American Journal of Theoretical and Applied Statistics
2023; 12(3): 51-65
http://www.sciencepublishinggroup.com/j/ajtas
doi: 10.11648/j.ajtas.20231203.13
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

SciencePG
Science Publishing Group

Monte Carlo Simulation and Derivation of Chi-Square Statistics

Tofik Mussa Reshid

Department of Statistics, Werabe University, Werabe, Ethiopia

Email address:
toffamr@gmail.com

To cite this article:
Tofik Mussa Reshid. Monte Carlo Simulation and Derivation of Chi-Square Statistics. *American Journal of Theoretical and Applied Statistics*. Vol. 12, No. 3, 2023, pp. 51-65. doi: 10.11648/j.ajtas.20231203.13

Received: April 6, 2023; **Accepted**: May 22, 2023; **Published**: June 27, 2023

Abstract: Computer simulation has become an important tool in teaching statistics. Teaching using computer simulation would enhance the understanding of the concept using visual illustrations. This paper describes how to use simulation in R-programming language to perform a chi-square test. We try to show the distribution of most commonly used chi-square statistics we often found in statistical methods in both derivation and simulation. In statistical methods in such cases as test of independency, test of goodness of fit, test of significance, log likelihood ratio test, significance test and model selection we use chi-square statistic. The approach of the paper will enhance the students' and researchers' ability to understand simulation and sampling distribution. The paper contains an expository discussion of chi-square statistic, its derivation and distribution and its derivatives such as t-distribution and F-distribution. We consider two chi-squares, the empirical chi-square statistic and the theoretical chi-square distribution. The empirical distribution of chi-square statistic agrees closely with the theoretical chi-square distribution for large simulations, only the empirical distribution near to zero has lower density compared to the theoretical one for one degree of freedom. This is because the theoretical chi-square distribution at 1 degree of freedom has infinite density near to zero, but for any number of simulation the empirical distribution has finite density near to zero. Chi-square itself turns to normal distribution as the degree of freedom is large.

Keywords: Chi-Square Distribution, Chi-Square Statistic, Likelihood Ratio, T-test, F-test, Simulation

1. Introduction

In inferential statistics chi-square distribution or chi-square tests are among the most useful and most widely used tests. Chi-square test has been applied in all research areas. The application of statistical test in scientific research has increased dramatically in recent years almost in every science [5]. Chi-square distribution is used in many occasions such as goodness of fit test, test of independence, test of homogeneity, hypothesis testing, confidence intervals, likelihood ratio test, log rank test, and Cochran-mantel Haenszel test. The main characteristics of these tests are present along with various problems related to their application.

Chi-square distribution is one of the most widely used distribution and most frequently encounter in statistical methods. Other most widely used distributions such as student's t-distribution and Fisher F-distribution are derived from chi-square distribution. Chi-square test is introduced by Karl Pearson was the subject of debate for much researches [3]. It is well known that Pearson chi-square is a family of tests with the following assumption (1) the data are randomly drawn from a population (2) the sample size is sufficiently large. There is no accepted cut-off for the sample size the minimum size varies from 20 to 50. [3]. (3) the values on the cell are adequate when no more than 1/5 of the expected values are smaller than five and there is no cell with zero counts [3].

The main aim of this paper is to create visual figure in students and in researchers regards to sampling distribution and Monte Carlo simulation of chi-square statistic. Teaching students using simulation is widely recommended but rarely evaluated. Students find the concept of sampling distribution difficult to understand. Specifically, this article tries to explore the concept of Monte Carlo simulation of chi-square statistic we often found in statistical methods. Which helps students and researchers the concept of simulation sampling distribution along with its derivation. It has immense role in research methodology. It tries to show the basic derivation of chi-square statistics and its derivatives such as t-distribution and F-distribution.

The objective of this paper is to show the proof of the distribution chi-square statistic and other distributions derived from chi-square distributions such as, t-distribution and F-distribution in addition to simulation. Simulation in this paper is such that drawing samples in an experiment over and over again it tends to reveal certain pattern. This paper drive the distribution of chi-square statistic we often found in statistical methods and the sampling distribution of the statistic using Monte Carlo simulation. It provides a tool for sampling distribution and simulation in R programing. It gives a good understanding for students in simulation and R syntax code. Copy the code and paste into R will run the program. It gives brief description to students and researchers how to make inference using chi-square distribution and how to handle computer simulations. As computer become more readily available to educators there is wide speculation that teaching inference via dynamic, visual simulations may make statistical inference more accessible to introductory students [13].

Simulation has wide application in statistical works. In this paper the distributional characteristics of chi-square statistics is analyzed using simulation. We consider two types of chi-squares. Chi-square statistics and chi-square distribution (theoretical distribution). The theoretical chi-square distribution is continuous distribution but chi-square statistic is obtained in discrete manner based on discrete difference between observed and expected values [3]. Chi-square statistics turns to theoretical chi-square distribution as the number of observations (in this case number of simulation) increases.

Chi-square statistic is frequently used in statistical analysis of experimental data. Using chi-square distribution (chi-square test) is widely used but rarely evaluated and proved. This paper has a tool for sampling distribution, simulation, and derivation for chi-square statistics. It presents solution to common problem when applying the chi square test of goodness of fit, test of independency, test of homogeneity, test of significance and model selection. It is used to handle how these inferences are drawn using chi-square distribution and chi-square test. The paper is prepared intended to aid students the concept of simulation, sampling distribution and derivation of chi-square statistics.

2. Material and Methods

Let Z is standard normal distribution then $X = Z^2$ is chi-square distribution with *1* degrees of freedom as its density described below. The proof is very simple, using probability transform we can compute the distribution of X as given in equation (1) below.

$$f(x) = \frac{1}{\Gamma(1/2)2^{\frac{1}{2}}} x^{-\frac{1}{2}} e^{-\frac{x}{2}}, \quad \text{for } x > 0 \qquad (1)$$

Let $Z_1, Z_2, \ldots Z_k$ are independent standard normal random variable, then $X = Z_1^2 + Z_2^2 + \ldots Z_k^2$ is chi-square distributed with *k* degrees of freedom, denoted as $\chi^2_{(k)}$. The chi-square distribution with *k* degrees of freedom for positive integer *k* is given by equation (2)

$$f(x) = \frac{1}{\Gamma(k/2)2^{\frac{k}{2}}} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}, \quad x \geq 0 \qquad (2)$$

Chi-square distribution with *k* degrees of freedom is special case of gamma distribution with parameter *k/2* and *2*. When *k=2*, chi-square distribution is exponential distribution with parameter *1/2*. The minimum value of chi-square is 0 and there is no maximum value. The maximum of chi-square density occurs at $n-2, \quad n \geq 2$. The density of chi-square distribution with *k* degrees of freedom for different *k* is given in *Figure 1*. The figure shows the density of chi-square distribution with different degrees of freedom for *k*-values of *1, 2, 3, 5, 12, 21,* and *30*.

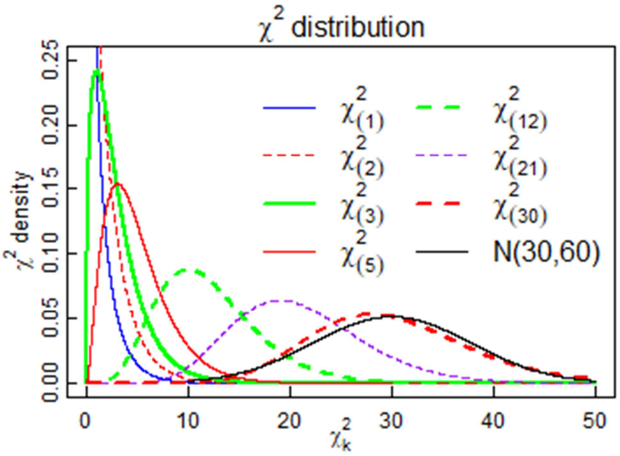

Figure 1. *Chi-square distribution for some values of k and normal approximation.*

Along horizontal axis is chi-square value and at the ordinate is its density of chi-square values. Chi-square distribution is highly skewed to right. But as the degrees of freedom increase it tends to normal distribution. We can see from the *Figure 1.* as degrees of freedom increases it looks like normal distribution. If the degrees of freedom *k* of chi square increases, it turns to normal distribution of of parameter mean *k* and variance *2k*. If *k >30* the maximum approximation error of chi-square by normal distribution is less than *0.7%*. If we have degrees of freedom *k >50* the maximum approximation error is less than *0.04%*. For large number of degrees of freedom *k*, the chi-square distribution may be approximated by normal distribution [1].

Chi-square can have a degree of freedom any value greater than zero. A random variable $\chi^2_{(k)}$ distribution especially, *k<1* can be represented by a random variable with Generalized Gaussian distribution [7].

Chi-square distribution has many important properties. These properties are derived from its density. *Table 1* shows some of the characteristics of theoretical chi-square distribution.

