Reducing Conservatism of Exact Small-Sample Inference for Discrete Data

Alan Agresti

Department of Statistics, University of Florida

What is “Exact” Discrete Inference?

- Use exact small-sample distributions (e.g., binomial), rather than large-sample approximations (e.g., normal), to obtain P-values and confidence intervals.
- For contingency tables, best known is “Fisher’s exact test” for 2×2 tables, which conditions on row and column margins and uses hypergeometric dist. to get P-value.
- Now a large literature on small-sample inference for contingency tables, including multi-way tables and models.
- Most literature for large tables uses conditional approach (Fisher) of eliminating nuisance parameters by conditioning on sufficient statistics.

Computations for "Exact" Inference

- Software now readily available, mainly for conditional approach, such as

  StatXact – contingency table methods

  LogXact – logistic regression

  $r \times c$ tables, stratified tables, dependent samples and clustered data, logistic and multinomial regression

- Almost all exact tests execute within a few seconds when $n < 30$, but computations grow exponentially in n.

  e.g., 5×6 table, margins (7, 7, 12, 4, 4), (4, 5, 6, 5, 7, 7): Up to 1.6 billion contingency tables have same margins and contribute to exact tests.

- For cases that are infeasible, fast and precise Monte Carlo approximations available.

“Exact” Inference is not Exact in terms of Error Rates

- For a parameter θ, H_0: $\theta = \theta_0$, let T = test statistic, t_{obs} = observed value, P-value = $P_{\theta_0}(T \geq t_{obs})$, nominal P(Type I error) = 0.05 (i.e., reject H_0 when P-value $\leq$ 0.05).
- Let 95% confidence interval (CI) be set of θ_0 for H_0: $\theta = \theta_0$ such that P-value $> .05$.

– Because of discreteness, error probabilities do *not* exactly equal nominal values.

ex.: If possible P-values for exact distribution are 0.031, 0.187, ..., (binomial $n = 5$, $\theta_0 = 0.50$) then *actual* size = 0.031.

For CI, inverting test with actual size $\leq .05$ for all θ_0 guarantees *actual* coverage probability $\geq$ 0.95.

– Inferences are *conservative* – actual error probabilities $\leq$ 0.05 nominal level.

Outline

- For small n, *large-sample* methods may work poorly yet *small-sample* ‘exact’ methods may be very conservative (and both true for larger n than you’d expect).
- Example: Small-sample CI for a binomial proportion
- Randomization and fuzzy inference for eliminating conservatism while maintaining exactness
- Quasi-exact inferences based on mid P-value
- Simple adjustments of popular large-sample CIs for proportions work well for small samples also
- Based partly on paper with Anna Gottard, Univ. of Firenze (to appear, *Comput. Statist. & Data Anal.*, 2007)

Example: T is binomial (n, π), $\quad \hat{\pi} = T/n$

Consider the popular 95% CI

$$\hat{\pi} \pm 2.0\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

Called *Wald CI*, since based on inverting Wald test;
i.e. values in CI are π_0 for H_0: $\pi = \pi_0$ satisfying

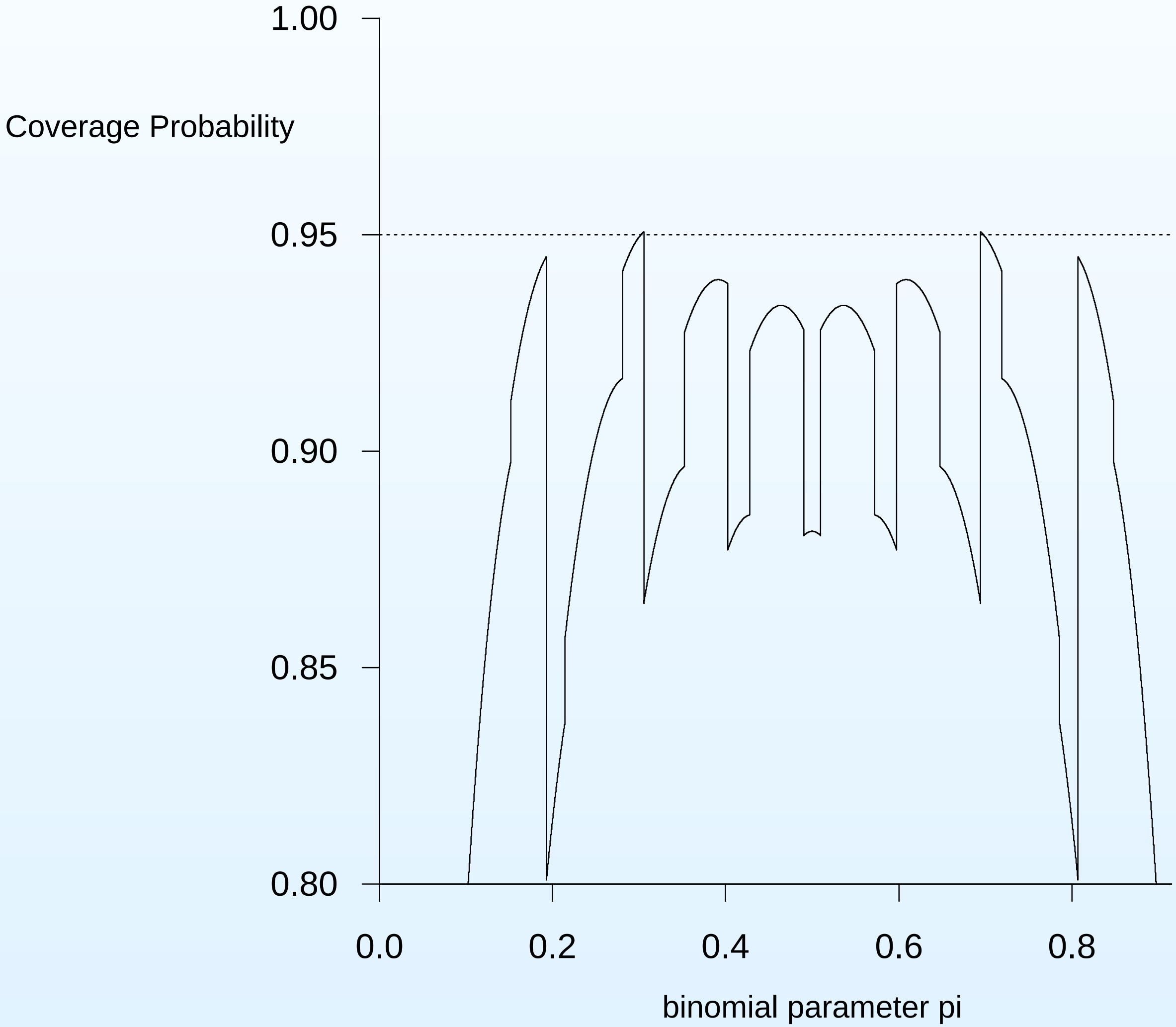
$$\frac{|\hat{\pi}-\pi_0|}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}} \leq 2.0$$

