Hypothesis Testing

Deciding What's True (Even If We're Just Guessing)

Let's Play A Game

A "Friendly" Introduction to Hypothesis Tests

Let's Play A Game

The semester is getting a little intense! You are a livinLet's break the tension with a friendly game.

Here are the rules:

- I have a new deck of cards. 52 Cards, 26 with Suits that are Red, 26 with Suits that are Black
- You draw a red-suited card, you give me a dollar
- You draw a black-suited card, I give you two dollars

Quick Questions

What is the expected number of dollars you will win playing this game?

Would you play this game?

Are We Forgetting Something?

Be Careful About Your Assumptions

Pause for a minute and think about what you are assuming is true when you play this game. For instance,

- You assume I'm going to shuffle the cards fairly
- You assume there are 52 cards in the deck
- You assume the deck has 26 red-suited cards in it
- You assume the deck has **a** red-suited card in it

How can we make sure the assumptions are safe??

- Shuffling assumption: watch me shuffle, make sure I'm not doing magic tricks, etc
- 52 Cards assumption: count the cards
- Red-suit assumption: Count the number of red cards

Whew! We can actually make sure all of our assumptions are good!

One Problem

I Refuse to Show You The Cards

Do You Trust Me?

Our Assumptions

I'm not going to show you all the cards. In other words, I refuse to show you the *population of possible outcomes*. This is justified: we are in a statistics course after all.

So, let's start with our unverifiable assumption: Is it safe to assume that this is a fair game. Why would we make this assumption?

- You trust that I'm (basically) an honest person (*assumption of decency*)
- You trust that I'm getting paid enough that I wouldn't risk cheating students out of money (*assumption of practicality*)
- You saw the deck was new (*manufacturer trust assumption*)
- You want it to be an fair game because you would win lots of money if it was (*assumption in self-interest*)

Our Assumptions

In statistical terminology, we wrap all these assumptions up into one assumption: our "**null hypothesis**" is that the game is not rigged - that the probability of you winning is 0.5

> **Null Hypothesis**
> The assumptions we are operate under in normal circumstances (i.e., what we believe is true). We wrap these assumptions up into a statistical/mathematical statement, but we will accept them unless we have reason to doubt them. We use the notation H_0 to refer to the null hypothesis.

In this case, we could say that the probability of winning is p and that would make our null hypothesis

$$H_0 : p = 0.5$$

Our Assumptions

Of course our assumptions could be wrong. We call the other assumptions our "alternative hypothesis":

> **Alternative Hypothesis**
> The conditions that we do require proof to accept. We would have to change our beliefs based on evidence. We use the notation H_A (or sometimes, H_1) to refer to the alternative hypothesis.

In this case, we could say that our alternative to believing the game is "fair" is to believe the game is not fair, or that the probability of winning is not 0.5. We write:

$$H_A : p = 0.5$$

A Compromise

I Won't Show You All The Cards

But I Will Let You Test The Game

Testing the Game

The test of whether or not the game is worth playing can be defined in term of whether or not our assumptions are true. In other words, we are going to test whether our null hypothesis is correct:

> **Hypothesis Tests**
> A **hypothesis test** is a way of checking if the outcomes of a random experiment are *statistically unusual* based on our assumptions. If we see really unusual results, then we have **statistically significant** evidence that allows us to **reject our null hypothesis**. If our assumptions lead to results that are not unusual, then we **fail to reject our null hypothesis**.

Testing the Game

So how can we test the game? What if we tried a single round of the game?

- What are the probabilities of the outcome of a single game?
- If we draw a single card do we have enough evidence that the game is fair?
- Do we have enough evidence that the game is rigged?

Based on a single round of the game, both of the possibel outcomes are pretty normal - that's not good enough.

If we draw a losing card, then we might be inclined to call the game unfair - even though a losing card is pretty common for a single round of the game

If we draw a winning card, then we might be inclined to call the game fair - even though a winning card may be common even when the game is not fair!

We can make lots of mistakes!!

The Mistakes We Might Make

We could of course be wrong: For instance, we could, just by random chance, see outcomes that are unusual for the assumptions we make and reject the assumptions even if (in reality they are true). This is called a "Type I Error"

> **Type I Error**
> When the results of a hypothesis test lead us to reject the assumptions, while the assumptions are actually true, we have committed a Type I Error.

