9/23/25

Stat 440 Topics & Equations for Lecture on Conf Intervals for Proportions

General topic: **Confidence intervals for one or more proportions based on SRS samples** of n units from $U = \{1, 2, \ldots, N\}$. Let A (and later also $B, C, \ldots$) be a subset of U with N_A elements (respectively $N_B, N_C, \ldots$). The true proportions to be estimated are

$$\pi_A = N_A/N \ , \qquad \pi_B = N_B/N \ , \qquad \textit{etc.}$$

General idea: for the different subpopulations define different **attributes**

$$y_i^A = I_{[i \in A]} \ , \qquad y_i^B = I_{[i \in B]} \ , \qquad \textit{etc.}$$

SRS sampling notations for Proportion Estimators

Based on the attributes $y_i^A,\ y_i^B,\ \ldots$ for $i = 1, \ldots, N$,

Population average: $\bar{y}_U^A \;=\; \frac{1}{N} \sum_{i=1}^N I_{[i \in A]} \;=\; N_A/N \;=\; \pi_A$

Sample average: $\bar{y}_S^A \;=\; \frac{1}{n} \sum_{i \in S} I_{[i \in A]} \;=\; \frac{|S \cap A|}{n} \;\equiv\; \frac{n_A}{n} \;=\; \widehat{\pi}_A$

Theoretical SRS Variance of $\bar{y}_S^A$ is $\frac{N-n}{n\,N}\, s^2_{y^A,U}$, where

$$s^2_{y^A,U} \;=\; \frac{1}{N-1} \sum_{i=1}^N (I_{[i \in A]} - \pi_A)^2 \;=$$

$$\frac{1}{N-1}\Big\{ N_A\,(1-\pi_A)^2 + (N - N_A)\,(-\pi_A)^2 \Big\}$$

$$s^2_{y^A,U} \;=\; \frac{1}{N^2\,(N-1)}\Big(N_A\,(N-N_A)^2 + N_A^2\,(N-N_A)\Big)$$

$$= \frac{N_A\,(N-N_A)}{N\,(N-1)} \;=\; \frac{N}{N-1}\,\pi_A\,(1-\pi_A)$$

And we conclude (*next-to-last formula, Thompson Sec. 5.1*):

$$\mathrm{Var}(\hat{\pi}_A) \;=\; \pi_A\,(1-\pi_A)\,\frac{N}{N-1}\,\Big(\frac{1}{n} - \frac{1}{N}\Big)$$

Similarly the sample variance is $s^2_{y^A,S} \;=\; \frac{n}{n-1}\,\hat{\pi}_A\,(1-\hat{\pi}_A)$.

So $\mathrm{Var}(\hat{\pi}_A)$ is unbiasedly estimated (*last formula in Sec. 5.1*) by

$$\widehat{\mathrm{Var}}(\hat{\pi}_A) \;=\; \hat{\pi}_A\,(1-\hat{\pi}_A)\,\frac{n}{n-1}\,\Big(\frac{1}{n} - \frac{1}{N}\Big)$$

Confidence Interval Formulas

First give the approximate large-sample CI: a **special case of CLT-based CI's from Chap. 3.**

Under conditions (large N, n, and π_A not too close to 0 or 1) guaranteeing CLT for $\bar{y}_S^A = \hat{\pi}_A$, **with probability** $\approx 1 - \alpha$

$$\pi_A \in \frac{n_A}{n} \pm z_{\alpha/2} \left[\pi_A (1 - \pi_A) \frac{N}{N-1} \left(\frac{1}{n} - \frac{1}{N} \right) \right]^{1/2}$$

Also with probability $\approx 1 - \alpha$,

$$\pi_A \in \frac{n_A}{n} \pm z_{\alpha/2} \left[\hat{\pi}_A (1 - \hat{\pi}_A) \frac{n}{n-1} \left(\frac{1}{n} - \frac{1}{N} \right) \right]^{1/2}$$

$$\approx \frac{n_A}{n} \pm t_{n-1,\alpha/2} \left[\frac{n_A (n - n_A)}{n^2 (n-1)} \cdot \frac{N-n}{N} \right]^{1/2}$$

In last formula, $t_{n-1,\alpha/2}$ replaces $z_{\alpha/2}$ by analogy with *iid* case, finite-sample (with-replacement or huge N), $\mathcal{N}(\mu, \sigma^2)$ attributes. *But* $t_{m-1} \approx \mathcal{N}(0,1)$ *whenever* $n \geq 50$.

