International Journal of Statistics and Probability; Vol. 9, No. 3; May 2020
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

A Paradox in Bland-Altman Analysis and a Bernoulli Approach

Steven B. Kim[1] & Jeffrey O. Wand[1]

[1] Mathematics and Statistics Department, California State University, Monterey Bay, Seaside, California, USA

Correspondence: Steven B. Kim, Mathematics and Statistics Department, California State University, Monterey Bay, 100 Campus Center, Seaside, CA 93955, USA. E-mail: stkim@csumb.edu

Received: February 15, 2020 Accepted: March 19, 2020 Online Published: March 25, 2020

doi:10.5539/ijsp.v9n3p1 URL: https://doi.org/10.5539/ijsp.v9n3p1

Abstract

A reliable method of measurement is important in various scientific areas. When a new method of measurement is developed, it should be tested against a standard method that is currently in use. Bland and Altman proposed limits of agreement (LOA) to compare two methods of measurement under the normality assumption. Recently, a sample size formula has been proposed for hypothesis testing to compare two methods of measurement. In the hypothesis testing, the null hypothesis states that the two methods do not satisfy a pre-specified acceptable degree of agreement. Carefully considering the interpretation of the LOA, we argue that there are cases of an acceptable degree of agreement inside the null parameter space. We refer to this subset as the paradoxical parameter space in this article. To address this paradox, we apply a Bernoulli approach to modify the null parameter space and to relax the normality assumption on the data. Using simulations, we demonstrate that the change in statistical power is not negligible when the true parameter values are inside or near the paradoxical parameter space. In addition, we demonstrate an application of the sequential probability ratio test to allow researchers to draw a conclusion with a smaller sample size and to reduce the study time.

Keywords: Bland-Altman analysis, limits of agreement, paradoxical space, sequential probability ratio test

1. Introduction

In medical, biological, health, and sport sciences, a new method of measurement is preferred if it is as or more reliable than the current standard method (gold standard). If a new method is attractive under practical considerations (e.g., cost, convenience), a small degree of disagreement between two methods of measurement may be acceptable. In the past, regression and correlation analyses were popular ways to assess the degree of agreement, but their drawbacks are now well known (Altman & Bland, 1983; Hopkins, 2000). Since the introduction of Bland and Altman analysis (Bland & Altman, 1986), the limits of agreement (LOA) have been widely used in practice because of its simple calculation, visualization, and interpretation.

In the Bland and Altman analysis, the difference between two measurements is modeled by a normal distribution with two parameters, mean μ and standard deviation σ, and the parameters of interest are $A = \mu - z\sigma$ and $B = \mu + z\sigma$, where z is the critical value calculated from the standard normal distribution (e.g., $z = 1.96$ for a probability of 0.95). It is recommended to prespecify acceptable limits based on clinical necessity, biological considerations, or other practical goals (Giavarina, 2015). For given predefined acceptable limits, denoted by $(-\delta, \delta)$, Lu et al. (2016) formulated hypothesis testing based on confidence intervals (CIs) for A and B. They proposed an iterative numerical approach to calculate the sample size for the hypothesis testing of a given significance level α, statistical power $1 - \beta$, null value δ, and true parameter values δ/σ and μ/σ. They demonstrated the accuracy of the proposed method of sample size calculation via simulation studies, and it is currently implemented in statistical software (MEDCALC, 2019).

In the formulation of the hypothesis test (Lu et al., 2016), the null hypothesis states that the degree of agreement between two methods of measurement is not acceptable (i.e., (A, B) is not within $-\delta$ and δ), and the alternative hypothesis states that it is acceptable (i.e., (A, B) is within $-\delta$ and δ). In this article, we propose a question on the interpretations of (A, B) and $(-\delta, \delta)$. In particular, based on the probabilistic interpretation of (A, B), we argue that the degree of agreement can be acceptable even when the null hypothesis is true. Throughout this paper, the *paradoxical parameter space* is defined as a set of values of (μ, σ) such that the degree of agreement is acceptable even though the null hypothesis is true. To resolve this paradox, we consider a simple alternative hypothesis test, referred to as the Bernoulli approach, which does not require the normality assumption or any distributional assumption. One caveat of the Bernoulli approach is that there is a minimum sample size required to perform a hypothesis test at a given significance level α. To overcome this caveat, the sequential probability ratio test (Wald, 1945; Wald, 1947) is applied in the context of the Bland and Altman analysis.

2. Methods

This section is structured as follows. The formulation of the hypothesis testing by Lu et al. (2016) is reviewed in Section 2.1. The hypothesis testing is formulated under the normality assumption, and it is referred to as the *normal approach* throughout the paper. In Section 2.2, a paradoxical case of the normal approach is introduced, and an alternative approach is considered in Section 2.3. In the new approach, a Bernoulli distribution (i.e., binary outcome) is used to address the paradox and to relax the normality assumption. We refer to this alternative approach as the *Bernoulli approach* throughout the paper. In Section 2.4, the required sample size is compared between the normal approach and the Bernoulli approach. One caveat of the Bernoulli approach is the minimum sample size requirement, and it is briefly discussed in Section 2.5. In Section 2.6, the application of sequential analysis (Wald, 1945; Wald, 1947) is discussed to overcome the caveat.

2.1 Normal Approach (Lu et al., 2016)

Let D_i be the difference between two measurements for $i = 1, 2, \ldots, n$, where n is a fixed sample size. Assume $D_1, \ldots, D_n$ are independent random variables, and assume $D_i \sim N(\mu, \sigma^2)$. Let $A = \mu - z_{1-\gamma/2}\sigma$ be the true lower limit of an acceptable degree of agreement and $B = \mu + z_{1-\gamma/2}\sigma$ be the true upper limit, where $z_{1-\gamma/2}$ is the $100(1-\gamma/2)$-th percentile of the standard normal distribution (e.g., $z_{0.975} = 1.96$). The probabilistic interpretation of (A, B) is $P(A \le D_i \le B) = 1 - \gamma$.

Let $\bar{D} = n^{-1}\sum_{i=1}^{n} D_i$ be the sample mean, and let $S_D^2 = (n-1)^{-1}\sum_{i=1}^{n}(D_i - \bar{D})^2$ be the sample variance. Point estimators for A and B are $\hat{A} = \bar{D} - z_{1-\gamma/2}S_D$ and $\hat{B} = \bar{D} + z_{1-\gamma/2}S_D$, respectively, and their standard errors can be estimated by

$$\widehat{\mathrm{SE}} = S_D\sqrt{\frac{1}{n} + \frac{(z_{1-\gamma/2})^2}{2(n-1)}}\,.$$

Given a confidence level $1-\alpha$, CIs for A and B are

$$(L_A, U_A) = \hat{A} \pm t_{1-\alpha/2,n-1}\widehat{\mathrm{SE}}\,,$$
$$(L_B, U_B) = \hat{B} \pm t_{1-\alpha/2,n-1}\widehat{\mathrm{SE}}\,,$$

respectively, where $t_{1-\alpha/2,\nu}$ is the $100(1-\alpha/2)$-th percentile of the t-distribution with ν degrees of freedom.

