Ανάλυση Δεδομένων με χρήση του Στατιστικού Πακέτου R

Δημήτρης Φουσκάκης,
Επίκουρος Καθηγητής,
Τομέας Μαθηματικών,
Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών,
Εθνικό Μετσόβιο Πολυτεχνείο.

Περιεχόμενα

- Εισαγωγή στη Στατιστική
- Εισαγωγή στο Στατιστικό Πακέτο R
- Περιγραφική Στατιστική
- Προσομοίωση
- Στατιστική Συμπερασματολογία
  - Ένα Δείγμα
  - Δύο Ανεξάρτητα Δείγματα
  - Δείγματα κατά Ζεύγη
  - Ποσοστά
  - Έλεγχος καλής προσαρμογής
  - Πίνακες Συνάφειας 2×2
- Ανάλυση Παλινδρόμησης
- Ανάλυση Διασποράς

Κατανομές στην R

- Στην R υπάρχουν πολλές συναρτήσεις που σχετίζονται με γνωστές κατανομές και υπολογισμό ποσοτήτων από αυτές. Κάθε συνάρτηση έχει ένα όνομα που αρχίζει με ένα από τα ακόλουθα γράμματα, τα οποία καθορίζουν το είδος της συνάρτησης:
  - r: Γεννήτρια τυχαίων αριθμών.
  - p: Συνάρτηση Κατανομής Πιθανότητας (σ.κ.π.) F(x).
  - d: Συνάρτηση Πυκνότητας Πιθανότητας (σ.π.π.) ή Συνάρτηση Μάζας Πιθανότητας (σ.μ.π.), f(x).
  - q: Υπολογισμός Ποσοστιαίων σημείων ή ισοδύναμα αντίστροφη Συνάρτηση Κατανομής Πιθανότητας $F^{-1}(x)$ (δηλαδή το σημείο x: $P(X \leq x) > q$ για καθορισμένο q).

Κατανομές στην R

Εντολή	Κατανομή	Εντολή	Κατανομή
beta	Βήτα	hyper	Υπεργεωμετρική
norm	Κανονική	unif	Ομοιόμορφη
pois	Poisson	cauchy	Cauchy
nbinom	Αρνητική Διωνυμική	weibull	Weibull
gamma	Γάμμα	chisq	X^2
t	Student	exp	Εκθετική
binom	Διωνυμική	geom	Γεωμετρική
f	Snedecor	mvnorm	Πολυμεταβλητή Κανονική

- Με την βοήθεια του help μπορείτε να δείτε τι παραμέτρους παίρνουν οι εν λόγω συναρτήσεις, π.χ. > help("dnorm")

Κατανομές στην R

- Παραδείγματα:

```
> pnorm(3,2,2)
[1] 0.6914625
```

Υπολογίζει την σ.κ.π. της Κανονικής κατανομής με μέσο 2 και τυπική απόκλιση (ΟΧΙ διασπορά) 2 στο σημείο 3.

```
> qgamma(0.3,1,1)
[1] 0.3566749
```

Βρίσκει το 0.3 ποσοστιαίο σημείο της Γάμμα κατανομής με παραμέτρους 1 και 1.

```
> dt(2,3)
[1] 0.06750966
```

Υπολογίζει την σ.π.π. της Student κατανομής με 3 βαθμούς ελευθερίας στο σημείο 2.

```
> runif(5,-2,2)
[1]  1.3448055 -0.4691324  1.2517269  1.5576504  0.9563447
```

Δημιουργεί 5 τυχαίους αριθμούς από την ομοιόμορφη στο (-2,2).

Κατανομές στην R

- Τα ορίσματα των συναρτήσεων μπορεί να είναι και διανύσματα, π.χ.

```
> dexp(1:5,2)
[1] 2.706706e-01 3.663128e-02 4.957504e-03 6.709253e-04
    9.079986e-05
```

Υπολογίζει την σ.π.π. της Εκθετικής κατανομής με παράμετρο 2 στα σημεία 1,2,3,4 και 5.

- Υπάρχουν για πολλές κατανομές προκαθορισμένες τιμές στις παραμέτρους, π.χ. η εντολή rnorm(50) (λείπουν οι τιμές για τις 2 παραμέτρους) γεννάει 50 τιμές από την Κανονική κατανομή με μέσο 0 και τυπική απόκλιση 1 (0 και 1 αντίστοιχα οι προκαθορισμένες τιμές).