Table 1. *Characteristics of chi-square random variable with k degrees of freedom.*

Property	**Formula**	$\chi_{(k)}{}^{2}$
Mean	$E(X)=\int_{0}^{\infty}\frac{1}{\Gamma\left(\frac{k}{2}\right)2^{\frac{k}{2}}}x^{\frac{k}{2}+1-1}e^{-\frac{x}{2}}dx$	k
Median	$\int_{0}^{m}f(x)dx=0.5$	$m\approx k\left(1-\frac{2}{9k}\right)^{3}$
Mode	$\frac{d}{dx}f(x')=0$	*x'=Max (k-2, 0)*
Variance	$Var(X)=\int_{0}^{\infty}\frac{1}{\Gamma\left(\frac{k}{2}\right)2^{\frac{k}{2}}}x^{\frac{k}{2}+2-1}e^{-\frac{x}{2}}dx-k^{2}$	$2k$
Moment	$E(X^{n})=\int_{0}^{\infty}x^{n}f(x)dx$	$2^{n}\frac{\Gamma(k/2+n)}{\Gamma(k/2)}$
Skewness	$E\left(\frac{X-k}{\sqrt{2k}}\right)^{3}$	$\sqrt{\frac{8}{k}}$
Kurtosis	$E\left(\frac{X-k}{\sqrt{2k}}\right)^{4}$	$\frac{12}{k}$
MGF	$M_{X}(t)=E\left(e^{tx}\right)$	$(1-2t)^{-\frac{k}{2}},\quad for\quad t<1/2$
CF	$\Phi(t)=E\left(e^{itx}\right)$	$(1-2it)^{-\frac{k}{2}},\quad for\quad t<1/2$

The mean of chi-square distribution is its degrees of freedom. The variance is twice its degrees of freedom. The median and the mode are less than its mean. Skewness and kurtosis tells us as the degree freedom of chi-square increase it behaves as normal distributions. Moment generating function and characteristic function implies if X_1 and X_2 are independent chi square distributions with k_1 and k_2 degrees of freedom, then the distribution of X_1+X_2 is chi-square distribution with k_1+k_2 degrees of freedom.

Let $X_1, X_2, ...X_k$ are *k* independent chi square distributions then $Y=X_1+X_2+....X_k$ is chi square distribution with degrees of freedom individual sum of degrees of freedom. This situation arises in the sampling distribution of the sample variance S^2. For example $(n-1)\frac{S^2}{\sigma^2}=\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{\sigma^2}\sim\chi_{(n-1)}{}^{2}$ is chi square distribution with *n-1* degrees of freedom. If the population mean μ is known then $\frac{\sum_{i=1}^{n}(xi-\mu)^{2}}{\sigma^2}\sim\chi_{(n)}{}^{2}$ is chi-square distribution with *n* degrees of freedom. We lost a degree of freedom when we use the sample mean rather than the true mean. The simulation for this is given in "sampling distribution and simulation in R" [6].

2.1. Goodness of Fit Test

The first type of chi-square test is the goodness of fit test. This is a test which makes statement or claim concerning the nature of the distribution for the whole population. This claim is called null hypothesis. The null hypothesis is a particular claim concerning how the data is distributed. This test is used to examine in order to see whether this distribution is consistent with hypothesized distribution of the population or not. The chi-square goodness of fit test begins by hypothesizing that the distribution of the variable behaves in a particular manner.

Suppose that a variable has a frequency distribution with *k* categories in to which the data has been grouped. For each category there is observed number of cases (O_i for $i=1,2,...k$) and expected number of cases (E_i for $i=1,2,...k$). Observed number of cases are frequency of the sample in each category. The expected number of cases are the number of observations calculated to be found in the sample if the hypothesized statement is true. The null hypothesis is that the observed number of cases is exactly equal to the expected number of cases in each category in the population.

The *n* observation in a random sample from population are classified in to *k* mutually exclusive classes. Let the population has *k* categories, for each category there is a probability p_i attached to it and the expectation of cases $\mathrm{E_i} = \mathrm{np_i}$ for $i = 1, 2, ...k$

We are about to test a hypothesis $H_0 : O_i = E_i$ For all i Vs $H_1 : O_i \neq E_i$ *for* atleast one i

To test this claim we sample the population and observe the number of cases in each category. If the null hypothesis is true, the number of cases in each category close to the expected number of cases. Then the statistics χ^2 calculated below in equation (3) is follow theoretical $\chi^2{}_{(k-1)}$ distribution. We are considering two chi-squares. Chi-square statistics which can be calculated from equation (3) and the theoretical chi-square distribution.

$$\chi^2 = \frac{\sum_{i=1}^{k}\left(O_i - E_i\right)^2}{E_i} \tag{3}$$

Where O_i is called observed number of frequency for i^{th} category and E_i is expected number of frequency for i^{th} category calculated from theoretical distribution (frequency based on the assumption of the null hypothesis). If the null hypothesis is true, the observed number of cases and the expected number of cases should be the same for all categories. The only assumption required for conducting this test is that each of the expected number of cases reasonably large. $E_i \geq 5$, for $i = 1, 2, ...k$

To prove χ^2 statistic calculated in equation (3) follow $\chi^2{}_{(k-1)}$ first we consider central limit theorem. From central limit theorem for sufficiently large sample size n

$$Z = \frac{estimate - parameter}{\sqrt{Var(estimate)}} \sim N(0,1) \tag{4}$$

Let p_i is the probability of i^{th} category and $\hat{p}_i$ is the estimate of p_i then plugin in to equation (4)

$$Z = \frac{\hat{p}_i - p_i}{\sqrt{p_i(1-p_i)/n}} \sim N(0,1) \tag{5}$$

If we assume $\frac{p_i{}^2}{n} \approx 0$ and multiplying both numerator and denominator of equation (5) by n

$$Z^2 = \frac{(n\hat{p}_i - np_i)^2}{np_i} \sim \chi^2{}_{(1)} \tag{6}$$

$n\hat{p}_i = O_i$ and $E_i = np_i$ hence this completes the proof

Or alternatively from Poisson approximation to binomial let $X_i \sim binom(n, p_i)$ where X_i the number of cases in i^{th} category then the distribution of X_i is approximated by Poisson (np_i). $E(X_i) = np_i$ and $Var(X_i) = np_i$. Then putting this in to equation (4) the same result will be obtained

$$Z = \frac{X_i - np_i}{\sqrt{np_i}} \sim N(0,1)$$

Then $Z^2 = \frac{(X_i - np_i)^2}{np_i} = \frac{(O_i - E_i)^2}{E_i} \sim \chi^2{}_{(1)}$ (7)

Hence $\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2{}_{(k-1)}$ The degrees of freedom is such that the number of independent terms in the summation equation. Since we have $\sum_{i=1}^{k} O_i = \sum_{i=1}^{k} E_i = n$ the number of terms in simplified equation is $k-1$. The degree freedom of the chi-square statistic is *(k-1)*. When the observed data does not conform to what expected on the basis of null hypothesis the difference between the observed and the expected value is large.

Theoretical distribution is a continuous distribution. But chi-square statistics is obtained in discrete manner on the basis of discrete difference between the observed and expected values. We calculate probabilities using both experimental and theoretical methods. Then at the end of the time when it is important to determine how well the experimental value matches the theoretical values.

To simulate the process in order to verify the probabilistic behavior of the resulting statistic, suppose we have a die of *k* faces that we are curious if it is fair or not. If it is fair, then the probability of each value should be the same with probability *1/k*. Hence the number of each face expected is *n/k*. where *k* is the number of faces of a die and *n* is the number of tosses. The R-syntax code for the simulation for k=6 and k=10 is given below.

```
f=function(N,K,n){#goodness of fit
c=matrix(0,N);o=matrix(0,N,n)
e=o;d=e
for(i in 1:N){
O=sample(1:n,K,rep=TRUE)
for(j in 1:n){
o[i,j]=sum(O==j)
e[i,j]=K/n
d[i,j]=((o[i,j]-e[i,j]))^2/e[i,j]}
c[i]=sum(d[i,])}
return(c)}
c=f(10000,100,6)
plot(density(c,bw=0.5),xlim=c(0.1,15),ylim=c(0,0.2),xlab=expression(chi^2),col="green",lty=2,lwd=2,main=expression(chi^2~"simulation for k=6 and k=10"))
lines(sort(c),dchisq(sort(c),5),lty=2,lwd=2,col="red")
c=f(10000,100,10)
lines(density(c,bw=0.5),col="black",lty=1,lwd=1)
lines(sort(c),dchisq(sort(c),9),lty=2,lwd=2,col="blue")
legend(10,0.22,c("simulation",expression(chi[(5)]^2),"simulation",expression(chi[(9)]^2)),col=c("green","red","black","blue"),lty=c(2,2,1,2),lwd=c(2,2,1,2))
```

Figure 2 is the simulation of *k* face die for *k=6* and *k=10* with equal probability in each case for *k* categories. The R-code for the simulation is given below. The sampling distribution of the statistic approaches to the theoretical chi-square distribution with *k-1* degrees of freedom.

