

Better One-Sided Coverage Intervals

Phillip S. Kott

pkott@rti.org

Senior Research Statistician

RTI International

(based on work with Yan K. Liu)

Outline

A Motivating Example

Why Wald Intervals Fail for a Proportion

A Two-sided Answer: Wilson Intervals

Stratified Samples

The Edgeworth Expansion and One-sided Intervals

Examples

Investigating the Asymptotics

Generalization

A Motivating Example:

Former President Bush's Approval among Blacks

Suppose $n = 100$, and $\hat{p} = \sum_S y_k / n = .02$ or 2%.

The usual Wald 95% confidence interval is

$$p \approx .02 \pm 2 \{ \hat{p}(1 - \hat{p})/100 \}^{1/2} \approx .02 \pm 2 \{ .14/10 \},$$

which includes 0.

Even a one-sided interval includes 0.

Why Wald Intervals Fail for a Proportion

\hat{p} is not normal.

It is neither continuous nor symmetric.

Worse: The estimated variance, $v(\hat{p}) \approx \hat{p}(1 - \hat{p})/n$,
is not the true variance, $V(\hat{p}) = p(1 - p)/n$.

The Wilson Interval

Solve $\frac{|\hat{p} - p|}{\{p(1-p)/n\}^{1/2}} \leq z_{1-\frac{\alpha}{2}}$ (e.g., $\alpha = 5\%$ or $.05$)

for p (by squaring both sides and solving the quadratic).

By being two-sided, the resulting Wilson (or score) interval effectively makes \hat{p} symmetric.

The Wilson interval is designed to cover p roughly $1-\alpha\%$ of the time, rather than *at least* $1-\alpha\%$ of the time. It is not a *confidence* interval.

The Wilson $1-\alpha\%$ coverage interval is approximately

$$\hat{p} + \frac{(1-2\hat{p})z_{1-\frac{\alpha}{2}}^2}{2n} \pm \left\{ \frac{z_{1-\frac{\alpha}{2}}^2 \hat{p}(1-\hat{p})}{n} + \frac{z_{1-\frac{\alpha}{2}}^4}{4n^2} \right\}^{1/2}$$

(e.g., $.02 + .02 \pm .035$)

Notice what happens when \hat{p} is 0 or 1.

Stratified Samples

Suppose the population is naturally divided into H mutually exclusive groups (strata), each with possibly its own p_h (e.g., H distinct sex/age groups).

The target of estimation is now

$$p = \sum_{h=1}^H W_h p_h,$$

where W_h is the fraction of the population in group h .

An unbiased estimator for p is

$$\hat{p} = \sum_{h=1}^H W_h \hat{p}_h, \quad \text{where } \hat{p}_h = \frac{1}{n_h} \sum_{k \in S_h} y_k$$

(or $\hat{p} = \frac{1}{n} \sum_S \omega_k y_k$, where $\omega_k = n W_h / n_h$ for k in group h .)

Unfortunately, the variance, $V(\hat{p}) = \sum W_h^2 p_h (1 - p_h) / n_h$,

in the pivotal,

$$\frac{\hat{p} - p}{\{V(\hat{p})\}^{1/2}}$$

is not a function of p .

One solution is to assume an *iid* model and replace

$$V(\hat{p}) \text{ with } V_{iid}(\hat{p}) = \left(\sum^H W_h^2 / n_h \right) p(1-p).$$

An alternative is to use the *idealized estimator*

under the general model:

$$v^*(\hat{p}) = v(\hat{p}) - B(\hat{p} - p),$$

where
$$B = \frac{\text{Cov}(v(\hat{p}), \hat{p})}{V(\hat{p})}$$

is consistently estimated by

$$b = \frac{m_3}{v(\hat{p})} = \frac{\sum_{h=1}^H \frac{W_h^3}{(n_h - 1)(n_h - 2)} \hat{p}_h(1 - \hat{p}_h)(1 - 2\hat{p}_h)}{\sum_{h=1}^H \frac{W_h^2}{n_h - 1} \hat{p}_h(1 - \hat{p}_h)}.$$

under “mild” conditions.

$v^*(\hat{p})$ is much more efficient than $v(\hat{p})$.

It is usually more efficient than the *ad hoc* construction:

$$v_{ah}(\hat{p}) = \frac{p(1-p)}{\hat{n}^*} = \frac{p(1-p)}{\hat{p}(1-\hat{p})} v(\hat{p}) \left(\approx v(\hat{p}) - \frac{1-2p}{n^*} (\hat{p} - p) \right).$$

But it can have an asymptotically relevant variance.

