On a Null Variance Estimator for the Mantel-Haenszel Risk Difference

Bernhard Klingenberg, Williams College, MA

Several 2x2 tables:

Table i:

	S	F	
Group 1	y_{i1}		n_{i1}
Group 2	y_{i2}		n_{i2}

Risk difference: $\delta = \pi_{i1} - \pi_{i2}$

Mantel-Haenszel estimator for common risk difference in several 2x2 tables:

$$\hat{\delta}_{MH} = \frac{\sum_{i=1}^{K} w_i (y_{i1}/n_{i1} - y_{i2}/n_{i2})}{\sum_{i=1}^{K} w_i} = \frac{\sum_{i=1}^{K} (n_{i2} y_{i1} - n_{i1} y_{i2})/n_{i+}}{\sum_{i=1}^{K} w_i}$$

Variance estimator?

- **Greenland-Robins**: Plug in sample proportions (default in software)
- **Sato (1989)**: under homogeneity $\mathrm{Var}[n_{i2} y_{i1} - n_{i1} y_{i2}] = E[\delta P + Q]$, so $\widehat{\mathrm{Var}}[\hat{\delta}_{\mathrm{MH}}] = \left(\hat{\delta}_{\mathrm{MH}} P + Q\right) / W^2$
- **New idea:** under homogeneity + null: $\widehat{\mathrm{Var}}_{\delta_0}[\hat{\delta}_{\mathrm{MH}}] = (\delta_0 P + Q)/W^2$

$P = \sum_i P_i$, $Q = \sum_i Q_i$ and $W = \sum_i w_i$

$$P_i = \frac{n_{i1}^2 y_{i2} - n_{i2}^2 y_{i1} + n_{i1} n_{i2} (n_{i2} - n_{i1})/2}{n_{i+}^2}$$

$$Q_i = \frac{y_{i1}(n_{i2} - y_{i2}) + y_{i2}(n_{i1} - y_{i1})}{2 n_{i+}}.$$

$$w_i = n_{i1} n_{i2} / n_{i+}$$

Sato interval (Wald-type interval ☹):

Invert $T = \dfrac{(\hat{\delta}_{MH} - \delta_0)^2}{\widehat{\mathrm{Var}}[\hat{\delta}_{MH}]}$

which leads to $\hat{\delta}_{\mathrm{MH}} \pm z_{\alpha/2} \sqrt{\widehat{\mathrm{Var}}[\hat{\delta}_{\mathrm{MH}}]}$

New interval:

Invert $T_0 = \dfrac{(\hat{\delta}_{MH} - \delta_0)^2}{\widehat{\mathrm{Var}}_{\delta_0}[\hat{\delta}_{MH}]}$

which leads to $\hat{\delta}_{\mathrm{Mid}} \pm \mathrm{ME}$ with

$$\hat{\delta}_{\mathrm{Mid}} = \hat{\delta}_{\mathrm{MH}} + 0.5 z_{\alpha/2}^2 (P/W^2)$$

$$\mathrm{ME} = \sqrt{\hat{\delta}_{\mathrm{Mid}}^2 - \hat{\delta}_{\mathrm{MH}}^2 + z_{\alpg/2}^2 (Q/W^2)}$$

```
> ## Myeloma data
> head(myeloma)
  y1 n1 y2 n2
1  3  4  1  3
2  3  4  8 11
3  2  2  2  3
...
```

Score interval (Miettinen & Nurminen):

Based on inverting score test statistic

$$S_{\mathrm{MN}} = \frac{\left(\sum_i \sum_j c_{ij} \tilde{\pi}_{ij} - \delta_0\right)^2}{\sum_i \left(\frac{n_{i+}}{n_{i+}-1}\right) \sum_j c_{ij}^2 \tilde{\pi}_{ij}(1 - \tilde{\pi}_{ij})/n_{ij}}$$

(no closed from for endpoints)

$\tilde{\pi}_{ij}$ are the restricted MLEs

$c_{i1} = w_i^*$ and $c_{i2} = -w_i^*$

$w_i^* = w_i / \sum_l w_l$

```
> strat.MHRD(myeloma) #new
$delta.MH
[1] 0.0572
$delta.Mid
[1] 0.05467
$pseudo.se
[1] 0.04076
$CI
[1,] -0.1019 0.2113
```

