Reproducibility of statistical test results based on $p$-value

Takashi Yanagawa
Biostatistics Center Kurume University
e-mail: yanagawa_takashi@kurume-u.ac.jp

Reproducibility is the essence of a scientific research. Focusing on two-sample problems we discuss in this paper the reproducibility of statistical test results based on $p$-values. First, demonstrating large variability of $p$-values it is shown that $p$-values lack the reproducibility, in particular, if sample sizes are not enough. Second, a sample size formula is developed to assure the reproducibility probability of $p$-value at given level by assuming normal distributions with known variance. Finally, the sample size formula for the reproducibility in general framework is shown equivalent to the sample size formula that has been developed in the Neyman-Pearson type testing statistical hypothesis by employing the level of significance and size of power.

Key words: Interquartile range, Median, Lower and upper quartiles, Neyman-Pearson type statistical test, Reproducibility probability

1. Introduction

$p$-value was introduced by K. Pearson in his chi-square test (1900) and its use promoted by R.A. Fisher as a useful tool for statistical inference from data (Fisher, R.A. 1925). Fisher suggested looked into data to check whether the treatment effect existed if $p$-value $\leq 0.05$, but not to consider data anymore if $p$-value $> 0.05$ (Ono, 2018). On the other hand, Neyman-and Pearson (1928) introduced a method of testing statistical hypotheses by establishing a null and alternative hypotheses, as well as introducing the first and second kind of errors, and by solving a mathematical problem of maximizing the power, which is defined as 1-probability of the second kind of error, keeping the probability of the first kind of error less than given level of significance. Fisher emphasized statistical inference, but Neyman-Pearson emphasized statistical decision; philosophical disputes between Fisher and Neyman on the difference of their ideas continued nearly 30 years until the death of Fisher in 1962. Apparently, their ways of testing are equivalently expressed as follows. Reject the null hypothesis if $p$-value $\leq \alpha$, and hold to reject it if $p$-value $> \alpha$, where $\alpha$ is called the level of significance.

Neyman-Pearson’s approach of testing statistical hypothesis is exclusively illustrated in mathematical statistics text books. We call it in this paper the Neyman-Pearson type test of
testing statistical hypothesis. On the other hand, $p$-value and statistical test based on $p$-value are exclusively illustrated in most biostatistics text books. The author of the present paper tried to bridge these two extremes by introducing the idea of *reproducibility of statistical test results based on $p$-value* in his booklet (Yanagawa, 2019), written in Japanese by giving no mathematical proof.

The purpose of the present paper is to describe it mathematically with rigorous proof and also in English so that to convey our findings not only to Japanese but also to overseas professionals.

Recently, after the publication of the booklet, Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories called for the entire concept of statistical significance to be abandoned (Amrhein, V., Greenland, S., McShane, B., 2019). I appreciate it the calling for give up of Neyman-Pearson type test and coming back to Fisher's statistical inference. If this is the case the 'determination of sample sizes', an important topic in designing scientific research that has been undertaken using the concept of statistical power in the framework of Neyman-Pearson type test, could drift on the air. However, it will be demonstrated in the present paper that the concept of reproducibility could replace it.

To present our idea simply, we focus the discussion in this paper on two-sample problems consisting of treatment group of sample size $n$ and control group of the same sample size; and letting $\Delta$ be the treatment effect, we consider testing statistical hypothesis $H_0 : \Delta = 0$ vs. $H_1 : \Delta = \delta$ ($\delta > 0$).

$p$-value is defined in the setup by

$$p = Pr(T \geq t_0 \mid H_0),$$

where $T$ is a test statistic and $t_0$ is the observed value of $T$.

As is seen in (1) $p$-value is a random variable. The distribution and density functions of $p$-value have been developed by several authors. We summarize it in section 2. Numerical evaluations of location and scale parameters of the distribution of $p$-value are given in section 3 to demonstrate the large variability of $p$-values. Also it is indicated in the section that statistical test results based on $p$-values lack the reproducibility, in particular, if sample sizes are not enough, because of the large variability of $p$-values. In section 4, by assuming two-sample normal distributions with known common variance, a sample size formula is developed to assure the reproducibility probability at given level. Finally in the section, the sample size formula for the reproducibility in general framework is shown equivalent to the sample size formula that has been developed in the Neyman-Pearson type test for testing statistical hypothesis by employing the level of significance and size of power.
2. Distribution of \( p \)-value

