Another insight into the null hypothesis significant testing

Gilbert Kirkorian

Gilbert Kirkorian is a retired professor of cardiology at the Université Claude Bernard - Lyon 1

Previous affiliations:

- UMR 5558, CNRS, Villeurbanne, France
- Université Claude Bernard - Lyon 1, Villeurbanne, France
- Hospices Civils de Lyon, Lyon, France

Address:

31, rue Denfert Rochereau
69004 Lyon
France

Email: gkirkorian@free.fr

Various versions of this paper have been rejected by the following journals:

Journal	Date of rejection	Manuscript number
Journal of the American Statistical Association	September 6th 2011	JASA T11-608
New England Journal of medicine	September 26th 2011	11-1014
Statistics in medicine	November 15th 2011	SIM-11-0692
Nature	August 14h 2015	2015-08-11175
Plos One	November 12th 2015	PONE-D-15-37643

Abstract

The analysis of clinical randomized studies traditionally refers to statistical tests which provide the probability of observing data under the null hypothesis in populations of patients randomly drawn from an infinite fictive population. However, the rationale for such an approach remains uncertain since, in reality, randomization is performed within a finite population and the definition of the null hypothesis is ambiguous. This paper proposes a method that analyses all possible combinations of two samples drawn from a finite population, and then enumerates all possible hypotheses and configurations. By using a combinational analysis exploring the weight of all possible and logical hypotheses and configurations of the observed data, this paper shows that it becomes possible to calculate the exact probability of each hypothesis or group of hypotheses for any observed configuration, hence the probability of effectiveness, harm, or no effect of an investigational intervention. Additionally, this method leads to question the real meaning of the null hypothesis, the confidence intervals, and the risk ratios.

Introduction

The usefulness of the null hypothesis significance testing as a method to infer new knowledge from experimental data has been a matter of intense debate over the past 90 years[1]. Opponents of the p-value promote either the use of confidence intervals[2-4], of risk ratios[5], or of a Bayesian approach[6], whereas others reject every inferential statistics as an a priori procedure[7]. Arguments in favour or against the null hypothesis significance testing are frequently conceptual and obscure, rarely based on formal demonstrations of its logic. In the traditional view, the analysis of epidemiological studies refers to statistical tests which provide the probability of obtaining results as extreme as those theoretically expected under the hypothesis of no effect, referred to as the null hypothesis or H0[8]. However, the observed data refer not only to those actually observed but also to data that have not been observed, those more extreme than those observed. Moreover, H0 is never well defined and H1, the so-called alternative hypothesis, is in fact an aggregation of a large number of hypotheses that are put together and presented, without any formal demonstration, as a single normal distribution. To infer the probability of effect from the p-value, Diamond and Forrester proposed applying the Bayes formula, but could not demonstrate how to calculate the probability of H1 given the data[9]. Inferring the probability of effect among a considerable number of alternative hypotheses given the data from the inverse probability of observing observed, and unobserved, data under a poorly defined null hypothesis remains highly hypothetical. To better understand the real meaning of the numerical values extracted from observed data in randomized studies, it is worth looking at them in a different way. In this paper, the method consists of the enumeration of all possible hypotheses compatible with the observed data within a finite population, incorporating the hypothesis of a constant individual behaviour independent of the allocation of patients in the treated and untreated groups.

Methods

In this paper, I propose a combinational analysis applied to a finite and fixed population of N patients, which will be called the source population, randomly separated into two groups of size m1 and m2: patients in group 1 are given a placebo whereas patients in group 2 are given a treatment under investigation. The two groups defined by the randomization process are only a pair among a considerable but finite number of possible and distinct paired groups which can be drawn from the source population, more precisely the combination of m1 or m2 subjects among N. For the sake of simplicity, endpoints are defined as deaths. At the end of the study, data are analysed as the number of deaths among patients left untreated in group 1 and the number of deaths among treated patients in group 2. The source population is considered to be composed of patients with a predefined spontaneous outcome and a predefined outcome under the treatment, independently of the patient group allocation. For each patient, three dimensions of possible outcomes are defined. The first dimension relates to the possible outcomes without treatment. For each patient, the spontaneous outcome is either death or survival. The second dimension relates to the outcomes of patients on treatment who would have died if they had had no treatment. In this case, for each patient, the two possible outcomes are death, the treatment is ineffective; or survival, the patient has been saved by the treatment. The third dimension relates to patients on treatment who would have survived had they received no treatment. In this circumstance, for each patient, the two possible outcomes under the treatment are death, the patient has been killed by the treatment; or survival, the treatment has no effect. It then becomes possible to systematically enumerate all possible hypotheses. If k is the number of spontaneous deaths in the source population varying from 0 to N, then the number of patients who might be saved by the treatment varies from 0 to k and the number of patients who might be killed by the treatment varies from 0 to (N-k). All possible hypotheses are considered to have an equal probability of occurring. It is possible to enumerate all possible configurations of the number of observed deaths in group 1 (from 0 to m1) and group 2 (from 0 to m2) as determined by the randomization process, from ‘0 - 0’ to ‘m1 - m2’. This model reproduces a hypergeometric distribution which describes a discrete probability distribution of the number of successes in a sequence of draws from a finite population without replacement in two dependent populations.

