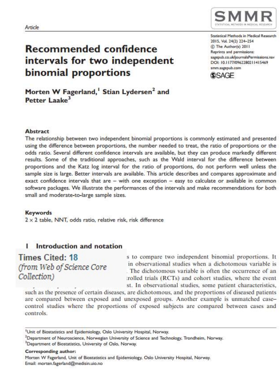
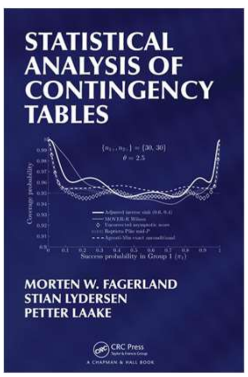
Analysis of 2x2 contingency tables: Hypothesis tests and confidence intervals

Different versions of the Pearson chi squared tests, the Fisher exact test, and Wald confidence intervals are widely used for contingency tables. Unfortunately, some of these methods are also commonly used in situations when they perform poorly, and better alternatives exist. I will present recommended methods for 2x2 tables in different situations.

Stian Lydersen
18 January 2018 and 29 January 2018
Revised 29 January 2018
http://folk.ntnu.no/slyderse/medstat/Pres18jan2018.pdf

Norwegian University of Science and Technology





Fagerland, M. W., Lydersen, S., & Laake, P.:
"Statistical Analysis of Contingency Tables."
Chapman and Hall/CRC, 2017.

www.contingencytables.com

Norsk:
Krysstabell (~~kontingenstabell~~)

Hypothesis tests for associations in 2x2 tables

TABLE 4.2
A possible result of Muriel Bristol's blind taste test

	Guessed		
Poured	Milk first	Tea first	Total
Milk first	3	1	4*
Tea first	1	3	4*
Total	4*	4*	8*

*Fixed by design

TABLE 4.3
Treatment of epinephrine in children with cardiac arrest (Perondi et al., 2004)

	Survival at 24h		
Treatment	Yes	No	Total
Standard dose	7 (21%)	27 (79%)	34*
High dose	1 (2.9%)	33 (97%)	34*
Total	8 (12%)	60 (88%)	68*

*Fixed by design

The number of successes n_{i1} in row number i is assumed bin(n_{i+} , π_i)

Expected counts mij under H0:

TABLE 4.4
Exposure to GADA for children with IPEX versus IPEX-like syndromes (Lampasona et al., 2013)

	Cases/controls		
GADA	**IPEX**	**IPEX-like**	**Total**
Positive	6.26 ~~9~~ (69%)	6.74 ~~4~~ (29%)	13 (48%)
Negative	6.74 ~~4~~ (31%)	7.26 ~~10~~ (71%)	14 (52%)
Total	13*	14*	27*

*Fixed by design

TABLE 4.1
The observed counts of a 2×2 table

	Success	**Failure**	**Total**
Group 1	n_{11}	n_{12}	n_{1+}
Group 2	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	N

The estimated expected cell counts of the 2×2 table under the null hypothesis are

$$m_{ij} = n_{i+}n_{+j}/N, \quad i,j = 1,2, \tag{4.5}$$

and the Pearson chi-squared test statistic is

$$T_{\text{Pearson}}(\mathbf{n}) = \sum_{i,j} \frac{(n_{ij} - m_{ij})^2}{m_{ij}} = \frac{N(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1+}n_{2+}n_{+1}n_{+2}}. \tag{4.6}$$

Asymptotically chi squared distributed with 1 d.f. under H0.
Cochran's criterion: OK if all m_ij ≥ 5.

Example Table 4.4:
T = 4.464, p=0.035

The asymptotic Pearson chi-squared test does not necessarily preserve the significance level, and may perform poorly in small samples.

