Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

A note on Wilson coverage intervals for proportions estimated from complex samples

by Phillip S. Kott

Release date: December 21, 2017

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at STATCAN.infostats-infostats.STATCAN@canada.ca

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-877-287-4369

Depository Services Program

• Inquiries line	1-800-635-7943
• Fax line	1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard table symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, December 2017
Vol. 43, No. 2, pp. 235-240
Statistics Canada, Catalogue No. 12-001-X

A note on Wilson coverage intervals for proportions estimated from complex samples

Phillip S. Kott[1]

Abstract

This note discusses the theoretical foundations for the extension of the Wilson two-sided coverage interval to an estimated proportion computed from complex survey data. The interval is shown to be asymptotically equivalent to an interval derived from a logistic transformation. A mildly better version is discussed, but users may prefer constructing a one-sided interval already in the literature.

Key Words: Effective sample size; Confidence interval; Logistic transformation.

1 Introduction

Brown, Cai and Dasgupta (2001) show that a method proposed by Wilson (1927) can produce reasonably well-behaved two-sided coverage intervals for a proportion under simple random sampling *with* replacement. Section 2 of this note discusses the theoretical foundations for extending this interval-construction method to estimated proportions computed from a complex survey. Section 3 shows that such a Wilson-type interval can be asymptotically equivalent to an interval derived from a logistic transformation. Section 4 offers some concluding remarks.

The term “coverage interval” is used here in place of the more common “confidence interval” because a 95% Wilson coverage interval does not attempt to cover the true proportion at least 95% of the time no matter what that proportion is. Instead, it merely tries to cover the true proportion 95% of the time for reasonable values of the true proportion. For some values it overcovers, for others it undercovers as shown in Brown et al. (2001). By limiting its applicability to two-sided coverage intervals, the Wilson methodology is (mostly) able to ignore the asymmetry of the distribution of an estimated proportion.

2 The extension

It is not hard to generalize Wilson coverage intervals (also called “score intervals”) to complex survey data. See, for example, Kott and Carr (1997). As with the Wilson itself, one simply solves this equation for the true proportion P :

$$\frac{(p-P)^2}{\left[\frac{P(1-P)}{n*}\right]} \le z^2_{1-\alpha/2}, \tag{2.1}$$

1. Phillip S. Kott, RTI International, 6110 Executive Blvd., Rockville, MD 20852, U.S.A. E-mail: pkott@rti.org.

where p is a consistent estimator for P under probability-sampling theory, and $z_{1-\alpha/2}$ is the Normal $z-$ score for $(1-\alpha/2)$ given the goal is to produce a $(1-\alpha)\,\%$ coverage interval (α is often set at 0.05). The missing piece to equation (2.1) is n^*, the so-called "effective sample size", which in the standard Wilson formulation is the sample size n. In our more general context, $n^* = p\,(1-p)/\mathrm{var}\,(p)$, where $\mathrm{var}\,(p)$ is a consistent estimator for the variance of p, $\mathrm{Var}\,(p)$.

In order to calculate n^*, we need both $p\,(1-p)$, and $\mathrm{var}\,(p)$ to be positive. In addition, let us assume that $1/n^* = O_P\,(1/n^a)$ for some positive $a \le 1$, $p - P = O_P\left(1/\sqrt{n^*}\right)$, $0 < \mathrm{Var}\,(p) = O\,(1/n^*)$, and $\mathrm{var}\,(p)/\mathrm{Var}\,(p)$ is $1 + O_P\left(1/\sqrt{n^*}\right)$. Note that the last three are always true under simple random sampling with replacement so long as $P\,(1-P) \ge B > 0$.

Dropping $O_P\left(1/[n^*]^{3/2}\right)$ terms, but allowing $p\,(1-p)$ to be small (effectively $o_p(1)$), one can derive this Wilson-like interval for P from equation (2.1):

$$p + \frac{1-2p}{n^*}\frac{z_{1-\alpha/2}^2}{2} - z_{1-\alpha/2}\left(\frac{p\,(1-p)}{n^*} + \frac{z_{1-\alpha/2}^2}{4\,(n^*)^2}\right)^{1/2} \le P$$

$$\le p + \frac{1-2p}{n^*}\frac{z_{1-\alpha/2}^2}{2} + z_{1-\alpha/2}\left(\frac{p\,(1-p)}{n^*} + \frac{z_{1-\alpha/2}^2}{4\,(n^*)^2}\right)^{1/2}. \tag{2.2}$$

We can call this the "complex-sampling Wilson coverage interval". WesVar (2007) computes a variant of this interval that does not drop $O_P\left(1/[n^*]^{3/2}\right)$ terms. It is dropped here because other terms of that size will be dropped later in this note.

