STAT 513 grad project

Karl Gregory

11/13/2019

Due: Thursday, December 5th, at class time. Do this project in R markdown and submit a pdf file.

Comparing tests of association based on contingency tables

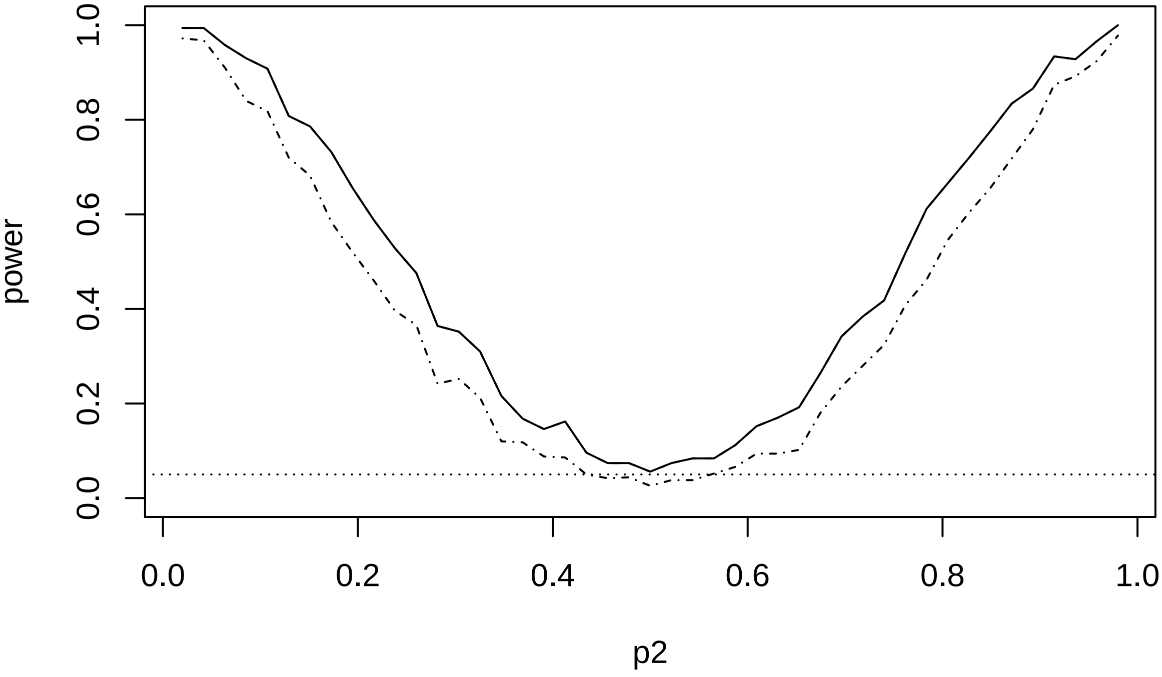
We have learned about three tests of association: The likelihood ratio test, Pearson's chi-squared test, and Fisher's exact test, where Fisher's exact test is only for 2×2 tables. The purpose of this project is to make some comparisons of these three tests based on some simulation studies.

Part 1: Power simulation on 2 x 2 tables

The following R code runs a simulation to find the power curves of the three tests.

TASK: (i) Explain in a few paragraphs what the code is doing. Your paragraphs should be written such that someone who reads them could reproduce the simulation. (ii) Interpret the output. Explain how the tests compare to each other based on this simulation.

```r
p1 <- .5
p2 <- seq(0.02,.98,length=45)
n <- 50
n1 <- n2 <- n/2
S <- 500
alpha <- 0.05

crit <- qchisq(1-alpha,1)
powerG <- powerChisq <- powerFisher <- numeric()
for(i in 1:length(p2))
{

    rejG <- rejChisq <- rejFisher <- numeric()

    s <- 1
    while(s <= S)
    {

        O <- matrix(0,2,2)

        O[1,1] <- rbinom(1,n1,p1)
        O[2,1] <- rbinom(1,n2,p2[i])

        O[1,2] <- n1 - O[1,1]
        O[2,2] <- n2 - O[2,1]
        if( sum(O == 0) != 0  ) next; # throw away tables with counts equal to zero

        E <- apply(O,1,sum) %*% t(apply(O,2,sum))/sum(O)
        G <- 2*sum(O*log(O/E))
        Chisq <- sum((O - E)^2/E)

        rejG[s] <- G > crit
        rejChisq[s] <- Chisq > crit

        # do Fisher's exact test
        Xobs <- O[1,1]
        R1 <- O[1,1] + O[1,2]
        C1 <- O[1,1] + O[2,1]
        N <- sum(O)

        obs.hyper.prob <- dhyper(Xobs, n = N - R1, m = R1, k = C1)
        all.hyper.probs <- dhyper(max(0,C1-(N-R1)):min(C1,R1),n = N - R1, m = R1, k = C1)
        rejFisher[s] <- sum(all.hyper.probs[all.hyper.probs<=obs.hyper.prob]) < alpha

        s <- s + 1

    }

    powerG[i] <- mean(rejG)
    powerChisq[i] <- mean(rejChisq)
    powerFisher[i] <- mean(rejFisher)

}


plot(powerG~p2,ylim=c(0,1),xlab="p2",ylab = "power",type="l")
lines(powerChisq~p2,lty=2)
lines(powerFisher~p2,lty=4)
abline(h=0.05,lty=3)
```



