Exact and Conservative Inference in Blocked Experiments with Binary Outcomes

Permutation and Causal Inference: Connections and Applications
IMSI
Chicago, IL

Philip B. Stark joint w/ Skip Garibaldi, Jiaxun Li, Jake Spertus, Mayuri Sridhar

24 August 2023

University of California, Berkeley

Stratification and Blocking

- statistical reasons
  - "take advantage of" (anticipated) within-stratum homogeneity (true for MSE but not necessarily for inference)
  - sometimes need stratumwise estimates/inferences

Stratification and Blocking

- statistical reasons
  - "take advantage of" (anticipated) within-stratum homogeneity (true for MSE but not necessarily for inference)
  - sometimes need stratumwise estimates/inferences
- logistics:
  - randomize independently at different centers
  - distribute work

Stratification and Blocking

- statistical reasons
  - "take advantage of" (anticipated) within-stratum homogeneity (true for MSE but not necessarily for inference)
  - sometimes need stratumwise estimates/inferences
- logistics:
  - randomize independently at different centers
  - distribute work
- analysis/inference:
  - stratification/blocking often ignored in clinical trial data
  - Fisher's exact test
  - Student's t
  - stratified surveys generally use normal approx.
  - texts: Kish; Cochran; Thompson; Levy & Lemeshow; Hansen, Hurwitz, & Madow; ...

Exact inference about binary populations from stratified samples

N items. G labeled "1." $N - G$ labeled "0." Partitioned into S strata.

Stratum s contains N_s items, of which G_s are labeled "1."

$N = \sum_{s=1}^{S} N_s$ and $G := \sum_{s=1}^{S} G_s$.

Draw simple random sample of size n_s from stratum s, independently across strata.

Y_s is the number of items labeled "1" in the sample from stratum s.

$\{Y_s\}_{s=1}^{S}$ are independent. Observed value of Y_s is y_s.

Seek hypothesis tests and confidence bounds for G.

COMMUNICATIONS IN STATISTICS
Theory and Methods
Vol. 33, No. 9, pp. 2245–2257, 2004

Estimation of Proportion of Success From a Stratified Population: A Comparative Study

A. Nanthakumar[1,*] and K. Selvavel[2]

[1]Department of Mathematics, SUNY-Oswego, Oswego, New York, USA
[2]Office of the Inspector General, Department of Defense, Arlington, Virginia, USA

ABSTRACT

This article attempts to compare the different interval estimation methods for a stratified population where each stratum represents a binomial population. We compare Wald, Wilson, modified Agresti–Coull and Clopper–Pearson type intervals for both "with" and "without" replacement sampling scheme. The Wilson type interval performs well when compared to other intervals, but it fails to achieve the coverage probability when the proportion of success in each of the stratum is near 0 or 1. None of these methods are reliable

when the proportion of success for each of the stratum is near the boundaries. Indeed, Wilson, Wald and Agresti–Coull intervals have coverage probabilities much below the nominal confidence level. The coverage probability for the modified Clopper–Pearson Type interval is much higher than the nominal confidence level.

Wright's (1991) method for CIs

- Add simultaneous LCBs for $(G_s)_{s=1}^S$ to get LCB; add simultaneous UCBs to get UCB.
  - Samples from different strata are independent: use Šidák's adjustment, $(1-\alpha)^{1/S}$.
  - Find CI for G_s by inverting hypergeometric tests using Y_s
  - General method: joint $1-\alpha$ confidence set for all the parameters $\{G_j\}_{j=1}^S$, then find a bound on functionals of interest over the joint set.
- Lots of slack:
  - unnecessarily constrains $S-1$ nuisance parameters
  - not tight "geometry" for the desired functional

JASA Th & Meth 1996

Exact Inference for Proportions From a Stratified Finite Population

John P. WENDELL and Josef SCHMEE

Auditors and others often encounter finite populations with a dichotomous characteristic from which they draw stratified samples. In auditing the dichotomy arises when a population item is classified as either in error or in compliance with some rule or regulation. Usually the proportion of errors is small. The auditing objective may require calculation of a p value for the sample outcome relative to a hypothesis, or a confidence bound for the proportion or total number of errors in the population. In sampling from L strata with hypotheses concerning the total number of errors in the population, the calculation of p values is not straightforward. The complication arises because the parameter of the null hypothesis does not completely specify the distribution of the test statistic. This distribution depends on an $(L - 1)$-dimensional nuisance parameter consisting of the number of errors in each stratum. Because confidence bounds can be obtained by inverting the hypothesis test, the same difficulty applies to calculating confidence bounds. This article tests H_1 using the maximum p value over the feasible set of nuisance parameters. It describes a fairly efficient search method for finding a global maximum p value. Confidence bounds are calculated by inverting the hypothesis test. The article also presents an heuristic expression for determining good starting values in the search for confidence bounds. The procedures are implemented on a standard statistical package and are available from StatLib. They seem to perform reasonably well with samples from a moderate number of strata with a small number of errors.

