The Waves and the Sigmas
(To Say Nothing of the 750 GeV Mirage)[*]

G. D'Agostini

Università "La Sapienza" and INFN, Roma, Italia

(giulio.dagostini@roma1.infn.it, http://www.roma1.infn.it/~dagos)

Abstract

This paper shows how p-values do not only create, as well known, wrong expectations in the case of flukes, but they might also dramatically diminish the 'significance' of *most likely* genuine signals. As real life examples, the 2015 first detections of gravitational waves are discussed. The March 2016 statement of the American Statistical Association, warning scientists about interpretation and misuse of p-values, is also reminded and commented. (The paper is complemented with some remarks on past, recent and future claims of discoveries *based on sigmas* from Particles Physics.)

1 Introduction

On February 11 the LIGO-Virgo collaboration announced the detection of Gravitational Waves (GW). They were emitted about one billion years ago by a Binary Black Hole (BBH) merger and reached Earth on September 14, 2015. The claim, as it appears in the 'discovery paper' [1] and stressed in press releases and seminars, was based on "$> 5.1\,\sigma$ significance." Ironically, shortly after, on March 7 the American Statistical Association (ASA) came out (independently) with a strong statement warning scientists about interpretation and misuse of p-values [2]. As promptly reported by Nature [3], "this is the first time that the 177-year-old ASA has made explicit recommendations on such a foundational matter in statistics, says executive director Ron Wasserstein. The society's members had become increasingly concerned that the P value was being misapplied in ways that cast doubt on statistics generally, he adds."

In June we have finally learned [4] that another 'one *and a half*' gravitational waves from Binary Black Hole mergers were also observed in 2015, where by the 'half' I refer to the October 12 event, highly *believed* by the collaboration to be a gravitational wave, although

---

[*]Note based on the invited talk *Claims of discoveries based on sigmas* at MaxEnt 2016 (Ghent, Belgium, 15 July 2016) and on seminars and courses to PhD students in the first half of 2016.

having only $1.7\,\sigma$ *significance* and therefore classified just as LVT (LIGO-Virgo Trigger) instead of GW. However, another figure of merit has been provided by the collaboration for each event, a number based on probability theory and that tells how much we <u>must</u> modify the relative beliefs of two alternative hypotheses in the light of the experimental information. This number, at my knowledge never even mentioned in press releases or seminars to large audiences, is the Bayes factor (BF), whose meaning is easily explained: if you considered *à priori* two alternative hypotheses equally likely, a BF of 100 changes your odds to 100 to 1; if instead you considered one hypothesis rather unlikely, let us say your odds were 1 to 100, a BF of 10^4 turns them the other way around, that is 100 to 1. You will be amazed to learn that even the "1.7 sigma" LVT151012 has a BF of the order of $\approx 10^{10}$, considered a very strong evidence in favor of the hypothesis "Binary Black Hole merger" against the alternative hypothesis "Noise". (Alan Turing would have called the evidence provided by such an huge 'Bayes factor,' or what I. J. Good would have preferred to call "Bayes-Turing factor" [5],[1] 100 *deciban*, well above the 17 deciban threshold considered by the team at Bletchley Park during World War II to be reasonably confident of having cracked the daily Enigma key [7].)

In the past I have been writing quite a bit on how 'statistical' considerations based on p-values tend to create wrong expectations in frontier physics (see e.g. [8] and [9]). The main purpose of this paper is the opposite, i.e. to show how p-values might relegate to the role of a possible fluke what is most likely a genuine finding. In particular, the solution of the apparent paradox of how a marginal '1.7 sigma effect' could have a huge BF such as 10^{10} (and virtually even much more!) is explained in a didactic way.

---

[1]Note that Eq. (1) in [5] clearly contains a typo, or it has got a problem in the scanning of the document, since $P(E\,|\,H)/P(E\,|\,H)$ makes no sense in that equation and it should have been $P(E\,|\,H)/P(E\,|\,\overline{H})$, where H and $\overline{H}$ stand for 'complementary' (formally "exhaustive, mutually exclusive") hypotheses. The equation should then read

$$\frac{O(H\,|\,E)}{O(H)} \;\; = \;\; \frac{P(E\,|\,H)}{P(E\,|\,\overline{H})},$$

where $O(H)$ and $O(H\,|\,E)$ are *prior* and *posterior* odds, i.e., respectively, $O(H\,|\,E) = P(H\,|\,E)/P(\overline{H}\,|\,E)$ and $O(H) = P(H)/P(\overline{H})$. Eq. (1) of [5] would then result into

$$\frac{P(H\,|\,E)/P(\overline{H}\,|\,E)}{P(H)/P(\overline{H})} \;\; = \;\; \frac{P(E\,|\,H)}{P(E\,|\,\overline{H})}$$

or

$$\frac{P(H\,|\,E)}{P(\overline{H}\,|\,E)} \;\; = \;\; \frac{P(E\,|\,H)}{P(E\,|\,\overline{H})} \cdot \frac{P(H)}{P(\overline{H})},$$

in words

$$\textbf{posterior odds} \;\; = \;\; \textbf{Bayes factor} \times \textbf{prior odds}$$

(For log representation of odds and Bayes factors see section 2 and appendix E of [6] and references therein, although at that time Turing's contributions, as well as 'bans' and 'decibans', were unknown to the author, who arrived at the same conclusion of Turing's 1 deciban as rough estimate of *human resolution* to *judgement leaning* and *weight of evidence* – table 1 in page 13 and text just below it.)

2 Preamble

Since this paper can be seen as the sequel of Refs. [12] and [9], with the basic considerations already expounded in [8], for the convenience of the reader I shortly summarize the main points maintained there.

- The "essential problem of the experimental method" is nothing but solving "a problem in the probability of causes", i.e. ranking in credibility the hypotheses that are considered to be possibly responsible of the observations, (quotes by Poincaré [13]). [2] [There is indeed no conceptual difference between "comparing hypotheses" or "inferring the value" of a physical quantity, the two problems only differing in the numerosity of hypotheses, *virtually* infinite in the latter case, when the physical quantity is *assumed*, for mathematical convenience,[3] to assume values with continuity.]

- The deep source of uncertainty in inference is due to the fact that (apparently) identical *causes* might produce different effects, due to *internal* (intrinsic) probabilistic aspects of the theory, as well as to *external* factors (think at measurement errors).

- Humankind is used to live – and survive – in conditions of uncertainty and therefore the human mind has developed a mental 'category' to handle it: *probability*, meant as degree of belief. This is also valid when we 'make science', since "it is scientific only to say what is more likely and what is less likely" (Feynman [15]).

- *Falsificationism* can be recognized as an attempt to extend the classical *proof by contradiction* of classical logic to the experimental method, but it *simply fails* when stochastic (either internal or external) effects might occur.

- The further extension of falsificationism from *impossible* effects to *improbable* effects is simply deleterious.

- The invention of p-values can be seen as an attempt to overcome the evident problem occurring in the case of a large number of effects (*virtually infinite* when we make measurements): any observation has a very small probability in the light of whatever hypothesis is considered, and then it 'falsifies' it.

- Logically the previous extension ("observed effect" $\rightarrow$ "all possible effects equally or less probable than the observed one") does not hold water. (But it seems that for many practitioners logic is optional – the reason why "p-values *often work*" [8] will be discussed in section 6.)

---

[2]Instead, "making statistics", i.e. to describe and summarize data, has never been the *primary* interest of physicists as well as of many other scientists, although it is certainly useful for a variety of reasons.

[3]"*No mathematical squabbles*" was John Skilling's mantra in his recent tutorial at MaxEnt 2016, in which he was stressing the importance to restart thinking, at least "initially", in terms of "finite target", "finite partitioning" and integers [14].

- In practice p-values are routinely misinterpreted by most practitioners and scientists, and incorrect interpretations of the data are spread around over the media[4] (for recent examples, related to the LHC presumptive 750 GeV di-photon signal (see e.g. [16, 17, 18, 19, 20] and footnote 31 for later comments.).

- The reason of the misunderstandings is that p-values (as well as other outcomes from other methods of the dominating 'standard statistics', including *confidence intervals* [8]), do not reply to the very question human minds *by nature* ask for, i.e. which hypothesis is more or less believable (or how likely the 'true' value of a quantity lies within a given interval). For this reason I am afraid p-values (or perhaps a new invention by statisticians) will still be misinterpreted and misused despite the 2016 ASA statement, as I will argue at the end of section 3.2).

- Given the importance of the previous point, for the convenience of the reader I report here verbatim the list of misunderstandings appearing in the Wikipedia at the **end of 2011** [9],[5] highlighting the sentences that mostly concern our discourse.

  1. "**The p-value is not the probability that the null hypothesis is true.** In fact, frequentist statistics does not, and cannot, attach probabilities to hypotheses. Comparison of Bayesian and classical approaches shows that a p-value can be very close to zero while the posterior probability of the null is very close to unity (if there is no alternative hypothesis with a large enough a priori probability and which would explain the results more easily). This is the Jeffreys-Lindley paradox.

