Group Sequential and Adaptive Designs

Christopher Jennison

Department of Mathematical Sciences,

University of Bath, UK

http://people.bath.ac.uk/mascj

Progress in Statistical Decision Theory, 2022

Celebration of Iain Johnstone's 65th birthday

Palo Alto

May 2022

Group Sequential and Adaptive Designs

While Iain and I were graduate students at Cornell, together with Bruce Turnbull, we wrote the paper.

> *Asymptotically optimal procedures for sequential adaptive selection of the best of several normal means.*

We presented this research at the *Purdue Symposium on Decision Theory and Related Topics* in 1981 and the paper was published in 1982.

I shall reflect on the results in this paper and subsequent research into sequential and adaptive methods, in particular for clinical trial designs.

First, some reminiscences:

Kiefer-Wolfowitz Conference, Cornell, 1983

ASYMPTOTICALLY OPTIMAL PROCEDURES FOR SEQUENTIAL ADAPTIVE SELECTION OF THE BEST OF SEVERAL NORMAL MEANS

Christopher Jennison[1], Iain M. Johnstone[2]

and Bruce W. Turnbull[1]

Departments of Operations Research and Mathematics
Cornell University
Ithaca, New York, U.S.A.

I. INTRODUCTION

Suppose we have $k(\geq 2)$ normal populations with common variance σ^2 and unknown means $\{\mu_i;\ 1 \leq i \leq k\}$. We wish to select a population with a "high" mean, the population with the highest mean is called the best population. Let $\mu_{[1]} \leq \mu_{[2]} \leq \ldots \leq \mu_{[k]}$ denote the ordered means. Bechhofer [1] formulated this problem with the following probability of correct selection (PCS) requirement:

(PCS 1) Whenever $\mu_{[k]} - \mu_{[k-1]} \geq \delta$,

$$P(\text{Select the best population}) \geq P^*,$$

where $\delta > 0$ and $1/k < P^* < 1$ are to be set by the experimenter.

Jennison, Johnstone and Turnbull, 1982 (JJT)

Asymptotically optimal procedures for sequential adaptive selection of the best of several normal means.

Given k populations from which observations have distributions

$$N(\mu_1, \sigma^2), \ldots, N(\mu_k, \sigma^2),$$

we wish to select the population with the largest mean.

Denoting the ordered means

$$\mu_{[1]} \leq \mu_{[2]} \leq \ldots \mu_{[k]},$$

we require

$$P\{\text{Select the best population}\} \geq P^* \text{ whenever } \mu_{[k]} - \mu_{[k-1]} \geq \delta.$$

Jennison, Johnstone and Turnbull, 1982 (JJT)

The selection procedure can be

Sequential:

Populations are eliminated one by one as the study proceeds

Adaptive:

Observations can be allocated in different proportions to the populations that remain

JJT defined a multi-stage procedure that achieves the lowest possible expected sample size in the limit as $P^* \to 1$.

For this procedure

$$\frac{E(\text{Total Sample Size})}{-\log(1-P^*)} \to \text{ a limit} \quad \text{as } P^* \to 1.$$

What happened next?

I shall give an overview of some subsequent developments in

Sequential,

Adaptive,

Multiple Testing

procedures.

I shall focus on applications in clinical trials.

Group sequential tests (GSTs)

Suppose a new treatment (Treatment A) is to be compared to a placebo or positive control (Treatment B) in a Phase 3 trial.

The treatment effect θ for the **primary endpoint** represents the advantage of Treatment A over Treatment B.

If $\theta > 0$, Treatment A is more effective.

