# Inverse Sampling for McNemar's Test 

Mark Von Tress ${ }^{1}$<br>${ }^{1}$ Alcon, Fort Worth TX 1<br>Correspondence: Mark Von Tress, 6201 South Freeway, Fort Worth, TX 76134-2099, E-mail: Mark.VonTress@ Alcon.com

Received: September 30, 2016 Accepted: November 3, 2016 Online Published: December 21, 2016
doi:10.5539/ijsp.v6n1p78 URL: https://doi.org/10.5539/ijsp.v6n1p78


#### 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.


Keywords: McNemar's test, inverse sampling

## 1. Introduction

This paper focuses on planning an experiment on paired binary responses. We want our hypothesis test comparing the probability of success in the two margins to control both the type I error and power. During the planning stage of the experiment, there are many possible $2 \times 2$ tables that may result, and each has a critical region. The size of the critical region under the null hypothesis is the type I error rate and the size of the critical region under the alternative hypothesis is the power. We create the critical regions for all possible tables so that the type I error is some fixed level, however, this means that the power in each table may vary. Generally, we speak of power of a discrete test as being the expected power over all the possible tables that can result in the experiment. This metric is okay if we wish to describe the average performance of the test when used repeatedly by the scientific community. However, for the individual experiment, only one table is observed, and we would like to have some assurance that size of the critical region is controlled under both the null and alternative hypotheses. Miettinen (1968) calls the average power "unconditional power" and power of an individual table "conditional power". We want to avoid the unlucky situation of making a type II error simply because conditional power is low for the table we observe in our experiment.

The paper examines conditional power for different sampling schemes and test statistics for the paired binary experiment. Inverse sampling may be used to bring the conditional power of McNemar's test (McNemar, 1947) 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 conditional power can be less than nominal if a small number of discordant events occur.

The paper first presents probability models, estimators and test statistics for paired binary experiments. Probability models for standard and inverse sampling schemes are developed. Large sample and exact tests are presented. Sample size estimation will be considered for the sampling schemes and test statistics. An example will be presented next. It illustrates the kinds of considerations needed in selecting a sampling scheme and test statistic. Tests are compared in terms of the percent of tables with conditional power less than nominal power and the standard deviation of the conditional power.

## 2. Probability Models

This section provides probability models, densities for the two sampling schemes, estimators, and some advice on sample size selection. Note that McNemar (1947) describes the large sample test under standard sampling. The exact test is an obvious extension of McNemar's test when the number of discordant events is small.

### 2.1 Probability Model for the Paired Binary Experiment

McNemar's test is used when two binary measurements are made on the same unit of analysis. Some example experiments
are before/after studies on the same subject, or when two judges rate the same experimental unit. The following table summarizes the responses where two treatments are given to a subject and the response to each treatment is positive or negative. There are 4 possible pairs of outcomes: $(1,1),(1,0),(0,1)$ and $(0,0)$. A concordant response is where the responses are the same, $N_{c}$, and a discordant event is where the responses differ, $N_{d}$.

Table 1. $2 \times 2$ Table Count Statistics in the Paired Binary Problem

|  | Treatment 0 |  |  |  |
| ---: | :---: | :---: | :--- | :---: |
| Treatment 1 | positive 1 | negative 0 | $N_{d}=X_{10}+X_{01}=r$ |  |
| positive 1 | $X_{11}$ | $X_{10}$ |  |  |
| negative 0 | $X_{01}=r-X_{10}$ | $X_{00}$ |  |  |

For standard sampling, the total number of subjects is fixed in advance, say $N=N_{c}+N_{d}$, and 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 of discordant events is fixed in advance, $r=N_{d}$, and sampling continues until $r$ is reached. In this case, the number of concordant events, $N_{c}$, is a random variable.

The usual treatment difference of interest is between the first row marginal probability and the first column 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 reparameterized into expressions involving the probability of a discordant pair, the treatment difference and the marginal row/column probabilities.

