Communicating uncertainty of scientific studies: focusing on 50 shades of gray rather than an accept-and-reject world

Sandra Hamel & Nigel G. Yoccoz
Department of Arctic and Marine Biology, UiT the Arctic University of Norway

/ BACKGROUND /

Evidence from most studies are based on a black and white determination of statistical significance, completely neglecting uncertainty. This deterministic thinking is problematic because statistical significance on its own tells nothing about the magnitude of the effect, its practical significance, and the uncertainty surrounding the evidence. Resulting in high risk for making an interpretation mistake or taking the wrong decision.

Statistical significance is often a sufficient way of evaluating evidence. Statistical significance, which ever form it takes, can always be reached if we try hard enough, e.g. increasing sample size, forking paths, p-hacking, HARKING/HARKing – Making Hypothesis/Justifying After the Results are Known. This can be simple because we rarely have an effect that is exactly 0 in anything we measure. It is important whether is the effect we obtained is large compared to measurement variability, and whether its magnitude is in the range expected and its sign in the direction expected.

/ BLACK OR WHITE ISSUES /

Deterministic thinking

When thinking of examples of deterministic evidence, hypothesis testing with p-values comes first in mind. P-values are typically used as a black or white proof of evidence, but they are actually a continuous measure of evidence.

Think about this comparison:

- P = 0.04 – typically described as significant
- P = 0.05 – typically described as NOT significant

BUT: These two P-values are providing similar evidence. The threshold for evidence versus no evidence is arbitrary and subjective, and most importantly it was never meant to be universal. This universally encourages mechanical, automatic, unthinking procedures.

Using no threshold should force us to think about the results and their impacts.

What is the meaning of a P-value?

Problems with misuse of P-values are still omnipresent in all fields of science. The blind use of P-values is persisting and is one of the main causes behind the recent replication crisis.

A common misconception:

A P-value is the probability that H_0 is true. This is a misconception because there is no evidence on how compatible the data observed are with a specific model or hypothesis.

H_0: Height belongs to a population of men

A mean value of 188 gives a P = 0.05. We reject H_0 because the other sex is on average younger giving that H_0 is false (true) at 0.05.

What if the alternative hypothesis is wrong?

H_0: Height belongs to a population of women

Pearson 1: evidence is actually higher for when H_0. Pearson 2: evidence is truly higher for H_0. Pearson 3: evidence is similar for both H_0 and H_0.

P-values & similar methods

Other methods proposed, e.g. Bayesian factors, Information Criteria like AIC, likelihood ratio, or confidence/Confidence intervals, will all suffer from the same flaw. If they are also used as a black and white threshold for evidence.

/ WELCOMING UNCERTAINTY /

We should not aim for certainty: It is unlikely achievable. Instead, we should acknowledge uncertainty.

In study design:

We need to design studies to try to minimize uncertainty as much as possible to be able to show convincing evidence. The aim is to meet the models implicit requirements, i.e. that effects are large compared with measurement variability. This has nothing to do with having a large or small sample size. It means we need to have high quality measurements.

In data analysis:

We should move towards a more mechanistic approach because it allows to highlight uncertainty in the system we are studying.

In interpretation and presentation of results:

We need to be aware of the risk of making the wrong decision due to Type I and Type II error. Using the magnitude and the sign of the effects obtained. We need to present effect sizes and their uncertainties, along with what the uncertainty means for the effect we are studying.

“A number with no context is just a number, a meaningless number”3

/ CONFIDENCE INTERVALS /

Precision of estimates

Confidence intervals (CI’s) are the range of values for a parameter (e.g. the mean) that will in theory include the real population value for a range of sample sizes drawn from the population distribution of results.

Choosing the right confidence interval level

The right level of CI should be chosen on the basis of the question we are aiming at, not necessarily presenting only the classical 95%.

Comparing two weather-related forecasts

The 5-day and 10-day forecast for temperatures in Mürren have a validity of about 10%. Because of the huge risk to human life in lack of such storms, we need as many details as possible on the uncertainty of the path.

/ RELIABILITY AND VALIDITY /

Judging the evidence

Interpretation should be done with the task of presenting and judging the evidence. A good design, a frequentist procedure should be robust in a range of frequency distributions, a Bayesian inference should be robust to alternative priors, and a model should provide valid predictions.

Evaluating and accounting for model uncertainty

Scientists working within the North American Waterfowl Management Plan developed a model to maximize the potential for harnessing wildlife in terms of wildlife population size over a long time frame. Because of uncertainty on how the population respond to hunting, they built 4 competing models with different assumptions for reproductive rates and density-dependence.

The first year, each model had an equal weight in the global model for predicting population size the next year. By the 5th time, the weight of the few models evolved depending on which models more reliably predicted yearly observations.

Now, remember that the best of a set of poor models is still a poor model. Therefore, they compared the observed population sizes with the average model prediction each year (arrows in the figure). They found a relatively good fit, indicating the global model provided valid predictions.

/ CONCLUSIONS – MOVING FORWARD /

1) From the beginning of a study, we need better awareness of these issues to prompt researchers to build better designs and make better measurements.

2) We need to emphasize on estimation and its realistic and biological importance:

- Focus on size, consistency and precision.
- Focus on the realism of the estimation based on the expected effect.

3) We work with continuous measures of evidence, so we need to avoid using a razor blade threshold. Threshold should be decided in line with the evidence or with the risk posed to science and society of false positive or negative results.

4) We need to make better use of statistics for inferring and judging evidence:

- Interpretation should be done by presenting evidence as a combination of function of the effect size and its uncertainty, the potential for type I or type II errors, and choice of the study.
- The deterministic approach can be useful for exploratory analyses or pilot studies, but we need to be honest about presenting and discussing results by pinpointing that we are exploring potential effects.

Statistics are much more a gradient of 50 shades of gray than the classic black and white vision: It is all about informed judgment, self-examination and criticism, and transparency5.