Hypothesis testing (one proportion): the Wheel of Destiny

Is there something really going on, or is it just luck?

STA130 Week 4

1. State hypotheses

In the Wheel of Destiny problem, we started with a question: "Is the new Wheel of Destiny

Two hypotheses:

Null hypothesis (H_0): The Wheel of Destiny spinner is fair

$$H_0 : p_{red} = 0.5$$

Alternative hypothesis (H_A or H_a or H_1): The Wheel of Destiny spinner is not fair

$$H_1 : p_{red} \neq 0.5$$

where p_{red} is the proportion of spins of the new Wheel of Destiny that land on red (if we spun it infinitely many times).

2. Calculate the test statistic from observed data

```
test_stat <- 32/50
```

3. Simulate samples assuming H_0 is true and calculate the statistic for each sample

Goal: Explore the distribution of values of the statistic (in this case, $\hat{p}_{red}$) we would observe if H_0 was true. What kind of results are common under the null? Which are unusual?

Simulation is a way to explore random events.

Previously, we used people-power to simulate values of $\hat{p}_{heads}$ (or equivalently $\hat{p}_{red}$) under the assumption of a 50/50 process (H_0)

We can use R to simulate values more quickly!

Simulating coin flips with R to get a sampling distribution

```
set.seed(5)
```

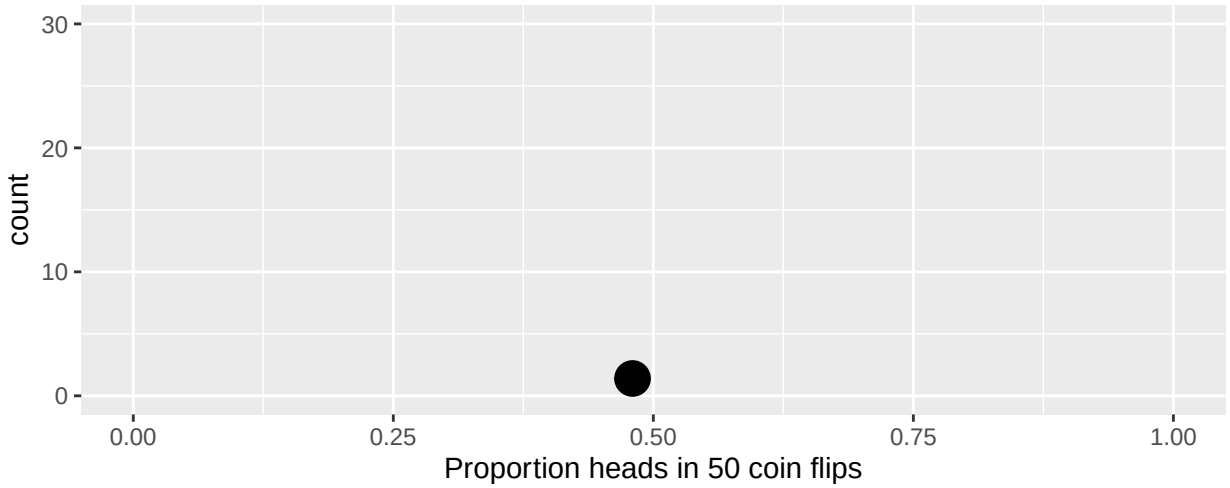
What will the following lines of R code do?

```
sum(flips == "heads")
sum(flips == "heads") / 50
mean(flips == "heads")
```

```
sim <- tibble(p_heads = mean(flips=="heads"))
print(as.numeric(sim))
```

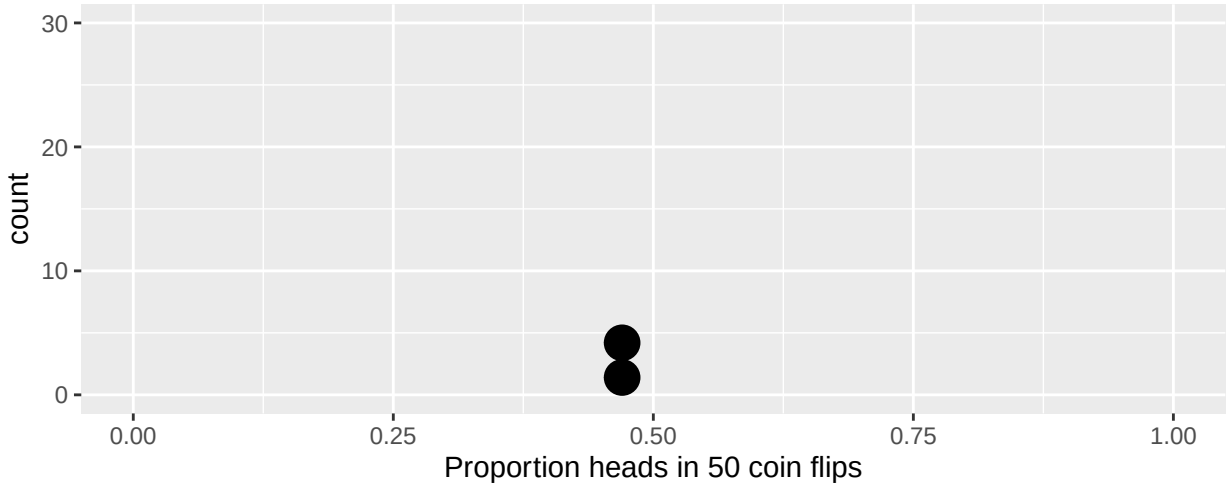
```
## [1] 0.48
```

```
sim %>% ggplot(aes(x=p_heads)) +
  geom_dotplot() + xlim(0, 1) + ylim(0,30) +
  labs(x="Proportion heads in 50 coin flips")
```



