Σύνθετα Δίκτυα

com+plex: with+ -fold (having parts)

Διδάσκων –
Δημήτριος Κατσαρός

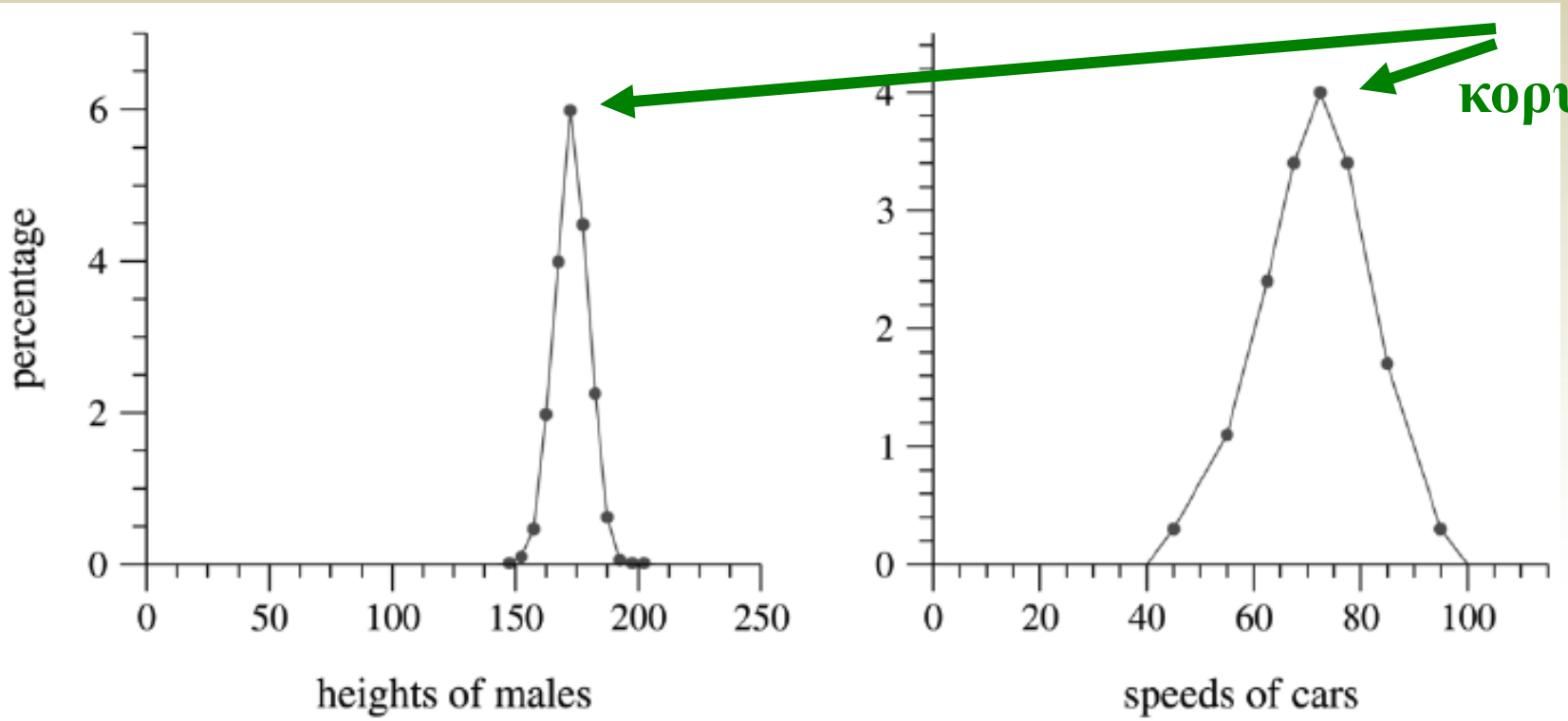
Διάλεξη 3η

Δυναμο-νόμοι

Power-laws

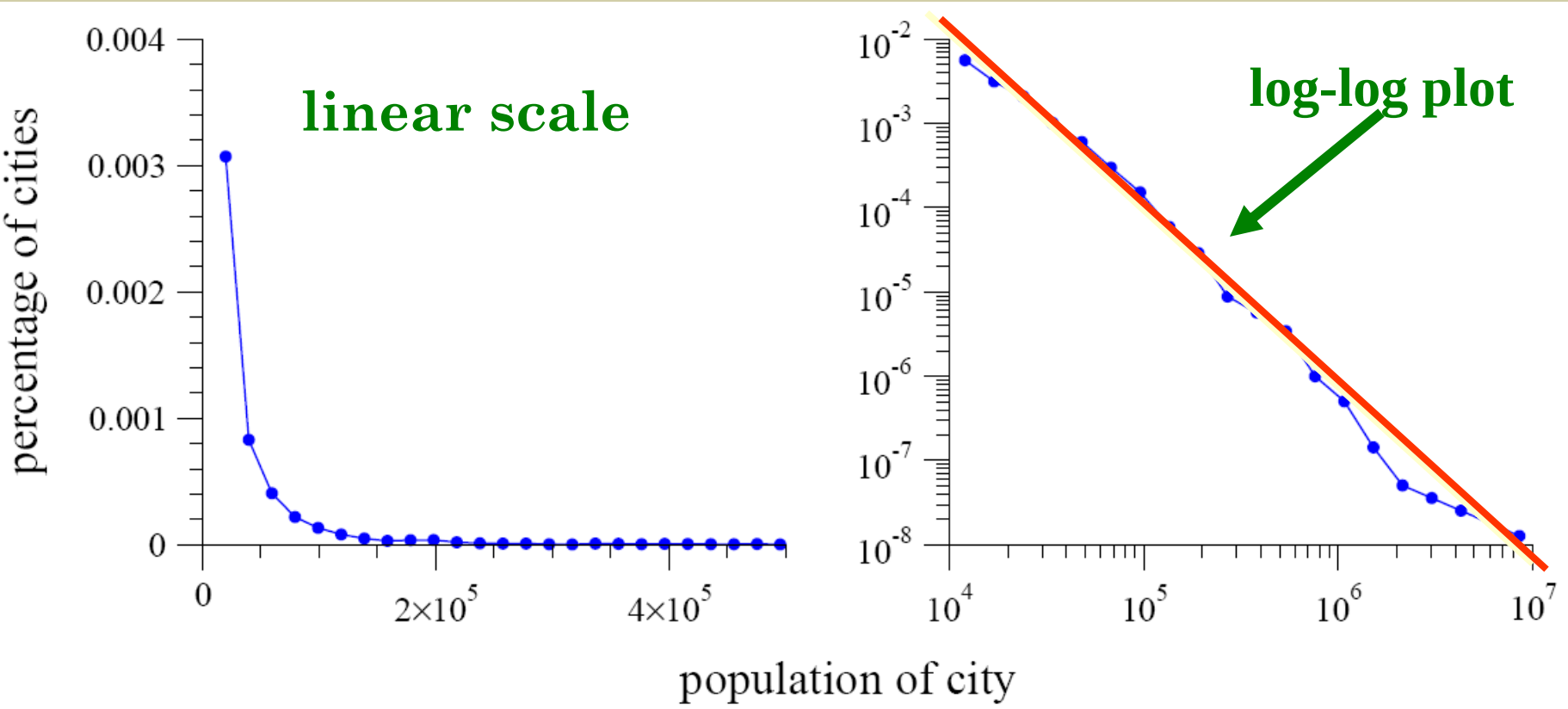
Η κλίμακα μετρούμενων ποσοτήτων

- Τα περισσότερα από τα πράγματα που θα μπορούσε κάποιος να μετρήσει έχουν “μέγεθος” (κλίμακα)
- Π.χ., Η κατανομή ύψους των ανθρώπων (Fig.: USA, males ‘59-’62)
- Η ωριαία ταχύτητα των αυτοκινήτων (Fig.: UK motorways, 2003)



