A Generalization of the «Lady-Tasting-Tea» Procedure to Link Qualitative and Quantitative Approaches in Psychiatric Research

Bruno Falissard[1,2,3,*], Daniel Milman[4,#] and David Cohen[4,5,#]

[1]*Inserm, U669, Paris, France*

[2]*Univ Paris-Sud and Univ Paris Descartes, UMR-S0669, Paris, France*

[3]*AP-HP, Hôpital Paul Brousse, Département de Santé Publique, Villejuif, France*

[4]*AP-HP, Service de Psychiatrie de L'enfant et de L'adolescent, Groupe Hospitalier Pitié-Salpêtrière, Paris, France*

[5]*CNRS UMR 8189 Psychologie et Neurosciences Cognitives et Université Paris VI, Paris, France*

Abstract: In Fisher's "The Design of Experiments", a trial was designed to test a lady's claim to be able to discriminate whether the milk or the tea was added first to a cup. In this trial, eight cups are poured, four with milk first and four with tea first. They are then presented in random order to a subject who has to divide them into two sets of 4, according his/her belief about the "treatment" received. The present paper generalizes this design so that a hypothesis concerning the existence of two sub groups in a set of psychiatric patient records (whether written, audiotaped or videotaped) can be tested rigorously from a statistical point of view. Tables are proposed to enable power and sample size calculations. A real example is presented; it shows that psycho-dynamically oriented professionals are able to discriminate seven healthy adults who have experienced a sibling's cancer during childhood or adolescence from seven matched controls. This method is particularly suited to small sample studies that explore elusive clinical hypotheses traditionally tackled with qualitative methodologies.

Keywords: Randomization, power calculation, permutation test, clinical records.

INTRODUCTION

It is perhaps in psychology that randomization was used in experiment for the first time [1, 2]. This was in 1884; C.S. Pierce and his student Joseph Jastrow were the experimenters; their objective was to refute a hypothesis made by Gustave Fechner about the existence, for each sense, of a nonzero threshold of intensity below which two sensations cannot be distinguished [3]. The experiment was based on the repeated presentation of a pair of weights; the subjects had to determine whether or not the first weight was the heavier one. To facilitate the interpretation of results, the authors decided to randomize the order of presentation of the two weights in batches of 25.

Curiously, the practice of randomization does not seem to have spread after this seminal publication; it is only after the work by Sir Ronald Aylmer Fisher that it came to be considered as a standard [1]. Even if Fisher had a professional interest in genetics and agriculture, the place he gave to psychological experiment also appears considerable. The second chapter of R. A. Fisher's book *The Design of Experiments* [4] is entitled "The Principle of Experimentation, Illustrated by a Psycho-physical Experiment". This chapter begins with an anecdote which occurred at a university tea party in the late 1920s [5]:

> "A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup. We will consider the problem of designing an experiment by means of which the assertion can be tested. Our experiment consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject in random order. [...] Her task is to divide the cups into two sets of 4, agreeing, if possible, with the treatments received."

Fisher insists on the importance of randomization. This process is "the only point in the experimental procedure in which the laws of chance […] have been explicitly introduced". This is essential from a statistical point of view since "the simple precaution of randomisation will suffice to guarantee the validity of the test of significance".

In the area of psychiatric research, a procedure of this type can be considered as a precursor of single-subject randomised experimental designs, which are

*Address correspondence to this author at the INSERM U669, Maison de Solenn, 97 Bd du Port Royal, 7569 Paris cedex 14, France; Tel: (33) 1 58412850; Fax: (33) 1 58412946; E-mail: bruno.falissard@gmail.com

[#]CO-Authors E-mail: daniel.milman75@orange.fr, david.cohen@psl.aphp.fr

E-ISSN: 1929-6029/13

sometimes used in the evaluation of behavioral therapies [6, 7]. If it is often difficult, for theoretical and practical reasons, to give a single subject a random succession of two "treatments" as in the lady-tasting-tea experiment, it is possible to achieve randomization by choosing at random the point at which one of the treatments will succeed to the other. A design of this sort is known as an “AB” design [7] where all A treatments are given first, then all B treatments. This design can be extended to “ABAB” designs or to multiple-baseline AB designs, where a small sample of subjects is studied instead of a single subject [8].

