Universal Inference for Testing Calibration of Mean Estimates within the Exponential Dispersion Family

Łukasz Delong* Mario Wüthrich†

Abstract

Calibration of mean estimates for prediction is a crucial property in many applications, particularly in the fields of financial and actuarial decision making. In this paper, we first review classical approaches for validating mean-calibration, and we discuss the Likelihood Ratio Test (LRT) within the Exponential Dispersion Family (EDF). Then, we investigate the framework of universal inference to test for mean-calibration. We develop a sub-sampled split LRT within the EDF that provides finite sample guarantees with universally valid critical values. We investigate type I error, power and e-power of this sub-sampled split LRT, and we compare it to the classical LRT. We propose a novel test statistics based on a sub-sampled split Lq-Likelihood Ratio Test (LqRT) to enhance the performance of the calibration test. A numerical analysis verifies that universal inference with the sub-sampled split LRT and the sub-sampled split LqRT is an attractive alternative to the classical LRT achieving a high power in detecting miscalibration in medium and large samples.

Keywords. Calibration of mean predictions, e-values, e-power, Exponential Dispersion Family, isotonic regression, split Likelihood Ratio Test, universal inference.

1 Introduction

Calibration is an important concept in forecasting which refers to a consistent alignment of predictions and observations. Intuitively, a predictive model is calibrated if it neither produces systematically too low nor too high forecasts for the response to be predicted. The notion of calibration was introduced by Murphy (1973) in the literature of meteorology, and it was further developed in the recent statistical literature; see, e.g., Dimitriadis et al. (2021), Gneiting & Resin (2023) and Pohle (2020). There is a consensus across most scientific communities that predictive models must be calibrated to be useful in practice, especially, in financial applications where a miscalibration can lead to a large financial loss. This paper considers testing calibration of mean estimates.

We denote the responses by Y. These responses are equipped with features (covariates, explanatory variables) $\mathbf{X}$ taking values in a feature space $\mathcal{X} \subseteq \mathbb{R}^q$. We are interested in (measurable) regression functions $\mu : \mathcal{X} \to \mathbb{R}$, $\mathbf{X} \mapsto \mu(\mathbf{X})$, serving as forecasts for Y, given the features $\mathbf{X}$.

Definition 1. *A regression function $\mu : \mathcal{X} \to \mathbb{R}$ is mean-calibrated for the pair $(Y, \mathbf{X})$, or simply calibrated, if*

$$\mathbb{E}\left[Y \mid \mu\left(\mathbf{X}\right)\right] = \mu\left(\mathbf{X}\right), \qquad \textit{almost surely.} \tag{1.1}$$

*University of Warsaw, Department of Statistics and Econometrics, l.delong@uw.edu.pl
†ETH Zürich, Department of Mathematics, mario.wuethrich@math.ethz.ch

In different streams of the literature, (mean-)calibration (1.1) is also known as auto-calibration or well-calibrated.

Calibration (1.1) implies that whenever the prediction $\mu(\mathbf{X})$ is used to forecast Y, it is on average correct. Clearly, this is a desired property of well calibrated predictive models. Calibration of mean estimates is particularly important in financial applications such as actuarial pricing. Lindholm & Wüthrich (2025) elaborate on necessary properties of actuarial pricing systems, calibration (1.1) being a crucial one because it implies that there is no systematic cross-subsidy within the pricing system since every price cohort is on average self-financing.

Since mean-calibration (1.1) is a local property, its validation is generally challenging. The following approaches were considered in the literature. Dimitriadis et al. (2021) and Gneiting & Resin (2023) suggested CORP reliability diagrams in a Bernoulli setting. A deviation of a CORP reliability diagram from its diagonal line is a sign of miscalibration of the mean estimates, this gives a visual calibration test. Delong & Wüthrich (2025) developed Monte Carlo/bootstrap techniques to estimate consistency bands for CORP reliability diagrams within the Exponential Dispersion Family (EDF), and Delong et al. (2025) derived analytical formulas for calibration bands of mean estimates within the EDF. Dimitriadis et al. (2021), Gneiting & Resin (2023) and Pohle (2020) proposed to decompose the validation loss (under strictly consistent loss functions) resulting in Murphy (1973)'s score decomposition with the three components of miscalibration, discrimination (resolution) and uncertainty. The resulting miscalibration statistics summarizes the deviation of a CORP reliability diagram from its diagonal line in a single number. Using the bootstrap techniques of Delong & Wüthrich (2025), this allows one to test for the null hypothesis of mean-calibration using the miscalibration statistics of Murphy's score decomposition. Denuit et al. (2024) derive a statistical test with a large sample limit distribution being based on the difference between Lorenz and concentration curves, and this is further discussed in simplified settings in Wüthrich (2025).

From the statistical viewpoint, the Likelihood Ratio Test (LRT) is the most popular statistical test in parametric and non-parametric hypothesis testing. The LRT has not been largely investigated so far in the context of mean-calibration. Our main new contribution is to show how the Likelihood Ratio Statistics (LRS) can be used for verifying mean-calibration. The LRS in classical inference coincides with the miscalibration statistics of Murphy's score decomposition within the EDF, up to a non-linear transformation. This seems very promising since both the LRT and Murphy's score decomposition have strong technical foundations in mathematical statistics and statistical machine learning. The most challenging part of the classical LRS and the miscalibration statistics in practical applications is to derive the critical value of the test; as noted in Dimitriadis et al. (2021) and Gneiting & Resin (2023). One can use Monte Carlo simulation or the bootstrap, but the need for simulations clearly limits a widespread application of the test in practice. An important point is that we do not expect that the classical LRS for testing mean-calibration has a standard χ^2-distribution. In fact, Banerjee (2007) derives the limit distribution of the LRS under the null hypothesis that the mean estimate for one chosen observation is calibrated. The limit distribution is based on a functional of Brownian motion, and its practical use is difficult (cumbersome).

In many practical situations, including the case of mean-calibration, the necessary regularity assumptions and large sample requirements for the classical LRT likely fail to hold, and the true large sample limit distribution of the test statistics remains intractable. Moreover, simulations of the test statistics under the null hypothesis may not be attractive to practitioners. As an

alternative, Wasserman et al. (2020) introduce *universal inference* and the *split LRT*. The idea of the split LRT is to split the data into two sets: a *training sample* used to train a model under the alternative hypothesis and a *validation sample* used to validate the null hypothesis against the alternative. Interestingly, thanks to this data splitting scheme, a universal critical value for the split LRT can be derived in rather general settings.

We emphasize the following general points related to universal inference:

(a) It is straightforward in applications since the critical values do not need to be determined by simulations of the test statistics under the null hypothesis.

(b) It works for a general non-parametric set-up as long as we can compute the likelihoods.

(c) It is valid on finite samples and no regularity conditions are required to obtain critical values.

(d) It is most useful in 'irregular' models for which valid tests (on finite samples and/or asymptotically) are intractable.

(e) However, the resulting tests are conservative in terms of type I errors, and they have a lower power than classical LRTs.

In this paper, we develop and promote universal inference for testing mean-calibration within the EDF. Even though many approaches for testing mean-calibration have been developed, see above discussion, we believe that none of them offers a straightforward and sufficiently practical application. At the same time, universal inference is gaining popularity in the statistical community; Wasserman et al. (2020) give numerous applications of universal inference in irregular parametric and non-parametric settings in which the split LRT is the first ever statistical test with finite sample guarantees, e.g., for testing the number of components in mixture models. Strieder & Drton (2022) observe empirically that the split LRT may be conservative. Dunn et al. (2022) study the performance of the split LRT in Gaussian models and they conclude that the power of the split LRT is reasonable compared to the classical LRT, even though the type I error is below the pre-assumed significance level. Strieder & Drton (2022) and Dunn et al. (2022) conclude that the choice of the splitting ratio equal to $1/2$, based on which the data set is split into two parts – training and validation samples – is a reasonable choice in all dimensions and should lead to the highest power of the split LRT (under some assumptions). Shi & Drton (2025) recently investigated the split LRT in Gaussian mixture models. Finally, Henzi et al. (2025) use the split LRT to validate the calibration of estimated probabilities in a binomial distribution setting. All this literature recommends to use a sub-sampled version of the split LRT and most conclusions are model-specific.

The goal of this paper is to generalize the binomial results of Henzi et al. (2025) to the whole EDF. We especially aim at studying properties of the sub-sampled split LRT within the EDF setting to verify mean-calibration of estimated models. Following Wasserman et al. (2020) and Dunn et al. (2022): *"Our methods may not be optimal, though we do not yet fully understand how close to optimal they are beyond special cases. Many basic questions remain unanswered about the universal LRT, since its power even in very simple settings remains unknown."* Our goal is to make a next step in this direction and to contribute to the field of universal inference by investigating the performance of the sub-sampled split LRT for the purpose of testing mean-calibration within the EDF, especially in the context of actuarial pricing. In this paper, in a theoretical framework and

in a numerical example, we consider the split LRT and the sub-sampled split LRT within the EDF, their type I error, power and e-power, compare the sub-sampled split LRT to the classical LRT, and we study the effect of the sampling of training and validation sets on the split LRT. Moreover, we propose a new testing procedure to verify mean-calibration within universal inference. The novel test statistics is based on the sub-sampled split LqRT and is motivated by the Lq-Likelihood Ratio Test (LqRT), see Ferrari & Yang (2010), Qin & Priebe (2017) – the terminology Lq comes from L^q-norms. All our results essentially draw on the newly developed framework of e-values and e-testing; an excellent reference is the book of Ramdas & Wang (2025). From our numerical analysis, we conclude that the sub-sampled split LRT and the so-called sub-sampled split mean power LqRT have a high and reasonable power in medium/large samples in the case of medium/large miscalibration errors, and the sub-sampled split mean power LqRT may perform better in some scenarios than the sub-sampled split LRT in terms of the power of rejecting the false null hypothesis against the true alternative for mean estimates.

This paper is structured as follows. In Section 2, we introduce our statistical set-up. In Section 3, we discuss classical approaches to test the mean-calibration. In Section 4, we focus on universal inference and present the main theoretical results of this paper. Finally, in Section 5, we study the split LRT and the split LqRT with numerical examples. The Supplementary contains some additional analyses related to the split LqRT.

2 Statistical set-up

We work within the EDF. We recall that the EDF is given by densities of the form

$$f(y) = \exp\left(\frac{y\theta - \kappa(\theta)}{\phi/v} + a(y, v/\phi)\right),$$

where θ denotes the *canonical parameter*, $\phi > 0$ is the *dispersion parameter*, $v > 0$ is a *case weight* and $a(\cdot, v/\phi)$ is a normalizing function w.r.t. the selected σ-finite measure on $\mathbb{R}$ so that f integrates to one. The canonical parameter θ is supported in the effective domain $\mathbf{\Theta}$, and the *cumulant function* $\kappa : \mathbf{\Theta} \to \mathbb{R}$ is smooth and convex on the interior of this effective domain.

A distribution from the EDF with tuple (θ, ϕ, v) is denoted by $EDF_{v,\phi}(\theta)$. The EDF includes the normal, binomial, Poisson, gamma and inverse Gaussian distributions; the specific distributional choice is characterized by the functional form of the cumulant function κ. The canonical parameter θ of the distribution $EDF_{v,\phi}(\theta)$ is our main object of interest (and to be estimated). It is well known that $Y \sim EDF_{v,\phi}(\theta)$ has expected value $\mu := \mathbb{E}[Y] = \kappa'(\theta)$. Based on the convexity of κ, there is a one-to-one correspondence between the mean parametrization μ and the canonical parameter θ of the selected member of the EDF. We freely switch between these two parametrizations.

We consider a test sample $\mathcal{D} = (y_i, \mathbf{x}_i, v_i)_{i=1}^n$ generated by i.i.d. observations $(Y_i, \mathbf{X}_i, V_i) \sim F_{(Y,\mathbf{X},V)}$. For each $i = 1, \ldots, n$, we assume that the conditional distribution of the response satisfies $Y_i|\mathbf{x}_i, v_i \sim EDF_{v_i,\phi}(\theta_i^*)$, and there exists a regression function f^* such that

$$\theta_i^* = f^*(\mathbf{x}_i) \quad \text{and} \quad \mu_i^* = \kappa'\big(\theta_i^*\big) = \kappa'\big(f^*(\mathbf{x}_i)\big);$$

the upper star * labels the true data generating model. Thus, the true conditionally expected

value of $Y_i|\mathbf{x}_i, v_i \sim EDF_{v_i,\phi}(\theta_i^*)$ is equal to μ_i^*, for all $i = 1, \ldots, n$. Equivalently, in the context of regression modeling, it is more common to assume that there exists a regression function g^* such that $\mu_i^* = g^*(\mathbf{x}_i)$ and $\theta_i^* = (\kappa')^{-1}(\mu_i^*)$, for all $i = 1, \ldots, n$; the function $(\kappa')^{-1}$ is the canonical link of the selected EDF member.

Under an unknown true model, we assume that the true regression function f^* has been estimated by a regression function $\widehat{f}$ being based on a learning sample $\mathcal{L} = (y_i^\ell, \mathbf{x}_i^\ell, v_i^\ell)_{i=1}^m$ coming from i.i.d. observations $(Y_i, \mathbf{X}_i, V_i) \sim F_{(Y,\mathbf{X},V)}$. The samples $\mathcal{D}$ and $\mathcal{L}$ are assumed to be independent and sampled from the same distribution $F_{(Y,\mathbf{X},V)}$. Given the estimated regression function $\widehat{f}$, we estimate the canonical parameters and the mean values, respectively, on the test sample $\mathcal{D}$ by

$$\widehat{\theta}_i = \widehat{f}(\mathbf{x}_i), \quad \widehat{\mu}_i = \kappa'\big(\widehat{\theta}_i\big), \quad i = 1, \ldots, n. \tag{2.1}$$

We assume that there exists a strictly increasing mapping $\mathcal{T}$ such that

$$\mathcal{T}(\widehat{f}(\mathbf{X})) = f^*(\mathbf{X}), \quad \text{almost surely,} \tag{2.2}$$

that is we assume strict comonotonicity between $\widehat{f}(\mathbf{X})$ and $f^*(\mathbf{X})$. We emphasize that this is stronger than comonotonicity which requires a mapping $\mathcal{T}$ that is non-decreasing. The assumption of strict comonotonicty is common in statistical inference for isotonic regression, see, e.g., Wright (1981) and (Wüthrich & Ziegel, 2024, Section 3.1). We need it for the correspondence between our formulations of the statistical hypothesis of mean-calibration.

The goal is to validate mean-calibration of these estimates (2.1) for the observations in the test sample $\mathcal{D}$. Let us point out that the member of the EDF and the dispersion parameter ϕ are assumed to be given, only the canonical parameters and the mean values of the conditional distributions of $\big(Y_i|\mathbf{x}_i, v_i\big)_{i=1}^n$ are estimated and validated. We aim at verifying the null hypothesis

$$\mathcal{H}_0 : \text{the mean predictions } \big(\widehat{\mu}_i\big)_{i=1}^n \text{ are calibrated,}$$

against the alternative hypothesis

$$\mathcal{H}_1 : \text{the mean predictions } \big(\widehat{\mu}_i\big)_{i=1}^n \text{ are not calibrated.}$$

We introduce our hypotheses more formally. Let $\widehat{\mu} : \mathcal{X} \mapsto \mathbb{R}$ denote the estimated regression function. We consider a measurable space $(\Omega, \mathcal{F})$, and we let $\mathcal{M}$ denote a set of all probability measures on $\big(\Omega, \mathcal{F}\big)$. The set $\mathcal{M}$ contains distributions of the i.i.d. observations $(Y_i, \mathbf{X}_i, V_i)_{i=1}^n$ such that, for all $i = 1, \ldots, n$, the variables $Y_i|\widehat{\mu}(\mathbf{X}_i), V_i \sim EDF_{V_i,\phi}\big(\Psi_i\big)$ for the given EDF member considered with dispersion parameter ϕ and for some canonical parameter $\Psi_i = \Psi(\mathbf{X}_i)$. Remark that without the strict comontonicity assumption (2.2), $Y_i|\widehat{\mu}(\mathbf{X}_i), V_i$ will not generally be a member of the EDF. Using the definition of calibration (1.1), we verify the null hypothesis

$$\mathcal{H}_0 = \Big\{\mathbb{P} \in \mathcal{M};\ \mathbb{E}_{\mathbb{P}}\big[Y_i|\widehat{\mu}\big(\mathbf{X}_i\big)\big] = \widehat{\mu}\big(\mathbf{X}_i\big),\ \text{for all } i = 1, \ldots, n\Big\},$$

against the alternative

$$\mathcal{H}_1 = \Big\{\mathbb{Q} \in \mathcal{M};\ \mathbb{E}_{\mathbb{Q}}\big[Y_i|\widehat{\mu}\big(\mathbf{X}_i\big)\big] \neq \widehat{\mu}\big(\mathbf{X}_i\big),\ \text{for some } i = 1, \ldots, n\Big\}.$$

Adapting to the notation of (Ramdas & Wang, 2025, Chapter 3.1), we introduce $\mathcal{P}$ and $\mathcal{Q}$ to denote the sets of probability measures from $\mathcal{M}$ which satisfy the null and the alternative hypothesis, respectively. Hence, the null (alternative) hypothesis is interpreted both as the statistical statement $\mathcal{H}_0$ ($\mathcal{H}_1$) and as the set of probability measures $\mathcal{P}$ ($\mathcal{Q}$). An element of $\mathcal{P}$ is denoted by $\mathbb{P}$, and an element of $\mathcal{Q}$ by $\mathbb{Q}$.

Remark 1. Within the EDF, there is the expected value $\kappa'(\theta) = \mathbb{E}[Y] = \mathbb{E}[Y|v]$, this is to say that the exposure $v > 0$ does not impact the expected value of Y within the EDF, only the higher moments depend on v. For this reason, within the EDF, in the null hypothesis and the alternative, conditioning on the weights V_i has no impact on the statements.