KEY WORDS: Attribute sampling; Confidence bound; Hypergeometric distribution; Nuisance parameters; p value; Statistical auditing.

2.2.2 P Values in Stratified Sampling with only M_t Specified. In many audit applications, only the hypothesized number of errors in the population, the error threshold M_t, is specified. Because $(M_{t_1}, \ldots, M_{t_L})$ is needed to calculate the outcome probabilities, $(M_{t_1}, \ldots, M_{t_L})$ becomes a vector of nuisance parameters with the restriction that $M_t = \sum_{i=1}^{L} M_{t_i}$. $(M_{t_1}, \ldots, M_{t_L})$ cannot be easily eliminated. Different specifications of the nuisance parameters yield different outcome probabilities and thus different p values. The problem of calculating the p value can be overcome by choosing a nuisance parameter that yields the most

(9, 1).)

The sample estimate of the number of errors $\hat{M}_{st}$ is 4.1667, versus the hypothesized M_t of 10. The variance $\hat{V}(\hat{M}_{st})$ is 4.9102. The resulting standardized z_{pnorm} is -2.6325, which corresponds to a p_{norm} of .0042.

2.2.4 Results Comparing p_{max} to p_{norm}. Table 2 presents the values of p_{max} and the normal distribution approximation p_{norm} for a selection of typical audit populations and sample results. The uncorrected normal distribution severely underestimates the actual p values in all cases investigated.

```
S = 2
N_1 = N_2 = 100
n_1 = 60
n_2 = 40
y_1 = y_2 = 1
```


Table 1. p Values for all (M_1, M_2)

M_1	M_2	p value	
1	9	.061574	
2	8	.067081	p_{max}
3	7	.062872	
4	6	.054091	
5	5	.043091	
6	4	.034080	
7	3	.025498	
8	2	.018467	
9	1	.012978	

upper bounds:

$$U_{st(W)} = \sum_{i=1}^{L} U'_{srs(i)},$$

where $U'_{srs(i)}$ is an $100(1-\gamma')\%$ upper confidence bound for stratum i based on SRS calculations and with $\gamma' = 1 - \sqrt[L]{1-\gamma}$.

The 95% upper confidence bound for M based on p_{max} for the example in Section 2.2.3 can be established by cal-

- P-value for pop total is max P-value over stratum totals that give that pop total: $S-1$-dimensional nuisance parameter
- Each P-value uses test statistic $\hat{p} := \frac{1}{N}\sum_{s=1}^{S} N_s y_s / n_s$, like norm approx
- CIs by inverting tests (CI includes all pop totals for which an allocation isn't rejected at level α)
- Maximizing the P-value over all allocations of G ones across S strata is combinatorial:
  - Feller's "bars and stars" $\binom{G+S-1}{S-1}$ ways to allocate G objects among S strata (some don't honor data or stratum sizes).
  - $S = 10$, $N_s = 400$, $G = 300 \implies \approx 6.3e+16$ allocations
  - search intractable when there are many 1s or more than a few strata
- Nonconvex objective: no guarantee numerical optimization will succeed
- W&S use exhaustive search & numerical optimization by descent from some number of random starting points.

duce the number of evaluations in applications with combinatorially larger spaces of points.

4.2 Finding U_{st}

This section presents four steps for the efficient calculation of an upper confidence bound for M.

Step 1 guesses a starting point. A good heuristic starting point is

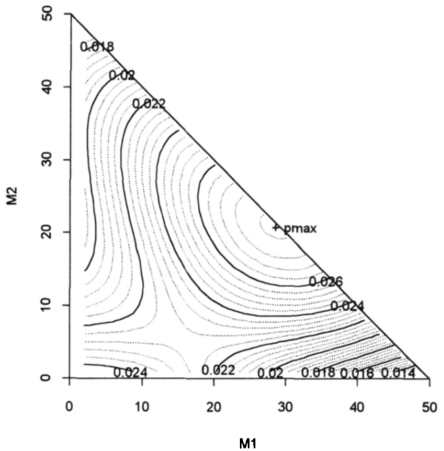

Figure 1. Contour Plot of p Values Over Nuisance Parameter Space for N = (500, 300, 200), n = (75, 50, 25), y = (2, 1, 0), and M_t = 50. p_{max} is .02768 and is found at (M_1, M_2, $M_t - M_1 - M_2$) = (28, 21, 1).

