Kevin Dunn <kgdunn@gmail.com>

[pystatsmodels] p-value problems persist...

8 messages

Warren Weckesser <warren.weckesser@gmail.com> Wed, Feb 12, 2014 at 1:28 PM
Reply-To: pystatsmodels@googlegroups.com
To: pystatsmodels@googlegroups.com

More p-value critique here:
http://www.nature.com/news/scientific-method-statistical-errors-1.14700

Warren

josef.pktd@gmail.com <josef.pktd@gmail.com> Wed, Feb 12, 2014 at 2:19 PM
Reply-To: pystatsmodels@googlegroups.com
To: pystatsmodels <pystatsmodels@googlegroups.com>

On Wed, Feb 12, 2014 at 1:28 PM, Warren Weckesser <warren.weckesser@gmail.com> wrote:

> More p-value critique here:
> http://www.nature.com/news/scientific-method-statistical-errors-1.14700

Interesting.
That's for most parts pretty much the worst case of "blame the tool" that I read in this discussion, instead of the attitude of the community. The last part is better.

It's just a number and not a magic or mystic creature. And it's a random number that is uniform distributed under the Null, and could be anything unconditionally.

If they would do a p-value correction for the multiplicity of data and method mining, then the p-value would have to be very, very small to maintain the level.

(back of the envelope:)
50 or a 100 decisions on the sample and methodology that can affect the test results, with a simple bonferroni correction your adjusted pvalue should be smaller than

```
>>> 0.05 / 50
0.001
>>> 0.05 / 100
0.0005
```

and then it's still just a p-value, although a corrected one.

BTW: I would expect that a 22% decrease in divorce rate has a significant effect on the income of divorce lawyers

Josef

> Warren

josef.pktd@gmail.com <josef.pktd@gmail.com> Wed, Feb 12, 2014 at 2:39 PM
Reply-To: pystatsmodels@googlegroups.com
To: pystatsmodels <pystatsmodels@googlegroups.com>

On Wed, Feb 12, 2014 at 2:19 PM, <josef.pktd@gmail.com> wrote:
>
>
>
> On Wed, Feb 12, 2014 at 1:28 PM, Warren Weckesser <warren.weckesser@gmail.com> wrote:
>>
>> More p-value critique here:
>> http://www.nature.com/news/scientific-method-statistical-errors-1.14700
>
>
> Interesting.
> That's for most parts pretty much the worst case of "blame the tool" that I read in this discussion, instead of the attitude of the community. The last part is better.
>
> It's just a number and not a magic or mystic creature. And it's a random number that is uniform distributed under the Null, and could be anything unconditionally.
>
> If they would do a p-value correction for the multiplicity of data and method mining, then the p-value would have to be very, very small to maintain the level.
>
> (back of the envelope:)
> 50 or a 100 decisions on the sample and methodology that can affect the test results, with a simple bonferroni correction your adjusted pvalue should be smaller than
>
> >>> 0.05 / 50
> 0.001
> >>> 0.05 / 100
> 0.0005
>
> and then it's still just a p-value, although a corrected one.

(Since I always feel I have to add the qualifiers to my answer.)

This wasn't supposed to be serious, however, I think this is correct
if the researcher looked at the results of 50 or 100 tests or versions
of the test and then reported the smallest p-value.
(and Bonferroni correction is conservative, so the actual family wise
error rate would most likely be smaller than 0.05.)

Josef
Being a statistician means never having to say you're certain.
(source ?)
[Quoted text hidden]

Sturla Molden <sturla.molden@gmail.com> Thu, Feb 13, 2014 at 6:22 AM
Reply-To: pystatsmodels@googlegroups.com
To: pystatsmodels@googlegroups.com

<josef.pktd@gmail.com> wrote:

> Interesting. That's for most parts pretty much the worst case of "blame
> the tool" that I read in this discussion, instead of the attitude of the
> community. The last part is better.

The p-value has been loathed by statisticans for decades. Anyone with more than superficial understanding of statistics knows the p-value bullshit and misleading. (The reasons for which have been explained hundreds of times before, so it serves no purpose to repeat them.)

