Chapter 3
Classes of Multiple Test Procedures

Abstract The aim of this chapter is a systematic overview of different classes of multiple tests. Procedures are distinguished by their structure, by the degree of detail of the underlying statistical model and by the type of error control that they provide. Major categories comprise margin-based multiple tests, multivariate multiple test procedures and closed test procedures. Subcategories are introduced where appropriate. We discuss specific examples and indicate computer implementations by means of flow diagrams and pseudo-code. Applications and references to later chapters illustrate which kind of multiple test procedure can be utilized for some standard types of multiple test problems which are relevant in practice. Precise references to the literature are collected for a deeper study of specific methods.

Although the literature on multiple test procedures (MTPs) is nowadays exponentially increasing over time, it is still possible to systematize the proposed methods according to some general categories. For instance, one class of methods only models the marginal distributions of the involved test statistics explicitly and combines these test statistics or, equivalently, corresponding p-values following probabilistic calculations. We call resulting procedures margin-based multiple test procedures. Different margin-based MTPs employ different qualitative assumptions on the dependency structure between test statistics or p-values, cf. our Chap. 2. Examples of this kind of procedures are discussed in Sect. 3.1.

Another class of MTPs considers the full joint distribution of all test statistics and relies on calculating or approximating quantiles of this joint distribution, for instance by resampling or by proving asymptotic normality by means of central limit theorems. We term such procedures multivariate multiple test procedures and discuss them in Sect. 3.2. A class of in a certain sense hybrid (neither purely margin-based nor entirely multivariate) multiple test procedures, which are specifically tailored to control the FWER in structured systems of hypotheses, is constituted by closed test procedures, which we will treat in Sect. 3.3.

Further criteria to distinguish MTPs are their structure (single-step or stepwise rejective), and the type of error control (k-FWER-controlling, FDR-controlling, FDX-controlling, etc.) that they provide. We exclude a distinction between frequentist

T. Dickhaus, *Simultaneous Statistical Inference*,
DOI: 10.1007/978-3-642-45182-9_3,

and Bayesian procedures here, because this work is not considered with Bayesian approaches to multiple hypotheses testing. As far as frequentist procedures are concerned, the aforementioned criteria in our opinion allow us to treat the majority of the most popular MTPs up to present.

One type of procedures which do not fit in a clear-cut way into the categories defined above is constituted by so-called augmentation procedures. Augmentation procedures for control of the k-FWER, the FDR or the FDX work in two stages: In the first stage, an FWER-controlling MTP is applied. In the second stage, a certain number of hypotheses not rejected by the procedure employed in the first stage is rejected additionally, whereby this number in general depends on the data and on probabilistic bounds. Although augmentation procedures have attracted some attention recently, we do not cover them in the present work. References for augmentation procedures include van der Laan et al. (2004; 2005), and Farcomeni (2009).

3.1 Margin-Based Multiple Test Procedures

The multiple tests discussed in this section only require that each marginal test φ_i can be calibrated to keep a local significance level $\alpha_{\text{loc.}}$ (say). The multiple test $\varphi = (\varphi_i : 1 \leq i \leq m)$ is then built up from these marginal tests by adjusting $\alpha_{\text{loc.}}$ for the multiplicity of the problem. This adjustment may be given by an explicit "correction for multiplicity" based on probabilistic considerations or in a data-dependent manner, for instance by defining $\alpha_{\text{loc.}}$ by the value of an order statistic of marginal p-values $p_1, \ldots, p_m$.

3.1.1 Single-Step Procedures

Single-step multiple test procedures carry out each individual test φ_i, $1 \leq i \leq m$, at (local) significance level $\alpha_{\text{loc.}}$, where $\alpha_{\text{loc.}}$ is the result of a multiplicity correction of α. In view of Theorem 2.1, single-step multiple tests are extremely easy to carry out in practice: Just calculate marginal p-values $p_1, \ldots, p_m$ and reject H_i if and only if $p_i < \alpha_{\text{loc.}}$. The choice of $\alpha_{\text{loc.}}$ depends on qualitative assumptions regarding the joint distribution of $(p_1, \ldots, p_m)$. Two classical procedures are the Bonferroni correction (or Bonferroni test) and the Šidák correction (or Šidák test).

Example 3.1 (Bonferroni correction, cf. Bonferroni (1935; 1936)). The Bonferroni correction is based on the union bound and consists in choosing $\alpha_{\text{loc.}} = \alpha/m$. It provides strong control of the FWER without any assumptions on the dependency structure among $(p_1, \ldots, p_m)$, because for a Bonferroni test φ, it holds for all $\vartheta \in \Theta$ that

$$\begin{aligned}\mathrm{FWER}_{\vartheta}(\varphi) &= \mathbb{P}_{\vartheta}(\bigcup_{i\in I_0(\vartheta)}\{\varphi_i = 1\})\\ &\leq \sum_{i\in I_0(\vartheta)} \mathbb{P}_{\vartheta}(\{\varphi_i = 1\})\\ &\leq m_0\alpha/m \leq \alpha.\end{aligned}$$

The inequality $\mathbb{P}(\bigcup_{i=1}^{m} A_i) \leq \sum_{i=1}^{m} \mathbb{P}(A_i)$ is referred to as Bonferroni inequality in the multiple testing literature.

The disadvantage of Bonferroni tests is that α/m is very small for large m. Therefore, Bonferroni tests have low multiple power if m is large. If joint independence of all m marginal p-values can be assumed, $\alpha_{\text{loc.}}$ can be chosen slightly larger than α/m.

Example 3.2 (Šidák correction, cf. Šidák 1967). The Šidák correction consists in choosing $\alpha_{\text{loc.}} = 1 - (1 - \alpha)^{1/m}$. It provides strong control of the FWER if $(p_1, \ldots, p_m)$ are jointly stochastically independent, because for a Šidák test φ, it then holds for all $\vartheta \in \Theta$ that

$$\begin{aligned}\mathrm{FWER}_{\vartheta}(\varphi) &= \mathbb{P}_{\vartheta}(\bigcup_{i\in I_0(\vartheta)}\{\varphi_i = 1\})\\ &= 1 - \mathbb{P}_{\vartheta}(\bigcap_{i\in I_0(\vartheta)}\{\varphi_i = 0\})\\ &= 1 - \prod_{i\in I_0(\vartheta)} \mathbb{P}_{\vartheta}(\{\varphi_i = 0\})\\ &\leq 1 - \prod_{i\in I_0(\vartheta)} (1-\alpha)^{1/m}\\ &= 1 - (1-\alpha)^{m_0/m}\\ &\leq 1 - (1-\alpha) = \alpha.\end{aligned}$$

As mentioned before, for all $m \in \mathbb{N}$ it holds $\alpha/m < 1-(1-\alpha)^{1/m}$, so that the more restrictive model assumptions made for a Šidák test allow one to increase multiple power uniformly. We may remark here that Šidák tests control the FWER under certain forms of positive dependence among $(p_1, \ldots, p_m)$, too. More details are provided in Chap. 4. Also asymptotically, it holds $m[1-(1-\alpha)^{1/m}] \to -\ln(1-\alpha) > \alpha = m\alpha/m$, $m \to \infty$, for any $\alpha \in (0, 1)$. However, also for the Šidák correction, we have $\alpha_{\text{loc.}} \to 0$, $m \to \infty$.