Table 2. 2x2 Table Probabilities in Paired Binary Problem

|  | Treatment 0 |  |  |  |
| ---: | :---: | :---: | :--- | :---: |
| Treatment 1 | positive 1 | negative 0 |  |  |
| positive 1 | $p_{11}=\left(p_{1}+p_{0} \cdot-\rho\right) / 2$ | $p_{10}=(\rho+\Delta) / 2$ | $p_{1}=p_{11}+p_{10}$ |  |
| negative 0 | $p_{01}=(\rho-\Delta) / 2$ | $p_{00}=1-\left(p_{1 \cdot}+p_{0} .+\rho\right) / 2$ | $q_{1} .=1-p_{1}$. |  |
|  | $p_{.0}=p_{11}+p_{01}$ | $q_{.0}=1-p_{.0}$ | 1 |  |

where

- $\rho=$ probability of a discordant pair, $p_{10}+p_{01}$
- $\Delta=$ treatment difference $p_{1} .-p_{0} .=p_{10}-p_{01}$

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.

### 2.1 Joint Distribution for Standard Sampling

Under standard sampling, the response pairs are 4-fold multinomial random variables. The multinomial distribution may be factored into 3 densities under the McNemar's parameterization (see Appendix 5.1). There is a binomial distribution for each diagonal in the table and a binomial distribution for the number of discordant pairs. The joint distribution under standard sampling is the product of three binomial densities:

$$
\begin{equation*}
f\left(x_{10}, x_{11}, N_{d}\right)=b i\left[x_{10} ; r, \frac{1}{2}\left(1+\frac{\Delta}{\rho}\right)\right] b i\left[x_{11} ; N-r, \frac{p_{1} \cdot+p_{0} \cdot-\rho}{2(1-\rho)}\right] b i\left[N_{d}=r ; N, \rho\right] \tag{1}
\end{equation*}
$$

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 (McNemar, 1947, Miettinen 1968, Newcombe, 1998):

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

Although the information about the treatment difference, $\Delta$, seems isolated in the distribution for $x_{10}$ the constraint $|\Delta| \leq$ $\rho$ makes $\rho$ a nuisance parameter that cannot be entirely removed from power calculations and information about $\rho$ 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.

### 2.2 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 5.2). The joint density is the product of two binomials and a negative binomial density

$$
\begin{equation*}
f\left(x_{10}, x_{11}, N_{c}\right)=b i\left[x_{10} ; r, \frac{1}{2}\left(1+\frac{\Delta}{\rho}\right)\right] b i\left[x_{11} ; N_{c}, \frac{p_{1}+p_{0}-\rho}{2(1-\rho)}\right] n b\left[N_{c} ; r, \rho\right] \tag{2}
\end{equation*}
$$

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 (McNemar, 1947, Miettinen 1968, Newcombe, 1998):
The test statistics are

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


### 2.3 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

- $\tilde{\rho}=r /\left(N_{c}+r\right)$
- $\tilde{p}_{10}=x_{10} /\left(N_{c}+r\right)$
- $\tilde{p}_{11}=x_{11} /\left(N_{c}+r\right)$
- $\tilde{\Delta}=2 \tilde{p}_{10}-\tilde{\rho}=\left(x_{10}-x_{01}\right) /\left(N_{c}+r\right)$

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 $\rho$ for the negative binomial distribution. Confidence intervals for $\Delta$ under standard sampling are discussed in (Newcombe, 1998). The simplest interval is $\hat{\Delta} \pm z_{1-\alpha / 2}$ se where $\left.\left.s e=\sqrt{( } N_{c} N_{d}+4 x_{10} x_{01}\right) / N^{3}\right)$. A similar confidence interval for $\Delta$ under inverse sampling is $\hat{\Delta} \pm z_{1-\alpha / 2} \sqrt{V}(\hat{\Delta})$ where the $V(\hat{\Delta})$ may be found in equation 8 of Appendix 5.3.

