# The Cochran-Mantel-Haenszel test and the Lucia de Berk case

Richard D. Gill*

July 4, 2007

**Abstract**

The asymptotic power of the Mantel-Haenszel test is compared to that of its competitors.

## 1 Introduction

Suppose we compare a *treatment* with a *control* in each of $S$ subpopulations or *strata*, where the outcome is binary: *success* or *fail*. In an investigation of a suspected serial killer nurse, the experimental units might be the shifts at a number of different wards of various hospitals and/or in different time periods. The treatment is "suspect is on duty". The success outcome is "incident": a death or reanimation or other medical crisis occurs on the ward `;-)`. We are interested in whether the treatment increases the probability of success.

The situation of our nursing example is of course *not* a double-blind randomized clinical trial, so the model should not be taken too seriously. It can however still be interesting to detect departures from the model, for whatever reason they might occur—the suspect might have been assigned more often than other nurses to difficult shifts, for instance a fully-licensed nurse gets shifts which one would hesitate to assign to a trainee-nurse; the suspect may tend to have more shifts in busy periods when all the beds on the ward are occupied, than in quiet periods where he is sent on a training course or takes vacation; the suspect might be a better nurse so that he alerts the emergency team when a more lazy nurse would not notice anything, with the result that an incident starts on his shift instead of the next one or not at all; he might just have bad luck to be on sick leave during a quiet period, fit and on duty and catching up with overtime during a busy period.

*Mathematical Institute, Leiden University

Suppose for $i = 1, \ldots, S$ that

$$\begin{aligned} X_i &\sim \text{Bin}(m_i, p_i) \\ Y_i &\sim \text{Bin}(n_i, q_i), \end{aligned} \tag{1}$$

all independent of one another. $X_i$ and $Y_i$ stand for the numbers of incidents in the suspect's $m_i$ shifts and the remaining $n_i$ shifts, respectively, in stratum (ward/hospital/time period) $i$. Define

$$\begin{aligned} \widehat{p_i} &= X_i/m_i, \\ \widehat{q_i} &= Y_i/n_i, \\ N_i &= m_i + n_i, \\ N &= \sum_i N_i. \end{aligned} \tag{2}$$

The Cochran-Mantel-Haenszel test of the null-hypothesis $p_i = q_i$ for all $i$ against the alternative $p_i \geq q_i$ for all $i$, with at least one inequality strict, is the one-sided test based on the weighted and normalized sum of estimated differences $p_i - q_i$

$$T_{\text{MH}} = N^{-\frac{1}{2}} \sum_{i=1}^{S} \frac{m_i n_i}{m_i + n_i} (\widehat{p_i} - \widehat{q_i}). \tag{3}$$

Up to a normalization constant, $T_{\text{MH}}$ is the score test statistic of the null-hypothesis $\rho = 1$ in the popular model of a common odds ratio

$$\frac{p_i/(1 - p_i)}{q_i/(1 - q_i)} = \rho \qquad \text{for all } i \tag{4}$$

and working with the conditional distribution of the data given the "column totals" $X_i + Y_i$. The alternative hypothesis is $\rho > 1$. Under this conditioning, and under the assumption of a constant odds ratio $\rho$, the distribution of the data only depends on $\rho$. The constant odds ratio model, also known as the proportional odds model, is not only attractive from a subject-matter point of view but also from a mathematical point of view, since it gives us an exponential family and consequently inferential problems with very nice properties.

One may display the data as a collection of 2 × 2 tables, see Table 1. We will also use an alternative notation where the $S$ strata are labelled with indices $s$, the $(i, j)$'th element of the $s$'th 2 × 2 table is denoted by $n_{sij}$, and addition over row or column elements is indicated by a dot, see Table 2.

