Central Limit Theorem

SSSA - Applied Statistics- Chiara Seghieri and Costanza Tortù

2023-10-02

Idea

Scenario:

Imagine you are a social scientist conducting surveys to understand public opinion on a particular policy issue. Respondents are asked a yes-or-no question about their support for the policy. The responses are binary, with 1 indicating support and 0 indicating non-support.

Now, you want to analyze the average level of support across different samples, and you're interested in the distribution of these sample averages.

Central Limit Theorem (CLT): The Central Limit Theorem states that, regardless of the shape of the original population distribution, the distribution of the sum (or average) of a large number of independent, identically distributed random variables will be approximately normally distributed.

Population Distribution:

The distribution of individual opinions in the population might be quite arbitrary. Some individuals strongly support the policy, some are strongly against it, and others are neutral.

Sampling:

You collect multiple random samples of a fixed size from the population, each sample containing respondents and their binary responses (support or no support). For each sample, you calculate the average level of support by summing the 1s (support) and dividing by the sample size.

Applying CLT

According to the Central Limit Theorem, as you collect more and more samples and calculate their averages, the distribution of these sample averages will approach a normal distribution, regardless of the shape of the original population distribution.

This is particularly powerful because it allows you to make probabilistic statements about the distribution of sample averages even if the underlying population distribution is not normal.

Recap

Central Limit Theorem (CLT):

If you have a sufficiently large sample size drawn from any population with a finite mean and finite variance, the distribution of the sample mean (or the sum of the sample values) will be approximately normally distributed, regardless of the shape of the original population distribution.

In mathematical terms, let X_1, X_2, X_n be a sequence of independent and identically distributed random variables with a mean μ and finite variance σ^2.If n is sufficiently large, the distribution of the sample mean $\overline{X}$ will be approximately normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$

