Good Confidence Intervals for Categorical Data Analyses

Alan Agresti

Professor Emeritus

Department of Statistics

University of Florida

Outline of my talk

- *Score test-based confidence interval* (CI) as alternative to Wald, likelihood-ratio-test-based large-sample intervals
- Relation of score-test-based inference to *Pearson chi-squared* statistic, its application for variety of categorical data analyses
- *Pseudo-score CI* for model parameters based on generalized Pearson statistic comparing models
- Good *small-sample "exact"* pseudo-score CIs
- An *"adjusted Wald"* pseudo-score method performs well for simple cases (e.g., proportions and their difference) by approximating the score CI

Inverting tests to obtain CIs

For parameter β, consider CIs based on inverting standard tests of H_0: $\beta = \beta_0$

(95% CI is set of β_0 for which P-value > 0.05)

Most common approach inverts one of three standard asymptotic chi-squared tests: Likelihood-ratio (Wilks 1938), Wald (1943), score (Rao 1948)

For log likelihood $L(\beta)$, denote

maximum likelihood (ML) estimate by $\hat{\beta}$

score $u(\beta) = \partial L(\beta)/\partial\beta$

information $\iota(\beta) = -E[\partial^2 L(\beta)/\partial\beta^2]$

Wald, likelihood-ratio, score large-sample inference

- Wald test: $[(\hat{\beta} - \beta_0)/SE]^2 = (\hat{\beta} - \beta_0)^2 \iota(\hat{\beta})$.
  e.g., 95% Wald CI is $\hat{\beta} \pm 1.96(SE)$
- Likelihood-ratio (LR) statistic: $-2[L(\beta_0) - L(\hat{\beta})]$
- Rao's score test statistic:

$$\frac{[u(\beta_0)]^2}{\iota(\beta_0)} = \frac{[\partial L(\beta)/\partial \beta_0]^2}{-E[\partial^2 L(\beta)/\partial \beta_0^2]}$$

  where the partial derivatives are evaluated at β_0
  (For canonical GLMs, this is standardized sufficient stat.)

The three methods are asymptotically equivalent under H_0.

In practice, Wald inference popular because of simplicity, ease of forming it using software output.

Examples of score-test-based inference

- Pearson chi-squared test of independence in two-way contingency table
- McNemar test for binary matched-pairs
- Cochran–Mantel–Haenszel test of conditional independence for stratified 2×2 tables
- Cochran–Armitage trend test for several ordered binomials
- $Y \sim$ binomial(n, π), $\hat{\pi} = y/n$

  Test of H_0: $\pi = \pi_0$ uses

  $z = \frac{\hat{\pi}-\pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \sim$ N(0,1) null distribution (or $z^2 \sim \chi_1^2$).

  Inverting two-sided test gives Wilson CI for π (1927).

  (Wald 95% CI is $\hat{\pi} \pm 1.96\sqrt{\hat{\pi}(1-\hat{\pi})/n}$.)

Wald inference can be poor for categorical data

- Hauck and Donner (1977) showed aberrant behavior in logistic regression when effect is strong.

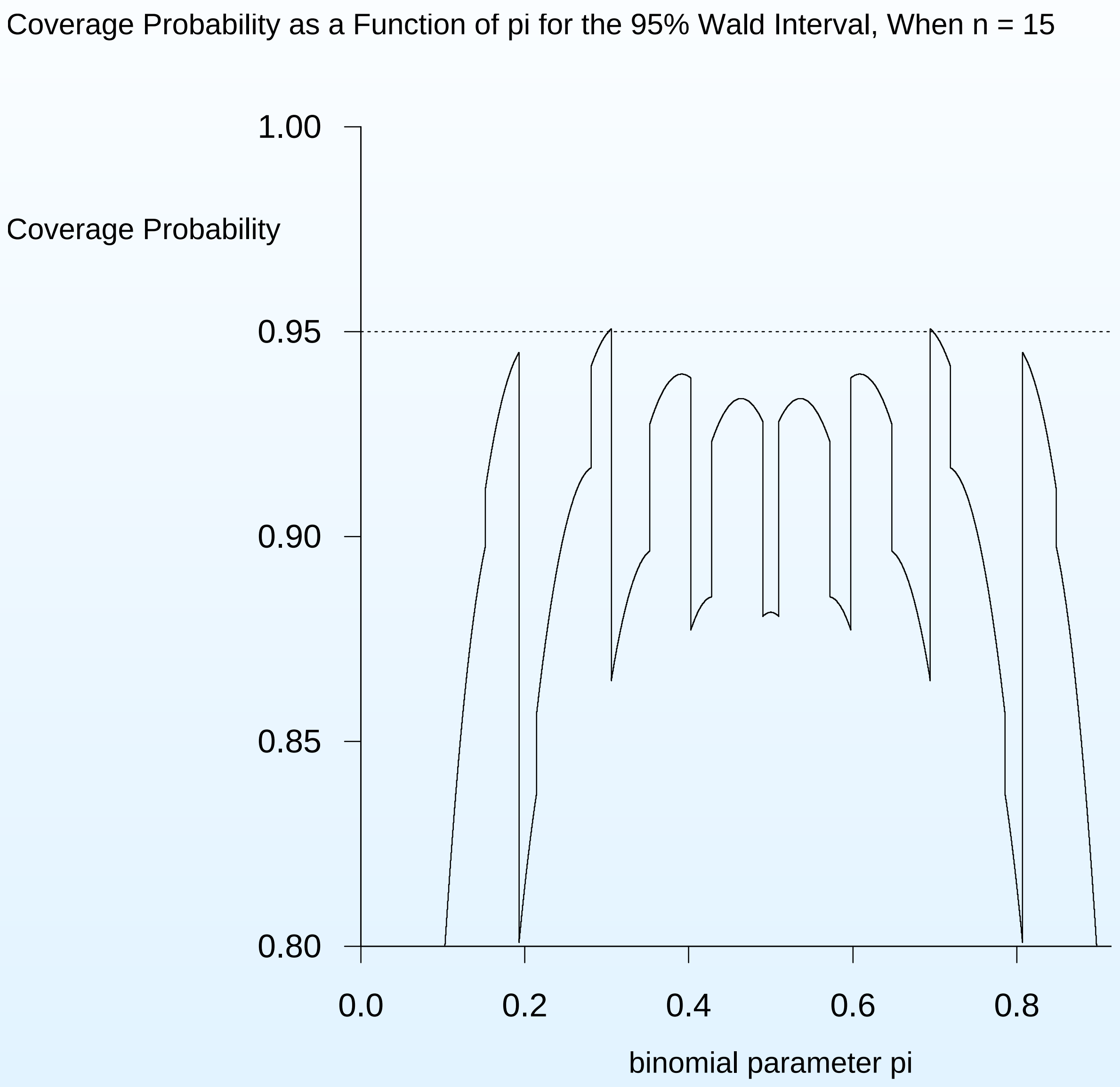
  Example: $Y \sim$ binomial(n, π) with $n = 25$

