

Comparing two treatments by decision theory

Nicholas T. Longford,
SNTL and Imperial College, London, United Kingdom

Abstract

Decision theory is applied to the general problem of comparing two treatments in an experiment with subjects assigned to the treatments at random. The inferential agenda covers collection of evidence about superiority, non-inferiority and average bioequivalence of the treatments. The proposed approach requires defining the terms ‘small’ and ‘large’ to qualify the magnitude of the treatment effect, and specifying the losses (or loss functions) that quantify the consequences of the incorrect conclusions. We argue that any analysis that ignores these two inputs is deficient, and so is any ad hoc way of taking them into account. Sample size calculation for studies intended to be analysed by this approach is also discussed.

Keywords: *Clinical trial, hypothesis test, loss function, treatment effect, verdict.*

1 Introduction

Hypothesis test, [1], is one of the great statistical inventions. Its special case, test of a null-hypothesis, is commonly applied as a means of comparing two treatments for a medical or some other condition. Hypothesis testing has come under a lot of criticism lately because its application has over the years become rather ossified and ritualised, [2] and [3]. Our negative view of it is summarised by the following statement: as a method, the hypothesis test is oblivious to the consequences of the two kinds of error (bad decision) that can be made, [4]. For other arguments, in the context of research in psychology, see [5]. We present a unified decision-theoretical framework for analysing a class of clinical trials that are at present dealt with almost exclusively by hypothesis testing.

Throughout, we assume that a novel treatment 2 is compared to an established (standard or placebo) treatment 1 in an experiment with random assignment of the recruited or selected subjects to treatments. We assume that the outcomes of such a study are, within the treatment groups $g = 1$ and 2, independent random samples from distributions with expectations μ_g and common variance σ^2 . These distributions are often assumed to be normal; this assumption is not central to our development.

The key hypothesis we consider is that the within-treatment expectations coincide, or equivalently, that the average treatment effect, defined as $\Delta = \mu_2 - \mu_1$, vanishes. In principle, Δ may attain any value in an interval that has zero in its interior, and there is no prior information that would justify a prejudice for $\Delta = 0$ over some other values of Δ . In particular, $|\Delta|$ may be quite small, and such a value of Δ belongs to the alternative. It would be singularly unwise to bet on the hypothesis $\Delta = 0$; simply there are too many alternative plausible values to the solitary zero.

The outcome of a test of the hypothesis $H_0 : \Delta = 0$ is either its rejection, which is interpreted as evidence that Δ differs from zero, or failure to reject the hypothesis, which corresponds to a state of ignorance. Rejection of H_0 is often taken for a justification to continue the research, business, production or some other agenda or activity as if Δ were substantially different from zero. This is well supported if the error rate, controlled by the study design, is sufficiently small and the qualifier ‘substantially’ is well defined. The rate of 5% (the level of significance) is adopted almost universally, rarely weighing its role in the planned use of this analysis.

Failure to reject H_0 is commonly regarded as a licence to pursue the outlined agenda

as if Δ were equal to zero, a gross logical inconsistency. The weaker conclusion that $|\Delta|$ is small (insubstantially different from zero) has no foundation either. This follows directly from the construction of a null-hypothesis test. The analyst is asked to adopt the posture of H_0 , that is, to provisionally presume its validity. A test statistic s and a rule for what amounts to an exceptional outcome (the rejection region) are defined. A common example of such a rule is $|s| > C$ for a suitable constant C . The presumed validity of H_0 is regarded as implausible if the realised value of s is exceptional. Such a value of s is interpreted as evidence against H_0 , as a statistical contradiction with H_0 . If the value of s is not exceptional, then H_0 is plausible, since no contradiction with it has been found. But the alternative, denoted H_A , is also plausible — no contradiction with H_A has been found either. In fact, by arranging a sufficient power for a specified alternative, the test is constructed with a provision for seeking such a contradiction only for some ‘large’ values of Δ .

In our perspective, the choice between $H_0 : \Delta = 0$ and a general alternative H_A of a continuum of values ($\Delta \neq 0$) is always for H_A , assuming that $\sigma^2 > 0$ and the null-distribution of s also has positive variance; $\text{var}(s; H_A) > 0$. Data-based evidence for H_0 is impossible to collect in this case, and choosing any one particular value (such as zero) from an uncountable set may be appropriate only when that specific value is supported by some prior information. In this context, a small value, such as 10^{-6} , is *equally* different from zero as a much greater value, say, 14.3.

Some experiments to compare two treatments, and clinical trials in particular, are expensive in financial terms and are associated with high ethical costs. The possibility of a conclusion that we have learned nothing (state of ignorance) is clearly unsustainable. In most cases, failure to reject H_0 is regarded as a failure of the study and the course of action selected following the analysis is the one that would be appropriate if H_0 were valid. This is sometimes justified by a reference to the design (setting the sample size), arguing that $|\Delta|$ is bound to be small. The key point in the design, to which this reference is made, is the power at a designated value Δ_0 of Δ . The power is usually prescribed to be at least 0.80 or 0.90. Could we afford the error rate of 10 or 20% if Δ had the value of Δ_0 ? This depends on the context, but this context rarely informs the design in a direct and effective way.

In the method we propose, two courses of action, A and B, are contemplated. The courses are complementary — one has to be chosen, and exclusive — it is impossible

to apply both of them. Action A is appropriate when the treatment effect is zero, and also when it is small (insubstantial). Action B is appropriate when the treatment effect is substantial. The task is to identify the appropriate action. The method requires two elements of input in addition to the data:

- the range (interval) of values of the treatment effect Δ for which action A, the same as for $\Delta = 0$, would be appropriate; action B is appropriate for the complement of values of the treatment effect;
- a numerical description of the consequences of the errors (losses or loss functions) of the two kinds that may be committed.

For the first element, we specify constants $\Delta_- < 0$ and $\Delta_+ > 0$ so that all values of Δ in the range (Δ_-, Δ_+) are small, and all values in its complement are large. This suggests testing the hypothesis that $\Delta \in (\Delta_-, \Delta_+)$ against the alternative $\Delta \notin (\Delta_-, \Delta_+)$. See [6] for a related proposal. It is not clear what should be declared as the hypothesis and the alternative. Both $\Delta \in (\Delta_-, \Delta_+)$ and its complement could be adopted as the hypothesis. In the context of average bioequivalence, the hypothesis is $\Delta \notin (\Delta_-, \Delta_+)$, because evidence is sought for its complement.

The values Δ_- and Δ_+ are intended as *sharp* limits. That is, even $\Delta = \Delta_- - 10^{-6}$ and $\Delta_+ + 10^{-6}$ belong to the alternative, despite very low power of the hypothesis that $\Delta \in (\Delta_-, \Delta_+)$ at these values. Informally, the two values would then be assigned to the hypothesis, just as $\Delta = \pm 10^{-6}$ is in the test of $H_0: \Delta = 0$. With a hypothesis test, there is a gray area to the left of Δ_- and to the right of Δ_+ , where failure to reject H_0 would be regarded as appropriate. This makes the limits Δ_- and Δ_+ *soft*. In contrast, in our perspective, Δ_- and Δ_+ are uncompromising limits for what is regarded as the hypothesis (failure) and the alternative (success).

The next section gives a concise background to decision theory; more detailed treatment of this topic can be found in [4] and [7] – [9]. The section introduces losses that characterise the consequences of the two kinds of error. Section 3 applies the theory to clinical trials for comparing two treatments, with some simplifying assumptions. Section 4 extends the theory to loss functions that are not constant. Section 5 extends the results to t -distributed (test) statistics. Section 6 outlines the Bayesian version of our proposal. Section 7 deals with design of clinical trials. A set of examples is analysed

in Section 8. The concluding section summarises our objections to hypothesis testing and contrasts them with the strengths of our proposal.

