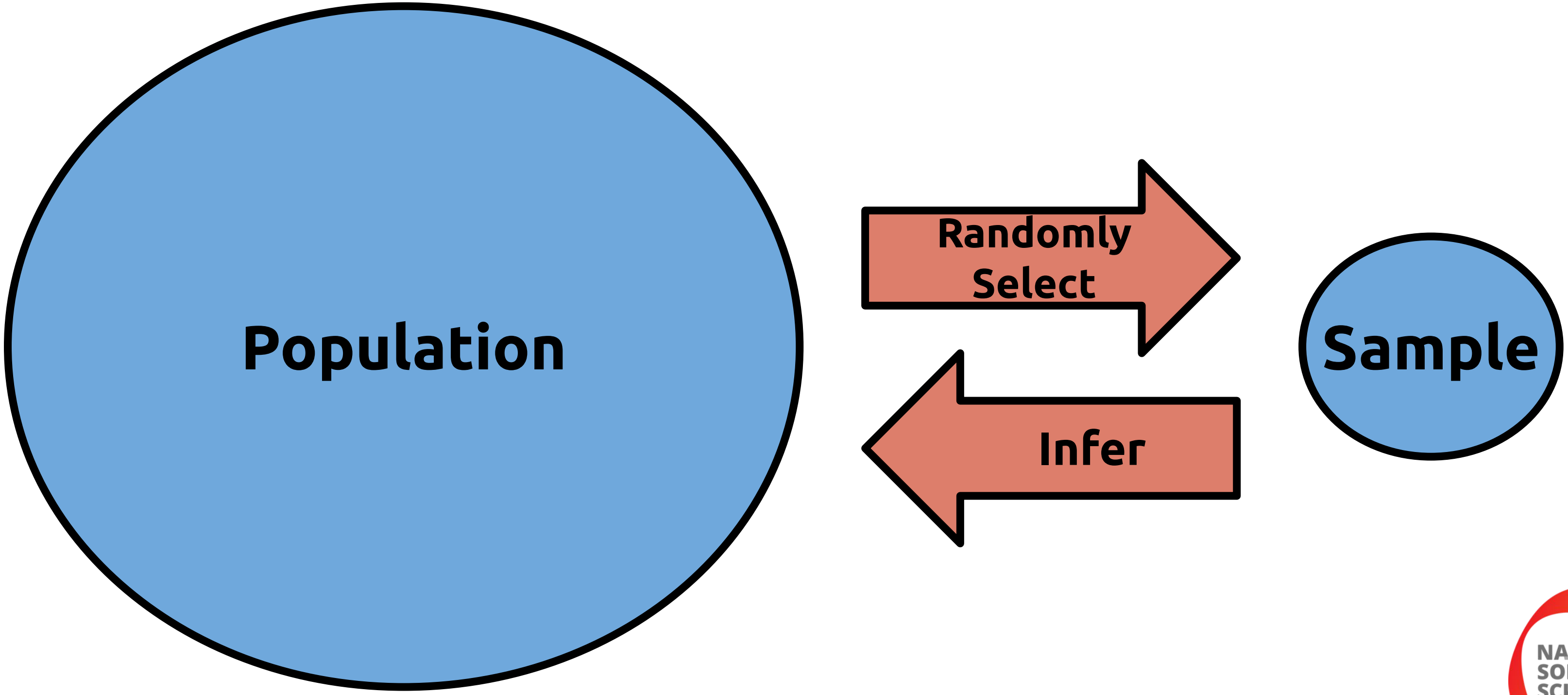
Introduction to Hypothesis Testing

NASHVILLE SOFTWARE SCHOOL

Recall: Populations and Samples



Hypothesis Testing

Goal: Test whether some hypothesis about a population parameter is true, by inspecting only a sample.

Sampling leads to variance and randomness.

You must be careful not to be fooled by this randomness into an incorrect conclusion.

Hypothesis Testing

The question we want to answer: “Given a sample and an apparent effect, what is the probability of seeing such an effect by chance?”