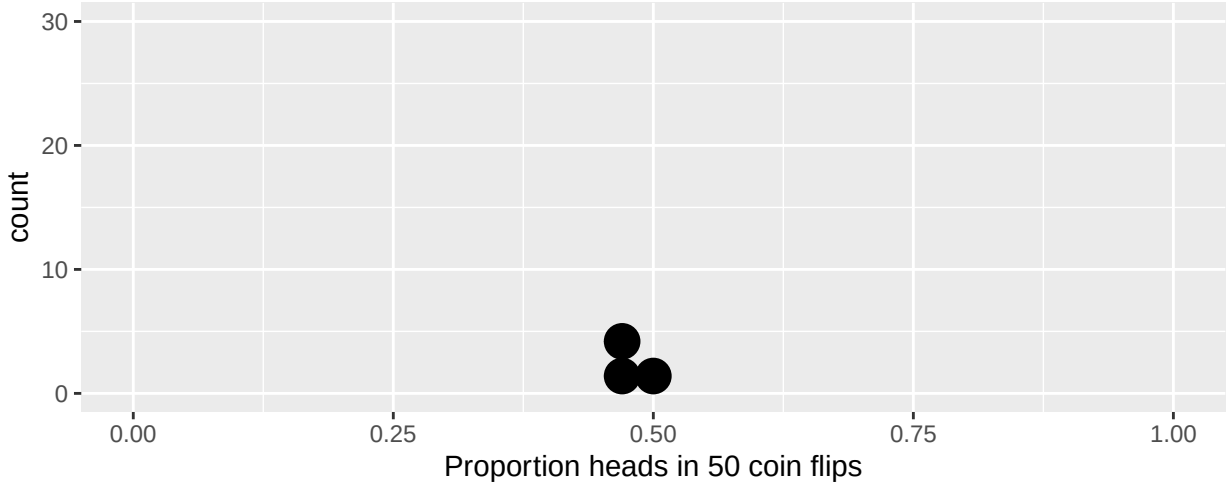
Add another simulation

```
## [1] 0.46
```



And another simulation

```
## [1] 0.5
```



We could keep going to build the sampling distribution (from simulation)... But we don't want to have to keep doing each simulation by hand!

Use `for` loops to generate many simulations

- Automate the process of generating many simulations
- Evaluate a block of code for each value of a sequence (for example, 1, 2, 3, ... 1000)
- The following `for` loop will evaluate SOME CODE 1000 times, for `i=1` and `i=2` and ... and `i=1000`
  - Note that code is within *curly* brackets

```
for (i in 1:1000)
{
  SOME CODE
}
```

####3A Set values for simulation.

```
n_observations <- 50 # number of observations (e.g. coin flips or spins)
repetitions <- 1000 # 1000 simulations
simulated_stats <- rep(NA, repetitions) # 1000 missing values to start
```

####3B. Automate simulation with a `for` loop.

```
for (i in 1:repetitions){
  new_sim <- sample(c("red", "black"),
                    size = n_observations,
                    prob = c(0.5, 0.5),
                    replace = TRUE)
  sim_p <- sum(new_sim == "red") / n_observations
```