Η κλίμακα μετρούμενων ποσοτήτων

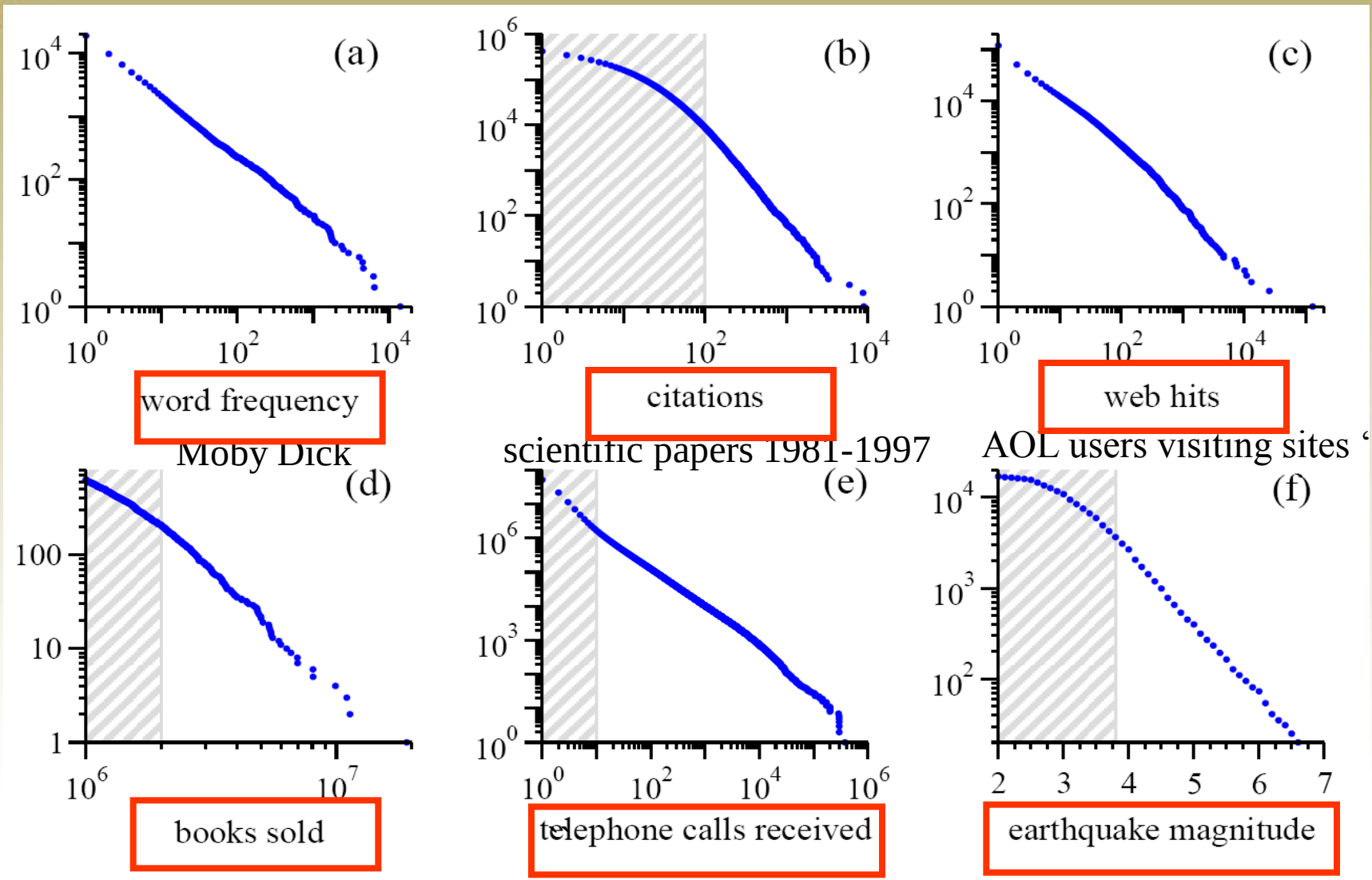
- Δεν είναι όμως όλες οι κατανομές συγκεντρωμένες γύρω από μια κορυφή
- Όλες οι USA πόλεις με πληθυσμό πάνω από 10000 κατοίκους



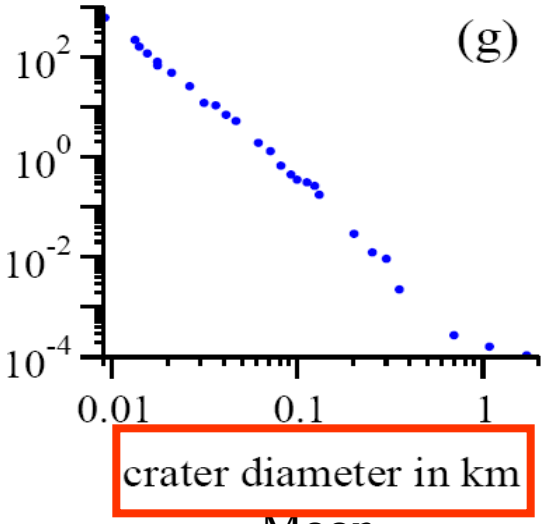
Ποια είναι κατανομή “παχιάς ουράς”?

- Δεξιά κύρτωση
  - Κατανομή κανονική (gaussian) (χωρίς παχιά ουρά)
    - π.χ., ύψος των ανδρών: με κέντρο γύρω στο 180cm
  - Κατανομή Zipfian ή δυναμο-νόμου (με παχιά ουρά)
    - π.χ., μεγέθη πληθυσμού πόλεων: NYC 8 εκατομ, αλλά, πάρα πολλές μικρές πόλεις
- Μεγάλη τιμή του κλάσματος max προς min
  - Ανθρώπινο ύψος
    - Υψηλότερος άνδρας: 272cm, κοντύτερος άνδρας: 56,6cm
      κλάσμα: 4.8
      από το Guinness Book των παγκοσμίων records
  - Πληθυσμοί πόλεων
    - NYC: 8 εκατομύρια, Duffield (Virginia): 52, *κλάσμα: 150000*

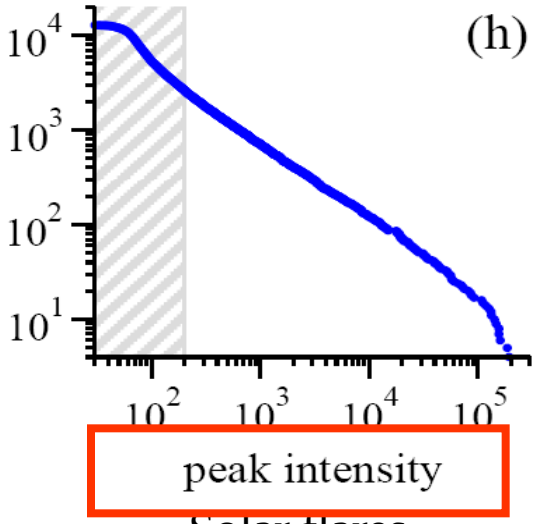
Δυναμο-νόμοι βρίσκονται παντού …



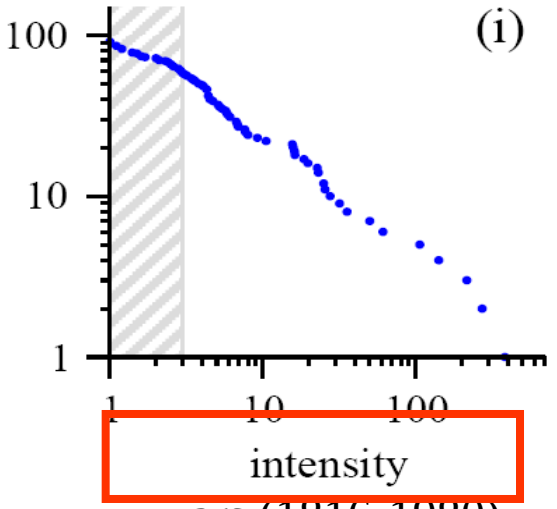
…περισσότεροι δυναμο-νόμοι



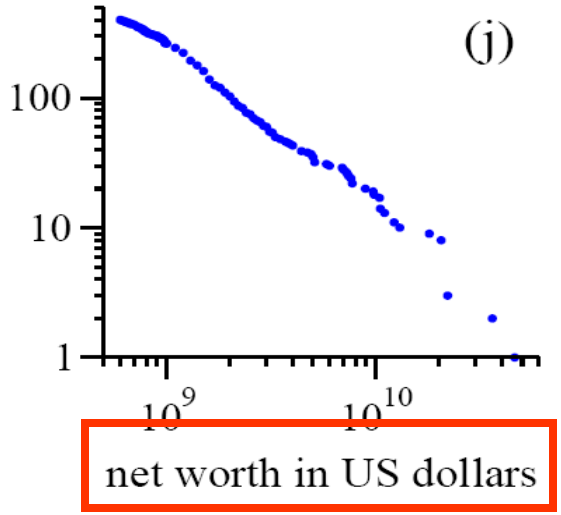
Moon



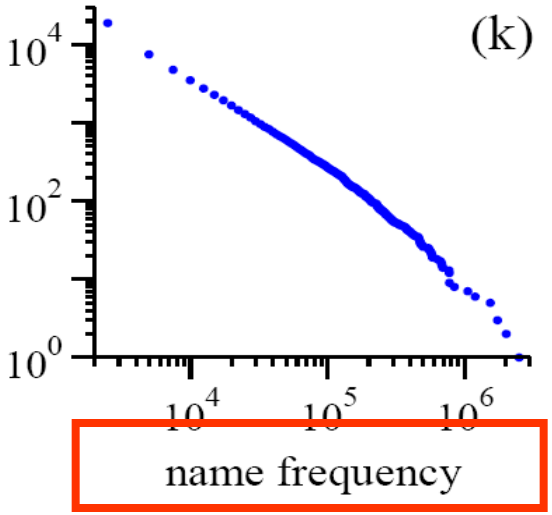
Solar flares



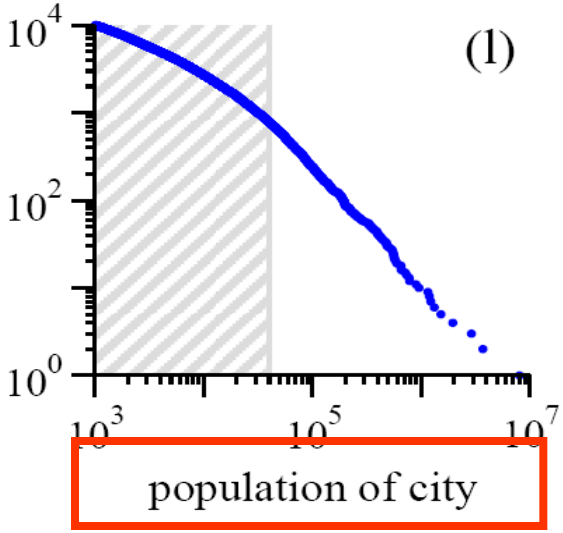
wars (1816-1980)



