Introduction to Statistical Thinking

Neba Nfonsang

University of Denver

Outline

- Statistical Inference
  - Hypothesis testing
  - Estimation
- Statistical Inference in Classical Statistics
  - Classical hypothesis testing
  - One-sample hypothesis testing using a simulation-based approach
  - Two-sample hypothesis testing using a simulation-based approach

Statistical Inference

- One of the main goals of statistics is to make inference.
- Statistical inference involves drawing conclusions about the population using sample data.
- Researchers and scientist prefer to use sample data because it is easier and more cost efficient to collect and manage sample data.
- However, researchers are usually interested in generalizing the results obtained from sample data to the population.

Statistical Inference

- If the analysis data is population data, there is no need for statistical inference.
- Statistical inference therefore involves drawing conclusions and making decisions about the population using sample data.
- There are two main parts to statistical inference:
  - Hypothesis Testing
  - Estimation

Statistical Inference

- Two main frameworks or approaches are used to achieve statistical inference. That is:
  - The frequentist framework (classical or traditional statistics)
  - The Bayesian framework (Bayesian statistics)
- These two approaches differ in several ways, including the goal of the analysis, how the results are obtained and interpreted.
- Though these approaches will sometimes arrive at the same results, the interpretation of the results are quite different. Bayesian analysis seems to be more intuitive.

Statistical (Frequentist) Thinking

- Classical statistics, also know as traditional or frequentist statistics is based on the idea of resampling or simulating several trials.
- For example, hypothesis testing in classical statistics is based on the sampling distribution, which is a theoretical distribution of samples simulated from the population.
- Let's examine the fundamental idea of hypothesis testing in classical statistics.

Classical Hypothesis Testing

- In traditional statistics, hypothesis testing tries to investigate whether an observed sample came from a hypothesized population.
- That is, a hypothesis test examines whether there is sufficient evidence from the data to show that the observed sample is consistent with the null hypothesis.
- A null hypothesis is a claim about the population, and this claim needs to be tested using sample data.

Case Study 1:
Classical Hypothesis Testing

Classical Hypothesis Testing

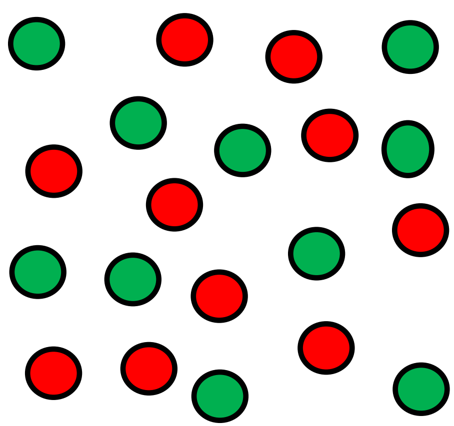
Example

- A florist claims that 50% of seeds produced by plant A are red while 50% are green (null hypothesis).
- How would you test this hypothesis using the classical framework? We need to collect some data and see how likely the data is consistent with the null hypothesis.
- Seeds need to be randomly selected from the population of seeds. Random selection ensures that the sample looks like the population. As such we can generalize the results we find from the sample to the population.

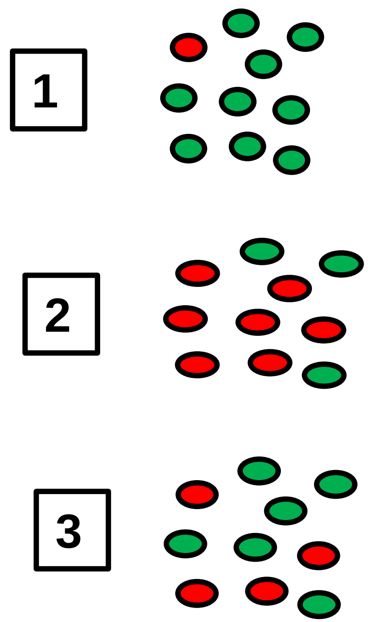
Classical Hypothesis Testing

Resembles the hypothesized population

Which observed sample is more consistent with the null hypothesis?



Hypothesized population with 50% Red, 50% Green



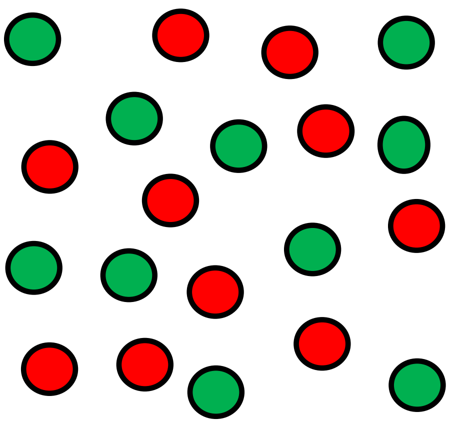
Observed samples

Use the phrase "more likely" or "less likely" to answer the questions below?

If each of these samples (1, 2, and 3) were the observed sample, how likely did the observed come from the hypothesized population (50% Red, 50% Green)?

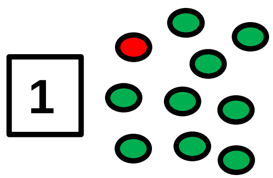
Classical Hypothesis Testing

Which observed sample is more consistent with the null hypothesis?

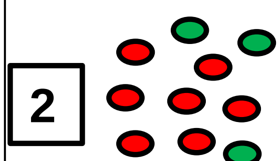


Hypothesized population with 50% R, 50% G

* R=Red, G=Green



This observed sample (10%R, 90%G) is less likely from the hypothesized population as it does not look a lot like the hypothesized population, so we reject the null hypothesis.

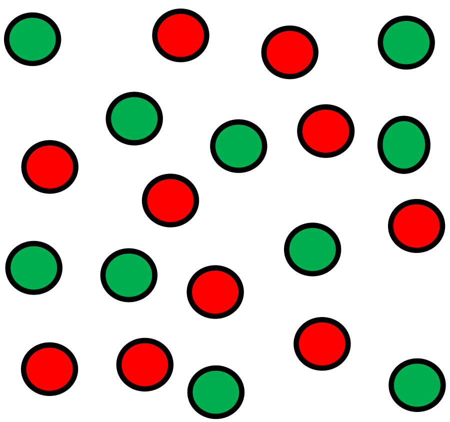


This observed sample (70%R, 30%G) is somewhat more likely to come from hypothesized population, so we fail to reject the null hypothesis.

Observed samples

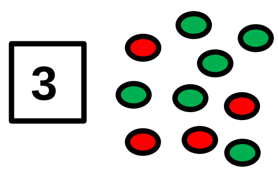
Classical Hypothesis Testing

Which observed sample is more consistent with the null hypothesis?



Hypothesized population with 50% R, 50% G

* R=Red, G=Green



Observed samples

This observed sample (40%R, 60%G) appears to be more likely from the hypothesized population. Therefore, this observed sample is consistent with the null hypothesis. We fail to reject the null hypothesis.

