2019

Sequential Testing in Reliability and Validity Studies With Repeated Measurements per Subject

Steven B. Kim

Jeffrey O. Wand

Follow this and additional works at: https://digitalcommons.csumb.edu/math_fac
Sequential Testing in Reliability and Validity Studies With Repeated Measurements per Subject

Steven B. Kim¹ & Jeffrey O. Wand¹

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

Correspondence: Steven B. Kim, Mathematics and Statistics Department, California State University, Monterey Bay, Seaside, CA 93955, USA. Tel: 1-831-582-3954. E-mail: stkim@csumb.edu

Received: November 19, 2018 Accepted: December 10, 2018 Online Published: December 24, 2018
doi:10.5539/ijsp.v8n1p120 URL: https://doi.org/10.5539/ijsp.v8n1p120

Abstract

In medical, health, and sports sciences, researchers desire a device with high reliability and validity. This article focuses on reliability and validity studies with \( n \) subjects and \( m \geq 2 \) repeated measurements per subject. High statistical power can be achieved by increasing \( n \) or \( m \), and increasing \( m \) is often easier than increasing \( n \) in practice unless \( m \) is too high to result in systematic bias. The sequential probability ratio test (SPRT) is a useful statistical method which can conclude a null hypothesis \( H_0 \) or an alternative hypothesis \( H_1 \) with 50% of the required sample size of a non-sequential test on average. The traditional SPRT requires the likelihood function for each observed random variable, and it can be a practical burden for evaluating the likelihood ratio after each observation of a subject. Instead, \( m \) observed random variables per subject can be transformed into a test statistic which has a known sampling distribution under \( H_0 \) and under \( H_1 \). This allows us to formulate a SPRT based on a sequence of test statistics. In this article, three types of study are considered: reliability of a device, reliability of a device relative to a criterion device, and validity of a device relative to a criterion device. Using SPRT for testing the reliability of a device, for small \( m \), results in an average sample size of about 50% of the fixed sample size for a non-sequential test. For comparing a device to criterion, the average sample size approaches to 60% approximately as \( m \) increases. The SPRT tolerates violation of normality assumption for validity study, but it does not for reliability study.

Keywords: sequential probability ratio test, reliability, validity, repeated measurements

1. Introduction

In medical, health, and sports sciences, researchers and practitioners want to use a highly valid and reliable device to conduct research or to make an important decision. Often there is a criterion devices (the standard), and researchers often test the validity and reliability of a new device against the criterion. Some researchers suggest using a correlation as a parameter of interest when two or more raters (devices) are compared (Prescott, 2018; Mokkink et al., 2010; Shrout and Fleiss, 1979). However, a correlation is difficult to interpret in the context of a research problem, and it may be controversial to set a threshold of satisfying correlation. In particular, a correlation depends on the heterogeneity of study participants (Hopkins, 2000). In this article, we model measurement error by a normal distribution, and we focus on the mean and the standard deviation (SD) of the normal model as our parameters of interest. From a statistical perspective, it is reasonable to quantify the validity by using the mean of measurement error and the reliability by the SD.

We can increase precision of parameter estimation and statistical power of hypothesis testing by increasing the sample size (i.e., the number of subjects) and/or the repetitions (i.e., the number of repeated measurements per subject). Hopkins (2000) suggested approximately 50 study participants and at least 3 trials, but the desired sample size and the number of repetitions depend on various factors. Sometimes, a too high number of repetitions may introduce systematic change in measurement (e.g., learning effect, fatigue, etc.) For some researchers, it can be difficult to recruit a large number of subjects, and most researchers would like to draw a valid conclusion in an efficient manner in terms of cost, time, and effort. For those researchers, sequential analysis can be a useful statistical method. Wald (1945) introduced the sequential probability ratio test (SPRT) which allows a researcher to terminate hypothesis testing with about 50% of the required sample size on average while preserving the desired significance level and statistical power. In some practical cases, the SPRT (which requires data monitoring after each new data point) can be a serious practical burden, and the SPRT has advanced to various forms of group sequential tests in clinical trials (Pocock, 1977; O’Brien & Fleming, 1979; Wang & Tsiatis, 1987; Jennison & Turnbull, 2000).

In some practical situations, it is more convenient to increase the number of repeated measurements per subject rather than increasing the sample size. Let \( n \) denote the sample size (i.e., the number of subjects) and \( m \) denote the number of repetitions per subject. Using these notations, there are \( n \times m \) data points denoted by \( Y_{ij} \) for \( i = 1, \ldots, n \) and \( j = 1, \ldots, m \). It
may be challenging and exhausting to perform \( n \times m \) sequential tests (after each observation for every subject), but it may be more practical to perform \( n \) sequential tests after each subject. In this article, we demonstrate the effect of increasing \( m \) for fixed \( n \) in a reliability and validity study, and we demonstrate the operating characteristics of the SPRT when it is done after each subject with \( m \) repeated measurements taken per subject. When \( m \) random variables \( (Y_{11}, \ldots, Y_{im}) \) are observed from the \( i \)th subject, we can apply the SPRT based on a known sampling distribution of \( g(Y_{11}, \ldots, Y_{im}) \), where \( g: \mathcal{R}^m \to \mathcal{R} \) (i.e., a single-value statistic which summarizes \( m \) data points). Thus, for reliability and validity studies which typically involve \( m \) repeated measurements per subject, the novelty of this article is that we apply the SPRT based on the sampling distribution of \( g(Y_{11}, \ldots, Y_{im}) \) rather than the SPRT based on the distribution of each \( Y_{ij} \) (as typically done in a SPRT). We provide simulation results that this SPRT, which is based on the likelihood function of the sampling distribution of \( g(Y_{11}, \ldots, Y_{im}) \), still preserves significance level and statistical power while significantly reducing the average sample size.

To this end, researchers who study reliability and validity of measurement devices can terminate their studies early with a substantially fewer number of subjects by taking \( m \) repeated measurements per subject.

This manuscript is structured as follows. Section 2 includes a normal error model, some terminology used throughout this article, and a brief review of SPRT. In Section 3, we focus on hypothesis testing for the reliability of a single device. In Section 4, we discuss hypothesis testing for comparing the reliability of a new device to a criterion (i.e., comparing SDs of measurement error in two devices). In Section 5, we discuss hypothesis testing for comparing the validity of a new device to a criterion (i.e., comparing means of measurement error in two devices).

2. Assumptions, Terminology and Review of SPRT

2.1 Normal Error Model

Suppose the value of a subject is measured by a device (e.g., body temperature, body mass index, force generated by a body part, etc.). Let \( \mu \) denote the true value of a subject (or simply the truth). Suppose the device measures the true value \( m \) times (repeated measurements). Let \( Y_j \) denote the observed value in the \( j \)th measurement which is not exactly equal to \( \mu \). We define the \( j \)th measurement error by \( \epsilon_j = Y_j - \mu \), the difference between the observed value \( Y_j \) and the truth. A typical probability model for measurement error is a normal distribution, denoted by \( \epsilon_j \sim N(\beta, \sigma^2) \).

We further assume \( \epsilon_1, \ldots, \epsilon_m \) are independent. There are two parameters of interest, the mean error \( \beta \) and the standard deviation (SD) \( \sigma \). The normal error model is graphically represented in Figure 1.

