Practice Session 8

Introduction:

Distribution of a Sample Proportion:

When selecting random samples of size n from a population with proportion p, the distribution of the sample proportions is centered at the population proportion p, and has a standard error given by:

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

and is reasonably normally distributed if $np \geq 10$ and $n(1-p) \geq 10$.

Confidence Interval for a Proportion:

If z^* is a standard normal endpoint to give the desired level of confidence, and if the sample size is large enough so that $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$, the confidence interval for a population proportion p is

$$\text{Sample statistic } \pm z^* \cdot SE$$

Where sample proportion based on a random sample of size n and the standard error are

$$\text{Sample statistic } = \hat{p}$$

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Hypothesis Test for a Proportion:

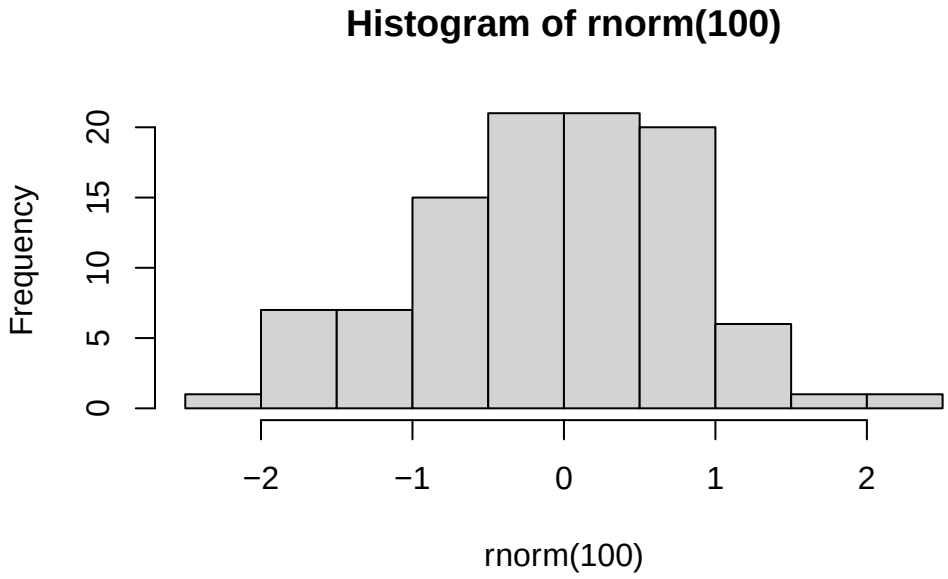
If the sample size is reasonably large so that $np_0 \geq 10$, then we can test $H_0 : p = p_0$ vs $H_a : p \neq p_0$ (or a one-tail alternative), and the standardized test statistic is:

$$z = \frac{\text{Statistic} \ - \ \text{Null Value}}{SE} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Practice with `rnorm`, `pnorm`, `qnorm`, and `mosaic::cnorm`

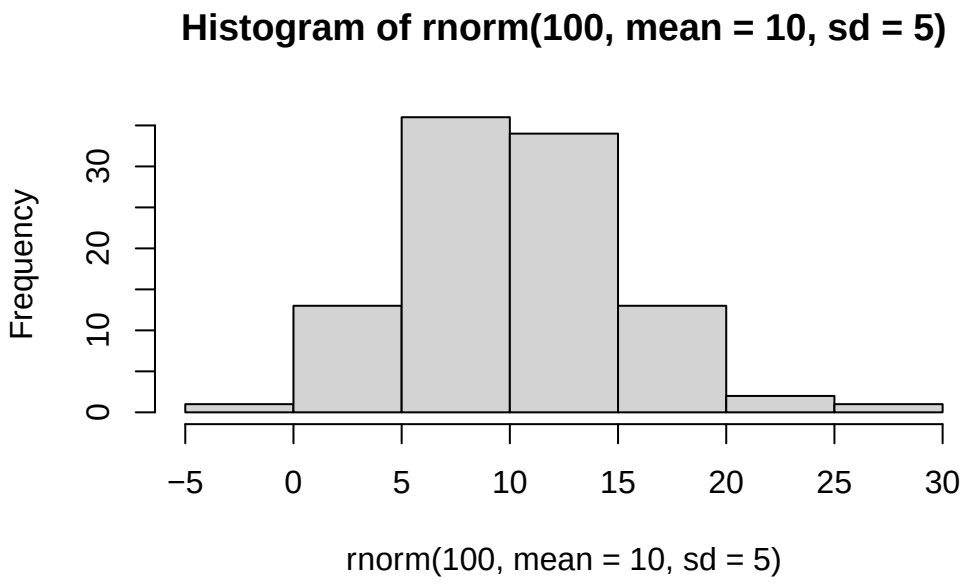
a.) Using the `rnorm`, generate 100 random observations from a normal distribution. What are the default mean and standard deviation values? Plot a histogram of the data.

```
hist(
  rnorm(100)
)
```



