ST 502 Final

Katelyn Settlemyre, Julia Farrell

2022-12-04

Introduction

This report will explore the Chi-Square test for homogeneity in detail. We will derive the likelihood ratio test (LRT) statistic used to conduct this test and explain the Pearson Chi-Square statistic that can be used as an approximation. Once the theory is well established, we then conduct a simulation. The goal is to determine how well the asymptotic rejection region performs at controlling the alpha level of the Pearson Chi-Square test and to determine the power of the asymptotic test when comparing certain alternative hypotheses.

The Chi-Square test for homogeneity is used in a specific case: comparing J multinomials with I classes (with $I, J \in \mathbb{N}$), where the researcher is interested in determining if the probabilities of each cell are the same across every multinomial.

Data Example

Use hospital data given to conduct a χ^2 test for homogeneity.

```
#create and print matrix of hospital data
rows <- rbind(a=c(41, 27, 51), b=c(36,3,40), c=c(169, 106, 109))
hospDat <- matrix(data=rows, nrow = 3, ncol = 3,
                   dimnames = list(c("A", "B", "C"),
                                   c("Surgical Site Infections",
                                     "Pneumonia Infections",
                                     "Bloodstream Infections")))
summary(hospDat)
```

```
##  Surgical Site Infections Pneumonia Infections Bloodstream Infections
##  Min.   : 36.0            Min.   :  3.00       Min.   : 40.00
##  1st Qu.: 38.5            1st Qu.: 15.00       1st Qu.: 45.50
##  Median : 41.0            Median : 27.00       Median : 51.00
##  Mean   : 82.0            Mean   : 45.33       Mean   : 66.67
##  3rd Qu.:105.0            3rd Qu.: 66.50       3rd Qu.: 80.00
##  Max.   :169.0            Max.   :106.00       Max.   :109.00
```

Now we will conduct a Chi-Square test for homogeneity using this sample data.
H_0: The distribution of infections is the same for each hospital
H_1: The distribution of infections is **not** the same for each hospital.

```
x = chisq.test(hospDat)
x
```

```
##
##  Pearson's Chi-squared test
##
## data:  hospDat
## X-squared = 30.696, df = 4, p-value = 3.531e-06
```

```
x$expected
```

```
##   Surgical Site Infections Pneumonia Infections Bloodstream Infections
## A                 50.29897             27.80756               40.89347
## B                 33.39175             18.46048               27.14777
## C                162.30928             89.73196              131.95876
```

The p-value of our chi-square test statistic is extremely small, indicating that we reject the null hypothesis. The data in this example provides support for the alternative hypothesis that the multinomials from these hospitals are not homogeneous.

Deriving the Likelihood Ratio Test

The goal is to derive the likelihood ratio test for a generalized case comparing J independent multinomial distributions, each with I categories.

Let there be J independent multinomial distributions, each with I categories, where $I, J \in \mathbb{N}$. We want to test the hypothesis $H_0 = \pi_{11} = \pi_{12} = ... = \pi_{1J}, \pi_{21} = \pi_{22} = ... = \pi_{1J}, \pi_{I1} = \pi_{I2} = ... = \pi_{IJ}$ vs. H_a : at least one probability differs

To derive the likelihood ratio test, we will initially look at the likelihood, which is just the product of the J multinomials.

Expected Counts Under H_0: $L(\pi'_{ij}s) = \prod_{j=1}^{J} \binom{0}{0} \cdot \pi_{ij}^{n_{ij}} \cdot \pi_{2j}^{n_{2j}} \cdot ... \cdot \pi_{IJ}^{n_{IJ}} \propto \prod_{j=1}^{J} \prod_{i=1}^{I} \pi_{ij}^{n_{ij}}$ subject to constraints ***

Under the null hypothesis,$\pi_{11} = \pi_{12} = ... = \pi_{1J}$, so we will replace this with the common value π_1. Similarly, we will continue forward considering the common values $\pi_1, ..., \pi_I$. There is one restriction on these probabilities to make them valid, which is $\pi_1 + \pi_2 + ... + \pi_I = 1$

Degrees of freedom for reference distribution:
$dim(\Omega) = J \cdot (I-1)$
$dim(\omega) = (I-1)$
$df = J(I-1) - (I-1) = (I-1)(J-1)$

.
. More derivation steps
.
.
The likelihood ratio test for the homogeneity hypothesis being tested is therefore given by $LRT = 2\sum_{j=1}^{J}\sum_{i=1}^{I} Obs_{ij} \cdot ln(\frac{Obs_{ij}}{Exp_{ij}})$

Simulation

Simulate test with data. Potentially make use of code already given in notes. Two multinomial case only, with 3 categories in each multinomial. Plot summaries of simulations. Set seed for reproduction purposes.