We wish to test the null hypothesis H_0: $\theta \leq 0$ against $\theta > 0$ with

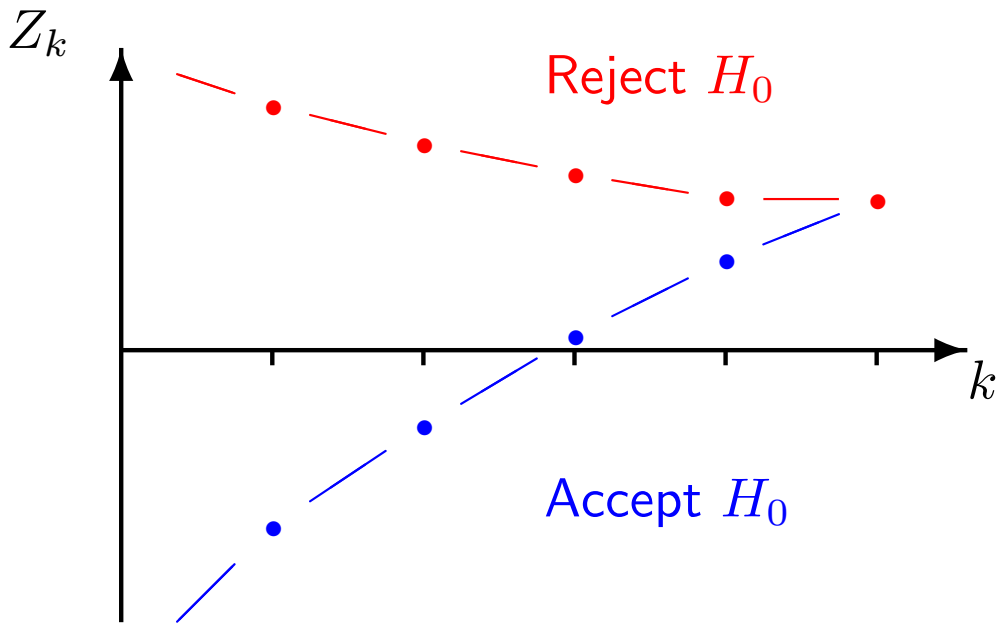
$$P_{\theta=0}\{\text{Reject } H_0\} = \alpha,$$

$$P_{\theta=\delta}\{\text{Reject } H_0\} = 1 - \beta.$$

In a group sequential trial, data are examined on a number of occasions to see if an early decision may be possible.

Group sequential tests (GSTs)

A typical boundary for a one-sided test, expressed in terms of standardised test statistics $Z_1, \ldots, Z_K$, has the form:



Crossing the upper boundary leads to early stopping for a positive outcome, rejecting H_0 in favour of $\theta > 0$.

Crossing the lower boundary implies stopping for "futility" with acceptance of H_0.

Benefits of group sequential testing

In order to test H_0: $\theta \leq 0$ against $\theta > 0$ with type I error probability α and power $1-\beta$ at $\theta = \delta$, a fixed sample size study needs information

$$\mathcal{I}_{fix} = \frac{\{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\}^2}{\delta^2},$$

where Φ is the standard normal CDF.

Information is (roughly) proportional to sample size in many clinical trial settings.

A GST with K analyses will need to be able to continue to a maximum information level $\mathcal{I}_K$, greater than $\mathcal{I}_{fix}$.

On average, the GST can stop earlier than this and expected information on termination, $E_\theta(\mathcal{I})$, will be considerably less than $\mathcal{I}_{fix}$, especially under extreme values of θ.

We call $R = \mathcal{I}_K / \mathcal{I}_{fix}$ the *inflation factor* of a group sequential test.

Optimal group sequential tests

We can seek a GST that minimises expected information $E_\theta(\mathcal{I})$ under certain values of the treatment effect, θ, with a given number of analyses K and inflation factor R.

Eales & Jennison (*Biometrika*, 1992) and Barber & Jennison (*Biometrika*, 2002) optimise designs for criteria of the form

$$\sum_i w_i \, E_{\theta_i}(\mathcal{I}) \quad \text{or} \quad \int f(\theta) \, E_\theta(\mathcal{I}) \, d\theta,$$

where f is a normal density.

These optimised designs could be used in their own right.

They also serve as benchmarks for other methods which may have additional useful features (e.g., error spending tests).

Computing optimal group sequential tests

In optimising a GST, we create a Bayes sequential decision problem, placing a prior on θ and defining costs for sampling and for making incorrect decisions.

Such a problem can be solved rapidly by dynamic programming.

We then search for the combination of prior and costs such that the solution to the (unconstrained) Bayes decision problem has the specified frequentist error rates α at $\theta = 0$ and β at $\theta = \delta$.

