Yang_Adam_PCE6

Adam Yang

Sampling Distributions

What is the difference between the sampling distribution of a statistic and the population distribution of a variable?

Review of the Central Limit Theorem

In this exercise, you will recreate the demonstration of the CLT seen in the async. Instead of using the Old Faithful data, you are to take random draws from a Bernoulli distribution.

Recall that a Bernoulli random variable with parameter p takes on just two values: 1, with probability p; and 0, with probability 1-p. We choose this variable because (1) it's very simple, and (2) its distribution is distinctly non-normal.

It turns out that (base) R doesn't have a Bernoulli function. To simulate draws from a Bernoulli variable, you can either

a. Use the sample command to select values from {0,1}

```
n=3
p = 0.5
sample(c(0,1), n, prob = c(1-p,p), replace = TRUE)
```

```
## [1] 0 0 0
```

b. Note that the Bernoulli distribution is a special case of the more general binomial distribution, with the binomial size parameter set to 1. R has an rbinom function that lets you draw from this distribution.

```
rbinom(3, size=1, prob=0.5)
```

```
## [1] 0 1 1
```

The Fair Coin

Using R, complete the following simulation exercise.

1. First, set p = 0.5 so your population distribution is symmeteric. Use a variable n to represent your sample size. Initially, set n = 3.

```
n=3
p = 0.5
sample(c(0,1), n, prob = c(1-p,p), replace = TRUE)
```

```
## [1] 0 0 0
```

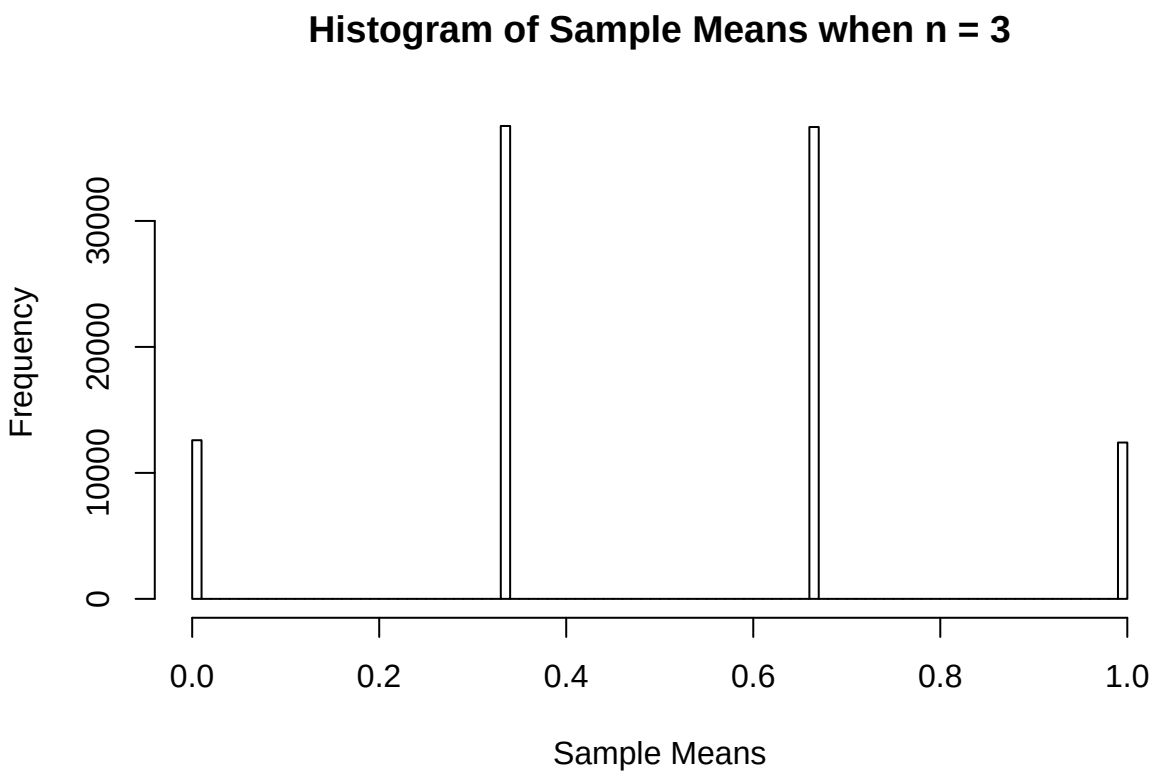
2. Write a function that simulates taking a random sample of n draws from a Bernoulli variable with parameter p, then return the sample mean.

```
execute_study = function(n, p){
  # your code here
  mean(sample(c(0,1), n, prob = c(1-p,p), replace = TRUE))
  }
```