In the sequel, we assume that the test sample $\mathcal{D}$ is random, but the learning sample $\mathcal{L}$ is observed and fixed. Consequently, all the following results are conditional on $\mathcal{L}$, and the estimated regression functions $\widehat{f}$ and $\widehat{\mu}$, respectively, are treated as given. For this reason, we do not further use $\mathcal{L}$ below, but we directly work on the estimated regression functions $\widehat{f}$ and $\widehat{\mu}$; in fact, we could also drop the hats in the notation, because $\widehat{f}$ and $\widehat{\mu}$ can be understood as generic regression functions that are (out-of-sample) validated for mean-calibration. We simplify the notation as follows. We set $\widehat{\Theta}_i = \widehat{f}(\mathbf{X}_i)$ for the given regression function $\widehat{f}$, and we directly focus on the test sample $\mathcal{D} = (y_i, \widehat{\theta}_i, v_i)_{i=1}^n$ generated by the i.i.d. observations $(Y_i, \widehat{\Theta}_i, V_i)$, where $(\widehat{\Theta}_i)_{i=1}^n$ denote the estimates of the true canonical parameters of the conditional distributions $(Y_i|\mathbf{X}_i, V_i)_{i=1}^n$ using the covariate transformations $\widehat{\Theta}_i = \widehat{f}(\mathbf{X}_i)$. Under the strict comonotonicity assumption (2.2) we can reformulate the null hypothesis of mean-calibration as

$$\mathcal{H}_0 = \Big\{\mathbb{P} \in \mathcal{M};\ Y_i|\widehat{\Theta}_i, V_i \overset{\mathbb{P}}{\sim} EDF_{V_i,\phi}(\widehat{\Theta}_i),\ \text{for all } i = 1, \ldots, n\Big\}, \tag{2.3}$$

and the alternative hypothesis by

$$\mathcal{H}_1 = \Big\{\mathbb{Q} \in \mathcal{M};\ Y_i|\widehat{\Theta}_i, V_i \overset{\mathbb{Q}}{\not\sim} EDF_{V_i,\phi}(\widehat{\Theta}_i),\ \text{for some } i = 1, \ldots, n\Big\}, \tag{2.4}$$

the canonical parameters are linked with the means by the identity $\widehat{\mu}(\mathbf{X}_i) = \kappa'(\widehat{\Theta}_i)$; this uses the one-to-one correspondence between mean parameters and canonical parameters within the EDF.

3 Classical inference for mean estimates

We start with classical approaches of testing the null hypothesis (2.3) against the alternative hypothesis (2.4). These methods serve as benchmarks for universal inference presented in Section 5, below.

3.1 Isotonic regression

Many modern approaches to validate mean-calibration rely on isotonic regression; see, e.g., Dimitriadis et al. (2021), Gneiting & Resin (2023), Henzi et al. (2025) and Delong & Wüthrich (2025). For the reader's convenience, we briefly recall isotonic regression.

Isotonic regression is a rank based, non-parametric regression that preserves monotonicity in pre-specified ranks $(\pi_i)_{i=1}^n$ of $(Y_i)_{i=1}^n$. Assume we have data points $(y_i, \pi_i, v_i)_{i=1}^n$ with positive case

weights $v_i > 0$. The aim is to fit a non-parametric regression model to the responses $(y_i)_{i=1}^n$ respecting the ranks of $(\pi_i)_{i=1}^n$. This motivates the isotonic estimation problem

$$\begin{aligned}\widehat{\boldsymbol{\gamma}} \quad &= \quad \underset{\boldsymbol{\gamma}=(\gamma_1,\ldots,\gamma_n)^\top \in \mathbb{R}^n}{\arg\min} \sum_{i=1}^n v_i \left(y_i - \gamma_i\right)^2, \\ &\qquad \text{subject to} \quad \gamma_j \leq \gamma_k \Longleftrightarrow \pi_j \leq \pi_k, \quad \text{for all } 1 \leq j,k \leq n.\end{aligned}$$

This isotonic estimation problem can be solved with the pool adjacent violators (PAV) algorithm; see Leeuw et al. (2009). Typically, the solution $\widehat{\boldsymbol{\gamma}} \in \mathbb{R}^n$ is interpolated by a step function.

3.2 CORP reliability diagram

Popular methods in practice to assess calibration are graphical ones based on reliability diagrams. Reliability diagrams visualize the property of calibration (1.1) by plotting an estimate of the conditionally expected value $\mathbb{E}[Y|\widehat{\mu}(\mathbf{x}_i)]$ against the mean estimate $\widehat{\mu}(\mathbf{x}_i)$ for all instances $1 \leq i \leq n$. Intuitively, systematic deviations from the diagonal indicate a lack of calibration.

Dimitriadis et al. (2021) and Gneiting & Resin (2023) popularized so-called CORP reliability diagrams (Consistent, Optimally binned, Reproducible, Pool adjacent violators algorithm) to assess calibration. Their idea is to use isotonic regression to estimate the conditionally expected value $\mathbb{E}[Y|\widehat{\mu}(\mathbf{x}_i)]$ based on the ranking generated by $(\widehat{\mu}(\mathbf{x}_i))_{i=1}^n$. This estimation step is called *recalibration step*; see Wüthrich & Ziegel (2024). For the purpose of our applications below, we shall apply the recalibration step to all observations in the test set.

Isotonic recalibration of mean estimates

- Consider the sample $\big(Y_i, \widehat{\mu}(\mathbf{X}_i), V_i\big)_{i\in\mathcal{D}}$ with case weights $V_i > 0$.
- Compute the isotonic regression $m \mapsto \widehat{\mu}^{\text{iso}}_{\mathcal{D}}(m)$ based on the sample $\big(Y_i, \widehat{\mu}(\mathbf{X}_i), V_i\big)_{i\in\mathcal{D}}$ with ranking $\big(\widehat{\mu}(\mathbf{X}_i)\big)_{i\in\mathcal{D}}$ (we interpolate by a step function).
- Define the recalibrated mean estimates

$$\widehat{\mu}^{\text{rc}}\big(\mathbf{X}_i\big) = \widehat{\mu}^{\text{iso}}_{\mathcal{D}}\Big(\widehat{\mu}\big(\mathbf{X}_i\big)\Big), \quad i \in \mathcal{D},$$

  as estimates of $\mathbb{E}[Y_i|\widehat{\mu}(\mathbf{X}_i)]$.

The CORP reliability diagram is obtained by the two-dimensional plot of $(\widehat{\mu}\big(\mathbf{X}_i\big), \widehat{\mu}^{\text{rc}}\big(\mathbf{X}_i\big))_{i=1}^n$ which should roughly lie on the diagonal to have calibration. CORP reliability diagrams can be supplemented by point-wise consistency or confidence bands; see, e.g., Dimitriadis et al. (2021) and Delong & Wüthrich (2025). The latter reference uses bootstrap and Monte Carlo techniques for generating point-wise consistency bands for the recalibrated mean estimates within the EDF.

3.3 Classical Likelihood Ratio Test

In statistics, the LRT is a main tool for hypothesis testing. The LRT can also be used in our test setting of the null hypothesis (2.3) against the alternative (2.4). The likelihood of the conditional responses $(y_i)_{i=1}^n$ within the EDF is available, and the MLE of the mean values of the responses (under a weaker assumption than (2.2), namely, that the mean values are non-decreasing w.r.t. the initial mean estimates $(\widehat{\mu}(\mathbf{x}_i))_{i=1}^n$) is given by fitting an isotonic regression to $(y_i, \widehat{\mu}(\mathbf{x}_i), v_i)_{i=1}^n$, see Banerjee (2007). The LRT to validate the null hypothesis (2.3) of mean-calibration against the alternative (2.4) for the observations in the test set $\mathcal{D}$ then takes the following form:

Classical LRT for mean-calibration within the EDF

- Consider the sample $\big(Y_i, \kappa'(\widehat{\Theta}_i), V_i\big)_{i\in\mathcal{D}}$ with case weights $V_i > 0$.
- Estimate the isotonic regression $m \mapsto \widehat{\mu}_{\mathcal{D}}^{\text{iso}}(m)$ based on $(Y_i, \kappa'(\widehat{\Theta}_i), V_i)_{i\in\mathcal{D}}$ using the ranks $(\kappa'(\widehat{\Theta}_i))_{i\in\mathcal{D}}$.
- Estimate the canonical parameter Ξ_i for all observations $i \in \mathcal{D}$ by

$$\widehat{\Xi}_i = (\kappa')^{-1}\left(\widehat{\mu}_{\mathcal{D}}^{\text{iso}}\big(\kappa'(\widehat{\Theta}_i)\big)\right), \quad i \in \mathcal{D}.$$

- Define the Likelihood Ratio Statistics (LRS) for testing $\mathcal{H}_0$ against $\mathcal{H}_1$

$$E^{LRT} = \frac{\mathcal{L}_{\mathcal{D}}\big(\widehat{\Xi}\big)}{\mathcal{L}_{\mathcal{D}}\big(\widehat{\Theta}\big)} = \prod_{i\in\mathcal{D}} E_{\widehat{\Xi}_i}\big(Y_i, \widehat{\Theta}_i, V_i\big) = \prod_{i\in\mathcal{D}} \exp\left(\frac{Y_i\,(\widehat{\Xi}_i - \widehat{\Theta}_i) - (\kappa(\widehat{\Xi}_i) - \kappa(\widehat{\Theta}_i))}{\phi/V_i}\right), \quad (3.1)$$

  where $(\widehat{\Xi}_i)_{i=1}^n$ and $(\widehat{\Theta}_i)_{i=1}^n$ are the estimates of the canonical parameters constructed under $\mathcal{H}_1$ and $\mathcal{H}_0$, respectively, and $\mathcal{L}_{\mathcal{D}}(\cdot)$ denotes the likelihoods of the observations in $\mathcal{D}$ under the two estimates of the canonical parameters.

- Reject the null hypothesis $\mathcal{H}_0$ at significance level $\alpha \in (0,1)$ if

$$E^{LRT} > F^{-1}_{E^{LRT}|\mathcal{H}_0}(1-\alpha). \quad (3.2)$$

In our case, $(\widehat{\Theta}_i)_{i=1}^n$ are fully determined under $\mathcal{H}_0$ – with the specific mean estimates $(\widehat{\mu}(\mathbf{X}_i))_{i=1}^n$ – and $(\widehat{\Xi}_i)_{i=1}^n$ are estimated under $\mathcal{H}_1$ with the recalibrated mean estimates using isotonic regression. The most challenging part of the classical LRT is to work out the quantiles of the test statistics E^{LRT} under the null hypothesis $\mathcal{H}_0$ to determine the critical value of the statistical test in (3.2). Banerjee (2007) shows in his model that the LRS under the null hypothesis has asymptotically a limit distribution that is different from a χ^2-distribution, but which can be characterized in terms of a functional of Brownian motion. His characterization of the limit distribution of the LRS is difficult to apply in practice. Moreover, Banerjee (2007) considers a simplified case where the null hypothesis meets mean-calibration only for one observation. We are not aware of any extensions which cover the null hypothesis of mean-calibration for all observations.

3.4 Murphy's score decomposition

It is well established that the predictive accuracy of forecast models should be evaluated with strictly consistent loss functions; see Gneiting & Raftery (2007). Let L denote the deviance loss function of the selected EDF (which is a strictly consistent loss function for mean estimation); we refer to Wüthrich & Merz (2023). The predictive accuracy of the mean estimates $(\widehat{\mu}(\mathbf{X}_i))_{i=1}^n$ on the test set $\mathcal{D}$ is evaluated by the score (smaller means better)

$$S(\mathbf{Y}, \widehat{\boldsymbol{\mu}}, \mathbf{V}) = \frac{1}{\sum_{i=1}^n V_i} \sum_{i=1}^n V_i \, L(Y_i, \widehat{\mu}(\mathbf{X}_i)), \tag{3.3}$$

with responses $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$, mean estimates $\widehat{\boldsymbol{\mu}} = \big(\widehat{\mu}(\mathbf{X}_1), \ldots, \widehat{\mu}(\mathbf{X}_n)\big)^\top$ and case weights $\mathbf{V} = (V_1, \ldots, V_n)^\top$ from the test set $\mathcal{D}$. Following Murphy (1973), Dimitriadis et al. (2021), Gneiting & Resin (2023) and Pohle (2020), we decompose the score into uncertainty (UNC), discrimination (DSC) and miscalibration (MCB) statistics

$$S(\mathbf{Y}, \widehat{\boldsymbol{\mu}}, \mathbf{V}) = \mathrm{UNC}(\mathbf{Y}, \mathbf{V}) - \mathrm{DSC}(\mathbf{Y}, \widehat{\boldsymbol{\mu}}, \mathbf{V}) + \mathrm{MCB}(\mathbf{Y}, \widehat{\boldsymbol{\mu}}, \mathbf{V}), \tag{3.4}$$

where

$$\begin{aligned} \mathrm{UNC}(\mathbf{Y}, \mathbf{V}) &= S(\mathbf{Y}, \bar{\boldsymbol{\mu}}, \mathbf{V}), \\ \mathrm{DSC}(\mathbf{Y}, \widehat{\boldsymbol{\mu}}, \mathbf{V}) &= S(\mathbf{Y}, \bar{\boldsymbol{\mu}}, \mathbf{V}) - S(\mathbf{Y}, \widehat{\boldsymbol{\mu}}^{\mathrm{rc}}, \mathbf{V}) \;\geq\; 0, \\ \mathrm{MCB}(\mathbf{Y}, \widehat{\boldsymbol{\mu}}, \mathbf{V}) &= S(\mathbf{Y}, \widehat{\boldsymbol{\mu}}, \mathbf{V}) - S(\mathbf{Y}, \widehat{\boldsymbol{\mu}}^{\mathrm{rc}}, \mathbf{V}) \;\geq\; 0, \end{aligned}$$

with $\bar{\boldsymbol{\mu}} = (\bar{\mu}, \ldots, \bar{\mu})^\top \in \mathbb{R}^n$ having weighted mean entries $\bar{\mu} = \sum_{i=1}^n V_i Y_i / \sum_{i=1}^n V_i$ not considering features, and $\widehat{\boldsymbol{\mu}}^{\mathrm{rc}} = (\widehat{\mu}^{\mathrm{rc}}(\mathbf{X}_1), \ldots, \widehat{\mu}^{\mathrm{rc}}(\mathbf{X}_n))^\top$ is the recalibrated version of $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}(\mathbf{X}_1), \ldots, \widehat{\mu}(\mathbf{X}_n))^\top$. Moreover, MCB $= 0$ if and only if $\widehat{\boldsymbol{\mu}}^{\mathrm{rc}} \equiv \widehat{\boldsymbol{\mu}}$, and DSC $= 0$ if and only if $\widehat{\boldsymbol{\mu}}^{\mathrm{rc}} \equiv \bar{\boldsymbol{\mu}}$ (this follows from the strict consistency of L). We observe that the miscalibration statistics MCB expresses deviations of the CORP reliability diagram from the diagonal in terms of the score under consideration. Hence, there is a close link to the techniques presented in Section 3.2. In our EDF deviance loss case, the miscalibration statistics for the mean estimates $(\widehat{\mu}(\mathbf{X}_i))_{i=1}^n$ takes the form

$$\mathrm{MCB}(\mathbf{Y}, \widehat{\boldsymbol{\mu}}, \mathbf{V}) = \frac{1}{\sum_{i=1}^n V_i} \sum_{i=1}^n \frac{Y_i \, (\widehat{\Xi}_i - \widehat{\Theta}_i) - (\kappa(\widehat{\Xi}_i) - \kappa(\widehat{\Theta}_i))}{\phi / V_i}, \tag{3.5}$$

where $(\widehat{\Theta}_i)_{i=1}^n$ are the predictions of the canonical parameters derived from $(\widehat{\mu}(\mathbf{X}_i))_{i=1}^n$, and $(\widehat{\Xi}_i)_{i=1}^n$ denote the canonical parameters derived from the recalibrated mean estimates, see Sections 3.2-3.3. We conclude that the miscalibration statistics from Murphy's score decomposition is related to the LRS for testing the null hypothesis (2.3) against the alternative (2.4) by the simple transformation

$$\mathrm{MCB}(\mathbf{Y}, \widehat{\boldsymbol{\mu}}, \mathbf{V}) = \log\left(\frac{\mathcal{L}_{\mathcal{D}}(\widehat{\Xi})}{\mathcal{L}_{\mathcal{D}}(\widehat{\Theta})}\right) \Big/ \left(\sum_{i=1}^n V_i\right). \tag{3.6}$$

Henceforth, we can equivalently use the LRS and the MCB to test calibration of the mean estimates within the EDF, i.e., under the corresponding deviance loss choice.

4 Universal inference for mean estimates

In this section, we develop our new approach to assess calibration (1.1) based on universal inference and the *split LRT*. This approach uses e-values and e-testing, and it is essentially based on the work of Ramdas & Wang (2025).

4.1 Split Likelihood Ratio Test

We first introduce basic tools of universal inference; for a comprehensive mathematical introduction, see Ramdas & Wang (2025). We work with classical parametric statistical models. Consider i.i.d. observations $\mathbf{Y} = (Y_1, \ldots, Y_n)$ from a distribution $\mathbb{P}_{\theta^*}$ that belongs to a collection of distributions $\{\mathbb{P}_\theta;\, \theta \in \mathbf{\Theta}\}$. A general aim is to test

$$\mathcal{H}_0 : \theta^* \in \mathbf{\Theta}_0 \qquad \text{against} \qquad \mathcal{H}_1 : \theta^* \in \mathbf{\Theta}_1 = \mathbf{\Theta} \setminus \mathbf{\Theta}_0,$$

for a non-empty null hypothesis set $\mathbf{\Theta}_0 \subsetneq \mathbf{\Theta}$. Consider the LRS, and assume that under the null hypothesis $\mathcal{H}_0$ the LRS satisfies the asymptotic behavior

$$2\log\left(\frac{\mathcal{L}_{\mathbf{Y}}\big(\widehat{\theta}_1\big)}{\mathcal{L}_{\mathbf{Y}}\big(\widehat{\theta}_0\big)}\right) \sim \chi^2\big(\dim(\mathbf{\Theta}_1) - \dim(\mathbf{\Theta}_0)\big), \qquad \text{as } n \to \infty,$$

where $\widehat{\theta}_1$ and $\widehat{\theta}_0$ are the MLEs of the parameters under $\mathcal{H}_1$ and $\mathcal{H}_0$, respectively, and $\mathcal{L}_{\mathbf{Y}}(\cdot)$ denotes the data likelihoods of $\mathbf{Y}$ under the given parameter. Such a LRS limit relies on a large sample asymptotics which requires regularity conditions. These conditions are likely not fulilled in many practical applications. Wasserman et al. (2020) introduce universal inference which yields tests and confidence sets for any statistical model and with finite sample guarantees without restrictive regularity conditions. In particular, Wasserman et al. (2020) propose to use the so-called split LRS, justification will be given below.

