TEST
https://doi.org/10.1007/s11749-021-00758-y

ORIGINAL PAPER

QANOVA: quantile-based permutation methods for general factorial designs

Marc Ditzhaus[1] · **Roland Fried**[1] · **Markus Pauly**[1]

Received: 14 May 2020 / Accepted: 25 January 2021

Abstract
Population means and standard deviations are the most common estimands to quantify effects in factorial layouts. In fact, most statistical procedures in such designs are built toward inferring means or contrasts thereof. For more robust analyses, we consider the population median, the interquartile range (IQR) and more general quantile combinations as estimands in which we formulate null hypotheses and calculate compatible confidence regions. Based upon simultaneous multivariate central limit theorems and corresponding resampling results, we derive asymptotically correct procedures in general, potentially heteroscedastic, factorial designs with univariate endpoints. Special cases cover robust tests for the population median or the IQR in arbitrary crossed one-, two- and higher-way layouts with potentially heteroscedastic error distributions. In extensive simulations, we analyze their small sample properties and also conduct an illustrating data analysis comparing children's height and weight from different countries.

Keywords Birth cohorts · IQR · Main and interaction effects · Median · Permutation tests

Mathematics Subject Classification 62G09 · 62G10 · 62G20

1 Introduction

Factorial designs are popular in various fields such as ecology, biomedicine and psychology (Cassidy et al. 2008; Mehta et al. 2010; Kurz et al. 2015) as they allow us to study interaction effects between different factors alongside their main effects. In fact,

The work of Marc Ditzhaus and Markus Pauly was funded by the *Deutsche Forschungsgemeinschaft* (Grant No. PA-2409 5-1).

✉ Marc Ditzhaus
marc.ditzhaus@tu-dortmund.de

[1] Department of Statistics, TU Dortmund University, Dortmund, Germany

Published online: 24 February 2021

Lubsen and Pocock (1994) pointed out that *"it is desirable for reports of factorial trials to include estimates of the interaction between the treatments."* The ANOVA–F-test is the most common tool for this but suffers from restrictive assumptions such as homoscedasticity and normality. Thus, several tests have been developed that allow for non-normal errors or are valid for heteroscedastic one- and two-way or even more general factorial designs (Johansen 1980; Brunner et al. 1997; Bathke et al. 2009; Pauly et al. 2015; Friedrich et al. 2017a, b; Harrar et al. 2019).

All these procedures describe effects by (contrasts of) means. This is in line with a phenomenon observed in various areas: Comparisons are mainly based upon means or variances but not on their robust counterparts. This can be explained in part by the simplicity and elegance gained by using linear or, under independence, additive statistics. Nevertheless, it contradicts the important role of statistics based on quantiles, like the median and the interquartile range (IQR), in data exploration and modeling, e.g., in boxplots or summary statistics. The interest in analyzing quantiles has led to the development of quantile regression, which is commonly established nowadays (Koenker and Hallock 2001; Koenker et al. 2019). However, as, e.g., stressed by Beyerlein (2014) *"it appears to be quite underused in medical research."* One reason may be that, although there exist several approaches for specific designs (Sen 1962; Potthoff 1963; Fung 1980; Hettmansperger and McKean 2010; Fried and Dehling 2011; Chung and Romano 2013), there does not exist an equal abundance of methods based on quantiles for general factorial designs. There are procedures, at least for the median, but they often require strong distributional assumptions (as symmetry) or, at least, an extension to factorial designs is missing. Therefore, the *main aims* of the present paper are to develop inference procedures (tests and compatible confidence regions)

(i) for the median, the IQR or any linear combination of quantiles.
(ii) for factorial designs to study robust main and interaction effects.
(iii) for general heterogeneous or heteroscedastic models beyond normality.
(iv) being theoretically valid and performing satisfactorily for finite samples.

To achieve these goals, we combine and extend the ideas of Chung and Romano (2013) (tests for equality of medians in one-way ANOVA models) and Pauly et al. (2015) (mean-based testing procedures in general factorial designs) to (simultaneously) infer arbitrary linear contrasts of general quantiles. In view of (ii) and (iv), we follow the idea of *permuting studentized Wald-type statistics* to obtain methods that are finitely correct in case of exchangeable data (e.g., under the null hypothesis of equal means/medians in the classic F-ANOVA normal model) *but also* asymptotically valid for general non-exchangeable settings. This alluring technique has originally been developed for special two-sample models (Neuhaus 1993; Janssen 1997; Janssen and Pauls 2003; Pauly 2011) and has recently displayed its full strength to obtain accurate methods in one-way (Chung and Romano 2013) and more general factorial designs (Pauly et al. 2015; Friedrich et al. 2017a; Smaga 2017; Harrar et al. 2019).

However, to derive the fore-mentioned theoretical evidence in our general quantile-based approaches we could not employ the methods derived in the previously mentioned papers. In fact, to overcome some technical difficulties that occur when jointly permuting sample quantiles, we had to take a detour in which we extended

some results for general permutation empirical processes and uniform Hadamard differentiability (van der Vaart and Wellner 1996) that are of own mathematical interest. Anyhow, this finally results in (i)–(iv), i.e., a flexible toolbox for inferring contrasts of different quantiles in factorial designs. In the special case of the median and its bootstrap-based variance estimator, we obtain the one-way permutation test derived in Chung and Romano (2013).

Outline: We first introduce the model, estimators for population quantiles and how to formulate null hypotheses in them to test for certain main or interaction effects. In Sect. 3, we state the theory to handle the joint asymptotics for sample quantiles and their covariance matrix estimators. As the latter are crucial to obtain the correct dependency structure, we study different approaches: kernel density estimators, bootstrapping or certain interval estimates. As they are mostly only known for the sample median, we explain in Sects. 3.1–3.3 how to extend them to our general situations. From these findings, we deduce three different asymptotically valid testing procedures. To improve their small sample performance, we consider their respective permutation versions in Sect. 4, prove asymptotic exactness and analyze their power under local and fixed alternatives. To compare the small sample behavior of the resulting tests, we conducted extensive simulations presented in Sect. 5. Finally, we illustrate the new methodology by analyzing a recent dataset on height and weight of children from different countries in Sect. 6. All proofs details to higher-way layouts and additional simulation results are deferred to supplement.

2 The setup

We consider a general model given by mutually independent random variables, e.g., corresponding to the outcome from independent patients in randomized clinical trials,

$$X_{ij} \sim F_i \quad (i = 1, \ldots, k;\ j = 1, \ldots, n_i) \tag{1}$$

with absolutely continuous distribution functions F_i and densities f_i. This setup allows to incorporate factorial structure of different kinds by adequately splitting up indices. To accept this, consider, e.g., a two-way design with factors A (a levels) and B (b levels). Setting $k = a \cdot b$, we split up the group index $i = (i_1, i_2)$ and model observations as $X_{i_1 i_2 j} \sim F_{i_1 i_2}$ ($i_1 = 1, \ldots, a$; $i_2 = 1, \ldots, b$). Factorial designs of more complexity can be incorporated similarly (Pauly et al. 2015).

Having the model fixed, we now turn to the parameters of interest: Choosing $m \in \mathbb{N}$ different probabilities $0 < p_1 < \ldots < p_m < 1$, we want to study inference methods for the corresponding quantiles

$$q_{ir} = F_i^{-1}(p_r) = \inf\{t \in \mathbb{R} : F_i(t) \geq p_r\} \quad (i = 1, \ldots, k; r = 1, \ldots, m). \tag{2}$$

Pooling them in $\mathbf{q} = (\mathbf{q}_1', \ldots, \mathbf{q}_k')' = (q_{11}, \ldots, q_{1m}, q_{21}, \ldots, q_{km})'$, we are particularly interested in testing the *QANOVA* null hypothesis $\mathcal{H}_0 : \mathbf{Hq} = \mathbf{0}_r$ for a contrast matrix $\mathbf{H} \in \mathbb{R}^{r \times km}$ of interest. Here, $\mathbf{H}$ is called a contrast matrix if $\mathbf{H}\mathbf{1}_{km} = \mathbf{0}_r$, where $\mathbf{1}_d$ and $\mathbf{0}_d$ are vectors in $\mathbb{R}^d$ consisting of 1's and 0's only. Choosing the con-

trast matrices in line with the design and the question of interest allows us to test various hypotheses about main and interaction effects. Moreover, we want to point out that respective confidence regions for corresponding contrasts of quantiles can be obtained straightforwardly by inverting the test procedures. In what follows, we will therefore focus on hypothesis testing but provide some exemplary confidence intervals in the context of the illustrative data analyses in Sect. 6. Turning back to the null hypothesis, we recall from general ANOVA that it is convenient to re-formulate it as $\mathcal{H}_0 : \mathbf{Tq} = \mathbf{0}_{km}$ for the unique projection matrix $\mathbf{T} = \mathbf{H}'(\mathbf{HH}')^{+}\mathbf{H}$ (Brunner et al. 1997; Pauly et al. 2015; Smaga 2017). Here, $\mathbf{A}^{+}$ denotes the Moore–Penrose inverse of the matrix $\mathbf{A}$. In fact, both matrices, $\mathbf{H}$ and $\mathbf{T}$, describe the same null hypothesis, while $\mathbf{T}$ has preferable properties as being symmetric and idempotent. To infer $\mathcal{H}_0$, we propose sensitive test statistics in the vector of corresponding sample quantiles. To introduce them, let $\widehat{F}_i(t) = n_i^{-1}\sum_{j=1}^{n_i} 1\{X_{ij} \leq t\}$ and $\widehat{F}(t) = n^{-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i} 1\{X_{ij} \leq t\}$ denote the group-specific and pooled empirical distribution function, respectively, where $n = \sum_{i=1}^{k} n_i$. Let $X_{1:n_i}^{(i)} \leq \ldots \leq X_{n_i:n_i}^{(i)}$ be the order statistics of group i. Then, the natural estimator of the quantile q_{ir} is

$$\widehat{q}_{ir} = \widehat{F}_i^{-1}(p_r) = \inf\{t \in \mathbb{R} : \widehat{F}_i(t) \geq p_r\} = X_{\lceil n_i p_r \rceil:n_i}^{(i)}. \tag{3}$$