b.) Now, change the mean and standard deviation to $\mu = 10$ and $\sigma = 5$. Plot a histogram of the data and compare.

```
hist(
  rnorm(100, mean = 10, sd = 5)
)
```



c.) Using the `pnorm` function with $\mu = 10$ and $\sigma = 5$, calculate the probability of observing a value greater than 13.

```
pnorm(13, mean = 10, sd = 5, lower.tail = F)
```

```
[1] 0.2742531
```

d.) Using the `pnorm` function with $\mu = 10$ and $\sigma = 5$, calculate the probability of observing a value less than 7.

```
pnorm(7, mean = 10, sd = 5)
```

```
[1] 0.2742531
```

e.) Using the `pnorm` function with $\mu = 10$ and $\sigma = 5$, calculate the probability of observing a value less 15 but greater than 11.

```
pnorm(15, mean = 10, sd = 5) - pnorm(11, mean = 10, sd = 5)
```

```
[1] 0.262085
```

f.) Using the `qnorm` function with $\mu = 10$ and $\sigma = 5$, find the population value that corresponds to the 90th percentile.

```
qnorm(0.9, mean = 10, sd = 5)
```

```
[1] 16.40776
```

g.) Using the `qnorm` function with $\mu = 10$ and $\sigma = 5$, find the population value that corresponds to the 20th percentile.

```
qnorm(0.2, mean = 10, sd = 5)
```

```
[1] 5.791894
```

h.) Using the `qnorm` function with $\mu = 10$ and $\sigma = 5$, find the population values corresponding to a two-sided interval that contains 95% of the population.

```
qnorm(0.975, mean = 10, sd = 5)
```

```
[1] 19.79982
```

```
qnorm(0.025, mean = 10, sd = 5)
```

```
[1] 0.2001801
```

i.) Using the `mosaic::cnorm` function with $\mu = 10$ and $\sigma = 5$, find the population values corresponding to a two-sided 95% interval.

```
suppressPackageStartupMessages({library(mosaic)})
mosaic::cnorm(0.95, mean = 10, sd = 5)
```

```
         lower    upper
[1,] 0.2001801 19.79982
```

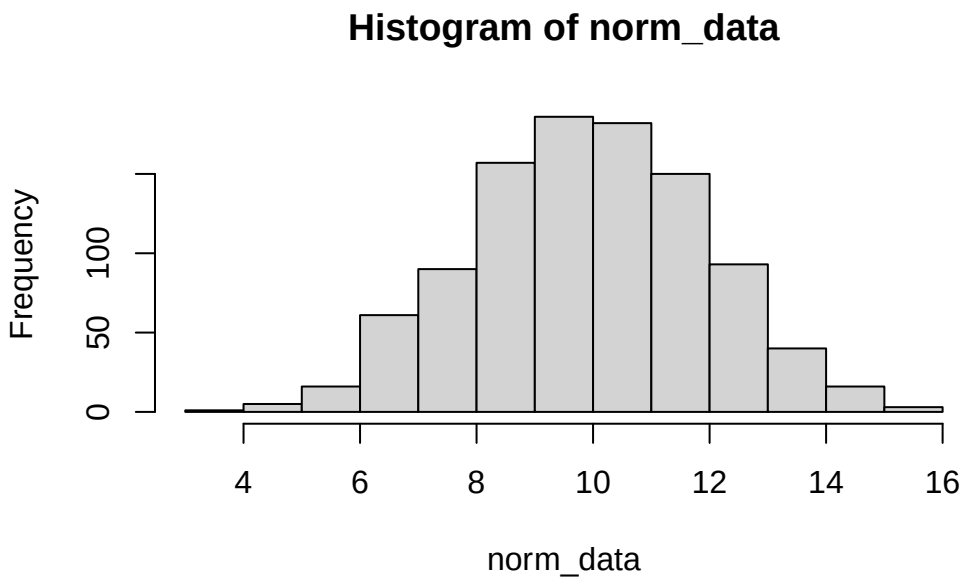
Z-Scores and the Standard Normal Distribution