The principle is illustrated by an example of a source population of six patients among whom three would die without treatment (patients 1, 2, and 3) and two would die under the treatment (patients 2 and 4). In other words, the treatment is capable of saving two patients, patients 1 and 3, and of killing one patient, patient 4, it has no effect in patients 2, 5 and 6. Overall, the treatment actually saves one patient on average if given to all patients of the source population. As a consequence of the randomization process, the numbers of patients with each outcome in each group will depend not only on their group allocation, but also on their individual response to the treatment. Based on this example, Table 1 discloses all possible pairs of samples of three patients among six and the number of deaths in group 1 (no treatment) and group 2 (on treatment) for each of them. Table 2 summarizes the number of samples for each possible configuration of 'group 1 – group 2'. Finally, Table 3 (a table of hypotheses and configurations) enumerates in horizontal lines all possible logical hypotheses and in vertical lines all possible configurations. For greater populations, the number of samples corresponding to a given hypothesis and a given configuration was calculated using a specific software according to a general formula (See Appendix).

Table 1: Details of a group of six patients whose outcome is predetermined, independent of their group allocation. Patients 1, 2, and 3 would die without treatment; patients 2 and 4 would die under the treatment. The table discloses all possible outcomes (number of patients who died) and the possible configurations according to the different possible allocations to no treatment (group 1) or to treatment (group 2).

	Group 1	Group 2	Deaths group 1	Deaths group 2
1	1 2 3	4 5 6	3	1
2	1 2 4	3 5 6	2	0
3	1 2 5	3 4 6	2	1
4	1 2 6	3 4 5	2	1
5	1 3 4	2 5 6	2	1
6	1 3 5	2 4 6	2	2
7	1 3 6	2 4 5	2	2
8	1 4 5	2 3 6	1	1
9	1 4 6	2 3 5	1	1
10	1 5 6	2 3 4	1	2
11	2 3 4	1 5 6	2	0
12	2 3 5	1 4 6	2	1
13	2 3 6	1 4 5	2	1
14	2 4 5	1 3 6	1	0
15	2 4 6	1 3 5	1	0
16	2 5 6	1 3 4	1	1
17	3 4 5	1 2 6	1	1
18	3 4 6	1 2 5	1	1
19	3 5 6	1 2 4	1	2
20	4 5 6	1 2 3	0	1

Table 2: Number of samples for each pair of observed deaths in groups 1 and 2. Same example as Table 1

Number of deaths	Number of paired samples
0 – 1	1
1 – 0	2
1 – 1	5
1 – 2	2
2 – 0	2
2 – 1	5
2 – 2	2
3 – 1	1

Table 3: Table of hypotheses and configurations for the example of a source population of six patients. Cells contain the number of calculated samples for each hypothesis (one hypothesis per line) and for each possible configuration of deaths observed in, respectively, group 1 and group 2 (upper row). k: number of deaths without treatment, i.e. of spontaneous deaths. Killed: number of deaths under treatment of patients who would have survived without it. Saved: number of patients surviving under treatment who would have died without it. Eff: Saved minus Killed, the average effect. All hypotheses are enumerated for k from 0 to 6 and for each possible individual effect; saved from 0 to k, killed from 0 to N-k.