A particular interest of the lady-tasting-tea procedure is that it offers a very simple design to test hypotheses that would be rather difficult to tackle using, for example, psychometric scales or questionnaires. We propose here to develop and adapt it to the field of psychological research. The question of statistical power will be particularly important since low power has been pointed to as a major limitation of single-subject experiments in general [9]. A real example will be presented; it focuses on a question formulated in the field of psychodynamic psychopathology.

METHODS

Objective

To test the hypothesis that two sets of records A and B (written, audiotaped or videotaped interviews) are distinguishable.

Procedure

1. n records (n even) are collected. n/2 records belong to set A and n/2 to set B.

2. The n records are presented in random order to k independent raters (the orders are different from rater to rater). The raters know that half of the records belong to A and the other half to B.

3. The raters are then asked to give their opinion about the likelihood that the records belong to A or B. In practice, for each record a given rater has to choose among 4 propositions: “it is certain that the record belongs to A”, “it is plausible that the record belongs to A”, “it is plausible that the record belongs to B”, “it is certain that the record belongs to B”. The raters can examine the records as long as they want, possibly several times in different orders. The raters work blind to one another's assessments.

4. Ratings are analyzed in the following way: for each record in group A, when a rater has considered that “A is certain” 2 is added to the score, 1 is added if “A is plausible”, 1 is subtracted if “B is plausible” and 2 is subtracted if “B is certain”. The same is applied to each record in group B, but here 2 is added to the score if “B is certain”, and so forth. A total score is then obtained from the responses given by all raters to all records. Finally a statistical test of hypothesis compares this total score to 0, and a permutation test is used ([10] p. 202). It should be noted that since the raters know that half of the records belong to A and the other half to B, the ratings cannot be considered as independent realizations of a random variable, so that the traditional Student t.test or Mann-Whitney test should not be used. In contrast, under the null hypothesis that A and B records are indistinguishable, all permutations of scores obtained for each record are equi-probable. So that a p-value (one-sided) can be estimated as the proportion of permutations of the n records for which the total score is higher than or equal to the total score obtained in the experiment ([10], p. 208). When a two-sided p-value is preferred, some authors suggest doubling the one-sided p-value; this is also the option proposed in the International Conference on Harmonization ICH E9 guidelines [11]. Two-sided p-values will be preferred in the rest of the paper.

Power

From a theoretical point of view, since in a randomization test the outcomes are regarded as fixed quantities and not as random variables with a given distribution the concept of statistical power is in itself questionable. But since this point is discussed by several authors [6], we will not enter into this fundamental debate and will focus on very practical considerations.

We will consider here three alternative hypotheses: one where raters have a sensitivity and a specificity of 0.8 for correctly allocating subjects to groups A and B; one where sensitivity and specificity are equal to 0.7 and one where they are equal to 0.6. It should be noted that when sensitivity and specificity are equal to 1 the raters have perfect discriminating ability, and when sensitivity and specificity are equal to 0.5 their discriminating ability is null, or rather it is comparable to an equiprobable random scoring.

For the case where sensitivity and specificity are equal to 0.7 (for example), data sets are simulated in the following way. For records belonging to group A, the k×n/2 ratings are randomly chosen from the following distribution {"2" with a probability of 0.35, "1" with a probability of 0.35, "-1" with a probability of 0.15 and "-2" with a probability of 0.15}. For records belonging to group B, the k×n/2 ratings correspond to -1 multiplied by a random permutation of the ratings obtained for group A. This procedure guarantees that there will be as many negative as positive ratings, which is what each rater is supposed, by construction, to produce.

For a series of n and k, 7000 random data sets are generated according each alternative hypothesis. The statistical power is estimated as the proportion of random data sets for which the two-sided permutation test at the 5% level rejects the null hypothesis. The number of 7000 guarantees that for a power of 75%, the half span of the 95% confidence interval for the estimated statistical power is equal to 0.01. This precision in the estimated power will be greater for higher power values and lower for smaller power values (up to 0.50). All computations were performed using R software version 2.4.1 [12].

