Evaluating Safety Signals in Clinical Trials: the Dirichlet-NBD Model Implementation with SAS

Yuqin Li, inVentiv Clinical Solutions, Indianapolis, IN

Xiaohai Wan, Eli Lilly and Company, Indianapolis, IN

ABSTRACT

In clinical research, it is often necessary to analyze a group of drug adverse events jointly as well as to analyze them individually. The Dirichlet-NBD model, a statistical model with the combination of both the Negative Binomial Distribution (NBD) and the Dirichlet Multinomial Distribution (DMD), can be applied to analyze adverse events jointly and individually. The model simultaneously fits incidence counts of each adverse event recorded over a period of time. In this paper, we describe an implementation of the Dirichlet-NBD model to analyze drug adverse events data with SAS®. First we give a brief introduction to the statistical theories behind the model. Then we describe the SAS macro developed using maximum likelihood method to estimate the parameters for both DMD and NBD. Based on the estimated distributional parameters, the SAS macro computes derived statistics which are useful to study either groups of adverse events or individual adverse event. These derived statistics are from marketing research literature and include metrics such as average event rate, event penetration, and event proportion. We also present an example to illustrate the application of the Dirichlet-NBD model to analyze MedDRA coded adverse events from a clinical study.

INTRODUCTION

The Dirichlet-NBD model is a well-known statistical model in marketing research [1, 2], but it is not often applied in clinical trial data analyses. Our objective of this paper is to introduce the Dirichlet-NBD model and demonstrate its application in clinical research. Safety data in clinical trials often is collected as adverse event occurrences for patients who participated in the study. Common statistical methods used to analyze adverse events data include Chi-square test and Fisher's exact test to compare event occurrences between treatment groups. The Dirichlet-NBD model provides additional statistics for data summarization and for statistical inference. These additional statistics include, but are not limited to, average event rate for patients with at least one event, average event rate for all patients, event penetration, and event repeat rate. To implement the Dirichlet-NBD model, we have three SAS macros in our program. The first macro estimates the distributional parameters for the Negative Binomial Distribution (NBD), and the second macro estimates distributional parameters for the Dirichlet Multinomial Distribution (DMD). The third macro calls the first two macros, and based on the estimated distributional parameters, computes the derived statistics. There are two popular approaches to estimate the parameters of DMD: method of moments and maximum likelihood. In this paper, we only include maximum likelihood method for better efficiency.

METHOD

THE MACRO %NBD

This macro estimates the parameters for NBD. It has three input parameters: indata, id, outdata. Indata is the dataset to be analyzed, id is the unique ID for each subject, and outdata is the output SAS dataset containing the NBD parameters. For the input dataset, all subjects should be included, even subjects without any events. In the macro we first prepare the data for PROC GENMOD, which is used to estimate the distributional parameters. The macro creates a variable called sumevt, representing the total event count for every patient. We use PROC GENMOD null model, i.e., the intercept-only model to get the parameter estimates. In this way we can get both the dispersion and intercept estimates, and from those we can calculate the alpha and beta parameters for NBD. The main part of this macro which obtains the parameter estimates for the NBD is the following:

```
proc genmod data=nbddata;
       model sumevt = / dist=negbin;
       ods output parameterestimates=pe;
       run;

proc transpose data=pe out=tpe;
       var estimate;
       id parameter;
       run;

data &outdata;
       set tpe;
       nb_k = 1/dispersion; ** alpha;
       nb_k2 = 1/dispersion - 1;
       nb_p = 1/(1+exp(intercept)*dispersion);
       beta=1/nb_p-1;  ** beta;
       nb_mean=nb_k*(1-nb_p)/nb_p;
       nb_var =nb_k*(1-nb_p)/nb_p**2;
       run;
```