![Normal Error Model](http://ijsp.ccsenet.org/images/ijsp-normal-error-model.png)

**Figure 1.** Graphical representation of a normal error model denoted by \( N(\beta, \sigma^2) \)

2.2 Terminology

We say a device is valid (or unbiased) when \( \beta = 0 \). When a device is invalid (i.e., \( \beta \neq 0 \)), overestimation refers to \( \beta > 0 \), and underestimation refers to \( \beta < 0 \). We say a device is more or less reliable (or more or less precise) when \( \sigma \) is smaller or larger, respectively, where \( \sigma \) takes some positive value. For comparing two devices, device 1 relative to device 0, let \( \beta_0 \) and \( \beta_1 \) denote the mean error of device 0 and the mean error of device 1, respectively. Without loss of generality, assume that device 1 is a new device being tested and device 0 serves as a criterion (known to be standard). We say device 1 is valid relative to device 0 when \( \beta_1 - \beta_0 = 0 \). Let \( \sigma_0 \) and \( \sigma_1 \) denote the SDs of device 0 and device 1, respectively. We say device 1 is as reliable as device 0 when \( \sigma_1/\sigma_0 = 1 \) and less reliable than device 0 when \( \sigma_1/\sigma_0 > 1 \). A study of \( \beta \) (or \( \beta_1 - \beta_0 \)) is referred to as a validity study, and a study of \( \sigma \) (or \( \sigma_1/\sigma_0 \)) is referred to as a reliability study.

121
2.3 Estimable Parameters

Consider a single device with the normal error assumption $\epsilon_j \sim N(\beta, \sigma^2)$. If the truth is known, a sequence of independent random variables $(\epsilon_1, \ldots, \epsilon_m)$ is observable, and unbiased estimators of $\beta$ and $\sigma^2$ are $\bar{\epsilon} = \frac{1}{m} \sum_{j=1}^{m} \epsilon_j$ and $S^2 = \frac{1}{m-1} \sum_{j=1}^{m} (\epsilon_j - \bar{\epsilon})^2$, respectively (Hogg and Tanis, 1997). If the truth is unknown, $(\epsilon_1, \ldots, \epsilon_m)$ is not observable, but $(Y_1, \ldots, Y_m)$ is observable instead. With $Y_j = \mu + \epsilon_j$ for $j = 1, \ldots, m$, we cannot estimate $\beta$, but we can estimate $\sigma^2$ by $S^2_Y = \frac{1}{m-1} \sum_{j=1}^{m} (Y_j - \bar{Y})^2$, where $\bar{Y} = \frac{1}{m} \sum_{j=1}^{m} Y_j$ is the sample mean which estimates $\mu + \beta$. In most practical cases, the truth is unknown, so only $\sigma$ can be estimated. For comparing two devices with normal error models $N(\beta_0, \sigma^2_0)$ for device 0 and $N(\beta_1, \sigma^2_1)$ for device 1, we can estimate $\beta_1 - \beta_0$ and $\sigma_1/\sigma_0$ even when the truth is unknown.

2.4 Practical Research Questions for Studying a Single Device

In most practical situations, the truth is unknown, so it is impossible to answer the research question “Is this device valid?” without a criterion. In other words, in absence of knowing the true value of a subject, we cannot conduct a validity study of a single device. Since there is no device with perfect reliability (i.e., $\sigma > 0$), it is challenging to address whether a device is reliable or not without some threshold. Therefore, a practical research question is “Do we have statistical evidence to reject $\sigma = \sigma_0$ in favor of $\sigma > \sigma_0$?” where $\sigma_0$ is a maximal acceptable SD. For example, a 100-meter sprint world record is rounded to two decimal places in seconds. In order to record the truth 9.58 seconds as 9.58 by a valid device with a probability 0.9973 (within 3 SDs from the truth by the empirical rule of a normal model), we want to find $\sigma_0$ such that $9.58 - 3\sigma_0 = 9.575$ or $9.58 + 3\sigma_0 = 9.5849$, so $\sigma_0 = 0.0016$.

2.5 Review of SPRT

Wald (1945) introduced the sequential probability ratio test (SPRT) which allows interim hypothesis testing after each observation. In the framework of SPRT, the sample size is a random variable, and we use an uppercase letter $N$ to denote the random sample size. The general procedure of SPRT is as follows. Let $L_{0i}$ and $L_{1i}$ denote the likelihood after the $i^{th}$ subject under a simple null hypothesis $H_0$ and a simple alternative hypothesis $H_1$, respectively. Let $\alpha$ and $1 - \zeta$ denote the fixed significance level and desired statistical power, respectively. Let $\Lambda_i = L_{1i}/L_{0i}$ be the likelihood ratio which serves as the test statistic for SPRT, and one of the following three decisions is made after observing each $i^{th}$ subject:

- **Case 1**: Terminate the study by concluding $H_0$ if $\Lambda_i \leq \frac{1}{\alpha}$.
- **Case 2**: Terminate the study by concluding $H_1$ if $\Lambda_i \geq \frac{1-\zeta}{\alpha}$.
- **Case 3**: Do not make any conclusion and continue the study if $\frac{1}{\alpha} \leq \Lambda_i \leq \frac{1-\zeta}{\alpha}$.

A SPRT is guaranteed to make a conclusion with a finite sample size (Wald, 1945). As compared with the fixed sample size under the most powerful test introduced by Neyman (1933), the SPRT often results in a saving of about 50% in the sample size on average (Wald, 1945).

3. Reliability Study of a Single Device With $n$ Subjects and $m$ Repetitions

3.1 Formulation of Hypothesis Testing

In a reliability study for a single device (no criterion device), suppose $n$ subjects are recruited and $m$ repeated measurements are taken per subject. In most practical situations, the truth varies from subject to subject (e.g., body mass index). Let $\mu_i$ denote the unknown fixed truth of the $i^{th}$ subject for $i = 1, \ldots, n$. We do not make any assumption about the distribution of $\mu_i$ because it is not needed in our discussion. Let $Y_{ij}$ denote the observed value of the $j^{th}$ measurement from the $i^{th}$ subject, and the $(i, j)^{th}$ measurement error is denoted by $\epsilon_{ij} = Y_{ij} - \mu_i$. As discussed in Section 2.1, we assume $\epsilon_{ij} \sim N(\beta, \sigma^2)$ and independence among $\epsilon_{ij}$’s for $i = 1, \ldots n$ and $j = 1, \ldots, m$. The null hypothesis is $H_0$: $\sigma = \sigma_0$, and the alternative hypothesis is $H_1$: $\sigma > \sigma_0$, where $\sigma_0$ is a maximal SD under practical considerations.

3.2 Exact Sampling Distribution

Let $(Y_1, \ldots, Y_m)$ be $m$ random variables observed from the $i^{th}$ subject. Let $\bar{Y}_j = \frac{1}{m} \sum_{j=1}^{m} Y_{ij}$ be the sample mean for the $i^{th}$ subject, and let $S^2_i = \frac{1}{m-1} \sum_{j=1}^{m} (Y_{ij} - \bar{Y}_j)^2$ be the sample variance which is an unbiased estimator for $\sigma^2$ based on the $m$ observations made from the $i^{th}$ subject. We have the exact sampling distribution $W_i = \frac{(m-1)S^2_i}{\sigma^2} \sim \chi^2_{m-1}$, where $\chi^2_{m-1}$ denotes the chi-square distribution with $m - 1$ degrees of freedom (Hogg and Tanis, 1997). Since $S^2_1, \ldots, S^2_n$
are independent random variables,

\[ W = \frac{n(m - 1)S^2}{\sigma^2} \sim \chi^2_{n(m-1)}, \]

where \( S^2 = \frac{1}{n} \sum_{i=1}^{n} S_i^2 \) is an unbiased estimator for \( \sigma^2 \) when all data points from \( n \) subjects are accumulated. For the purpose of hypothesis testing with \( H_0: \sigma = \sigma_0 \), we replace the unknown parameter \( \sigma \) by the null value \( \sigma_0 \).

The effect of increasing \( m \) for fixed \( n \) is as follows. Given the true value of \( \sigma \), at significance level \( \alpha \), the statistical power shall depend on the product \( n(m-1) \). If \( m = 2^j + 1 \) for \( j = 0, 1, 2, \ldots \), then for \( m^* = 2^{j+1} + 1 \)

\[ W = \frac{n(m^* - 1)S^2}{\sigma^2} \sim \chi^2_{n(m^*-1)}. \]

This follows from the fact that

\[ n(m - 1) = n(2^j) \]
\[ = n \left( \frac{2}{2} \right)(2^j) \]
\[ = n \left( \frac{2^{j+1} + 1 - 1}{2} \right) \]
\[ = n \left( \frac{m^* - 1}{2} \right). \]

In other words, if we increase the number of repeated measurements per subject from \( m \) to \( m^* \) for any \( j \) (e.g., from 2 to 3, from 3 to 5, from 5 to 9, and so on), we can maintain the statistical power with one half of \( n \).