Split Likelihood Ratio Test

- Partition the test sample $\mathbf{Y} = (Y_1, \ldots, Y_n)$ at random into two disjoint non-empty sub-samples $\mathbf{Y}_0$ and $\mathbf{Y}_1$.
- Let $\widehat{\theta}_1$ denote an estimator of the model parameter $\theta^* \in \mathbf{\Theta}_1$ under $\mathcal{H}_1$ based on sub-sample $\mathbf{Y}_1$ (e.g., the MLE constructed under $\mathcal{H}_1$).
- Let $\widehat{\theta}_0$ denote the MLE constructed under $\mathcal{H}_0$ based on $\mathbf{Y}_0$.
- Calculate the split LRS for testing $\mathcal{H}_0$ against $\mathcal{H}_1$

$$E^{sLRT} = \frac{\mathcal{L}_{\mathbf{Y}_0}\big(\widehat{\theta}_1\big)}{\mathcal{L}_{\mathbf{Y}_0}\big(\widehat{\theta}_0\big)}. \tag{4.1}$$

- Reject the null hypothesis $\mathcal{H}_0$ at a pre-assumed significance level $\alpha \in (0,1)$ if

$$E^{sLRT} \geq 1/\alpha. \tag{4.2}$$

The key difference between the classical LRS (3.1) and the split LRS (4.1) is that the former uses an in-sample likelihood ratio whereas the latter uses an out-of-sample likelihood ratio. We emphasize that when implementing the split LRT, a part of the sample $\mathbf{Y}$ is used for estimating the parameter of a model under the alternative $\mathcal{H}_1$ (training a model) and the remaining part is used for validating the null hypothesis $\mathcal{H}_0$ about the parameter of a model (testing a model). In other words, the test sample $\mathbf{Y}$ is split into a *training set* $\mathbf{Y}_1$ and a *validation set* $\mathbf{Y}_0$ for hypothesis testing.

The universal approach developed by Wasserman et al. (2020) considers a universal critical value at the selected significance level α for the statistics E^{sLRT} given by $1/\alpha$; we also refer to (Ramdas & Wang, 2025, Section 5.2). This is justified by the following theorem.

Theorem 1. *The split LRT* (4.2) *controls the type I error of rejecting* $\mathcal{H}_0$ *in favor of* $\mathcal{H}_1$ *at significance level* $\alpha \in (0,1)$*, i.e.,*

$$\sup_{\theta^* \in \mathbf{\Theta}_0} \mathbb{P}_{\theta^*}\left(E^{sLRT} \geq 1/\alpha\right) \leq \alpha.$$

Generally, one can expect that the type I error is much lower than α if the universal critical value $1/\alpha$ is used for rejecting the null hypothesis.

The proof of Theorem 1 is straightforward. It is based on Markov's inequality and the property that the expected value of the split LRS is less or equal to one under the null hypothesis; this comes from the fact that $\mathcal{L}_{\mathbf{Y}_0}(\theta^*)/\mathcal{L}_{\mathbf{Y}_0}(\widehat{\theta}_0) \leq 1$ for all $\theta^* \in \mathbf{\Theta}_0$, since $\widehat{\theta}_0$ is the MLE over $\mathbf{\Theta}_0$, for observations $\mathbf{Y}_0$. In other words, the split LRS (4.1) is an *e-variable* under the null hypothesis $\mathcal{H}_0$. For an e-variable, we recall Definition 1.2 of Ramdas & Wang (2025). In the definition below and in the sequel, we now interpret the null hypothesis $\mathcal{H}_0$ and the alternative $\mathcal{H}_1$ as sets of probability measures $\mathcal{P}$ and $\mathcal{Q}$, see also Section 2.

Definition 2. *Let* $\mathcal{P}$ *denote the set of probability measures representing the null hypothesis* $\mathcal{H}_0$*.*

- *An e-variable* E *for* $\mathcal{P}$ *is an* $[0,\infty]$*-valued random variable satisfying* $\mathbb{E}_{\mathbb{P}}[E] \leq 1$ *for all* $\mathbb{P} \in \mathcal{P}$*.*
- *An e-variable* E *is exact if* $\mathbb{E}_{\mathbb{P}}[E] = 1$ *for all* $\mathbb{P} \in \mathcal{P}$*.*

The split LRS involves statistical randomness (irreducible risk), due to the randomness in the observations $\mathbf{Y}$, and it involves algorithmic randomness, due to the partition of the data $\mathbf{Y}$ into the two sub-samples $\mathbf{Y}_0$ and $\mathbf{Y}_1$. In the classical LRS, only the statistical randomness is present. This motivates one to try to reduce the algorithmic randomness in the split LRS by (independently) repeating the partition of $\mathbf{Y}$ and averaging the resulting split LRSs. This suggests to consider a repeated independent sampling of $\mathbf{Y}_0$ and $\mathbf{Y}_1$. We define the *sub-sampled split LRS*

$$\overline{E}_B^{sLRT} = \frac{1}{B}\sum_{b=1}^{B} E_b^{sLRT}, \tag{4.3}$$

where E_b^{sLRT} denotes the split LRS of the b-th (independent) random partition of $\mathbf{Y}$. The test statistics $\overline{E}_B^{sLRT}$ is an e-variable, if E_b^{sLRT} are e-variables, and one immediately concludes with the following statement from Theorem 1.

Corollary 1. *Let $\mathcal{P}$ denote the set of probability measures representing the null hypothesis $\mathcal{H}_0$. Reject the null hypothesis $\mathcal{H}_0$ at a pre-assumed significance level $\alpha \in (0,1)$ if $\overline{E}_B^{sLRT} \geq 1/\alpha$. This test controls the type I error of rejecting $\mathcal{H}_0$ in favor of $\mathcal{H}_1$ at significance level $\alpha \in (0,1)$, i.e.,*

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\left(\overline{E}_B^{sLRT} \geq 1/\alpha\right) \leq \alpha.$$

In practical applications, we are interested in the power of the statistical tests. Traditionally, we could estimate the probability that the null hypothesis is rejected under the true alternative, as we would do in classical inference. However, in the context of e-variables, one rather relies on *e-power* as the suitable measure in universal inference; see (Ramdas & Wang, 2025, Chapter 3).

Definition 3. *Let $\mathcal{Q}$ denote the set of probability measures representing the alternative $\mathcal{H}_1$.*

- *The power of a test based on an e-variable E against a simple alternative $\mathbb{Q} \in \mathcal{Q}$ is $\mathbb{E}_{\mathbb{Q}}[E]$. An e-variable E is powered against $\mathcal{Q}$ if $\mathbb{E}_{\mathbb{Q}}[E] > 1$ for all $\mathbb{Q} \in \mathcal{Q}$.*
- *The e-power of an e-variable E against a simple alternative $\mathbb{Q} \in \mathcal{Q}$ is $\mathbb{E}_{\mathbb{Q}}[\log(E)]$, assuming the expectation exists. An e-variable E has a positive e-power against $\mathcal{Q}$ if $\mathbb{E}_{\mathbb{Q}}[\log(E)] > 0$ for all $\mathbb{Q} \in \mathcal{Q}$.*

Jensen's inequality implies $\mathbb{E}_{\mathbb{Q}}[\log(E)] > 0$ provides $\mathbb{E}_{\mathbb{Q}}[E] > 1$. The intuition behind these definitions is that the test statistics should be large under the true alternative to correctly reject a false null hypothesis. As discussed in (Ramdas & Wang, 2025, Section 3.8), the larger the e-power of a test statistics, the better the test is based on that statistics, and the e-power of the test statistics can be used as a comparative notion. The link between the e-power and the power of a statistical test is discussed in the next section; we also refer to (Ramdas & Wang, 2025, Proposition 5.3).

The sub-sampled split LRS is an example of an e-variable constructed by merging e-variables given by the split LRS. The main challenge in universal inference is to find an e-variable by merging basic e-variables so that the resulting test statistics has a large power.

4.2 Sub-sampled split LRT for mean-calibration within the EDF

The previous section on the sub-sampled split LRS was rather generic. We now apply this method to evaluate mean-calibration within the EDF. The results of Wasserman et al. (2020) motivate the following statistical test of the null hypothesis $\mathcal{H}_0$ of mean-calibration (2.3) against the alternative $\mathcal{H}_1$ given by (2.4) for an observed test sample $\mathcal{D}$.

Sub-sampled split LRT for mean-calibration within the EDF

- Consider the sample $(Y_i, \kappa'(\widehat{\Theta}_i), V_i)_{i\in\mathcal{D}}$ with case weights $V_i > 0$.
- Choose a fixed split ratio $s \in (0,1)$.

- Select a sub-sample $\mathcal{D}_0$ by randomly choosing $[ns]$ tuples from $(Y_i, \widehat{\Theta}_i, V_i)_{i\in\mathcal{D}}$ without replacement, and set $\mathcal{D}_1 = \{1, \ldots, n\} \setminus \mathcal{D}_0$.
- Fit an isotonic regression $m \mapsto \widehat{\mu}^{\text{iso}}_{\mathcal{D}_1}(m)$ on $(Y_i, \kappa'(\widehat{\Theta}_i), V_i)_{i\in\mathcal{D}_1}$ with ranks $(\kappa'(\widehat{\Theta}_i))_{i\in\mathcal{D}_1}$.
- Estimate Ξ_i for all instances $i \in \mathcal{D}_0$ by

$$\widehat{\Xi}_i = (\kappa')^{-1}\Big(\widehat{\mu}^{\text{iso}}_{\mathcal{D}_1}\big(\kappa'(\widehat{\Theta}_i)\big)\Big). \tag{4.4}$$

- Define the split LRS for testing $\mathcal{H}_0$ against $\mathcal{H}_1$

$$E^{sLRT} = \frac{\mathcal{L}_{\mathcal{D}_0}\big(\widehat{\Xi}\big)}{\mathcal{L}_{\mathcal{D}_0}\big(\widehat{\Theta}\big)} = \prod_{i\in\mathcal{D}_0} E_{\widehat{\Xi}_i}\big(Y_i, \widehat{\Theta}_i, V_i\big) = \prod_{i\in\mathcal{D}_0} \exp\left(\frac{Y_i\,(\widehat{\Xi}_i - \widehat{\Theta}_i) - (\kappa(\widehat{\Xi}_i) - \kappa(\widehat{\Theta}_i))}{\phi/V_i}\right), \tag{4.5}$$

  and, by repeating the previous items, the sub-sampled split LRS for testing $\mathcal{H}_0$ against $\mathcal{H}_1$

$$\overline{E}^{sLRT}_B = \frac{1}{B}\sum_{b=1}^{B} E^{sLRT}_b, \tag{4.6}$$

  where E^{sLRT}_b are conditional i.i.d. copies of E^{sLRT}, with the i.i.d. applying to the partitioning of the set $\mathcal{D}$ into $\mathcal{D}_0$ and $\mathcal{D}_1$ (i.e., this is repeated B times).
- Reject the null hypothesis $\mathcal{H}_0$ at a pre-assumed significance level $\alpha \in (0, 1)$ if

$$\overline{E}^{sLRT}_B \geq 1/\alpha. \tag{4.7}$$

For $B = 1$, the sub-sampled split LRS turns into the split LRS.

Theorem 2. *Consider the null hypothesis $\mathcal{H}_0$ of mean-calibration against the alternative $\mathcal{H}_1$.*

- *The split LRS* (4.5) *and the sub-sampled split LRS* (4.6) *are exact e-variables for the set $\mathcal{P}$ representing the null hypothesis $\mathcal{H}_0$.*
- *The split and sub-sampled split LRTs* (4.7) *control the type I error of rejecting $\mathcal{H}_0$ in favor of $\mathcal{H}_1$ at significance level $\alpha \in (0, 1)$.*

Proof: Select $\mathbb{P} \in \mathcal{P}$ from the null hypothesis $\mathcal{H}_0$ set given by (2.3); in fact, $\mathcal{P}$ is a singleton. We prove $\mathbb{E}_{\mathbb{P}}[E^{sLRT}] = 1$. We compute

$$\mathbb{E}_{\mathbb{P}}\left[E^{sLRT}\right] = \mathbb{E}_{\mathbb{P}}\left[\mathbb{E}_{\mathbb{P}}\left[E^{sLRT}\,\middle|\mathcal{D}_1, \left(\widehat{\Theta}_i, V_i\right)_{i\in\mathcal{D}_0}\right]\right].$$

Given $\{\mathcal{D}_1, (\widehat{\Theta}_i, V_i)_{i\in\mathcal{D}_0}\}$, the variables $\left(\widehat{\Xi}_i, \widehat{\Theta}_i, V_i\right)_{i\in\mathcal{D}_0}$ are fixed, and the responses $(Y_i)_{i\in\mathcal{D}_0}$ are conditionally independent. Using the moment generating function for the conditional distribution of $Y_i|\mathcal{D}_1, \widehat{\Theta}_i, V_i$, with canonical parameter $\widehat{\Theta}_i$ under $\mathcal{H}_0$, within the EDF, proves the first claim. The second assertion follows from the theory of e-variables and Theorem 1. □

The next proposition links the e-power and the classical power of the split LRT.

Proposition 1. *Let $\mathcal{D}$ denote the test sample of size $n = |\mathcal{D}|$, and split $\mathcal{D}$ into $\mathcal{D}_0$ and $\mathcal{D}_1$ with sizes $n_0 = |\mathcal{D}_0|$ and $n_1 = |\mathcal{D}_1|$. Consider the split LRS given by, see also* (4.5)*,*

$$E^{sLRT}_{[n_0]} = \prod_{i \in \mathcal{D}_0} E_{\widehat{\Xi}_i}\big(Y_i, \widehat{\Theta}_i, V_i\big). \tag{4.8}$$

Choose $\mathbb{Q} \in \mathcal{Q}$ from the set of probability measures representing the alternative hypothesis $\mathcal{H}_1$. If $\mathbb{E}_{\mathbb{Q}}\big[\log(E^{sLRT}_{[n_0]})\big|\mathcal{D}_1\big] > 0$ for all $\mathcal{D}_1$ with fixed size n_1, then $\mathbb{Q}\big(E^{sLRT}_{[n_0]} > 1/\alpha\big) \to 1$ as $n_0 \to \infty$ for any $\alpha > 0$. If $\mathbb{E}_{\mathbb{Q}}\big[\log(E^{sLRT}_{[n_0]})\big|\mathcal{D}_1\big] < 0$ for all $\mathcal{D}_1$ with fixed size n_1, then $\mathbb{Q}\big(E^{sLRT}_{[n_0]} > 1/\alpha\big) \to 0$ as $n_0 \to \infty$ for any $\alpha > 0$.

Proof: We analyze the probability

$$\mathbb{Q}\left(E^{sLRT}_{[n_0]} > 1/\alpha\right) = \mathbb{E}_{\mathbb{Q}}\left[\mathbb{Q}\left(E^{sLRT}_{[n_0]} > 1/\alpha \middle| \mathcal{D}_1\right)\right].$$

Given $\mathcal{D}_1$, the variables $(E_{\widehat{\Xi}_i}(Y_i, \widehat{\Theta}_i, V_i))_{i \in \mathcal{D}_0}$ are i.i.d., and we are in the classical setting of multiplicative i.i.d. e-variables. We deduce that

$$E^{sLRT}_{[n_0]}\big|\mathcal{D}_1 \approx e^{n_0 \mathbb{E}\left[\log\left(E_{\widehat{\Xi}_i}\left(Y_i, \widehat{\Theta}_i, V_i\right)\right)\big|\mathcal{D}_1\right]}, \tag{4.9}$$

for large $n_0 \to \infty$. From (Ramdas & Wang, 2025, Proposition 3.12) we conclude $\mathbb{Q}\big(E^{sLRT}_{[n_0]} > 1/\alpha\big|\mathcal{D}_1\big) \to 1$ as $n_0 \to \infty$ for any $\alpha > 0$. The result then follows by dominated convergence. □

The conclusion from Proposition 1 is that an a.s. positive conditional e-power of the split LRS (4.8) leads to an asymptotic power-1 test as desired, and an a.s. negative conditional e-power of the split LRS (4.8) leads to an asymptotic power-0 test.

4.3 Sub-sampled split LqRT for mean-calibration within the EDF

The e-variable based on the LRS (4.5) is not the only e-variable in our statistical setting of testing the null hypothesis (2.3) against the alternative (2.4). We introduce a novel test statistics in this section. The idea for this novel test statistics comes from robust likelihood and Lq-likelihood, see Ferrari & Yang (2010), Qin & Priebe (2017) and (Ramdas & Wang, 2025, Section 5.6.3).

Definition 4. *Choose $q \in (0, 1]$. Define the split Lq-Likelihood Ratio Statistics (LqRS) by*

$$\begin{aligned} E^{sLRT,q} &= \prod_{i \in \mathcal{D}_0} E^q_{\widehat{\Xi}_i}\big(Y_i, \widehat{\Theta}_i, V_i\big) \\ &= \prod_{i \in \mathcal{D}_0} \exp\left(\frac{q\, Y_i\left(\widehat{\Xi}_i - \widehat{\Theta}_i\right) - \left(\kappa\left(q\,\widehat{\Xi}_i + (1-q)\widehat{\Theta}_i\right) - \kappa\left(\widehat{\Theta}_i\right)\right)}{\phi/V_i}\right). \end{aligned} \tag{4.10}$$

Theorem 3. *Consider the null hypothesis $\mathcal{H}_0$ of calibration against the alternative $\mathcal{H}_1$. For any $q \in (0, 1]$, the split LqRS* (4.10) *is an exact e-variable for $\mathcal{P}$ representing the null hypothesis $\mathcal{H}_0$.*

The proof is analogous to the one of Theorem 2. Choosing $q = 1$ in (4.10), we rediscover the split LRS (4.5). The main key question is which choice $q \in (0, 1]$ provides us with the largest power of the split LqRT? The next two theorems shed more light on the power of the split LqRT.

Within the EDF and under the alternative $\mathcal{H}_1$, for every measure $\mathbb{Q} \in \mathcal{Q}$ from the set representing the alternatives, there exists an instance $i \in \{1, \ldots, n\}$ such that

$$Y_i|\widehat{\Theta}_i, V_i \overset{\mathbb{Q}}{\sim} EDF_{V_i,\phi}(\Pi_i) \overset{(\mathrm{d})}{\neq} EDF_{V_i,\phi}(\widehat{\Theta}_i).$$

That is, there is a true canonical parameter $\Pi_i \neq \widehat{\Theta}_i$ for some instance i under $\mathbb{Q}$. In the sequel, we write for the taken canonical parameters $\Pi = (\Pi_1, \ldots, \Pi_n)$ in the corresponding notation $\mathbb{Q} = \mathbb{Q}_\Pi$.

Theorem 4. *Choose an element $\mathbb{Q}_\Pi = \mathbb{Q} \in \mathcal{Q}$ from the set of measures representing the alternative $\mathcal{H}_1$, giving the canonical parameters $\Pi = (\Pi_1, \ldots, \Pi_n)$. For $i \in \mathcal{D}_0$, consider the function*

$$f_i(q) = \mathbb{E}_{\mathbb{Q}_\Pi}\left[E^q_{\widehat{\Xi}_i}\left(Y_i, \widehat{\Theta}_i, V_i\right)\middle|\mathcal{D}_1, (\widehat{\Theta}_i, V_i)_{i\in\mathcal{D}_0}\right], \quad q \in (0, 1]. \tag{4.11}$$

Assume $\widehat{\Xi}_i > \widehat{\Theta}_i$. This gives us the following property: if $\Pi_i > \widehat{\Theta}_i$, then $f_i(q) > 1$ for all $q \in (0, 1]$.

