Simulation Study on Exchangeability
and Significant Test on Survey Data
by
Yong Cao
Program in Statistical and Economic Modeling
Duke University

Date:_______________________
Approved:

___________________________
Surya Tapas Tokdar, Supervisor

___________________________
Kent P Kimbrough

___________________________
Sayan Mukherjee

Thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science in the Program in Statistical and Economic Modeling
in the Graduate School of Duke University

2015
ABSTRACT

Simulation Study on Exchangeability

and Significant Test on Survey Data

by

Yong Cao

Program in Statistical and Economic Modeling
Duke University

Date:_____________________

Approved:

___________________________
Surya Tapas Tokdar, Supervisor

___________________________
Kent P Kimbrough

___________________________
Sayan Mukherjee

An abstract of a thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the Program in Statistical and Economic Modeling in the Graduate School of Duke University

2015
Abstract

The two years of Master of Science in Statistical and Economic Modeling program is the most rewarding time ever in my life. This thesis acts as a portfolio of project and applied experience while I am enrolled in the Master of Science in Statistical and Economic Modeling program. This thesis will summarize my graduate study in two parts: Simulation Study of Exchangeability for Binary Data, and Summary of Summer Internship at Center for Responsible Lending. The project of Simulation Study of Exchangeability for Binary Data contains materials from a team project, which is jointly performed by Sheng Jiang, Xuan Sun and me. Abstracts for both projects are below in order.

(1) Simulation Study of Exchangeability for Binary Data

To investigate tractable Bayesian tests on exchangeability, this project considers special cases of nonexchangeable random sequences: Markov chains. Asymptotic results of Bayes factor (BF) are derived. When null hypothesis is true, Bayes Factor in favor of the null goes to infinity at geometric rate (true odds is not one half). When null hypothesis is not true, Bayes Factor in favor of the null goes to 0 faster than geometric rate. The results are robust under misspecifications. Simulation studies are employed to see the performance of the test when the sample size is small, prior beliefs change and true parameters change.
My summer internship deals with a survey data from Social Science Research Solution about auto financing. The dataset includes about one thousand valid responses and 114 variables for each response. My efforts on exploratory statistic analysis unfolded many interesting findings. For example, African Americans and Latinos are receiving 2.02% higher APR on average than white buyers, excluding the effects of relevant variables. And what’s more, a Fisher’s Exact Test of Significance is widely used to discover the significance of a series of variables. Results are presented in organized neat tables. Findings are included in weekly reports. One example finding is that warranty add-ons of a financed car has significant impacts on all three aspects of a loan, which is Annual Percent Rate, Loan Amount, and Monthly Payment.
## Contents

Abstract...............................................................................................................................................iv

List of Tables.........................................................................................................................................vii

List of Figures.........................................................................................................................................viii

Acknowledgements.................................................................................................................................ix

1. Simulation Study of Exchangeability for Binary Data ................................................................. 1
   1.1 Introduction........................................................................................................................................ 1
   1.2 Bayes Factor......................................................................................................................................... 2
   1.3 Generate Simulation Data .................................................................................................................. 3
   1.4 Simulation Study .................................................................................................................................. 3

2. Summary of Summer Internship at Center for Responsible Lending .......................................... 9
   2.1 Exploratory Analysis Results............................................................................................................... 9
   2.2 Fisher Exact Test of Significance ...................................................................................................... 12

Appendix A.............................................................................................................................................18

Appendix B.............................................................................................................................................21

References...............................................................................................................................................22
# List of Tables

Table 1: Example on Fisher Exact Test of Significance .......................................................... 13  
Table 2: Example of Fisher Exact Test of Significance ............................................................. 13  
Table 3: Summary of Significant of Add-ons ........................................................................... 15  
Table 4: Summary of Significant of Negotiation ................................................................. 16  
Table 5: Summary of Significance for Awareness of Negotiation ....................................... 17
List of Figures

Figure 1: Bayes Factor v/Sample Size ................................................................. 5
Figure 2: Bayes Factor vs Sample Size ............................................................... 5
Figure 3: Contour Plot of $P(BF<1)$ .................................................................. 6
Figure 4: Perspective Plot of $P(BF<1)$ ............................................................... 6
Figure 5: Bayes Factor vs different Prior under iid ............................................. 7
Figure 6: Bayes Factor vs different Prior under Markov Chain .......................... 7
Figure 7: Credit Tier vs Education Level .......................................................... 11
Figure 8: Credit Tier vs Race ........................................................................... 12
Acknowledgements