  - Model logit$(\pi) = \alpha$
  - Consider H_0: $\alpha = 0$ (i.e., $\pi = 0.50$)
  - Wald chi-squared statistic = $[\text{logit}(\hat{\pi})]^2 [n\hat{\pi}(1 - \hat{\pi})]$

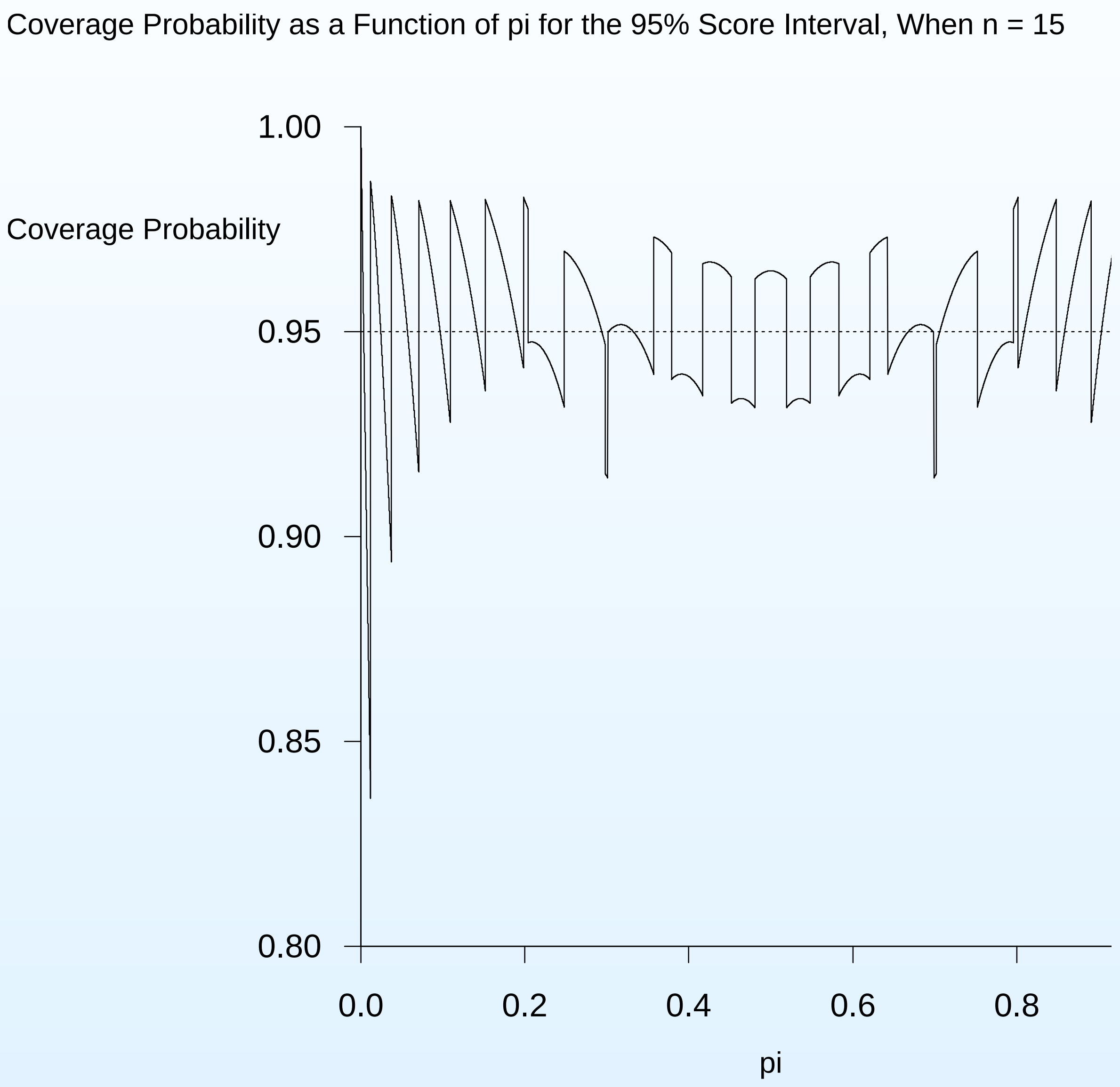
    = 11.0 when $y = 23$ ($\hat{\pi} = 0.92$)
    = 9.7 when $y = 24$ ($\hat{\pi} = 0.96$)
    (likelihood-ratio statistics are 20.7 and 26.3)

- Wald CI for π has coverage probability that is especially poor when π near 0 or 1.

