Chapter 2
Some Theory of *p*-values

Abstract Many multiple test procedures are formalized and carried out in practice by means of p-values. In this chapter, we formally introduce the notion of a p-value and its usage for testing a statistical hypothesis. Methods for computing p-values are discussed with respect to tests of Neyman-Pearson type and for discrete statistical models. In the context of testing multiple hypotheses, we introduce the concept of local significance levels. Randomized p-values are discussed for situations with multiple composite hypotheses and for discretely distributed test statistics. Some p-value models commonly used in multiple testing literature are explained. In view of stepwise rejective multiple test procedures, properties of order statistics of p-values are discussed for some of these models.

Many (stepwise) multiple tests are formalized and carried out by means of p-values corresponding to (marginal) test statistics. In the statistical literature, there exists an overwhelming debate whether p-values are suitable decision tools, cf. the references in Sect. 3.11 of Lehmann and Romano (2005). In this work, we pragmatically regard a p-value as a deterministic transformation of a test statistic which is particularly useful for multiple testing, because it provides a standardization. Every p-value is supported on the unit interval [0, 1], even if test statistics have drastically different scales.

Definition 2.1 (***p*-value**). Let $(\mathscr{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta\in\Theta})$ a statistical model and φ a (one-dimensional) non-randomized test for the single pair of hypotheses $\emptyset \neq H \subset \Theta$ versus $K = \Theta \setminus H$. Assume that φ is based on a real-valued test statistic $T : \mathscr{X} \to \mathbb{R}$. More specifically, let φ be characterized by rejection regions $\Gamma_\alpha \subset \mathbb{R}$ for any given significance level $\alpha \in (0, 1)$, such that $\varphi(x) = 1 \iff T(x) \in \Gamma_\alpha$ for $x \in \mathscr{X}$. Then, we define the p-value of an observation $x \in \mathscr{X}$ with respect to φ by

$$p_\varphi(x) = \inf_{\{\alpha : T(x)\in\Gamma_\alpha\}} \mathbb{P}^*(T(X) \in \Gamma_\alpha),$$

T. Dickhaus, *Simultaneous Statistical Inference*,
DOI: 10.1007/978-3-642-45182-9_2,

where the probability measure $\mathbb{P}^*$ is chosen such that

$$\mathbb{P}^*(T(X) \in \Gamma_\alpha) = \sup_{\vartheta \in H} \mathbb{P}_\vartheta(T(X) \in \Gamma_\alpha),$$

if H is a composite null hypothesis.

Remark 2.1.

(i) If H contains only one single element ϑ_0 (H is a simple hypothesis) and if $\mathbb{P}_H \equiv \mathbb{P}_{\vartheta_0}$ is continuous, it (typically) holds

$$p_\varphi(x) = \inf\{\alpha : T(x) \in \Gamma_\alpha\}.$$

(ii) In view of (3.3) in Lehmann and Romano (2005), we may regard the p-value as the "observed size" of φ.

(iii) Let Ω denote the domain of X. The mapping $p_\varphi(X) : \Omega \to [0, 1], \omega \mapsto p_\varphi(X(\omega))$, can be regarded as a random variable (under measurability assumptions). Often, there is no clear-cut distinction between the value $p_\varphi(x) \in [0, 1]$ and the random variable $p_\varphi(X)$. We will try to be as precise as possible with respect to this.

Definition 2.2. Under the assumptions of Definition 2.1, let the test statistic T fulfill the monotonicity condition

$$\forall \vartheta_0 \in H : \forall \vartheta_1 \in K : \forall c \in \mathbb{R} : \mathbb{P}_{\vartheta_0}(T(X) > c) \leq \mathbb{P}_{\vartheta_1}(T(X) > c). \tag{2.1}$$

Then, we call φ a test of (generalized) Neyman-Pearson type, if for all $\alpha \in (0, 1)$ there exists a constant c_α, such that

$$\varphi(x) = \begin{cases} 1, & T(x) > c_\alpha, \\ 0, & T(x) \leq c_\alpha. \end{cases}$$

In practice, the constants c_α are determined via $c_\alpha = \inf\{c \in \mathbb{R} : \mathbb{P}^*(T(X) > c) \leq \alpha\}$ with $\mathbb{P}^*$ as in Definition 2.1 ("at the boundary of the null hypothesis"). If H is simple and $\mathbb{P}_H$ continuous, we obtain $c_\alpha = F_T^{-1}(1-\alpha)$, where F_T denotes the cdf. of $T(X)$ under H.