richest individuals 2003



US family names 1990



US cities 2003

Κατανομή ενός δυναμο-νόμου

- Ευθεία γραμμή σε διάγραμμα με log-log άξονες

$$\ln(p(x)) = c - \alpha \ln(x)$$

- Exponentiate και τα δυο μέρη για να πάρουμε το *p(x)*, η πιθανότητα να παρατηρήσουμε ένα στοιχείο μεγέθους 'x' δίνεται από τη σχέση

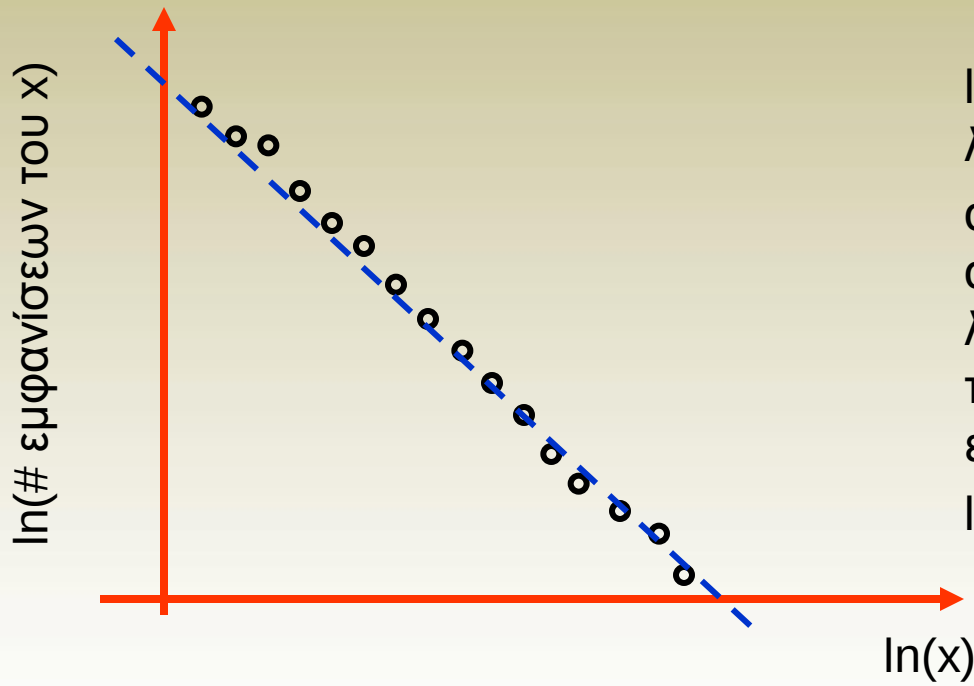
$$p(x) = Cx^{-\alpha}$$

σταθερά κανονικοποίησης (οι πιθανότητες πάνω σε όλα τα *x* πρέπει να αθροίζουν στο 1)

εκθέτης α του δυναμο-νόμου

Προσέγγιση κατανομών δυναμο-νόμων

- Η πιο κοινή, αλλά όχι πολύ ακριβής μέθοδος:
  - Βάζουμε σε κουτιά (binning) τις διαφορετικές τιμές του x, και δημιουργούμε ένα ιστόγραμμα συχνοτήτων (frequency histogram)



ln(*x*) είναι ο φυσικός λογάριθμος του x,

αλλά οποιαδήποτε άλλη βάση του λογαρίθμου θα δώσει τον ίδιο εκθέτη, επειδή

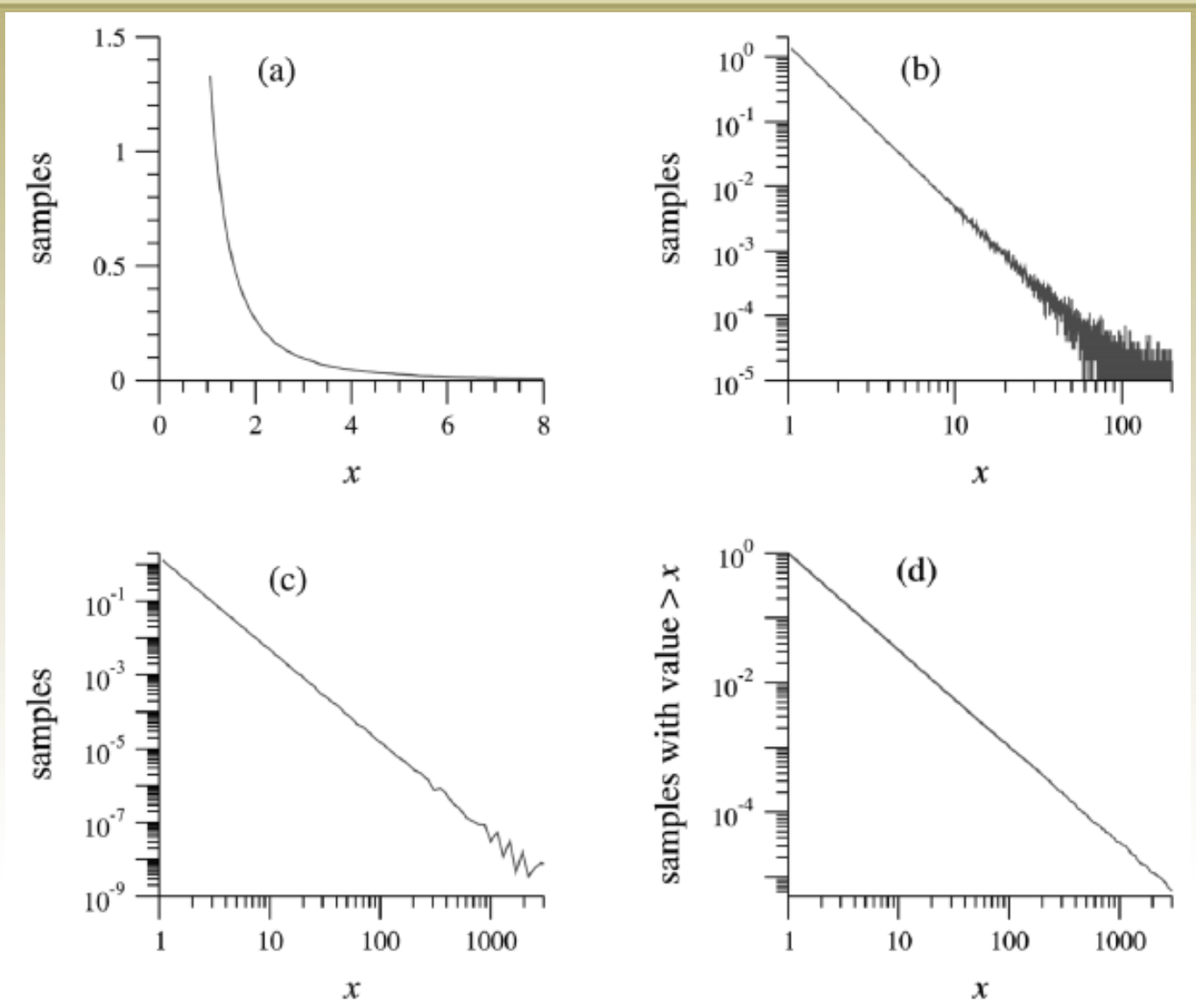
$\log_{10}(x) = \ln(x)/\ln(10)$

Το x μπορεί να αναπαριστά διάφορες ποσότητες, τον in-degree ενός κόμβου, το PageRank ενός κόμβου, το μέγεθος ενός σεισμού, τη συχνότητα μια λέξης σε κάποιο κείμενο, …

