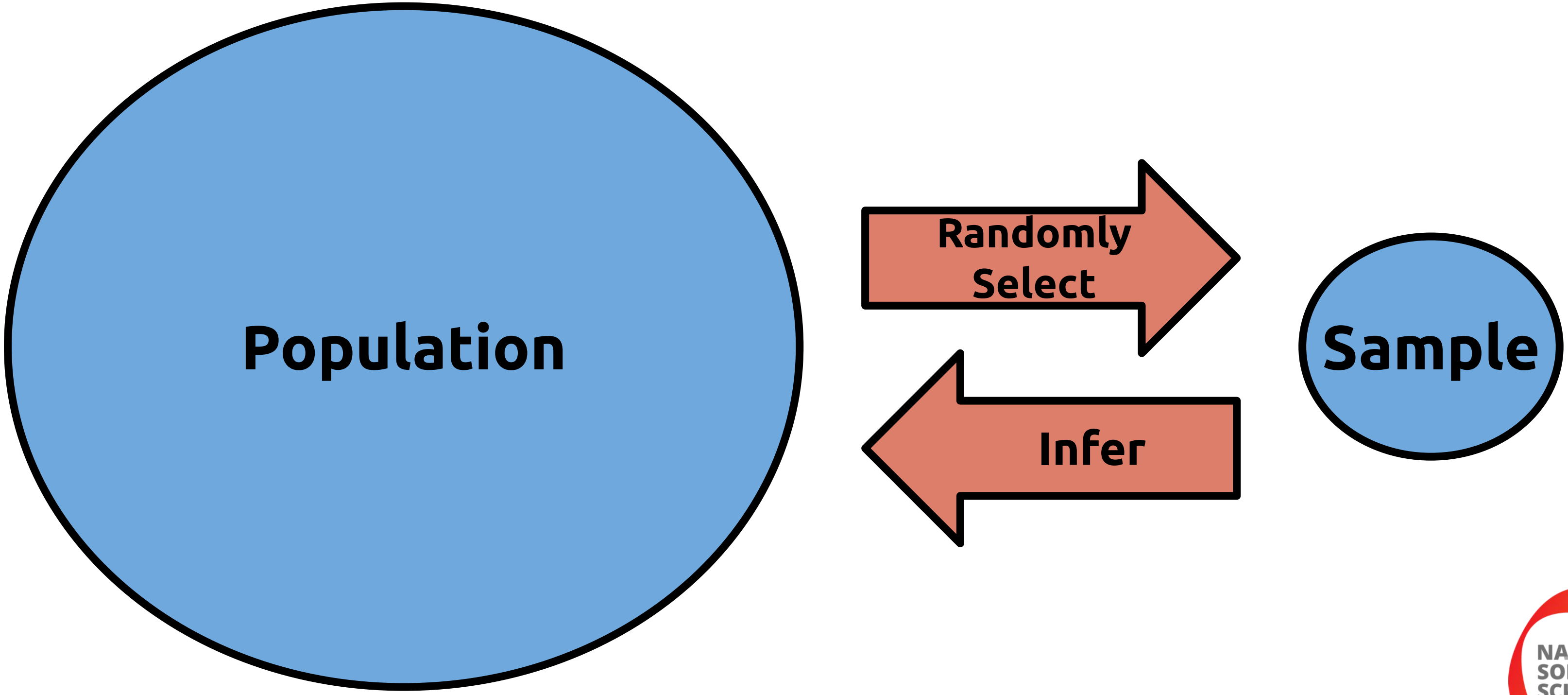
Introduction to Hypothesis Testing

Recall: Populations and Samples



NASHVILLE SOFTWARE SCHOOL

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?”

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.**

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 proceeding, I need to decide what my default position is.

I won’t change my mind from this position unless the data shows something that is very unlikely, under this assumption, to happen just due to chance.

Hypothesis Testing

Before proceeding, I need to decide what my default position is.

I won't change my mind from this position unless the data shows something that is very unlikely, under this assumption, to happen just due to chance.

Default Position: The coin is fair.

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.