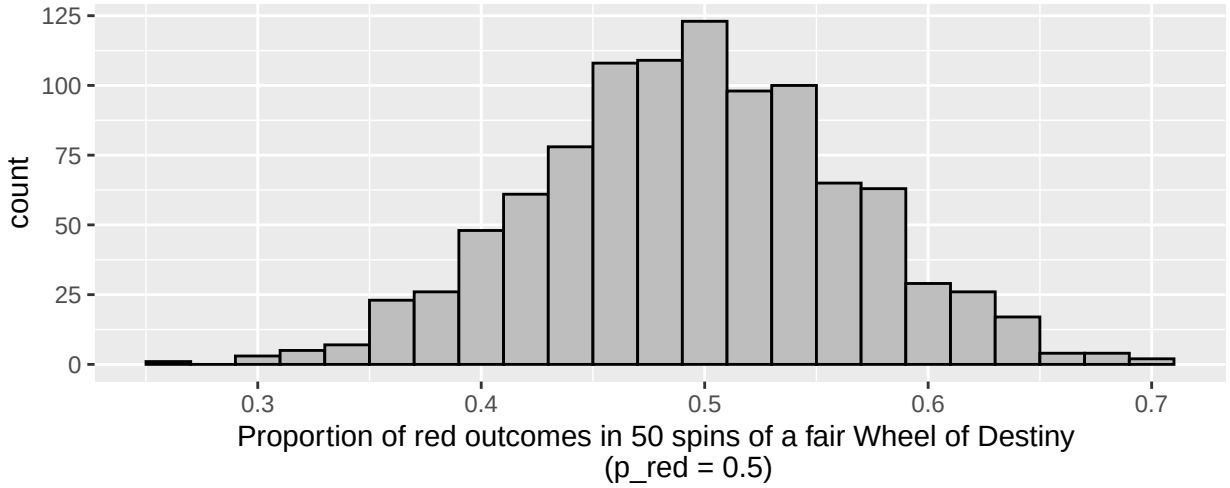
```
  simulated_stats[i] <- sim_p; # add new value to vector of results
}
```

```
sim <- tibble(p_red = simulated_stats)
```

3C Turn results into a data frame so we can use ggplot for plotting

3D Plot results Although samples are selected at random (so the value of the statistic is different for each sample) the distribution of all possible values of the statistic (i.e. the sampling distribution) has specific features.

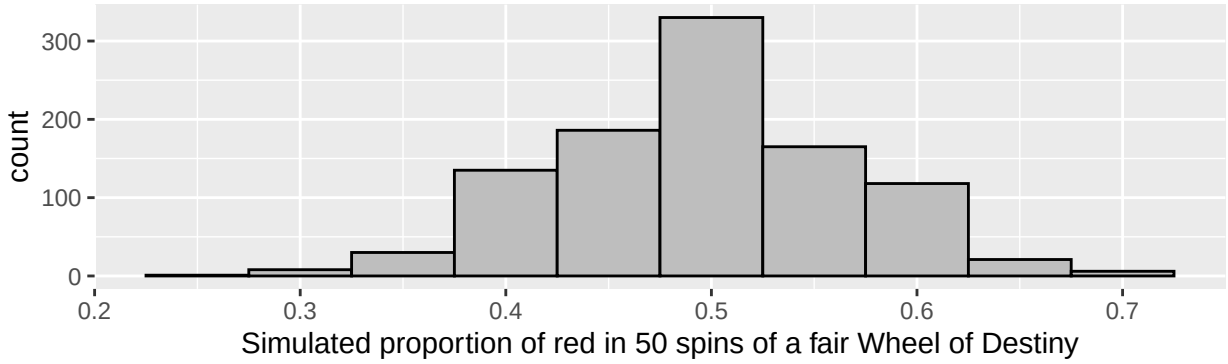
```
sim %>% ggplot(aes(x = p_red)) +
  geom_histogram(binwidth = 0.02, colour = "black", fill = "grey") +
  xlab("Proportion of red outcomes in 50 spins of a fair Wheel of Destiny
       (p_red = 0.5)")
```



4. Evaluate the evidence against H_0

I spun the Wheel of Destiny 50 times. In my sample (observed, not simulated!) I saw: $\hat{p}_{red} = 32/50 = 0.64$ (the proportion of our 50 spins which landed on red)

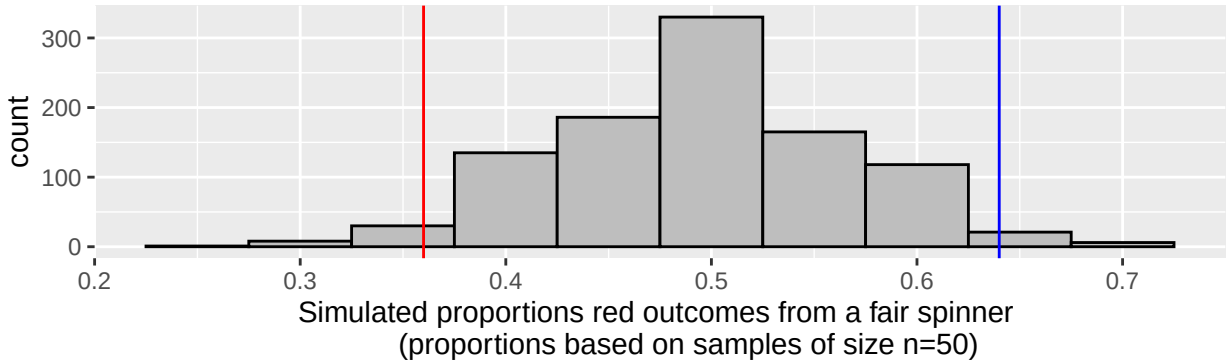
If the spinner really is fair, is our observed value of $\hat{p}_{red} = 0.64$ unusual? How unusual?



