July 17, 2025

Manuscript: *The statistical concepts of multiple imputation and bootstrap: illustration with cholera mortality data collected by John Snow*

Presentation: *The Variance Calculation Following Multiple Imputation: Illustration with Cholera Mortality Data Collected by John Snow*

This work brought together two different interests of mine: the history of epidemiology, and using 'minimalist' examples to teach statistical concepts and techniques.

I have long admired John Snow. Two aspects of his numerical work have always impressed me: his 'shoe leather' data collection and his use of existing (usually governmental) data.

Sadly, the current-day telling of his work, and its focus on the Broad Street Pump story, sidelines what I think is a much stronger piece of evidence for his waterborne theory of the spread[1] of cholera, namely the Grand Experiment in *South* London. Indeed he was in the middle of his South London data collection (of the numerators) when he was called away to the Broad Street outbreak. And he devoted less of his book to it (as Steven Johnson's book *The Ghost Map* so nicely tells it, it was Henry Whitehead who (later, after reading Snow's book) extracted the best evidence from the Broad Street episode.

Whenever I encounter a new statistical technique (such as GEE), I try to understand it using as simple a dataset as possible, and using another established method as a cross check. So, even though Rubin's formula seemed intuitive to me, I wanted to see what happened in the 'null' case. And I was very pleased to see that in this simple case where, if we *pretended* that the full denominators[2] were missing, and so had to be imputed, Rubin's formula lines up with Woolf's formula, which has been around since 1955.

If you would like to take revitalize it, I would be delighted to hear from you!

Sincerely,

James Hanley

webpage: https://jhanley.biostat.mcgill.ca | email: `james.hanley@mcgill.ca`

[1] Starting from his initial, and beautifully argued, pamphlet in 1849, he always wrote of the 'mode of *communication*' of cholera.

[2] JSM Attendees may not, but I still remember the vehemence of the attack, by a non-statistician, on what he perceived as my 'poor scholarship' in using the 'known' denominators (41,676 and 24,477) I used. I told him the two numbers I used as 'known' (or '*pretend* known') denominators were purely to illustrate a statistical concept. Later, when I looked into the critic's own publication on the subject, I was astounded at the lack of rigour/care displayed in his and his co-author's publication on the topic. My physician (also a colleague) was worried about what responding would do to my cardiac system. So, for that and other reasons, I left it be.
But in 2024, a properly-reviewed article that addressed the critics' 'scholarship' concluded that "Unfortunately [...] are wrong in their initial claims, their analysis, and their conclusions. Snow's statistical intuition (and his approach), were absolutely correct. In 1856 he did not have the tools to demonstrate the overwhelming influence of water, but applying modern statistical tools to Snow's approach and data fully vindicates Snow's claim for the overwhelming influence of water."

The statistical concepts of multiple imputation and bootstrap: illustration with cholera mortality data collected by John Snow

Juli Atherton[1], James A. Hanley[!]

[2]Department of Epidemiology, Biostatistics and Occupational Health, McGill University Montreal, Quebec, H3A 1A2, Canada

Correspondence to: Juli Atherton `juli.atherton@mcgill.ca`

For submission to American Journal of Epidemiology, and **presentation at JSM Vancouver**; draft May 07, 2010

Keywords: ....

ABSTRACT

Multiple imputation techniques are increasingly used in epidemiology. In order to appreciate the concepts and principles behind them, it is helpful to see how they work in a simple situation where we already have well-established formulae, such as 'Woolf's' formula for the standard error of the log of a cross-product ratio. This ratio is used to estimate a rate ratio when the relative sizes of the denominators in the compared rates are estimated (via a 'control' or 'denominator' series) rather than known. We illustrate and compare the standard error formula obtained by multiple imputation with that of Woolf. We do so using the exposure information in the 'numerator series' of 300 deaths collected by John Snow using data from the 'grand experiment' which exploited 'the intermixing of the water supply of the Southwark and Vauxhall Company with that of the Lambeth Company, over an extensive part of London'. For his denominators, Snow used the already-established numbers of customers served by these two companies. For the sake of this exposition, we pretend that he had to estimate their relative sizes himself from a sample survey - by the same combination of shoe-leather epidemiology and technical sophistication he used to arrive at the numerators – and to use multiple imputation to obtain the standard error for the log of the rate ratio.

1 INTRODUCTION

Multiple imputation techniques are increasingly used in epidemiology. They require specialized software and thus are seldom illustrated in statistical texts with hand-worked examples that would allow the end-user to appreciate the concepts behind them and some of the subtleties in their correct use. This note attempts to rectify this by using a simple and well known data example that allows the basic concepts not just to be understood but also to be tested against an already well-established – and epidemiologically-famous – formula: 'Woolf's' formula for the standard error of the log of a cross-product ratio. This ratio is used to estimate a rate ratio when the relative sizes of the denominators in the compared rates are estimated (via a 'control' or 'denominator' series) rather than known. We illustrate and compare the standard error formula obtained by multiple imputation with that of Woolf. We do so using the exposure information in the 'numerator series' of 300 deaths collected by John Snow using data from the 'grand experiment'

