Lab 01

Jyosna Philip

2023-11-20

Sampling Distribution and Standard Error

AIM: To obtain the sampling distribution and the Standard Error

Dataset Description:

The data gives the speed of cars and the distance taken to stop. Note that the data were recorded in the 1920s.

ANALYSIS

```
#dataset
cars
```

```
##    speed dist
## 1      4    2
## 2      4   10
## 3      7    4
## 4      7   22
## 5      8   16
## 6      9   10
## 7     10   18
## 8     10   26
## 9     10   34
## 10    11   17
## 11    11   28
## 12    12   14
## 13    12   20
## 14    12   24
## 15    12   28
## 16    13   26
## 17    13   34
## 18    13   34
## 19    13   46
## 20    14   26
## 21    14   36
## 22    14   60
## 23    14   80
## 24    15   20
## 25    15   26
```

```
## 26    15   54
## 27    16   32
## 28    16   40
## 29    17   32
## 30    17   40
## 31    17   50
## 32    18   42
## 33    18   56
## 34    18   76
## 35    18   84
## 36    19   36
## 37    19   46
## 38    19   68
## 39    20   32
## 40    20   48
## 41    20   52
## 42    20   56
## 43    20   64
## 44    22   66
## 45    23   54
## 46    24   70
## 47    24   92
## 48    24   93
## 49    24  120
## 50    25   85
```

```
dim(cars)
```

```
## [1] 50  2
```

Hence the cars dataset has 50 observations and 2 variables.

```
head(cars) #to print first six observation
```

```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
## 4     7   22
## 5     8   16
## 6     9   10
```

```
tail(cars) # get last six observations
```

```
##    speed dist
## 45    23   54
## 46    24   70
## 47    24   92
## 48    24   93
## 49    24  120
## 50    25   85
```

```r
head(cars)
```

```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
## 4     7   22
## 5     8   16
## 6     9   10
```

```r
tail(cars)
```

```
##    speed dist
## 45    23   54
## 46    24   70
## 47    24   92
## 48    24   93
## 49    24  120
## 50    25   85
```

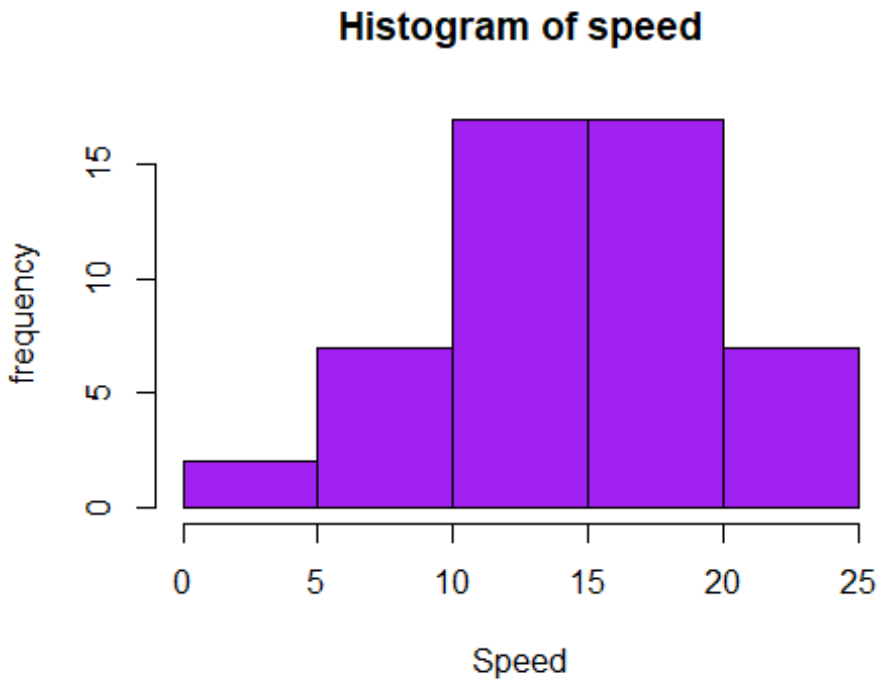
DESCRIPTIVE STATISTICS

```r
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Therefore, the variable speed varies from 4.0 to 25.0. Median speed is 15.0. There is equal number of values above and below median.

