Eur J Epidemiol
DOI 10.1007/s10654-015-0054-1

DIALOGUE

# The researcher and the consultant: from testing to probability statements

Ghassan B. Hamra[1] · Andreas Stang[2,3] · Charles Poole[4]

Received: 7 April 2015 / Accepted: 30 May 2015

**Abstract** In the first instalment of this series, Stang and Poole provided an overview of Fisher significance testing (ST), Neyman–Pearson null hypothesis testing (NHT), and their unfortunate and unintended offspring, null hypothesis significance testing. In addition to elucidating the distinction between the first two and the evolution of the third, the authors alluded to alternative models of statistical inference; namely, Bayesian statistics. Bayesian inference has experienced a revival in recent decades, with many researchers advocating for its use as both a complement and an alternative to NHT and ST. This article will continue in the direction of the first instalment, providing practicing researchers with an introduction to Bayesian inference. Our work will draw on the examples and discussion of the previous dialogue.

**Keywords** Significance testing · Probability statements · Bayesian inference · Confidence intervals

**Electronic supplementary material** The online version of this article (doi:10.1007/s10654-015-0054-1) contains supplementary material, which is available to authorized users.

✉ Ghassan B. Hamra
gbh28@drexel.edu; ghassan.b.hamra@drexel.edu

[1] Department of Environmental and Occupational Health, Drexel University School of Public Health, 3215 Market St., Philadelphia, PA 19104, USA

[2] Center of Clinical Epidemiology, Institute of Medical Informatics, Biometry and Epidemiology, University Hospital Essen, Essen, Germany

[3] Department of Epidemiology, School of Public Health, Boston University, Boston, MA, USA

[4] Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC, USA

## Introduction

In the first instalment of this series, Stang and Poole [1] provided an overview of Fisher significance testing (ST), Neyman–Pearson null hypothesis testing (NHT), and their unfortunate and unintended offspring, null hypothesis significance testing (NHST). In addition to elucidating the distinction between the first two and the evolution of the third, the authors alluded to alternative models of statistical inference; namely, Bayesian statistics. Interestingly, Bayes' theorem existed long before development of Fisher's ST or Neyman–Pearson's NHT [2]; the latter two have dominated statistical thinking, notably in the field of public health, for decades. Nonetheless, Bayesian inference has experienced a revival in recent decades, with many researchers advocating for its use as both a complement and an alternative to NHT and ST [3].

This article will continue in the direction of the first instalment, providing practicing researchers with an introduction to Bayesian inference. We will discuss the theoretical framework for Bayesian analyses and some practical considerations. We emphasize the distinction between the previously discussed models of inference and Bayesian methods by focusing on and explaining how researchers can integrate prior information into Bayesian inference, by design. We will continue the dialogue between researcher and consultant as our didactic format. Our work will draw on the examples and discussion of the previous dialogue. Thus, we encourage the reader to review the first dialogue of Stang and Poole [1].

*Researcher* I wanted to follow-up on our previous meeting. I'm uncertain that any of the modes of inference that we discussed are applicable to my research. I remember thinking to myself that no matter what mode of

Published online: 25 June 2015

inference I choose, it should not be a NHST. Therefore, I'm left with Fisher ST and Neyman–Pearson NHT. Neither of these is very attractive to me. I recall you mentioning Bayesian inference as a viable alternative. I'd like to learn more about that possibility.

*Consultant* As I recall, you're interested in the difference in five-year mortality risks among newly diagnosed skin melanoma patients between those with immunohistochemical factor A and those without that factor. I also remember that you consider a protective effect of factor A highly unlikely. You noted a body of existing research that led you to this view.

*Researcher* That's right. The evidence I have seen leads me to believe that a protective effect is much less plausible than a harmful one. I don't see any way to give this knowledge any credit in my research outside of a paragraph in the introduction or discussion of a manuscript. Can Bayesian inference allow me to integrate that information in a *quantitative* instead of a *qualitative* way?

*Consultant* Before we discuss Bayesian inference, why do you consider a protective effect so unlikely?

*Researcher* The existing literature clearly supports a harmful effect rather than a protective effect. In fact, I have never seen any evidence of a protective effect, though I suppose it is not impossible. The epidemiologic evidence varies in terms of study designs, periods of observation, and other features, but it all indicates an association in the direction of harm. Also, there is evidence of harm from toxicology and mechanistic research. How can Bayesian inference help me make use of this information in my study?