What we are testing for is **statistical significance.**

A set of measurements or observations is said to be statistically significant if it is **unlikely to have occurred by chance**.

Hypothesis Testing

Example: I have a coin which I suspect is not fair, meaning that I think it is more likely to land on one side of the other.

How can I test this?

One option is to flip it some number of times (say, 100) and observe what happens.

Hypothesis Testing

NASHVILLE SOFTWARE SCHOOL

Population of interest: All possible tosses of this particular coin

Parameter of interest: The probability of landing on heads

Randomly Select

Infer

Sample: The 100 coin tosses that I record

Statistic: The proportion of times the coin lands on heads in my sample

Hypothesis Testing

Before I can proceed, I need to decide what my default position is. Since I have no reason to think otherwise, I am going to assume that I do in fact have a fair coin.

I will only change my mind if my test reveals very compelling evidence - evidence so compelling that I would feel silly not to change my mind.

I will not change my mind about the coin being fair unless my test reveals something that is very unlikely to happen just due to chance.

Hypothesis Testing

In hypothesis testing, this default position is known as the **null hypothesis**, or H_0

If I see compelling enough evidence to change my mind, I will instead adopt the **alternative hypothesis,** H_1

Hypothesis Testing

Scenario 1:

Outcome	
Heads	47
Tails	53

Should I change from my default position that the coin is fair?

Probably not. There is variability in the proportion of times that it lands on heads, but we are not far from the expected 50/50 outcome.

Hypothesis Testing

Scenario 1:

Outcome	
Heads	47
Tails	53

Here, I do not have enough evidence to reject the null hypothesis.

I haven't *proven* the null hypothesis; I've just not rejected it.

Hypothesis Testing

Scenario 2:

Outcome	
Heads	38
Tails	62

Should I change from my default position that the coin is fair?

Here, it is harder to say, but it seems much less likely to be this far off from the expected 50/50. I'm much more skeptical that the coin is fair in this scenario.

Hypothesis Testing

Scenario 2:

Outcome	
Heads	38
Tails	62

I will reject the null hypothesis, in favor of the alternative hypothesis that the coin is <u>not</u> fair.

Again, I have not proven anything, but our evidence does not support the hypothesis that the coin is fair.

NASHVILLE SOFTWARE SCHOOL

Hypothesis Testing - Types of Errors

What could have gone wrong in the above example?

In scenario 2, we rejected the null hypothesis in favor of the alternative hypothesis.

If the coin really was fair, and we just had a particularly unlikely run of coin flips, then we would have committed what is called a **Type I Error**. That is, we incorrectly rejected the null hypothesis.

The coin was fair, but we concluded that it was not.

Hypothesis Testing - Types of Errors

What could have gone wrong in the above example?

In scenario 1, we did not reject the null hypothesis. We would have been wrong if in reality the coin was not fair.

This is an example of a **Type II Error**. That is, failing to reject the null hypothesis when in reality it is false.

Hypothesis Testing - Types of Errors

		Reality	
		H_0 is True	**H_0 is False**
Our Decision	**Do not Reject H_0**	Correct Decision	False Negative / Type II Error
	Reject H_0	False Positive / Type I Error	Correct Decision

NASHVILLE SOFTWARE SCHOOL

Hypothesis Testing

When doing hypothesis testing, we choose the null hypothesis H_0 so that it serves as the “default decision”.

This means that in the absence of compelling evidence, we can feel good about falling back on the null hypothesis.

Think of a hypothesis test as being like a trial. The default decision is to find the defendant *not* guilty unless the prosecution can present compelling enough evidence to change the jury’s mind.

Hypothesis Testing - Types of Errors

		Reality	
		H_0 is True: Not Guilty	H_0 is False: Guilty
Our Decision	Do not Reject H_0: Not Guilty	Correct Decision	False Negative / Type II Error
	Reject H_0: Guilty	False Positive / Type I Error	Correct Decision

Hypothesis Testing

The way that hypothesis testing is done, calibration is set according to the likelihood of a Type I error in the case that the null hypothesis is true.