```r
# To find the distribution of speed
hist(cars$speed,xlab = "Speed",ylab = "frequency", main = "Histogram of
speed", col="purple")
```




```
#finding standard deviation
sd(cars$speed)
```

```
## [1] 5.287644
```

A standard deviation close to zero indicates that data points are close to the mean. Here,the standard deviation obtained is 5.28 which indicates that the data points are not close to mean.

POPULATION

```
population = cars$speed
population
```

```
##  [1]  4  4  7  7  8  9 10 10 10 11 11 12 12 12 12 13 13 13 13 14 14 14 14
15 15
## [26] 15 16 16 17 17 17 18 18 18 18 19 19 19 20 20 20 20 20 22 23 24 24 24
24 25
```

SAMPLE OF SIZE=10

```
samplesize1=10
#Choosing a sample of size 10 from the population using simple random
sampling with replacement technique
s1=sample(population , samplesize1, replace = TRUE)
s1
```

```
##  [1]  4 17 11 15  7 14 13 24 12 13
```

```
mean(s1)

## [1] 13
```

Mean of sample 1 is 13

```
sd(s1)

## [1] 5.416026
```

Standard deviation of sample 1 is 5.42

STANDARD ERROR OF SAMPLE1

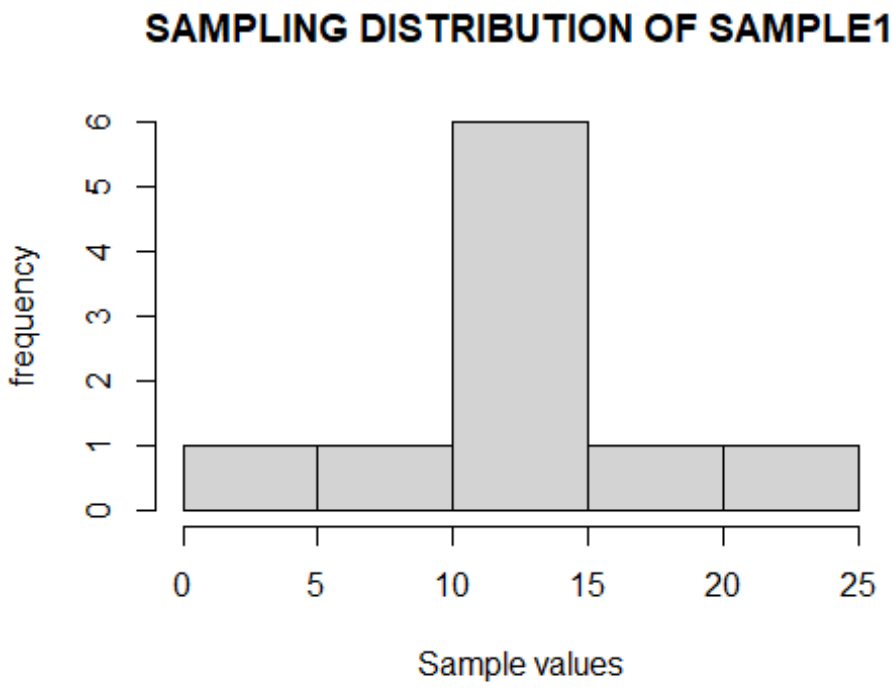
```
print(sd(s1)/sqrt(samplesize1)) #gets standard error

## [1] 1.712698

The standard error of sample 1 is 1.71 which is high.
```

SAMPLING DISTRIBUTION OF SAMPLE 1

```
hist(s1,xlab="Sample values",ylab="frequency",main="SAMPLING DISTRIBUTION OF
SAMPLE1")
```



We can observe that this graph doesn't give a great idea about how the sample is distributed, so we increase the sample size to 15.

SAMPLE OF SIZE=15

```
#choosing 15 observations in the sample
samplesize2=15
#sample
#Choosing a sample of size 15 from the population using simple random
sampling with replacement technique
s2=sample(population , samplesize2, replace = TRUE)
s2
```