Thanks so much to everybody who has supported me, helped me and advised me, especially: Surya T. Tokda, Kent P. Kimbrough, Sayan Mukherjee, Mike West, Lori Rauch, David B. Dunson, Sheng Jiang, Xuan Sun, Dalvin Davis.
1. Simulation Study of Exchangeability for Binary Data

To investigate tractable Bayesian tests on exchangeability, this paper considers special cases of nonexchangeable random sequences: Markov chains. Asymptotic results of Bayes factor (BF) are derived: when null hypothesis is true, BF in favor of the null goes to infinity at geometric rate (true odds is not one half); when null hypothesis is not true, BF in favor of the null goes to 0 faster than geometric rate. The results are robust under misspecifications. Simulation studies are employed to see the performance of the test when the sample size is small, prior beliefs change and true parameters change. To illustrate the testing procedure, we provide applications to real datasets: air quality data, game record of Duke Men’s Basketball team, and stock price.

1.1 Introduction

The definition of infinitely exchangeable is the following: for random variable sequence $x_i$ and any finite permutation of the indices $\pi(.)$, we have $p(x_1, ..., x_n) = p(x_{\pi(1)}, ..., x_{\pi(n)})$. By de Finetti theorem, it is equivalent to model the probability as $x_i \sim Ber(\theta)$, while $\theta \sim P$, where $P$ is some probability measure. It is natural to put a prior on the probability prior, but due to special structure of binary data, we can show that we

---

1 Adapted from paper coauthored by Sheng Jiang, Yong Cao and Xuan Sun.
Department of Statistical Science and Department of Economics, Duke University
do not need to put a prior on the de Finetti measure since only the mean of this measure comes into the marginal likelihood calculation.

However, there is no clear representation for nonexchangeable random sequences, since the only thing we know about them is there exists one such permutation that \( p(x_1, \ldots, x_n) \neq p(x_{\pi(1)}, \ldots, x_{\pi(n)}) \). To remedy this, we investigate a surrogate class of random sequences: Markov chains. The main reasons for choosing Markov chains are twofold: firstly, Markov chains are very flexible and represent a large portion of random phenomena in practice, so studying Markov chains is of great value even though they are only part of the nonexchangeable sequences; secondly, simplicity of Markov chains makes them tractable to analytically derive theoretical results.

The Bayesian testing procedure proposed here is the same as standard Bayesian tests: using Bayes factor as the criteria to decide rejection or acceptance. The contribution of this paper is that we carefully study the asymptotic properties of the test under surrogate alternative hypotheses and that we provide simulations to investigate the test’s small sample performance.

### 1.2 Bayes Factor

According to Sheng Jiang’s relentless effort on Bayes Factor, I get the Bayes Factor in favor of the null hypothesis that the series is identical and independent
distributed.

\[
BF_{01} = \frac{p(H_0|Y)}{p(H_1|Y)} = \frac{B(c_1, d_1)B(c_2, d_2)}{B(a, b)B(1 + y_1, 2 - y_1) \cdot B(n_{00} + c_1, n_{01} + d_1)B(n_{10} + c_2, n_{11} + d_2)}
\]

### 1.3 Generate Simulation Data

In order to obtain simulation data, R command “rbinom” is used to generate Binomial sample.

\[
x_i = \begin{cases} 
1 \text{ with probability } p \\ 
0 \text{ with probability } 1 - p 
\end{cases}
\]

And also a mechanism to generate Markov Chain data is designed.

\[
x_{i+1} = \begin{cases} 
x_i \text{ with probability } p \\ 
1 - x_i \text{ with probability } 1 - p 
\end{cases}
\]

### 1.4 Simulation Study

First simulation is with fixed sample size (500), vary probability (Binomial or Markov Chain Probability) from 0.5 to 0.88, and scatterplot of Bayesian Factors against probability is shown in Figure 1. It is obvious that, for Markov Chain values, BF shrinks below 1 dramatically fast and BF for independent values also increase as probability increase.

Second simulation fixed the probability to be 0.57, and then vary sample size from 50 to 1000 to see how the BF changes. Figure 2 suggests BF shrinks under 1 as sample size increase for Markov Chain values, but increases fast above 1 as sample size increase for independent values.
Additionally, Figure 1 and 2 shows the scatterplot of Bayes Factor for Binomial series on the left and first order Markov series on the right.

Overall, Figure 1 and Figure 2 suggest that, this BF works out significantly well for Markov Chain simulators. In probability increase, or sample size increase, it is very easy for BF to go less than 1. BF is very sensitive to both the probability parameter p and sample size parameter.