That is, we don't want to conclude that there is some effect when there is none, just like we wouldn't want to incorrectly find a person who is not guilty to be guilty.

The data must show us compelling enough evidence to reject the null hypothesis.

NASHVILLE SOFTWARE SCHOOL

Hypothesis Testing

How do we decide whether or not to reject the null hypothesis?

By quantifying how unlikely our sample would be if the null hypothesis were in fact true.

A **p-value** measures the probability, under the assumption of the null hypothesis, of obtaining a sample *at least as extreme* as what we observed.

p-values

To determine whether or not to reject the null hypothesis, we must establish a threshold for how extreme our observation is. This threshold is called the **significance level**.

Traditionally, the significance level used has been 0.05, meaning that if we calculate a *p*-value less than 0.05, we will reject the null hypothesis.

The significance level determines the chance of a Type I error (incorrectly rejecting the null hypothesis) *in the event that the null hypothesis is true.*

p-values

In the case of coin flips, we know what the exact data generation process would be under the assumptions of the null hypothesis: a binomial distribution with probability of success = 0.5 and $n = 100$.

We're testing if the coin is not a fair coin, so our alternative hypothesis is that probability of success ≠ 0.5.

We also need to specify our significance level. We'll use the standard 0.05 significance level here.

p-values

It would also be possible to test an alternative like probability of heads > 0.5 or probability of heads < 0.5. These are what are called **one-tailed tests**.

What we are testing, probability of heads ≠ 0.5, is called a **two-tailed test** since we are not specifying in which direction the coin is unbalanced, just that it is more likely to land on one of the sides.

