

Εισαγωγή στη θεωρία ακραίων τιμών

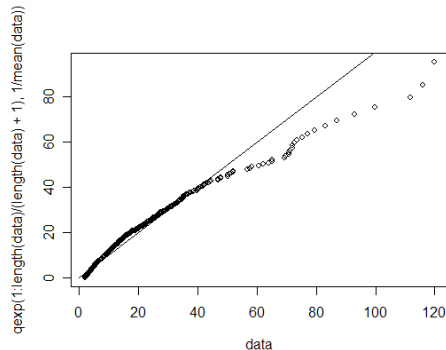
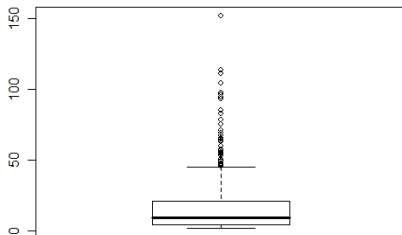
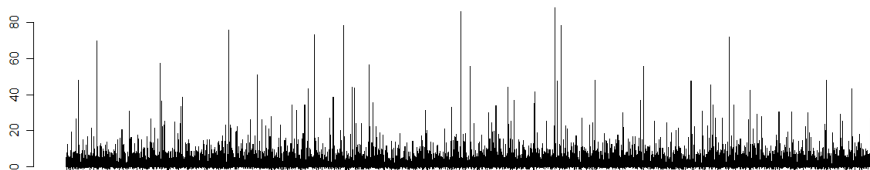
Αντικείμενο της θεωρίας ακραίων τιμών αποτελεί:

Η ανάπτυξη και μελέτη στοχαστικών μοντέλων με σκοπό την **επίλυση προβλημάτων** που σχετίζονται με την **εμφάνιση «πολύ μεγάλων» ή «πολύ μικρών» τιμών** σε τυχαία πειράματα.

Η θεωρία αυτή μπορεί χοντρικά να χωριστεί σε δύο μέρη:

- το **πιθανοθεωρητικό** μέρος που αφορά τη μελέτη της στοχαστικής συμπεριφοράς ακραίων παρατηρήσεων, και
- το **στατιστικό** μέρος που αποτελείται από μοντέλα ερμηνείας, εκτιμήσεων και προβλέψεων ακραίων συμβάντων με βάση πραγματικά δεδομένα.

Γενικό Πρόβλημα: Έχουμε τ.μ. $X_1, X_2, \dots, X_m \sim F$ (άγνωστη) και θέλουμε να μελετήσουμε τη μορφή της δεξιάς ουράς της F . Π.χ.



- Επειδή δεν γνωρίζουμε την κατανομή F από την οποία προέρχονται οι παρατηρήσεις $X_1, X_2, \dots$ θα βασιστούμε στην οριακή κατανομή G της μέγιστης παρατήρησης των $X_1, X_2, \dots, X_n$ η οποία δεν εξαρτάται από την άγνωστη F .

- Συγκεκριμένα, θέτουμε

$$M_n = \max\{X_1, X_2, \dots, X_n\}$$

η οποία όμως συγκλίνει (για $n \rightarrow \infty$) στο x_F (= **δεξιό άκρο** του στηρίγματος της F).

- Για να αποφύγουμε το παραπάνω εκφυλισμένο όριο, θεωρούμε την κανονικοποίηση

$$\frac{M_n - d_n}{c_n}$$

για κατάλληλα επιλεγμένες ακολουθίες d_n, c_n .

- Το ερωτήματα που προκύπτουν τώρα είναι :

- *πότε η παραπάνω τ.μ. συγκλίνει σε μια οριακή κατανομή G*
- *ποιά μπορεί να είναι αυτή η οριακή κατανομή G*

- Την απάντηση στο δεύτερο ερώτημα δίνει το ακόλουθο θεώρημα:

Το Θεμελιώδες Θεώρημα των *Fisher-Tippett*

Εστω $X_1, X_2, \dots$ μια ακολουθία από iid τ.μ. Αν υπάρχουν ακολουθίες $c_n > 0$, $d_n \in \mathbb{R}$ και μια μη-εκφυλισμένη κατανομή G ώστε

$$\frac{M_n - d_n}{c_n} = \frac{\max\{X_1, \dots, X_n\} - d_n}{c_n} \rightarrow_d G$$

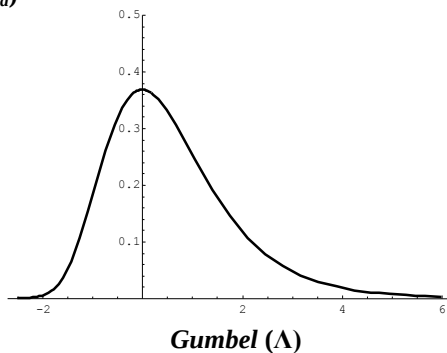
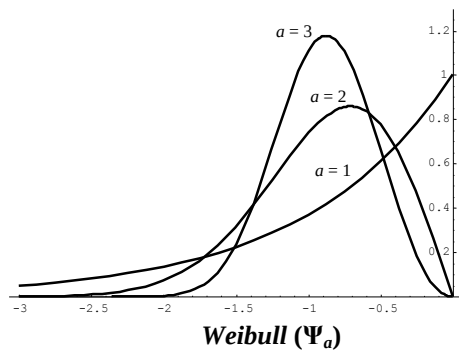
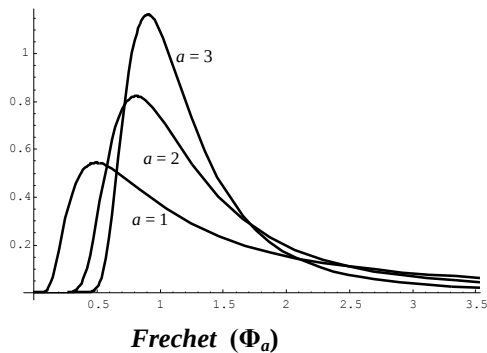
τότε η G θα είναι του ιδίου τύπου με μια από τις τρεις ακόλουθες κατανομές:

$$\text{Frechet: } \Phi_a(x) = \begin{cases} 0, & x \leq 0 \\ e^{-x^{-a}}, & x > 0 \end{cases}, \quad a > 0$$

$$\text{Weibull: } \Psi_a(x) = \begin{cases} e^{-(-x)^a}, & x \leq 0 \\ 1, & x > 0 \end{cases}, \quad a > 0$$

$$\text{Gumbel: } \Lambda(x) = e^{-e^{-x}}, \quad x \in \mathbb{R}.$$

- Οι κατανομές Φ_a , Ψ_a , Λ καλούνται **τυπικές κατανομές ακροτάτων** ενώ κατανομές που είναι του ιδίου τύπου με αυτές καλούνται **κατανομές ακροτάτων**.
- Δηλαδή $G = \Lambda((x - d)/c)$ ή $\Phi_a((x - d)/c)$ ή $\Psi_a((x - d)/c)$ για κάποια d, c .



Οι κατανομές Φ_a και Λ έχουν $x_F = \infty$, ενώ η Ψ_a έχει $x_F = 0$.

- Για τα πρώτο ερώτημα: τότε η $(M_n - d_n) / c_n$ συγκλίνει σε μια οριακή κατανομή G

Αν $X_i \sim F = 1 - \bar{F}$, τότε προκύπτουν τα παρακάτω

- Αν $x_F = \infty$ και $\bar{F}(x) = x^{-a} L(x)$ τότε

$$G = \Phi_a \text{ (Frechet)} \quad (\text{π.χ. Pareto, Cauchy, Loggamma} \in \text{MDA}(\Phi_a))$$

(L : συνάρτηση βραδείας κύμανσης, δηλαδή $L(tx) / L(x) \rightarrow 1$, για κάθε $t > 0$)

- Αν $x_F < \infty$ και $\bar{F}(x_F - x^{-1}) = x^{-a} L(x)$ (ισοδύναμα, $\bar{F}(y) = (x_F - y)^a L(\frac{1}{x_F - y})$)

$$G = \Psi_a \text{ (Weibull)} \quad (\text{π.χ. Uniform, Beta} \in \text{MDA}(\Psi_a))$$

- Αν $\bar{F}(x) = c(x) e^{-\int_z^x \frac{g(t)}{a(t)} dt}$ ($c(x) \rightarrow c_0, g(x) \rightarrow 1, a'(x) = 0$, για $x \uparrow x_F$), π.χ. $\bar{F}(x) = c(x) e^{-cx}$ τότε

$$G = \Lambda \text{ (Gumbel)} \quad (\text{π.χ. Gamma, Normal, Lognormal, Exponential} \in \text{MDA}(\Lambda))$$

ΚΕΦ. 2 Στατιστική ανάλυση ακραίων παρατηρήσεων

(μοντέλα ερμηνείας, εκτιμήσεων - προβλέψεων ακραίων συμβάντων με βάση πραγματικά δεδομένα)

- Θα προσπαθήσουμε, με βάση ιστορικά δεδομένα και μόνο, να δώσουμε απαντήσεις σε ερωτήσεις της μορφής:

- Ποια είναι η πιθανότητα τα επόμενα a χρονικά διαστήματα οι παρατηρήσεις να μην υπερβούν ένα προκαθορισμένο υψηλό κατώφλι u ;

- Ποια είναι η πρόβλεψη για την μεγαλύτερη παρατήρηση στα επόμενα a χρονικά διαστήματα;

- Ποια είναι η στάθμη u την οποία δεν θα υπερβεί καμία παρατήρηση στα επόμενα a χρονικά διαστήματα με πιθανότητα p ;

Κ.Ο.Κ.

- Οι βασικές μέθοδοι είναι:

1. Μέθοδος **Block Maxima** (βασίζομαστε στις μέγιστες τιμές ανά χρονικές περιόδους)

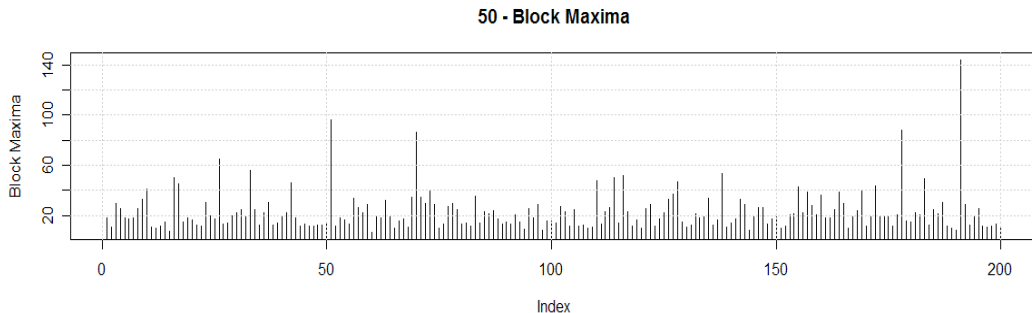
2. Μέθοδος **POT** (βασίζομαστε στις τιμές πάνω από ένα κατώφλι)

1. Η Μέθοδος Block Maxima

- Έστω παρατηρήσεις $X_1, X_2, \dots, X_m \sim F$ (άγνωστη)
- Μπορούμε να θεωρήσουμε ότι οι $X_1, X_2, \dots, X_m$, $m = kn$ χωρίζονται σε k το πλήθος υποσύνολα (*blocks*) από n παρατηρήσεις το καθένα.

$$X_1, X_2, \dots, X_n \mid X_{n+1}, X_{n+2}, \dots, X_{n+n} \mid \dots \mid X_{(k-1)n+1}, X_{(k-1)n+2}, \dots, X_{(k-1)n+n}$$

Συμβολίζουμε με $Y_1, Y_2, \dots, Y_k$, τις μέγιστες τιμές σε καθένα από αυτά τα k υποσύνολα, οι οποίες καλούνται **block maxima**



($n = 50, k = 200$)

- Από το Θ . *Fisher-Tippett*, εάν το n είναι αρκετά μεγάλο τότε

(κανονικοποιημένα) block maxima $\sim \Lambda((x - d)/c)$ ή

$\Phi_a((x - d)/c)$ ή

$\Psi_a((x - d)/c)$

ανάλογα με το αν $F \in \text{MDA}(\Lambda)$ ή $\text{MDA}(\Phi_a)$ ή $\text{MDA}(\Psi_a)$ αντίστοιχα.