Proof: Given $\{\mathcal{D}_1, (\widehat{\Theta}_i, V_i)_{i\in\mathcal{D}_0}\}$, the variables $\left(\widehat{\Xi}_i, \widehat{\Theta}_i\right)$ are fixed, and we can use the moment generating function for the conditional distribution of $Y_i|\mathcal{D}_1, \widehat{\Theta}_i, V_i$ with the canonical parameter Π_i under $\mathbb{Q}_\Pi \in \mathcal{Q}$ within the EDF. We derive from this moment generating function

$$\begin{aligned}
\frac{\phi}{V_i}\log(f_i(q)) &= \left(\kappa\left(\Pi_i + q\left(\widehat{\Xi}_i - \widehat{\Theta}_i\right)\right) - \kappa\left(\Pi_i\right) - \left(\kappa\left(q\,\widehat{\Xi}_i + (1-q)\,\widehat{\Theta}_i\right) - \kappa\left(\widehat{\Theta}_i\right)\right)\right) \\
&= \left(\kappa\left(\Pi_i + q\left(\widehat{\Xi}_i - \widehat{\Theta}_i\right)\right) - \kappa\left(\Pi_i\right) - \left(\kappa\left(\widehat{\Theta}_i + q\left(\widehat{\Xi}_i - \widehat{\Theta}_i\right)\right) - \kappa\left(\widehat{\Theta}_i\right)\right)\right) > 0,
\end{aligned}$$

the last inequality follows from strict convexity of the cumulant function κ. This leads to increasing differences. □

Remark that function (4.11) allows us to identify the alternative hypotheses (the probability measures) in (2.4) against which the e-variable (4.10) is powered; see Definition 3. We investigate a conditional version of (4.10) for one instance i. Under the assumption $\widehat{\Xi}_i > \widehat{\Theta}_i$, we should require that $\Pi_i > \widehat{\Theta}_i$ in order to obtain $f_i(q) > 1$ for all $q \in (0, 1]$. By complete symmetry, under the assumption $\widehat{\Xi}_i < \widehat{\Theta}_i$, we should require that $\Pi_i < \widehat{\Theta}_i$ to force that $f_i(q) > 1$ for all $q \in (0, 1]$. The conclusions from Theorem 4 are that the isotonic estimate $\widehat{\Xi}_i$ and the ground truth Π_i should be on the same side of the estimate $\widehat{\Theta}_i$ (specified in the null hypothesis) for the calibration test to have power against the true alternative Π_i. This property has been identified by Henzi et al. (2025) within the class of binomial models, and we have just extended it to the entire EDF.

From a practical point of view, if we under-estimate (over-estimate) the true value Π_i by the original estimate $\widehat{\Theta}_i$, i.e., if $\Pi_i > \widehat{\Theta}_i$ ($\Pi_i < \widehat{\Theta}_i$), then the split LqRT for any $q \in (0, 1]$ – including the split LRT – has power in rejecting the false $\widehat{\Theta}_i$ from the null hypothesis against the true alternative Π_i, if the isotonic regression estimate $\widehat{\Xi}_i$ uncovers this under-estimation (over-estimation) of Π_i with $\widehat{\Theta}_i$, i.e., if $\widehat{\Xi}_i > \widehat{\Theta}_i$ ($\widehat{\Xi}_i < \widehat{\Theta}_i$). This property may be hard to guarantee in small samples because it also relies on the accuracy on the isotonic recalibration step, but it is likely to hold in large samples where the isotonic regression attains good accuracy. If we look for an e-variable which is powered against the true alternative, then the split LRT and the split LqRT with any $q \in (0, 1]$ have the same property in this regard.

Theorem 5. *Choose an element $\mathbb{Q}_\Pi = \mathbb{Q} \in \mathcal{Q}$ from the set of measures representing the alternative $\mathcal{H}_1$, giving the canonical parameters $\Pi = (\Pi_1, \ldots, \Pi_n)$. For $i \in \mathcal{D}_0$, consider the function*

$$F_i(t) = \mathbb{E}_{\mathbb{Q}_\Pi}\Big[\log\left(E^q_{\widehat{\Xi}_i}\left(Y_i, \widehat{\Theta}_i, V_i\right)\right) \Big| \mathcal{D}_1, (\widehat{\Theta}_i, V_i)_{i\in\mathcal{D}_0} \Big], \quad q \in (0,1]. \tag{4.12}$$

Assume $\widehat{\Xi}_i > \widehat{\Theta}_i$ and $\Pi_i > \widehat{\Theta}_i$. We have the properties:

(a) If $\Pi_i > \widehat{\Xi}_i$, then $\arg\sup_{q\in(0,1]} F_i(q) = 1$.

(b) If $\Pi_i \in (\widehat{\Theta}_i, \widehat{\Xi}_i)$, then $\arg\sup_{q\in(0,1]} F_i(q) \in (0,1)$.

(c) There is a unique $\Pi_i^* \in (\widehat{\Theta}_i, \widehat{\Xi}_i)$ such that

$$\kappa'\left(\Pi_i^*\right) = \frac{\kappa\left(\widehat{\Xi}_i\right) - \kappa\left(\widehat{\Theta}_i\right)}{\widehat{\Xi}_i - \widehat{\Theta}_i}.$$

(d) If $\Pi_i > \Pi_i^*$, then we have $F_i(q) > 0$ for all $q \in (0,1]$.

(e) If $\Pi_i \in (\widehat{\Theta}_i, \Pi_i^*)$, then there exists $q^* \in (0,1)$ such that $F_i(q) > 0$ for all $t \in (0, q^*)$ and $F_i(q) < 0$ for all $q \in (q^*, 1]$.

Proof: Similar to the proof of Theorem 4, one derives

$$\frac{V_i}{\phi} F_i(q) = q\,\kappa'\left(\Pi_i\right)\left(\widehat{\Xi}_i - \widehat{\Theta}_i\right) - \left(\kappa\left(q\,\widehat{\Xi}_i + (1-q)\widehat{\Theta}_i\right) - \kappa\left(\widehat{\Theta}_i\right)\right). \tag{4.13}$$

We deduce that $F_i(0) = 0$. Moreover, the first derivative w.r.t. q is equal to

$$\begin{aligned} \frac{V_i}{\phi} F_i'(q) &= \kappa'\left(\Pi_i\right)\left(\widehat{\Xi}_i - \widehat{\Theta}_i\right) - \kappa'\left(q\,\widehat{\Xi}_i + (1-q)\widehat{\Theta}_i\right)\left(\widehat{\Xi}_i - \widehat{\Theta}_i\right) \\ &= \left(\widehat{\Xi}_i - \widehat{\Theta}_i\right)\left(\kappa'\left(\Pi_i\right) - \kappa'\left(\widehat{\Theta}_i + q\left(\widehat{\Xi}_i - \widehat{\Theta}_i\right)\right)\right), \end{aligned}$$

and from the second derivative we conclude concavity

$$\frac{V_i}{\phi} F_i''(q) = -\left(\widehat{\Xi}_i - \widehat{\Theta}_i\right)^2 \kappa''\left(\widehat{\Theta}_i + q\left(\widehat{\Xi}_i - \widehat{\Theta}_i\right)\right) < 0.$$

Moreover, the first derivative in zero satisfies

$$\frac{V_i}{\phi} F_i'(0) = \left(\widehat{\Xi}_i - \widehat{\Theta}_i\right)\left(\kappa'\big(\Pi_i\big) - \kappa'\big(\widehat{\Theta}_i\big)\right).$$

We immediately conclude that $F_i'(0) > 0$ and $F_i''(q) < 0$ for all $q \in (0,1]$ under the assumptions of the theorem. Hence, the function F_i is increasing in the neighbourhood of zero. The function F_i is increasing for all $q \in (0,1]$ if $F_i'(q) > 0$ for all $q \in (0,1]$. Using the concavity of F_i on $(0,1]$, the latter is equivalent to requiring

$$\frac{V_i}{\phi} F_i'(1) = \left(\widehat{\Xi}_i - \widehat{\Theta}_i\right)\left(\kappa'\big(\Pi_i\big) - \kappa'\big(\widehat{\Xi}_i\big)\right) > 0,$$

which is satisfied only if $\Pi_i > \widehat{\Xi}_i$. In this case, $\arg\sup_{q\in(0,1]} F_i(q) = 1$. This proves item (a).

Otherwise, if $\Pi_i < \widehat{\Xi}_i$, then $\arg\sup_{q\in[0,1]} F_i(q) \in (0,1)$, and the concave function F_i is increasing for $q \in (0, q^{\text{opt}})$ and decreasing for $q \in (q^{\text{opt}}, 1]$; this proves item (b). We investigate $F_i(1)$ and its sign. We have

$$\frac{V_i}{\phi} F_i(1) = \kappa'\big(\Pi_i\big) \left(\widehat{\Xi}_i - \widehat{\Theta}_i\right) - \left(\kappa\big(\widehat{\Xi}_i\big) - \kappa\big(\widehat{\Theta}_i\big)\right).$$

By the mean value theorem, there exists a unique $\Pi_i^* \in (\widehat{\Theta}_i, \widehat{\Xi}_i)$ such that

$$\kappa'\big(\Pi_i^*\big) \left(\widehat{\Xi}_i - \widehat{\Theta}_i\right) - \left(\kappa\big(\widehat{\Xi}_i\big) - \kappa\big(\widehat{\Theta}_i\big)\right) = 0;$$

recall that κ is strictly increasing; this proves item (c). If $\Pi_i > \Pi_i^*$, then $F_i(1) > 0$ and $F_i(q) > 0$ for all $q \in (0,1]$; this proves item (d). Otherwise, $F_i(1) < 0$ and the concave function F_i crosses zero from above at some point $q^* \in (0,1)$. This concludes the proof. □

Conclusions from Theorem 5. By symmetry, there is a similar result for $\widehat{\Xi}_i < \widehat{\Theta}_i$ and $\Pi_i < \widehat{\Theta}_i$. Function (4.12) allows us to identify the alternative hypotheses (the probability measures) in (2.4) for which our e-variables (4.10) have a positive e-power; see Definition 3. Consider a conditional version of (4.10) for one observation. As expected, the conditions of Theorem 5 are stronger than the ones of Theorem 4. From Theorem 4, we know that we should require the isotonic estimate $\widehat{\Xi}_i$ and the ground truth Π_i to be on the same side of the estimate $\widehat{\Theta}_i$, specified through the null hypothesis for a given data set. Under this assumption, the split LRT and the split LqRT for any $q \in (0,1]$ are powered against the true alternative. Yet, the tests have different e-powers against the true alternative. If $\widehat{\Xi}_i$ and Π_i are both larger (smaller) than $\widehat{\Theta}_i$, and in addition if $\Pi_i > \widehat{\Xi}_i$ ($\Pi_i < \widehat{\Xi}_i$) for a given data split – i.e., if we under-estimate (over-estimate) the true value Π_i with the isotonic regression $\widehat{\Xi}_i$ but we correctly identify with $\widehat{\Xi}_i$ that the original estimate $\widehat{\Theta}_i$ from the null hypothesis is under-estimated (over-estimated) – then the split LqRS with $q = 1$ (the split LRT) has the largest e-power. However, if $\widehat{\Xi}_i$ and Π_i are both larger (smaller) than $\widehat{\Theta}_i$, but if $\Pi_i < \widehat{\Xi}_i$ ($\Pi_i > \widehat{\Xi}_i$) for a given data split – i.e., if we over-estimate (under-estimate) the true value Π_i with the isotonic regression $\widehat{\Xi}_i$ but we still correctly identify with $\widehat{\Xi}_i$ that the original estimate $\widehat{\Theta}_i$ from the null hypothesis is under-estimated (over-estimated) – then the split LqRT with $q^{\text{opt}} \in (0,1)$ has a larger e-power than the split LqRT with $q = 1$ (the split LRT). From this we conclude that in small samples we should try to use the split LqRT for an optimal $q \in (0,1]$ to mitigate the inaccuracy of the isotonic regression on small samples. In large samples, we can safely choose $q = 1$ in the split LqRT (i.e., the split LRT) as the isotonic regression should sufficiently accurately estimate the true mean values based on the original mean estimates specified in the null hypothesis.

Theorem 5 can be better illustrated with the example of Figure 4.1. We study the e-power of the LqRT computed from a Poisson distributed response. Let the original estimate of the canonical parameter be equal to $\widehat{\theta} = \log(0.5)$, and its recalibrated version equal to $\widehat{\xi} = \log(0.7)$. We consider different true values of the canonical parameter π satisfying $\pi > \widehat{\theta} = \log(0.5)$. If the true value of the canonical parameter satisfies $\pi > \widehat{\xi} = \log(0.7)$, e.g., $\pi = \log(0.75)$, then the split LqRT with $q = 1$ (the split LRT) has the largest e-power; see Figure 4.1 (bottom-right). However, if the true value of the canonical parameter satisfies $\widehat{\theta} = \log(0.5) < \pi < \widehat{\xi} = \log(0.7)$, then the split

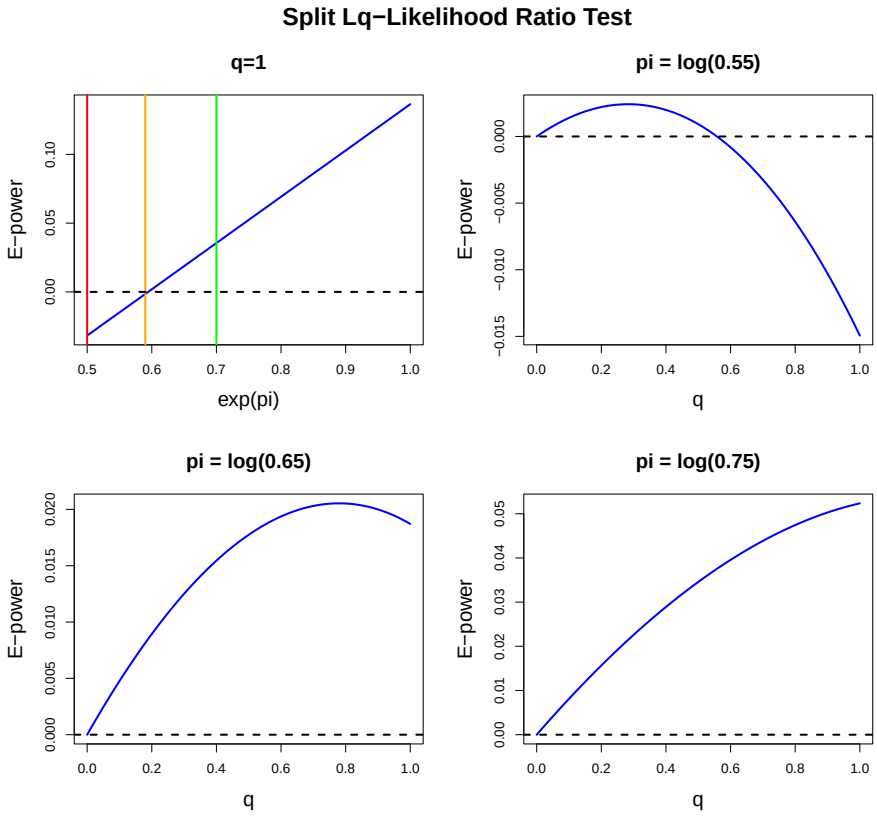

Figure 4.1: E-power of the split LqRT for Poisson responses as the function determined by (4.12) for $(\widehat{\theta}, \widehat{\xi}) = (\log(0.5), \log(0.7))$ being the original estimate and its recalibrated version.

LqRT with some $q^{\text{opt}} \in (0, 1)$ has the largest e-power, in particular, it has a larger e-power than the split LqRT with $q = 1$ (the split LRT). If the true value of the canonical parameter satisfies $\log(0.59) < \pi < \widehat{\xi} = \log(0.7)$, e.g., $\pi = \log(0.65)$, then the split LqRT with $q = 1$ has a positive e-power (but not optimal); see Figure 4.1 (bottom-left). If the true value of the canonical parameter satisfies $\widehat{\theta} = \log(0.5) < \pi < \log(0.59)$, e.g., $\pi = \log(0.55)$, then the split LqRT with $q = 1$ has negative e-power; see Figure 4.1 (top-right). Hence, if we over-estimate the true mean value with an isotonic regression fitted to the original mean estimate, given that the true mean value is under-estimated with the original mean estimate specified in the null hypothesis, then the split LqRT should be more powerful than the split LRT in rejecting the false null hypothesis (2.3) against the true alternative (2.4). If we severely over-estimate the true mean value with an isotonic regression fitted to the original mean estimate, given the true mean value is only slightly under-estimated with the original mean estimate specified in the null hypothesis, then the split LRT has a negative e-power, which may imply an asymptotic power-0 test, see Proposition 1.

From Theorems 4-5 we conclude that we should use the split LqRT with the optimal power parameter

$$q^* = \underset{q \in (0,1]}{\arg\max} \sum_{i \in \mathcal{D}_0} \mathbb{E}_{\mathbb{Q}_\Pi} \Big[\log \left(E^q_{\widehat{\Xi}_i} \left(Y_i, \widehat{\Theta}_i, V_i \right) \right) \Big| \mathcal{D}_1, (\widehat{\Theta}_i, V_i)_{i \in \mathcal{D}_0} \Big], \tag{4.14}$$

under the true alternative $\mathbb{Q}_\Pi$. Yet, Theorems 4-5 are of theoretical nature and an optimal choice

(4.14) cannot generally be found. The expected value $\mathbb{E}_{\mathbb{Q}_\Pi}[\cdot]$ is not known and we can only try to estimate it. However, this estimate, and the choice of q, would be accurate only in large samples, and in large samples the split LRT is good enough; the analyses are presented in the Supplementary.