1.1 John Snow's data

Snow's study exploited the mixing – "of the most intimate kind" – of the water supply of the Southwark and Vauxhall Company with that of the Lambeth Company, over several sub-districts, with a combined population at least 300,000 people. The main features have been reproduced in many epidemiology textbooks, and on the dedicated UCLA website, and so they will not be repeated here. We focus on the data in Table VII of his report,

entitled "The mortality from Cholera in the four weeks ending 5th August [1854]" and on a lesser-known aspect of the data in this table – how Snow obtained the key numerators and denominators that formed the basis for his early results: "Consequently, as 286 fatal attacks of cholera took place, in the first four weeks of the epidemic, in houses supplied by the former Company, and only 14 in houses supplied by the latter, the proportion of fatal attacks to each 10,000 houses was as follows. Southwark and Vauxhall xx Lambeth x. The cholera was therefore fourteen times as fatal at this period, amongst persons having the impure water of the Southwark and Vauxhall Company, as amongst those having the purer water from Thames Ditton."

"To turn this grand experiment to account," Snow set out "to learn the supply of water to each individual house where a fatal attack of cholera might occur." He obtained the addresses of persons dying of cholera in these districts with an intermingled supply, but quickly discovered that "the inquiry was necessarily attended with a good deal of trouble. There were very few instances in which I could at once get the information I required. It would, indeed, have been almost impossible for me to complete the inquiry, if I had not found that I could distinguish the water of the two companies with perfect certainty by ...". [1]

For his denominators, Snow used the already-established numbers of customers served by these two companies the previous year. However, for the sake of this exposition, *we will pretend that he had to estimate their relative sizes himself from a sample survey* - by the same combination of shoe-leather

[1]The reader who is unaware of how he did so can consult [online] Snow's description of his "high-tech" method.

epidemiology and technical sophistication he used to arrive at the numerators – and to use multiple imputation to obtain the (large-sample) standard error for the log of the rate ratio.

1.2 The sampling variability of the log of a rate ratio

To begin with, for simplicity, we will treat the "fourteen times as fatal" as an empirical rate (or incidence density) ratio with numerators that were subject to Poisson variability. We consider two possible denominator situations (i) where they are – or at least their ratio is – *known without error*, and (ii) when their ratio has to be *estimated* from a simple random sample of houses.

Snow's shoe-leather and high-tech epidemiology allowed him to classify the 300 informative cases[2] in the numerator series into c_1 exposed and c_0 unexposed cases. With the *known person-time denominators* PT_1 and PT_0, as in Snow's study, the rate ratio is estimated as

$$\text{rate ratio} = \frac{c_1/PT_1}{c_0/PT_0} = \frac{286/PT_1}{14/PT_0}$$

and, as is derived in the Appendix, the variance of the log of the rate ratio is estimated by

$$\frac{1}{\text{no. exposed cases}} + \frac{1}{\text{no. unexposed cases}} = \frac{1}{c_1} + \frac{1}{c_0} = \frac{1}{286} + \frac{1}{14} = (0.274)^2.$$

Thus, if a reviewer had requested it, Snow could have accompanied his

[2]There were 334 in all, but the other 34 received their water from pump-wells (4), the river Thames, ditches etc.(26), or unascertained sources(4).

rate ratio of *fourteen* by a 95% multiplicative margin of error of $\times/\div$ $\exp[\pm 1.96[0.274)] = 1.71$.

For the reminder of this note, we consider a scenario where John Snow *did not know the sizes of* PT_1 *and* PT_0, and thus had to enlist a 'denominator-assistant' to visit, and classify the water supply of, a simple random sample – a denominator series – of $d = 100$ homes who received water from one of the two companies. Suppose $d_1 = 63$ of these were supplied with the impure water of the Southwark and Vauxhall Company, and $d_0 = 37$ with the purer water from the Lambeth Company.

If we write $PT = PT_1 + PT_0$, then our estimates of PT_1 and PT_0 are $\widehat{PT_1} = (d_1/d) \times PT$ and $\widehat{PT_0} = (d_0/d) \times PT$ respectively, so that our estimate of the rate ratio is now

$$\text{rate ratio} = \frac{c_1/\widehat{PT_1}}{c_0/\widehat{PT_0}} = \frac{c_1/\{(d_1/d) \times PT\}}{c_0/\{(d_0/d) \times PT\}} = \frac{c_1/d_1}{c_0/d_0} = \frac{c_1/c_0}{d_1/d_0}.$$

and, as per Yule's derivation, the ('Woolf') variance of the log of this rate ratio estimate is

$$\left\{\frac{1}{c_1} + \frac{1}{c_0}\right\} + \left\{\frac{1}{d_1} + \frac{1}{d_0}\right\} = \left\{\frac{1}{286} + \frac{1}{14}\right\} + \left\{\frac{1}{63} + \frac{1}{37}\right\} = (0.343)^2.$$

Thus, if a reviewer had requested it, Snow could have accompanied his rate ratio of *twelve* [(286/63)/(14/37)] by a 95% multiplicative margin of error of $\times/\div$ $\exp[\pm 1.96[0.343)] = 1.96$, rather than the previous 1.71. The increased margin of error reflects the increased statistical uncertainly from having to use *estimated* rather than *known* denominators for Poisson-distributed nu-

merators.

For the reminder of this note, the standard error of 0.343 will serve as the gold standard against which to judge the performance of multiple imputation.

