REGULAR ARTICLE

The distribution of power-related random variables (and their use in clinical trials)

Francesco Mariani[1] · **Fulvio De Santis**[1] · **Stefania Gubbiotti**[1]

Abstract

In the hybrid Bayesian-frequentist approach to hypotheses tests, the power function, i.e. the probability of rejecting the null hypothesis, is a random variable and a pre-experimental evaluation of the study is commonly carried out through the so-called probability of success (PoS). PoS is usually defined as the expected value of the random power that is not necessarily a well-representative summary of the entire distribution. Here, we consider the main definitions of PoS and investigate the power related random variables that induce them. We provide general expressions for their cumulative distribution and probability density functions, as well as closed-form expressions when the test statistic is, at least asymptotically, normal. The analysis of such distributions highlights discrepancies in the main definitions of PoS, leading us to prefer the one based on the utility function of the test. We illustrate our idea through an example and an application to clinical trials, which is a framework where PoS is commonly employed.

Keywords Assurance · Bayesian power · Bayes risk · Predictive power

Mathematics Subject Classification 35C05 · 65L10 · 65L12 · 65L20 · 65L70

1 Introduction

Evaluation of success is a crucial step of an experiment, especially at the design stage. Under the standard frequentist approach to testing, success is evaluated by computing the power function $\eta_n(\cdot)$, the probability of rejecting the null hypothesis, at a fixed

✉ Francesco Mariani
 f.mariani@uniroma1.it

 Fulvio De Santis
 fulvio.desantis@uniroma1.it

 Stefania Gubbiotti
 stefania.gubbiotti@uniroma1.it

1 Department of Statistical Sciences, Sapienza University of Rome, Piazzale Aldo Moro n. 5, 00185 Rome, Italy

🍂 Springer

(design) value of the parameter of interest. It is well known that this approach lacks flexibility and does not account for uncertainty on the design value.

Conversely, under the hybrid Bayesian-frequentist paradigm, the parameter of interest θ is considered as a random variable, Θ, with distribution $\pi(\theta)$, often called design prior (in most of the cases a density). Consequently, the power function $\eta_n(\Theta)$ is a random variable as well: we will refer to it as *random power* (but it is also known as *predictive power*, see Spiegelhalter et al. 2004). In most of the cases, the experimental success is evaluated by the expected value of the random power, commonly known as *probability of success* (PoS). However, the concept of PoS has been highly debated in the literature: Kunzmann et al. (2021) is a recent review paper that identifies even 17 different definitions of it. The computation of PoS is a routine component of study planning and decision-making (Liu and Yu 2024) especially in clinical trials, where it is employed for several reasons, such as, to support funding approval from sponsor governance boards (Crisp et al. 2018; Wang et al. 2013), to compute the optimal sample size at the design stage (Kunzmann et al. 2021) or at interim analysis (Wang 2007), and to choose the clinical development plan (Temple and Robertson 2021).

However, PoS is just the expected value of a random variable, whose distribution may not be well-represented by its mean; the perils of entrusting solely on averaging are pointed out, for example, in Liu and Yu (2024) and Dallow and Fina (2011). Starting from the seminal paper of Spiegelhalter et al. (1986), several authors suggested to complement PoS with other alternative summaries (such as the median or other quantiles, see for instance Huson 2009) or with the whole distribution that PoS summarizes (Huson 2009; Rufibach et al. 2016; De Santis and Gubbiotti 2023). The latter, in fact, provides an overall indication of the chance of success of an experiment: the basic idea is that a test is well-designed if the distribution of the random power induced by the design prior assigns high density to large values of the power (i.e. values as close to one as possible). The whole density function of $\eta_n(\Theta)$ was studied in Rufibach et al. (2016) with the objective of providing recommendations and guidelines on the design prior choice, but Kunzmann et al. (2021), Liu (2010), Dallow and Fina (2011) argued that by definition $\eta_n(\Theta)$ and its mean are misleading in explaining the success of the experiment, since they represent the probability of rejecting the null hypothesis regardless of it being true or not. Recently, in Kunzmann et al. (2021) and De Santis et al. (2024), it was shown that PoS has essentially four main specifications, each representing the expected value of a power-related random variable (PrRV) based on a suitable function of $\eta_n(\Theta)$. Since PrRVs have the potential to provide an overall indication of the success of an experiment, the goal of this paper is to investigate their distributions.

We find out that the main definitions of PoS and the respective PrRVs are closely interconnected and their discrepancies, albeit few, discern which definition accounts for success more properly. This analysis also provides guidelines on prior choices, methods to evaluate whether PoS is a well-representative summary, as well as alternative tools to obtain an overall pre-experimental evaluation of the designed experiment, such as cumulative distribution functions (cdfs), probability density functions (pdfs) and quantiles.

The outline of the article is as follows. We introduce notation, setup and literature review on the main definitions of PoS in Sect. 2. In Sect. 3 we provide general results

for the density functions of the PrRVs, whose closed-form expressions are drawn under normality assumption of the test statistic. In addition, to illustrate our ideas, we replicate an example of Spiegelhalter et al. (1986) in a clinical trial where the log-hazard ratio is the endpoint. We show how the qualitative features of the PrRV densities can be employed to set the design prior parameters. We also sketch a simulation algorithm useful when explicit expressions are not available. An application to a two sample confirmatory phase III trial is illustrated in Sect. 4, while we come to conclusions in Sect. 5.

2 Setup and notation

In this section we formally introduce power and utility functions of a test, whose definitions are needed to discuss PoS and its specifications. Consider the statistical model $(\mathcal{X}, f_n(\cdot, \theta), \theta \in \Omega)$ for which $\theta \in \Omega$, $\Omega \subseteq \mathbb{R}$, is a parameter of interest, $\mathcal{X}$ is the sample space, $f_n(\cdot|\theta)$ the probability mass or pdf of the random sample of independent and identical distributed (iid) random variables $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)$, and $\mathbf{x}_n$ an element of $\mathcal{X}$ (i.e. observed values of $\mathbf{X}_n$). We denote with $\mathbb{P}_\theta(\cdot)$ and $\mathbb{E}_\theta(\cdot)$ the probability measure and the expected value with respect to the sampling distribution of $f_n(\cdot|\theta)$, respectively. Given $\Omega_i \subset \Omega$, $i = 0, 1$ a partition of Ω, consider the following testing problem on θ:

$$H_0 : \theta \in \Omega_0 \quad \text{vs} \quad H_1 : \theta \in \Omega_1. \tag{1}$$

We here focus on the one-sided hypotheses $\Omega_0 = (-\infty, \theta_0]$ and $\Omega_1 = \Omega_0^C = (\theta_0, +\infty)$, but implementation of the reversed one-sided test is straightforward. Let $\mathcal{X}_0$ and $\mathcal{X}_1$ be the two elements of a partition of $\mathcal{X}$, representing the acceptance and the rejection regions of H_0, respectively. The power function of the test is formally defined as:

$$\eta_n(\theta) = \mathbb{P}_\theta(\mathbf{X}_n \in \mathcal{X}_1). \tag{2}$$

In order to define the utility function of a test we adopt a decision-theoretic perspective, under which a test statistic is a decision function $d_n(\mathbf{x}_n) = a_i \mathbb{I}_{\mathcal{X}_i}(\mathbf{x}_n)$, $i = 0, 1$, where in general $\mathbb{I}_A(\cdot)$ is the indicator function of the set A, and a_i means accepting H_i, $i = 0, 1$. The utility function evaluates the quality of the random decision $d_n(\mathbf{X}_n)$ and it is defined as:

$$U_n(\theta) = 1 - \mathbb{E}_\theta[\mathbb{L}(\theta, d_n(\mathbf{X}_n))],$$

where $\mathbb{L}(\theta, a_i) = \mathbb{I}_{\Omega_j}(\theta)$ is the $0 - 1$ loss function of a_i, $i = 0, 1$. Note that $\mathbb{E}_\theta[\mathbb{L}(\theta, d_n(\mathbf{X}_n))]$ is the risk function of the decision.

