Ασκήσεις Προσομοίωσης - Σειρά 1

Α. Μπουρνέτας, ΠΜΣ 'Βιοστατιστική'

Για τα παραδείγματα που ακολουθούν θα χρησιμοποιήσουμε τις παρακάτω γεννήτριες τυχαίων αριθμών από το P (η παράμετρος n δηλώνει τον αριθμό παρατηρήσεων που θα δημιουργηθούν):

1. Ομοιόμορφη $U(a,b)$: runif(n,a,b)
2. Κανονική $N(m,s^2)$: rnorm(n,m,s)
3. t-Student t_N: rt(n,N)
4. Εκθετική $Exp(m)$: rexp(n,m)
5. Γάμμα $Gamma(k,L)$: rgamma(n,k,L)
6. Poisson $P(m)$: rpois(n,m)

1 Γεννήτριες Τυχαίων Αριθμών

ΠΡΟΒΛΗΜΑ 1. Να δημιουργηθεί ένα δείγμα τυχαίων αριθμών μεγέθους 20 από την εκθετική κατανομή Exp(5) (μέση τιμή 1/5). Να βρεθεί ο δειγματικός μέσος και η δειγματική τυπική απόκλιση, και να ελεγχθεί μέσω του test Kolmogorov-Smirnov η προσαρμογή στην κατανομή.

Χρησιμοποιούμε τη γεννήτρια του R:

```
> x=rexp(20,5)
> x
 [1] 0.6411876052 0.6056879519 0.0693769450 0.0586497131 0.3245735187
 [6] 0.4881836850 0.0304149647 1.1611939024 0.0004887973 0.0380583467

[11] 0.3832830292 0.2688941348 0.3314560719 0.1367555471 0.0195411063
[16] 0.5710361421 0.3417361855 0.1622198930 0.1101890027 0.2176947983
```

Ο έλεγχος για την εκθετική κατανομή με παράμετρο 5 δίνει

```
> ks.test(x,"pexp", 5)

        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.2527, p-value = 0.1301
alternative hypothesis: two-sided
```

Σημειώστε ότι η εντολή ks.test παίρνει όρισμα την αθροιστική συνάρτηση για την αντίστοιχη κατανομή που ελέγχεται, π.χ. για την εκθετική το όρισμα είναι "pexp".
Στο παράδειγμά μας η μηδενική υπόθεση δεν απορρίπτεται, επομένως δεν υπάρχει ένδειξη απόκλισης από τη ζητούμενη κατανομή. Το ίδιο θα συμβεί όμως και αν ελέγξουμε ως προς την εκθετική κατανομή με άλλες παραμέτρους, π.χ. για $\lambda = 4$

```
> ks.test(x,"pexp", 4)

        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.177, p-value = 0.5024
alternative hypothesis: two-sided
```

και πάλι δεν απορρίπτεται, ενώ για $\lambda = 6$ απορρίπτεται:

```
> ks.test(x,"pexp", 6)

        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.3074, p-value = 0.03562
alternative hypothesis: two-sided
```

Ιδανικά θα θέλαμε η μηδενική υπόθεση να μην απορρίπτεται για $\lambda - 4$, και να απορρίπτεται για όλες τις τιμές $\lambda \neq 4$. Η ανακρίβεια που παρατηρούμε οφείλεται στο ότι το μέγεθος δείγματος είναι μικρό. Αν επαναληφθεί το πείραμα με μεγαλύτερο μέγεθος, π.χ. 2000, τα αποτελέσματα θα είναι πολύ πιο ακριβή (δοκιμάστε το).

ΠΡΟΒΛΗΜΑ 2. Να δημιουργηθεί γεννήτρια τυχαίων αριθμών από την κατανομή Weibull(k, λ) χρησιμοποιώντας τη μέθοδο αντιστροφής.

Η κατανομή Weibull(k, λ) έχει συνάρτηση πυκνότητας πιθανότητας

$$f(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k}, \quad x \geq 0$$

και αθροιστική συνάρτηση κατανομής

$$F(x) = 1 - e^{-\left(\frac{x}{\lambda}\right)^k}, \quad x \geq 0.$$

Σύμφωνα με τη μέθοδο αντιστροφής, αν $U \sim U(0,1)$, τότε η X για την οποία ισχύει η εξίσωση $F(X) = U$ ακολουθεί την κατανομή F. Εδώ επομένως έχουμε

$$1 - e^{-\left(\frac{X}{\lambda}\right)^k} = U$$

και λύνοντας ως προς X παίρνουμε

$$X = \lambda \left(-\ln(1-U)\right)^{1/k}.$$

Ο κώδικας για τη δημιουργία της Weibull(4,5) (δηλαδή $k = 4, \lambda = 5$) φαίνεται παρακάτω μαζί με τον αντίστοιχο έλεγχο καλής προσαρμογής:

```
> x=runif(1000,0,1)
> y=5*(-log(1-x))^(1/4)
> ks.test(y, "pweibull", 4,5)

        One-sample Kolmogorov-Smirnov test

data:  y
D = 0.0222, p-value = 0.7061
alternative hypothesis: two-sided
```

ΠΡΟΒΛΗΜΑ 3. Να δημιουργηθεί γεννήτρια τυχαίων αριθμών από την τυχαία μεταβλητή $X \sim Beta(2,4)$ με τη μέθοδο αποδοχής-απόρριψης, χρησιμοποιώντας μόνο τη γεννήτρια ομοιόμορφης (0,1) κατανομής.

Η συνάρτηση πυκνότητας πιθανότητας της κατανομής $Beta(a,b)$ είναι

$$f(x) = Cx^{a-1}(1-x)^{b-1} = Ch(x), \quad 0 \le x \le 1,$$

όπου η σταθερά $C = 1/(B(a,b)$, δίνεται μέσω του ολοκληρώματος Βήτα. Ένα πλεονέκτημα της μεθόδου αποδοχής-απόρριψης είναι ότι δε χρειάζεται την ακριβή τιμή της σταθεράς κανονικοποίησης. Κατ' αρχήν πρέπει να επιλέξουμε τη βοηθητική κατανομή $g(x)$ από την οποία θα γίνει η δειγματοληψία. Επειδή η X παίρνει τιμές στο διάστημα $[0,1]$, μια καλή επιλογή είναι η ομοιόμορφη στο $[0,1]$ για την οποία υπάρχει εύκολα υλοποιήσιμη γεννήτρια. Επομένως παίρνουμε

