TREBALL DE FI DE MÀSTER

Màster Interuniversitari en Estadística i Investigació Operativa

Títol: La distribució Zipf Estesa segons la transformació Marshall-Olkin

Autora: Aina Casellas Torrentó

Directora: Marta Pérez Casany

Departament: Matemàtica Aplicada II

Universitat: Universitat Politècnica de Catalunya

Convocatòria: 29 de gener de 2013

UPC
Facultat de Matemàtiques i Estadística
UNIVERSITAT POLITÈCNICA DE CATALUNYA
UNIVERSITAT DE BARCELONA

Upc - Facultat de Matemàtiques i Estadística

Treball Final de Màster
Projecte Final de Carrera

La distribució Zipf Estesa segons la transformació Marshall-Olkin

---

Autora:
AINA CASELLAS TORRENTÓ

Directora:
MARTA PÉREZ CASANY
Departament de Matemàtica Aplicada II
Universitat Politècnica de Catalunya

29 de gener de 2013

Índex

Introducció

El treball portat a terme en aquesta Tesi Final de Màster / Projecte Final de Carrera ha consistit en definir i estudiar una generalització de la distribució de probabilitat discreta coneguda amb el nom de *distribució Zipf*. La distribució Zipf és uniparamètrica, i és molt utilitzada en àrees de coneixement tan diverses com poden ser: l'ecologia, el màrqueting, les xarxes socials, el tràfic a internet, les assegurances, la lingüística o la bioinformàtica.

La distribució Zipf és un cas particular de la distribució Power Law, de la que es diferencia principalment per tenir un suport fix. Les principals característiques de la distribució Zipf són que té per suport els enters positius majors o iguals a la unitat, que la unitat té una probabilitat en general força gran en la majoria del domini del seu paràmetre, i que aquesta probabilitat decreix ràpidament, però de forma que la distribució té una cua molt llarga. A més, si s'aplica la transformació log-log, és a dir si s'escriu el logaritme de la probabilitat com a funció del logaritme del valor observat, el resultat és una recta.

La Zipf és útil per a modelar conjunts de dades bàsicament de dos tipus: freqüències de freqüències o rànquings. Entendrem per una freqüència de freqüències aquella taula de freqüències tal que les observacions són també freqüències. Això passa per exemple en la modelització de l'aparició de paraules en un text. El que es modela és la taula que ens dóna el nombre de paraules que apareixen exactament i cops en el text. Òbivament hi haurà moltes paraules que apareixeran només un cop i molt poques que apareixeran moltes vegades, d'aquí que s'obtingui una elevada probabilitat en la unitat i una cua molt llarga. Les dades direm que són rànquings quan s'han ordenat en funció de la seva magnitud. En aquest cas la observació de l'1 correspondria a la dada més gran, la del dos la segona més gran i així successivament. Un exemple seria la variable que mesura la popularitat de les pàgines web. Fixat un conjunt de pàgines, si se

n'estudien el nombre de visites en un període de temps fixat, s'observa que hi ha una pàgina que rep una proporció molt elevada de visites i que hi ha moltes pàgines que en reben proporcions molt petites, d'aquí que es torni a observar una elevada probabilitat en l'1 i una cua llarga.

Malgrat l'elevada popularitat de la distribució Zipf, és conegut que sovint es detecten dos tipus de problemes de cara a l'ajust a dades reals. D'una banda, les probabilitats dels primers valors (a vegades només la de la unitat) són subestimades per la distribució, mentre que la probabilitat de la cua sovint és sobreestimada. A vegades només s'observa clarament una de les dues coses. Això ens va portar a pensar que fóra bo intentar definir una generalització de la distribució que li atorgués més flexibilitat, i així neix el treball que es presenta.

La generalització considerada en aquest treball és la obtinguda per mitjà de l'anomenada transformació de Marshall i Olkin. Aquesta transformació permet augmentar el nombre de paràmetres de la distribució en una unitat. I, tal com es veurà, és apropiada per tal de solventar els problemes esmentats anteriorment. El treball es presenta de la forma següent: en el Capítol 1 es parla de la distribució Power Law en general i de la Zipf en particular. El Capítol 2 està dedicat a la definició i estudi de la distribució generalitzada. En el Capítol 3 s'analitzen diversos conjunts de dades i es comparen els ajustos obtinguts amb les distribucions generalitzada i sense generalitzar. Finalment, al Capítol 4 es porten a terme simulacions per tal d'implementar el test de Kolmogorv-Smirnov discret per a la distribució presentada. El test s'aplica als conjunts de dades analitzats al capítol anterior, per tal de decidir si aquestes segueixen o no la distribució presentada. El document acaba amb la presentació de les principals conclusions del treball realitzat.

Objectius

El treball realitzat en aquesta Tesi Final de Màster / Projecte Final de Carrera es podria dividir en una part més teòrica/matemàtica i una part de caire més aplicat. Els objectius proposats en cadascuna d'aquestes parts són els següents:

Part teòrica

- Utilitzar la transformació de Marshall i Olkin per tal de definir una generalització biparamètrica de la distribució Zipf.
- Estudiar algunes de les propietats bàsiques de la distribució generalitzada.
- Interpretar quin és el paper que desenvolupa el paràmetre addicional, en el comportament de la distribució.

Part pràctica

- Ajustar la nova família de distribucions a conjunts de dades reals de diferents àmbits, per tal de veure'n la seva aplicació.
- Estudiar la bondat d'ajust del nou model.
- Comparar els ajustos obtinguts amb les famílies generalitzada i sense generalitzar, a través de diferents test d'hipòtesis.

Capítol 1

La distribució Power Law

1.1 La distribució Power Law

1.1.1 Introducció

La distribució Power Law (PL) és una distribució de probabilitat que depèn de dos paràmetres, es caracteritza per concentrar la major part de la probabilitat en els valors inicials i tenir una cua dreta molt llarga. Els paràmetres són α, el paràmetre d'escala, i x_{min}, la cota inferior del suport de la distribució. El suport és infinit, ja que la variable pot prendre valors més grans o iguals que x_{min}. La distribució existeix tan per a dades contínues com per a dades discretes, però en aquest treball ens centrarem en el cas discret.

Hi ha diverses situacions reals que s'adiuen amb la descripció de la PL. En general, les dades que corresponen a freqüències de freqüències s'ajusten de forma apropiada per aquesta distribució quan hi ha molts esdeveniments que s'observen poques vegades i pocs esdeveniments que s'observen molts cops. A la Figura 1.1 es pot apreciar la distribució de probabilitats PL amb paràmetres $x_{min} = 1$ i 5, i $\alpha = 1.5$ i 2.5. Seguidament s'exposen uns quants exemples de situacions reals que s'ajusten bé per una PL, classificats per temes.

- **Lingüística**: si pensem en la freqüència d'aparició de les paraules en un text, podem veure que el patró que segueix la distribució d'aquesta freqüència, com va demostrar G. K. Zipf [1] estudiant les paraules del llibre *Moby Dick* de Herman Melville, és el mateix que el de la PL. És a dir, sabem que s'observaran poques paraules que es repeteixin moltes vegades i n'hi haurà moltes que apareixeran molt poc. També s'ha demostrat que segueixen

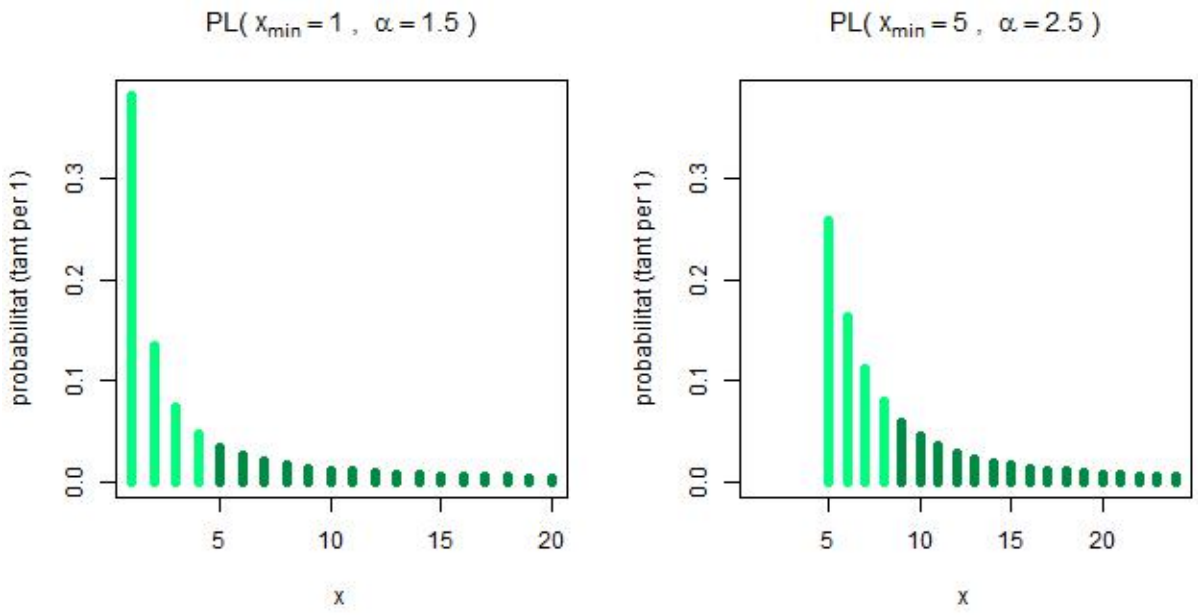

Figura 1.1: Distribucions $PL(x_{min}, \alpha)$ discretes amb paràmetres $x_{min} = 1$ i 5, i $\alpha = 1.5$ i 2.5.

aquesta distribució la freqüència de parelles de lletres en un text; el nombre de publicacions d'autors, ja que hi ha pocs autors que publiquen molt i molts que publiquen poc; el nombre de llibres per nombre de pàgines i la freqüència de cites d'un article, entre d'altres. També n'és un exemple el nombre de còpies venudes dels llibres més venuts (*bestsellers*) en un període de temps determinat. A l'article [2] s'estudia el cas dels *bestsellers* que hi va haver als Estats Units entre 1895 i 1965.

- **Demografia**: una altra situació que s'adapta al perfil de la PL és la que correspon a l'estudi del nombre d'habitants de les ciutats. Per una banda, hi ha poques ciutats que tenen molts habitants i, per l'altra, hi ha un nombre força elevat de ciutats i pobles que en tenen pocs. N'és un cas particular el nombre d'habitants de les ciutats dels Estats Units comptabilitzats al Cens de l'any 2000, que també s'estudia a l'article [2]. També es pot trobar el patró de la PL a la població de zones metropolitanes, de societats tribals, d'àrees regionals, etc.

- **Sismologia**: la intensitat d'un terratrèmol (amplitud màxima de moviment durant el terratrèmol segons l'escala de Richter) també s'ajusta de forma apropiada a la distribució PL. Així ho va demostrar Newman (veure [3]) amb les dades dels terratrèmols que hi va haver a Califòrnia entre el mes de gener de 1910 i el maig de 1992, amb dades procedents

del *Berkeley Earthquake Catalog*.

- **Biologia**: per exemple, la variable que ens dóna el nombre de gèneres[1] per nombre d'espècies segueix una distribució PL. Hi ha moltes espècies amb pocs gèneres i poques que en tinguin molts. També es distribueixen segons una PL els graus (per exemple el nombre de parelles d'interacció diferents) de proteïnes a la xarxa d'interacció de proteïnes del llevat *Saccharomyces cerevisiae* (veure [2]). També s'ha observat que segueixen el mateix patró els graus de metabòlits[2] en la xarxa metabòlica del bacteri *Escherichia coli* (veure [2]).

- **Internet i xarxes socials**: a Internet hi ha diversitat i llibertat per escollir, cosa que crea desigualtat. A més diversitat, més desigualtat. En sistemes on molta gent pot escollir lliurement entre diverses opcions, un petit grup del total rep una quantitat desproporcionada de trànsit (o atenció) mentre que un gran nombre de pàgines reben poca atenció, per tant es pot observar el patró de la PL. En són exemples el nombre de visites diàries a un lloc web; el nombre d'enllaços a pàgines web; el nombre de bytes de dades rebuts com a resultat de peticions web (HTTP) individuals, etc. En termes generals, aquest darrer exemple correspon a la distribució de la grandària dels arxius transmesos per Internet. I, en aquest cas, té sentit que els arxius de la majoria de transmissions tinguin una grandària petita i només els de poques transmissions tinguin una grandària molt gran.

  També es pot observar el comportament de la PL en el volum de comunicacions electròniques, que són inversament proporcionals a la distància geogràfica. A l'article [4] ho estudien per dos conjunts de dades: vincles de membres de Facebook i comunicacions via correu electrònic.

- **Economia**: l'economista V. Pareto va observar que la riquesa segueix, amb les seves paraules, un "desequilibri predicitible", ja que només el 20% de la població posseeix el 80% de la riquesa, el que s'anomena la Regla del 80/20. Amb aquesta mateixa idea també podem estudiar les empreses a qualsevol nivell geogràfic (estatal, continental, mundial, etc.), que es poden ordenar segons el nombre d'empleats, ingressos, beneficis, capitalització borsària, etc. Aquesta ordenació també es pot aplicar a certes indústries o bé a certes localitzacions

[1] Gènere: [biol] Categoria taxonòmica intermèdia entre la família i l'espècie que inclou espècies amb una sèrie de caràcters comuns (veure www.enciclopedia.cat).

[2] Metabòlit: [bioq] Nom genèric dels composts químics existents en els éssers vius que participen en les reaccions químiques del metabolisme intermediari o hi són produïts (veure www.enciclopedia.cat).

geogràfiques. El debat que es manté sobre les dades econòmiques és si aquestes en general es distribueixen segons una PL, una log-normal o alguna altra distribució.

Un cas particular que sí que es té certesa que segueix una PL en economia el trobem a l'article [3], on s'estudia la distribució de la riquesa (patrimoni net) dels habitants més rics dels Estats Units segons les dades publicades per la revista *Forbes* l'octubre de 2003.

- **Societat**: es pot observar el comportament de la PL en la intensitat (mesurada segons el nombre de morts en combat) de les guerres. A l'article [2] s'analitza la intensitat de les guerres civils entre 1816 i 1980. Un altre exemple és la severitat (mesurada com el nombre de morts directes) dels atacs terroristes, que s'estudia a [5] pels atacs a tot el món des del febrer de 1968 al juny de 2006.

Les situacions presentades demostren que la distribució PL és de gran utilitat a l'hora d'ajustar dades discretes no negatives, i d'aquí la seva importància. A més, també hem vist que es tracta d'una distribució amb la que es treballa molt actualment en l'àmbit de les noves tecnologies, la qual cosa li confereix una especial rellevància. A continuació passem a exposar-la i a esmentar-ne les principals característiques.

1.1.2 Distribució de probabilitat i funcions relacionades

El nom de Power Law prové del fet que les probabilitats evolucionen com una potència del valor observat. Com veurem a continuació, l'exponent és negatiu i superior a la unitat, cosa que indica que les probabilitats decreixen com a funció de la observació.

Si X és una variable aleatòria (v.a.) amb suport els reals més grans o iguals que un determinat valor $x_{min} > 0$, es diu que X es distribueix segons una PL de paràmetres x_{min} i α, ho denotem per $X \sim PL(x_{min}, \alpha)$, si i només si la seva funció de densitat és la següent:

$$p(x) = \begin{cases} \frac{\alpha-1}{x_{min}} \left(\frac{x}{x_{min}}\right)^{-\alpha} & \forall\, x \geq x_{min}, \\ 0 & \text{altrament.} \end{cases}$$

per a $x \in \mathbb{R}$ i $\alpha > 1$. Aquesta distribució també es coneix amb el nom de distribució de Pareto.

Tal com ja hem esmentat, en aquest projecte ens centrarem únicament en el cas de dades

discretes. Si X és una v.a. amb suport els enters més grans o iguals que un cert valor x_{min} no negatiu i diferent de zero, podem dir que es distribueix segons una PL discreta, també anomenada Pareto discreta, $X \sim PL(x_{min}, \alpha)$, si i només si la seva funció de probabilitat és:

$$p(x) = P(X = x) = \begin{cases} \frac{x^{-\alpha}}{\zeta(\alpha, x_{min})} & \forall\, x \geq x_{min}, \\ 0 & \text{altrament.} \end{cases} \tag{1.1}$$

on $x_{min} > 0$, $\alpha > 1$ i el denominador és la constant normalitzadora, que és la funció zeta de Hurwitz i és igual a:

$$\zeta(\alpha, x_{min}) = \sum_{k=x_{min}}^{\infty} k^{-\alpha} = \sum_{n=0}^{\infty} (n + x_{min})^{-\alpha}\,. \tag{1.2}$$

A partir de la funció de probabilitat podem veure que la distribució PL té la característica que el logaritme de la probabilitat és lineal com a funció del logaritme del valor esperat. N'hi ha prou amb prendre logaritmes a l'equació (1.1) per comprovar-ho.

$$\ln p(x) = \ln \frac{x^{-\alpha}}{\zeta(\alpha, x_{min})} \;\Rightarrow\; \ln p(x) = -\alpha \ln x - \ln \zeta(\alpha, x_{min})\,. \tag{1.3}$$

Per tant, si dibuixem la mateixa distribució que a la Figura 1.1 en escala logarítmica, obtenim una recta, tal i com es mostra a la Figura 1.2. On podem observar que a l'esquerra del gràfic els punts estan molt dispersos, al centre del gràfic els punts es troben mitjanament a prop i a la dreta hi ha un nombre elevat de punts. Per tant, hi distingim els primers valors de la variable, que concentren la major part de la probabilitat, i la cua, que és llarga i té probabilitats petites.

A continuació es calculen tres funcions importants en el càlcul de probabilitats en general i, en especial, en anàlisi de la supervivència. Per veure'n el comportament, s'han dibuixat per les dues distribucions PL utilitzades anteriorment i es poden trobar a la Figura 1.3. A més, s'interpreten aquestes tres funcions per a una variable amb distribució PL, en particular, aquesta variable correspon a les freqüències del nombre d'enllaços que tenen les pàgines web. Aplicant la definició de la distribució PL, intuïm que hi haurà moltes pàgines web que tinguin pocs enllaços i que n'hi haurà poques que en tinguin molts.

La distribució de probabilitat complementària, $P(x)$, es defineix com la probabilitat que la variable sigui superior o igual a un valor concret, seguint l'article [2]. Per tant, pel cas particular del nombre d'enllaços de les pàgines web, $P(x)$ serà la probabilitat que una pàgina web tingui

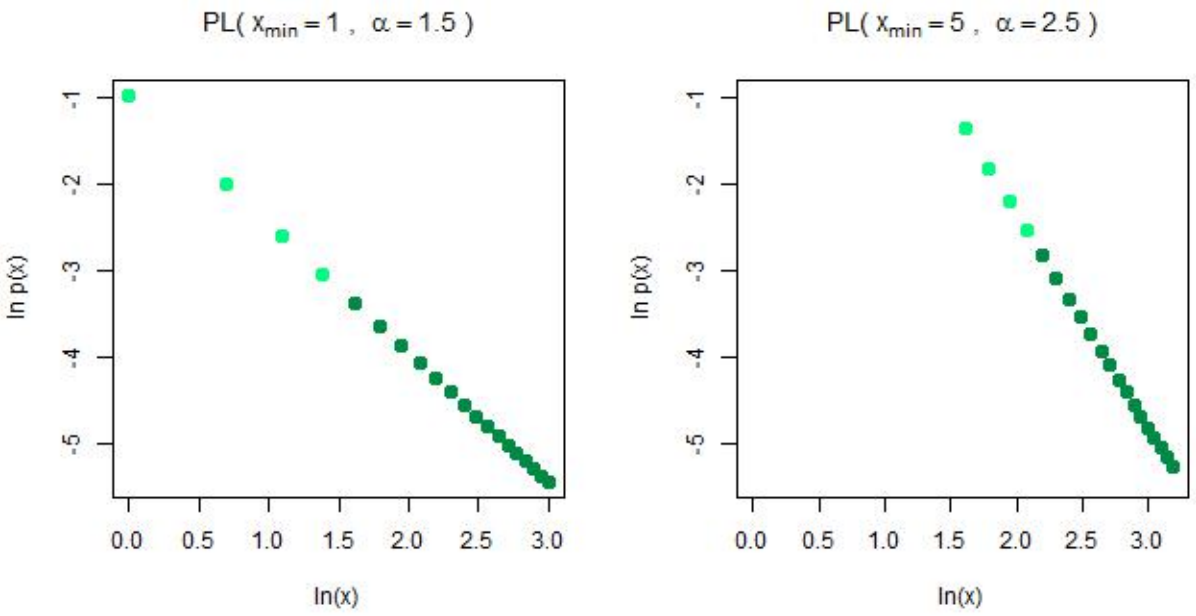

Figura 1.2: Distribució PL discreta amb $x_{min} = 1$ i $\alpha = 1.5$ representada en eixos logarítmics.

x o més enllaços.

$$P(x) = P(X \geq x) = \sum_{k=x}^{\infty} p(k) = \sum_{k=x}^{\infty} \frac{k^{-\alpha}}{\zeta(\alpha, x_{min})} = \frac{\zeta(\alpha, x)}{\zeta(\alpha, x_{min})}\,, \ \forall\, x \geq x_{min}\,. \tag{1.4}$$

Més endavant també necessitarem la funció de supervivència, $\overline{F}(x)$, que ens indica la probabilitat que la variable sigui superior a un valor concret $x \geq x_{min}$ i és igual a:

$$\overline{F}(x) = P(X > x) = \sum_{k=x+1}^{\infty} p(k) = \sum_{k=x+1}^{\infty} \frac{k^{-\alpha}}{\zeta(\alpha, x_{min})} = \frac{\zeta(\alpha, x+1)}{\zeta(\alpha, x_{min})}\,, \ \forall\, x \geq x_{min}\,. \tag{1.5}$$

És clar que per $x < x_{min}$, $\overline{F}(x) = 1$. Pel cas del nombre d'enllaços de les pàgines web, la funció de supervivència ens diu la probabilitat que una pàgina web tingui més de x enllaços.

La darrera funció que considerem és la funció de risc, $r(x;\alpha)$. La funció de risc d'una v.a. discreta es defineix com el quocient entre la funció de probabilitat (1.1) en x, i la funció de supervivència (1.5) en $x-1$. Així, per la PL discreta, pren la forma següent:

$$r(x;\alpha) = \frac{p(x)}{\overline{F}(x-1)} = \frac{x^{-\alpha}}{\zeta(\alpha, x_{min})} : \frac{\zeta(\alpha, x)}{\zeta(\alpha, x_{min})} = \frac{x^{-\alpha}}{\zeta(\alpha, x)}\,, \ \forall\, x \geq x_{min}\,. \tag{1.6}$$

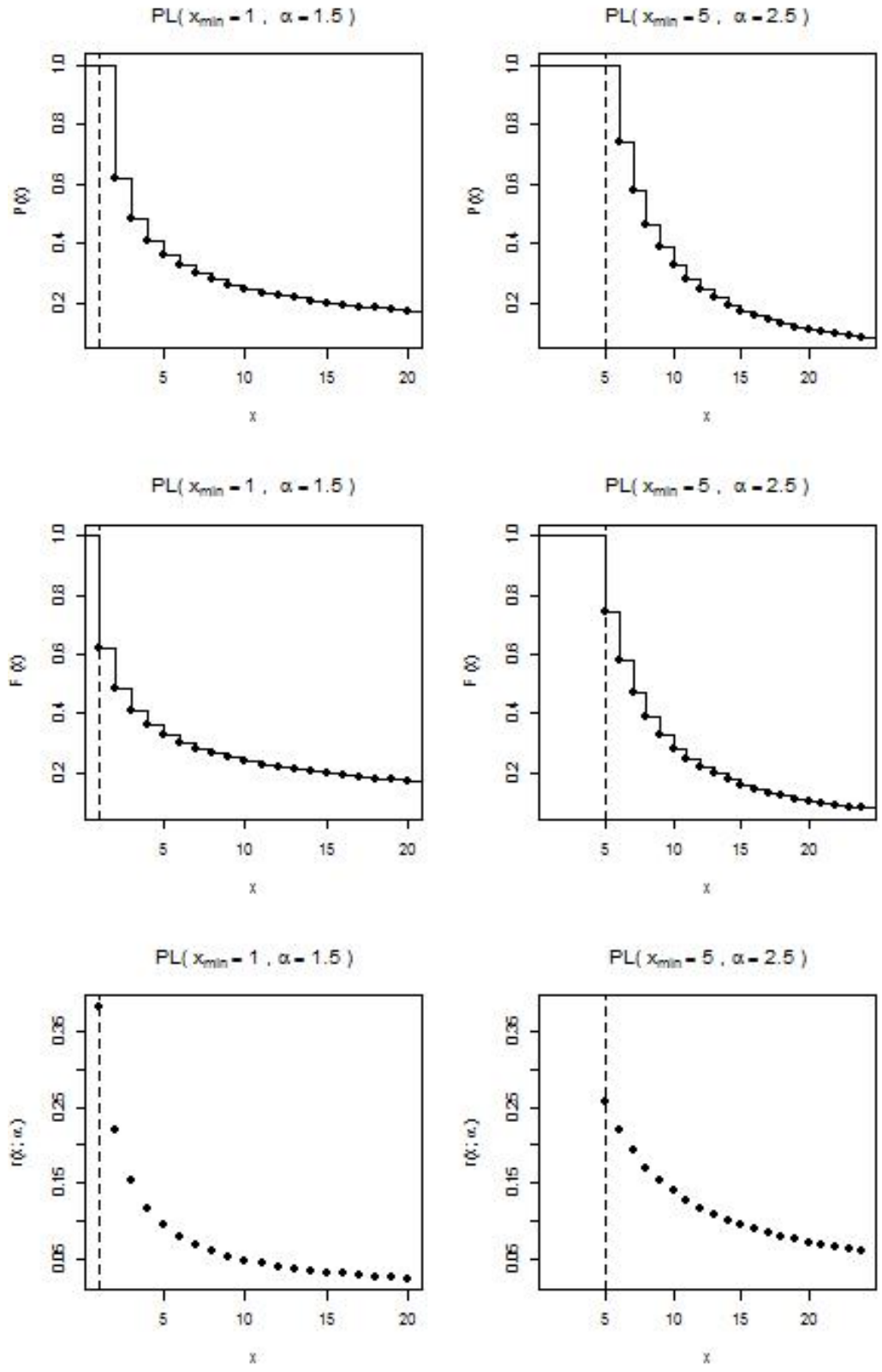

Figura 1.3: Per files, funcions de probabilitat complementàries (1.4), de supervivència (1.5) i de risc (1.6) per les distribucions $PL(x_{min}, \alpha)$ discretes amb paràmetres $x_{min} = 1$ i 5, i $\alpha = 1.5$ i 2.5.

La funció de risc es pot interpretar en termes de probabilitat en el cas discret però no en el cas continu. Així, si tornem a l'exemple dels enllaços de les pàgines web, $r(x;\alpha)$ ens indica la probabilitat que una pàgina web tingui x enllaços sabent que en té com a mínim x.
Observi's que aquesta definició de funció de risc també es pot veure en termes de la funció de supervivència com:

$$r(x;\alpha) = \frac{\overline{F}(x-1) - \overline{F}(x)}{\overline{F}(x-1)}\,.$$

Propietat 1.1. *La funció de risc de la PL discreta, $r(x;\alpha)$, és monòtona decreixent.*

Demostració. Hem de comprovar que es compleix que $r'(x;\alpha) \leq 0$, on $r'(x;\alpha)$ designa la primera derivada de $r(x;\alpha)$ respecte de x. Com que

$$r'(x;\alpha) = \frac{-\alpha\, x^{-\alpha-1}\,\zeta(\alpha,x) - \zeta'(\alpha,x)}{\left[\zeta(\alpha,x)\right]^2}\,, \tag{1.7}$$

i tenint en compte que

$$\zeta'(\alpha,x) = \left(\sum_{k=x}^{\infty} k^{-\alpha}\right)' = \sum_{k=x}^{\infty} -\alpha\, k^{-\alpha-1} = -\alpha \sum_{k=x}^{\infty} k^{-\alpha-1} = -\alpha\,\zeta(\alpha+1,x)\,,$$

podem reescriure (1.7) com

$$r'(x;\alpha) = \frac{-\alpha\, x^{-\alpha}}{\left[\zeta(\alpha,x)\right]^2}\left[\frac{\zeta(\alpha,x)}{x} + \alpha\,\zeta(\alpha+1,x)\right]\,.$$

El primer terme d'aquest producte és negatiu i el segon és positiu, per tant, el resultat de multiplicar-los és negatiu. De manera que hem demostrat que la derivada de la funció de risc és negativa i, en conseqüència, podem afirmar que es tracta d'una funció monòtona decreixent. □

Una definició alternativa a la funció de risc discreta que s'utilitza habitualment apareix en l'article [6], i concretament és igual a:

$$r^*(x;\alpha) = \left[\ln \frac{\overline{F}(x-1)}{\overline{F}(x)}\right]\,. \tag{1.8}$$

Aquesta nova definició de funció de risc aconsegueix que les propietats de la funció en el cas

discret i en el cas continu siguin molt semblants. Cosa que no es compleix amb la definició inicial.

Pel cas concret de la PL discreta i amb la definició (1.8),

$$r^*(x;\alpha) = \ln\left[\frac{\zeta(\alpha,x)}{\zeta(\alpha,x_{min}} : \frac{\zeta(\alpha,x+1)}{\zeta(\alpha,x_{min}}\right] = \ln\left[\frac{\zeta(\alpha,x)}{\zeta(\alpha,x+1)}\right] .$$

1.1.3 Moments

Les expressions de l'esperança i la variància de la v.a. discreta $X \sim PL(\alpha, x_{min})$ es dedueixen a continuació.