2 MI - in general

We first illustrate the general idea using a small dataset collected on 10 children. A described in Weisberg (1980), a catheter is passed into a major vein or artery at the femoral region and moved into the child's heart. The proper length of the introduced catheter has to be guessed by the physician. The aim of the data collected on the 10 children is to describe the relation between the catheter length and the patient's height. Unfortunately, information on height is missing in 3 subjects, but information on the child's weight is available on all 10.

[Figure 1 about here.]

Whereas the concern would be with a fitted prediction equation, say

$$\widehat{length} = b_0 + b_1 \times height,$$

we will for the same of the exposition, limit ourselves to the point and interval estimate of one of the 2 parameters, namely b_1

We have a few choices: we can estimate it using just the 7 complete observations, or we can 'impute' the 3 missing heights from a regression

equation linking height and weight, and then regress the 10 catheter lengths on the 10 heights (7 direct, 3 imputed). Or, we could ... (Juli)

To be filled in: the details as to how the values are imputed.

If we do so naively, the SE for the fitted b_1 is artificially low, since it does not reflect the fact that 3 of the heights are not exact, but inputed.

The way to reflect this uncertainty is to make several such datasets (C 'copies', say) , each one with 3 slightly different estimates for the 3 missing heights. See Figure 1. Each one yields a b_1 and an associated $SE(b_1)$ or $SE^2 = Var(b_1)$. Lets us call these variances $v_1, \ldots, v_C$. These are derived as part of the regression fitting, and are each shown as light blue squares at the bottom of the Figure. The C estimates of b_1 differ from one dataset to another; they are shown as black dots at the bottom of Fig 1. we denote their empirical variance by V, say, and show it as the red square.

The best estimate of b_1 is the average, $\overline{b_1}$, of the C estimates, and its associated $SE(\overline{b_1})^2 = Var(\overline{b_1})$ is an amalgam of 2 variances,

$$Var(\overline{b_1}) = \bar{v} + (1 + 1/C) \times V.$$

This is represented as the larger black square at the bottom right of Figure 1.

3 MI - Snow example

This is somewhat non-standard because – unlike in most case-control studies – we did not link the numerators and the denominators – different teams

assembles each, they did not consult each other, and they did not keep the addresses of the houses, just the numerator tallies of 286 and 14 and the (partial) denominator tallies of 63 and 37.

In this case, the entire numbers of houses of the two types are imputed from the 63 and 37. If it is known that there were a total of 66,153 homes, the single-inputation estimate is $(63/100) \times 66,153$ and $(37/100) \times 66,153$ so that the rate difference can be calculated as $\frac{286}{(63/100)\times 66,153} - \frac{14}{(37/100)\times 66,153}$, and the rate ratio can be calculated as

$$\frac{286}{(63/100) \times 66,153} \div \frac{14}{(37/100) \times 66,153} = \frac{286}{63} \div \frac{14}{37}.$$

Note that this rate ratio estimate does not require that we know what the total number of houses is, or what the sampling fraction was.

If we naively assume that the estimated denominators $(63/100) \times 66,153 = 41,676$ and $(37/100) \times 66,153 = 24,477$ were exactly correct, then the SE^2 for the log of the rate ratio is simply $1/286 + 1/14$.

However, they are not exact, and so this SE^2 is artificially low, since it does not reflect the fact that the 41,676 and 24,477 were inputed. Again, as in the catheter case, we form multiple versions of the denominators, reflecting the additional statistical uncertainty. Ten such copies are shown in Fig 2, along with 10 estimates of the log-rateRatio and 10 corresponding variances.

Fig 2 to come, but like Fig 1 –

key is that each v is $1/286 + 1/14$, and that the point estimated of log RateRatio differe from copy to copy by a factor that depends only on the 63

and 37, in fact the log-RateRatios have a variance of $1/63 + 1/37$. (it wont be exactly that in any 10 copies, but if no. of copies is large it will be close to that, and on average it will be that.

$$V = \text{Var of the } C \; logRateRatios = 1/63 + 1/37.$$

This fact is because the proportions are drawn from a beta distribution with parameters $\alpha = 63, \beta = 37$), and they produce this variance when they are converted into logRateRatios (proof to be added, based on variance of beta and log transformation.).

$$\overline{\widehat{logRR}} = \text{ave. of the } C \; \widehat{logRR}\,'s$$

$$v_1 = v_2 = \cdots = v_C = 1/286 + 1/14, \quad \text{so } \bar{v} = 1/286 + 1/14$$

So, with a large C so that $1/C$ is negligible, the formula is

$$Var(\overline{\widehat{logRR}}) = \bar{v} + V = 1/286 + 1/14 + 1/63 + 1/37.$$

SO, WOOLF and RUBIN (MI) formulae coincide. The $1/63 + 1/37$ is the price paid for estimating the denominators.

4 COMMENTS/DISCUSSION

Even though books say 5-10 copies, we suggest more. Otherwise, cannot V is not stable. (effort one should devote to getting a stable V depends on

relative sizes of $\bar{v}$ and V.)

Guidance re proper MI model. Juli.