Lemma 2.1. *Let φ a test of Neyman-Pearson type and assume that $\mathbb{P}^*$ does not depend on α. Then it holds*

$$p_\varphi(x) = \mathbb{P}^*(T(X) \geq t^*) \text{ with } t^* = T(x).$$

Proof. The rejection regions $\Gamma_\alpha = (c_\alpha, \infty)$ are nested. Therefore, $\inf\{\alpha : T(x) \in \Gamma_\alpha\}$ is attained in $[t^*, \infty)$. The assertion follows from Definition 2.1. □

If H is simple, $\mathbb{P}_H$ continuous, and φ of Neyman-Pearson type, Lemma 2.1 yields $p_\varphi(x) = 1 - F_T(t^*)$, with F_T as in Definition 2.2.

Theorem 2.1 (*p*-values as decision tools). *Let $\alpha \in (0, 1)$ a fixed given significance level and assume that $\mathbb{P}^*$ is continuous. Then we have the duality*

$$\varphi(x) = 1 \iff p_\varphi(x) < \alpha.$$

Proof. We restrict the proof to the case of tests of Neyman-Pearson type. The mapping $t \mapsto \mathbb{P}^*(T(X) > t)$ is decreasing in t. Moreover, due to the construction of c_α (see Definition 2.2), we must have $\mathbb{P}^*(T(X) > c_\alpha) \leq \alpha$ and $\mathbb{P}^*(T(X) > c) > \alpha$ for all $c < c_\alpha$. Altogether, this entails that $p_\varphi(x) < \alpha$ is equivalent to $t^* > c_\alpha$. The latter event characterizes rejection of H according to Definition 2.2. □

Remark 2.2.

(i) The advantage of p-values for testing is that they can be computed without prior specification of a significance level α. This is why all common statistics software systems implement statistical tests via the computation of p-values. However, for the purpose of decision making, pre-specification of α is inevitable.

(ii) The p-value gives an answer to the question "How probable are the observed data, given that the null hypothesis is true?". However, it does *not* answer the question "How probable is the validity of the null hypothesis, given the observed data?".

(iii) For some applications, it is more useful to consider isotone transformations of test statistics rather than antitone ones. Therefore, we remark here that $1-p_\varphi(X)$ is in the cases that are relevant for our work equal to the *distributional transform* of $T(X)$ as defined by Rüschendorf (2009). We will adopt this terminology in the remainder of this work.

Theorem 2.2. *Under the assumptions of Definition 2.1, assume that H is simple, $\mathbb{P}_H$ is continuous and φ is a test of Neyman-Pearson type. Then it follows*

$$p_\varphi(X) \underset{H}{\sim} UNI[0, 1].$$

Proof. The assertion is a consequence of the principle of quantile transformation. Making use of Lemma 2.1, we easily calculate

$$\begin{aligned}
\mathbb{P}_H(p_\varphi(X) \leq t) &= \mathbb{P}_H(1 - F_T(T(X)) \leq t) \\
&= \mathbb{P}_H(F_T(T(X)) \geq 1 - t) \\
&= \mathbb{P}(U \geq 1 - t) = 1 - \mathbb{P}(U \leq 1 - t) \\
&= 1 - (1 - t) = t,
\end{aligned}$$

where U denotes a standard uniform variate. □

Remark 2.3. In general, it holds that $p_\varphi(X)$ is under H stochastically not smaller than a standard uniform variate, i.e.,

$$\forall \vartheta \in H : \mathbb{P}_\vartheta(p_\varphi(X) \leq t) \leq t, \ t \in [0, 1]. \tag{2.2}$$

Occasionally, p-values are even defined via property (2.2) in the literature, without reference to test statistics or rejection regions at all; see, for instance, Definition 8.3.26 in the textbook by Casella and Berger (2002).