Wald CI for binomial parameter, (n = 15)



Score CI for binomial parameter, (n = 15)



Examples of score inference: Not so "well known"

- Difference of proportions for independent binomial samples

Consider H_0: $\pi_1 - \pi_2 = \beta_0$. Score statistic is square of

$$z = \frac{(\hat{\pi}_1 - \hat{\pi}_2) - \beta_0}{\sqrt{[\hat{\pi}_1(\beta_0)(1 - \hat{\pi}_1(\beta_0))/n_1] + [\hat{\pi}_2(\beta_0)(1 - \hat{\pi}_2(\beta_0))/n_2]}},$$

where $\hat{\pi}_1$ and $\hat{\pi}_2$ are sample proportions and $\hat{\pi}_1(\beta_0)$ and $\hat{\pi}_2(\beta_0)$ are ML estimates under constraint $\pi_1 - \pi_2 = \beta_0$.

(When $\beta_0 = 0$, z^2 = Pearson chi-squared for 2×2 table.)

Score CI for $\pi_1 - \pi_2$ inverts this test.
(Mee 1984, Miettinen and Nurminen 1985)

Aside: 2×2 tables with no 'successes' (meta-analyses)

- For significance tests (e.g., Cochran–Mantel–Haenszel and small-sample exact), no information about whether there is an association; data make no contribution to the tests.
- For estimation, no information about odds ratio or relative risk but there is about $\pi_1 - \pi_2$
(i.e., impact on practical, not statistical, significance).

	Response		Response	
Group	Success	Failure	Success	Failure
1	0	10	0	100
2	0	20	0	200

Score 95% CIs for $\pi_1 - \pi_2$: $(-0.16, 0.28)$, $(-0.02, 0.04)$
Wald 95% CIs for $\pi_1 - \pi_2$: $(0.00, 0.00)$, $(0.00, 0.00)$

Note: Not necessary to add constants to empty cells.

Examples of score inference: Not “well known” (2)

- Score CI for odds ratio for 2×2 table $\{n_{ij}\}$ (Cornfield 1956): For given β_0, let $\{\hat{\mu}_{ij}(\beta_0)\}$ have same row and column margins as $\{n_{ij}\}$ and

$$\frac{\hat{\mu}_{11}(\beta_0)\hat{\mu}_{22}(\beta_0)}{\hat{\mu}_{12}(\beta_0)\hat{\mu}_{21}(\beta_0)} = \beta_0.$$

  95% CI = set of β_0 satisfying

$$X^2(\beta_0) = \sum (n_{ij} - \hat{\mu}_{ij}(\beta_0))^2 / \hat{\mu}_{ij}(\beta_0) \leq 1.96^2$$

- Likewise, score CI applies to relative risk, logistic regression parameters, generic measure of association (Lang 2008), but is not found in standard software.

(Some R functions: www.stat.ufl.edu/∼aa/cda/software.html)

Relation of score statistic to Pearson chi-squared

For counts $\{n_i\}$ for a multinomial model and testing goodness of fit using ML fit $\{\hat{\mu}_i\}$ under H_0,
the score test statistic is the Pearson chi-squared statistic

$$X^2 = \sum \frac{(n_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

True for generalized linear models (Smyth 2003, Lovison 2005)

When model refers to parameter β, inverting Pearson chi-squared test of H_0: $\beta = \beta_0$ gives score CI
e.g., 95% CI for β is set of β_0 for which $X^2 \leq \chi^2_{1,0.05} = (1.96)^2$.

For small to moderate n, actual coverage probability closer to nominal level for score CI than Wald, usually LR.

Evidence that score-test-based inference is “good”

- Testing independence in two-way tables (Koehler and Larntz 1980), LR stat. $G^2 = 2 \sum n_{ij} \log(n_{ij}/\hat{\mu}_{ij})$
- CI for binomial parameter (Newcombe 1998)
- CI for difference of proportions, relative risk, odds ratio, comparing dependent samples (Newcombe 1998, Tango 1998, Agresti and Min 2005)
- Multivariate comparisons of proportions for independent and dependent samples (Agresti and Klingenberg 2005, 2006)
- Simultaneous CIs comparing binomial proportions (Agresti, Bini, Bertaccini, Ryu 2008)
- CI for ordinal effect measures such as $P(y_1 > y_2)$ (Ryu and Agresti 2008)

Score-test-based inference infeasible for some models

Example: 2006 General Social Survey Data responses to “How successful is the government in (1) Providing health care for the sick? (2) Protecting the environment?”

(1 = successful, 2 = mixed, 3= unsuccessful)

	y_2 = Environment			
y_1 = Health Care	1	2	3	Total
1	199	81	83	363
2	129	167	112	408
3	164	169	363	696
Total	492	417	558	1467

A marginal model for multivariate data

Cumulative logit marginal model for responses (y_1, y_2)

$$\text{logit}[P(y_1 \leq j)] = \alpha_j, \quad \text{logit}[P(y_2 \leq j)] = \alpha_j + \beta, \quad j = 1, 2.$$

designed to detect location shift in marginal distributions.

