Inverse Sampling for McNemars Test

Mark Von Tress*

Abstract

Inverse sampling for McNemars test is studied. Sampling is conducted until a pre-specified number of discordant pairs is observed instead of sampling until a pre-specified total number of pairs is observed. The joint likelihood is decomposed into a product of a negative binomial distribution for the number of pairs required to observe r discordant pairs, a binomial distribution for the number of successes in the concordant observations, and a binomial distribution for the number of successes in the discordant observations. Since inference in this problem is based on the discordant observations, inverse sampling controls the type II error when small numbers of discordant observations are observed and the exact binomial test is required. The control results from fixing the sample size for the exact binomial test. Standard sampling instead lets the sample size for the exact binomial test vary and then performs the test conditionally on the observed number of discordant pairs.

Key Words: McNemar's Test, Inverse Sampling

1. Introduction

Inverse sampling may be used to bring the power of McNemar's test closer to the nominal power. Inference is based on the number of discordant pairs. Inverse sampling controls the type II error when small numbers of discordant observations are observed and the exact binomial test is used. The control results from fixing the sample size for the exact binomial test. Standard sampling lets the sample size for the exact binomial test vary and then performs the test conditionally on the observed number of discordant pairs. The power can be less than nominal if a small number of discordant events occur.

1.1 Sampling Plans for a McNemar's Test

McNemar's test is used when two binary measuremtns are made on the same unit of analysis. There are 4 possible pairs of outcomes. The following table summarizes the responses where two treatments are given to a subject and the response to each treatment is positive or negative. A concordant response is where the responses are the same, N_c, and a discordant event is where the responses differ, N_d. For standard sampling, the total number of subjects is

	Treatment 0		
Treatment 1	positive 1	negative 0	$N_d = X_{10} + X_{01} = r$
positive 1	X_{11}	X_{10}	
negative 0	$X_{01} = r - X_{10}$	X_{00}	
			$N_c = X_{11} + X_{00}$

Table 1: 2x2 table count statistics in the paired binary problem.

fixed in advance, say $N = N_c + N_d$, an sampling continues until N is reached. The number of discordant events, r, is random and conditioned upon at the end of the study. For inverse sampling, the total number discordant events is fixed in andvance, $r = N_d$, and sampling continues until r is reached. In this case N_c is a random variable.

*Alcon Laboratories, Fort Worth, TX, Mark.VonTress@alcon.com

2. Probability Models

2.1 Probability Model for McNemar's Test

The usual treatment difference of interest is between the first row marginal probability and the first columm marginal probability. This subtraction removes p_{11} from the treatment comparison and isolates the treatment difference into the difference between the probabilities of discordant pairs. The cell probabilities may be reprarmaterized into expressions involving the probability of a discordant pair, the treatment difference and the marginal row/column probabilities.

	Treatment 0		
Treatment 1	positive 1	negative 0	
positive 1	$p_{11} = \frac{p_{1\cdot}+p_{0\cdot}-\rho}{2}$	$p_{10} = \frac{\rho+\Delta}{2}$	$p_{1\cdot} = p_{11} + p_{10}$
negative 0	$p_{01} = \frac{\rho-\Delta}{2}$	$p_{00} = 1 - \frac{p_{1\cdot}+p_{0\cdot}+\rho}{2}$	$q_{1\cdot} = 1 - p_{1\cdot}$
	$p_{\cdot 0} = p_{11} + p_{01}$	$q_{\cdot 0} = 1 - p_{\cdot 0}$	1

Table 2: 2x2 table probabilities in paired binary problem.

where

- ρ = probability of a discordant pair, $p_{10} + p_{01}$
- Δ = treatment difference $p_{1\cdot} - p_{\cdot 0} = p_{10} - p_{01}$

Note that this parameterization imposes the constraint that $|\Delta| \leq \rho$ to satisfy the requirement that p_{01} and p_{10} be in $[0, 1]$. That is to say, the absolute value of the treatment difference must be less than or equal to the probability of a discordant pair for this parameterization to be well formulated.