NASHVILLE SOFTWARE SCHOOL

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

		Reality	
		Coin is Fair	Coin is Not Fair
Our Decision	Coin is Fair		
	Coin is Not Fair		

Hypothesis Testing - Types of Errors

		Reality	
		Coin is Fair	**Coin is Not Fair**
Our Decision	**Coin is Fair**		
	Coin is Not Fair		

NASHVILLE SOFTWARE SCHOOL

Hypothesis Testing - Types of Errors

		Reality	
		Coin is Fair	Coin is Not Fair
Our Decision	Coin is Fair	Correct Decision	
	Coin is Not Fair		

Hypothesis Testing - Types of Errors

		Reality	
		Coin is Fair	**Coin is Not Fair**
Our Decision	**Coin is Fair**	Correct Decision	
	Coin is Not Fair		

Hypothesis Testing - Types of Errors

		Reality	
		Coin is Fair	**Coin is Not Fair**
Our Decision	**Coin is Fair**	Correct Decision	
	Coin is Not Fair	False Positive / Type I Error	

Hypothesis Testing - Types of Errors

		Reality	
		Coin is Fair	**Coin is Not Fair**
Our Decision	**Coin is Fair**	Correct Decision	
	Coin is Not Fair	False Positive / Type I Error	

NASHVILLE SOFTWARE SCHOOL

Hypothesis Testing - Types of Errors

		Reality	
		Coin is Fair	**Coin is Not Fair**
Our Decision	**Coin is Fair**	Correct Decision	
	Coin is Not Fair	False Positive / Type I Error	Correct Decision

Hypothesis Testing - Types of Errors

		Reality	
		Coin is Fair	**Coin is Not Fair**
Our Decision	**Coin is Fair**	Correct Decision	
	Coin is Not Fair	False Positive / Type I Error	Correct Decision

Hypothesis Testing - Types of Errors

		Reality	
		Coin is Fair	**Coin is Not Fair**
Our Decision	**Coin is Fair**	Correct Decision	False Negative / Type II Error
	Coin is Not Fair	False Positive / Type I Error	Correct Decision

Hypothesis Testing - Types of Errors

		Reality	
		Coin is Fair	**Coin is Not Fair**
Our Decision	**Coin is Fair**	Correct Decision	False Negative / Type II Error
	Coin is Not Fair	False Positive / Type I Error	Correct Decision

Hypothesis Testing - Types of Errors

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

Hypothesis Testing Steps

1. Assume the null hypothesis.

Hypothesis Testing Steps

1. Assume the null hypothesis.
2. Find the probability of observing a sample at least as extreme as the sample you have if the null is true.

Hypothesis Testing Steps

1. Assume the null hypothesis.
2. Find the probability of observing a sample at least as extreme as the sample you have if the null is true.
3. If the probability is low enough, reject the null in favor of the alternative.

Hypothesis Testing Steps

1. Assume the null hypothesis.
2. Find the probability of observing a sample at least as extreme as the sample you have if the null is true.
3. If the probability is low enough, reject the null in favor of the alternative.

Hypothesis Testing Steps

1. Assume the null hypothesis.
2. Find the probability of observing a sample at least as extreme as the sample you have if the null is true.
3. If the probability is low enough, reject the null in favor of the alternative.

This probability is called
the ***p*-value** for the test.

NASHVILLE SOFTWARE SCHOOL

Hypothesis Testing Steps

1. Assume the null hypothesis.
2. Find the probability of observing a sample at least as extreme as the sample you have if the null is true.
3. If the probability is **low enough**, reject the null in favor of the alternative.

Hypothesis Testing Steps

1. Assume the null hypothesis.
2. Find the probability of observing a sample at least as extreme as the sample you have if the null is true.
3. If the probability is **low enough**, reject the null in favor of the alternative.

This cutoff is called the
significance level for the test.

Significance Level

A very common significance level is 0.05.

This means reject the null if the *p*-value is less than 0.05.

Significance Level

A very common significance level is 0.05.

This means reject the null if the *p*-value is less than 0.05.

If the null is true, for 5% of samples, the *p*-value will be less than 0.05.

Significance Level

A very common significance level is 0.05.

This means reject the null if the *p*-value is less than 0.05.

If the null is true, for 5% of samples, the *p*-value will be less than 0.05.

So 5% of the time, the null will be incorrectly rejected (Type I Error).

Significance Level

A very common significance level is 0.05.

This means reject the null if the *p*-value is less than 0.05.