Enlightenment re the modern case-control study: numerator series coupled with denominator series that served to estimate the relative sizes of the 2 denominators. Nothing says that the two series need to be linked. On team can work on the numerators, another *independently* on the denominators. Just as Woolf himself did in his study of blood types and ulcers. He used an independent source for the denominators, and he had a series that was much much larger than the numerator series (wasteful if denominators have a big unit cost)

5 APPENDICES

5.1 The variance of the log of a rate, and of a rate ratio

Most textbooks give, but do not derive, these large-sample variance formula. The derivation relies on the 'Delta method' for a transformation (function) of a random variable. When the transformation is merely linear, the formula is exact, Thus, if the SD of a series of temperatures is 10C, then the SD of this same series is $10\times(9/5) = 18$C. The variance of 100 'square degrees C' is transformed to $100\times(9/5)^2$ on the 'square degrees F' scale.[3] We can think of this the $(9/5)$ as dF/dC, and so write the variance in the new scale as

$$Var[Temp_F] = Var[Temp_C] \times (dF/dC)^2.$$

In our application, we convert from rate to log(rate), and so the scaling factor is not constant over the range in question. In such non-linear transformations, it is customary to approximate the new variance using the value of the scaling factor (the derivative, hence the 'Delta') at the center of the distribution in the original scale.

We assume that the possible values for the numerators of the two compared rates – i.e. the numbers of cases, c_1 and c_0 – are governed by independent Poisson-distributions with expectations μ_1 and μ_0. We assume

[3] In everyday life, the concept of variance is not easily communicated. For example, if the average fertility is 1.6 children per woman, and the SD is say 1.2 1.6 children per woman, then the variance is 1.44 square children per square woman. The main reason we use the square of the SD, rather than the SD itself, is that variances add, whereas SDs do not.

that the absolute person-time denominators of the two compared rates are either known (PT_1 and PT_0), or at a minimum, that their ratio ($PT_{ratio} = PT_1/PT_0$) is known. Since PT_{ratio} is assumed to have no sampling variation or uncertainty, the variance (Var) of the random variable

$$\log(\text{rate ratio}) = \log\left(\frac{c_1/PT_1}{c_0/PT_0}\right) = \log(c_1) - \log(c_0) + \log(PT_{ratio})$$

is merely $\text{Var}[\log(c_1)] + \text{Var}[\log(c_0)]$. Now,

$$\text{Var}[\log(c_i)] \approx Var(c_i) \times \left(\frac{d\ log(\mu_i)}{d\mu_i}\right)^2.$$

From the Poisson model, we know that $\text{Var}(c_i) = \mu_i$, while the square of the derivative is $(1/\mu_i)^2$, so their product is $1/\mu_i$. Thus, the two large-sample variance components of the log of the rate ratio add to $1/\mu_1 + 1/\mu_0$. In practice, since the values of μ_1 and μ_0 are unknown, we substitute the observed values c_1 and c_0 for the corresponding expected values, and arrive at the variance formula

$$\widehat{Var}[\log(\text{rate ratio})] = \frac{1}{\text{no. exposed cases}} + \frac{1}{\text{no. unexposed cases}} = \frac{1}{c_1} + \frac{1}{c_0}.$$

5.2 'Woolf's' formula for the variance of the log of a ratio of cross-products of observed frequencies

Woolf used, but did not derive, the variance formula. It has been known at least since the 1900 paper by Yule, who derived a formula for the square root of the sampling variance of the cross-product ratio itself, rather than of its log. Ref. In the $a,\ b,\ c,\ d$ notation for frequencies used by epidemiologists,[4] his main target was the correlation-type statistic $Q = (ad - bc)/(ad + bc)$, rather than the cross-product ratio, ad/bc. His derivations relied on the same 'Delta method' mathematical statisticians use today, and began with the large-sample sampling variance of the log of a single odds, derived from the binomial-based variance of a single proportion $p = a/(a + b)$ derived from a sample of n $(= a + b)$, about its expected value P:

$$\mathrm{Var}\left[\log\left(\frac{p}{1-p}\right)\right] \approx Var(p) \times \left(\frac{d\ log\frac{P}{1-P}}{dP}\right)^2$$

From the binomial model, we know that $\mathrm{Var}(p) = P(1 - P)/n$. Using two applications of the chain rule – one for the odds and one for the log – the scaling factor in the second term on the right is found to be $\{P(1 - P)\}^{-1}$, so that the large-sample variance of the log of the odds is

$$\frac{P(1-P)}{n} \times \{P(1-P)\}^{-2} = \frac{1}{nP(1-P)} = \frac{1}{nP} + \frac{1}{n(1-P)}.$$

[4] Yule denoted the four frequencies by AB, $A\beta$, αB and $\alpha\beta$.

In practice, since the value of P is unknown, we substitute the observed values $np = a$ and $n(1-p) = b$ for the corresponding expected values. The variance of the log of the ratio of two independently estimated odds, a/b and c/d, is the sum of the two separate variances, $1/a + 1/b$ and $1/c + 1/d$.
In our application, the cross-product is

$$\text{rate-ratio estimate} = \frac{c_1/\widehat{PT_1}}{c_0/\widehat{PT_0}} = \frac{c_1/d_1}{c_0/d_0} = \frac{c_1/c_0}{d_1/d_0},$$

where c_1 and c_0 are the frequencies of exposed and unexposed in the case (numerator) series, and d_1 and d_0 are the frequencies of exposed and unexposed in the completely independent denominator series used to estimate the relative sizes of the person-time denominators PT_1 and PT_0. In this notation, the variance of the log of the rate ratio estimate is

$$\text{Var}[\log(\text{rate-ratio estimate})] = (1/c_1 + 1/c_0) + (1/d_1 + 1/d_0).$$

We have written it in this way to emphasize the separate variance components arising from the Poisson variation of the numbers c_1 and c_0, in the *case*-series[5] and the binomial-based variation of the $d_1 : d_0$ split in the *denominator* series.

