Taylor & Francis
Taylor & Francis Group

Check for updates

Bias-corrected estimators for proportion of true null hypotheses: application of adaptive FDR-controlling in segmented failure data

Aniket Biswas [a], Gaurangadeb Chattopadhyay[b] and Aditya Chatterjee[b]

[a]Department of Statistics, Dibrugarh University, Dibrugarh, India; [b]Department of Statistics, University of Calcutta, Kolkata, India

ABSTRACT

Two recently introduced model-based bias-corrected estimators for proportion of true null hypotheses (π_0) under multiple hypotheses testing scenario have been restructured for random observations under a suitable failure model, available for each of the common hypotheses. Based on stochastic ordering, a new motivation behind formulation of some related estimators for π_0 is given. The reduction of bias for the model-based estimators are theoretically justified and algorithms for computing the estimators are also presented. The estimators are also used to formulate a popular adaptive multiple testing procedure. Extensive numerical study supports superiority of the bias-corrected estimators. The necessity of the proper distributional assumption for the failure data in the context of the model-based bias-corrected method has been highlighted. A case-study is done with a real-life dataset in connection with reliability and warranty studies to demonstrate the applicability of the procedure, under a non-Gaussian setup. The results obtained are in line with the intuition and experience of the subject expert. An intriguing discussion has been attempted to conclude the article that also indicates the future scope of study.

1. Introduction

The current work considers a segmented failure dataset, where failure time or some similar entity of a particular component is available for a number of units but the units are operated or tested in different conditions, that may vary over space and time. Thus, the dataset is divided into several segments and the observations are available for each segment. The number of observations per segment (in order of tens or hundreds) might be much less than the number of segments (in order of hundreds or thousands), as the segmentation is done on the basis of time and space among other things. Thus the situation is quite similar to that of microarray datasets where thousands of genes are tested to identify differentially expressed genes based on gene expression levels of two small groups of subjects, viz.

---

treatment group and control group. For segmented failure dataset, similar kind of questions may arise regarding identification of segment(s) for which the failure patterns of that particular component is strikingly different (worse or better) from a benchmark, say the average. To answer this question, appropriate hypotheses for each segment are framed and tested simultaneously. While testing a large number of hypotheses, a control over the false discovery rate (FDR) [1] is desirable and the classical Benjamini–Hochberg algorithm [1] might be employed to achieve it. However, the power of Benjamini–Hochberg algorithm or any general step-up procedure can be improved by incorporating a conservative estimate of the proportion of true null hypotheses (π_0) or equivalently the number of the true null hypotheses [2,20].

Gaussian model assumption for failure data is inappropriate, especially when sample size corresponding to each segment or equivalently each test is small. On the contrary, in such situation, exponential distribution may be a reasonable primary model choice. Under such exponential setup, we modify both the estimators proposed in Cheng *et al.* [6] and Biswas [3] and find that these model-based estimators are more efficient than the existing π_0-estimators in practice. Application of adaptive Benjamini–Hochberg procedure has the ability to list the significantly different segments with respect to such time to event or equivalent entity of a certain component in our case study. For microarray datasets, the model-based approach is well established, especially under normality assumption [3,6]. In this article, we adapt and implement the same under the exponential model to a segmented failure data, where such model assumption is appropriate and an alternative model formulation may not be satisfactory to the desired extent. In what follows we introduce the parameter π_0 through the empirical Bayesian setup given in Storey [22].

Consider m similar but independent hypotheses are to be tested, viz. $H_1, H_2, \ldots, H_m$. For $H_i = 1$, the ith null hypothesis is true and for $H_i = 0$, false for any $i \in \{1, 2, \ldots, m\}$. Thus, H_i's are Bernoulli random variables with success probability $\pi_0 \in (0, 1)$. Let m_0 be the number of true null hypotheses. Thus, $m_0 = \sum_{i=1}^{m} H_i$ is a binomial random variable with index m and parameter π_0. Clearly, H_i's and hence m_0 remain latent and can never be realized in a given multiple testing scenario. As in case of single hypothesis testing problem, the test statistics $T_1, T_2, \ldots, T_m$, respectively, for $H_1, H_2, \ldots, H_m$ may be observed. For F_0 being the common distribution of $T_i | H_i = 1$ and F_1 being the same for $T_i | H_i = 0$, a two-component mixture model for T_i is

$$T_i \sim \pi_0 F_0 + (1 - \pi_0) F_1 \quad \text{for all } i = 1, 2, \ldots, m. \tag{1}$$

Thus, π_0 may be thought of as the mixing proportion of the null test statistics with the non-null test statistics when multiple tests are performed. In existing literature p-values are considered as test statistics since its use ensures similar nature of critical region, irrespective of the nature of hypotheses framed. Usually, a little abuse of notation is made while denoting p-value by p irrespective of whether it is a random variable or a realization on that. The distinction of usage ought to be understood as the situation demands. The marginal density function of p-value [13] is

$$f(p) = \pi_0 f_0(p) + (1 - \pi_0) f_1(p) \quad \text{for } 0 < p < 1, \tag{2}$$

where f_0 and f_1 are two p-value densities, respectively, under the null and alternative hypotheses. When the tested null is simple and the corresponding test statistic is absolutely continuous, $f_0(p)$ is simply 1, the density function of a uniform random variable

over $(0, 1)$ and the p-value under the alternative hypothesis is stochastically smaller than the uniform variate. In addition, the density estimation-based approaches for estimating π_0 impose certain restrictions on f_1 [9,13,17]. Often p-values under the alternative are modelled by parametric distributions [15,19] and π_0 is estimated using maximum likelihood methods. This requires the p-values to be independent among themselves which is rarely satisfied. Storey's estimator [22] is constructed on the basis of a tuning parameter $\lambda \in (0, 1)$ such that, $f_1(p) = 0$ for $p > \lambda$. This assumption introduces a conservative bias in the estimator that can be corrected or in practice, can be reduced as have been discussed in Cheng *et al.* [6]. The setup given therein for the applicability of the Gaussian model-based bias correction is discussed in Section 2. Biswas [3] has recently proposed an alternative model-based bias-corrected estimator for π_0 under the same setup. A comparative performance study of both the estimators with simulated microarray datasets has also been provided. There are several other works on estimation of π_0, not directly related to the current work. The interested readers are referred to Storey and Tibshirani [23], Wang *et al.* [26] and Tong *et al.* [25].

The remaining part of the article is structured as follows. In Section 2, we reproduce Storey's estimator and the recently introduced bias-corrected estimators from stochastic ordering approach which ties them in a yarn and may inspire further works in similar line. The next section is devoted to different testing scenarios and useful properties of respective non-null p-values. In Section 4, we briefly revisit the estimation algorithms and discuss adaptation of the π_0 estimates to Benjamini–Hochberg algorithm. Section 5 deals with performance comparison of the new estimators with existing ones through extensive simulation experiment. In Section 6, a real-life synthetic segmented failure dataset is presented, has been validated for applicability of the proposed methods, analysed to demonstrate the superior performance of adaptive algorithm with the new estimators along with proper justification of the findings. We conclude the article with a mention of a few limitations of the present work and a glimpse of the future direction of the study.

2. Methods of estimation

Let p denote a p-value corresponding to a simple null hypothesis testing problem with continuous test statistic. Thus, p has the support $(0, 1)$. Consider another random variable V on the same support $(0, 1)$ with the distribution function G. Then,

$$P(p \geq V) = \int_0^1 f(p)G(p)\, \mathrm{d}p. \tag{3}$$

In the following subsections, we take different choices for G and motivate different estimators for π_0 as mentioned in Section 1.

2.1. Storey's bootstrap estimator and related approaches

Consider V to be degenerate at some $\lambda \in (0, 1)$. Thus,

$$G(v) = \begin{cases} 1 & \text{for } v \geq \lambda \\ 0 & \text{for } v < \lambda. \end{cases} \tag{4}$$

Putting (2) and (4) in (3), we obtain

$$\bar{F}(\lambda) = \pi_0 (1 - \lambda) + (1 - \pi_0) Q(\lambda) \tag{5}$$

where F is the distribution function of p, $\bar{F} = 1 - F$ and Q is the survival function of non-null p-value. Assume,

- A1: For an appropriate choice of λ, $Q(\lambda) = P(p > \lambda | H = 0)$ i.e the probability of non-null p-value being greater than λ equals zero [22].

When parameter of interest under alternative hypothesis is substantially far from the same specified under null hypothesis or sample size is moderate to large, p-value tends to be smaller for consistent tests. Hence, even for moderate choice of λ, the probability of p-value under false null dominating λ, vanishes. This is a reasonable but crucial assumption in a sense that, violation of assumptions regarding the true value of the parameter of interest and sample size may not result in $Q(\lambda) = 0$. Thus, applying A1 in (5) we get

$$\pi_0 = \frac{\bar{F}(\lambda)}{(1 - \lambda)}. \tag{6}$$

Let $p_1, p_2, \ldots, p_m$ be the p-values corresponding to the m hypotheses tested or equivalently m realizations on p. Denote $W(\lambda) = \sum_{i=1}^{m} I(p_i > \lambda)$ (I denoting the indicator function) to be the number of p-values greater than λ. Putting the plug-in estimator of $\bar{F}(\lambda)$, i.e $W(\lambda)/m$ in (6), an estimator for π_0 depending upon the choice of λ may be suggested as

$$\hat{\pi}_0(\lambda) = \frac{W(\lambda)}{m (1 - \lambda)}. \tag{7}$$

For a given dataset, two different choices of λ would yield two different estimates and thus an optimum choice of λ for a given dataset is necessary. For a subjectively chosen set with possible values of $\lambda \in \Lambda$, where $\Lambda = \{0, 0.05, 0.10, \ldots, 0.95\}$; a bootstrap routine is given in Storey [22] and Storey et al. [24] to approximate the best λ. Thus, Storey's bootstrap estimator is: $\hat{\pi}_0^B = \hat{\pi}_0(\lambda_{best})$. In Storey and Tibshirani [23], natural cubic spline has been fitted to the $(\lambda, \hat{\pi}_0(\lambda))$ curve for smoothing and the evaluated value of the fit at $\lambda = 1$ (as motivated in Corollary 1 of [22]) is taken as the final estimate which we denote by $\hat{\pi}_0^P$.