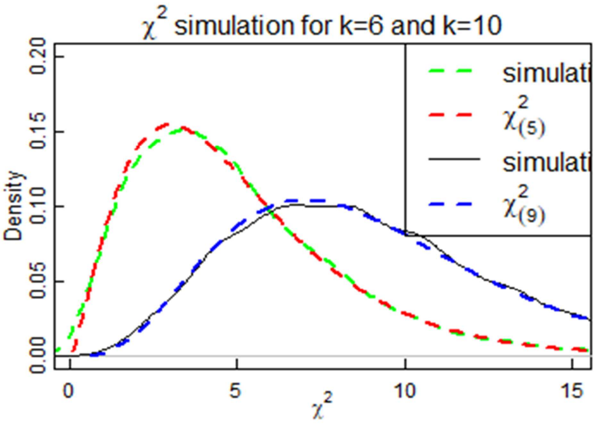

Figure 2. *Chi-square simulation for test of goodness of fit.*

The statistic χ^2 approaches to ${\chi^2}_{(k-1)}$ as the number of simulation is large. We have two chi-squares. The chi-square distribution ${\chi^2}_{(k-1)}$ is theoretically derived mathematical distribution. In contrast the chi-square statistic χ^2 is a discrete statistic based on finite number of possible values. If the number of simulations increase χ^2 statistics approaches to the theoretical chi-square distribution.

2.2. Test of Independency

Chi-square test of independency is used to determine the significant relationship between two or more qualitative variables. Qualitative data is where we collect data on individuals that are categories or names. Then we would count the number of individuals having particular quality. The chi-square test of independency allows the researcher to determine whether variables are independent of each other. The variables which are being examined can be measured at any level, nominal, ordinal, interval or ratio.

The data obtained from the sample is called observed number of cases. There is frequency of occurrences for each category in to which the data have been grouped. In the chi-square test the null hypothesis is makes a statement concerning how many cases are to be expected in each category if the hypothesis is correct.

Chi-square test of independency is a non-parametric statistical test used two or more variables are independent or associated. The distribution of chi-square statistic approaches to the theoretical χ^2 distribution. Chi-square test of homogeneity is used to determine whether frequency counts are identically distributed across different populations or across different groups of same population. This is the link between test of homogeneity and test of independence.

Suppose we have two populations, and certain event happen in the two populations. We want to know whether these populations are independent. In other word we want to test the equality of the events in the two population. We want to test a hypothesis H_0 : the two populations are independent (the events are homogeneous in the two populations) Vs H_1 : the two populations are dependent (the event is not homogeneous). If the null hypothesis is true, then the number of observations in each cell is equal to the expected number of cases. If we have *c* populations each have categories of *r*, then the statistic calculated in equation (8) is chi square distribution.

$$\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}\sim{\chi^2}_{(r-1)(c-1)} \qquad (8)$$

Where *r*=number of rows, *c*=number of columns. The degrees of freedom are the number of terms in simplified terms in equation (8). Since $\sum_{i=1}^{r}O_{ij}=\sum_{i=1}^{r}E_{ij}=n_j$, $\sum_{j=1}^{c}O_{ij}=\sum_{j=1}^{c}E_{ij}=m_i$ and $\sum_{j=1}^{c}n_j=\sum_{j=1}^{r}m_i=n$, where n_j is the number of j^{th} population and m_i is the total number of observations having i^{th} quality, *n* is the total number of all observations.

Therefore the number of degrees of freedom is $(r-1)(c-1)$

To prove the χ^2 statistic calculated in equation (8) follows ${\chi^2}_{(r-1)(c-1)}$ we consider the central limit theorem. Let $X_1\sim binom(n_1,p_1)$ and $X_2\sim binom(n_2,p_2)$ assume S_1 is the number of events (success) and F_1 the number of failures in population *1* and S_2 is the number of events and F_2 is the number of failures in population *2*. If the two populations are independent, then the number of success in each population are equal. i.e. $p_1=p_2$. Hence we expect equal number of success and failure in each population (homogeneously distributed). By substituting in to equation (4)

$$Z=\frac{(\hat{p}_1-\hat{p}_2)}{\sqrt{\left[\frac{1}{n1}+\frac{1}{n2}\right](p(1-p))}}\sim N(0,1) \qquad (9)$$

For large n_1 and n_2 then, substituting $\hat{p}_1=S_1/n_1$, $\hat{p}_2=S_2/n_2$ and $S=S_1+S_2$, $F=F_1+F_2$, $n=n_1+n_2=F+S$ then, after some approximations and some algebraic computation

$$Z^2=\frac{\left(S_1-\frac{n_1}{n_1+n_2}S\right)^2}{\frac{n_1}{n_1+n_2}S}+\frac{\left(S_2-\frac{n_2}{n_1+n_2}S\right)^2}{\frac{n_2}{n_1+n_2}S}+\frac{\left(F_1-\frac{n_1}{n_1+n_2}F\right)^2}{\frac{n_1}{n_1+n_2}F}+\frac{\left(F_2-\frac{n_2}{n_1+n_2}F\right)^2}{\frac{n_2}{n_1+n_2}F} \qquad (10)$$

There are four dependent terms in the equation. The number of terms in simplified equation is one. Therefore, the degree of freedom is one. For c population each having r categories equation (10) can be written as follows.

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)} \qquad (11)$$

Now we are to simulate χ^2 statistic for test of independency (homogeneity). Assume we have four dies of six faces each. We want to test whether the four dies are independent. In other words, the occurrence of each faces homogeneously distributed across four dies. If the dies are independently tossed the expected frequency will be row total times column total divided by grand total. If the four dies are tossed the number of values having characteristics is the same in all dies. The R code given as below.

```
f=function(N,n,c,r){ # test of independency
C=matrix(0,N)
for(i in 1:N){
p=matrix(0,n,c);o=matrix(0,r,c);e=o
for(j in 1:c){
p[,j]=sample(1:r,n,rep=TRUE)
for(k in 1:r){
o[k,j]=sum(p[,j]==k)}}
for(j in 1:c){
for(k in 1:r){
e[k,j]=sum(o[k,])*sum(o[,j])/sum(o)}}
C[i]=sum((o-e)^2/e)}
return(C)}
c=f(10000,40,6,4)
plot(density(c,bw=2),xlim=c(min(c),50),ylim=c(0,0.12),xl
ab=expression(chi^2),ylab=expression(chi^2~density),main=
expression(chi^2~"simulation for c=6 and r=4"))
lines(sort(c),dchisq(sort(c),15),lty=2,lwd=2,col="red")
legend(30,0.1,c("simulation",expression(chi[(15)]^2)),lty=
c(1,2),lwd=c(1,2),col=c("black","red"))
```

The following simulation is based five independent populations with four categories each. The simulation gives chi-square distribution with 12 degrees of freedom. Figure 3. shows the simulation of this phenomena of 10000 simulations compared with the theoretical chi-square distribution. The degrees of freedom is *(4-1) (6-1) =15*. The algorithm can work for any integer $c>1$ and $r>1$.

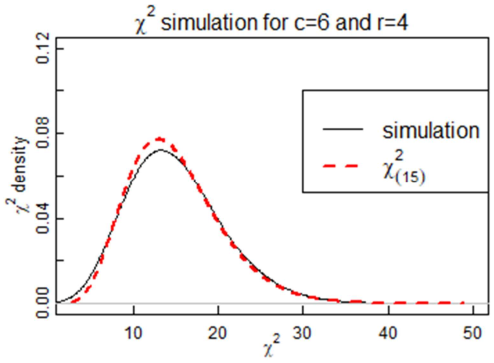

Figure 3. *Chi-square simulation for test of independency (homogeneity).*

For large simulation of c population with r categories each the χ^2 statistic calculated in equation (11) turns to theoretical $\chi^2_{(r-1)(c-1)}$

2.3. Hypothesis Testing

The other case is chi-square distribution arises in hypothesis testing. Let $X_1, X_2, ..., X_n$ be independent normal random variables distributed according to $\sim N(\mu,\sigma^2)$. To test a hypothesis $H_0:\sigma^2=\sigma_0{}^2$ vs $H_1:\sigma^2\neq\sigma_0{}^2$ we use chi-square test. We reject the null hypothesis if $\chi^2=\frac{(n-1)S^2}{\sigma^2}$ is large compared to $\chi^2_{(n-1)}$. If he population mean is known then the distribution of the statistic $\chi^2=\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2}$ follows chi-square distribution with n degrees of freedom. The simulation turns to chi-square distribution shown in "sampling distribution and simulation in R" [6] at α level of significance, we would reject the null hypothesis if the chi-square statistic is large compared to the theoretical one. Sometimes the distribution of the statistic difficult to extract. To test the hypothesis, in such cases, we use some approximations such as likelihood ratio test.

2.4. Likelihood Ratio Test

Now we consider the general case of hypothesis testing. Hypothesis can be tested using the logarithm of the ratio of likelihoods. This is one of the most useful method for complicated models. If the data are presented in group form, and if the alternative hypothesis is completely general, it is known that in large sample the chi-square statistic and the likelihood ratio test becomes equivalent [4].

