REVISITING FISHER'S TEA TASTING LADY

DONALD L. BENTLEY
POMONA COLLEGE

TECHNICAL REPORT NO. 91-009

SLAW is supported by a grant from the Exxon Corporation
Statistics in the Liberal Arts Workshop (SLAW) is a group of educators concerned with the teaching of statistics. The workshop was initially funded by the Sloan Foundation. Continuing support has been provided by the Exxon Corporation.

Participants in the summer of 1991 are:

Donald L. Bentley
Pomona College

George W. Cobb
Mt. Holyoke College

Janice Gifford
Mt. Holyoke College

Katherine T. Halvorsen
Smith College

Homer T. Hayslett, Jr.
Colby College

Gudmund Iversen
Swarthmore College

Robin H. Lock
St. Lawrence University

David S. Moore
Purdue University

Thomas L. Moore
Grinnell College

Norean Radke-Sharpe
Bowdoin College

Rosemary Roberts
Bowdoin College

Dexter Whittinghill III
Colby College

Jeffrey Witmer
Oberlin College
INTRODUCTION:

During a recent sabbatical leave in London, my wife and I were befriended by a retired dentist. When she invited us to have high tea at her house in Wimbledon one Sunday afternoon, we quickly accepted the opportunity to experience another British custom. We arrived at the appointed time, and after a short period of friendly conversation in her living room she sat us at an elegant dining table prepared with sandwiches and cakes. We began by passing the sandwiches, and then she asked "Would you care for milk in your tea?"

My mind started racing as I heard those words coming from this attractive, dark complexioned woman. Could it possibly be that she, those many years ago, had been the woman whom Fisher had asked to separate the famous eight cups of tea into those with the milk added first, and those with the tea added first? As the afternoon drew to a close my mind filled with questions. Suppose that this really was the woman. What were the conditions surrounding her original interview with Fisher, and more important, how would she respond if the experiment could be repeated now?
After a restless night I was off early the next morning by underground to Goodge St. Station, then across Tottenham Court Road to Mallet, up Mallet to the Royal Statistical Society Library at London University. After a brief search I located what I had come for: a copy of Fisher's *The Design of Experiments* with the original account of the tea tasting experiment. But as I read the account I felt uneasy, and upon rereading it my gaze came to rest upon the source of that uneasiness. Fisher's instructions to my friend had been as follows:

Our experiment consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgment in a random order. The subject has been told in advance of what the test will consist, namely that she will be asked to taste eight cups, that these shall be presented to her in a random order, that is in an order not determined arbitrarily by human choice, but by the actual manipulation of the physical apparatus used in games of chance, cards, dice, roulettes, etc., or, more expeditiously, from a published collection of random sampling numbers purporting to give the actual results of such manipulation. Her task is to divide the 8 cups into two sets of 4, agreeing, if possible, with the treatments received.

That evening as my wife and I sat in our Putney flat watching the darkness settle on the Thames, I began to fantasize what might happen if Fisher were to visit our friend today in an attempt to duplicate the experiment. Suddenly I found myself again sitting at her table, with Fisher between us. He was instructing her as to the structure of the experiment. There were to be eight cups of tea, four of which would have the milk added first, and four with the tea added first. "No, there must be five of each type." she replied. "You must understand that I am trying to demonstrate to
the F.D.A. that I can distinguish between those with milk, and those with tea added first. And they require that I demonstrate an ability distinguish, beyond guessing, at the p = 0.05 level. The extra cups will provide a bit more power to the test." Fisher reluctantly agreed, and placed the ten cups in front of her in a random order.

As my friend began to taste the various cups, Fisher turned to me with an explanation of what was occurring. "We are assuming that, if this woman has any ability to distinguish between cups, then each cup will be an all or nothing situation. Either she will definitely know whether milk or tea was added first, or she will have no idea. Further, we assume that her ability to distinguish is independent from cup to cup." He then proceeded to explain how she would arrive at her response and the subsequent statistical analysis. In the event that she identified all ten cups correctly we would have the following two by two table for analysis.

<table>
<thead>
<tr>
<th>ACTUAL FIRST</th>
<th>PICK FIRST</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MILK</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>TEA</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>

I quickly replied "Using Fisher's Exact Test for the analysis we would conclude that the chance of her arriving at such an array if she were guessing on each cup is only p = 0.004, well below the p
Fisher seemed pleased with my response.