For practical applications of the conclusions from Theorems 4 - 5, we now proceed as follows: in order to enhance the e-power and, consequently, the power of the split LRT, one could try to average or maximize over $q \in (0,1]$ in the split LqRS defined by (4.10). Select a finite subset $\mathcal{I} \subset (0,1]$ including $\{1\}$, and define the split mean power LqRS w.r.t. $\mathcal{I}$

$$E^{sLRT,\bar{q}} \;=\; \frac{1}{|\mathcal{I}|}\sum_{q\in\mathcal{I}} E^{sLRT,q}, \tag{4.15}$$

and *the split maximal power LqRS* w.r.t. $\mathcal{I}$

$$E^{sLRT,\tilde{q}} \;=\; \max_{q\in\mathcal{I}} E^{sLRT,q}. \tag{4.16}$$

We also introduce sub-sampled versions of the above test statistics by defining for i.i.d. repetitions

$$\overline{E}_B^{sLRT,\bar{q}} = \frac{1}{B}\sum_{b=1}^{B} E_b^{sLRT,\bar{q}}, \tag{4.17}$$

$$\overline{E}_B^{sLRT,\tilde{q}} = \frac{1}{B}\sum_{b=1}^{B} E_b^{sLRT,\tilde{q}}, \tag{4.18}$$

The variables (4.15) and (4.17) are e-variables and the standard theory of universal inference applies. However, (4.16) and (4.18) are no longer e-variables, which means that it is not immediately clear whether one can control type I errors for these test statistics.

Theorem 6. *Consider the null hypothesis $\mathcal{H}_0$ of calibration against the alternative $\mathcal{H}_1$, and reject $\mathcal{H}_0$ at a fixed significance level $\alpha \in (0,1)$ if a test statistics exceeds the value of $1/\alpha$. The split mean power LqRS* (4.15)*, the split maximal power LqRS* (4.16) *and the sub-sampled split mean power LqRS* (4.17) *control the type I error of rejecting $\mathcal{H}_0$ in favor of $\mathcal{H}_1$ at significance level $\alpha \in (0,1)$.*

Proof: The results for (4.15) and (4.17) are obvious since the test statistics are e-variables. We prove the result for (4.16). Select $\mathbb{P} \in \mathcal{P}$, where $\mathcal{P}$ is the set of probability measures representing the null hypothesis $\mathcal{H}_0$ defined by (2.3). The split LqRS is given by, see (4.10),

$$\begin{aligned} E^{sLRT,q} &= \prod_{i\in\mathcal{D}_0} \exp\left(\frac{q\,Y_i\left(\widehat{\Xi}_i - \widehat{\Theta}_i\right) - \left(\kappa\left(q\,\widehat{\Xi}_i + (1-q)\widehat{\Theta}_i\right) - \kappa\left(\widehat{\Theta}_i\right)\right)}{\phi/V_i}\right) \\ &= \exp\left(q\sum_{i\in\mathcal{D}_0}\frac{V_iY_i}{\phi}\left(\widehat{\Xi}_i - \widehat{\Theta}_i\right) - \sum_{i\in\mathcal{D}_0}\frac{V_i}{\phi}\left(\kappa\left(q\,\widehat{\Xi}_i + (1-q)\widehat{\Theta}_i\right) - \kappa\left(\widehat{\Theta}_i\right)\right)\right) = \exp\left(qZ - c_q\right), \end{aligned}$$

where we have defined

$$Z = \sum_{i\in\mathcal{D}_0}\frac{V_iY_i}{\phi}\left(\widehat{\Xi}_i - \widehat{\Theta}_i\right) \quad \text{and} \quad c_q = \sum_{i\in\mathcal{D}_0}\frac{V_i}{\phi}\left(\kappa\left(q\,\widehat{\Xi}_i + (1-q)\widehat{\Theta}_i\right) - \kappa\left(\widehat{\Theta}_i\right)\right).$$

Given $\{\mathcal{D}_1, \big(\widehat{\Theta}_i, V_i\big)_{i\in\mathcal{D}_0}\}$, first (i), the variables $\big(\widehat{\Xi}_i, \widehat{\Theta}_i, V_i\big)_{i\in\mathcal{D}_0}$ are fixed, and second (ii), the responses $(Y_i)_{i\in\mathcal{D}_0}$ are conditionally independent and EDF distributed with canonical parameters $\widehat{\Theta}_i$ under $\mathbb{P}$. The first item (i) implies that given $\{\mathcal{D}_1, \big(\widehat{\Theta}_i, V_i\big)_{i\in\mathcal{D}_0}\}$, the only random term in the split LRS $E^{sLRT,q}$ is the random variable Z. An immediate consequence is that the variables $\big(E^{sLRT,q}\big)_{q\in\mathcal{L}}$ are comonotonic, given $\{\mathcal{D}_1, \big(\widehat{\Theta}_i, V_i\big)_{i\in\mathcal{D}_0}\}$. This allows us to apply (Ramdas & Wang, 2025, Lemma 15.4) to show that

$$\begin{aligned}\mathbb{P}\Big(E^{sLRT,\tilde{q}} > 1/\alpha\Big) &= \mathbb{P}\Big(\max_{q\in\mathcal{I}} E^{sLRT,q} > 1/\alpha\Big) = \mathbb{E}_{\mathbb{P}}\Big[\mathbb{P}\Big(\max_{q\in\mathcal{I}} E^{sLRT,q} > 1/\alpha \Big| \mathcal{D}_1, \big(\widehat{\Theta}_i, V_i\big)_{i\in\mathcal{D}_0}\Big)\Big] \\ &= \mathbb{E}_{\mathbb{P}}\Big[\max_{q\in\mathcal{I}} \mathbb{P}\Big(E^{sLRT,q} > 1/\alpha \Big| \mathcal{D}_1, \big(\widehat{\Theta}_i, V_i\big)_{i\in\mathcal{D}_0}\Big)\Big] \leq \alpha,\end{aligned}$$

where the last step follows by Markov's inequality and by computing the conditional expected value of $E^{sLRT,q}$ equal to 1 for each $q \in \mathcal{I}$, see Theorem 3. This completes the proof. □

Unfortunately, a similar result does not obviously hold for the sub-sampled split maximal power LqRT (4.18) and it is an open question if $\mathbb{P}\big(\overline{E}_B^{sLRT,\tilde{q}} \geq 1/\alpha\big) \leq \alpha$ for $B > 1$ within the EDF.

5 Numerical analysis

We present a numerical analysis that studies the performance of classical and universal inference in testing mean-calibration. We start by describing the data generating mechanism.

5.1 Data set

We generate Poisson distributed responses Y_i with expected annual frequencies $\mu_i^* > 0$ being realistic choices for insurance claims counts modeling. Based on the real French motor insurance data set `freMTPL2freq` from the R package CASdatasets of Dutang et al. (2024), we select an expected annual frequency range of $[\mu_{\min}, \mu_{\max}] = [0.02, 0.25]$, the lower bound reflects excellent car drivers and the upper bound typically corresponds to inexperienced ones. This is a usual range of expected annual frequencies in car insurance pricing, it also reflects the usual class imbalance between insurance policyholders that suffer claims and those that do not suffer any claims.

For the test sample $\mathcal{D}$, we consider the sample sizes $n \in \{1000, 2000, 5000, 10000, 20000, 50000\}$. If the test sample $\mathcal{D}$ is 10% of the entire data (often 90% of the data is used for the learning sample $\mathcal{L}$ to fit the regression function $\widehat{f}$, see Section 2), then $n = 1000, 2000, 5000$ correspond to small insurance portfolios, $n = 10000, 20000$ to medium size portfolios and $n = 50000$ to a large portfolio, e.g., the `freMTPL2freq` data is of that latter size.

Then, we simulate the true expected annual frequencies $(\mu_i^*)_{i=1}^n$ of the test sample $\mathcal{D}$ for the selected sample sizes n. Since the expected frequency profile of an insurance motor portfolio is typically right-skewed, we simulate the portfolio from a beta distribution. Simulate i.i.d. random variables $R_i \sim \text{Beta}(a = 1.5, b = 5)$, and set for the true expected annual frequencies

$$\mu_i^* = \mu_{\min} + R_i \left(\mu_{\max} - \mu_{\min}\right), \qquad i = 1, \ldots, n. \tag{5.1}$$

Figure 5.1 shows the histogram of the selected true expected annual frequencies $(\mu_i^*)_{i=1}^n$ for the test sample $\mathcal{D}$ of sample size $n = 50000$. For simplicity, we set for all instances $V_i = 1$, i.e., we assume

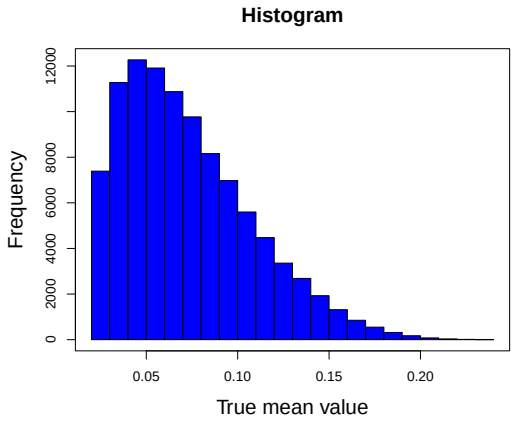

Figure 5.1: Histogram of the simulated true means in the test sample of sample size $n = 50000$.

a unit exposure. Finally, we simulate the Poisson claims by $Y_i \sim \text{Poi}(\mu_i^*)$ for all $i = 1, \ldots, n$. This provides us with the test sample $(Y_i, \log(\mu_i^*), V_i = 1)_{i=1}^n$; note that the cumulant function κ of the Poisson EDF distribution is the exponential one, which provides us with the (canonical) log-link mapping the means μ_i^* to the true canonical parameters $\theta_i^* = \log(\mu_i^*)$.

As described above, the general goal is to find these true mean parameters $(\mu_i^*)_{i=1}^n$ and canonical parameters $(\theta_i^* = \log(\mu_i^*))_{i=1}^n$, respectively. We usually proceed as follows. Based on a learning sample $\mathcal{L}$, we fit a regression model that provides us with estimated expected annual frequencies $\widehat{\mu}_i$; see Section 2. Our aim is to assess a potential miscalibration error in these estimates. For our synthetic example, we are going to consider four scenarios of increasing miscalibration, and we try to understand whether the above calibration tests are able to identify these scenarios. We select the slopes

$$\text{slope}_j = \begin{cases} 1.0, & j = 1 \text{ (no miscalibration)}, \\ 0.9, & j = 2 \text{ (small miscalibration)}, \\ 0.8, & j = 3 \text{ (medium miscalibration)}, \\ 0.7, & j = 4 \text{ (strong miscalibration)}. \end{cases}$$

Based on these slopes, we consider the four different scenarios $j = 1, 2, 3, 4$ of estimated means

$$\widehat{\mu}_i^j = \bar{\mu} + \text{slope}_j \, (\mu_i^* - \bar{\mu}), \qquad i = 1, \ldots, n. \tag{5.2}$$

where $\bar{\mu} = 0.075$ is the average mean value over the entire sample. The value of the slope in (5.2) represents the size of miscalibration of the mean estimates. These four scenarios of (mis)calibration are illustrated in Figure 5.2, and for the distribution of miscalibration one also needs to take into account the (right-skewed) sample distribution of Figure 5.1.

For each scenario $j = 1, 2, 3, 4$, our goal is to validate the following null hypothesis

$$\mathcal{H}_0 : \left\{ \mathbb{P} \in \mathcal{M}, \ \mathbb{E}_{\mathbb{P}}\big[Y_i | \widehat{\mu}_i^j\big] = \widehat{\mu}_i^j, \text{for all } i = 1, \ldots, n \right\}, \tag{5.3}$$

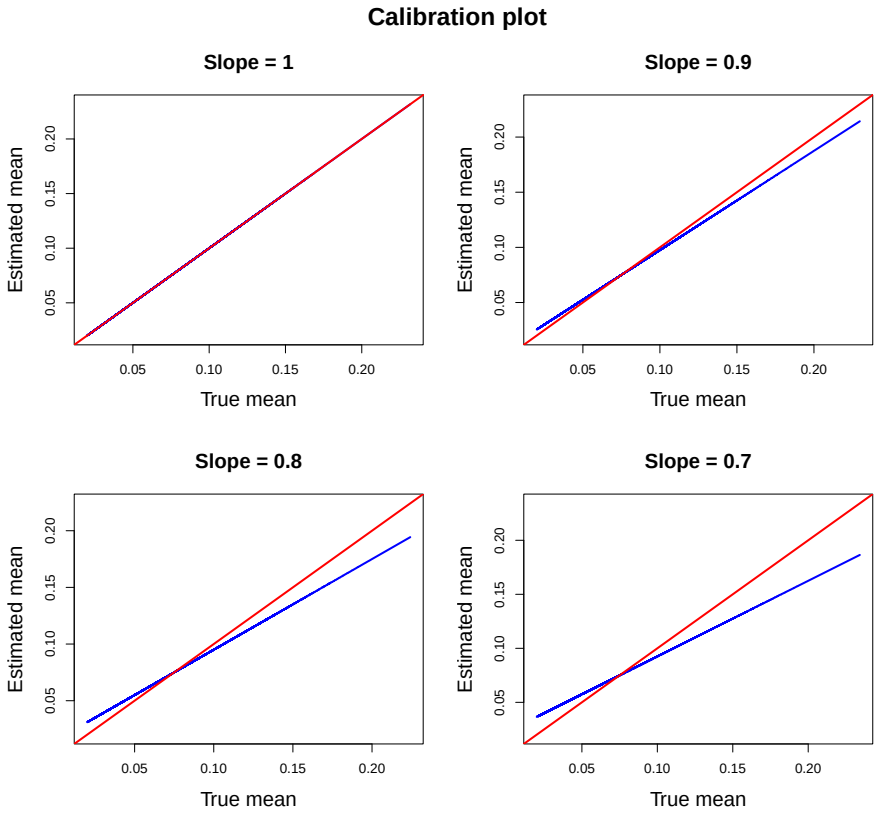

Figure 5.2: Estimated means $\widehat{\mu}_i^j$ vs. true means μ_i^* in the four considered calibration scenarios; the red line is the diagonal and the blue line indicates the misspecification relation.

against the alternative

$$\mathcal{H}_1 : \Big\{ \mathbb{Q} \in \mathcal{M},\ \mathbb{E}_{\mathbb{Q}}\big[Y_i | \widehat{\mu}_i^j\big] \neq \widehat{\mu}_i^j, \text{for some } i = 1, \ldots, n \Big\}. \tag{5.4}$$

For the calibration test for each $j = 1, 2, 3, 4$, we have a test sample $\mathcal{D} = (Y_i, \widehat{\mu}_i^j, V_i = 1)_{i=1}^n$ with Poisson responses generated by $Y_i \sim \text{Poi}(\mu_i^*)$, and the set of probability measures $\mathcal{M}$ contains distributions for conditionally i.i.d. responses such that $Y_i | \widehat{\mu}_i^j \sim \text{Poi}(\Psi_i)$ for some Ψ_i^j. The assumption of strict comonotonicity (2.2) is satisfied.

Since we would also like to understand the (irreducible) randomness implied by the (finite) test sample $\mathcal{D}$, the above steps of generating the true means $(\mu_i^*)_{i=1}^n$ and the test sample $\mathcal{D} = (Y_i, \widehat{\mu}_i^j, V_i = 1)_{i=1}^n$ are repeated 1000 times for the following analyses. Please also note that the vector of features $\mathbf{X}_i$ does not play any role in our experiments because we directly selected the heterogeneous means $\widehat{\mu}_i^j$, i.e., implicitly this includes the construction of the regression model.

5.2 CORP reliability diagram, Murphy's decomposition and classical LRT

For a suitable benchmark, we use the classical methods for testing the null hypothesis (5.3) of calibration against the alternative (5.4).

We begin with the CORP reliability diagram for mean estimates using the techniques from Section 3.2. The results for one random test sample $\mathcal{D}$, sample size $n = 50000$, are presented in

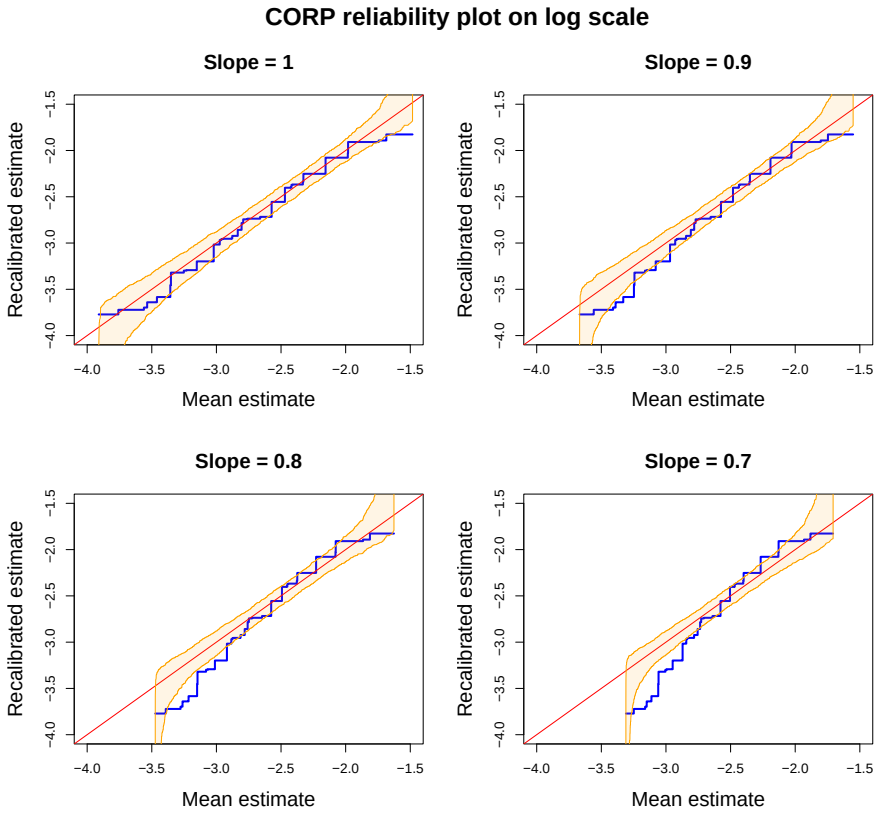

Figure 5.3: CORP reliability diagram (blue) for the mean estimates together with 95% point-wise consistency band (orange) for one random test sample of size $n = 50000$; the red line is diagonal.

Figure 5.3. Once the CORP reliability diagram is constructed for the mean estimates, we compute a 95% point-wise consistency band with 1000 simulations using the parametric bootstrap approach of Delong & Wüthrich (2025). Broadly speaking, under the null hypothesis of mean-calibration, we generate new (parametric bootstrap) observations $Y_i^\star \sim \mathrm{Poi}(\widehat{\mu}_i^j), i = 1, \ldots, n$, and re-fit a new isotonic regression to $(Y_i^\star, \widehat{\mu}_i^j, V_i = 1)_{i=1}^n$ to quantify the randomness in the recalibrated mean estimates under the assumption of mean-calibration. This is illustrated in Figure 5.3.