Still, the p-value is commonly used among researchers. I believe there are two reasons, none of which the author mentions:

First, the number one priority of any working scientist is to get papers published. One would think it be to arrive at the truth, but it's not. What counts in a grant application is your publications and their impact factor, not the correctness of the results. Grants are what feeds researchers. Rearchers produce p-values and interpret them religiusly because they know refrees like them equally much – not surprising, since the refrees are their peers. And satisfying refrees is the key to get a paper on print. The refree might be an idiot, but the the refrees judgement still counts. Rebuttal letters serve no real purpose, the editor will ask you to abide by the refrees comments anyway. Scientists who know the p-values are bullshit will therefore gladly produce them anyway, because that is what it takes to get a paper published. A dog does not bite the hand that feeds it. I see this as an insult to the scientific process, but this is how it works.

Second, programs like SAS, STATA and SPSS make it easy to produce p-values. They make it hard to do something else, so why bother? The majority of reaserches lack any understanding of statistics anyway. They push the number in and hit the button, and SPSS spits out these magical numbers. And in lack of thorough understanding, they are even mistaken for the effect-size.

Sturla

---

josef.pktd@gmail.com <josef.pktd@gmail.com> Thu, Feb 13, 2014 at 8:01 AM
Reply-To: pystatsmodels@googlegroups.com
To: pystatsmodels <pystatsmodels@googlegroups.com>

On Thu, Feb 13, 2014 at 6:22 AM, Sturla Molden <sturla.molden@gmail.com> wrote:
> <josef.pktd@gmail.com> wrote:
>
>> Interesting. That's for most parts pretty much the worst case of "blame
>> the tool" that I read in this discussion, instead of the attitude of the
>> community. The last part is better.
>
> The p-value has been loathed by statisticans for decades. Anyone with more
> than superficial understanding of statistics knows the p-value bullshit and
> misleading. (The reasons for which have been explained hundreds of times
> before, so it serves no purpose to repeat them.)

Maybe you wanted to use a more civilized language?

There is no such thing as "the" "statisticians" in this debate.
"Anyone with more than superficial understanding of statistics"

Sounds like the typical polemic between schools of thought to me.
(Did I tell you that I got into the Bayesian - Frequentist debate a long time ago - on both sides.)

p-values won't go away, they are just a statistic from your data analysis, one among many.
It's usage and importance might change.

>
> Still, the p-value is commonly used among researchers. I believe there are
> two reasons, none of which the author mentions:
>
> First, the number one priority of any working scientist is to get papers
> published. One would think it be to arrive at the truth, but it's not. What
> counts in a grant application is your publications and their impact factor,
> not the correctness of the results. Grants are what feeds researchers.
> Rearchers produce p-values and interpret them religiusly because they know
> refrees like them equally much - not surprising, since the refrees are
> their peers. And satisfying refrees is the key to get a paper on print. The
> refree might be an idiot, but the the refrees judgement still counts.
> Rebuttal letters serve no real purpose, the editor will ask you to abide by
> the refrees comments anyway. Scientists who know the p-values are bullshit
> will therefore gladly produce them anyway, because that is what it takes to
> get a paper published. A dog does not bite the hand that feeds it. I see
> this as an insult to the scientific process, but this is how it works.

Yes, that's what I meant with that the problem of the p-value is
mainly it's usage in some research communities.
I read over the last few years many editorials, letter to the editors
and proposals to change this.

That's also why we have a strong "reproducibility" debate, that mainly
addresses research practices.

>
> Second, programs like SAS, STATA and SPSS make it easy to produce p-values.
> They make it hard to do something else, so why bother? The majority of
> reaserches lack any understanding of statistics anyway. They push the
> number in and hit the button, and SPSS spits out these magical numbers. And
> in lack of thorough understanding, they are even mistaken for the
> effect-size.

chicken and egg problem.
software packages provide what user communities want (and buy)

You are welcome to contribute more calculations for effect size and
their confidence intervals to statsmodels to complement the p-values.
There are still many missing.

http://jpktd.blogspot.ca/2013/03/different-fields-different-problems.html
https://groups.google.com/d/msg/pystatsmodels/JMMB2CrWjDY/Ek9vyFKKXSIJ
https://groups.google.com/d/msg/pystatsmodels/djwdaOKs9lE/ZRGGBQ1jZzwJ
https://groups.google.com/d/msg/pystatsmodels/4GcqR565oDY/VaNp3CIjNV4J

Note also that all estimation models and associated test report effect
size (in terms of parameters), standard errors, t- or z-values, and
p-values. Discrete models like Logit and Poisson, provide additionally
"Margins" to make it easier to interpret the (non-linear) effects.

