APPENDIX

The R-Codes for various procedures that have been used in this thesis are presented in this Appendix. For a better understanding all codes are supplemented with an illustrative data set and hence required modification for data input in the respective form can be carried out for replication of each of these R-codes.

Chapter 2 – Classical Estimation

### Declaration of the 3 functions for each method: QH, GM, WALD, WALDCC, FS, WILSON.

### Define the vector of values from the matrix of given I x J Contingency table

```r
inpmat = c(56,72,73,59,62,87,58)
k = length(inpmat)
alpha=0.05

### QH METHOD
k = length(inpmat)
s = sum(inpmat)
chi = qchisq(1-alpha, df=k-1)
pi = inpmat/s
QH.UL = (chi + 2*inpmat + sqrt(chi*chi + 4*inpmat*chi*(1 - pi)))/(2*(chi+s))
QH.LL = (chi + 2*inpmat - sqrt(chi*chi + 4*inpmat*chi*(1 - pi)))/(2*(chi+s))
QH.WI = QH.UL - QH.LL #Length of the interval
QH.VL = prod(QH.WI)
round(cbind(QH.LL, QH.UL,QH.WI),3)
QH.VL
```

```r
### GM METHOD
k = length(inpmat)
s = sum(inpmat)
chi = qchisq(1-(alpha/k), df=1)
pi = inpmat/s
GM.UL = (chi + 2*inpmat + sqrt(chi*chi + 4*inpmat*chi*(1 - pi)))/(2*(chi+s))
GM.LL = (chi + 2*inpmat - sqrt(chi*chi + 4*inpmat*chi*(1 - pi)))/(2*(chi+s))
GM.WI = GM.UL - GM.LL #Length of the interval
round(cbind(GM.LL, GM.UL,GM.WI),3)
GM.VL = prod(GM.WI)
GM.VL
```
## WALDCC METHOD

\[ k = \text{length(inpmat)} \]

\[ s = \text{sum(inpmat)} \]

\[ \text{chi} = \text{qchisq}(1-\alpha, \text{df}=1) \]

\[ \text{pi} = \frac{\text{inpmat}}{s} \]

\[ \text{WALDCC.LL} = \text{pi} - \left( \sqrt{\text{chi} \times (\text{pi} \times (1-\text{pi})/s)} \right) - \left( 1/(2 \times s) \right) \]

\[ \text{WALDCC.UL} = \text{pi} + \left( \sqrt{\text{chi} \times (\text{pi} \times (1-\text{pi})/s)} \right) + \left( 1/(2 \times s) \right) \]

\[ \text{WALDCC.WI} = \text{WALDCC.UL} - \text{WALDCC.LL} \]

\[ \text{round} (\text{cbind(WALDCC.LL, WALDCC.UL, WALDCC.WI),3}) \]

\[ \text{WALDCC.VL} = \text{prod}(\text{WALDCC.WI}) \]

## WALD METHOD

\[ k = \text{length(inpmat)} \]

\[ s = \text{sum(inpmat)} \]

\[ \text{chi} = \text{qchisq}(1-\alpha, \text{df}=1) \]

\[ \text{pi} = \frac{\text{inpmat}}{s} \]

\[ \text{WALD.LL} = \text{pi} - \left( \sqrt{\text{chi} \times (\text{pi} \times (1-\text{pi})/s)} \right) \]

\[ \text{WALD.UL} = \text{pi} + \left( \sqrt{\text{chi} \times (\text{pi} \times (1-\text{pi})/s)} \right) \]

\[ \text{WALD.WI} = \text{WALD.UL} - \text{WALD.LL} \]

\[ \text{round} (\text{cbind(WALD.LL, WALD.UL, WALD.WI),3}) \]

\[ \text{WALD.VL} = \text{prod}(\text{WALD.WI}) \]

## FS METHOD

\[ k = \text{length(inpmat)} \]

\[ s = \text{sum(inpmat)} \]

\[ zval = \text{abs(qnorm}(1-(\alpha/2)) \]

\[ \text{pi} = \frac{\text{inpmat}}{s} \]

\[ \text{FS.LL} = \text{pi} - \left( \frac{zval}{2 \times \sqrt{s}} \right) \]

\[ \text{FS.UL} = \text{pi} + \left( \frac{zval}{2 \times \sqrt{s}} \right) \]

\[ \text{FS.WI} = \text{FS.UL} - \text{FS.LL} \]

\[ \text{round} (\text{cbind(FS.LL, FS.UL, FS.WI),3}) \]

\[ \text{FS.VL} = \text{prod}(\text{FS.WI}) \]

## WS METHOD

\[ k = \text{length(inpmat)} \]

\[ s = \text{sum(inpmat)} \]

\[ \text{chi} = \text{qchisq}(1-\alpha, \text{df}=1) \]

\[ \text{pi} = \frac{\text{inpmat}}{s} \]

\[ \text{WS.UL} = \frac{\text{chi} + 2 \times \text{inpmat} + \sqrt{\text{chi} \times \text{chi} + 4 \times \text{inpmat} \times \text{chi} \times (1 - \text{pi})}}{2 \times (\text{chi} + s)} \]

\[ \text{WS.LL} = \frac{\text{chi} + 2 \times \text{inpmat} - \sqrt{\text{chi} \times \text{chi} + 4 \times \text{inpmat} \times \text{chi} \times (1 - \text{pi})}}{2 \times (\text{chi} + s)} \]

\[ \text{WS.WI} = \text{WS.UL} - \text{WS.LL} \]

\[ \text{WS.VL} = \text{prod}(\text{WS.WI}) \]
round(cbind(WS.LL, WS.UL, WS.WI), 3)
WS.VL = prod(WS.WI)
WS.VL

## CONSOLIDATED RESULTS
round(cbind(WALD.LL, WALD.UL, WALDCC.LL, WALDCC.UL, WS.LL, WS.UL, QH.LL, QH.UL, GM.LL, GM.UL, FS.LL, FS.UL), 3)
round(cbind(WALD.WI, WALDCC.WI, WS.WI, QH.WI, GM.WI, FS.WI), 3)
cbind(WALD.VL, WALDCC.VL, WS.VL, QH.VL, GM.VL, FS.VL)

-----------------------------------------------

SG- code

t1 = proc.time()

sgp = function(c)
{
  #x = c(41, 5, 33)
  #x = c(11, 19, 30, 58, 67, 92, 118, 173, 297)
  #x = c(56, 72, 73, 59, 62, 87, 58)
  #x = c(0, 8, 15, 5)  ## INPUT as 1xk
  s = sum(x)  ## SUM(Cell_Counts)
  k = length(x)
  #c = 2
  alpha = 0.05

  b = x - c
  a = x + c

  ### FINDING FACTORIAL MOMENTS-TRUNCATED POISSON

  # fm1 = x[1]*(ppois(a[1]-1,x[1])-ppois(b[1]-2,x[1]))/(ppois(a[1],x[1])-ppois(b[1]-1,x[1]))
  fm1 = 0
  fm2 = 0
  fm3 = 0
  fm4 = 0
  for (i in 1:k)
  {
    fm1[i] = x[i]*(ppois(a[i]-1,x[i])-ppois(b[i]-2,x[i]))/(ppois(a[i],x[i])-ppois(b[i]-1,x[i]))
    fm2[i] = x[i]^2*(ppois(a[i]-2,x[i])-ppois(b[i]-3,x[i]))/(ppois(a[i],x[i])-ppois(b[i]-1,x[i]))
    fm3[i] = x[i]^3*(ppois(a[i]-3,x[i])-ppois(b[i]-4,x[i]))/(ppois(a[i],x[i])-ppois(b[i]-1,x[i]))
    fm4[i] = x[i]^4*(ppois(a[i]-4,x[i])-ppois(b[i]-5,x[i]))/(ppois(a[i],x[i])-ppois(b[i]-1,x[i]))
  }
}
## FINDING CENTRAL MOMENTS - TRUNCATED POISSON

\[ m_1 = 0 \]
\[ m_2 = 0 \]
\[ m_3 = 0 \]
\[ m_4 = 0 \]
\[ m_{4t} = 0 \]

for (i in 1:k)
{
  \[ m_{1i} = fm_{1i} \]
  \[ m_{2i} = fm_{2i} + fm_{1i} - (fm_{1i} \times fm_{1i}) \]
  \[ m_{3i} = fm_{3i} + fm_{2i} \times (3 - (3 \times fm_{1i})) + (fm_{1i} \times (3 \times fm_{1i} \times fm_{1i})) + (2 \times fm_{1i}^3) \]
  \[ m_{4i} = fm_{4i} + fm_{3i} \times (-12 \times fm_{1i}) + (6 \times fm_{1i}^2) + (6 \times fm_{1i}^3) + (3 \times fm_{1i}^4) \]
  \[ m_{4t_i} = m_{4i} - 3 \times m_{2i}^2 \] # Temporary Variable for next step
}

\[ s_1 = \text{sum}(m_{1}) \]
\[ s_2 = \text{sum}(m_{2}) \]
\[ s_3 = \text{sum}(m_{3}) \]
\[ s_4 = \text{sum}(m_{4t}) \]

## FINDING GAMMAS ---> EDGEOGHRT EXPANSION

\[ g_1 = \frac{s_3}{s_2^{3/2}} \]
\[ g_2 = \frac{s_4}{s_2^2} \]

## FINDING CHEBYSHEV-HERMITE POLYNOMIALS ---> EDGEOGHRT EXPANSION

\[ z = (s - s_1) / \sqrt{s_2} \]
\[ z^2 = z^2 \]
\[ z^3 = z^3 \]
\[ z^4 = z^4 \]
\[ z^6 = z^6 \]

\[ \text{poly} = 1 + g_1 \times (z^3 - (3 \times z))/6 + g_2 \times (z^4 - (6 \times z^2) + 3)/24 + (g_1^2) \times (z^6 - (15 \times z^4) + (45 \times z^2) - 15)/72 \]
\[ f = \text{poly} \times \text{exp}(-z^2/2)/\sqrt{2 \times \pi} \]

## FINDING PROBABILITY FUNCTION BASED ON 'c'

\[ pc = 0 \]

for (i in 1:k)
{
  \[ pc[i] = \text{ppois}(a[i], x[i]) - \text{ppois}(b[i] - 1, x[i]) \]
}

\[ pcp = \text{prod}(pc) \] # PRODUCT OF pc THAT HAS k ELEMENTS
\[ pps = 1/\text{dpois}(s, s) \] # POISSON PROB FOR s WITH PARAMETER AS s
\[ rp = pps \times pcp \times f/\sqrt{s_2} \] # REQUIRED PROBABILITY

rp
SG- Result

t=proc.time()

y=0
#x=c(41,5,33)
#x=c(3,8,10,5)
x=c(56,72,73,59,62,87,58)
#x=c(11,19,30,58,67,92,118,173,297)
s=sum(x)
alpha=0.05

if (min(x) > 5)
{
    M1=1
    M2=max(x-5)  #AS PER MAY _ JOHNSON PAPER
    #M2=s
    c=M1:M2
}

M=length(c)
for (i in 1:M)
{
    y[i]=round(sgp(c[i]) ,4)
}
#y
j=1
vc=0
while(j<=M){
    if (y[j]<1-alpha && 1-alpha < y[j+1])
        vc=j
    else
        vc=vc
    j = j+1
}
vc  ##REQUIRED VALUE OF C
delta=((1-alpha)-y[vc])/(y[vc+1]-y[vc])

##FINDING LIMITS
sp=x/s  #SAMPLE PROPORTION
LL=round(sp-(vc/s),4)  #LOWER LIMIT
UL=round(sp+(vc/s)+(2*delta/s),4)  #UPPER LIMIT

LLA=0
ULA=0
for (r in 1:length(x))
{
if ( LL[r]< 0) LLA[r] = 0 else LLA[r]=LL[r]
if (UL[r] > 1) ULA[r] = 1 else ULA[r]=UL[r]
}
cbind(LL,UL)

cbind(LLA,ULA)

diA=ULA-LLA  ##FIND LENGTH OF CIs
VOL=round(prod(diA),8)  ##PRODUCT OF LENGTH OF CIs
t=proc.time()
y=0
#x=c(41,5,33)
#x=c(3,8,10,5)
x=c(56,72,73,59,62,87,58)
#x=c(11,19,30,58,67,92,118,173,297)
s=sum(x)
alpha=0.05

if (min(x) > 5)
{
    M1=1
    M2=max(x-5)
    #M2=s  #AS PER MAY _ JOHNSON PAPER
    c=M1:M2
}
M=length(c)
for (i in 1:M)
{
y[i]=round(sgp(c[i]) ,4)
}
#y
j=1
vc=0
while(j<=M){
    if (y[j]<1-alpha && 1-alpha < y[j+1])
        vc=j  else
        vc=vc
    j = j+1
}
vc  ##REQUIRED VALUE OF C
delta=((1-alpha)-y[vc])/(y[vc+1]-y[vc])

