Bernoulli
2014, Vol. 20, No. 4, pp. 2102-2130.

Tail approximations for the Student $t-$, $F-$, and Welch statistics for non-normal and not necessarily i.i.d. random variables

Dmitrii Zholud*

Abstract

Let T be the Student one- or two-sample $t-$, $F-$, or Welch statistic. Now release the underlying assumptions of normality, independence and identical distribution and consider a more general case where one only assumes that the vector of data has a continuous joint density. We determine asymptotic expressions for $\mathbf{P}(T > u)$ as $u \to \infty$ for this case. The approximations are particularly accurate for small sample sizes and may be used, for example, in the analysis of High-Throughput Screening experiments, where the number of replicates can be as low as two to five and often extreme significance levels are used. We give numerous examples and complement our results by an investigation of the convergence speed - both theoretically, by deriving exact bounds for absolute and relative errors, and by means of a simulation study.

Keywords: Student's one- and two-sample $t-$test, $F-$test, Welch statistic, small sample size, high-throughput screening, SmartTail, FDR, pFDR, test power, non-normal population distribution, systematic effects, outliers, non-homogeneous data, dependent random variables.

AMS 2010 Subject Classifications: $\frac{\text{Primary-62G32;}}{\text{Secondary-62P10, 60F10;}}$.

1 Introduction

This article extends early results of (Bradley, 1952a) and (Hotelling, 1961) on the tails of the distributions of some popular and much used test statistics. We quantify the effect of non-normality, dependence, and non-homogeneity of data on the tails of the distribution of the Student one- and two-sample $t-$, $F-$ and Welch statistics. Our approximations are valid for samples of any size, but are most useful for very small sample sizes, e.g. when standard central limit theorem-based approximations are inapplicable.

1.1 Problem statement and main result

Let $\mathbf{X} \in \mathbb{R}^n$, $n \geq 2$, be a random vector and $T = T_n(\mathbf{X})$ be (i) the Student one-sample $t-$test statistic; or (ii) the Student two-sample $t-$test statistic; or (iii) the $F-$test statistic for comparison of variances (in fact the $F-$test results apply also to one-way ANOVA, factorial designs, a lack-of-fit sum of squares test, and an $F-$test for comparison of two nested linear models).

In this paper we study the asymptotic behavior of the tail distribution of T for small and fixed sample sizes. Let $g_0(\mathbf{x})$ be the true joint density of $\mathbf{X}$ under H_0 and $g_1(\mathbf{x})$ be the density under the

*Department of Mathematical Statistics
Chalmers University of Technology and University of Göteborg, Sweden.
E-mail: dmitrii@chalmers.se

alternative H_1. Define $\mathcal{G}$ as a set of continuous densities that satisfy the regularity constraints of Theorem 2.1, 3.1, or 5.1 for the three test statistics accordingly. Our main result is

Theorem 1.1. *For any fixed value of* n *and each of the three choices of* T*, there exists a functional* $K : \mathcal{G} \to \mathbb{R}^+$*, such that for all* $g_0, g_1 \in \mathcal{G}$ *the limit expression*

$$\frac{\boldsymbol{P}(T > u|H_1)}{\boldsymbol{P}(T > u|H_0)} = \frac{K_{g_1}}{K_{g_0}} + o(1) \quad as \quad u \to \infty \tag{1}$$

holds with constants $0 < K_{g_0} = K(g_0) < \infty$ *and* $0 < K_{g_1} = K(g_1) < \infty$*. The exact expressions for* $K(g)$ *are given in (4), (10) and (18) for the three choices of the test statistic* T*.*

Remark 1. Standard assumption in the use of any of the test statistics described above is that $\mathbf{X} \sim MVN(\mathbf{0}, \sigma^2 \mathbf{1}_n)$, where $MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. It is easy to check that $MVN(\mathbf{0}, \sigma^2 \mathbf{1}_n) \in \mathcal{G}$ and that $K(MVN(\mathbf{0}, \sigma^2 \mathbf{1}_n)) = 1$.

Further remarks on Theorem 1.1 are given in Supplementary Materials.

1.2 Motivation and applications

The questions addressed in this article have gained significant new importance through the explosive increase of High-Throughput Screening (HTS) experiments, where the number of replicates is often small, but instead thousands or millions of tests are performed, at extremely high significance levels. Studying extreme tails of test statistics under deviation from standard assumptions is crucial in HTS because of the following factors:

Extreme significance levels. HTS uses many thousands or even millions of biochemical, genetic or pharmacological tests. In order to get a reasonable number of rejections, the significance level of the tests is often very small, say, 0.001 or lower, and it is the extreme tails of the distribution of test statistics which are important.

Deviation from standard assumptions. HTS assays are often subject to numerous systematic and spatial effects and to large number of preprocessing steps. The resulting data may become dependent, non-normal, or non-homogeneous, yet common test statistics such as one- and two- sample $t-$ tests are still routinely computed under standard assumptions.

Test power. It is even less likely that the data follows any standard distribution under the alternative hypothesis. By quantifying the tail behavior of a test statistic under arbitrary distributional assumptions one can get more realistic estimators for the test power.

Error-control quantities. Given the scale of HTS experiments and necessity to make even larger investments into further research on positives detected through a HTS study, it is important to have realistic picture of the accuracy of such experiments. Consider, for example, estimation of $pFDR$, the positive False Discovery Rate, see (Storey, 2002, 2003, 2004). As of now, estimates of $pFDR$ are obtained under the assumption that the true null distribution equals the theoretical one, and this may lead to wrong decisions. In most cases, however, a sample from the null distribution can be obtained by conducting a separate experiment. One can then model the tail distribution of the test statistic, and apply e.g. methods of (Rootzén and Zholud, 2014), which account for deviations from the theoretical null distribution.

Small sample sizes. Due to economical constraints, numbers of replicates in an individual experiment in HTS are as small as two to five, which makes large sample normal approximations inapplicable. Even for moderate sample sizes, CLT-based approximations are not accurate in the tails and better approximations, such as those presented in this paper, are needed.

We now consider a HTS experiment which was the motivation for the present paper. Left panel of Figure 1 shows measured values of the *Logarithmic Strain Coefficient* (LSC) of the wildtype cells in a Bioscreen array experiment in yeast genome screening studies, see (Warringer and Blomberg, 2003) and (Warringer et al., 2003).

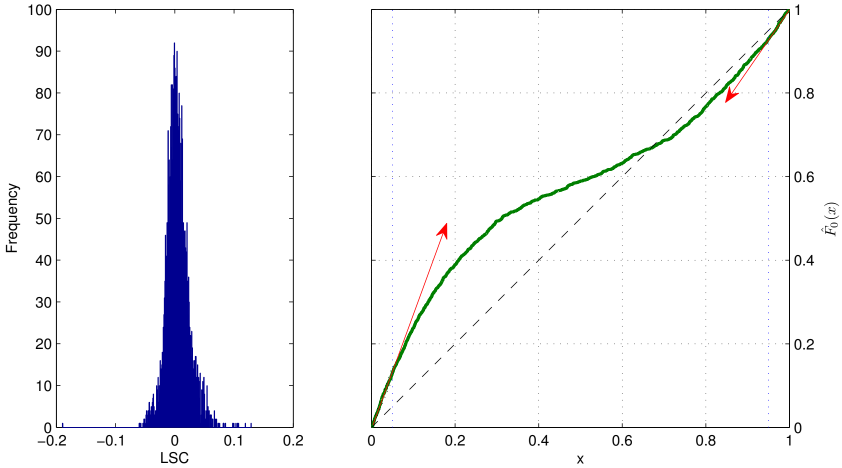

Figure 1. The Wild Type Data Set. *Left:* Histogram of 3456 LSC values from the wildtype dataset. *Right:* Empirical CDF of 1728 p-values obtained from one-sample $t-$test for pairs of LSC values.

The null hypothesis was that LSC of a wildtype yeast cell had normal distribution with mean zero and unknown variance. The experiment was made for quality control purposes, hence no treatment has been applied and the null hypothesis of mean zero was known to be true.

The histogram of the *LSC* values was skewed, see Figure 1, and we therefore plotted the empirical cumulative distribution function (CDF) of the 1728 $p-$values obtained from the LSC values. As expected, see the right panel in Figure 1, the distribution of the $p-$values was different from the theoretical uniform distribution.

Note, however, that both lower and upper tails of the plot approach straight lines, as indicated by the two arrows. This was in fact the starting point of the present article, and it later followed that such tail behavior is justified by Theorem 1.1, see Supplementary Materials.

In practical applications one needs to be able to compute or estimate the constant K_g. This can be done in a variety of ways.

Exact algebraic expression. For the case when components of $\mathbf{X}$ are i.i.d. random variables, constant K_g can be obtained directly from (4), (10) and (18) for the three choices of the test statistic T accordingly. We give numerous examples through Sections 2-5, and Supplementary Materials provides Wolfram Mathematica code to compute K_g for even more complicated cases, like e.g. Multivariate Normal case with $g \sim MVN(\mathbf{0}, \mathbf{\Sigma})$.

Numerical integration using quadratures. For an arbitrary multivariate density $g(\mathbf{x})$ and Student one- and two-sample $t-$ statistics, or $F-$ statistics with low degrees of freedom, K_g can be computed from (4), (10) or (18) using adaptive Simpson or Lobatto quadratures. We provide the corresponding MATLAB scripts in Supplementary Materials.

Numerical integration using Monte Carlo methods. For an $F-$statistic with the denominator that has more than two degrees of freedom, K_g can be computed using Monte Carlo integration, see Supplementary Materials. Monte Carlo methods are applicable to the case described above as well.

Simulations. The distribution tail of T can be estimated using simulations, see e.g. Section 7. In the current paper we used "brute-force" approach, but importance sampling techniques can be applied quite generally as well.

Estimation. If $g(\mathbf{x})$ is unknown but one instead has a sample from g, then K_g can be estimated as a slope of the graph of the CDF of the corresponding p-values in the origin of zero. In the yeast genome screening experiment, for example, K_g approximately equals the slope of the red arrow - theoretical justification of this fact is given in Supplementary Materials, and the estimation technique is similar to the Peak-Over-Threshold (POT) method in Extreme Value Theory, see e.g. the SmartTail software at www.smarttail.se and further examples in (Rootzén and Zholud, 2014).

Finally, the existence of K_g and its importance for questioning the logic behind some multiple testing procedures is discussed in (Zholud, 2011), I.3.

1.3 Literature review

There is enormous amount of literature on the behavior of the Student one- and two-sample $t-$ and $F-$ statistics under deviations from the standard assumptions. The overwhelming part of this literature is focused on normal approximations, that is, when $n \to \infty$. These are large sample approximations though, and are irrelevant to the topic of the present article.

For small and moderate sample sizes one would typically use Edgeworth expansion, see e.g. (Field and Ronchetti, 1990), (Hall, 1987) and (Gaen, 1949, 1950), or saddlepoint approximations, see e.g. (Zhou and Jing, 2006), (Jing et al., 2004) and (Daniels and Young, 1991). Edgeworth expansion improves the normal approximation but is still inaccurate in the tails. Saddlepoint approximations, on the other hand, can be very accurate in the tails, see e.g. (Jing et al., 2004), but the latter statement is based on purely empirical evidence and the asymptotic behavior of these approximations as $u \to \infty$ is not well studied. Furthermore, in practice one would require exact parametric form of the population density, and the use of saddle point approximations in statistical inference is questionable.

As for the approximations considered in this article, that is, when n is small and $u \to \infty$, the existing literature is very limited. This presumably can be explained by the fact that situations where one would need to test at significance levels of 10^{-3} and lower never arose, until present times. We focus on the most relevant works by (Bradley, 1952a,b) and (Hotelling, 1961).