Examples of specific hypotheses To give some examples of hypotheses covered within this framework, we first consider a one-way design. For $m = 1$, we obtain the k-sample null hypothesis

– *No group effect:* $\mathcal{H}_0 = \{\mathbf{P}_k\mathbf{q} = \mathbf{0}_k\} = \{q_1 = \ldots = q_k\}$ with $\mathbf{P}_k = \mathbf{I}_k - \mathbf{J}_k/k$.

Here, $\mathbf{I}_k \in \mathbb{R}^{k\times k}$ denotes the unit matrix, $\mathbf{J}_k = \mathbf{1}_k\mathbf{1}_k'$ and we suppressed the second index of the quantiles ($m = 1$). Choosing $p_1 = 1/2$ gives the null hypothesis of equal medians which reduces to the null hypothesis of equal means in case of symmetric error distributions.

Setting $k = ab$, we consider a two-way design with factors A (having levels $i_1 = 1, \ldots, a$) and B (with levels $i_2 = 1, \ldots, b$) and suppose that we like to formulate main and interaction effects in terms of quantiles:

– *No main effect A*: $\mathcal{H}_0 = \{\mathbf{H}_A\mathbf{q} = \mathbf{0}_{ab}\} = \{\bar{q}_{1\cdot} = \ldots = \bar{q}_{a\cdot}\}$,
– *No main effect of B*: $\mathcal{H}_0 = \{\mathbf{H}_B\mathbf{q} = \mathbf{0}_{ab}\} = \{\bar{q}_{\cdot 1} = \ldots = \bar{q}_{\cdot b}\}$,
– *No interaction effect*: $\mathcal{H}_0 = \{\mathbf{H}_{AB}\mathbf{q} = \mathbf{0}_{ab}\} = \{\bar{q}_{\cdot\cdot} - \bar{q}_{\cdot i_2} - \bar{q}_{i_1\cdot} + q_{i_1 i_2} \equiv 0\}$,

where $\mathbf{H}_A = \mathbf{P}_a \otimes (\mathbf{J}_b/b)$, $\mathbf{H}_B = (\mathbf{J}_a/a) \otimes \mathbf{P}_b$ and $\mathbf{H}_{AB} = \mathbf{P}_a \otimes \mathbf{P}_b$. Here, $\otimes$ is the Kronecker product and $\bar{q}_{i_1\cdot}$, $\bar{q}_{\cdot i_2}$ and $\bar{q}_{\cdot\cdot}$ are the means over the dotted indices. The latter hypotheses can also be described more lucid by utilizing an additive effects notation. To this end, we decompose the quantile $q_{i_1 i_2} = q^{\mu} + q_{i_1}^{\alpha} + q_{i_2}^{\beta} + q_{i_1 i_2}^{\alpha\beta}$ from group (i_1, i_2) into a general effect q^{μ}, main effects $q_{i_1}^{\alpha}$ and $q_{i_2}^{\beta}$ as well as an interaction effect $q_{i_1 i_2}^{\alpha\beta}$ assuming the usual side conditions $\sum_{i_1} q_{i_1}^{\alpha} = \sum_{i_2} q_{i_2}^{\beta} = \sum_{i_1} q_{i_1 i_2}^{\alpha\beta} = \sum_{i_2} q_{ij}^{\alpha\beta} = 0$. Then, the null hypotheses can be written as $\{\mathbf{H}_A\mathbf{q} = \mathbf{0}_{ab}\} = \{q_1^{\alpha} = \ldots q_a^{\alpha} = 0\}$ or $\{\mathbf{H}_{AB}\mathbf{q} = \mathbf{0}_{ab}\} = \{q_{i_1 i_2}^{\alpha\beta} \equiv 0 \text{ for all } i_1, i_2\}$. This methodology can be straightforwardly extended to higher-way layouts as described in supplementary material.

Beyond working with specific quantiles, it is also possible to infer hypotheses about linear combinations $\mathbf{c}'\mathbf{q}_i = \sum_{r=1}^{m} c_r q_{ir}$ of quantiles. Here, $\mathbf{c} \in \mathbb{R}^k$ is an arbitrary vector, e.g., choosing $c_1 = -c_2 = -1$ for $m = 2$ and setting $p_1 = 0.25$ and $p_2 = 0.75$ lead to the group-specific interquartile ranges $\mathbf{c}'\mathbf{q}_i = IQR_i$. To obtain similar hypothesis in these parameters as above, the contrast matrix has to be specified to $\widetilde{\mathbf{H}} = \mathbf{H} \otimes (c_1, \ldots, c_r)$, where $\mathbf{H}$ is one of the aforementioned contrast matrices. For example, $\mathbf{H} = \mathbf{P}_k$ together with the previous choices for $\mathbf{c}$ and p_1, p_2 gives the null hypothesis $\{IQR_1 = \cdots = IQR_k\}$ of equal IQRs among all k groups. However, the framework is much more flexible and even allows to infer hypotheses about IQRs and medians simultaneously by choosing $p_1 = 0.5$, $p_2 = 0.25$ and $p_3 = 0.75$ together with adequate contrast matrices.

3 Asymptotic results

To establish the joined asymptotic theory for the sample quantiles and their covariance matrix estimators, we assume non-vanishing groups throughout:

$$\frac{n_i}{n} \to \kappa_i > 0 \quad \text{as } \min(n_i : i = 1, \ldots, k) \to \infty. \tag{4}$$

Recall that the sample median will be asymptotically normal if the underlying density is positive and continuous in a neighborhood of the true median. This statement can be extended to the multivariate case (Serfling 2009), e.g., under the following assumption, which we consider throughout.

Assumption 1 Let F_i be continuously differentiable at q_{ir} with positive derivative $f_i(q_{ir}) > 0$ for every $r = 1, \ldots, m$ and $i = 1, \ldots, k$.

Proposition 1 (Theorem B in Sec. 2.3.3 of Serfling (2009))
We have

$$\sqrt{n}\left(\widehat{q}_{ir} - q_{ir}\right)_{r=1,\ldots,m} \xrightarrow{\mathrm{d}} \mathbf{Z}_i \quad (i = 1, \ldots, k), \tag{5}$$

where $\mathbf{Z}_i$ is a zero-mean, multivariate normal distributed random variable with non-singular covariance matrix $\mathbf{\Sigma}^{(i)}$ given by its entries

$$\mathbf{\Sigma}^{(i)}_{ab} = \kappa_i^{-1} \frac{1}{f_i(q_{ia}) f_i(q_{ib})} (p_a \wedge p_b - p_a p_b) \qquad (a, b \in \{1, \ldots, m\}). \tag{6}$$

In general, the covariance matrix is unknown and, thus, needs to be estimated. Let us suppose, for a moment, that a consistent estimator $\widehat{\mathbf{\Sigma}}^{(i)}$ for $\mathbf{\Sigma}^{(i)}$ is available. Then, we could define a *Wald-type statistic* for testing $\mathcal{H}_0 : \mathbf{Tq} = \mathbf{0}_r$

$$S_n(\mathbf{T}) = n(\mathbf{T}\widehat{\mathbf{q}})'(\mathbf{T}\widehat{\mathbf{\Sigma}}\mathbf{T}')^{+}\mathbf{T}\widehat{\mathbf{q}} \quad \text{with } \widehat{\mathbf{\Sigma}} = \bigoplus_{i=1}^{k} \widehat{\mathbf{\Sigma}}^{(i)}, \tag{7}$$

where $\oplus$ denotes the direct sum. By Proposition 1, the limiting covariance matrix $\mathbf{\Sigma} = \oplus_{i=1}^{k} \mathbf{\Sigma}^{(i)}$ is positive definite, which implies that the Moore–Penrose inverse $(\mathbf{T}\widehat{\mathbf{\Sigma}}\mathbf{T}')^{+}$ converges in probability to $(\mathbf{T}\mathbf{\Sigma}\mathbf{T}')^{+}$. Thus, $S_n(\mathbf{T})$ converges to $Z = \mathbf{Y}'(\mathbf{T}\mathbf{\Sigma}\mathbf{T}')^{+}\mathbf{Y}$ in distribution under $\mathcal{H}_0$, where $\mathbf{Y} \sim N(\mathbf{0}, \mathbf{T}\mathbf{\Sigma}\mathbf{T}')$. Moreover, the limit $Z = \mathbf{Y}'(\mathbf{T}\mathbf{\Sigma}\mathbf{T}')^{+}\mathbf{Y}$ is Chi-square distributed with $\operatorname{rank}(\mathbf{T}\mathbf{\Sigma}\mathbf{T}') = \operatorname{rank}(\mathbf{T}\mathbf{\Sigma}^{1/2}) = \operatorname{rank}(\mathbf{T})$ degrees of freedom (Rao and Mitra 1971, Theorem 9.2.2). We summarize this as

Theorem 1 *Under* $\mathbf{Tq} = \mathbf{0}_r$, $S_n(\mathbf{T})$ *converges in distribution to* $Z \sim \chi^2_{rank(\mathbf{T})}$.