```
##  [1] 12 16 12  7 19 10 17 18 13 14 22  7 11 24 11
```

```
mean(s2) #to get mean
```

```
## [1] 14.2
```

```
sd(s2)
```

```
## [1] 5.059644
```

The mean of sample 2 is 14.2 and Standard deviation of sample 2 is 5.05 which is not close to 0.

STANDARD ERROR

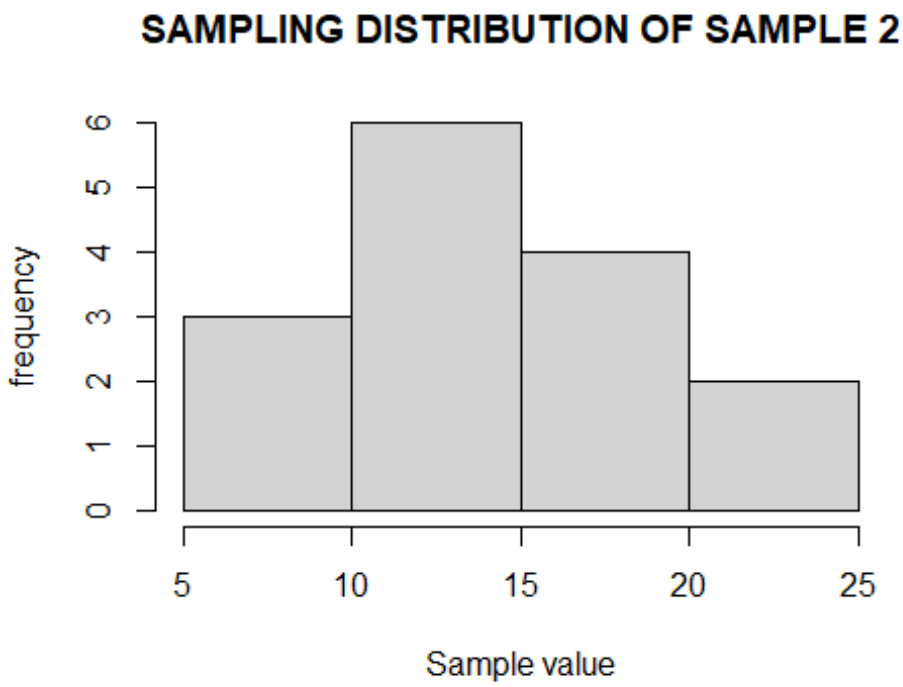
```
print(sd(s2)/sqrt(samplesize2)) #gets standard error
```

```
## [1] 1.306395
```

The standard error is 1.31.

SAMPLING DISTRIBUTION OF SAMPLE 2

```
#Find the sampling distribution of sample 2
hist(s2, xlab="Sample value", ylab= "frequency", main="SAMPLING DISTRIBUTION
OF SAMPLE 2")
```



We can observe that this graph also doesn't give a great idea about how the sample is distributed, so we use replicate function to replicate the statistic.

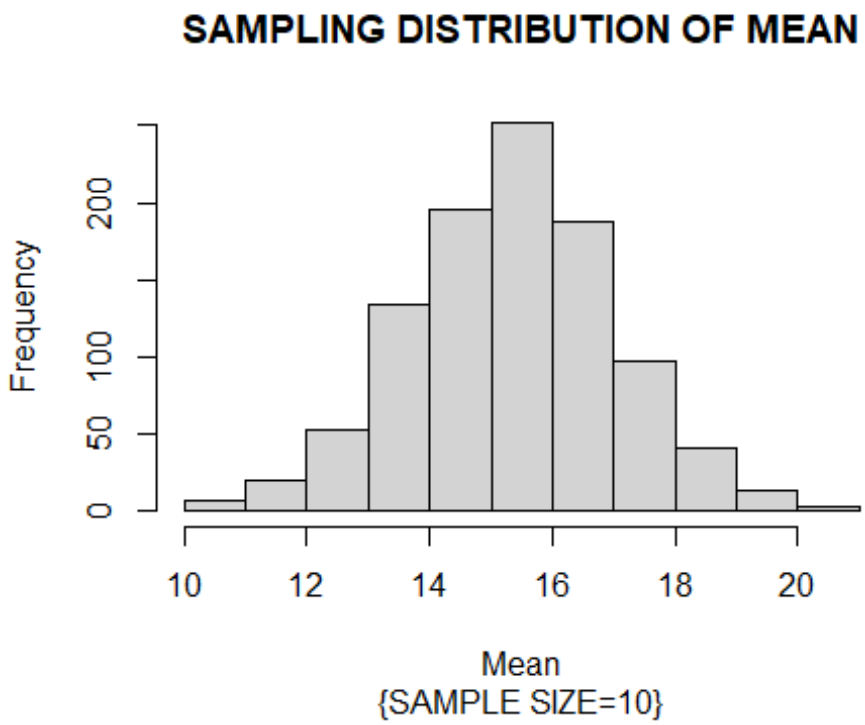
SAMPLING DISTRIBUTION OF MEAN USING REPLICATE() FUNCTION

SAMPLE SIZE=10