Next, another simulation computed P(BF<1) for each pair of Sample Size (from 50 to 1000) and p value (from 0.5 to 0.6). And Figure 3 shows the contour plot for P(BF<1). Figure 4, the perspective plots visualize the relationship between p-value, sample size and P(BF<1) more clearly.

Figure 3 and Figure 4 suggests P(BF<1) in Markov Chain is not very sensible to parameter probability, but very sensible to sample size. As sample size increase, P(BF< 1) increase rapidly. P(BF<1) in Binomial series is already very low even with sample size of 50.
Figure 1: Bayes Factor vs Sample Size

Figure 2: Bayes Factor vs Sample Size
Figure 3: Contour Plot of $P(BF<1)$

Figure 4: Perspective Plot of $P(BF<1)$
Figure 5: Bayes Factor vs different Prior under iid

Figure 6: Bayes Factor vs different Prior under Markov Chain
In terms of priors, I randomly generated a set of data, with sample size to be 200, probability to be 0.55, both for Binomial and Markov Chain situations. Figure 5 and Figure 6 separately show the Bayes Factor versus prior a, BF versus prior c, BF versus prior d, with other priors to be fixed as 20 in both Binomial and Markov Chain situations.
2. Summary of Summer Internship at Center for Responsible Lending

During summer internship, I worked on a survey data from Social Science Research Solution about auto financing. Over 4000 people were surveyed through mobile and landline phones, in both English and Spanish. The dataset includes 946 valid responses and 114 variables for each response. Variables include survey questions variables and background variables, including state, county, kids number, income, age, education level, race, sex, Annual Percent Rate, monthly payment, loan amount, loan length, down payment, trade-in value, awareness and practice of negotiation, etc. My efforts on exploratory statistic analysis unfolded many interesting findings. And what’s more, a Fisher’s Exact Test of Significance is widely used to discover the significance of a series of variables. Results are presented in organized neat tables and graphs.

2.1 Exploratory Analysis Results

Thanks to the larger dataset. A lot of interesting findings are discovered. Many of them will have great influence on campaign raised by Center for Responsible Lending.

- African Americans and Latinos receive 2.02% higher APR than white on average, exclude the effects of relevant variables.

- Auto loans for used cars gain 1.64% higher APR than new car loans; exclude the effects of relevant variables.
• Being in poor credit tier has average treatment effect of paying more interest as much as 7.56% of the loan amount.

• There is significant negative correlation between loan amount and APR. On average, every $10,000 more loans can drop APR down by 0.5323%.

• Loan over six years got significant higher APR than those less than six years.

• Three out of six add-ons surveyed suggest significant impact on APR, with markup of about additional 0.93%-1.35% on APR. (Extended Warranty, Guaranteed Automobile Protection and Credit life and Disability insurance)

• Analysis suggested that negotiation on APR, Monthly Payment and Down Payment didn’t help lower the APR, sometimes even increase it.

• Car buyers who negotiated on MSRP (tag price) and Add-ons’ price got significant lower APR than those did not (about 1.2% lower).

   Apart from the above findings, there are two visualization findings, which are interesting as well. Figure 7 suggests that car buyers with education level less than high school and technical or other school are very likely to be well below average credit rate than other education levels. Higher education always accompanies higher credit score in auto financing population, and drop loan APR together. And the proportion of well above average increase significantly as education level increase (less than high school -> technical/others -> high school -> some college -> graduated college -> graduate school).
The Figure 8 suggests the probability of well below average in white buyers is much less than other races. And around 0% of white buyers have well above average credit rate, which is significantly higher than other races.

Figure 7: Credit Tier vs Education Level
2.2 Fisher Exact Test of Significance

Due to the large amount of variables, I chose Fisher Exact Test of Significance to analyze the significance on the variable one by one. The Fisher Exact Test of Significance can be explained in example. Take Annual Percent Rate and Warranty as an example. Suppose I am interested in the difference between means for those buy and not buy warranty add-on.

The example in Table 1 has a Test statistic of \((4+2+2.9)/3\) - \((4+1.25+2.5)/3\) = 2.96 - 2.58 = 0.38%. Then a permutation is conducted for the Yes and No on warranty variable, which results in the Table 2.
Table 1: Example on Fisher Exact Test of Significance

<table>
<thead>
<tr>
<th>Individual</th>
<th>APR %</th>
<th>Warranty</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>YES 1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>YES 1</td>
</tr>
<tr>
<td>3</td>
<td>2.9</td>
<td>YES 1</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>NO 0</td>
</tr>
<tr>
<td>5</td>
<td>1.25</td>
<td>NO 0</td>
</tr>
<tr>
<td>6</td>
<td>2.5</td>
<td>NO 0</td>
</tr>
</tbody>
</table>

In Table 2, 8 out of 20 combinations show more extreme test-statistics than 0.3833. So the p-value should be roughly \(\frac{8}{20} = 40\%\). The impact is significant when the p-value is less than 5%.