Bradley covers the Student one-sample $t-$ statistic for i.i.d. non-normal observations, and also makes a somewhat less complete study of the corresponding cases for the Student two- sample $t-$test and the $F-$test of equality of variances. (Bradley, 1952b) derives the constant K_g from geometrical considerations, but does not state any assumptions on the underlying population density which ensure that the approximations hold. (Bradley, 1952a), on the other hand, gives assumptions on the population density, but these assumptions are insufficient, see Section 2 of Appendix A.

(Hotelling, 1961) studies the Student one-sample $t-$test for an "arbitrary" joint density of $\mathbf{X}$. Hotelling derives the constant K_g assuming that the limit in the left-hand side of (1) exists and that the function

$$D_n(\xi) = \int_0^\infty r^{n-1} g(r\xi_1, \cdots, r\xi_n) dr$$

is continuous for both densities g_0 and g_1. When it comes to the examples, however, the existence of the limit in (1) is taken for granted and the assumption of continuity of $D_n(\xi)$ is never verified.

Finally, a more detailed literature review that covers other approaches and meritable scientific works is given in Supplementary Materials.

The structure of this paper is as follows: Sections 2 - 5 contain main theorems and examples; Section 6 addresses the convergence speed and higher order expansions; Section 7 presents a simulation study. Appendix A includes the key lemma used in the proofs, in Section 1, and a discussion on the regularity conditions, in Section 2; Appendix B contains figures from the simulation study; and, finally, follows a brief summary of the Supplementary Materials that are available online.

2 One-sample $t-$statistic

Let $\mathbf{X} = (X_1, X_2, .., X_n),\ n \geq 2$, be a random vector that has a joint density g and define

$$T = \sqrt{n}(\overline{\mathbf{X}}/S),$$

where $\overline{\mathbf{X}}$ and S^2 are the sample mean and the sample variance of the vector $\mathbf{X}$. Introduce the unit vector $\mathbf{I} = (1/\sqrt{n}, 1/\sqrt{n}, .., 1/\sqrt{n})$, and assume that

$$g(x\mathbf{I}) > 0 \quad \text{for some} \quad x \geq 0 \tag{2}$$

and that

$$\int_0^\infty r^{n-1} \sup_{\substack{\|\boldsymbol{\xi}\|<\varepsilon,\\ \boldsymbol{\xi}\in L^\perp}} g\left(r\left(\mathbf{I}+\boldsymbol{\xi}\right)\right) dr < \infty \tag{3}$$

for some $\varepsilon > 0$, where L is the linear subspace of $\mathbb{R}^n$ spanned by the vector $\mathbf{I}$ and $L^\perp$ is its orthogonal complement. Finally, introduce the constant

$$K_g = 2\frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2})}\int_0^\infty r^{n-1} g\left(r\mathbf{I}\right) dr. \tag{4}$$

Theorem 2.1. *If g is continuous and satisfies (2) and (3), then*

$$\frac{\boldsymbol{P}\left(T>u\right)}{t_{n-1}(u)} = K_g + o(1) \quad as \quad u \to \infty, \tag{5}$$

where $t_{n-1}(u)$ is the tail of the $t-$distribution with $n-1$ degrees of freedom and $0 < K_g = K(g) < \infty$.

Proof. We use several variable changes to transform the right-hand side of

$$\mathbf{P}\left(T>u\right) = \int_{D_1} g(\mathbf{x})d\mathbf{x},$$

where $D_1 = \{\mathbf{x} : T > u\}$ and $d\mathbf{x}$ is the notation for $dx_1dx_2..dx_n$, to the form treated in Corollary 8.1.1 in Section 1 of Appendix A. Let $\mathbf{e_1}, \mathbf{e_2}, .., \mathbf{e_n}$ be the standard basis in $\mathbb{R}^n$ and A be an orthogonal linear operator which satisfies

$$A\mathbf{e_n} = \mathbf{I}. \tag{6}$$

Setting $\mathbf{x} = A\mathbf{y}$ we have that $\overline{\mathbf{X}} = y_n/\sqrt{n}$ and $S^2 = \sum_{i=1}^{n-1} y_i^2/(n-1)$, and hence

$$\mathbf{P}\left(T>u\right) = \int_{D_2} g(A\mathbf{y})d\mathbf{y},$$

where $D_2 = \left\{\mathbf{y} : \dfrac{y_n}{\sqrt{\frac{1}{n-1}\sum\limits_{i=1}^{n-1} y_i^2}} > u\right\}$.

Next, introducing new variables $y_i = (n-1)^{1/2} r t_i$ for $i \leq n-1$ and $y_n = r$, $r > 0$, applying Fubini's theorem, and recalling (6) we get

$$\mathbf{P}\left(T > u\right) = \int \cdots \int_{\sum t_i^2 < u^{-2}} G(\mathbf{t}) d\mathbf{t}, \tag{7}$$

where

$$G(\mathbf{t}) = (n-1)^{\frac{n-1}{2}} \int_0^\infty r^{n-1} g\left(r\left(\mathbf{I} + A\mathbf{v}(\mathbf{t})\right)\right) dr,$$

and

$$\mathbf{v}(\mathbf{t}) = (n-1)^{1/2}\left(t_1, t_2, .., t_{n-1}, 0\right).$$

Continuity of g and (3) ensure that G is continuous at zero, by the dominated convergence theorem, and Corollary 8.1.1 in Section 1 of Appendix A completes the proof. □

Assumption (2) ensures that $K_g > 0$ and the condition (3) holds if, for example, $K_g < \infty$ and g is continuous and has the asymptotic monotonicity property, see Lemma 9.2, Section 2, Appendix A.

Now consider the case when one of the assumptions (3) or (2) is violated. If (3) holds and (2) is violated, then (5) holds with $K_g = 0$, that is, the right tail of the distribution of T is "strictly lighter" than $t_{n-1}(u)$, the tail of the $t-$distribution with $n-1$ degrees of freedom. If, instead, (2) holds and (3) is violated, then, Theorem 9.1 in Section 2 of Appendix A shows that the right tail of the distribution of T is "at least as heavy" as $t_{n-1}(u)$, provided $K_g < \infty$, and "strictly heavier" than $t_{n-1}(u)$ if $K_g = \infty$.

We next consider two important corollaries - one concerning dependent Gaussian vectors, and another one that addresses the non-normal i.i.d. case.

Corollary 2.1.1 (Gaussian zero-mean case). *If* $\mathbf{X} \sim MVN(\mathbf{0}, \mathbf{\Sigma})$, *where* $\mathbf{\Sigma}$ *is a strictly positive-definite covariance matrix, then (5) holds with*

$$K_g = \frac{\left(\mathbf{I}\mathbf{\Sigma}\mathbf{I}^T\right)^{n/2}}{|\mathbf{\Sigma}|^{1/2}}.$$

Proof. Deriving the expression for K_g in (4) is straightforward. Note that $K_g < \infty$ since $\mathbf{\Sigma}$ is non-degenerate and $MVN(\mathbf{0}, \mathbf{\Sigma})$ has the asymptotic monotonicity property, see Section 2 of Appendix A. It then follows from Lemma 9.2 that the regularity constraint (3) holds, and so does (5). □

One possible application of Corollary 2.1.1 is to correct for the effect of dependency when using test statistic T. This is done by dividing the corresponding $p-$value by K_g.

Now consider the effect of non-normality. Assume that the elements X_i of the vector $\mathbf{X}$ are independent and identically distributed and let $h(x)$ be their common marginal density, so that $g(\mathbf{x}) = h(x_1)h(x_2)\cdots h(x_n)$.

Corollary 2.1.2 (i.i.d. case). *If* $h(x)$ *is continuous, and monotone on* $[L, \infty)$ *for some finite constant* L, *then (5) holds with*

$$K_g = 2\frac{(\pi n)^{\frac{n}{2}}}{\Gamma(\frac{n}{2})} \int_0^\infty r^{n-1} h\left(r\right)^n dr < \infty.$$

Proof. The monotonicity of $h(x)$ on $[L, \infty)$ implies that $g(\mathbf{x})$ has the asymptotic monotonicity property, see Section 2 of Appendix A, and the regularity assumption (3) hence follows from finiteness of K_g and Lemma 9.2. The finiteness of K_g, in turn, follows if we show that $rh(r) \to 0$ as $r \to \infty$.

Indeed, assume to the contrary that $\limsup rh(r) > 0$. Then there exists $\delta > 0$ and a sequence $\{r_k\}_{k=0}^{\infty}$ with $r_0 = L+1$ and such that $r_{k+1} > 2r_k$ and $r_k h(r_k) > \delta$ for any $k > 0$, and we get a contradiction since the monotonicity of $h(x)$ on $[L, \infty)$ then implies that $h(x)$ can not be a density,

$$\int_{L+1}^{\infty} h(r)dr \geq \sum_{k=1}^{\infty} (r_k - r_{k-1})\, h(r_k) > \delta \sum_{k=1}^{\infty} \frac{r_k - r_{k-1}}{r_k} = \infty. \qquad \square$$

Table 1. The constants K_g for the i.i.d case, Student one-sample $t-$test; $\Gamma(x)$, $B(x)$ and $M(a,b,x)$ are the Gamma, Beta and Kummer confluent hypergeometric function, see e.g. (Hayek, 2001).

Normal with mean $\mu \neq 0$ and standard deviation $\sigma > 0$
$M\left(\frac{1-n}{2}, \frac{1}{2}, -\frac{n\mu^2}{2\sigma^2}\right) + \frac{\mu}{\sigma}\frac{\sqrt{2n}\,\Gamma\left(\frac{1+n}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} M\left(1-\frac{n}{2}, \frac{3}{2}, -\frac{n\mu^2}{2\sigma^2}\right)$
Half-normal, and log-normal derived from a $N(\mu, \sigma^2)$
2^n and $\frac{n^{\frac{n-1}{2}}\sqrt{\pi}}{2^{\frac{n-3}{2}}\sigma^{n-1}\Gamma\left(\frac{n}{2}\right)}$
χ with $\nu > 0$, and χ^2 (and its inverse) with $\nu \geq 2$ d.f.
$\frac{2^n \pi^{n/2}\Gamma\left(\frac{n\nu}{2}\right)}{n^{\frac{n}{2}(\nu-1)}\Gamma\left(\frac{\nu}{2}\right)^n \Gamma\left(\frac{n}{2}\right)}$ and $\frac{2\pi^{n/2}\Gamma\left(\frac{n\nu}{2}\right)}{n^{\frac{n}{2}(\nu-1)}\Gamma\left(\frac{\nu}{2}\right)^n \Gamma\left(\frac{n}{2}\right)}$
F with $\mu > 0$ and $\nu > 0$ degrees of freedom
$\frac{2(\pi n)^{n/2}\Gamma\left(\frac{\mu n}{2}\right)\Gamma\left(\frac{\nu n}{2}\right)\Gamma\left(\frac{\mu+\nu}{2}\right)^n}{\Gamma\left(\frac{n}{2}\right)\left[\Gamma\left(\frac{\mu}{2}\right)\Gamma\left(\frac{\nu}{2}\right)\right]^n \Gamma\left(\frac{\mu+\nu}{2}n\right)}$
T with $\nu > 0$ d.f. and Cauchy
$\frac{n^{n/2}\Gamma\left(\frac{\nu n}{2}\right)}{\Gamma\left(\frac{(\nu+1)n}{2}\right)}\left(\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\right)^n$ and $\frac{n^{n/2}}{2^{n-1}\pi^{\frac{n-1}{2}}\Gamma\left(\frac{n+1}{2}\right)}$
Beta with shape parameters $\alpha > 1$ and $\beta > 1$
$\frac{2(\pi n)^{n/2}\Gamma(\alpha n)\Gamma\left(1+(\beta-1)n\right)}{B(\alpha,\beta)^n \Gamma\left(\frac{n}{2}\right)\Gamma\left(1+(\alpha+\beta-1)n\right)}$
Gamma (and its inverse) with shape $\alpha > 1$
$\frac{2n^{\frac{n}{2}(1-2\alpha)}\pi^{n/2}\Gamma(\alpha n)}{\Gamma(\alpha)^n \Gamma\left(\frac{n}{2}\right)}$
Uniform on interval $[a, b], b > 0$
$\frac{(\pi n)^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}+1\right)}\begin{cases}\left(\frac{b}{b-a}\right)^n & 0 \in [a,b] \\ \frac{b^n - a^n}{(b-a)^n} & [a,b] \subset [0,\infty)\end{cases}$
Centered exponential and exponential
$\frac{2\left(\frac{\pi}{n}\right)^{n/2}\Gamma(n)}{e^n \Gamma\left(\frac{n}{2}\right)}$ and $\frac{2\left(\frac{\pi}{n}\right)^{n/2}\Gamma(n)}{\Gamma\left(\frac{n}{2}\right)}$
Maxwell, and Pareto with $k > 0$ and scale $\alpha > 0$
$\frac{\left(\frac{4}{n}\right)^n \Gamma\left(\frac{3n}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}$ and $\frac{(\pi n)^{n/2}\alpha^{n-1}}{\Gamma\left(\frac{n}{2}+1\right)}$