Yates (1934) proposed the following modification of Equation 4.6:

$$\begin{aligned} T_{\text{YatesCC}}(\mathbf{n}) &= \sum_{i,j} \frac{\left(|n_{ij} - m_{ij}| - 1/2\right)^2}{m_{ij}} \\ &= \frac{N\left(|n_{11}n_{22} - n_{12}n_{21}| - N/2\right)^2}{n_{1+}n_{2+}n_{+1}n_{+2}}. \end{aligned} \tag{4.7}$$

The modification in Equation 4.7 is called a continuity correction, and the purpose of the correction is to obtain asymptotic P-values closer to exact P-values. The use of Yates's continuity correction—and other continuity corrections suggested in the literature—has been widely debated; see, for instance, the historical review in Hitchcock (2009). Many now consider such corrections to be no more than "interesting historic curiosities" (Hirji, 2006, p. 149).

Example Table 4.4:
T Yates CC = 2.984, p=0.084

Possible counts x_{ij} if the marginals are fixed (or conditioned on)

TABLE 4.1
The observed counts of a 2×2 table

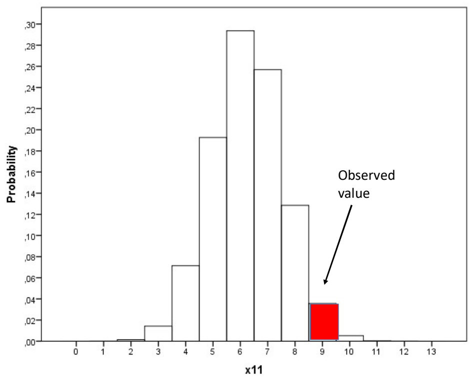
	Success	**Failure**	**Total**
Group 1	x_{11}		n_{1+}
Group 2			n_{2+}
Total	n_{+1}	n_{+2}	N

Additional notation:
$\mathbf{n} = \{n_{11}, n_{12}, n_{21}, n_{22}\}$: the observed table
$\mathbf{x} = \{x_{11}, x_{12}, x_{21}, x_{22}\}$: any possible table

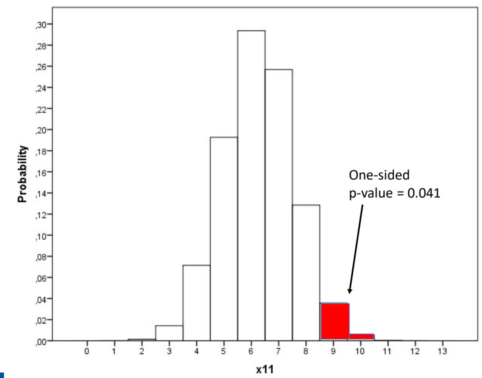
Fisher's exact test uses the following fact:
Conditional on the marginals, x1 is hypergeometrically distributed under H0

$$f(x_{11} \mid n_{1+}, n_{2+}, n_{+1}, n_{+2}) = \frac{\binom{n_{1+}}{x_{11}}\binom{n_{2+}}{n_{+1} - x_{11}}}{\binom{N}{n_{+1}}}. \tag{4.2}$$

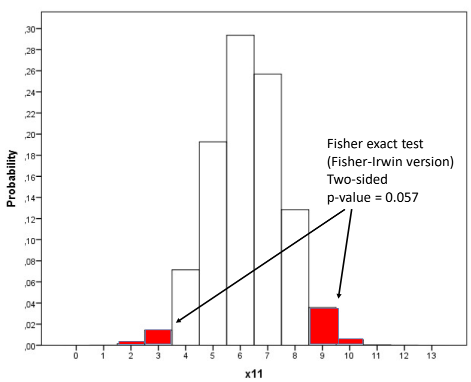
This distribution does not depend on the unknown common success probability (nuisance parameter) under H0!



P-value = P(Observed or more extreme)



P-value = P(Observed or more extreme)



What is more extreme (than the observed value)? Different test statistics (Norsk: testobservator) may be used to measure «distance» from the null hypothesis:

- Use the conditional point probabililty as in the Fisher-Irwin version of the Fisher exact test
- The Pearson chi squared statistic
- Etc

- In a one-sided test in a 2x2 table, the choice of (monotonely increasing) statistics does not matter.
- Also the case for twice the smallest tail two sided tests
- The choice of statistic matters for other two sided tests in 2x2 tables, and for rxc tables with r>2 or c>2.
- In 2x2 tables, the conditional probability (Fisher) and the Pearson statistics generally perform well.

