Journal Title
XX(X):2–13
©The Author(s) 0000
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

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

Duncan T. Wilson[1]

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 interventions[1], 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 scale[2]. 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 years[3], 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 interest[4]. 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 in[5], 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

[1] Leeds 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 explicitly[6] and implicitly[7], 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 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. 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. The main trial will only proceed if the result of the pilot test is statistically significant. The design and analysis of each trail will follow the standard N-P hypothesis testing formulation, and as such can be encapsulated by type I and II error rates α_1, β_1 (for the pilot) and α_2, β_2 (for the main trial).

We will consider two optimisation problems. Firstly, we jointly optimise over $\alpha_1, \beta_1, \alpha_2, \beta_2$. In order to provide a comparison with the conventional approach of not testing at all in pilots, we also optimise over only α_2, β_2 and fix $\alpha_1 = 1, \beta_1 = 0$.

3 Maximising expected utility

3.1 Attributes and value

We construct a utility function by first defining a *value* function. A value function $v(y)$ should encode our preferences on some attributes y under conditions of certainty, in the sense that $v(y) > v(v')$ if and only if we prefer y to y'. We first need to define the attributes we are interested in and the values they can take. We consider the following:

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

We use the notation $y \prec y'$ to mean we prefer y to y'. We want to define a 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 at. Then, our preferences can be characterised by a value function of the form

$$v(n, C, \delta) = k_n v_n(n) + k_c v_c(C) + k_d v_d(\mu),$$

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 now need to define the component value functions v_n, v_c, v_d and the corresponding weights k_n, k_c, k_d. First consider the functions. 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,

$$v(n, C, \delta) = k_n n + k_c C + k_d \mu.$$

We need to choose the scaling constants which will determine the relative importance of each attribute. Given that $k_n + k_c + k_d = 1$ (because the scale of the value function is arbitrary), we can deduce the parameter values given two sets of equivalence judgements. First, we ask for the change in outcome $\bar{d}$ that would be needed for us to judge that

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

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

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

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

- $k_d = 1/(1 + \hat{d} + \bar{d}/n_*)$,
- $k_n = -\lambda_d \bar{d}/n_*$,
- $k_c = \lambda_d \hat{d}$.

In our example we will take $n_* = 100, \bar{d} = 0.01$, meaning that we would be happy to "pay" a sample size of $n = 100$ for a guaranteed increase in the mean outcome of 0.01. We take $\hat{d} = 0.142$, meaning that if we knew the new treatment would improve the mean outcome by 0.142 then we would be indifferent as to whether or not it should be recommended over the control as the standard treatment, bearing in mind the costs associated with implementing such a change [*].

3.2 Utility

The value function represents our preferences under conditions of certainty, but in reality we are uncertain about both whether the current standard treatment will be changed (i.e. if the null hypothesis is rejected) and associated change in outcome. We therefore need to define a utility function to encode our preferences for distributions defined on our attributes, which will allow us to calculate expected utility and use this to guide our decisions.

To deduce 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:

1. $u(n, d, C) = 1 - e^{-\rho v(n,d,C)}, \rho > 0$,
2. $u(n, d, C) = v(n, d, C), \rho = 0$,
3. $u(n, d, C) = -1 + e^{-\rho v(n,d,C)}, \rho > 0$.

That is, the utility function over the scaler attribute V must be of the exponential form, which implies constant absolute absolute risk aversion. The coefficient of absolute risk aversion is

$$\frac{-u''(v)}{u'(v)} = \rho.$$

Assessment of u then just requires ρ. Because d is utility independent of n and C, we can think about gambles on d whilst ignoring the value of the other attributes. Doing so, we find the point d^* which is the certainty equivalent to a gamble which will give $d = 0$ and $d = 1.5$ each with probability 0.5. If

$$d^* = 0.5 * 0 + 0.5 * 1.5 = 0.75$$

[*] The value 0.142 was in this case arrived at based on the original design which gave a power of 0.8 to detect a difference of 0.2 with a one-sided type I error rate of 0.025. This corresponds to a power of 0.5 at the point $\mu \approx 0.142$, suggesting that this is the true point of indifference or equipoise. This point is described by [8] as "the only non-arbitrary point on the power curve".