Γέννηση με τεχνητό τρόπο ενός συνόλου δεδομένων που ακολουθούν δυναμο-νόμο

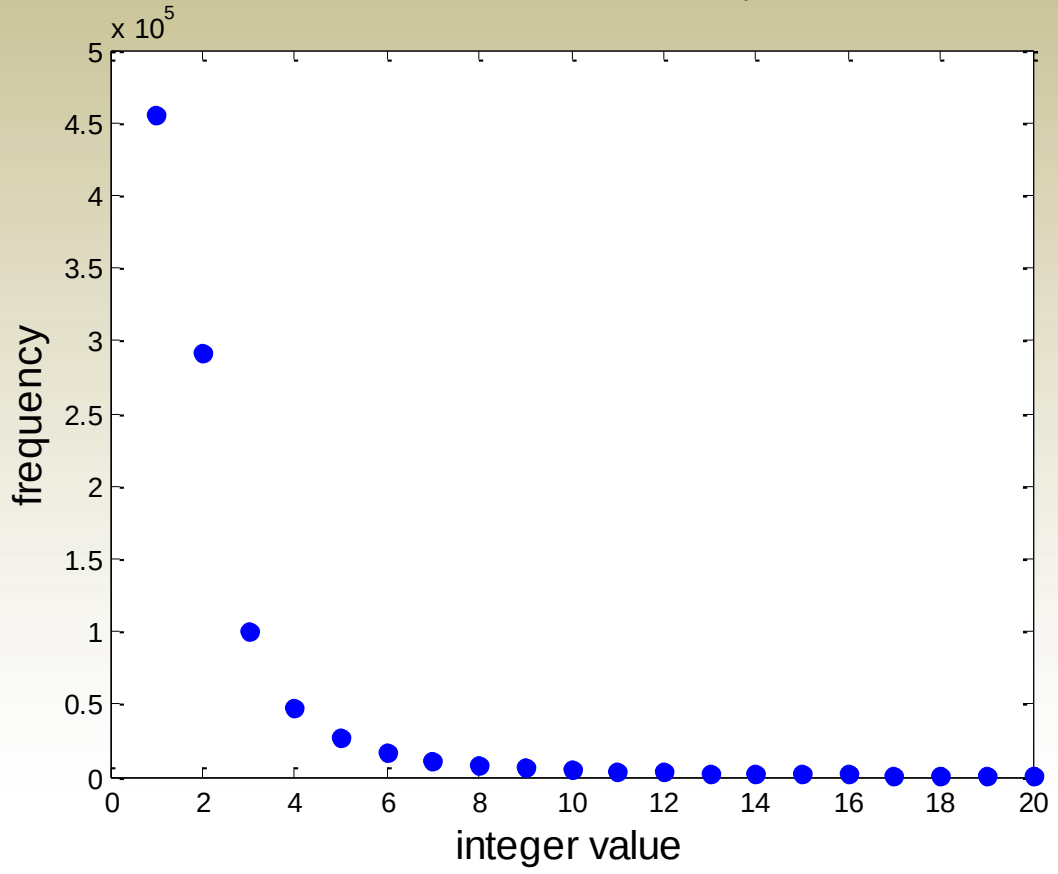
- Θα παράξουμε 1 εκατομμύριο τυχαίους αριθμούς από μια κατανομή δυναμο-νόμου με $\alpha = 2.5$
- Μπορεί να παραχθούν με την επονομαζόμενη 'transformation method'
- Γεννούμε τυχαίους αριθμούς r στο μοναδιαίο διάστημα $0 \leq r < 1$
- τότε, ο $x = (1-r)^{-1/(\alpha-1)}$ είναι τυχαίος πραγματικός αριθμός κατανεμημένος κατά έναν δυναμο-νόμο στην εμβέλεια $1 \leq x < \infty$

Γραφική αναπαράσταση

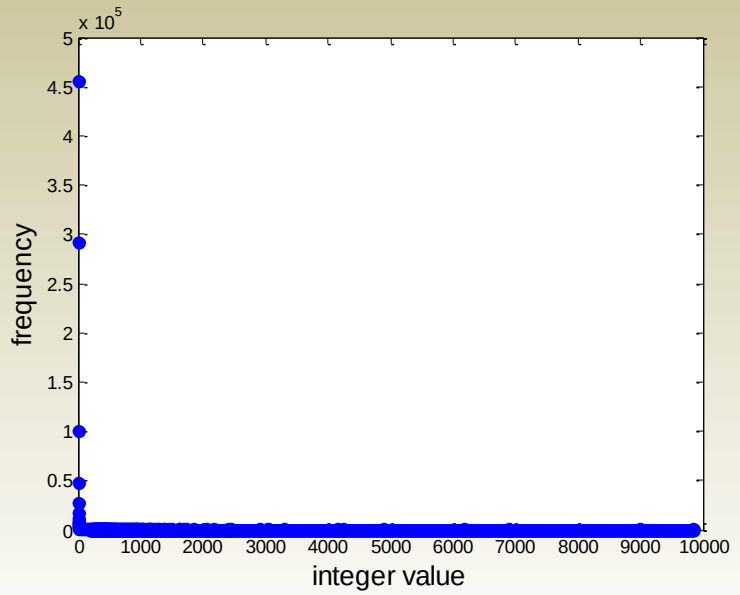


Γραμμική αναπαράσταση του straight binning των δεδομένων

- Ένας δυναμο-νόμος μπορεί να μην είναι άμεσα ορατός
- Φαίνεται μόνο εάν κοιτάξουμε στα μικρά “κουτιά”



