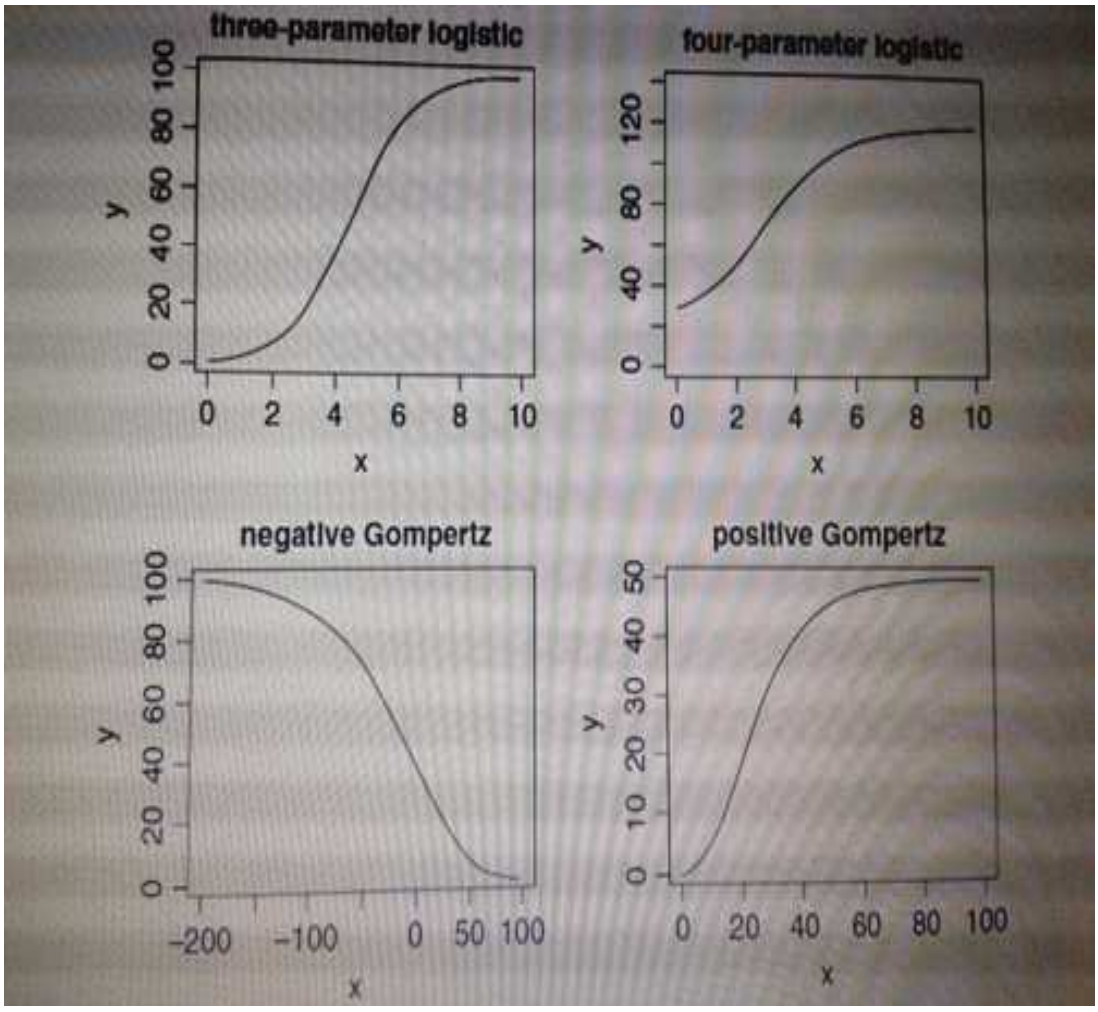
y=α*exp(b*(exp(cx)).

Το σχήμα της συνάρτησης εξαρτάται από τα σημάδια των παραμέτρων b και c. Για αρνητική σιγμοειδές, b είναι αρνητικό (εδώ -1) και το c θετικό (εδώ +0.02):

```
x<-200:100
y<-100*exp(-exp(0.02*x))
plot(x,y,type="l",main=" αρνητική  Gompertz")
```

Για ένα θετικό σιγμοειδές και οι δύο παράμετροι είναι αρνητικές:

```
x<- 0:100
y<- 50*exp(-5*exp(-0.08*x))
plot(x,y,type="l",main="θετική Gompertz")
```

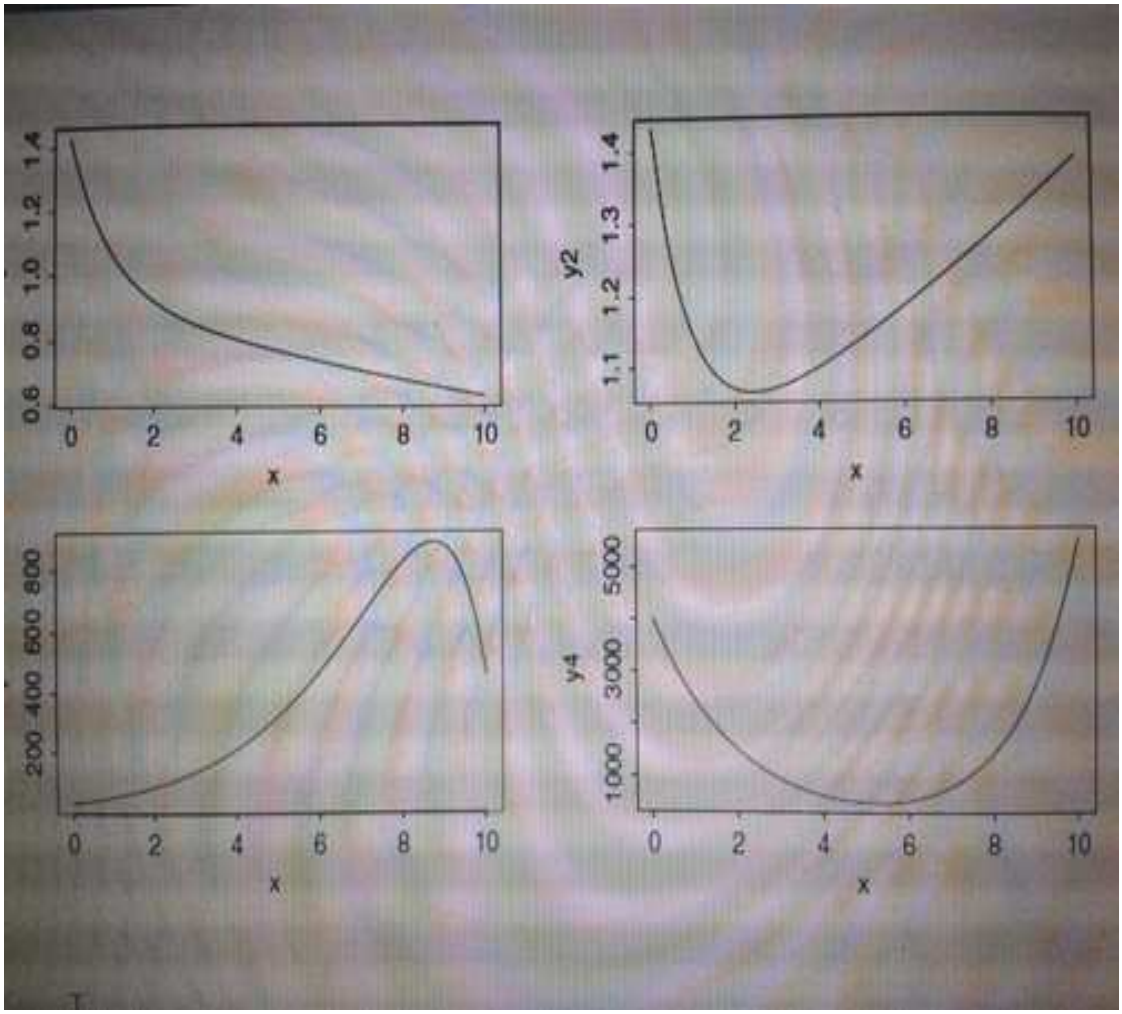


Εκθετικό μοντέλο

Αυτό είναι μια χρήσιμη μη-γραμμική συνάρτηση τεσσάρων παραμέτρων, η οποία είναι το άθροισμα των δύο εκθετικών συναρτήσεων του x:

y=αe^(bx) +ce^(dx).

Διάφορα σχήματα εξαρτώνται από τα σημάδια των παραμέτρων b, c και d:



το άνω αριστερό τμήμα του πίνακα εμφανίζει c θετική, b και d αρνητική; το επάνω δεξιό τμήμα του πίνακα δείχνει c και d θετικό, b αρνητικό; η κάτω αριστερή πλευρά του πίνακα δείχνει c και d αρνητικό, b θετικό; και το κάτω δεξί ταμπλό δείχνει c και b αρνητικό, d θετικό. Όταν b, c και d είναι όλα αρνητικά, αυτή η συνάρτηση είναι γνωστή ως **η πρώτης τάξης μοντέλο διαμερίσματος** στην οποία ένα φάρμακο χορηγείται σε χρόνο μηδέν περνά μέσα από το σύστημα με τη δυναμική του θίγονται από την εξάλειψη, την απορρόφηση και την κάθαρση.

Μεταμορφώσεις της ανταπόκρισης και επεξηγηματικές μεταβλητές

Έχουμε δει τη χρήση του μετασχηματισμού για τη γραμμικοποίηση της σχέσης μεταξύ της απόκρισης και των ερμηνευτικών μεταβλητών:
log(y) κατά x για τις εκθετικές σχέσεις;
log(y) κατά log(x) για δυναμικές συναρτήσεις;
exp(y) κατά x για λογαριθμικές σχέσεις;
1/y κατά 1/x για ασυμπτωτικές σχέσεις;
log(p/(1−p) κατά x για τα δεδομένα αναλογίας

Άλλοι μετασχηματισμοί είναι χρήσιμοι για τη σταθεροποίηση διακύμανσης:

y^1/2 να σταθεροποιήσει τη διακύμανση μέτρησης των δεδομένων ?
arcsin(y) να σταθεροποιηθεί η διακύμανση του ποσοστού των δεδομένων.

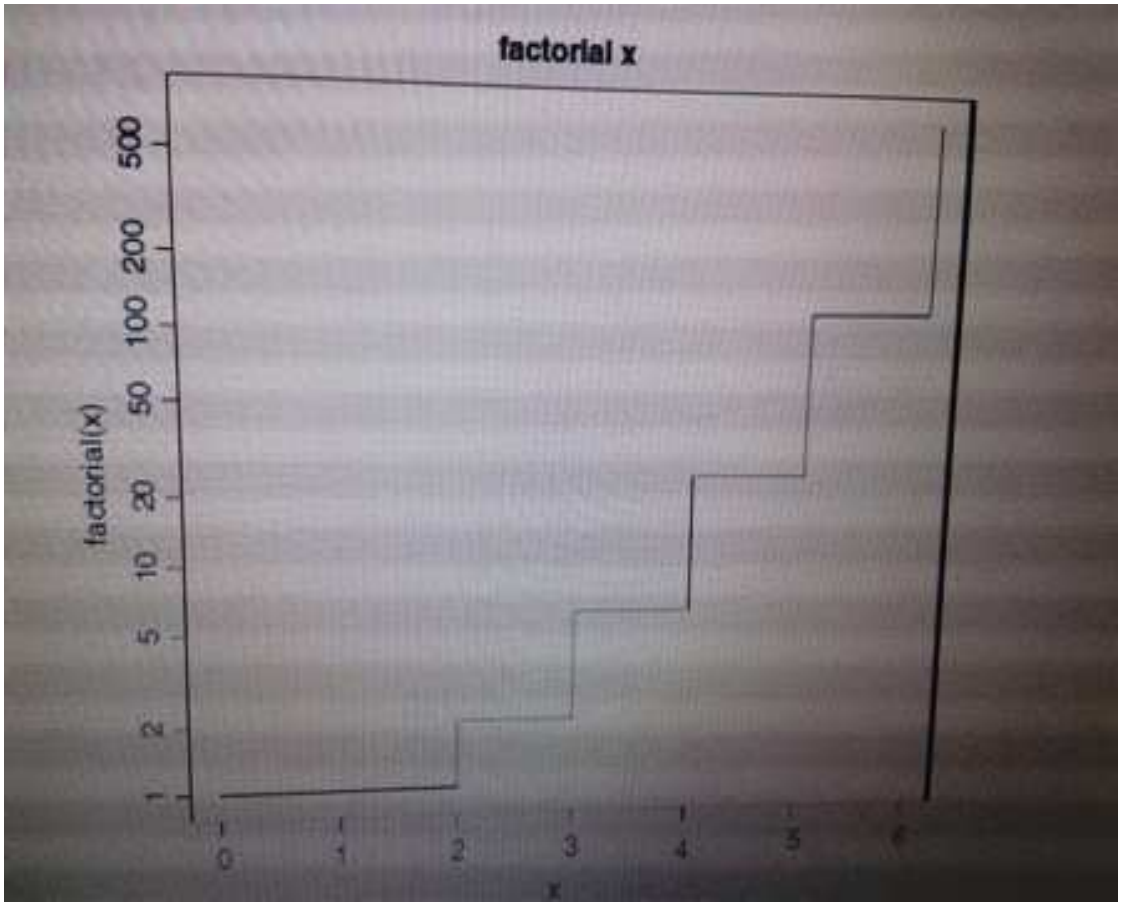
Συναρτήσεις πιθανότητας

Υπάρχουν πολλές συγκεκριμένες κατανομές πιθανότητας στο R (κανονική, Poisson, διωνυμική, κ.λπ.), και αυτές συζητούνται λεπτομερώς αργότερα. Εδώ θα δούμε τις βασικές μαθηματικές συναρτήσεις που ασχολούνται με στοιχειώδη πιθανότητα. Η συνάρτηση παραγοντικό δίνει τον αριθμό των μεταθέσεων από n αντικείμενα. Με πόσους τρόπους μπορούν να 4 στοιχεία να διευθετηθούν; Την πρώτη θέση θα μπορούσε να έχει οποιαδήποτε ένα από τα 4 στοιχεία σε αυτό, αλλά από τη στιγμή που έχουμε την ευκαιρία να επιλέξουμε το δεύτερο στοιχείο θα έχουμε ήδη καθορίσει λεπτομερώς το πρώτο στοιχείο, έτσι υπάρχουν μόνο 4-1=3 τρόποι για την επιλογή του δεύτερου στοιχείου. Υπάρχουν μόνο 4-2=2 τρόπους για την επιλογή το τρίτο στοιχείο, και από τη στιγμή που παίρνουμε στο τελευταίο στοιχείο δεν έχουμε βαθμούς ελευθερίας σε όλα: ο τελευταίος αριθμός πρέπει να είναι το ένα στοιχείο από τις τέσσερις που δεν έχουν χρησιμοποιηθεί σε θέσεις 1, 2 ή 3. Έτσι, με 4 είδη η απάντηση
είναι 4 ×(4-1) ×(4-2) ×(4-3) η οποία είναι 4 × 3 × 2 × 1 = 24 . Σε γενικές γραμμές, η συνάρτηση παραγοντικό(n) είναι δεδομένη από

$$n!=n(n-1)(n-2)*3*2.$$

Η συνάρτηση R είναι παραγοντική και μπορούμε στο γράφημα των τιμών του x από 0 έως 10 χρησιμοποιώντας το βήμα επιλογή τύπου = "s", σε γραφική παράσταση με λογαριθμική κλίμακα για τον άξονα y log = "y",

```
x<-0:6
plot(x,factorial(x),type="s",main="factorial x",log="y")
```



Η άλλη σημαντική βασική συνάρτηση για τον υπολογισμό της πιθανότητας της R είναι να διαλέξεις συνάρτηση που να υπολογίζει διώνυμους συντελεστές. Αυτά δείχνουν τον αριθμό των τρόπων που υπάρχουν από την επιλογή των x στοιχείων από n στοιχεία, όταν το στοιχείο μπορεί να είναι ένα από μόλις δύο τύπους (π.χ. είτε αρσενικό ή θηλυκό, μαύρο ή άσπρο, φερέγγυος ή αφερέγγυος). Ας υποθέσουμε ότι έχουμε 8 άτομα και θέλουμε να γνωρίζουμε πόσοι τρόποι υπάρχουν που 3 από αυτά θα μπορούσε να είναι άνδρες (και ως εκ τούτου 5 αυτά γυναίκες). Η απάντηση δίνεται από

$$\begin{pmatrix} n \\ x \end{pmatrix} = n! : x!(n-x)!,$$

έτσι με n=8 και x=3 παίρνουμε

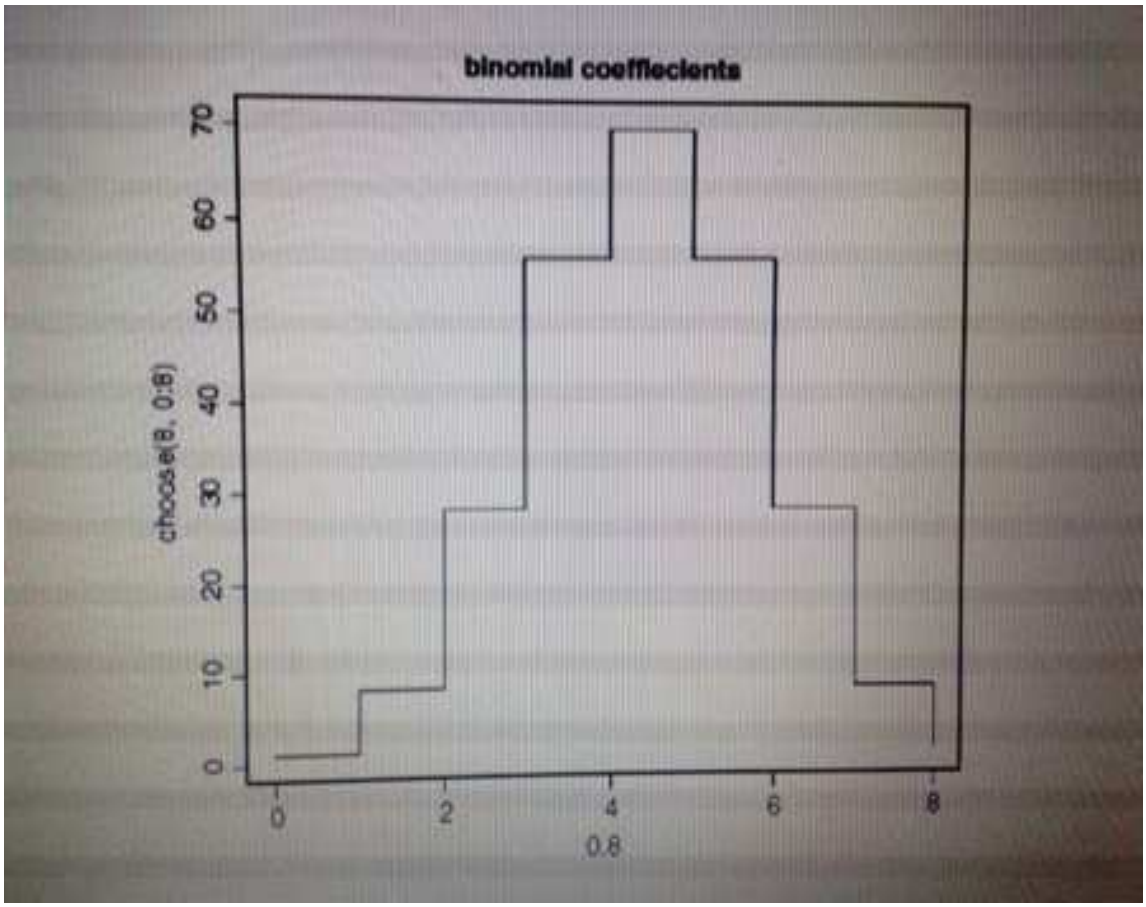
$$\begin{pmatrix} n \\ x \end{pmatrix} = 8! : 3!(8-3)! = 8*7*6:3*2=56$$

και μέσα στην R διαλέγουμε (8,3)

```
[1] 56
```

Προφανώς, υπάρχει μόνο ένας τρόπος που όλα τα 8 άτομα θα μπορούσαν να είναι άνδρας ή γυναίκα, έτσι υπάρχει είναι μόνο ένας τρόπος για να πάρει 0 ή 8 «επιτυχίες». Ένας άνδρας θα μπορούσε να είναι το πρώτο άτομο που

επιλέξατε, ή το δεύτερο ή το τρίτο, και ούτω καθεξής. Έτσι, υπάρχουν 8 τρόποι επιλογής 1 από 8.
Με το ίδιο σκεπτικό, πρέπει να υπάρχουν 8 τρόποι επιλογής 7 άνδρες από 8 άτομα (θα μπορούσε να είναι η μοναδική γυναίκα σε μία οποιαδήποτε από τις οκτώ θέσεις). Οι παρακάτω είναι μια γραφική παράσταση του αριθμού των τρόπων για την επιλογή των 0-8 ανδρών από 8 άτομα:
plot(0:8,choose(8,0:8),type="s",main="binomial coefficients")
[γραφική παράσταση (0:8, επιλέξτε (8,0:8), τύπος = "s", κυρίως = "διωνυμικοί συντελεστές")]



Συνεχείς Κατανομές Πιθανοτήτων
Το R έχει ένα ευρύ φάσμα από ενσωματωμένες κατανομές πιθανότητας, για κάθε μια από τις τέσσερις συναρτήσεις οι οποίες είναι διαθέσιμες: η συνάρτηση πυκνότητας πιθανότητας (η οποία έχει πρόθεμα d.)?, η αθροιστική πιθανότητα (p)? οι ποσοστημορίων της κατανομής (q)? και τυχαίων αριθμών που παράγονται από τη διανομή (r). Κάθε γράμμα μπορεί να είναι το πρόθεμα για τα ονόματα των συναρτήσεων R στον πίνακα 7.1 (π.χ. dbeta).

Πίνακας 7.1. Οι κατανομές πιθανοτήτων που υποστηρίζονται από την R. τις έννοιες της οι παράμετροι εξηγούνται στο κείμενο.

Συνάρτηση R	**Διανομή**	**Παράμετροι**
beta	Βήτα	Μορφή1, Μορφή 2
binom	Διωνυνική	μέγεθος του δείγματος, πιθανότητα
cauchy	Cauchy	Θέση, κλίμακα
exp	Εκθετική συνάρτηση	Τιμή (επιλεκτική)
chisq	chi-τετράγωνο	βαθμούς ελευθερίας
f	F του ψαρά	df1, df2
gamma	γάμμα	σχήμα
geom	γεωμετρικός	πιθανότητα
hyper	υπεργεωμετρική	m,n,k
lnorm	λογαριθμική	μέση τιμή, τυπική απόκλιση
logis	συμβολική λογική	θέση, κλίμακα
nbinom	αρνητική διωνυμική	το μέγεθος, η πιθανότητα
norm	κάθετη	μέση τιμή, τυπική απόκλιση
pois	Poisson	σημαίνει
signrank	Wilcoxon προσημασμένος βαθμός στατιστικής μεταβλητής	μέγεθος του δείγματος
t	t του σπουδαστή	βαθμούς ελευθερίας
unif	ομοειδής	ελάχιστος, μέγιστος (οπτικά)
weibull	Weibull	σχήμα
wilcox	Wilcoxon βαθμός αθροίσματος	m,n

Η αθροιστική συνάρτηση πιθανότητας είναι μια απλή έννοια: είναι μία καμπύλη σχήματος S που δείχνει, για κάθε τιμή του x, τη πιθανότητα απόκτησης μιας τιμής δείγματος που είναι μικρότερη ή ίση με x. Εδώ είναι τι μοιάζει με την κανονική κατανομή:

```
curve(pnorm(x),-3,3)
arrows(-1,0,-1,pnorm(-1),col="red")
arrows(-1,pnorm(-1),-3,pnorm(-1),col="green")
```

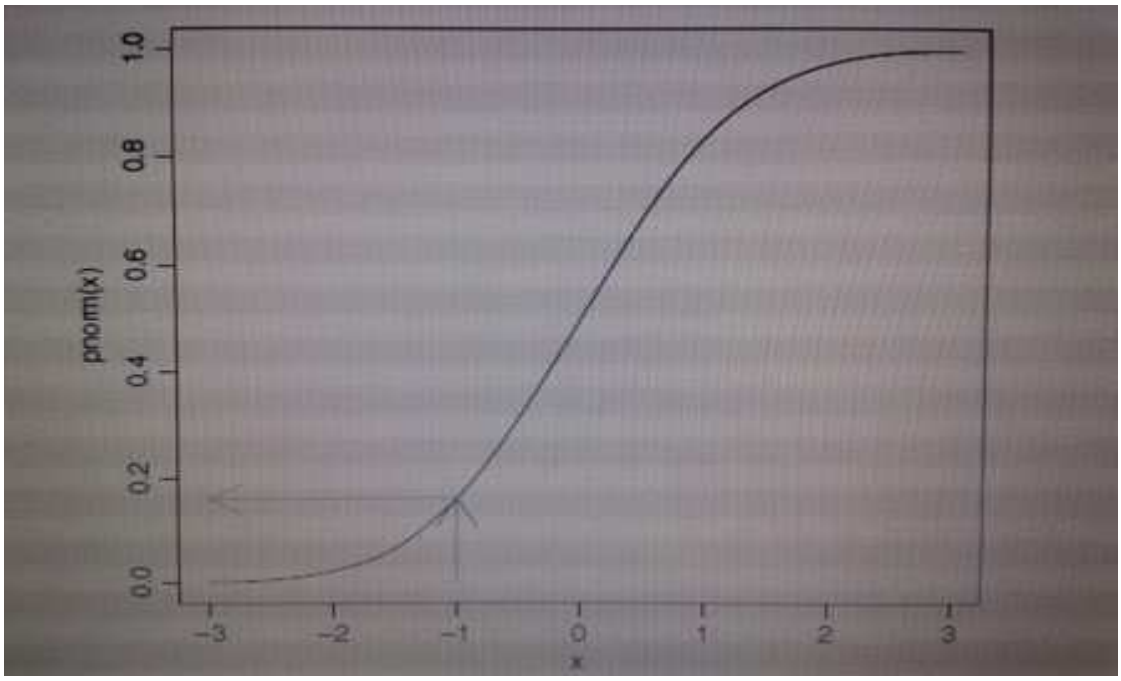
Η τιμή του x -1 οδηγεί μέχρι την αθροιστική πιθανότητα (κόκκινο βέλος) και η πιθανότητα συνδέεται με την απόκτηση τιμής αυτού του μεγέθους (-1) ή μικρότερη είναι επί του άξονα y (πράσινο βέλος).

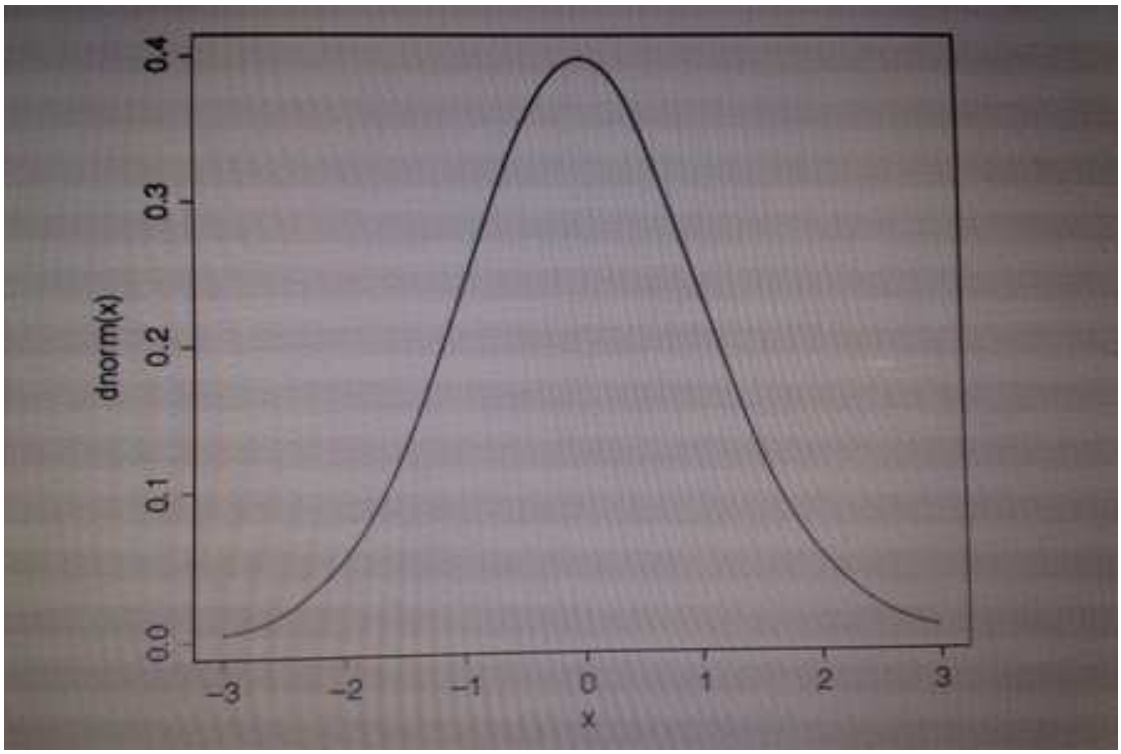
Η τιμή στον άξονα y είναι 0,158 655 3:

```
pnorm(-1)
[1] 0.1586553
```

Η πυκνότητα πιθανότητας είναι η κλίση της καμπύλης αυτής («παράγωγος» του). Μπορείτε να δείτε αμέσως ότι η κλίση δεν είναι ποτέ αρνητική. Η κλίση ξεκινάει από πολύ ρηχά μέχρι περίπου x = -2, αυξήσεις μέχρι μια κορυφή (στο x = 0 σε αυτό το παράδειγμα), τότε γίνεται πιο ρηχή, και γίνεται πολύ μικρό πράγματι παραπάνω για x = 2. Εδώ είναι ό, τι η συνάρτηση πυκνότητας της κανονικής (dnorm) μοιάζει με:

```
curve(dnorm(x),-3,3)
```

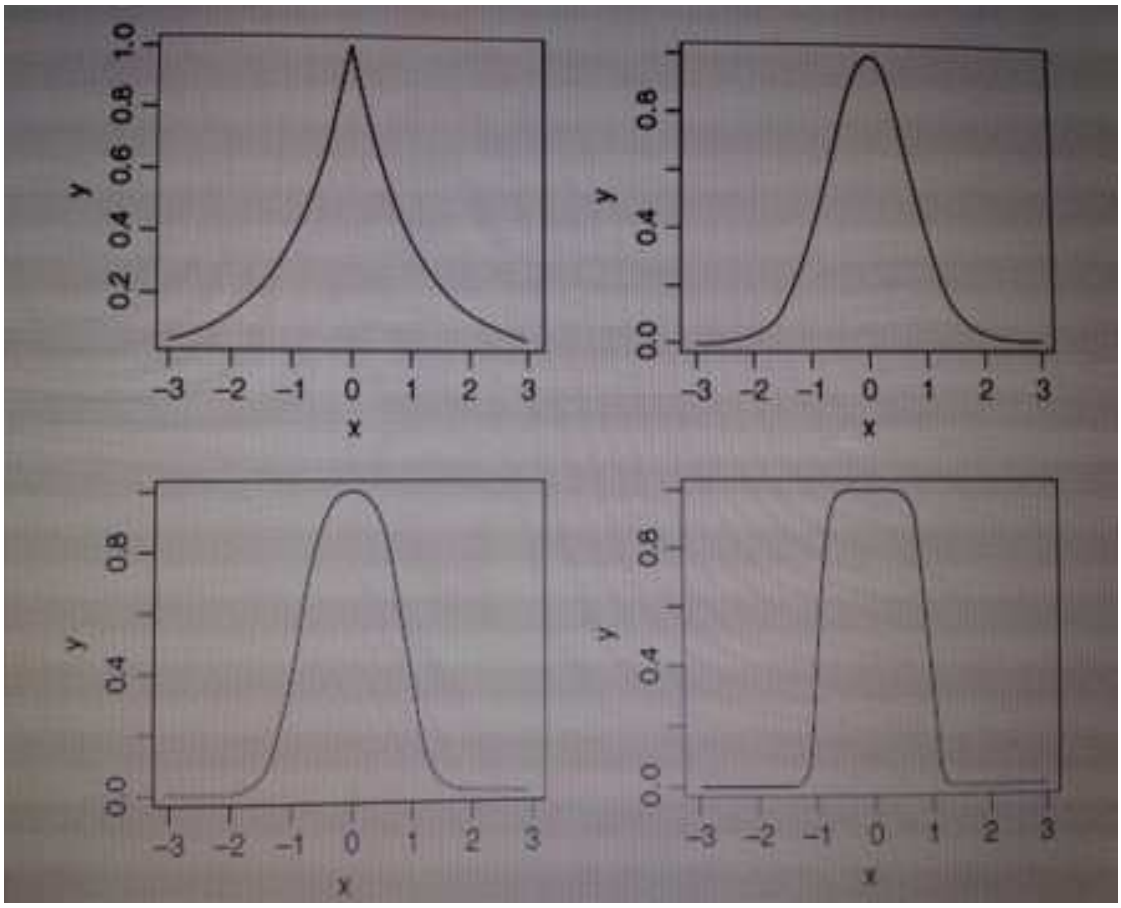

pnorm(x)
0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
3
x


dnorm(x)
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
x

Για μια διακριτή τυχαία μεταβλητή, όπως τη Poisson ή τη διωνυμική, η πυκνότητα πιθανότητας συνάρτησης είναι απλή: είναι απλά ένα ιστόγραμμα με τον άξονα y να ανεβαίνει ως πιθανότητες και όχι ως μετρήσεις, και τις διακριτές τιμές του x (0, 1, 2, 3,……) στον οριζόντιο άξονα.
Αλλά για μια συνεχή τυχαία μεταβλητή, ο ορισμός της συνάρτησης πυκνότητας πιθανότητας είναι πιο λογικός: δεν έχει πιθανότητες στον άξονα y, αλλά μάλλον την παράγωγο (την κλίση) της συνάρτησης αθροιστική πιθανότητα σε μία δεδομένη τιμή του x.

Κανονική κατανομή
Αυτή η κατανομή είναι κεντρικής σημασίας για τη θεωρία των παραμετρικών στατιστικών. Ας εξετάσουμε την ακόλουθη απλή εκθετική συνάρτηση:
y=exp(-|x|^m).
Καθώς η δύναμη (m) αυξάνεται στον εκθέτη, η συνάρτηση γίνεται όλο και περισσότερο σαν μια βηματική συνάρτηση. Οι παρακατω πινακες δειχνουν τη σχεση ανάμεσα στο x και το y για m=1,2,3 και 8, αντιστοιχως:

```
par(mfrow=c(2,2))
x<-seq(-3,3,0.01)
y<-exp(-abs(x))
plot(x,y,type="l")
y<-exp(-abs(x)^2)
plot(x,y,type="l")
y<-exp(-abs(x)^3)
plot(x,y,type="l")
y<-exp(-abs(x)^8)
plot(x,y,type="l")
```



Το δεύτερο από αυτά τα πάνελ (επάνω δεξιά), όπου y = exp(-x^2), είναι η βάση μιας εξαιρετικά σημαντικής και γνωστής συνάρτησης πυκνότητας πιθανότητας. Μόλις έχει κλιμακωθεί, έτσι ώστε το ολοκλήρωμα (η περιοχή κάτω από την καμπύλη από -∞ έως +∞) είναι η ενότητα, αυτή είναι η κανονική κατανομή.
Δυστυχώς, οι σταθερές κλιμάκωσης είναι μάλλον δυσκίνητες. Όταν η κατανομή έχει μέση 0 και τυπική απόκλιση 1 (η τυπική κανονική κατανομή), η εξίσωση γίνεται:
f(z)=(1/[(2π)^(1/2)])*e^(-z^2:2).
Ας υποθέσουμε ότι έχουμε μετρήσει τα ύψη από 100 ανθρώπους. Το μέσο ύψος ήταν 170 cm και η τυπική απόκλιση ήταν 8 cm (άνω αριστερό πλαίσιο, παρακάτω). Μπορούμε να ρωτήσουμε τρία είδη από ερωτήσεις για δεδομένα όπως αυτά: ποια είναι η πιθανότητα ότι ένα τυχαία επιλεγμένο άτομο θα είναι:

• κοντότερος από ένα συγκεκριμένο ύψος;
• ψηλότερος από ένα συγκεκριμένο ύψος;
• μεταξύ ενός συγκεκριμένου ύψους και ενός άλλου;

Η περιοχή κάτω από την πλήρη καμπύλη είναι ακριβώς 1; Ο καθένας έχει ένα ύψος μεταξύ μείον άπειρο και το συν άπειρο. Είναι αλήθεια, αλλά δεν είναι ιδιαίτερα χρήσιμο. Ας υποθέσουμε ότι θέλουμε να γνωρίζουμε την πιθανότητα ότι ένας από τους ανθρώπους μας, που επιλέχτηκε τυχαία από την ομάδα, θα

είναι λιγότερο από 160 εκατοστά ύψος. Πρέπει να μετατρέψουμε αυτό το ύψος σε μια τιμή z; Δηλαδή, θα πρέπει να μετατρέψουμε 160 εκατοστά σε μια μέτρηση από τυπικές αποκλίσεις από την μέση τιμή. Τι γνωρίζουμε για τη πρότυπη κανονική κατανομή; Έχει μια μέση τιμή μηδέν και τυπική απόκλιση 1. Έτσι, μπορούμε να μετατρέψουμε κάθε τιμή y, από μια κατανομή με μέση τιμή $\bar{y}$ και τυπική απόκλιση s πολύ απλά με υπολογισμό:

$z= (y- \bar{y})/s.$

Γι 'αυτό και τη μετατροπή 160 εκατοστών σε μια μέτρηση από τυπικές αποκλίσεις. Είναι λιγότερη από το μέσο ύψος (170 cm) έτσι ώστε η τιμή του θα είναι αρνητική:

$z= (160-170)/8=-1.25$

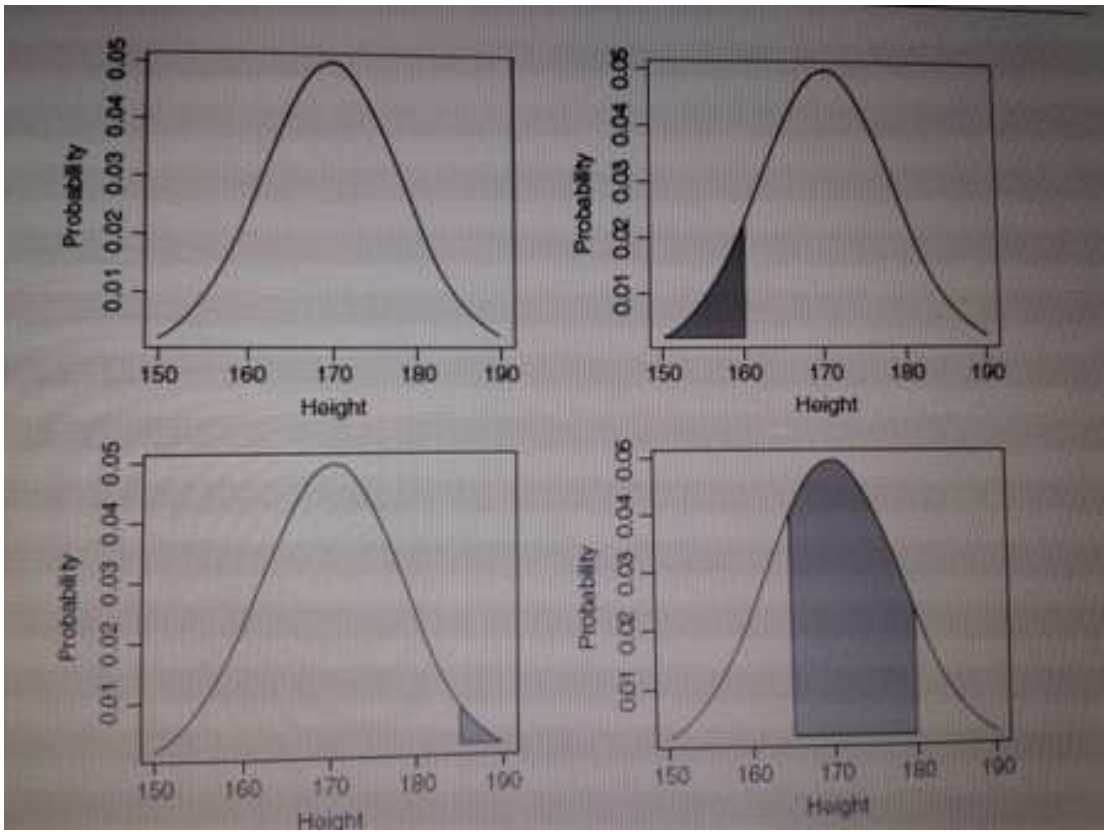
Τώρα πρέπει να βρούμε την πιθανότητα της τιμής της τυπικής κανονικής λαμβάνοντας μια τιμή -1,25 ή μικρότερη. Αυτή είναι η περιοχή κάτω από το αριστερό χέρι ουρά (το ολοκλήρωμα) της πυκνότητας συνάρτησης. Η συνάρτηση που χρειαζόμαστε για αυτό είναι pnorm: εμείς παρέχουμε αυτό με μια τιμή z (ή, πιο γενικά, με μέτρηση ποσότητας) και μας παρέχει την πιθανότητα που θέλουμε:

```
pnorm(-1.25)
[1] 0.1056498
```

Έτσι, η απάντηση στο πρώτο ερώτημα μας (η κόκκινη περιοχή, επάνω δεξιά) είναι μόλις πάνω από 10%. Στη συνέχεια, ποια είναι η πιθανότητα να επιλέξουμε έναν από τους ανθρώπους μας και να διαπιστώσουμε ότι αυτοί είναι ψηλότεροι των 185 cm (κάτω αριστερά); Τα πρώτα δύο μέρη της άσκησης είναι ακριβώς το ίδιο όπως πριν. Πρώτα μετατρέπουμε την τιμή μας από 185 εκατοστά σε μια σειρά από τυπικές αποκλίσεις:

$z=(185-170)/8=1.875$

Τότε θα ρωτήσουμε τι πιθανότητα συνδέεται με αυτό, χρησιμοποιώντας pnorm:




```
pnorm(1.875)
[1] 0.9696036
```

Αλλά αυτή είναι η απάντηση σε ένα διαφορετικό ζήτημα. Αυτή είναι η πιθανότητα ότι κάποιος θα είναι μικρότερος ή ίσος έως 185 εκατοστά ψιλός (δηλαδή ό, τι η pnorm συνάρτηση έχει γραφτεί για να το παρέχει). Το μόνο που χρειάζεται να κάνουμε είναι να επεξεργαστούμε το συμπλήρωμα αυτό:

```
1-pnorm(1.875)
[1] 0.03039636
```

Έτσι, η απάντηση στο δεύτερο ερώτημα είναι περίπου 3%.

Τέλος, θα μπορούσαμε να θέλουν να γνωρίζουμε την πιθανότητα επιλογής ενός προσώπου μεταξύ των 165 εκατοστών και 180 cm. Έχουμε λίγο περισσότερη δουλειά να κάνουμε εδώ, γιατί πρέπει να υπολογίσουμε δύο z τιμές:

z1= (165−170)/8=-0,625 και z2=(180-170)/8=1,25.

