Small Sample Inference

by

Inna Gerlovina

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biostatistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Alan E. Hubbard, Chair
Professor Mark J. van der Laan
Professor Martyn T. Smith

Fall 2016

Small Sample Inference

Abstract

Small Sample Inference

by

Inna Gerlovina

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Alan E. Hubbard, Chair

Multiple comparisons and small sample size, common characteristics of many types of "Big Data" including those that are produced by genomic studies, present specific challenges that affect reliability of inference. Use of multiple testing procedures necessitates estimation of very small tail probabilities and thus approximation of distal tails of a test statistic distribution. Results based on large deviation theory provide a formal condition that is necessary to guarantee error rate control given practical sample sizes, linking the number of tests and the sample size; this condition, however, is rarely satisfied. Using methods that are based on Edgeworth expansions (relying especially on the work of Peter Hall), we explore what it might translate into in terms of actual error rates. Our investigation illustrates how far the actual error rates can be from the declared nominal levels, indicating poor error rate control.

Edgeworth expansions, providing higher order approximations to the sampling distribution, also offer a promising direction for data analysis that could ameliorate the situation. In Chapter 1, we derive generalized expansions for studentized mean-based statistics that incorporate ordinary and moderated one- and two-sample t-statistics as well as Welch t-test. Fifth-order expansions are generated with our developed software that can be used to produce expansions of an arbitrary order. In Chapter 2, we propose a data analysis method based on these expansions that includes tail diagnostic procedure and small sample adjustment. Using the software algorithm developed for generating expansions, we also obtain results for unbiased moment estimation of a general order. Chapter 3 introduces a general linear combination (GLC) bootstrap, which is specifically tailored for small sample size. A stabilized variance version of GLC bootstrap, based on empirical Bayes approach, is developed for high-dimensional data. Applying these methods to clustering, we propose an inferential procedure that produces pairwise clustering probabilities.

To my daughter Masha.

Contents

List of Figures

List of Tables

Acknowledgments

There are two extraordinary people without whom I cannot imagine this journey.

I am infinitely grateful to my advisor, Alan Hubbard, for his insightful and gentle guidance during my research, for his support and encouragement, and for being a kind and amazing human being. I am thankful for his patience, his fresh angles and points of view, and for allowing my ideas to take shape while providing a solid ground for them to stand on and tactfully steering them in the right direction. I also have to thank Alan for his sense of humor and his interest in Russian politics and music theory - any statistical discussion benefits greatly from these topics.

I want to thank my dad, Boris Gerlovin, for his invaluable help, for our lively and fruitful discussions, and for always being just a phone call away with his "every problem can be solved" attitude. His endless intellectual curiosity and excitement about inner workings of things have always been an inspiration to me. And I am truly thankful to my mom, Marina Gerlovina, for being a nurturing and supportive force behind these discussions.

I would like to thank all our faculty, staff, and my fellow students for contributing to this great experience and for being a part of the special atmosphere that is Berkeley Biostatistics.

Finally, I want to thank my wonderful family and friends who stood with me through thick and thin, who provided help, advice, and food - at the first request or without one; who lent an ear, were my sounding board, who knew when to ask and when not to ask, and who rooted for me at every step of the process. Thank you all.

Introduction

Explosive proliferation of recordable information that keeps Big Data at the top of the buzzword list not only calls for continuing development of analytical methodology but also ignites heated debates in statistical and scientific literature as well as in the media. Following March 14, 2014 article in *Science* magazine [46] (also [45]) disecting Google Flu Trends failure, commentaries and opinion pieces on the subject appeared in such publications as *New York Times* [49], *Financial Times* [35], and *Harvard Business Reveiw* [27]. Discussions on misuse of statistical inference as well as reliability of conclusions have a very long history, their current and highest wave starting back in 2005 when the paper titled "Why Most Published Research Findings Are False" by John Ioannidis [38] came out and quickly became a famous catalyst on the topic - "an instant cult classic" [62]. This wave culminated with American Statistical Association issuing a statement on March 7, 2016 that calls forth a new era of statistical literacy and reliability/reproducibility of scientific findings [70]. With an advent of Big Data, the enthusiasm that accompanied it and the limitless possibilities it opened were counter-balanced by the growing distrust in published findings. Concerns outlined in Ioannidis' article did not disappear with technogolical advances and restructuring of social networking that brought about massive amounts of generated (produced and collected) data. Just the opposite - the number of discoveries in many fields were rising, and so was the skepticism about how trustworthy those results might be, reflected in opinion pieces and papers such as an eloquently titled "A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null" [25].

In describing data, adjectives *big* and *small* take on special meanings and are not necessarily mutually exclusive anymore - Big Data and small sample size can be the characteristics of the same dataset. Complexity of Big data often involves multiple comparisons and thus necessitates testing a great number of hypotheses simultaneously, which only exacerbates problems posed by the small sample size, making inference yet more challenging. Much quoted words from David Speigelhalter, professor at the Statistical Laboratory of the University of Cambridge, summarize the problem with blunt simplicity: "There are a lot of small data problems that occur in big data. They don't disappear because you've got lots of the stuff. They get worse." [35]

Keeping a traditional assumption of the gravity of type I errors compared to type II errors, one of the main points of rigorous statistical inference is to protect against the proliferation of distracting and potentially misleading false positive results. This is partly addressed by a variety of multiple testing procedures; still, many other factors contribute to reliability of the conclusions. Given these factors, reliability can itself be explored, quantified, and possibly improved by the choice of an analysis method. This notion relates to the "inference on inference" issue, which has also received attention in the literature recently; for example exploring the fact that summaries such as p-values and confidence intervals are themselves statistics and carry a certain amount of noise [28]. When "big data" means that the number of tests gets progressively larger and information available for testing each of the hypotheses gets yet more limited, finite sample departures from asymptotic distributions become critical, possibly leading to biased inference and inflated false positive error rates that extend far beyond nominal level.

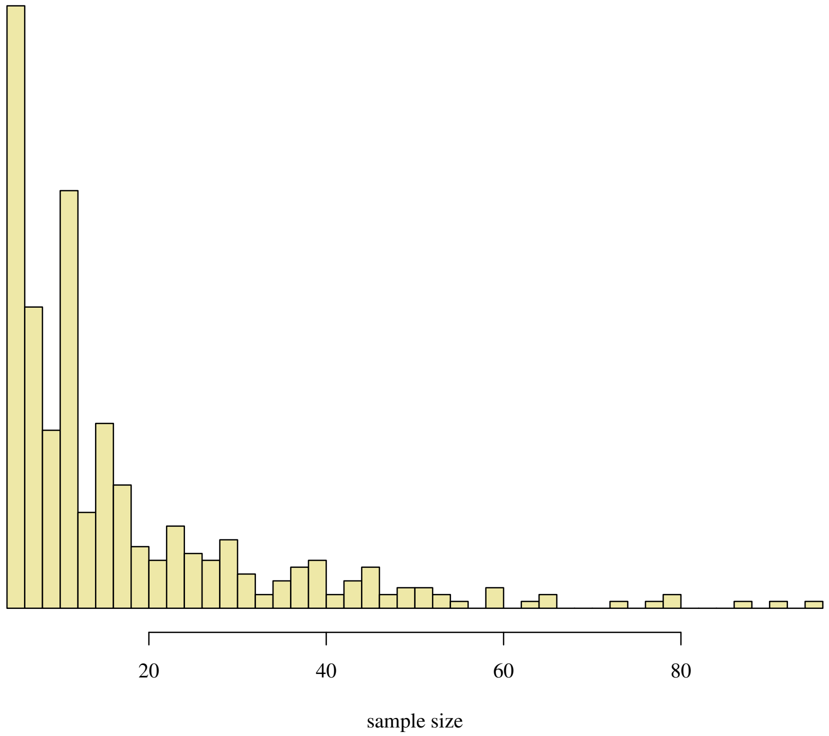
As an example of small sample high-dimensional data, consider a genomic study where the number of replicates, n, is very limited. Each replicate consists of many features, such as genes, RNA-seq, miRNA, etc.; analysis of such a study includes testing one or more hypotheses for each of the features bringing the number of tests, m, to hundreds of thousands or even millions. Accounting for multiple comparisons is commonly done by some multiple testing procedure aimed at controlling an error rate, either familywise error rate (FWER) or false discovery rate (FDR). All of these procedures move the critical value (rejection cut-off) further from the center of distribution, requiring estimation of distal tail probabilities of the test statistic distribution under the null; the greater the number of tests, the smaller the probabilities that need to be estimated for the null hypothesis to be rejected, and the more extreme tails of the distribution need to be approximated.

For a small sample size, true sampling distribution of asymptotically linear (and thus asymptotically normal) estimators might be quite far from normal. For large deviations - quantiles far away from the mean - it can potentially present an even bigger problem and adversely affect reliability of inference; even apparently small departures from normality might have a significantly amplified effect on the far tails resulting in poor approximation and, consequently, poor error rate control and ultimately faulty inferences.

How can we know the inference is to be trusted - what is needed to ensure a declared level of certainty? The most accurate inference is achieved using approaches with rigorous theoretical underpinnings, which in this case are provided by large deviation theory. The theory is used to determine the sample size that is large enough so that multivariate sampling distribution of a test statistic (such as sample average) is closely approximated by multivariate normal distribution at the appropriate quantiles (based on m) for proper error rate control. Some of the conditions that ensure the convergence of Cramer's large deviation expansion establish formal relationship between a sample size n and a quantile x of the distribution of an estimator: $x = o\big(n^{\frac{1}{6}}\big)$ [20, 52]. As noted above, in high-dimensional data the critical value (cut-off quantile) x_c is directly related to the number of tests m through a multiple testing procedure, and that leads to the condition that is necessary and sufficient to guarantee error rate control: For a Normal or Student's t approximation of the distribution of the test statistic, error rate control (either FWER or FDR) can be achieved if and only if $log(m) = o(n^{\frac{1}{3}})$ [69]. If this condition is satisfied, the actual error rates are close to the nominal ones and the null hypothesis is indeed rejected at the declared significance level α; failure to achieve it puts reported results into the territory of hope rather than probability.

How often is this condition satisfied in high-dimensional data we encounter in practice? To illustrate, we gauge the sample size that would make error rate control possible for a given number of tests. Suppose $m = 10,000$, which is relatively modest; then, if we want $log(m) = \frac{1}{10}n^{\frac{1}{3}}$, n should be $800,000$. As a comparison, sample sizes of recently uploaded datasets on GEO provide a striking reality check: sample sizes of the great majority of the studies are below 20 and very few reach above 80 as can be seen in Figure 0.1. Many of these studies are genome-wide and therefore the number of tests is usually high (starting at tens of thousands), which means error rate control is not guaranteed for such studies though it is still not clear how far off claimed results might be from the truth.

Figure 0.1: Sample size histogram - GEO datasets.



One way to avoid making assumptions on sampling distribution is to use permutation when it is relevant (testing for independence) - a nonparametric exact method. That, however, will present a problem of a different kind: the limited number of permutations available for a small sample will provide too coarse a grid to estimate very small probabilities. Then the smallest possible unadjusted p-value, which would be a result of a situation when observed data provides an extreme point of a permutation distribution, is $\frac{1}{\text{\# of permutations}}$. After multiple testing correction with large number of comparisons, the minimal adjusted p-value will be too large to pass the significance threshold in most cases and thus will never allow any signal detection.

Methods based on finite sample inequalities and bounds (such as Chebychev's inequality, Hoefding bound, and Berry-Esseen bound) provide ways to obtain reliable inference in finite samples - see, for example, exact confidence intervals for the mean, based on Bernstein and Bennett's inequalities [60], which guarantee coverage probability to be greater than or equal to the nominal level for all sample sizes and all possible data generating distributions that satisfy method's assumptions. These methods are generally conservative; in small samples, where power is already an issue, their ability to detect signal can be very limited. However, they might be preferable in situations where false positives are highly undesirable and error rate control is crucial.

Faced with the problems described above, we would like to be able to achieve more reliable inference while not giving up on gaining knowlegde in the reality of a small sample size. Since the data is so limited, the goal would be to extract maximum information from it and use the data in the most efficient way. Some natural ways to do that include resampling (where each data point is used repeatedly), calculating higher sample moments in addition to the first two and incorporating them into analysis, and, for high-dimensional data, borrowing information from the whole dataset for use in each individual test/feature. We employ these three approaches in proposed methods.

Higher moments are used in Edgeworth expansions to better approximate finite sampling distributions; these expansions are the focus of Chapters 1 and 2. To achieve our goal, it would be desirable to obtain a closer approximation to the distribution of interest (such as a true distribution of a test statistic or a null distribution for the hypothesis testing), and Edgeworth series provides the means to obtain higher-order approximations for such distirubution. Edgeworth series is an asymptotic expansion that originally extended the idea of a Central Limit Theorem providing an expansion for the distribution of a standardized sample mean, as well as a general framework for obtaining expansions of the same type for other sample statistics such as a studentized mean or a variance. It is a series of functions where the first function (which is usually denoted as a zero term) is a standard normal c.d.f. $\Phi(\cdot)$. Since it is a power series in $n^{-\frac{1}{2}}$, truncation after j'th term provides an approximation to the distribution of interest as the remainder is of the order of $n^{-\frac{j+1}{2}}$. This truncated series is usually called a j-term expansion or a $(j+1)$'th order approximation.

In Chapter 1, we derive expansions for a generalized version of mean-based statistics that incorporates various standardized and studentized means, including statistics for two-sample and Welch t-tests. Our developed software generates expansions of an arbitrary order that can be used in data analysis. Increasingly higher order approximations offer an illuminating tool for assessing possible discrepancies between actual and nominal error rates and give us hope for the approximation refinement and achieving more stabilized inference. These are the topics of Chapter 2, where empirical Edgeworth expansions are explored and adapted to small sample size and data analysis method is developed; this method includes tail diagnostic and small sample adjustment.

Resampling is the subject of Chapter 3 where we propose a generalization of non-parametric bootstrap - general linear combination (GLC) bootstrap. Two important versions of that generalization are specifically designed for small sample size: unbiased variance GLC and stabilized variance (for high-dimensional data) GLC bootstrap. They correct the variance of the bootstrap generating distirubtion and break the discreteness of the small sample size non-parametric resampling by introducing a controlled amount of noise to bootstrap distribution.

For both Edgeworth expansions and GLC bootstrap, we use moderated t-statistic [66], which is based on empirical Bayes methods that borrow information for individual tests from other features in a high-dimensional dataset, and is therefore very useful for small sample inference. It uses a hierarchical model that results in shrinkage of residual sample variances toward a common value (producing posterior variances) and therefore stabilizes feature-specific variances, addressing one of the main challenges of small sample inference. Thus, Edgeworth expansion for moderated t-statistic, which utilizes "external information" and reduces variability of the scaling factor, might be an especially attractive tool for small sample high-dimensional data analysis. In GLC bootstrap, we turn to posterior variance to implement a stabilized-variance version. When the number of tests is large, this method often results in increased power and reduced error rates.

Chapter 1

General order Edgeworth expansions for ordinary and moderated t-statistics

Higher-order approach, and especially developments based on Edgeworth expansions (EE), played an important role in statistical inference for over a century - in particular as a means to obtain more accurate approximation to the distribution of interest, to gain understanding and establish properties of methods like bootstrap (and develop inferential procedures in combination with these methods), and to compare different statistical procedures. While interest to asymptotic expansions has been sustained throughout much of this time, some advances in statistical theory and methodology brought renewed attention to EE - such as fundamental theoretical results of Bhattacharya and Ghosh [6], [7] and introduction of bootstrap [22]. More recently, proliferation of massive amounts of data, often with complicated structure, introduced specific challenges where higher-order inference procedures could be very beneficial - for example, multiple comparisons, small sample size, and high-dimensional data analysis that requires probability estimation in far tail regions as a consequence of some multiple testing procedure. For these challenges, Edgeworth expansions might offer a promising direction and become the basis for new tools with better error rate control.

Tremendous amount of research has been conducted on validity and derivation of EE for many tests, classes of estimators, and test statistics. Among them, to name just a few, are Hotelling T^2 test [42], [26], linear and non-linear regression models ([58], [44], Cox regression model [30], linear rank statistics [1], [10], [64], [41], M-estimators [44], U-statistics [9], [14], [36]. Expansions have been developed for various dependent data structures: Markov chains [5], martingales [50], autoregression and ARMA processes [67], [40]; they have also been used for sampling procedures (e.g. [72]). Some papers focus specifically on multivariate analysis [64], [26], [44], [2], [55].

Research establishing validity and theoretical properties of asymptotic expansions, starting with Cramer [20], has been the basis for developing EE. Classical Edgeworth expansion theory regarded a sum of independent identically distributed variables - a sample mean, where the mean is standardized, or scaled by a known standard deviation. This was followed by work of Petrov [52] that proved the results for sums of independent but not necessarily identically distributed random variables; later research extended EE to sums of independent and somewhat dependent random variables [64]. However, in order to use EE as an inferential tool, expansions for studentized, not standardized, statistics are needed as the variance is not normally known in practice and needs to be estimated - with t-statistic being the most important and commonly used one. First expansions for a studentized mean were derived by Chung [18] and included fourth order (3-term) expansion. Groundbreaking research by Bhattacharya and Ghosh [6] proved the validity of EE for any multivariate asymptotically linear estimator in general case. Their moment conditions for studentized mean required finite $2(k+2)$ moments for a k-term expansion (an expansion with a remainder $o\big(n^{-k/2}\big)$, also called $(k+1)$'th order EE). Next important development for t-statistic happened in 1987 when P. Hall introduced a special streamlined way of deriving EE specifically for an ordinary t-statistic, obtaining an explicit 2-term expansion [31], [33]. He proved the validity of a k-term EE for minimal moment conditions: $k+2$ finite moments, which is exactly the number of moments needed for the expansion, with a non-singularity condition on an original distribution. Work that followed was concerned with less resrictive and later optimal smoothness conditions in various cases (as well as results on Cramer condition) [12], [4], [3] and different dependence conditions for many different cases (most generally in [43]).

For many statistics and naturally for more general classes/groups of estimators, EE are presented in a general form, often in terms of cumulants of the distribution of the estimator or some intermediate statistics. As such, they are not immediately adaptable for practical implementation and would require additional steps for that. These steps can be further analytical processing, numerical methods such as numerical differentiation, or estimation of the cumulants of sampling distribution with the help of resampling methods such as jackknife [57] and Monte Carlo simulation [34]. Expansions for t-statistic presented by P. Hall [31] are expressed in terms of cumulants of the original distribution (equaled to scaled cumulants since unit variance is assumed), which is the classical form of EE for the sum/mean, and standard normal p.d.f. - exact algebraic expressions. We follow Hall's method [33], generalize it to various commonly used variants of t-statistic, including one- and two-sample t-tests as well as the Welch test, and present these results in an explicit ready-to-use form.

When sample size n is small or moderate, the difference between unbiased $s^2_{unb} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2$ and biased $s^2_b = \frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2$ variance estimates is not negligible, so it would be useful to obtain EE for a sample mean scaled by an inbiased standard error estimate. Historically, most expansions for t-statistic are developed for a biased estimate; Chung [18] mentions the unbiased version before switching to the biased one "for brevity", Hendriks at al [37] consider s^2_{unb} and suggest an approximated correction for it based on Taylor expansion. With our generalized t-statistic framework we are able to provide an exact correction resulting in EE for the unbiased version, as well as for the pooled variance in a two-sample t-test. Another useful and widely used in high-dimensional data analysis variant is a so-called moderated t-statistic (part of a more general moderated contrast estimators approach that also includes moderated F-statistic) based on empirical Bayes method [65], which also fits into our framework and for which results are presented (both one- and two-sample). In moderated t-statistic, the sample mean is scaled by a posterior variance; this approach uses a hierarchical model that leads to a shrinkage of feature-wise (e.g. gene-wise) residual sample variances towards a common value, thus producing far more stable inference when sample size is small. To accommodate these statistics and be able to incorporate them into a more general case, we introduce an adjustment to the traditional unit variance of the estimator, getting a form of characteristic function and therefore an expansion that is similar to the general non-unit variance expansion (see, for example [6], [43]). Using computer algebra, we get an exact expression for this adjustment in both one- and two-sample tests.

Throughout the paper, we adopt the terminology of $(k + 1)$'th order or k-term expansion, where normal approximation is a zero term. In most of the literature, expansions are derived up to the second or third order (Chung presents 3-term or fourth order expansion). With small samples and distributions that are far enough from Gaussian, especially highly skewed distributions, closer approximations and terms beyond second or third order might be desirable. Other advantages of having subsequent terms include providing insights into the error of lower-order approximations, or comparing different procedures based on these lower-order approximations [8]. However, obtaining these higher-order terms might be hard due to long expressions and tedious algebra. Blinnikov et al [11] propose an algorithm that helps obtain the terms for the classical standardized mean (sum of independent random variables); they calculate 12 terms using *Fortran.* Current technological capabilities allow us to develop a software that generates expansions for general case t-statistic of a desired order; our algorithm covers a standardized case as well, calculations for which are considerably faster than those for studentized statistics. Serving as a link between well developed theory and practical use, our results are directly available for practical application and the method is developed with the goal of producing user-friendly analysis tools. In the paper, some of the cases (with shorter expressions for Edgeworth expansions) are presented up to the fifth order; the terms for the other ones that are too long are not included here but are available as well and will be a part of a software package. Also, the general results can be used for any statistic if the moments of the sampling distribution are available (as functions of n).

This paper is organized as follows: section 1.1 outlines a roadmap that we follow to derive the expansions, where our main contribution comes in step 1 - finding moments of the distribution of the test statistic expressed in moments of the original distribution; the other steps are calculated with *Symbolic Python.* In section 1.2 we revisit standardized sample mean and two-sample difference in means. While expansion for a mean is well known, we demonstrate the use of the roadmap to get the results and describe the algorithm for finding moments of the mean, which will later be expanded for a more complicated studentized case. We also find scaled cumulants of the distribution of the two-sample difference in means, which will be helpful in reducing long expressions for the studentized version. Section 1.3 deals with general case: it details the steps, mainly step 1, including setup and derivation of the moments - generalizing Hall's method; it also includes an overview of the Fourier transform step that incorporates the variance adjustments and provides expressions for them; finally, that section provides general results that can be used for any estimator. Section 1.4 presents the results for the following statistics: *a)* ordinary t-statistic - one-sample: biased (results for two terms obtained by P. Hall) and unbiased variance estimates; two-sample: unequal variance, biased and unbiased variance estmates (Welch test), equal variance, unbiased variance estmate (pooled variance); *b)* moderated t-statistic - one-sample: posterior variance, unbiased variance estimate; two-sample: posterior variance, equal variance, unbiased variance estimate. We conclude with a simple illustration for higher-order approximations based on Edgeworth expansions of different orders (section 1.5).

1.1 Roadmap for deriving Edgeworth expansions

Let $X_1, \ldots, X_n$ be a sample of n i.i.d. random variables with mean μ_1 and variance σ^2 and let $\hat{\theta}$ be some normalized test statistic with c.d.f. $F_{\hat{\theta}}(\cdot)$. One of the common representations of Edgeworth expansion for $F_{\hat{\theta}}$ is

$$F_{\hat{\theta}}(x) = P\left(\hat{\theta} \leqslant x\right) = \Phi(x) + n^{-\frac{1}{2}} p_1(x)\phi(x) + n^{-1} p_2(x)\phi(x) + \ldots, \tag{1.1}$$

where polynomials $p_i(x)$ are expressed in terms of moments μ_j of distribution of X or standardized cumulants $\lambda_j = \frac{\kappa_j}{\sigma^j}$, where κ_j is a j'th cumulant of a data generating distribution. $\Phi(\cdot)$ and $\phi(\cdot)$ denote standard normal c.d.f. and p.d.f. respectively.

For a two-sample or multiple-sample test statistic, this expression should either be modified to incorporate sample sizes $n_1, n_2, \ldots$, or some summary measure n can be used to keep the same form. In his book [33], Hall uses $q_i(x)$ instead of $p_i(x)$ for EE for a studentized mean to distinguish it from that for a standardized mean. We will use $q_i(x)$ for a general case, also incorporating variance adjustment r^2 (section 1.3) to get the form:

$$F_{\hat{\theta}}(x) = \Phi\left(\frac{x}{r}\right) + \sum_{i=1}^{K} n^{-\frac{i}{2}}\, q_i(x)\phi\left(\frac{x}{r}\right) + o\left(n^{-\frac{K}{2}}\right). \tag{1.2}$$

For an arbitrary test statistic $\hat{\theta}$, to obtain an Edgeworth expansion of its distribution, the following roadmap can be used:

1. Find cumulants $\kappa_{m,\hat{\theta}}$'s of the distribution of $\hat{\theta}$ by deriving non-central moments $E(\hat{\theta}^m)$, $m = 1, 2, \ldots$.

2. Characteristic function $\varphi_{\hat{\theta}}$ of $\hat{\theta}$ is

$$\varphi_{\hat{\theta}}(t) = exp\left[\kappa_{1,\hat{\theta}} it + \kappa_{2,\hat{\theta}}\frac{(it)^2}{2} + \kappa_{3,\hat{\theta}}\frac{(it)^3}{3!} + \cdots + \kappa_{j,\hat{\theta}}\frac{(it)^j}{j!} + \cdots\right] \tag{1.3}$$

3. Taylor expansion of $\varphi_{\hat{\theta}}$ about 0:

$$\begin{aligned}\varphi_{\hat{\theta}}(t) =& 1 + \left(\kappa_{1,\hat{\theta}} it + \kappa_{2,\hat{\theta}}\frac{(it)^2}{2} + \kappa_{3,\hat{\theta}}\frac{(it)^3}{3!} + \cdots\right)\\ &+ \frac{1}{2}\left(\kappa_{1,\hat{\theta}} it + \kappa_{2,\hat{\theta}}\frac{(it)^2}{2} + \kappa_{3,\hat{\theta}}\frac{(it)^3}{3!} + \cdots\right)^2 + \cdots \end{aligned} \tag{1.4}$$

4. Collect the terms of $\varphi_{\hat{\theta}}$ according to the powers of $n^{-\frac{1}{2}}$; truncate the series to the desired order. **Note**: for two- or multiple-sample statistics, some kind of summary, such as an average of $n_1, n_2, \ldots$ can be used for that purpose.

5. Use Hermite polynomials (in place of $(it)^j$) to obtain Edgeworth expansion through Fourier transform.

6. Optional: to get the expansion in terms of scaled cumulants, substitute the moments of the original distribution with standardized cumulants λ_j.

Depending on the statistic and an order of approximation, calculations can be computationally intensive, so it might be helpful to start truncations at earlier steps. We will also address some calculation efficiency issues in later sections.

For a K-term (or, in other words, $(K+1)$'th-order) Edgeworth expansion for mean-based statistics, one needs $E(\hat{\theta}^m)$, $m = 1, \ldots, M$, where $M = K + 2$ - since j-th cumulant κ_j is of order $n^{-\frac{j-2}{2}}$ ([33] p.46) and k term will have a $n^{-\frac{k}{2}}$ factor (e.g. for a 4-term expansion we need to consider $m = 1, \ldots, 6$). For other statistics, expansions of the same order might involve a different number of moments of the original distribution: for example, K-term expansion for a sample variance will require $2(K + 2)$ moments/cumulants.

Therefore steps 2 and 3 can be rewritten as

2. $$\varphi_{\hat{\theta}}(t) = exp\left[\kappa_{1,\hat{\theta}} it + \kappa_{2,\hat{\theta}}\frac{(it)^2}{2} + \kappa_{3,\hat{\theta}}\frac{(it)^3}{3!} + \cdots + \kappa_{K+2,\hat{\theta}}\frac{(it)^{K+2}}{(K+2)!}\right]$$

3.

$$\begin{aligned}\varphi_{\hat{\theta}}(t) = & 1 + \left(\kappa_{1,\hat{\theta}} it + \kappa_{2,\hat{\theta}} \frac{(it)^2}{2} + \cdots + \kappa_{K+2,\hat{\theta}} \frac{(it)^{K+2}}{(K+2)!}\right) \\ & + \frac{1}{2}\left(\kappa_{1,\hat{\theta}} it + \kappa_{2,\hat{\theta}} \frac{(it)^2}{2} + \cdots + \kappa_{K+2,\hat{\theta}} \frac{(it)^{K+2}}{(K+2)!}\right)^2 \\ & \vdots \\ & + \frac{1}{K!}\left(\kappa_{1,\hat{\theta}} it + \kappa_{2,\hat{\theta}} \frac{(it)^2}{2} + \cdots + \kappa_{K+2,\hat{\theta}} \frac{(it)^{K+2}}{(K+2)!}\right)^K \end{aligned} \tag{1.5}$$

1.2 Standardized statistics - known variance

In this section we go over a sample mean and a two-sample difference in means scaled by their standard deviations. These statistics can be viewed as sums of random variables, and Petrov [52] proved validity of EE and derived the expansion for the sum of independent random variables for an arbitrary order. Blinnikov and Moessner [11] suggest a simplified algorithm for obtaining this expansion, calculating a twelfth-order expansion for a standardized mean with their Fortran program. Petrov's generalization allows us to easily modify the series for a sample mean to the two-sample (or possibly multiple-sample) case. With current software capabilities, we can calculate these expansions straightforwardly using the roadmap above within a short computational time. To generate expressions for moments of the distribution of the mean and calculate involved coefficients, we use a simple algorithm that will be extended in later sections for studentized statistics.

For generality, rewrite the Edgeworth series for such sum of random variables in a slightly different way:

$$F_{\hat{\theta}}(x) = \Phi(x) + r_1(x)\phi(x) + r_2(x)\phi(x) + r_3(x)\phi(x) + \cdots , \tag{1.6}$$

where r_i's have $n^{-\frac{i}{2}}$ factors embedded in them. This form makes it convenient to generalize to a two-sample difference in means where sample sizes n_x and n_y might be different. This can be expanded further using known explicit expansion for the sum and scaled cumulants $\lambda_{j,\hat{\theta}}$ of the distribution of the test statistic $\hat{\theta}$:

$$\begin{aligned} F_{\hat{\theta}}(x) = & \Phi(x) - \frac{1}{6}\lambda_{3,\hat{\theta}}(x^2-1)\phi(x) \\ & - \left[\frac{1}{24}\lambda_{4,\hat{\theta}}(x^3-3x) + \frac{1}{72}\lambda^2_{3,\hat{\theta}}(x^5-10x^3+15x)\right]\phi(x) + \cdots \end{aligned} \tag{1.7}$$

Sample mean

Consider a random variable X with the mean μ and variance σ^2.
$X_1, \dots, X_n$ is a random sample and the test statistic is

$$\hat{\theta} = \sqrt{n}\frac{\bar{X} - \mu}{\sigma} \sim F_{\hat{\theta}} \tag{1.8}$$

If we refer to the classical form of Edgeworth series (1.1), the polynomials $p_i(x)$ are:

$$p_1(x) = -\frac{1}{6}\lambda_3(x^2 - 1) \tag{1.9}$$

$$p_2(x) = \frac{1}{24}\lambda_4(x^3 - 3x) + \frac{1}{72}\lambda_3^2(x^5 - 10x^3 + 15x), \tag{1.10}$$

$$\begin{aligned} p_3(x) = -&\left[\frac{1}{120}\lambda_5(x^4 - 6x^2 + 3) + \frac{1}{144}\lambda_3\lambda_4(x^6 - 15x^4 + 45x^2 - 15)\right. \\ &\left.+ \frac{1}{1296}\lambda_3^3(x^8 - 28x^6 + 210x^4 - 420x^2 + 105)\right] \end{aligned} \tag{1.11}$$

$$\begin{aligned} p_4(x) = -&\left[\frac{1}{720}\lambda_6(x^5 - 10x^3 + 15x)\right. \\ &+ \frac{1}{1152}\lambda_4^2(x^7 - 21x^5 + 105x^3 - 105x) \\ &+ \frac{1}{720}\lambda_3\lambda_5(x^7 - 21x^5 + 105x^3 - 105x) \\ &+ \frac{1}{1728}\lambda_4\lambda_3^2(x^9 - 36x^7 + 378x^5 + 1260x^3 + 945x) \\ &\left.+ \frac{1}{31104}\lambda_3^4(x^{11} - 55x^9 + 990x^7 - 6930x^5 + 17325x^3 - 10395x)\right]. \end{aligned} \tag{1.12}$$

As mentioned earlier, $\lambda_j = \dfrac{\kappa_j}{\sigma^j}$ and κ_j's are the cumulants of the distribution of X. In particular:

$\kappa_3 = E(X - \mu)^3$; λ_3 is the skewness of the data generating distribution;
$\kappa_4 = E(X - \mu)^4 - 3\,\sigma^4$; λ_4 is the kurtosis.

For a simple standardized sample mean, $\lambda_{j,\hat{\theta}} = n^{-\frac{j-2}{2}}\lambda_j$ - to be used with EE representation in (1.7).

Expansion form presented in (1.7) also makes it explicit that each term of the expansion, having a factor of respective power of n embedded in $\lambda_{j,\hat{\theta}}$, should have only those combinations of λ's that yield that particular order - e.g. term 2 has only $\lambda_{3,\hat{\theta}}^2 = \left(n^{-\frac{1}{2}}\lambda_3\right)^2$ and $\lambda_{4,\hat{\theta}} = n^{-1}\lambda_4$. Going back to the notation with scaled cumulants λ_j of the original distribution, it translates into an observation that each p_j (as in (1.9) - (1.12)) is comprised of the terms that have only λ's associated with the correct order - thus, for example, p_3 has terms with factors λ_5, $\lambda_3\lambda_4$, and λ_3^3. This correspondence holds for standardized statistics but breaks down when we turn to statistics that incorporate estimated variance.

Following the roadmap, we get the same results. Step 1 calculations:

$$E\left(\hat{\theta}^m\right) = n^{\frac{m}{2}}\sigma^{-m}E\left[\left(\bar{X}-\mu\right)^m\right]; \tag{1.13}$$

$E(X-\mu) = 0$ and $\mu_j = E\left[(X-\mu)^j\right]$, $j = 2, \dots$ are central moments of the distribution of X. For simplicity, we can set $\mu = 0$ without any loss of generality. Then

$$E\left[\left(\bar{X}-\mu\right)^m\right] = E\left[\left(\bar{X}\right)^m\right] = \frac{1}{n^m}\sum_{i_1=1}^{n}\sum_{i_2=2}^{n}\cdots\sum_{i_m}^{n}E\left(X_{i_1}X_{i_2}\cdots X_{i_m}\right) \tag{1.14}$$

To find (1.14), we need to consider all the different combinations of ordered indices $i_1, i_2, \dots, i_m$; $i_j = 1, \dots, n$ for each j. There are n^m such combinations but many combinations yield the same $E(X_{i_1}\cdots X_{i_m})$ - for example, $E(X_2X_2X_2X_2X_2X_5X_5X_1X_1) = E(X_4X_3X_4X_6X_6X_3X_6X_6X_6) = \mu_2^2\mu_5$. Combinations that produce the same expectation form a set that we will call a *grouping*, and the problem therefore reduces to considering all the groupings (each producing a distinct expectation) and calculating their coefficients, which are the number of combinations in each set. Each product $X_{i_1}\cdots X_{i_m}$ can be broken into smaller products, or *groups*, of X's with the same indices such as $\{X_{i_j} : i_j = c\}$, $c = 1, \dots, n$. The number of groups ranges between 1 (when all the indices are the same: $i_1 = i_2 = \cdots = i_m$) and m (when all the indices are different: $i_1 \neq i_2 \neq \cdots \neq i_m$); sizes of these groups determine $E(X_{i_1}\cdots X_{i_m})$. Thus each grouping is fully characterized by the number of groups and the group sizes.

Let d denote the number of groups in one grouping G and $a_1, \dots, a_d$ - the numbers of X's in each group, $\sum_{u=1}^{d} a_u = m$; set of group sizes is unordered, so assigning indices to a's is arbitrary (e.g. decreasing). In the example above: $m = 9$, $d = 3$, $a_1 = 5$, $a_2 = 2$, and $a_3 = 2$. If $\sum_{u=1}^{d} I(a_u = 1) > 0$ (at least one group is of size 1), $E(X_{i_1}\cdots X_{i_m}) = 0$ since $E(X) = 0$ and there is no need to calculate a coefficient for this grouping, which is important in terms of computational efficiency; otherwise $E(X_{i_1}\cdots X_{i_m}) = \prod_{u=1}^{d}\mu_{a_u}$. Adding a subscript g to indicate a grouping G, we get

$$E\left[\left(\bar{X}\right)^m\right] = \sum_{all\ g} C_g \prod_{u=1}^{d_g}\mu_{a_{g,u}}, \tag{1.15}$$

where C_g is the coefficient for G, i.e. the number of combinations that yield $\{a_{g,u}\}$.

$$C_g = (n)_d \frac{\binom{m}{a_{g,1}}\binom{m - a_{g,1}}{a_{g,2}}\binom{m - a_{g,1} - a_{g,2}}{a_{g,3}} \cdots \binom{a_{g,d-1} + a_{g,d}}{a_{g,d-1}}}{s_{g,1}!\, s_{g,2}! \, \cdots}, \tag{1.16}$$

where $(n)_d = n\,(n-1)\cdots(n-d+1)$ and s_g's are the numbers of the same sized groups if there are any - e.g. for group sizes $a_1 = a_2 = 5$, $a_3 = a_4 = 4$, and $a_5 = a_6 = a_7 = 2$, we will get $s_1 = 2$, $s_2 = 2$, and $s_3 = 3$ (from these we can gather that $m = 24$, $d = 7$, and $E(X_1 \cdots X_m) = \mu_5^2\, \mu_4^2\, \mu_2^3$). Going back to our original example (group sizes $\{5, 2, 2\}$) - there is only one s_g: $s_{g,1} = 2$; the coefficient for that example is $C_g = n\,(n-1)\,(n-2)\frac{1}{2!}\binom{9}{5}\binom{4}{2}$.

One way of arriving at the expression for C_g could be the following: There are $\frac{(n)_d}{s_{g,1}!\, s_{g,2}! \, \cdots}$ ways to pick (unordered) indices that satisfy given group sizes (set $\{a_{g,u}\}$) and $\binom{m}{a_{g,1}}\binom{m - a_{g,1}}{a_{g,2}}\binom{m - a_{g,1} - a_{g,2}}{a_{g,3}} \cdots \binom{a_{g,d-1} + a_{g,d}}{a_{g,d-1}}$ ways to place these indices on m positions.

Our software generates expressions for $E\left[\left(\bar{X}\right)^m\right]$ for a given m using the method described above. To find all the possible groupings, we impose an ordering on them and use it to generate each consecutive grouping when the previous one is given, thus moving through a complete set of groupings from $\{a_1 = m\}$ to $\{a_1 = a_2 = \cdots = a_m = 1\}$. For example, in an agglomerative order, a grouping $\{5, 2, 2\}$ is preceded by $\{5, 2, 1, 1\}$ and followed by $\{5, 3, 1\}$.

The smallest number of groups is $d = 1$, which produces an order of $\frac{n}{n^m} = \frac{1}{n^{m-1}}$ (the highest order in the range); the largest d with a non-zero contribution to $E\left(\bar{X}^m\right)$ is $\lfloor \frac{m}{2} \rfloor$ (when the indices of X appear in pairs and there are no unpaired indices; when m is odd, one of the groups is of size 3), and the order it produces is $\frac{1}{n^{\lceil \frac{m}{2} \rceil}}$.

Thus we have obtained non-central moments of the distribution of $\hat{\theta}$, from which the cumulants are calculated (the rest of Step 1). Subsequent steps are calculated with *SymPy*, a Python library [39].

Two-Sample Difference in Means

As mentioned before, two-sample difference in means can be approached as a sum of independent random variables (unlike a similar case with studentized statistics) and therefore the general form (1.7) can be used. Then the only remaining task is to find $\lambda_{j,\hat{\theta}}$ for a two-sample case; a simple way to do it would be by considering the characteristic function of $\hat{\theta}$. For this case, we assume minimal restrictions and allow variances and other cumulants to be different for the two random variables that comrise the sample.

Let X, Y be independent random variables with means μ_x, μ_y, variances σ_x^2, σ_y^2, cumulants $\kappa_{j,x}$, $\kappa_{j,y}$, and scaled cumulants $\lambda_{j,x}$, $\lambda_{j,y}$. The sample is $X_1, \dots, X_{n_x}, Y_1, \dots, Y_{n_y}$. We wish to find an expansion for $F_{\hat{\theta}}$, with

$$\hat{\theta} = \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}.$$

Proceeding in the same way as the derivation of the original expansion does, we start with characteristic function of the distribution of $\bar{X} - \bar{Y}$ expressed in terms of its cumulants:

$$\begin{aligned}
\varphi_{\bar{X}-\bar{Y}}(t) &= \left[\varphi_X\left(\frac{t}{n_x}\right)\right]^{n_x} \left[\varphi_Y\left(-\frac{t}{n_y}\right)\right]^{n_y} \\
&= exp\left[n_x\left(\kappa_{1,x}\frac{it}{n_x} + \kappa_{2,x}\frac{(it)^2}{2\,n_x^2} + \cdots + \kappa_{j,x}\frac{(it)^j}{j!\,n_x^j} + \cdots\right)\right. \\
&\qquad \left. + n_y\left(\kappa_{1,y}\frac{(-it)}{n_y} + \kappa_{2,y}\frac{(-it)^2}{2\,n_y^2} + \cdots + \kappa_{j,y}\frac{(-it)^j}{j!\,n_y^j} + \cdots\right)\right] \\
&= exp\left[(\kappa_{1,x} - \kappa_{1,y})it + \left(\frac{\kappa_{2,x}}{n_x} + \frac{\kappa_{2,y}}{n_y}\right)\frac{(it)^2}{2} + \cdots + \left(\frac{\kappa_{j,x}}{n_x^{j-1}} + (-1)^j\frac{\kappa_{j,y}}{n_y^{j-1}}\right)\frac{(it)^j}{j!} + \cdots\right] \\
&= exp\left[\kappa_{1,\bar{X}-\bar{Y}}\,it + \kappa_{2,\bar{X}-\bar{Y}}\,\frac{(it)^2}{2} + \cdots + \kappa_{j,\bar{X}-\bar{Y}}\,\frac{(it)^j}{j!} + \cdots\right], \text{ where}
\end{aligned}$$

$$\begin{aligned}
\kappa_{1,\bar{X}-\bar{Y}} &= \mu_x - \mu_y \\
\kappa_{2,\bar{X}-\bar{Y}} &= \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y} \\
\kappa_{3,\bar{X}-\bar{Y}} &= \frac{\kappa_{3,x}}{n_x^2} - \frac{\kappa_{3,y}}{n_y^2}, \qquad \text{and so on.}
\end{aligned}$$

Terms for Edgeworth expansions are gathered according to their order - powers of $n^{-\frac{1}{2}}$ - so it would be helpful (in fact, necessary if we want to keep the same general form (1.1)) to have a single summary measure for representing sample size. We assume that n_x and n_y are comparable; to eliminate n_x and n_y and have a quantity n representing sample size, we introduce

$$n = \frac{n_x + n_y}{2}, \; b_x = \frac{n}{n_x}, \;\; \text{and } b_y = \frac{n}{n_y}. \tag{1.17}$$

Then $\kappa_{2,\bar{X}-\bar{Y}} = \frac{1}{n}\left(b_x\sigma_x^2 + b_y\sigma_y^2\right)$, $\kappa_{3,\bar{X}-\bar{Y}} = \frac{1}{n^2}\left(b_x^2\kappa_{3,x} - b_y^2\kappa_{3,y}\right)$, etc. Or, in terms of scaled cumulants $\lambda_{j,\bar{X}-\bar{Y}}$,

$$\lambda_{3,\bar{X}-\bar{Y}} = \frac{\kappa_{3,\bar{X}-\bar{Y}}}{\kappa_{2,\bar{X}-\bar{Y}}^{3/2}} = \frac{1}{n^{\frac{1}{2}}}\,\frac{\lambda_{3,x}\,b_x^2\sigma_x^3 - \lambda_{3,y}\,b_y^2\sigma_y^3}{\left(b_x\sigma_x^2 + b_y\sigma_y^2\right)^{\frac{3}{2}}}$$

$$\lambda_{4,\bar{X}-\bar{Y}} = \frac{\kappa_{4,\bar{X}-\bar{Y}}}{\kappa_{2,\bar{X}-\bar{Y}}^{2}} = \frac{1}{n}\,\frac{\lambda_{4,x}\,b_x^3\sigma_x^4 + \lambda_{4,y}\,b_y^3\sigma_y^4}{\left(b_x\sigma_x^2 + b_y\sigma_y^2\right)^{2}} \qquad \text{and in general}$$

$$\lambda_{j,\bar{X}-\bar{Y}} = \frac{1}{n^{\frac{j-2}{2}}}\,\frac{\lambda_{j,x}\,b_x^{j-1}\sigma_x^j + (-1)^j\,\lambda_{j,y}\,b_y^{j-1}\sigma_y^j}{\left(b_x\sigma_x^2 + b_y\sigma_y^2\right)^{\frac{j}{2}}} \tag{1.18}$$

Since $\hat{\theta} = \dfrac{(\bar{X}-\bar{Y}) - E(\bar{X}-\bar{Y})}{\left[Var(\bar{X}-\bar{Y})\right]^{1/2}} = \dfrac{(\bar{X}-\bar{Y}) - E(\bar{X}-\bar{Y})}{\kappa_{2,\bar{X}-\bar{Y}}^{1/2}}$,

$$\kappa_{j,\hat{\theta}} = \frac{\kappa_{j,\bar{X}-\bar{Y}}}{\kappa_{2,\bar{X}-\bar{Y}}^{j/2}}, \text{ and } \lambda_{j,\hat{\theta}} = \frac{\kappa_{j,\hat{\theta}}}{\kappa_{2,\hat{\theta}}^{j/2}} = \lambda_{j,\bar{X}-\bar{Y}} \text{ for all } j. \tag{1.19}$$

This shows that to get polynomials $p_i(x)$ for Edgeworth expansion in a standard form, one only needs to substitute $\dfrac{\lambda_{j,x}\,b_x^{j-1}\sigma_x^j + (-1)^j\,\lambda_{j,y}\,b_y^{j-1}\sigma_y^j}{\left(b_x\sigma_x^2 + b_y\sigma_y^2\right)^{\frac{j}{2}}}$ for λ_j, $j = 3, \ldots$ in (1.9), (1.10), etc. in a one-sample expansion.