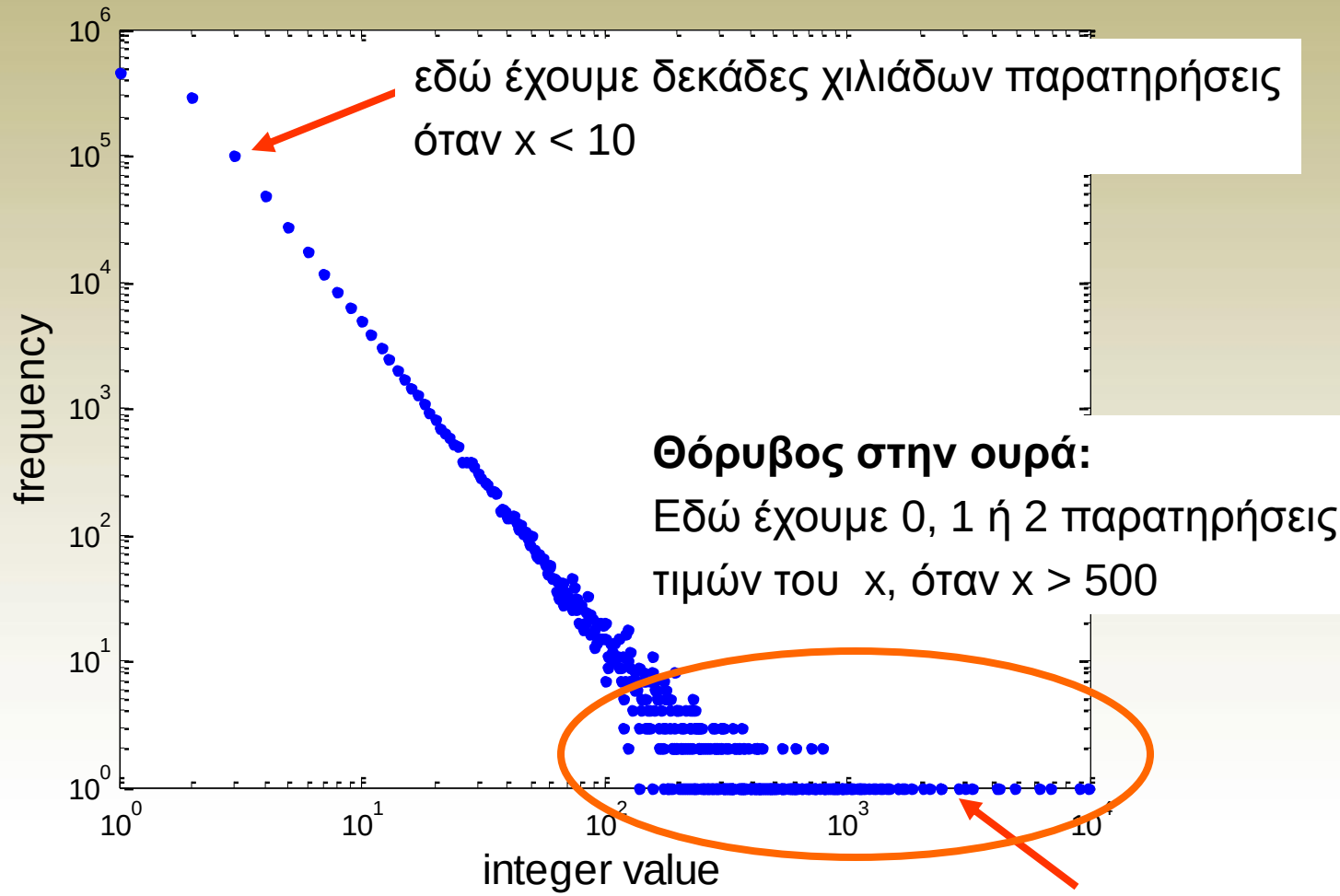
τα πρώτα λίγα κουτιά



ολόκληρη η εμβέλεια

Log-log αναπαράσταση του straight binning των δεδομένων

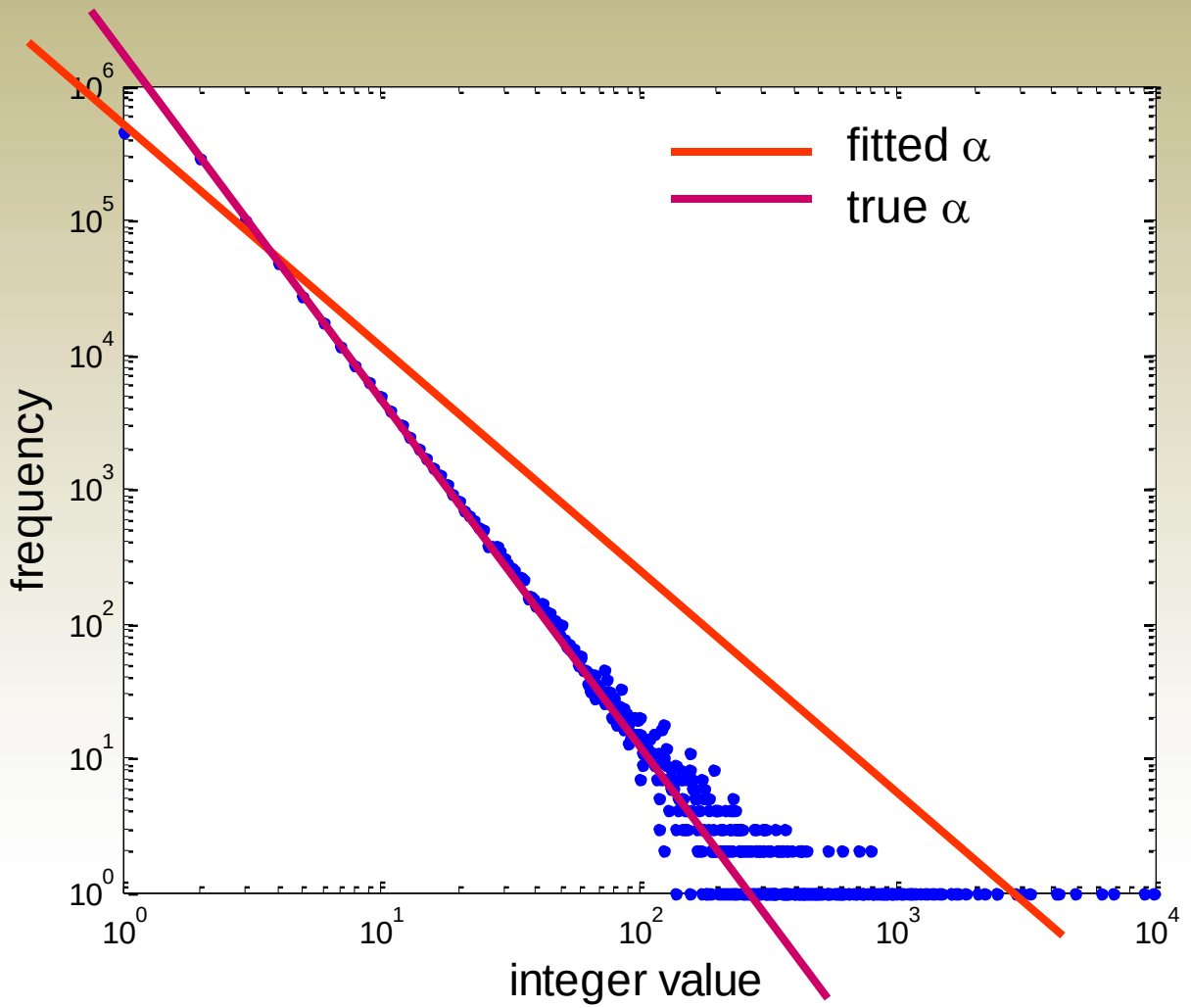
- Τα ίδια κουτιά, αλλά σε log-log κλίμακα



Στην πραγματικότητα δεν βλέπουμε όλες τις μηδενικές τιμές, γιατί log(0) = ∞

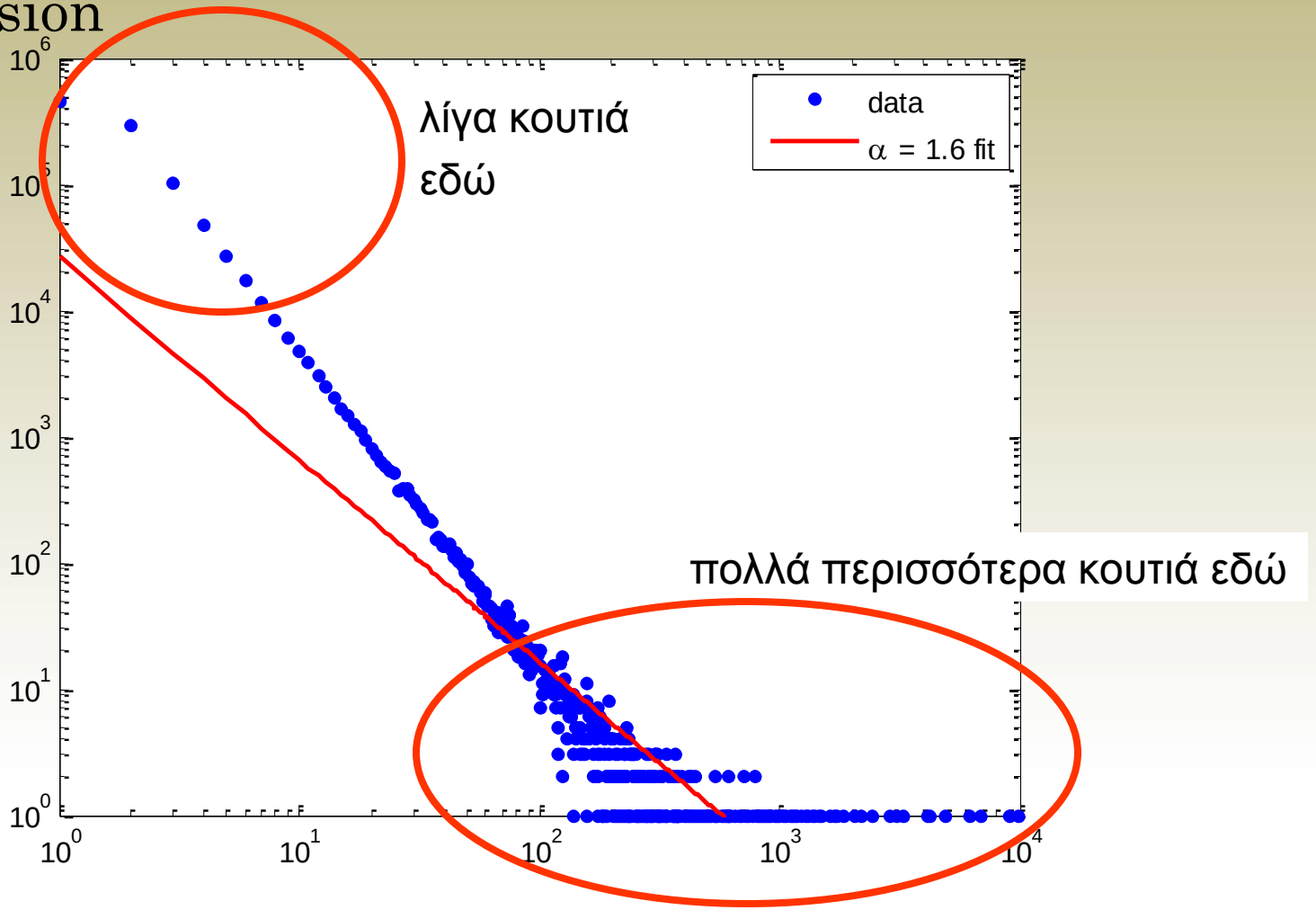
Log-log scale γράφημα του απλοϊκού binning των δεδομένων

Προσαρμόζοντας μια ευθεία γραμμή στα δεδομένα με την τεχνική των ελαχίστων τετραγώνων regression θα υπολογίσει τιμές για τον εκθέτη α οι οποίες είναι πολύ χαμηλές



Τι είναι λάθος με το απλοϊκό binning;