2 Decision theory

We consider two states, A and B, and two courses of action, a and b. The purpose of the analysis is to choose one of these actions. Neither choosing both actions nor choosing none of them is possible. The analysis concludes with the selected action, referred to as the *verdict* (of the analysis). The verdict-state pairs aA and bB correspond to correct choices and bA and aB to incorrect choices (errors). We associate the latter two kinds of error with losses that quantify the damage, harm and other (negative) consequences (ramifications). Suppose error bA (type I) is associated with the loss of L_{bA} units and error aB (type II) with the loss of L_{aB} units; see Table 1. The correct verdicts, L_{aA} and L_{bB} , are associated with no loss. In other words, L_{aB} is the loss (outlay, expense, or the like) with verdict a in excess of what would be incurred with the correct verdict b, *if* state B prevailed; L_{bA} is interpreted similarly. For the unit of loss we use the term *lossile*. In any realistic setting, both L_{bA} and L_{aB} are positive. We emphasise that they are losses *relative* to the correct verdict. We may define positive values of L_{aA} and L_{bB} ; then meaningful values of L_{aB} and L_{bA} are greater than L_{bB} and L_{aA} , respectively. Instead of losses, gains may be considered. Then we may set $G_{aB} = G_{bA} = 0$ for the gains with discordant verdict-state pairs, and set G_{aA} and G_{bB} to positive values. In what follows, we consider losses only, with $L_{aA} = L_{bB} = 0$.

TABLE 1

As a characteristic of the verdict, the loss has to have two key properties: additivity and linearity. They mean that incurring k times loss L_{bA} and h times loss L_{aB} is for all purposes equivalent to any other compound loss which adds up to $kL_{bA} + hL_{aB}$. That is, the lossile can be treated like a monetary currency, with which we intend to be frugal. The nature of the problem is not changed by scaling the units of loss so that $L_{aB} = 1$. In fact, all that matters is the loss ratio $R = L_{bA}/L_{aB}$.

We prefer to incur no loss, but that is impossible in the presence of uncertainty about the average treatment effect Δ . Therefore, we seek to minimise the expected loss. In the frequentist perspective, the effect Δ is fixed, and then the expected loss depends on Δ . We remove this dependence by applying the *fiducial argument*, [10]

and [11]. Suppose $\hat{\Delta}$ is an unbiased estimator of Δ . Prior to conducting the study, $\hat{\Delta}$ is a random variable. Once the value of $\hat{\Delta}$ is realised, we regard it as fixed, and also change the status of Δ from fixed to random. This amounts to inverting the equation $\hat{\Delta} = \Delta + \delta$, where δ is the error of estimation, to $\Delta = \hat{\Delta} - \delta$. When the original version of $\hat{\Delta}$ is normally distributed, $\hat{\Delta} \sim \mathcal{N}(\Delta, \tau^2)$, the fiducial argument yields the distributional identity $\Delta \sim \mathcal{N}(\hat{\Delta}, \tau^2)$. By studying its Bayesian equivalent, [12] showed that such inversion is appropriate when $\hat{\Delta}$ is normally distributed.

In studies with nontrivial sample sizes, we can justify the inversion by appealing to the central limit theorem for approximate normality of an estimator $\hat{\Delta}$ or Δ , especially when even the observations y have within-treatment distributions that are close to normal or their averages converge to normality quite rapidly with increasing sample size (e.g., binary samples with probabilities distant from zero or unity).

3 Clinical trials

In this section, we discuss three common types of clinical trials, classified by their purpose, or inferential agenda, to provide evidence of superiority, non-inferiority and average bioequivalence, in designs in which each subject can receive at most one treatment.

3.1 Superiority

In the established perspective, superiority corresponds to $\Delta > 0$. With a definition of ‘small’, it corresponds to $\Delta > \Delta_+$. So, we are in error if we conclude that $\Delta \leq \Delta_+$, even though $\Delta > \Delta_+$, or if we conclude that $\Delta > \Delta_+$, even though $\Delta \leq \Delta_+$. Any reasonable rule for choosing between the two options has the form $\hat{\Delta} \leq c$ for lack of superiority (verdict a) and $\hat{\Delta} > c$ for superiority (verdict b). We set the borderline c by minimising the expected loss.

Denote by $\phi(x; \mu, \tau)$ the density and by $\Phi(x; \mu, \tau)$ the distribution function of $\mathcal{N}(\mu, \tau^2)$. When $\mu = 0$ and $\tau = 1$, these two arguments of ϕ and Φ are dropped. In particular, $\phi(x; \hat{\Delta}, \tau) = \phi\{(x - \hat{\Delta})/\tau\}/\tau$. With verdict b, the expectation of the loss we incur is

$$\begin{aligned} Q_{\text{bA}} &= L_{\text{bA}} \int_{-\infty}^{\Delta_+} \phi(x; \hat{\Delta}, \tau) \, dx \\ &= L_{\text{bA}} \Phi(\Delta_+; \hat{\Delta}, \tau) = L_{\text{bA}} \Phi(z_+) , \end{aligned}$$

where $z_+ = (\Delta_+ - \hat{\Delta})/\tau$; we assume that τ is known. In practice, τ is estimated; this case is dealt with in Section 5. With verdict a, the expected loss is

$$Q_{aB} = L_{aB} \int_{\Delta_+}^{+\infty} \phi(x; \hat{\Delta}, \tau) dx = L_{aB} \{1 - \Phi(z_+)\}.$$

The expected loss Q_{bA} is a decreasing function of $\hat{\Delta}$, with respective limits $L_{bA} > 0$ and zero as $\hat{\Delta}$ diverges to $-\infty$ and $+\infty$. The expected loss Q_{aB} increases in $\hat{\Delta}$ from zero to L_{aB} . Therefore the *balance* equation $Q_{bA} = Q_{aB}$ always has a unique solution; in fact, it is,

$$c^* = \Delta_+ - \tau \Phi^{-1} \left(\frac{L_{aB}}{L_{aB} + L_{bA}} \right) = \Delta_+ - \tau \Phi^{-1} \left(\frac{1}{R+1} \right). \quad (1)$$

The corresponding value of $Q_{aB} = Q_{bA}$ is $Q^* = L_{aB} R / (R+1) = L_{bA} / (R+1)$. For $\hat{\Delta} < c^*$, $Q_{aB} < Q_{bA}$, so verdict a is appropriate. Otherwise b is preferred. Figure 1 presents an illustration for a particular setting. The established rule based on the relevant hypothesis test corresponds to $\Delta_+ = 0$ and $R = 19$ for $H_A: \Delta > 0$.

FIGURE 1

The relevance of Δ_+ and R to choosing between actions a and b is obvious from expression (1). For a given R and very large Δ_+ , when the standard for superiority ($\Delta > \Delta_+$) is extremely high, we would always choose action a. For given Δ_+ and R sufficiently small, we would always choose action b because the consequences of choosing a inappropriately are dire in comparison. Conversely, for sufficiently large R , verdict a is the wise choice because verdict b may be associated with enormous loss.

These extreme cases highlight the need to specify Δ_+ and R . In practice, this is a rather contentious exercise, but some extreme values of Δ_+ and R can always be ruled out. In a constructive approach, plausible ranges of values are specified for Δ_+ and R by elicitation from the client (or sponsor of the study). A very wide range for Δ_+ , such as $(0.1, 50)$ in a particular context, is specified first so that all values outside this range are indisputably implausible. The analyst and the client then negotiate over narrowing this range by gradually raising its lower limit or reducing the upper limit. The exercise is concluded when the client is no longer willing to change either limit. In this way, the plausible range is set as narrow as possible, while any value outside the range can be ruled out. Let the plausible ranges be $(\Delta_{+L}, \Delta_{+U})$ for Δ_+ and (R_L, R_U) for R .