```
#replicate() function in R is used to evaluate an expression N number of
times repeatedly.
#Here it is replicated 1000 times.
samp_dist1=replicate(1000,mean(sample(cars$speed,10,replace=TRUE)))
head(samp_dist1)

## [1] 15.4 14.9 15.8 15.7 14.8 16.3

#Finding sampling distribution using histogram
hist(samp_dist1, xlab="Mean", ylab= "Frequency", main="SAMPLING DISTRIBUTION
OF MEAN",sub="{SAMPLE SIZE=10}")
```




```
#VARIANCE
var(samp_dist1)

## [1] 2.80733

#STANDARD ERROR
sd(samp_dist1)

## [1] 1.675509
```

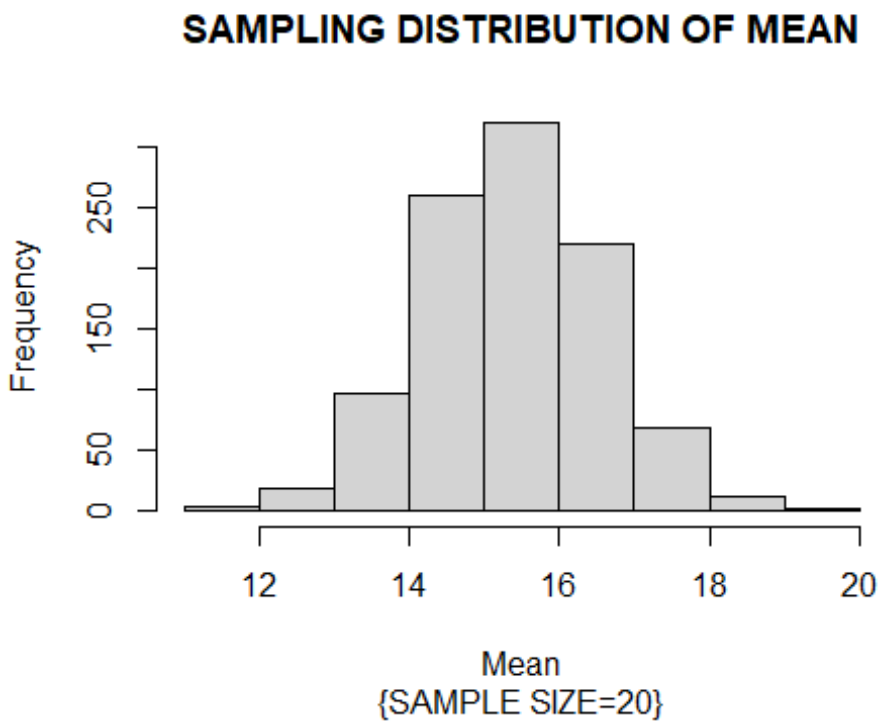
The variance and standard error of sampling distribution of mean is 2.8 and 1.67 respectively.

SAMPLE SIZE=20

```
samp_dist2=replicate(1000,mean(sample(cars$speed,20,replace=TRUE)))
head(samp_dist2)

