Binomial Tests

Galvanize

Objectives

By the end of this lesson, you will be able to:

- ❑ Describe the steps of the null hypothesis testing procedure.
- ❑ State a null hypothesis and alternative hypothesis.
- ❑ Calculate the p-value for a binomial test.
- ❑ State some common misinterpretations of p-values.

Probability vs. Statistics

Probability and statistics are closely related subjects, but there is a fundamental difference.

Probability

In probability we know the parameters of a distribution (associated with some random variable), and we would like to study properties of data generated from that distribution.

Example properties of random variables:

- Expectation value: $\mathrm{E}[X] = \sum_{i=1}^{k} x_i\, p_i = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k$

- Variance: $V[X] = E\left[(X - E[X])^2\right]$

If you know the parameters of the distribution, then you can compute the expectation value and variance.

Statistics

In **statistics** we have data generated from a random variable, and we would like to *infer* properties of its distribution.

A few points are evident:

- Independent and identically distributed data are important, as they allow us to pool information using data all generated from *indistinguishable* random variables.
- We can never know *exactly* the distribution that generated the data, we can only hope to approximate it.
- We *may* be able to quantify the uncertainty in our approximation (this is what much of classical statistics is about).

Binomial Test

A binomial test is a specific type of hypothesis test involving random variables that can be modeled by the binomial distribution (i.e. they fall into two categories)

In this lesson, we'll introduce the hypothesis testing process, but apply it only to binomial cases. Don't worry, we'll cover other types of hypothesis testing in a later lesson.

Fisher's Tea Experiment

Ronald Fisher's friend Muriel Bristol claims that she can tell, by actually drinking the beverage, whether milk was poured in first or second into a cup of tea. I.e. tea into the milk, or milk into the tea.

Fisher, being an upstanding skeptic, is skeptical, so devises an experiment to test her claim.

Discussion: How could we determine whether Muriel is telling the truth?

Fisher's Tea Experiment

Fisher's solution is as follows. He prepares six cups of tea, three with tea first and three with milk first. These cups are then given to Muriel arranged in a random order. For simplicity, I'll assume in my version that Muriel does *not know* that there are three of each.

He has Muriel drink each beverage, and attempt to guess if tea or milk were poured first. Her results are to the right.

Discuss: Would you be surprised if this were due to random guessing?

Cup	Result
1	Correct
2	Correct
3	Correct
4	Incorrect
5	Correct
6	Correct

Hypothesis Testing

The tea example was reported in Fisher's classic text *The Design of Experiments*, and it is a prototypical example of the logic behind null hypothesis testing.

1. **State a scientific question** - its answer should be yes or no
2. **State a null hypothesis** - the skeptic's answer to the question
3. **State an alternative hypothesis** - the non-skeptic's answer to the question
4. **Create a model** - the model assumes the null hypothesis is true
5. **Set a threshold** - decide how surprised you need to be to reject the null hypothesis
6. **Collect your data**
7. **Calculate a p-value** - the probability of finding a result equally or more extreme if the null hypothesis is true
8. **Compare the p-value to your stated rejection threshold**

Hypothesis Testing

Let’s apply the hypothesis testing process to Fisher’s experiment:

1. **State a scientific question** - its answer should be yes or no

   “Can Muriel predict when tea was poured before coffee?”

Hypothesis Testing

Let’s apply the hypothesis testing process to Fisher’s experiment:

2. **State a null hypothesis** - the skeptic’s answer to the question

   “Muriel has no better than a random chance of guessing correctly.”

- In statistics, the null hypothesis is give the label H_0.
- A null hypothesis can also be thought of as the “status quo” or “no difference” answer to the question.

Hypothesis Testing

Let’s apply the hypothesis testing process to Fisher’s experiment:

3. **State an alternative hypothesis** - the non-skeptic’s answer to the question

   “Muriel has a better-than-random chance of guessing correctly.”

- In statistics, the alternate hypothesis is give the label H_a.
- The alternative hypothesis can also be thought of as the “something changed” answer to the question.