For a small choice of λ in $\hat{\pi}_0(\lambda)$, the bias of the estimator is large while the variance is small. The situation is exactly opposite for large λ. It has been first noted by Jiang and Doerge [12] and they have suggested the use of multiple λ's instead of a single best choice, in some sense. For the time being assume a fixed set $S_\lambda = \{(\lambda_1, \lambda_2, \ldots, \lambda_k) : 0 < \lambda_1 < \lambda_2 < \cdots < \lambda_k < 1\}$ for a fixed k and equal width given by $(\lambda_{i+1} - \lambda_i)$ for $i = 1, 2, \ldots, k - 1$ such that A1 holds. Then, the average estimate based approach suggests $\hat{\pi}_0^A = (1/k) \sum_{i=1}^{k} \hat{\pi}_0(\lambda_i)$ to be an appropriate estimator for π_0. The authors have also suggested a change-point based algorithm to select S_λ.

2.2. Bias correction of Storey's estimator

Without the assumption A1, from (5), we get

$$\pi_0 = \frac{\bar{F}(\lambda) - Q(\lambda)}{(1 - \lambda) - Q(\lambda)} \tag{8}$$

for fixed λ. Cheng *et al.* [6] obtained (8) from a somewhat different motivation. Substituting plug-in estimator of $\bar{F}(\lambda)$ has already been discussed in Section 2.1. For estimating $Q(\lambda)$ following assumptions are necessary.

- A2: The availability of a common test for all the m hypotheses.
- A3: The data-arrays used for each test are generated from a known parametric family.
- A4: The closed form distribution of the test-statistics under the null are of a known family, enabling the calculation of the exact p-values.
- A5: The distribution of the non-null test-statistics and hence the non-null p-values are labelled by unknown effect sizes, which are different for each test.

A2 is generally true for microarray experiments and is also appropriate for the present setup. Cheng *et al.* [6] assumed normality for each expression level, such that A3 is valid. Time to events or its equivalent entities for each segment are assumed to be exponentially distributed, thus satisfying A3. Under normality, the test-statistics for usual single-sample or two-sample tests for the mean are normal under null. In this work, the test-statistics for single-sample test related to the exponential rate parameter is a χ^2 variate under null and for a two-sample problem the test-statistic is distributed as a F variate. Thus, A4 also holds good. As mentioned earlier test for H_i is performed by T_i and we introduce the notation δ_i to denote the effect size of the corresponding test. Non-null distribution of T_i and hence the non-null distribution of p_i is to be labelled by δ_i, $i = 1, 2, \ldots, m$. Hung *et al.* [11] have discussed properties of non-null p-values, where non-null distribution of the p-value for Z-test has been explored. For single sample and two sample t-test, similar discussion is available in Section 3 of Cheng *et al.* [6]. We will discuss such properties of non-null p-value for single and two sample problems under exponential setup in Section 3.

Let $\mathcal{I} = \{1, 2, \ldots, m\}$. Also let $\mathcal{T}$ denote the set of indices corresponding to the originally true null hypotheses i.e, $\mathcal{T} = \{i \in \mathcal{I} : H_i = 1\}$. Thus, the cardinality of $\mathcal{T}$, is m_0. Similarly denote the set of originally false null hypotheses by $\mathcal{F}$. Clearly, $\mathcal{F} = \mathcal{I} - \mathcal{T}$ with cardinality m_1. Each null p has the same distribution, uniform over $(0, 1)$; while the distribution of non-null p-values are different but they belong to the same family. Let $f_1^\delta(p)$ denote the distribution of p with effect size δ. Then for all $i \in \mathcal{F}$, $Q_{\delta_i}(\lambda) = \int_\lambda^1 f_1^{\delta_i}(p)\,dp$, probability of ith non-null p-value being greater than λ. Define, $Q(\lambda) = (1/m_1) \sum_{i \in \mathcal{F}} Q_{\delta_i}(\lambda)$, the average of non-null p-values greater than λ. To estimate $Q(\lambda)$, individual δ_i's are estimated by $\hat{\delta}_i$, $i \in \mathcal{F}$. In fact, $\hat{\delta}_i$'s are strongly consistent for δ_i for each $i \in \mathcal{F}$. The estimation of δ under different testing problem is discussed in Section 3. Each $Q_{\delta_i}(\lambda)$ is continuous in δ_i and thus, $Q_{\hat{\delta}_i}(\lambda)$ is strongly consistent for $Q_{\delta_i}(\lambda)$. Thus, a strongly consistent estimator for $Q(\lambda)$ is $\tilde{Q}(\lambda) = (1/m_1) \sum_{i \in \mathcal{F}} Q_{\hat{\delta}_i}(\lambda)$. In practice, $\mathcal{F}$ is unknown and hence $\tilde{Q}(\lambda)$ is unavailable. Assume $\hat{Q}(\lambda)$ to be a dummy for $\tilde{Q}(\lambda)$ such that $\tilde{Q}(\lambda) \geq \hat{Q}(\lambda)$ with probability 1. The computation of $\hat{Q}(\lambda)$ is discussed in detail in Section 4. Substituting the plug-in estimators for $\bar{F}(\lambda)$ and $Q(\lambda)$ in (8), we get $\tilde{\pi}_0^U(\lambda)$ (or $\hat{\pi}_0^U(\lambda)$), bias-corrected estimator for π_0 with fixed choice of λ. We now address the issues related to reduction in bias and over-correction in the following result.

Result 2.1: With the setup and notations introduced in Section 2.2, for all $\lambda \in (0, 1)$

     (a)   For $W(\lambda)/m \leq (1 - \lambda)$, $\tilde{\pi}_0^U(\lambda) \leq \hat{\pi}_0^U(\lambda) \leq \hat{\pi}_0(\lambda)$.

     (b) $\tilde{\pi}_0^U(\lambda) \rightarrow \pi_0$, almost surely.

Result 2.1 combines claims written in Sections 2 and Section 4.2 of Cheng *et al.* [6]. We have been able to prove Result 2.1 in a more direct way. Thus, the approach reduces conservative bias of Storey's primary estimator while refraining from over-correction.

     The situations $\hat{\pi}_0(\lambda) \leq 1$ and $\tilde{Q}(\lambda) \leq (1 - \lambda)$ are quite usual in multiple testing setup as the first one is a reasonable estimate of π_0 and $\tilde{Q}(\lambda)$ is a consistent estimate of $Q(\lambda)$, which is obviously less than $(1 - \lambda)$. If these do not hold good, $\hat{\pi}_0^U(\lambda)$ lies outside the parameter space and then we take the estimate to be the nearest boundary point.

     $\Lambda = \{0.20, 0.25, \ldots, 0.5\}$ is taken as in Jiang and Doerge [12] for similar purpose (see Section 2.2 in [6]) and we identify the following estimator as the bias and variance reduced estimator for π_0:

$$\hat{\pi}_0^U = \frac{1}{\#\Lambda} \sum_{\lambda_j \in \Lambda} \min\{1, \max\{0, \hat{\pi}_0^U(\lambda_j)\}\},$$

where $\#\Lambda$ denotes cardinality of Λ.

2.3. Estimator based on sum of all p-values

Instead of taking V degenerated at some fixed λ, assume $V \sim Uniform(0, 1)$. Putting $G(v) = v$ for $v \in (0, 1)$ in (3), we get

$$P(p \geq V) = \int_0^1 p f(p) \, dp = E(p) = \frac{\pi_0}{2} + (1 - \pi_0) e \tag{9}$$

since $p | H = 1 \sim Uniform(0, 1)$. In (9), we use e to denote expectation of non-null p-value: $e = E(p|H = 0)$. From (9), we get

$$\pi_0 = \frac{E(p) - e}{0.5 - e}. \tag{10}$$

To estimate π_0, both $E(p)$ and e are to be estimated. $E(p)$ can be estimated by the mean of observed p-values: $\bar{p} = (1/m) \sum_{i=1}^m p_i$. Define $e = (1/m_1) \sum_{i \in \mathcal{F}} e_{\delta_i}$. The average of expected p-values under the alternative, e can be estimated imitating the approach of estimating $Q(\lambda)$ with assumptions A2-A5. The corresponding estimator for π_0 has recently been introduced in Biswas [3] and computation of $e_{\delta_i} = E(p_i|i \in \mathcal{F})$ has been demonstrated for single and two sample t-tests therein. Since each e_{δ_i} is bounded and continuous in δ_i, following the discussion in Section 2.2, a strongly consistent estimator for e is $\tilde{e} = (1/m_1) \sum_{i \in \mathcal{F}} e_{\hat{\delta}_i}$, which cannot be realized in practice for obvious reason mentioned earlier and hence $\tilde{\pi}_0^E = (\bar{p} - \tilde{e})/(0.5 - \tilde{e})$ cannot be implemented. For $\hat{e}$ being a dummy of $\tilde{e}$ with $\hat{e} \leq \tilde{e}$ almost surely, an workable estimator for π_0 is $\hat{\pi}_0^E = (\bar{p} - \hat{e})/(0.5 - \hat{e})$.