Let $X_1,X_2,....,X_n$ be iid $f(x/\theta)$ and let $L(\theta/X)=\prod_{i=1} f(x/\theta)$ be the likelihood function

Let $\lambda(X)=\left[\frac{L(\theta)}{L(\hat{\theta})}\right]$ then $-2\log\lambda(X)$ is chi square distribution with the number of free parameters in the null hypothesis

Proof:

$$-2\log\lambda(X)=-2\log\left[\frac{L(\theta)}{L(\hat{\theta})}\right]=-2(\log L(\theta)-\log L(\hat{\theta})) \quad (12)$$

Let $l(\theta)=\log(L(\theta))$ then using Tayler expansion near to $\hat{\theta}$ $l(\theta)=l(\hat{\theta})+(\theta-\hat{\theta})l'(\hat{\theta})+\frac{1}{2!}(\theta-\hat{\theta})^2l''(\hat{\theta})+......$

Where we are going to ignore the higher order terms it becomes $l(\theta)\approx l(\hat{\theta})+(\theta-\hat{\theta})l'(\hat{\theta})+\frac{1}{2!}(\theta-\hat{\theta})^2l''(\hat{\theta})$ and we have $l'(\hat{\theta})=0$

$-2\log\lambda(X) = (\theta-\hat{\theta})^2 l''(\hat{\theta}) = \dfrac{(\theta-\hat{\theta})^2}{-\dfrac{1}{l''(\hat{\theta})}}$ where $l''(\hat{\theta})$ is called information number (Fisher information. Let $h(\hat{\theta})$ is the estimator of $h(\theta)$ We know from theorem of Cramer Rao lower bound the estimator of $h(\theta)$ has variance

$Var(h(\hat{\theta})) = \dfrac{(h'(\theta))^2}{-l''(\theta)}$ in our case $h(\theta)=\theta$ then $Var(\hat{\theta}) = \dfrac{(1)^2}{l''(\theta)}$

$-2\log\lambda(X) = \dfrac{(\theta-\hat{\theta})^2}{Var(\hat{\theta})} = \left[\dfrac{\theta-\hat{\theta}}{\sqrt{Var(\hat{\theta})}}\right]^2$ From central limit theorem $\dfrac{\theta-\hat{\theta}}{\sqrt{Var(\hat{\theta})}}$ is standard normal distribution. Hence the proof is completed. In a hypothesis $H_0:\theta=\theta_0$ Vs $H_1:\theta\neq\theta_0$, H_0 will be rejected if $-2\log\lambda(X)$ is large

To verify the theorem using simulation let us see the log likelihood ratio of some important distributions. Consider a binomial experiment as an example. Binomial experiment is a collection of Bernoulli trials. We are about to test hypothesis $H_0: p = p_0$ versus $H_0: p \neq p_0$ to test this we have likelihood ratio test in equation (13) below

$$-2\log\lambda(X) = 2n\left((\bar{x}-1)\log\left[\frac{1-p_0}{1-\bar{x}}\right] - \bar{x}\log\left[\frac{p}{\bar{x}}\right]\right) \quad (13)$$

The following simulation is based on the assumption the population is come from $p = 0.5$. assume we want to test $H_0: p = 0.5$ versus $H_1: p \neq 0.5$. If the null hypothesis is true, then the log likelihood ratio is chi-square distribution with *1* degrees of freedom. The algorithm given below can simulate this situation.

```
f=function(N,n){ #hypothesis test binomial
c=matrix(0,N)
for(i in 1:N){
x=sample(c(0,1),n,prob=c(0.5,0.5),rep=TRUE)
c[i]=2*n*((mean(x)-1)*log((1-0.5)/(1-mean(x)))-mean(x)*log(0.5/mean(x)))}
return(c)}
c=f(1000,100)
plot(density(c,bw=0.3),xlim=c(min(c)+0.2,6),ylim=c(0,0.6),main=expression(chi^2~"simulation for binomial"),xlab=expression(chi^2))
lines(sort(c),dchisq(sort(c),1),lty=2,lwd=2,col="red")
legend(3,0.5,c("simulation",expression(chi[(1)]^2)),lty=c(1,2),lwd=c(1,2),col=c("black","red"))
```

Figure 4. depicts the simulation of the log likelihood ratio for binomial parameter test.

The log likelihood ratio approximately chi-square distribution with 1 degree of freedom. The density of the simulation near to zero is lower than the theoretical density. This is due to the theoretical density near to zero is infinite. But the empirical density is based on finite number of simulations.

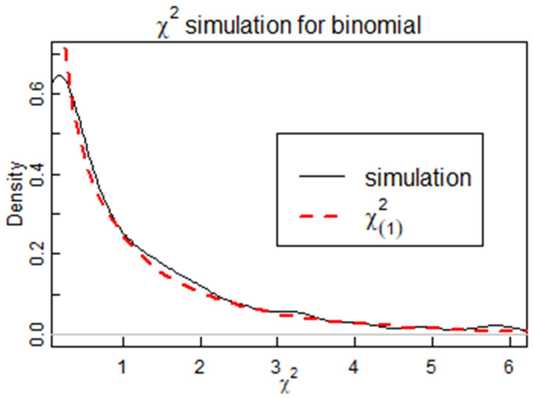

Figure 4. *Log likelihood ratio simulation for binomial parameter test.*

Now let us consider another example, a Poisson experiment. We are to test hypothesis $H_0:\lambda=\lambda_0$ versus $H_1:\lambda\neq\lambda_0$ to test this we have likelihood ratio test.

$$-2\log\lambda(X) = -2n(\bar{x}-\lambda_0) - 2n\bar{x}\log\left(\lambda_0/\bar{x}\right) \quad (14)$$

The following simulation is based on the assumption the population is come from $\lambda = 6$.

```
f=function(N,n){ #hypothesis test poisson
c=matrix(0,N);L0=c; L1=c
for(i in 1:N){
x=rpois(n,6)
L0[i]=exp(-n*6)*6^(sum(x))
L1[i]=exp(-n*mean(x))*mean(x)^(sum(x))
c[i]=-2*(log(L0[i]/L1[i]))}
return(c)}
c=f(1000,50)
plot(density(c,bw=0.3),xlim=c(min(c)+0.3,6),ylim=c(0,0.7),main=expression(chi^2~"simulation for poisson"),xlab=expression(chi^2))
lines(sort(c),dchisq(sort(c),1),lty=2,lwd=2,col="red")
legend(3,0.5,c("simulation",expression(chi[(1)]^2)),lty=c(1,2),lwd=c(1,2),col=c("black","red"))
```

Figure 5. shows the simulation of the likelihood ratio test for Poisson parameter test.

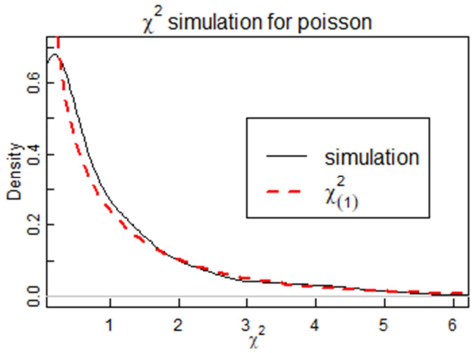

Figure 5. *Log likelihood ratio for Poisson parameter test.*

If the null hypothesis is true, then the log likelihood ratio in equation (14) is chi-square distribution with 1 degrees of freedom. A similar statement can be made as previous the empirical density near to zero is lower than the theoretical density.

Another important statistical test we often found in statistical method is the population mean for normal distribution. That is we are to test hypothesis $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. The following is the log likelihood ratio test for hypothesis testing for the population mean when sigma is known. The log likelihood ratio is:

$$-2\log\lambda(X) = \frac{\sum_{i=1}^{n}(x_i-\mu_0)^2}{\sigma^2} - \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sigma^2} \qquad (15)$$

We know that the first and the second term are chi-square distribution with *n* and *n-1* degrees of freedom respectively. But the two terms are not independent. Equation (15) can be simplified as

$-2\log\lambda(X) = \left(\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right)^2$ which is standard normal distribution square. The R code for simulation is

```
f=function(N,n){ #hypothesis for normal mean test sigma is known
c=matrix(0,N);L0=c; L1=c
for(i in 1:N){
x=rnorm(n,8,3)
c[i]=-sum((x-mean(x))^2)/9+sum((x-8)^2)/9}
return(c)}
c=f(1000,10)
plot(density(c,bw=0.3),xlim=c(min(c)+0.3,6),ylim=c(0,0.6),main=expression(chi^2~"simulation for normal"),xlab=expression(chi^2))
lines(sort(c),dchisq(sort(c),1),lty=2,lwd=2,col="red")
legend(3,0.5,c("simulation",expression(chi[(1)]^2)),lty=c(1,2),lwd=c(1,2),col=c("black","red"))
```

Figure 2. depicts the simulation of the log likelihood ratio for normal distribution mean test when the population variance is known.