In the particular context of testing linear contrasts in Gaussian models, Scheffé (1953) obtained the following result.

Theorem 3.1 (Scheffé (1953)). *Let $k \geq 3$ and $n_i \geq 2$ for all $1 \leq i \leq k$ be given integers and $X = (X_{ij} : 1 \leq i \leq k, 1 \leq j \leq n_i)$. Assume that all X_{ij} are stochastically independent and normally distributed, $X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$, where*

$\mu_i \in \mathbb{R}$, $1 \leq i \leq k$, and $\sigma^2 > 0$. For notational convenience, denote $n. = \sum_{i=1}^k n_i$. Consider the linear subspace

$$\mathscr{L} = \left\{ \sum_{j=1}^{q} h_j a^{(j)} \right\}$$

of $\mathbb{R}^k$ of dimension $q \leq k$, where $h_j \in \mathbb{R}$ for all $1 \leq j \leq q$ and $a^{(1)}, \ldots, a^{(q)} \in \mathbb{R}^k$ are linearly independent vectors. Then it holds for all $\mu \in \mathbb{R}^k$ and for all $\sigma^2 > 0$ that

$$\mathbb{P}_{(\mu,\sigma^2)}\left(\forall c \in \mathscr{L} : c^T \mu \in \left[c^T \hat{\mu} \mp \sqrt{q \widehat{Var}(c^T \hat{\mu}) F_{q,n.-k;\alpha}} \right]\right) = 1 - \alpha, \qquad (3.1)$$

where $\mu = (\mu_1, \ldots, \mu_k)^\top$, $\hat{\mu} = (\overline{X}_{1.}, \ldots, \overline{X}_{k.})^T$ (vector of empirical group means), and $\widehat{Var}(c^T \hat{\mu}) = s^2 \sum_{i=1}^k (c_i^2/n_i)$, with s^2 denoting the pooled unbiased estimator of σ^2, and $F_{q,n.-k;\alpha}$ the upper α-quantile of Fisher's F-distribution with q and $n. - k$ degrees of freedom.

Equation (3.1) yields a simultaneous $1-\alpha$ confidence region for *all* linear contrasts of group means defined by $\mathscr{L}$ in the considered analysis of variance model. By duality of tests and confidence regions (see Theorem 1.1), this also entails a multiple single-step test for such contrasts.

3.1.2 Stepwise Rejective Multiple Tests

An interesting other class of multiple test procedures are stepwise rejective tests. In contrast to single-step tests, here the hypotheses are ordered by a pre-defined criterion and tested one after the other, where testing can stop at every step due to the occurrence of a rejection or a non-rejection. This means that the test result for a particular pair of hypotheses H_i versus K_i depends on the data not only directly via the test statistic T_i or the p-value p_i, but also indirectly via potentially all other test statistics or p-values. The way the ordering among the hypotheses is defined leads to different subtypes of stepwise rejective multiple tests.

3.1.2.1 Step-Up-Down Tests

Step-up-down tests, introduced by Tamhane et al. (1998), rely on an ordering of the hypotheses $H_1, \ldots, H_m$ which is induced by the order statistics of marginal p-values $p_1, \ldots, p_m$.

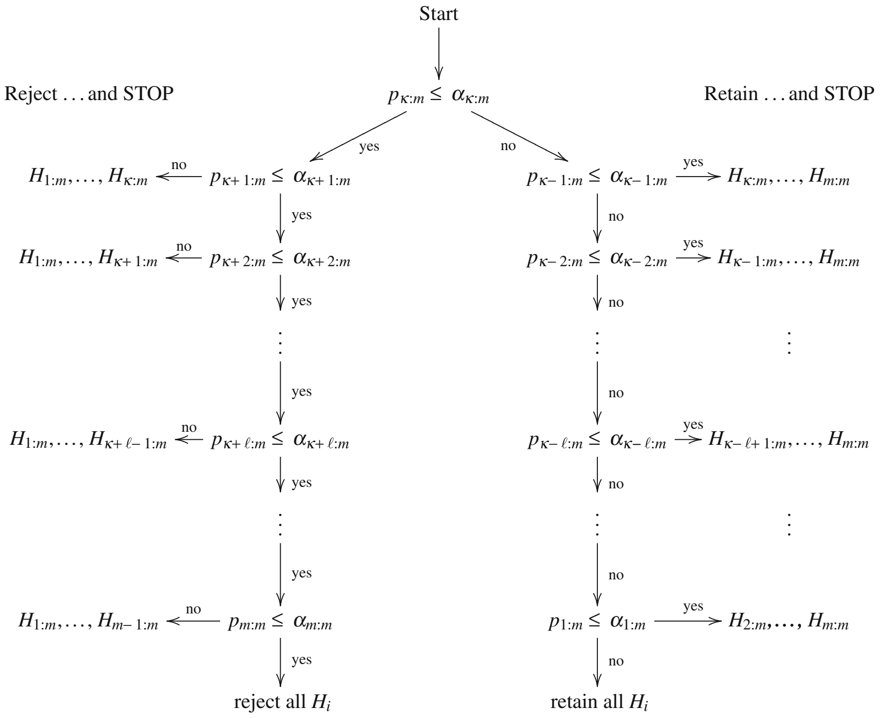
Start
Reject ... and STOP
$p_{\kappa:m} \leq \alpha_{\kappa:m}$
Retain ... and STOP
yes
no
$H_{1:m}, \ldots, H_{\kappa:m}$
no
$p_{\kappa+1:m} \leq \alpha_{\kappa+1:m}$
$p_{\kappa-1:m} \leq \alpha_{\kappa-1:m}$
yes
$H_{\kappa:m}, \ldots, H_{m:m}$
yes
no
$H_{1:m}, \ldots, H_{\kappa+1:m}$
no
$p_{\kappa+2:m} \leq \alpha_{\kappa+2:m}$
$p_{\kappa-2:m} \leq \alpha_{\kappa-2:m}$
yes
$H_{\kappa-1:m}, \ldots, H_{m:m}$
yes
no
yes
no
$H_{1:m}, \ldots, H_{\kappa+\ell-1:m}$
no
$p_{\kappa+\ell:m} \leq \alpha_{\kappa+\ell:m}$
$p_{\kappa-\ell:m} \leq \alpha_{\kappa-\ell:m}$
yes
$H_{\kappa-\ell+1:m}, \ldots, H_{m:m}$
yes
no
yes
no
$H_{1:m}, \ldots, H_{m-1:m}$
no
$p_{m:m} \leq \alpha_{m:m}$
$p_{1:m} \leq \alpha_{1:m}$
yes
$H_{2:m}, \ldots, H_{m:m}$
yes
no
reject all H_i
retain all H_i