If it is reasonable to drop $O_P\left(1/[n^*]^{3/2}\right)$ terms in deriving equation (2.2), one can also safely ignore the difference between $1/n$ and $1/(n-1)$. Under simple random sampling *without* replacement, $n^* = n/(1-f)$ (or $(n-1)/(1-f)$) where f is the sampling fraction. When f is very small, the distinction between with and without replacement sampling can be ignored.

Observe that under simple random sampling with replacement, the denominator of the pivotal appearing on the left-hand side of equation (2.1) has no variance at all. By contrast, the denominator in the traditional Wald pivotal, $\mathrm{var}\,(p) = p\,(1-p)/(n-1)$, can have considerable variance, especially when p or $1-p$ is small. That is why Wilson intervals have superior performance under simple random sampling, whether with or without replacement.

That superiority carries over to complex sampling (see, for example, Kott, Andersson and Nerman, 2001), where the pivotal's denominator is

$$\begin{aligned}\frac{P\,(1-P)}{n^*} = \mathrm{var}\,(p)\frac{P\,(1-P)}{p\,(1-p)} &= \mathrm{var}\,(p)\left[1 - \frac{(p-P)-(p^2-P^2)}{p\,(1-p)}\right]\\ &= \mathrm{var}\,(p)\left[1 - \frac{(p-P)-(p-P)\,(p+P)}{p\,(1-p)}\right]\\ &= \mathrm{var}\,(p) - \frac{1-2P}{n^*}\,(p-P) + \mathrm{O}_P\left(1/[n^*]^2\right),\end{aligned}$$

which is likely to have less variance than $\text{var}(p)$ in most applications. For an intuition into why this is so, observe that a putative variance estimator of the form $\text{var}_1(p) = \text{var}(p) - b(p - P)$ is minimized when $b = \text{Cov}[\text{var}(p), p] / \text{Var}(p)$. Under simple random sampling, whether with or without replacement, b is exactly $(1 - 2P)/n^*$.

Although the minimizing b is not exactly equal to $(1 - 2P)/n^*$, under more complex sampling designs, the optimal b is likely to be closer to $(1 - 2P)/n^*$ than to 0. It is thus not surprising that the variance of $\text{var}(p) - [(1 - 2P)/n^*](p - P)$ will usually be less than the variance of $\text{var}(p)$. Nevertheless, a slight improvement on the complex-sampling Wilson coverage interval can be made by replacing n^* in equation (2.2) by

$$\tilde{n} = [(1 - 2p)\,\text{var}(p)] / \text{cov}[\text{var}(p), p]$$

when $\text{cov}[\text{var}(p), p]$, a consistent estimator for $\text{Cov}[\text{var}(p), p]$, exists (see Kott et al., 2001).

As with the standard Wilson, the center of the complex-sample Wilson interval in equation (2.2) is slightly different from p when p is not $\frac{1}{2}$:

$$C = p + \frac{1 - 2p}{n^*} \frac{z_{1-\alpha/2}}{2}.$$

Its length L appears longer than the Wald's:

$$L = z_{1-\alpha/2} \left(\frac{p(1 - p)}{n^*} + \frac{z_{1-\alpha/2}^2}{4(n^*)^2} \right)^{1/2} > z_{1-\alpha/2} \left(\frac{p(1 - p)}{n^*} \right)^{1/2}.$$

When $P(1 - P) \geq B > 0$, however,

$$\begin{aligned} \left(\frac{p(1 - p)}{n^*} + \frac{z_{1-\alpha/2}^2}{4(n^*)^2} \right)^{1/2} &= \left(\frac{p(1 - p)}{n^*} \right)^{1/2} \left(1 + \frac{\frac{1}{4} z_{1-\alpha/2}^2}{n^* p(1 - p)} \right)^{1/2} \\ &= \left(\frac{p(1 - p)}{n^*} \right)^{1/2} + \text{o}_p\left(\frac{1}{n^*} \right). \end{aligned} \tag{2.3}$$