Given a pre-defined acceptable limit, $\delta > 0$, according to Lu et al. (2016), the null hypothesis is formulated as H_{01}: $A < -\delta$ or H_{02}: $B > \delta$ and the alternative hypothesis as H_{11}: $A \ge -\delta$ and H_{12}: $B \le \delta$. When H_{01} and H_{02} are rejected simultaneously, the two methods of measurement would be inferred to agree. In other words, the alternative hypothesis is concluded when both $L_A \ge -\delta$ and $U_B \le \delta$ happen.

Lu et al. (2016) examined the accuracy of their sample size formula and iterative method by Monte Carlo simulations of 10,000 replicates. For instance, when $\mu = 0$ and $\delta/\sigma = 2.8, 2.9, 3.0$, the calculated sample size was $n = 49, 40, 33$, respectively, for significance level $\alpha = 0.05$ and statistical power $1-\beta = 0.8$. Instead of Monte Carlo simulations, their sample size calculations can be evaluated by deriving the exact sampling distribution of (L_A, U_B). See the Appendix A for more details. By a numerical search for n in the aforementioned three scenarios, we found $n = 47, 38, 32$ from the exact sampling distribution, and these results are fairly close to $n = 49, 40, 33$, respectively. The sample size calculations were similar under other fixed parameter values, and the simulation-based iterative method by Lu et al. (2016) seems reasonably close to the numerical results based on the exact sampling distribution.

2.2 A Paradox in the Normal Approach

In what is to follow, the normality assumption is vital. Suppose that two methods of measurement agree such that $A = -\delta$ and $B = \delta$. If the normality assumption is true, we have that $P(A \le D_i \le B) = 1-\gamma$. If the normality assumption is violated, this interpretation may fail, and the consequence may depend on the degree of departure from the normality assumption. If the normality assumption is violated, a logarithmic transformation (in a case of skewness) or a nonparametric approach may be a suitable alternative (Bland & Altman, 1999; Grilo & Grilo, 2012).

Consider the case where $1-\gamma = 0.95$. If $A = -\delta$ and $B = \delta$, it means that $P(-\delta \le D_i \le \delta) = 0.95$. If $A > -\delta$ and $B < \delta$, then we may anticipate $P(-\delta \le D_i \le \delta) > 0.95$. For example, if $\delta = 0.1$, $\mu = 0.01$, and $\sigma = 0.04$, then

$$A = 0.01 - 1.96(0.04) = -0.0684$$
$$B = 0.01 + 1.96(0.04) = 0.0884\,,$$

and $P(-\delta \le D_i \le \delta) = 0.9848$ which was expected.

If $\delta = 0.1$, $\mu = 0.05$, and $\sigma = 0.03$, then $A = -0.0088$ and $B = 0.1088$. In this case, the null hypothesis is true, but

$P(-\delta \le D_i \le \delta) = 0.9522$ which is greater than $1-\gamma = 0.95$. According to the formulation of hypothesis testing, this scenario ($\delta = 0.1$, $\mu = 0.05$, and $\sigma = 0.03$) is not an acceptable degree of agreement, but it may be acceptable according to the probabilistic statement $P(-\delta \le D_i \le \delta) = 0.9522$.

Under the normality assumption $D_i \sim N(\mu, \sigma^2)$, for a given acceptable limit $\delta > 0$, the parameter space of (μ, σ) is the two-dimensional plane $(-\infty, \infty) \times (0, \infty)$, and it can be partitioned into two spaces: the null parameter space

$$\{(\mu, \sigma) : \mu - z_{1-\gamma/2}\sigma < -\delta \text{ or } \mu + z_{1-\gamma/2}\sigma > \delta\}$$

and the alternative parameter space

$$\{(\mu, \sigma) : \mu - z_{1-\gamma/2}\sigma \ge -\delta \text{ and } \mu + z_{1-\gamma/2}\sigma \le \delta\}.$$

Within the null parameter space, we define the *paradoxical parameter space* as the set of values of (μ, σ) such that (1) $\mu - z_{1-\gamma/2}\sigma < -\delta$ and $P(-\delta \le D_i \le \delta) > 1-\gamma$ or (2) $\mu + z_{1-\gamma/2}\sigma > \delta$ and $P(-\delta \le D_i \le \delta) > 1-\gamma$. In other words, the degree of agreement is not acceptable according to the parameter space of (μ, σ), but it may be acceptable according to the probabilistic perspective. Figure 1 demonstrates the paradoxical parameter space for $\delta = 0.1$ and $1-\gamma = 0.95$.

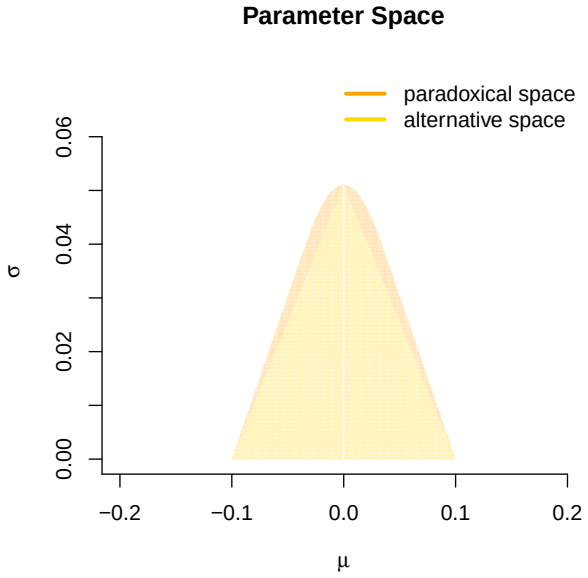

Figure 1. Demonstration of the paradoxical space for $\delta = 0.1$ and $1-\gamma = 0.95$ by numerical search. The white space is the null space

2.3 Bernoulli Approach to Address the Paradox

One simple approach to address the paradox under the normality assumption is a Bernoulli approach (which also addresses potential violation of the normality assumption). For a pre-specified value of $\delta > 0$, let $Y_i = 1$ if $-\delta \le D_i \le \delta$ and $Y_i = 0$ otherwise. By defining $\pi = P(-\delta \le D_i \le \delta)$, a hypothesis test can be formulated as H_0: $\pi \le 1-\gamma$ and H_1: $\pi > 1-\gamma$. If H_0 is true, the two methods of measurement agree with a probability at most $1-\gamma$. If H_1 is true, the two methods agree more often with a probability greater than $1-\gamma$.