1.3 *t*-statistic - General Case

With the possible goal of applying higher order approximations to inference in data analysis, the need arises to consider Edgeworth expansions for the estimators that incorporate random sample variance. By design, Edgeworth expansions, having standard normal c.d.f. as a base, are derived for the normalized - unit variance - statistics. However, as the true variance is usually unknown, the standardized mean-based estimators are not useful for data analysis; instead, their studentized versions, such as one or two-sample *t*-statistics, are commonly used in practice, and expansions for these statistics could have practical applications leading to potential data analysis methods that aim at closer approximations of distributions of interest.

In this section, we elaborate some of the steps of the roadmap and present the general results. Most attention is paid to step 1 in application to one- and two-sample *t*-statistics, where we propose a general framework that would incorporate different versions of these statistics, including the setup and obtaining the moments of sampling distributions in this generalized case. We introduce a variance adjustment and review step 5 that incorporates this adjustment. Then general results are presented; these results, as well as described step 5 (Fourier transform) apply to any estimator or test statistic (not just t-statistic), for which higher moments expressions are available or can be found.

Moments - General Case (Step 1)

We present Edgeworth expansions for a variety of one- and two-sample t-statistics. The difference between these statistics comes from the estimate of variance, which results in different expansions. The two main types of t-statistics that we consider are ordinary (section 1.4) and moderated (section 1.4). We start with an ordinary t-statistic with sample variance $s^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$. This is the most basic version of a sample variance and is the one for which a two-term expansion has been derived [31]. While this estimate of the variance is biased, it provides a good starting point; building on Peter Hall's derivation, we next turn to the unbiased estimate $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$, which is especially important for small n. For two-sample t-test we always assume different sample sizes and consider the cases with equal and unequal variances for the two samples; equal variance with pooled (residual) sample variance as its estimate yields a distribution of a t-statistic with smaller variance compared to the case with unequal variances and thus more power, so it is beneficial to provide expansions for this particular case as well. Moderated t-statistic is only considered with this pooled (residual) variance as it is the one that is typically used in practice ([65]). We start by looking at the scaling variance-related factor since this is the component that needs to be generalized.

One-sample setup

X is a random variable with cdf F_X.
$E(X) = \mu = 0$, $Var(X) = \sigma^2 = \mathcal{O}(1)$ (as mentioned previously, we can consider a mean-zero random variable without any loss of generality).
$X_i, \dots, X_n$ is a random sample.
One-sample mean-based test statistic can be most generally written as

$$\hat{\theta} = \frac{\bar{X}}{s_{\bar{X}}} = \frac{\sqrt{n}\,\bar{X}}{s}, \tag{1.20}$$

where s is different for different statistics - e.g. sample variance for ordinary t-statistics or posterior variance for empirical Bayes methods.

Let

$$\bar{X}_s = \frac{1}{n}\sum_{i=1}^n \left(X_i^2 - \sigma^2\right) = \frac{1}{n}\sum_{i=1}^n X_i^2 - \sigma^2 = \overline{X^2} - \sigma^2 = \mathcal{O}\left(n^{-\frac{1}{2}}\right) \tag{1.21}$$

We use notation $\overline{X^2}$ to denote the mean of X^2.

The general form that we introduce and that is useful for deriving Edgeworth expansions:

$$\hat{\theta} = n^{\frac{1}{2}} A^{-\frac{1}{2}} \bar{X} \left(1 + \gamma_1 - \gamma_2\right)^{-\frac{1}{2}}, \tag{1.22}$$

where

$$\gamma_1 = \frac{B}{A}\bar{X}_s = \mathcal{O}\left(n^{-\frac{1}{2}}\right), \text{ and} \tag{1.23}$$

$$\gamma_2 = \frac{B}{A}\bar{X}^2 = \mathcal{O}\left(n^{-1}\right). \tag{1.24}$$

$A = \mathcal{O}(1)$ and $B = \mathcal{O}(1)$ are constants that reflect the difference in the estimators. They do not depend on the sample and therefore will not be involved in expectations, so the way the moments of the test statistic are calculated is not dependent of their specific form. A is a quantity related to the variance of the original distribution and B is a supplementary adjustment in some way related to the sample size.

The expression (1.22) arises from the general case for s^2 that also makes it convenient to extract A and B for each specific case:

$$s^2 = A + B\left(\bar{X}_s - \bar{X}^2\right) = A\left(1 + \frac{B}{A}\bar{X}_s - \frac{B}{A}\bar{X}^2\right). \tag{1.25}$$

Two-sample setup

X, Y - centered random variables with cdf F_X and F_Y.

$$Var(X) = \sigma_x^2 = \mathcal{O}(1), \quad Var(Y) = \sigma_y^2 = \mathcal{O}(1). \tag{1.26}$$

The sample is $X_1, \ldots, X_{n_x}, Y_1, \ldots, Y_{n_y}$. As in section 1.2, define a single summary measure for sample size: $n = \dfrac{n_x + n_y}{2}$.

$$\hat{\theta} = \frac{\bar{X} - \bar{Y}}{s_{\bar{X}-\bar{Y}}} = \frac{\sqrt{n}\left(\bar{X} - \bar{Y}\right)}{s} \tag{1.27}$$

Note that for this case there is no straightforward interpretation for s^2; however, it is a useful construct that is analogous to the one-sample case.

Let

$$\bar{X}_s = \frac{1}{n_x}\sum_{i=1}^{n_x}\left(X_i^2 - \sigma_x^2\right) = \mathcal{O}\left(n_x^{-\frac{1}{2}}\right) \tag{1.28}$$

$$\bar{Y}_s = \frac{1}{n_y}\sum_{i=1}^{n_y}\left(X_i^2 - \sigma_y^2\right) = \mathcal{O}\left(n_y^{-\frac{1}{2}}\right). \tag{1.29}$$

The test statistic is

$$\hat{\theta} = n^{\frac{1}{2}} A^{-\frac{1}{2}} \left(\bar{X} - \bar{Y} \right) \left(1 + \gamma_1 - \gamma_2 \right)^{-\frac{1}{2}}, \tag{1.30}$$

where

$$\gamma_1 = \frac{B_x}{A} \bar{X}_s + \frac{B_y}{A} \bar{Y}_s = \mathcal{O}\left(n^{-\frac{1}{2}} \right), \text{ and} \tag{1.31}$$

$$\gamma_2 = \frac{B_x}{A} \bar{X}^2 + \frac{B_y}{A} \bar{Y}^2 = \mathcal{O}\left(n^{-1} \right). \tag{1.32}$$

Two-sample analog for a general case s^2 is

$$s^2 = A + B_x \left(\bar{X}_s - \bar{X}^2 \right) + B_y \left(\bar{Y}_s - \bar{Y}^2 \right) \tag{1.33}$$

$$= A \left[1 + \left(\frac{B_x}{A} \bar{X}_x + \frac{B_y}{A} \bar{Y}_s \right) - \left(\frac{B_x}{A} \bar{X}^2 + \frac{B_y}{A} \bar{Y}^2 \right) \right]. \tag{1.34}$$

Note that with this general form, the only difference between one- and two-sample cases is in $\bar{X}$ vs $\bar{X} - \bar{Y}$, which will allow us to make the same arguments for both cases and make the analytical calculations for $E\left(\hat{\theta}^m \right)$ very similar.

Find $E\left(\hat{\theta}^m \right)$ - one-sample case

$$\hat{\theta}^m = n^{\frac{m}{2}} A^{-\frac{m}{2}} \bar{X}^m \left(1 + \gamma_1 - \gamma_2 \right)^{-\frac{m}{2}} = n^{\frac{m}{2}} A^{-\frac{m}{2}} \bar{X}^m \left[1 + \sum_{k=1}^{\infty} a_{m,k} (\gamma_1 - \gamma_2)^k \right], \tag{1.35}$$

where

$$a_{m,k} = \frac{1}{k!} \prod_{j=0}^{k-1} \left(-\frac{m}{2} - j \right) = \frac{1}{k!\, 2^k} (-1)^k \prod_{j=0}^{k-1} (m + 2j) \tag{1.36}$$

From Taylor expansion of $(1 + \gamma_1 - \gamma_2)^{-\frac{m}{2}}$ and, subsequently, from $(\gamma_1 - \gamma_2)^k$ (1.35) we only need the terms with factors up to $n^{-\frac{M-2}{2}}$. Knowing the orders of γ_1 and γ_2 does not only allow us to use Taylor expansion in the first place, it also provides a tool to keep only the relevant terms of the expansion. One way would be to just substitute K for ∞ in (1.35); this will produce some extra terms that will be truncated at the later steps of the roadmap. However, these particular terms are longer than the rest and their calculations will be especially computationally expensive (more on it later), so it would be efficient to omit them from the beginning.

Start with grouping the terms by orders (powers of $n^{-\frac{1}{2}}$):

$$
\begin{aligned}
(1+\gamma_1-\gamma_2)^{-\frac{m}{2}} &= 1+\sum_{k=1}^{\infty} a_{m,k} \sum_{i=0}^{k} \binom{k}{i} (-1)^i \, \gamma_1^{k-i} \, \gamma_2^{i} \\
&= 1 + a_{m,1} \binom{1}{0} \gamma_1^1 \, \gamma_2^0 \\
&\quad + \left[a_{m,2} \binom{2}{2} \gamma_1^2 \, \gamma_2^0 - a_{m,1} \binom{1}{1} \gamma_1^0 \, \gamma_2^1 \right] \qquad \left(\mathcal{O}\left(n^{-1}\right) \right) \\
&\quad + \left[a_{m,3} \binom{3}{0} \gamma_1^3 \, \gamma_2^0 - a_{m,2} \binom{2}{1} \gamma_1^1 \, \gamma_2^1 \right] \qquad \left(\mathcal{O}\left(n^{-\frac{3}{2}}\right) \right) \\
&\quad + \left[a_{m,4} \binom{4}{0} \gamma_1^4 \, \gamma_2^0 - a_{m,3} \binom{3}{1} \gamma_1^2 \, \gamma_2^1 + a_{m,2} \binom{2}{2} \gamma_1^0 \, \gamma_2^2 \right] \quad \left(\mathcal{O}\left(n^{-2}\right) \right) \\
&\quad + \left[a_{m,5} \binom{5}{0} \gamma_1^5 \, \gamma_2^0 - a_{m,4} \binom{4}{1} \gamma_1^3 \, \gamma_2^1 + a_{m,3} \binom{3}{2} \gamma_1^1 \, \gamma_2^2 \right] \quad \left(\mathcal{O}\left(n^{-\frac{5}{2}}\right) \right) \\
&\quad \vdots \\
&= \Bigg(1 + \sum_{k=1}^{\infty} \Bigg[a_{m,k} \binom{k}{0} \gamma_1^k \, \gamma_2^0 - a_{m,k-1} \binom{k-1}{1} \gamma_1^{k-2} \, \gamma_2^1 + a_{m,k-2} \binom{k-2}{2} \gamma_1^{k-4} \, \gamma_2^2 \\
&\qquad\qquad - \cdots + \begin{cases} a_{m,\frac{k}{2}} (-1)^{\frac{k}{2}} \dbinom{\frac{k}{2}}{\frac{k}{2}} \gamma_1^0 \, \gamma_2^{\frac{k}{2}} \Bigg] \Bigg) & \text{- for even } k \\[2ex] a_{m,\frac{k+1}{2}} (-1)^{\frac{k-1}{2}} \dbinom{\frac{k+1}{2}}{\frac{k-1}{2}} \gamma_1^1 \, \gamma_2^{\frac{k-1}{2}} \Bigg] \Bigg) & \text{- for odd } k \end{cases} \\
&= 1 + \sum_{k=1}^{\infty} \sum_{i=0}^{\left\lfloor \frac{k}{2} \right\rfloor} a_{m,k-i} (-1)^i \binom{k-i}{i} \gamma_1^{k-2i} \, \gamma_2^{i} \qquad (1.37)
\end{aligned}
$$

From this, we can easily pick the orders that are needed for K terms of Edgeworth expansion and get

$$
\hat{\theta}^m = n^{\frac{m}{2}} A^{-\frac{m}{2}} \bar{X}^m \left[1 + \sum_{k=1}^{K} \sum_{i=0}^{\left\lfloor \frac{k}{2} \right\rfloor} a_{m,k-i} (-1)^i \binom{k-i}{i} \gamma_1^{k-2i} \, \gamma_2^{i} \right] \qquad (1.38)
$$

$$\begin{aligned}
E\left(\hat{\theta}^m\right) &= n^{\frac{m}{2}} A^{-\frac{m}{2}} \left[E\left(\bar{X}^m\right) + \sum_{k=1}^{K} \sum_{i=0}^{\left\lfloor \frac{k}{2} \right\rfloor} a_{m,k-i} (-1)^i \binom{k-i}{i} E\left(\bar{X}^m \gamma_1^{k-2i} \gamma_2^i\right) \right] \\
&= n^{\frac{m}{2}} A^{-\frac{m}{2}} \left[E\left(\bar{X}^m\right) + \sum_{k=1}^{K} \sum_{i=0}^{\left\lfloor \frac{k}{2} \right\rfloor} a_{m,k-i} (-1)^i \binom{k-i}{i} \frac{B^{k-i}}{A^{k-i}} E\left(\bar{X}^{m+2i} \bar{X}_s^{k-2i}\right) \right] \\
&= n^{\frac{m}{2}} A^{-\frac{m}{2}} \left[\rho_{m,0} + \sum_{k=1}^{K} \sum_{i=0}^{\left\lfloor \frac{k}{2} \right\rfloor} a_{m,k-i} (-1)^i \binom{k-i}{i} \frac{B^{k-i}}{A^{k-i}} \rho_{m+2i,k-2i} \right], \qquad (1.39)
\end{aligned}$$

where $\rho_{i,j} = E\left(\bar{X}^i \bar{X}_s^j\right)$.

Let

$$\begin{aligned}
\nu_{k,l} &= E\left[\bar{X}^k \left(\overline{X^2}\right)^l \right] \\
&= E\left[\left(\frac{1}{n} \sum_{i_1=1}^{n} X_{i_1} \frac{1}{n} \sum_{i_2=1}^{n} X_{i_2} \cdots \frac{1}{n} \sum_{i_k=1}^{n} X_{i_k} \right) \left(\frac{1}{n} \sum_{j_1=1}^{n} X_{j_1}^2 \cdots \frac{1}{n} \sum_{j_l=1}^{n} X_{j_l}^2 \right) \right] \\
&= \frac{1}{n^{k+l}} \sum_{i_1=1}^{n} \cdots \sum_{i_k=1}^{n} \sum_{j_1=1}^{n} \cdots \sum_{j_l=1}^{n} E(X_{i_1} \cdots X_{i_k} X_{j_1}^2 \cdots X_{j_l}^2), \qquad (1.40)
\end{aligned}$$

then

$$\begin{aligned}
\rho_{i,j} &= E\left(\bar{X}^i \bar{X}_s^j\right) = E\left[\bar{X}^i \left(\overline{X^2} - \sigma^2\right)^j \right] = E\left[\sum_{k=0}^{j} (-1)^k \binom{j}{k} \bar{X}^i \, \overline{X^2}^{j-k} \sigma^{2k} \right] \\
&= \sum_{k=0}^{j} (-1)^k \binom{j}{k} \sigma^{2k} \nu_{i,j-k}. \qquad (1.41)
\end{aligned}$$

To generate expressions for (1.40) ($\nu_{k,l}$), we extend the algorithm described in section 1.2 for equation (1.14). Now groups consist of X's and X^2's with the same indices: $\{X_{i_s}, X_{j_t}^2 \,:\, i_s = j_t = c\}$, $c = 1, \dots, n$, and are thus described not by a single number (group size) but by a pair (a, b), where a is the number of i's and b is the number of j's in the group. Consequently, a grouping in this version is characterized by a set of pairs $\{(a_u, b_u)\}$, $u = 1, \dots, d$; $\sum_{u=1}^{d} a_u = k$, $\sum_{u=1}^{d} b_u = l$, and its definition is different from the one in 1.2 since for given k and l there can be different groupings that yield the same expectation, e.g. groupings $\{(2,3),\ (3,0),\ (1,1)\}$, $\{(4,2),\ (1,1),\ (1,1)\}$, and $\{(0,4),\ (3,0),\ (3,0)\}$ will all produce $E(X_{i_1} \cdots X_{i_6} X_{j_1}^2 \cdots X_{j_4}^2) = \mu_3^2 \mu_8$. Analogously to the original version, if $\sum_{u=1}^{d} I(a = 1,\, b = 0) > 0$ (at least one pair in the grouping is $(1, 0)$),

$E(X_{i_1} \cdots X_{i_k} X_{j_1}^2 \dots X_{j_l}^2) = 0$; otherwise $E(X_{i_1} \cdots X_{i_k} X_{j_1}^2 \dots X_{j_l}^2) = \prod_{u=1}^{d} \mu_{a_u+2b_u}$.

Coefficient C_g for a grouping G is calculated in a similar way to 1.2 (equation (1.16)) with a few adjustments:

$$C_g = (n)_d \frac{\binom{k}{a_{g,1}}\binom{k-a_{g,1}}{a_{g,2}} \cdots \binom{a_{g,d-1}+a_{g,d}}{a_{g,d-1}}\binom{l}{b_{g,1}}\binom{l-b_{g,1}}{b_{g,2}} \cdots \binom{b_{g,d-1}+b_{g,d}}{b_{g,d-1}}}{s_{g,1}!\, s_{g,2}! \cdots}, \quad (1.42)$$

where $s_{g,1}, s_{g,2}, \dots$ are the numbers of the groups with same values for (a, b) (for a and b).

In this case the order ranges from $\frac{1}{n^{k+l-1}}$, when $i_1 = \dots = i_k = j_1 = \dots = j_k$ $(d = 1)$, to $\frac{1}{n^{\lceil \frac{k}{2} \rceil}}$, when all indices i_s appear in pairs if k is even ("extra" index joining one of the groups if k is odd), and all the j_t's are different from i_s's and each other $(d = \lfloor \frac{k}{2} \rfloor + l)$.

Figure 1.1 shows which $\nu_{k,l}$ are needed for different orders of Edgeworth expression. By grouping the terms by their orders and leaving out irrelevant terms in (1.38), we have cut out the area in the bottom right corner; even though the number of expressions is not comparatively large in that triangle, these are the longest expressions. To get an idea of how much computational time it saves, consider $K = 4$. Generating ν expressions for a grid (the rectangle, that would be the terms needed if we just substituted K for ∞ in the sum limits) instead of the shaded area increases computational time by a factor of more than 100.

Note that ρ's are sums that can be looked at as truncated power series in $n^{-\frac{1}{2}}$ and they contain terms that are of higher orders than needed for our goal terms of Edgeworth expansions, so they would need to be truncated. For computational efficiency, that can be done before proceeding to the next step as suggested by the roadmap.

Find $E\left(\hat{\theta}^m\right)$ - two-sample case

$$\hat{\theta}^m = n^{\frac{m}{2}} A^{-\frac{m}{2}} (\bar{X} - \bar{Y})^m (1 + \gamma_1 - \gamma_2)^{-\frac{m}{2}} \quad (1.43)$$

Using the same argument for truncation and leaving only terms of relevant orders as in the one-sample case, we get a similar expression:

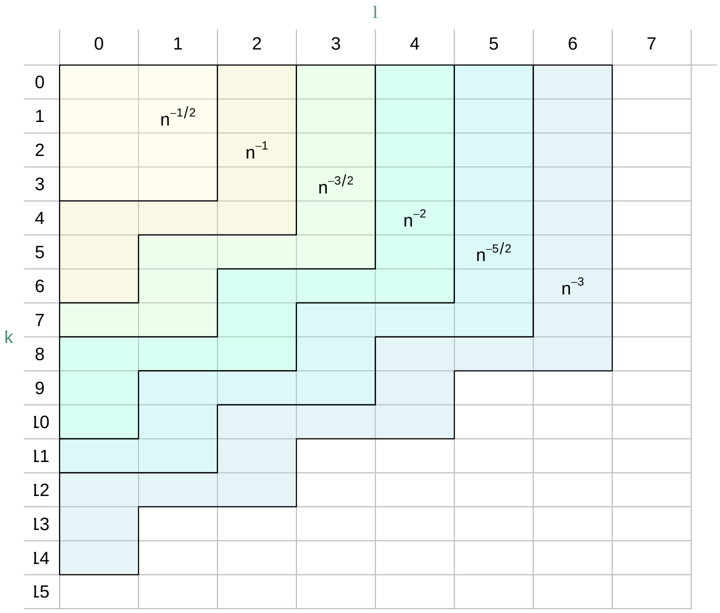


Figure 1.1: The grid showing which $\nu_{k,l}$ need to be caculated for various terms of Edgeworth expansion for a t-statistic, with terms indicated as powers of n. Combinations of k (rows) and l (columns) needed for a particular term also include all the combinations needed for previous terms as well.

$$
\begin{aligned}
\hat{\theta}^m &= n^{\frac{m}{2}} A^{-\frac{m}{2}} (\bar{X} - \bar{Y})^m \left[1 + \sum_{k=1}^{K} \sum_{i=0}^{\left\lfloor \frac{k}{2} \right\rfloor} (-1)^i a_{m,k-i} \binom{k-i}{i} \gamma_1^{k-2i}\, \gamma_2^i \right] \\
&= n^{\frac{m}{2}} A^{-\frac{m}{2}} \sum_{j=0}^{m} (-1)^j \binom{m}{j} \bar{X}^{m-j} \bar{Y}^j \left[1 + \sum_{k=1}^{K} \sum_{i=0}^{\left\lfloor \frac{k}{2} \right\rfloor} (-1)^i a_{m,k-i} \binom{k-i}{i} \gamma_1^{k-2i}\, \gamma_2^i \right], \qquad (1.44)
\end{aligned}
$$

where $a_{m,k}$ is the same as in (1.36).

$$
\begin{aligned}
E\left(\hat{\theta}^m\right) = & n^{\frac{m}{2}} A^{-\frac{m}{2}} \sum_{j=0}^{m} (-1)^j \binom{m}{j} \Bigg[E\left(\bar{X}^{m-j}\right) E\left(\bar{Y}^j\right) \qquad (1.45) \\
& + \sum_{k=1}^{K} \sum_{i=0}^{\left\lfloor \frac{k}{2} \right\rfloor} (-1)^i a_{m,k-i} \binom{k-i}{i} E\big(\bar{X}^{m-j}\, \bar{Y}^j\, \gamma_1^{k-2i}\, \gamma_2^i\big) \Bigg] \qquad (1.46)
\end{aligned}
$$

To get the expectation, we need to expand $\gamma_1^k\, \gamma_2^l$:

$$
\begin{aligned}
\gamma_1^k\, \gamma_2^l &= \frac{1}{A^k} (B_x \bar{X}_s + B_y \bar{Y}_s)^k\, \frac{1}{A^l} (B_x \bar{X}^2 + B_y \bar{Y}^2)^l \\
&= \frac{1}{A^{k+l}} \sum_{i=0}^{k} \sum_{j=0}^{l} \binom{k}{i} \binom{l}{j} B_x^{(k+l)-(i+j)} B_y^{i+j} \bar{X}^{2(l-j)} \bar{X}_s^{k-i}\, \bar{Y}^{2j}\, \bar{Y}_s^i. \qquad (1.47)
\end{aligned}
$$

$$\begin{aligned} E\left(\hat{\theta}^m\right) = & n^{\frac{m}{2}} A^{-\frac{m}{2}} \sum_{j=0}^{m} (-1)^j \binom{m}{j} \Bigg[E\left(\bar{X}^{m-j}\right) E\left(\bar{Y}^j\right) \\ & + \sum_{k=1}^{K} \sum_{i=0}^{\lfloor \frac{k}{2} \rfloor} (-1)^i a_{m,k-i} \binom{k-i}{i} A^{i-k} \sum_{u=0}^{k-2i} \sum_{v=0}^{i} \binom{k-2i}{u} \binom{i}{v} B_x^{(k-i)-(u+v)} B_y^{u+v} \\ & \times E\left(\bar{X}^{m-j+2(i-v)} \bar{X}_s^{k-2i-u}\right) E\left(\bar{Y}^{j+2v} \bar{Y}_s^u\right) \Bigg] \end{aligned} \tag{1.48}$$

Let

$$\begin{aligned} \rho_{i,j} &= E(\bar{X}^i \bar{X}_s^j), \\ \tau_{i,j} &= E(\bar{Y}^i \bar{Y}_s^j) \end{aligned} \tag{1.49}$$

and rewrite the previous equation as

$$\begin{aligned} E\left(\hat{\theta}^m\right) = & n^{\frac{m}{2}} A^{-\frac{m}{2}} \sum_{j=0}^{m} (-1)^j \binom{m}{j} \Bigg[\rho_{m-j,0}\, \tau_{j,0} \\ & + \sum_{k=1}^{K} \sum_{i=0}^{\lfloor \frac{k}{2} \rfloor} (-1)^i a_{m,k-i} \binom{k-i}{i} A^{i-k} \sum_{u=0}^{k-2i} \sum_{v=0}^{i} \binom{k-2i}{u} \binom{i}{v} B_x^{(k-i)-(u+v)} B_y^{u+v} \\ & \times \rho_{m-j+2(i-v),k-2i-u}\, \tau_{j+2v,u} \Bigg] \end{aligned} \tag{1.50}$$

Let

$$\begin{aligned} \nu_{x,k,l} &= E\left[\bar{X}^k \left(\overline{X^2}\right)^l\right] = \frac{1}{n_x^{k+l}} \sum_{i_1=1}^{n_x} \cdots \sum_{i_k=1}^{n_x} \sum_{j_1=1}^{n_x} \cdots \sum_{j_l=1}^{n_x} E(X_{i_1} \cdots X_{i_k} X_{j_1}^2 \cdots X_{j_l}^2), \text{ and} \\ \nu_{y,k,l} &= E\left[\bar{Y}^k \left(\overline{Y^2}\right)^l\right] = \frac{1}{n_y^{k+l}} \sum_{i_1=1}^{n_y} \cdots \sum_{i_k=1}^{n_y} \sum_{j_1=1}^{n_y} \cdots \sum_{j_l=1}^{n_y} E(Y_{i_1} \cdots Y_{i_k} Y_{j_1}^2 \cdots Y_{j_l}^2). \end{aligned} \tag{1.51}$$

then

$$\begin{aligned} \rho_{i,j} &= \sum_{k=0}^{j} (-1)^k \binom{j}{k} \sigma_x^{2k} \nu_{x,i,j-k}, \text{ and} \\ \tau_{i,j} &= \sum_{k=0}^{j} (-1)^k \binom{j}{k} \sigma_y^{2k} \nu_{y,i,j-k} \end{aligned} \tag{1.52}$$

For truncation of ρ's and τ's and the range of their orders, refer to section 1.3. ν_x's and and ν_y's are generated in exactly the same way as their analog for a one-sample t-statistic.

To distinguish terms of Edgeworth expansions for Studentized statistics from those with known variance, P. Hall used $q_j(x)$ in place of $p_j(x)$'s in (1.1). We take this lead but also use $q_j(x)$ to denote polynomials for the general case as well as for the statistics that incorporate unknown variance.

Fourier Transform - General Case (Step 5)

In classic Edgeworth expansions, standard normal zero term $\Phi(x)$ relies on the fact that in the log of characteristic function of $\hat{\theta}$ (1.3) the term associated with n^0 is $\frac{1}{2}\,(it)^2$ (carried over from $\kappa_{\hat{\theta},2}$ having a term consisting of 1). This holds for standardized means as well as for some versions of ordinary t-statistic. However, for a general case that we are considering, it is not necessarily true. For a one-sample statistic it is $\dfrac{\sigma^2}{2\,A}\,(it)^2$ and for a two-sample statistic $\dfrac{b_x\sigma_x^2 + b_y\sigma_y^2}{2\,A}\,(it)^2$. A is different for different statistics; and the term reduces to $\frac{1}{2}\,(it)^2$ only for ordinary one- and two-sample t-statistics using biased variance estimates. There could be various ways to address the issue; for example, if we look at the one-sample ordinary t-statistic with unbiased variance estimate, $\frac{\sigma^2}{A} = \frac{n-1}{n} = 1 - \frac{1}{n}$ (details in section 1.4), which means that part of this term actually belongs to the n^{-1} term, not the n^0; it is possible to transfer it to the "correct" term of the Edgeworth expansion. To do it in general case might not be straightforward and is not necessarily a correct way to approach the issue. Instead, we suggest using a variance adjustment that will accommodate any version of t-statistic.

The extra factor discussed above in the n^0 term of log of characteristic function suggests that we should consider a distribution of a still normalized test statistic with possible departures from unit variance in finite samples. Let r^2 denote this adjustment; note that $r^2 \to 1$ as $n \to \infty$. This is in fact a special case for a general non-unit variance EE (see, for example [6] and [43]). Starting with characteristic function, which is a special case of inverse Fourier transform, and following the usual steps for obtaining Edgeworth expansions for this general case through original Fourier transform, we get

$$\varphi(t) = \int_{-\infty}^{\infty} e^{itx} \frac{1}{\sqrt{2\pi r^2}} e^{-\frac{x^2}{2r^2}} = e^{-\frac{1}{2}t^2 r^2} \tag{1.53}$$

$$\begin{aligned} e^{-\frac{1}{2}t^2r^2} &= -\frac{1}{itr}\int_{-\infty}^{\infty} e^{itx}\phi^{(1)}\left(\frac{x}{r}\right)dx \\ &= -\frac{1}{(itr)^2}\int_{-\infty}^{\infty} e^{itx}\phi^{(2)}\left(\frac{x}{r}\right)dx \\ &\vdots \\ &= -\frac{1}{(itr)^k}\int_{-\infty}^{\infty} e^{itx}\phi^{(k)}\left(\frac{x}{r}\right)dx, \end{aligned} \tag{1.54}$$

where $\phi^{(k)}\left(\frac{x}{r}\right) = \frac{d^k}{dy^k}\,\phi(y)\Big|_{y=\frac{x}{r}}$. Then

$$\begin{aligned} (-it)^k e^{-\frac{1}{2}t^2r^2} &= \frac{1}{r^k}\int_{-\infty}^{\infty} e^{itx}\phi^{(k)}\left(\frac{x}{r}\right)dx \\ &= (-1)^k\,\frac{1}{r^k}\int_{-\infty}^{\infty} e^{itx}\,He_k\left(\frac{x}{r}\right)\phi\left(\frac{x}{r}\right)dx, \end{aligned} \tag{1.55}$$

where $He_k(x) = (-1)^k\,e^{\frac{x^2}{2}}\,\dfrac{d^k}{dx^k}\,e^{-\frac{x^2}{2}}$ are Hermite polynomials.

Therefore in Fourier transform $r^{-k}\,He_{k-1}\left(\frac{x}{r}\right)$ will be substituted for $(it)^k$. This is how we get the general Edgeworth expansion form for any normalized statistic (1.2).

The difference in expansions for different statistics are coming from $q_i(x)$ polynomials; most general results are provided in the next section and expressions for specific statistics are presented in sections 1.4 and 1.4.

Since $\varphi(t) = e^{-\frac{1}{2}t^2r^2}$, we can get explicit expressions for r^2 from $\kappa_{\hat{\theta},2}$ for one- and two-sample t-statistics:

$$r^2 = \frac{\sigma^2}{A} \qquad \text{for one-sample and} \tag{1.56}$$

$$r^2 = \frac{b_x\sigma_x^2 + b_y\sigma_y^2}{A} \qquad \text{for two-sample.} \tag{1.57}$$

The terms in (1.2) still represent orders of approximation but we need to keep in mind that there is a distinction and it would not be appropriate to directly compare these to the terms of the original EE representation as in (1.1) - the reason being that expression for A might include sample size n in some form, which means that some part of the term would belong to higher terms in the traditional form. Having said that, this general representation makes it convenient to compare approximations for different test statistics or different versions of statistics. Take the example discussed above and compare first order approximation (zero term, normal approximation) for a one-sample ordinary t-statistic with biased and unbiased variance estimate. For the same sample, biased estimate of the variance $s_b^2 = \frac{1}{n}\sum_{i=0}^{n}\left(X_i - \bar{X}\right)^2$ is smaller than unbiased s_{unb}^2 ($s_b^2 < s_{unb}^2$) and therefore $|t_b| > |t_{unb}|$, meaning that $Var(t_b) > Var(t_{unb})$. Zero terms of the expansions already reflect that: $min\left(\Phi(x), 1-\Phi(x)\right) > min\left(\Phi\left(\sqrt{\frac{n}{n-1}}\,x\right), 1-\Phi\left(\sqrt{\frac{n}{n-1}}\,x\right)\right)$. Also, $\phi(x) > \phi\left(\sqrt{\frac{n}{n-1}}\,x\right)$, which are the part of all the higher order terms.

A general remark about Edgeworth expansions for studentized statistics is in order. If we look at the first order approximation to the distribution of a test statistic, we can see that the variance of this approximation is less than or equal to 1 (can be less than 1 in the general formulation). This does not reflect the fact that in reality, for these statistics, $Var(\hat{\theta}) > 1$ (as an example consider $X \sim N(0, \sigma^2)$, for which $Var\left(\frac{\bar{X}}{s_{unb}}\right) = \frac{n-1}{n-3}$). However, higher terms of the expansion do address this problem and thicken the tails as will be shown in section 1.4.

Results - General Case

The expressions for higher order terms get progressively longer, so in order to be able to calculate truncated $F_{\hat{\theta}}(x)$ more efficiently, we break the cumulants of the distribution of $\hat{\theta}$ into smaller terms according to their order (power of $n^{-\frac{1}{2}}$), calculate them first, then use the values to find $q_i(x)$. As seen in, for example, [33], [8]:

$$\kappa_{\hat{\theta},j} = n^{-\frac{j-2}{2}}\left(k_{j,1} + n^{-1}k_{j,2} + n^{-2}k_{j,3} + \cdots\right) \qquad j \geqslant 1. \tag{1.58}$$

As mentioned previously, this method can be used for any test statistic, not just mean-based.

After splitting cumulants $\kappa_{\hat{\theta},j}$ into these terms ($k_{j,i}$), we proceed with the roadmap to get

$$q_1(x) = -\frac{1}{6\,r^3}\,k_{3,1}\,He_2\left(\frac{x}{r}\right) - \frac{1}{r}\,k_{1,2} \tag{1.59}$$

$$\begin{aligned} q_2(x) = &-\frac{1}{72\,r^6}\,k_{3,1}^2\,He_5\left(\frac{x}{r}\right) - \frac{1}{24\,r^4}\left(4\,k_{1,2}k_{3,1} + k_{4,1}\right)He_3\left(\frac{x}{r}\right) \\ &-\frac{1}{2\,r^2}\left(k_{1,2}^2 + k_{2,2}\right)He_1\left(\frac{x}{r}\right) \end{aligned} \tag{1.60}$$

$$
\begin{aligned}
q_3(x) = &- \frac{1}{1296\, r^9}\, k_{3,1}^3\, He_8\left(\frac{x}{r}\right) - \frac{1}{144\, r^7}\left(2\, k_{1,2}k_{3,1}^2 + k_{3,1}k_{4,1}\right) He_6\left(\frac{x}{r}\right) \\
&- \frac{1}{120\, r^5}\left(10\, k_{1,2}^2 k_{3,1} + 10\, k_{2,2}k_{3,1} + 5\, k_{1,2}k_{4,1} + k_{5,1}\right) He_4\left(\frac{x}{r}\right) \\
&- \frac{1}{6\, r^3}\left(k_{1,2}^3 + 3\, k_{1,2}k_{2,2} + k_{3,2}\right) He_2\left(\frac{x}{r}\right) - \frac{1}{r} k_{1,3}
\end{aligned} \tag{1.61}
$$

$$
\begin{aligned}
q_4(x) = &- \frac{1}{31104\, r^{12}}\, k_{3,1}^4\, He_{11}\left(\frac{x}{r}\right) - \frac{1}{5184\, r^{10}}\left(4\, k_{1,2}k_{3,1}^3 + 3\, k_{3,1}^2 k_{4,1}\right) He_9\left(\frac{x}{r}\right) \\
&- \frac{1}{5760\, r^8}\left(40\, k_{1,2}^2 k_{3,1}^2 + 40\, k_{2,2}k_{3,1}^2 + 40\, k_{1,2}k_{3,1}k_{4,1} + 5\, k_{4,1}^2 + 8\, k_{3,1}k_{5,1}\right) He_7\left(\frac{x}{r}\right) \\
&- \frac{1}{720\, r^6}\left(20\, k_{1,2}^3 k_{3,1} + 60\, k_{1,2}k_{2,2}k_{3,1} + 15\, k_{1,2}^2 k_{4,1} + 20\, k_{3,1}k_{3,2}\right. \\
&\qquad\qquad \left. + 15\, k_{2,2}k_{4,1} + 6\, k_{1,2}k_{5,1} + k_{6,1}\right) He_5\left(\frac{x}{r}\right) \\
&- \frac{1}{24\, r^4}\left(k_{1,2}^4 + 6\, k_{1,2}^2 k_{2,2} + 3\, k_{2,2}^2 + 4\, k_{1,3}k_{3,1} + 4\, k_{1,2}k_{3,2} + k_{4,2}\right) He_3\left(\frac{x}{r}\right) \\
&- \frac{1}{2\, r^2}\left(2\, k_{1,2}k_{1,3} + k_{2,3}\right) He_1\left(\frac{x}{r}\right),
\end{aligned} \tag{1.62}
$$

where $He_j(x)$ are Hermite polynomials, for example:

$He_1(x) = x$
$He_2(x) = x^2 - 1$
$He_3(x) = x^3 - 3x$
$He_4(x) = x^4 - 6x^2 + 3$
$He_5(x) = x^5 - 10x^3 + 15x$ and so on.

Since $k_{j,i}$'s do not depend on x, it is especially useful if $F(x)$ needs to be calculated for many values of x.

1.4 Results - Ordinary and Moderated t-statistics

Ordinary t-statistic

One-sample, biased variance estimate

$$
\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \sigma^2 + \frac{1}{n}\sum_{i=1}^{n}\left(X_i^2 - \sigma^2\right) - \bar{X}^2 \\
&= \sigma^2 + \bar{X}_s - \bar{X}^2
\end{aligned} \tag{1.63}
$$

Comparing this expression with (1.25), it is easy to see that

$$A = \sigma^2, \qquad B = 1, \qquad r^2 = 1 \tag{1.64}$$

In this particular case, $r^2 = 1$. Since this statistic is a pivotal quantity, σ^2 can be set to 1 as previously; this does not affect the results as the variance cancels out once we substitute σ^2 for A anyway. General results simplify to the following expressions for Edgeworth expression terms:

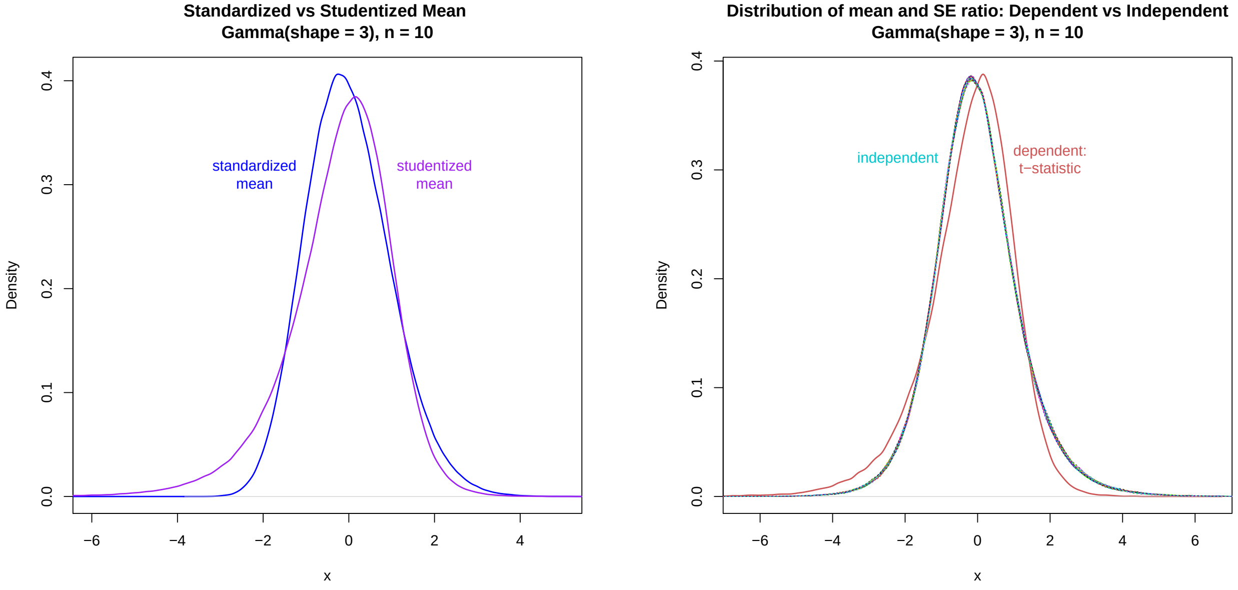
$$q_1(x) = \frac{1}{6}\,\lambda_3\left(2\,x^2 + 1\right) \tag{1.65}$$

$$q_2(x) = \frac{1}{12}\,\lambda_4\left(x^3 - 3\,x\right) - \frac{1}{18}\,\lambda_3^2\left(x^5 + 2\,x^3 - 3\,x\right) - \frac{1}{4}\left(x^3 + 3\,x\right) \tag{1.66}$$

$$\begin{aligned} q_3(x) = & -\frac{1}{40}\,\lambda_5\left(2\,x^4 + 8\,x^2 + 1\right) - \frac{1}{144}\,\lambda_3\lambda_4\left(4\,x^6 - 30\,x^4 - 90\,x^2 - 15\right) \\ & + \frac{1}{1296}\,\lambda_3^3\left(8\,x^8 + 28\,x^6 - 210\,x^4 - 525\,x^2 - 105\right) \\ & + \frac{1}{24}\,\lambda_3\left(2\,x^6 - 3\,x^4 - 6\,x^2\right) \end{aligned} \tag{1.67}$$

$$\begin{aligned} q_4(x) = & -\frac{1}{90}\,\lambda_6\left(2\,x^5 - 5\,x^3 - 15\,x\right) + \frac{1}{60}\,\lambda_3\lambda_5\left(x^7 + 8\,x^5 - 5\,x^3 - 30\,x\right) \\ & - \frac{1}{288}\,\lambda_4^2\left(x^7 - 21\,x^5 + 33\,x^3 + 111\,x\right) \\ & + \frac{1}{216}\,\lambda_3^2\lambda_4\left(x^9 - 12\,x^7 - 90\,x^5 + 36\,x^3 + 261\,x\right) \\ & - \frac{1}{1944}\,\lambda_3^4\left(x^{11} + 5\,x^9 - 90\,x^7 - 450\,x^5 + 45\,x^3 + 945\,x\right) \\ & + \frac{1}{48}\,\lambda_4\left(x^7 - 7\,x^5 + 9\,x^3 + 21\,x\right) - \frac{1}{72}\,\lambda_3^2\left(x^9 - 6\,x^7 - 12\,x^5 - 18\,x^3 - 9\,x\right) \\ & - \frac{1}{96}\left(3\,x^7 + 5\,x^5 + 7\,x^3 + 21\,x\right) \end{aligned} \tag{1.68}$$

These expansions are considerably different from the ones in section 1.2 as they capture some of the key differences between studentized and standardized estimators. To get some insight into these differences, we can turn to the Student's t-distribution with $n-1$ degrees of freedom, which was derived as a distribution of a t-statistic for a sample of n i.i.d. normally distributed random variables. Its derivation relies on a specific property unique to Gaussian distribution: independence of a sample mean and a sample standard error. Without normality, this is no longer the case, which can be easily seen with asymmetric distributions; in fact, dependence of the sample average and standard error is the cause for some important features of the sampling distribution of a studentized mean (which can present an additional challenge in trying to approximate that distribution). Consider a distribution of X that is skewed to the right, with the thin left and thick right tails. While the distribution of the standardized mean (scaled by a constant) is also skewed to the right, the distribution of the studentized mean (scaled by a random variable) is, in contrast, skewed to the left (Fig 1.2a). The reason for the "flip" stems from the fact that observations that contribute to a greater sample average, coming from the thicker tail, have greater dispersion as well, thus resulting in a smaller value for the t-statistic. Moreover, as can be seen in Fig 1.2b, the difference between thicker and thinner tails appears to be even more pronounced than that of a ratio with assumed independence (obtained with permutation/random pairings of averages and standard errors from different samples).



(a) Standardized vs studentized mean

(b) Distribution of $\bar{X}$ and s ratio: dependent vs independent

Figure 1.2

Edgeworth expansion for the studentized mean does not assume independence of sample mean and sample variance; as can be seen in Figure 1.3, with each added term, Edgeworth approximations get progressively closer to the true sampling distribution targeting the right shape and tail thickness.