##FINDING LIMITS
sp=x/s  #SAMPLE PROPORTION
LL = round(sp - (vc/s), 4) # LOWER LIMIT
UL = round(sp + (vc/s) + (2*delta/s), 4) # UPPER LIMIT

LLA = 0
ULA = 0

for (r in 1:length(x))
{
    if (LL[r] < 0) LLA[r] = 0 else LLA[r] = LL[r]
    if (UL[r] > 1) ULA[r] = 1 else ULA[r] = UL[r]
}
cbind(LL, UL)
cbind(LLA, ULA)

diA = ULA - LLA  # FIND LENGTH OF CIs
VOL = round(prod(diA), 8)  # PRODUCT OF LENGTH OF CIs
VOL

proc.time() – t

Chapter 3 – Bayesian Estimation

Bayes estimation with different hyper priors

### Initialize this before start new set of run
m = 0
l = 0
u = 0
diff = 0
g = c(1, 1, 2, 3, 4, 5, 6, 7, 8, 9, 12, 13, 14, 15, 16, 19)  # Given data set
s = sum(g)
k = length(g)
d = 5
s1 = floor(k/d)
d1 = runif(s1, 0, 1)  # First half of the vector
d2 = runif(k - s1, 1, 2)  # Second half of the vector
a = c(d1, d2)
p = g + a  # Prior for Dirichlet
dr = rdirichlet(10000, p)  # Posterior

for(j in 1:k)
Bayesian estimation with equal hyper priors

library(MCMCpack)
x=c(56,72,73,59,62,87,58) ####Input matrix
k=length(x)
n_r=10000 ##NO OF RUNS FOR SIMULATION OF POSTERIOR
prio=1 ##PRIOR WITH EQUAL VALUES FOR DIRICHLET
PARAMETERS
po=x+prio ##POSTERIOR DIRICHLET PARAMETERS
dr=rdirichlet(n_r,po)
a=0
l=0
u=0
dif=0
for(j in 1:k)
{
a[j]=round(mean(dr[,j]),4)
l[j]=round(quantile(dr[,j],0.025),4)
u[j]=round(quantile(dr[,j],0.975),4)
dif[j]=u[j]-l[j]
}
v=prod(dif)

Chapter 4 – Classical Testing of Hypotheses

fnRAP <-
function(a,nr,nc)
{
### nr = number of rows in original matrix
### nc = number of cols in original matrix
tot_tab=0
if ( (nr == 2 && nc > 2) || (nr > 2 && nc == 2) || (nr > 2 && nc > 2) ) {
mat=matrix(a,nr,nc, byrow=TRUE)
outd = as.data.frame(t(c(dim(mat), toString(c(1, 2, 3, 4, 5)), toString(c(1, 2, 3, 4, 5)), round(chisq.test(mat)$p.value, 4), chisq.test(mat)$p.value < 0.05)))
ori_con = round(chisq.test(mat)$p.value, 4) # ms_code-for original conclusion: test result of given matrix
outlist=list()

for (i in c(2:nr))
{
  outlist[[i]]=list()
  for (j in c(2:nc))
  {
    outlist[[i]][[j]]=list()
    rpermmat = t(combn(c(1:nr),i))
cpermmat = t(combn(c(1:nc),j))

    if(i==nr) rfact = 1 else rfact = factorial(nr)/(factorial(i) * factorial(nr-i))
    if(j==nc) cfact = 1 else cfact = factorial(nc)/(factorial(j) * factorial(nc-j))

    if(i==nr & & j==nc) {
      names(outd) = c("No. of rows", "No. of cols", "Selected rows", "Selected cols", "Pvalue", "Pvalue significant at 5%?")
      outlist[[i]][[j]][[1]] = mat
      outlist[[i]][[j]][[2]] = outd
    }
    else {
      for (ii in c(1:rfact))
      for (jj in c(1:cfact))
      {
        tempvar = round(chisq.test(mat [ c(rpermmat[ii,]), c(cpermmat[jj,]) ])$p.value, 4)
        if ((ori_con < 0.05 & & tempvar > 0.05) || (ori_con > 0.05 & & tempvar < 0.05)) #ms_code-for checking the reversal in conclusions
        {
          outlist[[i]][[j]][(ii-1)*cfact + jj] = tempvar
          outd = rbind(outd, t(c(dim(mat [ c(rpermmat[ii,]), c(cpermmat[jj,]) ])), toString(c(rpermmat[ii,])), toString(c(cpermmat[jj,])), tempvar, tempvar < 0.05)))
        }
      }
    }
  }
}

A-9
Chapter 5 – Bayesian Testing of Hypotheses.

#####Input the size of the contingency table
r=4  #no of rows
c=4  #no of columns

#####Input the original cell counts as a row vector
ns=c(1,3,10,6,2,3,10,7,1,6,14,12,0,1,9,11) #Cell counts row wise
data=matrix(ns,r,c,byrow=TRUE)

rt=0 #Calculating row totals
for (i in 1:r)
{
  rt[i]=sum(data[i,])
}

cr=0 #Calculating column totals
for (j in 1:c)
{
  ct[j]=sum(data[,j])
}
n=sum(ns)  #Grand Total

alp = 0
k=r*c
for (k1 in 1:k) #Prior for Dirichlet parameters-No association Model
{
  alp[k1]=0.5
}
a=sum(alp)

gam = 0
for (k2 in 1:r) #Prior for Dirichlet parameters-Association Model-Row
{
  gam[k2]=0.5
}
g=sum(gam)

del = 0
for (k3 in 1:c) #Prior for Dirichlet parameters-Association Model-Col
{
  del[k3]=0.5
}
d=sum(del)

# TERMS FOR LOG BAYES FACTOR FORMULA

## NUMERATOR AND DENOMINATOR OF LOG BAYES FACTOR FORMULA

nr=sum(lgamma(t1))+lgamma(a)+sum(lgamma(gam))+sum(lgamma(del))+lgamma(t2)+lgamma(t3)
dr=sum(lgamma(alp))+lgamma(g)+lgamma(d)+sum(lgamma(t4))+sum(lgamma(t5))+sum(lgamma(t6))
lbf=nr-dr #####LOG BAYES FACTOR B10
lbf
lbf1=-lbf #####LOG BAYES FACTOR B01
lbf1
exp(lbf) #####BAYES FACTOR B10
exp(lbf1) #####BAYES FACTOR B01
Acceptance mail_JDS-1248

jds-editor <jds-editor@nuk.edu.tw>

to me

Dear Dr. U. Sangeetha

I am happy to inform you that your article, "Sensitivity Analysis of Bayes Factor for Categorical Data with Emphasis on Sparse Multinomial Data" (JDS-1248) is accepted for publication by Journal of Data Science. I attach a "permission to publish" form for you to approve. This means we do not want your copyright, but merely ask your permission for us to print your article. We believe your effort to write this article is a lot more than the effort we review your article so it is simply not fair for us to ask you to transfer your copyright.

Please send your e-files (including the original Word or TEX and all *.eps or *.ps graphic files) to us. If you have done so, please ignore this message.

Thank you for your contribution to Journal of Data Science and we look forward to receiving further submissions from you.

Sincerely yours,

Wen-Jang Huang

================================
Prof. Wen-Jang Huang,
Editor, Journal of Data Science
Department of Applied Mathematics
National University of Kaohsiung
Kaohsiung, 81148, Taiwan, R.O.C.
E-mail: jds-editor@nuk.edu.tw
Web site: http://www.jds-online.com/
SENSITIVITY ANALYSIS OF BAYES FACTOR FOR CATEGORICAL DATA WITH
EMPHASIS ON SPARSE MULTINOMIAL DATA

U. Sangeetha\textsuperscript{a}, M. Subbiah\textsuperscript{b}, M.R. Srinivasan\textsuperscript{c}, B. Nandram\textsuperscript{d},

\textsuperscript{a} Department of Management Studies, SSN College of Engineering, Chennai. Email: usangee19@gmail.com
\textsuperscript{b} Department of Mathematics, L. N. Government College, Ponneri. Email: sisufive@gmail.com
\textsuperscript{c} Department of Statistics, University of Madras, Chennai. Email: mrsrin8@gmail.com
\textsuperscript{d} Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609, United States. Email: balnan@wpi.edu

Contact person: U. Sangeetha, Email: usangee19@gmail.com

Phone: 00-91-95000-79439

Running title: Bayes factors for Categorical data
SENSITIVITY ANALYSIS OF BAYES FACTOR FOR CATEGORICAL DATA WITH EMPHASIS ON SPARSE MULTINOMIAL DATA

Abstract
This article considers hypothesis testing using Bayes factor in the context of categorical data models represented in two-dimensional contingency tables. The study includes multinomial model for a general $I \times J$ table data. Other data characteristics such as low as well as polarized cell counts and size of the tables are also considered. The objective is to investigate the sensitivity of Bayes factor taking these features into account so as to understand the performance of non-informative priors itself. Consistency has been studied based on different types of data and using Dirichlet prior with eight different choices for multinomial model followed by a bootstrap simulation. Study has emphasized the reasonable choice of values for the parameters that normally represents the underlying physical phenomena, though partially vague in nature.

Keywords
Bayesian methods, Categorical data, Sparseness, Statistical computing.

1. Introduction
Categorical data are generally presented in a contingency table of size $I \times J$ with $I$ representing the number of rows and $J$ as number of columns. In most of the survey sampling, cell counts in $I \times J$ table follow a multinomial distribution and Chi square test of association is widely applied to understand the association between the variables. However, in practice, it is impossible to have an agreement about the appropriateness of the model to be used unless there is a well-
established theoretical framework or a mechanism underlying the problem. Therefore, it is imperative to take into account uncertainties in the model-building process, and so to start with, a set of competing models has to be considered with each model viewed as a different state of a random variable.

Marden (2000) has emphasized the need to expand the scope of hypotheses testing beyond p values and noted that Bayesian inference with automatic computing methods will be a promising approach. Statistical inference on parameter estimation based on Bayesian approach is conceptually straightforward in most of the problems. Once a prior distribution is defined and with a reasonable likelihood function, inference about parameters of interest is obtained from marginal posterior distributions and nuisance parameters are integrated. The role of probability in measuring the uncertainty at each stage of Bayesian estimation process is well established. In order to arrive at the posterior distribution of possible states of the model, prior distribution of each model is updated using the information contained in the data. Now, the inference is drawn from the entire posterior distribution of the most plausible model.

Johnson (2005) proposed methods based on Bayes factors for modeling the sampling distributions of standard test statistics. The study indicated the possible extensions to test statistics associated with categorical data. However, influence of priors and nature of subjectivity affects the sensitivity of results in a significant way. Vanpaemel (2010) has provided exhaustive list of studies that address the sensitivity of priors and has emphasized the importance of prior in sensitivity analysis. Particularly Hashemi (1997) and Nandram and Choi (2007) separated by a decade but similar in approach, have provided motivation to investigate deeply the Bayes factor
for categorical data under multinominal design, which is ubiquitous in many social science survey sampling designs and problems.

The above studies provided varying recommendations in literature on the use of Bayes factors for contingency tables based on the study design and size of the table (Upton, 1982; 1992). Recently, Hitchcock (2009) emphasized that hypotheses testing problems related to $I \times J$ tables require elaborate studies in understanding the elegance of Yates correction. Mirkin (2001) listed different ways to look at Chi square tests and observed that it is an appealing measure for studying the association. The relevance of study design and analyses based on $2 \times 2$ tables too need more careful attention (Upton, 1982; Campbell, 2007).

In this paper, algebraic form of Bayes factor, in the context of categorical data represented in two-dimensional contingency tables has been presented based on the underlying sampling design. The multinominal model for a general $I \times J$ table with two competing models of ‘no association’ and ‘association’ between categorical variables has been considered. The important aspects of this study relate to multinominal models on categorical data (i) to understand the sensitivity of priors and (ii) to incorporate the sparseness of data in the case of higher order $I \times J$ tables with zero and / or positive low counts, polarized cell counts and size of the tables. The sensitivity of priors has witnessed an active discussion in Bayes factor applications in both designs and sparseness of data has been well incorporated for $2 \times 2$ tables (Subbiah and Srinivasan, 2008).
In the following Section, a quick overview of Bayes factor in general and statistical details for categorical data is presented. In Section 3, comparative analysis based on data collected from literature that exemplify the above listed features are presented followed by a study using bootstrap simulation based on the data extracted and Section 4 has concluding remarks.

2. Bayes Factor

If there are several competing hypotheses or models about a system, then the set of models can be considered as mutually exclusive and exhaustive. A prior probability $p(H_i)$ ($i=1,2,\ldots,N$) can be assigned to each hypothesis such that $\sum p(H_i) = 1$, with $N$ denoting the number of hypotheses. After observing data $y$, the posterior probability of hypothesis $H_i$ is

$$p(H_i | y) = \frac{\sum_{i=1}^{N} p(H_i)p(y | H_i)}{\sum_{i=1}^{N} p(H_i)p(y | H_i)}$$

where $p(y|H_i)$ is the marginal density which is the expected value of all possible likelihoods.