$U_n(\theta)$ and $\eta_n(\theta)$ are related in the following way:

$$U_n(\theta) = [1 - \eta_n(\theta)]\mathbb{I}_{\Omega_0}(\theta) + \eta_n(\theta)\mathbb{I}_{\Omega_1}(\theta). \tag{3}$$

A graphical representation of $\eta_n(\theta)$ and of $U_n(\theta)$ is provided on the top panels of Fig. 1. We comment on that in the next section.

Table 1 Main definitions of PoS with the corresponding PrRV, cdf and pdf

PoS definition	PrRV	cdf	pdf
$e_P = \mathbb{E}_\pi[\eta_n(\Theta)]$	$P_n = \eta_n(\Theta)$	$\mathbb{F}_P(\cdot)$	$f_P(\cdot)$
$e_J = \mathbb{P}_\pi(\mathbf{X}_n \in \mathcal{X}_1, \Theta \in \Omega_1)$	$J_n = \eta_n(\Theta)\mathbb{I}_{\Omega_1}(\Theta)$	$\mathbb{F}_J(\cdot)$	$f_J(\cdot)$
$e_C = \mathbb{P}_\pi(\mathbf{X}_n \in \mathcal{X}_1 \mid \Theta \in \Omega_1)$	$C_n = \eta_n(\Theta) \mid \Theta \in \Omega_1$	$\mathbb{F}_C(\cdot)$	$f_C(\cdot)$
$e_U = \mathbb{E}_\pi[U_n(\Theta)]$	$U_n = U_n(\Theta)$	$\mathbb{F}_U(\cdot)$	$f_U(\cdot)$

2.1 Main definitions of PoS

According to the hybrid Bayesian-frequentist approach, the unknown parameter of interest is a random variable Θ, thus power and utility functions are random variables as well.

We respectively denote with $\mathbb{P}_\pi(\cdot)$, $\mathbb{F}_\pi(\cdot)$ and $\mathbb{E}_\pi(\cdot)$ probability measure, cdf and expected value with respect to $\pi(\cdot)$, the design prior density of Θ.

As discussed in the Sect. 1, our starting point is the specification of the main definitions of PoS, $e_i = e_i(n, \pi)$, with $i \in \{P, J, C, U\}$, identified in Kunzmann et al. (2021), De Santis et al. (2024), which are the expected values of some PrRVs. Table 1 reports these definitions and notation.

The specifications of PoS and the related PrRVs are deeply dependent on the choice of the design prior, and they all coincide when the design prior assumes values only on the alternative hypothesis. As the design prior assigns positive probability masses to θ values in Ω_0, differences in the definitions of PoS are not negligible and in Sect. 3 we show that $e_U \geq e_C \geq e_P \geq e_J$.

To facilitate the understanding of how the PrRVs vary from each other, in Fig. 1 we plot $\eta_n(\theta)$, $\eta_n(\theta)\mathbb{I}_{\Omega_1}(\theta)$, $\eta_n(\theta) \mid \theta \in \Omega_1$ and $U_n(\theta)$ as functions of θ deterministic. We refer to the setting of an example discussed in Spiegelhalter et al. (1986) and further developed in Sect. 3.1.1. We now provide comments on the main definitions of PoS.

- The value e_P (where P stands for power) is mostly known with the name of *assurance* (O'Hagan and Stevens 2001; O'Hagan et al. 2005), and according to Kunzmann et al. (2021) this is the most used definition of PoS. It is the expected value of the random power P_n, according to which the experiment success is the rejection of H_0, no matter of its truth. In other words, P_n treats the occurrence of a type I error as a success; hence, its expected value averages the power with an error, which reduces the chance of success.
- The joint probability to reject H_0, e_J (with J standing for joint probability), is developed in Brown et al. (1987) and Ciarleglio et al. (2015). Similarly to the random power, the random variable J_n assumes that the test success is only related to the proper selection of H_1, but not to the proper rejection of it. Consequently, e_J averages the power values when $\theta \in \Omega_1$ and 0 when $\theta \in \Omega_0$ reducing the chance of success.
- The expected conditional power e_C (where C stands for conditional) is introduced in Spiegelhalter et al. (2004): the concept of success is only related to the proper rejection of the null hypothesis, thus conditional on H_1 being true. Consequently,
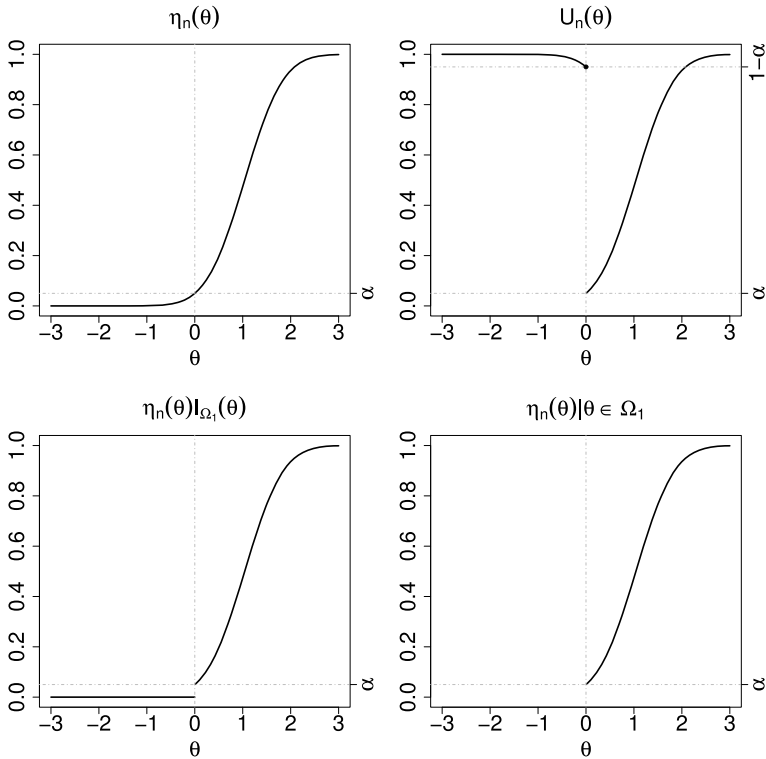
Fig. 1 Power and power-related functions of θ deterministic for $n = 10$ under the setting of the example in Sect. 3.1.1, which is when $X_i|\theta \sim N(\theta, \sigma^2)$ iid $i = 1, ..., n, \sigma^2 = 4$, $\alpha = 0.05$, $\theta_0 = 0$

the pdf $f_C(\cdot)$ is the density of the random power when Θ is restricted on the alternative hypothesis. Note that e_C can be also seen as $\mathbb{E}_\pi[\eta_n(\Theta)\mathbb{I}_{\Omega_1}(\Theta)/\pi_1]$, where $\pi_1 = \mathbb{P}_\pi(\Theta \in \Omega_1)$ and $\eta_n(\Theta)\mathbb{I}_{\Omega_1}(\Theta)/\pi_1$ is a linear transformation of J_n. However, the support of this random variable is $(\frac{\alpha}{\pi_1}, \frac{1}{\pi_1})$, where $\frac{1}{\pi_1} \geq 1$: this implies that its interpretation as an overall representation of the probability of success does not make sense. For this reason, we will consider only the definition of e_C as expected value of C_n provided in Table 1.

- The Bayes utility e_U (where U stands for utility) has been formally introduced in De Santis et al. (2024) as *uPoS*. Its advantages over e_i, $i \in \{P, J, C\}$, discussed in the aforementioned contribution, follow from the fact that the random variable U_n generalizes the concept of probability of success by considering successful the proper rejection of H_1 and of H_0.