L'esperança,

$$E(X) = \sum_{x\geq x_{min}} x\, p(x) = \sum_{x=x_{min}}^{\infty} x\, \frac{x^{-\alpha}}{\zeta(\alpha,x_{min})} = \frac{\zeta(\alpha-1,x_{min})}{\zeta(\alpha,x_{min})} .$$

En general, el moment d'ordre r, $r \geq 1$, es pot calcular de la forma següent:

$$E(X^r) = \sum_{x\geq x_{min}} x^r\, p(x) = \frac{\zeta(\alpha-r,x_{min})}{\zeta(\alpha,x_{min})} .$$

Utilitzant les dues igualtats anteriors s'obté l'expressió de la variància de la variable:

$$\begin{aligned} var(X) &= E(X^2) - E(X)^2 = \frac{\zeta(\alpha-2,x_{min})}{\zeta(\alpha,x_{min})} - \left(\frac{\zeta(\alpha-1,x_{min})}{\zeta(\alpha,x_{min})}\right)^2 = \\ &= \frac{\zeta(\alpha-2,x_{min})\,\zeta(\alpha,x_{min}) - \zeta(\alpha-1,x_{min})^2}{\zeta(\alpha,x_{min})^2} . \end{aligned}$$

1.1.4 Estimació dels paràmetres

En aquest apartat es presenten els estimadors dels paràmetres de la distribució PL discreta que es proposen a l'article *Power-law distributions in empirical data* [2].

Atès que la naturalesa dels dos paràmetres d'aquesta distribució és ben diferent, la seva estimació

es fa en dos passos. Primer s'ha d'estimar el paràmetre x_{min}, que després ens permetrà estimar el paràmetre d'escala α. Per tant, necessitem que l'estimació del primer paràmetre sigui bona per tal que la del segon també ho sigui.

A l'hora de treballar amb dades reals, sovint passa que si es dibuixa el logaritme de la probabilitat empírica com a funció del logaritme del valor observat, la gràfica és lineal a partir d'un cert valor, però no acostuma a ser-ho pels primers valors. Això porta als investigadors a seguir diferents estratègies a l'hora d'ajustar les dades. Una d'aquestes estratègies consisteix en dividir les dades observades en dos grups, de forma que el primer grup conté les observacions més petites que x_{min} i el segon, les més grans o iguals que x_{min}. Les observacions d'aquest segon grup són les que s'assumeix que es distribueixen segons una PL, la resta segueixen qualsevol altra distribució. Si subestimem el valor de x_{min}, estarem ajustant una distribució PL a dades que no la segueixen. Si sobreestimem x_{min}, perdem observacions que realment segueixen la distribució PL. Per tant, en ambdós casos trobarem estimadors esbiaixats. A l'article [2], en el que ens basem en aquest apartat, els autors estudien gràficament l'efecte de l'estimació de x_{min} sobre l'estimació del paràmetre d'escala. Arriben a la conclusió que subestimar x_{min} fa que l'estimació màxim versemblant d'α ($\hat{\alpha}$) es desviï ràpidament del valor real del paràmetre. En canvi, quan se sobreestima x_{min}, la diferència amb el valor real del paràmetre x_{min} ha de ser més gran perquè l'estimació del paràmetre d'escala es desviï significativament del valor real.

Els diferents mètodes que s'utilitzen per estimar x_{min} s'expliquen a continuació:

- **Mètode visual**. Podem detectar visualment on comença la PL de dues maneres:
  - Identificant el valor de x a partir del qual el dibuix que genera l'histograma de les dades amb eixos logarítmics és una recta.
  - Dibuixant $\hat{\alpha}$ com a funció de $\hat{x}_{min}$ i veure a partir de quin punt el valor d'$\hat{\alpha}$ s'estabilitza.

  Es tracta d'un mètode molt subjectiu i sensible al soroll.

- **Mètode basat en el *Bayesian Information Criterion* (BIC)**. Suposem que es fixa un valor $\hat{x}_{min}$ per a x_{min}, aleshores, per sobre de $\hat{x}_{min}$ ajustem una PL per a dades discretes i, per sota, assignem a cada valor la freqüència observada en aquest punt, és a dir, $P(X = k) = p_k$, $1 \leq k < \hat{x}_{min}$, on p_k representa la freqüència observada del valor k.

Si volem fer l'estimació dels paràmetres dividint les dades en dos grups, no podem aplicar directament el mètode de màxima versemblança perquè el nombre de paràmetres del model no és fixe, depèn de x_{min}. De fet, el nombre de paràmetres a estimar és exactament x_{min}, atès que hi ha $x_{min} - 1$ paràmetres inicials més el paràmetre α de la PL. Per tant, si utilitzem la versemblança, sempre la podrem millorar augmentant el nombre de paràmetres, cosa que no ens interessa perquè farem massa gran x_{min} i perdrem observacions de la PL. És a dir, ens cal aplicar un mètode que ens permeti utilitzar la versemblança i que penalitzi el fet d'afegir paràmetres al model.

Com a alternativa al mètode de màxima versemblança es proposa maximitzar el logaritme de la versemblança marginal, que és equivalent al BIC i té la forma:

$$\ln P(x|x_{min}) \simeq \mathcal{L} - \frac{1}{2} x_{min} \ln N \,,$$

on $\mathcal{L}$ és el valor màxim de la versemblança per cada x_{min} i N és el nombre total d'observacions.

El mètode BIC proposa maximitzanr aquesta funció respecte x_{min} per tal de trobar el valor estimat $\hat{x}_{min}$.

Val a dir que hi ha situacions en què el BIC té tendència a subestimar el valor de x_{min} perquè el nombre de paràmetres utilitzats per ajustar les dades que es troben per sota de x_{min} pot ser més petit que $x_{min} - 1$. Per tant, en aquestes situacions, el fet de subestimar x_{min} ens portarà a un estimador esbiaixat del paràmetre d'escala.

- **Mètode basat en les distàncies**. Escollirem el valor de $\hat{x}_{min}$ que faci que la funció de probabilitat estimada i l'empírica siguin el més semblants possible per sobre de $\hat{x}_{min}$. Quantificarem la distància entre les funcions de probabilitat utilitzant l'estadístic de Kolmogorov-Smirnov (KS), que és la distància màxima entre les dues funcions:

$$D = \max_{x \geq x_{min}} |S(x) - P(x)| \,,$$

on $S(x)$ i $P(x)$ són, respectivament, les distribucions de probabilitat complementàries

empírica i teòrica (1.4). El valor estimat $\hat{x}_{min}$ serà el valor que faci que D sigui mínima.

Suposem ara que el valor de x_{min} és conegut o que l'hem estimat a partir d'algun dels mètodes proposats anteriorment, i que ens proposem estimar el paràmetre d'escala. Alguns dels mètodes més utilitzats són els següents:

- **Histograma**. Es tracta de dibuixar l'histograma de les dades utilitzant eixos logarítmics, de manera que el que es dibuixarà serà la funció (1.3). Així, si les dades segueixen una distribució PL, el gràfic que obtindrem serà una recta i l'estimació d'α serà el pendent de la mateixa.

- **Màxima versemblança**. Si suposem que disposem d'una mostra de N observacions procedents d'una distribució PL, la funció de versemblança es pot escriure com:

$$\mathcal{L}(\alpha, x_{min}; x_1, \ldots, x_N) = \prod_{i=1}^{N} \frac{x_i^{-\alpha}}{\zeta(\alpha, x_{min})} = \zeta(\alpha, x_{min})^{-N} \left[\prod_{i=1}^{N} x_i\right]^{-\alpha} . \tag{1.9}$$

  El logaritme d'aquesta funció és

$$l(\alpha, x_{min}; x_1, \ldots, x_N) = -N \ln \zeta(\alpha, x_{min}) - \alpha \sum_{i=1}^{N} \ln x_i \, . \tag{1.10}$$

  Per trobar $\hat{\alpha}$, l'estimador màxim versemblant (MLE, *Maximum Likelihood Estimator*) del paràmetre d'escala, derivem el logaritme de la versemblança i l'igualem a zero. D'aquesta manera arribem a l'equació (1.11), que s'ha de resoldre numèricament.

$$\frac{\zeta'(\hat{\alpha}, x_{min})}{\zeta(\hat{\alpha}, x_{min})} = -\frac{1}{N} \sum_{i=1}^{N} \ln x_i \, , \tag{1.11}$$

  on $\zeta'(\hat{\alpha}, x_{min})$ denota la primera derivada de la funció $\zeta(\hat{\alpha}, x_{min})$ respecte el primer paràmetre.

  Una altra via per trobar $\hat{\alpha}$ és maximitzar directament la funció de versemblança (1.9) o el seu logaritme (1.10), que a voltes pot resultar ser més senzill que resoldre l'equació (1.11).

  Es pot demostrar que el MLE és consistent, és a dir, que la successió d'estimadors del paràmetre obtinguda a l'augmentar la grandària mostral convergeix en probabilitat al

valor real d'aquest. L'estimador també és asimptòticament eficient, o el que és el mateix, assoleix la cota de Cramér-Rao (variància mínima) per mides mostrals grans.

L'error estàndard de l'estimació obtinguda per màxima versemblança es calcula de la forma següent:

$$\sigma = \frac{1}{\sqrt{N\left[\frac{\zeta''(\hat{\alpha},x_{min})}{\zeta(\hat{\alpha},x_{min})} - \left(\frac{\zeta'(\hat{\alpha},x_{min})}{\zeta(\hat{\alpha},x_{min})}\right)^2\right]}}\,.$$

- **Màxima versemblança aproximada**. Si considerem que la variable observada prové d'una distribució contínua que ha estat arrodonida a l'enter més proper, podem trobar una expressió aproximada per a $\zeta(\alpha, x_{min})$ i la seva derivada, el quocient de les quals intervé en el procés de maximitzar el logaritme de la versemblança. Aquesta manera de procedir permet estimar el paràmetre d'escala a partir de:

$$\hat{\alpha} \simeq 1 + N\left[\sum_{i=1}^{N} \ln \frac{x_i}{x_{min} - \frac{1}{2}}\right]^{-1}\,.$$

Aquest estimador dóna bons resultats a la pràctica i és molt més senzill de calcular que l'estimador exacte.

L'equació que ens permet trobar l'estimació del paràmetre d'escala és gairebé la mateixa que en el cas continu (només difereixen en el terme $\frac{1}{2}$, que no apareix en el cas de dades contínues). És per això que, en aquest cas, se suggereix calcular l'error de l'estimació utilitzant la fórmula corresponent al cas continu, que és la següent:

$$\sigma = \frac{\hat{\alpha} - 1}{\sqrt{N}} + \mathrm{O}\left(1/N\right)\,.$$

- **Dades contínues**. Una altra manera d'estimar α és suposant que les dades són contínues i utilitzar l'estimador màxim versemblant per aquest cas, que és:

$$\hat{\alpha} \simeq 1 + N\left[\sum_{i=1}^{N} \ln \frac{x_i}{x_{min}}\right]^{-1}\,.$$

Aquesta alternativa no té gaire sentit perquè té la mateixa dificultat de càlcul que el cas anterior, i ens porta a resultats menys exactes que el MLE aproximat.

En l'article [7], es comparen diferents formes d'estimar el paràmetre α quan $x_{min} = 1$. L'article conclou que la màxima versemblança és molt millor que les basades en mètodes gràfics. L'estimador màxim versemblant és molt més robust que la resta.

1.1.5 Bondat de l'ajust

Per decidir si el model PL és adequat per tal d'ajustar les dades observades es poden aplicar diferents tests i/o calcular indicadors de bondat de l'ajust. Se n'ha escollit un de cada, el test χ^2 de Pearson i l'*Akaike Information Criterion* (AIC), i s'expliquen a continuació. Val a dir que se suposarà en ambdós casos que la variable amb la que es treballa és una PL discreta amb x_{min} fix i amb el paràmetre α estimat per màxima versemblança.

- **χ^2 de Pearson**. Donada una v.a. X i una mostra aleatòria simple (m.a.s.) $x_1, x_2, \ldots, x_N$ d'aquesta, volem saber si la distribució PL és adequada per les dades, per tant, plantegem el contrast següent:

$$H_0 : X \sim \text{distribució PL}$$
$$H_1 : X \nsim \text{distribució PL.}$$

  Pearson, l'any 1928 [8], proposava realitzar aquest test d'hipòtesi per mitja de l'estadístic χ^2 de Pearson, que es calcula com:

$$X^2 = \sum_{i=1}^{l} \frac{(o_i - e_i)^2}{e_i},$$

  on l és el nombre de nivells (categories que s'han observat), o_i correspon a la freqüència observada del nivell i, i e_i correspon a la freqüència esperada del mateix nivell.

  Cal tenir en compte que la freqüència observada a l'últim nivell es calcularà a partir de $N - \sum_{i=1}^{l-1} e_i$, amb N el nombre total d'observacions. I que per poder aplicar el test cal que $e_i \geq 5$, per a tot valor de i.

  Pearson va demostrar que, si la hipòtesi nul·la és certa, l'estadístic segueix una distribució χ^2_{l-p-1}, on p correspon al nombre de paràmetres de la distribució que hauran estat estimats

a partir de les dades. Cal tenir en compte que, quan els graus de llibertat (k) de la distribució χ^2 són més grans que 50, aquesta s'aproxima per la $Normal(k, 2k)$. Aquesta aproximació la utilitzarem en alguns dels casos reals que analitzarem al Capítol 3, atès que es tracta de conjunts de dades molt grans.

- ***Akaike Information Criterion* (AIC)**. Es tracta d'una mesura de bondat de l'ajust que ens permet comparar models que no han de ser necessàriament aniuats. Es calcula mitjançant la versemblança i penalitzant en funció del nombre de paràmetres estimats. Com més petit sigui l'AIC, més bo serà el model.

$$AIC = 2\,p - 2\,l(x_{min}, \hat{\alpha}; x_1, x_2, \ldots, x_N)\,,$$

on p és el nombre de paràmetres estimats, que en aquest cas és 1, x_{min} és fix, $\hat{\alpha}$ és el MLE del paràmetre de la distribució PL i $l(x_{min}, \hat{\alpha}; x_1, x_2, \ldots, x_N)$ és el valor màxim del logaritme de la versemblança.

Més endavant, també s'ha considerat un altre test per avaluar la bondat de l'ajust. Es tracta del test de Kolmogorov-Smirnov, que és no paramètric i, per tant, és d'una naturalesa molt diferent als dos mètodes explicats en aquest apartat. Per això i perquè requereix realitzar simulacions per tal d'establir els valors crítcs, se li ha dedicat un capítol sencer, el Capítol 4.

1.2 Distribució Zipf

En aquest projecte estudiem la distribució PL uniparamètrica que s'obté quan fixem $x_{min} = 1$, que serà la base de la generalització proposada en el proper capítol. La $PL(x_{min} = 1, \alpha)$ és una distribució que rep diversos noms, a continuació s'expliquen els que es poden trobar al llibre *Univariate Discrete Distributions* [9]: distribució Zipf, nom que li ve donat pel seu descobridor George Kingsley Zipf, que estudiava la freqüència d'aparició de les paraules en un text; Zipf-Estoup Law, el taquígraf Jean-Baptiste Estoup també va detectar el comportament de les freqüències de les paraules dels textos uns anys abans que ho fes Zipf; distribució zeta de Riemann o distribució zeta, ja que en el càlcul de la funció de probabilitat hi intervé la funció

zeta de Riemann. A partir d'ara ens referirem a la distribució que volem estudiar amb el primer d'aquests noms, distribució Zipf.

Si X és una v.a. tal que $X \sim Zipf(\alpha)$, fixant $x_{min} = 1$ a l'equació (1.1), s'obté que la funció de probabilitat de X és igual a:

$$p(x) = P(X = x) = \begin{cases} \frac{x^{-\alpha}}{\zeta(\alpha)} & \text{si } x = 1, 2, 3, 4, \dots \\ 0 & \text{altrament.} \end{cases}$$

on $\alpha > 1$, $\zeta(\alpha)$ és la funció zeta de Riemann, que és el cas particular de la funció zeta de Hurwitz (1.2) per a $x_{min} = 1$. És a dir,

$$\zeta(\alpha, 1) = \zeta(\alpha) = \sum_{k=1}^{\infty} k^{-\alpha}\,.$$

El motiu pel qual la distribució es defineix per a valors d'α majors que la unitat no és cap altre que el fet que la funció zeta de Riemann només convergeix quan $\alpha > 1$.
Com que es tracta d'un cas particular de la distribució PL, el logaritme de la probabilitat segueix sent lineal com a funció del valor esperat (1.3).

Pel que fa a les funcions relacionades amb l'anàlisi de la supervivència, la distribució de probabilitat complementària en aquest cas concret pren la forma:

$$P(x) = P(X \geq x) = \frac{\zeta(\alpha, x)}{\zeta(\alpha)}\,,\ \forall\, x \geq 1\,,$$

la funció de supervivència és igual a:

$$\overline{F}(x) = P(X > x) = \frac{\zeta(\alpha, x+1)}{\zeta(\alpha)}\,,\ \forall\, x \geq 1\,,$$

i la funció de risc, com que en cas general no es veu afectada pel valor de x_{min}, segueix tenint la mateixa expressió que podem trobar a l'equació (1.6).

Pel que fa als moments de la distribució, l'esperança de la Zipf és igual a:

$$E(X) = \frac{\zeta(\alpha - 1)}{\zeta(\alpha)}\,.$$

En general, el moment d'ordre r de la Zipf es calcula a partir de:

$$E(X^r) = \frac{\zeta(\alpha - r)}{\zeta(\alpha)},$$

i, per tant, la variància calculada a partir dels moments d'ordre u i ordre dos, serà igual a:

$$var(X) = \frac{\zeta(\alpha - 2)\,\zeta(\alpha) - \zeta(\alpha - 1)^2}{\zeta(\alpha)^2}.$$

Atès que, tal com s'ha esmentat abans, la funció zeta de Riemann només convergeix per a valors d'α més grans que la unitat, l'esperança existirà només per a valors d'$\alpha > 2$ i la variància, per valors d'$\alpha > 3$. Així doncs, la distribució Zipf amb $\alpha \in (1, 2)$ existeix però no té ni esperança ni variància finites. Si $\alpha \in (2, 3)$, té esperança finita però no té variància. I, si $\alpha > 3$, ambdós moments són finits.

A la Figura 1.4 es mostren els gràfics de la funció de probabilitat de la distribució Zipf amb paràmetres $\alpha = 1.1$, 1.5, 2.5 i 5. S'observa que, per a valors petits d'α, la major part de la probabilitat es concentra en els valors inicials i després trobem una cua dreta molt llarga, malgrat que en el dibuix no s'acabi d'apreciar del tot ja que hem tallat el suport a 10. A mesura que anem augmentant el valor d'α, la probabilitat tendeix a concentrar-se molt més als dos o tres primers valors, i la cua dreta segueix sent llarga però té menys probabilitat.

També es poden dibuixar les distribucions de la figura anterior en eixos logarítmics, de manera que, com s'ha explicat abans (1.3), obtindrem rectes. Aquestes gràfiques es poden veure a la Figura 1.5. Com a conseqüència del fet que la funció logaritme creix ràpidament al principi, però la pendent es va suavitzant fins a ser molt petita, la densitat de punts és molt poca al principi de les gràfiques, i aquesta va augmentant a mida que augmenta el valor de x. Si comparem els quatre gràfics observem que, a mesura que augmenta el paràmetre α, s'incrementa el valor de el pendent atès que aquest coincideix exactament amb α. Per tant, a mesura que augmenta α, augmenta la probabilitat al primer valor i disminueix la probabilitat a la cua.

L'estimació dels paràmetres pel cas particular de la distribució Zipf és més senzilla, ja que no haurem d'estimar el valor de x_{min} perquè ha estat fixat igual a la unitat. Tal com ja hem esmentat, quan es treballa amb una $PL(\alpha, x_{min})$, l'estimació de x_{min} és una etapa del procés

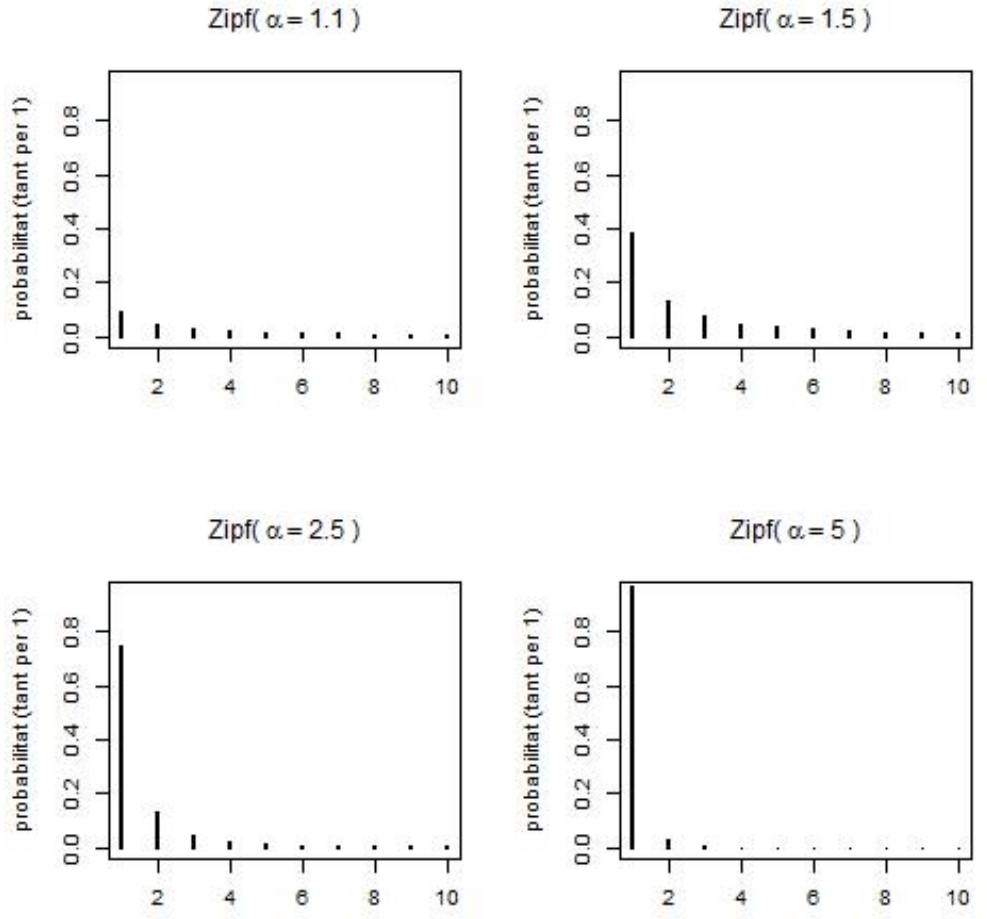

Figura 1.4: Funció de probabilitat de distribucions Zipf(α) amb $\alpha = 1.1, 1.5, 2.5$ i 5.

d'estimació que ens pot portar a resultats diversos en funció del mètode emprat i això afecta molt directament a l'estimació de l'altre paràmetre de la distribució. Per a la $Zipf(\alpha)$, només n'haurem d'estimar un, α. Ho farem aplicant el mètode de màxima versemblança. Donada una m.a.s. $x_1, x_2, \ldots, x_N$ d'una v.a. amb distribució $Zipf(\alpha)$, el logaritme de la funció de versemblança té la forma:

$$l(\alpha; x_1, \ldots, x_N) = -N \ln \zeta(\alpha) - \alpha \sum_{i=1}^{N} \ln x_i \,,$$

i caldrà trobar-ne el màxim numèricament.

Hem d'esmentar que, al tractar-se d'una distribució uniparamètrica, la seva flexibilitat és reduïda i és per això que té sentit generalitzar-la, que és l'objectiu d'aquest treball. Concretament,

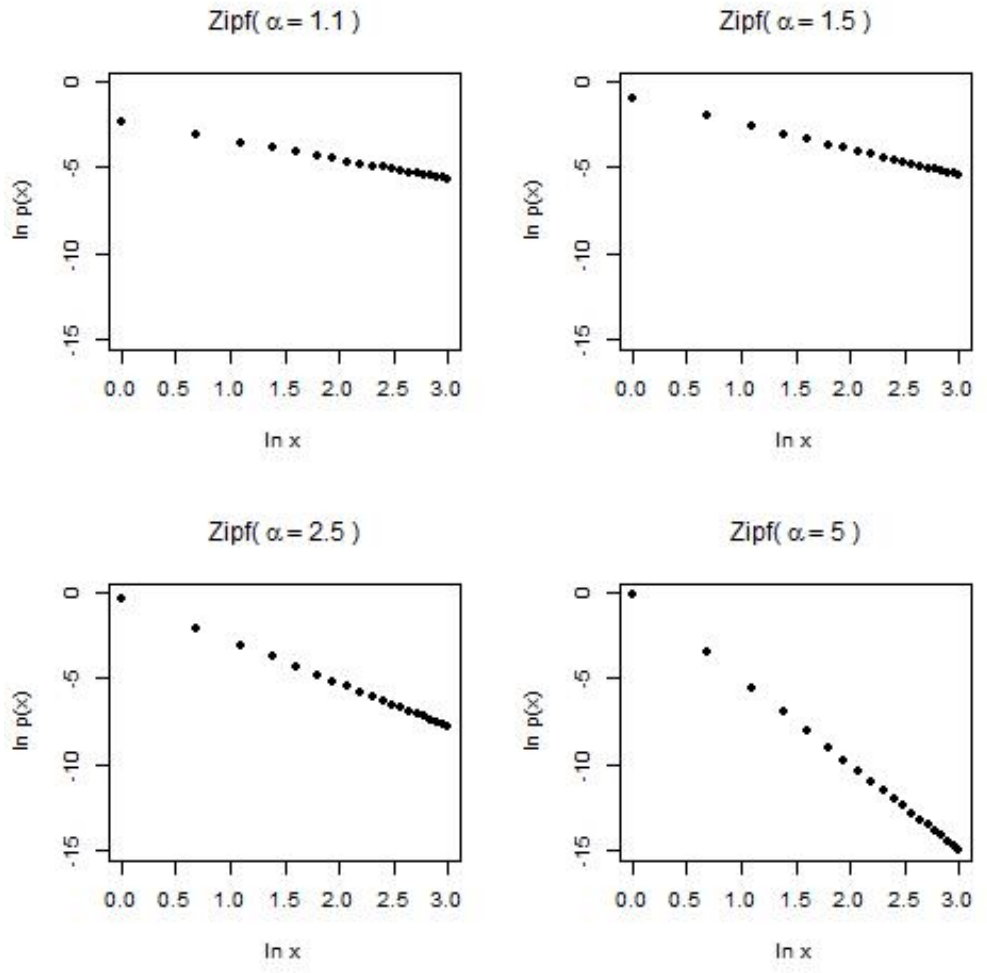

Figura 1.5: Funció de probabilitat de distribucions Zipf(α), dibuixada en eixos logarítmics, amb $\alpha = 1.1, 1.5, 2.5$ i 5.

el que s'observa a la bibliografia és que la distribució $Zipf(\alpha)$ ajusta bé la cua de la distribució empírica, però no els primers valors. Millorar l'ajust dels primers valors augmentant una unitat el nombre de paràmetres de la distribució és el que ens proposem en el proper capítol.

Capítol 2

Distribució Zipf Estesa

2.1 Transformació Marshall-Olkin

A. W. Marshall i I. Olkin presenten a l'article [10] un mètode que permet afegir un paràmetre a una família de distribucions per tal de generalitzar-la, procediment que anomenem transformació Marshall-Olkin (MO). Aquest mètode utilitza la funció de supervivència de la distribució per, mitjançant un nou paràmetre, β, definir una nova funció de supervivència, $\overline{G}$, que és la funció de supervivència de la distribució generalitzada. Es diu que la transformació MO generalitza la distribució inicial pel fet que aquesta darrera n'és un cas particular, tal com es veurà més endavant.

La generalització del model de probabilitats amb funció de supervivència $\overline{F}(x)$ proposada per Marshall i Olkin és el model de probabilitats que té funció de supervivència:

$$\overline{G}(x;\beta) = \frac{\beta\,\overline{F}(x)}{1-\overline{\beta}\,\overline{F}(x)} = \frac{\beta\,\overline{F}(x)}{F(x)+\beta\,\overline{F}(x)} \quad (-\infty < x < \infty, 0 < \beta < \infty)\,, \tag{2.1}$$

on $\overline{\beta} = 1-\beta$. Cal notar que els autors quan defineixen la transformació ho fan pensant en models de probabilitat continus, d'aquí que la variable X prengui valors reals.

Es pot observar que $\overline{G} = \overline{F}$ quan $\beta = 1$, amb la qual cosa, el model generalitzat conté el model inicial com a cas particular. Es tracta d'una transformació tal que si l'apliquem dos cops, continuem estant dins la mateixa família. Aquesta propietat els autors l'anomenen d'*estabilitat*.

La nova funció de risc associada a la distribució generalitzada s'expressa en funció de la funció

de risc original, $r_F(x)$, com:

$$r(x;\beta) = \frac{1}{1-\overline{\beta}\,\overline{F}(x)}\, r_F(x) \quad (-\infty < x < \infty)\,.$$

A l'article esmentat els autors demostren que la funció de risc de la distribució generalitzada compleix les següents condicions límit:

$$\lim_{x\to-\infty} r(x;\beta) = \lim_{x\to-\infty} r_F(x)/\beta, \quad \lim_{x\to\infty} r(x;\beta) = \lim_{x\to\infty} r_F(x)\,.$$

A més, l'expressió de $r(x;\beta)$ i la definició de $\overline{G}(x;\beta)$, permeten demostrar les següents fites:

$$\begin{aligned}
&r_F(x)/\beta \le r(x;\beta) \le r_F(x) && (-\infty < x < \infty,\ \beta \ge 1),\\
&r_F(x) \le r(x;\beta) \le r_F(x)/\beta && (-\infty < x < \infty,\ 0 < \beta \le 1),\\
&\overline{F}(x) \le \overline{G}(x;\beta) \le \overline{F}^{1/\beta}(x) && (-\infty < x < \infty,\ \beta \ge 1),\\
&\overline{F}^{1/\beta}(x) \le \overline{G}(x;\beta) \le \overline{F}(x) && (-\infty < x < \infty,\ \beta \ge 1),
\end{aligned}$$

que també figuren a l'article. També s'explica que el quocient $r(x;\beta)/r_F(x)$ és creixent quan $\beta \ge 1$ i decreixent quan $0 < \beta \le 1$.