Fig. 3.1 Decision rule of an SUD test. If $\kappa = m$ (SU test) and $p_{m:m} \leq \alpha_{m:m}$, all m null hypotheses are rejected. If $\kappa = 1$ (SD test) and $p_{1:m} > \alpha_{1:m}$, all m null hypotheses are retained

Definition 3.1 (Step-up-down test of order κ, cf. Finner et al. (2012)). Let $p_{1:m} < p_{2:m} < \cdots < p_{m:m}$ denote the ordered marginal p-values for a multiple test problem. For a tuning parameter $\kappa \in \{1, \ldots, m\}$ a step-up-down (SUD) test $\varphi^{\kappa} = (\varphi_1^{\kappa}, \ldots, \varphi_m^{\kappa})$ of order κ based on some critical values $\alpha_{1:m} \leq \cdots \leq \alpha_{m:m}$ is defined as follows. If $p_{\kappa:m} \leq \alpha_{\kappa:m}$, set $j^* = \max\{j \in \{\kappa, \ldots, m\} : p_{i:m} \leq \alpha_{i:m}$ for all $i \in \{\kappa, \ldots, j\}\}$, whereas for $p_{\kappa:m} > \alpha_{\kappa:m}$, put $j^* = \sup\{j \in \{1, \ldots, \kappa - 1\} : p_{j:m} \leq \alpha_{j:m}\}$ $(\sup \emptyset = -\infty)$. Define $\varphi_i^{\kappa} = 1$ if $p_i \leq \alpha_{j^*:m}$ and $\varphi_i = 0$ otherwise $(\alpha_{-\infty:m} = -\infty)$.

A step-up-down test of order $\kappa = 1$ or $\kappa = m$, respectively, is called step-down (SD) or step-up (SU) test, respectively. If all critical values are identical, we obtain a single-step test.

Figure 3.1 illustrates the decision rule of an SUD test schematically.

As we will discuss in Chap. 5, many commonly used step-up-down tests are margin-based and only employ qualitative assumptions regarding the joint distribution of test statistics or p-values. For instance, this holds true for the multiple tests by Holm (1979) (which are FWER-controlling step-down tests) and the famous linear step-up test by Benjamini and Hochberg (1995) for FDR control. However, there are remarkable exceptions, especially shortcuts of closed test procedures, cf. Sect. 3.3. The following obvious lemma can be used to compare different SUD tests which keep the same type I error criterion.

Lemma 3.1. *Consider two SUD tests $\varphi^{(1)}$ and $\varphi^{(2)}$ for the same multiple test problem $(\mathscr{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta}, \mathscr{H}_m)$. Assume that one of the following properties holds true.*

(a) The two tests $\varphi^{(1)}$ and $\varphi^{(2)}$ employ the same set of critical values and the tuning parameter κ_2 of $\varphi^{(2)}$ is larger than the tuning parameter κ_1 of $\varphi^{(1)}$.

(b) The two tests $\varphi^{(1)}$ and $\varphi^{(2)}$ employ the same tuning parameter κ and the critical values utilized in $\varphi^{(2)}$ are index-wise not smaller than the ones utilized in $\varphi^{(1)}$.

(c) Both tests $\varphi^{(1)}$ and $\varphi^{(2)}$ are single-step tests and the critical value utilized in $\varphi^{(2)}$ is larger than that utilized in $\varphi^{(1)}$.

Then, for any realization of $(p_1, \dots, p_m)^\top$, $\varphi^{(2)}$ rejects all hypotheses that are rejected by $\varphi^{(1)}$, and possibly more.

Hence, under the constraint of type I error control of given type and at given level, an optimal SUD test (with respect to multiple power, cf. Definition 1.4) is given by choosing κ and $\alpha_{1:m}, \dots, \alpha_{m:m}$ as large as possible. For instance, SU tests have higher (not smaller) multiple power than the corresponding SD tests (with the same set of critical values). On the other hand, the same holds true for the comparison with respect to the FWER. Let us mention that additional assumptions are required in order that more rejections entail larger FDR, cf. Theorem 5.7.

Notice that we implicitly used part (c) of Lemma 3.1 for the comparison of Bonferroni tests and Šidák tests. In Chap. 5, Lemma 3.1 will be used for discussing relationships between the dependency structure among $p_1, \dots, p_m$ and the choice of tuning parameters and critical values for SUD tests.

3.1.2.2 Fixed Sequence Multiple Tests

Similarly to step-up-down tests, fixed sequence multiple tests also rely on an ordering of the hypotheses $H_1, \dots, H_m$. However, the ordering is now not data-dependently given by the ordering of p-values or test statistics, but is pre-defined before testing starts, for instance by weighting the hypotheses for importance. With respect to control of the FWER, the following fixed sequence procedure is widely used.

Theorem 3.2. Let $(\mathscr{X}, \mathscr{F}, \mathscr{P}, \mathscr{H})$ with $\mathscr{H} = (H_i : 1 \leq i \leq m)$ denote a multiple test problem and assume that valid marginal p-values $p_1, \dots, p_m$ are at hand. Let $\alpha \in (0, 1)$ be a given constant and consider the multiple test φ defined by the following rule: Reject exactly hypotheses $H_1, \dots, H_{k^*}$, where

$$k^* = \max\{1 \leq i \leq m : p_j \leq \alpha \text{ for all } j = 1, \dots, i\}.$$

If k^* does not exist, retain all m null hypotheses. Then, φ strongly controls the FWER at level α.

Proof. First, consider the case $m = 2$. We have to distinguish four cases.

1. If both H_1 and H_2 are false, no type I error can occur, hence $\mathrm{FWER}_\vartheta(\varphi) = 0$ for such ϑ.
2. If only H_1 is true, $\mathrm{FWER}_\vartheta(\varphi) = \mathbb{P}_\vartheta(p_1(X) \leq \alpha) \leq \alpha$.
3. If only H_2 is true, $\mathrm{FWER}_\vartheta(\varphi) = \mathbb{P}_\vartheta(\{p_1(X) \leq \alpha\} \cap \{p_2(X) \leq \alpha\}) \leq \mathbb{P}_\vartheta(p_2(X) \leq \alpha) \leq \alpha$.
4. If both H_1 and H_2 are true, $\mathrm{FWER}_\vartheta(\varphi) = \mathbb{P}_\vartheta(p_1(X) \leq \alpha) \leq \alpha$.

