The objective is to estimate the average prevalence of correlated binary outcomes. Specifically, assuming that there are n patients and there are 3 exchangeable binary outcomes per patient denoted by $(Y_{i1}, Y_{i2}, Y_{i3}), i = 1, \cdots, n$, the parameter of interest is $P(Y_{ij} = 1) = \theta_0$. We want to estimate this parameter and construct its 95% confidence interval. The observed data can be summarized by (n_0, n_1, n_2, n_3), where n_j is the number of patients with exactly j positive responses. Under the exchangeability assumption:do we actually need the exchangeability assumption? The vectors are IID, so their sums are IID with sample space {0,1,2,3}, which we treat as the categories of a multinomial

$$(n_0, n_1, n_2, n_3) \sim MN(n, \boldsymbol{p_0}),$$

where $\boldsymbol{p_0} = (p_{00}, p_{10}, p_{20}, p_{30})$. Under this model, the parameter of interest

$$\theta_0 = \frac{1}{3}p_{10} + \frac{2}{3}p_{20} + \frac{3}{3}p_{30}.$$

To this end, we first consider a test statistic for testing the null hypothesis $H_0: \theta_0 = \theta$:

$$T(\theta) = \frac{\left|\frac{1}{3}\hat{p}_1 + \frac{2}{3}\hat{p}_2 + \frac{3}{3}\hat{p}_3 - \theta\right|}{\sqrt{\frac{1}{9}\tilde{p}_1^2 + \frac{4}{9}\tilde{p}_2^2 + \frac{9}{9}\tilde{p}_3^2 + \frac{4}{9}\tilde{p}_1\tilde{p}_2 + \frac{6}{9}\tilde{p}_1\tilde{p}_3 + \frac{12}{9}\tilde{p}_2\tilde{p}_3}},$$

where

$$\hat{p}_1 = \frac{n_1}{n}, \quad \hat{p}_2 = \frac{n_2}{n}, \quad \hat{p}_3 = \frac{n_3}{n}, \quad \tilde{p}_1 = \frac{n_1 + \frac{1}{4}}{n+1}, \quad \widetilde{p_2} = \frac{n_2 + \frac{1}{4}}{n+1}, \quad \tilde{p}_3 = \frac{n_3 + \frac{1}{4}}{n+1}.$$

The exact p-value can be calculated by

$$p_\theta = \sup_{\frac{1}{3}p_1 + \frac{2}{3}p_2 + \frac{3}{3}p_3 = \theta} p_\theta(\boldsymbol{p}) = \sup_{\frac{1}{3}p_1 + \frac{2}{3}p_2 + \frac{3}{3}p_3 = \theta} P^*(T^*(\theta) > T(\theta) | \, \boldsymbol{p})$$

where the probability is with respect to the random variable

$$T^*(\theta) = \frac{\left|\frac{1}{3}p_1^* + \frac{2}{3}p_2^* + \frac{3}{3}p_3^* - \theta\right|}{\sqrt{\frac{1}{9}\tilde{p}_1^{*2} + \frac{4}{9}\tilde{p}_2^{*2} + \frac{9}{9}\tilde{p}_3^{*2} + \frac{4}{9}\tilde{p}_1^*\tilde{p}_2^* + \frac{6}{9}\tilde{p}_1^*\tilde{p}_3^* + \frac{12}{9}\tilde{p}_2^*\tilde{p}_3^*}},$$

where

$$p_1^* = \frac{n_1^*}{n}, \quad p_2^* = \frac{n_2^*}{n}, \quad p_3^* = \frac{n_3^*}{n}, \quad \tilde{p}_1^* = \frac{n_1^* + \frac{1}{4}}{n+1}, \quad \widetilde{p_2^*} = \frac{n_2^* + \frac{1}{4}}{n+1}, \quad \tilde{p}_3^* = \frac{n_3^* + \frac{1}{4}}{n+1}.$$

and

$$(n_0^*, n_1^*, n_2^*, n_3^*) \sim MN\big(n, \boldsymbol{p} = (p_0, p_1, p_2, p_3)\big)$$

If p_θ is available, then we can reject or accept the hypothesis that $\theta_0 = \theta$ depending on if $p_\theta < 0.05$.