- The **p-value** is the probability of observing data that are **at least as unusual** (or **at least as extreme**) as the sample data, *under the assumption that* H_0 *is true.*
- We estimate the p-value as the proportion of values in the estimated sampling distribution that are as extreme or more extreme than the test statistic calculated from our observed sample data.
- For the *Wheel of Destiny* example:
  - Null hypothesis value: $p_{red} = 0.5$
  - Observed estimtate from the sample: $\hat{p}_{red} = \frac{32}{50}$
  - Values at least as extreme/unusual as the sample statistic: all values **greater or equal to** $\hat{p}_{red}$ and all values **less than or equal to** $0.5 - |\hat{p}_{red} - 0.5|$
    - i.e. values further away from the null value ($p_{red} = 0.5$) than the test statistic is (i.e. further than $|\hat{p}_{red} - 0.5|$ away from $p_{red} = 0.5$)
- This is a **two-sided test** because it considers differences from the null hypothesis that are both larger and smaller than what you observed.
  (It is also possible to carry out one-sided tests. They are useful in some specific applications.)

What proportion of simulated values are at least as far from the null value as $\hat{p}_{red}$?

```
test_stat <- 32/50
sim %>% ggplot(aes(x=p_red)) +
  geom_histogram(binwidth = 0.05, colour = "black", fill = "grey") +
  geom_vline(xintercept = 0.5  - abs(0.5 - test_stat), color = "red") +
  geom_vline(xintercept = 0.5 + abs(0.5 - test_stat), color = "blue") +
  labs(x = "Simulated proportions red outcomes from a fair spinner
       (proportions based on samples of size n=50)")
```



```
pvalue <- sim %>%
  filter(p_red >= 0.64 | p_red <= 0.36) %>%
  summarise(p_value = n() / repetitions)

as.numeric(pvalue)
```

```
## [1] 0.066
```

Another way to calculate the p-value is

```
pvalue <- sim %>%
  filter(abs(p_red - 0.5) >= abs(test_stat - 0.5)) %>%
  summarise(p_value = n() / repetitions)
pvalue
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.066
```

A small p-value tells us that there is only a small chance that we would observe a test statistic as far away from the null value of the parameter if H_0 were really true

Two reasons that can lead to a small p-value:

1. H_0 is actually true and we just observed an *unlikely extreme value* of the statistic
2. H_0 is not true

The smaller the p-value, the more we lean towards (2) - in other words, the smaller the p-value, the more "evidence" we have against H_0

5. Make a conclusion

.normalsize[Some guidelines for how small is small? This table tells you how to comment on the **strength of evidence against H_0**.

P-value	Evidence
p-value > 0.10	no evidence against H_0
$0.05 <$ p-value < 0.10	weak evidence against H_0
$0.01 <$ p-value < 0.05	moderate evidence against H_0
$0.001 <$ p-value < 0.01	strong evidence against H_0
p-value < 0.001	very strong evidence against H_0

]

Conclusion (based on $\hat{p}_{red} = \frac{32}{50} = 0.64$): Since the p-value is 0.066, we conclude that we have ***weak evidence against the null hypothesis*** that the Wheel of Destiny spinner is fair.

Sometimes, you'll see some conclusions talking about *statistical significance* (or a *statistically significance* difference)

- A significance level (α) set in advance determines the cut-off for how unusual/extreme te test statistic has to be (assuming H_0 is true) in order to reject the assumption that H_0 is true (i.e. to conclude statistical significance)
- α can be chosen to be any number, but typically $\alpha = 0.05$

It is **better** to report the p-value and comment on the *strength of evidence against* H_0 instead of only reporting whether the result is/isn't statistically significant

But somtimes we apply a rule like "Reject H_0 if p-value $\leq \alpha$"

If we had picked a significance level of

$$\alpha = 0.05$$

before starting our hypothesis test, what would we conclude?

We would ***fail to reject the null hypothesis*** that the new Wheel of Destiny is fair at the $\alpha = 0.05$ significance level. In other words, we cannot reject H_0 at the $\alpha = 0.05$ significance level.