Results are presented in Table **1**. They show that even for a weak alternative hypothesis (sensitivity = specificity = 0.6) some designs can lead to substantial

Table 1: Statistical power of the "testing on randomized templates" procedure for three alternative hypotheses (sensitivity and specificity of raters equal to 0.6 (above), 0.7 (medium), 0.8 (below)), for n records scored by k raters (two-sided test with a type 1 error equal to 0.05). Example: for n=2×7=14 records scored by 3 raters, the statistical power is equal to 0.88 for the alternative hypothesis that the raters discriminate the two groups with a specificity and a sensitivity of 0.8.

	N=2×5	n=2×6	n=2×7	n=2×8	n=2×9	n=2×10	n=2×12	n=2×15	n=2×20
k=1	0.14 0.24 0.40	0.14 0.24 0.41	0.17 0.31 0.52	0.18 0.34 0.59	0.20 0.38 0.63	0.22 0.41 0.68	0.22 0.44 0.73	0.24 0.50 0.82	0.28 0.62 0.90
k=2	0.18 0.36 0.62	0.20 0.41 0.69	0.22 0.45 0.77	0.24 0.49 0.82	0.26 0.54 0.86	0.27 0.59 0.89	0.30 0.67 0.94	0.34 0.74 0.97	0.40 0.85 0.99
k=3	0.20 0.43 0.75	0.23 0.52 0.83	0.25 0.58 0.88	0.28 0.65 0.93	0.30 0.68 0.95	0.33 0.72 0.97	0.37 0.80 0.98	0.43 0.87 >0.99	0.51 0.95 >0.99
k=5	0.26 0.60 0.91	0.29 0.70 0.96	0.34 0.77 0.98	0.27 0.82 0.99	0.40 0.86 0.99	0.43 0.89 >0.99	0.50 0.93 >0.99	0.58 0.97 >0.99	0.67 0.99 >0.99
k=7	0.30 0.72 0.97	0.38 0.81 0.99	0.42 0.87 >0.99	0.46 0.91 >0.99	0.50 0.94 >0.99	0.54 0.96 >0.99	0.61 0.98 >0.99	0.68 >0.99 >0.99	0.79 >0.99 >0.99
k=10	0.40 0.84 0.99	0.46 0.91 >0.99	0.52 0.95 >0.99	0.57 0.97 >0.99	0.62 0.98 >0.99	0.65 0.99 >0.99	0.73 >0.99 >0.99	0.80 >0.99 >0.99	0.90 >0.99 >0.99
k=14	0.48 0.93 >0.99	0.57 0.97 >0.99	0.63 0.99 >0.99	0.69 0.99 >0.99	0.74 >0.99 >0.99	0.78 >0.99 >0.99	0.93 >0.99 >0.99	0.91 >0.99 >0.99	0.96 >0.99 >0.99
k=20	0.59 0.98 >0.99	0.69 >0.99 >0.99	0.76 >0.99 >0.99	0.81 >0.99 >0.99	0.84 >0.99 >0.99	0.88 >0.99 >0.99	0.93 >0.99 >0.99	0.96 >0.99 >0.99	0.99 >0.99 >0.99
k=30	0.75 >0.99 >0.99	0.83 >0.99 >0.99	0.88 >0.99 >0.99	0.92 >0.99 >0.99	0.94 >0.99 >0.99	0.97 >0.99 >0.99	0.98 >0.99 >0.99	0.99 >0.99 >0.99	>0.99 >0.99 >0.99

statistical power (for example power = 0.88 for 20 raters and 2×9 records).

EXAMPLE

In a recent study the authors were interested in the long term psychological outcome of siblings of children with cancer. They hypothesized: 1/ that cancer would have a lasting traumatic effect on the siblings of children with the disease and 2/ that personal psychodynamic experience enhances the ability to discriminate between these siblings and controls [13].

To test these two hypotheses, seven healthy adults who had experienced a sibling's cancer during childhood or adolescence and seven matched controls were asked to give a 5-minute spontaneous free-association speech sample following specific instructions designed to activate a buffer zone between fantasy and reality. Three psycho-dynamically oriented professionals and three non-experienced professionals were randomly shown the videos and asked to classify them blind according to possible traumatic history (i.e. being siblings of children with cancer) using a -2/-1/1/2 response pattern.

