Treball final de grau

GRAU DE MATEMÀTIQUES

Facultat de Matemàtiques i Informàtica
Universitat de Barcelona

Llei de Zipf i Aplicacions

Autor: Pan Ye

Director: Dr. Jose Fortiana Gregori
Realitzat a: Departament de Matemàtiques i Informàtica

Barcelona, 27 de juny de 2018

Abstract

Zipf's law, formulated by George Kingsley Zipf, states that the word frequency is inversely proportinal to its Zipf rank with an exponent slightly larger than the unit. This statement is not only valid in a linguistic context, but also serves for most frequency and rank phenomena. Because of this, Zipf's law is known as the distribution of frequencies and ranks. Then, the objective of this work is to introduce this law, explaning its origin and its different formulations, and develop one of its applications, the distribution of frequencies in texts.

Resum

La llei de Zipf, formulada per George Kingsley Zipf, afirma que la freqüència de paraules és inversament proporcional als seus rangs de Zipf amb un exponent lleugerament més gran que la unitat. Aquesta afirmació no solament és valida en un context lingüístic, sinó que també serveix per a la majoria dels fenòmens de freqüències i rangs. Degut això, la llei de Zipf és coneguda com la distribució de freqüències i rangs. Aleshores, l'objectiu d'aquest treball és introduir aquesta llei, explicant el seu origen i les seves diferents formulacions, i desenvolupar una de les seves aplicacions, la distribució de freqüències en textos.

Agraïments

Primer de tot, vull agrair al meu tutor, Jose Fortiana Gregori, pels seus suports, dedicacions i correccions del treball. També, volia agrair el professor Ramon Ferrer i Cancho pels seus coneixements en aquest camp i facilitar-me les referencies, i el meu company Edagar Harris Exposito per la seva correcció.

Índex

1 Introducció

1.1 Manuscrit Voynich

El ***Manuscrit Voynich*** és un llibre escrit per un autor anònim durant el segle XV segons proves del carboni 14, en un alfabet no indentificat i un idioma desconegut.

Des de la seva reaparició en l'any 1912, nombrosos criptògrafs professionals i aficionats van intentar desxifrar-lo, incloent-hi destacats especialistes en desxiframent de la Segona Guerra Mundial. Malauradament, van fracassar i, d'allí, van sortir moltes teories per explicar el manuscrit Voynich. Alguns deien que era només un engany, mentre que uns altres deien que era genuí. Fins que l'any 2001, l'especialista Gabriel Landini va trobar que el Manuscrit Voynich segueix la **Llei de Zipf**: hi ha una relació entre la freqüència i la longitud de les paraules, en el seu article "*Evidence of linguistic structure in the Voynich Manuscript using spectral analysis*" (cfr. Reddy i Knight, 2011). Tretze anys després, Stephen Bax (2014), professor de Lingüística Aplicada de la universitat de Bedfordshire, en el Regne Unit va aconseguir per primera vegada descodificar parcialment alguns segments del mateix, deu paraules sobre un total de 37.919.

La llei de Zipf és una llei empírica formulada utilitzant mètodes estadístics i es refereix al fet que molts fenòmens físics i socials es poden aproximar amb una distribució zipfiana, una de les famílies de distribucions discretes del tipus potencial. La llei porta el nom del lingüista i filòleg estatunidenc George Kingsley Zipf (1902-1950), qui la va popularitzar i va intentar explicar-la. Va formalitzar les seves observacions per primera vegada l'any 1935 en el seu treball: "*The psycho-biology of language*". En aquesta monografia afirma que un cop ordenat les paraules de més freqüents a menys freqüents, aleshores la freqüència d'una paraula de la posició n es pot expressar com $f_n \approx 1/n^{\alpha}$, on α s'anomena l'exponent de Zipf i és lleugerament més gran que 1 per la distribució de les freqüències. D'una altra manera, la paraula més freqüent d'un text apareix el doble de vegades que la següent més freqüent, triple que la tercera més freqüent, quatre vegades més que la quarta, i així successivament.

L'expressió de la llei de Zipf té una forma potencial que pertany a la familia de la llei potencial. La llei potencial té una aplicació molt amplia, per posar alguns exemples, la intensitat de les erupcions solars i la intensitat de les guerres. Aquí, la intensitat de les guerres es defineix com el nombre de morts de tots els països participants en una guerra dividits per la població total d'aquests països i multiplicant-se per 1000.

L'ús de la llei de Zipf no es va aturar en el camp lingüístic, sinó que va estar en el progrés. L'any 1949, Zipf va observar patrons per a la freqüència d'ús de paraules en diversos idiomes i aquests patrons van ser coneguts com la relació de la freqüència i la seva posició com l'esmentada anteriorment. Els mateixos patrons també van aparèixer en la mida de les ciutats. Avui en dia, la llei de Zipf s'utilitza en internet, distribució d'ingressos d'empreses, contrasenyes textuals, etc. Malgrat això, volia enfocar el seu ús en l'ambit lingüístic, tant llengües humanes com no humanes. Un

estudi de la comunicació animal és un bon exemple de la llengues no humanes. D'una primera vista, no sembla que és tasca senzilla, però els experts lingüístics estadístics van trobar un ús apropiat en aquest camp. Per exemple, l'estudi de les seqüències de xiulades de dofins (Ferrer-i-Cancho&McCowan, 2012).

Finalment, l'objectiu principal d'aquest treball és introduir la llei de Zipf explicant què és i d'on procedeix. Després estudiarem una de les seves aplicacions: la distribució de freqüències de paraules. D'acord amb aquestes idees, en la continuació es detalla l'estructura de la memòria.

1.2 Estructura de la Memòria

Aquest treball consta de sis capítols.

1 El primer capítol, la introducció, consisteix en tres parts: la motivació del treball a través del Manuscrit Voynichhi, l'estructura de la memòria i, finalment, algunes abreviacions que farem servir durant tot treball.

2 Un cop introduït el treball, en el capítol 2 farem una recerca de la llei de Zipf estudiant algunes distribucions estretament relacionades per entendre que és la llei de Zipf.

3 En el capítol 3 introduïm la formació empírica de la llei i, a més, alguns models teòrics que es poden deduir la llei de Zipf. A més a més, explicarem amb detall sobre el model de *Random Typing* i la seva deficiència.

4 En el capítol 4 estudiem la distribució de freqüències de paraules en els textos. Per tant, introduïm el concepte de l'espectre de freqüències i els models LNRE. De manera que mostrem almenys una aplicació de la llei de Zipf concreta.

5 En el capítol següent farem una anàlisi del Corpus Brown que està basada en l'estudi de les distribucions de freqüències de paraules en textos com un exemple pràctic del capítol anterior.

6 En l'últim capítol farem una petita revisió del treball on concloure el tema que tractem.

1.3 Notacions:

· cdf (*cumulative density function*): funció de distribució de probabilitat

· pmf (*prbability mass function*): funció de massa de probabilitat

· pdf (*probability density function*): funció de densitat de probabilitat

· $\mathbb{E}[X]$: esperança d'una variable aleatòria X

· $\mu_k = \mathbb{E}[X^k]$: moment d'ordre k d'una variable aleatoria

2 Distribucions

Considerem primer dues preguntas: (1) què és la llei de Zipf i (2) d'on procedeix? Amb aquestes dues qüestions anem a introduir formalmente la llei de Zipf i estudiar algunes lleis de probabilitat relaciondes amb ella. Primer de tot, expliquem el cas general, la llei potencial, i algunes propietats importants d'ella.

2.1 Llei potencial

Una variable aleatòria contínua amb llei potencial té una pdf de la forma

$$f(x) = Cx^{-\alpha} \quad C \in \mathbb{R}, \quad \alpha > 0, \tag{2.1}$$

per x d'un interval $[a, b]$ contigut a $\mathbb{R}_+$. Observem que $f \uparrow +\infty$ per $x \to 0$, o sigui que per modelar dades reals 'a' haurà de ser estrictament positiu. També convindrà de vegades considerar lleis de probabilitat que tenen la cua com (2.1). Sovint, farem servir una llei d'aquesta família per modelar variables que prenen valors més gran que un x_{min}. D'aquesta manera podrem reescriure la pdf com:

$$f(x) = \begin{cases} 0 & \text{si } x < x_{min} \\ Cx^{-\alpha} & \text{si } x \geq x_{min} \end{cases} \tag{2.2}$$

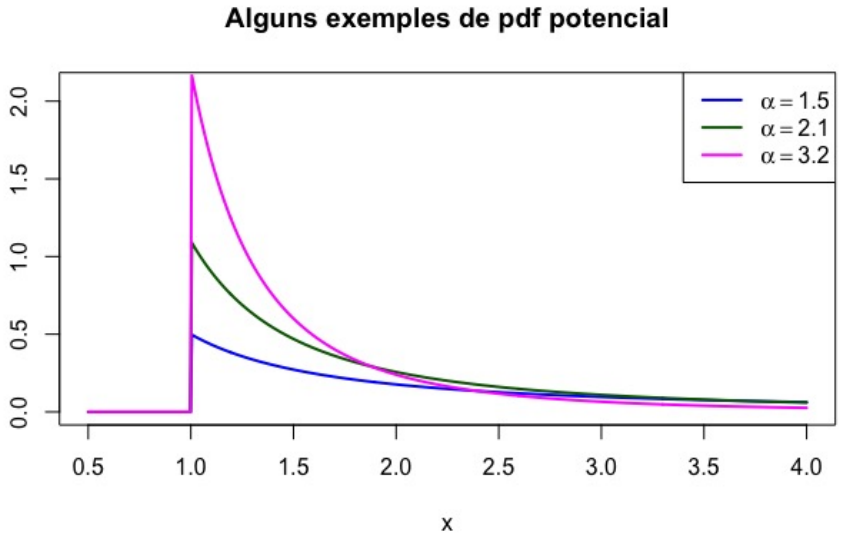


Figura 1: Funciones del tipus (2.2) amb exponents indicats.

Aquestes distribucions poden observar-se en una gran varietat de dades físiques, biològiques i fenòmens artificials. Per exemple: les freqüències de paraules en la majoria de llengües, el nombre de visitants d'un servidor durant un cert temps, el nombre de còpies de llibres venuts, la magnitud dels terratrèmols, la intensitat de les erupcions solars, la població de ciutats, etc. (Newman, 2005).

Invariància escalar Un dels atributs de la llei potencial és la invariància escalar. És a dir, la forma fucional (2.1) és invariant per canvi d'escala:

$$x \to kx \quad f(kx) = C(kx)^{-\alpha} = C_1 x^{-\alpha},$$

torna a ser de la mateixa família. Aquest comportament produeix una relació lineal en l'escala logarítmica en x i $f(x)$:

$$\log f(x) = \log Cx^{\alpha} = \alpha \log x + \log C,$$

independentment de l'escala de x. Així, podem estimar l'exponent α.

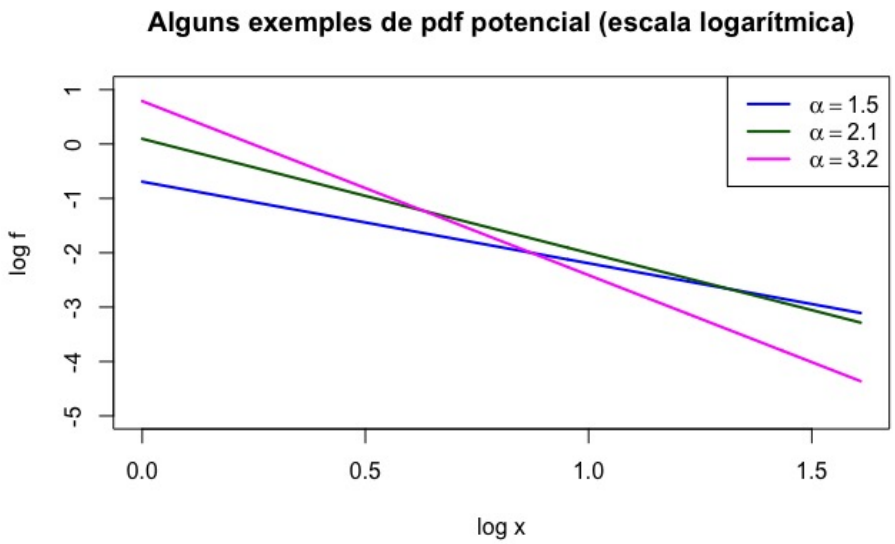


Figura 2: Representació de de la figura (1) en escala logarítmica.

Per aquest motiu, la distribució de la llei potencial també s'anomena, en anglès, *distribution scale free*. De fet, és l'única distribució que compleix la propietat d'invariància escalar.

Normalització Ens agradaría determinar la constant C i aquest valor ve donat per la condició de la pdf

$$1 = \int_{x_{min}}^{\infty} f(x)dx = C \int_{x_{min}}^{\infty} x^{\alpha} dx = \frac{C}{1-\alpha} \left[x^{-\alpha+1}\right]\Big|_{x_{min}}^{\infty}.$$

Observem que aquesta equació només té sentit si $\alpha > 1$. Una llei potencial amb exponent més petit que la unitat no pot ser normalitzada. A partir d'ara, sempre considerem $\alpha > 1$. Aleshores l'equació ens dóna

$$C = (\alpha - 1)x_{min}^{\alpha-1}$$

Sustituïm el valor de C a la funció (2.1) tenim $\forall x \geq x_{min}$

$$f(x) = \frac{\alpha - 1}{x_{min}} \left(\frac{x}{x_{min}} \right)^{-\alpha}. \tag{2.3}$$

Per modelar fenòmens reals, no totes les distribucions segueixen la llei potencial a partir de x_{min}, sinó que algunes la segueixen només en part del seu rang fins un x_{max}.

Moments Una altra propietat d'aquesta llei és la manca de mitjana per a alguns exponents. L'expresió de l'esperança de X ve donada per

$$\mathbb{E}[X] = \int_{x_{min}}^{\infty} x f(x) dx = C \int_{x_{min}}^{\infty} x^{-\alpha+1} dx = \frac{C}{2-\alpha} \left[x^{-\alpha+2} \right] \Big|_{x_{min}}^{\infty}. \tag{2.4}$$

Observem que aquesta expressió tendeix a infinit si $\alpha \leq 2$. Per aquests α, X no té una esperança finita. Alguns exemples de aquestes lleis potencial són la distribució de la intensitat d'erupcions solars amb l'exponent ≈ 1.83 i de la intensitat de guerres amb l'exponent ≈ 1.80 en la taula 1 de Newman (2005).