Then hypothesis $i$ relative to $j$ is of the form

$$\frac{p(H_i | y)}{p(H_j | y)} = \frac{\sum_{i=1}^{N} p(H_i)p(y | H_i)}{\sum_{i=1}^{N} p(H_i)p(y | H_i)} = \frac{p(H_i)}{p(H_j)} \frac{p(y | H_i)}{p(y | H_j)}$$

It could be observed that the posterior odds ratio is the product of the prior odds ratio and ratio of the marginal probabilities under each of the hypotheses. Then Bayes factor $B_{ij}$ is defined as
\[ B_{ij} = \frac{p(y | H_i)}{p(y | H_j)} = \frac{p(H_i | y)}{p(H_j | y)} = \frac{\text{posterior odds}}{\text{prior odds}} \]

\( B_{ij} \) is not affected by prior specifications and \( B_{ij} > 1 \) can be interpreted as the hypothesis \( H_i \) to be more plausible than \( H_j \) in the light of \( y \). However, the above interpretation holds only when both \( H_i \) and \( H_j \) are simple hypotheses. Berger and Delampady (1987), Kass (1993), Bernardo and Smith (1994), Kass and Raftery (1995), Goodman (1999), Delampady and Berger (1990), Lavine and Schervish (1999), Ghosh et al (2006) provide a better insight into the concept of Bayes factor. The explicit forms of Bayes factor for multinomial model is derived in the following subsection.

### 2.1 Bayes Factors for I × J multinomial model

In the case of general I × J tables, if \( x_{ij} \) (\( i=1,2,\ldots,I; j=1,2,\ldots,J \)) denotes the observed cell counts, with \( r_i = \sum_j x_{ij} \) is the row total \( c_j = \sum_i x_{ij} \) is the column total and \( n = \sum \sum x_{ij} \) is the grand total then the Multinomial likelihood is

\[
f(X | \theta) = \frac{n!}{\prod_i \prod_j x_{ij}} \prod_i \prod_j \theta_{ij}^{x_{ij}} \quad \text{and} \quad \sum_i \sum_j \theta_{ij} = 1.
\]

Also, the conjugate prior (Gelman et al, 2002) for the proportion parameter vector \( \theta = (\theta_{ij}) \) could be a multivariate generalization of Beta distribution known as Dirichlet \( (\alpha_{ij}) \) with \( \alpha_{ij} > 0 \) and density function is

\[
\pi(\theta) = \frac{\Gamma(\alpha)}{\prod_i \prod_j \Gamma(\alpha_{ij})} \prod_i \prod_j \theta_{ij}^{\alpha_{ij}-1} \quad \text{where} \quad \alpha = \sum_i \sum_j \alpha_{ij}.
\]
The pervasive inferential problem related to a categorical data summarized in contingency tables is testing the statistical independence of two categories of the categorical data. Model $H_0$ corresponds to the null hypothesis that there is no association between the two categories whereas Model $H_1$ takes that there is an association between the categories constituting $I \times J$ contingency table.

Then under $H_0$, the prior distribution $\pi_0(\theta)$ for the parameter $\theta = (\theta_{ij})$ is based on the law of independence $\theta_{ij} = \Pi_i \Psi_j$ where

$$\Pi_i = \text{Dirichlet}(\gamma_i) \text{ and } \Psi_j = \text{Dirichlet}(\delta_j)$$

Also for the prior $\pi_1(\theta)$ for model $H_1$ is $\theta = (\pi_{ij}) \sim \text{Dirichlet}(\alpha_{ij})$. Hence the marginal likelihood under the model $M_t$ ($t = 0, 1$) is

$$p(X | H_t) = \int f(X | \theta) \pi_t(\theta) d\theta$$

After suitable integration,

$$p(X | H_1) = \frac{n!}{\prod \prod x_{ij}!} \frac{\prod \prod \Gamma(n_{ij} + \alpha_{ij})}{\Gamma(n + \alpha)} \frac{\Gamma(\alpha)}{\prod \prod \Gamma(\alpha_{ij})}$$

$$p(X | H_0) = \frac{n!}{\prod \prod x_{ij}!} \frac{\Gamma(\gamma) \Gamma(\delta)}{\prod \Gamma(\gamma_i) \prod \Gamma(\delta_j)} \frac{\prod \Gamma(r_{ij} + \gamma_i) \prod \Gamma(c_{ij} + \delta_j)}{\Gamma(n + \gamma) \Gamma(n + \delta)},$$

where $\gamma = \Sigma \gamma_i; \delta = \Sigma \delta_j$

Hence, the Bayes factor for comparing these two models is

$$B_{01} = \frac{p(X | H_0)}{p(X | H_1)} = \frac{\prod \prod \Gamma(\alpha_{ij}) \Gamma \gamma \Gamma \delta \Gamma(n + \alpha) \prod \Gamma(r_{ij} + \gamma_i) \prod \Gamma(c_{ij} + \delta_j)}{\prod \prod \Gamma(n_{ij} + \alpha_{ij}) \Gamma \alpha \prod \Gamma \gamma_i \prod \Gamma \delta_j \Gamma(n + \gamma) \Gamma(n + \delta)}$$
However computing $B_{01}$ on log scale will alleviate the problem of overflow that may occur if it is computed directly. Kass and Raftery (1995) have provided appropriate guidelines for interpreting $B_{01}$ and $\log(B_{01})$ as the degree of evidence for $H_0$ and is as follows;

- $1 < B_{01} < 3$ indicates ‘$H_0$ is not worth more than a bare mention’
- $3 < B_{01} < 20$ indicates ‘$H_0$ is positive’
- $20 < B_{01} < 150$ indicates ‘strong evidence for $H_0$’
- $150 < B_{01}$ indicates ‘very strong evidence for $H_0$’

3. Comparative Data Analysis

Statistical literature on theory and applications of inferential procedures associated with contingency tables of two-dimensional categorical data provide a few essential characteristics necessary for comparative studies. These include order of tables (k=IJ), sample size (n), zero counts, notable polarized cell counts and positive low counts (cell counts not more than 6).

All $I \times J$ data sets are extracted from Agresti (2002), a classical book for social sciences that illustrate most of the issues in categorical data. In line with the above features, tables vary in sizes from $3 \times 3$ to $6 \times 4$ rectangular tables; 5% to 43% of the cells have low counts; grand total spans over a range of 96 to 3600; cell counts vary from 14 to 711. Also, to quantify the polarized cell counts a metric $v = \text{range}/k$ has been used to indicate the nature of distribution of values. The summary of the selected characteristics for the data sets are presented in Table 1.

Table 1

It is necessary to consider three cases of Dirichlet parameters as prior distributions for multinomial model; $\alpha_{ij}$ for model $H_1$ and $\gamma_i$’s and $\delta_j$’s for model $H_0$. In all the cases, an equal value for these parameters has been considered with eight choices (0.5 to 2.5) that include either
side of Uniform distribution ($\alpha_{ij} = 1$ in $\pi(\theta)$) and Table 2 presents the log Bayes factor in favor of $H_0$.

It can be noted that Bayes factors have shown a consistent pattern except in one case (VI data set) within these two groups of choice of Dirichlet parameters. These data sets are of reasonable size with no zeros and the total is of moderately high value (1660) with reasonably non-polarized counts. This may be a significant observation that choice of Dirichlet parameter (the only prior parameters) for a well-behaved data set is more critical. However, sparseness or polarized count compel to select values either less than 1 or greater than 1 for all the Dirichlet distribution parameters so that Bayes factors are not much sensitive for making conclusions based on the usual recommendations.

**Table 2**

Further, a simulation study has been carried out to supplement the findings. Since the study is focused to consider identified characteristics of $I \times J$ tables, bootstrap simulation are used for the computation of Bayes factor and testing the sensitivity of prior parameters. Based on each of the data sets, 1000 bootstrap samples are generated so that the noted features are expected to be consistently present in the samples. Estimates comprising mean together with its standard error (SE) and 95% limits for confidence interval as 2.5 and 97.5 percentiles are presented in Table 3. Also, Figures 1 and 2 depict the box plots to show the distribution of estimated Bayes factor from bootstrap samples based on data sets I–IV and V–VIII, respectively.
From the box plots it could be observed that values log Bayes factor are exactly on one side of zero with an exemption in data sets II and VI. Though they could be considered as less-polarized data set, low and / or zero counts would cause the changes in the estimates. This is further evident in the numerical summaries which indicate the changes in terms of positive and negative estimated values.

Interestingly, such changes are evident when prior values are near to 1 and this tends to provide a method on choice of prior parameters based on the study which deals with rare or non-rare phenomena. Further, among other data sets with moderately polarized and notable low counts such as data set VI, such directional changes in the estimated values are visible from Table 3 but may not be fully captured by box plots. All other tables with sufficiently large cell counts do not yield any such feature irrespective of polarized cell counts.

However, if a point estimate like mean value is compared for a data set over different choices of priors then no directional changes can be observed. Such a pattern prevails among all the data sets of distinct characteristics. This is consistent with the values of corresponding standard errors too and hence reporting point estimate may not be fully sufficient to study the sensitivity of
estimates over various choices of parametric values. The entire comparison is to demonstrate the sensitivity of priors in the values of Bayes factors attributed to data characteristics, especially when they are sparse in nature. Hence, sparse data occurrence could be reckoned by the researcher as a priori based on the problem under consideration and the way respondents might behave to the choices of variables. This partially elicited information will help to set the range for the prior parameters to estimate as well as consider the sensitivity analysis.

4. Conclusion
Two contrasting recommendations regarding Bayes factors for I × J tables and a 2 × 2 table provide ample scope to investigate the data model, sparseness and pattern of cell counts as it could affect the estimates and thereby the conclusions. Sensitivity of priors on Bayes factor has been accepted in principle; yet understanding the nature of data representing the physical phenomena is important. In such cases, even a controlled vague prior would reflect partial information which tends to reduce the extreme sensitivity as exhibited in Nandram and Choi (2007). Three examples considered in Nandram and Choi (2007) have ignored the distinct features of I × J tables as listed in the present study and could be well influenced by the specific choice of Dirichlet parameters.

Subsequently, Nandram et al (2013) have provided a test of independence related to data from a two-stage cluster sampling design; simulation study has revealed the fact that Bayes factor will not be sensitive to small changes in the uniform prior. This finding has been attributed to the nature of cell counts that are expected to be larger than zero but one or two cells can have zero counts.
The main objective of the paper is to carry out the inevitable sensitivity analysis for Bayes factor but with reasonable parameter choice related to the problem of interest as observed in Vanpaemel (2010). The present work considered the problem on sparse multinominal tables not only limited to zero counts but also low and / or polarized counts as polarized counts tend to affect estimation of multinominal probabilities in the sense of aberrations (May and Johnson, 2000).

Also, to substantiate the findings, Bootstrap samples have been used to study the sensitivity mainly to preserve the expected features of data / underlying phenomena. Computational tool has been provided for a variety of Dirichlet parameters beyond the usual uniform distribution. The procedure envisages a more flexible approach to handle priors for null hypothesis and replication of these procedures can be done through R codes presented in the Appendix.

Kruschke (2010) has pointed out the need to attempt mildly informed or consensually informed prior distributions rather than objective priors. The present study has made an attempt to divide the parameter space appropriately for choosing prior distributions in obtaining Bayes factor to test the independence related to data from a I x J contingency table. Such recommendations are suitable for a more realistic sensitivity analysis for Bayes factor computed for various choices of prior values. Hence the researchers are encouraged to adopt Bayes Factor with plausible priors that may be partially informative as a result of theoretical background of the problem. However, a more concrete way to define the distance between the cell counts could be attempted to study the effect on Bayes factors and redefine the recommendations in the analysis of contingency tables.
Acknowledgment

The authors would like to thank the referees and editor for the valuable comments and suggestions which helped to improve the article.