  2. **The p-value is not the probability that a finding is "merely a fluke."** As the calculation of a p-value is based on the assumption that a finding is the product of chance alone, it patently cannot also be used to gauge the probability of that assumption being true. This is different from the real meaning which is that the p-value is the chance of obtaining such results if the null hypothesis is true.

  3. The p-value is not the probability of falsely rejecting the null hypothesis. This error is a version of the so-called prosecutor's fallacy.

  4. The p-value is not the probability that a replicating experiment would not yield the same conclusion.

---

[4]Sometimes scientists say they reported "the right thing" (i.e. just the p-value), but it was journalist's fault to misinterpret them. But, as I have documented in my writings, often are the official statement of laboratories, of collaboration spokespersons, or of prominent physicists to confuse p-values with probabilities of hypotheses, as you can e.g. find in [9] and, more extensively, in `http://www.roma1.infn.it/~dagos/badmath/index.html#added`. A suggestion to laymen is that, "instead of heeding impressive-sounding statistics, we should ask what scientists actually believe" [21].

[5]As it is well known, the content of Wikipedia is variable with time. The reason I report here the list of misunderstandings as it appeared some years ago, and as it has been more ore less until the beginning of 2016 – I have no documented records, but I have been checking it from time to time, in occasion of seminars and courses and I had not realized major changes, like the reductions of the items from 7 to 5 – is that the present version has been clearly being influenced by the ASA statement of March 2016. (I report here all seven items, although I have to admit that I get lost after the third one – but you for you seven are still not enough see [22])

5. $(1 - \text{p-value})$ is not the probability of the alternative hypothesis being true.

6. The significance level of the test is not determined by the p-value. The significance level of a test is a value that should be decided upon by the agent interpreting the data before the data are viewed, and is compared against the p-value or any other statistic calculated after the test has been performed. (However, reporting a p-value is more useful than simply saying that the results were or were not significant at a given level, and allows the reader to decide for himself whether to consider the results significant.)

7. The p-value does not indicate the size or importance of the observed effect (compare with effect size). The two do vary together however – the larger the effect, the smaller sample size will be required to get a significant p-value."

- If we want to form our minds about which hypothesis is more or less probable in the light of all available information, then we need to base our reasoning on *probability theory*, understood as the *mathematics of beliefs*, that is essentially going back to the ideas of Laplace. In particular the updating rule, presently known as the *Bayes rule* (or Bayes theorem), should be probably better called *Laplace rule*, or at least Bayes-Laplace rule.

- The 'rule', expressed in terms of the alternative *causes* (C_i) which could possibly produce the *effect* (E), as originally done by Laplace,[6] is

$$P(C_i \,|\, E, I) \;\; = \;\; \frac{P(E \,|\, C_i, I) \cdot P(C_i \,|\, I)}{\sum_k P(E \,|\, C_k, I) \cdot P(C_k \,|\, I)} \,. \tag{1}$$

or, considering also $P(C_j \,|\, E, I)$ and taking the ratio of the two *posterior probabilities*,

$$\frac{P(C_i \,|\, E, I)}{P(C_j \,|\, E, I)} \;\; = \;\; \frac{P(E \,|\, C_i, I)}{P(E \,|\, C_j, I)} \times \frac{P(C_i \,|\, I)}{P(C_j \,|\, I)} \,, \tag{2}$$

where I stands for the *background information*, sometimes implicitly assumed.

- Important consequences of this rule – I like to call them Laplace's teachings [9], because they stem from his "*fundamental principle* of that branch of the analysis of chance that consists of reasoning a posteriori from events to causes" [23] – are:

  - It makes no sense to speak about how the probability of C_i changes if:
    1. there is no alternative cause C_j;
    2. the way how C_j might produce E is not properly modelled, i.e. if $P(E \,|\, C_j, I)$ has not been *somehow* assessed.[7]

---

[6]This is "Principle VI", expounded in simple words in [23], in which he calls 'principles' the principal rules resulting from his theory. Note also that Eq. (1) requires that hypotheses C_i form a 'complete class' (exhaustive and mutually exclusive), while Eq. (2) is more general, although it might require some care in its application, as pointed out in [24] [think e.g. at the hypotheses $H_1 = C_1 \cap C_2$ and $H_2 = C_2$, implying: i) $P(H_1) \leq P(H_2) \, \forall \, E$; ii) the calculation of $P(C_1 \,|\, E)$ and $P(C_2 \,|\, E)$ requires extra information].

[7]It does not matter if the assessment is done analytically, numerically, by simulation, or just by pure subjective considerations – what is important to understand is that without the slightest guess on what $P(E \,|\, C_j, I)$ could be, and on how much C_j is more or less believable, you cannot modify your 'confidence' on C_i, as it will be further reminded in section 6.

– The updating of the probability ratio depends only on the so called *Bayes factor*

$$\frac{P(E \mid C_i, I)}{P(E \mid C_j, I)}, \tag{3}$$

ratio of the probabilities of E given either hypotheses,[8] and *not on the probability of other events that have not been observed and that are even less probable than E* (upon which p-values are instead calculated).

– One should be careful not to confuse $P(C_i \mid E)$ with $P(E \mid C_i)$, and in general $P(A \mid B)$, with $P(B \mid A)$. Or, moving to continuous variables, $f(\mu \mid x)$ with $f(x \mid \mu)$, where: '$f()$' stands here, depending on the contest, for a *probability function* or for a *probability density function* (pdf): x and μ are symbols for observed quantity and 'true' value, respectively, the latter being in fact just the *parameter of the model we use to describe the physical world*.

– Cause C_i is *falsified* by the observation of the event E *only if C_i <u>cannot</u> produce it*, and not because of the smallness of $P(E \mid C_i, I)$.

– Extending the reasoning to continuous observables (generically called X) characterized by a pdf $f(x \mid H_i)$, the probability to observe a value in the *small* interval Δx is $f(x \mid H_i)\,\Delta x$. What matters, for the comparison of two hypotheses in the light of the observation $X = x_m$, is therefore the ratio of pdf's $f(x_m \mid H_i)/f(x_m \mid H_j)$, and not the smallness of $f(x_m \mid H_i)\,\Delta x$, which tends to zero as $\Delta x \to 0$. Therefore, *an hypothesis is*, strictly speaking, *falsified*, in the light of the observed $X = x_m$, *only if $f(x_m \mid H_i) = 0$*.

- Finally, I would like to stress that *falsificability is not a strict requirement for a theory to be accepted as 'scientific'*. In fact, in my opinion a weaker condition is sufficient, which I called *testability* in [12]: given a theory Th_i and possible observational data $\mathcal{D}$, it should be possible to model $P(\mathcal{D} \mid Th_i)$ in order to compare it with an alternative theory Th_j characterized by $P(\mathcal{D} \mid Th_j) \neq P(\mathcal{D} \mid Th_i)$.[9] This will allow to rank theories in probability in the light of empirical data and of any other criteria, like simplicity or aesthetics[10] without the requirement of falsification, that cannot be achieved, logically speaking, in most cases.[11]

---

[8]Eq. 3 is also known as "*likelihood* ratio", but I avoid and discourage the use of the '*l*-word', being a major source of misunderstanding among practitioners [8, 25], who regularly use the '*l*-function' as pdf of the unknown quantity, taking then (also in virtue of an unneeded 'principle') its argmax as *most believable value*, sticking to it in further 'propagations' [25]. (A recent, important example comes from two reports of the same organization, each using the '*l*-word' with two different meanings [26, 27].)

[9]For example String Theory (ST) supporters should tell us in what $P(\mathcal{D} \mid ST)$ differs from $P(\mathcal{D} \mid SM)$ from Standard Model, with $\mathcal{D}$ being past, present or future *observational data*.

[10]But we have to be careful with judgments based on aesthetics, which are unavoidably anthropic (and debates on aesthetics will never end, while ancient Romans wisely used to say that "de gustibus non disputandum est" and, as someone warned, "if you are out to describe the truth, leave elegance to the tailor." [29]. This is more or less what is going on in Particle Physics in the past years, after that nothing new has been found at LHC besides the highly expected observation of the Higgs boson in the final state, with many serious theorists humbling admitting that "Nature does not seem to share our ideas of *naturalness*."

[11]Think for example at all infinite numbers of Gaussian models $\mathcal{N}(\mu, \sigma)$ that might have produced the

3 ASA statement on statistical significance and p-values

3.1 *Ante factum*

The statement of the American Statistical Association on March this year did not arrive completely unexpected. Many scientists were in fact aware and worried of the "science's dirtiest secret", i.e. that "the 'scientific method' of testing hypotheses by statistical analysis stands on a flimsy foundation"[30]. Indeed, as Allen Caldwell of MPI Munich eloquently puts it (e.g. in [31]) "The real problem is not that people have difficulties in understanding Bayesian reasoning. The problem is that they do not understand the frequentist approach and what can be concluded from a frequentist analysis. What is not understood, or forgotten, is that the frequentist analysis relates only to possible data outcomes within a model context, and not probabilities of a model being correct. This misunderstanding leads to faulty conclusions."