Example: Scenario 1

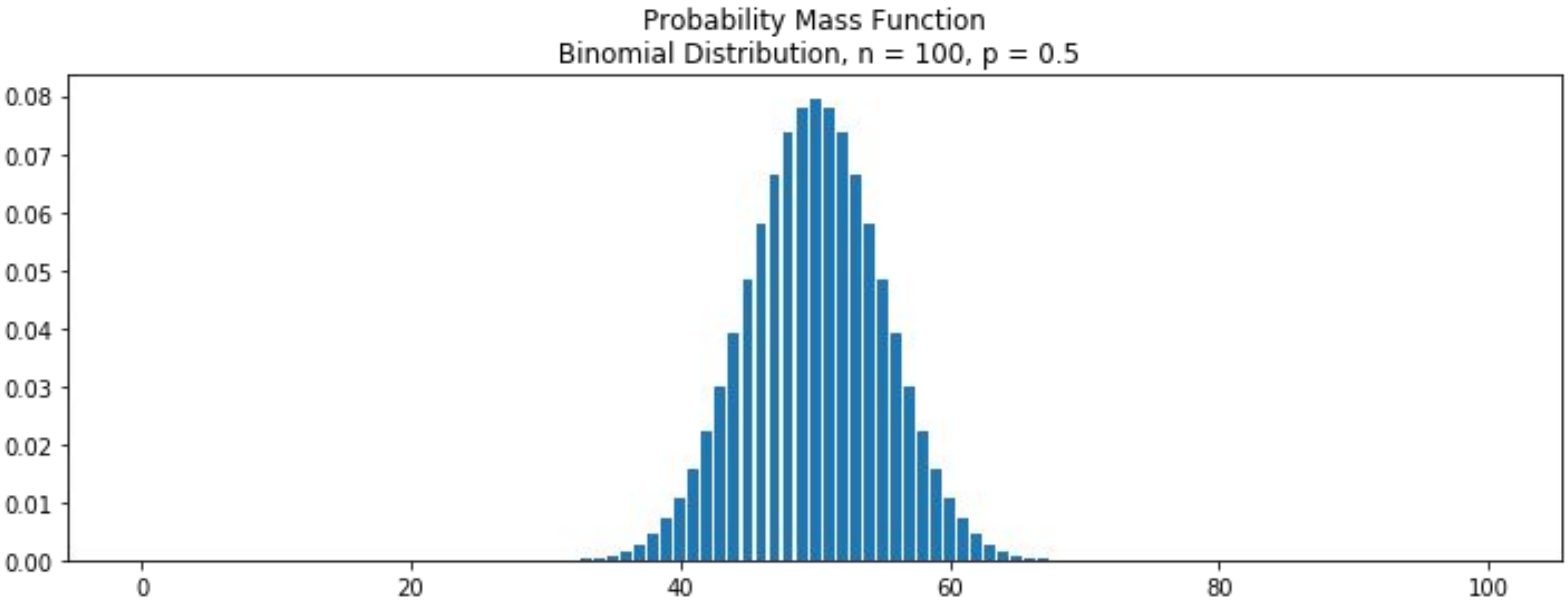
Outcome	
Heads	47
Tails	53

Let's look at the probability mass function if the null hypothesis (that the coin in fair) is true.

NASHVILLE SOFTWARE SCHOOL

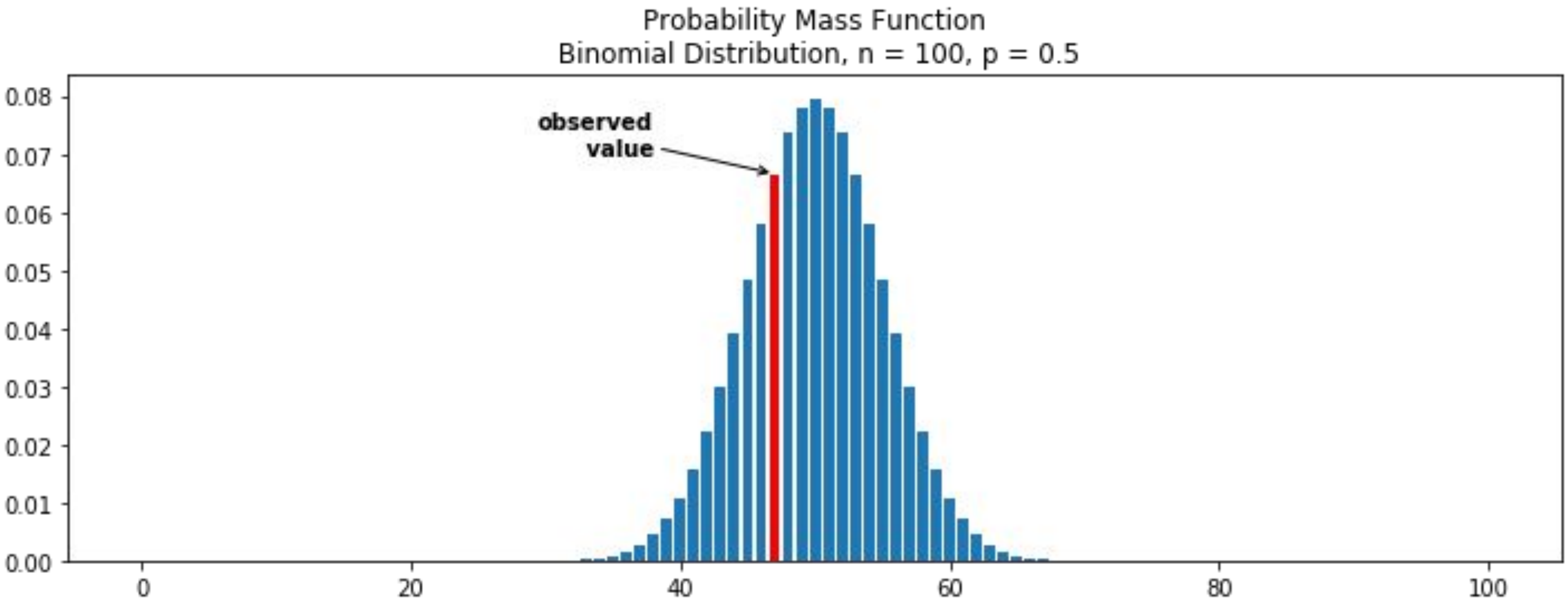
Example: Scenario 1

If the null hypothesis is true, here is what the pmf looks like:



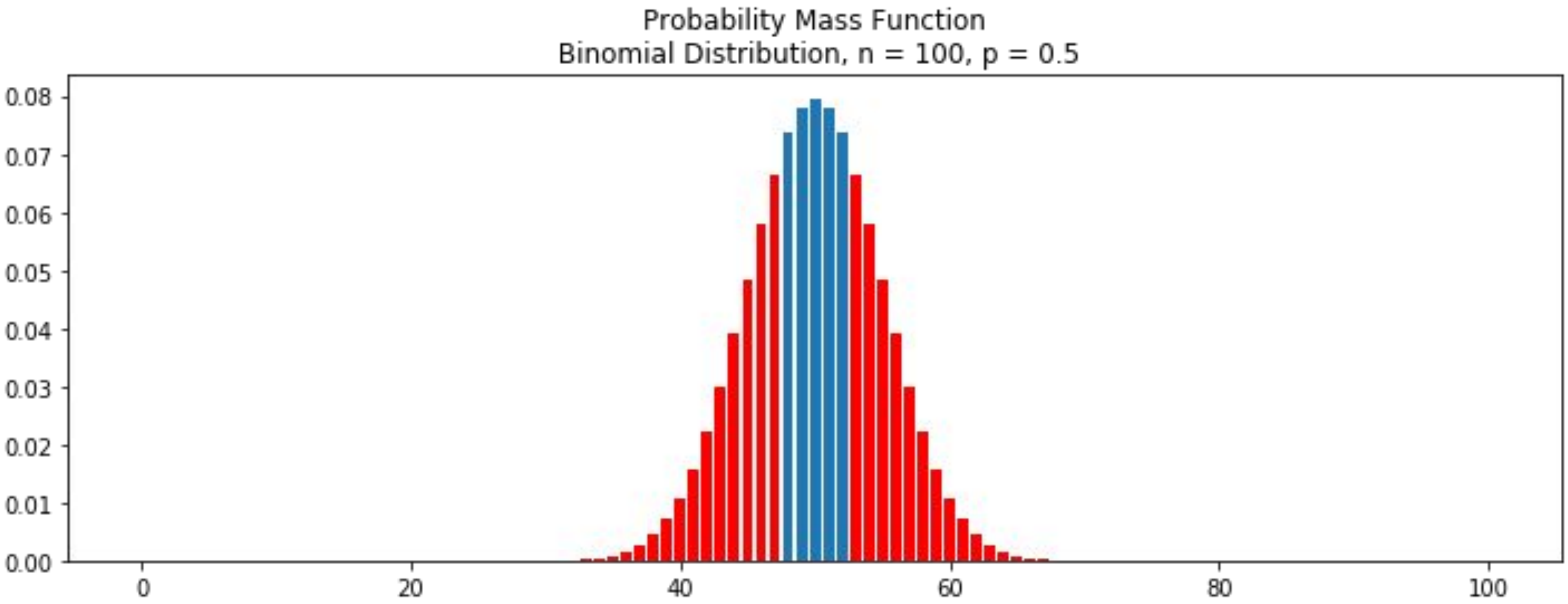
Example: Scenario 1

Let's see where our observed value lands.



Example: Scenario 1

And then let's look at all the possible values that are at least as extreme as what we observed. That is, cases where we get no more than 47 heads, or 53 or more heads.



Example: Scenario 1

Using the cumulative distribution function reveals that the likelihood of these outcomes is approximately 0.617.

This means that the *p*-value is 0.617.

This is not below our threshold of 0.05, so we will **not** reject the null hypothesis.

There is not enough evidence to conclude that the coin is not fair.