En l'article original, Marshall i Olkin apliquen la transformació presentada a la distribució exponencial, i comparen la distribució obtinguda amb d'altres distribucions biparamètriques utilitzades enlloc de l'exponencial, com són la Weibull, la gamma i la lognormal. Els autors també generalitzen la distribució Weibull mitjançant la seva transformació, obtenint una distribució triparamètrica de la qual n'esmenten les principals propietats.

Per acabar els comentaris sobre aquest article, volem esmentar que independetment de quina sigui la distribució original escollida, la família resultant gaudeix de la propietat de ser el que s'anomena *geometric-extreme stable*. Aquesta propietat afirma que si $X_1, X_2, \ldots, X_N$ és una successió de v.a. independents i idènticament distribuïdes (i.i.d.) amb distribució en la família (2.1), i N segueix una distribució geomètrica, llavors les v.a. $Y_1 = \min(X_1, X_2, \ldots, X_N)$ i $Y_2 = \max(X_1, X_2, \ldots, X_N)$ també tenen distribucions en la mateixa família.

Aquesta propietat es basa en el fet que la suma geomètrica de variables aleatòries amb distribució geomètrica també segueix una distribució geomètrica. Entenent per suma geomètrica la suma

de variables aleatòries i.i.d., on el nombre de sumands s'ha obtingut a partir de la distribució geomètrica. En general, aquesta propietat no la compleixen altres distribucions de probabilitat definides en els enters no-negatius, com la Poisson, per exemple, d'aquí que no es pugui afirmar que la família és *Poisson-extreme stable*.

Hi ha hagut diversos autors que han utilitzat la transformació MO per tal de generalitzar distribucions (veure [11], [12], [13] i [14]). En tots els articles que hem consultat, la distribució inicial és de tipus continu i s'utilitza bàsicament en anàlisi de supervivència, assegurances i en enginyeria en l'estudi de la fiabilitat de components. Aquest és el cas de les distribucions exponencial, Weibull, Lomax o Pareto contínua.

En aquest treball generalitzem la distribució Zipf, i és molt probable que contribueixi el primer cas de generalització d'una distribució discreta a partir de la transformació MO.

2.2 Zipf Estesa

2.2.1 Definició

Anomenarem distribució MOEZipf, *Marshall-Olkin Extended Zipf*, a la transformació de MO de la distribució Zipf. Si suposem que X és una v.a. amb distribució $Zipf(\alpha)$, la seva funció de supervivència $\overline{F}(x)$ és de la forma (1.5), i aplicant la transformació MO obtenim la nova funció de supervivència de la nova família de distribucions, que depèn de dos paràmetres i és igual a:

$$\overline{G}(x;\alpha,\beta) = \frac{\beta\,\overline{F}(x)}{1-\overline{\beta}\,\overline{F}(x)} = \frac{\beta\,\zeta(\alpha,x+1)}{\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x+1)}\,. \tag{2.2}$$

D'ara endavant, anomenarem Y a una v.a. que segueix una distribució MOEZipf de paràmetres α i β, i ho denotarem per $Y \sim MOEZipf(\alpha,\beta)$.

Com que es tracta d'una v.a. discreta, podem calcular la seva funció de probabilitat seguint el procés que es detalla a continuació.

$$
\begin{aligned}
P(Y=x) &= P(Y\geq x)-P(Y\geq x+1)=\overline{G}(x-1)-\overline{G}(x)= \\
&= \frac{\beta\,\zeta(\alpha,x)}{\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x)}-\frac{\beta\,\zeta(\alpha,x+1)}{\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x+1)}= \\
&= \frac{\beta\,\zeta(\alpha,x)\left[\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x+1)\right]-\beta\,\zeta(\alpha,x+1)\left[\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x)\right]}{\left[\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x)\right]\left[\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x+1)\right]}= \\
&= \frac{\beta\,\zeta(\alpha)\left[\zeta(\alpha,x)-\zeta(\alpha,x+1)\right]}{\left[\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x)\right]\left[\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x+1)\right]}= \\
&= \frac{\beta\,\zeta(\alpha)\,x^{-\alpha}}{\left[\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x)\right]\left[\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x+1)\right]}\,. \qquad (2.3)
\end{aligned}
$$

Cal observar que (2.3) també admet la següent expressió:

$$
P(Y=x)=\beta\,P(X=x)\,\frac{\zeta^2(\alpha)}{\left[\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x)\right]\left[\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x+1)\right]}\,, \qquad (2.4)
$$

d'on es dedueix que per x suficientment grans, atès que $\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x)\simeq\zeta(\alpha)$, s'obté que

$$
P(Y=x)\simeq\beta\,P(X=x)\qquad (x \text{ grans})\,.
$$

Una nova expressió alternativa per a (2.3) s'obté a continuació. Donat que

$$
\zeta(\alpha)\left[\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x)\right]=\zeta(\alpha)\left[\sum_{k=1}^{+\infty}k^{-\alpha}-(1-\beta)\sum_{k=x}^{+\infty}k^{-\alpha}\right]=
$$

$$
=\zeta(\alpha)\left[\sum_{k=1}^{x-1}k^{-\alpha}+\beta\sum_{k=x}^{+\infty}k^{-\alpha}\right]=\zeta(\alpha)\left[P(X\leq x-1)+\beta\,P(X>x-1)\right], \qquad (2.5)
$$

$$
P(Y=x)=\beta\,P(X=x)\,\frac{1}{\left[P(X\leq x-1)+\beta\,P(X>x-1)\right]\left[P(X\leq x)+\beta\,P(X>x)\right]}\,.
$$

Cal observar que de (2.5) es dedueix que valors de β superiors a la unitat, donen més pes a la cua que el que li dóna una Zipf, mentre que valors de β menors que la unitat augmenten el pes dels primers valors. Aquest resultat s'apreciarà als gràfics que es presenten més endavant.

La Figura 2.1 conté uns quants gràfics de la funció de probabilitat (2.3) per un valor d'α constant i diferents valors de β. Com s'ha dit abans, quan $\beta = 1$ obtenim la variable original. Així que, comparant amb la $Zipf(\alpha = 1.8) = MOEZipf(\alpha = 1.8, \beta = 1)$, podem arribar a la conclusió que per un α fix, si $0 < \beta < 1$, obtenim més probabilitat als valors inicials amb la generalització que amb la distribució inicial. En canvi, si $\beta > 1$, la inclusió del nou paràmetre disminueix la probabilitat als valors inicials i augmenta la probabilitat de la cua.

2.2.2 Comparació de la $MOEZipf(\alpha, \beta)$ i la $Zipf(\alpha)$

Propietat 2.1. *Sigui X una v.a. amb distribució $Zipf(\alpha)$ i Y una v.a. amb distribució $MOEZipf(\alpha, \beta)$. Es té:*

1. *Si $\beta = 1$, $P(Y = x) = P(X = x) \quad \forall\, x \geq 1$.*

2. *Si $\beta > 1$, $P(Y = x) \geq \frac{1}{\beta} P(X = x) \quad \forall\, x \geq 1$.*

3. *Si $0 < \beta < 1$, $P(Y = x) \geq \beta\, P(X = x) \quad \forall\, x \geq 1$.*

Demostració. Estudiem les tres situacions enumerades.

1. Quan $\beta = 1$, les probabilitats (2.3) corresponen a les d'una distribució Zipf amb paràmetre α, és a dir, $P(Y = x) = P(X = x)$.

2. Quan $\beta > 1$ ($\overline{\beta} < 0$), donat que $\zeta(\alpha, x) \leq \zeta(\alpha)$ per a tot $x \geq 1$, tenim que $\overline{\beta}\,\zeta(\alpha, x) \geq \overline{\beta}\,\zeta(\alpha)$. Per tant,

$$\begin{aligned} 0 \leq \zeta(\alpha) - \overline{\beta}\,\zeta(\alpha, x) \leq \beta\,\zeta(\alpha) \quad &\Longleftrightarrow \quad 0 \leq [\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha, x)][\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha, x+1)] \leq \beta^2\, [\zeta(\alpha)]^2 \\ &\Longleftrightarrow \quad \frac{x^{-\alpha}\,\beta\,\zeta(\alpha)}{[\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha, x)][\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha, x+1)]} \geq \frac{x^{-\alpha}}{\zeta(\alpha)} \frac{1}{\beta} \\ &\Longleftrightarrow \quad P(Y = x) \geq P(X = x) \frac{1}{\beta}. \end{aligned}$$

3. Quan $\beta < 1$ ($\overline{\beta} > 0$), donat que $\zeta(\alpha, x) \leq \zeta(\alpha)$ per a tot $x \geq 1$, tenim que $\overline{\beta}\,\zeta(\alpha, x) \leq \overline{\beta}\,\zeta(\alpha)$. Per tant, per tot valor enter positiu de x, $0 \leq \zeta(\alpha) - \overline{\beta}\,\zeta(\alpha, x) \leq \zeta(\alpha)$, cosa que

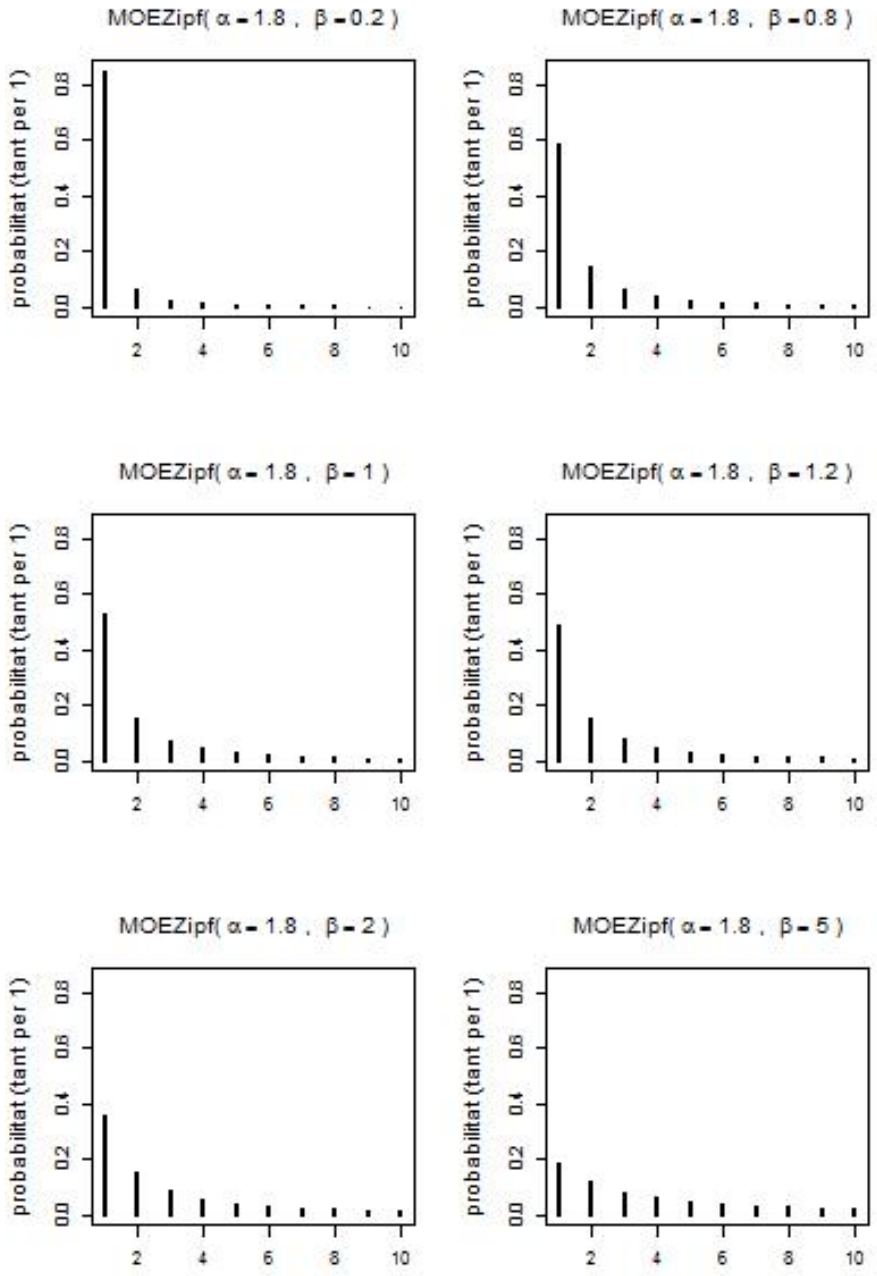

Figura 2.1: Funció de probabilitat de distribucions MOEZipf amb paràmetres $\alpha = 1.8$ i $\beta = 0.2, 0.8, 1, 1.2, 2$ i 5.

ens porta a:

$$\begin{aligned}\zeta^2(\alpha) &\geq [\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x)][\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x+1)]\\ &\iff \frac{x^{-\alpha}\beta}{\zeta(\alpha)} \leq \frac{x^{-\alpha}\beta\,\zeta(\alpha)}{[\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x)][\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x+1)]}\\ &\iff \beta P(X=x) \leq P(Y=x).\end{aligned}$$

□

Cal destacar que la introducció del paràmetre β implica una modificació important de les probabilitats dels primers valors de la variable respecte una distribució Zipf. En canvi, pels valors elevats de x, les probabilitats simplement es multipliquen per una constant, tal com veurem a la propietat que s'enuncia a continuació:

Propietat 2.2. *El paràmetre β es pot interpretar com la raó de probabilitats d'una $MOEZipf(\alpha,\beta)$ i una $Zipf(\alpha)$ per valors grans de x.*

Demostració. De (2.4) es dedueix que

$$\lim_{x\to+\infty}\frac{P(Y=x)}{P(X=x)} = \beta \lim_{x\to+\infty}\frac{\zeta^2(\alpha)}{\left[\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x)\right]\left[\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x+1)\right]} = \beta\,.$$

□

Tal com hem fet amb la distribució Zipf, en aquest cas també és interessant veure què s'obté si escrivim la funció de probabilitat en escala logarítmica. Així doncs, prenent logaritmes a (2.3) tenim que:

$$\ln P(Y=x) = \ln\frac{\beta\,\zeta(\alpha)\,x^{-\alpha}}{\left[\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x)\right]\left[\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x+1)\right]} =$$

$$= \ln\beta - \alpha\ln x + \ln\zeta(\alpha) - \ln\left[\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x)\right] - \ln\left[\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x+1)\right]\,. \qquad (2.6)$$

Propietat 2.3. *Independentment dels valors d'α i β, per a x gran, és a dir, a la cua de la distribució, el logaritme de la probabilitat és lineal com a funció del logaritme de x. Aquest resultat també és cert en un suport més ampli de la distribució si α és suficientment gran.*

Demostració. Per a x suficientment gran, $\zeta(\alpha,x)$ i $\zeta(\alpha,x+1)$ són propers a zero, amb la qual

cosa de (2.6) s'obté que:

$$\ln P(Y = x) \simeq \ln \frac{\beta}{\zeta(\alpha)} - \alpha \ln x, \tag{2.7}$$

per tant, a la cua de la distribució la $MOEZipf(\alpha, \beta)$ es comportarà com una recta de pendent $-\alpha$ i terme independent $\ln(\beta/\zeta(\alpha))$.
Observi's que per α grans, de l'expressió de la funció zeta de Hurwitz (1.2) s'obté que fins i tot per x petits, $x \geq 2$, $\zeta(\alpha, x)$ i $\zeta(\alpha, x + 1)$ són propers a zero, de manera que el logaritme de la probabilitat serà lineal en el logaritme de x gairebé des del principi. □

A la Figura 2.2 hem dibuixat en eixos logarítmics les probabilitats d'una distribució MOEZipf, amb α fix i diferents valors de β i la recta (2.7). Podem observar que quan $\beta = 1$, el logaritme de la probabilitat és una línia recta, ja que ens trobem en el cas particular de la distribució Zipf. Però que quan β s'allunya de l'1, els valors inicials discrepen força amb els de la recta, essent la discrepància més o menys acusada en funció de la proximitat de β a la unitat. Concretament, i coincidint amb el que s'ha vist a la Figura 2.1, es poden distingir dos casos. Quan $0 < \beta < 1$, els valors se situen per sobre de la recta (2.7), és a dir, la recta proposada sotaestima el valor real del logaritme de la probabilitat pels primers valors. En canvi, quan $\beta > 1$, els valors apareixen per sota, per tant, la recta sobreestima la probabilitat als valors inicials. A més, observem que quan β és gran, totes les probabilitats se situen per sota de la recta, per tant, per a valors elevats de β, s'haurà d'anar a uns valors de x molt grans per tal d'observar la linealitat.

A continuació veurem que l'ajust de la recta també depèn en certa manera del valor d'α. A la Figura 2.3, s'ha dibuixat la distribució MOEZipf per a diferents valors d'α i β, juntament amb la recta (2.7). De manera que podem veure quin és l'efecte dels dos paràmetres sobre la distribució. S'observa que:

- Quan $\beta = 1$, com calia esperar, l'ajust a la recta és perfecte en tots els casos, ja que estem en el cas particular de la distribució Zipf.

- Quan $0 < \beta < 1$, i per a tots els valors d'α, el logaritme de les probabilitats se situa per sobre de la recta i s'hi va acostant a mesura que el logaritme de x creix, per tant, la transformació MO concentra més probabilitat als valors inicials que la que esperem si

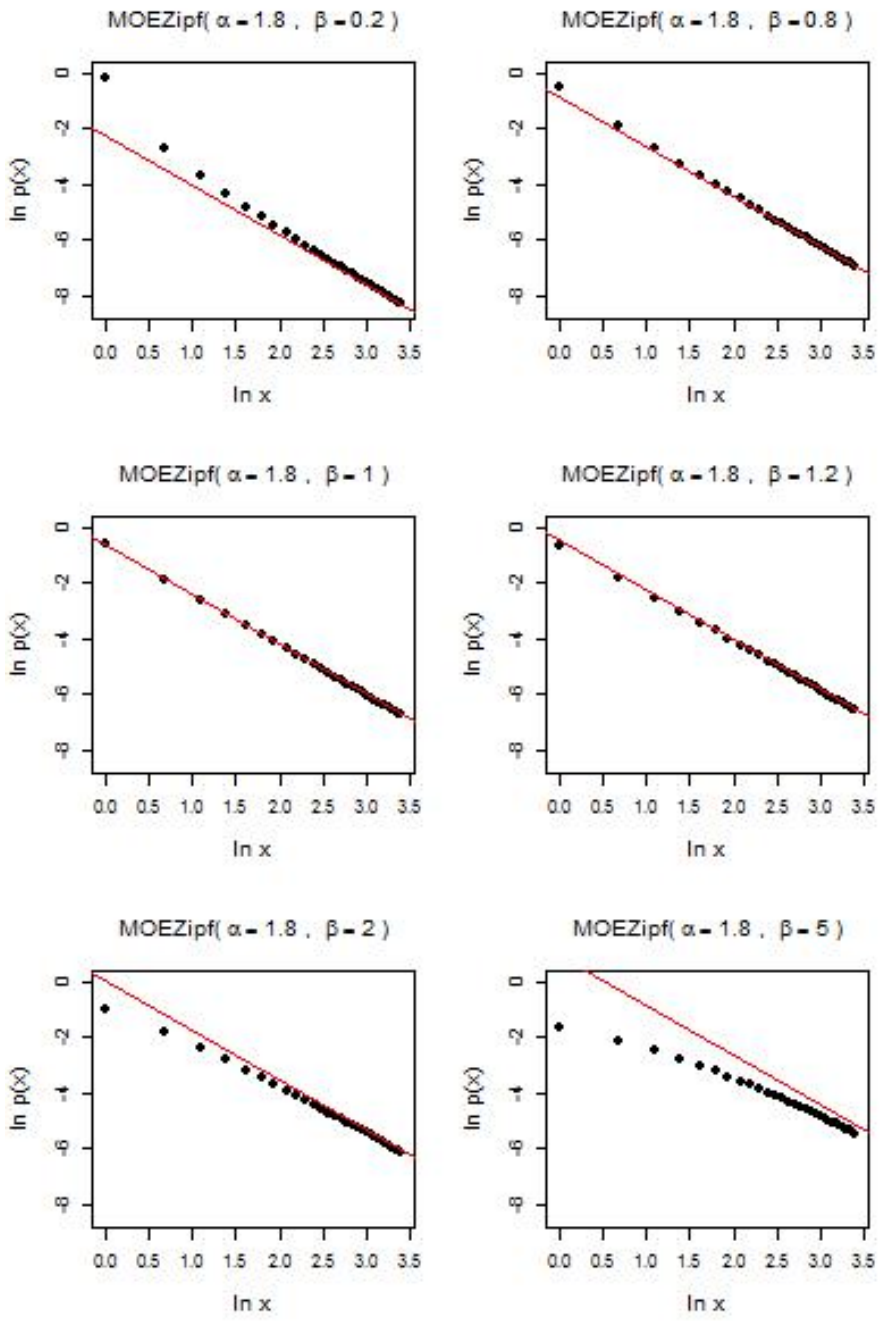

Figura 2.2: En negre, la funció de probabilitat de distribucions MOEZipf amb paràmetres $\alpha = 1.8$ i $\beta = 0.2, 0.8, 1, 1.2, 2$ i 5 representada en eixos logarítmics. I en vermell, la recta $\ln P(Y = x) = \ln \frac{\beta}{\zeta(\alpha)} - \alpha \ln x$ que li correspon.

utilitzem la recta (2.7).

- Per contra, quan $\beta > 1$, el logaritme de les probabilitats se situa per sota de la recta i s'hi va acostant, és a dir, la transformació concentra menys probabilitat als valors inicials que la donada per la recta (2.7).

- Quan α és petita l'ajust és molt dolent per totes les $\beta \neq 1$, concretament aquest ajust va pitjor a mesura que β augmenta. Això és degut al fet que, com que α és petita, els termes $\zeta(\alpha, x)$ i $\zeta(\alpha, x+1)$ s'acosten a zero molt lentament i, per tant, cal buscar valors de x molt grans perquè la recta (2.7) i les probabilitats coincideixin.

- Quan α és gran i independentment del valor de β, la recta proposada és útil per tots els valors excepte per $x = 1$, tal com es dedueix de les gràfiques de la última fila de la figura.

Així doncs, un cop estudiades les dues figures, podem concloure que la recta (2.7) ens serà útil per valors elevats de x en general, per a tots els valors de x excepte la unitat si α és gran, i per a qualsevol x si β és molt propera a la unitat.

També podem calcular la nova *funció de risc* $r(x; \alpha, \beta)$ de la variable a partir de la funció $r(x; \alpha)$ (1.6), de manera que trobem:

$$\begin{aligned} r(x;\alpha,\beta) &= \frac{1}{1-\overline{\beta}\,\overline{F}(x)} r(x;\alpha) = \frac{1}{1-\overline{\beta}\,\frac{\zeta(\alpha,x+1)}{\zeta(\alpha)}} \frac{x^{-\alpha}}{\zeta(\alpha,x)} = \\ &= \frac{\zeta(\alpha)}{\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x+1)} \frac{x^{-\alpha}}{\zeta(\alpha,x)}\,. \end{aligned}$$

La Figura 2.4 compara la funció de risc de la variable original (en negre) i la de la transformació corresponent (en vermell). Podem comprovar que es compleixen les propietats enunciades a l'apartat anterior.

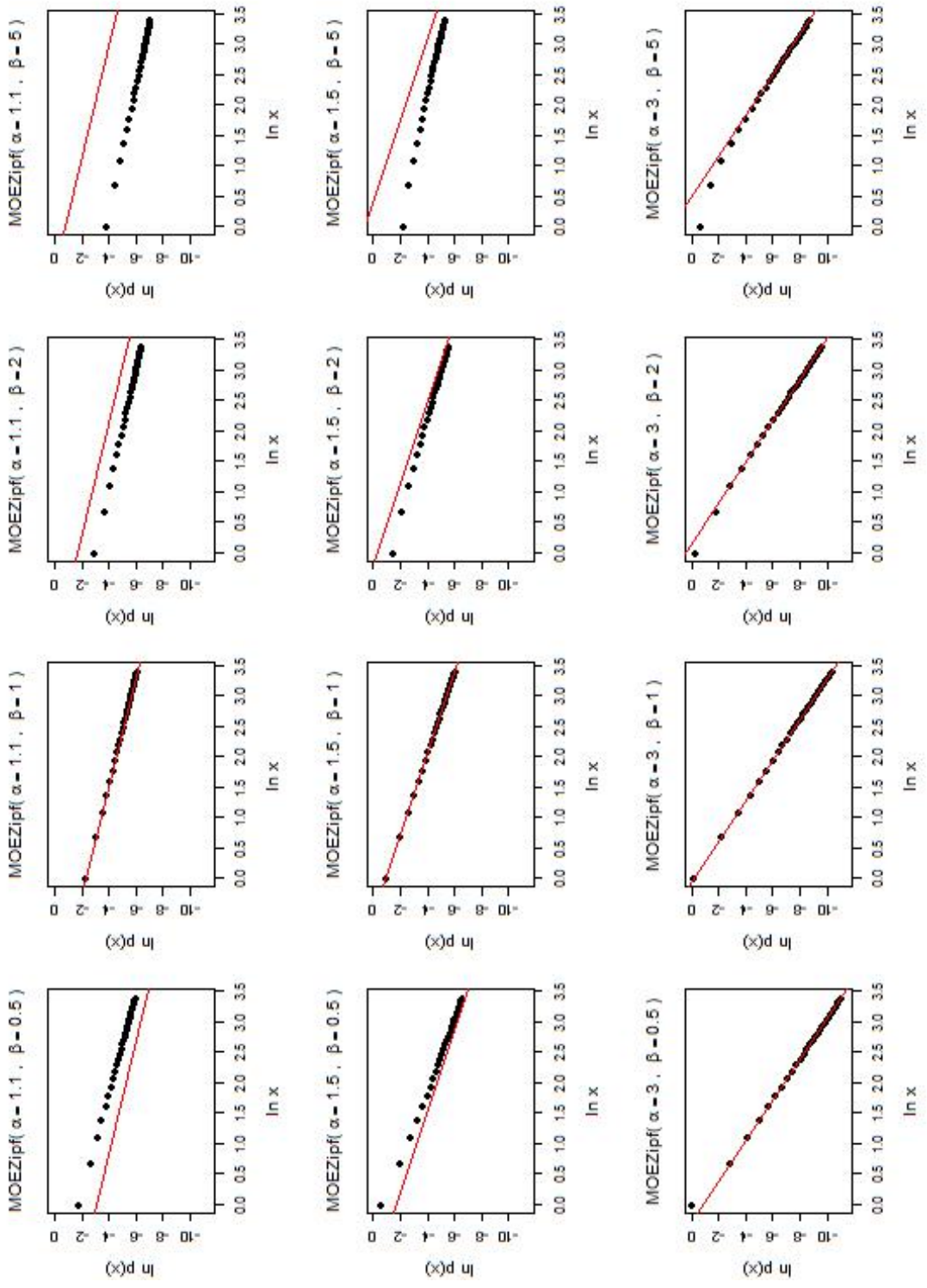

Figura 2.3: En negre, la funció de probabilitat de distribucions MOEZipf amb paràmetres $\alpha = 1.1, 1.5 i 3$ i $\beta = 0.5, 1, 2 i 5$ representada en eixos logarítmics. I en vermell, la recta $\ln P(Y = x) = \ln \frac{\beta}{\zeta(\alpha)} - \alpha \ln x$ que li correspon.

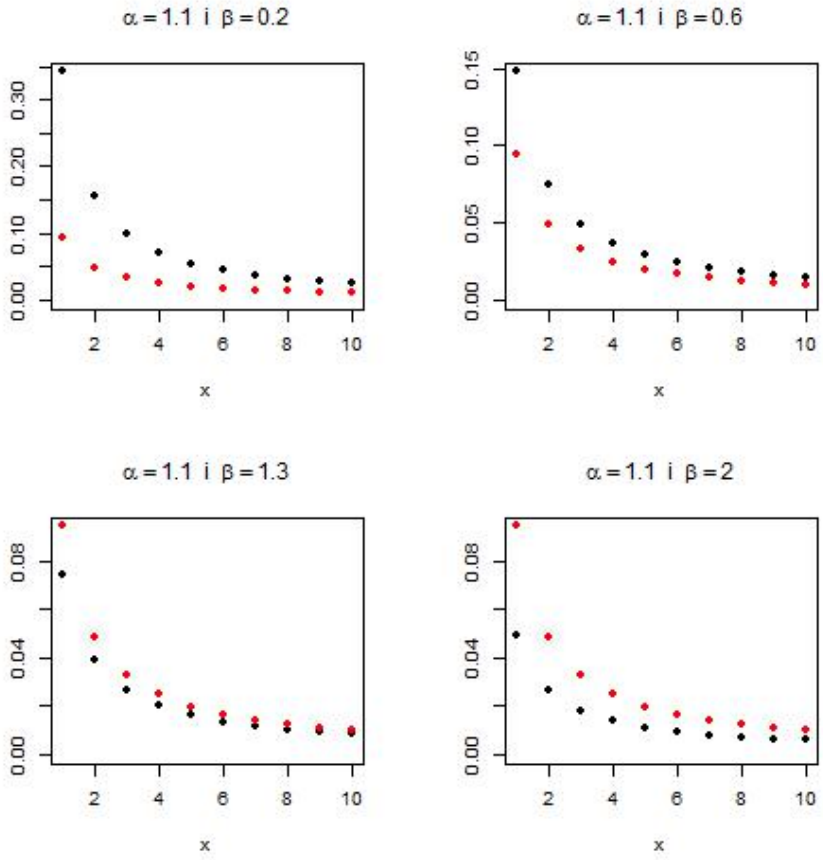

Figura 2.4: Funcions de risc de distribucions Zipf i MOEZipf amb paràmetres $\alpha = 1.1$ i $\beta = 0.2, 0.6, 1.3$i2.