Because the test is the score test in a one-parameter model, it is (conditionally) locally uniformly most powerful for testing the null hypothesis $\rho = 1$ against the alternative $\rho > 1$. Usually one makes a minor "continuity correction" to $T_{\text{MH}}$ and

Table 1: Data for stratum $i$

| | | |
|---|---|---|
| $X_i$ | $m_i - X_i$ | $m_i$ |
| $Y_i$ | $n_i - Y_i$ | $n_i$ |
| $X_i + Y_i$ | $(m_i + n_i) - (X_i + Y_i)$ | $m_i + n_i$ |

Table 2: Alternative notation, data for stratum $s$

| | | |
|---|---|---|
| $n_{s11}$ | $n_{s12}$ | $n_{s1\cdot}$ |
| $n_{s21}$ | $n_{s22}$ | $n_{s2\cdot}$ |
| $n_{s\cdot 1}$ | $n_{s\cdot 2}$ | $n_{s\cdot\cdot}$ |

divides it by the square root of the null-hypothesis uniform minimum variance unbiased estimate of its unconditional null-hypothesis variance, and compares it (one-sidedly) to the standard normal distribution. In principle however one can compare it to its null-hypothesis conditional distribution, which is of course just a permutation distribution: permutations of the assignments of treatment ("suspect is on duty") to experimental units (shifts) within strata, keeping the outcomes fixed. The just-mentioned unbiased estimate of the null hypothesis unconditional variance is constant under the permutation distribution and hence identical to the variance of the null hypothesis permutation distribution.

Under the null-hypothesis, and conditional on the column totals, the $X_i$ are independent and hypergeometrically distributed so that the test becomes an extension to the case of multiple independent $2 \times 2$ tables of Fisher's exact test for independence in a single $2 \times 2$ contingency table.

Mantel and Haenszel's contribution was to introduce modifications to a test statistic earlier proposed by Cochran, so as to extend good properties of the test to the situation where the number of strata is large, while the stratum sample sizes $N_i = m_i + n_i$ are very small. Their modification, in its two-sided form (the statistic should be compared to the chi-square distribution with one degree of freedom) and our alternative notation,

$$\chi^2_{\mathrm{MH}} \;=\; \frac{\left(\left|\sum_s \left(n_{s11} - \frac{n_{s\cdot 1}n_{s1\cdot}}{n_{s\cdot\cdot}}\right)\right| - \frac{1}{2}\right)^2}{\sum_s \frac{n_{s\cdot 1}\,n_{s\cdot 2}\,n_{s1\cdot}\,n_{s2\cdot}}{n^2_{s\cdot\cdot}(n_{s\cdot\cdot} - 1)}},$$

has excellent properties when $N = \sum_i N_i \to \infty$, independently of whether the number of the strata or the size of each stratum increases to infinity. The test is an ubiquitous workhorse in epidemiology, particularly in "matched pair" retrospective case-control studies, when each $m_i = n_i = 1$. Another common application (or rather, extension) is in survival analysis, with one stratum per failure time (the survivors up to that time instant), and $X_i + Y_i = 1$ for all $i$, when the statistic becomes the well known log rank test of Peto and Peto, also known as the score test of a proportional hazards alternative in the Cox regression model. Mantel and Haenszel's insight that one could formally use the analysis of a collection of independent $2 \times 2$ tables in this context, was a stroke of brilliance, and it had enormous impact.

In the alternative formulation, but going back to our original notation, we see that the test compares the total number of "incidents" $\sum X_i$ to the null hypothesis expected number, $\sum \frac{X_i+Y_i}{m_i+n_i} m_i$, using the permutation distribution to get the p-value. But the expected number is constant under permutations (or: constant, conditional on $X_i + Y_i$). Thus the test is nothing else than a permutation based test based on the statistic $\sum X_i$. The null hypothesis distribution respects the stratification by considering permutations within strata only.

Mantel and Haenszel's modifications to the Cochran statistic were (1) the use of the conditional permutation variance, which ensures good properties when one has many small strata, and (2) the introduction of the "continuity correction" $\frac{1}{2}$. We do not need to take account of either of these technical innovations in this study. The relation between the two forms of the statistic given above will be further elucidated, in the section after the next.

## 2 Asymptotic theory

In this and the next section I will compare the power of the Mantel-Haenszel test against that of various natural competitors in various situations, where I keep the number of strata fixed but let their size grow indefinitely. In order to obtain an interesting comparison I suppose that the larger the sample size, the smaller the effect of the treatment; so called-local alternatives. Effectively I will be computing the Pitman asymptotic relative efficiency between the different tests. The different situations are: (i) the $q_i$ are arbitrary, and in general all different; (ii) the $q_i$ are all equal. The statistician may or may not believe she is in situation (i) or (ii), and she may or may not be correct in her belief. In the situation when the $q_i$ are different but the statistician acts as though they are equal, I will again suppose that the larger the sample size, the smaller is the difference, in order to obtain an illuminating comparison. The asymptotics are not supposed to imply that in the real world, as we gather more data, effects actually get smaller in such a peculiar fashion. They simply serve to obtain informative approximations in the situation that sample sizes are large, so asymptotic approximations are good, but effects are relatively small, so there is an appreciable difference between the powers of more or less sensible test statistics.