Per $\alpha > 2$, l'esperança no és divergent, en aquest cas, podem calcular el valor utilitzant la fórmula (2.3)

$$\mathbb{E}[X] = \frac{C}{\alpha - 2} x_{min}^{2-\alpha} = \frac{\alpha - 1}{\alpha - 2} x_{min},$$

on la segona igualtat substituïm C per $(\alpha - 1) x_{min}^{\alpha-1}$. I els moments d'ordre k venen donats per l'expressió següent:

$$\mu_k = \int_{x_{min}}^{\infty} x^k f(x) dx = C \int_{x_{min}}^{\infty} x^{k-\alpha} dx = \frac{C}{k - \alpha + 1} \, x^{k-\alpha+1} \Big|_{x_{min}}^{\infty} \tag{2.5}$$

Ens fixem el cas $k = 2$, el moment d'ordre 2 divergeix si $\alpha \leq 3$, llavors la distribució de la llei potencial no té variància finita. Per $\alpha > 3$, el moment de segon ordre és finit i té el valor

$$\mu_2 = \mathbb{E}[X^2] = \frac{\alpha - 1}{\alpha - 3} x_{min}^2.$$

Podrem generalitzar el resultat pels moments d'ordre k. Els moments d'ordre k existeix si $k < \alpha - 1$ i és

$$\mu_k = \frac{\alpha - 1}{\alpha - 1 - k} x_{min}^k$$

I els moments d'ordres més grans divergeixen.

Màxim Denotem per $Pot(\alpha, x_{min})$ la llei potencial amb l'exponent α i un mínim valor x_{min}, i $F(x) = F(x; \alpha, x_{min})$ i $f(x) = f(x; \alpha, x_{min})$ la cdf i la pdf corresponent d'aquesta llei respectivament.

Ara bé, siguin $X_1, \cdots, X_n$ variables aleatòries iid (independents i idènticament distribuïdes) amb llei $Pot(\alpha, x_{min})$, f la pdf comú de $X_1, \cdots, X_n$ que té la forma (2.3) i F la cdf corresponent que té l'expressió següent:

$$F(x) = 1 - \int_x^\infty f(t)dt = 1 - \frac{C}{\alpha - 1} x^{\alpha-1} = 1 - \left(\frac{x}{x_{min}}\right)^{\alpha-1} \tag{2.6}$$

Definim $M_n = \max\{X_1, \cdots, X_n\}$. Sigui G_n la cdf de M_n i es calcula: per $t \in \mathbb{R}$

$$G_n(t) = P\{M_m \leq t\} = P\{X_1 \leq t, \cdots, X_n \leq t\}$$

$$= P\{X_1 \leq t\} \cdot P\{X_2 \leq t\} \cdots P\{X_n \leq t\} = (F(t))^n.$$

Aleshores $g_n(t)$, la pdf de M_n, és

$$g_n(t) = (G_n(t))' = n \cdot F(t)^{n-1} \cdot f(t).$$

Ara podrem calcular l'esperança de M_n:

$$\mathbb{E}[M_n] = \int_{x_{min}}^\infty x g_n(x)dx = n \int_{x_{min}}^\infty x f(x) F(x)^{n-1} dx$$

Utilitzant les fórmula (2.2) i (2.5), obtenim

$$\mathbb{E}[M_n] = n \int_{x_{min}}^\infty \frac{\alpha - 1}{x_{min}} \left(\frac{x}{x_{min}}\right)^{-\alpha} \left[1 - \left(\frac{x}{x_{min}}\right)^{-\alpha+1}\right]^{n-1} dx =$$

$$n(\alpha - 1) \cdot \int_{x_{min}}^\infty \left(\frac{x}{x_{min}}\right)^{-\alpha+1} \left[1 - \left(\frac{x}{x_{min}}\right)^{-\alpha+1}\right]^{n-1} dx =$$

$$n x_{min} \int_0^1 \frac{y^{n-1}}{(1-y)^{1/(\alpha-1)}} dy = n x_{min} B(n, (\alpha-2)/(\alpha-1)),$$

on hem utilitzant el canvi de variable $y = 1 - (x/x_{min})^{-\alpha+1}$ i la funció Beta $B(a, b)$ definida com

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \quad \text{on} \quad \Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx.$$

Farem servir l'aproximació asintòtica de la funció Beta. Donat un valor de a suficientment gran i un valor de b fixat, utilitzant l'aproximació de Stirling

$$\lim_{n \to +\infty} \frac{n!}{\sqrt{2\pi n}(n/e)^n} = 1$$

podem veure $B(a, b) \sim a^{-b}$. En molts casos, la mida n de la mostra per a les nostres distribucions és gran, així que

$$B(n, (\alpha - 2)/\alpha - 1) \sim n^{-(\alpha-2)/(\alpha-1)} \quad \text{i}$$

$$\mathbb{E}[X_{max}] \sim n^{1/(\alpha-1)}. \tag{2.7}$$

Aquesta aproximació ens permet completar els calculs del moments. Pel cas $2 < \alpha \leq 3$, el qual cobreix la majoria de les distribucions de la llei potencial observades a la vida real, veiem que divergeix el moment de segona odre en l'equació (3.4). En canvi, en la realitat totes les dades són finites i, en conseqüència, existeix una observació màxima x_{max}. Aleshores, utilitzant la formula (3.6) tenim

$$\mu_2 = \frac{C}{3-\alpha}\left[x^{-\alpha+3}\right]\Big|_{x_{min}}^{x_{max}} \sim n^{(3-\alpha)/(\alpha-1)}.$$

Si $\alpha = 5/2$, llavors la mitjana quadràtica de la mostra, i també la variancia, creix com $n^{1/3}$ quan la mida de la mostra augmenta.

La presència de valor màxim ens permet modelar les dades reals d'una forma més ajustada de manera que podem trobar patrons com esperança i variança.

Estimació del paràmetre Per acabar aquest tema, considerem l'estimació del paràmetre de l'exponent de la llei potencial d'unes dades reals.

Una de les propietats de la llei potencial és la dependencia lineal entre $\log x$ i $\log f(x)$, on el pendent de la recta és justament l'exponent de la llei. Aleshores, per detectar si unes dades observades segueixen o no a la distribució només cal veure la dependència en l'escala logarítmica. Un cop detectat, podem estimar directament el pendente de la recta que, com justifica Newman (2005), resulta ser un estimació amb biax.

Una altre manera d'estimar l'exponent és utilitzant l'estimació de màxima versemblança. Sigui $X = (x_1, \cdots, x_n)$ un vector aleatori amb els components independents i idènticamentdistribuïts seguint llei potencial de la forma (2.2). Tenim que:

$$f(x, \alpha) = \frac{\alpha - 1}{x_{min}}\left(\frac{x}{x_{min}}\right)^{-\alpha},$$

pels $x \geq x_{min}$. Aleshores la funció de versemblança, $\mathcal{L}(x_1, \cdots, x_n, \alpha)$, és

$$\mathcal{L}(x_1, \cdots, x_n, \alpha) = \prod_{i=1}^{n} f(x_i, \alpha) = \prod_{i=1}^{n} \frac{\alpha - 1}{x_{min}}\left(\frac{x_i}{x_{min}}\right)^{-\alpha}.$$

En la pràctica, és natural considerar la funció Log-versemblança

$$l(x) = \log \mathcal{L}(x_1, \cdots, x_n, \alpha) = \sum_{1}^{n}\left[\log(\alpha - 1) - \log x_{min} - \alpha \log \frac{x_i}{x_{min}}\right]$$

$$= n\log(\alpha - 1) - n\log(x_{min}) - \alpha \sum_{i=1}^{n} \log \frac{x_i}{x_{min}},$$

on log és logaritme natural. El métode consisteix en trobar el valor de α que faci la máxima $l(x)$. Derivant:

$$l'(\alpha) = \frac{n}{\alpha - 1} - \sum_{i=1}^{n} \log \frac{x_i}{x_{min}} = 0 \Rightarrow \hat{\alpha} = 1 + n\left[\sum_{i=1}^{n} \log \frac{x_i}{x_{min}}\right]^{-1}.$$

El màxim s'assoleix en $\hat{\alpha}$, on la segona derivada

$$l''(\alpha) = \frac{-n}{(\alpha-1)^2} < 0, \quad \forall \alpha > 1.$$

Aquest mètode ens dóna una estimació asimptòticament sense biax, però que en una mostra finita produeix una biax d'ordre $O(n^{-1})$. Malgrat això, si $n > 100$ el biax serà petit. A més a més, l'error estàndar d'aquest estimador és:

$$\sigma = \frac{\hat{\alpha}-1}{\sqrt{n}} + O(n^{-1}).$$

Relació amb la llei de Zipf La llei de Zipf és un cas particular de la llei potencial, un cas discret. A vegades, una variable aleatòria que segueix una llei potencial es diu que segueix la llei de Zipf o la distribució de Pareto. En aquest sentit, "llei de Zipf" i "llei de Pareto" són sinònims de llei potencial. En molts llocs, la distribució de Pareto es refereix com una versió contínua de la llei de Zipf, i la llei de Zipf com la versió discreta de Pareto.

2.2 Distribució de Pareto

Vilfredo Pareto va ser un enginyer, economista, sociòleg i filòsof italià. Al final del segle XIX, ell va observar el nombre d'individus amb els ingressos superiors a un cert nivell x es podia aproximar per la llei potencial amb una cert constant i un cert exponent. Posterioment es va descobrir que aquesta aproximació només era valida per a grans valors de x. En l'actualitat la distribució de Pareto s'utilitza en la descripció de fenòmens social, geofísics, actuarials i molts altres tipus d'observacions.

Distribució de Pareto clàssica i variants Donada una variable aleatòria X absolutament contínua, la funció de supervivència associada a X és

$$\bar{F}(x) = Pr(X > x) = \int_x^\infty f(x)dx = 1 - F(x).$$

on $f(x)$ i $F(x)$ són pdf i cdf de X respectivament.

En general, es defineix la distribució de Pareto clàssica (o Tipus I), utilitzant la funció de supervivència de la variable X, de la forma següent:

$$S_1(x) = Pr(X > x) = \begin{cases} \left(\frac{x}{\sigma}\right)^{-\alpha} & x \geq \sigma \\ 1 & x < \sigma \end{cases} \tag{2.8}$$

on σ és el possible mínim valor positiu de X que vam esmentar anteriorment. Aquesta distribució està caracteritzada per dos paràmetres σ i α i s'utilitza per modelar la distribució de riquesa. En aquest cas, el paràmetre α també s'anomena l'índex de Pareto. Es denota $X \sim Par(\sigma, \alpha)$.

En final del segle XIX, Pareto va suggerir tres variants de la seva distribució. La primera és la distribució de Pareto clàssica o tipus I en l'any 1895. La segona, en l'any 1896, va involucrar un nou paràmetre de localització a la funció de supervivència:

$$\hat{S}_2(x) = \left[1 + \frac{x-\mu}{\sigma}\right]^{-\alpha} \quad x \geq \mu.$$

Un cas especial és quan $\mu = 0$,

$$S_2(x) = \left[1 + \frac{x}{\sigma}\right]^{-\alpha} \quad x \geq 0.$$

sovint es coneix com la distribució Pareto tipus II, segons Kleiber&Kotz (2003). Aquesta distribució va ser redescoberta per Lomax uns 50 anys més tard en un context diferent. Per aquest motiu, també es coneguda com la distribució de K.S. Lomax. Observem que existeix una relació simple entre els ambdos tipus de la distribució de Pareto.

$$S_2(x) = \left[1 + \frac{x}{\sigma}\right]^{-\alpha} = \left[\frac{x+\sigma}{\sigma}\right]^{-\alpha} \quad x \geq 0 \Leftrightarrow S_1(y) = \left[\frac{y}{\sigma}\right]^{-\alpha} \quad y \geq \sigma,$$

amb $y = x + \sigma$. Aleshores

$$X \sim Par(II)(\sigma, \alpha) \Leftrightarrow X + \sigma \sim Par(\sigma, \alpha).$$

La tercera distribució proposada per Pareto en l'any 1897 és del tipus III que té la funcó de supervivència de la forma:

$$S_3(x) = \left[1 + \left(\frac{x-\mu}{\sigma}\right)^{1/\gamma}\right]^{-1},$$

on γ es un valor positiu. A més existeix la distribució de Pareto tipus IV, encara no sabem qui va proposar aquest model, que és una generalització del tipus III i té la forma

$$S_4(x) = \left[1 + \left(\frac{x-\mu}{\sigma}\right)^{1/\gamma}\right]^{-\alpha}.$$

Les distribucions del tipus I, II i III són casos particulars del tipus IV i presenten les següents relacions:

$$\begin{gathered} P(IV)(\sigma, \sigma, 1, \alpha) = P(I)(\sigma, \alpha) \\ P(IV)(\mu, \sigma, 1, \alpha) = P(II)(\mu, \sigma, \alpha) \\ P(IV)(\mu, \sigma, \gamma, 1) = P(III)(\mu, \sigma, \gamma). \end{gathered}$$

L'estudi de la distribució de Pareto no es va aturar aquí Feller (1971) va definir la variable Pareto per la transformació $U = Y^{-1} - 1$ d'una variable aleatòria beta Y, la seva pdf és

$$f(y) = \frac{y^{a-1}(1-y)^{b-1}}{B(a,b)}, \quad 0 < y < 1,\ a, b > 0,$$

on B(a,b) és la funció beta. Si

$$W = \mu + \sigma U^{\gamma}, \quad \sigma > 0, \gamma > 0,$$

aleshores W té una distribució de Feller-Pareto que es denota per $FP(\mu, \sigma, \gamma, a, b)$. Els casos especial d'aquesta distribució són

$$\begin{gathered} FP(\sigma, \sigma, 1, 1, \alpha) = P(I)(\sigma, \alpha) \\ FP(\mu, \sigma, 1, 1, \alpha) = P(II)(\mu, \sigma, \alpha) \\ FP(\mu, \sigma, \gamma, 1, 1) = P(III)(\mu, \sigma, \gamma) \\ FP(\mu, \sigma, \gamma, 1, \alpha) = P(IV)(\mu, \sigma, \gamma, \alpha). \end{gathered}$$

Principi de Pareto i les seves aplicacions En l'any 1896, Pareto va publicar el seu primer paper *Cours d'économie politique*, quan estava en la universitat de Lausanne. En aquest treball, ell va demostrar que el 80% de la terra a Italià era propietat del 20% de la població. Posteriorment, el pensador en temes de gestió de negocis Joseph M. Juan va suggerir el principi i el va anomenar en honor a Pareto. Fins ara, aquest principi es conegut com la regla 80/20 o el principi de Pareto. És una regla empírica habitual per als negocis i que també s'utilitza en la ciencia, el sofrware i l'altre camp. A més a més, té una demostració matemàtica donada a continuació utilitzant propietats de la llei potencial.