3 Two-sample $t-$statistic

In this section we cover the Student two-sample $t-$statistic. However, we first consider a more general case. For $n_1 \geq 2,\ n_2 \geq 2$, set $n = n_1 + n_2$ and let $\mathbf{X} = (X_1, X_2, .., X_n)$ be a random vector that has a multivariate joint density g. Further, let S_1 and S_2 be the sample variances of the vectors $(X_1, X_2, .., X_{n_1})$ and $(X_{n_1+1}, X_{n_1+2}, .., X_n)$ and define

$$T = \frac{\frac{1}{n_1}\sum\limits_{i=1}^{n_1} X_i - \frac{1}{n_2}\sum\limits_{i=n_1+1}^{n} X_i}{\sqrt{\alpha S_1^2 + \beta S_2^2}},$$

where α and β are some positive constants (to be set later). Next, define the two unit vectors

$$\mathbf{I}_1 = (1/\sqrt{n_1}, 1/\sqrt{n_1}, .., 1/\sqrt{n_1}, 0, 0, .., 0) \quad \text{and} \quad \mathbf{I}_2 = (0, 0, .., 0, 1/\sqrt{n_2}, 1/\sqrt{n_2}, .., 1/\sqrt{n_2}),$$

and let $\omega_0 = \arccos\left(\sqrt{n_2/n}\right)$. We assume that

$$g\left(r\left(\cos(\omega - \omega_0)\mathbf{I}_1 + \sin(\omega - \omega_0)\mathbf{I}_2\right)\right) > 0 \tag{8}$$

for some $r \geq 0$ and $\omega \in [-\pi/2,\ \pi/2]$, and that for some $\varepsilon > 0$

$$\int\limits_{-\pi/2}^{\pi/2} \cos(\omega)^{n-2} \int\limits_{0}^{\infty} r^{n-1} \sup_{\substack{\|\boldsymbol{\xi}\|<\varepsilon \\ \boldsymbol{\xi}\in L^{\perp}}} g\left(r\left(\cos(\omega - \omega_0)\mathbf{I}_1 + \sin(\omega - \omega_0)\mathbf{I}_2 + \boldsymbol{\xi}\right)\right) dr d\omega < \infty, \tag{9}$$

where L is a linear subspace of $\mathbb{R}^n$ spanned by the vectors $\mathbf{I}_1$ and $\mathbf{I}_2$, and $L^{\perp}$ is its orthogonal complement. Next, define the constant

$$K_g = C(n_1, n_2, \alpha, \beta) \int\limits_{-\pi/2}^{\pi/2} \cos(\omega)^{n-2} \int\limits_{0}^{\infty} r^{n-1} g\left(r\left(\cos(\omega - \omega_0)\mathbf{I}_1 + \sin(\omega - \omega_0)\mathbf{I}_2\right)\right) dr d\omega, \tag{10}$$

where the constant $C(n_1, n_2, \alpha, \beta)$ is given by

$$C(n_1, n_2, \alpha, \beta) = \frac{2\pi^{\frac{n-1}{2}} \left(\frac{n_1-1}{\alpha}\right)^{\frac{n_1-1}{2}} \left(\frac{n_2-1}{\beta}\right)^{\frac{n_2-1}{2}} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{\frac{n-2}{2}}}{\Gamma\left(\frac{n-1}{2}\right)(n-2)^{\frac{n-2}{2}}}.$$

Theorem 3.1. *If g is continuous and satisfies (8) and (9), then*

$$\frac{\boldsymbol{P}(T > u)}{t_{n-2}(u)} = K_g + o(1) \quad \textit{as} \quad u \to \infty, \tag{11}$$

where $t_{n-2}(u)$ is the tail of the $t-$distribution with $n-2$ degrees of freedom and $0 < K_g = K(g) < \infty$.

Proof. The proof is similar to the proof of Theorem 2.1. Let A be an orthogonal linear operator such that

$$A\mathbf{e}_{\mathbf{n_1}} = \mathbf{I}_1 \ \text{ and } \ A\mathbf{e}_{\mathbf{n}} = \mathbf{I}_2. \tag{12}$$

Changing coordinate system $\mathbf{x} = A\mathbf{y}$ gives

$$\frac{1}{n_1}\sum\nolimits_{i=1}^{n_1} X_i = y_{n_1}/\sqrt{n_1}, \quad \frac{1}{n_2}\sum\nolimits_{i=n_1+1}^{n} X_i = y_n/\sqrt{n_2},$$

$$S_1^2 = \sum\nolimits_{i=1}^{n_1-1} y_i^2/(n_1-1) \quad \text{and} \quad S_2^2 = \sum\nolimits_{i=n_1+1}^{n-1} y_i^2/(n_2-1)$$

and therefore

$$\mathbf{P}\left(T > u\right) = \int\limits_{\{\mathbf{x}: T>u\}} g(\mathbf{x})d\mathbf{x} = \int\limits_{D} g(A\mathbf{y})d\mathbf{y},$$

where $D = \left\{ \mathbf{y} : \left(\frac{1}{\sqrt{n_1}} y_{n_1} - \frac{1}{\sqrt{n_2}} y_n \right) \Big/ \left(\frac{\alpha}{n_1-1} \sum\limits_{i=1}^{n_1-1} y_i^2 + \frac{\beta}{n_2-1} \sum\limits_{i=n_1+1}^{n-1} y_i^2 \right)^{1/2} > u \right\}$.

Next, define $c_1(\omega)$ and $c_2(\omega)$ by

$$\frac{c_1(\omega)}{\sqrt{1/n_1+1/n_2}} = \sqrt{\frac{n_1-1}{\alpha}}\cos(\omega) \quad \text{and} \quad \frac{c_2(\omega)}{\sqrt{1/n_1+1/n_2}} = \sqrt{\frac{n_2-1}{\beta}}\cos(\omega),$$

and introduce new variables $t_1, t_2, .., t_{n-2}, r, \omega$ such that

$$\begin{aligned} y_i &= rc_1(\omega)t_i \quad \text{for} \quad i = 1, 2, .., n_1 - 1, \\ y_i &= rc_2(\omega)t_{i-1} \quad \text{for} \quad i = n_1+1, n_1+2, .., n-1, \\ y_{n_1} &= r\cos(\omega-\omega_0) \quad \text{and} \quad y_n = r\sin(\omega-\omega_0), \quad r > 0. \end{aligned}$$

The identity $\cos(\omega-\omega_0)/\sqrt{n_1} - \sin(\omega-\omega_0)/\sqrt{n_2} = \sqrt{1/n_1+1/n_2}\cos(\omega)$, Fubini's theorem, and (12) give

$$\mathbf{P}\left(T > u\right) = \int\limits_{\sum\limits_{i=1}^{n-2} t_i^2 < u^{-2}} \cdots \int G(\mathbf{t})d\mathbf{t}, \tag{13}$$

where

$$G(\mathbf{t}) = M \int\limits_{-\pi/2}^{\pi/2} \cos(\omega)^{n-2} \int\limits_{0}^{\infty} r^{n-1} g\left(r\left(\cos(\omega-\omega_0)\mathbf{I}_1 + \sin(\omega-\omega_0)\mathbf{I}_2 + A\mathbf{v}(\mathbf{t}, \omega-\omega_0) \right) \right) dr d\omega$$

with

$$\mathbf{v}(\mathbf{t}, \omega) = \left(c_1(\omega)t_1, .., c_1(\omega)t_{n_1-1}, 0, c_2(\omega)t_{n_1}, .., c_2(\omega)t_{n-2}, 0 \right)$$

and

$$M = \left(\frac{n_1-1}{\alpha}\right)^{\frac{n_1-1}{2}} \left(\frac{n_2-1}{\beta}\right)^{\frac{n_2-1}{2}} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{\frac{n-2}{2}}.$$

The finiteness of the integral in (9) and continuity of g imply the continuity of G at zero by the dominated convergence theorem, and Corollary 8.1.1 in Section 1 of Appendix A gives the asymptotic expression (11) with the constant K_g defined in (10). □

The assumption (8) ensures that $K_g > 0$, and the regularity constraint (9) can be verified directly, or using criteria in Section 2 of Appendix A.

Corollary 3.1.1 (Gaussian zero-mean case)**.** *If $X \sim MVN(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is a strictly positive-definite covariance matrix, then (11) holds with*

$$K_g = C(n_1, n_2, \alpha, \beta) \frac{\Gamma\left(\frac{n}{2}\right)}{2\pi^{\frac{n}{2}}|\mathbf{\Sigma}|^{1/2}} \int\limits_{-\pi/2}^{\pi/2} \frac{\cos(\omega)^{n-2}}{\left(\mathbf{v}(\omega)\mathbf{\Sigma}^{-1}\mathbf{v}(\omega)^T\right)^{n/2}} d\omega, \tag{14}$$

where $\mathbf{v}(\omega) = \cos(\omega-\omega_0)\mathbf{I}_1 + \sin(\omega-\omega_0)\mathbf{I}_2$.

Proof. Let λ be the smallest eigenvalue of $\mathbf{\Sigma}^{-1}$. Note that $\lambda > 0$, which implies that

$$g(\mathbf{x}) \leq \frac{1}{(2\pi)^{n/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{\lambda}{2}\|\mathbf{x}\|^2} < \frac{1}{\|\mathbf{x}\|^{n+1}}$$

for $\|\mathbf{x}\|$ large enough. Now, condition (9) holds according to Lemma 9.1 in Section 2 of Appendix A, and deriving K_g is a calculus exercise. □

The asymptotic expression for the distribution tail of the Student two-sample $t-$statistic is obtained by setting

$$\alpha = \frac{n_1 - 1}{n-2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \quad \text{and} \quad \beta = \frac{n_2 - 1}{n-2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right).$$

For the Gaussian zero-mean case the expression (14) then reduces to

$$\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})\sqrt{\pi}|\mathbf{\Sigma}|^{1/2}} \int_{-\pi/2}^{\pi/2} \frac{\cos(\omega)^{n-2}}{\left(\mathbf{v}(\omega)\mathbf{\Sigma}^{-1}\mathbf{v}(\omega)^T\right)^{n/2}} d\omega. \tag{15}$$

As expected, if $\mathbf{\Sigma} = \sigma^2 \mathbf{1}_n$ (recall, $\mathbf{1}_n$ is the identity matrix) and $\sigma^2 > 0$, then direct calculation shows that $K_g = 1$. A less trivial case is when the population variances are unequal. Substituting the diagonal matrix

$$\mathbf{\Sigma} = diag\{\underbrace{\sigma_1^2, .., \sigma_1^2}_{n_1}, \underbrace{\sigma_2^2, .., \sigma_2^2}_{n_2}\}$$

into (15), the latter, after some lengthy algebraic manipulations, takes form

$$\frac{\Gamma(\frac{n}{2}) n_1^{\frac{n}{2}-1} k^{n_2}}{n^{\frac{n-1}{2}} \Gamma(\frac{n-1}{2})\sqrt{\pi}} \left[\int_{-\infty}^{1} \frac{(1-x)^{n-2}}{(1+ck^2x^2)^{n/2}} dx + \int_{1}^{\infty} \frac{(x-1)^{n-2}}{(1+ck^2x^2)^{n/2}} dx \right],$$

where $k = \sigma_1/\sigma_2$ and $c = n_2/n_1$. The integrals can be computed by resolving the corresponding rational functions into partial fractions (n is even) or by expanding brackets in the numerator and integrating by parts (n is odd). We have computed K_g for sample sizes up to 6, see Table 2 below.