To calculate p_θ, we need to calculate $p_\theta(\boldsymbol{p})$ for all $\boldsymbol{p} = (p_0, p_1, p_2, p_3)$ such that $\frac{1}{3}p_1 + \frac{2}{3}p_2 + \frac{3}{3}p_3 = \theta$, which is infeasible in practice. Instead, we can select "dense" grid sets $\Omega = \{\boldsymbol{p_1}, \boldsymbol{p_2}, \boldsymbol{p_3}, \cdots, \boldsymbol{p_N}\}$ to cover the parameter space

$$\{(p_0, p_1, p_2, p_3) | p_0 \geq 0, p_1 \geq 0, p_2 \geq 0, p_3 \geq 0, p_0 + p_1 + p_2 + p_3 = 1\}$$

For a given θ, we can select a small number $\epsilon > 0$ and approximate p_θ by

$$\max_{\mathbf{p_k}: \left|\frac{1}{3}p_{k1} + \frac{2}{3}p_{k2} + \frac{3}{3}p_{k3} - \theta\right| < \epsilon} \hat{p}_\theta(\boldsymbol{p_k}),$$

where $\boldsymbol{p_k} = (p_{k0}, p_{k1}, p_{k2}, p_{k3})$. Here $\hat{p}_\theta(\boldsymbol{p_k})$ can be easily calculated by a Monte-Carlo method. Specifically, we can simulate a large number of $(n_0^*, n_1^*, n_2^*, n_3^*) \sim MN(n, \boldsymbol{p_k})$, and calculate the corresponding $T^*(\theta)$. $\hat{p}_\theta(\boldsymbol{p_k})$ is the proportion of $T^*(\theta)$s greater than observed $T(\theta)$.

The 95% confidence interval of θ_0 is then can be constructed as all θ with an estimated $p_\theta \geq 0.05$.

The entire procedure can be described in the following algorithm

- For $k = 1, \cdots, N$
  - Simulate $Z_{1k}, Z_{2k}, Z_{3k}, Z_{4k}$ from unit exponential distribution and let
    $$\boldsymbol{p_k} = \frac{(Z_{1k}, Z_{2k}, Z_{3k}, Z_{4k})}{Z_{1k} + Z_{2k} + Z_{3k} + Z_{4k}}.$$
  - For $b = 1, \cdots, B$
    - Simulate $(n_{bk0}^*, n_{bk1}^*, n_{bk2}^*, n_{bk3}^*) \sim MN(n, \boldsymbol{p_k})$.
    - Calculate $p_{bk1}^* = \frac{n_{bk1}^*}{n}$, $p_{bk2}^* = \frac{n_{bk2}^*}{n}$, $p_{bk3}^* = \frac{n_{bk3}^*}{n}$, $\tilde{p}_{bk1}^* = \frac{n_{bk1}^* + \frac{1}{4}}{n+1}$, $\tilde{p_{bk2}^*} = \frac{n_{bk2}^* + \frac{1}{4}}{n+1}$, $\tilde{p}_{bk3}^* = \frac{n_{bk3}^* + \frac{1}{4}}{n+1}$
    - Calculate $\theta_{bk}^* = \frac{1}{3}p_{bk1}^* + \frac{2}{3}p_{bk2}^* + \frac{3}{3}p_{bk3}^*$
    - Calculate $\sigma_{bk}^{*2} = \frac{1}{9}\tilde{p}_{bk1}^{*2} + \frac{4}{9}\tilde{p}_{bk2}^{*2} + \frac{9}{9}\tilde{p}_{bk3}^{*2} + \frac{4}{9}\tilde{p}_{bk1}^*\tilde{p}_{bk2}^* + \frac{6}{9}\tilde{p}_{bk1}^*\tilde{p}_{bk3}^* + \frac{12}{9}\tilde{p}_{bk2}^*\tilde{p}_{bk3}^*$ where does the variance formula come from? We just want transpose(contrast)*(var matrix of multinomial at p)*contrast, don't we? But this doesn't seem to match eg binomial terms on the diagonal.
- For $\theta \in \{\theta_1, \theta_2, \cdots, \theta_m\} \in [0, 1]$
  - Identify all $\boldsymbol{p_k}$ such that $\left|\frac{1}{3}p_{1k} + \frac{2}{3}p_{2k} + \frac{3}{3}p_{3k} - \theta\right| < \epsilon$
  - For each identified $\boldsymbol{p_k}$
    - Calculate $T_{bk}^*(\theta) = \frac{|\theta_{bk}^* - \theta|}{\sigma_{bk}^*}, b = 1, \cdots, B.$
  - Calculate the proportion $T_{bk}^*(\theta) > T(\theta)$ denoted by $\hat{p}_\theta(\boldsymbol{p_k})$ which θ is this? Is it the same θ in "For $\theta \in \{\theta_1, \theta_2, \cdots, \theta_m\} \in \cdots$"? Since θ_{bk}^* is sampled under the θ corresponding to $\boldsymbol{p_k}$, so shouldn't we use that θ to center θ_{bk}^*?
  - Calculate $\hat{p}_\theta = max_{\boldsymbol{p_k}} \hat{p}_\theta(\boldsymbol{p_k})$