[5]It turns out that if we condition on the sum c of two independent Poisson random variables c_1 and c_0, then $c_1 \mid c \sim Binomial[c,\ \mu_1/(\mu_1 + \mu_0)]$.

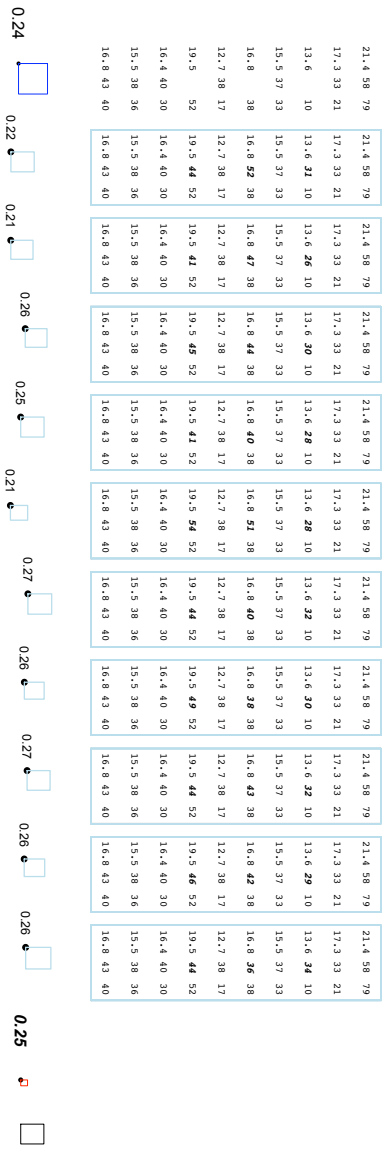


Figure 1: Original dataset (missing 3 heights) and 10 MI copies of the catheter data (mputed values in bold). colums: 1 = catheter length in inches, 2= height in inches, 3= weight in lbs. At bottom of original and of each copy, point est. of slope, and squared SE. At bottom right (red) black square = SE^2 = variance = sum of red square and ave. of blue squares.

The Variance Calculation Following Multiple Imputation: Illustration with Cholera Mortality Data Collected by John Snow

James Hanley and Juli Atherton

Epidemiology, Biostatistics & Occupational Health
McGill University

JSM – Vancouver

August 2 , 2010

Outline

- Multiple Imputation Variance Formula – standard example
- Var[Rate Ratio] in (simulated) case control study

  [case control study – incomplete denominators]
- Summary

standard example

(*simplified for sake of exposition*)

The Optimal Length of Insertion of Central Venous Catheters for Pediatric Patients

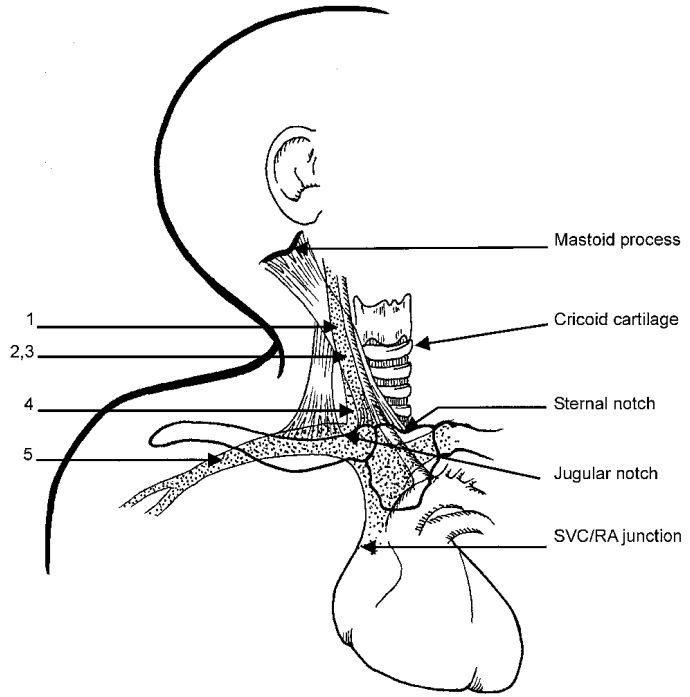


Figure 1. Surface and deep landmarks for right internal jugular (RIJ) and subclavian venipuncture. Puncture sites: 1 = high approach to RIJ used in this study—midway between mastoid process and sternal notch. 2,3 = middle approach using apex of muscular triangle formed by the sternal and clavicular heads of the sternocleidomastoid muscle, or lateral to the cricoid cartilage. 4 = low approach using the jugular notch as a landmark. 5 = subclavian vein puncture site used in this study—1 cm lateral to midpoint of clavicle for patient weighing <10 kg, 2 cm lateral if >10 kg. SVC/RA = supe-