Κατανομές στην R

- Μπορούμε να βρούμε και την 1-F. Π.χ. αν X~Student(10) τότε για να βρούμε την P(X≤2) πληκτρολογούμε

```
> pt(2,10)
[1] 0.963306
```

ενώ για να βρούμε την P(X>2) πληκτρολογούμε

```
> pt(2,10, lower.tail=FALSE)
[1] 0.03669402
```

Κατανομές στην R

- Υπάρχει η δυνατότητα οι προηγούμενες συναρτήσεις (με αρχικά p και d) να είναι σε λογαριθμική κλίμακα

```
> pnorm(3,2,2)
[1] 0.6914625
> pnorm(3,2,2, log=T)
[1] -0.3689464
> dt(2,3)
[1] 0.06750966
> dt(2,3, log=T)
[1] -2.695485
```

Γραφικές Παραστάσεις Κατανομών

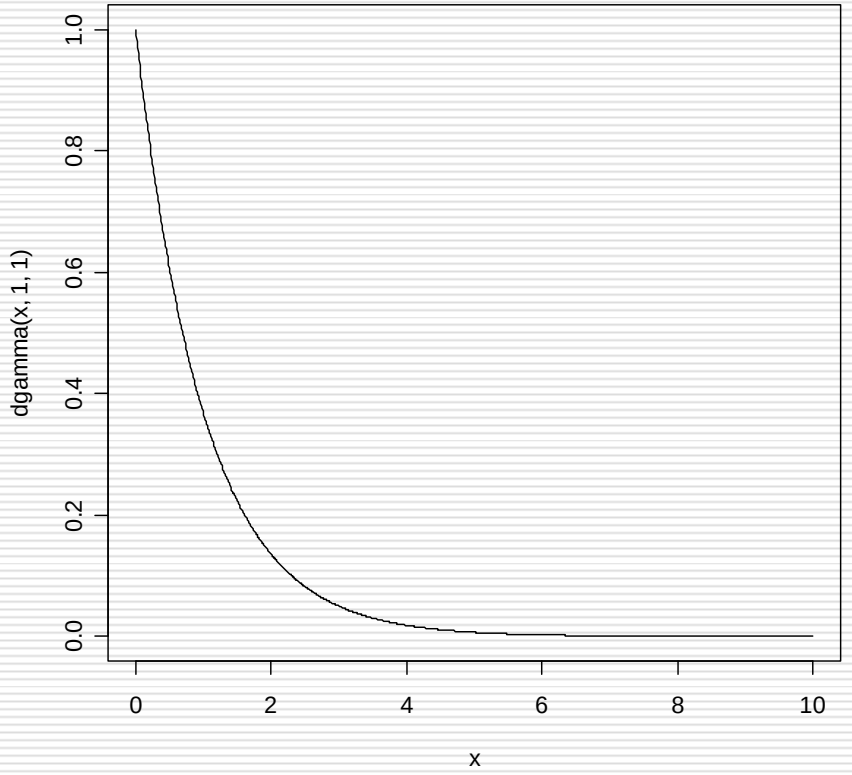
- Για να δούμε την γραφική παράσταση μιας σ.π.π. ή σ.μ.π. δημιουργούμε μια ακολουθία τιμών και παίρνουμε το γράφημα της σ.π.π. ή σ.μ.π. υπολογισμένης στην ακολουθία τιμών.

Γραφικές Παραστάσεις Κατανομών

☐ Παράδειγμα:

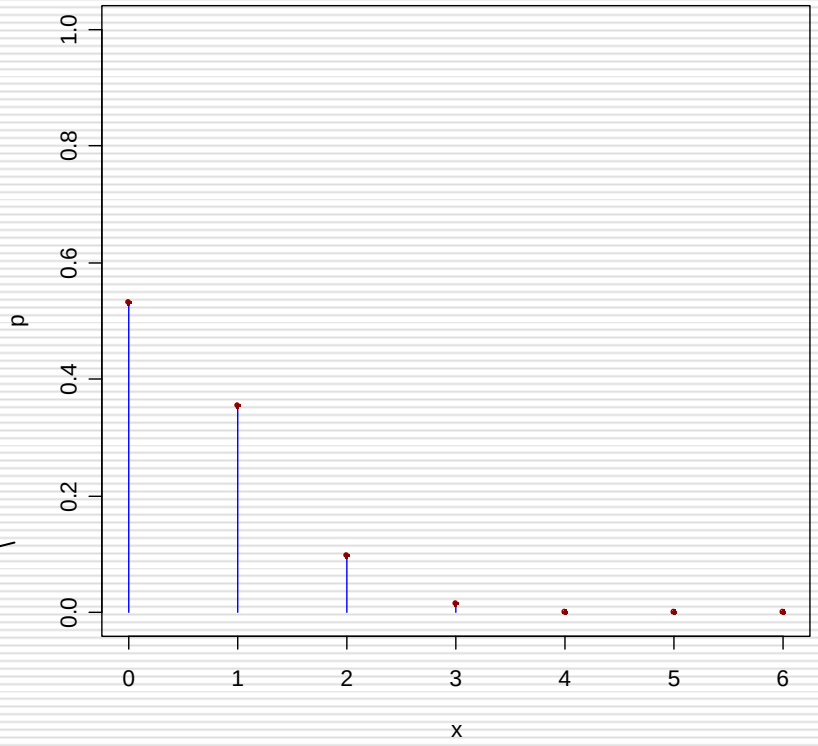
```
> x<-seq(0,10, 0.01)
> plot(x, dgamma(x,1,1), type='l')
```

Γάμμα κατανομή με παραμέτρους (1,1)



Γραφικές Παραστάσεις Κατανομών

```
> n<-6
> p<-0.1
> x<-0:6
> pr<-dbinom(x,n,p)
>plot(x,pr,type="h",xlim=c(0,6),
    ylim=c(0,1),
    col="blue",ylab="p")
>points(x,pr,pch=20,col="dark
    red")
```



Διωνυμική κατανομή με παραμέτρους n=6 και p=0.1

Γραφικός έλεγχος καταλληλότητας κατανομής

- Όπως αναφέραμε και στην εισαγωγή στην παραμετρική στατιστική υποθέτουμε ότι γνωρίζουμε την κατανομή του υπό μελέτη χαρακτηριστικού του πληθυσμού. Μπορούμε να ελέγξουμε γραφικά αυτήν την υπόθεση.
- Ο πιο απλός τρόπος είναι να κάνουμε το ιστόγραμμα των τιμών του δείγματός μας και να το συγκρίνουμε με την γραφική παράσταση της υποτιθέμενης κατανομής.

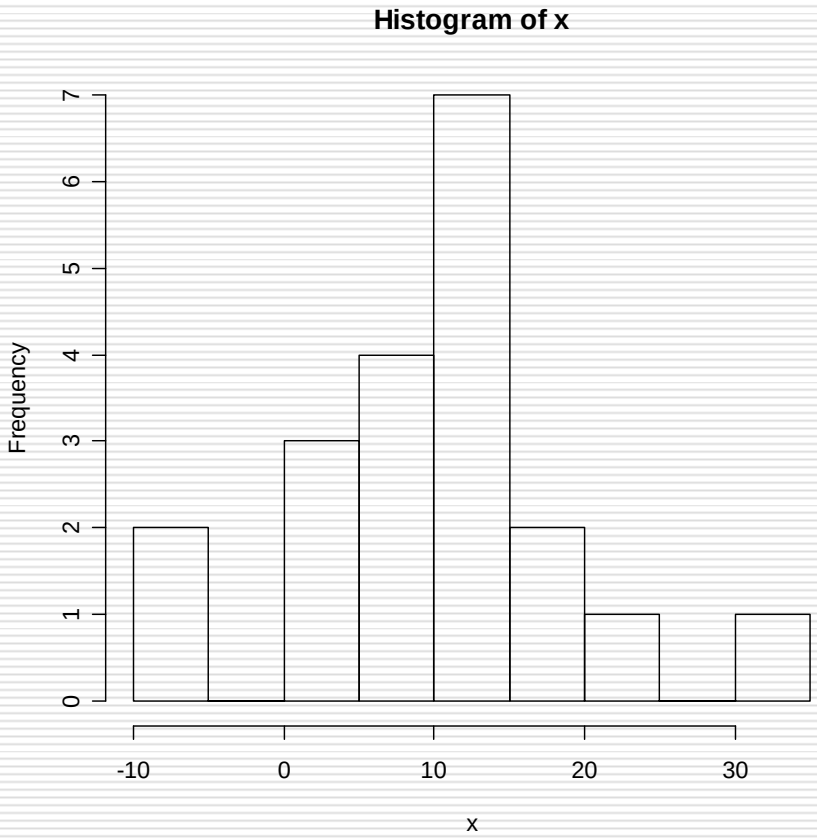
Γραφικός έλεγχος καταλληλότητας κατανομής

