STAT 140 - Lab 8

Sampling distributions and the Central Limit Theorem

YOUR NAME HERE

3/30/2020

Sampling distribution - Roosevelt Dimes

For this lab, we will be working with data about United States Roosevelt dimes (the dimes you typically see if you get change). I have compiled a data set with their mintages (number of dimes produced) by year, from 1948, when they were first produced, to 2018. We will be using these data to explore sampling distributions.

```
## Read in coin data
roosevelt_dimes <- read_csv("https://marievozanne.github.io/roosevelt_dimes.csv")
```

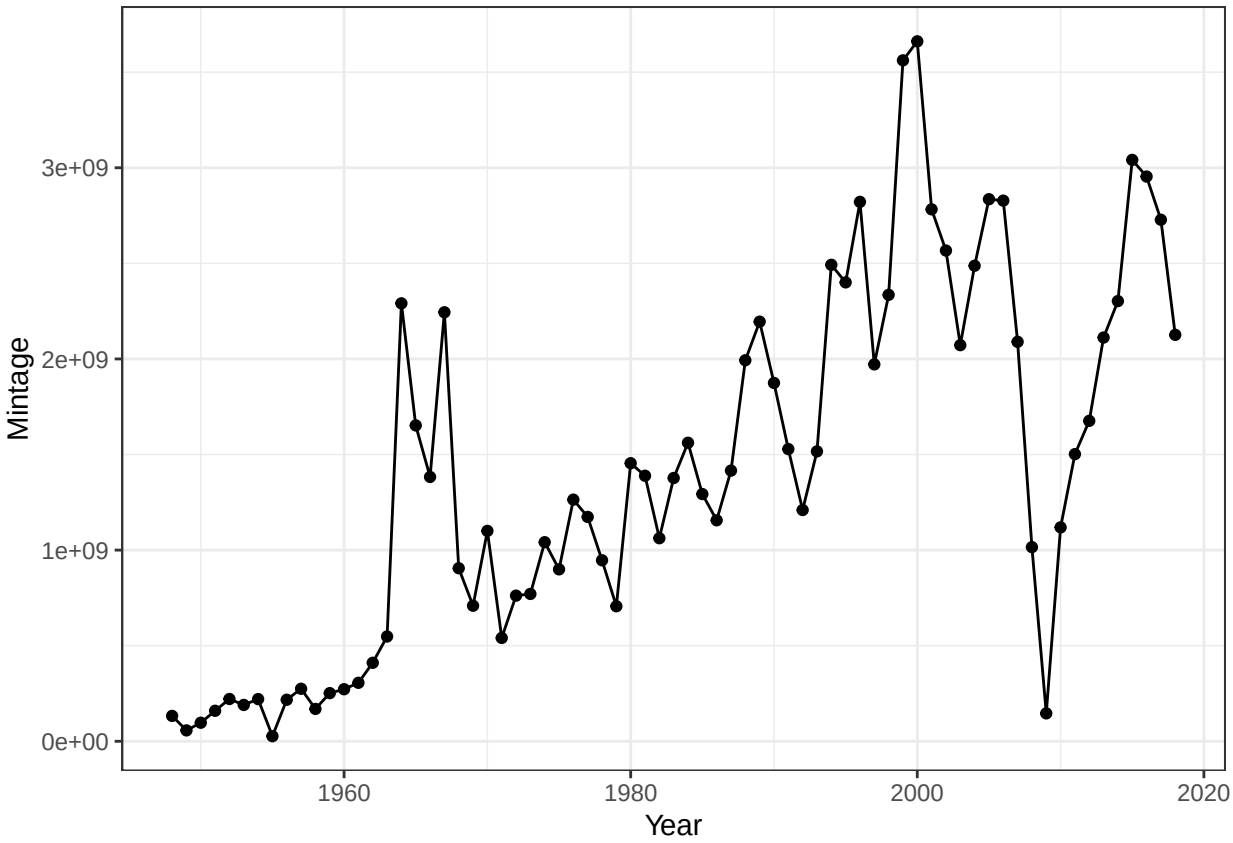
```
## Parsed with column specification:
## cols(
##   Year = col_double(),
##   Mintage = col_double()
## )
```

```
## Add column for probabilities
roosevelt_dimes$Probability <- roosevelt_dimes$Mintage/sum(roosevelt_dimes$Mintage)

## Add indicator (0/1) column for coins after 2000 - 0 if before, 1 otherwise
roosevelt_dimes$Ind_After2000 <- rep(0, nrow(roosevelt_dimes))
roosevelt_dimes[roosevelt_dimes$Year > 1999,]$Ind_After2000 <- 1
```