Next, we map these bootstrap results to test statistics by assessing the deviations of the CORP reliability diagram from the diagonal line under the Poisson deviance loss. In other words, we calculate the miscalibration statistics of Murphy's score decomposition (3.5). The values of the miscalibration statistics observed for one random test sample of size $n = 50000$ (green vertical lines) together with the bootstrap distributions of the miscalibration statistics under the null hypothesis of calibration of the mean estimates are presented in Figure 5.4; for technical details see Delong & Wüthrich (2025). From Figure 5.4 we conclude that the null hypothesis is (clearly) rejected at significance level $\alpha = 5\%$ in the medium and large miscalibration scenarios $j = 3, 4$ (the 95% quantile of the distribution of the test statistics under the null of mean-calibration corresponds to the red vertical line), but the test does not reveal the small miscalibration generated with $\text{slope}_2 = 0.9$, i.e., in that case the consistency band violations indicated in Figure 5.3 (top-right) are not sufficiently severe so that the calibration test could detect the miscalibration (reliably) on a sample size of $n = 50000$. In the case of the scenario $j = 1$ of mean-calibration, the test does

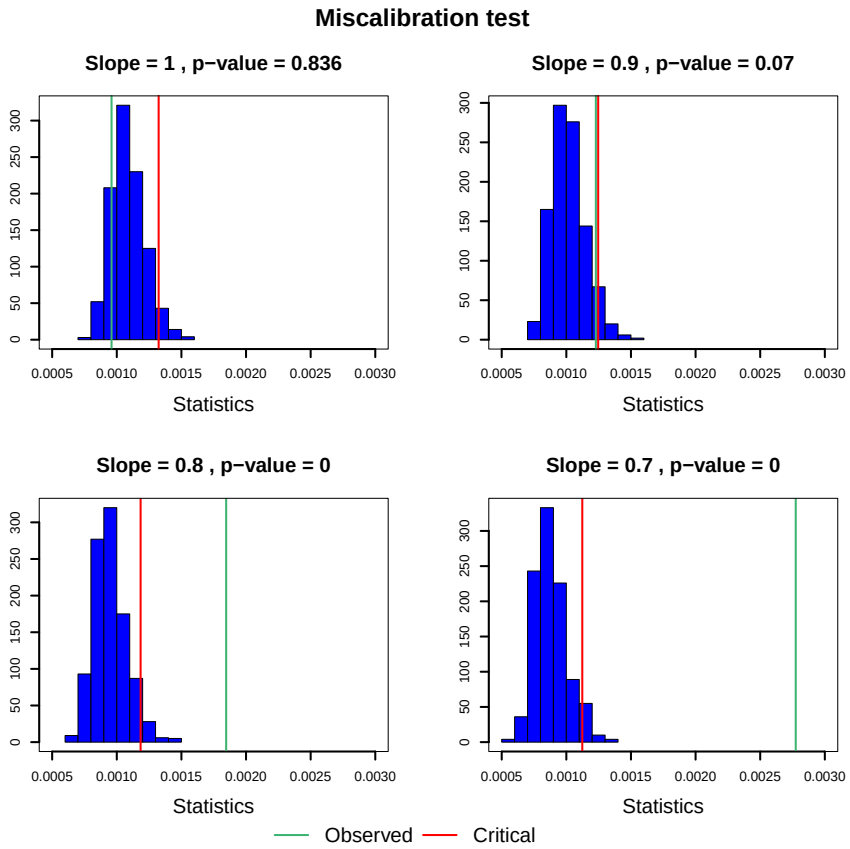

Figure 5.4: Distribution of the miscalibration statistics under the null hypothesis: green vertical line is the observed miscalibration statistics and the critical value at significance level $\alpha = 0.05$ corresponds to the red vertical line; sample size $n = 50000$.

(correctly) not reject the null hypothesis.

In Section 3.4, equation (3.6), we have shown the equivalence between Murphy's miscalibration statistics and the LRS. In the sequel, we work on the LRS scale, which allows us to directly connect to the universal inference proposals of Section 4. In particular, the results in Figure 5.4 then reflect the classical LRT for the given test sample $\mathcal{D} = (Y_i, \widehat{\mu}_i^j, V_i = 1)_{i=1}^n$. This classical LRT is based on one single test sample $\mathcal{D} = (Y_i, \widehat{\mu}_i^j, V_i = 1)_{i=1}^n$, as illustrated in Figures 5.3 and 5.4. In order to understand the power of the classical LRT, we repeat this experiment 1000 times by simulating 1000 i.i.d. test samples $\mathcal{D}$ to evaluate the power of this classical LRT. Figure 5.5 shows the results for significance level $\alpha = 5\%$, for the miscalibration scenarios $j = 2, 3, 4$, and the different sample sizes between 1000 and 50000 – by construction the type I error of this classical LRT is $\alpha = 5\%$. The large miscalibration scenario $j = 4$ can reliably be detected for sample sizes bigger than 10000, but the small miscalibration scenario $j = 2$ is difficult to detect, even a comparably large sample size of 50000 only provides a power of roughly 0.6; see Figure 5.5.

We conclude this section by confirming that classical inference provides us with suitable tools to detect medium and large miscalibrations in medium and large samples. Yet, these tools are based on simulating the LRS under the null hypothesis, which may hinder their widespread use practice.

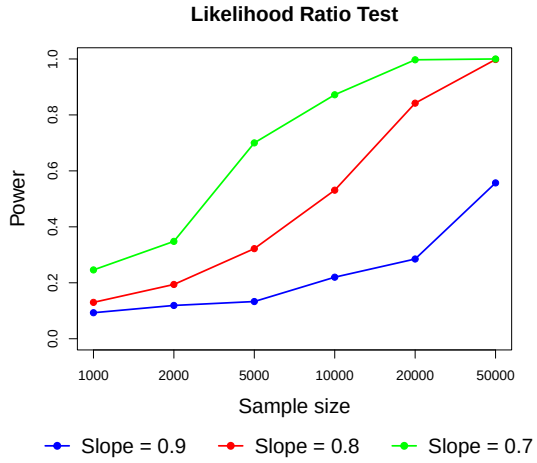

Figure 5.5: Power of the classical LRT at significance level $\alpha = 0.05$.

5.3 Universal inference: sub-sampled split LRT

This section applies the techniques from universal inference to test the null hypothesis (5.3) of mean-calibration against the alternative (5.4). We begin with the sub-sampled split LRS defined in (4.6), see Section 4.2. We choose split ratio $s = 0.5$ and number of sub-samples $B = 1000$ for the sub-sampled split LRS $\overline{E}_B^{sLRT}$, see (4.6). The split ratio $s = 0.5$ is advised by Dunn et al. (2022); other values are investigated at the end of this section.

We start by investigating the type I error of the sub-sampled split LRT. Since we use the universal critical value of $1/\alpha = 20$ at the significance level $\alpha = 0.05$, we expect the type I error of the test to be much lower than 0.05. Figure 5.6 confirms this expectation; the property of lower true type I errors has also been identified in other papers on universal inference. We choose the mean estimates from the scenario $j = 1$ for the null hypothesis and consider the significance levels $\alpha = 0.01, 0.05, 0.1$, providing (universal) critical values $1/\alpha$ for the test. The probabilities are estimated based on 1000 i.i.d. simulations of the test sample $\mathcal{D}$.

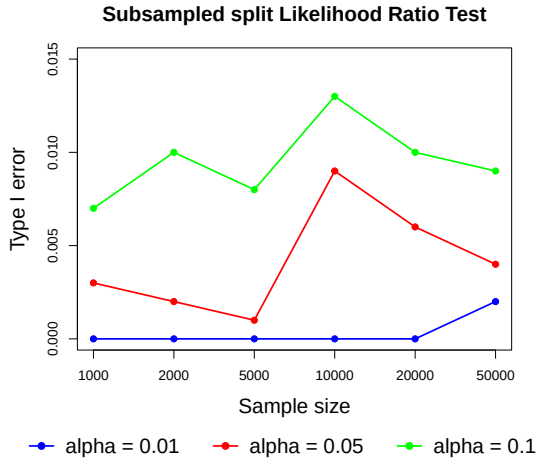

Figure 5.6: Type I error of the sub-sampled split LRT with the (universal) critical value equal to $1/\alpha$ at the pre-assumed significance levels $\alpha = 0.01, 0.05, 0.1$.

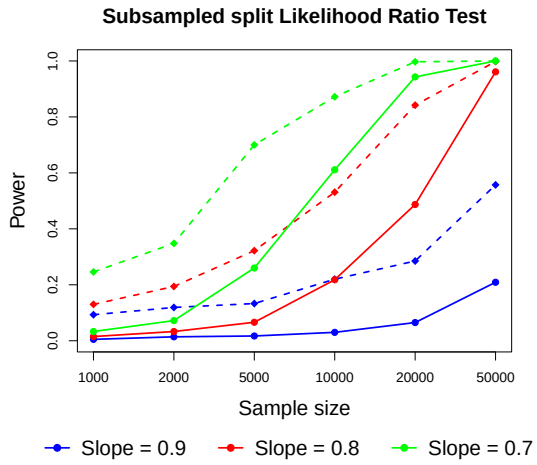

Figure 5.7: Power of the sub-sampled split LRT with critical value equal to $1/\alpha = 20$; the dashed lines present the power of the classical LRT at significance level $\alpha = 0.05$ (taken from Figure 5.5).

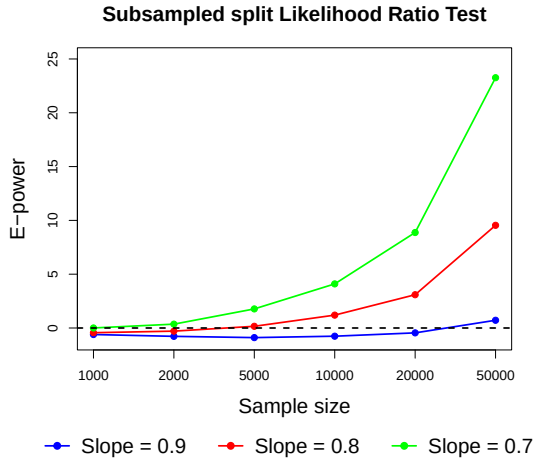

Figure 5.8: E-power of the sub-sampled split LRT with (universal) critical value $1/\alpha = 20$.

Fortunately, as our next results demonstrate, even though the type I error is below its targeted value, the power of the sub-sampled split LRT is reasonable good and large enough for many practical applications. Figure 5.7 shows the powers of the sub-sampled split LRTs at the pre-assumed significance level $\alpha = 0.05$ with the (universal) critical value $1/\alpha = 20$ for rejecting the null hypothesis. We choose the mean estimates from the scenario $j = 2, 3, 4$. The solid lines show the results of the sub-sampled split LRTs, and the dotted lines their classical LRT counterparts taken from Figure 5.5. The main conclusions of Figure 5.7 are that the medium and large miscalibrations can reliably be detected by the sub-sampled split LRT in sample sizes bigger than 20000, and the power of the test is comparably low for the miscalibration scenarios $j = 3, 4$ for smaller sample sizes and for the weak miscalibration $j = 2$ for all sample sizes. We also observe gaps in powers between the classical LRT and its sub-sampled split counterpart for weak miscalibration and for small sample sizes. We deduce that for small sample sizes one should rely on the classical LRT (being evaluated with Monte Carlo and bootstrap), and for medium/large sample sizes one can rely on

the sub-sampled split LRT to detect bigger miscalibration errors (which does not use simulations). Weak miscalibration of mean estimates is difficult to detect in general, i.e., with both the classical and the sub-sampled split LRT even in large sample sizes.

The e-power of the sub-sampled split LRT is investigated in Figure 5.8. The e-power is negative for small sample sizes and small/medium miscalibration sizes and gets positive for large sample sizes for all miscalibration sizes. By Definition 3, Proposition 1 and the discussions, we conclude that a positive e-power is a desired property of the sub-sampled split LRT in rejecting the false null hypothesis against the true alternative. Consequently, the sub-sampled split LRT gains more e-power, the larger the sample size and the larger the miscalibration size. This conclusion, of course, agrees with the conclusions from Figure 5.7. We summarize that the sub-sampled split LRT for different sample and miscalibration sizes behaves as expected.

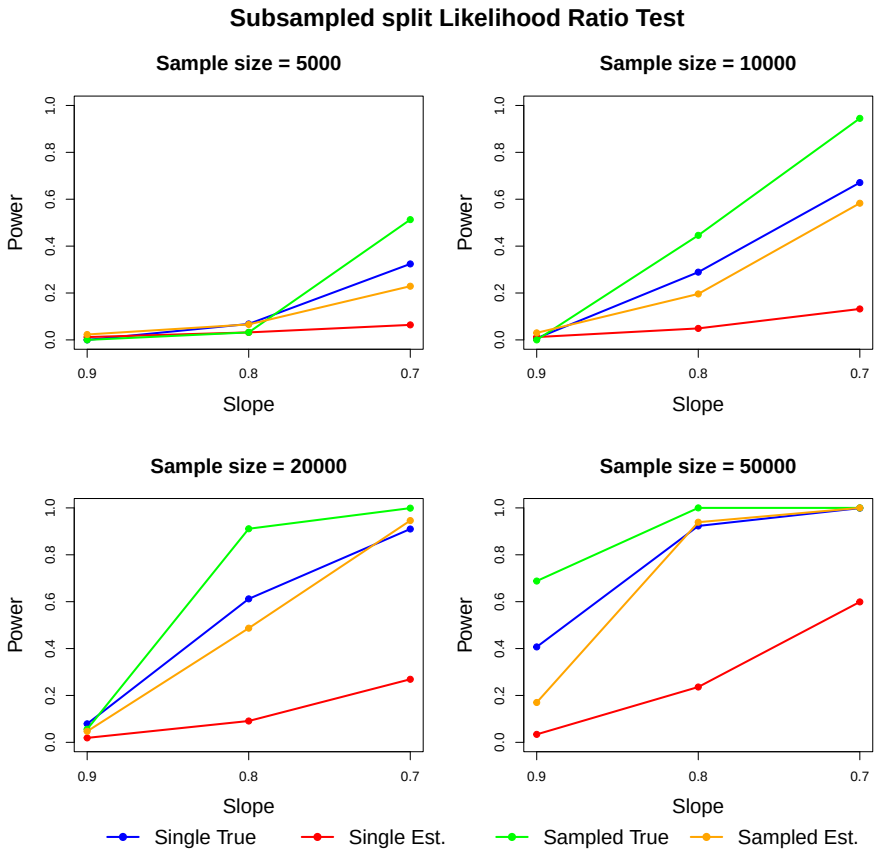

Figure 5.9: Power of the (sub-sampled) split LRTs for $B \in \{1, 1000\}$ with (universal) critical value $1/\alpha = 20$; we use the estimated means with isotonic regression (red/orange) and the true means (blue/green) under the alternative.

Figure 5.9 investigates four different test statistics. In red color it considers the split LRS E^{sLRT}, defined in (4.5), this corresponds to $B = 1$, and in orange color the sub-sampled split LRS $\overline{E}_B^{sLRT}$, defined in (4.6), for $B = 1000$. The difference between these two lines in Figure 5.9 shows the gain in statistical power from sub-sampling $B > 1$. Second, Figure 5.9 analyzes the inaccuracy implied by the isotonic regression (4.4) used to estimate the mean values under the alternative hypothesis. Working with synthetic data, we can replace these isotonic regression estimates by the

true means to evaluate the loss of power by this estimation step. The corresponding graphs are given in blue color ($B = 1$) and in green color ($B = 1000$). From these results, we conclude that the power of the split LRT is crucially depending on accuracy of the isotonic regression step, which can be critical in small samples. This conclusion is again in support of using the sub-sampled split LRT for medium/large sample sizes with medium/large miscalibration sizes, and the classical LRT (with Monte Carlo and bootstrap) otherwise (as splitting a small sample or a medium/large sample with a small miscalibration has a negative impact on the power of the test).

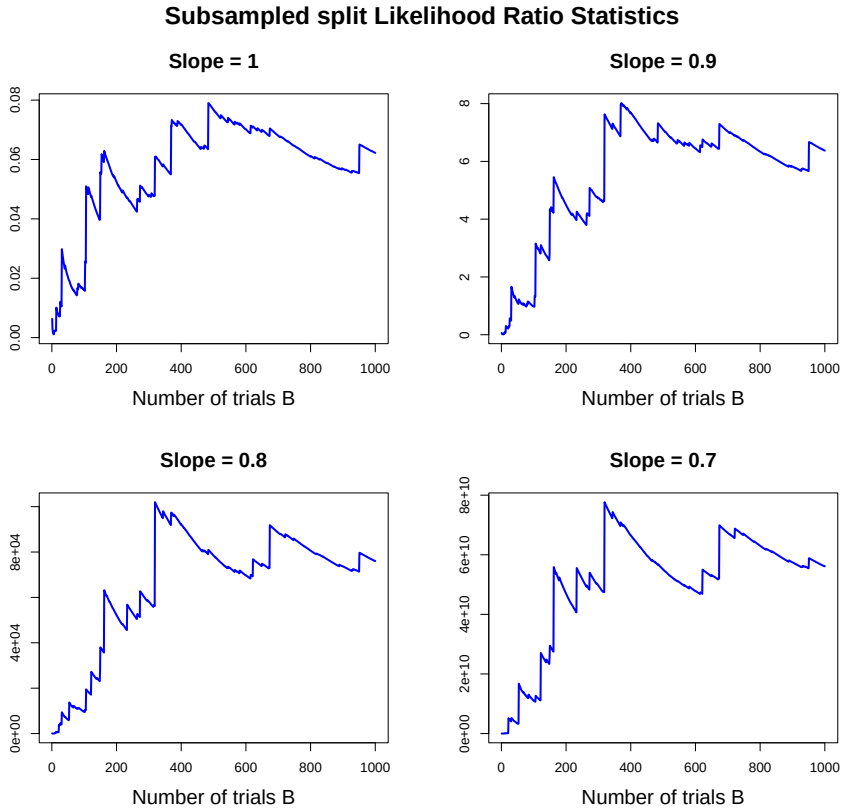

Figure 5.10: Running value of the sub-sampled split LRS in repeated sampling, $B \geq 1$, for one random test sample $\mathcal{D}$ of sample size $n = 50000$.

The above results of the sub-sampled split LRTs were based on $B = 1000$ sub-samples. In Figure 5.10 we investigate the running value of the sub-sampled split LRS for increasing numbers $B \geq 1$ of sub-samples in (4.6); this is for a single test sample $\mathcal{D}$ of sample size $n = 50000$. As expected, the larger the miscalibration, the bigger the value of the sub-sampled split LRS. This figure shows that it takes roughly 500 iterations for the running maximum to stabilize.