2.2 Joint Distribution for Standard Sampling

Under standard sampling, the response pairs are 4-fold multinomial random varaiables. The multinomial distribution may be factored into 3 densities under the McNemar's parameterization (See Appendix A). There is a binomial distribution for each diagonal in the table and a binomial distribution for the number of discordant pairs. The joint distribtuion under standard sampling is the product of three binomial densities:

$$f(x_{10}, x_{11}, N_d) = bi\left[x_{10}; r, \frac{1}{2}(1 + \frac{\Delta}{\rho})\right] bi\left[x_{11}; N - r, \frac{p_{1\cdot} + p_{0\cdot} - \rho}{2(1-\rho)}\right] bi\left[N_d = r; N, \rho\right]$$

Tables may be generated by first generating r, and then x_{11} and x_{10} using the generated value of r. The test statistics are well known [1]:

- large sample test statistic: $T = (x_{10} - x_{01})/\sqrt{x_{10} + x_{01}}$,
- exact test statistic: x_{10} in exact binomial test that $\pi = 0.5$, where $\pi = (\rho + \Delta)/(2\rho) = 0.5 \implies H_0 : \Delta = 0$

Although the information about the treatment difference, Δ, seems isolated in the distribution for x_{10} the constraint $|\Delta| \leq \rho$ makes ρ a nuisance parameter that cannot be entirely removed from power calculations and information about ρ is contained in other factors of the likelihood. The probability of a discordant pair is only removed from the model for x_{10} under the null hypothesis.

2.3 Joint Distribution for Inverse Sampling

There is a binomial distribution for each diagonal in the table and a negative binomial distribution for the number of concordant pairs (See Appendix B). The joint density is the product of two binomials and a negative binomial density

$$f(x_{10}, x_{11}, N_c) = bi\left[x_{10}; r, \frac{1}{2}(1 + \frac{\Delta}{\rho})\right] bi\left[x_{11}; N_c, \frac{p_1 + p_0 - \rho}{2(1-\rho)}\right] nb\left[N_c; r, \rho\right]$$

Tables may be generated by first generating N_c, and then x_{11} using the generated value of N_c. The value of x_{10} is generated independently of x_{11} and N_c. The test statistics are well known [1]:

The test statistics are

- large sample test statistic: $T = (2x_{10} - r)/\sqrt{r}$,
- small sample test statistic: x_{10} in exact binomial test that $\pi = 0.5$ where $\pi = (\rho+\Delta)/(2\rho) = 0.5 \implies H_0 : \Delta = 0$,
- N_c is random - number of concordant pairs needed until r discordant pairs have been observed

2.4 Maximum Likelihood Estimators

The sampling joint likelihoods differ by a normalization factor and which variables are random. The kernels for both likelihoods are symbolically identical. The estimators are

- $\tilde{\rho} = \frac{r}{N_c+r}$
- $\tilde{p}_{10} = \frac{x_{10}}{N_c+r}$
- $\tilde{p}_{11} = \frac{x_{11}}{N_c+r}$
- $\tilde{\Delta} = 2\tilde{p}_{10} - \tilde{\rho} = (x_{10} - x_{01})/(N_c + r)$

Under standard sampling, the $\tilde{\rho}$, $\tilde{p}_{10}$ and $\tilde{p}_{11}$ are just functions of cell probabilities since $N = N_c+r$. For inverse sampling, $\tilde{\rho}$ is the usual MLE for ρ for the negative binomial distribution. Confidence intervals for standard sampling are discussed in [2].

2.5 Sample Size Selection under Inverse Sampling

Sample size selection for standard sampling is discussed in [1] and [3]. Under inverse sampling, methods for the exact binomial sample size estimtaion may be used because r is fixed and $x_{10} \sim bi\left[r, \frac{1}{2}\left(1 + \frac{\Delta}{\rho}\right)\right]$. The hypothesis of interest is $H_0 : \Delta = 0$ vs $H_A : \Delta = \delta$ where $\delta = \frac{1}{2}\left(1 + \frac{\Delta}{\rho}\right)$. However, the detectable difference, δ, depends on the probability of a discordant event, ρ. This is also an issue in standard sampling. The dependency is further complicated by the fact that the size of treatment difference, Δ, must be less than ρ. Larger values of ρ reduce the value of δ and decrease the power of the exact binomial test.

