3. Bootstrap for Sampling Distribution Recovery

Dimitrios Ligas

Asymptotic Results

The approximate asymptotic standard error and 95% Confidence Interval for the MLE estimator of θ are:

```
##                     Asymptotic Results theta_hat mle
## theta_mle                                   0.032760
## Approx SE                                   0.006348
## Approx 2.5%-ile                             0.020318
## Approx 97.5%-ile                            0.045203
## 95% Approx CI Range                         0.024885
```

First, we use the nonparametric bootstrap technique to recover the sampling distribution of $\hat{\theta}_{MLE}$.

Define a routine to compute $\hat{\theta}_{MLE}$ for each generated sample based on the equation:

$\Sigma_{i=1}^{4} X_i \hat{\theta}_{MLE}^2 \ -(X_1 - 2X_2 - 2X_3 - X_4)\hat{\theta}_{MLE} \ - 2X_4 = 0$

```
theta_mle=function(X){
  a=sum(X)
  b=-(X[1]-2*X[2]-2*X[3]-X[4])
  c=-2*X[4]
  d=(b^2)-4*(a*c)
  l=sqrt(d)
  x1=(-b-l)/(2*a)
  x2=(-b+l)/(2*a)
  if((x2>0)&(x2<1)){
    theta=x2
    }else{
    theta=x1}
  return(theta)
}
```

Nonparametric Bootstrap:

