

Jackknife and Bootstrap for Traffic Problems

One needs to estimate p , the frequency of days with 0 traffic accidents on a certain highway. The data are collected. During 40 days, there are 26 days with 0 accidents, 10 days with 1 accident, and 4 days with 2 accidents.

Statistician A estimates p by the sample proportion $\hat{p} = \frac{26}{40} = 0.65$.

Statistician B argues that this method does not distinguish between the days with 1 accident and the days with 2 accidents, losing some valuable information. She suggests to model the number of accidents X by a Poisson distribution with parameter λ . Then $p = P(X = 0) = e^{-\lambda}$. Statistician B estimates λ with $\hat{\lambda} = \bar{X}$. Then $\hat{p}_{poisson} = \exp(-\hat{\lambda})$ is a new estimator of p , and it uses all the information contained in the data.

However, this estimator is *biased*. Both statisticians use Jackknife to reduce the bias, and together they obtain the jackknife estimator \hat{p}_{JK} .

Now we have three competing estimators - \hat{p} , $\hat{p}_{poisson}$, and \hat{p}_{JK} . Evaluate and rank them according to their bias and standard error?

Solution:

We can generate the data as follows

```
> x = c( rep(0,26), rep(1,10), rep(2,4) )
> X
[1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 2
2 2 2
```

Define the function `P.poisson = exp(- sample mean)` and apply Jackknife.

```
> P.poisson = function( X ){
+ return( exp(-mean(X)) ) }

> P.poisson(X)
[1] 0.6376282

> library(bootstrap)
> JK = jackknife(X,P.poisson)
> JK$jack.bias
[1] 0.003683256
> P.JK = P.poisson(X) - JK$jack.bias
> P.JK
[1] 0.6339449
```

Our three estimates are evaluated as $\hat{p} = 0.65$, $\hat{p}_{poisson} = 0.6376$, and $\hat{p}_{JK} = 0.6339$.

According to the bias... \hat{p} is unbiased; $\hat{p}_{poisson}$ has bias = 0.0037, and \hat{p}_{JK} has a lower bias than $\hat{p}_{poisson}$, because Jackknife reduces bias. Thus, according to the bias, the estimates are ranked as:

- (1) \hat{p}
- (2) \hat{p}_{JK}
- (3) $\hat{p}_{poisson}$

According to the standard deviation...

To use the bootstrap, define three functions for the sample proportion proposed by Statistician A, the estimator proposed by Statistician B, and for its jackknife version.

```
> P.hat = function( X, sample ){
+ return( mean(X[sample]==0) )}

> P.poisson = function( X, sample ){
+ return( exp(-mean(X[sample])) )}

> P.JK = function( X, sample ){
+ n = length(X)
+ P.i = rep(0,n)
+ for (i in 1:n){ P.i[i] = P.poisson(X[sample],-i) } # Leave one out
+ return( n*P.poisson(X[sample],) - (n-1)*mean(P.i) )} # Jackknife
formula
```

Now, use bootstrap to evaluate performance of these estimators from the given data.

```
> library(boot)
> boot( X, P.hat, R=10000 )
      original      bias    std. error
t1*    0.65 -0.0005225  0.07590366

> boot( X, P.poisson, R=10000 )
      original      bias    std. error
t1* 0.6376282 0.004304599 0.06729368

> boot( X, P.JK, R=10000 )
      original      bias    std. error
t1* 0.6339449 0.004085077 0.06780789
```

We knew that the first estimator \hat{p} is unbiased. However, it ignores some information, as it is correctly noticed in the problem. As a result, it has the highest standard error, among these three competing estimators. The second estimator $\hat{p}_{poisson}$ has some bias, but it has a lower standard error. Its jackknife version makes some bias reduction, but it is at the expense of a little higher standard error.

These results may vary somewhat because they are based on random bootstrap samples.

Summary

According to the lower bias, the estimators are ranked as

- (1) \hat{p} (best)
- (2) \hat{p}_{JK}
- (3) $\hat{p}_{poisson}$ (worst)

According to the lower standard deviation (standard error), the estimators are ranked as

- (1) $\hat{p}_{poisson}$ (best)
- (2) \hat{p}_{JK}
- (3) \hat{p} (worst)