The resulting design solves both the Bayes decision problem and the original frequentist problem.

Complete Class Theorem:

The Bayes problem is introduced as a computational device — we could have called it a Lagrangian approach. Nevertheless, this derivation demonstrates that an efficient frequentist design should also be a good Bayesian procedure (and vice versa).

Benefits of group sequential testing

One-sided GSTs minimising $\{E_0(\mathcal{I}) + E_\delta(\mathcal{I})\}/2$ for K equally sized groups, $\alpha = 0.05$, $1 - \beta = 0.95$ and $\mathcal{I}_{max} = R\mathcal{I}_{fix}$.

Minimum values of $\{E_0(\mathcal{I}) + E_\delta(\mathcal{I})\}/2$, as a percentage of $\mathcal{I}_{fix}$

	R					
K	1.01	1.05	1.1	1.2	1.3	*Minimum over R*
2	80.9	74.5	72.8	**73.2**	75.3	72.6 at R=1.13
3	76.3	69.3	66.5	**64.8**	64.8	64.7 at R=1.25
5	72.2	65.2	62.2	59.8	**59.0**	58.7 at R=1.41
10	69.1	62.1	59.0	56.3	**55.2**	54.3 at R=1.61
20	67.6	60.5	57.4	54.6	**53.3**	51.9 at R=1.80
∞ (SPRT)						49.0 at R=∞

Note: $E(\mathcal{I}) \searrow$ as $K \nearrow$ but with diminishing returns,
$E(\mathcal{I}) \searrow$ as $R \nearrow$ up to a point.

Benefits of group sequential testing

Most of the benefits of sequential testing can be achieved by a group sequential test with a small number of analyses and a modest increase in maximum sample size over a fixed sample design.

Many "sequential" clinical trials have $K = 2$, i.e., one interim analysis and a final analysis.

I would recommend designs with 4 or 5 analyses and an inflation factor of 1.05 or 1.1.

Relating results on optimal GSTs to our 1982 paper

JJT: Theory tells us a reduction to 25% of the fixed sample size is possible in the limit as $\alpha = \beta \to 0$.

The smallest number in the preceding table is 49%.

Even the SPRT (the best one can do), approaches this limit slowly.

$\alpha\ (= \beta)$	SPRT's $\{E_0(\mathcal{I}) + E_\delta(\mathcal{I})\}/2$ as a percentage of $\mathcal{I}_{fix}$
0.05	49.0
0.01	41.6
10^{-3}	36.1
10^{-4}	33.3
10^{-5}	31.6
10^{-6}	30.6

Relating results on optimal GSTs to our 1982 paper

JJT: We applied results of Schwarz and Berk that show one can construct sequential tests which are asymptotically efficient for all θ values.

For typical error rates, α and β, it is possible to construct group sequential designs which are efficient over a range of θ values.

For example, if a design with K analyses and inflation factor R is chosen to minimise

$$\int f(\theta)\, E_\theta(\mathcal{I})\, d\theta,$$

where $f(\theta)$ is a $N(\delta/2,\, (\delta/2)^2)$ density, then this design will be highly efficient for θ between 0 and δ.

Furthermore, one can specify an error spending design that will have very similar boundaries and, thus, similarly high efficiency.

Adaptive design and multiple testing

A change in philosophy for clinical trials

From around 2000, there has been a growing interest in adaptive design of clinical trials.

Prior to this, the philosophy for Phase 3 clinical trials was to pre-specify all aspects of the designs and stay as close as possible to this plan.

With a decline in major breakthroughs in the pharmaceutical industry and a rise in late phase trial failures, new approaches were encouraged — and methods were proposed to achieve this while protecting against the risk of false positive results.

In 2006, these developments made the front page of the Wall Street Journal:

Adaptive design in clinical trials

Wall Street Journal, July 2006:

FDA Signals it's Open to Drug Trials that Shift Midcourse

Adaptive designs may allow trials to be adjusted:

- Route more patients to the treatment that seems to work best
- Drop treatments that don't seem to be effective
- Add more of the type of patients ... reacting best to a particular treatment
- Merge two different phases of drug development into one trial

With views from:

Bob O'Neill , FDA

Paul Gallo, Novartis

Tom Fleming, Univ. Washington

Michael Krams, Wyeth

Don Berry, M. D. Anderson Cancer Center

Bruce Turnbull, Cornell University

Adaptive design and multiple testing

In JJT, the asymptotic setting implies large sample sizes and the opportunity to learn a great deal about the true parameter values during the course of a study.