2.2.3 Moments

Si Y és una v.a. tal que $Y \sim MOEZipf(\alpha, \beta)$, la seva esperança es pot escriure com:

$$\begin{aligned} E(Y) &= \sum_{x\geq 1} x\, P(Y=x) = \sum_{x=1}^{\infty} \frac{\beta\, \zeta(\alpha)\, x^{-\alpha+1}}{\left[\zeta(\alpha) - \overline{\beta}\, \zeta(\alpha, x)\right]\left[\zeta(\alpha) - \overline{\beta}\, \zeta(\alpha, x+1)\right]} = \\ &= \frac{1}{\beta\, \zeta(\alpha, 2) + 1} + \beta\, \zeta(\alpha) \sum_{x=2}^{\infty} \frac{x^{-\alpha+1}}{\left[\zeta(\alpha) - \overline{\beta}\, \zeta(\alpha, x)\right]\left[\zeta(\alpha) - \overline{\beta}\, \zeta(\alpha, x+1)\right]}\,. \end{aligned}$$

Proposició 2.1. *L'esperança de Y només és finita per a $\alpha > 2$.*

Demostració. Hem de veure per quins valors d'α convergirà aquesta esperança. Per fer-ho utilitzem el criteri de comparació per pas al límit que podem trobar explicat a la pàgina 231 del llibre *Análisis matemático* [15], que ens diu que donades dues sèries de termes positius $\sum_{n=1}^{\infty} a_n$

i $\sum_{n=1}^{\infty} b_n$, si el límit

$$\lim_{n\to+\infty} \frac{a_n}{b_n} = l\,,$$

és finit i diferent de zero, aleshores, a_n convergeix si i només si b_n convergeix.

En aquest cas, compararem l'esperança d'una v.a. $X \sim Zipf(\alpha)$, que sabem que convergeix per $\alpha > 2$, amb $E(Y)$. $\sum_{n=1}^{+\infty} a_n$ correspondrà a l'esperança de Y mentre que $\sum_{n=1}^{+\infty} b_n$ correspondrà a $E(X)$.

$$E(Y) = \sum_{n=1}^{\infty} a_n = \sum_{n=1}^{\infty} \frac{\beta\,\zeta(\alpha)\,n^{-\alpha+1}}{\left[\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha,n)\right]\left[\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha,n+1)\right]}\,,\quad E(X) = \sum_{n=1}^{\infty} b_n = \sum_{n=1}^{\infty} \frac{n^{-\alpha+1}}{\zeta(\alpha)}\,.$$

Calculant el límit s'obté que:

$$\lim_{n\to+\infty} \frac{a_n}{b_n} = \lim_{n\to+\infty} \frac{\beta\,\zeta(\alpha)\,n^{-\alpha+1}}{\left[\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha,n)\right]\left[\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha,n+1)\right]} : \frac{n^{-\alpha+1}}{\zeta(\alpha)} =$$

$$= \lim_{n\to+\infty} \frac{\beta\,\left[\zeta(\alpha)\right]^2}{\left[\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha,n)\right]\left[\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha,n+1)\right]} = \beta\,.$$

Com que per definició $0 < \beta < \infty$, el límit és finit i diferent de zero. I, per tant, $\sum_{n=1}^{\infty} a_n$ només convergirà quan $\sum_{n=1}^{\infty} b_n$ convergeixi, és a dir, l'esperança de Y serà finita només per $\alpha > 2$. □

Tornant a l'expressió de l'esperança, si tenim en compte que $\zeta(\alpha,x) - \zeta(\alpha,x+1) = \frac{1}{x^{\alpha}}$, quan α sigui gran, atès que:

$$\left[\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha,x)\right]\left[\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha,x+1)\right] \approx \left[\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha,x)\right]^2\,, \tag{2.8}$$

s'obté que:

$$E(Y) \simeq \frac{1}{\beta\,\zeta(\alpha,2)+1} + \beta\,\zeta(\alpha) \sum_{x=2}^{\infty} \frac{x^{-\alpha+1}}{\left[\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha,x)\right]^2}\,.$$

A la Figura 2.5 podem veure el comportament que té l'esperança quan és finita, com a funció dels dos paràmetres de la distribució MOEZipf. Al gràfic de l'esquerra s'observa que $E(Y)$ és decreixent com a funció d'α i, independentment del valor del paràmetre β, tendeix a 1 a mesura que α creix. De fet, si comparem amb la distribució Zipf (cas $\beta = 1$), la transformació MO només modifica el pendent de l'esperança, ja que la forma és la mateixa en tots els casos.

Concretament, quan $0 < \beta < 1$ ($\beta > 1$) l'esperança és superior (inferior) a la que correspondria a una Zipf amb el mateix α. Pel que fa al gràfic de la dreta, hi trobem que $E(Y)$ és creixent com a funció de β. Tot i així, el valor del paràmetre α té una gran influència en l'esperança i es pot tornar a veure que quan α creix, l'esperança tendeix a 1.

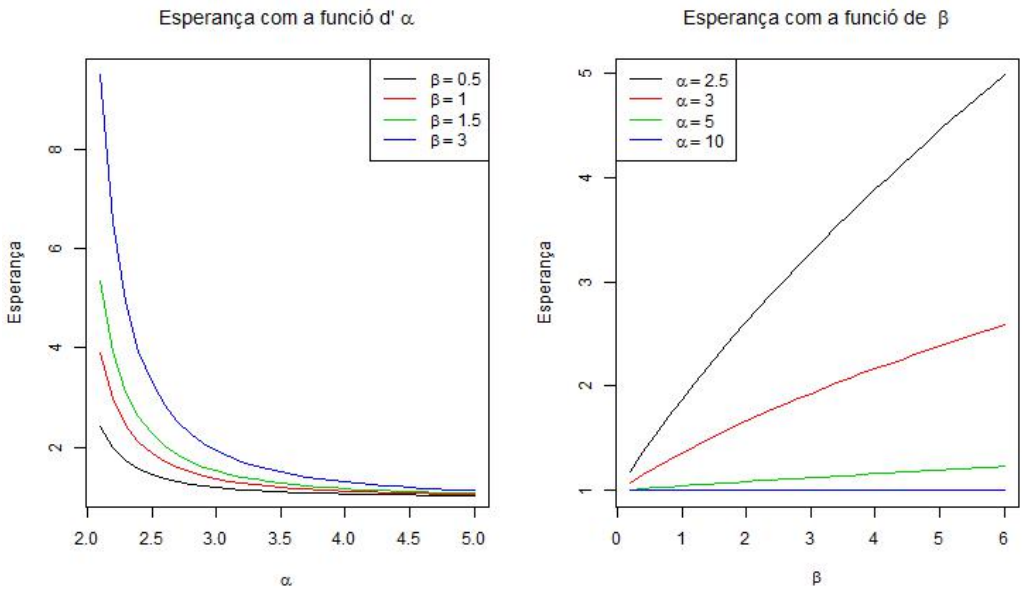

Figura 2.5: Evolució de l'esperança de distribucions MOEZipf en funció d'α (esquerra) amb $\beta = 0.5, 1, 1.5$ i 1, i com a funció de β amb $\alpha = 2.5, 3, 5$ i 10 (dreta).

Pel que fa a la variància,

$$var(Y) = E(Y^2) - E(Y)^2 = \sum_{x=1}^{\infty} \frac{\beta\,\zeta(\alpha)\,x^{-\alpha+2}}{\left[\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha,x)\right]\left[\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha,x+1)\right]} - \left[\sum_{x=1}^{\infty} \frac{\beta\,\zeta(\alpha)\,x^{-\alpha+1}}{\left[\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha,x)\right]\left[\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha,x+1)\right]}\right]^2$$

Proposició 2.2. *La variància d'una $MOEZipf(\alpha,\beta)$ és finita només si $\alpha > 3$.*

Demostració. De forma anàloga a com s'ha fet amb el moment d'ordre 1, es demostra comparant amb el moment d'ordre 2 d'una $Zipf(\alpha)$, que la $MOEZipf(\alpha,\beta)$ té moment d'ordre 2 finit només quan $\alpha > 3$. Atès que la variància és finita només quan el moment d'ordre 2 és finit, aquesta existirà quan $\alpha > 3$. □

Quan α és gran, també podem aplicar l'aproximació (2.8) a la variància, així que obtenim:

$$var(Y) \simeq \sum_{x=1}^{\infty} \frac{\beta\,\zeta(\alpha)\,x^{-\alpha+2}}{\left[\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha,x)\right]^2} - \left[\sum_{x=1}^{\infty} \frac{\beta\,\zeta(\alpha)\,x^{-\alpha+1}}{\left[\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha,x)\right]^2}\right]^2 .$$

També s'ha dibuixat la variància, per casos en que és finita, per veure com evoluciona en funció dels dos paràmetres de la distribució MOEZipf (veure Figura 2.6). Al gràfic de l'esquerra s'observa un dibuix molt similar al de l'esperança com a funció d'α. De fet, les diferències que trobem són que està definida per $\alpha > 3$, i que enlloc de tendir a 1, la variància tendeix a 0 quan α creix. Que la variància tendeixi a zero equival a dir que la variable tendeix a la degenerada en un punt. Tal com es veurà en la secció següent, per α tendint a infinit la variable tendeix a la degenerada en l'1. S'aprecia que el paràmetre β provoca el mateix efecte a la variància que el que hem descrit abans amb l'esperança, és a dir, manté la forma de la variància d'una Zipf modificant-ne només el pendent. Al gràfic de la dreta veiem que $var(Y)$ és creixent en funció de β, però, com abans, el fet que α sigui gran influeix molt en el valor de la variància i fa que tendeixi a zero independentment del valor de β.

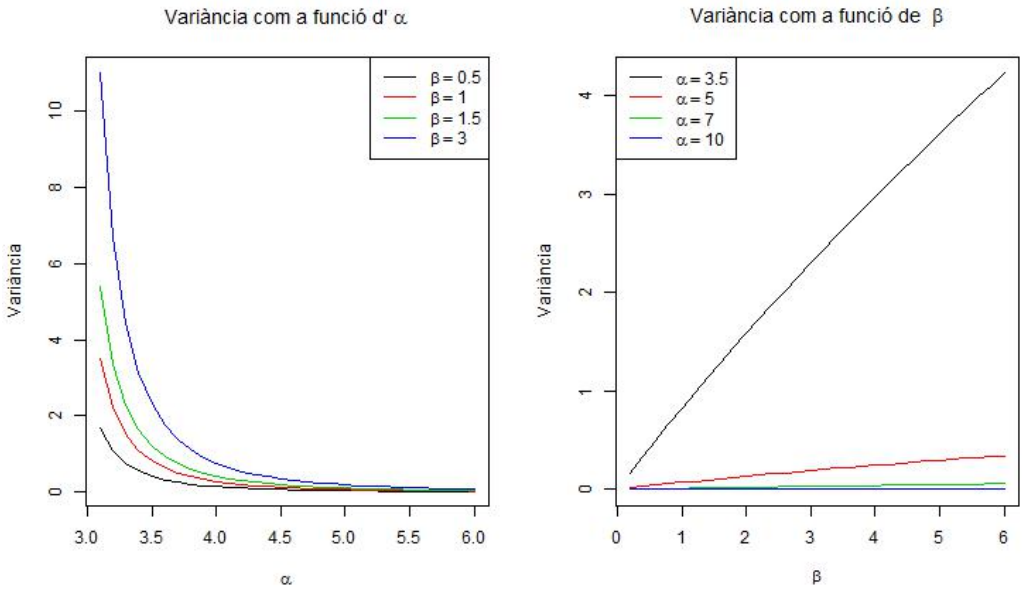

Figura 2.6: Evolució de la variància de distribucions MOEZipf en funció d'α (esquerra) amb $\beta = 0.5, 1, 1.5$ i 1, i com a funció de β amb $\alpha = 3.5, 5, 7$ i 10 (dreta).

2.2.4 Probabilitat de l'1 i distribució límit

Tal com hem vist a la Figura 1.4, per petit que sigui el valor d'α, la distribució Zipf concentra la major part de la probabilitat en el primer valor, l'1. És per això que en aquest apartat estudiem com afecta la transformació aplicada en el comportament de la probabilitat en aquest valor. És a dir, estudiem com evoluciona, com a funció de β, la probabilitat de l'1 d'una v.a. Y que es distribueix segons una MOEZipf. Utilitzant (2.3) i sabent que $\zeta(\alpha,1)=\zeta(\alpha)$ i $\zeta(\alpha,2)=\zeta(\alpha)-1$ trobem que:

$$\begin{aligned} P(Y=1) &= \frac{\beta\,\zeta(\alpha)\,1^{-\alpha}}{\left[\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,1)\right]\left[\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,1+1)\right]} = \frac{\beta\,\zeta(\alpha)}{\zeta(\alpha)\left(1-\overline{\beta}\right)\left[\zeta(\alpha)-\overline{\beta}\,(\zeta(\alpha)-1)\right]} \\ &= \frac{1}{\zeta(\alpha)-(1-\beta)\,(\zeta(\alpha)-1)} = \frac{1}{\zeta(\alpha)-\zeta(\alpha)+1+\beta\,\zeta(\alpha)-\beta} = \frac{1}{\beta\,(\zeta(\alpha)-1)+1} \\ &= \frac{1}{\zeta(\alpha,2)\,\beta+1}\,. \end{aligned} \tag{2.9}$$

La probabilitat de l'1 en funció de β té la forma que es pot veure a la Figura 2.7. On també s'indica el valor de la probabilitat quan $\beta=1$, que és la probabilitat a l'1 d'una distribució Zipf, és a dir, $1/\zeta(\alpha)$. Més concretament, si $\beta=1$ i suposant que $X\sim Zipf(\alpha)$,

$$P(Y=1)=\frac{1}{\zeta(\alpha,2)+1}=\frac{1}{\zeta(\alpha)}=P(X=1)\,.$$

Propietat 2.4. *Sigui Y una v.a. amb distribució $MOEZipf(\alpha,\beta)$. S'observa que $P(Y=1)$ és una funció decreixent de β que compleix:*

$$\lim_{\beta\to 0} P(Y=1)=1.$$

Demostració. A (2.9) s'observa que per un valor d'α fix, el denominador és una funció lineal i creixent de β, amb la qual cosa, la probabilitat decreix en funció de β.

Quan $\beta\to 0$, es té que:

$$\lim_{\substack{\alpha\ \text{fix}\\ \beta\to 0}} P(Y=1)=\lim_{\substack{\alpha\ \text{fix}\\ \beta\to 0}} \frac{1}{\zeta(\alpha,2)\,\beta+1}=1\,.$$

Hem vist que si $\beta\to 0$, la probabilitat de la distribució es concentra a l'1. Veiem que la

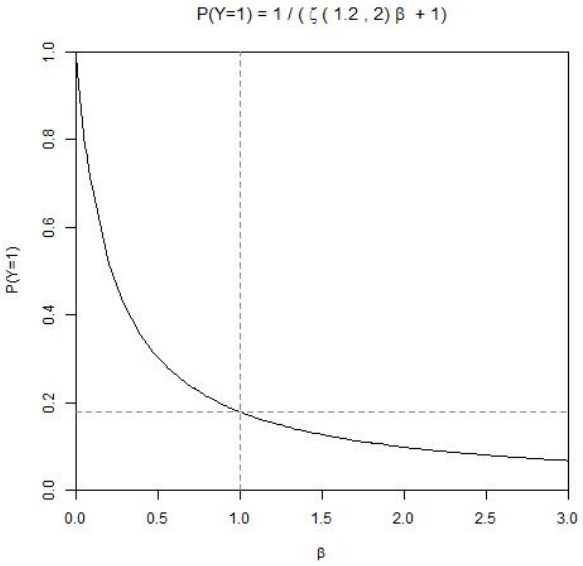

Figura 2.7: Funció $P(Y=1) = 1/(\zeta(\alpha,2)\beta+1)$ per $\alpha = 1.2$. Les línies discontínues senyalen el valor de la probabilitat a l'1 quan $\beta = 1$, $P(Y=1) = P(X=1) = 1/\zeta(\alpha)$.

probabilitat de tots els valors diferents d'1 tendeix a zero.

$$\begin{aligned}\forall\, x>1\,,\ \lim_{\substack{\alpha \text{ fix}\\ \beta\to 0}} P(Y=x) &= \lim_{\substack{\alpha \text{ fix}\\ \beta\to 0}} \frac{\beta\,\zeta(\alpha)\,x^{-\alpha}}{\left[\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x)\right]\left[\zeta(\alpha)-\overline{\beta}\,\zeta(\alpha,x+1)\right]} = \\ &= \frac{0}{\left[\zeta(\alpha)-\zeta(\alpha,x)\right]\left[\zeta(\alpha)-\zeta(\alpha,x+1)\right]} = 0\,.\end{aligned}$$

Per tant, quan $\beta \to 0$, obtenim una disribució que s'anomena degenerada a l'1, ja que hi concentra tota la probabilitat. La podem escriure com:

$$\delta_1 = \begin{cases} 1 & \text{si } x = 1 \\ 0 & \text{si } x > 1 \end{cases}$$

□

Propietat 2.5. *Sigui Y_n, $n \geq 1$ una successió de variables aleatòries tal que Y_n es distribueix segons una $MOEZipf(\alpha, \beta_n)$, essent $(\beta_n)_n$ una successió amb límit zero quan n tendeix a infinit. Llavors es compleix que*

$$Y_n \xrightarrow{n\to+\infty}_{\mathcal{L}} X,$$

on X és una v.a. que segueix una distribució degenerada en la unitat, i $\mathcal{L}$ indica convergència en llei, també anomenada convergència en distribució.

Demostració. La convergència en llei equival a la convergència de les funcions de distribució en tots els punts de continuïtat de la funció de distribució límit. Així doncs, tenint en compte (2.2) podem afirmar que per a tot x enter $x \geq 1$,

$$\lim_{n\to+\infty} \overline{G}(x;\alpha,\beta_n) = \lim_{n\to\infty} \beta_n \frac{\zeta(\alpha, x+1)}{\zeta(\alpha) - \overline{\beta_n}\zeta(\alpha, x+1)} = \frac{0\,\zeta(\alpha, x+1)}{\zeta(\alpha) - \zeta(\alpha, x+1)} = 0.$$

Si $F_n(x)$ és la funció de distribució de Y_n, atès que $F_n(x) = 1 - \overline{G}_n(x)$, llavors es té que:

$$\lim_{n\to+\infty} F_n(x) = 1 \ \forall x \geq 1,$$

i, per tant, les funcions de distribució convergeixen a la funció de distribució de la distribució degenerada en l'1 ja que aquesta és constant igual a la unitat.

□

Propietat 2.6. *Sigui Y una v.a. amb distribució $MOEZipf(\alpha,\beta)$, $P(Y=1)$ és una funció decreixent de β que compleix:*

$$\lim_{\beta\to+\infty} P(Y=1) = 0.$$

Demostració. Quan $\beta \to +\infty$, es té que:

$$\lim_{\substack{\alpha \text{ fix} \\ \beta\to+\infty}} P(Y=1) = \lim_{\substack{\alpha \text{ fix} \\ \beta\to+\infty}} \frac{1}{\zeta(\alpha,2)\,\beta + 1} = 0\,. \tag{2.10}$$

□

Per veure el comportament de la distribució MOEZipf quan $\beta \to +\infty$, es presenta la Figura 2.8. En aquesta figura hi trobem els gràfics en escala logarítimca de tres distribucions MOEZipf amb el mateix α i amb tres β molt grans. Així, observem que, per un α fix i a mesura que augmenta β, la recta del gràfic té tendència a ser més plana, és a dir, en disminueix el pendent. Això ens indica que va augmentant la probabilitat a la cua. Encara que també veiem que la probabilitat dels primers valors no deixa de ser mai superior a la resta. Per tant podem tenir una idea del que passa quan $\beta \to +\infty$. Sembla ser que, en el límit, s'obté una mena de distribució uniforme "infinita", ja que cada cop és més plana i va d'1 a infinit.

Veiem ara que, quan α tendeix a $+\infty$, la distribució també tendeix a la degenerada en la unitat.

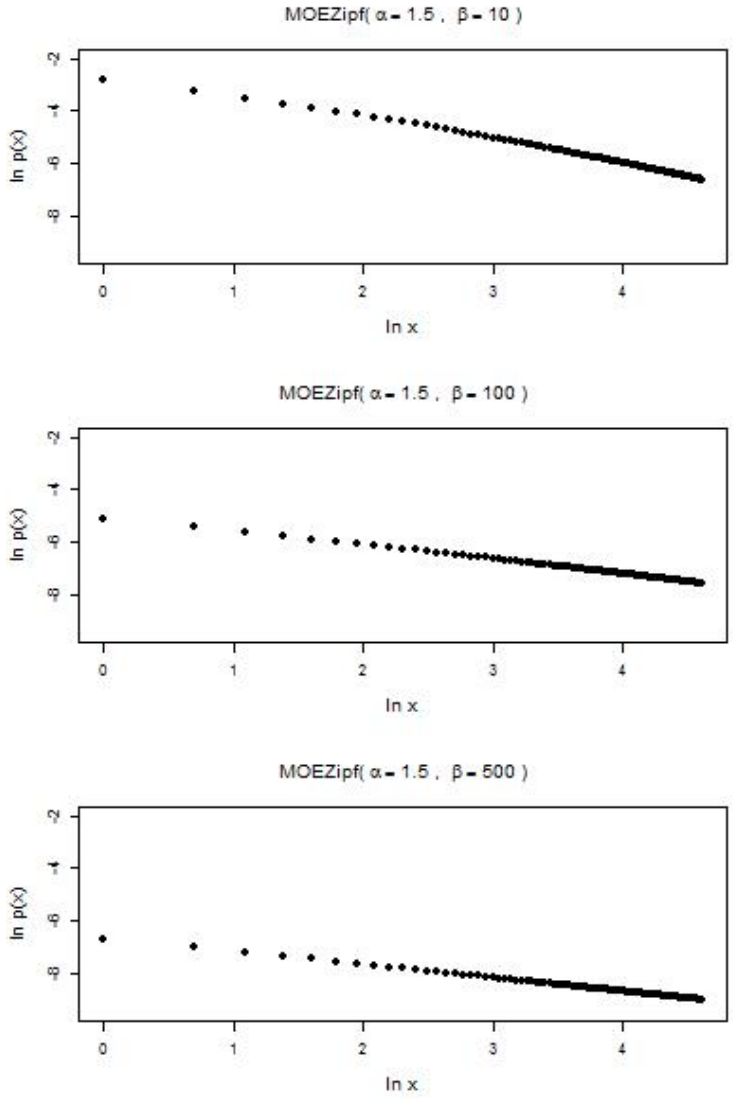

Figura 2.8: Funció de probabilitat de distribucions MOEZipf amb paràmetres $\alpha = 1.5$ i $\beta = 10, 100$ i 500 representada en eixos logarítmics.

Propietat 2.7. *Sigui Y_n, $n \geq 1$ una successió de variables aleatòries tal que Y_n es distribueix segons una $MOEZipf(\alpha_n, \beta)$, essent $(\alpha_n)_n$ una successió amb límit infinit. Llavors es compleix que*

$$Y_n \xrightarrow{n\to+\infty}_{\mathcal{L}} X,$$

on X és una v.a. que segueix una distribució degenerada en la unitat, i $\mathcal{L}$ indica convergència en llei, també anomenada convergència en distribució.

Demostració. Per a β fix i donat que $\zeta(\alpha_n, x+1)$ tendeix a zero quan α tendeix a infinit, i $\zeta(\alpha_n)$

tendeix a 1, tenim que

$$\lim_{n\to+\infty} \overline{G}(x;\alpha_n,\beta) = \lim_{n\to\infty} \beta \frac{\zeta(\alpha_n, x+1)}{\zeta(\alpha_n) - \overline{\beta}\zeta(\alpha_n, x+1)} = \lim_{n\to\infty} \beta \frac{0}{1} = 0.$$

Si $F_n(x)$ és la funció de distribució de Y_n, atès que $F_n(x) = 1 - \overline{G}_n(x)$, llavors es té que:

$$\lim_{n\to+\infty} F_n(x) = 1 \ \forall x \geq 1,$$

i, per tant, les funcions de distribució convergeixen a la funció de distribució de la distribució degenerada en l'1 ja que aquesta és constant igual a la unitat. □

2.2.5 Raó de dues probabilitats consecutives

En aquest apartat estudiem com evoluciona el quocient de probabilitats $\frac{p_{x+1}}{p_x}$ de la distribució MOEZipf per a diferents valors d'α i β. Si $p_x = P(Y = x)$, es pot escriure el quocient que volem estudiar com:

$$\frac{p_{x+1}}{p_x} = \frac{\beta\,\zeta(\alpha)\,(x+1)^{-\alpha}}{\left[\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha, x+1)\right]\left[\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha, x+2)\right]} : \frac{\beta\,\zeta(\alpha)\,x^{-\alpha}}{\left[\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha, x)\right]\left[\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha, x+1)\right]} =$$

$$= \left(\frac{x}{x+1}\right)^{\alpha} \frac{\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha, x)}{\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha, x+2)} \tag{2.11}$$

Quan $\beta = 1$ aquest quocient s'ha de correspondre amb el de la Zipf, que és:

$$\frac{P(X = x+1)}{P(X = x)} = \frac{(x+1)^{-\alpha}}{\zeta(\alpha)} : \frac{x^{-\alpha}}{\zeta(\alpha)} = \left(\frac{x}{x+1}\right)^{\alpha} \tag{2.12}$$

Aquest fet porta a la següent interpretació: les observacions que apareixen més vegades apareixeran aproximadament el doble de vegades que les que apareixen en el rang dos (en la segona posició), i el triple de vegades que les que apareixen en el rang tres (tercera posició) i així successivament. Això porta a afirmar que la freqüència observada d'un determinat element és inversament proporcional a la posició que ocupa en la taula de freqüències. Aquest resultat seria exacte si en prengués un suport finit de la distribució i el paràmetre d'escala fos igual a la unitat. Per comprovar-ho n'hi ha prou veient que de (2.12), suposant suport finit i $\alpha = 1$, es

dedueix que:

$$\frac{P(X=1)}{P(X=2)} = 2 \ , \frac{P(X=1)}{P(X=3)} = \frac{P(X=1)}{P(X=2)}\frac{P(X=2)}{P(X=3)} = 2 \cdot \frac{3}{2} = 3, \ etc.$$

d'on s'obté que

$$P(X=x) = \frac{P(X=1)}{x},$$

Aquesta proporció entre les probabilitats apareix amb elevada freqüència en la vida real, per exemple quan s'analitzen freqüències de paraules, o els rànquings de les ciutats en funció de la seva població, o els rànquings de determinats tipus de grups corporatius, etc.

A la Figura 2.9 presentem la forma que pren el quocient de les probabilitats consecutives (2.11) de les distribucions MOEZipf de paràmetres $\alpha = 1.1$, a l'esquerra, i $\alpha = 2.5$, a la dreta, i $\beta = 0.5, 1, 1.5$ i 2.5. S'observa que:

- En tots els casos, el valor que pren el quocient es troba entre zero i u, cosa que ens indica que cada probabilitat és més petita que la probabilitat immediatament anterior.
- p_{x+1}/p_x és sempre creixent, per tant, a mesura que x creix, la diferència entre una probabilitat i l'anterior va disminuint, de manera que aquest quocient tendeix a 1.
- A mesura que el paràmetre α creix, el quocient de les probabilitats pren valors més baixos. També veiem que, quan α és gran, les diferències del principi són més grans, però, alhora, l'apropament al quocient de la distribució Zipf és més ràpid.
- Pel que fa al paràmetre β, observem diferències entre els quocients de les quatre distribucions quan x és petita, però a mida que ens acostem a la cua, aquests acaben coincidint amb el quocient corresponent a la distribució Zipf (cas $\beta = 1$).
- Si $\beta < 1$, el quocient de les probabilitats és sempre inferior o igual al corresponent a la distribució Zipf original ($\beta = 1$). I, si $\beta > 1$, el quocient és més gran o igual que el de la Zipf que li correspon.

Aquest darrer resultat juntament amb que el límit dels quocients tendeix al quocient de la distribució Zipf consitueix la següent propietat que demostrarem.

Propietat 2.8. *El quocient de probabilitats consecutives compleix les següents propietats:*

1) $\forall\ \beta$, *si* $\beta > 1$, *llavors*

$$\frac{P(Y = x+1)}{P(Y = x)} > \frac{P(X = x+1)}{P(X = x)},$$

i si $0 < \beta < 1$ *la desigualtat és la contrària.*

2) $\forall\, \beta > 0$,

$$\lim_{x \to +\infty} \frac{P(Y = x+1)}{P(Y = x)} = \frac{P(X = x+1)}{P(X = x)}.$$

Demostració. Per demostrar el primer apartat n'hi ha prou observant que, atès que per a tot $\alpha > 1$ i per a tot $x \geq 1$, $\xi(\alpha, x+2) < \xi(\alpha, x)$, podem afirmar que si $0 < \beta < 1$,

$$1 > \frac{\xi(\alpha) - (1-\beta)\,\xi(\alpha, x)}{\xi(\alpha) - (1-\beta)\,\xi(\alpha, x+2)},$$

amb la qual cosa els quocients associats a la $MOEZipf(\alpha, \beta)$ seran tots superiors als associats a la $Zipf(\alpha)$. En el cas que el paràmetre β sigui superior a la unitat, tindrem exactament la desigualtat contrària.