- Για παράδειγμα ας υποθέσουμε ότι έχουμε τα εξής δεδομένα

```
> x
 1.2126793  0.1811545 12.6740008 12.1543243 18.2933201
 8.3776755 11.7624305 14.6615550 21.3756432  3.1409630
 8.7359266 13.5479024 10.9576792 17.4204680  7.3268606
13.8451238  8.7802663 34.4452373 -6.7201282 -5.2785726
```

- Το ιστόγραμμά τους μας λέει ότι η υπόθεση της κανονικότητας των εν λόγω δεδομένων δεν είναι παράλογη.

Γραφικός έλεγχος καταλληλότητας κατανομής

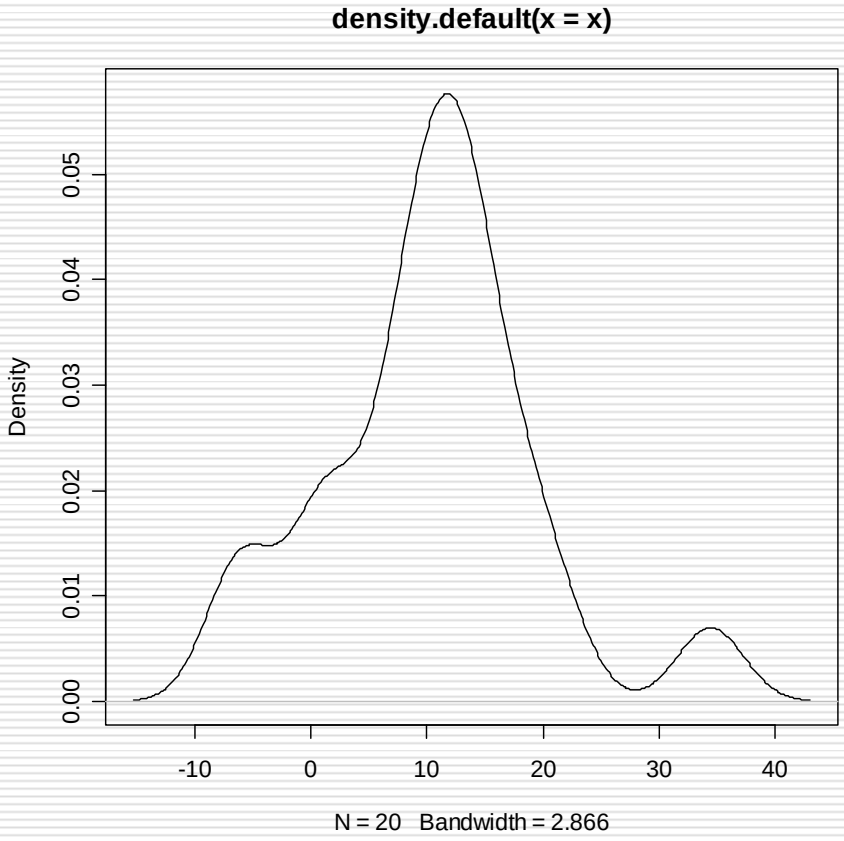


Γραφικός έλεγχος καταλληλότητας κατανομής

- Αντί για ιστόγραμμα μπορούμε να απεικονίσουμε την μη παραμετρική εκτιμήτρια της σ.π.π., με βάση τις παρατηρήσεις και με την βοήθεια της εντολής density στην R

```
> plot(density(x))
```

