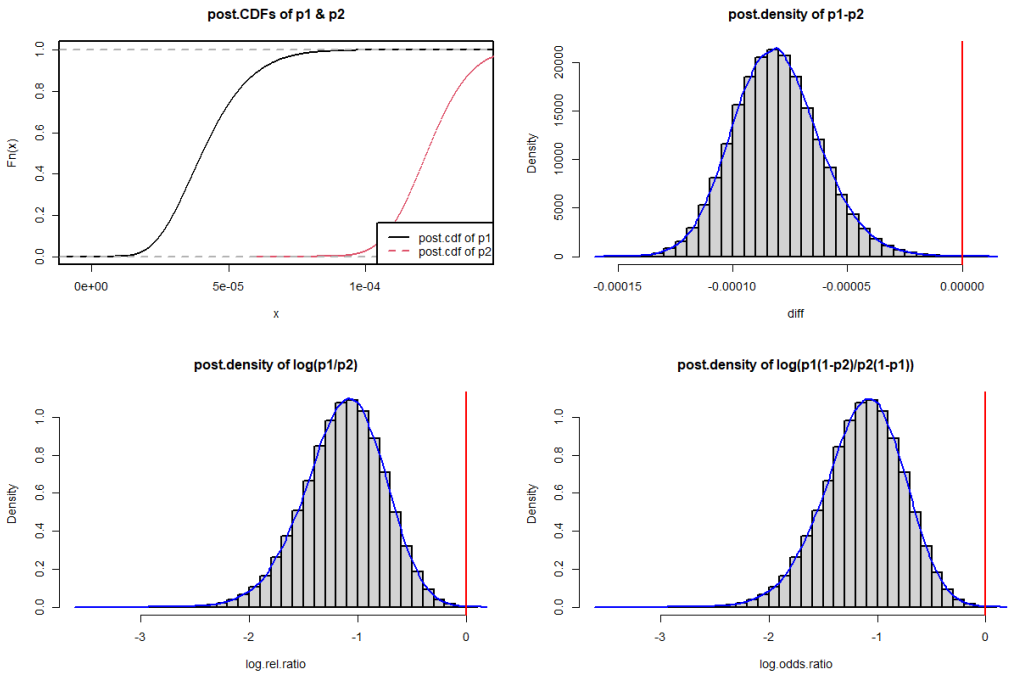
Background: I was asked to perform statistical analysis on data provided to me on behalf of PlayerPulse, to assess whether the proportion of ACL injuries in two populations, consisting of high school female athletes are statistically significantly different.

The data, to my understanding was captured from the following two groups.

- The first group, represents players (U15 – U19) from [redacted], who utilized the 'SoccerPulse' app as part of their training regimen and collected total exposures along with ACL injuries across 3 coaches from this club. The data collected was 8 ACL injuries across 203112 exposures. This equates to 8/203112 or a rate of 3.94 per 100,000 exposures. I designated this as group 1.
- The second group represents a study published at the NIH website which can be found on the website https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3867093/ . Data extracted from Table 1 for Girls' HS Soccer showed 96 ACL injuries across 786293 exposures. This equates to 96/786293 or .000122 ACL injuries per exposure, or a rate of 12.2 per 100,000 exposures. I designated this as group 2.

Analysis – Due to the fact that both proportions, when viewed on a per exposure basis are quite small (.0000394 vs. .000122), instead of using classical asymptotic analysis, I utilized Bayesian methods to identify the posterior probability distributions of difference = p1 – p2, lo relative difference = log (p1/p2), and log odds ratio = log (p1 (1 – p2)/p2(1 – p1). The R script used to analyze the data is included in the appendix.

The resulting graphs from the analysis are shown below. The top left plot provides the estimated cumulative distribution functions (CDFs) of the posterior distributions of p1 and p2. Other three plots present the estimated posterior densities of the difference, log relative ratio and the log odds ratio, respectively. The red vertical line indicated the value of no difference.