Το σημαντικό σημείο που πρέπει να κατανοήσετε είναι το εξής: θέλουμε την πιθανότητα της επιλογής ενός προσώπου μεταξύ των δύο αυτών τιμών z, έτσι ώστε να αφαιρέσουμε τη μικρότερη πιθανότητα από τη μεγαλύτερη πιθανότητα:

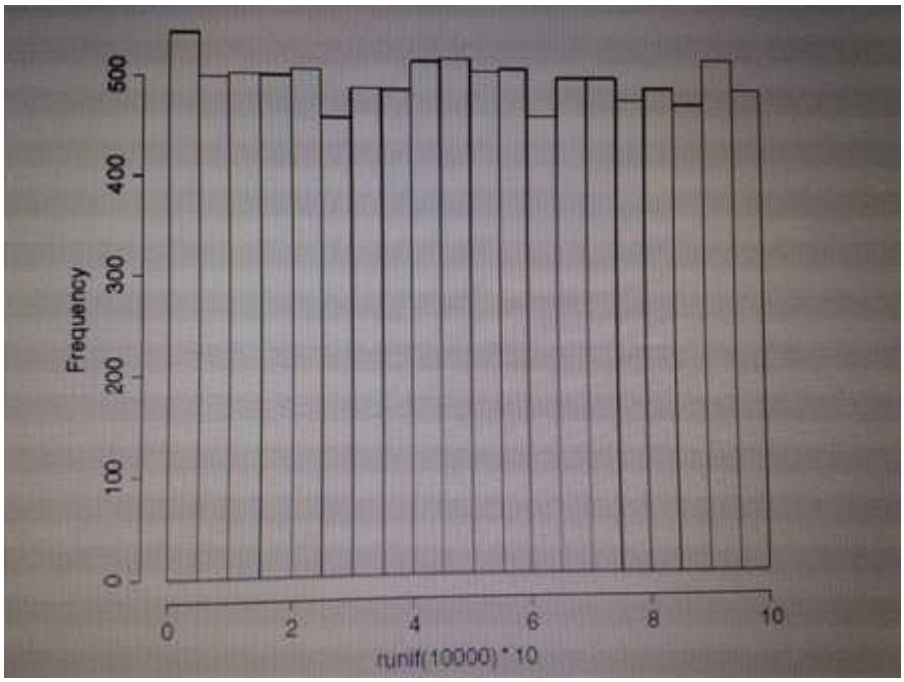
```
pnorm(1.25)-pnorm(-0.625)
[1] 0.6283647
```

Έτσι έχουμε 63% πιθανότητες επιλογής ενός ατόμου μεσαίου μεγέθους (ψηλότερο των 165 cm και κοντότερο από 180 cm) από το δείγμα αυτό με ένα μέσο ύψος 170 cm και τυπική απόκλιση 8 cm (κάτω δεξιά, παραπάνω).

Το κεντρικό οριακό θεώρημα

Εάν πάρετε επαναλαμβανόμενα δείγματα από έναν πληθυσμό με πεπερασμένη διακύμανση και να υπολογίσετε τους μέσους όρους, τότε οι μέσοι όροι θα πρέπει να διανέμονται κανονικά. Αυτό ονομάζεται το **κεντρικό οριακό θεώρημα**. Ας το αποδείξει για τους εαυτούς μας. Μπορούμε να πάρουμε πέντε ομοιόμορφα κατανεμημένους τυχαίους αριθμούς μεταξύ 0 και 10 και υπολογίζουμε το μέσο όρο. Ο μέσος όρος θα είναι χαμηλός όταν εμείς πάρουμε, ας πούμε, 2,3,1,2,1 και μεγάλος όταν παίρνουμε 9,8,9,6,8. Τυπικά, βέβαια, ο μέσος όρος θα είναι κοντά στο 5. Ας κάνουμε αυτό 10 000 φορές και να εξετάσουμε την κατανομή των 10 000 αριθμητικών μέσων. Τα δεδομένα είναι ορθογώνια (ομοιόμορφα) κατανέμονται στο διάστημα 0 έως 10, έτσι ώστε η κατανομή των ανεπεξέργαστων δεδομένων θα πρέπει να είναι επίπεδες κορυφές:

```
hist(runif(10000)*10,main="")
```

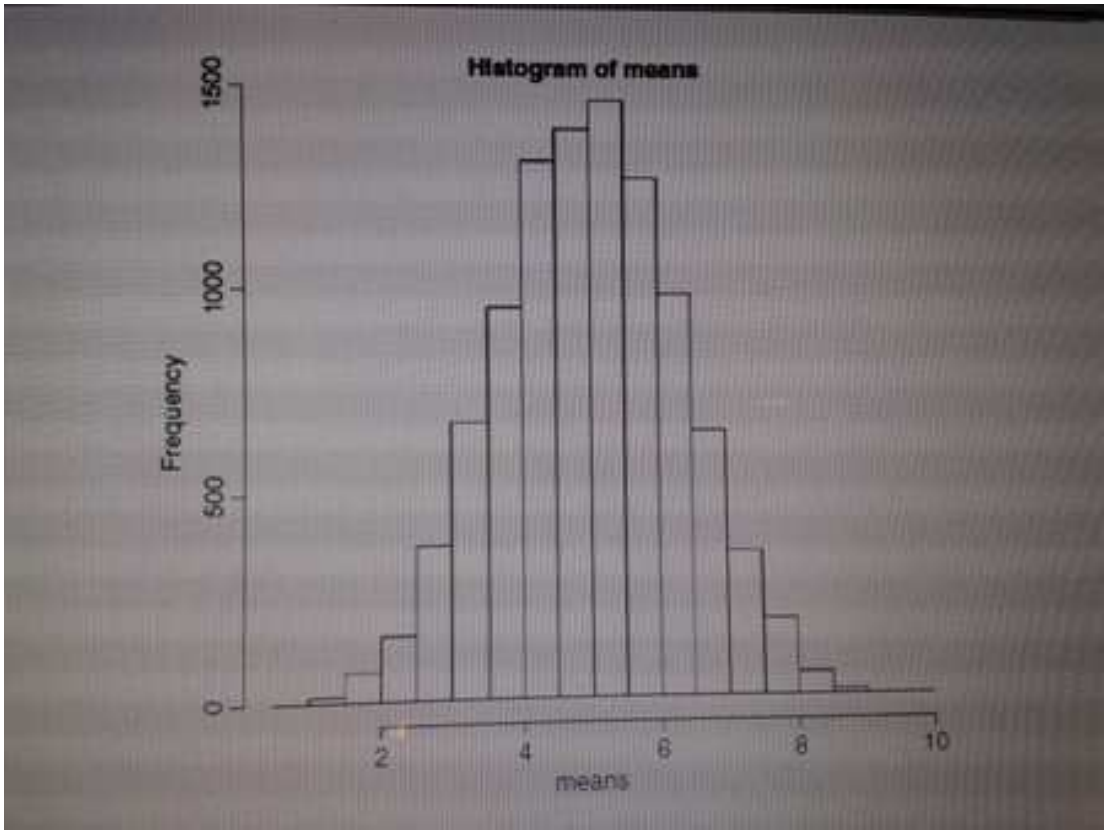


Τι γίνεται με την κατανομή των μέσων βάσει δείγματος, λαμβάνοντας μόλις 5 ομοιόμορφα κατανεμημένους τυχαίους αριθμούς;

```
means<-numeric(10000)

for (i in 1:10000){

means[i]<-mean(runif(5)*10)
```

```
}

hist(means,ylim=c(0,1600))
```

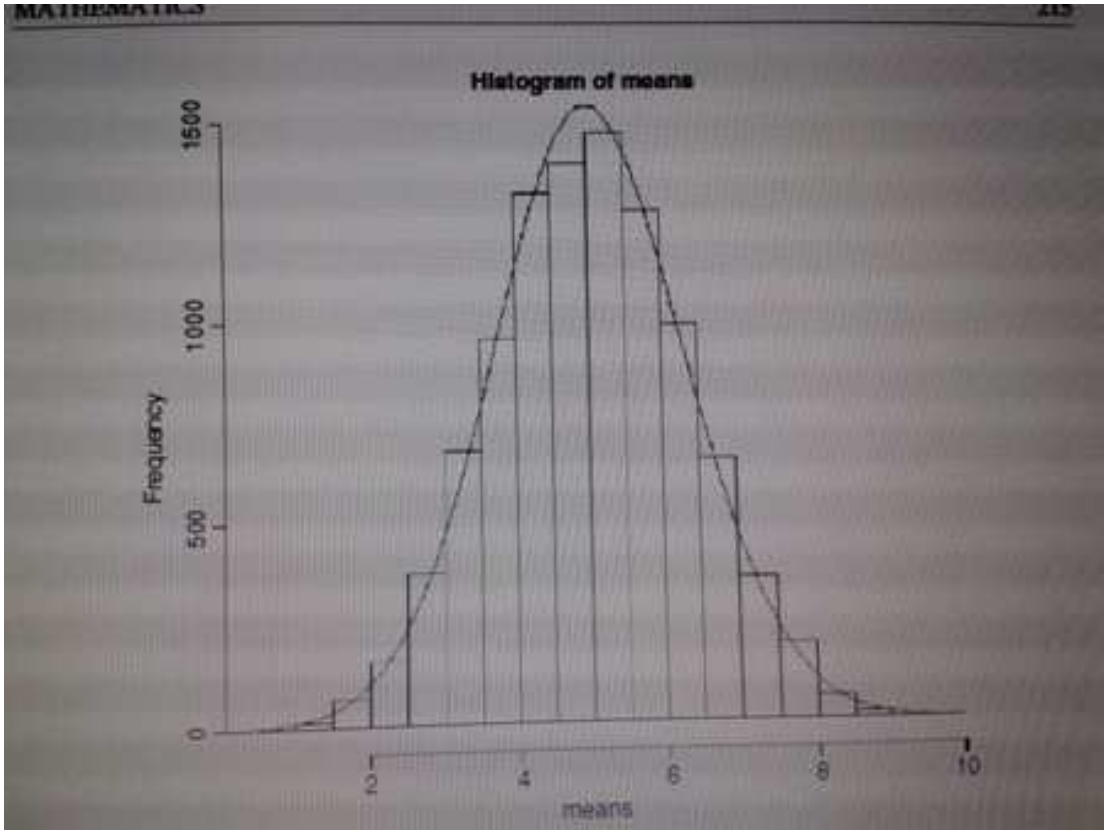
Ωραία, αλλά πόσο κοντά είναι αυτό σε μια κανονική κατανομή; Μια δοκιμή είναι να σχεδιάσετε μια κανονική κατανομή με τις ίδιες παραμέτρους στην κορυφή του ιστογράμματος. Αλλά τι είναι αυτές οι παράμετροι; Το κανονικό είναι δύο παραμέτρων κατανομή που έχει χαρακτηριστεί από τη μέση και την τυπική της απόκλιση. Μπορούμε να υπολογίσουμε τις δύο αυτές παραμέτρους από το δείγμα μας των 10 000 αριθμητικών μέσων (οι τιμές σας θα είναι ελαφρώς διαφορετικές λόγω της τυχαιοποίησης):




```
mean(means)
[1] 4.998581
sd(means)
[1] 1.289960
```

Τώρα έχουμε χρησιμοποιήσει αυτές τις δύο παραμέτρους της συνάρτησης πυκνότητας πιθανότητας της κανονικής κατανομής (dnorm) για να δημιουργήσουμε μια κανονική καμπύλη με ιδιαίτερο αριθμητικό μέσο μας και την τυπική απόκλιση. Για να σχεδιάσουμε την ομαλή γραμμή της κανονικής καμπύλης, θα πρέπει να δημιουργήσουμε μια σειρά τιμών για τον άξονα X? επιθεώρηση των ιστογράμματων δείχνουν ότι η λογική ορίων θα είναι από 0 έως 10 (τα όρια που επιλέξαμε για ομοιόμορφα κατανεμημένους τυχαίους αριθμούς μας). Ένας καλός εμπειρικός κανόνας είναι ότι για μια ομαλή καμπύλη θα χρειάζεται τουλάχιστον 100 τιμές, οπότε ας προσπαθήσουμε αυτό:

xv<-seq(0,10,0.1)
Υπάρχει μόνο ένα πράγμα που αφήνεται για να κάνει. Η συνάρτηση πυκνότητας πιθανότητας έχει αναπόσπαστο 1,0 (που είναι η περιοχή κάτω από την κανονική καμπύλη), αλλά είχαμε 10 000 δείγματα. Η κανονική κλίμακα συνάρτησης πυκνότητας πιθανότητας για την συγκεκριμένη περίπτωση μας, ωστόσο, εξαρτάται από το ύψος της υψηλότερης ράβδου (περίπου 1500 σε αυτήν την περίπτωση). Το ύψος, με τη σειρά του, εξαρτάται από το επιλεγέν πλάτος κουτιού? εάν διπλασιαστεί με το πλάτος του κουτιού θα υπάρχουν περίπου διπλάσιοι αριθμοί στο κουτί και η μπάρα θα είναι διπλάσια σε ύψος στον άξονα y. Για να πάρετε το ύψος των ράβδων για μια κλίμακα συχνότητας, ως εκ τούτου, πολλαπλασιάζουμε την συνολική συχνότητα, 10 000 από το πλάτος κουτιού, επί 0.5 για να πάρει 5000. Έχουμε πολλαπλασιάσει 5000 από την πυκνότητα πιθανότητας να πάρουμε το ύψος της καμπύλης. Τέλος, χρησιμοποιούμε τις γραμμές για να επικαλύψουν την ομαλή καμπύλη στο ιστόγραμμα μας:
yv<-dnorm(xv,mean=4.998581,sd=1.28996)*5000
lines(xv,yv)



Η εφαρμογή είναι εξαιρετική. Το κεντρικό οριακό θεώρημα λειτουργεί πραγματικά. Σχεδόν κάθε διανομή, έστω και μια «κακή συμπεριφορά» μια όπως την ομοιόμορφη κατανομή εμείς δουλέψαμε μαζί εδώ, θα παράγει μια κανονική κατανομή του δείγματος που λαμβάνετε μέσα από αυτήν.
Ένα απλό παράδειγμα της λειτουργίας του κεντρικού οριακού θεωρήματος περιλαμβάνει η χρήση των ζαριών.

Ρίξτε μια πεθαίνουν πολλές φορές, και κάθε ένας από τους έξι αριθμούς θα πρέπει να καταλήξουμε εξίσου συχνά: αυτό είναι ένα παράδειγμα μιας ομοιόμορφης κατανομής:

```
par(mfrow=c(2,2))
hist(sample(1:6,replace=T,10000),breaks=0.5:6.5,main="",xlab="one die")
```

Τώρα ρίξε δύο ζάρια και να προσθέσετε τις βαθμολογίες μαζί: αυτό είναι το αρχαίο παιχνίδι των ζαριών. Εκεί είναι 11 πιθανά αποτελέσματα από τουλάχιστον 2 σε κατ 'ανώτατο όριο 12. Η πιο πιθανό σκορ είναι 7 επειδή υπάρχουν 6 τρόποι με τους οποίους αυτό θα μπορούσε να συμβεί:

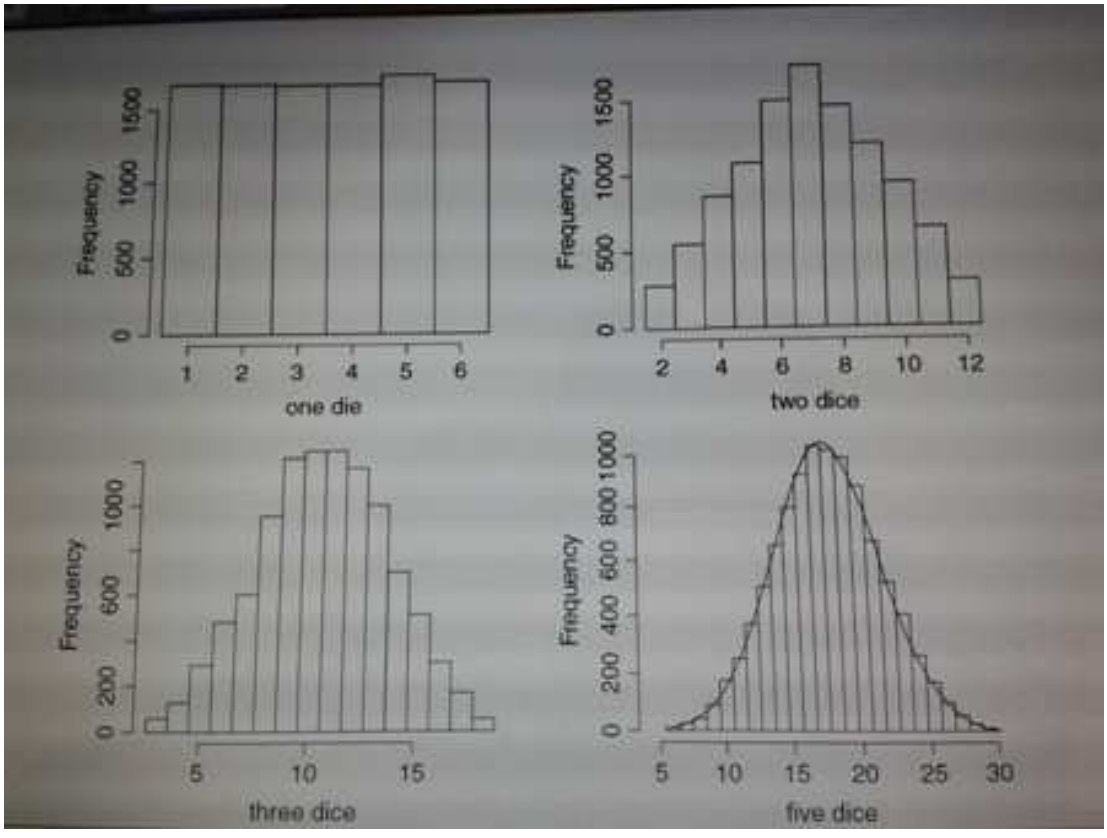
1,6 6,1 2,5 5,2 3,4 4,3

Για πολλές βολές ζαριών έχουμε μια τριγωνική κατανομή της βαθμολογίας, με επίκεντρο 7:

```
a<-sample(1:6,replace=T,10000)
b<-sample(1:6,replace=T,10000)
hist(a+b,breaks=1.5:12.5,main="", xlab="two dice")
```

Υπάρχει ήδη μια σαφή ένδειξη της κεντρικής τάσης και διάδοσης. Για τρία ζάρια παίρνουμε

```
c<-sample(1:6,replace=T,10000)
hist(a+b+c,breaks=2.5:18.5,main="", xlab="three dice")
```



και το σχήμα καμπάνας της κανονικής κατανομής έχει αρχίσει να αναδύεται. Μέχρι τη στιγμή που θα έχουμε σε πέντε ζάρια, τη διωνυμική κατανομή να είναι σχεδόν δυσδιάκριτη από την κανονική:

```
d<-sample(1:6,replace=T,10000)
e<-sample(1:6,replace=T,10000)
```

hist(a+b+c+d+e,breaks=4.5:30.5,main="", xlab="five dice")
Η ομαλή καμπύλη δίνεται από μια κανονική κατανομή με την ίδια μέση τιμή και τυπική απόκλιση:
mean(a+b+c+d+e)
[1] 17.5937
sd(a+b+c+d+e)
[1] 3.837668
lines(seq(1,30,0.1),dnorm(seq(1,30,0.1),17.5937,3.837668)*10000)

Μέγιστης πιθανότητας με την κανονική κατανομή

Η πυκνότητα πιθανότητας της κανονικής είναι:

f(y/μ,σ)=1:[σ*(2π)^1/2]*exp[-(y-μ)^2:2*(σ^2)],

το οποίο διαβάζεται όπως λέγεται η πιθανότητα του να πάρει ένα δεδομένο τιμή y , δεδομένου (|) μια μέση τιμή μ και μια μεταβλητή από σ^2, υπολογίζεται από αυτή τη μάλλον περίπλοκη εμφάνιση δύο παραμέτρων εκθετικής συνάρτησης. Για οποιοδήποτε δεδομένο συνδυασμό μ και σ^2 , δίνει μια τιμή μεταξύ 0 και 1. Υπενθυμίζουμε ότι η πιθανότητα είναι το προϊόν των πυκνοτήτων πιθανότητας, για κάθε μία από τις τιμές της μεταβλητής απόκρισης y. Έτσι, αν έχουμε n τιμές του y στο πείραμά μας, η συνάρτηση πιθανότητας είναι
L(μ,σ)=Π{1:[σ*(2π)^(1/2)]*exp[-(yi-μ)^2 :2(σ^2)]}, (όπου i=1 έως n)
όπου η μόνη αλλαγή είναι ότι το y έχει αντικατασταθεί από yi και πολλαπλασιάζουμε μαζί τις πιθανότητες για κάθε ένα από τα σημεία n δεδομένων. Υπάρχει ένα μικρό κομμάτι της άλγεβρας που μπορούμε να κάνουμε απλοποιήση αυτού: μπορούμε να απαλλαγούμε από το φορέα εκμετάλλευσης του προϊόντος, Π, σε δύο στάδια. Πρώτον, για το σταθερό όρο: ότι, πολλαπλασιάζεται με n φορές μόνη της, μπορεί απλά να γραφτεί ως 1:[σ*(2π)^(1/2)]^n. Δεύτερον, θυμήσου ότι το προϊόν ενός συνόλου αντιλογάριθμων (exp) μπορεί να γραφτεί ως ο αντιλογάριθμος του αθροίσματος των τιμών του xi όπως αυτό: Πexp(xi)=exp(Σxi). Αυτό σημαίνει ότι το προϊόν του δεξιού μέρους της έκφρασης μπορεί να γραφτεί ως
exp[-Σ(yi-μ)^2 :2(σ^2)], (όπου i=1 έως n)
ώστε να μπορούμε να ξαναγράψουμε την πιθανότητα της κανονικής κατανομής όπως:
L(μ,σ)= 1:[σ*(2π)^(1/2)]^n* exp[-1/2(σ^2)*Σ(yi-μ)^2], (όπου i=1 έως n)
Οι δύο παράμετροι μ και σ είναι άγνωστοι, και ο σκοπός της άσκησης είναι να χρησιμοποιήσεις στατιστικές μοντελοποίησης για τον καθορισμό μέγιστων τιμών πιθανότητας τους από τα δεδομένα (οι n διαφορετικές τιμές του y). Έτσι, πώς μπορούμε να βρούμε τις τιμές μ και σ που μεγιστοποιούν αυτή την πιθανότητα; Η απάντηση περιλαμβάνει λογισμό: Πρώτα βρίσκουμε την παράγωγο της συνάρτησης σε σχέση με τους παραμέτρους, έπειτα τη βάζουμε στο μηδέν, και τη λύνουμε.
Αποδεικνύεται ότι, λόγω της συνάρτησης exp στην εξίσωση, είναι ευκολότερο να λειτουργήσει ο log της πιθανότητας,

l(μ,σ)=-n/2*log(2π)-nlog(σ)-Σ(yi-μ)^2/2(σ^2),

και να μεγιστοποιήσει αυτό αντ 'αυτού. Προφανώς, οι τιμές των παραμέτρων που μεγιστοποιούν τη log-πιθανότητα l(μ,σ)=log L(μ,σ) θα είναι οι ίδιες με εκείνες που μεγιστοποιούν την πιθανότητα. Από τώρα και στο εξής, θα υποθέσουμε ότι το άθροισμα είναι πάνω από το δείκτη i από 1 έως n.
Τώρα για την ανάλυση(μαθηματική). Ξεκινάμε με τη μέση τιμή, μ. Η παράγωγος της log πιθανότητας σε σχέση με το μ είναι
dl/dμ=Σ(yi-μ)/σ^2.
Βάλε το παράγωγο στο μηδέν και λύσε την ως προς μ:
Σ(yi-μ)/σ^2=0 έτσι Σ(yi-μ)=0.
Παίρνουμε το άθροισμα απ' ευθείας το βάζουμε σε παρενθέσεις, και παρατηρούμε ότι Σμ=nμ,
Σyi-nμ=0 έτσι Σyi=nμ και μ= Σyi/n,
Η εκτίμηση μέγιστης πιθανότητας του μ είναι ο αριθμητικός μέσος όρος.
Στη συνέχεια βρίσκουμε την παράγωγο της log-πιθανότητας σε σχέση με το σ:
dl/dσ=-n/σ+Σ(yi-μ)^2/σ^3,
υπενθυμίζοντας ότι η παράγωγος logx είναι 1 / x και η παράγωγος της -1/x^2 είναι 2/x^3. Επίλυνοντας, παίρνουμε
-n/σ+Σ(yi-μ)^2/σ^3=0 έτσι Σ(yi-μ)^2= σ^3*(n/σ)= σ^2*n, οπότε: σ^2= Σ(yi-μ)^2/n.
Η μέγιστη εκτίμηση της πιθανότητας της μεταβλητής σ^2 είναι η μέση τετραγωνική απόκλιση από τις y τιμές από την μέση τιμή. Αυτή είναι μια προκατειλημμένη εκτίμηση της μεταβλητής, ωστόσο, επειδή κάνεις δεν λαμβάνει υπόψη το γεγονός ότι εκτιμήσαμε την τιμή του μ από τα δεδομένα. Για να ευρύνουμε την εκτίμηση, θα πρέπει να χάσουμε 1 βαθμό ελευθερίας ώστε να αντικατοπτρίζει το γεγονός αυτό, και να μοιραστούν το ποσό των τετράγωνων από n-1 και όχι από n (βλ. σελ. 52. και περιορισμένη μέγιστη πιθανότητα εκτίμησης στο κεφάλαιο 19).
Εδώ, εμείς επιλεγούμε η συνάρτηση πιθανότητας R είναι ενσωματωμένη στο πλαίσιο της κανονικής κατανομής. Η dnorm συνάρτηση πυκνότητας έχει την τιμή z (μια ποσοτικοποίηση) ως το όρισμα της. Επιλεκτικά ορίσματα προδιαγράφουν τη μέση τιμή και την τυπική απόκλιση (η προεπιλογή είναι η τυπική κανονική με μέση τιμή 0 και τυπική απόκλιση 1). Οι τιμές του z έξω από το πεδίο τιμών της συνάρτησης -3.5 σε +3.5 είναι πολύ απίθανο.

```
par(mfrow=c(2,2))
curve(dnorm,-3,3,xlab="z",ylab="Probability density",main="Density")
```

Η pnorm συνάρτηση πιθανότητας έχει επίσης μια τιμή z (μια ποσοτικοποίηση) ως το όρισμα της. Επιλεκτικά ορίσματα προδιαγράφουν τη μέση τιμή και την τυπική απόκλιση (η προεπιλογή είναι η τυπική κανονική με μέση τιμή 0 και τυπική απόκλιση 1). Δείχνει την αθροιστική πιθανότητα της τιμής z λιγότερη ή ίση από την καθορισμένη τιμή, και είναι μια καμπύλη σχήματος S:

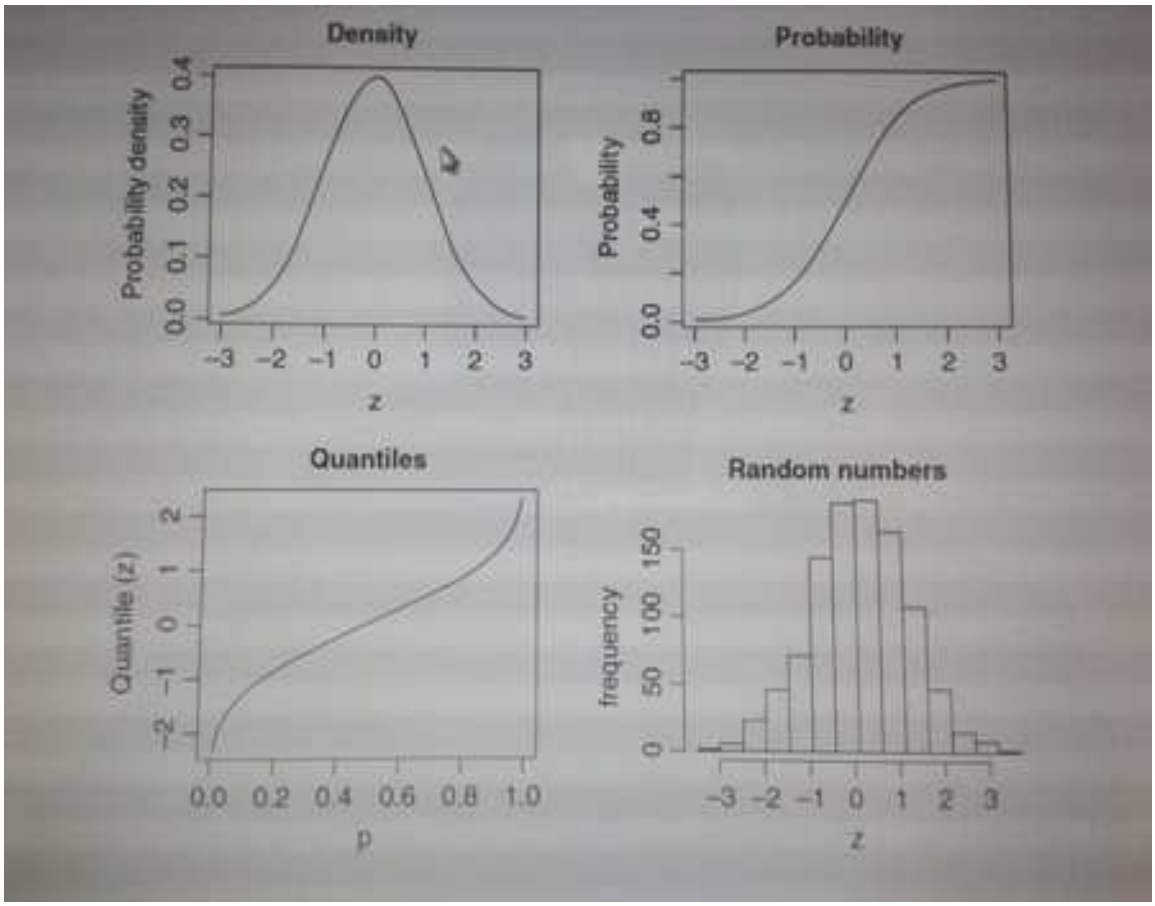
```
curve(pnorm,-3,3,xlab="z",ylab="Probability",main="Probability")
```

Ποσοτικοποιήσεις της κανονικής qnorm συνάρτησης έχουν αθροιστική πιθανότητα ως το όρισμα τους. Εκτελούν την αντίστροφη συνάρτηση του pnorm, επιστρέφοντας την τιμή του z, όταν προβλέπεται με πιθανότητα.

```
curve(qnorm,0,1,xlab="p",ylab="Quantile (z)",main="Quantiles")
```

Η κανονική τυχαία κατανομή εντολή rnorm γεννήτρια αριθμού παράγει τυχαίους πραγματικούς αριθμούς από μια κατανομή που καθορίζεται με μέση τιμή και τυπική απόκλιση. Το πρώτο όρισμα είναι ένας αριθμός των αριθμών

που θέλετε να παραχθούν: εδώ είναι 1000 τυχαίοι αριθμοί με μέση τιμή 0 και τυπική απόκλιση 1:



Οι τέσσερις συναρτήσεις (d, ρ, q και r) εργάζονται με παρόμοιους τρόπους με όλες τις άλλες διανομές πιθανότητας.

Δημιουργία τυχαίων αριθμών με την ακριβή μέση τυπική απόκλιση

Εάν χρησιμοποιείτε μια γεννήτρια τυχαίων αριθμών, όπως εντολή rnorm στη συνέχεια, φυσικά, το δείγμα που παράγετε δεν θα έχει ακριβώς την μέση τιμή και την τυπική απόκλιση που καθορίζετε, και δύο πίστες θα παράγουν φορείς με διαφορετικές μέσες τιμές και τυπικές αποκλίσεις. Υποθέστε ότι θέλουμε 100 συνηθισμένους τυχαίους αριθμούς με μέση τιμή ακριβώς 24 και μια τυπική απόκλιση με ακρίβεια 4:

```
yvals<-rnorm(100,24,4)
mean(yvals)
[1] 24.2958
sd(yvals)
[1] 3.5725
```

Ακριβής, αλλά δεν είναι σωστοί. Αν θέλετε να δημιουργήσετε τυχαίους αριθμούς με μια ακριβής μέση τιμή και τυπική απόκλιση, τότε κάντε τα εξής:

```
ydevs<-rnorm(100,0,1)
```

Τώρα αντισταθμίζεται το γεγονός ότι η μέση τιμή δεν είναι ακριβώς 0 (μηδέν) και η τυπική απόκλιση είναι δεν είναι ακριβώς 1 εκφράζοντας όλες τις τιμές ως αποκλίσεις από την μέση τιμή δείγματος που κλιμακώνεται σε μονάδες των αποκλίσεων προτύπου δείγματος:

```
ydevs<-(ydevs-mean(ydevs))/sd(ydevs)
```

Ελέγξτε ότι η μέση τιμή είναι μηδέν και η τυπική απόκλιση είναι ακριβώς 1:

```
mean(ydevs)
[1] -2.449430e-17
sd(ydevs)
[1] 1
```

Η μέση τιμή είναι όσο πιο κοντά στο μηδέν, όπως δεν κάνει καμία διαφορά, και η τυπική απόκλιση είναι ένα. Τώρα πολλαπλασιάστε αυτό το διάνυσμα από την επιθυμητή τυπική απόκλιση σας και προσθέσετε στην επιθυμητή μέση σας τιμή να πάρετε ένα δείγμα με ακριβώς την απαιτούμενη μέση τιμή και τυπική απόκλιση:

```
yvals<-24 + ydevs*4
mean(yvals)
[1] 24
sd(yvals)
[1] 4
```

Συγκρίνοντας τα δεδομένα με μια κανονική κατανομή

Διάφορα τεστ για ομαλότητα που περιγράφεται στη σελ.. 281. Εδώ ασχολούμαστε με το έργο της συγκρίσεως ενός ιστογράμματος πραγματικών δεδομένων με μια ομαλή κανονική κατανομή με την ίδια μέση τιμή και την τυπική απόκλιση, προκειμένου να ψάξουν για στοιχεία για μη κανονικότητα (π.χ. λοξότητα ή κύρτωση).

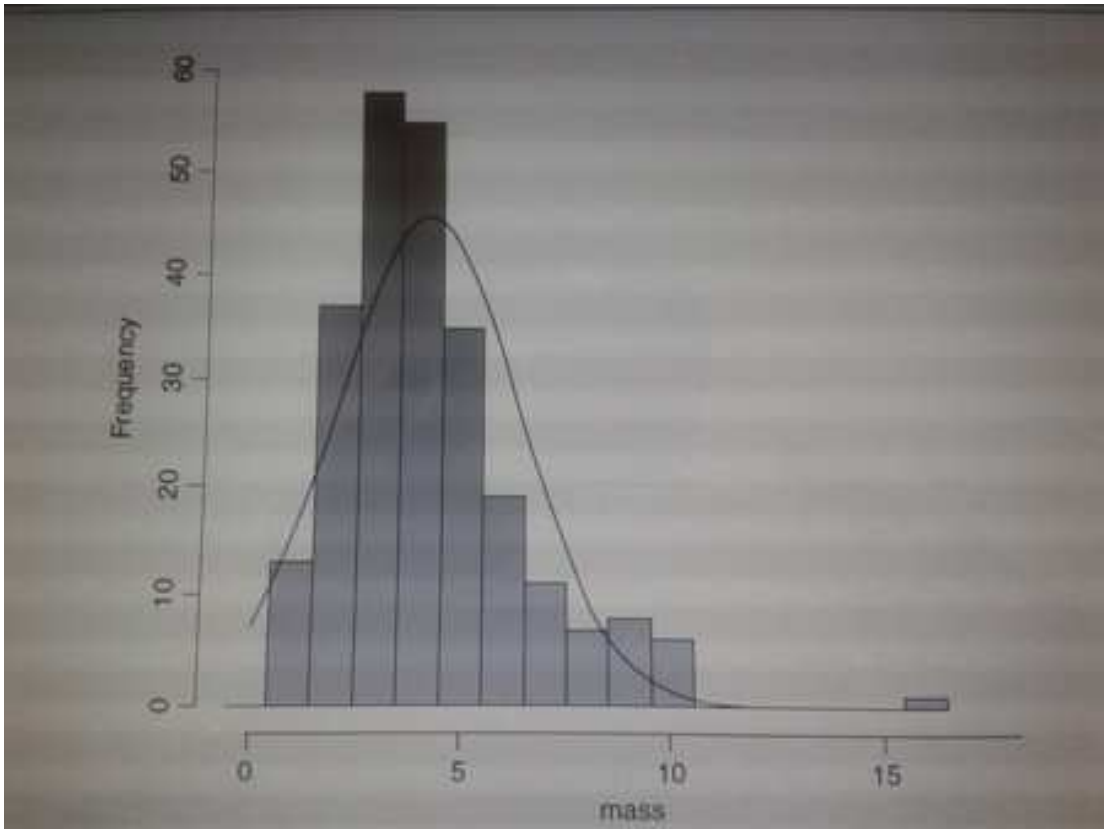
```
par(mfrow=c(1,1))
fishes<-read.table("c:\\temp\\fishes.txt",header=T)
attach(fishes)
names(fishes)
[1] "mass"
mean(mass)
[1] 4.194275
max(mass)
[1] 15.53216
```

Τώρα το ιστόγραμμα της μάζας των ιχθύων παράγεται, διευκρινίζοντας κάδους ακέραιος που είναι 1 γραμμάριο σε πλάτος, μέχρι ένα μέγιστο από 16.5 g:

```
hist(mass,breaks=-0.5:16.5,col="green",main="")
```

Για τους σκοπούς της επίδειξης, παράγουμε όλα όσα χρειαζόμαστε μέσα στις γραμμές συνάρτησης: Η αλληλουχία των x τιμών για αποτύπωση (0 έως 16), και το ύψος της συνάρτησης πυκνότητας (ο αριθμός Ψαριών (length(mass)) φορές η πυκνότητα πιθανότητας για κάθε μέλος αυτής της αλληλουχίας, για μια κανονική κατανομή με mean(mass) και τυπική απόκλιση sqrt(var(mass)) ως παραμέτρους του, όπως αυτό:

```
lines(seq(0,16,0.1),length(mass)*dnorm(seq(0,16,0.1),mean(mass),sqrt(var(mass))))
```



Η κατανομή των μεγεθών των ψαριών σαφώς δεν είναι φυσιολογική. Υπάρχουν πάρα πολλά ψάρια των 3 και 4 γραμμάριων, πολύ λίγα από 6 ή 7 γραμμάρια, και πάρα πολλά πραγματικά μεγάλα ψάρια (πάνω από 8 γραμμάρια). Αυτό το είδος ασύμμετρης κατανομής είναι πιθανώς καλύτερα να περιγράφεται από μια κατανομή γάμμα μιας κανονικής κατανομής (βλ. σελ. 231)..