Josef

>
> Sturla
>

Emanuele Olivetti <emanuele.olivetti@gmail.com> Thu, Feb 13, 2014 at 8:43 AM
Reply-To: pystatsmodels@googlegroups.com

To: pystatsmodels@googlegroups.com

Hi,

That short article is nice and well written, in my opinion. It does not sound at all "Frequentist vs Bayesian", but more Fisher vs Neyman-Pearson (Frequentist A vs Frequentist B). And even though I am not much keen to Frequentism, I don't understand why Fihserian significance testing has such wide adoption in almost all fields of experimental science while Neyman-Pearson hypothesis testing is barely known. The first approach models just the null hypothesis and so can attempt conclusions only about that, while the second approach explicitly model both null and alternative hypotheses, so that it is possible to argue about the alternative. My personal experience in experimental science teaches me that scientists really want to make statements about the alternative hypothesis. So... why Fisher?

Maybe one explanation is that it is usually way easier to model the null-hypothesis while it is a lot more difficult to model the alternative hypothesis. But you know, no free lunch... :D

Moreover, if you can afford to model the alternative too, why not going Bayesian? ;)

Best,

Emanuele

[Quoted text hidden]

---

josef.pktd@gmail.com <josef.pktd@gmail.com> Thu, Feb 13, 2014 at 10:16 AM
Reply-To: pystatsmodels@googlegroups.com
To: pystatsmodels <pystatsmodels@googlegroups.com>

On Thu, Feb 13, 2014 at 8:43 AM, Emanuele Olivetti
<emanuele.olivetti@gmail.com> wrote:
> Hi,
>
> That short article is nice and well written, in my opinion. It does not
> sound at all "Frequentist vs Bayesian", but more Fisher vs Neyman-Pearson
> (Frequentist A vs Frequentist B). And even though I am not much keen to
> Frequentism, I don't understand why Fihserian significance testing has such
> wide adoption in almost all fields of experimental science while
> Neyman-Pearson hypothesis testing is barely known. The first approach models
> just the null hypothesis and so can attempt conclusions only about that,
> while the second approach explicitly model both null and alternative
> hypotheses, so that it is possible to argue about the alternative. My
> personal experience in experimental science teaches me that scientists
> really want to make statements about the alternative hypothesis. So... why
> Fisher?

I have to say that I'm muddled about what the difference between
Fisher and Neyman-Pearson is. I never tried to really figure out the
difference.

Isn't `reject = (pvalue <=0.05)`
and `reject = (statistic >= threshold(0.05))
the same thing.

Why I like p-values:

p-values are just a continuous summary of the evidence in your data.
A p-value of 0.001 means it is very unlikely that you would have
gotten this sample if the Null Hypothesis were really true (unless you
biased your experiment).

As a summary statistic it doesn't imply any decision, or how you
interpret it. (no magic 0.05)

Reporting the rejection decision, converts a continuous variable into
a binary variable, and doesn't allow a user to have a different alpha.

Decision

In terms of decision theory, I'm almost completely Bayesian
http://www.springer.com/statistics/statistical+theory+and+methods/book/978-0-387-96098-2
(Once upon a time I was a theoretical economist, and all decisions of
individuals where based on beliefs. Beliefs also incorporate
information that is available to the individual and is not too
costly.)

Would I bet money based on an outcome where pvalue < 0.05?
No, not if I have a lot more information about the problem and the
data, and after merging it with my "prior" information.

To me it sounds a bit like the difference between Bayesian and
Likelihood based inference.
If you report a posterior, then you mix **your** prior with the data.
But if I don't agree with your prior, then I would prefer just to get
the information that is in the data, i.e. the likelihood.