Γραφικός έλεγχος καταλληλότητας κατανομής

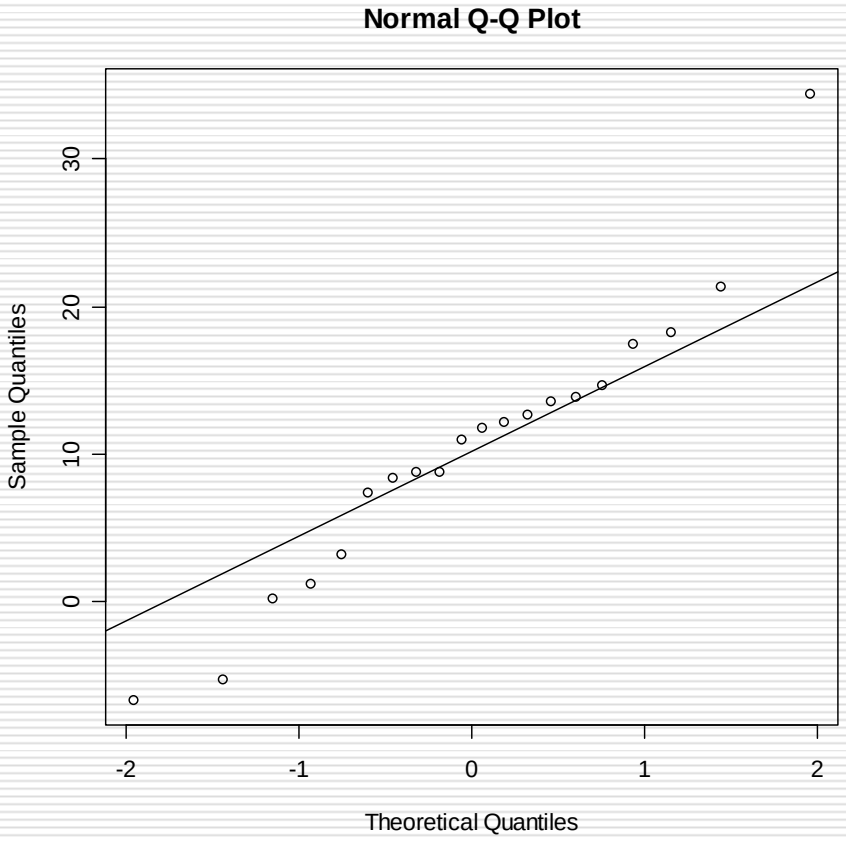


Γραφικός έλεγχος καταλληλότητας κατανομής

- Ειδικά για να ελέγξουμε αν τα δεδομένα μας προέρχονται από την Κανονική κατανομή κάνουμε μια γραφική παράσταση των δειγματικών ποσοστημορίων ως προς τα θεωρητικά ποσοστημόρια της Κανονικής Κατανομής (QQ - PLOT). Όσο πιο κοντά στην γραμμή, που αναπαριστά τα θεωρητικά ποσοστημόρια, είναι τα σημεία, που με την σειρά τους αναπαριστούν τα δειγματικά ποσοστημόρια, τόσο καλύτερη προσαρμογή έχουμε.

```
> qqnorm(x)
> qqline(x)
```

Γραφικός έλεγχος καταλληλότητας κατανομής



Ασθενής Νόμος των Μεγάλων Αριθμών (ΑΝΜΑ)

- Με βάση τον Α.Ν.Μ.Α. αν $X_1,\ldots,X_n$ είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με πεπερασμένη μέση τιμή μ, τότε

$$\bar{X} = n^{-1}\sum_1^n X_i \to \mu \quad \text{κατά πιθανότητα καθώς} \quad n \to \infty.$$

- Εφαρμογή: Έστω $X_i \sim$ Bernoulli(p). Τότε $\mu = P(X_i = 1) = p$, άρα από Α.Ν.Μ.Α.

$$\bar{X} \to p \quad \text{κατά πιθανότητα καθώς} \quad n \to \infty.$$

Ασθενής Νόμος των Μεγάλων Αριθμών (ΑΝΜΑ)

