

# Παράδειγμα

---

Για τα NB10 δεδομένα έχουμε το μοντέλο:

$$\begin{array}{llll} \nu & \sim & p(\nu) & \\ \sigma^2 & \sim & \text{SI-}\chi^2(\nu_0, \sigma_0^2) & \\ (\mu|\sigma^2) & \sim & N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right) & \text{or } N(\mu_0, \tilde{\sigma}^2) \\ (\lambda_i|\nu) & \stackrel{\text{IID}}{\sim} & \Gamma^{-1}\left(\frac{\nu}{2}, \frac{\nu}{2}\right) & \\ (y_i|\mu, \sigma^2, \lambda_i) & \stackrel{\text{indep}}{\sim} & N(\mu, \lambda_i \sigma^2) & \left. \vphantom{(\lambda_i|\nu)} \right\} (y_i|\mu, \sigma^2, \nu) \stackrel{\text{IID}}{\sim} t_\nu(\mu, \sigma^2) \end{array}$$

# Ερωτήματα

---

- Εκ των προτέρων κατανομή για το  $\mu$ . Δεν γνωρίζω τίποτα για το πραγματικό βάρος του NB10, άρα πρέπει να χρησιμοποιήσω μια διακεχυμένη (diffuse) εκ των προτέρων κατανομή. Η εντολή  $\mu \sim \text{dnorm}(0.0, 1.0E-6)$  στο WinBUGS μας λέει ότι το  $\mu$  κατανέμεται κανονικά με μέσο 0 και ακρίβεια (precision)  $10^{-6}$ , δηλ.  $SD = 1/\sqrt{10^{-6}} = 1000$ , οπότε μπορεί να πάρει οποιαδήποτε τιμή από -3000 μέχρι 3000.
- Όμοια θέλουμε μια διακεχυμένη εκ των προτέρων κατανομή για το  $\sigma^2$  ή ισοδύναμα για την ακρίβεια  $\tau = 1/\sigma^2$ . Συνηθίζεται να χρησιμοποιούμε  $\tau \sim \Gamma(\epsilon, \epsilon)$ , για μικρό  $\epsilon$ , π.χ.  $\epsilon = 0.001$ . Η συγκεκριμένη κατανομή είναι σχεδόν επίπεδη για το μεγαλύτερο σύνολο τιμών του συνόλου  $(0, \infty)$ , ενώ καθώς  $\tau \downarrow 0$ ,  $\Gamma(\epsilon, \epsilon) \uparrow \infty$ .
- Η ιδέα πίσω από τις διακεχυμένες εκ των προτέρων είναι να τις κατασκευάσουμε με τρόπο τέτοιο ώστε να είναι σχεδόν σταθερές στην περιοχή που η πιθανοφάνεια είναι μη αμελητέα.
- Τέλος για την εκ των προτέρων του  $\nu$ , μπορούμε να πούμε από το qqplot των δεδομένων πως οι βαθμοί ελευθερίας της  $t$ -κατανομής είναι σχετικά μικροί, οπότε χρησιμοποιούμε ομοιόμορφη κατανομή  $(2, 12)$  ( $\nu > 2$ , για να υπάρχει η διασπορά).

# Ερωτήματα

---

- Εύκολα με την βοήθεια της κλασικής στατιστικής μπορούμε να υπολογίσουμε ένα 95% διάστημα εμπιστοσύνης για το  $\mu$ : είναι το  $(\bar{y} \pm t_{n-1,0.975} s / \sqrt{n}) = (403.3, 405.9)$ , οπότε η πιθανοφάνεια για το  $\mu$  πρέπει να είναι “μη-αμελητέα” στο διάστημα (402, 407).

- Ακόμα γνωρίζουμε ότι  $\frac{[(n-1)s^2]}{\sigma^2} \sim \chi_{n-1}^2$  με  $s^2$  η αμερόληπτη δειγματική διασπορά που θεωρείτε τυχαία και  $\sigma^2$  σταθερά. Άρα ένα 99% Δ.Ε. για το  $\sigma^2$  είναι το

$$\left[ \frac{(n-1)s^2}{\chi_{n-1,0.005}^2}, \frac{(n-1)s^2}{\chi_{n-1,0.995}^2} \right].$$

# Εκ των προτέρων

---

```
> print(n<-length(nb10))
[1] 100
> print(s2<-var(nb10))
[1] 41.8201
> qchisq(0.005,99)
[1] 66.5101
> qchisq(0.995,99)
[1] 138.9868
> (n-1)*s2/qchisq(0.995,99)
[1] 29.78837
> (n-1)*s2/qchisq(0.005,99)
[1] 62.24904
> qchisq(0.995,99)/((n-1)*s2)
[1] 0.03357015
> qchisq(0.005,99)/((n-1)*s2)
[1] 0.01606451
```