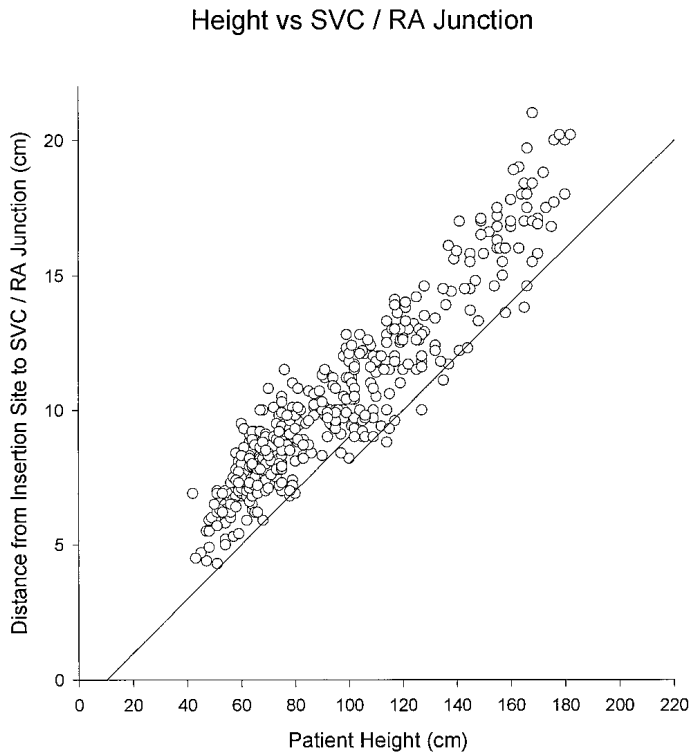


Figure 2. Plot of patient height versus distance from catheter insertion site to junction of superior vena cava (SVC) and right atrium (RA) for right internal jugular and right subclavian vein catheters. Solid lines represent recommendations for initial length of catheter insertion in centimeters: (patient height in cm/10) − 1 for patients ≤100 cm, and (patient height in cm/10) − 2 for patients >100 cm.

Use Patient Height as predictor of Optimal Length

Missing Information

Catheter Length (")	Height (")	Weight (lbs)
21.4		79
17.3		21
13.6	24	10
15.5	37	33
16.8	43	38
12.7		17
19.5	46	52
16.4	40	30
15.5	38	36
16.8	43	40

Catheter Length = a + b . Height

SE's, and their squares (Variances) ...

> Based on n = 7

Data from Weisberg text: n=7 ‘complete’ cases.

7 ‘complete’ + 3 inputed cases

Catheter Length (")	Height (")	Weight (lbs)
21.4	***65***	79
17.3	***30***	21
13.6	24	10
15.5	37	33
16.8	43	38
12.7	***29***	17
19.5	46	52
16.4	40	30
15.5	38	36
16.8	43	40

Catheter Length = a* + b* . Height

SE's, and their squares (Variances) ...

Based on n = 10

7 ‘complete’ + 3 inputed cases

Catheter Length (")	Height (")	Weight (lbs)
21.4	***65***	79
17.3	***30***	21
13.6	***24***	10
15.5	37	33
16.8	43	38
12.7	***29***	17
19.5	46	52
16.4	40	30
15.5	38	36
16.8	43	40

Catheter Length = a* + b* . Height

SE's, and their squares (Variances) ...