$$g(x) = 1, 0 \le x \le 1.$$

Η μέθοδος αποδοχής-απόρριψης λειτουργεί ως εξής: Έστω M μια σταθερά τέτοια ώστε

$$\frac{f(x)}{g(x)} \le M, \quad \forall 0 \le x \le 1.$$

Τότε ο αλγόριθμος προσομοίωσης είναι

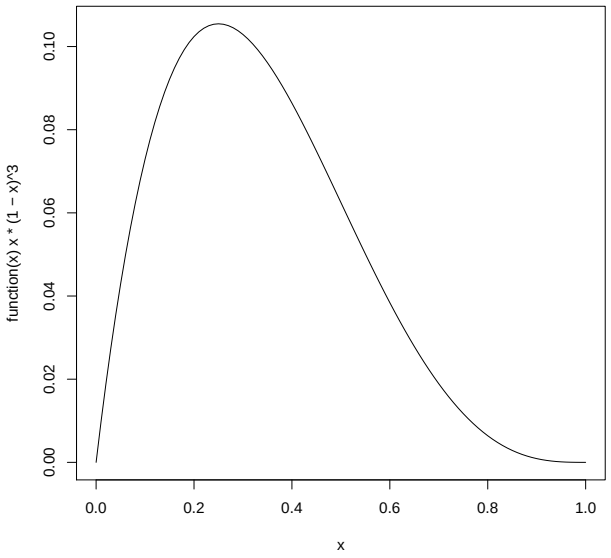
1. Δημιουργία X από την $g(x)$.
2. Δημιουργία $U \sim U(0,1)$.
3. Αν $U \le \frac{f(X)}{Mg(X)}$, η X γίνεται δεκτή - τέλος. Διαφορετικά επιστροφή στο βήμα 1.

Έστω $f(x) = Ch(x)$. Παρατηρούμε ότι η ανισότητα $\frac{f(x)}{g(x)} \leq M$ είναι ισοδύναμη με $\frac{h(x)}{g(x)} \leq MC$, ενώ η $U \leq \frac{f(X)}{Mg(X)}$ είναι ισοδύναμη με $U \leq \frac{h(X)}{MCg(X)}$. Επομένως η μέθοδος εφαρμόζεται ακριβώς με τον ίδιο τρόπο αν στη θέση της $f(x)$ χρησιμοποιήσουμε την αντίστοιχη $h(x)$ που δεν περιέχει τη σταθερά κανονικοποίησης, αν θεωρήσουμε ότι η νέα σταθερά είναι η MC.
Για να βρούμε τώρα τη σταθερά M τέτοια ώστε $\frac{h(x)}{g(x)} \leq M$, παίρνουμε τον μικρότερο αριθμό M που ικανοποιεί την ανισότητα για $0 \leq x \leq 1$, δηλαδή

$$M = \max_{0 \leq x \leq 1} \frac{h(x)}{g(x)} = \max_{0 \leq x \leq 1} \frac{x^1(1-x)^3}{1}.$$

Το γράφημα της συνάρτησης $x(1-x)^3$ φαίνεται στο Σχήμα 1

```
> plot(function(x) x*(1-x)^3, xlim=c(0,1))
```



Σχήμα 1: Διάγραμμα της $h(x) = x(1-x)^3, 0 \leq x \leq 1$.

Παρατηρούμε ότι η συνάρτηση έχει μέγιστο και αυτό μπορεί να υπολογιστεί με την εντολή

```
> optimize(function(x) x*(1-x)^3, lower = 0, upper = 1, maximum=T)
$maximum
[1] 0.2499993

$objective
[1] 0.1054687
```

Επομένως το μέγιστο προκύπτει για τιμή $x = 0.25$ και η μέγιστη τιμή είναι $M = 0.1055$. Αυτή θα χρησιμοποιήσουμε για τη δημιουργία της γεννήτριας.

Επομένως ο αλγόριθμος αποδοχής απόρριψης μπορεί να υλοποιηθεί με την παρακάτω συνάρτηση:

```
randbeta=function()
{
  x=runif(1);
  u=runif(1);
        while (u> x*(1-x)^3/0.1055)
  {
        x=runif(1);
        u=runif(1);
        }
  return(x);
}
```

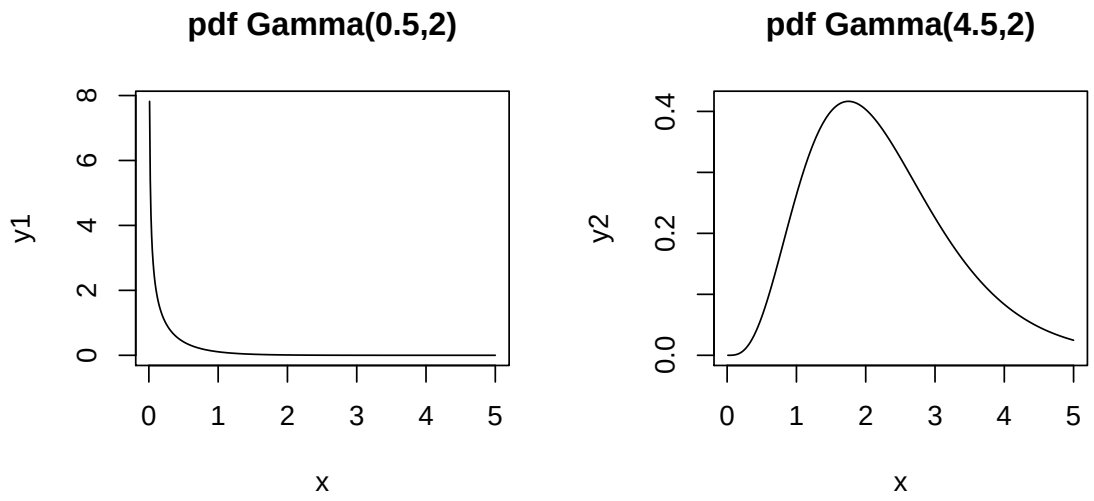
Δημιουργούμε ένα διάνυσμα 100 τιμών με αυτή τη συνάρτηση και ελέγχουμε για καλή προσαρμογή :

```
> N=100; x=rep(0,N); for (i in 1:N) { x[i]=randbeta() }

> ks.test(x, "pbeta", 2,4)

        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.0718, p-value = 0.6803
alternative hypothesis: two-sided
```

ΠΡΟΒΛΗΜΑ 4. Να δημιουργηθεί γεννήτρια τυχαίων αριθμών από την τυχαία μεταβλητή $X \sim Gamma(4,2)$, χρησιμοποιώντας μόνο τη γεννήτρια ομοιόμορφης (0,1) κατανομής.

