UiT THE ARCTIC UNIVERSITY OF NORWAY

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 around this evidence, resulting in high risk for making an interpretation mistake or taking the wrong decision.

Statistical significance is not 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/JARKing – Making Hypothesis/Justifying After the Results are Known. This is simply because we rarely have an effect that is exactly 0 in anything we measure. What is important is whether the effect we obtained is large compared to measurement variability, and whether its magnitude is in the range expected and its sign is in the direction expected.

/ BLACK OR WHITE ISSUES

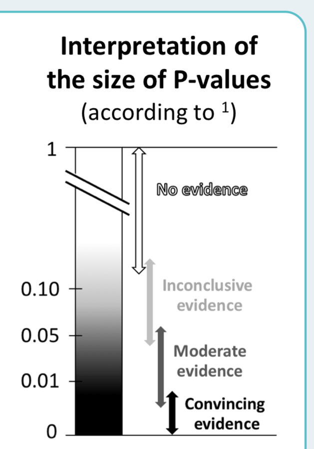
Deterministic thinking

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=0.06 –> 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 universality encourages mechanical, automatic, unthinking procedures[2].



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

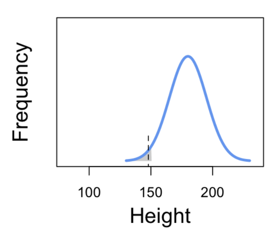
What is the meaning of a P-value?

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 cause behind the recent replication crisis.

A common misconception:
A P-value is ~~the~~ not the probability that H_0 is true ✗

A P-value only tells us the strength of evidence on how compatible the data observed are with a specific model or hypothesis.

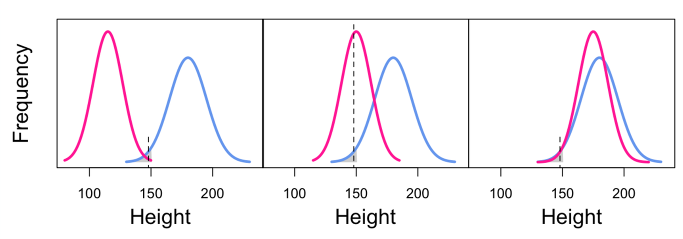
H_0: Height belongs to a population of men



A mean value of 148 gives a P<0.05. We reject H_0 because the chance of having this mean or lower giving that H_0 is true is below 0.05.

What if the alternative hypothesis is worst?

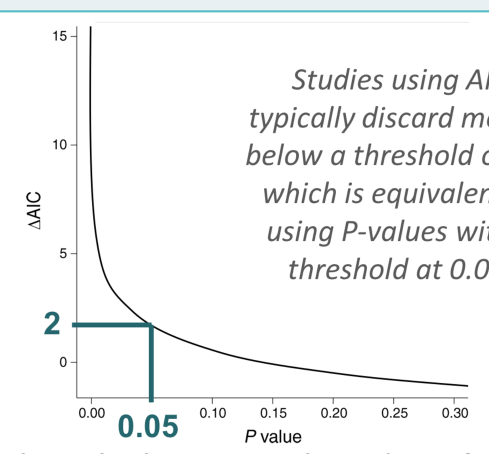
H_A: height belongs to a population of woman



Panel 1: evidence is actually higher for H_0 than H_A.
Panel 2: evidence is truly higher for H_A than H_0.
Panel 3: evidence is similarly for both H_0 and H_A.

P-values & similar methods

Other methods proposed, *e.g. Bayesian factors, Information Criteria like AIC, likelihood ratios, or credible/confidence intervals,* will all suffer from the same flaws if they are also used as a black and white threshold for evidence[3].



Relationship between P-value and ΔAIC for models with a large sample (reproduced with permission from [4])

Any method will be problematic if it is applied blindly to all situations. Replacing one problem with new one is definitely not the solution!

Why is the deterministic approach dominating?

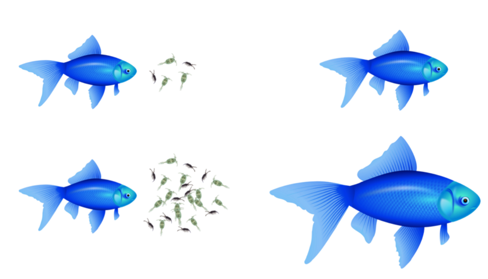
- ***Ease***: **less thinking required**
- ***Education*** **: this is what is taught!**
- ***Software***: **automaticity of inference**
- ***Publication***: **paper without significant results will often not be published**

/ 50 SHADES OF GREY

Moving from statistical significance to statistical estimation

Many H_0 hypothesis testing are irrelevant because they do not allow to answer the research goals or the complex questions we are asking nowadays.