Multinomial likelihood in terms of cell probabilities $\{\pi_{ij} = \mu_{ij}/n\}$ and cell counts $\{n_{ij}\}$

$$L(\boldsymbol{\pi}) \propto \pi_{11}^{n_{11}} \pi_{12}^{n_{12}} \cdots \pi_{33}^{n_{33}}$$

but model parameters refer to marginal probabilities.

LR and score inference comparing two models

Contingency table $\{n_i\}$ with ML fitted values $\{\hat{\mu}_i\}$ for a model and $\{\hat{\mu}_{i0}\}$ for simpler "null" model (e.g., with $\beta = \beta_0$):

H_0: Simpler model, H_a: More complex model

LR statistic (for multinomial sampling) is

$$G^2 = 2 \sum_i \hat{\mu}_i \log(\hat{\mu}_i / \hat{\mu}_{i0}).$$

Rao (1961) suggested Pearson-type statistic,

$$X^2 = \sum_i \frac{(\hat{\mu}_i - \hat{\mu}_{i0})^2}{\hat{\mu}_{i0}}.$$

Pseudo-score CI based on Pearson stat. (with E. Ryu)

For hypothesis testing, G^2 nearly universal, X^2 mostly ignored since Haberman (1977) results on sparse asymptotics.

Confidence intervals:

Popular to obtain *profile likelihood confidence intervals*: If $G^2 = G^2(\beta_0)$ is LR stat. for H_0: $\beta = \beta_0$, then 95% LR CI is

$$\{\beta_0\} \text{ such that } G^2(\beta_0) \leq \chi^2_{1,0.05}$$

e.g., in SAS, available with LRCI option in PROC GENMOD; in R, with confint() function applied to model object.

Since score CI often out-performs LR CI for simple discrete measures, as alternative to LR CI, could find *pseudo-score CI*:

$$\{\beta_0\} \text{ such that } X^2(\beta_0) \leq \chi^2_{1,0.05}$$ *Biometrika* 2010

Example: Cumulative logit marginal model

Cumulative logit marginal model

$$\text{logit}[P(y_1 \leq j)] = \alpha_j, \quad \text{logit}[P(y_2 \leq j)] = \alpha_j + \beta, \quad j = 1, 2.$$

Joe Lang (Univ. Iowa) has R function “mph.fit” for ML fitting of general class of models (*JASA* 2005, *Ann. Statist.* 2004).

For various fixed β_0, need to fit model with that constraint (using offset), giving $\{\hat{\mu}_{ij,0}\}$ to compare to $\{\hat{\mu}_{ij}\}$ in 3×3 table to find β_0 with $X^2(\beta_0) \leq \chi^2_{1,0.05}$.

95% pseudo-score CI is (0.2898, 0.5162).

Here, n large and results similar to LR CI of (0.2900, 0.5162).

Simulation studies show pseudo-score often performs better for small n.

Pseudo-score inference for discrete data

When independent $\{y_i\}$ for a GLM, a Pearson-type pseudo-score statistic is

$$X^2 = \sum_i \frac{(\hat{\mu}_i - \hat{\mu}_{i0})^2}{v(\hat{\mu}_{i0})} = (\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_0)' \hat{\mathbf{V}}_0^{-1} (\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_0),$$

$v(\hat{\mu}_{i0})$ = estimated null variance of y_i

$\hat{\mathbf{V}}_0$ = diagonal matrix containing $v(\hat{\mu}_{i0})$
(Lovison 2005)

Binary response: $v(\hat{\mu}_{i0}) = \hat{\mu}_{i0}(1 - \hat{\mu}_{i0})$

Pseudo-score CI potentially useful for discrete dist's other than binomial, multinomial, and for complex sample survey data (e.g., variances inflated from simple random sampling) and clustered data.

