Hypothesis Testing

Scenarios

- You flip a coin 10 times and it comes up heads 8 times
- Is the coin fair?
- You show 2 different versions of a web page to users and get some sort of feedback. For version A, the feedback is 65% positive. For version B, the feedback is 55% positive.
- Is version A preferred by users?
- You're doing trials of a new drug. 70% of people who take your drug get better within a week. 60% of those who take a placebo get better in a week.
- Is your drug effective?

Hypothesis Testing

- Lets us quantify the answers to those questions while making assumptions explicit
- Generally, we assume that our data is a set of samples from a population with a given probability distribution (this is our hypothesis)
- We compute the probability of drawing a sample like our data from that probability distribution
- If the probability is extremely low, we can conclude that our assumption about the population was wrong
- The probability also tells how likely we are to incorrectly reject our hypothesis

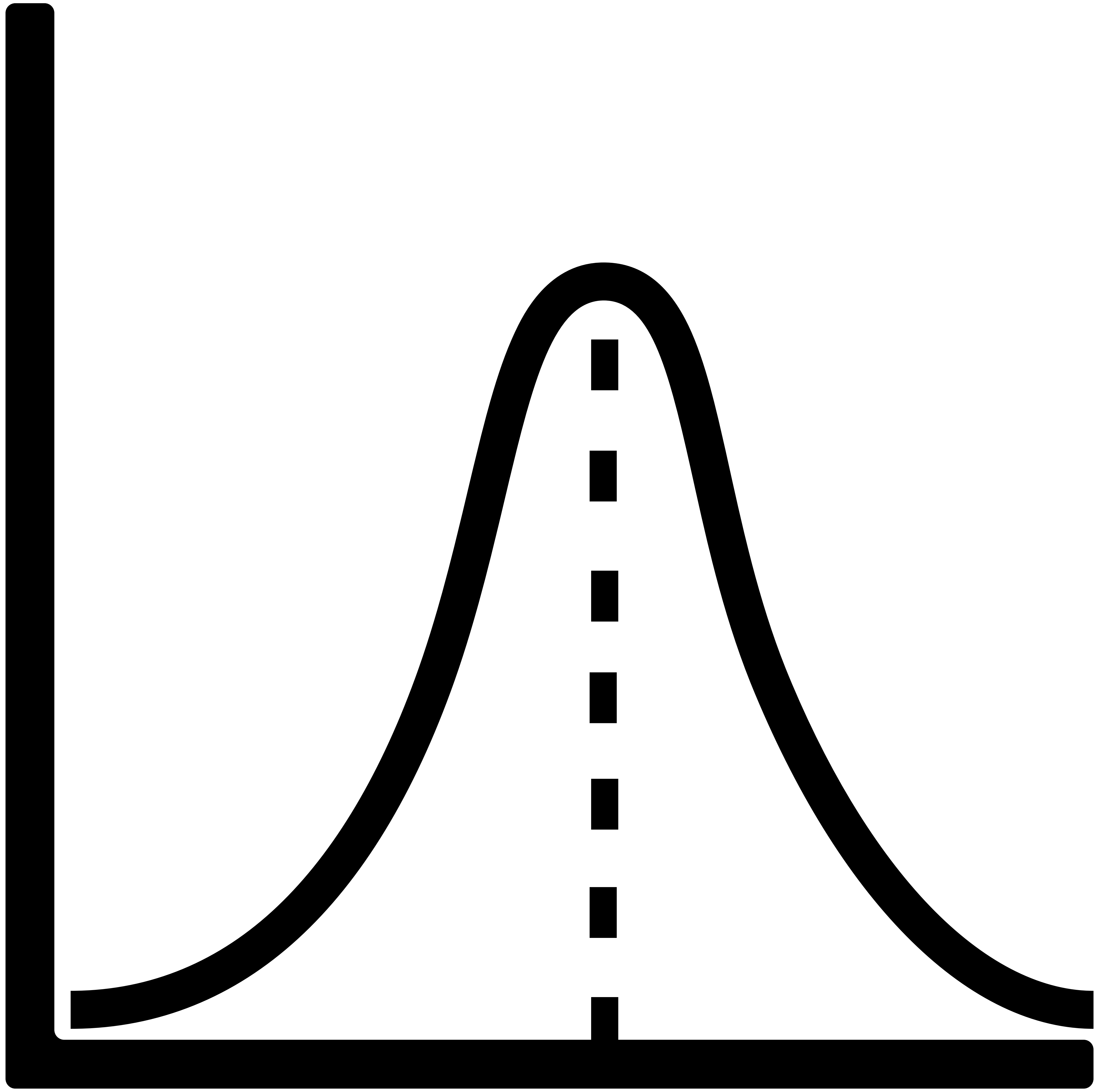
Coin flips

- If our coin is modeled as a bernoulli random variable with p = 0.5, how likely is it to flip it 10 times and get an answer as extreme as the one we saw
- Our hypothesis is that the coin is fair (p = 0.5)
- For this case, we might describe extreme as getting 8 or more flips with the same value
- We can easily compute this: 2 * binomial.cdf(0.5, 2)
  - = 2 * the probability of flipping 0, 1, or 2 heads in 10 flips
  - multiply by 2 to account for 0,1,2, 8, 9, 10 flips (the distribution is symmetric
- This probability is 0.109: ie there is a 11% chance of flipping 8+ heads or tails out of 10 with a fair coin. So this result was somewhat unlikely, but not so unlikely that we can conclude the coin is unfair
- In this case we do NOT reject our hypothesis. Note this does NOT allow us to conclude that the coin IS fair! All we can say is that the coin MIGHT be fair!