Thus, one can effectively use a design optimised for the true parameter values.

Results in JJT rely on a "mean path approximation" where it is argued that, with probability close to 1, the mean of a sample is very close to its expectation, even at an early point in the study.

Some of the comments promoting adaptive designs for current use appear to be based on similar reasoning.

However, in the non-asymptotic case, the picture can be very different.

Sample size re-estimation

Rather than use a group sequential test, some authors propose:

> Design a trial to achieve desired power at $\theta = \Delta_0$, an initial estimate of the true value of θ,
>
> Calculate an estimate $\widehat{\theta}$ at an interim analysis,
>
> Modify the sample size to achieve power in the case $\theta = \widehat{\theta}$.

Suppose a study is designed to test H_0: $\theta \leq 0$ vs $\theta > 0$ with type I error probability $\alpha = 0.025$ and power 0.9 when $\theta = 10$.

After observing half the data, the standard deviation of $\widehat{\theta}$ is 4.4.

So, if $\widehat{\theta} = 5.0$, a 95% confidence interval for θ is $(-3.6, 13.6)$.

Conclusion: Sequential tests do not gain their efficiency by "adapting to new estimates of θ".

Multi-arm multi-stage designs

Adaptive designs can contribute significantly when multiple hypotheses are to be tested:

- Seamless Phase 2 / Phase 3 designs
- Enrichment designs (shifting focus to a sub-population)
- Umbrella trials (multiple therapies for a single disease)
- Basket trials (one therapy, multiple diseases)

Even so, interim information on which to base adaptations tends to be noisy.

JJT's selection problem has much in common with the testing problem in an "umbrella trial" — and there are lessons to take from their 1982 solution.

Discussion and conclusions

What is the role of asymptotic theory?

Over the decades, numerical computation and simulation have played an increasing role — but exhaustive computation for all possible cases is not always an option.

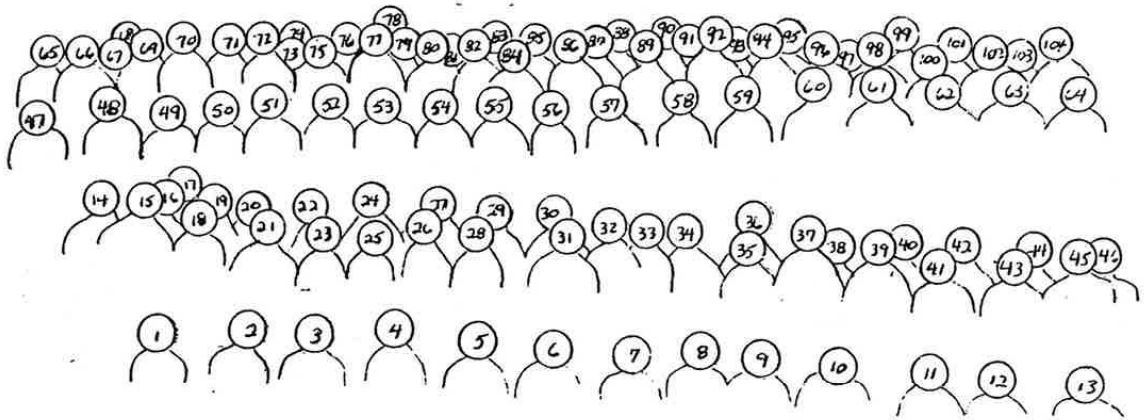
Theory is likely to reveal patterns and principles to guide us when choosing an approach to solve a new problem.

Presenting an impressive looking paper at a Purdue Symposium (or equivalent) is a good way to start your academic career!