Sample mean

Make a scatterplot of Mintage versus Year. Add a line connecting the points using `geom_line()`.

```
ggplot(data=roosevelt_dimes, aes(x=Year, y=Mintage)) +
  geom_point() +
  geom_line() +
  theme_bw()
```

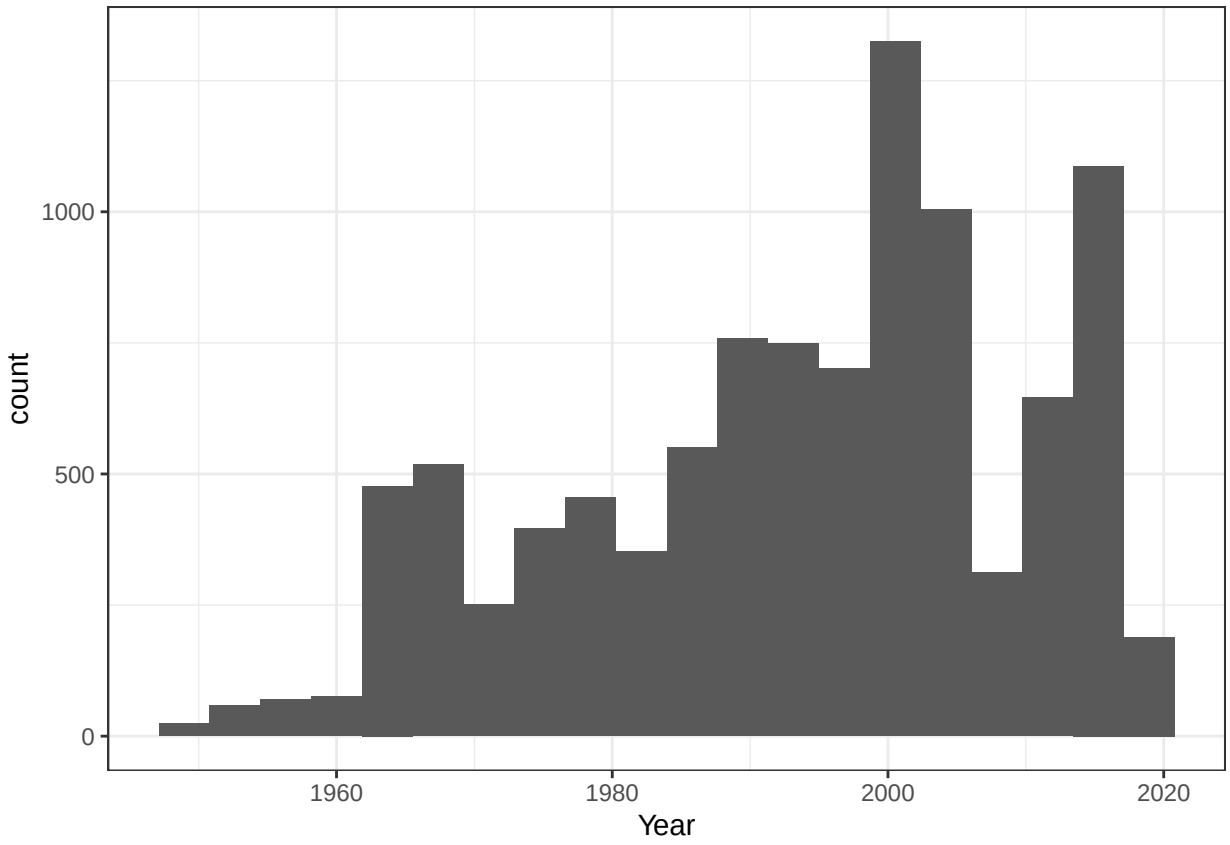


Use the following code to make a new data set of the years.

```
parent_Year <- data.frame(Year=sample(roosevelt_dimes$Year, prob=roosevelt_dimes$Probability, replace=TR
```

Using `parent_Year`, make a histogram of the dime mintage years. Does the distribution of mintage years appear to be normal?

```
ggplot(data=parent_Year, aes(x=Year)) +
  geom_histogram(bins=20) +
  theme_bw()
```