As discussed by Spiegelhalter et al. (2004) and Kunzmann et al. (2021), since the objective of a test is the rejection of H_0, the design prior is usually almost fully concentrated on the alternative hypothesis, thus differences between e_i, $i \in \{P, J, C, U\}$ and their related distributions are negligible. Nonetheless there are cases where accounting for a design distribution which assigns not negligible probability masses to Ω_0 is inescapable: this happens for instance in Temple and Robertson (2021), where the

design prior is a mixture of two normals defined on Ω_0 and Ω_1, respectively, or in Liu and Yu (2024) and Dallow and Fina (2011) where, instead of a design prior, a data driven design posterior is used in the interim analysis of a clinical trial. In fact, in these cases, discrepancies between the PrRVs under study and consequently in results, can not be ignored.

3 The power related random variables

We here provide the general expressions of the pdfs $f_P(y)$ and $f_i(y), i \in \{J, C, U\}$ as functions of $f_P(y)$.

Theorem 3.1 *Given the hypotheses in* (1) *and a size-α test with monotone increasing power function $\eta_n(\theta)$, the density functions of the PrRVs in Table* 1 *are:*

$$f_P(y) = \pi\left(\eta_n^{-1}(y)\right)\left|\frac{d}{dy}\eta_n^{-1}(y)\right|\mathbb{I}_{[0,1]}(y)$$
$$f_J(y) = \delta_0(y) + f_P(y)\mathbb{I}_{(\alpha,1]}(y)$$
$$f_C(y) = \frac{f_P(y)}{\pi_1}\mathbb{I}_{(\alpha,1]}(y)$$
$$f_U(y) = f_P(y)\mathbb{I}_{(\alpha,1]}(y) + f_P(1-y)\mathbb{I}_{(1-\alpha,1]}(y)$$

where $\eta_n^{-1}(\cdot)$ is the inverse function of $\eta_n(\cdot)$ and $\delta_0(y)$ is the Dirac delta function at 0.

Proof We start the proof deriving the expression of the cdf of the random power, which support is [0, 1]. Since P_n is assumed monotone increasing, $\mathbb{F}_P(y)$ can be written as:

$$\mathbb{F}_P(y) = \mathbb{P}_\pi[P_n \le y] = \mathbb{P}_\pi[\eta_n(\Theta) \le y] = \mathbb{P}_\pi\left[\Theta \le \eta_n^{-1}(y)\right] =$$
$$= \mathbb{F}_\pi[\eta_n^{-1}(y)].$$

The pdf $f_P(y)$ is obtained by deriving $\mathbb{F}_P(y)$ wrt y. Recall that $\frac{d}{dx}f^{-1}(x) = \frac{1}{f'(f^{-1}(x))}$, where $f'(x) = \frac{d}{dx}f(x)$. We have:

$$f_P(y) = \frac{d}{dy}\mathbb{F}_P(y) = \pi\left(\eta_n^{-1}(y)\right)\left|\frac{d}{dy}\eta_n^{-1}(y)\right|.$$

From (3) it follows that the generic expression of the cdf of U_n is:

$$
\begin{aligned}
\mathbb{F}_U(y) = \mathbb{P}_\pi[U_n \leq y] &= \\
&= \mathbb{P}_\pi[U_n \leq y, \Theta \in \Omega_0] + \mathbb{P}_\pi[U_n \leq y, \Theta \in \Omega_1] = \\
&= \mathbb{P}_\pi[P_n \geq 1 - y, \Theta \in \Omega_0] + \mathbb{P}_\pi[P_n \leq y, \Theta \in \Omega_1] = \\
&= \mathbb{P}_\pi[\Theta \geq \eta_n^{-1}(1 - y), \Theta \leq \theta_0] + \mathbb{P}_\pi[\Theta \leq \eta_n^{-1}(y), \Theta > \theta_0] = \\
&= \begin{cases} 0 & \text{for } y \leq \alpha \\ \mathbb{F}_P(y) - \pi_0 & \text{for } \alpha < y \leq 1 - \alpha \\ \mathbb{F}_P(y) - \mathbb{F}_P(1 - y) & \text{for } 1 - \alpha < y \leq 1 \\ 1 & \text{for } y > 1 \end{cases}
\end{aligned}
$$

where $\pi_0 = \mathbb{F}_\pi(\theta_0) = \mathbb{P}_\pi(\Theta \in \Omega_0)$. Note that $\mathbb{F}_U(y)$ is differentiable for $y \neq 1 - \alpha$, hence $\mathbb{F}_U(y)$ admits density function $f_U(y)$, that is:

$$
\begin{aligned}
f_U(y) = \pi\left(\eta_n^{-1}(y)\right) \left| \frac{d}{dy}\eta_n^{-1}(y) \right| \mathbb{I}_{(\alpha,1]}(y) + \\
+ \pi\left(\eta_n^{-1}(1 - y)\right) \left| \frac{d}{dy}\eta_n^{-1}(1 - y) \right| \mathbb{I}_{(1-\alpha,1]}(y).
\end{aligned}
\tag{4}
$$

Since $\mathbb{F}_U(y)$ is not differentiable at $y = 1 - \alpha$, then $f_U(y)$ has a jump discontinuity at that point. Derivations of cdf and pdf expressions for J_n and C_n are available in Appendix A. $\qquad\square$

Corollary 3.1 *[Stochastic order of the PrRVs] Given the hypotheses in (1) and a size-α test with monotone increasing power function $\eta_n(\theta)$, the PrRVs in Table 1 are in the following stochastic order*

$$
U_n \succeq C_n \succeq P_n \succeq J_n,
$$

that is, for any $y \in \mathbb{R}$,

$$
\mathbb{F}_U(y) \leq \mathbb{F}_C(y) \leq \mathbb{F}_P(y) \leq \mathbb{F}_J(y).
$$

Proof The proof is provided in the Supplementary Material. $\qquad\square$

Remark 3.1 The stochastic order of the PrRVs implies that:

(i) $e_U \geq e_C \geq e_P \geq e_J$;
(ii) $q_U^\gamma \geq q_C^\gamma \geq q_P^\gamma \geq q_J^\gamma$, $\forall \gamma \in [0, 1]$, where $q_i^\gamma = q_i^\gamma(n, \pi)$, $i \in \{P, J, C, U\}$, is the γ level quantile of the PrRV.

3.1 Normal models

In this section, we apply Theorem 3.1 when the test statistic has normal distribution. Following Spiegelhalter et al. (2004), consider the statistical model of Sect. 2 and assume that T_n is the sufficient statistic and that, at least asymptotically, $T_n|\theta \sim N(\theta, \frac{\sigma^2}{n})$. Then, the size-$\alpha$ uniformly most powerful (UMP) test statistic is

$$W(T_n, \theta_0) = \frac{\sqrt{n}}{\sigma}(T_n - \theta_0) \sim N(0, 1) \text{ under } H_0. \tag{5}$$

The random power of the test is:

$$P_n = \eta_n(\Theta) = \Phi\left(\frac{\Theta - \theta_0 + \frac{\sigma}{\sqrt{n}}z_\alpha}{\frac{\sigma}{\sqrt{n}}}\right), \tag{6}$$

where $\Phi(\cdot), \phi(\cdot)$ and $z_\gamma = \Phi^{-1}(\gamma)$ are the cdf, pdf and γ level quantile of the standard normal. As it was shown in Rufibach et al. (2016),

$$f_P(y) = \pi\left(\theta_0 + \frac{\sigma}{\sqrt{n}}(z_y - z_\alpha)\right) \times \frac{\sigma}{\sqrt{n}}\sqrt{2\pi}\exp\left\{\frac{1}{2}z_y^2\right\}\mathbb{I}_{[0,1]}(y). \tag{7}$$

Details are available in Appendix B. From Theorem 3.1, the expressions of $f_i(y)$, $i \in \{J, C, U\}$ are straightforward. Following Rufibach et al. (2016), we use Eq. (7) and Theorem 3.1 to derive closed-form expressions of $f_U(y)$ under three typical prior choices: (*i*) normal, (*ii*) truncated normal, (*iii*) uniform.