Thus, comparing $S_n(\mathbf{T})$ with the $(1-\alpha)$-quantile of the limiting null distribution defines an asymptotic exact level α test $\varphi_n = 1\{S_n(\mathbf{T}) > \chi^2_{\operatorname{rank}(\mathbf{T}),1-\alpha}\}$. As Proposition 1 is not restricted to the null hypothesis, we can even deduce that $n^{-1}S_n(\mathbf{T})$ always converges in probability to $(\mathbf{Tq})'(\mathbf{T}\mathbf{\Sigma}\mathbf{T})^{+}\mathbf{Tq}$. Since $\mathbf{Tq} \neq \mathbf{0}_{km}$ implies $(\mathbf{Tq})'(\mathbf{T}\mathbf{\Sigma}\mathbf{T}')^{+}\mathbf{Tq} > 0$ (see Supplement for a verification) consistency follows.

Theorem 2 *Under* $\mathcal{H}_1 : \mathbf{Tq} \neq \mathbf{0}_r$, $S_n(\mathbf{T})$ *converges in probability to* ∞.

It remains to find appropriate estimators $\widehat{\mathbf{\Sigma}}^{(i)}$ for the unknown covariance matrices. For that purpose, we examine different strategies: "Brute force" via plug-in of a kernel density estimator into (6) or using a different approach that first estimates the diagonal elements $\mathbf{\Sigma}^{(i)}_{aa}$ and then employs their following relationship with the remaining matrix elements:

$$\mathbf{\Sigma}^{(i)}_{ab} = \sqrt{\mathbf{\Sigma}^{(i)}_{aa}\mathbf{\Sigma}^{(i)}_{bb}}\,\frac{p_a \wedge p_b - p_a p_b}{\sqrt{(p_a - p_a^2)(p_b - p_b^2)}} \qquad (a, b \in \{1, \ldots, m\}). \tag{8}$$

In the latter case, we consider two ways for estimating the variances $\Sigma^{(i)}_{aa}$: Via bootstrapping (Efron 1979) or with the interval estimator proposed in Price and Bonett (2001). In the following, we explain all three possibilities in detail.

3.1 Kernel estimator

A popular way to estimate densities is so-called kernel density estimators, which are based on a Lebesgue density $K : \mathbb{R} \to [0, \infty)$ with $\int K(x)\,\mathrm{d}x = 1$ and a bandwidth $h_n \to 0$. For more flexibility, we allow for different choices within the groups and add the corresponding group index, i.e., we work with K_i and h_{ni}. Then, the kernel density estimator for f_i is given by

$$\widehat{f}_{K,i}(x) = (n_i h_{ni})^{-1} \sum_{i=1}^{n_i} K_i\Big(\frac{x - X_{ij}}{h_{ni}}\Big) \quad (i = 1, \ldots, k). \tag{9}$$

Nadaraya (1965) proved strong uniform consistency of (9), i.e., we have $\sup_{x\in\mathbb{R}} |\widehat{f}_{K,i}(x) - f_i(x)| \to 0$ with probability one, under:

Assumption 2 Let K_i be of bounded variation and f_i be uniformly continuous. Furthermore, suppose that $\sum_{n=1}^{\infty} \exp(-\gamma n h_{ni}^2)$ converges for any choice of γ.

Here, the convergence of the series $\sum_{n=1}^{\infty} \exp(-\gamma n h_{ni})$ is, e.g., implied by choosing $h_{n,i} = n_i^{-\theta}$ for some $\theta \in (0, 1/2)$. We further note that Schuster (1969) discussed necessary and sufficient conditions for the stated uniform consistency. In particular, all f_i need to be uniformly continuous. Moreover, the conditions on the bandwidths can be weakened when the kernel fulfills additional regularity conditions (Silverman 1978). Anyhow, combining Proposition 1 and the strong consistency of (9) yields consistency of the plug-in covariance matrix estimators.

Lemma 1 *Under Assumption 2, we have*

$$\widehat{\mathbf{\Sigma}}_{ab}^{(i),K} \equiv \frac{n}{n_i} \frac{p_a \wedge p_b - p_a p_b}{\widehat{f}_{K,i}(\widehat{q}_{ia}) \widehat{f}_{K,i}(\widehat{q}_{ib})} \to \mathbf{\Sigma}_{ab}^{(i)} \quad \textit{in probability}. \tag{10}$$

3.2 Bootstrap estimator

In their one-way tests for equality of medians, Chung and Romano (2013) used the bootstrap approach of Efron (1979) to estimate the sample median's asymptotic variance. We adopt this idea for general quantiles. Therefore, for every group i, let $X_{i1}^*, \ldots, X_{in_i}^*$ denote a bootstrap sample (drawn with replacement) from the observations $\mathbf{X}_i = (X_{ij})_{j=1,\ldots,n_i}$. From this, we can calculate bootstrap versions of all previous estimators which we indicate by a superscript *. The mean squared error of the bootstrapped sample quantile $\widehat{q}_{ir}^*$ given the data can be explicitly calculated using (3)

$$\begin{aligned}\left(\widehat{\sigma}_i^*(p_r)\right)^2 &\equiv \mathbb{E}(n_i(\widehat{q}_{ir}^* - \widehat{q}_{ir})^2 \mid \mathbf{X}_i) = n_i \sum_{j=1}^{n_i} \left(X_{ij} - \widehat{q}_{ir}\right)^2 P\left(\widehat{q}_{ir}^* = X_{ij} \mid \mathbf{X}_i\right) \\ &= n_i \sum_{j=1}^{n_i} \left(X_{j:n_i}^{(i)} - \widehat{q}_{ir}\right)^2 P_{ij}; \quad P_{ij} = P\left(X_{\lceil mp_r \rceil:n_i}^{(i),*} = X_{j:n_i}^{(i)} \mid \mathbf{X}_i\right).\end{aligned}$$

Following Efron (1979), the probabilities P_{ij} can be rewritten to

$$P_{ij} = P(B_{n_i,(j-1)/n_i} \leq \lceil n_i p_r \rceil - 1) - P(B_{n_i,j/n_i} \leq \lceil n_i p_r \rceil - 1),$$

where $B_{n,p}$ denotes a binomial distributed random variable with size parameter n and success probability p. In contrast to the standard jackknife method, the bootstrap median variance estimator $(\widehat{\sigma}_i^*(1/2))^2$ converges to $1/(4 f_i^2(F_i^{-1}(1/2)))$ as desired (Efron 1979). Moreover, a detailed proof for strong consistency of this estimator was given by Ghosh et al. (1984) under

Assumption 3 Let $\max_{i=1,\ldots,k} \mathbb{E}(|X_{i1}|^{\delta}) < \infty$ for some $\delta > 0$.