Classical Hypothesis Testing

The intuition of hypothesis testing

- We have seen the intuition of what hypothesis testing in classical statistics is trying to accomplish.
- Hypothesis testing is trying to compare an observed sample to a hypothesized population to understand how likely the observed sample came from the hypothesis population.
- However, the implementation of hypothesis testing in traditional statistics is less intuitive and slightly different from directly comparing an observed sample to a hypothesized population.

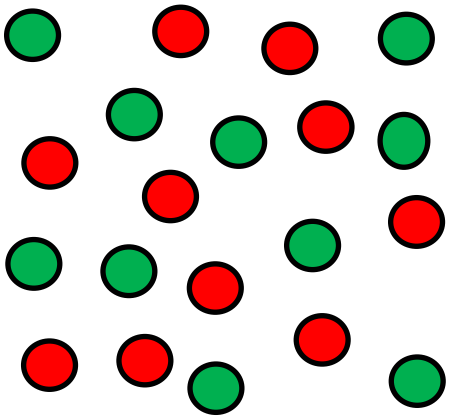
Classical Hypothesis Testing

The intuition of hypothesis testing

- Instead of comparing the observed sample to the hypothesized population directly, classical statistics compares the observed sample to a sampling distribution, theoretically generated from the population.
- A sampling distribution used in hypothesis testing can be viewed as a distribution of samples generated using the parameters of the hypothesized population.
- That means, a sampling distribution is a simulated distribution of samples from the hypothesized population.

Classical Hypothesis Testing

Replace the hypothesized population with a sampling distribution



Hypothesized population with 50% Red, 50% Green

To conduct the hypothesis test, we will now replace the hypothesized population with the sampling distribution (a distribution of simulated samples from the hypothesized population).

Each sample will be represented as the proportion of Red seeds in the sample.

Classical Hypothesis Testing

How to simulate or randomly draw a sample from the hypothesized population

```{r}
set.seed(1234)
sample(size = 10, x = c("R", "G"),
       replace=TRUE, prob = c(.5, .5))

```

```
 [1] "G" "R" "R" "R" "R" "R" "G" "G" "R" "R"
```

x is used to specifies the unique elements in the hypothesized population

prob specifies the proportion (probability of drawing) of each element in the population.

- This code randomly selects 10 seeds independently from the population with replacement, where the probability of selecting a Red or Green seed are the same (=0.5).

Classical Hypothesis Testing

How to simulate or randomly draw a sample from the hypothesized population

```{r}
set.seed(1234)
sample(size = 10, x = c(1, 0),
       replace=TRUE, prob = c(.5, .5))

```

```
[1] 0 1 1 1 1 1 0 0 1 1
```

- Instead of selecting “R” and “G”, let’s replace “R” with 1 and replace “G” with 0. This will help us to calculate the proportion of red in the sample easily.

Classical Hypothesis Testing

How to simulate or randomly draw a sample from the hypothesized population

```
```{r}
set.seed(1234)
sample(size = 10, x = c(1, 0),
       replace=TRUE, prob = c(.5, .5))

```
```

```
[1] 0 1 1 1 1 1 0 0 1 1
```

- What is proportion of Red seeds in the hypothesized population? **.5**
- What is the proportion of Red is the above sample drawn from the hypothesized population? **7/10 = 0.7**

Classical Hypothesis Testing

Compute test statistics after each sample is drawn

```{r}
set.seed(1234)
sample(size = 10, x = c(1, 0),
       replace=TRUE, prob = c(.5, .5))

```

```
[1] 0 1 1 1 1 1 0 0 1 1
```

Computing the proportion of 1's (Red seeds) in binary data taking values 0, and 1 is the same as the mean of the binary data.

The proportion of Red is 0.7.

- So, instead of using the ratio of Red to Green to describe the population or samples, we will instead use the proportion of Red (as test statistics).

Classical Hypothesis Testing

Generating a sampling distribution through resampling

- If we draw another sample from the population, the proportion of Red seeds will not necessarily be 0.3. Each draw is likely going to produce a different proportion.
- We need to repeat the sampling process several times (say 1000 times) and compute the proportion of Red in each of the sample drawn (this is a simulation).
- Compute proportion of Red seeds in each sample. The find the frequency (count) of each proportion. The relationship between the sample proportions and their corresponding counts constitute a sampling distribution.

Classical Hypothesis Testing

Generating a sampling distribution through resampling

```
```{r}
set.seed(1234)
sim.props <- replicate(n = 1000,
                       mean(sample(size = 10, x = c(0, 1),
                                   replace = TRUE, prob = c(.5, .5))))
sim.props
```
```

```
 [1] 0.3 0.6 0.6 0.6 0.4 0.5 0.6 0.6 0.6 0.7 0.9 0.5 0.5 0.3 0.3 0.5
[17] 0.7 0.5 0.5 0.1 0.4 0.3 0.8 0.4 0.4 0.5 0.6 0.5 0.5 0.4 0.4 0.7
[33] 0.7 0.6 0.5 0.3 0.4 0.6 0.7 0.3 0.4 0.4 0.7 0.3 0.1 0.3 0.5 0.6
[49] 0.4 0.3 0.2 0.4 0.5 0.3 0.3 0.5 0.4 0.5 0.5 0.3 0.7 0.6 0.4 0.4
[65] 0.4 0.5 0.4 0.5 0.6 0.6 0.6 0.6 0.6 0.3 0.7 0.3 0.4 0.4 0.6 0.3
```

The code replicates 1000 samples drawn from the hypothesized population, then computes the proportion of Red seeds in each sample. Here are a few of the 1000 sample proportions.

Classical Hypothesis Testing

A sampling distribution of proportions

First, the **table()** function is used to generate a frequency table.

Then, the **as.dataframe()** function is applied on the table to create a data frame

