Stat-581: Probability and Statistical Inference for Data Science
Assignment - 2 : Bootstrap and Jackknife

Name: Prakruti Joshi
NetID: phj15

---

1. Build a bias, standard deviation, and confidence interval estimator for the mean based on the bootstrap (use 10000 = nboot) and the jackknife.

Bootstrap

For bias correction in bootsrap, we subtract the first moment and divide the second moment from '*meanb'* in order to make the result independent from the sample:

```
[1] bootvec <- c(bootvec,(meanb-mean0)/(sdb/sqrt(n0)))
```

Now, the estimated mean is given by (Σmeanb/nboot). We can get the approximate estimate from bootvec (assuming sd0 ≅ sdb) by multiplying the sd and adding back mean0:

```
[1] bootestimate <- mean((sd0/sqrt(n0))*bootvec + mean0)
```

To get the exact estimate from resampling and simulation, we can define a new vector to store the '*meanb'* and the mean of the vector will be equal to the estimated.

```
[1] bootvec_estimate <- c(bootvec_estimate, meanb)
[2] bootestimate <- mean(bootvec_estimate)
```

Standard deviation and bias of bootstrap is calculated as:

```
[1] bootsd <- sd(bootvec)
[2] bootbias <- mean(bootbiasvec)
```

The Bootstrap function returns:

```
[1] list( bootestimate = bootestimate, bootbias = bootbias,
          bootsd = bootsd,  bootstrap.confidence.interval=c(LB,UB),
          normal.confidence.interval=c(NLB,NUB) )
```

Jackknife

The normal approximation based confidence intervals are good when n is large (because of CLT) or when the observations are normally distributed. When sample size is 30 or more, we consider the sample size to be large and by central limit theorem, the sample mean will converge to a normal distribution even if the sample does not come from a normal distribution.
If n<30, the skewness of the population could influence the shape of the sampling distribution and the sample mean may not converge to a normal. In this case, we will not assume that the sample distribution is normal. Instead, we will use a *t-distribution,* which is designed to give us a better interval estimate of the mean when we have a small sample size.
The t-distribution incorporates the fact that for smaller sample sizes the distribution will be more spread out using something called *degrees of freedom.* For confidence intervals, the degrees of freedom will always be `df = n-1`, or one less than the sample size. For every change in degrees of freedom, the t-distribution changes. The larger the sample size (n), the closer the t-distribution mimics the z-distribution in shape. We construct a confidence interval for small sample size in the same way as we do for a large sample, except we use the t-distribution instead of the z-distribution. [1][2]

Hence, for the jackknife normal approximation confidence Interval, I have used t-distribution for the normal approximation since the sample size is taken to be around 30.

```
[1] TLB <- mean(v1) - (sd0/sqrt(n0)) * qt(1-alpha/2, df = n0-1)
[2] TUB <- mean(v1) + (sd0/sqrt(n0)) * qt(1-alpha/2, df = n0-1)
[3] t.confidence.interval = c(TLB, TUB)
```

Standard deviation and bias of jackknife is calculated as:

```
[1] bootsd <- mean(jackvec) - mu0
[2] bootbias <- sd(jackvec)
```

The Jack knife function returns:

```
[1] list( mu0=mu0, jackestimate = mean(jackvec),
          jackbias= jackbias, jacksd=jacksd,
          normal.confidence.interval=c(NLB,NUB))
```

Result

Data used: (from normal distribution)

```
[1] v1 = rnorm(30, mean = 10, sd = 2)
```

Function parameters:

Sample size = 30
alpha = 0.05
nboot = 1000

Mean of the sample = 10.08564

	Estimate	**Bias**	**Standard Deviation**	**CI**
Bootstrap	10.09852	0.0128774	1.06027	[9.258455, 10.989392]
Jacknife	10.08564	5.329071e-15	2.252094	[9.244693, 10.92658]

Table 1: Result of the bias, standard deviation, and confidence interval estimator for the mean

