
Optimising tests of efficacy in external pilot trials using Bayesian statistical decision theory

Journal Title
XX(X):2–12
©The Author(s) 0000
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/

SAGE

Duncan T. Wilson¹

Abstract

Introduction: External pilot trials of complex interventions are often conducted in advance of a definitive trial to assess feasibility and to inform its design. The efficacy of the intervention is rarely assessed using a formal hypothesis test since it would have low power, given the small sample size of a pilot and assuming a conventional type I error rate (e.g. 0.025). By not testing efficacy, an external pilot will effectively have a type I error rate of 1, suggesting an infinite preference for type I errors over type II errors. As such a preference will never occur in practice, we consider methods for finding the optimal balance of between type I and II error rates in external pilots. **Methods:** We consider the problem of determining the sample size and type I error rate which maximise the expected utility of an external pilot trial testing intervention efficacy. We introduce a utility function which accounts for improvement in primary outcome, the cost of sampling, treatment costs, and the decision-maker's attitude to risk. We apply the method to the re-design of a pilot trial with a continuous primary outcome with known standard deviation and where uncertainty in the treatment effect is quantified using a normal prior distribution.

Timing of potential results: A study of the proposed method's properties under a range of values for the utility function and prior distribution parameters is to be completed by August 2019.

Potential relevance and impact: By viewing external pilot trial design from a Bayesian decision-theoretic viewpoint, we will provide a method for finding the optimal balance of type I and II error rates in external pilots. In particular, we will identify in which (if any) settings the current approach of not assessing efficacy is the optimal course of action.

Keywords

Clinical trial, pilot trial, external pilot, statistical decision theory, optimal design

1 Introduction

Feasibility and pilot studies form a key stage in the development and evaluation of complex interventions¹, being conducted prior to a planned clinical trial assessing the effectiveness of the intervention under study and aiming to inform that trials feasibility and optimal design. Pilot trials are a particular type of feasibility study which takes the same form as the planned main trial, but at a smaller scale². Pilots may be internal, in which case the pilot data is analysed together with the main trial data and there is a seamless transition between the two; or external, in which case there is a clear gap between the trials and the data are kept separate.

Several authors have sought to clarify the purpose, design and analysis of pilot trials over the past 20 years³, with a common point being that hypothesis tests of effectiveness should not be carried out as they will have low power to detect an effect of interest⁴. This criticism rests on two assumptions. Firstly, it assumes that the effect size of interest in the pilot trial is the same as, or at least not much larger than, that which is of interest in the main trial. This may not be the case when the endpoint being tested in the pilot is different from that in the main trial (for example, a surrogate endpoint); or when the idealised setting of the pilot implies a larger effect should be expected than when under the more pragmatic setting of the main trial. Secondly, it assumes that the type I error rate of the pilot test will be the same as, or at least very similar to, that used in the main trial (e.g. the conventional choice of 0.025 one-sided). Under these restrictions, the fact that pilot trials are typically smaller than the planned main trial will indeed imply they will have low power when testing efficacy. For example, suppose a standardised effect size of 0.3 is of interest, and the pilot will randomise 35 participants to each arm as recommended in⁵, then a test in the pilot with one-sided type I error rate of 0.025 will have a power of 0.23.

By rejecting a possible pilot test with error rates $\alpha = 0.025$ and $\beta = 1 - 0.23 = 0.77$ and instead recommending not testing at all, we effectively obtain a procedure with error rates $\alpha = 1, \beta = 0$. This decision reflects an infinite preference for minimising type II errors over type I errors in the pilot, a preference too extreme to be expected in practice. Even if the magnitude of preference is substantial, any finite level would lead to a different choice. For example, an alternative would be $\alpha = 0.9, \beta = 0.006$. In many cases, we might expect a decrease in type I error of 0.1 in exchange for an increase in type II error of 0.006 to be a quite reasonable trade-off and hard to argue against.