2.1 Randomized p-values

In Theorem 2.2, we assumed a simple null hypothesis H and that $\mathbb{P}_H$ is continuous. Hence, two potential sources of non-uniformity of p-values are discreteness of $\mathbb{P}_H$ and testing of composite null hypotheses. In this section, we demonstrate how randomization techniques can be used to remove or at least to diminish the conservativity that we have reported in (2.2) if the statistical model entails one of the aforementioned sources of non-uniformity of the p-values in the sense of Definition 2.1. As we will point out later, this is important for multiple testing, especially because many data-adaptive multiple tests require exactly uniformly distributed p-values under null hypotheses for a reasonable performance and fail to work properly if this assumption is violated.

2.1.1 Randomized p-values in Discrete Models

We started with non-randomized tests in Definition 2.1. Especially in discrete models, this leads to p-values that are stochastically larger than UNI[0, 1]. To meet the requirement of uniformity of the p-values under null hypotheses at least for the case of testing point hypotheses, p-values can be slightly modified in analogy to randomization of tests.

Definition 2.3 (Realized randomized p-value). Let a statistical model $(\mathscr{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ be given. Consider the two-sided test problem $H : \{\vartheta = \vartheta_0\}$ versus $K : \{\vartheta \neq \vartheta_0\}$ and assume the decision is based on the realization x of a discrete random variate $X \sim \mathbb{P}_\vartheta$ with values in $\mathscr{X}$. Moreover, let U denote a uniformly distributed random variable on [0, 1], stochastically independent of X. Then, a realized randomized p-value for testing H versus K is a measurable mapping $p^{\text{rand.}} : \mathscr{X} \times [0, 1] \to [0, 1]$ fulfilling that

$$\mathbb{P}_{\vartheta_0}(p^{\text{rand.}}(X, U) \leq t) = t \text{ for all } t \in [0, 1]. \tag{2.3}$$

The property (2.3) is an abstract mathematical requirement. For practical applications, the following theorem which is due to Klaus Straßburger makes the concept of realized randomized p-values fully usable. The proof of Theorem 2.3 can be found in Appendix II of Dickhaus et al. (2012).

Theorem 2.3. *Let $T : \mathscr{X} \to \mathbb{R}$ denote a statistic and let $f : \mathscr{X} \to \mathbb{R}_+$ be the pmf. of a discrete random variate X with values in $\mathscr{X}$, such that $f(x) > 0$ for all $x \in \mathscr{X}$. Moreover, let U denote a UNI$[0,1]$—distributed variate which is stochastically independent of X. Define*

$$\begin{aligned} p_T(x) &= \sum_{y:T(y)\leq T(x)} f(y), \quad \mathscr{W} = \{p_T(x) : x \in \mathscr{X}\}, \text{ and} \\ p_T^{\text{rand.}}(x,u) &= p_T(x) - u \sum_{y:T(y)=T(x)} f(y). \end{aligned} \tag{2.4}$$

Then it holds

$$\begin{aligned} \mathbb{P}(p_T(X) \leq t) &\leq t, \text{ for all } t \in [0,1], \\ \mathbb{P}(p_T(X) \leq t) &= t, \text{ for all } t \in \mathscr{W}, \\ \mathbb{P}(p_T^{\text{rand.}}(X,U) \leq t) &= t, \text{ for all } t \in [0,1]. \end{aligned} \tag{2.5}$$

If realized randomized p-values are constructed according to (2.4), the relationship between p-value and distributional transform given in part (iii) of Remark 2.2 remains to hold.

2.1.2 Randomized p-values for Testing Composite Null Hypotheses

Dickhaus (2013) proposed randomized p-values for testing composite null hypotheses as follows.