Let $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$ be the sample proportion of agreement between two methods of measurement with a given threshold $\delta > 0$. According to the large-sample theory,

$$Z = \frac{\bar{Y} - (1-\gamma)}{\sqrt{(1-\gamma)\gamma/n}} \dot{\sim} N(0,1), \tag{1}$$

and H_0 is rejected when $Z \ge z_{1-\alpha}$ for a fixed significance level α, and so the two methods of measurement would be inferred to agree.

2.4 Sample Size Calculation for a Desired Power

Consider a hypothesis test of the form H_0: $\pi = 1 - \gamma$ versus H_1: $\pi > 1 - \gamma$ (i.e., right-tail test). Based on the test statistic in equation (1), the statistical power $1 - \beta$ is defined as $1 - \beta = P(Z \geq z_{1-\alpha})$ for given $\pi = 1 - \gamma^*$, where $1 - \gamma^* > 1 - \gamma$. Then the required sample size is

$$n = \left(\frac{z_\beta \sqrt{(1-\gamma^*)\gamma^*} - z_{1-\alpha} \sqrt{(1-\gamma)\gamma}}{\gamma^* - \gamma} \right)^2 \tag{2}$$

for given α, $1 - \beta$, $1 - \gamma$, and $1 - \gamma^*$. See the Appendix B for a detailed explanation.

Under normality assumption, Lu et al. (2016) calculated the sample size needed for $\alpha = 0.05$; $1 - \beta = 0.8, 0.9$; $\delta/\sigma = 2, 2.1, \ldots, 3.0$; $\mu/\sigma = 0, 0.1, \ldots, 0.9$. Given $\delta > 0$, $\mu > 0$, and $\sigma > 0$, we can find

$$1 - \gamma^* = P(-\delta \leq D_i \leq \delta) = \Phi\left(\frac{(\delta - \mu)}{\sigma} \right) - \Phi\left(-\frac{(\delta + \mu)}{\sigma} \right)$$

and calculate the required sample size n using equation (2) as shown in table 1. In the table, the superscript * indicates a case when H_1 is true in the Bernoulli approach and H_{01} and H_{02} are true in the normal approach (i.e., a case in the paradoxical parameter space defined in Section 2.2). Tables 2 and 3 present the relative sample size in percent for comparing the Bernoulli approach to the normal approach. Table 2 is for the power of $1 - \beta = 0.8$, and table 3 is for $1 - \beta = 0.9$.

As δ/σ increases for fixed μ/σ, the Bernoulli approach requires more sample size than the normal approach. As μ/σ increases for fixed δ/σ (i.e., μ deviates from zero), the Bernoulli approach requires a smaller sample size than the normal approach. For large μ/σ, even under the normality assumption, there are extreme cases where the required sample size is only 2–3% in the Bernoulli approach when compared to the normal approach. On the other hand, in the scope of tables 2 and 3, the Bernoulli approach can require about 200% of the sample size required by the normal approach. In summary, the sample size requirement is very sensitive to how researchers formulate hypothesis testing particularly when the true parameter values are inside or near the paradoxical space.

Table 1. Values of $1 - \gamma^* = P(-\delta \leq D_i \leq \delta)$ given δ/σ and μ/σ. The superscript * indicates a case when H_1 is true in the Bernoulli approach and H_{01} and H_{02} are true in normal approach

	μ/σ									
δ/σ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
2.0	0.954	0.953^*	0.950^*	0.945	0.937	0.927	0.915	0.900	0.882	0.862
2.1	0.964	0.963	0.961^*	0.956^*	0.949	0.941	0.930	0.917	0.901	0.884
2.2	0.972	0.971	0.969	0.965^*	0.959^*	0.952^*	0.943	0.931	0.918	0.902
2.3	0.979	0.978	0.976	0.973	0.968^*	0.962^*	0.954^*	0.944	0.932	0.919
2.4	0.984	0.983	0.981	0.979	0.975	0.969^*	0.963^*	0.954^*	0.945	0.933
2.5	0.988	0.987	0.986	0.984	0.980	0.976	0.970^*	0.963^*	0.955^*	0.945
2.6	0.991	0.990	0.989	0.987	0.985	0.981	0.977	0.971^*	0.964^*	0.955^*
2.7	0.993	0.993	0.992	0.990	0.988	0.985	0.982	0.977	0.971^*	0.964^*
2.8	0.995	0.995	0.994	0.993	0.991	0.989	0.986	0.982	0.977	0.971^*
2.9	0.996	0.996	0.996	0.995	0.993	0.991	0.989	0.986	0.982	0.977
3.0	0.997	0.997	0.997	0.996	0.995	0.994	0.992	0.989	0.986	0.982

2.5 Minimum Sample Size for Bernoulli Approach

Regardless of the sample size in the Bernoulli approach, $\bar{Y} = 1$ is the strongest evidence against H_0 in favor of H_1, so the maximum value of the Z statistic in equation (1) is

$$Z = \frac{1 - (1 - \gamma)}{\sqrt{\frac{(1-\gamma)\gamma}{n}}} = \sqrt{n\left(\frac{\gamma}{1 - \gamma} \right)}$$

Since we need $Z \geq z_{1-\alpha}$ to reject H_0 in favor of H_1, the minimum sample size requirement in the Bernoulli approach is the smallest integer such that

$$n \geq \frac{(1 - \gamma)(z_{1-\alpha})^2}{\gamma}$$

Table 2. The relative sample size in percent when the Bernoulli approach is compared to the normal approach (bottom rows) for the power of $1 - \beta = 0.8$

	μ/σ									
δ/σ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
2.0	74									
2.1	83	10								
2.2	93	46	5							
2.3	105	74	30	3						
2.4	117	97	56	24	3					
2.5	132	116	81	48	21	2				
2.6	146	135	105	72	44	20	2			
2.7	162	153	129	97	68	42	19	2		
2.8	178	175	152	122	93	66	41	19	2	
2.9	198	190	173	146	118	90	65	41	18	2
3.0	221	211	195	170	145	116	89	64	40	18

Table 3. The relative sample size in percent when the Bernoulli approach is compared to the normal approach (bottom rows) for the power of $1 - \beta = 0.9$

	μ/σ									
δ/σ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
2.0	82									
2.1	90	11								
2.2	99	45	5							
2.3	109	72	29	3						
2.4	119	93	53	23	3					
2.5	131	111	75	45	20	2				
2.6	142	128	96	66	41	19	2			
2.7	156	145	116	87	62	39	18	2		
2.8	168	161	136	108	83	60	38	18	2	
2.9	184	175	153	129	105	80	59	37	17	2
3.0	198	191	176	149	125	103	80	58	37	17

which is derived from the strongest evidence for H_1: $\pi > 1 - \gamma$ (i.e., $\bar{Y} = 1$). For example, we need n to be at least 73 when $1 - \gamma = 0.95$ and $\alpha = 0.05$ in this asymptotic approximation.

