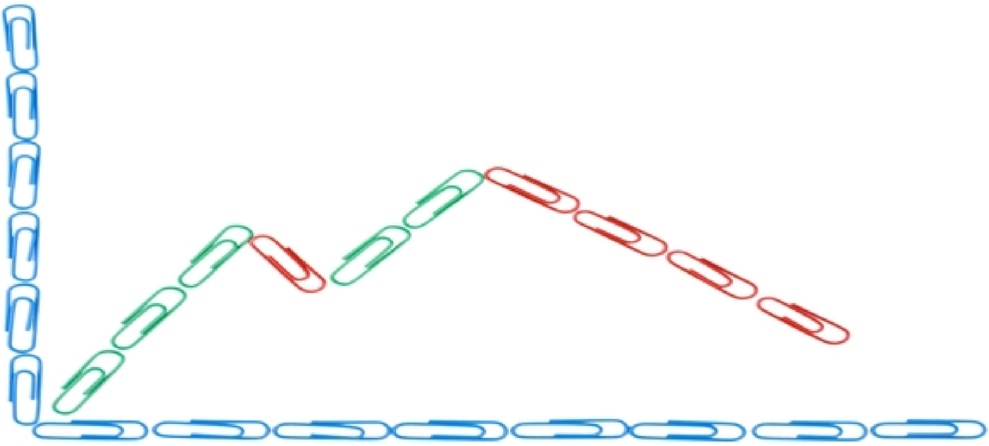
PURSUING DATA

# What is a p-value and how do you calculate it?

May 01



I had some challenges following along with the explanation that was given in class, so I decided to write one out for myself for future reference.

For context, we are using the bootstrapping methods (that I've referenced previously) for simulating null and sampling distributions (rather than standard statistical formulae) and so the methodology is a bit different than what would be done in traditional statistics courses.

## Definition

First of all, the definition of the p-value is: the probability of obtaining the observed statistic or a "more extreme" value (by extreme we just mean more in favour of the alternate hypothesis mean) if the null hypothesis is true. (If you need to know more about what null and alternate hypotheses are, Wikipedia has a pretty good definition).

**Note:** There's a whole other discussion to be had about what p-values mean if you assume that the alternate hypothesis is true, but that is for another day!

## What does this actually mean?

The definition is all well and good, but what does it mean when actually using it? There's two things we want to consider, how it informs decision making, and how to calculate it.

### Decision Making

The p-value helps us make a decision. Because of the way we construct our assumptions, when calculated, the p-value tells us the probability of committing a Type I error if the null hypothesis is true. (A Type I error is when you incorrectly reject the null hypothesis - usually we would consider making Type I errors to be 'bad,' so we want to make as few of them as possible, and so are happy to err on the side of caution and make this chance quite low)

A low p-value is often considered to be less than 0.05 in business and research, and 0.01 in medicine, but it could be any value appropriate to the situation. That is, if you get a p-value that is 0.05, this means that there is a 5% chance that a statistic that you observed came from a population where the null hypothesis is true. With this reasoning, at low p-values we typically reject the null hypothesis. That is, we act on the assumption that the observed statistic came from a population where the alternate hypothesis is true.

In standard methodology we don't just calculate the p-value and decide if it's low enough, we pick a threshhold beforehand (a priori - *sister wave*!), which is called the α level/value.

So if we calculate the p-value and it is below the α we make a decision to act as if the alternate hypothesis is true.

### Calculating

We know that we need to calculate a p-value. How do we do that? There are two things to consider, what do our null and alternate hypotheses say? (Which tells us generally what we'll need to calculate) And what values do we use to make the calculation? (How to actually do it).

At this point it is really better to start visualizing with a specific example, so here goes.

