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 response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, orange juice or ascorbic acid (a form of vitamin C and coded as VC).

# ANALYSIS

#dataset  
ToothGrowth

## len supp dose  
## 1 4.2 VC 0.5  
## 2 11.5 VC 0.5  
## 3 7.3 VC 0.5  
## 4 5.8 VC 0.5  
## 5 6.4 VC 0.5  
## 6 10.0 VC 0.5  
## 7 11.2 VC 0.5  
## 8 11.2 VC 0.5  
## 9 5.2 VC 0.5  
## 10 7.0 VC 0.5  
## 11 16.5 VC 1.0  
## 12 16.5 VC 1.0  
## 13 15.2 VC 1.0  
## 14 17.3 VC 1.0  
## 15 22.5 VC 1.0  
## 16 17.3 VC 1.0  
## 17 13.6 VC 1.0  
## 18 14.5 VC 1.0  
## 19 18.8 VC 1.0  
## 20 15.5 VC 1.0  
## 21 23.6 VC 2.0  
## 22 18.5 VC 2.0  
## 23 33.9 VC 2.0  
## 24 25.5 VC 2.0  
## 25 26.4 VC 2.0  
## 26 32.5 VC 2.0  
## 27 26.7 VC 2.0  
## 28 21.5 VC 2.0  
## 29 23.3 VC 2.0  
## 30 29.5 VC 2.0  
## 31 15.2 OJ 0.5  
## 32 21.5 OJ 0.5  
## 33 17.6 OJ 0.5  
## 34 9.7 OJ 0.5  
## 35 14.5 OJ 0.5  
## 36 10.0 OJ 0.5  
## 37 8.2 OJ 0.5  
## 38 9.4 OJ 0.5  
## 39 16.5 OJ 0.5  
## 40 9.7 OJ 0.5  
## 41 19.7 OJ 1.0  
## 42 23.3 OJ 1.0  
## 43 23.6 OJ 1.0  
## 44 26.4 OJ 1.0  
## 45 20.0 OJ 1.0  
## 46 25.2 OJ 1.0  
## 47 25.8 OJ 1.0  
## 48 21.2 OJ 1.0  
## 49 14.5 OJ 1.0  
## 50 27.3 OJ 1.0  
## 51 25.5 OJ 2.0  
## 52 26.4 OJ 2.0  
## 53 22.4 OJ 2.0  
## 54 24.5 OJ 2.0  
## 55 24.8 OJ 2.0  
## 56 30.9 OJ 2.0  
## 57 26.4 OJ 2.0  
## 58 27.3 OJ 2.0  
## 59 29.4 OJ 2.0  
## 60 23.0 OJ 2.0

dim(ToothGrowth)

## [1] 60 3

Hence the ToothGrowth dataset has 60 observations and 3 variables.

head(ToothGrowth) #to print first six observation

## len supp dose  
## 1 4.2 VC 0.5  
## 2 11.5 VC 0.5  
## 3 7.3 VC 0.5  
## 4 5.8 VC 0.5  
## 5 6.4 VC 0.5  
## 6 10.0 VC 0.5

tail(ToothGrowth) # get last six observations

## len supp dose  
## 55 24.8 OJ 2  
## 56 30.9 OJ 2  
## 57 26.4 OJ 2  
## 58 27.3 OJ 2  
## 59 29.4 OJ 2  
## 60 23.0 OJ 2

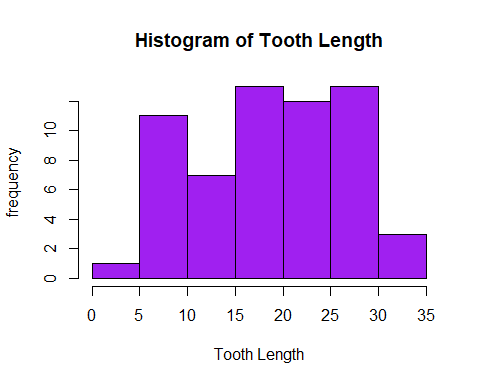
# DESCRIPTIVE STATISTICS

summary(ToothGrowth)

## len supp dose   
## Min. : 4.20 OJ:30 Min. :0.500   
## 1st Qu.:13.07 VC:30 1st Qu.:0.500   
## Median :19.25 Median :1.000   
## Mean :18.81 Mean :1.167   
## 3rd Qu.:25.27 3rd Qu.:2.000   
## Max. :33.90 Max. :2.000

Therefore, the variable len varies from 4.20 to 33.9. Median length is 19.25.

# To find the distribution of length  
hist(ToothGrowth$len,xlab = "Tooth Length",ylab = "frequency", main = "Histogram of Tooth Length", col="purple")



#finding standard deviation  
sd(ToothGrowth$len)

## [1] 7.649315

A standard deviation close to zero indicates that data points are close to the mean. Here,the standard deviation obtained is 7.64 which indicates that the data points are not close to mean.