Result 2.2: With the setup and notations introduced in Section 2.3

$$(a) \quad \text{For } \bar{p} \leq 0.5, \; \tilde{\pi}_0^E \leq \hat{\pi}_0^E.$$

$$(b) \quad \tilde{\pi}_0^E \; \rightarrow \; \pi_0, \text{ almost surely.}$$

The situations $\bar{p} \leq 0.5$ and $\tilde{e} \leq 0.5$ are very natural in multiple testing setup as $\bar{p}$ is consistent for $E(p)$, which is less than 0.5 and similarly $\tilde{e}$ is consistent for e which is also less than 0.5. If theses do not hold good, $\hat{\pi}_0^E$ lies outside the parameter space and then we take the estimator for π_0 as

$$\hat{\pi}_0^E = \min\left\{1, \max\left\{0, \frac{\bar{p} - \hat{e}}{0.5 - \hat{e}}\right\}\right\}.$$

Both the model-based bias-corrected estimators are shown to have conservative bias for estimating π_0. In Cheng *et al.* [6], $\hat{\pi}_0^U$ has been shown to outperform the robust estimators under reasonable model assumption, whereas under similar situation $\hat{\pi}_0^E$ outperforms, it in terms of mean square error, as empirically studied in Biswas [3] through extensive simulation study. Note that, both the estimators use an initial estimator for π_0 but the computation of $\hat{\pi}_0^E$ does not require flexible threshold tuning parameters owing to the fact that it uses all the p-values. To rule out the possibility of estimates taking value outside the parameter space under very unusual situation, $\hat{\pi}_0^E$ is taken to be equal to the nearest boundary point when it lies outside the parameter space. Proof of the results presented in this section are provided in the Appendix.

3. Properties of non-null *p*-values

To implement the bias-corrected estimators $\hat{\pi}_0^U$ and $\hat{\pi}_0^E$, appropriate estimates of the unknown quantities $Q(\lambda)$ and e are needed. To get explicit expressions for these quantities, we need to have the probability density functions $f_1^{\delta_i}(p)$ (for notational convenience we write this to be $f_{\delta_i}(p)$ henceforth) for each non-null p-value with effect size δ_i, $i \in \mathcal{F}$. The subscript i in effect sizes are not specified in this section for ease of notation. Thus, for different testing scenarios, we determine the probability density function $f_\delta(p)$, then $Q_\delta(\lambda)$ by integrating $f_\delta(p)$ from λ to 1 and finally obtain e_δ through the following results. As discussed in Section 2.2, $Q_\delta(\lambda)$ for fixed λ and e_δ are continuous in δ under each of the testing problems considered here.

Result 3.1: Assume $X_1, X_2, \ldots, X_n$ be a random sample of size n from an exponential distribution with mean θ. Consider the following testing problem:

$$H_0 : \theta = 1 \quad \text{versus } H_0 : \theta > 1. \tag{11}$$

For the corresponding likelihood ratio test

$$(a) \quad \delta = \theta \quad \text{and thus } \hat{\delta} = \min\{1, \bar{X}\}$$

$$(b) \quad f_\delta(p) = \frac{\frac{1}{\delta} f_{\chi_{2n}^2}\left(\frac{1}{\delta}\chi_{p,2n}^2\right)}{f_{\chi_{2n}^2}\left(\chi_{p,2n}^2\right)} \quad \text{for } 0 < p < 1$$

$$(c) \quad Q_\delta(\lambda) = F_{\chi^2_{2n}}\left(\frac{1}{\delta}\,\chi^2_{\lambda,2n}\right) \quad \text{for } 0 < \lambda < 1$$

$$(d) \quad e_\delta = E_{X \sim \chi^2_{2n}}\left[1 - F_{\chi^2_{2n}}\left(\frac{X}{\delta}\right)\right].$$

Here $f_{\chi^2_\nu}$, $F_{\chi^2_\nu}$ and $\chi^2_{p,\nu}$ denote the probability density function, the distribution function and the upper-p point of chi-square distribution with ν degrees of freedom, respectively.

Result 3.2: For $X_1, X_2, \ldots, X_n$ to be a random sample from exponential distribution with mean θ, consider the testing problem:

$$H_0 : \theta = 1 \quad \text{versus } H_1 : \theta \neq 1. \tag{12}$$

For the corresponding likelihood ratio test

$$(a) \quad \delta = \theta \quad \text{and thus } \hat{\delta} = \bar{X}$$

$$(b) \quad Q_\delta(\lambda) = F_{\chi^2_{2n}}\left[\frac{1}{\delta}\chi^2_{\frac{\lambda}{2},2n}\right] - F_{\chi^2_{2n}}\left[\frac{1}{\delta}\chi^2_{1-\frac{\lambda}{2},2n}\right] \quad \text{for } 0 < \lambda < 1$$

$$(c) \quad e_\delta = E_{X \sim \chi^2_{2n}(0,\mu)}\left[F_{\chi^2_{2n}}\left(\frac{X}{\delta}\right)\right] - E_{X \sim \chi^2_{2n}(\mu,\infty)}\left[F_{\chi^2_{2n}}\left(\frac{X}{\delta}\right)\right].$$

The notations used for stating Result 3.1 also remain relevant here. In addition to that, $\chi^2_\nu(a, b)$ denotes the truncated chi-squared distribution with degree of freedom ν and region of truncation being (a, b). Here μ denotes the median of χ^2_{2n} distribution.

Result 3.3: Let $X_1, X_2, \ldots, X_{n_1}$ and $Y_1, Y_2, \ldots, Y_{n_2}$ be two random samples of size n_1 and n_2, respectively, from exponential distribution with mean θ_1 and θ_2. Consider the testing problem

$$H_0 : \theta_2 = \theta_1 \quad \text{versus } H_1 : \theta_2 > \theta_1. \tag{13}$$

For the corresponding likelihood ratio test, we have the following.

$$(a) \quad \delta = \frac{\theta_2}{\theta_1} \quad \text{and} \quad \hat{\delta} = \min\left\{1, \frac{n_1 - 1}{n_1}\frac{\sum_{i=1}^{n_2} Y_i}{\sum_{i=1}^{n_1} X_i}\right\}$$

$$(b) \quad f_\delta(p) = \frac{\frac{1}{\delta}f_{F_{2n_2,2n_1}}\left(\frac{1}{\delta}F_{p,2n_2,2n_1}\right)}{f_{F_{2n_2,2n_1}}\left(F_{p,2n_2,2n_1}\right)} \quad \text{for } 0 < p < 1$$

$$(c) \quad Q_\delta(\lambda) = F_{F_{2n_2,2n_1}}\left(\frac{1}{\delta}F_{\lambda,2n_2,2n_1}\right) \quad \text{for } 0 < \lambda < 1$$

$$(d) \quad e_\delta = E_{X \sim F_{2n_2,2n_1}}\left[F_{F_{2n_2,2n_1}}\left(\frac{X}{\delta}\right)\right].$$

Here $f_{F_{\nu_1,\nu_2}}$, $F_{F_{\nu_1,\nu_2}}$ and F_{p,ν_1,ν_2}, respectively, denote the probability density function, the distribution function and the upper-p point of F distribution with ν_1 and ν_2 degrees of freedom.

Result 3.4: Consider $X_1, X_2, \ldots, X_{n_1}$ and $Y_1, Y_2, \ldots, Y_{n_2}$ to be independent random samples of size n_1 and n_2 from exponential distributions with mean θ_1 and θ_2, respectively. Consider the testing problem

$$H_0 : \theta_2 = \theta_1 \quad \text{versus } H_1 : \theta_2 \neq \theta_1. \tag{14}$$

For the corresponding likelihood ratio test, we have

$(a) \quad \delta = \dfrac{\theta_2}{\theta_1} \quad \text{and} \quad \hat{\delta} = \dfrac{n_1 - 1}{n_1} \dfrac{\sum_{i=1}^{n_2} Y_i}{\sum_{i=1}^{n_1} X_i}$

$(b) \quad Q_\delta(\lambda) = F_{F_{2n_2, 2n_1}} \left[\dfrac{1}{\delta} F_{\frac{\lambda}{2}, 2n_2, 2n_1} \right] - F_{F_{2n_2, 2n_1}} \left[\dfrac{1}{\delta} F_{1 - \frac{\lambda}{2}, 2n_2, 2n_1} \right]$

$(c) \quad e_\delta = E_{X \sim F_{2n_2, 2n_1}(0, \mu)} \left[F_{F_{2n_2, 2n_1}} \left(\dfrac{X}{\delta} \right) \right] - E_{X \sim F_{2n_2, 2n_1}(\mu, \infty)} \left[F_{F_{2n_2, 2n_1}} \left(\dfrac{X}{\delta} \right) \right].$

The notations used for Result 5 also remain relevant here. In addition to that, $F_{\nu_1, \nu_2}(a, b)$ denotes the truncated F-distribution with degrees of freedom ν_1, ν_2 and region of truncation (a, b). Here μ denotes the median of $F_{2n_2, 2n_1}$ distribution. Interested readers may find proof of the results in the Appendix.

4. Algorithms

Algorithm for computing $\hat{\pi}_0^U$ under normal model assumption is given in in Cheng *et al.* [6] and for $\hat{\pi}_0^E$ under same setup, see Biswas [3]. First, we reframe the algorithms under current setup to maintain readability and for making the proposed estimation methods readily available to the practitioners. For the sake of brevity we only consider the testing problem in Result 3.2 and use the corresponding non-null p-value properties here. For all the four situations discussed here, the following algorithms can be implemented with obvious modifications.

