0.1 1e Distribution of p-values

Now, we are going to imagine that we have a bunch of samples (each drawn either from the null distribution or the alternative distribution). We want to predict whether each sample was generated from H_0 or H_1 by looking at its p-value. As a reminder, the two hypothesis to consider are:

The null hypothesis:

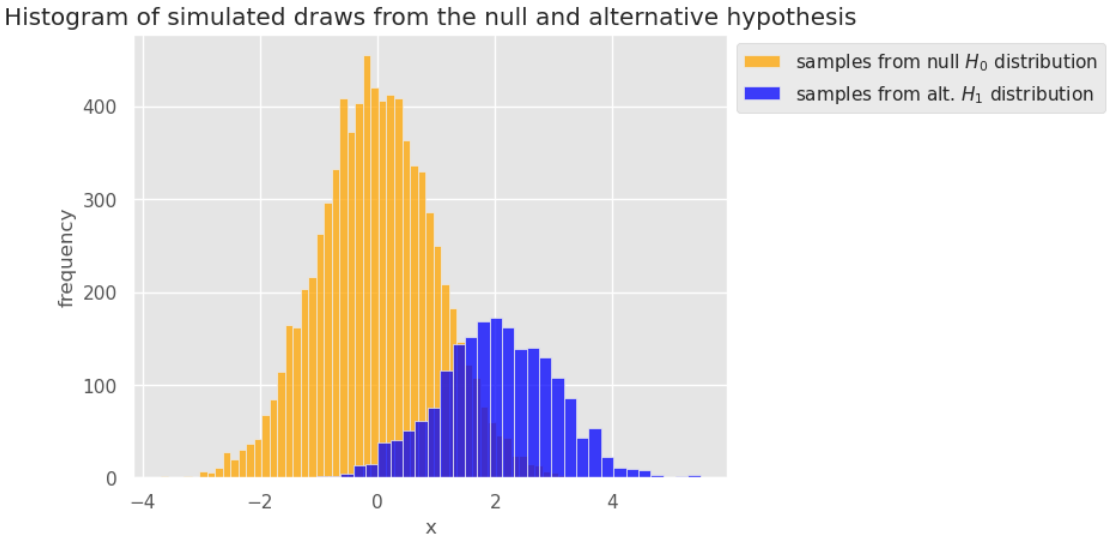
$$H_0 : X \sim \mathcal{N}(0, 1)$$

The alternative hypothesis:

$$H_1 : X \sim \mathcal{N}(2, 1)$$

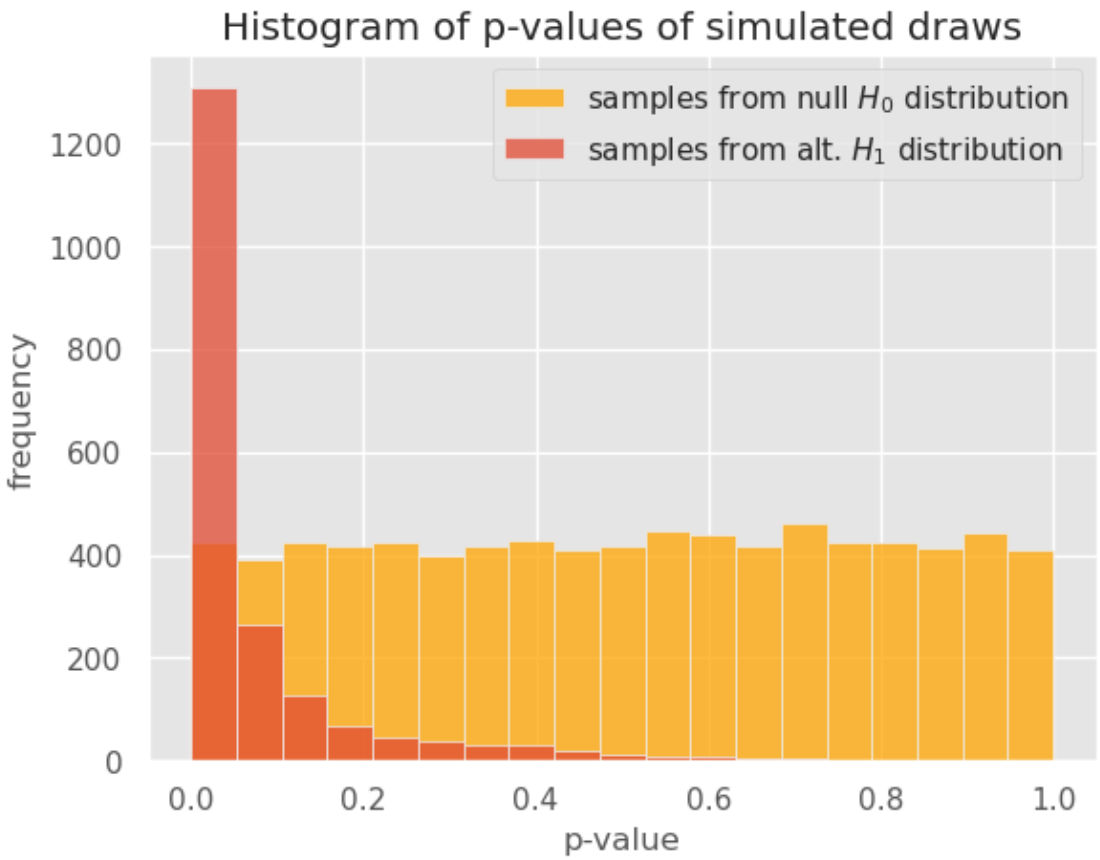
In the example below, we simulate $n = 10000$ draws, in which approximatively 80% come from the null distribution (Reality = 0), and approximately 20% come from the alternative distribution (Reality = 1).

In [21]:
```
# NOTE: you just need to run this cell to instantiate variables; don't change this code.

rs = np.random.RandomState(0)
n = 10000

# roughly 80% of the data comes from the null distribution
# true_values is an n-dimensional array of indicators, where "1" means that x is from the alte
true_values = rs.binomial(1, 0.2, n)

# null distribution is N(0, 1) and alternative distribution is N(2, 1)
x_obs = rs.randn(n) + 2*true_values

sns.histplot(x_obs[np.where(true_values == 0)],  label="samples from null $H_0$ distribution",
sns.histplot(x_obs[np.where(true_values == 1)],  label="samples from alt. $H_1$ distribution",

plt.title("Histogram of simulated draws from the null and alternative hypothesis")
plt.xlabel("x")
plt.ylabel("frequency")
plt.legend(bbox_to_anchor=(1,1));
```



Now, let's take these simulated draws, calculate their p-values, and visualize the p-values corresponding to each of these simulated draws.

In [22]:
```
# NOTE: you just need to run this cell and understand what it does; no code to modify or write
# calculate the p-values for each individual hypothesis
p_values = calculate_p_value(x_obs)

bins = np.linspace(0,1,num=20)
sns.histplot(p_values[np.where(true_values == 0)],  label="samples from null $H_0$ distributio
sns.histplot(p_values[np.where(true_values == 1)],  label="samples from alt. $H_1$ distributio
plt.legend(bbox_to_anchor=(1,1))
plt.title("Histogram of p-values of simulated draws")
plt.xlabel("p-value")
plt.ylabel("frequency");
```



Using the histogram of our calculated p-values, write down a few of your observations ($\leq$ 3 sentences). In particular, your observations should address: 1. The shape of the distribution of p-values under the null hypothesis, 2. how it contrasts with the shape of the p-values under the alternative hypothesis, 3. …and why, for any given p-value close to zero, we're able to state with high confidence that the data point was generated under the alternative hypothesis.

Type your answer here, replacing this text.

0.2 2e Conclusions

Finally, write a short ($\leq$ 4 sentences) summary comparing the three different methods from this problem.

Type your answer here, replacing this text.