If a researcher observes $Y_i = 1$ for $i = 1, 2, \ldots$ in a row or $Y_i = 0$ for $i = 1, 2, \ldots$ in a row, it may be tempting to terminate the study before reaching a fixed sample size n given α and $1 - \beta$. Armitage el al. (1969) demonstrated the inflation of Type I error rate when a researcher continually performs hypothesis testing during data collection. In the following section, we discuss a statistical method for drawing a valid conclusion in the middle of a study.

2.6 Sequential Probability Ratio Test for the Bernoulli Approach

In many practical situations, observations are made sequentially. Due to logistics (e.g., recruiting human subjects and scheduling), the time between two observations D_i and D_{i+1} can be long, and observing one data point can be expensive in terms both cost and labor. Furthermore, in the middle of a study, a researcher can be quite certain whether the degree of agreement is acceptable or not based on accumulated data. In such a case, the sequential probability ratio test (SPRT) can be considered to validly terminate the study before making all n observations (Wald, 1945; Wald, 1947). Particularly for the Bernoulli approach, a simple formula-based rule can be applied to terminate the study during data collection.

Suppose a researcher fixes $\delta > 0$, the maximum of $|D_i|$ which is acceptable. Let $\pi = P(|D_i| \leq \delta)$ be the parameter of interest. Let H_0: $\pi = 1 - \gamma$ be a simple null hypothesis and H_1: $\pi = 1 - \gamma^*$ be a simple alternative hypothesis, where $1 - \gamma^* > 1 - \gamma$ is chosen based on considerations by the researcher. Let α be a significance level and $1 - \beta$ be a statistical power desired by the researcher. Set $Y_i = 1$ if $|D_i| \leq \delta$, and $Y_i = 0$ otherwise. Let $S_m = \sum_{i=1}^{m} Y_i$ be the total number of observations that the two methods of measurement agree after the m^{th} observation. Then the likelihood ratio (for

comparing H_1 to H_0) is given by

$$W_m = \left(\frac{1-\gamma^*}{1-\gamma}\right)^{S_m} \left(\frac{\gamma^*}{\gamma}\right)^{m-S_m}.$$

The researcher makes one of the following decisions based on W_m, γ, γ^*, α, and β.

- Decision 0: If $W_m \geq (1-\beta)/\alpha$, stop the study by concluding H_1: $\pi = 1-\gamma^*$.
- Decision 1: If $W_m \leq \beta/(1-\alpha)$, stop the study by concluding H_0: $\pi = 1-\gamma$.
- Decision 2: Otherwise, make the $(m+1)^{\text{th}}$ observation.

Note that α and β should be small enough so that $(1-\beta)/\alpha > 1$ and $\beta/(1-\alpha) < 1$.

As a numerical example, consider a case with $1-\gamma = 0.95$, $1-\gamma^* = 0.98$, $\alpha = 0.05$, and $1-\beta = 0.9$. According to equation (2) in Section 2.4, the test would require $n = 322$ subjects. On average, however, one can make a conclusion with a *smaller* sample size by using the SPRT. In this study, H_0: $\pi = 0.95$ versus H_1: $\pi = 0.98$, the researcher can stop the study after the m^{th} observation when $W_m \geq 18$ (by concluding H_1) or $W_m \leq 0.1053$ (by concluding H_0). If two methods of measurement agree poorly so that the first four observations result in $Y_i = 0, 0, 1, 0$, then $W_m = 0.4, 0.16, 0.165, 0.066$ for $m = 1, 2, 3, 4$, and the study can be terminated after the fourth subject by concluding H_0 (further data collection is not needed). On the other hand, if the two methods result in $Y_i = 1$ for $i = 1, 2, \ldots, 93$, then $W_m = 1.032, 1.064, \ldots, 18.019$ for $m = 1, 2, \ldots, 93$, so the study can be terminated after the 93^{rd} subject by concluding H_1. In both cases, the sample size is substantially smaller than 322.

3. Simulations

In Section 2, we discussed the paradoxical parameter space in the normal approach, and we considered an alternative formulation of hypothesis testing based on the Bernoulli approach. We then discussed the application of SPRT to the Bland and Altman analysis, and saw that it allows a researcher to terminate the Bernoulli approach early if accumulative data strongly favors H_0 or H_1 over the other hypothesis. In Section 3.1, via Monte Carlo simulations, we compare the normal approach and the Bernoulli approach when the true values of (μ, σ) are inside or near the paradoxical parameter space. Note that our objective is not to argue that one approach is better than the other approach. The objective is to demonstrate the non-negligible difference when the true parameter values are in a neighborhood of the paradoxical space. In Section 3.2, we demonstrate the operating characteristics of the SPRT in the context of the Bland and Altman analysis, where a tested value of π is near the boundary of the parameter space (i.e., π close to one).

3.1 Normal Approach Versus Bernoulli Approach

Simulation studies were designed at $\delta = 0.1$, $1-\gamma = 0.95$, $n = 100, 500, 1000$, $\mu = 0.01, 0.03, 0.05$ and values of σ such that $B = \mu + z_{0.975}\sigma$ is close to $\delta = 0.1$ (i.e., $P(|D_i| \leq \delta)$ is close to 0.95). Each scenario was replicated 50,000 times to approximate the probability of concluding the alternative hypothesis under the normal approach and under the Bernoulli approach.

In most cases considered in this simulation study, the Bernoulli approach is more powerful than the normal approach for a given n. Table 4 provides the simulation results in a neighborhood of the paradoxical parameter space. Figure 2 graphically demonstrates this tendency for $\mu = 0.5$, $0.01 \leq \sigma \leq 0.04$, and $n = 100$ (left panel) and $n = 1000$ (right panel). The difference in $1-\beta$ between the normal approach and the Bernoulli approach is more significant when n is larger. Note that $\sigma = 0.0255$ is a case of the alternative hypothesis $B = 0.09998$ for the normal approach and $P(|D_i| \leq \delta) = 0.975$ for the Bernoulli approach. The probability of concluding the alternative hypothesis is about 0.05 under the normal approach (which is supposed to happen; nominal type I error rate), whereas it is already close to one under the Bernoulli approach when $n = 1000$ (see the right panel of figure 2). The simulation results demonstrate that the statistical power can be very different depending on whether researchers formulate the hypothesis testing based on the normal approach or based on the Bernoulli approach.