*Consultant* Bayesian inference is learning process, oriented more towards estimation than testing. While you can still do testing with Bayesian methods [4] the beauty of Bayesian inference is that you don't need to posit hypotheses before a study and see if your data and model reject or fail to reject them, as you would with NHT. Nor are you required to examine $p$ values as inverse measures of evidence against hypotheses, as in ST. Instead, you'll make probability statements that quantitatively summarize the credibility of possible values of the 5-year mortality risk difference you are estimating and then update those statements with the results of your new study. In this way, as more and more information becomes available, we can quantitatively combine it with previous information and, thus, learn more about the risk difference we are interested in researching.

*Researcher* That sounds intuitively appealing to me, much more so than testing the null hypothesis with a type I error rate or interpreting $p$ values, all of which seem distantly connected, at best, to my research goal of estimating that RD. How will I do these Bayesian calculations?

*Consultant* In three steps. First, you'll quantify your prior probability distribution, which is the quantitative summary of credible values of the RD that I mentioned. As the RD can be any of an infinite number of values between $-1$ and 1, the formal term is *probability density function*, but we can safely consider the distinction between probability and probability density a technicality.

Next, you'll conduct your study and obtain a risk difference estimate from that. This second step is where you will learn the degree to which your study data and model support each possible value of the RD, given the assumptions built into your statistical model. Think of this step as similar, if not identical, to any statistical analysis you've conducted *before* conducting a Bayesian analysis: your work would have started and stopped with step two.

Finally, you'll integrate the two using Bayes' theorem, [2, 5] which was developed for this purpose: to combine a prior credibility distribution with a support function to produce an updated, or posterior, credibility distribution.

*Researcher* You make it sound easy, almost *too* easy. Could you explain how I quantify the first two components to get to the third?

*Consultant* Why don't we simply dive in? If you have a calculator, or smartphone, we can conduct a Bayesian analysis right here and now.

*Researcher* Well then let's get started!

*Consultant* Very well. Suppose you have formulated a prior represented by the curve on this graph (Fig. 1). It's a normal distribution with a mean RD of 0.20, a variance of 0.01 and a standard error of 0.1. According to this distribution, you consider RD = 0.20 the most credible value, based on all the relevant scientific information of which you're aware. It also says that you're 95 % sure that the true RD is between 0.00 and 0.40. This means you're 2.5 % certain that the true RD is negative and 2.5 % certain that it is $> 0.40$.

*Researcher* This matches my current views quite well: I don't think it is impossible that factor A is protective, but I do believe a reduced 5-year mortality risk is highly unlikely. Other informed experts might have a different prior, but I believe it would be hard to defend a very different prior from this one. What next?

*Consultant* Suppose that you conduct a pilot study of 100 patients with newly diagnosed skin melanoma. You observed 10 deaths among 50 patients with factor A and 5 deaths among 50 patients without factor A in the first 5 years. Conveniently,

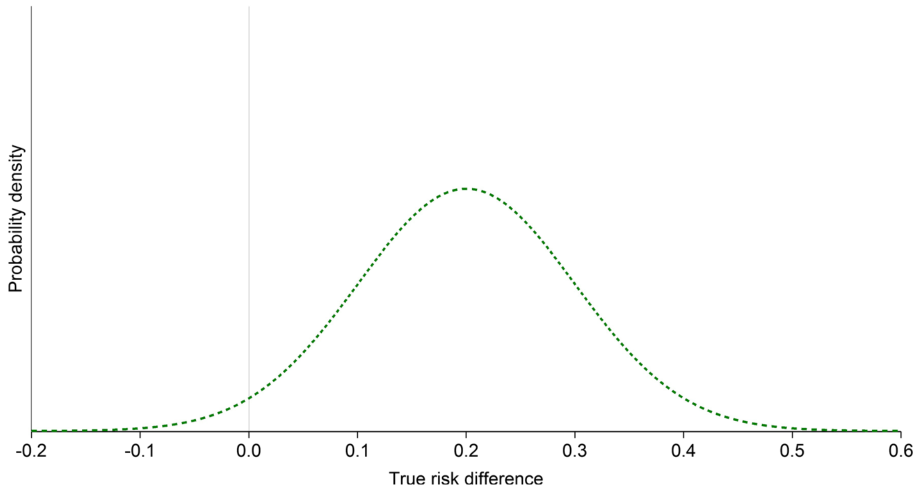

**Fig. 1** The probability density function summarizing the researcher's prior knowledge regarding the risk difference

the likelihood function is normal, as shown in this graph (Fig. 2). You'll notice right away that it follows a similar shape to your quantitative prior regarding the RD. The mean of the likelihood function is an RD of $10/50 - 5/50 = 0.20 - 0.10 = 0.10$. The estimated variance is 0.0051 [6].

*Researcher* I see that the curve for likelihood function reaches its highest point at the point estimate of RD = 0.10.