Under stratified sampling that variance can be estimated and the pivotal's *effective degrees of freedom* computed.

The general Wilson $1-\alpha\%$ coverage interval is now

$$\hat{p} + \delta_W \pm \left\{ z_{1-\frac{\alpha}{2}}^2 v(\hat{p}) + \delta_W^2 \right\}^{1/2},$$

where $\delta_W = \frac{z_{1-\frac{\alpha}{2}}^2 m_3}{2 v(\hat{p})}$.

(perhaps replacing $z_{1-\frac{\alpha}{2}}$ with a $t_{1-\frac{\alpha}{2}}$).

It cannot be used when \hat{p} is 0 or 1.

Use the *iid* version in its place?

One-Sided Intervals

Assume $\hat{\rho}$ can be approximated by a continuous distribution sharing its first three central moments.

Its Edgeworth expansion would then be

$$\Pr\left(\frac{\hat{\rho} - \rho}{\sqrt{V(\hat{\rho})}} \leq z\right) = \Phi(z) + (1/6)(1 - z^2)\varphi(z)\tau + O(1/n),$$

where $\Phi(z)$ and $\varphi(z)$ are the *cdf* and *pdf* of the standard normal distribution,

and the skewness coefficient of $\hat{\rho}$ is

$$\begin{aligned} \tau &= \frac{E[(\hat{p} - p)^3]}{\{E[(\hat{p} - p)^2]\}^{3/2}} = \frac{M_3}{[V(\hat{p})]^{3/2}} = \frac{\sum_{h=1}^H \frac{W_h^3}{n_h^2} p_h(1-p_h)(1-2p_h)}{\left[\sum_{h=1}^H \frac{W_h^2}{n_h} p_h(1-p_h) \right]^{3/2}} \\ &= \frac{O(1/n^2)}{[O(1/n)]^{3/2}} = O(1/n^{1/2}). \end{aligned}$$

The asymptotics need some mild conditions

(e.g., all $nW_h/n_h < U < \infty$; all $p_h(1-p_h) > L > 0$).

After a bit of Calculus and a lot of algebra:

$$\Pr\left((p - \hat{p})^2 - (1/3)(1 - z^2) \frac{M_3}{V(\hat{p})} (p - \hat{p}) + O(1/n^2) \leq z^2 V(\hat{p})\right) = \Phi(z).$$

Replacing $\frac{M_3}{V(\hat{p})}$ by b , and

$V(\hat{p})$ by $v(\hat{p}) - b(\hat{p} - p)$

leads to the one-sided $1-\alpha\%$ coverage intervals:

$$p \leq \hat{p} + \delta_E + \left\{ z_{1-\alpha}^2 v(\hat{p}) + \delta_E^2 \right\}^{1/2}, \text{ and}$$