Faulty conclusions based on p-values are countless in all fields of research, and frankly I am personally much more worried when they might affect our health[12] and security, or the future of our planet, rather then when they spread around unjustified claims of revolutionary discoveries or of possible failures of the so called Standard Model of Particle Physics [9].[13] For instance, "A lot of what is published is incorrect" reported last year *The Lancet*'s Editor-in-Chief Richard Horton [36]. This could be because, *looking around more or less 'at random'*, statistical 'significant results' will soon or later show up (as that of the last frame of an *xkcd* cartoon shown in Fig. 1 – see [37] for the full story); or because dishonest (or driven by wishful thinking, which in Science is more or less the same) researchers might do some *p-hacking* (see e.g. [38] and [39]) in order to make 'significant effects' appear – remember that "if you torture the data long enough, it will confess to anything" [40].

A special mention deserves the February 2014 editorial of David Trafimow, Director of *Basic and Applied Social Psychology* (BASP), in which he takes a strong position against "null hypothesis significance testing procedure (NHSTP)" because it "has been shown to be logically invalid and to provide little information about the actual likelihood of either the null or experimental hypothesis" [41]. In fact a large echo (see e.g. [42], [43] and [44]) had last year a second editorial, signed together with his Associate Director Michael Marks

---

observation $x_m = 4$. Since, strictly speaking, any Gaussian might produce any real value, it follows none of the ∞^2 models can be falsified. Nevertheless, every one will agree that $x_m = 4$ it is *more likely* to be attributed to model $\mathcal{N}(3, 1)$ than $\mathcal{N}(20, 1)$. But you cannot say that the observation $x_m = 4$ falsifies model $\mathcal{N}(20, 1)$!

[12]See e.g. [32, 33, 34, 35] (for instance Elisabeth Iorns' *comment* on New Scientist [33] reports that "more than half of biomedical findings cannot be reproduced" and "pharmaceutical company Bayer says it fails to replicate two-thirds of published drug studies" – !!!).

[13]Frankly I do not think that these claims hurt fundamental physics, which I consider quite healthy and (mostly) done by honest researchers. In fact, false alarms might even have positive effects inside the community, because they stimulate discussions on completely new possibilities and encourage new researches to be undertaken, as also recognized in the bottom line of de Rujula's cartoon of Fig. 2. My worries mainly concern negative reputation the field risks to gain and, perhaps even more, bad education provided to young people, most of which will leave pure research and will try to apply elsewhere the analysis methods they learned in searching for new particles and new phenomena.

Figure 1: A 'significant' result obtained *provando e riprovando* [37].

published on February 15, 2015, in which they announce that, after "a grace period allowed to authors", "from now on, BASP is banning the NHSTP" [45].

3.2 *Principia*

Moving finally to the content of the ASA statement, after a short introduction, in which it is recognized that "the p-value [. . .] is commonly misused and misinterpreted," and a reminder of what a p-value "informally"' is ("the probability under a specified statistical model that a statistical summary of the data [. . .]would be equal to or more extreme than its observed value") a list of six items, indicated as "principles", follows (the highlighting is original).

1. **P-values can indicate how incompatible the data are with a specified statistical model.**

   A p-value provides one approach to summarizing the incompatibility between a particular set of data and a proposed model for the data. The most common context is a model, constructed under a set of assumptions, together with a so-called "null hypothesis." Often the null hypothesis postulates the absence of an effect, such as no difference between two groups, or the absence of a relationship between a factor and an outcome. The smaller the p-value, the greater the statistical incompatibility of the data with the null hypothesis, if the underlying assumptions used to calculate the p-value hold. This incompatibility can be interpreted as casting doubt on or providing evidence against the null hypothesis or the underlying assumptions.

2. **P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance**

alone.

Researchers often wish to turn a p-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The p-value is neither. It is a statement about data in relation to a specified hypothetical explanation, and is not a statement about the explanation itself.

3. **Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.**

Practices that reduce data analysis or scientific inference to mechanical "bright-line" rules (such as "$p < 0.05$") for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision making. A conclusion does not immediately become "true" on one side of the divide and "false" on the other. Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis. Pragmatic considerations often require binary, "yes-no" decisions, but this does not mean that p-values alone can ensure that a decision is correct or incorrect. The widespread use of "statistical significance" (generally interpreted as "$p \leq 0.05$") as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.

4. **Proper inference requires full reporting and transparency**

P-values and related analyses should not be reported selectively. Conducting multiple analyses of the data and reporting only those with certain p-values (typically those passing a significance threshold) renders the reported p-values essentially uninterpretable. Cherry-picking promising findings, also known by such terms as data dredging, significance chasing, significance questing, selective inference, and "p-hacking," leads to a spurious excess of statistically significant results in the published literature and should be vigorously avoided. One need not formally carry out multiple statistical tests for this problem to arise: Whenever a researcher chooses what to present based on statistical results, valid interpretation of those results is severely compromised if the reader is not informed of the choice and its basis. Researchers should disclose the number of hypotheses explored during the study, all data collection decisions, all statistical analyses conducted, and all p-values computed. Valid scientific conclusions based on p-values and related statistics cannot be drawn without at least knowing how many and which analyses were conducted, and how those analyses (including p-values) were selected for reporting.

5. **A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.**

Statistical significance is not equivalent to scientific, human, or economic significance. Smaller p-values do not necessarily imply the presence of larger or more important effects, and larger p-values do not imply a lack of importance or even lack of effect. Any effect, no matter how tiny, can produce a small p-value if the sample size or measurement precision is high enough, and large effects may produce unimpressive p-values if the sample size is small or measurements are imprecise. Similarly, identical estimated effects will have different p-values if the precision of the estimates differs.

6. **By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.**

    Researchers should recognize that a *p*-value without context or other evidence provides limited information. For example, a *p*-value near 0.05 taken by itself offers only weak evidence against the null hypothesis. Likewise, a relatively large *p*-value does not imply evidence in favor of the null hypothesis; many other hypotheses may be equally or more consistent with the observed data. For these reasons, data analysis should not end with the calculation of a *p*-value when other approaches are appropriate and feasible.

These words sound as an admission of failure of much of the statistics teaching and practice in the past *many* decades. But yet I find their courageous statement still somehow unsatisfactory, and, in particular, the first principle is in my opinion still affected by the kind of 'original sin' at the basis of p-value misinterpretations and misuse. Many practitioners consider in fact a value occurring several (but often just a few) standard deviations from the 'expected value' (in the probabilistic sense) to be a 'deviance' from the model, which is clearly absurd: no value a model can yield can be considered an *exception* from the model itself (see also footnote 11 – the reason why "p-values *often work*" will be discussed in section 6). Then, moving to principle 2, it is not that "researchers *often wish* to turn a *p*-value into a statement about the truth of a null hypothesis" (italic mine), as if this would be an extravagant fantasy: reasoning in terms of degree of belief of whatever is uncertain is connatural to the 'human understanding' [46]: *all methods that do not tackle straight the fundamental issue of the probability of hypotheses, in the problems in which this is the crucial question, are destined to fail, and to perpetuate misunderstanding and misuse.*

4 The 'Monster' blessed by the 5 sigmas

Rumors that the LIGO interferometers had most likely detected a gravitation wave (GW) were circulating in autumn last year. Personally, the direct information I got quite late, at the beginning of December, was "we have seen a *Monster*", without further detail. Therefore, when a few days before February 11 quantitative rumors talked of 5.1 sigmas, I was disappointed and highly puzzled. How could a Monster have *only* just a bit more than five sigmas? Indeed in the past decades we have seen in Particle Physics several effects of similar statistical significance coming and going, as Alvaro de Rujula depicted already in 1985 in his famous *Cemetery of Physics* of Fig. 2 [48].[14] Therefore for many of us a five-sigma effect would have been something worth discussions or perhaps further investi-

---

[14]Finally he humorously summarized his very long experience in the '*de Rujula paradox*' [47]:

   If you *disbelieve* every result presented as having a 3 sigma,
   or 'equivalently' a 99.7% chance of being correct,
   you will turn out to be *right* 99.7% of the times.

('Equivalently' within quote marks is de Rujula's original, because he knows very well that there is no equivalence at all.)