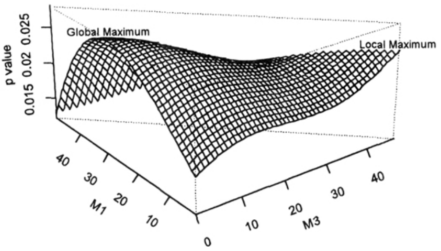

Figure 2. Grid Plot of p Values Over Nuisance Parameter Space for N = (500, 300, 200), n = (75, 50, 25), y = (2, 1, 0), and M_t = 50. The global maximum p value, p_{max}, is .02768 and is at (M_1, $M_t - M_1 - M_3$, M_3) = (28, 21, 1). The locally maximum p value is .02433 and is at (M_1, $M_t - M_1 - M_3$, M_3) = (2, 1, 47).

Table 2. Comparison of p_{max} and p_{norm} for Selected Cases With Computation Times for p_{max} in Seconds: $N = (N_1, \ldots, N_L)$, $n = (n_1, \ldots, n_L)$, $y_{obs} = (y_i, \ldots, y_L)$

N	n	y_{obs}	M_t	p_{max}	p_{norm}	Seconds
(200, 100)	(50, 25)	(0, 0)	15	.01194	0	12
(200, 100)	(50, 50)	(1, 0)	15	.07232	.00077	15
(2,000, 1,000)	(50, 50)	(1, 0)	150	.09958	.00268	16
(300, 200)	(75, 50)	(1, 1)	25	.02918	.00027	18
(300, 200)	(100, 100)	(3, 2)	25	.03514	.00496	23
(500, 500)	(100, 50)	(2, 1)	50	.08662	.00424	18
(5,000, 5,000)	(100, 50)	(2, 1)	500	.10908	.00676	28
(100, 100, 100)	(25, 25, 25)	(0, 0, 0)	15	.01205	0	44
(300, 200, 100)	(50, 50, 50)	(1, 1, 0)	30	.03775	.00103	63
(3,000, 2,000, 1,000)	(50, 50, 50)	(1, 1, 0)	300	.04902	.00255	97
(300, 200, 100)	(75, 50, 25)	(1, 1, 0)	30	.00931	.00004	75
(500, 300, 200)	(50, 50, 50)	(2, 1, 0)	50	.12760	.04756	117
(500, 300, 200)	(75, 50, 25)	(2, 1, 0)	50	.02768	.00137	45
(5,000, 3,000, 2,000)	(75, 50, 25)	(2, 1, 0)	500	.03639	.00272	85

Basic strategy: maximize P-value over a multidimensional nuisance parameter

- P-value for composite null is the maximum of the P-values of the simple nulls that comprise the composite.
- The individual P-values can be hard to find.
- Representing simple nulls as intersection hypotheses helps.
- Union-of-intersections tests:

$$H_G = \cup_{\mathbf{g}:\sum_s g_s = G} \cap_{s=1}^{S} H_{s,g_s}$$

- Test intersections by combining (independent) P-values.
  - Inspired by NPC to build multivariate tests from univariate tests

Different test statistic makes the optimization trivial!

Define

$$p_s(g_s) := \mathbb{P}\{Y_s \geq y_s || G_s = g_s\} = \sum_{y=y_s}^{g_s} \frac{\binom{g_s}{y}\binom{N_s - g_s}{n_s - y}}{\binom{N_s}{n_s}},$$

where $\binom{a}{b} := 0$ if $a \leq 0$ or $b > a$.

P-value for the most powerful test of the hypothesis $G_s = g_s$ against the alternative $G_s > g_s$.

Test the intersection hypothesis $G_s = g_s$, $s = 1, \ldots, S$ by combining (independent) stratumwise P-values, e.g., using Fisher's combining function.

If all S stratumwise nulls are true, the distribution of

$$X^2(\vec{g}) := -2\sum_{s=1}^{S} \log p_s(g_s)$$

is dominated by the chi-square distribution with $2S$ degrees of freedom. Let $\chi_d(z)$ denote the chance that a random variable with the chi-square distribution with d degrees of freedom is greater than or equal to z.

A conservative P-value for the allocation $\vec{g}$ is

$$P(\vec{g}) = \chi_{2S}(X^2(\vec{g})).$$

The allocation $\vec{g}$ of g ones across strata that maximizes the P-value minimizes minimizes $X^2(\vec{g})$ (maximizes $\sum_{s=1}^{S} \log p_s(g_s)$) and satisfies $\sum_s g_s = g$.

Let

$$a_s(j) := \begin{cases} \log p_s(y_s), & j = y_s \\ \log\left(p_s(j)/p_s(j-1)\right), & j = y_s + 1, \ldots N_s - (n_s - y_s). \end{cases}$$

Then $\log p_s(g_s) = \sum_{j=y_s}^{g_s} a_s(j)$ if $y_s \leq g_s \leq N - (n_s - y_s)$, and $\log p_s(g_s) = -\infty$ otherwise. Moreover,

$$X^2(\vec{g}) = -2\sum_{s=1}^{S} a_s(y_s) - 2\sum_{s=1}^{S}\sum_{j=y_s+1}^{g_s} a_s(j)$$

provided $y_s \leq g_s \leq N - (n_s - y_s)$, $s = 1, \ldots, S$; otherwise, it is infinite.

