Cross Validated

Misunderstanding a P-value?

Asked 6 years, 1 month ago Active 3 years, 3 months ago Viewed 8k times

18

34

So I've been reading a lot about how to correctly interpret a P-value, and from what I've read, the p-value says NOTHING about the probability that the null hypothesis is true or false. However, when reading the following statement:

> The p – value represents the probability of making a type I error, or rejecting the null hypothesis when it is true. The smaller the p value, the smaller is the probability that you would be wrongly rejecting the null hypothesis.

EDIT: And then 5 minutes later I read:

> Incorrect interpretations of P values are very common. The most common mistake is to interpret a P value as the probability of making a mistake by rejecting a true null hypothesis (a Type I error).

This confused me. Which one is correct? And can anyone please explain how to correctly interpret the p-value and how it properly relates back to probability of making a type I error?

hypothesis-testing p-value

Share Cite Edit Follow Flag

asked Aug 9 '15 at 5:51

rb612

544 4 16

1 `The p – value represents the probability of making a type I error, or rejecting the null hypothesis when it is true` The p – value represents the apriori probability of making a type I error, that is, of rejecting the null hypothesis under the assumption that it is true. – ttnphns Aug 9 '15 at 8:17

4 @Paul: the probability to reject the null conditional on the null being true is the probability of a type I error, this is not the same as a p-value. The proability of a type I error is equal (for continuous random variables) to the choosen significance level, see also my answer below. – user83346 Aug 9 '15 at 11:51

Yes, I see now, you are absolutely right. – Paul Aug 9 '15 at 15:38

4 @fcoppens The probability of a type I error is only equal to the pre-chosen level of alpha if you condition on the null hypothesis being true. In an unconditional case you do not know if the null is true or false and so you can only specify a probability of a type I error if you supply a prior probability for the truth of the null. – Michael Lew Aug 16 '15 at 21:20

@Michael Lew: This conditioning on the null is mentioned in my answer below ? – user83346 Aug 17 '15 at 4:03

|

4 Answers

Active | Oldest | Votes

27

Because of your comments I will make two separate sections:

p-values

In statistical hypothesis testing you can find 'statistical evidence' for the **alternative** hypothesis; As I explained in What follows if we fail to reject the null hypothesis?, it is similar to 'proof by contradiction' in mathematics.

So if we want to find 'statistical evidence' then we assume the opposite, which we denote H_0 of what we try to proof which we call H_1. After this we draw a sample, and from the sample we compute a so-called test-statistic (e.g. a t-value in a t-test).

Then, as we assume that H_0 is true and that our sample is randomly drawn from the distribution under H_0, we can compute the **probability of observing values that exceed or equal the value derived from our (random) sample. This probability is called the p-value.**

If this value is ''small enough'', i.e. smaller than the significance level thase we have choosen, then we reject H_0 and we consider to H_1 is 'statistically proven'.

Several things are important in this way of doing:

- we have derived probabilities under the assumption that H_0 is true
- we have taken a random sample from the distrubtion that was assumed under H_0
- we **decide** to have found evidence for H_1 if the test-statistic derived from the random sample has a low probability of being exceeded. So it is not impossible that it is exceeded while H_0 is true and in these cases we make a type I error.

So what is a type I error: a type I error is made when the sample, **randomly** drawn from H_0, leads to the conclusion that H_0 is false while in reality it is true.

Note that this implies that a **p-value is not the probability of a type I error**. Indeed, a type I error is a wrong decision by the test and the decision can only be made by comparing the p-value to the choosen significance level, with a p-value alone one can not make a decision, **it is only after comparing the p-value to the choosen significance level that a decision is made**, and as long as no decision is made, type I error is not even defined.

What then is the p-value ? The potentially wrong rejection of H_0 is due to the fact that we draw a random sample under H_0, so it could be that we have ''bad luck'' by drawing the sample, and that this ''bad luck'' leads to a false rejection of H_0. So the p-value (although this is not fully correct) is more like the probability of drawing a ''bad sample''. **The correct interpretation of the p-value is that it is the probability that the test-statistic exceeds or equals the value of the test-statistic derived from a randomly drawn sample under H_0**

False discovery rate (FDR)

As explained above, each time the null hypothesis is rejected, one considers this as 'statistical evidence' for H_1. So we have found new scientific knowledge, therefore it is called a **discovery**. Also explained above is that we can make false discoveries (i.e. falsely rejecting H_0) when we make a type I error. In that case we have a false belief of a scientific truth. We only want to discover really true things and therefore one tries to keep the false discoveries to a minimum, i.e. one will control for a type I error. It is not so hard to see that the probability of a type I error is the chosen significance level α. So in order to control for type I errors, one fixes an α-level reflecting your willingness to accept ''false evidence''.

Intuitively, this means that if we draw a huge number of samples, and with each sample we perform the test, then a fraction α of these tests will lead to a wrong conclusion. It is important to note that we're **'averaging over many samples'**; so same test, many samples.

If we use the **same sample** to do many **different tests** then we have a multiple testing error (see my anser on Family-wise error boundary: Does re-using data sets on different studies of independent questions lead to multiple testing problems?). In that case one can control the α inflation using techniques to control the **family-wise error rate (FWER)**, like e.g. a Bonferroni correction.

A different approach than FWER is to control the **false discovery rate (FDR)**. In that case one controls the number of false discoveries (FD) among all discoveries (D), so one controls $\frac{FD}{D}$, D is the number of rejected H_0.

So the **type I error probability** has to do with executing the same test on many different samples. For a huge number of samples the type I error probability will **converge to the number of samples leading to a false rejection divided by the total number of samples drawn**.