Table 2: Example of Fisher Exact Test of Significance

<table>
<thead>
<tr>
<th>Individual</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Test-statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Warranty</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.38333333</td>
</tr>
<tr>
<td>Warranty</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1.11666667</td>
</tr>
<tr>
<td>Warranty</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-0.71666667</td>
</tr>
<tr>
<td>Warranty</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.11666667</td>
</tr>
<tr>
<td>Warranty</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1.71666667</td>
</tr>
<tr>
<td>Warranty</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-0.1166667</td>
</tr>
<tr>
<td>Warranty</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.71666667</td>
</tr>
<tr>
<td>Warranty</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.61666667</td>
</tr>
<tr>
<td>Warranty</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1.45</td>
</tr>
<tr>
<td>Warranty</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>-0.3833333</td>
</tr>
</tbody>
</table>
Using the Fisher Exact test above, summarization of the findings is in the Table 3, 4 and 5. Be aware, sign indicate the difference between the mean values for those bought the adds-on and those didn’t is positive or negative. “Two” indicates the significance of two-side hypothesis test (null hypothesis: mean for those bought adds-on is equal to the mean for those who didn’t). Significance in two-side test suggests the means for with and without adds-on are significantly unequal. However one-side significance suggests the mean for with adds-on is significantly greater (+) (or less (-)) than the mean for those without adds-on.

The * suggests the significance. * indicates p-value 0.05-0.01. ** indicates p-value 0.01-0.001, *** indicates p-value less than 0.001. Blank indicates no significance.
Table 3: Summary of Significant of Add-ons

<table>
<thead>
<tr>
<th></th>
<th>APR</th>
<th>Loan Amount</th>
<th>Monthly Payment</th>
<th>Interest Paid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sign</td>
<td>Two</td>
<td>One</td>
<td>Sign</td>
</tr>
<tr>
<td>Warranty</td>
<td>+</td>
<td>**</td>
<td>**</td>
<td>+</td>
</tr>
<tr>
<td>GAP</td>
<td>+</td>
<td>*</td>
<td>*</td>
<td>+</td>
</tr>
<tr>
<td>Theft</td>
<td>+</td>
<td>-</td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>Credit life</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>-</td>
</tr>
<tr>
<td>Tire</td>
<td>+</td>
<td></td>
<td></td>
<td>+</td>
</tr>
<tr>
<td>Accessories</td>
<td>-</td>
<td></td>
<td></td>
<td>+</td>
</tr>
</tbody>
</table>

The Table 3 suggests:

1. Warranty add-ons have the widest impact on all three aspect of a loan. Warranty may include additional costs in auto loan.

2. Although accessories have very strong impact on monthly payment and loan amount, it is not persuasive to conclude influence on the loan. Since at many dealers, accessories are not as mandatory as warranty. Auto buyers can choose to buy depends on their financial ability. It is easy to believe people who bought expensive car (those with higher loan amount and can tolerate higher monthly payment) prefer buying accessories. No significance for accessories on APR, and even negative sign in APR reaffirms this conclusion, because people who afford expensive car always have good credit score, so therefore have comparatively lower APR on average.
The induced variable Interest Paid shows no significance at all. This maybe is caused by two reasons: banks earn the interests but dealers’ earning is not from interests, and the computation of Interest Paid has problem.

More significance tests are done for negotiation and awareness of negotiation variables in Table 4 and Table 5, but meaningful conclusions are hard to reach.

**Table 4: Summary of Significant of Negotiation**

<table>
<thead>
<tr>
<th>APR</th>
<th>Loan Amount</th>
<th>Monthly Payment</th>
<th>Interest Paid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sign</td>
<td>Two</td>
<td>One</td>
<td>Sign</td>
</tr>
<tr>
<td>APR</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Month Pay</td>
<td>+</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>MSRP</td>
<td>-</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>Trade-in</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Down Pay</td>
<td>+</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>Loan Term</td>
<td>+</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Add-ons cost</td>
<td>-</td>
<td>*</td>
<td>**</td>
</tr>
</tbody>
</table>
Table 5: Summary of Significance for Awareness of Negotiation