At a fixed π, actual coverage probability equals sum of

$$\binom{n}{t}\pi^t(1-\pi)^{n-t}$$

for all t such that CI contains π. (Figure: $n = 15$)

Coverage Probability as a Function of pi for the 95% Wald Interval, When n = 15



Small-sample CI

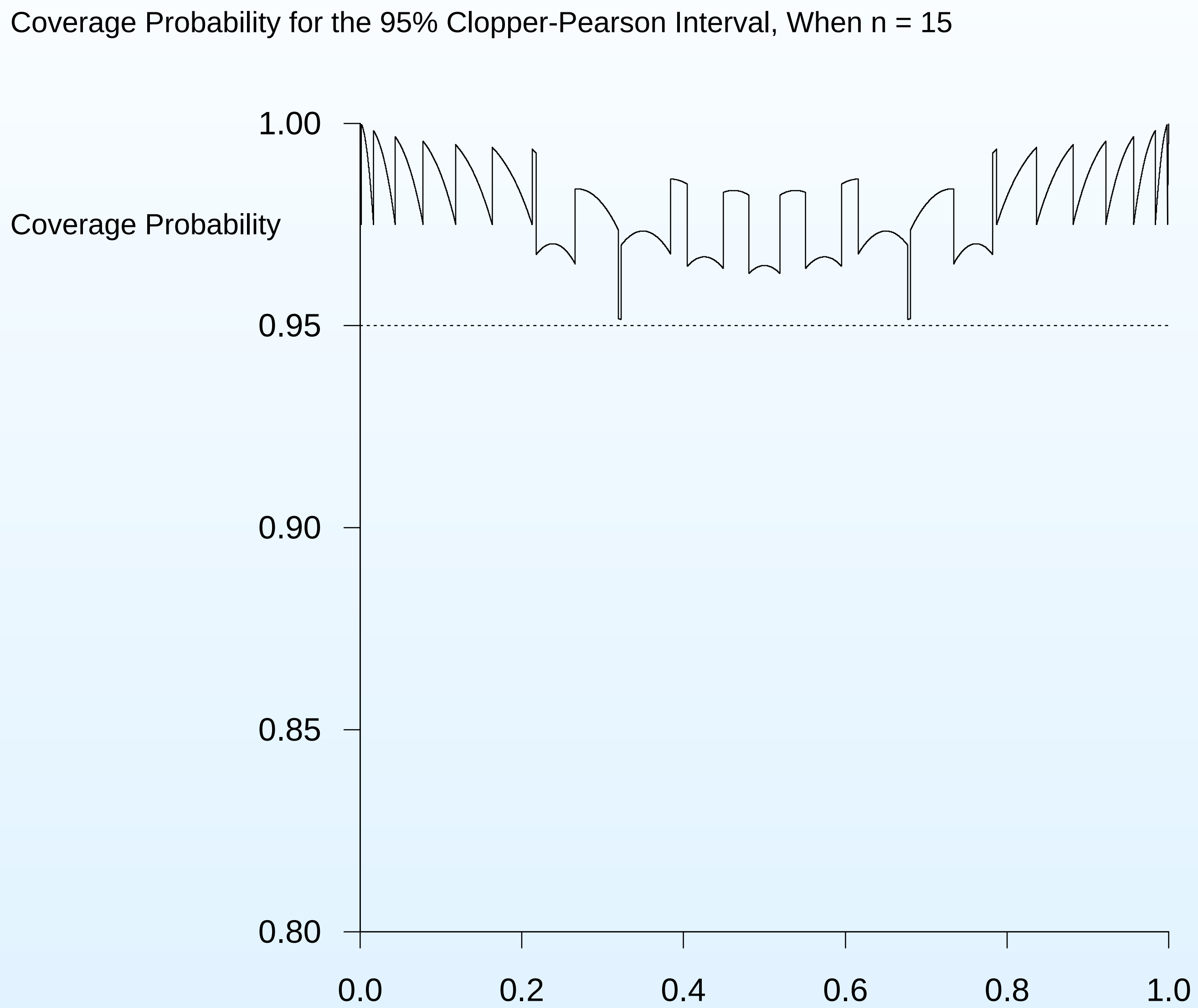
Best known small-sample ‘exact’ CI based on inverting binomial test (Clopper and Pearson, 1934)