3.2 SPRT in Bernoulli Approach

In this section, we consider the SPRT of the Bernoulli approach in the context of the Bland and Altman analysis. The simulations were designed at $\delta = 0.1$, $1-\gamma = 0.95$, and $1-\gamma^* = 0.96, 0.97, 0.98, 0.99$, $\alpha = 0.05$, and $1-\beta = 0.8, 0.9$. The required sample size n is calculated according to equation (2) in Section 2.4 (see tables 5 and 6). In the SPRT, the sample size for concluding H_0: $\pi = 1-\gamma$ (a simple null hypothesis) or H_1: $\pi = 1-\gamma^*$ (a simple alternative hypothesis) is a random variable, and the random sample size is denoted by N. By simulating each scenario 50,000 times, the probability of concluding H_1 (denoted by P_1), the average sample size (denoted by $E(N)$), and the probability that the SPRT requires

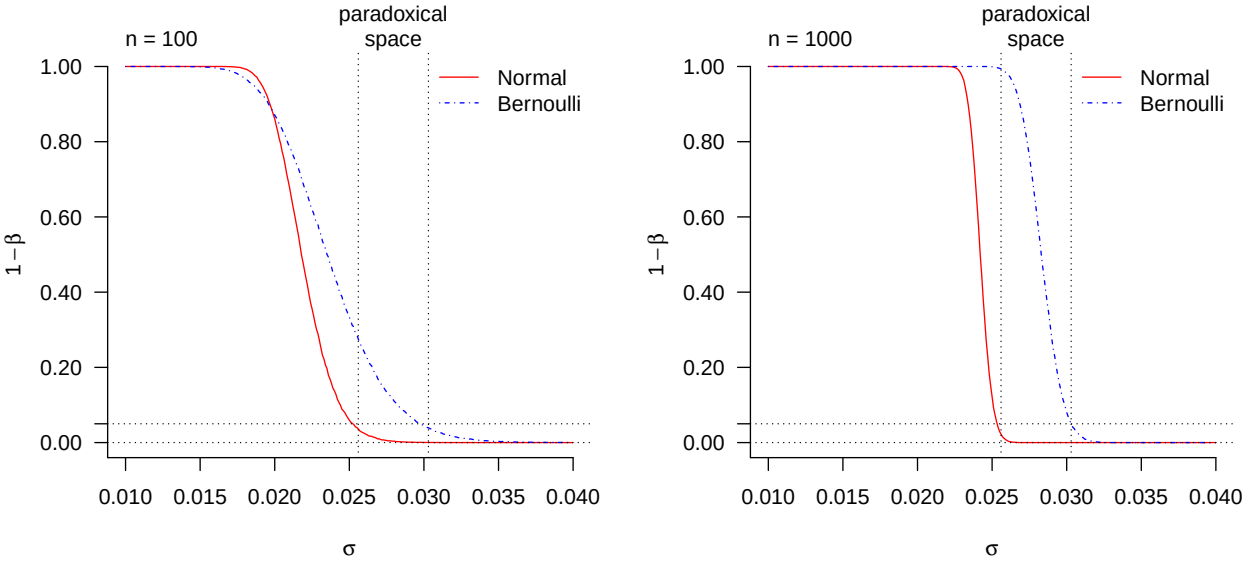

Figure 2. Difference in statistical power between the normal approach and the Bernoulli approach in the paradoxical space for $\delta = 0.1$, $1 - \gamma = 0.95$, and $n = 100$ (left) and $n = 1000$ (right)

Table 4. Comparing the probability of concluding the alternative hypothesis under the normal approach and under the Bernoulli approach based on simulations of 50,000 replicates per scenario. The superscript * indicates a case in the paradoxical parameter space

					Normal Approach			Bernoulli Approach		
μ	σ	B	$P(\lvert D_i\rvert \leq \delta)$	H	$n = 100$	$n = 500$	$n = 1000$	$n = 100$	$n = 500$	$n = 1000$
0.01	0.048	0.104	0.959	1^*	0.012	0.002	0.000	0.081	0.183	0.353
0.01	0.049	0.106	0.955	1^*	0.006	0.000	0.000	0.056	0.091	0.157
0.01	0.050	0.108	0.951	1^*	0.003	0.000	0.000	0.039	0.039	0.050
0.01	0.051	0.110	0.946	0	0.002	0.000	0.000	0.025	0.013	0.013
0.01	0.052	0.112	0.941	0	0.001	0.000	0.000	0.017	0.004	0.002
0.03	0.040	0.108	0.960	1^*	0.002	0.000	0.000	0.092	0.234	0.445
0.03	0.041	0.110	0.957	1^*	0.001	0.000	0.000	0.066	0.120	0.220
0.03	0.042	0.112	0.952	1^*	0.000	0.000	0.000	0.046	0.058	0.086
0.03	0.043	0.114	0.948	0	0.000	0.000	0.000	0.032	0.023	0.025
0.03	0.044	0.116	0.944	0	0.000	0.000	0.000	0.022	0.009	0.006
0.05	0.029	0.106	0.960	1^*	0.002	0.000	0.000	0.089	0.222	0.429
0.05	0.030	0.108	0.955	1^*	0.001	0.000	0.000	0.058	0.091	0.157
0.05	0.031	0.110	0.949	0	0.000	0.000	0.000	0.035	0.029	0.036
0.05	0.032	0.112	0.944	0	0.000	0.000	0.000	0.021	0.008	0.005

a larger sample size than n are approximated (denoted by $P^* \equiv P(N > n)$). The operating characteristics of the SPRT depend on $\pi = P(|D_i| \leq \delta)$, and simulation results are summarized in the tables 5 and 6.

Wald (1945) noted that the SPRT, in many cases, results in saving of about 50% in the number of observations (on average) as compared with the most powerful test. The average percent savings is defined as $(1 - E(N)/n) \times 100\%$, where n is the sample size needed without the SPRT (equation (2)). For $1 - \beta = 0.8$, when H_0: $\pi = 1 - \gamma$ is true, the SPRT results in a saving of 57%, 56%, 57%, and 62% on average when $1 - \gamma^* = 0.96$, 0.97, 0.98, and 0.99, respectively. When H_1: $\pi = 1 - \gamma^*$ is true, the respective average saving is 35%, 31%, 26%, and 27%. The probability of spending a sample size more than n (denoted by P^*) is about 0.06 – 0.08 when H_0 is true and 0.16 – 0.17 when H_1 is true.

For $1 - \beta = 0.9$, when H_0 is true, the SPRT results in the average saving of 74%, 66% 62%, and 65% when $1 - \gamma^* = 0.96$,

0.97, 0.98, and 0.99, respectively. When H_1 is true, the average saving is 42%, 40%, 34%, and 31%, respectively, which is slightly higher than for $1-\beta = 0.8$.