Another important feature of this expansion in contrast with the ones for standardized statistics is the cumulant order inconsistency inside the expansion terms $q_2(x), q_3(x), \ldots$: as can be seen in a two-sample difference in means expansion, λ_3 is associated with $n^{-\frac{1}{2}}$, λ_4 with n^{-1}, λ_5 with $n^{-\frac{3}{2}}$, and so on. While all the terms that comprise polynomials $p_1(x), p_2(x), \ldots$ respect that order, some of the terms comprising q_j do not. Again, reference to normal distribution might provide some intuition for the reason and effect of this difference. Consider

$$X \sim N(0,1).$$

$\hat{\theta} = \sqrt{n}\dfrac{\bar{X}}{s} \sim t_{n-1}$ and $\lambda_3 = \lambda_4 = \ldots = 0$.
Then the third order expansion $F_{\hat{\theta},2}$ reduces to

$$F_{\hat{\theta},2}(x) = \Phi(x) - n^{-1}\frac{1}{4}\left(x^3 + 3x\right)\phi(x).$$

True distribution of this test statistic (Student's t) has thicker tails than a normal distribution (zero term in expansion) and the one term that does not have a λ factor in Edgeworth expansion (part of $q_2(x)$) makes for thicker tails as well, distinguishing it from the expansion for a sample average and reflecting the fact that the varince is unknown and estimated. Irregular terms in subsequent orders are likely to contribute to the thickness of the tails as well, though this is harder to assess when some λ factor is present but is not of "regular" order. We will return to this issue again when we examine the results for the two-sample case.

One-sample, unbiased variance estimate

For a small sample size the difference between $s^2 = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ and $s^2_{unb} = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ might be important, so it is useful to get Edgeworth expansion for a sample mean scaled by the unbiased estimate of the variance.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \frac{n}{n-1}\left(\sigma^2 + \bar{X}_s - \bar{X}^2\right) = C\sigma^2 + C\left(\bar{X}_s - \bar{X}^2\right), \tag{1.69}$$

where $C = \dfrac{n}{n-1}$.

$$A = C\sigma^2, \qquad B = C, \qquad r^2 = \frac{1}{C}. \tag{1.70}$$

Plugging in these expressions into general results yields

$$q_1(x) = \frac{1}{6}\,\lambda_3\left(2\,Cx^2+1\right) \tag{1.71}$$

$$\begin{aligned} q_2(x) = & \frac{1}{12}\,\lambda_4\left(C^{\frac{3}{2}}x^3 - 3\,C^{\frac{1}{2}}x\right) - \frac{1}{18}\,\lambda_3^2\left(C^{\frac{5}{2}}x^5 + 2\,C^{\frac{3}{2}}x^3 - 3\,C^{\frac{1}{2}}x\right) \\ & - \frac{1}{4}\left(C^{\frac{3}{2}}x^3 + 3\,C^{\frac{1}{2}}x\right) \end{aligned} \tag{1.72}$$

If we define $x_1 = \sqrt{C}x$, then $q_j(x)$ for this statistic will match the previous ones (from section 1.4) with x_1 in place of x for all the terms, for example:

$$q_1(x) = \frac{1}{6}\,\lambda_3\left(2\,x_1^2+1\right). \tag{1.73}$$

Two-sample results - general notes

Expressions for two-sample t-statistics are fairly long and it is recommended to use a general form for calculating $F(x)$ (section 1.3). Still, it might be interesting to look at some of the results in a traditional form that we have used before - in terms of cumulants. This will allow comparison with one-sample statistics as well as between different two-sample statistics.

Taking a lead from section 1.2, we construct two-sample analogs of the cumulants that combine λ_{xj}, λ_{yj}, variances, and sample sizes into one summary measure. They are similar to λ_j's in section 1.2 but do not necessarily have a clear interpretation and are used as a convenient notation.

Let

$$\lambda_{3a} = \frac{B_x b_x\,\lambda_{x3}\,\sigma_x^3 - B_y b_y\,\lambda_{y3}\,\sigma_y^3}{A^{\frac{3}{2}}} \tag{1.74}$$

$$\lambda_{3b} = \frac{b_x^2\,\lambda_{x3}\,\sigma_x^3 - b_y^2\,\lambda_{y3}\,\sigma_y^3}{A^{\frac{3}{2}}} \tag{1.75}$$

Then

$$q_1(x) = \frac{3\left(b_x\sigma_x^2 + b_y\sigma_y^2\right)\lambda_{3a} - A\lambda_{3b}}{6\,Ar^5}\left(x^2-r^2\right) + \frac{\lambda_{3a}}{2\,r} \tag{1.76}$$

Note that the powers of constants are the same in both λ_{3a} and λ_{3b} and match those of $\lambda_{\bar{X}-\bar{Y}}$; if $B_x = b_x$ and $B_y = b_y$, $\lambda_{3a} = \lambda_{3b}$. For consecutive terms the number of possible combinations of various constants increases (while keeping the same corresponding power), so to produce a similar expression for $q_2(x)$ and higher terms we would have to enumerate all of them.

Below are some of the $k_{j,i}$'s for lower values of j and i for a general two-sample case; the expressions get progressively longer for subsequent terms.

$$k_{1,2} = -\frac{B_x b_x \mu_{x3} - B_y b_y \mu_{y3}}{2\,A^{\frac{3}{2}}} \tag{1.77}$$

$$k_{2,1} = r^2 = \frac{b_x \mu_{x2} + b_y \mu_{y2}}{A} \tag{1.78}$$

$$k_{3,1} = -\frac{3\,(B_x b_x \mu_{x3} - B_y b_y \mu_{y3})(b_x \mu_{x2} + b_y \mu_{y2})}{A^{\frac{5}{2}}} + \frac{b_x^2 \mu_{x3} - b_y^2 \mu_{y3}}{A^{\frac{3}{2}}}, \tag{1.79}$$

where μ_{xj} and μ_{yj} are j'th moments of the distributions of X and Y respectively ($\mu_{x2} = \sigma_x^2$, $\mu_{y2} = \sigma_y^2$).

Two-sample, unequal variances, biased variance estimates

$$\hat{\theta} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}, \tag{1.80}$$

where

$$s_x^2 = \frac{1}{n_x} \sum_{i=1}^{n_x} \left(X_i - \bar{X}\right)^2 \tag{1.81}$$

$$s_y^2 = \frac{1}{n_y} \sum_{i=1}^{n_y} \left(Y_i - \bar{Y}\right)^2 \tag{1.82}$$

Set $n = \dfrac{n_x + n_y}{2}$, $b_x = \dfrac{n}{n_x}$, and $b_y = \dfrac{n}{n_y}$ as in (1.17).

$$\begin{aligned} s^2 &= b_x s_x^2 + b_y s_y^2 = b_x \left(\sigma_x^2 + \bar{X}_s - \bar{X}^2\right) + b_y \left(\sigma_y^2 + \bar{Y}_s - \bar{Y}^2\right) \\ &= b_x \sigma_x^2 + b_y \sigma_y^2 + b_x \left(\bar{X}_s - \bar{X}^2\right) + b_y \left(\bar{Y}_s - \bar{Y}^2\right) \end{aligned} \tag{1.83}$$

Using two-sample variance template (1.33), we get

$$\begin{aligned} A &= b_x \sigma_x^2 + b_y \sigma_y^2 \\ B_x &= b_x \\ B_y &= b_y \\ r^2 &= 1. \end{aligned} \tag{1.84}$$

In this case, $\lambda_{3a} = \lambda_{3b} = \lambda_{3,\bar{X}-\bar{Y}}$ (from section 1.2) and the results are mirroring the ones for a one-sample t-statistic with biased variance estimate, with some important distinctions. As mentioned in section 1.4, Edgeworth expansions for studentized statistics have "regular" and "irregular" terms in regard to the cumulant order. For this particular test statistic, if we use $\lambda_{j,\bar{X}-\bar{Y}}$ (section 1.2) to combine cumulants of the distributions of X and Y as well as b_x and b_y, then the regular terms will match the one-sample analog (ordinary t-statistic, biased variance estimate, section 1.4) exactly. The irregular terms, where the power of $n^{-\frac{1}{2}}$ does not match the cumulant order, cannot therefore be collapsed into a simple expression that is analogous to a one-sample case.

$$q_1(x) = \frac{1}{6}\,\lambda_{3,\bar{X}-\bar{Y}}\left(2\,x^2+1\right), \qquad \text{no irregular terms;} \tag{1.85}$$

$$\begin{aligned} q_2(x) = \; & \frac{1}{12}\lambda_{4,\bar{X}-\bar{Y}}\left(x^3-3\,x\right) - \frac{1}{18}\lambda^2_{3,\bar{X}-\bar{Y}}\left(x^5+2\,x^3-3\,x\right) \\ & - \frac{\left(b_x^3\sigma_x^4+b_y^3\sigma_y^4\right)(x^3+3x)+2b_xb_y\sigma_x^2\,\sigma_y^2\,(b_x+b_y)\,x}{4(b_x\sigma_x^2+b_y\sigma_y^2)^2} \end{aligned} \tag{1.86}$$

If $n_x = n_y$, the last term reduces to $-\dfrac{\left(\sigma_x^4+\sigma_y^4\right)(x^3+3\,x)+4\,\sigma_x^2\sigma_y^2 x}{4\left(\sigma_x^2+\sigma_y^2\right)^2}$. If $\sigma_y^2=0$, it becomes $-\dfrac{1}{4}\,(x^3+3\,x)$, which is an exact match of the one-sample counterpart of this statistic; naturally, in that case all the irregular terms match the corresponding one-sample case.

Two-sample, unequal variances, unbiased variance estimates

$$s_x^2 = \frac{1}{n_x-1}\sum_{i=1}^{n_x}\left(X_i-\bar{X}\right)^2 \tag{1.87}$$

$$s_y^2 = \frac{1}{n_y-1}\sum_{i=1}^{n_y}\left(Y_i-\bar{Y}\right)^2 \tag{1.88}$$

By analogy with one-sample case, let $C_x = \dfrac{n_x}{n_x-1}$ and $C_y = \dfrac{n_y}{n_y-1}$. Then

$$\begin{aligned} s^2 &= C_xb_x\left(\sigma_x^2+\bar{X}_s-\bar{X}^2\right)+C_yb_y\left(\sigma_y^2+\bar{Y}_s-\bar{Y}^2\right) \\ &= C_xb_x\sigma_x^2+C_yb_y\sigma_y^2+C_xb_x\left(\bar{X}_s-\bar{X}^2\right)+C_yb_y\left(\bar{Y}_s-\bar{Y}^2\right) \end{aligned} \tag{1.89}$$

$$\begin{aligned}
A\ &= C_x b_x \sigma_x^2 + C_y b_y \sigma_y^2 \qquad (1.90)\\
B_x &= C_x b_x\\
B_y &= C_y b_y\\
r^2 &= \frac{b_x\sigma_x^2 + b_y\sigma_y^2}{C_x b_x \sigma_x^2 + C_y b_y \sigma_y^2}. \qquad (1.91)
\end{aligned}$$

Recall that for a one-sample t-statistic with unbiased variance estimate we were able to match the expression for a simplest case (biased variance estimate) by setting $x_1 = \sqrt{C}\,x$. In this case, we could define $x_1 = \sqrt{C_x}\,x$ and $x_2 = \sqrt{C_y}\,x$ but this still would not produce the same result because of different combinations of x_1, x_2, and their coefficients. The match occurs if $n_x = n_y$ and therefore $x_1 = x_2$.

Two-sample, equal variances, unbiased variance estimates

An assumption of equal variances allows for a more efficient estimator, so there is an advantage to use pooled (or residual) variance if that assumption is warranted. As an example of increased efficiency consider normally distributed X and Y: the distribution of a t-statistic with equal variances is Student's t-distribution with $n_x + n_y - 2$ degrees of freedom, while the distribution of t-statistic with unequal variances is approximately t with $\frac{\left(s_x^2/n_x + s_y^2/n_y\right)^2}{\frac{\left(s_x^2/n_x\right)^2}{n_x-1} + \frac{\left(s_y^2/n_y\right)^2}{n_y-1}}$ degrees of freedom, which is only equal $n_x + n_y - 2$ when $s_x = s_y$ (which can never happen) and $n_x = n_y$ - and is less than that otherwise.

$Var(X) = Var(Y) = \sigma^2 = \mathcal{O}(1).$

$$\hat{\theta} = \frac{\bar{X} - \bar{Y}}{\sqrt{\left(\frac{1}{n_x} + \frac{1}{n_y}\right)\frac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{n_x+n_y-2}}} \qquad (1.92)$$

$$\begin{aligned}
s^2 &= (b_x + b_y)\frac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{n_x+n_y-2} = (b_x+b_y)\frac{\sum_{i=1}^{n_x}\left(X_i - \bar{X}\right)^2 + \sum_{i=1}^{n_y}\left(Y_i - \bar{Y}\right)^2}{n_x+n_y-2}\\
&= \frac{b_x+b_y}{n_x+n_y-2}\Big[n_x\left(\sigma^2 + \bar{X}_s - \bar{X}^2\right) + n_y\left(\sigma^2 + \bar{Y}_s - \bar{Y}^2\right)\Big]\\
&= \frac{(b_x+b_y)(n_x+n_y)}{n_x+n_y-2}\sigma^2 + \frac{(b_x+b_y)n_x}{n_x+n_y-2}\left(\bar{X}_s - \bar{X}^2\right) + \frac{(b_x+b_y)n_y}{n_x+n_y-2}\left(\bar{Y}_s - \bar{Y}^2\right) \qquad (1.93)
\end{aligned}$$

Let $C_{xy} = \dfrac{n_x + n_y}{n_x + n_y - 2} = \dfrac{n}{n-1}$ - a two-sample adjustment analogous to C for one sample. Then

$$\begin{aligned}
A \; &= C_{xy}\,(b_x + b_y)\,\sigma^2\\
B_x &= C_{xy}\,b_y\\
B_y &= C_{xy}\,b_x\\
r^2 &= \frac{1}{C_{xy}}.
\end{aligned} \tag{1.94}$$

Note the similarities and the differences between this pooled variance statisitc and the previous two-sample case - now the two samples are meshed together and B_x includes b_y (instead of b_x as previously) and vice versa.

Setting $\sigma_x^2 = \sigma_y^2 = \sigma^2$ and expanding A, B_x, and B_y in the general case, we get

$$q_1(x) = \frac{3\,\lambda_{3a} - C_{xy}\lambda_{3b}}{6\,C_{xy}\,r^5}\left(x^2 - r^2\right) + \frac{\lambda_{3a}}{2\,r} \tag{1.95}$$

In this case, $\lambda_{3a} = 2\,(\lambda_{x3} - \lambda_{y3})$.

High-dimensional data: Moderated t-statistic

Moderated t-statistic, which uses empirical Bayes approach, became a great practical tool widely used in high-dimensional data analysis. In this case, a normalizing factor is a posterior variance that incorporates prior information consisting of s_0^2 and degrees of freedom d_0 - and a sample (residual) variance. For a feature (i.e. gene) g, the following prior distribution is assumed for its variance σ_g^2:

$$\frac{1}{\sigma_g^2} \sim \frac{1}{d_0 s_0^2}\,\chi^2_{d_0}. \tag{1.96}$$

The method uses a hierachical model, in which only two hyperparameters are estimated from the data. Estimators for these parameters have a closed form and are sufficiently stable due to the fact that high dimensionality provides extensive information from which only two parameters are estimated - even when the number of replicates (sample size) is small. This allows us to treat s_0 and d_0 as constants in our approach to deriving the expansion.

Posterior variance $\tilde{s}_g^2$ is a linear combination of s_0^2 and a sample/residual variance s_g^2:

$$\tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}, \tag{1.97}$$

where d_g is residual degrees of freedom. Because of that, moderated t-statistic can also be viewed as a generalization for any scaled mean-based statistic as it can be reduced to either standardized ($d_g = 0$) or studentized ($d_0 = 0$) version. If data is distributed normally, moderated t-statistic follows a t-distribution with augmented ($d_g + d_0$) degrees of freedom.

One-sample moderated t-statistic

Following our general case convention (1.20),

$$\hat{\theta} = \frac{\sqrt{n}\,\bar{X}_g}{s} = \frac{\sqrt{n}\,\bar{X}_g}{\tilde{s}_g}, \tag{1.98}$$

with $\tilde{s}_g^2$ as in (1.97) and $s_g^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \bar{X}\right)^2$. The model does not assume $E(X_g) = 0$ for all g but, as previously, we set the mean to zero for convenience (variable substitution can be used, for example). Then

$$\begin{aligned} s^2 &= \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g} = \frac{d_0 s_0^2}{d_0 + d_g} + \frac{d_g}{d_0 + d_g}\left(C\sigma_g^2 + C\left(\bar{X}_s - \bar{X}^2\right)\right) \\ &= \frac{d_0 s_0^2 + C d_g \sigma_g^2}{d_0 + d_g} + \frac{C d_g}{d_0 + d_g}\left(\bar{X}_s - \bar{X}^2\right) \end{aligned} \tag{1.99}$$

$$A = \frac{d_0 s_0^2 + C d_g \sigma_g^2}{d_0 + d_g} = \frac{d_0 s_0^2 + n\sigma_g^2}{d_0 + n - 1} \tag{1.100}$$

$$B = \frac{C d_g}{d_0 + d_g} = \frac{n}{d_0 + n - 1} \tag{1.101}$$

since $d_g = n - 1$ for a one-sample test.

With moderated t-statistic, σ^2 cannot be simply set to 1 since it is a part of a weighted average in A and the relation of σ^2 to s_0^2 is not fixed (it is also not known in practice and will have to be estimated). However, we can still cancel σ_g^2 out by rewriting $A = D\sigma_g^2$ and presenting the results in terms of $D = \frac{d_0 \frac{s_0^2}{\sigma_g^2} + C d_g}{d_0 + d_g}$. Then $r^2 = \frac{1}{D}$.

In this case, using a general form of expansion (section 1.3) is preferred from a computational standpoint but to give an idea of the inner structure of the expansions for moderated t-statistics we provide a few terms here:

$$q_1(x) = \frac{1}{6}\left(3\,Bx^2 - Dx^2 + 1\right)\lambda_3 \tag{1.102}$$

$$
\begin{aligned}
q_2(x) = &-\frac{1}{24}\,\lambda_4\Big[\Big(3\,B^2D^{-\frac{1}{2}} - 6\,BD^{\frac{1}{2}} + D^{\frac{3}{2}}\Big)x^3 + 3\,\Big(B^2D^{-\frac{3}{2}} + 2\,BD^{-\frac{1}{2}} - D^{\frac{1}{2}}\Big)x\Big] \\
&-\frac{1}{72}\,\lambda_3^2\Big[\Big(9\,B^2D^{\frac{1}{2}} - 6\,BD^{\frac{3}{2}} + D^{\frac{5}{2}}\Big)x^5 - 2\,\Big(9\,B^2D^{-\frac{1}{2}} - 18\,BD^{\frac{1}{2}} + 5\,D^{\frac{3}{2}}\Big)x^3 \\
&\qquad - 3\,\Big(3\,B^2D^{-\frac{3}{2}} + 6\,BD^{-\frac{1}{2}} - 5\,D^{\frac{1}{2}}\Big)\,x\Big] \\
&-\frac{1}{4}\quad\Big[B^2D^{-\frac{1}{2}}\,x^3 + \Big(B^2D^{-\frac{3}{2}} + 2\,BD^{-\frac{1}{2}}\Big)\,x\Big] \qquad (1.103)
\end{aligned}
$$

Note that combined orders of B and D are consistent with the power of x and match the order of C in the ordinary t-statistics with unbiased variance estimate case. That t-statistic can be considered a special case of the moderated t, where $d_0 = 0$, so there is no prior information, and $B = D = C$.

Two-sample moderated t-statistic

For this estimator, we only consider equal variance for X and Y since it has been developed for the residual (pooled) variance only:

$$\hat{\theta} = \frac{\bar{X} - \bar{Y}}{\sqrt{\left(\frac{1}{n_x} + \frac{1}{n_y}\right)\tilde{s}_g}}, \qquad (1.104)$$

where $\tilde{s}_g^2$ is as in (1.97) and $s_g^2 = \dfrac{\sum_{i=1}^{n_x}(X_i - \bar{X})^2 + \sum_{i=1}^{n_y}(Y_i - \bar{Y})^2}{n_x + n_y - 2}$.

Following our two-sample general convention (1.27) and using (1.93), we get

$$
\begin{aligned}
s^2 &= (b_x + b_y)\,\tilde{s}_g^2 = (b_x + b_y)\,\frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g} \\
&= \frac{(b_x + b_y)\,d_0 s_0^2}{d_0 + d_g} + \frac{d_g}{d_0 + d_g}\Big[C_{xy}(b_x + b_y)\sigma_g^2 + C_{xy}\,b_y\left(\bar{X}_s - \bar{X}^2\right) + C_{xy}\,b_x\left(\bar{Y}_s - \bar{Y}^2\right)\Big] \\
&= \frac{(b_x + b_y)\left(d_0 s_0^2 + C_{xy}\,d_g\sigma_g^2\right)}{d_0 + d_g} + \frac{C_{xy}\,d_g b_y}{d_0 + d_g}\left(\bar{X}_s - \bar{X}^2\right) + \frac{C_{xy} d_g b_x}{d_0 + d_g}\left(\bar{Y}_s - \bar{Y}^2\right). \qquad (1.105)
\end{aligned}
$$

Note that $d_g = n_x + n_y - 2$ in this case.

$$A = \frac{(b_x + b_y)\left(d_0 s_0^2 + C_{xy}\, d_g \sigma_g^2\right)}{d_0 + d_g} = D\,\sigma_g^2, \text{ where } D = \frac{(b_x + b_y)\left(d_0 \frac{s_0^2}{\sigma_g^2} + C_{xy}\, d_g\right)}{d_0 + d_g}$$

$$B_x = \frac{C_{xy}\, d_g b_y}{d_0 + d_g}$$

$$B_y = \frac{C_{xy}\, d_g b_x}{d_0 + d_g}$$

$$r^2 = \frac{b_x + b_y}{D}. \tag{1.106}$$

Using the general results and the fact that $\sigma_x^2 = \sigma_y^2 = \sigma^2$, we get

$$q_1(x) = \frac{3\,(b_x + b_y)\,\lambda_{3a} - D\,\lambda_{3b}}{6\,D\,r^5}\left(x^2 - r^2\right) + \frac{\lambda_{3a}}{2\,r} \tag{1.107}$$

Here $\lambda_{3a} = \dfrac{C_{xy}\, d_g b_x b_y\,(\lambda_{x3} - \lambda_{y3})}{D\,(d_0 + d_g)}$ and $\lambda_{3b} = \dfrac{b_x^2\,\lambda_{3x} - b_y^2\,\lambda_{3y}}{D}$.

1.5 Illustration of Higher-Order Approximations

To provide an illustration for higher-order approximations to the distribution of a t-statistic, we return to the simple example from section 1.4 with a skewed original distribution (Gamma with shape 3) and small sample size ($n = 10$). Fig 1.3 displays approximations of orders 1 through 5, and we look at ordinary t-statistics with biased and unbiased variance estimates.

Distributions of these versions of t-statistic are skewed to the left, and this is the side that we will focus on. Edgeworth expansions are not probability functions and do not have their properties - they are not necessarily monotonic everywhere and might not be bounded but 0 and 1. This behavior is usually localized in the thinner tail of the distribution and therefore EE are not very helpful there; it is clearly seen in the second order approximation (term 1) in the graph. We can also see that distribution of a biased version is more spread out in both true distributions and their Edgeworth expansion approximations - as discussed in section 1.3.

The difference between the normal approximation (term 0) and the true distribution is quite striking; subsequent orders improve approximation considerably. It appears that the third order is already fairly close to the truth; however, as we move further into the tail, we can see that approximation gets further from the distribution and higher order terms come

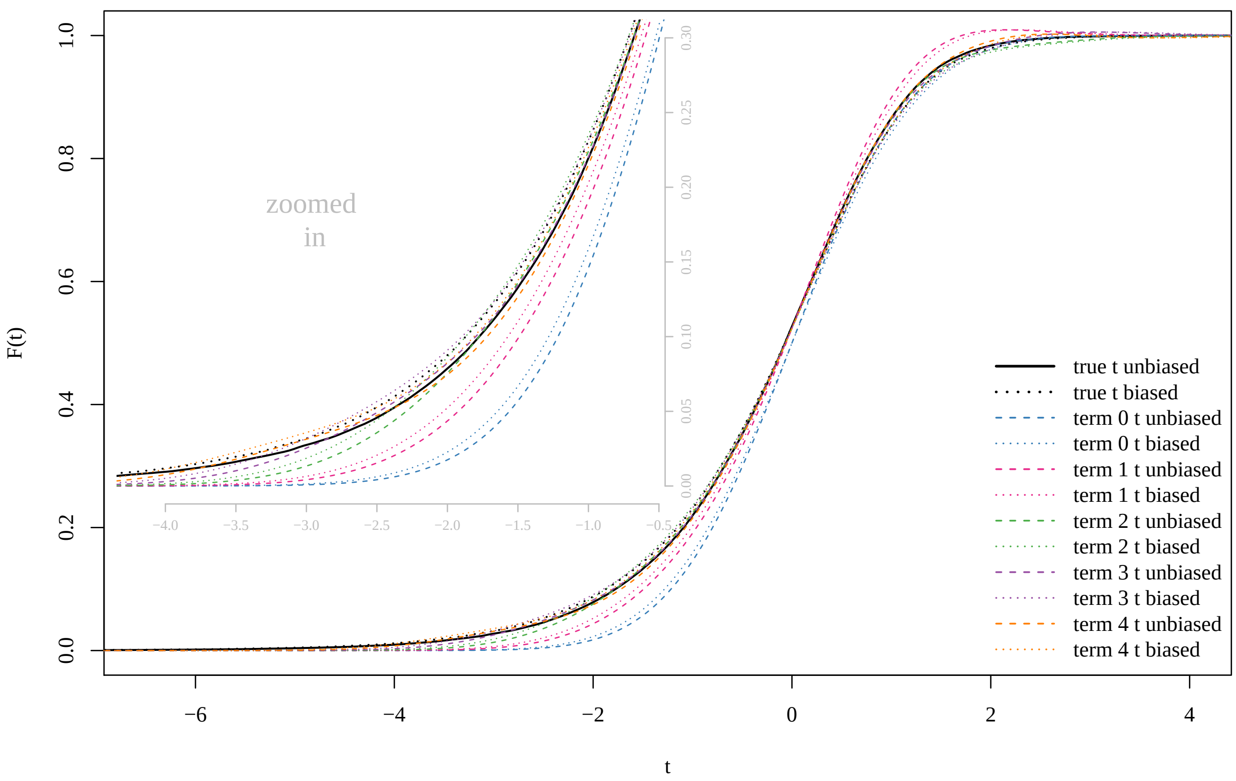


Figure 1.3: Ordinary t-statistic with biased and unbiased variance estimate, Edgeworth expansions approximation. $\Gamma(k = 3)$, $n = 10$.

into play proving the value of Edgeworth expansion of the orders beyond the second and even third.

Chapter 2

Edgeworth expansions approach to data analysis

In this chapter, we use Edgeworth expansions to assess error rates and their connection to sample size and number of tests, and to develop a data analysis method based on higher sample moments and cumulants. It would be illuminating to see what the difference between theory and practice in terms of sample size n translates into for actual error rate control: to explore the effect that some departures from normality might have on type I error rates and show the extent of the resulting discrepancies between actual and nominal error rates. We are specifically interested in how increasing the number of tests m affects the accuracy of the analyses involving multiple testing procedures, with special focus on situations when $m \gg n$. Edgeworth expansions provide a way of comparing cut-offs obtained with different orders of approximations for the same dataset and evaluating them under the true sampling distribution of a chosen parameter. In the reality of a finite sample, Edgeworth series allow us to get closer to the critical value that would be calculated had the true distribution been known by obtaining higher-order approximations of that distribution. As such, they also suggest a promising direction for data analysis where empirical moments could be substituted for the true moments in the higher-order terms.

It is important to keep in mind that Edgeworth expansion is not a probability function and is not guaranteed to have its properties. It does not necessarily converge as an infinite series: for a continuous random variable X, the condition needed for convergence is $E\left(e^{\frac{1}{4}\hat{\theta}^2}\right)$ ([20], [33] p. 45), which does not hold true even for a standardized mean of the exponential distribution ($X \sim Exp(\lambda)$, $\hat{\theta} = \frac{X-\mu}{\sigma}$). Truncated series might not be a monotonic function, its values might occasionally reach beyond $[0, 1]$ bound, and it does not necessarily integrate to 1. While the function in the regions where this happens cannot be used for assessment or inference, we will show that this usually happens in the “thin” tail of the distribution, and that the behavior of the function can itself be used as a diagnostic tool that might help infer the shape of the distribution we are trying to approximate. For the proposed method of using the substitution expansion for inference, these regions of non-monotonicity are easily detected and first order approximation is used in place of a higher-order one, ensuring that the inference will stay on the conservative side.

Sampling distribution of t-statistics in small samples and especially behavior of large deviations from the mean are explored in this Chapter, along with the performance of normal-based EE in the far tail regions. As a result, to reach extreme tails of the true distribution and get a closer approximation in those regions, we introduce a small sample adjustment that uses Student's t-distribution as a base for the expansions instead of standard normal. Small sample size is also the setting for investigation of moment estimates that are used in empirical Edgeworth expansions, where sample moments are substituted for the true ones. We use our software to produce unbiased estimates of moments and their products and powers of arbitrary order, and discuss the use of these estimates in Edgeworth expansions.

For explorations and simulations we choose some known distributions that yield analytical expressions for central moments/scaled cumulants that are the main "ingredients" of Edgeworth expansions, possibly with closed form distributions of sample means, which is helpful for calculation of very small probabilities that are involved in extreme tail estimation. Of particular interest to us are asymmetric/skewed distributions, especially their thicker tails as the thick tail of a sampling distribution is the region where a first order approximation can potentially fail to deliver error rate control producing faulty inference. As examples of such skewed distributions with parameters that allow straightforward manipulation of the the shape of the tails, we concentrate on gamma (which has a closed form for sample average) and log-normal distributions. For a skewed distribution with infinite support, which is useful to explore in regard to extreme tails, we pick a mixture-normal distribution that provides many parameters to choose from and has an analytical expression for $F_{\bar{X}}(x)$.

The chapter begins with Edgeworth expansions for standardized sample mean as a tool to explore how error rates change as n gets smaller and m grows; after that we move to the expansions for distributions of studentized test statistics. Performance of these expansions is first evaluated with known true moments and cumulants; then empirical expansions with plug-in estimates are introduced. Tail diagnostic exploration leads to the proposed data analysis procedure. Finally, we look at various types of simulations that include high-dimensional data and EE for moderated t-statistic, among others.

2.1 Dependence of error rates on sample size and number of tests

Using standardized sample mean - the most basic test statistic - as an example, we compare actual and nominal error rates for various distributions, choosing different sample sizes n

(starting with very small) and looking at the progressively smaller probabilities in the far tails that correspond to the number of tests m through some multiple testing procedure. These error rates will be compared across various orders of approximation, including first order (standard normal) with nominal significance level α.

To obtain an actual error rate for such a test given the order k, sample size n, number of tests m, significance level α, distribution of X, and multiple testing procedure:

1. calculate the unadjusted probability cut-off corresponding to α: e.g. for Bonferroni multiple testing correction $p = \dfrac{\alpha}{m}$ for the left tail and $p = \dfrac{1-\alpha}{m}$ for the right;
2. find the corresponding critical value (quantile) $q = F_{k,n}^{-1}(p)$, where $F_{k,n}$ is a k'th order Edgeworth expansion;
3. find *p-value* for this quantile that is based on the true sampling distribution F_0 of the estimator: $p_{true} = F_0(q)$ for the left tail and $p_{true} = 1 - F_0(q)$ for the right;
4. express it in terms of the actual error rate r: e.g $R = max(m\, p_{true},\, 1)$ if Bonferroni MTC is used.

Note that R needs to be bounded by 1 since it's a probability; a value of R that is greater than 1 has no statistical interpretation; however, in some cases it might be helpful to look at that unsubstituted value to gauge the "magnitude of a disaster" and compare it with other unsubstituted values.

For this case study we use $\alpha = 0.05$ and Bonferroni MTC. The choice of multiple testing procedure is not a subject of evaluation in our studies - it is just a basis for comparison; this particular choice is dictated by convenience since it is not a step-down procedure and therefore does not require sorting of *p-values* and can be easily reversed for step 4. It is also equivalent to FDR in case of the global null. Still, the results demonstrated with this MTP are probably more stark compared to other procedures since it reaches the largest deviations from the mean in critical values.

Moments, cumulants, and distribution of a standardized mean

As in the previous chapter, let X be a random variable with mean μ and variance σ^2, $X_1, X_2, \ldots, X_n$ - random sample, and $\hat{\theta} = \dfrac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$ - a test statistic.

Convenient distributions of X for this case study would be gamma and mixture-normal since the distributions of $\hat{\theta}$ in these cases have closed form.

Gamma distribution - sample average is also gamma distributed:

$$X \sim \Gamma(\alpha, \beta) - \frac{\alpha}{\beta}$$
$$\hat{\theta} \sim \Gamma(n\alpha, \sqrt{n\alpha}) - \sqrt{n\alpha},$$

where α is a shape parameter and β is a rate parameter.

Non-central moments of the distirubtion of X:

$$\mu'_j = \frac{\alpha + j - 1}{\beta}\, \mu'_{j-1},\ j = 2, 3, \ldots$$

Scaled cumulants:

$$\lambda_j = \frac{(j-1)!}{\alpha^{\frac{j-2}{2}}},\ j = 3, 4, \ldots$$

Mixture-normal (we will use bi-normal) distribution:

$$X \sim p\, N(\tilde{\mu}_1, \sigma_1^2) + (1-p)\, N(\tilde{\mu}_2, \sigma_2^2)$$

$$f_X(x) = p\left[\frac{1}{\sqrt{2\pi}\sigma_1} exp\left(-\frac{(x-\tilde{\mu}_1)^2}{2\sigma_1^2}\right)\right] + (1-p)\left[\frac{1}{\sqrt{2\pi}\sigma_2} exp\left(-\frac{(x-\tilde{\mu}_2)^2}{2\sigma_2^2}\right)\right]$$

By varying the distance between $\tilde{\mu}_1$ and $\tilde{\mu}_2$ and σ_1^2/σ_2^2, we can produce sampling distributions with different tail thickness. For convenience, we can set $E(X) = 0$ and specify the difference in means instead of specifying both means of the normal distributions in our mixture distribution: $\mu_d = \tilde{\mu}_2 - \tilde{\mu}_1$. Then $\tilde{\mu}_1 = -(1-p)\,\mu_d$ and $\tilde{\mu}_2 = p\,\mu_d$.

Let $a = \frac{\sqrt{n}}{\sigma}$, where $\sigma^2 = Var(X)$. Then $\hat{\theta} = a\bar{X}$ and

$$f_{\hat{\theta}}(x) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \frac{1}{\sqrt{2\pi}\sigma_k} exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right) \tag{2.1}$$

$$F_{\hat{\theta}}(x) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \Phi\left(\frac{x-\mu_k}{\sigma_k}\right), \tag{2.2}$$

where $\mu_k = \frac{k\tilde{\mu}_1 + (n-k)\tilde{\mu}_2}{\sqrt{n}\,\sigma} = \left(p - \frac{k}{n}\right)\frac{\sqrt{n}\,\mu_d}{\sigma}$ and $\sigma_k^2 = \frac{k\sigma_1^2 + (n-k)\sigma_2^2}{n\sigma^2}$.

Moments of the distribution of X:
Assume $\mu_1 = E(X) = 0$ and let $q = 1 - p$. Then

$$
\begin{aligned}
\mu_2 &= p\sigma_1^2 + q\sigma_2^2 + \mu_d^2\, pq \\
\mu_3 &= 3\,\mu_d\, pq\left(\sigma_2^2 - \sigma_1^2\right) + \mu_d^3\, pq\,(2\,p - 1) \\
\mu_4 &= 3\,\left(p\sigma_1^4 + q\sigma_2^4\right) + 6\,\mu_d^2\, pq\left(q\sigma_1^2 + p\sigma_2^2\right) + \mu_d^4\, pq\left(3\,p^2 - 3\,p + 1\right) \\
\mu_5 &= 15\,\mu_d\, pq\left(\sigma_2^4 - \sigma_1^4\right) + 10\,\mu_d^3\, pq\left(p^2\sigma_2^2 - q^2\sigma_1^2\right) + \mu_d^5\, pq\left(4\,p^3 - 6\,p^2 + 4\,p - 1\right) \\
\mu_6 &= 15\,\left(p\sigma_1^6 + q\sigma_2^6\right) + 45\,\mu_d^2\, pq\left(q\sigma_1^4 + p\sigma_2^4\right) + 15\,\mu_d^4\, pq\left(q^3\sigma_1^2 + p^3\sigma_2^2\right) \\
&\quad + \mu_d^6\, pq\left(5\,p^4 - 10\,p^3 + 10\,p^2 - 5\,p + 1\right)
\end{aligned}
$$

Derivation of (2.1) **- density of $\hat{\theta}$ (mixture-normal):**

In general, the sampling distribution of the mean of a random variable with a known theoretical density can be found using characteristic functions. The fact that characteristic function of the sum of random variables is the product of characteristic functions of these variables:

$$\varphi_{a_1X_1+\ldots+a_nX_n}(t) = \varphi_{X_1}(a_1t)\cdot\ldots\cdot\varphi_{X_n}(a_nt)$$

allows us to find a characteristic function of the sampling distribution of the mean. Using reverse Fourier transform will recover the density of that distribution, since characteristic function is a special case of Fourier transform.

Characteristic function for a univariate normal random variable $Y \sim N(\mu, \sigma^2)$ is $\varphi_Y(t) = exp\left(i\mu t - \dfrac{1}{2}\sigma^2 t^2\right)$, and we use linearity of integration to get a characteristic function for X:

$$\varphi_X(t) = p\, exp\left(i\tilde{\mu}_1 t - \frac{1}{2}\sigma_1^2 t^2\right) + (1-p)\, exp\left(i\tilde{\mu}_2 t - \frac{1}{2}\sigma_2^2 t^2\right)$$

Characteristic function of a sample mean:

$$\varphi_{\bar{X}}(t) = \left[\varphi_X\left(\frac{t}{n}\right)\right]^n = \left[p\,exp\left(\frac{i\tilde{\mu}_1 t}{n} - \frac{1}{2}\frac{\sigma_1^2 t^2}{n^2}\right) + (1-p)exp\left(\frac{i\tilde{\mu}_2 t}{n} - \frac{1}{2}\frac{\sigma_2^2 t^2}{n^2}\right)\right]^n$$

$$= \sum_{k=0}^{n}\binom{n}{k}p^k exp\left(\frac{ik\tilde{\mu}_1 t}{n} - \frac{k\sigma_1^2 t^2}{2n^2}\right)\cdot(1-p)^{n-k}exp\left(\frac{i(n-k)\tilde{\mu}_2 t}{n} - \frac{(n-k)\sigma_2^2 t^2}{2n^2}\right)$$

$$= \sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}exp\left(\frac{i\big(k\tilde{\mu}_1 + (n-k)\tilde{\mu}_2\big)t}{n} - \frac{\big(k\sigma_1^2 + (n-k)\sigma_2^2\big)t^2}{2n^2}\right)$$

$$= \sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}exp\Big(it\tilde{\mu}_k - \frac{1}{2}\tilde{\sigma}_k^2 t^2\Big),$$

where $\tilde{\mu}_k = \dfrac{k\tilde{\mu}_1 + (n-k)\tilde{\mu}_2}{n}$, and $\tilde{\sigma}_k^2 = \dfrac{k\sigma_1^2 + (n-k)\sigma_2^2}{n^2}$.

Then

$$\varphi_{\hat{\theta}}(t) = \varphi_{\bar{X}}(at) = \sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}exp\Big(ita\tilde{\mu}_k - \frac{1}{2}\tilde{\sigma}_k^2 a^2 t^2\Big)$$

$$= \sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}\varphi_{Y_k}(t),$$

where
$Y_k \sim N(\mu_k, \sigma_k^2)$, $\quad \mu_k = \dfrac{k\tilde{\mu}_1 + (n-k)\tilde{\mu}_2}{\sqrt{n}\,\sigma}$, and $\sigma_k^2 = \dfrac{k\sigma_1^2 + (n-k)\sigma_2^2}{n\sigma^2}$.

Using linearity of integration once again and the fact that there is a bijection between a probability distribution function and a characteristic function, we obtain:

$$f_{\hat{\theta}}(x) = \sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}\frac{1}{\sqrt{2\pi}\sigma_k}exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right).$$

Results

As noted previously, Edgeworth expansions are not distribution functions; they are not necessarily monotonic and bounded by 0 and 1. In case of sample mean for distributions being considered here, this happens in the left tail, and therefore higher-order approximations cannot be used for this tail. Since this tail is thinner than that of a Gaussian distribution, normal approximation will be conservative and control the error rate; it is then interesting to look at the right - thicker - tail with a one-side test to get a clear picture of the actual error rates.

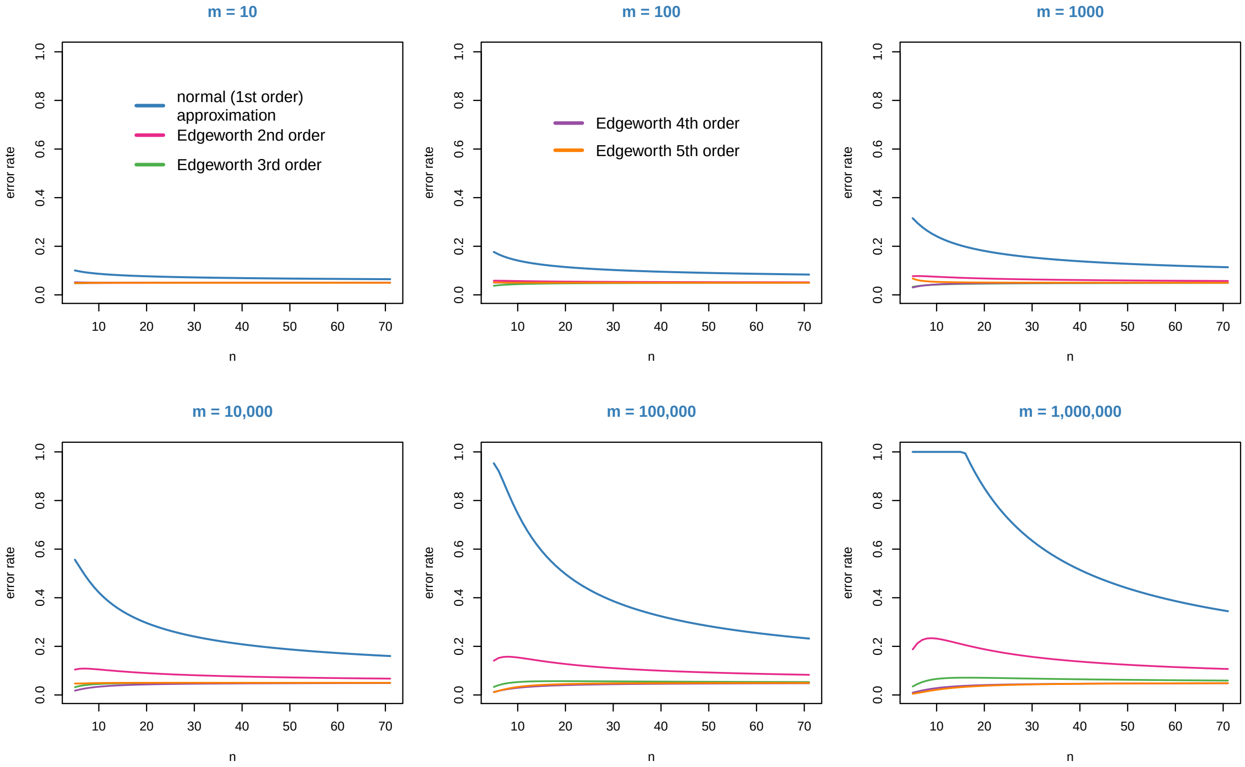


Figure 2.1: Standardized Mean, mixture normal distribution, $\alpha = 0.05$

Figure 2.1 illustrates dependence of the discrepancies between actual and nominal error rates on the sample size n for various number of tests m. Sample sizes range from 5 to 71 and numbers of tests are 10^k, $k = 1, \ldots, 6$. Distributions of X here is mixture-normal, with $\mu_d = 3$, $\sigma_1 = 0.5$, and $\sigma_2 = 2$. Actual error rates are obtained using first through fifth orders of approximations, first order (zero term EE) being a normal approximation. From these graphs, we can see that:

– as the number of tests m increases, normal (first order) approximation yields gradually increasing actual error rates, which means poor error rate control that eventually fails completely;

– as the sample size n decreases, actual error rates go up with normal approximation. For large m error rate control stays poor even as sample size grows;

– higher order approximations provide better and more stable error rate control that holds for most n and m.

Next, to take a more detailed look, we fix the sample size at $n = 10$ and plot the error rates against number of tests m, which reflect progressively more extreme critical values. For Figures 2.2 and 2.3 we use mixtrue-normal (with the same parameter values as before) and gamma ($\alpha = 10$) distributions respectively. Note that the x-axis is on the log scale and the number of tests ranges from 1 to a little over a million. Dotted line at $y = 0.05$) indicates where the error rate would need to be to match the nominal level. Normal approximation eventually fails to provide any error control; second order (term 1) delivers marked improvement and the third order already comes much closer to the nominal level. By varying parameters of the distributions, we change the shape of a sampling distribution and therefore actual error rates resulting from all the orders of approximation, but general conclusions still hold.

