**################################################################################################################## ZIPF LAW: Program Zipf.docx ################################################### #################################################################################################**

**## following program is stored in Zipf.docx**

**library(VGAM)**

**## needed for the Riemann zeta function**

**## to calculate the sum of the general harmonic series (p-series)**

**## generating data from a discrete Pareto distribution**

**set.seed(1) ## seed for the random number generator (for repeatability)**

**n=10000 ## consider max word frequency n**

**xx=dim(n)**

**prob=dim(n)**

**xx=c(1:n)**

**xx**

**alpha=1.25 ## selected alpha**

**for (i in 1:n) {**

**prob[i]=xx[i]^(-alpha)**

**}**

**prob=prob/sum(prob)**

**sum(prob)**

**plot(prob)**

**sampleDist = function(n) {**

**sample(xx, n, replace = T, prob)**

**}**

**N=100000 ## number of words in corpus: feel free to experiment with different numbers**

**x=sampleDist(N)**

**x**

**t=table(x)**

**dim(t)**

**freq=dim(dim(t))**

**for (i in 1 :dim(t)) {**

**freq[i]=t[[i]]**

**}**

**freq**

**xval=as.numeric(levels(factor(x)))**

**freq**

**xval**

**plot(freq~xval,ylab="number of words of given frequency",xlab="frequency of occurrence")**

**plot(freq~xval,xlim=c(0,50),ylab="number of words of given frequency",xlab="frequency of occurrence")**

**lnfreq=log(freq)**

**lnxval=log(xval)**

**plot(lnfreq~lnxval)**

**fit=lm(lnfreq~lnxval)**

**summary(lm(lnfreq~lnxval))**

**alphaLS=-fit$coef[2] ## LS estimate of alpha**

**alphaLS**

**plot(lnfreq~lnxval,ylab="logarithm of the number of words of given frequency",xlab="log frequency of occurrence")**

**abline(fit)**

**## omitting word frequencies that occur less than k times**

**k=2 ## this omits word frequencies that occur just once**

**ind=freq<2**

**yy=lnfreq[!ind]**

**xx=lnxval[!ind]**

**fit=lm(yy~xx)**

**summary(lm(yy~xx))**

**alphaLS=-fit$coef[2] ## LS estimate of alpha**

**alphaLS**

**plot(yy~xx,ylab="logarithm of the number of words of given frequency",xlab="log frequency of occurrence")**

**abline(fit)**

**## omitting word frequencies that occur less than k times**

**## maximum likelihood estimation**

**x**

**nuobs=length(x)**

**nuobs**

**sus=sum(log(x))**

**sus**

**obj=dim(200)**

**alpha=dim(200)**

**par=1.00**

**for (jjj in 1:200) {**

**par=par+0.01**

**alpha[jjj]=par**

**sum=zeta(par) ## calculates the Riemann zeta function**

**obj[jjj]=-nuobs\*log(sum)-par\*sus**

**}**

**alpha**

**obj**

**plot(obj~alpha,ylab="log-likelihood",xlab="alpha")**

**gg=obj==max(obj)**

**alpha[gg==TRUE] ## MLE of alpha**

**## using the R library poweRlaw**

**library(poweRlaw)**

**m = displ$new(x)**

**## estimate\_xmin(m)**

**## m$setXmin(1)**

**estimate\_pars(m)**

**## using the R library poweRlaw**

**## estimating xmin**

**estimate\_xmin(m)**

**m$setXmin(3)**

**estimate\_pars(m)**