A Second Source of Conservativism in Fisher's Exact Test for Rare Events

Mark Von Tress*

Abstract

Fisher's exact test (FET) is often recommended for small numbers of events of concern in a comparison of proportions, or when the asymptotic assumptions of chi-square or other z tests are not met. This test controls the type I error conservatively since the size of the critical region under the hypothesis may be smaller than the nominal type I error. A second source of conservativism is discussed where the size of the critical region under the alternative in the final observed table may be substantially smaller than the nominal power. This affects the design of a single study because we observe only one 2x2 table instead of many. This paper recommends examining the probability of having less than the nominal power and the variance of the conditional power function prior to study initiation to determine if inverse sampling may provide an advantage over standard sampling.

Key Words: 2x2 Tables, Fisher's Exact Test, Inverse Sampling

1. Introduction

FET is well known to be conservative because the actual size of the critical region under the null hypothesis may be less than or equal to the nominal type I error. The conditional power for many tables in the reference set of tables collected under standard sampling can also be considerably less than the nominal power for the study. This is a second source of conservativism for FET. The variance of the power function for FET is different for standard and inverse sampling. Randomized inverse sampling may be used to bring the power closer to the nominal power. These issues will be examined by an example and a theoretical reason for the issues will be proposed. More details and further developments in comparative randomized inverse sampling are found in Heimann, Von Tress and Gasparini [1].

1.1 Sampling Plans for a 2x2 FET

The same table is used for both sampling plans. It is a simple 2x2 layout of frequencies.

Table 1: 2x2 Table: Count Statistics in Binary Two-Sample Problems.

	RESPONSE		
TREATMENT	positive $+$	negative $-$	
active A	X	Y	$n_1 = X + Y$
control C	$r - X$	$S - Y$	$n_2 = r + S - X - Y$
	r	S	$N = r + S$

Under standard sampling, subjects are randomized to treatment groups. The number of subjects per group are set to n_1 and n_2. Then sampling proceeds until $N = n_1 + n_2$ is reached. In this case r, the total number of positive responses, is random and conditioned upon at the end of the study. In inverse sampling, subjects are randomized to treatment groups but the number of total positive responses in the study, r, is fixed in advance. Sampling proceeds until r positive events occur. The row marginal totals, n_1 and S, are random variables and are conditioned upon at the end of the study.

*Alcon Laboratories, 6201 S. Freeway, Fort Worth TX 76134-2099. I would like to thank Mauro Gasparini, Politecnico di Torino, and Günter Heimann, Novartis Pharma, for their collaboration in this work.

1.2 Motivating Example

The example is based on a study designed to detect a 50% reduction in the probability of a rare, but serious, adverse event.

- The adverse event occurs in about 1 in 1000 cases of a common procedure ($p_C = 0.001$).
- A treatment is proposed that would reduce that rate to 1 in 2000 cases ($p_A = 0.0005$).
- Randomization fraction $p_0 = 0.5858$ of subjects to active and the remainder to control.
- One-side type I error rate of $\alpha = 0.02375$ assuming a 95% utility index for stochastic curtailment and an overall type I error of $\alpha = 0.025$.
- Power of 0.92 assuming a 80% futility index for stochastic curtailment and an overall power of 0.9.
- Use $r = 98$ for an expected sample size of 138495 total surgeries.
- Use a one-sided Fisher's Exact Test.

The sample size was selected based on inverse sampling, and corresponds to the expected value of a negative binomial distribution for S with $r = 98$ events and a success rate of $\bar{p} = p_0 p_A + (1 - p_0)p_A = 0.5858 * 0.0005 + (1 - 0.5858) * 0.001$. The final sample size may be smaller or larger than this expected value.

The randomization fraction is also greater than 0.5 since it is anticipated that there will be fewer events in the active group, so more subjects need to receive the active treatment to increase the speed at which r is reached. Stochastic curtailment was also used in this example because stochastic curtailment can stop the study early on the basis of p-values.

1.3 Simulation Algorithm

Fisher's exact test is based on conditioning on the row or column marginal totals where the other margin was fixed at the beginning of the study. Standard sampling has a different distribution from inverse sampling, so tables must be simulated from different sampling distributions.