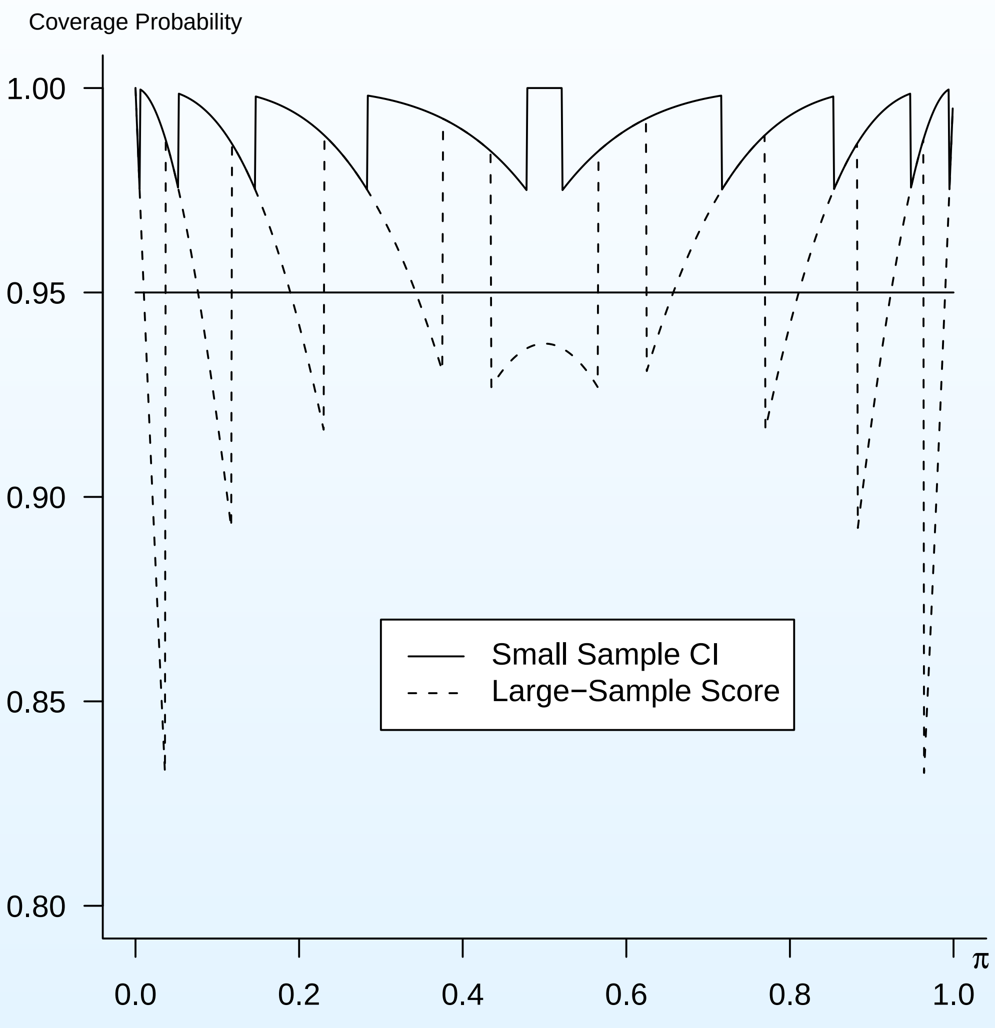
Small-sample methods (not “exact”)

Using score (or other) stat., can use small-sample distributions (e.g., binomial), rather than large-sample approximations (e.g., normal), to obtain P-values and confidence intervals.

- Because of *discreteness*, error probabilities do *not* exactly equal nominal values.
- For CI, inverting test with actual size $\leq .05$ for all β_0 guarantees *actual* coverage probability $\geq$ 0.95.
- Inferences are *conservative* – actual error probabilities $\leq$ 0.05 nominal level.
- Actual coverage prob varies for different β values and is unknown in practice.

Example: Binomial (n, π) with $n = 5$

Large-sample score CI vs. small-sample CI ($n = 5$)



Examples of small-sample CIs (95%)

Use *tail method*: Invert two separate one-sided tests each of size ≤ 0.025. (P-value = double the minimum tail probability)

1. Binomial parameter π

Clopper and Pearson (1934) suggest solution (π_L, π_U) to

$$\sum_{k=t_{obs}}^{n} \binom{n}{k} \pi_L^k (1-\pi_L)^{n-k} = 0.025$$

and

$$\sum_{k=0}^{t_{obs}} \binom{n}{k} \pi_U^k (1-\pi_U)^{n-k} = 0.025$$

Examples of small-sample CIs (2)

2. Logistic regression parameter

For subject i with binary outcome y_i, explanatory variables ($x_{i0} = 1, x_{i1}, x_{i2}, \ldots, x_{ik}$)

Model: logit$[P(y_i = 1)] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$

Use dist. of score stat. after eliminating nuisance para. by conditioning on sufficient stat's ($T_j = \sum_i y_i x_{ij}$ for β_j).

e.g., Bounds (β_{1L}, β_{1U}) of 95% CI for β_1 satisfy

$$P(T_1 \geq t_{1,obs} | t_0, t_2, \ldots, t_k; \beta_{1L}) = 0.025$$
$$P(T_1 \leq t_{1,obs} | t_0, t_2, \ldots, t_k; \beta_{1U}) = 0.025$$

Randomizing Eliminates Conservatism

- For testing $H_0 : \beta = \beta_0$ against $H_a : \beta > \beta_0$ using a test stat. T, a randomized test has P-value

$$P_{\beta_0}(T > t_{obs}) + \mathcal{U} \times P_{\beta_0}(T = t_{obs})$$

where $\mathcal{U}$ is a uniform(0,1) random variable.

- To construct CI with actual coverage probability 0.95,

$$P_{\beta_U}(T < t_{obs}) + \mathcal{U} \times P_{\beta_U}(T = t_{obs}) = 0.025$$

and

$$P_{\beta_L}(T > t_{obs}) + (1 - \mathcal{U}) \times P_{\beta_L}(T = t_{obs}) = 0.025.$$

Use randomized methods in practice?

- Randomized CI suggested by Stevens (1950), for binomial parameter.
- Pearson (1950): Statisticians may come to accept randomization *after* performing experiment just as they accept randomization *before* the experiment.
- Stevens (1950): “We suppose that most people will find repugnant the idea of adding yet another random element to a result which is already subject to the errors of random sampling. But what one is really doing is to eliminate one uncertainty by introducing a new one. ... It is because this uncertainty is eliminated that we no longer have to keep ‘on the safe side’, and can therefore reduce the width of the interval.”

Mid-P Pseudo-Score Approach

- Mid-P-value (Lancaster 1949, 1961): Count only $(1/2)P_{\beta_0}(T = t_{obs})$ in P-value; e.g., for $H_a : \beta > \beta_0$,

$$P_{\beta_0}(T > t_{obs}) + (1/2)P_{\beta_0}(T = t_{obs}).$$

- Unlike randomized P-value, depends only on data.
- Under H_0, ordinary P-value stochastically larger than uniform, E(mid-P-value)= 1/2.
- Sum of right-tail and left-tail P-values is $1 + P_{\beta_0}(T = t_{obs})$ for ordinary P-value, 1 for mid-P-value.