2.1 Distribution function and probability density function of \( p \)-value

Suppose two-sample problems consisting of treatment group of sample size \( n \) and control group with the same sample size as the treatment group; and letting \( \Delta \) be the treatment effect, consider testing statistical hypothesis \( H_0 : \Delta = 0 \) vs. \( H_1 : \Delta = \delta (\delta > 0) \) with test statistic \( T \). Let \( F_0 \) and \( F_1 \) be the probability distribution functions of test statistic \( T \) under \( H_0 \) and \( H_1 \), respectively. Distribution function and probability density function of \( p \)-value have been given by Hung, O’Neil, Bauer and Kohne (1997); Sackrowitz and Samuel-Cahn (1999); and Donahue (1999). We represent them in the following proposition.

**PROPOSITION 2.1**

1. \( p \)-value follows a uniform distribution on \((0, 1)\) under \( H_0 \). (Hung, O’Neil, Bauer and Kohne, 1997).

2. Denote by \( G_p(x) \) and \( g_p(x) \) the distribution function and probability density function of \( p \)-value under \( H_1 \), respectively, then \( G_p(x) \) and \( g_p(x) \) are given as follows. (Sackrowitz and Samuel-Cahn, 1999; Donahue, 1999).

\[
G_p(x) = 1 - F_0^{-1}\left(F_0^{-1}(1-x)\right) \quad (0 < x < 1),
\]

\[
g_p(x) = \frac{f_0\left(F_0^{-1}(1-x)\right)}{f_1\left(F_0^{-1}(1-x)\right)} \quad (0 < x < 1).
\]

3. The distribution function of \( F_0^{-1}(1-p) \) under \( H_1 \) is \( F_1(x) \).

**COROLLARY 2.1** Let \( X_1, X_2, \ldots, X_n \) be independently and identically distributed (i.i.d) random variables as normal distribution with mean 0 and variance \( \sigma^2 \); and let \( Y_1, Y_2, \ldots, Y_n \) be i.i.d random variables as normal distribution with mean \( \mu \) and variance \( \sigma^2 \), where \( \sigma \) is assumed known and the treatment effect is expressed by \( \delta = \mu / \sigma \). Furthermore, let the test statistic \( T \) be given by

\[
T = \frac{\bar{Y} - \bar{X}}{\sqrt{(2\sigma^2/n)}}.
\]

Then, it follows that

1. \( G_p(x) \) and \( g_p(x) \) are given as follows, where \( \Phi \) is the distribution function of the standard normal distribution, \( \Phi^{-1} \) is its inverse, \( \varphi \) is its density function, and \( \mu_n = \sqrt{n/2} \delta \).

\[
G_p(x) = 1 - \Phi\left(\Phi^{-1}(1-x) - \mu_n\right) \quad (0 < x < 1),
\]

\[
g_p(x) = \frac{\varphi\left(\Phi^{-1}(1-x) - \mu_n\right)}{\varphi\left(\Phi^{-1}(1-x)\right)} \quad (0 < x < 1).
\]

2. \( \Phi^{-1}(1-p) \) follows normal distribution \( N(\mu_n, 1) \) under \( H_1 \).
Location and dispersion parameters of the distribution of $p$-value

Probability density function given in (4) is illustrated in Figure 1 when $n = 25$ and $\mu = 0.176$, taking $\ln(p) = \log_e(p)$ in the horizontal line. The figure indicates that the distribution of $p$-value is not symmetric and has long right tail under $H_1$. Thus we shall employ the median (50% point of the distribution function) as a location parameter and the interquartile range, 75% point - 25% point, as a dispersion parameter of the distribution function of $p$-value, as employed in Box and Whisker plot (Tukey, 1977).

Putting $q_1 = 0.25$, $q_2 = 0.50$, $q_3 = 0.75$, the lower quartile (25% point), median and upper quartile (75% point) of the distribution of $p$-value are given by $\xi_i$, $i = 1, 2, 3$, respectively, satisfying

$$G_p(\xi_i) = q_i. \quad (5)$$

More precisely, they are given in the following proposition.

