Inferential Statistics: inference of the population mean

Inferential Statistics: inference of the population mean

Sample mean and Sample variance

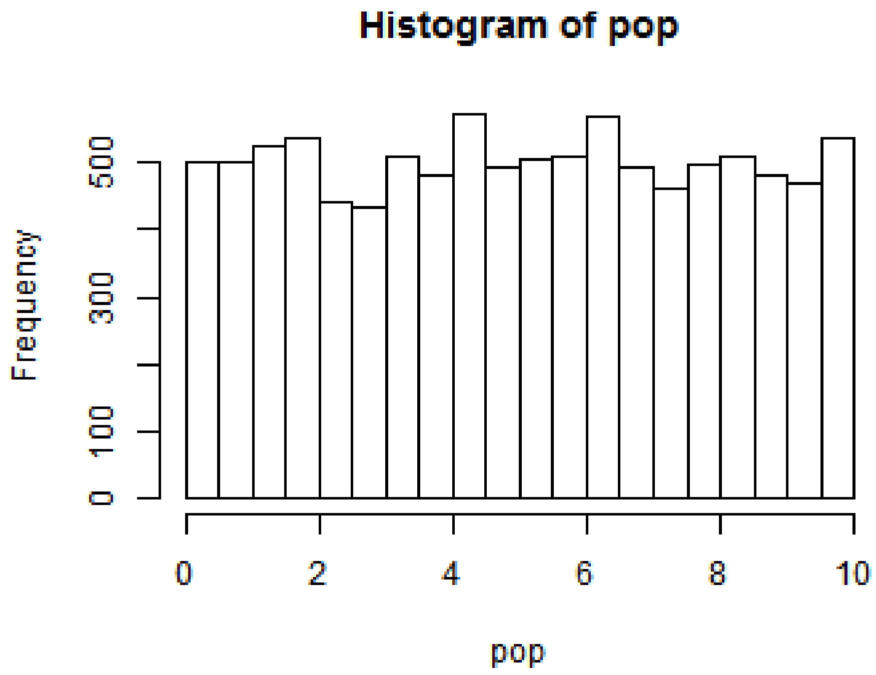
Given an unknown population I need to estimate some statistics but I can't afford in analyzing every member of the population, hence I get a random sample from the original population and I calculate statistics on that sample. The estimated variable (is this case the mean) is itself a random variable. Thanks to the Central Limit Theorem (CLT) whatever the original population distribution is, it states that the sampling distribution, is a normal distribution with a mean and a variance, where:

- Mean[sampling distribution] = population mean
- Standard error = population sd / sqrt(n)
  - or if the population sd is unknown is s / sqrt(n) where s is the sample standard deviation.

Thus the larger the sample size, the smaller the variance of the sampling distribution of the mean is.

The standard error of the mean is the standard deviation of the sampling distribution of the mean.

```
a <- 0
b <- 10
pop <- runif(n=10000, min=a, max=b) # pop is the original population
                                    # obtained by a Uniform r.v. [a, b]
hist(pop)
```



```
pop.mean <- (a + b)/2
pop.mean

## [1] 5

pop.sd <- sqrt(1/12*(b - a)^2)
pop.sd

## [1] 2.887
```

Now we pick up a random sample from the original population in order to estimate sample statistics.

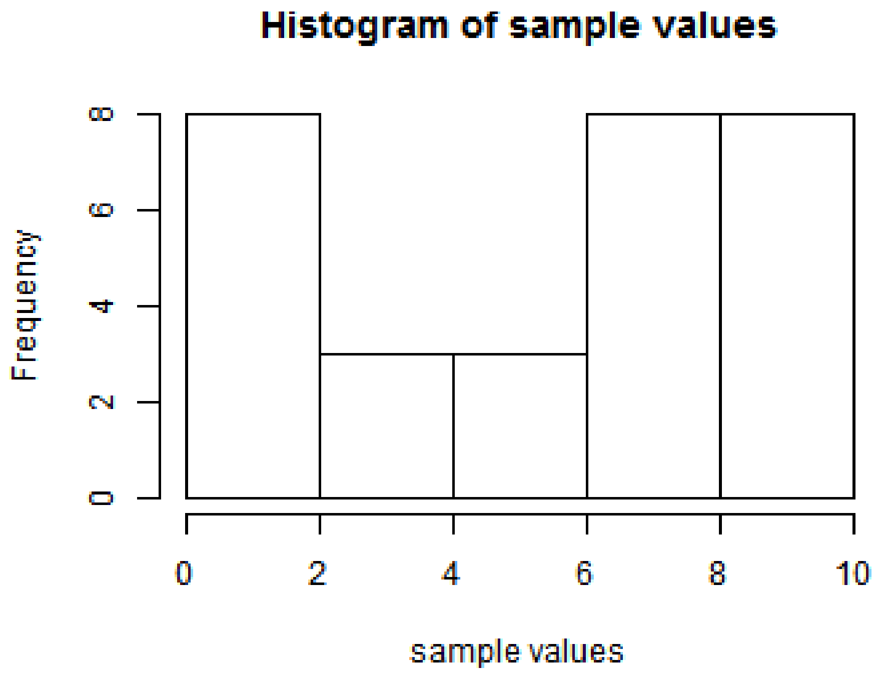
```
# sam is the sampling distribution
sam.size <- 30
sam <- sample(x=pop, size=sam.size, replace=FALSE)
sam.mean <- mean(sam)
sam.mean

## [1] 5.347

s <- sd(sam)
s # is the sample standard deviation

## [1] 3.402

hist(sam, main="Histogram of sample values", xlab = "sample values")
```