### 3.3 Power Analysis in Non-Sequential Test for \( \sigma \)

If a researcher specifies \( H_0: \sigma = \sigma_0 \) under practical considerations, it is reasonable to consider a one-sided alternative hypothesis \( H_1: \sigma > \sigma_0 \). For illustration purposes, consider the significance level \( \alpha = 0.05 \), the null value \( \sigma_0 = 0.05 \), and the alternative value \( \sigma_1 = 0.06 \). Let \( 1 - \zeta \) denote the statistical power. Table 1 provides required \( n \) for given \( m \) and desired \( 1 - \zeta \) for a non-sequential test. For \( \alpha = 0.05 \) and \( 1 - \zeta \), there are various designs \((m,n)\) such as \((2, 164)\), \((3, 82)\), and \((5, 41)\) to list a few.

**Table 1. Required sample size \( n \) for given \( 1 - \zeta \) and \( m \) in a non-sequential test at level \( \alpha = 0.05 \) with parameter values \( \sigma_0 = 0.05 \), and \( \sigma_1 = 0.06 \)**

<table>
<thead>
<tr>
<th>( 1 - \zeta )</th>
<th>( m = 2 )</th>
<th>( m = 3 )</th>
<th>( m = 4 )</th>
<th>( m = 5 )</th>
<th>( m = 6 )</th>
<th>( m = 7 )</th>
<th>( m = 8 )</th>
<th>( m = 9 )</th>
<th>( m = 10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.70</td>
<td>68</td>
<td>34</td>
<td>23</td>
<td>17</td>
<td>14</td>
<td>12</td>
<td>10</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>0.80</td>
<td>91</td>
<td>46</td>
<td>31</td>
<td>23</td>
<td>19</td>
<td>16</td>
<td>13</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>0.90</td>
<td>128</td>
<td>64</td>
<td>43</td>
<td>32</td>
<td>26</td>
<td>22</td>
<td>19</td>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td>0.95</td>
<td>164</td>
<td>82</td>
<td>55</td>
<td>41</td>
<td>33</td>
<td>28</td>
<td>24</td>
<td>21</td>
<td>19</td>
</tr>
<tr>
<td>0.99</td>
<td>244</td>
<td>122</td>
<td>82</td>
<td>61</td>
<td>49</td>
<td>41</td>
<td>35</td>
<td>31</td>
<td>28</td>
</tr>
</tbody>
</table>

#### 3.4 SPRT for \( \sigma \)

Under the assumption of \( H_0: \sigma = \sigma_0 \), we have the exact sampling distribution

\[ T_i = \frac{(m - 1)S_i^2}{\sigma_0^2} \sim \chi^2_{m-1}. \]

Since \( T_1, T_2, \ldots, T_i \) are independent observations, the likelihood under \( H_0 \) is given by

\[ L_{0i} = \prod_{i=1}^{j} \frac{1}{2^{m-1}} \cdot \frac{1}{\Gamma \left( \frac{m-1}{2} \right)} \cdot t_{h_i}^{m-1} - \frac{1}{2} e^{-\frac{1}{2} \{ t_{h_i} \}^2}. \]

Under the assumption of \( H_1: \sigma = \sigma_1 \), the exact sampling distribution is \( T_i \sim \text{Gamma} \left( \frac{m-1}{2}, \frac{\sigma_1^2}{2} \right) \), so the likelihood under \( H_1 \) is given by

\[ L_{1i} = \prod_{i=1}^{j} \frac{1}{2^{m-1}} \cdot \frac{1}{\Gamma \left( \frac{m-1}{2} \right)} \cdot \left( \frac{\sigma_0}{\sigma_1} \right)^{2(m-1)} \cdot t_{h_i}^{m-1} - \frac{1}{2} \left( \frac{\sigma_0^2}{\sigma_1^2} \right) \cdot e^{-\frac{1}{2} \{ t_{h_i} \}^2}. \]
The $i$th test statistic for SPRT is

$$\Lambda_i = \frac{L_{ii}}{L_{0i}} = \left(\frac{\sigma_0}{\sigma_1}\right)^{2(i-1)} e^{\frac{1}{2}1}\sum_{k=1}^{i} T_k$$

which is in terms of the sum $\sum_{k=1}^{i} T_k$.

To demonstrate operating characteristics of SPRT, assume $\alpha = 0.05$, $\sigma_0 = 0.05$, $\sigma_1 = 0.06$, and $1 - \zeta = 0.95$ (desired power). Using a simulation of 10,000 replicates, Tables 2 and 3 are generated. Tables 2 represents the probability of rejecting $H_0$, and Table 3 represents the average sample size $E(N)$. For $m \geq 2$, the probability of rejecting $H_0$ is slightly under $\alpha = 0.05$ when $H_0$: $\sigma = 0.05$ is true, and it is slightly above $1 - \zeta = 0.95$ when $H_1$: $\sigma = 0.06$ is true. When $H_0$ is true, comparing to required $n$ in a non-sequential test for given $m$, the average sample size is about 57–59% for $m = 2$ and $m = 3$, and it reaches up to about 65% as $m$ increases. When $H_1$ is true, the average sample size is about 49–50% for $m = 2$ and $m = 3$, and it reaches up to about 60% as $m$ increases. This tendency is graphically demonstrated in Figure 2. The average sample size $E(N)$ is significant even when $\sigma$ is slightly below $\sigma_0 = 0.05$ and when $\sigma$ is slightly above as demonstrated in Table 3. The R code for the simulation study is given in Appendix 1.

We can see that $E(N)$ is nearly halved when $m$ is increased in certain ways. If the number of repeated measurements has the form $m = 2^j + 1$ for $j = 0, 1, 2, \ldots$, the average sample size $E(N)$ seems to be halved if $m$ is increased to $m^* = 2^{j+1} + 1$. A researcher who plans a non-sequential test which requires $n = 164$ with $m = 2$ to detect a practical significant difference between $\sigma_0 = 0.05$ and $\sigma_1 = 0.06$ with $\alpha = 0.05$ and $1 - \beta = 0.95$; however, the SPRT with $m = 3$ would result in an expected sample size which is about one quarter of $n$.

Table 2. Probability of rejecting $H_0$ in SPRT designed for $\sigma_0 = 0.05$, $\sigma_1 = 0.06$, $\alpha = 0.05$ and $1 - \zeta = 0.95$

<table>
<thead>
<tr>
<th>$\sigma$</th>
<th>$m = 2$</th>
<th>$m = 3$</th>
<th>$m = 4$</th>
<th>$m = 5$</th>
<th>$m = 6$</th>
<th>$m = 7$</th>
<th>$m = 8$</th>
<th>$m = 9$</th>
<th>$m = 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.03</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.04</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.05</td>
<td>0.036</td>
<td>0.034</td>
<td>0.031</td>
<td>0.031</td>
<td>0.029</td>
<td>0.031</td>
<td>0.030</td>
<td>0.028</td>
<td></td>
</tr>
<tr>
<td>0.06</td>
<td>0.0951</td>
<td>0.956</td>
<td>0.955</td>
<td>0.959</td>
<td>0.960</td>
<td>0.962</td>
<td>0.963</td>
<td>0.962</td>
<td>0.963</td>
</tr>
<tr>
<td>0.07</td>
<td>1.000</td>
<td>0.999</td>
<td>1.000</td>
<td>1.000</td>
<td>0.999</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>0.08</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 3. Average sample size $E(N)$ in SPRT designed for $\sigma_0 = 0.05$, $\sigma_1 = 0.06$, $\alpha = 0.05$ and $1 - \zeta = 0.95$