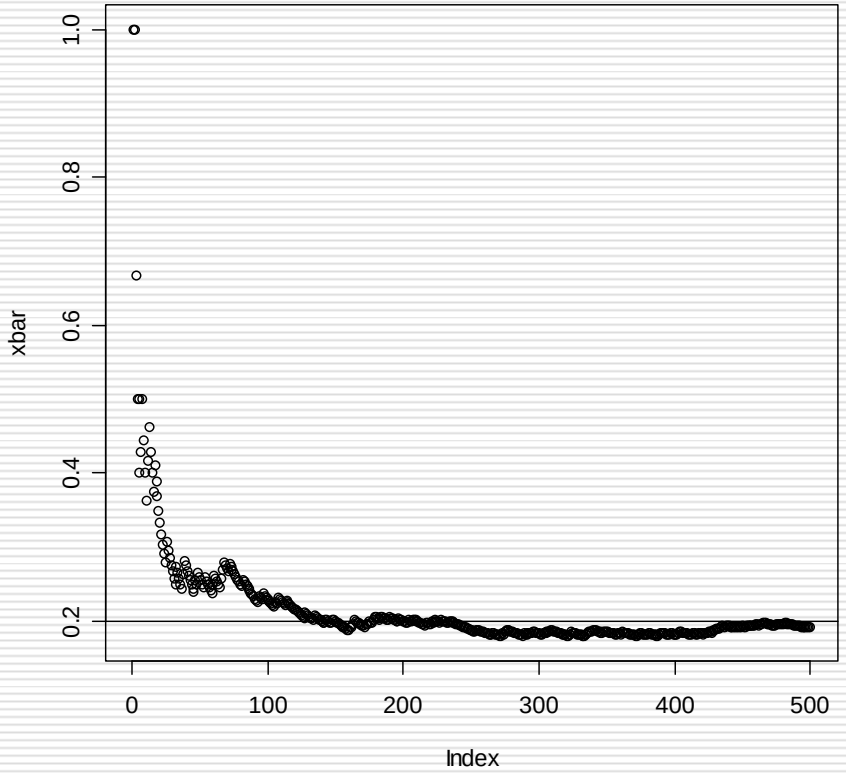
- Για να δούμε ότι πράγματι ισχύει ο Α.Ν.Μ.Α. ας υποθέσουμε ότι p=0.2, n=500 και ας τρέξουμε τον παρακάτω κώδικα:

```
> x<-rbinom(500,1,0.2)
>  xbar<-cumsum(x)/(1:500)
> plot(xbar)
> abline(h=0.2)
```

Ασθενής Νόμος των Μεγάλων Αριθμών (ΑΝΜΑ)

- Με την πρώτη εντολή προσομοιώνουμε (γεννάμε) τυχαίο δείγμα μεγέθους n=500 από την Bernoulli(0.2). Εν συνεχεία θεωρούμε την συνάρτηση του δειγματικού μέσου ως ακολουθία και απεικονίζουμε το γράφημά της εν λόγω ακολουθίας μαζί με την ευθεία y=0.2 για να ελέγξουμε την σύγκλιση.

Ασθενής Νόμος των Μεγάλων Αριθμών (ΑΝΜΑ)