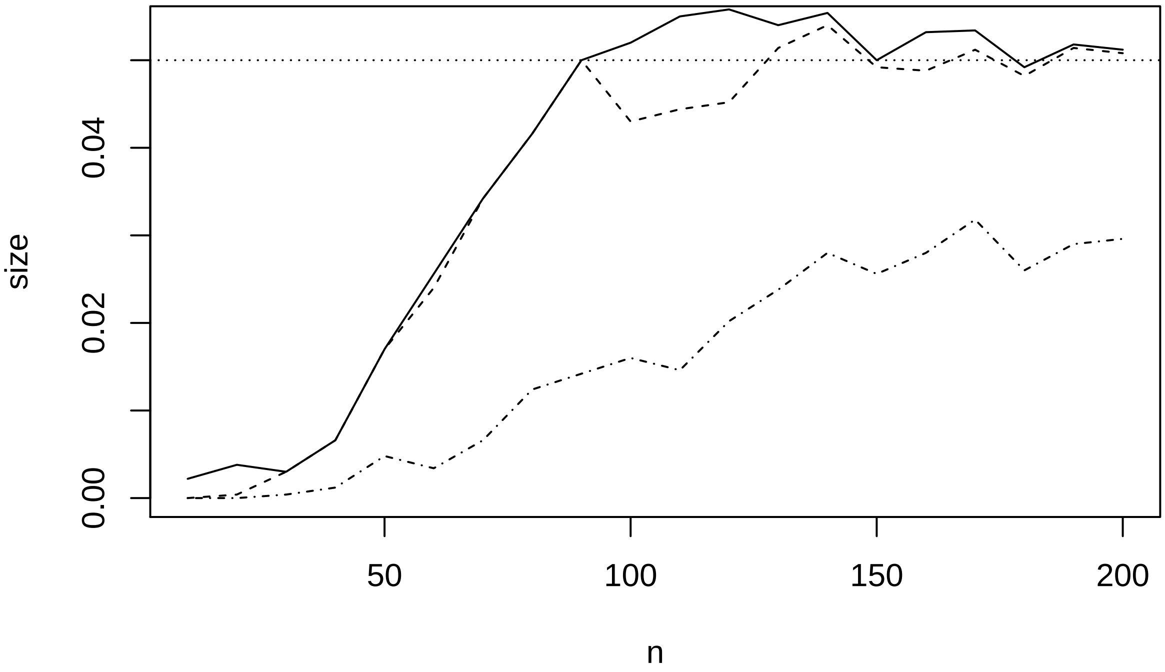
Part 2: Size simulation on 2 x 2 tables

Now write your own code to do a simulation study of the size of the three tests. Set $p_1 = p_2 = 0.10$ and let $Y_1 \sim \text{Binomial}(n_1, p_1)$ and $Y_2 \sim \text{Binomial}(n_2, p_2)$, where $n_1 = n_2 = n/2$ for $n = 10, 20, 30, \ldots, 200$. At each value of n create 5000 2×2 tables in the form

$$\begin{matrix} Y_1 & n_1 - Y_1 \\ Y_2 & n_2 - Y_2 \end{matrix}$$

If a table has entries equal to zero, skip it, i.e. generate a new one, as was done in the power simulation. Set the desired size to $\alpha = 0.05$. For small values of n, we do not expect the tests to achieve the desired size, but as n gets larger, we expect the sizes of the tests to approach α.

TASK: Give your code and produce from your simulation a plot like the one below, which plots the size of the three tests against $n = 10, 20, 30, \ldots, 200$. Your plot should look similar. Include in the plot a legend which indicates to which test each line corresponds.