If the null is true, for 5% of samples, the *p*-value will be less than 0.05.

So 5% of the time, the null will be incorrectly rejected (Type I Error).

Hence, the significance level is the chance of a Type I Error, *in the event that the null hypothesis is true.*

Coin-Flipping Example

Null Hypothesis:

NASHVILLE SOFTWARE SCHOOL

Coin-Flipping Example

Null Hypothesis:

$$P(\text{Heads}) = 0.5$$

Coin-Flipping Example

Null Hypothesis:

$$P(\text{Heads}) = 0.5$$

Alternative Hypothesis:

Coin-Flipping Example

Null Hypothesis:

$$P(\text{Heads}) = 0.5$$

Alternative Hypothesis:

$$P(\text{Heads}) \neq 0.5$$

Coin-Flipping Example

Null Hypothesis:

$$P(\text{Heads}) = 0.5$$

Alternative Hypothesis:

$$P(\text{Heads}) \neq 0.5$$

This is a **two-tailed** test, but we could also do a **one-tailed** version if we are claiming $P(\text{Heads}) > 0.5$ or $P(\text{Heads}) < 0.5$.

Coin-Flipping Example

Null Hypothesis:

$$P(\text{Heads}) = 0.5$$

Alternative Hypothesis:

$$P(\text{Heads}) \neq 0.5$$

Coin-Flipping Example

Our data will be gathered by flipping the coin 100 times and counting the number of heads.

Coin-Flipping Example

Our data will be gathered by flipping the coin 100 times and counting the number of heads.

Question: Under the null hypothesis, what distribution would the number of heads follow?

NASHVILLE SOFTWARE SCHOOL

Coin-Flipping Example

Our data will be gathered by flipping the coin 100 times and counting the number of heads.

Question: Under the null hypothesis, what distribution would the number of heads follow?

A binomial distribution with 100 trials and probability of success 0.5.

