# Question 2

Jin Barai

**Description of population:**

The data set is of Billboard Top 30 songs from 2010 to 2019. The variate we are interested is the beats per minute (tempo) of each song.

**Comparing the attributes using a sample design**

We will now find and compare the total and mean bpm using the sampling design - SRSWOR (Sampling Mechanism without Replacement). Then we will go on to understand the effect of changes such as sample size on these estimators.

```
spotify <- read.csv("./spotify.csv", header=TRUE)
popSpotify <- rownames(spotify)
srswor <- createSamplingMechanism(popSpotify)
set.seed(341)
sample_indx <- as.numeric((srswor(10)))
Spotify <- spotify[sample_indx, ]
```

- *Total Bpm*

```
totalBpm <- sum(spotify$bpm)
print(totalBpm)
```

```
## [1] 35876
```

- *Mean bpm*

```
aveBpm <- mean(spotify$bpm)
print(aveBpm)
```

```
## [1] 119.5867
```

- Marginal Inclusion probability $\pi_u$ for SRSWOR is $\frac{n}{N}$ where $n$=30 for all $u$
- The joint inclusion probability $\pi_{u,v}$ when sampling without replacement is $\frac{n(n-1)}{N(N-1)}$ for any pair $(u, v)$ and $u \neq v$

(i) We will now find and display the **Horvitz-Thompson Estimates** on the left and **True Population Value** on the right of the following attributes:

- *Total Bpm*:

```
y_u <- Spotify$bpm
pi_u <- pi[sample_indx]
c(sum(y_u/pi_u), sum(spotify$bpm))
```

```
## [1] 11890 35876
```

- *Average Bpm*:

```
y_u <- Spotify$bpm/N
pi_u <- pi[sample_indx]
c(sum(y_u/pi_u), sum(spotify$bpm/N))
```

```
## [1]  39.63333 119.58667
```

(ii) We will now find and display the **Estimate of the variance of the HT estimator** on the left and the **Estimate of the standard deviation of the HT estimator** also known as standard error on the right of the attributes:

- *Total Bpm*:

```
y_u <- Spotify$bpm
v <- estVarHT(sam = sample_indx, y_u, pi, pij)
c(v, sqrt(v))
```

```
## [1] 9163517.59    3027.13
```

- *Average Bpm*:

```
y_u <- Spotify$bpm/N
v <- estVarHT(sam = sample_indx, y_u, pi, pij)
c(v, sqrt(v))
```

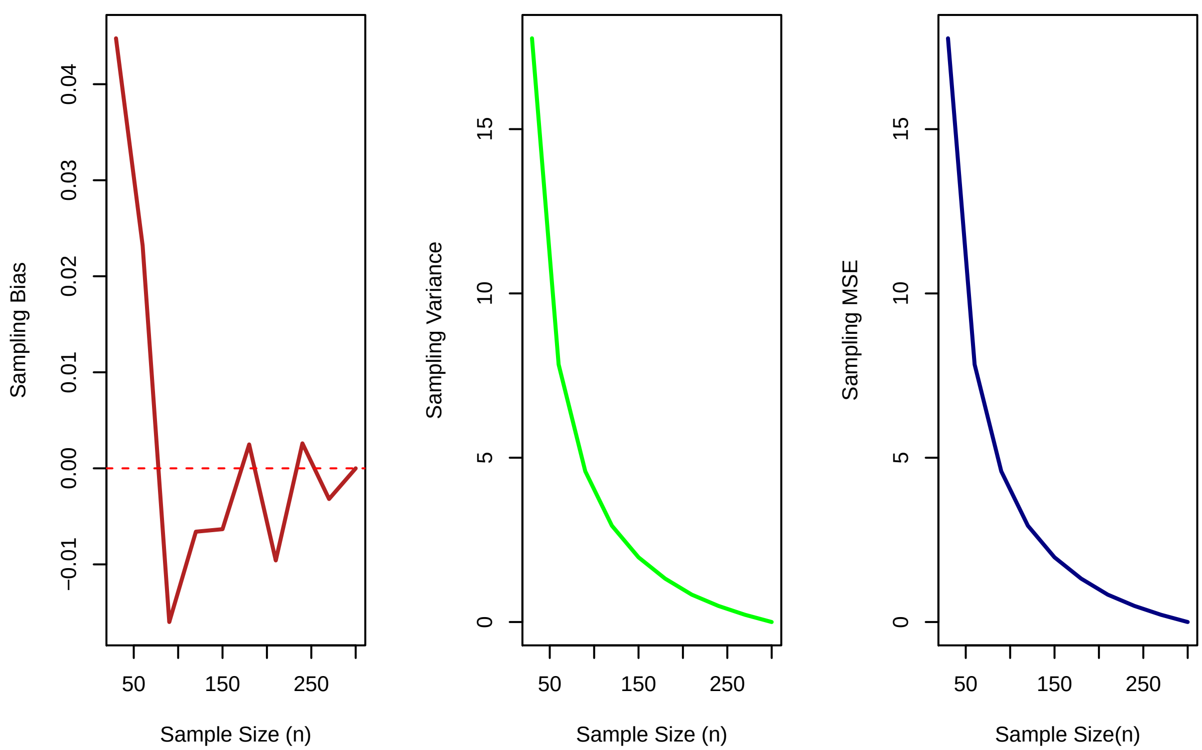
```
## [1] 101.81686  10.09043
```

### Effect of change in some parameters such as Sample Size

We will now consider the sample sizes $n \in \{30, 60, 90, .., 300\}$. For each of the sample size n, we will take 20,000 SRSWOR samples.

```
N <- dim(spotify)[1]
n <- seq(30,300,30)
bias <- rep(0,length(n))
variance <- rep(0,length(n))
mse <- rep(0,length(n))
for (i in 1:length(n)) {
  pi.vec <- rep(n[i]/N,N)
  pi.mat <- matrix((n[i]*(n[i]-1))/(N*(N-1)), nrow=N,ncol=N)
  avgBpmHT <- rep(0,20000)
  for (j in 1:20000){
```

```r
    srsSampIndex <- sample(N, n[i])
    y_u <- spotify$bpm[srsSampIndex]/N
    pi_u <- pi.vec[srsSampIndex]
    avgBpmHT[j] <- sum(y_u/pi_u)
  }
  bias[i]<-mean(avgBpmHT - aveBpm)
  variance[i] <- var(avgBpmHT)
  mse[i] <- mean((avgBpmHT - aveBpm)^2)
}
par(mfrow = c(1,3))
plot(n, bias, type="l", xlab="Sample Size (n)", ylab = "Sampling Bias", col="firebrick", lwd = 2)
abline(h=0, col="red", lty=2)
plot(n, variance, type="l", xlab="Sample Size (n)", ylab="Sampling Variance", col="green", lwd=2)
plot(n, mse, type="l", xlab="Sample Size(n)", ylab="Sampling MSE", col="navyblue", lwd=2)
```




**Conclusion**

- HT estimates become more accurate and precise as n increases
- We see that regardless of the sample size the estimated bias of the estimator fluctuates around 0. This is because the HT estimator is unbiased
- We can see that as the sample size increases, the estimated variance of the estimator decreases and approaches 0.
- We can aslo see that as sample size increases, the estimated MSE of the estimator decreases and approaches 0.