It is easy to check that the latter reasoning remains to hold true for $m > 2$. □

The obvious drawback of the multiple test φ from Theorem 3.2 is that, once a particular hypothesis cannot be rejected, the remaining not yet rejected hypotheses have to be retained without being tested explicitly. Wiens (2003) developed a method based on a Bonferroni-type adjustment of α that allows for continuing testing after potential non-rejections. Other related testing strategies for fixed sequences of (pre-ordered) hypotheses ensuring strict FWER control have been discussed by Westfall and Krishen (2001) and Bauer et al. (1998), among many others. Such methods are particularly important for clinical trials with multiple endpoints.

3.1.3 Data-Adaptive Procedures

From the calculations in Examples 3.1 and 3.2, it follows that the realized k-FWER of the investigated margin-based multiple tests crucially depends on the proportion $\pi_0 = m_0/m$ of true null hypotheses. In Chap. 5, we will show that the same holds true for the realized FDR of many classical step-up-down tests. Data-adaptive procedures aim at adapting to the unknown quantity π_0 in order to exhaust the type I error level better and, consequently, increase multiple power of standard procedures. Explicitly adaptive (plug-in) procedures employ an estimate $\hat{\pi}_0$ and plug $\hat{\pi}_0$ into critical values, typically replacing m by $m \cdot \hat{\pi}_0$. In view of Definition 1.4 and Lemma 3.1, this increases multiple power at least on parameter subspaces on which $\mathbb{P}_\vartheta(\hat{\pi}_0 < 1)$ is large.

Maybe, the still most popular though, as well, the most ancient estimation technique for π_0 is the one of Schweder and Spjøtvoll (1982). It relies on a tuning parameter $\lambda \in [0, 1)$. Denoting the empirical cumulative distribution function (ecdf) of m marginal p-values by $\hat{F}_m$, the proposed estimator from Schweder and Spjøtvoll (1982) can be written as

$$\hat{\pi}_0 \equiv \hat{\pi}_0(\lambda) = \frac{1 - \hat{F}_m(\lambda)}{1 - \lambda}. \tag{3.2}$$

Among others, Storey et al. (2004), Langaas et al. (2005), Finner and Gontscharuk (2009), Dickhaus et al. (2012) and Dickhaus (2013) have investigated theoretical properties of $\hat{\pi}_0$ and slightly modified versions of this estimator. There exist several possible heuristic motivations for the usage of $\hat{\pi}_0$. The simplest one considers a

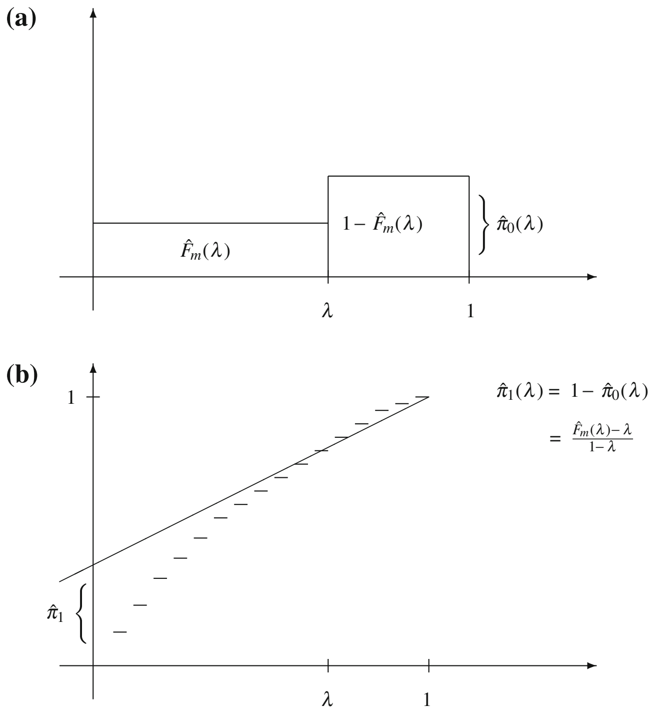

Fig. 3.2 Two graphical representations of the Schweder-Spjøtvoll estimator $\hat{\pi}_0(\lambda)$

histogram of the marginal p-values with exactly two bins, namely $[0, \lambda]$ and $(\lambda, 1]$. Then, the height of the bin associated with $(\lambda, 1]$ equals $\hat{\pi}_0(\lambda)$, see graph (a) in Fig. 3.2. A graphical algorithm for computing $\hat{\pi}_0$ connects the point $(\lambda, \hat{F}_m(\lambda))$ with the point $(1, 1)$. The offset of the resulting straight line at $t = 0$ equals $\hat{\pi}_1 = \hat{\pi}_1(\lambda) = 1 - \hat{\pi}_0(\lambda)$, see graph (b) in Fig. 3.2.

The following lemma is due to Dickhaus et al. (2012), see Lemma 1 in their paper.

Lemma 3.2. *Whenever $(p_1, \ldots, p_m)$ are valid p-values, i.e., marginally stochastically not smaller than UNI$[0, 1]$ under null hypotheses, the value of $\hat{\pi}_0$ is a conservative estimate of π_0, meaning that $\hat{\pi}_0$ has a non-negative bias. More specifically, it holds*

$$\mathbb{E}_\vartheta[\hat{\pi}_0(\lambda)] - \pi_0 \geq \frac{1}{m(1-\lambda)} \sum_{i \in I_1} \mathbb{P}_\vartheta(p_i > \lambda) \geq 0.$$

The data-adaptive Bonferroni plug-in (BPI) test by Finner and Gontscharuk (2009) replaces m by $m \cdot \hat{\pi}_0$ in the Bonferroni-corrected threshold for marginal p-values and the asymptotic version of the data-adaptive multiple test procedure by Storey et al. (2004) (STS test) replaces m by $m \cdot \hat{\pi}_0$ in Simes' critical values, cf. Sect. 5.3.

Another class of data-adaptive multiple tests is constituted by two-stage or multistage adaptive procedures, see Benjamini and Hochberg (2000) or Benjamini et al. (2006), for example. Such methods employ the number of rejections of a multiple test applied in the first stage in an estimator for m_0. This estimator is then used to calibrate the second stage test which leads to the actual decisions, where this principle may be applied iteratively. A third of class of methods is given by implicitly adaptive procedures. Here, the idea is to find critical values that automatically (for as many values of π_0 as possible) lead to full exhaustion of the type I error level. To this end, worst-case situations (i.e., LFCs) build the basis for the respective calculations. We will present some of such implicitly adaptive multiple tests in Sect. 5.5. Further estimation techniques for π_0 have also been proposed in the multiple testing literature. We defer the reader to the introduction in Finner and Gontscharuk (2009) for an overview.