THE MACRO %DMD

We did not find any SAS procedures in SAS 9.1 to estimate the parameters for DMD. Motivated by an R package [3], we use SAS IML to calculate the parameters for DMD. The following subroutines are the modules included in the macro:

dbbin_ab(x,n,a,b): Computes the Beta-Binomial probability density function;

loglik(x,t): Computes the log-likelihood function;

u(x,t): Computes the score function;

expfim(x,t): Computes the expected Fisher information matrix;

obsfim(x,t): Computes the observed Fisher information matrix;

dirmult(data, init, initscalar, epsilon, mode): Main module in the macro. It calls other modules to estimate the parameters in DMD. This module has the following parameters:
data: Name of the SAS dataset with adverse event counts. Rows represent patients and columns represent adverse events. If any rows or columns sum up to zero, they are removed.
init: Initial values for the gamma vector. Default is empty, implying the column-proportions are used as initial values.
initscalar: Initial value for theta. Default value is 30.
epsilon: Convergence tolerance. On termination, the difference of consecutive log-likelihood function values must be smaller than epsilon.
mode: This variable takes the macro variable &fimmode value. It takes values "obs" (default) or "exp" determining whether the observed or expected Fisher's Information Matrix (FIM) should be used in the Fisher Scoring algorithm. All other arguments will produce an error message, and the observed FIM will be used in the iterations.

The output from this module contains the following summaries:
gamma: A vector of estimates for the gamma parameters;
pi: A vector of π representing the estimated proportion for individual event;
theta: Estimated all-events polarization value.
The code to initiate gamma estimation is the following;

```
gamma = J(1,ncol(data)-1,0);
if init = 0 then gamma1 = data[+,]/(sum(data)/2)*initscalar;
else gamma1 = init;
do loop1=1 to ncol(data)-1;
```

```
        gamma[loop1]=gamma1[loop1];
    end;
```

The following code decides which FIM the program uses when it calculates gamma. Expected FIM is the matrix based on the beta-binomial probability density function. Another option is to use the observed FIM.

```
    if mode = '' then mode = 'obs';
    if (mode ^= 'obs' & mode ^= 'exp') then do;
        put @1 'Warning: Mode is not valid';
        mode = 'obs';
    end;
```

The following iteration step calculates the maximum log likelihood and DMD parameter gamma:

```
    lik1 = 0;
    lik2 = epsilon*10;
    ite = 1;
    conv=1;

    do while (conv=1);
        if (abs(lik2-lik1) < epsilon) then conv = 0;
        if mode = 'exp' then fim = expfim(data,gamma);
        else if mode = 'obs' then fim = -1 * obsfim(data, gamma);
        lik1 = loglik(data, gamma);
        rslt_u = (u(data,gamma))`;
        fim2=inv(fim);
        temphere = fim2*rslt_u;
        gamma = gamma + temphere`;
        do loop1 = 1 to ncol(gamma);
            if gamma[loop1] < 0 then gamma[loop1] = 0.01;
        end;

        lik2 = loglik(data,gamma);
        ite = ite + 1;

    end;

    sumgam = sum(gamma);
    theta = 1/(sumgam + 1); *all-events polarization;
    pi = gamma/sumgam;  *Proportion for individual event;
    a = gamma || pi || sumgam || theta;
    return (a);
```

THE MACRO %DIRICHLET_NBD

%dirichlet_nbd is a macro that calls macros %nbd and %dmd to calculate the parameters for both NBD and DMD. Based on the estimated distributional parameters, this macro further computes useful statistics based on marketing research literature to characterize adverse events. The input parameters for this macro are indata, subjectid, order and fimmode. Indata is the name of the SAS dataset containing event counts of each adverse event for every patient. The adverse event dataset must contain the following variables: subject_id, event1_count, event2_count, event3_count, etc. The values for each event count variable must be an integer greater than or equal to zero. SubjectID is the unique ID for each subject. Order is the group number we want to assign to this set of data, e.g., treatment group. Fimmode is the mode of Fisher's information matrix. It can be either 'exp' (expected Fisher's information matrix) or 'obs' (observed Fisher's information matrix).