(i) Normal design prior

Let $\Theta \sim N\left(\theta_d, \frac{\sigma^2}{n_d}\right)$, the pdf $f_U^N(y)$ is:

$$f_U^N(y) = f_P^N(y)\mathbb{I}_{(\alpha,1]}(y) + f_P^N(1-y)\mathbb{I}_{(1-\alpha,1]}(y) \tag{8}$$

where, as in Rufibach et al. (2016),

$$f_P^N(y) = \tau\phi(\Psi + \tau(z_y - z_\alpha))[\phi(z_y)]^{-1} =$$
$$= \tau\exp\left\{-\frac{1}{2}(\Psi + \tau(z_y - z_\alpha))^2 + \frac{1}{2}z_y^2\right\}, \tag{9}$$

where $\Psi = \frac{\theta_0 - \theta_d}{\frac{\sigma}{\sqrt{n_d}}}$ and $\tau = \sqrt{\frac{n_d}{n}}$.

(ii) Truncated normal design prior

Let $\Theta \sim TN\left(\inf = a, \sup = b, \theta_d, \frac{\sigma^2}{n_d}\right)$. The density $f_U^{TN}(y)$ is:

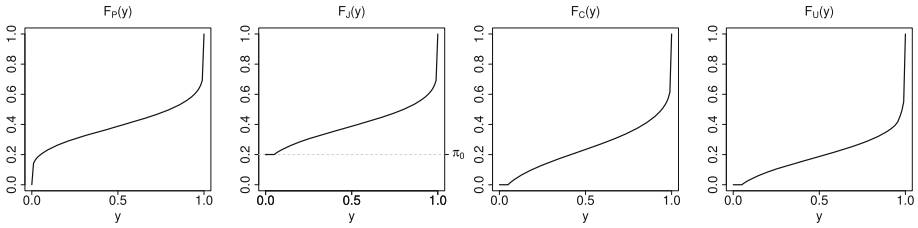$$f_U^{TN}(y) = f_P^{TN}(y)\mathbb{I}_{(\alpha,1]}(y) + f_P^{TN}(1-y)\mathbb{I}_{(1-\alpha,1]}(y)$$

Fig. 2 Cdfs of the PrRVs

where

$$f_P^{TN}(y) = \frac{f_P^N(y)}{\Phi\left(\frac{b-\theta_d}{\sigma/\sqrt{n_d}}\right) - \Phi\left(\frac{a-\theta_d}{\sigma/\sqrt{n_d}}\right)} \mathbb{I}_{[\eta_n(a),\eta_n(b)]}(y).$$

Specifically, if the design prior of Θ is a normal truncated on the alternative hypothesis $(\inf = \theta_0, \sup = \infty)$, then $f_P^{TN}(y) = \frac{f_P^N(y)}{1-\Phi(\theta_0)} I_{(\alpha,1]}(y)$.

(iii) Uniform design prior
Let $\Theta \sim Unif(a, b)$. The density $f_U^{UNIF}(y)$ is:

$$f_U^{UNIF}(y) = f_P^{UNIF}(y)\mathbb{I}_{(\alpha,1]}(y) + f_P^{UNIF}(1-y)\mathbb{I}_{(1-\alpha,1]}(y)$$

where

$$f_P^{UNIF}(y) = \frac{1}{b-a}\frac{\sigma}{\sqrt{n}}\sqrt{2\pi}\exp\left\{\frac{1}{2}z_y^2\right\}\mathbb{I}_{[\eta_n(a),\eta_n(b)]}(y).$$

3.1.1 Example on log-hazard ratio

As an example, we consider an application provided in Spiegelhalter et al. (1986), whose main objective is the design of a clinical trial to compare a new treatment to the standard one in terms of log-hazard ratio θ, with $\theta_0 = 0$, $\alpha = 0.05$ and known variance $\sigma^2 = 4$. The authors assume that a balanced trial is designed to have more than 0.8 of frequentist power at $\theta_d = 0.56$, which is reached for $n = 79$. We now consider a normal design prior $N(\theta_d, \sigma^2/n_d)$ where $n_d = 9$, which is the prior sample size needed to obtain a prior with 0.20 probability that θ is less than zero, thus $\pi_1 = 0.80$. In Figs. 2 and 3 we respectively provide a graphical representation of $\mathbb{F}_i(y)$ and of $f_i(y)$, $i \in \{P, J, C, U\}$. Visual inspection of the PrRVs cdfs confirms the result of Corollary 3.1. To complement cdfs and pdfs, we compute expectations e_i and quantiles q_i^γ, which are available in Table 2. The PrRVs C_n and U_n, in comparison with the other two, assign substantially higher densities to values of y close to 1, with medians equal to $q_C^{0.5} = 0.947$ and $q_U^{0.5} = 0.981$, while $q_P^{0.5} = 0.798$ and $q_J^{0.5} = 0.798$. As expected, U_n presents the highest density mass for high values of the power. Conversely, the cdfs and pdfs of P_n and J_n coincide for $y \in (\alpha, 1]$, but they differ for $y \in [0, \alpha]$: consequently, their γ quantiles coincide when $\gamma > \alpha$, but in terms of expected values, $e_J = 0.604$ is slightly lower than $e_P = 0.606$ as an implication of the fact that J_n is null in $[0, \alpha]$.
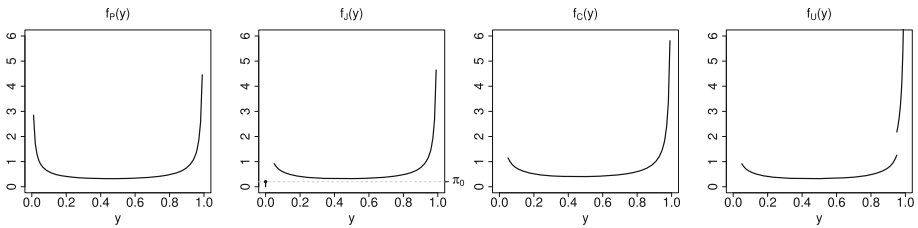
Fig. 3 Pdfs of the PrRVs

Table 2 Summaries of the PrRVs

i	e_i	$q_i^{0.25}$	$q_i^{0.5}$	$q_i^{0.75}$
P	0.606	0.123	0.798	1.000
J	0.604	0.123	0.798	1.000
C	0.758	0.545	0.947	1.000
U	0.803	0.680	0.981	1.000

3.1.2 Qualitative features of $f_U(y)$

As discussed in Sect. 3, the shape of the PrRVs pdfs indicates whether the experiment is well-designed: a large density for values close to 1 corresponds to high chances of conducting a well-designed experiment. The qualitative features of these density functions depend on the sample size and on the design prior, therefore they provide guidelines on the choice of the design prior parameters. We specifically focus on $f_U(y)$ assuming a normal design prior $\Theta \sim N(\theta_d, \sigma^2/n_d)$. From Eqs. (8) and (9), the qualitative features of $f_U(y)$ can be summarized as follows.

Result 3.1 *[Qualitative features of $f_U(y)$]*

- *For $y \in (\alpha, 1 - \alpha]$:*

    - *for $n_d = n$ ($\tau = 1$):*
        1. *$f_U(y)$ is strictly increasing for $\theta_d > \theta^\star$;*
        2. *$f_U(y)$ is constant for $\theta_d = \theta^\star$;*
        3. *$f_U(y)$ is strictly decreasing for $\theta_d < \theta^\star$;*
    - *for $n_d \neq n$ ($\tau \neq 1$):*
        1. *$f_U(y)$ is strictly increasing for $\theta_d > \theta^\star$;*
        2. *$f_U(y)$ is not strictly increasing for $\theta_d \leq \theta^\star$.*

- *For $y \in (1 - \alpha, 1]$:*

    - *$f_U(y)$ is strictly increasing for $\theta_d > \theta^\star$ when $n_d \leq n$ ($\tau \leq 1$),*

where $\theta^\star = \theta_0 + \kappa(n, n_d) \in \Omega_1$ and $\kappa(n, n_d) = \frac{\sigma}{\sqrt{n_d}} \max\{-z_\alpha; -2\sqrt{\frac{n_d}{n}} z_\alpha\}$.

