Simulation-based inference - hypothesis testing

Intro to Data Science

Shawn Santo

Today's agenda

- Introduce the hypothesis testing framework
- Perform statistical hypothesis tests using a simulation-based approach

Recall

Terminology

Population: a group of individuals or objects we are interested in studying

Parameter: a numerical quantity derived from the population (almost always unknown)

Sample: a subset of our population of interest

Statistic: a numerical quantity derived from a sample

Common population parameters of interest and their corresponding sample statistic:

Quantity	Parameter	Statistic
Mean	μ	$\bar{x}$
Variance	σ^2	s^2
Standard deviation	σ	s
Median	M	$\tilde{x}$
Proportion	p	$\hat{p}$

Statistical inference

Statistical inference is the process of using sample data to make conclusions about the underlying population the sample came from.

- **Estimation**: estimating an unknown parameter based on values from the sample at hand
- **Testing**: evaluating whether our observed sample provides evidence for or against some claim about the population

We will now move to testing hypotheses.

Testing

How can we answer research questions using statistics?

Statistical hypothesis testing is the procedure that assesses evidence provided by the data in favor of or against some claim about the population (often about a population parameter or potential associations).

Example:

The state of North Carolina claims that students in 8th grade are spending, on average, 200 minutes on Zoom each day. **What do you make of this statement? How would you evaluate the veracity of the claim?**

The hypothesis testing framework

1. Start with two hypotheses about the population: the null hypothesis and the alternative hypothesis.

2. Choose a (representative) sample, collect data, and analyze the data.

3. Figure out how likely it is to see data like what we observed, **IF** the null hypothesis were in fact true.

4. If our data would have been extremely unlikely if the null claim were true, then we reject it and deem the alternative claim worthy of further study. Otherwise, we cannot reject the null claim.

Two competing hypotheses

The **null hypothesis** (often denoted H_0) states that "nothing unusual is happening" or "there is no relationship," etc.

On the other hand, the **alternative hypothesis** (often denoted H_1 or H_A) states the opposite: that there is some sort of relationship (usually this is what we want to check or really think is happening).

In statistical hypothesis testing we always first assume that the null hypothesis is true and then see whether we reject or fail to reject this claim

1. Defining the hypotheses

The null and alternative hypotheses are defined for **parameters,** not statistics.

What will our null and alternative hypotheses be for this example?

- H_0: the true mean time spent on Zoom per day for 8th grade students is 200 minutes
- H_1: the true mean time spent on Zoom per day for 8th grade students is not 200 minutes

Expressed in symbols:

- $H_0 : \mu = 200$
- $H_1 : \mu \neq 200$,

where μ is the true population mean time spent on Zoom per day by 8th grade North Carolina students.

2. Collecting and summarizing data

With these two hypotheses, we now take our sample and summarize the data.

```
zoom_time <- c(299, 192, 196, 218, 194, 250, 183, 218, 207,
               209, 191, 189, 244, 233, 208, 216, 178, 209,
               201, 173, 186, 209, 188, 231, 195, 200, 190,
               199, 226, 238)
```

```
mean(zoom_time)
```

```
#> [1] 209
```

The choice of summary statistic calculated depends on the type of data. In our example, we use the sample mean: $\bar{x} = 209$.

Do you think this is enough evidence to conclude that the mean time is not 200 minutes?

3. Assessing the evidence observed

Next, we calculate the probability of getting data like ours, *or more extreme*, if H_0 were in fact actually true.

This is a conditional probability:

> Given that H_0 is true (i.e., if μ were *actually* 200), what would be the probability of observing $\bar{x} = 209$?"

This probability is known as the **p-value**.

4. Making a conclusion

We reject the null hypothesis if this conditional probability is small enough.

If it is very unlikely to observe our data (or more extreme) if H_0 were actually true, then that might give us enough evidence to suggest that it is actually false (and that H_1 is true).

What is "small enough"?

- We often consider a numeric cutpoint (the **significance level**) defined *prior* to conducting the analysis.
- Many analyses use $\alpha = 0.05$. This means that if H_0 were in fact true, we would expect to make the wrong decision only 5% of the time.

What can we conclude?

Case 1: $\text{p-value} \geq \alpha$:

If the p-value is α or greater, we say the results are not statistically significant and we **fail to reject** H_0.

Importantly, **we never "accept" the null hypothesis** -- we performed the analysis assuming that H_0 was true to begin with and assessed the probability of seeing our observed data or more extreme under this assumption.

Case 2: $\text{p-value} < \alpha$

If the p-value is less than α, we say the results are **statistically significant**. In this case, we would make the decision to **reject the null hypothesis**.

Similarly, **we never "accept" the alternative hypothesis**.

Ok, so what **isn't** a p-value?

> *"A p-value of 0.05 means the null hypothesis has a probability of only 5% of being true"*
>
> *"A p-value of 0.05 means there is a 95% chance or greater that the null hypothesis is incorrect"*

NO

p-values do **not** provide information on the probability that the null hypothesis is true given our observed data.

Ok, so what **isn't** a p-value?

Again, a p-value is calculated *assuming* that H_0 is true. It cannot be used to tell us how likely that assumption is correct. When we fail to reject the null hypothesis, we are stating that there is **insufficient evidence** to assert that it is false. This could be because...

- ... H_0 actually *is* true!
- ... H_0 is false, but we got unlucky and happened to get a sample that didn't give us enough reason to say that H_0 was false

Even more bad news, hypothesis testing does NOT give us the tools to determine which one of the two scenarios occurred.

What can go wrong?

Suppose we test a certain null hypothesis, which can be either true or false (we never know for sure!). We make one of two decisions given our data: either reject or fail to reject H_0.

We have the following four scenarios:

Decision	H_0 is true	H_0 is false
Fail to reject H_0	Correct decision	*Type II Error*
Reject H_0	*Type I Error*	Correct decision

It is important to weigh the consequences of making each type of error.

In fact, α is precisely the probability of making a Type I error. We will talk about this (and the associated probability of making a Type II error) in future lectures.

Let's conduct some hypothesis tests

- Create your personal private repository by clicking https://classroom.github.com/a/J6K9tCbW