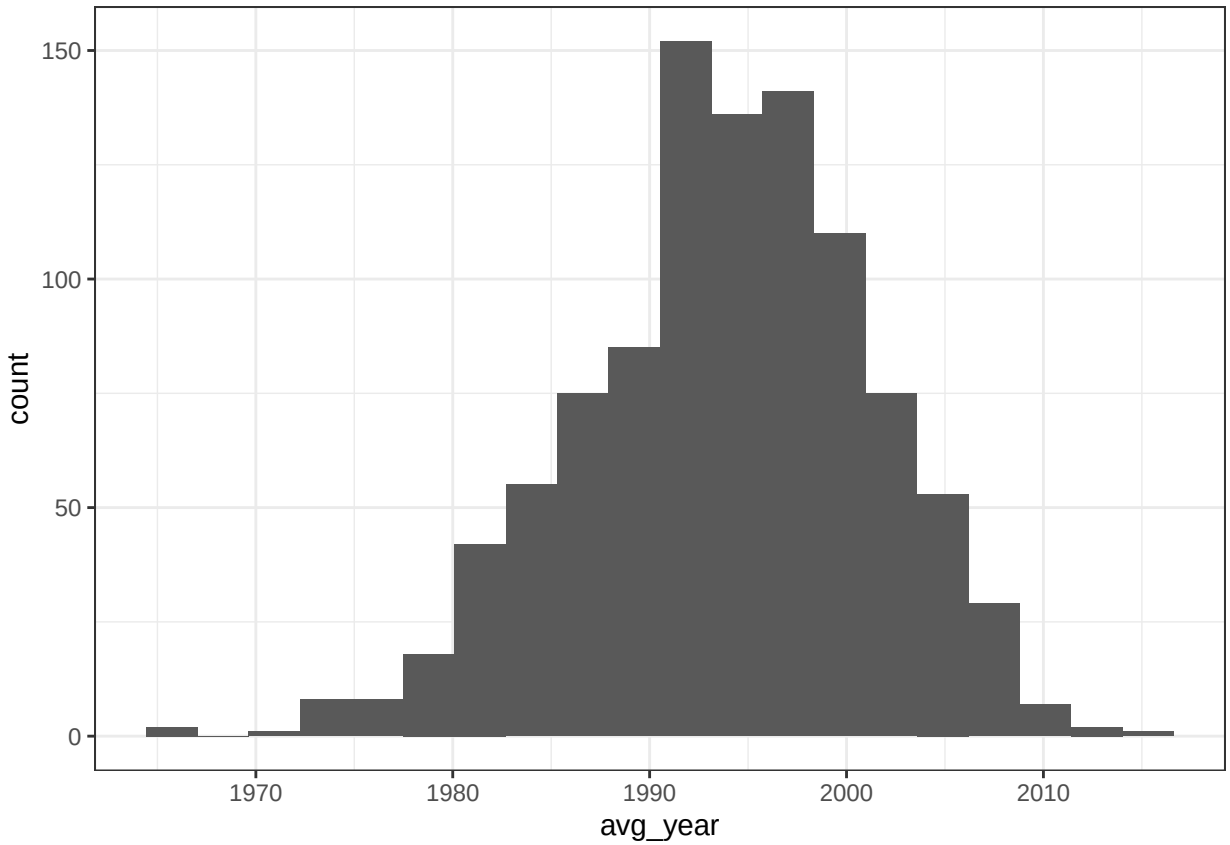
No, the distribution of mintage years does not appear to be normal. It is left skewed and possibly bimodal.

Use the following code to generate 1000 samples of size 5 and find the average mintage year for each sample. We will use this to plot a sampling distribution for the average mintage year for samples of size 5 (n=5).

```
avg_year <- rep(NA, 1000)
for (i in 1:1000){
  avg_year[i] <- mean(sample(roosevelt_dimes$Year, prob=roosevelt_dimes$Probability,
                             replace=TRUE, size=5))
}

sampling_AvgYear_n5 <- data.frame(avg_year=avg_year)
```

Using `sampling_AvgYear_n5`, make a histogram of the average dime mintage years. Does the sampling distribution appear to be normal?

```
ggplot(data=sampling_AvgYear_n5, aes(x=avg_year)) +
  geom_histogram(bins=20) +
  theme_bw()
```



This looks closer to normal than the parent distribution, but I would say it is still slightly left-skewed.