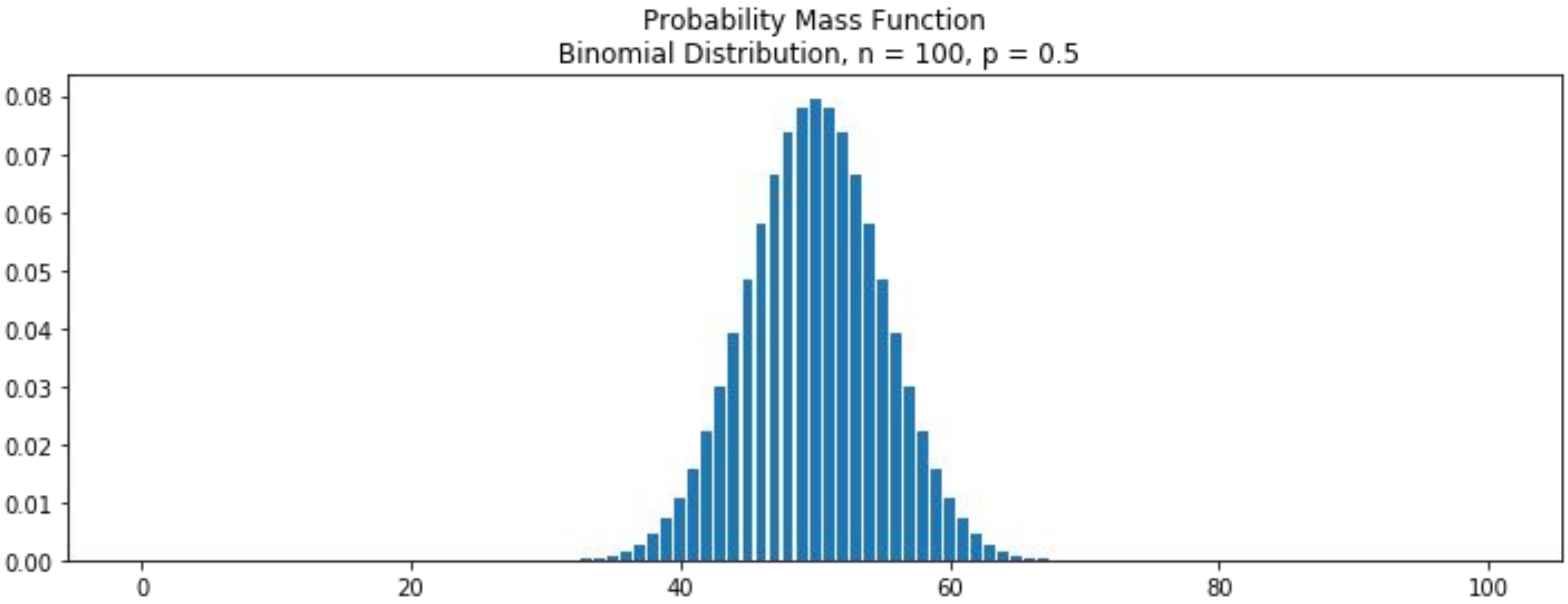
Example: Scenario 1

Outcome	
Heads	47
Tails	53

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

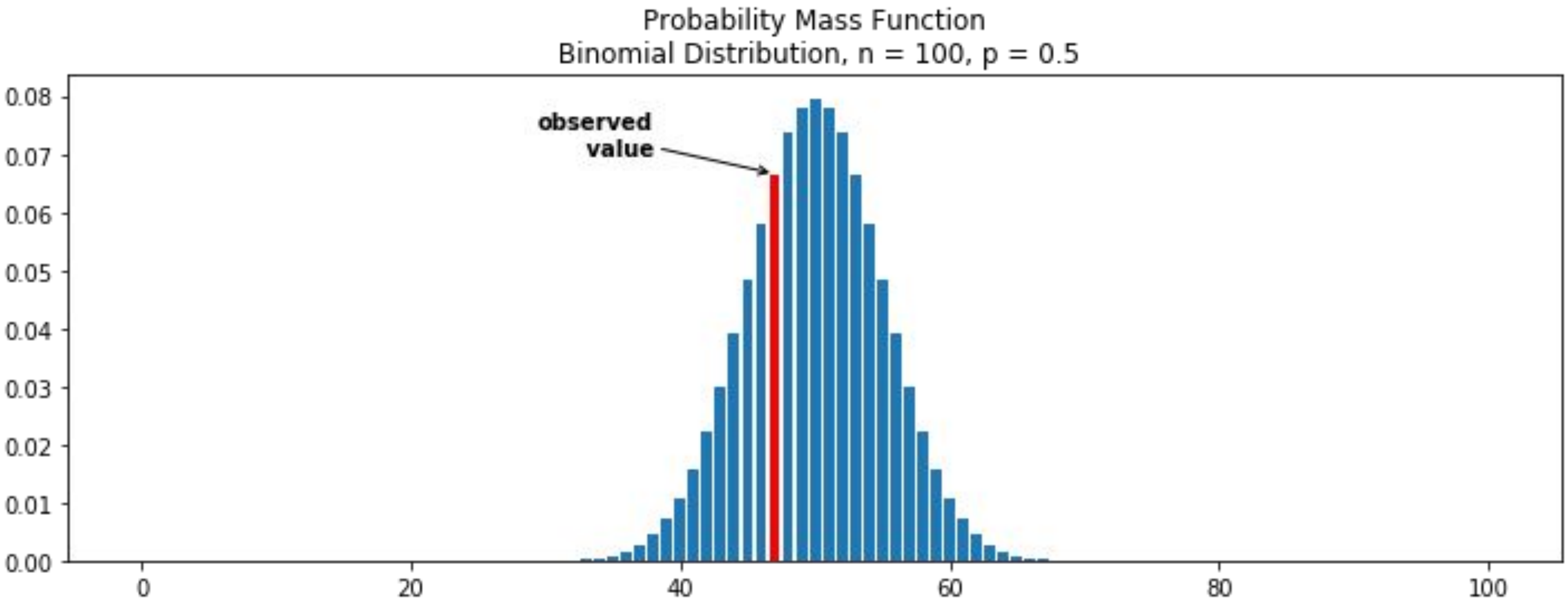
Example: Scenario 1

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