- Generate a table from the sampling distribution under the alternative distribution.
- Use the marginal totals to determine the critical values for the FET assuming the null distribution.
- Calculate the actual size of the critical region under the null hypothesis (type I error).
- Calculate the size of the critical region under the assumed alternative distribution (power).
- Collect r or n_1, Power and type I error. Repeat these steps many times and summarize the output.

The example figures plot the actual type I error rates and the power for the simulated tables.

1.4 Comparison of Example Conditional Power Functions

The power function for the two sampling methods were compared for the example problem. For standard sampling, the x-axis is the number of positive cases. For inverse sampling, the x-axis is the sample size under active treatement required to reach 98 events. The histograms are the relative frequency that a value of conditional power was observed for the x variable. Analytical expressions for the associated distributions are provided later in this paper. Figure 1 is for standard sampling. Note that the power increases as the value of r increases. In this figure, 62% of the tables have less than nominal power and the standard deviation of the power function is 3.3%. There is more than a 5% chance that the size of the critical region under the alternative can be less than 0.85 and a 1% chance that the power is less than 0.8 even though the expected power is 92%.

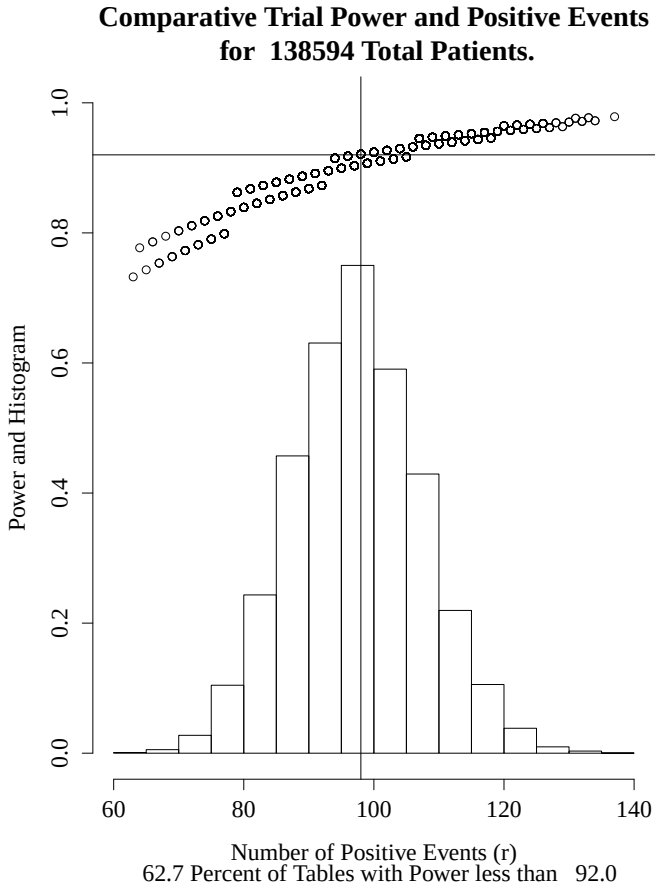

Figure 1: Power function for FET under standard sampling where the number of positive events is not fixed prior to the study.

Inverse sampling demonstrates better control of the power. The standard deviation of the power was 0.3% even though 47.8% of the tables had power less than the nominal value. It is interesting to compare the two figures. Figure 1 achieves nominal power once the number of positive cases exceeds the expected value of 98. This gives some indication of why the inverse sampling controls power. Inverse sampling forces the expected value of the number of positive cases to be achieved, so the nominal power should be obtained with some variation. However, inverse sampling may stop earlier than standard sampling if positive cases acrue rapidly.

