Vanligaste orden- en fråga om logaritmer

George Zipf, en amerikansk språkforskare, upptäckte i slutet av 1940-talet mönster i hur vanliga olika ord är. Om man tar en text och räknar hur många förekomster som finns av varje ord får man en rangordning. I svenska språket är till exempel ordet ***och*** vanligast och sedan kommer ***i*** och ***att***.

Om vi tittar närmare på frekvensen för många ord kan man se ett mönster: när rangordningen fördubblas så halveras frekvensen ungefär. Det vanligaste ordet förekommer dubbelt så ofta som ordet på andra plats och ordet på andra plats förekommer dubbelt så ofta som ordet på fjärde plats osv.

Enligt Zipf är frekvensen alltså omvänt proportionell mot rangordningen. Zipf påstod också att detta gäller för alla språk. Vi kan alltså uttrycka detta som

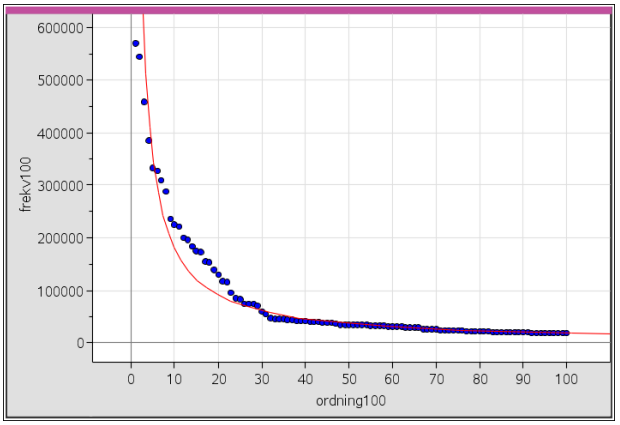
$f \approx \frac{k}{r}$ som kan skrivas som $f \approx k \cdot r^{-1}$

där r är rangordningen och k är någon konstant.

Vi ska här testa den saken på en lista med de vanligaste orden i svenska språket. Listan med ord kommer från en textdatabas som rymmer tjugo miljoner ord! Listan är ca 20 år och det har naturligtvis kommit en del nya ord men i början har det inte skett några större förändringar.

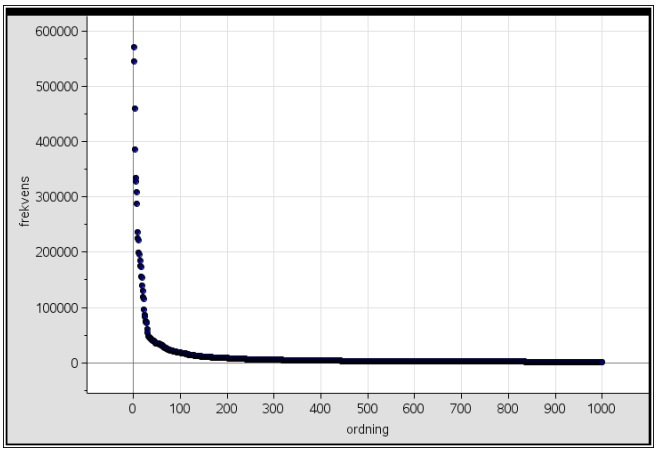
	A ordning	B frekvens	C orden
=			
1	1.	570041.	och
2	2.	544542.	i
3	3.	458872.	att
4	4.	385642.	det
5	5.	333312.	som
6	6.	327301.	en
7	7.	308952.	på
8	8.	287120.	är
9	9.	236020.	för
10	10.	224991.	av
11	11.	221132.	med

Vi ser att regeln inte verkar gälla för de allra vanligaste orden. Vi ritar först ett spridningsdiagram för de 100 första orden.