Proof The proof is provided in the Supplementary Material. □

The interpretation as applied to the experiment success is that we get a well-shaped density of the utility, thus a well-designed experiment, when $\theta_d > \theta^\star$. Since $\theta^\star$ is a

decreasing function of n and n_d, the choice of the design prior parameters is a matter of finding the good trade-off between θ_d, n_d, n. For instance, for low values of n_d a higher value of θ_d is required in order to ensure a good shape of the density. Note also that the equation of $\theta^\star$ helps in the choice of the design value n or n_d or θ_d, when the other two are given. This is particularly useful for example in the design of clinical trials when the prior sample size n_d is the historical and/or external data sample size, thus it is known at the design stage of the trial. Similar considerations about the shape of the pdf of U_n hold for the truncated normal design prior case.

Example on log-hazard ratio (continued)   We consider once again the example of Sect. 3.1.1 and we illustrate the qualitative features of $f_U(y)$ in Fig. 4, bottom panels, for $n_d = 29, 79, 129$ and for several values of θ_d chosen to be symmetric around the $\theta^\star$ value. Moreover, for the same design values, we plot $f_P(y)$ in Fig. 4, top panels.

As discussed above, in general, for low values of n_d a higher value of θ_d is required to ensure a well-shaped pdf. We note that, according to findings in Rufibach et al. (2016), in order to ensure a well-shaped $f_P(y)$, the design values may need to satisfy hard-to-meet criteria. Conversely, the conditions that the design values need to satisfy to ensure a well-shaped $f_U(y)$ are milder. For instance, when $n_d = 29$ (thus for n_d realistically low) and $\theta_d = 0.51$, $f_U(y)$ is substantially increasing (except for a very small set of values of y) and $q_U^{0.25} = 0.46$, while $f_P(y)$ presents a bad-shape (u-shape) and $q_P^{0.25} = 0.31$. As θ_d is reduced to 0.41, things for $f_P(y)$ go worst since $q_P^{0.25} = 0.17$, while $q_U^{0.25} = 0.38$. Our point is that the u-shape of $f_P(y)$ is not solely a consequence of the design prior choices: as largely debated in Sect. 2, P_n takes values close to 0 ($< \alpha$) with non-negligible probability (π_0), leading to a u-shape of $f_P(y)$ and therefore a reduced quantification of success even for realistic design prior parameters choices. The employment of U_n avoids this scenario.

3.2 Simulation algorithm

When closed-form expressions are not available, the PrRVs distributions can be simulated. Let $w_{1-\alpha}$ be the generic $1 - \alpha$ level quantile of the test statistic $W(T_n, \theta_0)$. The algorithm works as follows.

1: Draw M values $\theta^{(1)}, \cdots, \theta^{(k)}, \cdots, \theta^{(M)}$ from $\pi(\cdot)$;
2: for each $\theta^{(k)}$, $k = 1, \ldots, M$, draw N values $T_{n,1}^{(k)}, \cdots, T_{n,i}^{(k)}, \cdots T_{n,N}^{(k)}$ from the sampling distribution of $T_n | \theta^{(k)}$;
3: for each $k = 1, \ldots, M$, compute $W_i^{(k)} = W(T_{n,i}^{(k)}, \theta_0)$, $i = 1, \cdots, N$;
4: obtain M simulated values $\eta_n(\theta^{(1)}), \cdots, \eta_n(\theta^{(k)}), \cdots, \eta_n(\theta^{(M)})$ from the empirical distribution of the random power by computing:

$$\eta_n(\theta^{(k)}) = \frac{\sum_{i=1}^{N} \mathbb{I}(W_i^{(k)} > w_{1-\alpha})}{N}, \quad k = 1, \cdots, M;$$

5: obtain the empirical PrRV of interest:

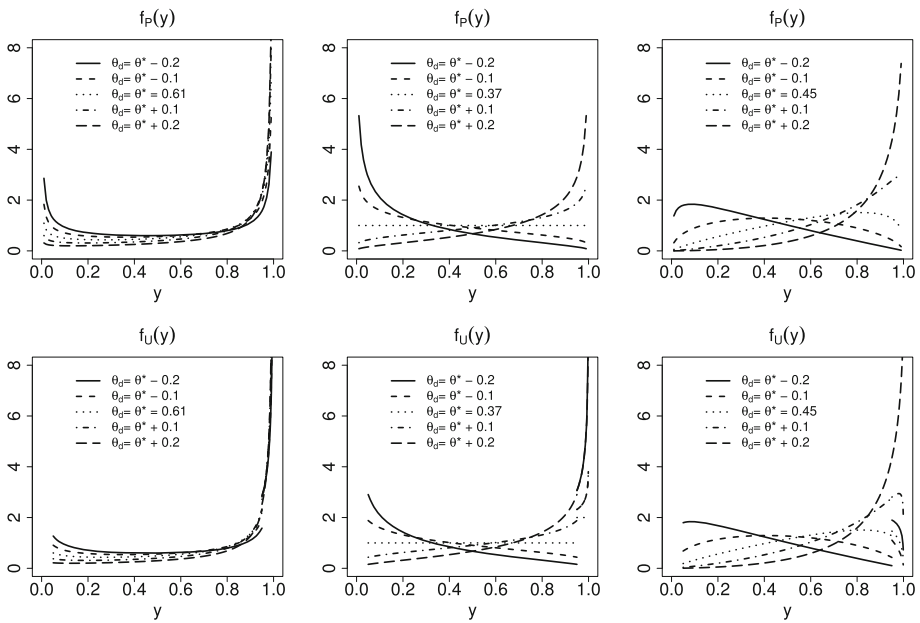- $P_n \approx \{\eta_n(\theta^{(k)}), \ k = 1, \cdots, M\}$,

Fig. 4 Qualitative features of $f_P(y)$ (top panels) and of $f_U(y)$ (bottom panels) for the example in Sect. 3.1.1 when $n_d = 29, 79, 129$ (left, central and right panels, respectively) and for several choices of θ_d

- $J_n \approx \{\eta_n(\theta^{(k)}) \times \mathbb{I}_{\Omega_1}(\eta_n(\theta^{(k)}), \ k = 1, \cdots, M\}$,
- $C_n \approx \{\eta_n(\theta^{(k)}) : \theta^{(k)} \in \Omega_1, \ k = 1, \cdots, M\}$,
- $U_n \approx \{(1 - \eta_n(\theta^{(k)}))\mathbb{I}_{\Omega_0}(\theta^{(k)}) + \eta_n(\theta^{(k)})\mathbb{I}_{\Omega_1}(\theta^{(k)}), \ k = 1, \cdots, M\}$.

From the empirical distributions of the PrRVs, it is easy to approximate their summaries: for instance, $e_P \simeq \sum_{k=1}^{M} \eta_n(\theta^{(k)})/M$. Note that if $T_n|\theta$ is exactly or asymptotically normal, steps 2, 3 and 4 are not necessary: once $\theta^{(1)}, \cdots, \theta^{(k)}, \cdots, \theta^{(M)}$ are drawn, the values $\eta_n(\theta^{(k)}), \ k = 1, \cdots M$ can be computed using Eq. (6). The R code is provided in the Supplementary Material. Note that the algorithm works also for the reversed one-sided hypotheses test (ie $\Omega_0 = [\theta_0, +\infty)$), point-null hypotheses test (ie $\Omega_0 = \{\theta_0\}, \Omega_1 = \{\theta_1\}$), and two-sided hypotheses test (ie $\Omega_0 = \theta_0, \Omega_1 = \Omega \backslash \{\theta_0\}$) under appropriate specifications of $\eta_n(\theta)$.