Figure 5.11 presents the histograms of the number of sampling trials B before the sub-sampled split LRS $\overline{E}_B^{sLRT}$ exceeds for its first time the universal critical value $1/\alpha = 20$. The plots show the histograms over the 1000 i.i.d. test samples $\mathcal{D}$, and the upper right corner indicates the power of the calibration test. E.g., the bottom-middle plot shows the case of sample size $n = 50000$ and miscalibration error $\text{slope}_3 = 0.8$. This test has a power of 0.978, and from the histogram we see that in almost half of the 1000 i.i.d. test samples $\mathcal{D}$ the critical value $1/\alpha = 20$ is exceeded with $B = 2$. We point out that here the power is reported for the sub-sampled split LRT with

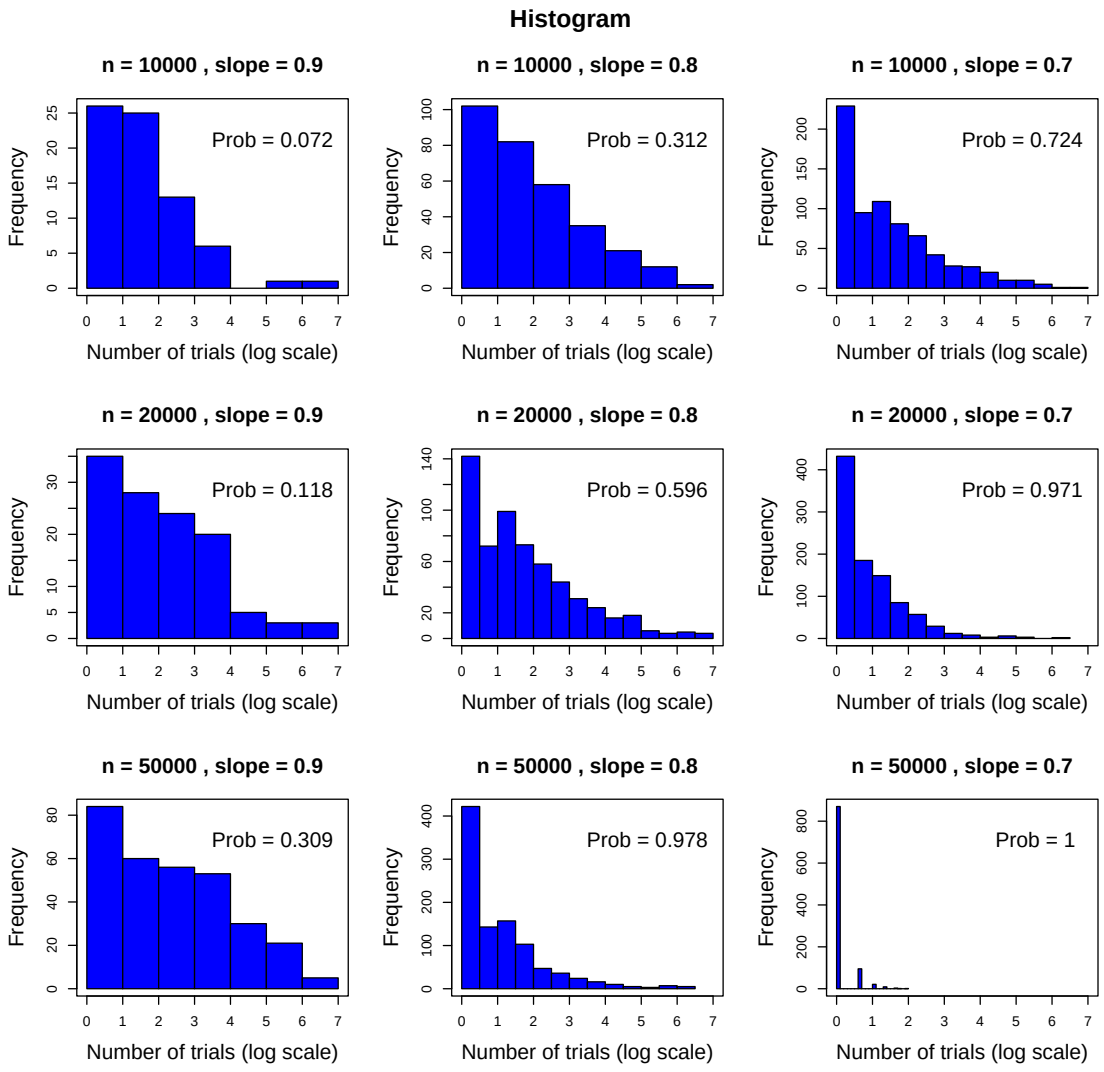

Figure 5.11: Histogram of the number of sampling trials B before the sub-sampled split LRS $\overline{E}_B^{sLRT}$ exceeds the universal critical value $1/\alpha = 20$ for the first time; "Prob" gives power of the calibration test (empirically determined with data-dependent B from at most 1000 i.i.d. test samples $\mathcal{D}$).

data-dependent B, where the sampling is terminated at the first time the running value of the sub-sampled split LRS exceeds 20, see (Ramdas & Wang, 2025, Section 5.3). From these plots we conclude that often roughly $B = 200$ to 300 sub-samples are sufficient in all our cases, and even much less for medium/large samples with medium/large miscalibrations sizes.

From Figure 5.9 we concluded that the loss in power of the split LRT compared to the classical LRT comes from the inaccuracy of the isotonic regression step. For the split LRT, the test sample $\mathcal{D}$ is partitioned at random into the training sample $\mathcal{D}_1$ used for estimating the isotonic regression step and the validation sample $\mathcal{D}_0$ used for performing the calibration test. Following Dunn et al. (2022), we selected a split ratio $s = 0.5$. Figure 5.12 benchmarks this choice against $s = 0.2, 0.3, 0.7, 0.8$. We run the simulations with $B = 200$. The results confirm the proposal of Dunn et al. (2022).

Finally, Figure 5.13 investigates the relation between the sub-sampled split LRS $\overline{E}_B^{sLRT}$ and the classical LRS E^{LRT}. The graph shows a scatter plot over 1000 i.i.d. test samples $\mathcal{D}$ of sample size

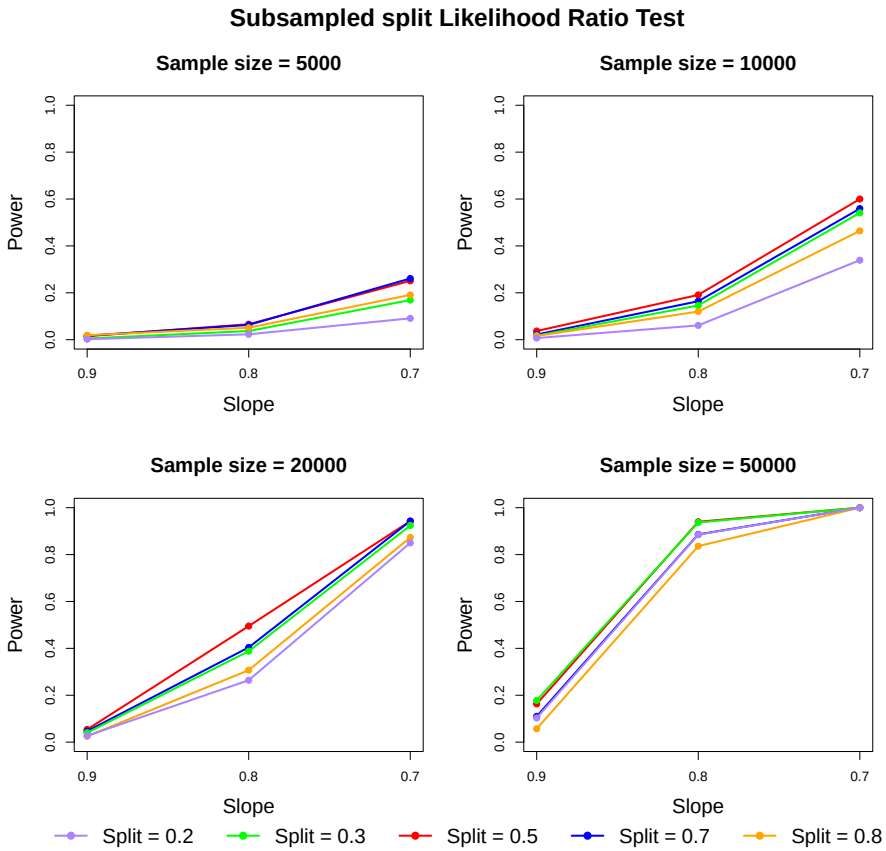

Figure 5.12: Power of the sub-sampled split LRT for different split ratios $s \in \{0.2, 0.3, 0.5, 0.7, 0.8\}$.

$n = 50000$ for the two LRS. Dunn et al. (2022) prove under a Gaussian assumption that on the log-scale the relation between the two statistics should be asymptotically linear with slope equal to 3/5. Interestingly, Figure 5.13 verifies this also in our setting with a slope close to 3/5 (equal to approximately 0.56). The potential proof of this conjecture is beyond the scope of this paper.

5.4 Universal inference: sub-sampled split LqRT

There are (at least) two known options to increase the power of the sub-sampled split LRT studied in the previous section. A first option is to stop the sub-sampling process once the sub-sampled split LRS exceeds the universal critical value $1/\alpha$ for the first time, see (Ramdas & Wang, 2025, Section 5.3). A second option is to introduce a randomization into the decision of rejecting the null hypothesis against the alternative, see (Ramdas & Wang, 2025, Section 2.4). These modifications are feasible in our setting since the sub-sampled split LRS (4.6) is an e-variable. We do not investigate the application of these two modifications because our focus is on understanding the performance of our novel test statistics, the split mean and maximal power LqRS (4.15) and (4.16), respectively, as well as their sub-sampled versions (4.17) and (4.18).

The solid lines in Figure 5.14 show the single sub-sampled versions of the statistical tests with $B = 1$: the benchmark is the split LRS E^{LRT}, and the split mean and maximal power LqRS show the results of the test statistics (4.15) and (4.16). We conclude that in any case these Lq-likelihood

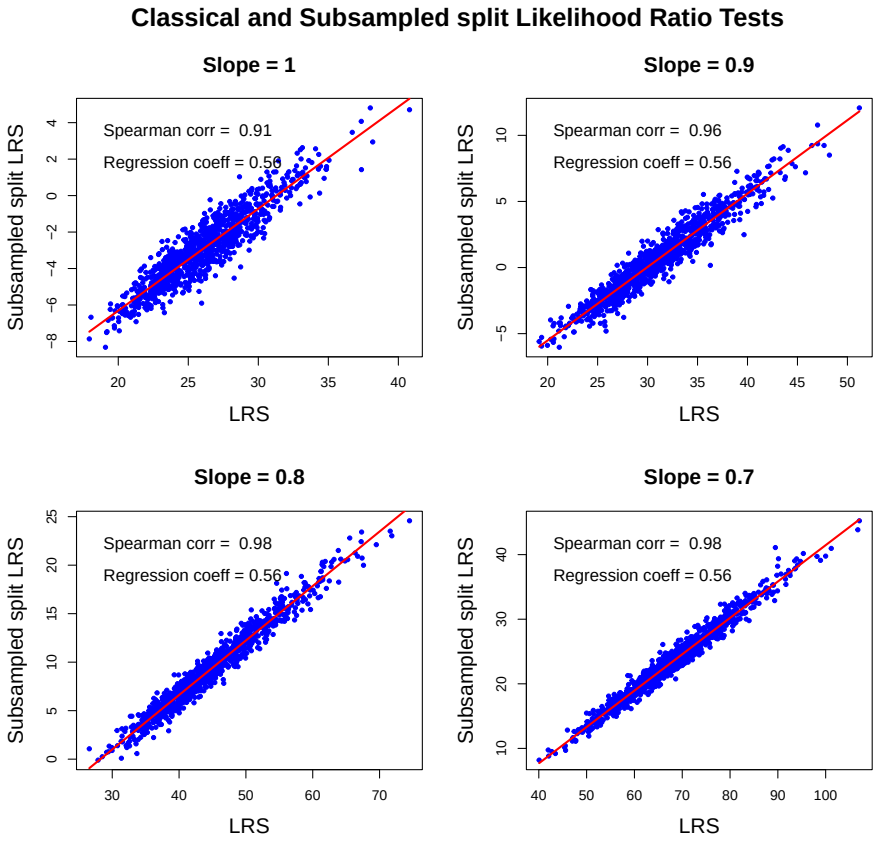

Figure 5.13: Relation between the sub-sampled split LRS $\overline{E}_B^{sLRT}$ and the classical LRS E^{LRT} on the log-scale for i.i.d. test samples $\mathcal{D}$ of sample size $n = 50000$.

versions lead to an increased power compared to the benchmark; $\mathcal{I}$ was selected by 10 uniformly distributed values q across $(0, 1)$. The results suggest that on medium and large sample sizes it is beneficial to work with the split mean power LqRT and the split maximal power LqRT to identify medium and large miscalibration, as in these cases the improvement in the power is the largest.

In Figure 5.14, we also look at the sub-sampled versions (dashed lines) of the statistical tests with $B = 1000$ sub-samples: the benchmark is the sub-sampled split LRS $\overline{E}_B^{LRT}$, and the sub-sampled split mean and maximal power LqRS show the results of the test statistics (4.17) and (4.18). We observe that the sub-sampled split maximal power LqRT has the largest power for all sample sizes and all miscalibration sizes. Unfortunately, this test statistics cannot be supported because it lacks the formal type I error bound, although our numerical results verify that the type I error bound our example holds, and the true type I error for the test with the critical value of 20 corresponding to the pre-assumed significance level $\alpha = 0.05$ fluctuates around 0.5% for all sample sizes considered (similar as for the sub-sampled split LRT, see Figure 5.6). Therefore, we shall only focus on the sub-sampled split mean power LqRT. The true type I error for this test with the critical value of 20 fluctuates around 0.2% (lower than for the sub-sampled split LRT). It is interesting to note that if we sub-sample $B = 1000$ training and validation sets, the sub-sampled split LRT has now a slightly larger power than the sub-sampled split mean power LqRT for all considered sample sizes and miscalibration errors. Hence, the ranking of these two tests based on their power has

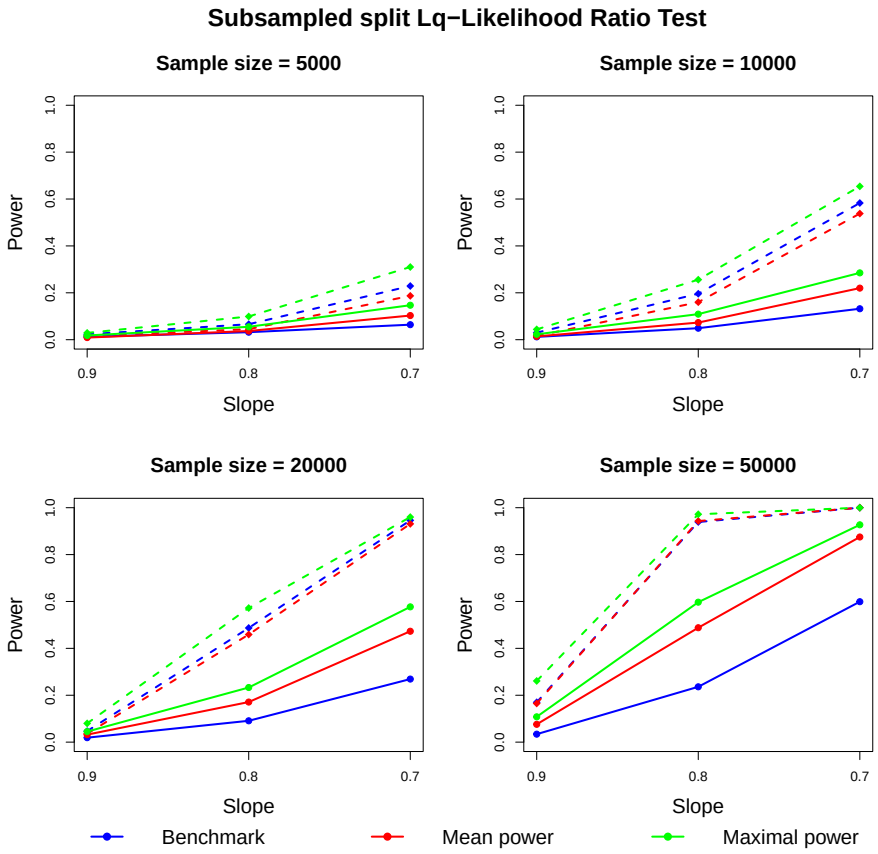

Figure 5.14: Power of the (sub-sampled) split LRT (Benchmark), the (sub-sampled) split mean power LqRT and the (sub-sampled) split maximal power LqRT with (universal) critical value $1/\alpha = 20$: the solid lines show the single sub-sample versions $B = 1$, and the dashed lines the sub-sampled versions $B = 1000$.

changed if we switch from $B = 1$ to $B = 1000$, see also discussion in the Supplementary. Table 5.4 shows that if we sample a lower number of partitions, e.g., if we set $B = 20$, the sub-sampled split mean power LqRT receives a slight advantage over the sub-sampled split LRT in terms of power. Moreover, their powers, despite using a much lower number of sub-samples, are at a similar level as the power of the sub-sampled split LRT based on $B = 1000$ sub-samples. From Figure 5.11 we know that $B = 1000$ is not needed to have a sound statistical power of a sub-sampled split test, and in practical applications we have recommended to use a lower B with a small effect on the power of the test. We conclude that the sub-sampled split mean power LqRT may be preferred over the sub-sampled split LRT, especially if we choose low B.

6 Conclusions and open problems

In this paper, we investigated the framework of universal inference to test mean-calibration within the EDF. We developed a sub-sampled split Likelihood Ratio Test (LRT) and its modifications based on the Lq-Likelihood Ratio Test (LqRT). In a case study, we confirmed that the sub-sampled split LRT and its modifications are appealing alternatives to the classical LRT. We identified

Sample size	Statistics	Miscalibration=0.9	Miscalibration=0.8	Miscalibration=0.7
10000	Benchmark $B = 1000$	0.03	0.22	0.61
10000	Mean power $B = 1000$	0.02	0.17	0.55
10000	Benchmark $B = 20$	0.02	0.17	0.54
10000	Mean power $B = 20$	0.01	0.14	0.53
20000	Benchmark $B = 1000$	0.06	0.49	0.94
20000	Mean power $B = 1000$	0.05	0.44	0.93
20000	Benchmark $B = 20$	0.05	0.40	0.90
20000	Mean power $B = 20$	0.05	0.41	0.92
50000	Benchmark $B = 1000$	0.21	0.96	1.00
50000	Mean power $B = 1000$	0.20	0.96	1.00
50000	Benchmark $B = 20$	0.14	0.89	1.00
50000	Mean power $B = 20$	0.16	0.94	1.00

Table 5.1: Power of the sub-sampled split LRTs (Benchmark) based on 1000 and 20 sub-samples and the sub-sampled split mean power LqRTs based on 1000 and 20 sub-samples with (universal) critical value $1/\alpha = 20$.

the beneficial impact of sub-sampling of training and validation sets on the split LRT, and we demonstrated that despite of a low type I error, the sub-sampled split LRT has a high and reasonable power on medium and large samples in the case of medium and large size miscalibration errors. In addition, our new so-called sub-sampled split mean power LqRT may perform better in some scenarios than the sub-sampled split LRT in terms of the power of rejecting the false null hypothesis against the true alternative for mean estimates.