Formatted

- Denote the resulting p values by $\hat{p}_{\theta_1}, \cdots, \hat{p}_{\theta_m}$ and the final 95% confidence interval for θ_0 can be constructed as

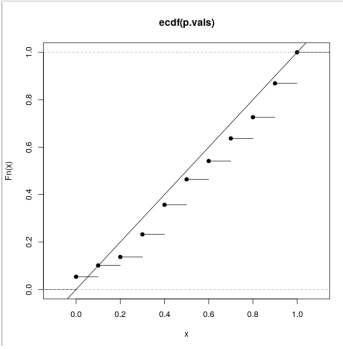
$$\left[\min_{\hat{p}(\theta_k)\geq 0.05} \theta_k \ , \max_{\hat{p}(\theta_k)\geq 0.05} \theta_k\right]$$

In addition, we may consider a different test statistic. In particular, when θ_0 is expected to be very close to 0 or 1, then it can be more appropriate to consider a test statistic in the form of

$$T(\theta) = \frac{\left|\frac{1}{3}\hat{p}_1 + \frac{2}{3}\hat{p}_2 + \frac{3}{3}\hat{p}_3 - \theta\right|}{\sqrt{\frac{1}{9}\tilde{p}_1^2 + \frac{4}{9}\tilde{p}_2^2 + \frac{9}{9}\tilde{p}_3^2 + \frac{4}{9}\tilde{p}_1\tilde{p}_2 + \frac{6}{9}\tilde{p}_1\tilde{p}_3 + \frac{12}{9}\tilde{p}_2\tilde{p}_3}} + \frac{\lambda_1|\hat{p}_1 - p_1|}{\sqrt{\tilde{p}_1(1-\tilde{p}_1)}} + \frac{\lambda_2|\hat{p}_2 - p_2|}{\sqrt{\tilde{p}_2(1-\tilde{p}_2)}}.$$

In my implementation, I find nice behavior for a range of true parameters values. However, there is a somewhat complicated relationship between the tuning parameters epsilon, the number of p vectors we sample from the simplex, and the density of the theta values in our grid.

I focus on a hard case here, a sample size N=10, with 4 multinomial categories as above, using the simple test statistic T, and the true p vector is close to the boundary. Results are still reasonable when the true p vector is (.1,.1,.1.7). Here is the empirical CDF of the p-values at the null in a simulation, which are close the expectd uniform CDF. For example empirical coverage is .946 at the alpha=.05 CI level.

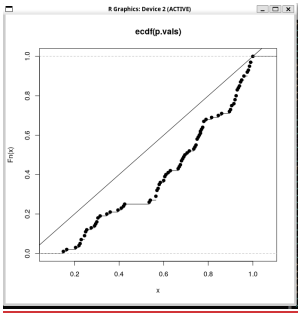


When the true p is (.05,.05,.05,.85), the coverage inflates to .98 unless adjustment is made to the epsilon. At p=(.01,.01,.01,.97), the coverage goes to 1. It becomes necessary to increase epsilon. Otherwise, this close to the boundary of the simplex, there will be no sample p's in an epsilon neighborhood. (The other option would be to drastically increase the number of sample p's we take.) The max over such a large neighborhood tends to be larger. We will probably have to give guidance on

the choice of tuning parameters; we can give simple suggestions relating the suspected distance from the edge to the required number of sample p points. How are your student's results?

I wondered about a different approach. Instead of sampling on the simplex everywhere and and looking at a neighbordhood of a theta value, maybe we can sample from the pre-image of theta. I.e., we start with a grid of theta values [theta_1,…theta_n], then for each theta_i, we sample m probabilities p_ij,j=1,…m, where each p_ij is a probability vector [p_ij0,…p_ij3] such that theta=0/3*p_ij1+…3/3*p_ij3.

Here is the case where the truth p=(.01,.01,.01,.97). With sample size N=50 the p-values at the null become much closer to uniform; coverage of a nominal 95% CI is observed to be 88%.



Technically this approach is a little more difficult since it requires sampling on the intersection of a hyperplane with the probability simplex—I think I figured out a way, but I am not sure if the sampling is uniform. On the other hand, the benefit of this approach is that the tuning parameters are much simpler: A parameter for the density of theta, and a parameter for the density of the probability vectors corresponding to a given theta value. Increasing one or both straightforwardly increases the accuracy of the CI. (We could of course also have the density of the probability vectors per theta depend on the value of theta to reflect prior knowledge of where theta is.) Whereas in the current approach, even for fixed grid size, epsilon cannot be too large or too small, there is an ideal range it must lie in. I can investigate this approach further if you think it is a good idea, or we can stick with the original plan, or we could present both sampling methods. I don't want to over-complicate the project, though, since it is already far along.