**PROPOSITION 2.2** When $q_i$ is given $\xi_i$ is given as follows.

$$\xi_i = 1 - F_n \left( F_i^{-1}(1-q_i) \right)$$

**COROLLARY 2.2** In the framework given in Corollary 2.1

(1) $\xi_i$ is given as follows.

$$\xi_i = 1 - \Phi \left( \mu_n + \Phi^{-1}(1-q_i) \right). \quad (6)$$

$i = 1, 2, 3$, where $\mu_n = \sqrt{n/2} \delta$

(2) $\xi_i$ is a monotone decreasing function of $n$, $i = 1, 2, 3$.

(3) The interquartile range is a monotone decreasing function of $n$.

### 3. Numerical evaluation of quartile and interquartile range

Assuming the framework given in Corollary 2.1, we numerically evaluate in this section the...
Quartile and interquartile range of the distribution of \( p \)-value that are given in the previous section.

3.1 Medians depend sharply on sample size

Figure 2 gives curves of median, lower and upper quartiles of the distribution of \( p \)-value when \( \delta = 0.30 \) for \( n = 20 \sim 200 \), taking \( n \) in the horizontal line. The figure shows that the median decreases rapidly with the increase of \( n \). In particular, it is 0.09, larger than 0.05, when \( n = 20 \), whereas it takes small value 0.0013, much smaller than 0.05, when \( n = 200 \), showing that the median depends sharply on sample sizes.

3.2 Interquartile range

From (6) the interquartile range of the distribution of \( p \)-value is given by

\[
\xi_3 - \xi_1 = \Phi(\mu_n + 0.67) - \Phi(\mu_n - 0.67).
\]

Table 1 gives values of lower and upper quartiles, interquartile range and interquartile ratio of the distribution of \( p \)-value when \( n = 20, 50, 100, 200; \delta = 0.1, 0.2, 0.3 \) obtained from this equation. Also, Figure 3 (A) gives the curve of interquartile range, and (B) gives interquartile ratio, i.e. lower quartile/upper quartile, taking \( n \) in horizontal line, when \( \delta = 0.1, 0.2, 0.3 \).

We get following findings from Table 1 and Figure 3.

- When \( n \) is small the interquartile range of the distribution of \( p \)-value is large despite values of \( \delta \); decreases as the increases of \( n \); and the speed of decreasing is large when \( \delta \) is large, for example, Figure 3 (A) shows when \( \delta = 0.1 \) it decreases slowly and linearly from 0.48 when \( n = 20 \) to 0.35 when \( n = 200 \), whereas when \( \delta = 0.3 \) it decreases exponentially from 0.34 when \( n = 20 \) to 0.01 when \( n = 200 \).

- Interval (25\%point, 75\%point) is a frequent range of random variations of a variable. Thus, a large interquartile range of distribution of \( p \)-value means the large variability of \( p \)-value. In other words, it means that the statistical test results based on \( p \)-value could be easily reversed.
i.e., loose the reproducibility, when sample sizes are small.

- When $n$ and $\delta$ are large, interquartile ranges are small and it appears that the decision based on $p$-value is stable. However, the interquartile ratio increases in $n$, in particular, when $\delta$ is large. For example, when $n=200$ and $\delta=0.3$, the interval (25% point, 75% point) is given by (0.0001, 0.01), showing the upper bound is 100 times larger than the lower bound, indicating $p$-values 0.01 and 0.001 are within the range of variability and have no significant difference. In other words, the statistical decision based on significance level 0.1% could be a random decision, and has no reproducibility even if $n=200$.

4. Reproducibility of statistical test results based on $p$-values

We consider sample sizes needed to assure the reproducibility of a statistical test results based on $p$-values.