Lemma 2 *Under Assumption 3, we have*

$$\widehat{\mathbf{\Sigma}}_{ab}^{(i),B} \equiv \frac{n}{n_i}\widehat{\sigma}_i^*(p_a)\widehat{\sigma}_i^*(p_b)\frac{p_a \wedge p_b - p_a p_b}{\sqrt{(p_a - p_a^2)(p_b - p_b^2)}} \xrightarrow{p} \mathbf{\Sigma}_{ab}^{(i)}.$$

3.3 Interval-based estimator

McKean and Schrader (1984) introduced an estimator for the sample median standard deviation based on a standardized confidence interval. Later, Price and Bonett (2001) suggested to modify this estimator to improve its performance in small sample size settings. Both estimators are consistent (Price and Bonett 2001) and can compete with the aforementioned bootstrap approach in simulations (McKean and Schrader 1984; Price and Bonett 2001) with a slightly better performance of the Price–Bonnet modification. While both papers only treat the median, extensions to general quantiles follow intuitively and have already been used, e.g., for the 25%- and 75%-quantile in Bonett (2006). The (extended) McKean–Schrader estimator for the standard deviation of the pth sample quantile, $p \in (0, 1)$, is given by

$$\widehat{\sigma}_i^{MS}(p) = n_i^{1/2}\frac{(X_{u_i(p):n_i}^{(i)} - X_{l_i(p):n_i}^{(i)})}{2z_{\alpha/2}},$$

where $\alpha \in (0, 1)$ and $l_i(p) = 1 \vee \lfloor n_i p - z_{\alpha/2}\sqrt{n_i}\sqrt{p(1-p)} \rfloor$ as well as $u_i(p) = n_i \wedge \lfloor n_i p + z_{\alpha/2}\sqrt{n_i}\sqrt{p(1-p)} \rfloor$ are the lower and upper limits of binomial intervals. Here, $z_{\alpha/2}$ denotes the $(1-\alpha/2)$-quantile of the standard normal distribution. Typically, $\alpha = 0.05$ is chosen leading to $z_{\alpha/2} \approx 1.96$. A brief discussion on the effect of the choice α on the estimator can be found in Price and Bonett (2001). In fact, the Price–Bonnet modification concerns the choice of α: They propose to replace it in the denominator by the following finite sample correction (for ease of notation we suppressed the dependency on i)

$$\alpha_n^*(p) = P\left(F_i^{-1}(p) \notin (X_{l_i(p):n_i}^{(i)}, X_{u_i(p):n_i}^{(i)})\right) = 1 - \sum_{i=l_i(p)+1}^{u_i(p)-1}\binom{n_i}{j}p^j(1-p)^{n_i-j}.$$

Clearly, $\alpha_n^*(p) \to \alpha$ by the central limit theorem. For large sample sizes, the benefit of the correction is negligible and may even lead to computational problems due to $\binom{n_i}{j} \gg 1$, especially for $j \approx n_i/2$. Thus, we only use the modifications for sample sizes smaller than 100 and recommend to set $\alpha_n^*(p) = \alpha$ for larger values ($n_i > 100$). Moreover, the simulations of Price and Bonett (2001) reveal that additionally adding $2n_i^{-1/2}$ to the denominator results in a slight reduction of bias and mean squared error. Altogether, we thus define their extended estimator for the respective standard

deviation as

$$\widehat{\sigma}_i^{\mathrm{PB}}(p) = n_i^{1/2} \frac{(X^{(i)}_{u_i(p):n_i} - X^{(i)}_{l_i(p):n_i})}{2z_{\alpha_n^*(p)/2} + 2n_i^{-1/2}}.$$

As explained above, this estimator is consistent for the variance, leading to:

Lemma 3 *We have for all $i = 1, \ldots, k$ and $a, b = 1, \ldots, m$:*

$$\widehat{\mathbf{\Sigma}}^{(i),PB}_{ab} = \frac{n}{n_i} \widehat{\sigma}_i^{PB}(p_a) \widehat{\sigma}_i^{PB}(p_b) \frac{p_a \wedge p_b - p_a p_b}{\sqrt{(p_a - p_a^2)(p_b - p_b^2)}} \xrightarrow{p} \mathbf{\Sigma}^{(i)}_{ab}. \tag{11}$$

Utilizing the different choices of covariance estimators results in three different versions of the asymptotic test φ_n. However, simulation results (Sect. 5) exhibit serious issues for small to moderate sample sizes which may be due to a rather poor χ^2-approximation to the test statistic. To tackle this problem, we propose the initially mentioned technique of permuting studentized statistics.

4 Permutation test

For a better finite sample performance, it is often advisable to replace the asymptotic critical value of the test, here the $(1-\alpha)$-quantile of the $\chi^2_{\mathrm{rank}(\mathbf{T})}$-distribution, by a resampling-based critical value. For the current problem, we promote the permutation approach, which leads to a finitely exact test under exchangeability, i.e., under $\widetilde{\mathcal{H}}_0 : F_1 = \ldots = F_k$. Moreover, the proper studentization within the Wald-type statistic makes it possible to transfer the consistency and asymptotic exactness (under $\mathcal{H}_0 : \mathbf{Tq} = 0$) of the tests φ_n to their permutation versions. To explain this, let $\mathbf{X}^\pi = (X^\pi_{ij})_{i=1,\ldots,k;\, j=1,\ldots,n_i}$ be a random permutation of the pooled data $\mathbf{X} = (X_{ij})_{i=1,\ldots,k;\, j=1,\ldots,n_i}$. As for Efron's bootstrap, we draw new samples from the pooled data, but now without replacement. In other words, we randomly permute the group memberships of the observations X_{ij}. Pooling the data affects our Assumptions 1 and 2 such that we need to replace the original distribution functions F_i and their densities f_i by their pooled versions $F = \sum_{i=1}^k \kappa_i F_i$ and $f = \sum_{i=1}^k \kappa_i f_i$ describing the (unconditional) distribution of X^π_{ij}. To be concrete, we postulate

Assumption 4 Let F be differentiable with uniformly continuous derivative f such that $f(F^{-1}(p_r)) > 0$ for all r, and K_i be a kernel fulling Assumption 2.

As in Chung and Romano (2013), it turned out that the asymptotic correctness of the permutation approach needs a certain convergence rate in (4):

$$\frac{n_i}{n} - \kappa_i = O(n^{-1/2}). \tag{12}$$

Theorem 3 *Under $\mathcal{H}_0 : \mathbf{Tq} = \mathbf{0}_{km}$ as well as under $\mathcal{H}_1 : \mathbf{Tq} \neq \mathbf{0}_{km}$, the permutation version $S_n^\pi(\mathbf{T})$ of $S_n(\mathbf{T})$ with any of the covariance estimators (10)–(11) always mimics its null distribution asymptotically, i.e.,*

$$\sup_{x\in\mathbb{R}}\left|P\left(S_n^\pi(\mathbf{T}) \leq x|\mathbf{X}\right) - \chi^2_{rank(\mathbf{T})}((-\infty, x])\right| \xrightarrow{p} 0. \tag{13}$$

Replacing the critical value $\chi^2_{\text{rank}(\mathbf{T}),1-\alpha}$ of the asymptotical tests with $c_n^\pi(\alpha)$, the $(1-\alpha)$-quantile of the conditional distribution function $x \mapsto P(S_n^\pi(\mathbf{T}) \leq x|\mathbf{X})$, leads to three different permutation tests $\varphi_n^\pi = 1\{S_n(\mathbf{T}) > c_n^\pi(\alpha)\}$. Under the assumptions given in Theorem 3, it follows that $c_n^\pi(\alpha)$ converges in probability to $\chi^2_{\text{rank}(\mathbf{T}),1-\alpha}$ irrespective whether the null hypothesis is true or not. Thus, we can deduce the asymptotic exactness of the permutation test and its consistency for general fixed alternatives (Janssen and Pauls 2003, Lemma 1 and Theorem 7). In addition, we prove in the next section that the permutation test has an asymptotic relative efficiency of 1 compared to the asymptotic test, i.e., the tests' asymptotic power values coincide for local alternatives.

Local alternatives To study local alternatives, we need to replace Model (1) with its local counterpart given by a triangular array of row-wise independent random variables $X_{nij} \sim F_{ni}$ $(i = 1, \ldots, k; j = 1, \ldots, n_i)$ with absolutely continuous distribution functions F_{ni}, corresponding densities f_{ni}, quantiles q_{nir} and quantile vector $\mathbf{q}_n = (q_{n11}, \ldots, q_{n1m}, q_{n21}, \ldots, q_{nkm})'$. Within this framework, we discuss local alternatives $\mathbf{Tq}_n = O(n^{-1/2})$, i.e., small perturbations of the null hypotheses, under the following additional regularity conditions:

Assumption 5 For every $i = 1, \ldots, k$, let F_i be an absolutely continuous distribution function with corresponding density f_i. Moreover, set $F = \sum_{i=1}^k \kappa_i F_i$.

(i) For some $M > 0$ let $\sqrt{n}|F_{ni}(x) - F_i(x)| \leq M$ for all $n \in \mathbb{N}$ and all $x \in \mathbb{R}$.
(ii) Suppose that f_i is continuous and positive at $q_{ir} = F_i^{-1}(p_r)$ and that f_{ni} converges uniformly to f_i in a compact neighborhood around q_{ir} for all r.
(iii) For the permutation approach, suppose additionally (12), Assumption 4 and uniform convergence of f_{ni} to f_i in a compact neighborhood around $q_r = F^{-1}(p_r)$ for every r.