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

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



Coverage Probability for the 95% Clopper-Pearson Interval, When n = 15
Coverage Probability
1.00
0.95
0.90
0.85
0.80
0.0
0.2
0.4
0.6
0.8
1.0

Discreteness and conservatism

- Discreteness implies finite set of possible P-values, not usually including 0.05, and *actual* coverage probability (i.e., sum of $\binom{n}{t}\pi^t(1-\pi)^{n-t}$ for all t such that CI contains π) cannot normally achieve *exactly* 0.95.
- Actual coverage prob. $\geq$ nominal coverage prob.
- If T has cdf $F(t;\theta)$, conservatism results from distribution of $F(T;\theta)$ (and P-value) stochastically larger than uniform (Casella and Berger 2001, pp. 77, 434)
- Actual coverage prob varies for different θ values and is unknown in practice.

Randomizing Eliminates Conservatism in Exact Tests

- In theory, (see, e.g., Lehmann) can set up critical function $\phi(t)$ for the probability of rejecting the null hypothesis
  - $\phi(t) = 1$ for t inside rejection region,
  - $\phi(t) = 0$ for t outside rejection region,
  - $\phi(t)$ on boundary of rejection region, such that size equals desired value.

ex. Suppose T is bin$(5, \pi)$, H_0: $\pi = 0.50$, H_a: $\pi > 0.50$,

Under H_0, $P(T = 5) = 0.031$, $P(T = 4) = 0.156$.

So, if $\phi(5) = 1$, $\phi(4) = 0.12$, then

P(reject H_0 | H_0 true) = 0.031 + 0.12(0.156) = 0.05.

Randomized P-value and CI

- For testing $H_0 : \theta = \theta_0$ against $H_a : \theta > \theta_0$ using T, a randomized test corresponds to using P-value

$$P_{\theta_0}(T > t_{obs}) + \mathcal{U} \times P_{\theta_0}(T = t_{obs})$$

  where $\mathcal{U}$ is a uniform(0,1) random variable.
- To construct CI with coverage probability 0.95,

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

  and

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

Use randomized methods in practice?

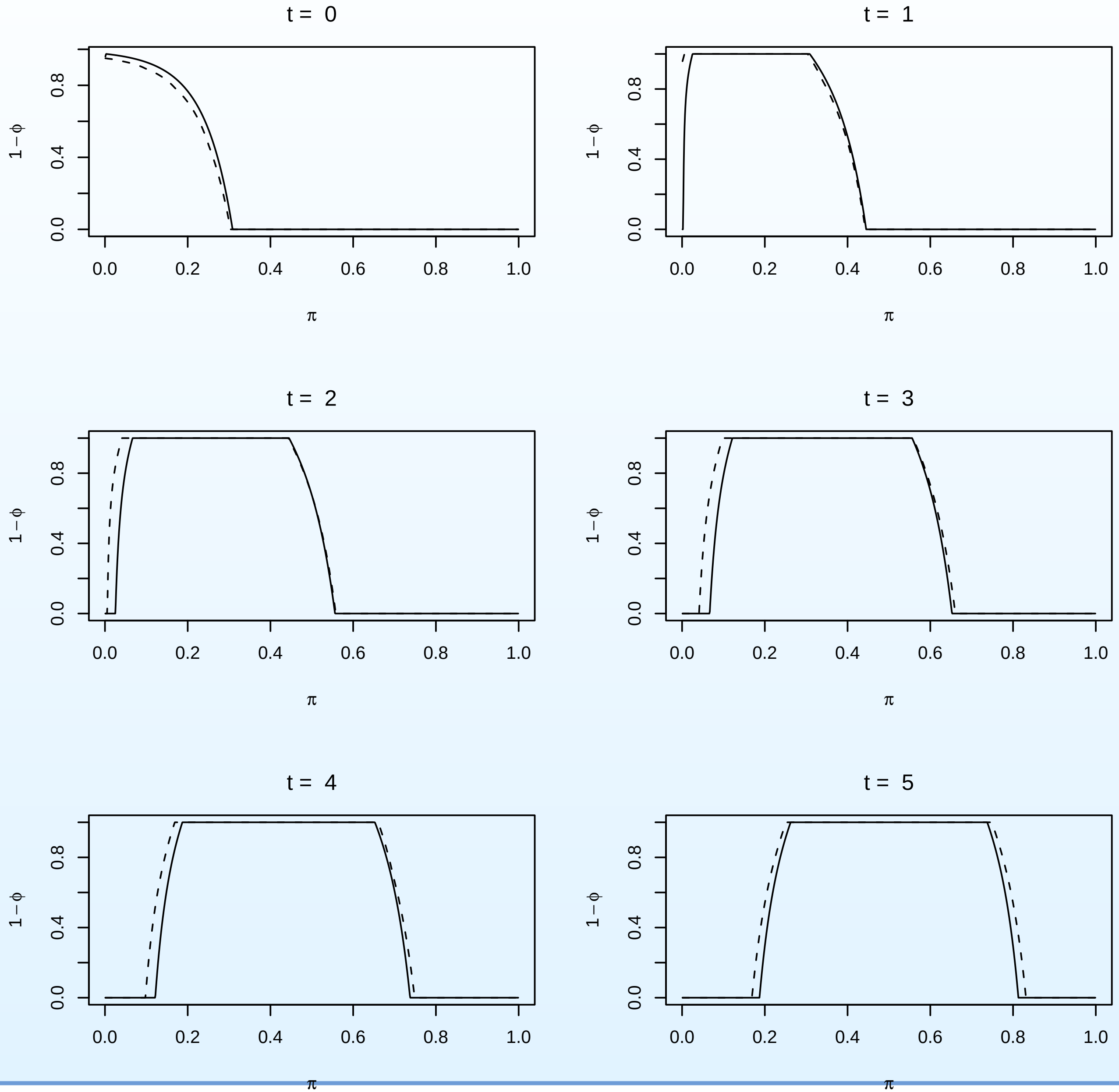
- Randomized CI suggested by Stevens (1950), for binomial parameter
- Pearson (1950): Statisticians may come to accept randomization *after* performing experiment just as they accept randomization *before* the experiment.
- Stevens (1950): “We suppose that most people will find repugnant the idea of 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.”