Exact Hypergeometric Confidence Interval

Under SRS, $n_A \sim$ Hypergeometric(N_A, $N - N_A$, n), so "exact" CI with coverage probability sure to be $\geq 1 - \alpha$ is $[k_1/N,\ k_2/N]$

with $\texttt{phyper}(k_1 - 1,\ \lceil N\pi_A \rceil,\ N - \lceil N\pi_A \rceil,\ n) \overset{<}{\approx} \alpha/2$

and $\texttt{phyper}(k_2,\ \lceil N\pi_A \rceil,\ N - \lceil N\pi_A \rceil,\ n) \overset{>}{\approx} 1 - \alpha/2$

Exact Hypergeometric CI, continued

This is an instance of a **test-based** CI: we would

accept the hypothesis $H_0: \ N_A \le N\,\pi$ at level $\alpha/2$ at $n_A = k$ whenever the cdf Phyper$_{N,\lceil N\pi \rceil,n}(k) \ \ge \ 1 - \alpha/2$

and we would

accept the hypothesis $H_0: \ N_A \ge N\,\pi$ at level $\alpha/2$ at $n_A = k$ whenever complementary cdf 1-Phyper$_{N,\lfloor N\pi \rfloor,n}(k-1) \ \ge \ 1-\alpha/2$

So the two-sided $1 - \alpha$ Confidence Interval is the set of $\pi = \pi_A$ values for which both inequalities hold for $k = n_A$

Sample Size Formulas (based on Large-Sample CIs)

If we know π_A **roughly** and want to find n large enough to bracket π_A within a level $1-\alpha$ (large-sample, approximate) confidence interval of specified half-width δ (e.g., 0.01), then

without finite-population correction, n must be at least the smallest integer for which

$$\delta \;\geq\; z_{\alpha/2}\left[\pi_A\,(1-\pi_A)\,\frac{N}{N-1}\cdot\frac{1}{n}\right]^{1/2}$$

equivalently $$n \;\geq\; n_0 \;\equiv\; \frac{z^2_{\alpha/2}\,\pi_A\,(1-\pi_A)}{\delta^2}\cdot\frac{N}{N-1}$$

Sample Size Formulas, continued

with finite-population correction, n must be at least the smallest integer for which

$$\delta \ \geq \ z_{\alpha/2} \left[\pi_A \, (1 - \pi_A) \, \frac{N}{N-1} \Big(\frac{1}{n} - \frac{1}{N} \Big) \right]^{1/2}$$

$$\Longrightarrow \qquad n \ \gtrapprox \ \left[\frac{1}{N} + \frac{1}{n_0} \right]^{-1} \ = \ n_0 \cdot \frac{N}{N + n_0}$$

With/without replacement, if $\pi_A \leq \pi_{max} \leq 1/2$, then $\pi_A(1 - \pi_A) \ \leq \ \pi_{max}(1 - \pi_{max})$, and get **conservative** CI and sample-size by replacing $\pi_A(1-\pi_A)$ with $\pi_{max}(1-\pi_{max})$ in n_0.

The default, most conservative interval uses $\pi_{max} = 1/2$.

Sample Size Formulas for Multiple Proportions

To estimate several proportions π_A, π_B, ..., then the usual idea is to meet the constraints for adequate sample size for all of them **separately**, $P(|n_A/n - N_A/N| \le \delta) \ge 1 - \alpha$, etc., then take

$$n_0 \;=\; \frac{z^2_{\alpha/2}\, N}{\delta^2\,(N-1)} \cdot \max\left\{\pi_G(1-\pi_G) : \; G = A,\, B,\, \ldots\right\}, \quad n \;\ge\; \frac{n_0\, N}{N + n_0}$$

The problem discussed in Sec. 5.4 of Thompson is different and less common, to find n large enough so that for **disjoint** subsets $A, B, \ldots$ whose union is U, the prob. is $>$ 0.95 that **all** of the proportions satisfy $|n_G/n \,-\, N_G/N| \,\le\, d$, for $G = A, b, \ldots,$.