Definition 2.4 (Dickhaus (2013)). Let p^{LFC} be a p-value which is constructed as in Definition 2.1 and let u denote the realization of a UNI$[0,1]$—distributed random variable U which is stochastically independent of X. Then, the randomized p-value $p^{\text{rand.}}$ is given by

$$p^{\text{rand.}}(x,u) = u\mathbf{1}_H(\hat{\theta}(x)) + G(p^{\text{LFC}}(x))\mathbf{1}_K(\hat{\theta}(x)),$$

where $\theta : \Theta \to \Theta'$ denotes a one-dimensional (possibly derived) parameter, $\hat{\theta}$ a consistent and (at least asymptotically for large sample sizes) unbiased estimator of θ, and G the conditional cdf of $p^{\text{LFC}}(X)$ given $\hat{\theta} \in K$ under the (or: any) LFC for the type I error probability of the test φ of $H \subset \Theta'$ versus $K = H \setminus \Theta'$ corresponding to p^{LFC}.

At least for one-sided tests of means in Gaussian models, Dickhaus (2013) showed that these p-values are valid and under null hypotheses stochastically not larger than the traditional, LFC-based ones. Alternative methods for multiple testing of composite null hypotheses are reviewed in the introduction of Dickhaus (2013).

2.2 p-value Models

In the context of multiple test problems, (marginal) p-values $p_1, \ldots, p_m$ can be computed for every individual pair of hypotheses H_i versus K_i, if marginal models can, at least under null hypotheses, be specified exactly (which is often a hard requirement). A broad class of multiple tests depend on the data only via $p_1, \ldots, p_m$ and combine them in a suitable way in order to control errors, based on probabilistic calculations. Hence, for the mathematical analysis of such multiple tests, it suffices to model the distribution of the vector $(p_1(X), \ldots, p_m(X))^\top$ of (random) p-values and to consider statistical models of the form $([0,1]^m, \mathscr{B}([0,1]^m), (\mathbb{P}_\vartheta : \vartheta \in \Theta))$. Especially in high-dimensional settings, often only qualitative assumptions on the joint distribution of $p_1, \ldots, p_m$ (regarded as random variables) are made which lead to a variety of standard p-value models which are frequently considered in multiple hypotheses testing.

2.2.1 The iid.-Uniform Model

If one can assume that all m p-values $p_1, \ldots, p_m$ are stochastically independent and that the marginal test problems H_i versus K_i, $1 \leq i \leq m$, are such that Theorem 2.2 applies for all of them, then the joint distribution of $p_1(X), \ldots, p_m(X)$ under the global hypothesis H_0 is fully specified, because under these assumptions $p_1(X), \ldots, p_m(X)$ are under H_0 distributed as a vector $(U_1, \ldots, U_m)^\top$ of m stochastically independent, identically UNI[0, 1]—distributed random variables. Moreover, if only (without loss of generality) hypotheses $H_1, \ldots, H_{m_0}$ are true for some $m_0 = m_0(\vartheta) \in \{1, \ldots, m\}$, then $p_1(X), \ldots, p_{m_0}(X)$ are distributed as $(U_1, \ldots, U_{m_0})^\top$. We call this p-value model the iid.-uniform model.

For certain classes of multiple test procedures, the iid.-uniform model already implies the distribution of V_m and hence suffices to calibrate such multiple tests with respect to FWER control. To illustrate this, assume that the multiple test φ is such that those hypotheses are rejected for which the corresponding p-value is smaller than some given threshold $\alpha_{\text{loc.}} \in (0, 1)$. We call $\alpha_{\text{loc.}}$ a local significance level and such multiple test procedures single-step tests, cf. Sect. 3.1.1. Then, assuming the iid.-uniform model for $p_1(X), \ldots, p_m(X)$, V_m is under $\vartheta \in \Theta$ binomially distributed with parameters $m_0(\vartheta)$ and $\alpha_{\text{loc.}}$. This leads to the following expression for the FWER of φ under ϑ.

$$\mathrm{FWER}_{\vartheta}(\varphi) = \mathbb{P}_{\vartheta}(V_m > 0) = 1 - \mathbb{P}_{\vartheta}(V_m = 0) = 1 - (1 - \alpha_{\mathrm{loc.}})^{m_0}. \tag{2.6}$$