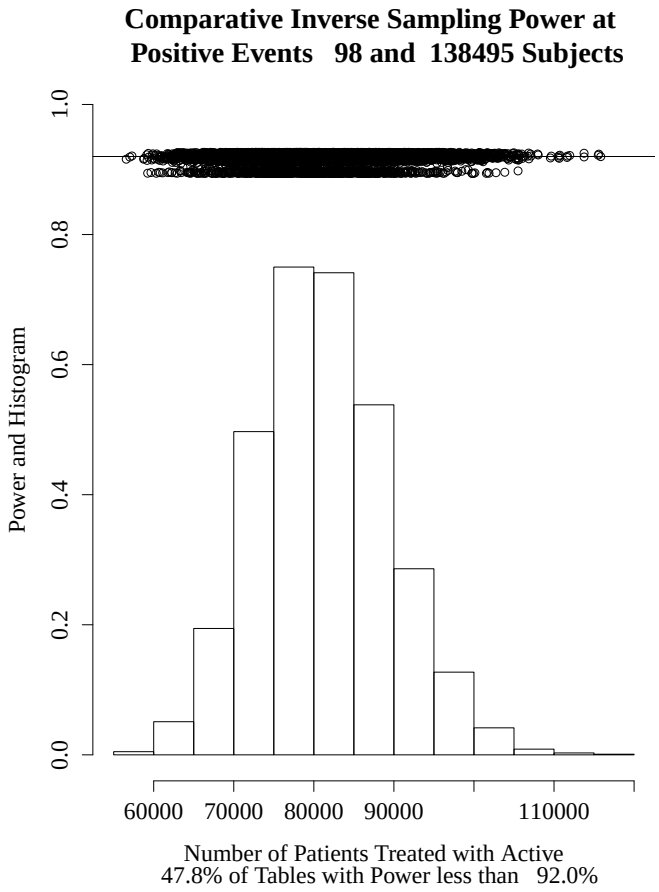

Figure 2: Power function for FET under inverse sampling where the total sample size is not fixed prior to the study.

1.5 Power vs Type I Error in Example

Intuitively, conditional power should increase as the size of the critical region increases, but this is not the case for the example under standard sampling (see Figure 3). This figure is somewhat surprising and indentifies a second source of conservativism in FET. One would assume that the power should increase as the actual size of the critical region increases if FET were simply conservative due to having a smaller type I error than nominal. For many table configurations, the power actually decreases as the type I error increases in standard sampling. This is a second source of conservativism. This problem is not as noticable for inverse sampling.

1.6 Comparisons of Larger Values of Proportions

The second source of conservativism is not as pronounced when comparing larger proportions, e.g. $p_A = 0.45$ and $p_C = 0.55$ (see Figure 4). This example requires fewer subjects to reach the expected number of positive cases. Also, the number of positive cases is more centrally distributed about the expected value. In this example, 53.2% of the tables have power less than the nominal value, but the variance is small and most of the tables have power greater than 90%. It also demonstrates a sawtooth form similar to the power function of the exact one-sample binomial test.

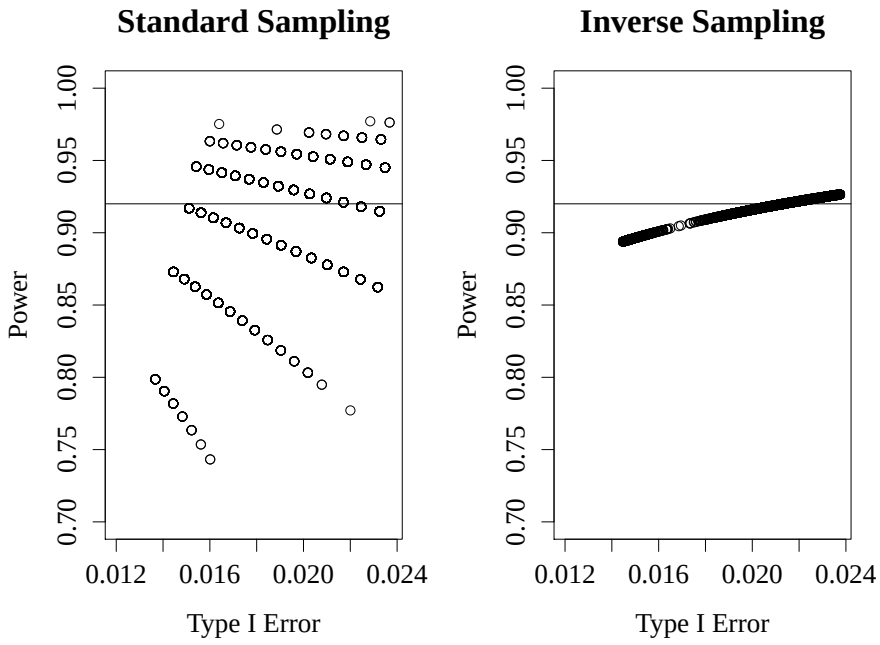

Figure 3: Plot of type I error versus Power for standard and inverse sampling. Power does not increase with type I error for standard sampling.