CI based on mid-P-value

- Mid-P CI based on inverting tests using mid-P-value:

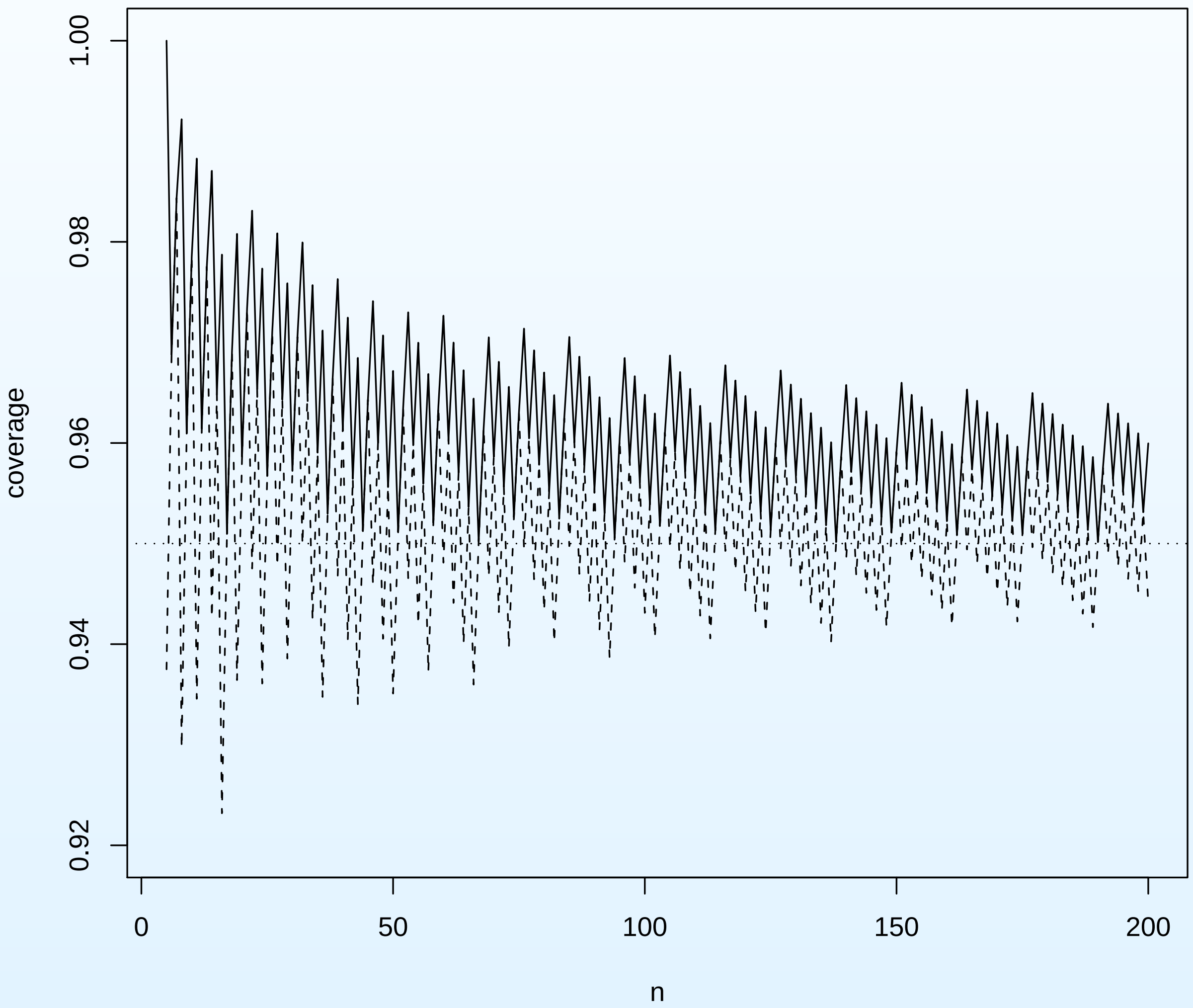
$$P_{\beta_L}(T > t_{obs}) + (1/2) \times P_{\beta_L}(T = t_{obs}) = 0.025.$$

$$P_{\beta_U}(T < t_{obs}) + (1/2) \times P_{\beta_U}(T = t_{obs}) = 0.025.$$

- Coverage probability not guaranteed $\geq$ 0.95, but mid-P CI tends to be a bit conservative.
- For binomial, how do Clopper–Pearson and mid-P CI behave as n increases?