Fuzzy Inference

To avoid arbitrariness of picking random number, Geyer and Meeden (2005) suggested *fuzzy inference*.

- For $H_0 : \theta = \theta_0$, construct critical function $\phi(t, \theta_0)$ having desired size $\alpha = 0.05$.
- For fixed t, $[1 - \phi(t, \theta)]$ is *fuzzy confidence interval* over space of θ, and for given θ, $[1 - \phi(T, \theta)]$ has unconditional coverage probability 0.95.
- Geyer and Meeden provided UMPU fuzzy inference, but computationally complex.
- Given t, plot fuzzy CI to portray inference while guaranteeing desired coverage probability. (Example for binomial with $n = 10$, t_{obs} = 0, 1, 2, 3, 4, 5)

Fuzzy 95% CI: Geyer-Meeden (- -) Agresti-Gottard (—)



Alternative but simpler fuzzy CI

- *Core* = set of θ for which $[1 - \phi(t, \theta) = 1]$.
- *Support* = set of θ for which $[1 - \phi(t, \theta) > 0]$.
- Agresti and Gottard (2005): Directly generalize Stevens (1950) randomized CI to fuzzy CI for exponential family
  - As U increases from 0 to 1, lower and upper endpoints are monotonically increasing.
  - $\mathcal{U}$ = 0: Lower bound = lower bound from conservative CI.
  - $\mathcal{U}$ = 1: Upper bound = upper bound from conservative CI.
  - Support: Ordinary conservative confidence interval (e.g., Clopper–Pearson CI for binomial).
  - Core: θ values that fall in every possible randomized CI – goes from θ_L when $\mathcal{U} = 1$ to θ_U when $\mathcal{U} = 0$.

Mid-P Quasi-Exact Approach

- *Mid-P-value* (Lancaster 1949, 1961): Count only $(1/2)P_{\theta_0}(T = t_{obs})$ in P-value; e.g., for $H_a : \theta > \theta_0$,

$$P_{\theta_0}(T > t_{obs}) + (1/2)P_{\theta_0}(T = t_{obs}).$$

- 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 $1 + P_{\theta_0}(T = t_{obs})$ for ordinary P-value, 1 for mid-P-value.
- Lancaster: Like uniform P-value for continuous r.v., can easily combine for several independent samples.
- Mid-P-value not probability of particular sample set, does not satisfy P_{θ_0}(P-value $\leq 0.05) \leq 0.05$.

CI based on mid-P-value

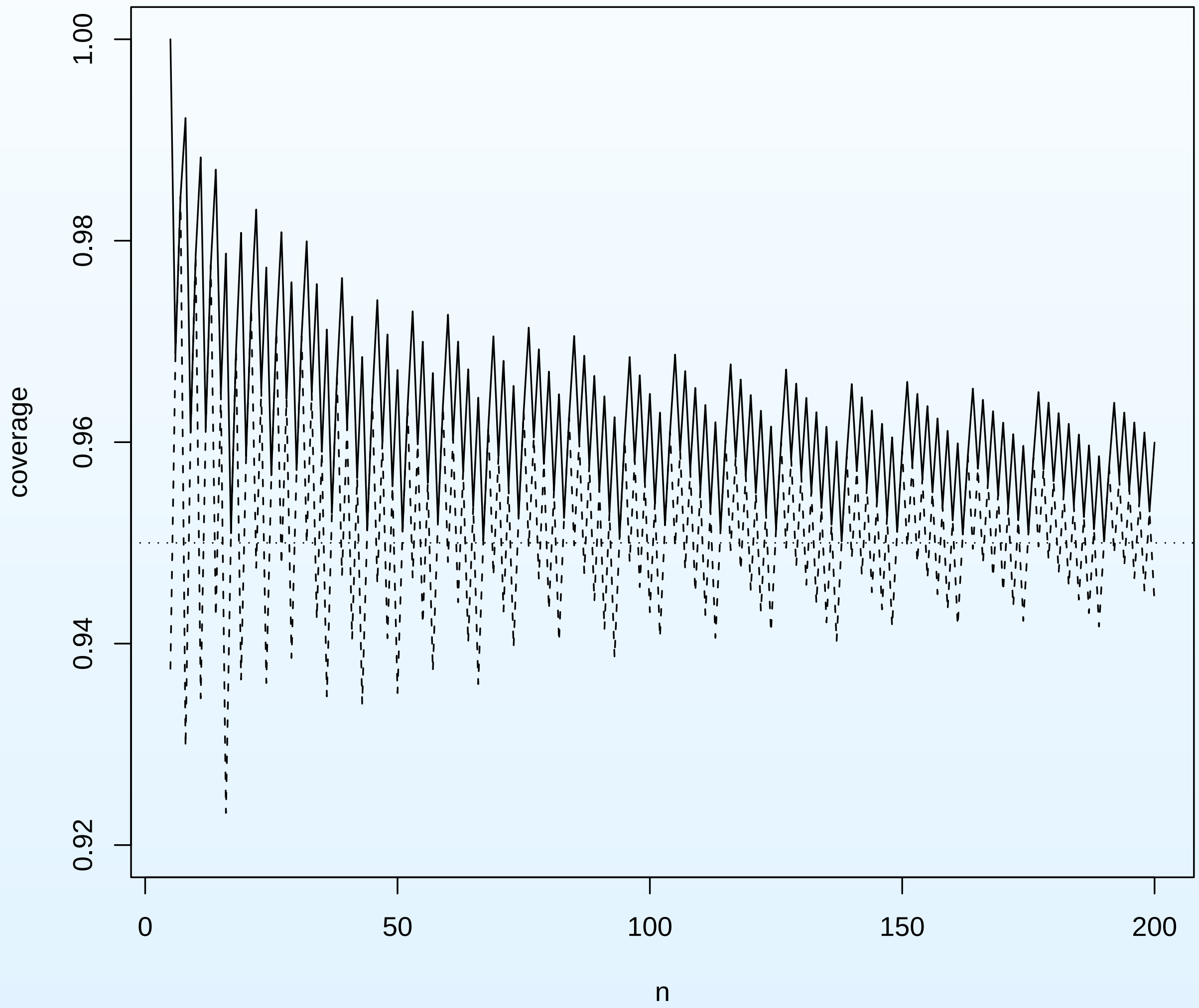
- *Mid-P CI* based on inverting tests using mid-P-value:

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

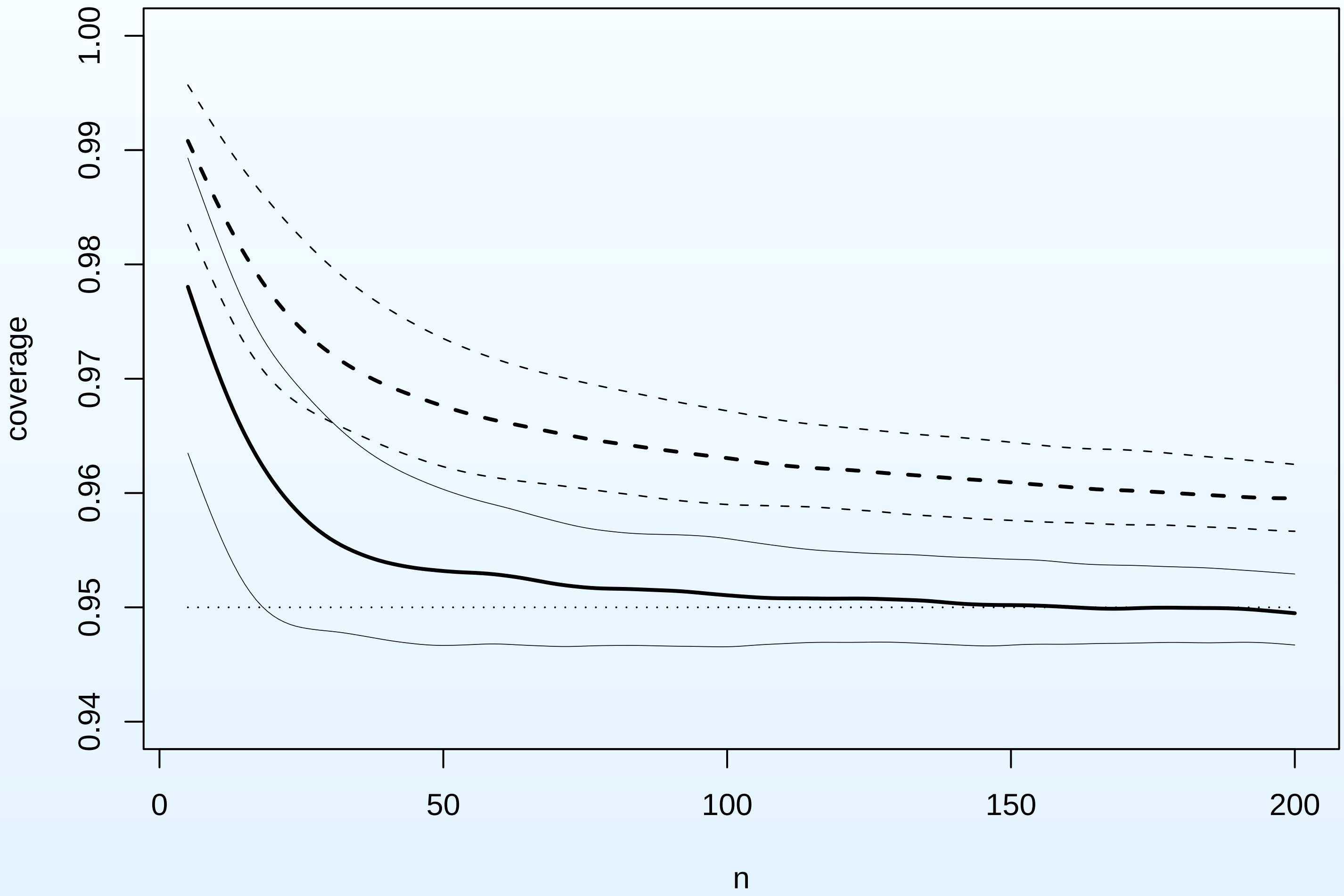
$$P_{\theta_U}(T < t_{obs}) + (1/2) \times P_{\theta_U}(T = t_{obs}) = 0.025.$$

- Coverage prob. not guaranteed $\geq$ 0.95, but mid-P CI tends to be a bit conservative.
- R function (A. Gottard) for mid-P binomial CI at www.stat.ufl.edu/$\sim$aa/cda/software.html
- For binomial, how do Clopper–Pearson and mid-P CI behave as n increases?