Άλλες κατανομές που χρησιμοποιούνται για έλεγχο υποθέσεων

Οι κύριες διανομές που χρησιμοποιούνται σε έλεγχο υποθέσεων είναι: η στατιστική x τετραγώνου, για τον έλεγχο υποθέσεων που αφορούν δεδομένα για το πλήθος? Η F συνάρτηση του Fisher, στην ανάλυση του τετραγώνου της τυπικής απόκλισης (ANOVA) για τη σύγκριση δύο τετραγώνων τυπικής απόκλισης? και t του Student, σε μικρά-δείγματα δουλειάς για τη σύγκριση των δύο εκτιμήσεων των παραμέτρων. Οι διανομές μας λένε ότι το μέγεθος του στατιστικού αποτελέσματος της δοκιμής που θα μπορούσε να αναμένεται από την τύχη μόνο όταν τίποτα δεν συνέβαινε (δηλαδή όταν η μηδενική υπόθεση ήταν αλήθεια). Λαμβάνοντας υπόψη το κανόνα ότι μια μεγάλη τιμή του στατιστικού αποτελέσματος της δοκιμής μας λέει ότι κάτι συμβαίνει, και ότι, συνεπώς, η μηδενική υπόθεση είναι ψευδής, αυτές οι διανομές καθορίζουν τι αποτελεί μια μεγάλη τιμή από τη στατιστική δοκιμή. Για παράδειγμα, αν κάνουμε ένα τεστ στατιστικής x τετραγώνου και το στατιστικό μας τεστ είναι 14,3 πάνω από 9 d.f.(βαθμούς ελευθερίας), εμείς χρειάζεται να ξέρουμε αν αυτή είναι μια μεγάλη τιμή (δηλαδή η μηδενική υπόθεση είναι ψευδής) ή μικρή τιμή (δηλαδή η μηδενική υπόθεση γίνεται αποδεκτή, ή τουλάχιστον δεν μπορεί να απορριφθεί). Στην παλιότερες ημέρες εμείς θα είχαμε βελτιώσει την τιμή στους πίνακες στατιστική x τετραγώνου. Θα

είχαμε εξετάσει τη σειρά με την ένδειξη 9 (οι βαθμοί της σειράς ελευθερίας) και τη στήλη από α= 0, 05. Αυτό είναι η συμβατική τιμή για την αποδεκτή πιθανότητα διάπραξης ενός λάθος Τύπου I: ότι είναι να πούμε ότι επιτρέπει 1 στις 20 πιθανότητες να απορρίπτει τη μηδενική υπόθεση όταν είναι όντως αλήθεια? βλ. σελ.. 317). Στις μέρες μας, απλά πληκτρολογήστε:

1-pchisq(14.3,9)
[1] 0.1120467

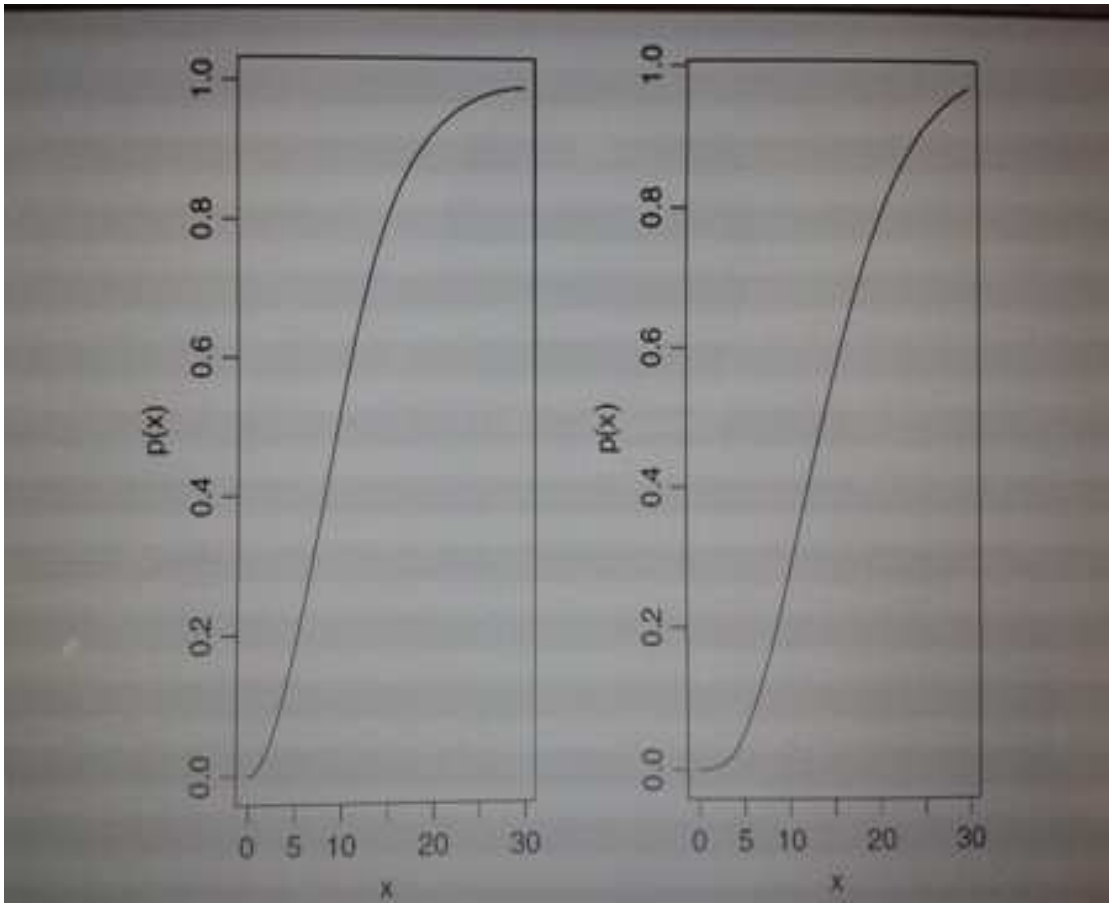
Αυτό δείχνει ότι το 14.3 είναι στην πραγματικότητα ένα σχετικά μικρός αριθμός όταν έχουμε 9 df (βαθμούς ελευθερίας). Εμείς θα συμπεράνουμε ότι δεν συμβαίνει τίποτα, διότι η τιμή της στατιστικής x τετραγώνου τόσο μεγάλη όσο η 14.3 έχει μια μεγαλύτερη από ό, τι μια πιθανότητα 11% που της απορρέουν από τύχη και μόνο. Θα θέλαμε την πιθανότητα να είναι μικρότερη από 5% πριν απορρίψουμε τη μηδενική υπόθεση. Έτσι, πόσο μεγάλο θα είναι το τεστ στατιστικής που πρέπει να είναι, για να μπορέσουμε να απορρίψουμε την μηδενική υπόθεση; Χρησιμοποιούμε qchisq να απαντήσουμε αυτό. Τα δύο ορίσματα είναι 1 -α και ο αριθμός των βαθμών ελευθερίας:
qchisq(0.95,9)
[1] 16.91898
Έτσι, το στατιστικό αποτέλεσμα της δοκιμής θα πρέπει να είναι μεγαλύτερο από το 16,92 για να μπορέσουμε να απορρίψουμε την μηδενική υπόθεση, όταν υπήρχαν 9 d.f. (βαθμοί ελευθερίας).
Θα μπορούσαμε να χρησιμοποιήσουμε pf και qf σε ένα ακριβώς ανάλογο τρόπο για F του Fisher. Έτσι, η πιθανότητα του να πάρει ένα ποσοστό διακύμανσης 2,85 από τύχη και μόνο όταν η μηδενική υπόθεση είναι αλήθεια, δεδομένου ότι έχουμε 8 d.f. στον αριθμητή και 12 d.f. στον παρονομαστή, είναι λίγο κάτω 5% (δηλαδή η αξία του είναι αρκετά μεγάλη για να μας επιτρέψει να απορρίψουμε τη μηδενική υπόθεση):
1-pf(2.85,8,12)
[1] 0.04992133
Σημειώστε ότι με pf, βαθμούς ελευθερίας στον αριθμητή (8) έρχονται πρώτα στη λίστα των ορισμάτων, ακολουθούμενη από d.f. στον παρονομαστή (12).
Ομοίως, με του φοιτητή t τις στατιστικές και pt και qt. Για παράδειγμα, η τιμή του t για δοκιμή σε πίνακες για ένα αμφίπλευρο έλεγχο στα α/ 2 = 0, 025 με d.f. = 10 είναι
qt(0.975,10)
[1] 2.228139

Στατιστική x τετραγώνου

Ίσως η πιο γνωστή από όλες τις στατιστικές κατανομές, εισήχθη σε γενιές από τα παιδιά στο σχολείο τα μαθήματα γεωγραφίας τους, και συνολικά παρεξηγημένη στη συνέχεια. Πρόκειται για μια ειδική περίπτωση της κατανομής γάμμα (σελ. 229), που χαρακτηρίζεται από μία μόνο παράμετρο, ο αριθμός των βαθμών ελευθερίας. Η μέση τιμή είναι ίση με τους βαθμούς ελευθερίας ν («nu»,προφέρεται «νέα»), και η διακύμανση είναι ίση με 2ν . Η συνάρτηση πυκνότητας μοιάζει με αυτό:
f(x)=1:[2^ν/2*Γ(ν/2)]*[x^((ν/2)-1)]*[1:e^(x/2)]
πού Γ είναι η συνάρτηση γάμμα (βλ. σελ. 201.). Η στατιστική x τετραγώνου είναι σημαντική γιατί πολλές τετραγωνικές μορφές ακολουθούν τη κατανομή

στατιστική x τετραγώνου με βάση την παραδοχή ότι ακολουθούν τα δεδομένα της κανονικής κατανομής. Ειδικότερα, η διακύμανση του δείγματος είναι μια μεταβλητή κλίμακα στατιστικής x τετραγώνου.
Οι στατιστικές αναλογίες πιθανότητας, επίσης, περίπου κατανέμονται ως στατιστική x τετραγώνου (βλ. το F διανομής, παρακάτω).
Όταν η αθροιστική πιθανότητα χρησιμοποιείται, ένα προαιρετικό τρίτο όρισμα μπορεί να παρέχεται περιγραφή μη τάσης παραμονής στο κέντρο. Εάν η μη κεντρική στατιστική x τετραγώνου είναι το άθροισμα από ν ανεξάρτητες κανονικά τυχαίες μεταβλητές, τότε η μη κεντρικότητας παράμετρος είναι ίση με το άθροισμα των τετραγώνων μέσα από τις συνήθεις μεταβλητές. Εδώ είναι οι αθροιστικές πιθανοτήτων γραφικές παραστάσεις για μια μη-κεντρικότητας παράμετρο (ncp) που βασίζεται σε τρία κανονικά μέσα (από 1, 1,5 και 2) και το άλλο με 4 μέσα και ncp = 10:

```
par(mfrow=c(1,2))
x<-seq(0,30,.25)
plot(x,pchisq(x,3,7.25),type="l",ylab="p(x)",xlab="x")
plot(x,pchisq(x,5,10),type="l",ylab="p(x)",xlab="x")
```



Η αθροιστική πιθανότητα στα αριστερά έχει 3 d.f.(βαθμούς ελευθερίας) και μη κεντρική παράμετρο (1^2 +1,5^2 +2^2 = 7,25), ενώ η διανομή στα δεξιά έχει 4 d.f. (βαθμούς ελευθερίας) και μη κεντρικότητας 10 (σημειώστε η πλέον αριστερή ουρά σε χαμηλές πιθανότητες).

Η στατιστική x-τετράγωνου χρησιμοποιείται επίσης για τη δημιουργία διαστημάτων εμπιστοσύνης για τις διακυμάνσεις του δείγματος. Η ποσότητα

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

είναι οι βαθμοί ελευθερίας (n-1) πολλαπλασιάζομενοι με το λόγο της διακύμανσης δείγματος s^2 στην άγνωστη διακύμανση του πληθυσμού σ^2. Αυτό ακολουθεί μια στατιστική x τετραγώνου κατανομή, ώστε να μπορούμε να θεσπίσουμε ένα 95% διάστημα εμπιστοσύνης για σ^2 ως ακολούθως:

$$\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{\alpha/2}}$$

Ας υποθέσουμε ότι η διακύμανση του δείγματος s^2 = 10,2 στις 8 d.f. (βαθμούς ελευθερίας). Στη συνέχεια, το διάστημα στο σ^ 2 δίνεται από 8*10.2/qchisq(.975,8)

[1] 4.65367
8*10.2/qchisq(.025,8)
[1] 37.43582
πράγμα που σημαίνει ότι μπορούμε να είμαστε 95% σίγουροι ότι η διακύμανση του πληθυσμού βρίσκεται στην περιοχή
4,65≤σ^2≤37,44

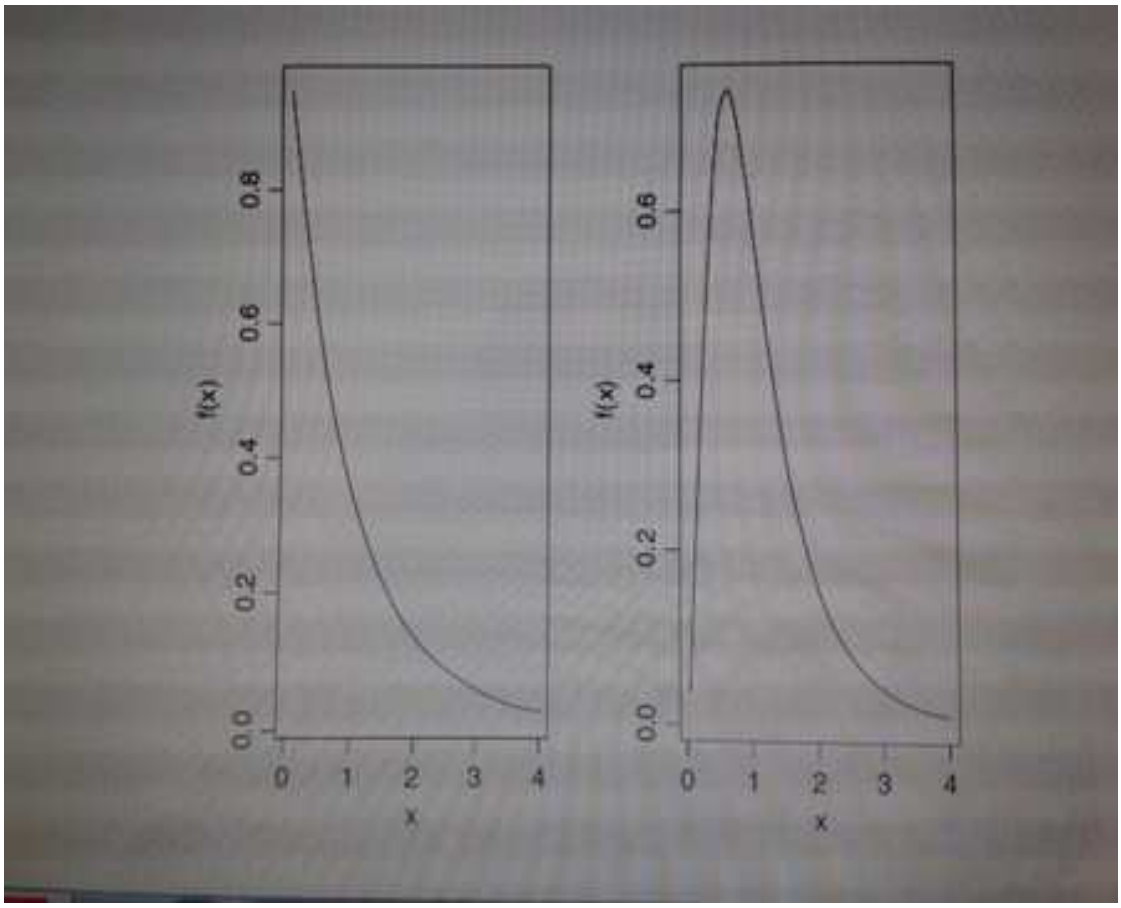
Συνάρτηση F του Fisher

Αυτό είναι το περίφημο τεστ του λόγου διακύμανσης που καταλαμβάνει την προτελευταία στήλη του κάθε ANOVA πίνακα. Ο λόγος της διακύμανσης μεταχείρισης διακύμανσης των σφαλμάτων ακολουθεί την κατανομή F, και συχνά θα θέλετε να χρησιμοποιήσετε την ποσότητα qf να αναζητήσετε κρίσιμες τιμές της F. Μπορείτε να καθορίσετε, προκειμένου, η πιθανότητα ελέγχου μιας ουράς (αυτό θα είναι συνήθως 0,95), τότε οι δύο βαθμοί ελευθερίας: αριθμητής πρώτα, και μετά παρονομαστής. Έτσι, το 95% της τιμής της F με 2 και 18 d.f. είναι

```
qf(.95,2,18)
[1]  3.554557
```

Αυτό είναι ό, τι η συνάρτηση πυκνότητας της F μοιάζει για 2 και 18 d.f. (αριστερά) και 6 και 18 d.f. (δεξιά):

```
x<-seq(0.05,4,0.05)
plot(x,df(x,2,18),type="l",ylab="f(x)",xlab="x")
plot(x,df(x,6,18),type="l",ylab="f(x)",xlab="x")
```



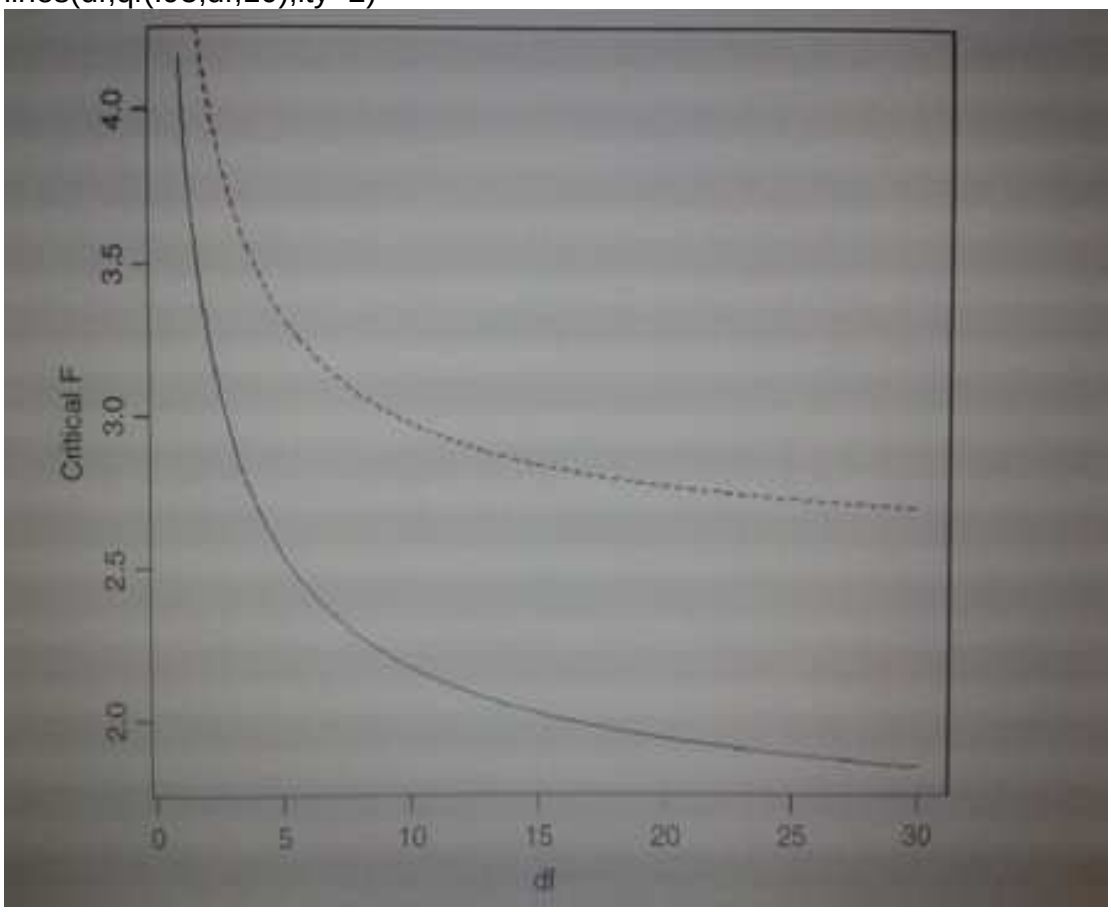
Η κατανομή F είναι μια δύο παραμέτρων κατανομής που ορίζεται από τη συνάρτηση πυκνότητας

$$f(x) = \frac{r\Gamma(1/2(r+s))}{s\Gamma(1/2r)\Gamma(1/2s)} \frac{(rx/s)^{(r-1)/2}}{[1+(rx/s)]^{(r+s)/2}}$$

όπου r είναι οι βαθμοί ελευθερίας του αριθμητή και s είναι οι βαθμοί ελευθερίας στον παρονομαστή. Η διανομή είναι το όνομά του R.A. Fisher, ο πατέρας της ανάλυσης των διακύμανσεων, και κύριος δημιουργός στην ποσοτική γενετική. Είναι κομβικής σημασίας για έλεγχο υποθέσεων, λόγω της χρήσης του για την εκτίμηση της σημασίας των διαφορών μεταξύ των δύο διακυμάνσεων. Το στατιστικό αποτέλεσμα της δοκιμής υπολογίζεται διαιρώντας την μεγαλύτερη διακύμανση από την μικρότερη διακύμανση. Οι δύο διακυμάνσεις είναι σημαντικά διαφορετικές όταν η αναλογία αυτή είναι μεγαλύτερη από την κρίσιμη τιμή της συνάρτησης F του Fisher. Οι βαθμοί ελευθερίας στον αριθμητή και στον παρονομαστή επιτρέπουν τον υπολογισμό του κρίσιμης τιμής του στατιστικού αποτελέσματος της δοκιμής. Όταν υπάρχει ένας μονός βαθμός ελευθερίας στον αριθμητή, η κατανομή είναι ίση με το τετράγωνο του t του Student; F = t^2. Έτσι, ενώ ο κανόνας του αντίχειρα για την κρίσιμη τιμή του t είναι 2, έτσι ο κανόνας του αντίχειρα για F = t^2 = 4. Για

να δείτε πόσο καλά δουλεύει ο κανόνας του αντίχειρα, μπορούμε να σχεδιάσομε κρίσιμες τιμές F εναντίον d.f.(βαθμών ελευθερίας) στον αριθμητή:

```
df<-seq(1,30,.1)
plot(df,qf(.95,df,30),type="l",ylab="Critical F")
lines(df,qf(.95,df,10),lty=2)
```

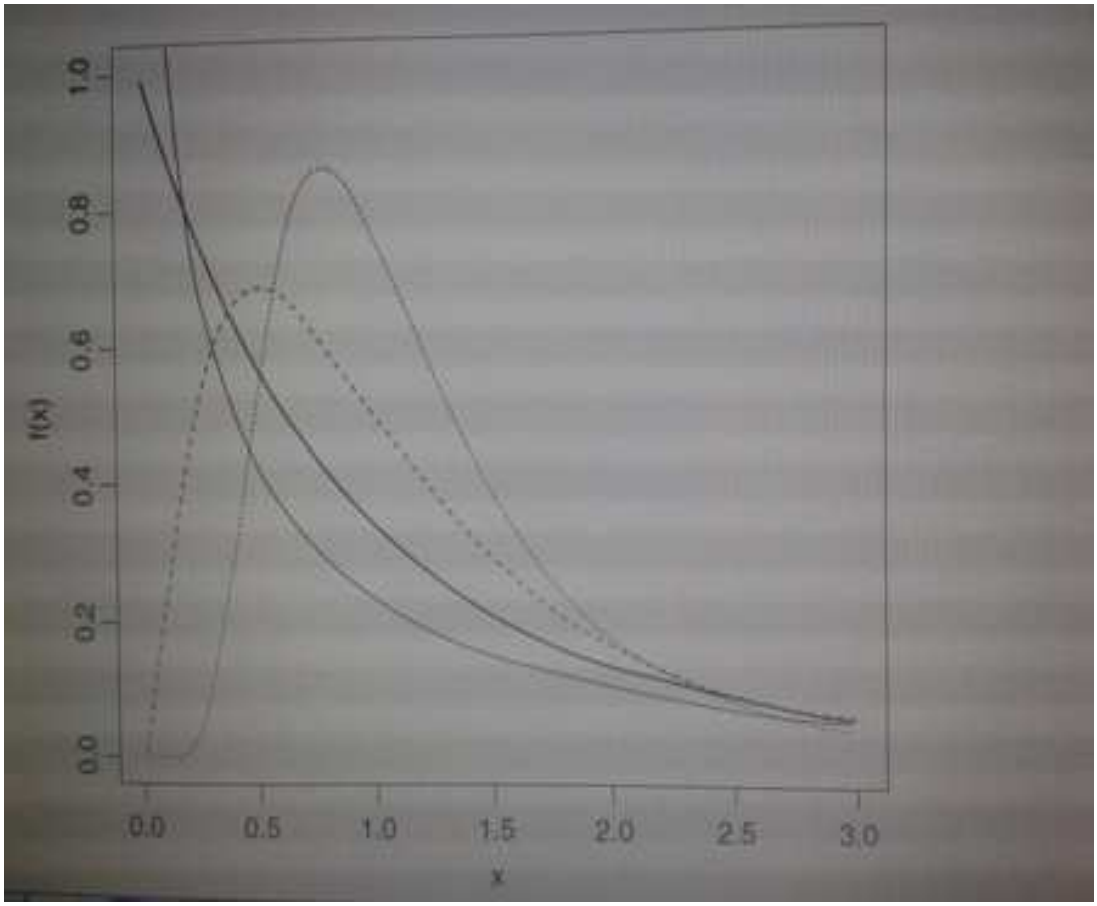


Θα δείτε ότι ο κανόνας του αντίχειρα (κρίσιμη F = 4) γρήγορα γίνεται πολύ μεγάλος μόλις ο df στον αριθμητή (στον άξονα x) είναι μεγαλύτερος από 2. Η χαμηλότερη (στερεή) γραμμή δείχνει τις κρίσιμες τιμές του F, όταν ο παρονομαστής έχει 30 d.f. και η ανώτερη (διακεκομμένη) γραμμή δείχνει την περίπτωση στην οποία ο παρονομαστής έχει 10 d.f.

Το σχήμα της συνάρτησης πυκνότητας της κατανομής F εξαρτάται από τους βαθμούς ελευθερίας στον αριθμητή.

```
x<-seq(0.01,3,0.01)
plot(x,df(x,1,10),type="l",ylim=c(0,1),ylab="f(x)")
lines(x,df(x,2,10),lty=6)
lines(x,df(x,5,10),lty=2)
lines(x,df(x,30,10),lty=3)
```

Η f(x) συνάρτηση πυκνότητας πιθανότητας μειώνεται μονότονα, όταν ο αριθμητής έχει 1 df ή 2d.f., αλλά αυξάνεται σε ένα μέγιστο για d.f. από 3 ή περισσότερους (5 και 30 φαίνονται εδώ): όλες οι γραφικές παραστάσεις έχουν 10 d.f. στον παρονομαστή.

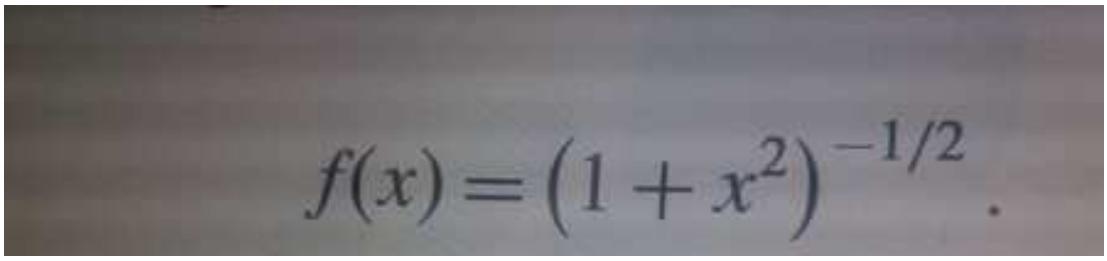


t του Student

Αυτή η περίφημη διανομή εκδόθηκε για πρώτη φορά από τον WS Gossett το 1908 με το ψευδώνυμο του «Student», διότι τότε ο εργοδότης του, η εταιρεία ζυθοποιίας Guinness στο Δουβλίνο, δεν θα επίτρεπε στους υπαλλήλους να τη δημοσιεύσουν με τα ονόματά τους. Είναι ένα μοντέλο με μία παράμετρο, r, με συνάρτηση πυκνότητας:

$$f(x) = \frac{r\Gamma(1/2(r+s))}{s\Gamma(1/2r)\Gamma(1/2s)} \frac{(rx/s)^{(r-1)/2}}{[1+(rx/s)]^{(r+s)/2}}$$

όπου -∞ <x <+∞. Αυτό φαίνεται πολύ περίπλοκο, αλλά αν έχουν αφαιρεθεί όλες οι σταθερές μακριά, μπορείτε να δείτε πόσο απλή είναι η βασική δομή πραγματικά

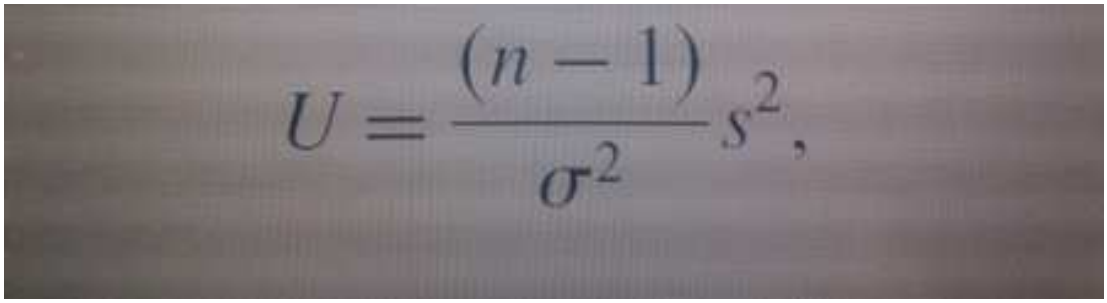


$$f(x) = \left(1 + x^2\right)^{-1/2}.$$

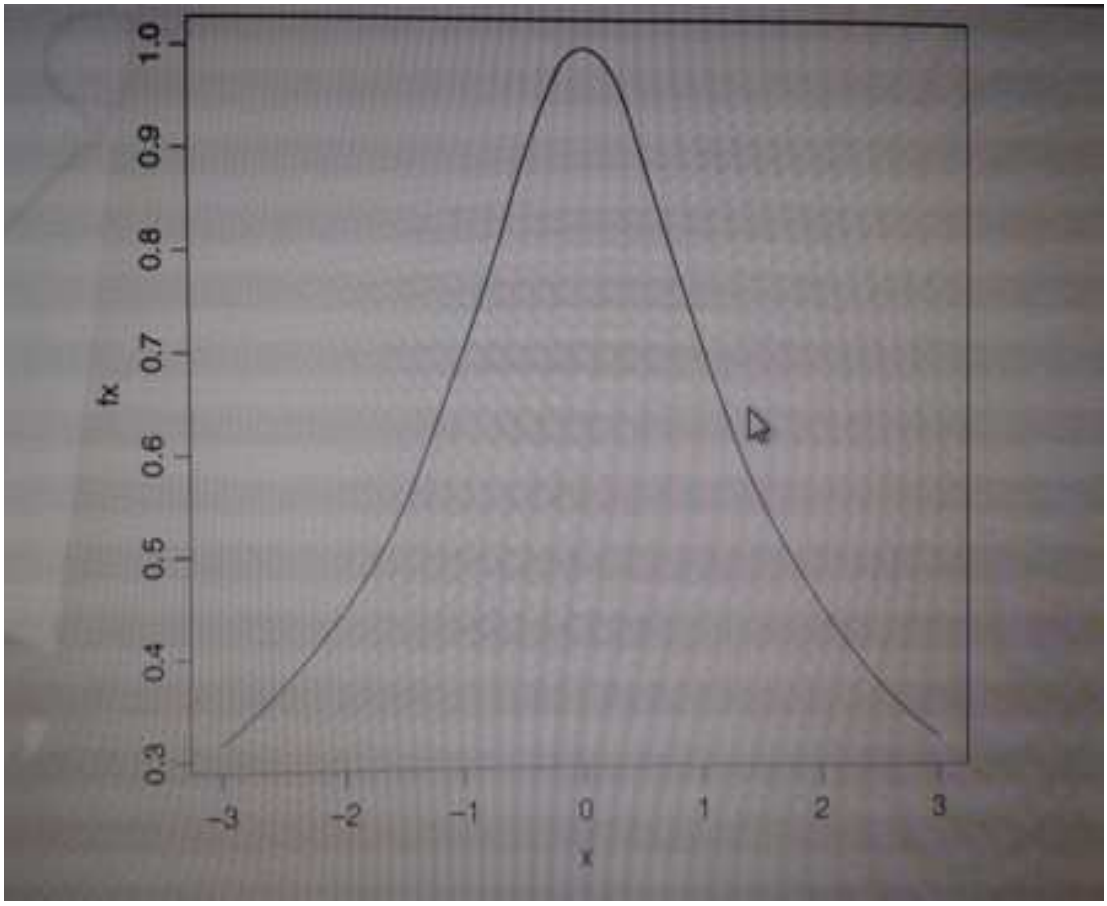
Μπορούμε να σχεδιάσετε αυτό για τιμές του x από -3 έως +3 ως εξής:

```
x<-seq(-3,3,0.01)
fx<-(1+x^2)^(-0.5)
plot(x,fx,type="l")
```

Το κύριο πράγμα που πρέπει να παρατηρήσετε είναι το πώς το πάχος είναι οι ουρές της κατανομής, σε σύγκριση με τη κανονική κατανομή. Η πληθώρα των σταθερών είναι απαραίτητο να κλιμακωθεί η συνάρτηση πυκνότητας έτσι ώστε αυτό το ολοκλήρωμα είναι 1. Αν ορίσουμε U ως



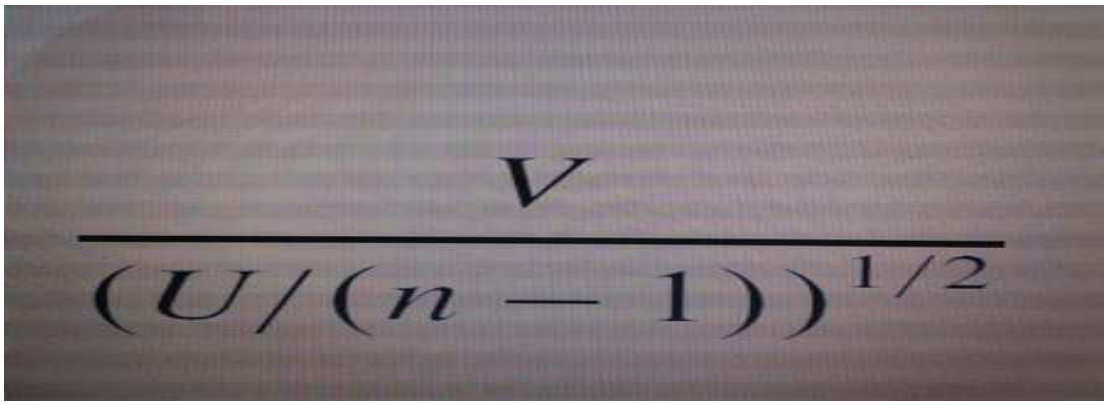
$$U = \frac{(n-1)}{\sigma^2} s^2,$$



τότε αυτή είναι στατιστική x τετράγωνου κατανεμημένη σε n-1 d.f. (βλ. ανωτέρω). Τώρα ορίζω V όπως

$$V \equiv \frac{n^{1/2}}{\sigma}(\bar{y} - \mu)$$

και σημειώστε ότι αυτό είναι κανονικά κατανεμημένο με μέση τιμή 0 και τυπική απόκλιση 1 (το πρότυπο κανονική κατανομή), έτσι



$$\frac{V}{(U/(n-1))^{1/2}}$$

είναι ο λόγος της κανονικής κατανομής και x τετραγώνου κατανομής. Ίσως να ήθελε να την συγκρίνει με την κατανομή της F (παραπάνω), η οποία είναι η αναλογία των δύο διανεμόντων x τετραγώνων τυχαίων μεταβλητών.

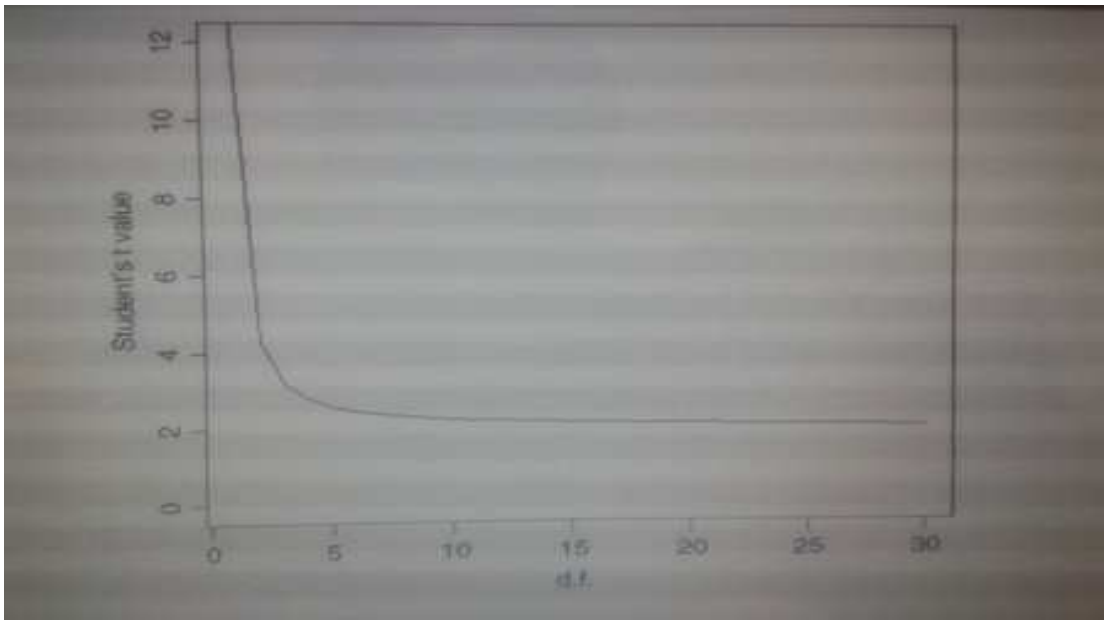
Σε ποιο σημείο έχει ο εμπειρικός κανόνας για το t του Student t= 2 καταρρεύσει τόσο σοβαρά ώστε η πραγματικότητα είναι παραπλανητική; Για να βρούμε αυτό έξω, θα πρέπει να σχεδιάζουμε την τιμή του t Student έναντι του μεγέθους του δείγματος (στην πραγματικότητα ενάντια βαθμών ελευθερίας) για τα μικρά δείγματα. Χρησιμοποιούμε qt (quantile της t) και καθορίζουμε την πιθανότητα σε δύο-ουρά τιμές από 0.975:

```
plot(1:30,qt(0.975,1:30), ylim=c(0,12),type="l",ylab="Student"s t
value",xlab="d.f.")
```

Όπως βλέπετε, ο κανόνας του αντίχειρα γίνεται μόνο πραγματικά αδύνατος για τους βαθμούς ελευθερίας λιγότερο από 5 ή περίπου τόσο. Για τους περισσότερους πρακτικούς λόγους t ≈ 2 είναι πραγματικά ένας καλός κανόνας εργασίας του αντίχειρα.

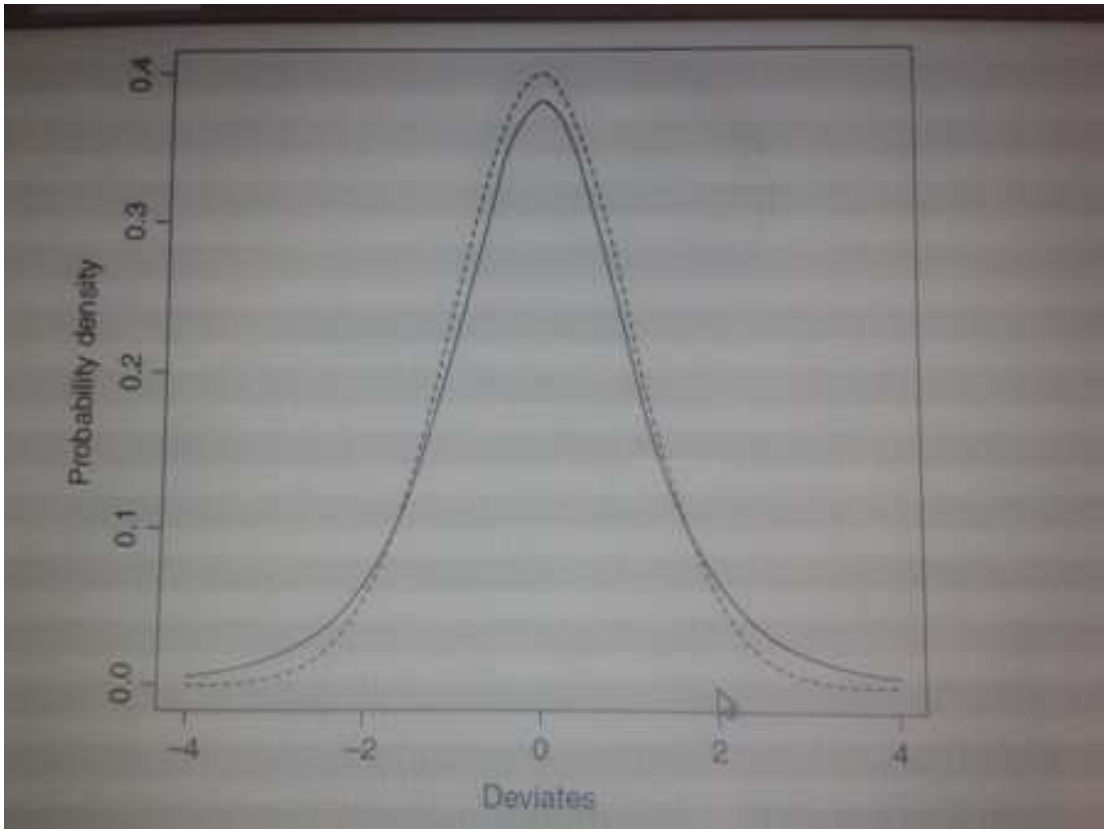
Έτσι με τι μοιάζει η κατανομή t , σε σύγκριση με μια κανονική; Ας αναδιατυπώσουμε την τυπική κανονική ως διακεκομμένη γραμμή (lty = 2):

```
xvs<-seq(-4,4,0.01)
plot(xvs,dnorm(xvs),type="l",lty=2,ylab="Probability density",xlab="Deviates")
```



Τώρα μπορούμε να επικαλύψουμε το t του Student με d.f. = 5 ως στερεά γραμμή για να δείτε τη διαφορά:

lines (xvs, dt (xvs, df = 5))

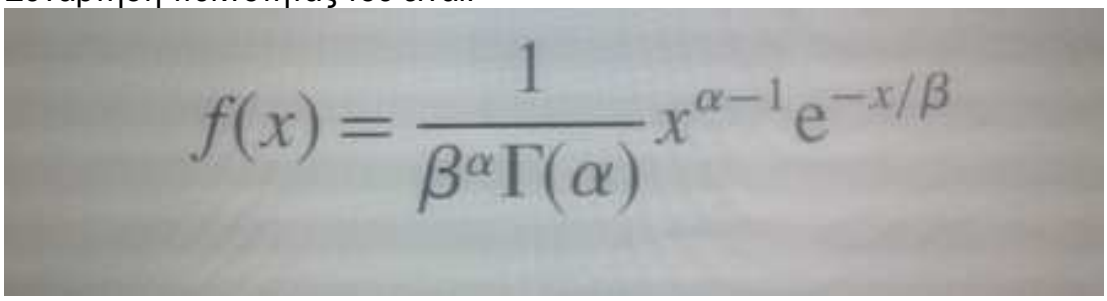


Η διαφορά μεταξύ της κανονικής (διακεκομμένης) και διανομών t του Student (συμπαγής γραμμή) είναι εκείνη η t κατανομή που έχει «παχύτερες ουρές». Αυτό σημαίνει ότι οι ακραίες τιμές είναι πιο πιθανές με μια t διανομή από ό, τι με μια κανονική, και τα διαστήματα εμπιστοσύνης είναι αντίστοιχα ευρύτερα. Έτσι, αντί για 95% διάστημα από ± 1, 96, με κανονική κατανομή θα πρέπει να έχουμε ένα 95% διάστημα από ± 2,57 για την κατανομή t του Student με 5 βαθμούς ελευθερίας:

```
qt(0.975,5)
[1] 2.570582
```

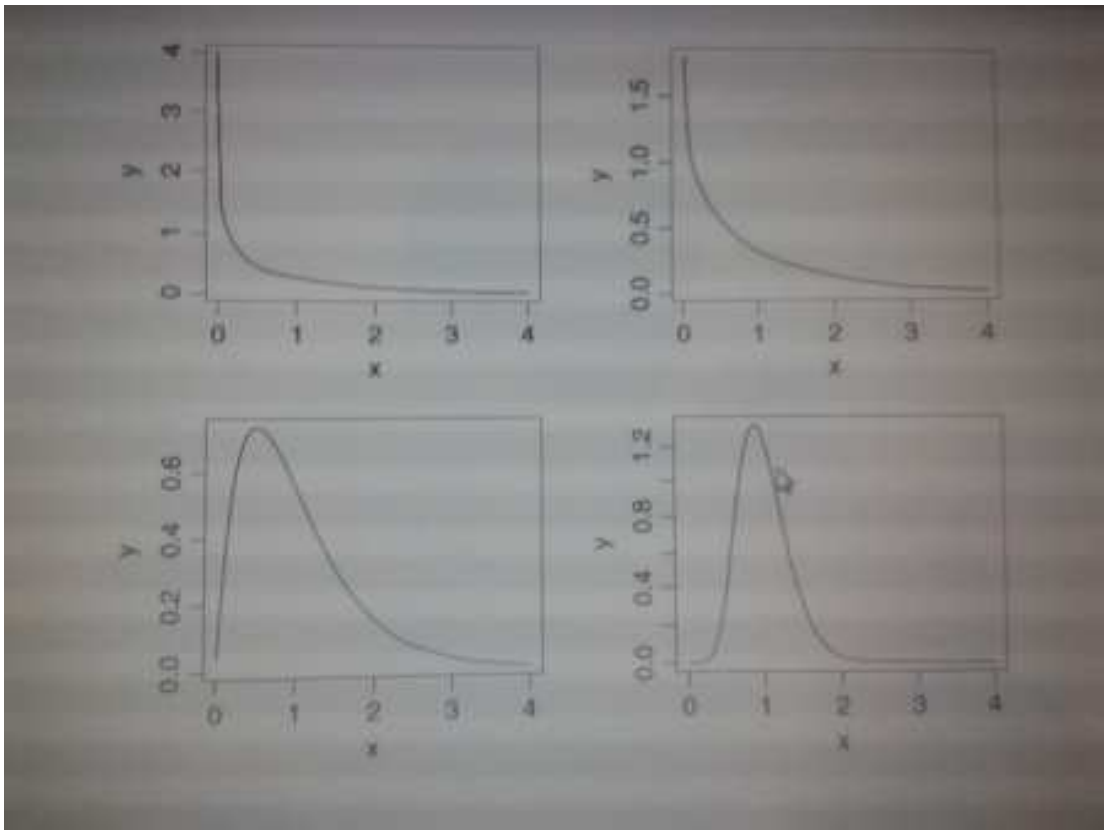
Η κατανομή γάμμα

Η κατανομή γάμα είναι χρήσιμη για την περιγραφή ενός ευρέως φάσματος διαδικασιών, όπου τα δεδομένα είναι θετικά ασύμετρα (δηλ. μη-φυσιολογικά, με μακρά ουρά στα δεξιά). Είναι μια δύο παραμέτρων κατανομή, όπου οι παράμετροι παραδοσιακά γνωστοί ως σχήμα και σταθερή αναλογία. Συνάρτηση πυκνότητας του είναι:



Πού α είναι η παράμετρος σχήματος και β^-1 είναι η παράμετρος συντελεστή (εναλλακτικά, β είναι γνωστός ως η κλίμακα παράμετρος). Ειδικές περιπτώσεις της κατανομής γάμμα είναι η εκθετική (α=1) και η στατιστική x τετραγώνου (α=v/2, β=2).
Για να δούμε την επίδραση της παραμέτρου σχήματος για την πυκνότητα πιθανότητας, μπορούμε να σχεδιάσουμε τη κατανομή γάμμα για διάφορες τιμές του συντελεστή σχήματος και πάνω από την περιοχή 0.01 έως 4:

```
x<-seq(0.01,4,.01)
par(mfrow=c(2,2))
y<-dgamma(x,.5,.5)
plot(x,y,type="l")
y<-dgamma(x,.8,.8)
plot(x,y,type="l")
y<-dgamma(x,2,2)
plot(x,y,type="l")
y<-dgamma(x,10,10)
plot(x,y,type="l")
```



Οι γραφικές παραστάσεις από πάνω αριστερά προς τα κάτω δεξιά δείχνουν διαφορετικές τιμές του α: 0,5, 0,8, 2 και 10. Σημείωση πώς α <1 παράγει μονότονες φθίνουσες συναρτήσεις και α> 1 παράγει κυρτωμένες καμπύλες που διέρχονται από την αρχή, με το βαθμό ασυμμετρίας να εξασθενεί όσο το α αυξάνεται.