If the information contained in the data is weak, then we wouldn't
change our belief by much. If the evidence is strong, then the prior
doesn't matter much, unless it's dogmatic.

P-values are like the likelihood, they summarize the evidence in the
data. Calling something statistically significant at 5% is a decision
and doesn't mean much by itself. It's still information, but very
coarse.

I see only partially your point about there being a difference in
terms of alternative in the two approaches:

p-values still depend on the alternative. They are different if we
have one-sided or two-sided t-test for example.
But the p-value itself doesn't imply a decision (and it's not a
sufficient statistic, so by itself it's not very informative).

If we follow a decision like reject if p-value <= alpha, then we
can calculate the type 1 and type 2 errors, the power against
different alternatives, based on the information in our sample.
That's a different piece of important information.

In my opinion, p-values are just another results statistic. We should
not assign to it an importance and a meaning that it doesn't have, but
I also don't see a need to throw it out with the bathwater.

>
> Maybe one explanation is that it is usually way easier to model the
> null-hypothesis while it is a lot more difficult to model the alternative
> hypothesis. But you know, no free lunch... :D

it's a lot more difficult
A simple t-test has a continuum of alternatives.

Which alternative is important?

It's a bit like eliciting prior information and making decisions.
That's up to the user, not the statistician.
The actual calculations might not be that difficult.
Or in equivalence tests, TOST: what's the threshold for considering two results to be equivalent for "practical" purposes? Based on my readings that's one of the difficult parts in applying equivalence tests.

(
Pitman local alternatives
http://www.jstor.org/discover/10.2307/3532049?uid=3739808&uid=2&uid=4&uid=3739256&sid=21103500322883
)

>
> Moreover, if you can afford to model the alternative too, why not going
> Bayesian? ;)

Because as a Bayesian you not only have to model the alternative, you also have to assign a prior probability (belief) to it - more work. ;)

What if your prior is "wrong"?

Advertising:
Robust Statistics: We get results for you, even if you don't know much about the true model (or don't have any strong beliefs about it).
You will get results without going "meta" and specify your uncertainty about your uncertainty about your uncertainty ... (*)

(Since I'm a frequentist: true model refers to fully specified likelihood.)

Josef
(*) hierarchical models ?
[Quoted text hidden]

Sturla Molden <sturla.molden@gmail.com> Fri, Feb 14, 2014 at 12:59 PM
Reply-To: pystatsmodels@googlegroups.com
To: pystatsmodels@googlegroups.com

On 13/02/14 14:43, Emanuele Olivetti wrote:
> So... why Fisher?

Fisherian hypothesis testing fits well with the hypothetico-deductive method of Karl Popper. It is only testing if observed data is consistent with a model. It is pure falsificationism, and consistent with how natural science usually is carried out, and is simply explained by the "theory of science" that researchers in the natural sciences are used to.

Fisherian confidence intervals says something about which values of the parameter would be deemed consistent with the data, and has a simple and understandable interpretation. (Unfortunately, most stats textbooks cites the NP interpretation of a confidence interval.)

Neyman-Pearson is a decision-theoretic approach, but its emphasis on keeping a low and fixed alpha-value often makes it arrive at unreasonable conclusions. And if you swap H0 and HA, it will often give you a different conclusion from the same data. The procedure can make you select the model least consistent with the data, depending on which is H0 and which is HA, and simply cannot be trusted. And that swapping the ordering of H0 and HA should make you prefer a different model given the same data is counterintuitive.

NP confidence intervals says something about the long run distribution of the intervals themselves, not the parameter, and we only have one sample from these intervals. NP confidence intervals have no interpretation that normal humans can understand, and they are invariably (and wrongfully) interpreted as Bayesian HDRs by most practitioners.

If you are a frequentist, chances are you will prefer the Fisherian approach. But if you want a decision-theoretic approach to hypothesis testing, the Bayesian method is almost always preferred to NP. The Bayesian prior is a minor nuisance compared to the inconsistencies of NP testing. If the prior matters in Bayesian statistics, it (almost always) means we need to sample more data. That is far easier to accept than a procedure that can give you different conclusions from the same data.

Also see here:
http://www.stat.ualberta.ca/~wiens/stat665/TAS%20-%20testing.pdf

Sturla