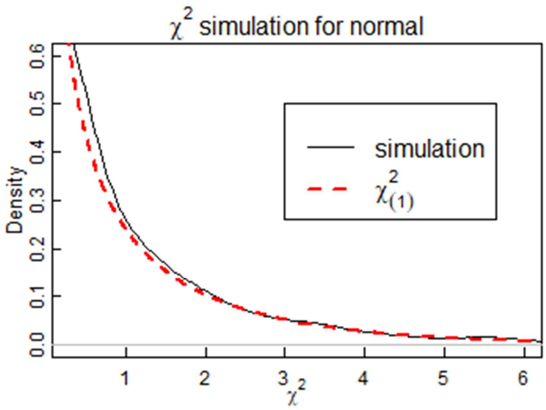

Figure 6. *The simulation of the log likelihood ratio for the population mean when sigma is known.*

When the population variance is known the likelihood ratio for normal distribution mean test is chi-square distribution with one degree of freedom.

In statistical method we often found testing population variance when the population mean is known. We are about to test the hypothesis $H_0 : \sigma^2 = \sigma_0{}^2$ versus $H_1 : \sigma^2 \neq \sigma_0{}^2$. The following is the log likelihood ratio test for hypothesis testing for the population variance when μ is known. The log likelihood ratio is:

$$-2\log\lambda(X) = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma_0{}^2} - n + n\log\left(\frac{\sigma_0{}^2}{\hat{\sigma}^2}\right) \text{ Where}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n} \qquad (16)$$

The algorithm for the simulation is given below.

```
f=function(N,n){ #hypothesis test normal var test mu is known
c=matrix(0,N)
for(i in 1:N){
x=rnorm(n,8,3)
sig=sum((x-8)^2)/n
c[i]=sum((x-8)^2)/9-n+n*log(9/sig)}
return(c)}
c=f(1000,10)
plot(density(c,bw=0.3),xlim=c(min(c)+0.3,6),ylim=c(0,0.6),main=expression(chi^2~"simulation for normal"),xlab=expression(chi^2))
lines(sort(c),dchisq(sort(c),1),lty=2,lwd=2,col="red")
legend(3,0.5,c("simulation",expression(chi[(1)]^2)),lty=c(1,2),lwd=c(1,2),col=c("black","red"))
```

Figure 7. depicts the simulation density for log likelihood ratio test for the population variance when the population mean is known. The simulation density is compared with the theoretical chi-square distribution with one degree of freedom.

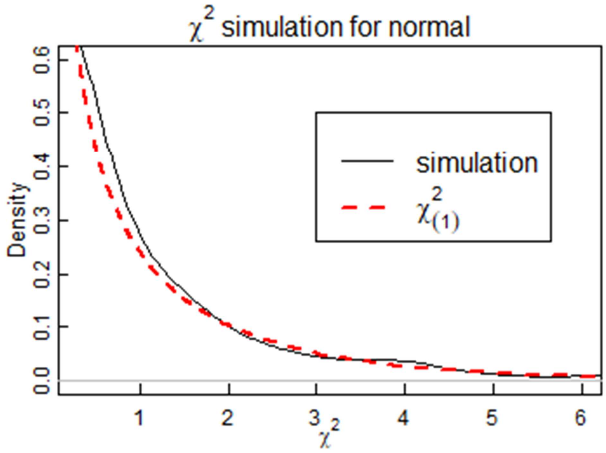

Figure 7. *The log likelihood ratio test simulation population variance when mu is known.*

When the population mean is known the log likelihood ratio for test of variance is chi-square distribution with 1 degree of freedom, only the density of the simulation near to zero is lower than the theoretical density.

The other case hypothesis testing of mean for normal distribution when the population variance is unknown. The hypothesis to be tested is $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ since σ^2 is unknown we substitute it with sample variance S^2. The following is the log likelihood ratio test for hypothesis testing for the population mean when sigma is unknown.

$$-2\log\lambda(X) = 1 - n + \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{S^2} \tag{17}$$

The algorithm for the simulation is given as:

```
f=function(N,n){ #to test mu sigma is unknown
c=matrix(0,N);
for(i in 1:N){
x=rnorm(n,8,3)
c[i]=-n+1+sum((x-8)^2)/var(x)}
return(c)}
c=f(10000,10)
plot(density(c,bw=0.3),xlim=c(min(c)+0.3,6),ylim=c(0,0.6
),main=expression(chi^2~"simulation for
normal"),xlab=expression(chi^2))
lines(sort(c),dchisq(sort(c),1),lty=2,lwd=2,col="red")
legend(3,0.5,c("simulation",expression(chi[(1)]^2)),lty=c(
1,2),lwd=c(1,2),col=c("black","red"))
```

Figure 8 depicts the simulation density of log likelihood ratio in equation (17) compared with the theoretical chi-square distribution. The population variance is unknown and estimated from sample whose divisor is *n-1*.

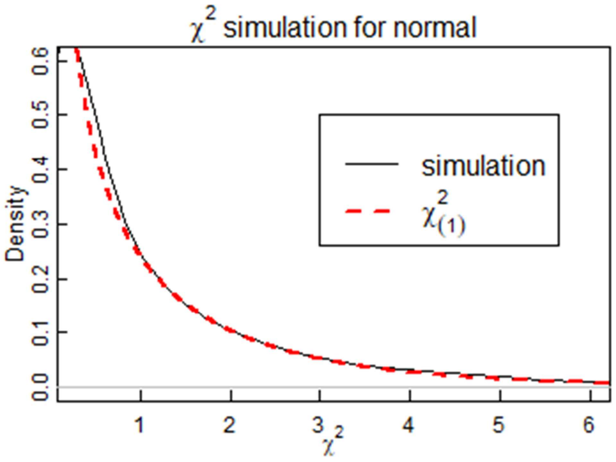

Figure 8. *The log likelihood ratio simulation for mean test.*

The log likelihood ratio for population mean test is chi-square distribution when the population variance is estimated from sample.

The other case we often found in statistical method is testing population variance when the population mean is unknown. Since the true mean is unknown we estimate from sample. The hypothesis to be test is $H_0 : \sigma^2 = {\sigma_0}^2$ versus $H_1 : \sigma^2 \neq {\sigma_0}^2$. The following is the log likelihood ratio test for hypothesis testing for the population variance when mu is unknown. The log likelihood ratio is:

$$-2\log\lambda(X) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{{\sigma_0}^2} - n + n\log\left(\frac{{\sigma_0}^2}{\hat{\sigma}^2}\right), \text{ Where}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n} \tag{18}$$

The R-code is given as below.

```
f=function(N,n){ #mu is unknown to test sigma
c=matrix(0,N);L0=c; L1=c
for(i in 1:N){
x=rnorm(n,8,3)
c[i]=sum((x-mean(x))^2)/9-n+n*log(9/((n-1)*var(x))*n)}
return(c)}
c=f(10000,10)
plot(density(c,bw=0.3),xlim=c(min(c)+0.3,6),ylim=c(0,0.6
),main=expression(chi^2~"simulation for
normal"),xlab=expression(chi^2))
lines(sort(c),dchisq(sort(c),1),lty=2,lwd=2,col="red")
legend(3,0.5,c("simulation",expression(chi[(1)]^2)),lty=c(
1,2),lwd=c(1,2),col=c("black","red"))
```

Figure 9 shows the simulation of log likelihood ratio density compared with the theoretical chi-square distribution. Since the true mean is unknown it is estimated from sample.

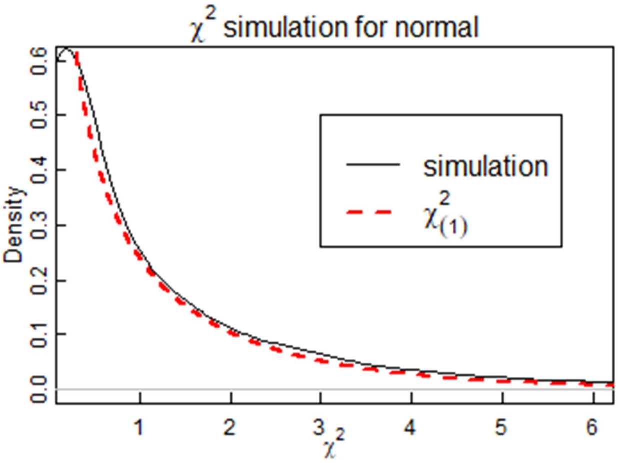

Figure 9. *Simulation for test of population variance when the true mean is unknown.*

When the true mean is unknown we estimate from sample. The log likelihood ratio in equation (18) for population variance test is chi-square distribution with *1* degree of freedom.

The importance of the log likelihood ratio is used even for more complicated models. The test can be used for testing several variables at the same time. If we have k variables the log likelihood ratio is chi-square distribution with k degrees of freedom.