Algorithm 4.1 (For computing $\hat{\pi}_0^U$):
- For all $i = 1, 2, \ldots, m$, estimate δ_i by $\hat{\delta}_i = \bar{X}_i$.
- For all $i = 1, 2, \ldots, m$ and for each $\lambda_j \in \Lambda$; estimate the upper tail probability $Q_{\delta_i}(\lambda_j)$ by $Q_{\hat{\delta}_i}(\lambda_j)$ given by

$$Q_{\hat{\delta}_i}(\lambda_j) = F_{\chi^2_{2n_i}} \left[\dfrac{1}{\hat{\delta}_i} \chi^2_{\frac{\lambda_j}{2}, 2n_i} \right] - F_{\chi^2_{2n_i}} \left[\dfrac{1}{\hat{\delta}_i} \chi^2_{1 - \frac{\lambda_j}{2}, 2n_i} \right].$$

where, n_i denotes the available sample size for testing ith hypothesis.
- Using an available estimator of π_0 as initial estimator $\hat{\pi}_0^I$, calculate $d = [m \times (1 - \hat{\pi}_0^I)]$, where $[\ ]$ denotes the usual box function. Arrange $Q_{\hat{\delta}_i}(\lambda_j)$'s in increasing order and denote the ith quantity in the list as $\hat{Q}_{(i)}(\lambda_j)$. Thus a conservative estimator for $Q(\lambda_j)$ is

$$\hat{Q}(\lambda_j) = \dfrac{1}{d} \sum_{i=1}^{d} \hat{Q}_{(i)}(\lambda_j).$$

- Given $\hat{Q}(\lambda_j) \ \forall \ \lambda_j \in \Lambda$, calculate

$$\hat{\pi}_0^U = \frac{1}{\#\Lambda} \sum_{\lambda_j \in \Lambda} \min \left\{ 1, \max \left\{ 0, \frac{\frac{W(\lambda_j)}{m} - \hat{Q}(\lambda_j)}{(1 - \lambda_j) - \hat{Q}(\lambda_j)} \right\} \right\}.$$

Algorithm 4.2 (For computing $\hat{\pi}_0^E$):
- For all $i = 1, 2, \ldots, m$, estimate δ_i by $\hat{\delta}_i = \bar{X}_i$.
- For all $i = 1, 2, \ldots, m$, estimate the mean of non-null p-value e_{δ_i} by

$$\hat{e}_{\delta_i} = E_{X \sim \chi^2_{2n_i}(0, \mu_i)} \left[F_{\chi^2_{2n_i}} \left(\frac{X}{\hat{\delta}_i} \right) \right] - E_{X \sim \chi^2_{2n_i}(\mu_i, \infty)} \left[F_{\chi^2_{2n_i}} \left(\frac{X}{\hat{\delta}_i} \right) \right].$$

where n_i denotes the available sample size for testing ith hypothesis and μ_i denotes the median of $\chi^2_{2n_i}$ distribution.
- Using an initial estimator of π_0 as initial estimator $\hat{\pi}_0^I$, calculate $d = [m \times (1 - \hat{\pi}_0^I)]$, as before. Arrange $\hat{e}_{\delta_i}$'s in increasing order and denote the ith quantity in the list as $\hat{e}_{(i)}$. Thus a conservative estimator for e is

$$\hat{e} = \frac{1}{d} \sum_{i=1}^{d} \hat{e}_{(i)}.$$

- Given $\hat{e}$, calculate

$$\hat{\pi}_0^E = \min \left\{ 1, \max \left\{ 0, \frac{\bar{p} - \hat{e}}{0.5 - \hat{e}} \right\} \right\}.$$

Note 1: The role of $\hat{\pi}_0^I$ is important in obtaining $\hat{Q}(\lambda)$ and $\hat{e}$. For $\hat{\pi}_0^I \geq \pi_0$, observe that $m_1 \geq d$. Clearly, $\hat{Q}(\lambda) \leq \tilde{Q}(\lambda)$ and $\hat{e} \leq \tilde{e}$. Thus, $\hat{\pi}_0^U \geq \tilde{\pi}_0^U$ and $\hat{\pi}_0^E \geq \tilde{\pi}_0^E$.

Note 2: For implementation of both the algorithms, we choose Storey's bootstrap estimator, given by $\hat{\pi}_0^B$ as the initial estimator. This choice seems reasonable albeit being non-universal and further research on this is warranted. The algorithms could also be implemented with other choices of $\hat{\pi}_0^I$. The performance analysis of the bias-corrected estimators under the current setup requires extensive simulation study, starting with different choices of the initial estimator. In fact, the algorithms could in principle be done several times, each time with the estimate of π_0 from the previous iteration. Obviously, this technique will become computation intensive for all practical purposes. We refrain from addressing these issues, as they are beyond the scope of the current study.

It has already been mentioned in Section 1 that Benjamini–Hochberg procedure for controlling the FDR is conservative. To understand this, we briefly discuss FDR and the algorithm for controlling it at a prefixed level $q \in (0, 1)$. While testing m hypotheses simultaneously, let R be the total number of rejected hypotheses by application of certain multiple testing algorithm. From the entire set of rejected hypotheses, some hypotheses may be originally true. These are categorized as false discovery and let V denote the total number of such false discoveries. Then the false discovery proportion (FDP) is defined as

$$FDP = \begin{cases} \dfrac{V}{R} & \text{if } R > 0 \\ 0 & \text{if } R = 0. \end{cases}$$

Note that, prior to the application of any algorithm both V and R are random variables and the expected value of FDP is termed as FDR. Let $p_{(1)} \le p_{(2)} \le \cdots \le p_{(m)}$ be the ordered sequence of the available p-values. Benjamini–Hochberg procedure identifies the largest k such that $p_{(k)} \le (k/m)q$ and rejects all hypotheses with corresponding p-value less than $p_{(k)}$ along with the hypothesis with p-value $p_{(k)}$. This procedure is conservative, as the implementation of the same ensures $FDR = \pi_0 q$ where, $\pi_0 = m_0/m$. To overcome this shortcoming, Craiu and Sun [7] worked with the following adaptive Benjamini–Hochberg procedure which uses an approximation of π_0.

Algorithm 4.3 (Implementing adaptive BH procedure to control FDR at level q):
- Let the p-value corresponding to the problem of testing H_i be p_i for $i = 1, 2, \ldots, m$. Arrange the available p-values in increasing order: $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$. Denote the corresponding hypotheses by $H_{(i)} : i = 1, 2, \ldots, m$.
- Given the dataset, estimate π_0. Let it be $\hat{\pi}_0$.
- Compute the adjusted p-value corresponding to $p_{(i)}$:

$$adj.p_{(i)} = \min\left\{ \hat{\pi}_0 \frac{m\, p_{(j)}}{j} : j \ge i \right\} \quad \forall\, i = 1, 2, \ldots, m.$$

- For all $i = 1, 2, \ldots, m$, reject $H_{(i)}$ if $adj.p_{(i)} \le q$.

Both adaptive BH procedure and Storey's q-value approach are justified to be equivalent in Craiu and Sun [7]. They have also emphasized that both the approaches require a good approximation of π_0. Less conservative estimators for π_0 are in demand since closer approximation of π_0 will bring superiority in the adaptive procedure by increasing the number of rejections while controlling FDR at level q, as evident from Algorithm 4.3. In numerical study, we use *adjust.p( )* function (available in *cp4p* library from Bioconductor) by Gianetto *et al.* [8] for obtaining the adjusted p-values.

5. Simulation study

We have conducted an extensive simulation study to investigate the performance of the bias-corrected estimators under different settings. The well-known and established estimators apart from the proposed $\hat{\pi}_0^U$ and $\hat{\pi}_0^E$, considered for performance comparison are listed below.

$\hat{\pi}_0^B$: Storey's bootstrap estimator (discussed in Section 2.1)
$\hat{\pi}_0^L$: Convest estimator [13]
$\hat{\pi}_0^A$: Jiang and Doerge's average estimator (discussed in Section 2.1)
$\hat{\pi}_0^P$: Natural cubic spline smoothing-based estimator (discussed in Section 2.1)
$\hat{\pi}_0^H$: Histogram-based estimator [16]
$\hat{\pi}_0^D$: A robust estimator of π_0 [18]
$\hat{\pi}_0^S$: Sliding linear model based model-based estimator (Wang *et al.* [26]).

5.1. Simulation setting

We imitate a segmented time to event dataset to generate artificial datasets. For this purpose, we choose $m = 100, 500, 1000$ segments. Two different settings are considered, balanced setting with sample size for each segment being equal to 35 and unbalanced setting with different sample sizes 15, 25, 35, 45, 55 for equal number of segments. For fixed $\pi_0 = 0.1, 0.2, \ldots, 0.9$, calculate $m_0 = [m \pi_0]$ and take $m_1 = m - m_0$. We set the mean failure time under null θ_0 as 'unity'. For m_0 randomly chosen numbers from $\mathcal{I} = \{1, 2, \ldots, m\}$ we fix $\theta = \theta_0 = 1$ and for the remaining cells in the array θ, we generate values through some stochastic mechanism ensuring that they are not equal to θ_0. For segments with better average lifetime $\theta \sim \text{Uniform}(1, 1.5)$ and for segments with poor average lifetime $\theta \sim \text{Uniform}(0.5, 1)$. We take the proportion of better (or poor) non-null mean lifetimes to be 0.5.

After generating the array of parameter θ, we generate a sample of size n_i from the exponential distribution with mean θ_i for all $i = 1, 2, \ldots, m$. Thus the dataset with m rows and varying number of columns is generated where each row correspond to observations from a particular segment and out of them m_0 (fixed by the choice of π_0) segments originally have mean lifetime equal to $\theta_0 = 1$. From each row of the dataset we obtain p-value by applying appropriate test and construct a p-value array of length m to compute the bias-corrected estimators from Algorithms 4.1 and 4.2. The other estimators are computed using *estim.pi0* R-function (available in *cp4p* library). Algorithm 4.3 also uses this array of p-values and an estimate of π_0 to identify the significantly different segments with control over FDR.