Take-Aways - The top left graph shows that the posterior CDF of p1 completely dominates that of p2 indicating with extremely high probability p1 < p2, and is highly likely to be smaller than p2 by a factor of almost 3 (as the posterior median of log relative ratio is -1.11). The top right graph captures the probability density of p1 – p2. There is almost no density to the right of the red vertical line, positioned at zero. Again, extremely strong evidence that p1 < p2. The bottom two graphs draw the same conclusion by providing the probability density for log(p1/p2) as well as log of the odds ratio.

95% Bayes Credible Interval

The 95% credible interval for log(p1/p2) is about (-1.9106, -0.4592), indicating that p2 can be larger than p1 by a factor ranging from exp(0.4592)=1.5 to exp(1.9106)=6.75 with 95% probability. While this may not appear great in magnitude if we extend this thinking to 100,000 exposures, we arrive at the following: on average there will about 3 (=exp(1.11)) times less ACL injuries in the group utilizing the SoccerPulse app.

In closing:

Clearly, there's overwhelming evidence that p1< p2, as the posterior distribution of p1 completely dominates that of p2. The posterior CDF plot indicates p1 is stochastically smaller than p2 which is a very strong case for a reduction in ACL injuries in the group using the SoccerPulse app.

Sujit K Ghosh

Professor
Department of Statistics
NC State University

Appendix – R code

```r
#Two sample binomial comparison
#By Sujit K. Ghosh (2024), NC State University

#Data:
#x=c(4,96); n=c(49536,786293)
x=c(8,96); n=c(203112,786293)

#Prior distribution parameters:
a=b=0.5 #Jeffrey's Beta prior

#Posterior samples from Beta:
#MC sample size
N=1.0e5
p=matrix(0,N,2)
for(j in 1:2){
  p[,j] <- rbeta(N,x[j]+a, n[j]-x[j]+b)}

diff <- p[,1]- p[,2]
log.rel.ratio <- log(p[,1]/p[,2])
log.odds.ratio <- log(p[,1]*(1-p[,2])/(p[,2]*(1-p[,1])))

#Posterior densities:
par(mfrow=c(2,2),lwd=2,cex=0.85)
plot(ecdf(p[,1]),main="post.CDFs of p1 & p2"); lines(ecdf(p[,2]),lty=2,col=2)
legend("bottomright",legend=c("post.cdf of p1", "post.cdf of p2"),lty=1:2,col=1:2)

hist(diff,breaks=30,freq = F,main="post.density of p1-p2")
lines(density(diff),col="blue"); abline(v=0,col="red")
#posterior probability p1-p2>0
mean(diff>0); round(quantile(diff,probs = c(0.025,0.05,0.5,0.95,0.975)),8)

## [1] 0.00018

##        2.5%          5%         50%         95%       97.5%
## -0.00011634 -0.00011102 -0.00008171 -0.00004819 -0.00004099

hist(log.rel.ratio,breaks=30,freq = F,main="post.density of log(p1/p2)")
lines(density(log.rel.ratio),col="blue"); abline(v=0,col="red")

#Posterior probability log.rel.ratio>0
mean(log.rel.ratio>0); round(quantile(log.rel.ratio,probs = c(0.025,0.05,0.5,0.95,0.975)),4)

## [1] 0.00018

##    2.5%      5%     50%     95%   97.5%
## -1.9106 -1.7691 -1.1140 -0.5567 -0.4592

hist(log.odds.ratio,breaks=30,freq = F, main="post.density of log(p1(1-p2)/p2(1-p1))")
lines(density(log.odds.ratio),col="blue"); abline(v=0,col="red")

#posterior probability log.odds.ratio>0
mean(log.odds.ratio>0); round(quantile(log.odds.ratio,probs = c(0.025,0.05,0.5,0.95,0.975)),4)

## [1] 0.00018

##    2.5%      5%     50%     95%   97.5%
## -1.9107 -1.7692 -1.1140 -0.5568 -0.4593
```