# Binary signal standard errors

Alex Reinhart and Nat DeFries

November 2021

## Contents



## 1 Notation

Let $n$ be the number of observations. Let:

$$Y_i \in \{0, 1\} = \text{respondent } i\text{'s response}$$
$$w_i = \text{respondent } i\text{'s weight.}$$

We normalize the weights so that $\sum_{i=1}^{n} w_i = 1$. Note that $Y_i$ is a boolean, since each respondent reports whether or not they fulfill some criteria (e.g. have been vaccinated).

## 2 The Jeffreys approach

The motivation for the Jeffreys approach is as follows. We assume

$$\sum_{i=1}^{n} Y_i \sim \text{Binomial}(n, p)$$
$$p \sim \text{Beta}(1/2, 1/2).$$

The Beta distribution is a conjugate prior here, and this particular beta is the Jeffreys prior (hence the name of this approach). Because of conjugacy, we have that

$$p \mid \sum_{i=1}^{n} Y_i \sim \text{Beta}\left(\frac{1}{2} + \sum_{i=1}^{n} Y_i, n + \frac{1}{2} - \sum_{i=1}^{n} Y_i\right).$$

This results in a posterior mean of

$$\hat{p}_{\text{bayes}} = \frac{\frac{1}{2} + \sum_{i=1}^{n} Y_i}{1+n}.$$

Now, if we were doing this in a completely Bayesian way, we'd use a Bayesian credible interval for our uncertainty. But this would not be symmetric and couldn't give a simple standard error. So we use the normal approximation: the standard error of a binomial with $n$ observations, *plus one fake observation*, would be

$$\hat{\text{se}}_{\text{bayes}} = \sqrt{\frac{\hat{p}_{\text{bayes}}(1-\hat{p}_{\text{bayes}})}{n+1}}.$$

The denominator is $n+1$ because the frequentist sampling distribution should account for the extra fake observation being added to every sample.

So when we proceed with our derivation, if the Jeffreys approach is our motivation, the unweighted case should yield this result.

## 3 Current status

### 3.1 The current code

In the current code (`binary.R`), the function `compute_binary_response` calculates the following:

$$\hat{p}_{\text{unadj}} = \sum_{i=1}^{n} w_i Y_i \tag{1}$$
$$\hat{\text{se}}_{\text{unadj}} = \text{NA} \tag{2}$$
$$n_{\text{eff}} = n\frac{\left(\frac{1}{n}\sum_{i=1}^{n} w_i\right)^2}{\frac{1}{n}\sum_{i=1}^{n} w_i^2} = \frac{\left(\sum_{i=1}^{n} w_i\right)^2}{\sum_{i=1}^{n} w_i^2} = \frac{1}{\sum_{i=1}^{n} w_i^2} \tag{3}$$

Then `jeffreys_binary` adjusts $\hat{p}_{\text{unadj}}$ by introducing an additional observation with $Y = 0.5$. $\hat{p}_{\text{adj}}$ is then used to produce a final standard error. It does this:

$$\hat{p}_{\text{code}} = \frac{n\hat{p}_{\text{unadj}} + \frac{1}{2}}{1+n} \tag{4}$$

$$\hat{\text{se}}_{\text{code}} = \sqrt{\frac{\hat{p}_{\text{adj}}(1-\hat{p}_{\text{adj}})}{n}} \tag{5}$$

Notably, neither the adjustment to $\hat{p}_{\text{unadj}}$ nor the standard error calculation use $n_{\text{eff}}$.

### 3.2 The documented calculation

The documentation produces the same $\hat{p}$ as the code. It then states that:

$$\hat{\text{se}}_{\text{unadj}} = \sqrt{\sum_{i=1}^{n} w_i^2 (Y_i - \hat{p}_{\text{unadj}})^2}, \tag{6}$$

which differs from the code. The documentation also makes no mention of how Jeffreys-adjusted estimates are calculated.

## 4 Deriving the right approach

Let's start over.

We can agree that

$$\hat{\text{se}}_{\text{unadj}} = \sqrt{\frac{\sum_{i=1}^{n} w_i^2 (Y_i - \hat{p}_{\text{unadj}})^2}{(\sum_{j=1}^{n} w_j)^2}}, \tag{7}$$

and that we want to prevent this from being zero, which would occur when $Y_i = \hat{p}$ for all $i$ such that $w_i > 0$. (For example, if $\hat{p} = 0$.)