5.2. Simulation results

Under different settings mentioned in Section 5.1, each experiment is repeated $N = 1000$ times and the estimators are compared through $MSE(\hat{\pi}_0) = 1(/N) \sum_{i=1}^{N} (\hat{\pi}_{0i} - \pi_0)^2$ and $Bias(\hat{\pi}_0) = (1/N) \sum_{i=1}^{N} (\hat{\pi}_{0i} - \pi_0)$. We also validate whether the adaptive BH algorithm using different π_0 estimators are conservative or not by simulating FDR values as a function of π_0 for the adaptive algorithms. Here we identify power of an multiple testing algorithm as the proportion of rejected nulls among the hypotheses in $\mathcal{F}$. The non-adaptive BH algorithm and its different adaptive versions are also compared with respect to power. The comparative study for $m = 100$ under balanced setting is provided in Figure 1. The results for other simulation settings are provided in Figures 1–5 of the supplementary material. From Figure 1, it can be pointed out that the bias-corrected estimators beat other estimators over a significant region of the parameter space (for $\pi_0 \in (0, 0.6)$) while $\hat{\pi}_0^U$ performs slightly better than $\hat{\pi}_0^E$. Thus, their performance may be considered to be approximately equivalent. Thus using the bias-corrected estimators for small to moderate values of π_0 brings significant improvement while for larger values of π_0, it remains a viable alternative. For MSE, similar comments may be made. Additionally, we point out that $\hat{\pi}_0^U$ really reduces the bias of $\hat{\pi}_0^B$ for a significant portion of the parameter space. As expected, the mean squared error for different estimators increase with increasing m/n ratio, while relative performance of the proposed bias-corrected estimators gets better when the same ratio increases. However, the gain from improved estimation of π_0 needs to be elaborated. Precise estimation of π_0 is used to apply adaptive algorithm for identifying significant segments as mentioned in Section 4. For lower to moderate values of π_0, the adaptive versions

result in substantial gain in power. Percentage relative gains in power of $\hat{\pi}_0^U$-adaptive BH over non-adaptive are 41%, 27%, 17% for π_0=0.2, 0.4, 0.6, correspondingly. Marginal gain is observed for larger values of π_0 (8% for π_0=0.8). It is evident that the bias-corrected estimators outperform the others for lower to moderate values of π_0, where it really matters as pointed out from Figure 1. For higher values of π_0, effect of bias correction is there but in a lesser extent. When almost all the null hypotheses are true, $Q(\lambda)$ and e are close to 0. Thus the bias correction does not work as effectively as it does for lower to moderate values of π_0.. FDR of all the adaptive BH algorithms are seen to be controlled below 0.1, while non-adaptive BH is the most conservative and the $\hat{\pi}_0^U$-adaptive BH is the least conservative one. Similar conclusions can be made from results of the other simulation settings reported in the supplementary material.

6. Data analysis

For the case-study we have considered the real-life data (with proper camouflaging, after taking care of the data confidentiality issue) used by Gupta *et al.* [10] in connection with reliability and warranty studies. The detailed description of the data is available there and we report only the relevant part of it, which is required in the present study. The date of failure of a particular component of automobiles along with the mileage at failure as reflected through the odometer readings are available. Although the entire data set cover two disjoint geographical regions, as reported in Gupta *et al.* [10], they may be further subdivided into failures corresponding to seven sub-regions, termed as zones. Owing to data confidentiality issue, let us number them from 1 to 7. We have considered failure data corresponding to a particular year and the mileage figures of the failures in successive calendar months across the zones as the response variable. The twelve calendar months are recorded as JAN, FEB, MAR, APR, MAY, JUN, JUL, AUG, SEP, OCT, NOV and DEC. Thus the entire dataset is inherently segmented by 7 different zones and 12 different months in a year. In other words, our synthetic dataset contains mileage at failure for 84 different month×zone segments. In total, little less than 3000 component failures have been reported in the year considered with varying number of warranty claims over the month×zone segments. Thus on an average around 35 failures are reported in each segment. In line with the discussion in Section 1, here we are primarily interested in identifying the segments which have significantly better or poor performance in terms of mileage at failure in comparison to a bench mark. Thus, appropriate hypotheses are needed to be formed and tested separately for each of the segments making way for application of adaptive FDR-controlling algorithms.

To validate our model-assumption we perform Kolmogorov-Smirnov's test with empirical *p*-value for exponential distribution (using R-function *ks.exp.test* available in *exptest* package) and find that, at level 0.05 exponentiality fails for only 18 out of 84 segments whereas at level 0.01 only 7 rejections are there. QQ-plot of some randomly chosen segments are given in Figure 6 of the supplementary material. As the sample sizes for most of the segments are moderate, we also check normality applying Shapiro-Wilks' test (using default R-function *shapiro.test*). At level 0.05, 59 out of 84 hypotheses gets rejected and at level 0.01, the number is 42. The first line of information justifies applicability of the model-based estimators for π_0 discussed in this article whereas the results from the normality test demonstrate the necessity of the modifications achieved through this work over the existing related works, usually done under normal model.
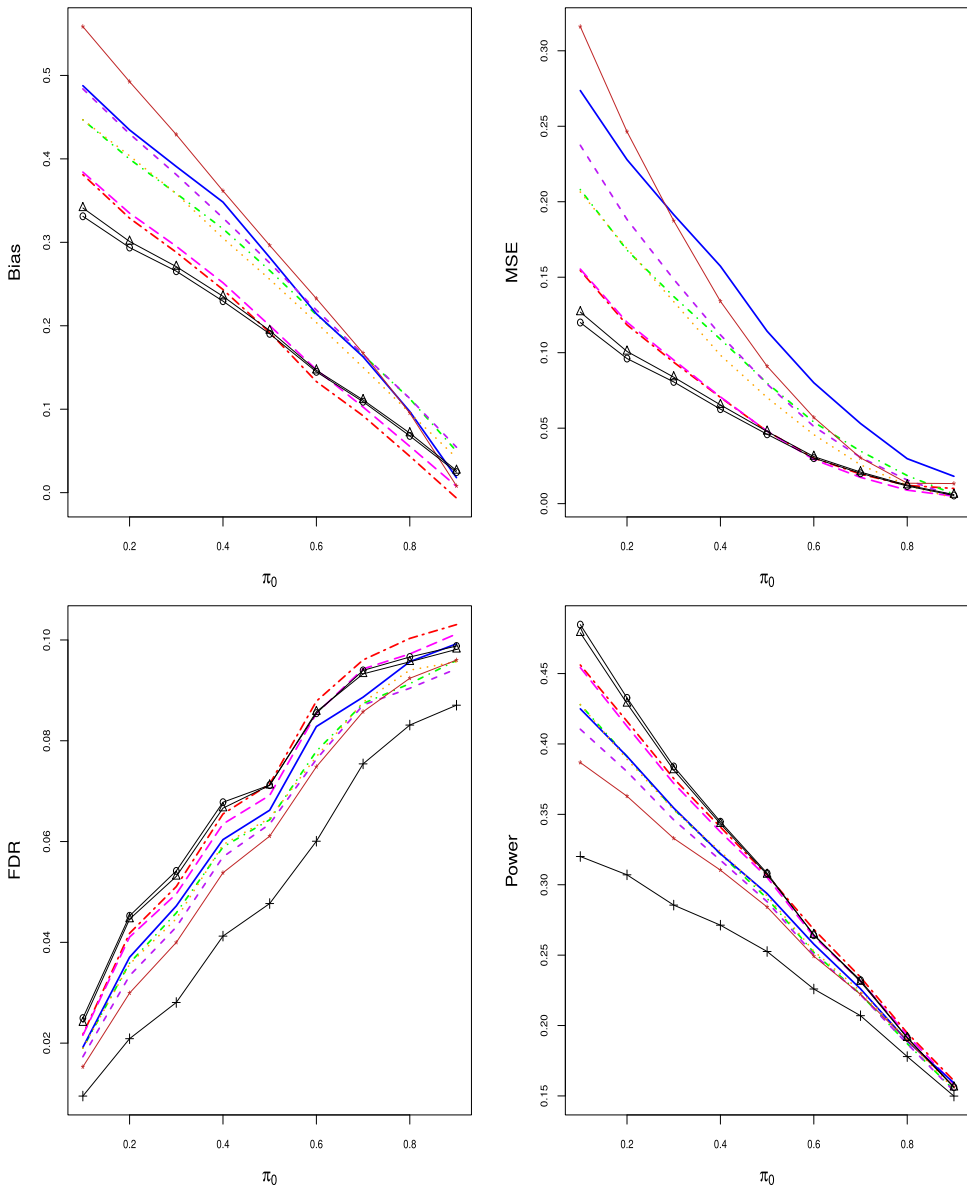
Figure 1. Bias and mean squared error of the estimators $\hat{\pi}_0^B$ (two dashed line), $\hat{\pi}_0^L$ (long-dashed line), $\hat{\pi}_0^A$ (dot-dashed line), $\hat{\pi}_0^P$ (solid line), $\hat{\pi}_0^H$ (dotted line), $\hat{\pi}_0^D$ (dashed line), $\hat{\pi}_0^S$ (∗-marked line), $\hat{\pi}_0^U$ (o-marked line), $\hat{\pi}_0^E$ (△-marked line) and comparison of false discovery rate and power between ordinary BH (+-marked line) and the adaptive BH with the specified estimators for $m = 100$ under balanced setting.

Now we consider framing of the appropriate hypotheses. We assume that the mileages at failure for the ith segment to be exponentially distributed with mean θ_i miles. Thus, θ_i's are the mean mileage to failure (MMTF) for the ith. segment, a quantity similar to mean time to failure (MTTF) in terms of the response variable 'mileage', for $i = 1, 2, \ldots, 84$. We consider, as an indicator of the bench mark, the MMTF of the entire dataset as our null hypothesis point, approximately given by $\theta_0 = 10973$ miles. This value as an benchmark

seems to be justified, as the warranty mileage limit for the data base is 36, 000 miles and it is well known that such failure data are usually positively skewed. According to the research question we then simultaneously test the following hypotheses:

$$H_{0i} : \theta_i = \theta_0 \quad \text{versus } H_{1i} : \theta_i \neq \theta_0 \text{ for } i = 1, 2, \ldots, 84. \quad (15)$$

The two-sided choice of the alternative hypotheses at all the segments needs clarification. In the absence of any prior knowledge about the functioning of the component, it is not possible to mark any segment to be better/worse than the overall benchmark in terms of MMTF. As a result, to be on the safe side, we have suggested the alternative hypotheses at all segments to be two-sided. This is very common in multiple testing situations. As an example, in microarray data analysis, the samples used as a reference are called control samples. The other samples exhibiting different phenotypic status are called treated samples. The gene expression levels among these groups may be different. To identify whether a gene is differentially expressed or not, we fix two-sided alternative [5].