References


Table 1: Details of the eight data sets considered for the comparative study of Bayes factor associated with two dimensional contingency tables

<table>
<thead>
<tr>
<th>Data</th>
<th>Size</th>
<th>Study Variables</th>
<th>Page No</th>
<th>n</th>
<th>v</th>
<th>% of zero counts</th>
<th>% of low counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>3 x 3</td>
<td>Education and religious belief</td>
<td>80</td>
<td>2726</td>
<td>60.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>II</td>
<td>4 x 4</td>
<td>Income and job Satisfaction</td>
<td>57</td>
<td>96</td>
<td>0.86</td>
<td>6.25</td>
<td>43.75</td>
</tr>
<tr>
<td>III</td>
<td>4 x 4</td>
<td>Opinions about premarital sex and teenage birth control</td>
<td>368</td>
<td>926</td>
<td>9.19</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>IV</td>
<td>5 x 5</td>
<td>First and second Purchase</td>
<td>446</td>
<td>521</td>
<td>6.200</td>
<td>4.00</td>
<td>20.00</td>
</tr>
<tr>
<td>V</td>
<td>3 x 3</td>
<td>Psychiatric thought and ascribed origin of Schizophrenia</td>
<td>83</td>
<td>282</td>
<td>9.89</td>
<td>0.00</td>
<td>11.11</td>
</tr>
<tr>
<td>VI</td>
<td>6 x 4</td>
<td>Mental health status and socio economic status</td>
<td>381</td>
<td>1660</td>
<td>5.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>VII</td>
<td>5 x 5</td>
<td>Father’s and son’s occupational status</td>
<td>447</td>
<td>3600</td>
<td>28.40</td>
<td>0.00</td>
<td>4.00</td>
</tr>
<tr>
<td>VIII</td>
<td>6 x 2</td>
<td>Daily average number of cigarettes and disease group</td>
<td>64</td>
<td>2714</td>
<td>46.92</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

v = range/k; n = sum of cell counts

---

Table 2: Bayes factors (in natural log scale) for the evidence of null hypothesis (no association) in I × J data sets under multinomial sampling model.

<table>
<thead>
<tr>
<th>Data No</th>
<th>0.5</th>
<th>0.75</th>
<th>0.9</th>
<th>1</th>
<th>1.5</th>
<th>1.75</th>
<th>2</th>
<th>2.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>II</td>
<td>5.738</td>
<td>5.159</td>
<td>5.045</td>
<td>5.021</td>
<td>5.186</td>
<td>5.346</td>
<td>5.519</td>
<td>5.865</td>
</tr>
<tr>
<td>III</td>
<td>-45.260</td>
<td>-46.703</td>
<td>-47.199</td>
<td>-47.437</td>
<td>-47.963</td>
<td>-47.974</td>
<td>-47.885</td>
<td>-47.506</td>
</tr>
<tr>
<td>VI</td>
<td>9.373</td>
<td>6.050</td>
<td>4.680</td>
<td>3.925</td>
<td>1.287</td>
<td>0.458</td>
<td>0.297</td>
<td>1.335</td>
</tr>
</tbody>
</table>
Table 3: Bootstrap estimates for the log Bayes factor in favor of null hypothesis of independence using eight distinct prior choices for Dirichlet parameters.

<table>
<thead>
<tr>
<th>Data</th>
<th>Bootstrap Estimates</th>
<th>Different choices for Dirichlet prior used in Bootstrap estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Mean</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>SE(Mean)</td>
<td>-194.041</td>
</tr>
<tr>
<td></td>
<td>SE(Mean)</td>
<td>0.425</td>
</tr>
<tr>
<td></td>
<td>97.5 percentile</td>
<td>3.778</td>
</tr>
<tr>
<td>III</td>
<td>Mean</td>
<td>-96.277</td>
</tr>
<tr>
<td></td>
<td>SE(Mean)</td>
<td>3.476</td>
</tr>
<tr>
<td>IV</td>
<td>Mean</td>
<td>-72.783</td>
</tr>
<tr>
<td></td>
<td>SE(Mean)</td>
<td>4.365</td>
</tr>
<tr>
<td></td>
<td>97.5 percentile</td>
<td>0.246</td>
</tr>
<tr>
<td></td>
<td>SE(Mean)</td>
<td>2.472</td>
</tr>
<tr>
<td></td>
<td>2.5 percentile</td>
<td>-124.088</td>
</tr>
<tr>
<td></td>
<td>97.5 percentile</td>
<td>0.425</td>
</tr>
<tr>
<td></td>
<td>SE(Mean)</td>
<td>1.884</td>
</tr>
<tr>
<td></td>
<td>97.5 percentile</td>
<td>9.634</td>
</tr>
<tr>
<td>VII</td>
<td>Mean</td>
<td>-1034.296</td>
</tr>
<tr>
<td></td>
<td>SE(Mean)</td>
<td>27.509</td>
</tr>
<tr>
<td></td>
<td>97.5 percentile</td>
<td>-430.798</td>
</tr>
<tr>
<td></td>
<td>SE(Mean)</td>
<td>16.361</td>
</tr>
<tr>
<td></td>
<td>2.5 percentile</td>
<td>-977.974</td>
</tr>
<tr>
<td></td>
<td>97.5 percentile</td>
<td>-74.202</td>
</tr>
</tbody>
</table>
Figure 1: Box plots for estimated Bayes factors from the bootstrap samples based on data sets I – IV. Each plot corresponds to eight different choices of prior parameters.
Figure 2: Box plots for estimated Bayes factors from the bootstrap samples based on data sets V – VIII. Each plot corresponds to eight different choices of prior parameters.
Appendix

R code for obtaining Bayes factor for the evidence of null or alternative hypothesis related to test of independence for contingency tables is presented in this Appendix. Input together with the respective symbols used in the code include number of rows (r) and columns (c) of the given contingency table, cell counts as a row vector (ns), and the prior parameters for no association (alp) and association (gam and del) models. Bayes factors in log scale favoring both statements together with the interpretation criterion will be provided; exponentiated version is also given as output.

```
#####Input the size of the contingency table
r=4  # no of rows
c=4  # no of columns

#####Input the original cell counts as a row vector
ns=c(1,3,10,6,2,3,10,7,1,6,14,12,0,1,9,11)  # Cell counts row wise
data=matrix(ns,r,c,byrow=TRUE)  # Calculating row totals
rt=0
for (i in 1:r)
{
    rt[i]=sum(data[i,])
}

ct=0  # Calculating column totals
for (j in 1:c)
{
    ct[j]=sum(data[,j])
}

n=sum(ns)  # Grand Total
alp = 0
k=r*c
for (k1 in 1:k)  # Prior for Dirichlet parameters-Independent Model
{
    alp[k1]=0.5
}

a=sum(alp)

#####Prior for Dirichlet parameters-Association Model-Row
gam = 0
for (k2 in 1:r)
```
{ 
  gam[k2]=0.5
} 

g=sum(gam) 

####Prior for Dirichlet parameters-Association Model-Col 

del = 0 

for (k3 in 1:c) 
{
  del[k3]=0.5 
}
d=sum(del) 

#########Terms for Log Bayes Factor Formula 

t1=ns+alp 

t2=n+g 

t3=n+d 

t4=n+a 

t5=rt+gam 

t6=ct+del 

#########Numerator and Denominator of Log Bayes Factor Formula 
	nr=sum(lgamma(t1))+lgamma(a)+sum(lgamma(gam))+sum(lgamma(del))+lgamma(t2)+lgamma(t3) 

dr=sum(lgamma(alp))+lgamma(g)+lgamma(d)+sum(lgamma(t4))+sum(lgamma(t5))+sum(lgamma(t6)) 

lbf=nr-dr 

#Log Bayes Factor B10 

lbf 

lbf1=-lbf 

#Log Bayes Factor B01 

lbf1 

exp(lbf) 

#Bayes Factor B10 

exp(lbf1) 

#Bayes Factor B01
Mathematical Analysis of propensity of aberration on the methods for interval estimation of the multinomial proportions

U.Sangeetha\(^1\), M.Subbiah\(^2\), M.R.Srinivasan\(^3\)

\(^1\)(Department of Management Studies, SSN College of Engineering, Chennai)
\(^2\)(Department of Mathematics, L. N. Government College, Poonneri)
\(^3\)(Department of Statistics, University of Madras, Chennai)

Abstract: The multinomial distribution has found applications in many practical situations that deal with a discrete random variable with \(k\) possibilities (\(k > 2\)). Interval estimation for the proportion parameter in a special case \(k = 2\), the binomial distribution has been studied extensively in literature. However for \(k > 2\), studies have focused the performance of estimation procedure mainly on coverage probabilities and yet there are other important aspects such as propensity of aberration in the limits of confidence intervals and computational issues. The present paper makes an attempt to look beyond coverage probabilities by marshalling the existing procedures in classical and Bayesian approaches for \(k > 2\). To alleviate the computational issues, a comprehensive R program is also made available to facilitate the implementation of the procedures in both classical and Bayesian statistical paradigms.

Keywords: Aberrations, Bayesian, Contingency Tables, Multinomial, Zero Width Intervals.

I. INTRODUCTION

Probability estimation has drawn an active research attention as it has found wider applications in many areas. Statistical methods for estimating single binomial proportion or associated arithmetic forms of two proportions have discussed extensively in statistical literature (Subbiah, 2009). This includes point and interval estimation procedures and related properties such as coverage probabilities, expected length and existence of unrealistic bounds (Newcombe, 1998 and Sweeting, 2004).

The problem of estimating simultaneous confidence intervals for multinomial proportion has witnessed an active discussion in terms of theoretical, computational and application aspects. Different procedures have been discussed extensively in the literature. Quesenberry and Hurst (1964), Goodman (1965), Fitzpatrick and Scott (1987), Sison and Glaz (1995), Glaz and Sison (1999) can be considered as an earlier work. Also similar to binomial proportion problems, Wald type intervals and Wilson intervals are also available. Further boot strap methods (Jhun and Jeong, 2000), power divergence methods (Hou et al, 2003) are discussed for developing simultaneous confidence intervals. Wang (2008) has proposed exact confidence coefficient; Chafai and Concordat (2009) has discussed the problem especially for small samples.

Bayesian procedures have also been discussed in obtaining simultaneous confidence coefficients for multinomial proportion. A conjugate prior model, multinomial – Dirichlet has been discussed extensively in many literature and Gelman et al (2002) provides the details of the relations and effect of hyper parameters; Tuyl et al (2009) have shown that the uniform prior could be considered as the consensus prior for multinomial parameters and recently, Komaki (2012) has developed a prior based on a specific Dirichlet prior that would be asymptotically minimax.

Further, the performance of methods is studied based on coverage probability length of intervals and computational flexibility; small and large sample sizes; polarized cell counts are also considered in this comparison. However, aberrations are not elaborately discussed especially on small samples though direct adjustments on such limits are recommended in practice. Computational issues have also been discussed in many studies and in particular, May and Johnson (1997, 2000) have listed the computational tools for most of the available methods, however the availability of these procedures are limited to selected software so that most of the applications might rely on a specific method or a software.

This paper has two fold objectives: discussion of possible aberration more generally for widely applicable procedures in classical and Bayesian approaches. Similar discussions for other estimation problems (for example, Newcombe, 1998) only illustrate the propensity of overshoot and a more general way to understand such existence has been attempted to classify the methods.

Secondly, limited availability of procedures like Sison and Glaz in selected software that may deter its usage (http://www.rforge.net/doc/packages/NCStats/gofCI.html) and an attempt has been made to develop comprehensive computational tools for most of the methods in both classical and Bayesian methods. R programming environment has been used to provide a tool for obtaining multinomial interval estimate based on...
Mathematical Analysis of propensity of aberration on the methods for interval estimation of the

Wald, Wilson, Queenberry and Hurst (QH), Goodman (GM), Fitzpatrick and Scott (FS), Sison and Glaz (SG) methods in classical inference and multinomial-Dirichlet distribution method in Bayesian procedure. Numerical examples are presented to compare these methods especially on the small counts. Findings of this work have encouraged to further investigate sparse data sets in large two dimensional categorical tables that are being analyzed under multinomial sampling.

II. METHODS

To depict the multinomial sampling which is generally represented as a contingency table following list of notations could be useful

$I$ = No. of rows; $J$ = No. of columns; $k$ = IxJ
$n_{ij}$ = Cell counts
$n = \sum n_{ij}$ (Size of the sample)
$\pi_i$ = Population proportion of $i^{th}$ cell where $i = 1, 2, \ldots, k$
$p_i$ = Sample proportion
MLE estimation of $\pi_i = \frac{n_i}{N}$
$(p_i)_B$ = Bayesian estimation of $\pi_i$, and a dot represents the prior
$A = \chi^2(\alpha, 1)$ i.e., the upper 100$(1-\alpha)$% point of the chi square distribution

Further Sison and Glaz (1995) have proposed a method to determine simultaneous confidence intervals for multinomial proportions; one method is based on independent Poisson random variable and associated central and factorial moments. In this method, cell counts are assumed to be independent Poisson random variable with parameter $\lambda_i = \pi_i n$. Since $\pi_i$ is unknown, $\lambda_i$ is estimated as $n p_i = n_i$

Also, $n_i$~ Poisson $[\lambda_i]$, and hence $p(b_i \leq Z_i \leq a_i) = \sum_{b_i}^{a_i} p(Z_i = b_i) = F(a_i) - F(b_i - 1)$ where $F$ is CDF of Poisson random variable.

