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"))))
```

```
##   Surgical Site Infections Pneumonia Infections Bloodstream Infections
## A                       41                   27                     51
## B                       36                    3                     40
## C                      169                  106                    109
```

```
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 of this section 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 determine if there is a difference between the probability's associated with each multinomial being tested, we look at the difference between the expected and observed counts $Obs_{ij} - Exp_{ij}$, where the expected counts assume the null hypothesis to be true. We need to look at the likelihood function, which is given by the product of the J multinomials.

The likelihood function is given by $L(\pi'_{ij}s) = \prod_{j=1}^{J} \binom{J}{n} \cdot \pi_{ij}^{n_{ij}} \cdot \pi_{2j}^{n_{2j}} \cdot ... \cdot \pi_{IJ}^{n_{IJ}} \propto \prod_{j=1}^{J} \prod_{i=1}^{I} \pi_{ij}^{n_{ij}}$. This likelihood function is subject to the constraint $\sum_{i=1}^{I} \pi_i = 1, \forall i \in I \forall j \in J$.

Note that the constraint dictates that the degrees of freedom of the whole space is given by $dim(\Omega) = J \cdot (I-1)$ with the dimension of each multinomial given by $dim(\omega) = (I-1)$. The expected counts under H_0 are $E(ij) = \frac{n_{\cdot j} n_{i\cdot}}{n}$.

Under the null hypothesis,$\pi_{11} = \pi_{12} = ... = \pi_{1J}$, so we will replace these with a 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 $\sum \pi_i = 1$

The degrees of freedom for this reference distribution are therefore:$df = J(I-1) - (I-1) = (I-1)(J-1)$.

Using the Lagrange Multiplier technique, we can find the maximized value for the likelihood ratio.

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