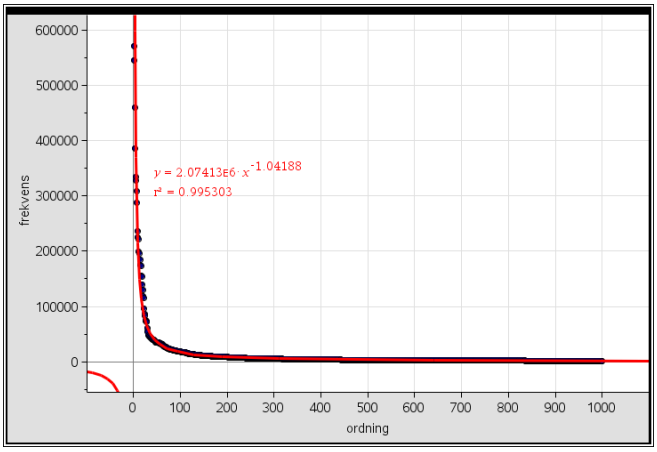


Här ser man tydligt avvikelsen i ett spridningsdiagram. Vi har lagt in en regressionskurva (potensform) för jämförelse.

hur ser det ut länge ner i listan? Vi ritar nu ett *spridningsdiagram* för hela listan med ordningstalet på den vågräta axeln och frekvensen på den lodräta.



Vi gör nu återigen en *regressionsanalys* på våra data. Y=a



Vi verkar få en bra anpassning efter denna modell. Vi får resultatet $f_r = 2.07413 \cdot 10^6 \cdot r^{-1.04188}$.

Många datapunkter ligger nära *x*-axeln så tar vi till ett trick. Vi *logaritmerar* vänster- och högerled i vår ekvation. Vi får:

$$\log f_r = \log(2.07413 \cdot 10^6 \cdot r^{-1.04188})$$

Första logaritmlagen ger:

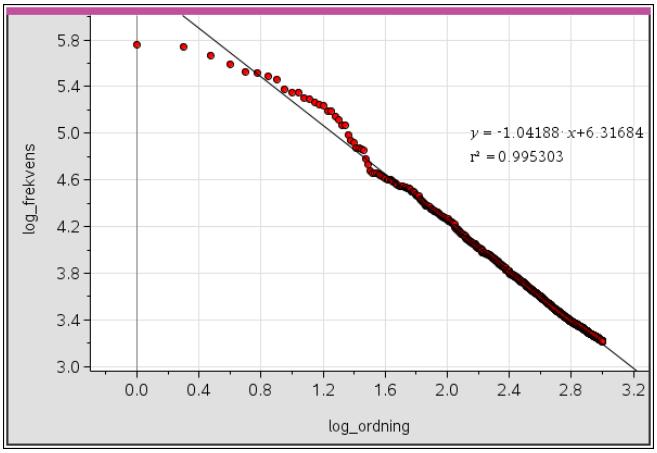
$$\log f_r = \log(2.07413 \cdot 10^6) + \log(r^{-1.04188})$$

Tredje logaritmlagen ger sedan:

$$\log f_r = \log(2.07413 \cdot 10^6) - 1.04188 \cdot \log(r)$$

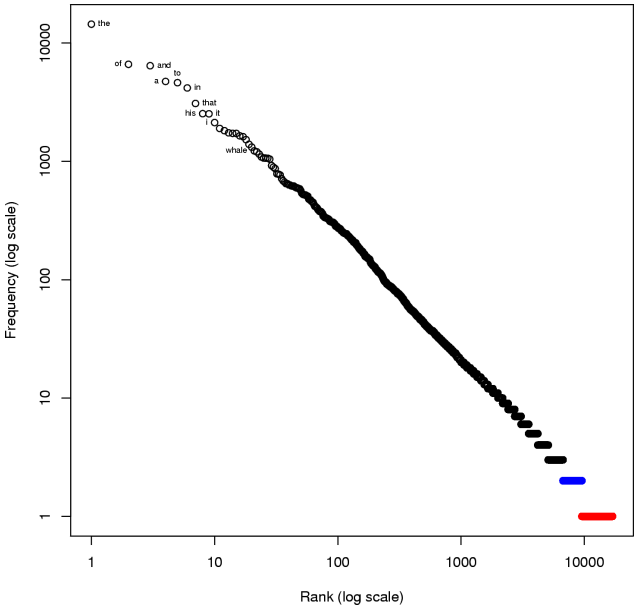
Detta betyder att om vi logaritmerar data i kolumnerna ordning och frekvens så kan vi anpassa efter ett linjärt samband. Se kalkylark och diagram på nästa sida.