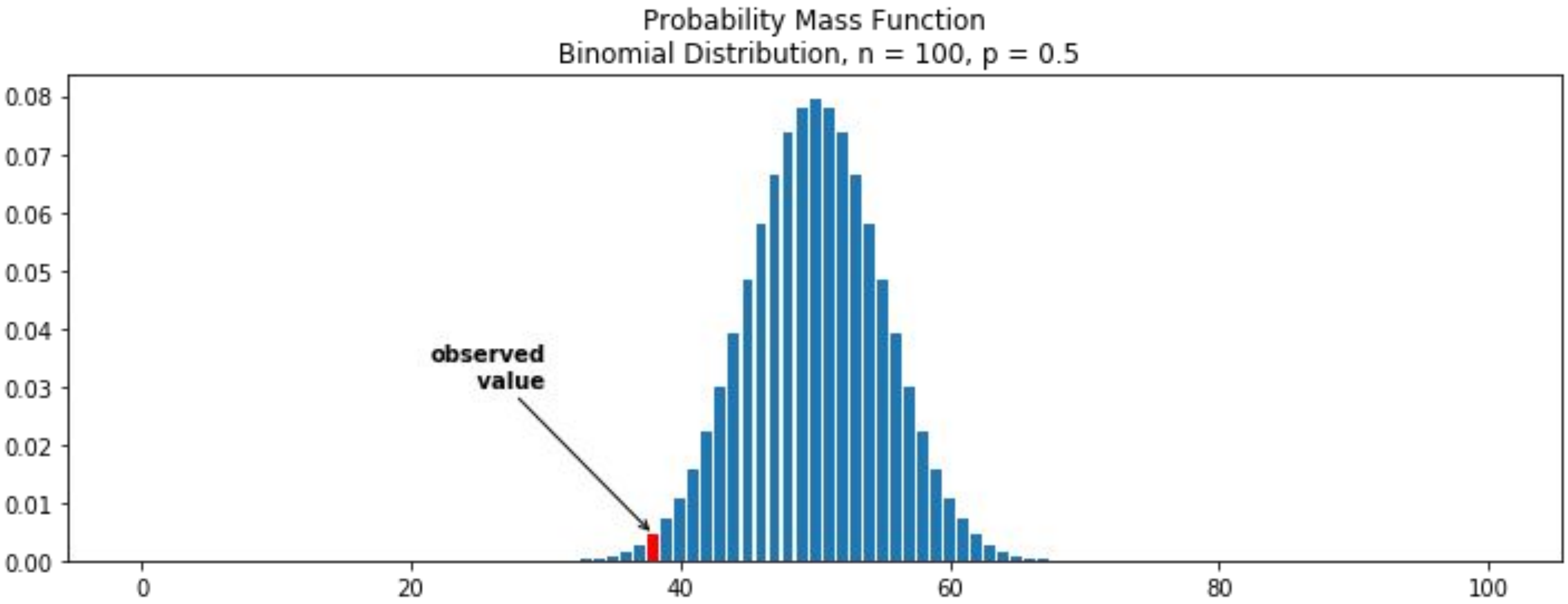
Example: Scenario 2

Outcome	
Heads	38
Tails	62

Let's look at the probability mass function if the null hypothesis that the coin in fair is true.