## The Example

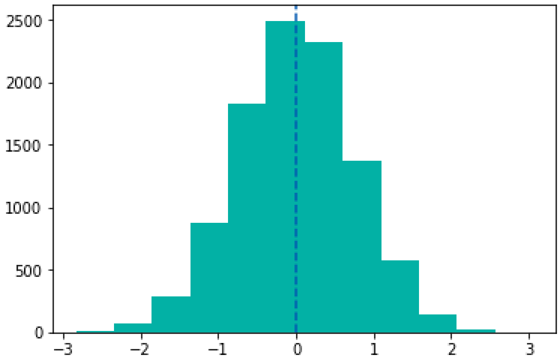
Let's say that you were looking at whether eldest childern have different IQs than the rest of the siblings.

While directionality is indicated, the general null hypothesis for this study is that there is no difference between the siblings. With this information we can simulate what the comparisons of siblings in the population would look like if the null hypothesis is true.

### Simulating the Population

To note, when conducting hypothesis testing, **we always simulate the null population** and then **compare to the observed statistic**.

To do this, we reference the null hypothesis. As we discussed above, the general null hypothesis is that there is no difference. That is, if we took the average IQ for all eldest children and compared that to the average IQ for all of their siblings, the difference would be zero. For the sake of argument, let's also simulate the population using a standard deviation of 0.75. (More technically, this would be the standard deviation of the sampling distribution of the differences that we created from the sample we took for our experiment) This allows us to set up the following representation of the null population.



The aqua space shows the spread of all of the differences in IQ between siblings that would be seen if the null hypothesis is true (the null population), and the blue dashed line shows the null mean (0). We can use this distribution to help us calculate our p-values.

## Hypotheses Version 1: Eldest Children have Higher IQs

In the first version we are going to create hypotheses in line with the directionality indicated by the study.

### Hypotheses

To test if eldest siblings have higher IQs, we would set up our hypotheses like this:

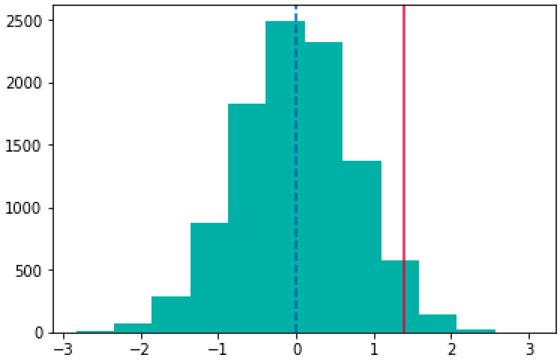
**$H_0$:** $\mu_{\text{eldest}} - \mu_{\text{non-eldest}} \leq 0$
**$H_1$:** $\mu_{\text{eldest}} - \mu_{\text{non-eldest}} > 0$

That is:

**Null Hypothesis:** The difference between the average IQ of eldest children and the average IQ of non-eldest children is equal to or less than zero.
**Alternate Hypothesis:** The difference between the average IQ of eldest children and the average IQ of non-eldest children is greater than zero.

### Calculation

Let's say that in our sample we observed that the average difference in IQ for eldest children compared to their siblings was 1.4 points. We can call this our observed mean. In terms of our null population distribution, it would fall here:



The red line shows the observed difference.

**We can now pick up from where we left off from when we started talking about how to calcalte the p-value.** Now that we have the distribution (aqua shaded area), the null mean (blue dashed line) and the observed mean (red line), we have everything we need to calculate the p-value. To do this are going to calculate the area for part of the aqua shaded distribution, but which part?

Here is where we consult our alternate hypothesis and look at the direction of the arrow. If the sign is greater than (with the tail pointing to the right), then we need to calculate the shaded area to the right of our observed value (the red line), or the proportion of values from our null distribution that are greater than the observed mean.

The code (using numpy, and where 'dist' is an array representation the distribution) for this would be as follows:

```
p_val = (dist > 1.4).mean()
```

This compares each value in the distribution to 1.4 and creates an array of these comparisons (For each comparison, if the null value is greater than the observed statistic, the result will be *True*, if not, it will be *False*). When calculating the mean, comparisons that were *True* are evaluated as 1 and comparisons that were *False* are evaluted as 0.

The value for 'p_val' wil be:

```
0.0294
```

Therefore, 2.94% of values from our null distribution fall to the right, or are above, our observed mean. Looking at how much of the aqua shading is to the right of the red line compared to the rest of the shading, this seems to be about right.