The question which we will consider in this paper is how to determine the optimal error rates for a test of effectiveness in a pilot trial, to be conducted prior to a main trial to decide if it should be done, and based on a normally

¹Leeds Institute of Clinical Trials Research, University of Leeds, Leeds, UK

Corresponding author:

Duncan T. Wilson, Clinical Trials Research Unit, Leeds Institute of Clinical Trials Research, University of Leeds, Leeds, LS2 9JT, UK
Email: d.t.wilson@leeds.ac.uk

distributed primary endpoint which will be used at both stages. The general idea of relaxing type I error has been suggested explicitly⁶ and implicitly⁷, although in neither case are methods for choosing an appropriate level described. More generally, there is a lack of formal methods for the choice of error rates - some have argued for minimising their sum, or for balancing them, but the lack of methods is reflected in the ongoing yet much criticised convention of choosing $\alpha = 0.025, \beta = 0.2$ for all trials, regardless of the consequences of each type of error or the unconditional probabilities of their occurrence.

One possible strategy for doing so is through Bayesian statistical decision theory. If we can define a suitable utility function, we can make decisions which maximise the expected value of that function with respect to prior distributions on all unknown parameters. This framework, as we will show, can be used flexibly to determine the optimal type I error rates and sample sizes of the pilot trial, main trial, or both simultaneously.

2 Problem

An external pilot trial followed by a confirmatory trial are to be designed. We will denote the pilot trial as stage $i = 1$, and the definitive trial as stage $i = 2$ of the overall programme. Both trials will be parallel group studies comparing an intervention to control with respect to the same normally distributed primary outcome with known standard deviation common to each arm (although see Section 6 for relaxations of some of these assumptions). We denote the true mean difference by μ , and assume that the primary analysis at both stages will be a test of the null hypothesis $H_0 : \mu = 0$. The tests will compare the sample mean difference between groups x_i to associated critical values d_i , such that the stage i test is passed if $x_i > d_i$. Denoting the per-arm sample size at each stage by n_i , the full design of the pilot and definitive trial programme can be summarised by n_1, d_1, n_2, d_2 .

Given some alternative hypothesis $H_1 : \mu = \mu^*$, we define the type I and II error rates for stage i in the usual way, i.e. $\alpha_i = Pr[x_i > d_i \mid \mu = \mu_0]$ and $\beta_i = Pr[x_i \leq d_i \mid \mu = \mu^*]$. An alternative summary of the pilot and definitive trial programme is then $\alpha_1, \beta_1, \alpha_2, \beta_2$. Our primary interest is in how to determine optimal values of these four design variables.

3 Maximising expected utility

We consider a Bayesian view of the frequentist design problem, with a normal prior distribution $\mu \sim N(m_0, s^2)$ representing our knowledge and uncertainty about the true treatment effect. Moreover, we will use Bayesian statistical decision theory to define optimal design variables as those which maximise the expected value, with respect to the prior $p(\mu)$, of a utility function $u(\cdot)$. We construct $u(\cdot)$ by first selecting the attributes of interest and then defining a *value* function over the space of these attributes under conditions of certainty. We then transform the value function into a utility function by considering preferences under uncertainty.

3.1 Attributes and value

We consider the following three attributes to be of interest:

- the overall sample size, $n = n_1 + n_2$
- the cost incurred by changing the current standard treatment, denoted $C \in \{-, +\}$
- the change in the average outcome, denoted d

We use the notation $y \prec y'$ to mean we prefer y to y' . We aim to specify a value function $v(n, C, \mu)$ such that

$$(n, C, \mu) \prec (n', C', \mu') \Leftrightarrow v(n, C, \mu) < v(n', C', \mu').$$

To construct the value function we first assume that each pair of attributes is *preferentially independent* of the remaining attribute - that is, the trade-offs between any two attributes when the remaining attribute is kept fixed do not depend on the level it is fixed at⁸. Then, our preferences can be characterised by

$$v(n, C, d) = k_n v_n(n) + k_c v_c(C) + k_d v_d(d),$$

where

1. $v_i(worst) = 0, v_i(best) = 1, i = 1, 2, 3;$
2. $0 < \lambda_i < 1, i = 1, 2, 3;$
3. $\sum_{i=1}^3 \lambda_i = 1.$