Obviously, the right-hand side of (2.6) is increasing in m_0. Therefore, under the iid.-uniform model, the FWER of a single-step test φ becomes largest for such ϑ for which $I_0(\vartheta) = I = \{1, \ldots, m\}$. In other words, all $\vartheta \in H_0$ are least favorable for the FWER of φ under the iid.-uniform model. This allows for a precise calibration of $\alpha_{\mathrm{loc.}}$ for strong FWER control (which is equivalent to weak FWER control here). The resulting single-step test is known as Šidák test and will be presented in Example 3.2. Since the full joint distribution of $(p_i(X) : i \in I_0(\vartheta))$ is completely specified in the iid.-uniform model, also the joint and the marginal distributions of the order statistics of the latter sub-vector of p-values can be derived and expressed in closed form. These distributions are important for calibrating step-up-down multiple test procedures. Such multiple tests reject hypotheses whose p-values are below a threshold which is determined data-dependently by the value of an order statistic of $(p_i(X) : 1 \leq i \leq m)$, see Sect. 3.1.2. Let us briefly recall the following facts.

Lemma 2.2. *Let $Y_1, \ldots, Y_m$ denote stochastically independent, identically distributed random variables driven by the probability measure $\mathbb{P}$, with cdf. F of Y_1. Assume that $\mathbb{P}^{Y_1}$ is absolutely continuous with respect to the Lebesgue measure λ and denote the order statistics of $(Y_1, \ldots, Y_m)^\top$ by $(Y_{1:m}, \ldots, Y_{m:m})^\top$. Then the following assertions hold true.*

$$\mathbb{P}(Y_{i:m} \leq y) = \sum_{j=i}^{m} \binom{m}{j} F(y)^j (1 - F(y))^{m-j},$$

$$\frac{d\mathbb{P}^{Y_{i:m}}}{d\mathbb{P}^{Y_1}}(y) = m \binom{m-1}{i-1} F(y)^{i-1} (1 - F(y))^{m-i}.$$

If $\mathbb{P}^{Y_1}$ has Lebesgue density f, then $\mathbb{P}^{Y_{i:m}}$ has Lebesgue density $f_{i:m}$, given by

$$f_{i:m}(y) = m \binom{m-1}{i-1} F(y)^{i-1} (1 - F(y))^{m-i} f(y). \tag{2.7}$$

Letting $\mu = \mathbb{P}^{Y_1}$, $(Y_{i:m})_{1 \leq i \leq m}$ has joint μ^m-density

$$(y_1, \ldots, y_m) \mapsto m! \mathbf{1}_{\{y_1 < y_2 < \cdots < y_m\}}.$$

If μ has Lebesgue density f, then $(Y_{i:m})_{1 \leq i \leq m}$ has λ^m-density

$$(y_1, \ldots, y_m) \mapsto m! \prod_{i=1}^{m} f(y_i) \mathbf{1}_{\{y_1 < y_2 < \cdots < y_m\}}.$$

Remark 2.4. Considering iid. UNI[0, 1]—distributed random variables $U_1, \ldots, U_m$ in Lemma 2.2, Eq. (2.7) shows that the order statistic $U_{i:m}$ has a Beta$(i, m - i + 1)$ distribution with

$$\mathbb{E}[U_{i:m}] = \frac{i}{m+1}, \quad \mathrm{Var}(U_{i:m}) = \frac{i(m-i+1)}{(m+1)^2(m+2)}.$$

For computing the joint cumulative distribution function of $(U_{1:m}, \ldots, U_{m:m})$, efficient recursive algorithms exist, for instance Bolshev's recursion and Steck's recursion (see Shorack and Wellner (1986), p. 362 ff.).