# POPULATION

population = ToothGrowth$len  
population

## [1] 4.2 11.5 7.3 5.8 6.4 10.0 11.2 11.2 5.2 7.0 16.5 16.5 15.2 17.3 22.5  
## [16] 17.3 13.6 14.5 18.8 15.5 23.6 18.5 33.9 25.5 26.4 32.5 26.7 21.5 23.3 29.5  
## [31] 15.2 21.5 17.6 9.7 14.5 10.0 8.2 9.4 16.5 9.7 19.7 23.3 23.6 26.4 20.0  
## [46] 25.2 25.8 21.2 14.5 27.3 25.5 26.4 22.4 24.5 24.8 30.9 26.4 27.3 29.4 23.0

## 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] 7.0 9.7 23.3 25.5 17.3 20.0 26.4 24.8 17.6 26.4

mean(s1)

## [1] 19.8

Mean of sample 1 is 19.8

sd(s1)

## [1] 6.942942

Standard deviation of sample 1 is 6.94

### STANDARD ERROR OF SAMPLE1

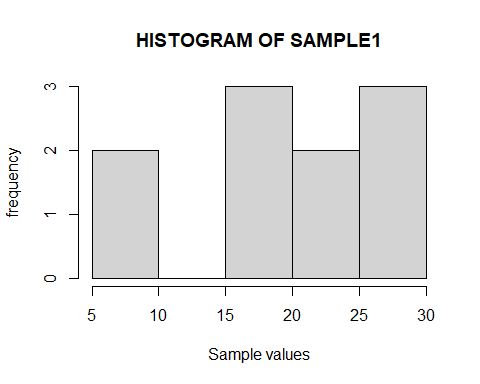
print(sd(s1)/sqrt(samplesize1)) #gets standard error

## [1] 2.195551

The standard error of sample1 is 2.19, which is high.

### SAMPLING DISTRIBUTION OF SAMPLE 1

hist(s1,xlab="Sample values",ylab="frequency",main="HISTOGRAM 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] 23.6 29.4 26.4 26.4 9.4 27.3 24.5 32.5 26.4 26.4 6.4 22.4 27.3 20.0 5.8

mean(s2) #to get mean

## [1] 22.28

sd(s2)

## [1] 8.345161

Mean of sample is 22.28 and Standard deviation of sample 2 is 8.34 which is not close to 0.

## STANDARD ERROR

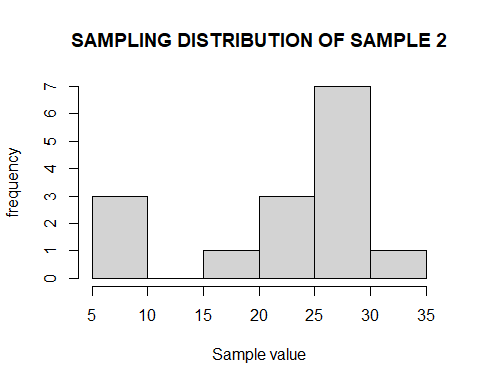
print(sd(s2)/sqrt(samplesize2)) #gets standard error

## [1] 2.154711

The standard error is 2.15

### 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.

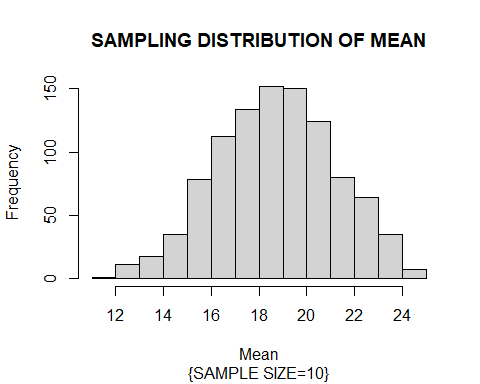
# SAMPLING DISTRIBUTION OF MEAN USING REPLICATE() FUNCTION

## 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(ToothGrowth$len,10,replace=TRUE)))  
head(samp\_dist1)

## [1] 20.85 18.63 18.76 19.36 19.82 18.93

#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] 6.110182

#STANDARD ERROR  
sd(samp\_dist1)

## [1] 2.471878

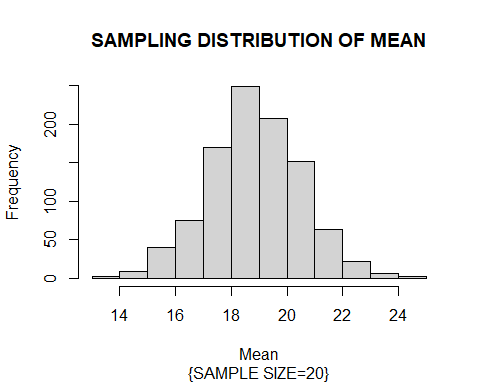
The variance and sampling distribution of mean is 6.11 and 2.47 respectively.