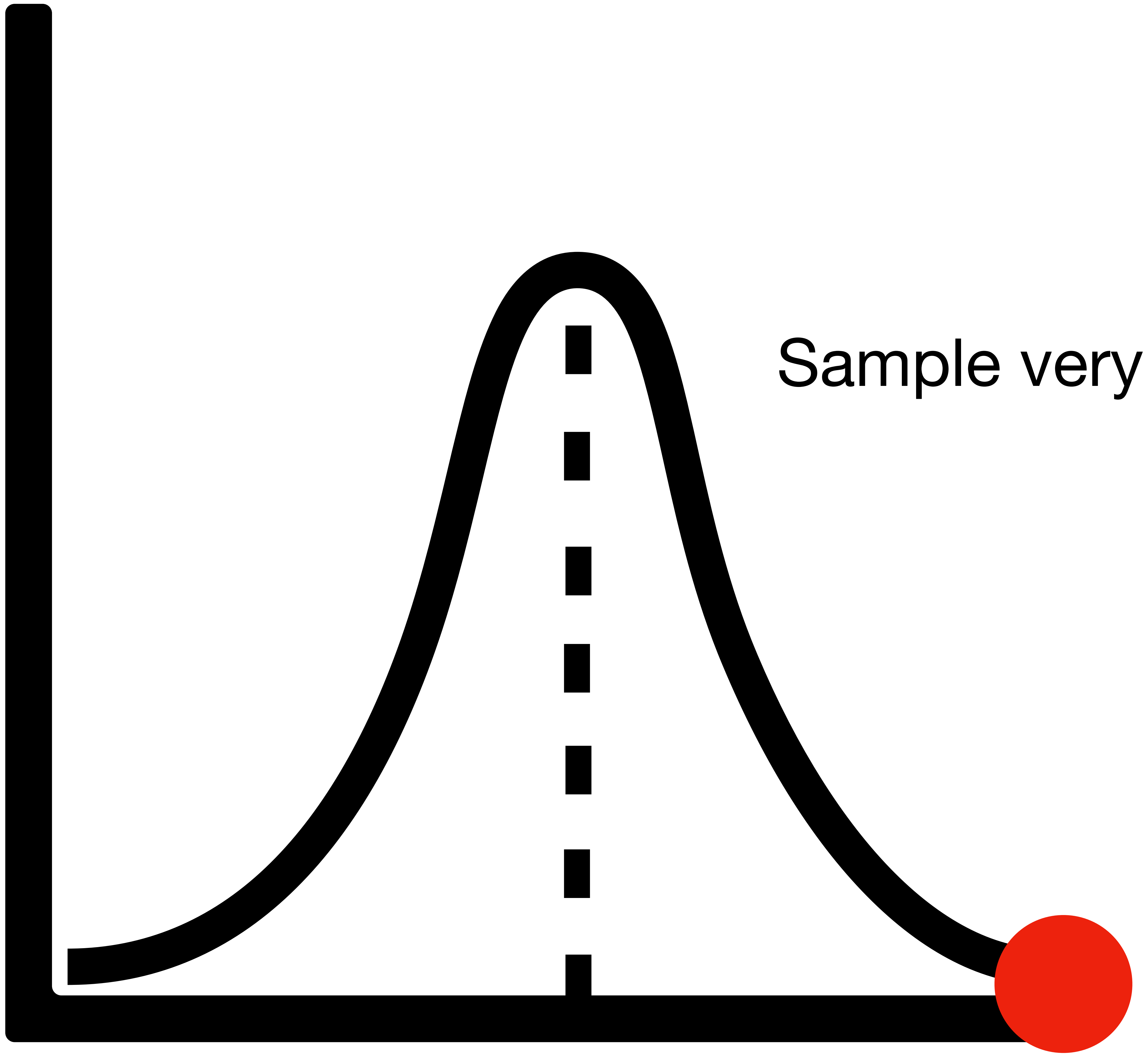
Lingo

- Null Hypothesis (H0): the assumed characteristics of the population, typically a probability distribution and its parameters (eg a normal distribution with mean 10 and variance 2, or poisson distribution with lambda 6.3).
- P-value: the probability of sampling data like ours from the distribution in our null hypothesis
- Significance: our cutoff value for rejecting the null hypothesis. If our P-value is less than the significance, we conclude that the null hypothesis is incorrect. Typical values are .05, .02, .01
- Remember, we cannot show that the null hypothesis was correct! We can only say that it is false or that it might be true! Typically when setting up a hypothesis test we pick H0 so that we learn something if it is false. For example, H0 might be “our drug has the same effectiveness as a placebo”
- For more complicated scenarios, we compute a “test statistic” which is some function of our assumptions + the data we have access to. If the null hypothesis is true, and we repeat the experiment a bunch of times, we expect the test statistic to come from a given probability distribution (often a normal distribution with mean 0, variance 1). Our P-value is how likely our computed test statistic is, if it’s supposed to come from that distribution