Based on what you know about the Central Limit Theoreom, can you say that the sampling distribution (for n=5) is approximately normal?

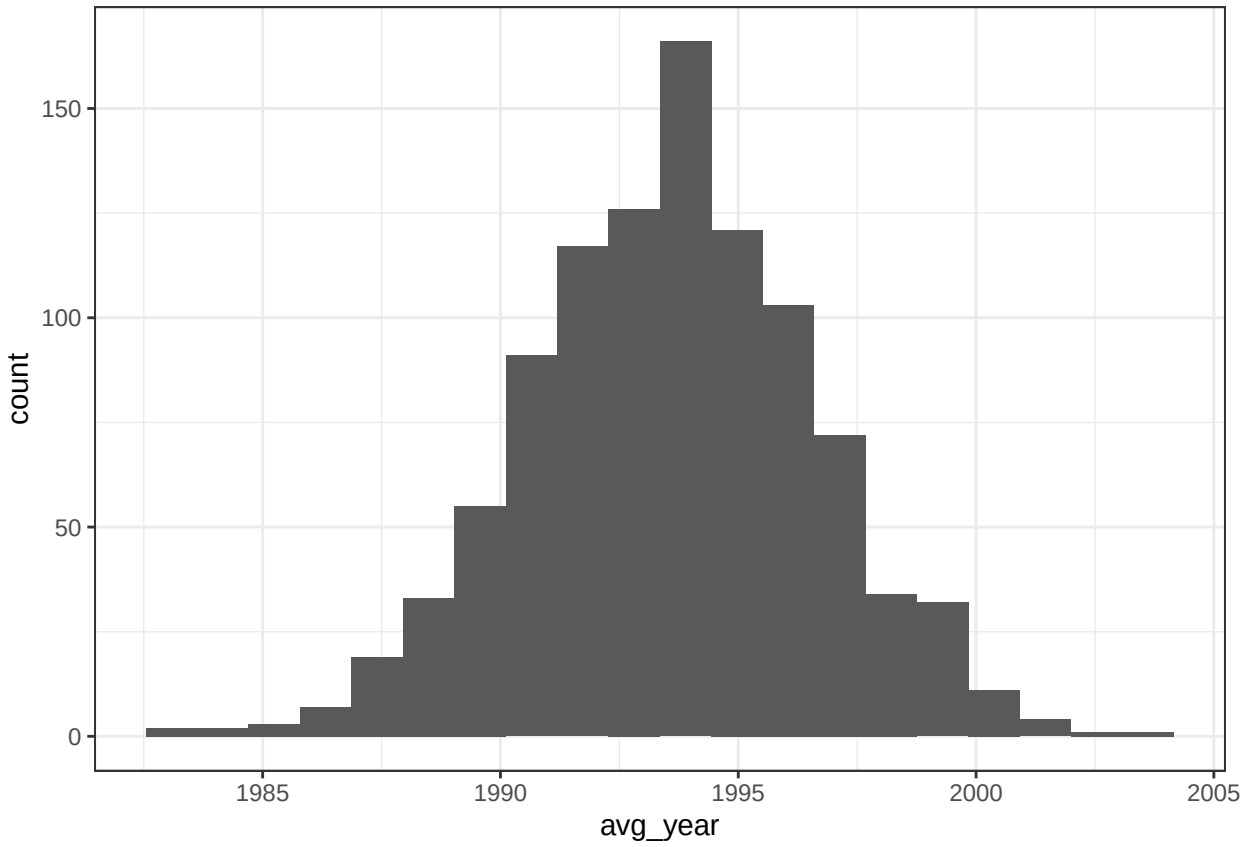
The Central Limit Theorem states that I need a sample size of at least 30 to say that the sampling distribution for the mean is approximately normal. Even though the histogram looks compelling, this technically is not approximately normal because $n = 5 < 30$. Remember, $n \geq 30$ is a rule of thumb that is applied to every situation, so there may be situations where sampling distributions look normal for smaller n, but we're going to stick with our rule. Thus, the sampling distribution is NOT approximately normal because the sample size is too small to satisfy the Central Limit Theorem.