Some prior information is needed for the value of ρ to help pick δ. References [1] and [3] suggest internal pilot studies to estimate ρ or external evidence to help with this. Note that most information about ρ is contained in $N_c \sim nb(r, \rho)$ and this distribution does not contain direct information about the treatment difference. Hence information about ρ may be gathered as the study progresses without unmasking the treatment codes. Futility might be declared if the current estimate of ρ is unlikely to exceed Δ.

3. Examples

All of the examples assume the same design: $p_{1\cdot} = 0.87$, $p_{\cdot 0} = 0.75$, $N = 104$, $\Delta = 0.12$, $\rho = 0.2$, type I error = 0.05 one-sided, type II error = 0.1. The number of pairs, $N = 104$ is the sample size recommended by NQuery Advisor using the other assumed values. The expected number of discordant pairs under standard sampling is $n_d = 21 = (\rho * N = 20.8)$. For inverse sampling, $r = 21$ will be assumed.

3.1 Large Sample Test

This section provides some examples of the effect of the sampling methods on the large sample tests. In particualar, the examples demonstrate what may happen to power and ρ in equations 5.4 and 6.2 of [1] which are used in Nquery Advisor. The conclusion is that inverse sampling produces a less variable conditional power function, which increases the ability to predict what the size of the final critical region will be under the alternative hypothesis.

The algorithm for generating tables in this example is as follows:

- Generate a table from the sampling distribution under the alternative distribution.
- Compute T, power (Φ in eqn 5.4) and ρ (ψ in eqn 6.2).
- Repeat these steps many times and summarize the output.

The distribution of T for both methods appear similar, which means they will have similar tests. The standard error of the difference is constant for inverse sampling and varible for standard sampling. Consequently, confidence interval width on the treatment difference will be constant for inverse sampling.

The difference between Figures 1 and 2 is remarkable. While both curves have declining power as the fraction of discordant events increases, the power may be much lower for standard sampling at values of $\tilde{\rho}$ smaller than 0.2. Also, while both sampling schemes have declining power with increasing $\tilde{\rho}$, inverse sample presents a more predictable decline that is in line with the fact that δ declines with increasing ρ. The reason for this is that the number of random variables in the test statistic, T, for each of the sampling mehtods. For inverse sampling, T only has one random variable, and has two under standard sampling.

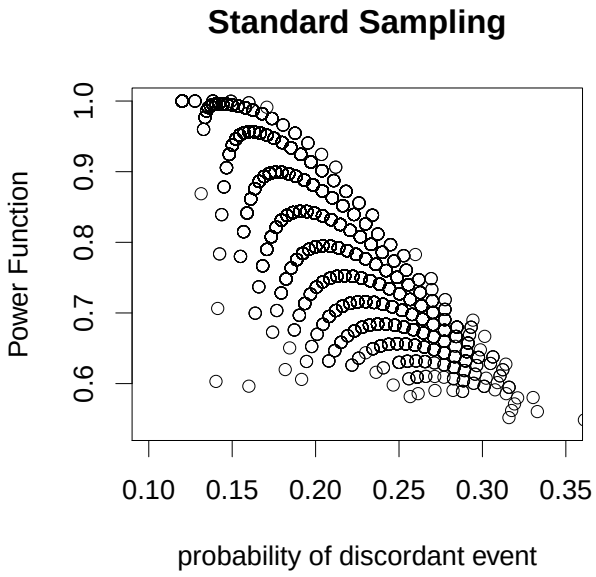

Figure 1: Plot of unconditional power by $\tilde{\rho}$ from [1] for standard sampling. 79% of tables have less than 0.9 power and standard deviation of power is 9.1%