\begin{table}[h]
\centering
\caption{Values of lower and upper quartiles, interquartile range and interquartile ratio of the distribution of $p$-value when $n = 20, 50, 100, 200$; $\delta = 0.1, 0.2, 0.3$.}
\begin{tabular}{|c|ccc|ccc|}
\hline
 & \multicolumn{3}{c|}{$n = 20$} & \multicolumn{3}{c|}{$n = 50$} \\
 & $\delta = 0.1$ & $\delta = 0.2$ & $\delta = 0.3$ & $\delta = 0.1$ & $\delta = 0.2$ & $\delta = 0.3$ \\
\hline
75\% point & 0.638 & 0.515 & 0.390 & 0.567 & 0.371 & 0.203 \\
25\% point & 0.162 & 0.096 & 0.053 & 0.121 & 0.047 & 0.015 \\
interquartile range & 0.476 & 0.419 & 0.337 & 0.446 & 0.324 & 0.188 \\
interquartile ratio & 3.9 & 5.3 & 7.4 & 4.7 & 7.8 & 13.5 \\
\hline
 & \multicolumn{3}{c|}{$n = 100$} & \multicolumn{3}{c|}{$n = 200$} \\
 & $\delta = 0.1$ & $\delta = 0.2$ & $\delta = 0.3$ & $\delta = 0.1$ & $\delta = 0.2$ & $\delta = 0.3$ \\
\hline
75\% point & 0.485 & 0.228 & 0.073 & 0.371 & 0.092 & 0.010 \\
25\% point & 0.084 & 0.019 & 0.003 & 0.047 & 0.004 & 0.000 \\
interquartile range & 0.401 & 0.209 & 0.070 & 0.324 & 0.088 & 0.010 \\
interquartile ratio & 5.8 & 12.3 & 27.9 & 7.8 & 24.2 & 81.7 \\
\hline
\end{tabular}
\end{table}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3.png}
\caption{Interquartile range and interquartile ratio}
\end{figure}
4.1 Definition of reproducibility

To begin with, we define the statistical test based on $p$-value for testing $H_0: \Delta = 0$ vs. $H_1: \Delta = \delta$ ($\delta > 0$).

**DEFINITION 4.1** Suppose that $\alpha$ ($0 < \alpha < 1$) is a given constant. We call the statistical test such that:

- If $p \leq \alpha$, then decide “treatment effect is significance”,
- if $p > \alpha$, then decide “no enough evidence to decide the treatment effect is significance”

the **statistical test based on $p$-value with the level of significance $\alpha$**.

**DEFINITION 4.2** Suppose that we repeat the test again with the same level of significance as the first test, independently to the first test, if and only if the first test results significance. If the second test leads the same result as the first test, we call the statistical test result has **reproducibility**.

**DEFINITION 4.3** Denote by $p$ and $p_2$ the $p$-values of the first and second statistical test, respectively. Then we call the conditional probability of $p_2 \leq \alpha$ given $p \leq \alpha$, namely $Pr(p_2 \leq \alpha | p \leq \alpha)$, the **reproducibility probability** of the statistical test based on $p$-value with the level of significance $\alpha$.

Since the definition of reproducibility includes the statistical independence of $p$ and $p_2$ it follows the following proposition.

**PROPOSITION 4.1** The reproducibility probability of statistical test with the level of significance $\alpha$ is given by $Pr(p_2 \leq \alpha)$.

4.2 Sample size for the reproducibility

Reproducibility probability $Pr(p_2 \leq \alpha)$ depends on $n$ and $\delta$. We have the following proposition. The proofs are given in Appendix.

**PROPOSITION 4.2** Assume the framework in Corollary 2.1. Then

1. the reproducibility probability of statistical test with the level of significance $\alpha$ for testing $H_0$: $\Delta = 0$ vs. $H_1$: $\Delta = \delta$ ($\delta > 0$) is an increasing function of $n$ for any fixed $\delta$, and
2. the reproducibility probability of statistical test with the level of significance $\alpha$ for testing $H_0$: $\Delta = 0$ vs. $H_1$: $\Delta = \delta$ ($\delta > 0$) is an increasing function of $\delta$ for any fixed $n$.

To assure the reproducibility of a statistical test with the level of significance $\alpha$, we may specify a large $\gamma_0$ ($0 < \gamma_0 < 1$), for example 0.80, and then decide $n$ to satisfy

$$Pr(p_2 \leq \alpha) = \gamma_0$$  \hspace{1cm} (8)

for given $\delta$ from Proposition 4.2 (1). We have the following proposition, whose proof is given in Appendix.

**PROPOSITION 4.3** Assume the framework in Corollary 2.1. Then, when $\delta$ is a given constant, sample size $n$ that satisfies (8) is obtained as follows.
where \( Z_{\alpha} \) is the \( a \times 100\% \) upper percent point of the standard normal distribution.

To apply Proposition 4.3 in practice under the set up given in Corporally 2.1, we need to specify the value of \( \delta \). It may be done as follows.