"Suppose," Fisher continued, "that she can only identify nine of the ten, say five with milk added first and four with tea. What then would be her response?" I pondered this for a moment, and then responded that she would be able to identify the unknown cup as having the tea added first since she would know that there were five of each kind. Since this would provide the correct identification for all ten cups we would have the same table as given above.

"Very good" replied Fisher. "Now suppose that she can only identify eight correctly. What would be the consequences of that?" After a little more thought I suggested the following. There are two possibilities. The first would be that she could identify five of one kind and three of the other. For example she detected which five had the milk first, but only three which had the tea first. By deduction she would know that the unidentified two had tea added first. This would again lead to the table above.

But if she identified four with milk first and four with tea first, then she could not definitively deduce the remaining two. Hence she would be required to guess which of the remaining two had the tea first. Her chance of being correct would be one half, and
therefore we would arrive at the above table with probability 0.5, and at the table below with a like probability.

<table>
<thead>
<tr>
<th></th>
<th>PICK FIRST</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTUAL</td>
<td></td>
</tr>
<tr>
<td>MILK</td>
<td>4</td>
</tr>
<tr>
<td>TEA</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2:

Again using Fisher's Exact Test, the probability that she would arrive at the latter table by chance if she were to guess on each cup is \( p = 0.1032 \) which is above the \( p = 0.05 \) maximum set by the F.D.A. In other words, by guessing on the two unknown cups she would only have a fifty-fifty chance of satisfying the F.D.A.'s criterion. "Again you are correct" responded Fisher. "And now it appears that our subject is through with her tasting. Let us see how well she has done."

Fisher then turned his attention to my friend. He asked if she was clear as to the structure of the experiment. My friend ignored the question, as she appeared to be lost in thought. Suddenly a smile came across her face, and she volunteered "I am ready to give my response to the experiment now." Fisher indicated that she should continue. "Cups number 2, 5, 8 and 9 had the tea added first. The remaining six cups had the milk added first." Fisher seemed puzzled by the response, and reminded her that there were five cups of each kind. My friend responded "I realize that, but I wish to
pick four as having the tea first and six with the milk first." Fisher appeared a bit irritated, but when he started to object my friend interrupted "Recall that in your 1935 Royal Statistical Society paper¹ you declared that the marginal information is ancillary, and hence Fisher's Exact Test is appropriate in 2x2 tables even when the row marginals are not fixed in advance."

Fisher stroked his beard for several moments, then grudgingly informed us that she was correct on the four chosen as having tea added first. Consequently she had incorrectly classified only one cup, number 4, which in fact had the tea added first. Hence we were led to the following table.

Table 3:

<table>
<thead>
<tr>
<th>ACTUAL FIRST</th>
<th>PICK FIRST</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MILK</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>TEA</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

Fisher, using his Exact Test, quickly calculated the probability of obtaining such an array by guessing on each cup, and announced an observed p = 0.0238. My friend smiled broadly as she heard Fisher proclaim that she had met the F.D.A.'s criterion. Fisher then apologized for having to make a hasty departure, and left without further comment.
I turned to my friend and confided in her that I was puzzled by her response. She knew that there were five cups of each kind, yet she guessed only four to have the tea added first. Her response was brilliant. "I was able to correctly classify six of the ten cups. Fortunately, I identified four cups of one type, that is I knew that cups 2, 5, 8 and 9 had tea added first. I also knew that cups 3 and 7 had milk added first. But I was unable to distinguish the remaining four cups. If I were to guess one of these four to have the tea first, and placed the remaining three as having milk first, my chance of meeting the F.D.A.'s criterion would only be 0.025. But by guessing all the unknown cups as having milk added first, I knew that I would incorrectly classify exactly one cup, and I would be guaranteed to arrive at the above table. I further knew that Fisher was committed to using his Exact Test even though the marginal totals were not fixed, that is regardless of how many cups I assigned as having tea added first. By splitting them into groups of four and six I had probability 1.0 that I would be below the F.D.A.'s p = 0.05 maximum. On the other hand, if I had been required to divide them into two groups of five each my chances of satisfying the F.D.A. would have dropped from 1.0 to 0.25."