Get the idea through simulations

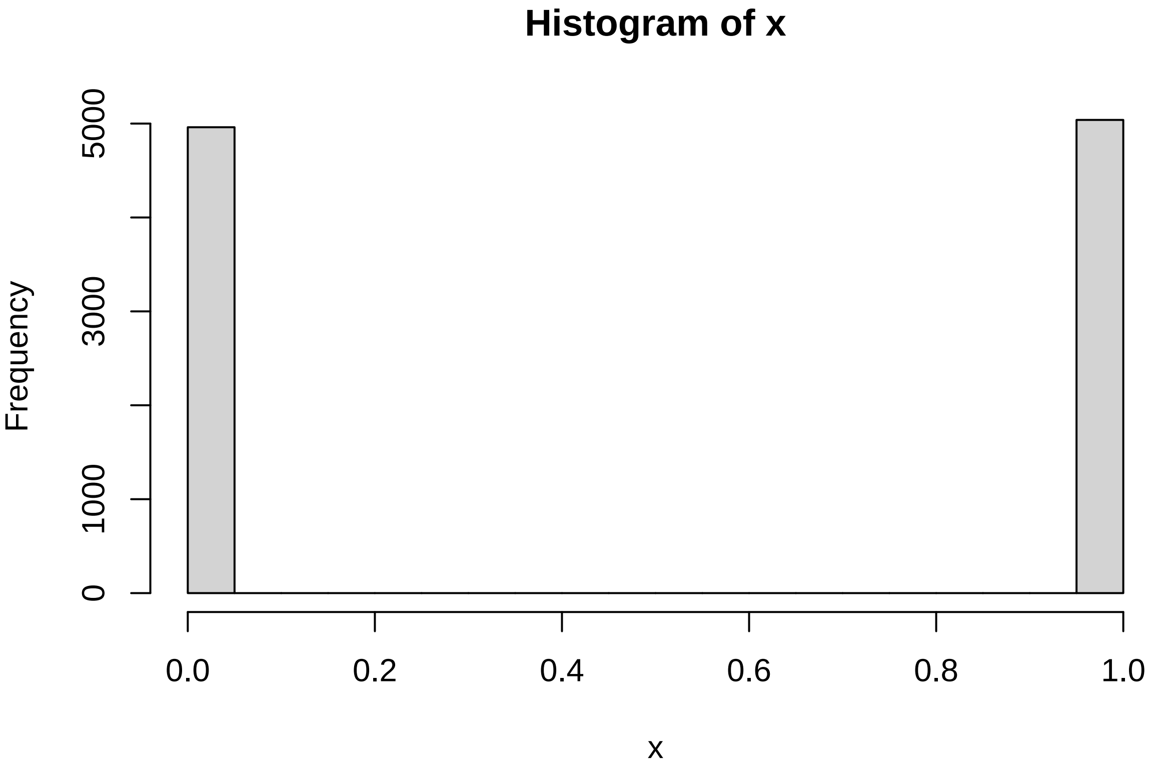
First step: simulate the data

Randomly generate N realizations of a binary variable X, whose probability of taking the value 1 is fixed and equal to 0.5. X represents the support to a given policy.

```
N <- 10000  # number of observations
x <- sample(0:1, N, replace = T) # generate a binary variable
```

Let's have a look at the distribution of X

```
hist(x)
```



The mean of X is

```
mean(x)
```

```
## [1] 0.5039
```

The standard deviation of X is

```
sd(x)
```

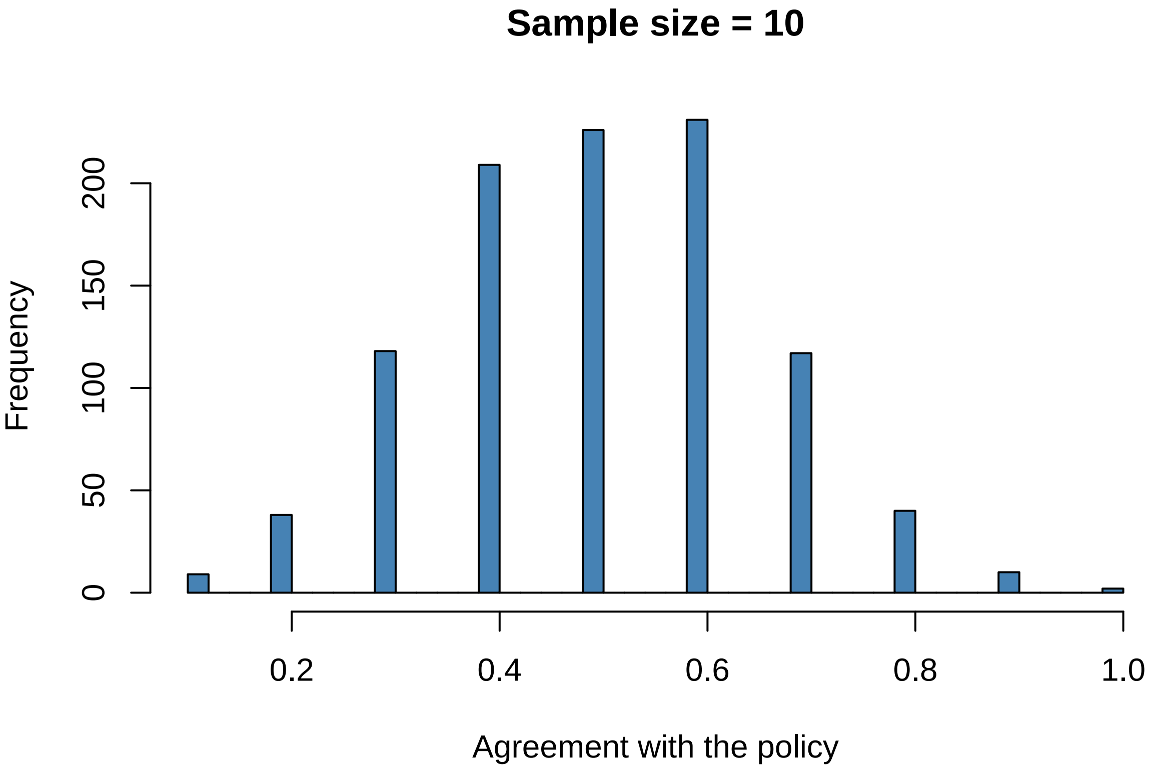
```
## [1] 0.5000098
```

Draw $k = 1000$ random samples of size n from X, compute the sample avergaes and plot the distribution