While (ii) and (iii) are local versions of the conditions for Model (1), condition (i) ensures the usual $\sqrt{n}$-convergence of F_{ni} to F_i. Anyhow, the tests' asymptotic power functions can be described by means of a non-central χ^2 distribution:

Theorem 4 *Under $\sqrt{n}\mathbf{Tq}_n \to \boldsymbol{\theta} \neq \mathbf{0}_{km}$, the asymptotic test φ_n and its permutation variant φ_n^π with any of the covariance estimators (10)–(11) have the same asymptotic power $P(Z > \chi^2_{rank(\mathbf{T}),1-\alpha}) > \alpha$, where Z is $\chi^2_{rank(\mathbf{T})}(\delta)$-distributed with non-centrality parameter $\delta = \boldsymbol{\theta}'(\mathbf{T}\boldsymbol{\Sigma}\mathbf{T}')^+\boldsymbol{\theta} > 0$.*

5 Simulations

To assess the tests' small sample performance, we complement our theoretical findings with numerical comparisons. For ease of presentation, we consider

1. **A one-way layout** in which we like to infer the null hypothesis $\mathcal{H}_0 : \{IQR_1 = \cdots = IQR_4\}$ of equal IQRs, i.e., as described at the end of Sect. 2 we choose probabilities $p_1 = 0.25$ and $p_2 = 0.75$ and specify the contrast matrix as $H = P_4 \otimes (-1, 1)$.
2. **A** 2×2 **layout** in which we test for the presence of main or interaction effects measured in terms of medians, i.e., setting $k = a \cdot b = 2 \cdot 2$ we infer the hypotheses $\mathcal{H}_0 : \{\mathbf{H}_A \mathbf{q} = \mathbf{0}_{ab}\}$ (no main median effect of factor A) and $\mathcal{H}_0 : \{\mathbf{H}_{AB} \mathbf{q} = \mathbf{0}_{ab}\}$ (no median $A \times B$ interaction effect), see Sect. 2.

In addition, we present detailed simulations for a **five-factor model** in supplement. Data were simulated within Model (1) via $X_{ij} = \mu_i + \sigma_i(\epsilon_{ij} - m_i) \sim F_i$, where we consider (a) balanced and unbalanced settings given by sample size vectors $\mathbf{n_1} = (15, 15, 15, 15)$ and $\mathbf{n_2} = (10, 10, 20, 20)$, respectively. (b) five different distributions for ϵ_{ij}: the standard normal distribution ($N_{0,1}$), Student's t-distribution with $df = 2, 3$ degrees of freedom (t_2, t_3), the Chi-square distribution with $df = 3$ (χ^2_3) and the standard log-normal distribution ($LN_{0,1}$). All distributions were centered by subtracting the respective median m_i from ϵ_{ij}. (c) a homoscedastic setting $\boldsymbol{\sigma_1} = (\sigma_1, \ldots, \sigma_4) = (1, 1, 1, 1)$ and heteroscedastic designs $\boldsymbol{\sigma_2} = (1, 1.25, 1.5, 1.75)$ and $\boldsymbol{\sigma_3} = (1.75, 1.5, 1.25, 1)$. Together with $\mathbf{n}_2$, the latter represent a positive and negative pairing, respectively.

The simulations were conducted by means of the computing environment R (R Core Team 2020), version 3.5.0, generating $N_{\text{sim}} = 5000$ simulation runs and $N_{\text{perm}} = 1999$ permutation iterations for each setting. The nominal level was set to $\alpha = 5\%$. We compare the type-1 error rate as well as the power values of our tests below. In both cases, we include all three variance estimation strategies introduced in Sects. 3.1–3.2. For the kernel density estimation, we choose the classical Gaussian kernel with a bandwidth according to Silverman's rule-of-thumb (Silverman 1986, Eq. (3.31)), where we applied the function *bw.nrd0* from the R package *stats* to determine the latter.

In case of the 2×2-median design, these methods are additionally compared with the current state-of-the-art tests for regression parameters in quantile regression: From the R package *quantreg* (Koenker et al. 2019), we choose the rank inversion method by Koenker and Machado (1999) for non-iid errors, the default choice in *quantreg*, and the wild bootstrap approach of Feng et al. (2011). For a fair comparison, we include the main factors A and B and their interaction in the respective regression model. Hence, regression parameters β_A, β_B and β_{AB} are estimated, and corresponding p-values for testing $\mathcal{H}_0 : \beta_A = 0$ (no main effect A) and $\mathcal{H}_0 : \beta_{AB} = 0$ (no interaction effect) are derived by both *quantreg* approaches.

Type-1 error In this subsection, we discuss the type-1 error control of all procedures. To simulate under the corresponding null hypotheses, we set $\mu_i = \mu_{i_1 i_2} = 0$ in the 2×2-median-based cases and restrict to the homoscedastic setting $\boldsymbol{\sigma} = \boldsymbol{\sigma}_1$ for the 4-sample IQR testing question. The standard error of the estimated sizes in case of $N = 5000$ simulation runs is 0.3% if the true type-1 error probability is 5%, i.e., estimated sizes outside the interval [4.4%,5.6%] deviate significantly from the nominal 5% significance level.

Table 1 Type-1 error rate in % (nominal level $\alpha = 5\%$) for testing the median null hypothesis of no main effect in the 2×2 design for the rank-based (Rank) and wild bootstrap (Wild) quantile regression approach as well as all asymptotic and permutation tests using the interval-based (Int), kernel density (Kern) and bootstrap (Boot) approach for estimating the covariance matrix

Distr	**n**	σ	Asymptotic			Permutation			Quantile reg.	
			Int	Kern	Boot	Int	Kern	Boot	Rank	Wild
$N_{0,1}$	$\mathbf{n_1}$	$\boldsymbol{\sigma_1}$	2.6	4.2	3.3	**4.9**	**5.1**	**5.2**	6.9	6.6
		$\boldsymbol{\sigma_2}$	3.0	**4.4**	3.3	**5.5**	5.8	**5.5**	5.7	7.3
	$\mathbf{n_2}$	$\boldsymbol{\sigma_1}$	2.2	**4.8**	3.6	**5.0**	**5.6**	**5.6**	2.5	4.1
		$\boldsymbol{\sigma_2}$	2.0	4.0	3.0	5.7	**4.7**	**4.7**	3.5	3.6
		$\boldsymbol{\sigma_3}$	2.5	**5.4**	4.0	6.2	6.4	6.3	3.0	**5.2**
t_3	$\mathbf{n_1}$	$\boldsymbol{\sigma_1}$	1.7	2.7	2.3	**5.1**	**5.2**	**5.2**	6.4	**5.5**
		$\boldsymbol{\sigma_2}$	2.0	2.9	2.6	**5.5**	**5.2**	**4.9**	**5.6**	6.0
	$\mathbf{n_2}$	$\boldsymbol{\sigma_1}$	0.8	3.0	2.1	**4.5**	**4.5**	3.5	3.2	**4.4**
		$\boldsymbol{\sigma_2}$	1.1	3.1	2.4	6.4	**4.7**	**5.2**	3.1	2.8
		$\boldsymbol{\sigma_3}$	0.7	3.7	2.7	**5.8**	6.5	6.4	3.1	3.9
$LN_{0,1}$	$\mathbf{n_1}$	$\boldsymbol{\sigma_1}$	**4.9**	4.0	2.0	**5.4**	5.8	5.7	**4.6**	6.5
		$\boldsymbol{\sigma_2}$	**5.2**	3.8	1.8	5.8	5.8	6.0	**4.6**	7.6
	$\mathbf{n_2}$	$\boldsymbol{\sigma_1}$	3.1	3.0	1.7	**4.8**	**4.7**	**4.8**	2.8	3.6
		$\boldsymbol{\sigma_2}$	3.4	3.4	2.1	5.9	**5.3**	**5.4**	3.0	3.7
		$\boldsymbol{\sigma_3}$	3.8	**4.2**	2.5	6.6	6.8	6.3	3.0	**5.4**
χ^2_3	$\mathbf{n_1}$	$\boldsymbol{\sigma_1}$	**5.1**	**5.0**	3.2	**5.5**	**5.6**	5.8	5.7	7.7
		$\boldsymbol{\sigma_2}$	**4.4**	**4.7**	2.7	**5.1**	**5.5**	**5.2**	**5.4**	7.7
	$\mathbf{n_2}$	$\boldsymbol{\sigma_1}$	3.6	**4.5**	2.8	**5.0**	**5.1**	**5.2**	**3.0**	**4.8**
		$\boldsymbol{\sigma_2}$	3.3	3.7	2.6	**5.1**	**4.6**	**4.8**	3.3	3.2
		$\boldsymbol{\sigma_3}$	**4.6**	5.7	3.5	7.2	7.2	6.4	3.2	**5.6**

Values inside the 95% binomial interval [4.4, 5.6] are printed bold

The observed type-1 error rates for the 2×2-median design are displayed in Table 1 for testing the hypothesis of no main effect. It is readily seen that all asymptotic tests are rather conservative with type-1 errors reaching down to 1.7% for the bootstrap-based and 0.7% for the interval-based approaches, respectively. This conservativeness is less pronounced for the test based upon the kernel density variances estimator that exhibits values between 2.7% and 5.7% and a reasonable good error control in case of the standard normal and χ^2_3 distribution except for the settings with positive variance pairing. In contrast, all permutation methods control the type-1 error level reasonably well except for the situations with a skewed distribution and negative pairing. Here, we find error rates up to 7.2% for the tests based upon the interval- and kernel-based variance estimators. For the two quantile regression methods from the R package *quantreg* (Koenker et al. 2019), the observations are diverse: The rank-based approach tends to conservative test decisions in case of unbalanced sample sizes with observed error rates in the range 2.5−3.5%. However, in case a balanced homoscedastic design with symmetric errors, a slight liberality (6.4−6.9%) is detected. For all other settings,

