STAT_512_sp_2018_Lec_06_R_supplement

Karl Gregory

2/26/2018

A simulation to illustrate consistency

Consider drawing independent realizations $X_1, X_2, \dots$ from the Bernoulli(p) distribution, one by one, and computing the value of $\hat{p}_n = (X_1 + X_2 + \cdots + X_n)$ for every $n \geq 1$, so that the sequence $\{\hat{p}_n\}_{n\geq 1}$ is given by $\hat{p}_1 = X_1/1$, $\hat{p}_2 = (X_1 + X_2)/2$, $\hat{p}_3 = (X_1 + X_2 + X_3)/3$, and so on. By the weak law of large numbers, the sequence of estimators $\{\hat{p}_n\}_{n\geq 1}$ is consistent for p, meaning that for any $\epsilon > 0$, $P(|\hat{p}_n - p| < \epsilon)$ can be made arbitrarily close to 1 by choosing a large enough n.

The following R code simulates $S = 100$ sequences of estimators $\{\hat{p}_n\}_{n=1}^{2000}$, so that the largest sample size is $n = 2000$. Each of the blue and red lines traces a single realization of the sequence of estimators $\{\hat{p}_n\}_{n=1}^{2000}$. The sequences traced by blue lines are those for which $\hat{p}_{2000}$ was within

$$\epsilon = z_{.025}\sqrt{\frac{0.5(1-0.5)}{2000}}$$

of the true value $p = 0.5$; the remaining sequences are traced by red lines. The black line traces the simulated value of $P(|\hat{p}_n - p| < \epsilon)$, for $n = 1, \dots, 2000$; that is, for each $n = 1, \dots, 2000$, the height of the black line is the proportion of the $S = 100$ sequences for which $|\hat{p}_n - p| < \epsilon$. We see that he height of the black line tends to 1 as n increases, illustrating the consistency of the sequence of estimators $\{\hat{p}_n\}_{n=1}^{2000}$ for p.

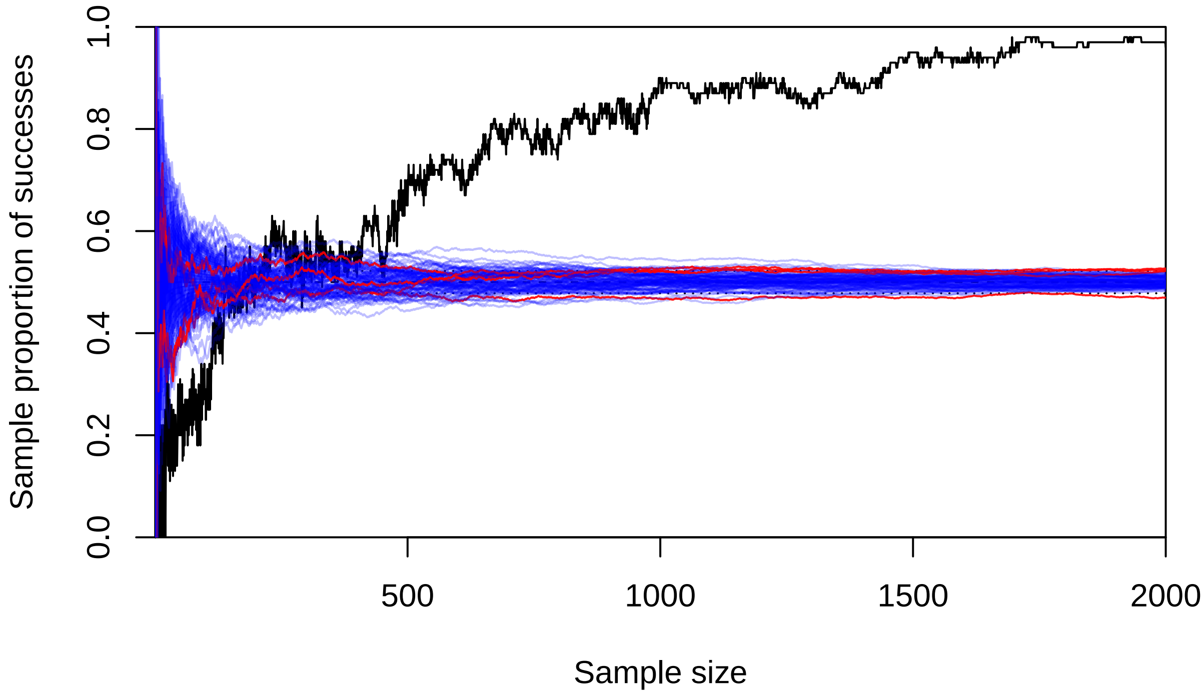
```r
# make bernoulli draws and compute values of \hat p_n sequences
p <- .5
max_n <- 2000
S <- 100
bern_mat <- matrix(rbinom(max_n*S,1,prob=p),max_n,S)
p_hat_mat <- apply(bern_mat,2,cumsum)/row(bern_mat) # each column a sequence of estimators

# choose some \epsilon
alpha <- .05
epsilon <- qnorm(1-alpha/2)*sqrt(p*(1-p)/max_n)

# compute the proportion of sequences for which \hat p_n is within epsilon of p at each n
prop_within_epsilon <- function(x,epsilon,p){ mean(abs(x-p) < epsilon) }
p_within <- apply(p_hat_mat,1,FUN=prop_within_epsilon,epsilon,p)

# make plots
plot(NA,yaxs="i",xaxs="i",xlim=c(1,max_n),ylim=c(0,1),
     xlab="Sample size",ylab="Sample proportion of successes")
lines(p_within~c(1:max_n))
abline(h=.5,lty=2)
abline(h=.5+epsilon,lty=3)
abline(h=.5-epsilon,lty=3)

for(s in 1:S){
    within_epsilon <- abs(p_hat_mat[max_n,s]-p) < epsilon
    lines(p_hat_mat[,s]~c(1:max_n),col=ifelse(within_epsilon,rgb(0,0,1,.25),rgb(1,0,0,.9)))
}
```


Sample proportion of successes
0.0
0.2
0.4
0.6
0.8
1.0
500
1000
1500
2000
Sample size