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



Quartiles of coverage probabilities, when π uniform, for C-P (- -) and mid-P (—) CIs



u-P-value and related CI

- *u-P CI* based on inverting tests using u-P-value:

$$P_{\theta_L}(T > t_{obs}) + u \times P_{\theta_L}(T = t_{obs}) = 0.025.$$

$$P_{\theta_U}(T < t_{obs}) + u \times P_{\theta_U}(T = t_{obs}) = 0.025.$$

- Now, u fixed rather than random.
- For given discrete problem, could choose u so that mean coverage (for some distribution over parameter) = 0.95.
- For binomial, coverage pictures (as function of π) look like mid-P CI, but with occasional poor coverages.

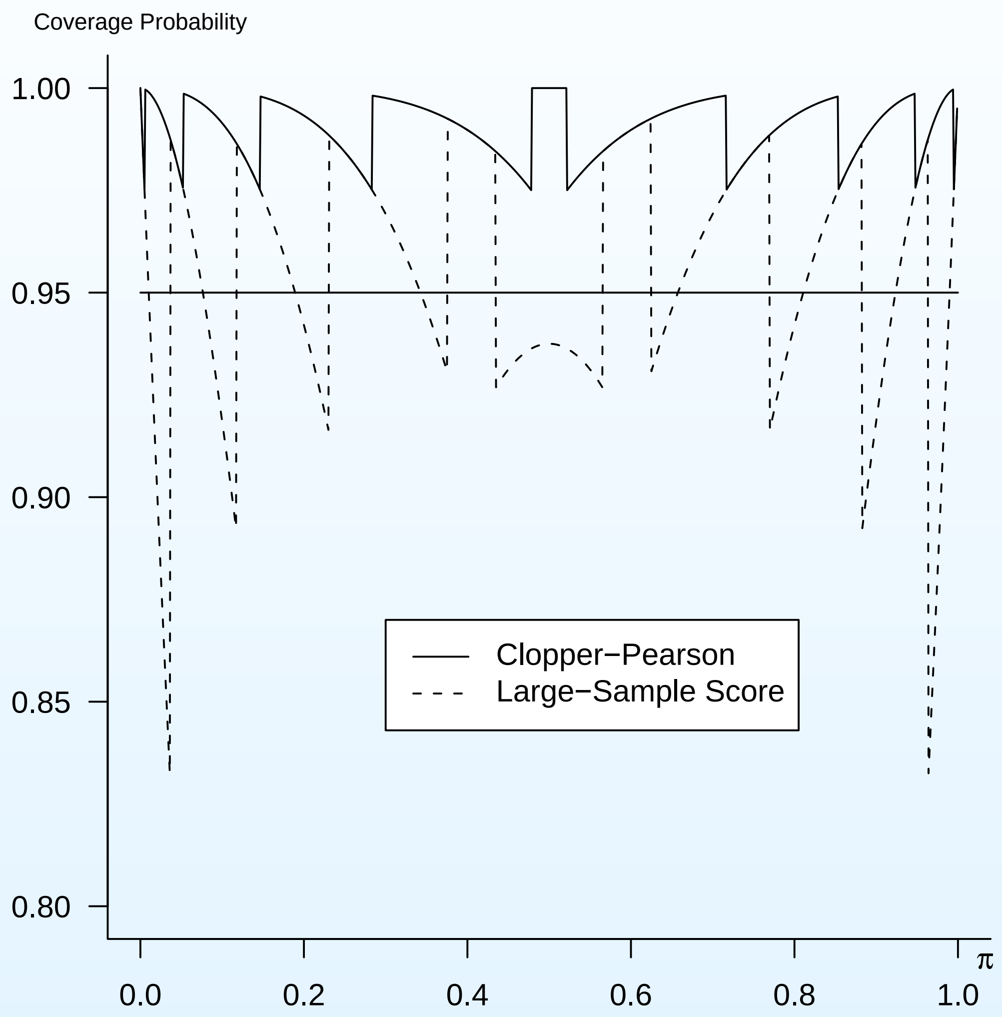
Improving Large-Sample CIs for Use with Small n

- Usual large-sample tests are Wald, likelihood-ratio, score
- Simplest approach is Wald CI, $\hat{\theta} \pm$ 2.0(std. error)
  - Proportion: $\hat{\pi} \pm 2.0\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$
- Wald methods for proportions, differences of proportions, etc., usually poor, especially near boundary of parameter space
- Closer to nominal levels by inverting score test

  Proportion (Wilson 1927):

$$\frac{|\hat{\pi} - \pi_0|}{\sqrt{\pi_0(1 - \pi_0)/n}} \leq 2.0$$

Score CI vs. Clopper-Pearson CI ($n = 5$)



95% CIs for a binomial proportion

Wald CI: $\hat{\pi} \pm 2.0\sqrt{\hat{\pi}(1-\hat{\pi})/n}$

Score CI: Inverting $|\hat{\pi} - \pi_0|/\sqrt{\pi_0(1-\pi_0)/n} = 2.0$,

$$\frac{\hat{\pi} + \frac{2}{n} \pm 2\sqrt{[\hat{\pi}(1-\hat{\pi}) + 1/n]/n}}{1 + 4/n}$$

