Propagation of Error for F_1

William Webber
University of Maryland
wew@umd.edu

April 23, 2013

1 Estimating F_1

		R	
		1	0
C	1	R_1	$N_1 - R_1$
	0	R_0	$N_0 - R_0$

Figure 1: Contingency table of true and false positives and negatives, defined by intersection of the sets classified as relevant (C) and actually relevant (R).

The contingency table in Figure 1 describes the population values of our classifier. Then F_1 is:

$$F_1 = \frac{2R_1}{R_1 + R_0 + N_1} \tag{1}$$

We draw a sample from Figure 1 to estimate classifier effectiveness. We independently sample from the set of classified-relevant and classified-irrelevant documents to derive estimates $\hat{R}_1$ and $\hat{R}_0$, and hence our estimate of F_1:

$$\hat{F}_1 = \frac{2\hat{R}_1}{\hat{R}_1 + \hat{R}_0 + N_1} \tag{2}$$

noting that N_1 is known, not estimated. What we now require is an expression for $\widehat{\mathrm{Var}}(\hat{F}_1)$.

2 Propagation of error

If:

$$X = f(A, B) \tag{3}$$

then the theory of propagation of error states:

$$\mathrm{Var}(X) = \left(\frac{\partial f}{\partial A}\sigma_A\right)^2 + \left(\frac{\partial f}{\partial B}\sigma_B\right)^2 + 2\frac{\partial f}{\partial A}\frac{\partial f}{\partial B}\mathrm{Cov}_{AB} \ , \tag{4}$$

Let $X = \hat{F}_1$, $A = \widehat{R}_1$, and $B = \widehat{R}_0$. Because the classified-relevant and classified-irrelevant strata are independently sampled, $\mathrm{Cov}_{AB} = 0$, and the third term of Equation (4) drops out.

We now evaluate the partial derivatives in Equation (4). Let $f(A, B)$ be Equation (2). Write $f_A(A)$ for $f(A, B)$ where B is fixed, so that:

$$f'_A \equiv \frac{\partial f}{\partial A} \; . \tag{5}$$

Then write:

$$f_A(A) = g_A(A) \cdot h_A(A) \tag{6}$$

where:

$$g_A(A) = 2A \tag{7}$$

and

$$h_A(A) = \frac{1}{A + B + N} \tag{8}$$

writing N as shorthand for N_1. Then, by the product rule:

$$f'_A(A) = (g_A(A) \cdot h_A(A))' = g'_A(A) \cdot h_A(A) + g_A(A) \cdot h'_A(A) \; . \tag{9}$$

Now:

$$g'_A(A) = 2 \tag{10}$$

and:

$$h'_A(A) = -\frac{1}{(A + B + N)^2} \tag{11}$$

(which can be verified by the chain rule). So:

$$f'_A(A) = \frac{2}{A + B + N} - \frac{2A}{(A + B + N)^2} \; . \tag{12}$$

Next, write:

$$f_B(B) = g_B(B) \cdot h_B(B) \tag{13}$$

where:

$$g_B(B) = 2A \tag{14}$$

and

$$h_B(B) = \frac{1}{A + B + N} \; . \tag{15}$$

Now:

$$g'_B(B) = 0 \; , \tag{16}$$

and

$$h'_B(B) = -\frac{1}{(A + B + N)^2} \; ; \tag{17}$$

the latter by analogy with $h'_A(A)$. So

$$f'_B(B) = -\frac{2A}{(A + B + N)^2} \; . \tag{18}$$

Plugging Equation (12) and Equation (18) into Equation (4) (and reverting to the latter notation), we find:

$$\begin{aligned}
\widehat{\mathrm{Var}}(\hat{F}_1) &= \left(\frac{2}{\widehat{R}_1 + \widehat{R}_0 + N_1} - \frac{2\widehat{R}_1}{\left(\widehat{R}_1 + \widehat{R}_0 + N_1\right)^2} \right)^2 \widehat{\mathrm{Var}}(\widehat{R}_1) \\
&\quad + \left(\frac{2\widehat{R}_1}{\left(\widehat{R}_1 + \widehat{R}_0 + N_1\right)^2} \right)^2 \widehat{\mathrm{Var}}(\widehat{R}_0) \\
&= \frac{4}{\left(\widehat{R}_1 + \widehat{R}_0 + N_1\right)^4} \left[\left(\widehat{R}_0 + N_1\right)^2 \widehat{\mathrm{Var}}(\widehat{R}_1) + \widehat{R}_1^2 \widehat{\mathrm{Var}}(\widehat{R}_0) \right] \quad (19)
\end{aligned}$$

(noting that the minus sign in Equation 18 goes inside the parentheses and then gets "squared out"), where:

$$\widehat{\mathrm{Var}}(\widehat{R}_1) = N_1^2 \left(\frac{(r_1/n_1) \cdot (1 - r_1/n_1)}{n_1} \right) = N_1^2 \cdot \frac{r_1}{n_1^2} \cdot \left(1 - \frac{r_1}{n_1}\right) \quad (20)$$

(r_1 and n_1 being sample values, but N_1 being the population value), and similarly for $\widehat{\mathrm{Var}}(\widehat{R}_0)$. We can also add a finite population adjustment:

$$\widehat{\mathrm{Var}}(\widehat{R}_1) = N_1^2 \cdot \frac{r_1}{n_1^2} \cdot \left(1 - \frac{r_1}{n_1}\right) \cdot \left(1 - \frac{n_1}{N_1}\right) \quad (21)$$

Equation 19 gives our expression for the sampling variance of F_1.