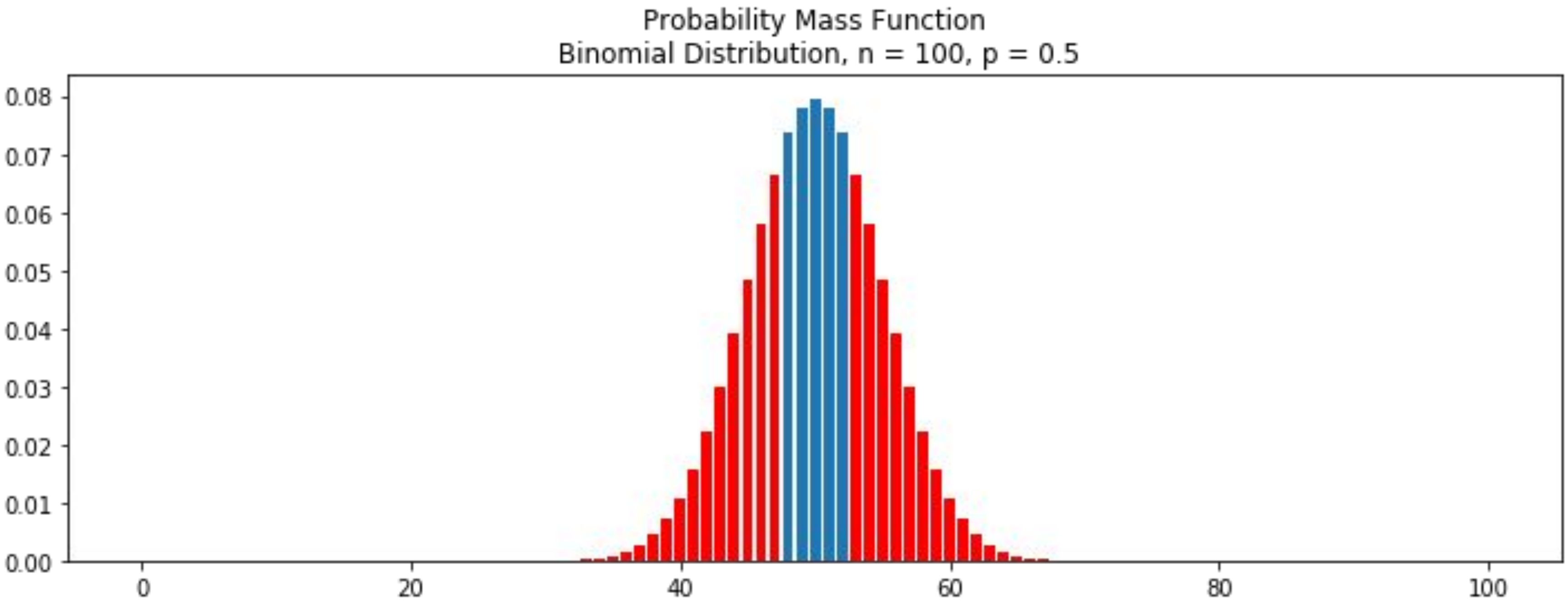
Example: Scenario 1

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.

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.

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 - the observation was within the expected range due to chance.