Wald method simplest to explain, but poor performance

Score CI better, but messy to explain when teaching basic statistics in classroom or consulting environment

Simpler way to view the score CI

Score CI has form $M \pm 2s$ 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 *adjusted CI*

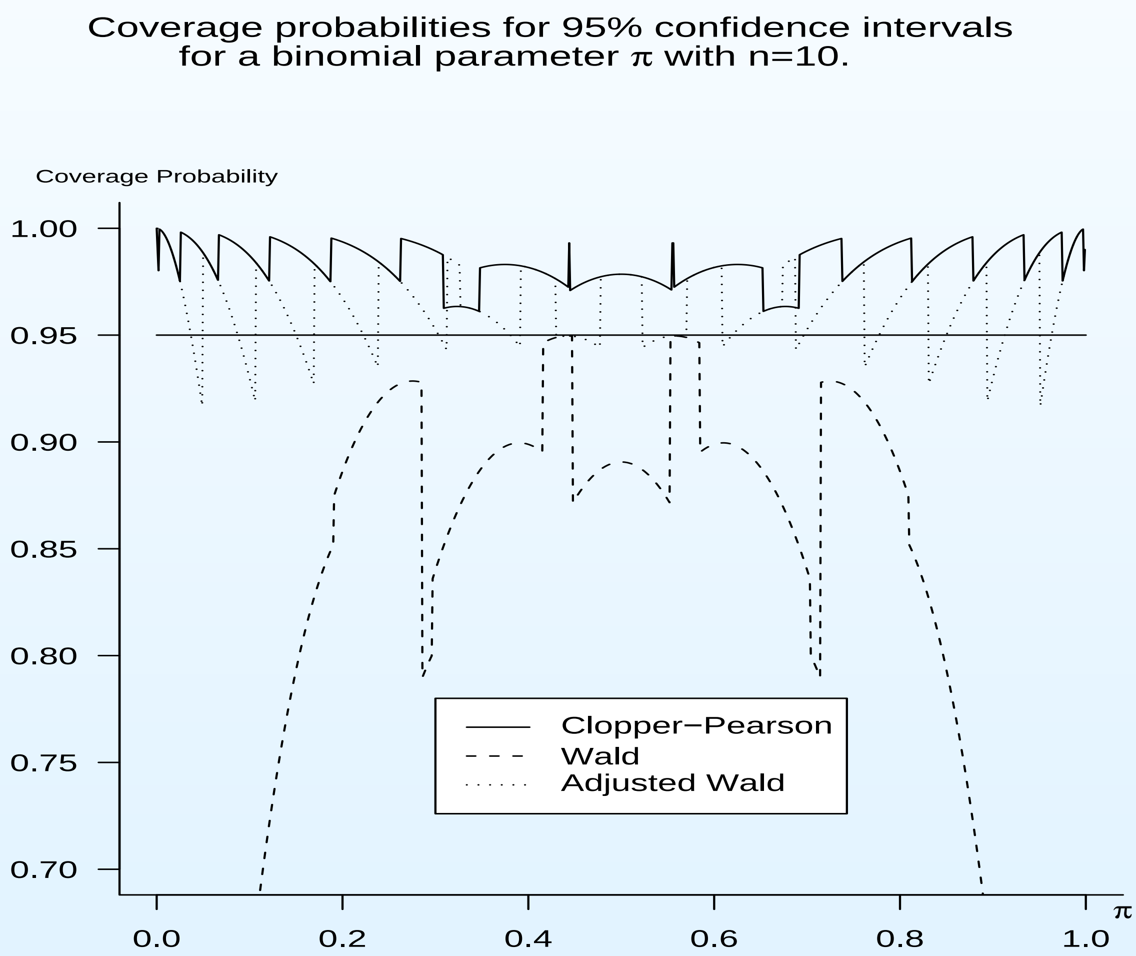
$$\tilde{\pi} \pm 2.0\sqrt{\tilde{\pi}(1-\tilde{\pi})/\tilde{n}}$$