Για την κατανομή Γάμμα $\Gamma(n,a)$ με n ακέραιο αριθμό γνωρίζουμε ότι μπορεί να γραφτεί σαν άθροισμα εκθετικών με παράμετρο a: Αν $X_1,\ldots,X_n$ ανεξάρτητες και ισόνομες τυχαίες μεταβλητές που ακολουθούν $Exp(a)$, τότε η $Y = X1 + \ldots + X_n$ ακολουθεί $Y \sim Gamma(n,a)$. Επομένως ένας εύκολος τρόπος να δημιουργήσουμε ένα τυχαίο αριθμό από αυτή την κατανομή είναι να δημιουργήσουμε n ανεξάρτητους τυχαίους αριθμούς από την εκθετική κατανομή με παράμετρο a και να πάρουμε το άθροισμά τους. Γνωρίζουμε ότι για να δημιουργήσουμε εκθετική κατανομή από την ομοιόμορφη $U \sim U(0,1)$ αρκεί να πάρουμε $X = -\log(1-U)/a$. Η παρακάτω συνάρτηση δημιουργεί μια τυχαία παρατήρηση από την $Gamma(n,a)$ με n ακέραιο.

```
randgammaexp=function(n,a)
{
  u=runif(n,min=0, max=1);
  x=-log(1-u)/a;
  y=sum(x);
  return(y);
}
```



Σχήμα 2: Πυκνότητα πιθανότητας της κατανομής Γαμμα(α,2) για $a = 0.5$) και $a = 4.5$.

Δημιουργούμε ένα διάνυσμα 100 τιμών με αυτή τη συνάρτηση για παραμέτρους (4,2) και ελέγχουμε για καλή προσαρμογή:

```
> N=100; x=rep(0,N); for (i in 1:N) { x[i]=randgammaexp(4,2)}
> ks.test(x, "pgamma", 4,2)

        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.0844, p-value = 0.4749
alternative hypothesis: two-sided
```

ΠΡΟΒΛΗΜΑ 5. Να δημιουργηθεί γεννήτρια τυχαίων αριθμών από την τυχαία μεταβλητή $X \sim Gamma(4.5, 2)$, χρησιμοποιώντας μόνο τη γεννήτρια ομοιόμορφης (0,1) κατανομής.

Για την κατανομή $Gamma(a, \lambda)$ με a όχι ακέραιο, δεν μπορούμε να εφαρμόσουμε τη μέθοδο του προηγούμενου παραδείγματος. Εδώ μπορούμε να εφαρμόσουμε την μέθοδο αποδοχής-απόρριψης. Παρατηρούμε πρώτα τη μορφή της συνάρτησης πυκνότητας πιθανότητας $f(x), x \geq 0$, για δύο περιπτώσεις της παραμέτρου $a = 0.5 < 1$ και $a = 4.5 > 1$, όπως φαίνεται στο Σχήμα 2.
Όπως μπορεί να αποδειχθεί και μαθηματικά, για $a < 1$ η συνάρτηση πυκνότητας πιθανότητας απειρίζεται ασυμπτωτικά για μικρές τιμές $x \to 0$, ενώ όταν $a >= 1$ είναι πάντα φραγμένη. Για την πρώτη περίπτωση, η διαδικασία εύρεσης κατάλληλης κατανομής g για την εφαρμογή της μεθόδου αποδοχής-απόρριψης είναι αρκετά πολύπλοκη και ξεφεύγει από τα πλαίσια του μαθήματος [1] Για το

[1] Οι ενδιαφερόμενοι μπορούν να δούν την εργασία Kundu, D. and Gupta, R. D. *A Convenient Way of Generating Gamma Random Variables Using Generalized Exponential Distribution*, URL:http://home.iitk.ac.in/ kundu/paper120.pdf

λόγο αυτό εδώ θα περιοριστούμε στην περίπτωση $a > 1$ και θα δημιουργήσουμε μια γεννήτρια που βασίζεται στη μέθοδο αποδοχής απόρριψης.
Η συνάρτηση πυκνότητας πιθανότητας της κατανομής $Gamma(a, \lambda)$ είναι

$$f(x) = Cx^{a-1}e^{-\lambda x} = Ch(x),$$

επομένως για $a = 4.5, \lambda = 2$,

$$h(x) = x^{3.5}e^{-2x}.$$

Η ακριβής τιμή της σταθεράς C δε θα μας απασχολήσει γιατί όπως έχουμε δει (βλ. Παράδειγμα 2) ο αλγόριθμος αποδοχής-απόρριψης δεν απαιτεί την τιμή αυτή. Η βοηθητική κατανομή που θα επιλέξουμε πρέπει να παίρνει τιμές στο διάστημα $[0, \infty)$, και να έχει εύκολη γεννήτρια τυχαίων αριθμών. Μια προφανής επιλογή είναι η εκθετική κατανομή, γενικά με παράμετρο μ. Το ερώτημα όμως είναι με ποια είναι η κατάλληλη τιμή της παραμέτρου μ. Παρατηρούμε ότι αν π.χ. πάρουμε ως g την εκθετική κατανομή με παράμετρο $\mu > 2$, π.χ. $\mu = 4$, δηλαδή θέσουμε $g(x) = 4e^{-4x}$, τότε ο λόγος $\frac{h(x)}{g(x)} = \frac{x^{3.5}e^{2x}}{4}$, και η ποσότητα αυτή δεν έχει μέγιστο στο διάστημα $[0, \infty)$, επειδή απειρίζεται για $x \to \infty$. Πρέπει επομένως να επιλέξουμε μια εκθετική κατανομή με παράμετρο τέτοια ώστε το παραπάνω κλάσμα να μην απειρίζεται. Για να συμβεί αυτό πρέπει να θέσουμε $\mu < 2$, έστω π.χ. $\mu = 1$, δηλαδή $g(x) = e^{-x}$. Τότε

$$\frac{h(x)}{g(x)} = x^{3.5}e^{-x}.$$

Για να εφαρμόσουμε τον αλγόριθμο πρέπει να υπολογίσουμε τη σταθερά

$$M = \max_{0 \le x < \infty} \frac{h(x)}{g(x)} = \max_{0 \le x \le 1} x^{3.5}e^{-x}.$$

Το γράφημα της συνάρτησης $x^{3.5}e^{-x}$ φαίνεται στο Σχήμα 3.