An allocation of g ones across strata is inconsistent with the data unless $g_s \geq y_s$, $s = 1, \ldots, S$.

How to allocate the remaining $g - \sum_s y_s$ ones to maximize the P-value (equivalently, to minimize $X^2(\vec{g})$)?

Let b_k denote the kth largest element of the bag

$$\{\!\{ a_s(j) \}\!\}_{j=y_s+1}^{N_s-(n_s-y_s)}{}_{s=1}^{S}$$

with ties broken arbitrarily. Define $\tilde{g}_y := g - \sum_{s=1}^{S} y_s$.

Proposition. For every $\vec{g}$ with $\sum_s g_s = g$,

$$X^2(\vec{g}) \geq X^2_*(g) := \begin{cases} -2\left(\sum_{s=1}^{S} a_s(y_s) + \sum_{k=1}^{\tilde{g}_y} b_k\right), & \sum_s y_s \leq g \leq N - \sum_s (n_s - y_s) \\ \infty, & \text{otherwise.} \end{cases}$$

Proof. Any $\vec{g}$ for which $X^2(\vec{g})$ is finite includes the first sum and a sum of $\tilde{g}_y$ elements of $\{b_k\}$; the latter is at most the sum of the $\tilde{g}_y$ largest elements of $\{b_k\}$. $\square$

Proposition: For $j \in y_s + 1, \ldots, N_s - (n_s - y_s)$, $a_s(j)$ is monotone decreasing in j. (Equivalently, $p_s(j)$ is concave in j.)

Implies the bound is sharp: if $a_s(i)$ is a term in the second sum for some $i > y_s + 1$, so is every $a_s(j)$, $y_s \leq j \leq i - 1$: the second sum corresponds to an allocation $\vec{g}$ of g ones across the S strata, with $y_s \leq g_s \leq N_s - (n_s - y_s)$.

Among all allocations of g 1s, this one minimizes the tail probability, because it corresponds to exponentiating the smallest sum of logs (the largest negative sum of logs). $\square$

Theorem: If $\sum_s y_s \leq g \leq N - \sum_s (n_s - y_s)$,

$$P(g) \leq \chi_d(X_*^2(g)).$$

A "greedy" approach finds a conservative P-value:

- Add the S values $\wr a_s(x_k)\wr_{s=1}^{S}$ to the $g - g_y$ largest elements of $\wr b_k \wr$.
- Upper tail probability of the chi-square distribution with $2S$ degrees of freedom for -2 times the sum is a conservative P-value for the hypothesis $G = g$.
- A conservative upper $1 - \alpha$ confidence bound for G is the largest g for which $P(g) \geq \alpha$.

Special case of maximizing a weakly concave function over a polymatroid. Rado-Edmonds Theorem guarantees the greedy algorithm succeeds.
(Componentwise concavity implies weak concavity over $\mathcal{J} \subset \mathbb{Z}^S$.)

Same greedy approach gives lower bound on spending for lottery wins.

Operation count

- Calculate $a_s(j)$ and $a_s(j+1)$ for all j ($2S$ function evaluations)
- Evaluate $a_s(\cdot)$ once for each remaining step for the stratum a 1 is added to ($g - \sum_s y_s - 1$ evaluations), if $g_s < N_s$.
- When a 1 is allocated, have to find a largest element of $\{a_s(g_s+1)\}_{s=1}^{S}$.
  - Sort at the first step in $O(S \ln S)$ operations,
  - Update sort as elements are replaced in $O(S(g - \sum_s y_s - 1))$ operations

Comparison to Wendell & Schmee (1996)

				P-values	
N	n	observed	g	Greedy	WS
$[200, 100]$	$[50, 25]$	$[0, 0]$	15	0.06482	0.01194
$[200, 100]$	$[50, 50]$	$[0, 20]$	60	0.01686	0.03340
$[300, 200]$	$[75, 50]$	$[1, 1]$	25	0.09105	0.02918
$[300, 200]$	$[75, 50]$	$[0, 15]$	100	0.00703	0.00563
$[300, 300]$	$[50, 50]$	$[0, 20]$	200	0.00039	0.00106
$[5000, 5000]$	$[100, 50]$	$[2, 1]$	500	0.21563	0.10908
$[5000, 5000]$	$[100, 50]$	$[10, 0]$	1000	0.04454	0.04493
$[15000, 5000, 1000]$	$[150, 30, 10]$	$[3, 2, 0]$	2000	0.02123	*
$[50000, 15000, 5000, 1000]$	$[500, 150, 30, 10]$	$[5, 3, 2, 0]$	2750	0.02735	*