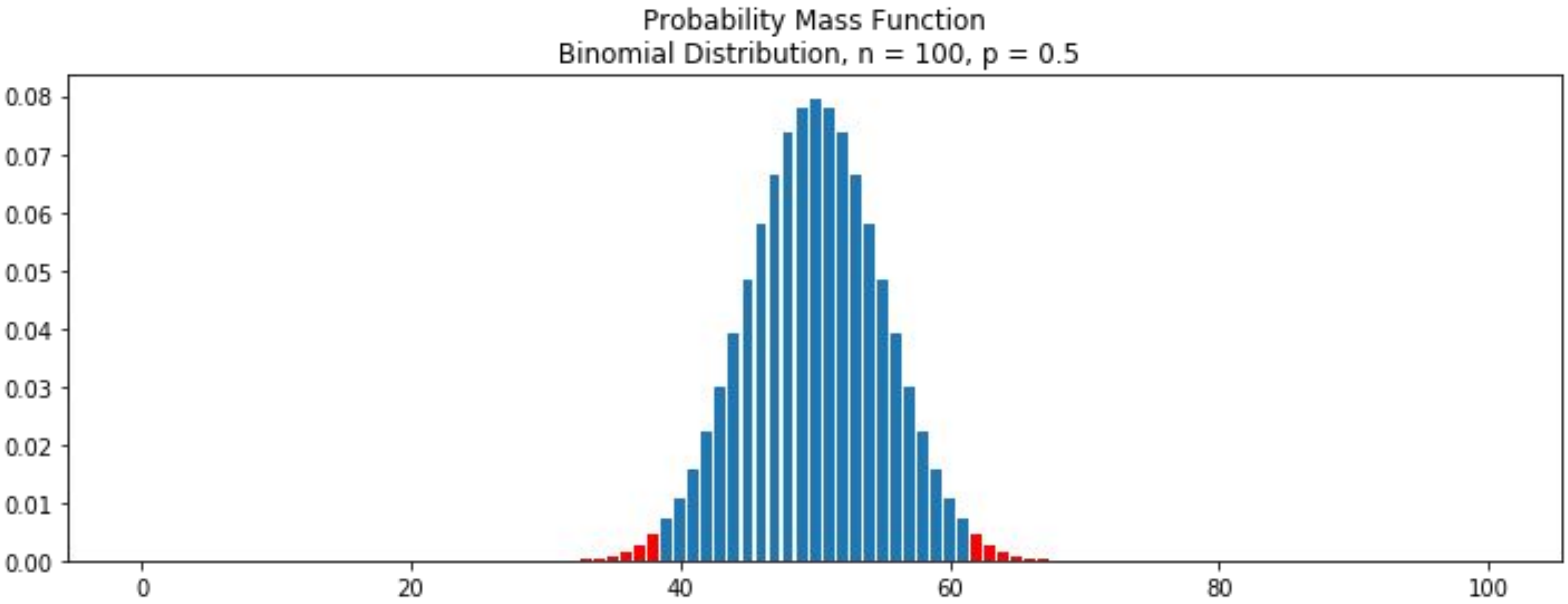
Example: Scenario 2

Let’s see where our observed value lands.



Example: Scenario 2

And then let’s look at all the possible values that are at least as extreme as what we observed. That is, cases where we get no more than 38 heads, or 62 or more heads.



Example: Scenario 2

Now, the cdf reveals a *p*-value of only 0.021.

This probability is below the threshold value of 0.05, so in this case, we can reject the null hypothesis.

It seems unlikely that the extremeness of our observation was due only to random chance.

There is statistically significant evidence that the coin is not fair.

Hypothesis Testing Recap:

- Create your null and alternative hypothesis and choose a significance level.
  - Null hypothesis, H_0 is the skeptical view/the effect is not present in the population
  - Alternative hypothesis, H_1 is that the effect you are testing is present in the population
- Assume that the null hypothesis is true, and choose a statistic to calculate.
- Determine/estimate how your chosen statistic is distributed under the null hypothesis.
- Find the *p*-value: calculate how often you would see a sample statistic as extreme or more extreme than the one you observed.
- If the *p*-value is less than the significance level, reject the null, otherwise, do not reject the null.

Cautions about *p*-values

The use of *p*-values has become more controversial in recent years due to how often they are either misused or misunderstood.