Lemma 2.2 can be used to calibrate a step-up-down multiple test procedure φ for weak FWER control under the assumption of an iid.-uniform model for the p-values $p_1(X), \ldots, p_m(X)$. However, if $\vartheta \notin H_0$, the FWER of φ typically depends on the distribution of $(p_j(X) : j \in I_1(\vartheta))$, too. The same holds true for the FDR of φ, because the distribution of R_m certainly relies on that of $(p_j(X) : j \in I_1(\vartheta))$. This shows that some assumptions on the p-value distribution under alternatives are also needed to study the behavior of multiple tests operating on p-values, even for the sole purpose of type I error rate control according to Definition 1.2. A generalization of Steck's recursion to two populations has been derived by Blanchard et al. (2014). This generalization can for instance be used for calibrating multiple tests for FDR control if a fixed alternative p-value distribution is assumed and all m p-values are stochastically independent.

2.2.2 Dirac-Uniform Configurations

It seems that the term "Dirac-uniform configuration" was used for the first time by Finner and Roters (2001). A Dirac-uniform configuration is characterized by three distributional assumptions regarding the joint distribution of $(p_1, \ldots, p_m)$.

Definition 2.5 (Dirac-uniform configuration). The value of the parameter ϑ is called a Dirac-uniform configuration if the following three distributional properties hold.

1. All m_0 marginal p-values corresponding to true null hypotheses are stochastically independent and identically distributed as UNI[0, 1].
2. The random vector $(p_i(X) : i \in I_0(\vartheta))$ is stochastically independent of the random vector $(p_j(X) : j \in I_1(\vartheta))$.
3. For all $j \in I_1$, $p_j(X)$ follows a Dirac distribution with point mass 1 in zero, meaning that p_j is almost surely equal to zero.

Of course, in practice it is unrealistic to assume that effect sizes are so large that p-values are almost surely equal to zero under alternatives. Therefore, Dirac-uniform configurations are not useful for modeling real-life data. This is why we do not term them "models". They are technical devices for deriving upper bounds for the FWER or the FDR of multiple testing procedures. For the mathematical analysis of multiple tests under independence assumptions, Dirac-uniform configurations are important tools, because they are, for fixed m_0, often LFCs for the FWER and/or the FDR of multiple tests if p-values are independent. In particular, if φ is a step-up-down test, its

FWER and FDR typically become maximum if parameter values under alternatives are extreme in the sense that p-values under alternatives are as small as possible. As a consequence, control of the respective error rate by φ under Dirac-uniform configurations (which are LFCs) entails that φ controls the error rate also under all other (often more realistic) values of the parameter ϑ of the model.

Furthermore, analytic calculations for the FWER and the FDR are very straightforwardly possible under Dirac-uniform configurations, because it holds (almost surely) that $R_m = V_m + m_1$ for a stepwise rejective multiple test φ, if ϑ is a Dirac-uniform configuration. This is because the m_1 null hypotheses with indices in I_1 are almost surely rejected by φ due to their p-values which are almost surely equal to zero. Consequently, the joint distribution of V_m and R_m is already determined by that of V_m which in turn can be expressed in terms of the joint distribution of order statistics of m_0 iid. UNI[0, 1]—distributed random variables, and the respective Bolshev's or Steck's recursions suffice for the type I error calibration of φ. The latter reasoning will play an important role in Chap. 5 where we will provide more details.

2.2.3 Two-Class Mixture Models

In contrast to the models discussed before, two-class mixture models are often used as models for real-life data. They still have a tractable structure.

Definition 2.6. The joint distribution of $(p_1, \ldots, p_m)$ is called a two-class mixture model, if the following two properties hold.

1. All m_0 p-values corresponding to true null hypotheses are marginally distributed with cdf. F_0, where F_0 is stochastically lower-bounded by UNI[0, 1].
2. All m_1 p-values corresponding to false null hypotheses are marginally distributed with cdf. F_1.