$*$ calculation hadn't finished in 5 minutes

Directions to explore

- other P-value combining functions that yield weak concavity, so greedy algorithm still works
- base stratumwise tests on E-values from test supermartingales
  - product of independent E-values is an E-value for the intersection null
  - predictable interleaving of terms from stratum test supermartingales is a test supermartingale for the intersection
    - choose stratum test SMs for each null
    - choose interleaving: "gang of bandits" problem
  - no adjustment for # strata needed
  - works for bounded populations, not only binary populations
  - sequential validity: can sample until CI is as short as desired
  - generally need guardrails to keep an E-value from approaching 0 in stata w true nulls
  - generally, order of data matters

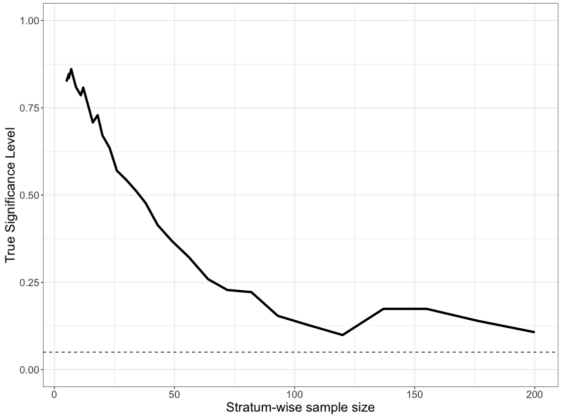

Figure 1: Estimated significance level of a stratified, one-sample t-test of the null hypothesis $H_0 : \mu \leq \eta_0$. The nominal level of the test is 5% and the true level was estimated at every sample size by 1000 simulations. The solid black line plots the true significance level (y-axis) of the test against a range of sample sizes within each of two equally-sized strata (x-axis). For example, when 50 samples are taken from each of the strata (a total sample size of 100) the true level of the test is around 35%. Both strata are mixture distributions: each sample has a 99% probability to be drawn from a truncated normal centered at $\mu_k = 0.505$ with standard deviation $\sigma_k = 0.001$ and a 1% probability of being identically 0. The true population mean is $\mu = 0.49995$ and the null mean is $\eta_0 = 0.5$. For the test to be valid for this population, the true level should always be below the dashed line at $\alpha = 5\%$.

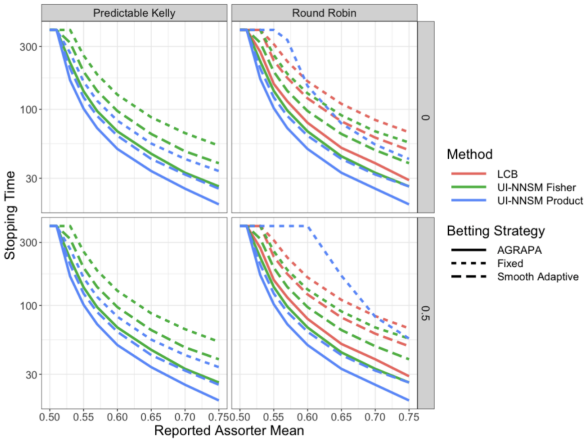

Figure 2: Stopping times (y-axis; $\log_{10}$ scale) against global reported assorter mean (x-axis) for stratified ballot-level comparison audits without CVR error. There are two strata of equal size $N_1 = N_2 = 200$, and equal assorter mean $\bar{A}_1^c = \bar{A}_2^c = \bar{A}_c$ displayed on the x-axis. Three different bets are displayed as linetypes. The strategy for testing is either to combine lower confidence bounds (LCB) akin to Wright's method, or union-intersection testing by nonnegative supermartingales (UI-NNSM).

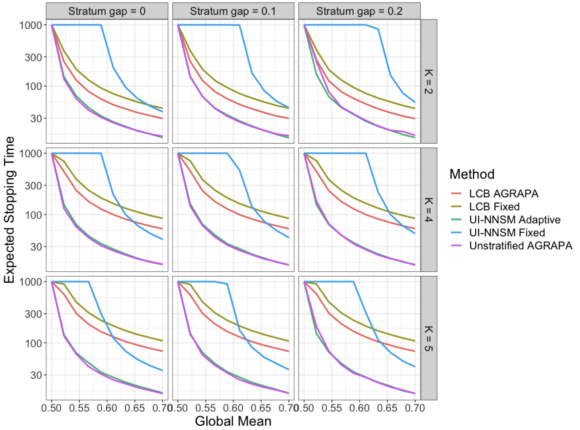

Figure 3: Expected stopping times (y-axis; $\log_{10}$ scale) of various sequential-stratified tests (line colors) of the null $H_0 : \mu \leq 1/2$ against a range of true global means (x-axis) at level 0.05. Populations consist of $N_1 = N_2 = 500$ units within each stratum, drawn from truncated Gaussian distributions with standard deviation $\sigma = 0.05$. The number of strata K varies over the rows. The global mean μ is on the x-axis, while the columns correspond to the largest gap between stratum means, with other means spaced linearly between the largest and smallest. LCB = lower confidence bound; UI-NNSM = union-intersection nonnegative supermartingale