a.) Generate 1000 random observations from a normal distribution with mean 10 and standard deviation 2. Save the result to the variable `norm_data`

```r
norm_data = rnorm(1000, mean = 10, sd = 2)
```

b.) Create a histogram of `norm_data`.

```r
hist(norm_data)
```



c.) Convert `norm_data` into a vector of Z-scores using the following formula. Call this vector `z_data`.

```r
z_data = (norm_data - 10) / 2
```

d.) Compute the mean and standard deviation of `z_data`. How do they compare to the parameter values you defined in part (a)?

```r
mean(z_data)
```

```
[1] -0.04191159
```

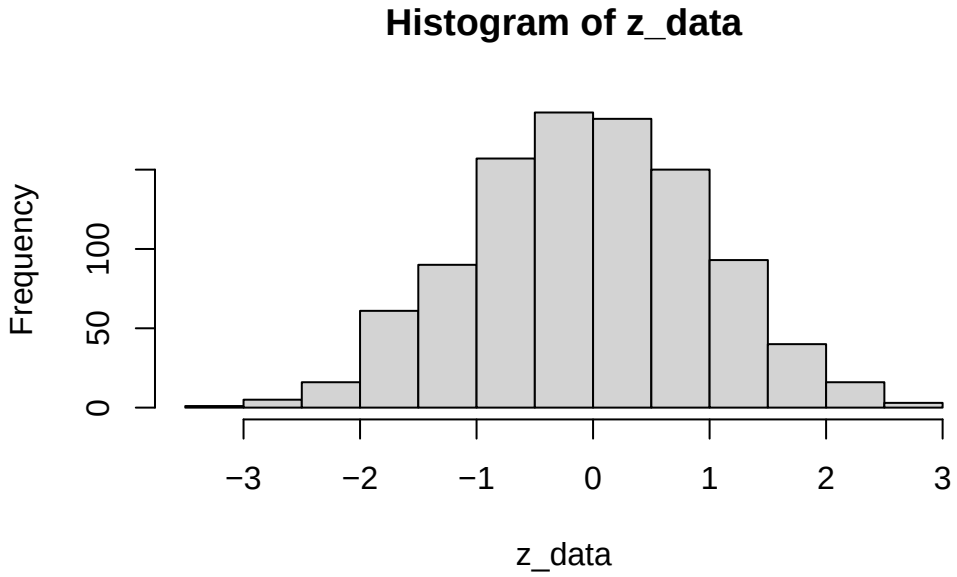
```r
sd(z_data)
```

```
[1] 0.9948422
```

It seems that `z_data` has a mean of 0, and a standard deviation of 1.

e.) Create a histogram of `z_data`. How does its center and spread compare to the histogram you created in part (b)?

```
hist(z_data)
```



This histogram is centered at 0, and has a smaller spread than the first histogram.

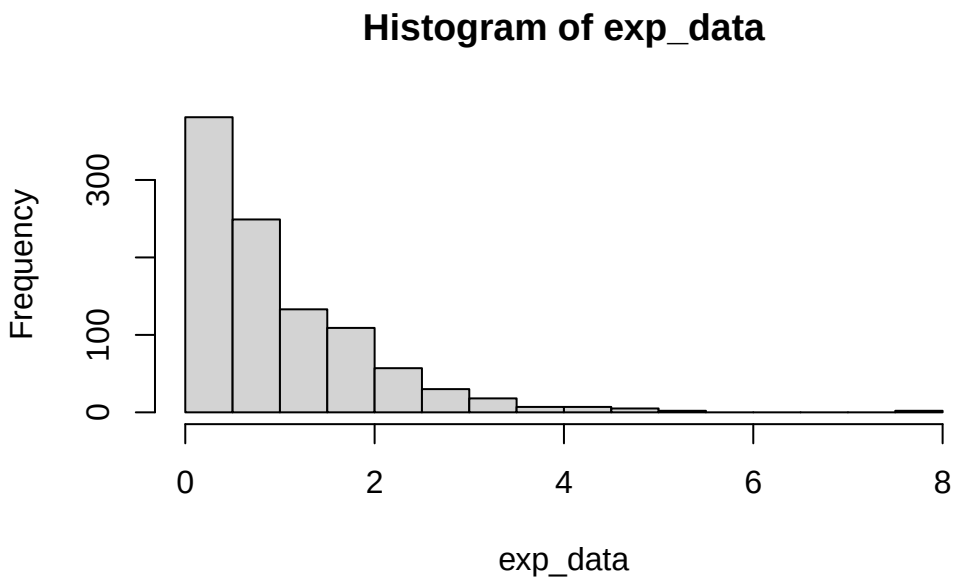
Introduction to the Central Limit Theorem

