Count 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 proportion, since each respondent reports the proportion of their household that is sick.

2 Current status

2.1 The current code

In the current code (`count.R`), the function `compute_count_response` calculates the following:

$$\hat{p} = \sum_{i=1}^{n} w_i Y_i \tag{1}$$

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

$$n_{\text{eff}} = n \frac{\left(\frac{1}{n}\sum_{i=1}^{n} w_i\right)^2}{\frac{1}{n}\sum_{i=1}^{n} w_i^2} = \frac{\left(\sum_{i=1}^{n} w_i\right)^2}{\sum_{i=1}^{n} w_i^2} = \frac{1}{\sum_{i=1}^{n} w_i^2}. \tag{3}$$

(Evidently the n_{eff} code can be simplified in the knowledge that the weights sum to 1.)

Then `jeffreys_se` adjusts this to produce a final weighted standard error by introducing an additional observation. This additional observation has $Y = 0.5$. It does this:

$$\hat{\text{se}}_{\text{code}} = \frac{1}{1 + n_{\text{eff}}} \sqrt{\left(\frac{1}{2} - \hat{p}\right)^2 + n_{\text{eff}}^2 \hat{\text{se}}_{\text{unadj}}^2}. \tag{4}$$

2.2 The documented calculation

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

$$\hat{\text{se}}_{\text{unadj}} = \sqrt{\sum_{i=1}^{n} w_i^2 (Y_i - \hat{p})^2}, \tag{5}$$

which is the same as in the code as well. But it then states the following adjusted standard error:

$$\hat{\text{se}}_{\text{doc}} = \sqrt{\left(\frac{1}{1 + n_{\text{eff}}}\right)^2 \left(\frac{1}{2} - \hat{p}\right)^2 + n_{\text{eff}} \hat{\text{se}}_{\text{unadj}}^2}, \tag{6}$$

again using the same definition of n_{eff}.

So how is $\hat{\text{se}}_{\text{code}}$ different from $\hat{\text{se}}_{\text{doc}}$? Let's try to pull factors out of $\hat{\text{se}}_{\text{doc}}$ to make the comparison clearer:

$$\hat{\text{se}}_{\text{doc}} = \frac{1}{1 + n_{\text{eff}}} \sqrt{\left(\frac{1}{2} - \hat{p}\right)^2 + (1 + n_{\text{eff}})^2 n_{\text{eff}} \hat{\text{se}}_{\text{unadj}}^2}. \tag{7}$$

That $(1 + n_{\text{eff}})^2$ term is what Nat found to be inconsistent with the code, since expanding it out leads to several terms not present in the code. (Nat's derivation in the GitHub comments started from (6) but without the n_{eff} multiplying $\hat{\text{se}}_{\text{unadj}}^2$, since he thought that to be a typo.)

2.3 Derivation from covid-19 repo's aggregations-setup.R

The `jeffreys_se` function in `aggregations-setup.R` contains a derivation in code comments, which we translate into notation here. For the purposes of this derivation, let's **not** assume that the weights are normalized. We'll apply assumptions at the end.

In the general case,

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

$$n_{\text{eff}} = \frac{(\sum_{i=1}^{n} w_i)^2}{\sum_{i=1}^{n} w_i^2}. \tag{9}$$

Rearranging,

$$\left(\sum_{j=1}^{n} w_j\right)^2 \hat{\text{se}}_{\text{unadj}}^2 = \sum_{i=1}^{n} w_i^2 (Y_i - \hat{p})^2. \tag{10}$$

Applying the Jeffreys adjustment on both sides by adding an extra term to the sums with $Y_0 = 0.5$ and weight w_0,

$$\left(w_0 + \sum_{j=1}^{n} w_j\right)^2 \hat{\text{se}}_{\text{final}}^2 = w_0^2 \left(\frac{1}{2} - \hat{p}\right)^2 + \sum_{i=1}^{n} w_i^2 (Y_i - \hat{p})^2, \tag{11}$$

and rearranging, we obtain

$$\hat{\text{se}}_{\text{final}} = \sqrt{\frac{w_0^2}{(w_0 + \sum_{j=1}^{n} w_j)^2} \left(\frac{1}{2} - \hat{p}\right)^2 + \frac{1}{(w_0 + \sum_{j=1}^{n} w_j)^2} \sum_{i=1}^{n} w_i^2 (Y_i - \hat{p})^2} \tag{12}$$

$$= \sqrt{\frac{w_0^2}{(w_0 + \sum_{j=1}^{n} w_j)^2} \left(\frac{1}{2} - \hat{p}\right)^2 + \frac{(\sum_{j=1}^{n} w_j)^2}{(w_0 + \sum_{j=1}^{n} w_j)^2} \hat{\text{se}}_{\text{unadj}}^2} \tag{13}$$

$$= \frac{1}{w_0 + \sum_{j=1}^{n} w_j} \sqrt{w_0^2 \left(\frac{1}{2} - \hat{p}\right)^2 + \left(\sum_{j=1}^{n} w_j\right)^2 \hat{\text{se}}_{\text{unadj}}^2}. \tag{14}$$