Blocked/stratified experiments

Comparison of adaptive pacing therapy, cognitive behaviour therapy, graded exercise therapy, and specialist medical care for chronic fatigue syndrome (PACE): a randomised trial

P D White, K A Goldsmith, A L Johnson, L Potts, R Walwyn, J C DeCesare, H L Baber, M Burgess, L V Clark, D L Cox, J Bavinton, B J Angus, G Murphy, M Murphy, H O'Dowd, D Wilks, P McCrone, T Chalder, M Sharpe*, on behalf of the PACE trial management group†*

Summary

Background Trial findings show cognitive behaviour therapy (CBT) and graded exercise therapy (GET) can be effective treatments for chronic fatigue syndrome, but patients' organisations have reported that these treatments can be harmful and favour pacing and specialist health care. We aimed to assess effectiveness and safety of all four treatments.

Methods In our parallel-group randomised trial, patients meeting Oxford criteria for chronic fatigue syndrome were recruited from six secondary-care clinics in the UK and randomly allocated by computer-generated sequence to receive specialist medical care (SMC) alone or with adaptive pacing therapy (APT), CBT, or GET. Primary outcomes were fatigue (measured by Chalder fatigue questionnaire score) and physical function (measured by short form-36 subscale score) up to 52 weeks after randomisation, and safety was assessed primarily by recording all serious adverse events, including serious adverse reactions to trial treatments. Primary outcomes were rated by participants, who were necessarily unmasked to treatment assignment; the statistician was masked to treatment assignment for the analysis of primary outcomes. We used longitudinal regression models to compare SMC alone with other treatments, APT with CBT, and APT with GET. The final analysis included all participants for whom we had data for primary outcomes. This trial is registered at http://isrctn.org, number ISRCTN54285094.

Findings We recruited 641 eligible patients, of whom 160 were assigned to the APT group, 161 to the CBT group, 160 to the GET group, and 160 to the SMC-alone group. Compared with SMC alone, mean fatigue scores at 52 weeks were 3·4 (95% CI 1·8 to 5·0) points lower for CBT (p=0·0001) and 3·2 (1·7 to 4·8) points lower for GET (p=0·0003), but did not differ for APT (0·7 [−0·9 to 2·3] points lower; p=0·38). Compared with SMC alone, mean physical function scores were 7·1 (2·0 to 12·1) points higher for CBT (p=0·0068) and 9·4 (4·4 to 14·4) points higher for GET (p=0·0005), but did not differ for APT (3·4 [−1·6 to 8·4] points lower; p=0·18). Compared with APT, CBT and GET were associated with less fatigue (CBT p=0·0027; GET p=0·0059) and better physical function (CBT p=0·0002; GET p<0·0001). Subgroup analysis of 427 participants meeting international criteria for chronic fatigue syndrome and 329 participants meeting London criteria for myalgic encephalomyelitis yielded equivalent results. Serious adverse reactions were recorded in two (1%) of 159 participants in the APT group, three (2%) of 161 in the CBT group, two (1%) of 160 in the GET group, and two (1%) of 160 in the SMC-alone group.

Interpretation CBT and GET can safely be added to SMC to moderately improve outcomes for chronic fatigue syndrome, but APT is not an effective addition.

Lancet 2011; 377: 823–36

Published Online
February 18, 2011
DOI:10.1016/S0140-6736(11)60096-2

See Comment page 786

*Authors contributed equally

†Members listed at end of paper

Wolfson Institute of Preventive Medicine, Barts and The London School of Medicine, Queen Mary University of London, London, UK (Prof P D White *MD*, J C DeCesare *BSc*, H L Baber *BSc*, L V Clark *PhD*); **Mental Health and Neuroscience Clinical Trials Unit, Institute of Psychiatry, King's College London, London, UK** (K A Goldsmith *MPH*, L Potts *MSc*, R Walwyn *MSc*); **Medical Research Council Biostatistics Unit, Institute of Public Health, University of Cambridge, UK** (A L Johnson *PhD*); **Medical Research Council Clinical Trials Unit, London, UK** (A L Johnson); **South London and Maudsley NHS Foundation Trust, London, UK** (M Burgess *PhD*); **Faculty of Health and Well Being, University of Cumbria, Lancaster, UK** (D L Cox *PhD*); **Nuffield Department of**

Methods

Study design and participants

PACE was a parallel, four group, multicentre, randomised trial, with outcomes assessed up to 52 weeks after randomisation for patients with chronic fatigue syndrome.[10] We recruited 641 participants from consecutive new outpatients attending six specialist chronic fatigue syndrome clinics in the UK National Health Service between March 18, 2005, and Nov 28, 2008, and completed outcome data collection in January, 2010.

of consent. A database programmer undertook treatment allocation, independently of the trial team. The first three participants at each of the six clinics were allocated with straightforward randomisation. Thereafter allocation was stratified by centre, alternative criteria for chronic fatigue syndrome[12] and myalgic encephalomyelitis,[13] and depressive disorder (major or minor depressive episode or dysthymia),[14] with computer-generated probabilistic minimisation. Once notified of treatment allocation by the Clinical Trials Unit, the research assessor informed the

naire and 11 for short form-36). Prorating involved calculating the mean value of the item scores present and replacing the missing values with that score.

