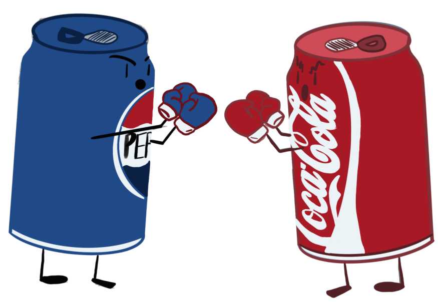
# Day 1 - Coke vs Pepsi

## Introduction

Statistics is the science of quantifying uncertainty. At its core, statistics involves gathering data and using those data to address research questions while accounting for the uncertainty in the data collection process. To illustrate the concept of **statistical inference**, let's consider the age old questions: **Can Coke and Pepsi be discerned?**

Today, each of you will receive two unmarked cups, labeled only A and B. One cup contains Coke, and the other contains Pepsi. Flip a coin to randomly determine which cup you drink first: if the coin lands heads up, drink cup A first. If tails up, drink cup B first.



Coke vs. Pepsi from The Olaf Messenger.

1. Which cup (A or B) do you think had Coke?

2. Based only on **your** answer to the previous question, are you convinced that Coke and Pepsi can be discerned? Why or why not?

3. How might we build stronger evidence that people are able to discern Coke from Pepsi?

## Towards statistical inference

4. In your groups, calculate the proportion of individuals who correctly identified Coke?

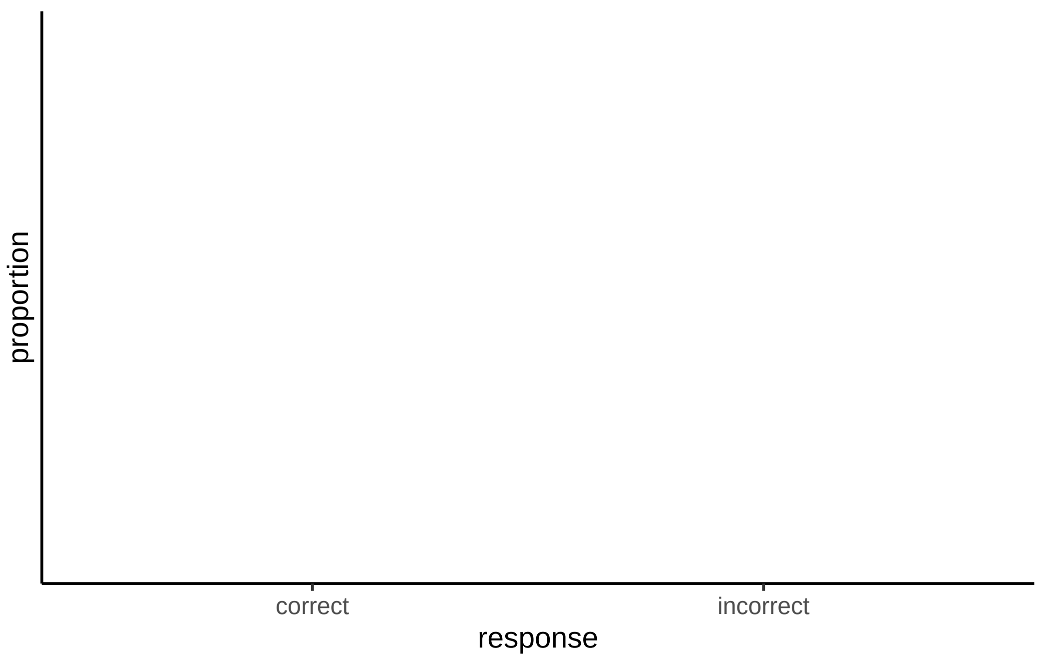
5. If Coke and Pepsi cannot be discerned, what proportion of your group would you expect to correctly identify Coke?

6. Compare your group's proportion to what you would expect if Coke and Pepsi cannot be discerned. Are you convinced that we can discern Coke from Pepsi? Why or why not?

## Calculating an observed statistic

7. Calculate the proportion of the class that correctly identified Coke. Plot this proportion as a **relative frequency bar plot.**

```
library(tidyverse)
tibble(response = c("correct", "incorrect"), proportion = c(NA, NA)) |>
  ggplot(aes(x = response, y = proportion)) +
  geom_bar(stat = "identity", alpha = 0) +
  theme(
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank(),
    panel.background = element_rect(fill = NA),
    axis.line = element_line()
  )
```




8. Compare the class proportion to what you would expect if Coke and Pepsi cannot be discerned. Are you convinced we can discern Coke from Pepsi? Why or why not?