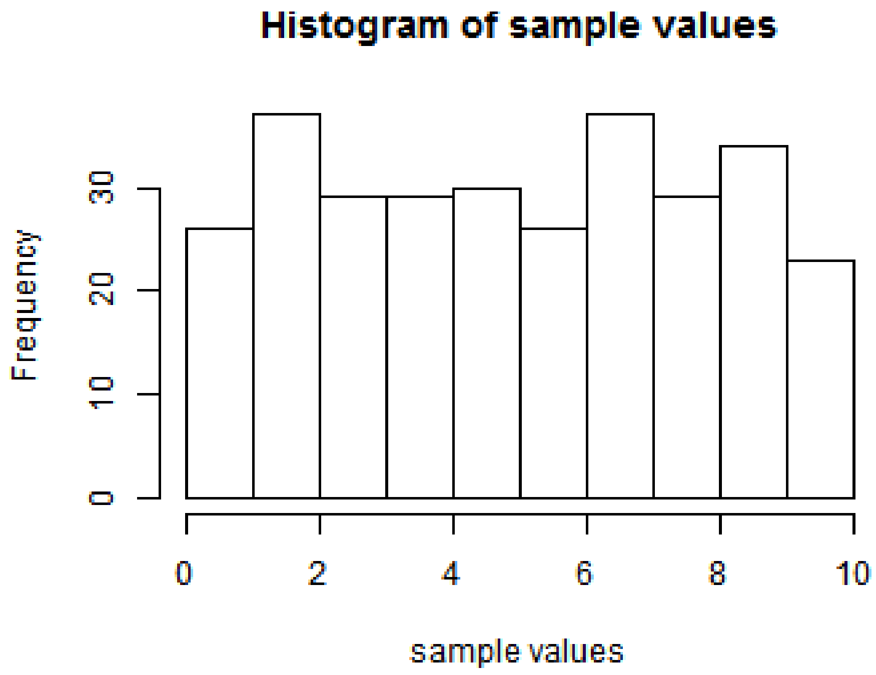
The sample mean statistic is a Normal random variable, the 95% confidence interval of the sample statistic is:

```
ll <- sam.mean - qnorm(0.025, lower.tail = FALSE)*s/sqrt(sam.size)
ul <- sam.mean + qnorm(0.025, lower.tail = FALSE)*s/sqrt(sam.size)
```

The sample mean lies between [4.13, 6.5648] with a 95% of probability. The sample mean obtained by a sample of size n=30 is 5.3474 with a standard error of 0.6211

If the sample size is getting higher such as 300, the confidence interval will be narrower than before.

```
## [1] 2.825
```

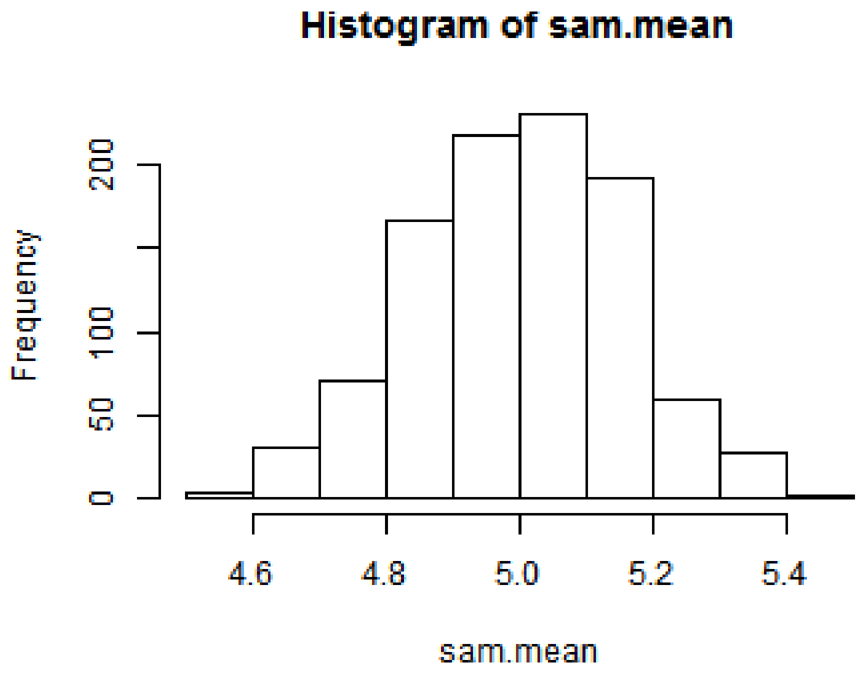


The sample mean lies between [4.6161, 5.2554] with a 95% of probability. The sample mean obtained by a sample of size n=300 is 4.9357 with a standard error of 0.1631