Null hypothesis

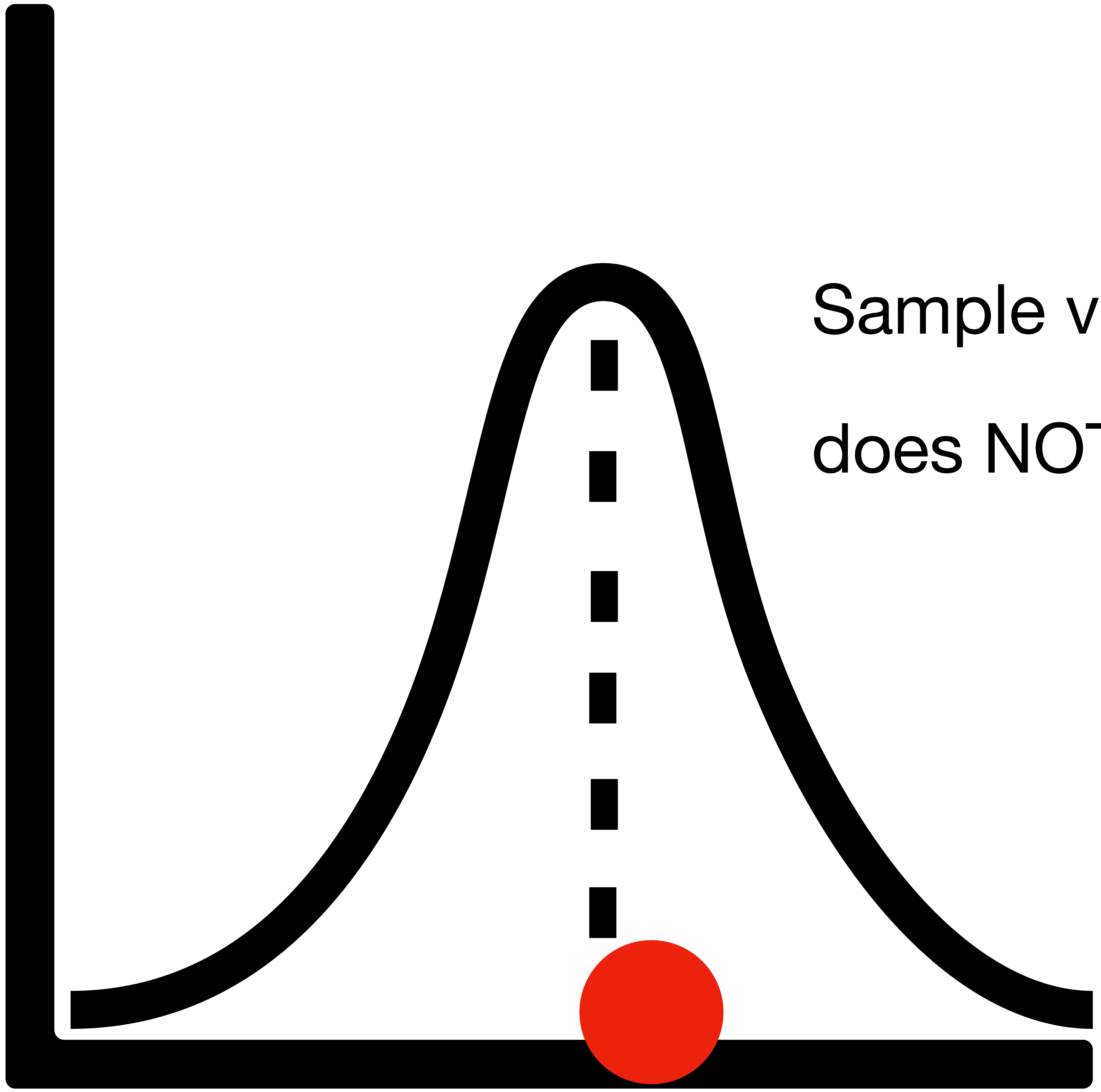


Null hypothesis



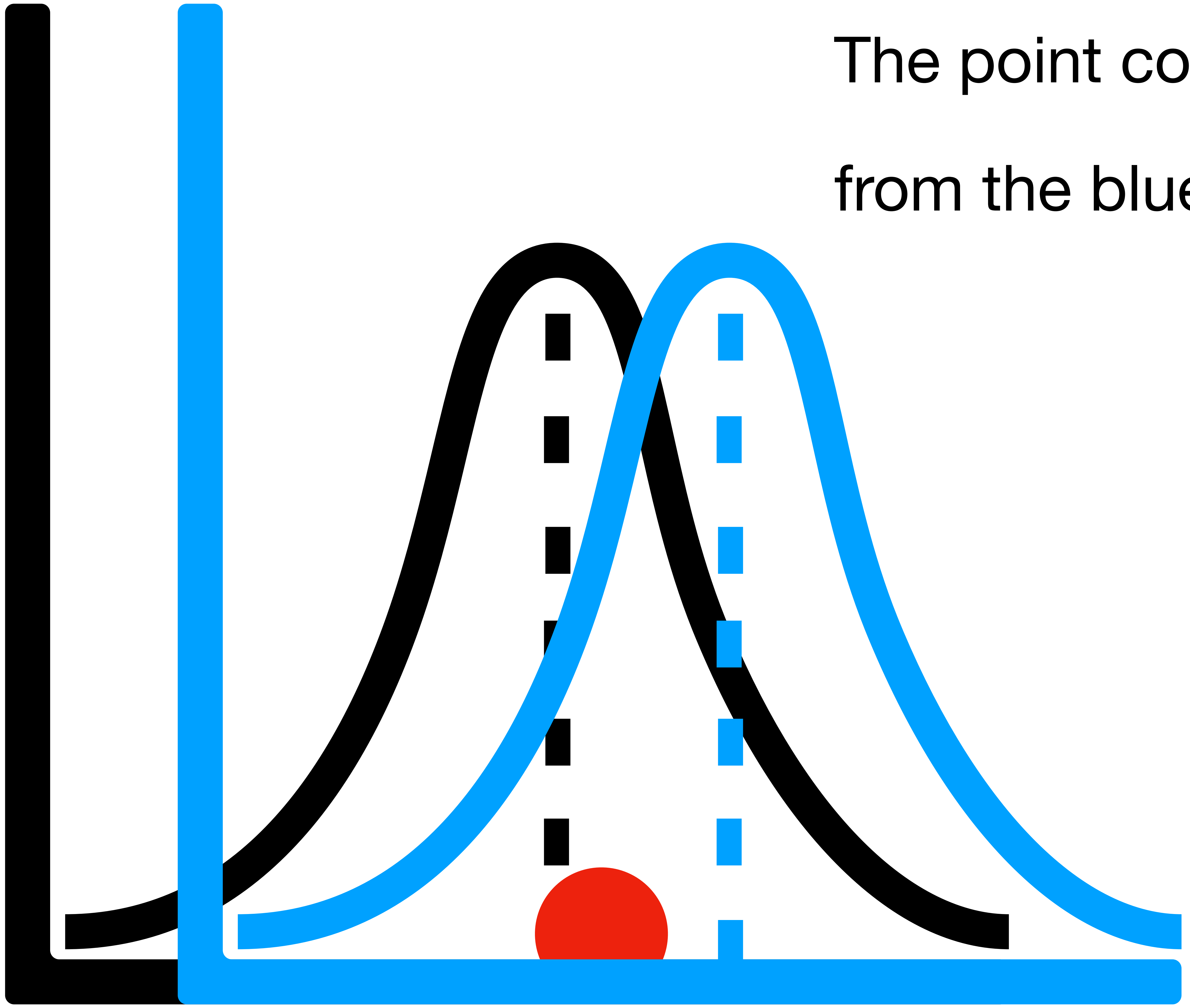
Sample very unlikely, can reject H0

Null hypothesis



Sample very likely, but this does NOT prove H0!

Null hypothesis



The point could have come

from the blue distribution too