As the true value of $\pi = P(|D_i| \leq \delta)$ decreases from the null value $1-\gamma = 0.95$, the expected saving increases with a lower Type I error rate. As it increases from the alternative value $1-\gamma^*$, the expected saving increases with a higher statistical power. When it is between the null value and the alternative value, the expected saving decreases, still below n, and the probability of concluding H_1 is between the fixed significance level $\alpha = 0.05$ and desired power $1-\beta$.

Table 5. Resulting P_1, $E(N)$, and $P^* = P(N > n)$ in the simulation studies of the SPRT (50,000 replications per scenario) for $\alpha = 0.05$ and $1-\beta = 0.8$

	$1-\gamma^* = 0.96$ $n = 2740$			$1-\gamma^* = 0.97$ $n = 631$			$1-\gamma^* = 0.98$ $n = 253$			$1-\gamma^* = 0.99$ $n = 123$		
π	P_1	$E(N)$	P^*	P_1	$E(N)$	P^*	P_1	$E(N)$	P^*	P_1	$E(N)$	P^*
0.935	0.000	349	0.000	0.002	128	0.003	0.007	61	0.012	0.015	33	0.017
0.940	0.000	464	0.001	0.006	155	0.011	0.015	71	0.022	0.023	37	0.028
0.945	0.005	679	0.011	0.018	198	0.033	0.027	84	0.044	0.035	41	0.040
0.950	0.050	1158	0.084	0.052	262	0.082	0.054	101	0.074	0.052	47	0.060
0.955	0.353	1927	0.234	0.144	352	0.157	0.101	123	0.120	0.084	54	0.082
0.960	0.819	1748	0.167	0.342	444	0.231	0.187	149	0.169	0.129	62	0.116
0.965	0.967	1146	0.021	0.613	470	0.246	0.319	175	0.222	0.195	72	0.154
0.970	0.996	798	0.000	0.833	413	0.168	0.508	192	0.249	0.298	82	0.192
0.975	1.000	601	0.000	0.941	334	0.068	0.703	191	0.235	0.432	91	0.229
0.980	1.000	479	0.000	0.983	265	0.012	0.853	174	0.171	0.588	97	0.243
0.985	1.000	399	0.000	0.996	215	0.001	0.941	149	0.082	0.753	97	0.226
0.990	1.000	342	0.000	1.000	179	0.000	0.983	126	0.020	0.886	91	0.159
0.995	1.000	299	0.000	1.000	153	0.000	0.998	105	0.001	0.967	80	0.065

Table 6. Resulting P_1, $E(N)$, and $P^* = P(N > n)$ in the simulation studies of the SPRT (50,000 replications per scenario) for $\alpha = 0.05$ and $1-\beta = 0.8$

	$1-\gamma^* = 0.96$ $n = 3717$			$1-\gamma^* = 0.97$ $n = 833$			$1-\gamma^* = 0.98$ $n = 322$			$1-\gamma^* = 0.99$ $n = 148$		
π	P_1	$E(N)$	P^*	P_1	$E(N)$	P^*	P_1	$E(N)$	P^*	P_1	$E(N)$	P^*
0.935	0.000	493	0.000	0.002	177	0.002	0.007	84	0.010	0.013	42	0.020
0.940	0.000	656	0.000	0.005	218	0.008	0.012	98	0.022	0.020	47	0.031
0.945	0.004	973	0.006	0.016	280	0.030	0.026	117	0.044	0.032	54	0.050
0.950	0.051	1708	0.086	0.049	378	0.086	0.051	140	0.079	0.050	61	0.076
0.955	0.427	2825	0.260	0.158	516	0.178	0.104	175	0.137	0.082	72	0.112
0.960	0.908	2130	0.122	0.400	650	0.263	0.204	216	0.206	0.129	84	0.158
0.965	0.994	1246	0.006	0.722	637	0.246	0.377	251	0.265	0.212	98	0.216
0.970	1.000	836	0.000	0.916	503	0.132	0.601	266	0.284	0.328	112	0.274
0.975	1.000	625	0.000	0.980	371	0.031	0.804	247	0.240	0.485	124	0.322
0.980	1.000	500	0.000	0.997	282	0.003	0.927	205	0.145	0.672	127	0.340
0.985	1.000	416	0.000	1.000	225	0.000	0.980	165	0.052	0.832	121	0.316
0.990	1.000	356	0.000	1.000	187	0.000	0.996	133	0.008	0.940	105	0.220
0.995	1.000	311	0.000	1.000	160	0.000	1.000	110	0.000	0.990	87	0.083

4. Example

We revisit the data presented in Bland and Altman (1999). They compared the measurement of systolic blood pressure (SBP) by a human observer and the measurement by an automatic machine. For illustration purposes, the first measurement of each method is considered as in Bland and Altman (1999). There are a total of 85 subjects in the original data, and it seems evident that the human measurement (denoted by $J1$) is lower than the machine measurement (denoted by $S1$), on average, as shown in the Bland-Altman plot at the left panel of figure 3. In addition, it seems that the difference $(J1 - S1)$ does not follow a normal distribution.

Suppose an absolute difference of $\delta = 10$ or higher can severely affect a clinical decision, and suppose clinicians want for $P(-10 \leq J1 - S1 \leq 10)$ to be at least $1-\gamma = 0.95$. Based on the sample of size $n = 85$, we observed only 31 cases where

$-10 \leq J1 - S1 \leq 10$, and the Z statistic for the Bernoulli approach is as extreme as

$$Z = \frac{31/85 - 0.95}{\sqrt{\frac{0.95(1-0.95)}{85}}} = -24.76\,.$$

The conclusion of disagreement (i.e., unacceptable degree of agreement between $J1$ and $S1$) is valid without the normality assumption.

For the application of SPRT (Section 2.6) with $\alpha = 0.01$, $1-\beta = 0.99$, $1-\gamma = 0.95$, and $1-\gamma^* = 0.98$, the trajectory of $\ln(W_m)$ is shown at the right panel of figure 3. Using the logarithmic transformation of W_m, H_1: $\pi = 0.98$ is concluded when $\ln(W_m) \geq \ln[(1-\beta)/\alpha] = \ln(99) = 4.595$, and H_0: $\pi = 0.95$ is concluded when $\ln(W_m) \leq \ln[\beta/(1-\alpha)] = \ln(1/99) = -4.595$. As shown in table 7, H_0: $\pi = 0.95$ is concluded after observing $m = 6$ subjects (the absolute differences exceeded $\delta = 10$ six times in a row) with $\ln(W_m) = -5.498$. Note that the conclusion of H 0 does not necessarily imply that ?? = 0.95 exactly. It implies that ?? does not exceed 0.95 (i.e., poor agreement), so the true value of ?? may be substantially lower than 0.95. Using the sequential analysis, such a large sample size $n = 85$ may not be needed, and the same conclusion (i.e., unacceptable degree of agreement between $J1$ and $S1$) could be drawn with $m = 6$ subjects (7% of $n = 85$) without the normality assumption.