	A ordning	B frekvens	C orden	D log_ordning	E log_frekvens	F
=				=log(ordning)	=log(frekvens)	
1	1.	570041.	och	0.	5.75591	
2	2.	544542.	i	0.30103	5.73603	
3	3.	458872.	att	0.477121	5.66169	
4	4.	385642.	det	0.60206	5.58618	
5	5.	333312.	som	0.69897	5.52285	
6	6.	327301.	en	0.778151	5.51495	
7	7.	308952.	på	0.845098	5.48989	
8	8.	287120.	är	0.90309	5.45806	
9	9.	236020.	för	0.954243	5.37295	
10	10.	224991.	av	1.	5.35217	
11	11.	221132.	med	1.04139	5.34465	



Det visar sig att denna regel också gäller för många andra saker, t.ex. inkomstfördelningar och storlek på städer t.ex. nu förtiden har man modifierat lagen en aning.

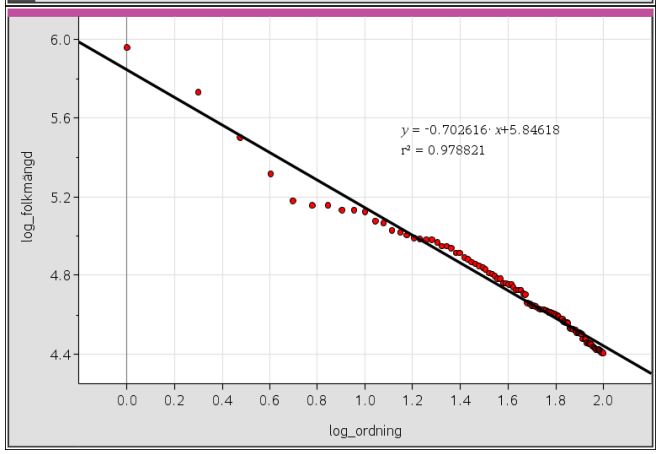
Här har vi ytterligare ett exempel på ord. Diagrammet visar ett log-log-diagram för orden i romanen *Moby Dick* för olika ord. Ordet *whale* kommer på 21:a plats.



Vi ska också titta närmare på de 100 största städerna i Sverige och göra samma analys som för vanligaste orden.

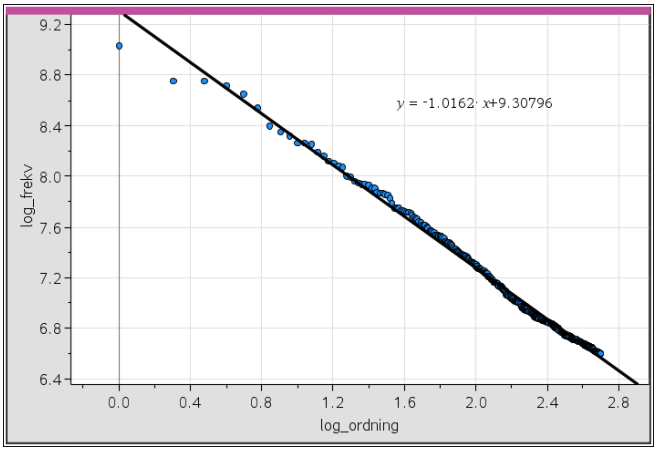
	A ordningstal	B kommun	C folkmängd	D log_ordning	E log_folkmängd	F
=				=log('ordningstal)	=log('folkmängd)	
1	1.	Stockholm	911989.	0.	5.95999	
2	2.	Göteborg	541145.	0.30103	5.73331	
3	3.	Malmö	318107.	0.477121	5.50257	
4	4.	Uppsala	207362.	0.60206	5.31673	
5	5.	Linköping	151881.	0.69897	5.1815	
6	6.	Västerås	143702.	0.778151	5.15746	
7	7.	Örebro	142618.	0.845098	5.15417	
8	8.	Helsingborg	135344.	0.90309	5.13144	
9	9.	Norrköping	135283.	0.954243	5.13124	
10	10.	Jönköping	132140.	1.	5.12103	
11	11.	Umeå	119613.	1.04139	5.07778	
12	12.	Lund	115968.	1.07918	5.06434	

D1 =0.



I problem 3 har vi nu en lista med engelska ord. Här är tanken att eleverna ska göra sin egen analys.

Så här blir slutresultatet.



Referenser:

https://shadycharacters.co.uk/2015/10/zipfs-law/
https://www.youtube.com/watch?app=desktop&v=fCn8zs912OEt%3D0m00s
https://www.wikiwand.com/sv/Zipfs_lag
https://www.sketchengine.eu/english-word-list/