Use the following code to generate 1000 samples of size 30 and find the average mintage year for each sample. We will use this to plot a sampling distribution for the average mintage year for samples of size 30 (n=30).

```
avg_year <- rep(NA, 1000)
for (i in 1:1000){
  avg_year[i] <- mean(sample(roosevelt_dimes$Year, prob=roosevelt_dimes$Probability,
                             replace=TRUE, size=30))
}

sampling_AvgYear_n30 <- data.frame(avg_year=avg_year)
```

Using `sampling_AvgYear_n30`, make a histogram of the average dime mintage years. Does the sampling distribution appear to be normal?

```
ggplot(data=sampling_AvgYear_n30, aes(x=avg_year)) +
  geom_histogram(bins=20) +
  theme_bw()
```



Yes, the sampling distribution for the mean appears normal when $n = 30$. It is symmetric (more symmetric than that for $n = 5$).

Find the mean and standard deviation of `sampling_AvgYear_n30`. Save them as `mean_n30` and `sd_n30`, respectively.

```
mean_n30 <- mean(sampling_AvgYear_n30$avg_year)
sd_n30 <- sd(sampling_AvgYear_n30$avg_year)
```

Based on what you know about the Central Limit Theoreom, can you say that the sampling distribution (for n=30) is approximately normal? What are the mean and the standard error for this distribution?

Yes, since $n = 30$ and the samples are independent, the sampling distribution for the mean is approximately normal. The mean is 1993.583 and the standard deviation is 2.997.

Use the following code to generate 1000 samples of size 50 and find the average mintage year for each sample. We will use this to plot a sampling distribution for the average mintage year for samples of size 50 (n=50).

```
avg_year <- rep(NA, 1000)
for (i in 1:1000){
  avg_year[i] <- mean(sample(roosevelt_dimes$Year, prob=roosevelt_dimes$Probability,
                             replace=TRUE, size=50))
}

sampling_AvgYear_n50 <- data.frame(avg_year=avg_year)
```

Find the mean and standard deviation of `sampling_AvgYear_n50`. Save them as `mean_n50` and `sd_n50`, respectively. How do they compare to `mean_n30` and `sd_n30`?

```
mean_n50 <- mean(sampling_AvgYear_n50$avg_year)
sd_n50 <- sd(sampling_AvgYear_n50$avg_year)
```

The means for $n = 30$ and $n = 50$ are about the same: 1993.583 and 1993.381, respectively. We expect this - they should be centered around the population mean. The standard deviation for $n = 30$ is greater than that for $n = 50$: $2.997 > 2.261$. A larger sample size allows us to be more precise about the mean, so there is less variability from sample to sample with regard to mean for a larger sample size.