Table 2 Type-1 error rate in % (nominal level $\alpha = 5\%$) for the four-sample IQR testing problem of our asymptotic and permutation tests using the interval-based (Int), kernel density (Ker) and bootstrap (Boo) approach for estimating the covariance matrix

Distr	$\mathbf{n}_1$ (balanced)						$\mathbf{n}_2$ (unbalanced)					
	Asymptotic			Permutation			Asymptotic			Permutation		
	Int	Ker	Boo	Int	Ker	Boo	Int	Ker	Boo	Int	Ker	Boo
$N_{0,1}$	1.3	**4.6**	1.2	**5.2**	**5.1**	**5.2**	1.0	**5.0**	1.1	**4.7**	**4.8**	**4.9**
t_2	0.6	3.7	0.3	**5.1**	**5.1**	**5.1**	0.4	**4.8**	0.7	**5.0**	**5.0**	**5.2**
t_3	0.9	3.7	0.6	**4.9**	**5.0**	**5.3**	0.5	**4.4**	0.9	4.5	**4.6**	**5.1**
$LN_{0,1}$	7.5	8.5	1.6	**5.2**	**4.9**	**4.6**	4.3	10.2	1.6	**4.8**	**5.2**	**4.7**
χ^2_3	**5.1**	6.2	1.8	**4.5**	**4.8**	**4.7**	3.8	8.1	1.7	**5.1**	**5.0**	**4.8**

Values inside the 95% binomial interval [4.4, 5.6] are printed in bold

the decisions are accurate. In comparison, the wild bootstrap strategy is liberal for almost all balanced settings (with observed error rates up to 7.7%) and conservative for all positive pairings (2.8−3.7%). Overall, the permutation procedure that uses a bootstrap variance estimator exhibits the most robust type-1 error control with values ranging from 4.7−6.4%.

Summarizing the results for the interaction tests presented in supplement, we get a similar impression for the wild bootstrap quantile regression strategy and the six Wald-type procedures. For them, the only major difference is that the permutation methods also exhibit a fairly well error control for the settings with skewed distributions and negative pairing. However, the results for the rank-based quantile regression method are partially different: While the type-1 error rate is still accurate for balanced sample sizes, the decisions become very liberal in the unbalanced scenarios with estimated type-1 error rates between 6.1% and 10.1%.

The type-1 error rates in the situation of the four-sample testing problem of equal IQRs are presented in Table 2. Here, the finite sample behavior of the asymptotic tests becomes even more extreme: For the symmetric distributions, the type-1 error rates are between 0.4% and 1.3% for the interval-based estimator and between 0.3 and 1.2% for the bootstrap approach, i.e., very conservative. In contrast, the decisions for the kernel-based method are quite accurate with values between 3.7% and 5.0%. Switching to skewed distribution, however, the type-1 error rates increase, leading to very liberal decisions in the log-normal case with values up to 10.2% for the kernel-based and 7.5% for the interval-based tests. Here, only the bootstrap-based method remained very conservative. In comparison, all permutation counterparts lead to satisfactory type-1 error control close to the 5% level. Due to the extreme behavior of the asymptotic tests in this setting, we conducted additional simulation results in supplement. Therein, all asymptotic tests for equality of IQRs more or less approach the 5% level for larger group-specific sample sizes $n_i \geq 150$.

Power behavior Due to the diverse behavior of the asymptotic tests and the rank-based quantile regression method under the null hypotheses and for ease of presentation, we solely focus on permutation tests and the wild bootstrap quantile regression strategy here. The results for the asymptotic tests are presented in supple-

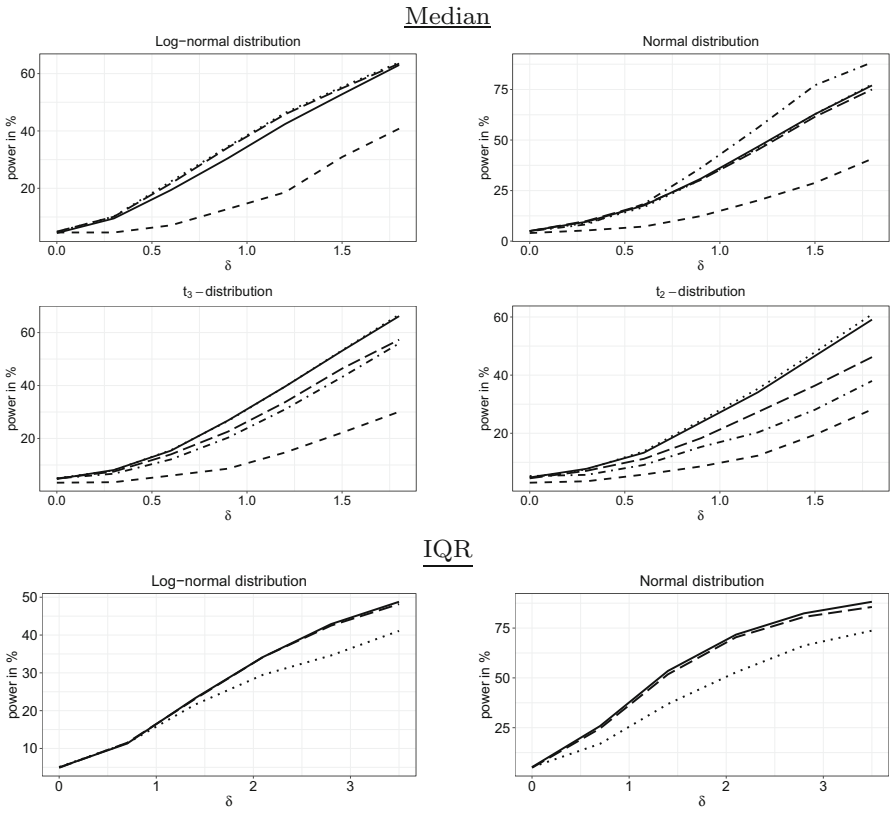

Fig. 1 Power curves for the 2×2-median testing problem (first two rows) and for the four-sample IQR testing problem (last row) of the permutation PBK test (dash-dotted), the wild bootstrap quantile regression test (dashed) and the three permutation tests based on interval-based (long-dashed), kernel density (dotted) and bootstrap (solid) covariance matrix estimation, respectively, for $\mathbf{n} = \mathbf{n_2}$, $\boldsymbol{\sigma} = \boldsymbol{\sigma_1}$ and shift alternatives $\boldsymbol{\mu} = (0, 0, 0, \delta)$ (median) or scale alternatives $\boldsymbol{\sigma} = (1, 1, 1, 1 + \delta)$ (IQR)

ment, and apart from their different levels under $\mathcal{H}_0$, their power curves run almost parallel to the ones of the permutation version.

To achieve a scenario under the alternative in the 2×2-median test setting, we disturbed the respective null setup by adding a shift parameter $\delta = \mu_{2,2}$ to the last group. In addition to the aforementioned methods, we considered the permutation Wald-type test (PBK) of Pauly et al. (2015) which was developed for testing means in general factorial designs. Their procedure is implemented in the R package *GFD* (Friedrich et al. 2017b). For a fair comparison, we included their PBK test just for the cases where mean and median coincide, i.e., for the symmetric distributions. The results for the procedures inferring a main effect are presented in Fig. 1, while the corresponding power curves of the interaction tests are shown in supplement. Studying Fig. 1, we observe that the PBK test leads to higher power values compared to our tests for the normal distribution settings but is less powerful under the t_2- and t_3-distributions. An explanation may be given by the (asymptotic) efficiencies of the

location estimators: While the sample mean is more efficient than the sample median under normal distributions, the situation is reversed for the two more heavy-tailed t-distributions. A comparison among the median-based permutation tests shows that the interval-based approach leads to lower power values than the other two methods for both t-distributions, while the bootstrap approach is slightly less powerful than the other two for the skewed log-normal distribution. Under normality, however, the tests' power functions are almost identical. In comparison, the wild bootstrap quantile regression method has considerably less power than all other methods for testing main median effects. The power curves for the interaction effects presented in supplement show a similar pattern for almost all tests. The only exception is the wild bootstrap approach which exhibits a similar power behavior as the permutation tests. Moreover, it is slightly advantageous for shift alternatives with $\delta > 1$.

To obtain alternatives for the four-sample IQR testing problem, we consider scale alternatives $\boldsymbol{\sigma} = (1, 1, 1, 1+\delta)$. For ease of presentation, we only show the results for normal as well as lognormal distributions here. The resulting power curves are plotted in Fig. 1. We can observe that the kernel density approach leads to lower power values compared to the other two methods.