3.2 Multivariate Multiple Test Procedures

The basic idea behind multivariate multiple test procedures is to incorporate the dependency structure of the data explicitly into the multiple test and thereby optimizing its power. The general reason why this is often possible is that margin-based procedures which control a specific multiple type I error rate have to provide this multiple type I error control generically over a potentially very large family of dependency structures. Hence, if it is possible to derive or to approximate the particular dependency structure for the data-generating distribution at hand, this information may be helpful to fine-tune a multiple test for this specific case. This is particularly important for applications from modern life sciences, because the data there are often spatially, temporally, or spatio-temporally correlated as we will demonstrate in later chapters. Three alternative ways to approximate dependency structures are resampling (Sect. 3.2.1), proving asymptotic normality by means of central limit theorems (Sect. 3.2.2), and fitting copula models (Sect. 3.2.3).

3.2.1 Resampling-Based Methods

It is fair to say that the basic reference for resampling-based FWER control is the book by Westfall and Young (1993), who introduced simultaneous and step-down multiple tests based on resampling under the assumption of subset pivotality (see Definition 4.3, basically meaning that the joint distribution of test statistics corresponding to true null hypotheses does not depend on the distribution of the remaining test statistics such that resampling under the global hypothesis H_0 is not only providing weak, but also strong FWER control). This assumption has been criticized as too restrictive such that (among others) Troendle (1995) and Romano and Wolf (2005a, b) generalized the methods of Westfall and Young (1993) to dispense with subset pivotality.

FDR-controlling (asymptotic) multiple tests based on resampling have been derived by Yekutieli and Benjamini (1999), Troendle (2000), and Romano et al. (2008). The resampling methods developed by Dudoit and van der Laan (2008) (see also the references therein) provide a general framework for controlling a variety of error rates (some of which we have introduced in Definitions 1.2 and 1.3), with particular emphasis on applications in genetics. While resampling often only asymptotically (for the sample size n tending to infinity) reproduces the true data distribution, Arlot et al. (2010) provide an in-depth study of resampling methods that control the FWER strictly for finite n.

3.2.2 Methods Based on Central Limit Theorems

Asymptotic normality of moment and maximum likelihood estimators are classical results in mathematical statistics, see, for instance, Chap. 12 by Lehmann and Romano (2005) or Chap. 5 by Van der Vaart (1998). We will discuss the special cases of multiple linear regression models and of generalized linear models in Chap. 4. If the vector T of test statistics for a given multiple test problem is (a transformation of) such an asymptotically normal point estimator, the asymptotic distribution of T can be derived and utilized for calibrating the multiple test. This has been demonstrated, for instance, by Hothorn et al. (2008) and Bretz et al. (2010) in general parametric models. For particular applications in genetic association studies (cf. Chap. 9), central limit theorems for multinomial distributions, together with positive dependency properties of multivariate chi-square distributions, have been exploited by Moskvina and Schmidt (2008) and Dickhaus and Stange (2013) (see also the references therein).

3.2.3 Copula-Based Methods

As discussed in Chap. 2, p-values are under certain assumptions uniformly distributed on [0, 1] under null hypotheses. In particular, this holds true in many models which are typically used in life science applications. One example is the problem of multiple testing for differential gene expression, see Chap. 10. Hence, according to Theorem 2.4, in such cases it suffices to estimate the (often unknown) copula of $p_1(X), \ldots, p_m(X)$ in order to calibrate a multivariate multiple test procedure operating on these p-values. In particular, parametric copula models are convenient, because the dependency structure can in such models be condensed into a low-dimensional copula parameter. A flexible class of copula models is constituted by the family of Archimedean copulae.

Definition 3.2 (Archimedean copula). The joint distribution of the random vector $(p_i(X) : 1 \leq i \leq m)$ under $\vartheta \in \Theta$ is given by an Archimedean copula with copula generator ψ, if for all $(t_1, \ldots, t_m)^\top \in [0, 1]^m$,

$$\mathbb{P}_{\vartheta,\psi}(p_1(X) \leq t_1, \ldots, p_m(X) \leq t_m) = \psi\left(\sum_{i=1}^{m} \psi^{-1}\left(F_{p_i(X)}(t_i)\right)\right), \tag{3.3}$$

where $F_{p_i(X)}$ denotes the marginal cdf of $p_i(X)$ under $\vartheta \in \Theta$.

Dickhaus and Gierl (2013) demonstrated the usage of Archimedean copula models for FWER control, while Bodnar and Dickhaus (2013) are considered with FDR control under Archimedean p-value copulae. If the generator ψ only depends on a copula parameter η (say), standard parametric estimation approaches can be employed to estimate η. Two plausible estimation strategies are the maximum likelihood method (see, e. g., Hofert et al. (2012)) or the method of moments (referred to as "realized copula" method by Fengler and Okhrin (2012)). For the latter approach, the "inversion formulas" provided in the following lemma are helpful.

Lemma 3.3. *Let X and Y two real-valued random variables with marginal cdfs F_X and F_Y and bivariate copula C_η, depending on a copula parameter η. Let $\sigma_{X,Y}$, $\rho_{X,Y}$ and $\tau_{X,Y}$ denote (the population versions of) the covariance, Spearman's rank correlation coefficient and Kendall's tau, respectively, of X and Y. Then it holds:*

$$\sigma_{X,Y} = f_1(\eta) = \int_{\mathbb{R}^2} \left[C_\eta\{F_X(x), F_Y(y)\} - F_X(x)F_Y(y)\right] dx\, dy, \tag{3.4}$$

$$\rho_{X,Y} = f_2(\eta) = 12 \int_{[0,1]^2} C_\eta(u, v)\, du\, dv - 3, \tag{3.5}$$

$$\tau_{X,Y} = f_3(\eta) = 4 \int_{[0,1]^2} C_\eta(u, v)\, dC_\eta(u, v) - 1. \tag{3.6}$$

Proof. Equation (3.4) is due to Höffding (1940), Eq. (3.5) is Theorem 5.1.6. in Nelsen (2006) and (3.6) is Theorem 5.1.3 in Nelsen (2006). □

The "realized copula" method for empirical calibration of a one-dimensional parameter η of an m-variate copula essentially considers every of the $m(m-1)/2$ pairs of the m underlying random variables $X_1, \ldots, X_m$, inverts (3.4) each time with respect to η, replaces the population covariance by its empirical counterpart and aggregates the resulting $m(m-1)/2$ estimates in an appropriate way. More specifically, Fengler and Okhrin (2012) define for $1 \leq i < j \leq m$: $g_{ij}(\eta) = \hat{\sigma}_{ij} - f_1(\eta)$, set $\mathbf{g}(\eta) = (g_{ij}(\eta))_{1 \leq i < j \leq m}$, and propose to estimate

$$\hat{\eta} = \arg\min_{\eta} \mathbf{g}^\top(\eta)\mathbf{W}\mathbf{g}(\eta)$$

for an appropriate weight matrix $\mathbf{W} \in \mathbb{R}^{\binom{m}{2} \times \binom{m}{2}}$. In this, $\hat{\sigma}_{ij}$ denotes the empirical covariance of X_i and X_j. Indeed, any of the functions $f_\ell, \ell = 1, 2, 3$ corresponding to relationships (3.4)–(3.6) may be employed in this realized copula method. Moreover, they may be combined to estimate two- or three-dimensional copula parameters η.