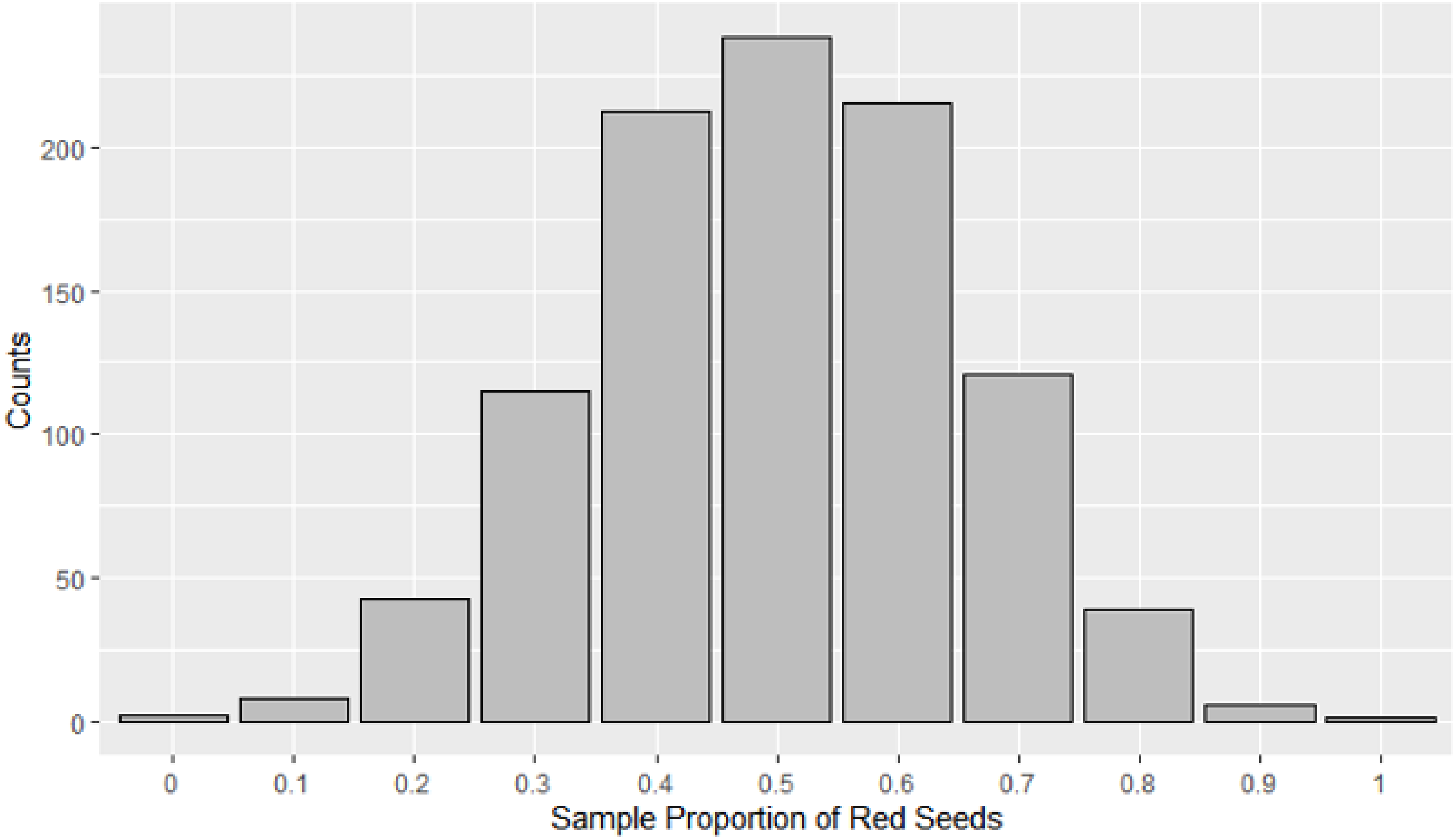
This table of sample proportion vs frequency of proportions constitute the sampling distribution of sample proportion.

```{r}
count.data= data.frame(table(sim.props))
count.data
```

Sample_Proportion <fctr>	Frequency <int>
0	2
0.1	8
0.2	43
0.3	115
0.4	212
0.5	238
0.6	215
0.7	121
0.8	39
0.9	6

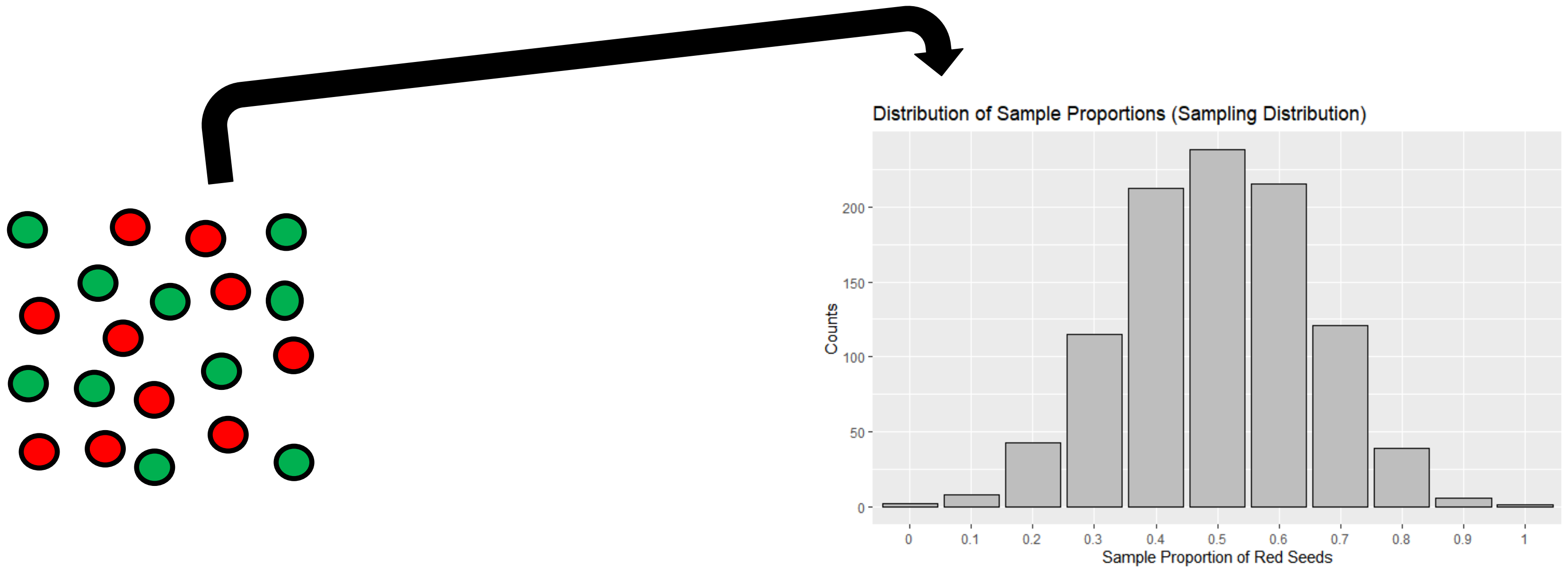
Classical Hypothesis Testing

A visual plot of the sampling distribution of proportions



Classical Hypothesis Testing

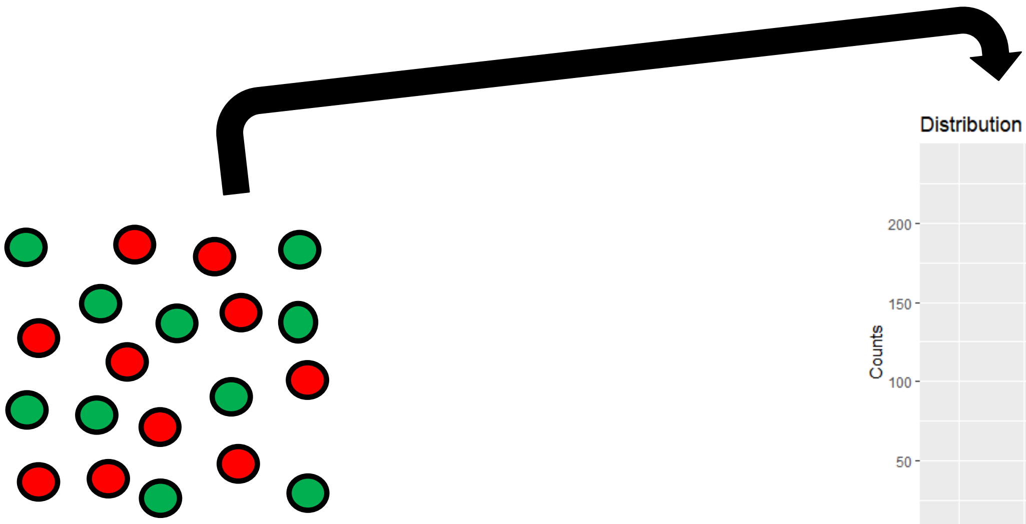
Using the sampling distribution as the null distribution

