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

Εισαγωγικά

- Όπως είδαμε και στην εισαγωγή, στην Στατιστική συνήθως ενδιαφερόμαστε να εκτιμήσουμε ένα άγνωστο μέγεθος, το οποίο καλείται παράμετρος και το οποίο συνήθως συνοψίζει κατά κάποιον τρόπο τις τιμές της υπό μελέτης μεταβλητής στον πληθυσμό, π.χ. τη μέση της τιμή. Η εκτίμησή μας γίνεται με την βοήθεια κατάλληλα επιλεγμένων δειγματοσυναρτήσεων, συναρτήσεων δηλαδή του δείγματος, οι οποίες καλούνται (σημειακές) εκτιμήτριες. Ο τρόπος επιλογής εκτιμητριών γίνεται είτε (α) με βάση την λογική, π.χ. αν θέλουμε να εκτιμήσουμε την μέση τιμή στον πληθυσμό μας ακούγεται λογικό να χρησιμοποιήσουμε ως εκτιμήτρια την μέση τιμή του δείγματος (plug in principle), είτε (β) με βάση διάφορες ιδιότητες, π.χ. η εκτιμήτρια μας θέλουμε να έχει μέση τιμή ίση με την ποσότητα όπου εκτιμά (αμεροληψία) είτε (γ) με βάση κάποιο κριτήριο κατασκευής (π.χ. εκτιμήτριες μέγιστης πιθανοφάνειας).
- Να υπενθυμίσουμε εδώ ότι οι εκτιμήτριες ως δειγματοσυναρτήσεις είναι τυχαίες μεταβλητές και άρα η ίδια εκτιμήτρια συνήθως παίρνει άλλη τιμή όταν παρατηρούμε άλλα δεδομένα. Συνήθως θέλουμε η εκτιμήτρια μας να έχει μικρή μεταβλητότητα (δηλαδή διασπορά), έτσι ώστε το τυπικό σφάλμα εκτίμησης (η τυπική απόκλιση της εκτιμήτριας) να είναι μικρό, δηλαδή οι τιμές της εκτιμήτριας μας να μην μεταβάλλονται πολύ από δείγμα σε δείγμα.

Εκτιμήτριες Μέγιστης Πιθανοφάνειας

- Το πιο διαδεδομένο κριτήριο κατασκευής εκτιμητριών είναι αυτό της μέγιστης πιθανοφάνειας (maximum likelihood). Πιθανοφάνεια καλείται η από κοινού σ.π.π. ή σ.μ.π. του τυχαίου δείγματος $X_1,...,X_n$. Αν το χαρακτηριστικό που μας ενδιαφέρει προέρχεται από ένα πληθυσμό με σ.π.π ή σ.μ.π. $f(x;\boldsymbol{\theta})$ όπου $\boldsymbol{\theta}$ είναι το διάνυσμα των αγνώστων παραμέτρων, τότε επειδή το τυχαίο δείγμα αποτελείται από ανεξάρτητες και ισόνομες τ.μ. η πιθανοφάνεια θα είναι ίση με

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta})$$

Εκτιμήτριες Μέγιστης Πιθανοφάνειας

- Η παραπάνω συνάρτηση είναι μια συνάρτηση του **θ** και μας δίνει την πιθανότητα το δείγμα μας να προέρχεται από την υποτιθέμενη κατανομή με παράμετρο **θ**. Συνεπώς μπορούμε να διαλέξουμε το **θ** έτσι ώστε να μεγιστοποιείται αυτή η πιθανότητα. Οι εκτιμήτριες που παίρνουμε με τον τρόπο αυτόν καλούνται Εκτιμήτριες Μέγιστης Πιθανοφάνειας (Ε.Μ.Π.) και τις συμβολίζουμε με $\hat{\boldsymbol{\theta}}$.