  (from Agresti and Gottard 2007)

Clopper-Pearson (—) and mid-P (- -) CIs for $\pi = 0.50$



Simple approximations to score CIs often work well

Example: Binomial proportion

Finding all π_0 such that $\frac{|\hat{\pi}-\pi_0|}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} < 2$

provides 95% score CI of form $M \pm 2s$
(approximating 1.96 by 2) with

$$M = \left(\frac{n}{n+4}\right)\hat{\pi} + \left(\frac{4}{n+4}\right)\frac{1}{2} = \frac{t_{obs}+2}{n+4}$$

$$s^2 = \frac{1}{n+4}\left[\hat{\pi}(1-\hat{\pi})\left(\frac{n}{n+4}\right) + \frac{1}{2}\frac{1}{2}\left(\frac{4}{n+4}\right)\right]$$

Adjusted Wald CI approximates score CI

For 95% CI, this suggests an adjusted Wald CI (*plus 4 CI*)

$$\tilde{\pi} \pm 2.0\sqrt{\tilde{\pi}(1-\tilde{\pi})/\tilde{n}}$$

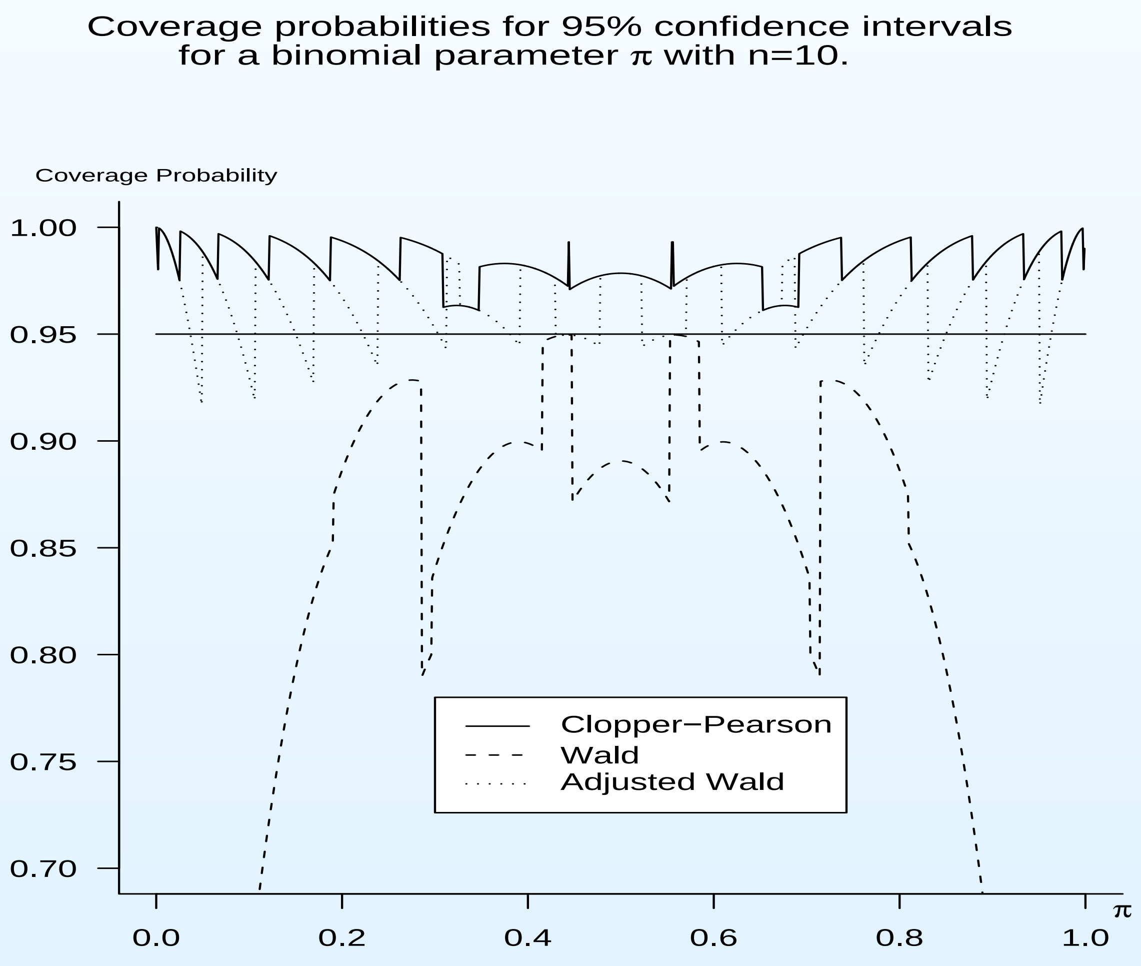
with $\tilde{\pi} = \frac{t_{obs}+2}{n+4}$ and $\tilde{n} = n + 4$.