(I) Αν $F \in \text{MDA}(\Lambda)$ τότε $x_F < \infty$ ή $x_F = \infty$ και η F έχει λεπτή ή μέτρια βαριά ουρά (η ουρά της ίδιας της Λ μειώνεται εκθετικά).

(II) Αν $F \in \text{MDA}(\Phi_a)$ τότε $x_F = \infty$ και η F έχει βαριά ουρά. Μάλιστα $\bar{F}(x) = x^{-a} L(x)$

(III) Αν $F \in \text{MDA}(\Psi_a)$ τότε $x_F < \infty$ με ουρά $\bar{F}(x) = (x_F - x)^a L(\frac{1}{x_F - x})$.

- Μια πρώτη προσέγγιση:

(1) **Προσδιορίζουμε** από ποια από τις τρεις κατανομές ακροτάτων προέρχονται τα block maxima $Y_1, Y_2, \dots, Y_k$ (ανάλογα με το MDA που ανήκουν τα X_i)

(2) **Εκτιμούμε** τις παραμέτρους a (σχήματος), d (θέσης) και c (κλίμακας).

ΠΑΡΑΤΗΡΗΣΗ: Η παραπάνω διαδικασία ακολουθείται όταν υπάρχει βεβαιότητα για τον τύπο της οριακής κατανομής (Λ , Φ ή Ψ) των X_i . Διαφορετικά, χρειάζεται μια μέθοδος για την επιλογή της κατάλληλου τύπου της οριακής κατανομής ακροτάτων.

ΜΕΙΟΝΕΚΤΗΜΑ: Μια λανθασμένη επιλογή του τύπου της οριακής κατανομής ακροτάτων μπορεί να οδηγήσει στη συνέχεια σε εντελώς εσφαλμένα αποτελέσματα.

- Συνήθως ακολουθείται μια **ασφαλέστερη** διαδικασία:
 - Οι τρεις οικογένειες κατανομών ακροτάτων ενοποιούνται **σε μία οικογένεια κατανομών** η οποία καλείται κατανομή **GEV**
 - Αν το n είναι αρκετά μεγάλο, μπορούμε να θεωρήσουμε ότι τα block maxima $Y_1, Y_2, \dots, Y_k \sim \text{GEV}$ και να *εκτιμήσουμε* τις παραμέτρους της.

ΠΛΕΟΝΕΚΤΗΜΑΤΑ:

- (i) Δεν είμαστε αναγκασμένοι να επιλέξουμε a-priori κάποιον από τους τρεις τύπους, Λ, Φ, Ψ , και να παραμείνουμε σε αυτόν.
- (ii) Η αβεβαιότητα που ενδεχομένως θα υπάρχει για τον τύπο της οριακής κατανομής ακροτάτων θα εκφράζεται φυσιολογικά μέσα από την αβεβαιότητα της εκτίμησης της παραμέτρου σχήματος ξ (π.χ. από το εύρος διαστήματος εμπιστοσύνης για το ξ).

Θεώρημα *Fisher-Tippett* (ενοποιώντας τους τρεις τύπους κατανομών ακροτάτων)

Έστω $X_1, X_2, \dots$ μια ακολουθία από ανεξ. ισον. (iid) τ.μ. Αν υπάρχουν ακολουθίες $c_n > 0$, $d_n \in \mathbb{R}$ και μια μη-εκφυλισμένη κατανομή G ώστε

$$\frac{M_n - d_n}{c_n} = \frac{\max\{X_1, \dots, X_n\} - d_n}{c_n} \rightarrow_d G$$

τότε η G θα είναι της μορφής

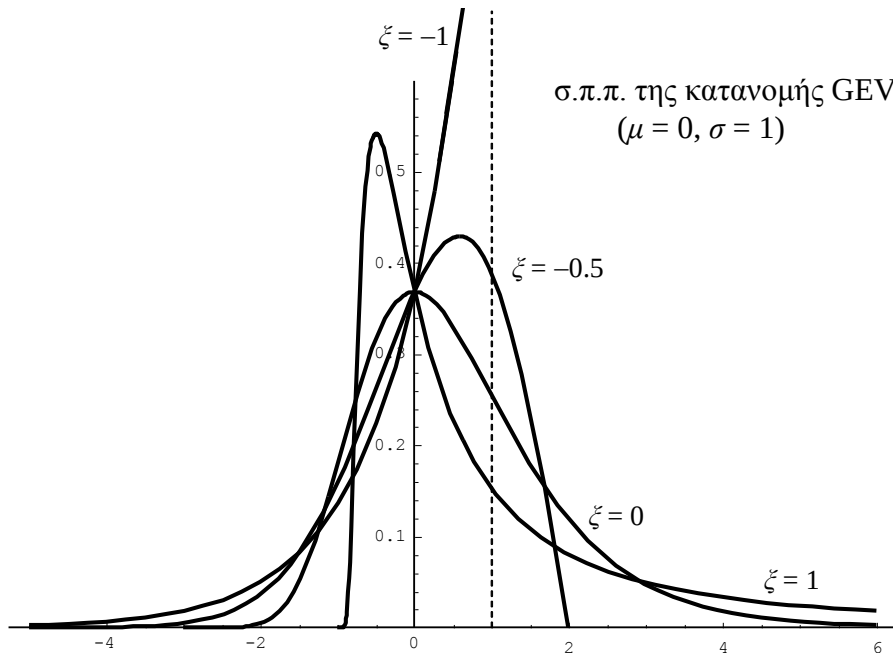
$$G(x; \mu, \sigma, \xi) = e^{-\left(1 + \xi \frac{x - \mu}{\sigma}\right)^{\frac{1}{\xi}}}, \quad x: 1 + \xi \frac{x - \mu}{\sigma} > 0$$

όπου $\mu, \xi \in \mathbb{R}$, $\sigma > 0$.