	k	Killed	Saved	Eff	00	01	02	03	10	11	12	13	20	21	22	23	30	31	32	33	
1	0	0	0	0	20																20
2	0	1	0	-1	10	10															20
3	0	2	0	-2	4	12	4														20
4	0	3	0	-3	1	9	9	1													20
5	0	4	0	-4		4	12	4													20
6	0	5	0	-5			10	10													20
7	0	6	0	-6				20													20
8	1	0	0	0		10			10												20
9	1	0	1	1	10				10												20
10	1	1	0	-1		6	4		4	6											20
11	1	1	1	0	6	4			4	6											20
12	1	2	0	-2	3	6	1		1	6	3										20
13	1	2	1	-1	6	3	6	1	1	6	3										20
14	1	3	0	-3		6	3			3	6	1									20
15	1	3	1	-2	1	1	6	3		3	6	1									20
16	1	4	0	-4			6				6	4									20
17	1	4	1	-3		4	4	6			6	4									20
18	1	5	0	-5			10					10									20
19	1	5	1	-4				10				10									20
20	2	0	0	0			4			12			4								20
21	2	0	1	1		4			6	6			4								20
22	2	0	2	2	4				12				4								20
23	2	1	0	-1			3	1		6	6		1	3							20
24	2	1	1	0		3	1		3	6	3		1	3							20
25	2	1	2	1	3	1			6	6			1	3							20
26	2	2	0	-2			2	2		2	8	2		2	2						20
27	2	2	1	-1		2	2		1	5	5	1		2	2						20
28	2	2	2	0	2	2			2	8	2			2	2						20

29	2	3	0	-3			1	3			6	6			3	1					20
30	2	3	1	-2		1	3			3	6	3			3	1					20
31	2	3	2	-1	1	3				6	6				3	1					20
32	2	4	0	-4				4				12				4					20
33	2	4	1	-3			4				6	6				4					20
34	2	4	2	-2		4					12					4					20
35	3	0	0	0				1			9			9			1				20
36	3	0	1	1			1			6	3		3	6			1				20
37	3	0	2	2		1			3	6			6	3			1				20
38	3	0	3	3	1				9				9				1				20
39	3	1	0	-1				1			6	3		3	6			1			20
40	3	1	1	0			1			4	4	1	1	4	4			1			20
41	3	1	2	1		1			2	5	2		2	5	2			1			20
42	3	1	3	2	1				6	3			3	6				1			20
43	3	2	0	-2				1			3	6			6	3			1		20
44	3	2	1	-1			1			2	5	2		2	5	2			1		20
45	3	2	2	0		1			1	4	4			4	4	1			1		20
46	3	2	3	1	1				3	6				6	3				1		20
47	3	3	0	-3				1				9				9				1	20
48	3	3	1	-2			1				6	3			3	6				1	20
49	3	3	2	-1		1				3	6				6	3				1	20
50	3	3	3	0	1					9					9					1	20
51	4	0	0	0								4			12			4			20
52	4	0	1	1							3	1		6	6		1	3			20
53	4	0	2	2						2	2		2	8	2		2	2			20
54	4	0	3	3					1	3			6	6			3	1			20
55	4	0	4	4					4				12				4				20
56	4	1	0	-1								4			6	6			4		20
57	4	1	1	0							3	1		3	6	3		1	3		20
58	4	1	2	1						2	2		1	5	5	1		2	2		20
59	4	1	3	2					1	3			3	6	3			3	1		20
60	4	1	4	3					4				6	6				4			20
61	4	2	0	-2								4				12				4	20
62	4	2	1	-1							3	1			6	6			1	3	20
63	4	2	2	0						2	2			2	8	2			2	2	20
64	4	2	3	1					1	3				6	6				3	1	20
65	4	2	4	2					4					12					4		20
66	5	0	0	0												10			10		20
67	5	0	1	1											6	4		4	6		20
68	5	0	2	2										3	6	1	1	6	3		20
69	5	0	3	3									1	6	3		3	6	1		20
70	5	0	4	4									4	6			6	4			20
71	5	0	5	5									10				10				20
72	5	1	0	-1												10				10	20
73	5	1	1	0											6	4			4	6	20
74	5	1	2	1										3	6	1		1	6	3	20

75	5	1	3	2									1	6	3			3	6	1	20
76	5	1	4	3									4	6				6	4		20
77	5	1	5	4									10					10			20
78	6	0	0	0																20	20
79	6	0	1	1															10	10	20
80	6	0	2	2														4	12	4	20
81	6	0	3	3													1	9	9	1	20
82	6	0	4	4													4	12	4		20
83	6	0	5	5													10	10			20
84	6	0	6	6													20				20
					69	99	99	69	99	153	153	99	99	153	153	99	69	99	99	69	1680

