# Question 3

### What is a Sample?

Sometimes, it's not possible to calculate an attribute for the entire population because of various reasons - such as the entire population is too large, or it's not completely available to us, or even if the attribute is too complex. In the such scenarios, we would take a subset $S$ of $n << N$ units of the population to calculate the attribute. The result that we get based on the sample is an **estimate** $a(S)$ of the population attribute $a(P)$ such that,

$$a(S) = a(\hat{P})$$

While using a sample, we consider 2 important aspects - **Sample Error and Fisher consistency**

### Sample Error

We can define sample error as the difference between the actual value of the estimate $a(S)$ and the quantity being estimated (or the true value $a(P)$).

$$SampleError = a(S) - a(P)$$

Let's look at example. We will look at the differences between some attributes. The sample we will use here would be of size $N = 40$.

```
temp <- read.csv("temperature.csv", header = TRUE)
set.seed(341)
temp.jan <- temp$JAN
s = sample(length(temp$JAN), 40)
c(mean(temp.jan[s]) - mean(temp.jan), median(temp.jan[s]) - median(temp.jan),
sd(temp.jan[s]) - sd(temp.jan), IQR(temp.jan[s]) - IQR(temp.jan))
```

```
## [1] 0.05957590 0.04300000 0.01524549 0.05325000
```

We prefer an attribute with low sampling errors.

### Fisher Consistency

Ronald A. Fisher identified a consistency that, when the sample reaches the size of the population, the sample error approaches zero/non-existent. This consistency is known as Fisher consistency.

Now, we will look at the same differences between the attributes as shown above but with different sample sizes

```
s <-sample(length(temp.jan),60)
c(mean(temp.jan[s]) - mean(temp.jan), median(temp.jan[s]) - median(temp.jan),
sd(temp.jan[s]) - sd(temp.jan), IQR(temp.jan[s]) - IQR(temp.jan))
```

```
## [1]  0.0220925703  0.0190000000  0.0003452442 -0.0070000000
```

```
s <-sample(length(temp.jan),166)
c(mean(temp.jan[s]) - mean(temp.jan), median(temp.jan[s]) - median(temp.jan),
sd(temp.jan[s]) - sd(temp.jan), IQR(temp.jan[s]) - IQR(temp.jan))
```

```
## [1] 0 0 0 0
```

We observe from the above results that as the sample size $N$ increases, the difference between the sample estimate and the population decreases. Hence as Sample Size -> Population, the sample error -> 0

### All possible samples

In a population $P$ of size $N$ and a sample of size $n$, the number of different possible samples $S$ are $\binom{N}{n}$

Hence, we can calculate the attribute for each possible sample. Let's try it out in the following example. We will find the average length of sharks of all possible samples.

```
sharks <- read.csv("sharks.csv")
popSharks <- rownames(sharks)
sharks = na.omit(sharks)
popSharksAus <- popSharks[sharks$Australia == 1]
samples <- combn(popSharksAus, 5)
m <- ncol(samples)
avg.Samples <- apply(samples, MARGIN=2, FUN=function(s) {
mean(sharks[s, "Length"])
})
# Printing the average in the first 10 samples, and the last sample
avg.Samples[c(1:10, m)]
```