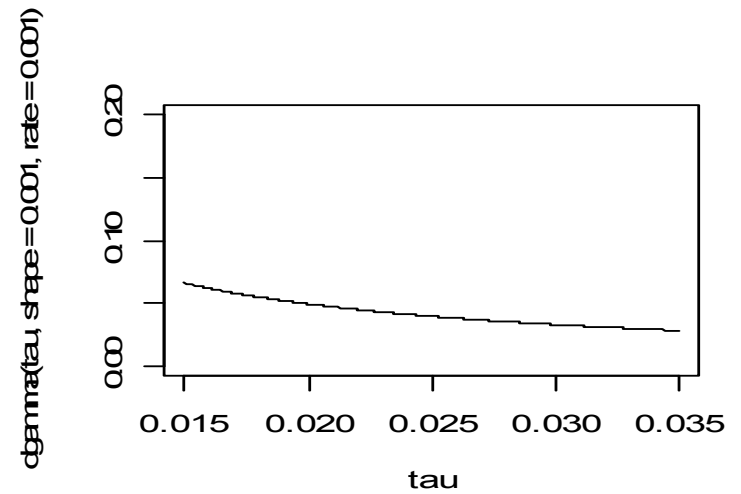
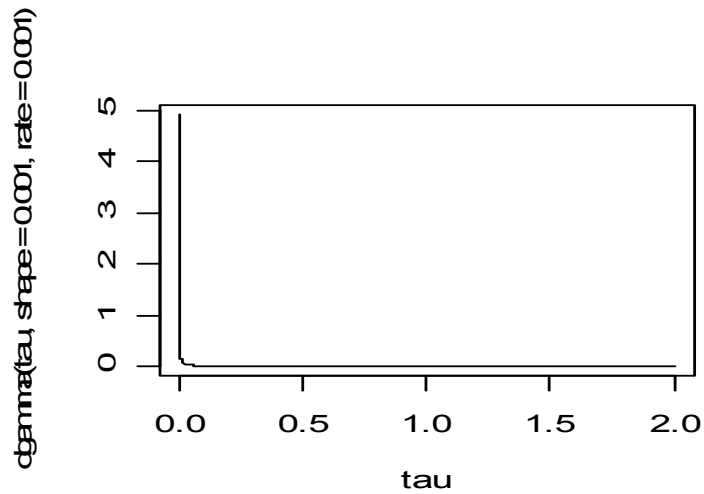
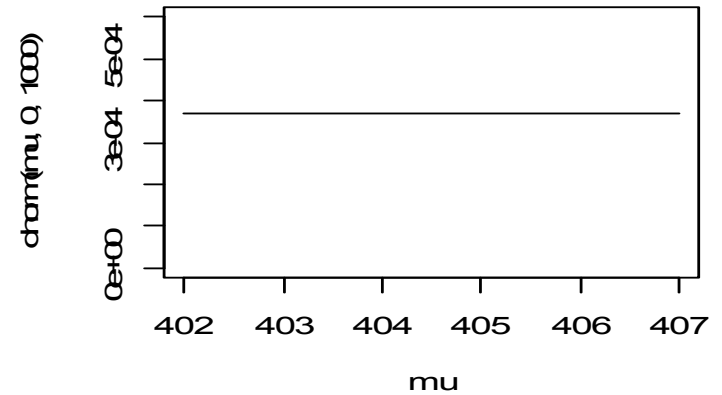
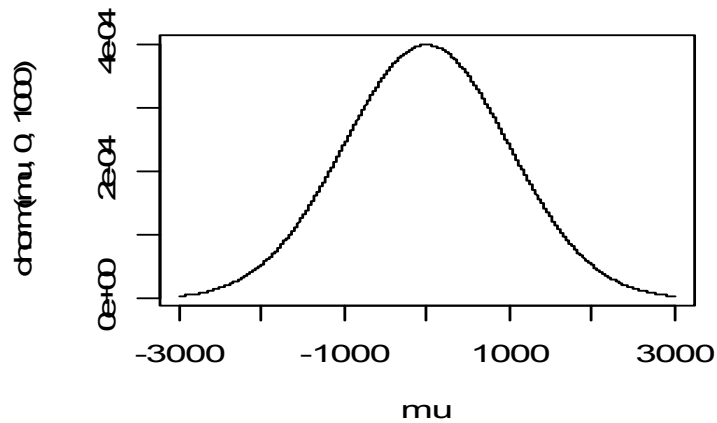
Οπότε η πιθανοφάνεια για το  $\tau$  πρέπει να είναι “μη-αμελητέα” στο διάστημα (0.015, 0.035).

# Εκ των προτέρων

---

```
> par(mfrow=c(2,2))
> mu<-seq(from=-3000,to=3000, length=10000)
> plot(mu,dnorm(mu,0,1000),type='l')
> mu<-seq(from=402,to=407, length=1000)
> plot(mu,dnorm(mu,0,1000), ylim=c(0,0.0006),type='l')
> tau<-seq(from=0.0, to=2.0, length=10000)
> plot(tau,dgamma(tau,shape=0.001,rate=0.001),type='l')
> tau<-seq(from=0.015, to=0.035, length=1000)
> plot(tau,dgamma(tau,shape=0.001,rate=0.001),type='l',
ylim=c(0,0.20))
```

# Εκ των προτέρων



# Αρχικές τιμές

---

- $\mu = 404.59$  (δειγματικός μέσος).
- $\nu = 5$ .
- $\text{Var}(t_\nu(\mu, \sigma^2)) = \sigma^2[\nu/(\nu-2)]$ , οπότε για  $\nu=5$  και δειγματική διασπορά 41.82 παίρνουμε  $\tau = 1/\sigma^2 \approx 0.04$ .