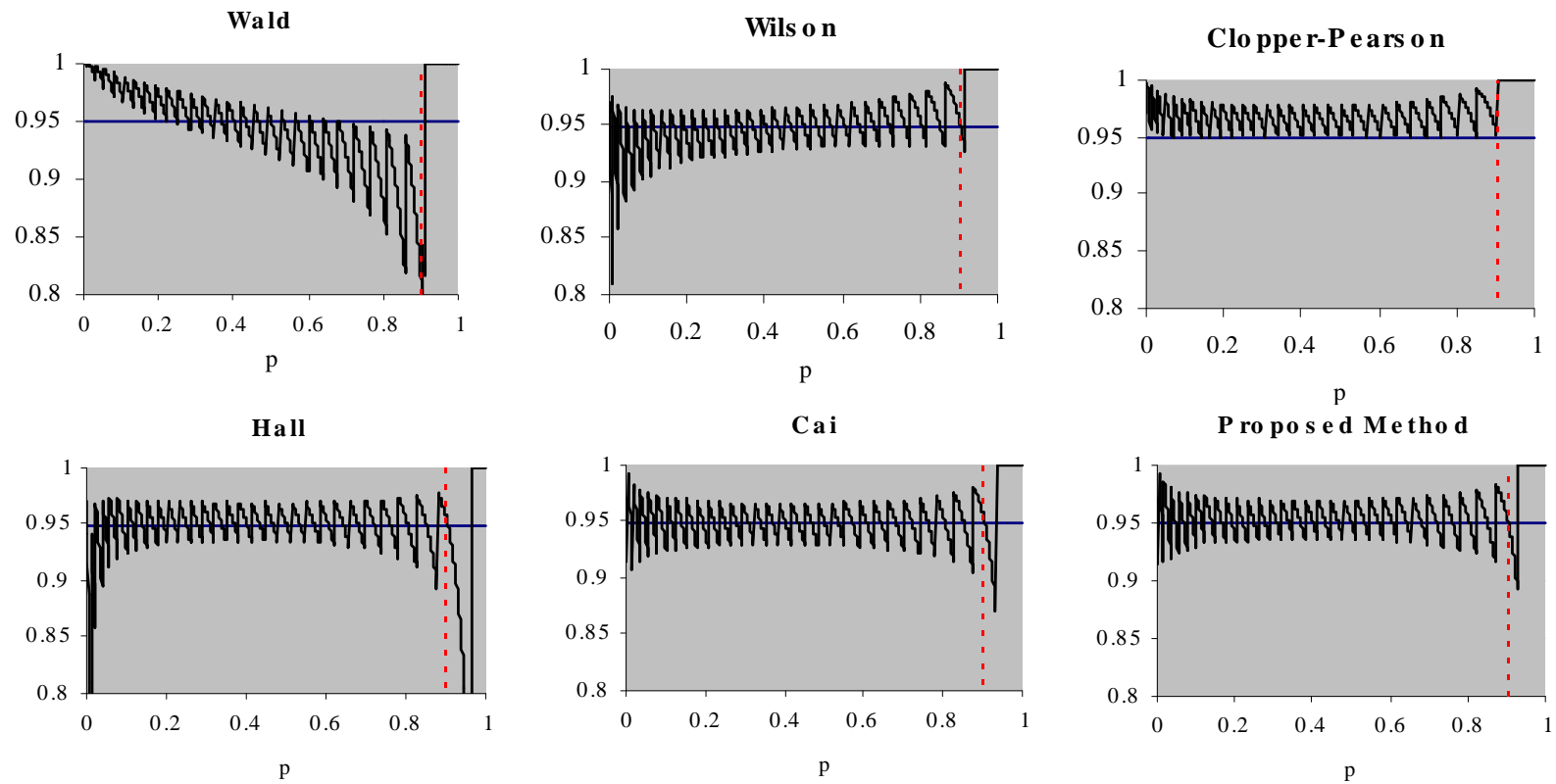
$$p \geq \hat{p} + \delta_E - \left\{ z_{1-\alpha}^2 v(\hat{p}) + \delta_E^2 \right\}^{1/2}$$

where

$$\delta_E = \left(\frac{1 - z_{1-\alpha}^2}{6} + \frac{z_{1-\alpha}^2}{2} \right) \frac{m_3}{v(\hat{p})} = \left(\frac{1}{6} + \frac{z_{1-\alpha}^2}{3} \right) \frac{m_3}{v(\hat{p})}.$$

An Example:

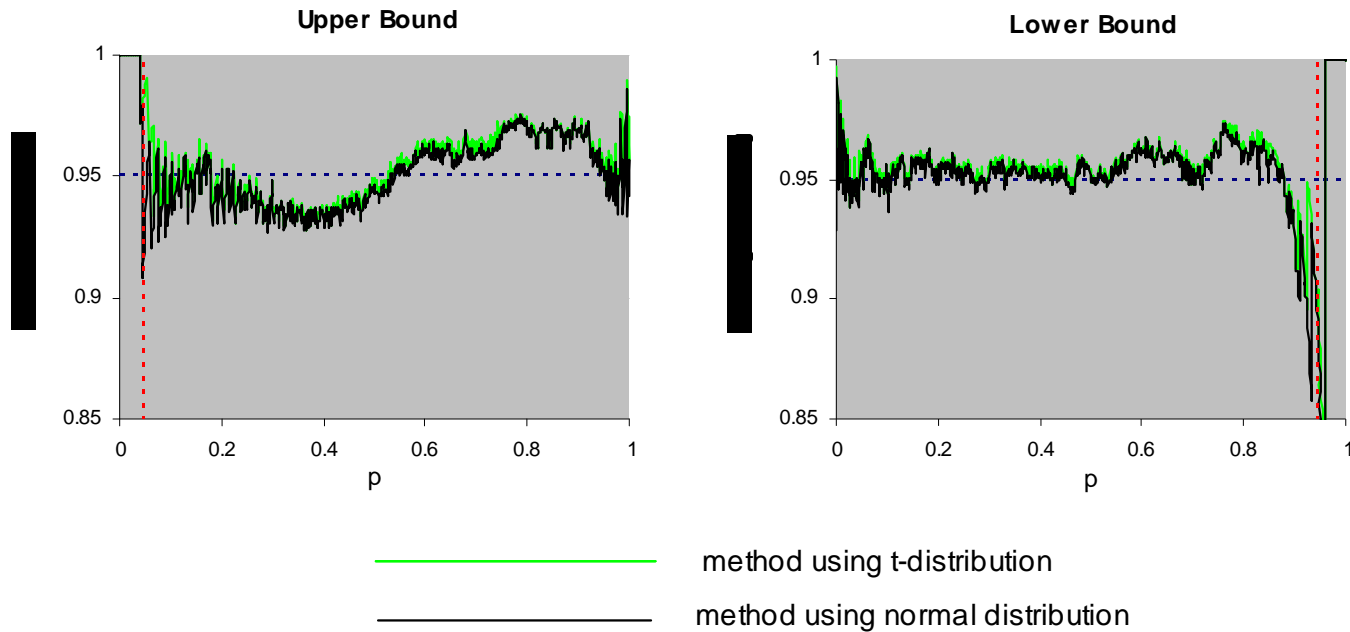
Coverages with a nominal 95% lower bound
under simple random sampling with $n = 30$