<table>
<thead>
<tr>
<th>$\sigma$</th>
<th>$m = 2$</th>
<th>$m = 3$</th>
<th>$m = 4$</th>
<th>$m = 5$</th>
<th>$m = 6$</th>
<th>$m = 7$</th>
<th>$m = 8$</th>
<th>$m = 9$</th>
<th>$m = 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.03</td>
<td>23.8</td>
<td>12.1</td>
<td>8.3</td>
<td>6.3</td>
<td>5.2</td>
<td>4.4</td>
<td>3.8</td>
<td>3.4</td>
<td>3.1</td>
</tr>
<tr>
<td>0.04</td>
<td>35.6</td>
<td>18.2</td>
<td>12.4</td>
<td>9.4</td>
<td>7.7</td>
<td>6.5</td>
<td>5.6</td>
<td>5.0</td>
<td>4.5</td>
</tr>
<tr>
<td>0.05</td>
<td>93.9</td>
<td>48.2</td>
<td>32.8</td>
<td>25.0</td>
<td>20.4</td>
<td>17.0</td>
<td>14.8</td>
<td>13.2</td>
<td>11.8</td>
</tr>
<tr>
<td>0.06</td>
<td>80.5</td>
<td>41.4</td>
<td>28.0</td>
<td>21.3</td>
<td>17.4</td>
<td>14.7</td>
<td>12.8</td>
<td>11.3</td>
<td>10.2</td>
</tr>
<tr>
<td>0.07</td>
<td>29.5</td>
<td>15.0</td>
<td>10.3</td>
<td>7.9</td>
<td>6.5</td>
<td>5.5</td>
<td>4.8</td>
<td>4.3</td>
<td>3.9</td>
</tr>
<tr>
<td>0.08</td>
<td>17.1</td>
<td>8.9</td>
<td>6.2</td>
<td>4.8</td>
<td>3.9</td>
<td>3.4</td>
<td>3.0</td>
<td>2.7</td>
<td>2.4</td>
</tr>
</tbody>
</table>

4. Testing Reliability of One Device Relative to Criterion Device

4.1 Formulation of Hypothesis Testing

Suppose reliability of a new device (called device 1) is tested against a criterion device (called device 0) based on $n$ subjects with $m$ repeated measurements taken per subject. Let $\mu_i$ denote the truth of the $i$th subject for $i = 1, \ldots, n$. Assume $\mu_i$ varies from subject to subject, and the value of $\mu_i$ is fixed and unknown. Let $e_{ijk}$ denote the measurement error in the $(i, j)^{th}$ observation by device $k$ for $i = 1, \ldots, n$, $j = 1, \ldots, m$, and $k = 0, 1$. Assume $e_{ijk} \sim N(\beta_{ik}, \sigma^2)$ and independence among all $e_{ijk}$'s. A researcher observes the value of $Y_{ijk} = \mu_i + e_{ijk}$. It is impossible to observe $\mu_i$ and $e_{ijk}$ separately, but we can estimate $\sigma^2_k$ for $k = 0, 1$. In this section, the parameter of interest is $\tau = \sigma_1 / \sigma_0$ which quantifies the reliability of device 1 relative to device 0. We say device 1 is as reliable as device 0 if $\tau = 1$, and we say device 1 is less reliable than device 0 if $\tau > 1$. In terms of percentage, the SD of device 1 is $(\tau - 1)100\%$ greater than the SD of device 0. The null hypothesis is $H_0$: $\tau = \tau_0$, and the alternative hypothesis is $H_1$: $\tau = \tau_1$.

4.2 Exact Sampling Distribution

Let $(Y_{i0}, \ldots, Y_{im})$ be $m$ random variables generated by device 0 and $(Y_{i1}, \ldots, Y_{im})$ be $m$ random variables generated by device 1 when the two devices measured the $i$th subject for $i = 1, \ldots, n$. Let $\bar{Y}_{ik} = \frac{1}{m}\sum_{j=1}^{m} Y_{ijk}$ be the sample mean for
Figure 2. Average sample size of SPRT (%) with respect to $m$ (repetitions per subject)

the $i^{th}$ subject measured by device $k$. From the $i^{th}$ subject, an unbiased estimator for $\sigma_k^2$ is $S_{ik}^2 = \frac{1}{m-1} \sum_{j=1}^{m} (Y_{ijk} - \bar{Y}_{ik})^2$ for $k = 0, 1$, and we have the exact sampling distribution

$$W_i = \left( \frac{S_{i0}^2}{S_{i1}^2} \right) \left( \frac{\sigma_0}{\sigma_1} \right)^2 \left( \frac{1}{\tau_1^2} \right) \sim F_{m-1,m-1},$$

where $F_{m-1,m-1}$ denotes the F distribution with $m - 1$ numerator degrees of freedom and $m - 1$ denominator degrees of freedom. Further note that

$$W = \left( \frac{S_{i0}^2}{S_{i1}^2} \right) \left( \frac{1}{\tau_1^2} \right) \sim F_{n(m-1),n(m-1)},$$

where $S_{i1}^2 = \frac{1}{n} \sum_{i=1}^{n} S_{ik}^2$ is an unbiased estimator for $\sigma_k^2$ for $k = 0, 1$ when all data points from $n$ subjects are combined.

4.3 Power Analysis in Non-Sequential Test for $\tau$

For illustration purposes, consider the significance level $\alpha = 0.05$, the null value $\tau_0 = 1$, and the alternative value $\tau_1 = 1.2$. Let $1 - \zeta$ denote statistical power. Table 4 presents required $n$ for given $1 - \zeta$ for $n \geq 5$ and $m = 2, 3, \ldots, 10$. There are various designs $(m, n)$ for achieving $1 - \zeta = 0.95$ such as $(2, 327)$, $(3, 164)$, and $(5, 82)$. As seen in Section 3.3, a similar pattern of halving $n$ is observed.

Table 4. Required sample size $n$ for given $1 - \zeta$ and $m$ in a non-sequential test at level $\alpha = 0.05$ with parameter values $\tau_0 = 1$, and $\tau_1 = 1.2$

<table>
<thead>
<tr>
<th>$1 - \zeta$</th>
<th>$m = 2$</th>
<th>$m = 3$</th>
<th>$m = 4$</th>
<th>$m = 5$</th>
<th>$m = 6$</th>
<th>$m = 7$</th>
<th>$m = 8$</th>
<th>$m = 9$</th>
<th>$m = 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.70</td>
<td>143</td>
<td>72</td>
<td>48</td>
<td>36</td>
<td>29</td>
<td>24</td>
<td>21</td>
<td>18</td>
<td>16</td>
</tr>
<tr>
<td>0.80</td>
<td>187</td>
<td>94</td>
<td>63</td>
<td>47</td>
<td>38</td>
<td>32</td>
<td>27</td>
<td>24</td>
<td>21</td>
</tr>
<tr>
<td>0.90</td>
<td>259</td>
<td>130</td>
<td>87</td>
<td>65</td>
<td>52</td>
<td>44</td>
<td>37</td>
<td>33</td>
<td>29</td>
</tr>
<tr>
<td>0.95</td>
<td>327</td>
<td>164</td>
<td>109</td>
<td>82</td>
<td>66</td>
<td>55</td>
<td>47</td>
<td>41</td>
<td>37</td>
</tr>
<tr>
<td>0.99</td>
<td>476</td>
<td>238</td>
<td>159</td>
<td>119</td>
<td>96</td>
<td>80</td>
<td>68</td>
<td>60</td>
<td>53</td>
</tr>
</tbody>
</table>
4.4 SPRT for $\tau$