There are two ways to achieve this that would fit the Jeffreys idea:

1. Transform $\hat{p}$: recalculate it as if there were one additional observation $i = 0$ with $Y_0 = 1/2$, then calculate the standard error according to the equation above while including that extra observation in the sum.

2. Leave $\hat{p}$ untouched but add the extra term to the sum anyway.

### 4.1 Option 1: Using counts approach

We choose the latter route because while we *do* intend to perform Jeffreys adjustment of $\hat{p}_{\text{unadj}}$ for binary indicators not all necessary data is available when we're performing the adjustment in `jeffreys_binary`.

To use approach 1 above, we need access to $\hat{p}_{\text{adj}}$ and the full set of weights and responses (plus the fake observation) to plug into (7), like

$$\hat{\text{se}}_{\text{adj}} = \sqrt{\frac{\sum_{i=1}^{n} w_i^2 (Y_i - \hat{p}_{\text{adj}})^2}{(\sum_{j=1}^{n} w_j)^2}}. \tag{8}$$

However, in `jeffreys_binary` we only have the summary statistics available so we can't do this calculation directly. And we can't do it earlier in `compute_binary_response` because we don't have $\hat{p}_{\text{adj}}$ (and Jeffreys adjustment isn't meant to be performed there).

Instead, $\hat{\text{se}}_{\text{unadj}}$ and $\hat{\text{se}}_{\text{adj}}$ are defined as for counts indicators, and $\hat{p}$ is adjusted separately (using $n_{\text{eff}}$ instead of $n$).

$$\hat{p}_{\text{final}} = \frac{n_{\text{eff}} \hat{p}_{\text{unadj}} + \frac{1}{2}}{1 + n_{\text{eff}}} \tag{9}$$

$$\hat{\text{se}}_{\text{final}} = \frac{1}{1 + n_{\text{eff}}} \sqrt{\left(\frac{1}{2} - \hat{p}_{\text{unadj}}\right)^2 + n_{\text{eff}}^2 \hat{\text{se}}_{\text{unadj}}^2} \tag{10}$$

### 4.2 Option 2: Use p * (1-p) formula

(5) generalizes in a natural way to the weighted case (see `https://stats.stackexchange.com/a/159220`) to

$$\hat{\text{se}} = \sqrt{\hat{p}(1 - \hat{p}) \frac{\sum_{i=1}^{n} w_i^2}{(\sum_{i=1}^{n} w_i)^2}} \tag{11}$$

Adding in the fake observation with weight $w_0 = 1/n_{\text{eff}}$,

$$\hat{\text{se}}_{\text{final}} = \sqrt{\hat{p}_{\text{adj}}(1 - \hat{p}_{\text{adj}}) \frac{(\frac{1}{n_{\text{eff}}})^2 + \sum_{i=1}^{n} w_i^2}{(\frac{1}{n_{\text{eff}}} + \sum_{i=1}^{n} w_i)^2}} \tag{12}$$

$$= \sqrt{\hat{p}_{\text{adj}}(1 - \hat{p}_{\text{adj}}) \frac{1 + n_{\text{eff}}^2 \sum_{i=1}^{n} w_i^2}{(1 + n_{\text{eff}} \sum_{i=1}^{n} w_i)^2}} \tag{13}$$

Since the original weights sum to 1, this simplifies to

$$\hat{\text{se}}_{\text{final}} = \sqrt{\hat{p}_{\text{adj}}(1-\hat{p}_{\text{adj}})\frac{1+n_{\text{eff}}^2\frac{1}{n_{\text{eff}}}}{(1+n_{\text{eff}})^2}} \tag{14}$$

$$= \sqrt{\hat{p}_{\text{adj}}(1-\hat{p}_{\text{adj}})\frac{1+n_{\text{eff}}}{(1+n_{\text{eff}})^2}} \tag{15}$$

$$= \sqrt{\frac{\hat{p}_{\text{adj}}(1-\hat{p}_{\text{adj}})}{1+n_{\text{eff}}}} \tag{16}$$

Conveniently, it doesn't require having all of the weights and responses available. This does not agree with (5).

# 5 Conclusion

Implement option 2.