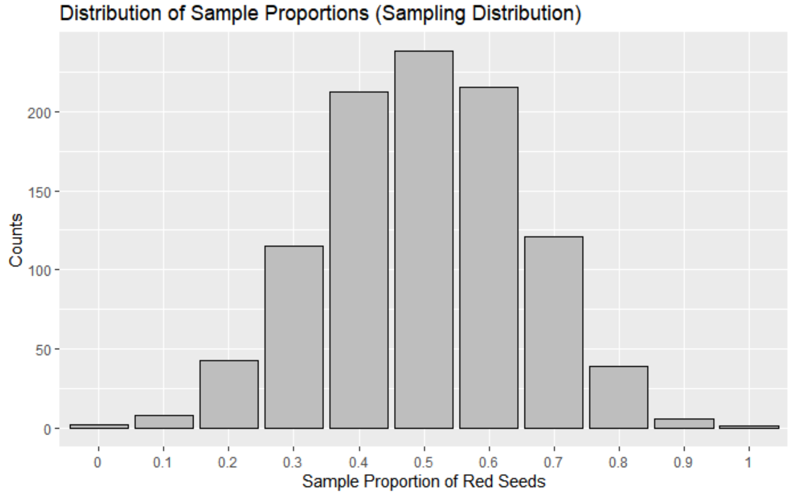


The hypothesized population can now be replaced by a distribution of samples (sampling distribution) theoretically drawn from the hypothesized population. This sampling distribution represents the null distribution.

Classical Hypothesis Testing

The Central Limit theorem





There is a relationship between the hypothesized population and the sampling distribution. This relationship is called the central limit theorem.

Classical Hypothesis Testing

The Central Limit theorem

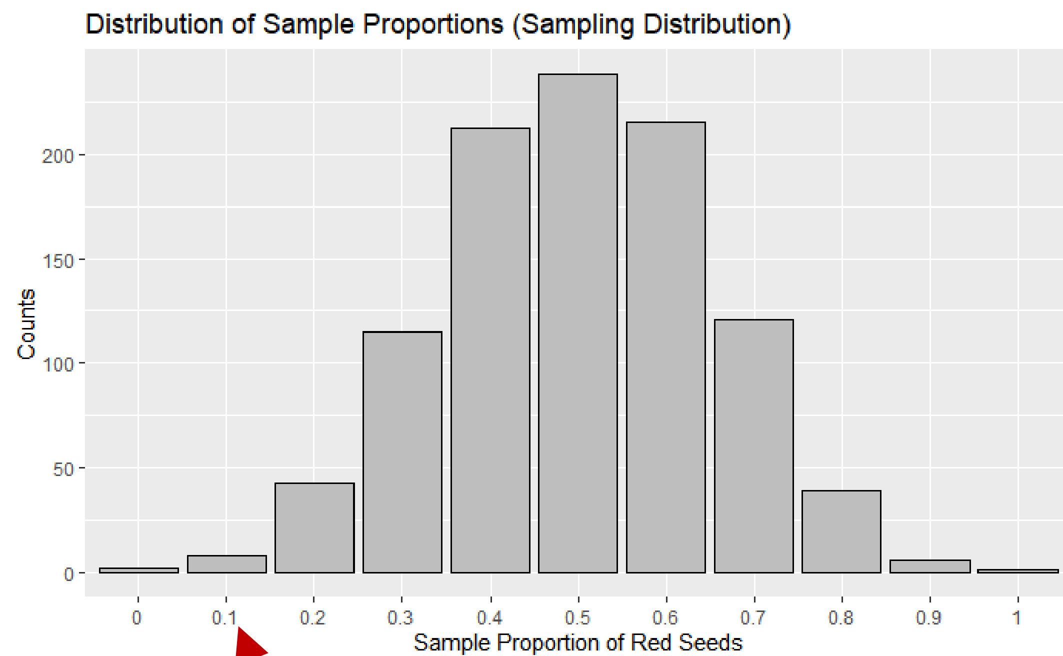
- The central limit theorem says that if we randomly draw samples (with sample size n) several times from a population with mean mu and standard deviation sigma, then the distribution of the sample statistic will be nearly normal with a mean of mu and standard deviation of sigma/sqrt(n), if n is large enough.
- Hence if we can use the characteristics of the population to understand the characteristics of the sampling distribution, which will be used to replace the population.

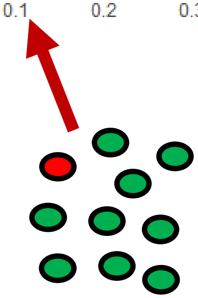
Classical Hypothesis Testing

How likely is the observed sample from the hypothesized population?

Given this null distribution (or null hypothesis), how likely is it to observe a sample with a proportion of 0.1% Red see?

The probability of observing a sample with a proportion of 0.1 under the null distribution is 8/1000 = 0.008. We can conclude that it is less likely to observe this sample given the null hypothesis, hence we reject the null hypothesis.





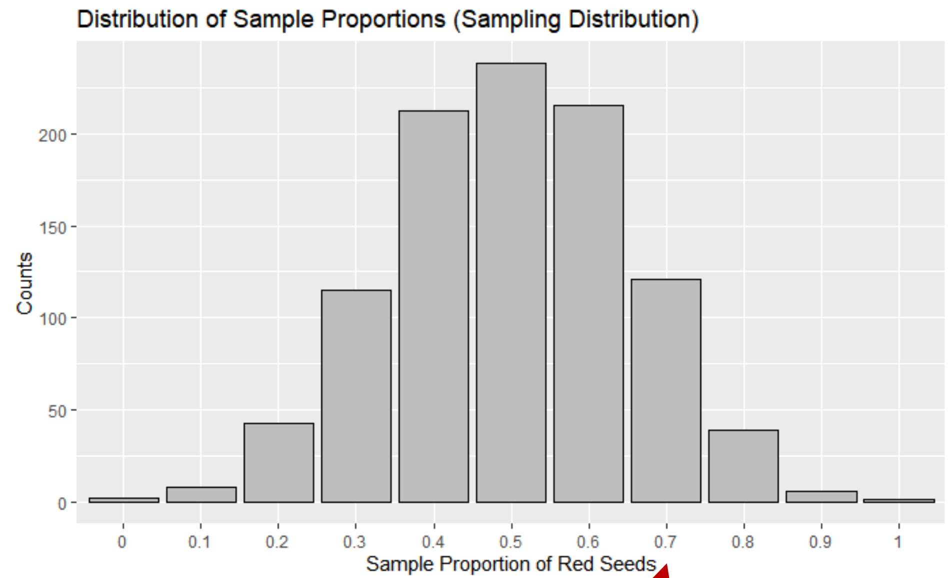
Observed sample

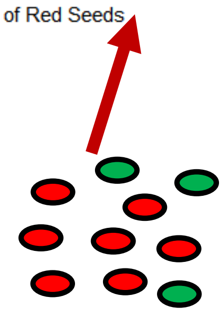
Classical Hypothesis Testing

How likely is the observed sample from the hypothesized population?