Η μέση τιμή της κατανομής είναι αβ, η τυπική απόκλιση είναι α*(β^2), η ασυμμετρία είναι 2/(α^1/2) και η κύρτωση είναι 6/α. Έτσι, για την εκθετική κατανομή έχουμε μια μέση τιμή β, μια διακύμανση β^2, μια ασυμμετρία 2 και μία κύρτωση 6, ενώ για την στατιστική x τετράγωνου κατανομή έχουμε ένα μέσο όρο v, μια διακύμανση των 2v μια ασυμμετρία της 2*[(2/v)^(1/2)] και μια κύρτωση του 12/v. Παρατήρησε επίσης ότι

1/β=mean/ variance,

Shape=1/β X mean.

Μπορούμε να απαντήσουμε τώρα ερωτήματα όπως αυτό: ποια είναι η τιμή 95% κβάντισης που αναμένεται από μια κατανομή γάμμα με αριθμητικό μέσο=2 και διακύμανση=3; Αυτό σημαίνει ότι το ποσοστό είναι 2/3 και η κατάσταση είναι 4/3 έτσι:

```
qgamma(0.95,2/3,4/3)
[1] 1.732096
```

Μία σημαντική χρήση της κατανομής γάμα είναι στην περιγραφή συνεχής μέτρησης δεδομένων τα οποία δεν διανέμονται κανονικά. Εδώ είναι ένα παράδειγμα όπου τα δεδομένα μάζας σώματος για 200 ψάρια απεικονίζονται ως ιστόγραμμα και μία κατανομή γάμα με την ίδια μέση τιμή και διακύμανση επικάθονται ως μία ομαλή καμπύλη:

```
fishes<-read.table("c:\\temp\\fishes.txt",header=T)
attach(fishes)
names(fishes)
[1] "mass"
```

Πρώτον, έχουμε υπολογίσει τις δύο τιμές των παραμέτρων για την κατανομή γάμμα:

```
rate<-mean(mass)/var(mass)
shape<-rate*mean(mass)
rate
[1] 0.8775119
shape
[1] 3.680526
```

Θα πρέπει να γνωρίζουμε τη μεγαλύτερη τιμή της μάζας, προκειμένου να φτιάξουν τα κουτιά από το ιστόγραμμα:

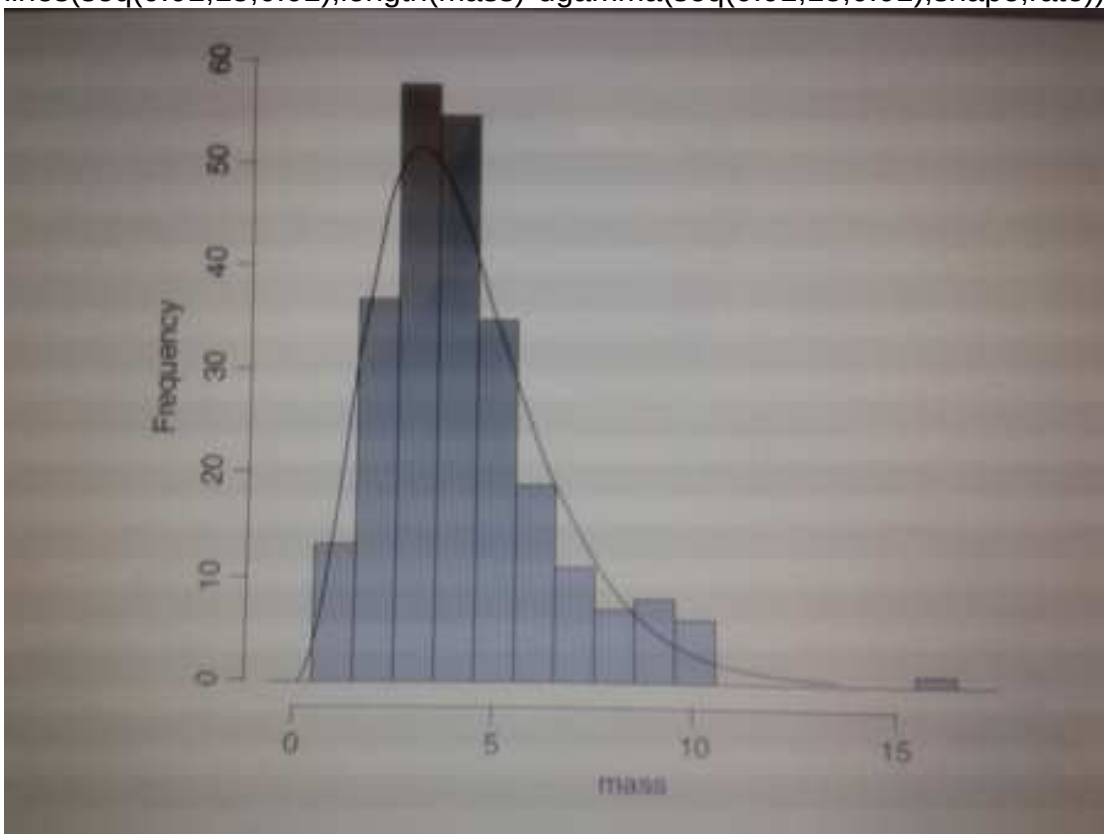
```
max(mass)
[1] 15.53216
```

Τώρα μπορούμε να σχεδιάσουμε το ιστόγραμμα, χρησιμοποιώντας τα σημεία διάσπασης στο 0,5 για να πάρει ακέραιο με επίκεντρο εμπόδιο μέχρι κατ' ανώτατο όριο 16,5 για να φιλοξενήσει μεγαλύτερα ψάρια μας:

```
hist(mass,breaks=-0.5:16.5,col="green",main="")
```

Η συνάρτηση πυκνότητας της κατανομής γάμα επικαλύπτεται με γραμμές όπως αυτή:

```
lines(seq(0.01,15,0.01),length(mass)*dgamma(seq(0.01,15,0.01),shape,rate))
```



Η προσαρμογή είναι πολύ καλύτερη από ό, τι όταν προσπαθήσαμε να χωρέσει μια κανονική κατανομή σε αυτά τα ίδια δεδομένα νωρίτερα (βλ. σελ.. 221).

Η εκθετική κατανομή

Αυτή είναι μία παράμετρος διανομής που είναι μια ειδική περίπτωση της κατανομής γάμμα. Χρησιμοποιείται πολύ στην ανάλυση επιβίωσης, συνάρτηση πυκνότητας της δίνεται σελ. 792 και η χρήση του στην ανάλυση επιβίωσης εξηγείται στη σελ. 802. Η γεννήτρια τυχαίου αριθμού της εκθετικής είναι χρήσιμο για προσομοιώσεις Monte Carlo του χρόνου μέχρι θανάτου όταν ο κίνδυνος (ο στιγμιαίος κίνδυνος θανάτου) είναι σταθερός με την ηλικία. Μπορείτε να καθορίσετε τον κίνδυνο, ο οποίος είναι το αντίστροφο της μέσης ηλικίας θανάτου:

```
rexp(15,0.1)
[1] 9.811752 5.738169 16.261665 13.170321 1.114943
[6] 1.986883 5.019848 9.399658 11.382526 2.121905
[11] 10.941043 5.868017 1.019131 13.040792 38.023316
```

Πρόκειται για 15 τυχαίες ζωές με μια αναμενόμενη τιμή του 1/0,1 = 10 χρόνια; Δίνουν μια μέση τιμή δείγματος 9,66 έτη.

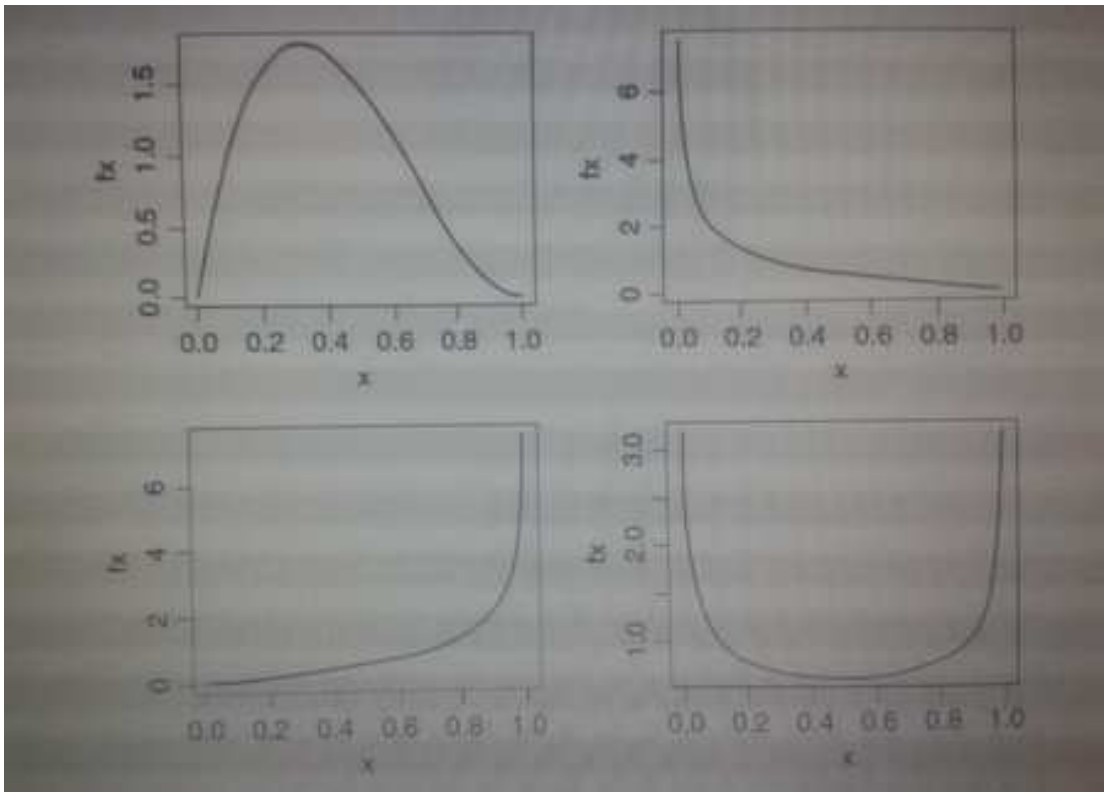
Η κατανομή βήτα

Αυτή έχει δυο θετικές σταθερές, α και β, και x είναι περιορισμένο 0≤x≤1:

$$f(x)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}.$$

Στην R έχουμε δημιουργήσει μια οικογένεια από συναρτήσεις πυκνότητας όπως αυτό:

```
x<-seq(0,1,0.01)
fx<-dbeta(x,2,3)
plot(x,fx,type="l")
fx<-dbeta(x,0.5,2)
plot(x,fx,type="l")
fx<-dbeta(x,2,0.5)
plot(x,fx,type="l")
fx<-dbeta(x,0.5,0.5)
plot(x,fx,type="l")
```



Το σημαντικό σημείο είναι το κατά πόσον οι παράμετροι είναι μεγαλύτερες ή μικρότερες από 1. Όταν και οι δύο είναι μεγαλύτερες από 1 παίρνουμε μια n-σχηματισμένη καμπύλη οποία γίνεται όλο και πιο ασύμμετρη όπως b>α (κορυφή αριστερά). Αν 0<α<1 και b>1 τότε η πυκνότητα είναι αρνητική (κορυφή δεξιά), ενώ για α>1 και 0<b<1 η πυκνότητα είναι θετική (κάτω αριστερά). Η συνάρτηση είναι σχήματος U όταν και τα δύο α και b είναι θετικά κλάσματα. Αν α=β=1, τότε παίρνουμε την ομοιόμορφη κατανομή [0,1].
Εδώ είναι 20 τυχαίοι αριθμοί από την Βήτα κατανομή με παραμέτρους σχήματος 2 και 3:
rbeta (20,2,3)
[1] 0.5820844 0.5150638 0.5420181 0.1110348 0.5012057 0.3641780
[7] 0.1133799 0.3340035 0.2802908 0.3852897 0.6496373 0.3377459
[13] 0.1743189 0.4568897 0.7343201 0.3040988 0.5670311 0.2241543
[19] 0.6358050 0.5932503

Cauchy

Πρόκειται για μια μακρά ουρά δύο παραμέτρων διανομή, που χαρακτηρίζεται από μια παράμετρο τοποθεσίας α και μια παράμετρο κλίμακας b. Είναι πραγματικές τιμές, συμμετρικές για α (η οποία είναι επίσης στατιστικός μέσος του), και είναι μια περιέργεια σε ότι έχει αρκετά μακριές ουρές ώστε η προσδοκία δεν υπάρχει - πράγματι, δεν έχει στιγμές σε όλα (εμφανίζεται συχνά σε αντι-παραδείγματα στα μαθηματικά βιβλία). Ο αρμονικός μέσος μιας μεταβλητής με θετική πυκνότητα σε 0 τυπικά είναι κατανεμημένος ως Cauchy, και η κατανομή Cauchy εμφανίζεται επίσης στη θεωρία της κίνησης Μπράουν (π.χ. τυχαίους περιπάτους). Η γενική μορφή της κατανομής είναι

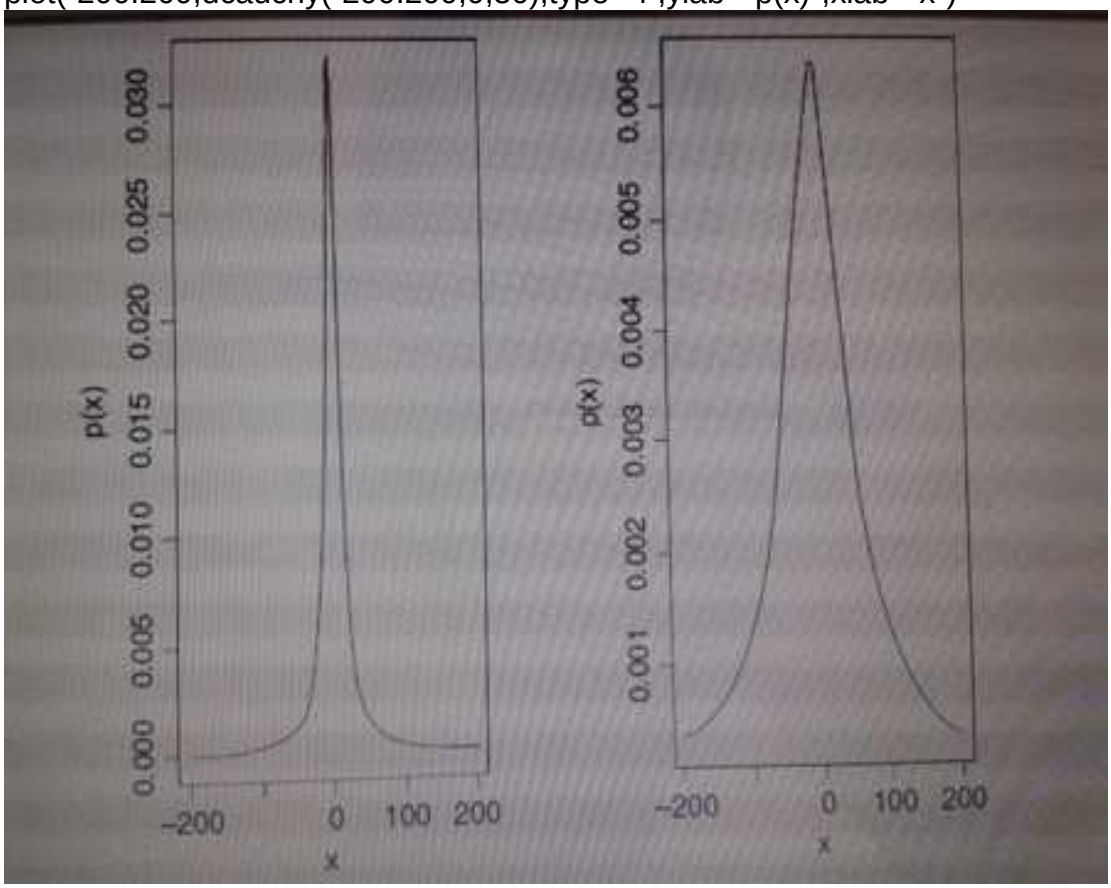
$$f(x) = \frac{1}{\pi b(1 + ((x-a)/b)^2)},$$

για -∞<x<∞ . Υπάρχει επίσης μια εκδοχή μιας παραμέτρου, με α=0 και b=1, η οποία είναι γνωστή ως το κριτήριο της διανομής Cauchy και είναι η ίδια με τη διανομή t του Stutent με ένα βαθμό ελευθερίας:

$$f(x) = \frac{1}{\pi(1 + x^2)},$$

για -∞<x<∞ .

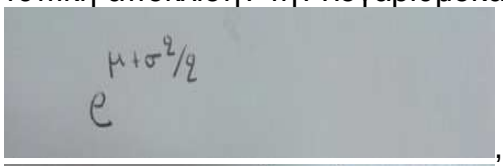
```
par(mfrow=c(1,2))
plot(-200:200,dcauchy(-200:200,0,10),type="l",ylab="p(x)",xlab="x")
plot(-200:200,dcauchy(-200:200,0,50),type="l",ylab="p(x)",xlab="x")
```



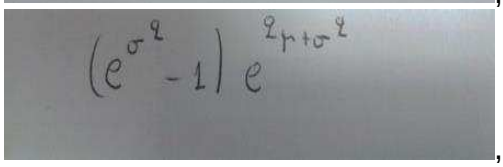
Σημείωσε η πολύ μακριά, παχιά ουρά της διανομής Cauchy. Η αριστερά συνάρτηση πυκνότητας έχει κλίμακα=10 και η δεξιά γραφική παράσταση έχει κλίμακα=50; και οι δύο έχουν τοποθεσία=0.

Η λογαριθμοκανονική κατανομή

Η λογαριθμοκανονική κατανομή παίρνει τιμές στη θετική πραγματική γραμμή. Εάν ο λογάριθμος της α λογαριθμικής αποκλίνει έχει ληφθεί, το αποτέλεσμα είναι μια κανονική απόκλιση, από 'δω το όνομα. Οι αιτήσεις για την κανονική λογαριθμική περιλαμβάνουν την κατανομή των μεγεθών των σωματιδίων στα αδρανή υλικά, οι ροές των πλημμυρών, οι συγκεντρώσεις των ρύπων του αέρα, και οι χρόνοι αποτυχίας. Η συνάρτηση κινδύνου της λογαριθμικής αυξάνεται για μικρές τιμές και στη συνέχεια μειώνεται. Ένα μίγμα από ετερογενή στοιχεία που έχουν μεμονωμένα μονότονες κινδύνους μπορεί να δημιουργήσει μια τέτοια συνάρτηση κινδύνου.
Πυκνότητα, αθροιστική πιθανότητα, ποσοτικοποιήσεις και τυχαία παραγωγή ενέργειας για την λογαριθμοκανονική κατανομή απασχολούν την dlnorm συνάρτηση σαν αυτή:
dlnorm(x, meanlog=0, sdlog=1)
Η μέση τιμή και η τυπική απόκλιση είναι προαιρετική, με προεπιλογή meanlog=0 και sdlog=1. Σημειώστε ότι αυτά δεν είναι η μέση τιμή και την τυπική απόκλιση? την λογαριθμοκανονική κατανομή έχει μέσος

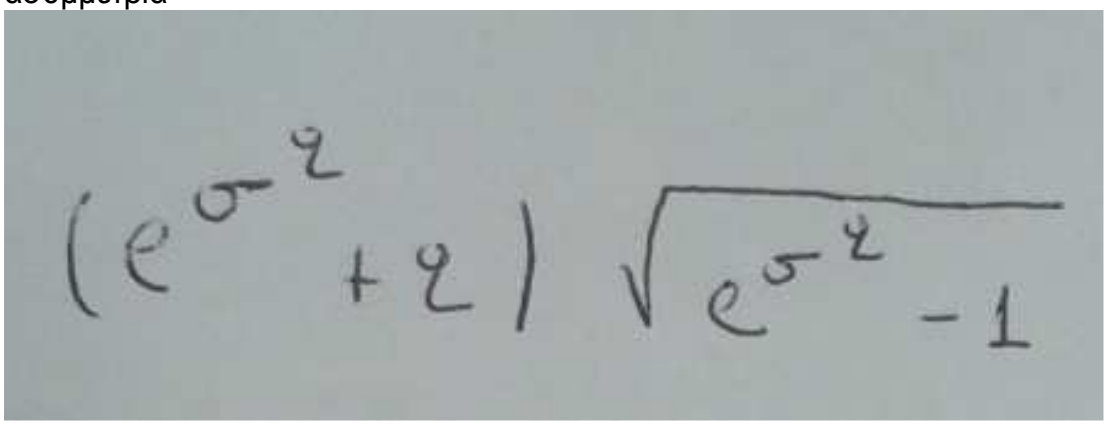


, διακύμανση

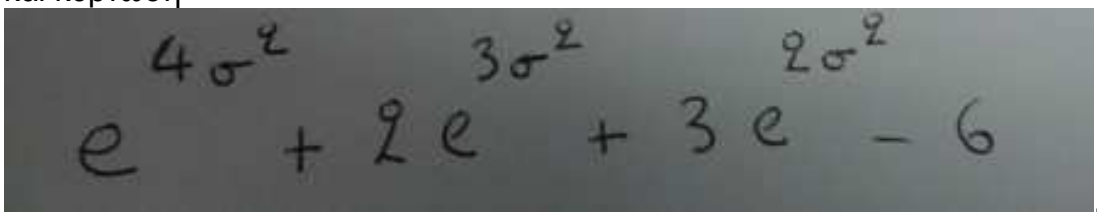


,

ασυμμετρία

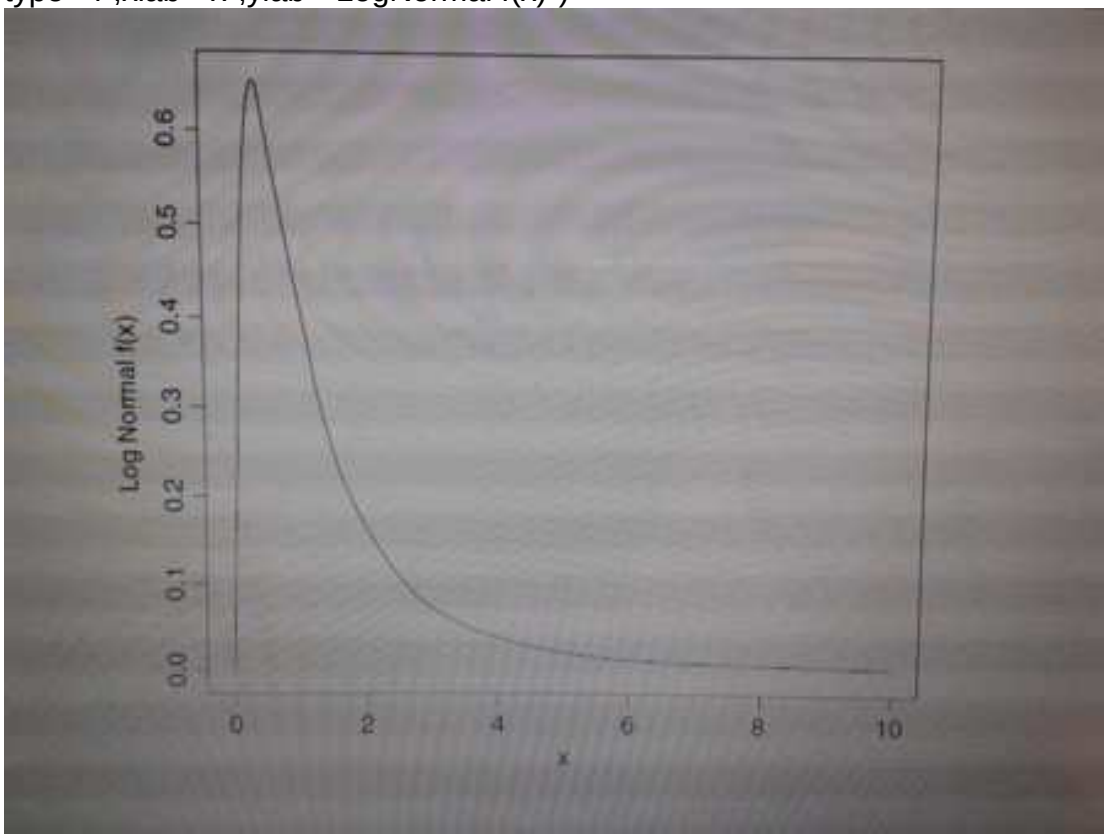


και κύρτωση



.

```
par(mfrow=c(1,1))
plot(seq(0,10,0.05),dlnorm(seq(0,10,0.05)),
type="l",xlab="x",ylab="LogNormal f(x)")
```



Η εξαιρετικά μεγάλη ουρά και υπερβολική θετική κλίση είναι χαρακτηριστικές της λογαριθμικής κατανομής. Λογαριθμικός μετασχηματισμός παρακολουθείται με ανάλυση με φυσιολογικά σφάλματα είναι συχνά κατάλληλος για τα δεδομένα όπως αυτά.

Η λογιστική διανομή

Η λογιστική είναι η κανονική συνάρτηση συνδέσμου σε γενικευμένα γραμμικά μοντέλα με σφάλματα διωνυμικής και περιγράφεται αναλυτικά στο κεφάλαιο 16 σχετικά με την ανάλυση του μεγέθους των δεδομένων. Η αθροιστική πιθανότητα είναι μια συμμετρική σχήματος S διανομή που οριοθετείται από το ανωτέρω 1 και κάτω από 0. Υπάρχουν δύο τρόποι γραφής της αθροιστικής πιθανότητας εξίσωσης:

$$p(x) = \frac{e^{a+bx}}{1+e^{a+bx}}$$

και

$$p(x) = \frac{1}{1+\beta e^{-\alpha x}}$$

Το μεγάλο πλεονέκτημα της πρώτης μορφής είναι ότι η γραμμική μορφοποίηση κάτω από τη λογαριθμική απόδοση μετατροπής (βλ. σελ.. 572), έτσι ώστε

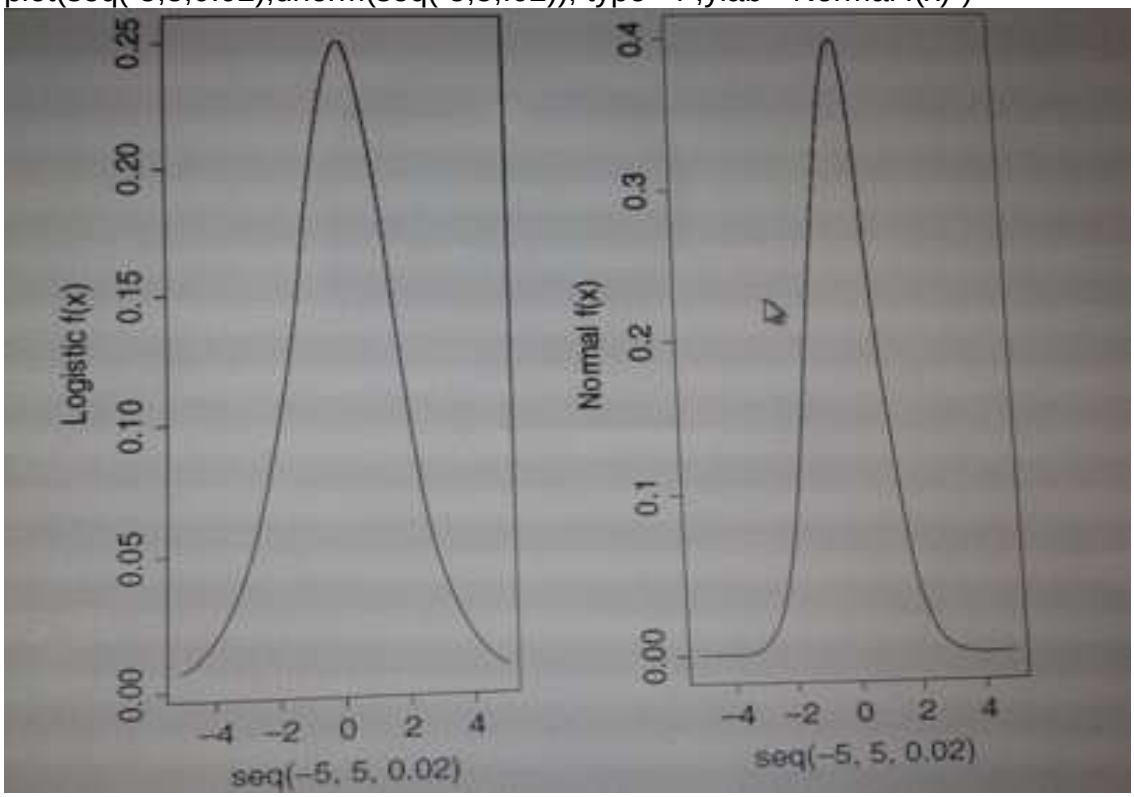
$$\ln\left(\frac{p}{q}\right) = a + bx,$$

όπου ρ είναι η πιθανότητα επιτυχίας και q =(1 – p) είναι η πιθανότητα της αποτυχίας.

Η λογιστική είναι μια μονότροπη, συμμετρική κατανομή στην πραγματική γραμμή με τις ουρές που είναι μεγαλύτερες από την κανονική κατανομή. Συχνά χρησιμοποιείται σε καμπύλες ανάπτυξης μοντέλων, αλλά έχει επίσης χρησιμοποιηθεί σε μελέτες βιολογικής δοκιμασίας και άλλες εφαρμογές. Ένα κίνητρο για τη χρήση της συμβολικής λογικής με καμπύλες αύξησης είναι ότι

η λογιστική συνάρτηση κατανομής f(x) έχει την ιδιότητα ότι η παράγωγος της f(x) σε σχέση με το x είναι ανάλογη με [f(x)-A][B-f(x)] με A<B. Η ερμηνεία είναι ότι ο ρυθμός ανάπτυξης είναι ανάλογος με το ποσό που έχει ήδη αναπτυχθεί, φορές που ισοδυναμεί με τη παραγωγή που ακόμη αναμένεται .

```
 par(mfrow=c(1,2))
plot(seq(-5,5,0.02),dlogis(seq(-5,5,.02)), type="l",ylab="Logistic f(x)")
plot(seq(-5,5,0.02),dnorm(seq(-5,5,.02)), type="l",ylab="Normal f(x)")
```



Εδώ, η συμβολική λογική συνάρτηση πυκνότητας dlogis (αριστερά) σε σύγκριση με μια αντίστοιχη κανονική dnorm συνάρτηση πυκνότητας (δεξιά) χρησιμοποιεί την προεπιλεγμένη μέση τιμή 0 και την τυπική απόκλιση 1 και στις δύο περιπτώσεις. Σημειώστε οι πολύ παχιές ουρές της συμβολικής λογικής (ακόμα σημαντική πιθανότητα στις ± 4 τυπικές αποκλίσεις. Σημειώστε επίσης τη διαφορά στις κλίμακες των δύο αξόνων y (0,25 για τη συμβολική λογική, 0,4 για την κανονική).

Η log-logistic διανομή

Το log-logistic είναι ένα πολύ ευέλικτο τεσσάρων παραμέτρων μοντέλο για την περιγραφή της ανάπτυξης ή αποσύνθεσης διεργασιών:

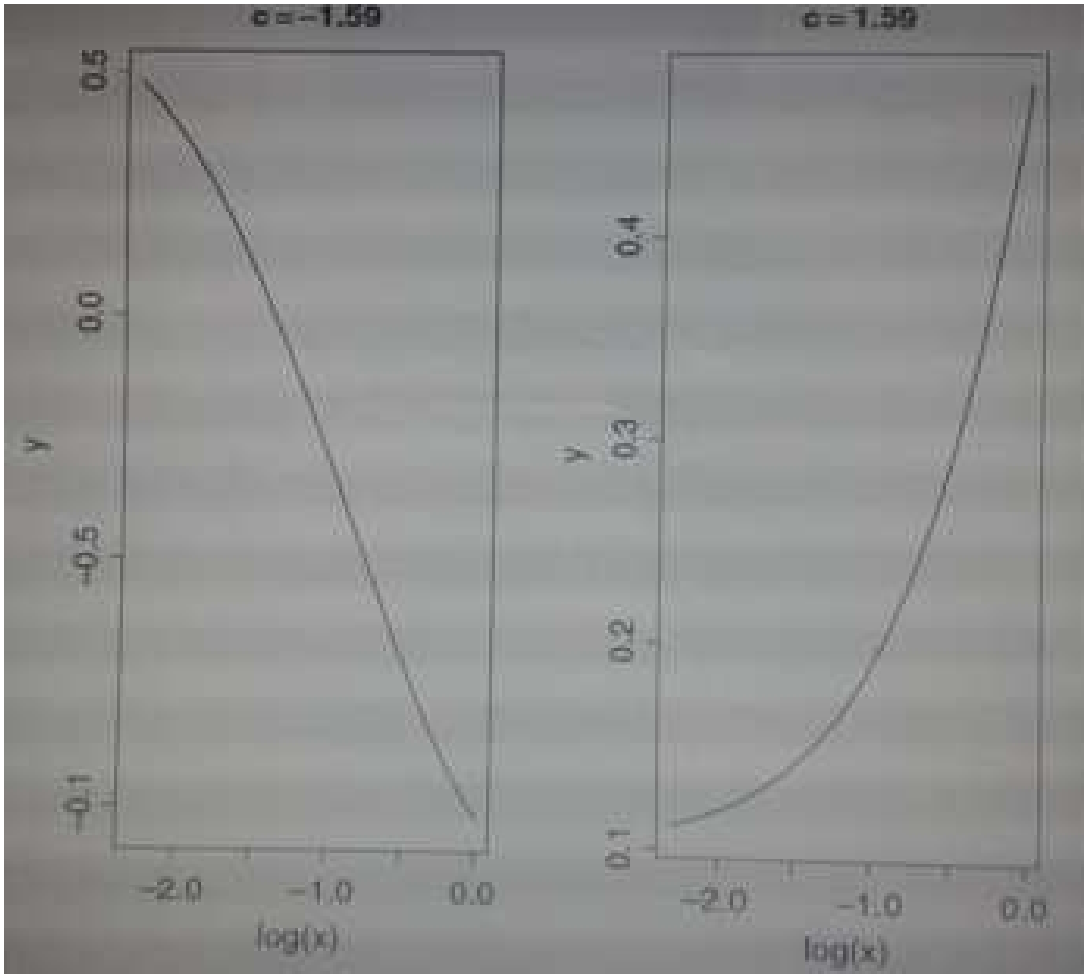
$$y = a + b\left[\frac{\exp(c(\log(x) - d)}{1 + \exp(c(\log(x) - d)}\right].$$

Εδώ είναι δύο περιπτώσεις. Η πρώτη είναι ένα αρνητικό σιγμοειδές με c = -1, 59 και α = -1, 4:

```
x<-seq(0.1,1,0.01)
y<- -1.4+2.1*(exp(-1.59*log(x)-1.53)/(1+exp(-1.59*log(x)-1.53)))
plot(log(x),y,type="l", main="c = -1.59")
```

Για το δεύτερο έχουμε c = 1, 59 και α=0,1:

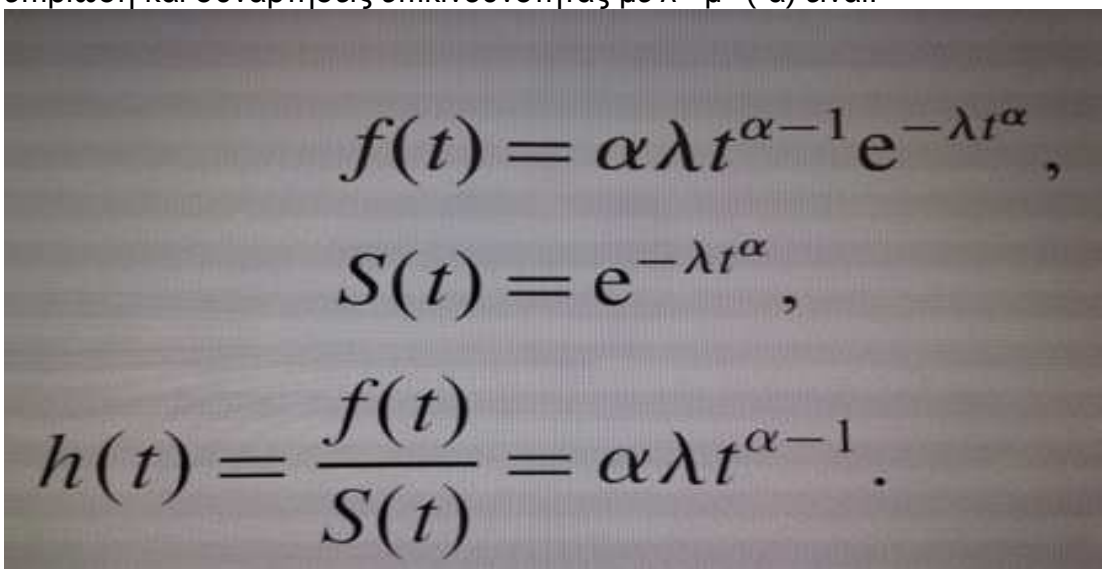
```
y<-0.1+2.1*(exp(1.59*log(x)-1.53)/(1+exp(1.59*log(x)-1.53)))
plot(log(x),y,type="l",main="c = -1.59")
```



Η κατανομή Weibull:

Η προέλευση της Weibull κατανομής είναι πιο αδύναμος κρίκος ανάλυσης. Εάν υπάρχουν r συνδέσεις σε μια αλυσίδα, και τα πλεονεκτήματα του κάθε συνδέσμου Zi είναι ανεξάρτητα διανέμονται σε (0,∞) στη συνέχεια, η κατανομή των πιο αδύναμων κρίκων V = min(Zj) προσεγγίζει την κατανομή Weibull, καθώς ο αριθμός των συνδέσεων αυξάνεται.
Η Weibull είναι ένα δύο παραμέτρων μοντέλο που έχει την εκθετική κατανομή ως ειδική περίπτωση. Η τιμή του στις δημογραφικές μελέτες και ανάλυση

επιβίωσης είναι ότι επιτρέπει για το ποσοστό των θανάτων να αυξάνεται ή να μειώνεται με την ηλικία, έτσι ώστε και οι τρεις τύποι της καμπύλης επιβίωσης μπορούν να αναλυθούν (όπως εξηγείται στη σελίδα. 802). Η πυκνότητα, επιβίωση και συναρτήσεις επικινδυνότητας με λ =μ^ (-α) είναι:

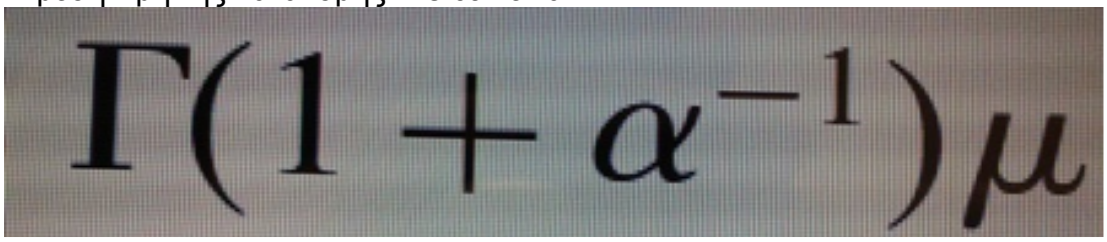


$$f(t) = \alpha\lambda t^{\alpha-1} e^{-\lambda t^{\alpha}},$$

$$S(t) = e^{-\lambda t^{\alpha}},$$

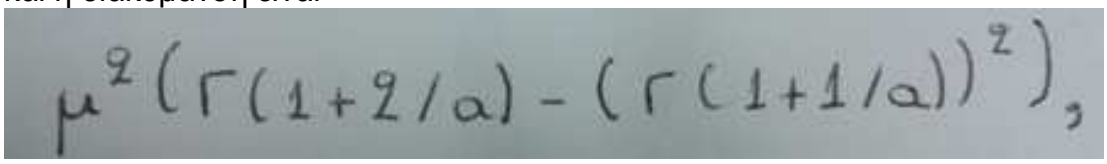
$$h(t) = \frac{f(t)}{S(t)} = \alpha\lambda t^{\alpha-1}.$$

Η μέση τιμή της κατανομής Weibull είναι



$$\Gamma(1+\alpha^{-1})\mu$$

και η διακύμανση είναι



$$\mu^2\left(\Gamma(1+2/\alpha) - \left(\Gamma(1+1/\alpha)\right)^2\right),$$

και η παράμετρος α περιγράφει τη μορφή της συνάρτησης κινδύνου (το φόντο για τον προσδιορισμό των εξισώσεων πιθανότητας δίνεται από Aitkin και άλλα (1989, σσ. 281-283). Για α=1 (η εκθετική κατανομή), η επικινδυνότητα είναι σταθερή, ενώ για α>1 ο κίνδυνος αυξάνεται με την ηλικία και για α<1 ο κίνδυνος μειώνεται με την ηλικία.

Επειδή η Weibull, λογαριθμική κανονική συνάρτηση και λογαριθμική συμβολική λογική έχουν όλες θετική ασυμμετρία, είναι δύσκολο να γίνει διάκριση μεταξύ τους με μικρά δείγματα. Αυτό είναι ένα σημαντικό πρόβλημα, διότι κάθε διανομή έχει διαφορετικό σχήμα συναρτήσεων επικινδυνότητας, και θα είναι δύσκολο, συνεπώς, να γίνει διάκριση μεταξύ διαφορετικών υποθέσεων σχετικά με την ηλικία-εξειδίκευσης των ποσοστών θανάτου. Σε μελέτες επιβίωσης, η οικονομία απαιτεί ότι ταιριάζουμε την εκθετική και όχι η Weibull, εκτός εάν η παράμετρος σχήματος α είναι σημαντικά διαφορετική από 1.

Εδώ είναι μια οικογένεια από τρεις Weibull κατανομές με α=1,2 και 3 (διακεκομμένες, παύλες και συμπαγείς γραμμές, αντίστοιχα). Σημειώστε ότι για μεγάλες τιμές του α η κατανομή γίνεται συμμετρική, ενώ για α≤1 η κατανομή έχει ρυθμό του σε t=0.