Figure 2: Alvaro de Rujula's Cemetery of Physics [48], with graves indicating 'false alarms' in frontier physics, and not old physics ideas faded out with time, like epicycles, phlogiston or aether.

gations but certainly not a Monster.[15] This impression was very evident from the reaction many people had *after* seeing the wave form. "Came on, this is not a five-sigma effect", commented several colleagues, more or less using the same words, "these are *hundreds* of sigmas!", a colored expression to say that just by eye the hypothesis Noise was beyond any imagination.[16]

The reason of the 'monstrosity' of GW150914 was indeed in Table 1 of the accompanying paper on *Properties of the binary black hole merger GW150914* [28]: a Bayes factor "BBH merger" Vs "Noise"[17] of about 5×10^{125} (yes, *five times ten to one-hundred-twenty-*

---

[15]"And the July 2012 5-sigma Higgs boson?", you might argue. Come on! That <u>was</u> the Higgs boson, the highly expected missing tessera to give sense to the amazing mosaic of the Standard Model, whose mass had already been somehow inferred from other measurements, although with quite large uncertainty (see e.g. [49, 50]). For this reason the 2011 data were sufficient to many who had followed this physics since years (and not sticking to the 5-sigma dogma) to be highly confident that the Higgs boson was finally observed in a final state diagram [9]. Instead, some of those who were casting doubt on the possibility of observing the Higgs are the same who were giving credit to the December 2015 $\gamma\gamma$ 750 GeV excess at LHC (and some even to the Opera's superluminar neutrinos!). I hope they will learn from the double/triple lesson.

[16]And indeed we have also learned that the only serious alternative hypothesis taken into account and investigated in detail was that of a sabotage!

[17]To be precise, the competing hypotheses are "BBH-merger & Noise" Vs "only Noise".

five). This means that, no matter how small the odds in favor of a BBH merger were and even casting doubt on the evaluation of the Bayes factor,[18] the posterior odds would be extraordinary large, the probability of noise being smaller than Shakespeare's drop of water identically recovered from the sea.[19]

5   Cinderella and her sisters

The results of the full observing run of the Advanced LIGO detectors (September 12, 2015, to January 19, 2016) have been presented on June 8 [4], slightly updating some of the February's digits. Figure 3 summarizes detector performances and results, with some important numbers (within this context) reminded in the caption.

---

[18]At this point a 'technical' remark is in order, which is indeed also conceptual and sheds some light on the difficulty of the calculation and possible uncertainties on the resulting value. Given the hypotheses H_0 and H_1 and data $\mathcal{D}$, the Bayes factor H_1 Vs H_0 is

$$\frac{P(\mathcal{D} \,|\, H_1, I)}{P(\mathcal{D} \,|\, H_0, I)} \,,$$

where for sake of simplicity we identify H_1 with "BBH merger" and H_0 with "Noise". Now the question is that *there is not a single, precisely defined, hypothesis "BBH merger"*. And the same is true also for the 'null hypothesis' "Noise". This is because each hypothesis comes with free parameters. For example, in the case of "BBH merger", the conditional probability of $\mathcal{D}$ depends on the masses of the two black holes (m_1 and m_2), on their distance from Earth (d) and so on, i.e. $P(\mathcal{D} \,|\, H_1, m_1, m_2, d, \ldots, I)$. The same holds for the Noise, because there is no such a thing as "the Noise", but rather a noise model with many parameters obtained monitoring the detectors. So in general, for the generic hypothesis H we have

$$P(\mathcal{D} \,|\, H, \underline{\theta}, I) \,,$$

in which $\underline{\theta}$ stands for the set of parameters of the hypothesis H. But what matters for the calculation of the Bayes factor is $P(\mathcal{D} \,|\, H, I)$, and this can be evaluated from probability theory taking account all possible values of the set of parameters $\underline{\theta}$, weighting them by the pdf $f(\underline{\theta} \,|\, H, I)$, i.e. 'simply' as

$$P(\mathcal{D} \,|\, H, I) \;\; = \;\; \int_{\underline{\Theta}} P(\mathcal{D} \,|\, H, \underline{\theta}, I) \, f(\underline{\theta} \,|\, H, I) \, d\underline{\theta} \,. \qquad (F.1)$$

But the game can be not simple at all, because i) this integral can be very difficult to calculate; ii) the result, and then the BF, depends on the prior $f(\underline{\theta} \,|\, H, I)$ about the parameters, which have to be properly modeled from the physics case. A rather simple example, also related to gravitational waves, is shown in [51] and helped dumping down claims of GW detection based on p-values, resulting in fact in ineffective Bayes factors Signal Vs Noise of the order of the unity, with values depending on the model considered. The calculations of the BF's published by the LIGO-Virgo Collaboration are *much* more complicate than those of [51] (see [28] and [4] and references therein, in particular [52]), and they have highly benefitted of Skilling's *Nested Sampling* algorithm [53]. And, for the little I can understand of BBH mergers, the priors on the parameters appear to have been chosen safely, so that the resulting BF's seem very reliable.

[19]William Shakespeare, *The Comedy of Errors*:

> *For know, my love, as easy mayst thou fall*
> *A drop of water in the breaking gulf,*
> *And take unmingled thence that drop again,*
> *Without addition or diminishing,*

12

Figure 3: The Monster (GW150914), Cinderella (LVT151012) and the third sister (GW151226), visiting us in 2015 (Fig. 1 of [4] – see text for the reason of the names). The published 'significance' of the three events (Table 1 of [4]) is, in the order, "$> 5.3\,\sigma$", "$1.7\,\sigma$" and "$> 5.3\,\sigma$", corresponding to the following p-values: 7.5×10^{-8}, 0.045, 7.5×10^{-8}. The log of the Bayes factors are instead (Table 4 of [4]) approximately 289, 23 and 60, corresponding to Bayes factors about 3×10^{125}, 10^{10} and 10^{26}.

The busy plot on the left side shows the sensitivity curves of the two interferometers (red and blue curves, with plenty of resonant peaks) and how the three signals fall inside them (bands with colors matching the wave forms of the right plot). In short, the two curves tell us that a signal of a given frequency can be distinguished from the noise if its amplitude is above them. Therefore all initial parts of the waves, when the black holes begin to spiral around each other at low frequency, are unobservable, and the bands below $\approx 20\,\mathrm{Hz}$ are extrapolations from the physical models. Later, when the frequency increases, the wave enters the sensitivity range,[20] which extends up to a given frequency, after which we 'loose' it. The lower and upper boundary frequencies depend on the amplitude of the signal, as it also happens in acoustics.

The plot on the right shows finally the 'waves'[21] from the instant they enter the optimal 30 Hz sensitivity region (the acoustic analogy depicted in footnote 20 might help):

---

[20]In analogy, imagine someone communicating to us using an audio signal, whose frequency changes with time, from infrasounds to ultrasounds. We can ear the signal only when it is in the acoustic region, conventionally in the range between 20 and 20,000 Hz, although depending from person to person. And, since this sensitivity window is not sharp, close to its edges loud sounds are better eared than quiet ones.

[21]To be more precise, these *are not* data points, but rather the 'adapted filters' that best match them, and therefore they could provide a too optimistic impression of what has really being detected. Therefore we have to use the Bayes factors provided by the collaboration, rather than intuitive judgement based on these wave forms.

- The wave indicated by GW150914 (the 'Monster', with GW standing for gravitational wave and 150914 for the detection date, September 14, 2015) is characterized by high amplitude, but short duration in the sensitivity region, because it fades out at a few hundred hertz.

- GW151226 instead, although of smaller intensity, has a longer 'life' (about 1.7 seconds) in the 'audible' spectrum, and therefore the signature of a BBH merger is also very recognizable.

- Then there is the October 12 event, LVT151012, which has an amplitude comparable to that of GW151226, but smaller duration. It has, nevertheless, about 20 oscillations in the sensitivity region, an information that, combined with the peculiar shape of the signal (remarkably the crests get closer as time passes, while the amplitude increases, until something 'catastrophic' seems to happen) and the fact that two practically 'identical' and 'simultaneous' signals have been observed by the two interferometers 3000 km apart, makes the experts highly confident that this is also a gravitational wave.

However, even if at a first sight it does not look dissimilar from GW151226 (but remember that the waves in Fig. 3 do not show raw data!), the October 12 event, hereafter referred as *Cinderella*, is not ranked as GW, but, more modestly, as LVT, for LIGO-Virgo Trigger. The reason of the downgrading is that *'she' cannot wear a "> 5σ's dress" to go together with the 'sisters' to the 'sumptuous ball of the Establishment.'* In fact Chance has assigned 'her' only a poor, unpresentable $1.7\,\sigma$ ranking, usually considered in the Particle Physics community not even worth a mention in a parallel session of a minor conference by an undergraduate student.[22] But, despite the modest 'statistical significance', experts are highly confident, because of physics reasons[23] (and of their understanding of background), that this is also a gravitational wave radiated by a BBH merger, much more than the 87% quoted in [4].[24]

---

[22]Note how the quoted p-value of 0.045 associated to it is just below the (in-)famous 0.05 "significance" threshold reminded in the xkcd cartoon of Fig. 1. I hope it is so just by chance and that no "p-value ≤ 0.05" requirement was applied to the data, then filtering out other possible good signals.

[23]Detecting something that has good reason to exist , because of our understanding of the Physical World (related to a network of other experimental facts and theories connecting them!), is quite different from just observing an unexpected bump, possibly due to background, even if with small probability, as already commented in footnote 15. [And remember that whatever we observe in real life, if seen with high enough resolution in the N-dimensional phase space, <u>had</u> *very small* probability to occur! (imagine, as a *simplified* example, the pixel content of any picture you take walking on the road, in which N is equal to five, i.e two plus the RGB code of each pixel).]