## [1] 13.95 15.05 16.75 15.20 14.00 15.45

#Finding sampling distribution using histogram
hist(samp_dist2, xlab="Mean", ylab= "Frequency", main="SAMPLING DISTRIBUTION
OF MEAN",sub="{SAMPLE SIZE=20}")
```




```
#VARIANCE
var(samp_dist2)

## [1] 1.358157

#STANDARD ERROR
sd(samp_dist2)

## [1] 1.1654
```

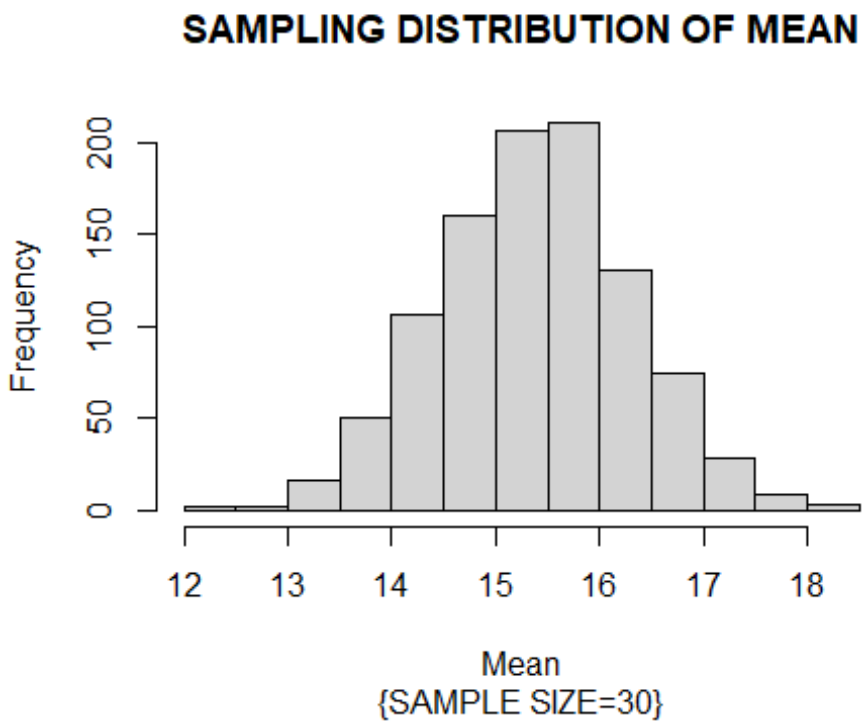
The variance and standard error are respectively 1.35 and 1.16. This is less than the values obtained for sample size=10.

SAMPLE SIZE=30

```
samp_dist3=replicate(1000,mean(sample(cars$speed,30,replace=TRUE)))
head(samp_dist3)

## [1] 15.83333 15.60000 13.90000 15.56667 14.36667 17.33333

#Finding sampling distribution using histogram
hist(samp_dist3, xlab="Mean", ylab= "Frequency", main="SAMPLING DISTRIBUTION
OF MEAN",sub="{SAMPLE SIZE=30}")
```




```
#VARIANCE
var(samp_dist3)

## [1] 0.8679117

#STANDARD ERROR
sd(samp_dist3)