Further, the central factorial moments $\mu_r(\pi)$ of truncated Poisson variable $Y_i$ and the central moments (up to r=4) of $Y_i$ are calculated to obtain $p(w=n)$. Then the limits of $100(1-\alpha)$% confidence limits are obtained as $LL = p_i \frac{C}{n}$ and $UL = p_i + \frac{C + 2\delta}{n}$ where $C$ is a positive integer that satisfies the condition $p(C) \leq 1-\alpha \leq p(C+1)$ with $p(C) = p[p_i C/n \leq \pi_i \leq p_i + C/n] = 1-\alpha$ and $\delta = \left( \frac{1-\alpha}{\rho(C+1)} \right) \frac{\rho(C)}{\rho(C+1)}$.

The Bayesian approach for estimating cell probabilities for a way contingency table under multinomial sampling could be outlined with a two-step likelihood-prior combination. If the random vector $n=(n_1, n_2, \ldots, n_k)$ has a multinomial distribution with n trials and cell probabilities $\pi_1, \pi_2, \ldots, \pi_k$ then the joint pmf of $(n_1, n_2, \ldots, n_k)$ is

$$f(n_1, n_2, \ldots, n_k) = \frac{n!}{n_1! n_2! \ldots n_k!} \pi_1^{n_1} \pi_2^{n_2} \ldots \pi_k^{n_k}$$
The conjugate prior (Gelman et al., 2002) for the proportion parameters \( \theta = (\pi_1, \pi_2, \ldots, \pi_k) \) could be a multivariate generalization of Beta distribution known as Dirichlet \((\alpha_1, \alpha_2, \ldots, \alpha_k)\) with \( \alpha_i > 0 \) and PDF is
\[
p(\theta) = \frac{\Gamma(\sum \alpha_i)}{\prod \Gamma(\alpha_i)} \pi_1^{\alpha_1-1} \pi_2^{\alpha_2-1} \ldots \pi_k^{\alpha_k-1}, \quad \sum \pi_i = 1.
\]
Hence, the posterior distribution will be
\[
\Pi(\theta | X) \propto \pi_1^{a_1+x_1-1} \ldots \ldots \pi_k^{a_k+x_k-1} = \text{Dirichlet} (\alpha_1+x_1, \alpha_2+x_2, \ldots, \alpha_k+x_k) \text{ up to normalizing constant.}
\]
For a non-informative prior, widely chosen values for \( (\alpha_1, \alpha_2, \ldots, \alpha_k) \) includes \( \alpha_i = 1 \) corresponds to the uniform prior distribution, Jeffreys’s prior \( \alpha_i = 1 \) and \( \alpha_i = 1.41 \) corresponds to asymptotic prior (Komaki, 2012) distribution.

### III. CONDITIONS FOR ABBERRATIONS

In most of the studies, overshoot or zero width intervals have been illustrated based on the low (zero or positive) cell counts. However, a more general scenario can be derived by considering the mathematical form of each method to indicate the possible chances of aberrations. This section includes the possibilities of cells with zero counts and the range of cell counts that may result with over shooting estimated values for the associated lower and upper limits (LL & UL) of confidence intervals.

In Queenberry and Hurst, if \( n_i = 0 \) for some \( i \) then \( \text{LL} = \frac{(\bar{A}_i - A_{i+1})}{\sqrt{A_{i+1} + 2 A_i n_i}} = 0 \) and \( \text{UL} = \frac{A_1}{A_{n+k}} \) that yield fixed limits for the cells with zero count.

Also LL < 0 is true
\[
\text{only if } A_1 + 2 n p_i - \sqrt{A_1 (A_1 + 4 n p_i (1 - p_i))} < 0
\]
\[
\text{only if } (A_1 + 2 n p_i)^2 < A_1^2 + 4 n A_i p_i (1 - p_i)
\]
\[
\text{only if } A_1^2 + 4 n^2 p_i^2 + 4 n p_i A_i < A_1^2 + 4 n A_i p_i - 4 n A_i p_i^2
\]
\[
\text{only if } 4 n^2 p_i^2 < -4 n A_i p_i^2
\]
\[
\text{only if } n < -A_1 \text{ which is not possible.}
\]

Hence QH never have negative lower limits.

Similarly UL > 1 is true only if \( (2 n p_i + A_i)^2 + \frac{A_1 (A_1 + 4 n A_i p_i (1 - p_i))}{2 (A_1 + n)} \) and subsequent algebra will yield a condition \( 0 > n (1 - p_i) + A_i (1 - p_i) \) which is not possible (since \( n, (1 - p_i), A_i \) are positive) and hence QH will never provide UL which are more than 1. Since, Goodman and Wilson differ from Queenberry and Hurst only in the \( \chi^2 \) value (\( A_1, A_i, A_k \)) similar conclusions can be made regarding these two procedures.

In Fitzpatrick and Scott method, when \( n_i = 0 \) corresponding LL and UL will be \( \frac{\bar{A}_i}{\sqrt{2 n}} \) that obviously yield a negative LL.

Also LL < 0 only if
\[
p_i < \frac{A_1 + 1 - n_i}{\sqrt{2 n}}
\]
\[
0 < n_i < \sqrt{\frac{A_1 - 1}{2 n}}
\]

This could be possible and such a case could also be attributed towards polarized cell counts with varying \( k \). For example, if \( n = 36, 1 \leq n_i \leq 5 \) will have negative lower limits for a 95% confidence interval. Further, in this case if \( k \) is small say 4 and \( n_i < 5 \) for at least one \( i \) (say \( i = 1 \)) remaining three cells must be added up to 32. A hypothetical example (1,1,1,12,9) would bring a negative Lower Limit for the cell \( i = 1 \).

Also, in this case, UL > 1 only if \( p_i > 1 - \frac{Z_{n_i/2}}{\sqrt{2 n}} > 1, \quad p_i > 1 - \frac{Z_{n_i/2}}{\sqrt{2 n}} \) only if \( n_i > n - \frac{Z_{n_i/2}}{\sqrt{2 n}} \). This case is also possible in practice and that might indicate the extremely distant values in the data set. Though FS has a symmetry property, aberration is an issue which would exist in possible cases. Hypothetical data set is (73, 1, 1, 6) will bring Upper Limit > 1 in first case and negative Lower Limit in all other 3 cases.

In Wald Type, when \( n_i = 0 \), both Lower Limit and Upper Limit are 0 which is an existence of zero width interval. Also, LL < 0 only if
\[
p_i < \frac{A_1 (1 - n_i)}{n_i} < 0
\]
\[
p_i < \frac{A_1 (1 - n_i)}{n_i}
\]
\[
p_i < \frac{A_1 (1 - n_i)}{n_i}
\]
\[
n_i < A_1 - A_n
\]
Mathematical Analysis of propensity of aberration on the methods for interval estimation of the

\[ n_i < \frac{n}{2(n + A)} A = 1 - \frac{A}{n + A} \]

(1)

Also, \( \text{UL} > 1 \) Only if

\[ p_i + \frac{\frac{\lambda_{n-1, p_i}}{n} \cdot \frac{1}{2}}{n} > 1 \]

\[ \frac{\lambda_{n-1, p_i}}{n} > (1 - p_i)^2 \]

\[ \frac{\lambda_{n-1, p_i}}{n} > 1 - p_i \]

\[ \frac{\lambda_{n-1, p_i}}{n} > 1 - p_i \]

\[ \text{An} > n \left( \frac{n}{n - 1} \right) \text{and } \text{LL} = n(n - 1) \]

However if \( n \) is large then \( \text{AF} \rightarrow 1 \) thereby \( n_i \rightarrow n \) in \( i^{th} \) cell. This means \( n_i \rightarrow 0 \) as \( j \neq i \) and could be an illustration of polarized cell counts.

In Wald CC method, if \( n_i = 0 \), then limits for the corresponding intervals are \( \mp \frac{1}{2n} \) that yield an obvious negative limits. Also, \( \text{LL} < 0 \) only if

\[ p_i \mp \frac{\frac{\lambda_{n-1, p_i}}{n} \cdot \frac{1}{2}}{n} < 0 \]

\[ \frac{n^2 + 1}{4n^2} \cdot \frac{n}{n + A} < \frac{p_i}{n} \cdot \frac{A}{n} \]

\[ \frac{n^2 + 1}{4n^2} \cdot \frac{n}{n + A} < \frac{(n_1n_2)}{n} \]

\[ n_i^2 \left[ n + A \right] - n_i \left[ n + nA \right] + \frac{2}{4} < 0 \]

\[ n_i^2 < n_i < n_i^* \text{ where } n_i^* = \frac{n(n + A) + \frac{n^2 + 2n + 2}{2}}{n(n + A)} \]

For, Sison and Glaz method \( p_i = \frac{n}{n} \) \( \text{LL} = \frac{n}{n} \cdot \frac{C}{n} \) and hence whenever \( C > n_i \) (for some \( i \)), LL could be negative in \( i^{th} \) confidence interval. Similarly, \( p_i \mp \frac{\frac{\lambda_{n-1, p_i}}{n} \cdot \frac{1}{2}}{n} > 1 \) only if \( n_i + C + 2\delta > n \), only if \( n_i > n - (C + 2\delta) \) where \( \delta \) depends on the distance between \( p(C), 1 - \alpha \) and \( p(C + 1) \), such that \( \delta \leq 1 \) so that \(-\delta \geq -1 \Rightarrow n_i > n - (C + 2\delta) > (n-1) - C \) or equivalently \( C > (n-1) - n_i \) For any \( i \) then \( \text{UL} \) of \( i^{th} \) CI will be more than 1.

IV. Data Analysis

To illustrate the notion of estimation and the cases in which negative lower limits occur, four data sets (referred to as I - IV) have been considered and presented in TABLE 1. The first data set have been used in Chafai and Concordet (2009) which deal with the difference in behavior of male and female students with respect to smoking habits; second one from Szyda et al (2008) that has been used to analyze DNA pooling experiments with arctic fox; a data set from Sison and Glaz (1995), Hou et al (2003) is the third set; The fourth one is a partial collection of data sets from the data repository (www.fars-nhtsa.gov/states/statesAlcohol.aspx) that deals with fatal crashes and related information from National Highway Traffic Safety Administration’s (NHTSA) Fatality Analysis Reporting System (FARS) and its supplemental files. FARS has a collection of most or all fatal motor vehicle crashes that occur in United States of America. Further, alcohol related files are made up of the counts of crashes and fatalities due to three different levels of blood level consumptions (BAC equals to 0.00, between 0.01-0.07 and 0.08 or more). However, only nine tables for the year 2009 with wider range of cell counts have been included for the analysis.

The data sets are collected in such a way to describe the presence of zeros; sparse non zero counts and has wider range. Also, minimum count in I and III data sets is 3 and 56 respectively and that of IV varies from 2 to 14 over the nine tables. It has been observed that some procedures yield obvious negative lower limits whereas Wald method provides zero width intervals (ZWI) for zero counts. However, the main interest of the study lies on the aberrations that could occur for non-zero counts. In the case of data set I, uniformly three methods W, WCC, FS and SG yield negative lower limits when the cell count is 3 and such a result does not exist in the other methods. In fact QH, GM, and WS can be eliminated from such comparison as they never exhibit the characteristic of aberration and similar conclusions can be made for the Bayesian procedures too.
Mathematical Analysis of propensity of aberration on the methods for interval estimation of the

For data set II, Wald type intervals do not yield negative estimates for positive counts, whereas SG and FS result with negative lower limits when the cell counts are 5, 14 and 10; further, when the cell count is 16, SG has zero as lower limit but FS has a negative lower limit. In fact, the choice of c in the SG procedure will decide such values when any cell count is equal to C, lower limit is zero for the lesser cell counts it is negative. This kind of behavior warrants a deeper insight into deciding the sparse nature among positive cell counts and when cell counts are highly polarized.

Similar conclusions are visible in the case of all nine tables of data set IV but a careful observation throws the light for SG when the range of cell counts are very high as much as 176 does not result with a negative lower limit for the cell count 14 in contrast to data set II. Hence, a quick decision could not be arrived regarding the performance of SG when high polarized count exist (May and Johnson, 2000) in the given data set; but the value of C is a notably influencing number for such cases in the estimated values. On the other hand, with larger values (more than 56) present in data set III, none of the cells have aberrations though range is notably high.

Further, TABLE 2 and TABLE 3 summarize the interval estimates (LL: lower limit and UL: upper limit) for the multinomials proportions related to the data set III to facilitate a comparison of the seven classical with three Bayesian methods. Bayesian methods yield least volume when compared to all other classical methods and the estimates are quite closer to the methods due to Wilson and Wald CC. Comparisons among classical procedures have been done already in many studies such as Hou et al (2003), this work does not include similar effort among non-Bayesian methods. But an emphasis is made to note the advantage of using Bayesian direct posterior estimation from Dirichlet distribution with standard objective priors that will yield better results and involve lesser computation and never yield implausible estimates under different circumstances of cell counts and size of the I x J tables.