[24]To understand how much people believe on a scientific statement it is often useful, besides proposing bets [9], to ask about the complementary hypothesis. For example when I see a 90% C.L. upper limit on a quantity, I ask "do you really believe 10% that the value is <u>above</u> that limit", or, even more embarrassing, "please use your method to evaluate the 50% C.L. upper limit, then, whatever number comes out, tell me if you really believe 50-50 that the value could be in either side of the limit, and be ready to accept a bet with 1 to 1 odds in the direction <u>I</u> will choose." (To learn more about the absurdities of 'frequentistic coverage' and also about limits derived from 'objective Bayesian methods,' see section 10.7 and chapter 13 of [8].) In the case of this 87% probability that LVT151012 is a GW from BBH merger the question to ask is "do you

Indeed the most useful number experimentalists can provide to the scientific community to quantify how the experimental data alone favor the 'Signal' hypothesis is the Bayes factor, as expounded in the preamble. And this factor is very large also for Cinderella: $\approx 10^{10}$. This means that, even if your initial odds Signal Vs Noise were one to one million, the observation of the LIGO interferometers turns them into 10,000 to 1, i.e. a probability of BBH merger of 99.99%.[25]

Now the question is, how can a modest 1.7σ effect be compatible with a Bayes factor as large as 10^{10}? The solution to this apparent paradox will be given in the next section, but I anticipate the answer: *p-values and BF's are two different things, and there is no simple, general rule, inside probability theory, that relates them.*

6 P-values Vs Bayes factors

Having discussed at length this topic elsewhere (see in particular sections 1.8, and 10.8 of [8]), I sketch here the main points, with the help of some plots. This is obviously a didactic example and does not enter at all into the (very complicate and CPU time consuming) details of the analysis of the interferometer data (see footnote 18). In particular a direct observation will be considered, while in general hypothesis tests are performed on a *statistic* chosen with *large freedom*.[26] So we just consider here simple models H_i that could produce the quantity x according to pdf's $f(x \mid H_i, I)$.

- As reminded above, according to probability theory what matters for the update of relative beliefs is the ratio of the pdf's. For example the observation $x_m = 5$ shown in the upper plot of Fig. 4 modifies our beliefs in favor of H_3, with respect to H_1 and H_2, *no matter the size of the area under the pdf's right of x_m.*

- In particular H_2 is ruled out ('falsified') because, being $f(x_m \mid H_2) = 0$, it cannot produce the observation, despite it provides the highest probability of $X > x_m$.[27]

---

really believe 13%, i.e. about 1 to 7, that this event <u>is not</u> a gravitational wave due to a BBH merger?" (and we should not accept any answer which is, even partially, based to the smallness of the sigmas.) As a matter of fact I find this 87% beyond my understanding, because such a probability has to depend on the prior probability of BBH mergers. For this reason I will focus in the sequel only on Bayes factors and how they (*do not simply*) relate to p-values.

[25]Note that this probability depends on set of hypotheses taken in account. If another, alternative physical hypothesis H^* to explain the LIGO signals is considered, than the Bayes factor of H^* Vs "BBH merger" has to be evaluated, and the absolute probabilities re-calculated accordingly.

[26]It is perhaps important to remind that, among other problems, p-values are affected by arbitrarity of the test variable used (see e.g. [54]), as well by the chosen subset of data. With some experience I have developed my *golden rule*:

        The more exotic is the name of the test, the less believe the result.

The rationale is that I'm pretty sure that several more common tests have been discarded before arriving to that which provided the desired significance.

[27]Note that, contrary to the similar probabilities for the models H_1 and H_3, this 13% is not a p-value, because $f(x \mid H_2) \geq f(x_m \mid H_2) \ \forall x > x_m$, while a p-value implies an integral on 'less probable' values.

Figure 4: Several models that could have produced the observed value of x_m [8].

- It follows that, if the values of pdf's $f(x_m \,|\, H_i)$ are equal for all H_i, as in the lower plot of Fig. 4, then the experiment is irrelevant and we hold our beliefs, *independently of* how far x_m occurs from the expected values $E[X \,|\, H_i]$, or of the size of the area left or right x_m.

- The reason why p-values 'often work' (and can then be useful *alarm bells* when getting experiments running, or validating freshly collected data), is quite simple.

  - Small p-values are normally associated to small values of the pdf, as shown in the upper plot of Fig. 5.

  - It is then *conceivable* an alternative hypothesis H_1 such that $f(x_m \,|\, H_1) \gg f(x_m \,|\, H_0)$, as shown in the bottom plot of Fig. 5.

16

Figure 5: Pdf's of X given the null hypothesis H_0 and the alternative hypothesis H_1.

- Then, **if** this is the case, the observed x_m *would push our beliefs towards* H_1, in the sense $\mathrm{BF}(H_1 : H_0) = \frac{f(x_m \mid H_1)}{f(x_m \mid H_0)} \gg 1$.

- **BUT** we need to take into account also the priors odds $P(H_1 \mid I)/P(H_0 \mid I)$.

- In the extreme case such a conceivable H_1 <u>could not exist</u>, or it could be <u>not believable</u>,[28] or it could be just <u>ad hoc</u>, as it happens in recent years, with a plethora of 'theorists' who give credit to any fluctuation. If this is the case, as it is often the case in frontier physics, then

$\Rightarrow$ $P(H_1 \mid I)/P(H_0 \mid I) \to 0$

$\Rightarrow$ *the smallness of the p-value is irrelevant!*

(Note that if, instead of the smallness of the value of the pdf, the rational were really the smallness of the area below the pdf, than the absurd situation might arise in which one could choose a "rejection area" anywhere, as shown in chapter 1 of [8].)

---

[28]For the distinction between what is *conceivable* ("Nothing is more free than the imagination of man") and what is *believable* a reference to David Hume [46] is a must.

- Finally, in order to understand the apparent paradox of large p-value and indeed very large BF, think at a very predictive model H_1, whose pdf of the observable x overlaps with that of H_0, like in the upper plot of Fig. 6. We clearly see that $f(x_m \,|\, H_1) \gg f(x_m \,|\, H_0)$, thus resulting in a Bayes factor highly in favor of H_1, although the p-value calculated from the null hypothesis H_0 would be absolutely *insignificant*. Something *like that* occurs in the analysis of the gravitational wave analysis, the case of Cinderella being the most striking one.[29]

- And 'paradoxically' – this is just a colloquial term, since there is no paradox at all – large deviations from the expected value of x given H_0, corresponding to small p-values, are those which favor H_0, if H_1 and H_0 are the only hypotheses in hand, as shown in the bottom plot of the same figure. Now, in the light of these examples, I simply re-propose you the following sentence from the first principle of the ASA's statement "The smaller the *p*-value, the greater the statistical incompatibility of the data with the null hypothesis, if the underlying assumptions used to calculate the *p*-value hold." [2] As you can now understand, it is not a matter of assumptions concerning H_0, but rather on whether alternative hypotheses to H_0 are conceivable and, more important, believable!

---

[29] I would like to remind that this is just an academic example to show that effects of this kind are possible and, as far as the GW analysis, I rely on the LIGO-Virgo collaboration for the evaluation of p-values and Bayes factors. I am not arguing at all that there could be mistakes in the calculation of the p-values, but rather that it is the interpretation of the latter to be troublesome. Finally, people mostly used to perform χ^2 tests must have already realized that the example does not apply *tout court* to what they do, because in that case H_1 is usually 'richer' than H_0 and it has then a higher level of adaptability. Therefore the observed value of χ^2 decreases (with a 'penalty' that frequentists quantify with a reduced number of degree of freedom). As a consequence, the measured value of the test variable is different under the two hypothesis, and, in order to distinguish them, let us indicate the first by χ_0^2 and the second by χ_1^2. What instead still holds, of the example sketched in the text, is that the adaptability of H_1 makes the p-value calculated from $f(\chi_1^2 \,|\, H_1)$ larger that that calculated from $f(\chi_0^2 \,|\, H_0)$,

$$\int_{\chi_{1_m}^2}^{\infty} f(\chi_1^2 \,|\, H_1) \, d\chi_1^2 \;>\; \int_{\chi_{0_m}^2}^{\infty} f(\chi_0^2 \,|\, H_0) \, d\chi_0^2 \,,$$

and therefore H_1 'gets preferred' to H_0. But, as stated in the text, the alternative hypothesis H_1 could be hardly believable, and therefore its 'nice' p-value will not affect the credibility of H_0. This almost regularly happens when suspicions against H_0 only arise from *event counting* in a particular variable, *without any specific physical signature*. [As a side remark, I would like to point out, or to remind, that one of the nice features of the Bayes factor calculated integrating over the prior parameters of the model, as sketched in footnote 18, is that models which have a large numbers of parameters, whose possible values *a priori* extend over a large (hyper-)volume, are suppressed by the integral (F.1) with respect to 'simpler' models. This effect is known as *Bayesian Occam's razor* and is independent from other considerations which might enter in the choice of the priors. Those interested to the subject are invited to read chapter 28 of David MacKay's great book [55].]