Suppose that as $N \to \infty$, the $q_i$ are fixed but $p_i$ depends on $N$ via the restriction

$$\rho = 1 + \delta/\sqrt{N}. \tag{5}$$

(More generally, for each $i$, both $p_i$ and $q_i$ depend on $N$, but are related to one another through the odds ratio $\rho$; and they converge to some limit as $N \to \infty$). Suppose that as $N \to \infty$

$$\frac{m_i}{N_i} \to \alpha_i, \quad \frac{n_i}{N_i} \to \beta_i, \quad \frac{N_i}{N} \to \mu_i. \tag{6}$$

Note that $\sum_i \mu_i = 1$ and $\alpha_i + \beta_i = 1$ for all $i$. Note also that

$$\begin{aligned} \frac{p}{1-p} &= (1+\delta)\frac{q}{1-q} \\ \Longrightarrow \quad p(1-q) &= (1+\delta)q(1-p) \\ \Longrightarrow \quad p(1-q+q+\delta q) &= (1+\delta)q \\ \Longrightarrow \quad p &= \frac{(1+\delta)q}{1+\delta q} = q + q(1-q)\delta + \mathscr{O}(\delta^2). \end{aligned} \tag{7}$$

Using the Bernoulli-de Moivre normal limit to the binomial, one discovers that, in distribution,

$$T_{\mathrm{MH}} \to \mathscr{N}\Big(\sum \mu_i \alpha_i \beta_i q_i (1-q_i)\, \delta, \ \sum \mu_i \alpha_i \beta_i q_i (1-q_i)\Big). \tag{8}$$

The quality of the test is measured by $\sum \mu_i \alpha_i \beta_i q_i (1-q_i)$ (the larger, the better). The form of the limiting distribution comes from its likelihood derivation and the fact that, given $S$ independent $\mathscr{N}(\sigma_i^2 \delta, \sigma_i^2)$ observations, the optimal test of $\delta = 0$ is based on their sum.

Suppose that all $q_i$ are equal, and hence (because of proportional odds) all $p_i$ are equal. In that situation one could merge the $S$ strata into one. The Mantel-Haenszel test with a single stratum is asymptotically the same as the one-sided chi-square test, or the Fisher exact test for one $2 \times 2$ table (for more details on this, see the next section). One finds that the asymptotic distribution of the pooled-data statistic is $\mathscr{N}(\overline{\alpha}\overline{\beta}q(1-q)\delta, \overline{\alpha}\overline{\beta}q(1-q))$, and the quality of the test is measured by $\overline{\alpha}\overline{\beta}q(1-q)$, where $\overline{\alpha} = \sum_i \mu_i\alpha_i$, $\overline{\beta} = \sum_i \mu_i\beta_i$, and $q = q_i$ for all $i$. When both are legitimate, the stratified Mantel-Haenszel test has lower power than the pooled one-stratum test except when $\alpha_i = \alpha$, $\beta_i = \beta$ for all $i$. In this case, and only then, $\overline{\alpha}\overline{\beta}q(1-q) = \overline{\alpha\beta}q(1-q)$ where $\overline{\alpha\beta} = \sum_i \mu_i\alpha_i\beta_i$. Thus, *if* merging is legitimate, the Mantel-Haenszel test loses power, *except* when the design is "balanced": the treatment is applied to the same proportion of experimental units in each stratum.

If Lucia de Berk is a full time nurse at all different wards/time-periods, one can expect the balance condition, $\alpha_i$ independent of $i$, to hold, i.e., she will be on duty on the same proportion of shifts in all strata, since she will work on average the same number of hours per year in each ward/time-period. Thus we can expect not to lose power by proper stratification properly, even if it is unncessary. However the $q_i$ might well be different if we are pooling different kinds of wards, or if due to changing situations in the hospital, the incident rate varies over time. In that case the "all-strata-pooled" statistic is illegitimate. We investigate what can go wrong in this situation, in the next section.