Recommendation Summarizing the findings, we recommend the use of the permutation methods over their asymptotic counterparts as they show a much better type-1 error control in case of small and moderate sample sizes ($n_i \leq 200$). However, there is no general recommendation for choosing between the three permutation versions as their power behavior (slightly) differed with respect to underlying settings, e.g., for comparing IQRs the interval- and bootstrap-based approaches performed better, while the kernel method exhibits the largest power for testing medians in a 2×2 design with heavy tails. In comparison with quantile regression, the advantage of the proposed factorial design approach is the simple incorporation of interaction effects without a loss in power. This benefit can be seen in the power simulations for the 2×2-median design, as shown in Fig. 1, and becomes even more pronounced in higher-way layouts, see Green et al. (2002) and Green (2012) and the additional simulation results in supplement.

6 Illustrative data analysis

A typical everyday situation in which we are confronted with quantiles is *percentile curves* for child heights and weights. We re-analyzed growth and weight data of children from five sites (Brazil, India, Guatemala, the Philippines, South Africa), which was provided to us by the COHORTS group (Richter et al. 2012). Both, height and weight, were converted to z-scores regarding the WHO child standards (de Onis et al. 2007). Having a comparison of percentile curves in mind, we test for effects in the 25%-, 50%- and 75%-quantile simultaneously. This also demonstrates the flexibility of the proposed methodology. For illustrative purposes, we focus on the following subgroups:

Example 1 We compare the birth weight of firstborns from the countries (factor A) Brazil and South Africa including both genders (factor B). To avoid confounding

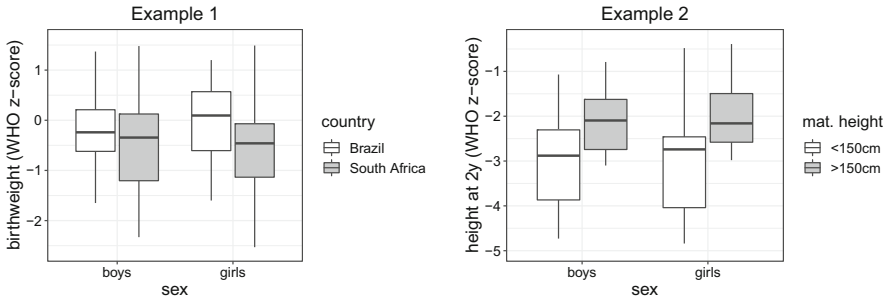

Fig. 2 Group-wise boxplots (outliers are not displayed) for the birth weight data from Example 1 (left) and the height data from Example 2 (right)

Table 3 For the effect of the country on the birth weight (Example 1) and the maternal height on the height at 2 years, the p values (in %) are shown for our asymptotic and permutation approach using the interval-based (Int), kernel density (Ker) and bootstrap (Boo) strategies for covariance matrix estimation

	Asymptotic			Permutation		
	Int	Ker	Boo	Int	Ker	Boo
Example 1	10.54	8.63	9.83	3.80	4.60	3.50
Example 2	10.95	9.19	8.43	3.30	6.75	3.60

effects regarding age, education or marital status, we restrict our analysis to 30-year-old or younger married mothers with a comparable education level of 9 completed school years. The $n = 173$ children are divided into $n_1 = 65$ boys and $n_2 = 46$ girls from Brazil and $n_3 = 36$ boys and $n_4 = 26$ girls from South Africa. We would like to infer whether there are differences between the countries regarding the boys' and girls' birth weight, respectively.

Example 2 We investigate the effect of the mother's height (factor A) on the children's height at the age of 2 years. Both sexes (factor B) are included. We restrict to firstborns of unmarried mothers from the Philippines. For this analysis, we divide the women into the groups "small" and "tall" consisting of the women, respectively, being smaller and taller than the median height of 150cm. The group "small" consists of data for $n_1 = 8$ boys and $n_2 = 13$ girls, and in the group "tall," there are data for $n_3 = 12$ boys and $n_4 = 11$ girls.

To get a first graphical impression, the group-specific box plots are presented in Fig. 2. In both cases, it appears that factor A (country and maternal height, respectively) leads to a shift of all three empirical quantiles of the children's height and weight. To infer this conjecture, we like to check for a main effect of factor A regarding the three quantiles $\mathbf{q}_i = (q_{i1}, q_{i2}, q_{i3})^\top$, $i = 1, \ldots, 4$ corresponding to the probabilities $(p_1, p_2, p_3) = (0.25, 0.5, 0.75)$ simultaneously. That is, we test $\mathcal{H}_0 : \{\mathbf{q}_1 + \mathbf{q}_2 = \mathbf{q}_3 + \mathbf{q}_4\}$. The p values of all three asymptotic and permutation tests (ignoring multiplicity) are summarized in Table 3.

It is apparent that the asymptotic and permutation test leads to different decisions at the nominal level $\alpha = 5\%$. In fact, the seemingly present effect from Fig. 2 is

Table 4 Point estimates $\widehat{\theta}$ for the difference $\theta_{i_2} = q_{2i_2} - q_{1i_2}$ of the countries' median with respect to sex for Example 1 together with permutation-based 95% confidence intervals

Gender	$\widehat{\theta}$	Int	Ker	Boo
Boys	−0.15	[−0.44, 0.14]	[−0.48, 0.18]	[−0.43, 0.13]
Girls	−0.56	[−1.04, −0.08]	[−1.10, −0.02]	[−1.02, −0.10]

Here, Int (interval-based), Ker (kernel density) and Boo (bootstrap) indicate the applied covariance matrix estimation technique

not detected by any asymptotic tests with p values around 8–10%. In contrast, the p values of the permutation approaches are, except for the kernel density method in Example 2, less than 5%. To investigate the reasons why these decisions are different, we run additional simulations for the three-quantile testing problem under the sample size settings of Example 2. The results are presented in supplement and may explain the above decisions to some extent: As in Sect. 5, the asymptotic tests are quite conservative with type-1 error rates ranging between 0.8% and 4.2%. Moreover, the permutation kernel density approach is less powerful than the other two permutation methods under shift alternatives for skewed distributions.

Beyond hypothesis testing, the theoretical results can also be used to formulate asymptotically valid confidence regions for contrasts of quantiles by inverting the corresponding tests. We exemplify this for the difference between two quantiles as effect parameter of interest. To this end, consider Example 1 and encode factor A (country) and factor B (gender) as follows: $i_2 = 1$ for the boys, $i_2 = 2$ for the girls, $i_1 = 1$ for Brazil and $i_1 = 2$ for South Africa. Then, for a fixed gender i_2, the asymptotic correct z- and permutation-$(1-\alpha)$-confidence intervals for the difference $\theta_{i_2} = q_{2i_2} - q_{1i_2}$ of the countries' quantiles are

$$\left[(\widehat{q}_{2i_2} - \widehat{q}_{1i_2}) \pm \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{\widehat{\sigma}^2_{1i_2} + \widehat{\sigma}^2_{2i_2}}\right], \quad \left[(\widehat{q}_{2i_2} - \widehat{q}_{1i_2}) \pm \frac{c^{\pi}_{ni_2}(\alpha/2)}{\sqrt{n}}\sqrt{\widehat{\sigma}^2_{1i_2} + \widehat{\sigma}^2_{2i_2}}\right],$$

respectively. Here, $\widehat{\sigma}^2_{i_1 i_2} = \widehat{\mathbf{\Sigma}}^{(i_1 i_2)}_{11}$ is an estimator for the asymptotic variance of $\sqrt{n}(\widehat{q}_{i_1 i_2} - q_{i_1 i_2})$ using one of our strategies from Sects. 3.1–3.3 and $c^{\pi}_{ni_2}(\alpha/2)$ is the $(1-\alpha/2)$-quantile of the permutation distribution of $\sqrt{n}(\widehat{q}^{\pi}_{2i_2} - \widehat{q}^{\pi}_{1i_2})(\widehat{\sigma}^{2,\pi}_{1i_2} + \widehat{\sigma}^{2,\pi}_{i_2})^{-1/2}$. To illustrate the application, we calculated the 95% permutation-based confidence intervals for the median difference separately for gender in Table 4. Ignoring multiplicity, we see that all three permutation procedures agree on a significant difference in the girl's median birth weight (at level $\alpha = 5\%$) but do not find a corresponding effect for the boys.

7 Discussion

While an abundance of methods exists for inferring means and mean vectors in general heterogeneous factorial designs (Johansen 1980; Brunner et al. 1997; Bathke et al.

2009; Zhang 2012; Konietschke et al. 2015; Pauly et al. 2015; Harrar et al. 2019), there are not so many methods for the analysis of medians or quantiles. To this end, we combined the idea of studentized permutations from heteroscedastic mean-based (Pauly et al. 2015) and one-way median-based ANOVA (Chung and Romano 2013) to establish flexible methods for inferring quantiles in general factorial designs which we coin QANOVA. In fact, we proposed three permutation methods in Wald-type statistics that only differ in the way the covariance matrix is estimated. All of them are applicable to construct confidence regions and to test null hypotheses about arbitrary contrasts of different quantiles.