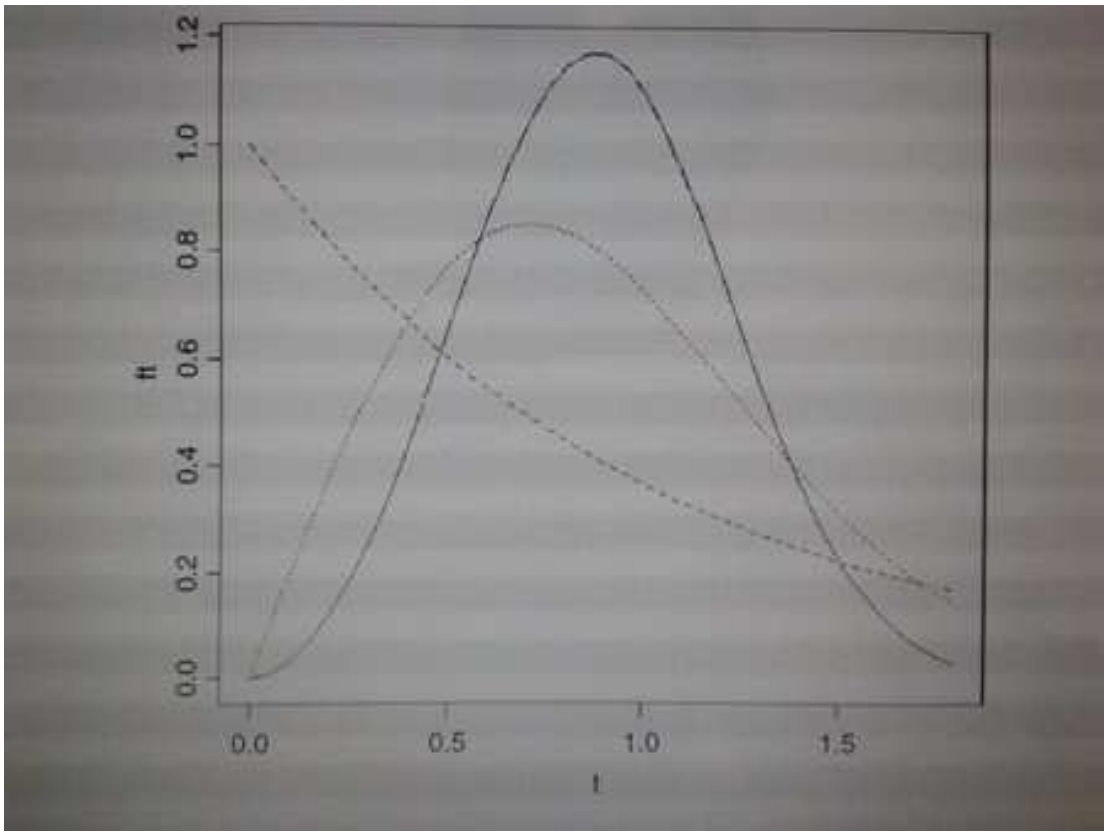
```
a<-3
l<-1
t<-seq(0,1.8,.05)
ft<-a*l*t^(a-1)*exp(-l*t^a)
plot(t,ft,type="l")
a<-1
ft<-a*l*t^(a-1)*exp(-l*t^a)
lines(t,ft,type="l",lty=2)
a<-2
ft<-a*l*t^(a-1)*exp(-l*t^a)
lines(t,ft,type="l",lty=3)
```

Πολυδιάστατη κανονική κατανομή

Αν θέλετε να δημιουργήσετε δύο (ή περισσότερα) ανύσματα ή κανονικά κατανεμημένους τυχαίους αριθμούς που συσχετίζονται με το ένα ή το άλλο σε ένα καθορισμένο βαθμό, τότε εσείς χρειάζεστε τη συνάρτηση mvrnorm από τη βιβλιοθήκη MASS:

```
library(MASS)
```

Ας υποθέσουμε ότι θέλουμε δύο διανύσματα από 1000 τυχαίους αριθμούς το κάθε ένα. Το πρώτο δυάνυσμα έχει μια μέση τιμή από 50 και το δεύτερο έχει μία μέση τιμή από 60. Η διαφορά από εντολή rnorm είναι ότι θα πρέπει να καθορίσετε το γινόμενο τους, καθώς και τις τυπικές αποκλίσεις για κάθε ξεχωριστή μεταβλητή. Αυτό επιτυγχάνεται με ένα θετικό-οριστικό συμμετρικό πίνακα που προσδιορίζει το γινόμενο πίνακα από τις μεταβλητές.



```
xy<-mvrnorm(1000,mu=c(50,60),matrix(c(4,3.7,3.7,9),2))
```

Μπορούμε να ελέγξουμε πόσο κοντά είναι τα τετράγωνα τυπικής απόκλισης σε συγκεκριμένες τιμές μας:

```
var(xy)

             [,1]     [,2]
[1,]  3.983063  3.831880
[2,]  3.831880  8.922865
```

Δεν είναι κακό: είπαμε το γινόμενο θα πρέπει να είναι 3,70 και τα δεδομένα προσομοίωσης είναι 3,83. Εξάγουμε τα δύο ξεχωριστά διανύσματα x και y και σχεδιάζουμε αυτά να εξετάσουν τη συσχέτιση

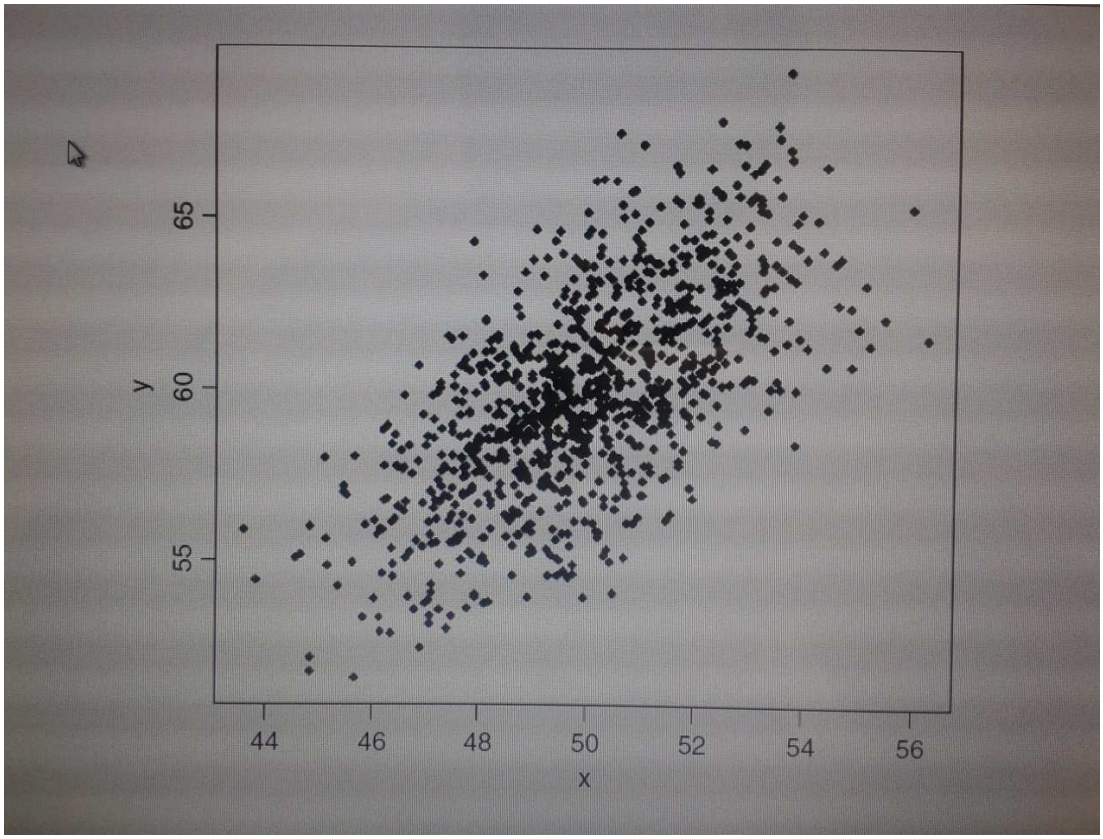
```
x<-xy[,1]
y<-xy[,2]
plot(x,y,pch=16,ylab="y",xlab="x")
```

Αξίζει κοιτάζοντας τα τετράγωνα τυπικής απόκλισης από x και y με περισσότερη λεπτομέρεια :

```
var(x)
[1] 3.983063
var(y)
[1] 8.922865
```

Εάν τα δύο δείγματα ήταν ανεξάρτητα, τότε το τετράγωνο τυπικής απόκλισης του αθροίσματος των δύο μεταβλητών θα είναι ίσο με το άθροισμα των δύο τετράγωνων τυπικής απόκλισης. Είναι αυτή η περίπτωση εδώ;

```
var(x+y)
[1] 20.56969
```

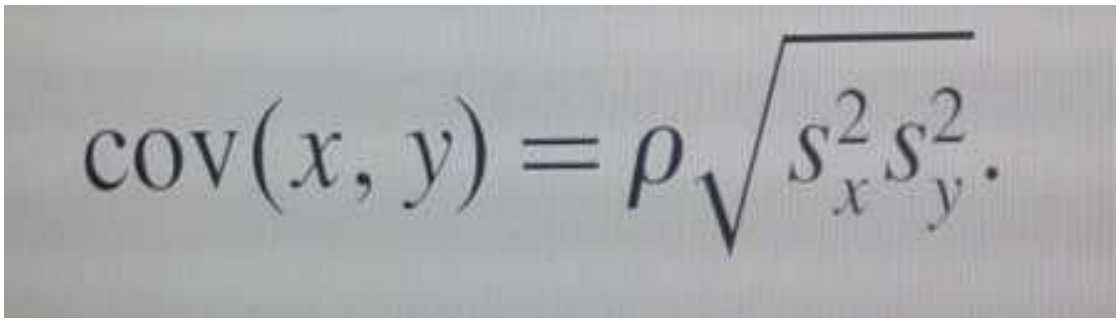



```
var(x)+var(y)
[1] 12.90593
```

Όχι, δεν είναι. Το τετράγωνο τυπικής απόκλισης του αθροίσματος (20,57) είναι πολύ μεγαλύτερο από το άθροισμα των τετραγώνων τυπικής απόκλισης (12.91). Αυτό συμβαίνει επειδή τα x και y συσχετίζονται θετικά? μεγάλες τιμές του χ τείνουν να συνδέονται με τις μεγάλες τιμές του y και αντίστροφα. Αυτό είναι έτσι, θα περιμέναμε το τετράγωνο τυπικής απόκλισης της διαφοράς μεταξύ x και y να είναι μικρότερη από το άθροισμα των δύο τετραγώνων τυπικής απόκλισης:

```
 var(x-y)
[1] 5.242167
```

Όπως είχε προβλεφθεί, το τετράγωνο τυπικής απόκλισης της διαφοράς (5,24) είναι πολύ μικρότερο από το άθροισμα των τετραγώνων τυπικών αποκλίσεων (12,91). Καταλήγουμε στο συμπέρασμα ότι το τετράγωνο τυπικής απόκλισης του αθροίσματος των δύο μεταβλητών είναι μόνο ίσο με το τετράγωνο τυπικής απόκλισης της διαφοράς των δύο μεταβλητών, όταν οι δύο μεταβλητές είναι ανεξάρτητες. Τι γίνεται με το γινόμενο των x και y; Βρήκαμε αυτό ήδη με την εφαρμογή της var συνάρτησης με την matrix *xy* (ανωτέρω). Εμείς καθορίσαμε ότι το γινόμενο θα πρέπει να είναι 3,70 στην καλούμενη πολυμεταβλητή κανονική κατανομή, και η διαφορά μεταξύ 3.70 και 3.831 880 απλά οφείλεται στη τυχαία επιλογή των σημείων. Το γινόμενο συνδέεται με τις αυτόνομα τετράγωνα τυπικής απόκλισης, μέσω του συντελεστή συσχέτισης ρ ως εξής (βλ. σελ. 310.):



$$\operatorname{cov}(x, y) = \rho \sqrt{s_x^2 s_y^2}.$$

Για το παράδειγμά μας, αυτό ελέγχει ως εξής, όπου η τιμή δείγμα ρ είναι cor(x,y)

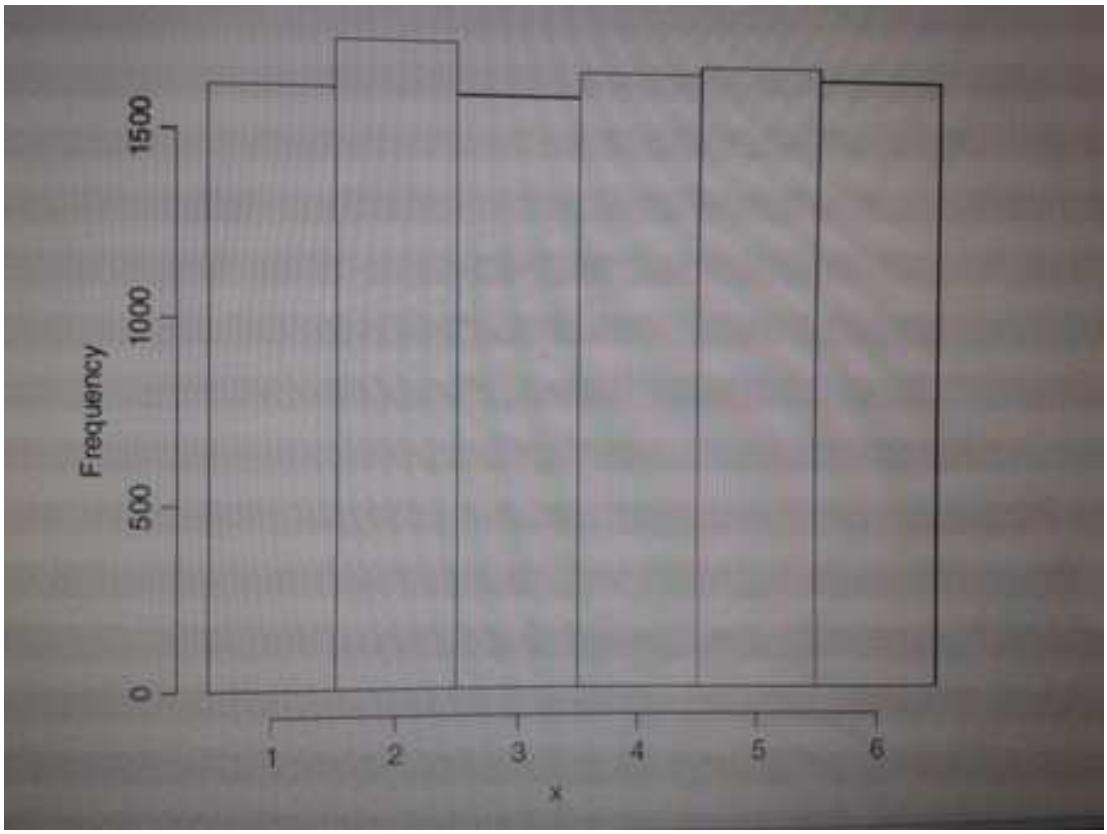
```
cor(x,y)*sqrt(var(x)*var(y))
[1] 3.83188
```

που παρατηρείται στο γινόμενο μας μεταξύ x και y μαζί ρ=0,642 763 5.

Η ομοιόμορφη κατανομή

Αυτή είναι η διανομή που η γεννήτρια τυχαίων αριθμών στο κομπιουτεράκι σας ελπίζει να μιμηθούν. Η ιδέα είναι να δημιουργήσει αριθμούς μεταξύ 0 και 1, όπου κάθε πιθανός πραγματικός αριθμός σε αυτό το διάστημα έχει ακριβώς την ίδια πιθανότητα να παραχθεί. Αν έχετε σκεφτεί αυτό, θα έχουν συμβεί σε σας ότι κάτι δεν πάει καλά εδώ. Υπολογιστές παράγουν αριθμούς με παρακάτω συνταγές. Εάν ακολουθείτε μια συνταγή, τότε η έκβαση είναι προβλέψιμη. Αν η έκβαση είναι προβλέψιμη, τότε πώς μπορεί να είναι τυχαία; Ως John von Neumann είπε κάποτε: «Όποιος χρησιμοποιεί αριθμητικές μεθόδους για να παράγει τυχαίους αριθμούς είναι σε μια κατάσταση της αμαρτίας ». Αυτό εγείρει το ερώτημα ως προς το τι, ακριβώς, ένας υπολογιστής που παρήγαγε τυχαίο αριθμό είναι. Η απάντηση αποδεικνύεται ότι είναι επιστημονικώς πολύ ενδιαφέρουσα και πολύ σημαντική για την μελέτη της κρυπτογράφησης (για παράδειγμα, κάθε ακολουθία ψευδοτυχαίων αριθμών που παράγεται από μία γραμμική αναδρομή είναι επισφαλής, διότι από μια επαρκώς μακρά υποακολουθία των εξόδων, μπορεί κάποιος να προβλέψει την υπόλοιπη από τις έξοδους). Αν σας ενδιαφέρει, αναζητήστε τον ανεμοστρόβιλο Mersenne σε απευθείας σύνδεση. Εδώ θα ασχοληθούμε μόνο με το πόσο καλά η σύγχρονη ψευδο-γεννήτρια τυχαίων αριθμών εκτελεί. Εδώ είναι το αποτέλεσμα της R συνάρτησης runif προσομοιώνοντας την ρίψη 6 όψεων που πεθαίνουν 10 000 φορές: το ιστόγραμμα θα πρέπει να είναι επίπεδο:

```
x<-ceiling(runif(10000)*6)
table(x)
x
   1    2    3    4    5    6
1620 1748 1607 1672 1691 1662
hist(x,breaks=0.5:6.5,main="")
```



Αυτό είναι αξιοσημείωτα κοντά στη θεωρητική προσδοκία, αντανακλώντας την πολύ υψηλή αποδοτικότητα των τυχαίων αριθμών της γεννήτριας R.
Δοκιμάστε χαρτογράφηση 1 000 000 σημείων για να ψάξουν για κενά:

```
x<-runif(1000000)
y<-runif(1000000)
plot(x,y,pch=16)
```

Αυτό παρείχε ένα απόλυτα στερεό μαύρο χάρτη για μένα: δεν υπήρχαν τρύπες στο σχέδιο, έτσι δεν υπήρχαν ζεύγη αριθμών που δεν παράγονται σε αυτή την ανάλυση της κλίμακας (pch=16). Για ένα πιο πλήρη έλεγχο μπορούμε να υπολογίζουμε τη συχνότητα των συνδυασμών αριθμών: με 36 κελιά, η αναμενόμενη συχνότητα είναι 1 000 000/36 = 27 777, 78: αριθμοί ανά κύτταρο. Χρησιμοποιούμε τη συνάρτηση για την παραγωγή 36 κομμένων κουτιών αποθήκευσης:

```
table(cut(x,6),cut(y,6))
```

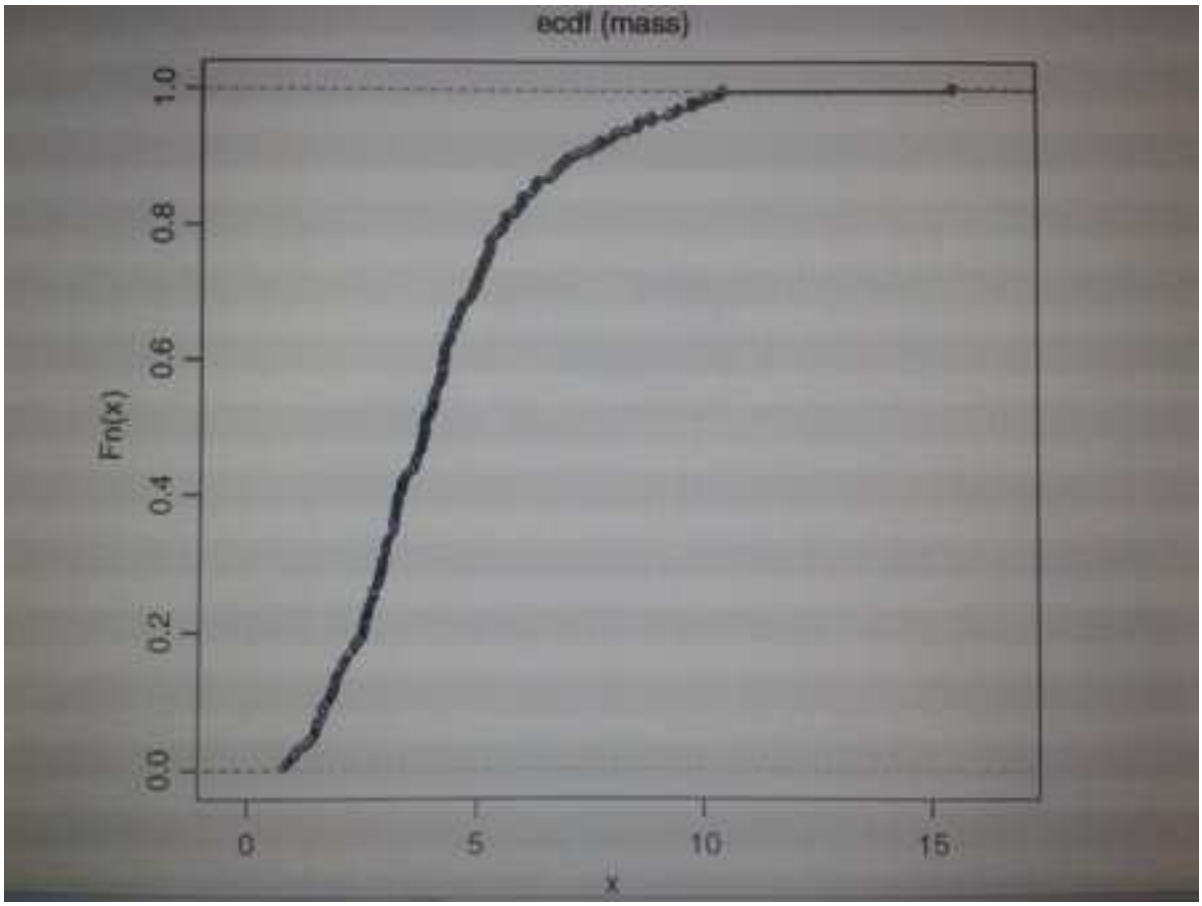
	(-0.001,0.166]	(0.166,0.333]	(0.333,0.5]	(0.5,0.667]	(0.667,0.834]	(0.834,1]
(-0.001,0.166]	27541	27795	27875	27851	27664	27506
(0.166,0.333]	27908	28033	27975	27859	27862	27600
(0.333,0.5]	27509	27827	27991	27689	27878	27733
(0.5,0.667]	27718	28074	27548	28062	27777	27760
(0.667,0.834]	27820	28084	27466	27753	27784	27454
(0.834,1]	27463	27997	27982	27685	27571	27906

Όπως μπορείτε να δείτε οι παρατηρούμενες συχνότητες είναι εξαιρετικά κοντά στην πρόβλεψη.

Χάραξη εμπειρικής αθροιστικής συνάρτησης κατανομής

Η ecdf συνάρτηση χρησιμοποιείται για να υπολογίσετε ή να σχεδιάσετε μια εμπειρική συνάρτηση αθροιστικής κατανομής. Εδώ είναι σε δράση για τα δεδομένα ψάρια (σελ. 220 και 230):

```
fishes<-read.table("c:\\temp\\fishes.txt",header=T)
attach(fishes)
names(fishes)
[1]   "mass"
plot(ecdf(mass))
```



Η έντονη θετική λοξότητα στα δεδομένα είναι προφανές από το γεγονός ότι η αριστερή πλευρά της σωρευτικής κατανομής είναι πολύ πιο απότομη από τη δεξιά πλευρά (βλ. και σελ.. 230).

Διακριτές κατανομές πιθανότητας

Η κατανομή Bernoulli

Αυτή είναι η διανομή υποκείμενων δοκιμών με μια δυαδική μεταβλητή απόκρισης.

Την απάντηση παίρνει μια από τις μόλις δύο τιμές: είναι 1 με πιθανότητα ρ (μια «επιτυχία») και είναι 0 με πιθανότητα 1-ρ (μια «αποτυχία»). Η συνάρτηση πυκνότητος δίδεται από την:

$$p(X) = p^x(1-p)^{1-x}$$

Ο ορισμός του στατιστικού τετραγώνου τυπικής απόκλισης είναι η προβλεψη της x^2 μείον το τετράγωνο της πρόβλεψης του x:

$$\sigma^2 = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2.$$

Μπορούμε να δούμε πώς αυτό λειτουργεί με μια απλή κατανομή όπως η Bernoulli. Υπάρχουν μόλις δύο αποτελέσματα σε f(x): μια επιτυχία, που x=1 με πιθανότητα ρ και μια αποτυχία, που x=0 με πιθανότητα 1-ρ. Έτσι, η προσδοκία του x είναι

$$\mathrm{E}(X) = \sum xf(x) = 0 \times (1-p) + 1 \times p = 0 + p = p$$

και η προσδοκία του x^2 είναι:

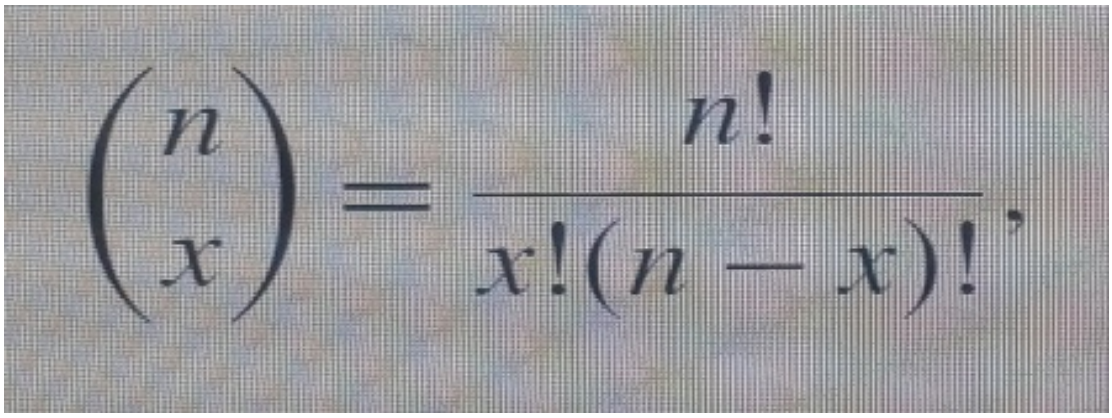
$$\mathrm{E}(X^2) = \sum x^2 f(x) = 0^2 \times (1-p) + 1^2 \times p = 0 + p = p.$$

έτσι ώστε η διακύμανση του Bernoulli είναι

$$\mathrm{var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = p - p^2 = p(1-p) = pq.$$

Η διωνυμική κατανομή

Αυτό είναι μια παραμέτρος διανομής η οποία ρ περιγράφει την πιθανότητα επιτυχίας σε μια δυαδική δοκιμή. Η πιθανότητα από x επιτυχίες έξω από n προσπάθειες δίνεται από το γινόμενο μαζί με τη πιθανότητα παρατήρησης μιας ειδικής σχέσης και του αριθμού των τρόπων που παίρνεται από αυτή την σχέση.
Χρειαζόμαστε έναν τρόπο γενίκευσης του αριθμό των τρόπων για να πάρει x τεμάχια έξω από n τεμάχια. Η απάντηση είναι η συνδυαστική φόρμουλα



$$\binom{n}{x} = \frac{n!}{x!(n-x)!},$$

όπου το «θαυμαστικό» σημαίνει «παραγοντικό». Για παράδειγμα, 5!=5×4×3×2=120. Αυτή η φόρμουλα έχει τεράστια πρακτική χρησιμότητα. Σας δείχνει ταυτόχρονα, για παράδειγμα, πόσο απίθανο είναι να κερδίζετε την Εθνική Λοταρία στην οποία καλείστε να επιλέξετε έξι αριθμούς μεταξύ 1 και 49. Μπορούμε να χρησιμοποιήσουμε την ενσωματωμένη συνάρτηση παραγοντικό για το σκοπό αυτό

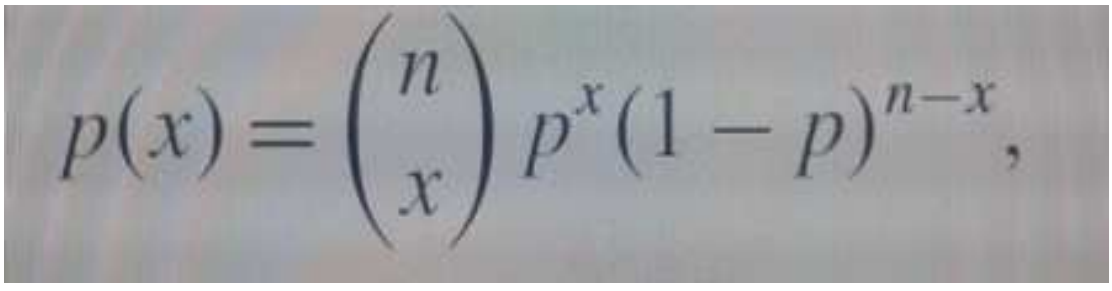
```
factorial(49)/(factorial(6)*factorial(49-6))
[1]      13983816
```

η οποία είναι περίπου 1 στις 14 εκατομμύρια πιθανότητες να κερδίσει το τζακ ποτ. Είστε πιο πιθανό να πεθάνουν μεταξύ των τιμών αγοράς του εισιτηρίου σας και άκουσε το αποτέλεσμα της κλήρωσης. Όπως είδαμε (σελ. 11), υπάρχει μια ενσωματωμένη συνάρτηση R για τη συνδυαστική συνάρτηση

```
choose(49,6)
[1]   13983816
```

και χρησιμοποιούμε την συνάρτηση Choose από εδώ.

Η γενική μορφή της διωνυμικής κατανομής δίνεται από

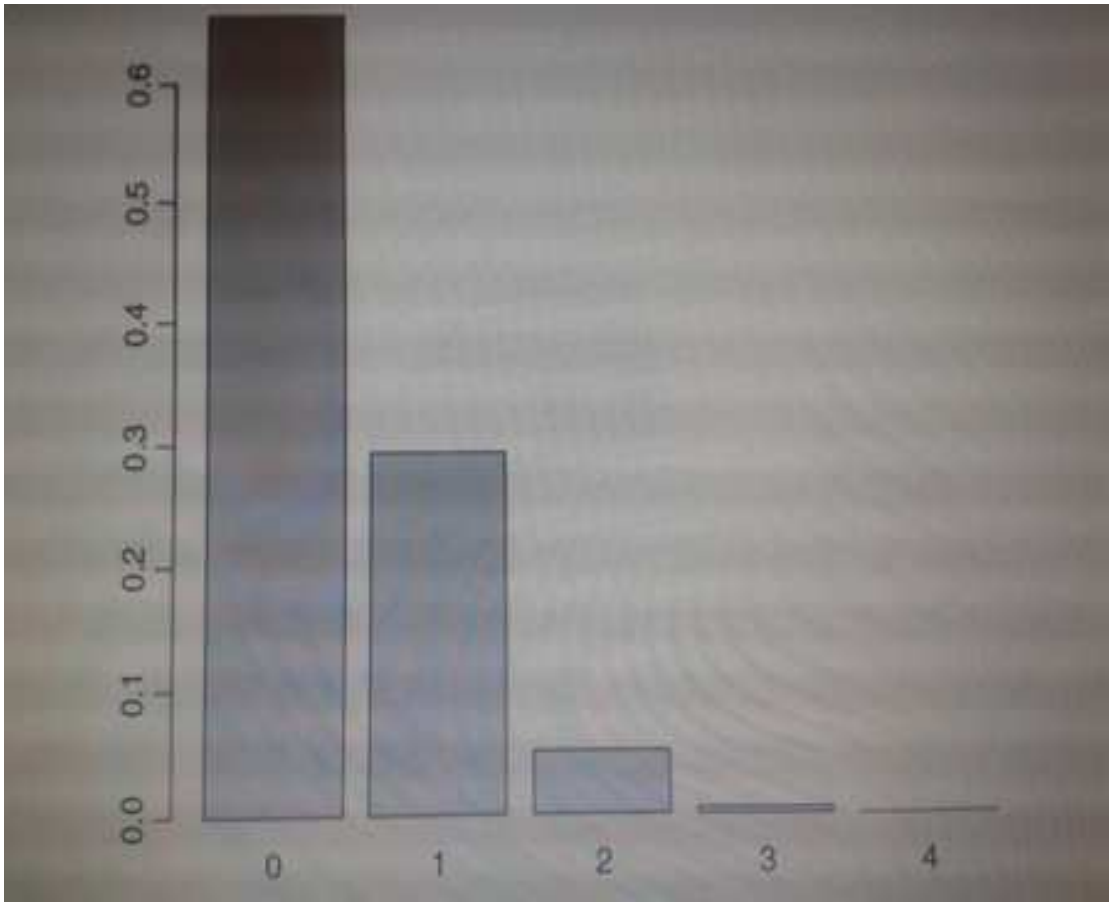


$$p(x) = \binom{n}{x} p^x (1-p)^{n-x},$$

χρησιμοποιώντας τον συνδυαστικό παραπάνω τύπο. Η μέση τιμή της διωνυμική κατανομής είναι np και η διακύμανση είναι np(1−p).

Αφού 1-ρ είναι μικρότερο από 1, είναι προφανές ότι το τετράγωνο τυπικής απόκλισης είναι μικρότερο από το μέσο όρο για την διωνυμική κατανομή (εκτός, βεβαίως, στην οριακή περίπτωση, όταν ρ = 0 και το τετράγωνο τυπικής απόκλισης είναι 0). Είναι εύκολο να απεικονίσει την κατανομή για συγκεκριμένες τιμές των n και p.

```
p<-0.1
n<-4
x<-0:n
px<-choose(n,x)*p^x*(1-p)^(n-x)
barplot(px,names=as.character(x))
```

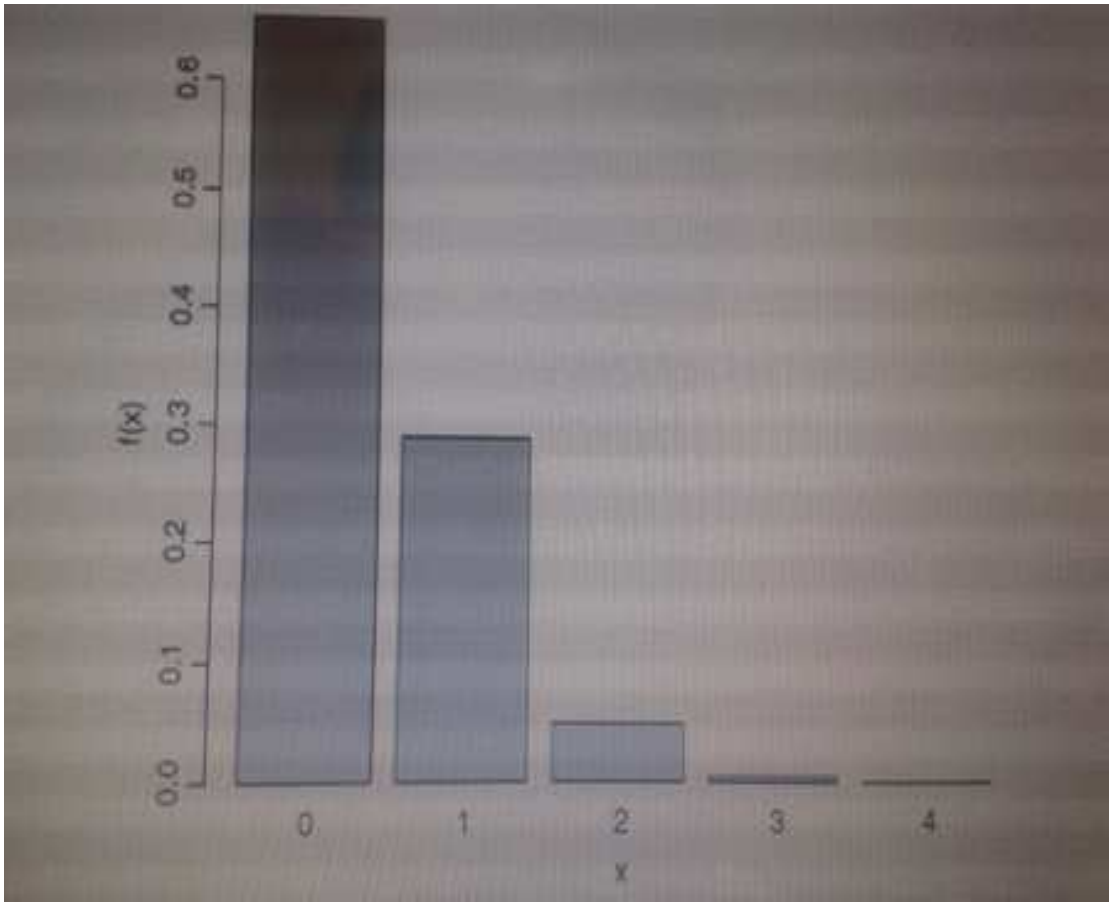


Οι τέσσερις υπάρχουσες συναρτήσεις διανομής για τη διωνυμική στο R (πυκνότητα, αθροιστική πιθανότητα, ποσοστιαίων και τυχαία παραγωγή) που χρησιμοποιούνται σαν αυτό:

dbinom(x, size, prob)

Η συνάρτηση πυκνότητας δείχνει την πιθανότητα για την καθορισμένη καταμέτρηση x (π.χ. ο αριθμός των παρασιτισμένων ψαριών) από ένα δείγμα μεγέθους n , με πιθανότητα επιτυχίας = prob. Έτσι, αν έχουμε πιάσει τέσσερα ψάρια, όταν το 10% είναι παράσιτα στον πληθυσμό του γονέα, έχουμε size = 4 και prob = 0, 1, έτσι μια γραφική παράσταση της πυκνότητας πιθανότητας έναντι αριθμού παρασιτών ψαριών μπορεί να λαμβάνεται σαν αυτό:

```
barplot(dbinom(0:4,4,0.1),names=as.character(0:4),xlab="x",ylab="f(x)")
```

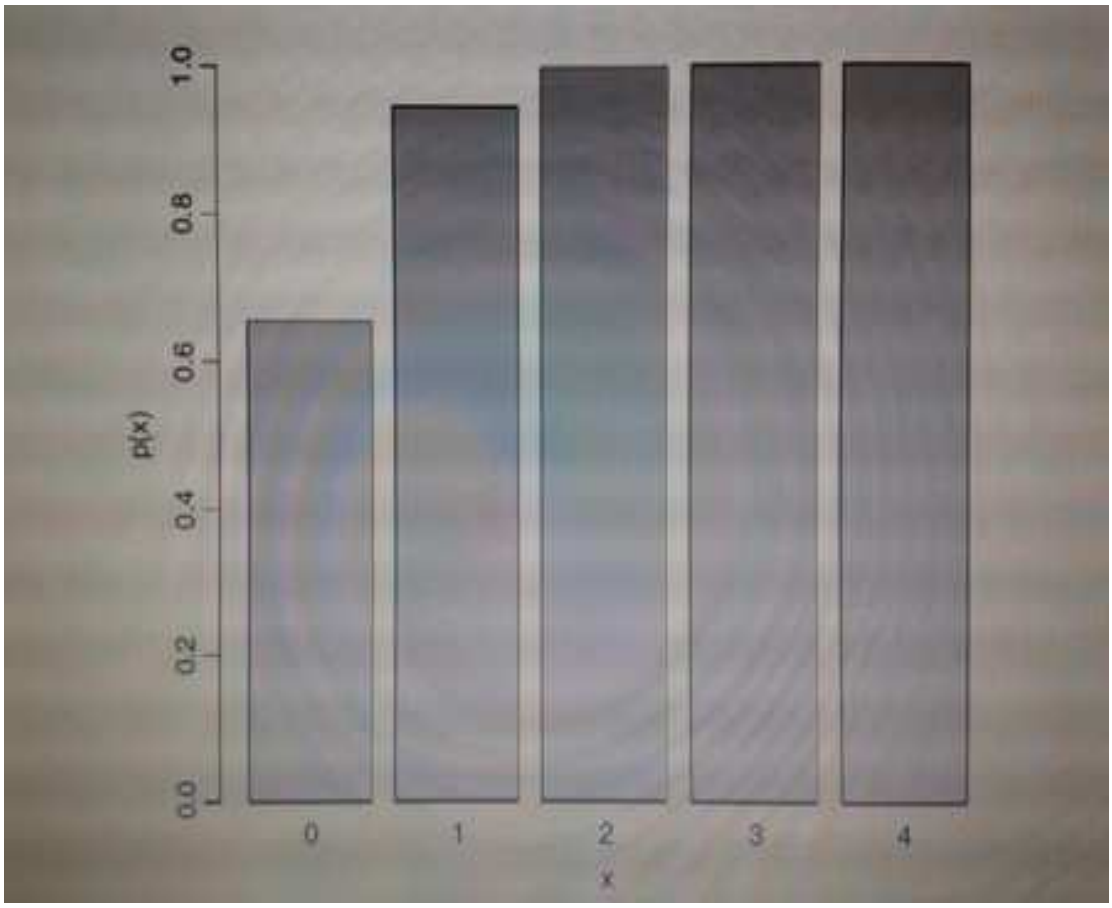


Ο πιο πιθανός αριθμός παρασίτων ψαριών είναι 0. Σημειώστε ότι μπορούμε να δημιουργήσουμε την ακολουθία των τιμών x θέλουμε να σχεδιάσουμε (0:4 σε αυτή την περίπτωση) εντός της συνάρτησης πυκνότητας.
Η αθροιστική πιθανότητα δείχνει το άθροισμα των πυκνοτήτων πιθανότητας μέχρι και μαζί με p(x), σχεδιάζουμε την αθροιστική πιθανότητα απέναντι στον αριθμό των επιτυχιών, για ένα δείγμα n = το μέγεθος και την πιθανότητα = prob. Το ψαροειδές μας σχέδιο μοιάζει με αυτό:

```
barplot(pbinom(0:4,4,0.1),names=as.character(0:4),xlab="x",ylab="p(x)")
```

Αυτό λέει ότι η πιθανότητα να πάρει 2 ή λιγότερα παράσιτα ψάρια έξω από ένα δείγμα από 4 είναι πολύ κοντά στο 1.
Η συνάρτηση ποσοστιαίων σημείων ζητά «με καθορισμένη πιθανότητα p (συνήθως 0.025 και 0.975 για στατιστικό τεστ με δύο τιμές 95% δοκιμές), τι είναι ο αναμενόμενος αριθμός των ψαριών που θα αλιεύονται σε ένα δείγμα μεγέθους n και πιθανότητα= prob ; ». Έτσι, για παράδειγμα μας, το στατιστικό τεστ με δύο τιμές (λογοπαίγνιο όχι σκόπιμο) χαμηλότερο και ανώτερο 95% αναμένεται να πιάσουν τα παράσιτα ψάρια είναι

```
qbinom(.025,4,0.1)
[1]  0
qbinom(.975,4,0.1)
[1]  2
```



πράγμα που σημαίνει ότι με βεβαιότητα 95% θα πιάσουμε μεταξύ 0 και 2 παράσιτα ψάρια έξω από το 4 αν επαναλάβουμε την άσκηση δειγματοληψίας. Είμαστε πολύ πιθανό να πάρουμε 3 ή περισσότερα παράσιτα ψάρια έξω από ένα δείγμα από 4, εάν η παράσιτων αναλογία είναι πραγματικά 0.1.

Αυτό το είδος του υπολογισμού είναι πολύ σημαντικό στους υπολογισμούς κατανάλωσης ρεύματος το οποίο μας ενδιαφέρει να διαπιστωθεί κατά πόσον ή όχι επιλεγμένο μέγεθος δείγματος (n = 4 σε αυτήν την περίπτωση) είναι ικανό να κάνει τη δουλειά που ζητάμε από αυτό. Ας υποθέσουμε ότι το θεμελιώδες ερώτημα της έρευνας μας είναι το κατά πόσον ή όχι το παράσιτο είναι παρόν σε μία δεδομένη λίμνη. Αν βρείτε ένα ή περισσότερα παράσιτα ψαριών, τότε η απάντηση είναι σαφώς «ναι». Αλλά πόσο πιθανό είναι εμείς να χάσομε την αλίευση οποιαδήποτε παράσιτων ψαριών και για αυτό το λόγο καταλήγουμε, λανθασμένα, ότι τα παράσιτα δεν είναι παρόντες στη λίμνη; Με έξω μέγεθος δείγματος n = 4 και p = 0,1 έχουμε μια πιθανότητα να λείπει το παράσιτο 0,9 για κάθε ψάρι που αλιεύεται και ως εκ τούτου σε μια πιθανότητα 0,96 ^ 4 = 0,6561 της λείπει τελείως η εύρεση του παράσιτου. Αυτό είναι προφανώς ανεπαρκές. Πρέπει να σκεφτούμε και πάλι για το μέγεθος του δείγματος. Ποιο είναι το μικρότερο δείγμα, n, που καθιστά την πιθανότητα των αγνοούμενων των παράσιτων συνολικά λιγότερη από 0.05;

Πρέπει να λύσουμε

Λαμβάνοντας λογάριθμους, $0,05 = 0,9^n$

$\log(0,05) = n \log(0,9)$

έτσι

n= log(0,05)/ log(0,9)= 28.433 16

πράγμα που σημαίνει ότι για να κάνει το ταξίδι μας αξίζει τον κόπο εμείς θα πρέπει να κρατήσουμε το ψάρεμα μέχρις ότου έχουμε περισσότερα από 28 μη παρασιτικά ψάρια, πριν απορρίψουμε την υπόθεση ότι ο παρασιτισμός είναι παρόν σε ποσοστό 10%. Φυσικά, θα χρειαστεί ένα πολύ μεγαλύτερο δείγμα για να απορρίψουμε μια υπόθεση της παρουσίας σε πολύ χαμηλότερο ρυθμό.