At a first glance, this model seems very restrictive, because all p-values under null hypotheses share the same marginal distribution and the same holds true for all p-values under alternatives. However, the following trick, mentioned for instance by Genovese and Wasserman (2004) and Farcomeni (2007), considerably extends the applicability of two-class mixture models: Even if it can not be assumed that all $p_i(X)$ with $i \in I_1$ share the same marginal distribution, we may at least assume that their marginal cdfs all belong to a class $\{F_\xi : \xi \in \Xi\}$. In addition, we may be able to put a prior distribution ν on Ξ. If these two requirements are fulfilled, let F_1 be defined by $F_1(t) = \int_\Xi F_\xi(t)\nu(d\xi)$. In an analogous manner, one can proceed for constructing the marginal distribution function F_0 under null hypotheses, if the multiple test problem does not already imply a fixed marginal distribution of p-values under null hypotheses, for instance UNI[0, 1]. However, let us mention here that putting a prior on the parameter space under null hypotheses is problematic from the classical (frequentist) viewpoint toward statistics.

Notice that Definition 2.6 only specifies the marginal distributions of p-values. As far as the dependency structure in two-class mixture models is concerned, one often assumes weak dependency in the sense of Definition 5.2, meaning that the ecdfs of $(p_i : i \in I_0)$ and $(p_j : j \in I_1)$ converge for $m \to \infty$ to F_0 and F_1, respectively, in the Glivenko-Cantelli sense. This gives enough structure to the statistical model for an asymptotic analysis of the behavior of multiple tests operating on such p-values.

2.2.4 Copula Models Under Fixed Margins

The following well-known theorem provides a convenient way to separate the models for the marginal distributions of $p_1(X), \ldots, p_m(X)$ (which are, at least under null hypotheses, often already implied by the test problems H_i versus K_i, $1 \leq i \leq m$, see Theorem 2.2) from a model regarding the dependency structure among the p-values.

Theorem 2.4 (Sklar (1959, 1996)). *Let $Y = (Y_1, \ldots, Y_m)^\top$ denote a random vector with values in $\mathbb{R}^m$ and with joint cdf F_Y and marginal cdfs $F_{Y_1}, \ldots, F_{Y_m}$. Then there exists a function $C : [0,1]^m \to [0,1]$, called the copula of Y, such that for all $y = (y_1, \ldots, y_m)^\top \in \bar{\mathbb{R}}^m$, it holds*

$$F_Y(y) = C(F_{Y_1}(y_1), \ldots, F_{Y_m}(y_m)).$$

If all m marginal cdfs are continuous, then the copula C is unique.

According to Theorem 2.4, the dependency structure among $p_1(X), \ldots, p_m(X)$ can be modeled by modeling their copula. Furthermore, if Theorem 2.2 applies for all marginal test problems H_i versus K_i, $1 \leq i \leq m$, the copula of the p-values coincides under the global hypothesis H_0 with the cdf of $p_1(X), \ldots, p_m(X)$. The latter fact is extremely useful for constructing simultaneous test procedures based on p-values, cf. Sect. 4.4. In particular, parametric copula models can be used as regularized models for the dependency structure of $p_1(X), \ldots, p_m(X)$, especially in cases where m is large such that the "curse of dimensionality" prohibits modeling or reliably estimating the full joint distribution of the data or the p-values, respectively. Regularization here means that the copula parameter is of low dimension. Of course, in practice this will typically only yield an approximation of the true dependency structure.

2.2.5 Further Joint Models

Assume that all marginal tests φ_i, $1 \leq i \leq m$, for a given multiple test problem $(\mathscr{X}, \mathscr{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta}, \mathscr{H}_m)$ are of (generalized) Neyman-Pearson type in the sense of Definition 2.2, with (marginal) test statistics $T_1, \ldots, T_m$. Then, in order to calibrate

the multiple test φ by multivariate techniques, it is often convenient to consider the joint distribution of $T_1, \ldots, T_m$ directly. On the p-value scale, however, the resulting adjustment for multiplicity of the overall significance level α (for the FWER or the FDR) is explicitly given, leading to a better interpretability of φ. Therefore, it may be of interest to derive the joint distribution of the p-values $p_1, \ldots, p_m$ corresponding to $T_1, \ldots, T_m$ by transformation of measures. To give a specific example, assume that $X = (X_1, \ldots, X_m)^\top$ follows a multivariate normal distribution with mean $\mu = (\mu_1, \ldots, \mu_m)^\top$ and covariance matrix Σ. For ease of exposition and without loss of generality, assume that all diagonal elements of Σ are equal to one. Furthermore, assume that the m null hypotheses $H_i : \{\mu_i = 0\}$ with two-sided alternatives $K_i : \{\mu_i \neq 0\}$ are of interest, $1 \leq i \leq m$. Suitable test statistics are given by $T_i = |X_i|$, $1 \leq i \leq m$. Following Lemma 2.1 and utilizing symmetry properties of the standard normal law, the marginal p-values corresponding to the test statistics $T_1, \ldots, T_m$ are given by