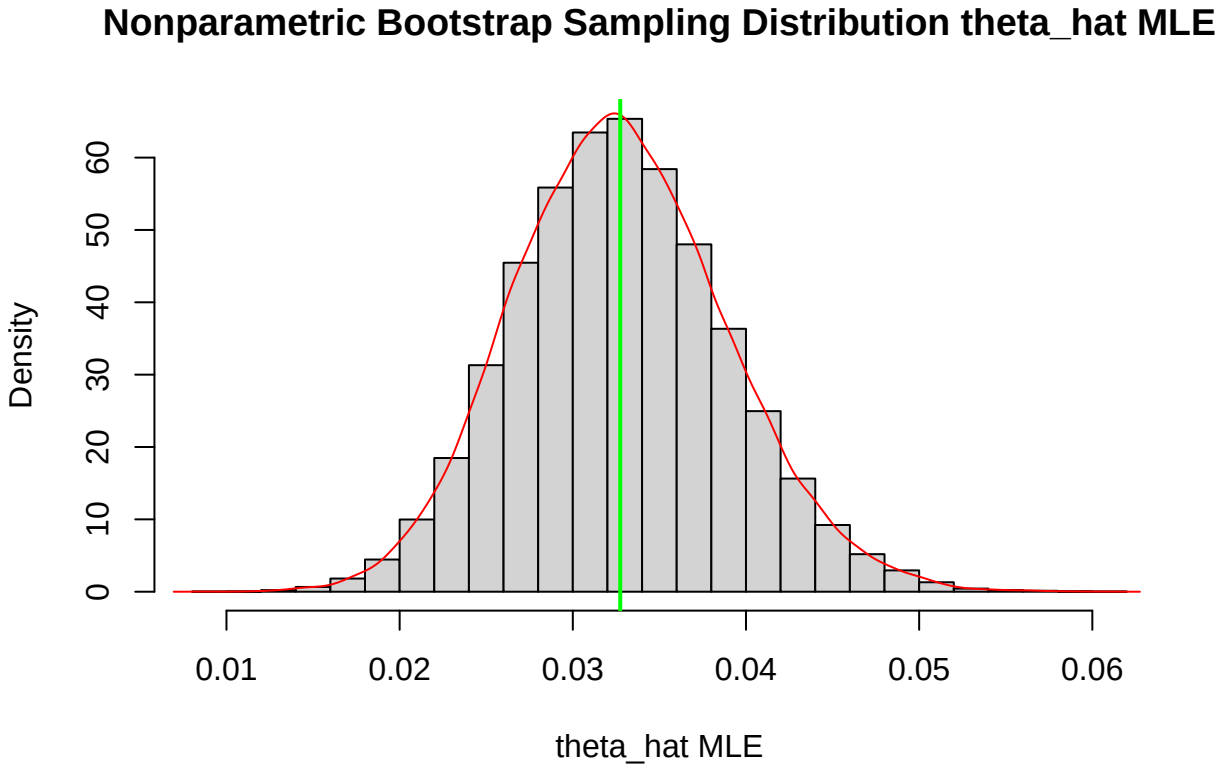
```
#Nonparametric bootstrap for sampling distribution recovery
#Load the data
set.seed(784)
X=c(1495,750,729,26)
#Sample size
n=sum(X)
#Empirical Probability Distribution of the Multinomial Cells:
p_emp=X/n
theta_hat=theta_mle(X)
#Pseudocategories:
categories=c("A","B","C","D")
S=50000 #Number of level 1 bootstrap samples
B=500 #Number of level 2 bootstrap samples
```

```
theta=rep(0,S)
theta_boot=rep(0,B)
t_ratio=rep(0,S)
#Resample data of the same size using the empirical distribution:
#Generate samples for level 1 bootstrap
samples=replicate(S,sample(categories,n,replace=TRUE,prob=p_emp))
#Generate samples for level 2 bootstrap
boot_samples=replicate(B,sample(categories,n,replace=TRUE,prob=p_emp))
#Preparing for bootstrap:
countA=colSums(samples=="A")
countA_boot=colSums(boot_samples=="A")
countB=colSums(samples=="B")
countB_boot=colSums(boot_samples=="B")
countC=colSums(samples=="C")
countC_boot=colSums(boot_samples=="C")
countD=colSums(samples=="D")
countD_boot=colSums(boot_samples=="D")
for (i in 1:S) {
  #Pass the counts/frequency of each category for every sample and find the MLE:
  Y=c(countA[i],countB[i],countC[i],countD[i])
  theta[i]=theta_mle(Y)
  #Level 2 Bootstrap for the t-intervals:
  for (j in 1:B) {
    Z=c(countA_boot[j],countB_boot[j],countC_boot[j],countD_boot[j])
    theta_boot[j]=theta_mle(Z)
    }
  se=sd(theta_boot)
  t_ratio[i]=(theta[i]-theta_hat)/se
}
mean_theta=mean(theta)
se=sd(theta)
#Empirical bootstrap percentiles:
q=c(0.025,0.975)
s=as.matrix(theta)
Q=apply(s,2,quantile,q)
boot_range=Q[2]-Q[1]
#Revised bootstrap percentiles:
revised_boot=c(2*theta_hat-Q[2],2*theta_hat-Q[1])
revised_range=revised_boot[2]-revised_boot[1]
#Bootstrap t-intervals:
t_interval=c(theta_hat-quantile(t_ratio,0.975)*se,theta_hat-quantile(t_ratio,0.025)*se)
t_range=t_interval[2]-t_interval[1]
nonparam_summary=round(rbind(mean_theta,se,Q[1],revised_boot[1],t_interval[1],Q[2],revised_boot[2],t_in
colnames(nonparam_summary)="Nonparametric Sampling Distribution"
rownames(nonparam_summary)=c("Mean","Boot SE","Boot 2.5%-ile","Revised Boot 2.5%-ile","2.5% t-Percentil
```

The results of the nonparametric bootstrap are:

```
##                          Nonparametric Sampling Distribution
## Mean                                                0.032740
## Boot SE                                             0.006156
## Boot 2.5%-ile                                       0.021240
## Revised Boot 2.5%-ile                               0.020176
## 2.5% t-Percentile                                   0.019855
```

```
## Boot 97.5%-ile                          0.045345
## Revised Boot 97.5%-ile                  0.044280
## 97.5% t-Percentile                      0.044574
## Boot CI 95% Range                       0.024105
## Revised Boot CI 95% Range               0.024105
## t-interval 95% Range                    0.024719
```