Compute how likely it is to observe the sample below given this null distribution.

Therefore, the probability of observing a sample with proportion 0.7 under the null distribution is 121/1000 = 0.121. Hence is more likely to observe a sample with a proportion of 0.7. The observed sample is consistent with the null hypothesis





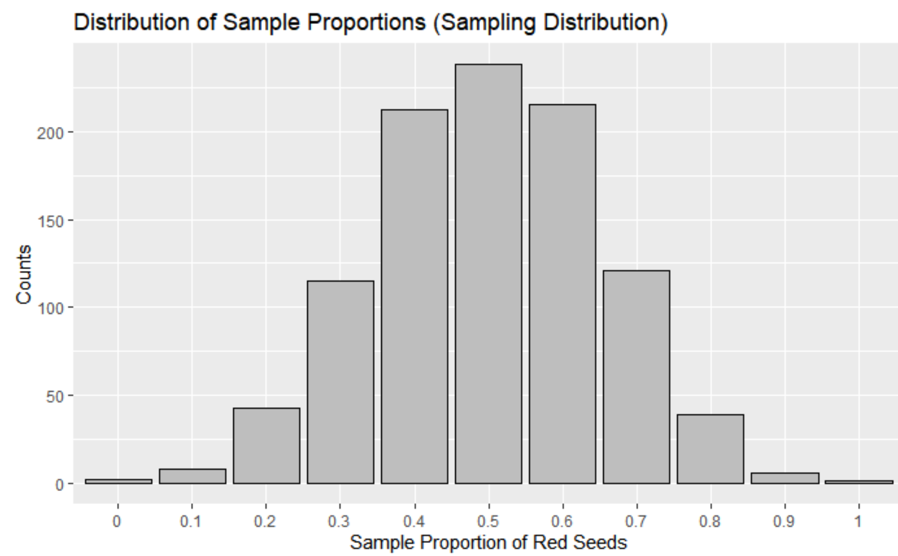
Observed sample

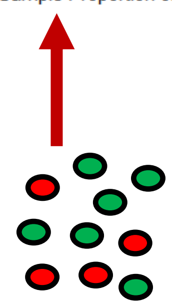
Classical Hypothesis Testing

How likely is the observed sample from the hypothesized population?

Compute how likely it is to observe the sample below given this null distribution.

The probability of observing a sample with proportion 0.4 under the null distribution is 212/1000 = 0.212. It is more likely to observe a sample with a 0.4 proportion of Red seeds given the null hypothesis. Hence, we fail to reject the null hypothesis.



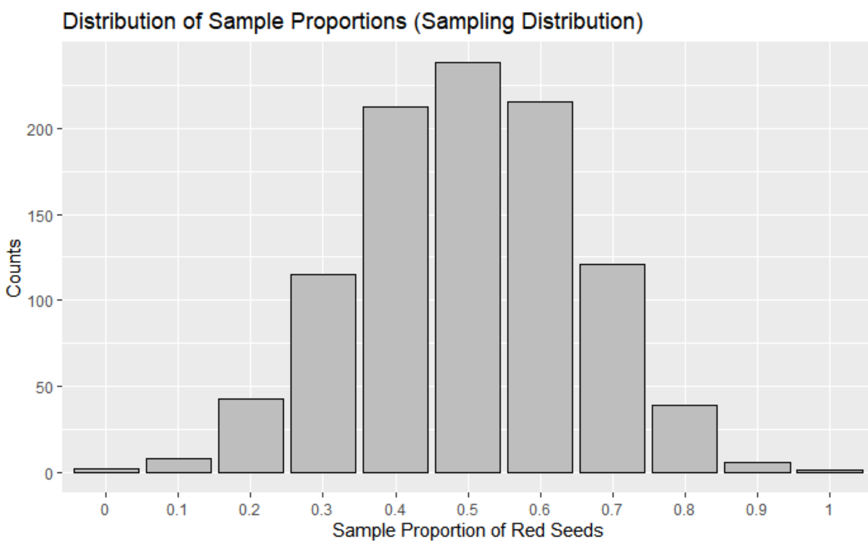


Observed sample

Classical Hypothesis Testing

How likely is the observed sample from the hypothesized population?

- This problem is a nice one because the proportions are discrete or take a finite number of values between 0 and 1.
- Suppose the proportions are continuous as it is in most real-life situations, the count of each proportion (sample) will be nearly zero resulting to a probability of nearly 0.0 for any sample under the null hypothesis.



This problem then leads us to the idea using of p-values instead.

Classical Hypothesis Testing

Dealing with continuous data (issues and the way out)

- So far, we have been using the **relative frequency of observed sample** to compute **how likely it is to observe the sample under the null hypothesis.** This can work fine when the test statistic is discrete (or finite).
- When we are dealing with continuous data as it is in most cases, the relative frequency of each test statistic in the sampling distribution is almost zero, so this will result to relative frequencies of nearly 0.0 for each sample in the sampling distribution, which does not make sense!

Classical Hypothesis Testing

Dealing with continuous data (issues and the way out)

- Instead of using relative frequencies of observed sample under the null distribution**, the relative frequency of the observed sample or samples more extreme than the observed samples** will be used. This is the real definition of p-value.
- Note that **relative frequency** is **probability** and is used to measures plausibility of the observed sample (and other extreme samples) under the null hypothesis.

Classical Hypothesis Testing

P-Value and hypothesis testing

- So, p-value is the probability of observing a sample or more extreme samples under (given) the null hypothesis.
- P-value = P(observed sample or more extreme samples | Null Hypothesis)
- P-value is used to evaluate the strength of evidence from data to understand whether the observed sample is consistent with the null hypothesis.

Classical Hypothesis Testing

Does the frequentist approach mean what it says and says what it mean?

- In the examples we have seen so far to show the intuition behind hypothesis testing, we have used the **probability of the observed sample** under the null hypothesis. This is what the frequentist is really trying to do.
- However, the complication of having a relative frequency being zero for each sample statistic when the test statistics is continuous makes us use p-value instead.
- P-value is however not the same as probability of **probability of the observed sample** under the null hypothesis.

Classical Hypothesis Testing

Does the frequentist approach mean what it says and says what it mean?