Figure 6: Pdf's of X given the null hypothesis H_0 and the alternative hypothesis H_1 (case of overlapping pdf's).

6.1 Playing with simulations

I hope it is now clear the reason why p-values and Bayes factors have in principle nothing to do with each other, and why p-values are not only responsible of unjustified claims of discoveries, but might also relegate genuine signals to the level of fluke, or reduce their 'significance', the word now used as normally understood and not with the 'technical meaning' of statisticians. But since I know that many might not be used with the reasoning just shown, I made a little R script [56], so that those who are still sceptical can run it and get a feeling of what is going on.

```
# initialization
mu.H0 <- 0; sigma.H0 <- 1
mu.H1 <- 0; sigma.H1 <- 1e-3
p.H1  <- 1/2
mu    <- c(mu.H0, mu.H1)
sigma <- c(sigma.H0, sigma.H1)

# simulation function
simulate <- function() {
  M <- rbinom(1, 1, p.H1); x <- rnorm(1, mu[M+1], sigma[M+1])
  x <- rnorm(1, mu[M+1], sigma[M+1])
  p.val <- 2 * pnorm(mu[1] - abs(x-mu[1]), mu[1], sigma[1])
  BF <- dnorm(x, mu[2], sigma[2]) / dnorm(x, mu[1], sigma[1])
  lBF <- dnorm(x, mu[2], sigma[2], log=TRUE) - dnorm(x, mu[1], sigma[1], log=TRUE)
  cat(sprintf("x = %.5f  =>  p.val = %.2e,  BF = %.2e  [ log(BF) = %.2e ]\n",
              x, p.val, BF, lBF))
  return(M)
}
```

By default H_0 is simply a standard Gaussian distribution ($\mu = 0$ and $\sigma = 1$), while H_1 is still a Gaussian centered in 0, with a very narrow width ($\sigma = 1/1000$). The prior odds are set at 1 to 1, i.e. $P(H_1) = P(H_0) = 1/2$. Each call to the function simulate() prints the values that we would get in a real experiment (x, p-value, Bayes factor and its log) and returns the true model (0 or 1), stored in a vector variable for later check. In this way you can try to infer what was the real cause of x before knowing the 'truth' (in simulations we can, in physics we cannot!). Here are the results of a small run, with x = 12 chosen in order to fill the page, thus postponing the solution to the next one.

```
> set.seed(150914); n=12; M <- rep(NA, n); for(i in 1:n) M[i] <- simulate()
x = -0.00079  =>  p.val = 9.99e-01,  BF = 7.29e+02  [ log(BF) = 6.59e+00 ]
x = -0.62293  =>  p.val = 5.33e-01,  BF = 0.00e+00  [ log(BF) = -1.94e+05 ]
x = -0.00029  =>  p.val = 1.00e+00,  BF = 9.57e+02  [ log(BF) = 6.86e+00 ]
x = -0.00162  =>  p.val = 9.99e-01,  BF = 2.68e+02  [ log(BF) = 5.59e+00 ]
x = -0.39258  =>  p.val = 6.95e-01,  BF = 0.00e+00  [ log(BF) = -7.71e+04 ]
x = -0.82578  =>  p.val = 4.09e-01,  BF = 0.00e+00  [ log(BF) = -3.41e+05 ]
x = 0.00073  =>  p.val = 9.99e-01,  BF = 7.69e+02  [ log(BF) = 6.64e+00 ]
x = -0.00012  =>  p.val = 1.00e+00,  BF = 9.93e+02  [ log(BF) = 6.90e+00 ]
x = 0.22295  =>  p.val = 8.24e-01,  BF = 0.00e+00  [ log(BF) = -2.48e+04 ]
x = -0.00022  =>  p.val = 1.00e+00,  BF = 9.76e+02  [ log(BF) = 6.88e+00 ]
x = 0.00117  =>  p.val = 9.99e-01,  BF = 5.07e+02  [ log(BF) = 6.23e+00 ]
x = -1.03815  =>  p.val = 2.99e-01,  BF = 0.00e+00  [ log(BF) = -5.39e+05 ]
```

And *the winners are*:

```
> M
 [1] 1 0 1 1 0 0 1 1 0 1 1 0 0 1 1 0 0 0 0 1 0 1 1
```

It should not be any longer a surprise that the best figure to discriminate between the two models is the Bayes factor and not the p-value.[30] You can now play with the simulations, varying the parameters. If you want to get a situation yielding Bayes factors of $\mathcal{O}(10^{10})$ you can keep the standard parameters of H_0, fixing instead `mu.H1` at 1.7 and `sigma.H1` at $\approx 4 \times 10^{-10}$. Then you can choose `p.H1` at wish and run the simulation. (You also need to change the numbers of digits of x, replacing "`%.5f`" by "`%.11f`" inside `sprintf()`.)

7 Conclusions

Uncritical or wishful use of p-values can be dangerous, not to speak of unscrupulous p-hacking. While years ago these criticisms were raised by a minority of thorny Bayesians, now the effect on the results in several fields of science and technology is felt as a primary issue.[31] The statement of the American Statistical Association is certainly commendable in

---

[30]If you don't like how the p-value is calculated in the script, because you might argue about one-side or two-sides tail(s), you are welcome to recalculate it, but the substance of the conclusions will not change.

[31]In the meanwhile it seems that particle physicists are hard in learning the lesson and the number of graves in the *Cemetery of physics* (Fig. 2) has increased since 1985, the last *funeral* being recently celebrated in Chicago on August 5, with the following obituary for the *dear departed*: "The intriguing hint of a possible resonance at 750 GeV decaying into photon pairs, which caused considerable interest from the 2015 data, has not reappeared in the much larger 2016 data set and thus appears to be a statistical fluctuation" [57]. And de Rujula's *dictum* (footnote 14) gets corroborated. Someone would argue that this incident has happened because the sigmas were only about three and not five. But it is not a question of sigmas, but of Physics, as it can be understood by those who in 2012 incorrectly turned the 5σ into 99,99994% "discovery probability" for the Higgs [58], while in 2016 are sceptical in front of a 6σ claim ("if I have to bet, my money is on the fact that the result will not survive the verifications" [59]): the famous "du sublime au ridicule, il n'y a qu'un pas" seems really appropriate! (Or the less famous, outside Italy, "siamo uomini o caporali!?") Seriously, the question is indeed that, now that predictions of New Physics around what should have been a *natural* scale substantially all failed, the only 'sure' scale I can see seems Planck's scale. I really hope that LHC will surprise us, but hoping and believing are different things. And, since I have the impression that are too many nervous people around, both among experimentalists and theorists, and because the number of possible histograms to look at is quite large, after the *easy bets* of the past years (against CDF peak and against superluminar neutrinos in 2011; in favor of the Higgs boson in 2011; against the 750 GeV di-photon in 2015, not to mention that against Supersymmetry going on since it failed to predict new phenomenology *below* the Z_0 – or the W? – mass at LEP, thus inducing me more than twenty years ago to gave away all SUSY Monte Carlo generators I had developed in order to optimize the performances of the HERA detectors.)* I can serenely bet, as I keep saying since July 2012, that **the first 5-sigma claim from LHC will be a fluke**. (I have instead little to comment on the sociology of the Particle Physics theory community and on the validity of 'objective' criteria to rank scientific value and productivity, being the situation self evident from the hundreds of references in a review paper which even had in the front page a fake PDG entry for the particle [60] and other amenities you can find on the web, like [61].)

* **Note added**: on August 22, 2016 a supersymmetry bet among theorists has been settled in Copenhagen, *declaring winners those who betted against supersymmetry* [62]. But I do not think all SUSY supporters will agree, because some of them seem to behave like the guy who said (reference missing) "I will not die, and nobody will be able to convince me of the opposite" – try to convince a dead man he died!

addressing the issue, but it is in my opinion unsatisfactory not admitting that the question is inherent to all statistical methods that refuse the very idea of probability of hypotheses, or of "probability of causes", i.e. what Poincaré used to call "the essential problem of the experimental method."

While I had experienced several times in the past, including this winter, claims of possible breaking discoveries in Particle Physics simply due to misinterpretations of p-values, for the first time I have realized of a case in which judgements based on p-values strongly reduce the 'significance' of important results. This happens with the gravitational wave events reported this year by the LIGO-Virgo collaboration, and in particular with the October 12 events timidly reported as a LIGO-Virgo Trigger ('Cinderella'), because of its 1.7 sigmas, in spite of the huge Bayes factor of about 10^{10}, that should instead convince any hesitating physicist about its nature of a gravitational wave radiated by a Binary Black Hole merger, especially in the light of the other, more solid two events ('the two sisters').[32] I hope than that LVT151012 will be upgraded to GW151012 and that in future searches the Bayes factor will become the principal figure of merit to rank gravitational wave candidates.