Για $\xi = 0$ θεωρούμε ως G το όριο της για $\xi \rightarrow 0$, δηλαδή,

$$G(x, \mu, \sigma, 0) = e^{-e^{\frac{x - \mu}{\sigma}}}, \quad x \in \mathbb{R}.$$

Η συγκεκριμένη παραμετρική οικογένεια κατανομών με σ.κ. G καλείται **κατανομή GEV** (*Generalized Extreme Value Distribution*) με παραμέτρους $\mu, \xi \in \mathbb{R}, \sigma > 0$.



- Για κατάλληλες τιμές των παραμέτρων μ, σ, ξ προκύπτουν και πάλι οι τρεις τύποι κατανομών ακροτάτων:

(I) (**Frechet**) Αν $\xi > 0$ και θέσουμε $\xi = 1/a, \sigma = c/a, \mu = c + d$, τότε

$$G(x) = e^{-\left(1 + \xi \frac{x-\mu}{\sigma}\right)^{\frac{1}{\xi}}} = e^{-\left(\frac{x-d}{c}\right)^a} = \Phi_a\left(\frac{x-d}{c}\right), \quad \frac{x-d}{c} > 0 \Leftrightarrow 1 + \xi \frac{x-\mu}{\sigma} > 0$$

(II) (**Weibull**) Αν $\xi < 0$ και θέσουμε $\xi = -1/a, \sigma = c/a, \mu = -c + d$, τότε

$$G(x) = e^{-\left(1 + \xi \frac{x-\mu}{\sigma}\right)^{\frac{1}{\xi}}} = e^{-\left(-\frac{x-d}{c}\right)^a} = \Psi_a\left(\frac{x-d}{c}\right), \quad \frac{x-d}{c} < 0 \Leftrightarrow 1 + \xi \frac{x-\mu}{\sigma} > 0$$

(III) (**Gumbel**) Αν $\xi = 0$ (δηλ. θεωρήσουμε ότι $\xi \rightarrow 0$) και θέσουμε $\sigma = c, \mu = d$, τότε

$$G(x) = e^{-\left(1 + \xi \frac{x-\mu}{\sigma}\right)^{\frac{1}{\xi}}} \xrightarrow{\xi \rightarrow 0} e^{-e^{-\frac{x-\mu}{\sigma}}} = \Lambda\left(\frac{x-d}{c}\right), \quad x \in \mathbb{R}$$

Επομένως η κατανομή GEV:

- για $\xi > 0$ έχει $x_F = \infty$, με **βαριά ουρά**: $\overline{G}(x) = x^{-1/\xi} L(x)$ (αν $\xi \uparrow$, γίνεται βαρύτερη)

- Αν π.χ. $\xi = 0.5$ τότε $\overline{G}(x) = \frac{L(x)}{x^2}$ όπου $L \in \mathfrak{R}_0$ (π.χ. σταθερή)

- Αν π.χ. $\xi = 2$ τότε $\overline{G}(x) = \frac{L(x)}{\sqrt{x}}$ όπου $L \in \mathfrak{R}_0$ (π.χ. σταθερή)

- για $\xi = 0$ έχει $x_F = \infty$, με **λεπτή ουρά**: $\overline{G}(x) = c(x)e^{-x}$ (όπου $c(x) \rightarrow 1$ για $x \rightarrow \infty$)

- για $\xi < 0$ έχει $x_F = \mu - \sigma/\xi < \infty$, με ουρά:

$$\overline{G}(x_F - x^{-1}) = x^{1/\xi} L(x) \Leftrightarrow \overline{G}(y) = (x_F - y)^{-1/\xi} L\left(\frac{1}{x_F - y}\right)$$

- Αν π.χ. $\xi = -0.5$ τότε $\overline{G}(y) = (x_F - y)^2 L\left(\frac{1}{x_F - y}\right)$ όπου $L \in \mathfrak{R}_0$

- Αν π.χ. $\xi = -2$ τότε $\overline{G}(y) = \sqrt{x_F - y} L\left(\frac{1}{x_F - y}\right)$ όπου $L \in \mathfrak{R}_0$

- Από το Θεώρημα Fisher-Tippett: Τα κανονικοποιημένα *block maxima*,

$$\frac{Y_1 - d_n}{c_n}, \frac{Y_2 - d_n}{c_n}, \dots, \frac{Y_k - d_n}{c_n} \sim \text{GEV}$$

Δηλαδή,

$$P\left(\frac{Y_i - d_n}{c_n} \leq x\right) = P\left(\frac{M_n^{(i)} - d_n}{c_n} \leq x\right) \approx G(x; \mu_0, \sigma_0, \xi),$$

- Από την παραπάνω σχέση, θέτοντας $y = c_n x + d_n$, προκύπτει ότι

$$P(Y_i \leq y) \approx G\left(\frac{y - d_n}{c_n}; \mu_0, \sigma_0, \xi\right) = e^{-\left(1 + \xi \frac{y - d_n - c_n \mu_0}{c_n \sigma_0}\right)^{\frac{1}{\xi}}} = G(y; \mu_n, \sigma_n, \xi)$$