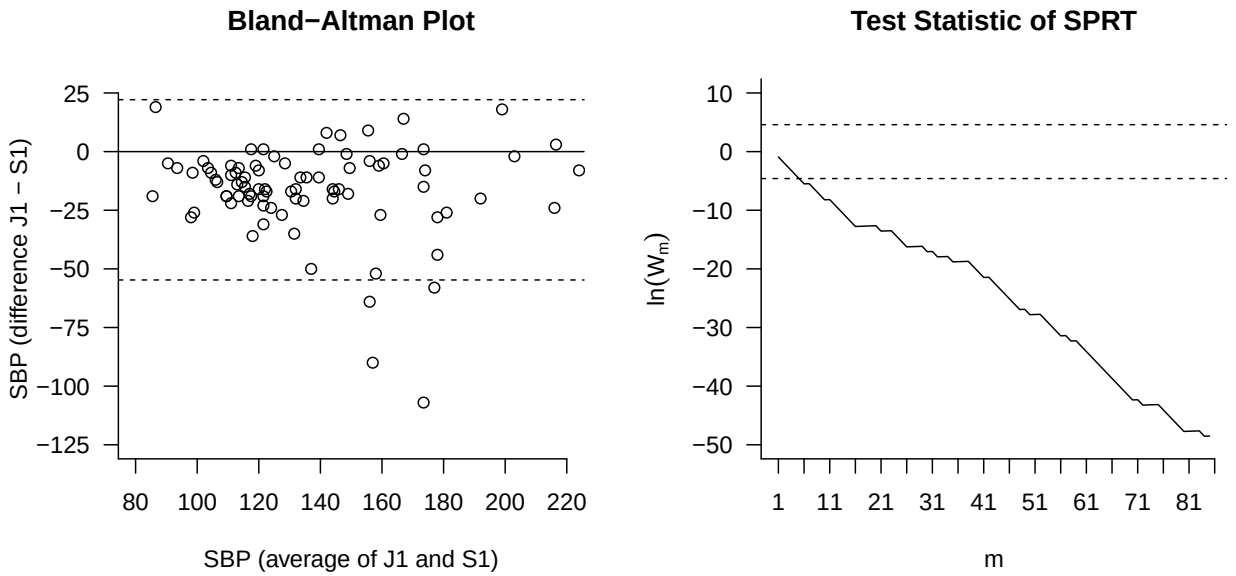

Figure 3. The Bland-Altman plot (left) and the trajectory of $\ln(W_m)$ for SPRT with respect the subject number m (right)

Table 7. The first $m = 6$ subjects in the data (Bland & Altman, 1999) which lead to the conclusion of H_0 in the SPRT

m	J1	S1	Difference ($J1 - S1$)	Binary Outcome	S_m	$\ln(W_m)$
1	100	122	–22	0	0	–0.916
2	108	121	–13	0	0	–1.833
3	76	95	–19	0	0	–2.749
4	108	127	–19	0	0	–3.665
5	124	140	–16	0	0	–4.581
6	122	139	–17	0	0	–5.498

5. Summary

In this article, we discussed the paradoxical parameter space in the normal approach, and the Bernoulli approach (which does not require the normality assumption) was considered based on the probabilistic interpretation of the Bland and Altman analysis given δ for $\pi = P(|D_i| \leq \delta)$. The authors emphasize that it is not reasonable to argue one approach is better than the other because the partition of the parameter space for the null and alternative hypotheses is not the

same. However, researchers should consider carefully whether they want to formulate their hypothesis testing based on the normal approach or based on the Bernoulli approach, and the statistical power can be very different even when both approaches satisfy $P(|D_i| \leq \delta) = 0.95$ or higher.

Given $\delta > 0$, when two methods agree because both $|\mu|$ and σ are small, the Bernoulli approach may require a larger sample size than the normal approach for fixed α and $1 - \beta$. If (μ, σ) belong to the paradoxical parameter space, the Bernoulli approach can require a substantially smaller sample size. Furthermore, the Bernoulli approach does not require any distributional assumption on D_i, and the interpretation of H_0: $\pi = 1-\gamma$ and H_1: $\gamma > 1-\gamma$ may be more straightforward for researchers and practitioners than the interpretation of H_{01}, H_{02}, H_{11}, and H_{12} of the normal approach.

In most studies of comparing two methods of measurement, a sample $(D_1, \ldots, D_n)$ is observed in a sequential manner. In such cases, the application of SPRT can save in sample size and in time of study. The application of sequential analysis is not new to medical and health sciences, however, and in some practical cases, it may not be feasible to calculate W_m for each $m = 1, 2, \ldots$, so group sequential analyses may be suitable alternative methods (Pocock, 1977; O'Brien & Fleming, 1979; Koepcke, 1989; Jennison & Turnbull, 2000). In other practical cases, multiple measurements are taken per subject to compare reliability and validity of two methods of measurement, and the SPRT can be applied in such cases (Kim & Wand, 2019).

There are some shortcomings of the SPRT in practice. Some funded research may require providing data on a pre-specified number of subjects, and a research team may be hired for a specific period of time, both of which could be impacted by a short SPRT and the random sample size. Additionally, if a study is terminated too early, though the small amount of information obtained might be sufficient to draw a conclusion, it would suffer from lack of precision in the parameter estimation.

References

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A (General)*, *132*(2), 235-244. https://doi.org/10.2307/2343787

Altman, C. W., & Bland, J. M. (1983). Measurement in medicine: the analysis of method comparison studies. *The Statistician*, *32*, 307-317. https://doi.org/10.2307/2987937

Bland, J. M., & Altman, C. W. (1986). Statistical method for assessing agreement between two methods of clinical measurement. *Lancet*, *327*, 307-310.

Bland, J. M., & Altman, C. W. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, *8*, 135-160. https://doi.org/10.1177/096228029900800204

Giavarina, D. (2015). Understanding Bland Altman analysis. *Biochemia Medica*, *25*(2), 141-151. https://doi.org/10.11613/BM.2015.015

Grilo, L. M., & Grilo, H. L. (2012). Comparison of clinical data based on limits of agreement. *Biometrical Letters*, *49*(1), 45-56. https://doi.org/10.2478/bile-2013-0003

Hopkins, W. G. (2000). Measures of reliability in sprts medicine and science. *Sports Medicine*, *30*(1), 1-15. https://doi.org/10.2165/00007256-200030010-00001

Jennison, C., & Turnbull, B. W. (2000). *Group Sequential Methods With Applications to Clinical Trials*, Boca Raton, FL: Chapman & Hall/CRC.