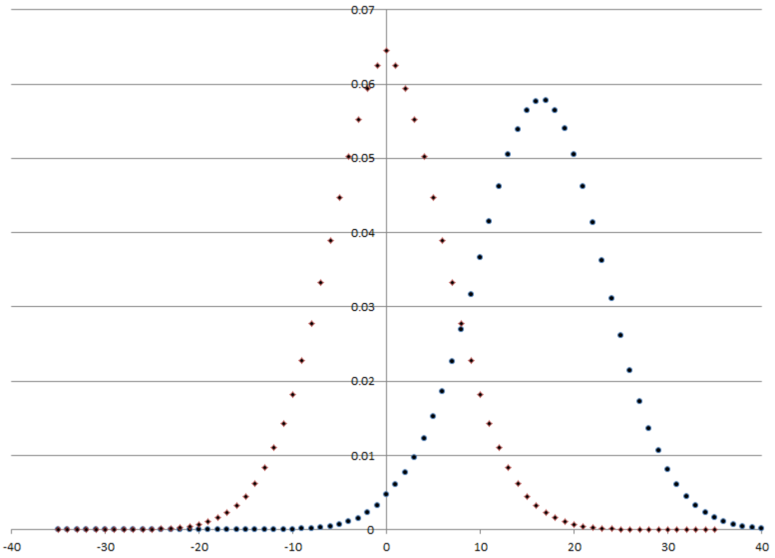
Results

Table 3 allows a general analysis illustrated by the example of a small population of six patients. An example with a greater number of patients cannot be easily displayed since the number of hypotheses and configurations steeply increases with the size of the population.

Horizontally, each line corresponds to one hypothesis of the number of spontaneous deaths in the source population, from 0 to 6, and for each potential effect of the treatment. For each hypothesis or each group of hypotheses, the probability of any configuration or group of configurations can be calculated by dividing the numbers of samples in the corresponding cell or cells by the sum of the corresponding line or lines. For instance, line 41 reproduces the situation of the example displayed in Tables 1 and 2, with a hypothesis of three spontaneous deaths, one patient who would have otherwise survived is killed by the treatment, two patients who would have died without treatment are saved. Overall, the treatment saves one patient. The table of hypotheses and configurations allows a deeper insight into the meaning of the null hypothesis. Based on the example of an observed configuration of three deaths in group 1 and two deaths in group 2, a '3 – 2' configuration, and under the hypothesis that the treatment is ineffective, i.e. that it neither saves nor kills any patient, the number of deaths in the source population of 6 is 3+2 = 5. This situation is displayed on line 66 where only configurations '2 - 3' and '3 - 2' are compatible, each having a 50% probability of occurring. One can consequently calculate the probability of a '3 – 2' configuration under the null hypothesis as 10/(10+10) = 50%. In a population of greater size, the null hypothesis line would show the distribution of possible configurations under the null hypothesis, and allow the calculation of the p-value in a finite population. There are, however, other potential null hypotheses within the same configuration, which refer to a null effect on average; the number of patients saved being offset by the number of patients killed. Lines 45, 57, 63, and 73 illustrate such situations. Moreover, there can be several other true null hypotheses depending on the draw, i.e. on the observed configurations for a same source population. For example, if a '2 – 1' configuration is drawn, then the null hypothesis would be the one displayed in line 35, with a 45% (9/20) probability to observe a '2 – 1' configuration. Obviously, the null hypothesis is not an invariant of a given source population.