<table>
<thead>
<tr>
<th></th>
<th>APR</th>
<th>Loan Amount</th>
<th>Monthly Payment</th>
<th>Interest Paid</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sign</td>
<td>Two</td>
<td>One</td>
<td>Sign</td>
</tr>
<tr>
<td>APR</td>
<td>-</td>
<td>***</td>
<td>***</td>
<td>+</td>
</tr>
<tr>
<td>Month Pay</td>
<td>-</td>
<td>***</td>
<td>***</td>
<td>+</td>
</tr>
<tr>
<td>MSRP</td>
<td>-</td>
<td>***</td>
<td>***</td>
<td>+</td>
</tr>
<tr>
<td>Trade-in</td>
<td>-</td>
<td>***</td>
<td>***</td>
<td>+</td>
</tr>
<tr>
<td>Down Payment</td>
<td>-</td>
<td>***</td>
<td>***</td>
<td>+</td>
</tr>
<tr>
<td>Loan Term</td>
<td>-</td>
<td>***</td>
<td>***</td>
<td>+</td>
</tr>
<tr>
<td>Add-ons cost</td>
<td>-</td>
<td>***</td>
<td>***</td>
<td>+</td>
</tr>
</tbody>
</table>
Appendix A Code for Simulation Study

### Define Function for log(the power of X)

```r
logpower=function(x){
  logpower=0
  for (i in 1:x){
    logpower=logpower+log(i)
  }
  return(logpower)
}
```

### Define Markov Chain generator for length n, probability p

```r
rmk=function(n,p){
  v=1
  for (i in 2:n){
    if(runif(1,0,1)<p){
      v=c(v,v[i-1])
    } else{
      v=c(v,1-v[i-1])
    }
  }
  return(v)
}
```
### Function for Bayesian Factor, a c d is priors

```r
BF = function(data, a, c, d) {
    b = a
    c1 = c2 = c
    d1 = d2 = d

    # count the transition types and number of 1, 0
    n = length(data)
    n1 = sum(data)
    n0 = n - n1
    data2 = data[1:n-1] + data[2:n]
    n11 = sum(data2 == 2)
    n00 = sum(data2 == 0)
    n01 = sum(data2 == 1 & data[1:n-1] == 0)
    n10 = sum(data2 == 1 & data[1:n-1] == 1)
    p1 = n1 / n
    p0 = n0 / n
    p11 = n11 / (n11 + n10)
    p10 = n10 / (n11 + n10)
    p01 = n01 / (n01 + n00)
}
```
p00=n00/(n01+n00)

### Compute Bayesian Factor

if (a<3 | c<3 | d<3){

left=beta(c1,c2)*beta(d1,d2)/(beta(a,b)*beta(1+data[1],2-data[1]))

logright=logpower(a+n1+1)+logpower(b+n0-1)-logpower(a+b+n1+n0-1)-
logpower(n00+c1-1)-logpower(n01+c2-1)-logpower(n10+d1-1)-logpower(n11+d2-1)+logpower(n11+n10+d1+d2-1)+logpower(n01+n00+c1+c2-1)

value=left*exp(logright)
}

else{

logbf=(logpower(c1-1)+logpower(c2-1)-logpower(c1+c2-1)+logpower(d1-1)+logpower(d2-1)-logpower(d1+d2-1))- 
(logpower(a-1)+logpower(b-1)-logpower(a+b-1)+log(beta(1+data[1],2-data[1])))+logpower(a+n1-1)+logpower(b+n0-1)-logpower(a+b+n1+n0-1)-
logpower(n00+c1-1)-logpower(n01+c2-1)-logpower(n10+d1-1)-logpower(n11+d2-1)+logpower(n11+n10+d1+d2-1)+logpower(n01+n00+c1+c2-1)

value=exp(logbf)

}

return(value)
Appendix B Code for Fisher Exact Test of Significance

```r
var1=label2[,62]
var2=label2[,68]

fisher<-function(var1,var2){
  select=(!is.na(var1) & !is.na(var2))
  VAR1=var1[select]
  VAR2=var2[select]

  test=mean(VAR1[VAR2=='Yes'])-mean(VAR1[VAR2!='Yes'])
  k=0
  one=0

  for (i in 1:100000){
    sam=sample(VAR2)
    test2=mean(VAR1[sam=='Yes'])-mean(VAR1[sam!='Yes'])
    if (abs(test2)>=abs(test)) {k=k+1}
    if (sign(test2)==sign(test) & abs(test2)>=abs(test)) {one=one+1}
  }

  return(list(mean(VAR1[VAR2=='Yes']),mean(VAR1[VAR2!='Yes']),test,k/100000,one/100000))
}
```
References