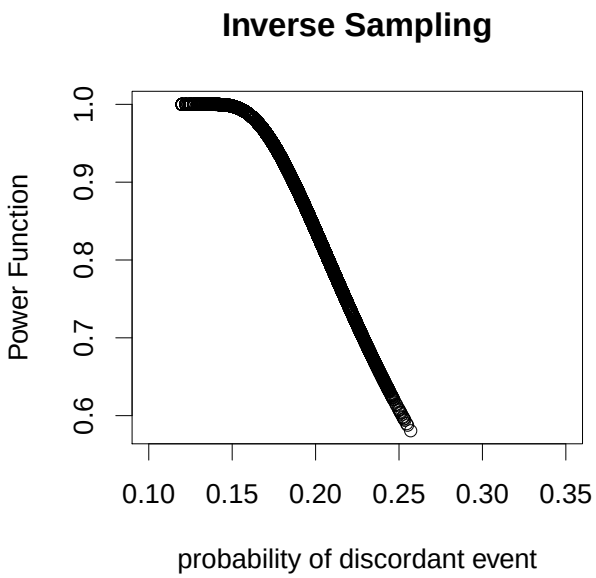

Figure 2: Plot of unconditional power by $\tilde{\rho}$ from [1] for inverse sampling. 61% of tables have less 0.9 power and standard deviation of power is 9.2%

3.2 Examine Effect of Sampling Methods on Exact Test

This section provides some examples of the effect of the sampling methods on the exact tests. In particualar, the examples demonstrate what may happen to power and ρ.

The algorithm is

- Generate a table from the sampling distribution under the alternative distribution.
- Use the number of discordant events to determine the critical values for the test assuming the null distribution.
- Calculate the size of the critical region under the assumed alternative distribution. (power)
- Calculate the proportion of discordant events.
- Repeat these steps many times plot $\tilde{\rho} = N_d/N$ versus power.

For standard sampling, $N = 104$ pairs will be used and the expected number of discordant pairs is $E[N_d] = 21 (= \rho * N = 20.8)$. The detectable difference, δ, is 0.804 to get 90% power in a 1 sided exact binomial test with a 5% chance of a type I error. To detect a treatment difference of $\Delta = 0.12$, then $\rho = 0.1974$, which is approximately 0.2. Note that the critical region size will vary as the number of discordant pairs varies. For instance, if $n_d = 21$ then the critical region is to reject the null hypothesis if $x_{10} \geq 15$. However if $n_d = 18$, the critical region is $x_{10} \geq 12$ and the size of the critical region assuming the alternative hypothesis, $\delta = 0.804$, is reduced to 0.6631.

For inverse sampling, sampling continues until $r = 21$ discordant pairs are observed, so the expected sample size will vary around $N = 104$ pairs. However, the critical region will stay fixed since r is fixed at rejecting the null hypothesis if $x_{10} \geq 15$. The detectable difference will remain at $\delta = 0.804$. As a consequence, there is only one test to be planned for, and its power will remain constant.

Figure 3 and 4 illustrate these observations. The power function decreases as the probability of discordant events decreases in standard sampling. Still, it is a smoother power function than that of the large sample test. This indicates, that the exact test is probably a better test to use than T in terms of the predictability of the size of the critical region under the alternative hypothesis.

Figure 4 demonstrates that power is controlled regardless of the number of concordant pairs that are observed.

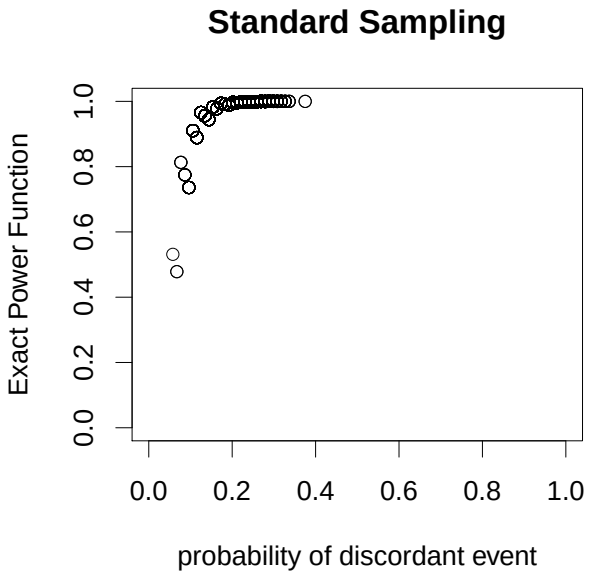

Figure 3: Plot of conditional power by $\tilde{\rho}$ for exact test in standard sampling. Variation in r allows low power if ρ is small.