On the other hand, if we stratify very finely by time period, because we suspect strong time variation in the rate of incidents, and if Lucia de Berk works strongly different proportions of time in different time periods (because of holidays, leave for following a course, sick leave) *and* if we are wrong in our suspicion of time-variation in the incident rate, then we will lose power compared to the test which we could have used.

Finally, the asymptotic theory we give here depends on the sample size per stratum growing large, and it does not give a good picture in the case of very many small strata, when a different kind of asymptotics should be followed.

# 3 Comparison with competing testing methods

In the previous section I compared Mantel-Haenzel to pooled chi-square in the situation that the pooled statistic is legitimate. We saw that asymptotic power was lost, except in the case of a "balanced design", all $\alpha_i$ equal. That is actually good news since the balance condition will often be at least approximately true. Now I allow for different $q_i$ over the strata. I compare (1) the Fisher combination method, (2) the standard chi-square test for stratified $2 \times 2$ tables, (3) quick and dirty pooling by adding, even if the $q_i$ are different.

In the usual chi-squared test for a 2 × 2 table, the quantity "observed minus expected", $O - E$, in each of the four cells of the table is equal, up to a sign, since row and column sums are reproduced exactly by the "expected" counts. On the other hand,

$$\begin{aligned}\widehat{p} - \widehat{q} &= \frac{X}{m} - \frac{Y}{n} \\ &= \frac{X}{m} + \frac{X}{n} - \frac{X+Y}{m+n}\frac{m+n}{n} \\ &= \frac{m+n}{mn} X - \frac{m+n}{mn}\frac{m}{m+n}(X+Y) \\ &= \frac{m+n}{mn}(O-E).\end{aligned} \tag{9}$$

We see that both Mantel-Haenszel test and the various chi-quare tests are built up from exactly the same quanties $\widehat{p} - \widehat{q}$, in some cases based on the pooled data, in other cases per stratum. The difference lies in *at what stage*, and *how* the strata are pooled.

## 3.1 Fisher combination method

The Fisher method is based on the sum of the logarithms of the $p$-values of the chi-square tests for each table separately (in their one-sided versions). Now the $p$-value of a one-sided test based on a statistic $T_i \sim \mathscr{N}(\sigma_i^2\delta, \sigma_i^2)$, is distributed as $1-\Phi^{-1}(\delta\sigma_i + Z)$ where $Z$ is standard normally distributed. Hence asymptotically the Fisher combination statistic $-2\sum_i \log(p\text{-value in stratum } i)$, to be compared with the $\chi^2_{2k}$ distribution, is asymptotically equivalent to $-2\sum_i \log(1-\Phi^{-1}(Z_i + \delta_i\sigma_i))$ where the $Z_i$ are independent standard normally distributed. Note that minus twice the logarithm of a uniform distributed random variable is $\chi^2_2$-distributed. There is not much that can be said about this method, except that it is less powerful against the proportional odds alternative than the Mantel-Haenszel test, which is based on $\sum T_i$, and which gives the uniformly most powerful test. The Fisher combination method combines the *wrong* function of each statistic, and does this moreover with the *wrong* weights. This problem does not go away if we consider different one-parameter families of alternatives to the proportional odds model. The strata are combined as if each stratum is equally important, but the relative weight to be given to each stratum will depend on the stratum parameters and relative sample sizes.

Essentially, the Fisher combination method is a method of last resort, when we know nothing at all about the relation between the different statistics on which the different p-values are to be compared. We just have the p-values and know nothing else. Then there is not much more that we can do ...

## 3.2 Stratified chi-squared test

The usual chi-sqare test based on summing a chi-squared (one degree of freedom) test statistic for each separate stratum, is asymptotically equivalent to a test of $\delta_i = 0$ for all $i$ given $T_i \sim \mathcal{N}(\sigma_i^2\delta_i, \sigma_i^2)$. The test is based on the sum of the squares of the $T_i$. This is an omnibus test, having some sensitivity to all possible departures from the null hypothesis. Implicitly it weights the strata so as to maximize the minimum loss of power against alternatives with equal envelope power function: it a so-called most stringent test (minimizes the maximum shortcoming, pointwise over the alternative, relative to the test of maximal power against each point in the alternative, separately). Clearly it will have severely reduced power against most alternatives of interest which are at the least "one-sided"( i.e., we are only interested in $\delta_i \geq 0$), if not one-dimensional (all $\delta_i$ equal).