- The frequentist approach has been criticized for not meaning what it say. Intuitively the frequentist is trying to evaluate how likely an observed sample is under the null hypothesis.
- In implementing a hypothesis test, the frequentist is instead evaluating how likely an observed sample plus unobserved samples more extreme than the observed sample will exist under the null hypothesis.
- So, let’s see this implementation of hypothesis testing using a p-value.

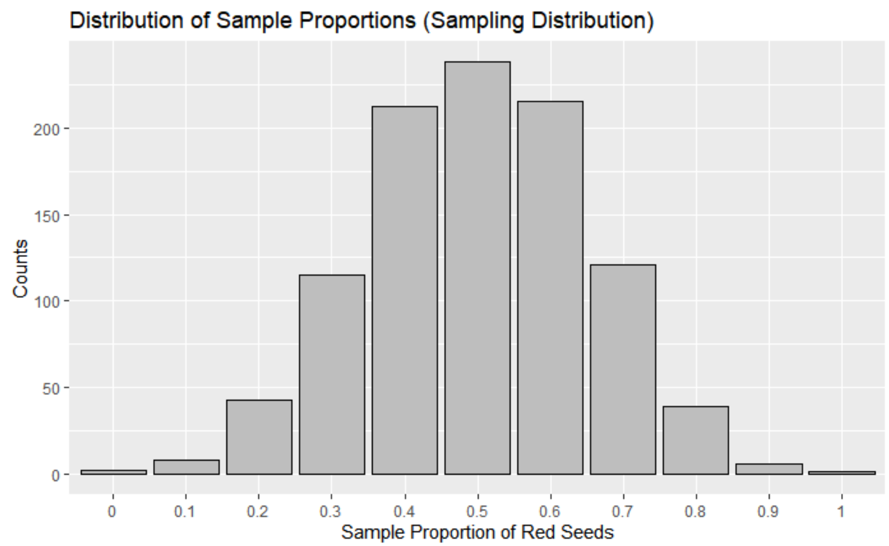
Classical Hypothesis Testing

H1: the proportion of Red seeds produced by plant A is less than 0.5 (lower tail test).

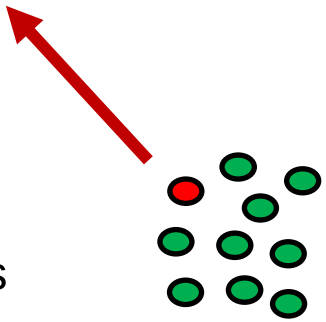
To conduct a lower tail hypothesis test, we need to count the number of extreme samples in the lower tail starting from the observed sample.

For this case, this would be samples with sample proportion of 0.1 and 0.0.

Sample_Proportion <fctr>	Frequency <int>
0	2
0.1	8



H1 = alternative hypothesis



Observed sample

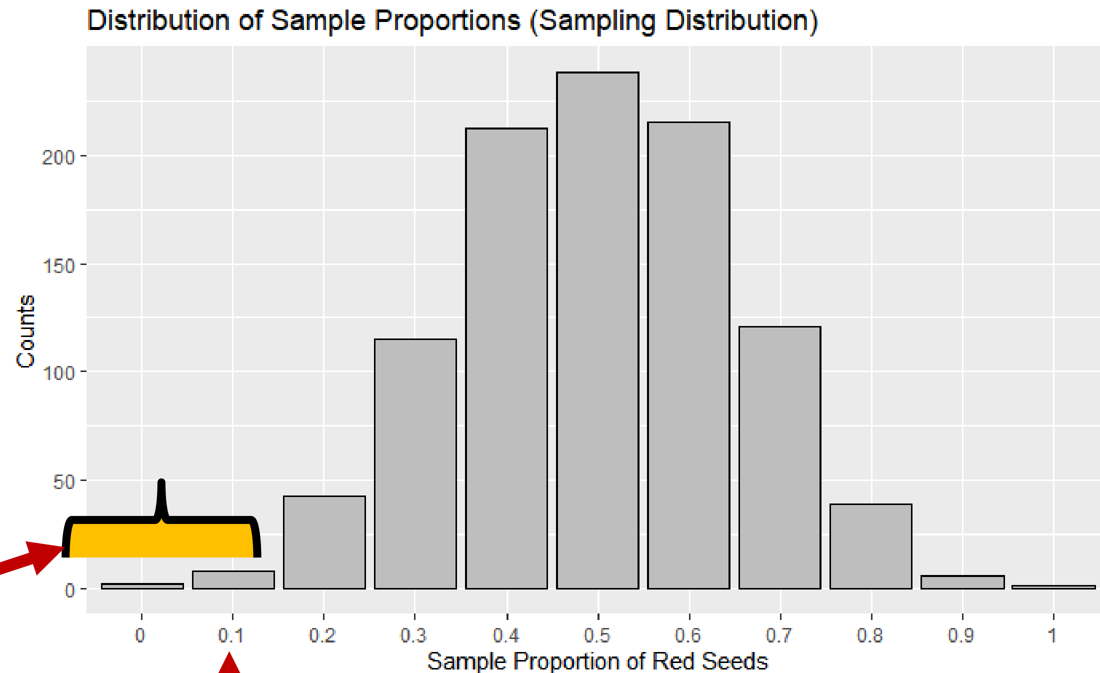
Classical Hypothesis Testing

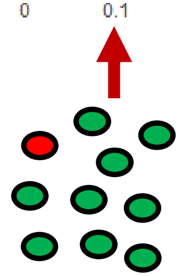
H1: the proportion of Red seeds produced by plant A is less than 0.5 (lower-tailed test).

8 of the observed samples were generated under the null distributions and 2 unobserved samples with a proportion of 0.0 were more extreme than the observed sample.

How likely are these samples?

p-value = (8+2)/1000
= 0.01





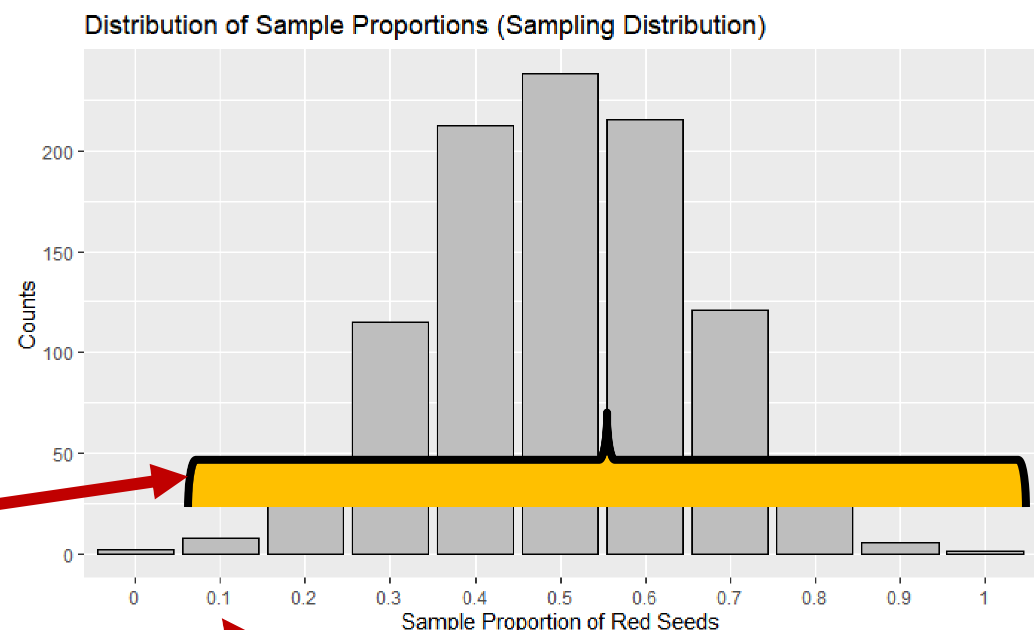
Observed sample

Classical Hypothesis Testing

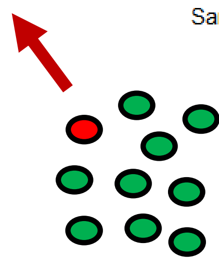
H1: the proportion of Red seeds produced by plant A is greater than 0.5 (upper-tailed test).