Per tal de demostrar el segon apartat de la propitat, n'hi ha prou amb prendre límit a (2.11) i observar que els dos termes del producte tendeixen a la unitat quan x tendeix a infinit. □

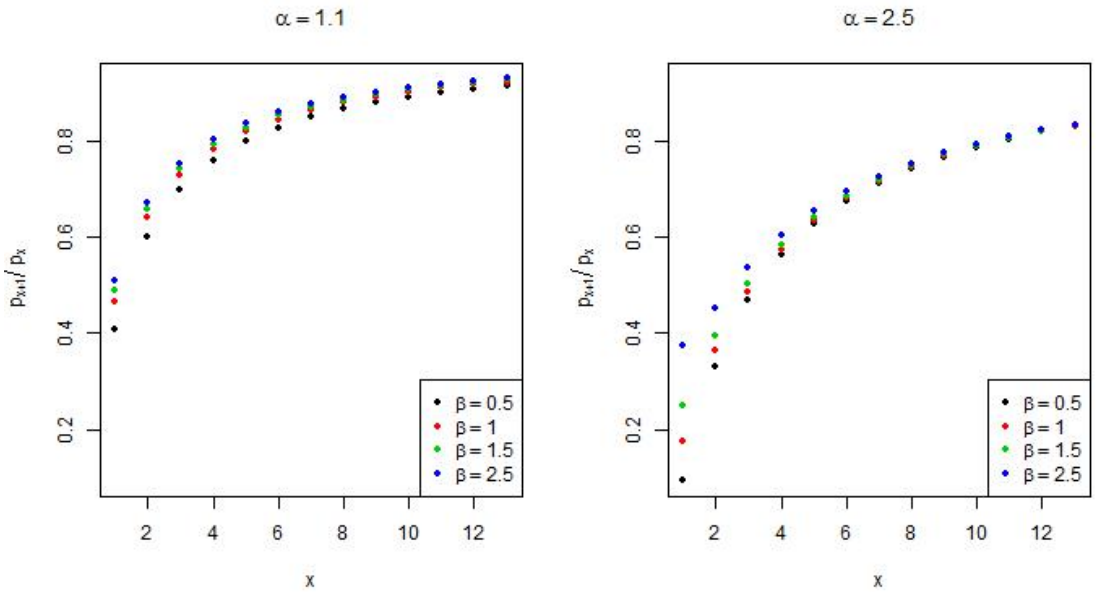

Figura 2.9: Evolució del quocient p_{x+1}/p_x de distribucions MOEZipf amb paràmetres $\alpha = 1.1$, a l'esquerra, i $\alpha = 2.5$, a la dreta, i $\beta = 0.5, 1, 1.5$ i 2.5, en negre, vermell, verd i blau, respectivament.

2.2.6 Estimació de paràmetres

En aquest apartat es plantegen dos mètodes per a l'estimació dels dos paràmetres de la distribució presentada. Suposant que disposem d'una m.a.s. $x_1, x_2, \ldots, x_N$ d'una v.a. Y amb distribució $MOEZipf(\alpha, \beta)$, els mètodes que es proposen són:

- **Primer mètode d'estimació**. Aquest mètode es basa en un cert paral·lelisme existent entre la distribució MOEZipf i les distribucions zero-modificades que s'expliquen tot seguit.

  Existeix una transformació que s'utilitza per v.a. discretes no negatives i serveix per modificar la probabilitat al primer valor (que normalment és el zero). Suposem que Z és una v.a. discreta qualsevol amb suport els enters no negatius, aleshores la *distribució zero-modificada* (Zm) d'aquesta variable té funció de probabilitat:

$$P(Zm = z) = \begin{cases} w + (1-w)\,P(Z=0) & \text{si } z = 0\,. \\ (1-w)\,P(Z=z) & \text{si } z > 0\,; \end{cases}$$

  on el paràmetre w és un pes que pot prendre valors entre $-\frac{P(Z=0)}{1-P(Z=0)}$ i 1. Es pot veure que quan $w = 0$, s'obté la distribució inicial. Quan $-\frac{P(Z=0)}{1-P(Z=0)} \leq w < 0$, disminueix la probabilitat al zero respecte la distribució inicial. I, en canvi, quan $0 < w \leq 1$, augmenta la probabilitat al zero. Per $w = 1$ obtenim la distribució degenerada al zero. I quan $w = -\frac{P(Z=0)}{1-P(Z=0)}$, el resultat és la distribució inicial truncada al zero.

  Atès que l'objectiu de modificar una distribució de probabilitat en zero és el d'adaptar la probabilitat teòrica en zero a l'empírica afegint un paràmetre, un dels mètodes usats per estimar els dos paràmetres de la distribució en el cas que Z tingui distribució uniparamètrica, consisteix en igualar l'esperança a la mitjana aritmètica i la probabilitat en zero empírica a la teòrica. Se'n pot trobar més informació al capítol 8 del llibre *Univariate Discrete Distributions* [9].

  Per tant, té sentit definir un mètode d'estimació dels paràmetres de la MOEZipf semblant al que acabem d'explicar. Pel nostre cas, com que s'han d'estimar dos paràmetres, α i β, també necessitem dues equacions. Té sentit prendre la primera que forci que la probabilitat a l'1 sigui la proporció observada d'aquest valor. I, la segona, que l'esperança sigui igual

a la mitjana mostral. Així doncs, les estimacions d'α i β s'obtindran solventant el sistema d'equacions següent:

$$\begin{cases} \frac{1}{\zeta(\alpha,2)\,\beta+1} = \frac{f_1}{N}\,, \\ E(Y) = \overline{Y}\,; \end{cases}$$

on f_1 és la freqüència observada a l'1. De la primera equació podem aïllar β i obtenim:

$$\beta = \frac{N - f_1}{f_1\,\zeta(\alpha,2)} \iff \overline{\beta} = 1 - \beta = \frac{f_1\,\zeta(\alpha) - N}{f_1\,\zeta(\alpha,2)}\,.$$

Substituint a la segona equació l'expressió de β, el sistema d'equacions que hem de resoldre queda reduït a la següent equació, que només depèn d'α i que caldrà resoldre numèricament:

$$\frac{N - f_1}{f_1\,\zeta(\alpha,2)}\,\zeta(\alpha) \sum_{x=1}^{\infty} \frac{x^{-\alpha+1}}{\left[\zeta(\alpha) - \frac{f_1\,\zeta(\alpha)-N}{f_1\,\zeta(\alpha,2)}\,\zeta(\alpha,x)\right]\left[\zeta(\alpha) - \frac{f_1\,\zeta(\alpha)-N}{f_1\,\zeta(\alpha,2)}\,\zeta(\alpha,x+1)\right]} = \overline{Y}\,.$$

A les estimacions obtingudes mitjançant aquest mètode les notarem per $\tilde{\alpha}$ i $\tilde{\beta}$. S'ha utilitzat R per a programar una rutina que permet resoldre l'equació anterior. S'ha fet servir la funció `optimize()` que permet optimitzar funcions d'un únic paràmetre i que utilitza el mètode de Brent per fer-ho, el codi es pot trobar a l'apartat d'estimació de paràmetres de l'Apèndix B.

- **Màxima versemblança**. La funció de versemblança corresponent a una mostra $x_1, x_2, \ldots, x_N$ de $Y \sim MOEZipf(\alpha,\beta)$ és igual a:

$$\mathcal{L}(\alpha,\beta;x_1,x_2,\ldots,x_N) = \prod_{i=1}^{N} \frac{\beta\,\zeta(\alpha)\,x_i^{-\alpha}}{\left[\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha,x_i)\right]\left[\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha,x_i+1)\right]} =$$

$$= \beta^N\,\zeta^N(\alpha)\left[\prod_{i=1}^{N} x_i\right]^{-\alpha} \prod_{i=1}^{N} \frac{1}{\left[\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha,x_i)\right]\left[\zeta(\alpha) - \overline{\beta}\,\zeta(\alpha,x_i+1)\right]}\,.$$

El nostre objectiu és trobar els valors d'α i β, que denotarem per $\hat{\alpha}$ i $\hat{\beta}$, que facin màxima aquesta funció. Els trobarem maximitzant el logaritme de la versemblança,

$l(\alpha, \beta; x_1, \ldots, x_N)$, que és més senzill de maximitzar i té la forma següent:

$$l(\alpha, \beta; x_1, \ldots, x_N) = N \ln \beta + N \ln \zeta(\alpha) - \alpha \sum_{i=1}^{N} \ln x_i \\ - \sum_{i=1}^{N} \ln \left[\zeta(\alpha) - \overline{\beta}\, \zeta(\alpha, x_i)\right] - \sum_{i=1}^{N} \ln \left[\zeta(\alpha) - \overline{\beta}\, \zeta(\alpha, x_i + 1)\right]$$

Per aquest cas, també s'ha utilitzat R per a programar una rutina que ens permeti optimitzar l'equació anterior. Com que es tracta d'una funció de dos paràmetres, s'ha utilitzat la funció `optim()`. El mètode que s'ha utilitzat permet fixar cotes pels paràmetres i es basa en un mètode quasi-Newton, que s'anomena BFGS.

2.2.7 Significació del paràmetre β

Per estudiar la bondat de l'ajust de la distribució MOEZipf a unes dades concretes, aplicarem els mètodes explicats a la subsecció 1.1.5 del primer capítol, que són el test χ^2 de Pearson i l'AIC. Però, evidentment, també ens interessarà avaluar si la distribució generalitzada ens permet obtenir estimacions significativament més bones que les obtingudes amb la distribució original. A tal efecte utilitzarem un test clàssic que permet comparar models aniuats i que és el **test de raó de versemblances** (*Likelihood Ratio Test*, LRT). En el nostre cas compararem l'ajust de la distribució Zipf amb l'ajust a la distribució MOEZipf, és a dir, farem el contrast següent:

$$\begin{array}{ll} H_0 : X \sim \text{distribució Zipf} & H_0 : \beta = 1 \\ \iff & \\ H_1 : X \sim \text{distribució MOEZipf} & H_1 : \beta \neq 1 \end{array}$$

Donada una m.a.s. $x_1, x_2, \ldots, x_N$ d'una v.a. X i un cop trobades $\hat{\alpha}_1$, estimació màxim versemblant sota el model Zipf, i $(\hat{\alpha}_2, \hat{\beta}_2)$, estimació màxim versemblant sota el model MOEZipf, es defineix l'estadístic de raó de versemblances com (2.13),

$$\lambda = -2 \left[\ln \mathcal{L}(\hat{\alpha}_1; x_1, x_2, \ldots, x_N) - \ln \mathcal{L}(\hat{\alpha}_2, \hat{\beta}_2; x_1, x_2, \ldots, x_N)\right], \tag{2.13}$$

on $\mathcal{L}(\hat{\alpha}_1; x_1, x_2, \ldots, x_N)$ és la versemblança màxima obtinguda mitjançant la distribució Zipf, i

$\mathcal{L}(\hat{\alpha}_2, \hat{\beta}_2; x_1, x_2, \ldots, x_N)$, la versemblança màxima associada a la distribució que s'ha proposat en aquest capítol, la MOEZipf. Si la hipòtesi nul·la és certa, aquest estadístic es distribueix segons una distribució χ^2_1.

Capítol 3

Anàlisi de dades reals

3.1 Introducció

En aquest capítol es presenten i analitzen quatre conjunts de dades reals d'àrees molt diverses. S'apliquen els mètodes explicats als capítols anteriors per tal d'ajustar les distribucions Zipf i MOEZipf, amb l'objectiu de veure si la transformació proposada en aquest treball permet obtenir ajustos significativament millors que els obtinguts mitjançant la distribució original.

Per a ajustar la distribució Zipf hem estimat el paràmetre α utilitzant el mètode de màxima versemblança. Els paràmetres de la distribució MOEZipf s'han estimat de dues maneres diferents. El primer mètode consisteix en igualar la probabilitat empírica i la teòrica de l'1, i la mitjana al valor esperat. El segon mètode és el de màxima versemblança. Els ajustos s'han realitzat amb el paquet estadístic R. El codi es pot trobar a l'Apèndix B.

En tots els casos, per a poder aplicar el test χ^2 de bondat de l'ajust, s'han agrupat les darreres categories de manera que el nombre de valors observats sigui superior o igual que 5. Així, s'intenta garantir que es compleixi una de les condicions del test, que diu que els valors esperats han de ser superiors o iguals que 5 per cada categoria. Podrem comparar els ajustos obtinguts amb les dues distribucions, ja que l'agrupació de categories és la mateixa en ambdós casos. Per a l'aplicació d'aquest test, s'ha fixat el nivell de confiança al 95%. Com s'ha explicat a l'apartat 1.1.5, els p-valors corresponents al test χ^2 es calculen tenint en compte l'aproximació de la distribució χ^2 per la Normal quan els graus de llibertat superen 50.

En capítols anteriors s'ha esmentat un altre test de bondat de l'ajust, el test de Kolmogorov-Smirnov. Com s'ha dit llavors, tota la informació referent a aquest test es troba al Capítol 4.

En tots els casos, per veure si la distribució MOEZipf és significativament més adequada que la distribució Zipf s'ha realitzat el test de raó de versemblances (LRT) que, si la hipòtesi nul·la és certa, es distribueix segons una χ^2_1. Per tant, si fixem el nivell de confiança al 95%, el valor crític amb el qual haurem de comparar aquest estadístic és 3.84.

Tots els exemples estudiats corresponen a conjunts de dades molt grans, el menor té 9101 observacions i el més gran, 225409. El fet que actualment es disposi de grans volums de dades de les quals cal extreure'n informació, fa possible disposar de mostres de grandària molt elevada, cosa que era inviable fa només pocs anys.

3.2 Terrorisme

Si definim la *severitat* dels atacs terroristes com el nombre de morts produïts en un atac, té sentit pensar que dita severitat és una v.a. amb distribució PL discreta. Això és així perquè s'observaran pocs atacs amb moltes víctimes mortals i molts atacs amb poques víctimes. Per tant, si ens limitem als atacs terroristes en els que almenys hi ha hagut una víctima mortal, tindrà sentit ajustar les dades a través de la distribució Zipf.

Disposem del nombre de morts dels atacs terroristes amb almenys una víctima mortal que hi va haver a tot el món des del febrer de 1968 fins el juny de 2006. Són dades procedents de diferents bases de dades que van ser recopilades i estudiades pels autors de l'article [5]. En aquest article, s'explica breument la distribució PL i s'ajusta a tres variables diferents, que són: el *nombre de vícitmes mortals per atac*, el *nombre de ferits per atac* i la *suma d'ambdues*. També s'ulilitzen les dades per estudiar l'evolució del terrorisme al llarg del temps, per veure si el comportament de la serveritat dels atacs terroristes és el mateix segons si es van produir dins o fora de l'OCDE i per comparar els atacs segons el tipus d'arma que es va utilitzar.

Un altre article que analitza aquestes dades és [2], i les facilita al lloc web [16]. En aquest

article presenten el cas continu i el cas discret de la distribució $PL(x_{min}, \alpha)$, expliquen diferents mètodes d'estimació pels paràmetres, la manera d'avaluar la bondat de l'ajust i finalment ho apliquen a 24 conjunts de dades diferents, entre els que hi ha les referents al terrorisme. Per aquestes dades, estimen el paràmetre x_{min} mitjançant un mètode que fa mínima la distància entre la funció de probabilitat estimada i l'empírica. Es tracta del mètode basat en les distàncies que hem explicat al primer capítol d'aquest treball. Concretament, el valor que estimen per x_{min} és 12 ± 4, de manera que la distribució Zipf no s'inclou en aquesta estimació. El paràmetre α s'estima per màxima versemblança. En l'article, els autors també defineixen un indicador de bondat de l'ajust basat en l'estadístic de Kolmogorov-Smirnov. Calculen el p-valor d'aquest estadístic per mitjà de simulacions i, per les dades corresponents al terrorisme, arriben a la conclusió que no es pot rebutjar la distribució $PL(x_{min}, \alpha)$. A més, també utilitzen un test de raó de versemblances per hipòtesis que no estan aniuades, definit per Vuong [17], per comparar l'ajust per la distribució PL amb el que s'obté utilitzant altres distribucions, entre les quals consideren la Poisson, la log-normal o l'exponencial. Tot i que la log-normal i la *stretched exponential* (modificació de l'exponencial que l'estira en els extrems) també són distribucions plausibles, el test de Vuong dóna un suport moderat a la distribució PL.

A la Taula 3.1 hi trobem les dades d'aquest exemple. Hi figuren el nombre d'atacs terroristes observats, o_i, amb mortalitat i, així com els ajustos de les dades per les dues distribucions d'interès. Cal observar que la major part de les observacions es concentra als primers valors, sent 1 el valor mínim de morts en un atemptat. Concretament, trobem que més de la meitat dels 9101 atacs amb víctimes mortals que hi ha hagut han tingut només una víctima. És a dir, més de la meitat de les observacions de la variable es concentren a la unitat. A la taula no podem veure el comportament total de la cua perquè, com en tots els exemples, es troba agrupada. En aquest cas concret, s'han agrupat les observacions més grans o iguals a 28. Les freqüències observades pels atacs amb més de 55 víctimes són totes iguals a 0, a 1 o a 2. I els tres atacs amb més víctimes en van tenir 331, 400 i 2749. Aquest últim atemptat correspon a l'atemptat de l'11 de setembre de 2001 a Nova York. Tal com tothom ja sap, es tracta d'un atemptat excepcional, i per aquest motiu en aquest treball s'ha decidit ajustar les dades de mortalitat en atacs terroristes amb i sense aquest atemptat.

Comencem considerant el conjunt complet de les dades, és a dir, el que inclou l'11 de setembre

Morts		Zipf $\hat{\alpha} = 1.897$		MOEZipf $\hat{\alpha} = 2.160$, $\hat{\beta} = 1.752$	
i	o_i	e_i	d_i	e_i	d_i
1	4802	5192.51	29.369	4777.11	0.130
2	1600	1393.74	30.525	1634.92	0.746
3	750	645.73	16.836	778.22	1.023
4	444	374.10	13.061	445.64	0.006
5	287	244.96	7.213	285.61	0.007
6	190	173.32	1.605	197.34	0.273
7	148	129.37	2.683	143.85	0.120
8	96	100.41	0.194	109.15	1.584
9	85	80.30	0.275	85.44	0.002
10	92	65.75	10.479	68.55	8.018
11	60	54.87	0.479	56.14	0.266
12	64	46.52	6.566	46.75	6.367
13	43	39.97	0.230	39.49	0.312
14	32	34.72	0.214	33.76	0.092
15	39	30.46	2.392	29.18	3.307
16	29	26.95	0.156	25.45	0.496
17	34	24.02	4.143	22.37	6.042
18	17	21.55	0.962	19.81	0.400
19	17	19.45	0.309	17.66	0.025
20	17	17.65	0.024	15.83	0.086
21	18	16.09	0.227	14.27	0.975
22	18	14.73	0.726	12.92	1.995
23	17	13.54	0.886	11.75	2.342
24	9	12.49	0.974	10.73	0.280
25	17	11.56	2.564	9.84	5.218
26	5	10.73	3.058	9.05	1.809
27	8	9.99	0.395	8.34	0.014
≥ 28	163	295.50	59.411	191.83	4.332
	9101				

Taula 3.1: Per cada *nombre de morts en atacs terroristes*, i, trobem la freqüència observada, o_i, les esperades, e_i, i els valors dels termes i-èssims de la suma del X^2, d_i, corresponents a les distribucions Zipf i MOEZipf.

de 2001. A la Taula 3.2 trobem els paràmetres estimats per les dues distribucions analitzades i els indicadors de bondat de l'ajust explicats anteriorment. Pel que fa a les estimacions de la distribució MOEZipf, observem que les del paràmetre α són molt semblants en ambdós mètodes, ja que coincideixen fins a les dècimes. Les estimacions de β ja no s'assemblen tant, atès que només coincideixen en les unitats. En tots dos casos, però, aquest últim paràmetre és més gran que 1, per tant, la distribució objecte d'aquest projecte fa que es concentri menys probabilitat als valors inicials respecte a una distribució Zipf amb el mateix paràmetre α. Pel que fa a la bondat

de l'ajust, veiem que el fet d'afegir el paràmetre β, ens porta a un ajust considerablement millor de les dades, ja que, la log-versemblança creix i, per tant, disminueix l'AIC. També decreix considerablement el valor de l'estadístic X^2. Aquest experimenta una reducció d'un 76,39% al passar de la distribució Zipf a la MOEZipf amb els paràmetres estimats, en ambdós casos, per màxima versemblança. L'estadístic del LRT també ens indica que la distribució MOEZipf proporciona un ajust molt més bo per a les dades que el proporcionat per la distribució Zipf. Aquest estadístic pren el valor 180.63, que és molt superior al valor crític que li correspon.

Distribució	Paràmetre	Estimació	log-versemblança	X^2	p-valor	AIC
Zipf	$\hat{\alpha}$	1.897	-16713.71	195.956	0.000	33429.42
MOEZipf	$\tilde{\alpha}$	2.198	-16625.34	53.983	0.001	33254.68
(1r mèt. est.)	$\tilde{\beta}$	1.823				
MOEZipf	$\hat{\alpha}$	2.160	**-16623.39**	**46.266**	0.006	**33250.79**
(màx. versem.)	$\hat{\beta}$	1.752				

Taula 3.2: Resum de l'ajust de la variable *nombre de morts en atacs terroristes* a distribucions Zipf i MOEZipf. Per cada distribució trobem el valor dels paràmetres estimats, la log-versemblança, el valor de l'estadístic X^2 amb el p-valor corresponent i l'AIC.

Si tornem a la Taula 3.1, podem trobar les freqüències esperades, e_i, per la distribució Zipf i la MOEZipf. Aquesta útlima, amb els paràmetres estimats per màxima versemblança, que és quan s'obtenen millors ajustos. A més, per totes dues distribucions també s'inclou el valor d_i, que correspon al terme i-èssim de l'estadístic X^2. Tant en aquesta taula com en la Figura 3.1, que conté els observats i esperats amb les dues distribucions en escala logarítmica, podem observar que ajustant amb la distribució MOEZipf s'aconsegueix capturar molt millor el comportament dels primers valors de la variable. De fet, pel que fa als set primers valors, l'ajust de la MOEZipf és notablement millor que el de la Zipf. A la part central de la distribució, els ajustos són molt semblants. I a la cua, un cop agrupada, torna a ser notablement millor la MOEZipf que la Zipf. Aquest és un comportament semblant al que s'observarà en els exemples posteriors.

Tal com s'ha explicat abans, l'atemptat amb més víctimes mortals, el que correspon a l'11-S, s'allunya molt de la resta d'observacions, ja que es tracta d'un cas excepcional. És per això que s'han tornat a ajustar les dades sense aquesta observació per tal de veure si es tracta d'una obervació influent. A la Figura 3.2 hi trobem les funcions de probabilitat teòriques que hem ajustat pel mètode de màxima versemblança per la distribució Zipf i la MOEZipf amb el nou

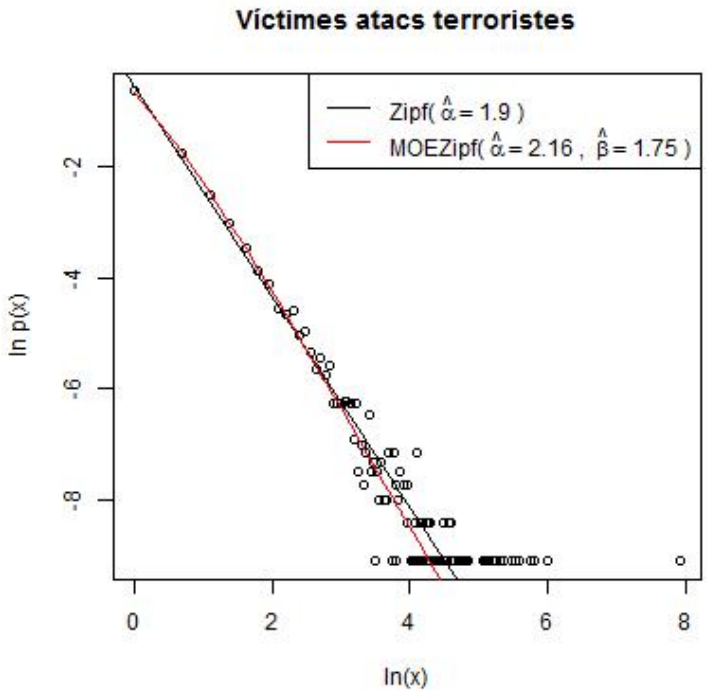

Figura 3.1: Gràfic, en eixos logarítmics, de la funció de probabilitat empírica del nombre d'atacs terroristes en funció del nombre de morts amb l'ajust obtingut per la distribució Zipf (en negre) i la distribució MOEZipf (en vermell).

conjunt de dades. Els paràmetres estimats són molt similars als que hem obtingut anteriorment. També són força semblants els indicadors de bondat de l'ajust. Les taules referents a aquests ajustos, atès que canvien molt poc de les obtingudes anteriorment, es poden trobar a l'Apèndix A.

3.3 Lingüística

Tal com hem explicat al primer capítol d'aquesta memòria, la freqüència d'aparició de les paraules en un text sovint s'ajusta mitjançant una distribució PL discreta. Això es deu al fet que sabem que hi haurà moltes paraules que apareixeran poc i n'hi haurà algunes, sovint poques, que apareixeran molt. A més, atès que la freqüència mínima d'aparició d'una paraula en un text és 1, té sentit considerar la distribució Zipf com a distribució de freqüències de paraules.

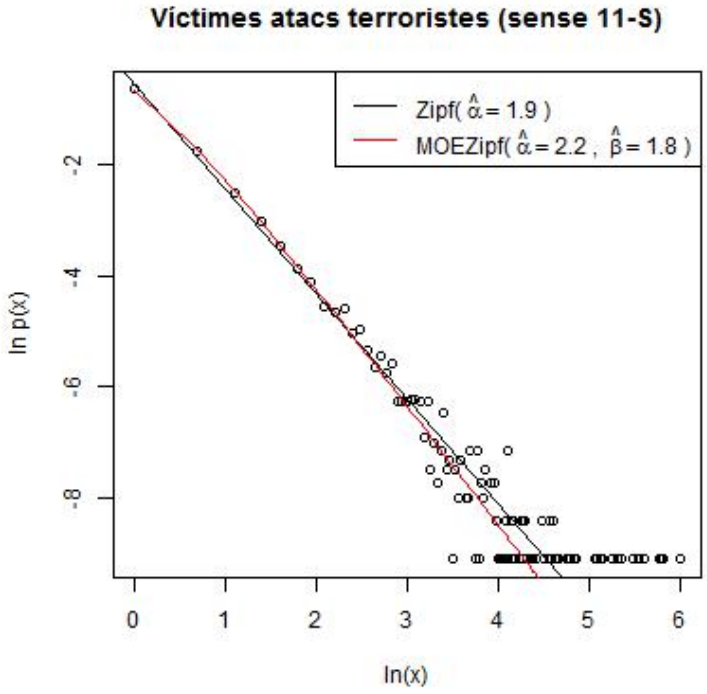

Figura 3.2: Gràfic, en eixos logarítmics, de la funció de probabilitat empírica del *nombre d'atacs terroristes* (sense tenir en compte l'atemptat de l'11-S) en funció del nombre de morts amb l'ajust obtingut per la distribució Zipf (en negre) i la distribució MOEZipf (en vermell).

En el present apartat s'estudien les dades corresponents a la freqüència d'aparició de les paraules de la novel·la *Moby Dick* de l'escriptor Herman Melville. Aquest conjunt de dades el va estudiar G. K. Zipf a finals dels anys 40 (veure [1]). En aquesta referència es va definir la distribució que porta el seu nom i, per tant, és d'especial rellevància. Les dades també s'han estudiat recentment, al 2009 en l'article [2]. Formen part dels 24 conjunts de dades que s'estudien en aquest article, on es proposa ajustar-les a través d'una $PL(x_{min}, \alpha)$, tal com s'ha explicat a la secció anterior. La distribució PL que s'estima, amb $x_{min} = 7 \pm 2$ (no inclou la distribució Zipf), proporciona un bon ajust per les dades i, quan es compara amb l'obtingut mitjançant d'altres distribucions, s'arriba a la conclusió que és la que proporciona el millor ajust. El fet que s'hagin estudiat recentment prova que són unes dades clàssiques però, a la vegada, d'actualitat. Per aquest motiu, s'ha decidit incloure-les també en aquest projecte. El conjunt de dades pot trobar-se a la pàgina web [16] d'aquest segon article. També han treballat aquestes dades els autors de l'article [18], l'objectiu del qual és comparar el comportament dels textos reals i els textos generats aleatòriament. Una de les conclusions de l'article és que els textos aleatoris perden la forma de distribució Zipf quan s'estudien paraules d'una llargada concreta, cosa que

no passa amb els textos reals.

Tal com es pot veure a la Taula 3.3, el conjunt de dades compleix les característiques de la distribució Zipf. Primer, observem que el nombre mínim d'aparicions de les paraules en aquesta novel·la és 1. També es veu que hi ha una gran concentració d'observacions als primers valors, de fet, gairebé tres quartes parts de les observacions es troben en els tres primers valors. Després, trobem molt poques observacions pels valors molt elevats, és a dir, poques paraules apareixen moltes vegades. Aquesta última característica no s'observa directament de la taula, ja que les observacions de la cua s'han agrupat a partir de 53. On sí que s'aprecia, però, és a la Figura 3.3, on trobem que pels valors grans del logaritme de x, el logaritme de la probabilitat empírica pren valors molt petits. Per tant, la probabilitat en aquests valors és molt propera a zero. A més, de la mateixa figura també es desprèn que la forma del logaritme de la probabilitat respecte del logaritme de x s'assembla a una recta. Així que tenim unes dades que s'adiuen amb la descripció de la distribució Zipf, ja que concentren molta probabilitat als primers valors, després tenen una cua llarga on s'hi concentra molt poca probabilitat i el gràfic en escala logarítmica és pràcticament una recta. Cal dir que els tres valors més grans observats en aquest conjunt de dades són: 6260, 6414 i 14086.

A la Taula 3.3 hi podem trobar les freqüències esperades, e_i, corresponents a les dues distribucions amb els paràmetres estimats per màxima versemblança. També hi ha el valor d_i, que correspon al sumand i-èssim de l'estadístic X^2. Podem observar grans diferències entre les dues distribucions sobretot pels primers 7 valors, pels que aconseguim un ajust notablement millor utilitzant la distribució MOEZipf. Pel que fa als valors centrals, no hi ha grans diferències entre els esperats amb les dues distribucions. I, en referència a la cua també aconseguim una millor estimació utilitzant la MOEZipf.