Psycho-dynamically oriented professionals (1) were able to recognize, beyond levels attributable to chance, healthy adults who had experienced a sibling's cancer, without explicit knowledge of this history (p=.002); and (2) discriminated better than inexperienced professionals (p=.003), who were unable to make such decisions beyond levels attributable to chance (p=.68). Of course, these results should be discussed more in depth, but this is not the scope of the present paper which focuses mainly on methodological considerations.

The R script used to perform these analyses is presented in the appendix.

DISCUSSION

The World Psychiatric Association (WPA) President's workplan for 2005-2008 was centered on "psychiatry for the person" which "promotes a contextualized and integrative perspective, seeking to articulate science and humanism in the service of the wholeness of the person who consults" [14].

However, when adapted to research, this laudable ambition to integrate science and humanism raises methodological issues, and if qualitative research can be interesting in this perspective, there are several drawbacks when it is used alone, among which the questionable generalizability of its results and their possible refutability [15]. The procedure proposed here thus appears as an example of a methodology that can enable the statistical testing of hypotheses that could be difficult to tackle using traditional tools like scales, questionnaires or cognitive tests.

Of course, the randomization of experimental materials is not a new idea in psychiatric or in psychological research [16]. We have even seen, in the introduction, that it goes back to a very early proposition. But, to our knowledge, there is no formal presentation of a method that enables power and sample size calculations.

Like all methods, the proposed procedure has advantages and limitations. Among the advantages: it can deal with (very) small samples of records; the methodology is clear-cut, with straightforward statistical inferences; the possibilities for applications are wide. Two protocols based on this methodology are presently underway. The first is a randomized controlled trial among children with an Attention Deficit with Hyperactivity Disorder (ADHD). This trial compares "treatment as usual" (psychological treatment and medication) to "treatment as usual" plus psychodynamic therapy. The objective is to show that after one year of treatment, the two groups are distinguishable on the basis of a general 5-minutes videotaped interview. The second protocol concerns psychological autopsies of state employees. The procedure will be used to show that records obtained from the last days of the lives of state employees who died by accident are distinguishable from the records obtained from the last days of the lives of state employees who committed suicide.

One limitation of the procedure presented here is that although it is based on a randomized procedure, it can be prone to problems of interpretation. Indeed, the randomization process makes it possible to say with a 5% error that records from groups A and B are distinguishable; but it does not say *why* these records are distinguishable. In the example presented above, if a subject from group A (siblings of children with cancer) says explicitly during the interview that he/she had this particular history, the experiment is of course no longer valid.

This limitation can be tackled in several ways. First, the question introducing the interview needs to be as general as possible; in the present situation it was "*[…]*

you could talk about the importance you assign to your dreams but also about how you relate to art, painting, music, or sculpture, and about how much room you give to all these feelings in your everyday life. […]”. In addition, an experimenter needs to verify that the explicit content (verbatim) of the interview is not informative. Finally, if a second group of raters is used and if the hypothesis tested concerns a difference between the two groups of raters, then the experiment is less limited by the explicit content problem (since it is present in a same way in the two situations). It should be noted that a limitation of this kind is present in many other randomized experiments, for example in randomized controlled trials on medications when there is no blinding, or when the blinding is likely to be invalidated due to the presence of specific side-effects.

In all events, if there is a need to generate hypotheses about the process which enabled discrimination of subjects from groups A and B, this can be done through qualitative interviews with the raters and a content analysis of these interviews.

In conclusion, provided that 1) cases and controls are selected in a careful manner; 2) raters are experienced and motivated by the study so that they will take all the time necessary to provide carefully thought-out ratings; 3) the starting question addressed to the subjects is worded in a way that does not generate an explicit content bias; 4) written records are drafted by a person blinded to group membership (when written records are used); 5) the duration of the audiotaped or videotaped interviews is optimal (5 minutes seems to be sufficiently informative and acceptable in practice [17]); then the procedure can be powerful in producing results with a level of evidence that is difficult to achieve using alternative methodologies.