- Αρκετά συχνά δουλεύουμε με την λογαριθμική πιθανοφάνεια για λόγους ευκολίας υπολογισμών

$$l(\theta) = \log(\mathrm{L}(\boldsymbol{\theta})) = \sum_{i=1}^{n} \log\left(\mathrm{f}(\mathrm{x}_i;\theta)\right)$$

Εκτιμήτριες Μέγιστης Πιθανοφάνειας

- Για την μεγιστοποίηση της παραπάνω συνάρτησης συνήθως δουλεύουμε αναλυτικά, παίρνουμε τις μερικές παραγώγους ως προς κάθε συνιστώσα του **θ,** τις εξισώνουμε με το μηδέν και λύνουμε το σύστημα που προκύπτει. Εν συνεχεία με την βοήθεια των δευτέρων μερικών παραγώγων ελέγχουμε αν το σημείο είναι πράγματι σημείο μεγίστου.
- Με την βοήθεια της R μπορούμε επίσης να δούμε το γράφημα της πιθανοφάνειας και να βρούμε με την βοήθειά του το μέγιστο.

Εκτιμήτριες Μέγιστης Πιθανοφάνειας

- Παράδειγμα 1: Έστω ότι ο αριθμός Χ των πετρελαιοφόρων που φθάνουν κάθε μέρα σε ένα λιμάνι είναι τ.μ. με κατανομή Poisson(λ), όπου λ άγνωστο. Τα παρακάτω δεδομένα αποτελούν τις παρατηρήσεις από τυχαίο δείγμα 20 ημερών

  9 4 5 5 7 13 8 3 6 5 4 5 10 5 5 4 3 3 4 7

Εκτιμήτριες Μέγιστης Πιθανοφάνειας

- Τότε

$$l(\lambda) = \log(L(\lambda)) = \sum_{i=1}^{n} \log\left(e^{-\lambda}\lambda^{x_i} / x_i!\right) = -n\lambda + \sum_{i=1}^{n} x_i \log\lambda - \log\prod_{i=1}^{n} x_i!$$

- Εύκολα προκύπτει ότι η Ε.Μ.Π. του λ τότε είναι $\hat{\lambda} = \bar{X}.$
- Για τις συγκεκριμένες παρατηρήσεις η τιμή της εκτιμήτριάς μας είναι

$$\hat{\lambda} = \bar{x} = 5.75.$$

Εκτιμήτριες Μέγιστης Πιθανοφάνειας

- Η επόμενη συνάρτηση στην R υπολογίζει την λογαριθμική Poisson πιθανοφάνεια για συγκεκριμένο δείγμα και για 10000 διαφορετικές τιμές του λ.

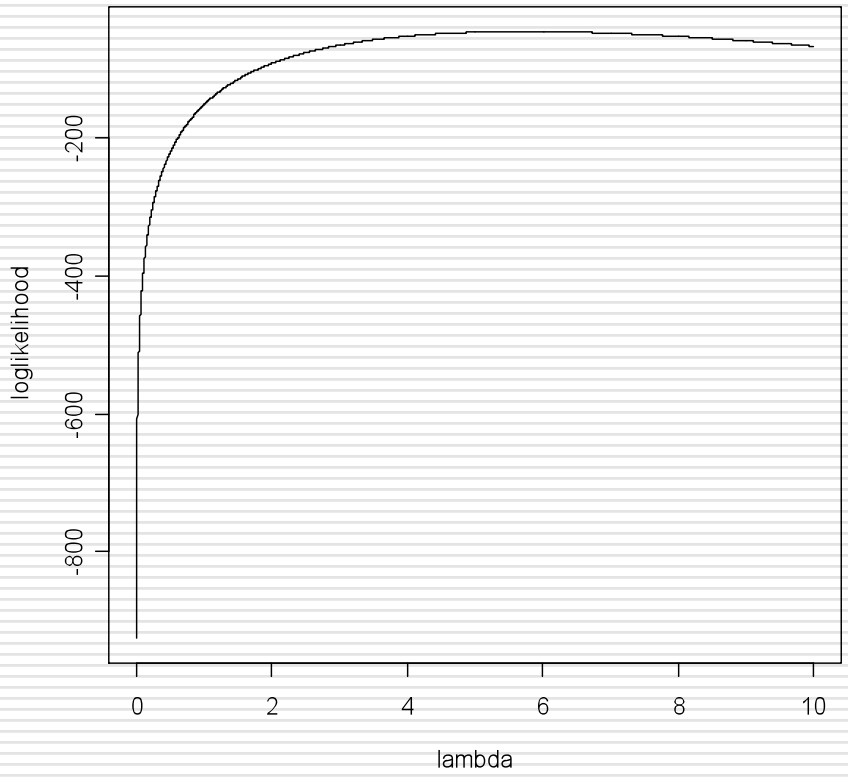
```
> lambda<-seq(0.001, 10, length=10000)
> poisson_loglikelihood<-function(data, lambda){
   results<-rep(NA,10000)
   for(i in 1:10000){
     results[i]<-sum(dpois(data,lambda[i],log=T))
    }
    return(results)
    }
```