Likelihood ratio tests are performed for each of the hypotheses after scaling the original observations by θ_0, maintaining equivalence of the test and the corresponding p-values along with effect sizes for each test are stored for further use. Table 1 of supplementary material provides details under the following heads:

- *segment*: This column provides serial number of the segment, 1 to 84 such that segment i is for ith month of zone 1 for $i = 1(1)12$, segment $12 + i$ is for ith month of zone 2 for $i = 1(1)12$ and so on for the 7 zones in order as mentioned in the first paragraph of this section.
- *n*: Provides available sample size for each segment.
- *pval*: Provides the obtained p-value corresponding to common likelihood ratio test performed for each segment.
- *del*: Provides maximum likelihood estimates of the effect sizes corresponding to each test.

These array of values can be readily fed into Algorithms 4.1–4.3 to get $\hat{\pi}_0^U$, $\hat{\pi}_0^E$ and the list of rejected hypotheses when adaptive FDR-controlling algorithm is applied with different π_0-estimates. Estimate of π_0 along with the corresponding list of rejected hypotheses using the estimators already mentioned in this article are also reported. The estimated π_0-values using different estimators are reported in the second column of Table 1.

In Table 2, we indicate the segments that are found to be significantly different from the average in terms of the mean mileage at of the designated component failure when adaptive BH-algorithm with different π_0-estimators and non-adaptive(N/A) BH-algorithm are applied to control FDR at level $q = 0.05, 0.1$. For visual display, we plot the adjusted p-values for non-adaptive Benjamini–Hochberg algorithm and its adaptive version using $\hat{\pi}_0^U$ with cut-off $q = 0.1$ (see Figure 7 of the supplementary material). From Table 2, it is evident that, the adaptive BH-algorithm using the proposed methods has the ability to identify a larger number of segments with significant variation from benchmark by controlling the FDR at the same level, compared to the non-adaptive BH-algorithm as well as adaptive BH-algorithm using existent estimators for π_0.

Table 1. Estimate of π_0 using different estimators under different model assumptions for the synthetic data.

Estimators	Exponential model	Normal model
$\hat{\pi}_0^B$	0.5555	0.4961
$\hat{\pi}_0^L$	0.5565	0.4559
$\hat{\pi}_0^A$	0.5761	0.5111
$\hat{\pi}_0^P$	0.6074	0.5707
$\hat{\pi}_0^H$	0.5555	0.4497
$\hat{\pi}_0^D$	0.6662	0.5730
$\hat{\pi}_0^S$	0.8065	0.6786
$\hat{\pi}_0^U$	0.4842	0.0000
$\hat{\pi}_0^E$	0.5096	0.0000

Table 2. Significantly different segments identified by adaptive-BH algorithm with different estimators for π_0 for the synthetic data.

Segment	Zone	Month	mean:$\bar{X}$	95%CI(θ)	$\hat{\pi}_0^B$	$\hat{\pi}_0^L$	$\hat{\pi}_0^A$	$\hat{\pi}_0^P$	$\hat{\pi}_0^H$	$\hat{\pi}_0^D$	$\hat{\pi}_0^S$	$\hat{\pi}_0^U$	$\hat{\pi}_0^E$	N/A
16	2	APR	1.44	(1.12, 1.93)	2	2	2	2	2	2	2	2	2	2
18	2	JUN	1.29	(1.01, 1.70)	0	0	0	0	0	1	0	1	0	0
20	2	AUG	1.47	(1.11, 2.04)	2	2	2	2	2	2	1	2	2	0
21	2	SEP	1.33	(1.01,1.82)	0	0	0	0	0	0	0	1	0	0
23	2	NOV	1.69	(1.30, 2.29)	2	2	2	2	2	2	2	2	2	2
38	4	MAR	0.55	(0.37, 0.92)	1	1	1	1	1	0	0	1	1	0
39	4	APR	0.50	(0.34, 0.79)	2	2	2	2	2	2	2	2	2	2
40	4	MAY	0.58	(0.40, 0.92)	1	1	1	1	1	0	0	1	1	0
41	4	JUN	0.63	(0.44, 0.95)	0	0	0	0	0	0	0	1	1	0
46	4	OCT	0.64	(0.45, 0.98)	0	0	0	0	0	0	0	1	0	0
52	5	APR	1.77	(1.10, 3.33)	1	1	1	1	1	1	0	1	1	0
56	5	AUG	1.82	(1.14, 3.32)	1	1	1	1	1	1	1	2	2	1
62	6	FEB	0.66	(0.53, 0.84)	2	2	2	2	2	2	2	2	2	2
63	6	MAR	0.60	(0.48, 0.78)	2	2	2	2	2	2	2	2	2	2
64	6	APR	1.27	(1.03, 1.61)	1	1	1	1	1	0	0	1	1	0
69	6	SEP	0.67	(0.54, 0.85)	2	2	2	2	2	2	2	2	2	2
74	7	FEB	0.20	(0.11, 0.51)	2	2	2	2	2	2	2	2	2	2
76	7	APR	0.48	(0.20, 0.48)	0	0	0	0	0	0	0	1	0	0

Notes: Values in fourth and fifth columns are reported after scaling the original variable by $\theta_0 = 10973.30$. For columns 6 to 15, 2 indicates that the segments are found significant for $q = 0.05$ and trivially for $q = 0.1$ both while 1 indicates same only for $q = 0.1$ and 0 indicates negation of the previous two statements. The segments not reported in this table are not found to be significantly different from the the overall average, taken as the null hypothesis point.

From the domain knowledge (not to be mentioned explicitly, owing to confidentiality issue), it is known that the functioning of the automobile component under consideration is likely to be influenced by the climate condition, reflected through the effect of the month, as well as by the effect of the zone of their usual operation. The effect of climate on the functioning of the automobiles is well known and has also been reported in Lawless [14]. For simplicity and demonstration purpose, we assume that each automobile is used only in the designated zone where the failure is reported. Although, we have used the two-sided alternative, as being done in any multiple testing problem, the corresponding confidence interval falling entirely below or above of the scaled null hypothesis point of 'unity', indicates the actual one-sided alternative for which the respective significance appears. Thus, the MMTFs of zone 4 are consistently and significantly smaller than the

benchmark value (the null hypothesis point) indicating usage related adverse problem of the automobiles and this problem is persistent in the first or second quarter of the year indicating a transition from colder to warmer climate or the fourth quarter of the year indicating the transition from warmer to colder climate. Interestingly for zone 2, exactly the converse situation is prevailed and this seemingly high MMTF might not be due to the climate condition and on the contrary may be attributed to better usage scenario. For zone 5, better usage scenario is evident at least in two months, although weather related issues might not be associated with such improvement. The findings for zone 6 is heavily dependent on climate condition especially during the advent of spring where a significant decrease in MMTF is identified followed by significant increase in MMTF just after. Again during the fall a significant decrease in MMTF is found establishing the climate dependence of failure data. For zone 7, climate plays an adverse role during the end of the winter and start of the summer. The data corresponding to remaining two zones, do not reveal any deviation from the usual usage pattern and/or are not affected by extreme climate conditions. It is to be noted that for almost all the zones, the month of April becomes significant concerning either betterment or worsening of the scenario in comparison to the benchmark. On the other hand the two months, viz. December and January never become markedly different from the benchmark at all the locations. It might be attributed to the fact that in the winter, the relatively colder temperature does not affect all the zones, while a transition in temperature, as observed in April, may play a decisive role in operating conditions in almost all the zones. Zones 1 and 3 never figure in the list and no marked deviation from the benchmark in any climate condition (non-rejection of the null hypothesis at all seasons) is observed. This homogeneity might be attributed to the fact that these two zones correspond to a relatively warmer climate and hence climate dependence on the operating conditions are not present here. Although, we have to suppress the zone identity for confidentiality issue, the findings are as corroborated by the domain knowledge experts.

To conclude this section, we emphasize the appropriateness of the model-based bias-correction approach. We try to explore a 'what-if' type scenario and try to assess the validity of the findings if we assume the mileages at failure in each segment to be normally distributed, instead of the exponential assumption. Our main objective remains same, i.e to identify the significantly different segments with respect to MMTF values. If we assume that the mileages at failure for the ith segment be normally distributed with mean θ_i and variance σ_i^2, the testing problem is still the same as in (15). We perform single sample both-sided t-test for each of the segments and obtain the array of p-values over all the segments. Computation of robust estimates may be done similarly as mentioned, but for the bias-corrected estimators we follow algorithms given in Cheng $et\ al.$ [6] (for $\hat{\pi}_0^U$) and Biswas [3] (for $\hat{\pi}_0^E$) instead of Algorithms 4.1 and 4.2 for obvious reasons. The estimates of π_0 under normality assumption are reported in the third column of Table 1. The robust estimators are seen to underestimate π_0. When the rates of exponential distributions are considered as the means of normal distributions, the sample means being the estimators under both the model assumptions overestimate the normal means. Since the overestimation of normal means makes the null hypotheses appear false, the observed p-values are less compared to those under the exponential case. The robust estimators of π_0 are non-decreasing functions of p-values and hence the underestimation of π_0. The bias-corrected estimators get disrupted owing to the inappropriate model assumption and hence misleading effect size of test, upper tail probability and expectation of non-null p-values. The

problem of overestimation of normal means transcends to overestimation of effect sizes for the single sample t-tests. The inflated estimates of effect sizes result in large estimates of $Q(\lambda)$ and e. As a result, the numerator in $\hat{\pi}_0^U$ and $\hat{\pi}_0^E$ usually turns out to be zero or negative. Hence both the bias-corrected estimates in Table 1 are zero. Thus, appropriate model-based bias-correction seems to be appropriate and efficient by bringing out more power in adaptive algorithms, while the findings may be misleading when not applied with adequate confidence on underlying model assumption. As a result, the necessary modification of bias-correction technique under exponential model seems to be the only way out, particularly while dealing with multiple testing problem arising from segmented failure data, usually encountered in survival and reliability studies.