Let \( \delta_0 \) be the least value of \( \delta (> 0) \) that the treatment effect is considered effective, then it follows from Proposition 4.2 (2) that the reproducibility probability with respect to \( \delta \) is larger than that with respect to \( \delta_0 \). Therefore, we may specify \( \delta \) in Corporally 2.1 as \( \delta = \delta_0 \), since it is conservative.

### 4.3 Relationship with the sample size determined in the Neyman-Pearson type statistical test

Note that sample size formula given in Proposition 4.3 is identical to that given in the Neyman-Pearson type test of testing statistical hypothesis (see, for example, Armitage and Colton 2005, p. 4696). It holds true in general two-sample problems; for example, it holds true when \( \sigma \) is an unknown constant in the framework of Proposition 2.1. We show it in this section.

Consider two-sample problems for testing

\[
H_0 : \Delta = 0 \text{ vs. } H_1 : \Delta = \delta \quad (\delta > 0)
\]

based on statistic \( T \).

In the Neyman-Pearson type test for testing \( H_0 \) vs. \( H_1 \), sample size is determined by specifying the level of significance and power of the test. Namely, letting \( \alpha_1 \) be the level of significance and \( \gamma_1 \) be the pre-specified power of the test, sample size \( n \) is determined by solving

\[
Pr (T > C | \Delta = 0) = \alpha_1, \quad Pr (T > C | \Delta = \delta) = \gamma_1
\]

for a given constant \( \delta \). We have the following proposition. The proof of the proposition is given in Appendix.

**PROPOSITION 4.4** Assume that \( \alpha_1 = \alpha \) and \( \gamma_1 = \gamma_0 \) where \( \alpha \) and \( \gamma_0 \) are parameters introduced in the previous subsection for computing sample sizes to assure reproducibility probability of \( p \)-values. Then if \( \delta = \delta_0 \) the sample size \( n \) given by equations (9) is equivalent to \( n \) given in (8).

The proposition shows that, when \( \delta \) is given, keeping power of Neyman-Pearson type test is equivalent to keeping reproducibility of statistical test result based on \( p \)-value.

### 5. Discussion

Statisticians have long been satisfied with the sample size determination by using the power of statistical test; no question has asked about the relationship of the power of statistical test and the reproducibility of research findings. It is shown in this paper that to keeping the power of Neyman-Pearson type test at given level is equivalent to keeping the reproducibility of statistical
test result based on $p$-value at the same level.

Not to mention, but the reproducibility of research finding is a crucial concern of scientists in scientific research. A finding is established by statistical test, meaning that the reproducibility of test results based on $p$-value is a key stone of the reproducibility of research finding. Therefore, it might be more natural and straightforward to teach scientists, and students as well, the importance of designing sample sizes in scientific research from the view point of reproducibility of statistical test result than teach it by using the power of statistical test.

An early concept of $p$-value is seen in the 1710 Arbuthnot paper that compared the ratios of male and female birth. The fact indicates the concept of $p$-value was natural, straightforward and accepted in the scientific society 300 years ago. It is unsophisticated and easy to be grasped by people living today, too. On the other hand, the formulation of testing statistical hypothesis by Neyman-Pearson is mathematically beautiful, but sophisticated and error-prone; actually, it is reported that 51% of 791 scientific papers committed errors in the interpretation of statistical significance (Amrhein, Greenland and McShane, 2019).

It might be better to give up the teaching of Neyman-Pearson type test in elementary classes in statistics, together with the sample size determination using the power of statistical test that was developed in the framework of Neyman-Pearson type test. Alternatively, we encourage the teaching of unsophisticated Fisher’s statistical inference based on $p$-value and the importance of designing sample sizes in scientific research from the view point of reproducibility of statistical test results based on $p$-values. It could be useful to avoid the abuse of $p$-values, issued recently in the Statement of the ASA (Wasserstein, R.L. and Lazar, N. A., 2016) and also in the paper in Nature by Amrhein, Greenland, McShane and more than 800 signatories (Nature, 2019).

Acknowledgements
The author express his deep gratitude to the Editor and reviewers of this paper for their valuable comments which were non-dispensable for improving the early version of this paper.