Example output from SPSS

Operation type * Severe nausea Crosstabulation

			Severe nausea		
			No	Yes	Total
Operation type	CABG	Count	39	12	51
		% within Operation ty	76,5%	23,5%	100,0%
	Other	Count	3	5	8
		% within Operation ty	37,5%	62,5%	100,0%
Total		Count	42	17	59
		% within Operation ty	71,2%	28,8%	100,0%

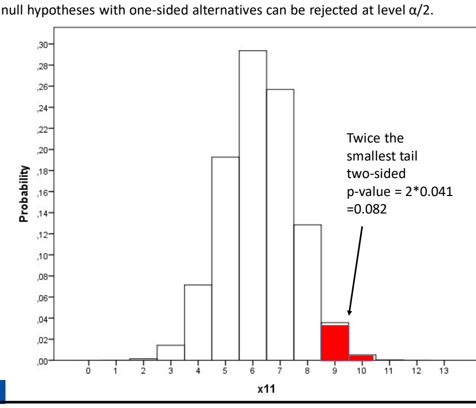
Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	5,120[b]	1	,024	,037	,037
Continuity Correction[a]	3,397	1	,065		
Likelihood Ratio	4,620	1	,032	,090	,037
Fisher's Exact Test				,037	,037
N of Valid Cases	59				

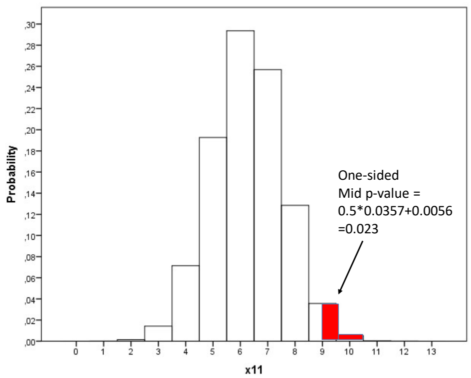
a. Computed only for a 2x2 table

b. 1 cells (25,0%) have expected count less than 5. The minimum expected count is 2,31.

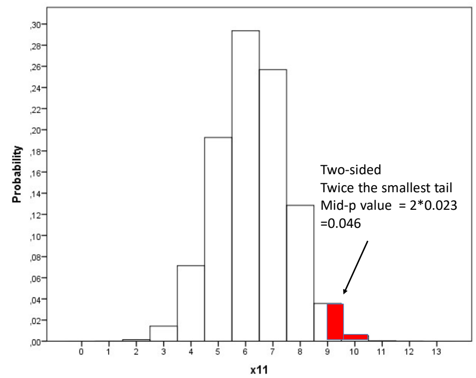
Twice the smallest tail approach:
Rejecting a two-sided alternative hypotheses at level α implies that both the null hypotheses with one-sided alternatives can be rejected at level α/2.



Mid-P value = 0.5 P(Observed) + P(more extreme)



Mid-P value = 0.5 P(Observed) + P(more extreme)



Mid-p tests and Mid p confidence intervals:

- Solid theoretical justifications (Fagerland, Lydersen, Laake, 2017, page 28-29, and references therein)
- Reduces the conservatism of exact methods
- Do not preserve nominal significance level (tests) and nominal coverage (confidence intervals), but the violations are usually not serious
- An ideal p-value is U(0,1) under H0. Exact p-values for categorical data are right skewed. Mid-p values have expectation 0.5 and is approximately U(0,1)
- In most cases the mid-p approach gives methods with better properties than those based on asymptotic normal theory. A notable exception is testing for equality of paired binomial distributions, where the McNemar asymptotic test is better than the McNemar mid-P test (Fagerland, Lydersen, Laake, 2017, page 29)