We assume that the cost of sampling is constant, and so v_n is linear, and similarly assume linearity in the value of change in outcome, v_d . Attribute C only has two levels. We use identity value functions for all attributes, which will keep the notation simple. So,

$$v(n, C, d) = k_n n + k_c C + k_d d.$$

The scale of the value function is arbitrary, so we can set $k_n + k_c + k_d = 1$ without loss of generality. We can then deduce the parameter values given two set equivalence judgements. First, we ask for the change in outcome \bar{d} that would be needed for us to judge that

$$(n = 0, C = +, d = 0) \sim (n = n_*, C = +, d = \bar{d}).$$

That is, what change in outcome would we want to see to justify an increase in sample size from 0 to n_* ? Secondly, we ask for the change in outcome \hat{d} that would be needed to judge that

$$(n, C = -, d = 0) \sim (n, C = +, d = \hat{d}).$$

That is, what change in outcome would justify the moving from the current standard to the new treatment. We then have the following values for the scaling parameters:

$$\begin{aligned} k_d &= 1/(1 + \hat{d} + \bar{d}/n_*) \\ k_n &= -k_d \bar{d}/n_* \\ k_c &= k_d \hat{d}. \end{aligned}$$

3.2 Utility

The value function represents our preferences under conditions of certainty, but in reality we are uncertain about the values of all three attributes which will be obtained. To accommodate this uncertainty, we move from a value function to a *utility* function, where the latter encodes our preferences under uncertainty. A utility function allows us to compare probability distributions over the attributes, as opposed to only fixed points, and choose the distribution which has the largest expected utility.

To define the form of our utility function, we first argue that we have one attribute (the change in outcome) which is *utility independent* of the other two attributes. This means that our preferences for gambles on the change of outcome, with the other attributes kept fixed at some level, does not depend on those levels. So, regardless of how much we have sampled, or whether or not we have incurred the cost of changing the standard treatment or not, our preferences for gambles on the change in outcome will be the same. Given this assumption together with the additive form of the value function, the utility function must have one of the following forms:

$$u(n, C, d) = \begin{cases} 1 - e^{-\rho v(n, C, d)}, & \rho > 0 \\ v(n, C, d), & \rho = 0 \\ -1 + e^{-\rho v(n, C, d)}, & \rho < 0 \end{cases}$$

That is, the utility function over the scalar attribute $v(n, C, d)$ must be of the exponential form, which implies constant absolute risk aversion. Given the value function, the utility function is then defined by ρ . A value of $\rho = 0$ implies a risk-neutral attitude, whereas $\rho > 0$ (< 0) implies a risk-averse (risk-seeking) attitude.

To elicit ρ we first note that d is utility independent of n and C and so we can think about gambles on d whilst ignoring the value of the other attributes. For a given value of ρ we then consider the value d^* such that obtaining d^* with certainty would have the same expected utility as a gamble which will result in d_{min} with probability 0.5 and d_{max} with probability 0.5:

$$u(n, C, d^*) = 0.5u(n, C, d_{min}) + 0.5u(n, C, d_{max}).$$

In the special case of risk-neutrality, $\rho = 0 \Leftrightarrow d^* = 0.5d_{min} + 0.5d_{max}$. For $\rho \neq 0$, we have

$$d^* = -\frac{1}{\rho} \ln (0.5e^{-\rho d_{min}} + 0.5e^{-\rho d_{max}}).$$

For example, setting (arbitrarily) $d_{min} = 0$, $d_{max} = 1$, a value of $\rho = 2$ implies $d^* = 0.283$. That is, $\rho = 2$ implies an indifference between obtaining a guaranteed change in outcome of 0.283 and a simple 50/50 gamble between no change at all and a change of 1. To find the exact value of ρ for the problem at hand, we could plot the function $d^*(\rho)$ over a range of ρ .