n=10

```
sample10 <- c()
k = 1000
for (i in 1:k){
sample10[i] = mean(sample(x, 10, replace=TRUE))
}

hist(sample10, col ='steelblue', xlab='Agreement with the policy', main='Sample size = 10',breaks = 40)
```



The mean of the distribution of sample averages is

```
mean(sample10)
```

```
## [1] 0.504
```

The standard deviation of the distribution of sample averages is

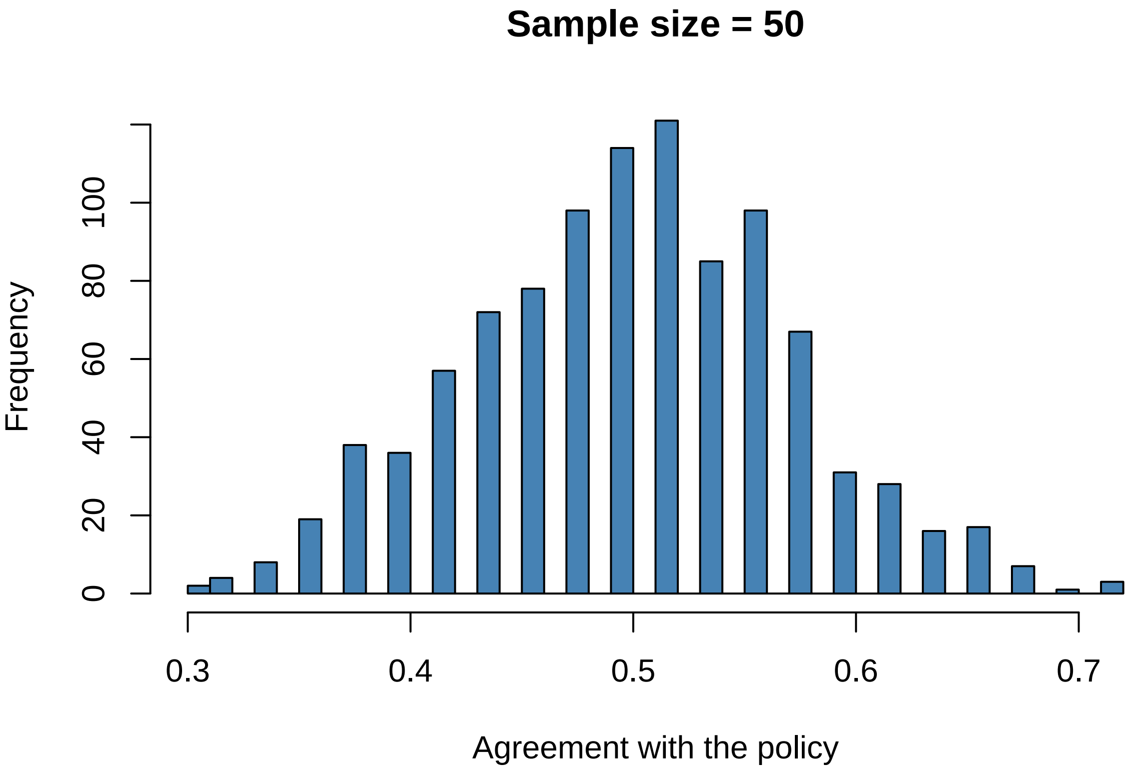
```
sd(sample10)
```

```
## [1] 0.1561037
```

n=50

```
sample50 <- c()
k = 1000
for (i in 1:k){
sample50[i] = mean(sample(x, 50, replace=TRUE))
}

hist(sample50, col ='steelblue', xlab='Agreement with the policy', main='Sample size = 50', breaks = 40)
```