3 The logistic transformation

The complex-sampling Wilson coverage interval turns out to be very similar to this two-sided coverage interval derived using a logistic transformation (see Brown et al., 2001):

$$f^{-1}\left\{ f(p) - z_{1-\alpha/2} \sqrt{\text{var}[f(p)]} \right\} \leq P \leq f^{-1}\left\{ f(p) + z_{1-\alpha/2} \sqrt{\text{var}[f(p)]} \right\}, \tag{3.1}$$

where $f(p) = \log(p) - \log(1 - p)$, and $\text{var}[f(p)] = \left[f'(p) \right]^2 \text{var}(p) = [1/p + 1/(1 - p)]^2 p(1 - p)/n^* = 1/[n^* p(1 - p)]$. The original rationale for this interval appears to be that it has this desirable property: it cannot contain values less than 0 or greater than 1, which would be nonsensical for a proportion.

The left-hand side of equation (3.1) can be rewritten as $g\,(x-h)$, where

$$g\,(y) = f^{-1}\,(y) = [1+\exp\,(-y)]^{-1},\; x = f\,(p) = \log\left(\frac{p}{1-p}\right),$$

and

$$h = \frac{z_{1-\alpha/2}}{\sqrt{n^*\,p\,(1-p)}}\,.$$

The first and second derivatives of $g\,(y)$ are $g'\,(y) = g\,(y)\,[1-g\,(y)]$, and $g''\,(y) = g\,(y)\,[1-g\,(y)]\,[1-2g\,(y)]$. Invoking the mean value theorem, there is an h^* between 0 and h such that

$$\begin{aligned}
g\,(x-h) &= g\,(x) - g'\,(x)\,h + \frac{1}{2}\,g''\,(x-h^*)\,h^2 \\
&= p - p\,(1-p)\,\frac{z_{1-\alpha/2}}{\sqrt{n^*\,p\,(1-p)}} \\
&+ \frac{1}{2}\left[1+\left(\frac{1-p}{p}\right)e^{h^*}\right]^{-1}\left\{1-\left[1+\left(\frac{1-p}{p}\right)e^{h^*}\right]^{-1}\right\}\left\{1-2\left[1+\left(\frac{1-p}{p}\right)e^{h^*}\right]^{-1}\right\}\frac{z^2_{1-\alpha/2}}{n^*\,p\,(1-p)} \\
&= p - p\,(1-p)\,\frac{z_{1-\alpha/2}}{\sqrt{n^*\,p\,(1-p)}} \\
&+ \frac{1}{2}\,\frac{p}{1+(1-p)\,(e^{-h^*}-1)}\,\frac{(1-p)-(1-p)\,(e^{h^*}-1)}{1+(1-p)\,(e^{h^*}-1)}\,\frac{(1-2p)-(1-p)\,(e^{h^*}-1)}{1+(1-p)\,(e^{h^*}-1)}\,\frac{z^2_{1-\alpha/2}}{n^*\,p\,(1-p)}\,,
\end{aligned}$$

using

$$\left[1+\left(\frac{1-p}{p}\right)e^{h^*}\right]^{-1} = \frac{p}{1+(1-p)\,(e^{h^*}-1)}\,.$$

An analogous derivation can be made for the right-hand side of equation (3.1).

Consequently,

$$\begin{aligned}
p + \frac{1-2p}{n^*}\,\frac{z^2_{1-\alpha/2}}{2} - z_{1-\alpha/2}\left(\frac{p\,(1-p)}{n^*}\right)^{1/2} + o_P\left(\frac{1}{n^*}\right) &\le P \\
\le\; p + \frac{1-2p}{n^*}\,\frac{z^2_{1-\alpha/2}}{2} + z_{1-\alpha/2}\left(\frac{p\,(1-p)}{n^*}\right)^{1/2} &+ o_P\left(\frac{1}{n^*}\right).
\end{aligned}$$

After invoking the asymptotic equality in equation (2.3) and dropping $o_P\,(1/n^*)$ terms, the last set of inequalities is equivalent to Wilson interval in equation (2.2) so long as n^* is sufficiently large and $P\,(1-P) > 0$, the latter meaning that the true proportion is neither 0 or 1.

4 Some concluding remarks

The asymptotic equivalence of a coverage interval based on a logistic transformation to the theoretically grounded Wilson interval is the main contribution of this paper. Although in the asymptotic framework, $P(1-P)$ is fixed and positive as n^* grows large, in practice it is the size of $p(1-p)n^*$ that matters when comparing the Wilson-type and logistic-transformation intervals. This requires that $P(1-P)$ not be too small.