όπου $\mu_n = d_n + c_n \mu_0$, $\sigma_n = c_n \sigma_0$, και άρα και τα μη-κανονικοποιημένα *block maxima*

$$Y_1, Y_2, \dots, Y_k \sim \text{GEV}.$$

Συνοψίζουμε: Προκειμένου να μελετήσουμε την κατανομή της μέγιστης παρατήρησης $M_n = \max\{X_1, X_2, \dots, X_m\}$ πράττουμε τα εξής:

- Χωρίζουμε τα $X_1, X_2, \dots, X_{kn}$ σε k ομάδες (blocks) και λαμβάνουμε τις μέγιστες παρατηρήσεις $Y_i = M_n^{(i)}, i = 1, 2, \dots, k$, μέσα σε κάθε ομάδα (block maxima).

$$\begin{array}{ccccccc} X_1, X_2, \dots, X_n & | & X_{n+1}, X_{n+2}, \dots, X_{n+n} & | & \dots & | & X_{(k-1)n+1}, X_{(k-1)n+2}, \dots, X_{(k-1)n+n} \\ Y_1 & & Y_2 & & \dots & & Y_k \end{array}$$

- Αν το n είναι αρκετά μεγάλο, τα **block maxima**

$$Y_1, Y_2, \dots, Y_k \sim \mathbf{GEV}(\mu_n, \sigma_n, \xi)$$

- **Εκτιμούμε** τις παραμέτρους μ_n, σ_n, ξ της GEV (οι οποίες, όπως είδαμε, «περιέχουν» και τις σταθερές κανονικοποίησης).

- Όταν έχουμε ένα συγκεκριμένο σύνολο δεδομένων το n θεωρείται σταθερό. Για το λόγο αυτό στη συνέχεια θα γράφουμε τις παραμέτρους μ_n, σ_n χωρίς τον δείκτη n .

Εκτίμηση των παραμέτρων της GEV

- Έχουν προταθεί αρκετές μέθοδοι για αυτό το σκοπό:
 - μέσω κατάλληλων γραφημάτων (παραλλαγές των Q-Q plots),
 - με τη μέθοδο των ροπών
 - μέσω κατάλληλων συναρτήσεων διατεταγμένων παρατηρήσεων (*order statistics*)
 - μέσω εκτιμητριών *μέγιστης πιθανοφάνειας*.

- Θα βασιστούμε κυρίως στη μέθοδο *μέγιστης πιθανοφάνειας* που ασυμπτωτικά, δίνει τις καλύτερες δυνατές εκτιμήτριες (ελαχίστης διασποράς) που ακολουθούν γνωστή κατανομή (πολυδιάστατη κανονική).

- Οι έλεγχοι υποθέσεων και τα δ.ε. για τις παραμέτρους μπορούν σχετικά εύκολα να γίνουν χρησιμοποιώντας τη γνωστή (profile) Deviance Function.

Ε.Μ.Π. των μ, σ, ξ της κατανομής GEV.

- Έστω τα block maxima $Y_1, Y_2, \dots, Y_k \sim \text{GEV}$. Η log-likelihood των Y_i , θα είναι ($\xi \neq 0$)

$$\begin{aligned} l(\mu, \sigma, \xi) &= \ln \prod_{i=1}^k G'(Y_i; \mu, \sigma, \xi) = \sum_{i=1}^k \ln G'(Y_i; \mu, \sigma, \xi) \\ &= \sum_{i=1}^k \ln I(Y_i, \mu, \sigma, \xi) - k \ln \sigma - \left(\frac{1}{\xi} + 1\right) \sum_{i=1}^k \ln \left(1 + \xi \frac{Y_i - \mu}{\sigma}\right) - \sum_{i=1}^k \left(1 + \xi \frac{Y_i - \mu}{\sigma}\right)^{-\frac{1}{\xi}} \end{aligned}$$

(Το άθροισμα των $\ln I(Y_i, \mu, \sigma, \xi)$ είναι ίσο με 0 αν τα μ, σ, ξ , ικανοποιούν την $1 + \xi(Y_i - \mu)/\sigma > 0$ για όλα τα Y_i , και $-\infty$ διαφορετικά).

- Στην περίπτωση που $\xi = 0$ αντίστοιχα θα είναι

$$l(\mu, \sigma, 0) = -k \ln \sigma - \sum_{i=1}^k \frac{Y_i - \mu}{\sigma} - \sum_{i=1}^k e^{-\frac{Y_i - \mu}{\sigma}}.$$

- Οι ε.μ.π. $\hat{\mu}, \hat{\sigma}, \hat{\xi}$ είναι οι τιμές των μ, σ, ξ που μεγιστοποιούν την $l(\mu, \sigma, \xi)$.
- Δυστυχώς, **δεν υπάρχουν αναλυτικές εκφράσεις για τις ε.μ.π.** Δεδομένου όμως ενός συγκεκριμένου τυχαίου δείγματος $y_1, \dots, y_k$ μπορούμε να βρούμε τις τιμές που μεγιστοποιούν την $l(\mu, \sigma, \xi)$ με **αριθμητικές μεθόδους βελτιστοποίησης** (π.χ. Newton-Raphson)

Ιδιότητες των ε.μ.π.

