Odd Ration Comparison Coarse Solution

Definition 1: Odds is a statistic calculated by dividing the probability of success by the probability of failures.

Definition 2: An odds ratio (OR) is a statistic that quantifies the strength of the association between two events, A and B. The odds ratio is defined as the ratio of the odds of A in the presence of B and the odds of A in the absence of B, or equivalently (due to symmetry), the ratio of the odds of B in the presence of A and the odds of B in the absence of A. (Wikipedia)

Theorem 1: Let o be an odds ratio that is a realization of the odds random variable O. Then O follows a Normal asymptotic distribution

$$\log(O) \sim N\left(\log(o), \sigma^2_{\log(o)}\right)$$

, where an estimate of $\sigma^2_{\log(o)}$ is that

$$\sigma^2_{\log(o)} \approx \left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right)$$

, where n_{ij} is the entry of the contingency table at the $i'th$ row, $j'th$ column. It means that when the sample size goes to infinity, $\log(O)$ converges to $N\left(\log(o), \sigma^2_{\log(o)}\right)$.

(Resources: https://stats.stackexchange.com/questions/1455/what-is-the-distribution-of-or-odds-ratio)

Theorem 2: If $X \sim N(x, \sigma_x^2), Y \sim N(y, \sigma_y^2)$, then the sum of the two random variables is $Z = X + Y \sim N\left(x + y, \sigma_x^2 + \sigma_y^2 + 2cov(X,Y)\right)$.

(Resources: https://math.stackexchange.com/questions/2252244/sum-of-two-dependent-normally-distributed-variables)

Corollary 1: For any two odds ratios o_1 and o_2 coming from the odds random variables O_1 and O_2, we may assume that

$$\log\left(\frac{O_1}{O_2}\right) \sim N\left(\log\left(\frac{o_1}{o_2}\right), \sigma^2_{\log(o_1)} + \sigma^2_{\log(o_2)} - cov(\log(O_1), \log(O_2))\right)$$

Proof:

By Theorem 1, $\log(O_1) \sim N\left(\log(o_1), \sigma^2_{\log(o_1)}\right)$ and $\log(O_2) \sim N\left(\log(o_2), \sigma^2_{\log(o_2)}\right)$ asymptotically.

We assume that $\log(O_1) \sim N\left(\log(o_1), \sigma^2_{\log(o_1)}\right)$ and $\log(O_2) \sim N\left(\log(o_2), \sigma^2_{\log(o_2)}\right)$ for sufficiently large sample size.

Then, by Theorem 2,

$$\log\left(\frac{O_1}{O_2}\right) = [\log(O_1) - \log(O2)] \sim \left[N\left(\log(o_1), \sigma^2_{\log(o_1)}\right) - N\left(\log(o_2), \sigma^2_{\log(o_2)}\right)\right]$$

$$= N\left(\log(o_1) - \log(o_2), \sigma^2_{\log(O_1)} + \sigma^2_{\log(O_2)}\right.$$

$$\left. + cov(\log(O_1), -\log(O_2))\right)$$

$$= N\left(\log\left(\frac{o_1}{o_2}\right), \sigma^2_{\log(o_1)} + \sigma^2_{\log(o_2)} - cov(\log(O_1), \log(O_2))\right)$$

Corollary 2: Given two odds o_1 and o_2 from the following two-by-two tables:

	yes	no
yes	n^1_{11}	n^1_{12}
no	n^1_{21}	n^1_{22}

	yes	no
yes	n^2_{11}	n^2_{12}
no	n^2_{21}	n^2_{22}

, and let O_1, O_2 be the random variables from which o_1 and o_2 come from, then the Maximum Likelihood Estimate (MLE) of $E\left[\log\left(\frac{O_1}{O_2}\right)\right]$ is just $\log\left(\frac{o_1}{o_2}\right)$, and the MLE of $Var\left[\log\left(\frac{O_1}{O_2}\right)\right]$ is **at most**

$$\sigma^2_{\log(O_1)} + \sigma^2_{\log(O_2)} + \sqrt{\sigma^2_{\log(O_1)} + \sigma^2_{\log(O_2)} + 2\sigma_{\log(O_1)}\sigma_{\log(O_2)}}$$

, where $\sigma^2_{\log(O_1)} = \left(\frac{1}{n^1_{11}} + \frac{1}{n^1_{12}} + \frac{1}{n^1_{21}} + \frac{1}{n^1_{22}}\right)$, $\sigma^2_{\log(O_2)} = \left(\frac{1}{n^2_{11}} + \frac{1}{n^2_{12}} + \frac{1}{n^2_{21}} + \frac{1}{n^2_{22}}\right)$.

Proof:

We let o_1 be a realization of an odd random variable O_1, let o_2 be a realization of an odd random variable O_2.

Then, by Corollary 3, we can assume that $\log\left(\frac{O_1}{O_2}\right) \sim N\left(\log\left(\frac{o_1}{o_2}\right), \sigma^2_{\log(O_1)} + \right.$

$\left.\sigma^2_{\log(O_2)} - cov(\log(O_1), \log(O_2))\right)$.