Hypothesis Testing

A statistician would state these hypotheses in a more technical fashion:

$H_0 : p = 0.5$ $\qquad\qquad$ $H_a : p > 0.5$

- The null hypothesis always includes the "=". (i.e. the status-quo)
- Hypothesis tests can be categorized into two types:
  - One-tailed test: $H_0 : p = 0.5$ $\qquad$ $H_a : p > 0.5$
  - Two-tailed test: $H_0 : p = 0.5$ $\qquad$ $H_a : p \neq 0.5$

Check for Understanding

By the end of this lesson, you will be able to:

- ❏ Describe the steps of the null hypothesis testing procedure.
- ❏ State a null hypothesis and alternative hypothesis.

1. What is the difference between a null hypothesis and an alternative hypothesis?

State a null and alternative hypothesis for each question and if it would result in a one-tail or two-tail test.

2. Did more than 30% of registered voters vote in the last election?
3. Does 80% of the US population know the Nike brand?

Hypothesis Testing

Let’s apply the hypothesis testing process to Fisher’s experiment:

4. **Create a model** - the model assumes the null hypothesis is true

   “If Muriel has no ability to discern which liquid is poured first, then her answers are essentially a coin flip. They would have a binomial distribution with a probability of 0.5.”

Hypothesis Testing

Let's apply the hypothesis testing process to Fisher's experiment:

5. **Set a threshold** - decide how surprised you need to be to reject the null hypothesis.

   Would you be surprised if 25% of the population could do as well or better than Muriel by random guessing? 5%? 0.05%?

   In statistics, this "surprise threshold" you set is called a **significance level** and symbolized by α (alpha).

   A commonly accepted threshold of α is 0.05 (5%).

Hypothesis Testing

Let's apply the hypothesis testing process to Fisher's experiment:

6. **Collect your data**

Cup	Result
1	Correct
2	Correct
3	Correct
4	Incorrect
5	Correct
6	Correct

Hypothesis Testing

Let's apply the hypothesis testing process to Fisher's experiment:

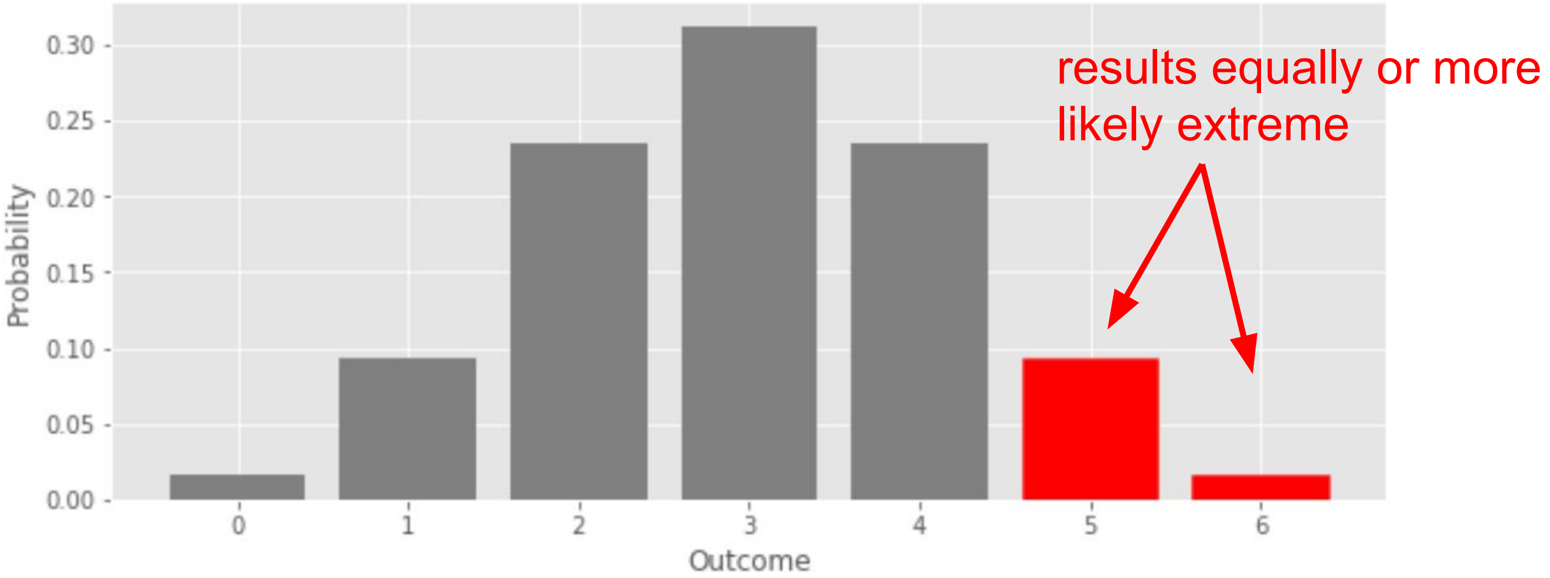
7. **Calculate a p-value** - the probability of finding a result equally or more extreme if the null hypothesis is true