A la Taula 3.4 trobem les estimacions dels paràmetres de les distribucions Zipf i MOEZipf, i els indicadors de bondat de l'ajust que s'han explicat anteriorment. Pel que fa a la distribució Zipf, si apliquem el test χ^2 veiem que el valor de l'estadístic és molt gran i ens porta a un p-valor que és pràcticament zero. En canvi, per la distribució MOEZipf i amb ambdós mètodes d'estimació, podem acceptar que la distribució és apropiada per les dades en base al test χ^2, atès que no tenim evidències per rebutjar la hipòtesi nul·la (p-valor > 0.05). Els valors estimats

Paraules		Zipf $\hat{\alpha} = 1.775$		MOEZipf $\hat{\alpha} = 1.944$, $\hat{\beta} = 1.523$	
i	o_i	e_i	d_i	e_i	d_i
1	9161	9813.85	43.430	9119.97	0.185
2	3085	2867.96	16.425	3179.21	2.792
3	1629	1396.52	38.701	1598.39	0.586
4	926	838.12	9.214	962.11	1.355
5	627	564.04	7.027	643.55	0.426
6	469	408.11	9.084	461.30	0.128
7	361	310.43	8.238	347.22	0.547
8	300	244.93	12.382	271.02	3.098
9	232	198.73	5.571	217.57	0.957
10	179	164.83	1.218	178.61	0.001
11	165	139.18	4.789	149.33	1.645
12	146	119.27	5.993	126.75	2.925
13	115	103.47	1.285	108.96	0.334
14	90	90.72	0.006	94.70	0.234
15	100	80.26	4.853	83.09	3.440
16	74	71.58	0.082	73.51	0.003
17	56	64.28	1.066	65.51	1.380
18	70	58.08	2.449	58.76	2.152
19	51	52.76	0.059	53.00	0.076
20	46	48.17	0.098	48.06	0.089
...	...	...	...	...	...
40	11	14.08	0.673	12.72	0.233
41	10	13.47	0.895	12.13	0.375
42	17	12.91	1.296	11.58	2.535
43	11	12.38	0.154	11.07	0.000
44	8	11.89	1.271	10.59	0.633
45	9	11.42	0.513	10.14	0.128
46	12	10.98	0.094	9.72	0.535
47	7	10.57	1.208	9.32	0.580
48	6	10.19	1.720	8.95	0.974
49	10	9.82	0.003	8.60	0.226
50	8	9.47	0.229	8.28	0.009
51	11	9.15	0.376	7.97	1.156
52	6	8.84	0.911	7.67	0.365
≥ 53	388	588.67	68.408	422.19	2.769
	18855				

Taula 3.3: Per cada *nombre d'aparicions de paraules en un text*, i, trobem la freqüència observada, o_i, les esperades, e_i, i els valors dels termes i-èssims de la suma del X^2, d_i, corresponents a les distribucions Zipf i MOEZipf.

dels paràmetres d'aquesta segona distribució pels dos mètodes són molt semblants. En tots dos casos $\beta > 1$, cosa que fa que la probabilitat que assigna la MOEZipf als valors inicials sigui

més petita que la corresponent a una distribució Zipf amb el mateix α. Si es comparen els dos ajustos corresponents a la MOEZipf, el millor és l'obtingut mitjançant l'estimació màxim versemblant dels paràmetres, atès que és la que aconsegueix la log-versemblança més gran i el X^2 i l'AIC més petits (veure Taula 3.4). De fet, al passar de l'ajust de la distribució Zipf al de la MOEZipf aconseguim una reducció del 79.64% del valor de l'estadístic X^2. L'estadístic del LRT ens permet confirmar que l'ajust de la distribució MOEZipf és molt millor que el de la distribució Zipf, ja que pren el valor $227.16 \gg 3.84$, cosa que ens porta a rebutjar la hipòtesi nul·la que diu que $\beta = 1$.

Distribució	Paràmetre	Estimació	log-versemblança	X^2	p-valor	AIC
Zipf	$\hat{\alpha}$	1.775	-40196.00	272.38	0	80394.00
MOEZipf	$\tilde{\alpha}$	1.908	-40086.28	62.96	0.097	80176.56
(1r mèt. est.)	$\tilde{\beta}$	1.429				
MOEZipf	$\hat{\alpha}$	1.944	**-40082.42**	**55.45**	0.293	**80168.83**
(màx. versem.)	$\hat{\beta}$	1.523				

Taula 3.4: Resum de l'ajust de la variable *nombre d'aparicions de les paraules* del llibre *Moby Dick* mitjançant les distribucions Zipf i MOEZipf. Per cada distribució trobem el valor dels paràmetres estimats, la log-versemblança, el valor de l'estadístic X^2 amb el p-valor corresponent i l'AIC.

A la Figura 3.3 també podem veure el comportament de les dues distribucions estudiades: l'ajust més precís de la distribució MOEZipf als primers valors, les poques diferències als valors centrals i certes diferències als valors elevats, que no hem vist anteriorment perquè havíem agrupat la cua. És important veure que una β igual a 1.5 provoca una corbatura al primer tram de la recta que la fa capaç d'adaptar-se molt millor als primers valors de la distribució. Com que aquest paràmetre és superior a la unitat, la corba és lleugerament cóncava.

3.4 Correu electrònic

Ara que les xarxes socials estan tan de moda, podem pensar les relacions entre individus com un graf (Figura 3.4) en el que els nodes corresponen als individus i les arestes serveixen per a connectar-los entre sí. Pensat així, a cada node se li associa un valor que correspon al seu ordre, és a dir, al nombre de connexions que té. La taula de freqüències dels ordres dels nodes té sentit ajustar-la amb la distribució Zipf, si considerem només els nodes amb ordre més gran o igual a

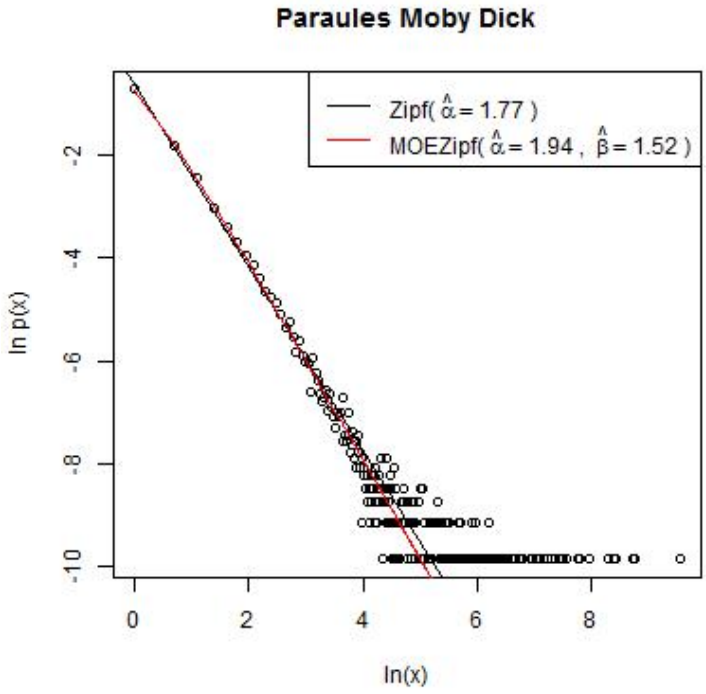

Figura 3.3: Gràfic, en eixos logarítmics, de la funció de probabilitat empírica de la freqüència d'aparició de les paraules del llibre *Moby Dick* amb l'ajust obtingut per la distribució Zipf (en negre) i la distribució MOEZipf (en vermell).

1. En aquesta secció s'ajusta un conjunt de dades que es podria representar mitjançant un graf on els nodes són adreces de correu electrònic, i les arestes serveixen per connectar les adreces entre les quals s'ha enviat almenys un e-mail en un període de temps fixat.

En altres paraules, la variable que estudiarem en aquest apartat és la *nombre de relacions establertes per correu electrònic*. Entenent com a relació el fet que dues adreces electròniques s'enviïn almenys un correu en un període de temps fixat. Podem suposar que hi haurà moltes adreces amb poques relacions i algunes amb moltes.

El conjunt de dades que utilitzarem es pot trobar a l'enllaç [19]. Són dades sobre la xarxa de correus electrònics d'una institució de recerca europea gran, recollides entre el mes d'octubre de 2003 i el mes de maig de 2005. Aquest conjunt de dades destaca pel seu gran volum, atès que té una grandària de 225409 dades.

Aquestes dades apareixen a la tesi [20], on s'estudien xarxes grans. Més específicament, en la tesi desenvolupen models que permeten explicar els processos que governen l'evolució d'aquestes xarxes, ajusten aquests models a xarxes reals i els utilitzen per generar grafs realistes i/o donar

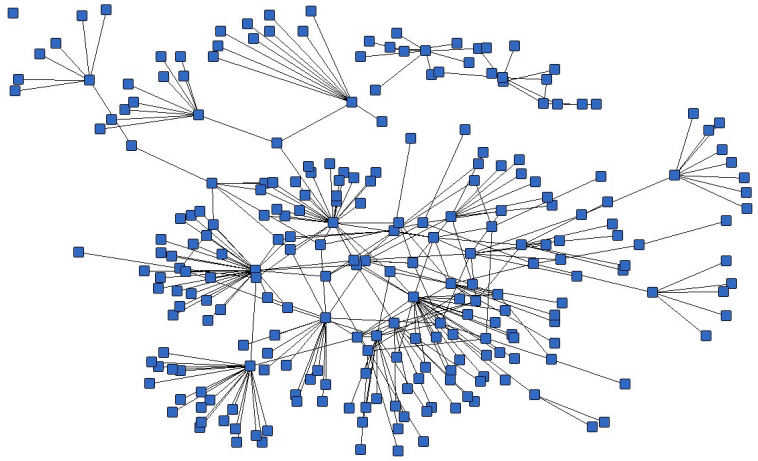

Figura 3.4: Exemple de graf (Font: www.wonderbabiesco.org/SNA.php). Podem pensar que els quadrats blaus corresponen a adreces de correu electrònic, i que les arestes que els uneixen indiquen que ambdues adreces han intercanviat informació durant un període de temps determinat.

explicacions formals sobre les seves propietats. És en aquest punt que utilitzen una relació que anomenen *Densification Power Law* (DPL), que considera que el nombre d'arestes d'un graf en un instant t és proporcional al nombre de nodes al mateix instant t elevat a un exponent, a. És a dir:

$$A_t = k \cdot N_t^a \ ,$$

on A_t i N_t són, respectivament, el nombre d'arestes i nodes a l'instant t. Els autors també ajusten la distribució PL del grau dels nodes, i analitzen quina relació té amb la DPL. També estudien l'existència de patrons "locals" i estructures de propagació en les xarxes, per veure'n la influència que tenen en la informació o en la propagació de virus a la xarxa. Un altre article que també estudia aquest tipus de dades és [21], on els autors demostren que les xarxes formades per persones que intercanvien correus electrònics compleixen tan les propietats de xarxes de "món petit"[1] com les de xarxes lliures d'escala[2].

Com es pot veure, la Taula 3.5 conté les dades observades i ajustades per la distribució Zipf i la MOEZipf estimant els paràmetres per màxima versemblança. El nombre mínim de relacions per correu electrònic és 1, cosa que, juntament amb el fet que hi ha moltes adreces que tenen poques relacions i algunes que en tenen moltes, ens permet suggerir inicialment la distribució Zipf com

[1] En anglès, *small-world network*. És una xarxa on la distància entre dos nodes escollits aleatòriament és proporcional al logaritme del nombre de nodes de la xarxa (veure www.wikipedia.org).

[2] En anglès, *scale-free network*. És una xarxa, la distribució del grau dels nodes de la qual és la PL (veure www.wikipedia.org).

a distribució apropiada per aquest conjunt de dades. Cal destacar que les adreces que només tenen una relació representen el 85% de les 225409 adreces observades. La freqüència observada, o_i, decreix ràpidament pels primers valors de i i, després, el decreixement es va desaccelerant. A la taula, les observacions que corresponen a valors elevats de la v.a. es troben agrupades. Concretament, s'han agrupat les dades a partir del 65. Ens hem de fixar en la Figura 3.5 per veure que la cua de la distribució empírica és llarga i concentra poques observacions. En aquesta figura, a l'igual que en els exemples anteriors, trobem que, pels valors elevats del logaritme de x, el logaritme de la probabilitat empírica és molt petit, cosa que ens indica que hi ha només algunes adreces que tenen un nombre molt elevat de relacions. Més específicament, les tres adreces amb més relacions en tenen 854, 871 i 930. A la figura també s'observa que el logaritme de la probabilitat empírica fa una mica de corba al principi i després tendeix a ser lineal. Aquesta forma s'adiu amb la obtinguda mitjançant la distribució MOEZipf amb un paràmetre β inferior a la unitat, que és exactament el que s'obtindrà.

A la Taula 3.6 hi trobem les estimacions i els indicadors de bondat de l'ajust obtinguts, mitjançant els mètodes plantejats en capítols anteriors, per les dues distribucions que estudiem. La distribució estimada en el cas de la distribució Zipf té el paràmetre $\alpha > 2$, cosa que ens indica que té esperança finita, resultat que no es donava en cap dels dos exemples anteriors. Pel que fa a la distribució MOEZipf, els paràmetres estimats per ambdós mètodes són bastant propers, especialment pel paràmetre β. Tal com havíem intuït a la vista del gràfic, l'estimació de β és inferior a 1, cosa que implica una menor concentració de probabilitat als primers valors de la variable respecte una distribució Zipf amb el mateix α. Pel que fa a la bondat de l'ajust, el p-valor que correspon a l'estadístic X^2 és pràcticament zero en tots tres casos. De manera que, segons aquest estadístic, no podem acceptar cap de les tres disribucions com a adequada per les dades. Tot i així, si comparem la bondat dels diferents ajustos, observem que el que ens permet obtenir més bons resultats és l'ajust de la distribució MOEZipf estimant els paràmetres pel mètode de màxima versemblança. Encara que mitjançant el test χ^2 hàgim de rebutjar la distribució MOEZipf com a vàlida, afegint un paràmetre (β) hem arribat a reduir el valor de l'estadístic en un 93.74% respecte la distribució Zipf i hem trobat la versemblança màxima i l'AIC mínim. Comparant les log-versemblances de la distribució Zipf i MOEZipf mitjançant el LRT ($4730.78 \gg 3.84$), també podem arribar a la conclusió que l'ajust a la distribució MOEZipf

Relacions		Zipf $\hat{\alpha} = 2.968$		MOEZipf $\hat{\alpha} = 2.284$, $\hat{\beta} = 0.390$	
i	o_i	e_i	d_i	e_i	d_i
1	192117	186505.931	168.810	192335.439	0.248
2	18763	23841.508	1081.779	17104.718	160.769
3	5461	7157.350	402.049	5895.277	31.991
4	2537	3047.718	85.583	2888.120	42.687
5	1489	1571.728	4.354	1683.396	22.449
6	896	914.941	0.392	1089.660	34.418
7	634	579.051	5.214	756.801	19.926
8	445	389.597	7.879	552.934	21.069
9	355	274.670	23.493	419.726	9.981
10	277	200.918	28.810	328.274	8.009
11	236	151.418	47.247	262.989	2.770
12	171	116.959	24.969	214.884	8.962
13	134	92.230	18.917	178.498	11.093
14	126	74.022	36.500	150.362	3.947
15	102	60.317	28.806	128.194	5.352
16	90	49.803	32.443	110.442	3.784
17	71	41.603	20.773	96.024	6.521
18	87	35.112	76.680	84.167	0.095
19	73	29.907	62.094	74.308	0.023
20	52	25.684	26.964	66.029	2.981
...	...	...	...	...	...
50	5	1.693	6.458	8.064	1.164
51	5	1.597	7.255	7.706	0.950
52	15	1.507	120.796	7.371	7.895
53	10	1.424	51.634	7.057	1.228
54	12	1.347	84.216	6.761	4.060
55	9	1.276	46.754	6.483	0.977
56	8	1.210	38.120	6.221	0.509
57	8	1.148	40.911	5.974	0.687
58	6	1.090	22.119	5.741	0.012
59	7	1.036	34.331	5.521	0.396
60	6	0.986	25.510	5.312	0.089
61	5	0.938	17.578	5.115	0.003
62	9	0.894	73.473	4.928	3.364
63	5	0.853	20.169	4.751	0.013
64	8	0.814	63.453	4.583	2.548
≥ 65	529	26.067	9703.538	225.771	407.261
	225409				

Taula 3.5: Per cada *nombre de relacions per correu electrònic*, i, trobem la freqüència observada, o_i, les esperades, e_i, i els valors dels termes i-èssims de la suma del X^2, d_i, corresponents a les distribucions Zipf i MOEZipf.

és molt millor, i que el paràmetre β és significativament diferent de la unitat.

Distribució	Paràmetre	Estimació	log-versemblança	X^2	p-valor	AIC
Zipf	$\hat{\alpha}$	2.968	-156765.21	13714.84	0	313532.42
MOEZipf	$\tilde{\alpha}$	2.126	-154526.75	968.34	0	309057.51
(1r mèt. est.)	$\tilde{\beta}$	0.321				
MOEZipf	$\hat{\alpha}$	2.284	**-154399.82**	**858.27**	0	**308803.64**
(màx. versem.)	$\hat{\beta}$	0.390				

Taula 3.6: Resum de l'ajust de la variable *nombre de relacions establertes per correu electrònic* a distribucions Zipf i MOEZipf. Per cada distribució trobem el valor dels paràmetres estimats, la log-versemblança, el valor de l'estadístic X^2 amb el p-valor corresponent i l'AIC.

A la Taula 3.5, podem comparar les freqüències observades, o_i, amb les freqüències esperades, e_i, per la distribució Zipf i la MOEZipf, amb els paràmetres estimats per màxima versemblança. A més, per cada freqüència esperada també hi trobem el valor d_i, que correspon a l'element i de l'estadístic X^2. Pel primer valor, l'observat i l'esperat per la MOEZipf són molt propers. De fet, pels quatre primers valors de la variable, l'ajust obtingut mitjançant la distribució MOEZipf és millor que el de la Zipf. Pels quatre valors següents, la situació és just al contrari. I, a partir de 9, l'ajust utilitzant la distribució MOEZipf torna a ser millor. Podem arribar a aquestes mateixes conclusions observant la Figura 3.5, on detectem a simple vista que la distribució MOEZipf captura molt millor el comportament de les dades, tan a l'inici com al final de la distribució.

3.5 Cites

L'últim conjunt de dades que considerem conté informació sobre el nombre de cites d'articles que apareixen en d'altres articles. En concret, la informació que conté indica quins articles se citen en un total de 32158 articles. Aquests pertanyen a un tema concret, la fenomenologia de física de partícules, que és la part teòrica de la física de partícules [3] (o física d'altes energies). Les dades, com les de l'apartat anterior, apareixen a la tesi [20] i es poden trobar a l'enllaç [22]. La informació es basa en els articles publicats a arXiv.org entre del gener de 1993 i l'abril de 2003, i es refereix únicament a les cites entre articles d'aquesta temàtica, de manera que no inclou informació sobre les cites relacionades amb articles que pertanyen a camps diferents del

[3] Física de partícules: Disciplina de la física que s'encarrega de l'estudi de les partícules constituents de la matèria i la radiació i de les interaccions entre elles (veure www.wikipedia.org).

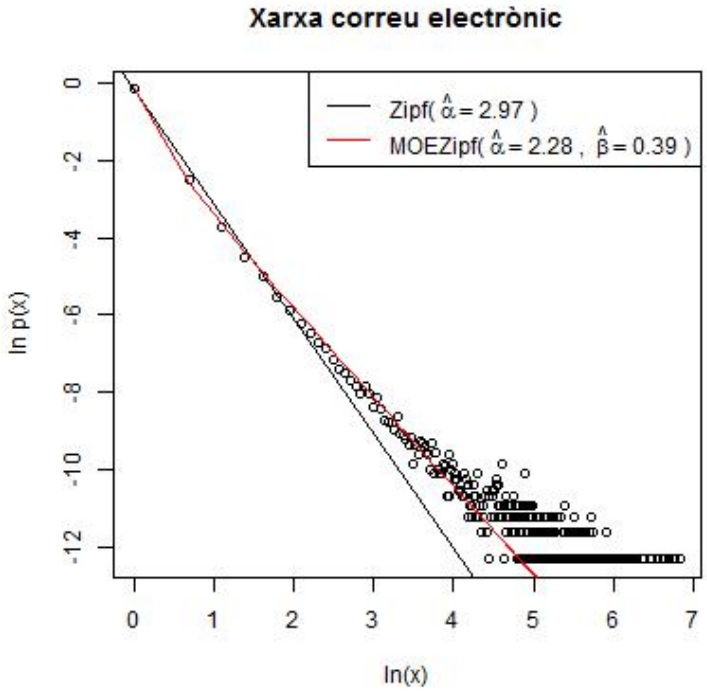

Figura 3.5: Gràfic, en eixos logarítmics, de la funció de probabilitat empírica del *nombre de relacions establertes per correu electrònic* amb l'ajust obtingut per la distribució Zipf (en negre) i la distribució MOEZipf (en vermell).

que és objecte d'estudi.

D'aquest conjunt de dades es poden extreure dues v.a. que iniciament té sentit que es distribucixen segons una PL. Per una banda, tindrem la v.a. *nombre d'articles citats en un article*. Que es comporta com una PL, ja que suposem que hi haurà molts articles que tinguin poques cites i pocs articles que en tinguin moltes. Per l'altra banda, també podrem considerar la v.a. *nombre de vegades que se cita un article*. Que també tindrà un comportament similar al de la PL, ja que sovint es publiquen molts articles que es referencien poc en d'altres articles, i pocs articles que es referencien molt.

El primer autor que va estudiar la xarxa de cites entre articles va ser D. J. de Solla Price [23], que va veure que el nombre de cites noves que obté un article és proporcional al nombre de cites que té. Ho va anomenar el fenomen de l'"avantatge acumulatiu" o bé "del ric esdevé més ric". La segona variable que estudiem en aquest apartat, el nombre de cites que rep un article, és la que s'ha estudiat més, ja que és una variable que permet veure l'impacte que tenen les publicacions, i calcular el que s'anomena *Factor d'Impacte* d'una revista, indicador de la

categoria de la mateixa. Per exemple, a l'article [24] l'autor treballa amb dues variables, que són el *nombre de cites que reben els articles publicats en un any concret* i els *articles publicats en una revista concreta durant 20 anys*. Ajusta mostres d'aquestes dues variables mitjançant la distribució Zipf, estimant α per mínims quadrats. També ajusta les dades mitjançant la distribució *stretched exponential*. Però per les dades estudiades, cap de les dues distribucions s'ajusta prou bé a la cua. És per això que l'autor arriba a la conclusió que la distribució del nombre de vegades que se citen els articles no es pot explicar amb una única funció en tot el rang del recompte de cites. Com a solució, en l'article es proposa estimar dues distribucions Zipf, una per la cua i l'altra per la resta de dades. L'estimació de la distribució de la cua la fa a partir del dibuix en eixos logarítmics, els rangs dels M articles més citats i el seu nombre de cites, aleshores, estima el pendent de la recta obtinguda i el transforma, de manera que obté l'estimació del paràmetre α de la distribució Zipf que correspon als valors elevats del recompte de cites. També s'estudien aquest tipus de dades a l'article [25], on s'explica que s'ajusten bé a la distribució Zipf-Mandelbrot, que no és altra que la distribució PL discreta i també rep aquest nom perquè és una generalització de la distribució Zipf que va proposar Mandelbrot.

3.5.1 Cites que apareixen als articles

En aquesta subsecció estudiem la primera v.a. que s'ha definit, és a dir, la que correspon al *nombre de cites que apareixen als articles*. A la Taula 3.7, podem veure les freqüències observades, és a dir, el nombre d'articles, o_i, que inclouen un total de i cites. El nombre d'articles que inclouen més de 83 referències s'han agrupat. La distribució Zipf pot semblar apropiada inicialment, perquè el nombre mínim de cites que apareixen als articles és 1, existeix una concentració elevada d'observacions als primers valors de la variable, tot i que no se centra tant en el primer valor com ho feien les variables estudiades en els exemples anteriors, i la cua és llarga i concentra molt poca probabilitat. A la Figura 3.6 es pot veure que, a excepció del primer valor, el dibuix que formen els punts al principi és bastant pla. També es compleix que hi ha molt pocs articles que tenen un nombre molt elevat de cites, ja que el logaritme de la probabilitat empírica pren valors molt baixos quan el logaritme de x és gran. La freqüència de l'1 en aquest cas representa el 8.18% del total de les observacions, una probabilitat notablement més petita que les observades en els exemples anteriors. Els tres articles que tenen més cites en

tenen 322, 376 i 411.

A la Taula 3.8 hi trobem resumits els ajustos obtinguts mitjançant les distribucions Zipf i MOEZipf de forma similar als exemples anteriors. Per la distribució Zipf, l'ajust que aconseguim mitjançant el mètode de màxima versemblança no és gens bo, ja que el valor de l'estadístic X^2 és molt gran. Per la distribució MOEZipf, tenim l'estimació pel primer mètode i pel de màxima versemblança. El paràmetre α estimat per aquesta distribució és bastant proper en tots dos casos. En canvi, el paràmetre β és molt diferent, 32.558 pel primer mètode i 56.506 per màxima versemblança. Aquests valors són molt superiors a la unitat, cosa que provocarà una corbatura molt marcada en el gràfic de la probabilitat en escala logarítmica, tal com s'aprecia en la Figura 3.6. El fet que β sigui molt superior a la unitat implica que la probabilitat dels valors inicials és molt inferior a la que correspondria a una distribució Zipf amb el mateix α. Si comparem els tres ajustos, el que dóna més bons resultats és, tal com ha passat amb els altres exemples, el de la distribució MOEZipf mitjançant màxima versemblança, ja que li corresponen el X^2 i l'AIC més petits. De fet, aconseguim que el X^2 es redueixi en un 94.96% en passar de la distribució Zipf a la MOEZipf. Tot i aconseguir aquesta gran reducció, el p-valor associat al X^2 de la distribució MOEZipf estimant per màxima versemblança segueix sent pràcticament zero i, per tant, ens porta a rebutjar la hipòtesi nul·la que les dades segueixen aquesta distribució. El LRT pren un valor de $31098.54 \gg 3.84$, la qual cosa indica que la inclusió del paràmetre β té un efecte positiu i significatiu en l'ajust de les dades, respecte a l'obtingut amb la distribució Zipf.

A la Taula 3.7 apareixen les freqüències esperades, e_i, per les dues distribucions objectes d'estudi estimades per màxima versemblança i d_i, el terme que els correspon de l'estadístic X^2. S'observa que per gairebé tots els nombres de cites d'un article, i, la diferència entre observats i esperats és més petita per la distribució MOEZipf que per la Zipf. O, el que és el mateix, que el terme d_i corresponent a la MOEZipf és més petit en gairebé tots els casos. També ho podem apreciar a la Figura 3.6, on l'ajust de les dades assolit amb la distribució MOEZipf és molt millor que el de la distribució Zipf en tot el rang de valors. Amb la distribució MOEZipf s'aconsegueix capturar la corba accentuada que dibuixen les dades en eixos logarítmics. En canvi, la distribució Zipf ens porta a una recta que té molt poc a veure amb les dades.

Cites en articles		Zipf $\hat{\alpha} = 1.387$		MOEZipf $\hat{\alpha} = 2.581$, $\hat{\beta} = 56.506$	
i	o_i	e_i	d_i	e_i	d_i
1	2632	10085.633	5508.494	1726.809	474.500
2	2295	3856.279	632.110	2288.483	0.019
3	2193	2197.486	0.009	2457.230	28.413
4	1995	1474.462	183.768	2407.542	70.691
5	1868	1081.976	571.024	2246.692	63.831
6	1683	840.217	845.357	2039.179	62.213
7	1681	678.477	1481.338	1821.565	10.847
8	1453	563.766	1402.598	1612.962	15.864
9	1399	478.795	1768.562	1422.167	0.377
10	1187	413.697	1445.494	1252.213	3.396
11	1095	362.468	1480.412	1103.098	0.059
12	1044	321.260	1625.951	973.365	5.126
13	968	287.502	1610.693	860.975	13.304
14	855	259.418	1367.361	763.758	10.900
15	781	235.744	1261.132	679.650	15.114
16	745	215.558	1300.387	606.781	31.485
17	647	198.173	1016.511	543.517	19.703
18	556	183.069	759.701	488.446	9.343
19	501	169.842	645.690	440.367	8.348
20	521	158.179	832.219	398.263	37.825
...	...	...	...	...	...
65	14	30.843	9.198	26.981	6.246
66	11	30.197	12.204	25.982	8.639
67	16	29.573	6.230	25.033	3.260
68	19	28.972	3.432	24.132	1.091
69	16	28.391	5.408	23.274	2.273
70	10	27.830	11.423	22.458	6.911
71	9	27.288	12.256	21.681	7.417
72	11	26.764	9.285	20.940	4.718
73	14	26.256	5.721	20.234	1.921
74	5	25.766	16.736	19.560	10.838
75	12	25.290	6.984	18.917	2.529
76	10	24.830	8.857	18.302	3.766
77	8	24.384	11.009	17.715	5.328
78	5	23.951	14.995	17.154	8.611
79	6	23.532	13.062	16.616	6.783
80	7	23.125	11.244	16.102	5.145
81	8	22.730	9.545	15.610	3.710
82	6	22.346	11.957	15.137	5.516
≥ 83	198	4723.466	4335.766	797.308	450.479
	32158				

Taula 3.7: Per cada *nombre de cites que poden aparèixer en un article*, i, trobem la freqüència observada, o_i, les esperades, e_i, i els valors dels termes i-èssims de la suma del X^2, d_i, corresponents a les distribucions Zipf i MOEZipf.