# WinBUGS

---

```
model extend;
{
mu~dnorm(0.0,1.0E-6)
tau~dgamma(0.001,0.001)
nu~dunif(2.0,12.0)

for(i in 1:n) {
y[i] ~dt(mu,tau,nu)
}

sigma <-1.0/sqrt(tau)

y.new ~ dt(mu,tau,nu)
}

#initial values

list(mu=404.59,tau=0.04, nu=5.0)

#data

list(y= c(375, 392, 393, 397, 398, 398, 399, 399, 399, 399, 399, 399, 399, 399, 399, 400, 400, 400, 400, 401,
401, 401, 401, 401, 401, 401, 401, 401, 401, 401, 401, 401, 402, 402, 402, 402, 402, 402, 402,
402, 403, 403, 403, 403, 403, 403, 403, 404, 404, 404, 404, 404, 404, 404, 404, 404, 404, 405, 405,
405, 405, 405, 406, 406, 406, 406, 406, 406, 406, 406, 406, 406, 406, 406, 406, 406, 406, 407, 407, 407,
407, 407, 407, 407, 407, 408, 408, 408, 408, 408, 408, 409, 409, 409, 409, 409, 409, 410, 410, 410,
410, 411, 412, 412, 412, 413, 415, 418, 423, 437), n=100)
```