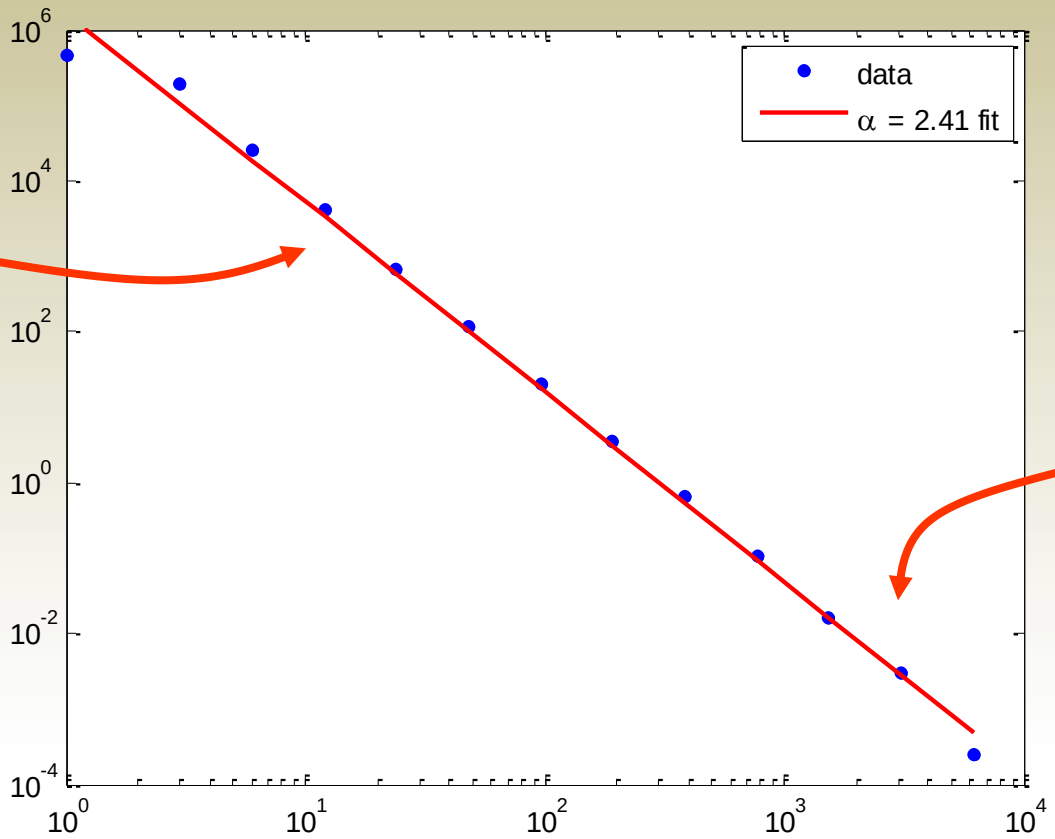
Ο θόρυβος στην ουρά κυρτώνει το αποτέλεσμα του regression



Πρώτη λύση: Λογαρίθμηση

- Βάζουμε (bin) τα δεδομένα σε εκθετικά μεγαλύτερα (wider) κουτιά: 1, 2, 4, 8, 16, 32, …
- Κανονικοποιούμε με το πλάτος κάθε κουτιού

ισαπέχοντα σημεία



λιγότερος θόρυβος στην ουρά της κατανομής

μειονέκτημα: το binning εξομαλύνει τα δεδομένα, αλλά χάνει πληροφορίες

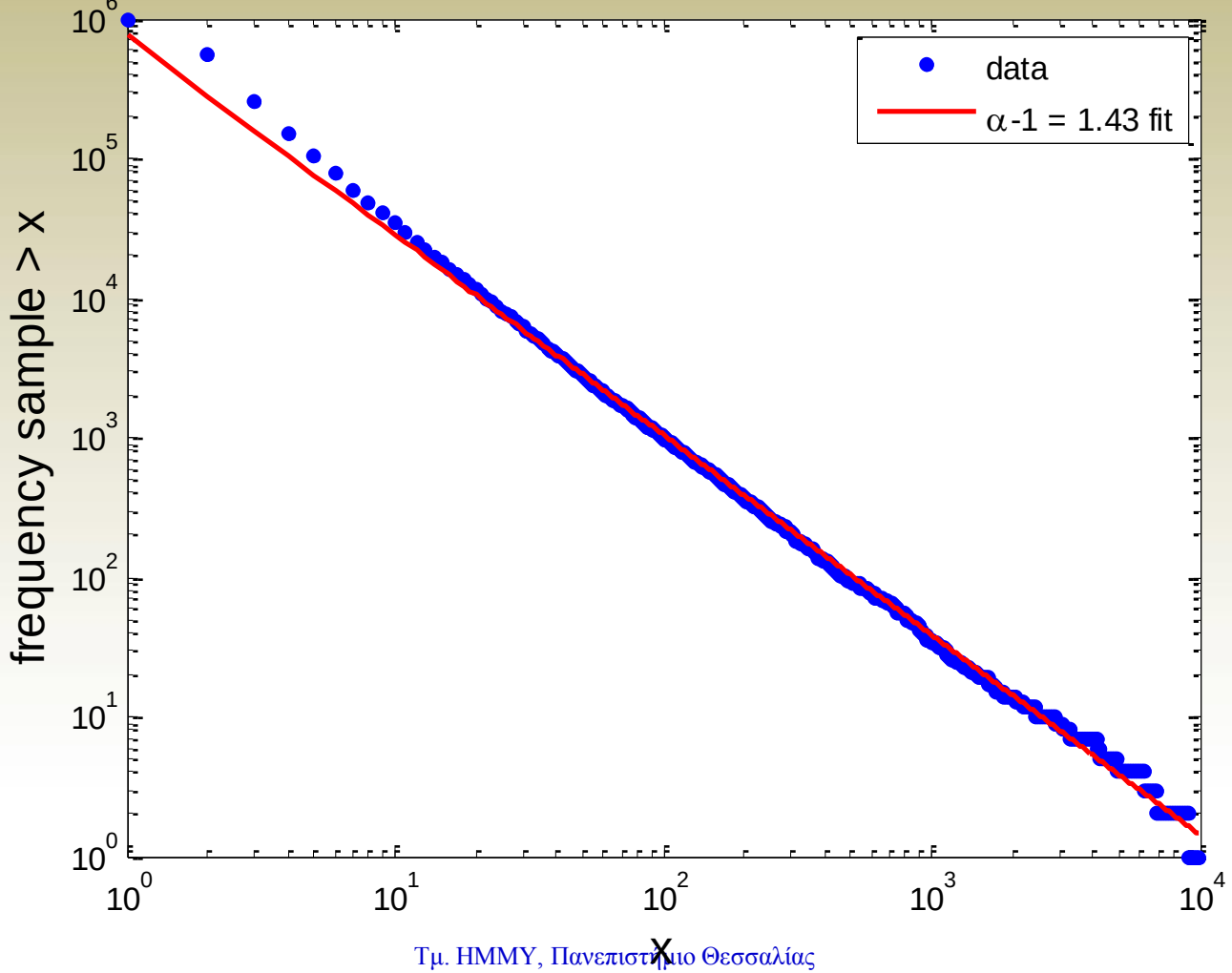
Δεύτερη λύση: Συσσωρευτικά κουτιά (bins)

- Χωρίς απώλεια πληροφορίας
  - Χωρίς ανάγκη για bin, έχει τιμή για κάθε παρατηρούμενη τιμή του x
- Αλλά, τώρα έχουμε cumulative κατανομή
  - Δηλ., πόσες από τις τιμές του x είναι τουλάχιστον X
  - Η cumulative probability της probability distribution ενός δυναμο-νόμου είναι επίσης δυναμο-νόμος με εκθέτη **α - 1**

$$\int cx^{-\alpha} = \frac{c}{1-\alpha} x^{-(\alpha-1)}$$

Προσαρμογή με regression στην cumulative κατανομή

fitted εκθέτης (2.43) πολύ κοντά στον πραγματικό (2.5)

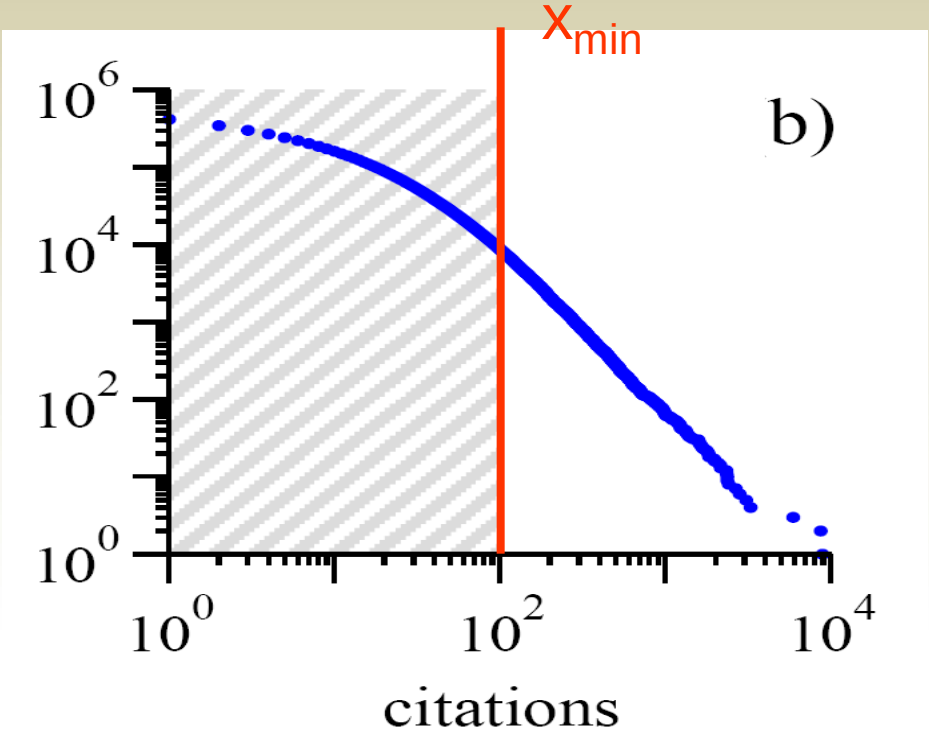


Από πού να ξεκινήσει η προσαρμογή;