Οι τυχαίοι αριθμοί προκύπτουν από το διωνυμική κατανομή, όπως αυτοί. Το πρώτο όρισμα είναι ο αριθμός των τυχαίων αριθμών που θέλουμε. Το δεύτερο όρισμα είναι το μέγεθος του δείγματος (n = 4) και το τρίτο είναι η πιθανότητα επιτυχίας (p = 0,1).

```
rbinom(10,4,0.1)
[1]  0  0  0  0  0  1  0  1  0  1
```

Εδώ επαναλαμβάνεται η δειγματοληψία των 4 ψαριών δέκα φορές. Έχουμε 1 παράσιτο ψάρι από 4 σε τρεις περιπτώσεις, και 0 παράσιτα ψάρια για τις υπόλοιπες επτά περιπτώσεις. Εμείς ποτέ δεν πιάσαμε 2 ή περισσότερα παράσιτα ψάρια σε οποιαδήποτε από αυτά τα δείγματα των 4.

Η γεωμετρική κατανομή

Ας υποθέσουμε ότι μια σειρά από ανεξάρτητες διαδρομές Bernoulli με πιθανότητα p πραγματοποιούνται κατά περιόδους 1, 2, 3,…. Τώρα ας W είναι ο χρόνος αναμονής μέχρι να συμβεί η πρώτη επιτυχία. Έτσι

$$P(W>x)=(1-p)^x,$$

πράγμα που σημαίνει ότι

$$P(W =x)=P(W>x-1)-P(W>x)$$

Η συνάρτηση πυκνότητας, ως εκ τούτου, είναι

$$f(x)=p(1-p)^{(x-1)}$$

```
fx<-dgeom(0:20,0.2)
barplot(fx,names=as.character(0:20),xlab="x",ylab="f(x)")
```

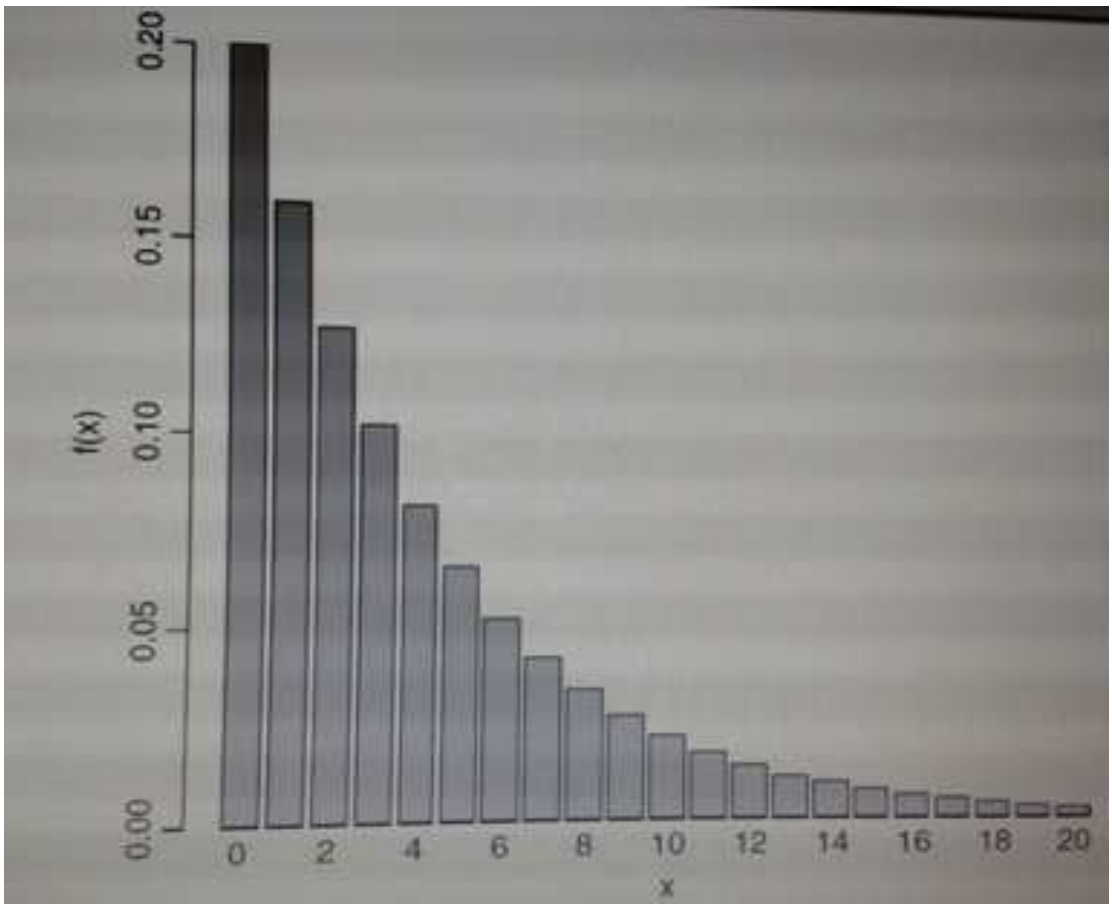
Για τη γεωμετρική κατανομή,

• η μέση τιμή είναι (1-p):p,

• το τετράγωνο τυπικής απόκλισης είναι (1-p):(p^2)

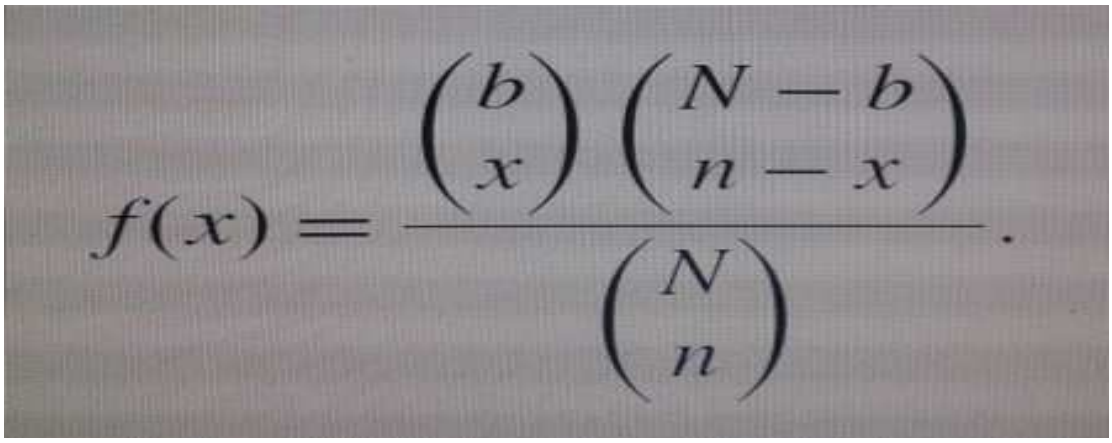
Η γεωμετρική έχει μια πολύ μεγάλη ουρά. Εδώ είναι 100 τυχαίοι αριθμοί από μια γεωμετρική κατανομή με p = 0,1: Ο τρόπος λειτουργίας είναι 0, αλλά απομακρυσμένες τιμές τόσο μεγάλες όσο 43 και 51 έχουν δημιουργηθεί.

```
table(rgeom(100,0.1)
 0  1  2  3  4  5  7  8  9 10 11 12 13 14 15 16 17 18 19 21
13  8  9  6  4 12  5  4  3  4  3  6  1  1  3  1  1  3  1  1
22 23 25 26 30 43 51
 1  2  1  3  2  1  1
```



Η υπεργεωμετρική κατανομή

«Μπάλες σε δοχεία» είναι το κλασικό είδος του προβλήματος επιλύεται με αυτή την κατανομή. Η συνάρτηση πυκνότητας της υπεργεωμετρικής είναι



$$f(x) = \frac{\binom{b}{x}\binom{N-b}{n-x}}{\binom{N}{n}}.$$

Ας υποθέσουμε ότι υπάρχουν N χρωματιστές μπάλες στην περίφημη στατιστική της στάμνας: b από αυτές είναι μπλε και r = N - b από αυτές είναι κόκκινες. Τώρα, ένα δείγμα από n μπάλες έχει αφαιρεθεί από τη στάμνα? Αυτό είναι δειγματοληψία χωρίς αντικατάσταση. Τώρα η f(x) δίνει την πιθανότητα ότι x αυτές τις n μπάλες είναι μπλε.

Οι ενσωματωμένες συναρτήσεις για την υπεργεωμετρική χρησιμοποιούνται σαν αυτό:
dhyper(q, b,r,n),
rhyper(m, b,r,n).

Εδώ

• q είναι ένα διάνυσμα των τιμών μιας τυχαίας μεταβλητής που αντιπροσωπεύει τον αριθμό από μπλε μπάλες από ένα δείγμα μεγέθους n προέρχεται από μια στάμνα που περιέχει b μπλε μπάλες και r αυτές κόκκινου χρώματος.

• b είναι ο αριθμός από μπλε μπάλες στο δοχείο. Αυτό θα μπορούσε να είναι ένα διάνυσμα με μη αρνητικά ακέραια στοιχεία

• r είναι ο αριθμός από κόκκινες μπάλες στη στάμνα = N - b. Αυτό θα μπορούσε επίσης να είναι ένα διάνυσμα με μη αρνητικά ακέραια στοιχεία

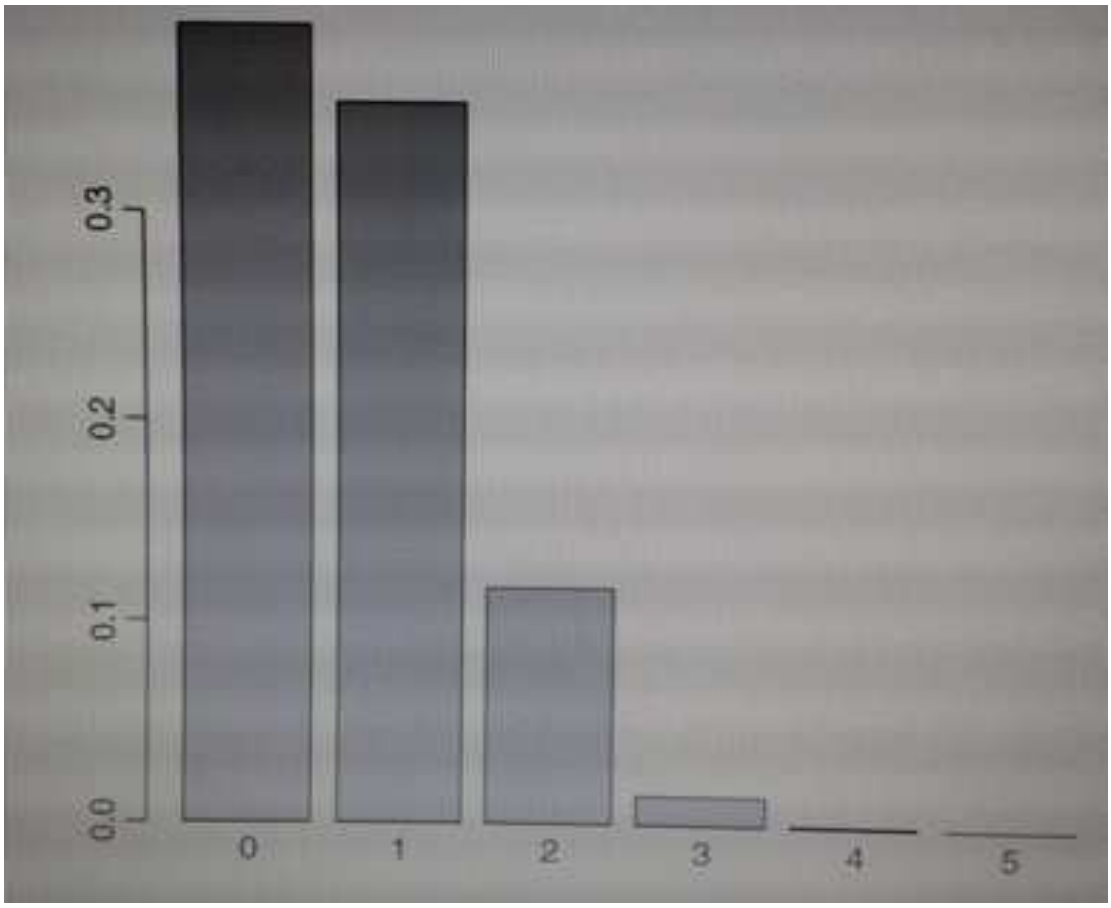
• n είναι ο αριθμός των σφαιρών που έλκονται από μια στάμνα με b μπλε και r κόκκινες μπάλες. Αυτό μπορεί να είναι ένα διάνυσμα όπως b και r.

• p είναι ένα διάνυσμα από πιθανότητες με τιμές μεταξύ 0 και 1.

• m είναι ο αριθμός υπεργεωμετρικά κατανεμημένων τυχαίων αριθμών που πρόκειται να παραχθούν.

Αφήστε τη στάμνα να περιέχει N = 20 μπάλες εκ των οποίων 6 είναι μπλε και 14 είναι κόκκινες. Παίρνουμε ένα δείγμα n = 5 μπάλες έτσι x θα μπορούσε να είναι 0, 1, 2, 3, 4 ή 5 από αυτές τις μπλε, αλλά επειδή η ποσοστιαία αναλογία των μπλε είναι μόνο 6/20 οι υψηλότερες συχνότητες είναι πολύ απίθανες. Το παράδειγμά μας αξιολογείται ως εξής:

```
ph<-numeric(6)
for(i in 0:5) ph[i]<-dhyper(i,6,14,5)
barplot(ph,names=as.character(0:5))
```



Είμαστε πολύ απιθανό να πάρουμε περισσότερες από 2 κόκκινες μπάλες από 5. Το πιο πιθανό αποτέλεσμα είναι ότι έχουμε πάρει 0 ή 1 κόκκινη μπάλα από 5. Μπορούμε να προσομοιώσουμε ένα σύνολο Monte Carlo δοκιμές από 5 μεγέθη. Εδώ είναι οι αριθμοί από κόκκινες μπάλες που λαμβάνονται σε 20 επιτεύγματα του παραδείγματός μας

```
rhyper(20,6,14,5)
[1] 1 1 1 2 1 2 0 1 3 2 3 0 2 0 1 1 2 1 1 2
```

Η διωνυμική κατανομή είναι μια οριακή περίπτωση της υπεργεωμετρικής που προκύπτει ως N, b και r άπειρη προσέγγιση με τέτοιο τρόπο ώστε b / N προσεγγίσεις p, και r / N προσεγγίσεις 1 - p (βλέπε σελ. 242.). Αυτό είναι επειδή, όπως οι αριθμοί λαμβάνονται τεράστιοι, το γεγονός ότι είμαστε δειγματοληψία χωρίς αντικατάσταση καθίσταται άνευ σημασίας. Η διωνυμική κατανομή προϋποθέτει δειγματοληψία με αντικατάσταση από ένα πεπερασμένο πληθυσμό, ή δειγματοληψία χωρίς αντικατάσταση από έναν

άπειρο πληθυσμό. ή δειγματοληψία χωρίς αντικατάσταση από ένα άπειρο πληθυσμό.

Η πολυωνυμική κατανομή

Ας υποθέσουμε ότι υπάρχουν t πιθανές εκβάσεις από μια πειραματική δοκιμή, και το αποτέλεσμα i έχει πιθανότητα p_i. Τώρα, επιτρέψτε n ανεξάρτητες δοκιμές όπου $n=n_1+n_2+\ldots+n_t$ και να ρωτήσω ποια είναι η πιθανότητα της απόκτησης του διανύσματος των περιστατικών Ni του i-οστού αποτελέσματος:

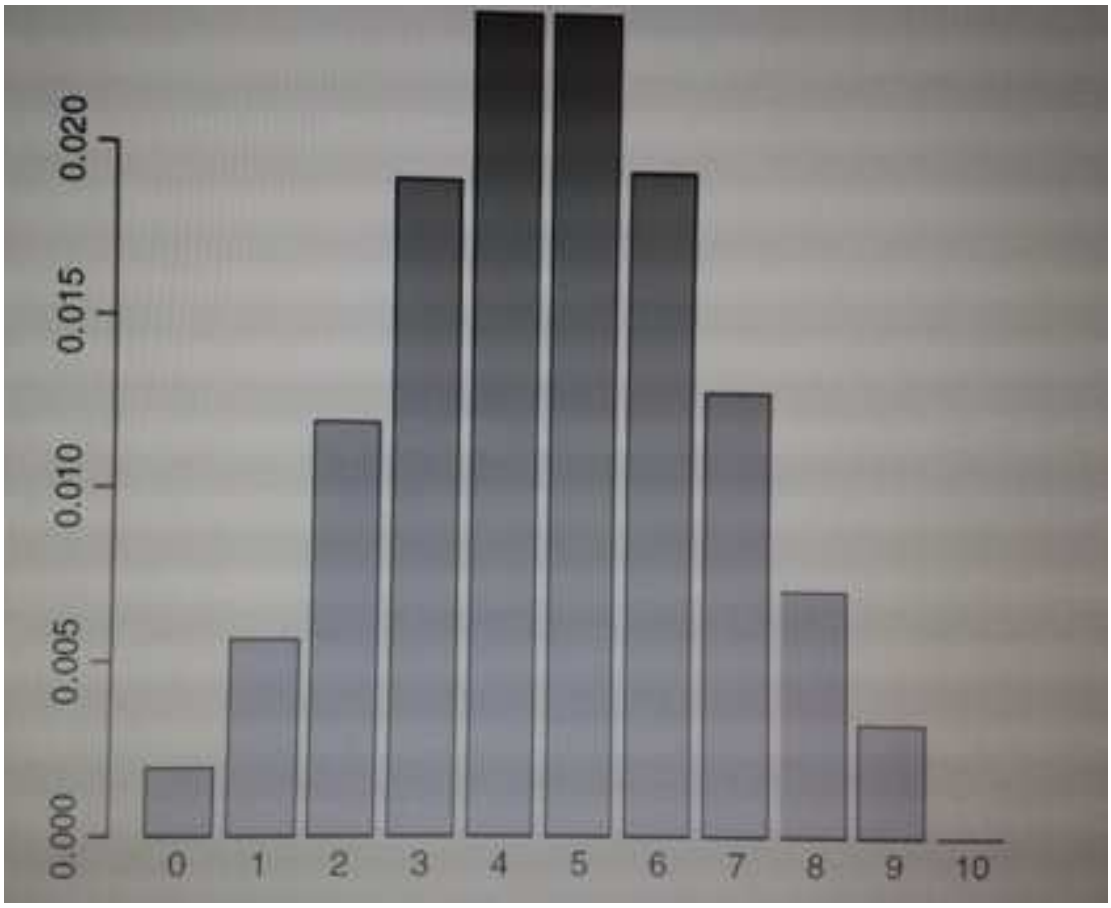
$$P(N_i = n_i) = \frac{n!}{n_1!n_2!n_3!\ldots n_t!} p_1^{n_1} p_2^{n_2} p_3^{n_3} \ldots p_t^{n_t},$$

όπου i πηγαίνει από 1 έως t. Πάρτε ένα παράδειγμα με τρία αποτελέσματα, όπου το πρώτο αποτέλεσμα είναι διπλάσιες πιθανότητες από τα άλλα δύο ($p_1=0,5$, $p_2=0,25$ και $p_3=0,25$). Κάνουμε 4 δοκιμές με $n_1=6$, $n_2=5$, $n_3=7$ και $n_4=6$, οπότε $n=24$. Θα πρέπει να αξιολογήσουμε τον τύπο για i = 1, 2 και 3 (επειδή υπάρχουν τρεις πιθανές εκβάσεις). Είναι λογικό να ξεκινήσουμε γράφοντας μια συνάρτηση που ονομάζεται multi να πραγματοποιήσει τους υπολογισμούς για κάθε αριθμό των επιτυχιών σε α, b και c για τα τρία αποτελέσματα δίνονται τρεις πιθανότητες μας 0.5, 0.25 και 0.25:

```
multi<-function(a,b,c) {
      factorial(a+b+c)/(factorial(a)*factorial(b)*factorial(c))*.5^a*.25^b*.25^c}
```

Τώρα βάλτε τη συνάρτηση σε ένα βρόχο για να λειτουργήσει η πιθανότητα του να πάρει τα απαιτούμενα πρότυπα της επιτυχίας, psuc, για τα τρία αποτελέσματα. Εμείς απεικονίζουμε μόνο μία περίπτωση, στην οποία το τρίτο αποτέλεσμα καθορίζεται σε τέσσερις επιτυχίες. Αυτό σημαίνει ότι η πρώτη και δεύτερη περιπτώσεις διαφέρουν βηματικά μεταξύ 19 και 1 και 9 και 11, αντιστοίχως:

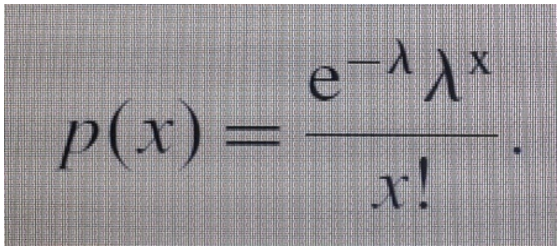
```
psuc<-numeric(11)
for (i in 0:10) psuc[i]<-multi(19-i,1+i,4)
barplot(psuc,names=as.character(0:10))
```



Το πιο πιθανό αποτέλεσμα είναι ότι θα πάρει 19-4 = 15 ή 19-5 = 14 επιτυχίες του τύπου 1 σε μια δοκιμή μεγέθους 24 με πιθανότητες 0.5, 0.25 και 0.25, όταν ο αριθμός των επιτυχιών της τρίτης περίπτωσης ήταν 4 από 24. Μπορείτε εύκολα να τροποποιήσετε τη συνάρτηση για να ασχοληθεί με άλλες πιθανότητες και άλλο αριθμό από αποτελέσματα.

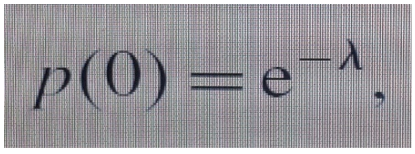
Η κατανομή Poisson

Αυτή είναι μια από τις πιο χρήσιμες και σημαντικές από τις διακριτές κατανομές πιθανοτήτων για την περιγραφή μέτρησης των δεδομένων. Ξέρουμε πόσες φορές συνέβη κάτι (π.χ. κλωτσιές από τα άλογα του ιππικού, ελάφρυνση από απεργίες, χτυπήματα από βόμβα), αλλά δεν έχουμε τρόπο να γνωρίζουμε πόσες φορές αυτό δεν συνέβη. Η Poisson είναι μιας-παραμέτρου κατανομή με την ενδιαφέρουσα ιδιότητα ότι το τετράγωνο τυπικής της απόκλισης είναι ίσο με το μέσο όρο της. Πάρα πολλές διαδικασίες παρουσιάζουν το τετράγωνο τυπικής απόκλισης που αυξάνεται με τη μέση τιμή, συχνά ταχύτερα από γραμμικά (βλ. την αρνητική διωνυμική κατανομή κατωτέρω). Η συνάρτηση πυκνότητας του Poisson δείχνει την πιθανότητα απόκτησης μιας καταμέτρησης του x, όταν η μέση μέτρηση ανά μονάδα είναι λ:



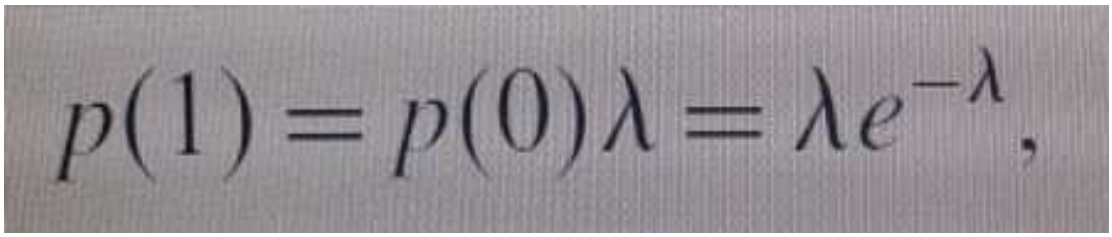
$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}.$$

Η μηδενική διάρκεια του Poisson (η πιθανότητα απόκτησης μιας καταμέτρησης από μηδέν) λαμβάνεται θέτοντας x = 0:



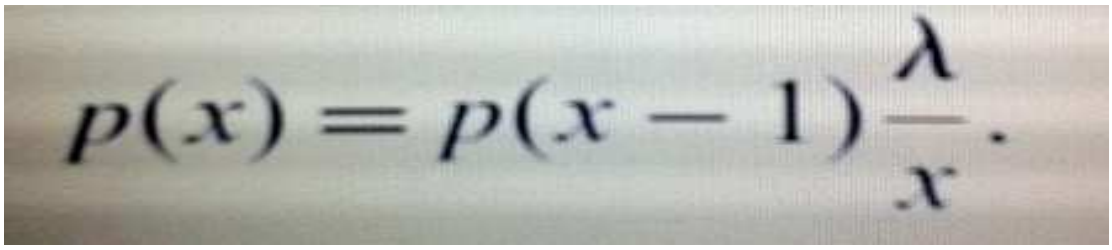
$$p(0) = e^{-\lambda},$$

το οποίο είναι απλά ο αντιλογάριθμος μείον τη μέση τιμή. Δεδομένου p(0), είναι σαφές ότι p(1) είναι ακριβώς



$$p(1) = p(0)\lambda = \lambda e^{-\lambda},$$

και οποιαδήποτε μεταγενέστερη πιθανότητα λαμβάνεται εύκολα από τον πολλαπλασιασμό της προηγούμενης πιθανότητας από τη μέση τιμή και διαιρώντας με την καταμέτρηση, έτσι:



$$p(x) = p(x-1)\frac{\lambda}{x}.$$

Οι συναρτήσεις για την πυκνότητα, αθροιστικής κατανομής, ποσοτικοποιήσεις και γεννήτρια τυχαίων αριθμών της κατανομής Poisson που λαμβάνεται με

dpois(x, lambda)
ppois(q, lambda)
qpois(p, lambda)
rpois(n, lambda)

όπου λάμδα είναι ο μέσος αριθμός ανά δείγμα.
Η κατανομή Poisson κατέχει κεντρική θέση σε τρεις τελείως διαφορετικές περιοχές των στατιστικών:

• στην περιγραφή των τυχαίων χωρικών προτύπων ανά σημείο (βλ. σελ. 749).?
• καθώς τη συχνότητα κατανομής των αριθμών των σπάνιων αλλά ανεξάρτητα αποτελέσματα (βλ. σελ. 208).?
• καθώς τη κατανομή του σφάλματος στο GLMS για τον αριθμό δεδομένων (βλ. σελ.. 527).
Αν θέλαμε 600 προσομοίωσεις μετρήσεων από μια κατανομή Poisson με μέση τιμή, ας πούμε, 0.90 αιμοσφαίρια ανά διαφάνεια, απλά πληκτρολογήστε:
count<-rpois(600,0.9)
Μπορούμε να χρησιμοποιήσουμε πίνακα για να δείτε τις συχνότητες που παράγονται σε κάθε μέτρηση:

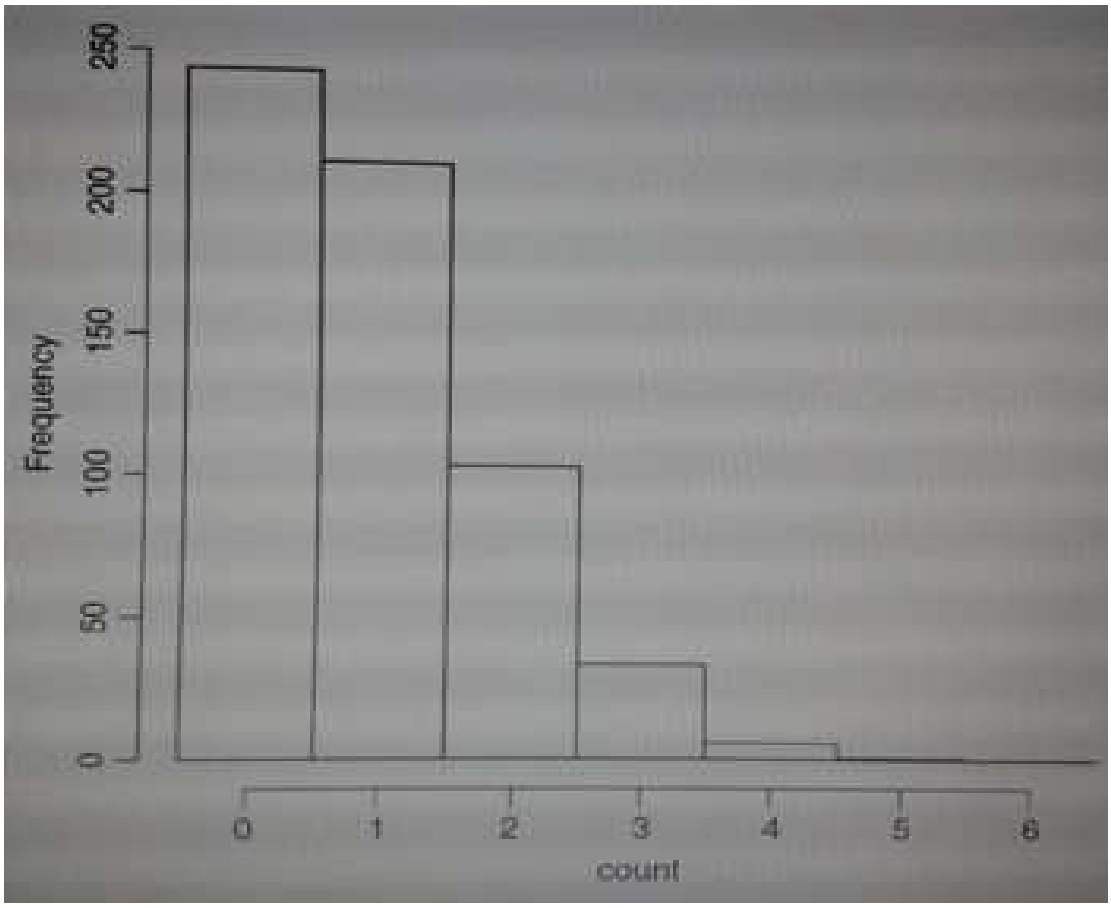
table(count)

καταμέτρηση

```
  0   1   2  3 4 5
244 212 104 33 6 1
```

ή 'ιστορικό' για να δείτε ένα ιστόγραμμα των μετρήσεων:

hist(count,breaks = - 0.5:6.5,main="")



Σημειώστε τη χρήση του διανύσματος των σημείων υποδιαίρεσης στις ακέραιες προσαυξήσεις από -0, 5 για τη δημιουργία ακέραιων κουτιών αποθήκευσης για τις ράβδους ιστογράμματος.

Η αρνητική διωνυμική κατανομή

Αυτή η διακριτή, δύο παραμέτρων κατανομή είναι χρήσιμη για την περιγραφή της κατανομής των δεδομένων μέτρησης, όπου το τετράγωνο τυπικής απόκλισης είναι συχνά πολύ μεγαλύτερο από τη μέση τιμή. Οι δύο παράμετροι είναι η μέση τιμή μ και η συσσώρευση k παράμετρος, που δίνεται από

$$k = \frac{\mu^2}{\sigma^2 - \mu}$$

Όσο μικρότερη είναι η τιμή του k, τόσο μεγαλύτερος είναι ο βαθμός της συσσωμάτωσης. Η συνάρτηση πυκνότητας είναι

$$p(x) = \left(1 + \frac{\mu}{k}\right)^{-k} \frac{(k+x-1)!}{x!(k-1)!} \left(\frac{\mu}{\mu+k}\right)^{x}.$$

Ο μηδενικός όρος βρίσκεται θέτοντας x = 0 και απλούστευση:

$$p(0) = \left(1 + \frac{\mu}{k}\right)^{-k}$$

Διαδοχικοί όροι στην κατανομή μπορούν στη συνέχεια να υπολογιστούν επαναληπτικά από

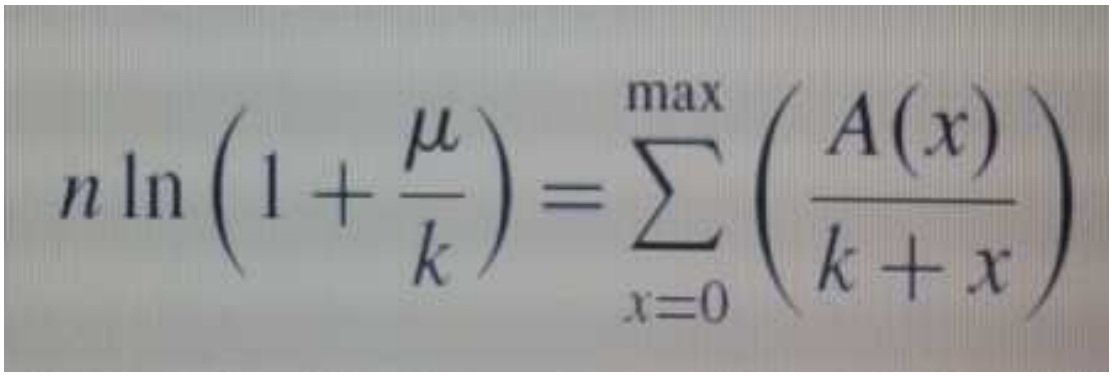
$$p(x) = p(x-1)\left(\frac{k+x-1}{x}\right)\left(\frac{\mu}{\mu+k}\right).$$

Μια πρώτη εκτίμηση της τιμής του k μπορεί να ληφθεί από τη μέση τιμή του δείγματος και το τετράγωνο τυπικής απόκλισης

$$k \approx \frac{\bar{x}^2}{s^2 - \bar{x}}.$$

Επειδή k δεν μπορεί να είναι αρνητική, είναι σαφές ότι στην αρνητική διωνυμική κατανομή δεν θα πρέπει να τοποθετούνται δεδομένα, όπου το τετράγωνο τυπικής απόκλισης είναι μικρότερο από τη μέση τιμή.

Η εκτίμηση μέγιστης πιθανότητας k βρίσκεται αριθμητικά, με την επανάληψη προοδευτικά όλο και πιο τελειοποιημένων τιμών του k μέχρι η αριστερή και η δεξιά πλευρά την ακόλουθη εξίσωση είναι ίσες:



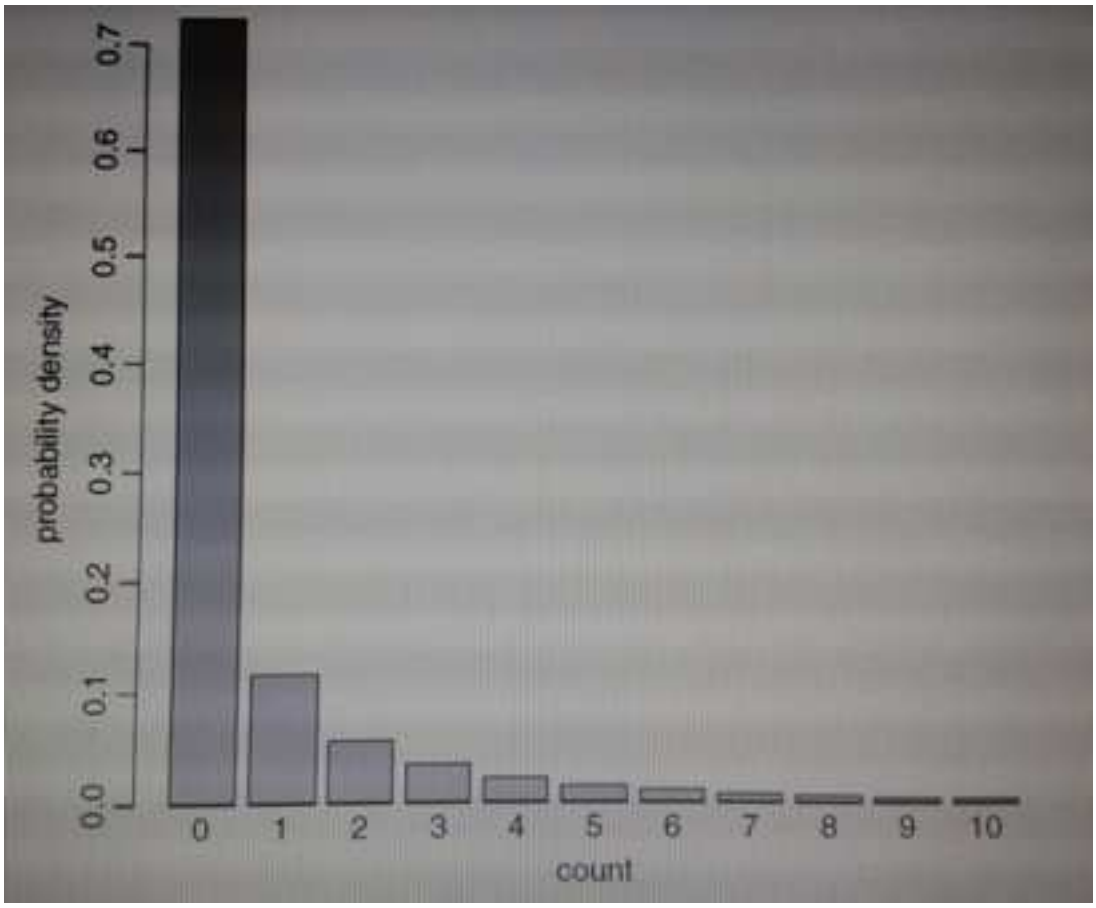
$$n \ln\left(1+\frac{\mu}{k}\right)=\sum_{x=0}^{\max}\left(\frac{A(x)}{k+x}\right)$$

όπου το διάνυσμα A(χ) περιέχει το σύνολο συχνότητων με τιμές μεγαλύτερες από x. Θα μπορούσατε να γράψετε μια συνάρτηση για να επεξεργαστεί τις πυκνότητες πιθανότητας όπως αυτό:

```
negbin<-function(x,u,k) (1+u/k)^(-
k)*(u/(u+k))^x*gamma(k+x)/(factorial(x)*gamma(k))
```

να χρησιμοποιήσετε τη συνάρτηση για την παραγωγή ενός ραβδογραφήματος με πυκνότητες πιθανότητας για μια σειρά από τιμές x (δηλαδή 0-10, για μια κατανομή που καθορίζεται με μέση τιμή και η παράμετρος συγκέντρωσης (ας πούμε μ = 0, 8, k = 0, 2), όπως αυτή

```
xf<-sapply(0:10, function(i) negbin(i,0.8,0.2))
barplot(xf,names=as.character(0:10),xlab="count",ylab="probability density")
```



Υπάρχει και ένας άλλος, εντελώς διαφορετικός τρόπος θεώρησης της αρνητικής διωνυμικής κατανομής.
Εδώ, η μεταβλητή απόκρισης είναι ο χρόνος αναμονής Wr για την επιτυχία rth:

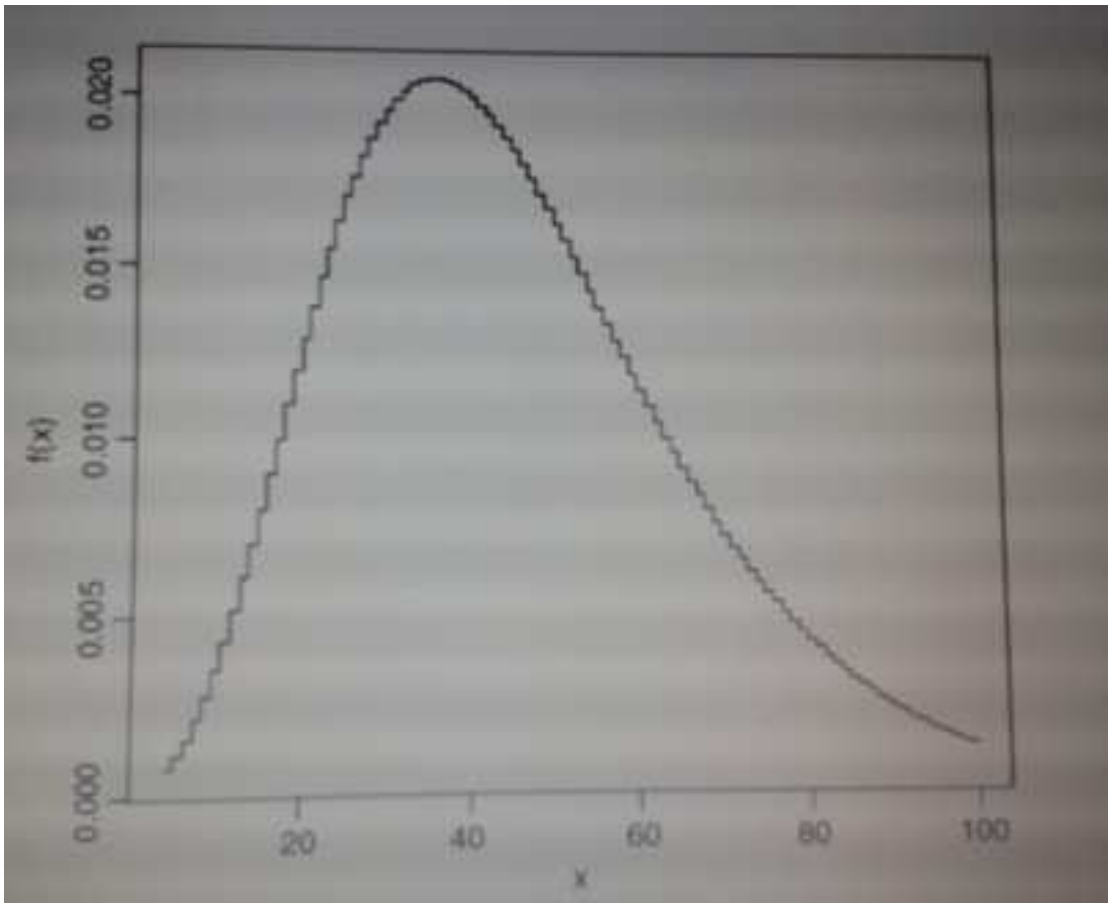
$$f(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$$

Είναι σημαντικό να συνειδητοποιήσουμε ότι το x ξεκινά στο r και αυξάνει από εκεί (προφανώς, η επιτυχία rth δεν μπορεί να συμβεί πριν από την προσπάθεια rth). Η dnbinom συνάρτηση αντιπροσωπεύει τον αριθμό των αποτυχιών x (π.χ. ουρές ρίχνοντας νόμισμα) πριν επιτυγχάνονται επιτυχίες μεγέθους (ή κεφάλια από ρίψη κέρματος), όταν η πιθανότητα επιτυχίας (ή της κεφαλής) είναι η πιθανότητα:

dnbinom(x, size, prob)

Ας υποθέσουμε ότι μας ενδιαφέρει η κατανομή του χρόνου αναμονής μέχρι την 5η επιτυχία που συμβαίνει σε μια αρνητική διωνυμική διαδικασία, με p = 0,1. Ξεκινάμε την ακολουθία των x τιμών σε 5

plot(5:100,dnbinom(5:100,5,0.1),type="s",xlab="x",ylab="f(x)")



Αυτό δείχνει ότι ο πιο πιθανός χρόνος αναμονής για την 5η επιτυχία, όταν η πιθανότητα επιτυχίας είναι 1/10, είναι περίπου 31 δοκιμές μετά την 5η δοκιμή. Σημειώστε ότι η αρνητική διωνυμική κατανομή είναι αρκετά έντονα λοξή προς τα δεξιά.