When the value of $\hat{\Delta}$ is established we find the borderlines $c^* = c^*(\Delta_+, R)$, the solutions of equation (1), for Δ_+ and R in their respective plausible ranges. Suppose these plausible values of c^* are in the range (c_L^*, c_U^*) . The borderline c^* is an increasing function of both Δ_+ and R ; therefore $c_L^* = c^*(\Delta_{+L}, R_L)$ and $c_U^* = c^*(\Delta_{+U}, R_U)$. If $\hat{\Delta} \notin (c_L^*, c_U^*)$, then we have an unequivocal verdict a (when $\hat{\Delta} < c_L^*$) or verdict b (when $\hat{\Delta} > c_U^*$). Otherwise, when $\hat{\Delta} \in (c_L^*, c_U^*)$, we arrive at an *impasse* (ambiguity); for some plausible pairs Δ_+ and R one verdict and for other pairs the other verdict is preferable. This shows how important is elicitation of the values of Δ_+ and R . On the one hand, integrity of the analysis is breached if the declared plausible ranges exclude some values of Δ_+ or R that should not be ruled out. On the other hand, the plausible ranges should be as narrow as possible, to reduce the ambiguity interval (c_L^*, c_U^*) , and with it the chances of an *impasse*. But *impasse* can be ruled out outright only when the plausible ranges degenerate into single values of Δ_+ and R . We stress that Δ_+ and R are not values that can be estimated; they are meant to reflect the perspectives, priorities and values (judgement) of the client.

In clinical trials for superiority, R may be quite large because the consequences of permitting a poor treatment to be prescribed (or made available for purchase) are much more grave than rejecting a good treatment; $L_{bA} \gg L_{aB}$. The loss ratio characterises the consumers' (public's) position, which is represented by the regulatory agency. It is not constant across medical conditions, and its value may be revised over time, in response to new treatments and changes in the societal values and priorities. Setting the value of Δ_+ should be guided by clinical considerations — it is the mandatory improvement that the novel treatment should exhibit over its established competitor or placebo. Smaller improvements are regarded as not useful.

3.2 Non-inferiority

Trials for non-inferiority seek to provide evidence that treatment 2 is no worse than treatment 1. This corresponds to the inequality $\Delta > \Delta_-$, so the rule given by (1) applies to this setting, with $\Delta_- < 0$ substituted for Δ_+ . The discussion of plausible values of Δ_- and R carries over directly.

Non-inferiority accommodates a slight negative effect of the novel treatment related to the established treatment. By setting the value of Δ_- , or its plausible range, the (undisputed) advantages of the new treatment, such as easier administration, more

convenient regimen, absence of a problematic compound and the like, are set against possibly (slightly) lower efficacy.

Increased value of R could be compensated by reducing the value of Δ_- . Such a tradeoff in setting the values of Δ_- (or Δ_+) and R should never be considered, because these parameters refer to distinct factors of the study; $\Delta_{\pm}$ divides the possible values of Δ to useful and not useful, and R quantifies the client's aversity to error bA relative to aB.

3.3 Average bioequivalence

Average bioequivalence corresponds to $\Delta \in (\Delta_-, \Delta_+)$. Therefore, the expected loss with verdict a (absence of bioequivalence) is

$$\begin{aligned} Q_{aB} &= L_{aB} \int_{\Delta_-}^{\Delta_+} \phi(x; \hat{\Delta}, \tau) dx \\ &= L_{aB} \{ \Phi(z_+) - \Phi(z_-) \} , \end{aligned}$$

where $z_- = (\Delta_- - \hat{\Delta})/\tau$. The expected loss with verdict b (bioequivalence) is

$$\begin{aligned} Q_{bA} &= L_{bA} \left\{ \int_{-\infty}^{\Delta_-} \phi(x; \hat{\Delta}, \tau) dx + \int_{\Delta_+}^{+\infty} \phi(x; \hat{\Delta}, \tau) dx \right\} \\ &= L_{bA} \{ 1 - \Phi(z_+) + \Phi(z_-) \} . \end{aligned}$$

The balance equation $Q_{aB} = Q_{bA}$, which defines the borderline between the two verdicts, is equivalent to

$$\Phi(z_+) - \Phi(z_-) = \frac{L_{bA}}{L_{aB} + L_{bA}} = \frac{R}{R + 1} . \quad (2)$$

Let $\Delta_0 = \frac{1}{2}(\Delta_+ + \Delta_-)$. In Appendix A, we show that the difference $\Delta\Phi = \Phi(z_+) - \Phi(z_-)$ is an increasing function of $\hat{\Delta}$ for $\hat{\Delta} < \Delta_0$ and decreasing for $\hat{\Delta} > \Delta_0$. Therefore equation (2) has up to two solutions. It has no solution when $\Delta\Phi < R/(R + 1)$ for all $\hat{\Delta}$, including $\hat{\Delta} = \Delta_0$. That is the case when

$$\frac{\Delta_+ - \Delta_-}{2\tau} < \Phi^{-1} \left(\frac{2R + 1}{2R + 2} \right) . \quad (3)$$

In this case, the study should not be conducted because verdict a is determined prior to its conduct. The ratio R is too high, the interval (Δ_-, Δ_+) in which bioequivalence applies is too narrow, or the sampling variance τ^2 is too large. Figure 2 presents an

illustration. For the setting given in its caption, the breakpoint value of R , for which there is a single solution, $c^* = \Delta_0 = 0$, is 79.5. It is obtained by solving equation (3) for R , with the inequality replaced by the equal sign.

FIGURE 2

Neither the limits $\Delta_{\pm}$ nor the ratio R should be adjusted to ensure that the study be meaningful to conduct. These three parameters are intended solely for the definition of bioequivalence. They specify the leeway around $\Delta = 0$ and characterize the aversity to the inappropriate verdict of bioequivalence (b). The variance τ^2 is reduced by conducting a larger study. The equivalence range (Δ_-, Δ_+) does not have to be symmetric, but there is rarely any rationale for setting $\Delta_- \neq -\Delta_+$.

4 Loss functions

The constant losses L_{aB} and L_{bA} can be replaced by functions of Δ , to reflect the dependence of the consequences on the magnitude of the error. A realistic loss function L_{bA} in a superiority trial is decreasing in the range $(-\infty, \Delta_+)$ because greater harm is caused by a more ineffective treatment. Similarly, loss L_{aB} may be set to an increasing function in $(\Delta_+, +\infty)$ when greater benefit would be derived from a more effective treatment.

Suitable and easy-to-interpret functions for L_{bA} are the linear loss $d_{bA}(\Delta_+ - \Delta)$ and quadratic loss $d_{bA}(\Delta_+ - \Delta)^2$, where d_{bA} is a positive constant and $\Delta < \Delta_+$. Their counterparts for L_{aB} are $d_{aB}(\Delta - \Delta_+)$ and $d_{aB}(\Delta - \Delta_+)^2$ for $\Delta > \Delta_+$. Let $\Phi^{(1)}(z)$ be the primitive function of $\Phi(z)$ and $\Phi^{(2)}(z)$ the primitive function of $2\Phi^{(1)}(z)$, such that their limits as $z \rightarrow -\infty$ are both equal to zero. It is easy to check by differentiation that $\Phi^{(1)}(z) = z\Phi(z) + \phi(z)$ and $\Phi^{(2)}(z) = (1 + z^2)\Phi(z) + z\phi(z)$, using the identity $\phi'(z) = -z\phi(z)$.