For an upper tail hypothesis test, we will be counting the observed sample and other samples above the observed sample under the null hypothesis.

How likely are these samples?



p-value = (8 + 43+115 + 212 + 238 + 215 + 121 + 39 + 6)/1000 = 0.997

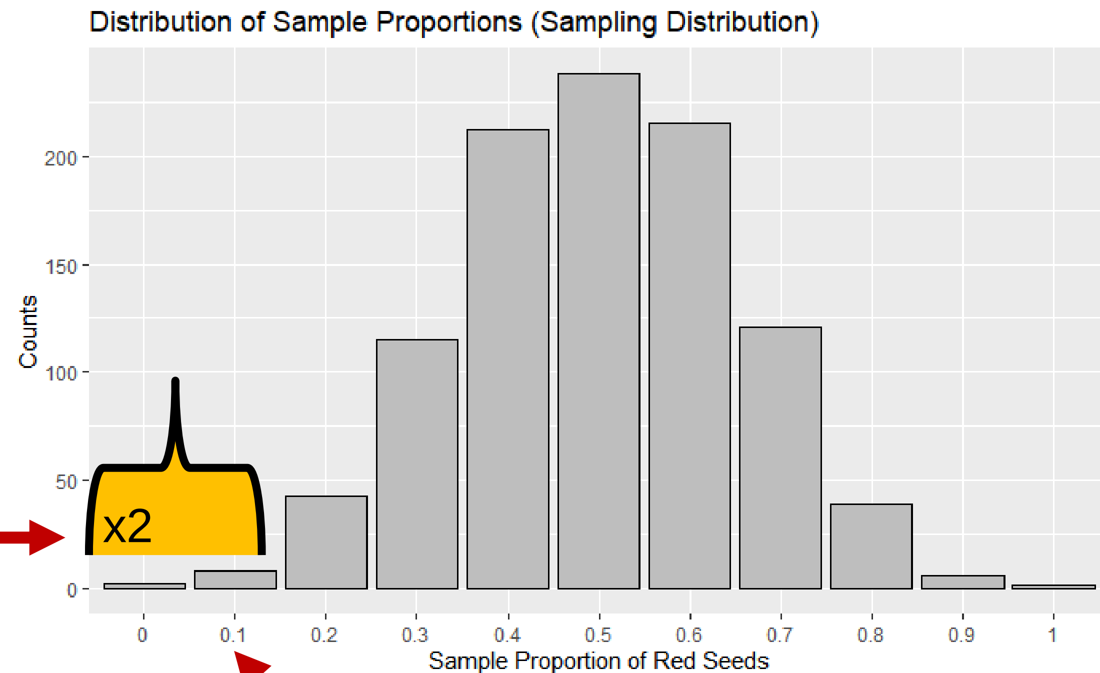


Observed sample

Classical Hypothesis Testing

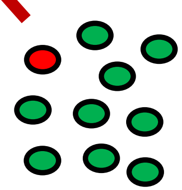
H1: the proportion of Red seeds produced by plant A is not equal to 0.5 (two-tailed test).

For a two-sided hypothesis test, we can compute the upper tail and lower tail p-value and take the one that is smaller and multiply by 2.



How likely are these samples?

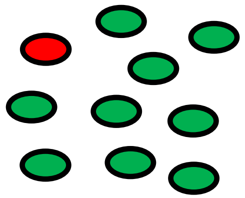
Upper tail p-value = 0.997
Lower tail p-value = = 0.01
Two tail p-value = 0.01x2 = 0.02



Observed sample

Classical Hypothesis Testing

H1: the proportion of Red seeds produced by plant A is not equal to 0.5 (two-tailed test)



Observed sample with x (number of red seeds) = 1 and n (sample size) = 10.

```{r}
prop.test(x = 1 , n = 10, p=0.5, alternative = "two.sided")
```

```
        1-sample proportions test with continuity correction

data:  1 out of 10, null probability 0.5
X-squared = 4.9, df = 1, p-value = 0.02686
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.005242302 0.458846016
sample estimates:
  p
0.1
```

This is a one-sample hypothesis test of proportion in R. The two-sided p-value here is similar to the one we obtained for the simulation.

Case Study 2:
Classical Hypothesis Testing

Classical Hypothesis Testing

Observed data

	decision		
	promoted	not promoted	Total
male	18	6	24
female	17	7	24
Total	35	13	48

You are trying to examine whether promotion decisions at your workplace are gender biased. You randomly selected 48 employees (24 male and 24 female) to conduct a study. The sample data is tabulated as shown above.

Classical Hypothesis Testing

Research question

Is the proportion of promoted males significantly different from the proportion of promoted females?

	decision		
	promoted	not promoted	Total
male	18	6	24
female	17	7	24
Total	35	13	48

Classical Hypothesis Testing

Research question and null hypothesis (H0).

- Research question:
  - Is the proportion of promoted males significantly different from the proportion of promoted females?
- Null hypothesis:
  - The proportion of promoted males is the same as the proportion of promoted female?
- The null hypothesis could also be stated as:
  - The difference between the proportion of promoted males and proportion of promoted females is zero.