H_0: A fish eating more will not differ in mass compared with a fish eating less



Trivial, we EXPECT the fish to be bigger. What is of interest is how much bigger, e.g. depending on the amount or type of food

We need to focus on statistical estimation, its uncertainty, along with what is expected[5,6].

On their own, P-values, Bayes factors or any other information criteria do not say anything about the magnitude of an effect, merely whether or not an effect is more or less likely to exist.

Which of these two results provide support for a stronger effect, P=0.01 or P=0.001?

NEITHER!
To answer which effect is stronger, we must evaluate each effect size and its uncertainty.

Type S and Type M errors

Gelman & Carlin[5] proposed design calculations focused on Type S and M errors, which are based on *expected effect size* (**EXP**) – typically assessed from empirical studies or data – the *standard error* (**SE**) obtained in the study, and the *level of significance* (**α**) used.

Type S - sign	***Type M - magnitude***
Probability of an estimate being in the wrong direction	Factor by which the magnitude of an effect might be overestimated

They referred to a study[7] comparing sex ratio of first child and attractiveness in humans, reporting a difference (±SE) of 8% (±3.3%), p=0.015. BUT:

Literature suggests effect size are rarely over 1%

What are the Type S and Type M errors?

EXP: 1%
SE: 3.3%
α: 0.05

Type S = 19%
Type M= factor of 8

Comparing some studies reporting effect sizes of reproduction on survival in birds

Reference	Sample Size	Effect Size	St. Err.
Nur 1988	216	-0.07	0.041
Nur 1988	113	0.15	0.070
Dijkstra *et al.* 1990	155	-0.198	0.108
Korpimäki 1988	65	0.135	0.368
De Steven 1980	40	0.187	0.328
Askenmo 1979	111	-1.33	0.532
Boyce & Perrins 1987	461	-0.005	0.030
Reid 1987	656	-0.094	0.042
Pettifor 1993	489	-0.044	0.045
Pettifor 1993	229	-0.312	0.098

This effect size is rather large

The mean effect size for all these studies is -0.05
Design calculations for Askenmo 1979:

Type S = 61%
Type M= factor of 25

/ 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 minimise 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 M and Type S errors, i.e. in 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.

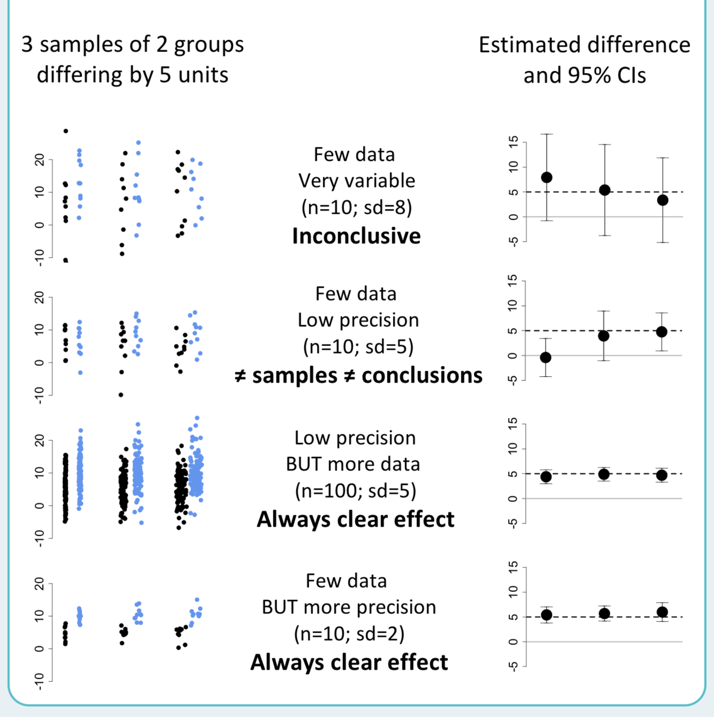
"A number with no context is just a number, a meaningless number"[8]

/ CONFIDENCE INTERVALS

Precision of estimates

Confidence intervals (CIs) are the range of values for a estimate (e.g. the mean) that will in theory include the real population value for a percentage of samples drawn from the population.