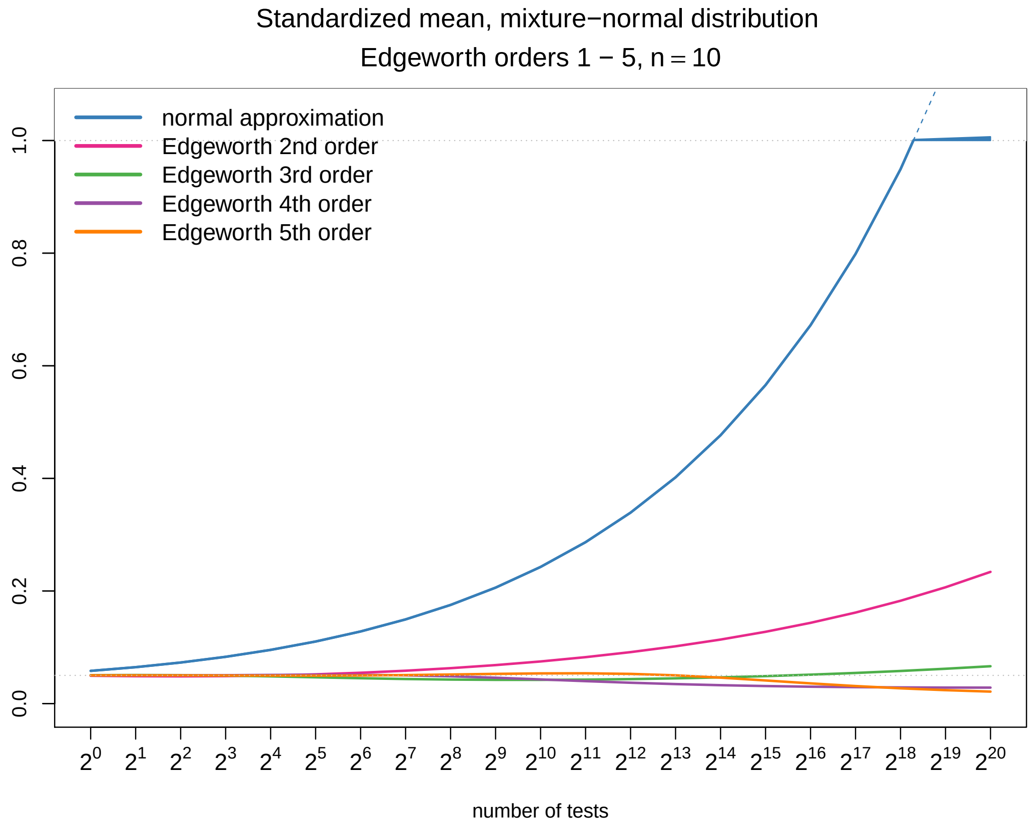


Figure 2.2: Actual vs nominal error rates: Edgeworth approximations for a standardized mean, mixture-normal distribution

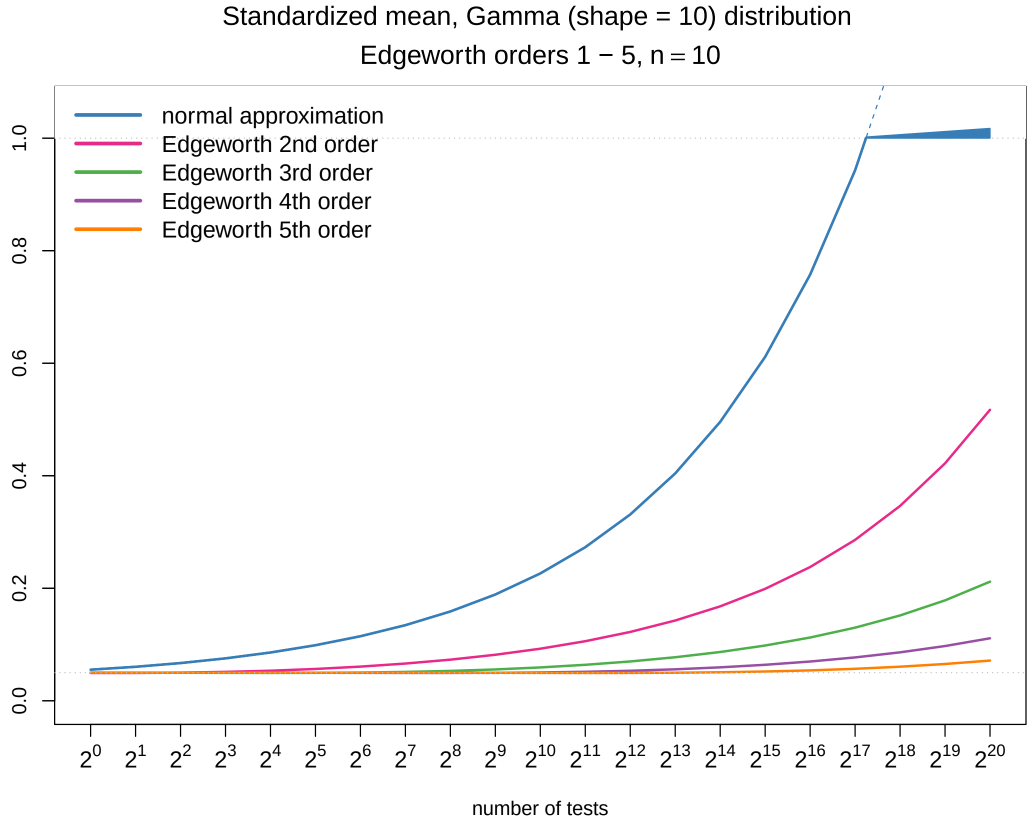


Figure 2.3: Error rates: Edgeworth approximations for a standardized mean, gamma distribution

2.2 Performance of EE for t-statistic when n is small

Turning our attention to Edgeworth expansions for studentized statistics is a next step toward data analysis - these statistics can be calculated entirely from data and are the ones generally used in practice. For this study, however, we will still use true cumulants and not their estimates for the expansions. This will serve as an intermediate step to gauge how close higher-order approximations for test statistics are before we proceed to simulations. Small sample size is the angle from which we investigate the challenges that studentized statistics pose in regard to approximations and the differences between large deviations in distributions of standardized and studentized sample means. This section will also cover small sample adjustment that we propose to use specifically for the statistics that incorporate variance estimates.

Null distribution - coarsened simulation

Apart from normal random variables, a closed form for sampling distribution of a t-statistic might not exist or be easily obtainable (as mentioned Chapter 1, dependence of sample mean and sample variance pose a challenge). Obtaining a sampling distribution that would provide reliable estimates for extremetly small probabilities through simple simulation is unfeasible as it would necessitate keeping in memory a great number of simulated test statistics (at least 10^{10} would be needed to reach the critical values in extreme tails corresponding to $m = 10^6$). Therefore we use what we call *coarsened simulation* that provides sufficient resolution for the tail quantiles (for instance, estimate $F_{k,n}^{-1}(p)$ with accuracy up to the 4th decimal place). It amounts to the following simple procedure that can use manageable amount of computer memory with coarsening the information obtained from simulations:

1. Divide support range for test statistic into regions and bins that reflect the focus of our investigation: extreme tail regions where all the observations will be saved, one large bin in the center of the distribution around zero, and the rest of the tails that are divided into reasonably fine grid (e.g. into intervals of length 10^{-4} or progressively finer division moving away from the mean).

2. With each simulation result, save it if it falls into a designated extreme tail region (a ray on the number line) - or increment the count in the corresponding bin otherwise. Then, without saving the great majority of simulations, we will know the amount of observations that fall into the center region (which can be designed to contain most of the results), the number of observations that fall into every smaller bin in the tails, and will have all the important far tail observations - the only ones that are saved, a very small proportion of the number of simulations.

This simulation scheme allows the calculation of probabilities for given values of test statistics with desired quantile accuracy - the statistic will be rounded according to the bin specification. It can also be adjusted depending on available amount of memory and focus regions.

Exploration: Normal vs Student's t-distrubution

As with standardized mean, Edgeworth expansions will be used for the thicker tail of a sampling distribution; note that for the previously considered distributions of X that have a thicker right tail, it is a left tail of the sampling distribution of t-statistic that is thicker.

As an extension of a central limit theorem, classic Edgeworth series is based on a standard normal distribution, which provides a zero term (first order approximation) as well as a factor for all the consecutive terms in the form of density. Standardized mean is scaled by a constant and therefore falls under statistics directly covered by CLT. For a studentized mean, while the limiting distribution is still normal, the finite sample situation is more complicated, including the rate of convergence. In contrast to a standardized mean, sampling distribution of a studentized mean/t-statistic is not necessarily mean-zero and unit-variance. The variance is greater than 1 and increases as the sample size decreases; sampling distribution and its variance also depend on the choice of a scaling factor for t-statistic, variety of which is outlined in Chapter 1.

Gaussian data generating distribution provides closed form distributions for both standardized and studentized means in finite sample and is a "best case scenario" - a well behaved "nice" symmetrical distribution. Therefore, comparing standardized and studentized means of a sample of normally distributed random variables will be our starting point in understanding the behavior of large deviations of a test statistic in a small sample.

Let $X \sim N(\mu, \sigma^2)$. We want to compare the distributions of $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$, which is $N(0,1)$, and $\frac{\sqrt{n}(\bar{X}-\mu)}{s}$, which is t_{n-1}, focusing on large deviations from zero for various sample sizes.

Figure 2.4 shows the difference between quantiles corresponding to tail probabilities of $0.05/m$, $m = 2^j$, where $j = 0, 1, \dots 10$ for standard normal and t-distribution with $\nu = n-1$ degrees of freedom. For comparison, we add gamma distribution with shape equal to n (that would also correspond to sum/mean of n exponentially distributed X). We can see that for $n = 10$, the difference in quantiles is dramatic; when $p = \frac{0.05}{2^{10}}$ (2^{10} would be the number of tests in multiping comparison analysis), the corresponding quantiles range from around 5.4 in standard normal to around 15.5 in t-distribution. The difference in quantiles for various sample sizes is illustrated by Figure 2.5; it can be seen that only when n gets to 50, quantiles become somewhat close. For a very small size $n = 5$, the quantile for $p = \frac{0.05}{2^{10}}$ approaches 90 (!), nowhere near that of either normal or gamma distribution. In other words, far tails of Student's t-distribution with small value for degrees of freedom are very thick and that gives us a reason to believe that this would be true for any sampling distribution of a t-statistic when the size is small.

Parameters of Student's t-distribution are functions of sample size, with degrees of freedom ν and $Var(t) = \frac{\nu}{\nu-2} \geqslant 1$ for $\nu > 2$ (asymptotically equal to 1 but quite a bit larger for small samples). That means that a zero term in Edgeworth expansion for a studentized mean will underestimate the variance; then the question is will higher-order terms of Edgeworth expansion based on standard normal distribution be able to get sufficiently close to the true (in our case Student's distribution based) quantiles in the far tails? Will they approach the increased variance of that distribution? Figure 2.6 demonstrates different order expansions for $n = 10$. It can be seen that while adding higher order terms does thicken the tails and gets the resulting quantiles farther into the tails, for a small sample size these quantiles fall short of approaching the true ones in the target areas.

Now we would like to investigate what happens with asymmetric distributions with varying tail thickness. It can be seen in Figure 2.7 a) and c) that Edgeworth expansion captures the shape of the sampling distribution by "reversing" the tails (recall from Chapter 1 that if the original distribution is skewed to the right, the sampling distribution of a t-statistic is skewed to the left). It is also clear that, as expected, expansions based on the normal

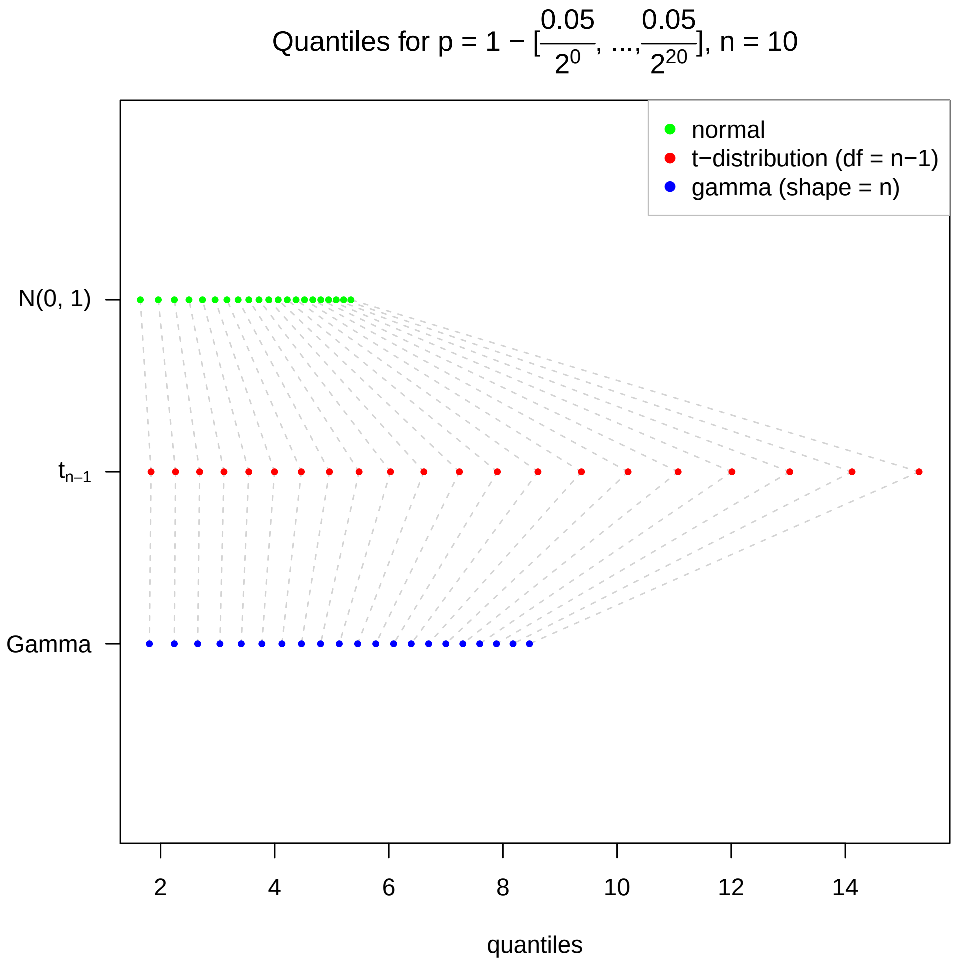


Figure 2.4: Comparison of tail quantiles for normal, gamma, and t-distributions as possible approximations for a sampling distribution of a mean-based statistic

distribution do not approach large deviations from the mean for small samples. Student's distribution approximation, on the other hand, while symmetric, is much closer to the true quantiles, especially in the thicker tails; changing the "base" of EE to t-distribution instead of standard normal seems to provide better approximation in the tails, especially with higher orders - 4-term (5th order) expansion performs especially well - Figure 2.7 b) and d). Figure 2.8 displays similar results for $\Gamma(\alpha = 3)$ distribution; sampling distribution's left tail is not as thick as with mixture-normal (which has particularly thick left tail with chosen parameters in small sample) but the general trend is the same. Finally, we go back to original normal distribution and look at the normal and t-based expansions to see how close they are to true sampling distribution. Note that in this case $\lambda_j = 0,\ j = 3, \ldots$ and the reason even terms deviate from first order is the presence of irregular terms inside $q_2(x), q_4(x)$, etc, that do not

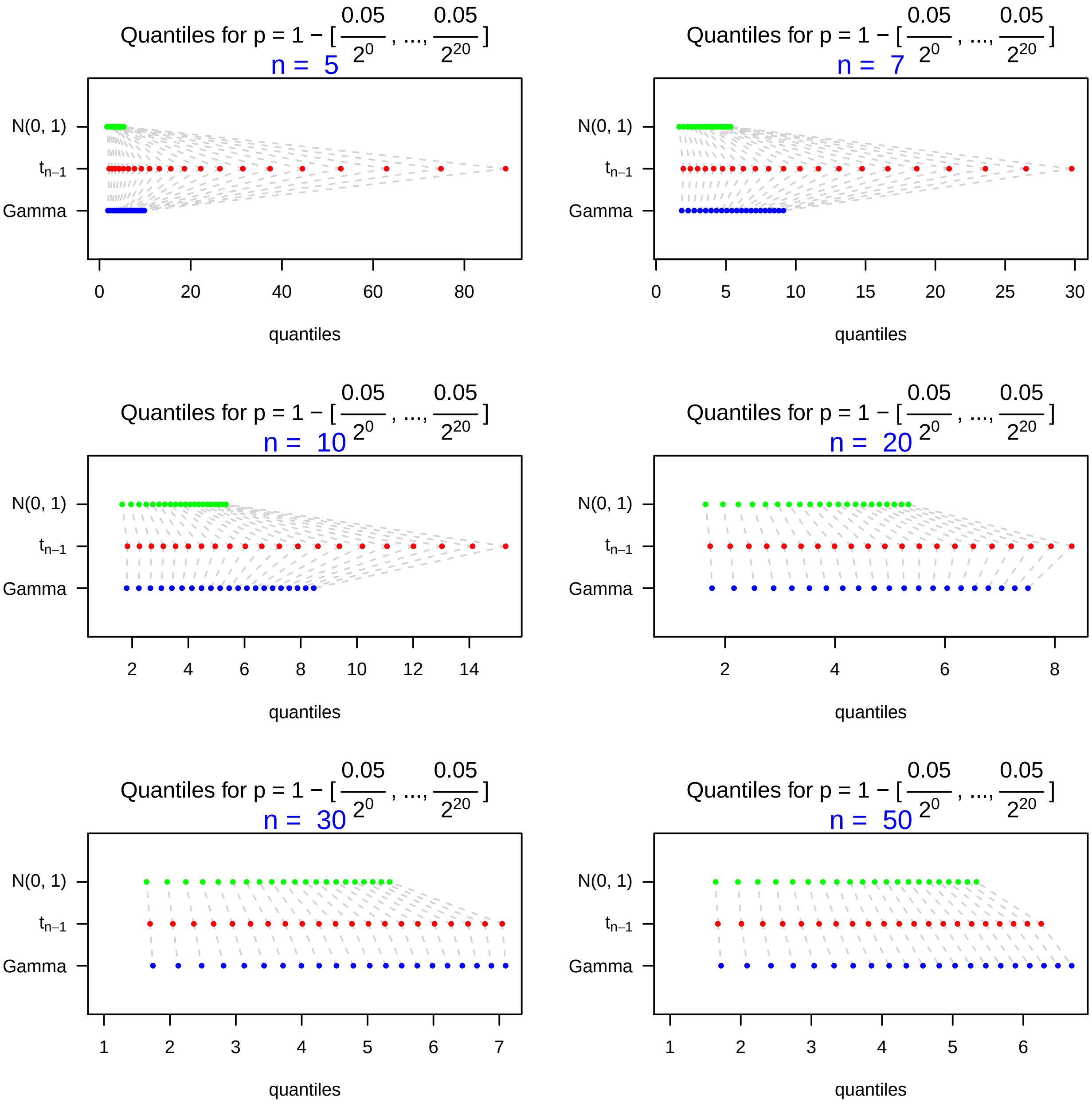
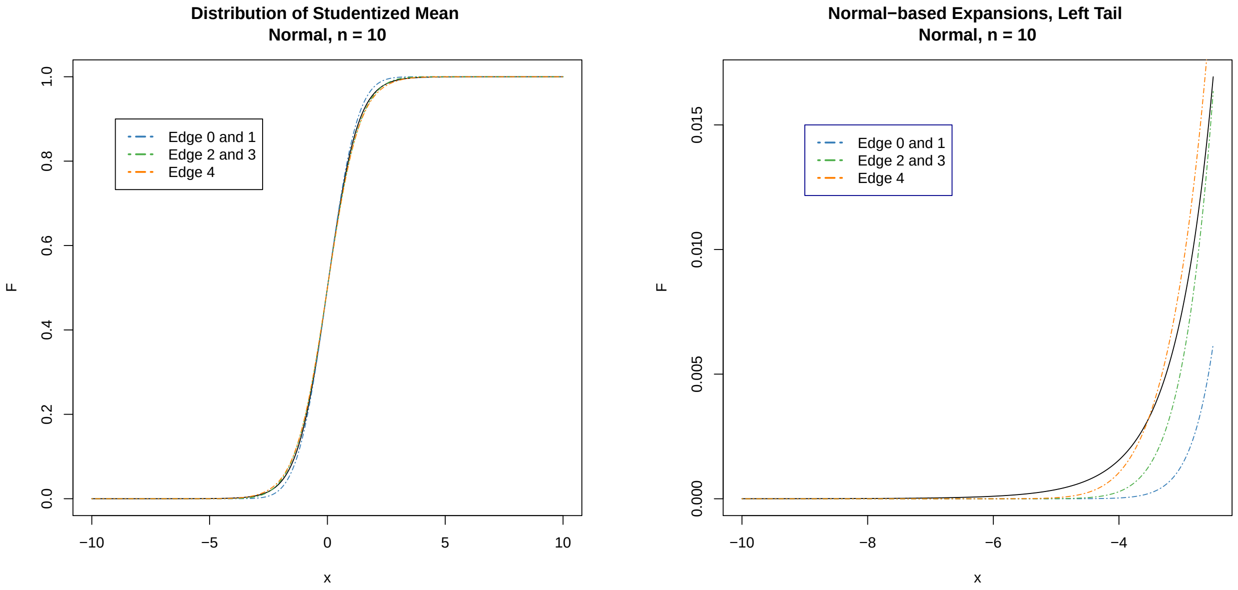


Figure 2.5: Tail quantiles for different sample sizes: normal and t-distribution approximations and gamma distribution

have cumulant factors (see Chapter 1). It appears that t-based Edgeworth approximations somewhat distort the general shape of the distribution close to the center but provide much better approximation for far tails.



(a) cdf

(b) Left tail zoomed in

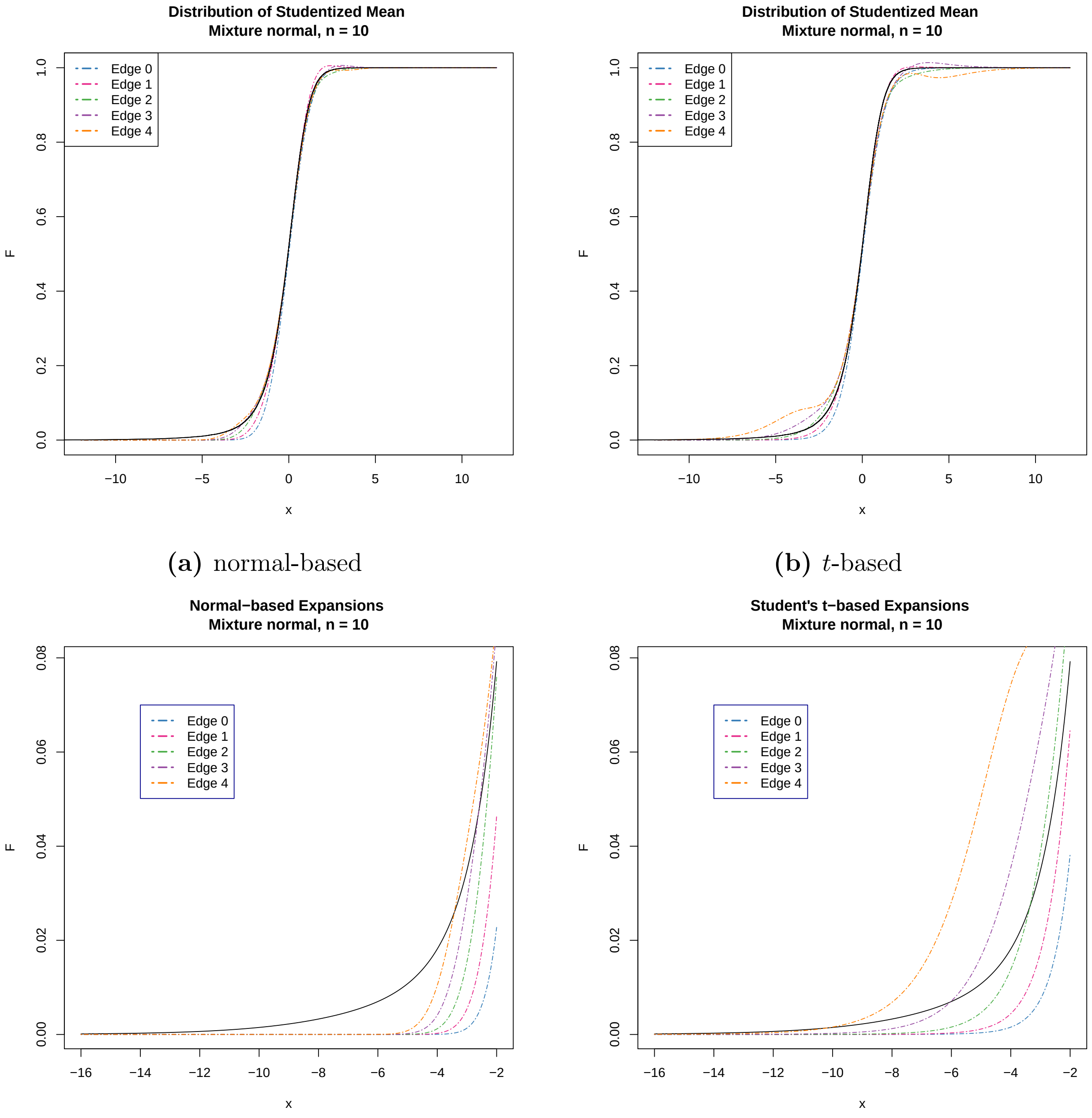
Figure 2.6: Sampling distribution of a studentized mean for normally distributed random variables: Student's t-distribution (truth) vs normal-based Edgeworth approximations

Small sample adjustment

We propose a small sample adjustment that takes the Edgeworth expansion framework but uses Student's t-distribution as a base. While there is no rigorous theoretical support for this procedure, some theoretical arguments can be made to justify the adjustment - for example, Zholud [73] argues that distribution of t-statistic for large deviations and small sample size can be approximated by $K\, t_{n-1}$, where K is a constant dependent on the original distribution. We also provide heuristic justification for t-distribution-based expansions, aiming at the increased inference reliability in practice when the sample size is small. The adjustment procedure is as follows (note that asymptotically it converges to a traditional normal-based expansions):

- Substitute Student's t_{n-1} c.d.f. for $\Phi(\cdot)$ as a zero term in the expansion.
- Substitute Student's t_{d_j} p.d.f. for $\phi(\cdot)$ in subsequent terms. Adjust degrees of freedom d_j in these higher-order terms j in a controlled way to guarantee monotonicity and convergence to 0 and 1 in the far tails. To do that, set $d_1 = d_0 + 2$ and $d_j = d_{j-1} + 3$, $j = 2, 3, \ldots$ based on the maximal power of quantile x in higher-order terms and the increase of this power by 3 in each consecutive term (Chapter 1).

This adjustment aims to provide closer approximation to the features of the true distribution that are most relevant for inference in multiple comparison analysis - critical values



(a) normal-based

(b) *t*-based

(c) normal-based, zoomed in

(d) *t*-based, zoomed in

Figure 2.7: Normal and t-based expansions for mixture normal distribution

located in the distal tails. Justification for the adjustment follows considerations presented in 2.2 and can be summarized in a few points:

- If distribution of X is skewed to the right, then sampling distribution of a studentized

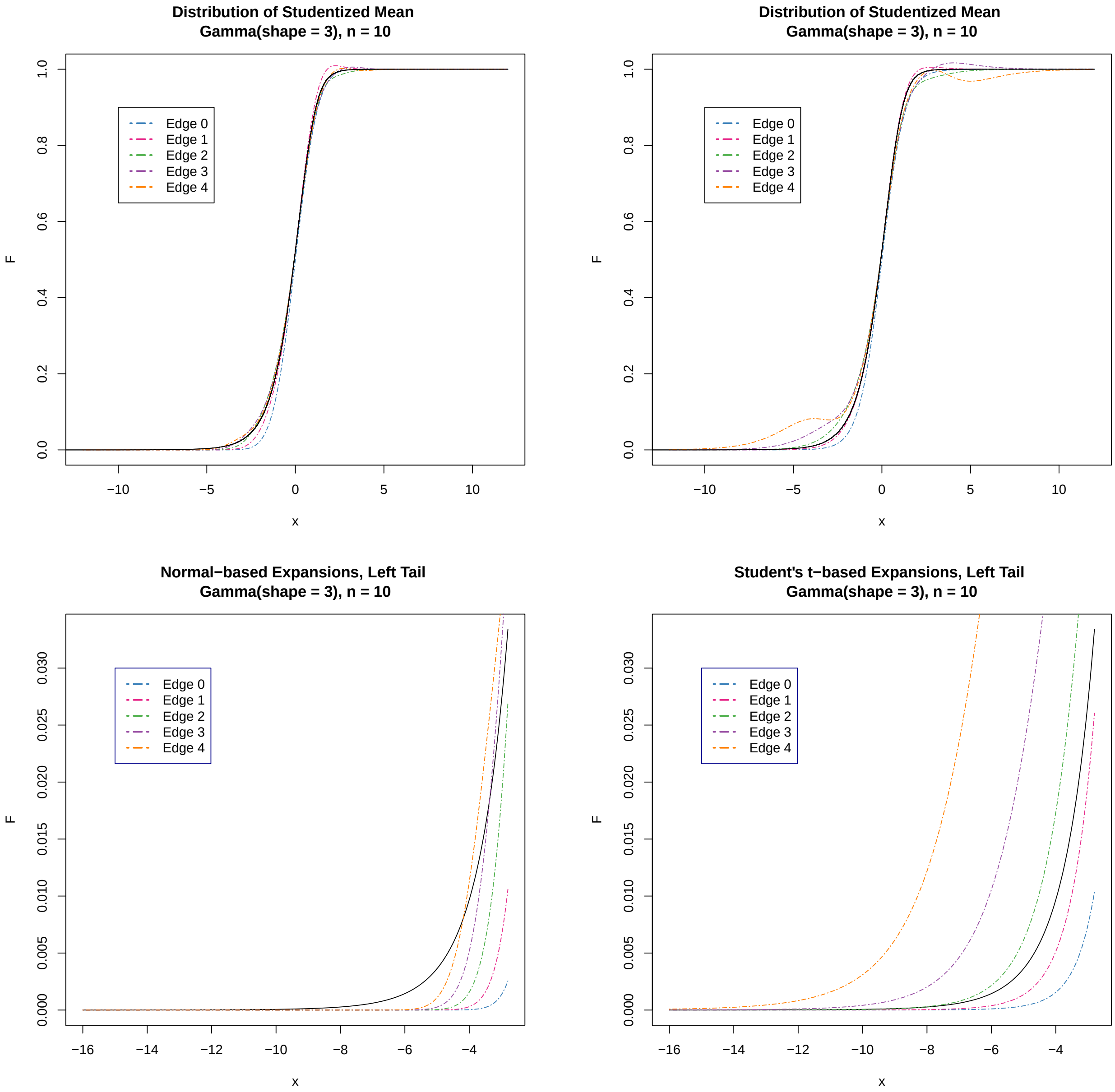


Figure 2.8: Normal and t-based expansions for Gamma distribution

mean is skewed to the left (unlike distribution of $\bar{X}$) due to the dependence of $\bar{X}$ and s^2. Edgeworth expansions go after the right shape of this distribution.

– The difference between standard normal and true sampling distribution of a studentized mean is great in small samples, especially in the tails - as illustrated by our “best case scenario” model with original normal distribution, where the sampling distribution is exactly t. Gaussian-based expansions do not reach targeted quantiles of the sampling

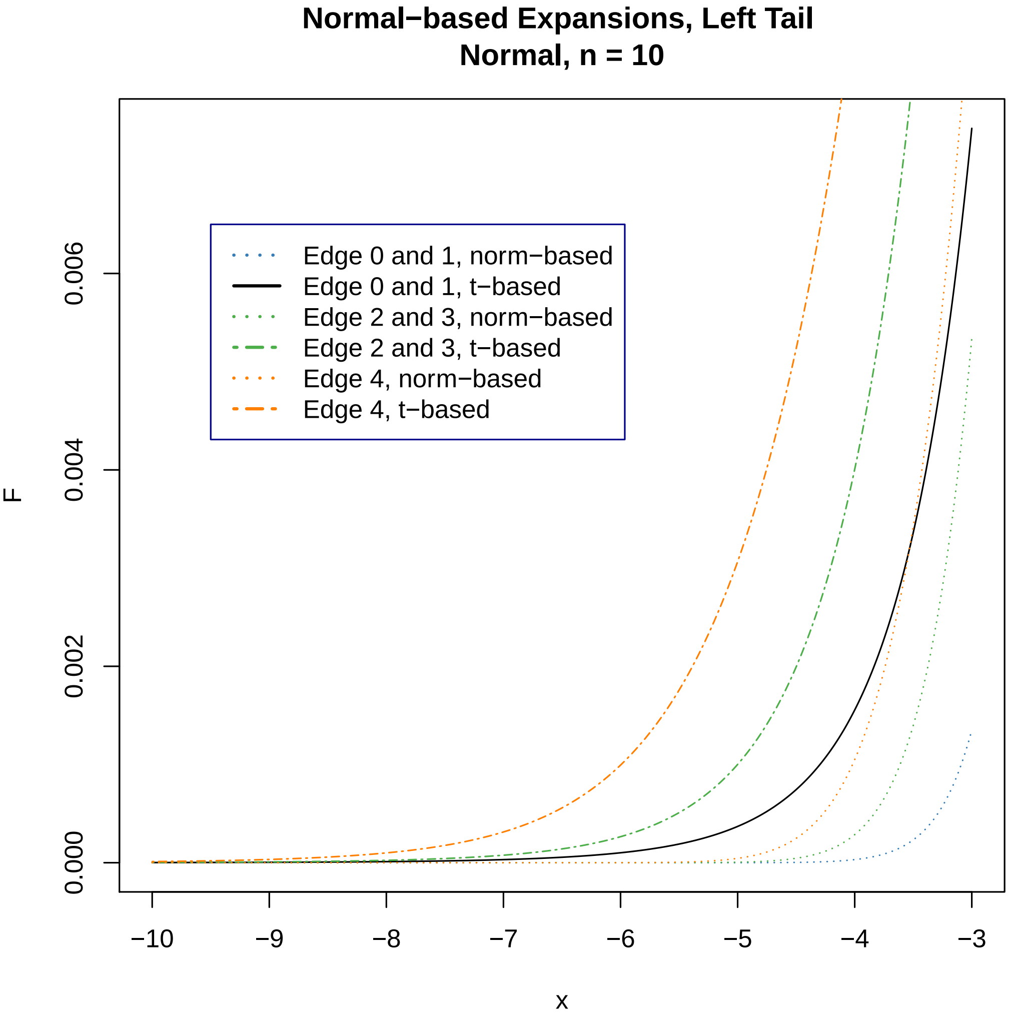


Figure 2.9: Normal and t-based expansions for normal distribution

distribution. In addition, recall that for unbiased estimates of variance in studentized mean, the "base" distribution is less spread out than standard normal or t_{n-1} due to the adjustment r^2 (Chapter 1).

– t-distribution is much closer to the truth; asymptotically it is equivalent to standard normal and provides the best available tool to get in the ballpark of the desired quantiles (without being too conservative) thus offering a valid starting point as an analog of the Gaussian approximation for studentized statistics in finite samples.

– By combining Student's t-distribution with Edgeworth expansions we are able to get the correct shape of the distribution (but of course no better than the sample itself) in the desired regions of the tails. Note that it is specifically targeted for the large deviations: quantiles around the mean are well approximated by the normal-based expansions and it is not recommended to use t-based expansions for them.

Results

Similarly to what we have seen in section 2.1, truncated Edgeworth expansions for the region of a thinner tail are non-monotonic and not bounded by 0 and 1 - and therefore cannot be used for probability estimation. For studentized statistics, thickness of tails is reversed and so for our considered distributions the right tail is the thinner one. Demonstrated results use true moments of distribution of X and small sample adjustment.

Figure 2.10 displays actual error rates for the sample size 10 and the number of tests $m = 1, 2, 4, \ldots, 1048576$ with $\Gamma(\alpha = 3)$ data generating distribution. The blue line gives the rates for Student's t-distribution approximation that is customarily used in data analysis (at this sample size, normal approximation will not be anywhere near the truth at the far tails). It can be seen that for the values of m that are over a thousand, there is virtually no error rate control; while the rates are truncated at 1, the thickness of that truncation line at the highest values of m indicates that it is indeed "worse" than no control (thickness of the line reflects visual projection of the values greater than 1). Starting with second-order approximations, Edgeworth expansions show markedly improved results, and forth and fifth orders are well below the nominal line ($y = 0.05$), which gives hope for a more reliable inference that could be achieved with incorporating higher empirical moments into data analysis.

Figure 2.11 shows results for log-normal($\sigma^2 = 0.4$) distribution and Figure 2.12 - again for $\Gamma(\alpha = 3)$ but with moderated t-statistic (for hyperparameters, we use preselected fixed values for d_0 and s_0^2). Note that for moderated t-statistic, degrees of freedom for Student's t-distribution are augmented and so all the terms have been changed accordingly, with the value of prior degrees of freedom d_0 added to degrees of freedom in first order (t_{d_0+n-1} c.d.f.) and all the consecutive terms.

A few observations relevant to the data analysis procedure going forward:

- for some distributions, higher-order expansions even for the thicker tail are not monotonic, though still inside the bounds - e.g. fifth-order expansion for $\Gamma(\alpha = 3)$;
- 4th and 5th orders in considered cases are quite conservative.

2.3 Empirical EE and data analysis

Expansions for various studentized statistics in Chapter 1 were developed with the goal of using them in practice; we can apply them with a small sample adjustment directly to data analysis. In previous section, we looked at the statistics that assume that the variance is

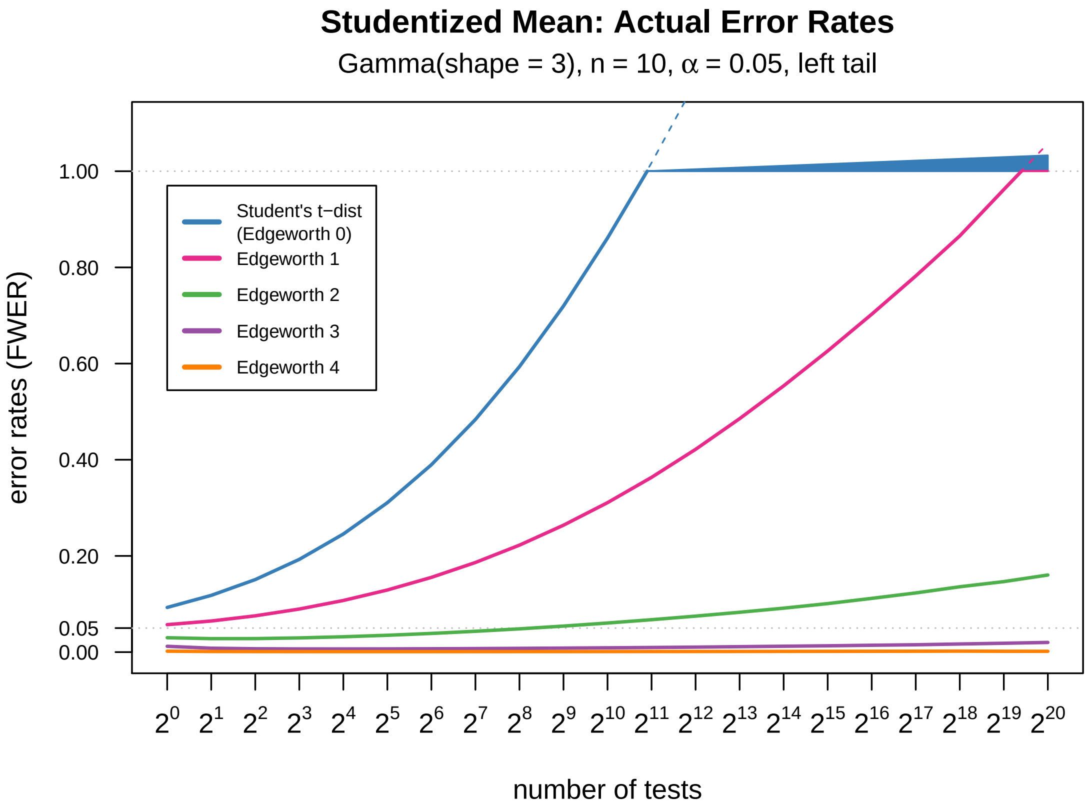


Figure 2.10: Error rates for studentized mean, gamma distribution, $\alpha = 0.05$

not known as they incorporate its estimate; however, in Edgeworth expansions for those statistics we still used true cumulants. To use these expansions in data analysis, moments and/or cumulants of the original distribution need to be estimated and substituted for the true ones. There are various ways to do that and different estimates can be used - for example, P. Hall [32] suggests replacing standardized moments that can be used to calculate cumulants with the following estimates:

$$\hat{\mu}_j = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^j}{\hat{\sigma}^j},$$

calling the result empiric Edgeworth approximations (note that these quantities are also studentized).

With application to data analysis, there are two issues that need to be considered: smaller sample size makes for greater variability of the moment estimators; we can also expect higher moments to be more variable and thus not necessarily very useful for approximation, limiting the number of usable terms in Edgeworth expansions. The second issue is that, as was

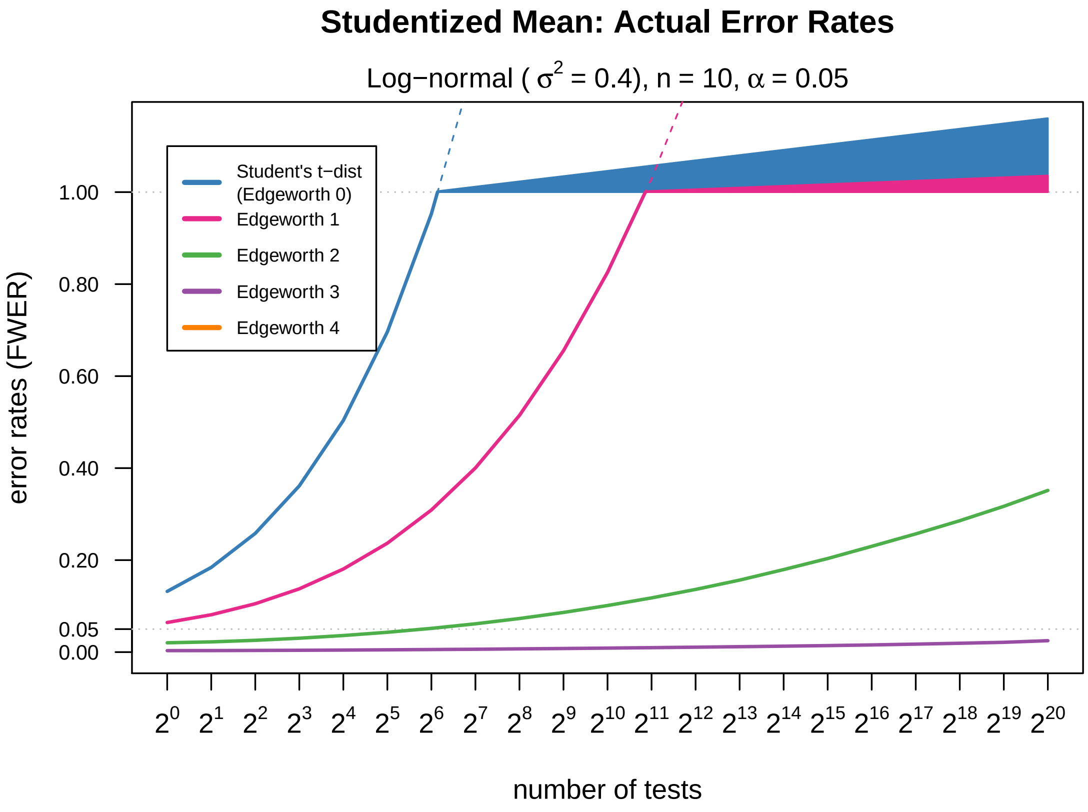


Figure 2.11: Error rates for studentized mean, log-normal distribution, $\alpha = 0.05$

mentioned earlier, we can use Edgeworth expansions for thicker tails only, leaving normal or Student's approximation for the thinner ones. In practice we don't know the data generating distribution, and therefore it is unknown if the distribution is symmetric and which tails are thicker or thinner than that of a Gaussian distribution (our point of reference in regard to expansions). This characteristic is going to be inferred from the information available - the sample. Before using Edgeworth expansions, the method performs "tail diagnostic", which itself relies on the series, exploring behavior of the tails for each higher-order approximation. Monotonicity and boundedness between $(0, 1)$ will inform on usability and tail thickness and that in turn will decide if and where higher-order approximation should be used.