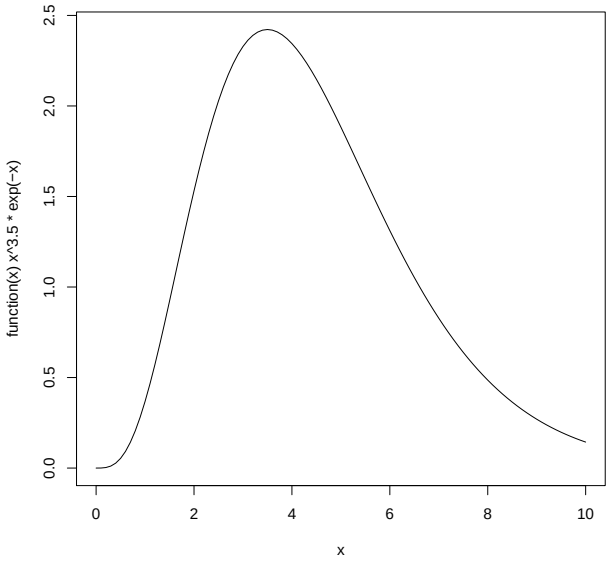
```
> plot(function(x) x^3.5*exp(-x), xlim=c(0,10))
```

Παρατηρούμε ότι η συνάρτηση μεγιστοποιείται μέσα στο διάστημα $[0, 10]$ και επομένως μπορούμε να χρησιμοποιήσουμε το διάστημα αυτό στην εντολή optimize του R, η οποία λειτουργεί μόνο με φραγμένα διαστήματα.

```
> optimize(function(x) x^3.5*exp(-x), interval=c(0,10), maximum=T )
$maximum
[1] 3.499999

$objective
[1] 2.422186
```

Βλέπουμε ότι η μέγιστη τιμή είναι $M = 2.4222$. Με βάση τα παραπάνω ο αλγόριθμος για τη δημιουργία της Γαμμα(4.5,2) με τη μέθοδο αποδοχής απόρριψης είναι:

1. Δημιουργία X από την $g(x) = e^{-x}$, δηλαδή
   - Δημιουργία $U_1 \sim U(0, 1)$



Σχήμα 3: Γράφημα της συνάρτησης $x^{3.5}e^{-x}$

- $X = -\log(1 - U_1)$.

2. Δημιουργία $U_2 \sim U(0, 1)$.
3. Αν $U_2 \leq x^{3.5}e^{-x}/2.4222$, η X γίνεται δεκτή - τέλος. Διαφορετικά επιστροφή στο βήμα 1.

Ο αλγόριθμος μπορεί να υλοποιηθεί με την παρακάτω συνάρτηση:

```
randgammaaccreject=function()
{
  u1=runif(1);
        x=-log(1-u1);
  u2=runif(1);
        while (u2> x^3.5 *exp(-x)/2.4222)
  {
  u1=runif(1);
        x=-log(1-u1);
  u2=runif(1);
        }
  return(x);
}
```

Δημιουργούμε ένα διάνυσμα 100 τιμών με αυτή τη συνάρτηση για παραμέτρους (4,2) και ελέγχουμε για καλή προσαρμογή:

```
> N=100; x=rep(0,N); for (i in 1:N) x[i]=randgammaaccreject()
> ks.test(x, "pgamma", 4.5,2)
```

```
        One-sample Kolmogorov-Smirnov test

data:  x
D = 0.0876, p-value = 0.4262
alternative hypothesis: two-sided
```

ΠΡΟΒΛΗΜΑ 6. Να δημιουργηθεί γεννήτρια τυχαίων αριθμών από τη διακριτή τυχαία μεταβλητή X που παίρνει τιμές στο σύνολο $A = \{a_1, \ldots, a_n\}$ με πιθανότητες $P(X = a_i) = p_i, i = 1, \ldots, n$.

Ο πιο συνηθισμένος και εύκολος τρόπος για τη δημιουργία τυχαίων αριθμών από μια αυθαίρετη διακριτή κατανομή είναι η δημιουργία $U \sim U(0,1)$ και

$$X = \begin{cases} a_1, & \text{αν} \quad U \leq p_1 \\ a_2, & \text{αν} \quad p_1 < U \leq p_1 + p_2 \\ \vdots & \\ a_n, & \text{αν} \quad p_1 + \ldots + p_{n-1} < U \leq 1 \end{cases}$$

Ο αλγόριθμος αυτός υλοποιείται με την παρακάτω συνάρτηση, που παίρνει ως ορίσματα το διάνυσμα τιμών $(a_1, \ldots, a_n)$ και το διάνυσμα πιθανοτήτων $(p_1, \ldots, p_n)$.

```
randdiscrete=function(a,p)
{
  u=runif(1);
  i=0;
  s=0;
  while(u>s)
  {
    i=i+1;
    s=s+p[i];
  }
  return(a[i]);
}
```

Δημιουργούμε 1000 παρατηρήσεις από τη διακριτή κατανομή με $a = (1, 2, 3, 4)$ και $p = (0.2, 0.3, 0.2, 0.3)$ και υπολογίζουμε τις σχετικές συχνότητες κάθε τιμής στο δείγμα. Επίσης κάνουμε έλεγχο χ^2 καλής προσαρμογής του εμπειρικού διανύσματος συχνοτήτων με το θεωρητικό διάνυσμα p.

```
> a=c(1,2,3,4); p=c(0.2,0.3,0.2,0.3)
> N=1000; x=rep(0,N); for (i in 1:N) x[i]=randdiscrete(a,p)
> xt=table(x)
> xt
x
  1   2   3   4
181 296 226 297
> xfreq=xt/N
> xfreq
```

```
x
    1     2     3     4
0.181 0.296 0.226 0.297
> chisq.test(xfreq, p=p)

        Chi-squared test for given probabilities

data:  xfreq
X-squared = 0.0053, df = 3, p-value = 0.9999
```

Η μηδενική υπόθεση δεν απορρίπτεται, επομένως δεν υπάρχει ένδειξη διαφοράς από την κατανομή p.

ΠΡΟΒΛΗΜΑ 7. Να δημιουργηθεί γεννήτρια τυχαίων αριθμών από τη διακριτή ομοιόμορφη κατανομή στο σύνολο $A = \{a_1, \ldots, a_n\}$, δηλαδή από την τυχαία μεταβλητή X που παίρνει τιμές στο σύνολο A με ίσες πιθανότητες $p_i = 1/n, i = 1, \ldots, N$.