4.4.7 Exact Unconditional Tests

The probability of observing an arbitrary table $\mathbf{x}$ under H_0 is then given by Equation 4.3 with $\pi = \pi_1 = \pi_2$, which we rewrite slightly to

$$f(\mathbf{x} \mid \pi, \mathbf{n}_+) = \binom{n_{1+}}{x_{11}} \binom{n_{2+}}{x_{21}} \pi^{x_{11}+x_{21}} (1-\pi)^{N-x_{11}-x_{21}},$$

where $\mathbf{n}_+ = \{n_{1+}, n_{2+}\}$ denotes the fixed row sums. An explicit expression for the exact unconditional P-value is

$$P\text{-value} = \max_{0 \le \pi \le 1} \left\{ \sum_{\Omega(\mathbf{x}|\mathbf{n}_+)} I\left[T(\mathbf{x}) \ge T(\mathbf{n})\right] \cdot f(\mathbf{x} \mid \pi, \mathbf{n}_+) \right\}, \qquad (4.14)$$

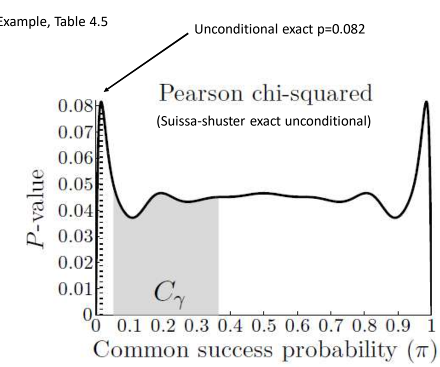
where $\Omega(\mathbf{x}|\mathbf{n}_+)$ denotes the set of all tables with row sums equal to $\mathbf{n}_+$.

TABLE 4.5
Genotype and presence of XFS in the eyes (Ritland et al., 2007)

	XFS		
Genotype	Yes	No	Total
CHRNA4-CC	0	16 (18%)	16 (18%)
CHRNA4-TC/TT	15 (17%)	57 (65%)	72 (82%)
Total	15 (17%)	73 (83%)	88* (100%)

*Fixed by design

Example, Table 4.5



The Berger and Boos Procedure

It may be argued that maximizing π over the entire nuisance parameter space, $0 \leq \pi \leq 1$, is unreasonable because the interval contains values that are highly unlikely in light of the observed data. This argument was the crux of Fisher's criticism (Fisher, 1945) of Barnard's proposition of the exact unconditional test. The Berger and Boos procedure is a remedy (Berger and Boos, 1994). It restricts the nuisance parameter space to C_γ: a $100(1-\gamma)\%$ exact confidence interval for π, where γ is taken to be very small. To make sure that the actual significance level is bounded by the nominal level, the value of γ is added to the P-value:

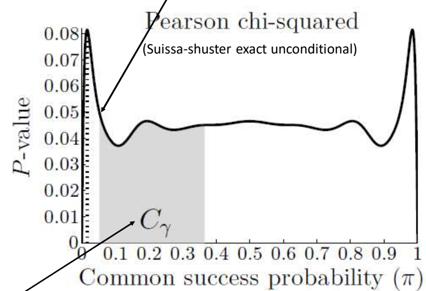
$$P\text{-value} = \max_{\pi \in C_\gamma} \left\{ \sum_{\Omega(\mathbf{x}|\mathbf{n}_+)} I\left[T(\mathbf{x}) \geq T(\mathbf{n})\right] \cdot f(\mathbf{x} \,|\, \pi, \mathbf{n}_+) \right\} + \gamma.$$

For the Suissa-Shuster exact unconditional test, Lydersen et al. (2012b) found $\gamma = 0.0001$ to be approximately optimal under rather general conditions. In addition to avoiding computation over unrealistic values of the nuisance parameter, Berger (1996) states two other advantages of using the Berger and Boos procedure: (i) maximization over C_γ is computationally easier than over $0 \leq \pi \leq 1$; and (ii) the resulting test can have higher power than the ordinary exact unconditional test. The user manual of the software package StatXact also notes that using the Berger and Boos procedure provides greater computational stability (StatXact 11, 2015, p. 528). A common method to form the confidence interval C_γ is to use the Clopper-Pearson exact interval (see Section 2.4.7).