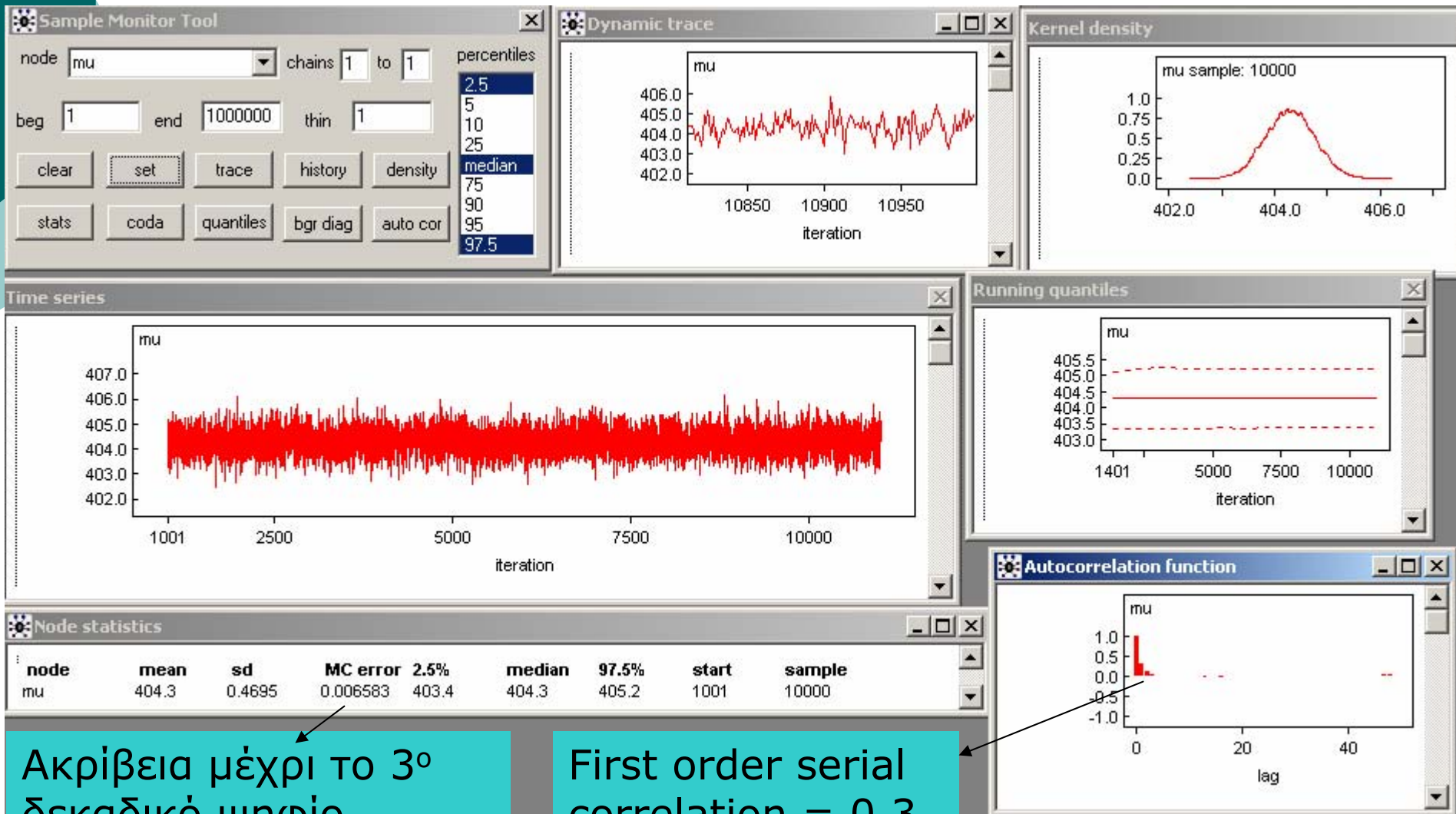


# WinBUGS

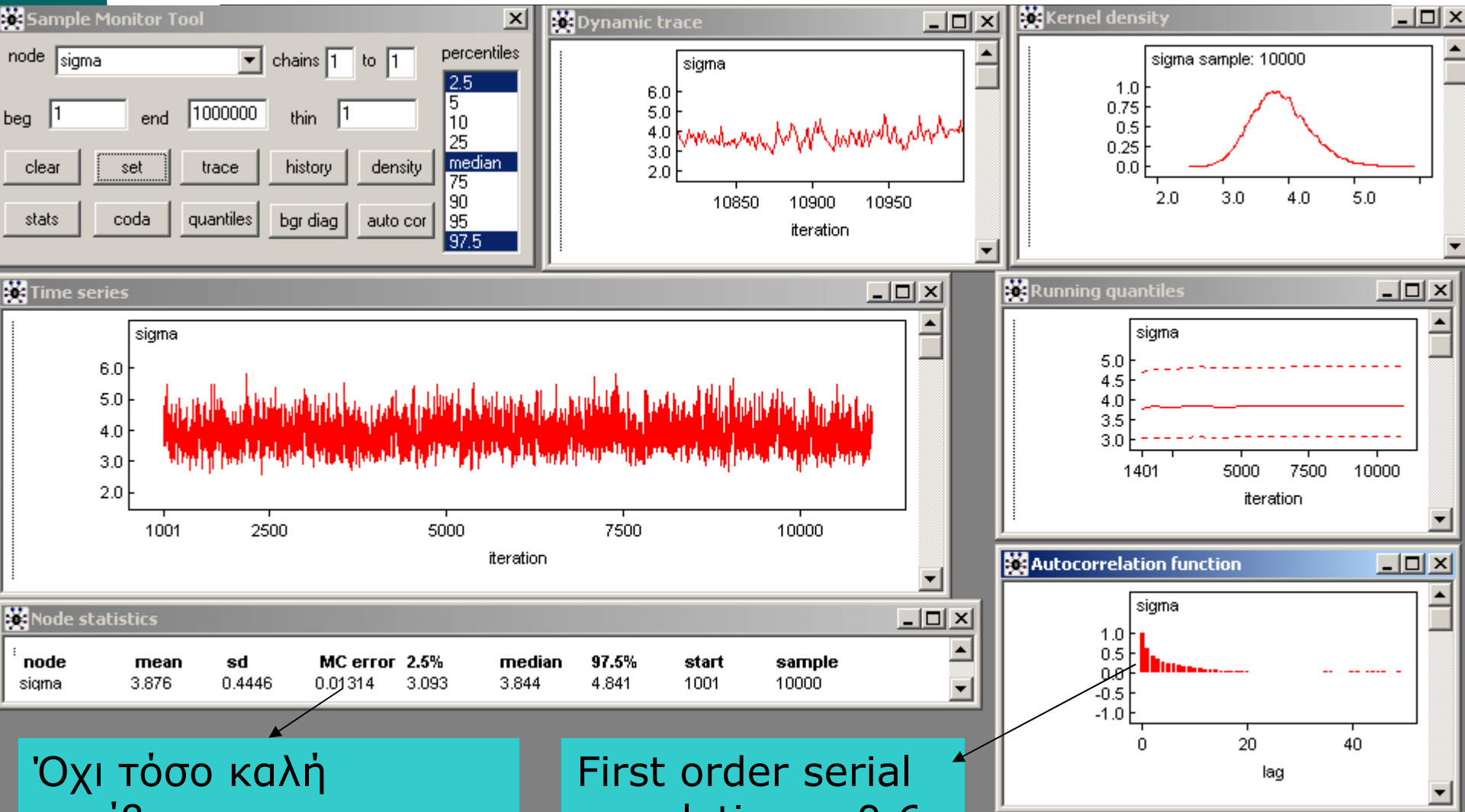
---

- Burn-in 1000 επαναλήψεις.
- 10000 επαναλήψεις (m).