Θα μπορούσαμε να χρησιμοποιήσουμε τη γενική συνάρτηση του προηγούμενου παραδείγματος. Όμως στην ειδική περίπτωση που όλες οι τιμές της X είναι ισοπίθανες μπορούμε να δημιουργήσουμε μια πιο εύκολη και γρήγορη γεννήτρια ως εξής: Έστω $U \sim U(0,1)$ και $I = 1 + \lfloor nU \rfloor$, όπου $\lfloor z \rfloor$ είναι το ακέραιο μέρος του αριθμού z, δηλαδή ο μεγαλύτερος ακέραιος μικρότερος ή ίσος του z (π.χ. $\lfloor 9.8 \rfloor = 9, \lfloor 2 \rfloor = 2$).
Στον παραπάνω τύπο παρατηρούμε ότι $0 < U < 1$, επομένως $0 < nU < n$, και η τυχαία μεταβλητή $\lfloor nU \rfloor$ μπορεί να πάρει τις τιμές $0, 1, \ldots, n-1$. Επίσης η πιθανότητα π.χ. $P(\lfloor nU \rfloor) = 0) = P(0 \leq nU < 1) = P(0 \leq U \leq \frac{1}{n}) = \frac{1}{n}$), και το ίδιο εύκολα δείχνεται για όλες τις υπόλοιπες τιμές. Επομένως η τυχαία μεταβλητή $I = 1 + \lfloor nU \rfloor$ είναι διακριτή ομοιόμορφη με τιμές $1, 2, \ldots, n$. Για να πάμε από την I στη X, αρκεί να δημιουργήσουμε τη I όπως είδαμε, και να θέσουμε $X = a_I$, δηλαδή να πάρουμε ως X το I-οστό στοιχείο του διανύσματος τιμών της X.
Στο R το ακέραιο μέρος υπολογίζεται με τη συνάρτηση floor. Ο αλγόριθμος υλοποιείται στην παρακάτω συνάρτηση:

```
randdiscunif=function(a)
{
  n=length(a);
  u=runif(1);
  i=1+floor(n*u);
  x=a[i];
  return(x);
}
```

Επίσης επειδή στο R ένα διάνυσμα μπορεί να έχει συνιστώσες μη αριθμητικά στοιχεία, μπορούμε εύκολα να χρησιμοποιήσουμε αυτή και την προηγούμενη συνάρτηση για να δημιουργήσουμε γεννήτριες από κατηγορικές μεταβλητές. Για παράδειγμα αν η X είναι μια κατηγορική μεταβλητή χρώματος με επίπεδα blue green orange yellow, όλα ισοπίθανα, μπορούμε να δημιουργήσουμε παρατηρήσεις από αυτήν όπως φαίνεται παρακάτω:

```
> a=c("blue", "green", "orange", "yellow")
> a
[1] "blue"   "green"  "orange" "yellow"
> N=1000; x=rep(0,N); for (i in 1:N) x[i]=randdiscunif(a)
> xt=table(x)
> xfreq=xt/N
> xfreq
x
  blue  green orange yellow
 0.235  0.261  0.259  0.245
> chisq.test(xfreq,p=rep(1/4,4))

        Chi-squared test for given probabilities

data:  xfreq
X-squared = 0.0018, df = 3, p-value = 1
```

2 Προσομοίωση Monte Carlo

ΠΡΟΒΛΗΜΑ 8. Έστω $\theta = E(X^2 \log(1+X))$, όπου $X \sim Exp(4)$. Να βρεθεί μια εκτίμηση του θ βασισμένη σε 1000 σενάρια προσομοίωσης, και να υπολογιστεί το αντίστοιχο διάστημα εμπιστοσύνης 95%.

Έστω $Y = X^2 \log(1+X))$, επομένως $\theta = E(Y)$. Δημιουργούμε $N = 1000$ ψευδοτυχαίες παρατηρήσεις $Y_1, \ldots, Y_N$ από την κατανομή της Y. Αυτό μπορεί να γίνει εύκολα, παίρνοντας $X_j \sim Exp(4)$ με την εντολή rexp και θέτοντας $Y_j = X_j^2 \log(1+X_j)$. Έχοντας το τυχαίο (στην πραγματικότητα ψευδοτυχαίο) δείγμα από την κατανομή της Y, μπορούμε να υπολογίσουμε το δειγματικό μέσο $\bar{Y}_N$, τη δειγματική τυπική απόκλιση s_Y και το διάστημα εμπιστοσύνης

$$\bar{Y}_N \pm t_{N-1,0.025} \frac{s_Y}{\sqrt{N}}.$$

Τα παραπάνω υλοποιούνται ως εξής:

```
> N=1000
> x=rexp(N,4)
> y=x^2*log(1+x)
> ym=mean(y)
> ym
[1] 0.06676881
> halflength=qt(0.975, N-1)*sy/sqrt(N)
> halflength
[1] 0.01495501
> halflength=qt(0.975, N-1)*sy/sqrt(N)
> halflength
```

```
[1] 0.01495501
> confint=c(ym-halflength, ym+halflength)
> confint
[1] 0.05181380 0.08172382
```

Βλέπουμε επομένως ότι η εκτίμηση του θ είναι ίση με 0.0668 και το 95% διάστημα εμπιστοσύνης $[0.0518, 0.0817]$. Το μισό μήκος του διαστήματος είναι ίσο με 0.0150, το οποίο αν συγκριθεί με την εκτίμηση είναι $0.0150/0.0667 = 0.223$ δηλαδή περίπου 22% της μέσης τιμής. Σίγουρα δεν πρόκειται για πολύ ακριβή εκτίμηση, αλλά αυτό μπορεί να βελτιωθεί παίρνοντας μεγαλύτερο μέγεθος δείγματος.

ΠΡΟΒΛΗΜΑ 9. Έστω ότι ρίχνουμε 10 δίκαια ζάρια, τα αποτελέσματά τους είναι $Y_1, \ldots, Y_n$ και θέτουμε $S_1 = Y_1 + \ldots + Y_n, S_2 = Y_1^2 + \ldots + Y_n^2, X = \frac{S_2}{S_1}$. Θέλουμε να εκτιμήσουμε την πιθανότητα $p = P(X > 4)$.
Να σχεδιαστεί ένα πείραμα προσομοίωσης Monte Carlo για την εκτίμηση του p όταν $n = 10$, με ένα διάστημα εμπιστοσύνης 95% το οποίο να έχει μισό μήκος το πολύ 0.01, έτσι ώστε το p να εκτιμηθεί με ακρίβεια ± 0.01.