Sample proportion

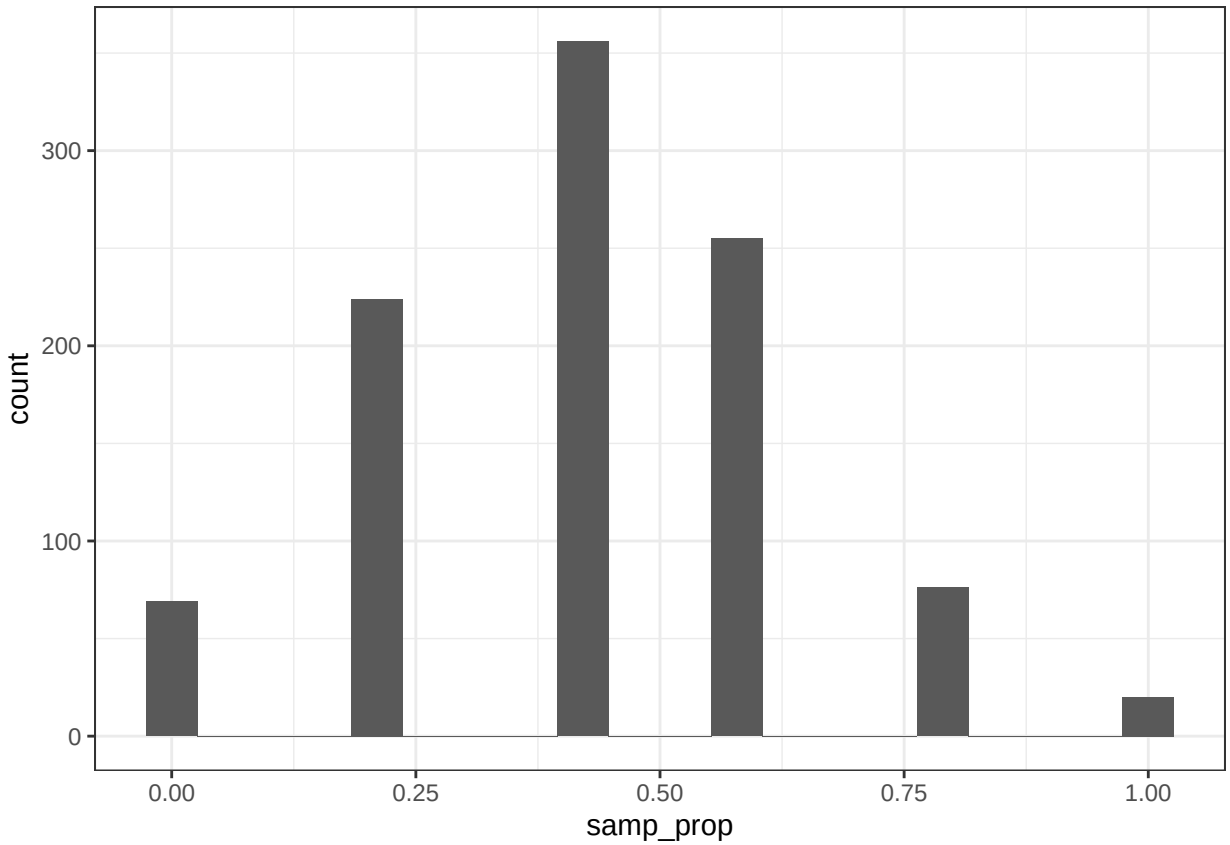
In the exercises above, we looked at sampling distributions for the sample mean. We can also look at sampling distributions for the sample proportion. For this, we will use the `Ind_After2000` variable.

Use the following code to generate 1000 samples of size 5 and find the proportion of successes (coin minted in 2000 or later). We will use this to plot a sampling distribution for the sample proportion for samples of size 5 (n=5).

```
n <- 5

samp_prop <- rep(NA, 1000)
for (i in 1:1000){
  samp_prop[i] <- mean(sample(roosevelt_dimes$Ind_After2000, prob=roosevelt_dimes$Probability,
                             replace=TRUE, size=n))
}

sampling_samp_prop_n5 <- data.frame(samp_prop=samp_prop)
```

Using `sampling_samp_prop_n5`, make a histogram of the average dime mintage years. Does the sampling distribution appear to be normal? Check your conditions, $np \geq 10$ and $n(1-p) \geq 10$‘ to see if the conditions for approximate normality (and the Central Limit Theorem) are satisfied.

```
ggplot(data=sampling_samp_prop_n5, aes(x=samp_prop)) +
  geom_histogram(bins=20) +
  theme_bw()
```




```
## The success probability is the sum of the probabilities
## of all the years greater than or equal to 2000
p <- sum(roosevelt_dimes$Ind_After2000*roosevelt_dimes$Probability)

succ_5 <- 5*p
fail_5 <- 5*(1-p)
```