Proof:

$$\log\lambda(X) = -2\log\left[\frac{L(\theta_1,\theta_2,......\theta_k)}{L(\hat{\theta}_1,\hat{\theta}_2,.......\hat{\theta}_k)}\right] = -2(\log L(\theta_1,\theta_2,......\theta_k) - \log L(\hat{\theta}_1,\hat{\theta}_2,.......\hat{\theta}_k))$$

Using tailor expansion $l(\theta_1,\theta_2,....\theta_k) = \log L(\theta_1,\theta_2,......\theta_k)$ near to $\hat{\theta}1.....\hat{\theta}_k$

$$l(\theta_1,\theta_2,....\theta_k) =$$
$$\sum_{i1=0}^{\infty}.....\sum_{ik=0}^{\infty}\frac{(\theta_1-\hat{\theta}_1)^{i_1}}{i_1!}\frac{(\theta_2-\hat{\theta}_2)^{i_2}}{i_2!}.....\frac{(\theta_k-\hat{\theta}_k)^{i_k}}{i_k!}\frac{\partial^{i_1+i_2+...i_k}}{\partial\theta_1^{i_1}\partial\theta_2^{i_2}...\partial\theta_k^{i_k}}l(\hat{\theta}_1.....\hat{\theta}_k)$$

We take the approximation up to second order terms. Higher order terms such as third degree and higher are ignored. And terms involve in the first derivative are zero. And the cross products in the first derivative are zero because the log likelihood function is substituted at its critical values.

$$-2\log\lambda(X) = \frac{(\theta_1-\hat{\theta}_1)^2}{\cfrac{-1}{\cfrac{\partial^2}{\partial\theta_1^2}l(\theta_1,...\theta_k)}} + \frac{(\theta_2-\hat{\theta}_2)^2}{\cfrac{-1}{\cfrac{\partial^2}{\partial\theta_2^2}l(\theta_1,...\theta_k)}} + \frac{(\theta_3-\hat{\theta}_3)^2}{\cfrac{-1}{\cfrac{\partial^2}{\partial\theta_3^2}l(\theta_1,...\theta_k)}}.......\frac{(\theta_k-\hat{\theta}_k)^2}{\cfrac{-1}{\cfrac{\partial^2}{\partial\theta_k^2}l(\theta_1,...\theta_k)}} \tag{19}$$

Each terms are chi square distribution with one degrees of freedom. The degree of freedom is the number of parameters in the model. Hence the proof is completed.

To verify the distribution of the log likelihood ratio is chi-square distribution using simulation, let us consider the log likelihood ratio distribution for two parameters.

Consider a normal distribution. Let $X_1, X_2....X_n$ are normally and independently distributed. Sometimes we are interested in testing the mean and the variance simultaneously. We want to test a hypothesis $H_0: \mu = \mu,\ \ \sigma^2 = \sigma^2$ Vs $H_1: \mu \neq \mu,\ \ \sigma^2 \neq \sigma^2$ simultaneously. The following is the log likelihood ratio test for two parameters.

$$-2\log\lambda(X) = -2\log\left(\frac{L(\mu,\sigma^2)}{L(\hat{\mu},\hat{\sigma}^2)}\right) = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2} - n - n\log\left(\frac{\hat{\sigma}^2}{\sigma^2}\right) \tag{20}$$

The R-code for the simulation is given as:

```
f=function(N,n,mu,sig){ #mu and sigma are unknown simultaneous test
c=matrix(0,N)
for(i in 1:N){
x=rnorm(n,mu,sig)
c[i]=sum((x-mu)^2)/sig^2-n-n*log(sum((x-mean(x))^2)/(n*sig^2))}
return(c)}
c=f(1000,50,8,3)
plot(density(c,bw=0.3),xlim=c(min(c)+0.3,22),ylim=c(0,0.5),xlab=expression(chi^2),ylab=expression(chi^2~density),main=expression(chi^2~"simulation mu and sigma unknown"))
lines(sort(c),dchisq(sort(c),2),lty=2,lwd=2,col="red")
legend(7,0.4,c("simulation",expression(chi[(2)]^2)),lty=c(1,2),lwd=c(1,2),col=c("black","red"))
```

Figure 10 is the density of the simulation of the log likelihood ratio for both of the normal parameters. The density of the simulation in equation (20) compared with chi-square distribution with 2 degrees of freedom.

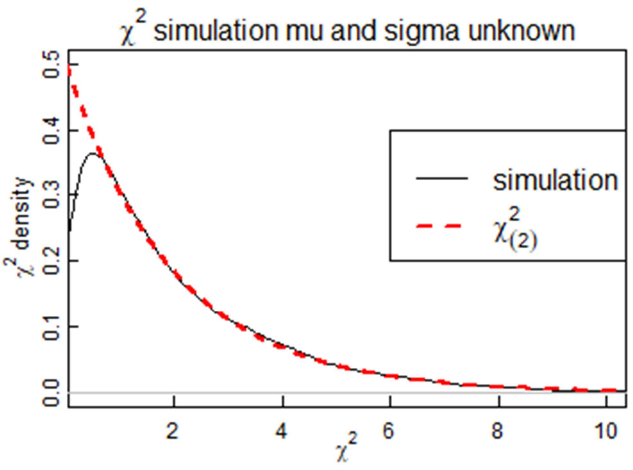

Figure 10. Log likelihood simulation for population mean and variance simultaneous test.

The simulation agrees with the theoretical $\chi^2_{(2)}$. Only the simulation has low density near to zero relative to the theoretical density. This problem cannot be solved increasing the simulation. The density of the simulation near to zero never coincides with the theoretical density. This problem is not unique to simulation. Even drawing samples from $\chi^2_{(1)}$ and $\chi^2_{(2)}$ and plotting the density do not agree to the theoretical distributions near to zero.

2.5. Significance Test

The importance of log likelihood ratio is it used to test significance of the parameter. In other words, it is used for model selection. To verify the theorem of the likelihood ratio distribution let us consider two models, Multiple linear regression and complicated logistic regression. We are about to test the significance of extra variable in the alternative hypothesis.

Consider test of significance (model selection) for linear regression. Suppose Y is the dependent variable and X_1 and X_2 are independent (explanatory) variables. we want to test the significance relation between the response variable and dependent variable. Suppose we are interested in testing only X_1 is significantly related to Y . and X_2 is not important predictor for Y . That is, we want to test a hypothesis.

H_0 : The model is $\hat{Y} = \beta_{00} + \beta_{10} X_1$ Vs the model is $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

In other words we are about to test $H_0 : \beta_0 = \beta_{00}, \beta_1 = \beta_{10}, \beta_2 = 0$ Vs $H_1 : \beta_0 = \beta_0, \beta_1 = \beta_1, \beta_2 \neq 0$

The log likelihood ratio is

$$-2\log\lambda(Y) = \frac{\sum_{i=1}^{n}\left(y_i - \beta_{00} - \beta_{10}X_1\right)^2}{\sigma^2} - \frac{\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 X_1 - \beta_2 X_2\right)^2}{\sigma^2} \tag{21}$$

The following algorithm gives the simulation of this situation.

```
f=function(N,K,n){ #regression
c=matrix(0,N)
X1=runif(K,1,6); X2=rnorm(K,2,1);Y=rnorm(K,8,3)
K=1:K
for(i in 1:N){
k=sample(K,n)
x1=X1[k];x2=X2[k]; y=Y[k]
m0=lm(y~x1);b00=m0$coeff[1];b10=m0$coeff[2]
m=lm(y~x1+x2);b0=m$coeff[1];b1=m$coeff[2];b2=m$coeff[3]
c[i]=-sum((y-b0-b1*x1-b2*x2)^2)/9+sum((y-b00-b10*x1)^2)/9}
return(c)}
c=f(1000,500,50)
plot(density(c,bw=0.3),xlim=c(min(c)+0.3,5),ylim=c(0,0.5),xlab=expression(chi^2),ylab=expression(chi^2~density),main=expression(chi^2~"simulation for significance test"))
lines(sort(c),dchisq(sort(c),1),lty=2,lwd=2,col="red")
legend(3,0.4,c("simulation",expression(chi[(1)]^2)),lty=c(1,2),lwd=c(1,2),col=c("black","red"))
```

The null hypothesis for the simulation is only X_1 is important predictor for Y and the alternative hypothesis is both X_1 and X_2 are important predictors. Figure 11. shows the simulation for log likelihood ratio for regression significance test.

For significance test the log likelihood ratio in equation (21) follows chi-square distribution with *1* degree of freedom. Generally, the degree of freedom is the number of free parameters in the null hypothesis. The ratio of population size to the sample size determine the convergence of the simulation.