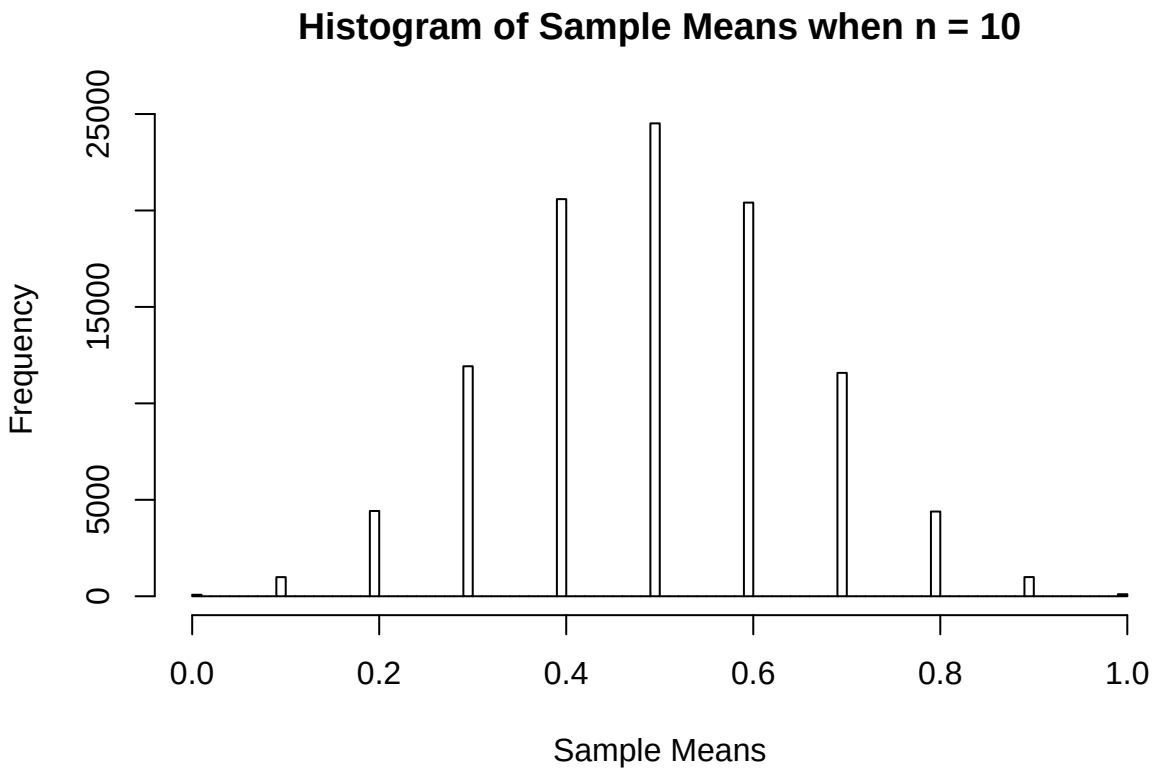
3. Write code that runs your function 100,000 times, storing all of the resulting sample means. Note that this would not be possible for a real-world study - this is just a thought experiment. Create a histogram of your result. Compute the standard deviation of the result. What does your histogram represent?

It seems like my sample means can only be a/n where a = 0,1,. . . ,n. The histograms distributions do look more and more normal as n gets larger.

