# A3Q2

## Summary:

In real World, with the explosion of data, deriving a complete analysis of a large population seems to be impossible. That's why **Samples** come into our mind. We take a subset $\mathcal{S}$ of the population $\mathcal{P}$ and estimate the attribute of the population based on the sample. So, we have $a(\mathcal{S})=\widehat{a(\mathcal{P})}=a(\hat{\mathcal{P}})$.

If we have a sample with size $n$ and a population with size $N$, then:

- $SampleError = a(\mathcal{S}) - a(\mathcal{P})$
- Fisher Consistency: As $n$ approaches $N$, the sample error should get closer to zero.

There are many ways to select a sample of size $n$ from population $N$, actually, we can find $\binom{N}{n}$ possible samples $\mathcal{S}$ of size $n$. For those samples, we cancalculate their sample errors and average sample error using formulas below:

$$Sample\ Error: a(\mathcal{S}) - a(\mathcal{P}) = \frac{1}{n}\sum_{u\in\mathcal{S}} y_u - \frac{1}{N}\sum_{u\in\mathcal{P}} y_u$$

$$Average\ Sample\ Error = \frac{1}{N_{\mathcal{S}}}\sum_{i=1}^{N_{\mathcal{S}}}[a(\mathcal{S}_i) - a(\mathcal{P})]$$

Size of the sample has a huge impact on the sample error. We expect that with sample size increases, the attribute should concentrate around the true value, which matches the Fisher consistency. To measure the consistency, we can check if the absolute value of the difference of sample attribute and population attribute is within a certain sample error:

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

## Example:

we can load the temperature.csv data from course website and find all possible samples for first 5 january temperatures:

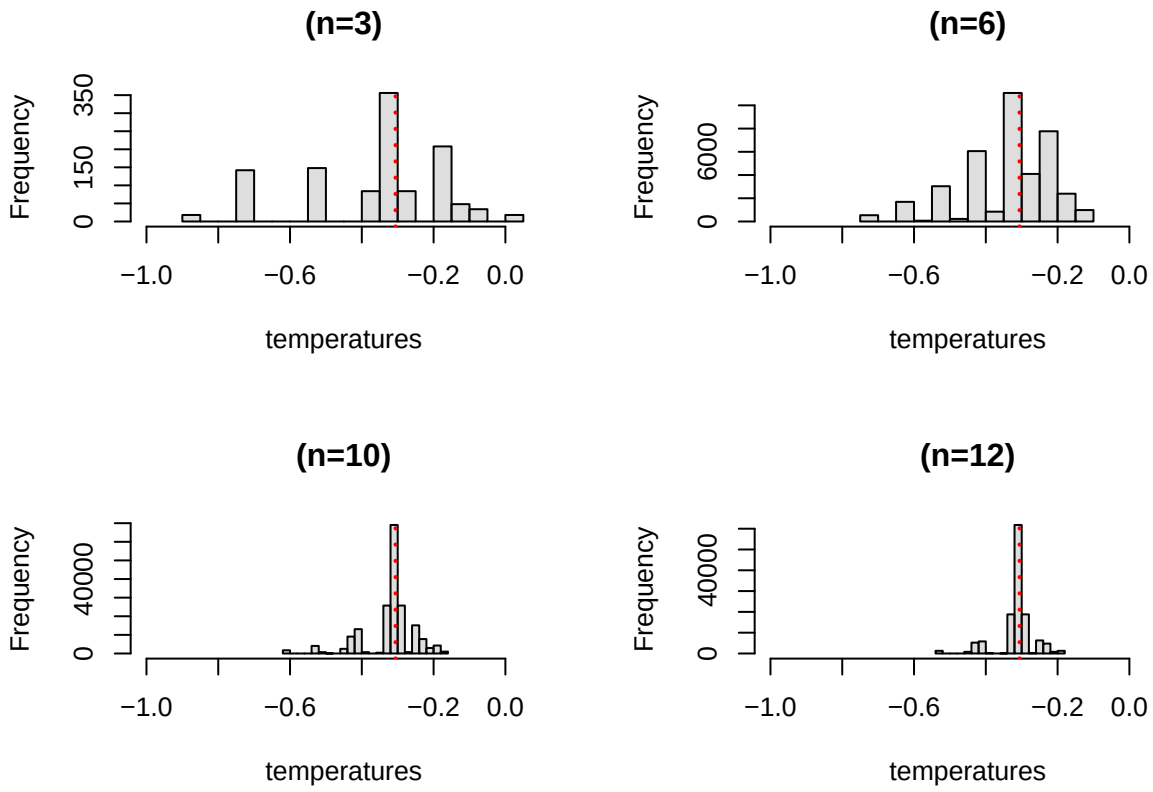
```
temperature <- read.csv('temperature.csv',header=T)
combn(temperature$JAN[1:5],2)
```

```
##        [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]   [,8]   [,9]  [,10]
## [1,] -0.702 -0.702 -0.702 -0.702 -0.303 -0.303 -0.303 -0.308 -0.308 -0.177
## [2,] -0.303 -0.308 -0.177 -0.360 -0.308 -0.177 -0.360 -0.177 -0.360 -0.360
```

As we can see, there are 10 samples generated, which matches the number of $\binom{5}{2}$.

we can use first 20 January data and sample sizes 3, 5, 10, 12, we are going to plot sample median:

```r
slice <- temperature$JAN[0:20]
Plot.Median = function(attr,pop=slice){
  indx=1:length(pop)
  n=c(3, 6, 10, 12)
  par(mfrow=c(2,2))
  for(i in 1:4){
    #generating all possible samples
    samples <- combn(indx, n[i])
    popMedian <- attr(pop)
    sampMedian <- apply(samples, MARGIN = 2, FUN = function(s){attr(pop[s])})
    hist(sampMedian, col=adjustcolor("grey", alpha = 0.5),
       main=paste("(n=",as.character(n[i]),")",sep=''),
       xlab="temperatures",breaks=20,xlim=c(-1,0.1))
  abline(v=popMedian, col="red", lty=3, lwd=2)
  }
}
Plot.Median(median)
```




**Comment:** From the data, we can see that most of medians for each graph are close to the true median, despite that there are more outliers for top 2 graphs then bottom 2 graphs. We can there are gaps between each bar as median can be one of the observed value in the data. From those four graphs, there's a clear trend that with n increases, the medians are getting closer and closer to the true median and the graphs are more symmetric, which means that more medians are concentrated around the true value. This matches the definition of Fisher consistency. As n gets larger, the corresponding sample error reduces.