The code to call %NBD and %DMD is as follows:

```
    %nbd(indata=&indata, id=&subjectid, outdata=nbd_pe&order);
    %dmd(indata=&indata, id=&subjectid, init=0, initscalar=30, epsilon= 10**(-12),
    fimmode=&fimmode, outdata=dmd_pe&order);
```

The derived parameters computed in %DIRICHLET_NBD are the following:

Average event rate: The adverse event rates averaged for the events. There is also a separate event rate for each event. The overall average event rate is the sum of each event rate.

Event polarization: if the overall average event rate for all patients is S, then event polarization is 1/(S+1).

Average event rate for AE patients: The event rate averaged over all patients who experienced at least one event.

Event penetration: The proportion of patients who had at least one event. There is a separate penetration for each individual adverse event and all events.

Loyalty: For each type of event, the proportion of patients who only experienced that event.

Event share: The ratio of total expected average frequency of a given event divided by total average frequency of all types of events (which is the sum of all gammas).

Event proportion: the ratio of total expected average frequency of a given event divided by total average frequency of all types of the events for patients with this event.

Repeat rate: The proportion of patients with a given event who will possibly have the same event next time.

Average unique events: average number of unique events for patients who experienced at least one event.

The following is the code used to calculate these derived parameters:

```
nc=(ncol(dmd_pe)-2)/2;
call symput('nc',strip(char(nc)));
col1 = 'gamma1':"gamma&nc";
col2 = "pi1":"pi&nc" ;
col3 =  'sumgamma' ;
col4 = 'theta';
col5 = col1 || col2 || col3 || col4;
mattrib dmd_pe colname=(col5) label={'DMD Parameter'};
mattrib nbd_pe colname=({'intercept' 'Dispersion' 'alpha' 'nbk2' 'nb_p'  'beta'
'nb_mean' 'nb_var'}) label={'NBD Parameter'};
alpha = nbd_pe[3];
beta = nbd_pe[6];
sumgam = dmd_pe[&nc*2+1];
gamma = dmd_pe[1:&nc] //  dmd_pe[&nc*2+1];
gamma = gamma`;

** define the column names for the derived measures;
col1 = "event1":"event&nc";
col2 = "all events";
cname = col1 || col2;

/* event share;    */
share = dmd_pe[&nc+1:&nc*2];
temp = 1;
share = share // 1;
share = share`;
mattrib share colname=(cname) label={'event share'};

/* average event rate for all patients   */
avger=gamma*alpha*beta/sumgam;
mattrib avger colname=(cname) label={'average event rate'};

/* penetration, event proportion   */
k=0;
p=exp(-(alpha)*log(1+beta));
penetration=J(1,&nc+1,0);
loyal=J(1,&nc+1, 0);
scr=J(1,&nc+1, 0);
```

```
do while (p<0.99999);
    k=k+1;
    pk=exp(log(gamma(alpha+k))+k*log(beta)-log(gamma(alpha))-log(gamma(k+1))-
    (alpha+k)*log(1+beta));
    p=p+pk;
    do j=1 to &nc;
    penetration[j]=penetration[j]+pk*(1-exp(log(gamma(sumgam))+log(gamma(sumgam-
    dmd_pe[j]+k))-    log(gamma(sumgam+k))-log(gamma(sumgam-dmd_pe[j]))));
    loyal[j]=loyal[j]+pk*exp(log(gamma(sumgam))+log(gamma(dmd_pe[j]+k))-
    log(gamma(sumgam+k))- log(gamma(dmd_pe[j])));
    scr[j]=scr[j]+pk*k*(1-exp(log(gamma(sumgam))+log(gamma(sumgam-dmd_pe[j]+k))-
    log(gamma(sumgam+k))-log(gamma(sumgam-dmd_pe[j]))));
    end;
end;

do j = 1 to &nc;
       loyal[j]=loyal[j]/penetration[j];
       scr[j]=avger[j]/scr[j];
end;
loyal[&nc+1]=.;
scr[&nc+1]=.;

*all events penetration;
penetration[&nc+1] = 1 - 1 / ((1+ beta)##alpha);
mattrib penetration colname=(cname) label={'Penetration'};
mattrib loyal colname=(cname) label={'loyalty'};

/* average event rate for AE patients;       */
evtfreq = J(1,&nc+1,.);
evtfreq = avger/penetration;
evtfreq[&nc+1] = alpha*beta/penetration[&nc+1];
mattrib evtfreq colname=(cname) label={'average event rate for AE patients'}

/* Average event types    */
portfolio = J(1,&nc+1,.);
temp=0;
do i=1 to &nc;
       temp=temp+penetration[i];
end;
portfolio[&nc+1]= temp/penetration[&nc+1];
mattrib portfolio colname=(cname) label={'portfolio'};

/* repeat rate        */
repeatrate = J(1,&nc+1,.);
do loop = 1 to &nc;
       repeatrate[loop] = (dmd_pe[loop]+1)/(dmd_pe[2*&nc+1] + 1);
end;
mattrib repeatrate colname=(cname) label={'repeat rate'};

/* event polarization    */
catpolarization = J(1, &nc+1, .);
catpolarization[&nc+1] = dmd_pe[2*&nc + 2];
mattrib catpolarization colname=(cname) label={'polarization'};
```