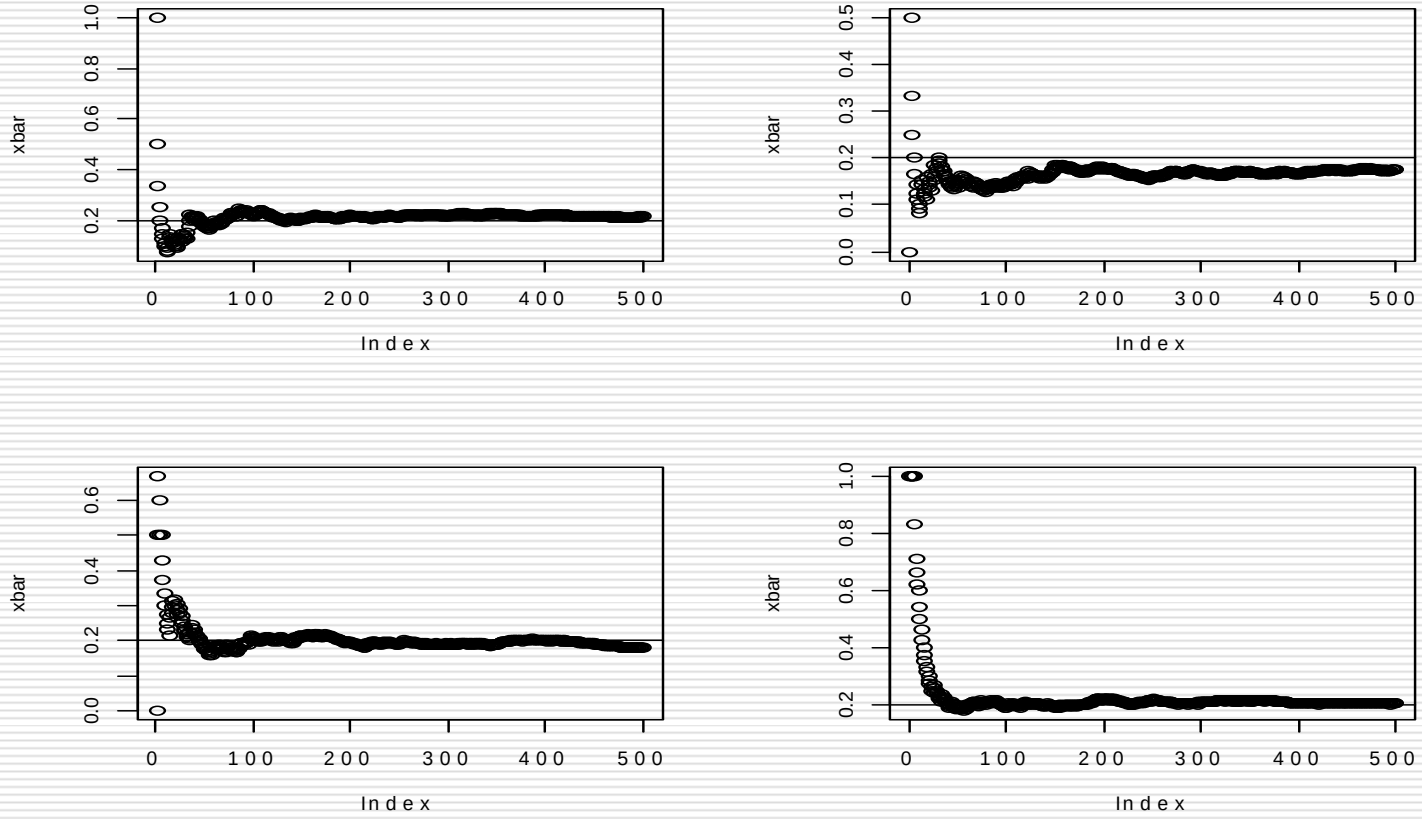


Ασθενής Νόμος των Μεγάλων Αριθμών (ΑΝΜΑ)

- Για να ελέγξουμε την σύγκλιση επαναλαμβάνουμε την διαδικασία 4 φορές.

```
> par(mfrow=c(2,2))
> i<-1
> for(i in 1:4)
 {
 x<-rbinom(500,1,0.2)
 xbar<-cumsum(x)/(1:500)
 plot(xbar)
 abline(h=0.2)
 i<-i+1
 }
```

Ασθενής Νόμος των Μεγάλων Αριθμών (ΑΝΜΑ)

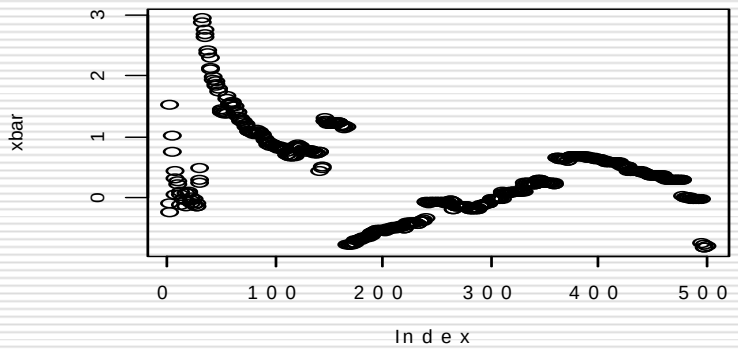


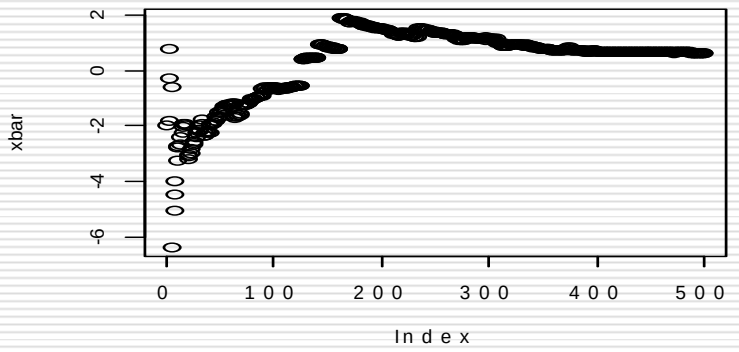
Ασθενής Νόμος των Μεγάλων Αριθμών (ANMA)