a.) Using the `rexp()` function, generate 1000 observations from an exponential distribution with `rate = 1` (the default). Save this vector as the variable `exp_data`.

```
exp_data = rexp(1000, rate = 1)
```

b.) Create a histogram of `exp_data`. How would you describe its shape?

```
hist(exp_data)
```



The histogram appears to be heavily right-skewed.

c.) Using the "do it" function, generate 10000 samples of size $n = 10$ from an exponential distribution with `rate = 1`. Compute the mean at each iteration. Save the output of "do it" to the variable `sampling_dist_exp_10`.

```
library(SDS1000)
```

```
Attaching package: 'SDS1000'
```

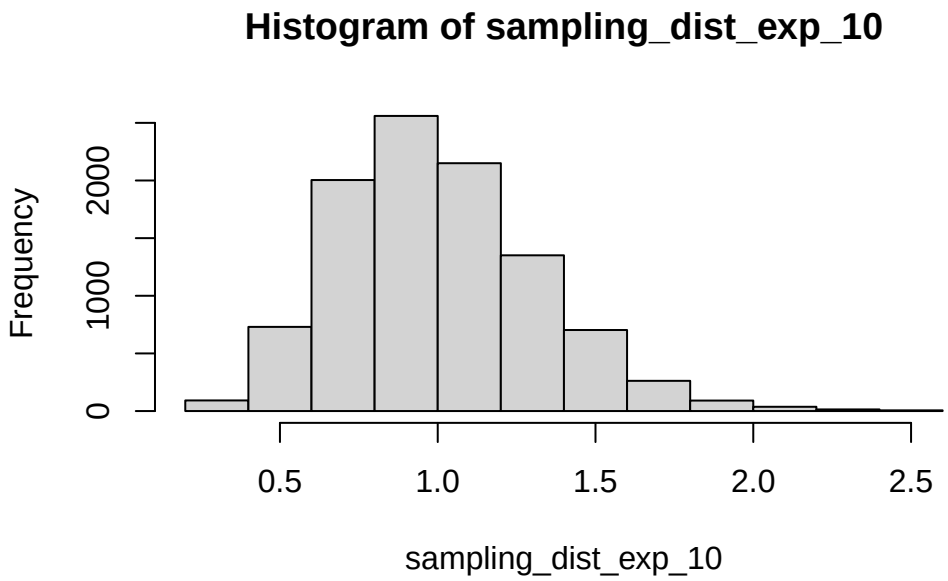
```
The following object is masked from 'package:mosaic':

    shuffle
```

```
sampling_dist_exp_10 = do_it(10000) * {
  curr_sample = rexp(10, rate = 1)
  mean(curr_sample)
}
```

d.) Create a histogram of `sampling_dist_exp_10`. How would you describe its shape?

```r
hist(sampling_dist_exp_10)
```



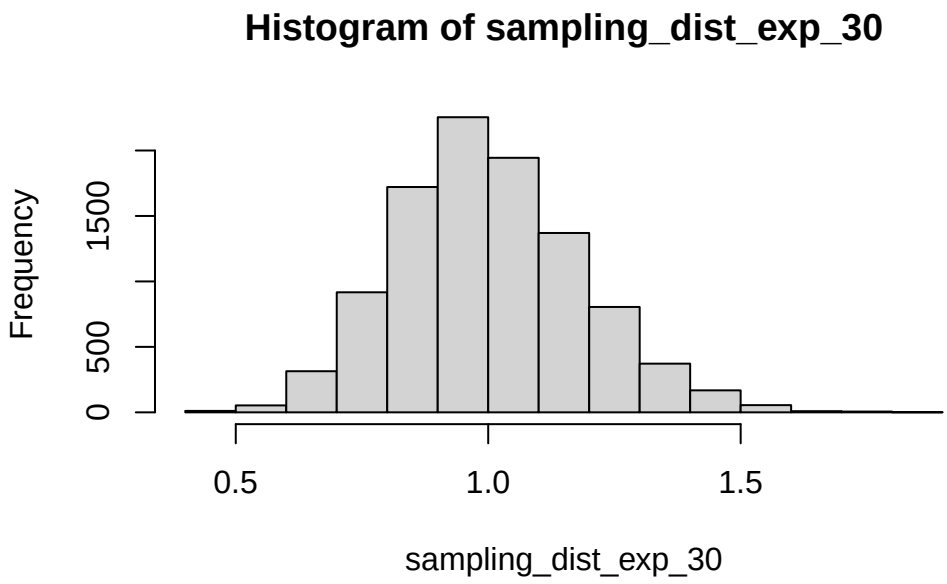
This distribution is somewhat skewed, but appears to be more normal than the original histogram.