Assuming the weights are normalized such that $\sum_{j=1}^{n} w_j = 1$, this becomes

$$\hat{\text{se}}_{\text{final, normalized}} = \frac{1}{w_0 + 1} \sqrt{w_0^2 \left(\frac{1}{2} - \hat{p}\right)^2 + \hat{\text{se}}_{\text{unadj}}^2} \tag{15}$$

Now we must choose the weight w_0 for the pseudo-observation. Let $w_0 = 1/n_{\text{eff}}$. This

results in the following:

$$\hat{\text{se}}_{\text{final,normalized}} = \frac{1}{1/n_{\text{eff}} + 1}\sqrt{\frac{1}{n_{\text{eff}}^2}\left(\frac{1}{2} - \hat{p}\right)^2 + \hat{\text{se}}_{\text{unadj}}} \tag{16}$$

$$= \frac{n_{\text{eff}}}{1 + n_{\text{eff}}}\sqrt{\frac{1}{n_{\text{eff}}^2}\left(\frac{1}{2} - \hat{p}\right)^2 + \hat{\text{se}}_{\text{unadj}}} \tag{17}$$

$$= \frac{1}{1 + n_{\text{eff}}}\sqrt{\left(\frac{1}{2} - \hat{p}\right)^2 + n_{\text{eff}}^2\hat{\text{se}}_{\text{unadj}}^2}. \tag{18}$$

This matches the current implementation in (4).

3 Deriving the right approach

So what's a typo and what's not a typo? Let's start over.

We can agree that

$$\hat{\text{se}}_{\text{unadj}} = \sqrt{\frac{\sum_{i=1}^n w_i^2(Y_i - \hat{p})^2}{(\sum_{j=1}^n w_j)^2}}, \tag{19}$$

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.

We chose the latter route because we expected $\hat{p} \approx 0$ for CLI signals, and Jeffreys adjustment of $\hat{p}$ would cause large amounts of bias for values very close to zero.

So we want to add an extra term to the sum. But what should its weight w_0 be? Here the documentation is not specific: it says this extra observation has "weight assigned to appear like a single effective observation according to importance sampling diagnostics."

Let's just use w_0 and see what values of w_0 match the code or documentation. If we

start with (19), we have

$$\left(\sum_{i=1}^{n} w_i\right)^2 \hat{\text{se}}^2_{\text{unadj}} = \sum_{i=1}^{n} w_i^2 (Y_i - \hat{p})^2 \tag{20}$$

$$\left(w_0 + \sum_{i=1}^{n} w_i\right)^2 \hat{\text{se}}^2_{\text{final}} = w_0^2 \left(\frac{1}{2} - \hat{p}\right)^2 + \sum_{i=1}^{n} w_i^2 (Y_i - \hat{p})^2 \tag{21}$$

$$\hat{\text{se}}_{\text{final}} = \frac{1}{w_0 + \sum_{i=1}^{n} w_i} \sqrt{w_0^2 \left(\frac{1}{2} - \hat{p}\right)^2 + \sum_{i=1}^{n} w_i^2 (Y_i - \hat{p})^2} \tag{22}$$

3.1 Option 1: Proceed with stated weight

If

$$w_0 = \frac{1}{1 + n_{\text{eff}}}, \tag{23}$$

and $\sum_{i=1}^{n} w_i = 1$, then starting with (22), we obtain:

$$\hat{\text{se}}_{\text{final}} = \frac{1}{\frac{1}{1+n_{\text{eff}}} + 1} \sqrt{\left(\frac{1}{1 + n_{\text{eff}}}\right)^2 \left(\frac{1}{2} - \hat{p}\right)^2 + \sum_{i=1}^{n} w_i^2 (Y_i - \hat{p})^2} \tag{24}$$

$$= \frac{1}{\frac{1}{1+n_{\text{eff}}} + 1} \sqrt{\left(\frac{1}{1 + n_{\text{eff}}}\right)^2 \left(\frac{1}{2} - \hat{p}\right)^2 + \hat{\text{se}}^2_{\text{unadj}}} \tag{25}$$

$$= \frac{1}{\frac{1+1+n_{\text{eff}}}{1+n_{\text{eff}}}} \sqrt{\left(\frac{1}{1 + n_{\text{eff}}}\right)^2 \left(\frac{1}{2} - \hat{p}\right)^2 + \hat{\text{se}}^2_{\text{unadj}}} \tag{26}$$

$$= \frac{1 + n_{\text{eff}}}{2 + n_{\text{eff}}} \sqrt{\left(\frac{1}{1 + n_{\text{eff}}}\right)^2 \left(\frac{1}{2} - \hat{p}\right)^2 + \hat{\text{se}}^2_{\text{unadj}}} \tag{27}$$

$$= \frac{1}{2 + n_{\text{eff}}} \sqrt{\left(\frac{1}{2} - \hat{p}\right)^2 + (1 + n_{\text{eff}})^2 \hat{\text{se}}^2_{\text{unadj}}}. \tag{28}$$

This doesn't match either documentation or code.

