Activity 10.4

In a randomized experiment, researchers tested whether adding an additional anticlotting drug, dipyridamole, would be more effective at preventing stroke than aspirin alone.

	Number of patients	Number of strokes
Aspirin alone	1649 n_1	206 X_1
Aspirin + dipyridamole	1650 n_2	157 X_2

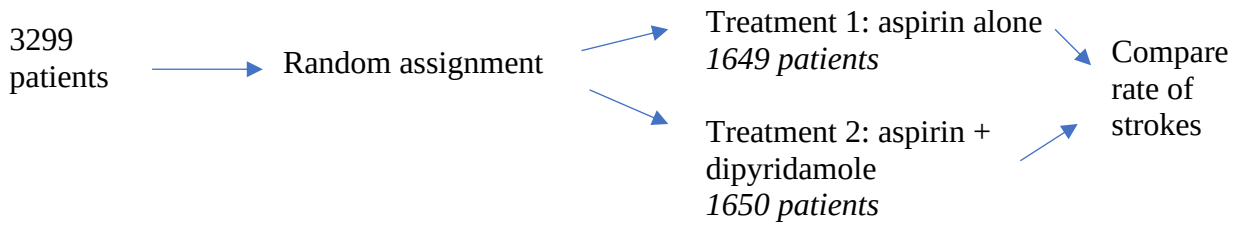
$\hat{p}_1 = \frac{X_1}{n_1}$

$\hat{p}_2 = \frac{X_2}{n_2}$

Describe and sketch the design of the randomized experiment.

We have a total number of 3299 patients who will be randomly assigned to the two treatment groups. Since the sample sizes in both groups are roughly the same, the researchers most likely did not flip a coin to assign each patient to a treatment group, since the coin method would not guarantee each sample sizes for each treatment group. The following method, using a random number generator, allows us to pre-determine the sample size for each treatment group.

1) Label the patients from 1 to 3299.
2) Use a random number generator to output 1649 distinct integers from 1 to 3299.
3) The patients whose labels correspond with the outputted integers are assigned to the aspirin alone group.
4) The remaining 1650 patients are assigned to the aspirin + dipyridamole group.



Define the hypothesis test to determine if adding dipyridamole will significantly reduce the risk of stroke than aspirin alone at the $\alpha = 0.05$ level.

We can compare the rate of strokes in the two treatment groups, using a two-sample z test for proportions.

Let p_1 = proportion of patients in the aspirin alone group who get a stroke
Let p_2 = proportion of patients in the aspirin + dipyridamole group who get a stroke

$$H_0 = p_1 - p_2 = 0$$
$$H_a = p_1 - p_2 > 0$$

The test is one-sided, since we are testing if adding dipyridamole will be *more* effective (i.e., *lower* rate of strokes) than aspirin alone.

Describe the simulation process to estimate the p-value of this test.

Under H_0, we assume that the rate of strokes will be the same in both treatment groups. In other words, under this assumption, regardless of which treatment group each patient was

assigned to, he or she will have the same outcome. In each simulation, we will create a new random assignment of the 3299 patients and assume that their health outcomes remain the same.

The following are the steps of each simulation. Run the code activity_10_4.R and make sure you can match the lines of code with the following procedures.

1) In our experiment, 363 patients had a stroke. We label these patients from 1 to 363.
2) The patients who did not have a stroke are labeled from 364 to 3299.
3) Use a random number generate to output 1649 distinct integers from 1 to 3299. Patients whose labels correspond with these outputted integers are assigned to the aspirin alone group. The remaining 1650 patients are assigned to the aspirin + dipyridamole group.
4) Calculate X_1 = the number of patients in the aspirin alone group who have a stroke. This number will be the number of the outputted 1649 integers that are less than or equal to 363. Thus, we can calculate $\hat{p}_1 = \frac{X_1}{n_1} = \frac{X_1}{1649}$.
5) Calculate X_2= the number of patients in the aspirin + dipyridamole alone group who have a stroke. We know that $X_2 = 363 - X_1$. Thus, we can calculate $\hat{p}_2 = \frac{X_2}{n_2} = \frac{X_2}{1650}$.
6) Calculate the difference in the sample proportions $\hat{p}_1 - \hat{p}_2 = \frac{X_1}{1649} - \frac{X_2}{1650}$.

We repeat the simulation process above many times and plot the histogram of the simulated $\hat{p}_1 - \hat{p}_2$. Suppose we simulate this process 10,000 times. The estimated p-value is the number of simulated $\hat{p}_1 - \hat{p}_2$ that are greater than or equal to our observed $\hat{p}_1 - \hat{p}_2 = \frac{206}{1649} - \frac{157}{1650} \approx 0.0298$ divided by 10,000. This is the last line of code in activity_10_4.R. Compare this estimated p-value with the p-value obtained using 2-PropZTest on your calculator. The results should be very similar.

The simulated histogram represents an approximate **randomization distribution** of $\hat{p}_1 - \hat{p}_2$. We do not use the term **sampling distribution**, because in this case, we do not have independent random samples from two populations. Instead, we have two treatment groups in a randomized experiment. Thus, instead of repeated sampling from two populations, we simulate different possible random assignments for the experiment.

The code in activity_10_4.R will show that the randomization distribution of $\hat{p}_1 - \hat{p}_2$ coincides with the sampling distribution of $\hat{p}_1 - \hat{p}_2$, which has the theoretical approximation

$$\hat{p}_1 - \hat{p}_2 \sim \mathcal{N}\left(0, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

that is further approximated as

$$\hat{p}_1 - \hat{p}_2 \sim \mathcal{N}\left(0, \sqrt{\frac{\hat{p}_c(1-\hat{p}_c)}{n_1} + \frac{\hat{p}_c(1-\hat{p}_c)}{n_2}}\right)$$

under H_0, where $\hat{p}_c = \frac{X_1+X_2}{n_1+n_2}$.