e.) Repeat part (c), except now increase the size of each sample to $n = 30$. Save the output of "do it" to the variable `sampling_dist_exp_30`.

```r
sampling_dist_exp_30 = do_it(10000) * {
  curr_sample = rexp(30, rate = 1)
  mean(curr_sample)
}
```

f.) Create a histogram of `sampling_dist_exp_30`. How does it compare to the histogram in part (b)? How does it compare to the histogram in part (d)?

```r
hist(sampling_dist_exp_30)
```

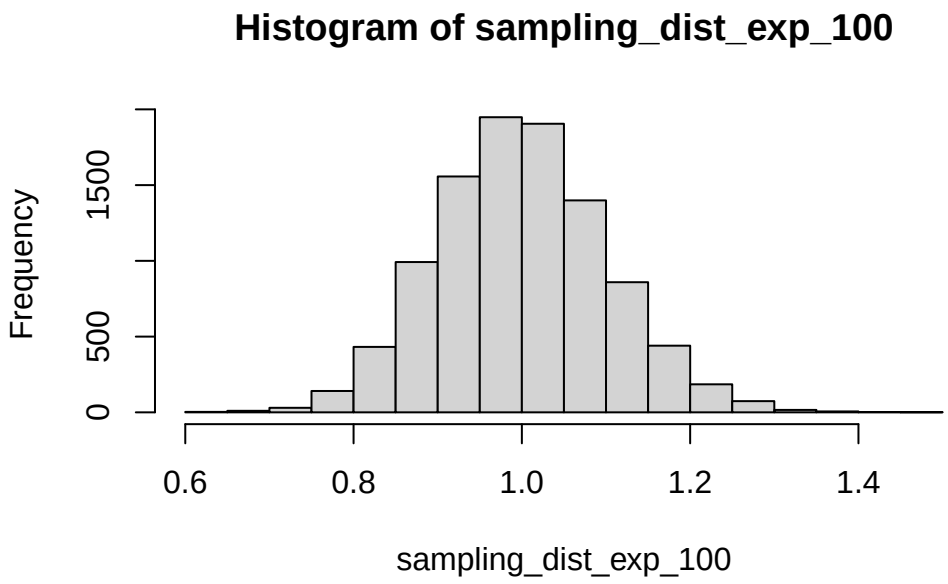


This histogram is much more normal looking than the histogram in part (b) and (d).

g.) Finally, repeat the sampling process for $n = 100$. Comment on the shape of the histogram.

```
sampling_dist_exp_100 = do_it(10000) * {
  curr_sample = rexp(100, rate = 1)
  mean(curr_sample)
}
```

```
hist(sampling_dist_exp_100)
```



This histogram is very normal and much tighter (less spread out) than the histogram in part (f).

Confidence Interval

Survival of ICU Patients in the dataset `ICUAdmissions` includes information on 200 patients admitted to an Intensive Care Unit. One of the variables, Status, indicates whether each patient lived (indicated with a 0) or died (indicated with a 1).

Construct and interpret a 95% confidence interval for the proportion of ICU patients who live. You can find the `ICUAdmissions` data in the `Lock5Data` library.

```
#your code here
```

Answers:

- a.) Load the Data

```
library(SDS1000)
library(Lock5Data)
data(ICUAdmissions)
```

We see in the dataset `ICUAdmissions` that 160 of the patients in the sample lived and 40 died. The sample proportion who live is $\hat{p} = 160/200 = 0.80$.

Since there are more than 10 in the living and the dying groups in the sample, we may use the normal approximation to construct a confidence interval for the proportion who live.