It should be noted that the sample of a very limited size might not be representative of an original distribution and no method can glean the information beyond the sample. The proposed method aims to learn more about the underlying distribution but it can only do as well as the sample itself. By using Edgeworth approximations on supposedly thicker tails and Student's t-distribution on the thinner ones, we ensure that there is no loss in error rate control compared to standard asymptotic methods, while possibly gaining reliability with

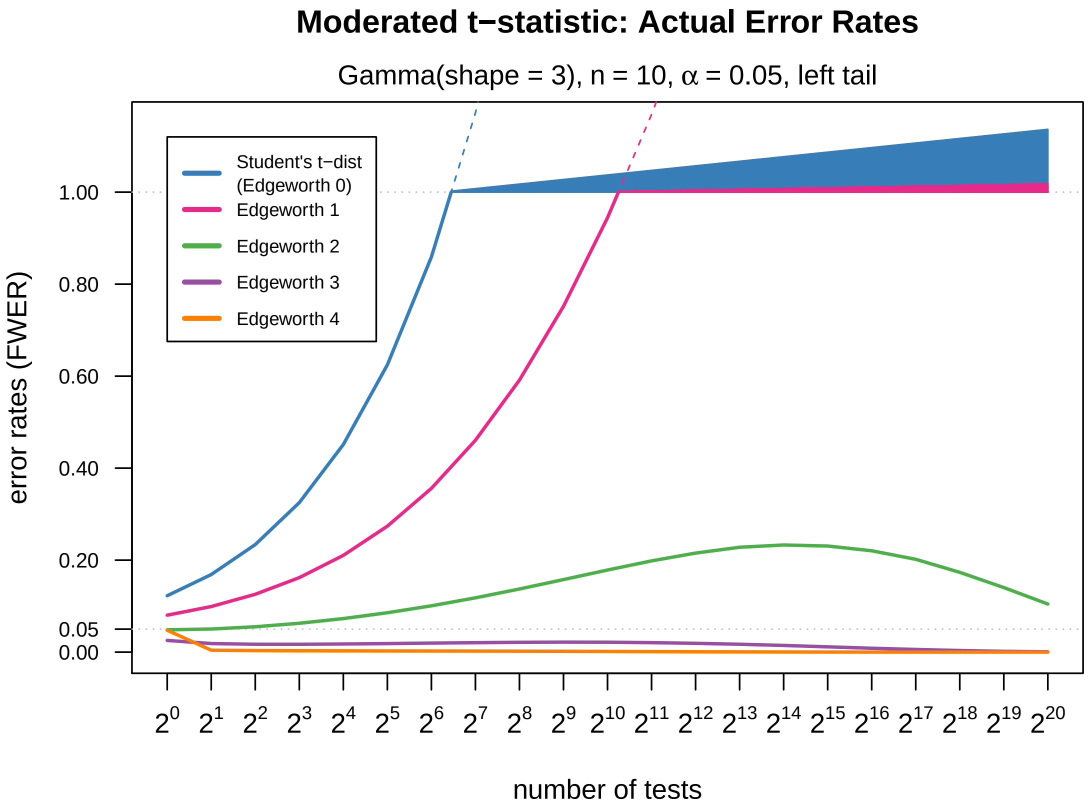


Figure 2.12: Error rates for a moderated t-statistic, gamma data generating distribution

higher-order approximations.

Substitution - moment estimates

Edgeworth expansions use central moments or cumulants of data generating distributions, which are usally unknown in practice, so they need to be estimated. As discussed earlier, for a small sample size the difference between simple biased and unbiased estimates should be considered as it might affect the results. This difference is bigger for higher moments; however, in asymptotic expansions it is offset by the smaller weight/contribution of higher order terms where they are used. For moment estimates, we adopt notation M_k for unbiased central moments, $E(M_k) = \mu_k$, and $m_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$ for simple biased estimates.

Expressions for expansion terms include not just "standalone" moments but their powers and products, as well as ratios of their combinations (denominators mainly consist of powers of variance); use $M(\cdot)$ to denote an unbiased estimate of an expression inside the parentheses, e.g. $E\left[M(\mu_3^2)\right] = \mu_3^2$. While unbiased estimates of products and integer powers of moments are possible to obtain, that is not the case with ratios and roots. Of course, such biased estimates, like square root of sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, are widely used in practice and it remains a question of how much "unbiasedness" is reasonable to use in more complex situations. This question warrants a separate investigation, and the answers probably depend on many factors - distributions, statistics, sample size, etc. Adding to the complexity is the fact that since unbiased estimate of the ratio cannot be obtained, simplifying expressions should also be questioned - consider, for example, scaled sixth cumulant:

$$\lambda_6 = \frac{\kappa_6}{\mu_2^3} = \frac{\mu_6 - 15\mu_2\mu_4 - 10\mu_3^2 + 30\mu_2^3}{\mu_2^3} = \frac{\mu_6}{\mu_2^3} - 15\,\frac{\mu_4}{\mu_2^2} - 10\,\frac{\mu_3^2}{\mu_2^3} + 30$$

For a closest estimate, it is natural to consider the ratio of an unbiased cumulant estimate $M(\kappa_6)$ and an unbiased scaling factor $M(\mu_2^3)$. Then, is $\dfrac{M(\mu_2\mu_4)}{M(\mu_2^3)}$ preferrable to $\dfrac{M(\mu_4)}{M(\mu_2^2)}$ for the second term? This issue becomes even more complicated when we consider long expressions for Edgeworth expansions with their various algebraic representations and simplifications.

In addition to these questions, EE for generalized t-statistics (see Chapter 1) are set up in such a way that even if we get unbiased estimates for $k_{j,i}$'s, their subsequent algebraic manipulations include multiplication and powers, so estimates for the whole expressions in numerators become biased. Trying to get the "maximum" amount of unbiasedness might be computationally unfeasible; besides, all the simplifications affect the results, so in the end it is still an arbitrary decision.

Extending the method we used to derive general order Edgeworth expansions for t-statistics, we can obtain unbiased estimates for arbitrary moments, as well as their products and integer powers using our software. Unlike the second central moment (variance) and the third, where biased moment estimates are scaled by constants to obtain unbiased ones, subsequent moment estimates contain combinations of lower moments. Below are the unbiased estimates of the moments and their combinations up to sixth order (see [19] and [59] for unbiased estimates up to fourth order).

$$M_3 = \frac{n^2}{(n-1)(n-2)}\,m_3$$

$$M_4 = -\frac{3\,n\,(2\,n-3)}{(n-1)(n-2)(n-3)}\,m_2^2 + \frac{n\,(n^2-2\,n+3)}{(n-1)(n-2)(n-3)}\,m_4$$

$$M(\mu_2^2) = \frac{n\,(n^2-3\,n+3)}{(n-1)(n-2)(n-3)}\,m_2^2 - \frac{n}{(n-2)(n-3)}\,m_4$$

$$M_5 = -\frac{10\,n^2}{(n-1)(n-3)(n-4)}\,m_2 m_3 + \frac{n^2\,(n^2-5\,n+10)}{(n-1)(n-2)(n-3)(n-4)}\,m_5$$

$$M(\mu_2\mu_3) = \frac{n^2\,(n^2-2\,n+2)}{(n-1)(n-2)(n-3)(n-4)}\,m_2 m_3 - \frac{n^2}{(n-2)(n-3)(n-4)}\,m_5$$

$$\begin{aligned} M_6 =& \frac{15\,n^2\,(3\,n-10)}{(n-1)(n-2)(n-3)(n-4)(n-5)} m_2^3 - \frac{40\,n\,(n^2-6\,n+10)}{(n-1)(n-2)(n-3)(n-4)(n-5)} m_3^2 \\ &- \frac{15\,n\,(n^3-8\,n^2+29\,n-40)}{(n-1)(n-2)(n-3)(n-4)(n-5)} m_2 m_4 + \frac{n\,(n^4-9\,n^3+31\,n^2-39\,n+40)}{(n-1)(n-2)(n-3)(n-4)(n-5)} m_6 \end{aligned}$$

$$\begin{aligned} M(\mu_3^2) =& -\frac{3\,n^2\,(3\,n^2-15\,n+20)}{(n-1)(n-2)(n-3)(n-4)(n-5)} m_2^3 + \frac{n\,(n^4-8\,n^3+25\,n^2-10\,n-40)}{(n-1)(n-2)(n-3)(n-4)(n-5)} m_3^2 \\ &+ \frac{3\,n\,(2\,n^3-5\,n^2-5\,n+20)}{(n-1)(n-2)(n-3)(n-4)(n-5)} m_2 m_4 - \frac{n\,(n^2-n+4)}{(n-2)(n-3)(n-4)(n-5)} m_6 \end{aligned}$$

$$\begin{aligned} M(\mu_2^3) =& \frac{n^2\,(n^2-7\,n+15)}{(n-1)(n-3)(n-4)(n-5)} m_2^3 - \frac{3\,n\,(n^2-5\,n+10)}{(n-1)(n-3)(n-4)(n-5)} m_2 m_4 \\ &+ \frac{2\,n}{(n-3)(n-4)(n-5)} m_6 - \frac{2\,n\,(3\,n^2-15\,n+20)}{(n-1)(n-2)(n-3)(n-4)(n-5)} m_3^2 \end{aligned}$$

$$\begin{aligned} M(\mu_2\mu_4) =& -\frac{3\,n^2\,(2\,n-5)}{(n-1)(n-3)(n-4)(n-5)} m_2^3 + \frac{4\,n\,(n^2-5\,n+10)}{(n-1)(n-3)(n-4)(n-5)} m_3^2 \\ &- \frac{n\,(n^2-3\,n+8)}{(n-2)(n-3)(n-4)(n-5)} m_6 + \frac{n\,(n^4-9\,n^3+53\,n^2-135\,n+120)}{(n-1)(n-2)(n-3)(n-4)(n-5)} m_2 m_4 \end{aligned}$$

An important consideration that should factor into a decision of which estimators to use should be variability of the denominator in algebraic expressions that come up in Edgeworth expansions (and as a general rule). An example above for possible estimates for λ_6 provides a simple illustration. The cumulant is scaled by μ_2^3; to substitute this unknown quantity, a variety of estimators can be used: m_2^3, $[M(\mu_2)]^3$, or $M(\mu_2^3)$, to name a few. While expression for $M(\mu_2)$ (and thus its cube) contains m_2 only, the expression for $M(\mu_2^3)$ includes m_4 and m_6 as well. These higher-order quantities may be highly variable, especially in the small sample and therefore the whole ratio becomes highly sensitive to the small values of estimates in the denominator that can inflate λ_6 dramatically, increasing variability of the ratio to the point of unusability. Our numeric exploration of stability of these complicated algebraic expressions indicates that for relatively small sample sizes it might be indeed preferable to use lower-order estimators in place of parameters in denominators ("power of mean" instead of "mean of power"). In fact, even simple biased moment estimates might perform better overall (have smaller mean squared errors) because of their relative stability.

Another approach, as mentioned earlier, could be to use estimates for standardized moments, but as the estimates are themselves studentized, there is no way to make them unbiased (however unbiased numerators and unbiased denominators can be used in the same way as described above, which brings up the same challenges).

Tail diagnostic - exploration

Tail behavior of truncated Edgeworth expansions can be informative - not just indicating if a certain order of approximation can be used, but also helping to infer the degree of asymmetry and thickness of the tails of a sampling distribution. Given a set of (estimated) moments or cumulants, we can look at the resulting higher-order approximations and assess departures from general probability function characteristics, such as monotonicity and containment inside $[0, 1]$ bounds. Figure 2.13 gives an example of the thin right tail of a t-statistic (for $\Gamma(\alpha = 3)$ distribution) and its t-based approximations. It can be seen that 2-term expansion (3rd order) is "well-behaved", 4-term is not monotonic but respects the bounds and 1- and 3-term expansions are both non-monotonic and out of bounds. For the purposes of this exploration, we assign 1 to the order of expansion if the characteristic, such as monotonicity, matches that of a probability function and 0 otherwise, separately for each tail - e.g. for $\Gamma(\alpha = 3)$ it will be "1 1 1 0   0 1 0 0" (terms 1 - 4).

To look at the tail behavior and its effect on error rates in empirical Edgeworth expansions, we draw multiple samples from a variety of distributions, look at the patterns and different combinations of tail diagnostic results, and summarize error rates within each group. Each simulated sample provides a set of moments and cumulants that yields a diagnostic combination for higher orders (like the one above for gamma distribution); to get the combination, we divide a tail region into small intervals and evaluate monotonicity and bounds. Actual error rates are calculated using the same steps as before (described in section 2.1), and then for each order of approximation the summary is produced. Since there are many very small values for probability that are important for our investigation, we calculate mean and variance of error rates on a log scale, thus preventing this information from being obscured by large values. In the figures presented, one standard deviation above and one below the mean are marked by shaded areas.

In Supplementary materials 1, we present the results in interactive graphics, where the user can choose from different orignal distributions and their parameters, and look at the figure for each combination while adjusting y-scale to zoom in and out. Distributions covered

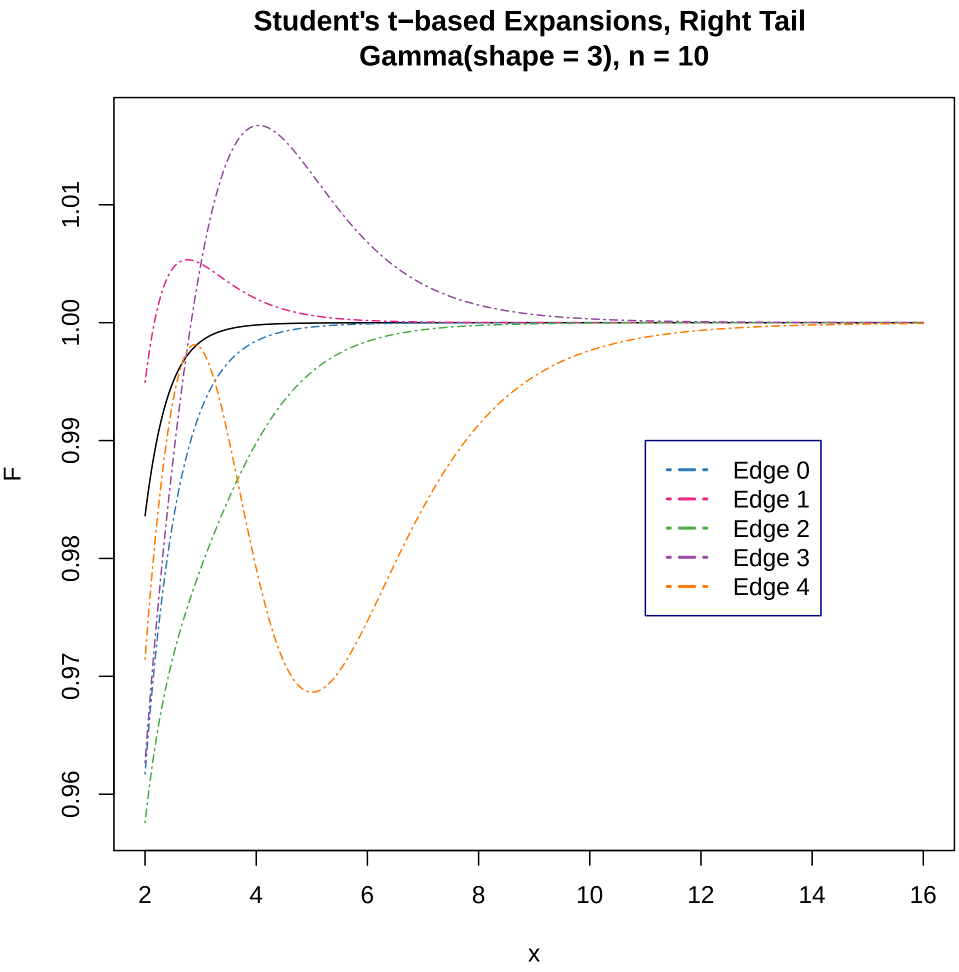


Figure 2.13: Edgeworth expansions: thin tail behavior, gamma distribution

are gamma with shapes $\alpha = 3$ and $\alpha = 50$, log-normal with $\sigma^2 = 0.4$ and $\sigma^2 = 0.5$, mixture-normal with $p = 0.5$, $\mu_d = 3$, $\sigma_1^2 = 0.5$, $\sigma_2^2 = 2$, and standard normal. Here we present some of these plots as examples; note that the order of values on x-axis for left tail is reversed to provide a visual analogy to tail quantiles spreading out from the mean. In Figures 2.14 and 2.15 there are several combinations based on $\Gamma(\alpha = 3)$ distribution, which is skewed but still fairly nice and unimodal. Combinations a) and b) in figure 2.14 represent the most common groups, with thick left and thin right tails. In a) the tails are well defined and error rates resulting from higher-order approximations provide great improvement and are not spread out, while combination in b) indicates less contrast between the tails with higher error rates and more variability in them. Tails in combination c) are both well-behaved - and we see that error rates for 3-term expansion (and to a lesser degree 1-term) vary a lot and in many cases might even be less conservative than the first order (zero term). Combination in Figure 2.15 indicates that the sample is not representative, and 1-term expansion is very anti-conservative and highly variable.

Figures 2.16 and 2.17 show some of combinations for $\Gamma(\alpha = 50)$ (near symmetrical) and normal distributions; they illustrate the point that expansions that are anti-conservative ("trying" to make the tails thinner) might also produce error rates that vary dramatically.

In a different simulation study, we compare expansions for ordinary and moderated t-statistics by looking at three characteristics separately for each simulated sample - the characteristics being monotonicity, respecting $[0, 1]$ bounds, and being no less conservative than first order approximation at every point in the tail region. For these simulations, we use randomly chosen distributions and group the samples by unique combinations of the three conditions. Observations and conclusions from these tail diagnostic investigations are summarized in the next section along with the data analysis procedure that was devised based on those observations.

Proposed procedure

When the function provided by truncated EE is not monotonic and especially not bounded by 0 and 1, it cannot be used to provide an approximation to the distribution of interest. In general, non-bounded function is always non-monotonic, but in some cases the function can be bounded but not monotonic. In these cases, some measures could be taken to smooth out non-monotonic regions of the function but this would involve arbitrary decisions about smoothing methods and tuning parameters - and in the end might result in unreliable probability estimates (in our simulations, this approach did not yield very interesting results). As it happens, great majority of samples have one well-behaved (monotonic and bounded) tail and one that is unusable. In a small fraction of cases, both tails are well-behaved; the situation where both tails are not "nice" has never been observed in our investigation.

It turns out that a well-behaved (monotonic and bounded) function for certain orders (like 4th order, or 3-term expansion) indicates a thick tail, while the opposite is true for the thin one. This was determined by evaluating tail behavior of EE for known data generating distributions, using true moments and comparing EE with the distribution of the test statistic (obtained through high-volume coarsened simulation described in section 2.2). Since we are aiming to protect against false positives by providing closer approximation in situations where first-order approximation is anti-conservative - where a tail of test statistic's distribution is "thicker" than normal or Student's t-distribution - the expansion would deliver better approximations exactly where they are needed. As for the thinner tail, first order approximation can be used, as usual. To take it a step further, we will use "zero-term" approximation whenever a higher-order expansion yields a tail that is thinner than normal/t distribution at any point, thus guaranteeing that our method's results are **never less con-**

servative than those of a traditional first order approximations.

EE "tail behavior" might slightly differ for various test statistics (e.g. for ordinary and moderated t-statistics) but there are common trends and features that allow us to generalize and propose a rule for "tail diagnostic" that would determine if a specific higher order expansion should be used for each tail. We describe some of these trends below, followed by the proposed rule/algorithm that is used in our data analysis method.

1. Terms that are indicative of a thin tail are odd terms - like 1 and 3. Second term is usually "nice" and thus not useful as a diagnostic tool; moreover, even in a "thin" tail it can be quite conservative, which makes it not reasonable to use in that situation. Subsequent even terms are not good indicators either, but as the order increases, they are more likely to have characteristics that distinguish them from distribution-like functions - non-monotonicity and falling beyond $[0, 1]$ interval, even in the thicker tail (e.g. 4'th term for ordinary t-statistic's left tail in $\Gamma(\alpha = 3)$ distribution).

2. In the majority of cases, term 1 already points at the thin tail; however, in other cases that term is well-behaved and term 3 is the one that determines that tail. In these latter cases, 1-term expansion is still less conservative than normal/t approximation. Therefore, checking behavior of the term 1 only is not always sufficient to determine if a higher-order approximation should be used.

3. Sometimes (relatively rare), both tails are well-behaved in higher-order expansion functions (apparently when sample points to a symmetric or near-symmetric distribution); however, in one of the tails, some higher orders (odd ones) are still less conservative than first-order approximation, so we choose to use regular normal/t approximation for this tail. And very rarely (in moderated statistics only) both tails are less conservative. Why should not higher order approximations be used in cases when they are "nice" but less conservative? Apart from intention to guarantee "never less conservative" results, our investigation (section 2.3) indicates that these cases are unstable and fairly volatile - the terms that try to approximate thin tail vary greatly in their approximations and resulting error rates from sample to sample, undermining the goal of reliability/stability (also keeping in mind that small samples might not be representative of the true data generating distribution - as some examples illustrate).

4. For thick tails, not all the terms are necessarily "nice", as was mentioned above - it is usually monotonicity that is affected. Once a term strays from distribution-like behavior, the subsequent terms usually follow the same pattern and cannot be used either. This rarely applies to terms lower than 4'th (fifth order) but in some cases 3-term expansions are already unusable. In other rare cases, the higher-order function in a thicker tail is well-behaved but less conservative than first order; it would be reasonable to assume that this points to a near-normal or at least near-symmetric distribution.

5. Comparing expansions for the distributions of ordinary and moderated t-statistics for the same samples, we observed that general tail diagnostic results are mostly similar, though occasionally "thick" and "thin" tails switch (another sign of near-symmetric distribution). Expansions for moderated t are sometimes "nicer" for higher orders than those for ordinary t-statistic; not surprisingly, there are more less-conservative cases for moderated t as well - including very rare cases where all but 1-term expansions are less conservative and even those where comparison with first order indicates that both tails are "thin" (this last situation has never been observed in ordinary t-statistic expansions).

Based on described investigation, our tail diagnostic procedure follows these simple rules: each tail is checked for monotonicity, boundedness by 0 and 1, and being less conservative than first order approximation. If at least one of these conditions is not satisfied, this order expansion is not used; none of the subsequent orders for this tail are used either. Here are the steps in which this procedure is implemented:

1. Preliminary steps:
   - non-sample specific: each tail is divided into intervals by points at which EE functions will be evaluated; first-order approximation at these points is calculated;
   - sample specific: calculation of sample moments and quantities that do not depend on a quantile x, such as λ_j and $k_{j,i}$ (see Chapter 1).
2. For each tail separately: starting from center-most quantiles and going outward, check quantiles for three conditions using increasingly higher order expansions. If conditions are not satisfied for a specific order, mark this order and all the higher ones as unusable (see possible representation below):

```
/* left tail */
nice[ ] = True; // (for all orders being considered)
for each point x going outward:
  for each increasing order k = 2, ... (term 1 and higher):
    if nice[k]:
      calculate F_k(x);
      if non-monotonic or non-bounded or less conservative:
        nice[k and higher] = False;
        if k = 2:
          stop; // (break from the outer loop)
/* right tail */
repeat the process;
```

If all the higher-order approximations are marked as unusable, only a first order (normal or t-distribution) approximation is used for data analysis. For a thicker tail, how higher-order results are presented can be up to the user - for example, if some higher orders are not available, the result of the last available order can be repeated in their place. A more detailed option might be to use the last available order if the function is not well-behaved and the first order approximation if it is nice but less conservative.

While the described procedure goes through a lot of steps, the actual computational time is reasonable: "irregularities" in the functions are usually close to the center and are discovered early in the process; for a thin tail, second order (term 1) always detects it and stops the process; and once an order is determined to be unusable, there is no need to calculate function values for higher orders.

2.4 Simulation

Using the data analysis procedure described above, we run several types of simulations to assess the performance of Edgeworth expansion based inference.

Error rates from empirical EE and known true distribution

In this simulation scheme, we repeat the process of calculating actual error rates used throughout this chapter. It requires knowing the true sampling distribution of a test statistic, and the critical values are obtained with Edgeworth expansions that use empirical moments and/or cumulants from the sample. Decisions of which higher-order approximations should be used are based on a tail diagnostic procedure (section 2.3).

First, we look at the results obtained from batches of samples drawn from the same distribution - this will allow us to see the extent of variability in resulting error rates, as we did in tail diagnostic exploration. Interactive graphics in Supplementary materials 2 contain results for the distributions used previously: gamma with $\alpha = 3$ and $\alpha = 50$, log-normal with $\sigma^2 = 0.4$ and $\sigma^2 = 0.5$, mixture-normal and standard normal. Figures 2.18 and 2.19 give an example for $\Gamma(\alpha = 3)$ one sized tests - left and right sides of the plots represent left- and right-sided tests with significance level 0.05; adjusted values for the rates are not truncated in this particular example.

"Zoomed-in" picture is presented in Figures 2.20 and 2.21 - to see the rates for the right-side test. It shows increased variability for higher-order terms. Since mean and variance are calculated on a log scale, one standard deviation above the log-based mean appears to rise above first-order rates; in reality, the rates never go above first-order approximation since the procedure is designed to prevent it.

Disclaimer: For an ordinary t-statistic with unbiased variance estimate (as well as with moderated t-statistic in high-dimensional simulation below), we are using generalized Edgeworth expansions from Chapter 1; therefore the first-order approximation (zero-term expansion) is not exactly t-distribution (recall the zero term being $t_{n-1}(x/r)$, where $r^2 = \frac{n-1}{n}$ for an ordinary t-statistic with unbiased estimate of σ^2). This estimate is less conservative than $t_{n-1}(x)$, a straightforward t-distribution approximation, which makes sense because t that incorporates $s_{unbiased}$ is less spread out than t scaled by s_{biased} (see Chapter 1). To be consistent in our comparison, we compare higher-order approximations with zero-term of the same expansion, not with $t_{n-1}(x)$.

Next step for the same simulation procedure is to draw from a kind of generalized skewed distribution; for that we employ a "grid simulation", using gamma and log-normal distributions with wide range of parameters: from 1 to 50 for shape parameter in gamma and from 0.001 to 1 for log-variance in log-normal distribution. For each of those distributions we first obtained a t-statistic sampling distribution through coarsened simulation. Results are then combined and summarized as means across the distributions for each order separately; they can be seen in Figure 2.23, where they are compared with a different simulation scheme.

Hypothesis testing

Instead of using EE-based critical values and known distiribution of a test statistic, this simulation mimics a real data analysis situation with hypothesis testing, where all we have is a sample and a null hypothesis, and the analysis is expected to produce a binary result - either reject or not reject the null hypothesis. To reach extremely small probabilities, we run 10^{10} simulations (the same as when obtaining sampling distributions with coarsened simulation).

The steps for this simulation procedure are:

1. draw a sample from a null distribution;
2. calculate a t-statistic and empirical moments/cumulants;
3. using empirical moments, perform tail diagnostic;
4. based on this diagnostic, calculate $p-values$ for each number of tests m using first-order and higher-order approximations;
5. for each order's adjusted $p-value$ (Bonferroni MTP), make a decision to either reject or not reject the null hypothesis for a given significance level.

The actual error rate for each order of approximation and each number of tests is the proportion p of rejected hypotheses multiplied by the number of tests: $R = min(pm, 1)$, which can be then compared with the nominal significance level.

Figure 2.22 shows the results for this simulation and compares them with the previous section's simulation for a two-sided test with $\Gamma(\alpha = 3)$ original distribution. For the hypothesis testing simulation, there is no measure of spread - each combination of number of tests m

and order k provides only one number without showing variation between the samples. The results are surprising and differ from the previous simulation quite a bit. While first-order approximation matches the previous result (as it should), higher-order approximations do not decrease error rates nearly as much as the previous simulation shows.

Next we run a bigger model simulation with drawing from a random distribution, which would be somewhat analogous to a grid simulation in previous chapter. For that, first we randomly select a distribution family (from gamma and log-normal with probability 0.5 each), then draw a value of a parameter for that distribution from continuous uniform with *min* and *max* that match the ranges specified for the grid earlier. Figure 2.23 compares results for these two simulations. The plot for grid simulation does not display the spread because for such a wide range of distributions and parameters it would be very wide and non-informative. The comparison confirms previous observation - the first-order results are similar in two simulation schemes, and while higher-order approximations do reduce error rates, this improvement is not as striking as would be expected from the critical value simulation.

The important conceptual difference between the two simulation schemes is that in the first one we run separate simulations to a) obtain sampling distribution of the test statistic (only t-statistic is of interest in each simulated sample) and b) produce empirical Edgeworth expansions (only moments/cumulants from each sample are used). In a hypothesis testing scheme, each sample provides both test statistic and a set of moments/cumulants that are used in EE. This brings us to the issue of dependence of a t-statistic and sample moments, which can offer an interesting insight into the nature of extreme values for t-statistic in small samples and might point to a promising new direction for methods to protect against false positives.

Dependence of t-statistic and empirical moments

Under the null, t-statistic and empirical moments (as well as scaled cumulants, λ's) are not independent: the subset of samples that produce extreme t-statistics will on average have cumulants that differ from those of the whole set of samples. In hypothesis testing we are mostly interested in that particular subset as it produces false positives and drives error rates. When the sample size is small, the main reasons pushing the value of t-statistic away from the center to the extremes are: 1) null hypothesis is false (true positive), 2) most of the observations are from the far tail of the distribution by chance (false positive), and 3) most of the observations are clustered together, so even if the mean is reasonably close to the center, it is amplified by a small sample variance (false positive). The second scenario is highly unlikely and the large number of false positives indeed come from the third one.

In fact, in a group of samples producing false positive results, sample variance will be smaller in general and λ's will be less representative of the data generating distribution than those in the set of all the samples. For a two-sample t-test, the treatment group might be distinguished by a smaller variance than the control group under the null. This suggests that if we had some idea of a likely range for the variance, an unusually small value of sample variance in conjunction with the large absolute value of t-statistic might raise a flag for a potential false positive. High-dimensional data might offer an opportunity to do that (which could explain good performance of empirical Bayes methods even when the assumptions are not true); taking this idea further, it might be helpful to extend the comparison to higher moments and cumulants. Two-sample t-tests might be another situation where t-statistics that are inflated due to uncharacteristically small variance can be detected and investigated further.

High-dimensional simulation

For high-dimensional simulations, we use several approaches:

1. all the features (genes) are simulated from the same known distribution, a proportion of them is shifted to a random distance to introduce "signal";
2. the features in the same simulated dataset come from different randomly selected distributions and parameters, with similar shifts for signal;
3. data is generated from a real dataset using a convex pseudo-data generating technique.

We compare performances of ordinary and moderated t-statistics across higher-order approximations. Empirical EE for moderated t-statistic require one extra substitution in expressions for "constants" A and D (Chapter 1) - they include unknown feature-specific variance σ_g^2, for which we plug in a value of a sample variance s_g^2. These constants are part of the terms of expansions and, most importantly, of a variance adjustment r^2 that we use for generalized EE. This variance adjustment is part of the quantile argument to zero-term's c.d.f. $t_{n-1}(\frac{x}{r})$ as well as to all the p.d.f. factors in subsequent terms. In our simulations and application of the EE-based analysis method, it appears, surprisingly, that this adjustment seriously weakens the performance of all the orders in moderated t-statistic. Compare the results in Figure 2.24 that applies expansions with and without the adjustment to the data simulated from $\Gamma(\alpha = 3)$. There is a very slight increase in power from using an adjustment; in terms of error rates (FDR), the results for ordinary t-statistics are almost the same but for moderated t they are increased dramatically, making its performance poor compared to that of an ordinary t. To find the root of this problem, a separate investigation would need to be conducted; the issue might stem from σ_g^2 substitution or from using r^2 in general, but in any case this adjustment is not arbitrary and is a result of analytical derivation.

We look at the few simulations with "no-adjustment" expansions as they provide interesting results (Figure 2.25). Moderated t-statistics perform very well compared to ordinary t, both in terms of error rate control and power. They seem to be especially helpful when all the features come from the same distribution as the empirical prior information is particularly valuable and is a strong stabilizing factor for individual features. As expected, higher-order approximations improve error rate control, though not much for moderated t in plot a) where it is already in a good range for the first order, but also lose in power somewhat.

Figures 2.26 and 2.27 show expansions with the variance adjustment r^2 for FWER and FDR. As pointed out above, results for moderated t, while increasing the power, present higher error rates. Higher-order expansions help with error rate control quite a bit - Figure 2.26 b) is especially interesting in that regard as the improvement in FWER comes with almost not loss in power. For FDR in convex pseudo-data (Figure 2.27) fourth and fifth orders do not help much with error rate control.

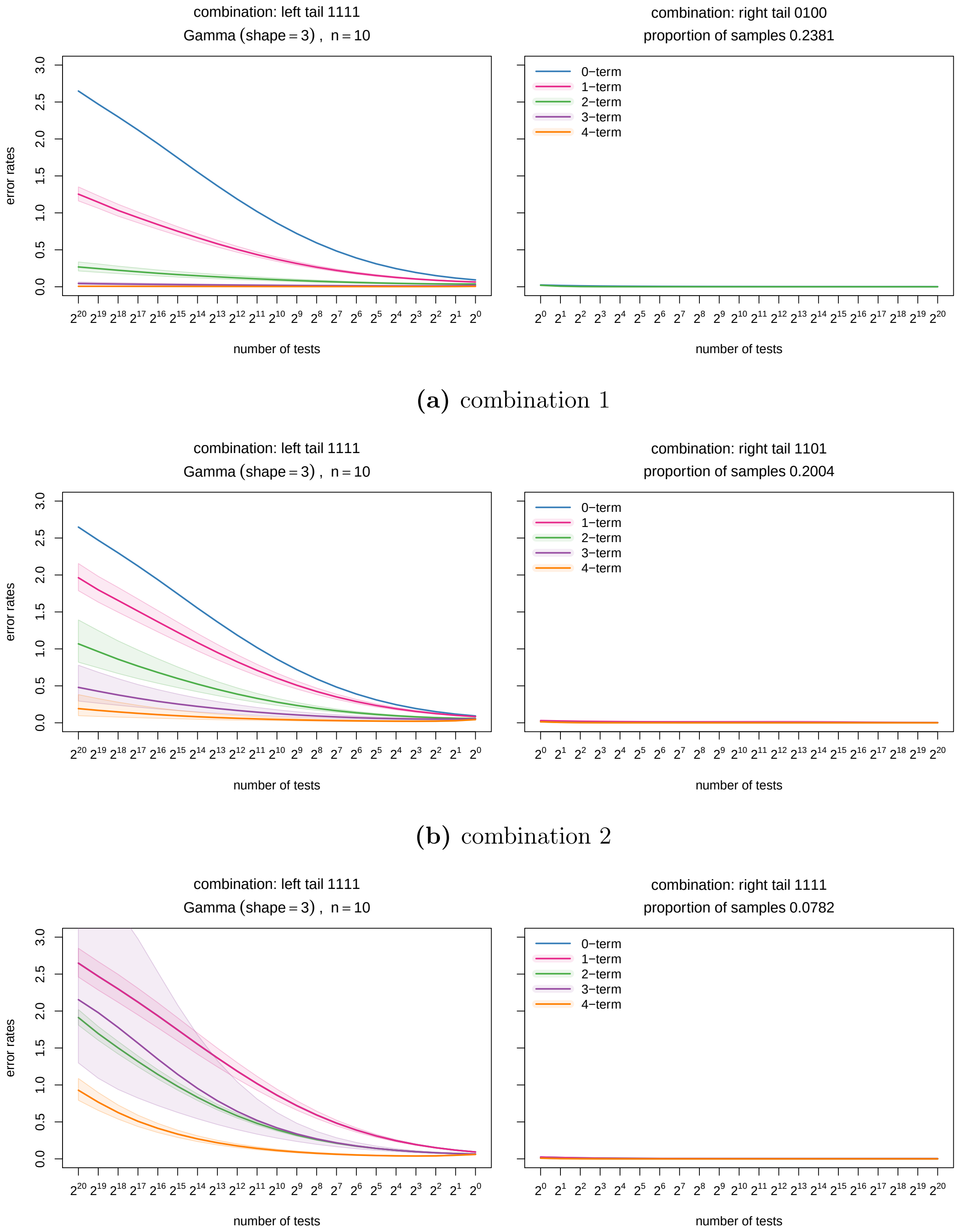
2.5 Discussion/future research

Edgeworth expansions offer many interesting possibilities for data analysis - and the more these possibilities are researched, the more questions, challenges, and insights leading to new promising directions appear. We address some of these challenges and issues to be figured out and outline a few intriguing new paths and opportunities for continuing research.

When using empirical higher orders, it is important to be mindful of the relationship between sample size n and order k. While rigorous summation of conditions and guidelines would require more research, we offer a few thoughts on this relationship. Some formalism is on the surface: moment estimates require $n \geqslant k - 1$ condition for a k'th order estimate; this can go beyond $k - 1$ if a more complex generalized form of expansion is used. Another partially formal consideration is finite moments condition and relation of t-statistic's sampling distribution to Student's t-distribution with ν degrees of freedom, where $\mu_k < \infty$ if and only if $k < \nu$. Finally, as our preliminary study of unbiased moment estimates suggested, small sample estimates of higher-order moments and moment-containing expressions can be quite unstable resulting in empirical quantities such as scaled cumulants that are far from the truth. That brings us to the question of how helpful high-order approximations are to the inference in very small samples. There are indications that EE might actually be more useful for inference in moderate to large sample sizes, where they can reliably refine inferential procedures.

Investigation of estimate instability brings up another possibility. Scaling, or "studentization" is a source of great variability in small samples (this will be also expanded on in Chapter 3) - in test statistics and analysis methods. By setting up a generalized version of Edgeworth expansions, we open up an opportunity for developing the idea further and research expansions for non-normalized (non-scaled) statistics, which might eliminate division from algebraic expressions for the terms thus stabilizing them significantly. Variability might still come into play in this scenario but it seems to be a direction worth exploring.

Researching the issue of dependence of t-statistic and empirical moments suggested another promising venue for inference that can use sample moments to detect/flag potential false positives. High-dimensional data such as produced by genomic experiments has the advantage of additional information that can be borrowed for individual feature inference; empirical Bayes methods shrink variance estimates for different features toward a common value, which is especially helpful for small sample inference - not only stabilizing the variance but also enlarging the variance precisely for the group of samples that are in danger of producing false positives: samples with extreme t-statistics where the null hypothesis is true. If empirical Bayes methods are extended to higher moments, that can be used to further identify potential false positives, leading to even more fine-tuned methods that improve error-rate controls without sacrificing the power. These methods can combine both empirical Bayes approach and Edgeworth expansions. They might potentially be even more helpful in two-sample t-tests (even without equal variances assumption), where detecting specific differences between control and treatment groups' statistics would assist in identifying potential false positives and result in better error rate control.



(a) combination 1

(b) combination 2

(c) combination 6

Figure 2.14: Tail diagnostic groups: selected combinations

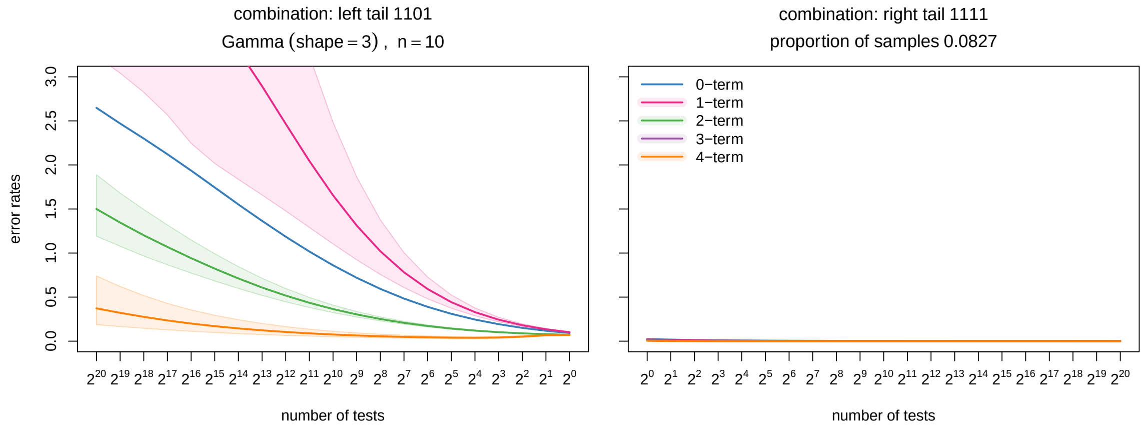


Figure 2.15: Tail diagnostic - sample not representative

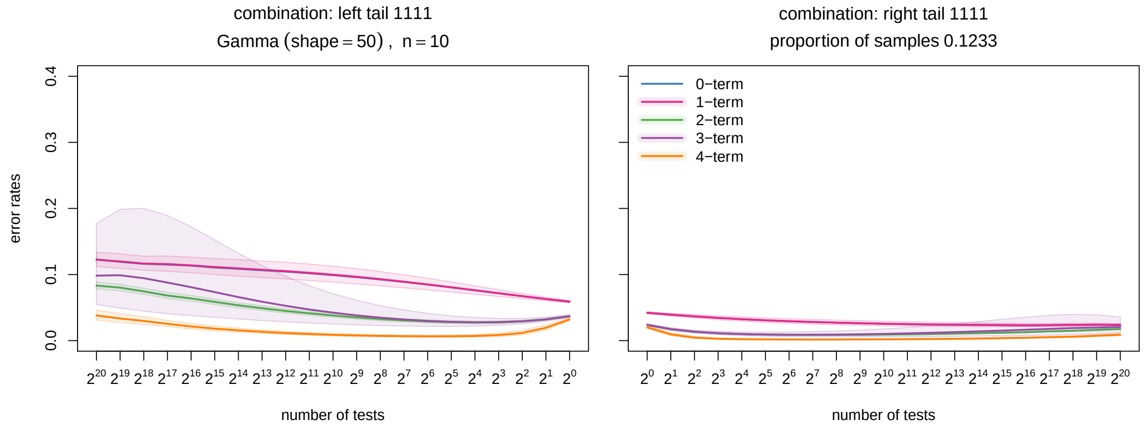


Figure 2.16: Tail diagnostic - near symmetric distribution

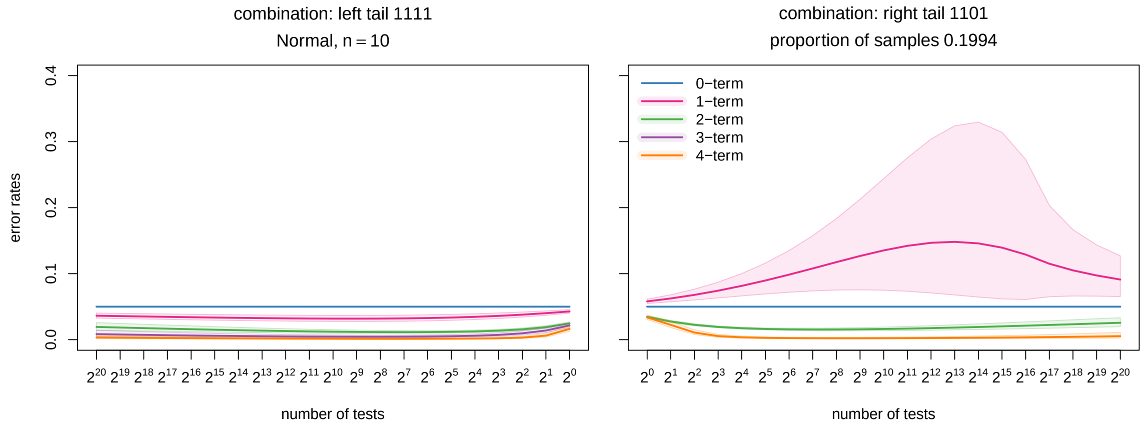


Figure 2.17: Tail diagnostic - normal distribution

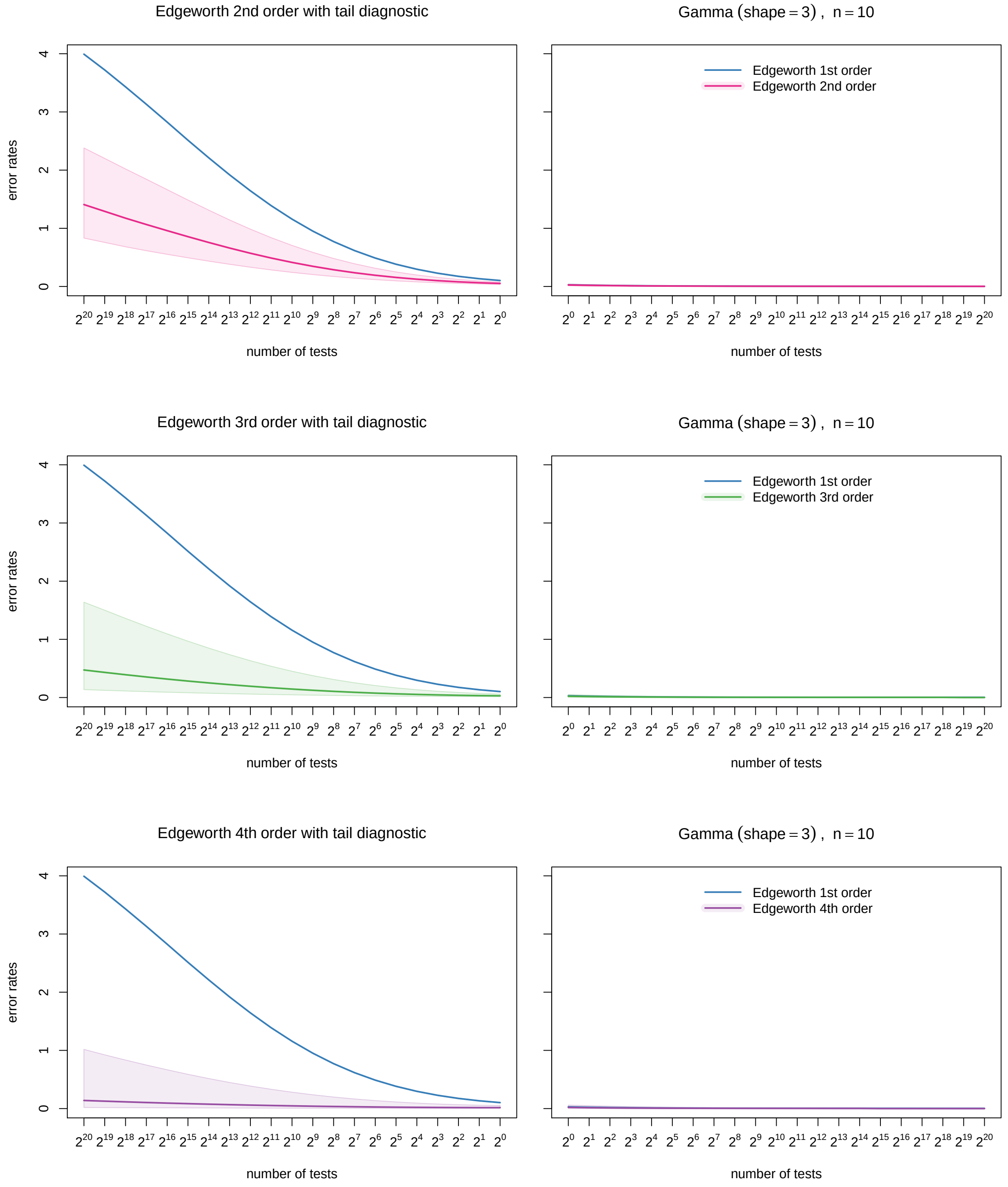
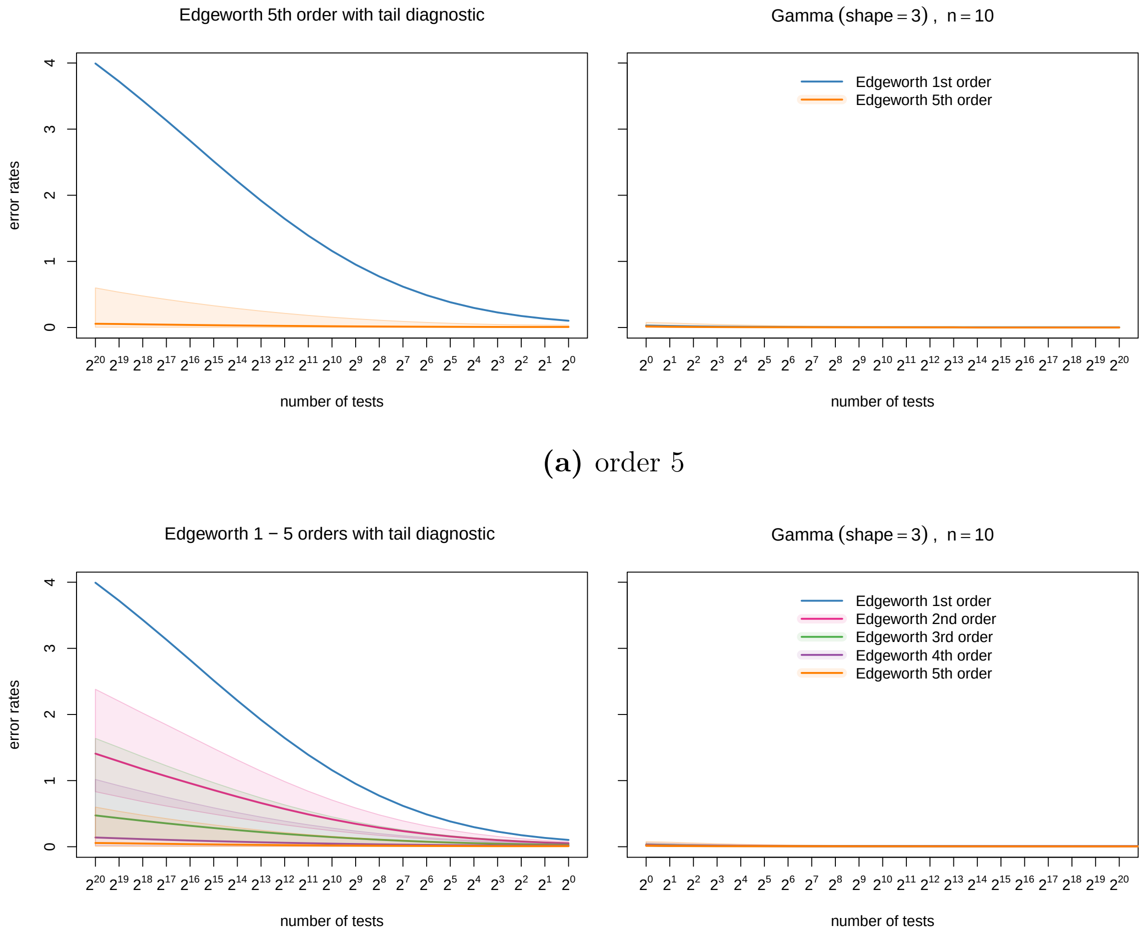