Parametric Bootstrap:

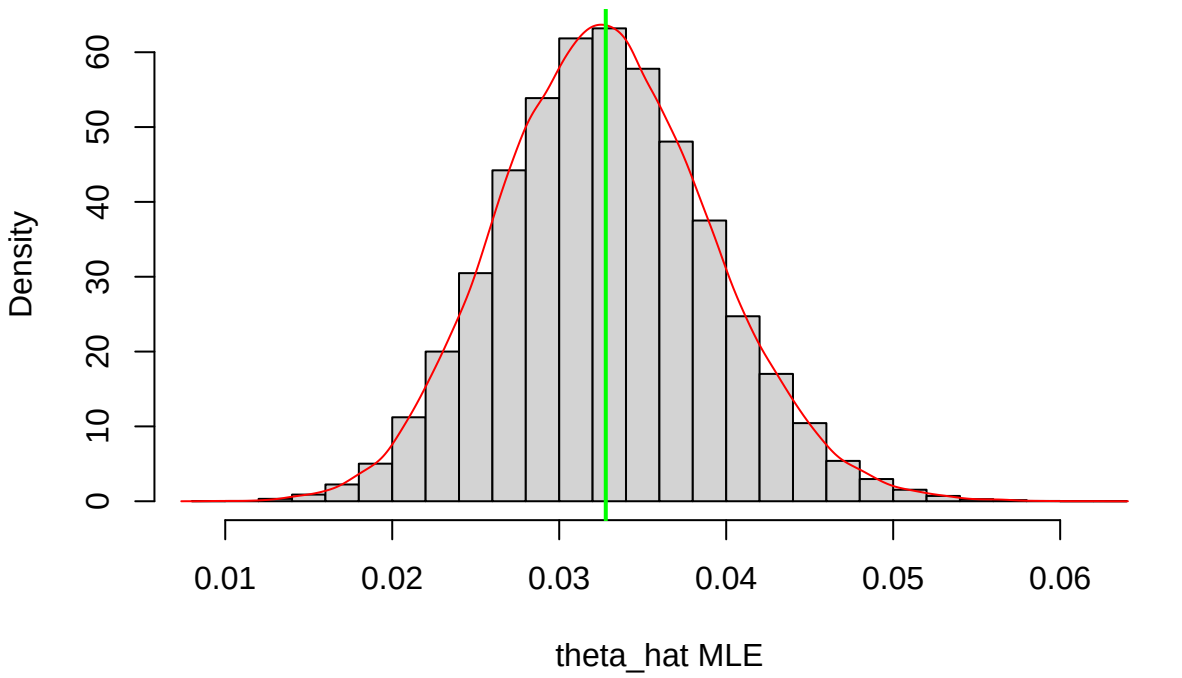
Next we use the parametric bootstrap technique to recover the sampling distribution of $\hat{\theta}_{MLE}$:

```
#Parametric Bootstrap for the Sampling Distribution of theta MLE
#Load the data:
set.seed(784)
X=c(1495,750,729,26)
n=sum(X)
#Initial MLE estimator given the observed data:
theta_hat=theta_mle(X)
#Preparing Bootstrap Parameters:
S=50000 # Number of Level 1 bootstrap samples
B=500 # Number of Level 2 bootstrap samples
#Multinomial Cell Probabilities:
p1=0.25*(2+theta_hat)
p2=p3=0.25*(1-theta_hat)
p4=0.25*theta_hat
p=c(p1,p2,p3,p4)
#Generate level 1 bootstrap samples:
```

```
samples=rmultinom(S,n,p)
#Generate level 2 bootstrap samples:
boot_samples=rmultinom(B,n,p)
#Obtain the MLE estimator for each generated sample:
theta=rep(0,(dim(samples)[2]))
theta_boot=rep(0,B)
t_ratio=rep(0,(dim(samples)[2]))
#Initial Value of the MLE estimator:
theta[1]=theta_hat
for (i in 2:S) {
  theta[i]=theta_mle(samples[,i])
  #Level 2 Bootstrap for the t-intervals:
  for (j in 1:B) {
    theta_boot[j]=theta_mle(boot_samples[,j])
  }
  se=sd(theta_boot)
  t_ratio[i]=(theta[i]-theta_hat)/se
}
#Summarizing the parametric bootstrap sampling distribution:
mean_theta=mean(theta)
se=sd(theta)
#Empirical bootstrap percentiles:
q=c(0.025,0.975)
s=as.matrix(theta)
Q=apply(s,2,quantile,q)
boot_range=Q[2]-Q[1]
#Revised bootstrap percentiles:
revised_boot=c(2*theta_hat-Q[2],2*theta_hat-Q[1])
revised_range=revised_boot[2]-revised_boot[1]
#Bootstrap t-intervals:
t_interval=c(theta_hat-quantile(t_ratio,0.975)*se,theta_hat-quantile(t_ratio,0.025)*se)
t_range=t_interval[2]-t_interval[1]
param_summary=round(rbind(mean_theta,se,Q[1],revised_boot[1],t_interval[1],Q[2],revised_boot[2],t_inter
colnames(param_summary)="Parametric Sampling Distribution"
rownames(param_summary)=c("Mean","Boot SE","Boot 2.5%-ile","Revised Boot 2.5%-ile","2.5% t-Percentile",
```