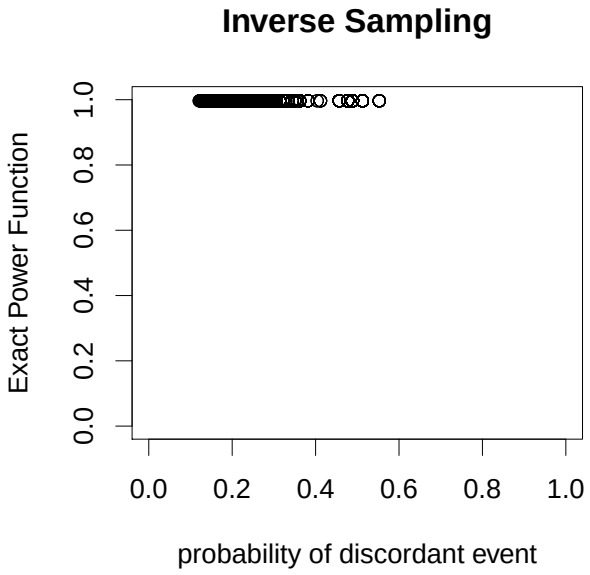

Figure 4: Plot of conditional power by $\tilde{\rho}$ for exact test in inverse sampling. Power is constant, but overall sample size varies.

4. Summary

Inverse sampling can control the size of the critical region under the alternative (power) for tables that result in a paired binary experiment. Inverse sampling also removes the variability of the number of discordant events at the expense of a variable overall sample size.

One remaining issue is to develop interim analyses for inverse sampling. It is straighforward to apply well known procedures for the one sample binomial test based on the number of discordant pairs [4]. However, the number of concordant pairs may be infinite. To stop based on futility, one could do something Bayesian based on the constraint that $|\Delta| \leq \rho$. A conjugate beta could be placed on ρ, say $\rho \sim Beta(\alpha, \beta)$ and $N_c \sim NB(r, \rho)$. The hyperparameters would be selected based on the proposed design and sample size calculation. In particular, one would select the hyperparameters on the basis of how precisely ρ can be estimated under the proposed design. It's easy to show that the posterior distribution of ρ is also $\rho|n_c \sim Beta(\alpha + n_c, \beta + r)$. One could the compute the posterior probability that the constraint will be met, $p(\rho > |\Delta||n_c)$, and stop for futility if that probability is small. More work is needed to formulate something like this based on frequentist testing, but the principles would be similar. The advantage of this test is that it does not require unmasking the treatment code. This test may also work for standard sampling since most of the information about ρ is contained in a binomial density, but the posterior distribution would have an extra random variable in it, $\beta^* = \beta + n_d$.

A. Factorization of the Joint Density for Standard Sampling

The sampling distribution is multinomial. The multinomial coefficient may be factored into 3 binomial coeficients by multiplying and dividing by $n_d!n_c!$,

$$\binom{n}{x_{10}x_{01}x_{11}x_{00}} = \binom{n_d}{n_{10}}\binom{n_c}{x_{11}}\binom{n}{n_d}$$

The desired factorization results from multiplying and dividing by $\rho^{n_d}(1-\rho)^{n-n_d}$, and substituting the parameterization from Table 2:

$$\begin{aligned} P\left[x_{10}, x_{11}, n_c\right] &= \binom{n_d}{x_{10}}\left(\frac{\rho+\Delta}{2\rho}\right)^{x_{10}}\left(1-\frac{\rho+\Delta}{2\rho}\right)^{n_d-x_{10}} \\ &\quad \binom{n_c}{x_{11}}\left(\frac{p_{1\cdot}+p_{0\cdot}-\rho}{2(1-\rho)}\right)^{x_{11}}\left(1-\frac{p_{1\cdot}+p_{0\cdot}-\rho}{2(1-\rho)}\right)^{n_c-x_{11}} \\ &\quad \binom{n}{n_d}\rho^{n_d}(1-\rho)^{n-n_d} \end{aligned}$$

B. Derivation and Factorization of the Joint Density for Inverse Sampling

I attribute this method of proof to Günter Heimann and Mauro Gasparini The notation in Table 1 and Table 2 will be used in this proof. Under inverse sampling, the experiment completes with a discordant pair, (1,0) or (0,1), which are mutually exclusive events. For both cases, there are $\binom{n-1}{n-r}$ ways to partition the $n-r$ concordant events among the $n-1$ events. Also, there are $\binom{n-r}{x_{11}}$ ways to partition the concordant successes among the $n-r$ concordant events. Each partioning has probability $p_{11}^{x_{11}}p_{00}^{n-r-x_{11}}$.