2. Probability Models

2.1 Probability Model for 2x2 Table

The probability model for the response frequencies is given below. It allows a randomization fraction, p_0, which form the row marginal probabilities. The overall probability of a positive case, $\bar{p}$, is the column 1 marginal probability. It is the convext combination of the response probabilities, p_A and p_C, where the weights are based on the randomization fraction.

Table 2: 2x2 Table: Probabilities in Binary Two-sample Problems.

	RESPONSE		
TREATMENT	positive $+$	negative $-$	
active A	$p_0 p_A$	$p_0(1-p_A)$	p_0
control C	$(1-p_0)p_C$	$(1-p_0)(1-p_C)$	$1-p_0$
	$\bar{p} = p_0 p_A + (1-p_0)p_C$	$1-\bar{p}$	1

Parameters:

- p_A - probability of positive response in Active group
- p_C - probability of positive response in Control group
- p_0 - randomization fraction
- $\bar{p}$ - overall probability of positive response

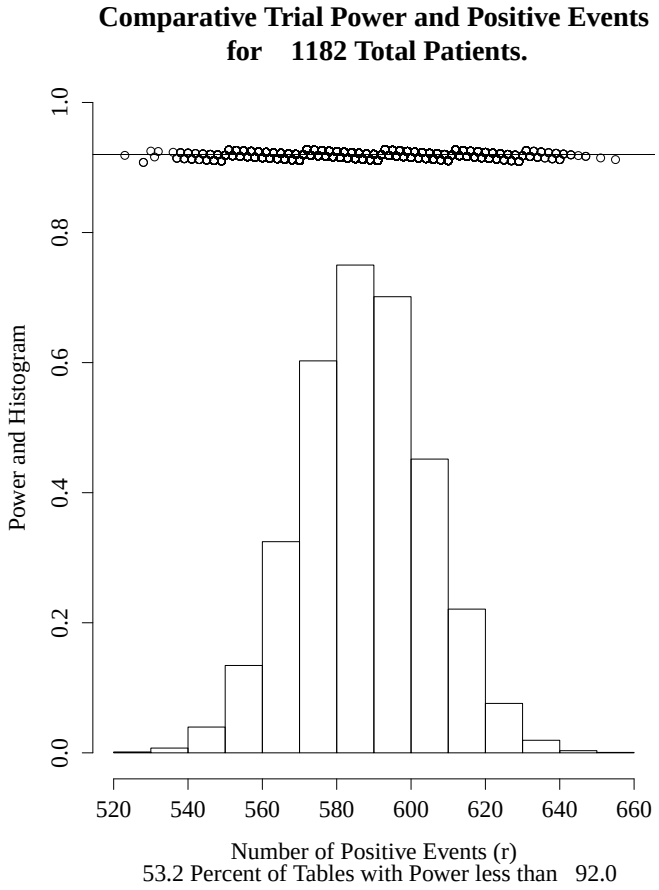

Figure 4: Power for FET under standard sampling is smooth when the proportions being compared are not small: $p_A = 0.45$ and $p_C = 0.55$.

2.2 Joint Distribution for Standard Sampling

Under standard sampling, r is a random variable, $r = x_A + x_C$. There is a binomial distribution for each row in Table 2:

$$f(x_A, r|n_1) = b(x_A; n_1, p_A)b(r - x_A; n_2, p_C)$$

where $b(z; n, \pi)$ is a binomial density for z with sample size n and success probability π. See [4] for details.