In the particular context of estimating p-value copulae in multiple testing models, it is infeasible to actually draw independent replications of the vector $(p_i(X) : 1 \leq i \leq m)$ from the target population, because this would essentially mean to carry out the entire experiment several times. Hence, one typically employs resampling methods for estimating the dependency structure among the p-values, namely the parametric bootstrap or permutations if $H_1, \ldots, H_m$ correspond to marginal two-sample problems. Pollard and van der Laan (2004) compared both approaches and argued that the permutation method reproduces the correct null distribution only under some conditions. However, if these conditions are met, the permutation approach is often superior to bootstrapping (see also Westfall and Young (1993) and Meinshausen et al. (2011)). Furthermore, it is important to notice that both bootstrap and permutation-based methods estimate the joint distribution of $(p_i(X) : 1 \leq i \leq m)$ under the global null hypothesis H_0. Hence, the assumption that η is a nuisance parameter which does not depend on ϑ is an essential prerequisite for the applicability of such resampling methods for estimating η.

3.3 Closed Test Procedures

An important class of FWER-controlling multiple tests which do not exactly fall into one of the categories "margin-based" and "multivariate" is constituted by closed test procedures, introduced by Marcus et al. (1976).

Theorem 3.3. Let $\mathscr{H} = \{H_i : i \in I\}$ denote a $\cap$-closed system of hypotheses and $\varphi = (\varphi_i : i \in I)$ a coherent multiple test for $(\mathscr{X}, \mathscr{F}, \mathscr{P}, \mathscr{H})$ at local level α. Then, φ is a strongly FWER-controlling multiple test at FWER level α for $(\mathscr{X}, \mathscr{F}, \mathscr{P}, \mathscr{H})$.

Proof. Let $\vartheta \in \Theta$ with $I_0(\vartheta) \neq \emptyset$. Since $\mathscr{H}$ is $\cap$-closed, there exists an $i \in I$ with $H_i = \bigcap_{j \in I_0(\vartheta)} H_j$, and $\vartheta \in H_i$. Hence, for all $j \in I_0(\vartheta)$, we have $H_j \supseteq H_i$. Now, coherence of φ entails $\{\varphi_i = 1\} \supseteq \bigcup_{j \in I_0(\vartheta)} \{\varphi_j = 1\}$. We conclude that

$$\text{FWER}_\vartheta(\varphi) = \mathbb{P}_\vartheta\left(\bigcup_{j \in I_0(\vartheta)} \{\varphi_j = 1\}\right) \leq \mathbb{P}_\vartheta(\{\varphi_i = 1\}) \leq \alpha,$$

because φ_i is a level α test. □

Notice that there is no restriction at all regarding the explicit form of the local level α tests φ_i in Theorem 3.3. One is completely free in choosing these tests. The decisive property of φ, however, is coherence. Not all multiple tests fulfill this property in the first place. This leads to the closed test principle, a "general solution to multiple testing problems" (Sonnemann (2008)).

Theorem 3.4 (Closure Principle, see Marcus et al. (1976), Sonnemann (2008)). Let $\mathscr{H} = \{H_i : i \in I\}$ denote a $\cap$-closed system of hypotheses and $\varphi = (\varphi_i : i \in I)$ an (arbitrary) multiple test for $(\mathscr{X}, \mathscr{F}, \mathscr{P}, \mathscr{H})$ at local level α. Then, we define the closed multiple test procedure (closed test) $\bar{\varphi} = (\bar{\varphi}_i : i \in I)$ based on φ by

$$\forall i \in I : \bar{\varphi}_i(x) = \min_{j : H_j \subseteq H_i} \varphi_j(x).$$

It holds:

(a) The closed test $\bar{\varphi}$ strongly controls the FWER at level α.
(b) For all $\emptyset \neq I' \subset I$, the "restricted" closed test $\bar{\varphi}' = (\bar{\varphi}_i : i \in I')$ is a strongly (at level α) FWER-controlling multiple test for $\mathscr{H}' = \{H_i : i \in I'\}$.
(c) Both tests $\bar{\varphi}$ and $\bar{\varphi}'$ are coherent.

Proof. The assertions follow immediately from the definitions of $\bar{\varphi}$ and $\bar{\varphi}'$ by making use of Theorem 3.3. □

Remark 3.1.

(a) The closed test $\bar{\varphi}$ based on φ rejects a particular hypothesis $H_i \in \mathscr{H}$ if and only if φ rejects H_i and all hypotheses $H_j \in \mathscr{H}$ of which H_i is a superset (implication).
(b) If $\mathscr{H}$ is not $\cap$-closed, then one can extend $\mathscr{H}$ by adding all missing intersection hypotheses, leading to the $\cap$-closed system of hypotheses $\bar{\mathscr{H}}$. If there are ℓ elementary hypotheses in $\mathscr{H}$, then $\bar{\mathscr{H}}$ can consist of up to $2^\ell - 1$ hypotheses. However, as we will demonstrate by specific examples, it is typically not necessary to test all elements in $\bar{\mathscr{H}}$ explicitly.
(c) Theorem 3.3 shows that under certain assumptions a multiple test at local level α is a strongly FWER-controlling multiple test at level α. Of course, the reverse statement is always true.
(d) If $\mathscr{H}$ is disjoint in the sense that $\forall i, j \in I, i \neq j : H_i \cap H_j = \emptyset$, and φ is a multiple test for $(\mathscr{X}, \mathscr{F}, \mathscr{P}, \mathscr{H})$ at local level α, then φ automatically strongly controls the FWER at level α, because φ is coherent and $\mathscr{H}$ is $\cap$-closed by the respective definitions. Often, there exist many possibilities for partitioning Θ in disjoint subsets, leading to the more general partitioning principle, see Finner and Strassburger (2002).
(e) If $I = \Theta$ and $H_\vartheta = \{\vartheta\}$ for all $\vartheta \in \Theta$, and if $\varphi = (\varphi_\vartheta : \vartheta \in \Theta)$ is a multiple test at local level α, then φ strongly controls the FWER at level α.

A nice application of the closed test principle is the problem of directional or type III errors, cf. Finner (1999) and references therein.