The expected loss with the linear loss function is

$$\begin{aligned} Q_{bA}^{(1)}(\hat{\Delta}) &= d_{bA} \int_{-\infty}^{\Delta_+} (\Delta_+ - x) \phi(x; \hat{\Delta}, \tau) dx \\ &= d_{bA} \tau \int_{-\infty}^{z_+} (z_+ - z) \phi(z) dz \\ &= d_{bA} \tau \int_{-\infty}^{z_+} \Phi(z) dz = d_{bA} \tau \Phi^{(1)}(z_+) , \end{aligned}$$

obtained by integration by parts. For the quadratic loss function, the expected loss is

$$\begin{aligned}
Q_{\text{bA}}^{(2)}(\hat{\Delta}) &= d_{\text{bA}} \int_{-\infty}^{\Delta_+} (\Delta_+ - x)^2 \phi(x; \hat{\Delta}, \tau) dx \\
&= d_{\text{bA}} \tau^2 \int_{-\infty}^{z_+} (z_+ - z)^2 \phi(z) dz \\
&= 2d_{\text{bA}} \tau^2 \int_{-\infty}^{z_+} (z_+ - z) \Phi(z) dz = d_{\text{bA}} \tau^2 \Phi^{(2)}(z_+) ,
\end{aligned}$$

after two-fold integration by parts. Note that both $Q_{\text{bA}}^{(1)}$ and $Q_{\text{bA}}^{(2)}$ are decreasing functions of $\hat{\Delta}$, converging to zero as $\hat{\Delta} \rightarrow +\infty$.

The expected loss following verdict a is derived similarly, exploiting the identities for the expectation and variance of the normal distribution;

$$\begin{aligned}
Q_{\text{aB}}^{(1)}(\hat{\Delta}) &= d_{\text{aB}} \tau \left\{ -z_+ + \Phi^{(1)}(z_+) \right\} \\
Q_{\text{aB}}^{(2)}(\hat{\Delta}) &= d_{\text{aB}} \tau^2 \left\{ 1 + z_+^2 - \Phi^{(2)}(z_+) \right\} .
\end{aligned}$$

These functions are increasing and converge to zero as $\hat{\Delta} \rightarrow -\infty$. The loss functions for the two kinds of error may be from different classes (e.g., one constant and the other quadratic). A pair of such expected loss functions intersects at a unique value c^* of $\hat{\Delta}$. For $\hat{\Delta} < c^*$ verdict a, and for $\hat{\Delta} > c^*$ verdict b is preferred.

In general, a loss function is non-negative and non-decreasing with $|\Delta - \Delta_+|$. It may have some discontinuities. The variety of the loss functions introduced, constant, linear and quadratic, can be enhanced by their linear combinations with positive coefficients. The quadratic loss function with $d_{\text{aB}} = d_{\text{bA}} = 1$ does not correspond to the loss implied by the mean squared error because $\hat{\Delta}$ is associated with no loss even when it differs a lot from Δ , so long as it leads to the correct verdict, when $\hat{\Delta}$ and Δ are both smaller or both greater than c^* .

For an increasing function f , the threshold Δ_- (or Δ_+) for Δ is equivalent to the threshold $f(\Delta_{\pm})$ for $f(\Delta)$. This reduces somewhat the onus on defining the estimator $\hat{\Delta}$. A sufficient statistic, when it exists, linearly transformed to be unbiased for $\hat{\Delta}$, is the obvious choice for $\hat{\Delta}$. In some cases, a nonlinear transformation of $\hat{\Delta}$ may make normality more palatable, but a bias correction should then be applied.

5 Distributions other than normal

Uncertainty about the sampling variance τ^2 can be addressed by declaring a plausible range for it. When $\hat{\Delta}$ has a noncentral t distribution and $\hat{\Delta}$ would be normally

distributed if τ^2 were known, this requires a small licence, because the support of the estimator of τ^2 is the entire interval $(0, +\infty)$. When available, prior information that reduces this support to a finite interval may be used. With a plausible range (τ_-^2, τ_+^2) , we have to find the borderline c^* only for the limits of this range, or for their combinations with the plausible values for the other parameters.

The fiducial argument is applied by inverting the identity $(\hat{\Delta} - \Delta)/\hat{\tau} = T$, where T is a random variable with t distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom. Inversion yields the identity $\Delta = \hat{\Delta} + \hat{\tau}T$, so all the evaluations of the expected losses carry over from Section 3, with the normal distribution replaced by the Student t . The counterparts of the functions $\Phi^{(1)}$ and $\Phi^{(2)}$ for the t distributions with ν degrees of freedom, denoted respectively by $\Psi_\nu^{(1)}$ and $\Psi_\nu^{(2)}$, are given in Appendix B. Note that the value of $\hat{\tau}^2$, the estimate of τ^2 , has to be available. The expressions in Section 3 are adapted for t distribution by substituting $\hat{\tau}$ for τ in the definitions of z_+ and z_- and Ψ_ν for Φ . In Section 4, $\Psi_\nu^{(1)}$ and $\Psi_\nu^{(2)}$ are substituted for $\Phi^{(1)}$ and $\Phi^{(2)}$, respectively.

6 The Bayesian paradigm

Arguably, decision theory is applied more naturally in the context of a Bayesian analysis, [7] and [9]. We have focused on the frequentist paradigm because it is more common in the analysis of clinical trials, especially when prior information is not permitted in their analysis. We can switch from the frequentist to the Bayesian paradigm simply by replacing the fiducial distribution in Section 3 by the posterior distribution of Δ given $\hat{\Delta}$. Conversely, application of the fiducial argument can be regarded as a frequentist way of switching the roles of Δ and $\hat{\Delta}$, a substitute for the Bayes' theorem. Prior information can be related to additional data, [4]. That renders the prior as important an input to the analysis as the data itself. However, using a noninformative prior, pretending that no information is available, erodes the potential of Bayesian methods because then they merely seek to reproduce the results of their frequentist analogues.

Formulating priors in clinical trials is often avoided because it would entail a difficult process of questionable integrity. However, no expensive study is conducted in an informational vacuum in which there is no prior information. In a common perspective, there is a strong prior information that the outcome will be rejection of the null-hypothesis, but the regulatory guidelines dictate that such information has to be ignored in the analysis. A consequence of this is objectivity, but it is accompanied by

a waste of resources. However, a prior distribution declared without integrity would distort the outcome of the analysis.

7 Study design

In the established perspective, the purpose of the study design is to ensure that the planned test of a hypothesis, with a prescribed test size, has sufficient power (for a given alternative $\Delta = \Delta_+$), or the adopted estimator has mean squared error smaller than a set threshold. In our perspective, a study is designed to balance the conflicting aims of spending least resources and making decisions with least expected loss. The expected loss Q^* vanishes when the correct verdict is issued in every (hypothetical) replication of the study, but the outlay for such a study may be prohibitive. We assume that the values of $\Delta_{\pm}$ and R are fixed, not affected by the design.

We consider a clinical trial for superiority. For a given value of Δ , the (conditional) expected loss is

$$\begin{aligned} Q_+(\Delta_+, R; \Delta) &= \int_{c^*}^{+\infty} \phi(\hat{\Delta}; \Delta, \tau) d\Delta = 1 - \Phi\left(\frac{c^* - \Delta}{\tau}\right) \\ &= 1 - \Phi\left\{\frac{\Delta_+ - \Delta}{\tau} - \Phi^{-1}\left(\frac{1}{R+1}\right)\right\} \\ Q_-(\Delta_+, R; \Delta) &= R \int_{-\infty}^{c^*} \phi(\hat{\Delta}; \Delta, \tau) d\Delta = R \Phi\left\{\frac{\Delta_+ - \Delta}{\tau} - \Phi^{-1}\left(\frac{1}{R+1}\right)\right\} \end{aligned} \quad (4)$$

when $\Delta < \Delta_+$ and $\Delta > \Delta_+$, respectively. Note that the probabilities involved in these expressions are not fiducial. In the Bayesian paradigm, they are considered as conditional on Δ , and the conditioning is removed by integrating over the prior distribution of Δ . In the frequentist perspective, we apply the same procedure, motivated as averaging over the plausible values of Δ . Let the prior distribution of Δ be given by density $p(\Delta)$. Then the a priori expected loss is

$$Q(\Delta_+, R) = \int_{-\infty}^{\Delta_+} Q_-(\Delta_+, R; \Delta) p(\Delta) d\Delta + \int_{\Delta_+}^{+\infty} Q_+(\Delta_+, R; \Delta) p(\Delta) d\Delta.$$

These integrals are approximated by numerical quadrature. There is no closed-form expression for them even when p is a normal density.