In case 1, (1,0), there are $\binom{r-1}{x_{10}-1}$ ways to partition the remaining x_{10} discordant events with success in the active group, each having probability $p_{10}^{x_{10}-1}p_{01}^{r-x_{10}}$. Similarly, in case 2, (0,1), there are $\binom{r-1}{x_{10}}$ ways to partition the remaining x_{10} discordant events with success in the active group, each having probability $p_{10}^{x_{10}}p_{01}^{r-1-x_{10}}$.

The joint probabilty function may now be written as

$$\begin{aligned} P\left[x_{10}, x_{11}, n\right] &= p_{10}\binom{n-1}{n-r}\binom{n-r}{x_{11}}p_{11}^{x_{11}}p_{00}^{n-r-x_{11}}\binom{r-1}{x_{10}-1}p_{10}^{x_{10}-1}p_{01}^{r-x_{10}} \\ &+ p_{01}\binom{n-1}{n-r}\binom{n-r}{x_{11}}p_{11}^{x_{11}}p_{00}^{n-r-x_{11}}\binom{r-1}{x_{10}}p_{10}^{x_{10}}p_{01}^{r-1-x_{10}} \end{aligned}$$

since the cases are mutually exclusive. The likelihood may reduced to common factors and simplified as

$$\begin{aligned} P\left[x_{10}, x_{11}, n\right] &= \left[\binom{r-1}{x_{10}-1} + \binom{r-1}{x_{10}}\right] \\ &\quad \binom{n-1}{n-r}\binom{n-r}{x_{11}} p_{11}^{x_{11}} p_{00}^{n-r-x_{11}} p_{10}^{x_{10}} p_{01}^{r-x_{10}} \\ &= \binom{n-1}{n-r}\binom{n-r}{x_{11}} p_{11}^{x_{11}} p_{00}^{n-r-x_{11}} \\ &\quad \binom{r}{x_{10}} p_{10}^{x_{10}} p_{01}^{r-x_{10}} \end{aligned}$$

since $\binom{a-1}{b-1} + \binom{a-1}{b} = \binom{a}{b}$. The desired factorization results from multiplying and dividing by $\rho^r(1-\rho)^{n-r}$, substituting the parameterization from Table 2, and recalling that $n_c = n - r$:

$$\begin{aligned} P\left[x_{10}, x_{11}, n_c\right] &= \binom{r}{x_{10}} \left(\frac{\rho+\Delta}{2\rho}\right)^{x_{10}} \left(1 - \frac{\rho+\Delta}{2\rho}\right)^{r-x_{10}} \\ &\quad \binom{n_c}{x_{11}} \left(\frac{p_{1\cdot} + p_{0\cdot} - \rho}{2(1-\rho)}\right)^{x_{11}} \left(1 - \frac{p_{1\cdot} + p_{0\cdot} - \rho}{2(1-\rho)}\right)^{n_c - x_{11}} \\ &\quad \binom{n_c + r - 1}{n_c} \rho^r (1-\rho)^{n_c} \end{aligned}$$

References

[1] Miettinen, O.S. (1968) The Matched Pairs Design in the Case of All-or-None Responses, *Biometrics*, **24**, 339-352

[2] Newcombe RG (1998) Improved Confidence Intervals for the difference between binomial proportions based on paired data, *Statistics in Medicine*, **17**, 2635-2650.

[3] Connet, JE, Smith, JA and McHugh, RB (1987) Sample Size and Power for Pair-Matched Case Control Studies, *Statistics in Medicine*, **6**, 53-57

[4] Jennison C, Turnbull BW, *Group Sequential Methods with Applications to Clinical Trials* Chapman&Hall/CRC, New York, 2000