Example 3.3 (Two-sample t-test). Assume that we can observe $X = (X_{ij})$ for $i = 1, 2$ and $j = 1, \ldots, n_i$, that all X_{ij} are stochastically independent and

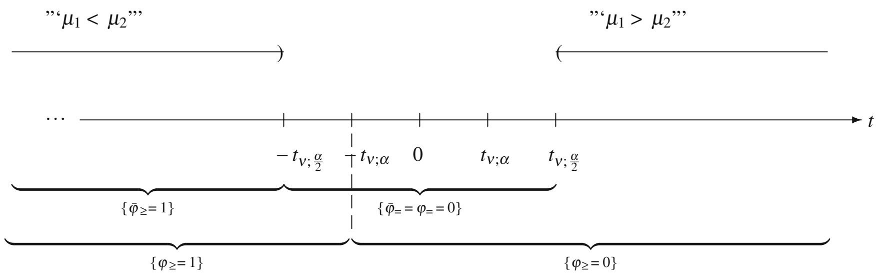

Fig. 3.3 Closed test for $\{H_=, H_\leq, H_\geq\}$ in the two-sample Gaussian model

$X_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$ with unknown variance $\sigma^2 > 0$. Consider the hypothesis $H_= : \{\mu_1 = \mu_2\}$. The two-sample t-test $\varphi_=$ (say) for testing $H_=$ is based on the test statistic

$$T(X) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\,\frac{\bar{X}_{1.} - \bar{X}_{2.}}{S}, \quad \text{where } S^2 = \frac{1}{\nu}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.})^2, \quad \nu = n_1 + n_2 - 2,$$

and is given by

$$\varphi_=(x) = \left\{ \begin{array}{cccc} 1 & & > & \\ & |T(x)| & & t_{\nu;\alpha/2} \\ 0 & & \leq & \end{array} \right\},$$

where $t_{\nu;\alpha/2}$ denotes the upper $\alpha/2$-quantile of Student's t-distribution with ν degrees of freedom. Let us restrict our attention to the case $\alpha \in (0, 1/2)$. The problem of directional or type III errors can be stated as follows. Assume that $H_=$ is rejected by $\varphi_=$. Can one then infer that $\mu_1 < \mu_2$ ($\mu_1 > \mu_2$) if $T(x) < -t_{\nu;\alpha/2}$ ($T(x) > t_{\nu;\alpha/2}$)? There is the possibility of an error of the third kind, namely, that $\mu_1 < \mu_2$ and $T(x) > t_{\nu;\alpha/2}$ ($\mu_1 > \mu_2$ and $T(x) < -t_{\nu;\alpha/2}$). The formal mathematical solution to this problem is given by the closed test principle. We add the two hypotheses $H_\leq : \{\mu_1 \leq \mu_2\}$ and $H_\geq : \{\mu_1 \geq \mu_2\}$ and notice that $H_= = H_\leq \cap H_\geq$. Level α tests for $H_\leq$ and $H_\geq$ are given by one-sided t-tests, say

$$\varphi_\leq(x) = \left\{ \begin{array}{cccc} 1 & & > & \\ & T(x) & & t_{\nu;\alpha} \\ 0 & & \leq & \end{array} \right\}, \quad \varphi_\geq(x) = \left\{ \begin{array}{cccc} 1 & & < & \\ & T(x) & & -t_{\nu;\alpha} \\ 0 & & \geq & \end{array} \right\}.$$

We construct the closed test $\bar{\varphi} = (\bar{\varphi}_\leq, \bar{\varphi}_=, \bar{\varphi}_\geq)$, given by $\bar{\varphi}_= = \varphi_=$, $\bar{\varphi}_\leq = \varphi_=\varphi_\leq$, $\bar{\varphi}_\geq = \varphi_=\varphi_\geq$.

Due to the nestedness of the rejection regions of $\varphi_\leq$ and $\bar{\varphi}_\leq$ ($\varphi_\geq$ and $\bar{\varphi}_\geq$), see Fig. 3.3, it follows from Theorem 3.4 that type III errors are automatically controlled at level α, hence, one-sided decisions after two-sided testing are allowed in this model. The argumentation further shows that this is generally true for likelihood ratio test statistics, provided that the model implies an isotone likelihood ratio.

The presumably most intensively studied application of closed test procedures, however, is the context of analysis of variance models, where linear contrasts regarding the group-specific means are of interest. Since this field of application has already deeply been studied in earlier books (Hochberg and Tamhane (1987), Hsu (1996)), we abstain from covering it here. Closed test-related multiple testing strategies for systems of hypotheses with a tree structure have been worked out by Meinshausen (2008) and Goeman and Finos (2012); see also the references in these papers. In the latter case, power can be gained by exploiting the logical restrictions among the hypotheses which are given by the tree structure. This has some similarities to the methods considered by Westfall and Tobias (2007).

Acknowledgments The material on copula-based multiple tests originated from joint work with Taras Bodnar, Jakob Gierl and Jens Stange. Helmut Finner and Klaus Straßburger taught me everything I know about closed test procedures and the partitioning principle. Special thanks are due to Mareile Große Ruse for programming the LaTeX code used for the figures.

References

Arlot S, Blanchard G, Roquain E (2010) Some nonasymptotic results on resampling in high dimension. II: Multiple tests. Ann Stat 38(1):83–99. doi:10.1214/08-AOS668

Bauer P, Röhmel J, Maurer W, Hothorn L (1998) Testing strategies in multi-dose experiments including active control. Stat Med 17(18):2133–2146

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

Benjamini Y, Hochberg Y (2000) On the adaptive control of the false discovery rate in multiple testing with independent statistics. J Edu Behav Stat 25:60–83

Benjamini Y, Krieger AM, Yekutieli D (2006) Adaptive linear step-up procedures that control the false discovery rate. Biometrika 93(3):491–507

Bodnar T, Dickhaus T (2013) False Discovery Rate Control under Archimedean Copula. arXiv:1305.3897

Bonferroni CE (1935) Il calcolo delle assicurazioni su gruppi di teste. Studi in onore Salvatore Ortu Carboni 13–60.

Bonferroni CE (1936) Teoria statistica delle classi e calcolo delle probabilita. Pubbl. d. R. Ist. Super. di Sci. Econom. e Commerciali di Firenze 8. Firenze: Libr. Internaz. Seeber.