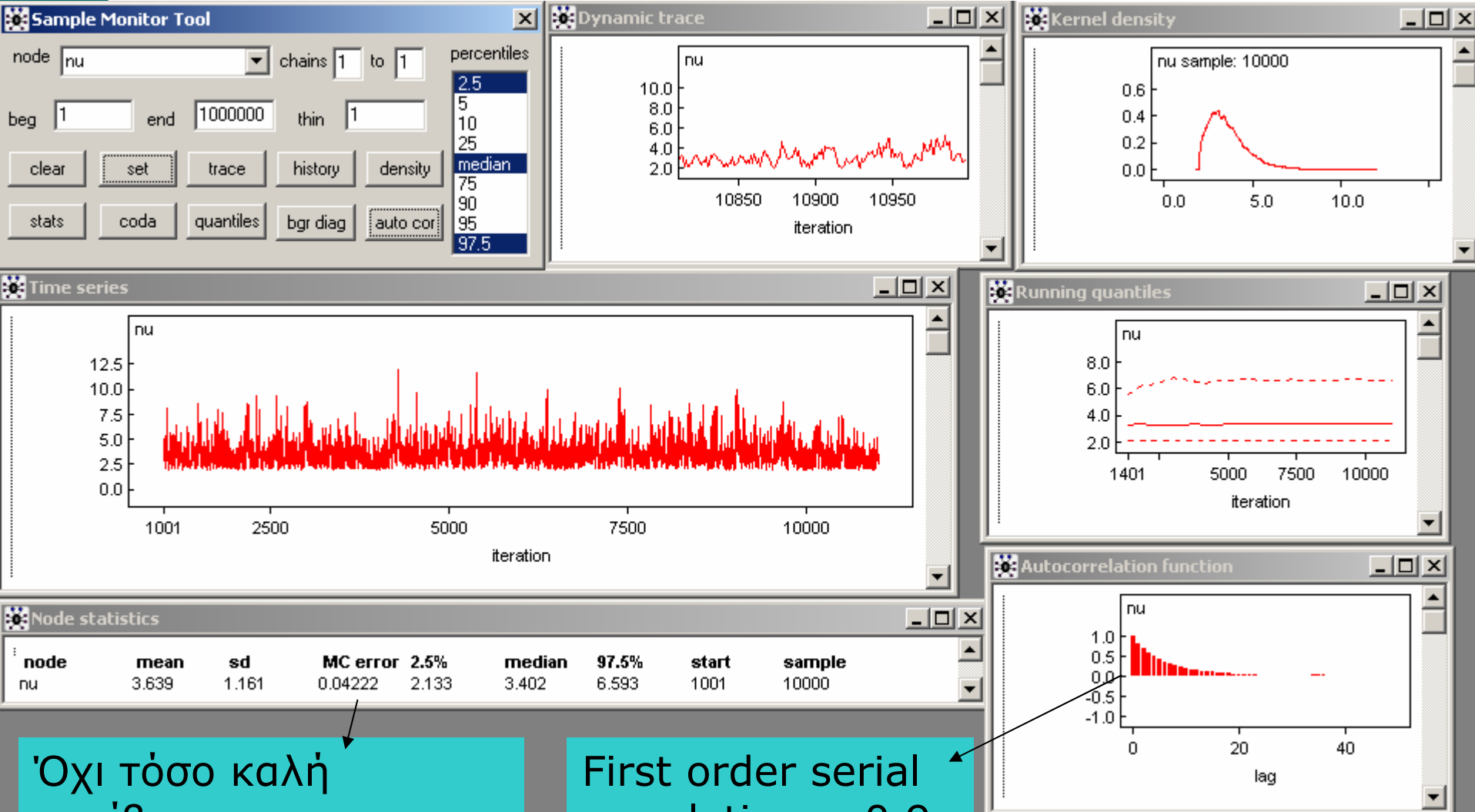
# WinBUGS



# WinBUGS



# WinBUGS



# WinBUGS

---

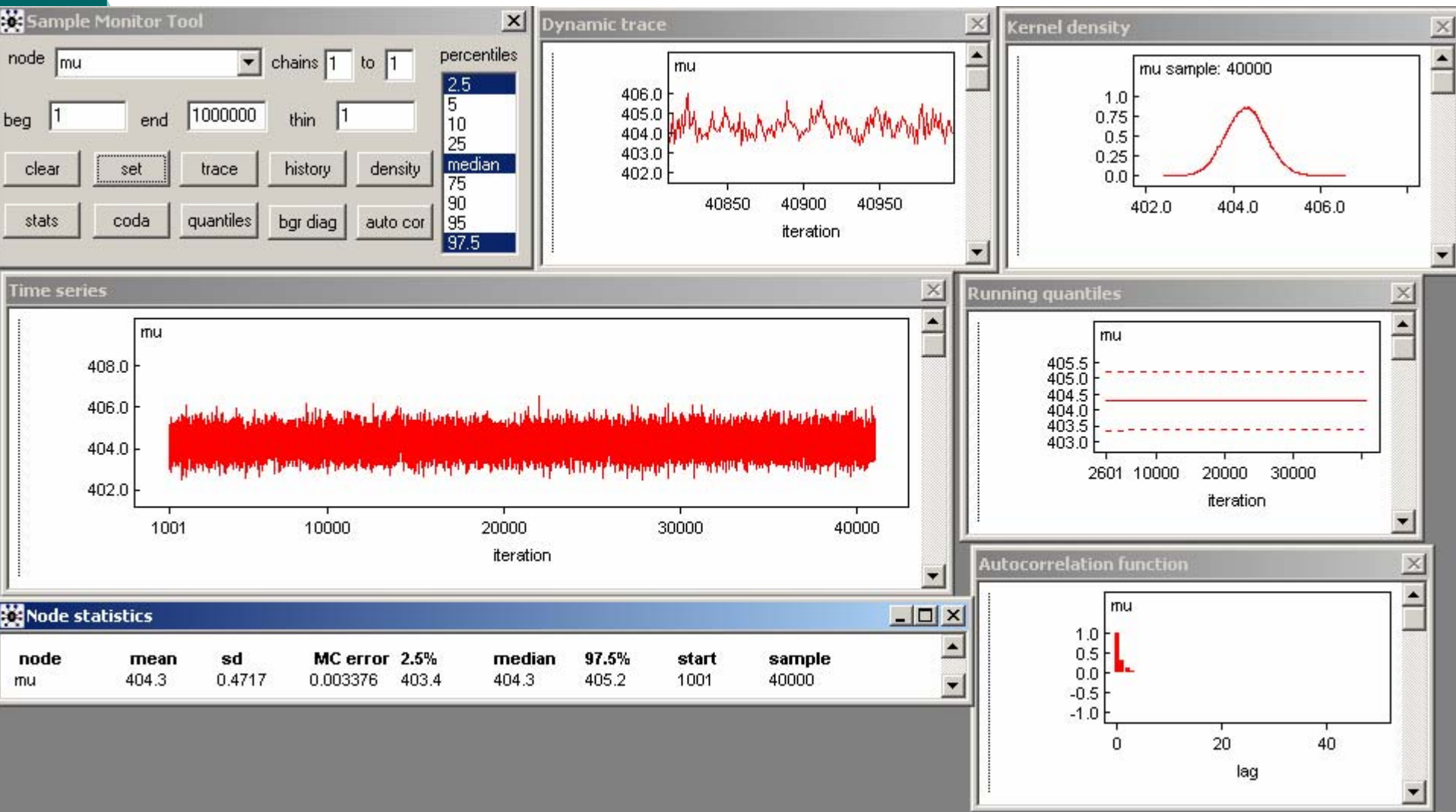
- Μπορούμε να υπολογίσουμε προσεγγιστικά το  $\rho_1$  από τον ακόλουθο τύπο (π.χ. για το  $\nu$ ):

$$\widehat{SE}(\bar{\theta}^*) = \frac{\hat{\sigma}_\theta}{\sqrt{m}} \sqrt{\frac{1+\hat{\rho}_1}{1-\hat{\rho}_1}} \Rightarrow \hat{\rho}_1 = \frac{m[\widehat{SE}(\bar{\theta}^*)]^2 - \hat{\sigma}_\theta^2}{m[\widehat{SE}(\bar{\theta}^*)]^2 + \hat{\sigma}_\theta^2} \Rightarrow$$

$$\hat{\rho}_1 = \frac{10000[0.04222]^2 - 1.161}{10000[0.04222]^2 + 1.161} \approx 0.86$$

- Παρατηρούμε ότι μίξη της αλυσίδας είναι ικανοποιητική για το  $\mu$ , και όχι τόσο καλή για το  $\sigma$  και το  $\nu$ . Για τον λόγο αυτό προσθέτω άλλες 30000 επαναλήψεις.