3.3 Expected utility

Denote by G_i the event of a positive test result at stage $i = 1, 2$, and by \bar{G}_i its complement. In this case, $G_i = x_i > d_i$. The utility is now

$$\begin{aligned} u(\mu, x_1, x_2) = & 1 - \exp(-\rho(k_d\mu + k_n(n_1 + n_2))I[G_1 \& G_2] \\ & - \rho(k_n(n_1 + n_2) + k_c)I[G_1 \& \bar{G}_2] \\ & - \rho(k_n n_1 + k_c)I[G_1]). \end{aligned} \quad (1)$$

The expected utility conditional on μ is

$$\begin{aligned} E_{x_1, x_2, | \mu}[u(\mu, x_1, x_2)] = & Pr[G_1 \& G_2 | \mu] \left(1 - e^{\rho(k_d\mu + k_n(n_1 + n_2))}\right) + \\ & Pr[G_1 \& \bar{G}_2 | \mu] \left(1 - e^{-\rho(k_n(n_1 + n_2) + k_c)}\right) + \\ & Pr[\bar{G}_1 | \mu] \left(1 - e^{-\rho(k_n n_1 + k_c)}\right), \end{aligned} \quad (2)$$

Since the sample means are distributed as $x_i | \mu \sim N(\mu, 2\sigma^2/n_i)$ and are independent conditional on μ , the relevant probabilities are easily calculated. We are then left with integrating out the μ :

$$E[u(\mu, x_1, x_2)] = \int E_{x_1, x_2, | \mu}[u(\mu, x_1, x_2)]p(\mu)d\mu, \quad (3)$$

where $p(\mu)$ is the normal prior density. Because we are integrating with a normal density weighting function, we use Gauss-Hermite quadrature (as implemented in the ‘fastGHQuad’ R package⁹) to evaluate the integral efficiently and accurately. Specifically, Gauss-Hermite approximates integrals of the form

$$\int_{-\infty}^{\infty} e^{-x^2} f(x) dx \approx \sum_{i=1}^n w_i f(x_i).$$

To use when integrating over a normal distribution $y \sim N(\mu, \sigma^2)$, we use the transform $y = \sqrt{2}\sigma x + \mu$ so the integration points are now $x_i\sqrt{2}\sigma + \mu$ and the weights are $w_i/\sqrt{\pi}$.

For the special case where pilot trial OCS are fixed at $\alpha_1 = 1, \beta_1 = 0$ and only the confirmatory trial OCs are to be determined, the exact expected utility can be derived without the need for any numerical integration. Full details are given in the appendix.

3.4 Optimisation

4 Illustration

We consider the problem described in¹⁰, where a pilot trial is to be conducted prior to a planned main trial, with a shared normally distributed primary endpoint with known standard deviation of $\sigma = 1$ and a MCID of $\delta = 1$. The main trial is planned to have a sample size of $n_2 = 22$, giving OCs

Table 1. Caption.

Problem	n_1	n_2	α_1	$1 - \beta_1$	α_2	$1 - \beta_2$
Unrestricted	9	35	0.31	0.058	0.005	0.059
No pilot test	0	25	1.00	0.000	0.007	0.142
Conventional main trial OCs	10	21	0.22	0.071	0.025	0.099

$\alpha_2 = 0.025, 1 - \beta_2 = 0.9$. The true treatment effect has a normal prior with $\mu_0 = 0, \sigma_0 = 1$. Note that this represents an archetypal sceptical view of the treatment effect, and that it implies a prior belief that the effect will be greater than the MCID with probability 0.159.

To apply our method, we first set $\hat{d} = 0.605$ as this corresponds to the point of 50% power in the planned main trial. We suppose that the decision maker is risk averse, with $\rho = 1$. We further suppose that the cost of sampling leads to a choice of $\bar{d} = 0.05$ (i.e. a sample size of 100 would be considered justifiable in exchange for a known treatment effect of 0.05, absent any considerations of treatment cost).

We consider three optimisation problems. First, we optimise jointly over the pilot and main trial programme ('unrestricted'). Then, we optimise only the main trial whilst fixing $\alpha_1 = 1, \beta_1 = 0$ ('no pilot test'). Finally, we fix the main trial OCs at $\alpha_2 = 0.025, 1 - \beta_2 = 0.9$, as was done in¹⁰, and optimise over only the pilot trial OCs ('conventional main trial OCs'). The results are given in Table 1. We see that our model suggests a considerably larger sample size than the conventional $n_2 = 21$. When we allow for testing in the pilot, the optimal type I error rate is much larger than conventional levels, while the type I error in the subsequent main trial is substantially smaller. The algorithm takes around 1 second to converge to a solution in the general unrestricted case; for the special case of no pilot test, convergence takes around 0.003 seconds.