Table 2. Constants K_g for the Student two-sample $t-$test, variances unequal.

$n_2 \backslash n_1$	$n_1 = 2$	$n_1 = 3$	$n_1 = 4$	$n_1 = 5$	$n_1 = 6$
$n_2 = 2$	$\frac{k^2+1}{2k}$	$\frac{(2k^2+3)^{3/2}}{5\sqrt{5}k^2}$	$\frac{(k^2+2)^2}{9k^3}$	$\frac{(2k^2+5)^{5/2}}{49\sqrt{7}k^4}$	$\frac{(k^2+3)^3}{64k^5}$
$n_2 = 3$	$\frac{(3k^2+2)^{3/2}}{5\sqrt{5}k}$	$\frac{(k^2+1)^2}{4k^2}$	$\frac{(3k^2+4)^{5/2}}{49\sqrt{7}k^3}$	$\frac{(3k^2+5)^3}{512k^4}$	$\frac{(k^2+2)^{7/2}}{27\sqrt{3}k^5}$
$n_2 = 4$	$\frac{(2k^2+1)^2}{9k}$	$\frac{(4k^2+3)^{5/2}}{49\sqrt{7}k^2}$	$\frac{(k^2+1)^3}{8k^3}$	$\frac{(4k^2+5)^{7/2}}{2187k^4}$	$\frac{(2k^2+3)^4}{625k^5}$
$n_2 = 5$	$\frac{(5k^2+2)^{5/2}}{49\sqrt{7}k}$	$\frac{(5k^2+3)^3}{512k^2}$	$\frac{(5k^2+4)^{7/2}}{2187k^3}$	$\frac{(k^2+1)^4}{16k^4}$	$\frac{(5k^2+6)^{9/2}}{14641\sqrt{11}k^5}$
$n_2 = 6$	$\frac{(3k^2+1)^3}{64k}$	$\frac{(2k^2+1)^{7/2}}{27\sqrt{3}k^2}$	$\frac{(3k^2+2)^4}{625k^3}$	$\frac{(6k^2+5)^{9/2}}{14641\sqrt{11}k^4}$	$\frac{(k^2+1)^5}{32k^5}$

Note also that for odd sample sizes the exact distribution of the Student two-sample $t-$statistic is known, see (Ray and Pitman, 1961). The closed form expressions for (14) or (15) for an arbitrary covariance matrix $\mathbf{\Sigma}$ is unknown, but for fixed n one can compute K_g numerically. In most cases it is also possible to obtain the exact expression for K_g using Mathematica software. Examples are given in Supplementary Materials.

4 Welch statistic

The Welch statistic differs from the Student two-sample $t-$statistic in that it has $\alpha = 1/n_1$ and $\beta = 1/n_2$, see the definition of T in the previous section. Welch statistic relaxes the assumption of equal variances and its distribution under the null hypothesis of equal means is instead approximated by the Student $t-$distribution with ν degrees of freedom, where

$$\nu = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{S_1^4/(n_1^2(n_1-1)) + S_2^4/(n_2^2(n_2-1))}$$

is estimated from the data. Welch approximation performs poorly in the tail area because it has wrong asymptotic behavior, cf. Corollary 3.1.1. The accuracy of our asymptotic approximation and its relation to the exact distribution of the Welch statistic for odd sample sizes, see (Ray and Pitman, 1961), is discussed in Supplementary Materials. We also study the accuracy of our approximations using simulations, see Section 7.

Finally, Table 3 presents constants K_g for the Welch statistic under standard assumptions. Here constant k stands for the ratio σ_1/σ_2.

Table 3. Constants K_g for the Welch $t-$test, variances unequal.

$n_2 \backslash n_1$	$n_1 = 2$	$n_1 = 3$	$n_1 = 4$	$n_1 = 5$
$n_2 = 2$	$\frac{k^2+1}{2k}$	$\frac{(2k^2+3)^{3/2}}{9k^2}$	$\frac{3\sqrt{\frac{3}{2}}(k^2+2)^2}{16k^3}$	$\frac{4(2k^2+5)^{5/2}}{125k^4}$
$n_2 = 3$	$\frac{(3k^2+2)^{3/2}}{9k}$	$\frac{(k^2+1)^2}{4k^2}$	$\frac{(3k^2+4)^{5/2}}{50\sqrt{5}k^3}$	$\frac{4(3k^2+5)^3}{1215k^4}$
$n_2 = 4$	$\frac{3\sqrt{\frac{3}{2}}(2k^2+1)^2}{16k}$	$\frac{(4k^2+3)^{5/2}}{50\sqrt{5}k^2}$	$\frac{(k^2+1)^3}{8k^3}$	$\frac{3\sqrt{\frac{3}{35}}(4k^2+5)^{7/2}}{1715k^4}$
$n_2 = 5$	$\frac{4(5k^2+2)^{5/2}}{125k}$	$\frac{4(5k^2+3)^3}{1215k^2}$	$\frac{3\sqrt{\frac{3}{35}}(5k^2+4)^{7/2}}{1715k^3}$	$\frac{(k^2+1)^4}{16k^4}$
$n_2 = 6$	$\frac{25\sqrt{\frac{5}{3}}(3k^2+1)^3}{216k}$	$\frac{25\sqrt{\frac{5}{7}}(2k^2+1)^{7/2}}{343k^2}$	$\frac{25\sqrt{\frac{5}{2}}(3k^2+2)^4}{16384k^3}$	$\frac{4(6k^2+5)^{9/2}}{177147k^4}$

5 $F-$statistic

In this section we study the tails of the distribution of an F-statistic for testing the equality of variances. Similar results can also be obtained for an F-test used in one-way ANOVA, lack-of-fit sum of squares, and when comparing two nested linear models in regression analysis. Define random vectors $\mathbf{X} = (X_1, X_2, .., X_{n_1})$ and $\mathbf{Y} = (Y_1, Y_2, .., Y_{n_2})$, $n_1 \geq 2$ and $n_2 \geq 2$, and let $g(\mathbf{x}, \mathbf{y})$ be the joint density of the vector $(\mathbf{X}, \mathbf{Y})$. Now set $n = n_1 + n_2$ and define

$$T = S_1^2/S_2^2,$$

where S_1 and S_2 are the sample variances of $\mathbf{X}$ and $\mathbf{Y}$ respectively. Let $s_1(\mathbf{x})$ denote the sample standard deviation of the vector $\mathbf{x} \in \mathbb{R}^{n_1}$ and define the unit vector $\mathbf{I} = \left(1/\sqrt{n_2}, 1/\sqrt{n_2}, .., 1/\sqrt{n_2}\right)$. We assume that

$$s_1(\mathbf{x})\, g\left(\mathbf{x}, r\mathbf{I}\right) > 0 \tag{16}$$

for some $\mathbf{x}$ and r, and that the integral

$$\int_{\mathbb{R}^{n_1}} \cdots \int s_1(\mathbf{x})^{n_2-1} \int_{-\infty}^{\infty} \max_{\substack{\|\boldsymbol{\xi}\|<\varepsilon,\\ \boldsymbol{\xi}\in L^{\perp}}} g\left(\mathbf{x}, r\mathbf{I} + s_1(\mathbf{x})\boldsymbol{\xi}\right) dr d\mathbf{x} \tag{17}$$

is finite for some $\varepsilon > 0$, where L is a linear subspace spanned by vector $\mathbf{I}$ and $L^{\perp}$ is its orthogonal complement. Finally, define the constant

$$K_g = \frac{\Gamma\left(\frac{n_1-1}{2}\right)\left(\pi(n_1-1)\right)^{\frac{n_2-1}{2}}}{\Gamma\left(\frac{n-2}{2}\right)} \underset{\mathbb{R}^{n_1}}{\int \cdots \int} s_1(\mathbf{x})^{n_2-1} \int_{-\infty}^{\infty} g\left(\mathbf{x}, r\mathbf{I}\right) dr d\mathbf{x}. \tag{18}$$

Theorem 5.1. *If g is continuous and satisfies (16) and (17), then*

$$\frac{\boldsymbol{P}(T > u)}{F_{n_1-1,n_2-1}(u)} = K_g + o(1) \quad as \quad u \to \infty, \tag{19}$$

where $F_{n_1-1,n_2-1}(u)$ is the tail of the F-distribution with parameters n_1-1 and n_2-1 and $0 < K_g = K(g) < \infty$.

Corollary 5.1.1 (Gaussian zero-mean case, independent samples)**.** *If X and Y are independent zero-mean Gaussian random vectors with strictly non-degenerate covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, then (19) holds with*

$$K_g = C \underset{\mathbb{R}^{n_1}}{\int \cdots \int} \frac{s_1(\mathbf{x})^{n_2-1}}{\left(1 + \mathbf{x}\mathbf{\Sigma}_1^{-1}\mathbf{x}^T\right)^{n/2}} d\mathbf{x}, \tag{20}$$

where the constant C is given by

$$C = \frac{(n-2)(n_1-1)^{\frac{n_2-1}{2}}\Gamma\left(\frac{n_1-1}{2}\right)|\mathbf{I}\mathbf{\Sigma}_2\mathbf{I}^T|^{1/2}}{2\pi^{\frac{n_1+1}{2}}|\mathbf{\Sigma}_1|^{1/2}|\mathbf{\Sigma}_2|^{1/2}}.$$

The proofs of Theorem 5.1 and Corollary 5.1.1 are given in Supplementary Materials. Now consider the asymptotic power of the $F-$statistic.

Corollary 5.1.2 (Asymptotic Power)**.** *If X and Y are independent zero-mean Gaussian random vectors with covariance matrices $\sigma_1^2\mathbf{1}_{n_1}$ and $\sigma_2^2\mathbf{1}_{n_2}$, $\sigma_1^2+\sigma_2^2 > 0$, then*

$$\lim_{u\to\infty} \frac{\boldsymbol{P}(T > u)}{F_{n_1-1,n_2-1}(u)} = \left(\frac{\sigma_1}{\sigma_2}\right)^{n_2-1}. \tag{21}$$

Proof. Changing variables $\mathbf{x} = \sigma_1 B\mathbf{y}$, where B is an orthogonal operator such that

$$B\mathbf{e_{n_1}} = \left(1/\sqrt{n_1}, 1/\sqrt{n_1}, .., 1/\sqrt{n_1}\right),$$

the integral on the right-hand side of (20) takes form

$$\sigma_1^{n-1}\left(\frac{1}{n_1-1}\right)^{\frac{n_2-1}{2}} \underset{\mathbb{R}^{n_1}}{\int \cdots \int} \frac{\left(\|\mathbf{y}\|^2 - y_{n_1}^2\right)^{\frac{n_2-1}{2}}}{(1+\|\mathbf{y}\|^2)^{n/2}} d\mathbf{y},$$

and is then evaluated by passing to spherical coordinates. □

A careful reader may note that (21) follows from the asymptotic expansion of

$$\mathbf{P}\left(T > u\right) = F_{n_1-1,n_2-1}\left((\sigma_2/\sigma_1)^2 u\right)$$

in terms of $F_{n_1-1,n_2-1}(u)$. Our aim was just to show that despite the seeming complexity of the expression (18), the constant K_g can be evaluated directly, at least for some standard densities. It is also possible to compute K_g numerically, see the (MATLAB, 2010) scripts in Supplementary Materials.