Vertically, Table 3 displays all possible configurations of the number of observed deaths in group 1 and group 2 and the number of the corresponding samples for each hypothesis. Given any observed configuration, it becomes possible to calculate the probability of any given hypothesis or group of hypotheses by dividing the number of samples in the cell or cells common to, respectively, this or these hypotheses and to the precise observed configuration by the total number of samples of the observed configuration in the bottom line of the corresponding column. It thus becomes possible to calculate the exact probability of any hypothesis, be it H0 or any H1 or groups of H1, given the data, i.e. given the observed configuration. For example, based on the '3 – 2' configuration in Table 3, the probability of H0 can be calculated as 1/24 or 4.2% since 24 hypotheses are compatible with this configuration. Furthermore, one can calculate the probability of H0 given the '3 – 2' configuration as 10/99 or 10.1%. For each cell, but not for each group of cells, it appears that the probabilities can be calculated according to the Bayes formula from the probability of the corresponding data and the corresponding hypothesis. Similarly, one can calculate the probability of each hypothesis. For instance, by adding all +1 effect samples in the '3 – 2' configuration column and dividing the sum by the total number of samples for this configuration, one obtains the probability of the treatment saving one patient. In this example and a '3 – 2' configuration, lines 46, 58, 64, 67, 74, and 79 show a +1 efficacy with 1+2+3+6+6+10= 28 compatible samples, thus giving a 28/99=28.3% probability. This can be computed from -6 to +6, giving for each configuration the distribution of the hypotheses, i.e. the distribution of effect. Using the same logic, one can calculate that given a '3 – 2' configuration, the probability on average of no effect is 20.2%, the probabilities of saving one, two, three, or four patients are, respectively, 28.3%, 26.3%, 14.1%, and 4.0%, the probabilities of killing one, two, three, or four patients are respectively 6.1%, 1.0%, 0, and 0%. Using the same mathematical methods, examples with a population of greater size are shown in the figure. They refer to a population of 200 patients divided into 2 groups of 100. The number of paired groups is the combination of 100 patients among 200, or 9.10^{58}, the number of calculated hypotheses is 1,373,701. The curves display the distribution of probabilities from killing 40 patients on the left to saving 40 patients on the right. The curve on the left displays the distribution of probabilities when the number of observed deaths is 10 in both groups. The number of null hypotheses on average is 1900. The probability that the treatment on average kills at least one patient is 46.8%, that it saves at least one patient is 46.8%, the probability of no effect is 6.5%. The curve on the right displays the distribution of probabilities when the number of observed deaths is 15 in the untreated group and 6 in the treated group. The number of null hypotheses on average is 1861. The probability that the treatment globally kills at least one patient is 1.0%, that it saves at least one patient is 98.5%. The probability that the treatment is ineffective is 0.5%.

Figure: Example of a population of 200 patients randomly divided into two groups of 100. The curves display the distribution of the probabilities from killing 40 patients to saving 40 patients. The curve on the left is obtained when the number of observed deaths is 10 in both groups. The probability that the treatment on average kills at least one patient is 46.8%, that it saves at least one patient is 46.8% and the probability of no effect is 6.5%. The curve on the right is obtained when the number of observed deaths is 15 in the untreated group and 6 in the treated group. The probability that the treatment globally kills at least one patient is 1.0%, that it saves at least one patient is 98.5%. The probability of no effect is 0.5%.



One point deserves further attention. In the table of hypotheses and configurations, the distribution of hypotheses for the same number of deaths observed in a group depends on the group. For example, if the number of deaths observed in group 1 is one, the highest probabilities in configurations '1 – 0', '1 – 1', '1 – 2', and '1 – 3' are concentrated on the hypothesis of two spontaneous deaths in the source population. Conversely, the highest probabilities of observing '0 – 1', '1 – 1', '2 – 1', and '3 – 1' in group 2 are equally spread among the hypotheses zero, two, four, and six of spontaneous deaths.

Discussion

Using very simple principles of calculation, it is shown here that when data are analysed from a finite population, the probability of any hypothesis or set of hypotheses simply derives from the process of listing all possible and logical hypotheses given the data. Since the source population is finite, the 2 groups are dependent on each other and there is a finite number of distinct dependent paired groups. The table of hypotheses and configurations helps analyse the whole process and calculate any desired exact probability. By generating a probability of effectiveness, ineffectiveness, or harm of an investigational treatment given the observed data, the model described here provides a continuous set of probabilities which is closer to the uncertainty of treatment effect than the 0.05 cut off p-value traditionally and arbitrarily associated with the null hypothesis testing procedure. The examples are particularly persuasive. They show that, for a given result, one can interpret very precisely the results of an experiment in terms of probability of effect. The curve on the left of the figure is of particular interest. It shows that similar results in the untreated and treated groups cannot be formally interpreted as no effect but rather as a similar probability of harm and effectiveness, with a small probability of no effect. One may question the ability and the legitimacy of a model based on probabilities calculated in a finite population. However, in the absence of any mathematical constraint, there is no reason not to explore data in this way. Ronald A. Fischer himself proposed a similar method in an experiment aimed at testing if a lady could discriminate whether tea or milk was added first to a cup[10], although he did not propose to explore as many hypotheses. One can add that, in the real world, studied samples are not drawn from well identified target populations of infinite size but rather from the same finite population. As it can be seen in the table of hypotheses and configurations, in the model of a finite population, hypotheses and data are internal to the source population and bound to each other. Therefore, there is no reason to infer the null hypothesis from external or a priori knowledge. However, a principle similar to the principle of conditional probabilities can be applied, not only at the level of the null hypothesis, but also at the level of any other hypothesis or group of hypotheses by varying their weight, i.e. their probability. For example, one can compute that some a priori categories of patients, men or women, aged or young, or any others, and for a predefined amount, can be saved, killed or unaffected by the treatment.