4 Application to a clinical trial

We here consider an application to a two samples confirmatory Phase III trial, which aims to show efficacy of a drug in treating Restless Legs Syndrome. This example was introduced in Muirhead and Soaita (2012) and discussed in Eaton et al. (2013). We exclude from the study the pdf $f_J(y)$ as its behavior almost coincides with the one of $f_P(y)$. It is assumed that $T_n \sim N(\theta, \frac{4\sigma^2}{n})$, where θ is the difference in treatment effects and σ^2 is the common variance in the two groups. Superiority of the experimental treatment is declared by rejecting $H_0 : \theta \leq \theta_0 = 0$ when $\alpha = 0.025$. The authors

assume a known variance $\sigma^2 = 64$, and a clinically meaningful difference $\theta_d = 4$. In this case, $\eta_n(\theta_d)$ reaches the desired power value of 0.80 at $n = 128$. At the design stage of the trial, we want to obtain a pre-experimental evaluation of the trial supposing three different prior beliefs about θ:

- neutral: $\Theta \sim N(\theta_d, \frac{4\sigma^2}{n_d})$, with $n_d = 4$. In this case, $\pi_1 = 0.69$;
- pessimistic: $\Theta \sim Unif(a, b)$, with $a = -3, b = 5$. In this case, $\pi_1 = 0.625$;
- optimistic: $\Theta \sim TN(a, b, \theta_d, \frac{4\sigma^2}{n_d})$, with $n_d = 4, a = \theta_0 = 0, b = +\infty$. In this case, $\pi_1 = 1$.

In Table 3 we provide values of e_i and of q_i^γ for $\gamma = 0.25, 0.5, 0.75, i \in \{P, C, U\}$ and $n = 64, 128, 256$. As the sample size increases, expected values and quantiles of the PrRVs increase for each prior belief. As expected, regardless of the sample size, the stochastic order of the PrRVs is never violated. For fixed $n = 128$, in Fig. 5 we report priors (*left panels*) and induced PrRV densities (*right panels*). As a consequence of Corollary 3.1, in the neutral and pessimistic scenarios, summaries of $f_P(y)$ and of $f_C(y)$ assume lower values than the ones of $f_U(y)$. This can be also seen from their pdfs, which are u-shaped. On the other hand, the u-shape of $f_U(y)$ is less marked. In the optimistic scenario, the design prior is entirely concentrated on Ω_1, ie $\pi_1 = 1$. As discussed in Sect. 2.1, this is the case where the definitions of the PrRVs (thus including their cdfs, pdfs, expected values and quantiles) coincide.

Finally, we consider the normal neutral design prior and show expected value and quantiles as functions of n in Fig. 6. This reveals interesting aspects. First, according to Corollary 3.1, $e_U \geq e_C \geq e_J$ and $q_U^\gamma \geq q_C^\gamma \geq q_J^\gamma, \gamma = 0.25, 0.5, 0.75$. Second, the 0.25 quantile of P_n never increases with n meaning that, even when the sample size suddenly increases, the u-shape of $f_P(y)$ can not be avoided, thus P_n still presents high probability masses for low values of the power. In conclusion, note that stochastic order among the PrRVs also implies that sample sizes chosen using summaries of U_n would be less than or equal to those based on summaries of the others PrRVs.

5 Conclusion

In the hybrid Bayesian-frequentist approach to hypotheses testing, the probability of success, PoS, is commonly employed to evaluate the design of an experiment. In its simplest and most common definition, PoS is the expected value of the random power. Two criticisms arise:

1. PoS is just the mean of an entire probability distribution, thus may not be a well representative summary of it (Spiegelhalter et al. 1986; Huson 2009; Rufibach et al. 2016; Liu and Yu 2024);
2. the random power does not properly account for the success of the experiment, as it is defined as the random probability of rejecting H_0, being it true or not (Kunzmann et al. 2021; De Santis et al. 2024).

To address criticism 1, in the seminal paper of Spiegelhalter et al. (1986) it is proposed to complement PoS with other summaries of the random power (such as quantiles), or with the whole distribution itself. This is accomplished in Rufibach et al. (2016),
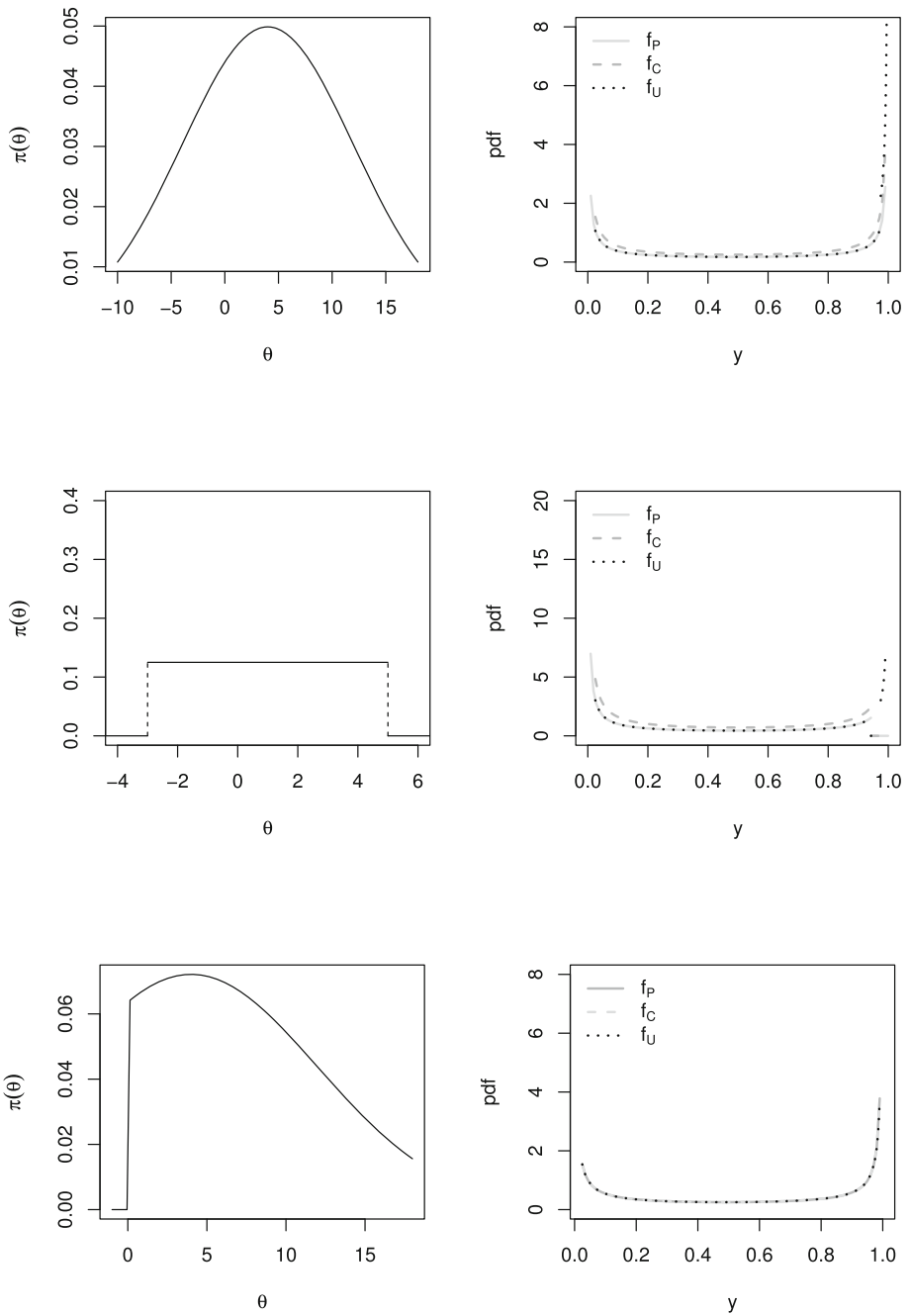
Fig. 5 Normal, uniform and truncated normal prior (left panels) and induced $f_i(y), i \in \{P, C, U\}$ (right panels)