7. Discussion

We have approached the problem of estimating π_0 and thus construction of adaptive FDR-controlling procedure from suitable model assumptions and a common test for all the hypotheses to be tested. Within the framework suggested in Cheng *et al.* [6] and Biswas [3], we have tried to develop methods for estimation of π_0 under exponential model and presented a simple adaptive Benjamini–Hochberg algorithm in a spirit similar to Craiu and Sun [7], which is shown to be more efficient than its counterparts for simulated as well as real-life synthetic data. The current work also motivates the Storey's bootstrap estimator for π_0 and the π_0-estimator based on sum of all p-values based on $P(p \geq V)$. The cases of V being degenerated at some λ and V being uniformly distributed over $(0, 1)$ have also been discussed. This may motivate other choices of V for further study of model-based π_0 estimators. The study on V having negatively skewed density function over $(0, 1)$ is presently under consideration, which tries to give more importance to the p-values corresponding to true null hypotheses and the construction of new estimators in future. In the current work, it has been assumed that the p-values corresponding to the true null hypotheses are uniformly distributed. However, if there are composite null hypotheses as in one-sided hypothesis testing scenarios, p-values corresponding to the true null hypotheses are stochastically larger than the uniform variate. Superuniform p-values make the proposed estimators conservative due to the increased value of $W(\lambda)$ and $\bar{p}$. The results and methods proposed in this work do not address the issues related to superuniformity of null p-values. Though the results presented in the current work strengthens the foundations of bias-corrected estimation of π_0 in general, the distinguishing feature of this work lies on the innovative application of multiple testing procedure to segmented failure data. To the best of our knowledge, such procedures have never been applied to answer such interesting research questions framed in Section 6 related to large-scale industrial data. In this work, however, we have focused on presenting and motivating a simple yet powerful technique of identifying significantly different segments in terms of the performance of automobile and exploring the effect of zone of operation coupled with climate, that too under exponential model assumption. The synthetic data explored in this work pose several other issues that may be solved by the application of modified methods, which are to be formulated in future.

This analysis of the real-life synthetic data is based on one year data and may be carried out on the basis of monthly or even weekly data associated with the component failures. Owing to the limited number of such failure data in each segment, one has to use the

standard failure models like exponential or Weibull. Instead, if one uses the usual Gaussian model to describe the failure pattern, then one is expected to commit a gross mistake and consequently, a false perception on the MMTF may be reached. This issue has been addressed with the same failure data. Instead of the exponential model, the normal probability model has been used and the test for equality of respective means in all 84 segments with the same null hypothesis point representing the bench mark, as being done in the usual multiple testing procedure, has been attempted. Interestingly, the test for normality at majority of the 84 segments fails miserably and hence conclusion on the basis of the test for MMTF with reference to the benchmark under normality will give a wrong signal about the true status of MMTF in those segments.

This work only focuses on controlling FDR by adaptive Benjamini–Hochberg algorithms with two new estimators for the proportion of true null hypotheses. It would be interesting to study control of family wise error rate (FWER) by adaptive procedures [21] with the π_0-estimators discussed here. We have demonstrated that the new π_0-estimators devise conservative procedures through simulation experiments. The proof is also done in asymptotic setting. Still, it is desirable to prove the same under finite sample considerations and future research on this aspect is warranted. Sarkar [20] and Blanchard and Roquain [4] provide sufficient conditions on $\hat{\pi}_0$ for proving control over FDR by the corresponding adaptive algorithms. Almost all the recently proposed estimators for π_0 including the two taken up in this work lack such structural simplicity and hence the desired result can only be verified through finite sample simulation experiments [3,5,24,26]. The estimators taken up in this work need an initial estimate of π_0. The current work does not focus on a simulation-based choice of the initial estimate. However, the choices are not expected to be universal and in this regard one may follow the routine presented in Biswas [3] for identifying the working initial estimator. The choice of initial estimate in the current work is justified by Cheng *et al.* [6] and Biswas [3]. As the proposed method makes assumption regarding the distribution of the mileage to failure data, we should accept the fact that, the proposed estimator is not universally suitable in all situations. At the same time, multiple testing problem in a non-Gaussian framework seems to be novel and may cover all parametric models for scenarios where non-negative valued random variables seem to be appropriate. In such a framework, we have introduced two simple estimators for π_0 which simultaneously reduces the bias and variance of the existing estimator over a relatively important part of the parameter space. The behaviour of such estimator is studied through extensive simulation studies and the new estimator is shown to be more precise under some practical assumptions in comparison to those available in the existing literature. Involvement of numerical or Monte-Carlo integration for each segment makes the proposed method rather computation intensive. This extra labour is expected to be compensated by the gain in precision of the analysis, thus meaningfully addressing the multiple testing problem in a non-Gaussian setup.

Acknowledgments

Disclosure statement

No potential conflict of interest was reported by the authors.

Code availability statement

The necessary R codes for computing the estimators $\hat{\pi}_0^U$ and $\hat{\pi}_0^E$ are available at https://github.com/aniketstat/EstPi0Exp2021.

ORCID

Aniket Biswas 🆔 http://orcid.org/0000-0002-1711-2928

References

[1] Y. Benjamini and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, J. R. Statist. Soc. Ser. B (Statist. Methodol.) 57 (1995), pp. 289–300.

[2] Y. Benjamini, A.M. Krieger, and D. Yekutieli, *Adaptive linear step-up procedures that control the false discovery rate*, Biometrika 93 (2006), pp. 491–507.

[3] A. Biswas, *Estimating the proportion of true null hypotheses based on sum of p-values and application in microarray data*, Commun. Stat. Simul. Comput. (2020), pp. 1–15. doi:10.1080/03610918.2020.1800036.

[4] G. Blanchard and E. Roquain, *Adaptive false discovery rate control under independence and dependence*, J. Mach. Learn. 10(12) (2009), pp. 2837–2871.

[5] J.J. Chen, S.J. Wang, C.A. Tsai, and C.J. Lin, *Selection of differentially expressed genes in microarray data analysis*, Pharmacogenomics. J. 7 (2007), pp. 212–220.

[6] Y. Cheng, D. Gao, and T. Tong, *Bias and variance reduction in estimating the proportion of true-null hypotheses*, Biostatistics 16 (2015), pp. 189–204.

[7] R.V. Craiu and L. Sun, *Choosing the lesser evil: trade-off between false discovery rate and non-discovery rate*, Stat. Sin. 18 (2008), pp. 861–879.

[8] Q.G. Gianetto, F. Combes, C. Ramus, and M.Q.G. Gianetto, Package 'cp4p'. (2019). Available at https://cran.r-project.org/web/packages/cp4p/cp4p.pdf.

[9] Z. Guan, B. Wu, and H. Zhao, *Nonparametric estimator of false discovery rate based on bernštein polynomials*, Stat. Sin. 18 (2008), pp. 905–923.

[10] S.K. Gupta, S. De, and A. Chatterjee, *Some reliability issues for incomplete two-dimensional warranty claims data*, Reliab. Eng. Syst. Saf. 157 (2017), pp. 64–77.

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

[12] H. Jiang and R.W. Doerge, *Estimating the proportion of true null hypotheses for multiple comparisons*, Cancer. Inform. 6 (2008), pp. 25–32.

[13] M. Langaas, B.H. Lindqvist, and E. Ferkingstad, *Estimating the proportion of true null hypotheses, with application to DNA microarray data*, J. R. Statist. Soc. Ser. B (Stat. Methodol.) 67 (2005), pp. 555–572.

[14] J.F. Lawless, *Statistical analysis of product warranty data*, Int. Stat. Rev. 66 (1998), pp. 41–60.

[15] A. Markitsis and Y. Lai, *A censored beta mixture model for the estimation of the proportion of non-differentially expressed genes*, Bioinformatics 26 (2010), pp. 640–646.

[16] D. Nettleton, J.G. Hwang, R.A. Caldo, and R.P. Wise, *Estimating the number of true null hypotheses from a histogram of p-values*, J. Agric. Biol. Environ. Stat. 11 (2006), pp. 337–356.

[17] I. Ostrovnaya and D.L. Nicolae, *Estimating the proportion of true null hypotheses under dependence*, Stat. Sin. 22 (2012), pp. 1689–1716.

[18] S. Pounds and C. Cheng, *Robust estimation of the false discovery rate*, Bioinformatics 22 (2006), pp. 1979–1987.

[19] S. Pounds and S.W. Morris, *Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values*, Bioinformatics 19 (2003), pp. 1236–1242.
[20] S.K. Sarkar, *On methods controlling the false discovery rate*, Sankhyā: Indian J. Stat. Ser. A 70 (2008), pp. 135–168.
[21] S.K. Sarkar, W. Guo, and H. Finner, *On adaptive procedures controlling the familywise error rate*, J. Stat. Plan. Inference 142 (2012), pp. 65–78.
[22] J.D. Storey, *A direct approach to false discovery rates*, J. R. Stat. Soc. Ser. B (Stat. Methodol.) 64 (2002), pp. 479–498.
[23] J.D. Storey and R. Tibshirani, *SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays*, in *The Analysis of Gene Expression Data*, Springer, New York, NY, 2003, pp. 272–290.
[24] J.D. Storey, J.E. Taylor, and D. Siegmund, *Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach*, J. R. Statist. Soc. Ser. B (Stat. Methodol.) 66 (2004), pp. 187–205.
[25] T. Tong, Z. Feng, J.S. Hilton, and H. Zhao, *Estimating the proportion of true null hypotheses using the pattern of observed p-values*, J. Appl. Stat. 40 (2013), pp. 1949–1964.
[26] H.Q. Wang, L.K. Tuominen, and C.J. Tsai, *SLIM: A sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures*, Bioinformatics 27 (2010), pp. 225–231.