*Consultant* Yes, it does. That's why it's called the *maximum likelihood estimate.*

*Researcher* So, 0.10 is the value that's most likely to be the true value?

*Consultant* No, 0.10 is just the value to which this pilot study gives the most support. We haven't calculated the value you should consider most likely to be the true value, now that this study has been added to the relevant scientific information.

*Researcher* Ah, I see. I'll find that value on the curve that represents my posterior credibility distribution. How do I combine these two curves to get that one?

*Consultant* In this case, we simply combine the two estimates with inverse-variance weighting. We can do this *only* because both prior and likelihood are normal and, when combined, form a normal distribution [7]. The weights for the means of the prior and the likelihood function are $1/0.01 = 100$ and $1/0.0051 = 196$, respectively. Hence, the mean of your posterior credibility distribution is $\text{RD} = [100(0.20) + 196(0.10)]/(100 + 196) = 0.13$. The posterior variance is simply the reciprocal of the sum of the weights: $1/(100 + 196) = 0.0034$. Your posterior distribution is shown on this graph (Fig. 3). It says you're now 99 % sure the effect is harmful, up from approximately 98 %. You're 50 % certain the true RD is between 0.09 and 0.17, whereas you were 50 % certain it was between 0.13 and 0.27 before the study. Your prior 95 % credibility interval

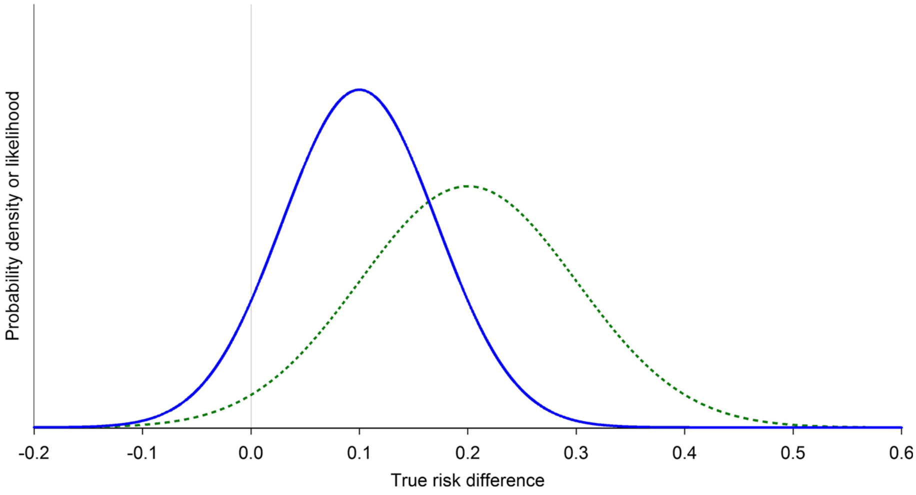

**Fig. 2** The likelihood function of the risk difference estimated from the pilot study data and model (*solid blue line*), along with the researcher's prior probability density function (*dashed green line*). (Color figure online)

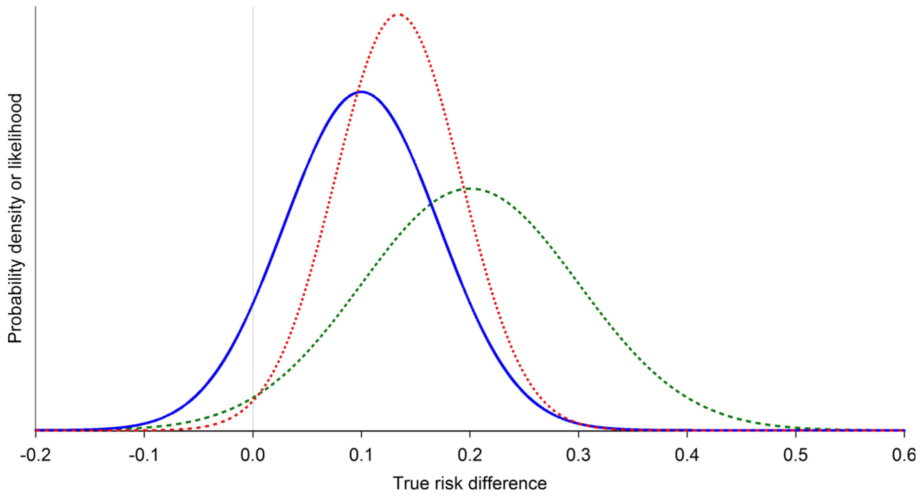

**Fig. 3** The posterior distribution of the risk difference estimate (*dotted red line*), which is the product of the researcher's prior probability (*dashed green line*) and the support function of the risk difference (*solid blue line*). (Color figure online)

for the true RD was 0.00–0.40 before the study. Now it's 0.02–0.25.