Εκτιμήτριες Μέγιστης Πιθανοφάνειας

- Μπορούμε λοιπόν να κάνουμε ένα γράφημα της εν λόγω συνάρτησης και να δούμε εμπειρικά που μεγιστοποιείται.

```
> x<-c(9,4,5,5,7,13,8,3,6,5,4,5,10,5,5,4,3,3,4,7)
> results<-poisson_loglikelihood(x, lambda)
> plot(lambda, results, xlab="lambda", ylab="loglikelihood",
type="l")
```

Εκτιμήτριες Μέγιστης Πιθανοφάνειας



Εκτιμήτριες Μέγιστης Πιθανοφάνειας

- Για να βρούμε εμπειρικά το μέγιστο γράφουμε την εντολή

```
> lambda[order(results)[10000]]
[1] 5.75
```

- Η εντολή order(results)[10000] μας επιστρέφει τη θέση που υπάρχει η μεγαλύτερη $l(\lambda)$, και άρα η εντολή lambda[order(results)[10000]] μας επιστρέφει πράγματι το λ που μεγιστοποιεί την $l(\lambda)$.
- Η παραπάνω τιμή δεν σημαίνει κατ' ανάγκη ότι είναι το μέγιστο, καθώς έχουμε πάρει μόνο ένα αριθμό σημείων λ. Στην περίπτωσή μας είναι πράγματι το μέγιστο αφού

```
> mean(x)
[1] 5.75
```

Εκτιμήτριες Μέγιστης Πιθανοφάνειας

- Παράδειγμα 2: Ο χρόνος Χ (σε λεπτά) αναμονής στην ουρά για να εξυπηρετηθείτε από το ταμείο μιας τράπεζας έστω ότι είναι τ.μ. ομοιόμορφα κατανεμημένη στο διάστημα (0,θ), με θ άγνωστη παράμετρο. Έστω ότι έχουμε τις παρακάτω παρατηρήσεις 12 χρόνων αναμονής

5 0 2 10 6 4 3 8 9 7 8 4

Εκτιμήτριες Μέγιστης Πιθανοφάνειας

- Η σ.π.π. της ομοιόμορφης κατανομής στο διάστημα (0, θ) είναι:

$$f(x;\theta) = \theta^{-1}, \quad x \in (0,\theta), \ \theta > 0$$

- Η πιθανοφάνεια τότε είναι

$$L(\theta) = \prod_{i=1}^{n} I_{(0,\theta)}(x_i) \qquad \text{οπου}$$

$$I_{(0,\theta)}(x) = \begin{cases} 1 & \text{αν } x \in (0,\theta) \\ 0 & \text{αν } x \notin (0,\theta) \end{cases}$$

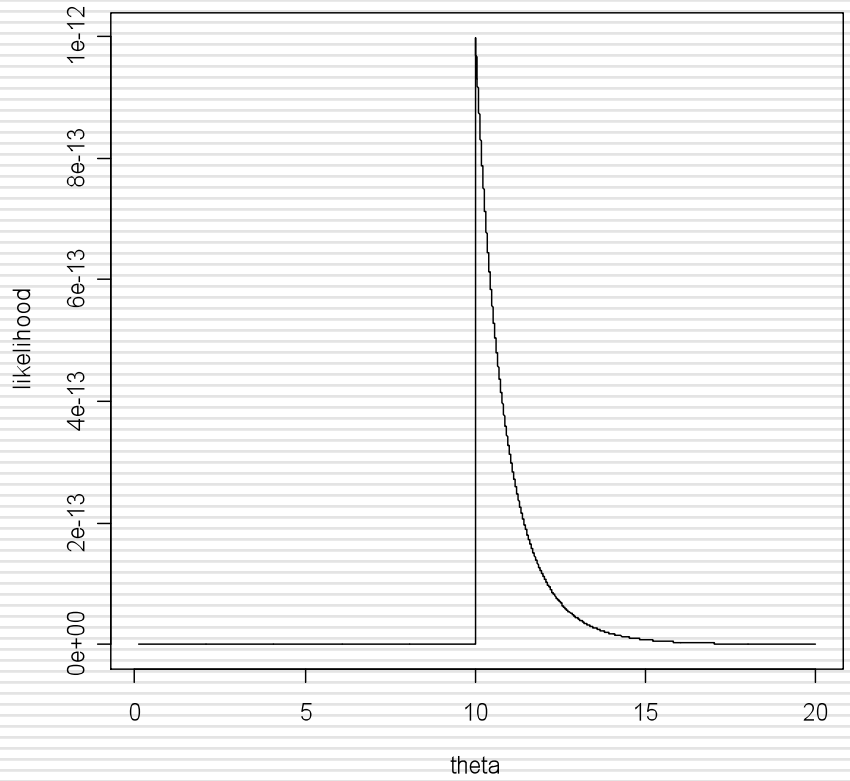
είναι η δείκτρια του συνόλου (0,θ).