NTNU

Example, Table 4.5

Unconditional exact with BB (γ=0.0001) p=0.0499



99.99% CI for the common success probability

NTNU

Evaluation criteria for hypothesis tests:

1. The actual significance level (ASL) should ideally equal the nominal significance level (usually 5%). If the ASL level is lower, say 2% or 3%, the test is conservative. If the ASL is higher, the test is liberal.
2. Among tests with acceptable ASL, we prefer the one with highest power.

27 NTNU

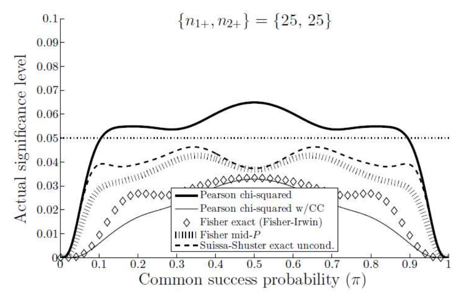


FIGURE 4.4
Actual significance levels of five commonly used tests

NTNU

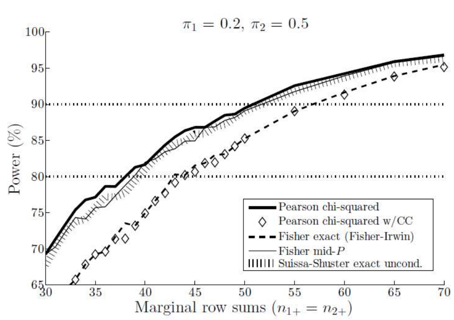


FIGURE 4.10
Power of five commonly used tests

NTNU

Fisher's exact test:
Advantage: Always preserves level
Disadvantage: Conservative (lower power than necessary)

Pearson's asymptotic test:
Advantage: Not conservative
Disadvantage: Does not always preserve level.

Pearson's asymptotic test with Yates' continuity correction:
Disadvantage: Does not always preserve level
Disadvantage: Conservative

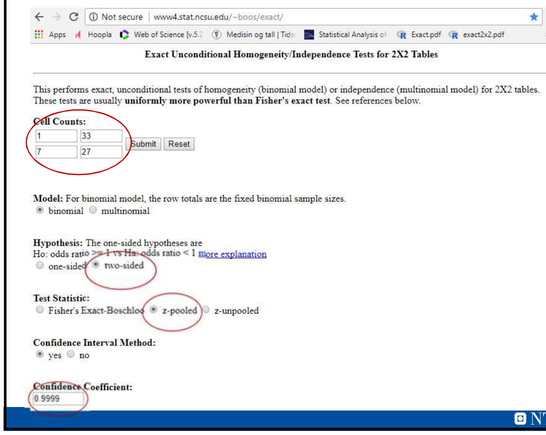
Fisher (and other) Mid-p:
Advantage: Not conservative
Disadvantage: Does not always preserve level (but seldom large violations)

Exact unconditional tests:
Advantage: Not conservative
Advantage: Always preserves level
Disadvantage: Computer-intensive in moderate and large sample sizes

NTNU

Exact unconditional tests are available in

- StatXact (Cytel software)
- R package «Exact»
- http://www4.stat.ncsu.edu/~boos/exact/
- …



Not secure | www4.stat.ncsu.edu/~boos/exact/

Exact Unconditional Homogeneity/Independence Tests for 2X2 Tables

This performs exact, unconditional tests of homogeneity (binomial model) or independence (multinomial model) for 2X2 tables. These tests are usually **uniformly more powerful than Fisher's exact test**. See references below.