Figure 2.18: Edgeworth expansions of orders 2 - 4: data analysis with tail diagnostic



(a) order 5

(b) orders 1 - 5

Figure 2.19: Edgeworth expansions of order 5 and all orders combined: data analysis with tail diagnostic

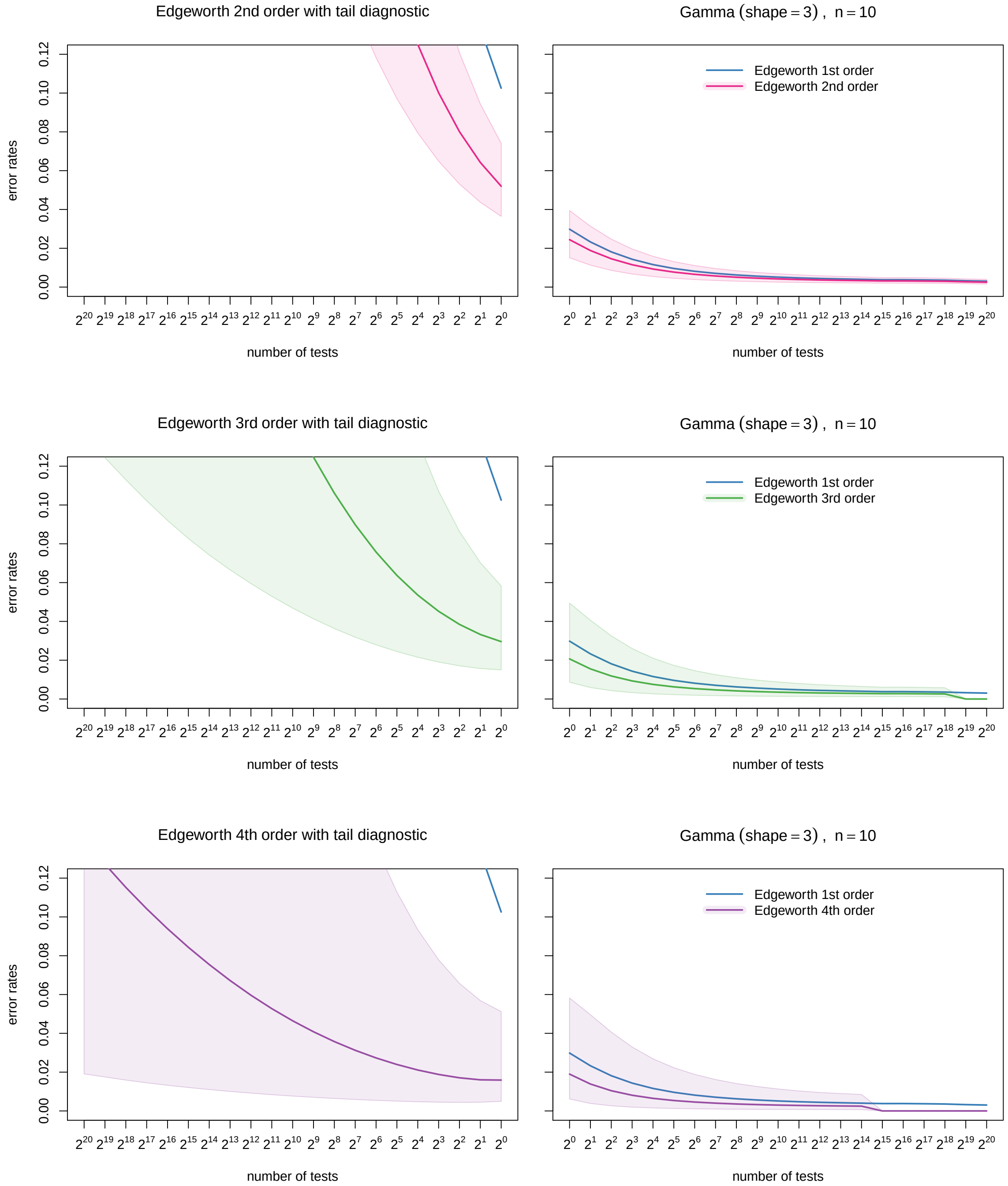
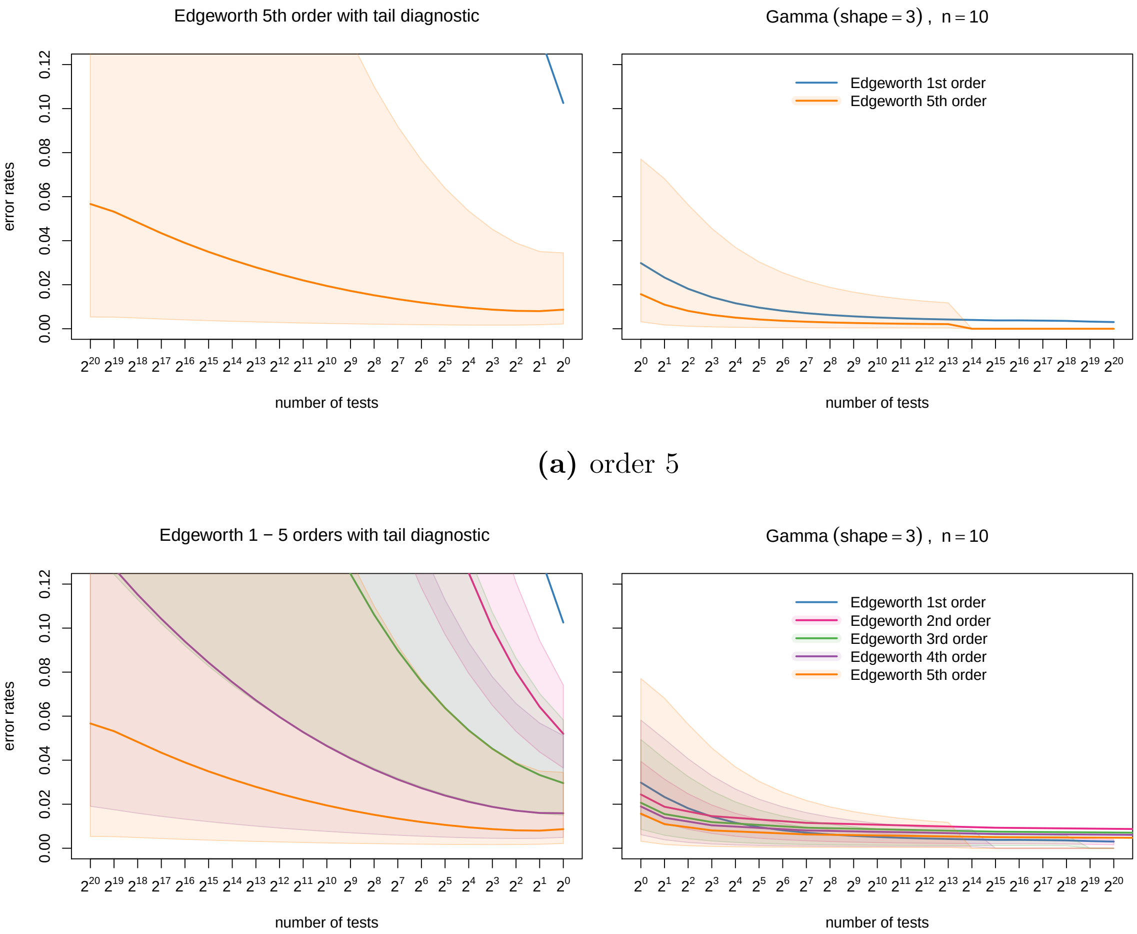


Figure 2.20: Edgeworth expansions of orders 2 - 4 zoomed in



(a) order 5

(b) orders 1 - 5

Figure 2.21: Edgeworth expansions of order 5 and all orders combined zoomed in

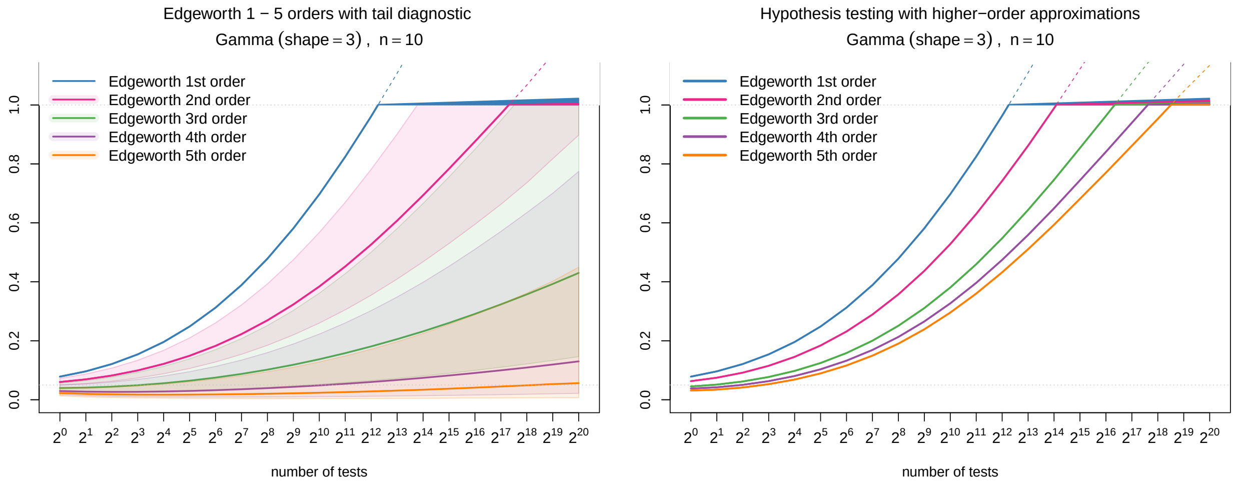


Figure 2.22: $\Gamma(\alpha = 3)$: critical values vs hypothesis testing simulation

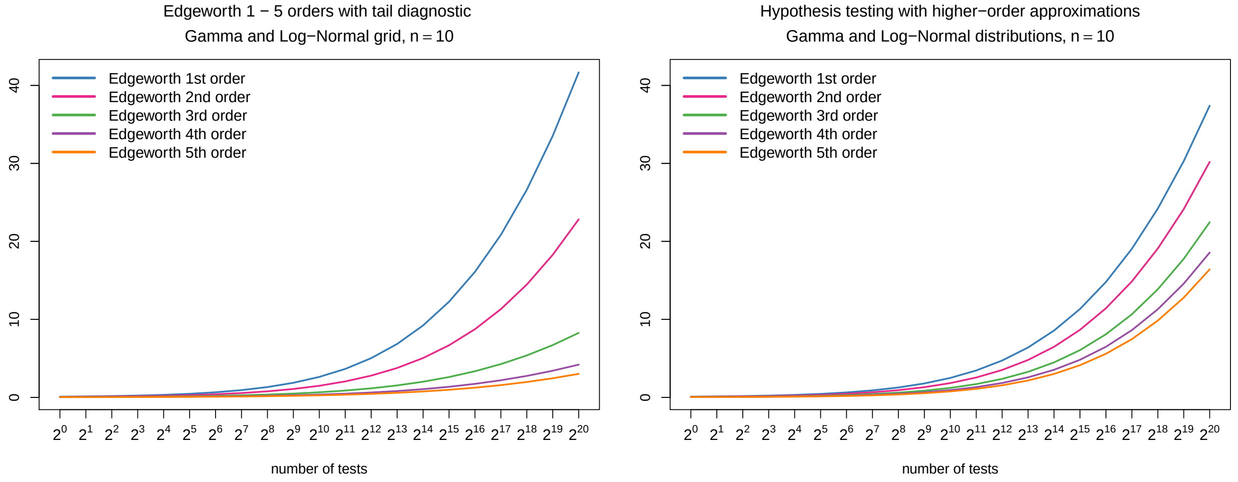
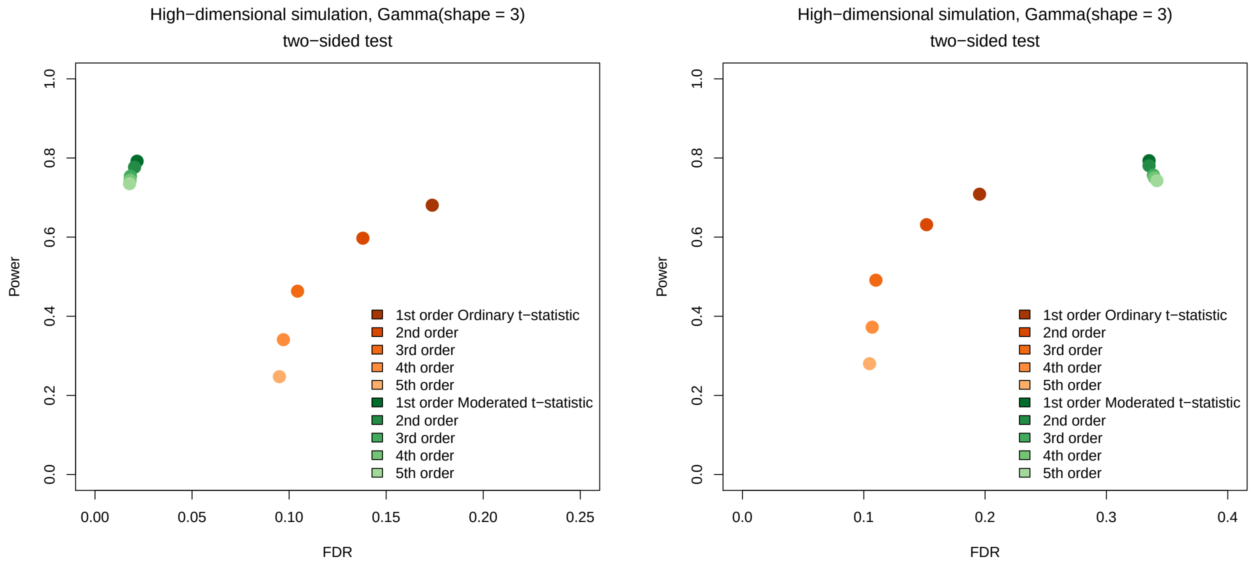


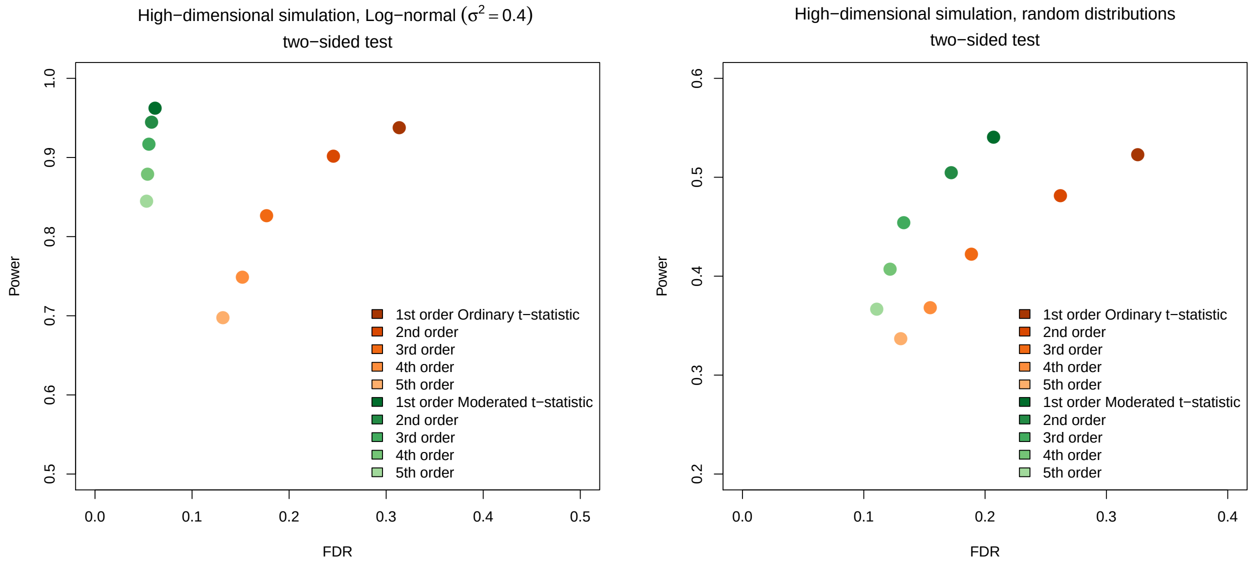
Figure 2.23: Grid vs hypothesis testing simulation



(a) no variance adjustment

(b) with variance adjustment

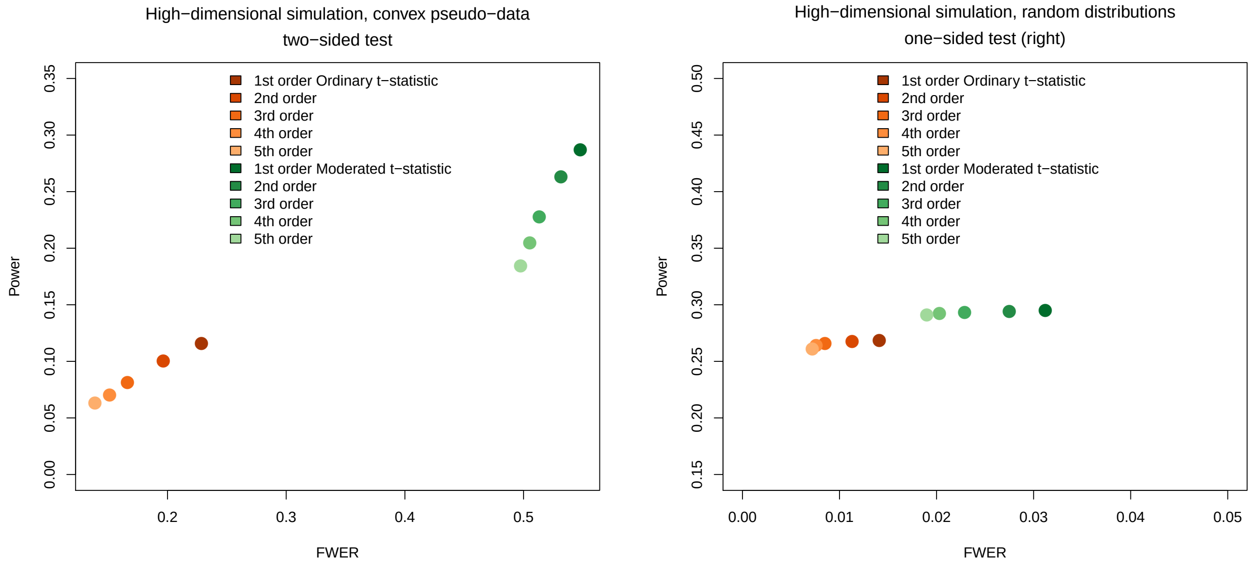
Figure 2.24: High-dimensional simulation, same distribution for all features: with and without adjustment



(a) same distribution for all features

(b) different distribution for each feature

Figure 2.25: High-dimensional simulation: no variance adjustment r^2



(a) **(b)**

Figure 2.26: High-dimensional simulation, family-wise error rate (with variance adjustment r^2)

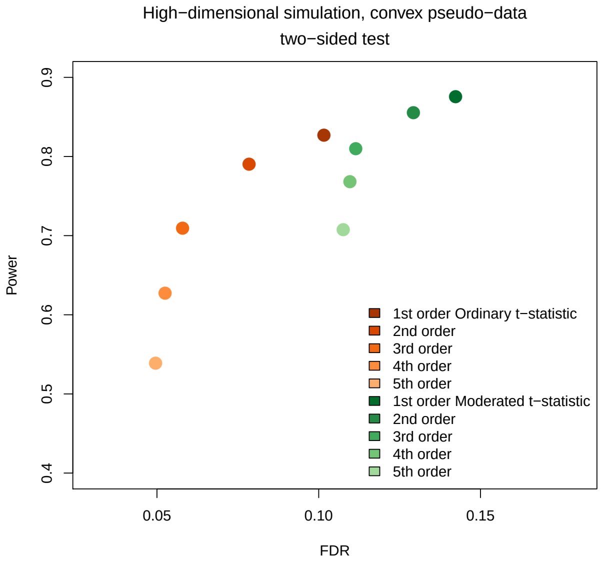


Figure 2.27: High-dimensional simulation, false discovery rate (with variance adjustment r^2)

Chapter 3

GLC Bootstrap

Ever since bootstrap was proposed in 1979 by B. Efron [22], it has remained one of the most widely used techniques for deriving inference in statistical analysis. Its introduction, followed by extensive research, lead to the development of great number of statistical methods based on it, including extensions, modifications, and improvements of the original technique. Technological advances allowed these computer-intensive methods virtually limitless applications to many fields of research. Justifying its name, bootstrapping became a truly “go to” family of methods and a natural solution for non-parametric inference. Among many advantages that ensure widespread popularity of this resampling technique are simplicity of the concept, versatility, and a broad range of possible applications. With the use of Edgeworth expansions, some important properties of bootstrap have been explored and established, such as the rate of convergence - showing that bootstrap improves on traditional first-order asymptotic approaches, providing higher-order approximations to the distribution of a test statistic [33].

These qualities make bootstrapping an attractive approach for small sample inference; at the same time, sample size poses unique challenges to resampling in general and non-parametric bootstrap in particular. Two main problems are discreteness and biased variance of the bootstrap generating distribution. Since an empirical distribution of a small sample is very discrete and the number of data points to draw from is very limited, resulting bootstrap samples will have even fewer distinct values (occasionally consisting of one repeated value only) and the method will break down. As for the bias, we will show that the variance of bootstrap samples is consistently underestimated (more so for very small samples as the bias is a function of sample size), which might affect the inference considerably, further reducing reliability of the results. Developing a bootstrap method specifically designed for small sample size that would ameliorate these problem is the goal of this Chapter.

We propose a generalization of the non-parametric bootstrap principal that smoothes empirical distribution by adding a controlled amount of noise and eliminates systematic variance bias for finite samples: General Linear Combination (GLC) bootstrap. It is an extension of the convex bootstrap [16] and pseudo-data generation idea [13]; by introducing a tuning parameter, it makes it possible to regulate the variance of bootstrap generating distribution. First, this tuning parameter is used to match bootstrap variance to the sample variance (which is the esimated variance of the data-generating distribution); later, for stabilized variance bootstrap, we provide a solution to "prescribe" posterior variance to bootstrap generating distributions of individual features in high-dimensional data analysis - using empirical Bayes method and moderated t-statistic [65], [66]. We show that GLC bootstrap can be used with correlated data, mostly preserving correlation structure. We apply the method to clustering and devise a way to assess reliability of clustering results by calculating probability that two given features/elements end up in the same cluster. In the last section, these methods are applied to yeast genomic data, in the study that involves mutant deletion strains - parallel deletion analysis (PDA).

3.1 General set up

The following set up includes non-parametric bootstrap as a special case, as well as other modifications, such as a convex bootstrap. Note that it is defined [designed] for continuous data.

Let $X' \sim P_0$, where P_0 is a true unknown data generating distribution. The data, original sample, is $x_1, \ldots, x_n$.

We introduce intermediate random variables and a tuning parameter d:

$X \sim P_n$, where $P(X = x_i) = \frac{1}{n}$, $i = 1, 2, \ldots, n$

$Z \sim U(0, d)$, $Z \perp X$, $d \in \mathbb{R}$.

Consider drawing a pair (X_1, X_2) with replacement and define a new "bootstrap" random variable $X^* \sim P_0^*$:

$$X^* = ZX_1 + (1 - Z)X_2, \tag{3.1}$$

and thus X^* is a linear combination of the pair.

The tuning parameter d is a constant and the key component of the set up. Its different values determine the version of the bootstrap. At this point, d can be restricted to the interval $-1 \leqslant d \leqslant 1$, but later (for stabilized variance version), we will relax this restriction and change it to $-\infty < d \leqslant 3/4$. Since in the general case d can be negative, to respect the order of the Uniform distribution bounds, the distribution of Z can be writen as $Z \sim U(min(0, d), max(0, d))$. Important special cases of this generalized set up include:

– Non-parametric bootstrap: $d = 0$.

- Convex bootstrap: $d = 1$ [16];
  other versions: $0 < d < 1$ [13].
- GLC bootstrap: $d = \frac{3}{4} - \frac{3}{4}\sqrt{\frac{3n+5}{3n-3}}$.

Being the function of the sample size n in the last case only, parameter d corrects the bias in variance of the bootstrap variable X^*.

Bias in Variance of X^*

First, we wish to find $Var(X^*)$ and compare it to $Var(X')$.
Define first two moments of random variable X:

$$\mu \equiv E\left(X\right) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\nu \equiv E\left(X^2\right) = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

First two moments of Z are $E\left(Z\right) = \frac{d}{2}$ and $E\left(Z^2\right) = \frac{d^2}{3}$.

Keeping in mind that X_1 and X_2 are independent, we get:

$$\begin{aligned} Var(X^*) &= Var(ZX_1 + (1-Z)X_2) \\ &= E\left[\Big(ZX_1 + (1-Z)X_2\Big)^2\right] - \Big[E\Big(ZX_1 + (1-Z)X_2\Big)\Big]^2 \end{aligned}$$

$$E\big(ZX_1 + (1-Z)X_2\big) = E(ZX_1) + E(X_2) - E(ZX_2) = E(X) = \mu$$

$$\begin{aligned} E\left[\big(ZX_1 + (1-Z)X_2\big)^2\right] &= E\Big[Z^2X_1^2 + 2X_1X_2Z\left(1-Z\right) + (1-Z)^2X_2^2\Big] \\ &= \nu E\left(Z^2\right) + 2\mu^2 E\left(Z - Z^2\right) + \nu E\left(1 - 2Z + Z^2\right) \\ &= \nu\frac{d^2}{3} + \mu^2 d - 2\,\mu^2\frac{d^2}{3} + \nu - \nu d + \nu\frac{d^2}{3} \\ &= \nu\left(\frac{2}{3}\,d^2 + 1 - d\right) + \mu^2\left(d - \frac{2}{3}\,d^2\right) \end{aligned}$$

$$Var(X^*) = \nu\left(\frac{2}{3}\,d^2 - d + 1\right) + \mu^2\left(d - \frac{2}{3}\,d^2\right) - \mu^2 = \left(\nu - \mu^2\right)\left[\frac{2}{3}\,d^2 - d + 1\right] \qquad (3.2)$$

Let $\bar{x}$, a sample mean, and s^2, a sample variance, be unbiased estimates of $E(X')$ and $Var(X') = \sigma^2$ respectively:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \mu \qquad \text{(by definition of } \mu\text{).}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n}{n-1}\,(\nu - \mu^2)$$

Now we compare these unbiased esimates of mean and variance of data-generating distribution P_0 with those of bootstrap-generating distribution P_0^*. As established above:

$$EX^* = \mu. \tag{3.3}$$

Thus, the mean of bootstrap sample is unbiased. Compare the variance:

$$Var(X^*) = \frac{n-1}{n}\,\mathbf{s^2}\left(\frac{2}{3}\,d^2 - d + 1\right). \tag{3.4}$$

As the sample size grows,

$$\lim_{n\to\infty} Var(X^*) = s^2\left(\frac{2}{3}\,d^2 - d + 1\right).$$

Therefore, asymptotically for $Var(X^*)$ to be unbiased, we need

$$\frac{2}{3}\,d^2 - d + 1 = 1,$$

which holds when $d = \left\{0,\, \frac{3}{2}\right\}$.

If $0 < d < \frac{3}{2}$, then $\frac{2}{3}\,d^2 - d + 1 < 1$ and $Var(X^*) < \sigma^2$.

Non-parametric Boostrap

Since $d = 0$, it is asymptotically unbiased. For finite samples, however,

$$Var(X^*) = \frac{n-1}{n}s^2$$

So, bootstrap generating distribution has a variance that is smaller than original sample variance by a factor of $\frac{n-1}{n}$. In other words, resulting bootstrap samples on average will have smaller variance than the original sample. In relatively small samples that can noticably affect results of the bootstrap procedure. It also means that we are "sampling" from a distribution that is somewhat different from our estimate $\hat{P}$ of the true distribution P_0.

Convex Boostrap

This version of bootstrap is asymptotically biased since $\dfrac{2}{3}\,d^2 - d + 1 < 1$. The biggest bias occurs when $d = \dfrac{3}{4}$; in this case s^2 is scaled by the factor of $\dfrac{5}{8}$. Scaling factors for $d = 1$ and $d = 0.5$, which are the most frequently used values, both equal $\dfrac{2}{3}$. This is a very serious departure from the sample variance, and it holds for any sample size.

This asymptotic bias factor is compounded by the finite sample bias that is present as well, further decreasing $Var(X^*)$.

Modified Convex Boostrap

Since the convex bootstrap has significant bias even asymptotically, a modified version of it has later been developed by the same authors [17]. Asymptotically it converges to non-parametric bootstrap: as $n \to \infty$, the contribution of the second variable of the $(X_1,\, X_2)$ pair disappears.

The set up for modified convex bootstrap is a little different from the general set up we introduced previously:
Data (original sample): $x_1, \ldots, x_n$,

$X \sim P_n$, where $P(X = x_i) = \frac{1}{n}$, $i = 1, 2, ..., n$
$Z \sim U(0,\, d),\; Z \perp X,\; 0 \leqslant d \leqslant 1$

Draw a pair (X_1, X_2) and define a new bootstrap random variable X^*:

$$X^* = Z^{1/n} X_1 + (1 - Z^{1/n}) X_2$$

Going through the same steps as before, we find $Var(X^*)$:

$E(X^*) = \mu$

$$E(X^{*2}) = E\left[Z^{\frac{2}{n}} X_1^2 + 2\, Z^{\frac{1}{n}}\left(1 - Z^{\frac{1}{n}}\right) X_1 X_2 + \left(1 - Z^{\frac{1}{n}}\right)^2 X_2^2\right]$$

$$E\big(Z^{\frac{1}{n}}\big) = \int_0^d z^{\frac{1}{n}}\,\frac{1}{d}\,dz = \frac{1}{d}\,\frac{z^{\frac{1}{n}+1}}{\frac{1}{n}+1}\bigg|_0^d = \frac{d^{\frac{1}{n}}}{\frac{1}{n}+1}$$

$$E\big(Z^{\frac{2}{n}}\big) = \int_0^d z^{\frac{2}{n}}\,\frac{1}{d}\,dz = \frac{1}{d}\,\frac{z^{\frac{2}{n}+1}}{\frac{2}{n}+1}\bigg|_0^d = \frac{d^{\frac{2}{n}}}{\frac{2}{n}+1}$$

$$\begin{aligned}
Var(X^*) &= \nu E\big(Z^{\frac{2}{n}}\big) + 2\,\mu^2\Big[E\big(Z^{\frac{1}{n}}\big) - E\big(Z^{\frac{2}{n}}\big)\Big] + \nu\Big[1 - 2\,E\big(Z^{\frac{1}{n}}\big) + E\big(Z^{\frac{2}{n}}\big)\Big] - \mu^2 \\
&= \nu\Big[2\,E\big(Z^{\frac{2}{n}}\big) - 2\,E\big(Z^{\frac{1}{n}}\big) + 1\Big] - \mu^2\Big[2\,E\big(Z^{\frac{2}{n}}\big) - 2\,E\big(Z^{\frac{1}{n}}\big) + 1\Big] \\
&= \big(\nu - \mu^2\big)\cdot C, \qquad (3.5)
\end{aligned}$$

where $C = \dfrac{2\,d^{\frac{2}{n}}}{\frac{2}{n}+1} - \dfrac{2\,d^{\frac{1}{n}}}{\frac{1}{n}+1} + 1$.

As $n \to \infty$, $C \to 1$ as does $\frac{n-1}{n}$, so asymptotically bootstrap variance is unbiased. However, for finite samples the bias might be relatively large - for example, if $n = 10$ and $d = 1$, original sample variance will be scaled by a factor of 0.76. The worst case (smallest value of C) occurs when $d = 0.5$ - then for $n = 10$ the variance of bootstrap generating distribution will be decreased by the factor of 0.68 compared to the data sample variance. Due to this, Chernick [15] notes that modified convex bootstrap does not perform well in practice.

GLC Boostrap

To remove both asymptotic and finite sample bias, we need to make

$$\frac{n-1}{n}\left(\frac{2}{3}\,d^2 - d + 1\right) = 1 \qquad (3.6)$$

For d, we pick one of the roots of the quadratic equation (3.6):

$$d = \frac{3}{4} - \frac{3}{4}\sqrt{\frac{3n+5}{3n-3}} \qquad (3.7)$$

d can be viewed as a tuning parameter that targets bootstrap variance of a random variable. Variance can be the crucial component of the inference; however, there might be reasons to target other characteristic of the distribution (e.g. skewness). In that case, the approach would be similar but the process/result might be more complicated, depending on the characteristic.

Features of GLC bootstrap

It follows from (3.7) that d is always negative and $\lim\limits_{n\to\infty} d = 0$. Moreover,

$$\begin{aligned}
d &= \frac{3}{4}\left(1 - \sqrt{\frac{3n+5}{3n-3}}\right) = -\left(\frac{1}{n} + \frac{1}{3\,n^2} + \frac{5}{9\,n^3} + \dots\right) \\
&= -\frac{1}{n}\left(1 + \frac{1}{3\,n} + \frac{5}{9\,n^2} + \dots\right) = -n^{-1} - \mathcal{O}(n^{-2}) \qquad (3.8)
\end{aligned}$$

This shows that d is approximated by $-\frac{1}{n}$, and the approximation is good for larger n (it makes sense intuitively as we are trying to "stretch" the range by $1/n$). These are crucial features of GLC bootstrap; in contrast, the other root of (3.6) is greater than 3/2 and $\lim_{n\to\infty} d = 3/2$.

The linear combination of X_1 and X_2 is no longer convex; in fact, the new generated bootstrap value is always going to be outside of the segment formed by the pair of sampled observations. Expression (3.7) for d also means that variable Z will take small negative values, and therefore X^* will be close to one of the original observations. This is illustrated by Figure 3.1: values for Z are plotted on x-axis, and values of (X_1, X_2) - on y-axis. In this example, $X^* > X_2 > X_1$.

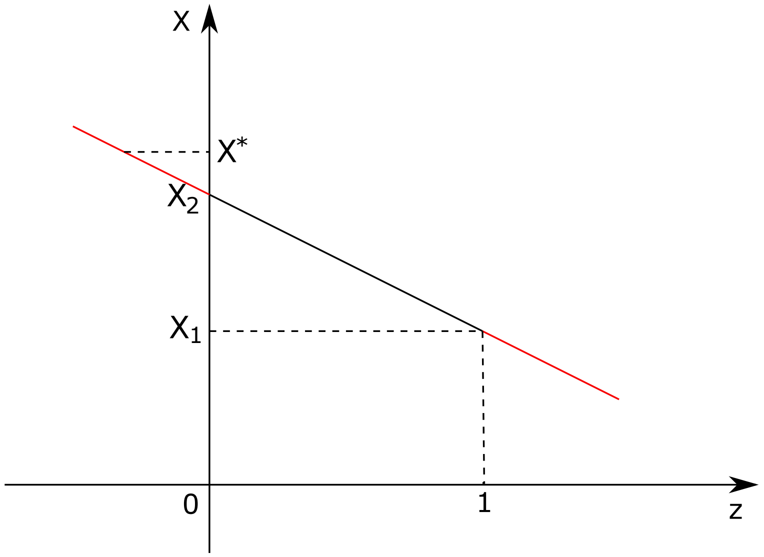


Figure 3.1: GLC bootstrap: non-convex linear combination

It naturally follows that the bounds of a bootstrap sample are beyond the actual range of the original data sample, though by a small amount:

$$\begin{aligned}\inf\{X^*\} &= (1-d)X_{(1)} + dX_{(n)} = X_{(1)} + d(X_{(n)} - X_{(1)})\\ \sup\{X^*\} &= dX_{(1)} + (1-d)X_{(n)} = X_{(n)} - d(X_{(n)} - X_{(1)}),\end{aligned}$$

where $X_{(1)}$ and $X_{(n)}$ are the minimum and maximum of the sample.

We can view GLC unbiased sampling as essentially a non-parametric bootstrap with some added random noise (or slight data perturbation). Asymptotically GLC bootstrap converges to non-parametric bootstrap since $d \to 0$ as $n \to \infty$, which means that $Z = 0$ is a constant and $X^* = X_2$. This is specific to the negative root of (3.6); the other root would not have yielded these properties. Another important feature is that the smaller the sample size n is, the more randomness is involved in generating a bootstrap sample, which helps to smooth a parameter distribution and make up for sparsity of the original data (pseudo-data generating feature). As the sample size increases, added noise decreases and bootstrap values get closer and closer to original data values.

Density of GLC bootstrap generating distribution

For GLC bootstrap, the density is not fully continuous - there is still point mass of $1/n^2$ at each data point, compared with $1/n$ for non-parametric bootstrap, which makes repeated points in GLC bootstrap sample considerably less likely.

The density of the bootstrap generating distirubtion:

$$f(x) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} A(x, x_i, x_j), \tag{3.9}$$

where

$$A(x, x_i, x_j) = \begin{cases} \delta(x - x_i) & \text{for } i = j \\ \dfrac{1}{|d\,(x_j - x_i)|}\, I\Big(x \in \big[x_i, x_i + d\,(x_j - x_i)\big]\Big) & \text{for } i \neq j \end{cases}$$

x_i's denote observations in the original sample and $\delta(\cdot)$ is Dirac delta function.

Figure 3.2 shows the graphs of the density for two different sample sizes: $n = 5$ and $n = 15$ with one sample for each n simulated from $U(-1,\, 1)$ distribution. Vertical dotted lines indicate data points x_i that have point mass ($f(x_i) = \infty$). Note higher density values concentrated around these data points. The density for larger n is smoother but the peaks around x_i's are more pronounced because of the smaller value of $|d|$.

Alternative set up - drawing without replacement

When we defined a bootstrap variable in the general set up, the pair (X_1, X_2) was drawn with replacement. Alternatively, the pair can be drawn without replacement; in that case, unless $d = 0$, there is zero probability that a bootstrap sample will contain any original data values. For GLC bootstrap with this set up, the expression for d is different and the density is continuous, without non-zero point mass at any point.

Compared to with-replacement set up, absolute value of d in this case is greater for the same sample size and, consequently, the bounds are wider. If we represent d as a function of n: $d_w(\cdot)$ for a with-replacement set up and $d_{wo}(\cdot)$ for without-replacement, we get the tuning

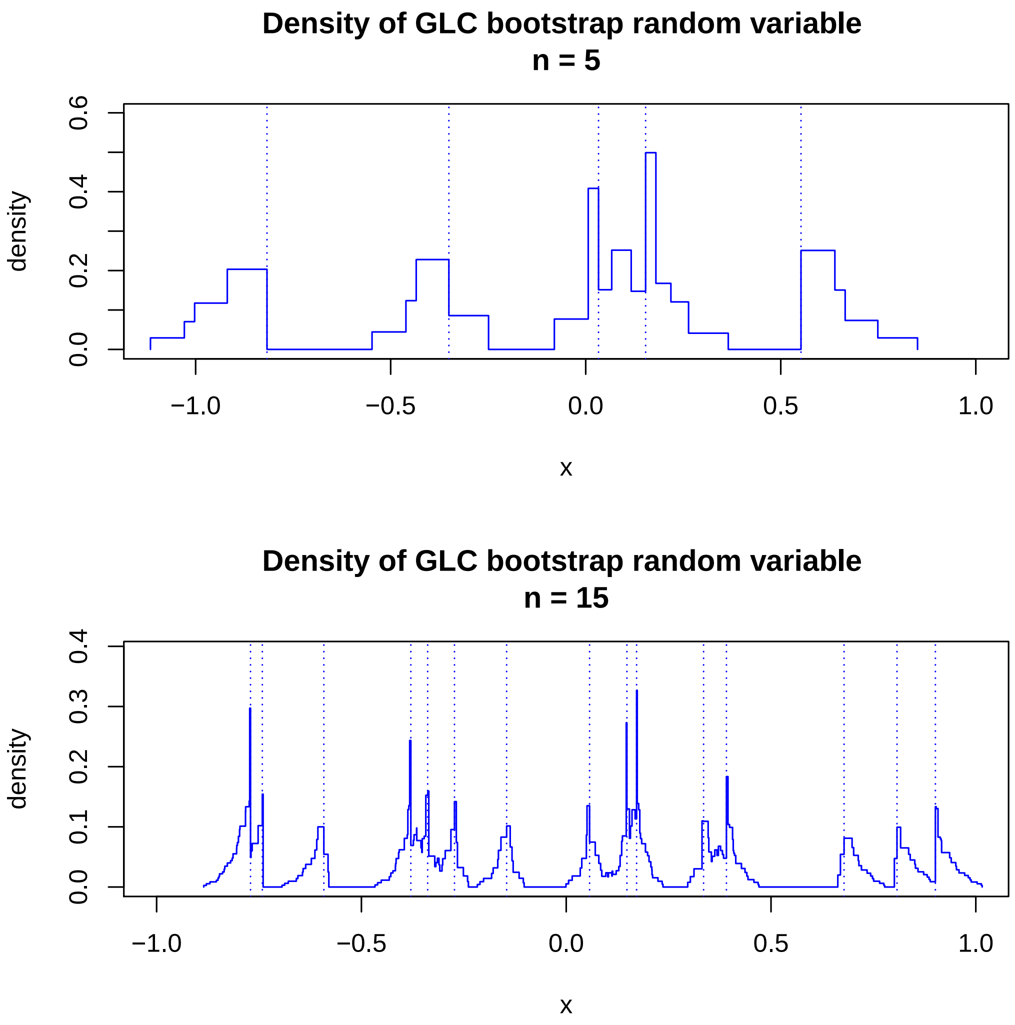


Figure 3.2: Density of GLC bootstrap

parameter for this alternative set up in the following way:

$$d_{wo}(n) = d_w(n+1) \tag{3.10}$$

Conditions

In general, for any bootstrap to work, two basic conditions have to be satisfied: continuity of an estimator $T(P)$ in the neighborgood of P_0 and convergence of its sampling distribution. We claim that GLC bootstrap works - its distribution converges to the true distribution as $n \to \infty$ - under the same conditions as non-parametric bootstrap.