Κατ' αρχήν δημιουργούμε μια συνάρτηση που προσομοιώνει ένα σενάριο του πειράματος, δηλαδή τη ρίψη n ζαριών (το n μετά θα τεθεί ίσο με 10), και τον υπολογισμό της τιμής της X από αυτές τις ρίψεις. Για την προσομοίωση κάθε ζαριού μπορούμε να χρησιμοποιήσουμε τη συνάρτηση randdiscrunif που δημιουργήσαμε σε προηγούμενο παράδειγμα. Η συνάρτηση που μας ενδιαφέρει εδώ και προσομοιώνει ένα σενάριο φαίνεται παρακάτω. Έχει ως όρισμα τον αριθμό ζαριών n που ρίχνονται.

```
dicethrow=function(n)
{
  Y=rep(0,n)
  for (i in 1:n)
      Y[i]=randdiscunif(1:6);
  S1=sum(Y);
  S2=sum(Y^2);
  X=S2/S1;
}
```

Για την εκτίμηση του $p = P(X > 4)$ από συγκεκριμένο αριθμό σεναρίων N μπορούμε να δημιουργήσουμε το διάνυσμα παρατηρήσεων $X_1, \ldots, X_N$ και από αυτές το διάνυσμα δυαδικών μεταβλητών $R_1, \ldots, R_N$, όπου $R_j = 1(X_j > 4)$. Γνωρίζουμε ότι οι $R_1, \ldots, R_N$ είναι α.ι.τμ. με κατανομή Bernoulli(p), επομένως $p = E(R)$, άρα το p μπορεί να εκτιμηθεί με τη μεθοδολογία του προηγούμενου προβλήματος. Για παράδειγμα, αν $n = 10, N = 1000$, έχουμε:

```
> N=1000; n=10; R=rep(0,N);
> for (i in 1:N)
+ {
+ X=dicethrow(n);
+ R[i]=(X>4);
```

```
+ }
> mR=mean(R);
> sR=sd(R);
> sR;
[1] 0.4477404
> hl=qt(0.975,N-1)*sR/sqrt(N);
> hl;
[1] 0.02778439
> confint=c(mR-hl, mR+hl);
> confint;
[1] 0.6952156 0.7507844
```

Βλέπουμε ότι για $N = 1000$ η ακρίβεια της εκτίμησης είναι μικρότερη από την επιθυμητή, γιατί το μισό μήκος του Δ.Ε. είναι $hl = 0.0289 > 0.01$. Για να πετύχουμε επιθυμητή ακρίβεια, μπορούμε να σκεφτούμε ως εξής: Από τα προηγούμενα αποτελέσματα έχουμε βρεί ότι η δειγματική τυπική απόκλιση της R είναι $s_R = 0.4478$. Προσεγγιστικά θεωρούμε ότι αυτή είναι η πραγματική τιμή της τυπικής απόκλισης σ_R. Επίσης, επειδή το τελικό μέγεθος δείγματος θα είναι πάνω από 1000, η κρίσιμη τιμή της t που χρησιμοποιείται στον τύπου του ΔΕ $(1-\alpha)100\%$ είναι προσεγγιστικά ίση με $t_{\alpha/2,N-1} \approx z_{1-\alpha/2}$, δηλαδή το άνω $1-\alpha/2$ ποσοστημόριο της κανονικής κατανομής. Επομένως ο τύπος για το μισό μήκος του ΔΕ γίνεται με πολύ καλή προσέγγιση:

$$hl = z_{1-\alpha/2}\frac{s_R}{sqrt(N)}.$$

Για να πετύχουμε το επιθυμητό μήκος hl, λύνουμε την παραπάνω έκφραση ως προς N και βρίσκουμε:

$$N = \frac{z^2_{1-\alpha/2}s^2_R}{hl^2}.$$

Επομένως συνεχίζουμε το προηγούμενο πρόγραμμα για να βρούμε το απαιτούμενο μέγεθος δείγματος, θέτοντας $\alpha = 0.05, hl = 0.01$:

```
> N=qnorm(0.975)^2*sR^2/0.01^2
> N
[1] 7701.029
```

Βλέπουμε ότι μια καλή προσέγγιση για το απαιτούμενο μέγεθος δείγματος είναι $N \approx 7700$ σενάρια. Επαναλαμβάνουμε λοιπόν την προηγούμενη διαδικασία εκτίμησης με $N = 7700$:

```
> N=7700; n=10; R=rep(0,N);
> for (i in 1:N)
+ {
+ X=dicethrow(n);
+ R[i]=(X>4);
+ }
> mR=mean(R);
> mR
[1] 0.7245455
```

```
> sR=sd(R);
> sR;
[1] 0.446772
> hl=qt(0.975,N-1)*sR/sqrt(N);
> hl;
[1] 0.009980608
> confint=c(mR-hl, mR+hl);
> confint;
[1] 0.7145648 0.7345261
```

Τώρα το μισό μήκος είναι $hl = 0.01$ όπως απαιτείται. Επομένως μια εκτίμηση για τη ζητούμενη πιθανότητα είναι $\hat{p} = 0.7245$ και το διάστημα εμπιστοσύνης 95% είναι $[0.7146, 0.7345]$.

ΠΡΟΒΛΗΜΑ 10. Έστω $X \sim N(\mu, \sigma^2)$ με άγνωστη μέση τιμή και διασπορά. Συγκεντρώνεται ένα δείγμα μεγέθους n από αυτή την κατανομή και γίνεται ο αμφίπλευρος έλεγχος υπόθεσης

$$H_0 : \mu = 0, \quad H_1 : \mu \neq 0$$

με επίπεδο σημαντικότητας α. Έστω ότι χρησιμοποιούμε τον έλεγχο t, σύμφωνα με τον οποίον υπολογίζουμε το δειγματικό μέσο $\bar{X}_n$, τη δειγματική τυπική απόκλιση s_n, το στατιστικό ελέγχου $t = \frac{\bar{X}_n}{s_n/\sqrt{n}}$ και με βάση τα παραπάνω ο κανόνας απόφασης είναι

$$H_0 \text{ απορρίπτεται αν και μόνο αν } |t| > t_{\alpha_2, n-1}.$$

Η ισχύς p του ελέγχου, η οποία είναι η πιθανότητα απόρριψης της H_0, δεδομένου ότι δεν ισχύει η H_0, είναι συνάρτηση των n, μ, σ, α, δηλαδή

$$p(\mu, \sigma, n, \alpha) = P(\text{απορρ. } H_0 | \mu, \sigma),$$

για τιμές $\mu \neq 0$ όπου ο έλεγχος πραγματοποιείται σε επίπεδο σημαντικότητας α.
Ζητείται να εκτιμηθεί η ισχύς του ελέγχου μέσω προσομοίωσης Monte Carlo και να γίνουν τα γραφήματα της μεταβολής του p ως προς $\mu, -2 \leq \mu \leq 2$, για τις εξής περιπτώσεις: (α) $\sigma = 1, n = 10$, (β) $\sigma = 1, n = 100$, (γ) $\sigma = 5, n = 10$, (δ) $\sigma = 5, n = 100$.