Cell Counts:

1	33
7	27

Submit Reset

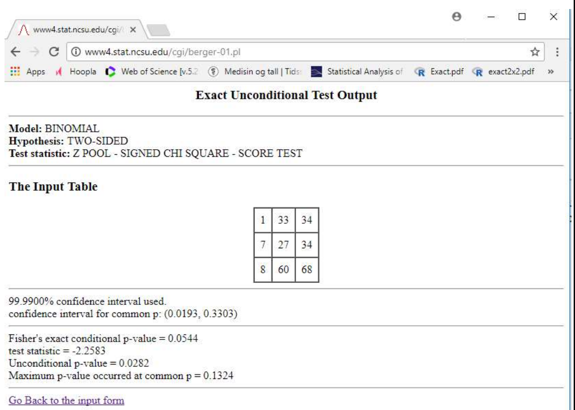
Model: For binomial model, the row totals are the fixed binomial sample sizes.
binomial multinomial

Hypothesis: The one-sided hypotheses are
Ho: odds ratio >= 1 vs Ha: odds ratio < 1 more explanation
one-sided two-sided

Test Statistic:
Fisher's Exact-Boschloo z-pooled z-unpooled

Confidence Interval Method:
yes no

Confidence Coefficient:
0.9999



www4.stat.ncsu.edu/cgi/berger-01.pl

Exact Unconditional Test Output

Model: BINOMIAL
Hypothesis: TWO-SIDED
Test statistic: Z POOL - SIGNED CHI SQUARE - SCORE TEST

The Input Table

1	33	34
7	27	34
8	60	68

99.9900% confidence interval used.
confidence interval for common p: (0.0193, 0.3303)

Fisher's exact conditional p-value = 0.0544
test statistic = -2.2583
Unconditional p-value = 0.0282
Maximum p-value occurred at common p = 0.1324

Go Back to the input form

TABLE 4.24
Recommended tests and confidence intervals (CIs) for 2 × 2 tables

Analysis	Recommended methods	Sample sizes
Tests for association	Fisher mid-P*	all
	Suissa-Shuster exact unconditional†	small/medium
	Fisher-Boschloo exact uncond.†	small/medium
	Pearson chi-squared*	large

*These methods have closed-form expression
†Preferably with the Berger and Boos procedure ($\gamma = 0.0001$)

Confidence intervals in 2x2 tables

Effect measures in 2x2 tables,
comparing two binomial probabilities π_1 and π_2 :

Probabillity difference
(risk difference, absolute risk reduction, attributable risk)

$$\Delta = \pi_1 - \pi_2$$

Ratio of probabilities (risk ratio, relative risk)

$$\phi = \pi_1 / \pi_2$$

Odds ratio

$$\theta = \frac{\pi_1 / (1 - \pi_1)}{\pi_2 / (1 - \pi_2)}$$

Difference between probabilities:
Maximum likelihood estimate (difference between sample proportions):

$$\hat{\Delta} = \hat{\pi}_1 - \hat{\pi}_2 = \frac{n_{11}}{n_{1+}} - \frac{n_{21}}{n_{2+}}.$$

The traditional Wald confidence interval for Δ is based on the asymptotic normal distribution of $\hat{\Delta}$:

$$\hat{\Delta} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_{1+}} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_{2+}}}, \qquad (4.19)$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentile of the standard normal distribution.

A continuity corrected version of the Wald interval, due to Yates (1934), can be expressed as shown in Fleiss et al. (2003, p. 60):

$$\hat{\Delta} \pm \left[z_{\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_{1+}} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_{2+}}} + \frac{1}{2}\left(\frac{1}{n_{1+}} + \frac{1}{n_{2+}}\right)\right]. \qquad (4.20)$$

Agresti-Caffo (2000) CI for the original table is simply the Wald CI for the table with one failure and one success added to each group:

TABLE 4.1
The observed counts of a 2×2 table

	Success	Failure	Total
Group 1	n_{11} +1	n_{12} +1	n_{1+} +2
Group 2	n_{21} +1	n_{22} +1	n_{2+} +2
Total	n_{+1} +2	n_{+2} +2	N +4