2. Build a simulator that draws n samples from a lognormal distribution (rlnorm) and builds both the central limit theorem based confidence interval, and compares it to the coverage rate for the 3 bootstrap and the jackknife normal based confidence interval (with bias correction (how would you do that)) (confidence interval based on the 1st program). (1000 simulation runs minimum)

- *Normal approximation*
- *Pivotal CI*

Bias Correction:

1. Bootstrap

For bias correction, we subtract the first moment and divide the second moment from '*meanb'* in order to make the result independent from the sample. Hence its distribution won't depend on the parameters.

```
[1] bootvec<-c(bootvec,(meanb-mean0)/(sdb/sqrt(n0)))
```

2. Jackknife

Bias corrected Jackknife estimate is given by:

$$\hat{\theta}_{\text{Jack}} = n\hat{\theta} - (n-1)\hat{\theta}_{(.)}$$

This formula for Jackknife comes from the interpretation of Jackknife in terms of pseudo values for the estimator.

$$ps_i(X) = n\phi_n(X_1, X_2, \ldots, X_n) - (n-1)\phi_{n-1}(\,(X_1, \ldots, \ldots, X_n)_{[i]})$$

$X_{[i]}$ means the sample $X = (X_1,X_2,...,X_n)$ with the ith value X_i deleted from the sample, so that $X_{[i]}$ is a sample of size $n - 1$. Note,

$$ps_i(X) = \phi_n(X) + (n-1)\big(\phi_n(X) - \phi_{n-1}(X_{[i]})\big)$$

so that $ps_i(X)$ can be viewed as a bias-corrected version of $\varphi_n(X)$ determined by the trend in the estimators $\varphi_n(X)$ from $\varphi_{n-1}(X_{[i]})$ to $\varphi_n(X)$. [3]

Implementation:

```
[1] jackvec<-c(jackvec, n1*(mu0)-(n1-1)*mua)
```

Result

Boot Coverage	**CLT coverage**	**Jackknife (t) coverage**
0.872	0.820	0.832

Table 2: Coverage rates, sample size = 30, simultations = 1000

3. Compare the coverage rates for the bootstrap confidence interval, the jackknife normal approximation confidence interval and the central limit theorem based confidence interval. For sample sizes 10, 30, and 100 alpha=0.05 (95% confidence)

Result

Index	n_sample	boot.cov	norm.cov	t.cov
1	10	0.853	0.778	0.800
2	30	0.877	0.838	0.845
3	100	0.882	0.868	0.871