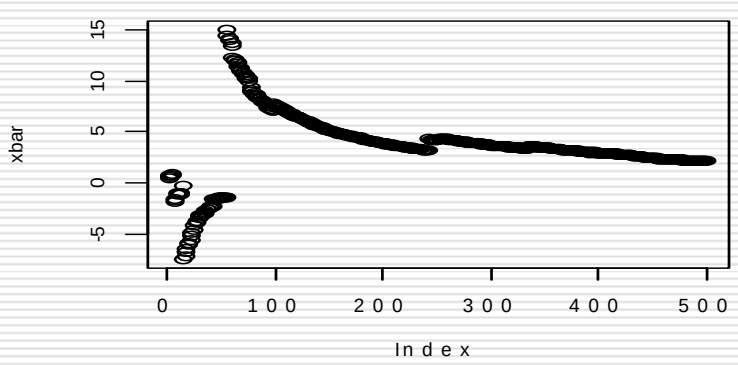
- Όταν η μέση τιμή της θεωρητικής κατανομής δεν υπάρχει τότε ο Α.Ν.Μ.Α. δεν ισχύει προφανώς. Παράδειγμα κατανομής της οποίας δεν υπάρχει η μέση τιμή είναι η Cauchy. Με την βοήθεια της R βλέπουμε ότι δεν υπάρχει σύγκλιση.

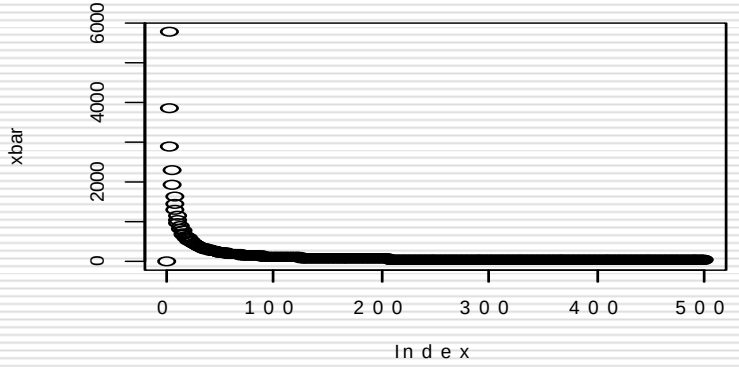
```
> par(mfrow=c(2,2))
> i<-1
> for(i in 1:4)
 {
 x<-rcauchy(500)
 xbar<-cumsum(x)/(1:500)
 plot(xbar)
 i<-i+1
 }
```

Ασθενής Νόμος των Μεγάλων Αριθμών (ΑΝΜΑ)









Κεντρικό Οριακό Θεώρημα

- Με βάση το Κ.Ο.Θ. αν $X_1,\ldots,X_n$ είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με πεπερασμένη μέση τιμή μ και διασπορά σ^2, τότε

$$S_n = \frac{\sum_1^n X_i - n\mu}{\sigma\sqrt{n}} \to Z \sim N(0,1) \ \text{κατά νόμο καθώς} \ n \to \infty.$$

- Ισοδύναμα

$$\bar{X} \to Y \sim N(\mu, \sigma^2/n) \ \text{κατά νόμο καθώς} \ n \to \infty.$$

Κεντρικό Οριακό Θεώρημα

- Εφαρμογή: Έστω τυχαίο δείγμα μεγέθους 150 από την κατανομή Poisson παραμέτρου $\lambda=2$ (τότε $\lambda=\mu=\sigma^2=2$).

```
> poisson.clt<-function(k,n,l)
{
Sn<-rep(NA,k)
i<-1
for(i in 1:k)
{
x<-rpois(n,l)
Sn[i]<-(sum(x)-n*l)/(sqrt(n*l))
}
return(Sn)
}
```

Κεντρικό Οριακό Θεώρημα

- Στην παραπάνω συνάρτηση δημιουργούμε k δείγματα μεγέθους n από την Poisson με παράμετρο l, και για κάθε δείγμα υπολογίζουμε το S_n. Ας ελέγξουμε, με βάση τις k τιμές για το S_n, αν πράγματι η κατανομή του είναι Κανονική. Επιλέγουμε k=300.

```
> run<-poisson.clt(300,150,2)
> par(mfrow=c(1,2))
> hist(run)
> qqnorm(run)
> qqline(run)
```

Κεντρικό Οριακό Θεώρημα