5 Evaluation

What do we want to show here? Obvious comparison to make is with the current no-testing approach. Could also look at Stallard's paper

6 Extensions

6.1 Internal pilots

Internal pilot trials are distinguished from external pilots by their data being used at the final analysis, with a seamless gap between the pilot and main trial stages. Extending our problem to the internal pilot setting, we continue to conduct a first test based on the pilot sample mean difference x_1 , but now follow this by a test based on the weighted combination of pilot and main trial samples. Specifically, the final analysis is of the overall sample mean difference

$$x_t = \frac{n_1 x_1}{n_1 + n_2} + \frac{n_2 x_2}{n_1 + n_2}.$$

We can now apply equation 2 in the internal pilot by defining $G_1 = x_1 > d_1$ and $G_2 = x_t > d_2$. The relevant probabilities can be calculated by noting that the pair x_1, x_t , conditional on μ , follow a bivariate normal distribution. Specifically, the appendix will show that

$$\begin{pmatrix} x_1 \\ x_t \end{pmatrix} \mid \mu \sim BN \left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \frac{2\sigma^2}{n_1} & \frac{2\sigma^2}{n_1+n_2} \\ \frac{2\sigma^2}{n_1+n_2} & \frac{2\sigma^2}{n_1+n_2} \end{pmatrix} \right).$$

The probabilities in equation 2 can now be calculated using, for example, the R package ‘mvtnorm’¹¹. Expected utility can then be calculated as before, integrating the conditional expected utility over the normal prior $p(\mu)$ using quadrature.

6.2 Binary endpoint

Suppose we have a binary primary endpoint, with probability of occurrence of τ_C in the control arm and τ_I in the intervention arm, and that we are interested in the absolute difference $\mu = \tau_C - \tau_I$ (interested in the sense that this is what features in our utility). We use Beta priors to reflect our prior uncertainty: $\tau_k \sim \text{Beta}(a_k, b_k)$ for $k = C, I$. We can extend our approach to this scenario by returning to equation 2 and noting that, now conditioning on the pair of probabilities (τ_C, τ_I) , the probabilities of test success/failure at each stage can still be easily calculated. Now, rather than integrating the conditional expectation over a single normal prior density, we must integrate over the two independent beta distributions for (τ_C, τ_I) . Unable to use Gauss-Hermite quadrature, we can instead use more general numerical integration methods such as the adaptive procedure implemented in the R package ‘cubature’¹². This is, however, slower by a factor of 10^3 , suggesting the time needed to solve the optimisation problem will be around 15 minutes.

6.3 Unknown variance

When the variance of the primary outcome σ^2 is unknown, we can extend the method as follows. Firstly, we note that the tests at each stage of the design will no longer be comparing sample means, but rather t-statistics incorporating a pooled estimate of the variance:

$$t_i = \frac{x_i}{\sqrt{2\hat{\sigma}^2/n_i}}.$$

Conditional on μ, σ^2 , these will follow non-central t-distributions with degrees of freedom $2n_i - 2$ and non-centrality parameter $\mu/\sqrt{2\sigma^2/n_i}$. As such, the relevant probabilities in equation 2 (now conditioning on μ and σ^2) can be readily calculated. Given this conditional utility, we can proceed as before by numerically integrating over the prior distribution, now a joint prior $p(\mu, \sigma^2) = p(\mu)p(\sigma^2)$ assuming μ and σ^2 are independent. As in the case of a binary endpoint, we can use a numerical method like the ‘cubature’ package to do the integration, noting that computation time will again be increased.

7 Discussion

Acknowledgements

Acknowledgements.

Declaration of conflicting interests

The Authors declare that there is no conflict of interest.

Funding

This work was supported by the Medical Research Council [grant number MR/N015444/1].