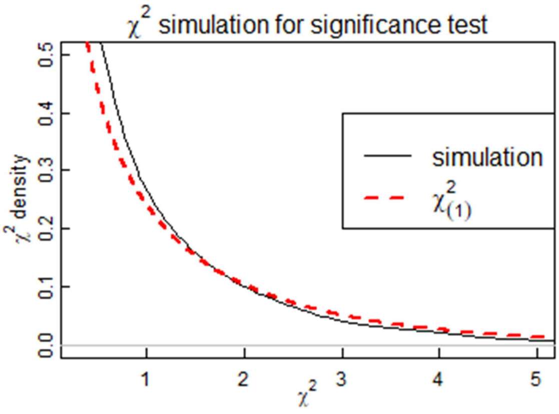

Figure 11. *Chi-square simulation for significance test for linear regression.*

Let us consider logistic regression whose log likelihood ratio distribution is difficult to find. Logistic regression is by far the most widely used tool for relating a binary response to a family of explanatory variables [8]. Let us take complicated model whose distribution is difficult to derive directly. But in likelihood ratio we can find its distribution. In likelihood ratio we can test the significance of the variable related to the other variable. In methods like survival analysis and regressions we need to test whether a variable is significantly related to the response variable. Suppose Y is response variable X_1, X_2, and X_3 are explanatory variables we need to test the significance of the two variables (say X_2 *and* X_3). The hypothesis to be tested is H_0 : only significant variable is X_1 Vs H_1 : all variables X_1, X_2 *and* X_3 are significant. In other word we are about model selection. The same hypothesis is $H_0 : \beta_0 = \beta_{00}, \beta_1 = \beta_{10}, \beta_2 = 0, \beta_3 = 0$ Vs $H_1 : \beta_0 = \beta_0, \beta_1 = \beta_1, \beta_2 \neq 0, \beta_3 \neq 0$

The log likelihood ratio is

$$-2\log\lambda(X) = -2\left[\frac{\sum_{i=1}^{n} y_i \log\left(\frac{e^{\beta_{00}+\beta_{10}X_{1i}}}{1+e^{\beta_{00}+\beta_{10}X_{1i}}}\right)+\sum_{i=1}^{n}(y_i-1)\log\left(1+e^{\beta_{00}+\beta_{10}X_{1i}}\right)}{\sum_{i=1}^{n} y_i \log\left(\frac{e^{\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\beta_3X_{3i}}}{1+e^{\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\beta_3X_{3i}}}\right)+\sum_{i=1}^{n}(y_i-1)\log\left(1+e^{\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\beta_3X_{3i}}\right)}\right] \quad (22)$$

The log likelihood ratio is chi-square distribution with degrees of freedom the number of free parameters in the null hypothesis. If the null hypothesis is true, then the log likelihood ratio follows chi-square distribution with *2* degrees of freedom. The R-code is given as follows

```
f=function(N,K,n){ #logistic regression
c=matrix(0,N);l0=c; l1=c
X1=runif(K,1,3);X2=rbinom(K,3,0.5);X3=rnorm(K,0,1)
Y=sample(c(0,1),K,prob=c(0.8,0.2),rep=TRUE)
for(i in 1:N){
k=sample(1:K,n)
y=Y[k];x1=X1[k];x2=X2[k];x3=X3[k]
m0=glm(y~x1,family=binomial(link="logit"))
b00=m0$coeff[1];b10=m0$coeff[2]
m=glm(y~x1+x2+x3,family=binomial(link="logit"))
b0=m$coeff[1];b1=m$coeff[2];b2=m$coeff[3];b3=m$coeff[4]
l0[i]=sum(y*log((exp(b00+b10*x1)/(1+exp(b00+b10*x1)))))+sum((y-1)*log(1+exp(b00+b10*x1)))
l1[i]=sum(y*log((exp(b0+b1*x1+b2*x2+b3*x3)/(1+exp(b0+b1*x1+b2*x2+b3*x3)))))+sum((y-1)*log(1+exp(b0+b1*x1+b2*x2+b3*x3)))
c[i]=-2*(l0[i]-l1[i])}
return(c)}
c=f(1000,800,100)
plot(density(c,bw=0.3),xlim=c(min(c)+1,15),ylim=c(0,0.31),main="Logistic regression simulation",xlab=expression(chi^2),ylab="density")
lines(sort(c),dchisq(sort(c),2),lty=2,lwd=2,col="red")
legend(8,0.2,c("simulation",expression(chi[(2)]^2)),lty=c(1,2),lwd=c(1,2),col=c("black","red"))
```

Figure 12 shows the simulation of the log likelihood ratio for 1000 simulation for logistic regression significance test.

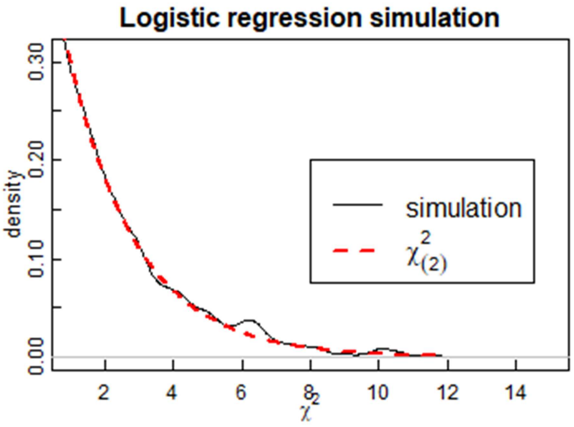

Figure 12. *Chi-square simulation for significance test for logistic regression.*

For sufficiently large simulation the log likelihood ratio turns to chi-square distribution. The convergence depends on not only the number of simulation; the number of population we draw samples determine its convergence. For smaller sample *n<50* the linear model does not converge. For smaller value of the ratio of sample size to population size the log likelihood ratio follows chi-square distribution with *2* degrees of freedom.

2.6. Derived Distribution

t-distribution and F-distribution are most widely used statistic used in estimation and hypothesis testing. Both are derived from chi-square distribution. Next we consider these distributions.

t-distribution

t-distribution is most widely used distribution in statistical method. Let Z is standard normal distribution and X is chi square distribution with k degrees of freedom. Then $t = Z/\sqrt{\frac{X}{k}}$ is t-distribution with k degrees of freedom. It is developed by William Sealy Gosset (1908)

Proof: Let $t = Z/\sqrt{X/k}$ and $v = X/k$ this implies $Z = t\sqrt{v}$ and $X = kv$ then the Jacobean is $|J| = k\sqrt{v}$ then

$$f(t,v) = \frac{1}{\Gamma(k/2)2^{\frac{k}{2}+1}\sqrt{\pi}} k^{\frac{k}{2}} v^{\frac{k-1}{2}} e^{-\left(\frac{k}{2}+\frac{t^2}{2}\right)v}$$ then

$$f(t) = \int_0^{\infty} f(t,v)dv$$

The integration is easy that can be computed using gamma function or gamma distribution.

$$f(t) = \frac{\Gamma((k+1)/2)}{\Gamma(k/2)\sqrt{k\pi}} \frac{1}{\left(1+t^2/k\right)^{\frac{k+1}{2}}} \quad (23)$$

The equation in (23) is called the density of student's t-distribution.

In many occasions we found t-distribution such as the distribution of the sample mean when the population variance is unknown. In testing (or estimating) population mean when the population variance is unknown the sample variance is the denominator. The simulation is given in "sampling distribution and simulation in R" [6] In so many other occasions we found t-distribution such as two population mean equality, correlation coefficient test and so on. One thing draw our attention here is in testing equality of two population mean when the variances are unknown and unequal. In majority of cases in which mean test are required we have no a prior knowledge of the variance of the population [9]. In testing mean difference in two populations when the samples are come from population of different variance and unknown, then chi-square distribution is in the denominator. Our aim here is to show the statistic *t* given next is t- distribution $t=\frac{(\bar{X}-\bar{Y})-\Delta}{\sqrt{Var(\bar{X}-\bar{Y})}}$ where Δ hypothesized mean difference between the two population (groups in a population). If the variances are different and unknown, we estimate from samples. Then the estimated variance is in the denominator.

$t=\frac{(\bar{X}-\bar{Y})-\Delta}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$ to find the density of this statistics is difficult. But we use Satterthwaite approximation. It is easy to change the denominator in to chi-square distribution.

Therefore, we try to change it t-distribution. If we divide both the numerator and denominator by *a* where $a^2=Var(\bar{X}-\bar{Y})=\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}$ the *t* will be

$t=\frac{[(\bar{X}-\bar{Y})-\Delta]/a}{\sqrt{\frac{S_1^2}{a^2n_1}+\frac{S_2^2}{a^2n_2}}}\approx\frac{Z}{\sqrt{\frac{X}{k}}}$ for some value of *k,* the numerator is standard normal distribution. Let us change the denominator in to chi-square distribution

$$\frac{X}{k}=\frac{\sigma_1^2}{(n_1-1)a^2n_1}\left[\frac{(n_1-1)S_1^2}{\sigma_1^2}\right]+\frac{\sigma_2^2}{(n_2-1)a^2n_2}\left[\frac{(n_2-1)S_2^2}{\sigma_2^2}\right]$$

this is two chi-square distributions multiplied by constants

$a_1=\frac{\sigma_1^2}{(n_1-1)a^2n_1}$ and $a_2=\frac{\sigma_2^2}{(n_2-1)a^2n_2}$