Είναι εύκολο να δημιουργήσει αρνητικά διωνυμικά δεδομένα χρησιμοποιώντας την γεννήτρια τυχαίων αριθμών:

rnbinom(n, size, prob)

Ο αριθμός των απαιτούμενων τυχαίων αριθμών είναι n. Όταν η δεύτερη παράμετρος, το μέγεθος, έχει οριστεί σε 1 γίνεται γεωμετρική κατανομή (βλέπε παραπάνω). Η τελευταία παράμετρος, prob, είναι η πιθανότητα

επιτυχίας ανά δοκιμή, p. Εδώ έχουμε δημιουργήσει 100 μετρήσεις με ένα μέσο όρο 0,6:

```
count<-rnbinom(100,1,0.6)
```

Μπορούμε να χρησιμοποιήσουμε πίνακα για να δείτε τη συχνότητα των διαφορετικών μετρήσεων:

```
table(count)
```

```
 0  1  2  3  5  6
65 18 13  2  1  1
```

Είναι λογικό να βεβαιωθείτε ότι η μέση τιμή είναι πραγματικά 0,6 (ή πολύ κοντά σε αυτή)

```
mean(count)
```

```
[1] 0.61
```

Το τετράγωνο τυπικής απόκλισης θα είναι ουσιαστικά μεγαλύτερο από τη μέση τιμή

```
var(count)
```

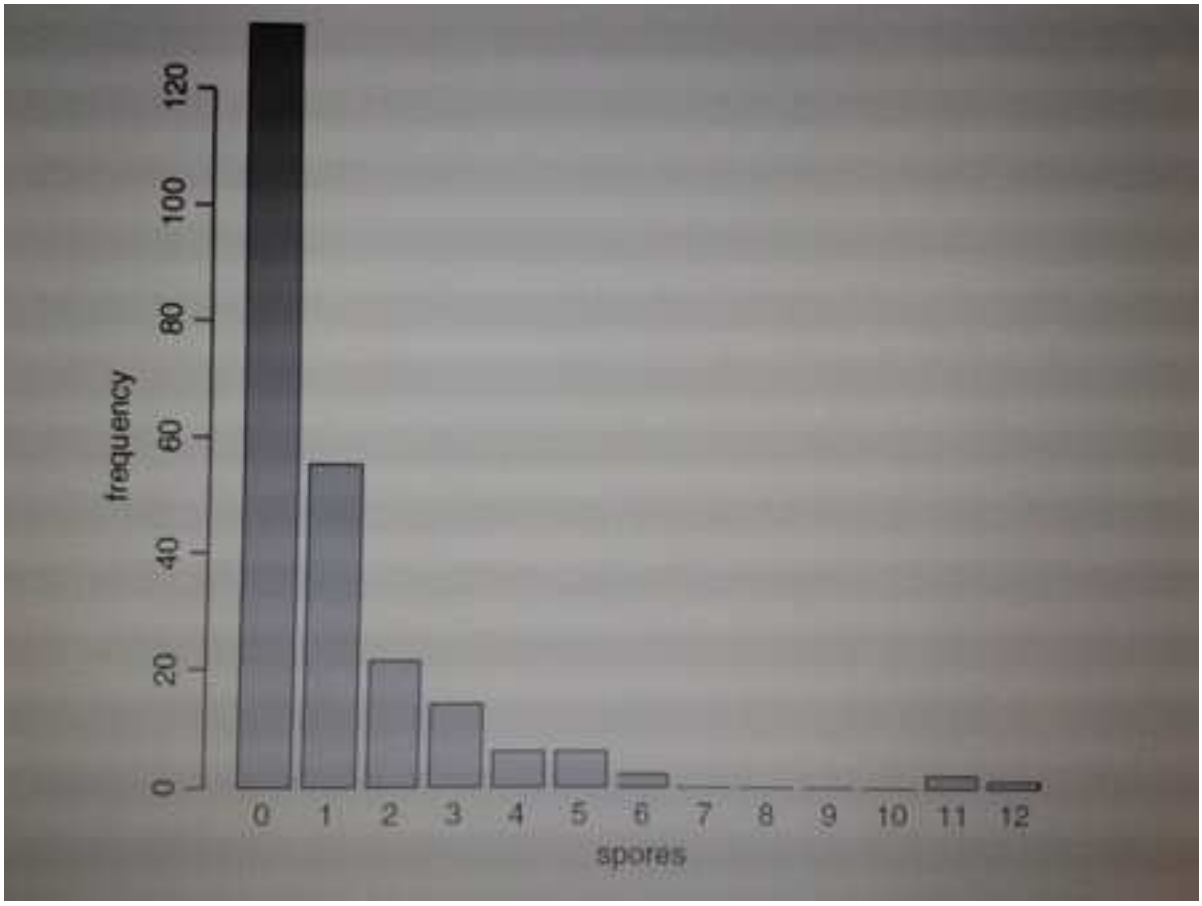
```
[1] 1.129192
```

και αυτό δίνει μια εκτίμηση του κ

0,61^2/(1,129-0,61)= 0,717.

Τα παρακάτω δεδομένα δείχνουν τον αριθμό των σπορίων που υπολογίζονται σε 238 θαμμένες γυάλινες διαφάνειες. Ενδιαφερόμαστε για το αν τα δεδομένα αυτά περιγράφονται καλά από μια αρνητική διωνυμική κατανομή. Αν είναι θα θέλαμε να βρούμε τη μέγιστη εκτίμηση της πιθανότητας συνάθροισης της παραμέτρου k.

```
x<-0:12
freq<-c(131,55,21,14,6,6,2,0,0,0,0,2,1)
barplot(freq,names=as.character(x),ylab="frequency",xlab="spores")
```



Θα ξεκινήσουμε εξετάζοντας το τετράγωνο τυπικής απόκλισης - μέση αναλογία των μετρήσεων. Εμείς δεν μπορούμε να χρησιμοποιήσουμε μέση τιμή και το τετράγωνο τυπικής απόκλισης άμεσα, επειδή τα δεδομένα μας είναι συχνότητες μετρήσεων, μάλλον μετρούν τους εαυτούς τους.
Αυτό είναι εύκολο να διορθωθεί: χρησιμοποιούμε rep να δημιουργήσει ένα διάνυσμα y το οποίο μετράει κάθε μέτρηση (x) επαναλαμβάνοντας το σχετικό αριθμό των φορών (freq). Τώρα μπορούμε να χρησιμοποιήσουμε μέση τιμή και το τετράγωνο τυπικής απόκλισης άμεσα:

```
y<-rep(x,freq)
mean(y)
```

[1] 1.004202

```
var(y)
```

[1] 3.075932

Αυτό δείχνει ότι τα δεδομένα συγκεντρώνονται σε μεγάλο βαθμό (η αναλογία του μέσου τετραγώνου τυπικής απόκλισης είναι περίπου 3, υπενθυμίζοντας ότι θα είναι 1 αν τα δεδομένα ήταν κατανομή Poisson). Πρόχειρη εκτίμηση μας k είναι συνεπώς

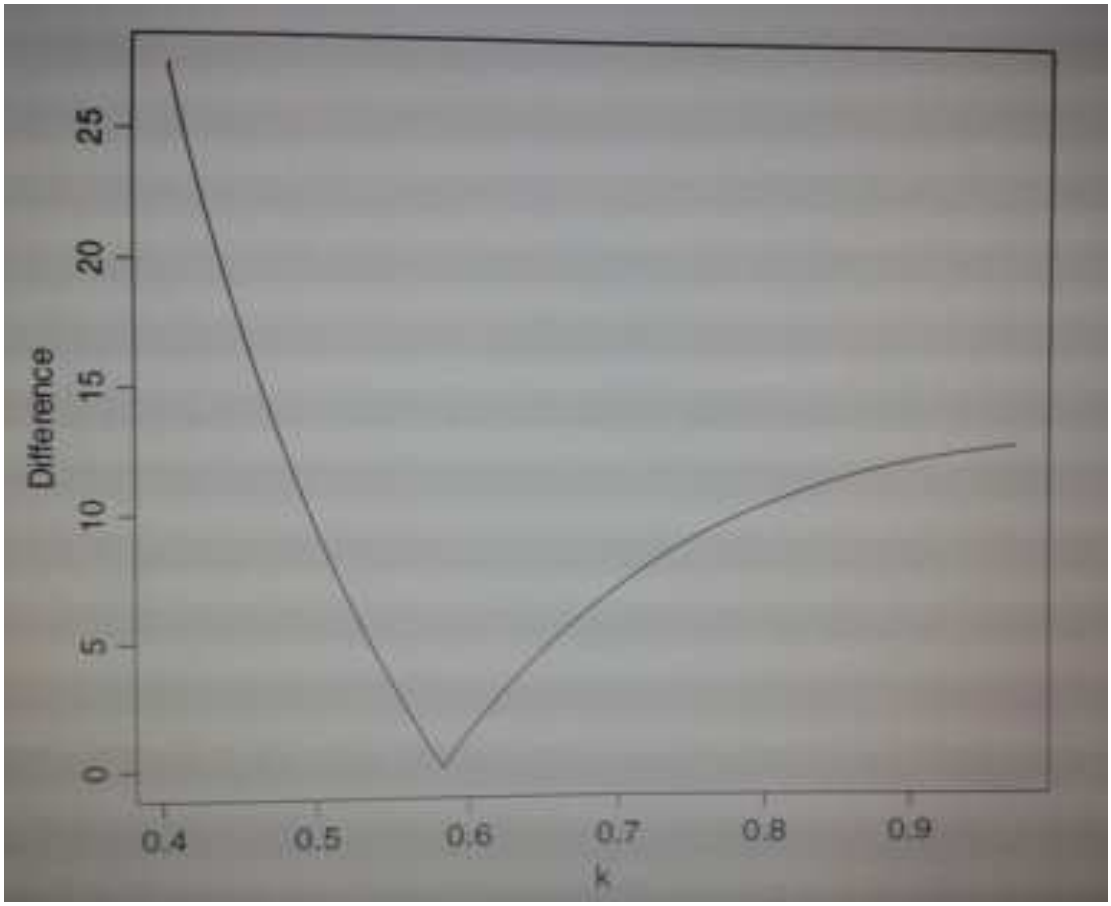
```
mean(y)^2/(var(y)-mean(y))

[1] 0.4867531
```

Εδώ είναι μια συνάρτηση που λαμβάνει ένα διάνυσμα των συχνοτήτων από μετρήσεις x(μεταξύ 0 και το μήκος length(x) - 1) και υπολογίζει την εκτίμηση μέγιστης πιθανότητας της παραμέτρου συσσωμάτωσης, k:

```
kfit <-function(x)
{
lhs<-numeric()
rhs<-numeric()
y <-0:(length(x) - 1)
j<-0:(length(x)-2)
m <-sum(x * y)/(sum(x))
s2 <-(sum(x * y^2) - sum(x * y)^2/sum(x))/(sum(x)- 1)
k1 <-m^2/(s2 - m)
a<-numeric(length(x)-1)
for(i in 1:(length(x) - 1)) a[i] <-sum(x [- c(1:i)])
i<-0
for (k in seq(k1/1.2,2*k1,0.001)) {
i<-i+1
lhs[i] <-sum(x) * log(1 + m/k)
rhs[i] <-sum(a/(k + j))
}
k<-seq(k1/1.2,2*k1,0.001)
plot(k, abs(lhs-rhs),xlab="k",ylab="Difference",type="l")
d<-min(abs(lhs-rhs))
sdd<-which(abs(lhs-rhs)==d)
k[sdd]
}
```

Μπορούμε να την δοκιμάσουμε με δεδομένα για το πλήθος σπορίων μας.

```
kfit(freq)
[1]      0.5826276
```



Η ελάχιστη διαφορά είναι κοντά στο μηδέν και εμφανίζεται σε περίπου k = 0,55. Η εκτύπωση δείχνει ότι η εκτίμηση μέγιστης πιθανότητας k είναι 0.582 (στα 3 δεκαδικά ψηφία που προσομοιώνουμε; Τα τελευταία 4 δεκαδικά ψηφία (6276) είναι χωρίς νόημα και δεν θα εκτυπωθεί σε μια πιο τελειοποιημένη συνάρτηση).
Πώς θα περιέγραφε τα δεδομένα της μέτρησης μας μια αρνητική διωνυμική κατανομή με μέση τιμή 1,0042 και μια τιμή k των 0.583; Οι αναμενόμενες συχνότητες λαμβάνοται πολλαπλασιάζοντας την πυκνότητα πιθανότητας (ανωτέρω) δια του συνολικού μεγέθους του δείγματος (238 ολισθαίνει στην περίπτωση αυτή).

```
nb<-238*(1+1.0042/0.582)^(-0.582)*factorial(.582+(0:12)-1)/
        (factorial(0:12)*factorial(0.582-1))*(1.0042/(1.0042+0.582))^(0:12)
```

Θα πρέπει να συγκρίνουμε τις παρατηρούμενες και αναμενόμενες συχνότητες με τη χρήση ραβδογραφήματος. Πρέπει να εναλλάσσονται οι παρατηρούμενες και αναμενόμενες συχνότητες. Υπάρχουν τρία στάδια στη διαδικασία:

- Ενώσετε τις παρατηρούμενες και αναμενόμενες συχνότητες σε μια εναλλασσόμενη ακολουθία.
- Δημιουργία λίστας των ετικετών για να αναφέρουμε τις ράβδους (εναλλασσόμενα κενά και μετρήσεις).

• Δημιουργήστε ένα μύθο για να περιγράψει τα διαφορετικά χρώματα ράβδου (στο γράφημα).

Ο συνεχόμενος κατάλογος των συχνοτήτων (που ονομάζονται και οι δύο) γίνεται σαν αυτό, βάζοντας τις 13 παρατηρούμενες μετρήσεις (freq) στις μονές μπάρες και τις 13 αναμενόμενες μετρήσεις (nb) στις οριζόντιες αριθμημένες ράβδους (σημειώστε τη χρήση του %% modulo να κάνει αυτό):

```
both<-numeric(26)
both[1:26 %% 2 != 0]<-freq
both[1:26 %% 2 == 0]<-nb
```

Τώρα μπορούμε να σχεδιάσουμε το συνδυασμένο ραβδογράφημα:

```
barplot(both,col=rep(c(1,0),13),ylab="frequency")
```

Επειδή δύο γειτονικές ράβδοι αναφέρονται στην ίδια μέτρηση (οι παρατηρούμενες και αναμενόμενες συχνότητες) δεν θέλουμε να χρησιμοποιούν ενσωματωμένα στην barplot ονόματα ορίσματος για την επισήμανση των ράβδων (που θα ήθελε να γράψει μια ετικέτα σε κάθε ράβδο, 26 ετικέτες σε όλες). Αντ 'αυτού, θέλουμε να γράψουμε την καταμέτρηση μόνο μία φορά για κάθε ζεύγος ράβδων, που βρίσκεται ανάμεσα στις παρατηρούμενες και αναμενόμενες ραύδους, χρησιμοποιώντας as.character(0:12). Αυτό είναι μια δουλειά για τη συνάρτηση mtext, που γράφει το κείμενο στο περιθώριο του σχεδίου. Πρέπει να καθορίσουμε το περιθώριο μέσα στο οποίο θέλουμε να γράψουμε. Το κάτω περιθώριο στην πλευρά = 1. Το μόνο ελαφρώς δύσκολο πράγμα είναι να επεξεργαστεί τις συντεταγμένες x για τις 13 ετικέτες κατά μήκος του άξονα. Για να δείτε πώς ο άξονας x έχει κλιμακωθεί από το ραβδόγραφημα της συνάρτησης, να κατανείμει το ραβδόγραφημα της συνάρτησης (όπως παραπάνω) σε ένα όνομα δυανύσματος, στη συνέχεια, ελέγξτε το περιεχόμενό της:

```
xscale<-barplot(both,col=rep(c(1,0),13),ylab="frequency")
as.vector(xscale)
 [1]  0.7  1.9  3.1  4.3  5.5  6.7  7.9  9.1 10.3 11.5 12.7 13.9 15.1
[14] 16.3 17.5 18.7 19.9 21.1 22.3 23.5 24.7 25.9 27.1 28.3 29.5 30.7
```

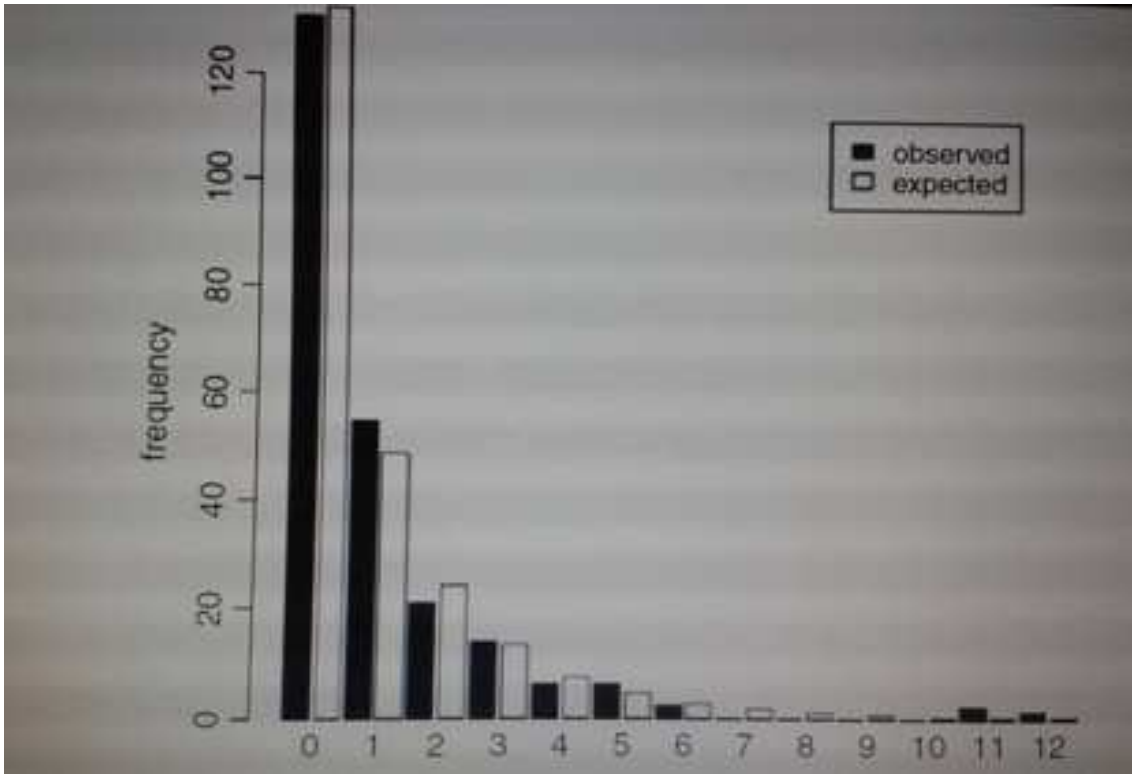
Εδώ μπορείτε να δείτε ότι η αριστερή πλευρά της πρώτης ράβδου είναι στο x = 0,7 και την 26η ράβδο είναι στη θέση x = 30,7. Λίγος πειραματισμός θα δείξει ότι θέλουμε να σβήσει η πρώτη ετικέτα στο x = 1, 4 και στη συνέχεια σε διαστήματα 2,4 (δύο ράβδοι διαχωρίζονται κοντά, για παράδειγμα, 11,5-9,1 = 2,4). Θα καθορίσουμε την ακολουθία seq(1.4,30.2,2.4) ως όρισμα σε ποσοστό εντός mtext:

```
mtext(as.character(0:12),side=1,at=seq(1.4,30.2,2.4))
```

Η προεπιλογή που χρησιμοποιείται εδώ είναι για mtext να γράψει τις ετικέτες σε αριθμό γραμμής 0 μακριά από την πλευρά της στην ερώτηση: αν θέλετε να το αλλάξετε αυτό, προσθέστε το όρισμα γραμμή= 1 έως mtext.

Η συνάρτηση legend δημιουργεί ένα μύθο για να δείξει ποιοί ράβδοι αντιπροσωπεύουν τις παρατηρούμενες συχνότητες (μαύρο στην προκειμένη περίπτωση) και οι οποίες αντιπροσωπεύουν τις αναμενόμενες, αρνητικές διωνυμικές συχνότητες (ανοικτές ράβδοι). Απλά κάντε κλικ στο κουμπί όταν ο δρομέας βρίσκεται στη θέση όπου θέλετε την επάνω αριστερή γωνία του πλαισίου legend να είναι:

legend(locator(1),c("observed","expected"),fill=c("black","white"))



Η εφαρμογή είναι πολύ κοντά, οπότε μπορούμε να είμαστε μετριοπαθώς σίγουροι για την περιγραφή των παρατηρούμενων μετρήσεων ως αρνητική διωνυμική κατανομή. Η ουρά της παρατηρούμενης κατανομής είναι μάλλον πιο παχιά από την αναμενόμενη αρνητική διωνυμική ουρά, οπότε ίσως να θέλουμε να μετρήσουμε την έλλειψη εναρμόνισης μεταξύ παρατηρημένων και αναμενόμενων κατανομών. Ένας απλός τρόπος να γίνει αυτό είναι να χρησιμοποιήσετε στατιστική x-τετράγωνου του Pearson προσέχοντας να χρησιμοποιούν μόνο τις περιπτώσεις όπου η αναμενόμενη συχνότητα nb είναι μεγαλύτερη από 5:

sum(((freq-nb)^2/nb)[nb > 5])

[1] 1.634975

Αυτό βασίζεται σε πέντε νόμιμες συγκρίσεις

```
sum(nb>5)

[1]  5
```

και ως εκ τούτου σε 5-ρ-1 = 2 d.f. γιατί έχουμε υπολογίσει p = 2 παραμέτρους από τα δεδομένα στην εκτίμηση της αναμενόμενης κατανομής (η μέση τιμή και k της αρνητικής διωνυμικής) και χάνουμε ένα βαθμό ελευθερίας για έκτακτη ανάγκη (ο συνολικός αριθμός των μετρήσεων πρέπει να προσθέσετε έως και 238). Υπολογιζόμενη τιμή της στατιστικής x-τετραγώνου= 1,63 είναι πολύ μικρότερη από την τιμή στους πίνακες

```
qchisq(0.95,2)

[1]  5.991465
```

έτσι δεχόμαστε την υπόθεση ότι τα δεδομένα μας δεν είναι σημαντικά από μια αρνητική διωνυμική με μέση τιμή = 1,0042 και k = 0,582.

Η Wilcoxon rank-sum (βαθμωτού-αθροίσματος) στατιστική

Αυτή η συνάρτηση υπολογίζει την κατανομή της στατιστικής Wilcoxon rank-sum (επίσης γνωστή ως Mann-Whitney), και επιστρέφει τιμές για την ακριβή πιθανότητα σε διακριτές τιμές του q:

```
dwilcox(q, m, n)
```

Εδώ q είναι ένα διάνυσμα ποσοτικοποίησης, m είναι ο αριθμός των παρατηρήσεων στο x δείγμα (ένας θετικός ακέραιος όχι μεγαλύτερος από 50), και το n είναι ο αριθμός των παρατηρήσεων στο δείγμα y (επίσης ένας θετικός ακέραιος όχι μεγαλύτερος από 50). Η Wilcoxon rank-sum στατιστική είναι το άθροισμα των βαθμών του x στο συνδυασμένο δείγμα c (x, y). Η Wilcoxon rank-sum στατιστική παίρνει τιμές W μεταξύ των ορίων:

$$\frac{m(m+1)}{2} \leq W \leq \frac{m(m+2n+1)}{2}.$$

Αυτή η στατιστική μπορεί να χρησιμοποιηθεί για μια μη-παραμετρική δοκιμασία της τοποθεσίας μετατόπισης μεταξύ της δημιουργίας απογόνων των πληθυσμών x και y.

Άλγεβρα Πινάκων

Υπάρχει ένα ολοκληρωμένο σύνολο συναρτήσεων για τη διαχείριση των πινάκων στο R. Θα ξεκινήσουμε με ένα πίνακα που ονομάζεται α που έχει

τρεις γραμμές και δύο στήλες. Τα δεδομένα τυπικά τέθηκαν σε πίνακα στήλη, έτσι ώστε οι τρεις πρώτοι αριθμοί (1, 0, 4) να πάνε στη στήλη 1 και οι δεύτεροι τρεις αριθμοί (2, -1, 1), να πάνε στη στήλη 2:

```
a<-matrix(c(1,0,4,2,-1,1),nrow=3)
a
```

```
     [,1] [,2]
[1,]    1    2
[2,]    0   -1
[3,]    4    1
```

Δεύτερος πινακάς μας, που ονομάζεται b, έχει τον ίδιο αριθμό στηλών όσες ο Α έχει γραμμές (δηλαδή τρεις σε αυτή την περίπτωση). Εισαγωγή στηλών, οι δύο πρώτοι αριθμοί (1, -1) πηγαίνουν στη στήλη 1, οι δεύτεροι δύο αριθμοί (2, 1) πηγαίνουν στη στήλη 2, και οι δύο τελευταίοι αριθμοί (1, 0), πηγαίνουν στη στήλη 3:

```
b<-matrix(c(1,-1,2,1,1,0),nrow=2)
b
```

```
     [,1] [,2] [,3]
[1,]    1    2    1
[2,]   -1    1    0
```

Πολλαπλασιασμός πινάκων

Για να πολλαπλασιάστει ένας πίνακας με ένα άλλο πίνακα πάρετε τις γραμμές του πρώτου πίνακα και τις στήλες του δεύτερου πίνακα. Βάλτε την πρώτη γραμμή από τη πλευρά α δίπλα στην πλευρά με την πρώτη στήλη του b:

```
a[1,]
```

```
[1] 1 2
```

```
b[,1]
```

```
[1]  1 -1
```

και λύνει το πρόβλημα δίνοντας έμφαση στο γινόμενο:

```
a[1,]*b[,1]
```

```
[1] 1 -2
```

στη συνέχεια, προσθέστε τα μερικά γινόμενα

```
sum(a[1,]*b[,1])

[1]    -1
```

Το άθροισμα των μερικών γινομένων είναι -1 και αυτό είναι το πρώτο στοιχείο του γινομένου του πίνακα. Στη συνέχεια, βάλτε την πρώτη γραμμή του α με τη δεύτερη στήλη του b :

```
a[1,]
[1]       1       2
b[,2]
[1]       2       1
a[1,] *b[,2]
[1]       2        2

sum(a[1,]*b[,2])

[1]     4
```

έτσι ώστε τα μερικά γινόμενα είναι 2, 2, και το άθροισμα των μερικών γινομένων είναι 2 +2 = 4. Έτσι, πηγαίνει 4 στη γραμμή 1 και στήλη 2 ως απάντηση. Στη συνέχεια, πάρτε την τελευταία στήλη του b και ταιριάξετε αυτή απέναντι στη πρώτη γραμμή του α:

```
a[1,]*b[,3]

[1]      1     0
sum(a[1,]*b[,3])

[1]     1
```

έτσι ώστε το άθροισμα των μερικών γινομένων είναι 1 +0 = 1. Αυτό πηγαίνει στη γραμμή 1, στήλη 3 ως απάντηση. Και ούτω καθεξής. Επαναλαμβάνουμε αυτά τα βήματα για την γραμμή 2 του πίνακα α(0, -1) και στη συνέχεια και πάλι για την γραμμή 3 του πίνακα α(4, 1) για να ληφθεί ο πλήρες πίνακας της απάντησης. Στην R, το σύμβολο για του πολλαπλασιασμού πίνακα είναι% *%. Εδώ είναι η πλήρης απάντηση:

```
a%*%b

          [,1]       [,2]      [,3]
[1,]       -1          4         1
[2,]        1         -1         0
[3,]        3          9         4
```

όπου μπορείτε να δείτε τις τιμές που υπολογίζονται με το χέρι (-1, 4, 1), στην πρώτη γραμμή.

Είναι σημαντικό να καταλάβουμε ότι με πίνακες α φορές b δεν είναι το ίδιο όπως b φορές α.

Ο πίνακας που προκύπτει από πολλαπλασιασμό έχει τον αριθμό των γραμμών του πίνακα στα αριστερά (α έχει 3 γραμμές στην παραπάνω περίπτωση). Αλλά ο b έχει μόλις δύο γραμμές, έτσι τον πολλαπλασιασμό

```
b%*%a
```

```
     [,1] [,2]
[1,]    5    1
[2,]   -1   -3
```

παράγει ένα πίνακα με 2 γραμμές. Η τιμή 5 στη γραμμή 1 στήλη 1 ως απάντηση είναι το άθροισμα των μερικών γινομένων (1 × 1) +(2 × 0) +(1 × 4) = 1 +0 +4 = 5.

Διαγώνιοι πινάκες

Για να δημιουργήσετε ένα διαγώνιο πίνακα 3 γραμμών και 3 στηλών, με 1s στη διαγώνιο χρησιμοποίησε η διαγώνια συνάρτηση σαν αυτή:

```
(ym<-diag(1,3,3))
```

```
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1
```

Μπορείτε να μεταβάλλετε τις τιμές των διαγώνιων στοιχείων του πίνακα όπως αυτό:

```
diag(ym)<-1:3
ym
```

```
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    2    0
[3,]    0    0    3
```

ή να εξάγει ένα διάνυσμα που περιέχει τα διαγώνια στοιχεία ενός πίνακα όπως αυτό:

```
diag(ym)
[1] 1 2 3
```

Μπορεί να θέλετε να εξαγάγετε τη διαγώνιο από ένα τετράγωνο τυπικής απόκλισης – το γινόμενο απόκλισης δύο τυχαίων από τον πίνακα:

```
M <-cbind(X=1:5, Y=rnorm(5))
var(M)
```

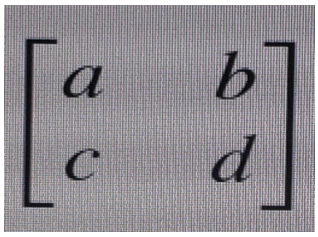
```
          X         Y
X 2.5000000 0.0434632
Y 0.0434632 0.8805603
```

```
diag(var(M))
```

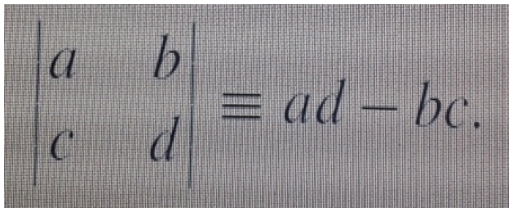
```
        X         Y
2.5000000 0.8805603
```

Ορίζουσα

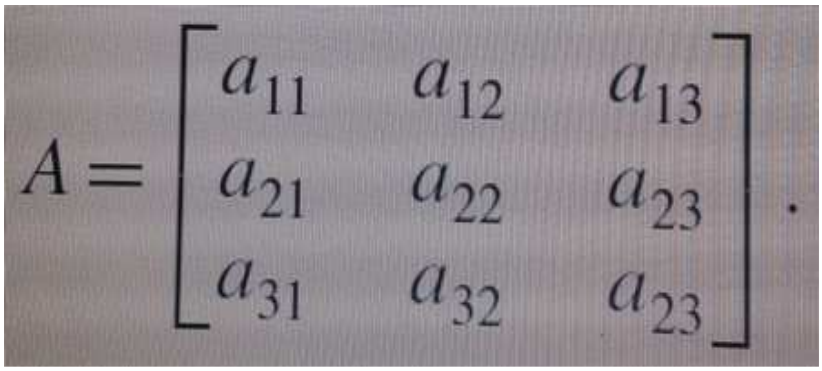
Η ορίζουσα του τετραγωνικού 2 × 2 πίνακα



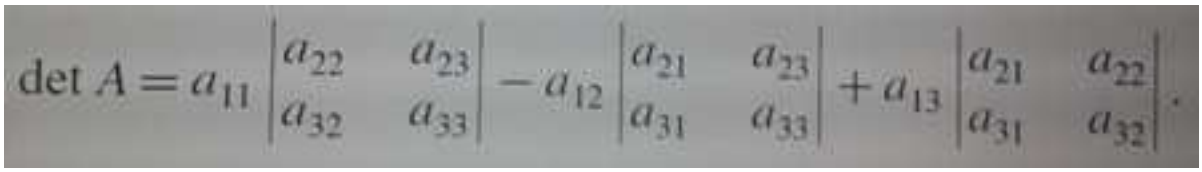
ορίζεται για τυχόν αριθμούς α, b, c και d όπως



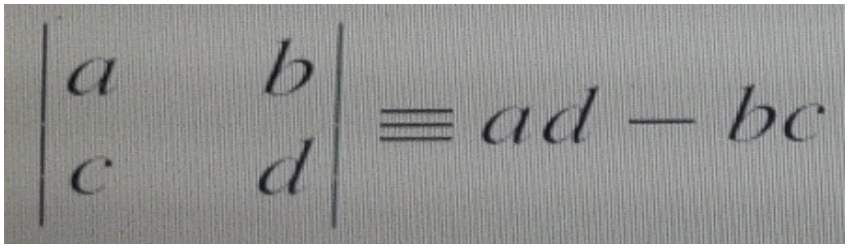
Ας υποθέσουμε ότι ο A είναι ένας τετραγωνικός πίνακας της τάξης (3 × 3):



Στη συνέχεια, η τρίτης τάξης ορίζουσα του A ορίζεται να είναι ο αριθμός



Εφαρμόζοντας τον κανόνα

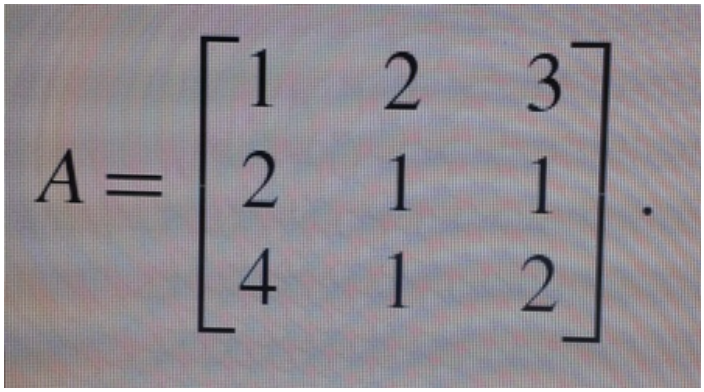


$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} \equiv ad - bc$$

σε αυτή την εξίσωση δίνει

detA=a11a22a33−a11a23a32+a12a23a31−a12a21a33+a13a21a32−a13a22a31.

Πάρτε ένα αριθμητικό παράδειγμα:



$$A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 1 \\ 4 & 1 & 2 \end{bmatrix}.$$

Αυτό έχει ορίζουσα

detA=(1×1×2)−(1×1×1)+(2×1×4)−(2×2×2)+(3×2×1)−(3×1×4)
=2−1+8−8+6−12=−5.
Εδώ είναι το παράδειγμα της R με την ορίζουσα det της συνάρτησης:

```
A<-matrix(c(1,2,4,2,1,1,3,1,2),nrow=3)
A
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    2    1    1
[3,]    4    1    2
```

```
det(A)

[1]   -5
```

Το μεγάλο πράγμα για τις ορίζουσες είναι ότι εάν υπάρχει κάποια γραμμή ή στήλη της ορίζουσας που πολλαπλασιάζεται με ένα αριθμό λ, τότε η τιμή της ορίζουσας πολλαπλασιάζεται με λ (επειδή ένας συντελεστής λ θα εμφανίζεται σε κάθε ένα από τα γινόμενα). Επίσης, αν όλα τα στοιχεία μιας γραμμής ή στήλης είναι μηδέν τότε η ορίζουσα |A| = 0. Και πάλι, αν όλα τα αντίστοιχα στοιχεία των δύο γραμμών ή
στηλών της |A| είναι ίσα τότε |A| = 0.
Για παράδειγμα, εδώ είναι η κατώτατη γραμμή του A πολλαπλασιασμένη επί 3:

```
B<-A
B[3,]<-3*B[3,]
B
      [,1]    [,2]    [,3]
[1,]     1       2       3
[2,]     2       1       1
[3,]    12       3       6
```

και εδώ είναι η ορίζουσα:

```
det(B)
[1]  -15
```

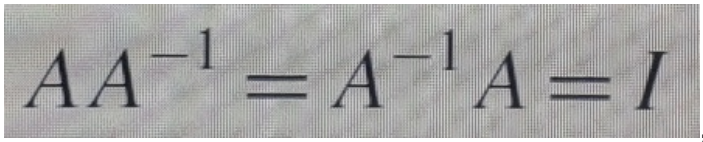
Εδώ είναι ένα παράδειγμα, όταν όλα τα στοιχεία της στήλης 2 είναι μηδέν, οπότε det C = 0:

```
C<-A
C[,2]<-0
C
      [,1]    [,2]    [,3]
[1,]     1       0       3
[2,]     2       0       1
[3,]     4       0       2


det(C)
[1]    0
```

Αν ορίζουσα του A≠0 τότε οι γραμμές και οι στήλες του A πρέπει να είναι γραμμικά ανεξάρτητες. Αυτή η σημαντική η έννοια έχει επεκταθεί σε όρους συντελεστών αντίθεσης στη σελ. 372.

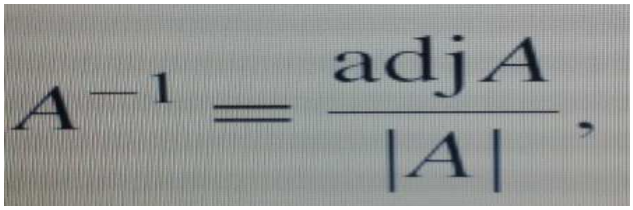
Αντίστροφος ενός πίνακα

Η λειτουργία της διαίρεσης δεν ορίζεται για τους πίνακες. Ωστόσο, για ένα τετράγωνικο πίνακα που έχει |A|≠0 μπορεί να οριστεί ένας πολλαπλασιαστικός αντίστροφος πίνακας συμβολίζεται με A^-1. Αυτός η πολλαπλασιαστικός αντίστροφος A^-1 είναι μοναδικός και έχει την ιδιότητα ότι

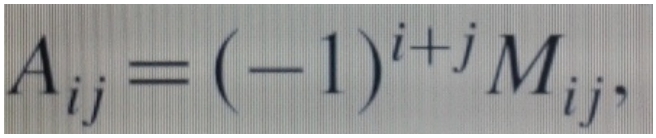

$$AA^{-1} = A^{-1}A = I$$

,

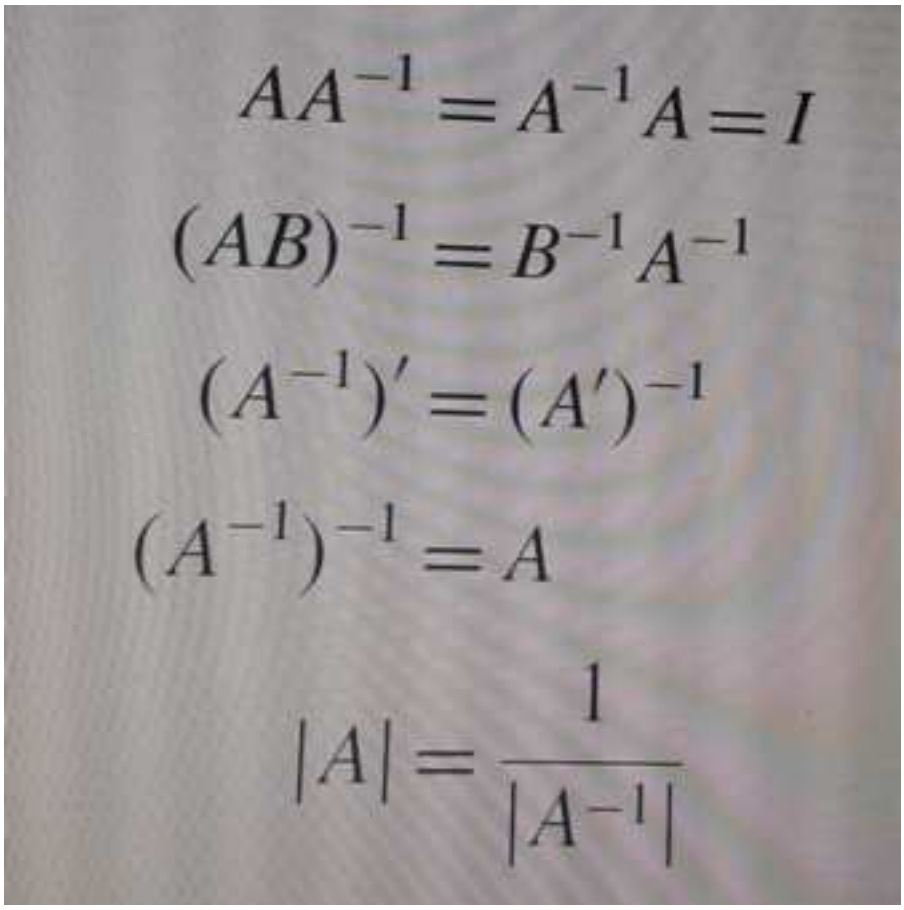
που I είναι ο μοναδιαίος πίνακας. Έτσι εάν A είναι ένας τετραγωνικός πίνακας για τον οποίο |A|≠0 ο πίνακας είναι αντιστρέψιμος ορισμένος από τη σχέση


$$A^{-1} = \frac{\text{adj}\,A}{|A|},$$

όπου ο συζυγής πίνακας του A (adjA) είναι ο πίνακας των συμπαραγόντων του A. Οι συμπαράγοντες του A υπολογίζονται ως


$$A_{ij} = (-1)^{i+j} M_{ij},$$

όπου Mij είναι οι «υποορίζουσες» των στοιχείων aij (αυτά είναι οι καθοριστικοί παράγοντες των πινάκων του A από τους οποίους i γραμμή και j στήλη έχουν διαγραφεί). Οι ιδιότητες του αντιστρόφου πίνακα μπορούν να προβλεφθούν για δύο μη μοναδικούς τετραγωνικούς πίνακες, A και B, της ίδιας τάξης ως εξής:



$$AA^{-1} = A^{-1}A = I$$

$$(AB)^{-1} = B^{-1}A^{-1}$$

$$(A^{-1})' = (A')^{-1}$$

$$(A^{-1})^{-1} = A$$

$$|A| = \frac{1}{|A^{-1}|}$$

Εδώ είναι η έκδοση της R του αντιστρόφου του 3 × 3 πίνακα A (παραπάνω) χρησιμοποιώντας τη συνάρτηση ginv από τη βιβλιοθήκη MASS

```
library(MASS)
ginv(A)
```

```
                    [,1]  [,2]  [,3]
[1,]  -2.000000e-01   0.2   0.2
[2,]  -2.224918e-16   2.0  -1.0
[3,]   4.000000e-01  -1.4   0.6
```

όπου ο αριθμός στη γραμμή 2 στήλη 1 είναι μηδέν (εκτός για το σφάλμα στρογγυλοποίησης).