2.3 Joint Distribution for Inverse Sampling

The joint distribution for inverse sampling allows randomization and study masking since r is the sum of positive events in both treatment groups, rather than the number of events in each treatment group. The joint sampling distribution is the product of a negative binomial for S, a binomial density for X and a binomial density for Y conditional on S. See [1][2][3] for more details.

The expected sample size for this study was determined from the negative binomial distribution: $E[N] = r + E[S] = r + r(1 - \bar{p})/\bar{p}$. The distribution of S can be used to stop the study. Large values of S provide evidence of the null since $\bar{p}$ becomes close to p_A as S increases and p_A remains fixed.

2.4 Distribution of X after Conditioning

Both sampling methods result in an extended hypergeometric distribution conditional on row and column marginal totals [1][2][3][4][5].

$$f(x|n_1, r, \theta) = \frac{\binom{n_1}{x}\binom{n_2}{r-x}\theta^x}{\sum_{j=a}^{b}\binom{n_1}{j}\binom{n_2}{r-j}\theta^j}, \quad x = a, \ldots, b$$

Parameters:

- $a = \max\{0, r - n_2\}$
- $b = \min\{n_1, r\}$
- $\theta = p_A(1 - p_C)/(p_C(1 - p_A))$ is the odds ratio

When $p_A = p_C$, $f(x|n_1, r, 1)$ is a hypergeometric distribution, which is the null distribution of FET.

3. Hypothesis Testing

3.1 Conditional Power Function for an Individual Table

For this paper, the hypothesis of interest has been $H_0 : p_A \geq p_c$ vs $H_1 : p_A < p_c$, or $H_0 : \theta \geq 1$ vs $H_1 : \theta < 1$. Conditional power is defined as the size of the critical region, C, for an observed table under the alternative distribution: $P_{H_A}[x \in C]$. These have analytical forms for both sampling methods in this paper.

The conditional power function is given by $R_N(\theta|r, s, n_1) = \sum_{x \in C} f(x|n_1, r, \theta)$. For the current example, the power function is $P[x \leq x_{crit}|\theta]$. The power function may be computed using the R package described in [6]. The expected value of the conditional power and its variance also have analytical expressions.

3.2 Expected Power Under Standard Sampling

Expected power is a weighted average of conditional power functions from the reference set of tables (see Section 5, Remarks of [4]).

$$\begin{aligned} Q_N(p_A, p_C|n_1) &= \sum_{r=0}^{N}\sum_{x \in C} f(x, r|n_1, \theta) \\ &= \sum_{r=0}^{N}\left[\sum_{x \in C} f(x|n_1, r, \theta)\right] f(r|n_1) \\ &= \sum_{r=0}^{N} R_N(\theta|r, n_1) f(r|n_1) \\ &= E_{(r|n_1)}[R_N(\theta|r, n_1)] \end{aligned}$$

$$\begin{aligned} f(r|n_1) &= f(r, n_1)/f(n_1) \\ &= \sum_x b(x; n_1, p_A) b(r - x; n_2, p_C) \end{aligned}$$

The weighting function, $f(r|n_1)$, is the analytical form for the histogram in Figure 1. Note that $Q_N(p_A, p_C|n_1)$ is the usual target value for power in FET. It is obtained with repeated

use of the FET and $R_N(\theta|r, n_1)$ is an unbiased estimator of $Q_N(p_A, p_C|n_1)$. See the remarks in [4] for details. However, as observed in Figure 1, the value of R_N may be unacceptably far from Q_N with sufficient probability to give standard sampling an unacceptable chance of producing a severely underpowered 2x2 table.

3.3 Expected Power Under Inverse Sampling

The following is an analogous derivation to the remarks in [4], but for inverse sampling

$$\begin{aligned} Q_N(p_A, p_C|r) &= \sum_{n_1=0}^{r+s} \sum_{x \in C} f(x, n_1|r, s, \theta) \\ &= \sum_{n_1=0}^{N} \left[\sum_{x \in C} f(x|r, s, n_1, \theta) \right] f(n_1|r, s) \\ &= \sum_{n_1=0}^{N} R_N(\theta|r, s, n_1) f(n_1|r, s) \\ &= E_{(n_1|r,s)}[R_N(\theta|r, s, n_1)] \end{aligned}$$