Then, the Maximum Likelihood Estimate (MLE) of $E\left[\log\left(\frac{O_1}{O_2}\right)\right]$ is just $\log\left(\frac{o_1}{o_2}\right)$, while the MLE of $Var\left[\log\left(\frac{O_1}{O_2}\right)\right]$ is calculated by

$$Var\left[\log\left(\frac{O_1}{O_2}\right)\right] = \sigma^2_{\log(O_1)} + \sigma^2_{\log(O_2)} + cov(\log(O_1), \log(O_2))$$

By Theorem 1,

$$\sigma^2_{\log(O_1)} = \left(\frac{1}{n^1_{11}} + \frac{1}{n^1_{12}} + \frac{1}{n^1_{21}} + \frac{1}{n^1_{22}}\right)$$

$$\sigma^2_{\log(O_2)} = \left(\frac{1}{n^2_{11}} + \frac{1}{n^2_{12}} + \frac{1}{n^2_{21}} + \frac{1}{n^2_{22}}\right)$$

$$cov(\log(O_1), \log(O_2)) = \sqrt{\sigma^2_{\log(O_1)} + \sigma^2_{\log(O_2)} + 2\rho\sigma_{\log(O_1)}\sigma_{\log(O_2)}}$$

, where ρ is the correlation coefficient between $\log(O_1)$ and $\log(O_2)$.
Since $0 \leq \rho \leq 1$, we know that

$$cov(\log(O_1), \log(O_2)) = \sqrt{\sigma^2_{\log(O_1)} + \sigma^2_{\log(O_2)} + 2\rho\sigma_{\log(O_1)}\sigma_{\log(O_2)}}$$

$$\leq \sqrt{\sigma^2_{\log(O_1)} + \sigma^2_{\log(O_2)} + 2\sigma_{\log(O_1)}\sigma_{\log(O_2)}}$$

$$= \sqrt{\left(\frac{1}{n^1_{11}} + \frac{1}{n^1_{12}} + \frac{1}{n^1_{21}} + \frac{1}{n^1_{22}}\right) + \left(\frac{1}{n^2_{11}} + \frac{1}{n^2_{12}} + \frac{1}{n^2_{21}} + \frac{1}{n^2_{22}}\right) + 2\sqrt{\left(\frac{1}{n^1_{11}} + \frac{1}{n^1_{12}} + \frac{1}{n^1_{21}} + \frac{1}{n^1_{22}}\right)}\sqrt{\left(\frac{1}{n^2_{11}} + \frac{1}{n^2_{12}}}}$$

Therefore,

$$Var\left[\log\left(\frac{O_1}{O_2}\right)\right]$$

$$= \left(\frac{1}{n^1_{11}} + \frac{1}{n^1_{12}} + \frac{1}{n^1_{21}} + \frac{1}{n^1_{22}}\right) + \left(\frac{1}{n^2_{11}} + \frac{1}{n^2_{12}} + \frac{1}{n^2_{21}} + \frac{1}{n^2_{22}}\right) + cov(\log(O_1), \log(O_2))$$

$$\leq \left(\frac{1}{n^1_{11}} + \frac{1}{n^1_{12}} + \frac{1}{n^1_{21}} + \frac{1}{n^1_{22}}\right) + \left(\frac{1}{n^2_{11}} + \frac{1}{n^2_{12}} + \frac{1}{n^2_{21}} + \frac{1}{n^2_{22}}\right)$$

$$+ \sqrt{\left(\frac{1}{n^1_{11}} + \frac{1}{n^1_{12}} + \frac{1}{n^1_{21}} + \frac{1}{n^1_{22}}\right) + \left(\frac{1}{n^2_{11}} + \frac{1}{n^2_{12}} + \frac{1}{n^2_{21}} + \frac{1}{n^2_{22}}\right) + 2\sqrt{\left(\frac{1}{n^1_{11}} + \frac{1}{n^1_{12}} + \frac{1}{n^1_{21}} + \frac{1}{n^1_{22}}\right)}\sqrt{\left(\frac{1}{n^2_{11}} + \frac{1}{n^2_{12}}}}$$

For convenience, say

$$Var\left[\log\left(\frac{O_1}{O_2}\right)\right] \leq \sigma^2_{\log(O_1)} + \sigma^2_{\log(O_2)} + \sqrt{\sigma^2_{\log(O_1)} + \sigma^2_{\log(O_2)} + 2\sigma_{\log(O_1)}\sigma_{\log(O_2)}}$$

, where $\sigma^2_{\log(O_1)} = \left(\frac{1}{n^1_{11}} + \frac{1}{n^1_{12}} + \frac{1}{n^1_{21}} + \frac{1}{n^1_{22}}\right)$, $\sigma^2_{\log(O_2)} = \left(\frac{1}{n^2_{11}} + \frac{1}{n^2_{12}} + \frac{1}{n^2_{21}} + \frac{1}{n^2_{22}}\right)$.

Based on corollary 2, we can implement an algorithm to calculate the widest 95% confidence interval of $\frac{O_1}{O_2}$ for sufficiently large sample size.