Εδώ είναι η προτελευταίος κανόνας $(A^{-1})^{-1} = A$ αξιολογούμενος από την R:

```
ginv(ginv(A))
```

```
     [,1]  [,2]  [,3]
[1,]    1     2     3
[2,]    2     1     1
[3,]    4     1     2
```

και εδώ είναι και ο τελευταίος κανόνας $|A| = \frac{1}{|A^{-1}|}$:

```
1/det(ginv(A))
[1]  -5
```

Χαρακτηριστικές ρίζες και χαρακτηριστικά ανύσματα

Έχουμε ένα τετραγωνικό πίνακα Α και δύο διανύσματα στήλη Χ και Κ, όπου

$$AX=K,$$

και θέλουμε να ανακαλύψουμε τον αριθμό πολλαπλασιαστή τέτοιο ώστε

$$AX=\lambda X.$$

Αυτό είναι ισοδύναμο με (A-λI)X=0, που I είναι ο μοναδιαίος πίνακας. Αυτό μπορεί να έχει μόνο μία μη-μηδενική λύση, όταν ο καθοριστικός παράγοντας που σχετίζεται με το συντελεστή πίνακα Α εξαφανίζεται, οπότε θα πρέπει να έχουμε

$$|A-\lambda I|=0.$$

Όταν επεκταθεί, αυτός ο καθοριστικός δημιουργεί μια αλγεβρική εξίσωση n βαθμού στον λ ονομάζεται η **χαρακτηριστική εξίσωση**. Έχει n ρίζες λ1,λ2,….,λn, καθέμια από τις οποίες ονομάζεται **χαρακτηριστική ρίζα**. Η αντίστοιχη λύση διάνυσματος Xi λέγεται **χαρακτηριστικό άνυσμα** του Α αντίστοιχου του λi.
Εδώ είναι ένα παράδειγμα από την πληθυσμιακή οικολογία. Ο πίνακας Α δείχνει τη δημογραφία από διαφορετικές ηλικιακές κατηγορίες: η πρώτη γραμμή δείχνει γονιμότητα (ο αριθμός των γυναικών που γεννήθηκαν ανά γυναίκες κάθε ηλικίας) και οι υπο-διαγώνιοι δείχνουν τα ποσοστά επιβίωσης (το κλάσμα της μιας ηλικίας τάξης που επιβιώνει στην επόμενη ηλικιακή τάξη). Όταν αυτοί οι αριθμοί είναι σταθεροί ο πίνακας είναι γνωστός ως **πίνακας Leslie**. Ελλείψει εξάρτησης από την πυκνότητα οι σταθερές τιμές παραμέτρων στον Α θα οδηγήσουν είτε στην εκθετική αύξηση του συνολικού μεγέθους του πληθυσμού (εάν λ1> 1) ή μια εκθετική μείωση (εάν; λ1 <1), αφού οι αρχικές μεταβάσεις στη δομή ηλικίας έχουν απόσβεση μακριά. Μόλις εκθετική ανάπτυξη έχει επιτευχθεί, τότε η ηλικιακή διάρθρωση, όπως φαίνεται από το ποσοστό των ατόμων σε κάθε ηλικιακή κατηγορία, θα είναι μια σταθερά. Αυτό είναι γνωστό ως το πρώτο χαρακτηριστικό διάνυσμα.
Εξετάστε το Leslie πίνακα, L, ο οποίος πρόκειται να πολλαπλασιαστεί με ένα πίνακα στήλης ηλικιακοκατασκευαστικών μεγεθών πληθυσμού, n:

```
L<-c(0,0.7,0,0,6,0,0.5,0,3,0,0,0.3,1,0,0,0)
L<-matrix(L,nrow=4)
```

Σημειώνεται ότι τα στοιχεία του πίνακα που εγγράφονται στη στήλη, όχι η γραμμή-διαβασμένη ακολουθία. Μπορούμε να διασφαλίσουμε ότι ο πίνακας Leslie είναι σωστά συμμορφώσιμος:

```
L

     [,1] [,2] [,3] [,4]
[1,]  0.0  6.0  3.0    1
[2,]  0.7  0.0  0.0    0
[3,]  0.0  0.5  0.0    0
[4,]  0.0  0.0  0.3    0
```

Η ανώτατη γραμμή περιέχει τη συγκεκριμένη ηλικία γονιμοτήτων (π.χ. 2-ετών παράγει 6 θηλυκούς απογόνους ετησίως), και η υπο-διαγώνιος περιέχει τις ικανότητες επιβίωσης (70% του 1ενός έτους γίνεται 2 ετών, κλπ.). Τώρα τα μεγέθη του πληθυσμού σε κάθε ηλικία πάνε σε ένα διάνυσμα στήλης, n:

```
n<-c(45,20,17,3)
n<-matrix(n,ncol=1)

n
     [,1]
[1,]   45
[2,]   20
[3,]   17
[4,]    3
```

Τα μεγέθη πληθυσμού το επόμενο έτος σε καθεμία από τις τέσσερις κατηγορίες ηλικίας λαμβάνονται με πολλαπλασιασμό πινάκων, %*%

```
L %*% n

      [,1]
[1,] 174.0
[2,]  31.5
[3,]  10.0
[4,]   5.1
```

Μπορούμε να ελέγξουμε αυτό τον μακρύ δρόμο. Ο αριθμός των νεαρών επόμενου έτους (το πρώτο στοιχείο του n) είναι το άθροισμα όλων των βρεφών που γεννήθηκαν πέρυσι:

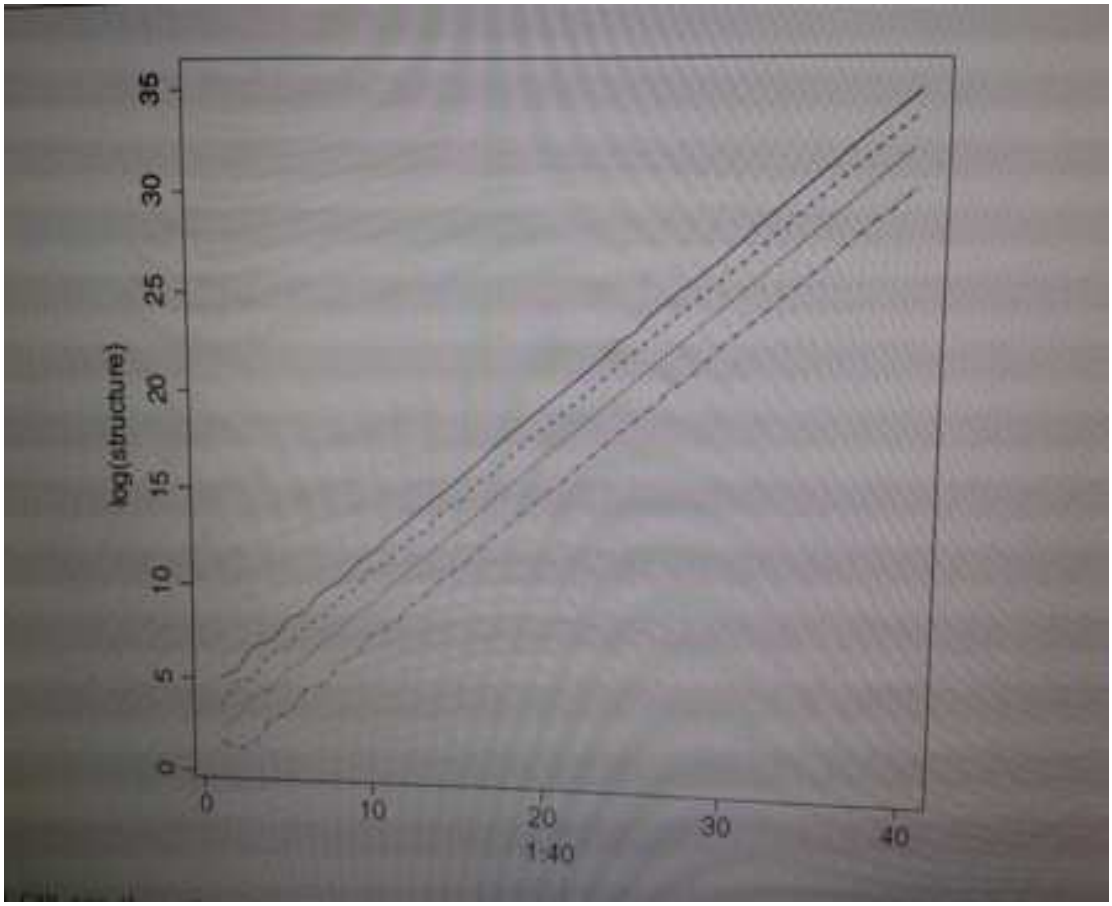
```
45*0+20*6+17*3+3*1

[1]  174
```

Γράφουμε μια συνάρτηση για την εκτέλεση του πολλαπλασιασμού πινάκων, δίνοντας πληθυσμό του επόμενου έτους ως διάνυσμα μιας συνάρτησης του τρέχοντος έτους που είναι:

```
fun<-function(x) L %*% x
```

Τώρα μπορούμε να προσομοιώσουμε τη δυναμική του πληθυσμού σε μια περίοδο αρκετά μεγάλου χρονικού διαστήματος (ας πούμε, 40 γενιές) για την ηλικιακή διάρθρωση να προσεγγίσει τη σταθερότητα. Εφ 'όσον ο ρυθμός αύξησης του πληθυσμού λ> 1, ο πληθυσμός θα αυξηθεί εκθετικά, αφού η ηλικιακή διάρθρωση έχει σταθεροποιηθεί:

```
n<-c(45,20,17,3)
n<-matrix(n,ncol=1)
structure<-numeric(160)
dim(structure)<-c(40,4)
for (i in 1:40) {
n<-fun(n)
structure[i,]<-n
}
matplot(1:40,log(structure),type="l")
```



Μπορείτε να δείτε ότι μετά από κάποιες αρχικές παροδικές διακυμάνσεις, η ηλικιακή διάρθρωση έχει περισσότερο ή λιγότερο σταθεροποιηθεί από το έτος 20 (οι γραμμές για το log μέγεθος του πληθυσμού καταγραφής των νεαρών (πάνω γραμμή), 1 -, 2 - και 3-ετών είναι παράλληλες). Μέχρι το έτος 40 ο πληθυσμός αυξάνεται εκθετικά σε μέγεθος, πολλαπλασιάζοντας με μια σταθερά του λ κάθε χρόνο.

Ο ρυθμός αύξησης του πληθυσμού (το ανά έτος ποσοστό πολλαπλασιασμού, λ) Προσεγγίζεται από την αναλογία του συνολικού μεγέθους του πληθυσμού στην 40η και 39η χρονιά:

```
sum(structure[40,])/sum(structure[39,])

[1]  2.164035
```

και η κατά προσέγγιση σταθερή δομή ηλικίας λαμβάνεται από την 40η τιμή του n:

```
structure[40,]/sum(structure[40,])

[1]      0.709769309  0.230139847  0.052750539  0.007340305
```

Οι ακριβείς τιμές του ρυθμού αύξησης του πληθυσμού και την σταθερή κατανομή ηλικίας λαμβάνεται με άλγεβρα πινάκων: είναι η κυρίαρχη χαρακτηριστική ρίζα και κυρίαρχο χαρακτηριστικό διάνυσμα, αντίστοιχα. Χρησιμοποιήστε τη συνάρτηση eigen εφαρμόζεται στο πίνακα Leslie, L, όπως αυτό:

```
eigen(L)

$values
[1] 2.1694041+0.0000000i  -1.9186627+0.0000000i  -0.1253707+0.0975105i
[4] -0.1253707-0.0975105i
$vectors

                    [,1]                 [,2]                          [,3]
[1,] -0.949264118+0i  -0.93561508+0i  -0.01336028-0.03054433i
[2,] -0.306298338+0i   0.34134741+0i  -0.03616819+0.14241169i
[3,] -0.070595039+0i  -0.08895451+0i   0.36511901-0.28398118i
[4,] -0.009762363+0i   0.01390883+0i  -0.87369452+0.00000000i


                                 [,4]
[1,]      -0.01336028+0.03054433i
[2,]      -0.03616819 -0.14241169i
[3,]       0.36511901+0.28398118i
[4,]      -0.87369452+0.00000000i
```

Η κύρια χαρακτηριστική ρίζα είναι 2,1694 (σε σύγκριση με την εμπειρική προσέγγιση μας 2,1640 μετά από 40 χρόνια). Η σταθερή ηλικιακή κατανομή δίνεται από τον πρώτο χαρακτηριστικό διάνυσμα, το οποίο πρέπει να μετατρέψουμε σε αναλογίες

```
eigen(L)$vectors[,1]/sum(eigen(L)$vectors[,1])

[1] 0.710569659+0i 0.229278977+0i 0.052843768+0i 0.007307597+0i
```

Αυτό μπορεί να συγκριθεί με την προσέγγιση μας (παραπάνω), στην οποία η αναλογία στην πρώτη ηλικιακή κατηγορία ήταν 0,70977 μετά από 40 χρόνια (αντί για 0,71057).

Πίνακες σε στατιστικά μοντέλα

Ίσως η κύρια χρήση των πινάκων στο R είναι σε στατιστικούς υπολογισμούς, στη γενίκευση της ο υπολογισμός των αθροισμάτων των τετραγώνων και των

αθροισμάτων των γινομένων (βλ. σελ.. 388 για το φόντο). Εδώ είναι τα δεδομένα που χρησιμοποιούνται στο κεφάλαιο 10 για να εισαγάγουν τον υπολογισμό των αθροισμάτων των τετραγώνων στη γραμμική εμπειρική σχέση μεταβλητών:

```
numbers<-read.table("c:\\temp\\tannin.txt",header=T)
attach(numbers)
names(numbers)

[1]       "growth"        "tannin"
```

Η μεταβλητή απόκρισης είναι η ανάπτυξη (y) και η επεξηγηματική μεταβλητή είναι η συγκέντρωση tannin (x) στη διατροφή μιας ομάδας προνύμφες (κάμπιες) εντόμων. Χρειαζόμαστε το περίφημο πέντε (βλ. σελ. 270.): Το άθροισμα των τιμών του y,

```
growth

[1]  12  10   8  11   6   7   2   3   3

sum(growth)

[1]      62
```

Το άθροισμα των τετραγώνων των τιμών y

```
growth^2

[1]  144  100   64  121   36   49   4   9   9

sum(growth^2)

[1]       536
```

το άθροισμα των τιμών x,

```
tannin

[1]  0   1   2   3   4   5   6   7   8

sum(tannin)

[1]      36
```

το άθροισμα των τετραγώνων των τιμών x,

```
tannin^2

[1] 0 1 4 9 16 25 36 49 64

sum(tannin^2)

[1]  204
```

και, τέλος, για τη μέτρηση του γινομένου μεταξύ x και y, χρειαζόμαστε το άθροισμα των γινομένων,

```
growth*tannin

[1] 0 10 16 33 24 35 12 21 24

sum(growth*tannin)

[1]  175
```

Μπορείτε να δείτε αμέσως ότι για πιο πολύπλοκα μοντέλα (όπως η πολλαπλή εμπειρική σχέση μεταβλητών), είναι σημαντικό να είναι σε θέση να γενικεύσουν και την απλοποίηση της διαδικασίας αυτής. Αυτό είναι όπου οι πίνακες έρχονται σε. Πολλαπλασιασμό πίνακα που περιλαμβάνει τον υπολογισμό των αθροισμάτων των γινομένων, όπου ένα διάνυσμα γραμμής πολλαπλασιάζεται με ένα διάνυσμα στήλης του ίδιου μήκους για να ληφθεί μία μόνο τιμή. Έτσι, θα πρέπει να είμαστε σε θέση να εξασφαλίσουμε το απαιτούμενο άθροισμα των γινομένων, 175, χρησιμοποιώντας το σύμβολο πολλαπλασιασμού πίνακα % *% στη θέση του τακτικού σύμβολου πολλαπλασιασμού:

```
growth%*%tannin

[,1]

[1,]  175
```

Αυτό δουλεύει μια χαρά. Αλλά τι γίνεται με τα αθροίσματα των τετραγωνικών; Σίγουρα αν χρησιμοποιούμε πολλαπλασιασμό πινάκα στο ίδιο διάνυσμα θα έχουμε ένα αντικείμενο με πολλές γραμμές (9 σε αυτήν την περίπτωση). Δεν είναι έτσι.

```
growth%*%growth

[,1]

[1,]  536
```

Η R έχει εξαναγκάσει το αριστερό διάνυσμα της ανάπτυξης σε ένα διάνυσμα γραμμής προκειμένου να επιτευχθεί το επιθυμητό αποτέλεσμα. Μπορείτε να παρακάμψετε αυτό, αν για κάποιο λόγο που ήθελε την απάντηση να έχει 9 γραμμές, καθορίζοντας το μεταφέρουν t () του δεξιού διανύσματος ανάπτυξης,

```
growth%*%t(growth)

      [,1]  [,2] [,3]  [,4] [,5] [,6] [,7] [,8] [,9]
[1,]  144  120   96   132   72   84   24   36   36
[2,]  120  100   80   110   60   70   20   30   30
[3,]   96   80   64    88   48   56   16   24   24
[4,]  132  110   88   121   66   77   22   33   33
[5,]   72   60   48    66   36   42   12   18   18
[6,]   84   70   56    77   42   49   14   21   21
[7,]   24   20   16    22   12   14    4    6    6
[8,]   36   30   24    33   18   21    6    9    9
[9,]   36   30   24    33   18   21    6    9    9
```

αλλά, φυσικά, δεν είναι αυτό που θέλουμε. Έλλειψη της R είναι ό, τι χρειαζόμαστε. Έτσι, αυτό θα πρέπει επίσης να εργαστεί για την εξασφάλιση του αθροίσματος των επεξηγηματικών μεταβλητών των τετραγωνικών (πινάκων):

```
tannin%*%tannin

     [,1]
[1,]  204
```

Μέχρι στιγμής, όλα καλά. Αλλά πώς θα αποκτήσουμε τα αθροίσματα χρησιμοποιώντας πολλαπλασιασμό πίνακα; Το τέχνασμα εδώ είναι να πολλαπλασιάσει ο πίνακας του διανύσματος με ένα διάνυσμα του 1s: εδώ είναι το άθροισμα των τιμών του y,

```
growth%*%rep(1,9)

     [,1]
[1,]   62
```

και το άθροισμα των τιμών x,

```
tannin%*%rep(1,9)

     [,1]
[1,]   36
```

Τέλος, μπορούμε να χρησιμοποιήσουμε τον πολλαπλασιασμό πίνακα να φτάσει στο μέγεθος του δείγματος, n; Το κάνουμε αυτό με πίνακα πολλαπλασιάζοντας ένα διάνυσμα γραμμή 1s με ένα διάνυσμα στήλη 1s. Αυτή η μάλλον περίεργη λειτουργία παράγει το σωστό αποτέλεσμα, με την πρόσθεση της από εννέα 1s που προκύπτουν από τις εννέα επαναλήψεις του υπολογισμού 1 × 1:

```
rep(1,9)%*%rep(1,9)
```

```
     [,1]
[1,]    9
```

Αλλά πώς θα πάρουμε όλα τα περίφημα πέντε σε ένα ενιαίο πίνακα; Το πράγμα που πρέπει να κατανοήσουμε είναι η διάσταση του εν λόγω πίνακα. Θα πρέπει να περιέχει τα αθροίσματα, καθώς και τα αθροίσματα των γινομένων. Έχουμε δύο μεταβλητές (growth and tannin) και ο πολλαπλασιασμός του πίνακα τους παράγει μια μοναδική βαθμωτή (αριθμητική) τιμή (βλ. παραπάνω). Κατάλληλο για να πάρει τα αθροίσματα των τετραγώνων, καθώς και τα αθροίσματα των γινομένων χρησιμοποιούμε cbind να δημιουργήσουμε ένα 9 × 2 πίνακα σαν αυτό:

```
a<-cbind(growth,tannin)
a
```

```
      growth tannin
[1,]     12      0
[2,]     10      1
[3,]      8      2
[4,]     11      3
[5,]      6      4
[6,]      7      5
[7,]      2      6
[8,]      3      7
[9,]      3      8
```

Για να αποκτήσετε έναν πίνακα αποτελεσμάτων με 2 γραμμές αντί για 9 γραμμές θα πρέπει να πολλαπλασιάσουμε την μεταφορά από πίνακα α με πίνακα α:

```
t(a)%*%a
```

```
        growth tannin
growth     536    175
tannin     175    204
```

Αυτό είναι εντάξει όσο πάει, αλλά μας έδωσε μόνο τα αθροίσματα των τετραγώνων (536 και 204) και το άθροισμα των γινομένων (175). Πώς θα

πάρετε τα αθροισματα επίσης; Το κόλπο είναι να δεσμεύσει μια στήλη από 1s πάνω αριστερά του πίνακα α:

```
b<-cbind(1,growth,tannin)
b
```

```
     [,1] growth tannin
[1,]    1     12      0
[2,]    1     10      1
[3,]    1      8      2
[4,]    1     11      3
[5,]    1      6      4
[6,]    1      7      5
[7,]    1      2      6
[8,]    1      3      7
[9,]    1      3      8
```

Θα εξετάσουμε καλύτερα αν η πρώτη στήλη είχε ένα όνομα μεταβλητής: ας το ονομάσουμε «δείγμα»:

```
dimnames(b)[[2]] [1]<-"sample"
```

Τώρα για να πάρετε ένα συνοπτικό πίνακα των αθροισμάτων, καθώς και τα αθροίσματα των γινομένων, πολλαπλασιάζουμε το πίνακα b από τον ίδιο. Θέλουμε η απάντηση να έχει τρεις γραμμές (και όχι από εννέα), έτσι ώστε εμείς πολλαπλασιάζουμε το πίνακα μεταφοράς του b (ο οποίος έχει τρεις γραμμές) με b (ο οποίος έχει εννέα γραμμές):

```
t(b)%*%b
```

```
       sample growth tannin
sample      9     62     36
growth     62    536    175
tannin     36    175    204
```

Έτσι εκεί το έχετε. Όλα τα περίφημα πέντε, συν το μέγεθος του δείγματος, σε ένα ενιαίο πολλαπλασιασμό πίνακα.

Στατιστικά μοντέλα στη σύστημα συμβόλων πίνακα

Συνεχίζουμε αυτό το παράδειγμα για να δείξει πώς η άλγεβρα πίνακα χρησιμοποιείται για να γενικεύσουν τις διαδικασίες που χρησιμοποιούνται σε γραμμική μοντελοποίηση (όπως τέτοια εμπειρική σχέση μεταβλητών ή ανάλυση τετραγώνου τυπικής απόκλισης) με βάση τις τιμές του διάσημου πέντε. Θέλουμε να είμαστε σε θέση να καθορίσουμε τις εκτιμήσεις παραμέτρων (όπως η τομή και η κλίση της γραμμικής εμπειρικής σχέσης μεταβλητών) και να κατανείμει το συνολικό άθροισμα των τετραγώνων μεταξύ

απόκλισης από το μέσο όρο εξηγείται από το μοντέλο (*SSR*) και ανεξήγητη απόκλιση από το μέσο όρο (SSE). Εκφράζονται σε όρους πίνακα, το μοντέλο γραμμικής εμπειρικής σχέσης μεταβλητών είναι
Y =Xb+e,

και θέλουμε να προσδιοριστεί η ελαχίστων τετραγώνων εκτίμηση του b, δίνεται από

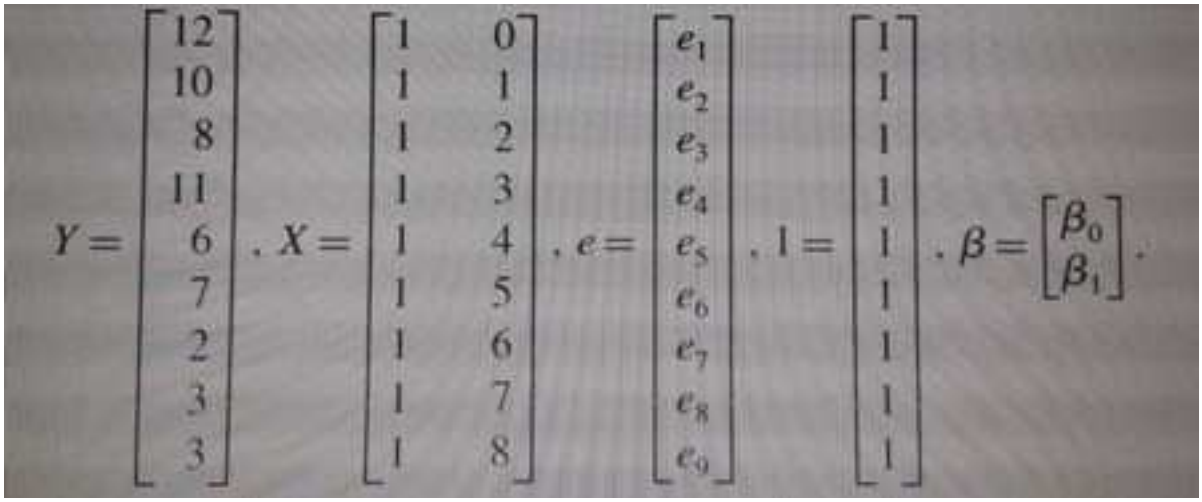
b=(X'X)−1 X'Y ,

και στη συνέχεια να εκτελέσει την ανάλυση του τετραγώνου τυπικής απόκλισης

b'X'Y'

Θα εξετάσουμε κάθε ένα από αυτά με τη σειρά.

Η μεταβλητή απόκρισης Y, 1 και τα λάθη e είναι απλά n×1 διανύσματα στήλη, X είναι ένας n×2 πίνακας και β είναι ένα 2×1 διάνυσμα από συντελεστές όπως ακολούθως:



Το διάνυσμα y και το διάνυσμα 1 δημιουργούνται όπως αυτό :

```
Y<-growth
one<-rep(1,9)
```

Το μέγεθος του δείγματος δίνεται από τον τύπο 1'1 (μεταφορά από διάνυσμα 1 φορές το διάνυσμα 1):

```
t(one) %*% one

     [,1]
[,1]    9
```

Το διάνυσμα από επεξηγηματική μεταβλητή X έχει δημιουργηθεί με πρόσδεση μιας στήλης από αυτές προς τα αριστερά

```
X<-cbind(1,tannin)
X
```

```
      tannin
[1,]  1      0
[2,]  1      1
[3,]  1      2
[4,]  1      3
[5,]  1      4
[6,]  1      5
[7,]  1      6
[8,]  1      7
[9,]  1      8
```

Σε αυτή τη σημείωση

$$\sum y^2 = y_1^2 + y_2^2 + \cdots + y_n^2 = Y'Y,$$

```
t(Y)%*%Y
```

```
     [,1]
[,1]  536
```

$$\sum y = n\bar{y} = y_1 + y_2 + \cdots + y_n = 1'Y$$

```
t(one)%*%Y
```

```
     [,1]
[,1]   62
```

$$\left(\sum y\right)^2 = Y'11'Y$$

```
t(Y) %*% one %*% t(one) %*% Y
```

```
      [1,]
[1,]  3844
```

Για το πίνακα των ερμηνευτικών μεταβλητών, εμείς βλέπουμε ότι X'X δίνει ένα 2×2 πίνακα που περιεχει n, Σx και Σ(x^2). Οι αριθμητικές τιμές είναι εύκολο να βρεθούν χρησιμοποιώντας πολλαπλασιασμό πίνακα %*%

```
t(X)%*%X
```

```
                tannin
            9       36
tannin     36      204
```

Σημείωσε ότι X'X (ένας 2×2 πίνακας) είναι εντελώς διαφορετικός από XX' (ένας 9×9 πίνακας). Ο πίνακας X'Y δίνει ένα 2×1 πίνακα που περιέχει Σy και το άθροισμα των γινομένων Σxy:

```
t(X)%*%Y
```

```
            [,1]
              62
tannin       175
```

Τώρα, με την όμορφη συμμετρία των κανονικών εξισώσεων

$$b_0 n + b_1 \sum x = \sum y,$$

$$b_0 \sum x + b_1 \sum x^2 = \sum xy,$$

μπορούμε να γράψουμε την εμπειρική σχέση μεταβλητών απευθείας σε μορφή πίνακα, όπως

$$X'Xb = X'Y$$

Επειδή εμείς είδη έχουμε τους αναγκαίους πίνακες από την αριστερή και δεξιά πλευρές. Για να βρείτε τις ελαχίστων τετραγώνων τιμές των παραμέτρων b θα πρέπει να διαιρέσετε τις δύο πλευρές με X'X. Αυτό περιλαμβάνει τον υπολογισμό του αντίστροφου του X'X πίνακα. Το αντίστροφο ισχύει μόνο όταν ο πίνακας είναι τετράγωνος και όταν ορίζουσα είναι μη-μοναδική. Ο αντίστροφος περιλαμβάνει $-\bar{x}$ και Σ(x^2) όπως οι όροι του, με

$$SSX = \sum (x - \bar{x})^2,$$

το άθροισμα των τετραγώνων των διαφορών μεταξύ των τιμών x και των μέσων τιμών x, ή n.SSX ως παρονομαστής:

$$(X'X)^{-1} = \begin{bmatrix} \frac{\sum x^2}{n\sum (x-\bar{x})^2} & \frac{-\bar{x}}{\sum (x-\bar{x})^2} \\ \frac{-\bar{x}}{\sum (x-\bar{x})^2} & \frac{1}{\sum (x-\bar{x})^2} \end{bmatrix}.$$

Όταν κάθε στοιχείο από ένα πίνακα έχει ένα κοινό παράγοντα, μπορεί να ληφθεί εκτός του πίνακα. Εδώ, ο όρος 1 /n(SSX) μπορεί να ληφθεί έξω για να δώσει

$$(X'X)^{-1} = \frac{1}{n\sum (x-\bar{x})^2} \begin{bmatrix} \sum x^2 & -\sum x \\ -\sum x & n \end{bmatrix}.$$

Υπολογίζοντας την αριθμητική τιμή από αυτό είναι εύκολη η χρήση της συνάρτησης του πίνακα ginv:

```
library(MASS)
ginv(t(X)%*%X)
```

```
                [,1]          [,2]
[1,]  0.37777778  -0.06666667
[2,] -0.06666667   0.01666667
```

Τώρα μπορούμε να λύσουμε τις κανονικές εξισώσεις

(X'X)−1(X'X)b=(X'X)−1X'Y,

Χρησιμοποιώντας το γεγονός ότι (X'X)−1(X'X)=I εξασφαλίζεις το σημαντικό γενικό αποτέλεσμα:

b=(X'X)−1X'Y,

ginv(t(X)%*%X)%*%t(X)%*%Y

```
     [,1]
[1,] 11.755556
[2,] -1.216667
```

που θα αναγνωρίσει από υπολογισμούς με το χέρι μας ως την τομή και κλίση αντίστοιχα (βλ. σελ. 392). Οι υπολογισμοί ANOVA είναι ως εξής. Ο συντελεστής(σταθερός όρος) διόρθωσης είναι

CF=Y'11'Y/n

```
CF<-t(Y) %*% one %*% t(one)%*% Y/9
CF
     [,1]
[1,] 427.1111
```

Το συνολικό άθροισμα των τετραγώνων, SSY, είναι Y'Y-CF:

```
t(Y)%*%Y-CF
     [,1]
[1,] 108.8889
```

Το άθροισμα των τετραγώνων εμπειρικής σχέσης μεταβλητών, SSR, είναι b'X'Y-CF:

```
b<-ginv(t(X)%*%X)%*%t(X)%*%Y
t(b)%*%t(X)%*%Y-CF
     [,1]
[1,] 88.81667
```

και το σφάλμα αθροίσματος των τετραγώνων, SSE είναι Y'Y-b'X'Y

```
t(Y) %*% Y - t(b) %*% t(X) %*% Y
     [,1]
[1,] 20.07222
```

Θα πρέπει να ελέγξετε τα ψηφία αυτά κατά τους υπολογισμούς με το χέρι επί σελ. 396. Προφανώς, αυτός δεν είναι ένας λογικός τρόπος για να πραγματοποιήσει μια απλή γραμμική εμπειρική σχέση μεταβλητών, αλλά εξηγεί πώς να γενικεύσουμε τους υπολογισμούς για τις περιπτώσεις που έχουν δύο ή περισσότερες συνεχής επεξηγηματικές μεταβλητές.

Επίλυση συστημάτων γραμμικών εξισώσεων χρησιμοποιώντας πίνακες

Ας υποθέσουμε ότι έχουμε δύο εξισώσεις που περιέχουν δύο άγνωστες μεταβλητές:

$$3x+4y=12,$$
$$x+2y=8.$$

Μπορούμε να χρησιμοποιήσουμε τη συνάρτηση solve(επίλυσης) για να βρούμε τις τιμές των μεταβλητών, αν εμείς παρέχουμε αυτή με δύο πίνακες:

• ένας τετραγωνικός πίνακας A που περιέχει τους συντελεστές (3, 1, 4 και 2, στις στήλες);
• ένα kv διάνυσμα στήλης που περιέχει τις γνωστές τιμές (12 και 8).

Θέτουμε τους δύο πίνακες σαν κι αυτόν (στήλη, ως συνήθως)

```
A<-matrix(c(3,1,4,2),nrow=2)
A
```

```
     [,1] [,2]
[1,]    3    4
[2,]    1    2
```

```
kv<-matrix(c(12,8),nrow=2)
kv
```

```
     [,1]
[1,]   12
[2,]    8
```

Τώρα μπορούμε να λύσουμε τις ταυτόσημες εξισώσεις

```
solve(A,kv)
```

```
     [,1]
[1,]   -4
[2,]    6
```

να δώσει x=-4 και y=6 (το οποίο μπορείς εύκολα να επαληθεύσεις με το χέρι). Η συνάρτηση είναι πιο χρήσιμη όταν υπάρχουν πολλές ταυτόσημες εξισώσεις που πρέπει να επιλυθούν.

Ανάλυση

Οι κανόνες της διαφοροποίησης και της ολοκλήρωσης είναι γνωστές στην R. Θα χρησιμοποιήσεις αυτές στη μοντελοποίηση (π.χ. κατά τον υπολογισμό των τιμών εκκίνησης σε μη-γραμμική εμπειρική σχέση μεταβλητών) και χρησιμοποιώντας optim για την αριθμητική ελαχιστοποίηση. Διαβάστε τα

αρχεία βοήθειας για την D και την integrate (ολοκλήρωση) να κατανοήσουν τα όρια αυτών των συναρτήσεων.

Παράγωγοι

Η συνάρτηση R για συμβολικές και αλγοριθμικές παράγωγους απλών εκφράσεων είναι D. Εδώ είναι μερικά απλά παραδείγματα για να σας δώσουν την ιδέα. Δείτε επίσης; deriv.

```
D(expression(2*x^3),"x")
2 * (3 * x^2)
D(expression(log(x)),"x")
1/x
D(expression(a*exp(-b * x)),"x")
-(a * (exp(-b * x) * b))
D(expression(a/(1+b*exp(-c * x))),"x")
a * (b * (exp(-c * x) * c))/(1 + b * exp(-c * x))^2
trig.exp <-expression(sin(cos(x + y^2)))
D(trig.exp, "x")
-(cos(cos(x + y^2)) * sin(x + y^2))
```

Ολοκληρώματα

Η συνάρτηση R είναι ολοκληρώσιμη. Εδώ είναι μερικά απλά παραδείγματα για να σας δώσουμε την ιδέα:

```
integrate(dnorm,0,Inf)
0.5 with absolute error < 4.7e-05
integrate(dnorm,-Inf,Inf)
1 με απόλυτο σφάλμα < 9.4e-05
integrate(function(x) rep(2, length(x)), 0, 1)
2 με απόλυτο σφάλμα < 2.2e-14
integrand <-function(x) {1/((x+1)*sqrt(x))}
integrate(integrand, lower = 0, upper = Inf)
3.141593 με απόλυτο σφάλμα < 2.7e-05
xv<-seq(0,10,0.1)
plot(xv,integrand(xv),type="l")
```

Η περιοχή κάτω από την καμπύλη είναι π=3,141593.

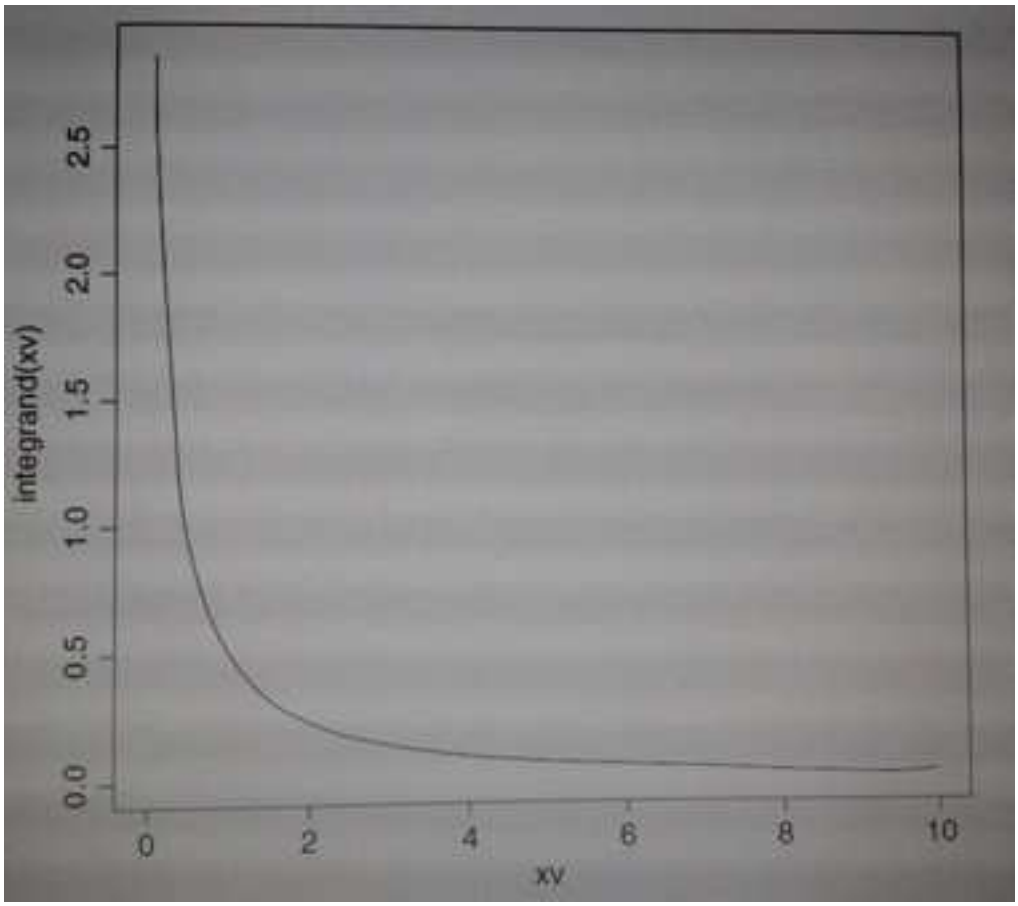
Διαφορικές Εξισώσεις

Πρέπει να λύσουμε ένα σύστημα συνήθων διαφορικών εξισώσεων (Σ.Δ.Ε.) και να επιλέξετε να χρησιμοποιήσετε το κλασικό Runge-Kutta τέταρτης τάξης συνάρτηση ολοκλήρωσης rk4 από το πακέτο odesolve:

```
install.packages("odesolve")
library(odesolve)
```

Το παράδειγμα αφορά ένα απλό οριακό μέσον φυτών φυτοφάγων ζώων όπου V = βλάστηση και N = πληθυσμός φυτοφάγων. Πρέπει να καθορίσουμε δύο διαφορικές εξισώσεις: μία για τη βλάστηση (dV / dt) και μία για τον πληθυσμό φυτοφάγων (dN / dt):

$$\frac{dV}{dt} = rV\left(\frac{K - V}{K}\right) - bVN,$$

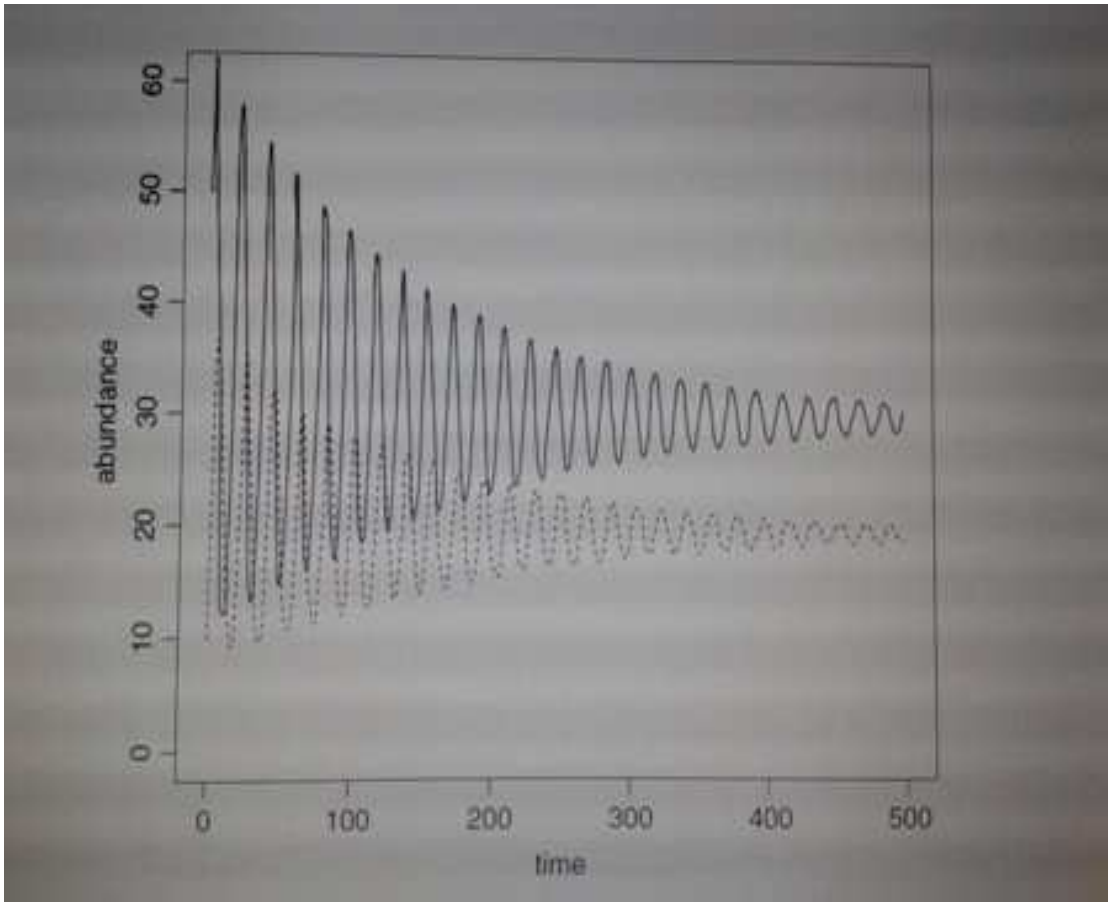


Τα βήματα που εμπλέκονται στην επίλυση αυτών των διαφορικών εξισώσεων (Σ.Δ.Ε.) στην R έχουν ως εξής:

• Ορίστε μια συνάρτηση (που ονομάζεται phmodel σε αυτήν την περίπτωση).
• Ονομάστε το μεταβλητές απόκρισης V και N από x [1] και x [2].
• Γράψτε την εξίσωση της βλάστησης, όπως dv όπως έχει ανωτέρω.
• Γράψτε την εξίσωση των φυτοφάγων όπως dn όπως έχει ανωτέρω.
• Συνδυάστε αυτά τα διανύσματα σε μια λίστα που ονομάζεται res.
• Δημιουργήστε μια χρονική σειρά άλλη μια φορά κατά την οποία να επιλύσετε τις εξισώσεις.
• Εδώ, t είναι από 0 έως 500 σε βήματα του 1.
• Ρυθμίστε τις τιμές των παραμέτρων parms.
• Ρυθμίστε τις τιμές εκκίνησης για V και N στο y και xstart.
• Χρησιμοποιήστε rk4 να δημιουργήσετε ένα πλαίσιο στοιχείων με τις V και N χρονοσειρές.

```
phmodel <-function(t, x, parms) {
v<-x[1]
n<-x[2]
with(as.list(parms), {
dv<-r*v*(K-v)/K - b*v*n
dn<-c*v*n – d*n
res<-c(dv, dn)


list(res)
})}

times <-seq(0, 500, length=501)

parms <-c(r=0.4, K=1000, b=0.02, c=0.01, d=0.3)

y<-xstart <-c(v=50, n=10)
output <-as.data.frame(rk4(xstart, times, phmodel, parms))

plot (output$time, output$v,
ylim=c(0,60),type="n",ylab="abundance",xlab="time")
lines (output$time, output$v)
lines (output$time, output$n,lty=2)
```

Η έξοδος δείχνει αφθονία φυτών ως μια σταθερή γραμμή έναντι στο χρόνο και την αφθονία φυτοφάγων ως διακεκομμένη γραμμή:



Τα εκθέματα του συστήματος απόσβεσης ταλαντώσεων σε ένα σταθερό σημείο ισορροπίας στο οποίο dV / dt και dN / dt είναι και τα δυο ίσα με μηδέν, οπότε ισορροπία αφθονίας φυτών = d / c = 0,3/0,01 = 30 και η ισορροπία αφθονίας φυτοφάγων = r (K-V *) / bK = 19,4.