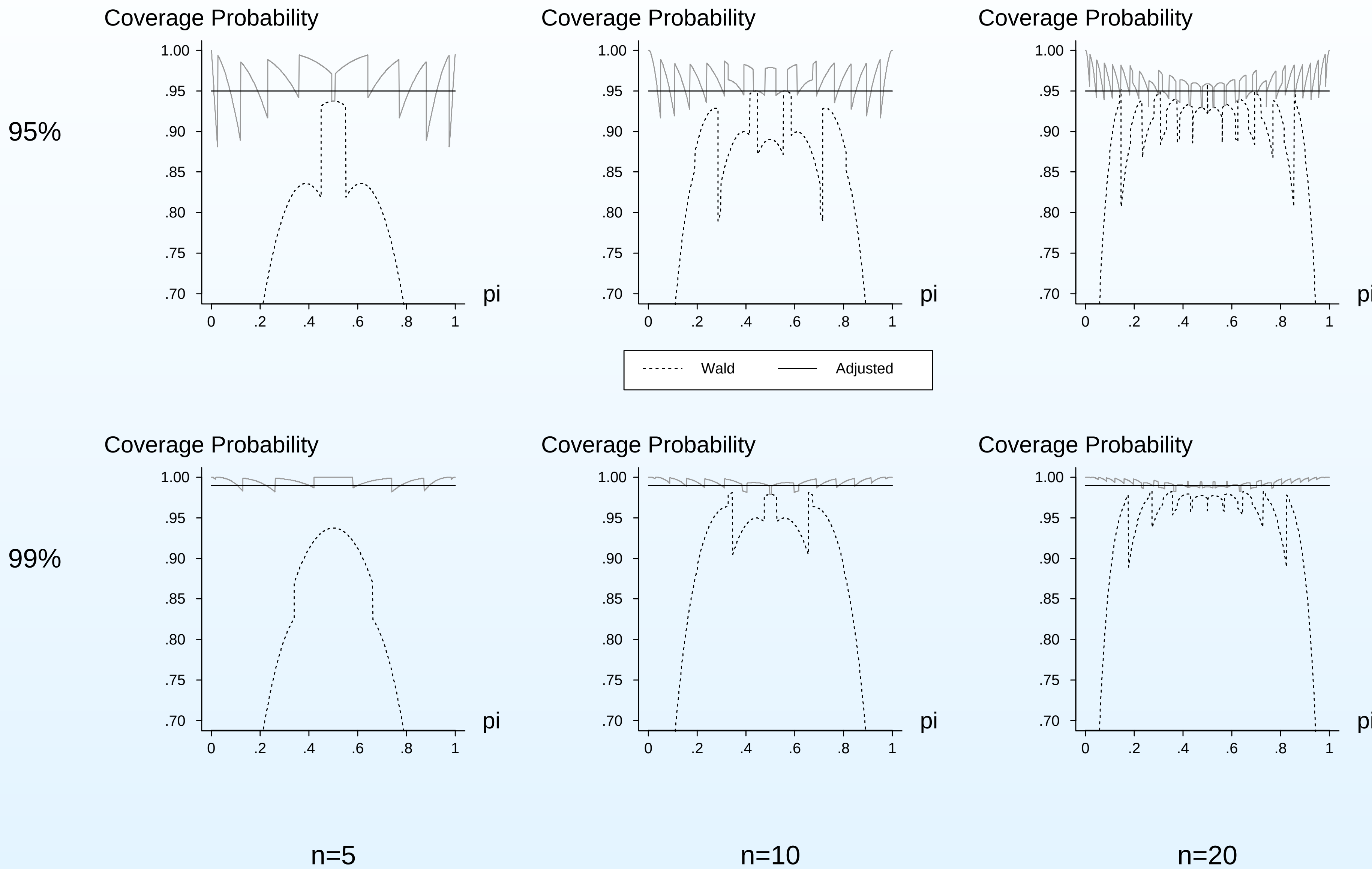
with $\tilde{\pi} = \frac{t_{obs}+2}{n+4}$ and $\tilde{n} = n + 4$

Midpoint same as 95% score CI, but wider (Jensen's inequality)
In fact, simple adjustments of Wald improve performance dramatically:

- *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 “Add 2+2” 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

Some comparisons (95% CI)

- For all n tremendous improvement for π near 0 or 1.

  e.g., Brown, Cai, and Das Gupta (2001):
  n_0 required such that cov. prob. ≥ 0.94 for all $n \geq n_0$ is

π	Wald	Adjusted
0.01	??	

Some comparisons (95% CI)

- For all n tremendous improvement for π near 0 or 1.

  e.g., Brown, Cai, and Das Gupta (2001):
  n_0 required such that cov. prob. ≥ 0.94 for all $n \geq n_0$ is

π	Wald	Adjusted
0.01	7963	1

Some comparisons (95% CI)

- n_0 required such that cov. prob. ≥ 0.94 for all $n \geq n_0$ is

π	Wald	Adjusted
0.01	7963	1
0.10	646	11
0.20	292	89
0.30	245	78
0.50	194	94

Comments

- 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.
- Shrinkage form of adjusted intervals suggests intervals resulting from Bayesian approach also perform well in a frequentist sense.

  Single proportion: Brown et al. (2001)

  Comparing proportions: Agresti and Min (2005)

Sample size guidelines (well ...)

Finally, an embarrassing difficulty with ordinary large-sample Wald CIs is sample size guidelines for their use.

Advantage of *add two successes and two failures* adjusted intervals is decent performance for (nearly) **all** n.

In fact, you don't need any data !!! :-)

Single-sample: $\tilde{\pi} = (t_{obs} + 2)/(n + 4) = 2/4$

95% adjusted CI is $.5 \pm 2\sqrt{(.5)(.5)/4}$, or (0, 1).

Two-sample: $\tilde{\pi}_1 = 1/2$ and $\tilde{\pi}_2 = 1/2$
95% adjusted CI is

$$(.5 - .5) \pm 2\sqrt{[(.5)(.5)/2] + [(.5)(.5)/2]},$$ or $(-1, +1)$.

Bibliography (selective to AA and colleagues)

Agresti, A. (1992), A survey of exact inference for contingency tables, *Statistical Science*.

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

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 confi dence intervals for parameters in discrete distributions, *Biometrics*.

Agresti, A. (2001). Exact inference for categorical data: recent advances and continuing controversies, *Stat. Medicine*.

Agresti, A., and Min, Y. (2002). Unconditional small-sample confi dence intervals for the odds ratio, *Biostatistics*.

Agresti, A. (2002). Dealing with discreteness *Stat. Methods Medical Res*.

Agresti, A., and Min, Y. (2005). Bayesian CIs for 2x2 contingency tables, *Biometrics*.

Agresti, A., and Gottard, A. (2007). Nonconservative exact small-sample inference for discrete data, *Comput. Statist. & Data Anal.*