Bretz F, Hothorn T, Westfall P (2010) Multiple Comparisons Using R. Chapman and Hall/CRC

Dickhaus T (2013) Randomized p-values for multiple testing of composite null hypotheses. J Stat Plann Infer 143(11):1968–1979

Dickhaus T, Gierl J (2013) Simultaneous test procedures in terms of p-value copulae. Global Science and Technology Forum (GSTF). In: Proceedings on the 2nd Annual International Conference on Computational Mathematics, Computational Geometry and Statistics (CMCGS 2013), vol 2, pp 75–80

Dickhaus T, Stange J (2013) Multiple point hypothesis test problems and effective numbers of tests for control of the family-wise error rate. Calcutta Statist Assoc Bull, to appear

Dickhaus T, Strassburger K, Schunk D, Morcillo-Suarez C, Illig T, Navarro A (2012) How to analyze many contingency tables simultaneously in genetic association studies. Stat Appl Genet Mol Biol 11(4):Article 12

Dudoit S, van der Laan MJ (2008) Multiple testing procedures with applications to genomics. Springer Series in Statistics. Springer, New York

Farcomeni A (2009) Generalized augmentation to control false discovery exceedance in multiple testing. Scand J Stat 36(3):501–517

Fengler MR, Okhrin O (2012) Realized Copula. SFB 649 Discussion Paper 2012–034, Sonderforschungsbereich 649, Humboldt-Universität zu Berlin, Germany, available at http://sfb649.wiwi.hu-berlin.de/papers/pdf/SFB649DP2012-034.pdf

Finner H (1999) Stepwise multiple test procedures and control of directional errors. Ann Stat 27(1):274–289. doi:10.1214/aos/1018031111

Finner H, Gontscharuk V (2009) Controlling the familywise error rate with plug-in estimator for the proportion of true null hypotheses. J Roy Stat Soc B 71(5):1031–1048

Finner H, Strassburger K (2002) The partitioning principle: a powerful tool in multiple decision theory. Ann Stat 30(4):1194–1213

Finner H, Gontscharuk V, Dickhaus T (2012) False discovery rate control of step-up-down tests with special emphasis on the asymptotically optimal rejection curve. Scand J Stat 39:382–397

Goeman JJ, Finos L (2012) The inheritance procedure: multiple testing of tree-structured hypotheses. Stat Appl Genet Mol Biol 11(1):Article 11

Hochberg Y, Tamhane AC (1987) Multiple comparison procedures. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics. Wiley, New York

Hofert M, Mächler M, McNeil AJ (2012) Likelihood inference for Archimedean copulas in high dimensions under known margins. J Multivariate Anal 110:133–150. doi:10.1016/j.jmva.2012.02.019

Höffding W (1940) Maßstabinvariante Korrelationstheorie. Schr math Inst u Inst angew Math Univ Berlin 5:181–233

Holm SA (1979) A simple sequentially rejective multiple test procedure. Scand J Stat Theory Appl 6:65–70

Hothorn T, Bretz F, Westfall P (2008) Simultaneous inference in general parametric models. Biom J 50(3):346–363

Hsu JC (1996) Multiple comparisons: theory and methods. Chapman and Hall, London

van der Laan MJ, Dudoit S, Pollard KS (2004), Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. Stat Appl Genet Mol Biol 3: Article15

van der Laan MJ, Birkner MD, Hubbard AE (2005), Empirical Bayes and resampling based multiple testing procedure controlling tail probability of the proportion of false positives. Stat Appl Genet Mol Biol 4:Article29

Langaas M, Lindqvist BH, Ferkingstad E (2005) Estimating the proportion of true null hypotheses, with application to DNA microarray data. J Roy Stat Soc B 67(4):555–572. doi:10.1111/j.1467-9868.2005.00515.x

Lehmann EL, Romano JP (2005) Testing statistical hypotheses. Springer Texts in Statistics, 3rd edn. Springer, New York

Marcus R, Peritz E, Gabriel KR (1976) On closed test procedures with special reference to ordered analysis of variance. Biometrika 63(3):655–660

Meinshausen N (2008) Hierarchical testing of variable importance. Biometrika 95(2):265–278. doi:10.1093/biomet/asn007

Meinshausen N, Maathuis MH, Bühlmann P (2011) Asymptotic optimality of the Westfall-Young permutation procedure for multiple testing under dependence. Ann Stat 39(6):3369–3391. doi:10.1214/11-AOS946

Moskvina V, Schmidt KM (2008) On multiple-testing correction in genome-wide association studies. Genet Epidemiol 32:567–573

Nelsen RB (2006) An introduction to copulas. Springer series in statistics. 2nd edn. Springer,New York

Pollard KS, van der Laan MJ (2004) Choice of a null distribution in resampling-based multiple testing. J Stat Plann Infer 125(1–2):85–100. doi:10.1016/j.jspi.2003.07.019

Romano JP, Wolf M (2005a) Exact and approximate stepdown methods for multiple hypothesis testing. J Am Stat Assoc 100(469):94–108. doi:10.1198/016214504000000539

Romano JP, Wolf M (2005) Stepwise multiple testing as formalized data snooping. Econometrica 73(4):1237–1282. doi:10.1111/j.1468-0262.2005.00615.x

Romano JP, Shaikh AM, Wolf M (2008) Control of the false discovery rate under dependence using the bootstrap and subsampling. Test 17(3):417–442. doi:10.1007/s11749-008-0126-6

Scheffé H (1953) A method for judging all contrasts in the analysis of variance. Biometrika 40:87–110

Schweder T, Spjøtvoll E (1982) Plots of P-values to evaluate many tests simultaneously. Biometrika 69:493–502

Šidák Z (1967) Rectangular confidence regions for the means of multivariate normal distributions. J Am Stat Assoc 62:626–633. doi:10.2307/2283989

Sonnemann E (2008) General solutions to multiple testing problems. Translation of "Sonnemann, E. (1982). Allgemeine Lösungen multipler Testprobleme. EDV in Medizin und Biologie 13(4), 120–128". Biom J 50:641–656

Storey JD, Taylor JE, Siegmund D (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. J. Roy. Stat. Soc. B, Stat Methodol 66(1):187–205

Tamhane AC, Liu W, Dunnett CW (1998) A generalized step-up-down multiple test procedure. Can J Stat 26(2):353–363. doi:10.2307/3315516

Troendle JF (1995) A stepwise resampling method of multiple hypothesis testing. J Am Stat Assoc 90(429):370–378. doi:10.2307/2291163

Troendle JF (2000) Stepwise normal theory multiple test procedures controlling the false discovery rate. J Stat Plann Infer 84(1–2):139–158. doi:10.1016/S0378-3758(99)00145-7

Van der Vaart A (1998) Asymptotic statistics. Cambridge Series in Statistical and Probabilistic Mathematics, 3. Cambridge University Press: Cambridge. doi:10.1017/CBO9780511802256

Westfall PH, Krishen A (2001) Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. J Stat Plann Infer 99(1):25–40. doi:10.1016/S0378-3758(01)00077-5

Westfall PH, Tobias RD (2007) Multiple testing of general contrasts: truncated closure and the extended Shaffer-Royen method. J Am Stat Assoc 102(478):487–494. doi:10.1198/016214506000001338

Westfall PH, Young SS (1993) Resampling-based multiple testing: examples and methods for p-value adjustment. Wiley Series in Probability and Mathematical Statistics, Applied Probability and Statistics. Wiley, New York

Wiens BL (2003) A fixed sequence Bonferroni procedure for testing multiple endpoints. Pharmaceut Statist 2:211215

Yekutieli D, Benjamini Y (1999) Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. J Stat Plann Infer 82(1–2):171–196. doi:10.1016/S0378-3758(99)00041-5