Table 3: Coverage rates, sample size = 10, 30 and 100, simultations = 1000

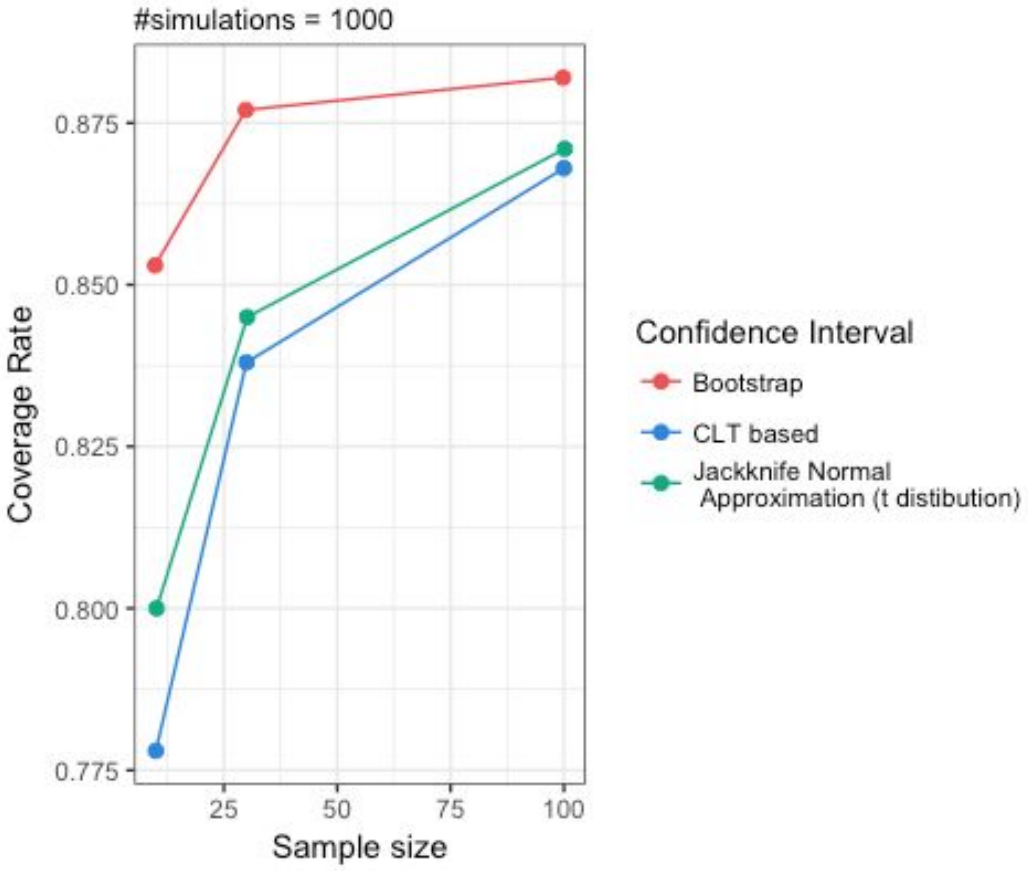

Figure 1: Coverage rate vs sample size

From the results, we can see that

1. Coverage rate increases with increase in the sample size.
2. Coverage rate for bootstrap is the highest amongst all three.
3. Since the sample size is not that large, t-distribution better captures the data than the normal distribution.
4. As the size of sample increases from 10, 30, 100, the difference between the coverage of t and normal confidence intervals reduces.

Code Snippet:

```
[1] df <-  data.frame(n_sample = numeric(), boot.cov = numeric(), [1]
[2] norm.cov = numeric(), t.cov = numeric())
[3] sample_size = c(10,20,30)
[4] nsim = 1000
[5] for (n_sample in sample_size)
[6] {
[7]   coverage.list <- Sim.func(mu.val=3,n=n_sample, nsim= nsim)
[8]   boot.cov<-coverage.list$boot.coverage
[9]   norm.cov<-coverage.list$norm.coverage
[10]  t.cov <- coverage.list$t.coverage
[11]  l <- list(n_sample = n_sample,boot.cov = boot.cov, norm.cov =
[12]            norm.cov, t.cov=t.cov )
[13]   df <- rbind(df, l)
[14] }
```

4. For the standard deviation of the normal distribution, estimate the bias of the sample standard deviation when dividing by n, compare the bootstrap and the jackknife (1000 simulations).

Data

```
[1] v1 <- rnorm(1000) //standard normal
```

Bootstrap implementation for standard deviation

```
[1] for( i in 1:nboot) {
[2]     vecb<-sample(vec0,replace=T)
[3]     sdb<-sqrt(var(vecb))
[4]     bootvec<-c(bootvec, sdb/sqrt(n0))
[5]     bootbiasvec<-c( bootbiasvec, sdb-sd0 )
[6]   }
[7]   bootsd <- sd(bootvec)
[8]   bootbias<-mean(bootbiasvec)
[9]   bootestimate <- sqrt(n0) * mean(bootvec)
[10] }
```

Simulating 1000 samples

```
[1] Sim.func<-function(mu.val=3,n=30,nsim=1000){
[3]   boot.estimate<-NULL
[4]   boot.bias<-NULL
[5]   boot.sd <- NULL
[6]   jack.sd <- NULL
[7]   for(i in 1:nsim){   #run simulation
[8]       vec.sample <- rnorm(n)  #sample the simulation vector
[9]       boot.list<-my.bootstrapci.sd(vec.sample,alpha = 0.05)
[10]      boot.estimate<- c(boot.estimate, boot.list$bootestimate)
[11]      boot.bias<-c(boot.bias, boot.list$bootbias)
[12]      boot.sd <- c(boot.sd, boot.list$bootsd)
[13]      jack.list <- Jackknife_confidence_interval(vec.sample, sd)
[14]      jack.sd <- c(jack.sd, jack.list$jacksd)
[15]   }
[16]   list(avg.boot.estimate = (sum(boot.estimate)/nsim),
[17]       avg.boot.bias = (sum(boot.bias)/nsim),
[18]       avg.boot.sd = (sum(boot.sd)/nsim),
[19]       avg.jack.sd = (sum(jack.sd)/nsim),
[20]       boot.sd = boot.sd, jack.sd = jack.sd)  }
```