### Conclusion

Because we are doing some pretty casual research, let's assume that our $\alpha$ value is 0.05.
Our calculated p-value is below this and so we would reject the null hypothesis and say, "**Yes, on average, eldest siblings do have a greater IQ than their younger siblings!**"

## Hypotheses Version 2: Eldest Children have Lower IQs

Now let's switch it up and do the opposite version of the above hypotheses.

### Hypotheses

So this time we are looking at if eldest siblings have lower IQs, we would set up our hypotheses like this:

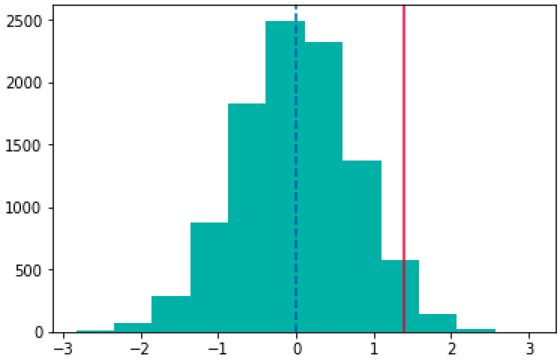
**$H_0$:** $\mu_{eldest} - \mu_{non\text{-}eldest} \geq 0$
**$H_1$:** $\mu_{eldest} - \mu_{non\text{-}eldest} < 0$

That is:

**Null Hypothesis:** The difference between the average IQ of eldest children and the average IQ of non-eldest children is equal to or greater than zero.
**Alternate Hypothesis:** The difference between the average IQ of eldest children and the average IQ of non-eldest children is less than zero.

### Calculation

For the sake of simplicity, let's keep all of the previous values the same - the average difference for the null hypothesis is 0, the standard deviation of the sampling distribution is 0.75, and the difference in IQ that is observed between eldest children and their siblings for our sample is 1.4. The chart will therefore look exactly the same:



But, because of our change in the hypotheses, what we calculate for the p-value will be different.

This time, because the alternate hypothesis talks about "less than 0" (with the tail of the sign pointing to the left), the area that we are going to calculate is everything shaded aqua that is to the left of the red line, or the proportion of values from our null distribution that are less than the observed mean.

To do this, we switch the direction of the sign in our code:

```
p_val = (dist < 1.4).mean()
```

This time, the value calculated for 'p_val' wil be:

```
0.9706
```

Again, when we look at how much aqua is to the left of the red line, this seems to make sense.

As a side note, because all we did was switch the signs, the total of this and the previously calculated p-value is one.

### Conclusion

This p-value is WELL above our previously established α level and so in this case we would fail to reject the null hypothesis.

Our conclusion would be, "**On average, the IQ of eldest children is greater than or equal to that of their younger siblings.**"

## Hypotheses Version 3: There is a difference in the IQs of eldest children and their younger siblings

Note that what is particular about this hypothesis is that a direction isn't specified, we are only looking for a difference. The difference could be higher or lower, and so we will need to consider that in our calculations.

### Hypotheses

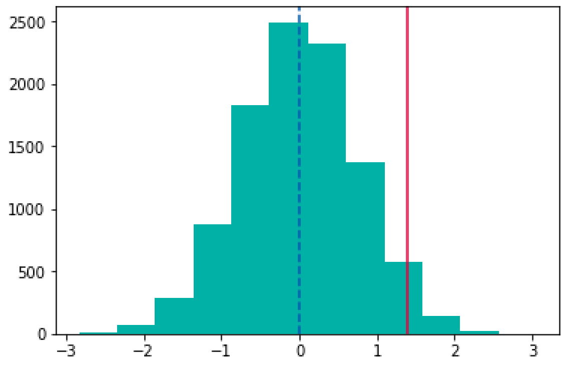
**$H_0$:** $\mu_{eldest} - \mu_{non\text{-}eldest} = 0$
**$H_1$:** $\mu_{eldest} - \mu_{non\text{-}eldest} \neq 0$

That is:

**Null Hypothesis:** There is no difference between the average IQ of eldest children and the average IQ of non-eldest children.
**Alternate Hypothesis:** There is a difference between the average IQ of eldest children and the average IQ of non-eldest children.