Στο πρόβλημα αυτό ένα σενάριο προσομοίωσης είναι πολύ γενικότερο από ότι στα δύο προηγούμενα προβλήματα. Εδώ, για συγκεκριμένες τιμές των n, μ, σ, α, προσομοιώνουμε την ίδια τη λειτουργία του ελέγχου t σε τυχαία δείγματα μεγέθους n που προκύπτει από την κανονική κατανομή $N(\mu, \sigma^2)$, στο επίπεδο σημαντικότητας α. Κάθε πραγματοποίηση του ελέγχου σε διαφορετικό δείγμα αποτελεί ένα σενάριο του οποίου το αποτέλεσμα είναι μια τυχαία μεταβλητή $Z \in \{0, 1\}$ που παίρνει τιμή 1 αν στο συγκεκριμένο σενάριο απορριφθεί η μηδενική υπόθεση και 0 διαφορετικά. Επομένως αν το σενάριο υλοποιηθεί με παραμέτρους (n, μ, σ, α), ισχύει ότι

$$p(\mu, \sigma, n, \alpha) = E(Z),$$

δηλαδή η ισχύς μπορεί να εκτιμηθεί μέσω προσομοίωσης N σεναρίων, από τη στατιστική ανάλυση του 'υπερδείγματος' $Z_1, \ldots, Z_N$, ακριβώς όπως και στα προηγούμενα παραδείγματα.

Η παρακάτω συνάρτηση υλοποιεί την προσομοίωση ενός σεναρίου, δηλαδή δημιουργεί δείγμα μεγέθους n από την κατάλληλη κατανομή, υπολογίζει τα στατιστικά του ελέγχου πάνω στο δείγμα και αποφασίζει αν απορρίπτεται ή όχι η μηδενική υπόθεση $H_0 : \mu = 0$, δίνοντας ως αποτέλεσμα τη δυαδική μεταβλητή Z.

```
ttestpower=function(m,s,n,a)
{
      ta=qt(1-a/2,n-1);
      x=rnorm(n,m,s);
      mx=mean(x);
      sx=sd(x);
      t=mx/(sx/sqrt(n));
      if (abs(t)>ta) Z=1 else Z=0;
      return(Z);
}
```

Μπορούμε τώρα να χρησιμοποιήσουμε τη συνάρτηση αυτή για να αξιολογήσουμε την ισχύ του ελέγχου κάτω από διάφορες υποθέσεις. Ως πρώτο παράδειγμα, θα εκτιμήσουμε την ποσότητα $p(1, 1, 10, 0.05)$, δηλαδή την ισχύ του ελέγχου για $n = 20, \alpha = 0.05$, όταν οι πραγματικές τιμές των άγνωστων παραμέτρων είναι $\mu = 1, \sigma = 1$. Θα χρησιμοποιήσουμε εκτίμηση Monte Carlo με 1000 σενάρια. Αυτό μπορεί να γίνει με τον παρακάτω κώδικα:

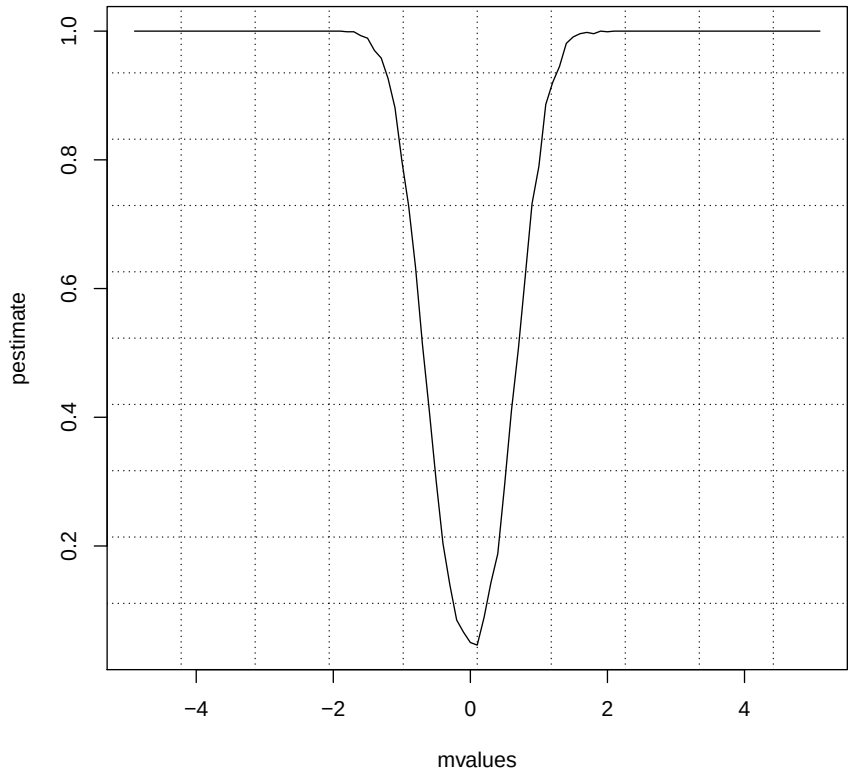
```
> m=1;s=1;n=20;a=0.05;
> N=1000; z=rep(0,N); for (i in 1:N) z[i]=ttestpower(m,s,n,a)
> mz=mean(z);
> mz;
[1] 0.989
> sz=sd(z);
> sz;
[1] 0.1043546
> hl=qt(1-a/2,n-1)*sz/sqrt(n);
> confint=c(mz-hl, mz+hl);
> confint
[1] 0.9401605 1.0378395
```

Η εκτίμηση της ισχύος είναι $\hat{p} = 0.989$, δηλαδή αν το δείγμα μεγέθους 20 προέρχεται από κανονική (1,1) κατανομή και ο έλεγχος γίνει στο επίπεδο σημαντικότητας 5%, τότε η πιθανότητα να απορριφθεί η μηδενική κατανομή είναι περίπου 99%, δηλαδή περίπου στο 99% των δειγμάτων θα προκύπτει στατιστικά σημαντική ένδειξη ότι η μέση τιμή είναι διάφορη του μηδενός.
Στο δεύτερο παράδειγμα θα δημιουργήσουμε την καμπύλη της ισχύος συναρτήσει της μέσης τιμής μ, για $n = 10, \sigma = 1, \alpha = 0.05$, όταν η μέση τιμή μεταβάλλεται στο διάστημα $(-5, 5)$. Για το σκοπό αυτό θα επαναλάβουμε την προηγούμενη διαδικασία για διαφορετικές τιμές του μ στο παραπάνω διάστημα (με βήμα 0.1), θα συγκεντρώσουμε τις εκτιμήσεις $\hat{p}(\mu)$ και θα σχηματίσουμε το διάγραμμα της συνάρτησης ισχύος.

```
n=10;s=1;a=0.05;
mmin=-5; mmax=5; delta=0.1; nmvalues=1+floor((mmax-mmin)/delta);
```

```
pestimate=rep(0,nmvalues);
mvalues=rep(0,nmvalues);
for (j in (1:nmvalues))
{
  m=mmin+j*delta;
  N=1000; z=rep(0,N); for (i in 1:N) {z[i]=ttestpower(m,s,n,a)};
  mvalues[j]=m;
  pestimate[j]=mean(z);
}
plot(pestimate~mvalues, type="l");
grid(nx=10, ny=10);
```