APPENDIX: R SCRIPT FOR THE STATISTICAL OF ANALYSES OF THE EXAMPLE

```
# data for the 3 psychoanalysts

data.1 <- matrix(nrow=14,ncol=3)

# ratings of first rater

data.1[,1] <- c(2,-2,1,1,-2,2,1,-2,2,2,-2,-2,-1,-1)

# ratings of second rater

data.1[,2] <- c(1,-2,2,-1,-1,2,2,-1,2,2,-1,1,-2,-1)

# ratings of third rater

data.1[,3] <- c(1,2,1,-1,-1,1,2,-1,2,2,-1,-1,-1,-1)

# computation of scores (summation of ratings by records)

x1 <- apply(data.1,1,sum)

# unblinding of the records: 1 for records belonging to group A, 0 for group B

m <- c(1,0,1,0,0,1,1,0,1,1,0,0,0,1)

library(coin)

# p value of the permutation test

pvalue(oneway_test(x1~factor(m),distribution="exact",alternative="two.sided"))

#

# data for the 3 other clinicians

data.2 <- matrix(nrow=14,ncol=3)

data.2[,1] <- c(-2,2,-2,1,-2,2,-1,2,2,2,-2,-2,2,-1)

data.2[,2] <- c(-1,2,1,-1,2,-1,1,-1,2,-1,2,-1,2,-1)

data.2[,3] <- c(1,2,1,-1,-1,1,2,-1,2,2,-1,-1,-1,-1)

x2 <- apply(data.2,1,sum)

pvalue(oneway_test(x2~factor(m),distribution="exact",alternative="two.sided"))

# are psychoanalysts "better" than controls?

x3 <- x1-x2

pvalue(oneway_test(x3~factor(m),distribution="exact",alternative="two.sided"))
```

REFERENCES

[1] Hall NS. R. A. Fisher and his advocacy of randomization. J Hist Biol 2007; 40(2): 295-25. http://dx.doi.org/10.1007/s10739-006-9119-z

[2] Stigler SM. Mathematical statistics in the early states. Ann Statist 1978; 6(2): 239-65. http://dx.doi.org/10.1214/aos/1176344123

[3] Peirce C, Jastrow J. On small differences of sensations. Memoirs Natl Acad Sci 1884 1885; 3: 75-83.

[4] Fisher SRA. The design of experiments. Hafner 1951.

[5] Salsburg D. The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century. Holt Paperbacks 2002.

[6] Edgington ES. Randomization Tests 3e. Marcel Dekker 1995.

[7] Edgington ES. Randomized single-subject experimental designs. Behav Res Therapy 1996; 34(7): 567-74. http://dx.doi.org/10.1016/0005-7967(96)00012-5

[8] Marascuilo LA, Busk PL. Combining statistics for multiple-baseline AB and replicated ABAB designs across subjects. Behav Assessm 1988; 10(1): 1-28.

[9] Ferron J, Ware W. Analyzing Single-Case Data: The Power of Randomization Tests. J Exper Educ 1995; 63(2): 167-78. http://dx.doi.org/10.1080/00220973.1995.9943820

[10] Efron B, Tibshirani R. An Introduction to the Bootstrap. Chapman & Hall 1993.

[11] Neuhäuser M. The Choice of α for One-Sided Tests. Drug Informat J 2004; 38: 57-60. http://dx.doi.org/10.1177/009286150403800108

[12] R Development Core Team. R: A language and environment for statistical computing [Internet]. 2004; Available from: http://www.R-project.org

[13] Cohen D, Milman D, Venturyera V, Falissard B. Psychodynamic Experience Enhances Recognition of Hidden Childhood Trauma. PLoS ONE 2011; 6(4): e18470. http://dx.doi.org/10.1371/journal.pone.0018470

[14] Mezzich JE. Institutional consolidation and global impact: towards a psychiatry for the person. World Psychiatry 2006; 5(2): 65-6.

[15] Myers M. Qualitative Research and the Generalizability Question: Standing Firm with Proteus. Qualitative Report 2000; 4(3/4).

[16] Hagemann N, Strauss B, Leißing J. When the Referee Sees Red. Psychol Sci 2008; 19(8): 769-71. http://dx.doi.org/10.1111/j.1467-9280.2008.02155.x

[17] Kadouri A, Corruble E, Falissard B. The improved Clinical Global Impression Scale (iCGI): development and validation in depression. BMC Psychiatry 2007; 7: 7. http://dx.doi.org/10.1186/1471-244X-7-7

Received on 31-12-2012 Accepted on 05-03-2013 Published on 30-04-2013

http://dx.doi.org/10.6000/1929-6029.2013.02.02.02