EXAMPLE

In this example we use the macro to process two datasets representing two different treatment groups. The macro will output the results side-by-side so that we can compare the results between the two groups. There are 3 types of adverse events in the data.

```
%dirichlet_nbd(indata=group1_pt, subjectid=usubjid, order=1, fimmode=exp);
%dirichlet_nbd(indata=group2_pt, subjectid=usubjid, order=2, fimmode=exp);
```

Table 1. Results:

parameters	therapy	event1	event2	event3	all events
Result from NBD:					
Alpha	group1				3.22729
	group2				2.83569
Beta	group1				0.62840
	group2				0.70030
Result from DMD:					
Gamma	group1	8.535	4.049	5.335	17.919
	group2	5.887	2.737	3.826	12.450
Result from DIRCHLET_NBD					
Loyalty	group1	0.357	0.221	0.253	
	group2	0.367	0.230	0.271	
Average event rate	group1	0.966	0.458	0.604	2.028
	group2	0.939	0.437	0.610	1.986
Average unique events	group1				1.654
	group2				1.631
Event proportion	group1	0.587	0.424	0.468	
	group2	0.590	0.427	0.480	
Average event rate for AE patients	group1	1.721	1.362	1.462	2.558
	group2	1.730	1.375	1.494	2.552
Penetration	group1	0.561	0.336	0.413	0.793
	group2	0.543	0.317	0.408	0.778
Event polarization	group1				0.053
	group2				0.074
Repeat rate	group1	0.504	0.267	0.335	
	group2	0.512	0.278	0.359	
Event share	group1	0.476	0.226	0.298	1.000
	group2	0.473	0.220	0.307	1.000

CONCLUSION

This paper demonstrates the development and application of the %dirichlet_nbd macro to calculate multiple statistics to help characterize adverse events occurrence in clinical studies.

REFERENCES

*1. Cam Rungie (*2003) How to Estimate the Parameters of the Dirichlet Model using Likelihood Theory in Excel, Marketing Bulletin, 2003, 14, Technical Note 3.
2. Cam Rungie and Gerald Goodhardt (2004) Calculation of Theoretical Brand Performance Measures from the Parameters of the Dirichlet Model, Marketing Bulletin, 2004, 15, Technical Note 2.
3. Torben Tvedebrink (2009-09-24) Estimation in Dirichlet-Multinomial distribution, R package 'dirmult'.

CONTACT INFORMATION

Yuqin Li
inVentiv Clinical Solutions, LLC.
Farm Bureau Building
225 South East Street, Suite 200
Indianapolis, IN 46202
Email: hli@inventivclinical.com
y.helen.li@gmail.com