Table 3 Expected values and quantiles for different design priors and sample sizes

n	Design prior	i	e_i	$q_i^{0.25}$	$q_i^{0.5}$	$q_i^{0.75}$
64	Normal	P	0.505	0.004	0.516	0.997
		C	0.728	0.440	0.950	1.000
		U	0.812	0.740	0.997	1.000
	Uniform	P	0.181	0.007	0.072	0.323
		C	0.286	0.091	0.238	0.466
		U	0.552	0.169	0.516	0.993
	Truncated normal	P	0.726	0.435	0.948	1.000
		C	0.726	0.435	0.948	1.000
		U	0.726	0.435	0.948	1.000
128	Normal	P	0.560	0.002	0.806	1.000
		C	0.810	0.730	0.999	1.000
		U	0.866	0.952	1.000	1.000
	Uniform	P	0.283	0.004	0.104	0.564
		C	0.451	0.141	0.424	0.756
		U	0.654	0.290	0.809	0.996
	Truncated normal	P	0.810	0.730	0.999	1.000
		C	0.810	0.730	0.999	1.000
		U	0.810	0.730	0.999	1.000
256	Normal	P	0.598	0.000	0.980	1.000
		C	0.867	0.951	1.000	1.000
		U	0.907	0.996	1.000	1.000
	Uniform	P	0.379	0.002	0.166	0.847
		C	0.607	0.240	0.708	0.963
		U	0.751	0.512	0.977	0.999
	Truncated normal	P	0.867	0.951	1.000	1.000
		C	0.867	0.951	1.000	1.000
		U	0.867	0.951	1.000	1.000

where the pdf of P_n is derived and studied under normality assumptions. To overcome criticism 2, multiple alternative definitions of PoS have been proposed in the literature (and deeply reviewed in Kunzmann et al. 2021). However, these alternative definitions are still affected by criticism 1.

Here, our starting point are four main definitions of PoS identified in De Santis et al. (2024), which can be seen as expected values of PrRVs. In the proposed analysis, we aim to provide tools useful to avoid criticism 1 and, at the same time, we investigate which definition of PoS overcomes criticism 2.

Specifically, we provide general expressions of the cdf and pdf of the PrRVs under investigation, as well as closed-form expressions under normality assumption of the test statistic; moreover, we sketch a simulation algorithm useful when explicit formulas are not available. We show our ideas through an illustrative example on log-hazard ratio and an application to a two arms Phase III clinical trial.
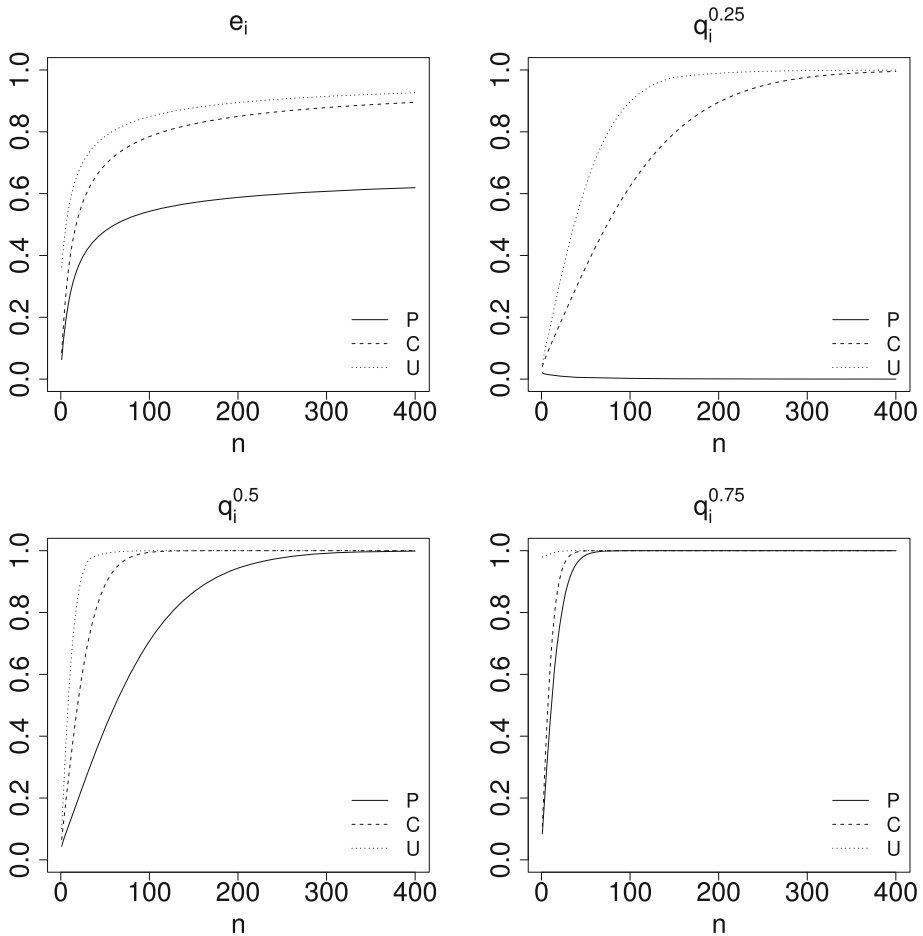
Fig. 6 Expected values and quantiles as a function of n when the design prior is $N(4, \frac{4 \cdot 64}{4})$

When the design prior assigns null or negligible probability masses to Ω_0 as, for instance, in the optimistic design prior case of Sect. 4, the discrepancies between the PrRVs are null or negligible. This is also noted in Kunzmann et al. (2021), with specific reference to the definitions of PoS, ie the expected values of the PrRVs.

On the other hand, when the design prior assigns not negligible probability masses to Ω_0, the use of U_n to address the evaluation of the experiment success is crucial. In fact, in this case we care about both rejecting H_0 when it is false, and not rejecting H_0 when it is true. U_n accounts for both the aspects; conversely, when C_n is considered, the interpretation of success is only related to the proper rejection of H_0; finally, when we consider P_n or J_n, the power of the test under Ω_1 is mixed up with the type I error and 0 values, respectively. This is debated in the application of Sect. 4 through the pessimistic and neutral design prior cases, where graphical representations of f_P (and

consequently f_J) show that it never loses its u-shape. Not by chance, the 0.25 quantile of f_P tends to 0 as n increases.

Finally, we note that the PrRVs are likely to be skewed, thus the median is usually a better representative summary of the entire distribution than the expected value. Similar conclusions are drawn in Liu and Yu (2024) and Huson (2009) as well. This has an impact on sample size determination (see De Santis et al. 2024).

Appendix A

Here, we continue the proof Theorem 3.1 deriving cdf and pdf of J_n and C_n.

