Unit 6 Pre-Class Warm-up

Shan He

Fall 2017

Sampling Distributions

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

The population distribution of a variable is the overall distribution of the values of the variables in the whole population. The sampling distribution of a statistic is the distribution of a sample statistic from samples of a certain size taken from the whole population.

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), 3, prob = c(1-p,p), replace = TRUE)
```

```
## [1] 1 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.

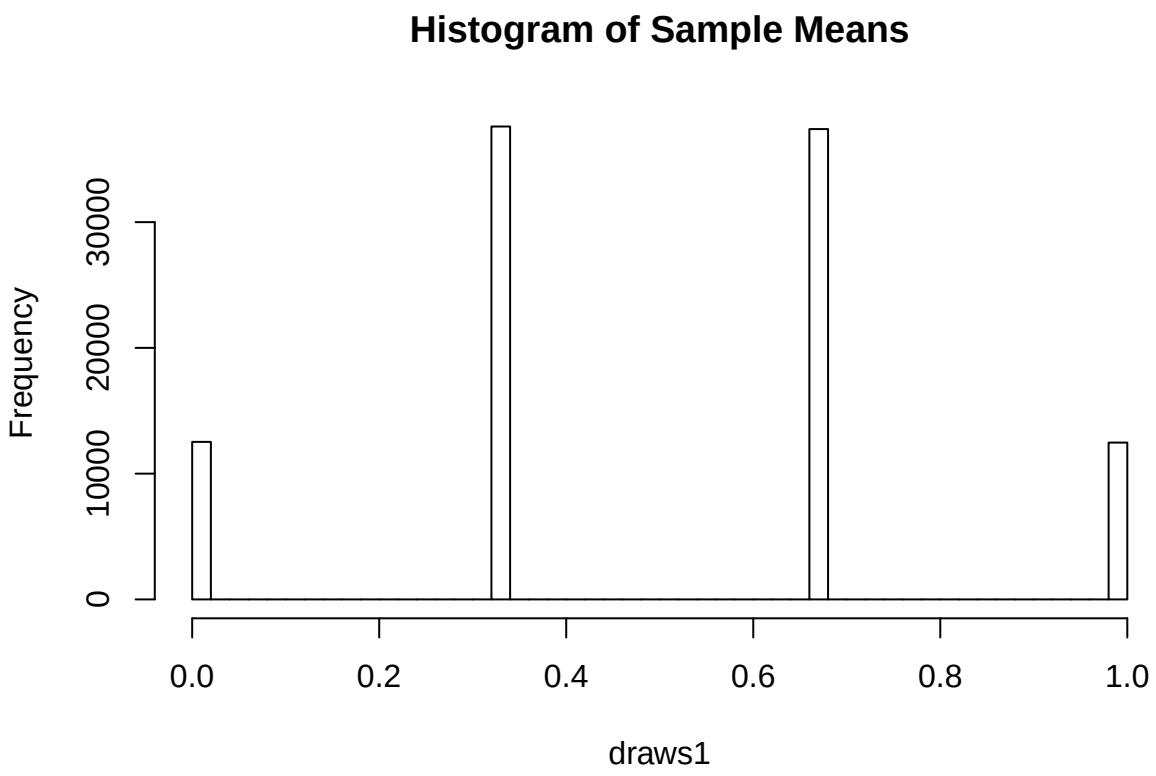
1. First, set p = 0.5 so your population distribution is symmeteric. Use a variable n to represent your sample size. Initially, set $n = 3$.
2. Write a function that simulates taking a random sample of n draws from a Bernoulli variable with parameter p, then returns the sample mean.

```
execute_study = function(n, p){
  means = 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?

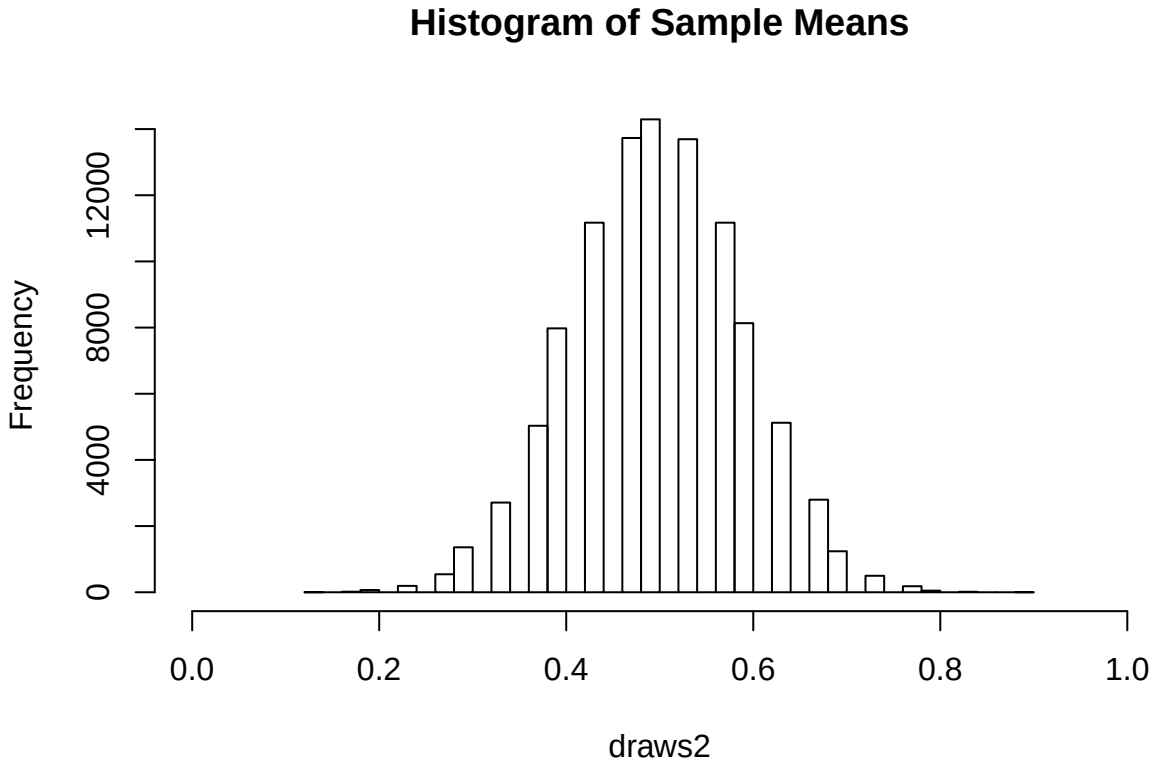
```
#replicate function 100,000 times.
draws1 <- replicate(100000, execute_study(3, 0.5))

#histogram of 100,000 replications
hist(draws1, breaks = 50, xlim = c(0,1), main = "Histogram of Sample Means")
```




```
#replicate function 100,000 times. Increased n to 30
draws2 <- replicate(100000, execute_study(30, 0.5))

#histogram of 100,000 replications
hist(draws2, breaks = 50, xlim = c(0,1), main = "Histogram of Sample Means")
```




```
#standard deviation for draws1
mean(draws1)
```

```
## [1] 0.4993733
```

```
sd(draws1)
```

```
## [1] 0.2886567
```

```
#standard deviation for draws2
mean(draws2)
```

```
## [1] 0.499964
```

```
sd(draws2)
```

```
## [1] 0.09088401
```

The histogram represents the sampling distribution of mean for the Bernouli distribution. When n = 3, it has a mean of 0.50 and a standard deviation of 0.29. As n increases to 30, it appears to be a normal distribution with a mean of 0.50 and a standard deviation of 0.09.