# WinBUGS

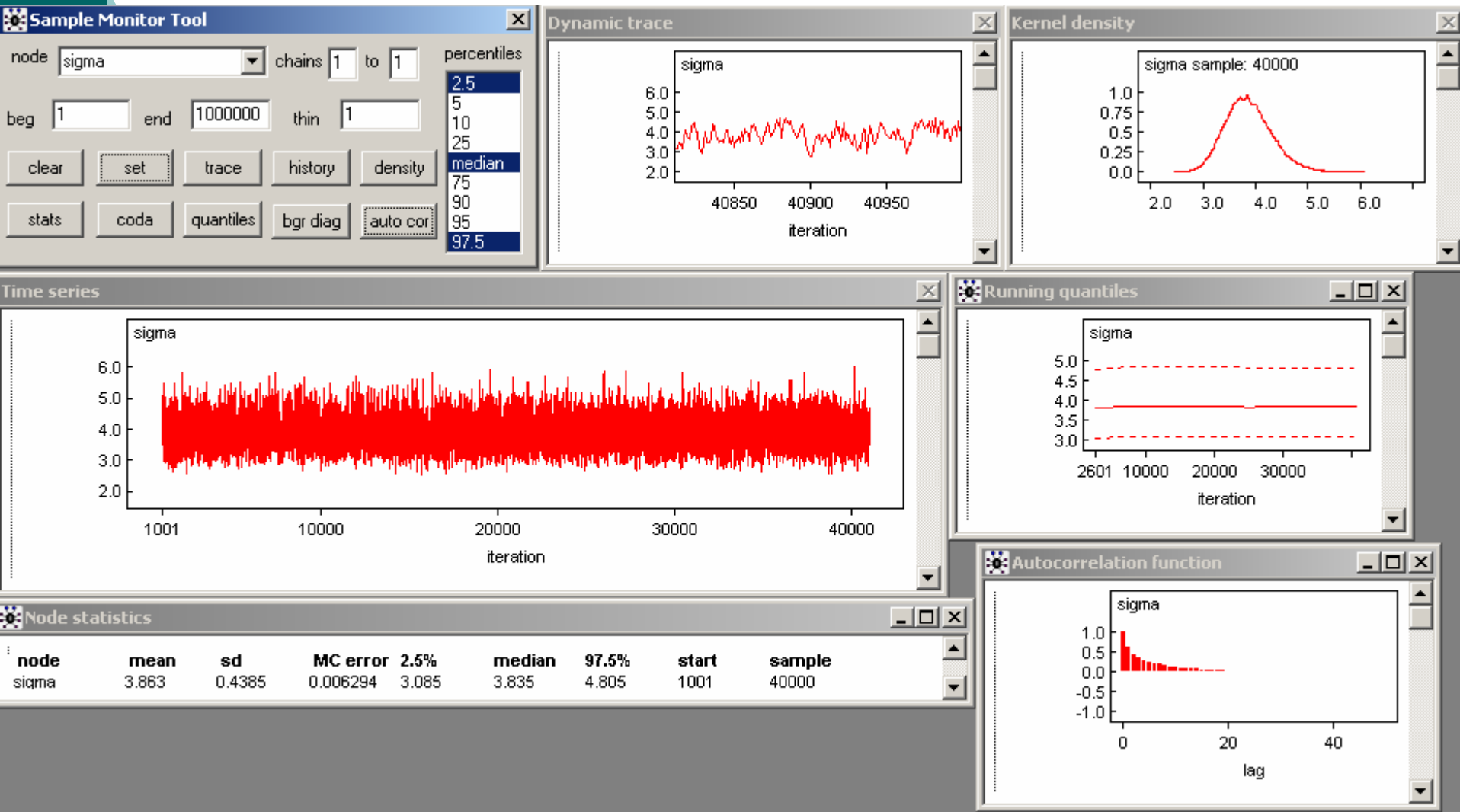


# WinBUGS

---

- Όπως αναμενότανε η μίξη της αλυσίδας είναι πολύ ικανοποιητική για το  $\mu$ . Μπορούμε να θεωρήσουμε ότι το  $\mu$  είναι γύρω στο 404.3 (με σφάλμα MCSE 0.003),  $\pm 0.47$  (posterior SD). Ένα 95% εκ των υστέρων διάστημα για το  $\mu$  είναι το (403.4, 405.2), ενώ η κατανομή του είναι συμμετρική όπως αναμενότανε.