---

[32]The last point deserves a comment, because someone would object that the three events are "independent" and, "having nothing to do with each other, we have to prove one by one 1) first, that it is a gravitational wave, and *then* 2) that it comes from BBH merger." In reality it is consistency of many things, including the fact that the values of the inferred parameters fall in the expected region, that makes us to believe that they are gravitational waves <u>and</u> come from a BBH merger. This is because Physics, meant as a Science, i.e. an activity of our minds to understand the Physical World, can be viewed as a large *network* of experimental facts and models, connecting each other ("a matrix of beliefs", as historian Galison puts it [63]). For this reason it is very hard, or even impossible, to accommodate in the overall picture a new observation that breaks dramatically the net, like the 2011 'superluminar neutrinos.' Not by chance the title of the February 11 paper was *Observation of Gravitational Waves from a Binary Black Hole Merger* stressing *both observations at once* (or if you like 'discoveries' – but I don't want to enter into the question of what is 'discovery' and what is 'observation', and I find it commendable that the collaboration used low profile terminology). Therefore, after the first event we feel highly confident that events of that kind, with masses of that order of magnitude do exist, and with this respect the three events are not independent, if we refer to probabilistic independence. More precisely they are *positively correlated*, i.e.

$$P(E_2 = \text{``BBHm's GW''} \,|\, E_1 = \text{``BBHm's GW''}, I) \quad > \quad P(E_2 = \text{``BBHm's GW''} \,|\, I)$$
$$P(E_1 = \text{``BBHm's GW''} \,|\, E_2 = \text{``BBHm's GW''}, I) \quad > \quad P(E_1 = \text{``BBHm's GW''} \,|\, I),$$

and so on. This effect, indeed rather intuitive, can been shown to occur in a quantitative way, modelling Galison's matrix of beliefs with a (simplified) *probabilistic network* 'Bayesian network'. For this reason our belief that also Cinderella is a gravitational wave from a BBH merger increases in the light that also the sisters are objects of the same kind. Note that this corroboration effect acts on the priors, while the Bayes factor should only contain the experimental information. But this is not exactly true, due to role that the priors on the model parameters play in the calculation of the Bayes factor via the integral $(F.1)$ of footnote 18. As soon as we start getting information about the BBH merger parameters the prior pdf $f(\underline{\theta} \,|\, H, I)$ to analyze the next events becomes less 'diffuse' than how they initially were, thus increasing the value of the integral ($\rightarrow$ "Occam razor") and then the resulting Bayes factor. (For a toy model showing the effect of mutually corroborating hypotheses see e.g. the Bayesian network described in Appendix J of [6].)

Note added: it is interesting to remark how, after six months from the first announcement, with much emphasis on the sigmas to prove its origin (plus Bayes factors), the Monster is finally considered 'self evident', or more precisely, "strong enough to be apparent, without using any waveform model, in the filtered detector strain data" [64]. So proceeds Science: the 'matrix of belief' has been clearly extended.

I finally conclude with some questions asked at the end of talk on which this paper is based.

- *Which Bayes factor would characterize the 750 GeV excess?*
  The result depends on the model to explain the excess[33] and an answer came the week after MaxEnt 2016 by Andrew Fowlie [66]. For the model considered he got a BF *around* 10, the exact value being irrelevant: a weak indication, but nothing striking to force sceptics to change substantially their opinion.[34]

- *Could have CDF at Fermilab claimed to have observed the Higgs boson if they had done a Bayesian analysis?*
  I am quite positive they could have it, also because the prior on the possible values of the Higgs mass was not so vague and well matching the value found later, and therefore the Bayes Factor would have been rather high (and the prior probability of a possible manifestation of the boson in the final state was high too).

References

[1] B. P. Abbott et al. (LIGO Scientific Collaboration and Virgo Collaboration), *Observation of Gravitational Waves from a Binary Black Hole Merger*, PRL **116**, 061102 (2016), https://dcc.ligo.org/public/0122/P150914/014/LIGO-P150914_Detection_of_GW150914.pdf

[2] R. L. Wasserstein and N. A. Lazara, *The ASA's Statement on p-Values: Context, Process, and Purpose*, The American Statistician, **70:2** (2016) 129-133, DOI: 10.1080/00031305.2016.1154108, http://dx.doi.org/10.1080/00031305.2016.1154108

---

[33]As an example from Particle Physics of model dependent Bayes factors see [65].

[34]A side question is how an experimental team can report the Bayes factor, since it depends on the alternative model. Obviously it cannot (one of "Laplace's teachings"), but they provide Bayes factors using 'popular' models, or it could just report the integral which appears in the denominator, and provide informations that allows other physicists to evaluate the numerator, depending on the their model.

[3] M. Baker, *Statisticians issue warning over misuse of P values*, Nature, **531** (2016) 151.

[4] B. P. Abbott et al (LIGO Scientific Collaboration and Virgo Collaboration), *Binary Black Hole Mergers in the first Advanced LIGO Observing Run*, `http://arxiv.org/abs/arXiv:1606.04856`.

[5] I. J. Good, *A list of properties of Bayes-Turing factors*, document declassified by NSA in 2011, `https://www.nsa.gov/news-features/declassified-documents/tech-journals/assets/files/list-of-properties.pdf`.

[6] G. D'Agostini, *A defense of Columbo (and of the use of Bayesian inference in forensics): A multilevel introduction to probabilistic reasoning*, arXiv:1003.2086, `http://arxiv.org/abs/1003.2086`.

[7] S. B. McGrayne, *The theory that would not die: How Bayes' rule cracked the enigma code, hunted down russian submarines, and emerged triumphant from two centuries of controversy*, Yale University Press 2012. (Video of the presentation of the book at Google available at `https://www.youtube.com/watch?v=8oD6eBkjF9o`.)

[8] G. D'Agostini, *Bayesian reasoning in data analysis – a critical introduction*, World Scientific 2003.

[9] G. D'Agostini, *Probably a discovery: Bad mathematics means rough scientific communication*, `http://arxiv.org/abs/1112.3620`.

[10] `https://en.wikipedia.org/wiki/Misunderstandings_of_p-values`.

[11] `http://en.wikipedia.org/wiki/P-value`
`https://en.wikipedia.org/wiki/Misunderstandings_of_p-values`.

[12] G. D'Agostini, *From Observations to Hypotheses: Probabilistic Reasoning Versus Falsificationism and its Statistical Variations*, 2004 Vulcano Workshop on Frontier Objects in Astrophysics and Particle Physics, Vulcano (Italy), `http://arxiv.org/abs/physics/0412148`

[13] H. Poincaré, *"Science et Hypothèse"*, 1905.

[14] J. Skilling, *Introductory tutorial* at the MaxEnt 2016, July 10-15, 2016 Ghent, Belgium, `http://www.maxent2016.org`.

[15] See e.g. `https://www.youtube.com/watch?v=EYPapE-3FRw`.

[16] D. Overbye, *Physicists in Europe Find Tantalizing Hints of a Mysterious New Particle*, The New York Times, December 15, 2015, `http://www.nytimes.com/2015/12/16/science/physicists-in-europe-find-tantalizing-hints-of-a-mysterious-new-particle.html?_r=0`.

[17] J. Parsons, *CERN announces potential discovery of a new Higgs Boson particle at the Large Hadron Collider*, Mirror, December 17, 2015, `http://www.mirror.co.uk/news/technology-science/science/cern-announces-potential-discovery-new-7027421`

[18] M. Delmastro, *Qualcosa di nuovo da LHC? Solo il tempo lo dirà*, Le Scienze 19 dicembre 2015, `http://www.lescienze.it/news/2015/12/19/news/qualcosa_di_nuovo_a_lhc_solo_il_tempo_lo_dira_-2900622/`

[19] F. Flam, *Lies, Damned Lies and Physics*, Bloomberg View, December 30, 2015, `https://www.bloomberg.com/view/articles/2015-12-30/lies-damned-lies-and-physics`.

[20] B. Crew, *Evidence of a new particle that could break the standard model of physics is mounting* , Science Alert, March 21, 2016, `http://www.sciencealert.com/evidence-of-a-new-particle-that-could-break-the-standard-model-of-physics-is-mounting`.

[21] Ph. Ball, *I'd put a tenner – but not a ton – on the Higgs-Boson existing*, The Guardian, December 23, 2011, `https://www.theguardian.com/commentisfree/2011/dec/23/critical-scientist-higgs-boson`.

[22] S. Goodman, *A Dirty Dozen: Twelve P-Value Misconceptions*, Seminars in Hematology **45** (2008) 135, `http://www.perfendo.org/docs/BayesProbability/twelvePvaluemisconceptions.pdf`.

[23] P.S. Laplace, *Essai philosophique sur les probabilités*, 1814, `http://books.google.it/books?id=JrEWAAAAQAAJ` (the English quotes in this paper are taken from A.I. Dale's translation, Springer-Verlag, 1995).