$$p_i(x) = 2(1 - \Phi(T_i(x))), \quad 1 \leq i \leq m, \tag{2.8}$$

where Φ denotes the cdf. of the standard normal distribution.

Hence, if the calibration of φ results in a threshold c_α for the T_i, then equivalently φ_i rejects H_i if $p_i(x) < 2(1 - \Phi(c_\alpha)) = \alpha_{\text{loc.}}$ (say). The value $\alpha_{\text{loc.}}$ can thus be regarded as a multiplicity-adjusted local significance level.

Under H_i, $p_i(X)$ is marginally UNI[0, 1]—distributed, see Theorem 2.2. Moreover, the joint cdf of $(p_i(X) : 1 \leq i \leq m)$ under μ and Σ is given by

$$\begin{aligned} u = (u_1, \ldots, u_m)^\top \in [0,1]^m \mapsto\ & \mathbb{P}_{(\mu,\Sigma)}(p_i(X) \leq u_1, \ldots, p_m(X) \leq u_m) \\ =\ & \mathbb{P}_{(\mu,\Sigma)}(\forall 1 \leq i \leq m : T_i \geq \Phi^{-1}(1 - u_i/2)). \end{aligned}$$

The latter probability can easily be computed by employing numerical routines for multivariate normal distributions, cf. Genz and Bretz (2009). Multiple tests for Gaussian means play an important role in many practical applications, for instance in the context of localized comparisons in analysis of variance models. Applications in genetics are discussed in Chaps. 9 and 10.

Acknowledgments Parts of Sect. 2.1 originated from joint work with Klaus Straßburger. I am grateful to Mette Langaas and Øyvind Bakke for inviting me and for their hospitality during my visit to Norwegian University of Science and Technology (NTNU), for many fruitful discussions and for critical reading.

References

Blanchard G, Dickhaus T, Roquain E, Villers F (2014) On least favorable configurations for step-up-down tests. Statistica Sinica 24(1):1–23

Casella G, Berger RL (2002) Statistical inference, 2nd edn. Brooks/Cole, Cengage Learning

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

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

Farcomeni A (2007) Some results on the control of the false discovery rate under dependence. Scand J Stat 34(2):275–297. doi:10.1111/j.1467-9469.2006.00530.x

Finner H, Roters M (2001) On the false discovery rate and expected type I errors. Biom J 43(8):985–1005

Genovese C, Wasserman L (2004) A stochastic process approach to false discovery control. Ann Stat 32(3):1035–1061. doi:10.1214/009053604000000283

Genz A, Bretz F (2009) Computation of multivariate normal and t probabilities. Lect Notes Stat 195. Springer, Berlin. doi:10.1007/978-3-642-01689-9

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

Rüschendorf L (2009) On the distributional transform, Sklar's theorem, and the empirical copula process. J Stat Plann Inference 139(11):3921–3927. doi:10.1016/j.jspi.2009.05.030

Shorack GR, Wellner JA (1986) Empirical processes with applications to statistics. Wiley Series in Probability and Mathematical Statistics. Wiley, New York

Sklar A (1959) Fonctions de répartition à n dimensions et leurs marges. Publ Inst Stat Univ Paris 8:229–231

Sklar A (1996) Random variables, distribution functions, and copulas—a personal look backward and forward. In: Distributions with fixed marginals and related topics, IMS Lecture Notes-Monograph Series, Volume 28, Institute of Mathematical Statistics, Hayward, CA, pp 1–4