Goal:

Process:

```r
#set seed for reproduction
set.seed(17)

#sample sizes
n1 <- c(20,30,50,100)
n2 <- n1
nCombos <- expand.grid(n1=n1, n2=n2)

#probabilities
p1 <- c(1/3, 1/3, 1/3) #equal
p2 <- c(0.1, 0.3, 0.6) #mixed 1
p3 <- c(0.1, 0.1, 0.8) #mixed 2

#Use M=50000 random tables (start lower, end up here)
M <- 500

#Add 0.5 to any expected counts that end up being 0 so as to avoid the divide by 0 case

#use rmultinom(1, size, prob) to generate 1 multinomial sample


n<-c(4,5,10,15,20,25,30,40,50,60,75)

#significance levels being tested
alpha<-c(0.1,0.05,0.01)

#generate many simulated data sets from a normal distribution (this
#is the assuming H0 true part), see observed distribution of H0, use this
#to make cutoffs
#number of data sets
B<-100
#B<-10000

#matrix to keep all correlation values in
corvals<-matrix(nrow=length(n),ncol=B)
for (i in 1:length(n)){
for (j in 1:B){
data<-sort(rnorm(n[i]))
corvals[i,j]<-cor(data,qnorm((1:n[i]-0.5)/n[i]))
}
}
#now each row contains sampled values from the distr. of r under H0, use
#these to find sample quantiles used for comparison
apply(FUN=quantile,X=corvals,MARGIN=1,alpha)
```

```
##          [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]
## 10% 0.8868666 0.9050370 0.9360873 0.9485507 0.9580354 0.9668500 0.9679189
## 5%  0.8468157 0.8858370 0.9218150 0.9408830 0.9523740 0.9631108 0.9658574
## 1%  0.8229198 0.8711964 0.8842723 0.9375656 0.9339893 0.9544932 0.9510339
##          [,8]      [,9]     [,10]     [,11]
## 10% 0.9767617 0.9771458 0.9818093 0.9887854
## 5%  0.9748631 0.9743809 0.9798538 0.9876859
## 1%  0.9694850 0.9640915 0.9728193 0.9849244
```

```r
#now transpose to get table in book
table<-t(apply(FUN=quantile,X=corvals,MARGIN=1,alpha))
rownames(table)<-n
table
```

```
##          10%        5%        1%
## 4  0.8868666 0.8468157 0.8229198
## 5  0.9050370 0.8858370 0.8711964
## 10 0.9360873 0.9218150 0.8842723
## 15 0.9485507 0.9408830 0.9375656
## 20 0.9580354 0.9523740 0.9339893
## 25 0.9668500 0.9631108 0.9544932
## 30 0.9679189 0.9658574 0.9510339
## 40 0.9767617 0.9748631 0.9694850
## 50 0.9771458 0.9743809 0.9640915
## 60 0.9818093 0.9798538 0.9728193
## 75 0.9887854 0.9876859 0.9849244
```

```r
#######################################################
##Now investigate power of test under uniform distribution
#Here data truly comes from uniform, use alpha=0.1, n=10
n<-10
# # of data sets
B<-10000
#vector to store cor results in
corvalues<-rep(0,B)
for (i in 1:B){
data<-sort(runif(n))
corvalues[i]<-cor(data,qnorm((1:n-0.5)/n))
}
#cut off from table, so see how often our observed value
#is less than this as that would imply we reject
sum(corvalues<0.9346399)/B
```

```
## [1] 0.1242
```

```r
###############################33
#same but use n=20
n<-20
# # of data sets
B<-10000
#vector to store cor results in
corvalues<-rep(0,B)
for (i in 1:B){
data<-sort(runif(n))
corvalues[i]<-cor(data,qnorm((1:n-0.5)/n))
}
#cut off is 0.9606222 from table, so see how often our observed value
#is less than this as that would imply we reject
sum(corvalues<0.9599344)/B
```

```
## [1] 0.2337
```

```
############Poor power for this alternative with small samples :(
```