Appendix

Proof of Result 2.1: Consider,

$$g(x) = \frac{a - x}{b - x}.$$

Note that, g is non-increasing in x for $a \le b$. Let $a = W(\lambda)/m$ and $b = (1 - \lambda)$. Since, $0 \le \hat{Q}(\lambda) \le \tilde{Q}(\lambda)$, $g(0) \ge g(\hat{Q}(\lambda)) \ge g(\tilde{Q}(\lambda))$, which proves (*a*). Now,

$$\tilde{\pi}_0^U(\lambda) = \frac{\frac{W(\lambda)}{m} - \tilde{Q}(\lambda)}{(1 - \lambda) - \tilde{Q}(\lambda)}.$$

$\hat{\delta}_i \xrightarrow{a.s} \delta_i \; \forall i = 1, 2, \ldots, m$ as $\min\{n_1, n_2, \ldots, n_m\} \to \infty$. Thus,

$$\tilde{\pi}_0^U(\lambda) \xrightarrow{a.s} \frac{\frac{W(\lambda)}{m} - Q(\lambda)}{(1 - \lambda) - Q(\lambda)} = \tilde{\pi}_0, \quad \text{say.}$$

Note that, $W(\lambda)/m \xrightarrow{a.s} \bar{F}(\lambda)$ as $m \to \infty$. Hence $\tilde{\pi}_0 \xrightarrow{a.s} (\bar{F}(\lambda) - Q(\lambda))/((1 - \lambda) - Q(\lambda)) = \pi_0$ as $m \to \infty$. ∎

Proof of Result 2.2: We consider g as in Result 2.1 and assume $a = \bar{p}$, $b = 0.5$. Since, $\hat{e} \le \tilde{e}$, $g(\hat{e}) \ge g(\tilde{e})$, which proves (*a*). Here,

$$\tilde{\pi}_0^E = \frac{\bar{p} - \tilde{e}}{0.5 - \tilde{e}}.$$

$\hat{\delta}_i \xrightarrow{a.s} \delta_i \; \forall i = 1, 2, \ldots, m$ as $\min\{n_1, n_2, \ldots, n_m\} \to \infty$. Thus,

$$\tilde{\pi}_0^E \xrightarrow{a.s} \frac{\bar{p} - e}{0.5 - e} = \tilde{\pi}_0, \quad \text{say.}$$

Note that, $\bar{p} \xrightarrow{a.s} E(p)$ as $m \to \infty$. Hence, $\tilde{\pi}_0 \xrightarrow{a.s} (E(p) - e)/(0.5 - e) = \pi_0$ as $m \to \infty$. ∎

Proof of Result 3.1: The likelihood ratio test corresponding to the hypothesis in Result 2.1 uses the test-statistic $T = 2 \sum_{i=1}^n X_i \sim \theta \chi_{2n}^2$. Effect size of the test $\delta = \theta$. As $E(T) = 2n\delta$, an unbiased estimator of δ is $\hat{\delta} = T/2n = \bar{X}$, the sample mean.

As we reject H_0 for larger observed value of T, the corresponding p-value is defined as $p = P_{H_0}(\chi^2_{2n} > T) = 1 - F_{\chi^2_{2n}}(T)$ since under H_0, $T \sim \chi^2_{2n}$. Therefore, $p \sim Uniform(0, 1)$, under H_0. Under H_1, $T \sim \delta\chi^2_{2n}$ and therefore the density function of T, labelled by δ is

$$f_\delta(t) = \frac{1}{\delta}f_{\chi^2_{2n}}\left(\frac{t}{\delta}\right) \quad \text{for } t > 0. \tag{A1}$$

From the relation between T and p, $t = F^{-1}_{\chi^2_{2n}}(1 - p) = \chi^2_{p,2n}$. The corresponding absolute Jacobian of transformation is $f_{\chi^2_{2n}}(\chi^2_{p,2n})$. Thus from (A1), the density function of p labelled by δ is

$$f_\delta(p) = \frac{\frac{1}{\delta}f_{\chi^2_{2n}}(\chi^2_{p,2n})}{f_{\chi^2_{2n}}(\chi^2_{p,2n})} \quad \text{for } 0 < p < 1. \tag{A2}$$

For $\lambda \in (0, 1)$ upper tail probability labelled by δ, using (A2) in expression of $Q_\delta(\lambda)$ we get

$$Q_\delta(\lambda) = \int_\lambda^1 \frac{\frac{1}{\delta}f_{\chi^2_{2n}}(\frac{1}{\delta}\chi^2_{p,2n})}{f_{\chi^2_{2n}}(\chi^2_{p,2n})} \, dp = I, \quad \text{say.}$$

By change of variable from p to v such that $v = (1/\delta)\chi^2_{p,2n}$ we get

$$I = F_{\chi^2_{2n}}\left(\frac{1}{\delta}\chi^2_{\lambda,2n}\right).$$

which proves the result in (c). For an explicit expression for expected p-value under the false null, we apply

$$e_\delta = \int_0^1 F_{\chi^2_{2n}}\left(\frac{1}{\delta}\chi^2_{p,2n}\right) dp$$

By change of variable from p to v such that $v = \chi^2_{p,2n}$, we get

$$e_\delta = \int_0^\infty F_{\chi^2_{2n}}\left(\frac{v}{\delta}\right) f_{\chi^2_{2n}}(v) \, dv$$

$$= E_{X \sim \chi^2_{2n}}\left[F_{\chi^2_{2n}}\left(\frac{X}{\delta}\right)\right].$$

∎

Proof of Result 3.2: The corresponding likelihood ratio test uses the same test-statistic as in Result 3.1 and thus part (a) of Result 3.2 follows directly from part (a) of Result 3.1. For the next part, it should be noted that due to two-sided alternative hypotheses, the corresponding p-value is defined through T, where

$$p = 2 \min\left\{P(\chi^2_{2n} > T), P(\chi^2_{2n} < T)\right\}$$
$$= 2 \min(p^*, 1 - p^*), \quad \text{say.}$$

Here $p^* = P(\chi^2_{2n} > T)$ is the p-value defined for the testing problem in Result 2.1. Thus from part (b) of Result 3.1, we have

$$f_\delta(p^*) = \frac{\frac{1}{\delta}f_{\chi^2_{2n}}(\chi^2_{p^*,2n})}{f_{\chi^2_{2n}}(\chi^2_{p^*,2n})}. \tag{A3}$$

Now for any $\lambda \in (0, 1)$,

$$Q_\delta(\lambda) = P(p > \lambda)$$

$$= P\left(\frac{\lambda}{2} < p^* < 1 - \frac{\lambda}{2}\right)$$

$$= \int_{\frac{\lambda}{2}}^{1 - \frac{\lambda}{2}} \frac{\frac{1}{\delta} f_{\chi_{2n}^2}(\chi_{p^*,2n}^2)}{f_{\chi_{2n}^2}(\chi_{p^*,2n}^2)} \, dp^* \quad [\text{using (A3)}]$$

$$= \int_{\frac{1}{\delta}\chi_{1-\frac{\lambda}{2},2n}^2}^{\frac{1}{\delta}\chi_{\frac{\lambda}{2},2n}^2} f_{\chi_{2n}^2}(v) \, dv \quad \text{by taking} \quad v = \frac{1}{\delta}\chi_{p^*,2n}^2,$$

which proves the result in (b).

$$e_\delta = \int_0^1 Q_\delta(\lambda) \, d\lambda$$

$$= \int_0^1 F_{\chi_{2n}^2}\left[\frac{1}{\delta}\chi_{\frac{p}{2},2n}^2\right] dp - \int_0^1 F_{\chi_{2n}^2}\left[\frac{1}{\delta}\chi_{1-\frac{p}{2},2n}^2\right] dp$$

$$= I_1 - I_2, \quad \text{say}.$$

Now, we consider the problem of evaluating the integral I_1. By change of variable from p to v such that $v = \chi_{\frac{p}{2},2n}^2$, we get

$$I_1 = 2 \int_{\chi_{\frac{1}{2},2n}^2}^{\infty} F_{\chi_{2n}^2}\left(\frac{v}{\delta}\right) f_{\chi_{2n}^2}(v) \, dv$$

$$= E_{X \sim \chi_{2n}^2(0,\mu)}\left[F_{\chi_{2n}^2}\left(\frac{X}{\delta}\right)\right]$$

Following the same steps for evaluating I_1, I_2 can also be evaluated and thus the result in (c). ■

Proof of Result 3.3: The likelihood ratio test corresponding to the hypothesis in Result 3.3 uses the test-statistic $T = \sum_{i=1}^{n_2} Y_i / \sum_{i=1}^{n_1} X_i \sim (\theta_2/\theta_1) F_{2n_2,2n_1}$. Effect size of the test $\delta = \theta_2/\theta_1$. As $E(T) = \delta[n_1/(n_1 - 1)]$, an unbiased estimator of δ is $\hat{\delta} = [(n_1 - 1)/n_1]T$ and thus the result in (a) follows. Rest of the proof follows from the proof of Result 3.1 with obvious changes. ■

Proof of Result 3.4: For the testing problem in Result 3.4, the likelihood ratio test uses the same test-statistic as in Result 3.3. Since the critical region is two-sided, the corresponding p-value is similarly defined as in Result 3.2. One can follow the steps elaborated through the proof of Result 3.2 and use Result 3.3 to easily prove Result 3.4. ■