Agresti and Caffo (2000) proposed a simple, yet effective procedure for computing a confidence interval: add one success and one failure (pseudo-frequencies, see Section 2.4.4) in each sample and calculate the Wald confidence interval on the resulting data:

$$\tilde{\pi}_1 - \tilde{\pi}_2 \pm z_{\alpha/2}\sqrt{\frac{\tilde{\pi}_1(1-\tilde{\pi}_1)}{\tilde{n}_{1+}} + \frac{\tilde{\pi}_2(1-\tilde{\pi}_2)}{\tilde{n}_{2+}}}, \qquad (4.21)$$

where

$$\tilde{n}_{1+} = n_{1+} + 2,\ \tilde{n}_{2+} = n_{2+} + 2,\ \tilde{\pi}_1 = (n_{11}+1)/\tilde{n}_{1+},\ \tilde{\pi}_2 = (n_{21}+1)/\tilde{n}_{2+}.$$

Newcombe (1998b) proposed a confidence interval for Δ formed as a combination of the Wilson score (Section 2.4.3) confidence limits for π_1 and the Wilson score confidence limits for π_2. Denote the interval for π_1 by (l_1, u_1) and the interval for π_2 by (l_2, u_2). The Newcombe hybrid score confidence interval (L, U) for Δ is given by

$$L = \hat{\Delta} - \sqrt{(\hat{\pi}_1 - l_1)^2 + (u_2 - \hat{\pi}_2)^2} \qquad (4.22)$$

and

$$U = \hat{\Delta} + \sqrt{(\hat{\pi}_2 - l_2)^2 + (u_1 - \hat{\pi}_1)^2}. \qquad (4.23)$$

Wilson score CI:

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

Example, two binomials:
Treatment of children with cardiac arrest.
(Perondi et al, NEJM, 2004)

Epinephrine treatment	survival at 24 hours yes	no	Total
High dose	1	33	34
Standard dose	7	27	34
total	8	60	68

Fisher's exact test, two-sided p=0.054
Exact z-pooled (Suissa & Shuster) unconditional test: p=0.028

Estimated probability difference: 1/34 – 7/34 = -0.176
95% CI: Wald: -0.324 to -0.029
Agresti-Caffo: -0.322 to -0.012
Newcombe hybrid score: -0.340 to -0.019

Evaluation criteria for confidence intervals:

1. Coverage probability: Ought to not dip (much) below the nominal coverage (usually 95%). This is the primary criterion.
2. Interval width: Among intervals with similar coverage, we prefer the narrower.
3. Interval location: We say that an interval is located too distally if it is located too far out from the centre of symmetry for the effect measure (the midpoint). If the interval is located too close to the midpoint, we say that the interval is too mesially located. A 1−α confidence interval has ideal location if both the left and right non-coverage are equal to α/2.

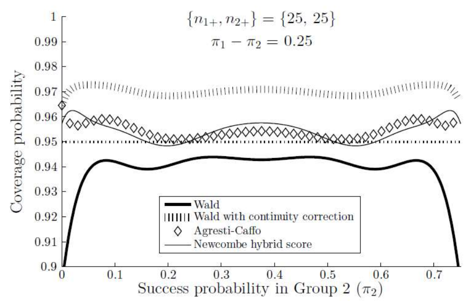


FIGURE 4.17
Coverage probabilities of four confidence intervals of closed-form expression for the difference between probabilities

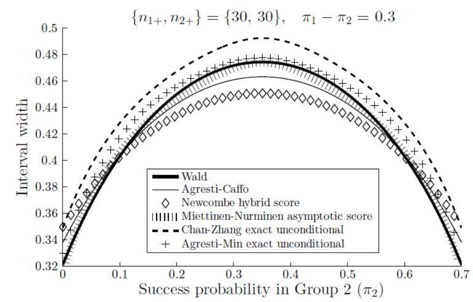


FIGURE 4.22
Expected interval width of six confidence intervals for the difference between probabilities