6 Second and higher order approximations

In this section we discuss the speed of convergence in Theorem 1.1. Let T be one of the test statistics defined in Sections 2, 3 and 5 and let $t_k(u)$ be the Student $t-$distribution tail with k degrees of freedom and $F_{m,k}(u)$ be the $F-$distribution tail with parameters m and k. For an arbitrary continuous multivariate density $g = g_1(\mathbf{x})$, assume that conditions (3), (9) and (17) hold, and define the constant K_g by (4), (10) and (18) for the three tests respectively. For the Student $t-$ statistic the function $G(\mathbf{t})$ is given by (7) and (13), and for the $F-$ statistic see the corresponding formula in the proof of Theorem 5.1 in Supplementary Materials. Finally, with the standard notation ∇f for the gradient of a scalar function f, and a parameter α which can take values 1 or 2, define

$$d_{\alpha,m,k}(u) = \frac{1}{u^{\frac{\alpha(k+1)}{2}}} \left[C_1 \sup_{\|\mathbf{x}\| \le u^{-\frac{\alpha}{2}}} \left\|\nabla G(\mathbf{x})\right\| + C_2 \frac{K_g}{\alpha} \frac{1}{u^{\frac{\alpha}{2}}} \right] \tag{22}$$

with the constants C_1, C_2 (which depend on m and k), see Lemma 8.1 (B), Section 1, Appendix A.

Lemma 6.1 (Absolute error bound)**.** *If $G(\mathbf{t})$ is differentiable in some neighborhood of zero, then for any $u > 0$ the following inequalities*

$$\begin{aligned} |\boldsymbol{P}(T > u) - K_g\, t_{n-1}(u)| &\le d_{2,1,n-1}(u), \\ |\boldsymbol{P}(T > u) - K_g\, t_{n-2}(u)| &\le d_{2,1,n-2}(u), \\ |\boldsymbol{P}(T > u) - K_g\, F_{n_1-1,n_2-1}(u)| &\le d_{1,n_1-1,n_2-1}(u), \end{aligned}$$

hold for the Student one- and two-sample $t-$ and $F-$ statistics accordingly.

Proof. The first two inequalities follow from (5), (11) and Corollary 8.1.1, Section 1, Appendix A, and for the $F-$statistic we use Lemma 8.1 (B) with $\alpha = 1$ and $\sqrt{u}$ instead of u. □

Below follows the asymptotic formula for the relative error. For convenience we denote the distribution tail of T under the null hypothesis $H_0:\ g_0 \sim MVN(\mathbf{0}, \sigma^2 \mathbf{1}_n)$ by $t(u)$.

Lemma 6.2 (Relative error decrease rate)**.** *If $G(\mathbf{t})$ is twice differentiable in some neighborhood of zero, then*

$$\frac{\boldsymbol{P}(T > u) - K_g t(u)}{\boldsymbol{P}(T > u)} = \frac{C_3}{u^\alpha}(1 + o(1)),$$

where

$$C_3 = \frac{\alpha\, k\, B\left(\frac{m}{2}, \frac{k}{2}\right)}{2\left(\frac{k}{m}\right)^{k/2}} \frac{L_{G,\alpha}}{K_g},$$

the triple (α, m, k) is set to $(2,1,n-1)$, $(2,1,n-2)$ and $(1, n_1, n_2)$ for the Student one- and two- sample $t-$ and $F-$ statistics respectively, and the constant $L_{G,\alpha}$ is defined in Lemma 8.1 (C) in Section 1 of Appendix A.

Proof. The result follows from formulas (5), (11) and (19) for $\mathbf{P}\,(T > u)$, Lemma 8.1 (C) in Section 1 of Appendix A, and formula (29). □

The bounds and asymptotic expressions for the case of an arbitrary null hypothesis H_0 are derived using basic calculus

$$\mathbf{P}\,(T > u|H_1) - (K_{g_1}/K_{g_0}) \times \mathbf{P}\,(T > u|H_0) = \Big(\mathbf{P}\,(T > u|H_1) - K_{g_1} t(u)\Big)$$

$$-(K_{g_1}/K_{g_0}) \times \Big(\mathbf{P}\,(T > u|H_0) - K_{g_0} t(u)\Big),$$

and the absolute error of the approximation (1) is thus bounded by the linear combination of the absolute errors considered in Lemma 6.1 above.

For the relative error we replace the two probabilities $\mathbf{P}(T > u|H_1)$ and $\mathbf{P}(T > u|H_0)$ by their second order expansions given by Lemma 8.1 (C) in Section 1 of Appendix A, and then use (29). Lemma 8.1 can also be generalized to obtain higher order series expansion for $\mathbf{P}(T > u)$ as $u \to \infty$.

7 Simulation study

Let T be one of the test statistics considered in the previous sections and $t(u)$ be the distribution tail of T under $H_0: \ g \sim MVN(\mathbf{0}, \mathbf{1}_n)$. Next, we choose the sample size, specify the density $g(\mathbf{x})$, and simulate N random vectors $\mathbf{X} \sim g$. For each vector $\mathbf{X}$ we compute $t^* = T(\mathbf{X})$, the value of the test statistic T, and two p-values $p^R = t(t^*)$ and $p^C = K_g\, t(t^*) = K_g p^R$. Finally, we plot the empirical CDF of p^R and p^C over the range $I(r) = [0, 1/r]$, where the *Zoom Factor* (Z.F.) parameter r determines the tail region of interest. Here $N = 10,000 \times r$ so that $I(r)$ contains approximately $10,000$ p-values (as if they were uniformly distributed) - this is to ensure that the tails of the distribution of the p-values p^R and p^C are equally well approximated by the corresponding CDFs in all the tail regions. The letters "R" and "C" in the notation for the p-values stand for "Raw", i.e. computed using $t(u)$, and "Corrected", i.e. computed using $K_g t(u)$.

For the i.i.d case, let $h(x)$, the marginal density of the vector $\mathbf{X}$, be either *Uniform*$(-1, 1)$, *Standard normal*, *Centered exponential*, *Cauchy*, or $t-$ density with 2 or 5 degrees of freedom. The constant K_g was either evaluated explicitly in Mathematica or computed numerically in MATLAB, see Supplementary Materials. Figures 2, 3 and 4 in Appendix B show empirical CDFs for different sample sizes and Zoom Factor r varying between 20 and $1,000,000$. One can see that our approximations are very accurate in the tail regions for all the three test statistics, all sample sizes, and densities $h(x)$ considered in the study. Note also that the convergence speed is better for smaller sample sizes - this is in agreement with the bounds for the absolute error in Lemma 6.1, see Section 6.

Next, we computed p-values for the Welch statistic and compared them with the p-values obtained using the expression (14) in Corollary 3.1.1. Here "Raw" p-values are obtained using the Welch approximation and the notation is p^W. According to the plots in the top row of Figure 5, it may seem that the p-values p^W are uniformly distributed. However, if one "zooms in" to the tail region, see the plots in the middle row of Figure 5, it is clear that the p-values obtained using Welch approximation deviate significantly from the theoretical uniform distribution, while the corrected p-values p^C follow the diagonal line precisely. The advantage of using our tail approximations is fully convincing at Zoom Factor $100,000$, see the bottom row of Figure 5.

Finally, we made similar plots for even more peculiar scenarios where the data was dependent and non-stationary, see e.g. Figure 6. Our approximations were very accurate in all considered cases.

Appendix A: Supplementary theorems and lemmas

This Appendix is split into two parts - the first one introduces the key lemma which is used in Sections 2, 3 and 5, and the second contains useful notes on the regularity constraints (replacing them by simpler criteria that can be used in practice) and shows how to weaken the assumption of continuity of the density $g(\mathbf{x})$.

1. Asymptotics of an integral of a continuous function over a shrinking ball

It was shown that the tails of the distribution of the Student one- and two-sample $t-$, Welch, and $F-$ statistics are determined by the asymptotic behavior of an integral of some function (different for each of the tests) over a shrinking ball.

Let $G(\mathbf{x}),\ \mathbf{x} \in \mathbb{R}^k$ be some real-valued function and consider the asymptotic behavior of

$$F(u) = \int \cdots \int_{\sum x_i^2 < u^{-2}} G(\mathbf{x})\, d\mathbf{x} \tag{23}$$

for fixed k and $u \to \infty$.

Lemma 8.1. *Set $f(u) = \alpha^{-1} F_{m,k}(u^2)$, where $F_{m,k}(\cdot)$ is the tail of the $F-$distribution with $m \geq 1$ and $k \geq 2$ degrees of freedom, and let $\mathrm{Vol}(B_k)$ be the volume of the unit k-ball B_k and $B(x,y)$ be the Beta function. The parameters α and m will be set later. With the above notation we have:*

(A) If G is continuous at zero, then

$$\frac{F(u)}{f(u)} = K_{G,\alpha} + o(1) \quad \text{as} \quad u \to \infty, \tag{24}$$

where

$$K_{G,\alpha} = \frac{\alpha\, k\, B\left(\frac{m}{2}, \frac{k}{2}\right)}{2\left(\frac{k}{m}\right)^{k/2}} \mathrm{Vol}(B_k) G(\mathbf{0}). \tag{25}$$

(B) If G is differentiable in some neighborhood of zero, then for any $u > 0$

$$|F(u) - K_{G,\alpha}\, f(u)| \leq \frac{C_1}{u^{k+1}} \sup_{\|\mathbf{x}\| \leq u^{-1}} \left\|\nabla G(\mathbf{x})\right\| + C_2 \frac{K_{G,\alpha}}{\alpha} \frac{1}{u^{k+2}}, \tag{26}$$

where

$$C_1 = \mathrm{Vol}(B_k) \quad \text{and} \quad C_2 = \frac{k(k+m)}{m(k+2)} \frac{\left(\frac{k}{m}\right)^{k/2}}{B\left(\frac{m}{2}, \frac{k}{2}\right)}, \tag{27}$$

and $\nabla G(\mathbf{x})$ is a gradient of G evaluated at point $\mathbf{x}$.

(C) If G is twice differentiable in some neighborhood of zero, then

$$u^{k+2} \left(F(u) - K_{G,\alpha}\, f(u)\right) = L_{G,\alpha} + o(1) \quad \text{as} \quad u \to \infty, \tag{28}$$

where

$$L_{G,\alpha} = C_1 \frac{tr\left(Hess\left(G(\mathbf{0})\right)\right)}{2(k+2)} - C_2 \frac{K_{G,\alpha}}{\alpha},$$

$tr(A)$ is the trace of a square matrix A, and $Hess\left(G(\mathbf{x})\right)$ is the Hessian matrix of G evaluated at point $\mathbf{x}$. Constants C_1 and C_2 are given by (27).

Proof. The first statement follows from the asymptotic expansion for the $F-$distribution tail

$$f(u) = \frac{2\left(\frac{k}{m}\right)^{k/2}}{\alpha\, k\, B\left(\frac{m}{2}, \frac{k}{2}\right)} \left[\frac{1}{u^k} - \frac{k^2(k+m)}{2m(k+2)} \frac{1}{u^{k+2}}\right] + o\left(\frac{1}{u^{k+2}}\right). \tag{29}$$

Indeed, changing variables $\mathbf{x} = \mathbf{y}/u$ we write

$$F(u) = \int \cdots \int_{\sum x_i^2 < u^{-2}} G(\mathbf{x}) d\mathbf{x} = \frac{1}{u^k} \int \cdots \int_{B_k} G(\mathbf{y}/u) d\mathbf{y}. \tag{30}$$

Continuity of G at zero implies uniform convergence of $G(\mathbf{y}/u)$ to $G(\mathbf{0})$ over the ball B_k, and thus

$$F(u) = \mathrm{Vol}\,(B_k)\, G(\mathbf{0}) \frac{1}{u^k}\,(1 + o(1)). \tag{31}$$

Dividing (31) by (29) we get that the value of $K_{G,\alpha}$ in (24) coincides with (25).