- Οι $\hat{\mu}, \hat{\sigma}, \hat{\xi}$ θα ακολουθούν ασυμπτωτικά μία *πολυδιάστατη κανονική* με μέση τιμή $(\mu, \sigma, \xi)^T$ και πίνακα διασποράς, τον αντίστροφο του πίνακα πληροφορίας, ο οποίος εκτιμάται από τον

$$\hat{I} = - \begin{pmatrix} \frac{\partial^2}{\partial \mu^2} l(\hat{\mu}, \hat{\sigma}, \hat{\xi}) & \frac{\partial^2}{\partial \mu \partial \sigma} l(\hat{\mu}, \hat{\sigma}, \hat{\xi}) & \frac{\partial^2}{\partial \mu \partial \xi} l(\hat{\mu}, \hat{\sigma}, \hat{\xi}) \\ \frac{\partial^2}{\partial \sigma \partial \mu} l(\hat{\mu}, \hat{\sigma}, \hat{\xi}) & \frac{\partial^2}{\partial \sigma^2} l(\hat{\mu}, \hat{\sigma}, \hat{\xi}) & \frac{\partial^2}{\partial \sigma \partial \xi} l(\hat{\mu}, \hat{\sigma}, \hat{\xi}) \\ \frac{\partial^2}{\partial \xi \partial \mu} l(\hat{\mu}, \hat{\sigma}, \hat{\xi}) & \frac{\partial^2}{\partial \xi \partial \sigma} l(\hat{\mu}, \hat{\sigma}, \hat{\xi}) & \frac{\partial^2}{\partial \xi^2} l(\hat{\mu}, \hat{\sigma}, \hat{\xi}) \end{pmatrix}$$

- Αν $s_{i,i}$ είναι το i -διαγώνιο στοιχείο του $\hat{I}^{-1}$, προσεγγιστικά, ισχύει ότι

$$\hat{\mu} \sim N(\mu, s_{1,1}), \quad \hat{\sigma} \sim N(\sigma, s_{2,2}), \quad \hat{\xi} \sim N(\xi, s_{3,3}),$$

και άρα τα

$$(\hat{\mu} \pm \sqrt{s_{1,1}} z_{a/2}), (\hat{\sigma} \pm \sqrt{s_{2,2}} z_{a/2}), (\hat{\xi} \pm \sqrt{s_{3,3}} z_{a/2})$$

είναι (προσεγγιστικά) διαστήματα εμπιστοσύνης συντελεστού $1 - a$ για τις παραμέτρους μ, σ, ξ αντίστοιχα.

Profile Deviance Function

- Για την κατασκευή δ.ε., μπορούμε εναλλακτικά να χρησιμοποιήσουμε τις *profile Deviance Functions* $D_p(\mu)$, $D_p(\sigma)$, $D_p(\xi)$.

- Αν θέλουμε δ.ε. για το ξ , αποδεικνύεται ότι (ασυμπτωτικά)

$$D_p(\xi) = 2(l(\hat{\mu}, \hat{\sigma}, \hat{\xi}) - \max_{\mu, \sigma} l(\mu, \sigma, \xi)) \sim \chi_1^2.$$

και επομένως, το σύνολο

$$\{\xi : D_p(\xi) \leq \chi_1^2(a)\} = \{\xi : \max_{\mu, \sigma} l(\mu, \sigma, \xi) \geq l(\hat{\mu}, \hat{\sigma}, \hat{\xi}) - \chi_1^2(a)/2\},$$

είναι ένα δ.ε. συντελεστού $1 - a$ για το ξ .

- Το παραπάνω δεν λύνεται ως προς ξ και για αυτό κατασκευάζεται το γράφημα της συνάρτησης $\max_{\mu, \sigma} l(\mu, \sigma, \xi)$ για διάφορες τιμές του ξ σε ένα διάστημα γύρω από το $\hat{\xi}$ και προστίθεται στο η οριζόντια γραμμή με τεταγμένη $l(\hat{\mu}, \hat{\sigma}, \hat{\xi}) - \chi_1^2(a)/2$.

ΣΤΑΘΜΗ ΑΠΟΔΟΣΗΣ

- **Στάθμη απόδοσης z_p (για $1/p$ χρονικές περιόδους απόδοσης):** το κατώφλι z_p το οποίο υπερβαίνει ένα block maximum με πιθανότητα p .

$$z_p = G^{-1}(1-p) = \begin{cases} \mu - \frac{\sigma}{\xi} (1 - (-\ln(1-p))^{-\xi}), & \xi \neq 0 \\ \mu - \sigma \ln(-\ln(1-p)), & \xi = 0 \end{cases}$$

Ισοδύναμα: Το κατώφλι z_p το οποίο υπερβαίνουν τα block maxima κατά μέσο όρο κάθε $1/p$ blocks.

ΕΚΤΙΜΗΣΗ ΤΗΣ ΣΤΑΘΜΗΣ ΑΠΟΔΟΣΗΣ

- **Σημειακή Εκτίμηση.** Ισχύει ότι $z_p = G^{-1}(1-p; \mu, \sigma, \xi) := g(\mu, \sigma, \xi)$. Επομένως

$$\hat{z}_p = g(\hat{\mu}, \hat{\sigma}, \hat{\xi}) = \begin{cases} \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}} (1 - e^{\hat{\xi} w_p}), & \hat{\xi} \neq 0 \\ \hat{\mu} + \hat{\sigma} w_p, & \hat{\xi} = 0 \end{cases}, \quad w_p = -\ln(-\ln(1-p)).$$

Δ.Ε. ΓΙΑ ΤΟ z_p ΜΕΣΩ ΤΗΣ PROFILE DEVIANCE FUNCTION

- Πραγματοποιούμε νέα παραμετροποίηση ώστε η z_p να είναι μια από τις παραμέτρους της GEV (π.χ. θέτουμε $\mu = z_p + \frac{\sigma}{\xi}(1 - (-\ln(1-p))^{-\xi})$).
- Αν $l'(z_p, \sigma, \xi)$ η πιθανοφάνεια των block maxima ως προς την νέα παραμετροποίηση, τότε

$$D_p(z_p) = 2(l'(\hat{z}_p, \hat{\sigma}, \hat{\xi}) - \max_{\sigma, \xi} l'(z_p, \sigma, \xi)) \sim \chi_1^2$$

- Συνεπώς, το σύνολο

$$\{z_p : D_p(z_p) \leq \chi_1^2(a)\} = \{z_p : \max_{\mu, \sigma} l'(z_p, \sigma, \xi) \geq l'(\hat{z}_p, \hat{\sigma}, \hat{\xi}) - \chi_1^2(a)/2\}$$

είναι ένα δ.ε. συντελεστού $1 - a$ για το z_p .