Care is required in the choice of the prior distribution p . A noninformative (dispersed) prior is rarely appropriate because it gives too much weight to values of Δ distant from Δ_+ , for which the choice between the two verdicts is associated with small expected loss Q_+ or Q_- . An example is given in Section 8.3.

The cost of the study, D , is usually an increasing function of the sample size n . The linear function is often appropriate; $D = C + C_1 n_1 + C_2 n_2$, where C is the cost of the set-up and management, C_1 and C_2 are the unit costs within the treatment groups and n_1 and n_2 are the sample sizes within the groups. The unit costs C_1 and C_2 are often the same, but ethical issues associated with one of the treatments and the lack of equipoise can be reflected by unequal values. The linear cost ignores any economies of scale. The sample sizes are often equal, $n_1 = n_2$, or the ratio n_2/n_1 is set in advance, so that the search for the optimal sample size is a univariate problem.

Usually D increases steeply as a function of the sample size(s), whereas Q decreases steeply at first but gradually more and more slowly with increasing sample sizes. Therefore there is a critical sample size at which the differentials $\partial D/\partial n$ and $\partial Q/\partial n$ cancel out. For smaller sample sizes, an increase of the sample size entails lower cost than the reduction of the expected loss; for greater sample sizes the reward for an increase of the sample size is negative. This critical size is the target of sample size calculation.

8 Examples

This section illustrates the proposed method on examples. Throughout, we assume that greater values of the outcome variable correspond to higher efficacy of the treatment.

8.1 Superiority

Suppose first that the estimate of the average treatment effect is $\hat{\Delta} = 2.34$, the associated standard error, established with precision, is $\tau = 1.00$, and the estimator $\hat{\Delta}$ is normally distributed. The conventional analysis for superiority of the novel treatment 2 over the reference treatment 1 would conclude by claiming evidence that the novel treatment is superior, based on the (one-sided) p value $1 - \Phi^{-1}(2.34) = 0.010$.

In the approach we propose, the values of the smallest effect that would be useful, Δ_+ , and the loss ratio R have to be specified, or plausible ranges given for them. Suppose $\Delta_+ = 0.2$ and $R = 25$. Then the borderline c^* between the two contemplated verdicts (a — no superiority, and b — superiority), given by equation (1), is 1.969, so verdict b is issued, in agreement with the conclusion based on the hypothesis test.

Suppose next that there is uncertainty about both Δ_+ and R , and the respective plausible ranges (0.1, 0.4) and (15, 50) are declared for them. The corresponding

plausible range for the borderline c^* is $(1.634, 2.462)$, and so the verdict is equivocal (ambiguous) because $\hat{\Delta}$ is within this range.

Figure 3 displays the borderline c^* as a function of R for a set of plausible values of Δ_+ . The strip delimited by the plausible borderline functions $c^*(R; \Delta_{+L})$ and $c^*(R; \Delta_{+U})$ for $R_L \leq R \leq R_U$ can be regarded as a gray zone. In our case, $\hat{\Delta}$ (horizontal dashes in the diagram) is within the plausible range of c^* for some plausible values of R , so the verdict is equivocal. When both Δ_+ and R are close to their respective upper bounds, 0.40 and 50, such as $\Delta_+ = 0.35$ and $R = 48$, verdict a applies ($c^* > \hat{\Delta}$), whereas verdict b applies, for instance, for all plausible values of Δ_+ when $R = 30$.

FIGURE 3

The ambiguity might have been avoided if narrower plausible ranges were declared. For example, if R_U were set to 40 instead of 50, and Δ_{+U} to 0.3 instead of 0.4 (dotted lines in the diagram), the gray zone, drawn by a darker shade of gray, would be entirely below the estimate, and the verdict of superiority would be unequivocal.

After an equivocal result, the analyst may turn to the sponsor of the analysis and ask for a revision (reduction) of the plausible ranges. Such a revision may be tainted by having been informed by the data. It is clearly preferable to conduct the elicitation exercise before data collection patiently and thoroughly to conclude with as narrow a range of plausible values as possible, while maintaining integrity — excluding from the plausible ranges only values that can be ruled out. The elicited values of Δ_+ and R are also useful for sample size calculation; see Section 8.3.

Suppose next that the estimator $\hat{\Delta}$ has t distribution with 38 degrees of freedom (e.g., based on $n_1 = n_2 = 20$) and $\hat{\tau} = 1.0$, but all the other details are unchanged. The p value is 0.012, so the conventional method would conclude with the claim of evidence of superiority of the novel treatment. For $\Delta_+ = 0.2$ and $R = 25$, the borderline is $c^* = 2.017$, leading to verdict b. For $\Delta_+ = 0.4$ and $R = 50$, $c^* = 2.536$, and therefore verdict a applies. The verdict with the plausible ranges $\Delta_+ \in (0.1, 0.4)$ and $R \in (15, 50)$ is equivocal, whereas with $(0.1, 0.3)$ and $(15, 40)$ it would be unequivocally b. The (numerical) results for t_{38} and normally distributed estimator $\hat{\Delta}$ differ only slightly, especially when related to the width of the plausible range of the borderline c^* .

8.2 Average bioequivalence

Suppose in a clinical trial for average bioequivalence $\hat{\Delta} = 0.72$ and $\tau = 1.00$, and bioequivalence corresponds to $\Delta_+ = -\Delta_- = 0.25$. We assume first that $\hat{\Delta}$ is normally distributed. The loss functions are $L_{aB} = 1$ when we incorrectly conclude that $\Delta \in (\Delta_-, \Delta_+)$, and

$$\begin{aligned} L_{bA} &= R \{1 + (\Delta - \Delta_+)^2\} \\ L_{bA} &= R \{1 + (\Delta_- - \Delta)^2\} \end{aligned}$$

when $\Delta > \Delta_+$ and $\Delta < \Delta_-$, respectively, but we conclude that $\Delta \in (\Delta_-, \Delta_+)$. The expected losses are

$$\begin{aligned} Q_{aB} &= \Phi(z_+) - \Phi(z_-) \\ Q_{bA} &= R\tau^2 \{1 + \Delta_+^2 + \Phi_2(z_-) - \Phi_2(z_+)\} . \end{aligned}$$

The balance equation $Q_{aB} = Q_{bA}$ is solved by the Newton-Raphson algorithm. It converges rapidly, requiring 5–8 iterations for any reasonable convergence criterion and a wide range of settings. For $\Delta_+ = 0.25$, $R = 10$ and $\tau = 1.0$, we obtain the solution $c^* = 0.598$, so the verdict is a, absence of bioequivalence, because $\hat{\Delta} \notin (-c^*, c^*)$.

Suppose next that the plausible ranges of Δ_+ and R are $(0.15, 0.30)$ and $(6, 15)$, respectively. Then the plausible range of the borderline is $(0.564, 0.639)$; the lower limit is obtained as the solution for $\Delta_+ = 0.15$ and $R = 15$, and the upper limit for $\Delta_+ = 0.30$ and $R = 6$. The solution c^* would be close to $\hat{\Delta}$ for $(R = 2.5, \Delta_+ = 0.30)$ or $(R = 6, \Delta_+ = 0.50)$.