Histogram of the sample mean distribution

In order to obtained the distribution of the sample mean, we need to extract many random samples from the population and calculate the sample mean for each one. If we run many trials we can calculate the mean of the sample mean distribution.

```
trials <- 1000
sam.size <- 300
sam.mat <- matrix(NA, nrow = trials, ncol = sam.size)
# each row of sam.mat will contain a sample of the population values (sam.size)
sam.mat <- t(sapply(1:trials, FUN=function(i) sample(x=pop, size=sam.size, replace=F)))
sam.mean <- rowMeans(sam.mat)
hist(sam.mean)
```



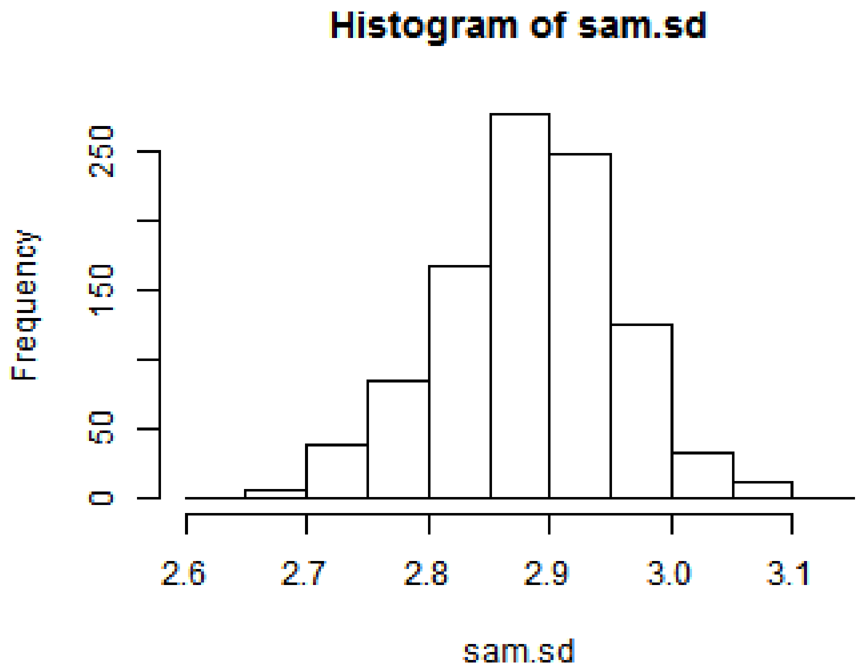
```
sam.sd <- apply(sam.mat, 1, FUN=sd)
```

With 1000 trials the 95% confidence interval is [4.6739, 5.3266]. The sample mean is 5.0003 with a standard error of 0.1665

Histogram of the sample variance distribution

In a similar way we can obtain the distribution of the sample variance (in this chart the sample standard deviation)

```
hist(sam.sd)
```



With 1000 trials the 95% confidence interval is [2.8754, 2.892]. The sample standard deviation is 2.8837 with a standard error of 0.0042

T Confidence Intervals

When population variance is unknown or the sample size is small (< 30) in order to detect a confidence interval, a better choice is to use a Student-t distribution to model the sample distribution. The confidence interval will be larger than those detected through a Normal distribution. For larger sample size, z and t statistics will provide the same reliability factor so we can default to the standard normal distribution and z-statistic.

For alpha of 5% (i.e. a 95% confidence interval), the reliability factor (ZÎ±/2) is 1.96. Given a sample size of 16, a sample mean of 20 and population standard deviation of 25, a 95% confidence interval would be:

```
20 + c(-1, 1)* 1.96*(25/sqrt(16))
```

```
## [1]  7.75 32.25
```

In short, for this sample size and for these sample statistics, we would be 95% confident that the actual population mean would fall in a range from 7.75 to 32.25. But a **more conservative approach** is to proceed with the construction of the confidence interval through a **Student-t distribution**. We must first calculate degrees of freedom, which for sample size 16 is equal to n - 1 = 15. Using an alpha of 5% (95% confidence interval), our confidence interval is

```
20 + c(-1, 1) * qt(p = 0.025, df = 15, lower.tail=F) * (25/sqrt(16))
```

```
## [1]  6.678 33.322
```

which gives a range minimum of 6.678 and a range maximum of 33.32.

As before, we can reduce this range with

1. larger samples and/or
2. reducing allowable degree of confidence