[24] N. Fenton, D. Berger, D. Lagnado, M. Neil and A. Hsu, *When 'neutral' evidence still has probative value (with implications from the Barry George Case)*, Science and Justice **54** (2014) 274, `http://www.scienceandjusticejournal.com/article/S1355-0306(13)00059-2/abstract`.

[25] G. D'Agostini, *Asymmetric Uncertainties: Sources, Treatment and Potential Dangers*, arXiv:physics/0403086, `http://arxiv.org/abs/physics/0403086`.

[26] European Network of Forensic Science Institutes, *ENFSI Guidelines for the Evaluative Reporting in Forensic Science*, March 8, 2015, `http://www.enfsi.eu/sites/default/files/documents/external_publications/m1_guideline.pdf`

[27] European Network of Forensic Science Institutes, *Best Practice Manual for the Forensic Examination of Digital Technology*, ENFSI-BPM-FIT-01, November 2015. `http://www.enfsi.eu/sites/default/files/documents/1._forensic_examination_of_digital_technology_0.pdf`

[28] B. P. Abbott et al (LIGO Scientific Collaboration and Virgo Collaboration), *Properties of the Binary Black Hole Merger GW150914*, PRL **116**, 241102 (2016), `https://dcc.ligo.org/LIGO-P1500218/public`

[29] `https://en.wikiquote.org/wiki/Truth#H`

[30] T. Siegfried, T. (2010), *Odds Are, It's Wrong: Science Fails to Face the Shortcomings of Statistics*, Science News **177**, 26, `https://www.sciencenews.org/article/odds-are-its-wrong`

[31] A. Caldwell, Lectures at the School on Bayesian analysis in Physics and Astronomy, Stellenbosch, South Africa, 23-26 November 2013.

[32] G. Naik, *Scientists' Elusive Goal: Reproducing Study Results*, The Wall Street Journal, December 2, 2011, `http://www.wsj.com/articles/SB10001424052970203764804577059841672541590`.

[33] E. Iorns, *Is medical science built on shaky foundations?*, New Scientist, 12 September 2012, `https://www.newscientist.com/article/mg21528826-000-is-medical-science-built-on-shaky-foundations/`.

[34] P. Jump, *More than half of psychology papers are not reproducible*, Times Higher Education, August 27, 2015, `https://www.timeshighereducation.com/news/more-half-psychology-papers-are-not-reproducible`.

[35] P. Jump, *Reproducing results: how big is the problem?*, Times Higher Education, September 3, 2015, `https://www.timeshighereducation.com/features/reproducing-results-how-big-is-the-problem`.

[36] R. Horton, em Offline: What is medicine's 5 sigma?, The Lancet **385** (2015) 1380, `http://www.thelancet.com/journals/lancet/article/PIIS0140-6736%2815%2960696-1/fulltext`

[37] xkcd, *Significant*, `http://xkcd.com/882/`

[38] L. D. Nelson, *False-positives, p-hacking, statistical power, and evidential value*, BITSS 2014 Summer Institute, June 2014, `https://bitssblog.files.wordpress.com/2014/02/nelson-presentation.pdf`.

[39] A. Charpentier, *P-hacking, or cheating on a p-value*, R-bloggers, June 2015, `https://www.r-bloggers.com/p-hacking-or-cheating-on-a-p-value/`

[40] `https://en.wiktionary.org/wiki/If_you_torture_the_data_long_enough,_it_will_confess_to_anything`

[41] D. Trafimow, *Editorial*, Basic and Applied Social Psychology *36* (2014) 1. `http://www.tandfonline.com/doi/full/10.1080/01973533.2014.865505`

[42] S. Novella, *Psychology Journal Bans Significance Testing*, Science-Based Medicine, February 25, 2015, `https://www.sciencebasedmedicine.org/psychology-journal-bans-significance-testing/`

[43] A. Gelman, *Psych journal bans significance tests; stat blogger inundated with emails*, February 26, 2015, `http://andrewgelman.com/2015/02/26/psych-journal-bans-significance-tests-stat-blogger-inundated-with-emails/`

[44] J. Berger, Ph. Dawid, J. Kadane, T. O'Hagan, L. Piricchi, Ch. P. Robert and D. Szucs, contributions to *Banning null hypothesis significance testing*, ISBA Bullettin 22, March 2015, 5, `https://bayesian.org/sites/default/files/fm/bulletins/1503.pdf`

[45] D. Trafimow and M Marks, *Editorial*, Basic and Applied Social Psychology *37* (2015) 1, `http://www.tandfonline.com/doi/full/10.1080/01973533.2015.1012991`

[46] D. Hume, *A Treatise of Human Nature, 1739*; *An Enquiry Concerning Human Understanding*, 1748.
(Also available as audiobooks at `LibriVox`, with links to the online texts:
`https://librivox.org/treatise-of-human-nature-vol-1-by-david-hume/`;
`https://librivox.org/an-enquiry-concerning-human-understanding-by-david-hume/`.)

[47] A. de Rujula, private communication, December 2011.

[48] A. de Rujula, *"Snapshots of the 1985 high energy physics panorama"*, Proc. of the Int. Europhys. Conf. on High-Energy Physics, Bari (Italy), July 1995, L. Nitti and G. Preparata eds.

[49] G. D'Agostini and G. Degrassi, *Constraints on the Higgs Boson Mass from Direct Searches and Precision Measurement*, Eur.Phys.J. **C10** (1999) 663, `http://link.springer.com/article/10.1007%2Fs100529900171` (arXiv:hep-ph/9902226, `http://arxiv.org/abs/hep-ph/9902226`).

[50] G. D'Agostini and G. Degrassi, *Constraining the Higgs boson mass through the combination of direct search and precision measurement results*, arXiv:hep-ph/0001269, `http://arxiv.org/abs/hep-ph/0001269`.

[51] P. Astone, G. D'Agostini and S. D'Antonio, *Bayesian model comparison applied to the Explorer-Nautilus 2001 coincidence data*, Class.Quant.Grav. **20** (2003) S769-S784 (arXiv:gr-qc/0304096, `http://xxx.lanl.gov/abs/gr-qc/0304096`).

[52] J. Veitch and A. Vecchio, *Bayesian coherent analysis of in-spiral gravitational wave signals with a detector network*, Phys. Rev. **D 81** (2010) 062003 (arXiv:0911.3820).

[53] J. Skilling, *Nested Sampling for General Bayesian Computation*, Bayesian Analysis **1** (2006) 833, `http://www.mrao.cam.ac.uk/~steve/maxent2009/images/skilling.pdf`, `https://en.wikipedia.org/wiki/Nested_sampling_algorithm`.

[54] G. K. Kanji, *100 statistical tests*, SAGE Publications Ltd, 2006.

[55] D. J.C. MacKay, *Information theory, Inference and learning algorithms*, Cambridge University Press, 2003, `http://www.inference.phy.cam.ac.uk/itila/book.html`

[56] R Core Team (2016). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`. (Script at `http://www.roma1.infn.it/~dagos/prob+stat.html`.)

[57] Email to the CERN users by the CERN DG Office, August 5, 2016.

[58] Corriere della Sera, *"Trovata la particella di Dio" – Una caccia lunga mezzo secolo*, July 3, 2012, `http://www.corriere.it/scienze/12_luglio_03/trovata-particella-di-dio-caccia-lunga-mezzo-secolo-giovanni-caprara_b967689e-c4d0-11e1-a141-5df29481da70.shtml`.

[59] Repubblica, *Da un laboratorio ungherese spunta la quinta forza*, May 25, 2016, `http://www.repubblica.it/scienze/2016/05/25/news/modello_standard_forze_fondamentali_cern_lhc_particelle_fondamentali_materia_oscura_bosone-140567449/`.

[60] A. Strumia. *Interpreting the 750 GeV digamma excess: a review*, CERN-TH-2016-131, `http://arxiv.org/abs/1605.09401`.

[61] *Game of Thrones: 750 GeV edition*, Résonances, June 18 2016, `http://resonaances.blogspot.co.uk/2016/06/game-of-thrones-750-gev-edition.html`.

[62] N. Walchover, *Supersymmetry bet settled with cognac*, Quanta Magazine, August 22, 2016, `https://www.quantamagazine.org/20160822-supersymmetry-bet-settled-cognac/`.

[63] P.L. Galison, *How experiments end*, The University of Chicago Press, 1987.

[64] P.B: Abbot et al., LIGO-Virgo Collaboration, *The basic physics of the binary black hole merger GW150914*, arXiv:1608.01940, `http://arxiv.org/abs/1608.01940`.

[65] D. Ghosh, M. Nardecchia and S. A. Renner, *Hint of lepton flavour non-universality in B meson decays*, J. High Energ. Phys. (2014) 131, `http://link.springer.com/article/10.1007/JHEP12(2014)131` `http://arxiv.org/pdf/1408.4097.pdf`

[66] A. Fowlie, *Bayes-factor of the ATLAS diphoton excess*, arXiv:1607.06608, `http://arxiv.org/abs/1607.06608`.