Per qualsevol llei potencial amb un exponent $\alpha > 1$, vam demostrar que l'esperança està ben definida. Aleshores, existeix un punt $x_{1/2}$ tal que passa això:

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

on $f(x) = Cx^{-\alpha}$. Com que aquesta integral està ben definida, obtenim:

$$\begin{aligned} \frac{C}{\alpha - 1}x_{1/2}^{1-\alpha} = \frac{C}{2(\alpha-1)}x_{min}^{1-\alpha} \quad &\Rightarrow \quad \left(\frac{X_{1/2}}{x_{min}}\right)^{1-\alpha} = \frac{1}{2} \\ \Rightarrow \quad x_{1/2} = 2^{1/(\alpha-1)}x_{min}. \end{aligned} \tag{2.9}$$

Així que, per exemple, en el cas de la distribució de la riquesa, segons les dades recogides per la revista *Forbes* en l'any 2003 sobre la riquesa total de les persones més riques dels Estats Units, l'exponent α estimada és 2.09 (taula 1 de Newman (2005)). Ara bé, com que l'exponent és més gran que la unitat, llavors la integral $\int_y^{\infty} xf(x)dx$ està ben defnida per $\forall y \geq x_{min}$. Considem el quocient següent:

$$\frac{\int_{x_{1/2}}^{\infty} xf(x)dx}{\int_{x_{min}}^{\infty} xf(x)dx} = \left(\frac{x_{1/2}}{x_{min}}\right)^{2-\alpha} = 2^{-(\alpha-2)(\alpha-1)}, \tag{2.10}$$

tenim en compta que hem utilitzat la formula (3.8). Aquest quocient és justament la fracció de la riquessa total de la meitat més rica. Sustituïm el valor de $\alpha = 2.09$ a l'expressió (3.9) obtenim és $2^{-0.083} \simeq 94\%$ de la riquesa està en mans d'aquestes persones.

Més general, la fracció de la població amb una riquesa personal superior a x utilitzant (2.6) és

$$\bar{F}(x) = \int_x^\infty f(y)dy = (x/x_{min})^{-\alpha+1} \tag{2.11}$$

i la fracció de la riquesa total en aquestes persones és

$$W(x) = \frac{\int_x^\infty yf(y)dy}{\int_{x_{min}}^\infty yf(y)dy} = \left(\frac{x}{x_{min}}\right)^{-\alpha+2}, \tag{2.12}$$

assumint de nou que $\alpha > 2$. Eliminant x/x_{min} entre l'equation (2.11) i (2.12), trobem que la fracció W de la riquesa en mans del F més ric de la població és

$$W(\bar{F}) = \bar{F}^{(\alpha-2)(\alpha-1)}, \quad \bar{F} \in [0,1], \tag{2.13}$$

on l'equation (2.10) és un cas especial. Això torna a tenir una forma de llei potencial amb un exponent positiu.

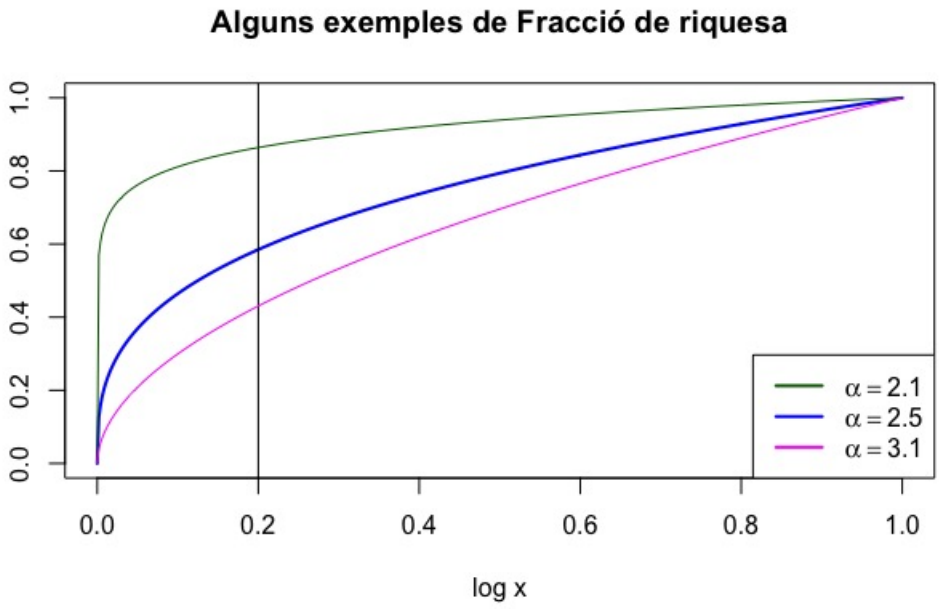


Figura 3: Algunes exemples de (2.13).

Utilitzant l'exponent $\alpha = 2.09$ (Newman, 2005), podrem calcular la fracció de riquesa de les persones més riques, 20% que repreenta de la població:

$$W(x_{1/5}) = \left(\frac{1}{5}\right)^{(\alpha-2)/(\alpha-1)} \simeq 88\%$$

Tenim en compta que el valor estimat de α depèn de les dates observades. En canvi, si fixem la distribució del PIB mundial de l'any 1989, els més rics (20% de la població) tenian un pes 82.70% de tots ingressos com indicada en la taula següent:

TABLE 3.2
Global economy, 1989
Percentage of total

	Income	Trade	Domestic investment	Domestic Savings
Poorest 20%	1.40	0.95	1.25	0.98
Second 20%	1.85	1.35	2.62	2.53
Third 20%	2.30	2.53	2.92	2.59
Fourth 20%	11.75	13.94	12.65	13.39
Richest 20%	82.70	81.23	80.56	80.51

Taula 1: Dades recollides pel Programa de les Nacions Unides per al Desenvolupament i publicades en *Human Development Report 1992*.

La idea principial d'aquest principi és afirmar que una minoria de causes, ingressos o esforços generalment condueixen a la majoria del resultats, producions o recompenses. Com les seves aplicacions, per exemples, l'autor Richard Koch, en el seu llibre *The 80/20 Rule*, va il·lustar algunes aplicacions pràctiques en la gestió empresarial i vida. En la ciencia de computació, el pricipi es pot aplicar als esforços d'optimització. Com a exemple, Microsoft va tenir la cuenta que si es solucionava el 20% dels bugs més reportats, s'eliminaran el 80% del errors relacionats i fallades en un sistema determinat.

2.3 Distribució Zeta

En la teoria de probabilitat i estadística, la distribució zeta és una distribució de probabilitat discreta definida sobre els nombres naturals amb la pmf

$$P_s(X = k) = \frac{k^{-s}}{\zeta(s)} \quad k \in \mathbb{N}, \quad \zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s},$$

on $\zeta : \mathbb{C} \to R$ és la funció zeta de Riemann i observem que ζ no està ben definida en $s = 1$. Aquesta funció juga un paper important en la teoria de nombres, especialment en la distribució dels nombres primers. Encara més, la seva funció de distribució de probabilitat ve donada per l'expressió:

$$P_s(X \leq k) = \frac{H_{k,s}}{\zeta(s)}, \quad H_{k,s} = \sum_{i=1}^{k} \frac{1}{i^s},$$

on $H_{k,s}$ és el k-èsim nombre harmònic generalitzat.

En moltes ocasions, la distribució de zeta es coneguda com a sinònim de la llei de Zipf.

3 Llei de Zipf

Un cop introduït les distribucions relacionades amb la llei de Zipf, és el moment de contestar la segona pregunta que vam plantejar en el capítol anterior: d'on procedeix?

La llei de Zipf és una famosa llei empírica formulada per Zipf en un context llingüístic. Aquesta llei afirma que donat un corpus de llenguatge, la freqüència de qualsevol paraula és inversamente proporcional al seu rang en la taula de freqüències. Quasi la majoria de la llengua natural segueix aquesta llei, a més a més, existeix almenys una no natural com Esperanto (una llengua auxiliar planificada per l'oftalmòleg polonès Ludwik Lejzer Zamenhof). En particular, la llei de Zipf és un cas discret de la llei potencial.

Abans de formular la llei de Zipf és covenient introduir algunes magnituds. Siguin

- w: una paraula qualsevol;
- N: nombre total de les paraules;
- $f = f(\omega)$: la freqüència de la paraula ω (el nombre d'ocurrències de la paraula ω);
- $F = F(f)$: el nombre de paraules ω_i tal que la seva freqüència és igual a f;
- $r = r(\omega)$: rank de la paraula ω.

Després d'ordenar la llista de paraules de la forma descendent respecte a les freqüències, podrem definir el *rank* $r(\omega)$ com la posició de la paraula ω en la llista ordenada. Finalment, suposem que cada paraula té un rank diferent.

Per la definició anterior, la freqüència $f(\omega)$ és una funció monòtona decreixent respecte el rank $r(\omega)$. Però, en cap moment ens diu què tipus de funció decreixent és. Llavors el Zipf va estudiar la correspondència específica entre aquestes quantitats en les seves monografies on va proposar una forma després d'observar molts textos llargs en anglès:

$$f(w) \propto \frac{1}{r(w)}. \tag{3.1}$$

Més concret, la freqüència de la paraula w en la posició n és $Cn^{-\alpha}$, amb $\alpha > 0$ i C constant.

3.1 Formulació empírica de la llei

Siguin:

- $N \in \mathbb{N}$ és el nombre total d'elements en una població;
- $k = 1, \cdots, N$, ranks corresponents;

· $\alpha > 1$, l'exponent de Zipf.

Segon el valor de N podem diferenciar dos diferents casos: cas general (quan $N = \infty$) i cas particular ($N < \infty$). Llavors la formulació empírica de la llei pel cas general és considerar la relació següent:

$$p_k = Ck^{-\alpha}, \quad C \in \mathbb{R}, \tag{3.2}$$

on p_k és la probabilitat dels elements amb el rank de Zipf k. De manera semblant a la llei potencial, podem suposar que k sigui estrictament positiu per evitar el problema. Aleshores, la constant C es donaria per la condició de normalització

$$1 = \sum_{k=1}^{\infty} p_k = C \sum_{i=1}^{\infty} k^{-\alpha} = C\zeta(\alpha) \Rightarrow C = \zeta(\alpha)^{-1}$$

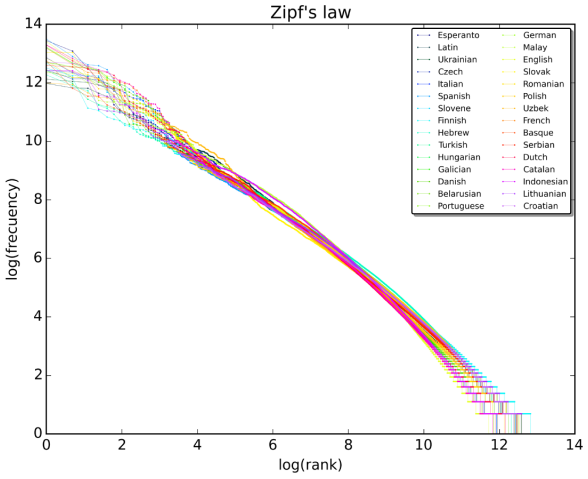
on $\zeta(\alpha)$ és la funció zeta de Riemman que esmentada anteriorment. Com que la llei de Zipf és un cas particular de la llei potencial, així doncs, existeix un k_{min} i podem obtenir una expressió equivalent: $\forall k \geq k_{min}$,

$$p_k = \frac{k^{-\alpha}}{\zeta(\alpha, k_{min})}, \quad \zeta(\alpha, k_{min}) = \sum_{n=k_{min}}^{\infty} n^{-\alpha}. \tag{3.3}$$

En la vida real, o sigui en les mostres observades, el nombre total N sempre és finit. Aleshores, es tracta del cas particular. En aquest cas, la probabilitat de la freqüència d'elements en el rank k d'una població de N elements, $f(k; \alpha, N)$, utilitzant la formulació general serà:

$$f(k; \alpha, N) = \frac{k^{-\alpha}}{\sum_{n=1}^{N} n^{-\alpha}}. \tag{3.4}$$

Les dades de la figura següent van ser extretes per Sergio Jimenez i provenen de les primeres 10 milions de paraules en 30 vikipedies. Representa la relació entre log(frecuency) i log(rank) de 30 llengües. Es veu que totes aquestes llengües de la figura segueixen a la fórmula (3.1), una recta enuna escala Log-Log, i té una pendent aproximada a $\alpha = 1$.



La llei de Zipf no només s'observa en la llingüística, sinó que també en l'ecologia, la sociologia, l'economia i la física. Es va trobar posterioment un millor estimació d'exponent en l'estudi de la relació entre freqüència i rank de les ciutats, $\alpha = 1.07$.

llei de Benford La llei de Benford, també anomenada llei de primer dígit, és una distribució de probabilitat que descriu la distribució de les freqüències dels dígits de la majoria dels conjunts de les dades extretes de la vida real. Es diu que un conjunt de xifres satisfà la llei de Benford si el dígit principal n ($n \in \{1, \cdots, 9\}$) es produeix amb una probabilitat

$$P(n) = \log(\frac{n+1}{n}),$$

on log refereix a $\log_{10}$. A més Pietronero i uns altres científics van descriure aquesta probabilitat com la integral següent:

$$P(n) = \int_n^{n+1} N^{-1} dN = \int_n^{n+1} d(\log N) = \log \frac{n+1}{n}.$$

Relacionada amb la llei de Zipf, ambdues lleis estudien la freqüència del conjunts. En el cas de Benford, els autors anteriors van proposar una densitat de probabilitat entre n i $n+1$ com n^{-1}, és just el cas limit de la llei de Zipf, $\alpha \to 1$.

3.2 Alguns fonaments teòrics de la llei

Segons Fedorowicz (1982) hi ha quatre grans escoles de pensament sobre els fonaments teòrics de la llei de Zipf:

1. màxima entropia de Mandelbrot que dóna lloc la llei de Zipf-Mandelbrot: un enfocament teòric de la informació per estudiar l'estructura estadística del llenguatge;
2. derivació de la funció beta de Simon;
3. derivació dels processos estocàstics d'Hill, Woodruffe, etc.;
4. i finalment la distribució acumulativa d'avantatges de Price, que manipular els models d'ocupació clàssics per produir una distribució hiperbòlica.

La màxima entropia de Mandelbrot i la derivació de Simon, de la dècada dels anys cinquanta del segle XX, que van ser els primers mètodes per aconseguir deduir la llei de Zipf.