### 2.4 Sample Size Selection under Inverse Sampling

Sample size selection for standard sampling is discussed in Miettinen (1968) and Connet et. al. (1987). Under inverse sampling, methods for the exact binomial sample size estimation may be used because $r$ is fixed and $x_{10} \sim b i\left[r, \frac{1}{2}\left(1+\frac{\Delta}{\rho}\right)\right]$. 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, $\delta$, depends on the probability of a discordant event, $\rho$. This is also an issue in standard sampling. The dependency is further complicated by the fact that the size of treatment difference, $\Delta$, must be less than $\rho$. Larger values of $\rho$ reduce the value of $\delta$ and decrease the power of the exact binomial test.
Some prior information is needed for the value of $\rho$ to help pick $\delta$. Miettinen (1968) and Connet et. al. (1987) suggest internal pilot studies to estimate $\rho$ or external evidence to help with this. Note that most information about $\rho$ is contained in $N_{c} \sim n b(r, \rho)$ and this distribution does not contain direct information about the treatment difference. Hence information about $\rho$ may be gathered as the study progresses without unmasking the treatment codes. Futility might be declared if the current estimate of $\rho$ is unlikely to exceed a pre-specified value of $\Delta$.

## 3. Examples

All of the examples assume the same design: $p_{1}=0.87, p_{.1}=0.75, r=23, \Delta=0.12, \rho=0.2$, type I error $=0.05$ one-sided, and type II error $=0.1$. The number of discordant pairs, $r=23$, is the smallest sample size that gives an exact
binomial test under inverse sampling with $90 \%$ power. By choosing $r=23$ for inverse sampling, the fixed sample size for standard sampling is $N=111$ so that ceiling $(\rho * N)=r$. Ten thousand $2 \times 2$ tables were generated for each sampling method assuming these parameters.

### 3.1 Effect of Sampling Methods on Large Sample Test

This section provides some examples of the effect of the sampling methods on the large sample tests. In particular, the examples demonstrate what may happen to condtional power and $\rho$ in equations 4.3 and 6.2 of Miettinen (1968). Note that a popular sample size calculator program, nQuery Advisor (nQuery Advisor 6.0, 2005, Method POT1), computes unconditional power using equation 5.3 from Miettinen (1968). The conclusion is that inverse sampling produces a smaller probability of having a table with less than $90 \%$ conditional power.

The algorithm for generating tables in this example is as follows:
Generate a table from the sampling distribution under the alternative distribution.
Compute $T$, the conditional power function $(\Pi(\delta \mid r)$ in eqn 4.3) and the nuisance parameter $\rho$ ( $\psi$ in eqn 6.3) of Meittinen (1968).

Repeat these steps many times and summarize the output.
The distribution of $T$ for both methods appear similar, which means they will reach similar decisions. The standard error of the difference is constant for inverse sampling and variable for standard sampling. Consequently, confidence interval width on the treatment difference will be constant for inverse sampling.


Figure 1. Conditional power for tables for large sample test under standard sampling


Figure 2. Conditional power for tables for large sample test under inverse sampling
The difference between Figures 1 and 2 is remarkable. While both curves have declining power as the fraction of discor-
dant 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 sampling presents a more predictable decline that agrees with the fact that $\delta$ declines with increasing $\rho$. The reason is the number of random variables in the test statistic, $T$, are different for the sampling methods. For inverse sampling, $T$ only has one random variable, and has two under standard sampling.
Under standard sampling $31.9 \%$ of tables have less than $90 \%$ conditional power and the standard deviation of the conditional power is $5.4 \%$ (Figure 1). Under inverse sampling $18.8 \%$ of tables have less than $90 \%$ power and the standard deviation of power is $5.0 \%$ (Figure 2).