$$\begin{aligned} f(n_1|r, s) &= f(n_1 = x + y, s)/f(s) \\ &= \sum_x b(x; r, p_1) b(y = n_1 - x; s, p_2) \\ &= \left(\frac{p_2}{1 - p_2} \right)^{n_1} (1 - p_1)^r (1 - p_2)^s \sum_x \binom{s}{n_1 - x} \binom{r}{x} \theta^x \end{aligned}$$

where $f(s) = nb(s; r, \bar{p})$, $p_1 = p_0 p_A/\bar{p}$ and $p_2 = p_0(1 - p_A)/(1 - \bar{p})$. The weighting function, $f(n_1|r, s)$, is the analytical form for the histogram in Figure 2. It is derived by first marginalizing $f(x, y = n_1 - x, s)$ over x to get $f(y = n_1 - x, s)$ and then getting $f(n_1|r, s) = f(y = n_1 - x, s)/f(s)$.

3.4 Variance of Power Functions

Before final conditioning, $R_N(\theta|r, n_1)$ is a random variable with some of its values greater than Q_N and others less than Q_N. The variance of the power function for standard sampling is given by

$$\sigma^2_{\mathbf{r}} = \sum_{\mathbf{r=0}}^{\mathbf{N}} [R_N(\theta|r, n_1) - Q_N(p_A, p_C|\mathbf{r})]^2 \mathbf{f}(\mathbf{r}|\mathbf{n_1})$$

The variance of the power function for inverse sampling is given by

$$\sigma^2_{\mathbf{n_1}} = \sum_{\mathbf{n_1=0}}^{\mathbf{N}} [R_N(\theta|r, n_1) - Q_N(p_A, p_C|\mathbf{n_1})]^2 \mathbf{f}(\mathbf{n_1}|\mathbf{r}, \mathbf{s})$$

The major differences between these variances are highlighted in bold face. The biggest difference is in the weighting functions. These variances are difficult to compare analytically, but may be simulated easily.

Note that the sample size, or required number of positive cases, may be increased to help reduce the probability of having lower than nominal power. However, for rare events, standard sampling requires larger increases than inverse sampling because the variance of conditional power is greater in standard sampling.

4. Summary

- Conclusions
  - For rare events, standard sampling may result in 2x2 tables that have conditional power considerably lower than the nominal power of the study.
  - Randomized inverse sampling can improve unconditional power by reducing the variability of the size of the critical region in the reference set of tables possible in a 2x2 setting.
- Recommendations
  - Examine the probability that a table will have less than nominal power
  - Examine the variability of the power function of tables before selecting a sampling plan.
- Additional Conjecture
  - The results of this paper may extend to other forms of conditional testing such as tests based on the posterior distribution or tests adjusted for random covariates. It seems that the reasoning behind the derivation of Q_N, σ_r^2 and $\sigma_{n_1}^2$ applies to conditional testing in general.
  - The results are an example of where a post randomization conditioning variable must be close to its expected value for conditional inference to generalize to a population in a meaningful way.
  - This conjecture will be studied in further research [1].

References

[1] Heimann G., Von Tress M. and Gasparini, M. (2013). Unconditional and conditional inference in clinical trials with small event rates under inverse sampling. *Submitted.*

[2] G. J. Rudolph (1967). A quasi-multinomial type of contingency table. *South African Journal of Statistics,* **1**, 59–65.

[3] Y.M. Bishop, S.E. Fienberg and P.W. Holland (1975). Discrete Multivariate Analysis, Theory and Applications. *Springer,* 2007 reprint, section 13.8, pp 454-457.

[4] W. L. Harkness and L. Katz (1964). Comparison of the Power Functions for the Test of Independence in 2×2 Contingency Tables. *Annals of Mathematical Statistics* Volume 35, Number 3, 1115-1127.

[5] W. L. Harkness (1965). Properties of the extended hypergeometric distribution, *Annals of Mathematical Statistics*. Volume 36, 938–945.

[6] Agner Fog (2011). BiasedURN: Biased Urn model distributions. URL http://www.agner.org/random/ http://www.r-project.org, R package version 1.04. [p6]