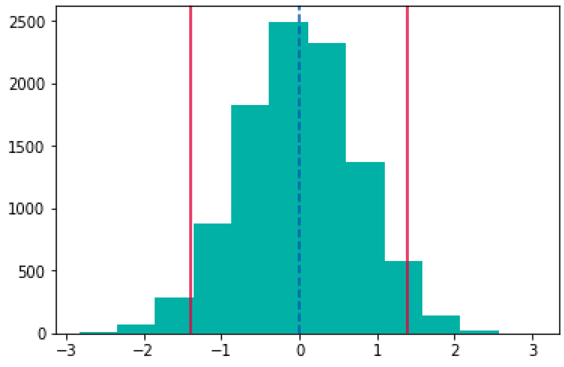
### Calculation

Again, we're keeping everything the same. As a reminder, the chart will look like this:



However, because of our hypothesis, here's where things get a bit different. We need to test any difference that we've observed between the null mean (0) and the observed mean (1.4) in both directions.

In other words, if we added an extra line to our chart to represent the difference between the null mean and the observed mean on the other side, we need to calculate the area that is outside both of the red lines.



Here's how we do that.

First, we look at the direction of the observed mean in relation to the null mean. In our case, the observed mean is greater than the null mean. So first of all, we are going to calculate all of the values of the null distribution that are higher than the observed mean, or the right/upper tail (we've done this before in our first example).

```
p_upper = (dist > 1.4).mean()
```

The value of 'p_upper' will be:

```
0.0294
```

Then we need to look at the other side/tail. For us that will be the left/lower tail. The calculation goes like this:

```
diff_means = 0 - 1.4
lower_compare = 0 + diff_means
p_lower = (dist < lower_compare).mean()
```

The first line calculates the difference between the null mean and the observed mean. The second line adds this to the null mean to get the difference on the other side. And then the comparison is made for all values in the distribution that are lower than the line on the left side.

The value for 'p_lower' will be:

```
0.0321
```

To find the total p-value, we add 'p_upper' and 'p_lower' together.

```
p_val = p_upper + p_lower
```

The p-value will be:

```
0.0615
```

### Conclusion

Again, our α level is 0.05. Our p-value (0.0615) is greater than this and so we would fail to reject the null hypothesis.

We would say, "**On average, there is no difference between the IQ of eldest children and their younger siblings.**"

One of the interesting things here is the impact of not selecting a direction in our hypothesis. If we select a direction, we have more 'space' within which to find a p-value that is lower than our αα level when compared to when we are just looking for a difference. (This is why when I did my stats education they strongly suggested that we used hypotheses that looked for a difference rather than a direction. You had to be VERY sure that you expected a certain direction if you wanted to pick that to prevent a greater chance of Type I errors.)

## Final Consideration - What does it MEAN?

Here's where things can get interesting. When we are doing null hypothesis statistical testing (NHST) people can get very focused on p-values. A result being significant is everything, and if the result is not significant, the work is often put to the side.

But p-values are not the only things that we should consider in reality. Let's think about the current situation, let's say that the study used the methodology of our first example and so the greater average IQ for eldest siblings was found. What would you do as an employer? **Would you include questions of sibling order in your hiring process to give you a better chance of hiring a more intelligent workforce?**

If you look at the size of the difference (called effect size), we're talking about less than 2 points of difference in IQ (In a range for most people that is 70 to 130). Can you imagine the lawsuits that could be directed towards your company if you declined to hire someone because they weren't an older sibling? In fact, this article suggests that in 4 out of 10 cases, the later-born is still smarter than their older sibling. That's an awful lot of potential lawsuits for what would seem to be a very tiny potential increase in productivity. It would seem that the downfalls associated with making decisions based on this information would be MUCH greater than not using this information.

This is why our instructor encourages us to consider not just statistical significance but also practical significance of our hypothesis testing.

Home

« Bootstrapping

Understanding Homelessness »