Kim, S. B., & Wand, J. O. (2019). Sequential testing in reliability and validity studies with repeated measurements per subject. *International Journal of Statistics and Probability*, *8*(1), 120-134. https://doi.org/10.5539/ijsp.v8n1p120

Koepcke, W. (1989). Analyses of group sequential clinical trials. *Controlled Clinical Trials, 10*(4), 222-230. https://doi.org/10.1016/0197-2456(89)90060-3

Lu, M. J., Zhong, W. H., Liu, Y. X., Miao, H. Z., Li, Y. C., & Ji, M. H. (2016). Sample size for assessing agreement between two methods of measurement by Bland-Altman method. *The International Journal of Biostatistics*, *12*(2). https://doi.org/10.1515/ijb-2015-0039

MEDCALC. (2019). *Sample size calculation: Bland-Altman plot*. Retrieved from https://www.medcalc.org/manual/sampling_blandaltman.php

O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, *35*(3), 549-556. https://doi.org/10.2307/2530245

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrica*, *64*(2), 191-199.

https://doi.org/10.2307/2335684

Shieh, G. (2018). The appropriateness of Bland-Altman's approximate confidence intervals for limits of agreement. *BMC Medical Research Methodology*, *18*(1), 45. https://doi.org/10.1186/s12874-018-0505-y

Wald, A. (1945). Sequential test of statistical hypotheses. *Annals of Mathematical Statistics*, *16*(2), 117-186. https://doi.org/10.1214/aoms/1177731118

Wald, A. (1947). *Sequential Analysis*. New York, NY: John Wiley and Sons.

Appendix A: The Exact Sampling Distribution of (L_A, U_B) and Power

Let $(X, Y) = (\bar{D}, S_D)$ and $(V, W) = (L_A, U_B)$. Since $X \sim N(\mu, \sigma^2/n)$ and

$$Z = \frac{(n-1)Y^2}{\sigma^2} \sim \chi^2_{n-1},$$

the probability density functions (PDFs) of X and Z are

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2/n}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2/n}}$$
$$f_Z(z) = \frac{1}{2^{\frac{n-1}{2}}\,\Gamma\left(\frac{n-1}{2}\right)}\, z^{\frac{n-1}{2}-1}\, e^{-z/2}$$

for $-\infty < x < \infty$ and $z > 0$. Using the Jacobian transformation, we see that the PDF of Y is

$$\begin{aligned} f_Y(y) &= f_Z\left(\frac{(n-1)y^2}{\sigma^2}\right)\left|\frac{dz}{dy}\right| \\ &= \frac{1}{2^{\frac{n-1}{2}}\,\Gamma\left(\frac{n-1}{2}\right)}\left(\frac{(n-1)y^2}{\sigma^2}\right)^{\frac{n-1}{2}-1} e^{-\frac{(n-1)y^2}{2\sigma^2}}\left|\frac{2(n-1)y}{\sigma^2}\right| \end{aligned}$$

for $y > 0$. Since X and Y are independent random variables, their joint PDF is the product of f_X and f_Y, so

$$f_{XY}(x, y) = \frac{n^{\frac{1}{2}}(n-1)^{\frac{n-1}{2}}}{2^{\frac{n}{2}-1}\,\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)\sigma^n}\, y^{n-2}\, e^{-\frac{n(x-\mu)^2+(n-1)y^2}{2\sigma^2}}.$$

Note that $V = X - aY$ and $W = X + aY$, where a is a constant as

$$a = z_{1-\gamma/2} + t_{1-\alpha/2,n-1}\sqrt{\frac{1}{n} + \frac{(z_{1-\alpha/2})^2}{2(n-1)}}.$$

Since $(X, Y) \to (V, W)$ is a one-to-one transformation in two-dimensional space, where $X = (V + W)/2$ and $Y = (W - V)/(2a)$, the bivariate transformation can be done by the Jacobian method and, hence, we obtain

$$f_{VW}(v, w) = f_{XY}\left(\frac{v+w}{2},\ \frac{w-v}{2a}\right)|J|,$$

where

$$|J| = \left|\frac{dx}{dv}\frac{dy}{dw} - \frac{dy}{dv}\frac{dx}{dw}\right| = \frac{1}{2a}.$$

To this end, the exact sampling distribution of (V, W) is given by the joint PDF

$$f_{VW}(v, w) = \frac{n^{1/2}(n-1)^{(n-1)/2}(w-v)^{n-2}e^{-g(v,w)}}{2^{n/2-1}\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)\sigma^n(2a)^{n-1}},$$

$$g(v, w) = \frac{n\left(\frac{v+w}{2} - \mu\right)^2 + (n-1)\left(\frac{w-v}{2a}\right)^2}{2\sigma^2}$$

for $-\infty < v < w < \infty$, and the exact power is given by

$$1-\beta = \int_{-\delta}^{\delta}\int_{-\delta}^{w} f_{VW}(v,w)\,dv\,dw\,.$$

Appendix B: The Sample Size Calculation in the Bernoulli Approach

In the Bernoulli approach, the statistical power is given by

$$\begin{aligned}
1-\beta &= P\left(\frac{\bar{Y}-(1-\gamma)}{\sqrt{(1-\gamma)\gamma/n}} > z_{1-\alpha}\right)\\
&= P\left(\bar{Y} > z_{1-\alpha}\sqrt{(1-\gamma)\gamma/n}+(1-\gamma)\right)\\
&= P\left(\frac{\bar{Y}-(1-\gamma^*)}{\sqrt{(1-\gamma^*)\gamma^*/n}} > \frac{z_{1-\alpha}\sqrt{(1-\gamma)\gamma/n}+(1-\gamma)-(1-\gamma^*)}{\sqrt{(1-\gamma^*)\gamma^*/n}}\right)\\
&= 1-\Phi\left(z_{1-\alpha}\sqrt{\frac{(1-\gamma)\gamma}{(1-\gamma^*)\gamma^*}}+\frac{n(\gamma^*-\gamma)}{\sqrt{(1-\gamma^*)\gamma^*}}\right),
\end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Since $\Phi(z_p) = p$, we can express

$$z_\beta = z_{1-\alpha}\sqrt{\frac{(1-\gamma)\gamma}{(1-\gamma^*)\gamma^*}}+\frac{\sqrt{n}(\gamma^*-\gamma)}{\sqrt{(1-\gamma^*)\gamma^*}}\,.$$