El primer model, la màxima entropia de Mandelbrot, està basat en la teoria de la informació, que es una proposta teòrica presentada per Shannon y Weaver a finals de la dècada dels anys 1940. L'entropia, en aquest context, és una magnitud que mesura la informació proveïda per una font de dades. Com deia Mandelbrot: "*De fet, els detalls matemàtics mostraran que totes les variants del criteri de mímim*

esforç porten a la mateixa família "canònica" de lleis per a entitats concretes. Anem a classificar-les en l'ordre de freqüència decreixent. Aleshores, la freqüència p_k de la k-sìma entitat en aquesta classificació ha de ser donada per

$$p_k = P(k+m)^{-B}, \tag{3.5}$$

on P, m, B són certes constantes positives, ⋯" (Mandelbrot, 1953, p.491)

Sobre el segon model, com deia en Fedorowicz (1982), Simon va seguir treballant a partir de l'obra de Zipf, descrivint un conjunt de distribucions asimètriques deduides empíricament i va presentar en termes de freqüències de paraules. Ell va mostrar que la distribució de paraules en un text es comporta d'acord amb la següent equació:

$$f(k) = CB(k,\alpha), \text{tal que } \sum_{k=1}^{\infty} f(k) = 1,$$

on k i $\alpha > 1$ són constants i $B(k,\alpha)$ és la funció Beta de k i α:

$$B(k,\alpha) = \int_0^1 \lambda^{k-1}(1-\lambda)^{\alpha-1} d\lambda = \frac{\Gamma(k)\Gamma(\alpha)}{\Gamma(k+\alpha)}.$$

Utilitzant la propietat de la funció gamma:

$$\forall\alpha \quad \lim_{k\to\infty} \frac{\Gamma(k)}{\Gamma(k+\alpha)} \approx k^{-\alpha}.$$

Obtendrem

$$f(k) = C\Gamma(\alpha) r^{-\alpha}.$$

Sobre les derivacions a partir de processos estòcastics, la derivació d'Hill (1970, 1975) utilitzant la forma de Bose-Einstein del model d'ocupació clàssica amb un nombre aleatòria de cel·les és un dels més representatiu. Malgrat això, per la seva complexitat i tecnisimes, no donarem els detalls. En Fedorowicz (1982) hi ha una breu explicació i, sobre tot, arriba una forma importante de la llei de Zipf:

$$f(k) = \frac{1}{k(k+1)}. \tag{3.6}$$

De la manera paral·lela, Price (1970) va deduir la mateixa fórmula (3.6) utilitzant la *Cumulative Advantage Distribution* que és una derivació a partir d'una modificació del model de Pólya.

3.3 Model de *Random Typing*

Un dels primers desenvolupaments en la teoria de les lleis potencials va ser l'estudi de la distribució de freqüència i rang en llengües naturals de Mandelbrot (1953) i va ser coneguda com la llei de Zipf-Mandelbrot. Posteriorment, Miller (1957) va deduir aquesta llei utilitzant un simple experiment que, avui dia, es conegut com el model de *Random typing* o, també, *intermittent silence model*.

L'experiment consisteix en considerar un mico prement aleatòriament les tecles d'una màquina d'escriure de manera que (1) la probabilitat de prémer la barra espaiadora és p_0 i totes les restes tecles amb una probabilitat de $p_1 = 1 - p_0$ i (2) el mico mai ha de tocar la barra espaiadora dues vegades seguides. En les sortides del mico, podrem trobar les "paraules" de i lletres, on $i = 1, 2, 3, \cdots$, separades per un espai blanc. Aquí, considerem les paraules com a elements formats per un conjunt de lletres, encara que no tinguin sentits. Per tant, la probabilitat d'una paraula de longitud i és:

$$P_i = p_0 \cdot p_1^{i-1}, \quad i = 1, 2, 3, \cdots,$$

tal que aquesta probabilitat disminuirà exponencialment a mesura que i augmenta.

Ara considerem que hi ha A tecles diferents en la màquina d'escriure, excloent la barra espaiadora. Aleshores el nombre de diferents paraules possibles de longitud i ha de ser A^i. Per tant, la probabilitat d'una paraula particular de longitud i, denotem $p(\omega, i)$, ha de ser

$$p(\omega, i) = \frac{P_i}{A^i} = p_0 \cdot p_1^{i-1} A^{-i} = \frac{p_0}{p_1} e^{-i[\log A - \log p_1]}, \tag{3.7}$$

on log és logaritme natural.

Com que hi ha A tecles diferents disponibles, llavors hi ha d'haver A paraules d'una lletra, $A + A^2$ paraules de longitud $i \leq 2$, $A + A^2 + A^3$ de $i \leq 3$, etc. Per tant, una expressió general del nombre de paraules de longitud $\leq k$ és

$$\sum_{j=1}^{k} A^j = \frac{A(1 - A^k)}{1 - A}, \quad (A \neq 1).$$

Ara suposem que reordenem totes les diferents paraules respecte a la longitud. Llavors, les paraules d'una sola lletra estaran en un rang entre 1 i A, les de dues lletres estaran entre $A + 1$ i $A(1 - A^2)/(1 - A)$, etc. De manera que $r(\omega, i)$, rang mitjà de la paraula ω de longitud i, tindrà una expressió següent:

$$r(\omega, i) = \left(\frac{A(1 - A^{i-1})}{1 - A} + 1 + \frac{A(1 - A^i)}{1 - A} \right) /2 = \frac{A^i(A + 1)}{2(A - 1)} - \frac{A + 1}{2(A - 1)} \tag{3.8}$$

Reescriviu la fórmula (3.8) com segueix:

$$\frac{2(A - 1)}{A + 1} \cdot \left[r(\omega, i) + \frac{A + 1}{2(A - 1)} \right] = e^{i \log A}. \tag{3.9}$$

També reescrivim l'equació (3.7):

$$p(\omega, i) = \frac{p_0}{p_1} \left(e^{i \log A} \right)^{-(1 - \log p_1 / \log A)}, \tag{3.10}$$

substituïm (3.9) a (3.10), obtenim:

$$p(\omega, i) = \frac{p_0}{p_1} \left[\frac{2(A - 1)}{A + 1} \cdot \left[r(\omega, i) + \frac{A + 1}{2(A - 1)} \right] \right]^{-(1 - \log p_1 / \log A)}.$$

Com que p_0, $p_1 = 1 - p_0$ i A són constants, aleshores podrem arribar una nova fórmula definint unes noves constants:

$$p(\omega) = b[r(\omega) + c]^{-d} \quad d > 1. \tag{3.11}$$

Aquesta equació és justament la llei de Zipf-Mandelbrot (eq, 3.5).

El model de Miller està basant en la suposició del fet que les probabilitats de totes les lletres (excloent la barra espaiadora) són la mateixa, però en la realitat això no passarà. Degut això, alguns científics vam estudiar les relacions de les probabilitats i, el principi del segle XXI, Conrad&Mitzenmacher (2004) van descriure el cas de probabilitats desiguals mitjaçant els mètodes analítics.

Deficiència del model El model de Miller només afirma que els rangs mitjans dels textos aleatoris segueixen a la llei de Zipf-Mandelbrot i no ens diu com es porten les restes. En realitat, la distribució de freqüències d'aquest model té una forma d'escala, com s'il·lustra en la figura següent:

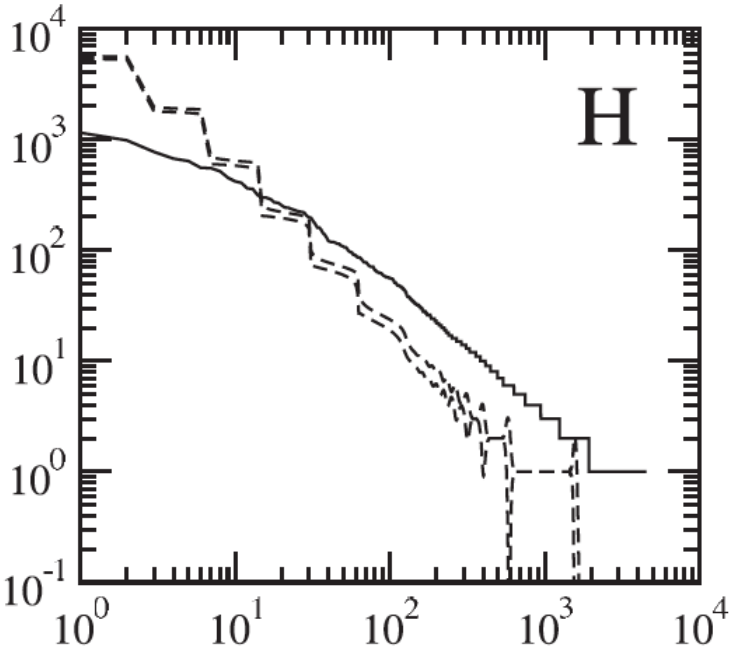


Figura 4: Figura treta de Ferre-i-Cancho&Elvevåg (2010). L'histograma de rang de text en anglès (Hamlet, Shakespeare) contra text aleatori.

Degut això, Mitzenmacher (2004) va estudiar una generalització de la llei potencial i va definir com següent: sigui f_j, la fracció asimptòtica de la j-èsima paraula més freqüent, segueix a una llei potencial si existeixen constants positives c_1, c_2, α tal que $c_1 j^{-\alpha} \leq f_j \leq c_2 j^{-\alpha}$ per un j suficientment gran.

4 Distribució de freqüències en text

En aquesta secció mostrem com la distribució de Zipf es presenta en un *corpus* de text real. Un corpus lingüístic és un conjunt, en general molt ampli, d'exmples reals d'ús d'una llengua. La llei de Zipf es pot caracteritzar per la freqüència de les paraules, també existeix una altra manera d'expressar-la. L'objectiu d'aquest capítol és introduir el concepte de l'espectre de freqüències i els models LNRE (*Large Number of Rare Event* en anglès) que farem servir per a l'anàlisi del Corpus Brown en el capítol següent.

4.1 Espectre de freqüències

Una manera de definir la llei de Zipf és basant la relació entre la freqüència de paraula i la seva posició (rank), també existeix una altra manera que és mitjaçant l'espectre de freqüencies que definirem més endavant. Per la intuició, hi ha una relació directa entre ambdos conceptes.

Suposem que tenim un corpus després de la *Tokenization* (el procés de delimitar i possiblement classificar segments d'una cadena de caràcters en unitat "to token" amb significant propi), és a dir, hem trobat una correspondència de cada símbol al tipus correspondent. D'ara endavant, farem servir el terme símbol com un equivalent aproximat al significant token. Per tant, podem comptar el nombre total N de símbols d'un corpus, la mida de corpus, i el nombre de tipus de paraules, la mida de vocabulari (V). La *Tokenization* també correspon al processament de llenguatge natural, les seves dificultats depenent del tipus de llenguatge.

D'aquesta manera formarà el nostre punt de partida, una llista de freqüències, per a qualsevol anàlisi addicional. Ara, considerem l'exemple de Baroni (2006), una petita lista de freqüències:

tipus	f	tipus	f	tipus	f
again	2	barks	6	her	1
and	3	dog	3	that	2
another	1	friends	1	this	1
bark	1	he	1	will	1
with	1				

Taula 2: La llista de freqüències

Les informacions en una llista de freqüències poden ser reorganitzades en dues maneres: la llista de freqüències ordenades i la llista d'espectre de freqüències. Ambdues són molt útils per estudiar la distribució de freqüències en textos.

Definició 4.1. *Una llista de freqüències ordenades és una parella* $(r_i, f_i = f(i, N))$*, on* f_i *és un valor de freqüències tal que compleix* $f_i \geq f_j$ *si* $i < j$ *i* $r_i = i$ *és la seva posició corresponent en la llista, que s'anomena el rank de Zipf o, simplement, el rank.*

En aquest sentit, f_i es defineix també com la freqüència en una mostra de N símbols d'una paraula amb el rank de Zipf i. A més a més, s'anomena la *distribució de freqüències ordenades de Zipf* a la distribució associada a f_i.

De fet, una llista de freqüències ordenades és una reordenació de la llista de freqüències amb nombre total N. Després del proces de la reordenació de la llista, també hi ha una assignació entre les paraules ω_j i els r_i's, en el cas que dues paraules tinguin la mateixa freqüència, l'assiginació d'ordre seria arbitrària. En el nostre exemple anterior, *barks* seria assignada l'ordre 1 a causa de tener una freqüència absoluta més alta. A més *and* i *dog*, l'ordre 2 i 3. En la taula següent, s'il·lustra la llista de freqüències ordenades de l'exemple anterior:

r	f	r	f	r	f
1	6	6	1	10	1
2	3	7	1	11	1
3	3	8	1	12	1
4	2	9	1	13	1
5	2				

Taula 3: la llista de freqüències ordenades

Definició 4.2. *Un espectre de freqüències o una distribució de freqüències agrupades, en aquest context, és una llista obtinguda a partir de la llista de freqüències ordenades, una llista de parelles* $(f_i, V(f_i))$ *on* $V(f_i)$ *és el nombre de paraules que apareixen amb freqüència* f_i.

Per exemple, l'espectre corresponent de la taula anterior és:

f	V(f)
1	8
2	2
3	2
6	1

Taula 4: la llista d'espectre

Veiem que la primera fila de la taula (3) ens indica que hi ha 8 paraules amb freqüència 1 (V(1)=8; *another, bark, friends, he, her, this, will, with*). La segona fila, 2 paraules amb freqüència 2 (V(2)=2; *again, that*), etc.

En general, es representa amb (m, $V_m = V(m, N)$) on m index de la classe de freqüència i V_m: el nombre de tipus de la clase de freqüència m i s'expressa com:

$$V_m = V(m, N) = \sum_{i=1}^{V(N)} \mathbb{1}_{\{f_i = m\}}.$$

D'aquesta manera, podem expressar N i $V(N)$ en els termes de m i V_m com:

$$N = \sum_m mV_m \quad \text{i} \quad V(N) = \sum_m V_m.$$

Una noció que està estretament relacionada amb l'espectre de freqüències V_m s'anomena la distribució empírica de tipus estructural (*empirical structural type distribution* en anglès).

Definició 4.3. *La distribució empírica de tipus estructural és una funció $g_m = g(m, N)$, que especifica el nombre de diferents tipus de paraules que es produeixen m vegades o més en una mostra de mida N, i es defineix com:*

$$g_m = g(m, N) = \sum_{i=1}^{V(N)} \mathbb{1}_{\{f_i \geq m\}}.$$

Aleshores les relacions entre la distrbució empírica de tipus estructural i l'espectre de freqüències són:

$$V_m = \sum_{i=1}^{V(N)} \mathbb{1}_{\{f_i \geq m+1\}} - \sum_{i=1}^{V(N)} \mathbb{1}_{\{f_i \geq m\}} = g_m - g_{m+1},$$

$$\text{i} \quad g_m = \sum_{i=1}^{V(N)} \mathbb{1}_{\{f_i \geq m\}} = \sum_{i \geq m} V_m.$$

A més, la distrbució empírica de tipus estructural és justament la inversa de la distribució de freqüències ordenades de Zipf:

$$g(m, N) = z \Leftrightarrow f(z, N) = m.$$

Per posar un exemple, utilitzem les dades de la taula 2, tenim

$$g(6, N) = 1 \quad \text{i} \quad f(1, N) = 6.$$

En les proximes seccions, introduim alguns models per la distribució de freqüències de paraules seguint el llibre de Baayen, *Word Frequency Distribution.*

4.2 Models no paramètrics

Per simplificar el problema, podem suposar que l'ús de paraules és una selecció aleatòria a partir d'una població amb una probabilitat fixada. Sota aquesta hipòtesi, és convenient considerar el model d'urna per la distribució de freqüències de paraules. Tot i així, no ens oblidem que l'aleatorietat i la independència d'una llengua no es pot grantitzar sota aquest simple experiment probabilístic, a causa de la gramàtica d'una llengua, la concordança en un text, etc. La idea principal és, per

un costat, entendre la dinàmica de la distribució de freqüències sobre la suposició bàsica del model d'urna i, per l'altre, construeix una base pels models de gran nombre d'esdeveniments rars (LNRE).