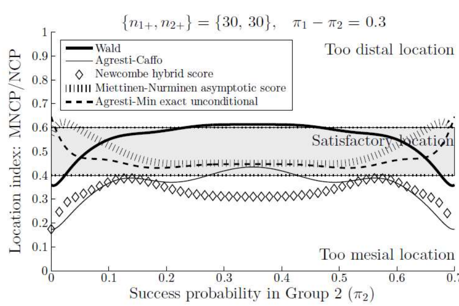
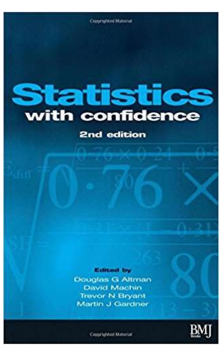


FIGURE 4.23
Location, as measured by the MNCP/NCP index, of five confidence intervals for the difference between probabilities

TABLE 4.24
Recommended tests and confidence intervals (CIs) for 2 × 2 tables

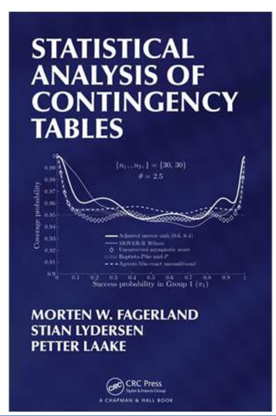
Analysis	Recommended methods	Sample sizes
Tests for association	Fisher mid-P*	all
	Suissa-Shuster exact unconditional†	small/medium
	Fisher-Boschloo exact uncond.†	small/medium
	Pearson chi-squared*	large
CIs for difference between probabilities	Agresti-Min exact unconditional†	small/medium
	Agresti-Caffo*	medium/large
	Newcombe hybrid score*	medium/large
	Miettinen-Nurminen asympt. score	medium/large
	Wald*	large
CIs for number needed to treat	The reciprocals of the limits of the recommended intervals for the difference between probabilities	
CIs for ratio of probabilities	Adjusted inverse sinh*	all
	MOVER-R Wilson*	all
	Koopman asymptotic score	all
	Agresti-Min exact unconditional†	small/medium
	Katz log*	large
CIs for odds ratio	Adjusted inverse sinh*	all
	MOVER-R Wilson*	all
	Baptista-Pike mid-P	all
	Agresti-Min exact unconditional†	small/medium
	Woolf logit*	large

*These methods have closed-form expression
†Preferably with the Berger and Boos procedure ($\gamma = 0.0001$)



Altman, D.G., Machin, D., Bryant, T.N., & Gardner, M.J. 2000. Statistics with Confidence, 2 ed. Bristol, BMJ books.

Includes software for recommended CI



Tables with r rows and c columns:

- 1x2
- 1xc
- 2x2
- Ordered rx2 and ordered 2xc
- Unordered, singly ordered and doubly ordered rxc
- Paired 2x2 and paired cxc
- 2x2xk and other stratified tables

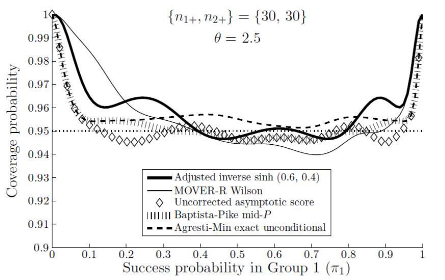


FIGURE 4.42
Coverage probabilities of the best performing intervals for the odds ratio

Fagerland, M. W., Lydersen, S., & Laake, P.: "Statistical Analysis of Contingency Tables." Chapman and Hall/CRC, 2017.

- "This book should be a very useful reference for anyone who wants an overview of the relevant literature (much of it quite recent) or who routinely needs to analyze contingency tables." Alan Agresti.
- "I highly recommended it for masters and doctoral students in statistics … and other fields requiring the analysis of discrete data." Karim F. Hirji
- "I strongly recommend the book both to statisticians and to researchers in health and social disciplines." Robert G. Newcombe
- "… an essential book to own if you analyse low-dimensional contingency tables." John McDonald
- "This book is encyclopaedic in its coverage and would be useful to graduate students and all applied statisticians who are always dealing with contingency tables." Michael J. Campbell

www.contingencytables.com

Available in a bookstore near you …