If the estimator $\hat{\Delta}$ has t distribution with 38 degrees of freedom, and $\hat{\tau} = 1.00$, the results are altered only slightly. For example, the plausible range of the borderline c^* for $\Delta_+ \in (0.15, 0.30)$ and $R \in (6, 15)$ is $c^* \in (0.547, 0.621)$ and for $\Delta_+ = 0.25$ and $R = 10$ we obtain $c^* = 0.581$.

8.3 Sample size calculation

Suppose a study is planned for superiority of one treatment over another, with $\Delta_+ = 0.25$ and $R = 25$ and the variance of the outcomes within the treatment groups is $\sigma^2 = 1$. We consider two normal priors for the treatment effect Δ ; $\mathcal{N}(0.5, 0.2^2)$ and $\mathcal{N}(1, 0.5^2)$, referred to as priors 1 and 2, respectively. The projected costs of the study

are linear with $C = 1.5$ and $C_1 = C_2 = 0.015$. Figure 4 presents the loss functions Q (thin curves), the cost function D (linear), and the ‘total’ functions $Q + D$ for $n = n_1 + n_2$ ($n_1 = n_2$) up to 300 for the two prior distributions. At the optimal sample size, the total $Q + D$ attains its minimum. For the prior with variance 0.2^2 this is attained at $n = 104$, whereas for the prior variance 0.5^2 it is at $n = 172$, marked by vertical dashes.

FIGURE 4

The total $Q + D$ is very flat in the vicinity of its minimum. For example, minimum of the total for prior 1, marked ‘Q1 + D’ in the diagram, is 18.23 (at $n = 172$), but its value is less than 2% higher for $n \in (124, 248)$, marked in the diagram by vertical dots. Nevertheless, the optimal sample sizes for the two prior distributions differ substantially. The sample size does not depend on the baseline cost C , but the unit costs C_1 and C_2 are important factors. Lower unit costs result in greater sample size. For example, if the common unit cost $C_1 = C_2$ is reduced to 0.01, the optimal sample sizes are 244 and 140 for the respective priors 1 and 2; if the unit cost is increased to 0.020, the optimal sample sizes are 132 and 80.

The established method, insisting on power of 0.80 at Δ_+ , while maintaining the test size of 0.05, yields the sample size $n = 396$. The comparison with our proposal is skewed because the roles Δ_+ in the two approaches are not directly comparable. Ignoring the consequences of the two kinds of error, as well as the cost of the study, are distinct weaknesses of the established method. Smaller sample size is an accident of the example; with $C_1 = C_2 = 0.005$, the optimal sample sizes are 472 and 264 with the respective priors 1 and 2. Also, the optimal sample sizes are greater for priors with higher density in the vicinity of Δ_+ .

Sample size calculation is easily repeated for several plausible values of Δ_+ , R and other parameters, when plausible ranges for them are declared. The prospect of an impasse raises additional problems. To address them, a penalty (or a plausible range of penalties) for it has to be declared, to introduce a realistic incentive to reduce the likelihood of an impasse. See [13] for a solution based on simulations.

9 Discussion

Hypothesis testing has over the years suffered from a failure common to great inventions — its assumptions, rationale and caveats have been gradually forgotten or dismissed as a distraction. The scientific environment in which hypothesis tests are conducted has evolved over the years, but the practice of hypothesis testing has remained largely impervious to these changes, adhering to conventions that lack a profound basis.

Difficulties with and deficiencies of hypothesis testing have been addressed selectively. For example, control of the error rates in a set of tests (multiplicity) can be enforced by the Bonferroni correction and its improvements, e.g., [14] and [15], culminating in the simple rule introduced by [16]. The source of the difficulties is that probability, as used in hypothesis testing for the rates of error of the two kinds, has an intricate calculus. In contrast, loss and expectation are additive, so making several decisions raises no problems additional to making each decision separately. The difficulty is moved to specifying the loss (function) so that its scale is linear and additive, which is not a data-analytical activity.

The practice of statistics has in many areas been distorted by the selective reporting of results of hypothesis tests, reporting such results for every conceivable comparison, highlighting those that turned out nominally significant, comparing significances, [17], regarding the rejection of a hypothesis as a success and failure to reject as a failure of the study, and converting every inferential problem into a hypothesis test. Also, the outcome of a test is often interpreted uncritically as the choice between two courses of action, one appropriate for the hypothesis and the other for the alternative. Even though the error rates are controlled, they are rarely small enough to be ignored. In decision theory, the two courses of action are treated symmetrically in one respect — the roles of the two actions are interchangeable, and asymmetrically in another — the two kinds of error are associated with different losses or loss functions.

We conclude by restating our position that any comparison of two treatments (or conditions) is a matter of choosing between two options, and corresponding courses of action, and hypothesis testing is poorly suited for this task because it cannot take into account the consequences of the two kinds of error. Similar arguments against using hypothesis testing in studies of bioequivalence are presented by [18], but they are relevant in a much wider context. In the decision-theoretical approach presented in this paper, the consequences are an integral part of the input, together with a sharp

declaration of what constitutes the two options (small or large value of the average treatment effect). The focus on (expected) loss is much closer to the management perspective of dealing with limited resources for experimentation and converting them near-optimally to (imperfect) knowledge than is the focus on hypothetical probabilities. Control should be applied to the losses resulting from uncertainty, not to the error rates.

We have assumed throughout that the average treatment effect Δ is the parameter of interest. It represents the individual-level treatment effects, as defined in the Rubin's causal model, [19] and [20], imperfectly, presuming that these effects are constant (treatment homogeneity) or that their variation can be ignored. Problems arise when the individual-level treatment effects attain both positive and negative values. First, a treatment may be beneficial to a majority, but harmful to some. Second, the recruitment process may conspire to misrepresent the distribution of the individual-level treatment effects. Without some nontrivial and difficult-to-verify assumptions, there is no way of confirming treatment homogeneity empirically, but maybe subject-matter expertise can inform about the plausibility of insubstantial variation of the individual-level treatment effects.

References

- [1] Lehmann E, Romano J. *Testing Statistical Hypotheses* (3rd edn). Springer-Verlag: New York, 2005.
- [2] Gigerenzer G, Krauss S, Vitouch O. The null ritual. What you always wanted to know about significance testing but were afraid to ask. In *The Sage Handbook of Quantitative Methodology for the Social Sciences*, Kaplan D (ed). Sage: Thousand Oaks, CA, 2004; pp. 391–408.
- [3] Gelman A, Loken E. The statistical crisis in science. *American Scientist* 2014; **102**:460–465.
- [4] Longford NT. *Statistical Decision Theory*. Springer-Verlag: Heidelberg, Germany, 2013.
- [5] Trafimow D, Marks M. Editorial. *Basic and Applied Social Psychology* 2015; **36**:1–2.

- [6] Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics* 1987; **15**:657–680.
- [7] Berger JO. *Statistical Decision Theory and Bayesian Analysis* (2nd edn). Springer-Verlag: New York, 1985.
- [8] DeGroot MH. *Optimal Statistical Decisions* (2nd edn). McGraw-Hill: New York, 2004.
- [9] Lindley DV. *Making Decisions* (2nd edn). Wiley: Chichester, UK, 1985.
- [10] Fisher RA. *Statistics Methods and Scientific Inference*. Oliver and Boyd: Edinburgh, UK, 1956.
- [11] Seidenfeld T. R.A. Fisher’s fiducial argument and Bayes’ theorem. *Statistical Science* 1992; **7**:358–368.
- [12] Lindley DV. Fiducial distributions and Bayes’ theorem. *Journal of the Royal Statistical Society Series B* 1958; **20**:102–107.
- [13] Longford NT. Sample size calculation for comparing two normal random samples using equilibrium priors. *Communications in Statistics – Simulation and Computation* 2014; **43**:2149–2161.
- [14] Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 1979; **6**:65–70.
- [15] Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 1988; **75**:800–802.
- [16] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* 1995; **57**:289–300.
- [17] Gelman A, Stern H. The difference between “significant” and “not significant” is not itself statistically significant. *The American Statistician* 2006; **60**:328–331.
- [18] Lindley DV. Decision analysis and bioequivalence trials. *Statistical Science* 1998; **13**:136–141.