then $\rho = 0$ and the utility function is additive. If the left-hand side is is greater then $\rho > 0$ (risk averse), and if the right hand side is greater, $\rho < 0$ (risk neutral). Either way, we can determine ρ by setting the utilities equal and solving

$$u(n, d^*, C) = 0.5u(n, 0, C) + 0.5u(n, 1.5, C).$$

3.3 Expected utility

Denote pilot and main sample size by n_1, n_2 and respective sample means x_1, x_2. The critical values against which these are compared are d_1, d_2. The utility is now

$$\begin{aligned} u(\mu, x_1, x_2) = 1 - exp(&- \rho(k_d\mu + k_n(n_d 1 + n_2))I[x_1 > d_1, x_2 > d_2] \\ &- \rho(k_n(n_1 + n_2) + k_c)I[x_1 > d_1, x_2 < d_2] \\ &- \rho(k_n n_1 + k_c)I[x_1 < d_1]). \end{aligned} \tag{1}$$

It has not been possible to obtain a closed-form expression for the expected utility over the joint distribution of μ, x_1, x_2. However, the expected utility conditional on μ is

$$\begin{aligned} E_{x_1,x_2,|\mu}[u(\mu, x_1, x_2)] =& Pr[x_1 > d_1, x_2 > d_2 \mid \mu] \left(1 - e^{\rho(k_d v_d(\mu) + k_n v_n(n_1+n_2)}\right) + \\ & Pr[x_1 > d_1, x_2 < d_2 \mid \mu] \left(1 - e^{-\rho(k_n v_n(n_1+n_2)+k_c)}\right) + \\ & Pr[x_1 < d_1 \mid \mu] \left(1 - e^{-\rho(k_n v_n(n_1)+k_c)}\right). \end{aligned} \tag{2}$$

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:

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

where $p(\mu)$ is the normal prior density. Because we are integrating with a normal density weighting function, we can use Gauss-Hermite quadrature (as implemented in the 'fastGHQuad' R package[9]) 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.

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

3.4 Optimisation

4 Illustration

We consider the problem described in[10], 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 $\alpha_2 = 0.025, 1 - \beta_2 = 0.9$. 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 in[10], 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

6 Extensions

6.1 Internal pilots

For internal pilot trials, the pilot data will be used together with the subsequent main trial data in the final assessment of effectiveness. As such, an internal pilot will typically lead to higher power than an external pilot, and may therefore be preferred whenever there is no need for a clear gap between the pilot and main trials. The method described in this paper can be adapted for the internal pilot case. First, note that the sample mean used in the final test is the weighted sum of the sample means x_1, x_2 of the pilot and main trials:

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

The appendix will show that the sample means to be tested now follow a bivariate normal distribution (conditional on μ):

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

We can now proceed as before, replacing the x_2, d_2 in equation 2 with x_t, d_t and evaluating the probabilities $Pr[x_1 > d_1, x_t > d_t \mid \mu]$ and $Pr[x_1 > d_1, x_t < d_t \mid \mu]$ with respect to the above bivariate normal, using the R package 'mvtnorm'[11].

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. We are interested in the absolute difference in these probabilities, $\tau_C - \tau_I$. We use Beta priors to reflect our prior uncertainty: $\tau_k \sim Beta(a_k, b_k)$ 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 intervalÂ 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. Willan AR and Pinto EM. The value of information and optimal clinical trial design. *Statistics in Medicine* 2005; 24(12): 1791–1806. DOI:10.1002/sim.2069. URL `http://dx.doi.org/10.1002/sim.2069`.

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

Denote the sample mean in the trial by x, and the critical value in the test by d. The utility function is denoted by $u(\mu, x)$. For the case $\rho = 0$, the utility function is the value function. The expected utility is then

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

The two probabilities follow from the marginal distribution of the sample mean, which is normal with mean μ_0 and variance $\sigma_0^2 + \frac{2\sigma^2}{n}$. So, 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:

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

We are then left with calculating the expected value of this posterior mean with respect to the distribution of x conditional on $x > d$. This is a truncated normal with mean

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

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

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

Since the second term doesn't include μ, we can take the expectation operator out. For the first term we have:

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

The expectation here is over the posterior distribution for μ given x, which (assuming σ^2 is known) is normal by conjugacy. Specifically

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

Note that this follows from the normal likelihood for the sample difference, $x|\mu \sim N(\mu, 2\sigma^2/n)$. It 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:

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

The first term is equal to

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

For the expectation, we again use a moment generating function. Here, the sample difference x is normally distributed $N(\mu_0, \sigma_x^2 = \sigma^2 + \frac{2\sigma^2}{n})$ 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

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

So, for our case we have

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

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

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

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 $x_t|\mu\ N\left(\mu, \frac{2\sigma^2}{n_1+n_2}\right)$. To fully specify the joint distribution of (x_1, x_t) we require the covariance:

$$cov(x_1, x_t) = E[x_1 x_t] - E[x_1]E[x_t].$$

By the law of total expectation,

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

Then,

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