Let

$$T_i = \left( \frac{S_{ij}^2}{S_{i0}^2} \right) \left( \frac{1}{\tau_0} \right) = \frac{S_{ij}^2}{S_{i0}^2}$$

be the test statistic based on $m$ observations made on the $i^{th}$ subject, where a researcher often sets $\tau_0 = 1$ for $H_0$. Under the assumption of $H_0$: $\tau = \tau_0$, we have the exact sampling distribution $T_i \sim F_{m-1,m-1}$. Since $n$ test statistics are independent, the likelihood under $H_0$ is given by

$$L_{0i} = \prod_{h=1}^{i} \frac{\Gamma(m-1)}{(\Gamma(\frac{m-1}{2}))} t_h^{m-2} (1 + t_h)^{-(m-1)}.$$ 

Under the assumption of $H_1$: $\tau = \tau_1$,

$$T_i^* = \left( \frac{S_{ij}^2}{S_{i0}^2} \right) \left( \frac{1}{\tau_1} \right) \sim F_{m-1,m-1}.$$ 

By letting $c = (\tau_1/\tau_0)^2$, and using the Jacobian transformation, the PDF of $T_i = cT_i^*$ is written as

$$f(t_i) = \frac{\Gamma(m-1)}{(\Gamma(\frac{m-1}{2}))} t_i^{m-2} \left( 1 + \frac{t_i}{c} \right)^{-(m-1)} \left( \frac{1}{c} \right)^{\frac{m-1}{2}},$$

which is known as the generalized F distribution, denoted by $G3F(\alpha, \beta, \lambda)$ with $\alpha = \beta = (m - 1)/2$ and $\lambda = 1/c$ (Pham-Gia and Duong, 1989). Therefore, the likelihood under $H_1$ is given by

$$L_{1i} = \prod_{h=1}^{i} \frac{\Gamma(m-1)}{(\Gamma(\frac{m-1}{2}))} t_h^{m-2} \left( 1 + \frac{t_h}{c} \right)^{-(m-1)} \left( \frac{1}{c} \right)^{\frac{m-1}{2}},$$

and the $i^{th}$ test statistic for SPRT is

$$\Lambda_i = \frac{L_{1i}}{L_{0i}} = c^{-\frac{m+1}{2}} \left( \prod_{h=1}^{i} \epsilon_h \right)^{m-1},$$

where $\epsilon_h = \frac{1}{1 + t_h/c}$. Assuming $\tau_0 = 1$, an alternative form of the $i^{th}$ test statistic is

$$\Lambda_i = \left( \frac{1}{\tau_1} \right)^{(m-1)} \prod_{h=1}^{i} \left( \frac{S_{h0}^2 + S_{h1}^2}{S_{h0}^2 + (S_{h1}/\tau_1)^2} \right)^{m-1}.$$ 

The operating characteristics of the SPRT for $\tau$ were studied using a simulation of 10,000 replicates with $\alpha = 0.05$, $\tau_0 = 1$, $\tau_1 = 1.2$, and $1 - \zeta = 0.95$. Tables 5 and 6 represent the probability of rejecting $H_0$ and $E(N)$, respectively. For small $m$, particularly $m = 2$, the SPRT based on the exact sampling distribution of $T_i$ did not follow the usual characteristics of Wald’s SPRT (which often results in about 50% average sample size while preserving $\alpha$ and $1 - \zeta$). When $m = 2$ and $H_0$: $\tau = 0.1$ is true, $H_0$ is rejected with a probability 0.073 which is slightly greater than the fixed $\alpha = 0.05$. When $m = 2$ and $\tau = 1.2$ is true, the resulting statistical power is 0.915 which is lower than the fixed $1 - \zeta = 0.95$. Furthermore, when $m = 2$ and $\tau$ is near $1$ and $1.2$, $E(N)$ is greater than $n = 327$ (the sample size required for a non-sequential test) which defeats the purpose of SPRT. On the other hand, for $m \geq 3$, the Type I error probability is at most $\alpha = 0.05$, and statistical power is at least $1 - \zeta = 0.95$. For large $m$, as shown in Figure 3, the average sample size is close to 60% when $\tau = 1$ and $\tau = 1.2$, but it does not seem to go below 60% as $m$ increases. (We simulated up to $m = 19$.) The R code for the simulation study is given in Appendix 2.

5. Testing Validity of One Device Relative to Criterion Device

5.1 Formulation of Hypothesis Testing

Suppose the validity of a new device (device 1) is tested against a criterion device (device 0) with $n$ subjects and $m$ repeated measurements per subject. Let $\mu_{ij}$ denote the truth of the $i^{th}$ subject in the $j^{th}$ measurement. In other words, the truth may vary from subject to subject and from trial to trial within subject. Let $\epsilon_{ij}$ denote the measurement error in the $(i, j)^{th}$ measurement by device $k$. Assume $\epsilon_{ijk} \sim N(\beta_0, \sigma^2)$ and independence among all $\epsilon_{ijk}$'s. A researcher observes the value of $Y_{ijk} = \mu_{ij} + \epsilon_{ijk}$. Even though it is impossible to observe $\mu_{ij}$ and $\epsilon_{ijk}$ separately, we can still estimate $\beta_1 - \beta_0$. (We cannot
Table 5. Probability of rejecting \( H_0 \) in SPRT designed for \( \tau_0 = 1, \tau_1 = 1.2, \alpha = 0.05 \) and \( 1 - \zeta = 0.95 \)

<table>
<thead>
<tr>
<th>( \tau )</th>
<th>( m = 2 )</th>
<th>( m = 3 )</th>
<th>( m = 4 )</th>
<th>( m = 5 )</th>
<th>( m = 6 )</th>
<th>( m = 7 )</th>
<th>( m = 8 )</th>
<th>( m = 9 )</th>
<th>( m = 10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.9</td>
<td>0.003</td>
<td>0.002</td>
<td>0.001</td>
<td>0.001</td>
<td>0.002</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>1.0</td>
<td>0.073</td>
<td>0.048</td>
<td>0.044</td>
<td>0.040</td>
<td>0.041</td>
<td>0.042</td>
<td>0.040</td>
<td>0.038</td>
<td>0.038</td>
</tr>
<tr>
<td>1.1</td>
<td>0.502</td>
<td>0.515</td>
<td>0.527</td>
<td>0.536</td>
<td>0.527</td>
<td>0.533</td>
<td>0.533</td>
<td>0.533</td>
<td>0.538</td>
</tr>
<tr>
<td>1.2</td>
<td>0.915</td>
<td>0.953</td>
<td>0.955</td>
<td>0.958</td>
<td>0.960</td>
<td>0.960</td>
<td>0.963</td>
<td>0.961</td>
<td>0.961</td>
</tr>
<tr>
<td>1.3</td>
<td>0.991</td>
<td>0.997</td>
<td>0.997</td>
<td>0.997</td>
<td>0.997</td>
<td>0.997</td>
<td>0.998</td>
<td>0.997</td>
<td>0.997</td>
</tr>
<tr>
<td>1.4</td>
<td>0.999</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 6. Average sample size \( E(N) \) in SPRT designed for \( \tau_0 = 1, \tau_1 = 1.2, \alpha = 0.05 \) and \( 1 - \zeta = 0.95 \)