My friend excused herself as she began to clear the cups from her table. I reviewed what had transpired. If she had been able to identify five cups of one kind, she would be able to classify all ten cups regardless of how many of the other type she could identify, and her observed p value would be p = 0.0040. If she
were able to identify four of one kind and four or fewer of the other she would be able to have an observed $p = 0.0238$. Similarly, if she identified exactly three of one kind and three or fewer of the other she could guarantee $p = 0.0833$, two of one kind and two or fewer of the other a $p = 0.2222$, and one of one kind and one or fewer of the other a $p = 0.5000$.

Suddenly a hand upon my shoulder brought me back to my flat at Putney. My wife was standing over me and asked "Would you care for a cup of tea with your shortbread?" "No thanks" I responded. "I'll stick with hot chocolate."

RESULTS:
The above narrative is intended to stimulate thought about the use of Fisher's Exact Test in settings other than where both sets of marginal totals are fixed. A common application is the clinical trial where interest is in the comparison of two binomial probabilities, such as comparing a treatment with a control procedure. Textbooks commonly advise\(^2\), and statistical packages warn\(^3\) that the Chi Squared Test is suspect when expected numbers in individual cells are below five.

A suggested remedy for the small expected value situation is to apply Yates Correction Factor. However, the main justification for the use of this statistic seems to be that it leads to a $p$ value close to that obtained by Fisher's Exact Test\(^4\). This then leads
us back to questioning the rationale for using Fisher's Exact Test in all applications.

In the tea tasting example, the row marginals are fixed. Fisher, in his original presentation which is quoted above, also requires the column marginals to be fixed, although in later writings he relaxes this condition. We assume that the lady has probability $p'$ of identifying the type of cup (tea or milk added first) with certainty, and probability $(1-p')$ of not having any indication as to type. In this latter case she would have probability 0.5 of guessing correctly on each cup. Further, the positive identifications of the different cups are stochastically independent events.

The null hypothesis for the experiment is $H_0: p' = 0$, and the alternative is $\text{Alt: } p' > 0$. The two strategies above will be denoted "pick 5" for the case where the lady guesses as necessary to make both column marginal totals equal to five (Tables 1 and 2), and "pick 4" for the case where she guesses as necessary to create marginal totals of four and six. In this latter case the column of four will be that of the type which contained the larger set of positively identified cups (for example, Table 3).

Of interest is a comparison of the power curves between the two procedures. The discussion in the introduction was based on the nonrandomized application of Fisher's Exact Test, where size of
The power curves for these nonrandomized tests are presented in Figure 1. As can be seen, over much of the range of the probability \( p' \), the power of the "pick 4" test exceeds that of the "pick 5" test by over 0.3 (for example, 0.54 versus 0.19 at probability \( p' = 0.5 \)). However, some of this disparity could be attributed to the fact that, under the null hypothesis, the probability of rejecting using "pick 5" is \( p = 0.004 \), while using "pick 4" it is \( p = 0.0238 \). Hence, differences in power can, at least in part, be attributed to the difference in sizes of the tests.

Though there seems to be general agreement that randomization is aesthetically inappropriate with Fisher's Exact Test, the power curves for "pick 4" and "pick 5" procedures using the randomized test are given in Figure 2. Again the power for the "pick 5" test is below that of the "pick 4" test. The lady, by adjusting her presentation of results using the knowledge that Fisher is going to perform his analysis as though both sets of marginals are fixed, along with a knowledge of the actual marginals corresponding to types of cups of tea, can increase her chances of rejection under the alternative hypothesis.

**SUMMARY:**
The intent in writing this article is to add to the wealth of interesting debate which has recently taken place concerning the appropriate methods for analyses of 2X2 tables. Intuition
suggests that, had Fisher not instructed the lady as to the numbers of each kind of cup he was presenting, then the lady would have to guess with probability 0.5 of being correct on each cup she could not identify.

It seems reasonable that, by making use of the information as to the distribution of numbers of cups of each kind, the lady should be able to increase the power of the test. However, there is something anti-intuitive about a test which motivates the lady to choose marginals other than according to that actual distribution in order to increase her power.


Fig 1: Nonrandomized Tests

- \text{pick 4}
- \text{pick 5}
Fig 2: Randomized Tests

- **power** vs. **probability**

- Solid line: pick 4
- Dashed line: pick 5