Εκτιμήτριες Μέγιστης Πιθανοφάνειας

- Ο αναλυτικός τρόπος με την παράγωγο δεν μπορεί να δουλέψει στο εν λόγω παράδειγμα οπότε θα ζητήσουμε την βοήθεια της R για να βρούμε το μέγιστο.

```
> theta<-seq(0.1, 20, length=10000)
> uniform_likelihood<-function(data, theta){
    results<-rep(NA,10000)
    for(i in 1:10000){
        results[i]<-prod(dunif(data,0, theta[i]))
    }
    return(results)
    }
> y<-c(5, 0, 2, 10, 6, 4, 3, 8, 9, 7, 8, 4)
> results<-uniform_likelihood(y, theta)
> plot(theta, results, xlab="theta", ylab="likelihood",
    type="l")
```

Εκτιμήτριες Μέγιστης Πιθανοφάνειας



Εκτιμήτριες Μέγιστης Πιθανοφάνειας

- Η πιθανοφάνεια όπως περιμέναμε είναι 0 όταν $\theta < y_i$, για τουλάχιστον ένα i, ενώ όταν $\theta \geq y_i$ (για όλα τα i = 1,..,n) είναι μια φθίνουσα συνάρτηση του θ. Θέλουμε λοιπόν την μικρότερη τιμή του θ που να ικανοποιεί όμως την ανισότητα $\theta \geq y_i$ (για όλα τα i = 1,..,n), και άρα το μέγιστο (Ε.Μ.Π.) είναι το $\hat{\theta} = \max_{i=1,..n} \{y_i\}$.

Εκτιμήτριες μέγιστης πιθανοφάνειας

Μια άλλη εκτιμήτρια (αμερόληπτη μάλιστα) της άγνωστης παραμέτρου θ είναι και η

$$\tilde{\theta} = 2\,\frac{X_1 + X_2 + \cdots + X_n}{n}$$

Από τον νόμο των μεγάλων αριθμών το δεξί μέλος συγκλίνει στο $2E_\theta[X]=2\theta/2=\theta$ όταν το n πάει στο άπειρο . Μπορούμε να την συγκρίνουμε με την Ε.Μ.Π. και να αποφασίσουμε ποια είναι καλύτερη;

Εκτιμήτριες μέγιστης πιθανοφάνειας

Ας παίξουμε το ακόλουθο παιχνίδι: ας ζητήσουμε από τον υπολογιστή να διαλέξει τυχαία μια τιμή για την θ. Χωρίς να την δούμε (για μας θα είναι η άγνωστη παράμετρος) θα προσπαθήσουμε να την εκτιμήσουμε με τη βοήθεια των δύο εκτιμητριών, προσομοιώνοντας δείγματα από μια τ.μ. με ομοιόμορφη κατανομή στο διάστημα (0,θ).