Who's who in the Kiefer-Wolfowitz Conference photo

JACK KIEFER - JACOB WOLFOWITZ
MEMORIAL
STATISTICAL RESEARCH CONFERENCE

July 6-9, 1983 at Cornell University



(Participants)

1. R. Kulkarni, U. North Carolina
2. E. Seiden, Hebrew U.
3. H.L. Hwang, Northern Illinois
4. J. Hwang, Cornell
5. L. Weiss, Cornell
6. S. Blumenthal, U. Illinois, Urbana
7. H.K. Liu, Cornell
8. P.Y. Chen, Syracuse U.
9. M. Cecce, Cornell
10. C. Gatsonis, Rutgers
11. J.P.M. Schalkwijk, Eindhoven
12. P. Velleman, Cornell
13. R. Adler, Technion
14. C.F. Wu, Wisconsin
15. J. Srivastava, Colorado State
16. R. Smith, Imperial College
17. N. Kiefer, Cornell
18. M. Chow, Cornell
19. W. Sitonik, Cornell
20. G. Legall, Cornell
21. L. Hsu, Santa Barbara
22. W. Piegorsch, Cornell
23. H.F. Wu, Santa Barbara
24. D. Umbach, Cornell
25. B. Taylor, Sherbrooke
26. J. Marden, U. Illinois, Urbana
27. M. Perlman, U. Washington
28. V. Cullinan, Cornell
29. T. Berger, Cornell
30. L. Brown, Cornell
31. A. Cohen, Rutgers
32. W. Strawderman, Rutgers
33. S. Hedayat, U. Illinois, Chicago
34. I. Blumen, Cornell
35. H. Levine, Columbia
36. B. Hajek, U. Illinois, Urbana
37. M. Woodfoofe, #. Michigan
38. S. Searle, Cornell
39. Y.C. Yao, MIT
40. F. Pukelsheim, Freiburg
41. S.P. Lin, Rotkland Research Institute
42. T. Mount, Cornell
43. C. Srinivasan, Kentucky
44. I. Johnstone, Stanford
45. H. Teicher, Rutgers
46. C.C. Heyde, CSIRO
47. G. Casella, Cornell
48. T. Hayter, Cornell
49. A. Tamhane, Northwestern
50. S. Menjoge, Kentucky
51. S. Mishra, U. South Alabama
52. G. Shorrack, Ecole Polytechnique
53. R. Shorrack
54. I. Olkin, Stanford
55. R. Bechhofer, Cornell
56. A. Ash, Boston U.
57. D. Raghavarao, Temple U.
58. W. Federer, Cornell
59. T. Mitchell, Oak Ridge
60. G. Lorden, Cal. Tech.
61. D. Siegmund, Stanford
62. M. Sobel, Santa Barbara
63. W. Studden, Purdue
64.
65. Y. Grize, Cornell
66. P. Huber, Harvard
67. A. Shapiro, Unisa
68. T. Green, Cornell
69. D. Robson, Cornell
70. T.L. Lai, Columbia
71.
72. C. Hagwood, U. Virginia
73. K.F. Yu, Yale
74. A. Ruhkin, Purdue
75. S. Gupta, Purdue
76. D. Whittinghill, Purdue
77. A. Dvoretsky, Hebrew U.
78. T. Cover, Stanford
79. W. Notz, Purdue
80. N. Heckman, SUNY, Stony Brook
81. T. Fine, Cornell
82. S. Schwager, Cornell
83. W.J. Hall, Rochester
84. A. Benjamin, Cornell
85. T. Santner, Cornell
86. B. Turnbull, Cornell
87. K. Mieske, U. Illinois, Chicago
88. J. Sacks, Northwestern
89. C. Jennison, Durham
90. H. Wynn, Imperial College
91. D. Bancroft, Consumer's Union
92. C. McCulloch, Cornell
93. C. Blyth, Queens U.
94. R. Wolpert, Duke
95. J. Bondar, Carleton
96. R. Farrell, Cornell
97. L. Kuo, SUNY, Stony Brook
98. S. Ghosh, U.C., Riverside
99. J. Berger, Purdue
100. C.S. Cheng, Berkeley
101. M. Harel, IUT de Limoges
102. J. Shen, U. Cincinnati
103. G. Constantine, Indiana U.
104. R. Ahlswede, Bielefeld