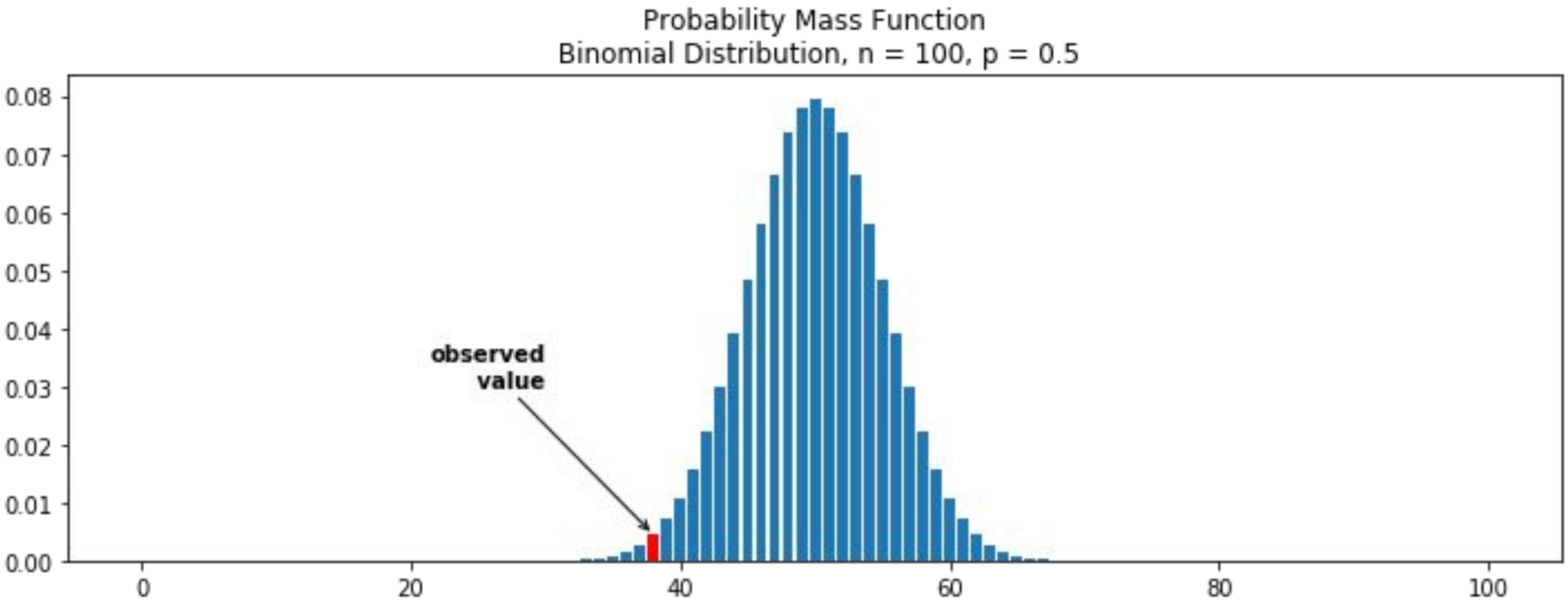
Example: Scenario 2

Outcome	
Heads	38
Tails	62

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

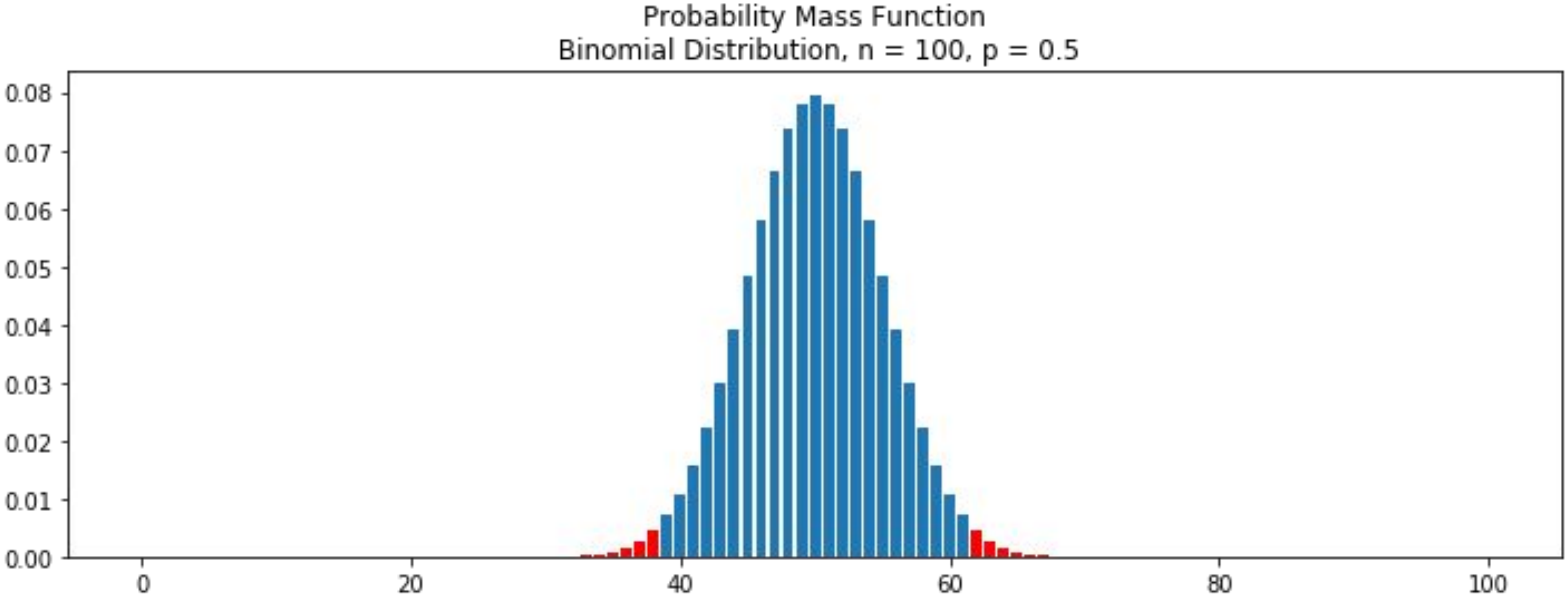
Example: Scenario 2

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.

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.

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.

NASHVILLE SOFTWARE SCHOOL

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:

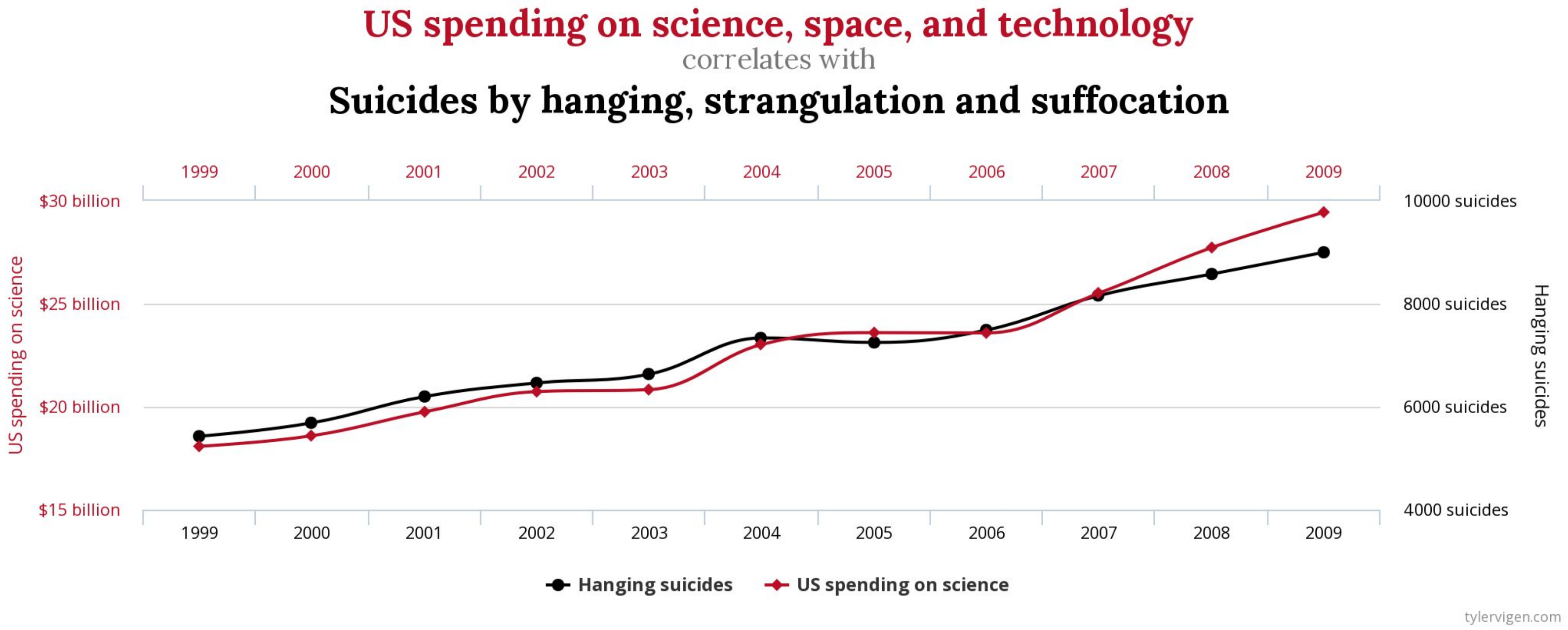
- *p*-values do not give the likelihood that the result is due to chance
- *p*-values only summarize the data, assuming the null hypothesis is true! They do not say how likely the result is to be true.
- *p*-values say nothing about the size of an effect. Statistical significance is not the same as *practical* significance.
- A low *p*-value does not prove 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>.