- Συνήθως κατασκευάζεται το γράφημα της $\max_{\mu, \sigma} l'(z_p, \sigma, \xi)$ και προστίθεται η οριζόντια γραμμή με τεταγμένη $l'(\hat{z}_p, \hat{\sigma}, \hat{\xi}) - \chi_1^2(a)/2$.

ΕΛΕΓΧΟΣ ΚΑΛΗΣ ΠΡΟΣΑΡΜΟΓΗΣ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΣΤΗΝ GEV

- Θα πρέπει να βεβαιωθούμε ότι τα block maxima προσαρμόζονται ικανοποιητικά σε κάποια GEV, ώστε να έχει νόημα η εκτίμηση των παραμέτρων μ, σ, ξ .
- Χρησιμοποιούμε Q-Q plot (ή P-P plot) των block maxima ως προς την GEV $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ το οποίο αποτελείται από τα σημεία

$$(Y_{(i)}, G^{-1}(\frac{i}{k+1}; \hat{\mu}, \hat{\sigma}, \hat{\xi})) = (Y_{(i)}, \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}} (1 - (-\ln \frac{i}{k+1})^{-\hat{\xi}})), i = 1, 2, \dots, k$$

όπου $Y_{(1)} < Y_{(2)} < \dots < Y_{(k)}$ είναι τα διατεταγμένα block maxima.

- Εάν τα block maxima ακολουθούν μία GEV θα πρέπει τα k σημεία του Q-Q plot να βρίσκονται κοντά στην διαγώνιο.

ΕΚΤΙΜΗΣΗ ΤΩΝ μ, σ, ξ ΜΕΣΩ Q-Q PLOT

- Το Q-Q plot των block maxima ως προς την $GEV(0,1,\hat{\xi})$ αποτελείται από τα σημεία

$$(Y_{(i)}, -\frac{1}{\hat{\xi}}(1 - (-\ln \frac{i}{k+1})^{-\hat{\xi}})), i = 1, 2, \dots, k$$

- Εάν τα block maxima ακολουθούν μία GEV θα πρέπει τα k σημεία του παραπάνω Q-Q plot να βρίσκονται «κοντά» στην ευθεία:

$$y = (x - \mu)/\sigma.$$

- Δηλαδή **τα μ, σ μπορούν πρόχειρα να εκτιμηθούν** από το γράφημα

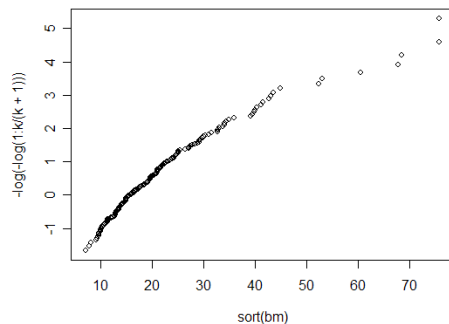
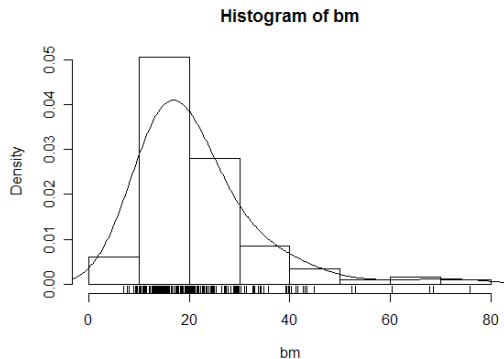
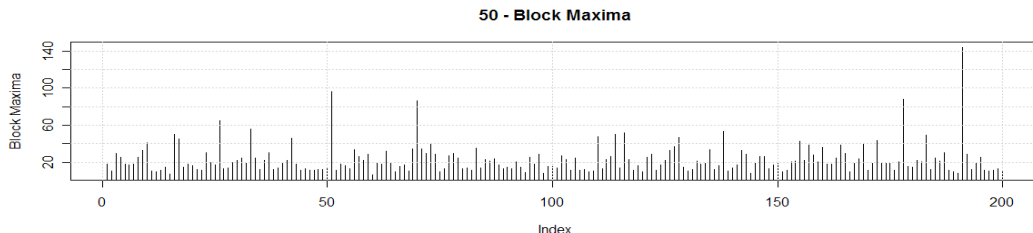
- **Πρόχειρη εκτίμηση του ξ :** Δοκιμάζουμε διάφορες τιμές του ξ μέχρι να δούμε για ποια τιμή, τα k σημεία περίπου ευθυγραμμίζονται.

- Συνήθως ξεκινάμε από το $\xi = 0$ (Q-Q plot των block maxima ως προς την $\Lambda(0,1)$):

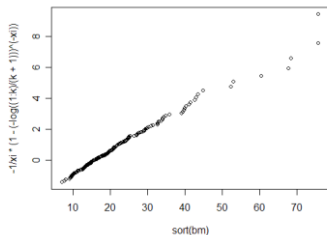
$$(Y_{(i)}, -\ln(-\ln \frac{i}{k+1})), i = 1, 2, \dots, k$$

- Εάν το γράφημα των σημείων είναι κοίλο (*concave*) τότε δοκιμάζουμε $\xi > 0$
- Εάν το γράφημα των σημείων είναι κυρτό (*convex*) τότε δοκιμάζουμε $\xi < 0$

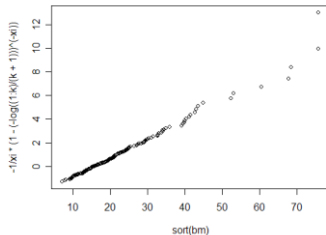
Παράδειγμα 1. Θεωρούμε το παράδειγμα με τις $nk = 10000$ τιμές των X_i που έχουμε χρησιμοποιήσει παραπάνω. Κρατάμε τα block maxima $Y_1, Y_2, \dots, Y_k$ με $k = 200, n = 50$:



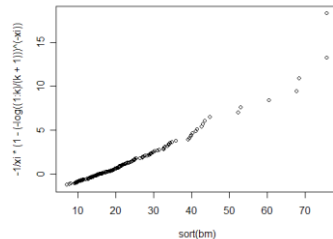
- Το γράφημα των σημείων στο Gumbel Q-Q plot φαίνεται να είναι κοίλο και υποπετυόμαστε ότι το ξ πιθανόν να είναι θετικό.
- Μπορούμε φτιάξουμε το Q-Q plot της GEV με $\mu = 0$, $\sigma = 1$, δοκιμάζοντας διάφορες τιμές του ξ έως ότου τα σημεία περίπου ευθυγραμμιστούν:



GEV Q-Q plot ($\xi = 0.2$)



GEV Q-Q plot ($\xi = 0.3$)



GEV Q-Q plot ($\xi = 0.4$)

- Μάλλον $\xi \approx 0.3$. Επίσης, τα σημεία θα βρίσκονται πάνω στην ευθεία $y = (x - \mu)/\sigma$. Από εδώ προκύπτει εύκολα από το μεσαίο παραπάνω σχήμα ότι το $\mu \approx 16$, $\sigma \approx 7$.

Μέθοδος μέγιστης πιθανοφάνειας.

- Η εκτίμηση των παραμέτρων μ , σ και ξ από τα block maxima μέσω της μεθόδου μέγιστης πιθανοφάνειας είναι (από το πακέτο ExtRemes του R):

$$\hat{\mu} = 15.9402659, \hat{\sigma} = 6.2675298 \quad \hat{\xi} = 0.2559513$$

με εκτιμήσεις των τυπικών αποκλίσεων τους

$$0.50655041, \quad 0.41630444 \quad \text{και} \quad 0.06223748$$

αντίστοιχα.

- Μπορούμε τώρα να κατασκευάσουμε και τα αντίστοιχα (προσεγγιστικά) δ.ε. συντελεστού $1 - \alpha$:

$$(\hat{\mu} \pm \sqrt{s_{1,1}} z_{\alpha/2}), (\hat{\sigma} \pm \sqrt{s_{2,2}} z_{\alpha/2}), (\hat{\xi} \pm \sqrt{s_{3,3}} z_{\alpha/2}).$$

Π.χ. για το ξ το δ.ε. 95% θα είναι το (0.133966, 0.377937).

Δ.Ε. μέσω της profile Deviance Function

- Ένα δ.ε. 95% για το ξ μέσω της *profile Deviance Function* θα είναι το

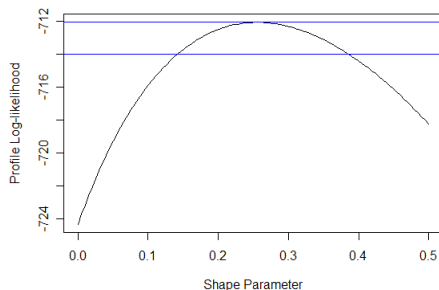
$$\{\xi : D_p(\xi) \leq \chi_1^2(a)\}, \quad \text{όπου} \quad D_p(\xi) = 2(l(\hat{\mu}, \hat{\sigma}, \hat{\xi}) - \max_{\mu, \sigma} l(\mu, \sigma, \xi)) \sim \chi_1^2$$

- Κατασκευάζουμε το γράφημα της συνάρτησης (Profile log-likelihood)

$$\max_{\mu, \sigma} l(\mu, \sigma, \xi) \quad \text{για} \quad \xi \in (0.1, 0.5)$$

(κατασκευάζεται αριθμητικά π.χ. λαμβάνοντας 100 τιμές του ξ μέσα στο $(0, 0.5)$)

- προσθέτουμε την οριζόντια γραμμή με τεταγμένη $l(\hat{\mu}, \hat{\sigma}, \hat{\xi}) - \chi_1^2(a)/2$:

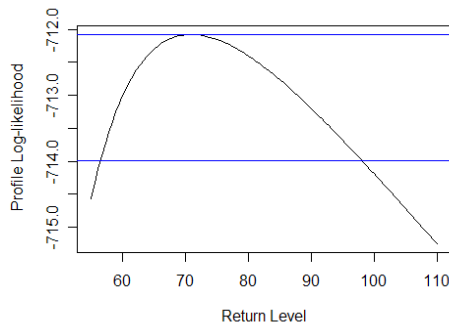
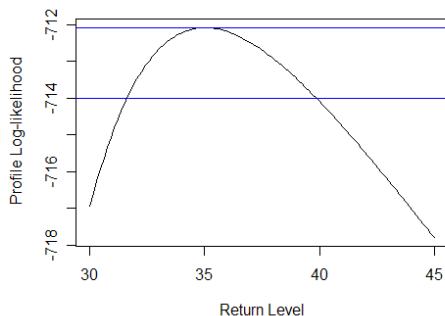


Εκτίμηση της στάθμης απόδοσης z_p (για $1/p$ χρονικές περιόδους)

(1 block = μία χρονική περίοδος)

$$\hat{z}_p = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}} (1 - e^{\hat{\xi} w_p}) \approx 15.94 - \frac{6.267}{0.256} (1 - e^{0.256 w_p})$$

όπου $w_p = -\ln(-\ln(1-p))$. Επίσης μπορούμε να εκτιμήσουμε την στάθμη απόδοσης από την profile log-likelihood: (για $1/p = 10$ και 100 αντίστοιχα):



- ένα block maximum υπερβαίνει την τιμή 72 με πιθανότητα $p = 1/100$ ή ισοδύναμα
- ένα block maximum υπερβαίνει την τιμή 72 κατά μέσο όρο κάθε 100 χρονικές περιόδους.

Διαγνωστικά γραφήματα (καλή προσαρμογή των block maxima στην GEV)