- [19] Holland PW. Statistics and causal inference. *Journal of the American Statistical Association* 1986; **81**:945–960.
- [20] Imbens GW, Rubin DB. *Causal Inference for Statistics, Social and Biomedical Sciences*. Cambridge University Press: New York, 2015.

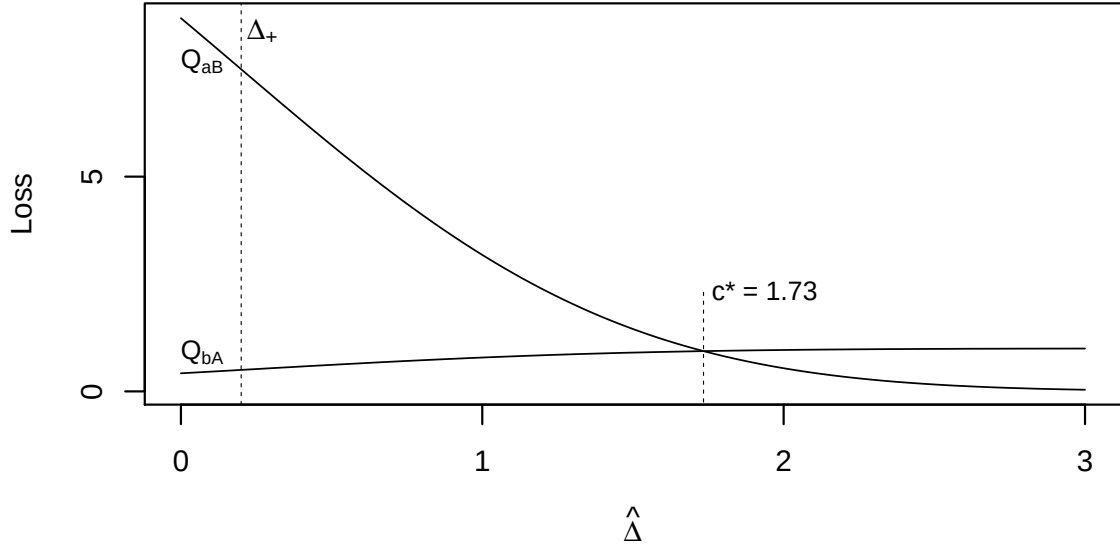


Figure 1: Deciding about superiority with minimum expected loss. The setting is: $L_{aB} = 15$, $L_{bA} = 1$, $\Delta_+ = 0.2$ and $\tau^2 = 1$. The solution is $c^* = 1.73$, with $Q^* = 0.94$.

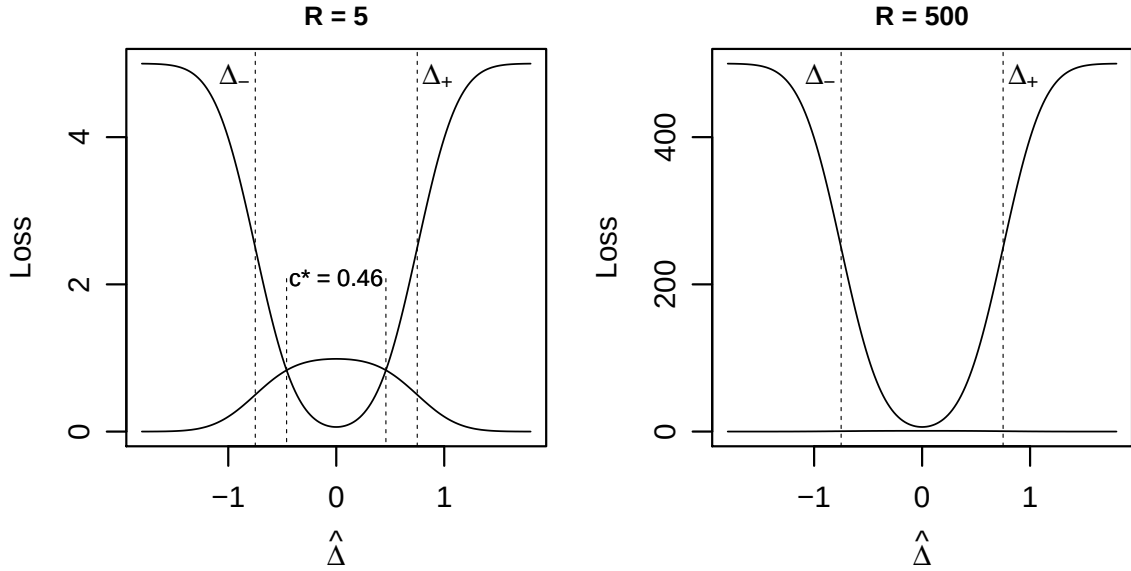


Figure 2: Inference about average bioequivalence with minimum expected loss. The setting is: $\Delta_- = -0.75$, $\Delta_+ = 0.75$, $\tau = 0.3$ and $L_{aB} = 1$ common to the two panels, and R ($= L_{bA}$) given in the title of each panel. The solutions in the left-hand panel are $c^* = \pm 0.460$, with $Q^* = 0.833$; there is no solution in the right-hand panel.

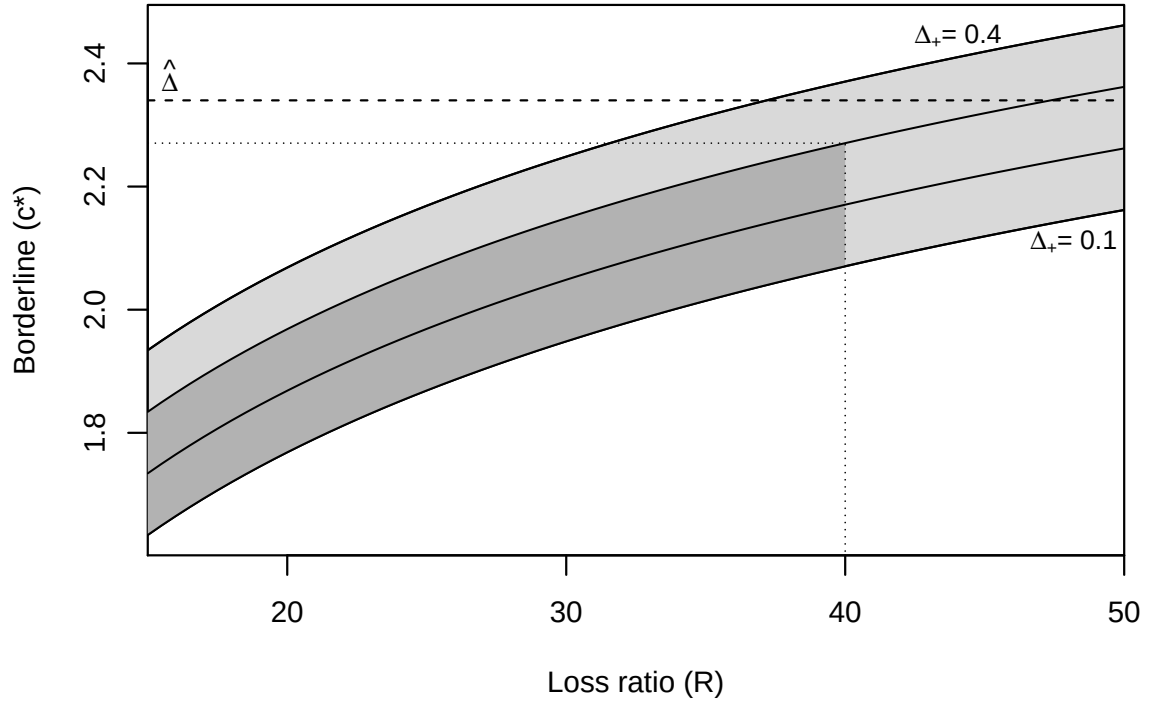


Figure 3: Borderline c^* as a function of R for $\Delta_+ = 0.1, 0.2, 0.3$ and 0.4 . The gray zone for the plausible ranges $\Delta_+ \in (0.1, 0.4)$ and $R \in (15, 50)$ is marked by gray color, and the narrower gray zone, for $\Delta_+ \in (0.1, 0.3)$ and $R \in (15, 40)$ by darker shade of gray. The estimate $\hat{\Delta}$ is marked by horizontal dashes.