We summarised continuous variables with mean (SD) or median (IQR) and categorical variables with frequencies and proportions. Differentiation of treatment compared independent ratings of therapy sessions with actual treatment. We calculated the inter-rater reliability (κ and 95% CI) between the two assessors. We used Kruskal-Wallis tests for comparisons of therapy received, therapeutic alliance, and manual adherence. We compared categorical variables with Fisher's exact test.

A clinically useful difference between the means of the primary outcomes was defined as 0·5 of the SD of these measures at baseline,[31] equating to 2 points for Chalder fatigue questionnaire and 8 points for short scores of the UK working age population of 84 (–24) for physical function (score of 60 or more).[32,33]

We estimated differences between treatment groups for both primary outcomes with mixed linear regression models with Kenward-Roger adjusted standard errors. Covariates were treatment group, baseline value of outcome, time, and stratification factors (centre, present depressive disorder, and alternative criteria for chronic fatigue syndrome and myalgic encephalomyelitis; all as stratified at entry). Time by treatment interaction terms were included to allow extraction of contrasts at 52 weeks. Models for the primary outcomes and the clinical global impression incorporated random intercepts and slopes over time by participant and main health-care practitioner (doctor or therapist who saw the participant most frequently, or, if equal, the first practitioner to see the participant) to allow for clustering of outcomes within

	Adaptive pacing therapy (n=159)	Cognitive behaviour therapy (n=161)	Graded exercise therapy (n=160)	Specialist medical care alone (n=160)	p value*
Treatment received					
Therapy sessions attended†	13 (12–15)	14 (12–15)	13 (12–14)	..	0·17
Specialist medical care sessions attended‡	3 (3–4)	3 (3–4)	3 (3–4)	5 (3–6)	0·0001
Adequate treatment§	143 (90%)	140 (87%)	136 (85%)	142 (89%)	0·56
Antidepressant at baseline	63 (40%)	57 (35%)	74 (46%)	66 (41%)	..
Antidepressant at 24 weeks¶	53 (34%)	45 (29%)	61 (40%)	60 (39%)	0·19
Antidepressant at 52 weeks¶	41 (27%)	47 (31%)	48 (31%)	61 (39%)	0·11
Hypnotic at baseline	6 (4%)	9 (6%)	6 (4%)	5 (3%)	..
Hypnotic at 24 weeks¶	3 (2%)	7 (5%)	5 (3%)	6 (4%)	0·61
Hypnotic at 52 weeks¶	5 (3%)	4 (3%)	3 (2%)	7 (5%)	0·62
Non-allocated treatment	8 (5%)	4 (3%)	7 (4%)	22 (14%)	0·0005
Dropouts from treatment	11 (7%)	17 (11%)	10 (6%)	14 (9%)	0·50
Views before treatment					
Treatment is logical	134 (84%)	115 (71%)	135 (84%)	79 (49%)	<0·0001
Confident about treatment	114 (72%)	91 (57%)	112 (70%)	65 (41%)	<0·0001
Views after treatment					
Satisfied with treatment¶	128 (85%)	117 (82%)	126 (88%)	76 (50%)	<0·0001
Dissatisfied with treatment¶	4 (3%)	7 (5%)	2 (1%)	17 (11%)	0·0010
Therapeutic alliance‖	6·5 (6·0–6·5)	6·5 (5·5–6·8)	6·5 (5·5–7·0)	..	0·96
Adherence to manual**	6·0 (6·0–6·5)	6·0 (5·0–6·5)	6·5 (6·0–6·5)	..	0·35

Data are median (IQR) or n (%). *p values across all groups. †86% of sessions were received face-to-face and 14% by telephone. ‡94% of sessions were received face-to-face and 6% by telephone. §Adequate treatment was ten or more sessions of therapy or three or more sessions of specialist medical care alone. ¶Percentages exclude missing data. ‖Scored 1–7 (1=poor, 7=excellent). **Scored 1–7 (1=not at all, 7=very much so).

Table 2: **Treatment details**

Exact inference in binary trials with binary outcomes

Neyman potential outcomes model: potential outcomes fixed before randomization, revealed by randomization.

