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

Goal: • Determine how well the asymptotic rejection region performs at controlling α • Determine the power of the asymptotic test when comparing certain alternative situations

Process:

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

#sample sizes
n1<-c(20,30,50,100)

# number of classes & multinomials being compared
I = 3
J = 2
```

```r
# three cases being tested
p1 = c(1/3, 1/3, 1/3)
p2 = c(1/10, 3/10, 6/10)
p3 = c(1/10, 1/10, 8/10)

# function to create two multinomials,using sample size and probabilities input
### sampleSize: (int)
### probs: (list of doubles) length of p split determines number of classes. All = 3 for this simulatio
simulateMultinomial <- function(sampleSize, probs) {
  multiGen1 <- rmultinom(1, sampleSize, probs);
  multiGen2 <- rmultinom(1, sampleSize, probs);
  multiGen <- cbind(multiGen1, multiGen2)
  multiGenT = t(multiGen)
  # Chi Square calculated by R
  p <- chisq.test(multiGenT)
  x <- p$statistic
  xapprox <- qchisq(p=0.05, df=(I-1)*(J-1))
  return(c(x, xapprox))
}
simulateMultinomial(n1[1], p1)
```

```
## X-squared
## 0.9645933 0.1025866
```

```r
p1XVals <- list()
for (i in 1:4) {
  append(p1XVals,simulateMultinomial(n1[i], p1))
}
print(p1XVals)
```

```
## list()
```