Εκτιμήτρια μέγιστης πιθανοφάνειας

Στην μεταβλητή theta εκχωρούμε ένα τυχαίο αριθμό στο διάστημα (0,100). Αυτή θα είναι η άγνωστη για εμάς παράμετρος θ

> theta<-runif(1,0,100)

Προσομοιώνουμε τώρα 100000 δείγματα από μια τ.μ. με ομοιόμορφη κατανομή στο (0,θ)

> x<-runif(100000,0,theta)

Υπολογίζουμε στη συνέχεια τις εκτιμήσεις της άγνωστης (σε εμάς) παραμέτρου θ που δίνουν οι δύο εκτιμήτριες,

$$\hat{\theta} = \max X_i \qquad \tilde{\theta} = 2\bar{X}$$

Εκτιμήτρια μέγιστης πιθανοφάνειας

Ας δούμε τα αποτελέσματα....

```
> a<-c(max(x),2*mean(x))
> a
[1] 49.19720 49.24383
```

Και ας δούμε ποια τιμή είχε εκχωρήσει ο υπολογιστής στην theta…

```
> theta
[1] 49.19777
```

Η Ε.Μ.Π. έδωσε μια εκτίμηση πιο κοντά στην πραγματική τιμή αλλά αυτό θα μπορούσε να είναι συμπτωματικό. Ας επαναλάβουμε το πείραμα.

Εκτιμήτρια μέγιστης πιθανοφάνειας

```
> x<-runif(100000,0,theta)
> a<-c(max(x),2*mean(x))
> a
[1] 49.19703 49.02390
```

Βλέπουμε ότι η ΕΜΠ έδωσε πάλι καλύτερη εκτίμηση της πραγματικής τιμής της theta, και ότι δεν έχει μεταβληθεί πολύ σε σχέση με την προηγούμενη εκτίμηση που πείραμε.

Φαίνεται ότι οι εκτιμήσεις της ΕΜΠ έχουν μικρότερη διασπορά. Θα επαναλάβουμε λοιπόν πολλές φορές το προηγούμενο πείραμα και θα εξετάσουμε τα δείγματα των εκτιμήσεων που δίνουν οι δύο εκτιμήτριες.

Εκτιμήτρια μέγιστης πιθανοφάνειας

```
> y<-rep(NA,1000)
> z<-rep(NA,1000)
> for (i in 1:1000) {x<-runif(100000,0,theta)
+ y[i]<-max(x)
+ z[i]<-2*mean(x) }
> summary(y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  49.19   49.20   49.20   49.20   49.20   49.20
> summary(z)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  48.93   49.13   49.19   49.19   49.25   49.47
```

Η ΕΜΠ βλέπουμε ότι έχει μικρότερη μεταβλητότητα και μάλλον θα την προτιμούσαμε σαν εκτιμήτρια της θ.

Διαστήματα Εμπιστοσύνης

- Οι (σημειακές) εκτιμήσεις δεν μας δίνουν κάποια πληροφορία σχετικά με την ακρίβεια ή το σφάλμα εκτίμησης.
- Είναι λοιπόν χρήσιμο να προσδιορίσουμε, μέσω των εκτιμητριών και των τυπικών τους σφαλμάτων, ένα διάστημα το οποίο θα περιέχει την άγνωστη τιμή της παραμέτρου με καθορισμένη πιθανότητα, έστω γ.
- Σκοπός μας δηλαδή είναι να βρούμε δυο ποσότητες u και v (u<v) έτσι ώστε $P(u \leq \theta \leq v)=\gamma=1-\alpha$.
- Το [u,v] καλείται διάστημα εμπιστοσύνης (Δ.Ε.) με συντελεστή εμπιστοσύνης (σ.ε.) $\gamma=1-\alpha$.
- Το διάστημα εμπιστοσύνης του θ προσδιορίζεται με βάση την κατανομή της εκτιμήτριας του θ από το τυχαίο δείγμα, συνεπώς οι τιμές u και v είναι τυχαίες μεταβλητές. Αυτό σημαίνει ότι από διαφορετικό δείγμα ίδιου μεγέθους ενδέχεται να προκύψουν διαφορετικά Δ.Ε. για το θ.