References

1. Craig P, Dieppe P, Macintyre S et al. Developing and evaluating complex interventions: the new medical research council guidance. *BMJ: British Medical Journal* 2008; 337. DOI:10.1136/bmj.a1655.
2. Eldridge SM, Lancaster GA, Campbell MJ et al. Defining feasibility and pilot studies in preparation for randomised controlled trials: Development of a conceptual framework. *PLOS ONE* 2016; 11(3): e0150205. DOI:10.1371/journal.pone.0150205. URL <http://dx.doi.org/10.1371/journal.pone.0150205>.
3. Lancaster GA, Dodd S and Williamson PR. Design and analysis of pilot studies: recommendations for good practice. *Journal of Evaluation in Clinical Practice* 2004; 10(2): 307–312. DOI:10.1111/j..2002.384.doc.x. URL <http://dx.doi.org/10.1111/j..2002.384.doc.x>.
4. Wilson DT, Walwyn RE, Brown J et al. Statistical challenges in assessing potential efficacy of complex interventions in pilot or feasibility studies. *Statistical Methods in Medical Research* 2015; 25(3): 997–1009. DOI:10.1177/0962280215589507. URL <http://dx.doi.org/10.1177/0962280215589507>.
5. Teare M, Dimairo M, Shephard N et al. Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: a simulation study. *Trials* 2014; 15(1): 264. DOI:10.1186/1745-6215-15-264. URL <http://www.trialsjournal.com/content/15/1/264>.
6. Lee E, Whitehead A, Jacques R et al. The statistical interpretation of pilot trials: should significance thresholds be reconsidered? *BMC Medical Research Methodology* 2014; 14(1): 41. DOI:10.1186/1471-2288-14-41. URL <http://www.biomedcentral.com/1471-2288/14/41>.
7. Cocks K and Torgerson DJ. Sample size calculations for pilot randomized trials: a confidence interval approach. *Journal of Clinical Epidemiology* 2013; 66(2): 197 – 201. DOI:<http://dx.doi.org/10.1016/j.jclinepi.2012.09.002>. URL <http://www.sciencedirect.com/science/article/pii/S0895435612002740>.
8. French S and Insua DR. *Statistical Decision Theory*. Number 9 in Kendall's Library of Statistics, Oxford University Press, 2000.
9. Blocker AW. *fastGHQuad: Fast 'Rcpp' Implementation of Gauss-Hermite Quadrature*, 2018. URL <https://CRAN.R-project.org/package=fastGHQuad>. R package version 1.0.

10. Stallard N. Optimal sample sizes for phase II clinical trials and pilot studies. *Statistics in Medicine* 2012; 31(11-12): 1031–1042. DOI:10.1002/sim.4357. URL <http://dx.doi.org/10.1002/sim.4357>.
11. Genz A, Bretz F, Miwa T et al. *mvtnorm: Multivariate Normal and t Distributions*, 2017. URL <https://CRAN.R-project.org/package=mvtnorm>. R package version 1.0-6.
12. Narasimhan B, Johnson SG, Hahn T et al. *cubature: Adaptive Multivariate Integration over Hypercubes*, 2018. URL <https://CRAN.R-project.org/package=cubature>. R package version 2.0.3.

Appendix

Expected utility for a single trial

The utility function is $u(\mu, x)$, where μ is the true difference and x is the sample mean difference. Note that

$$x \mid \mu \sim N\left(\mu, \frac{2\sigma^2}{n}\right).$$

Given a normal prior $\mu \sim N(\mu_0, \sigma_0^2)$, the posterior distribution of μ is also normal. Specifically

$$\mu \mid x \sim N\left(\mu_1 = \sigma_1^2 \left(\frac{\mu_0}{\sigma_0^2} + \frac{xn}{2\sigma^2}\right), \sigma_1^2 = 1 / \left(\frac{1}{\sigma_0^2} + \frac{n}{2\sigma^2}\right)\right).$$

The marginal distribution of the sample mean difference is

$$x \sim N\left(\mu_0, \sigma_x^2 = \sigma_0^2 + \frac{2\sigma^2}{n}\right).$$

For the case $\rho = 0$, the utility function is the value function. The expected utility is then

$$\begin{aligned} E[u(\mu, x)] &= E_x [E_{\mu|x}[u(\mu, x)]] \\ &= E_x [I\{x > d\}(k_d\mu_1 + k_n n) + I\{x \leq d\}(k_n n + k_c)] \\ &= k_n n + Pr(x > d)k_d E_{x|x>d}[\mu_1] + Pr(x \leq d)k_c, \end{aligned}$$