Another Example:

Three strata: $n_1 = 10$; $n_2 = 20$; $n_3 = 30$;

$$p_1 = p - p(1-p); p_2 = p; p_3 = p + p(1-p).$$



Investigating the Asymptotics

$$\begin{aligned}
 p \leq \hat{p} + \left(\frac{1}{6} + \frac{z_{1-\alpha}^2}{3} \right) \frac{m_3}{v(\hat{p})} + \left\{ z_{1-\alpha}^2 v(\hat{p}) + \left[\left(\frac{1}{6} + \frac{z_{1-\alpha}^2}{3} \right) \frac{m_3}{v(\hat{p})} \right]^2 \right\}^{1/2} \\
 \downarrow \qquad \qquad \qquad \downarrow \qquad \qquad \qquad \downarrow \\
 O_P(1/n) \quad + \quad \{ O_P(1/n) \quad + \quad O_P(1/n^2) \}^{1/2} \\
 O_P(1/n) \quad + \quad O_P(1/n^{1/2}) \quad + \quad O_P(1/n^{3/2})
 \end{aligned}$$

so long as $1/[nv(\hat{p})] = O_P(1)$.

Should the last term (the δ_E^2) be dropped?

It does not appear in Hall (*Biometrika*, 1982)
or Cai (*JSPI*, 2004).

If $1/\{nV(\hat{p})\}$ is too large, the Edgeworth expansion fails
because it needs

$$\tau = \frac{M_3}{\{V(\hat{p})\}^{3/2}} = \frac{1}{n^{1/2}} \frac{n^2 M_3}{\{nV(\hat{p})\}^{3/2}} \leftarrow O(1)$$

to be small.

Generalization

Suppose we have a sampling scheme and a consistent estimator $\hat{\theta}$ for θ such that

$$\text{Cov}(v(\hat{\theta}), \hat{\theta}) = E\left[(\hat{\theta} - \theta)^3\right].$$

Furthermore, suppose $v(\hat{\theta})$ is a consistent estimator for $E\left[(\hat{\theta} - \theta)^2\right]$, and m_3 is a consistent estimator for $E\left[(\hat{\theta} - \theta)^3\right]$.

(Counterexample: stratified *srs* when *fpc* matters).

We can then derive these one-sided $1-\alpha\%$ coverage intervals:

$$\theta \leq \hat{\theta} + \delta_E + \left\{ z_{1-\alpha}^2 v(\hat{\theta}) + \delta_E^2 \right\}^{1/2}, \text{ and}$$

$$\theta \geq \hat{\theta} + \delta_E - \left\{ z_{1-\alpha}^2 v(\hat{\theta}) + \delta_E^2 \right\}^{1/2}$$

where $\delta_E = \left(\frac{1}{6} + \frac{z_{1-\alpha}^2}{3} \right) \frac{m_3}{v(\hat{\theta})}$

under mild conditions.

Examples :

The difference between two domain proportions based on data from a complex sample (assuming a least three primary sampling units selected per first-stage stratum).

The total error in an stratified simple random auditing sample where only a small number of sampled elements have positive values (i.e., errors).

Concluding Remark

The two papers with Yan Liu can be found in a single document on my NASS website:

<http://www.nass.usda.gov/research/OD5.htm>

Thank You!