To find the distribution of X is quite difficult. If X_1 and X_1 are chi-square distributions with k_1 and k_2 degrees of freedom then the distribution of X_1+X_2 is chi square distribution with k_1+k_2 degrees of freedom. However the distribution of $a_1X_1+a_2X_2$, where a_i's are known constants, is quite difficult. But we might think $a_1X_1+a_2X_2$, it does seem reasonable to assume that $a_1X_1+a_2X_2\approx\frac{\chi^2(k)}{k}$ for some value of *k* will provide good approximation called Satterthwaite. So that $t=\frac{[(\bar{X}-\bar{Y})-\Delta]/a}{\sqrt{\chi^2(k)/k}}\approx t_{(k)}$

Now we assume $a_1X_1+a_2X_2\approx\frac{\chi^2(k)}{k}$. If we take expectation both sides $a_1E(X_1)+a_2E(X_2)=\frac{1}{k}E(\chi^2(k))=1$

This equation cannot solve *k* but it tells us restriction on the first moment. So taking both sides second moment

$$E[(a_1X_1+a_2X_2)^2]=\frac{1}{k^2}E[(\chi^2(k))^2]=\frac{2}{k}+1$$

Solving for *k*, $k=\frac{2}{E[(a_1X_1+a_2X_2)^2]-1}$ (24)

This may give negative degrees of freedom. The degree of freedom is obtained but one can be negative. We might suppose that Satterthwaite was aghast this possibility [2]. He worked much harder as follows.

$$E[(a_1X_1+a_2X_2)^2]=Var((a_1X_1+a_2X_2))+[E[a_1X_1+a_2X_2]]^2$$

$$E[(a_1X_1+a_2X_2)^2]=E[(a_1X_1+a_2X_2)]^2\left[\frac{Var((a_1X_1+a_2X_2))}{E[(a_1X_1+a_2X_2)]^2}+1\right]$$

So he uses the first moment restriction $E[a_1X_1+a_2X_2]=1$ and write the variance

$$E[(a_1X_1+a_2X_2)^2]=\left[\frac{Var((a_1X_1+a_2X_2))}{E[(a_1X_1+a_2X_2)]^2}+1\right]$$

Now substitute this in *k*

$$k=\frac{2}{\frac{Var((a_1X_1+a_2X_2))}{E[(a_1X_1+a_2X_2)^2]}}=\frac{2[E[(a_1X_1+a_2X_2)\ ]]^2}{Var((a_1X_1+a_2X_2))} \tag{25}$$

now the degrees of freedom is found

$$t=\frac{(\overline{X}-\overline{Y})-\Delta}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}=\frac{[(\overline{X}-\overline{Y})-\Delta]/a}{\sqrt{a_1X_1+a_2X_2}}\quad \text{where } a_1=\frac{\sigma_1^2}{(n_1-1)a^2n_1},a_2=\frac{\sigma_2^2}{(n_2-1)a^2n_2},a=\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}$$

Therefor the degree of freedom can be found using equation (25). Since X_1 is chi-square with n_1-1 degrees of freedom and X_2 is chi-square with n_2-1 degrees of freedom and X_1 and X_2 are independent substituting in equation (25)

$$k=\frac{2[[(a_1(n1-1)+a_2(n2-1))\ \ ]]^2}{((a_1^2(2(n_1-1))+a_2^2(2(n_2-1)))}=\frac{\left[\frac{\sigma_1^2}{a^2n_1}+\frac{\sigma_2^2}{a^2n_2}\right]^2}{\left[\frac{\sigma_1^2}{(n_1-1)a^2n_1}\right]^2(n_1-1)+\left[\frac{\sigma_1^2}{(n_2-1)a^2n_2}\right]^2(n_2-1)}$$

$$k=\frac{\frac{1}{a^4}\left[\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right]^2}{\frac{1}{a^4}\left[\frac{\sigma_1^2}{(n_1-1)n_1}\right]^2(n_1-1)+\frac{1}{a^4}\left[\frac{\sigma_1^2}{(n_2-1)n_2}\right]^2(n_2-1)}=\frac{1}{\frac{c_1^2}{n_1-1}+\frac{c_2^2}{n_2-1}} \tag{26}$$

Where $c_i=\dfrac{\sigma_i^2/n_i}{\sigma_1^2/n_1+\sigma_2^2/n_2}$ for $i=1,2$

This completes the proof. The simulation is given in article “Sampling distribution and simulation in R” in [6].

F –distribution

The ratio of two chi-square distributions divided by their corresponding degrees of freedom is F distribution. Let *X* and *Y* be two chi–square random variables with *m* and *n* degrees of freedom, then the probability density function of $u=\dfrac{X/m}{Y/n}$ derived below in equation (27) is called F-distribution. It is named after Ronald Fisher and George W. Snedecor. If we let $v=Y/n$

$f(u,v)=\dfrac{1}{\Gamma(m/2)\Gamma(n/2)2^{\frac{m+n}{2}}}m^{\frac{m}{2}}u^{\frac{m}{2}-1}n^{\frac{n}{2}}v^{\frac{m}{2}+\frac{n}{2}-1}e^{-(1/2(mu+n))v}$ then $f(u)=\int_0^\infty f(u,v)dv$ this integration can be computed easily if we use gamma function or gamma distribution.

$$f(u)=\frac{\Gamma((m+n)/2)}{\Gamma(m/2)\Gamma(n/2)2^{\frac{m+n}{2}}}m^{\frac{m}{2}}n^{\frac{n}{2}}u^{\frac{m}{2}-1}\left(\frac{2}{mu+n}\right)^{\frac{m}{2}+\frac{n}{2}}$$

Simplifying this it will give

$$f(u)=\frac{\Gamma((m+n)/2)}{\Gamma(m/2)\Gamma(n/2)}\left(\frac{m}{n}\right)^{\frac{m}{2}}\frac{u^{\frac{m-2}{2}}}{(1+m/n\times u)^{\frac{m+n}{2}}} \tag{27}$$

F-distribution arises in so many statistical tests. The simulation is given in in article “Sampling distribution and simulation in R” [6].

3. Discussion

A solid understanding of inferential statistics is of major importance of for designing and interpreting results in any scientific discipline. However, students are prone to misconceptions regarding this topic [12]. The use of computer simulation in the teaching of statistical method can help undergraduate students under difficult or abstract statistics concepts [10]. One way to use simulation is to allow students to experiment with a simulation and to discover the important principles on their own [11].

4. Conclusion

We have developed the distributional characteristics of chi-square statistic proof and simulation. It is tried to verify the probabilistic behavior of the resulting empirical chi-square distribution. In this paper we try to verify the sampling distribution of chi-square statistic that we often found in statistical methods using simulation. The aim is to aid students and researchers by giving a brief description of simulation, sampling distribution and its derivation. We try to derive the distribution of empirical chi-square statistic and its derivatives such as student's t-distribution and F-distribution. The derivation and simulation of chi-square statistic we found in many occasions such as test of goodness of fit, test of homogeneity, test of independency, model selection and so on. Chi-square statistic for contingency table follows chi square distribution with *(r-1)(c-1)* degrees of freedom. The log likelihood ratio for *k* variables follows chi-square distribution with *k* degrees of freedom. For model selection the log likelihood ratio follows chi-square distribution with degrees of freedom the number of free parameters in the null hypothesis.

References

[1] Christian Walck, 1998: Hand Book On Statistical Distribution For Experimentalists.

[2] Cassela George et al, 2001: Statistical Inference (2nd edition) Duxbury. ISBN 0-534-24312-6; pp. 102.

[3] Sorana D. Bolboala, et al, 2011: Pearson-Fisher Chi-square statistic Revisited; information journal, 528-545.

[4] William G. Cochran, 1952, the chi-square test of goodness of fit, The Anals Of Mathematical Statistics, vol. 23 no. 3 315-345.

[5] Teshome Hailemeskel Abebe, 2019: The derivation and choice of Approprate test Statistic (Z, T, F and Chi-square tests) in Research Methodology Mathematics letters Vol. 5 no. 3 2019, pp. 33-40. doi: 10.11648/j.ml.20190503.11.

[6] Reshid TM, 2020: Sampling distribution and simulation in R: International journal of statistics And Mathematics, 7 (2): 154-163.

[7] Simon J. A. Malham 2008: chi-square simulation of the ICR process and the Heston Model.

[8] Pragyasur et al 2017: the Likelihood ratio test in High Dimensional logistic regression in asymptotically a rescaled chi-square, Maths. ST. Arxiv.

[9] R. A Fisher, M. A. 1925: Application of "Students' distribution, Rothamsted experimental station, Metron, 90-104.

[10] Xuemao Zhang, et al, 2019: Using r as a simulation tool in teaching Introductory statistics, international electronic journal of mathematics education.

[11] David M. Lane (2015): Simulations of the Sampling Distribution of the Mean Do Not Necessarily Mislead and can facilitate learning, journal of statistics education, Volume 23, Number 2.

[12] Ana Elisa Castro Sotos et al, 2007: Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education, Elsevier, educational research review 98–113.

[13] Moore, D. S. (1997): New Pedagogy and New Content: The case of Statistics. International Statistical Review 65, 123–65.

[14] "Student" (William Seally Gosset, 1908): The probable error of a mean, Biometrika 6 (1): 1–25. doi:10.1093/biomet/6.1.1.