*Researcher* This is exactly what I am trying to accomplish with my current work: take the data and model from my pilot study and combine it with the existing evidence, because there is a wealth of information regarding factor A and mortality risk. It's nice that rather than expounding about how my study compares with other research, I can actually provide a quantitative integration of the two. How can I be sure that other researchers would have come to the same conclusions as I regarding the prior and the posterior?

*Consultant* You can never be sure of this. In fact, were you to give the same pool of information to multiple researchers to formulate a numerical prior, I would be surprised if they came up with *exactly* the same results. Equally, I'm confident many, well-informed researchers would come up with pretty similar priors.

*Researcher* What if someone comes up with a different prior than myself and, subsequently, a different posterior credible interval? What if I missed an important piece of evidence that someone else catches?

*Consultant* Don't worry too much. The important thing is to make what you did explicit and, thereby, expose it to widespread scrutiny by other experts. If it makes you feel better, it has been demonstrated that a *reasonable* prior will often return a result that is an improvement on what you would obtain otherwise [8]. In contrast, a default, flat prior assumes that a risk difference of 1 in 10 is just as plausible as a risk difference of 1 in 1,000,000 [2]. This prior is implicit in traditional, frequentist analysis, since it's analogous to saying before a study that you have no quantitative belief about the RD you're estimating. Of course, this is ridiculous, because knowing the prevalence of disease in the general population *alone* will give us some guidance about a plausible range of values for the risk difference. An informative (i.e., not flat) prior enables a researcher to assign higher prior probabilities to RD values in ranges known to be more realistic based on the general body of scientific information [9].

*Researcher* That's reassuring. One thing I noticed as we were going through this example is that the 95 % posterior credibility interval, while not equal to the 95 % confidence interval *sounds* a whole lot like it. I assume there's something to distinguish the two.

*Consultant* Aside from the fact that the interpretation of a confidence interval is *not,* as many believe, a probability statement, the major distinction is that a frequentist-based confidence interval does not permit you to incorporate the relevant, existing scientific information into the data analysis. It relies solely on your pilot study.

*Researcher* So, in the case that I had a completely flat prior, are the credible and confidence intervals not essentially the same?

*Consultant* Essentially, yes. But think about the implications of that. If you had *absolutely no prior information to contribute to your study, and were willing to assume any value of the RD is plausible*, then you would effectively have a flat prior with an infinite variance. In this case, the likelihood function and the posterior credible interval would be approximately equal.

*Researcher* Good point. I can't imagine this ever being the case. However, wouldn't such an approach, not imposing my prior beliefs, allow me to remain objective? This is a major concern of mine with regard to any research I do: the perception that I'm inappropriately imposing my beliefs on a research study.

*Consultant* Subjectivity in the appraisal of the existing evidence, before and after the study at hand, is going to be expressed anyway. It can either be expressed qualitatively, with adjectives and adverbs, in the Introduction and Discussion sections of the paper, where anything goes, or it can be expressed quantitatively in the Methods and Results sections, where transparency and coherence are the orders of the day.

Regardless of the decision to use a Bayesian approach, one would be hard pressed to think of any data analysis as objective. For example, specified type I and desirable type II error rates are nearly always 5 and 20 %, respectively; however, their frequency in research doesn't make them objective or defensible. Take variable selection as another example. No doubt you will be adjusting your analyses for important confounders. We often start with a set of variables with which we think we can best adjust confounding, such as age. This process is based on background knowledge. Again, different researchers will reach different conclusions about how to appropriately control confounding, [10, 11] or even how to model their exposure-disease relationship. Thus, this process is inherently subjective, from selecting the initial set of variables to consider to the decision of how to conduct regression modelling.

This is not to say that a Bayesian approach is immune to abuse and inappropriate use of priors. The most important thing is that your subjective prior is, in fact, based on existing evidence [4]. Further, one should err on the side of ascribing less confidence in their prior by assigning it a larger variance [12] and should very clearly describe how the prior is formulated; doing so will ensure that other researchers can compare their priors to yours. Finally, always report your prior probability distribution and study results along with your posterior distribution in the results of all of your work to allow full transparency and scrutiny by fellow researchers.

## Discussion

We hope that this second dialogue has shown that Bayesian inference is a more intuitive mode of statistical inference compared to frequentist inference in the form of ST, NHT, and NHST. We have provided a simple example to illustrate that a Bayesian approach can be as simple as a few hand calculations, though this is not always the case.