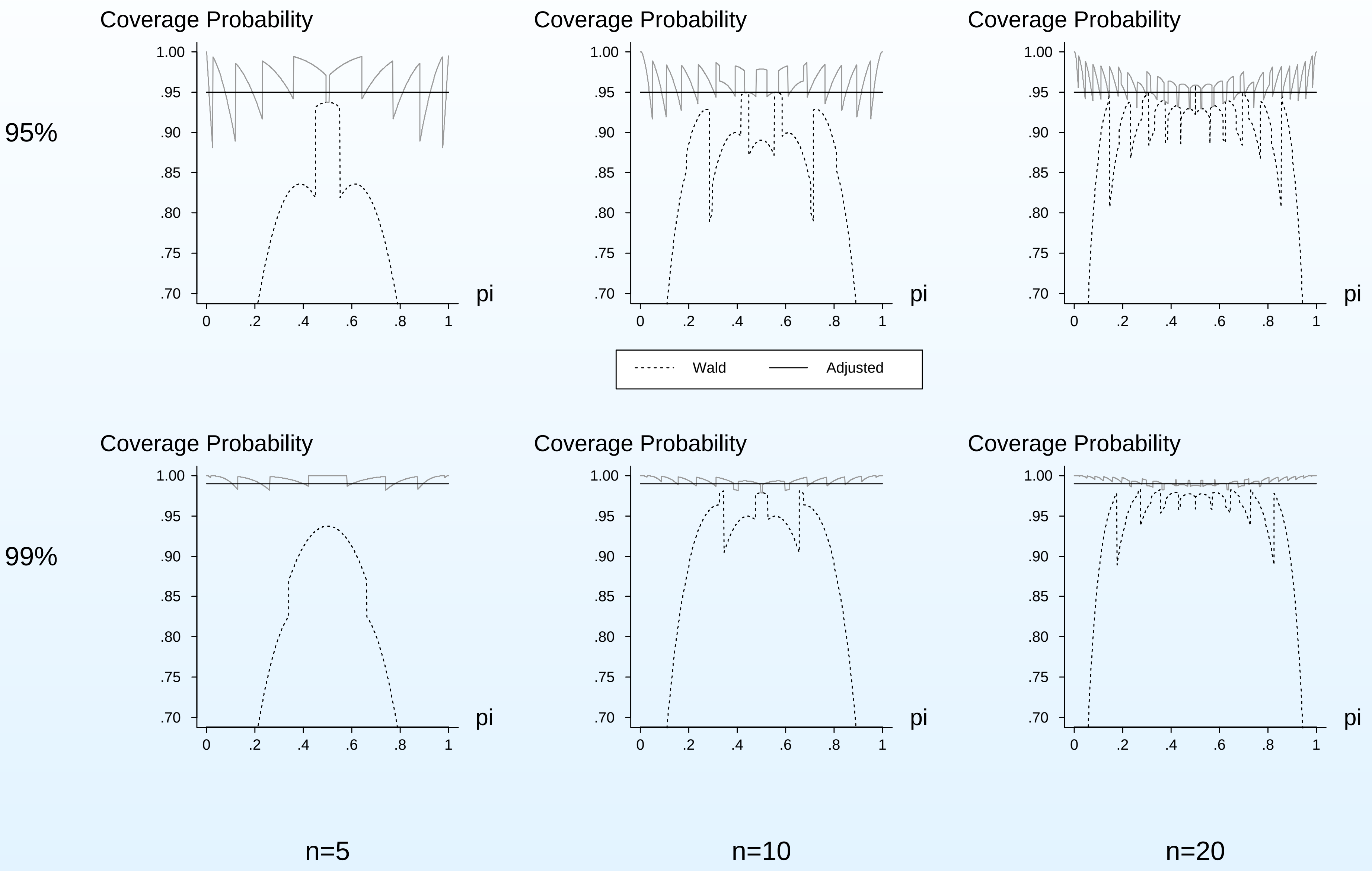
Midpoint same as 95% score CI, but wider (Jensen's inequality).

In fact, simple adjustments of Wald improve performance dramatically and give performance similar to score CI:

- *Proportion*: Add 2 successes and 2 failures before computing Wald CI (Agresti and Coull 1998)
- *Difference*: Add 2 successes and 2 failures before computing Wald CI (Agresti and Caffo 2000)
- *Paired Difference*: Add 2 successes and 2 failures before computing Wald CI (Agresti and Min 2005)

Clopper-Pearson, Wald, and “Plus 4” CI $(n = 10)$





95%
99%
Coverage Probability
1.00
.95
.90
.85
.80
.75
.70
pi
0
.2
.4
.6
.8
1
Wald
Adjusted
n=5
n=10
n=20

“Good” CIs shrink midpoints

- Poor performance of Wald intervals due to centering at $\hat{\pi}$, $(\hat{\pi}_1 - \hat{\pi}_2)$ rather than being too short.
- Wald CI has greater length than adjusted intervals unless parameters near boundary of parameter space.
- Intervals resulting from Bayesian approach can also perform well in frequentist sense.

  Single proportion: Brown et al. (2001)

  Comparing proportions: Agresti and Min (2005). Using independent Jeffreys beta(0.5, 0.5) priors gives frequentist performance similar to score CI.
- Many score CIs, mid-P corrections, Bayes CIs available in R at www.stat.ufl.edu/∼aa/cda/software.html.

Summary

- Full model saturated: Score confidence interval inverts goodness-of-fit test using Pearson chi-squared statistic.
- Full model unsaturated: Pseudo-score method of inverting Pearson test comparing fitted values available when ordinary score CI infeasible and may perform better than profile likelihood CI for small n.
- Good small-sample CI inverts score test with mid-P-value.
- For proportions and their differences, pseudo-score method that adjusts Wald method by adding 4 observations performs well.

Some references

Agresti, A., and Coull, B. (1998). Approximate better than ‘exact’ for CIs for binomial parameters, *American Statistician*.

Agresti, A., and Caffo, B. (2000). Effective CIs for proportions and difference of proportions result from adding two successes and two failures, *Amer. Stat.*

Agresti, A., and Min, Y. (2001). On small-sample confidence intervals for parameters in discrete distributions, *Biometrics*.

Agresti, A., and Gottard, A. (2007). Nonconservative exact small-sample inference for discrete data, *Computational Statistics & Data Analysis*.

Agresti, A., and Ryu, E. (2010). Pseudo-score inference for parameters in discrete statistical models. *Biometrika*.

Lovison G. (2005). On Rao score and Pearson X^2 statistics in generalized linear models. *Statistical Papers*.

Rao, C. R. (1961). A study of large sample test criteria through properties of efficient estimates. *Sankhya*.

Smyth, G. K. (2003). Pearson’s goodness of fit statistic as a score test statistic. In *Science and Statistics: A Festschrift for Terry Speed*, IMS Lecture Notes–Monograph Series.