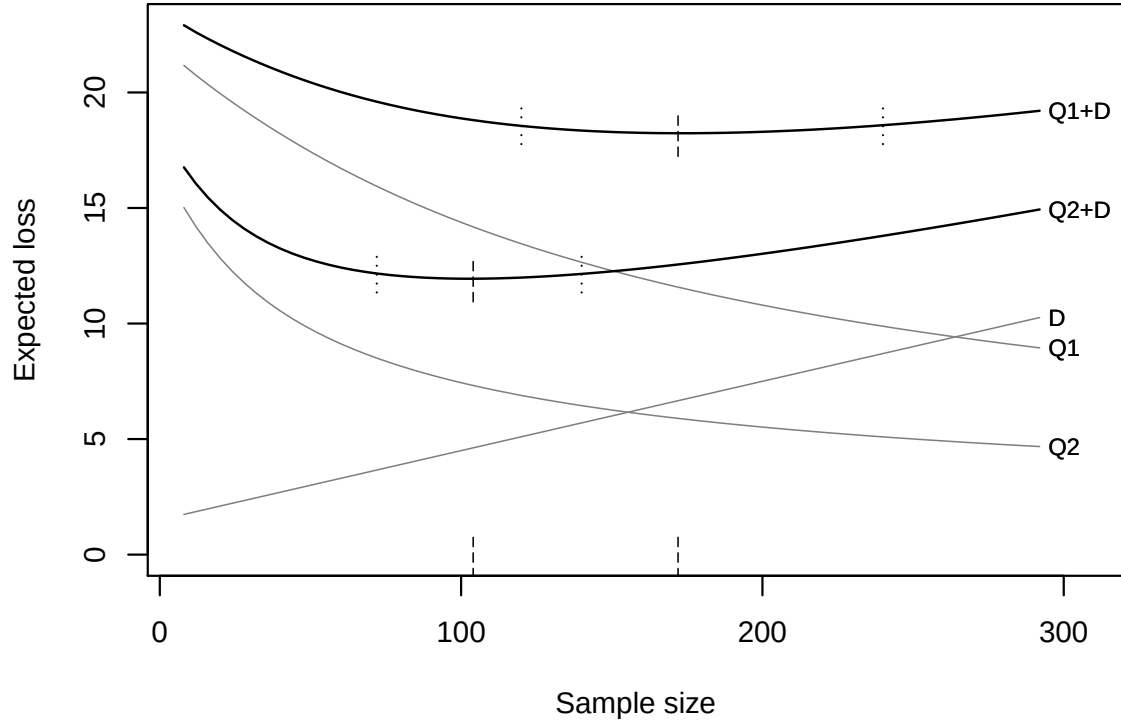


Figure 4: Sample size calculation by minimising the total $Q + D$ of the expected loss and cost of the study; normal priors $\mathcal{N}(0.5, 0.2^2)$ (prior 1) and $\mathcal{N}(1, 0.5^2)$ (prior 2). The minima are marked by vertical dashes and the ranges of indifference, where the total $Q + D$ is less than 2% greater are delimited by vertical dots.

Table 1: States, verdicts and losses.

Verdicts	States	
	A	B
a	0	L_{aB}
b	L_{bA}	0

Appendices

Appendix A

This appendix discusses solving equation (2). The difference $\Delta\Phi = \Phi(z_+) - \Phi(z_-)$ is positive throughout and converges to zero as $\hat{\Delta}$ diverges to $\pm\infty$. Its first-order differential,

$$\frac{\partial\Delta\Phi}{\partial\hat{\Delta}} = \frac{1}{\tau} \{\phi(z_-) - \phi(z_+)\} ,$$

is positive for $\hat{\Delta} < \Delta_0 = \frac{1}{2}(\Delta_+ + \Delta_-)$, and negative for $\hat{\Delta} > \Delta_0$, so $\Delta\Phi$ attains its unique maximum at Δ_0 . The difference $\Delta\Phi$ is smaller than $R/(R+1)$ throughout if it is smaller even at $\hat{\Delta} = \Delta_0$, where

$$\Delta\Phi_0 = \Phi\left(\frac{\Delta_+ + \Delta_-}{2\tau}\right) - \Phi\left(-\frac{\Delta_+ + \Delta_-}{2\tau}\right) = 2\Phi\left(\frac{\Delta_0}{\tau}\right) - 1.$$

Equation (2) has no solution when $\Delta\Phi_0 < R/(R+1)$, that is, when

$$\Phi\left(\frac{\Delta_0}{\tau}\right) < \frac{2R+1}{2R+2}.$$

The solutions of (2) are found by the Newton-Raphson algorithm. It converges rapidly in a wide range of settings; any reasonable convergence criterion is satisfied after 5–8 iterations.

With increasing R the solutions $c_-^* < \Delta_0$ and $c_+^* > \Delta_0$ converge to Δ_0 . When they exist, the solutions are symmetric around Δ_0 . Increased Δ_+ results in increased c_+^* and reduced Δ_- in reduced c_-^* .

Appendix B

The primitive function of the distribution function of the Student t distribution is derived in this appendix.

Denote by ψ_ν the density and by Ψ_ν the distribution function of the t distribution with ν degrees of freedom;

$$\psi_\nu(x) = \frac{\Gamma_2(\nu+1)}{\sqrt{\nu\pi}\Gamma_2(\nu)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

where Γ_2 is the half-gamma function, defined as $\Gamma_2(z) = \Gamma(\frac{1}{2}z)$. Denote $\gamma_\nu = \sqrt{1 - 2/\nu}$ and assume that $\nu > 4$. It can be checked by differentiation that a primitive function

for $x\psi_\nu(x)$ is $-\psi_{\nu-2}(\gamma_\nu x)/\gamma_\nu$. This implies that

$$\Psi_\nu^{(1)}(x) = x\Psi_\nu(x) + \frac{1}{\gamma_\nu} \psi_{\nu-2}(\gamma_\nu x)$$

is a primitive function for Ψ_ν . A primitive function for $2\Psi_\nu^{(1)}$ is

$$\Psi_\nu^{(2)}(x) = x^2\Psi_\nu(x) + \frac{1}{\gamma_\nu^2} \Psi_{\nu-2}(\gamma_\nu x) + \frac{x}{\gamma_\nu} \psi_{\nu-2}(\gamma_\nu x) .$$

The functions $\Psi_\nu^{(1)}(x)$ and $\Psi_\nu^{(2)}(x)$ converge to zero for $x \rightarrow -\infty$. As $\nu \rightarrow +\infty$, these functions converge pointwise to their respective normality counterparts, $\Phi^{(1)}$ and $\Phi^{(2)}$.