## 3.3 Chi-squared test based on pooled data

Suppose now we pool the strata, but

$$p_i = q + \frac{\theta_i}{\sqrt{N}} + \frac{q(1-q)}{\sqrt{N}}\delta + \mathcal{O}(N^{-1})$$

and

$$q_i = q + \frac{\theta_i}{\sqrt{N}} + \mathcal{O}(N^{-1}),$$

where the $\theta_i$ are arbitrary. The mean of $\sqrt{N}(\sum X_i/\sum m_i - \sum Y_i/\sum n_i)$ is asymptotically

$$\begin{aligned}
&\frac{\sum m_i\theta_i}{\sum m_i} + q(1-q)\delta - \frac{\sum n_i\theta_i}{\sum n_i} \\
&\quad = q(1-q)\delta + \frac{\sum(m_i/N_i)(N_i/N)\theta_i}{\sum(m_i/N_i)(N_i/N)} - \frac{\sum(n_i/N_i)(N_i/N)\theta_i}{\sum(n_i/N_i)(N_i/N)} \\
&\quad = q(1-q)\delta + \frac{\sum\alpha_i\mu_i\theta_i}{\sum\alpha_i\mu_i} - \frac{\sum\beta_i\mu_i\theta_i}{\sum\beta_i\mu_i} + \mathcal{O}(1).
\end{aligned}$$

Recall that $\sum_i \mu_i = 1$, and $\alpha_i + \beta_i = 1$ for all $i$. Write $\overline{\alpha} = \sum_i \mu_i\alpha_i$, $\overline{\alpha\theta} = \sum_i \mu_i\alpha_i\theta_i$, $\overline{\beta} = \sum_i \mu_i\beta_i$, $\overline{\beta\theta} = \sum_i \mu_i\beta_i\theta_i$, $\overline{\theta} = \sum_i \mu_i\theta_i$. Then we can write the asymptotic mean of our test statistic as

$$
\begin{aligned}
& q(1-q)\delta + \frac{\overline{\beta}\,\overline{\alpha\theta} - \overline{\alpha}\,\overline{\beta\theta}}{\overline{\alpha}\,\overline{\beta}} \\
&= q(1-q)\delta + \frac{\sum_i \mu_i(\overline{\beta}\alpha_i - \overline{\alpha}\beta_i)\theta_i}{\overline{\alpha}\,\overline{\beta}} \\
&= q(1-q)\delta + \frac{\sum_i \mu_i\theta_i\Big((1-\overline{\alpha})\alpha_i - \overline{\alpha}(1-\alpha_i)\Big)}{\overline{\alpha}\,\overline{\beta}} \\
&= q(1-q)\delta + \frac{\sum_i \mu_i(\theta_i - \overline{\theta})(\alpha_i - \overline{\alpha})}{\overline{\alpha}\,\overline{\beta}}
\end{aligned}
$$

Even if $\delta = 0$ we can have an asymptotic mean unequal to 0 and hence an incorrect test. This is Simpson's paradox raising its ugly head. If all the $\alpha_i$ are equal, the situation which we call a balanced design, then the asymptotic mean is $q(1-q)\delta$, independent of the $\theta_i$. The "dangerous situation" occurs when there is a positive correlation between the $\alpha_i$ and the $\theta_i$, i.e., the treatment is applied relatively more often in those strata where the probability of success is larger.

In terms of our nurse: Lucia works relatively more hours on wards/time periods where the incident rate is higher.

# 4 Conclusion

The Mantel-Haenszel test has been the work-horse of epidemiological and medical statistical research for 30 years. It is incredibly simple and intuitive and has wonderful properties. That no single statistician ever thought of using it in the Lucia de Berk case is a sorry reflection on the amateurism which afflicted the use not only of statistical but of all scientific evidence in this sorry case, in which the *only* evidence was indirect, complex, scientific evidence. The amateurism of the scientists involved was compounded by the fact that the interdisciplinary combination of the evidence was a task solely borne by the judges. No single expert witness took any interest whatsoever, in observing how his or her expertise was used by the judges. In fact they all took pride in delegating this task entirely to the wise judges. Those judges became convinced by amateur psychology and amateur statistics that Lucia de Berk was a monster and an evil killer. The scientific experts willingly supplied the legal facts necessary in order to put her away.

# 5 References

To be completed:  Mantel-Haenszel, Cochran, Peto and Peto, Cox, Fisher, Simpson paradox