The Thing about *P* Values

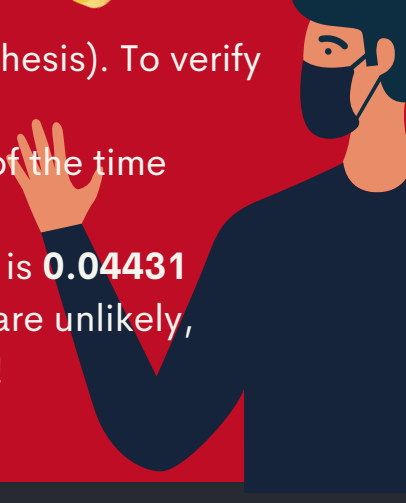
IT'S VERY POSSIBLE THAT P VALUES ARE NOT TELLING YOU WHAT YOU THINK THEY ARE TELLING YOU!

HUH?

- $p < 0.05$ means that if the null hypothesis is true, we would be less than 5% likely to see this particular data set (or more extreme)
  - Because 5% is unlikely, we conclude that the null is unlikely. Thus, we reject the null and accept the alternate hypothesis
- It is **NOT** the probability that you are making the wrong decision!
- It does **NOT** mean that if you repeat the experiment, you would have significant results 95% of the time! (Gigerenzer et al., 2004)

LET'S USE AN EXAMPLE!

- A friend gives you a coin that you think is weighted (alternative hypothesis). To verify your hunch, you flip the coin 100 times
  - If the coin is fair (null hypothesis), it should land on heads ~50% of the time
- The coin lands on heads **59 times**
  - The chances of a fair coin landing on heads at least 59/100 times is **0.04431**
- Because the chances are less than 5%, you decide that these results are unlikely, and thus the coin must NOT be fair. Ergo, the coin must be weighted!



HOW DID THIS COME TO BE?

- *P* values were invented by Ronald Fisher in the 1920s, as a quick a way to judge whether something is worth exploring
- Egon Pearson and Jerzy Newman introduced the use of false positives and negatives
- Other researchers had combined these methods together, but they were never designed to be used this way (Nuzzo, 2014)

WHY DOES IT MATTER?

- *P* values do not tell you much about the false positive rate
  - There is an 11% false positive rate for $p < 0.01$; the rate goes up to 29% for $p < 0.05$ (Goodman, 2001)
- Inflated false positive rates resulting from the use of *P* values are likely an important contributor to the replication crisis (Ioannidis, 2005)
  - 47/53 cancer studies could not be replicated (Begley & Ellis, 2012)
- *P* values alone do not identify clinically significant results

WHAT CAN WE DO?

- Combine *P* values with additional pertinent information
  - Use confidence intervals
    - The range of values within which the population parameter likely lies
  - Report effect sizes
    - A "statistically significant" effect may not matter practically (may not be biologically/clinically important)
- Use Bayesian statistics
  - Usually researchers are interested in the odds of the hypothesis; *p* values do not represent this, but Bayesian odds ratios do!

CITATIONS