## SAMPLE SIZE=20

samp\_dist2=replicate(1000,mean(sample(ToothGrowth$len,20,replace=TRUE)))  
head(samp\_dist2)

## [1] 20.160 15.965 18.515 20.670 21.280 14.090

#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] 2.813825

#STANDARD ERROR  
sd(samp\_dist2)

## [1] 1.677446

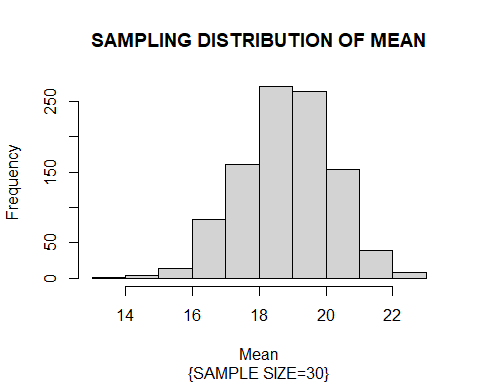
The variance and standard error are respectively 2.813 and 1.677. This is less than the values obtained for sample size=10.

## SAMPLE SIZE=30

samp\_dist3=replicate(1000,mean(sample(ToothGrowth$len,30,replace=TRUE)))  
head(samp\_dist3)

## [1] 19.51333 16.85000 18.64333 20.39333 20.87667 18.91333

#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] 1.904947

#STANDARD ERROR  
sd(samp\_dist3)

## [1] 1.380198

The variance and standard error are respectively 1.9 and 1.38. 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.

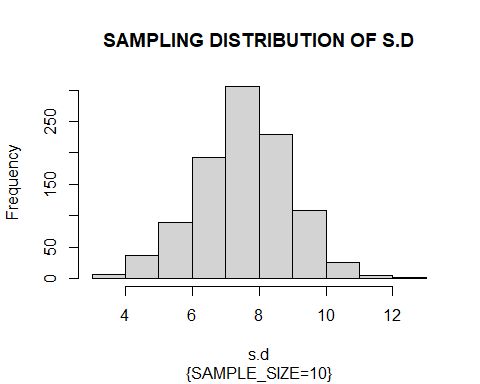
# SAMPLING DISTRIBUTION OF STANDARD DEVIATION

## SAMPLE SIZE=10

samp\_dist\_sd1=replicate(1000,sd(sample(ToothGrowth$len,10,replace=TRUE)))  
head(samp\_dist\_sd1)

## [1] 8.593311 9.960282 8.618204 5.778591 7.580677 3.798889

#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] 1.8843

#STANDARD ERROR  
sd(samp\_dist\_sd1)

## [1] 1.372698

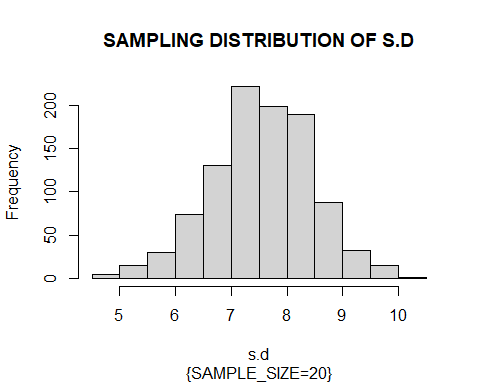
The variance and standard error are 1.88 and 1.37 respectively.

## SAMPLE SIZE=20

samp\_dist\_sd2=replicate(1000,sd(sample(ToothGrowth$len,20,replace=TRUE)))  
head(samp\_dist\_sd2)

## [1] 4.680854 7.181722 8.937449 9.137113 9.055792 7.647800

#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] 0.8280268

#STANDARD ERROR  
sd(samp\_dist\_sd2)

## [1] 0.9099598

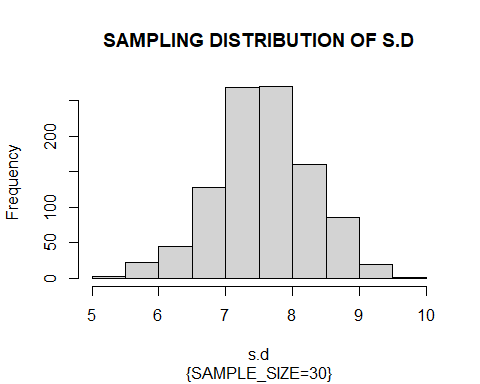
The variance and standard deviation are 0.82 and 0.9 respectively. This is less than the values obtained for sample size=10.

## SAMPLE SIZE=30

samp\_dist\_sd3=replicate(1000,sd(sample(ToothGrowth$len,30,replace=TRUE)))  
head(samp\_dist\_sd3)

## [1] 7.058765 8.375254 7.373996 8.132527 7.811115 7.637934

#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] 0.5192984

#STANDARD ERROR  
sd(samp\_dist\_sd3)

## [1] 0.7206236

The variance and standard deviation are 0.51 and 0.72 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 standardard 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.