Brown et al. (2001) show empirically that under simple random sampling (with $n=50$), coverage intervals derived from the logistic transformation tend to be larger than corresponding Wilson intervals for small values of $P(1-P)$. Kott and Liu (2009) make the same observation for one-sided intervals based on complex samples, supporting the notion that it is a better choice.

The asymptotic equivalence of the logistic-transformation interval with the Wilson interval explains the former's empirical superiority in the literature (e.g., in Brown et al., 2001) to an analogous interval constructed using an arcsine transformation. Because $\arcsin(p)$ has a constant large-sample variance under simple random sampling no matter the true value of P (so long as $P(1-P)>0$), it has been hoped that the arcsine transformation would be ideal for interval construction.

Better than a Wilson interval, but not yet incorporated into any software package I know of, is the one-sided coverage intervals for P derived using an Edgeworth expansion on $p-P$ in Kott and Liu (2009). That method produces this two-sided interval:

$$p+\frac{1-2p}{\tilde{n}}\left(\frac{1}{6}+\frac{z_{1-\alpha/2}^2}{3}\right)-z_{1-\alpha/2}\left(\operatorname{var}(p)+\left[\frac{1-2p}{\tilde{n}}\left(\frac{1}{6}+\frac{z_{1-\alpha/2}^2}{3}\right)\right]^2\right)^{1/2}\le P$$
$$\le p+\frac{1-2p}{\tilde{n}}\left(\frac{1}{6}+\frac{z_{1-\alpha/2}^2}{3}\right)+z_{1-\alpha/2}\left(\operatorname{var}(p)+\left[\frac{1-2p}{\tilde{n}}\left(\frac{1}{6}+\frac{z_{1-\alpha/2}^2}{3}\right)\right]^2\right)^{1/2},$$

where $\tilde{n}=[(1-2p)\operatorname{var}(p)]\,/\,\operatorname{cov}[\operatorname{var}(p),\ p]$, and $\operatorname{cov}[\operatorname{var}(p),\ p]$, a consistent estimator for $\operatorname{Cov}[\operatorname{var}(p),\ p]$, exists and equals a consistent estimator for the third moment of p. Note that $\operatorname{cov}[\operatorname{var}(p),\ p]$ doesn't exist for designs with only two primary sampling units per stratum. Moreover, it is not a consistent estimator for the third moment of p when finite population correction matters.

Observe that $\tilde{n}$ again replaces n^*. In addition, $1/6+z_{1-\alpha/2}/3$ replaces $z_{1-\alpha/2}/2$, which means that the center will often be closer to the p using this interval rather than the Wilson. The good coverage properties of this interval, like the Wilson, breaks down when the skewness coefficient of $p\left(\mathrm{E}\left[(p-P)^3\right]/[\operatorname{Var}(p)]^{3/2}\right)$ gets too large in absolute value, how large has yet to be determined.

Finally, SAS/STAT (SAS Institute Inc., 2010) offers a Wilson coverage interval for estimated proportions in its SURVEYFREQ procedure. The procedure's method of adjusting the effective sample size, which can – and should – be turned off, is not related to the $\tilde{n}$ discussed here. Instead, it is based on an ad-hoc t – adjustment that sadly is not related to the variance of the denominator variance of the Wilson pivotal.

Acknowledgements

The author thanks Per Gösta Andersson for introducing me to this area of research and an anonymous referee for correcting errors in a previous version of the manuscript. Remaining errors are my own.

References

Brown, L.D., Cai, T. and Dasgupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 16, 101-133.

Kott, P.S., and Carr, D.A. (1997). Developing an estimation strategy for a pesticide data program. *Journal of Official Statistics*, 13, 367-383.

Kott, P.S., and Liu, Y.K. (2009). One-sided coverage intervals for a proportion estimated from a stratified simple random sample. *International Statistical Review/Revue Internationale de Statistique*, 77, 251-265.

Kott, P.S., Andersson, P.G. and Nerman, O. (2001). Two-sided coverage intervals for small proportion based on survey data. Presented at Federal Committee on Statistical Methodology Research Conference, Washington, DC. http://fcsm.sites.usa.gov/files/2014/05/2001FCSM_Kott.pdf.

SAS Institute Inc. (2010). *SAS/STAT® 9.22 User's Guide*. Cary, NC: SAS Institute Inc. http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_surveyfreq_a0000000252.htm.

WesVar (2007). *WesVar® 4.3 Users' Guide*, B28-B29.

Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22, 209-212.