$n\hat{p} = 200 * 0.8 = 160 \geq 10$ and $n(1 - \hat{p} = 200 * (1 - 0.8) = 40 \geq 10$

For 95% confidence the standard normal endpoint is $z^* = 1.96$, so we compute the interval with:

$$0.80 \pm 1.96\sqrt{\frac{0.80(1-0.80)}{200}} = 0.80 \pm 0.055 = (0.745, 0.855)$$

We are 95% sure that the proportion of ICU patients (at this hospital) who live is between 74.5% and 85.5%.

Is B a Good Choice on a Multiple-Choice Exam?

Multiple-choice questions on Advanced Placement exams have five options: A, B, C, D, and E. A random sample of the correct choice on 400 multiple-choice questions on a variety of AP exams shows that B was the most common correct choice, with 90 of the 400 questions having B as the answer.

Does this provide evidence that B is more likely to be the correct choice than would be expected if all five options were equally likely? Show all details of the test. The data are available in `APMultipleChoice` from the `Lock5Data` library.

```
#your code here
```

Answers:

```
library(SDS1000)
library(Lock5Data)
data(APMultipleChoice)
```

the sample proportion of smokers is $\hat{p} = 90/400 = 0.225$.

The hypotheses are:

$$H_0: \quad p = 0.20$$
$$H_a: \quad p \neq 0.20$$

The test statistic:

$$z = \frac{\text{Sample statistic } - \text{ Null paramete}}{SE} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.225 - 0.20}{\sqrt{\frac{0.2(0.8)}{400}}} = 1.25.$$

The p-value

```
z <-(0.225-0.2)/ sqrt( 0.2*0.8/400)
z
```

```
[1] 1.25
```

```
pv <- pnorm(1.25, lower.tail= F)
pv
```

```
[1] 0.1056498
```

This is a two-tail test, so we need to double the p-value. We see that the p -value is $2 * (0.01056) = 0.2112$. This p-value leads us to fail to reject H_0. We could not find strong evidence to conclude that the proportion of smokers is not 20%.

Incentives for Quitting Smoking: Do They Work?

With no incentives, the proportion of smokers trying to quit who are still abstaining six months later is about 0.06. Participants in the study were randomly assigned to one of four different incentives, and the proportion successful was measured six months later. Of the 498 participants in the group with the least success, 47 were still abstaining from smoking six months later. We wish to test to see if this provides evidence that even the smallest incentive works better than the proportion of 0.06 with no incentive at all.

1.) State the null and alternative hypotheses, and give the notation and value of the sample statistic.

2.) Use a randomization distribution and the observed sample statistic to find the p-value.

```
#your code here
```

3.) Give the mean and standard error of the normal distribution that most closely matches the randomization distribution, and then use this normal distribution with the observed sample statistic to find the p-value.

```
#your code here
```

4.) Use the standard error found from the randomization distribution in part (2) to find the standardized test statistic, and then use that test statistic to find the p-value using a standard normal distribution.

```
#your code here
```

5.) Compare the p-values from parts (2), (3), and (4). Use any of these p-values to give the conclusion of the test.

```
#your code here
```

Answers:

1.) This is a test for a single proportion. Using p for the proportion of people who can successfully quit smoking using the incentive, the hypotheses are:

$$\begin{aligned} H_0: &\quad p = 0.06 \\ H_a: &\quad p > 0.06 \end{aligned}$$

The sample statistic is $\hat{p} = 47/498 = 0.094$.

2.) A randomization distribution for proportions in samples of size 498 when $p = 0.06$ is shown below. The proportion of these samples at or above the original sample $\hat{p} = 0.094$ gives a p -value of 0.0016 .

```
#your code here
```

3.) The standard error in the randomization distribution above is $SE = 0.011$ and the null hypothesis is $p = 0.06$, so we use a $N(0.06, 0.011)$ distribution to find the p-value. The area above the original proportion of $\hat{p} = 0.094$ in the figure on the left below gives a p-value of 0.0010 .

```
#your code here
```

4.) The null hypothesis proportion is 0.06 and the standard error from the randomization distribution in (2) is 0.011, so the standardized test statistic is:

$$z = \frac{\text{Sample Statistic - Null Parameter}}{SE} = \frac{0.094 - 0.06}{0.011} = 3.091$$

The area beyond this value in the right tail of a standard normal distribution (shown on the right above) is 0.001, so we have

$$\text{p-value } = 0.001$$

5.) The p-values are very similar, as we expect. In every case, at a 5% level, we reject H_0. We have strong evidence to conclude that the proportion of smokers who quit after incentives is more than 0.06, so incentives appear to help people quit smoking.