A partir d'allí, podem passar a un model de Poisson i, posterior, amb la introducció de la distribució de tipus estructural podem arribar a una forma integral de les expressions.

4.2.1 Model d'urna

Suposem que en una urna conté S diferents tipus de paraules ω_i, $i = 1, \cdots, S$. Per cada paraula ω_i associem una probabilitat π_i, $i = 1, \cdots, S$, de ser utilitzada, tal que compleix $\sum_{i=1}^{S} \pi_i = 1$. El mostreig de paraules consisteix en una selecció aleatòria amb la probabilitat indicada d'una paraula en l'urna amb el seu reemplaçament.

Què és la probabilitat d'una paraula ω_i amb una freqüència m en una mostra de mida N, $Pr(f(i, N) = m)$? Podem considerar la nostra mostra de N símbols com una seqüència de N proves amb m èxits (ω_i es va utilitzada) i $N - m$ fracassos (ω_i no es va utilitzada). La probabilitat d'una particular seqüència de prova és igual a $\pi_i^m(1 - \pi_i)^{N-m}$. Així doncs,

$$Pr(f(i, N) = m) = \binom{N}{m} \pi_i^m (1 - \pi_i)^{N-m}, \tag{4.1}$$

ja que tenim $\binom{N}{m}$ maneres de seleccionar m successos en N proves.

LLavors donat el model d'urna, la freqüència d'una paraula ω_i amb la probabilitat π_i en una mostra de N símbols és una distribució binomial (N, π_i). Utilitzant les propietats de la binomial, obtenim la freqüència esperada de ω_i en la mostra N:

$$\mathbb{E}[f(i, N)] = N\pi_i.$$

Utilitzant l'equació (4.1), podem obtenir les expressions següents:

$$\mathbb{E}[V(m, N)] = \mathbb{E}[\sum_{i=1}^{S} \mathbb{1}_{\{f(i,N)=m\}}] = \sum_{i=1}^{S} \mathbb{E}[\mathbb{1}_{\{f(i,N)=m\}}]$$

$$= \sum_{i=1}^{S} Pr(f(i, N) = m) = \sum_{i=1}^{S} \binom{N}{m} \pi_i^m (1 - \pi_i)^{N-m},$$

$$\mathbb{E}[V(N)] = \mathbb{E}[\sum_{i=1}^{N} V(m, N)] = \sum_{m=1}^{N} \mathbb{E}[V(m, N)]$$

$$= \sum_{m=1}^{N} \sum_{I=1}^{S} \binom{N}{m} \pi_i^m (1 - \pi_i)^{N-m}.$$

L'expressió de $\mathbb{E}[V(N)]$ es pot simplificar canviant la perspectiva. Considerem el cas que la paraula ω_i no apareix en la mostra, llavors la probabilitat corresponent és igual a $(1-\pi_i)^N$ i el seu complementari, la probabilitat d'una paraula que es produeix almenys una vegada en la mostra, és

$$1-(1-\pi_i)^N.$$

El nombre de tipus en una mostra és el nombre de tipus tal que es produeix almenys una vegada. Per tant,

$$\mathbb{E}[V(N)] = \mathbb{E}[\sum_{i=1}^{N} \mathbb{1}_{\{f(i,N)>0\}}] = \sum_{i=1}^{S} \mathbb{E}[\mathbb{1}_{\{f(i,N)>0\}}]$$

$$= \sum_{i=1}^{S} Pr(f(i,N)>0) = \sum_{i=1}^{S}(1-(1-\pi_i)^N) = S - \sum_{i=1}^{S}(1-\pi_i)^N.$$

De fet, les mostres de textos són generalment gran i la probabilitat de paraula petita, per tant, utilitzem l'aproximació de Poisson a la binomial podrem simplificar les expressions de $\mathbb{E}[V(N)]$ i $\mathbb{E}[V(m,N)]$.

Donada l'aproximació

$$\binom{N}{m}\pi_i^m(1-\pi_i)^{N-m} \approx \frac{(N\pi_i)m}{m!}e^{-N\pi_i},$$

podem reescriure $\mathbb{E}[V(N)]$ i $\mathbb{E}[V(m,N)]$ com segueix:

$$\mathbb{E}[V(m,N)] = \sum_{i=1}^{S}\frac{(N\pi_i)m}{m!}e^{-N\pi_i} \quad \text{i} \quad \mathbb{E}[V(N)] = \sum_{i=1}^{S}(1-e^{-N\pi_i}).$$

Tenint en compte que la probabilitat de Poisson d'una paraula ω_i que no apareix en una mostra de mida N és

$$Pr(f(i,N)=0) = \prod_{j=1}^{N} e^{-\pi_i} = e^{-\sum_{j=1}^{N}\pi_i} = e^{-N\pi_i}.$$

4.2.2 Distribució de tipus estructural

La importància de definir la distribució de tipus estructural és reescriure el model de Poisson en una forma integral. D'aquesta manera comporta expressions matemàtiques més pràctiques.

En la secció anterior, vam introduir el concepte de la distribució empírica de tipus estructural $g(m,N)$:

$$g(m,N) = \sum_{i=1}^{V(N)} \mathbb{1}_{\{f(i,N)\geq m\}}.$$

Anàlogament, podem definir la distribució de tipus estructural $G(\pi)$ que correspon a les probabilitats de la població.

Definició 4.4. *La distribució de tipus estructural* $G(\pi)$

$$G(\pi) = \sum_{i=1}^{S} \mathbb{1}_{\{\pi_i \geq \pi\}}$$

és el nombre de tipus en la població amb una probabilitat més gran o igual a π*.*

Denotem el nombre de tipus amb la probabilitat π per $V(\pi)$,

$$V(\pi) = \sum_{i=1}^{S} \mathbb{1}_{\{\pi_i = \pi\}},$$

a més podem reordenar les probabilitats π's per les que compleixen $V(\pi) > 0$, de manera que $\pi_j < \pi_{j+1}$. Aleshores, els salts a les probabilitats π_j, $j = 1, \cdots, k$, $k \leq S$ són donats per

$$\Delta G(\pi_j) = G(\pi_j) - G(\pi_{j+1}).$$

En altres paraules, $\Delta G(\pi_j)$ es denota el nombre de paraules de la població amb la probabilitat π_j.

A continuació, podem replantejar-nos les expressions de l'espectre de freqüències i la mida del vocabulari esperat en una forma integral:

$$\begin{aligned}
\mathbb{E}[V(m,N)] &= \sum_{i=1}^{S} \frac{(N\pi_i)^m}{m!} e^{-N\pi_i} = \sum_{j=1}^{k} \frac{(N\pi_j)^m}{m!} e^{-N\pi_j} \Delta G(\pi_j) \\
&= \int_0^\infty \frac{(N\pi)^m}{m!} e^{-N\pi} dG(\pi), \qquad (4.2) \\
\mathbb{E}[V(N)] &= \sum_{i=1}^{S} (1 - e^{-N\pi_i}) = \sum_{j=1}^{k} (1 - e^{-N\pi_j}) \Delta G(\pi_j) \\
&= \int_0^\infty (1 - e^{-N\pi}) dG(\pi).
\end{aligned}$$

Tinguem en compte que $dG(\pi) \neq 0$ si i només si $dG(\pi)$ és un interva infinitesimal que conté els π_j's, on $dG(\pi) = \Delta G(\pi)$. El canvi d'una suma a una integral està basat en la integral de Riemann-Stieltjes, o la integral de Stieltjes, que és una generalització de la integral de Riemann. Siguin $f(x)$ i $\alpha(x)$ funcions reals acotades definides en un interval tancat $[a, b]$. Donada una partició del interval

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b,$$

i considerem la suma de Riemann

$$\sum_{i=0}^{n-1} f(\xi_i)[\alpha(x_{i+1}) - \alpha(x_i))]$$

amb $\xi_i \in [x_i, x_{i+1}]$. Si la suma tendeix a un valor fix l quan $\max(x_{i+1} - x_i) \to 0$, aleshores S s'anomena la integral de Stieltjes. La integral de Stieltjes de f respecte a α és denota

$$\int_a^b f(x)d\alpha(x) \quad \text{o simplement} \quad \int_a^b f d\alpha.$$

A més si f és continua i α' és integral de Riemann en el interval indicat, aleshores

$$\int f(x)d\alpha(x) = \int f(x)\alpha'(x)dx. \tag{4.3}$$

En el cas de l'expressió per $\mathbb{E}[V(N)]$, per exemple, $f(x) = 1 - e^{Nx}$ (hem canviat la variable π per x).

Fent un canvi de prespectiva, passem a una integral en $[0, 1]$ a $(0, \infty)$. Com que el paràmetre d'una distribució de $Poisson(\lambda)$ d'una variable aleatòria pot ser interpretat com la taxa a la qual es produeix un esdeveniment particular. Suposem que hem escollit N_0 com la unitat de mesura de símbols i definim $\lambda_i = N_0\pi_i$, aleshores λ_i és equivalent a dir que la paraula w_i apareix $N_0\pi_i$ vegades en un interval de temps de N_0 símbols. Per tant, existeix una bijecció entre $[0, 1]$ a $[0, \infty)$ i reescriure $N/N_0 = t$ tindrem les expressions per l'espectre de freqüències i la mida de vocabulari esperat com segueix:

$$\mathbb{E}(V(m, t)) = \int_0^\infty \frac{(t\lambda)^m}{m!} e^{-t\lambda} dG(\lambda)$$

$$\mathbb{E}(V(t)) = \int_0^\infty (1 - e^{-t\lambda}) dG(\lambda).$$

En aquestes dues expressions $\lambda = N_0\pi$ es denota com la freqüència absoluta d'una paraula en una mostra de N_0 símbols. Malgrat això, com que una sola paraula és la unitat de mesura més òbvia, llavors és més convenient formular les expressions en termes de N i π que en termes de t i λ.

Per qualsevol funció, la seva primera derivada indica la rapidesa de les variacions. Tanmateix, podem fixar la derivada de $\mathbb{E}[V(N)]$ que expressa la taxa de creixement de vocabularis despés d'un mostreig de N símbols:

$$\frac{d}{dN}\mathbb{E}[V((N)] = \frac{d}{dN}\int_0^\infty (1 - e^{-N\pi}) dG(\pi) = \int_0^\infty -\pi \cdot -e^{-N\pi} dG(\pi)$$

$$= \frac{1}{N}\int_0^\infty N\pi e^{N\pi} dG(\pi) = \frac{\mathbb{E}[V(1, N)]}{N}.$$

Aquesta expressió ens porta a la definició següent:

Definició 4.5. *La taxa de creixement $\mathscr{P}(N)$ de vocaburaris, la velocitat a la qual augmenta la mida del vocabulari si augmenta la mida de la mostra, es defineix mitjaçant:*

$$\mathscr{P}(N) = \frac{d}{dN}\mathbb{E}[V(N)] = \frac{\mathbb{E}[V(1, N)]}{N}.$$

En general es denota aquesta taxa com simplement $\mathscr{P}$ i s'anomena, en un context lingüístic, **la mesura de productivitat**. Per una altra banda, La taxa de creixement dels elements individuals de l'espectre $\mathbb{E}[V(m,N)]$ també es pot obtenir diferenciant-se en la variable N.

$$\begin{aligned}
\frac{d}{dN}\mathbb{E}[V(m,N)] &= \frac{d}{dN}\int_0^\infty \frac{(N\pi)^m}{e^{-N\pi}} dG(\pi) \\
&= \int_0^\infty \left(mN^{m-1}\frac{\pi^m}{m!}e^{-N\pi} - \frac{(N\pi)^m}{m!}\pi e^{-N\pi} \right) dG(\pi) \\
&= \frac{1}{N}\int_0^\infty \left(m\frac{(N\pi)^m}{m!}e^{-N\pi} - (m+1)\frac{(N\pi)^{m+1}}{(m+1)!}e^{-N\pi} \right) dG(\pi) \\
&= \frac{1}{N}\left(m\mathbb{E}[V(m,N) - (m+1)\mathbb{E}[V(m+1,N)] \right). \qquad (4.4)
\end{aligned}$$

Denotem N_m^*: la mida de la mostra a la qual els elements de l'espectre $\mathbb{E}[V(m,N)]$ assoleix el seu màxim. En el cas de hàpax legòmena, el seu màxim assoleix és justament quan el nombre de hàpax legòmena és doble que el nombre de dilegòmena (les paraules que surten dues vagades). Ja que el nombre de hàpax assoleix el seu màxim quan

$$\frac{d}{dN}\mathbb{E}[V(1,N)] = 0,$$

utilitzant l'equació (4.4), obtenim

$$\frac{\mathbb{E}[V(1,N)] - 2\mathbb{E}[V(2,N)]}{N} = 0,$$

per tant, en N_1^*, $\mathbb{E}[V(1,N)] = 2\mathbb{E}[V(2,N)]$. De la manera sembla, obtenim una fórmula general:

$$\mathbb{E}[V(m,N_m^*)] = \frac{m}{m+1}\mathbb{E}[V(m+1,N_m^*)].$$

4.3 Model paramètric: LNRE

La teoria de gran quantitat d'esdeveniments rars, en anglès *Large Numbers of Rare Events* (LNRE), va introduir per Khmaladze en l'any 1987. Des d'alguns punts de vista, la presència d'un gran nombre d'esdeveniments rars és una característica fonamental de la natura, tant com en lingüística, en química, en demogràfica, o, etc. En particular, en qualsevol anàlisi estadística dedicada a l'estudi de la varietat de paraules en el gran text, un ha de tractar aquest tipus d'esdeveniments que són les paraules de baixa freqüència. Aquests tipus de paraules contribueixen una petita part comparant-los amb el nombre total d'observacions, en canvi, dins de tots els diferents tipus observats és bastant significatiu. Aquests esdeveniments rars normalment són molt importants.

El model LNRE està basant en el model d'urna, aproximant $G(\pi)$ per la integral d'una funció contínua que s'anomena la densitat de tipus estructural $g(x)$

$$G(\pi) = \int_\pi^\infty g(x)dx.$$

A partir d'ara, utilitzem la variable π per la funció g i ρ per la funció G només per la comoditat. A més utilitzem el límit d'integració superior $+\infty$ només per una forma matemàtica més elegant, encara que totes probabilitats de tipus es cauen en el rang $0 \leq \pi \leq 1$. Finalment, $g(x)$ compleix

$$\int_0^\infty \pi g(\pi) d\pi = 1, \text{a causa de } \sum_{i=1}^{S} \pi_i = 1 \tag{4.5}$$

i la mida poblacional de tipus es donada per $S = \int_0^\infty g(\pi) d\pi$.