The **FDR** has to do with many tests on the same sample and for a huge number of tests it will **converge to the number of tests where a type I error is made (i.e. the number of false discoveries) divided by total the number of rejections of H_0 (i.e. the total number of discoveries)**.

Note that, comparing the two paragraphs above:

1. The context is different; one test and many samples versus many tests and one sample.
2. The denominator for computing the type I error probability is clearly different from the denominator for computing the FDR. The numerators are similar in a way, but have a different context.

The FDR tells you that, if you perform many tests on the same sample and you find 1000 discoveries (i.e. rejections of H_0) then with an FDR of 0.38 you will have 0.38×1000 false discoveries.

Share Cite Edit Follow Flag

edited May 21 '18 at 0:57
Greenstick

answered Aug 9 '15 at 7:00
user83346


132 5

5 The correct interpretation of the p-value is that it is the probability that the test-statistic exceeds the value of the test-statistic derived from a randomly drawn sample under H0 Is so? Isn't it "equals or exceeds"? P-value is the prob that under true H0 we observe the difference or association *this or stronger* than the actually observed. – ttnphns Aug 9 '15 at 8:11

@ttnphns For a continuous test statistic there is no difference because the measure of a point is zero. For a discrete test statistic you are right (+1). I changed the text accordingly. – user83346 Aug 9 '15 at 8:17

1 You draw a very useful distinction between P-values and type I error rates, but I think you need to be more wary of the word "proven". Adding the modifier "statistically" does not soften it sufficiently, in my opinion. – Michael Lew Aug 16 '15 at 21:23

1 You have dealt with evidence as if it has only a binary state: exist and not exist. In the standard understanding of non-statistical evidence the word concept has a graded existence, and it is more complicated than a single dimension of strength can capture. The difficulty comes from the incompatibility of error rate considerations with ordinary interpretations of evidence. I would be very interested to read any account that captures non-binary interpretation of 'evidence' within the framework of FDR. (I haven't seen one yet.) – Michael Lew Aug 16 '15 at 21:25

1 Thank you for the correction. I made the pertinent change last night and credited your post. – Antoni Parellada Feb 17 '16 at 17:32

4

The first statement is not strictly true.

From a nifty paper on the misunderstanding of significance: (http://myweb.brooklyn.liu.edu/cortiz/PDF%20Files/Misinterpretations%20of%20Significance.pdf)

> "[This statement] may look similar to the definition of an error of Type I (i.e., the probability of rejecting the H0 although it is in fact true), but having actually rejected the H0, this decision would be wrong if and only if the H0 were true. Thus the probability "that you are making the wrong decision" is p(H0) and this probability... cannot be derived with null hypothesis significance testing. "

More simply, in order to assess the probability that you have incorrectly rejected H0 you require the probability that H0 is true which you simply cannot obtain using this test.

Share Cite Edit Follow Flag

edited Jun 11 '18 at 19:50

kjetil b halvorsen ♦
58.9k 23 134 431

answered Aug 9 '15 at 6:09

Henry B
111 2

Thank you! So when I'm reading the first part of statisticsdonewrong.com/p-value.html, the author concludes the FDR is 38%, so therefore the probability of a type I error is 38%? – rb612 Aug 9 '15 at 6:14

FDR is False Discovery rate and it is very different from type I error, so the answer to your question in no. FDR has to do with multiple testing, i.e. when you perform multiple tests on the same sample, see stats.stackexchange.com/questions/164181/... . FDR is an alternative to Familywise Error Rate, but

to explain that the number of characters in a comment is too limited. – user83346 Aug 9 '15 at 7:05

I added a second section in my answer for explaining FDR. – user83346 Aug 9 '15 at 8:01

1 Just as it is not possible to determine the probability of H0 being true without a prior, it is not possible to determine FDR without a prior. Be careful in how you interpret the FDR papers, because the priors used in them may not necessarily be relevant to your own experimental circumstances. – Michael Lew Aug 16 '15 at 21:33

2

The correct interpretation of a p-value is the *conditional* probability of an outcome at least as conductive to the alternative hypothesis as the observed value (at least as "extreme"), *assuming the null hypothesis is true*. Incorrect interpretations generally involve either a *marginal* probability or a switching of the condition:

$$\text{p-value} = \mathbb{P}(\text{At least as extreme as observed outcome}|H_0) \neq \mathbb{P}(\text{Type I error}).$$

Share Cite Edit Follow Flag edited Jun 12 '18 at 1:55

answered Jun 12 '18 at 1:48

Ben
78.1k 2 127 321

-1

The p-value allows us to determine whether the null hypothesis (or the claimed hypothesis) can be rejected or not. If the p-value is less than the significance level, α, then this represents a statistically significant result, and the null hypothesis should be rejected. If the p-value is greater than the significance level, α, then the null hypothesis cannot be rejected. This is the whole reason of looking up the p-value if you're using the table or using an online calculator, such as this one, p-value calculator, to find the p-value from the test statistic.

Now I know that you mentioned type I and type II errors. This really has nothing to do with the p-value. This has to do with the original data, such as the sample size used and the values obtained for the data. If the sample size is too small, for instance, this can lead to a type I error.

Share Cite Edit Follow Flag edited Jun 12 '18 at 0:59

answered Jun 11 '18 at 19:07

user1445657
7 1

2 -1. I'm sorry to welcome you to our site with a downvote, but this answer is plainly incorrect: it simply is not the case that the p-value is the probability of truth of the null hypothesis. This is amply discussed in many threads about p-values and hypothesis tests, such as stats.stackexchange.com/questions/31. – whuber ♦ Jun 11 '18 at 20:02

1 I modified the original answer a little to make it more precise. – user1445657 Jun 12 '18 at 1:02