$$P(\text{Data As Or More Extreme} \mid \text{Fisher's Assumption}) = \binom{6}{5}0.5^5 0.5^1 + \binom{6}{6}0.5^6$$

$= 0.11$

Hypothesis Testing

Let's look at a graph of a binomial PMF for six guesses (n=6):



Hypothesis Testing

Let's apply the hypothesis testing process to Fisher's experiment:

8. **Compare the p-value to your stated rejection threshold**

p-value = 0.11 $\alpha = 0.05$

When the p-value is greater than the threshold, we are not surprised. Therefore, at the $\alpha = 0.05$ significance level (threshold), we would <u>not</u> reject the null hypothesis that Muriel's chance of getting a correct answer is better than a guess.

Note we did not "accept" the null hypothesis. If we had Muriel make more guesses, there is a chance we would reject the null hypothesis in the future.

Check for Understanding

By the end of this lesson, you will be able to:

- [ ] Describe the steps of the null hypothesis testing procedure.

1. What does a significance level represent in hypothesis testing?
2. What does a p-value represent?
3. What is wrong with each of the following statements?
   - "The p-value of 0.02 is lower than our chosen significance level of 0.05, therefore we do not reject the null hypothesis."
   - "The p-value of 0.03 is higher than our chosen significance level of 0.01, therefore we accept the null hypothesis."

A Hypothesis Test … with code

Matt is all about consistency in skateboarding, and is learning to kickflip. He does not want to move onto another trick until he can cleanly land a kickflip 80% of the time.

A Hypothesis Test … with code

Matt is all about consistency in skateboarding, and is learning to kickflip. He does not want to move onto another trick until he can cleanly land a kickflip 80% of the time.

With an elbow partner, do the first five steps of a hypothesis test.

A Hypothesis Test … with code

Matt is all about consistency in skateboarding, and is learning to kickflip. He does not want to move onto another trick until he can cleanly land a kickflip 80% of the time.

1. Can Matt land a kickflip with a success probability of at least 80%?
2. H_0: $p = 0.8$
3. H_a: $p > 0.8$
4. Under the null hypothesis, $X \sim \text{Binomial}(n, 0.8)$
5. It won't be so harmful for Matt to start practicing a new trick if he rejects the null hypothesis incorrectly, so let's be less skeptical than usual: $\alpha = 0.2$

Let’s collect some data.

Matt landed 84 out of 100 kickflips.

Calculate a p-value

```
prob_equal_or_more_extreme = 1 - binomial.cdf(83)
print("Probability of Observing Data More Equal or More Extreme than Ac
    prob_equal_or_more_extreme))
```