3.2 Option 2: Proceed with different weight

If

$$w_0 = \frac{1}{n_{\text{eff}}}, \tag{29}$$

and $\sum_{i=1}^{n} w_i = 1$, then starting with (22), we obtain:

$$\begin{aligned}
\hat{\text{se}}_{\text{final}} &= \frac{1}{1/n_{\text{eff}} + 1} \sqrt{\frac{1}{n_{\text{eff}}^2} \left(\frac{1}{2} - \hat{p}\right)^2 + \sum_{i=1}^{n} w_i^2 (Y_i - \hat{p})^2} \\
&= \frac{n_{\text{eff}}}{1 + n_{\text{eff}}} \sqrt{\frac{1}{n_{\text{eff}}^2} \left(\frac{1}{2} - \hat{p}\right)^2 + \sum_{i=1}^{n} w_i^2 (Y_i - \hat{p})^2} \\
&= \frac{1}{1 + n_{\text{eff}}} \sqrt{\left(\frac{1}{2} - \hat{p}\right)^2 + n_{\text{eff}}^2 \hat{\text{se}}_{\text{unadj}}^2},
\end{aligned}$$

which matches $\hat{\text{se}}_{\text{code}}$ in (4).

3.3 Option 3: Forget a term

We started from $\hat{\text{se}}_{\text{unadj}}$ in (19). But suppose we started by normalizing the weights to sum to 1, meaning we started with this version instead:

$$\hat{\text{se}}_{\text{unadj}} = \sqrt{\sum_{i=1}^{n} w_i^2 (Y_i - \hat{p})^2}. \tag{30}$$

Notice the absence of a denominator. When we proceed to make the Jeffreys version, that denominator does not get moved to the left-hand side, and so when we add the fake term with $Y_0 = 0.5$ and weight w_0, that term is **not** added to the left-hand side. Instead, we get this:

$$\hat{\text{se}}_{\text{final}}^2 = w_0^2 \left(\frac{1}{2} - \hat{p}\right)^2 + \sum_{i=1}^{n} w_i^2 (Y_i - \hat{p})^2 \tag{31}$$

$$\hat{\text{se}}_{\text{final}} = \sqrt{w_0^2 \left(\frac{1}{2} - \hat{p}\right)^2 + \hat{\text{se}}_{\text{unadj}}^2} \tag{32}$$

If we now substitute in $w_0 = 1/(1 + n_{\text{eff}})$, we get

$$\hat{\text{se}}_{\text{final}} = \sqrt{\frac{1}{(1 + n_{\text{eff}})^2} \left(\frac{1}{2} - \hat{p}\right)^2 + \hat{\text{se}}_{\text{unadj}}^2} \tag{33}$$

$$= \frac{1}{1 + n_{\text{eff}}} \sqrt{\left(\frac{1}{2} - \hat{p}\right)^2 + (1 + n_{\text{eff}})^2 \hat{\text{se}}_{\text{unadj}}^2}. \tag{34}$$

then this matches the documented $\hat{\text{se}}_{\text{doc}}$–except for the documentation's excess n_{eff} term that Nat identified as a typo.

4 Conclusion

The documentation contains two mistakes: it has an extra n_{eff}, and it did not include the denominator in $\hat{\text{se}}_{\text{unadj}}$, causing it to miss an extra term.

The code does not match the documentation's claim that $w_0 = 1/(1 + n_{\text{eff}})$. It avoids the missing denominator, but defines $w_0 = 1/n_{\text{eff}}$.

The question is then "What should w_0 be?" To resolve this, let's consider the case where all weights are equal, i.e. $w_i = w$. Then

$$n_{\text{eff}} = \frac{n^2 w^2}{n w^2} = n. \tag{35}$$

If all weights are equal and sum to 1 before the fake observation is added, $w = 1/n$. it only makes sense for $w_0 = w$. We could only obtain this result if $w_0 = 1/n_{\text{eff}} = 1/n$.

If all weights are equal and sum to 1 *after* the fake observation is added, $w = 1/(n+1)$, i.e. $w_0 = 1/(n_{\text{eff}} + 1)$.

The code normalizes before the fake observation is added, so it makes sense that $w = 1/n_{\text{eff}}$.

In conclusion, the code is correct and the documentation is wrong. Option 2 is the way to go.