Distribució	Paràmetre	Estimació	log-versemblança	X^2	p-valor	AIC
Zipf	$\hat{\alpha}$	1.387	-129721.65	34803.65	0	259445.30
MOEZipf	$\tilde{\alpha}$	2.492	-114774.82	3112.24	0	229553.64
(1r mèt. est.)	$\tilde{\beta}$	32.558				
MOEZipf	$\hat{\alpha}$	2.581	**-114172.38**	**1755.04**	0	**228348.77**
(màx. versem.)	$\hat{\beta}$	56.506				

Taula 3.8: Resum de l'ajust de la variable *nombre de cites que apareixen en un article* a distribucions Zipf i MOEZipf. Per cada distribució trobem el valor dels paràmetres estimats, la log-versemblança, el valor de l'estadístic X^2 amb el p-valor corresponent i l'AIC.

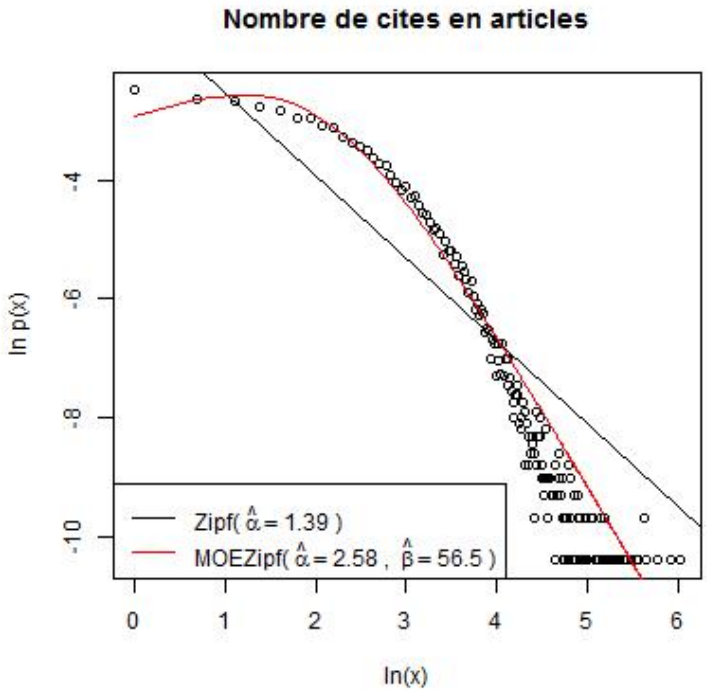

Figura 3.6: Gràfic, en eixos logarítmics, de la funció de probabilitat empírica del *nombre de cites que apareixen en un article* amb l'ajust obtingut per la distribució Zipf (en negre) i la distribució MOEZipf (en vermell).

3.5.2 Vegades que se cita un article

En aquesta secció estudiem la variàble corresponent al *nombre de vegades que se cita un article*. A la Taula 3.9 hi trobem la freqüència observada, o_i, per cada nombre de cites a un article. Veiem que el nombre mínim de cites que rep un article és 1. A més, la major part de les observacions es concentren als valors inicials, indicant que la majoria dels articles són citats poques vegades. La cua és llarga i concentra molt poca probabilitat, per tant, hi ha molt pocs articles que se citen moltes vegades. Aquesta última característica no la podem observar directament de la taula,

atès que les observacions de la cua es troben agrupades a partir de 119. És per això que ens hem de fixar en la Figura 3.7, on podem veure que el logaritme de la funció de probabilitat empírica decreix ràpidament quan augmenta el logaritme de x. Això ens indica que les probabilitats a la cua són molt properes a zero. Val a dir, que el dibuix que fa el logaritme de la probabilitat empírica ens fa pensar que la distribució Zipf no serà prou bona per les dades, ja que, tot i que tendeix a ser lineal a la cua, pels primers valors té una forma encorbada. Aquesta corba farà que el paràmetre β estimat per la MOEZipf sigui superior a la unitat.

A la Taula 3.10, de forma similar als exemples anteriors, podem veure els ajustos obtinguts per les distribucions Zipf i MOEZipf. Per una banda, s'observa que l'ajust per la distribució Zipf és molt dolent, ja que el valor de l'estadístic X^2 ens porta a un p-valor que és pràcticament zero. Per altra banda, pel que fa a la distribució MOEZipf cal esmentar que les estimacions del paràmetre β amb els dos mètodes emprats s'assemblen poc. En general, el millor ajust per les dades l'aconseguim mitjançant l'estimació màxim versemblant a la distribució MOEZipf, ja que és el que aconsegueix la log-versemblança màxima i els X^2 i AIC mínims, com ha anat passant en tots els exemples considerats. Amb aquesta estimació, la MOEZipf aconsegueix reduir el valor de l'estadístic X^2 en un 93.80% al passar de la distribució Zipf a la MOEZipf. Com a l'exemple anterior, tot i aconseguir una gran millora del valor de l'estadístic, aquest no ens permet acceptar que la distribució és adequada per les dades perquè el seu p-valor segueix sent pràcticament zero. El LRT també ens indica la superioritat de l'ajust per la distribució MOEZipf envers el de la distribució Zipf. En aquest cas, el valor de l'estadístic de raó de versemblances és 13283.74, que és molt superior al valor crític que li correspon.

Per cada nombre de cites, i, a la Taula 3.9, podem trobar les freqüències esperades, e_i, per les distribucions estudiades amb els paràmetres corresponents estimats per màxima versemblança, i el terme d_i associat a l'estadístic X^2. S'observa que pels primers valors de la variable, les freqüències esperades segons la distribució MOEZipf són més properes al valor real. Pels valors centrals de la distribució, l'ajust segons la distribució generalitzada segueix sent molt millor que el que trobem amb la distribució Zipf. I, pel que fa a la cua, la distribució MOEZipf també captura molt millor el comportament de les dades. Per tant, en aquest cas, la MOEZipf ajusta les dades millor que la Zipf en tot el seu rang.

A la Figura 3.7 podem apreciar gràficament que l'ajust per la distribució MOEZipf és millor

Cites d'articles		Zipf $\hat{\alpha} = 1.421$		MOEZipf $\hat{\alpha} = 2.161$, $\hat{\beta} = 13.058$	
i	o_i	e_i	d_i	e_i	d_i
1	4262	9464.171	2859.477	3646.107	104.035
2	3079	3534.665	58.741	3199.822	4.562
3	2351	1986.734	66.788	2629.706	29.538
4	1935	1320.121	286.395	2146.643	20.866
5	1584	961.422	403.156	1764.095	18.386
6	1279	742.002	388.633	1464.781	23.563
7	1079	596.047	391.318	1229.684	18.465
8	1039	493.037	604.571	1043.311	0.018
9	859	417.058	468.310	893.946	1.366
10	735	359.071	393.580	772.899	1.858
11	668	313.592	400.537	673.745	0.049
12	583	277.122	337.618	591.700	0.128
13	569	247.330	418.353	523.172	4.014
14	483	222.611	304.579	465.435	0.663
15	429	201.823	255.715	416.401	0.381
16	421	184.139	304.679	374.448	5.787
17	360	168.941	216.074	338.310	1.391
18	326	155.762	186.058	306.985	1.178
19	331	144.244	241.797	279.673	9.420
20	302	134.105	210.198	255.732	8.371
...	...	...	...	...	...
100	15	13.623	0.139	10.873	1.566
101	12	13.432	0.153	10.650	0.171
102	14	13.245	0.043	10.433	1.220
103	15	13.063	0.287	10.222	2.233
104	10	12.885	0.646	10.018	0.000
105	6	12.711	3.543	9.819	1.486
106	13	12.541	0.017	9.627	1.182
107	11	12.374	0.153	9.439	0.258
108	10	12.212	0.401	9.257	0.060
109	6	12.053	3.040	9.080	1.045
110	6	11.898	2.923	8.908	0.950
111	6	11.746	2.811	8.741	0.860
112	6	11.597	2.701	8.578	0.775
113	7	11.451	1.730	8.420	0.239
114	8	11.309	0.968	8.266	0.009
115	7	11.169	1.556	8.116	0.153
116	8	11.033	0.834	7.970	0.000
117	5	10.899	3.193	7.827	1.021
118	10	10.768	0.055	7.689	0.695
≥ 119	307	3013.465	2430.741	798.865	302.844
	28230				

Taula 3.9: Per cada *nombre de cites a un article*, i, trobem la freqüència observada, o_i, les esperades, e_i, i els valors dels termes i-èssims de la suma del X^2, d_i, corresponents a les distribucions Zipf i MOEZipf.

Distribució	Paràmetre	Estimació	log-versemblança	X^2	p-valor	AIC
Zipf	$\hat{\alpha}$	1.421	-105839.81	13172.05	0	211681.61
MOEZipf	$\tilde{\alpha}$	2.214	-99490.01	1615.25	0	198984.03
(1r mèt. est.)	$\tilde{\beta}$	11.677				
MOEZipf	$\hat{\alpha}$	2.161	**-99197.93**	**816.62**	0	**198399.87**
(màx. versem.)	$\hat{\beta}$	13.058				

Taula 3.10: Resum de l'ajust de la variable *nombre de vegades que se cita un article* a distribucions Zipf i MOEZipf. Per cada distribució trobem el valor dels paràmetres estimats, la log-versemblança, el valor de l'estadístic X^2 amb el p-valor corresponent i l'AIC.

que l'ajust per la distribució Zipf. El logaritme de la funció de probabilitat estimada per la distribució MOEZipf s'ajusta molt bé al dibuix que formen les dades, ja que captura perfectament la corbatura inicial i, després, és lineal per a valors elevats del logaritme de x. Pel que fa a la distribució Zipf, la recta ajustada queda allunyada dels valors en pràcticament tot el rang i sobreestima gairebé tots els valors de la cua.

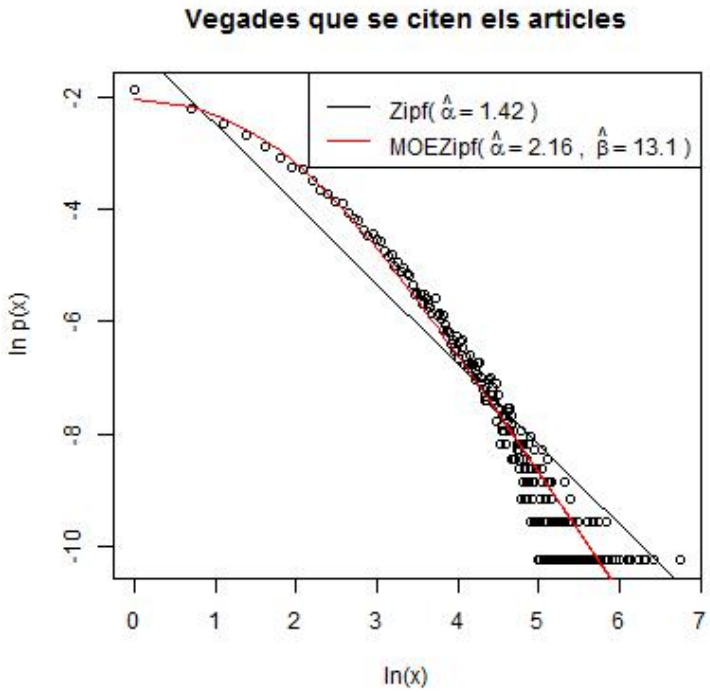

Figura 3.7: Gràfic, en eixos logarítmics, de la funció de probabilitat empírica del *nombre de cites que apareixen en articles* amb l'ajust obtingut per la distribució Zipf (en negre) i la distribució MOEZipf (en vermell).

Capítol 4

Test de Kolmogorov-Smirnov discret

4.1 Introducció

En aquest capítol hem utilitzat el test de Kolmogorov-Smirnov (KS) en la seva versió discreta, per tal de veure si els conjunts de dades ajustats en el Capítol 3 segueixen o no la distribució MOEZipf. Aquest test, a diferència del χ^2, és no paramètric i s'utilitza tan per a comparar la distribució de dues mostres, com per a decidir si una mostra prové o no d'una distribució teòrica determinada, com és el nostre cas. Un avantatge del test χ^2 respecte del de KS és que un cop fixat el nombre de nivells, la seva distribució no depèn de la distribució que s'estigui testant en la hipòtesi nul·la. Això no passa amb el test de KS discret. Ara bé, aquest últim té en compte l'ordre natural de les observacions, cosa que el primer no.

Els motius d'utilitzar aquest test han estat bàsicament dos. D'una banda, el fet d'observar que d'altres autors (veure [7]) l'han utilitzat pel cas concret de la Zipf. De l'altra, el fet que per tal de poder aplicar el test χ^2, en tots els casos hem hagut d'agrupar els valors a partir d'un determinat valor molt llunyà al darrer valor observat. Tanmateix, es coneix que l'acceptació de la distribució especificada en la hipòtesi nul·la del test χ^2 depèn en gran mesura de l'agrupació que s'hagi dut a terme. A l'article esmentat, els autors busquen la millor manera d'avaluar la bondat de l'ajust quan treballem amb la ditribució Zipf i hem pensat que fóra bo fer el paral·lelisme pel cas de la MOEZipf.

El test de KS pel cas discret es defineix de la forma següent:

Si $x_1, \ldots, x_N$ és una m.a.s. d'una v.a. Y amb funció de distribució F_N, i es vol portar a terme

el test d'hipòtesi:

$$H_0: \quad F_N = F$$
$$H_1: \quad F_N \neq F\,,$$

on F és la funció de distribució d'una determinada distribució teòrica, cal calcular l'estadístic:

$$KS_d = \sqrt{N} \max_{x \in J} |F_N(x) - F(x)|\ ,$$

on $F_N(x)$ és la funció de distribució acumulada empírica, N és la mida de la mostra i J és el conjunt de punts de discontinuïtat de $F(x)$.

En el cas discret, la distribució de KS_d depèn de $F(x)$ tal com ja s'ha dit, i també de la grandària mostral (N). Per aquest motiu, serà necessari calcular via simulacions el valor crític associat a KS_d pels valors de N que sigui necessari. La realització de les simulacions s'ha fet seguint la mateixa estratègia que s'utilitza a l'article [2], on s'utilitza la versió contínua de l'estadístic, però nosaltres hem utilitzat la versió discreta d'aquest. S'ha escollit aquesta estratègia perquè és la més còmoda d'implementar des del punt de vista de les simulacions. Els passos a seguir s'expliquen en detall a la secció següent.

4.2 Algorisme

L'estratègia per tal de testar si un determinat conjunt de dades prové o no d'una distribució $MOEZipf(\alpha, \beta)$ és la següent. Donada una m.a.s.,

1. S'estimen per màxima versemblança els paràmetres α i β.

2. Es generen un nombre gran de m.a.s. de la distribució $MOEZipf(\hat{\alpha}, \hat{\beta})$. En el nostre cas, se n'han generat 1000.

3. Per a cada mostra obtinguda es calcula el valor de KS_d.

4. Amb els valors de KS_d de totes les mostres, es busca el que deixa per sota una probabilitat igual al nivell de confiança del test. Aquest valor el notem per VC. El nivell de confiança

del test l'hem fixat a 0.95.

5. Si el valor de KS_d obtingut amb la mostra inicial, que denotem per KS_d^*, compleix que $KS_d^* \geq VC$, es rebutja la hipòtesi de que la mostra prové d'una MOEZipf, en cas contrari es conclou que no hi ha indicis per a rebutjar aquesta distribució de probabilitat com a una possible distribució de les dades.

Per tal de poder realitzar les simulacions que s'acaben d'explicar, ha estat necessari programar l'algorisme. S'ha fet utilitzant R i el codi es pot trobar a l'Apèndix B. La notació i l'algorisme resumit figuren a continuació.

Notació

g: funció definida per vectors de probabilitats acumulades que retorna, per cada probabilitat, el valor de la v.a. x que li correspon.

$K[j]$: valor de l'estadístic KS que correspon a la mostra j, $\forall\, j = 1, \ldots, nsim$

VC: valor crític

Dades

X	//Conjunt de dades//
α i β	//Paràmetres estimats pel conjunt de dades X//

Inicialització

$n \leftarrow length(X)$	//Mida de la mostra a generar//
$nsim \leftarrow 1000$	//Nombre de mostres a generar//
$\mathbf{F} \leftarrow P(Y \leq y;\, \alpha, \beta),\ y = 1, \ldots, m$	//Funció teòrica de probabilitat acumulada//

Per a $j \leftarrow 1$ **fins a** $nsim$ **fer**

Generar $u_1, \ldots, u_n$	//Vector de valors entre 0 i 1//
$x_1, \ldots, x_n = g(u_1, \ldots, u_n)$	//Mostra//
$\mathbf{S} = s_1, \ldots, s_m = P(Z \leq z),\ z = 1, \ldots, m$	//Funció empírica de probabilitat acumulada//
$K[j] = \sqrt{n}\, \max \lvert\mathbf{F} - \mathbf{S}\rvert$	//Estadístic KS de la mostra j//

Fi per a

Trobar VC t.q. $P(K \leq VC) = 0.95$ //Valor crític obtingut//

Mitjançant aquest algorisme, es poden calcular els valors crítcs dels estadístics de KS per a distribucions Zipf i MOEZipf. Per la primera de les distribucions, només cal que fixem el paràmetre $\beta = 1$. Noti's que la grandària de les mostres simulades depèn de la mida del conjunt de dades. Per tant, el temps d'execució de l'algorisme vindrà determinat per aquesta mida. Com a conseqüència d'això, no s'han pogut calcular els valors crítics corresponents al conjunt de dades del correu electrònic. Ja que, com s'ha destacat en el capítol anterior, es tracta d'un conjunt de dades molt gran (N=225409). Per a la resta de conjunts de dades analitzats en el capítol 3 es presenten els resultats obtinguts en la secció següent.

4.3 Resultats

A la Taula 4.1 figuren els valors de l'estadístic de KS per a cada conjunt de dades analitzat, així com els valors crítics associats a les dues distribucions estudiades en aquest treball. Aquests valors crítics obtinguts mitjançant simulacions els trobem entre parèntesi.

Dades	Zipf	MOEZipf
Terrorisme ($N = 9101$)	4.093 (1.025)	**0.579** (1.073)
Terrorisme sense 11-S ($N = 9100$)	4.114 (1.007)	**0.595** (1.110)
Paraules ($N = 18855$)	4.754 (1.236)	**0.548** (1.146)
Nombre de cites en articles ($N = 32158$)	50.296 (10.172)	5.181 (1.219)
Cites a un article ($N = 28230$)	33.674 (7.509)	3.666 (1.377)

Taula 4.1: Estadístics de KS_d per les distribucions Zipf i MOEZipf. Per cada conjunt de dades disposem del valor de l'estadístic i, entre parèntesis, el valor crític que li correspon. En negreta figuren els estadístics KS_d que no són significatius.

S'observa que per a tots els conjunts de dades es rebutja la distribució Zipf com a adequada per a les dades. També s'observa que per a les dades del terrorisme (amb i sense l'11-S) i les corresponents al nombre de paraules en el llibre *Moby Dick*, no es pot rebutjar la distribució MOEZipf ajustada, ja que el valor de l'estadístic surt inferior al valor crític obtingut. Les dues distribucions es rebutgen en el cas dels dos exemples relacionats amb cites. Observi's que aquests dos exemples contenen un nombre molt més elevat d'observacions. Per a mostres de mida gran, el test considerat en aquest capítol és més difícil d'acceptar. Això es deu al fet que per calcular el valor crític hem generat mostres de la mateixa mida que el conjunt de dades. I, encara que

l'estadístic es multipliqui pel terme $\sqrt{N}$, quan la mida mostral és molt gran, les diferències entre les disribucions empírica i teòrica arriben a ser tan petites que els valors crítics també prenen valors petits.

En tots els casos, tal com passava amb el χ^2 de Pearson, el valor de l'estadístic KS_d es redueix moltíssim en passar de la distribució Zipf a la MOEZipf. Cosa que, encara que els estadístics no siguin directament comparables, implica que amb la generalització presentada ajustem, en general, qualsevol conjunt de dades de forma molt més satisfactòria.

Conclusions

La transformació Marshall-Olkin ha resultat ser apropiada per tal de generalitzar la distribució Zipf. Concretament, el paràmetre addicional de la distribució MOEZipf li atorga més flexibilitat especialment de cara a adaptar les probabilitats dels primers valors de la variable. Tanmateix aquest paràmetre es pot interpretar com la raó de probabilitats de la distribució MOEZipf i la Zipf en el límit, és a dir per a valors grans de x. Pel que fa a l'esperança de la distribució, variant els valors del paràmetre β, aquesta no canvia de forma, segueix essent una funció convexa amb límit la unitat quan α creix. El que fa variar el paràmetre β és el pendent en cada punt. S'ha provat que la distribució límit quan β tendeix a zero és la distribució degenerada en la unitat. Respecte al quocient de dues probabilitats consecutives, la generalització permet quocients superiors o inferiors als obtinguts amb una Zipf amb el mateix paràmetre α. Els quocients tendeixen, però, als de la Zipf quan el valor de x tendeix a infinit. Tot i que no s'ha demostrat, s'observa que el fet que el paràmetre β sigui superior o inferior a la unitat determina que el logaritme de la probabilitat sigui una funció cóncava o convexa del logaritme de x.

La distribució proposada demostra ser molt útil de cara a l'ajust a dades reals. Dels dos mètodes d'estimació considerats, el màxim versemblant és el que dóna lloc a millors ajustos. Els conjunts de dades estudiats provenen d'àmbits molt diversos, i en tots ells s'ha vist una millora de la bondat d'ajust molt significativa si s'utilitza la distribució generalitzada. Les reduccions dels valors dels estadístics χ^2 de Pearson van des d'un 76% a un 94%. Les reduccions més elevades s'obtenen amb els conjunts de dades més grans. A partir del test de Kolmogorov-Smirnov discret, si bé en tots els conjunts de dades estudiats es rebutja la distribució Zipf com a distribució apropiada per les dades, no es pot rebutjar la distribució MOEZipf. Aquesta només es rebutja en els casos en que la grandària mostral és molt elevada (de l'ordre de 30.000).

Creiem que la distribució definida en aquest treball constitueix una molt bona alternativa bi-

paramètrica a la distribució Zipf. Per tal de donar a conèixer aquesta nova distribució s'ha escrit un article titulat *The Marshall-Olkin Extended Zipf Distribution* que, en l'actualitat, està sotmès a consideració per a ser publicat en una revista.

Apèndix A

Ajust terrorisme sense 11-S

A continuació es presenten les taules corresponents a l'ajust de les dades del terrorisme sense tenir en compte l'atemptat de l'11-S.

Distribució	Paràmetre	Estimació	log-versemblança	X^2	p-valor	AIC
Zipf	$\hat{\alpha}$	1.898	-16698.12	197.184	0.000	33398.24
MOEZipf	$\tilde{\alpha}$	2.234	-16612.12	65.295	0.000	33228.23
(1r mèt. est.)	$\tilde{\beta}$	1.905				
MOEZipf	$\hat{\alpha}$	2.163	**-16606.42**	**46.843**	0.001	**33216.84**
(màx. versem.)	$\hat{\beta}$	1.760				

Taula A.1: Resum de l'ajust de la variable *nombre de morts dels atacs terroristes* a distribucions Zipf i MOEZipf. Per cada distribució trobem el valor dels paràmetres estimats, la log-versemblança, el valor de l'estadístic X^2 amb el p-valor corresponent i l'AIC. No s'inclou l'atac terrorista de l'11-S.

Morts		Zipf $\hat{\alpha} = 1.897$		MOEZipf $\hat{\alpha} = 2.160$, $\hat{\beta} = 1.752$	
i	o_i	e_i	d_i	e_i	d_i
1	4802	5194.422	29.646	4775.841	0.143
2	1600	1393.565	30.580	1636.882	0.831
3	750	645.467	16.929	779.031	1.082
4	444	373.867	13.156	445.922	0.008
5	287	244.774	7.285	285.669	0.006
6	190	173.167	1.636	197.296	0.270
7	148	129.237	2.724	143.768	0.125
8	96	100.302	0.184	109.050	1.562
9	85	80.207	0.286	85.331	0.001
10	92	65.668	10.559	68.451	8.101
11	60	54.800	0.493	56.036	0.280
12	64	46.457	6.624	46.653	6.450
13	43	39.909	0.239	39.399	0.329
14	32	34.672	0.206	33.682	0.084
15	39	30.416	2.423	29.100	3.368
16	29	26.909	0.162	25.374	0.518
17	34	23.984	4.183	22.306	6.130
18	17	21.518	0.949	19.752	0.383
19	17	19.419	0.301	17.603	0.021
20	17	17.618	0.022	15.779	0.094
21	18	16.059	0.235	14.219	1.005
22	18	14.702	0.740	12.874	2.041
23	17	13.512	0.900	11.708	2.392
24	9	12.464	0.963	10.689	0.267
25	17	11.534	2.590	9.795	5.299
26	5	10.707	3.042	9.006	1.782
27	8	9.967	0.388	8.307	0.011
≥ 28	162	294.678	59.738	190.476	4.257
	9100				

Taula A.2: Per cada *nombre de morts en atacs terroristes*, i, trobem la freqüència observada, o_i, les esperades, e_i, i els valors dels termes i-èssims de la suma del X^2, d_i, corresponents a les distribucions Zipf i MOEZipf. No s'inclou l'atac terrorista de l'11-S.

Apèndix B

Codi R

En aquest aparta s'inclouen tots els fitxers de sintaxi de R que s'han utilitzat per la realització del treball.

B.1 Gràfics

Per generar els gràfics dels dos primers capítols d'aquest treball, s'han utilitzat els fitxers R següents:

zeta_x.R: Funció $\zeta(\alpha, x)$. En R disposem de la funció zeta de Riemann, `zeta()`, que es pot trobar al paquet VGAM. Però ens ha fet falta programar la funció zeta de Hurwitz.

```
##### Funció zeta(alpha,x)
library(VGAM)

zeta_x<-function(alpha,x){
aux<-0
if(x==1) {
zeta(alpha)
} else{
zeta(alpha)-sum((1:(x-1))^(-alpha))
}}
```

Zipf.R: Gràfics descriptius de la distribució Zipf.

```
##### Gràfics descriptius de la distribució PL
setwd("D:/ACASELLAS/upcNOU/PROJECTE/Scripts")
source("zeta_x.R")

a<-c(1.1,1.5)
xmin<-c(1,5)
prImg0<-matrix(c(rep(0,2*20)),nrow=20)
```

```
for(j in 1:2){
for (i in 1:20){
prImg0[i,j]<-((i+xmin[j]-1)^(-a[1+j]))/zeta_x(a[1+j],xmin[j])
}}

### Gràfic distribucions PL(x_min,alpha)
jpeg('grafs//img0.jpg', width = 650, height = 350, units = "px")
par(mfrow=c(1,2))
for(j in 1:length(xmin)){
plot(xmin[j]:(xmin[j]+19),prImg0[,j], type="h", lwd=5,ylab="probabilitat (tant
  per 1)", xlab="x",xlim=c(1,19+xmin[j]),ylim=c(0,max(prImg0)),
  main=bquote("PL("~  x[min] == .(xmin[j]) ~ ", " ~ alpha == .(a[1+j]) ~ ")"),
  col=c(rep("springgreen1",4), rep("springgreen4", 16)))
}
dev.off()

### Gràfic distribucions PL(x_min,alpha) en eixos logarítmics
jpeg('grafs//img0ln.jpg', width = 650, height = 350, units = "px")
par(mfrow=c(1,2))
for(j in 1:length(xmin)){
plot(log(xmin[j]:(xmin[j]+19)),log(prImg0[,j]),pch=19, lwd=4,ylab="ln p(x)",
  xlab="ln(x)", xlim=log(c(1,19+xmin[j])),ylim=log(c(min(prImg0),max(prImg0))),
  main=bquote("PL("~  x[min] == .(xmin[j]) ~ ", " ~ alpha == .(a[1+j]) ~ ")"),
  col=c(rep("springgreen1",4), rep("springgreen4", 16)))
}
dev.off()

### Gràfic de funcions de probabilitat complementària (P), supervivència (Fbar)
### i risc (r) de distribucions PL(x_min,alpha)
FbarImg0<-1-apply(prImg0,2,cumsum)
PImg0<-rbind(c(1,1),FbarImg0[1:19,])
rImg0<-matrix(c(rep(NA,2*20)),nrow=20)
for(j in 1:2){
 for(i in 1:20){
  rImg0[i,j]<-(i-1+xmin[j])^(-a[j+1])/zeta_x(a[j+1],i-1+xmin[j])
}}

jpeg('grafs//Img0fun.jpg',width = 450, height = 700, units = "px")
par(mfrow=c(3,2))
  for(j in 1:2){
   plot(stepfun((xmin[j]):(xmin[j]+19),c(1,FbarImg0[,j],f=0)), do.points=T, pch=16,
 cex=1, ylab=bquote(bar(F)~"(x)"), xlim=c(1,19+xmin[j]),ylim=c(min(FbarImg0),1),
   main=bquote("PL("~  x[min] == .(xmin[j]) ~ ", " ~ alpha == .(alpha[1+j]) ~ ")"))
 abline(v=xmin[j],lty=2)
  }
  for(j in 1:2){
   plot(stepfun((xmin[j]+1):(xmin[j]+19),PImg0[,j],f=0), do.points=T, pch=16,cex=1,
 ylab="P(x)", xlab="x", xlim=c(1,19+xmin[j]),ylim=c(min(PImg0),1),
   main=bquote("PL("~  x[min] == .(xmin[j]) ~ ", " ~ alpha == .(alpha[1+j]) ~ ")"))
 abline(v=xmin[j],lty=2)
  }
  for(j in 1:2){
   plot(xmin[j]:(19+xmin[j]),rImg0[,j], ylab=bquote("r(x; " ~ alpha ~ ")"),xlab="x",
   xlim=c(1,19+xmin[j]),ylim=c(min(rImg0),max(rImg0)),pch=19,
   main=bquote("PL("~  x[min] == .(xmin[j]) ~ "," ~ alpha == .(alpha[1+j]) ~ ")"))
 abline(v=xmin[j],lty=2)
```

```
  }
dev.off()

### Gràfic de 4 PL(x_min=1, a)
a<-c(1.1,1.5,2.5,5)
p<-matrix(c(rep(NA,10*length(a))),nrow=10)

for(j in 1:length(a)){
for (i in 1:10){
p[i,j]<-(i^(-a[j]))/zeta(a[j])
}}

jpeg('grafs//img1.jpg')
par(mfrow=c(2,2))
for(k in 1:length(a)){
plot(1:10,p[,k],type="h",ylab="probabilitat (tant per 1)", xlab="", ylim=c(0,0.95),
 main=bquote("Zipf(" ~ alpha == .(a[k]) ~ ")"),lwd=2)
}
dev.off()

### Gràfic de 4 PL(x_min=1, a) en eixos logarítmics
p<-NULL
p<-matrix(c(rep(NA,20*length(a))),nrow=20)

for(j in 1:length(a)){
for (i in 1:20){
p[i,j]<-(i^(-a[j]))/zeta(a[j])
}}

jpeg('grafs//img1ln.jpg')
par(mfrow=c(2,2))
for(k in 1:length(a)){
plot(log(1:20),log(p[,k]),ylim=c(-15,0),ylab="ln p(x)", xlab="ln x",
 main=bquote("Zipf(" ~ alpha == .(a[k]) ~ ")"),pch=16, lwd=1)
}
dev.off()
```