Statistics
in Medicine

Research Article

Received 3 January 2014, Accepted 13 November 2014 Published online 4 December 2014 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6384

Randomization inference for treatment effects on a binary outcome

Joseph Rigdon and Michael G. Hudgens[*†]

Two methods are developed for constructing randomization-based confidence sets for the average effect of a treatment on a binary outcome. The methods are nonparametric and require no assumptions about random sampling from a larger population. Both of the resulting $1-\alpha$ confidence sets are exact in the sense that the probability of containing the true treatment effect is at least $1-\alpha$. Both types of confidence sets are also guaranteed to have width no greater than one. In contrast, a previously proposed asymptotic confidence interval is not exact and may have width greater than 1. The first approach combines Bonferroni-adjusted prediction sets for the attributable effects in the treated and untreated. The second method entails inverting a permutation test. Simulations are presented comparing the two randomization-based confidence sets with the asymptotic interval as well as the standard Wald confidence interval and a commonly used exact interval for the difference in binomial proportions. Results show for small to moderate sample sizes that the permutation confidence set attains the narrowest

Commentary

Statistics in Medicine

(wileyonlinelibrary.com) DOI: 10.1002/sim.6764

Published online in Wiley Online Library

Exact confidence intervals for the average causal effect on a binary outcome

Xinran Li[a] **and Peng Ding**[b*†]

Based on the physical randomization of completely randomized experiments, in a recent article in *Statistics in Medicine*, Rigdon and Hudgens propose two approaches to obtaining exact confidence intervals for the average causal effect on a binary outcome. They construct the first confidence interval by combining, with the Bonferroni adjustment, the prediction sets for treatment effects among treatment and control groups, and the second one by inverting a series of randomization tests. With sample size n, their second approach requires performing $O(n^5)$ randomization tests. We demonstrate that the physical randomization also justifies other ways to constructing exact confidence intervals that are more computationally efficient. By exploiting recent advances in hypergeometric confidence intervals and the stochastic order information of randomization tests, we propose approaches that either do not need to invoke Monte Carlo or require performing at most $O(n^2)$ randomization tests. We provide technical details and R code in the Supporting Information. Copyright © 2016 John Wiley & Sons, Ltd.

Theorem 1

A potential table $\boldsymbol{N}$ is compatible with the observed table $\boldsymbol{n}$ if and only if

$$\max\left\{0, n_{11}-N_{10}, N_{11}-n_{01}, N_{+1}-n_{10}-n_{01}\right\} \leqslant \min\left\{N_{11}, n_{11}, N_{+1}-n_{01}, n-N_{10}-n_{01}-n_{10}\right\}.$$

Fast computation of exact confidence intervals for randomized experiments with binary outcomes

P. M. Aronow Haoge Chang Patrick Lopatto*

Abstract

Given a randomized experiment with binary outcomes, exact confidence intervals for the average causal effect of the treatment can be computed through a series of permutation tests. This approach requires minimal assumptions and is valid for all sample sizes, as it does not rely on large-sample approximations such as the central limit theorem. We show that these confidence intervals can be found in $O(n \log n)$ permutation tests in the case of balanced designs, where the treatment and control groups have equal sizes, and $O(n^2)$ permutation tests in the general case. Prior to this work, the most efficient known constructions required $O(n^3)$ such tests in the balanced case [Li and Ding, 2016], and $O(n^4)$ tests in the general case [Rigdon and Hudgens, 2015]. Our results thus facilitate exact inference as a viable option for randomized experiments far larger than those accessible by previous methods.

Blocked binary experiment with binary outcomes

N subjects in all; N_s in block s.

n_s in block s assigned active treatment, $m_s := N_s - n_s$ assigned placebo. Assignment independent across blocks.

N_{1+}: # subjects whose response to treatment would be 1, $N_{1+,s}$ in block s

N_{+1}: # subjects whose response to placebo would be 1, $N_{+1,s}$ in block s

ATE: $\tau := (N_{1+} - N_{+1})/N$.

$n_{11,s}$: # subjects in block s who received active treatment and responded 1

$n_{01,s}$: # subjects in block s who received placebo treatment and responded 1

$n_{11,s} \sim \text{Hyp}(N_s, N_{1+,s}, n_s)$; $n_{01,s} \sim \text{Hyp}(N_s, N_{+1,s}, m_s)$

CIs for ATE

- Enumerate & test all blocked potential outcome tables consistent w/ results
  - Test statistic? Does $|\hat{\tau} - \tau|$ make sense? Analogous to WS: doesn't use stratum heterogeneity
- Use Li & Ding or Aronow et al. to find CIs for ATE within blocks, then combine using Šidák (analogous to Wright's method)
- Use Li & Ding or Aronow et al. to find a P-value within blocks, then combine across blocks (union of intersections test, again)
  - Exploit Aronow et al. $O(n_s \log n_s)$ result in the balanced blocks
- Apply the greedy approach to finding $1 - \alpha/2$ LCB for N_{1+} and UCB for N_{+1}, subtract, divide by N.
  - With UI-NNSM approach, can make inferences about ATE for bounded treatments