V. CONCLUSION

Methods for estimating simultaneous multinomial proportions have found research attention in recent times but not specific to the unrealistic confidence intervals. For example Lee et al (2011) have pointed out the implausible limits for confidence interval as unrealistic. Though a simple practical remedy is visible such as truncating the limits suitably, a systematic method to understand such impractical estimates has not been developed at least in the case of multinominal proportions. This paper makes an attempt to derive suitable mathematical expression for the propensity of aberrations in the method for estimating multinominal proportions.

In the present work, seven classical procedures and Bayesian method with three priors are compared for the interval estimation of multinominal proportions. For the large samples and for an introductory level Wald type intervals can be suggested which are computationally simpler yet do not posses essential other properties (such as coverage probability) so as not to recommend for practical applications especially for low cell counts. Though FS can be equally compared to Wald type intervals for its mathematical simplicity, this method is also not a favorable choice when the cell counts are small. Wilson intervals and similar forms QH and GM are providing plausible set of limits for all kind of data; but conservatism is an issue when k is comparatively large.

However, SG performs better in terms of achieving required coverage probability in all cases by the virtue of its mathematical form; yet it suffers a limitation in terms of propensity of aberration whenever the cell counts are polarized. Hence a careful judgment is needed to handle such limits which could occur even when a cell count is as large as 14. The availability of SG has also made wide spread in this work, and hence among the method considered in classical statistics, SG can be recommended with a caution about the range of cell counts. This paper provides a beforehand indication for the propensity of aberrations in confidence limits.

On the other hand, Bayesian Multinomial-Dirichlet methodology has been revisited with two well known objective priors (uniform and Jeffreys) together with a recently discussed asymptotic prior. It has been observed that Bayesian methods are quite flexible in modeling, mathematical treatment, computation and its uniform behavior over a variety of data, and never warrant corrections due to over shoot of the estimates. Further, advances in simulation techniques encourage adopting Bayesian objective methods to obtain interval estimates for multinominal proportions. Together with the existing literature on multinominal proportions, this study provides a consolidated view to select the procedures based on some criteria such as size of the contingency table (k), nature of cell counts and its range to indicate polarization, computational flexibility, propensity of aberrations, and desirable frequentist properties. However, the choice between Bayesian and classical methods need to be handled similar to other situations of data analyses.

REFERENCES


www.iosrjournals.org 27 | Page
Table 1: Illustrative data sets and their basic characteristics

<table>
<thead>
<tr>
<th>Data sets</th>
<th>Size of the original table</th>
<th>No. of Cells</th>
<th>Zero entries</th>
<th>Range of table totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>2 x 2</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>II</td>
<td>4 x 5</td>
<td>20</td>
<td>12</td>
<td>60</td>
</tr>
<tr>
<td>III</td>
<td>1 x 7</td>
<td>7</td>
<td>0</td>
<td>31</td>
</tr>
<tr>
<td>IV</td>
<td>1 x 3, 9 tables</td>
<td>3</td>
<td>0</td>
<td>46, 64, 75, 176, 42, 74</td>
</tr>
</tbody>
</table>

Table 2: Comparison of six 95% confidence interval resulted from classical procedures for the data set III

<table>
<thead>
<tr>
<th>Cell</th>
<th>Wald</th>
<th>Wald CC</th>
<th>Wilson</th>
<th>Quenby-Hurst</th>
<th>Goodman</th>
<th>Fitzpatrick-Scott</th>
<th>Sison-Glaz</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LL</td>
<td>UL</td>
<td>LL</td>
<td>UL</td>
<td>LL</td>
<td>LL</td>
<td>LL</td>
</tr>
<tr>
<td>1</td>
<td>0.090</td>
<td>0.149</td>
<td>0.089</td>
<td>0.130</td>
<td>0.094</td>
<td>0.153</td>
<td>0.076</td>
</tr>
<tr>
<td>2</td>
<td>0.121</td>
<td>0.187</td>
<td>0.120</td>
<td>0.188</td>
<td>0.124</td>
<td>0.190</td>
<td>0.104</td>
</tr>
<tr>
<td>3</td>
<td>0.123</td>
<td>0.189</td>
<td>0.122</td>
<td>0.190</td>
<td>0.126</td>
<td>0.192</td>
<td>0.106</td>
</tr>
<tr>
<td>4</td>
<td>0.096</td>
<td>0.156</td>
<td>0.095</td>
<td>0.158</td>
<td>0.099</td>
<td>0.160</td>
<td>0.081</td>
</tr>
<tr>
<td>5</td>
<td>0.102</td>
<td>0.164</td>
<td>0.101</td>
<td>0.165</td>
<td>0.105</td>
<td>0.167</td>
<td>0.087</td>
</tr>
<tr>
<td>6</td>
<td>0.151</td>
<td>0.222</td>
<td>0.150</td>
<td>0.223</td>
<td>0.154</td>
<td>0.224</td>
<td>0.131</td>
</tr>
<tr>
<td>7</td>
<td>0.094</td>
<td>0.154</td>
<td>0.093</td>
<td>0.155</td>
<td>0.097</td>
<td>0.157</td>
<td>0.080</td>
</tr>
</tbody>
</table>

Volume: 4.07 x 10^{-7}, 5.14 x 10^{-8}, 4.08 x 10^{-7}, 2.52 x 10^{-7}, 3.59 x 10^{-8}, 5.11 x 10^{-9}
3.24 x 10^{-8}

Table 3: Comparison of three 95% confidence interval resulted from Bayesian procedures for the data set III

<table>
<thead>
<tr>
<th>Cell</th>
<th>Bayesian-1</th>
<th>Bayesian-2</th>
<th>Bayesian-3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LL</td>
<td>UL</td>
<td>LL</td>
</tr>
<tr>
<td>1</td>
<td>0.093</td>
<td>0.151</td>
<td>0.092</td>
</tr>
<tr>
<td>2</td>
<td>0.123</td>
<td>0.188</td>
<td>0.123</td>
</tr>
<tr>
<td>3</td>
<td>0.125</td>
<td>0.191</td>
<td>0.124</td>
</tr>
<tr>
<td>4</td>
<td>0.099</td>
<td>0.159</td>
<td>0.099</td>
</tr>
<tr>
<td>5</td>
<td>0.104</td>
<td>0.164</td>
<td>0.109</td>
</tr>
<tr>
<td>6</td>
<td>0.152</td>
<td>0.223</td>
<td>0.153</td>
</tr>
<tr>
<td>7</td>
<td>0.097</td>
<td>0.156</td>
<td>0.096</td>
</tr>
</tbody>
</table>

Volume: 3.76 x 10^{-7}, 3.96 x 10^{-7}, 3.53 x 10^{-7}
A study on further characteristics of contingency tables with the R Package - RAP.

U. Sangeetha, M. Subbiah, M.R. Srinivasan

Abstract — The ubiquitous Chi square test statistic for association between two or more categorical variables provides ample scope to investigate its characteristic in terms of methodological or application point of view. Many studies have pointed out its appropriateness in case of sparse tables and also there are few attempts to understand the category wise associations through partitioning Chi square distribution using G² statistic. This work attempts to study the exhaustive possibilities of forming sub tables from the given contingency table to study the category wise association through Chi square test statistic; particularly the tables which exhibit reversal association pattern (RAP) when compared to original conclusion. This computer intensive effort necessitates developing an R package called RAP for complete enumeration of sub tables. Further, the simulation study has observed that this behavior of RAP persistently exists among 2 x 2 tables and this software can be used to understand one more characteristic of Chi square statistic and a supporting tool to fix sub tables for partitioning schema for an academic exercise or for typical application studies.

Index Terms — Chi-Square tests, Partitioning, Association, Categorical variables, R packages

1 INTRODUCTION

Categorical data analysis had found applications in many fields such as medicine and social science (Agresti, 1992, Tang, et al, 2012). Such data consist of frequency counts of observations occurring in the response categories. For two categorical variables with I and J levels respectively, a contingency or cross-classification table is generally used; each cell of the table counts the number of cases for the simultaneous occurrence of row and column variables. Most of the statistical analyses related to categorical data presented in a contingency tables deal with testing independence of the categorical variables. In this attempt many studies have focused the issues of sparse contingency tables especially the presence of zero or small counts (Koehler and Larntz, 1980, Brown and Fuchs 1983, Haberman 1988, Gorman et al 1990, Maiste and Weir, 2004, Burman 2004, Campbell 2007, Hashino 2012). Ratio of sample size of the table to the number of cells is invariably considered as a tool to understand sparseness beyond the presence of zeros and small counts.

Recently Rapallo (2012) has provided the methods for outlier patterns in contingency tables, using distance between the cell counts. Also such distance plays significant role in estimation of multinomial probabilities as noted in May and Johnson (2000) with respect to a method due to Sison and Glaz (1995). Apart from this structural metric of a contingency table, statistical studies have focused on the use of Chi-square test as a measure of association (Mirkin, 2001). Under multinomial sampling in two-way contingency tables, Pearson Chi square statistic has found an extensive usage to test the null hypothesis of statistical independence.

$$H_0 : \pi_{ij} = \pi_{i+} \pi_{+j}$$ for all $i=1,2,\ldots, I$ and $j=1,2,\ldots, J$ where $\{\pi_{ij}\}$ is the joint probability distribution of both categorical variables and marginal distributions are the row and column totals denoted by $\pi_{i+} = \sum_j \pi_{ij}$ and $\pi_{+j} = \sum_i \pi_{ij}$.

The graph below shows the number of articles containing the phrase “Chi square Statistic for independence” between 2000 and 2012. The search is limited to Science Direct and Wiley Publications. It is also observed that numbers are doubled when the search phrase includes Chi square, Chi square test for association.

| Li. Sangeetha – Asst. Professor, Department of Management Studies, SSN College of Engineering, Kalavakkam, Chennai. Email: usanger19@gmail.com |
| M. Subbiah – Asst. Professor, Department of Mathematics, L.N.Govt. College, Ponneri. Email: sisufive@gmail.com |
| M. R. Srinivasan – Professor & Head, Department of Statistics, University of Madras, Chennai. Email:mrsrin8@gmail.com |
In spite of its theoretical popularity (Mirkin 2001) and computational ease, Chi square test faces a warning about its usage for small samples or sparse large contingency table (Berkson 1938, Cochran 1954, Campbell, 2007, Agresti, 1992). Also such situations are indicated in many statistical software with inbuilt warning messages.

Hence the objective of this work is twofold; to work out a metric for classifying a contingency table based on the polarized cell counts and to develop an R package to understand the category-to-category association supplementing statistical inference for contingency tables. This work could help the practitioners to classify the sparseness of the given contingency table and compare with the association results of its all possible tables. Since the exhaustive enumeration involves 

\[
\binom{I}{2} \binom{J}{2} \binom{J}{3} \cdots \binom{J}{J}
\]

attempts (I: no. of rows; J: no. of columns), a convenient procedure in the R package shortlists the sub table which reverse the association compared to the original table.

This article has brought out one such feature of Chi square test for independence based on the way category to category association behaves when compared to over all association. Along to the pair wise comparisons in ANOVA models, this work attempts to observe the relationship between possible association that could be exhibited between the levels of categorical bivariates. A systematic R package has been developed to implement the study that involves an exhaustive enumeration and calculations.

2 MOTIVATION

Agresti (1992) and few many studies have indicated the partitioning of Chi squared statistic. This is mainly to understand the component wise association aspects. A partition could help to show an association to indicate the differences between certain categories. Two studies can be considered for illustrating the notion of partitioning Chi square statistic; Example 1 deals with most Influential School of Psychiatric Thought and Ascribed Origin of Schizophrenia (Agresti, 1992) and following table presents the actual data.

<table>
<thead>
<tr>
<th>Biogenic</th>
<th>Environmental</th>
<th>Combination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eclectic</td>
<td>90</td>
<td>12</td>
</tr>
<tr>
<td>Medical</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>Psychoanalytic</td>
<td>19</td>
<td>13</td>
</tr>
</tbody>
</table>

Example 1

Example 2 investigates whether there is evidence to indicate a difference in the distribution of preference across the four state universities; this can be accessed from www.biostat.umn.edu/~dipankar/bmtry711.11/lecture_10 .pdf and the details are provided in the table

<table>
<thead>
<tr>
<th>State</th>
<th>Bargaining agent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>University 101</td>
</tr>
<tr>
<td>1</td>
<td>42</td>
</tr>
<tr>
<td>2</td>
<td>31</td>
</tr>
<tr>
<td>3</td>
<td>26</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
</tr>
</tbody>
</table>

Example 2

However, the partitioning procedure need not be unique combinations of sub tables yet it requires a careful way of construction of sub tables. Hence an attempt has been made to obtain all possible sub tables exhaustively and a scope to pick sub tables for a suitable partitioning schemes.

3 REVERSAL ASSOCIATION PATTERN (RAP)

An exhaustive enumeration of all possible sub tables will bring more information about category to category associations together with overall association. This approach needs a large number of sub tables that are reckoned using following details. Let I, J be the number of rows and number of columns. The problem is to find the number of sub tables with \( 2 \leq i \leq I, 2 \leq j \leq J \) with the assumption that original table is also considered as a sub table of itself.