# WinBUGS



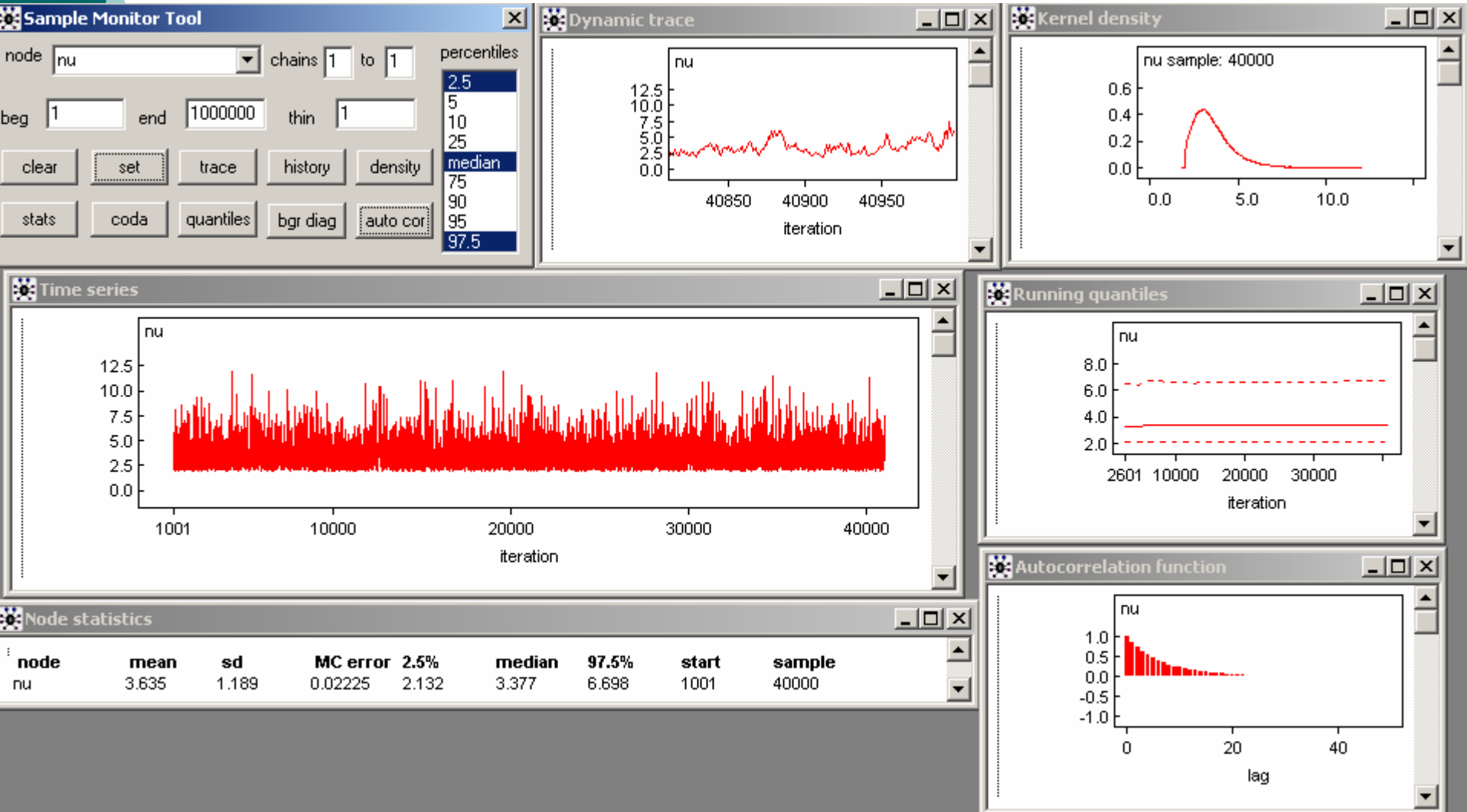


# WinBUGS

---

- Η μίξη της αλυσίδας είναι τώρα ικανοποιητική για το  $\sigma$ . Μπορούμε να θεωρήσουμε ότι το  $\sigma$  είναι γύρω στο 3.86 (με σφάλμα MCSE 0.006),  $\pm 0.44$  (posterior SD). Ένα 95% εκ των υστέρων διάστημα για το  $\sigma$  είναι το (3.08, 4.81), ενώ η κατανομή του έχει θετική ασυμμετρία (ουρά στα δεξιά) γεγονός αναμενόμενο αφού το  $\sigma$  είναι παράμετρος μεταβλητότητας.