To justify this claim, we turn to one of the more general formal results: Consistency of Bootstrap theorem by Enno Mammen [48] (adapting the notation for P fixed for every n, i.i.d. case):

1 *Consider a sequence $X_1, \dots, X_n$ of i.i.d. random variables with distribution P_0. For a function g consider $\hat{T}_n = T(\hat{P}_n) = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$. Consider a bootstrap sample $X_1^*, \dots, X_n^*$ with empirical distribution $\hat{P}_n^*$. Denote $\hat{T}_n^* = T(\hat{P}_n^*)$. Then for $t_0 = T(P_0)$ the following assertions are equivalent:*

1. *$\hat{T}_n$ is asymptotically normal: There exist σ_n with*
$$d_\infty(\mathcal{L}(\hat{T}_n - t_0), N(0, \sigma_n^2)) \to 0.$$
2. *The normal approximation with estimated variance works:*
$$d_\infty(\mathcal{L}(\hat{T}_n - t_0), N(0, \hat{S}_n^2)) \xrightarrow{p} 0,$$
where $\hat{S}_n^2 = \frac{1}{n^2}\sum_{i=1}^{n}(g(X_i) - \hat{T}_n)^2$.
3. *Bootstrap works:*
$$d_\infty(\mathcal{L}(\hat{T}_n - t_0), \mathcal{L}^*(\hat{T}_n^* - \hat{T}_n)) \xrightarrow{p} 0.$$

Here d_∞ denotes the Kolmogorov distance and $\mathcal{L}^*(\dots)$ is the conditional law $\mathcal{L}(\dots|X_1, \dots, X_n)$.

The reasoning for this justification is that GLC bootstrap asymptotically converges to non-parametric bootstrap; the rate of convergence is faster than the rate of convergence of non-parametric bootstrap to normality.

Let $\hat{P}_n^{**}$ be a glc bootstrap generating distribution and $\hat{T}_n^{**} = T(\hat{P}_n^{**})$.
From the proof of the theorem [48] we have:
$$d_\infty\big(\mathcal{L}(\hat{T}_n - t_0), \mathcal{L}^*(\hat{T}_n^* - \hat{T}_n)\big) = \mathcal{O}\Big(\frac{1}{\sqrt{n}}\Big)$$

Recall from GLC bootstrap definition (3.8): $d = -\frac{1}{n} - \mathcal{O}\Big(\frac{1}{n^2}\Big)$.

If it can be shown that this feature implies that $d_\infty\big(\mathcal{L}^*(\hat{T}_n^* - \hat{T}_n), \mathcal{L}^{**}(\hat{T}_n^{**} - \hat{T}_n)\big) = \mathcal{O}\Big(\frac{1}{n}\Big)$ or at least $\mathcal{O}\Big(\frac{1}{\sqrt{n}}\Big)$, then
$$d_\infty\big(\mathcal{L}(\hat{T}_n - t_0), \mathcal{L}^{**}(\hat{T}_n^{**} - \hat{T}_n)\big) = \mathcal{O}\Big(\frac{1}{\sqrt{n}}\Big).$$

Even more generalized conditions are formulated in Davison et al [21]:

2 *Let F be a true cumulative distribution function surrounded by a neighborhood $\mathcal{N}$ in a suitable space of distributions, and that as $n \to \infty$, $\hat{F}$ eventually falls into $\mathcal{N}$ with probability one. $G_{F,n}$ is the distribution function we wish to estimate. Then in order for a bootstrap c.d.f. $G_{\hat{F},n}$ to approach $G_{F,n}$ as $n \to \infty$, three conditions must hold:*

1. *For any $A \in \mathcal{N}$, $G_{A,n}$ must converge weakly to a limit $G_{A,\infty}$;*
2. *This convergence must be uniform on $\mathcal{N}$;*
3. *The function mapping A to $G_{A,\infty}$ must be continuous.*

Note that here the estimator is not required to be asymptotically normal but needs to have some limiting distribution in the neighborhood of F. There is no proof provided for these conditions, so it might be hard to prove that they would also hold for GLC bootstrap; however the same argument about convergence of GLC to regular non-parametric bootstrap might be applied in this case as well.

Simulations

First, we check if GLC bootstrap indeed achieved the intended goal of removing the bias in variance from bootstrap samples. For each simulation, a two-variable sample $X_1, \ldots, X_{n_x}, Y_1, \ldots, Y_{n_y}$ is drawn; based on that sample, three bootstrap procedures are performed: non-parametric bootstrap, convex, and GLC. Each procedure yields a variance for the bootstrap esimator $\hat{\psi}^*$ - we will look at these variances across all the simulations.

$X \sim 0.5\, N(2,\, 1) + 0.5\, U(2,\, 3)$,
$Y \sim 0.7\, Exp(1) + 0.3\, N(0.5,\, 0.5)$, $X \perp Y$,
$n_x = 13$, $n_y = 9$,
$\psi_0 = E(X - 0.8\, Y + 0.3\, X^2 - 0.2\, XY)$.

In Figure 3.3, historgrams of scaled variances for three procedures are displayed. There is no difference in their shapes (green line overlay follows χ^2 distribution), but the means for non-parametric and convex bootstrap show bias - the variance is underestimated, while there is no bias in GLC. Since these quantities are normalized for convenience of comparison, the means are the same while the true variance appears as shifting (it is scaled by a different factor in each case). The fourth graph in Figure 3.3 provides examples of bootstrap densities of $\hat{\psi}^*$ for some of the simulations with solid line for the true sampling distribution of the estimator.

Next, we look at the shape of bootstrapped distribution of an estimator and assess how well it approximates the true sampling distribution. An example:

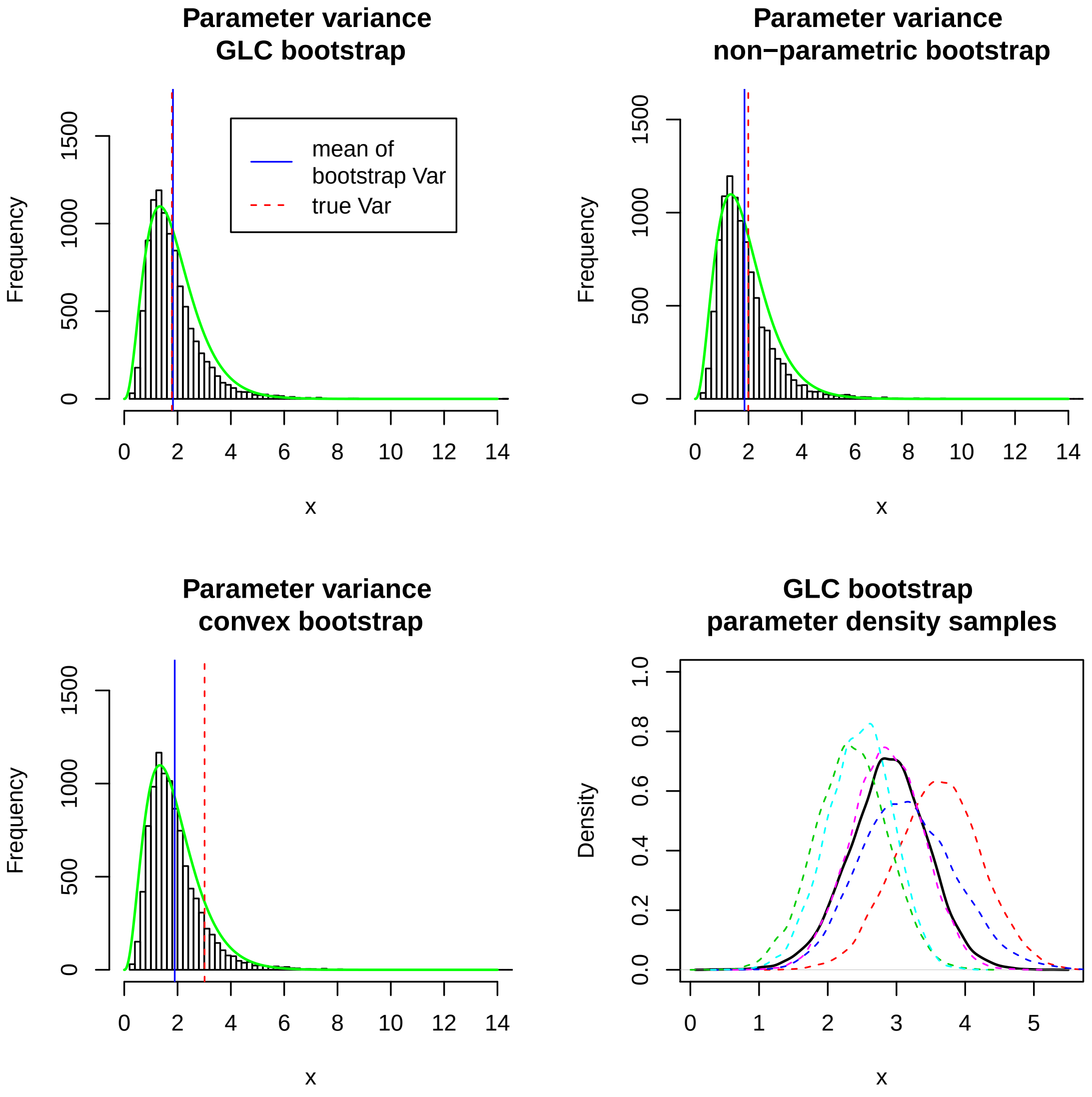
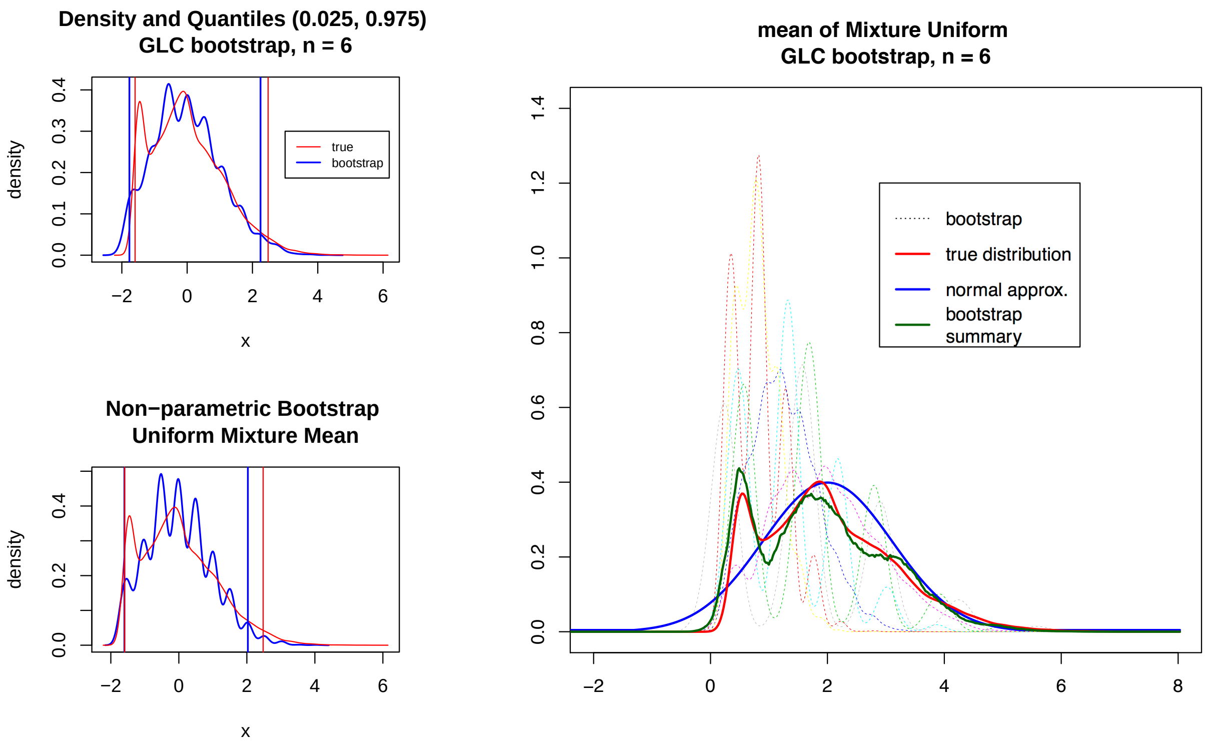


Figure 3.3: Estimated variances of sampling distribution: variance bias in non-parametric, convex, and GLC bootstrap

$X \sim \dfrac{2}{3}\,U(0,\,1) + \dfrac{1}{3}\,U(0,\,10);$
$X_1, \ldots, X_n$ - random sample, $n = 6$;
$\hat{\psi} = \bar{X} - E(X); \quad \hat{\psi}^* = \bar{X}^* - \hat{\psi}.$

Figure 3.4a compares the densites of $\hat{\psi}^*$ obtained with GLC (top) and non-parametric (bottom) bootstrap from the same sample. The density of $\hat{\psi}$ (true sampling distribution) is plotted in red and the density of $\hat{\psi}^*$ (bootstrap) in blue. The graphs also show 0.025 and

0.975 quantiles of these distributions. We can see that GLC bootstrap produces a much smoother distribution than non-parametric for this sample size - the result of extreme discreteness of non-parametric bootstrap for such small samples. There is also an improvement in terms of the tail quantiles for GLC bootstrap. Figure 3.4b displays an informal summary across many simulations, each simulation producing its own density of $\hat{\psi}^*$. Examples of some of these densities are included (dotted lines). As before, the true sampling distribution is in red; it is quite different from its normal approximation (in blue). Green line represents the summary of all the bootstrapped densities and seems to be close to the true one.



(a) **(b)**

Figure 3.4: Sampling distribution of the mean of mixture-uniform random variables - bootstrap vs true

Similarly, in Figure 3.5 we compare true sampling distribution with normal approximation and bootstrap summary - but for a highly skewed sampling distribution:

$X \sim 0.5\, N(3,\, 1) + 0.5\, N(-3,\, 3^2)$, $n = 6$;

$\hat{\psi} = \displaystyle\sum_{i=1}^{n} X_i^3$.

As previously, dotted lines represent c.d.f.'s for bootstrap sampling distributions; their summaries are in solid lines: light green for non-parametric bootstrap, light blue for convex, and dark green for GLC bootstrap. In the bottom left graph we can see that none of the bootstrap versions approximate the center and the thin tail of this distribution well; however, GLC bootstrap gets close to the true distribution in the left, thicker, tail. Here we also start exploring the performance of studentized bootstrap; it has been established that in certain cases, studentized non-parametric bootstrap has faster convergence rate than non-studentized one [33]. It does seem to be closer to the truth in the central region of the distribution but is extremely conservative in the right tail (bottom right graph).

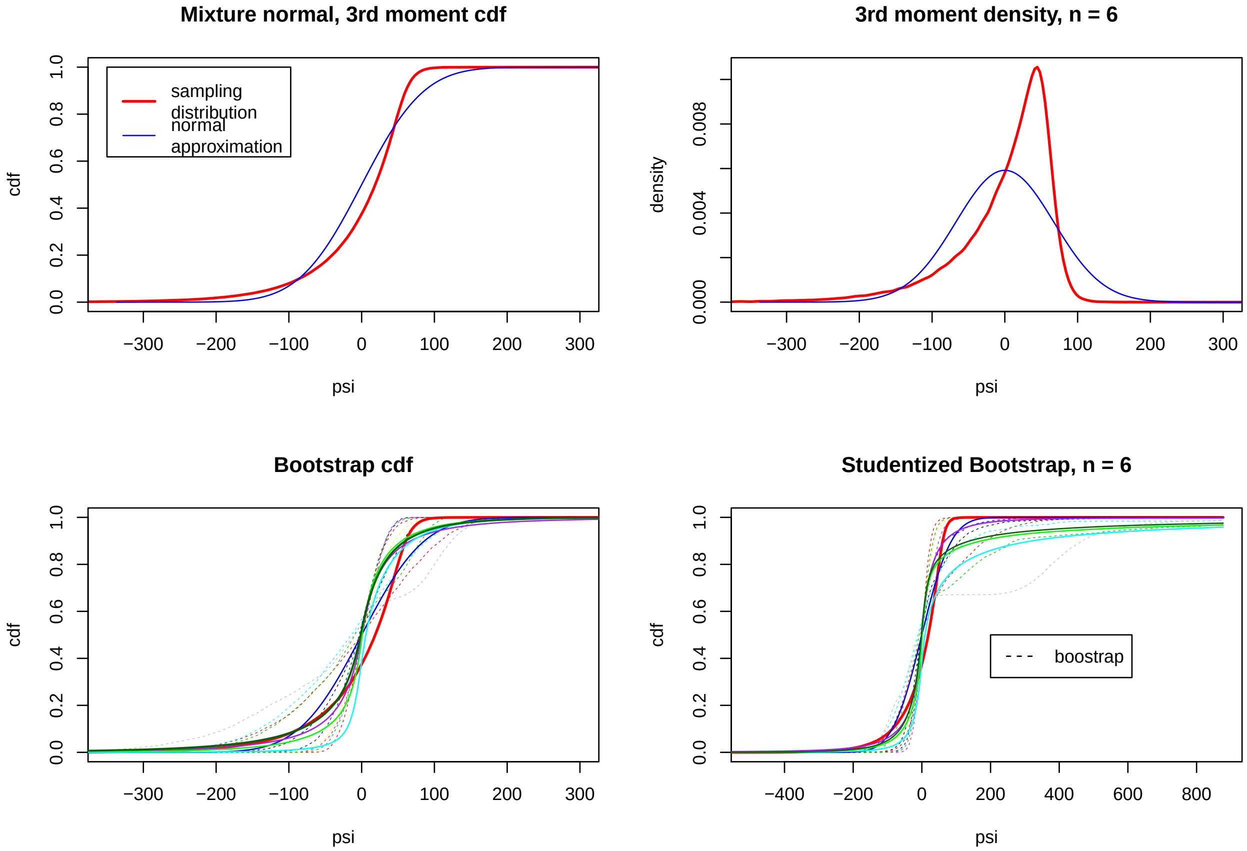


Figure 3.5: Highly skewed sampling distribution (3rd moment of mixture-normal random variables) - examples of bootstrap sampling distributions and their summaries across simulations

In Figure 3.6, we look at the tails more closely, comparing non-parametric, convex, and GLC bootstrap and zooming-in mid-tail regions. In addition to normal, we also include t-distribution approximation since for small samples the distinction is important (as explored in Chapter 2).

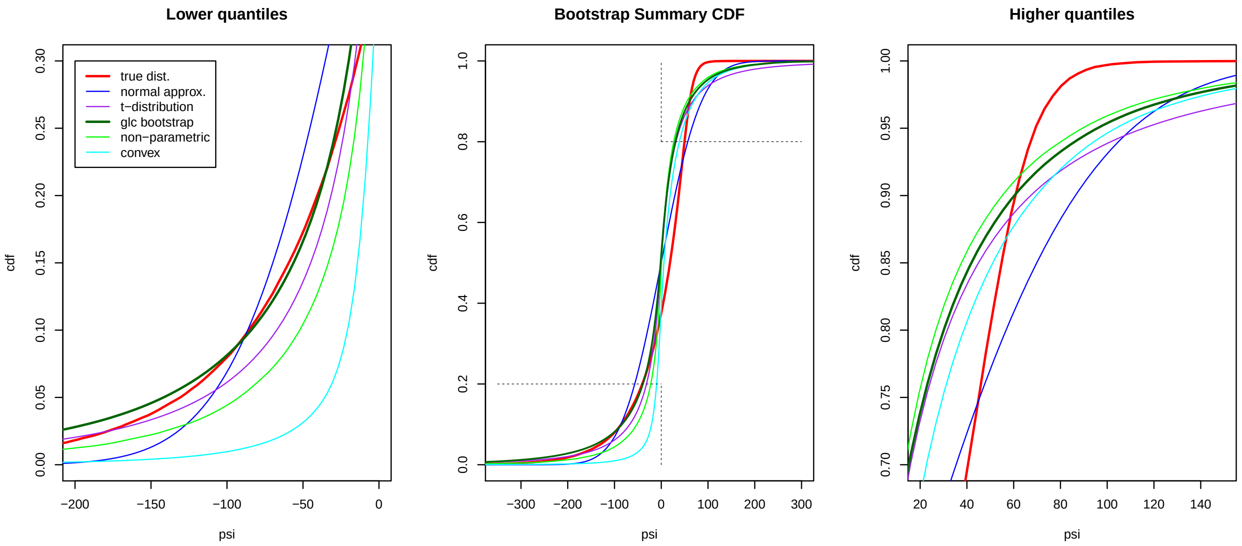


Figure 3.6: Bootstrap approximations: mid-tails of the distribution

In the next example, we explore the bias for several estimators - four central moments and tail quantiles (Table 3.1). $X \sim Exp(1)$, $n = 6$. The table includes results for studentized versions of the three bootstrap methods. For second, third, and fourth moments, the values for studentized bootstrap are wild, indicating great volatility of these estimators. The reason for that is the fact that for small sample size, bootstrap sample variance might get very small, driving up the values for studentized estimators. This continues the discussion in Chapter 2 about studentization in small samples. Table 3.2 presents "p-values" corresponding to 2.5% two-sided quantiles of the "true" sampling distribution across these bootstrap methods and t-distribution approximation (we want the numbers to be close to 0.025 but on the larger side to be conservative). GLC shows good results for both sides, as well as the regular t-distribution approximation (which seems to be quite conservative).

	0.025	0.975	mean	variance	3rd mom	4th mom
t-distribution	-0.2858	-0.0243	0	0.1052	-0.0555	1.4397
glc bootstrap	0.0139	-0.1986	0	-0.0034	-0.0174	0.0830
non-parametric	0.0712	-0.2643	0	-0.0304	-0.0256	0.0259
convex bootstrap	0.1619	-0.3889	0	-0.0759	-0.0407	-0.0507
glc Studentized	-1.6370	-0.1912	0	1.8209	-76.2534	18150.8135
np Studentized	-1.8539	-0.1947	0	36.1117	1334.9312	5.4 e^9
convex Studentized	-1.2439	-0.1758	0	1.1071	-77.2361	55222.0647
true values	-0.6320	0.9426	0	0.1664	0.0559	0.1113

Table 3.1: Quantile and central moments estimation (Exponential distribution)

	lower	upper
t-distribution	0.0719	0.0323
glc bootstrap	0.0476	0.0207
non-parametric	0.0369	0.0160
convex bootstrap	0.0191	0.0080
glc Studentized	0.1333	0.0194
np Studentized	0.1346	0.0195
convex Studentized	0.1165	0.0197
true	0.0250	0.0250

Table 3.2: Probabilities for 95% confidence interval endpoints

Another look at the lower and upper 95% bootstrapped quantiles for different methods and confidence intervals produced by these methods is presented in Figure 3.7. $X \sim Exp(0.01)$, $n = 6$, $\hat{\psi} = \bar{X}$. The true sampling distribution is skewed and we would like to see if bootstrap methods reflect that. In Figure 3.7a we can see that bootstrap averaged centered quantiles (0.025 and 0.975) indeed capture the asymmetry but fall short of reaching the true quantiles, with GLC bootstrap coming closer than the other ones. Studentized versions reach far beyond the true quantile for the thin tail while matching the results of non-studentized GLC for the thicker one. We also include the results for exact confidence intervals that use Bennett's inequality [60]. In Figure 3.7b we examine the Confidence Intervals $[l, u]$ for significance level $\alpha = 0.05$:

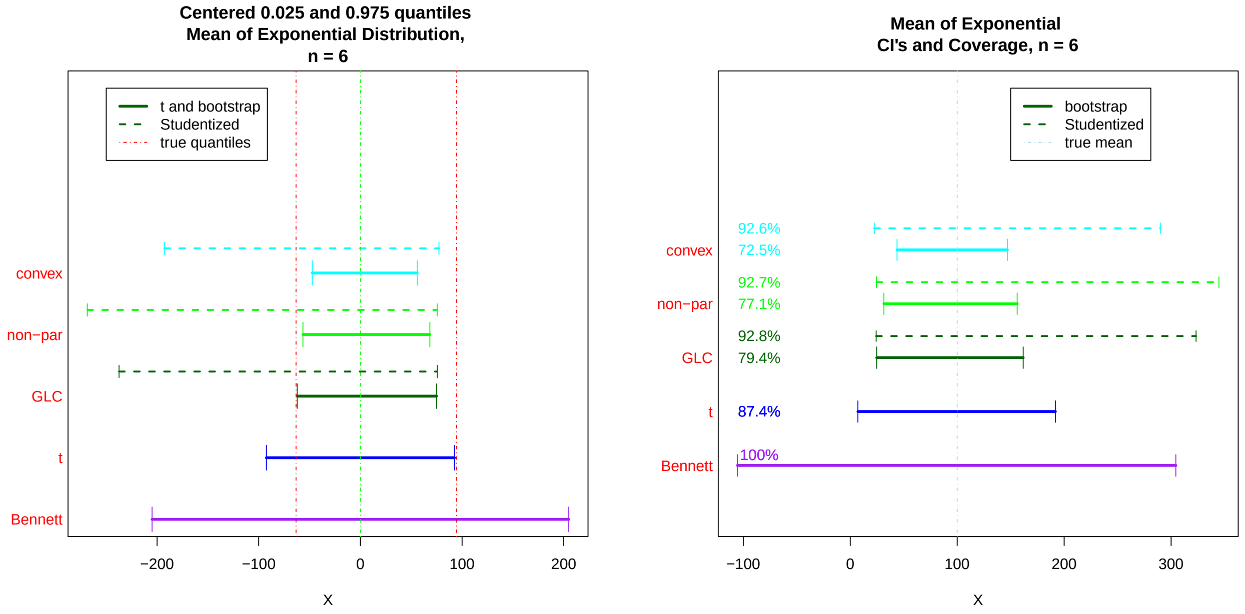
$$1 - 2\alpha = P(l \leqslant \hat{\psi} - \psi_0 \leqslant u) = P(\hat{\psi} - u \leqslant \psi_0 \leqslant \hat{\psi} - l)$$

approximated by bootstrap CI:

$$\left[\hat{\psi} - \left(\hat{\psi}^*_{[1-\alpha]} - \hat{\psi}\right),\ \hat{\psi} - \left(\hat{\psi}^*_{[\alpha]} - \hat{\psi}\right)\right], \text{ or } \left[2\hat{\psi} - \hat{\psi}^*_{[1-\alpha]},\ 2\hat{\psi} - \hat{\psi}^*_{[\alpha]}\right] \tag{3.11}$$

Note that the confidence intervals are flipped compared to the quantiles as follows from (3.11), which does not matter when methods that produce symmetric confidence intervals are used, but which is important for bootstrap and other higher-order approaches. None of the methods quite reach the nominal level of 95% except the inequality-based CI that has 100% coverage but is much wider than the other CI's.

Finally, Figure 3.8 shows coverage for different sample sizes across considered methods. Sample size significantly affects coverage probability both in normal (t-distribution) approximations and even more so in bootstrap. GLC fares slightly better than other kinds of bootstrap. Occasional fluctuations in all the lines (except inequality-based CI) are due to the randomness of simulation; they do not obscure the general trend and still provide comparison between the methods since all of them are applied to the same simulated samples.



(a) (b)

Figure 3.7: Quantiles and confidence intervals for different methods

3.2 Stabilized variance GLC bootstrap

In the previous section, we have described a mechanism that allows us to "prescribe" a variance to bootstrap samples through the tuning parameter d. So far, we have used it to get the variance of bootstrap generating distribution to match a sample variance s^2. However, when n is small, s^2 varies considerably from sample to sample. Suppose we have some "external" information about the true variance that would allow us to get closer to the true value and reduce variability of the variance estimate. A more stable robust estimate that would utilize this external information could be then "prescribed" to the bootstrap generating distribution using the same tuning parameter. In high-dimensional data, where features (such as genes) are analyzed in parallel, this information comes from the full dataset and can then be used for analysis of each individual feature.

The idea of taking advantage of the parallel structure of this kind of analysis and borrowing information from the whole collection of features (genes) for a more stable/robust inference took shape in application of the empirical Bayes methods to microarray data. Various approaches have been proposed, such as non-parametric empirical Bayes [24], [23] and parametric empirical Bayes [47] methods. Stabilized variance GLC bootstrap is based on moderated t- and F-statistics developed by G. Smyth [65] and implemented in R package *limma* [66].

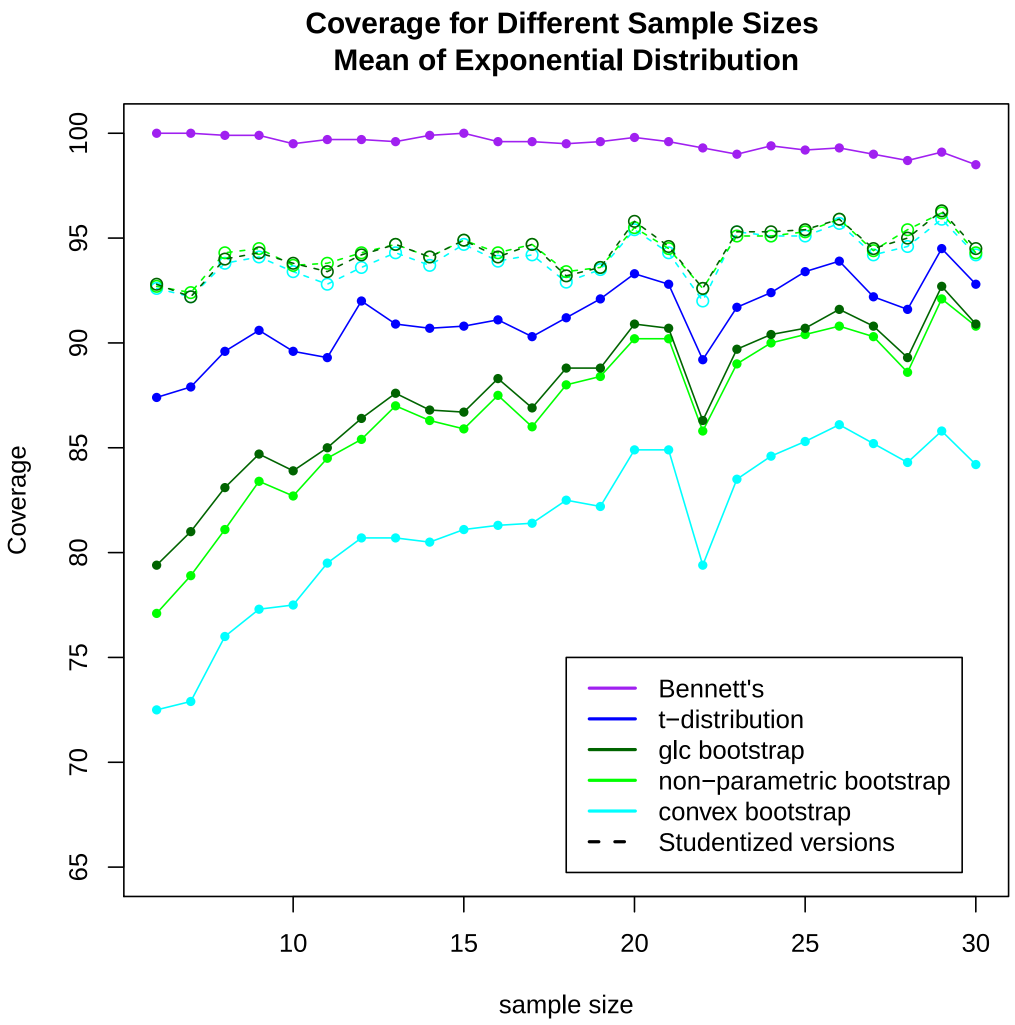


Figure 3.8: Coverage vs sample size

“Prescribed” Variance

The set up for moderated t-statistic is a hierarchical model that assumes different variances for different features and specifies the underlying distribution for these unknown variances σ_g^2 (subscript indicating a feature, e.g. a gene). Parameters of that distribution, or hyper-parameters, constitute prior information and, in contrast with fully Bayesian methods, are estimated from data - not specified upfront by the user. Only two hyperparameters df_0 and s_0^2 are estimated from the whole dataset, which makes them very stable when the number of features is large. Coefficient estimate $\hat{\beta}_g$ for a feature g is scaled by the posterior variance $\tilde{\sigma}_g^2$, which is a linear combination of the prior variance s_0^2 and residual feature-specific variance

s_g^2:

$$\tilde{\sigma}_g^2 = \frac{df_0\, s_0^2 + df_g\, s_g^2}{df_0 + df_g},$$

where df_0 and df_g are prior and residual degrees of freedom respectively (we change notation for degrees of freedom from previous chapters to df to avoid confusion with GLC bootstrap tuning parameter d). Prior estimated variance s_0^2 is usually smaller than $\frac{1}{G}\sum_{g=1}^{G} s_g^2$.

We "assign" the posterior variance $\tilde{\sigma}_g^2$ to a bootstrap random variable: $Var(X_g^*) = \tilde{\sigma}_g^2$. Let d_g be a tuning parameter for a feature g. Then

$$Var\left(X_g^*\right) = \left(\nu_g - \mu_g^2\right)\left(\frac{2}{3}\, d_g^2 - d_g + 1\right) = \tilde{\sigma_g}^2. \tag{3.12}$$

Solving (3.12) for d_g and taking the first root, we get

$$d_g = \frac{3}{4} - \sqrt{\frac{9}{16} - \frac{3}{2}\left(1 - \frac{n\tilde{\sigma_g}^2}{(n-1)\, s_g^2}\right)} \tag{3.13}$$

Now $-\infty < d_g \leqslant 3/4$. The upper bound 3/4 is introduced for situations when $\dfrac{\tilde{\sigma_g}^2}{s_g^2} \geqslant \dfrac{5\,(n-1)}{8\,n}$ (sample variance is much larger than posterior variance) and the expression under the square root is negative. In these situations we simply set $d_g = 3/4$. This rarely happens (and, as discussed in Chapter 2, is unlikely to happen for potential false positives), so the hope is that it does not introduce much bias into this working model.

In this variant of GLC bootstrap, the tuning parameter is not just a function of n - it is also a function of hyperparameters and, more importantly, of feature-specific s_g^2, which means that for each gene/feature g the value of d_g is different. While regular non-parametric bootstrap preserves correlation (the proof is below), this set-up would not. What if features are correlated and we need to preserve that correlation for further analysis - for example, clustering? Note that in this case there is an apparent contradiction with the model assumptions [65]; however, even though these assumptions are not always correct or even realistic, the method is shown to perform well in practice (see also Chapter 2 for discussion on this topic).

Correlation structure

To try to preserve most of the important characteristics of the correlation structure we use the following procedure (assume each row is a feature and each column is an observation):

1. Calculate d_g for all G features;

2. Draw a pair of vectors $\mathbf{X_1}$ and $\mathbf{X_2}$ (columns);
3. Draw Z from $U(0,\, 1)$;
4. For each feature g, the bootstrap variable is

$$X_g^* = d_g Z\, X_{1g} + (1 - d_g Z)\, X_{2g}. \tag{3.14}$$

Compare sample correlation of vectors $\mathbf{x}$ and $\mathbf{y}$ (two features - rows in the dataset) with the correlation of corresponding bootstrap random variables X^* and Y^*:

$X \sim P_{x,n}$, $P(X = x_i) = \frac{1}{n}$, $i = 1, 2, \ldots, n$
$Y \sim P_{y,n}$, $P(Y = y_i) = \frac{1}{n}$, $i = 1, 2, \ldots, n$
$Z \sim U(0,1)$, $Z \perp \{X, Y\}$

$X^* = d_x Z X_1 + (1 - d_x Z) X_2$
$Y^* = d_y Z Y_1 + (1 - d_y Z) Y_2.$

Similarly to variance calculation in the general set up, we introduce some notation:

$$\mu_x \equiv \frac{1}{n}\sum_{i=1}^{n} x_i,\ \mu_y \equiv \frac{1}{n}\sum_{i=1}^{n} y_i,\ \nu_x \equiv \frac{1}{n}\sum_{i=1}^{n} x_i^2,\ \nu_y \equiv \frac{1}{n}\sum_{i=1}^{n} y_i^2,\ \nu_{xy} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i y_i.$$

Sample correlation:

$$r_{xy} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} = \frac{\nu_{xy} - \mu_x\mu_y}{\sqrt{(\nu_x - \mu_x^2)(\nu_y - \mu_y^2)}}$$

Bootstrap random variable correlation:

$$corr(X^*, Y^*) = \frac{Cov(X^*, Y^*)}{\sqrt{Var(X^*)Var(Y^*)}} = \frac{E(X^*Y^*) - EX^*EY^*}{\sqrt{Var(X^*)Var(Y^*)}}$$

$$\begin{aligned}
E(X^*Y^*) &= E\Big[\big(d_x Z X_1 + (1 - d_x Z)X_2\big)\big(d_y Z Y_1 + (1 - d_y Z)Y_2\big)\Big] \\
&= E\Big[d_x d_y Z^2 X_1 Y_1 + Z(1 - d_x Z)d_y X_2 Y_1 \\
&\quad + d_x Z(1 - d_y Z)X_1 Y_2 + (1 - d_x Z)(1 - d_y Z)X_2 Y_2\Big] \\
&= \frac{1}{3}\, d_x d_y \nu_{xy} E\big(Z^2\big) + d_y \mu_x \mu_y \left(\frac{1}{2} - \frac{1}{3}\, d_x\right) \\
&\quad + d_x \mu_x \mu_y \left(\frac{1}{2} - \frac{1}{3}\, d_y\right) + \nu_{xy}\left(1 - \frac{1}{2}\, d_x - \frac{1}{2}\, d_y + \frac{1}{3}\, d_x d_y\right) \\
&= \nu_{xy}\left[\frac{2}{3}\, d_x d_y - \frac{1}{2}\, d_x - \frac{1}{2}\, d_y + 1\right] - \mu_x\mu_y\left[\frac{2}{3}\, d_x d_y - \frac{1}{2}\, d_x - \frac{1}{2}\, d_y\right]
\end{aligned}$$

$$Cov(X^*, Y^*) = E(X^*Y^*) - \mu_x\mu_y = (\nu_{xy} - \mu_x\mu_y)\left[\frac{2}{3}\,d_xd_y - \frac{1}{2}\,d_x - \frac{1}{2}\,d_y + 1\right] \tag{3.15}$$

$$\begin{aligned}
corr(X^*, Y^*) &= \frac{(\nu_{xy} - \mu_x\mu_y)\left[\frac{2}{3}\,d_xd_y - \frac{1}{2}\,d_x - \frac{1}{2}\,d_y + 1\right]}{\sqrt{(\nu_x - \mu_x^2)\left[\frac{2}{3}\,d_x^2 - d_x + 1\right](\nu_y - \mu_y^2)\left[\frac{2}{3}\,d_y^2 - d_y + 1\right]}} \\
&= \underbrace{\frac{\nu_{xy} - \mu_x\mu_y}{\sqrt{(\nu_x - \mu_x^2)(\nu_y - \mu_y^2)}}}_{r_{xy}}\;\underbrace{\frac{\frac{2}{3}\,d_xd_y - \frac{1}{2}\,d_x - \frac{1}{2}\,d_y + 1}{\sqrt{\left[\frac{2}{3}\,d_x^2 - d_x + 1\right]\left[\frac{2}{3}\,d_y^2 - d_y + 1\right]}}}_{\text{d-factor}} \\
&= r_{xy} \cdot \text{d-factor}
\end{aligned} \tag{3.16}$$

To see how much of the correlation structure is preserved, we would need to examine "d-factor" and assess how close its value is to 1 in various scenarios. It is a function of d_x and d_y and its values are plotted against these tuning parameters in Figure 3.9, which shows that "d-factor" is above 0.75 for the most probable ranges of d_x and d_y, and for many combinations it is in fact above 0.95. Expression (3.16) also shows that for non-parametric bootstrap ($d = 0$) and for cases when $d_x = d_y$, d-factor = 1 and so the correlation structure is completely preserved. Then we would also like to know how d-factor varies depending on the true and sample correlation - in other words, will highly correlated features retain their relationship better than uncorrelated ones in the bootstrap sample? This issue is explored in Figure 3.10. In this example, for each value of correlation $\rho_{X,Y}$, a number of simulations are performed. In each simulation, a pair of samples ($n = 6$) is drawn from multivariate normal distribution with this given correlation, and then d-factor and their sample correlation r_{xy} are calculated. It apprears that for correlated features, d-factor is reasonably close to 1 (red and purple lines indicate its mean and median across the simulations) and thus their correlation structure is mainly preserved. These pairs of features are usually the ones of interest. For highly correlated samples, d_x and d_y are also correlated, which drives d-factor up (green line). Light blue line reflects how variable sample correlation is depending on the original "true" correlation; it predictably shows that the less the absolute value of this true correlation is, the more variability there is in the sample correlation.

Simulations for high-dimensional data

Coverage and performance measures

In high-dimensional setting, we look at the coverage probability using three different approaches (Figure 3.11): quantiles yielded by bootsrapped sampling distribution (green-colored boxplots), normal approximation with bootstrap-estimated variance (blue-colored

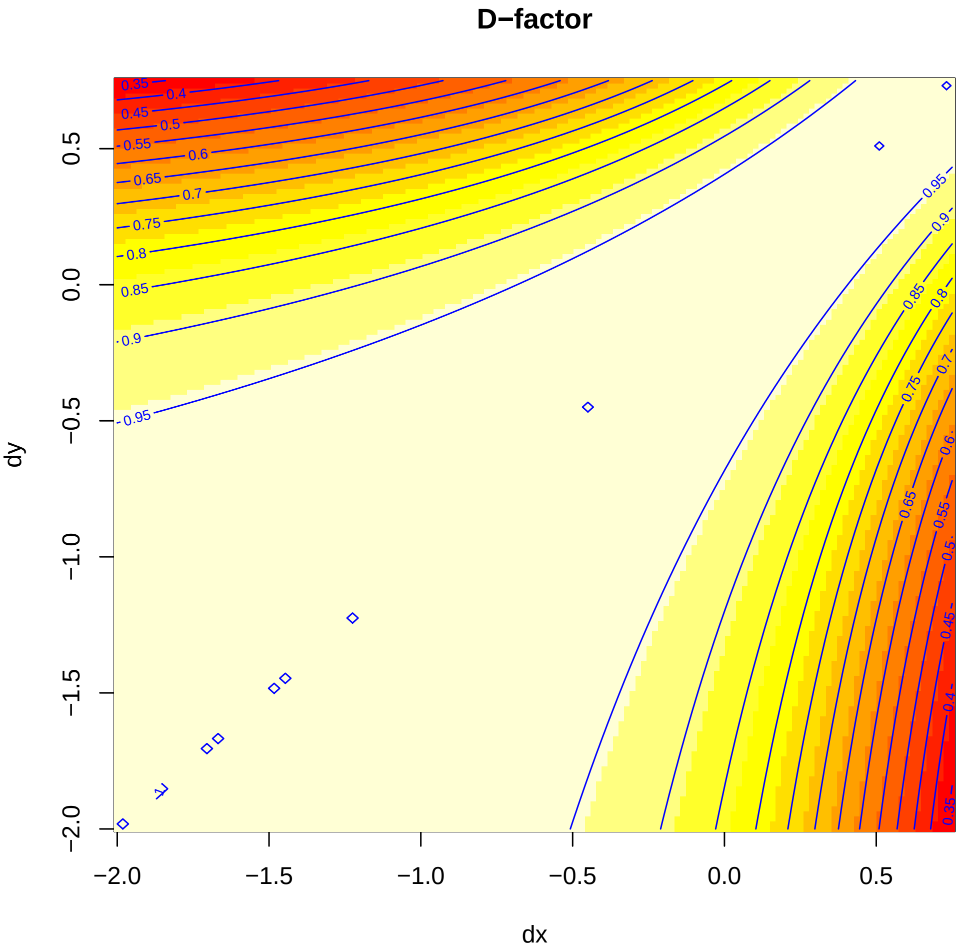


Figure 3.9: Values for correlation "d-factor" as a function of tuning parameters d_x and d_y for two correlated features x and y

boxplots), and t-distribution approximation with bootstrap-estimated variance (yellow boxplots). For comparison, we also include inequalities-based methods, and regular and moderated t-statistic approximations (with augmented degrees of freedom, df_0+df_g, for moderated t-statistic). Bootstrap methods include regular GLC, stabilized variance GLC (marked as GLC limma), non-parametric, and convex. For this and the next two figures, simulated data are generated with a convex pseudo-data algorithm applied to the real age data (similar to the procedure used for some simulations in Chapter 2). Interestingly, the mean coverage for moderated t-statistic is about the same (even a little higher) than for regular t, though the probabilities are more spread out. This provides a clue for the performance of GLC-limma method: the results are more spread out as well and the graph shows improved coverage compared to the regular GLC. As with one-dimensional case, other bootstrap methods' cov-

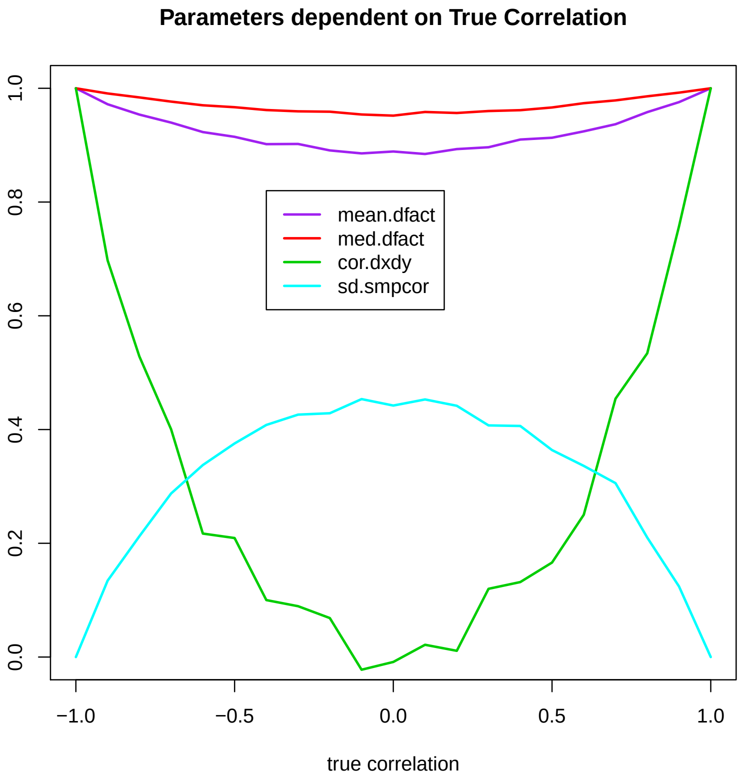


Figure 3.10: D-factor vs correlation between two features

erages are lower and there is almost 100% coverage with inequality-based methods.