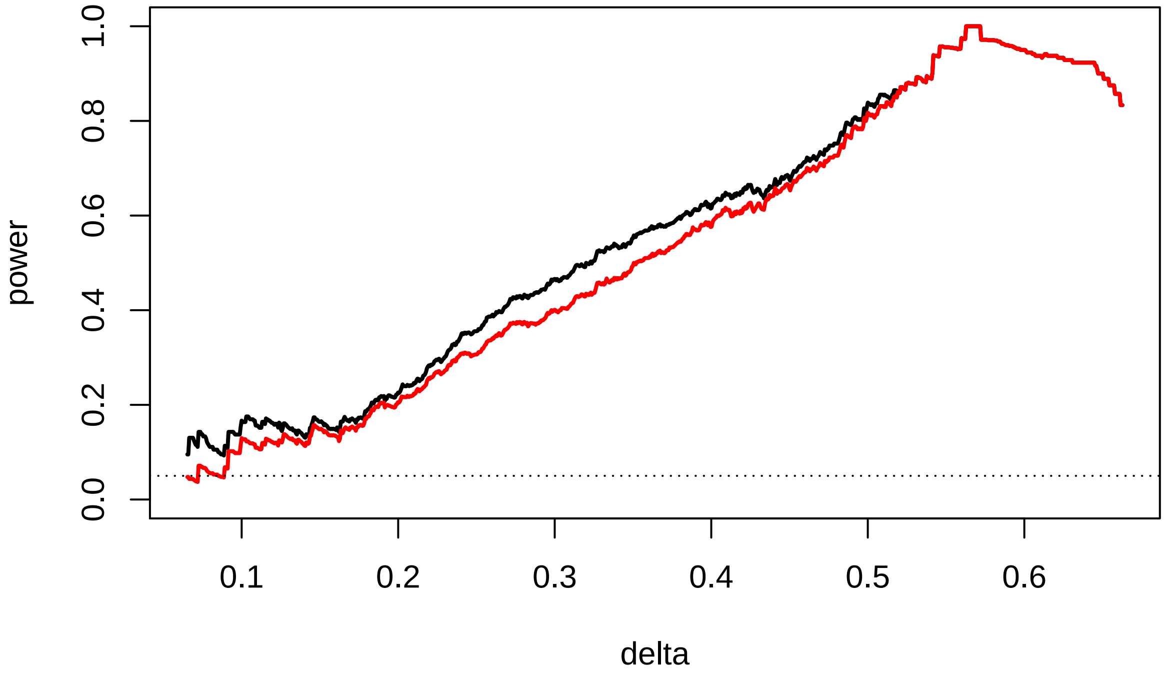
Part 3: Power simulation on K by M tables

The following R code runs a power simulation which proceeds as follows:

1. Generate cell probabilities $p_{kj}, k = 1, \ldots, K, j = 1, \ldots, M$; in addition, compute δ, which is a measure of how untrue the null hypothesis is.
2. Generate a $K \times M$ table of counts based on the cell probabilities.
3. Perform the likelihood ratio test and Pearson's chi-squared test of the null hypothesis of no association.

At the end of the simulation, the probability of rejection by the two tests is regressed in a nonparametric way upon the values of δ to create power curves.

```
max.pairwise.diff <- function(x)
{

    return( max( outer(x , x ,"-") ) )

}

K <- 3
M <- 3
nk <- 20
S <- 1000
alpha <- 0.05

G <- Chisq <- numeric()
rejG <- rejChisq <- numeric()
delta <- numeric()
crit <- qchisq(1-alpha,(K-1)*(M-1))
s <- 1
while(s <= S)
{
    nvec <- rep(nk,K)
    zmat <- matrix(runif(K*M,.1,.9),K,M)
    pmat <- zmat / apply(zmat,1,sum)
    delta[s]  <- max( apply(pmat,2,max.pairwise.diff) )

    O <- matrix(NA,K,M)
    for(k in 1:K)
    {

        O[k,] <- rmultinom(1,nvec[k],prob=pmat[k,])

    }

    if(sum(O==0)!=0) next;

    E <- apply(O,1,sum) %*% t(apply(O,2,sum))/sum(O)
    G[s] <- 2*sum(O*log(O/E))
    Chisq[s] <- sum((O - E)^2/E)

    rejG[s] <- G[s] > crit
    rejChisq[s] <- Chisq[s] > crit

    s <- s + 1

}

plot(NA,ylim=c(0,1),xlim=range(delta),xlab="delta",ylab="power")
lines(ksmooth(y=rejG,x=delta,bandwidth=.1),lwd=2)
lines(ksmooth(y=rejChisq,x=delta,bandwidth=.1),col="red",lwd=2)
abline(h=0.05,lty=3)
```



TASK: Interpret the results. Which of the two tests appears to achieve greater power?

Part 4: Data analysis

TASK: Find your own data set and perform the likelihood ratio test, Pearson's chi-squared test, and, if your data are a 2×2 table, Fisher's exact test of the null hypothesis of no association. Explain your data set and report your results, supplying R code.