Classical Hypothesis Testing

Observed statistics: difference between male and female promotion rates

		decision		
		promoted	not promoted	Total
gender	male	21	3	24
	female	14	10	24
	Total	35	13	48

Observed proportion of promoted males = 21/24 = 0.875
Observed proportion of promoted females = 14/24 = 0.583
Observed difference in proportion of promoted males and females = 0.75 – 0.71 = 0.292

Classical Hypothesis Testing

Generate the simulated or sampling distribution of test statistic

```
```{r}
sim.diff <- replicate(1000,
                      mean(sample(size = 24, x = c(0, 1), replace = TRUE, prob = c(.5, .5))) -
                      mean(sample(size = 24, x = c(0, 1), replace = TRUE, prob = c(.5, .5))))
sim.diff
```
```

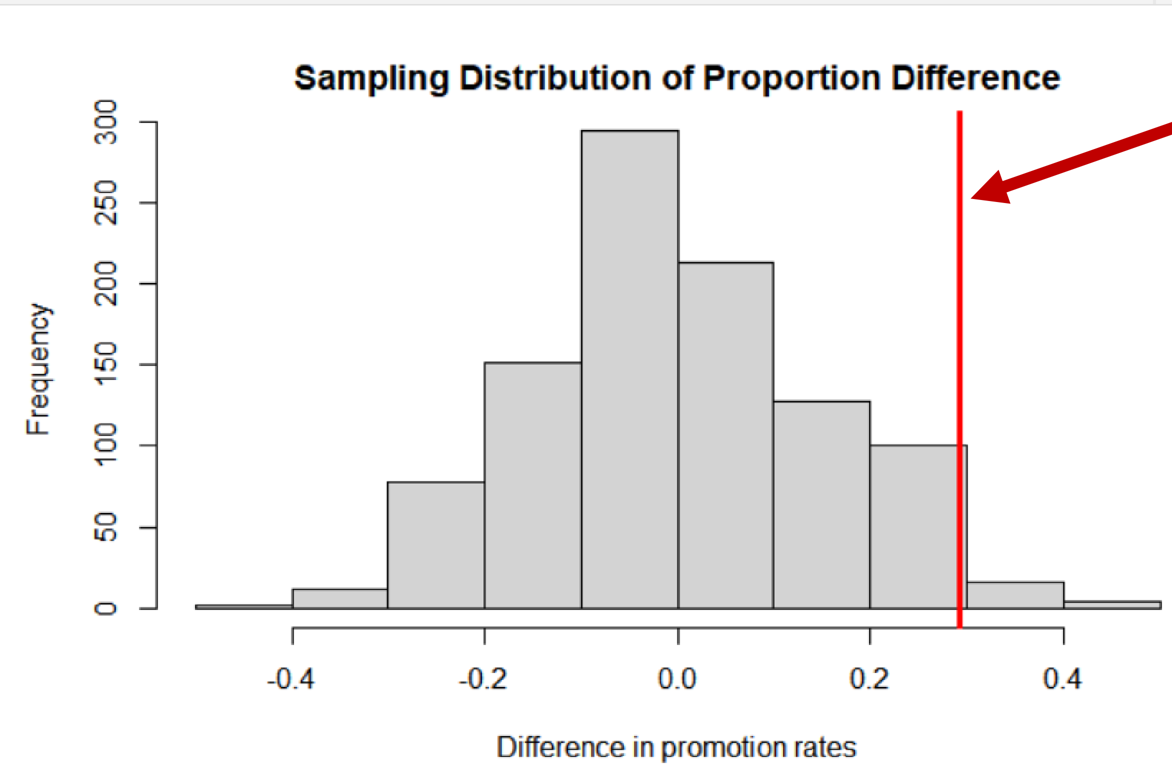
```
   [1] -0.20833333 -0.20833333 -0.25000000  0.16666667  0.16666667  0.04166667  0.16666667
   [8] -0.04166667  0.12500000  0.00000000 -0.29166667 -0.12500000  0.04166667 -0.04166667
  [15] -0.37500000  0.04166667 -0.20833333 -0.08333333  0.20833333  0.16666667 -0.25000000
  [22]  0.04166667 -0.08333333  0.08333333  0.20833333  0.00000000 -0.08333333  0.25000000
  [29] -0.08333333  0.00000000 -0.04166667 -0.12500000 -0.04166667  0.08333333  0.25000000
```

Since we are hypothesizing that male and female promotion rates are equal, we will randomly draw 24 male and 24 female from a population where the promotion rates of male and female are equal, compute the difference in promotion rates and repeat the process.

Classical Hypothesis Testing

Sampling distribution of difference in male and female proportions

```
```{r}
# observed difference
obs.diff = 0.292
hist(sim.diff, xlab = "Difference in promotion rates",
     main = "Sampling Distribution of Proportion Difference" )
abline(v=obs.diff, lw=3, col="red")
```
```



The observed difference seems to be far from the hypothesized zero difference

Classical Hypothesis Testing

Computing two-sided p-value

```
## Compute the two-sided p-value
```{r}
mean (sim.diff>=obs.diff)*2
```

[1] 0.04
```

The upper tail p-value is smaller, so we multiply it by 2 to get the two-sided p-value.

A very small p-value such as 0.04 indicates that the observed sample (including other extreme samples than the observed sample) is less likely to be observed under the null hypothesis so reject the null hypothesis. It appears there is discrimination against women when it comes to promotion decision as promotion rates for females are much lower than those for males.

Exercise

- You found a claim on your University's website that the average salary of graduates from the University is $70, 000 (with a standard deviation of $40,000). You decided to test this claim, so you randomly collected data from 30 graduates of the University.
- State the test statistics and state the null hypothesis
- Generate the sampling distribution
- Visualize the sampling distribution and the observed sample
- Compute a two-side p-value and interpret the p-value
- Draw a conclusion about your hypothesis test (see code snippet on the next slide)

Exercise (continues…)

Use this code to generate 1000 samples from the hypothesized population and compute the mean of each sample. Visualize the distribution and the observed sample, then compute the p-value, interpret the p-value and draw your conclusion

```
## Test claim about salaries of graduate students

```{r}
replicate(1000, mean(rnorm(n=30, mean=70000, sd=40000)))
```

 [1] 72891.55 74039.01 72394.61 70842.04 71953.63 65232.20 63100.04 57908.74 73587.59 68549.77
[11] 73330.09 72074.11 73770.89 74570.71 62201.03 65870.10 62395.61 64315.16 74785.40 75440.41
[21] 70964.15 65086.62 72790.51 76721.79 60390.51 70079.97 64521.36 67840.45 72532.76 73268.45
[31] 72007.62 58115.85 59952.70 74625.61 63528.20 72015.91 76077.96 84667.22 65059.98 65283.67
[41] 72424.23 61076.60 69942.37 80063.91 85220.53 65910.81 83563.70 76455.02 59088.88 70557.15
[51] 89241.92 72369.22 77062.08 66912.31 76859.80 65997.40 78337.77 69963.00 72469.87 66553.30
[61] 71468.81 77126.20 65283.38 69566.78 74723.17 78204.98 78901.93 76315.72 69126.29 87840.94
```

These are just a few out of 1000 sample means

Conclusion

- We have examined the foundation of statistical thinking from the frequentist perspective.
- The frequentist approach to inference is based on the idea resampling or replication.
- Resampling is used to generate a sampling distribution for hypothesis testing. The observed sample is examined under the null distribution to determine the plausibility of the observed sample under the null distribution.
- The p-value is intended to evaluate the plausibility of a sample but in practice, p-value is evaluating the plausibility of the observed sample and other extreme samples.

Conclusion

- Small p-values indicate that:
  - The observed sample is less likely under the null hypothesis
  - It is less likely that the observed sample belongs to the hypothesized population.
  - There is strong or sufficient evidence that the observed sample is inconsistent with the null hypothesis.
  - The null hypothesis needs to be rejected in support of the alternative hypothesis.