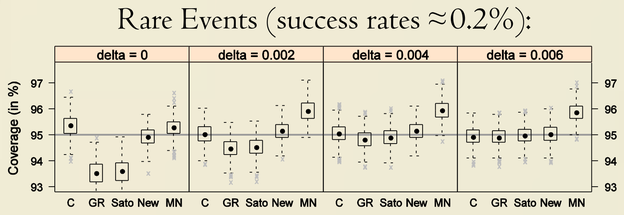
```
> strat.MHRD(myeloma, method="GR")
$delta.MH
[1] 0.0572
$se.delta.MH
[1] 0.06319
$CI
[1] -0.0667  0.1810
```

```
> strat.MHRD(myeloma, method="Sato")
$delta.MH
[1] 0.0572
$se.delta.MH
[1] 0.07989
$CI
[1] -0.0994  0.2137
```

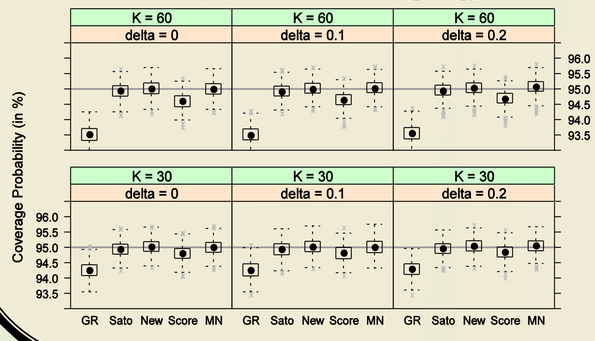
```
> #Miettinen-Nurminen
> strat.MHRD.MN(myeloma)
$delta.MH
[1] 0.0572
$CI
[1] -0.0997  0.2065
```

Simulations:

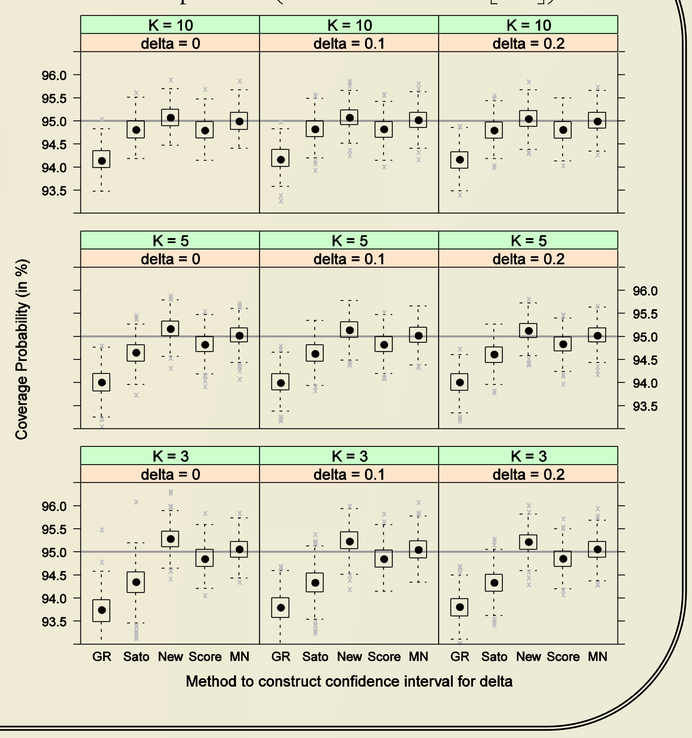
Rare Events (success rates ≈0.2%):



K strata with both large and small sample sizes (success rates ~ U[0,1]):



K strata with fixed but unbalanced (3:1) sample size (success rates ~ U[0,1]):



Conclusion:

Confidence interval based on new variance estimator has several desirable properties:

1. Leads to closed-form interval
2. Easy formula
3. Asymetric
4. Applies to matched pairs (as good as complicated Tango interval)
5. Excellent coverage performance for
   a. sparse data, including very rare events
   b. small sample sizes
   c. situations with both very large and small strata
   d. large or small number of strata

But: Does not work under substantial heterogeneity

Contact the author at: bklingen@williams.edu