This model offers a better understanding of the flaws in the interpretation of the traditional null hypothesis significance testing. An important consideration concerns the null hypothesis. It is frequently described as a situation of treatment ineffectiveness without any further precision. The table of hypotheses and configurations shows that a considerable number of null hypotheses on average can be defined for the same configuration of data, that the traditional approach ignores and implicitly incorporates among the H1 hypotheses. However, these null hypotheses comply with the usual calculation of the p-value neither by the z method nor by the $\chi 2$ method which are based on a one-dimensional approximately normal distribution applied to a single population composed of the sum of the 2 samples defined by the randomization process. As a matter of fact, the real null hypothesis is the only hypothesis which allows the calculations of a probability in a single dimension since both groups are then drawn from a single, well identified, source population. One can suggest that this is the reason why Ronald A. Fischer promoted such a calculation and did not suggest any calculation of probability of hypotheses. More specifically, this model can help reproduce the calculation of the p-value in a finite population by computing the number of samples on the null hypothesis line. It also shows that the null hypothesis is not an invariant for a given source population; it necessarily varies from one draw to the other, and with the configuration, unless the treatment is really ineffective, which cannot be known from the data. The same reasoning applies to any other hypothesis, for example to any inferiority hypothesis since several hypotheses are compatible with a same effect. For example, in a situation like Table 3, lines 4, 14, 17, 29, 33, and 47 show different hypotheses for a -3 effect. Obviously, the traditional null hypothesis analyses data on a specific line of the table of data and configurations, neglecting important information that can be extracted from vertical lines which are more consistent with the real question of the probability of hypotheses.

Importantly, the table of hypotheses and configurations helps understand the notions of point estimate and confidence intervals. The point estimate corresponds to the cell of highest probability in the corresponding column of the observed data, while the distribution of the samples illustrates the notion of confidence intervals. It can be easily seen that the confidence intervals reflect the probability of hypotheses given any observed configuration whereas the p-value actually analyzes data horizontally and infers data rather than hypotheses. However, there is a major flaw in the calculation of the confidence intervals in the treated group with the traditional model. As it can be judged from the table of hypotheses and configurations, the distribution of the hypotheses, and consequently the confidence intervals, are not the same in group 1 and in group 2 for the same number of deaths. This seems in contradiction with the central limit theorem, which implies that a same proportion observed in different samples drawn from a same source population should be associated with the same inferred point estimate and confidence interval. However, it should be noted that the untreated and the treated groups cannot be analysed similarly. Indeed, all samples of patients left untreated can be considered drawn from a source population of patients which does not vary, in which the number of events is independent of the draw. On the contrary, and as long as the treatment alters the outcome of at least one patient if given to all patients of the source population, the samples of treated patients refer to a source population which is not constant, in which the number of events vary from one draw to the other, according to the variable proportions of patients with predefined individual responses to the treatment. H1 is not a single hypothesis but a collection of a considerable number, not only of hypotheses but also of compositions of the source population which vary from one draw to the other. The treated group should obviously be analysed as tri-dimensional with a source population whose composition varies from one draw to the other. Consequently, the central limit theorem cannot apply similarly to the placebo group and to the treated group; so that the traditional formulas used to calculate the point estimate and the confidence interval of a sample drawn from a constant source population cannot be used when samples are drawn from a population of treated patients. As a consequence, the risk ratios and their distribution as traditionally calculated cannot be used as a substitute for the p-value. This probably explains why it has never been possible to calculate a probability of hypothesis given the data with the traditional null hypothesis significance testing procedure. For the same reasons, the notion of non-inferiority threshold cannot be well defined using the traditional approach since it corresponds to varying configurations of the treated group from one draw to the other.