The results of the parametric bootstrap are:

```
##                          Parametric Sampling Distribution
## Mean                                             0.032789
## Boot SE                                          0.006341
## Boot 2.5%-ile                                    0.020870
## Revised Boot 2.5%-ile                            0.019853
## 2.5% t-Percentile                                0.020507
## Boot 97.5%-ile                                   0.045668
## Revised Boot 97.5%-ile                           0.044650
## 97.5% t-Percentile                               0.044048
## Boot CI 95% Range                                0.024798
## Revised Boot CI 95% Range                        0.024798
## t-interval 95% Range                             0.023541
```



Comparing the results of the two bootstrap methods with the asymptotic results:

```
##                     Asymptotic Results theta_hat mle
## theta_mle                                   0.032760
## Approx SE                                   0.006348
## Approx 2.5%-ile                             0.020318
## Approx 97.5%-ile                            0.045203
## 95% Approx CI Range                         0.024885
```

```
##                            Nonparametric Sampling Distribution
## Mean                                                  0.032740
## Boot SE                                               0.006156
## Boot 2.5%-ile                                         0.021240
## Revised Boot 2.5%-ile                                 0.020176
## 2.5% t-Percentile                                     0.019855
## Boot 97.5%-ile                                        0.045345
## Revised Boot 97.5%-ile                                0.044280
## 97.5% t-Percentile                                    0.044574
## Boot CI 95% Range                                     0.024105
## Revised Boot CI 95% Range                             0.024105
## t-interval 95% Range                                  0.024719
```

```
##                            Parametric Sampling Distribution
## Mean                                               0.032789
## Boot SE                                            0.006341
## Boot 2.5%-ile                                      0.020870
## Revised Boot 2.5%-ile                              0.019853
```

```
## 2.5% t-Percentile                    0.020507
## Boot 97.5%-ile                       0.045668
## Revised Boot 97.5%-ile               0.044650
## 97.5% t-Percentile                   0.044048
## Boot CI 95% Range                    0.024798
## Revised Boot CI 95% Range            0.024798
## t-interval 95% Range                 0.023541
```