Το γράφημα που παράγεται φαίνεται στο Σχήμα 4



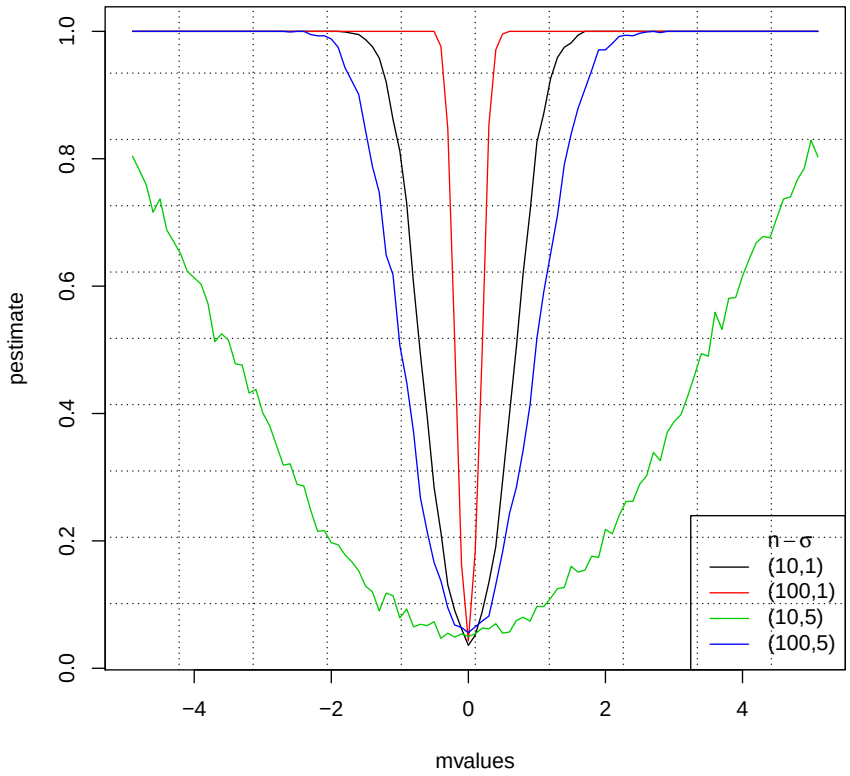
Σχήμα 4: Συνάρτηση ισχύος ως συνάρτηση του μ, για $\sigma = 1, n = 10, \alpha = 0.05$

Από το σχήμα παρατηρούμε τα εξής: Πρώτα για $\mu = 0$, η πιθανότητα απόρριψης είναι περίπου 0.05. Αυτό φυσικά είναι αναμενόμενο, επειδή όταν $\mu = 0$, η πιθανότητα απόρριψης ταυτίζεται με την πιθανότητα σφάλματος τύπου 1, δηλαδή το $\alpha = 0.05$. Για τιμές του $\mu \neq 0$, ιδανικά θα θέλαμε ο έλεγχος να απορρίπτει πάντα τη μηδενική υπόθεση, δηλαδή η ισχύς να είναι όσο το δυνατό πιο κοντά στο 1. Αυτό συμβαίνει εδώ για τιμές αρκετά μακριά από το μηδέν (περίπου

$|\mu| > 2$), ενώ για τιμές κοντά στο μηδέν η πιθανότητα απόρριψης πέφτει γρήγορα. Αυτό σημαίνει ότι με ένα δείγμα μεγέθους 10, και τη συγκεκριμένη διασπορά $\sigma^2 = 1$, ο έλεγχος δεν έχει καλή δυνατότητα να διακρίνει τη διαφορά από το μηδέν όταν η μέση τιμή είναι $|\mu| < 2$. Για να βελτιωθεί η συμπεριφορά του ελέγχου θα πρέπει να αυξηθεί το μέγεθος δείγματος. Στο τελευταίο παράδειγμα εφαρμογής αυτής της προσομοίωσης θα απαντήσουμε στο αρχικό ερώτημα της άσκησης, δηλαδή θα δημιουργήσουμε τα γραφήματα της ισχύος ως προς μ για κάθε μια από τις περιπτώσεις (α)-(δ). Για να το κάνουμε αυτό θα επαναλάβουμε τον προηγούμενο κώδικα δημιουργίας του διαγράμματος 4 φορές, μια για κάθε περίπτωση τιμών των n, σ, και θα προσθέσουμε όλα τα διαγράμματα στο ίδιο σχήμα για να γίνει η σύγκριση.

```
a=0.05;
ncases=4;
sval=c(1,1,5,5);
nval=c(10,100,10,100);
for (icase in 1:ncases)
{
    n=nval[icase];
    s=sval[icase];
    mmin=-5; mmax=5; delta=0.1; nmvalues=1+floor((mmax-mmin)/delta);
    pestimate=rep(0,nmvalues);
    mvalues=rep(0,nmvalues);
    for (j in (1:nmvalues))
    {
  m=mmin+j*delta;
  N=1000; z=rep(0,N); for (i in 1:N) {z[i]=ttestpower(m,s,n,a)};
  mvalues[j]=m;
  pestimate[j]=mean(z);
    }
    if (icase==1)
    {
  plot(pestimate~mvalues, type="l", col=1);
       grid(nx=10, ny=10, col="black");
    }
    else
    {
  lines(pestimate~mvalues, type="l", col=icase);
    }
}
legend("bottomright", c(expression(n-sigma), "(10,1)", "(100,1)", "
   (10,5)", "(100,5)"),col=0:4, lty=c(0,1,1,1,1))
```

Τα γραφήματα φαίνονται στο Σχήμα 5



Σχήμα 5: Συνάρτηση ισχύος ως συνάρτηση του μ, για $\alpha = 0.05$, $(n, \sigma) = (10, 1), (100, 1), (10, 5), (100, 5)$