Una definició diferent de la funció g ens porta a un model diferent. Per exemple, en Baayen (2001) hi ha tres families del model LNRE com el Lognomal-LNRE, la inversa de Gauss-Poisson generalitzada LNRE i la familia zipfiana LNRE. En aquest secció només ens preocupem els models de LNRE relacionats amb la familia de llei de Zipf, en particular la llei de Zipf-Mandelbrot.

4.3.1 Zones de LNRE

Tornem a la distribució de freqüències, primer de tot, definim $P(i, N)$ la freqüència relativa de la paraula ω_i d'una mostra de mida N. Aleshores, segons la llei de gran nombre, per qualsevol distribució de probabilitat

$$(\pi_i,\ 1 \leq i \leq S)$$

amb una mida de vocabularis S finita la freqüència relativa mostral convergeix a la probabilitat poblacional si $N \to \infty$

$$\lim_{N \to \infty} P(i, N) = \pi_i.$$

Com una conseqüència simple

$$\lim_{N \to \infty} V(m, N) = 0 \quad \forall m.$$

Malgrat això, les distribucions de freqüències de paraules són distribucions de LNRE, distribucions caracteritzades per la presència d'un gran nombre de paraules amb molt baixes probabilitats d'ocurrència. En el Corpus Brown, per exemple, més de 40% de tots tipus té una freqüència relativa mostral de 0.0000001. A causa d'això, la mida mostral N cal ser extremadament gran per a que sorgeixin les propietats asimptòtiques de la distribució. En la pràctica, quasi totes les mostres de paraules estan en una zona que s'anomena la *zona de LNRE*.

Definició 4.6. *Una zona de LNRE és el rang de la mida mostral on la mida de vocabulari encara està augmentant, i on el nombre de hàpax legòmena, dilegòmena, etc., són no menyspreables.*

Sigui $\nu(N) = (f(1, N), f(2, N), \cdots, f(S, N))$ el vector de freqüències del tipus de paraules com es va realitzar en una mostra de N símbols. Obtindrem una successió dels vectors augmentant el valor de N:

$$\{\nu(N)\},\ N = 1, 2, 3, \cdots.$$

Inicialment, la majoria de les freqüències de paraules seran zero, però si incrementem la mida de mostra, apareixeran més i més paraules amb freqüències no nul·les. D'aquí, sorgeix la primera definició de *successió de LNRE*.

Definició 4.7. *Una successió $\{\nu(N)\}$ és una successió amb un gran nombre d'esdeveniments rars, o una successió de LNRE, si*

$$\lim_{N\to\infty} \frac{\mathbb{E}[V(1,N)]}{N} > 0.$$

D'acord a aquesta definició, tenim una distribució de LNRE en el cas que la taxa de creixement de vocabularis sigui major que zero, fins i tot quan N s'incrementa indefinidament. Així i tot, si fixem un nombre finit del tipus S i les probabilitats π_i, $i = 1, 2, \cdots, S$, aleshores

$$\lim_{N\to\infty} f(i,N) = \infty,$$

en conseqüència,

$$\lim_{N\to\infty} \mathbb{E}[V(1,N)] = 0 \quad \text{i} \quad \lim_{N\to\infty} \mathbb{E}[V(N)] = S.$$

En aquesta situació, la taxa de creixement es convertirà en zero. Encara que per distribucions amb S infinit, una taxa no nul·la per $N \to \infty$ no està garantida. La segona definició de Khmaladze és més suau:

Definició 4.8. *Una successió $\{\nu(N)\}$ és una successió amb un gran nombre d'esdeveniments rars, o una successió de LNRE, si*

$$\lim_{N\to\infty} \frac{\mathbb{E}[V(1,N)]}{\mathbb{E}[V(N)]} > 0 \quad i \quad \lim_{N\to\infty} \mathbb{E}[V(N)] = \infty.$$

Per la comoditat, anomenem zona de LNRE $d1$ i $d2$ si es compleix la definició (4.7) i (4.8) respectivament. No hi ha una equivalència entre ambdues definicions, la $d1 \Rightarrow d2$, pero no viceversa. Per exemple, en Baayen (2001), sabem que $V(m,N)$ segueix a la llei de Zipf de la forma

$$V(m,N) \propto \frac{1}{m(m+1)}.$$

Més en concret,

$$V(m,N) \approx \frac{V(N)}{m(m+1)} \Leftrightarrow \frac{V(m,N)}{V(N)} \approx \frac{1}{m(m+1)},$$

perquè el nombre de hàpaxs tendeix a tenir la meitat de la mida de vocabularis. Per tant, en el cas $m = 1$,

$$\lim_{N\to\infty} \frac{V(1,N)}{V(N)} = \frac{1}{2}.$$

Veiem que es compleix $d2$ i no $d1$. De fet

$$\frac{N}{\mathbb{E}[V(N)]} = \sum_{m=1}^{\infty} \frac{m\mathbb{E}[V(m,N)]}{\mathbb{E}[V(N)]} \to \infty, \quad n \to \infty$$

i per tant,

$$\frac{V(1,N)}{N} \leq \frac{V(N)}{N} \to 0.$$

També hi ha unes condiciones sobre $d1$ i $d2$ que estan detallades en Khmaladze (1988).

4.3.2 LNRE de Zipf-Mandelbrot

La distribució de freqüències obtinguda sota el model d'urna, que es defineix anteriorment, és molt similiar a la llei de Zipf. Rouault va demostrar que, sota condicions moltes generales, les probabilitats poblacional de tipus amb baixa freqüència satisfà la llei de Zipf-Mandelbrot

$$\pi_i = \frac{C}{(i+b)^a}, \text{ amb } a > 1 \text{ i } b > 0.$$

La distribució de tipus estructural corresponent a la llei de Zipf-Mandelbrot és la funció esglaonada amb $G(\pi_i) = i$. Resoldre l'equació anterior per i:

$$\pi_i = \frac{C}{(i+b)^a} \Leftrightarrow \frac{C}{\pi_i} = (i+b)^a \Leftrightarrow i = \frac{C^{1/a}}{\pi_i^{1/a}} - b,$$

obtenim

$$G(\pi) = \frac{C^{1/a}}{\pi^{1/a}} - b, \text{ per } \pi = \pi_i, \tag{4.6}$$

i $G(\pi)$ és constant entre (π_i, π_{i+1}). Diferenciant l'equació (4.6) obtenim una forma de la densitat del tipus:

$$g(\pi) := \begin{cases} C \cdot \pi^{-\alpha-1} & 0 \leq \pi \leq B \\ 0 & \text{altres cassos} \end{cases} \tag{4.7}$$

amb dos paràmetres lliures $0 < \alpha < 1$ i $B > 0$. Tinguem en compte que el paràmetre α ve donat per l'expressió $\alpha = 1/a$ i la cota superior B és necessària, ja que el model prediria els tipus amb probabilitat $\pi > 1$ altrament. La constant C ve donada per la normalització:

$$1 = \int_0^{\infty} \pi g(\pi) d\pi = \int_0^{B} \pi \cdot C \cdot \pi^{-\alpha-1} d\pi = \int_0^{\infty} C\pi^{-\alpha} d\pi$$

$$= C \left[\frac{\pi^{1-\alpha}}{1-\alpha} \right]_0^B = C \cdot \frac{B^{1-\alpha}}{1-\alpha},$$

on obtenim

$$C = \frac{1-\alpha}{B^{1-\alpha}}. \tag{4.8}$$

La definició de g ens porta a un model de Zipf-Mandelbrot (ZM model) amb una població infinita, ja que $S = \int_0^B g(\pi)d\pi = \infty$. En aquest cas, la seva distribució de tipus estructural serà:

$$G(\rho) = \int_\rho^B g(\pi)d\pi = C \cdot \int_\rho^B \pi^{-\alpha-1}d\pi = C \cdot \left[\frac{\pi^{-\alpha}}{-\alpha}\right]_\rho^B$$

$$= \frac{C \cdot \rho^{-\alpha}}{\alpha} - \frac{C \cdot B^{-\alpha}}{\alpha} \overset{(4.8)}{=} \frac{C/\alpha}{\rho^\alpha} - \frac{1-\alpha}{B \cdot \alpha}$$

que és idèntica a l'equació (4.6) fent el canvi $a = \alpha^{-1}$ i $b = (1-\alpha)B^{-1}\alpha^{-\alpha}$ per qualsevol valor ρ on $G(\rho) \in \mathbb{N}$. També no s'oblida que la constant C que apareix en l'equació (4.6) i (4.7) no és la mateixa. Per tant, utilitzant la transformació d'integral de Stieltjes (4.3) i la definició de la densitat de tipus (4.7), podem obtenir una extenció contínua de la llei de Zipf-Mandelbrot sobre l'expressió d'esperança d'elements espectrals (4.2):

$$\mathbb{E}[V(m,N)] = \int_0^\infty \frac{(N\pi)^m}{m!}e^{-N\pi}dG(\pi) = \int_0^\infty \frac{(N\pi)^m}{m!}e^{-N\pi}g(\pi)d\pi$$

$$= \frac{C}{m!}\int_0^B (N\pi)^m e^{-N\pi}\pi^{-\alpha-1}d\pi \overset{(N\pi=t)}{=} \frac{C}{m!}\int_0^{NB} t^m e^{-t}\left(\frac{t}{N}\right)^{-\alpha-1}\frac{1}{N}dt$$

$$= \frac{C \cdot N^{-\alpha}}{m!}\int_0^{NB} t^{m-\alpha-1}e^{\cdot t}dt \approx \frac{C \cdot N^\alpha}{m!}\int_0^\infty t^{m-\alpha-1}e^{-t}dt$$

Utilitzem l'aproximació en l'última línea, ja que $NB \gg m$ i on la integral

$$\int_{NB}^\infty t^{m-\alpha-1}e^{-t}dt$$

és suficientment petita. Llavors, $\mathbb{E}[V(m,N)]$ es redueix a una funció gamma

$$\int_0^\infty t^{m-\alpha-1}e^{-t}dt = \Gamma(m-\alpha)$$

i obtenim

$$\mathbb{E}[V(m,N)] = \frac{C \cdot N^\alpha}{m!} \cdot \Gamma(m-\alpha). \tag{4.9}$$

El càlcul de $\mathbb{E}[V(N)]$ involucra la integral impròpia de Riemann per una integració per parts:

$$\mathbb{E}[V(N)] = \int_0^\infty (1-e^{-N\pi})g(\pi)d\pi \approx C \cdot N^\alpha \int_0^\infty (1-e^{-t})t^{-\alpha-1}dt$$

$$= C \cdot N^\alpha \cdot \lim_{A\to 0}\left(\int_A^\infty t^{-\alpha-1}dt - \int_A^\infty e^{-t}t^{-\alpha-1}dt\right)$$

integral directament la primera part i utilitzem la integració per part en la segona, $u = e^{-t}$ i $dv = t^{-\alpha-1}dt$, obtenim

$$= C \cdot N^\alpha \lim_{A\to 0}\left(\left[\frac{t^{-\alpha}}{-\alpha}\right]_A^\infty - \left[\frac{e^{-t}\cdot t^{-\alpha}}{-\alpha}\right]_A^\infty + \int_0^\infty \frac{e^{-t}\cdot t^{-\alpha}}{\alpha}dt\right)$$

$$= C \cdot N^{\alpha} \lim_{A \to 0} \left(\underbrace{(1 - e^{-A}) \cdot \frac{A^{-\alpha}}{\alpha}}_{=O(A^{1-\alpha}) \to 0)} + \underbrace{\frac{\Gamma(1-\alpha, A)}{\alpha}}_{\to \Gamma(1-\alpha)/\alpha} \right)$$

on $\int_A^{\infty} e^{-t} t^{-\alpha} dt = \Gamma(1-\alpha, A)$ és la funció gamma incompleta superior. Per tant,

$$\mathbb{E}[V(N)] = C \cdot N^{\alpha} \cdot \frac{\Gamma(1-\alpha)}{\alpha}. \tag{4.10}$$

Com conseqüències de (4.9) i (4.10), obtenim les relacions de recurrència

$$\frac{\mathbb{E}[V(m+1, N)]}{\mathbb{E}[V(m, N)]} = \frac{\Gamma(m+1-\alpha)}{(m+1)!} \cdot \frac{m!}{\Gamma(m-\alpha)} = \frac{m-\alpha}{m+1},$$

i

$$\frac{\mathbb{E}[V(m, N)]}{\mathbb{E}[V(N)]} = \frac{\alpha \cdot \Gamma(m-\alpha)}{\Gamma(m+1) \cdot \Gamma(1-\alpha)}$$

de la qual és independent de la mida de mostral N. El segon quocient és justament l'element espectral relatiu que va definir Baayen (2001, p. 94) en els models LNRE:

$$\alpha(m) = \frac{\mathbb{E}[V(m, N)]}{\mathbb{E}[V(N)]},$$

ja que no existeix una expressió completa de la distribució de tipus estructural per aquesta expresió

$$\mathbb{E}[V(m, N)] = \frac{\mathbb{E}[V(N)]}{m(m+1)}.$$

LNRE de Zipf-Mandelbrot finit Encara que el model ZM és teòricament ben fundat com un model per seqüències de caràcters aleatoris, però la suposició d'una mida infinita de vocabulari no és realista per dades de llenguatge natural. D'acord amb aquesta idea, és més convenient introduir una cota inferior $A > 0$ a la densitat de tipus:

$$g(\pi) := \begin{cases} C \cdot \pi^{-\alpha-1} & A \leq \pi \leq B \\ 0 & \text{altre cas} \end{cases}$$

en la qual implica que no existeix ningun tipus de paraules amb la probabilitat $\pi < A$ en la població. A causa de la normalització, obtenim la constant C

$$C = \frac{1-\alpha}{B^{1-\alpha} - A^{1-\alpha}},$$

i la mida de la població

$$S = \frac{C}{\alpha} \cdot (A^{-\alpha} - B^{-\alpha}) = \frac{1-\alpha}{\alpha} \cdot \frac{A^{-\alpha} - B^{-\alpha}}{A^{1-\alpha} - B^{1-\alpha}}.$$

De nou, obtenim una distribució de tipus estructral mateixa que abans, amb $G(\rho) = S$ per $\rho \leq A$. Després tenim

$$\mathbb{E}[V(m, N)] = \frac{C}{m!} \cdot N^{\alpha} \cdot \Gamma(m-\alpha, NA),$$

$$\mathbb{E}[V(N)] = C \cdot N^{\alpha} \cdot \frac{\Gamma(1-\alpha, NA)}{\alpha} + \frac{C}{\alpha \cdot A^{\alpha}}(1 - e^{NA}).$$

I finalment, no existeix una expressió simple per la relació de recurrència ni l'element d'espectre relatiu.