1. The random variable $J_n = \eta_n(\Theta)\mathbb{I}_{\Omega_1}(\Theta)$ is a mixture of a point mass distribution on $y = 0$, and of the random power, that is an absolutely continuous random variable, i.e.

$$J_n = \begin{cases} P_n \text{ if } \Theta \in \Omega_1 \\ 0 \text{ if } \Theta \in \Omega_0 \end{cases}.$$

The support of J_n is $\{0\} \cup (\alpha, 1]$ and its cdf is:

$$\mathbb{F}_J(y) = \mathbb{P}_\pi[J_n \le y] = \mathbb{P}_\pi[\eta_n(\Theta)\mathbb{I}_{\Omega_1}(\Theta) \le y, (0 \le y \le \alpha) \cup (\alpha < y \le 1)] =$$
$$= \mathbb{P}_\pi[\eta_n(\Theta)\mathbb{I}_{\Omega_1}(\Theta) \le y, 0 \le y \le \alpha] + \mathbb{P}_\pi[\eta_n(\Theta)\mathbb{I}_{\Omega_1}(\Theta) \le y, \alpha < y \le 1] =$$
$$= \begin{cases} 0 & y < 0 \\ \pi_0 & 0 \le y \le \alpha \\ \mathbb{F}_P(y) & \alpha < y \le 1 \\ 1 & y > 1 \end{cases}$$

It is straightforward to see that:

$$f_J(y) = \delta_0(y) + f_P(y)\mathbb{I}_{(\alpha,1]}(y),$$

where $\delta_0(y)$ can be informally thought of as a probability density function which is 0 everywhere for $y \in [0, \alpha]$ except at $y = 0$, where it is $1 - \int_\alpha^1 f_J(y)dy$ (see, for instance, Campbell and Gustafson 2023).

2. The PrRV $C_n = \eta_n(\Theta)|\Theta \in \Omega_1$ is a truncation of P_n on $y \in (\alpha, 1]$. Its cdf is:

$$\mathbb{F}_C(y) = \mathbb{P}_\pi[\eta_n(\Theta) \le y|\Theta \in \Omega_1] =$$
$$= \frac{\mathbb{P}_\pi[\eta_n(\Theta) \le y, \Theta \in \Omega_1]}{\pi_1} =$$
$$= \frac{\mathbb{P}_\pi[\eta_n(\Theta) \le y, \alpha < y \le 1]}{\pi_1} =$$
$$= \frac{\mathbb{F}_P(y) - \pi_0}{\pi_1}.$$

The pdf $f_C(y)$ can be either obtained from deriving $\mathbb{F}_C(y)$ wrt y or by noting that $f_C(y)$ is the pdf $f_P(y)$ truncated on $(\alpha, 1]$. Thus it is:

$$f_C(y) = f_{P|(\alpha,1]}(y) = \frac{f_P(y)}{\int_\alpha^1 f_P(y)dy} \mathbb{I}_{(\alpha,1]}(y)$$

where

$$\int_\alpha^1 f_P(y)dy = \mathbb{P}_\pi[\alpha < \eta_n(\Theta) \le 1] =$$
$$= \mathbb{P}_\pi[\eta_n^{-1}(\alpha) < \Theta \le \eta_n^{-1}(1)] =$$
$$= \mathbb{P}_\pi[\theta_0 < \Theta \le +\infty] =$$
$$= \pi_1.$$

Appendix B

From Eq. (6), it is easy to check that:

$$\eta_n^{-1}(y) = \theta_0 + \frac{\sigma}{\sqrt{n}}(z_y - z_\alpha),$$

and since $\frac{d}{dy}z_y = \frac{d}{dy}\Phi^{-1}(y) = [\phi(z_y)]^{-1}$,

$$\left|\frac{d}{dy}\eta_n^{-1}\right| = \frac{\sigma}{\sqrt{n}}[\phi(z_y)]^{-1} = \frac{\sigma}{\sqrt{n}}\sqrt{2\pi}\exp\left\{\frac{1}{2}z_y^2\right\}.$$

From the expression of $f_P(y)$ in Theorem 3.1, we obtain:

$$f_P(y) = \pi\left(\theta_0 + \frac{\sigma}{\sqrt{n}}(z_y - z_\alpha)\right) \times \frac{\sigma}{\sqrt{n}}\sqrt{2\pi}\exp\left\{\frac{1}{2}z_y^2\right\}\mathbb{I}_{[0,1]}(y),$$

that is Eq. (7).

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Brown BW, Herson J, Atkinson EN, Rozell ME (1987) Projection from previous studies: a Bayesian and frequentist compromise. Control Clin Trials 8(1):29–44. https://doi.org/10.1016/0197-2456(87)90023-7

Campbell H, Gustafson P (2023) Bayes factors and posterior estimation: two sides of the very same coin. Am Stat 77(3):248–258. https://doi.org/10.1080/00031305.2022.2139293

Ciarleglio MM, Arendt CD, Makuch RW, Peduzzi PN (2015) Selection of the treatment effect for sample size determination in a superiority clinical trial using a hybrid classical and Bayesian procedure. Contemp Clin Trials 41:160–171. https://doi.org/10.1016/j.cct.2015.01.002

Crisp A, Miller S, Thompson D, Best N (2018) Practical experiences of adopting assurance as a quantitative framework to support decision making in drug development. Pharm Stat 17(4):317–328. https://doi.org/10.1002/pst.1856

Dallow N, Fina P (2011) The perils with the misuse of predictive power. Pharm Stat 10(4):311–317. https://doi.org/10.1002/pst.467

De Santis F, Gubbiotti S (2023) On the limit distribution of the power function induced by a design prior. Stat Pap. https://doi.org/10.1007/s00362-023-01462-9

De Santis F, Gubbiotti S, Mariani F (2024) Rethinking probability of success as Bayes utility. Submitted

Eaton M, Muirhead R, Soaita A (2013) On the limiting behavior of the probability of claiming superiority in a Bayesian context. Bayesian Anal 8:221–232. https://doi.org/10.1214/13-BA809

Huson L (2009) The Bayesian bootstrap in a predictive power analysis. Case Stud Bus Ind Gov Stat 3(1):18–22

Kunzmann K, Grayling M, Lee K, Robertson DS, Rufibach K, Wason JMS (2021) A review of Bayesian perspectives on sample size derivation for confirmatory trials. Am Stat 75(4):424–432. https://doi.org/10.1080/00031305.2021.1901782

Liu F (2010) An extension of Bayesian expected power and its application in decision making. J Biopharm Stat 20(5):941–953. https://doi.org/10.1080/10543401003618967

Liu CC, Yu RX (2024) Epistemic uncertainty in Bayesian predictive probabilities. J Biopharm Stat 34(3):394–412. https://doi.org/10.1080/10543406.2023.2204943

Muirhead RJ, Soaita AI (2012) On an approach to Bayesian sample sizing in clinical trials. https://doi.org/10.48550/arXiv.1204.4460

O'Hagan A, Stevens JW (2001) Bayesian assessment of sample size for clinical trials of cost-effectiveness. Med Decis Mak 21(3):219–230. https://doi.org/10.1177/0272989X0102100307

O'Hagan A, Stevens JW, Campbell MJ (2005) Assurance in clinical trial design. Pharm Stat 4(3):187–201. https://doi.org/10.1002/pst.175

Rufibach K, Burger HU, Abt M (2016) Bayesian predictive power: choice of prior and some recommendations for its use as probability of success in drug development. Pharm Stat 15:438–446. https://doi.org/10.1002/pst.1764

Spiegelhalter DJ, Freedman LS, Blackburn PR (1986) Monitoring clinical trials: conditional or predictive power? Control Clin Trials 7(1):8–17. https://doi.org/10.1016/0197-2456(86)90003-6

Spiegelhalter DJ, Abrams KR, Myles JP (2004) Bayesian approaches to clinical trials and health-care evaluation. Wiley, New York

Temple JR, Robertson JR (2021) Conditional assurance: the answer to the questions that should be asked within drug development. Pharm Stat 20(6):1102–1111. https://doi.org/10.1002/pst.2128

Wang M (2007) Sample size reestimation by Bayesian prediction. Biom J 49(3):365–377. https://doi.org/10.1002/bimj.200310273

Wang Y, Fu H, Kulkarni P, Kaiser C (2013) Evaluating and utilizing probability of study success in clinical development. Clin Trials 10(3):407–413. https://doi.org/10.1177/1740774513478229