## [1] 0.9316178
```

The variance and standard error are respectively 0.86 and 0.93. This is less than the values obtained for sample size=20. Hence, it is clear from the result that as the sample size increases the variance decreases. It is known that the variance is inversely proportional to the precision. Hence, it can be concluded that precision increases as sample size increases.

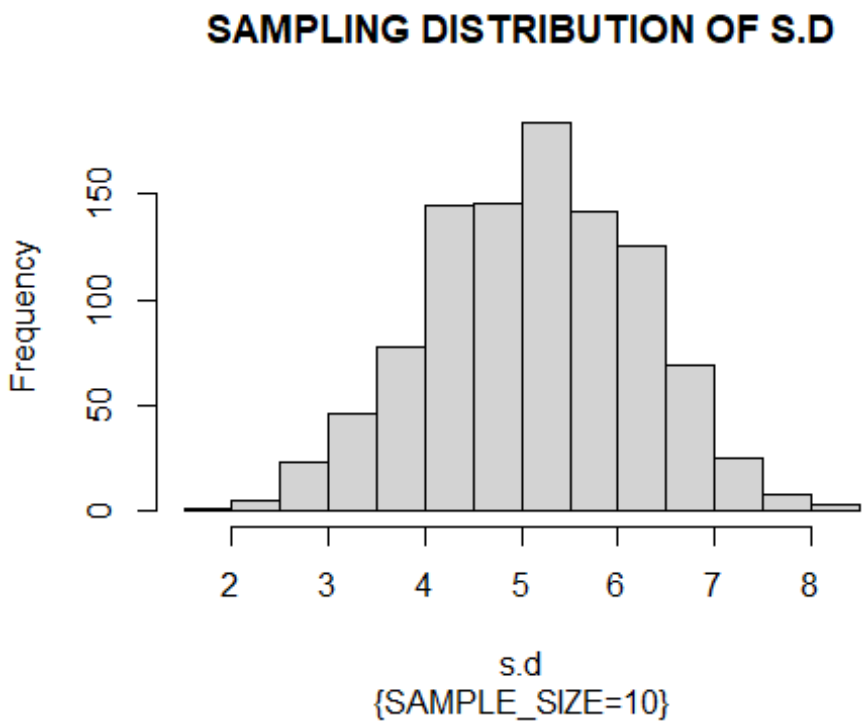
SAMPLING DISTRIBUTION OF STANDARD DEVIATION

SAMPLE SIZE=10

```
samp_dist_sd1=replicate(1000,sd(sample(cars$speed,10,replace=TRUE)))
head(samp_dist_sd1)

## [1] 5.334375 4.289522 6.415260 5.108816 5.275731 5.557777

#histogram of samp_dist
hist(samp_dist_sd1,xlab="s.d",ylab="Frequency",main="SAMPLING DISTRIBUTION OF
S.D",sub="{SAMPLE_SIZE=10}")
```




```
#VARIANCE
var(samp_dist_sd1)

## [1] 1.193303

#STANDARD ERROR
sd(samp_dist_sd1)

## [1] 1.092384
```

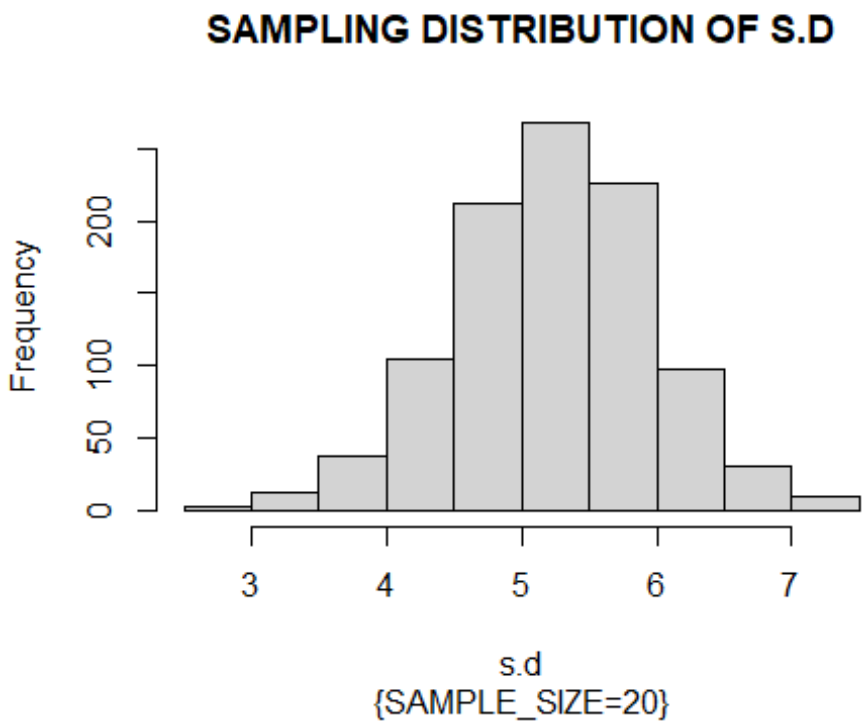
The variance and standard error are 1.19 and 1.09 respectively.

SAMPLE SIZE=20