The number of ways 2 rows can be selected is \( \binom{1}{2} \). For each of these choices there are \( \binom{2}{2} \binom{J}{J} \cdots \binom{J}{2} \binom{J}{J} \) ways of selecting 2 columns, 3 columns etc. Hence number of sub tables with exactly 2 rows
Similarly the number of sub tables with exactly 3 rows

\[ \binom{J}{3} \binom{I}{J-1} \]

Hence the number of sub tables (including original)

\[ \binom{1}{3} \binom{J}{J-1} = \binom{1}{2} \binom{J}{2} - \binom{1}{2} \binom{J}{J-1} \]

Therefore, number of sub tables required for finding reversal pattern

\[ (2^I - 1)(2^J - 1) - 1 \]

Naturally, the number of sub tables will increase in the order of \( I + J \) however, advances in computations make such task achievable and hence a tool in the open source environment R (R core development team), has been developed. This R package RAP will help the user to identify the sub tables which reflect an association that reverse the overall association between the given two variables. Reversal aspects are based on the usual level of significance (5%) followed in tests for independence.

4 Results

Initially the procedure of RAP has been obtained for the two illustrative data sets and the results are displayed in Tables 1 and 2.

<table>
<thead>
<tr>
<th>S.No</th>
<th>No. of rows</th>
<th>No. of cols</th>
<th>Selected rows</th>
<th>Selected cols</th>
<th>Pvalue</th>
<th>P value significant at 5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1,2,3</td>
<td>1,2</td>
<td>0.0002</td>
<td>TRUE</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1,2</td>
<td>1,2</td>
<td>0.9503</td>
<td>FALSE</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1,2</td>
<td>1,3</td>
<td>0.5032</td>
<td>FALSE</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2</td>
<td>1,2</td>
<td>2,3</td>
<td>0.5032</td>
<td>FALSE</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>2</td>
<td>1,3</td>
<td>1,3</td>
<td>0.5032</td>
<td>FALSE</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>2</td>
<td>2,3</td>
<td>1,2</td>
<td>0.5032</td>
<td>FALSE</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>2</td>
<td>2,3</td>
<td>2,3</td>
<td>0.5032</td>
<td>FALSE</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>2</td>
<td>1,2</td>
<td>2,3</td>
<td>0.5032</td>
<td>FALSE</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>S.No</th>
<th>K</th>
<th>Size</th>
<th>n</th>
<th>Median of the proportion of 2x2 sub tables</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>3x3</td>
<td>160</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>3x3</td>
<td>2726</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
<td>4x4</td>
<td>926</td>
<td>0.6</td>
</tr>
<tr>
<td>4</td>
<td>24</td>
<td>6x4</td>
<td>2714</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>25</td>
<td>5x5</td>
<td>3600</td>
<td>0.625</td>
</tr>
</tbody>
</table>
From Table 4 it could be observed that the sparseness measure based on sum of the cell counts is more sensitive to that is derived from range based measure; this could be mainly due to the fact that when the given data set is sampled through bootstrap method, cell counts would change that directly affect their sum; however the study is focused to replicate the given data sets samples are not based on random sampling techniques such as Monte Carlo samples. Hence the re-sampling counts tend to repeat the values and range is mostly a robust measure in such repeated circumstances. This is to substantiate the notion that when data likelihood is based on independent identical samples its characteristics are closely fixed by the respective samples with replacements except the sum of the counts which is quite sensitive to a slight change. There by the present work through the extensive simulation study has made an attempt to establish that range based measure can be considered as a better tool to portray the distinct pattern of data dispersion in a contingency table.

**CONCLUSION**

Usual contingency table is limited to chi square or probability estimation. Sparse nature which plays a role in statistical inference theory is also a relatively important area. But beyond that nature of cell counts and association at micro level has motivated to propose a measure and develop a computing tool to understand the micro association. However beyond the usual sparseness metric using sample sizes and cell counts, the distance between the cell counts do draw active attention to classify the tables. A new metric has been proposed for understanding the extent to which the cell counts are dispersed with respect to the size of the table; this is further exemplified by a set of motivating examples and bootstrap samples for its sensitivity.

Also, another objective of this paper includes a main feature of Chi square test for independence between bivariate categorical variables. This attempt can be considered as a primitive approach for understanding category to category association that is similar to the exhaustive enumeration involved in LSD approach for pair wise comparisons in ANOVA models. Also, similar reversal effect in summary measure has provided a procedure to classify the sparse 2 x 2 data (Subbiah and Srinivasan, 2008). However, the practical implementation in general I x J table requires a large number of sub tables and their association. Hence the notion is supplemented. With RAP, R Package to understand the category wise association more easily studies in academic classroom examples or typical association could make use of this tool to understand the important aspect of Chi squared test statistic for independence.

**REFERENCE**


Estimation of confidence intervals for Multinomial proportions of sparse contingency tables using Bayesian methods

U. Sangeetha*, M. Subbiah**, M.R. Srinivasan***

* Department of Management Studies, SSN College of Engineering, Chennai.
** Department of Mathematics, L. N. Government College, Ponnneri.
*** Department of Statistics, University of Madras, Chennai.

Abstract - Multinomial distribution, widely used in applications with discrete data, witnessed varieties of competing intervals from frequentist to Bayesian methods, still prove to be interesting in the case of zero counts or sparse contingency tables. The methods commonly recommended in both approaches are considered based on its influence of zero counts, polarizing cell counts, and aberrations. The inference based on comparative study shows that Bayesian approach, with an appropriate prior could be a good choice in dealing with a sparse data set without any imputation for zero values.

Index Terms: Bayesian inference, Coverage probabilities, Dirichlet distributions, Multinomial distributions, sparse data.

I. INTRODUCTION

The cell of contingency table contains frequency of outcomes of categorical response variables and its number denotes the dimension and size is determined by number of categories related to each of the variable. Generally, inferential methods for categorical data assume multinomial or Poisson sampling models. The observed counts \{ni; i=1,2,...,k\} could be considered as k levels of a single categorical variable or for k=IJ cells of a two way categorical variables with levels I and J. Agresti (1992) has explained the different sampling k models and in particular, the present work is based on the multinomial distribution \(\{\pi_1,\pi_2,...,\pi_k\}\). Maximum likelihood estimates (MLE) of cell probabilities can be derived easily as sample cell proportions but interval estimation of multinomial probabilities too has drawn then active attention.

The impact of sparseness provides an ample scope to have a comparative study among these methods as well as Bayesian procedures. Agresti and Yana (1987) have stated that the asymptotic approximations may be quite poor for sparse table, even for a large N. Further Szyda et al (2008) observed that sparseness could occur even when k is relatively large. Subbiah and Srinivasan (2008) have studied the nature of sparseness in a 2x2 table based on a summary measure. Also, recent developments have favored Bayesian approaches as more suitable methods to handle sparseness as compared to three standard recommendations while handling sparse or zero counts (Agresti, 1992, Subbiah et al 2008).

The objective of this paper is to draw comparisons that include Bayesian approach with non-informative priors for underlying parameters. Study envisages use of typical 2x2 data sets in the literature and a large contingency table (Szyda etal, 2008). Frequentist property of coverage probabilities for Bayesian approach have also been studied and compared with the available results of classical approaches using recent computational tools. The following section provides a comprehensive list of active methods in the literature considered for comparison of confidence intervals for multinomials proportions.

II. CONFIDENCE INTERVALS FOR MULTINOMIAL PROPORTIONS

In the case of Bayesian inference, Dirichlet distribution is the widely used and recommended conjugate prior distribution for the multinomial probability parameters (Gelman et al, 2000). However, to obtain posterior distribution, a relationship between Gamma distribution and Dirichlet distribution has been used and presented as

\[
Y_1 \sim \text{Gamma} (a_0,1), \quad V = \sum_{i=1}^{k} Y_1 \sim \text{Gamma} (a_0,1) \quad \text{where} \quad a_0 = \sum_{i=1}^{k} a_i
\]

Then \(\left(\frac{Y_1}{V}, \frac{Y_2}{V}, ..., \frac{Y_k}{V}\right) \sim \text{Dirichlet} (a_1, a_2, ..., a_k)\).

With a proper choice of hyper parameters \{a_i\} a complete Bayesian scheme can be implemented. However, recent advances in the Monte Carlo simulations, posterior summaries can directly be obtained from simulating Dirichlet distribution. The typical scheme (MD) is

\[
\begin{align*}
n_i & \sim \text{Multinomial} (N,\{\pi_1, \pi_2, ..., \pi_k\}) \\
\pi & \sim \text{Dirichlet}(\theta_1, \theta_2, ..., \theta_k) \quad \text{so that}
\end{align*}
\]

www.ijisrp.org
Setting $\theta_j = 1$ ($j = 1, 2, \ldots, k$) will yield a uniform density and Tuyl et al (2009) have favoured this choice as a better non informative prior for $\{\theta_j\}_{j=1}^k$.

Further, a simulation study has been carried out to compare the performance of the intervals in terms of repeated experiments. Bayesian estimates obtained from incorporating objective priors might require such a test based on frequentist approach. Agresti and Min (2005) have attempted this in evaluating the Bayesian confidence intervals for binomial proportions. The corresponding procedure for multinomial proportions includes following steps:

1) Consider any data set with cell count $\{n_1,n_2,\ldots,n_k\}$
2) Compute its MLE $p = \{n_i/N\}$ and assume $p$ as population parameter
3) Simulate Multinomial$(N, p)$ for L times
4) Obtain confidence interval using the required methods
5) Coverage Probability = (Number of intervals in (iv) that include $p$) / L

Similar attempts have been made for classical approaches or Bootstrap intervals in literature that are cited earlier in this paper. This work includes Bayesian methods by considering contingency tables with non-zero but low counts and has an appreciable distance between the counts. However for comparison purpose other standard procedures such as QH-Quesenberry and Hurst (1964), GM-Goodman (1965), FS- Fitzpatrick and Scott (1987), SG- Sison and Glaz (1995) and methods due to central limit theorem (CLT) and its continuity corrected version (CLT-CC) have also been considered.

### III. MOTIVATING DATA SETS

If $X$ and $Y$ denote two categorical response variables, $X$ with $I$ categories and $Y$ with $J$ categories leading to $k = IJ$ possible combinations that can be represented in a contingency or cross-classification table with cells contain frequency counts of outcomes for a sample. As a case of a hypothetical example, suppose that a clinical trial is undertaken to compare the effect of a new drug or other therapy with the current standard drug or therapy. Ignoring side effects and other complications, the response for each patient is assumed to be simply “success” or “failure.” For a single stand-alone experiment, the observed data can be shown in the following table:

<table>
<thead>
<tr>
<th></th>
<th>Success</th>
<th>Failure</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>a</td>
<td>b</td>
<td>m</td>
</tr>
<tr>
<td>Control</td>
<td>b</td>
<td>d</td>
<td>n</td>
</tr>
<tr>
<td>Total</td>
<td>r</td>
<td>s</td>
<td>N</td>
</tr>
</tbody>
</table>

Sparse tables often contain cells having zero counts and such cells are called empty cells. Contingency tables are referred to as sparse when many cells have small frequencies besides some of them being zeros too. It is extremely important to describe the location of zero cells in the 2 x 2 table, as the same is also crucial in studying the nature of sparsity and could affect the analysis. Sparsity is not restricted to the tables with smaller sample sizes alone but could also occur with large sample size due to high concentration of frequencies in certain cells and poor or none in other cells. The impact of sparsity is felt in estimation of summary measures like odds ratio, computational complexity and asymptotic approximations. Even for large contingency tables, due to the small sample size and the resulting sparseness of the data table, the asymptotic distributions of the tests may not be relied in hypothesis testing (Szyda et al 2008).

The characteristics of the data sets (referred to as I to X) collected from the published literature with 2 x 2 tables are summarized to provide the length and breadth of the sparsity in the data sets. Table 2 provides the details of source and distribution of zero cells. Apart from zero cells, proportion of non-zero cell counts with frequency less than six is also described, so that the sparse nature of the data sets are completely described. Also, to understand the spread of counts in individual tables minimum and maximum of range calculated for each table in a data set is presented. This value provides a quick view of polarization of counts; for example data set V shows a very high range so that cell counts are extremely different in their sizes. Zero minimum indicates equal cell counts in a data set (Kishore, 2007), whereas Efron (1996) has a table with zero in all the four cells. Also, based on Subbiah and Srinivasan (2008) nature of sparseness of each of these data sets has been classified to indicate the typical real time data variability among the collected literature and the results are summarized in the same table.
he user d extreme non-sical methods are available. However, very limited or no studies have included Bayesian method in this comparison. Table 3 provides results from one data set as an illustrative case and subsequently observations from the comparative analysis have been presented.