- Μερικά σύνολα δεδομένων υπακούουν σε δυναμο-νόμο μόνο στην ουρά
- Μετά το binning ή παίρνοντας την cumulative distribution, μπορούμε να κάνουμε προσαρμογή στην ουρά
- Άρα, αρκεί να επιλέξουμε x_{min} την τιμή του x όπου θεωρούμε ότι ξεκινά ο δυναμο-νόμος
- Φυσικά το x_{min} πρέπει να είναι μεγαλύτερο από 0, επειδή το $x^{-\alpha}$ απειρίζεται στο $x = 0$

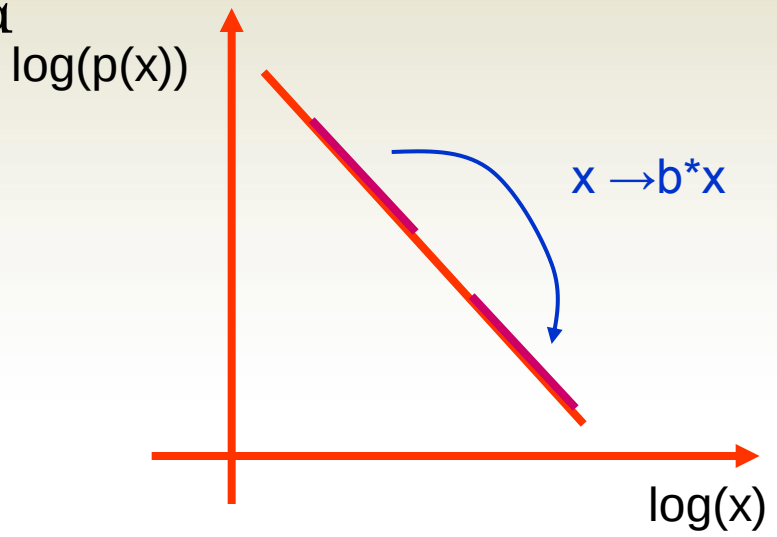
Παράδειγμα

- Κατανομή των αναφορών (citations) σε άρθρα
- Ο δυναμο-νόμος είναι εμφανής μόνο στην ουρά
  - $x_{min} > 100$ citations



Τι σημαίνει να είσαι άνευ κλίμακας (scale-free)?

- Ένας δυναμο-νόμος φαίνεται ίδιος ανεξάρτητα από την κλίμακα κάτω από την οποία τον παρατηρούμε (2 μέχρι 50 ή 200 μέχρι 5000)
- Είναι αληθές μόνο για κατανομές δυναμο-νόμων!
- $p(bx) = g(b)\ p(x)$
  - το σχήμα της κατανομής δεν αλλάζει εκτός από μια πολλαπλασιαστική σταθερά
- $p(bx) = (bx)^{-\alpha} = b^{-\alpha}\ x^{-\alpha}$



Μερικοί εκθέτες για πραγματικά δεδομένα

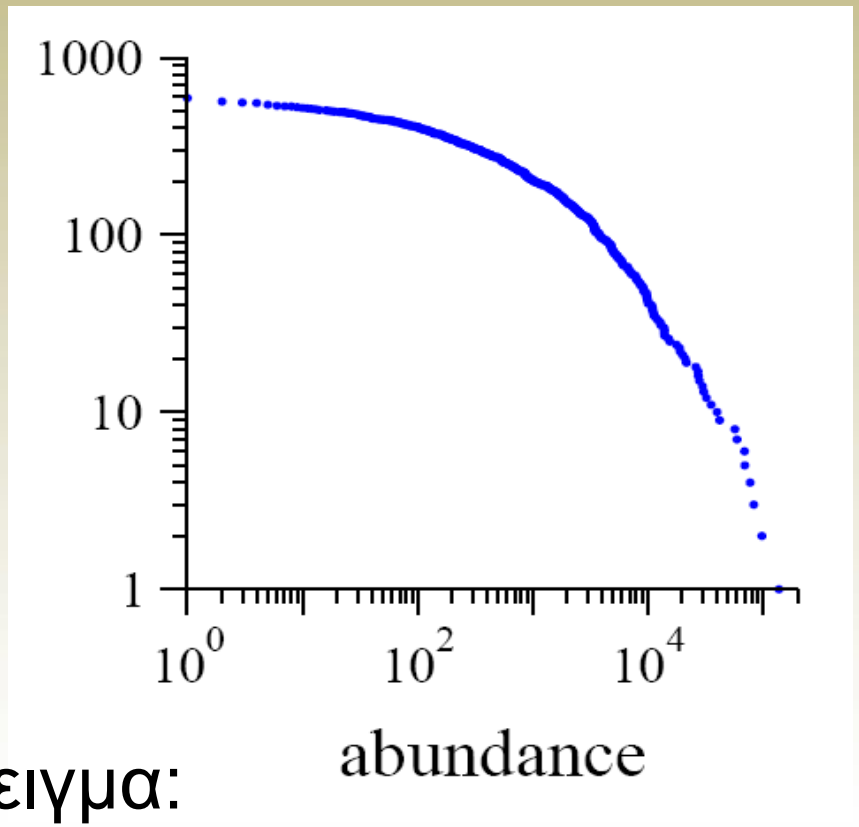
	x_{min}	exponent α
frequency of use of words	1	2.20
number of citations to papers	100	3.04
number of hits on web sites	1	2.40
copies of books sold in the US	2 000 000	3.51
telephone calls received	10	2.22
magnitude of earthquakes	3.8	3.04
diameter of moon craters	0.01	3.14
intensity of solar flares	200	1.83
intensity of wars	3	1.80
net worth of Americans	$600m	2.09
frequency of family names	10 000	1.94
population of US cities	40 000	2.30

Πολλά δίκτυα υπακούουν σε δυναμο-νόμους

	exponent α (in/out degree)
film actors	2.3
telephone call graph	2.1
email networks	1.5/2.0
sexual contacts	3.2
WWW	2.3/2.7
internet	2.5
peer-to-peer	2.1
metabolic network	2.2
protein interactions	2.4

Δεν είναι όλα δυναμο-νόμοι

- αριθμός εμφανίσεων 591 bird species στην North American Bird survey το 2003



- άλλο παράδειγμα:
  - Μέγεθος πυρκαγιών (σε acres)