### 3.2 Effect of Sampling Methods on Exact Test

This section provides some examples of the effect of the sampling methods on the exact tests. In particular, the examples demonstrate what may happen to power and $\rho$.

## 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 (conditional power)
Calculate the proportion of discordant events.
Repeat these steps many times plot $\hat{\rho}=N_{d} / N$ versus power.
For standard sampling, $N=111$ pairs were used and the expected number of discordant pairs is $E\left[N_{d}\right]=23$ (= ceiling $(\rho *$ $N)$ ). The detectable difference, $\delta$, is 0.80 to get $90 \%$ power in a 1 sided exact binomial test with a $5 \%$ chance of a type I error. The critical region is to reject the null hypothesis if $x_{10} \geq 16$. Note that the critical region size varies as the number of discordant pairs varies under standard sampling, but not under inverse sampling.

For inverse sampling, we will sample to $r=23$ discordant pairs are observed, so the expected sample size will vary around $N=111$ pairs. However, the critical region will stay fixed since $r$ is fixed. The detectable difference will remain at $\delta=0.8$. As a consequence, there is only one test to be planned for, and it's power will remain constant.
Figures 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 3. Conditional power for tables for exact test under standard sampling


Figure 4. Conditional power for tables for exact test under inverse sampling

### 3.3 Conditional Power Comparisons

This section presents the sort of considerations needed for planning a study. While this example shows a weakness of using a paired binary experiment, not all proposed designs will have these weaknesses to the same degree. The large sample test may actually be acceptable under standard sampling and different assumptions about sample size, parameters and type I/II errors.
The following table contains descriptive statistics for the four tests in this example.The first row in the next table is the average of the conditional power, which estimates unconditional power. The second row is the percent of tables with conditional power less than the nominal value of $90 \%$. The last row is the standard deviation of the conditional power.

While the unconditional power of the large sample test under standard sampling compared favorably with that of the exact test under inverse sampling, there was a $31.9 \%$ chance that the conditional power was less than the nominal power for standard sampling. The exact test under standard sampling had less than nominal unconditional power, $87.2 \%$, since the number of discordant pairs was allowed to vary. It also had a $60.8 \%$ chance of having conditional power less than $90 \%$. The large sample test under inverse sampling had $94.5 \%$ unconditional power, but an $18.8 \%$ chance of having less than $90 \%$ power. The standard deviations of the conditional power functions were comparable for the first 3 tests in the table, but was $0 \%$ for the exact test under inverse sampling.

Table 3. Descriptive Statistics for Conditional Power for the 4 Tests in the Example.

|  | Standard Sampling |  | Inverse Sampling |  |
| :--- | :---: | :---: | :---: | :---: |
| Statistic | Large Sample | Exact | Large Sample | Exact |
| Average or Unconditional Power (\%) | 92.3 | 87.2 | 94.5 | 92.8 |
| \% of tables with $<90 \%$ power | 31.9 | 60.8 | 18.8 | 0.0 |
| standard deviation (\%) | 4.4 | 7.5 | 5.0 | 0.0 |
| Total Sample Size (mean, std) | $111(0)$ | $111(0)$ | $92(23)$ | $92(23)$ |

## Empirical CDF of $\mathbf{N}$



Total Sample Size

Figure 5. Empirical distribution function of total sample size under inverse sampling

Another thing to consider for inverse sampling is the total sample size. In this example, the average sample size was $N=92$, which is a savings over standard sampling. There was a $79 \%$ chance that the sample size was less than 111. The median was $N=97$ with an interquartile range of 26 .

