"Lady Tasting Tea"[1]

R. A. Fisher was one of the founding fathers of modern statistics. One of his early, and perhaps the most famous, experiments was to test an English lady's claim that she could tell whether milk was poured before tea or not. Here's an account of the seemingly trivial event that had the most profound impact on the history of modern statistics, and hence arguably, modern quantitative science (Box 1978).

> Already, quite soon after he had come to Rothamstead, his presence had transformed one commonplace tea time to an historic event. It happened one afternoon when he drew a cup of tea from the urn and offered it to the lady beside him, Dr. B. Muriel Bristol, an algologist. She declined it, stating that she preferred a cup into which the milk had been poured first. "Nonsense," returned Fisher, smiling, "Surely it makes no difference." But she maintained, with emphasis, that of course it did. From just behind, a voice suggested, "Let's test her." It was William Roach who was not long afterward to marry Miss Bristol. Immediately, they embarked on the preliminaries of the experiment, Roach assisting with the cups and exulting that Miss Bristol divined correctly more than enough of those cups into which tea had been poured first to prove her case.
>
> Miss Bristol's personal triumph was never recorded, and perhaps Fisher was not satisfied at that moment with the extempore experimental procedure. One can be sure, however, that even as he conceived and carried out the experiment beside the trestle table, and the onlookers, no doubt, took sides as to its outcome, he was thinking through the questions it raised: How many cups should be used in the test? Should they be paired? In what order should the cups be presented? What should be done about chance variations in the temperature, sweetness, and so on? What conclusion could be drawn from a perfect score or from one with one or more errors?

The real scientific significance of this experiment is in these questions. These are, allowing incidental particulars, the questions one has to consider before designing an experiment. We will look at these questions as pertaining to the "lady tasting tea," but you can imagine how these questions should be adapted to different situations.

- *What should be done about chance variations in the temperature, sweetness, and so on?* Ideally, one would like to make all cups of tea identical except for the order of pouring milk first or tea first. But it is never possible to control all of the ways in which the cups of tea can differ from each other. If we cannot control these variations, then the best we can do–we do mean the "best"– is by randomization.

- *How many cups should be used in the test? Should they be paired? In what order should the cups be presented?* The key idea here is that the number and ordering of the cups should allow a subject ample opportunity to prove his or her abilities and keep a fraud from easily succeeding at correctly discriminating the the order of pouring in all the cups of tea served.

- *What conclusion could be drawn from a perfect score or from one with one or more errors?* If the lady is unable to discriminate between the different orders of pouring, then by guessing alone, it should be highly unlikely for that person to determine correctly which cups are which for all of the cups tested. Similarly, if she indeed possesses some skill at differentiating between the orders of pouring, then it may be unreasonable to require her to make no mistakes so as to distinguish her ability from a pure guesser.

An actual scenario described by Fisher and told by many others as the "lady tasting tea" experiment is as follows.

[1] Adapted from *Stat Labs: Mathematical statistics through applications* by D. Nolan and T. Speed, Springer-Verlag, New York, 2000

- For each cup, we record the order of actual pouring and what the lady says the order is. We can summarize the result by a table like this:

		Order of Actual Pouring: Tea first	Order of Actual Pouring: Milk first	
Lady says	Tea first	a	b	$a+b$
	Milk first	c	d	$c+d$
		$a+c$	$b+d$	n

  Here n is the total number of cups of tea made. The number of cups where tea is poured first is $a+c$ and the lady classifies $a+b$ of them as tea first. Ideally, if she can taste the difference, the counts b and c should be small. On the other hand, if she can't really tell, we would expect a and c to be about the same.

- Suppose now that to test the lady, 8 cups of tea are prepared, 4 tea first, 4 milk first, and she is informed of the design (that there are 4 cups milk first and 4 cups tea first). Suppose also that the cups are presented to her in random order. Her task then is to identify the 4 cups milk first and 4 cups tea first.

  This design fixes the row and column totals in the table above to be 4 each. That is,

$$a+b = a+c = c+d = b+d = 4.$$

  With these constraints, when any one of a, b, c, d is specified, the remaining three are uniquely determined:

$$b = 4-a, \quad c = 4-a, \quad \text{and} \quad d = a$$

  In general, for this design, no matter how many cups (n) are served, the row total $a+b$ will equal $a+c$ because the subject knows how many of the cups are "tea first" (or one kind as supposed to the other). So once a is given, the other three counts are specified.

- We can test the discriminating skill of the lady, if any, by randomizing the order of the cups served. If we take the position that she has no discriminating skill, then the randomization of the order makes the 4 cups chosen by her as tea first equally likely to be any 4 of the 8 cups served. There are $\binom{8}{4} = 70$ (in R, `choose(8,4)`–see also Devore pp. 71–73) possible ways to classify 4 of the 8 cups as tea first. If the subject has no ability to discriminate between two preparations, then by the randomization, each of these 70 ways is equally likely. Only one of 70 ways leads to a completely correct classification. So someone with no discriminating skill has 1/70 chance of making no errors.

- It turns out that, if we assume that she has no discriminating skill, the number of correct classifications of tea first ("a" in the table) has "hypergeometric" probability distribution (see `help(dhyper)` in R or Devore pp. 128–129). There are 5 possibilities: 0, 1, 2, 3, 4 for a and the corresponding probabilities (and R commands for computing the probabilities) are tabulated below.

Number of correct calls	R command	Probability
0	`dhyper(0,4,4,4)`	1/70
1	`dhyper(1,4,4,4)`	16/70
2	`dhyper(2,4,4,4)`	36/70
3	`dhyper(3,4,4,4)`	16/70
4	`dhyper(4,4,4,4)`	1/70

- With these probabilities, we can compute the p-value for the test of the hypothesis that the lady cannot tell between the two preparations. Recall that the p-value is the probability of observing a result as extreme or more extreme than the observed result assuming the null hypothesis. If she makes all correct calls,the p-value is $1/70$ and if she makes one error (3 correct calls) then the p-value is $1/70 + 16/70 \approx 0.24$.

The test described above is known as "Fisher's exact test."

References

Box, J. F. (1978). *R. A. Fisher: The Life of a Scientist*. John Wiley & Sons, Inc., New York.