```
samp_dist_sd2=replicate(1000,sd(sample(cars$speed,20,replace=TRUE)))
head(samp_dist_sd2)

## [1] 6.781360 4.221187 4.758372 4.168806 6.074970 4.840618

#histogram of samp_dist
hist(samp_dist_sd2,xlab="s.d",ylab="Frequency",main="SAMPLING DISTRIBUTION OF
S.D",sub="{SAMPLE_SIZE=20}")
```




```
#VARIANCE
var(samp_dist_sd2)

## [1] 0.5350735

#STANDARD ERROR
sd(samp_dist_sd2)

## [1] 0.7314872
```

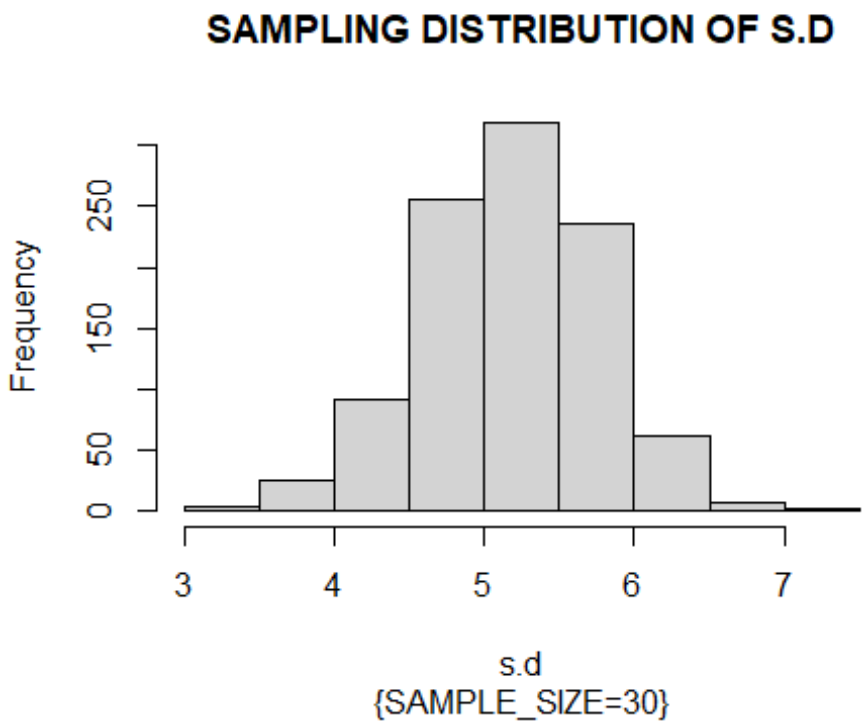
The variance and standard deviation are 0.5 and 0.73 respectively. This is less than the values obtained for sample size=10.

SAMPLE SIZE=30

```
samp_dist_sd3=replicate(1000,sd(sample(cars$speed,30,replace=TRUE)))
head(samp_dist_sd3)

## [1] 4.231926 4.806126 5.424932 5.481840 5.422813 4.695437

#histogram of samp_dist
hist(samp_dist_sd3,xlab="s.d",ylab="Frequency",main="SAMPLING DISTRIBUTION OF
S.D",sub="{SAMPLE_SIZE=30}")
```




```
#VARIANCE
var(samp_dist_sd3)

## [1] 0.3446647

#STANDARD ERROR
sd(samp_dist_sd3)

## [1] 0.5870815
```

The variance and standard deviation are respectively. This is less than the values obtained for sample size=20. Therefore, the variance of statistic(standard deviation) reduces with increase in sample size. Hence, it is clear from the result that as the sample size increases the variance decreases. It is known that the variance is inversely proportional to the precision. Hence, it can be concluded that precision increases as sample size increases.

CONCLUSION

The variance and standard error of the sampling distribution of both mean and standard deviation of the target variable reduced when the sample size increased. Hence an increase in sample size increases precision.