Apart from these ten 2 x 2 tables, another contingency table (Szyda et al 2008) has been considered whose size is 4 x 5 of which 12 cells (60%) are zero where as minimum and maximum among remaining non-zero cells are 5 and 66 respectively. This data illustrates the presence of more zeros and extreme non-zero counts with high range even in a large size tables. These observations among many such real time studies provide a notion for comparative study using relevant characteristics which are prevalent in data sets summarized in contingency tables.

IV. RESULTS

Bayesian data analysis can be referred to posterior inference given a fixed model and data and computation has been carried out in WinBUGS and R. However, sufficient search indicates non availability of classical methods in open sources and these methods are implemented using Macros in EXCEL except SG which is obtained through SAS.

Results from the computations include lower and upper limits of 95% confidence intervals calculated from the closed form classical methods. 2.5 and 97.5 percentiles from posterior samples are used to obtain lower and upper limits of Bayesian confidence intervals after a run of 50000 single MCMC chain with burn-in of initial 50% and convergence has also been monitored using kernel density. However, Table 3 provides results from one data set as an illustrative case and subsequently observations from the comparative analysis have been presented. This data set has as many characteristic as desired in explaining the performance of these procedures; especially, under sparseness, low non-zero counts and the impact on corresponding results.

The comparisons are based mainly on length of intervals (shorter or wider), aberrations; many studies have considered coverage probability as a tool for comparing performance of intervals. However, very limited or no studies have included Bayesian method in this comparison and this study has considered Bayesian MD procedure and compare with existing results. The data characteristics such as sparseness in terms of presence of zeros and low cell counts range of cell counts in a table and size of the table. Though computation tools become abundant in the present scenario, these procedures require a keen attention in the availability to the user community.

### Table 3: Comparison of seven simultaneous confidence interval procedures with α = 0.05 for five different 2 x 2 tables

<table>
<thead>
<tr>
<th>Data No</th>
<th>Source of data sets</th>
<th>Zero entries</th>
<th>Positive entries &lt; 6</th>
<th>Range of table totals</th>
<th>No of tables with nature of Sparseness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>%</td>
<td>No</td>
<td>%</td>
</tr>
<tr>
<td>I</td>
<td>Kishore (2007)</td>
<td>5</td>
<td>18</td>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>II</td>
<td>Agresti (1990)</td>
<td>7</td>
<td>35</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>III</td>
<td>Smith et al (1995)</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>V</td>
<td>Sweeting et al (2004)</td>
<td>2</td>
<td>7</td>
<td>6</td>
<td>21</td>
</tr>
<tr>
<td>VI</td>
<td>Efron (1996)</td>
<td>16</td>
<td>10</td>
<td>43</td>
<td>26</td>
</tr>
<tr>
<td>IX</td>
<td>Cochran (1954)</td>
<td>4</td>
<td>25</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>X</td>
<td>Warn et al (2002)</td>
<td>17</td>
<td>9</td>
<td>18</td>
<td>10</td>
</tr>
</tbody>
</table>

Apart from these ten 2 x 2 tables, another contingency table (Szyda et al 2008) has been considered whose size is 4 x 5 of which 12 cells (60%) are zero where as minimum and maximum among remaining non-zero cells are 5 and 66 respectively. This data illustrates the presence of more zeros and extreme non-zero counts with high range even in a large size tables. These observations among many such real time studies provide a notion for comparative study using relevant characteristics which are prevalent in data sets summarized in contingency tables.
In terms of length of intervals for the data sets (I to X), SG (63%) and QH (31%) yield wider intervals compared to other methods. SG has the maximum length in most of the cases where range of cell counts are markedly as high as 6821. Even in such polarized tables, only small count cells have this property and QH produces long intervals for other cells of corresponding tables. Data set IV, VI, VII, and other tables are available with the presence of zeros in different position of four cells. This property is apparent in the data set III in which all cell counts are non-zero. MD provides a wider interval only in this case for the low counts and QH has shared this for other larger data sets with which this study has made extended comparisons.

While considering other methods, no case has an interval with maximum length due to FS. However, two methods based on CLT share this property in almost all cases in similar cases though CLT-CC yield wider interval in slightly more cases. However, these methods possess a feature in that for zero cells they provide intervals of zero length which is due to the presence of sample proportion values and except this case, MD has not shown this property in any of the other tables considered for the comparisons that are presented here and other data sets with which this study has made extended comparisons.

In the data set V, possibly a rare table with an extreme characteristic in that cell counts are too wider (10, 45870, 40, 66163) is available. MD provides a wider interval only in this case for the low counts and QH has shared this for other larger values and except this case, MD has not shown this property in any of the other tables considered for the comparisons that are presented here and other data sets with which this study has made extended comparisons.

While considering other methods, no case has an interval with maximum length due to FS. However, two methods based on CLT share this property in almost all cases in similar cases though CLT-CC yield wider interval in slightly more cases. However, these methods possess a feature in that for zero cells they provide intervals of zero length which is due to the presence of sample proportion in their mathematical form. But it has been observed that for tables with all low counts so that total is also marginally low, wider intervals could be due to CLT methods; a single table in data I and data set II that has uniformly low counts and at least one zero cell in all the tables illustrate this observation.

---

### CLT

<table>
<thead>
<tr>
<th></th>
<th>LL</th>
<th>UL</th>
<th>LL</th>
<th>UL</th>
<th>LL</th>
<th>UL</th>
<th>LL</th>
<th>UL</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.000</td>
<td>0.000</td>
<td>0.170</td>
<td>0.920</td>
<td>0.000</td>
<td>0.000</td>
<td>0.080</td>
<td>0.830</td>
</tr>
<tr>
<td>0^a</td>
<td>0.562</td>
<td>0^a</td>
<td>0.562</td>
<td>0.000</td>
<td>0.000</td>
<td>0.139</td>
<td>0.861</td>
<td></td>
</tr>
<tr>
<td>0.139</td>
<td>0.861</td>
<td>0.000</td>
<td>0.000</td>
<td>0.435</td>
<td>0^a</td>
<td>0^a</td>
<td>0.673</td>
<td></td>
</tr>
<tr>
<td>0.061</td>
<td>0.772</td>
<td>0^a</td>
<td>0.283</td>
<td>0.139</td>
<td>0.861</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>0.104</td>
<td>0.580</td>
<td>0.000</td>
<td>0.000</td>
<td>0.288</td>
<td>1.141</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
</tbody>
</table>

^aLower limit is less than zero ^bUpper limit is greater than one
In the case of shorter intervals, FS dominates uniformly in all the four cells of all data sets considered for the study; 72%, 95%, 60% and 94% of occasions are the supportive numerals for this property. In each case, CLT methods immediately succeed FS in this property but this may be due to its feature already mentioned and hence could be avoided from comparison. Surprisingly SG yield shorter intervals in two tables of the data set V where counts are extremely varying in nature (range 8632 and 66153). No other methods exhibit this property in any of cases considered for the comparative study.

It is observed that aberrations exist in three procedures due to CLT based methods and FS. But those cells cannot be identified with any particular characteristic of a cell like zero count. In the case of zero counts these methods will yield a degenerate case with lower and upper limits are same value. This feature is an obvious outcome of their mathematical forms. Also, a closer look of CLT indicates that the procedure will be resulted with a smoothing by the chi square value whenever cell counts are zero. This kind of smoothing would encourage the recommendation of Bayesian procedure as observed in Agresti (1992). Also, from Table 3 it can be observed that upper limit can also have estimates that are not possible for a proportion; in limits of CLT based intervals exceed one where as SG yields exactly one as upper limit where the observed proportion is quite nearer to one and as low as five.

Further nature of sparseness has been considered in understanding the performance of these methods in term of extreme lengths; three classifications of sparseness also demonstrate this behavior. QH and SG perform uniformly across these classifications and CLT based intervals provide wider intervals even in the case of mild sparse as well as all the four cells are with low counts. However, because FS dominates uniformly while comparing shorter intervals, nature of sparseness has not been considered in those cases.

The analysis schemes have been extended to a data set that has a 4 x 5 contingency table (Syzdaetal, 2008) with many zeros. Results have shown that no major changes in terms of longest interval are visible when compared to 2 x 2 tables. QH has dominated uniformly over all non zero cells in the table followed by SG and GM. However, unlike the case 2 x 2 table, this behavior does not distinguish between low or high non zero counts. Also, FS yields smaller intervals in all cases that may not be a required feature for an interval estimator. Bayesian method yields a better compromise estimates when compared to these methods with extreme values. The inevitable 0.0 as estimates for zero counts in the case of CLT methods are apparent for this data set too. But CLT-CC yields a negative lower limit for a case where the count is five. Hence when table size (k) or total counts (N) become large, the negative lower limit could appear in the case of cell counts over and above five.

Also, the outcome of simulation studies indicate a consistent behavior of Bayesian confidence intervals when compared to classical approach though MD intervals are uniformly narrower than other counterparts and achieves coverage probability less than 0.95. Agresti and Coull (1998) have pointed out that such property can also be preferred in certain cases and very wider intervals which may tend to provide very high coverage probability in most of the cases. This attempt includes another data of size 1 x 7 (Quesenberry and Hurst,1964) that has been used almost in all similar studies that is beyond the data sets considered in this Section. Figure 1 presents the illustrative details of the consistent behavior of MD and the extreme performance of QH, GM, and SG; CLT methods and FS are not considered for this comparison based on their performance that is observed earlier.

![Figure 1: A comparison of coverage probabilities for the nominal 95% QH, GM, SG and MD intervals for multinomial proportions](www.ijsrp.org)
All these data sets represent the varied feature of contingency tables so that methods can be compared for the performance of the methods. Extreme range of cell counts where classical methods are consistently closer to one whereas MD provides around 85% as its coverage probability in all such cases. From the figure it can be observed that there is a reversal tendency in the case of second set where cell counts are low and classical procedures have as low as 45%. Similar effect is also the case when cell counts are high but of notably apart from each other. Hence, in the absence of a perfect definition of sparseness in a general I x J tables, MD has a consistent behavior in terms of coverage probability though the numerical value is below the nominal value.

V. Conclusions
Multinomial proportions have found many applications and Burda et al (2008) have provided illustrative cases for multinomial discrete models. Lee et al (2011) have applied the methods for constructing confidence intervals for multinomial proportions to the design process of grain tracing and recall system. Kern (2006) has used a pig data to illustrate the Bayesian inference on multinomial probabilities. The present study has emphasized the need to understand, implement and compare the procedures in obtaining confidence intervals for multinomial proportions, especially when the data is sparse in nature.

In general, comparison and subsequent recommendation of any statistical procedures are based on their performance, wider availability to the users, computational issues, and aberrations. Further sparseness also plays an important role in deciding the procedure to be adopted. Agresti (1990) has stated that for sampling zeros, it is not sensible to use 0.0 as the best estimate of a probability. In view of this, classical methods, which yield 0.0 as estimator either because of its form or through auto corrections, need not be recommended if the data sets do have more zeros. Hence zero counts irrespective of other cell counts need a careful investigation in using an estimation method. Bayesian procedures even with a non informative prior yield estimates in such situations similar to its performance uniformly over 2 x 2 tables considered as a basic form in many categorical data studies.

Such similarity is consistently visible when the size of the table increases and range of the cell counts is relatively higher. However a classical method needs careful choice when k or N changes also when the cell counts are comparatively different. Further it may be easier in the present scenario to obtain a computation tool or mechanism, but still the availability of these classical methods is restricted to Wald type intervals. However, CLT based methods are not recommended when data is sparse irrespective of zero or non zero counts and size of the table.

Bayesian procedure has distinct advantages in obtaining the confidence limits without any aberrations; possessing acceptable frequentist coverage probabilities and practically important in computational flexibility and availability. It has been observed that exact inference plays important role in statistical inference of discrete data; however, for sparse data large sample chi-square statistics are often unrealistic (Agresti and Coull, 1998). More importantly, all of these conclusions are drawn for sparse data which is more realistic even for large contingency tables. Hence, this comparative study has emphasized the need to apply Bayesian methods with an objective prior for estimating multinomial proportions in categorical data with presence of zero or low cell counts and has an appreciable difference between the cell counts; in some cases the number of categories also plays a role to choose between methods. Bayesian methods have been identified as unified approach to handle varied situations of cell frequencies that would generally arise in the analyses of contingency tables and its applications.

References


AUTHORS

First Author – U. Sangeetha, M.Sc., M.Phil. Department of Management Studies, SSN College of Engineering, Chennai. usangee19@gmail.com.

Second Author – M. Subbiah, M.Sc., Ph.D. Department of Mathematics, L. N. Government College, Ponneri. sisufive@gmail.com.

Third Author – M.R. Srinivasan, M.Sc., MBA., Ph.D. Department of Statistics, University of Madras, Chennai. mrsvasan8@hotmail.com.

Correspondence Author – U. Sangeetha, usangee19@gmail.com, 0091-95000-79439