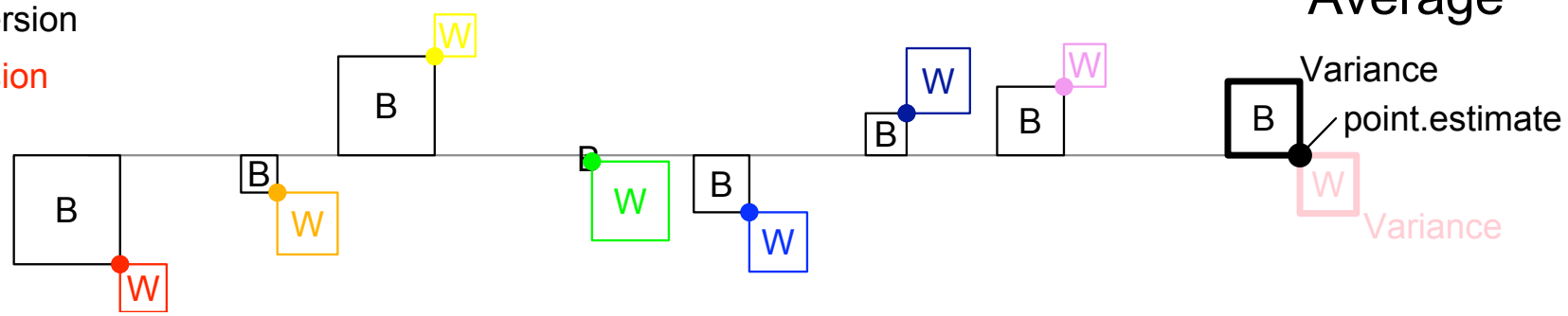
Based on n = 10

INSTEAD: Multiple Imputation

version:	original	1	2	3	4	5	6	7
	21.4 79	21.4 ***80*** 79	21.4 ***70*** 79	21.4 ***56*** 79	21.4 ***63*** 79	21.4 ***67*** 79	21.4 ***60*** 79	21.4 ***58*** 79
	17.3 21	17.3 ***20*** 21	17.3 ***29*** 21	17.3 ***42*** 21	17.3 ***30*** 21	17.3 ***22*** 21	17.3 ***31*** 21	17.3 ***37*** 21
	13.6 24 10	13.6 24 10	13.6 24 10	13.6 24 10	13.6 24 10	13.6 24 10	13.6 24 10	13.6 24 10
	15.5 37 33	15.5 37 33	15.5 37 33	15.5 37 33	15.5 37 33	15.5 37 33	15.5 37 33	15.5 37 33
	16.8 43 38	16.8 43 38	16.8 43 38	16.8 43 38	16.8 43 38	16.8 43 38	16.8 43 38	16.8 43 38
	12.7 17	12.7 ***26*** 17	12.7 ***34*** 17	12.7 ***26*** 17	12.7 ***37*** 17	12.7 ***26*** 17	12.7 ***30*** 17	12.7 ***24*** 17
	19.5 46 52	19.5 46 52	19.5 46 52	19.5 46 52	19.5 46 52	19.5 46 52	19.5 46 52	19.5 46 52
	16.4 40 30	16.4 40 30	16.4 40 30	16.4 40 30	16.4 40 30	16.4 40 30	16.4 40 30	16.4 40 30
	15.5 38 36	15.5 38 36	15.5 38 36	15.5 38 36	15.5 38 36	15.5 38 36	15.5 38 36	15.5 38 36
	16.8 43 40	16.8 43 40	16.8 43 40	16.8 43 40	16.8 43 40	16.8 43 40	16.8 43 40	16.8 43 40

B=Between.version

W=Within.version



MI Estimate and its Variance, if k versions

Est. = Ave. of version-specific point-estimates

Var.* = $\frac{k+1}{k} \times$ Var. of version-specific point-estimates (**B**etween')
+
Average of version-specific variances (**W**ithin)

Var.* (usually) $<$ Var. from 'complete-case' analysis

---

* Rubin, D.B. (1987) Multiple Imputation for Nonresponse in Surveys. Wiley.

MI Variance Formula: Reality Check

Consider a missing-data example where ...

- imputation provides no additional information.
- have an independently-arrived-at variance formula.

simulated case control study

(*actual numerator data*)

John Snow's 'Grand Experiment'

- Exploited mixing – "of most intimate kind" – of water supply of Southwark and Vauxhall Company ("1") with that of Lambeth Company ("0"), over several sub-districts, with combined population $>$ 300,000 [$>$ 66,000 homes].
- "c_1 = 286 **fatal** attacks of **cholera** ('cases') took place, in 1st 4 weeks of 1854 epidemic, in houses supplied by former Company, and only c_0 =14 in houses supplied by latter:
- XX times as fatal at this period, amongst persons having the impure water ("1") of the Southwark and Vauxhall Company, as amongst those having the purer water("0")."

Denominators (D_0, D_1) from 1853 Report

With the ***known denominators*** D_1 and D_0

$$\text{rate ratio} = \frac{c_1/D_1}{c_0/D_0} = \frac{286/D_1}{14/D_0}$$

Variance of the log of the rate ratio is estimated by*

$$\frac{1}{c_1} + \frac{1}{c_0} = \frac{1}{286} + \frac{1}{14} = 0.075.$$

---

* Extra-Poisson variation was minimal.

What if Snow had to *estimate* the Denominators? ...

- using a 'denominator-assistant' to visit, and classify the water supply of, a simple random sample – a denominator series – of $d = 100$ homes from the total of D homes.

- Suppose

  $d_1 = 63$ supplied with the impure water;
  $d_0 = 37$ supplied with the purer water.

and used ...

Single imputation?

Known total of $D = 66,153$ homes:

$$\widehat{D_1} = \tfrac{63}{100} \times 66,153 = 41,676 \quad \widehat{D_0} = \tfrac{37}{100} \times 66,153 = 24,477$$

$$\widehat{RateRatio} = \frac{286}{41,676} \div \frac{14}{24,477} = \frac{286}{63} \div \frac{14}{37} = 12.0$$

[Note: estimate does not require that we know what the total number of houses is, i.e., what the sampling fraction was.]