As computational methods have advanced, researchers have been able to explore and more easily implement Bayesian procedures. The most obvious benefit of a Bayesian procedure is that it allows researchers to integrate existing scientific information of relevance to parameters of interest directly into data analysis. Bayesian hierarchical modelling is a commonly used method where parameters are modelled in a way that assumes, a priori, that they are exchangeable, or share a common mean and variance. However, many more advanced procedures have been implemented, such as adjustment for unmeasured confounders, information bias, and selection bias [13–15]; imposing order constraints on parameters based on evidence from toxicology [16]; and restricting risk ratios estimates based on the knowledge that the predicted probability of the outcome cannot exceed 1.0 [17]. While difficult, these procedures are likely within the skill sets of many biostatisticians, who may aid practicing researchers in implementation.

There are many published examples of implementing Bayesian methods without advanced statistical software [2, 15]. Maclehose and Hamra provide a worked example using inverse variance weighting [5]. Cole et al. [18] illustrate a Bayesian approach without the need for Markov Chain Monte Carlo methods. We provide, as an eSupplement, an example of implementing a Bayesian analysis using SAS statistical software using a 2 × 2 table example. Sullivan and Greenland also provide a worked example of implementing a Bayesian procedure in SAS statistical software [19, 20].

Finally, we would like to further emphasize the importance of formulating and explicitly describing the derivation of defensible priors. It is helpful for other researchers with substantive experience to understand that the specified prior has a strong foundation based on existing evidence.

**Conflict of interest** None.

## References

1. Stang A, Poole C. The researcher and the consultant: a dialogue on null hypothesis significance testing. Eur J Epidemiol. 2013;28(12):939–44.
2. Greenland S. Bayesian perspectives for epidemiological research: I. Foundations and basic methods. Int J Epidemiol. 2006;35(3):765–75.
3. Goodman SN. Introduction to Bayesian methods I: measuring the strength of evidence. Clin Trials. 2005;2(4):282–290; discussion 301–284, 364–278.
4. Goodman SN. Toward evidence-based medical statistics. 2: the Bayes factor. Ann Intern Med. 1999;130(12):1005–13.
5. MacLehose R, Hamra G. Applications of Bayesian methods to epidemiologic research. Curr Epidemiol Rep. 2014;1(3):103–9.
6. Rothman KJ, Greenland S, Lash TL. Modern epidemiology. 3rd ed. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008.
7. MacLehose RF, Hamra GB. Applications of Bayesian methods to epidemiologic research. Curr Epidemiol Rep. 2014;1:103–9.
8. Greenland S. Principles of multilevel modelling. Int J Epidemiol. 2000;29(1):158–67.
9. Hamra GB, MacLehose RF, Cole SR. Sensitivity analyses for sparse-data problems—using weakly informative Bayesian priors. Epidemiology. 2013;24(2):233–9.
10. Greenland S. Modeling and variable selection in epidemiologic analysis. Am J Public Health. 1989;79(3):340–9.

11. Greenland S. Invited commentary: variable selection versus shrinkage in the control of multiple confounders. Am J Epidemiol. 2008;167(5):523–9.
12. Greenland S. Bayesian perspectives for epidemiological research II. Regression analysis. Int J Epidemiol. 2007;36(1):195–202.
13. Lash TL, Fox MP, Fink AK. Applying quantitative bias analysis to epidemiologic data. Dordrecht; New York: Springer; 2009.
14. Steenland K, Greenland S. Monte Carlo sensitivity analysis and Bayesian analysis of smoking as an unmeasured confounder in a study of silica and lung cancer. Am J Epidemiol. 2004;160(4):384–92.
15. Greenland S. Bayesian perspectives for epidemiologic research. III. Bias analysis via missing-data methods (vol 38, p. 1662, 2009). Int J Epidemiol. 2010;39(4):1116.
16. Hamra G, Richardson D, Maclehose R, Wing S. Integrating informative priors from experimental research with Bayesian methods: an example from radiation epidemiology. Epidemiology. 2013;24(1):90–5.
17. Chu H, Cole SR. Estimation of risk ratios in cohort studies with common outcomes: a Bayesian approach. Epidemiology. 2010;21(6):855–62.
18. Cole SR, Chu HT, Greenland S, Hamra G, Richardson DB. Bayesian posterior distributions without Markov chains. Am J Epidemiol. 2012;175(5):368–75.
19. Sullivan SG, Greenland S. Bayesian regression in SAS software. Int J Epidemiol. 2013;42(1):308–17.
20. Sullivan SG, Greenland S. Bayesian regression in SAS software. Int J Epidemiol. 2014;43(3):974.