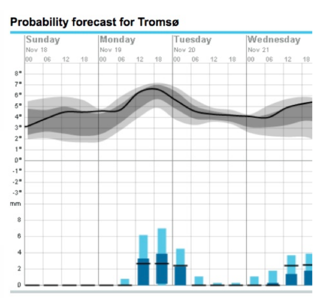
CIs give the precision around the estimates



Choosing the right confidence interval level

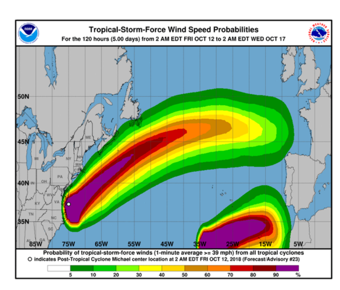
The right level of CI should be chosen on the basis of the question we are aiming to answer, not necessarily presenting only the classical 95%.

Comparing two weather-related forecasts



The daily weather forecast for Tromsø (yr.no), showing the 50% (dark colors) and 80% (light colors) CIs for t° and precipitations. This is enough to decide if we should take a rain jacket!

The 5-day wind forecast for cyclone Michael (noaa.gov), showing all the CIs from 5 to 100%. Because of the huge risk to human life of such storms, we need as many details as possible on the uncertainty of the path.



/ RELIABILITY AND VALIDITY

Judging the evidence

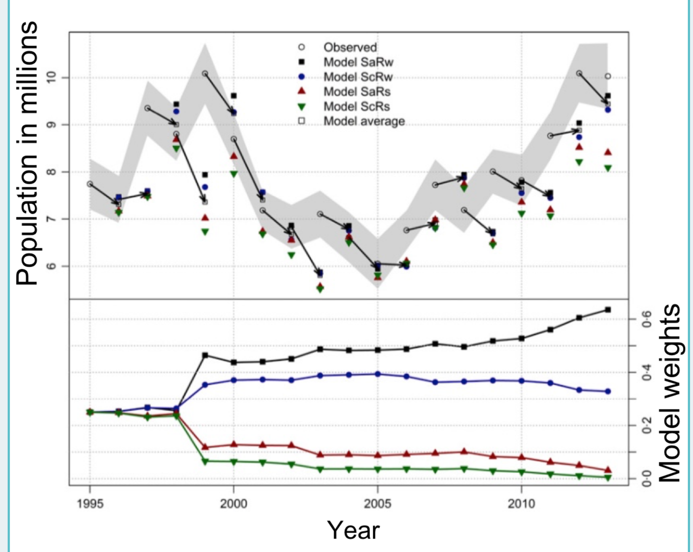
Interpretation should be done with the task of presenting and judging the evidence. With a good design, a frequentist procedure should be robust in a range of frequency distributions, a Bayesian inference should be robust to alternative priors, and a model should provide valid predictions.

Evaluating and accounting for model uncertainty

Scientists working within the North American Waterfowl Management Plan developed a model to maximize the potential for harvesting mallards in relation to changes in population size over a long time frame[9]. Because of uncertainty on how the population respond to harvest, 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. With time, the weight of the 4 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 size with the average model prediction each year (arrows in the figure). They found a relatively good fit, indicating the global model provided **valid predictions**.



Observations of mallards population size (grey: 95% CI) and predictions according to four different models, with the model weights evolving with time as greater evidence is accumulating in support of some models. (reproduced with permission from [9])

/ 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 direction of results.
- 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 combined function of the effect size and its uncertainty, the potential for type M or Type S errors, and the design of the study.
- The deterministic approach can be useful for exploratory analyses or pilot studies, but we need to be honest when presenting and discussing results by pinpointing that we are exploring potential effects.

Statistics are much more a gradient of 50 shades of grey than the classic black and white vision: it is all about informed judgment, self-examination and criticism, and transparency[10].

1. Ramsey, F. & Schafer, D. 2013. The statistical sleuth: a course in methods of data analysis. Duxbury Press
2. Mayo, D.G., & Cox, D.R. 2006. Optimality, 77-97, Institute of Mathematical Statistics
3. Van Helden, J. 2016. Nature Methods 13, 605-606
4. Murtaugh, P.A. 2014. Ecology 95, 611-617
5. Gelman, A., & Carlin, J. 2014. Perspectives on Psychological Science 9, 641-651
6. Claridge-Chang, A., & Assam, P.N. 2016 Nature Methods 13, 108-109
7. Kanazawa, S. 2007. Journal of Theoretical Biology 244, 133-140
8. Stigler, S.M. 2016. The seven pillars of statistical wisdom, Harvard University Press
9. Nichols, J., et al. 2015. Journal of Applied Ecology 52, 539-543
10. Beninger, P.G., et al. 2012. Journal of Experimental Marine Biology and Ecology 426-427, 97-108

UNIVERSITETET I TROMSØ