Our case study considered Poisson responses, but the authors also performed numerical examples for gamma responses with similar conclusions. As expected, the power of the tests deteriorates as the dispersion coefficient of the gamma responses increases, which makes the calibration validation challenging in highly skewed and dispersed data settings (lower signal-to-noise ratios). Based on our findings, we believe that universal inference may offer a practical and powerful tool for testing mean-calibration.

Related to the previous paragraph, a key ingredient of our statistical procedure is the estimation of the dispersion parameter. In our case study of Poisson responses the dispersion parameter is equal to one. In general two-parameter families, the impact and assessment of the dispersion estimation on universal inference to test calibration is an open question. Another open question is the behavior of universal inference to validate mean-calibration in the case the ranking of the risks is not fully correct. Finally, the type I error for our sub-sampled split maximal power LqRT is worth investigating as this test statistics achieves the highest power in our case study.

References

Banerjee, M. (2007). Likelihood based inference for monotone response models *The Annals of Statistics* **3**, 931-956.

Delong, Ł., Wüthrich, M.V. (2025). Isotonic regression for variance estimation and its role in mean estimation and model validation. *North American Actuarial Journal* **29**, 563-591.

Delong, Ł., Gatti, S., Wüthrich, M.V. (2025). Calibration bands for mean estimates within the exponential dispersion family. *arXiv*:2503.18896.

Denuit, M., Huyghe, J., Truffin, J., Verdebout, T. (2024). Testing for auto-calibration with Lorenz and concentration curves. *Insurance: Mathematics and Economics* **117**, 130-139.

Dimitriadis, T., Gneiting, T., Jordan, A.I. (2021). Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences of the United States of America* **118**, 1-10.

Dimitriadis, T., Dümbgen, L., Henzi, A., Puke, M., Ziegel, J. (2023). Honest calibration assessment for binary outcome predictions. *Biometrika* **110**, 663-680.

Dunn, R., Ramdas, A., Balakrishnan, S., Wasserman, L. (2022). Gaussian universal likelihood ratio testing. *Biometrika* **110**, 319-337.

Dutang, C., Charpentier, A., Gallic, E., Siharath, J. (2024). Package CASdatasets. *https://dutangc.perso.math.cnrs.fr/RRepository/*.

Ferrari, D., Yang, Y. (2010). Maximum Lq-likelihood estimation. *The Annals of Statistics* **38**, 753-783.

Gneiting, T., Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**, 359-378.

Gneiting, T., Resin, J. (2023). Regression diagnostics meets forecast evaluation: Conditional calibration, reliability diagrams and coefficient of determination. *Electronic Journal of Statistics* **17**, 3226-3286.

Henzi, A., Puke, M., Dimitriadis, T., Ziegel, J. (2025). A safe Hosmer-Lemeshow test. *The New England Journal of Statistics in Data Science* **2**, 175-189.

Leeuw, de J., Hornik, K., Mair, P. (2009). Isotone optimization in R: pool-adjacent-violators algorithm (PAVA) and active set methods. *Journal of Statistical Software* **32/5**, 1-24.

Lindholm, M., Wüthrich, M.V. (2025). The balance property in insurance pricing. *Scandinavian Actuarial Journal*, in press.

Murphy, A.H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology* **12**, 595-600.

Pohle, M.-O. (2020). The Murphy decomposition and the calibration-resolution principle: A new perspective on forecast evaluation. *arXiv*:2005.01835.

Qin, Y., Priebe, C.E. (2017). Robust hypothesis testing via Lq-likelihood. *Statistica Sinica* **27**, 1793-1813.

Ramdas, A., Wang, R. (2025). Hypothesis testing with e-values. *arXiv*:2410.23614.

Shi, H., Drton, M. (2025). On universal inference in Gaussian mixture models. *arXiv*:2407.19361.

Strieder, D., Drton, M. (2022). On the choice of the splitting ratio for the split likelihood ratio test. *Electronic Journal of Statistics* **16**, 6631-6650.

Wasserman, L., Ramdas, A., Balakrishnan, S. (2020). Universal inference. *The Proceedings of the National Academy of Sciences of the United States of America* **117**, 16880-16890.

Wright, F.T. (1981). The asymptotic behavior of monotone regression estimates. *The Annals of Statistics* **9**, 443-448.

Wüthrich, M.V. (2025). Auto-calibration tests for discrete finite regression functions. *European Actuarial Journal* **15**, 335-341.

Wüthrich, M.V., Merz, M. (2023). *Statistical Foundations of Actuarial Learning and its Applications*. Springer.

Wüthrich, M.V., Ziegel, J. (2024). Isotonic recalibration under a low signal-to-noise ratio. *Scandinavian Actuarial Journal* **2024**, 279-299.

Supplementary: Additional analyses of the power of the split LqRT

The optimal power parameter q for the split LqRT is given by (4.14). In our numerical example we can calculate $\mathbb{E}_{\mathbb{Q}_\Pi}[\cdot]$ by replacing Π_i with $\log(\mu_i^*)$ in formula (4.13). The optimal q^* is then found by optimizing (4.14) over a finite set $\mathcal{A}$. In practice, we have to estimate this parameter from data.

SPLIT LQRT WITH ESTIMATED POWER PARAMETER

- We modify the algorithm for the split LRT: In addition to the split ratio $s \in (0,1)$, choose a fixed split ratio $r \in (0,1)$.
- Construct a sub-sample $\mathcal{D}_{0,0}$ by randomly selecting $[[ns]r]$ triples from $(Y_i, \widehat{\Theta}_i, V_i)_{i\in\mathcal{D}_0}$ without replacement, and set $\mathcal{D}_{0,1} = \mathcal{D}_0 \setminus \mathcal{D}_{0,0}$.
- Find $q \in \mathcal{A}$ with the rule

$$\widehat{q}^* = \underset{q\in(0,1]}{\arg\max} \sum_{i\in\mathcal{D}_{0,1}} \log\left(E^q_{\widehat{\Xi}_i}\left(Y_i, \widehat{\Theta}_i, V_i\right)\right).$$

- Define the split LqRS for testing $\mathcal{H}_0$ against $\mathcal{H}_1$

$$E^{sLRT,\widehat{q}^*} = \prod_{i\in\mathcal{D}_{0,0}} E^{\widehat{q}^*}_{\widehat{\Xi}_i}\left(Y_i, \widehat{\Theta}_i, V_i\right). \tag{.1}$$

One can show that both E^{sLRT,q^*} and $E^{sLRT,\widehat{q}^*}$ are exact e-variables, this is similar to Theorem 2.

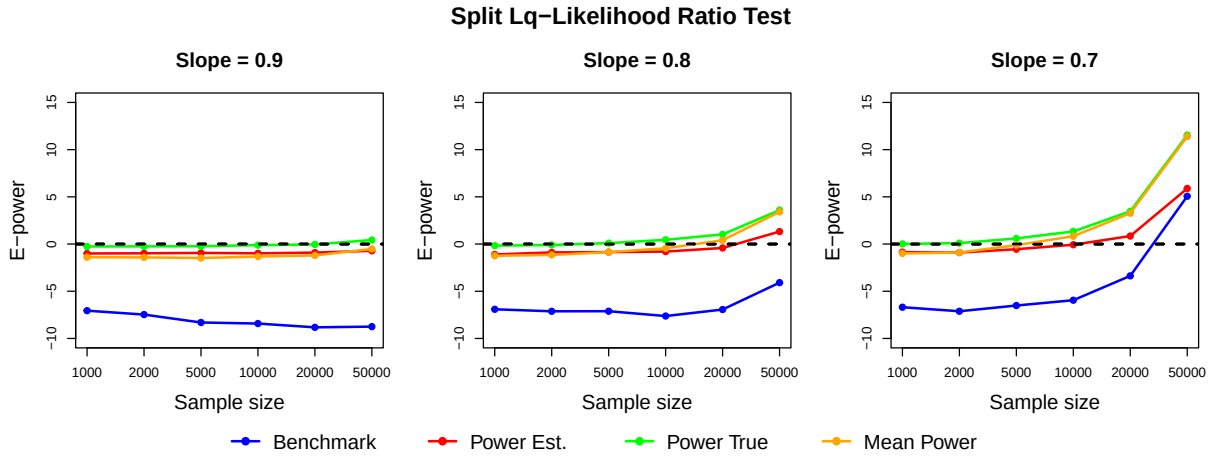

Figure .1: E-power of the split LRT (Benchmark), the split LqRT with the estimated optimal power parameter (Power Est.), the split LqRT with the true optimal power parameter (Power True), and the split mean power LqRT with (universal) critical value $1/\alpha = 20$ for $B = 1$.

Figure .1 illustrates the conclusions from Theorems 4-5 and investigates the e-power of the various split test statistics for $B = 1$: the benchmark is the split LRS, the split LqRS with the estimated optimal power parameter and the true optimal power parameter are given by the split LqRS (4.10) with, respectively, q^* and $\widehat{q}^*$, and the split mean power LqRS show the results of the

test statistics 4.15. The split LqRT with q^* achieves the largest e-power in all cases, followed by the split mean power LqRT and the split LqRT with $\widehat{q}^*$, whereas the split LRT has significantly lower e-power in all cases. In Figure .2 we look at the statistical power of the test statistics. For a single data split, in the case of medium/large miscalibration sizes in medium/large sample sizes, we see clear advantages of applying the split LqRT with q^*. A similarly good performance is also observed for the split mean power LqRT. In large samples it is also beneficial to apply the split LqRT with $\widehat{q}^*$ since the estimated optimal power parameter becomes close to the true optimal power parameter q^*. The split LRT falls behind the other test statistics in all cases. These conclusions have also been confirmed in larger sample sizes of 70000 and 100000; the gap between the red and the blue curve increases and the red curve approaches the green and orange curves as the sample size increases for all miscalibration sizes.

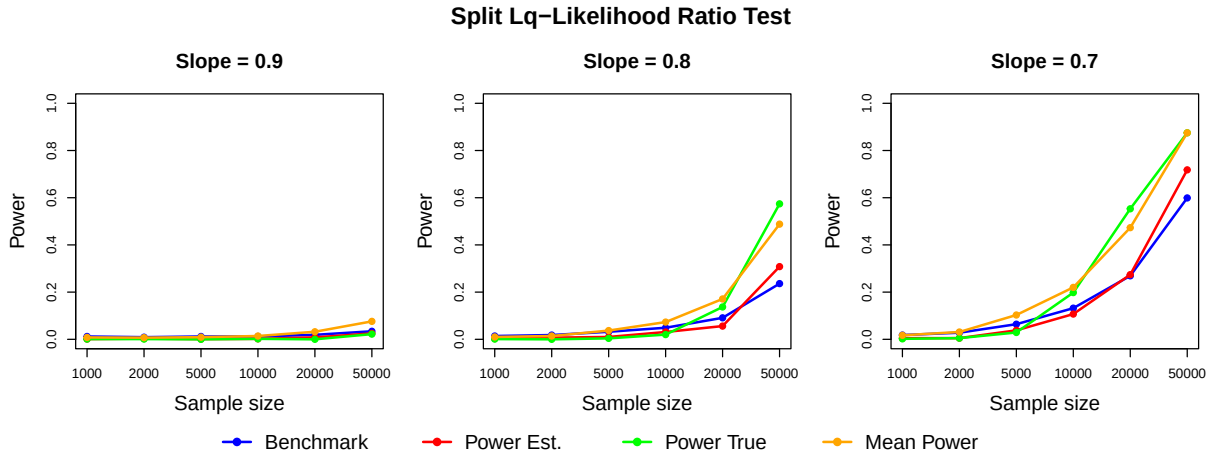

Figure .2: Power of the split LRT (Benchmark), the split LqRT with the estimated optimal power parameter (Power Est.), the split LqRT with the true optimal power parameter (Power True), and the split mean power LqRT with (universal) critical value $1/\alpha = 20$ for $B = 1$.

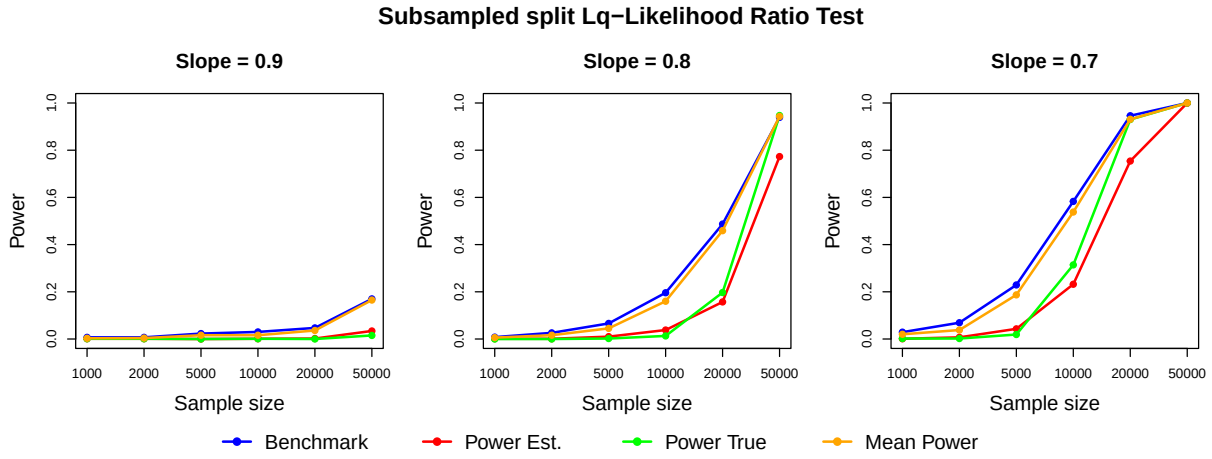

Figure .3: Power of the sub-sampled split LRT (Benchmark), the sub-sampled split LqRT with the estimated optimal power parameter (Power Est.), the sub-sampled split LqRT with the true optimal power parameter (Power True), and the sub-sampled split mean power LqRT with (universal) critical value $1/\alpha = 20$ for $B = 1000$.

Next, we consider the power of our sub-sampled split test statistics for $B = 1000$ in Figure .3. In all cases the power of the sub-sampled split LRT is now the best, followed closely by the sub-sampled split mean power LqRT. The power of the sub-sampled split LqRTs with q^* and $\widehat{q}^*$ is lower except in large samples. Hence, the ranking of the test statistics based on their power is different for $B = 1$ and $B = 1000$. The puzzling thing is that the power of the sub-sampled split LqRTs with q^* and $\widehat{q}^*$ improves only slightly and moderately in small/medium sample sizes compared to the case with $B = 1$, whereas the power of the split LRT and the mean power LqRT improves much more due to sub-sampling in all sample sizes (for all miscalibration sizes).

Consider the e-power of a sub-sampled split test statistics denoted by E. Its i.i.d. replications are denoted by $(E_b)_{b=1}^B$. It is known that

$$\mathbb{E}\left[\log\Big(\frac{1}{B}\sum_{b=1}^{B} E_b\Big)\right] \geq \mathbb{E}\Big[\log(E_1)\Big] = \mathbb{E}\left[\log\Big(\Big(\prod_{b=1}^{B} E_b\Big)^{1/B}\Big)\right],$$

by convexity of the logarithm. In the inequality above we compare the arithmetic mean with the geometric mean, that is,

$$\frac{1}{B}\sum_{b=1}^{B} E_b \quad \text{against} \quad \Big(\prod_{b=B}^{b} E_b\Big)^{1/B},$$

and, intuitively, the larger the variance of E_b, the larger the difference in the arithmetic and the geometric mean. In Figure .4 we study the standard deviation of our split test statistics. We notice that the standard deviation of the split LRT is the largest, followed by the split mean power LqRT, the standard deviations of the split LqRT with q^* and $\widehat{q}^*$ are much lower. This observation may explain the change in the ranking of the test statistics based on their power. The standard deviation of a split test statistics must reach a certain level of volatility in order to observe significant benefits from sub-sampling.

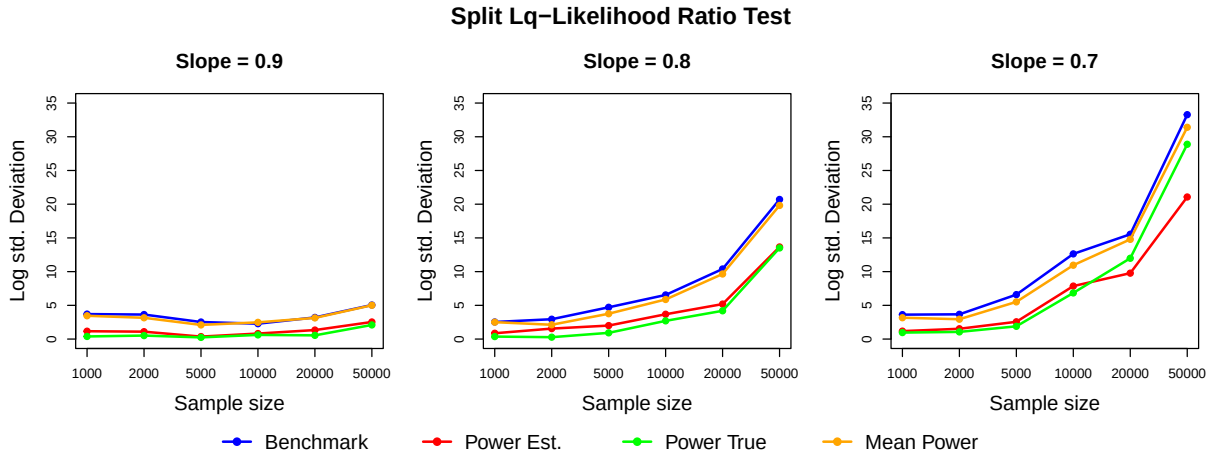

Figure .4: Logarithm of standard deviation of the power of the split LRT (Benchmark), the split LqRT with the estimated optimal power parameter (Power Est.), the split LqRT with the true optimal power parameter (Power True), and the split mean power LqRT with (universal) critical value $1/\alpha = 20$.