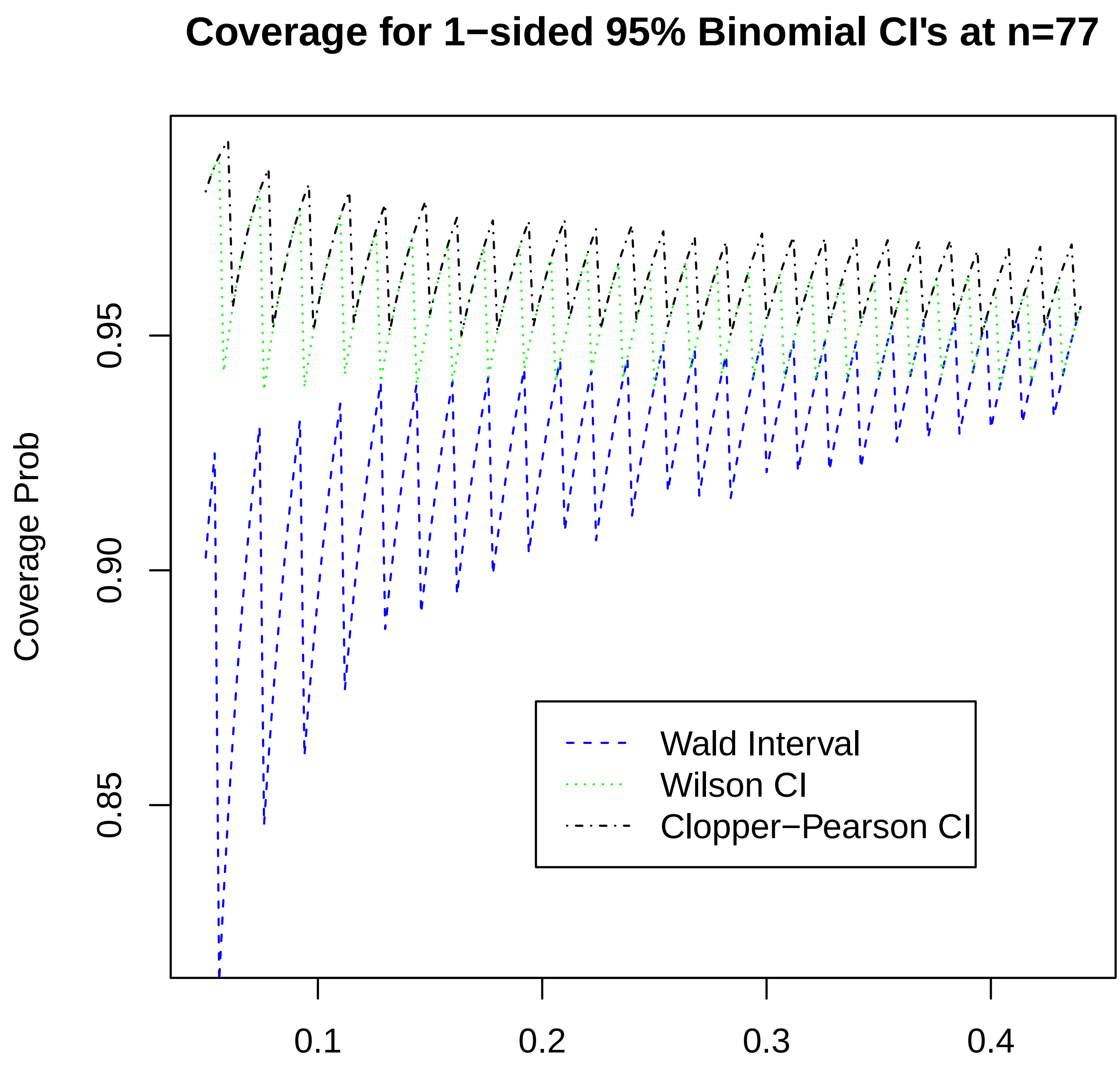
Conf Int's with Better Coverage

The (large-sample approximate) confidence intervals we use are called **Wald CIs**. Even in the $N \to \infty$ or with-replacement case (no **fpc**) they are known to be erratic for surprisingly large sample sizes n, depending on true $p = N_A/N$. See picture at `https://www.math.umd.edu/~slud/s701/RScripts/BinomialCvrg_n77.pdf` (included on next slide) for an illustration with $n = 77$.

Wald and (improved-coverage) **Wilson** CIs are respectively:

$$\left\{ p : \frac{(\hat{\pi}_A - p)^2}{\hat{\pi}_A(1-\hat{\pi}_A)\frac{N-n}{n(N-1)}} \le z^2_{\alpha/2} \right\} \ , \quad \left\{ p : \frac{(\hat{\pi}_A - p)^2}{p(1-p)\frac{N-n}{n(N-1)}} \le z^2_{\alpha/2} \right\}$$

Wilson involves solving quadratic inequality.


Coverage for 1−sided 95% Binomial CI's at n=77
Coverage Prob
0.85
0.90
0.95
0.1
0.2
0.3
0.4
Wald Interval
Wilson CI
Clopper−Pearson CI