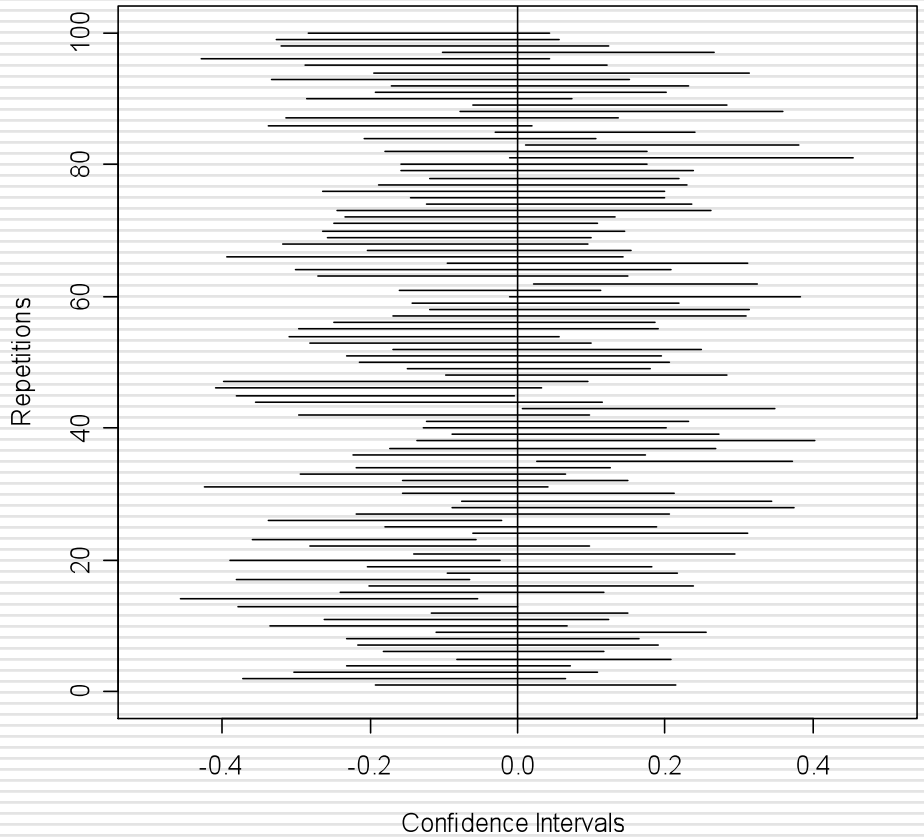
Διαστήματα Εμπιστοσύνης

- Το εύρος του Δ.Ε. εξαρτάται από το τυπικό σφάλμα της εκτιμήτριας του θ και τον συντελεστή εμπιστοσύνης. Όσο μεγαλύτερο είναι το τυπικό σφάλμα της εκτιμήτριας του θ τόσο μεγαλύτερο εύρος έχει το Δ.Ε. Επίσης όσο μεγαλύτερο συντελεστή εμπιστοσύνης έχουμε τόσο μεγαλύτερο εύρος έχει το Δ.Ε.
- Το τυπικό σφάλμα των εκτιμητριών είναι αντιστρόφως ανάλογο του μεγέθους του δείγματος n. Συνεπώς όσο το n αυξάνει τόσο το εύρος του Δ.Ε. θα μειώνεται.

Διαστήματα Εμπιστοσύνης

- Η πραγματική ερμηνεία ενός Δ.Ε. με σ.ε. γ είναι η ακόλουθη. Σε μια σειρά κατασκευών διαστημάτων εμπιστοσύνης μιας παραμέτρου, με ανεξάρτητα δείγματα του αυτού μεγέθους, ένα ποσοστό 100 γ% των διαστημάτων αυτών "αναμένεται" να περιέχουν την αληθή τιμή της παραμέτρου θ.

Διαστήματα Εμπιστοσύνης



Διαστήματα Εμπιστοσύνης

- Παράδειγμα: Έστω $X_1,\ldots,X_n$ τυχαίο δείγμα από πληθυσμό με μέση τιμή μ (άγνωστη) και διασπορά σ^2 (γνωστή). Μια λογική εκτιμήτρια για το μ είναι ο δειγματικός μέσος $\bar{X}$.
- Για μεγάλο n από το Κ.Ο.Θ. ξέρουμε ότι

$$\bar{X} \dot{\sim} N(\mu, \sigma^2/n).$$

- Συνεπώς το τυπικό σφάλμα της εκτιμήτριάς μας είναι $\sigma/\sqrt{n}$.