Result

Sample size = 30
Simulations = 1000

Boot.sd	**Jack.sd**
0.02204055	0.710938

Table4: Bias of bootstrap and jackknife estimator for, sample size = 30, simultations = 1000

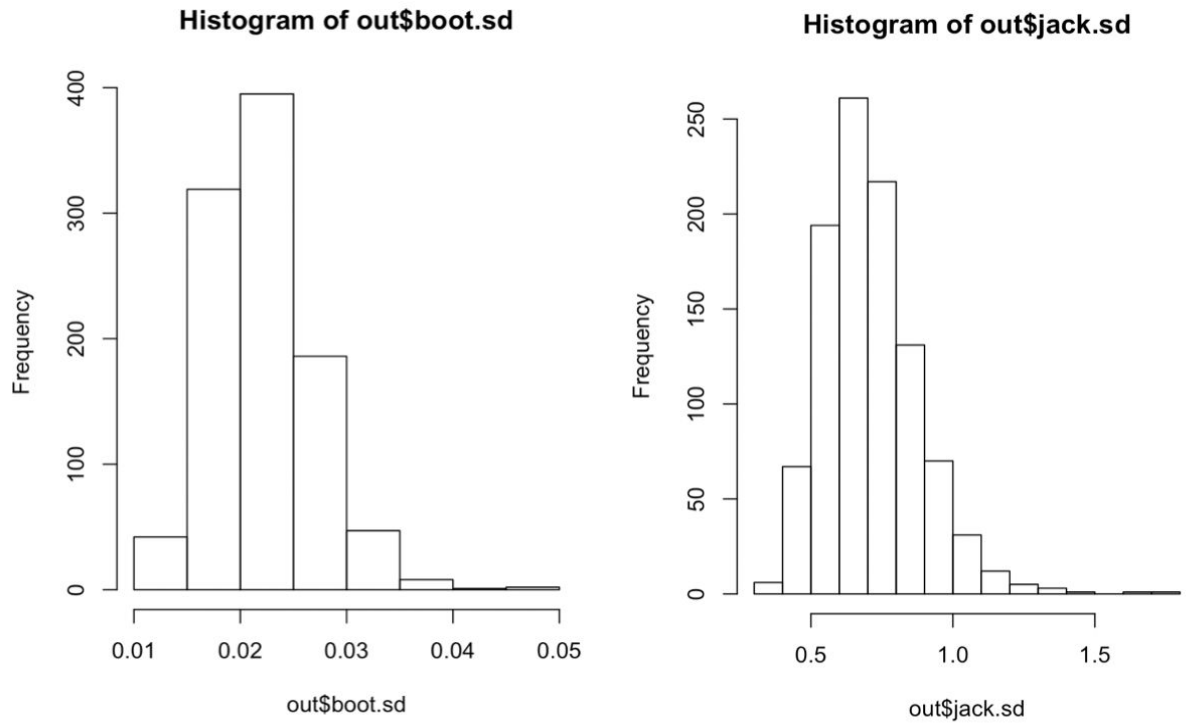

Figure 2: Histogram of bias of standard deviation estimators of bootstrap and jackknife

- Standard deviation of bootstrap estimate of bias is lesser than the jackknife estimate of the bias for a standard normal distribution.

References:

[1] https://pages.wustl.edu/montgomery/articles/2757
[2] https://newonlinecourses.science.psu.edu/stat506/node/8/
[3] https://www.math.wustl.edu/~sawyer/handouts/Jackknife.pdf