<table>
<thead>
<tr>
<th>( \tau )</th>
<th>( m = 2 )</th>
<th>( m = 3 )</th>
<th>( m = 4 )</th>
<th>( m = 5 )</th>
<th>( m = 6 )</th>
<th>( m = 7 )</th>
<th>( m = 8 )</th>
<th>( m = 9 )</th>
<th>( m = 10 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>110.1</td>
<td>40.8</td>
<td>24.4</td>
<td>17.5</td>
<td>13.6</td>
<td>11.2</td>
<td>9.6</td>
<td>8.3</td>
<td>7.4</td>
</tr>
<tr>
<td>0.9</td>
<td>315.6</td>
<td>64.3</td>
<td>38.9</td>
<td>27.6</td>
<td>21.4</td>
<td>17.6</td>
<td>15.0</td>
<td>12.9</td>
<td>11.5</td>
</tr>
<tr>
<td>1.0</td>
<td>1110.8</td>
<td>141.6</td>
<td>76.7</td>
<td>54.6</td>
<td>42.1</td>
<td>34.6</td>
<td>29.0</td>
<td>25.8</td>
<td>22.4</td>
</tr>
<tr>
<td>1.1</td>
<td>1674.0</td>
<td>295.4</td>
<td>133.9</td>
<td>91.0</td>
<td>72.1</td>
<td>59.2</td>
<td>50.3</td>
<td>44.0</td>
<td>38.6</td>
</tr>
<tr>
<td>1.2</td>
<td>1277.3</td>
<td>144.0</td>
<td>77.5</td>
<td>53.9</td>
<td>41.9</td>
<td>34.4</td>
<td>29.7</td>
<td>25.5</td>
<td>22.7</td>
</tr>
<tr>
<td>1.3</td>
<td>589.6</td>
<td>73.8</td>
<td>44.3</td>
<td>31.3</td>
<td>24.5</td>
<td>19.9</td>
<td>17.1</td>
<td>15.0</td>
<td>13.1</td>
</tr>
<tr>
<td>1.4</td>
<td>242.5</td>
<td>51.8</td>
<td>31.6</td>
<td>22.2</td>
<td>17.4</td>
<td>14.2</td>
<td>12.0</td>
<td>10.4</td>
<td>9.3</td>
</tr>
</tbody>
</table>

Figure 3. Average sample size of SPRT (%) with respect to \( m \) (repetitions per subject)

Let \( D_{ij} = Y_{ij1} - Y_{ij0} \) be the difference in the observed values when device 1 is compared to device 0. Since both \( Y_{ij0} \) and \( Y_{ij1} \) have the same target \( \mu_{ij} \), the difference is \( D_{ij} = \epsilon_{ij1} - \epsilon_{ij0} \sim N(\theta, \delta^2) \), where \( \theta = \beta_1 - \beta_0 \) and \( \delta^2 = \sigma_0^2 + \sigma_1^2 \). Let \( \bar{D}_i = \frac{1}{m} \sum_{j=1}^{m} D_{ij} \) which is an unbiased estimator for \( \theta \) based on \( m \) data points obtained from the \( i^{th} \) subject. Let
$S^2 = \frac{1}{m-1} \sum_{j=1}^{m} (D_{ij} - \bar{D}_i)^2$, which is an unbiased estimator for $\delta^2$ based on $m$ data points obtained from the $i^{th}$ subject. We have the exact sampling distribution

$$T_i = \frac{\bar{D}_i - \theta}{S_i / \sqrt{m}} \sim T_{m-1},$$

where $T_{m-1}$ denotes the T distribution with $m-1$ degrees of freedom. Further note that

$$T = \frac{\bar{D} - \theta}{S / \sqrt{n}} \sim T_{n-1},$$

where $\bar{D} = \frac{1}{n} \sum_{i=1}^{n} D_i$ and $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (D_i - \bar{D})^2$.

### 5.3 Power Analysis in Non-Sequential Test for $\theta$

For illustration purposes, consider the significance level $\alpha = 0.05$, the null value $\theta_0 = 0$, the alternative value $\theta_1 = 0.1$, and the SDs $\sigma_0 = 0.16$ and $\sigma_1 = 0.20$ for devices 0 and 1, respectively, so that $\delta^2 = 0.16^2 + 0.2^2$. Let $1 - \zeta$ denote statistical power. Table 7 provides required sample size $n$ for given $1 - \zeta$ and $m$ for a non-sequential test (for a one-sided test and a two-sided test).

Table 7. Required sample size $n$ for given $1 - \zeta$ and $m$ in a non-sequential test at level $\alpha = 0.05$ with parameter values $\theta_0 = 0$, $\theta_1 = 0.1$, $\sigma_0 = 0.16$, and $\sigma_1 = 0.2$

<table>
<thead>
<tr>
<th>$1 - \zeta$</th>
<th>$m = 2$</th>
<th>$m = 3$</th>
<th>$m = 4$</th>
<th>$m = 5$</th>
<th>$m = 6$</th>
<th>$m = 7$</th>
<th>$m = 8$</th>
<th>$m = 9$</th>
<th>$m = 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.70</td>
<td>17</td>
<td>12</td>
<td>10</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>0.80</td>
<td>22</td>
<td>15</td>
<td>12</td>
<td>10</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>0.90</td>
<td>30</td>
<td>21</td>
<td>16</td>
<td>13</td>
<td>11</td>
<td>10</td>
<td>9</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>0.95</td>
<td>37</td>
<td>26</td>
<td>20</td>
<td>16</td>
<td>14</td>
<td>12</td>
<td>11</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>0.99</td>
<td>54</td>
<td>36</td>
<td>28</td>
<td>23</td>
<td>19</td>
<td>17</td>
<td>15</td>
<td>14</td>
<td>12</td>
</tr>
</tbody>
</table>

### 5.4 SPRT for $\theta$

Let

$$T_i = \frac{\bar{D}_i - \theta_0}{S_i / \sqrt{m}}$$

be the test statistic based on $m$ random variables $(D_{i1}, \ldots, D_{im})$ observed from the $i^{th}$ subject. Under the assumption of $H_0$: $\theta = \theta_0$, we have the exact sampling distribution $T_i \sim T_{m-1}$. Since $T_1, T_2, \ldots, T_i$ are independent random variables, the likelihood under $H_0$ is given by

$$L_{0i} = \prod_{k=1}^{i} \frac{\Gamma\left(\frac{m}{2}\right)}{\sqrt{(m-1)\pi} \Gamma\left(\frac{m-1}{2}\right)} \left(1 + \frac{\theta_0^2}{m-1}\right)^{-\frac{m}{2}}.$$

Under the assumption of $H_1$: $\theta = \theta_1$, the exact sampling distribution is $T_i \sim T_{m-1,c}$, the non-central T distribution with $m-1$ degrees of freedom and non-centrality parameter

$$c = \frac{\sqrt{m(\theta_1 - \theta_0)}}{\sqrt{\sigma_0^2 + \sigma_1^2}}.$$

The likelihood under $H_1$ is given by $L_{1i} = \prod_{k=1}^{i} f(t_k)$, where $f$ is the PDF of $T_{m-1,c}$ (Lenth, 1989; Johnson et al., 1995). Then the $i^{th}$ test statistic for SPRT is $L_{1i} / L_{0i}$.

For a simulation study of 10,000 replicates, we let $\alpha = 0.05$, $\theta_0 = 0$, $\theta_1 = 0.1$, $\sigma_0 = 0.16$ for device 0, $\sigma_1 = 0.2$ for device 1, and $1 - \zeta = 0.95$. Tables 8 and 9 represent the probability of rejecting $H_0$ and $E(N)$, respectively. For $m \geq 2$, the probability of Type I error is at most $\alpha = 0.05$ when $\theta = \theta_0 = 0$ and statistical power is at least $1 - \zeta = 0.95$ when $\theta = \theta_1 = 0.1$. When $H_0$: $\theta = 0$ is true, the average sample size is about 66% for $m = 2$, and it approaches to 55% as $m$ increases. When $H_1$: $\theta = 0.1$ is true, it is about 69% for $m = 2$ and approaches to 58% as $m$ increases. When the true value of $\theta$ is at the exact midpoint of the null value 0 and the alternative value 0.1, the average sample size is greater than the fixed sample size $n$ for a non-sequential test. When the true value of $\theta$ is greater than 0.1, the average sample size is significantly lower than 50% of $n$ (e.g., as low as 25 – 28% when $\theta = 0.2$) as shown in Figure 4. The R code for this simulation study is given in Appendix 3.
6. Impact of Violating Normality Assumption

In practice, the true distribution of measurement error may deviate from a normal distribution. To investigate the impact of violating the normality assumption on the probability of rejecting $H_0$ and the average sample size $E(N)$, measurement error is simulated from a Gumbel distribution and mixed exponential distributions with a gradual increase of skewness (see Appendix 4 for more detail). When studying relative validity (Section 5.4), values of model parameters are chosen to satisfy $\sigma_0 = 0.16$, $\sigma_1 = 0.20$, and zero mean for the null case, $\theta = 0$, for the Gumbel and mixed exponential distributions. In a mixed exponential distribution, the parameter values are gradually altered to generate different skewness. For the alternative case $\theta = 0.1$, a distribution with zero mean is shifted by 0.1 for device 1. As shown in Tables 10 and 11 of Appendix 4, the Type I error probability is below the fixed $\alpha = 0.05$, the statistical power is above the fixed $1 - \zeta = 0.95$, and the average sample size $E(N)$ is close or slightly lower than the case of normality. However, when studying reliability (Section 3.4) and relative reliability (Section 4.4), the Type I error probability is about $2 - 3$ times $\alpha = 0.05$, and the statistical power is below $1 - \zeta$. Researchers conducting reliability studies should be aware of the impact of violating the normality assumption, and reasons and potential remedies remain in our future study.