The parametric bootstrap standard error of $\hat{\theta}_{MLE}$ is 0.006341 and almost equal to the approximate asymptotic standard error $SE[\hat{\theta_{MLE}}] = 0.006348$. The nonparametric standard error of $\hat{\theta}_{MLE}$ is 0.006156, it is significantly smaller than the corresponding asymptotic one, and it seems that the nonparametric bootstrap method underestimates the inherent uncertainty of the MLE estimator. For the confidence intervals, the approximate asymptotic 95% CI is: $0.020318 \leq \theta \leq 0.045203$. We compare this with the 95% Bootstrap t-intervals, which is the most reliable method for confidence intervals approximation compared to the empirical 95% bootstrap percentiles and the revised 95% bootstrap percentiles. The parametric 95% bootstrap t-interval is: $0.020507 \leq \theta \leq 0.044048$ and the nonparametric bootstrap t-interval is: $0.019855 \leq \theta \leq 0.044574$. The nonparametric lower and upper bounds are both lower than the approximate asymptotic bounds, which means that the nonparametric sampling distribution is somewhat skewed. The parametric lower bound is greater than the approximate asymptotic and the upper bound is lower than the corresponding asymptotic one. The parametric lower bound is closer to the asymptotic lower bound, while the nonparametric upper bound is closer to the asymptotic upper bound. Overall, the parametric bootstrap t-intervals are narrower than the nonparametric ones.

Most appropriate bootstrap method:

The nonparametric bootstrap method approximates the sampling distribution of $\hat{\theta}_{MLE}$ by resampling the observed data without making any assumptions about the underlying parametric likelihood model that generated the data, and does not require us to know a priori the specific probability distribution. In contrast, the parametric bootstrap method approximates the sampling distribution of $\hat{\theta}_{MLE}$ by making the assumption that the data given the parameters are distributed according to a parametric likelihood $X|\theta \sim f(X|\theta)$. Then, it substitutes for the unknown parameter θ the MLE estimator and draws S many samples from the likelihood $X|\hat{\theta_{MLE}} \sim f(X|\hat{\theta_{MLE}})$. In my opinion, in this specific situation the most appropriate method to approximate the sampling distribution of $\hat{\theta}_{MLE}$ is the parametric bootstrap for the following reasons:

1. For this specific case we truly know that the likelihood of the data is $X = (X1, X2, X3, X4)|\theta \sim Multinomial(3000, 4, g(\theta))$. Therefore, our assumption about the probability distribution that generated the data is unlikely to be incorrect. Since we do know that the data follows this probability law, there is no reason why we should not try to approximate the underlying distribution.

2. Since θ is unknown, we draw S many samples from $X|\hat{\theta_{MLE}} \sim f(X|\hat{\theta_{MLE}})$. From the asymptotic theory of MLE, we know that as the sample size $n \to \infty \quad \mathbb{E}[\hat{\theta_{MLE}}] \to \theta$. The sample size in this case is $n = 3000$ which is adequately large enough so that we can expect that the MLE estimator approximates the true value of the parameter θ reasonably well. As a result, it may be reasonable to anticipate that $X|\hat{\theta_{MLE}} \sim f(X|\hat{\theta_{MLE}}) \xrightarrow{d} X|\theta \sim f(X|\theta)$ so that we are actually sampling from the true likelihood of the data.

3. In the nonparametric resampling, we cannot generate samples beyond the empirical distribution that we have already observed. In this specific case, for the 4th category "Sugar-White", the counts observed are $X4 = 26$, a very small number compared to the other categories, suggesting that X4 most probably constitutes an outlier. Using the empirical pmf, category 4 will be resampled with a constant probability $p_4 = 0.25\frac{26}{3000} \approx 0.0087$, while in the parametric bootstrap category 4 will be resampled with a constant probability $p_4 = 0.25\hat{\theta_{MLE}} \approx 0.0082$. Thus, nonparametric bootstrap will overweight the probability of occurrence of category 4. In the total of the nonparametric samples, category 4 was sampled 1,298,858 times, while in the total of the parametric samples category 4 was sampled 1,229,742 times. Hence, nonparametric bootstrap has a tendency to lead in a more skewed sampling distribution.

4. Lastly, because the nonparametric bootstrap does not explore the whole multinomial probability distribution of the data, it tends to underestimate the variance of the multivariate random variable X and so it underestimates the uncertainty of $\hat{\theta}_{MLE}$, which is reflected in the smaller standard error.