Now assume G is differentiable in some neighborhood of zero and consider the Lagrange form of the Taylor expansion of $G(\mathbf{y}/u)$. The latter and (30) give

$$|F(u) - K_{G,\alpha}\, f(u)| \leq \frac{1}{u^k}\left|\mathrm{Vol}\,(B_k)\, G(\mathbf{0}) - u^k K_{G,\alpha}\, f(u)\right| + \frac{1}{u^{k+1}}\left|\underset{B_k}{\int\cdots\int} \nabla G\left(\xi(\mathbf{y})\mathbf{y}\right)\mathbf{y}^T d\mathbf{y}\right|,$$

where $0 \leq \xi(\mathbf{y}) \leq 1/u$. The second summand in the right-hand side above is bounded by

$$\frac{1}{u^{k+1}}\mathrm{Vol}\,(B_k) \sup_{B_k} \|\nabla G(\mathbf{x}/u)\|,$$

and the bound for the remaining summand follows from (29), where we note that $f(u)$ is bounded by the two successive partial sums in its alternated series (29) and that the factors before $\mathrm{Vol}\,(B_k)\, G(\mathbf{0})$ in the expression for $K_{G,\alpha}$ and before the square brackets in (29) cancel out. The last step is to use formulas (25) and (27) to express $\mathrm{Vol}\,(B_k)\, G(\mathbf{0})$ in terms of $K_{G,\alpha}$ and C_2.

We move on to the proof of (28). Taylor expansion for $G(\mathbf{y}/u)$ yields

$$F(u) = \frac{1}{u^k}\mathrm{Vol}\,(B_k)\, G(\mathbf{0}) + \frac{1}{u^{k+2}}\underset{B_k}{\int\cdots\int} \frac{\mathbf{y} Hess\,(G\,(\mathbf{0}))\,\mathbf{y}^T}{2} d\mathbf{y} + o\left(\frac{1}{u^{k+2}}\right),$$

where we took into account that the integral of the odd function $\nabla G(\mathbf{0})\mathbf{y}$ over the ball B_k is zero. Neglecting odd terms in $\mathbf{y} Hess\,(G\,(\mathbf{0}))\,\mathbf{y}^T$ we have

$$\underset{B_k}{\int\cdots\int} \mathbf{y} Hess\,(G\,(\mathbf{0}))\,\mathbf{y}^T d\mathbf{y} = \sum \underset{B_k}{\int\cdots\int} \frac{\partial^2 G(\mathbf{0})}{\partial^2 y_i} y_i^2 d\mathbf{y} =$$

$$= \left(\sum \frac{\partial^2 G(\mathbf{0})}{\partial^2 y_i}\right) \underset{B_k}{\int\cdots\int} \frac{\sum y_i^2}{k} d\mathbf{y} = \mathrm{Vol}\,(B_k)\,\frac{tr(Hess(G(\mathbf{0}))}{k+2},$$

where the last integral was computed using spherical coordinates. Substituting the second order Taylor expansion for $F(u)$ and expression for $f(u)$ in (29) into the left-hand side of (28) we get $L_{G,\alpha}$. □

Note that the expression $\alpha^{-1} K_{G,\alpha}$ does not depend on α and thus the right-hand side of (26) and (28) depends only on the integrand G in (23) and parameters m and k.

Corollary 8.1.1. *Let $t_k(u)$ be the Student $t-$distribution tail with k degrees of freedom. If G is continuous at zero, then*

$$\frac{F(u)}{t_k(u)} = K_{G,2} + o(1) \quad as \quad u \to \infty,$$

where $K_{G,2}$ is given by (25) with $m = 1$. Statements (B) and (C) also hold for $f(u) = t_k(u)$, provided $m = 1$ and $\alpha = 2$.

Proof. Note that $t_k(u) = \frac{1}{2}F_{1,k}(u^2)$ and apply Lemma 8.1. □

2. A note on the regularity constraints and the continuity assumption

The aim of this section is to replace the technical constraints (3), (9) and (17) of Theorems 2.1, 3.1 and 5.1 by simpler criteria, and to weaken the assumption of continuity of the multivariate density $g(\mathbf{x})$ of the data vector $\mathbf{X}$.

The nature of the regularity constraints (3), (9) and (17) becomes clear if one notes that all the proofs share a common part, which is to apply Lemma 8.1 (A) or Corollary 8.1.1 to the representation for the distribution tail of the test statistic T, see (7) and (13), and then to use dominated convergence theorem to show that the corresponding function $G(\mathbf{t})$ is continuous at zero. The only purpose of the regularity constraints is to ensure that the limiting and integration operations are interchangeable, and that the resulting constant K_g is finite. Omitting the regularity assumptions (3), (9) and (17) we immediately obtain

Theorem 9.1 ("liminf" analogue of Theorems 2.1, 3.1 and 5.1)**.** *Let T be the Student one- or two-sample $t-$statistic or an $F-$statistic and let $t(u)$ be the distribution tail of T under the null hypothesis $H_0:\ g \sim MVN(\mathbf{0}, \sigma^2 \mathbf{1}_n)$, where $\sigma^2 > 0$ and $\mathbf{1}_n$ is the identity matrix. If g is continuous, then*

$$\liminf_{u\to\infty} \frac{\boldsymbol{P}(T > u)}{t(u)} \geq K_g,$$

where the constant K_g is given by (4), (10) and (18) accordingly, though it may not be finite.

Next, we give the sufficient (but not necessary) conditions for the regularity constraints of Theorems 2.1, 3.1 and 5.1 to hold. One may expect that formulas (5), (11) and (19) hold when g is continuous and K_g is finite, but proving or disproving this claim is not easy and it remains an open problem.

Lemma 9.1. *If $g(\mathbf{x})$ is bounded and there exist positive constants R, C and δ such that*

$$g(\mathbf{x}) \leq \frac{C}{\|\mathbf{x}\|^{n+\delta}} \quad \text{for} \quad \|\mathbf{x}\| > R, \tag{32}$$

then the assumptions (3), (9) and (17) of Theorems 2.1, 3.1 and 5.1 hold.

Proof. The integrals in (3), (9) and (17) will be estimated by partitioning the integration domain into several disjoint parts D_i and D_j^* and analyzing the integrals over these sets separately. For non-compact domains D_j^* the integrand will be estimated from above using the bound (32) and showing that this bound is integrable. The integrability over the compact domains D_i follows from the fact that $g(\mathbf{x})$ is bounded. In the notation below let $G(r)$, $G(\omega, r)$ and $G(\mathbf{x}, r)$ be the integrands in (3), (9) and (17) accordingly.

Student's one-sample $t-$statistic: Set $D_1 = [0, R]$ and $D_1^* = [R, \infty]$. Since $\|\mathbf{I}\| = 1$ and $\mathbf{I}$ and $\boldsymbol{\xi}$ are orthogonal, we have $\|r\,(\mathbf{I} + \boldsymbol{\xi})\,\|^2 = r^2(1 + \|\boldsymbol{\xi}\|^2) \geq r^2$ and the bound (32) gives

$$\int\limits_{D_1^*} G(r)dr < \int\limits_R^\infty \frac{C}{r^{1+\delta}} dr < \infty.$$

Student's two-sample $t-$statistic: Setting $D_1 = [-\pi/2, \pi/2] \times [0, R]$ and $D_1^* = [-\pi/2, \pi/2] \times [R, \infty]$ and noting that $\mathbf{I}_1$, $\mathbf{I}_2$ and $\boldsymbol{\xi}$ are mutually orthogonal we get

$$\|r\left(\cos(\omega - \omega_0)\mathbf{I}_1 + \sin(\omega - \omega_0)\mathbf{I}_2 + \boldsymbol{\xi}\right)\|^2 = r^2(1 + \|\boldsymbol{\xi}\|^2) \geq r^2,$$

where we used the fact that $\|\mathbf{I}_1\| = \|\mathbf{I}_2\| = 1$. Now the bound (32) implies

$$\int\limits_{D_1^*} G(\omega, r)dr < \int\limits_{-\pi/2}^{\pi/2} \cos(\omega)^{n-2} d\omega \times \int\limits_R^\infty \frac{C}{r^{1+\delta}} < \infty.$$

$F-$*statistic*: Consider the following partition of $\mathbb{R}^{n_1+1}$: $D_1 = \Big\{(\mathbf{x}, r) : \|\mathbf{x}\| \le R\ ,\ |r| \le R\Big\}$, $D_1^* = \Big\{(\mathbf{x}, r) : \|\mathbf{x}\| \le R\ ,\ |r| > R\Big\}$, and $D_2^* = \Big\{(\mathbf{x}, r) : \|\mathbf{x}\| > R\Big\}$. Since $\mathbf{I}$ and $\boldsymbol{\xi}$ are orthogonal and $\|\mathbf{I}\| = 1$ we have $\|\ (\mathbf{x}, r\mathbf{I} + s_1(\mathbf{x})\boldsymbol{\xi})\ \|^2 = \|\mathbf{x}\|^2 + r^2 + s_1(\mathbf{x})^2\|\boldsymbol{\xi}\|^2 \ge \|\mathbf{x}\|^2 + r^2$, and then

$$\int\limits_{D_1^*}\cdots\int G(\mathbf{x}, r)\,dr\,d\mathbf{x} < \int\limits_{\|\mathbf{x}\|\le R}\cdots\int s_1(\mathbf{x})^{n_2-1}d\mathbf{x}\ \times \int\limits_{|r|>R}\frac{C}{|r|^{n+\delta}}dr < \infty$$

and

$$\int\limits_{D_2^*}\cdots\int G(\mathbf{x}, r)\,dr\,d\mathbf{x} < \int\limits_{\|\mathbf{x}\|>R}\cdots\int\int\limits_{-\infty}^{\infty}\frac{s_1(\mathbf{x})^{n_2-1}}{\left(\|\mathbf{x}\|^2 + r^2\right)^{\frac{n+\delta}{2}}}dr\,d\mathbf{x} <$$

$$< \int\limits_{\|\mathbf{x}\|>R}\cdots\int\frac{s_1(\mathbf{x})^{n_2-1}}{\|\mathbf{x}\|^{n-1+\delta}}d\mathbf{x}\ \times \int\limits_{-\infty}^{\infty}\frac{1}{(1 + r^2)^{n/2}}dr < \infty,$$

where the integral in the last inequality is computed passing to spherical coordinates. □

Note that in the i.i.d case the condition (32) is equivalent to the existence of the $n - 1 + \delta$ moment of the marginal density $h(x)$. For the Student one-sample $t-$test, however, the criterium of Lemma 9.1 is “too strict”, see below.

Definition. Multivariate density $g(\mathbf{x})$ has the asymptotic monotonicity property if there exists a constant M such that for any $1 \le i \le n$ and any constants $c_j,\ j \ne i$, the function $f(x) = g(c_1, .., c_{i-1}, x, c_{i+1}, .., c_n)$ is monotone on $[M, \infty)$.