Conclusion

This paper proposes a new way of considering probabilities in randomized studies based on a finite source population rather than on an infinite fictive population, and on an exact probability of effect given the actually observed data rather than on a probability to observe observed and unobserved data given the hypothesis that the data reflect the absence of individual effect. The table of hypotheses and configurations allows a deeper insight into the way data and hypotheses arrange, hence a better understanding of the meaning of different traditional, but generally poorly understood concepts. By giving an exact probability of effect, this model provides a continuous set of probabilities which is closer to the uncertainty of treatment effect than the traditional 0.05 cut off p-value. In any case, it should be remembered that a probability inferred from any population is only a representation of the reality of this particular population, particularly of what would happen if the same source population could be analysed with and without treatment. Consequently, it cannot be easily generalized to other populations without some arbitrary external assumptions that require explicit substantiation.

Finally, from these considerations, this paper helps explain some of the reasons of the recent statement of the American Statistical Association on statistical significance and p-values, that insisted on the fact that "the p-value was never intended to be a substitute for scientific reasoning".[11]

Acknowledgements

I would like to thank Denis Auroux, University of California, who devised a general formula for the calculation of the number of samples for a given hypothesis and configuration

References

1/ Nuzzo R. P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. Nature. 2014;506: 150-152. doi: 10.1038/506150a

2/ Rozeboom WW. The fallacy of the null hypothesis significance test. Psychological Bulletin. 1960;57: 416-428

3/ Cohen J. The earth is round (p<.05). American psychologist. 1994;49: 997-1003

4/ Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J 1986;292: 746–750

5/ Bland JM, Altman DG. The odds ratio. BMJ. 2000;320:1468

6/ Goodman SN. Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy. Ann. Internal Med. 1999;130: 995–1004

7/ Trafimow D, Marks M. Editorial, Basic and Applied Social Psychology. 2015;37: 1-2, Doi:10.1080/01973533.2015.1012991

8/ Bailar III JC, Mosteller F. Medical Uses of statistics. NEJM Books. Boston, Massachusetts. 1992

9/ Diamond GA, Forrester J.S. Clinical trials and statistical verdicts: probable grounds for appeal. Annals of intern Med 1983;98: 385-394

10/ Fisher, RA. The Design of Experiments. Second edition. Oliver and Boyd, Edinburgh, London. 1937

11/ Ronald L. Wasserstein & Nicole A. Lazar (2016): The ASA's statement on p-values: context, process, and purpose, The American Statistician, DOI:10.1080/00031305.2016.1154108

Appendix

Computation of the number of possible combinations for each hypothesis and each configuration for a source population randomly divided into two groups. Let’s define 0: the patient is alive, x: the patient dies, n: size of the source population, m1: size of group 1 (untreated patients), m2: size of group 2 (treated patients), p: number of patients who will die spontaneously and who will not be saved by the treatment (x-x), q: number of patients who will die spontaneously and who will be saved by the treatment (x-0), r: patients who will be alive without treatment and who will be killed by the treatment (0-x), k: number of deaths in the source population if all patients are left untreated (k = p + q), l: number of deaths in the source population if all patients are treated (l = p + r), v: number of deaths observed in group 1, w: number of deaths observed in group 2, i: number of x-x among the m1 patients in group 1. The number of samples for a given value of i can be calculated by observing that group 1 consists of i patients with an x-x response, v-i patients with an x-0 response, l-w-i patients with an 0-x response, and m1-v-l+w+i patients with an 0-0 response. The number of samples comprising i patients among the p number of patients with an x-x response is C(p,i). The number of samples comprising v-i patients among the q number of patients with an x-0 response is C(q,v-i). The number of samples comprising l-w-i patients among the r patients with an 0-x response is C(r,l-w-i). The number of samples comprising m1-v-l+w+i patients with a 0-0 response is C(n-p-q-r, m1-v-l+w+i). For a given i value, the number of samples is:

$$C(p,i)*C(q,v-i)*C(r,l-w-i)*C(n-p-q-r,m1-v-l+w+i)$$

Finally, the number of samples which give a v-w response is the sum of the samples as calculated above from i = max(0, v-q, l-w-r, v+l-w-m1) to i = min(p, v, l-w, v+l-w-m1+n-p-q-r)