The two probabilities follow from the marginal distribution of the sample mean. For example,

$$Pr[x > d] = 1 - \Phi\left(\frac{d - \mu_0}{\sqrt{\sigma_0^2 + \frac{2\sigma^2}{n}}}\right)$$

The expectation is of μ_1 , the mean of the posterior distribution on μ given x , with respect to the distribution of x conditional on $x > d$. This is a truncated normal with mean

$$E_{x|x>d}[\mu_1] = \mu_0 + \sigma_x \phi\left(\frac{d - \mu_0}{\sigma_x}\right) / \left(1 - \Phi\left(\frac{d - \mu_0}{\sigma_x}\right)\right).$$

For the case $\rho > 0$, we again start with the law of total expectation giving $E[u(\mu, x)] = E_x [E_{\mu|x}[u(\mu, x)]]$. Then,

$$\begin{aligned} E_{\mu|x}[u(\mu, x)] &= E_{\mu|x}[1 - e^{-\rho(k_d\mu + k_n n)I\{x > d\} - \rho(k_n n + k_c)I\{x < d\}}] \\ &= I\{x > d\}E_{\mu|x}[1 - e^{-\rho(k_d\mu + k_n n)}] + I\{x < d\}E_{\mu|x}[1 - e^{-\rho(k_n n + k_c)}] \end{aligned}$$

For the first term we have:

$$E_{\mu|x}[1 - e^{-\rho(k_d\mu + k_n n)}] = 1 - e^{-\rho k_n n} E_{\mu|x}[e^{-\rho k_d \mu}]$$

This takes the form as a moment generating function: for a normally distributed $Y \sim N(a, b^2)$,

$$E[e^{tY}] = e^{ta + \frac{b^2 t^2}{2}}.$$

This gives us

$$E_{\mu|x}[1 - e^{-\rho(k_d\mu + k_n n)}] = 1 - e^{-\rho k_n n} \left(e^{t\mu_1 + \frac{\sigma_1^2 t^2}{2}} \right),$$

where $t = \rho k_d$. This then all gives an expected utility conditional on x of

$$E_{\mu|x}[u(\mu, x)] = I\{x > d\} \left[1 - e^{-\rho k_n n} \left(e^{t\mu_1 + \frac{\sigma_1^2 t^2}{2}} \right) \right] + I\{x < d\} \left[1 - e^{-\rho(k_n n + k_c)} \right].$$

We now need to take the expectation of this over the marginal distribution for x :

$$\begin{aligned} E[u(\mu, x)] &= E_x [E_{\mu|x}[u(\mu, x)]] \\ &= E_x \left[I\{x > d\} \left(1 - e^{-\rho k_n n} e^{t\sigma_1^2 \left(\frac{\mu_0}{\sigma_0^2} + \frac{x n}{2\sigma^2} \right) + \frac{\sigma_1^2 t^2}{2}} \right) \right] + E_x \left[I\{x < d\} \left(1 - e^{-\rho(k_n n + k_c)} \right) \right]. \end{aligned}$$

The first term is equal to

$$Pr[x > d] \left(1 - e^{-\rho k_n n} e^{\frac{\sigma_1^2 t^2}{2}} e^{\frac{t\sigma_1^2 \mu_0}{\sigma_0^2}} \right) E_x \left[e^{\frac{t\sigma_1^2 x n}{2\sigma^2}} \right].$$

For the expectation, we again use a moment generating function. Here, the sample difference x is normally distributed $N(\mu_0, \sigma_x^2)$ but we have conditioned on $x > d$ and so have a truncated normal. The moment generating function for $Y \sim N(a, b^2), Y > d$ is

$$E[e^{rY}] = \left(e^{ar + \frac{b^2 r^2}{2}} \right) \left(\frac{1 - \Phi\left(\frac{d-a}{b} - br\right)}{1 - \Phi\left(\frac{d-a}{b}\right)} \right)$$