## 4. Summary

Inverse sampling can control the size of the critical region under the alternative for tables that may 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 fairly easy to apply well known procedures for the one sample binomial test based on the number of discordant pairs (Jennison \& Turnbull, 2000). 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 $\rho$, say $\rho \sim \operatorname{Beta}(\alpha, \beta)$ and $N_{c} \sim N B(r, \rho)$. 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 $\rho$ can be estimated under the proposed design. It's easy to show that the posterior distribution of $\rho$ given $n_{c}$ is also $\rho \mid n_{c} \sim \operatorname{Beta}\left(\alpha+r, \beta+n_{c}\right)$. One could compute the posterior probability that the constraint will be met, $P\left[\rho>\mid \Delta \| n_{c}\right]$, and stop for futility if that probability is small. 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 $\rho$ is contained in a binomial density. However, the posterior distribution would condition on a second random variable, $n_{d}$ in $\alpha^{*}=\alpha+n_{d}$. More work is needed to formulate something like this based on frequentist testing, but the principles would be similar.

## 5. Appendix

### 5.1 Factorization of the Joint Density for Standard Sampling

The sampling distribution is multinomial. The multinomial coefficient may be factored into 3 binomial coefficients by multiplying and dividing by $n_{d}!n_{c}!$,

$$
\begin{equation*}
\binom{n}{x_{10} x_{01} x_{11} x_{00}}=\binom{n_{d}}{n_{10}}\binom{n_{c}}{x_{11}}\binom{n}{n_{d}} \tag{3}
\end{equation*}
$$

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

$$
\begin{align*}
P\left[x_{10}, x_{11}, n_{c}\right]= & \binom{n_{d}}{x_{10}}\left(\frac{\rho+\Delta}{2 \rho}\right)^{x_{10}}\left(1-\frac{\rho+\Delta}{2 \rho}\right)^{n_{d}-x_{10}} \\
& \binom{n_{c}}{x_{11}}\left(\frac{p_{1}+p_{0} .-\rho}{2(1-\rho)}\right)^{x_{11}}\left(1-\frac{p_{1} \cdot+p_{0} \cdot-\rho}{2(1-\rho)}\right)^{n_{c}-x_{11}}  \tag{4}\\
& \binom{n}{n_{d}} \rho^{n_{d}}(1-\rho)^{n-n_{d}}
\end{align*}
$$

### 5.2 Derivation and Factorization of the Joint Density for Inverse Sampling

I attribute this method of proof to Günter Heimann and Mauro Gasparini (Heimann, 2015). 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 partition 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 probability function may now be written as

$$
\begin{align*}
P\left[x_{10}, x_{11}, n\right] & =p_{10}\binom{n-1}{n-r}\binom{n-r}{x_{11}} p_{11}^{x_{11}} p_{00}^{n-r-x_{11}}\binom{r-1}{x_{10}-1} p_{10}^{x_{10}-1} p_{01}^{r-x_{10}}  \tag{5}\\
& +p_{01}\binom{n-1}{n-r}\binom{n-r}{x_{11}} p_{11}^{x_{11}} p_{00}^{n-r-x_{11}}\binom{r-1}{x_{10}} p_{10}^{x_{10}} p_{01}^{r-1-x_{10}}
\end{align*}
$$

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

$$
\begin{align*}
P\left[x_{10}, x_{11}, n\right] & =\left[\binom{r-1}{x_{10}-1}+\binom{r-1}{x_{10}}\right]\binom{n-1}{n-r}\binom{n-r}{x_{11}} p_{11}^{x_{11}} p_{00}^{n-r-x_{11}} p_{10}^{x_{10}} p_{01}^{r-x_{10}} \\
& =\binom{n-1}{n-r}\binom{n-r}{x_{11}} p_{11}^{x_{11}} p_{00}^{n-r-x_{11}}\binom{r}{x_{10}} p_{10}^{x_{10}} p_{01}^{r-x_{10}} \tag{6}
\end{align*}
$$