# WinBUGS

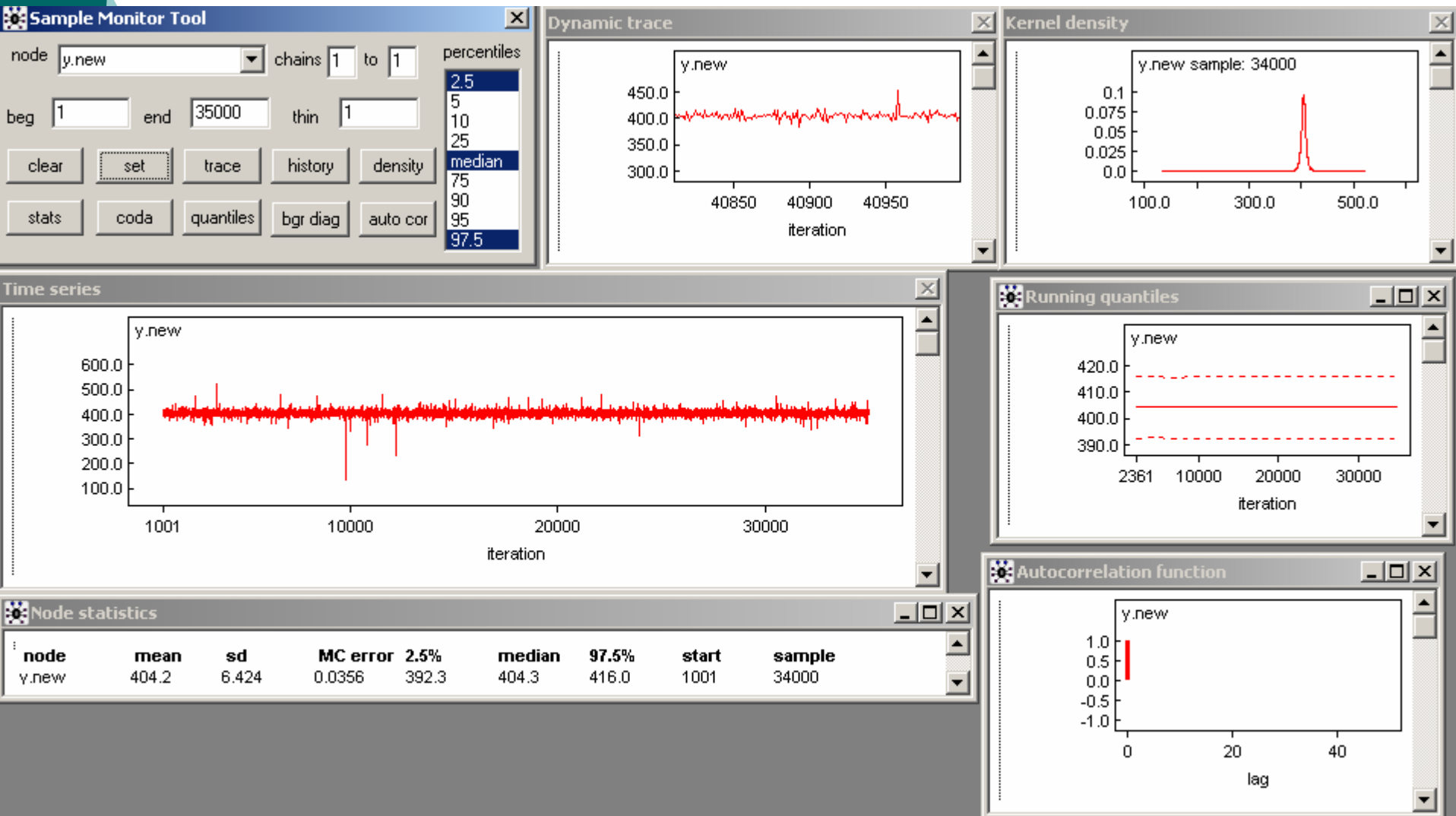


# WinBUGS

---

- Η μίξη της αλυσίδας για το  $n$  ακόμα δεν είναι ικανοποιητική. Αλλά μιας και ο σκοπός μας στην συγκεκριμένη ανάλυση είναι το  $\mu$ , και όπως θα δούμε και παρακάτω βασιζόμενοι στο γεγονός ότι τα  $\mu$  και  $n$  είναι σχεδόν ασυσχέτιστα στην εκ των υστέρων, μπορούμε να μην προσθέσουμε και άλλες επαναλήψεις.
- Αν παρά τα προαναφερθέντα θέλαμε να παρουσιάζαμε τον εκ των υστέρων μέσω του  $n$  με MCSE 0.01 (ακρίβεια δηλαδή σχεδόν στο 3<sup>ο</sup> δεκαδικό) τότε θα έπρεπε να πολλαπλασιάσουμε τον αριθμό των επαναλήψεων με τον παράγοντα  $(0.02225/0.01)^2 \approx 5$ , δηλαδή να είχαμε συνολικά 200000 επαναλήψεις.

# WinBUGS



# WinBUGS

---

- Η εκ των υστέρων προβλεπτική (**predictive**) κατανομή του  $y_{n+1}$ , έχει μέσο 404.2 και διασπορά 6.42, αρκετά κοντά δηλαδή στον δειγματικό μέσο και διασπορά (404.6 και 6.5 αντίστοιχα). Αλλά η κατανομή έχει μεγάλες ουρές (heavy tails), μιας και υπάρχουν ορισμένες προσομοιωμένες τιμές μικρότερες του 300 και μεγαλύτερες του 500, αρκετά δηλαδή απομακρυσμένες από το κέντρο των δεδομένων.

# Σύγκριση με το Κανονικό Μοντέλο

---

- Η εκ των υστέρων διασπορά για το  $\mu$ , η μόνη παράμετρος άμεσα συγκρινόμενη μεταξύ του κανονικού και του  $t$  μοντέλου για τα NB10 δεδομένα, ήταν 0.47 για το  $t$  μοντέλο και 0.64 για το κανονικό, δηλαδή το διάστημα εμπιστοσύνης για το  $\mu$  στο κανονικό μοντέλο είναι περίπου 40% ευρύτερο από αυτό στο  $t$  - μοντέλο.
- Μεταβαίνοντας από το κανονικό μοντέλο στο  $t$  αυξάνουμε την αβεβαιότητα μας, μιας και δουλεύοντας με το κανονικό μοντέλο είναι σαν να έχουμε μία  $t$  κατανομή με βαθμούς ελευθερίας  $\infty$ , ενώ αντιθέτως στο  $t$  μοντέλο το  $n$  είναι άγνωστο.
- Παρατηρούμε λοιπόν, πως παρόλο που υπάρχει αύξηση της αβεβαιότητας μας για την μετάβαση από το κανονικό στο  $t$  μοντέλο, η αβεβαιότητα μας για το  $\mu$  μειώνεται. Το γεγονός αυτό είναι αρκετά σπάνιο και συμβαίνει στην περίπτωση μας για δύο λόγους. (α) το μοντέλο  $t$  εκφράζει καλύτερα τα δεδομένα μας και (β) το κανονικό μοντέλο είναι αρκετά "συντηρητικό" όσον αφορά την ακρίβεια των παραμέτρων θέσης.