## Building a null distribution

9. Using coins, how might we simulate the proportion of the class that would correctly identify Coke, assuming everyone is just guessing?

10. The following code chunk simulates the count of students in a 29 student class that would correctly identify Coke, if they were all just guessing. How does this simulated count compare to the observed count in this class? Are you convinced that this class can discern Coke from Pepsi? Why or why not?

```r
rbinom(n = 1, size = 29, prob = .5)
```

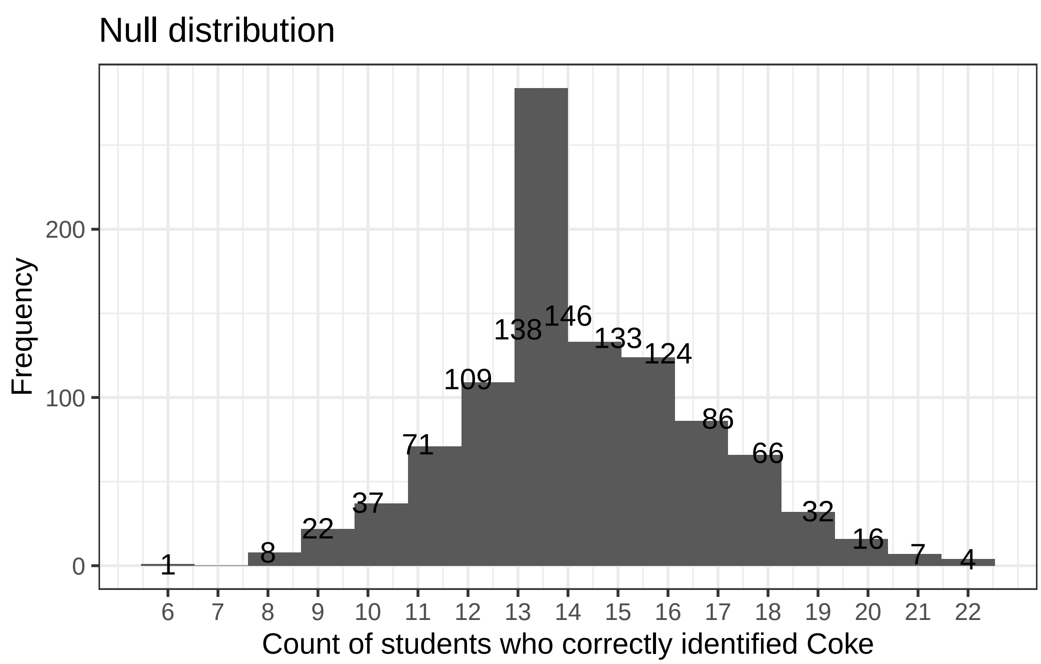
```
[1] 13
```

11. Why should we be wary of drawing conclusions based on one simulated scenario? How many simulations are required to feel good about the comparison we are drawing?

## Quantifying evidence

The code chunk below simulates counting the number of heads resulting from flipping 29 fair coins 1,000 times and plots the results as a **histogram.**

```
set.seed(09092025)
count_tbl <- tibble(freq = rbinom(n = 1000, size = 29, prob = .5))
count_tbl |>
  ggplot(aes(x = freq)) +
  geom_histogram(bins = length(unique(count_tbl$freq))) +
  scale_x_continuous(
    breaks = seq(min(count_tbl$freq), max(count_tbl$freq), 1)
  ) +
  geom_text(
    stat = "count",
    aes(label = ..count..),
    position = position_stack(vjust = 1.02),
    size = 4
  ) +
  labs(
    title = "Null distribution",
    x = "Count of students who correctly identified Coke",
    y = "Frequency"
  ) +
  theme_bw()
```



12. The above **distribution** is comprised of 1000 simulated values. What does a single simulated value on this distribution represent?

13. What proportion of simulations resulted in even more students correctly identifying Coke than our class? Interpret this proportion. What does it mean in terms of evidence against the assumption that everyone in the class is just guessing?

```
mean(count_tbl$counts >= #OBS STAT#)
```

14. Is it *possible* that we could observe our class result by chance, even if everyone was just guessing?

```
# what about 1,000,000 simulations?
knitr::opts_chunk$set(size = "footnotesize")
set.seed(09092025)
table(rbinom(1e6, 29, prob = .5))[-c(1:12)]
```

```
    15     16     17     18     19     20     21     22     23     24     25
144867 126146  96082  64490  37366  18521   8016   2881    892    233     35
    26     27
     6      1
```

## Wrapping up and looking ahead

This activity demonstrated how statisticians quantify evidence against hypotheses using data; the type of quantitative reasoning demonstrated in this process is extremely powerful, and will be a skill we look to refine throughout this semester.