Next, we want to compare type I and type II error rates across the methods described above; Figure 3.12 displays power and False Discovery Rate for each one. In general, the points for best performance would be located in the upper left region of the graph. Stabilized GLC outperforms other bootstrap methods; results for permutation and regular t are almost the same, and moderated t (empirical Bayesian) outperforms them in both power and FDR. While these non-bootstrap methods' error rate control is better than that of bootstrap analysis, the power is significantly lower. The issue of power is important for small sample sizes; for some cases a reversal of the traditional approach - "at a given type I error rate control, choose the method with the highest power" - might be considered. Instead we could formulate the task as "at a given power level choose the method that provides the smallest False Discovery Rate".

To summarize performance assessment in a single measure, we turn to Matthews Correlation Coefficient (MCC), Accuracy, and F-score:

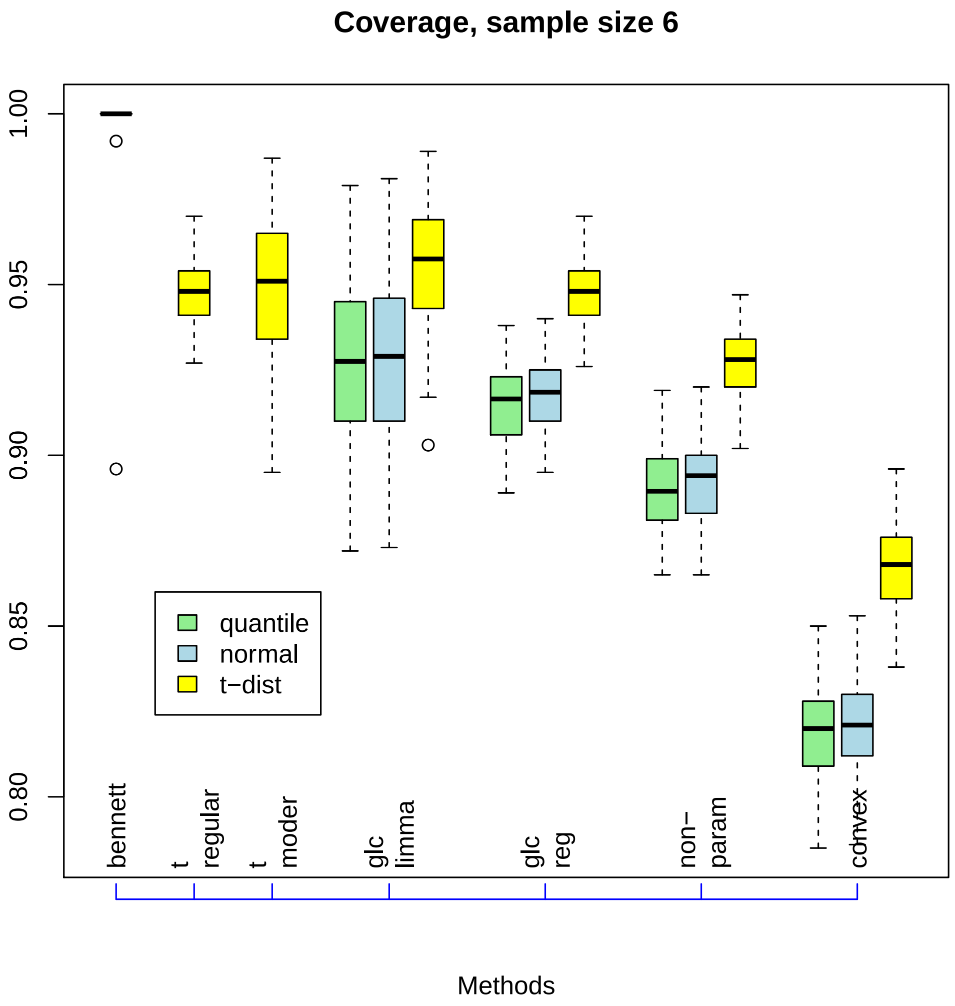


Figure 3.11: Coverage for sample mean: high-dimensional convex pseudo-data simulation

$$MCC = \frac{tp \cdot tn - fp \cdot fn}{\sqrt{(tp+fp)(tp+fn)(tn+fp)(tn+fn)}},$$

$$Acc = \frac{tp+tn}{tp+tn+fp+fn},$$

where tp, fp, tn, and fn are the number of true positives, false positives, true negatives, and false negatives respectively.

$$F = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

where $Precision = \frac{tp}{tp+fp}$ and $Recall = \frac{tp}{tp+fn}$.

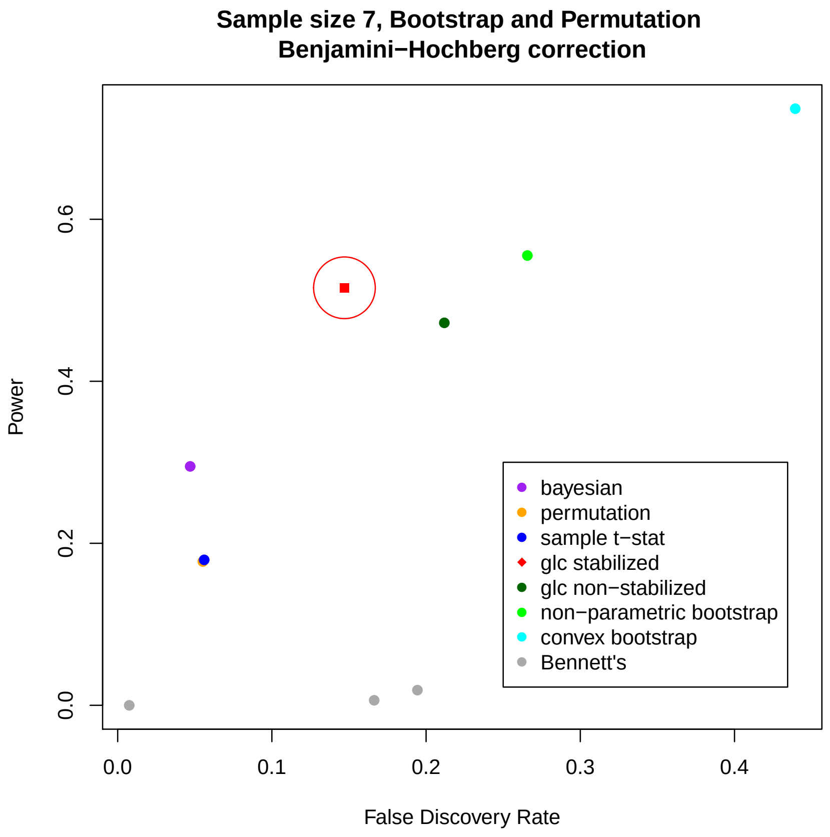


Figure 3.12: Power and FDR for log-ratio estimates: convex pseudo-data

These measures are presented in Figure 3.13 (the values are plotted along y-axis). With each assessment statistic, stabilized GLC bootstrap appears to score the highest.

Correlation Structure and Clustering

Having established theoretically that bootstrap preserves the correlation structure of data, we now explore the use of bootstrap in clustering methods, applied to datasets where features are correlated. Bootstrap is essential to obtaining inference for parameters, for which no measure of accuracy can be calculated analytically - such as cluster parameter. Resampling can provide a “degree of uncertainty” and indicate how reliable the clustering results are [56]. We implement the following measure of reliability: for each bootstrap-generated dataset $\mathbf{X}^*_\mathbf{b}$, $b = 1, \ldots, B$, the clustering procedure is performed and for each pair (u, v) of elements/features, the event $A^{(b)}_{uv} = I\,(u \text{ and } v \text{ are in the same cluster})$ is recorded [68]. Bootstrap probability $P_{uv} = \frac{1}{B}\sum_{b=1}^{B} A^{(b)}_{uv}$ of these elements being in the same cluster indicates reliability of the clustering result for the pair. This method will be used later for data analysis (section 3.3).

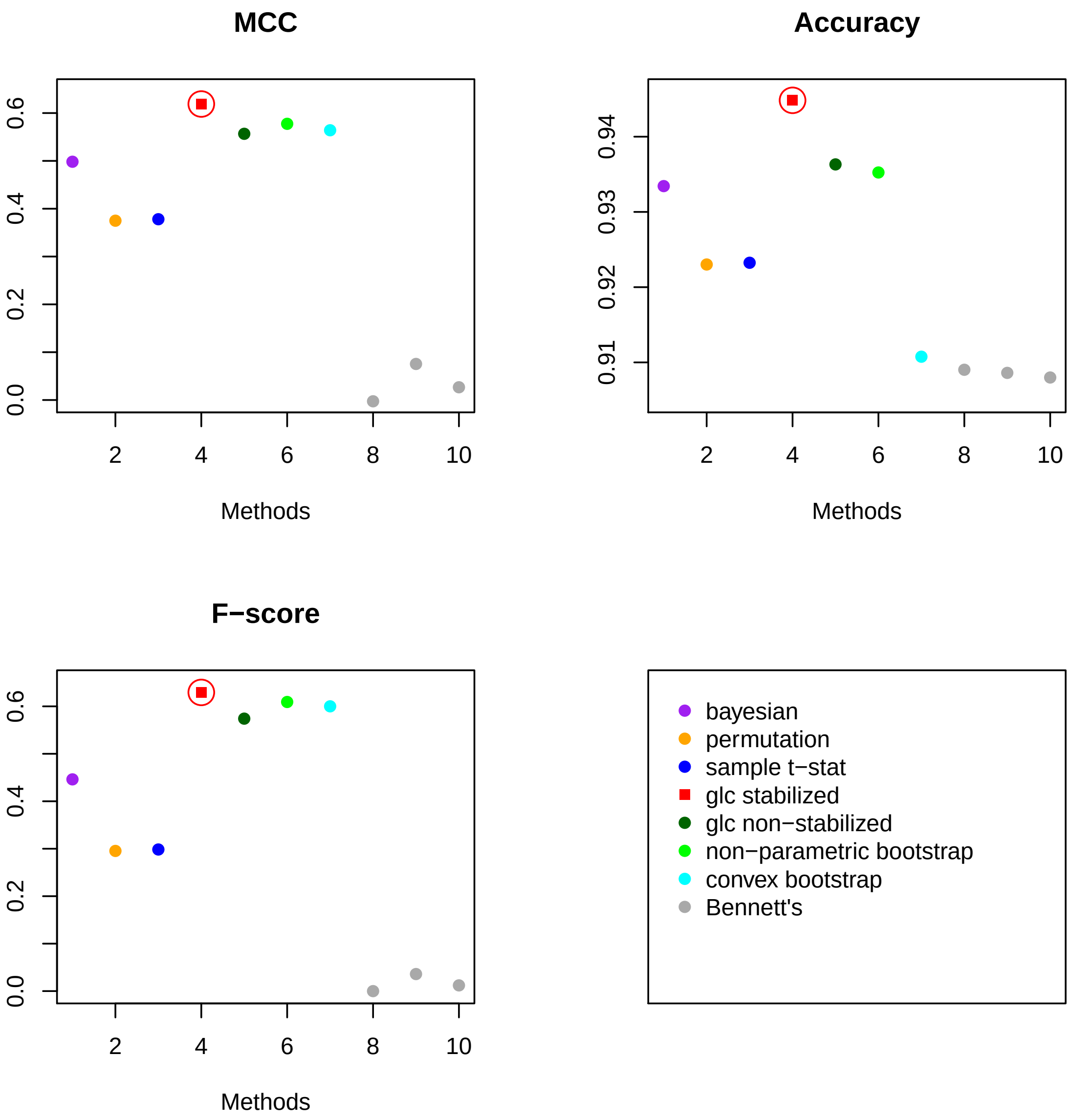


Figure 3.13: Performance measures for different methods

For the example in Figure 3.14, we simulate 30 elements from a multivariate normal distribution with provided covariance matrix (top left). This covariance matrix is the unknown "truth" that the analysis is trying to uncover from the sample. The other three graphs are based on one sample: distance matrix, adjacency matrix of original PAM clustering (1 for each pair of elements if they are in the same cluster and 0 otherwise), and GLC bootstrap-based probability. While neither distance matrix, nor adjacency matrix resemble covariance matrix very closely, the bootstrap seems to recover more of the underlying correlation structure and get closer to the truth.

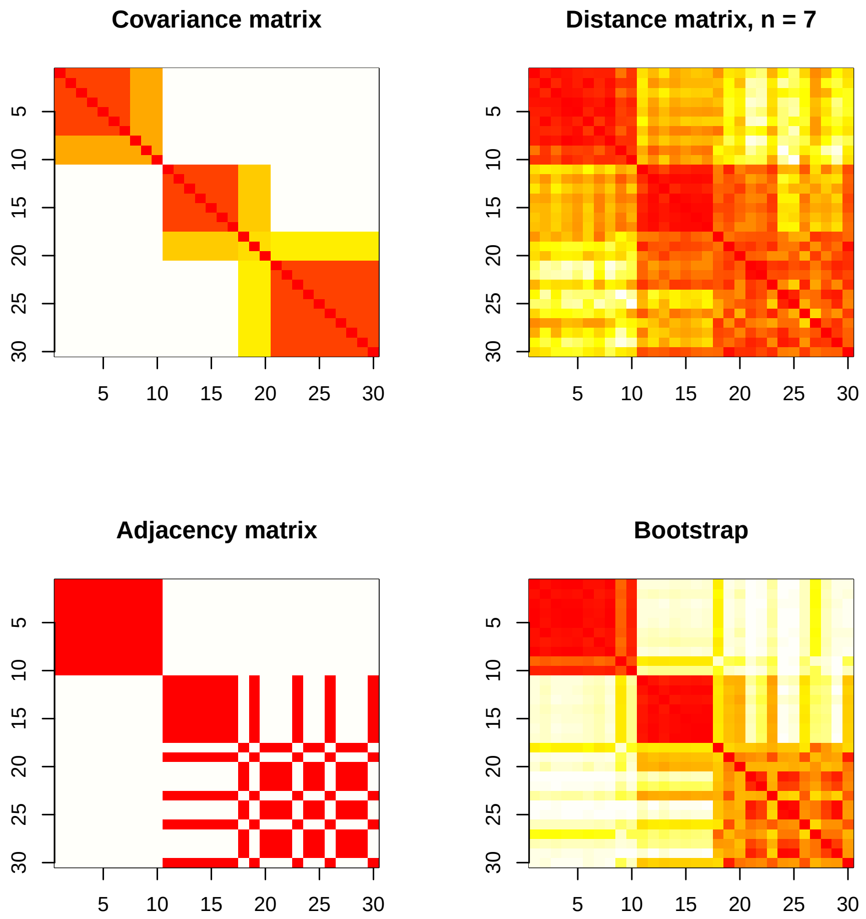


Figure 3.14: Clustering and bootstrap: multivariate normal simulation. The truth is represented by the covariance matrix, which is estimated by the matrix of bootstrap pairwise probabilities.

Figure 3.15 illustrates another clustering scenario: data is simulated using mixed effects model. There are 40 elements, $n = 7$, and in this setting there are true cluster memberships for each element with the number of clusters $K = 5$. GLC, non-parametric, and convex bootstrap results are included. Again, bootstrap methods seem to get closer to the true clusters; all three methods produce similar results with GLC bootstrap yielding slightly lower probabilities for some pairs and thus, perhaps, separating true clusters marginally better.

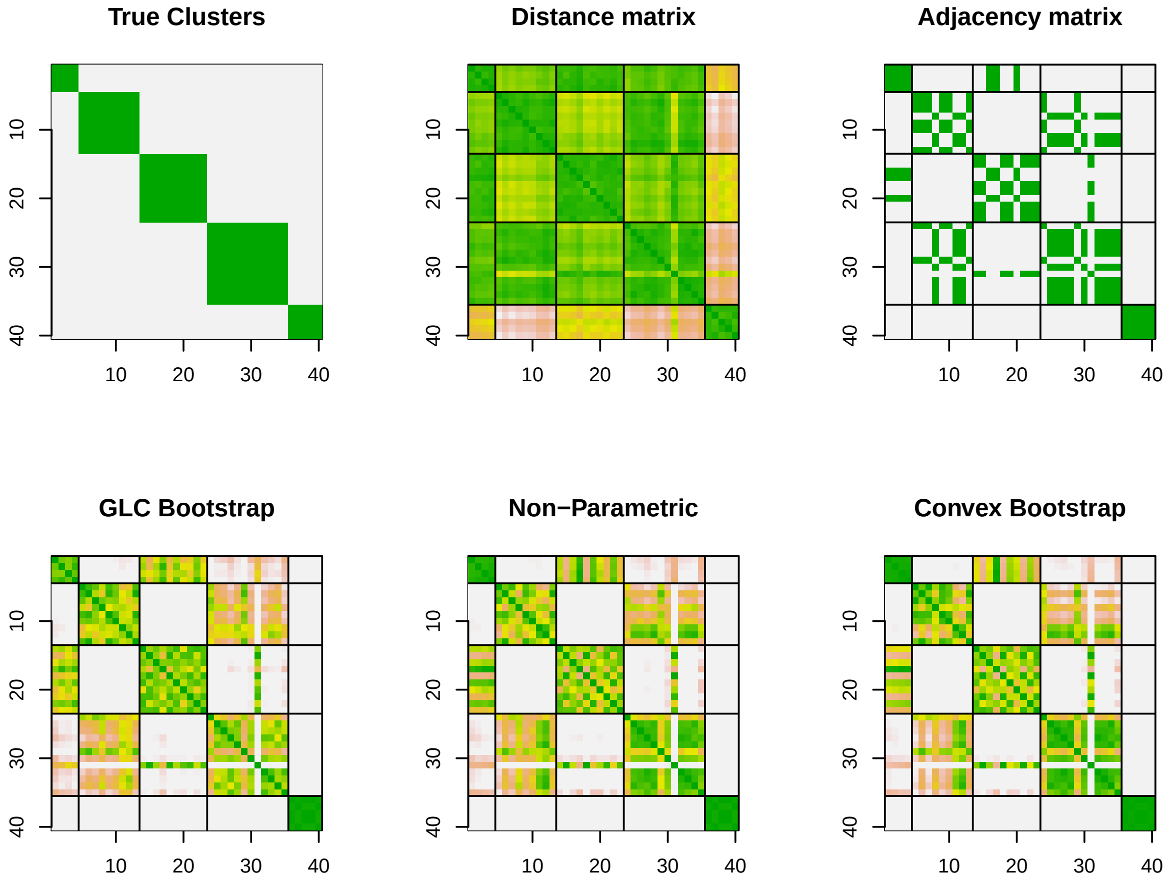


Figure 3.15: Clustering and bootstrap: mixed effects model simulation. Bootstrap pairwise probability matrices (GLC, non-parametric, and convex) recover some of the information about true clusters.

3.3 Data analysis

In this section, we present the experimental data from the study conducted by Chris Vulpe's Toxico-Genomic laboratory at UC Berkeley. The study became a motivating example for this work; it produced small sample high-dimensional data and presented complex questions about functional genomics that posed specific challenges for statistical analysis, requiring multi-step analysis procedure and novel inference approaches [51]. The experiments involve tagged mutant deletion strains of *Saccharomyces cerevisiae* that are grown together in selected toxic conditions and assayed with custom-designed molecular barcode array Tag4 [54].

Yeast PDA - functional genomics

Gene disruption is an important tool for inferring gene's function: viability of a deletion strain in various conditions (fitness profile) provides information about biological function of the corresponding gene. Each deletion strain has a precisely generated null mutation (a deletion of one gene from start to stop codon). That gene is replaced with a kanamycin-resistance gene (KanMX4) and unique molecular barcodes, or "tags", that label the strain. The development of a mutant collection that would allow the strains to be pooled and analyzed in parallel in competitive growth conditions was a significant break-through that opened new possibilities for experiments and analysis. This collection became a basis for numerous experiments aimed at uncovering new functional relationships between genes and deepening understanding about their biological functions.

A first pilot study involving 11 mutant strains was published in 1996 [63], just a few months after yeast genome sequence has been made public. It described a "molecular barcoding" approach, in which each deletion strain was labeled with a unique molecular tag - a 20 base-pair sequence. These tagged strains were grown together in selective conditions and their rate of growth (surviving ability), or "fitness", was then assessed through hybridizing the tags to high-density oligonucleotide arrays, which would show a relative abundance of each strain in the pool. This barcoding approach made a whole-genome parallel analysis possible, and an organized international effort had been mounted with the goal of creating a collection of deletion strains for all annotated yeast genes. Within a matter of a few years, large-scale collections have been constructed [71], [29], which eventually included deletion-mutants of nearly all annotated genes of Saccharomyces cerviciae (96% of open reading frames). In these collections, each strain has been labeled with two molecular barcodes (UPTAG and DOWNTAG) instead of one; both are unique for this strain and provide a measure of strain's abundance (see great graphical illustrations in [61], [53], and [54]). Since the development of these collections, they have been very extensively used in experimental practice [61] and rich experimental data libraries based on Parallel Deletion Analysis are growing all the time.

Research questions and analysis procedure

Questions posed by the study can be translated into two analysis goals:

1. Determine strains that are characterized by diminished growth in the presence of a toxicant (differential strain sensitivity analysis or DSSA). The set of sensitive strains will be different for each chemical.

2. Recognize sets of genes that are functionally related (genes that act together) - find strains that display similar survival pattern, phenotype, when interrogated under various toxic conditions. This question will be approached through a clustering procedure.

The diagram in Figure 3.16 illustrates the structure of the data and steps of the statistical analysis procedure. Control group contains 12 experiments; each of the chemical compounds has 3 treatment doses with 3 experiments per dose. DSSA is performed separately for each compound with control group (the same control group is used for all the chemicals but normalized results are different each time because control and treatment groups are normalized together). DSSA for each toxicant produces a set of significant (sensitive) genes; these genes are eventually combined in a union set that contains all the strains that were determined to be sensitive in at least one growth condition. A special distance metric is devised to compare the strains across the chemical compouns, which is then used to cluster the strains (genes). Hierarchical Ordered Partitioning and Collapsing Hybrid (HOPACH) clustering algorithm [68] is used for clustering procedure.

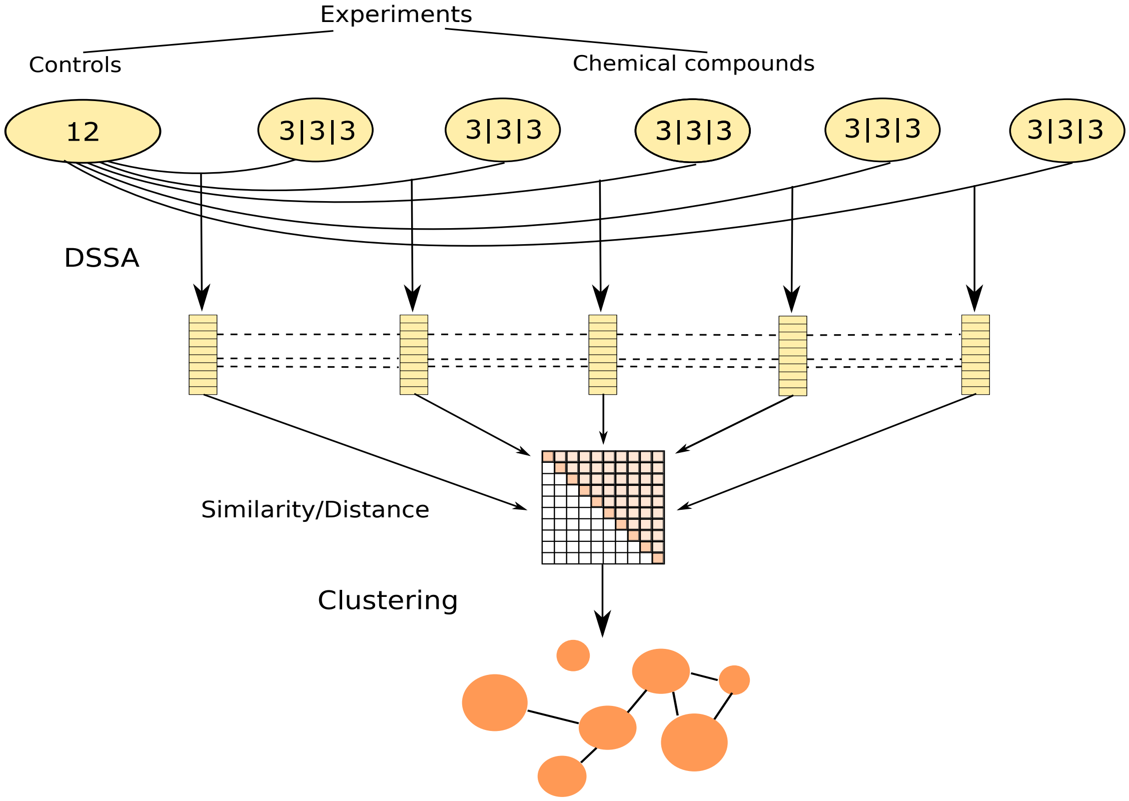


Figure 3.16: Yeast PDA data: steps of statistical analysis

It would be desirable to get some measure of accuracy for the final results, such as reliability of the clusters (as discussed in section 3.2): to find out how often two genes are clustered together, which also answers the second scientific question about functional relationships between the genes; this approach can be applied more generally to categorization problems and methods including trees (recursive partitioning) and networks. GLC bootstrap addresses the challenges of the statistical analysis of this study - small sample size and obtaining inference for parameter that does not have an analytical expression for variability - while also incorporating multi-step analysis procedure. To perform bootstrap in this case, we need to start from the original data and repeat all the steps of the analysis with each bootstrap sample (to be precise, this means a set of bootstrapped datasets of each group of controls and one chemical compound). Thus, bootstrap assesses reliability of not just clustering procedure, but all the intermediate steps, including estimated distances between the genes. Note that for each set of bootstrapped samples, there will be a variation in the resulting set of sensitive genes that participate in the clustering.

Final results

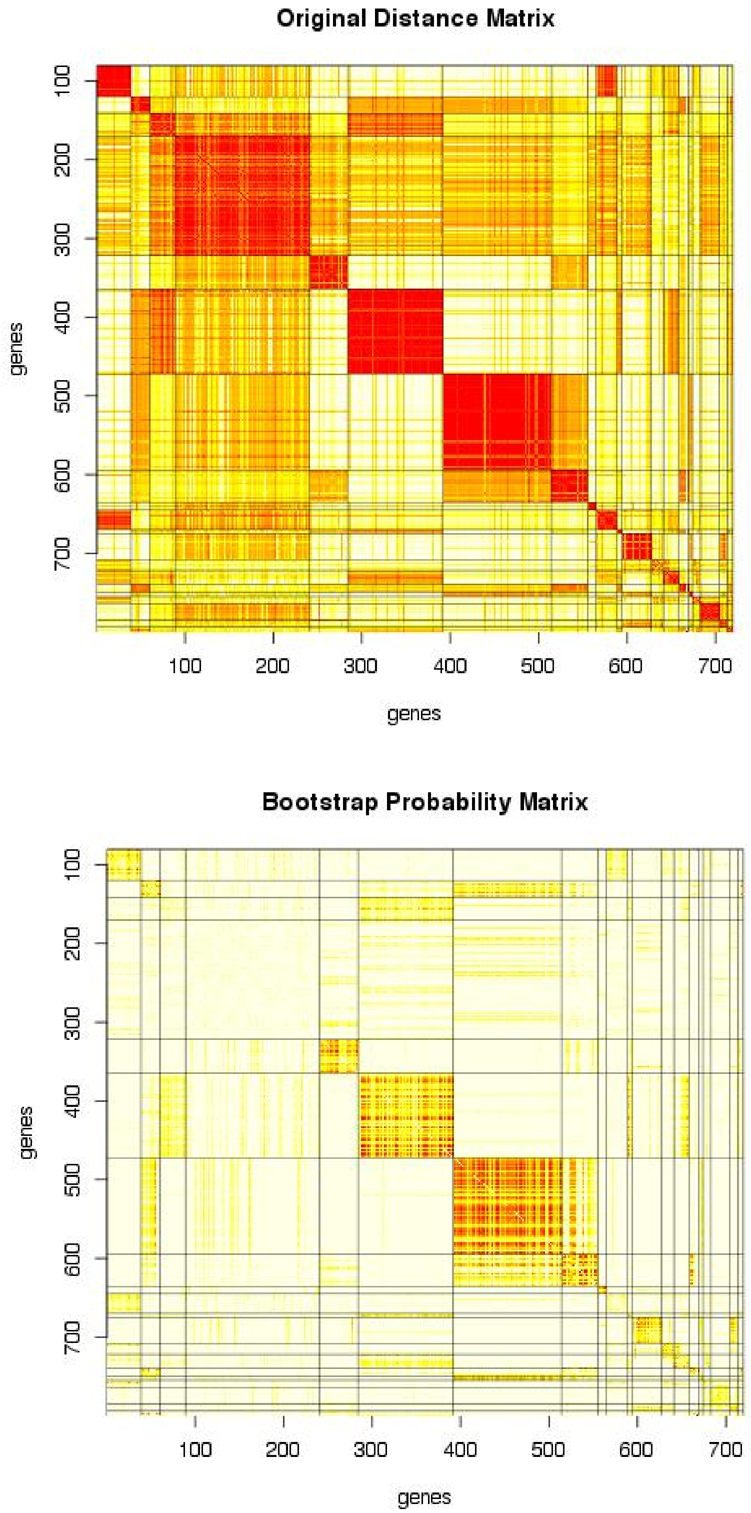
Figure 3.17 presents final bootstrap clustering results: original clusters and distances between the genes (top) and estimated accuracy of the clusters and strength of relationship between each pair of genes (bottom). The results are displayed only for the genes that ended up in a final set that was clustered in the processing of the original data. These results are quite interesting: while some of the clusters show high reliability and strong persistent pair relationships, others seem to be much weaker and a few do not show any signal at all, including a big cluster in the upper left region of the graph, indicating that these clusters are not reliable. There was no way of deducing these conclusions from the original results without bootstrapping all the steps of the analysis.

3.4 Conclusion

The set up for general linear combination bootstrap is convenient and flexible enough to allow various implementations suited for different scenarios through the use of a tuning parameter. GLC bootstrap provides a way to use resampling-based inference when the sample size is small, breaking the discreteness of the non-parametric bootstrap and eliminating bias in bootstrap variance. With added challenge of high-dimensionality, where empirical Bayes methods can increase power and improve error rate control, the GLC set up allows these methods to be integrated with the bootstrap, combining the advantages of both approaches. We have demonstrated that GLC bootstrap methods, including the stabilized variance version, produce better approximations to the sampling distribution, increase coverage for confidence intervals, and score well compared to other methods according to various performance measures in terms of power and error rate control. However, these advantages apply to estimators that are not studentized; when the sample size is small, studentized bootstrap appears to be unstable for all the considered methods and therefore is not recommended. Our results also show that high-dimensional data analysis benefits considerably from the use of stabilized variance bootstrap (GLC-limma version).

With the proposed algorithm for preserving correlation structure of the bootstrap sample, stabilized variance GLC bootstrap can be used to assess reliability of clustering results. Moreover, bootstrapping can be applied to multi-step analysis thus accommodating a wide range of procedures; removing bias in bootstrap variance might be crucial for some of the steps in that analysis. While pseudo-data generation such as convex bootstrap can be used for analysis of stand-alone categorization problems, it might not perform adequately when categorization/clustering is a part of a larger multi-step procedure (the data analysis section presents one such example). Previous steps, such as DSSA or gene differential expression analysis would suffer from variance bias; in these cases, GLC bootstrap allows estimation of all the steps of the process to be conducted without distortion of the results.

Adding to the long list of bootstrap methods, GLC bootstrap addresses a specific challenge of non-parametric resampling-based inference in small samples. It can be applied to continuous data only; this approach, however, might be generalized to other types of data if an appropriate data transformation is performed (for example, influence curves are calculated for all the observations). The method can also be used for exploration and simulation purposes when predetermined variance is desired for the samples. To assure accessibility of GLC bootstrap for wide variety of uses, its user-friendly implementation will be available as a part of a software package.



Original Distance Matrix

Bootstrap Probability Matrix

Figure 3.17: Data analysis results

Bibliography

[1] Willem Albers, Peter J Bickel, and Willem R van Zwet. “Asymptotic expansions for the power of distribution free tests in the one-sample problem”. In: *The Annals of Statistics* (1976), pp. 108–156.

[2] Niall H Anderson, Peter Hall, and DM Titterington. “Edgeworth expansions in very-high-dimensional problems”. In: *Journal of statistical planning and inference* 70.1 (1998), pp. 1–18.

[3] Gutti Jogesh Babu and ZD Bai. “Edgeworth expansions of a function of sample means under minimal moment conditions and partial Cramér’s condition”. In: *Sankhyā: The Indian Journal of Statistics, Series A* (1993), pp. 244–258.

[4] ZD Bai and C Radhakrishna Rao. “Edgeworth expansion of a function of sample means”. In: *The Annals of Statistics* (1991), pp. 1295–1315.

[5] Patrice Bertail and Stéphan Clémençon. “Edgeworth expansions of suitably normalized sample mean statistics for atomic Markov chains”. In: *Probability theory and related fields* 130.3 (2004), pp. 388–414.

[6] Rabi N Bhattacharya and Jayanta K Ghosh. “On the validity of the formal Edgeworth expansion”. In: *The Annals of Statistics* (1978), pp. 434–451.

[7] Rabindra N Bhattacharya and Ramaswamy Ranga Rao. *Normal approximation and asymptotic expansions*. Vol. 64. SIAM, 1986.

[8] PJ Bickel. “Edgeworth expansions in nonparametric statistics”. In: *The Annals of Statistics* (1974), pp. 1–20.

[9] PJ Bickel, F Götze, and WR Van Zwet. “The Edgeworth expansion for U-statistics of degree two”. In: *The Annals of Statistics* (1986), pp. 1463–1484.

[10] PJ Bickel, WR van Zwet, et al. “Asymptotic Expansions for the Power of Distributionfree Tests in the Two-Sample Problem”. In: *The Annals of Statistics* 6.5 (1978), pp. 937–1004.

[11] Sergei Blinnikov and Richhild Moessner. “Expansions for nearly Gaussian distributions”. In: *Astronomy and Astrophysics Supplement Series* 130.1 (1998), pp. 193–205.

[12] M Bloznelis and H Putter. “One term Edgeworth expansion for Student’st statistic”. In: *Probability Theory and Mathematical Statistics: Proceedings of the Seventh Vilnius Conference*. Vilnius, Utrecht: VSP/TEV. 1999, pp. 81–98.

[13] Leo Breiman. “Using convex pseudo-data to increase prediction accuracy”. In: *breast (Wis)* 699.9 (1998), p. 2.

[14] Herman Callaert, Paul Janssen, and Noel Veraverbeke. “An Edgeworth expansion for U-statistics”. In: *The Annals of Statistics* (1980), pp. 299–312.

[15] Michael R Chernick. *Bootstrap methods: A guide for practitioners and researchers*. Vol. 619. John Wiley &amp; Sons, 2011.

[16] MR Chernick, VK Murthy, and CD Nealy. “Application of bootstrap and other resampling techniques: evaluation of classifier performance”. In: *Pattern Recognition Letters* 3.3 (1985), pp. 167–178.

[17] MR Chernick, VK Murthy, and CD Nealy. “Estimation of error rate for linear discriminant functions by resampling: Non-Gaussian populations”. In: *Computers &amp; Mathematics with Applications* 15.1 (1988), pp. 29–37.

[18] Kai-Lai Chung. “The approximate distribution of Student’s statistic”. In: *The Annals of Mathematical Statistics* (1946), pp. 447–465.

[19] Harald Cramér. *Mathematical Methods of Statistics (PMS-9)*. Vol. 9. Princeton university press, 2016.

[20] Harald Cramér. “On the composition of elementary errors: First paper: Mathematical deductions”. In: *Scandinavian Actuarial Journal* 1928.1 (1928), pp. 13–74.

[21] Anthony Christopher Davison and David Victor Hinkley. *Bootstrap methods and their application*. Vol. 1. Cambridge university press, 1997.

[22] B Efron. “Bootstrap methods: Another look at the jackknife”. In: *Annals of Statistics* 7 (1979), pp. 1–26.

[23] Bradley Efron et al. “Robbins, empirical Bayes and microarrays”. In: *The annals of Statistics* 31.2 (2003), pp. 366–378.

[24] Bradley Efron et al. “Empirical Bayes analysis of a microarray experiment”. In: *Journal of the American statistical association* 96.456 (2001), pp. 1151–1160.

[25] Christopher J Ferguson and Moritz Heene. “A vast graveyard of undead theories publication bias and psychological science’s aversion to the null”. In: *Perspectives on Psychological Science* 7.6 (2012), pp. 555–561.

[26] Yasunori Fujikoshi. “An Asymptotic Expansion for the Distribution of Hotelling’sT 2-Statistic under Nonnormality”. In: *Journal of Multivariate Analysis* 61.2 (1997), pp. 187–193.

[27] K Fung. “Google Flu Trends’ failure shows good data> big data”. In: *Harvard Business Review* (2014).

[28] Andrew Gelman. “Commentary: P values and statistical practice”. In: *Epidemiology* 24.1 (2013), pp. 69–72.

[29] Guri Giaever et al. “Functional profiling of the Saccharomyces cerevisiae genome”. In: *nature* 418.6896 (2002), pp. 387–391.

[30] Minggao Gu. “On the Edgeworth expansion and bootstrap approximation for the Cox regression model under random censorship”. In: *Canadian Journal of Statistics* 20.4 (1992), pp. 399–414.

[31] Peter Hall et al. “Edgeworth expansion for Student’s t statistic under minimal moment conditions”. In: *The Annals of Probability* 15.3 (1987), pp. 920–931.

[32] Peter Hall. “On the relative performance of bootstrap and Edgeworth approximations of a distribution function”. In: *Journal of Multivariate Analysis* 35.1 (1990), pp. 108–129.

[33] Peter Hall. *The bootstrap and Edgeworth expansion*. Springer Science & Business Media, 2013.

[34] Peter Hall, Michael A Martin, and Shan Sun. “Monte Carlo approximation to Edgeworth expansions”. In: *The Canadian Journal of Statistics/La Revue Canadienne de Statistique* (1999), pp. 579–584.

[35] Tim Harford. *Big data: are we making a big mistake? FT Magazine*. 2014.

[36] Roelof Helmers. “On the Edgeworth expansion and the bootstrap approximation for a Studentized U-statistic”. In: *The Annals of Statistics* (1991), pp. 470–484.

[37] Harrie Hendriks, Pieta C IJzerman-Boon, Chris AJ Klaassen, et al. “Student’s t-Statistic under Unimodal Densities”. In: *Austrian journal of statistics= Österreichische Zeitschrift für Statistik* 35.2&amp;3 (2006), pp. 131–141.

[38] John PA Ioannidis. “Why most published research findings are false”. In: *PLoS Med* 2.8 (2005), e124.

[39] David Joyner et al. “Open source computer algebra systems: SymPy”. In: *ACM Communications in Computer Algebra* 45.3/4 (2012), pp. 225–234.

[40] Yoshihide Kakizawa. “Valid Edgeworth Expansions of Some Estimators and Bootstrap Confidence Intervals in First-order Autoregression”. In: *Journal of Time Series Analysis* 20.3 (1999), pp. 343–359.

[41] Willibrordes Cornelis Maria Kallenberg. “Interpretation and manipulation of Edgeworth expansions”. In: *Annals of the Institute of Statistical Mathematics* 45.2 (1993), pp. 341–351.

[42] Yutaka Kano. “An asymptotic expansion of the distribution of Hotelling’s T 2-statistic under general distributions”. In: *American Journal of Mathematical and Management Sciences* 15.3-4 (1995), pp. 317–341.

[43] SN Lahiri et al. "Edgeworth expansions for studentized statistics under weak dependence". In: *The Annals of Statistics* 38.1 (2010), pp. 388–434.

[44] Soumendra Nath Lahiri. "On Edgeworth Expansion and Moving Block Bootstrap for StudentizedM-Estimators in Multiple Linear Regression Models". In: *Journal of Multivariate Analysis* 56.1 (1996), pp. 42–59.

[45] David Lazer et al. "Google Flu Trends still appears sick: An evaluation of the 2013-2014 flu season". In: *Available at SSRN 2408560* (2014).

[46] David Lazer et al. "The parable of Google Flu: traps in big data analysis". In: *Science* 343.14 March (2014).

[47] Ingrid Lönnstedt and Terry Speed. "Replicated microarray data". In: *Statistica sinica* (2002), pp. 31–46.

[48] Enno Mammen. *When does bootstrap work?: asymptotic results and simulations*. Vol. 77. Springer Science &amp; Business Media, 2012.

[49] Gary Marcus and Ernest Davis. "Eight (no, nine!) problems with big data". In: *The New York Times* 6.04 (2014), p. 2014.

[50] Per Aslak Mykland. "Asymptotic expansions for martingales". In: *The Annals of Probability* (1993), pp. 800–818.

[51] Matthew North et al. "Genome-wide functional profiling reveals genes required for tolerance to benzene metabolites in yeast". In: *PloS one* 6.8 (2011), e24205.

[52] Valentin Petrov. *Sums of independent random variables*. Vol. 82. Springer Science & Business Media, 2012.

[53] Sarah E Pierce et al. "A unique and universal molecular barcode array". In: *Nature methods* 3.8 (2006), pp. 601–603.

[54] Sarah E Pierce et al. "Genome-wide analysis of barcoded Saccharomyces cerevisiae gene-deletion mutants in pooled cultures". In: *Nature protocols* 2.11 (2007), pp. 2958–2974.

[55] Margus Pihlak. "Using Edgeworth expansion approximating two-and three-dimensional probability distribution functions". In: ().

[56] Katherine S Pollard and Mark J van der Laan. "Statistical inference for simultaneous clustering of gene expression data". In: *Mathematical Biosciences* 176.1 (2002), pp. 99–121.

[57] Hein Putter, Willem R van Zwet, et al. "Empirical Edgeworth expansions for symmetric statistics". In: *The Annals of Statistics* 26.4 (1998), pp. 1540–1569.

[58] Maher B Qumsiyeh. "Edgeworth expansion in regression models". In: *Journal of Multivariate Analysis* 35.1 (1990), pp. 86–101.

[59] Lorenzo Rimoldini. "Weighted skewness and kurtosis unbiased by sample size and Gaussian uncertainties". In: *Astronomy and Computing* 5 (2014), pp. 1–8.

[60] Michael A Rosenblum and Mark J van der Laan. "Confidence intervals for the population mean tailored to small sample sizes, with applications to survey sampling". In: *The international journal of biostatistics* 5.1 (2009).

[61] Bart Scherens and Andre Goffeau. "The uses of genome-wide yeast mutant collections". In: *Genome biology* 5.7 (2004), p. 1.

[62] DA Shaywitz. "Science and shams". In: *Boston: Globe* (2006).

[63] Daniel D Shoemaker et al. "Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar–coding strategy". In: *Nature genetics* 14.4 (1996), pp. 450–456.

[64] Ib M Skovgaard. "On multivariate Edgeworth expansions". In: *International Statistical Review/Revue Internationale de Statistique* (1986), pp. 169–186.

[65] GK Smyth. "Statistical Applications in Genetics and Molecular Biology Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microar". In: *Stat. Appl. Genet. Mol. Biol* 3.1 (2004), pp. 1–25.

[66] Gordon K Smyth. "Limma: linear models for microarray data". In: *Bioinformatics and computational biology solutions using R and Bioconductor*. Springer, 2005, pp. 397–420.

[67] Masanobu Taniguchi. "Validity of Edgeworth expansions of minimum contrast estimators for Gaussian ARMA processes". In: *Journal of Multivariate Analysis* 21.1 (1987), pp. 1–28.

[68] Mark J Van der Laan and Katherine S Pollard. "A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap". In: *Journal of Statistical Planning and Inference* 117.2 (2003), pp. 275–303.

[69] Qiying Wang and Peter Hall. "Relative errors in central limit theorems for Student's t statistic, with applications". In: *Statistica Sinica* (2009), pp. 343–354.

[70] Ronald L Wasserstein and Nicole A Lazar. "The ASA's statement on p-values: context, process, and purpose". In: *The American Statistician* (2016).

[71] Elizabeth A Winzeler et al. "Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis". In: *science* 285.5429 (1999), pp. 901–906.

[72] Ali S Yousef. "Constructing a Three-Stage Asymptotic Coverage Probability for the Mean Using Edgeworth Second-Order Approximation". In: *International Conference on Mathematical Sciences and Statistics 2013*. Springer. 2014, pp. 53–67.

[73] Dmitrii Zholud et al. "Tail approximations for the Student t-, F-, and Welch statistics for non-normal and not necessarily iid random variables". In: *Bernoulli* 20.4 (2014), pp. 2102–2130.