7. Discussion

As devices for medical and sports science are being more accurate and precise, detecting superiority or inferiority of a new device requires a large number of subjects. When it is appropriate under practical considerations, it would be cost
effective to increase \( m \), the number of repetitions, and to implement SPRT based on the exact sampling distribution of a test statistic calculated after each subject. The SPRT for \( \sigma \) does not require any assumption about \( \beta \), and the average sample size is about 57–58% for small \( m \) under \( H_0 \) and about 49–50% for small \( m \) under \( H_1 \). However, the SPRT for \( \sigma \) cannot be recommended with \( m = 2 \) because it results in Type I error rate greater than \( \alpha \) and statistical power lower than \( 1 - \xi \) with an average sample size being greater than the sample size required for a non-sequential test. The SPRT for \( \tau \) does not require any assumption about \( \beta_0 \) and \( \beta_1 \), and the average sample size approaches to about 60% under \( H_0 \) or \( H_1 \) as \( m \) increases. The SPRT for \( \theta \) does not require any assumption about \( \sigma_0 \) and \( \sigma_1 \), and the average sample size approaches to 55–60% under \( H_0 \) or \( H_1 \) as \( m \) increases. For any SPRT considered in this article, the average sample size becomes significantly lower with a slight deviation from the null value or the alternative value.

For the SPRT for \( \theta \), under the model assumptions, the truth \( \mu_{ij} \) may vary from subject to subject and from trial to trial. If it is the case for testing \( \sigma \) or \( \tau \), the variation of \( \mu_{ij} \) within subject (between trials) becomes a part of the reliability measure. For example, if a device measures walking speed of a subject \( m \) times and the variance of \( \mu_{ij} \) is \( \nu^2 \) within subject, the SPRT for \( \tau \) would test for \( \sqrt{(\sigma_i^2 + \nu^2)/(\sigma_0^2 + \nu^2)} \) which is always smaller than \( \sigma_1/\sigma_0 \). Therefore, the reliability of truth within subject becomes critical when we compare reliability of two devices using the SPRT.

Wald (1945) approximated the average savings of SPRT would be about 50% in the case when the random variable is normally distributed. In particular, he analytically approximated the expected value of the sample size under both the null and alternative hypothesis in order to approximate average savings. We are especially interested in why the SPRT based on the exact sampling distribution results in the varying savings of sample size in the reliability and validity studies with \( m \) repeated measurements. In the future, we would also like to extend Wald’s calculation to the sampling distributions considered in this article (e.g., \( \chi^2 \), non-central T, or F distribution).

There are other parameters of interest when two or more devices are compared such as intraclass correlation and Cronbach’s alpha (Cronbach, 1951; Shrout, 1979). Kistner and Muller (2004) provided the exact sampling distributions under normality assumption and general covariance structure. Therefore, we may be able to use the sampling distributions of their estimators to study the operating characteristics of SPRT. Jin et al. (2013) derived group sequential testing with normality assumption and general covariance structure. Therefore, we may be able to use the sampling distributions of their estimators to study the operating characteristics of SPRT. Jin et al. (2013) derived group sequential testing with normality assumption and general covariance structure. Therefore, we may be able to use the sampling distributions of their estimators to study the operating characteristics of SPRT.

References

https://doi.org/10.1007/BF02310555


https://doi.org/10.2165/00007256-200030010-00001


Appendix 1

The R code for the simulation study in Section 3.4 is provided below.

```r
m = 3
n = 1000
alpha = 0.05
power = 0.95
beta = 1 - power
mu = 100
sigma.true = 0.06
sigma0 = 0.05
sigma1 = 0.06
n.sim = 10000
N = D = rep(NA, n.sim)
c = (sigma1 / sigma0)^2
a = (m - 1) / 2
b = 1 / (2 * c)
bound.lower = beta / (1 - alpha)
bound.upper = (1 - beta) / alpha
for (k in 1:n.sim) {
  full = matrix(rnorm(n * m, mu, sigma.true), nrow=n, ncol=m)
samp.var = apply(full, 1, var)
test.stat = (m - 1) * samp.var / sigma0^2
f0 = cumprod(dchisq(test.stat, m - 1))
f1 = cumprod(dgamma(test.stat, a, b))
sprt.stat = f1 / f0
index = which(sprt.stat <= bound.lower | sprt.stat >= bound.upper)
index = c(index, n)
N[k] = temp = min(index, na.rm = TRUE)
D[k] = ifelse(sprt.stat[temp] > 1, 1, 0)
}
mean(N)
mean(D)
```
Appendix 2
The R code for the simulation study in Section 4.4 is provided below.

```r
m = 5
n = 1000
alpha = 0.05
power = 0.95
beta = 1 - power
sigma0.true = 1
sigma1.true = 1.2
tau0 = 1
tau1 = 1.2
n.sim = 10000
N = D = rep( NA, n.sim )
c = ( tau1 / tau0 ) ^ 2
bound.lower = beta / ( 1 - alpha )
bound.upper = ( 1 - beta ) / alpha
for ( k in 1:n.sim ) {
  y0 = matrix( rnorm( n * m, 0, sigma0.true ), nrow=n, ncol=m )
y1 = matrix( rnorm( n * m, 0, sigma1.true ), nrow=n, ncol=m )
samp0.var = apply( y0, 1, var )
samp1.var = apply( y1, 1, var )
test.stat = samp1.var / samp0.var * ( 1 / tau0 ) ^ 2
f0 = cumprod( df( test.stat, m - 1, m - 1 ) )
f1 = cumprod( df( test.stat / c, m - 1, m - 1 ) / c )
sprt.stat = f1 / f0
index = which( sprt.stat <= bound.lower | sprt.stat >= bound.upper )
index = c( index, n )
N[k] = temp = min( index, na.rm=TRUE )
D[k] = ifelse( sprt.stat[temp] > 1, 1, 0 )
}
mean(N)
mean(D)
```
Appendix 3

The R code for the simulation study in Section 5.4 is provided below.

```r
m = 5
n = 1000
alpha = 0.05
power = 0.95
beta = 1 - power
sigma0 = 0.16
sigma1 = 0.2
theta0 = 0
theta1 = 0.1
theta.true = 0.1
n.sim = 10000
N = D = rep(NA, n.sim)
delta = sqrt(m) * (theta1 - theta0) / sqrt(sigma1^2 + sigma0^2)
bound.lower = beta / (1 - alpha)
bound.upper = (1 - beta) / alpha
for (k in 1:n.sim) {
  full = matrix(rnorm(n * m, theta.true, sqrt(sigma1^2 + sigma0^2)), nrow=n, ncol=m)
  samp.var = apply(full, 1, var)
  samp.mean = apply(full, 1, mean)
  test.stat = (samp.mean - theta0) / sqrt(samp.var / m)
  f0 = cumprod(dt(test.stat, m - 1))
  f1 = cumprod(dt(test.stat, m - 1, delta))
  sprt.stat = f1 / f0
  index = which(sprt.stat <= bound.lower | sprt.stat >= bound.upper)
  index = c(index, n)
  N[k] = temp = min(index, na.rm=TRUE)
  D[k] = ifelse(sprt.stat[temp] > 1, 1, 0)
}
mean(N)
mean(D)
```