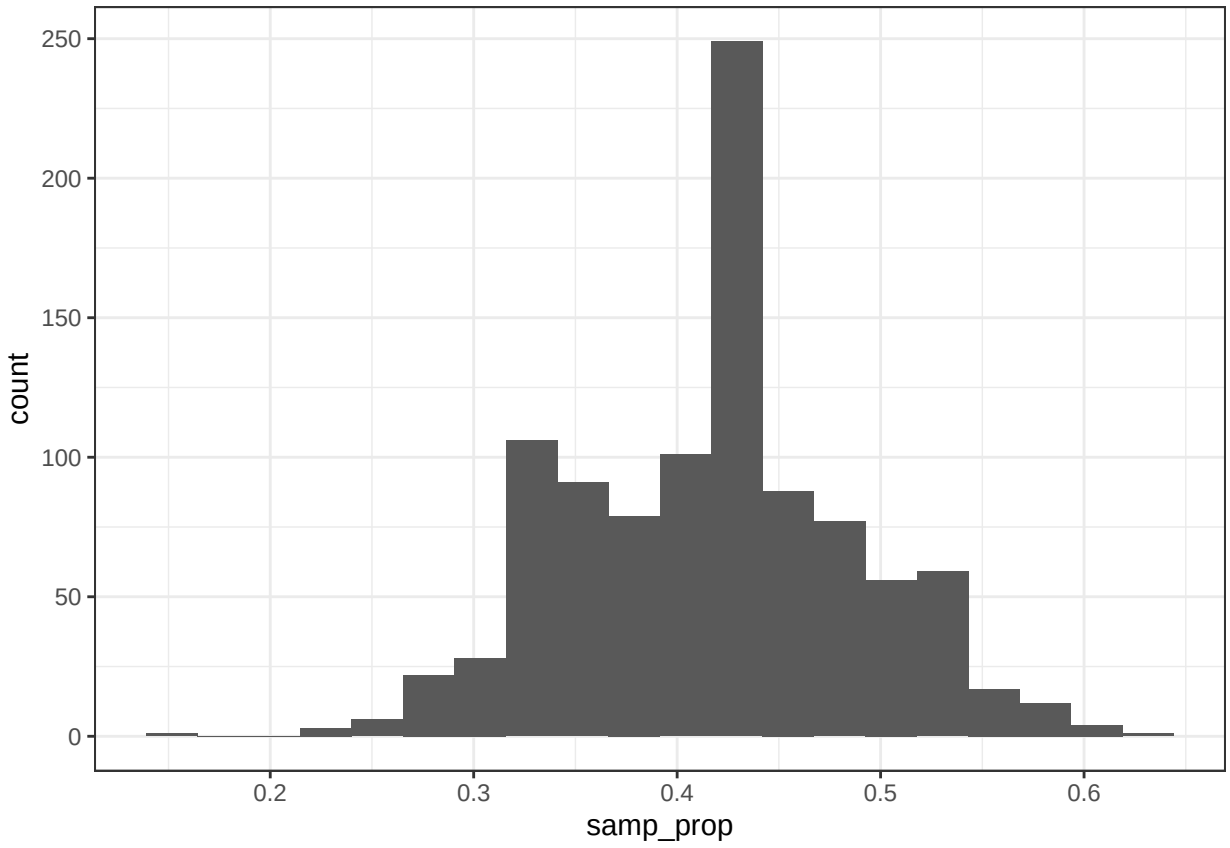
No, the sampling distribution does not appear normal. It is right skewed, and it also has large gaps (this is always a characteristic of the sampling distribution for the proportion, but the gaps get smaller for larger n). The conditions are not satisfied: $np = 2.09$ and $n(1-p) = 2.91$, which are both less than 10. The samples are independent.

Use the following code to generate 1000 samples of size 50 and find the proportion of successes (coin minted in 2000 or later). We will use this to plot a sampling distribution for the sample proportion for samples of size 50 (n=50).

```
n <- 50

samp_prop <- rep(NA, 1000)
for (i in 1:1000){
  samp_prop[i] <- mean(sample(roosevelt_dimes$Ind_After2000, prob=roosevelt_dimes$Probability,
                              replace=TRUE, size=n))
}

sampling_samp_prop_n50 <- data.frame(samp_prop=samp_prop)
```

Using `sampling_samp_prop_n50`, make a histogram of the average dime mintage years. Does the sampling distribution appear to be normal? Check your conditions, $np \geq 10$ and $n(1-p) \geq 10$‘ to see if the conditions for approximate normality (and the Central Limit Theorem) are satisfied.

```
ggplot(data=sampling_samp_prop_n50, aes(x=samp_prop)) +
  geom_histogram(bins=20) +
  theme_bw()
```




```
succ_50 <- 50*p
fail_50 <- 50*(1-p)
```

Yes, the sampling distribution does appear normal. The conditions are satisfied: $np = 20.9 \geq 10$ and $n(1-p) = 29.1 \geq 10$. The samples are independent.