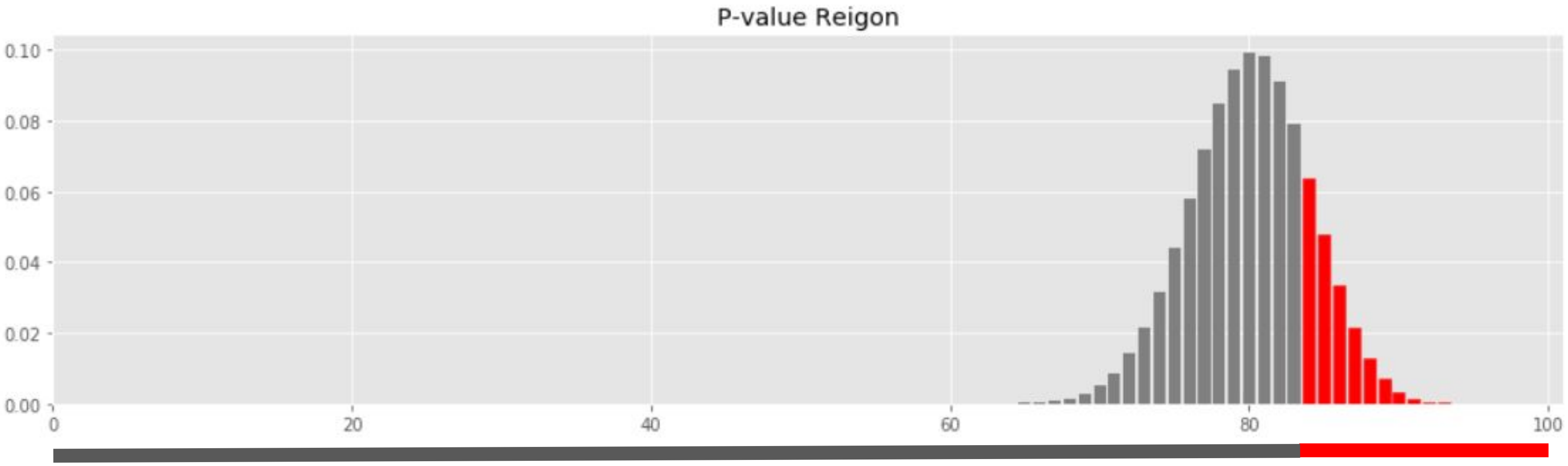
```
Probability of Observing Data More Equal or More Extreme than Actual:
0.19
```

A couple of questions arise here:

- Why are we calculating 1 - CDF and not just the CDF?
- Why is the argument for the CDF 83 when we observed 84 successful kickflips?

Calculate a p-value

Let’s look at a graph of the PMF of the binomial distribution (n=100, p=0.8):



binomial.cdf(83)

84 or more kickflips if H_0 is true

Compare p-value to threshold

We calculated a p-value of 0.19. We chose a significance level of 0.2. Should we reject the null hypothesis?

Compare p-value to threshold

We calculated a p-value of 0.19. We chose a significance level of 0.2. Should we reject the null hypothesis?

Yes. We can reject the null hypothesis at our chosen significance level. Matt can start working on new tricks.

Check for Understanding

By the end of this lesson, you will be able to:

- ❏ Calculate the p-value for a binomial test.

What code would you use to calculate the p-value for each of the following scenarios?

1. Your null hypothesis is that 60% of buses are late. Your alternative hypothesis is that more than 60% are late. Today you observe that 35 out of 50 buses are late.
2. Your null hypothesis is that 30% of all registered voters voted. Your alternative hypothesis is that fewer than 30% voted. A poll of 100 registered voters states that 23 of them voted.

Misconceptions about p-values

- **True or False Null Hypothesis**

  The interpretation of the p-value does not mean that the null hypothesis is true or false.

  It does mean that we have chosen to reject or fail to reject the null hypothesis at a specific statistical significance level based on empirical evidence and the chosen statistical test.

  You are limited to making probabilistic claims, not crisp binary or true/false claims about the result.

Source: Machine Learning Mastery

Misconceptions about p-values

- **p-value as Probability**

  A common misunderstanding is that the p-value is a probability of the null hypothesis being true or false given the data.

  Instead, the p-value can be thought of as the probability of the data given the pre-specified assumption embedded in the statistical test.

  The p-value is a measure of how likely the data sample would be observed if the null hypothesis were true.

Source: [Machine Learning Mastery]

Objectives

By the end of this lesson, you will be able to:

- ❑ State some common misinterpretations of p-values.

How would you correct each of the following statements?

1. “We obtained a p-value of 0.08. Therefore, at a 0.05 significance level the null hypothesis is true.”
2. “The p-value of 0.18 indicates that there is an 18% chance that the null hypothesis is true.”