MOEZipf.R: Gràfics descriptius de la distribució MOEZipf.

```
##### Gràfics descriptius de la distribució MOEZipf
setwd("D:/ACASELLAS/upcNOU/PROJECTE/Scripts")
source("zeta_x.R")

a<-c(rep(1.8,6))
b<-c(0.2,0.8,1,1.2,2,5)

### Gràfic de 6 MOEZipf(alpha=1.8,beta)
g<-matrix(c(rep(NA,10*length(a))),nrow=10)

for(j in 1:length(a)){
for (i in 1:10){
g[i,j]<-(b[j]*zeta(a[j])*(i^(-a[j])))/
((zeta(a[j])-(1-b[j])*zeta_x(a[j],i))*(zeta(a[j])-(1-b[j])*zeta_x(a[j],i+1)))
}}
```

```
jpeg('grafs//img3.jpg', width = 400, height = 600, units = "px")
par(mfrow=c(3,2))
for(k in 1:length(a)){
plot(1:10,g[,k],type="h",ylab="probabilitat (tant per 1)", ylim=c(0,0.85),xlab="",
cex.lab=1.25, main=bquote("MOEZipf(" ~ alpha == .(a[k]) ~ ", " ~  beta == .(b[k])
~ ")"),lwd=2)
}
dev.off()

### Gràfic de 6 distribucions MOEZipf(alpha=1.8,beta) en escala logarítmica amb
### la recta que aproxima el logaritme de la probabilitat
g<-NULL
g<-matrix(c(rep(NA,30*length(a))),nrow=30)

for(j in 1:length(a)){
for (i in 1:30){
g[i,j]<-(b[j]*zeta(a[j])*(i^(-a[j])))/
((zeta(a[j])-(1-b[j])*zeta_x(a[j],i))*(zeta(a[j])-(1-b[j])*zeta_x(a[j],i+1)))
}}

jpeg('grafs//img3ln.jpg', width = 400, height = 600, units = "px")
par(mfrow=c(3,2))
for(k in 1:length(a)){
plot(log(1:30),log(g[,k]),ylab="ln p(x)", ylim=c(-8.5,0),xlab="ln x",
cex.lab=1.25, main=bquote("MOEZipf(" ~ alpha == .(a[k]) ~ ", "
 ~  beta == .(b[k]) ~ ")"),pch=16, lwd=1)
abline(a=log(b[k]/zeta(a[k])), b=-a[k], col="red")
}
dev.off()

### Gràfic de 12 distribucions MOEZipf(alpha,beta) en escala logarítmica i amb
### la recta que aproxima el logaritme de la probabilitat
a<-NULL; b<-NULL; g<-NULL
a<-c(rep(1.1,4),rep(1.5,4),rep(3,4))
b<-c(rep(c(0.5,1,2,5),3))
g<-matrix(c(rep(NA,30*length(a))),nrow=30)

for(j in 1:length(a)){
for (i in 1:30){
g[i,j]<-(b[j]*zeta(a[j])*(i^(-a[j])))/
((zeta(a[j])-(1-b[j])*zeta_x(a[j],i))*(zeta(a[j])-(1-b[j])*zeta_x(a[j],i+1)))
}}

jpeg('grafs//img3ln2.jpg', width = 750, height = 550, units = "px")
par(mfrow=c(3,4))
for(k in 1:length(a)){
plot(log(1:30),log(g[,k]),ylab="ln p(x)", ylim=c(-11.4,0),xlab="ln x",
cex.lab=1.25, main=bquote("MOEZipf(" ~ alpha == .(a[k]) ~ ", "
~  beta == .(b[k]) ~ ")") ,pch=16, lwd=1)
abline(a=log(b[k]/zeta(a[k])), b=-a[k], col="red")
}
dev.off()

### Gràfic de la funció de risc de distribucions MOEZipf(alpha=1.1, beta) comparades
### amb la d'una Zipf(alpha=1.1)
a<-NULL; b<-NULL
```

```
a<-c(rep(1.1,4))
b<-c(0.2,0.6,1.3,2)
rF<-matrix(c(rep(NA,10*length(a))),nrow=10)
rG<-matrix(c(rep(NA,10*length(a))),nrow=10)

for(j in 1:length(a)){
for(i in 1:10){
rF[i,j]<-(i^(-a[j]))/zeta_x(a[j],i+1)
rG[i,j]<-zeta(a[j])/((zeta(a[j])-(1-b[j])*zeta_x(a[j],i+1))*(i^a[j])*zeta_x(a[j],i+1))
}}

jpeg('grafs//img4.jpg')
par(mfrow=c(2,2))
for(j in 1:length(b)){
plot(1:10, rG[,j],pch=19, ylim=c(0,max(rG[,j],rF[,j])),main=bquote(alpha == .(a[j]) ~
" i " ~ beta == .(b[j])), ylab="", xlab="x")
points(1:10,rF[,j], pch=19, col=2)
}
dev.off()

### Gràfics de l'evolució de l'esperança de la distribució MOEZipf(alpha,beta) com a
### funció d'alpha i de beta
# funció que calcula la mitjana
fnMitjana<-function(a,b){
aux<-1; i<-1; suma<-0; xi_a<-zeta(a)
while(aux>0.0001){
aux<- zeta(a)*b*i^(-a+1)/((xi_a-(1-b)*zeta_x(a,i))*(xi_a-(1-b)*zeta_x(a,i+1)))
suma<-suma+aux
i<-i+1
}
return(suma)
}
# esperança com a funció d'alpha
a<-NULL; b<-NULL
a<-(21:50)/10
b<-c(0.5,1,1.5,3)
Ea<-matrix(rep(NA,30*length(b)),ncol=length(b))
for(i in 1:length(b)){
Ea[,i]<-sapply(a,fnMitjana,b=b[i])
}
jpeg('grafs//img3E.jpg', width = 750, height = 450, units = "px")
par(mfrow=c(1,2))
plot(a,Ea[,4],pch=" ",main=bquote("Esperança com a funció d'"~alpha),
xlab=bquote(alpha),ylab="Esperança")
for(i in 1:length(b)){
lines(a,Ea[,i], col=i)
}
legend("topright", legend=c(as.expression(bquote(beta == .(b[1]))),
as.expression(bquote(beta == .(b[2]))),
as.expression(bquote(beta == .(b[3]))),
as.expression(bquote(beta == .(b[4])))), col=1:4,lty=1)

# esperança com a funció de beta
a<-NULL; b<-NULL
a<-c(2.5,3,5,10)
b<-(1:30)/5
```

```
Eb<-matrix(rep(NA,30*length(a)),ncol=length(a))
for(i in 1:length(a)){
Eb[,i]<-sapply(b,fnMitjana,a=a[i])
}
plot(b,Eb[,1],pch=" ",main=bquote("Esperança com a funció de "~beta),
xlab=bquote(beta),ylab="Esperança",ylim=c(min(Eb),max(Eb)))
for(i in 1:length(a)){
lines(b,Eb[,i], col=i)
}
legend("topleft", legend=c(as.expression(bquote(alpha == .(a[1]))),
as.expression(bquote(alpha == .(a[2]))),
as.expression(bquote(alpha == .(a[3]))),
as.expression(bquote(alpha == .(a[4])))), col=1:4,lty=1)
dev.off()

### Gràfic de l'evolució, com a funció de beta, de la probabilitat a l'1 d'una
### distribució MOEZipf(alpha=1.2,beta)
a<-NULL; b<-NULL
a<-1.2
C<-zeta(alpha)-1
fBeta<-function(beta) 1/(C*beta+1)
x<-sort(c(0,runif(200)*3),decreasing=T)

jpeg('grafs//img5.jpg')
plot(c(0,3), c(0,1), type = "n",xaxs="i",yaxs="i", xlab=bquote(beta),
ylab="P(Y=1)",
main=bquote("P(Y=1) = 1 / (" ~ zeta ~ "(" ~ .(a) ~ ", 2)"  ~beta ~" + 1)"))
lines(x,fBeta(x))
abline(h=1/(C+1),lty=2, col = "cornsilk4")
abline(v=1,lty=2, col = "cornsilk4")
dev.off()

### Gràfic de 3 distribucions MOEZipf(alpha=1.5,beta) amb betes molt grans
a<-NULL; b<-NULL; g<-NULL
a<-c(rep(1.5,3))
b<-c(10,100,500)
g<-matrix(c(rep(NA,100*length(a))),nrow=100)

for(j in 1:length(a)){
for (i in 1:100){
g[i,j]<-(b[j]*zeta(a[j])*(i^(-a[j])))/
((zeta(a[j])-(1-b[j])*zeta_x(a[j],i))*(zeta(a[j])-(1-b[j])*zeta_x(a[j],i+1)))
}}

jpeg('grafs//img7.jpg', width = 400, height = 600, units = "px")
par(mfrow=c(3,1))
for(k in 1:length(a)){
plot(log(1:100),log(g[,k]),ylab="ln p(x)", ylim=c(-9.5,-2),xlab="ln x",cex.lab=1.25,
 main=bquote("MOEZipf(" ~ alpha == .(a[k]) ~ ", " ~  beta == .(b[k]) ~ ")")
 ,pch=16, lwd=1)
}
dev.off()

### Gràfic de l'evolució del quocient de les probabilitats consecutives de distribucions
### MOEZipf(alpha, beta)
a<-NULL; b<-NULL; g<-NULL
```

```
a<-c(1.1,2.5)
b<-c(0.5,1,1.5,2.5)
n<-20
g<-matrix(c(rep(NA,n*length(b)*length(a))),nrow=n)
m<-1

for(k in 1:length(a)){
for(j in 1:length(b)){
for (i in 1:n){
g[i,m]<-(b[j]*zeta(a[k])*(i^(-a[k])))/
((zeta(a[k])-(1-b[j])*zeta_x(a[k],i))*(zeta(a[k])-(1-b[j])*zeta_x(a[k],i+1)))
}
m<-m+1
}}

g2_g1<-matrix(c(rep(NA,(n-1)*length(a)*length(b))),nrow=(n-1))
m<-1

for(j in 1:(length(a)*length(b))){
for(i in 1:(n-1)){
g2_g1[i,j]<-g[i+1,j]/g[i,j]
}}

jpeg('grafs//img8.jpg', width = 700, height = 400, units = "px")
m<-0
par(mfrow=c(1,2))
for(k in 1:length(a)){
plot(g2_g1[1:13,k],type="n", ylim=c(min(g2_g1[1:13,]),max(g2_g1[1:13,])),
    ylab=bquote(p[k+1]/p[k]), xlab=c(""),main=bquote(alpha == .(a[k])))
legend("bottomright", legend=c(as.expression(bquote(beta == .(b[1]))),
    as.expression(bquote(beta == .(b[2]))),as.expression(bquote(beta ==
    .(b[3]))),as.expression(bquote(beta == .(b[4])))), col=1:4, pch=20)
for(j in 1:length(b)){
points(g2_g1[,j+m], col=j,pch=20)
}
m<-m+length(b)}
dev.off()
```

B.2 Estimació dels paràmetres

El codi programat en R per trobar les estimacions dels paràmetres de les dues distribucions estudiades és el següent:

```
setwd("D:/ACASELLAS/upcNOU/PROJECTE/Scripts")
# executem l'arxiu zeta_x.R per disposar de la funció Zeta(alpha,x)
source("zeta_x.R")

##############################################################################
###################### INTRODUCCIÓ DE LES DADES ##############################
# 1. Paraules Moby Dick
# 21. Severitat atacs terroristes
# 22. Severitat atacs terroristes (sense l'11-S)
# 3. Xarxa email
# 41. Xarxa de cites (nombre de cites en articles)
# 42. Xarxa de cites (nombre vegades que se citen els articles)
```

```
k<-1 #escollir conjut de dades

if(k==1){
paraules<-read.table('dades//words.txt')
x<-unique(rev(paraules$V1))
dades<-as.data.frame(table(paraules))
Freq2<-tabulate(rev(paraules$V1))

} else if(k==21){
terrorism<-read.table('dades//terrorism.txt')
x<-unique(terrorism$V1)
dades<-as.data.frame(table(terrorism))
Freq2<-tabulate(terrorism$V1)

 } else if(k==22){
terrorism<-read.table('dades//terrorism.txt')
x<-unique(terrorism$V1[-dim(terrorism)[1]])
dades<-as.data.frame(table(terrorism))
Freq2<-tabulate(terrorism$V1)

}else if(k==3){
email<-read.table('dades//Email.txt')

# recompte del nombre de connexions de cada adreça electrònica
connexions<-as.data.frame(table(email[,1]))

dades<-as.data.frame(table(connexions[,2]))
x<-sort(unique(as.numeric(connexions[,2])))
Freq2<-tabulate(as.numeric(connexions[,2]))

} else if(k==41){
cites<-read.table('dades//Cit-HepPh.txt')

# recompte del nombre de cites en articles
n_cites<-as.data.frame(table(cites[,1]))

dades<-as.data.frame(table(n_cites[,2]))
x<-sort(unique(as.numeric(n_cites[,2])))
Freq2<-tabulate(as.numeric(n_cites[,2]))

} else if(k==42){
cites<-read.table('dades//Cit-HepPh.txt')

# recompte del nombre de vegades que se citen els articles
n_cites<-as.data.frame(table(cites[,2]))

dades<-as.data.frame(table(n_cites[,2]))
x<-sort(unique(as.numeric(n_cites[,2])))
Freq2<-tabulate(as.numeric(n_cites[,2]))
}

Freq<-as.numeric(dades[,2])
Prob<-Freq[which(Freq>0)]/sum(Freq)

N<-sum(Freq) # mida mostra
```

```
f1<-Freq[1] # freqüència de l'1

##############################################################################
######################### ESTIMACIÓ DE PARÀMETRES ############################

######### Distribució Zipf (MLE)
## logaritme de la versemblança d'una Zipf per dades agregades
fnZ<-function(alpha,x,Freq){
n<-sum(Freq)
-n*log(zeta(alpha))-alpha*sum(Freq*log(x))
}

## Maximitzem el logaritme de la versemblança d'una Zipf
l_Z<-optimize(fnZ,x=x,Freq=Freq, c(1, 100), maximum=T, tol=0.0001)

# paràmetre estimat
a_Z<-l_Z$maximum

######### Distribució MOEZipf (MLE)
## logaritme de la versemblança d'una MOEZipf per dades agregades
# funció auxiliar
fn_aux<-function(a,b,x){
log(zeta(a)-(1-b)*zeta_x(a,x))
}

fnMOEZ<-function(param,x,Freq){
n<-sum(Freq)
a<-param[1]
b<-param[2]
res<-n*log(b)+n*log(zeta(a))-a*sum(Freq*log(x))-sum(Freq*sapply(x,fn_aux,a=a,b=b))
                                              -sum(Freq*sapply(x+1,fn_aux,a=a,b=b))
return(res)
}

## Maximitzem el logaritme de la versemblança d'una MOEZipf
param.start<-c(2,2)
l_MOEZ<-optim(param.start, fn=fnMOEZ, x=x, Freq=Freq, method = c("L-BFGS-B"),
lower = c(1.00001,0.00001), upper = c(10,100), control = list(fnscale=-1))

# paràmetres estimats
a_MOEZ<-l_MOEZ$par[1]
b_MOEZ<-l_MOEZ$par[2]

######### Distribució MOEZipf (1r mètode d'estimació - Sitema d'equacions)
## funció que calcula l'esperança de les dades per un alpha donat i retorna
## la diferència entre aquesta i la mitjana mostral
fnMitjana<-function(a,x,Freq){
aux<-1; i<-1; suma<-0
f1<-Freq[1]
N<-sum(Freq)
xi_a<-zeta(a)
xi_a2<-zeta_x(a,2)

while(aux>0.0001){
aux<- i^(-a+1)/((xi_a-(((f1*xi_a-N)/(f1*xi_a2))*zeta_x(a,i)))*
                (xi_a-(((f1*xi_a-N)/(f1*xi_a2))*zeta_x(a,i+1))))
```

```
suma<-suma+aux
i<-i+1
}
return(abs(((N-f1)/(f1*zeta_x(a,2))*zeta(a)*suma)-sum(x*Freq)/sum(Freq)))
}

## Minimitzem la diferència que calcula la funció anterior, de manera que trobem
## una solució aproximada (el càlcul de l'esperança no és exacte) a l'equació
## plantejada en la definició del mètode.
res_SistEq<-optimize(fnMitjana,x=x,Freq=Freq, c(1.8, 5), tol=0.0001)

# paràmetres estimats
a_SistEq<-res_SistEq$minimum
b_SistEq<-(N-f1)/(f1*zeta_x(a_SistEq,2))

# versemblança obtinguda
ab_SistEq<-c(a_SistEq,b_SistEq)
l_SistEq<-fnMOEZ(ab_SistEq,x,Freq)

######### Gràfic dels ajustos
## Funció que calcula la probabilitat puntual d'una distribució MOEZipf(alpha,beta)
logY<-function(a,b,x){
log(b)-a*log(x)+log(zeta(a))
-log(zeta(a)-(1-b)*zeta_x(a,x))-log(zeta(a)-(1-b)*zeta_x(a,x+1))
}

## vector de probabilitats esperades segons la distribució MOEZipf (MLE)
valors<-rep(NA,max(x))
for(i in 1:max(x)){
valors[i]<-logY(a_MOEZ,b_MOEZ,i)
}

## Gràfic en eixos logarítmics que conté les probabilitats observades i les probabilitats
## esperades per la distribució Zipf i la MOEZipf, totes dues amb els paràmetres
## estimats per màxima versemblança
# options(digits=3)
jpeg('grafs//img10_DADES.jpg', width = 400, height = 400, units = "px")
plot(log(unique(x)),log(Prob), xlab="ln(x)",ylab="ln p(x)", main="TÍTOL GRÀFIC")
abline(a=-log(zeta(a_Z)), b=-a_Z)
lines(log(1:max(x)),valors,col=2)
legend("topright", legend=c(as.expression(bquote("Zipf(" ~ hat(alpha) ==.(a_Z)~ ")")),
  as.expression(bquote("MOEZipf(" ~ hat(alpha) ==.(a_MOEZ)~ ", " ~hat(beta)==.(b_MOEZ)~
")"))), col=1:2,lty=1)
dev.off()
```

B.3 Bondat de l'ajust

Per calcular la bondat dels diferents ajustos, s'ha creat una taula amb les freqüencies observades i esperades (per cada distribució ajustada), a partir d'aquesta s'ha agrupat la cua i s'ha creat una nova taula, que és la que apareix en aquest document i la que ens permet calcular el test χ^2 de bondat de l'ajust:

```
## Funció de probabilitat d'una distribució MOEZipf(alpha,beta)
PY<-function(a,b,x){
  x^(-a)*b*zeta(a)/((zeta(a)-(1-b)*zeta_x(a,x))*(zeta(a)-(1-b)*zeta_x(a,x+1)))
```

```
}

## Taula de l'ajust (la freqüència esperada de l'últim nivell es calcula fent
## la diferència del total d'observacions menys la suma de freqüències esperades
## per la resta de categories)
## X: distribució Zipf per màxima versemblança
## Y1: distribució MOEZipf pel 1r mètode
## Y2: distribució MOEZipf per màxima versemblança
taula<-matrix(c(rep(NA,max(x)*9)),ncol=9)

for(i in 1:(max(x)-1)){
auxX<-N*PY(a_Z,1,i)
auxY1<-N*PY(a_SistEq,b_SistEq,i)
auxY2<-N*PY(a_MOEZ,b_MOEZ,i)
taula[i,]<-c(i,Freq2[i],auxX,(Freq2[i]-auxX)^2/auxX,
           auxY1,(Freq2[i]-auxY1)^2/auxY1, auxY2,(Freq2[i]-auxY2)^2/auxY2)
}

auxXmaxx<-N-sum(taula[1:(max(x)-1),3])
auxY1maxx<-N-sum(taula[1:(max(x)-1),5])
auxY2maxx<-N-sum(taula[1:(max(x)-1),7])
taula[max(x),]<-c(max(x),Freq2[max(x)],auxXmaxx,
     (Freq2[max(x)]-auxXmaxx)^2/auxXmaxx,
      auxY1maxx, (Freq2[max(x)]-auxY1maxx)^2/auxY1maxx,
      auxY2maxx, (Freq2[max(x)]-auxY2maxx)^2/auxY2maxx)
colnames(taula)<-c("i", "o_i", "e_iX", "X2X", "e_iY_1", "X2Y_1", "e_iY_2", "X2Y_2")

## Taula de l'ajust amb les categories més elevades agrupades a partir de la categoria
## ``cua'' (la primera que té una freqüència observada inferior a 5)
cua<-which(taula[,2]<5)[1]
taula2<- matrix(c(rep(NA,cua*9)),ncol=9)
taula2[1:(cua-1),]<-taula[1:(cua-1),]
taula2[cua,]<-c(cua,sum(taula[cua:max(x),2]),sum(taula[cua:max(x),3]),
   (sum(taula[cua:max(x),2])-sum(taula[cua:max(x),3]))^2/sum(taula[cua:max(x),3]),
   sum(taula[cua:max(x),5]),
   (sum(taula[cua:max(x),2])-sum(taula[cua:max(x),5]))^2/sum(taula[cua:max(x),5]),
   sum(taula[cua:max(x),7]),
   (sum(taula[cua:max(x),2])-sum(taula[cua:max(x),7]))^2/sum(taula[cua:max(x),7]))
colnames(taula2)<-colnames(taula)

## Bondat de l'ajust - Test X^2 de Pearson
gdlX<-dim(taula2)[1]-1-1
gdlY<-dim(taula2)[1]-2-1

X2X<-sum(taula2[,4])
X2X; 1-pchisq(X2X,gdlX); 1-pnorm(X2X,gdlX,sqrt(2*gdlX))

X2Y1<-sum(taula2[,6])
X2Y1; 1-pchisq(X2Y1,gdlY); 1-pnorm(X2Y1,gdlY,sqrt(2*gdlY))

X2Y2<-sum(taula2[,8])
X2Y2; 1-pchisq(X2Y2,gdlY); 1-pnorm(X2Y2,gdlY,sqrt(2*gdlY))

## Bondat de l'ajust - AIC
AICX<-2*1-2*l_Z$objective
AICY1<-2*2-2*l_SistEq
```

```
AICY2<-2*2-2*l_MOEZ$value

## Significació de beta (Zipf vs. MOEZipf) - LRT
lambda<- -2*(l_Z$objective -l_MOEZ$value)
lambda; qchisq(0.95,1)
```

B.4 Test de Kolmogorov-Smirnov

A continuació s'exposa el codi utilitzat per a realitzar les simulacions que ens han permès trobar els valors crítics corresponents al test de KS per cada conjunt de dades.

```
setwd("D:/ACASELLAS/upcNOU/PROJECTE/Scripts")
source("zeta_x.R")

# Carreguem l'espai de treball del conjunt de dades que ens interessi: 1terrorisme.RData
# (atacs terroristes), 1terrorismeSense11S.RData (atacs terroristes sense 11-S),
# 1words.RData (paraules Moby Dick), 1email.RData (correu electònic), 1citesFan.RData
# (cites que apareixen en un article), 1citesReb.RData (vegades que se cita un article)
load("D:\\ACASELLAS\\upcNOU\\PROJECTE\\Scripts\\ajustos\\1terrorisme.RData")

m<-max(1000,max(x))  # mida de F
n<-N  # mida de les mostres
nsim<-1000  # nombre de simulacions
a<-a_MOEZ; b<-b_MOEZ # paràmetres estimats per la MOEZipf
a<-a_Z; b<-1 # paràmetres estimats per la Zipf

# generem el vector de les probabilitats teòriques acumulades
f<-c(rep(NA,m))
for(i in 1:m){
f[i]<-(b*zeta(a)*(i^(-a)))/
((zeta(a)-(1-b)*zeta_x(a,i))*(zeta(a)-(1-b)*zeta_x(a,i+1)))
}
F<-cumsum(f)

# generació d'estadístics KS a partir de mostres de distribucions MOEZipf(a,b)
K<-c(rep(NA,nsim)) # vector d'estadístics KS
for(j in 1:nsim){
U<-runif(n)
X<-sapply(U,function(x) which(F>=x)[1])
S<-cumsum(tabulate(X,nbins=m)/(n-sum(is.na(X))))
K[j]<-sqrt(n)*max(abs(F-S))
}
VC<-quantile(K,0.95) # valor crític obtingut

# valor de l'estadístic KS pel conjunt de dades triat
KS<-sqrt(n)*max(abs(F[1:min(m,length(Freq2))]-cumsum(Freq2[1:min(m,length(Freq2))]/N)))
```

Bibliografia

[1] G. K. Zipf, *Human behavior and the principle of least effort.* Addison-Wesley Press, 1949.

[2] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Review*, vol. 51, pp. 661–703, 2009.

[3] M. E. J. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemporary Physics*, vol. 46, pp. 323–351, 2006.

[4] J. Goldenberg and M. Levy, "Distance Is Not Dead: Social Interaction and Geographical Distance in the Internet Era," 2009.

[5] A. Clauset, M. Young, and K. S. Gleditsch, "On the Frequency of Severe Terrorist Events," *Journal of Conflict Resolution*, vol. 51, pp. 58–88, 2007.

[6] M. Xie, O. Gaudoin, and C. Bracquemond, "Redefining failure rate function for discrete distributions," *International Journal of Reliability, Quality and Safety Engineering*, vol. 9, pp. 275–286, 2002.

[7] M. L. Goldstein, S. A. Morris, and G. G. Yen, "Problems with Fitting to the Power-Law Distribution," *The European Physical Journal B - Condensed Matter and Complex Systems*, vol. 41, no. 2, pp. 255–258, 2004.

[8] J. Neyman and E. S. Pearson, "On the use and interpretation of certain test criteria for purposes of statistical inference," *Biometrika*, vol. 20A, pp. 263–294, 1928.

[9] N. L. Johnson, S. Kotz, and A. W. Kemp, *Univariate Discrete Distributions.* John Wiley & Sons, 1992.

[10] A. W. Marshall and I. Olkin, "A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families," *Biometrika*, vol. 84, pp. 641–652, 1997.

[11] M. E. Ghitany, E. K. Al-Hussaini, and R. A. Al-Jarallah, "Marshall-Olkin extended weibull distribution and its application to censored data," *Journal of Applied Statistics*, vol. 32, pp. 1025–1034, 2005.

[12] M. E. Ghitany, F. A. Al-Awadhi, and L. A. Alkhalfan, "Marshall-Olkin Extended Lomax Distribution and Its Application to Censored Data," *Communications in Statistics - Theory and Methods*, vol. 36, pp. 1855–1866, 2007.

[13] H.-C. Yeh, "The Generalized Marshall-Olkin Type Multivariate Pareto Distributions," *Communications in Statistics - Theory and Methods*, vol. 33, pp. 1053–1068, 2004.

[14] K. K. Jose, "Marshall-Olkin Family of Distributions and their applications in reliability theory, time series modeling and stress-strength analysis." http://isi2011.congressplanner.eu/pdfs/950092.pdf. 58th World Statistics Congress, International Statistical Institute (ISI).

[15] T. M. Apostol, *Análisis matemático*. Barcelona: Reverté, 2a ed., 1977.

[16] A. Clauset, "Power law distributions in empirical data." http://tuvalu.santafe.edu/~aaronc/powerlaws/data.htm. [Última visita 19-01-2013].

[17] Q. H. Vuong, "Likelihood ratio tests for model selection and non-nested hypotheses," *Econometrica*, vol. 57, pp. 307–333, 1989.

[18] R. Ferrer and R. V. Solé, "Zipf's Law and Random Texts," *Advances in Complex Systems*, vol. 5, pp. 1–6, 2002.

[19] J. Leskovec, "EU email communication network." http://snap.stanford.edu/data/email-EuAll.html. [Última visita 19-01-2013].

[20] J. Leskovec, *Dynamics of Large Networks*. PhD thesis, School of Computer Science - Carnegie Mellon University, 2006.

[21] H. Ebel, L.-I. Mielsch, , and S. Bornholdt, "Scale-free topology of e-mail networks," *Phys. Rev. E*, vol. 66, 2002.

[22] J. Leskovec, "High-energy physics citation network." http://snap.stanford.edu/data/cit-HepPh.html. [Última visita 19-01-2013].

[23] D. J. de Solla Price, "Networks of Scientific Papers," *Science*, pp. 510–515, 1965.

[24] S. Redner, "How Popular is Your Paper? An Empirical Study of the Citation Distribution," *European Physical Journal B 4*, vol. 149, pp. 131–134, 1998.

[25] Z. K. Silagadze, "Citations and the Zipf-Mandelbrot's law," *Complex Systems*, vol. 11, pp. 487–499, 1997.