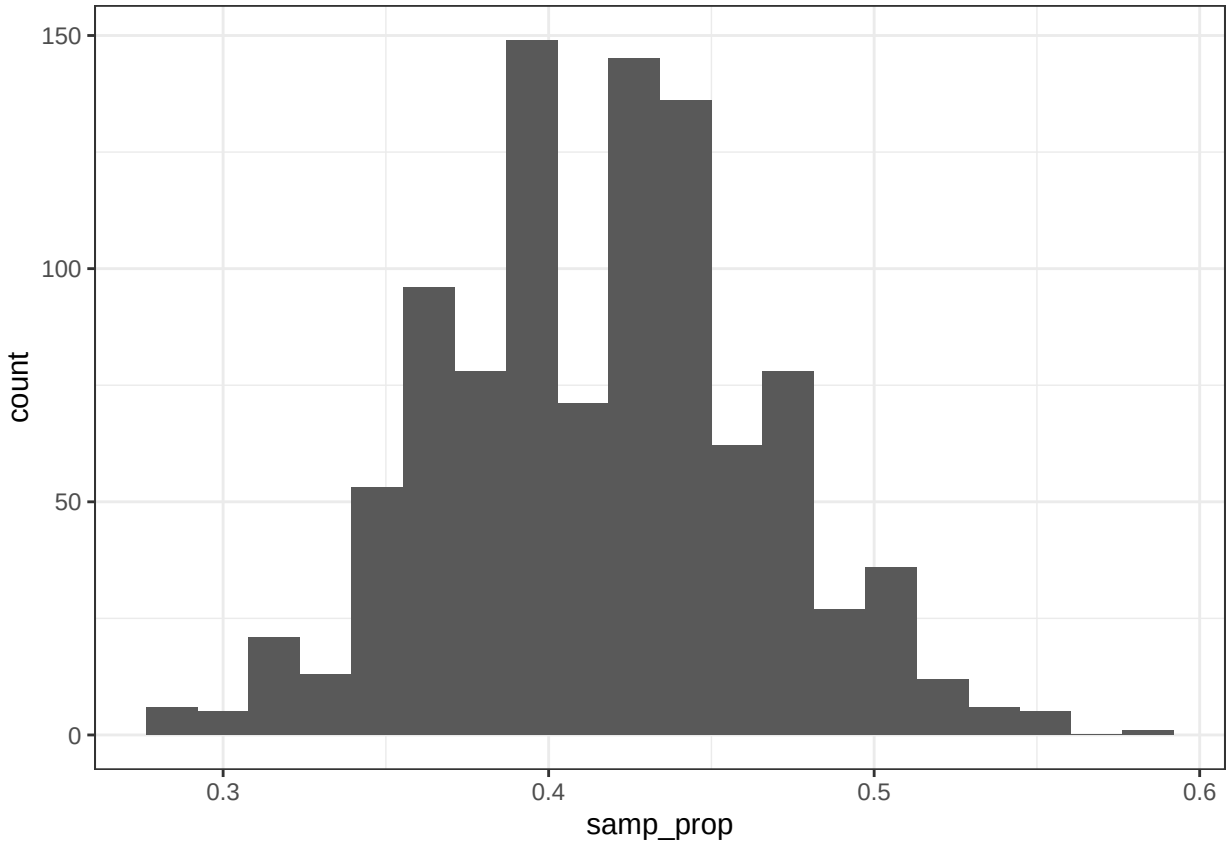
Use the following code to generate 1000 samples of size 100 and find the proportion of successes (coin minted in 2000 or later). We will use this to plot a sampling distribution for the sample proportion for samples of size 100 (n=100).

```
n <- 100

samp_prop <- rep(NA, 1000)
for (i in 1:1000){
  samp_prop[i] <- mean(sample(roosevelt_dimes$Ind_After2000, prob=roosevelt_dimes$Probability,
```

```
                              replace=TRUE, size=n))
}

sampling_samp_prop_n100 <- data.frame(samp_prop=samp_prop)
```

Using `sampling_samp_prop_n100`, make a histogram of the average dime mintage years. Does the sampling distribution appear to be normal? Check your conditions, $np \geq 10$ and $n(1-p) \geq 10$ to see if the conditions for approximate normality (and the Central Limit Theorem) are satisfied.

```
ggplot(data=sampling_samp_prop_n100, aes(x=samp_prop)) +
  geom_histogram(bins=20) +
  theme_bw()
```




```
succ_100 <- 100*p
fail_100 <- 100*(1-p)
```

Yes, the sampling distribution does appear normal. The conditions are satisfied: $np = 41.8 \geq 10$ and $n(1-p) = 58.2 \geq 10$. The samples are independent.

Independence assumption

An underlying assumption of the Central Limit Theorem is that the observations in each sample are independent. Based on what you know about the sample function in R, do you think the independence assumption

is satisfied?

Yes, the sample function randomly samples from the population, so independence is satisfied.