Lemma 9.2. *If K_g is finite and $g(\mathbf{x})$ is bounded and has the asymptotic monotonicity property, then the assumption (3) holds.*

Proof. Setting ε equal to $(2\sqrt{n})^{-1}$ and using asymptotic monotonicity property we get that the integral in (3) is bounded by

$$\int\limits_0^{2M\sqrt{n}} r^{n-1}\sup_{\|\boldsymbol{\xi}\|<\frac{1}{2\sqrt{n}}} g\left(r\left(\mathbf{I} + \boldsymbol{\xi}\right)\right)dr + \int\limits_{2M\sqrt{n}}^{\infty} r^{n-1}g\left(r\frac{\mathbf{I}}{2}\right)dr < \infty.$$

The first summand is finite owing to the boundness of g and the finiteness of the second summand is equivalent to the finiteness of K_g. □

Asymptotic monotonicity and finiteness of K_g are very mild constraints. For the i.i.d case of the Student one-sample $t-$test, for example, Lemma 9.2 implies that the statement of Theorem 2.1 holds for any continuous marginal density $h(\mathbf{x})$ that has monotone tails and such that $K_g < \infty$, and the latter assumption is weaker than the assumption of existence of the first moment and holds even for such heavy tailed densities as Cauchy.

Unfortunately there is no asymptotic monotonicity criterium analogue for the case of the Student two-sample $t-$ and $F-$statistics, and the constant K_g in (10) and (18) may be infinite for some heavy-tailed densities, cf. (Bradley, 1952a).

Finally, in the proofs of Theorems 2.1, 3.1 and 5.1 one may have used the “almost everywhere” version of the dominated convergence theorem. For the Student one-sample $t-$statistic the assumption of continuity of g can be replaced by the assumption that $g(\mathbf{x})$ is continuous function of $\mathbf{x} \in \mathbb{R}^n$ a.e. on the set of points $\mathbf{x} = r\mathbf{I},\quad r > 0$, for the Student two-sample $t-$statistic - on the set of points $\mathbf{x} = r\left(\cos(\omega - \omega_0)\mathbf{I}_1 + \sin(\omega - \omega_0)\mathbf{I}_2 + \mathbf{z}\right)$, where $r > 0$ and $\omega \in [-\pi/2\ ,\ \pi/2]$, and for the $F-$statistic - on the set of points $\mathbf{x} = \mathbb{R}^{n_1} \times r\mathbf{I},\ r \in \mathbb{R}$. Here a.e. means almost everywhere with respect to the Lebesque measure induced by the measure of the linear space L in (3), (9) and (17).

Appendix B: Figures

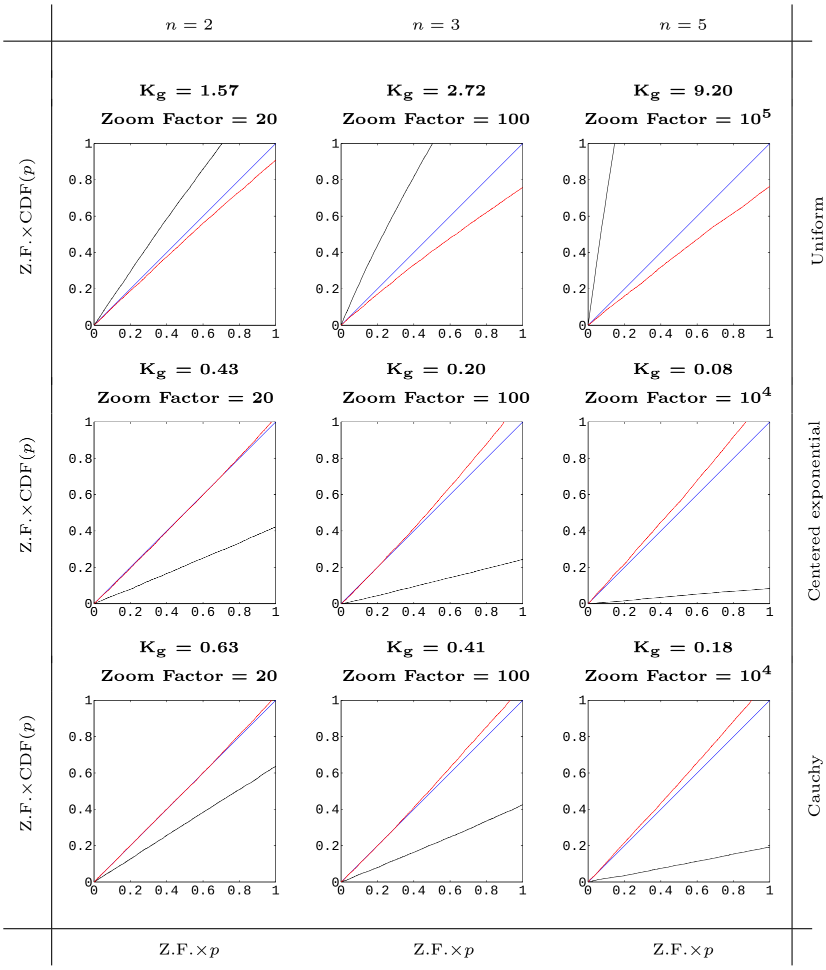

Figure 2. The eCDF of the p-values for the Student one-sample $t-$test. The empirical CDFs of the raw and corrected p-values p^R and p^C are shown in black and red accordingly. The top, middle and bottom rows correspond to the $Uniform(-1,1)$, *Centered exponential* and *Cauchy* densities, and left, middle and right columns correspond to sample sizes $n = 2$, $n = 3$ and $n = 5$. The blue diagonal line is the theoretical uniform distribution. The axes are scaled according to the Zoom Factor (Z.F.) parameter r in the title of the graphs.

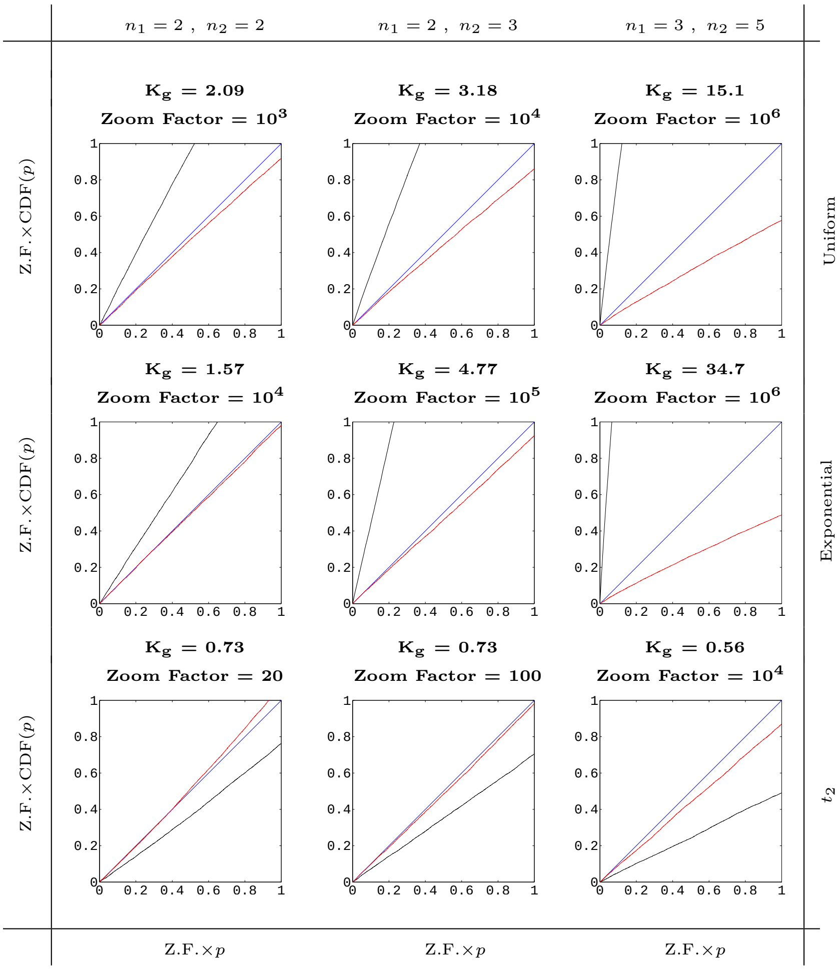

Figure 3. The eCDF of the p-values for the Student two-sample $t-$test. The empirical CDFs of the raw and corrected p-values p^R and p^C are shown in black and red accordingly. The top, middle and bottom rows correspond to the $Uniform(-1,1)$, *Exponential* and t_2 densities, and left, middle and right columns correspond to sample sizes $(n_1 = 2, n_2 = 2)$, $(n_1 = 2, n_2 = 3)$, and $(n_1 = 3, n_2 = 5)$. The blue diagonal line is the theoretical uniform distribution. The axes are scaled according to the Zoom Factor (Z.F.) parameter r in the title of the graphs.

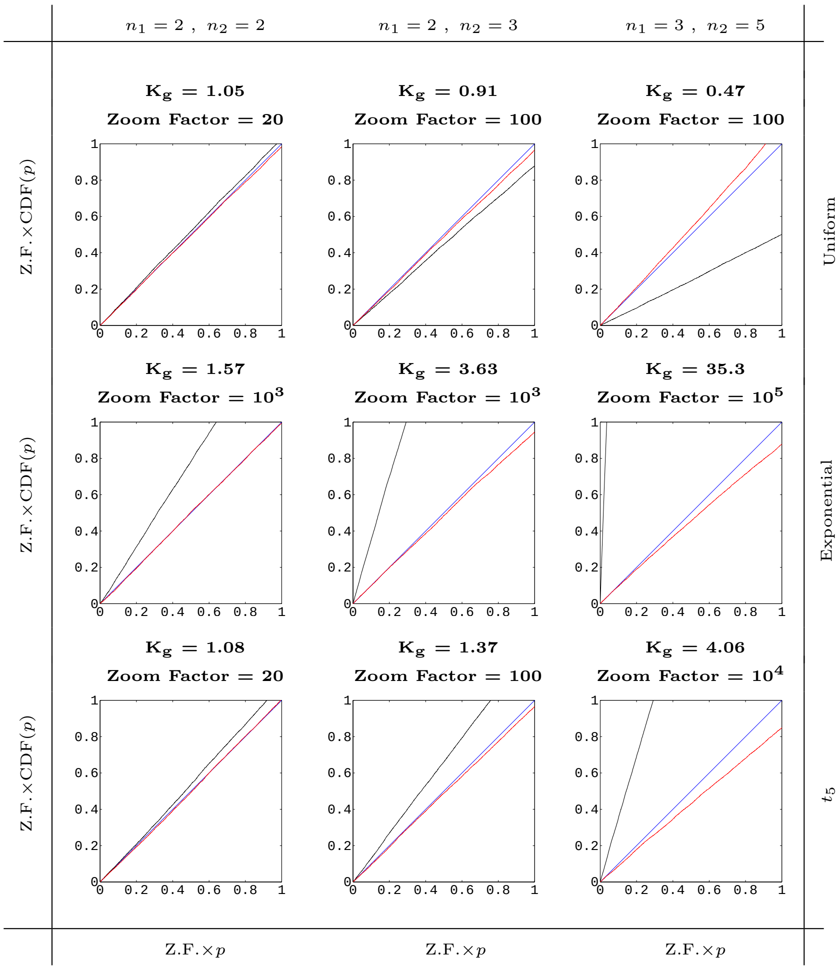

Figure 4. The eCDF of the p-values for the F−test (equality of variances). The empirical CDFs of the raw and corrected p-values p^R and p^C are shown in black and red accordingly. The top, middle and bottom rows correspond to the $Uniform(-1,1)$, *Exponential* and t_5 densities, and left, middle and right columns correspond to sample sizes $(n_1 = 2, n_2 = 2)$, $(n_1 = 2, n_2 = 3)$, and $(n_1 = 3, n_2 = 5)$. The blue diagonal line is the theoretical uniform distribution. The axes are scaled according to the Zoom Factor (Z.F.) parameter r in the title of the graphs.