5 Anàlisi del Corpus Brown

La intenció d'aquest capítol és fer una petita anàlisi sobre el Corpus Brown que involucran els conceptes del capítol anterior. L'eina que utilitzem és el paquet `zipfR` en el llenguatge R que implementa els models "*Large Number of Rare Event*". `zipfR` que s'està desenvolupant per Evert&Baroni (2006).

El Corpus Brown (*The Brown University Standard Corpus of Present-Day American English*) va ser el primer corpus llegit per ordinador per a la investigació lingüística sobre l'anglès modern. Va ser compilat per W. Nelson Francis i Henry Kucera a la Universitat Brown de la dècada dels seixanta i conté 500 mostres de text en anglès, que totalitzen aproximadament un milió de paraules, compilades a partir d'obres publicades als Estats Units el 1961. A més conté dos tipus de fons: prosa informativa (reportatge, text de ciència, etc.) i prosa imaginativa (ficció, humor, aventura, etc.). Hi ha diferents versions, la versió que vam utilitzar està dins del paquet `zipfR`. URL de Corpus Brown:

http://www.helsinki.fi/varieng/CoRD/corpora/BROWN/

En la taula següent, podem veure el rank superior i inferior al Corpus Brown, i les paraules corresponents: Fixem-nos en la taula, podem veure que els primers 10 tipus

més freqüent			menys freqüent		
r	f	paraula	rang de rank	f	exemples arbitraris
1	69836	the	7731 - 8272	10	schedules, polynominals, bleak
2	36365	of	8273 - 8922	9	tolerance, shaved, hymn
3	28826	and	8923 -9703	8	decreased, abolish, irresistible
4	26126	to	9704 - 10783	7	immunity, cruising, titan
5	23157	a	10784 - 11985	6	geographic, lauro, portrayed
6	21314	in	11986 - 13690	5	grigori, slashing, developer
7	10777	that	13691 - 15991	4	sheath, gaulle, ellipsoids
8	10182	is	15992 - 19627	3	mc, initials, abstracted
9	9968	was	19628 - 26085	2	thar, slackening, deluxe
10	9801	he	26086 - 45215	1	beck, encompasses, second-place

Taula 5: una part de Corpus Brown

de paraules més freqüents contribueixen 24% del recompte de símbol total al Brown (246,352 ocurrències per sobre 1,006,770 símbols en total), en canvi, només ocupen 0.02% de vocabularis. La imatge és molt diferent de la part inferior, les paraules que es repeteixen 4 vegades o menys contribuïen quasi 70% dels vocabularis. En canvi, aquest 70% de tipus només conta un 5% del recompte de símbol total al Brown (52,158 sobre 1,006,770). Els elements de freqüències més baixes són, per descomptat, paraules de continguts. Aleshores si fixem el complementari de aquest 70% de tipus, obtenim les següents dades:

30% de vocabularis contribueix al **95%** de paraules observades.

Ara bé, si fixem les paraules amb una freqüència igual o inferior a 10, aquest conjunt de dades ocupa un 83% del vocabulari i 10% d'observat. El seu complementari contribueix **18%** del vocabulari i **90%** d'observat. Aquestes dades compleixen **el principi de Pareto** o conegut com la regla de 20/80. La idea principal d'aquest principi és una petita part de causes produeixen la majoria de resultats. I la llei que está relacionada amb aquest principi és la llei de Pareto, com que estem en el cas discret, aleshores es tracta de la llei de Zipf. La distribució en text és un dels grans camps on s'aplica aquesta llei.

5.1 Informació general i espectre de freqüència

L'eina que utilitzem és el paquet `zipfR` en el llenguatge R. En el mateix paquet podem trobar les dades de freqüència de Corpus Brown, contenen els objectes `Brown.tfl`, `Brown.spc` i `Brown.emp.vgc` que són de les classes tfl (*Type Frequency Lists*, la lista de freqüència), spc (*Frequency Spectra*, l'espectre de freqüènecies) i vgc (*Vocabulary Growth Curves*, les corbes de creixement de vocabularis) respectivament. També existeix una funció `tfl2spc` que converteix una llista de freqüència a una llista de l'espectre de freqüències.

A partir d'ara, tots els anàlisis posteriors estan basats en l'espectre de freqüència, que és una estructura més important en `zipfR`. Un espectre de freqüències resumeix una distribució de freqüències en termes del nombre de tipus (V_m) per la seva classe de freqüència (m). Per exemple, els primers filas del `Brown.spc` són:

```
> Brown.spc
m     Vm
1    1 19130
2    2  6458
3    3  3636
4    4  2301
5    5  1705
6    6  1202
7    7  1080
8    8   781
9    9   651
10 10    541
...

N      V
1006770 45215
```

Observem que en el nostre corpus hi ha 19130 paraules que només han aparegut una vegada, que corresponen als hàpax legòmen, 6458 paraules que surten dues vegades, etc. En el secció del model LNRE, vam dir que els esdeveniments rars normalment són molt importants. En aquest cas, podem veure que els hàpaxs dins aquest corpus només tenen un pes de

```
> Vm(Brown.spc,1)/N(Brown.spc)
[1] 0.01900136
```

de totes les paraules observades, en canvi, ocupen un

```
> Vm(Brown.spc,1)/V(Brown.spc)
[1] 0.4230897
```

de tots tipus de paraules.

El concepte que relacionat amb els hàpaxs és la mesura de productivitat $\mathscr{P}$ i podem calcular la pel Corpus Brown (CB):

```
> Vm(Brown.spc,1)/N(Brown.spc)
[1] 0.01900136
```

La taxa d'apareixement de nou tipus de paraules respecte a l'augment de la mostra per CB és bastant baixa.

En el paquet `zipfR` també hi han diverses funciones útils per explorar la distribució de freqüències. Abans d'això, podem obtenir una informació general de l'espectre mitjaçant la funció `summary`:

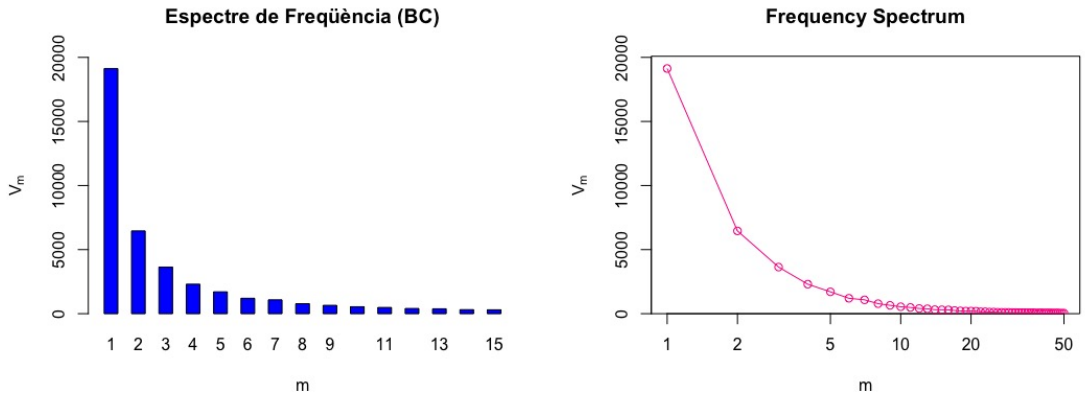
```
> summary(Brown.spc)
zipfR object for frequency spectrum
Sample size:     N  = 1006770
Vocabulary size: V  = 45215
Class sizes:     Vm = 19130 6458 3636 2301 1705 1202 1080 781 ...
```

En la taula anterior, podem veure les informacions com la mida de la mostra objecta, la mida de vocabularis i els elements d'espectre. A més a més, també s hi han les funciones `N`, `V` i `Vm` per donar les informacions anteriors:

```
> N(Brown.spc)
[1] 1006770
> V(Brown.spc)
[1] 45215
> Vm(Brown.spc,1)
[1] 19130
> Vm(Brown.spc,1:5)
[1] 19130  6458  3636  2301  1705
```

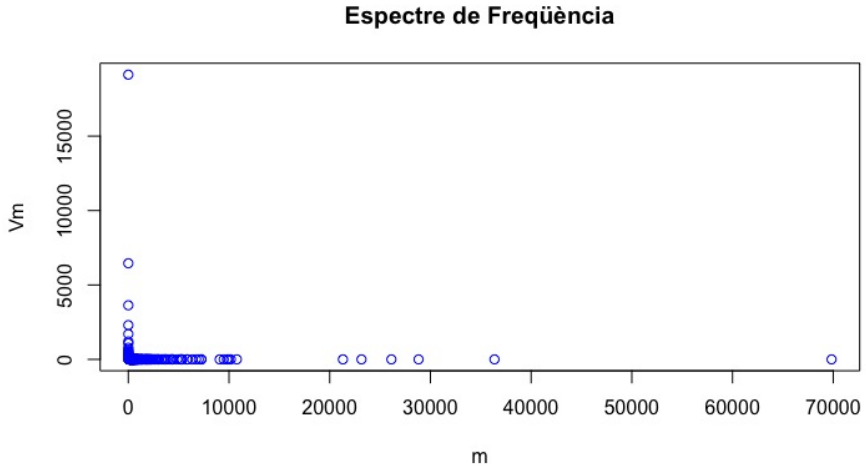
En el cas de la funció `Vm`, cal indicar la classe de freqüència m. En cas que m sigui un vector, la funció retornarà un vector indicant els nombres de tipus de les classes correspostes.

Una manera de visualitzar l'espectre de freqüència és mitjaçant les gràfiques:



La diferència entre dues figures és l'escala d'eix m, en la parte dreta, hem aplicat un argument extra `log="x"` que ens mostra els primers 50 elements d'espectre amb l'eix m en una escala logarítmica.

També podem visualitzar l'espectre de freqüència complet:



A partir de la figura, una visualització completa no és necessari, ja que la majoria de punts està situada a la part esquerra i la visualització completa no dóna cap informació extra. Com esmentat anterior, l'espectre de freqüència es caracteritza sovint pels valors alts que corresponent a les classes de freqüència més baixa i una llarga cua de classe de freqüència amb només un membre. Per tant un espectre complet en una escala no logarítmica tendrà una forma L com mostra en la figura anterior.

5.2 Corbes de creixement de vocabularis i interpolació

La idea de les corbes de creixement es veure la rapidesa del creixement de la mida de vocabularis, $V(N)$. Per exemple:

· mostra: a b b c a a b a

· $N = 1$, $V = 1$, $V_1 = 1$ ($V_2 = 0, \cdots$)

· $N = 3$, $V = 2$, $V_1 = 1$ ($V_2 = 1$, $V_3 = 0, \cdots$)

· $N = 5$, $V = 3$, $V_1 = 1$ ($V_2 = 2$, $V_3 = 0, \cdots$)

· $N = 8$, $V = 3$, $V_1 = 1$ ($V_2 = 0$, $V_3 = 1$, $V_4 = 1, \cdots$)

A partir de $N = 5$, la mida de vocabulari deixa de créixer i el hàpax sempre manté el mateix nombre. A més, les corbes també utilitzem per comparar diferents corpus. Considerem els dos "corpus": `a b a b a b a b` i `a a a a b b b b`: ells tenen el mateix espectre de freqüència, però diferents corbes de creixement.

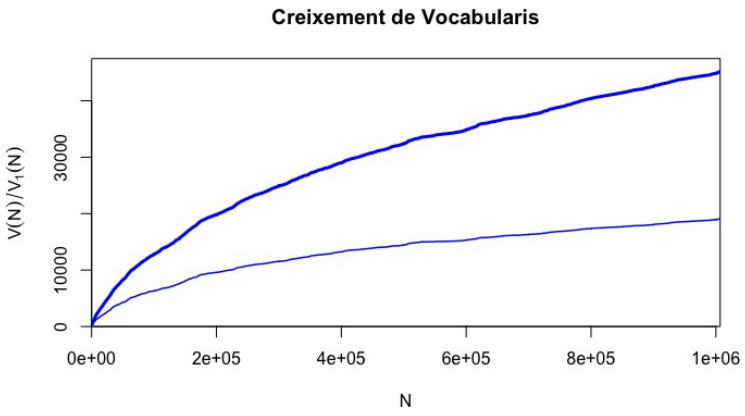
En el nostre cas, tenim les dades necessàries per dibuixar les corbes de creixement empírica del BC, `Brown.emp.vgc`. Les primeres files d'aquest objecte són:

```
> head(Brown.emp.vgc)
     N    V   V1
1 1000  442  311
2 2000  760  489
3 3000 1069  686
4 4000 1322  809
5 5000 1524  912
6 6000 1732 1032
```

Això indica que, després dels primers 1,000 símbols en el corpus objecte, veiem 442 diferents tipus de vocabularis, 331 d'ells ocorren només una sola vegada. Utilitzant la funció `summary` per obtenir una informació general de l'objecte, quantes mostres estan incloses en la corba de creixement:

```
> summary(Brown.emp.vgc)
zipfR object for vocabulary growth curve
1007 samples for N = 1000 ... 1006770
Spectrum elements included up to m = 1
```

Utilitzant les informacions obtingudes, podem dibuixar les corbes de V i V_1:



La línia més fina correspon a V_1. Observem que ambdues corbes creixen amb una velocitat bastant ràpida en l'inici, després es redueixen. A més, sembla que la corba de V_1 té una asímptota horitzontal, és a dir, tendeix a un valor.

En canvi, les corbes de creixement no són llises, ja que reflecteix totes les peculiaritats a causa de la distribució no aleatòria de paraules i textos en un corpus. Una corba més llisa pot ser obtinguda per una interpolació binomial que va proposar Baayen (2001).

La idea és: suposem que els $f(i, N_0)$ símbols de la paraula ω_i són distribuïts aleatòriament sobre els N_0 símbols d'un text donat. Dividim aquest text en dues parts: P_1 amb $N < N_0$ símbols i P_2 amb $N_0 - N$. La probabilitat tal que un particular símbol de ω_i ocurre en P_1 és N/N_0. El nombre de vegades que ω_i ocurre en P_1 és una variable aleatòria de binomial amb paràmetre $f(i, N_0)$ i N/N_0. LLavors

$$Pr(f_{N_0}(i, N) = m) = \binom{f(i, N_0)}{m} \left(\frac{N}{N_0}\right)^m \left(1 - \frac{N}{N_0}\right)^{f(i,N_0)-m},$$

on $f_{N_0}(i, N)$ és la freqüència de ω_i condicionada en la mostra gran de mida N_0.