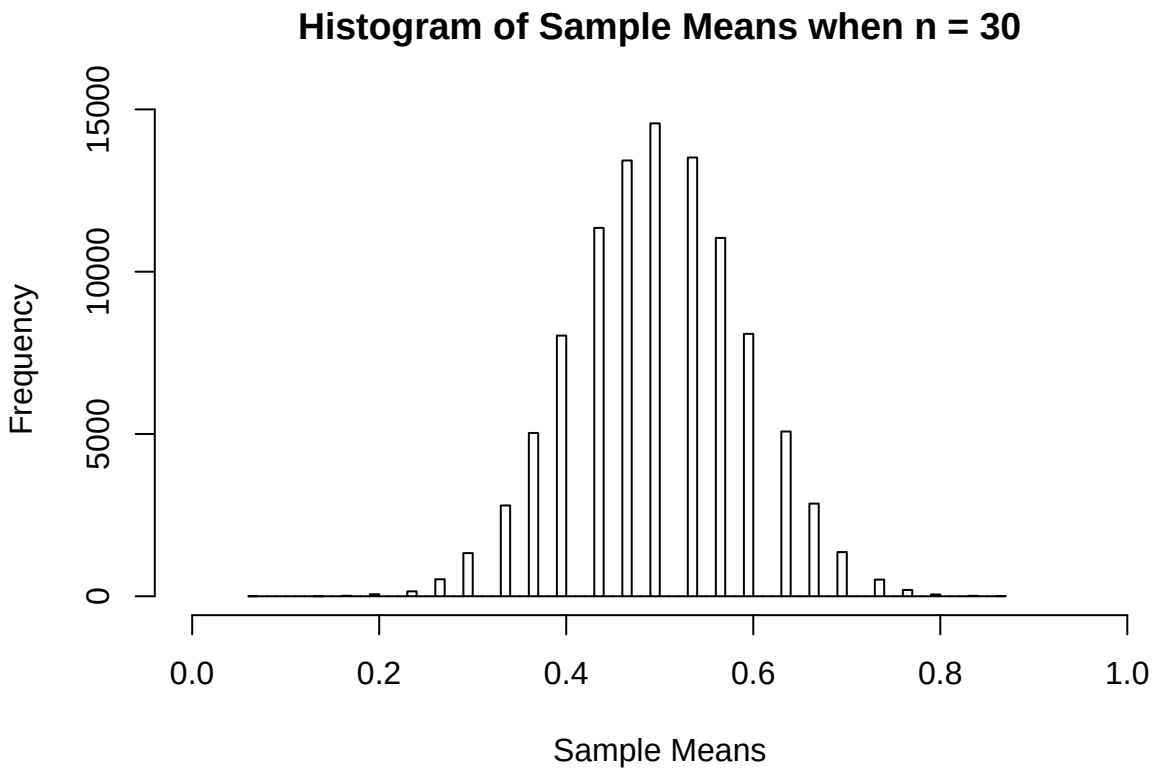
```
n = 3
p = 0.5
num_runs <- 100000
draws <- replicate(num_runs, execute_study(n,p))
hist(draws, breaks = 100, xlim = c(0,1),
     main = "Histogram of Sample Means when n = 3",
     xlab = "Sample Means")
```




```
n = 10
p = 0.5
num_runs <- 100000
draws <- replicate(num_runs, execute_study(n,p))
hist(draws, breaks = 100, xlim = c(0,1),
     main = "Histogram of Sample Means when n = 10",
     xlab = "Sample Means")
```




```
n = 30
p = 0.5
num_runs <- 100000
draws <- replicate(num_runs, execute_study(n,p))
hist(draws, breaks = 100, xlim = c(0,1),
     main = "Histogram of Sample Means when n = 30",
     xlab = "Sample Means")
```


Histogram of Sample Means when n = 30
Frequency
0
5000
10000
15000
0.0
0.2
0.4
0.6
0.8
1.0
Sample Means