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 $n_{c}=n-r$ :

$$
\begin{align*}
P\left[x_{10}, x_{11}, n_{c}\right]= & \binom{r}{x_{10}}\left(\frac{\rho+\Delta}{2 \rho}\right)^{x_{10}}\left(1-\frac{\rho+\Delta}{2 \rho}\right)^{r-x_{10}} \\
& \binom{n_{c}}{x_{11}}\left(\frac{p_{1 \cdot}+p_{0} \cdot-\rho}{2(1-\rho)}\right)^{x_{11}}\left(1-\frac{p_{1}+p_{0} \cdot-\rho}{2(1-\rho)}\right)^{n_{c}-x_{11}}  \tag{7}\\
& \binom{n_{c}+r-1}{n_{c}} \rho^{r}(1-\rho)^{n_{c}}
\end{align*}
$$

### 5.3 A Wald Confidence Interval for $\Delta$ Under Inverse Sampling

The delta rule can be used to derive a simple confidence interval on the treatment difference under inverse sampling. Recall that the estimate of the treatment difference under inverse sampling is $\hat{\Delta}=\left(2 x_{10}-r\right) /\left(N_{c}+r\right)$ which is a ratio of two independent random variables. The independence is apparent from equation 7 above since the density functions for $x_{10}$ and $n_{c}$ do not depend on each other. An estimate of the variance of the numerator is $V_{x}=4 r x_{10}\left(r-x_{10}\right)$ since $X_{10}$ is a binomial random variable. An estimate of the variance of the denominator is $V_{n}=\left(r+n_{c}\right) n_{c} / r$ since $N_{c}$ is a negative binomial random variable. The estimates are construced by replacing parameters in variances by their mle's. Hence, an estimate of the variance of $\hat{\Delta}$ is

$$
\begin{align*}
V(\hat{\Delta}) & =\frac{4 V_{x}}{\left(n_{c}+r\right)^{2}}+\frac{\left(2 x_{10}-r\right)^{2} V_{n}}{\left(n_{c}+r\right)^{4}} \\
& =\frac{4 r x_{10}\left(r-x_{10}\right)}{\left(n_{c}+r\right)^{2}}+\frac{\left(2 x_{10}-r\right)^{2} n_{c}}{\left(n_{c}+r\right)^{3} r} \tag{8}
\end{align*}
$$

The confidence interval is $\hat{\Delta} \pm z_{1-\alpha / 2} \sqrt{V}(\hat{\Delta})$.

## Acknowledgments

I would like to acknowledge Günter Heimann and Mauro Gasparini for their collaboration on another publication (Heiman$\mathrm{n}, 2015$ ) that produced the factorization method of proof for the inverse sampling density.
This work was presented by the author at the 2015 Joint Statistical Meetings in Seattle.

## References

Heimann, G., Von Tress, M., \& Gasparini, M. (2015). Exact and asymptotic inference in clinical trials with small event rates under inverse sampling. Statistics in Medicine, 34, 2708-2724. http://dx.doi.org/10.1002/sim. 6511
Connet, J. E., Smith, J. A. \& McHugh, R. B. (1987). Sample size and power for pair-matched case control studies. Statistics in Medicine, 6, 53-57 http://dx.doi.org/10.1002/sim. 4780060107
Jennison, C., \& Turnbull, B. W. (2000). Group Sequential Methods with Applications to Clinical Trails, New York, NY: Chapman \& Hall/CRC
Miettinen, O. S. (1968). The Matched pairs pesign in the case of all-or-none responses. Biometrics, 24, 339-352 http://dx.doi.org/10.2307/2528039
McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12, 153-157. http://dx.doi.org/10.1007/BF02295996
Newcombe R. G. (1998). Improved confidence intervals for the difference between binomial proportions based on paired data. Statistics in Medicine, 17, 2635-2650. http://dx.doi.org/10.1002/(SICI)1097-0258(19981130)17:22<2635::AID-SIM954>3.0.CO;2-C
nQuery Advisor 6.0 (2005). Statistical Solutions. Stonehill Corporate Center, Suite 104, 999 Broadway, Saugus, MA 01906, USA.

## Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.
This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).