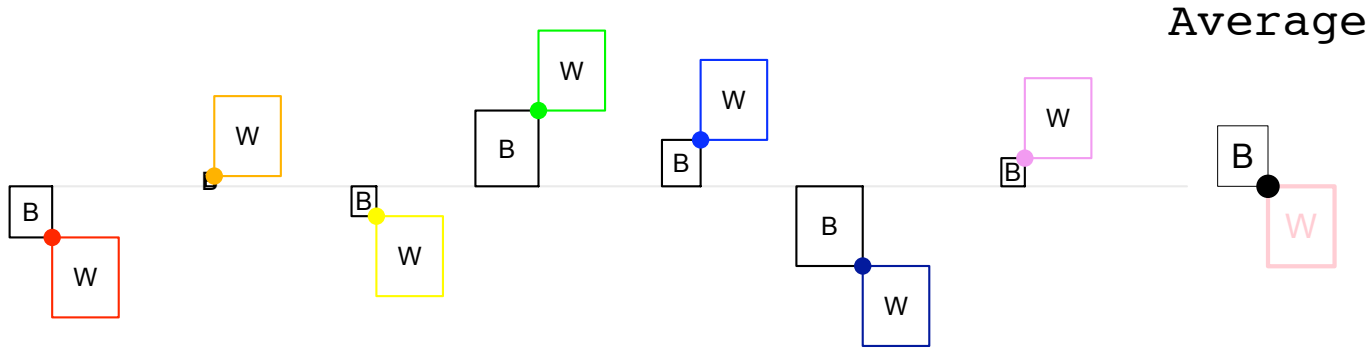
IF we (naïvely) assume that the $41,676$ and $24,477$ are correct, then the SE^2 for the log of the rate ratio is simply $1/286 + 1/14$.

However, they are not correct, and so this SE^2 is artificially low: it does not reflect fact that the 41,676 and 24,477 were inputed.

Multiple imputation? ...

Denominator Estimates, and Associated Rate Ratios

	Version:	1	2	3	4	5	6	7	
# Cases,	IMPURE:	286	286	286	286	286	286	286	NUMERATORS
# Cases,	PURE:	14	14	14	14	14	14	14	
prop'n	PURE:	0.674	0.626	0.658	0.572	0.597	0.695	0.612	
	IMPURE:	44570	41412	43499	37852	39465	45968	40454	DENOMINATORS
	PURE:	21583	24741	22654	28301	26688	20185	25699	
	RateRatio:	9.9	12.2	10.6	15.3	13.8	9	13	
W=Var[logRR]:		0.075	0.075	0.075	0.075	0.075	0.075	0.075	



W = Within-version Variance = 1/286 + 1/14 = 0.075

(large k) Average Between-version Variance = 1/63 + 1/37 = 0.043

MI: "Within" Component of Var[log RR]

The estimated denominators are treated as correct, and the numerators 286 and 14 are same from version to version.

SO...

Each $W = Var_{Within.version} = 1/c_1 + 1/c_0 = 1/286 + 1/14$

SO...

Average$[W] = 1/c_1 + 1/c_0 = 1/286 + 1/14$

MI:"Between" Component of Var[log RR]

$\widehat{D_1}$ and $\widehat{D_0}$ differ from version to version ...

$$B = Var_{between.version} = Var\left[\log\left\{\frac{\widehat{D_1}}{\widehat{D_0}}\right\}\right] = Var\left[\log\left\{\frac{\widehat{p}}{1-\widehat{p}}\right\}\right]$$

$$\widehat{p} \sim Beta(63+\alpha,\ 37+\beta) \rightarrow Var[\widehat{p}] = \frac{\frac{63+\alpha}{100+\alpha+\beta}\frac{37+\beta}{100+\alpha+\beta}}{101+\alpha+\beta} \approx \frac{63\times 37}{100^3}$$

$$Var\left[\log\left\{\frac{\widehat{p}}{1-\widehat{p}}\right\}\right] \approx \frac{63\times 37}{100^3}\times\left[\frac{1}{\frac{63}{100}\times\frac{37}{100}}\right]^2 = \frac{100}{63\times 37} = \frac{1}{63}+\frac{1}{37}.$$

$\frac{k+1}{k}$ negligible if k large enough. $d.f.[t] = (k-1)\left[1+\frac{k}{k+1}\frac{W}{B}\right]$

What if Snow had access to "Woolf's" 1955 Fomula?

- $\widehat{D_1} = (d_1/d) \times D\,; \quad \widehat{D_0} = (d_0/d) \times D$

$$\widehat{RR} = \frac{c_1/\widehat{D_1}}{c_0/\widehat{D_0}} = \frac{c_1/\{(d_1/d) \times D\}}{c_0/\{(d_0/d) \times D\}} = \frac{c_1/d_1}{c_1/d_0} = \frac{286/63}{14/37} = 12.0$$

- ('Woolf') variance of the log of this rate ratio estimate is

$$\left\{\frac{1}{c_1} + \frac{1}{c_0}\right\} + \left\{\frac{1}{d_1} + \frac{1}{d_0}\right\} = \left\{\frac{1}{286} + \frac{1}{14}\right\} + \left\{\frac{1}{63} + \frac{1}{37}\right\}.$$

Woolf merely used the formula. I have traced the proof back to Yule, 1900.

Summary

- Simulated case-control study with no additional informative data items: no $\downarrow$ in variance by imputing missing values.
- As $\frac{k+1}{k} \rightarrow 1$, Rubin variance formula $\rightarrow$ 'Woolf' formula.
- 'Controls' in 'c-c' study serve as 'denominator' series.
- Cohort studies and case-control studies are two versions of the same 'etiologic' study:
  - 'cohort' study: denominators are known
  - 'case-control' study: estimate (impute) denominators; additional variance is the additional price.

FUNDING & CO-ORDINATES

Natural Sciences and Engineering Research Council of Canada

Le Fonds québécois de la recherche sur la nature et les technologies

James.Hanley@McGill.CA

http://www.biostat.mcgill.ca/hanley