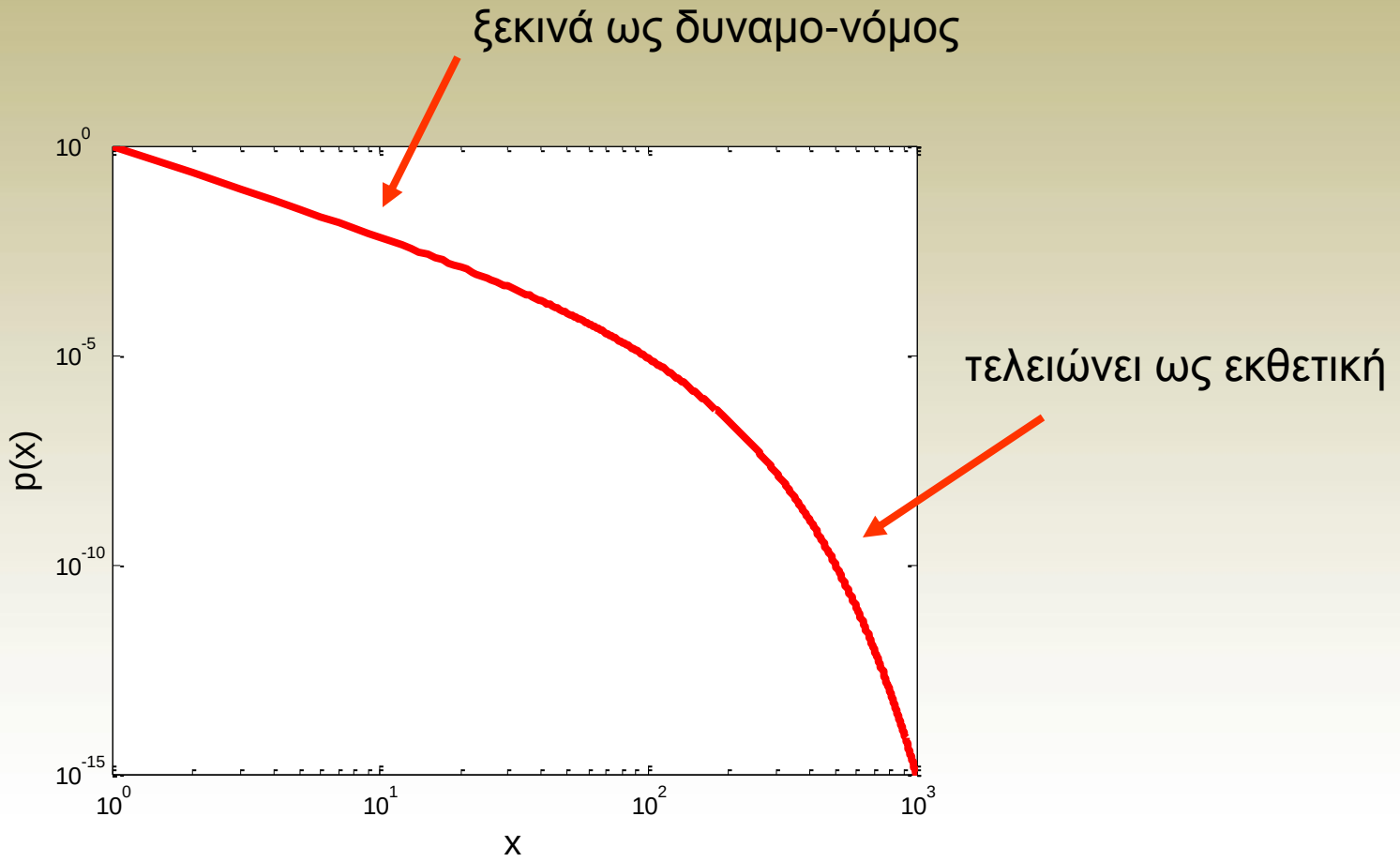
Source: MEJ Newman, 'Power laws, Pareto distributions and Zipf's law'

Δεν υπακούουν όλα τα δίκτυα σε κάποιον δυναμο-νόμο

- frequent email communication
- power grid
- Roget's thesaurus
- company directors …

Ακόμα μια δημοφιλής κατανομή: Δυναμο-νόμος με εκθετικό κατώφλι

- $p(x) \sim x^{-a}\, e^{-x/\kappa}$



αλλά, θα μπορούσε να είναι lognormal ή διπλή εκθετική …

Zipf & Pareto: Η σχέση με δυναμο-νόμους

- Ο George Kingsley Zipf, καθηγητής γλωσσολογίας στο Harvard, επεδίωξε να προσδιορίσει το "μέγεθος" της 3ης ή 8ης ή 100ης πιο συνηθισμένης λέξης
  - Το "μέγεθος" αναφέρεται στη συχνότητα χρήσης της λέξης στην αγγλική γλώσσα, και όχι στο μέγεθος της λέξης

- Ο νόμος του Zipf μας λέει ότι το "μέγεθος" του r-οστού σε εμφανίσεις γεγονός είναι αντιστρόφως ανάλογο του βαθμού (rank) του:

$$y \sim r^{-\beta}$$, με το β κοντά στη μονάδα

Zipf & Pareto: Η σχέση με δυναμο-νόμους

- Ο ιταλός οικονομολόγος Vilfredo Pareto μελέτησε την κατανομή του εισοδήματος
- Ο νόμος του Pareto εκφράζεται με όρους της cumulative κατανομής
  - Η πιθανότητα ότι κάποιος κερδίζει Χ ή περισσότερα δίνεται από τη σχέση:

$$P[X > x] \sim x^{-k}$$

- Εδώ, αναγνωρίζουμε το k απλά ως α -1, όπου α είναι ο power-law exponent

Πώς πάμε από τον Zipf στον Pareto?

- Η φράση «Η ***r***-οστή μεγαλύτερη πόλη έχει ***n*** κατοίκους" είναι ισοδύναμη με το να πούμε ότι "***r*** πόλεις έχουν ***n*** ή περισσότερους κατοίκους"
- Αυτός είναι ο ορισμός της κατανομής κατά Pareto, εκτός του ότι οι άξονες x και y ανταλλάσσονται
  - για Zipf, ***r*** είναι στον x άξονα, και ***n*** είναι στον y άξονα
  - για Pareto, ***r*** είναι στον y άξονα, και ***n*** είναι στον άξονα x
- Απλά αντιστρέφοντας του άξονες
  - εάν ο rank exponent είναι ***β***, δηλαδή

  $n \sim r^{-\beta}$ για Zipf, (n = income, r = rank of person with income n)
  - τότε ο Pareto exponent είναι $1/\beta$ έτσι ώστε

  $r \sim n^{-1/\beta}$ (n = income, r = number of people whose income is n or higher)

Zipf's Law και μεγέθη πόλεων (~1930)

Rank(k)	City	Population (1990)	Zips's Law $10{,}000{,}000/k$	Modified Zipf's law: (Mandelbrot) $5{,}000{,}000/\left(k-2/5\right)^{3/4}$
1	Now York	7,322,564	10,000,000	7,334,265
7	Detroit	1,027,974	1,428,571	1,214,261
13	Baltimore	736,014	769,231	747,693
19	Washington DC	606,900	526,316	558,258
25	New Orleans	496,938	400,000	452,656
31	Kansas City	434,829	322,581	384,308
37	Virgina Beach	393,089	270,270	336,015
49	Toledo	332,943	204,082	271,639
61	Arlington	261,721	163,932	230,205
73	Baton Rouge	219,531	136,986	201,033
85	Hialeah	188,008	117,647	179,243
97	Bakersfield	174,820	103,270	162,270

source: Luciano Pietronero

Ο κανόνας 80/20 (**Η αρχή του Pareto)**

- Ο Joseph M. Juran παρατήρησε ότι 80% της γης στην Ιταλία κατεχόταν από το 20% του πληθυσμού της

- Το ποσοστό W του πλούτου που βρίσκεται στα χέρια των P του πληθυσμού δίνεται από τη σχέση

$$W = P^{(\alpha-2)/(\alpha-1)}$$

- Παράδειγμα: Ο πλούτος στις US: $\alpha = 2.1$
  - Οι πλουσιότεροι 20% του πληθυσμού κατέχουν το 86% του πλούτου

Ασκήσεις

- **Άσκηση 1.**

  Έστω ότι κάποιος δυναμο-νόμος είναι της μορφής $p(x)= Cx^{-\alpha}$, με $\alpha>0$. Υποθέτουμε ότι υπάρχει μια ελάχιστη τιμή του x, έστω η x_{min}, πάνω από την οποία ισχύει ο δυναμο-νόμος. Να υπολογιστεί η τιμή της σταθεράς C ως συνάρτηση των x_{min} και α.

Ασκήσεις

- **Άσκηση 2.**

  Έστω ότι κάποιος δυναμο-νόμος είναι της μορφής $p(x)= Cx^{-\alpha}$, με $\alpha>0$. Υποθέτουμε ότι υπάρχει μια ελάχιστη τιμή του x, έστω η x_{min}, πάνω από την οποία ισχύει ο δυναμο-νόμος. Να βρεθεί η μέση τιμή $<x>$.

  Εξερευνήστε την επίδραση της τιμής του α.

- **Άσκηση 3.**

  Για τον ίδιο δυναμο-νόμο, ποια είναι η median τιμή, το $x_{1/2}$, δηλαδή εκείνο το x που διαιρεί την κατανομή στη μέση, έτσι ώστε οι μισές μετρούμενες τιμές να είναι μικρότερες του $x_{1/2}$ και οι άλλες μισές μεγαλύτερες του $x_{1/2}$;

$$\int_{x_{1/2}}^{\infty} p(x)dx = \frac{1}{2}\int_{x_{\min}}^{\infty} p(x)dx$$

Γέννηση τυχαίων μεταβλητών

Ο Η/Υ μας έχει γεννήτρια ψευδοτυχαίων με LCG

Παράγει: U ~ U(0,1) [διαιρώντας με INT_MAX]

Γέννηση με την μέθοδο του Inverse Transform

1. Generate $U \sim U(0,1)$
2. Return $X = F^{-1}(U)$

- **Γέννηση Exponential με mean β**

1. Generate $U \sim U(0,1)$
2. Return $X = -\beta \ln(U)$

- **Γέννηση Normal $N(0,1)$**

1. Generate U_1 and U_2 as IID $U(0,1)$, let $V_i = 2U_i - 1$ for i=1,2 and let $W = V_1^2 + V_2^2$
2. If $W>1$, go back to step 1. Otherwise, let $Y = \text{sqrt}\{(-2 \ln W)/W\}$, $X_1 = V_1 Y$, and $X_2 = V_2 Y$. Then, X_1 and X_2 are IID $N(0,1)$ random variates