Appendix 4

To consider the case when the normality assumption is invalid, measurement error $\epsilon$ is generated from a Gumbel distribution with the PDF

$$f(\epsilon) = \frac{1}{\gamma_2} \exp\left\{-\left(e^{-\epsilon} + \frac{\epsilon - \gamma_1}{\gamma_2}\right)\right\}$$

for $-\infty < \gamma_1 < \infty$ and $\gamma_2 > 0$ and a mixed exponential (ME) distribution with the PDF

$$f(\epsilon) = w \frac{1}{\gamma_1} e^{-\epsilon/\gamma_1} I_{\epsilon\geq0} + (1 - w) \frac{1}{\gamma_2} e^{\epsilon/\gamma_2} I_{\epsilon<0}$$

for $\gamma_1 > 0$, $\gamma_2 > 0$, and $0 < w < 1$. The parameter values are chosen to match with the simulation scenarios in Section 5.4. For the ME distribution, the parameter values are gradually changed to test different skewness (while preserving the effect size in each simulation scenario).
Table 10. Probability of rejecting $H_0$ in SPRT designed for $\theta_0 = 0, \theta_1 = 0.1, \sigma_0 = 0.16, \sigma_1 = 0.20$, $\alpha = 0.05$ and $1 - \zeta = 0.95$ when errors are generated from normal and other distributions

<table>
<thead>
<tr>
<th>Error Model</th>
<th>Skewness</th>
<th>$\theta$</th>
<th>$m = 2$</th>
<th>$m = 3$</th>
<th>$m = 4$</th>
<th>$m = 5$</th>
<th>$m = 6$</th>
<th>$m = 7$</th>
<th>$m = 8$</th>
<th>$m = 9$</th>
<th>$m = 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>0.00</td>
<td>0.00</td>
<td>0.044</td>
<td>0.038</td>
<td>0.035</td>
<td>0.033</td>
<td>0.036</td>
<td>0.029</td>
<td>0.030</td>
<td>0.024</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>0.962</td>
<td>0.964</td>
<td>0.966</td>
<td>0.968</td>
<td>0.970</td>
<td>0.973</td>
<td>0.973</td>
<td>0.974</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gumbel</td>
<td>1.14</td>
<td>0.00</td>
<td>0.024</td>
<td>0.023</td>
<td>0.023</td>
<td>0.024</td>
<td>0.026</td>
<td>0.020</td>
<td>0.021</td>
<td>0.020</td>
<td>0.019</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>0.967</td>
<td>0.971</td>
<td>0.971</td>
<td>0.973</td>
<td>0.972</td>
<td>0.976</td>
<td>0.976</td>
<td>0.977</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ME</td>
<td>0.00</td>
<td>0.00</td>
<td>0.040</td>
<td>0.036</td>
<td>0.031</td>
<td>0.033</td>
<td>0.027</td>
<td>0.029</td>
<td>0.027</td>
<td>0.025</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>0.987</td>
<td>0.983</td>
<td>0.981</td>
<td>0.982</td>
<td>0.978</td>
<td>0.980</td>
<td>0.982</td>
<td>0.982</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ME</td>
<td>1.85</td>
<td>0.00</td>
<td>0.021</td>
<td>0.020</td>
<td>0.022</td>
<td>0.022</td>
<td>0.020</td>
<td>0.018</td>
<td>0.018</td>
<td>0.014</td>
<td>0.017</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>0.993</td>
<td>0.991</td>
<td>0.987</td>
<td>0.988</td>
<td>0.987</td>
<td>0.984</td>
<td>0.985</td>
<td>0.985</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ME</td>
<td>5.66</td>
<td>0.00</td>
<td>0.007</td>
<td>0.008</td>
<td>0.006</td>
<td>0.006</td>
<td>0.007</td>
<td>0.006</td>
<td>0.008</td>
<td>0.007</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.999</td>
<td>0.999</td>
<td>0.998</td>
<td>0.998</td>
<td>0.997</td>
<td>0.997</td>
<td>0.995</td>
</tr>
</tbody>
</table>

Table 11. Average sample size $E(N)$ in SPRT designed for $\theta_0 = 0, \theta_1 = 0.1, \sigma_0 = 0.16, \sigma_1 = 0.20$, $\alpha = 0.05$ and $1 - \zeta = 0.95$ when errors are generated from normal and other distributions

<table>
<thead>
<tr>
<th>Error Model</th>
<th>Skewness</th>
<th>$\theta$</th>
<th>$m = 2$</th>
<th>$m = 3$</th>
<th>$m = 4$</th>
<th>$m = 5$</th>
<th>$m = 6$</th>
<th>$m = 7$</th>
<th>$m = 8$</th>
<th>$m = 9$</th>
<th>$m = 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>0.00</td>
<td>0.00</td>
<td>24.4</td>
<td>15.8</td>
<td>11.8</td>
<td>9.5</td>
<td>7.9</td>
<td>6.9</td>
<td>6.1</td>
<td>5.5</td>
<td>5.0</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>25.4</td>
<td>16.1</td>
<td>12.2</td>
<td>9.8</td>
<td>8.2</td>
<td>7.1</td>
<td>6.3</td>
<td>5.7</td>
<td>5.2</td>
<td></td>
</tr>
<tr>
<td>Gumbel</td>
<td>1.14</td>
<td>0.00</td>
<td>22.2</td>
<td>14.4</td>
<td>11.0</td>
<td>9.0</td>
<td>7.5</td>
<td>6.6</td>
<td>5.9</td>
<td>5.3</td>
<td>4.9</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>24.5</td>
<td>15.7</td>
<td>11.9</td>
<td>9.5</td>
<td>8.0</td>
<td>7.0</td>
<td>6.1</td>
<td>5.5</td>
<td>5.1</td>
<td></td>
</tr>
<tr>
<td>ME</td>
<td>0.00</td>
<td>0.00</td>
<td>24.9</td>
<td>15.9</td>
<td>11.7</td>
<td>9.3</td>
<td>8.0</td>
<td>6.8</td>
<td>6.1</td>
<td>5.5</td>
<td>5.0</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>21.0</td>
<td>14.1</td>
<td>10.8</td>
<td>8.8</td>
<td>7.5</td>
<td>6.6</td>
<td>5.8</td>
<td>5.4</td>
<td>4.9</td>
<td></td>
</tr>
<tr>
<td>ME</td>
<td>1.85</td>
<td>0.00</td>
<td>21.4</td>
<td>13.8</td>
<td>10.6</td>
<td>8.5</td>
<td>7.3</td>
<td>6.4</td>
<td>5.7</td>
<td>5.2</td>
<td>4.8</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>17.4</td>
<td>12.4</td>
<td>9.7</td>
<td>8.1</td>
<td>6.9</td>
<td>6.2</td>
<td>5.5</td>
<td>5.1</td>
<td>4.6</td>
<td>4.6</td>
</tr>
<tr>
<td>ME</td>
<td>5.66</td>
<td>0.00</td>
<td>17.4</td>
<td>11.6</td>
<td>8.8</td>
<td>7.3</td>
<td>6.2</td>
<td>5.5</td>
<td>5.0</td>
<td>4.6</td>
<td>4.3</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>8.7</td>
<td>6.2</td>
<td>5.1</td>
<td>4.4</td>
<td>3.9</td>
<td>3.6</td>
<td>3.4</td>
<td>3.2</td>
<td>3.0</td>
<td></td>
</tr>
</tbody>
</table>

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).