So, for our case we have

$$E_x \left[e^{\frac{t\sigma_1^2 x n}{2\sigma^2}} \right] = \left(e^{\mu_0 r + \frac{\sigma_x^2 r^2}{2}} \right) \left(\frac{1 - \Phi\left(\frac{d-\mu_0}{\sigma_x} - \sigma_x r\right)}{1 - \Phi\left(\frac{d-\mu_0}{\sigma_x}\right)} \right),$$

where $r = \frac{t\sigma_1^2 n}{2\sigma^2}$. Finally, putting everything together, we get

$$\begin{aligned} E[u(\mu, x)] &= E_x [E_{\mu|x}[u(\mu, x)]] \\ &= \left[1 - \Phi \left(\frac{d - \mu_0}{\sqrt{\sigma_0^2 + \frac{2\sigma^2}{n}}} \right) \right] \left[1 - e^{-\rho k_n n} e^{\frac{\sigma_1^2 t^2}{2}} e^{\frac{t\sigma_1^2 \mu_0}{\sigma_0^2}} \left(e^{\mu_0 r + \frac{\sigma_x^2 r^2}{2}} \right) \left(\frac{1 - \Phi(\frac{d - \mu_0}{\sigma_x} - \sigma_x r)}{1 - \Phi(\frac{d - \mu_0}{\sigma_x})} \right) \right] + \\ &\quad \Phi \left(\frac{d - \mu_0}{\sqrt{\sigma_0^2 + \frac{2\sigma^2}{n}}} \right) \left[1 - e^{-\rho(k_n n + k_c)} \right]. \end{aligned}$$

The case $\rho < 0$ follows immediately.

Internal pilot sampling distribution

Recall that we have sample means from both the first and second stage of an internal pilot design, x_1, x_2 , with the combined sample mean $x_t = \frac{n_1 x_1}{n_1 + n_2} + \frac{n_2 x_2}{n_1 + n_2}$ being used at the final testing stage. The distribution of the combined mean, conditional on μ is $x_t \mid \mu \sim N\left(\mu, \frac{2\sigma^2}{n_1 + n_2}\right)$. To fully specify the joint distribution of (x_1, x_t) we require the covariance:

$$\text{cov}(x_1, x_t) = E[x_1 x_t] - E[x_1]E[x_t].$$

By the law of total expectation,

$$\begin{aligned} E[x_1 x_t] &= E_{x_1} (E_{x_1 x_t | x_1} [x_1 x_t]) \\ &= E_{x_1} (x_1 E_{x_t | x_1} [x_t]) \\ &= E_{x_1} \left(x_1 E_{x_t | x_1} \left[\frac{n_1 x_1}{n_1 + n_2} + \frac{n_2 x_2}{n_1 + n_2} \right] \right) \\ &= E_{x_1} \left(\frac{n_1 x_1^2}{n_1 + n_2} + \frac{n_2 x_1 \mu}{n_1 + n_2} \right) \\ &= \frac{n_1}{n_1 + n_2} E[x_1^2] + \frac{n_2}{n_1 + n_2} E[x_1] \mu \\ &= \frac{n_1}{n_1 + n_2} (\text{var}(x_1) + E[x_1]^2) + \frac{n_2}{n_1 + n_2} \mu^2 \\ &= \frac{n_1}{n_1 + n_2} \left(\frac{2\sigma^2}{n_1} + \mu^2 \right) + \frac{n_2}{n_1 + n_2} \mu^2. \end{aligned}$$

Then,

$$\begin{aligned} \text{cov}(x_1, x_t) &= \frac{n_1}{n_1 + n_2} \left(\frac{2\sigma^2}{n_1} + \mu^2 \right) + \frac{n_2}{n_1 + n_2} \mu^2 - \mu^2 \\ &= \frac{2\sigma^2}{n_1 + n_2} + \frac{n_1 \mu^2}{n_1 + n_2} + \frac{n_2 \mu^2}{n_1 + n_2} - \mu^2 \\ &= \frac{2\sigma^2}{n_1 + n_2}. \end{aligned}$$