REFERENCES
Arbuthnot, J. (1710). An argument for Divine Providence, taken from the constant regularity observed in the births of both sexes, Philosophical Transactions of the Royal Society of London, 27 (325–336): 186–190.
Armitage, P. and Colton, T. (eds), (2005). Encyclopedia of Biostatistics, Second edition, Wiley.
Amrhein, V., Greenland, S., McShane, B. (2019). Scientists rise up against statistical significance. Nature 567 (7748), 305–307.
Bhattacharya, B., and Habtzghid, D. (2002). Median of the $p$ value under the alternative hypothesis. The American Statistician, 56, 201–206.
Donahue R.M.J. (1999). A note on information seldom reported via the p value. The American Statistician, 53, 303–306.

Fisher, R. A. (1925). Statistical Methods for Research Workers, Oliver and Boyd.

Hung, H.M.J., O’Neill, R.T., Bauer, P., and Kohne, K. (1997). The behavior of the p-value when the alternative hypothesis is true. Biometrics, 53, 11–22.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, Philos. Mag. 50, 157–175.

Lehmann, E.L. (1986). Testing Statistical Hypotheses, John Wiley and Sons, Inc.

Neyman, J. and Pearson, E.S. (1928). On the use and interpretation of certain test criteria. Biometrika 20 A, 175–240.

Ono, Y. (2018). On Fisher’s significance test, unpublished monograph (in Japanese).

Sackrowitz, H., and Samuel-Cahn, E. (1999). P values as random variables - Expected p values. The American Statistician, 53, 326–331.

Tukey, J.W. (1977). Exploratory Data Analysis, Addison-Wesley.

Wasserstein RL, Lazar NA. (2016): Editorial: The ASA’s statement on p-values: Context, process, and purpose.

The American Statistician 70, 129–133.

Yanagawa, T. (2019). p-value (in Japanese), Kindai Kagakusha.

APPENDIX

Proof of PROPOSITION 4.3 : Formula (8) may be represented as follows.

\[ Pr\left(\hat{p} \leq \alpha \mid \delta\right) = \gamma_{p} \]

(10)

Since the framework in Corollary 2.1 is assumed we have \( \hat{p} = 1 - \Phi(t_{0}) \), and \( t_{0} \) is a realized value of \( T \), we may represent

\[ Pr\left(\hat{p} \leq \alpha \mid \delta\right) = Pr\left(1 - \Phi(T) \leq \alpha \mid \delta\right) \]

\[ = Pr\left(T \geq \Phi^{-1}(1 - \alpha) \mid \delta\right). \]

Therefore from (10)

\[ Pr\left(T \geq \Phi^{-1}(1 - \alpha) \mid \delta\right) = \gamma_{p} \]
Furthermore, since
\[ Pr \left( T \leq x \bigg| \delta \right) = \Phi (x - \mu_n), \]
where \( \mu_n = \sqrt{n/2} \delta \). It follows that
\[ \Phi^{-1}(1-\alpha) - \mu_n = \Phi^{-1}(1-\gamma_0). \]
Thus from \( \mu_n = \sqrt{n/2} \delta \) we have
\[ n = 2 \left( \Phi^{-1}(1-\alpha) - \Phi^{-1}(1-\gamma_0) \right)^2 / \delta^2. \]
Since \( \Phi^{-1}(1-\alpha) = Z_\alpha \) and \( \Phi^{-1}(1-\gamma_0) = Z_{\gamma_0} \), we have the desired result.

**Proof of Proposition 4.4:** Two equations given in (9) may be represented as follows.
\[ F_\circ(C) = 1 - \alpha, \quad F_\circ(C) = 1 - \gamma_0 \]
Thus the equations reduce to
\[ F_1 \left( F_\circ^{-1}(1-\alpha) \right) = 1 - \gamma_0. \quad (11) \]
On the other hand we have
\[ p = Pr \left( T > t_0 \bigg| H_0 \right) = 1 - F_\circ(t_0). \]
Since \( t_0 \) is a realization of \( T \) we have from (8)
\[ \gamma_0 = Pr \left( \phi \leq \alpha \bigg| H_1 \right) = Pr \left( 1 - F_\circ(T) \leq \alpha \bigg| H_1 \right) \]
\[ = 1 - F_1 \left( F_\circ^{-1}(1-\alpha) \right), \]
namely
\[ F_1 \left( F_\circ^{-1}(1-\alpha) \right) = 1 - \gamma_0. \quad (12) \]
Therefore, sample size \( n \) determined by equation (11) is identical to the sample size \( n \) determined by equation (12), if \( \alpha_1 = \alpha \) and \( \gamma_1 = \gamma_0 \). This completes the proof of Proposition 4.4.