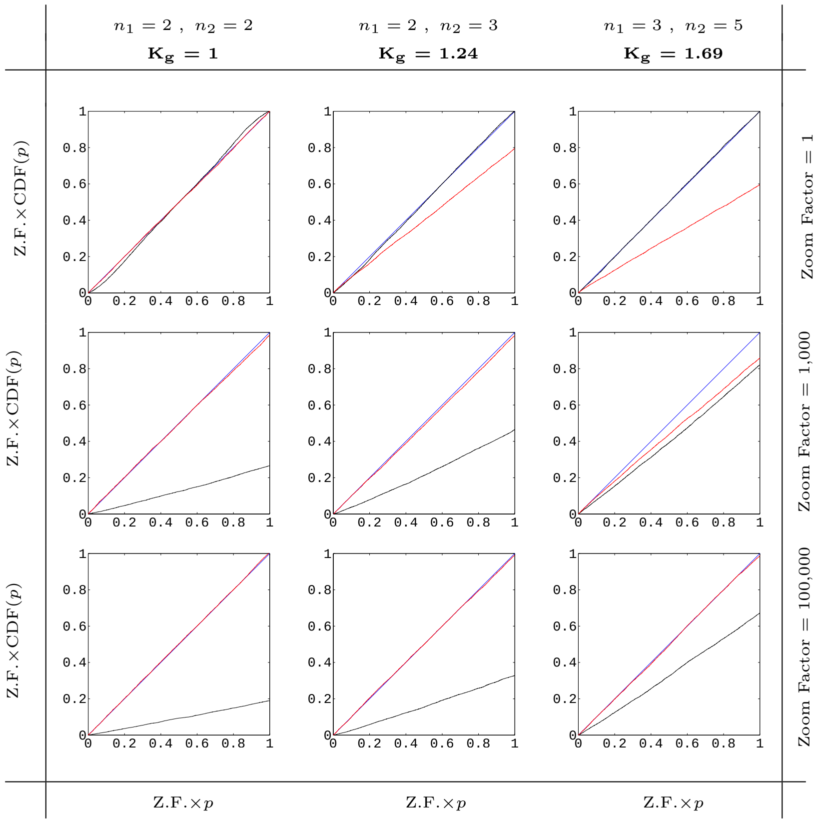

Figure 5. The distribution tails of the p-values for the Welch test. The empirical CDFs of the raw (Welch-Satterthwaite) and corrected p-values p^R and p^{WS} for the *Standard Normal* density are shown in black and red accordingly. The top, middle and bottom rows correspond to the different values of the Zoom Factor (Z.F.) parameter r shown on the right, and the axes are scaled accordingly. The left, middle and right columns correspond to sample sizes $(n_1 = 2, n_2 = 2)$, $(n_1 = 2, n_2 = 3)$, and $(n_1 = 3, n_2 = 5)$. The blue diagonal line is the theoretical uniform distribution.

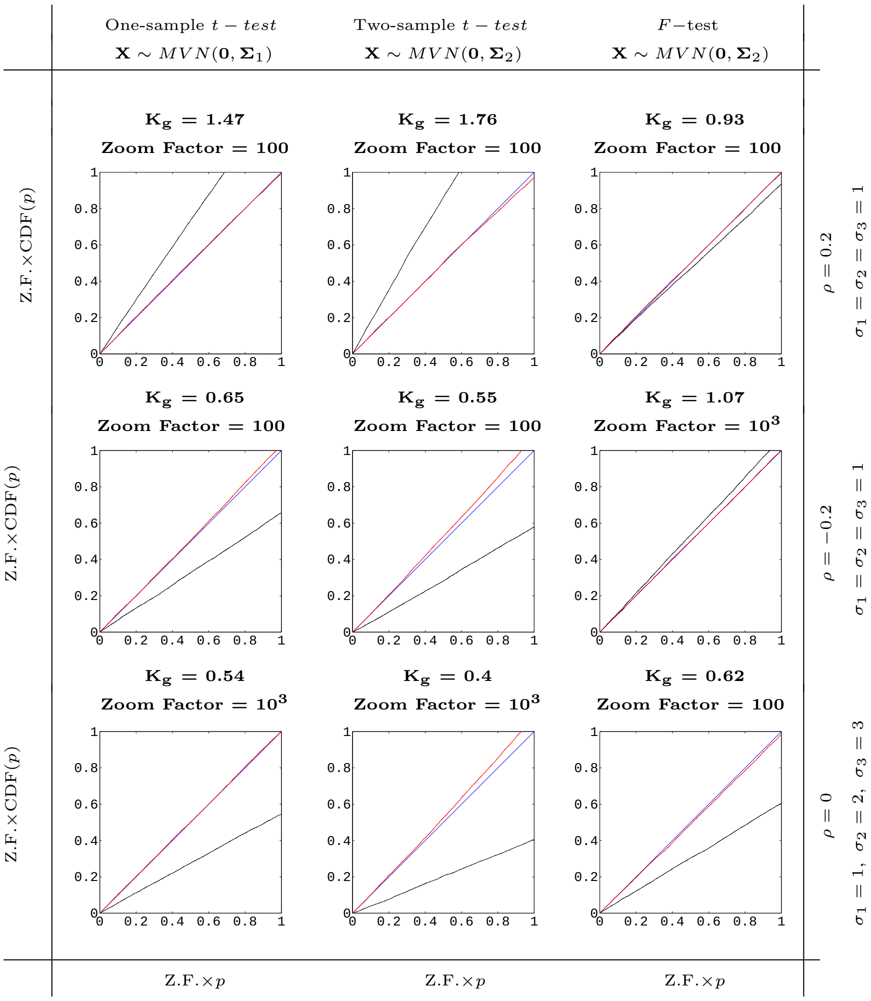

Figure 6. The effect of dependency and non-homogeneity of data on $\mathbf{P}(T > u)$ as $u \to \infty$. The empirical CDF of raw (black) and corrected (red) p-values. Analogue of Figures 2, 3 and 4 for dependent (top row - positively correlated observations; middle row - negatively correlated observations) and non-homogeneous (bottom row, unequal variances) data. Multivariate normal case with covariance matrices

$$\mathbf{\Sigma}_1 = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & 0 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \rho\sigma_2\sigma_3 \\ 0 & \rho\sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix}, \quad \mathbf{\Sigma}_2 = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & 0 & 0 & 0 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma_1^2 & \rho\sigma_1\sigma_2 & 0 \\ 0 & 0 & \rho\sigma_1\sigma_2 & \sigma_2^2 & \rho\sigma_2\sigma_3 \\ 0 & 0 & 0 & \rho\sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix}.$$

Supplementary Materials

MATLAB scripts

[*OST*/*TST*/*WELCH*/*F*]+*ComputeKg.m* - compute K_g for the Student one- and two- sample $t-$, Welch, and $F-$ statistics using adaptive Simpson or Lobatto quadratures. Here g is an arbitrary multivariate density.[1]

[*TST*/*WELCH*/*F*]+*ComputeKgIS*+.m - the same as above but for the case where samples are independent.[2]

[*OST*/*TST*/*WELCH*/*F*]+*ComputeKgIID*+.m - the same as above but assuming that the samples consist of i.i.d. random variables.[2]

RunSimulation+[*IID*/*MVN*]+.m - perform simulation study for i.i.d. and dependent/non-homogeneous cases, see Section 7 and Appendix B.

Wolfram Mathematica scripts

[*OST*/*TST*/*WELCH*/*F*]+*ComputeKg.nb* - compute the exact expression for K_g for an arbitrary multivariate density g and given sample size(s). We include a number of examples, such as evaluation of K_g for the zero-mean Gaussian case with an arbitrary covariance matrix $\mathbf{\Sigma}$; the "unequal variances" case for the Student two-sample $t-$ and Welch statistics; and evaluation of K_g for the densities considered in the simulation study.

OSTComputeKgIID.nb - verifies the constants in Table 1 for the i.i.d. case of the Student one-sample $t-$statistic.

TSTExactPDF.nb and *WELCHExactPDF.nb* - the exact distribution for the Student two-sample $t-$ and Welch statistics for odd sample sizes, see (Ray and Pitman, 1961).

Other Materials

Supplementary-Materials.pdf - Remarks on Theorem 1.1 and its application to real data; extended version of the literature review; comparison of the result of Theorem 1.1 with the exact distribution of the Welch statistic; proof of Theorem 5.1.

The Supplementary Materials are available at www.zholud.com

[1] For the $F-$statistic we use Monte Carlo integration.
[2] For the $F-$statistic and $n_1 > 3$ we use Monte Carlo integration.

Acknowledgements

The author thanks Holger Rootzén for assistance and fruitful discussions and the associate editor for helpful comments which led to improvement of the presentation of the present paper.

References

R.A. Bradley. The distribution of the t and F statistics for a class of non-normal populations. *Virginia J. Sci.*, 3:1–32, 1952a.

R.A. Bradley. Corrections for non-normality in the use of the two-sample $t-$ and $F-$tests at high significance levels. *Ann. Math. Statist.*, 23:103–113, 1952b.

H.E. Daniels and G.A. Young. Saddlepoint approximation for the Studentized mean, with an application to the bootstrap. *Biometrika*, 78:169–179, 1991.

C. Field and E. Ronchetti. *Small Sample Asymptotics.* IMS, Hayward, CA, 1990.

A.K. Gaen. The distribution of "Student's" t in random samples of any size drawn from non-normal universes. *Biometrika*, 36:353–369, 1949.

A.K. Gaen. The distribution of the variance ratio in random samples of any size drawn from non-normal universes. *Biometrika*, 37:236–255, 1950.

P. Hall. Edgeworth expansion for Student's t statistic under minimal moment conditions. *Ann. Probab.*, 15: 920–931, 1987.

S.I. Hayek. *Advanced Mathematical Methods in Science and Engineering.* Marcel Dekker, New York, NY, 2001.

H. Hotelling. The behavior of some standard statistical tests under non-standard conditions. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1:319–360, 1961.

B.-Y. Jing, Q.M. Shao, and W. Zhou. Saddlepoint approximation for Student's $t-$statistic with no moment conditions. *Ann. Statist.*, 32(6):2679–2711, 2004.

Mathematica. *Version 8.0.* Wolfram Research, Inc., Champaign, IL, 2010.

MATLAB. *Version 7.10.0 (R2010a).* The MathWorks, Inc., Natick, Massachusetts, 2010.

W.D. Ray and A.E.N.T. Pitman. An exact distribution of the Fisher-Behrens-Welch statistic for testing the difference between the means of two normal populations with unknown variance. *Journal of the Royal Statistical Society. Series B (Methodological)*, 23(2):377–384, 1961.

H. Rootzén and D. Zholud. Efficient estimation of the number of false positives in high-throughput screening. *Submitted to Biometrika*, 2014.

SmartTail, 2013. *url:* www.smarttail.se - software for the analysis of false discovery rates in high-throughput screening experiments. *Username:* Bernoulli, *Password:* PrkQ27.

J.D. Storey. A direct approach to false discovery rates. *J. R. Statist. Soc. B*, 64(3):479–498, 2002.

J.D. Storey. The positive false discovery rate: a Bayesian interpretation and the $q-$value. *The annals of Statistics*, 31(6):2013–2035, 2003.

J.D. Storey. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Statist. Soc. B*, 66(1):187–205, 2004.

J. Warringer and A. Blomberg. Automated screening in environmental arrays allows analysis of quantitative phenotypic profiles in saccharomyces cerevisiae. *Yeast*, 20:53–67, 2003.

J. Warringer, E. Ericson, L. Fernandez, O. Nerman, and A. Blomberg. High-resolution yeast phenomics resolves different physiological features in the saline response. *Proc Natl Acad Sci USA*, 100(26):15724–15729, 2003.

D. Zholud. *Extreme Value Analysis of Huge Datasets: Tail Estimation Methods in High-Throughput Screening and Bioinformatics.* PhD thesis, Göteborg University. ISBN: 978-91-628-8354-6, 2011.

W. Zhou and B.-Y. Jing. Tail probability approximations for Student's $t-$statistics. *Probability Theory and Related Fields*, 136(4):541–559, 2006.