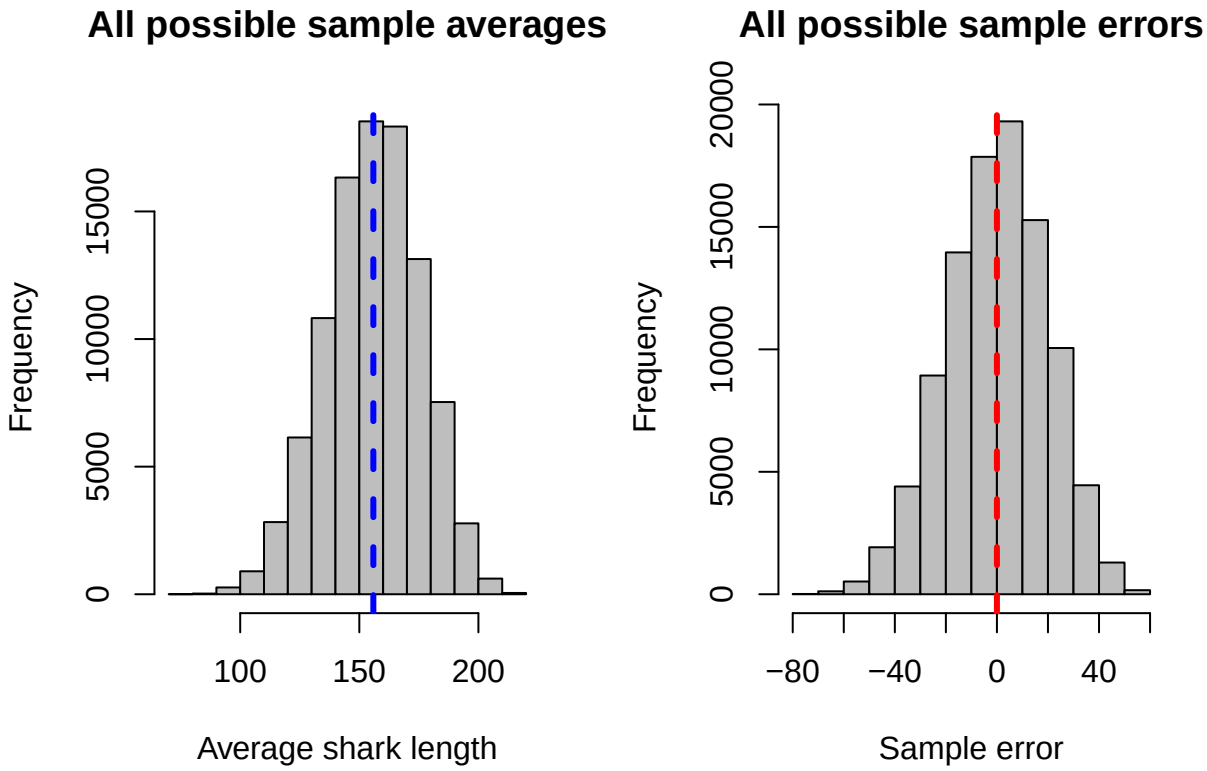
```
##  [1] 142.6 146.6 129.8 142.2 142.2 161.8 154.0 158.0 156.6 139.4 196.8
```

Let us also look at the sample error for all possible samples

```
## [1] -13.292857  -9.292857 -26.092857 -13.692857 -13.692857  40.907143
```

Now, let's look at it graphically.

```
par(mfcol=c(1,2))
av <- mean(sharks[popSharksAus,"Length"])
# Histogram of sample mean length
hist(avg.Samples, xlab="Average shark length", main="All possible sample averages", col="grey")
abline(v=av, col="blue", lwd=3, lty=2)
# Histogram of sample errors
hist(sampleErrors, xlab="Sample error", main="All possible sample errors", col="grey")
abline(v=0, col="red", lwd=3,lty=2)
```




**Consistency & Effect of Sample Size**

As the sample size increases, the sample approaches the population and attribute values will concentrate around the true value. We can quanity this concentration by,

$$|a(S) - a(P)| = \left| \frac{1}{n} \sum_{u \in S} y_u - \frac{1}{N} \sum_{u \in P} y_u \right| < C, \ \ for \ c > 0$$

The above stands for the absolute difference between the sample attribute and the population attribute.

Now, for each n, the possible set of samples will be

$$P_S(n) = \{S : S \subset P \ and \ |S| = n\}$$

For any $c > 0$,

$$P_a(c, n) = \{S : S \subset P_S(n) \ and \ |a(S) - a(P)| < c\}$$

and we define the proportion as

$$P_a(c, n) = \frac{|P_a(c, n)|}{|P_S(n)|}$$