# STAT_512_sp_2018_Lec_05_R_supplement

*Karl Gregory*

*2/20/2018*

## Comparing two estimators of the success probability

Let $X_1, \ldots, X_n$ be a random sample from the Bernoulli($p$) distribution, where $p \in (0,1)$ is unknown. Set $Y = X_1 + \cdots + X_n$ and consider the two estimators of $p$ given by $\hat{p} = Y/n$ and $\tilde{p} = (Y+2)/(n+4)$. We have

$$\text{MSE}\,\hat{p} = \frac{p(1-p)}{n} \quad \text{and} \quad \text{MSE}\,\tilde{p} = \left(\frac{n}{n+4}\right)^2 np(1-p) + \left(\frac{2-4p}{n+4}\right)^2.$$

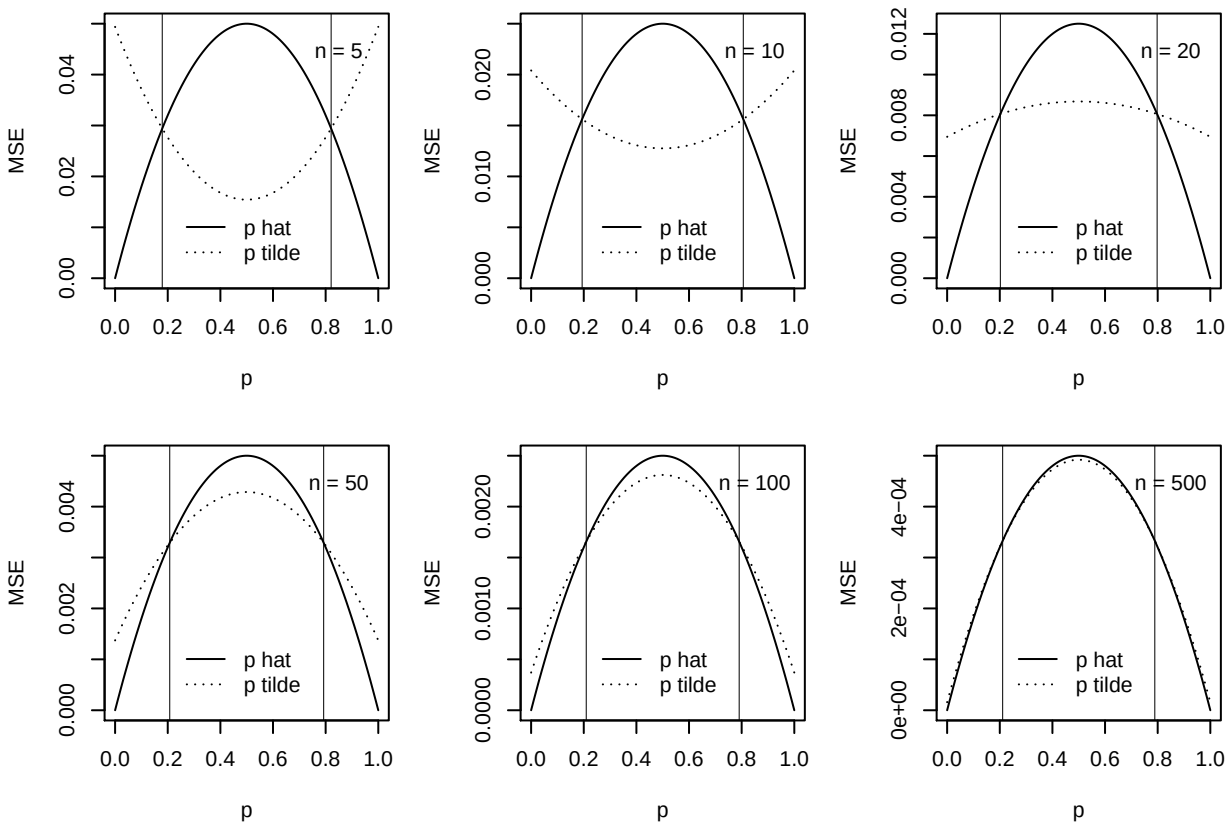
Which estimator has a smaller MSE depends on the value of the unknown parameter $p$. The following R code plots, for different sample sizes $n$, the MSE of $\hat{p}$ and $\tilde{p}$ across all values of the true parameter $p \in (0,1)$.

```
# define functions to compute MSE of \hat p and \tilde p:
MSE.p1 <- function(p,n){p*(1-p)/n}
MSE.p2 <- function(p,n){(n/(n+4))^2*p*(1-p)/n+((2-4*p)/(n+4))^2}
#MSE.p3 <- function(p,n){(n/(n+2))^2*p*(1-p)/n+((1-2*p)/(n+2))^2}

# create a sequence of values between 0 and 1 for the true parameter p
p.seq <- seq(0,1,length=100)

# choose 6 sample sizes n
nn <- c(5,10,20,50,100,500)
par(mfrow=c(2,3),mar=c(5.1,4.1,1.1,1.1))
for(i in 1:length(nn))
{
   # apply MSE.p1 and MSE.p2 to sequence of true values for p at current sample size
    MSE.p1.vals <-  sapply(p.seq,MSE.p1,n=nn[i])
    MSE.p2.vals <-  sapply(p.seq,MSE.p2,n=nn[i])
    #MSE.p3.vals <- sapply(p.seq,MSE.p3,n=nn[i])

    # plot the MSE of \hat p and the MSE of \tilde p against p
    ylims <- range(MSE.p1.vals,MSE.p2.vals)
    plot(MSE.p1.vals~p.seq,ylim=ylims,type="l",xlab="p",ylab="MSE")
    lines(MSE.p2.vals~p.seq,lty=3)
    #lines(MSE.p3.vals~p.seq,lty=3,col="red")
    legend(x=.2,y =grconvertY(.5,from="nfc",to="user"),lty=c(1,3),
           legend=c("p hat","p tilde"),bty="n")
    text(x=.85,y=grconvertY(.85,from="nfc",to="user"),labels=paste("n = ",nn[i],sep=""))
    abline(v = .5 - .5 *sqrt(1 - 2 * nn[i]/(3*nn[i]+2)),lwd=.5)
    abline(v = .5 + .5 *sqrt(1 - 2 * nn[i]/(3*nn[i]+2)),lwd=.5)
    #abline(v = .5 - .5 *sqrt(1 -  nn[i]/(2*nn[i]+1)),lwd=.5,col="red")
    #abline(v = .5 + .5 *sqrt(1 -  nn[i]/(2*nn[i]+1)),lwd=.5,col="red")
}
```




We can determine that when the value of the true parameter $p$ is in the interval

$$\left(1/2 - 1/2\sqrt{1 - 2\left(\frac{n}{3n+2}\right)}, 1/2 + 1/2\sqrt{1 - 2\left(\frac{n}{3n+2}\right)}\right),$$

the estimator $\tilde{p}$ has a lower MSE than $\hat{p}$. The difference is very large at smaller sample sizes, but is very small at larger sample sizes. If one has a small sample size and one believes that the true parameter is not too close to 0 or 1, then $\tilde{p}$ is a better estimator in terms of MSE than $\hat{p}$.