Διαστήματα Εμπιστοσύνης

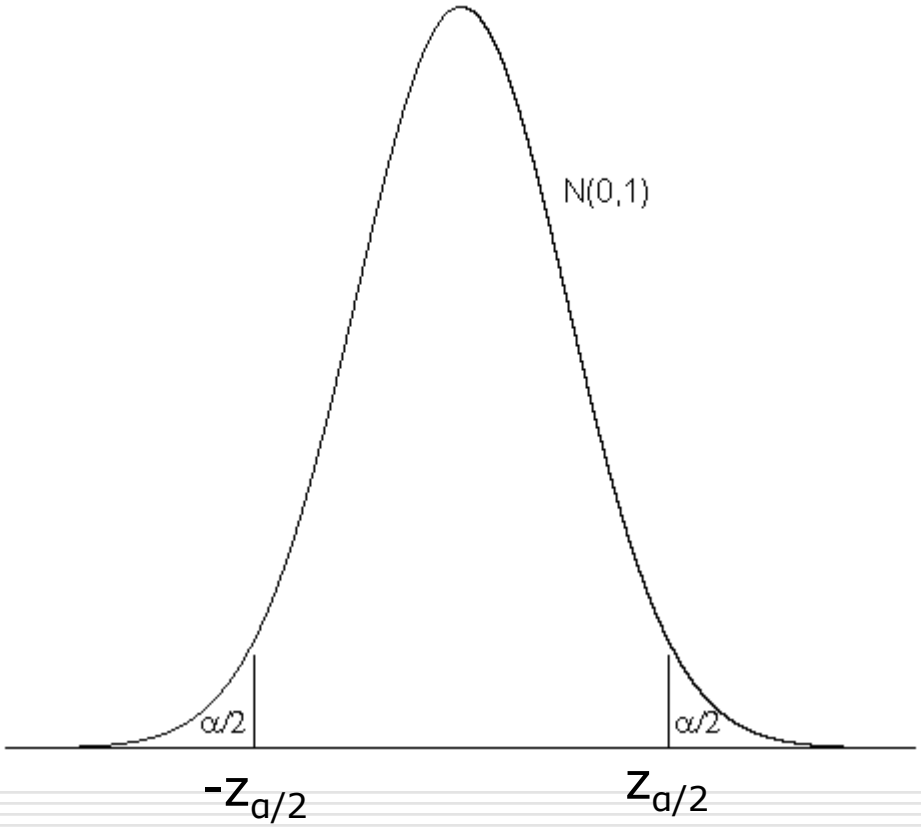
- Τότε όμως η τ.μ.

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

και συνεπώς αν $z_{a/2}$ το σημείο εκείνο της τυποποιημένης Κανονικής κατανομής για το οποίο $P(Z>z_{a/2})=a/2$ έχουμε

$$P(-z_{a/2}<Z<z_{a/2})=\gamma=1-a.$$

Διαστήματα Εμπιστοσύνης



Διαστήματα Εμπιστοσύνης

- Λύνοντας την ανισότητα $-z_{\alpha/2}<Z<z_{\alpha/2}$ ως προς μ προκύπτει το ακόλουθο γ% Δ.Ε. για το μ

$$\left(\bar{X}-z_{\alpha/2}\sigma/\sqrt{n},\ \bar{X}+z_{\alpha/2}\sigma/\sqrt{n}\right).$$

- Η πιθανότητα

$$P\left(\bar{X}-z_{\alpha/2}\sigma/\sqrt{n},\ \bar{X}+z_{\alpha/2}\sigma/\sqrt{n}\right)=\gamma=1-\alpha$$

δεν εκφράζει την πιθανότητα το μ να πάρει τιμές στο εν λόγω διάστημα, διότι το μ είναι μια σταθερά (αν και άγνωστη), αλλά είναι η πιθανότητα το εν λόγω διάστημα να περιέχει την πραγματική τιμή του μ.

Διαστήματα Εμπιστοσύνης

- Αν $\bar{x}$ είναι η τιμή του $\bar{X}$ σε συγκεκριμένο δείγμα τότε θα εκτιμήσουμε το μ με το διάστημα

$$\left(\bar{x} - z_{\alpha/2}\sigma/\sqrt{n},\ \bar{x} + z_{\alpha/2}\sigma/\sqrt{n}\right)$$

το οποίο με πιθανότητα γ περιέχει το άγνωστο μ.