The Mistakes We Might Make

A common example of this is found in criminal court:

- We assume that a individual accused of a crime is innocent (our assumption)
- After examinig the evidence, we conclude that it is there is no reasonable doubt the person is not innocent (in other words, we reject the assumption because it is very unlikely to be true based on our evidence).
- If the person truly is innocent, then we have committed a Type I error (rejecting assumptions that were true).

The Mistakes We Might Make

We could also make a different error: we could choose not to reject the assumptions when in reality the assumptions are wrong.

> **Type II Error**
> When the results of a hypothesis test lead us to fail to reject the assumptions, while the assumptions are actually false, we have committed a Type II Error.

The Mistakes We Might Make

Again, if we consider the example of criminal court:

- We assume that a individual accused of a crime is innocent (our assumption)
- After examinig the evidence, we conclude that it is there is **not** evidence beyond a reasonable doubt the person is not innocent (in other words, the evidence is not enough to reject our assumption because it is still reasonable to doubt the accused's guilt).
- If the person truly is not innocent, then we have committed a Type II error (failing to reject assumptions that were false).

In general, we want to make sure that a Type I error is unlikely. To take the example of court again,

- We commit a Type II error: a guilty person goes free
- We commit a Type I erro: an innocent person goes to jail; the guiilty person is still free

The Mistakes We Might Make

Let's go back to my game: We assume I am an honest person (i.e., we assume that the probability of winning a single game is $p = 0.5$)

Type I Error: Rejecting True Assumptions

- We gather evidence
- Looking at our evidence, we decide that the game was not fair even though it was.
- Fallout: you slander me, you disparge me, we have a fight, there is ill will in the class.

Type II Error: Failing to Reject False Assumptions

- We gather evidence
- Looking at our evidence, we decide that the game was fair even though it was not.
- Fallout: you play the game and lose some money.

Ideally, we won't make either error. However, we can only base our decision of our evidence we can gather - the truth is out of our grasp!

Gathering Statistical Evidence

Okay, so we don't want to make either error - that means we need good evidence.

Like we talked about before, even if the game is fair one test round of the game would not be enough to make a good decision since drawing a red-suited card and drawing a black-suited card are both pretty normal for a single round of the game.

But what if we played the game 10 times in a row? After 10 rounds, do you think we would have enough evidence to make a decision about our assumption?

If we assume the null hypothesis, then we can make some assumptions about what results are likely and what results are unlikely. We describe the likelihood of the results that we actually get using a **p-value**

> **p-value**
> After gathering evidence (aka, data) we can determine the probability that we would have gotten the evidence we did if our assumptions were true. That probabiliity is called the p-value. If the p-value is really, really small that means that the assumptions we started with are pretty unlikely and we reject our assumptions. If the p-values is not small, then the evidence collected (aka, the data) is pretty normal for our assumptions and we fail to reject our assumptions.

My Game

In other words, we collect evidence and determine a way to measure the whether or not our data was unusual *if our assumptions are true*.

If we have a very, very low chance of

- seeing both our results and
- having true assumptions then we reject the assumptions

Going along with the terminology we have introduced, if we have a small p-value the we reject our null hypothesis.

Gathering Statistical Evidence

In this game, if we assume that the game is fair, we have

- two outcomes: success (winning) and failure (losing)
- a constant chance of a successful outcome ($p = 0.5$, assuming the game is fair)
- independent rounds of the game (assuming fair shuffle, which we can check)

In other words, if we test the game 10 times we can model the number of successful outcomes as binomial: For X = the total number of wins,

$$P(X = x) = \frac{10!}{x!}(10 - x)!(0.5)^x(1 - 0.5)^{10-x}$$

This gives us a way of getting our p-value

Let's Test the Game

Gathering Statistical Evidence

We played the game. Let's figure out whether our results were unusual or not.

Again, we assume the game is fair and have decided that the number of times we win will follow a binomial distribution with probabiliity function

$$P(X = x) = \frac{10!}{x!}(10 - x)!(0.5)^x(1 - 0.5)^{10-x}$$

Now we need to make a conclusion: do we accept or reject our assumptions? What do we consider unusual? Is it fair to decide after we play?

Summary

- Sometimes we can know if something is true or not by examining the truth directly, but not always
- When we can't examine the truth, we need to test what we believe to be true
- A statistical test is a tool for testing our assumptions about what we believe
  - We state our assumed belief (generally our current beliefs, or the ethical beliefs, or the beliefs we hope are true, ...)
  - We come up with a way of collecting data that could validate or invalidate our assumption
  - We measure how likely it was that we would have gathered the data we did if our assumptions were correct
  - We reject the assumptions if our data is very unlikely we are our current beliefs