The mean of the distribution of sample averages is

```
mean(sample50)
```

```
## [1] 0.50342
```

The standard deviation of the distribution of sample averages is

```
sd(sample50)
```

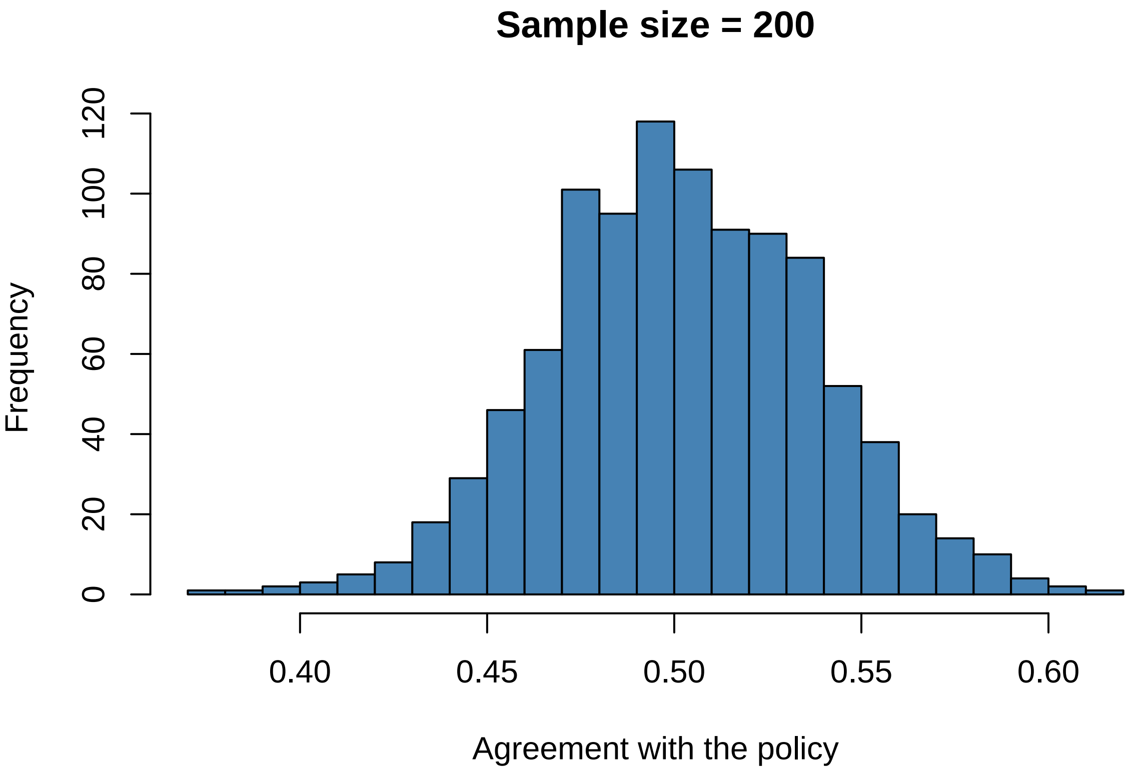
```
## [1] 0.07290829
```

n=200

```
sample200 <- c()
k = 1000
for (i in 1:k){
```

```
sample200[i] = mean(sample(x, 200, replace=TRUE))
}

hist(sample200, col ='steelblue', xlab='Agreement with the policy', main='Sample size = 200',breaks = 3
```



The mean of the distribution of sample averages is

```
mean(sample200)
```

```
## [1] 0.50461
```

The standard deviation of the distribution of sample averages is

```
sd(sample200)
```

```
## [1] 0.03573549
```

What happened?

We'll notice that as we use larger and larger sample sizes, the sample standard deviation gets lower and smaller.

Take home message: the bigger is the sample size the better it is!