The resulting procedures are finitely exact under exchangeability of the data and shown to be asymptotically valid. In doing so, we had to extend some results about permutation empirical processes and uniform Hadamard differentiability (van der Vaart and Wellner 1996) that are of own mathematical interest. From them, we could deduce the asymptotic validity as well as results about the procedures' asymptotics under fixed and local alternatives. In the special case of the median, these results even reveal new insights into one-way permutation test of Chung and Romano (2013).

In addition to the theoretical findings, we analyzed the procedures in extensive simulations. Our results indicate an accurate type-1 error control for the permutation methods in almost all settings. Only in case of skewed distributions and small unbalanced samples with a heteroscedastic negative pairing, a slight liberality was found when testing for main effects. Beyond this, we can recommend all three permutation methods with clear conscience. Currently, we work on implementing them within an R-package. We are confident that the current results can be transferred to questions about related quantile-based estimands, e.g., coefficients of quartile variation (Bonett 2006) as well as to complex ANOVA settings with correlated variables, e.g., quantile-based repeated measurements or complex MANOVA designs.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11749-021-00758-y.

Acknowledgements The authors are grateful to the editor, the associate editor and the two referees for their comments that substantially improved the paper's quality. Moreover, the authors thank the COHORT investigators (Richter et al. 2012) for providing their data, which were collected in five different studies (Victora and Barros 2006; Adiar 2007; Richter et al. 2007; Stein et al. 2008; Bhargava et al. 2009). Here, the authors are especially grateful to Linda Richter for helping with the communication between all sites.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Adiar L (2007) Size at birth and growth trajectories to young adulthood. Am J Hum Biol 19:327–337

Bathke A, Schabenberger O, Tobias R, Madden L (2009) Greenhouse-Geisser adjustment and the ANOVA-type statistic: cousins or twins? Am Stat 63:239–246

Beyerlein A (2014) Quantile regression-opportunities and challenges from a user's perspective. Am J Epidemiol 180(3):330–331

Bhargava S, Sachdev H, Fall C, Osmond C, Lakshmy R, Barker D, Biswas S, Ramji S, Prabhakaran D, Reddy K (2009) Relation of serial changes in childhood body-mass index to impaired glucose tolerance in young adulthood. N Engl J Med 250:865–875

Bonett D (2006) Confidence interval for a coefficient of quartile variation. Comput Stat Data Anal 50:2953–2957

Brunner E, Dette H, Munk A (1997) Box-type approximations in nonparametric factorial designs. J Am Stat Assoc 92:1494–1502

Cassidy J, Clarke S, Díaz-Rubio E, Scheithauer W, Figer A, Wong R, Koski S, Lichinitser M, Yang TS, Rivera F (2008) Randomized phase III study of capecitabine plus oxaliplatin compared with fluorouracil acid plus oxaliplatin as first-line therapy for metastatic colorectal cancer. J Clin Oncol 26:2006–2012

Chung E, Romano J (2013) Exact and asymptotically robust permutation tests. Ann Stat 41:484–507

de Onis A, Onyango A, Borghi E, Siyam A, Nishida C, Siekmann J (2007) Development of a WHO growth reference for school-aged children and adolescents. Bull World Health Organ 85:660–667

Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7:1–26

Feng X, He X, Hu J (2011) Wild bootstrap for quantile regression. Biometrika 98(4):995–999

Fried R, Dehling H (2011) Robust nonparametric tests for the two-sample location problem. Stat Methods Appl 20:409–422

Friedrich S, Brunner E, Pauly M (2017a) Permuting longitudinal data in spite of the dependencies. J Multivariate Anal 153:255–265

Friedrich S, Konietschke F, Pauly M (2017b) GFD: an R package for the analysis of general factorial designs. J Stat Softw 79:1–18

Fung K (1980) Small sample behaviour of some nonparametric multi-sample location tests in the presence of dispersion differences. Stat Neerl 34:189–196

Ghosh M, Parr W, Singh K, Babu GJ (1984) A note on bootstrapping the sample median. Ann Stat 12:1130–1135

Green S (2012) Factorial designs with time to event endpoints, 3rd edn. Handbook of statistics in clinical oncology. Chapman and Hall/CRC, Boca Raton, pp 201–210

Green S, Liu PY, O'Sullivan J (2002) Factorial design considerations. J Clin Oncol 20(16):3424–3430

Harrar S, Ronchi F, Salmaso L (2019) A comparison of recent nonparametric methods for testing effects in two-by-two factorial designs. J Appl Stat 46:1649–1670

Hettmansperger T, McKean J (2010) Robust nonparametric statistical methods, 2nd edn. CRC Press, Boca Raton

Janssen A (1997) Studentized permutation tests for non-iid hypotheses and the generalized Behrens–Fisher problem. Stat Probab Lett 36:9–21

Janssen A, Pauls T (2003) How do bootstrap and permutation tests work? Ann Stat 31:768–806

Johansen S (1980) The Welch–James approximation to the distribution of the residual sum of squares in a weighted linear regression. Biometrika 67:85–92

Koenker R, Hallock K (2001) Quantile regression. J Econ Perspect 15(4):143–156

Koenker R, Machado JA (1999) Goodness of fit and related inference processes for quantile regression. J Am Stat Assoc 94(448):1296–1310

Koenker R, Portnoy S, Ng PT, Zeileis A, Grosjean P, Ripley BD (2019) Package 'quantreg': quantile regression in R

Konietschke F, Bathke A, Harrar S, Pauly M (2015) Parametric and nonparametric bootstrap methods for general MANOVA. J Multivariate Anal 140:291–301

Kurz A, Fleischmann E, Sessler D, Buggy D, Apfel C, Akça O, Investigators FT, Fleischmann E, Erdik E, Eredics K (2015) Effects of supplemental oxygen and dexamethasone on surgical site infection: a factorial randomized trial. Br J Anaesth 115:434–443

Lubsen J, Pocock S (1994) Factorial trials in cardiology: pros and cons. Eur Heart J 15:585–588

McKean J, Schrader R (1984) A comparison of methods for studentizing the sample mean. Commun Stat B 13:751–773

Mehta S, Tanguay JF, Eikelboom J, Jolly S, Joyner C, Granger C, Faxon D, Rupprecht HJ, Budaj A, Avezum A (2010) Double-dose versus standard-dose clopidogrel and high-dose versus low-dose aspirin in individuals undergoing percutaneous coronary intervention for acute coronary syndromes (CURRENT-OASIS 7): a randomised factorial trial. Lancet 376:1233–1243

Nadaraya E (1965) On non-parametric estimates of density functions and regression curves. Theory Probab Appl 10:186–190

Neuhaus G (1993) Conditional rank tests for the two-sample problem under random censorship. Ann Stat 21:1760–1779

Pauly M (2011) Discussion about the quality of f-ratio resampling tests for comparing variances. TEST 20:163–179

Pauly M, Brunner E, Konietschke F (2015) Asymptotic permutation tests in general factorial designs. J R Stat Soc Ser B Stat Methodol 77:461–473

Potthoff R (1963) Use of the Wilcoxon statistic for a generalized Behrens–Fisher problem. Ann Math Stat 34:1596–1599

Price R, Bonett D (2001) Estimating the variance of the sample median. J Stat Comput Simul 68:295–305

R Core Team (2020) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Rao C, Mitra S (1971) Generalized inverse of matrices and its applications. Wiley, New York

Richter L, Norris S, Pettifor J, Yach D, Cameron N (2007) Cohort profile: Mandela's children: the 1990 birth to twenty study in South Africa. Int J Epidemiol 36:504–511

Richter L, Victora C, Hallal P, Adair L, Bhargava S, Fall C, Lee N, Martorell R, Norris S, Sachdev H, Stein A, Group C (2012) Cohort profile: the consortium of health-orientated research in transitioning societies. Int J Epidemiol 43:621–626

Schuster E (1969) Estimation of a probability density function and its derivatives. Ann Math Stat 40:1187–1195

Sen P (1962) On studentized non-parametric multi-sample location tests. Ann Inst Stat Math 14:119–131

Serfling R (2009) Approximation theorems of mathematical statistics. Wiley, New York

Silverman B (1978) Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. Ann Stat 6:177–184

Silverman B (1986) Density estimation. Chapman and Hall, London

Smaga Ł (2017) Diagonal and unscaled wald-type tests in general factorial designs. Electron J Stat 11:2613–2646

Stein A, Melgar P, Hoddinott J, Martorell R (2008) Cohort profile: the Institute of Nutrition of Central America and Panama (INCAP) nutrition trial cohort study. Int J Epidemiol 37:716–720

van der Vaart A, Wellner J (1996) Weak convergence and empirical processes. Springer series in statistics. Springer, New York

Victora C, Barros F (2006) Cohort profile: the 1982 Pelotas (Brazil) birth cohort study. Int J Epidemiol 35:237–242

Zhang JT (2012) An approximate degrees of freedom test for heteroscedastic two-way ANOVA. J Stat Plan Inference 142:336–346