See, for example, this Nature editorial: https://www.nature.com/articles/d41586-019-00874-8

Cautions about *p*-values

Important:

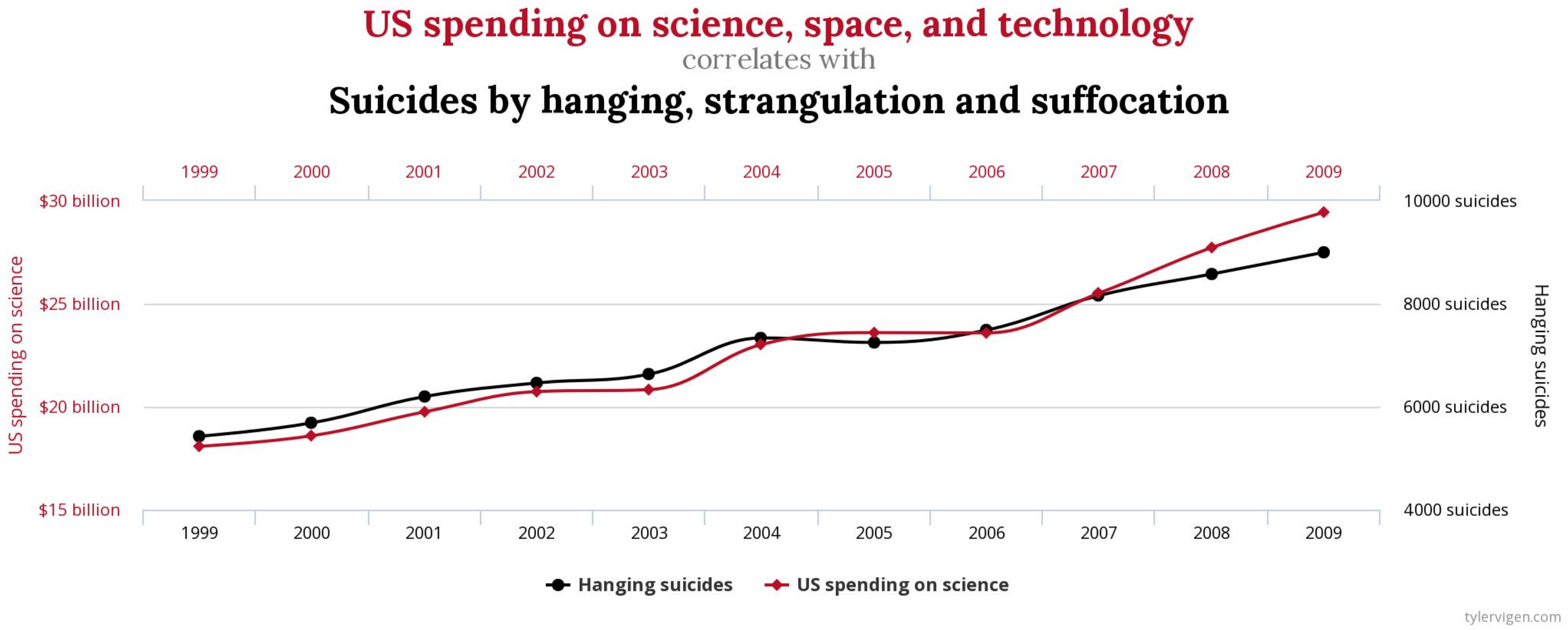
- *p*-values do not give the likelihood that the result is due to chance
- *p*-values only summarize the data, <u>assuming the null hypothesis is true</u>! They do not say how likely the result is to be true.
- *p*-values say nothing about the size of an effect. Statistical significance is not the same as *practical* significance.
- A low *p*-value does not <u>prove</u> the alternative. Ronald Fisher, the inventor of the *p*-value, only meant for "statistical significance" to be an informal index.

Cautions about *p*-values

Another easy mistake to make with *p*-values is the **multiple comparisons/multiple testing** problem. When doing many simultaneous comparisons across a dataset, the chances increase of seeing a “statistically significant” effect which is just due to random sampling error.

See this xkcd comic: https://xkcd.com/882/ or this FiveThirtyEight interactive: https://fivethirtyeight.com/features/science-isnt-broken/#part1

Cautions about *p*-values



Cautions about *p*-values

When doing hypothesis testing, it is important to distinguish between exploratory analysis and hypothesis testing.

Hypothesis testing must be deliberate, which a specific hypothesis in mind prior to looking at the data.

It is not valid to first look for potential effects in a dataset and then test those effects <u>using the same data</u>.