Podrem obtenir les expressions de $\mathbb{E}[V_{N_0}(N)]$ i $\mathbb{E}[V_{N_0}(m, N)]$, la mida de vocabulari condicional i els elements d'espectre condicional per una mostra de mida N donat l'espectre de freqüència per N_0 símbols, analogament com en el model d'urna:

$$\begin{aligned}\mathbb{E}[V_{N_0}(m, N)] &= \sum_{i=1}^{V(N_0)} \binom{f(i, N_0)}{m} \left(\frac{N}{N_0}\right)^m \left(1 - \frac{N}{N_0}\right)^{f(i,N_0)-m} \\ &= \sum_{k \geq m} V(k, N_0) \binom{k}{m} \left(\frac{N}{N_0}\right)^m \left(1 - \frac{N}{N_0}\right)^{k-m},\end{aligned}$$

i

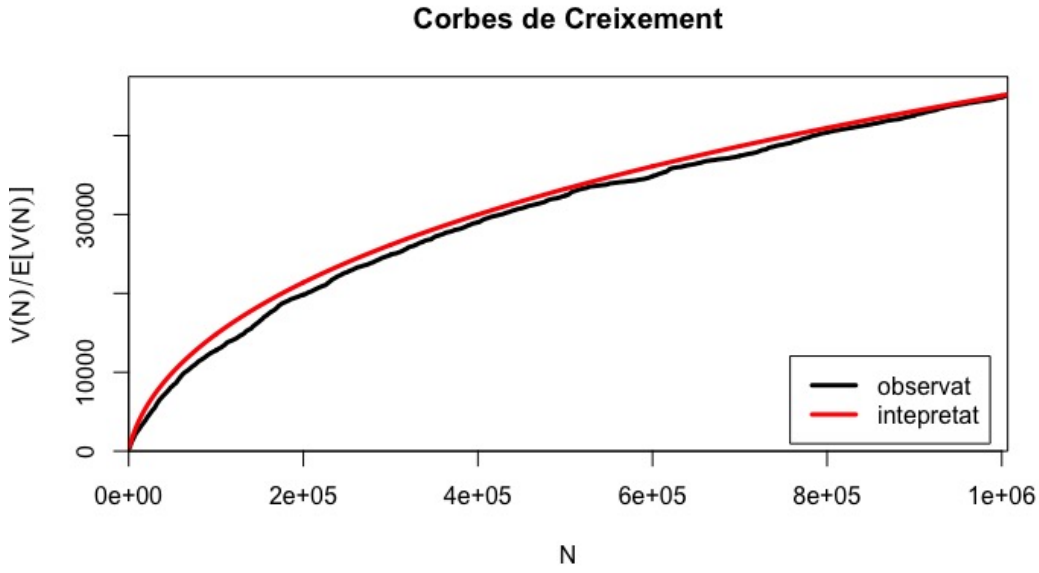
$$\begin{aligned}\mathbb{E}[V_{N_0}(N)] &= \sum_{i=1}^{V(N_0)} \left(1 - (1 - \frac{N}{N_0})^{f(i,N_0)}\right) = \sum_{m=1}^{N_0} V(m, N_0) \left(1 - (1 - \frac{N}{N_0})^m\right) \\ &= V(N_0) + \sum_{m=1}^{N_0} (-1)^{m-1} V(m, N_0) \left(\frac{N}{N_0} - 1\right)^m.\end{aligned}$$

Només volia il·lustrar les formules i no el tecnicisme.

Per tant, podem obtenir la corba de creixement interpolada a partir de la llista d'espectre de freqüència del CB:

```
> Brown.bin.vgc<-vgc.interp(Brown.spc, N(Brown.emp.vgc),m.max=1)
> head(Brown.bin.vgc)
     N         V        V1
1 1000  561.3796  464.2796
2 2000  990.4850  802.1776
3 3000 1371.5155 1090.0857
4 4000 1720.5995 1344.1823
5 5000 2045.5511 1573.1559
6 6000 2351.1401 1782.4742
```

A partir de les dades obtingudes, anem a dibuixarles:



La corba interpolada és més suau que l'empírica. Malgrat això, si les dades rellevants estan disponibles, és recomanable fer un cop d'ull a les corbes empíriques, ja que podrien revelar la presència de forts patrons no aleatoris en les dades que invaliden els supòsits en la base de l'estimació del model estadístic.

5.3 Els models LNRE

Ara bé, podem aplicar els models ZM-LNRE (Zipf-Mandelbrot LNRE model) i FZM-LNRE (finit Zipf-Mandelbrot LNRE model) per estimar els paràmetres del models i, també, veure la diferència entre dos model, ja que vam esmentar que una mida infinita de vocabulari no és realista per ZM-LNRE.

```
> summary(Brown.zm)
Zipf-Mandelbrot LNRE model.
Parameters:
Shape:            alpha = 0.5233222
Upper cutoff:         B = 0.002012022
[ Normalization:      C = 9.194397 ]
Population size: S = Inf
Sampling method: Poisson, approximations are allowed.

Parameters estimated from sample of size N = 1006770:
V       V1      V2      V3      V4      V5
Observed: 45215 19130.00 6458.00 3636.00 2301.00 1705.00 ...
Expected: 45215 23662.01 5639.58 2775.95 1718.78 1195.13 ...
```

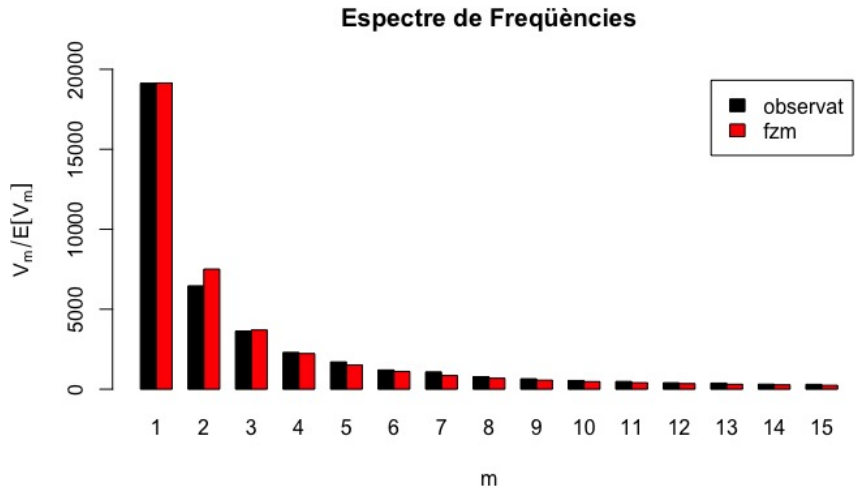
```
Goodness-of-fit (multivariate chi-squared test):
      X2 df p
8635.828 14 0


> summary(Brown.fzm)
finite Zipf-Mandelbrot LNRE model.
Parameters:
Shape:            alpha = 0.5999534
Lower cutoff:         A = 1.600374e-07
Upper cutoff:         B = 0.002496772
[ Normalization:      C = 4.49267 ]
Population size: S = 89168.36
Sampling method: Poisson, approximations are allowed.

Parameters estimated from sample of size N = 1006770:
V       V1      V2      V3   V4      V5
Observed: 45215 19130.00 6458.00 3636.00 2301 1705.00 ...
Expected: 45215 19141.63 7508.39 3701.65 2229 1515.99 ...

Goodness-of-fit (multivariate chi-squared test):
     X2 df             p
989.265 13 3.532034e-203
```

Fixem ambdos resums, veiem que tant el pendent com la cota superior són bastant similars. A més, el valor esperat del hàpax és el mateix a l'observat en tot cas. En canvi, el model FZM té unes estimacions més ajustes que el ZM i també té un $p - value$ suficientment petit que indica la seva significació. S'utilitza el test de χ^2 multivariant per la mesura de la bondat de l'ajust (Baayen, 2001, section 3.3). D'acord amb aquesta sortida, tenim la densitat segueix:

$$g(\pi) := \begin{cases} 4.4927 \cdot \pi^{-0.6000-1} & 1.6 * 10^{-7} \leq \pi \leq 0.0025 \\ 0 & \text{altre cas} \end{cases}$$

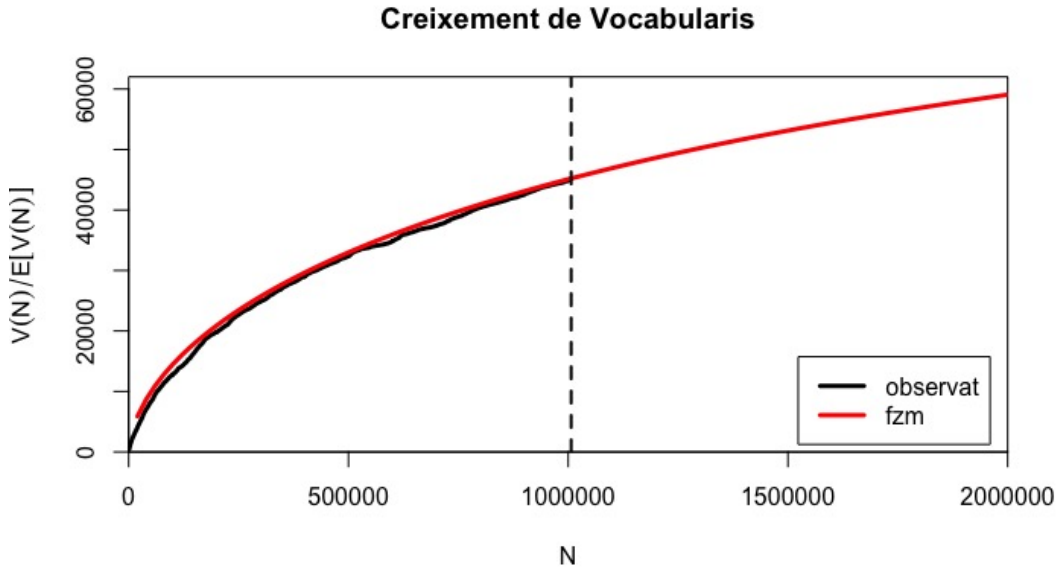
A més a més, el $p - value$ ens dóna la significació d'aquests paràmetres.

Ara bé, només fixem el model de FZM-LNRE. Basant aquest model, no existeix paraules amb probabilitat menys que $1.6 \cdot 10^{-7}$. Encara més, podem comparar els espectres de freqüències en un histograma:



Podem veure que la diferència entre elements espectrals observats i predits és petita en la majoria.

El model FZM ara es pot utilitzar per obtenir estimacions de V i V_m per mides de mostra arbitràries. Per exemple, podem generar una corba de creixement per $\mathbb{E}[V]$ per N igual a 2 milions:



Segons la figura anterior, podrem predir el nombre de vocabularis per una mostra de N igual a 2 milions.

6 Conclusions

L'objectiu d'aquest treball és introduir la llei de Zipf i explicar una de les seves aplicacions. Primer de tot, vam introduir el context històric d'aquesta llei. De seguida, vam estudiar algunes distribucions relacionades a la llei de Zipf com la llei potencial (la família general de la llei de Zipf), la distribució de Pareto i la distribució de Zeta. Un cop explicant l'origen de la llei Zipf, vam explicar algunes formulacions teòrics d'ella com el model *Random Typing* i la seva possible deficiència. En capítol 4 vam introduir el concepte de l'espectre de freqüències, els model LNRE i vam profunditzar un dels seus usos, la distribució de freqüències en textos, mitjançant l'anàlisi del Corpus Brown a través dels models LNRE en llenguatge R en l'últim capítol.

Referències

Arnold, B.C. (2015). *Pareto Distribution*. Wiley Online Library.

Baayen, R. Harald (2001). *Word Frequency Distributions*. Vol. 18. Springer Science & Business Media

Bax, S. (2014). "A proposed partial decoding of the Voynich script", *University of Bedfordshire, http://stephenbax. net/wp-content/uploads/2014/01/Voynich-a-provisionalpartial-decoding-BAX. pdf*.

Baroni, M. (2006). "39 Distributions in text". (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.79.7893)

Conrad, R. and Mitzenmacher, M. (2004). "Power laws for monkeys typing randomly: the case of unequal probabilities". In *IEEE Transactions on information theory*. Vol. 50, Num. 7, p. 1403-1414. IEEE.

Evert, Stefan. (2004). "Asimple LNRE model for random character sequences". In *proceedings of JADT*. Vol. 2004.

Evert, Stefan and Baroni, Marco (2007). "zipfR: Word frequency distributions in R". In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Session*, Prague, Czech Republic.

Fedorowicz, Jane (1982). "The theoretical foundation of Zipf's law and its application to the bibliographic database environment". In *Journal of the Association for Information Science and Technology*. Vol. 33, Num. 5, p. 285-293, Wiley Online Library.

Feller, W. (1971). *An introduction of Probability Theory and its Application*. ed. II, p. 50, New York: John Wiley & Sons.

Ferrer-i-Cancho, R. and McCowan, B. (2012). "The span of correlations in dolphin whistle sequences". In *Journal of Statistical Mechanics: Theory and Experiment*. Vol. 2012, Num. 06, p. P06002. IOP Publishing

Ferrer-i-Cancho, R. and Elvevåg, B. (2010). "Random texts do not exhibit the real Zipf's law-like rank distribution". In *PLoS One*. Vol. 5, Num. 3, p. e94111. Public Library of Science.

Hill, Bruce M. (1970). "Zipf's law and prior distributions for the composition of a population". In *Journal of the American Statistical Association*. Vol. 65, num. 331, p. 1200-1232. Taylor & Francis.

Hill, Bruce and Woodroofe, Michael (1975). "Stronger forms of Zipf's law". In *Journal of the American Statistical Association*. Vol. 70, Num. 349, p. 212-219. Taylor & Francis

Khmaladze, Estate V. (1988). "The statistical analysis of a large number of rare

events". In *Department of Mathmatical Statistics*. CWI.

Kleiber C. & Kotz S. (2003). *Statistical size distributions in economics and actuarial sciences*. (https://leseprobe.buch.de/images-adb/7d/4a/7d4a6289-48c1-4f39-aff0-87000a85af32.pdf) Vol. 470, p. 59-106. New York: John Wiley & Sons.

Koch, Richard (1998). *The 80/20 Principle: the Secreto of Achieving More with Less*. New York: Doubleday. ISBN 9780385491747

Mandelbrot, Benoit (1953). "An Information Theory of the Statistical Structure of Language". In *Communication theory*. Vol. 84, p. 486-502. Betterworth.

Miller, George A. (1957). "Some effects of intermittent silence". In *The American journal of psychology*. Vol. 70, Num. 2, p. 311-314. JSTOR.

Mitzenmacher, M. (2004). "A brief history of generative models for power law and lognormal distributions". In *Internet mathematics*. Vol. 1, Num. 2, p. 226-251. Taylor & Francis

Newman, M.E.J. (2005). "Power laws, Pareto distributions and Zipf's law". In *Conteporary Physics*. Vol. 46, Num. 5, p. 323-351. Taylor & Francis.

Price, Derek de Solla (1976). "A general theory of bibliometric and other cumulative advantage processes". In *Journal of the Association for Information Science and Technology*. Vol. 27, Num. 5, p. 292-306. Wiley Online Library

Reddy, S. and Knight, K. (2011). "What we know about the Voynich manuscript". In *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities*, Association for Computational Linguistics, p. 78-86.