mathematics

MDPI

Article

Parity-Based Statistics and Combinatorial Identities

Andrew L. Rukhin

Department of Mathematics and Statistics, University of Maryland at Baltimore County, Baltimore, MD 21250, USA; arukhin@umbc.edu

Abstract

Notable discrete probability laws appear as posterior distributions in the estimation of the common mean with heterogeneous variances. These probabilities, which are defined by an arbitrary set of distinct real numbers, also arise in seemingly unrelated areas of polynomial approximation and statistical physics. The corresponding orthogonal polynomials possess an interesting self-duality property, which invites the study of statistical distributions based on data parity. These distributions provide novel, intriguing formulas for the classical hypergeometric function along with several combinatorial identities.

Keywords: Dirichlet distribution; hypergeometric function; lagrange interpolation formula; parity-based sum; self-dual orthogonal polynomials

MSC: 05A99; 33C05; 41A10; 60C99; 62F15

1. Introduction: Common Mean and Unknown Heterogeneous Uncertainties

This paper arose from statistical analysis of a sequence, $x_1, \ldots, x_n$, in the setting of heterogeneous research synthesis, with x_j representing the estimate of the common mean (say, treatment effect), as reported by the j-th study. No conditions are imposed on the unknown accuracies, which cannot be assumed equal. The main statistical challenge is the estimation of the common mean, treated as a shift parameter, when the standard deviations are considered to be unknown nuisance scale parameters.

In some applications, the uncertainty appraisals are either missing or utterly unreliable. The difficulty in accurately valuing the variances of systematic errors, whether due to specific laboratory conditions or hospital protocols, is well acknowledged by data scientists.

The issue of underreported uncertainties, particularly those that stem from asymptotic normal theory, which presupposes large data sets, is prevalent in metrology. Furthermore, the challenge of reproducibility within individual centers may be exacerbated by the nature of the employed measuring instruments.

Another source of artificially small uncertainties may be due to the removal of outliers for purely mathematical reasons. By eliminating "unrepresentative" or "spurious" data points, one typically is left with a part of the sample that is unrealistically accurate. See [1] for further motivation.

Unlike classical statistical models, the scenario suggested here does not require accompanying estimates of uncertainty of x_j. Our investigation focuses on the special "self-dual" weights that define the discrete posterior distribution for the unknown mean, set against a non-informative objective prior for both the mean and independent variances.

This line of inquiry, initiated in [2] under the assumption of normality, grapples with the lack of variance information, causing several statistical complications. For instance, the

Check for updates

Academic Editor: Niansheng Tang

Received: 26 April 2026
Revised: 11 June 2026
Accepted: 21 June 2026
Published: 5 July 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

classical maximum likelihood estimator cannot be determined uniquely, as the likelihood function reaches infinity at each data point. Nevertheless, the problem is well-defined. Indeed, estimating the common mean requires determining at most n parameters, the mean itself and, say, $\omega_i = \sigma_i^{-2} / \sum_j \sigma_j^{-2}$, which belong to the unit simplex of dimension $n-1$. Statistical practice needs to determine only $(n-1)$ weights to form the common mean estimator, say, $\sum_{i=1}^{n} \omega x_i$. In some applications, there is additional information that allows further dimension reduction.

This paper is motivated by mathematical statistics, but its goal is to explore the mathematical aspects of the issues arising in the statistical problem. Indeed, it is addressed to a general mathematical audience. The main contribution is the construction of a unified probabilistic framework for rank-induced parity distributions and the derivation of related moment and combinatorial formulas, which show the link to the Gauss hypergeometric function.

More specifically, Section 2 examines the polynomial approximation problem over the set $\{x_1, \ldots, x_n\}$, and orthogonal polynomials, which are less deviant from zero on this set. The required formulas for the distributions of parity-based sums are established in Section 3. The deep connection between these distributions and the hypergeometric function is demonstrated in Section 4. We present a self-contained proof of formulas for the specific value of this classical function. Obviously, these formulas are known to specialists, but the author failed to find an easy reference. Our approach yields seemingly new combinatorial identities, (39), (48), (53)–(55), (57) in Sections 4 and 5. Some useful expressions for partial derivatives are given in Section 6.

2. Self-Dual Probabilities and Orthogonal Polynomials Least Deviating from Zero

We initiate our discussion with a problem that, at first glance, appears unrelated to the main focus. Namely, assuming that all x's are distinct, the best polynomial approximation of a function f over a finite set $\{x_1, \ldots, x_n\}$ is sought.

The optimal uniform approximation by a polynomial of degree $n-1$ (or higher) is achieved through the classical Lagrange interpolation polynomial $L(x)$, which coincides with f on this set, i.e., $L(x_j) = f_j = f(x_j)$ for $j = 1, \ldots, n$. If the polynomial's degree is $n-2$, the best approximation is derived by subtracting from $L(x)$ a specific multiple of the oscillating polynomial that alternates its sign at each successive x_j, attaining the same absolute value at these points.

The probabilities,

$$w_i = \frac{1}{\prod_{j \neq i} |x_i - x_j|} \left[\sum_k \prod_{j \neq k} \frac{1}{|x_k - x_j|} \right]^{-1} \tag{1}$$

$$= \frac{s_i}{\prod_{j \neq i} (x_i - x_j)} \left[\sum_k s_k \prod_{j \neq k} \frac{1}{(x_k - x_)|} \right]^{-1},$$

are related to the Lagrange formula and to the *parities* of x_i, $i = 1, \ldots, n$,

$$s_i = \operatorname{sign} \left(\prod_{j \neq i} (x_i - x_j) \right). \tag{2}$$

It is known (see Theorem 1.15, [3]) that the approximation error coincides with the absolute value of the average, taken under (1), of products $s_i f_i$. Thus,

$$|\sum_i s_i f_i w_i| = \min_R \max_i |f_i - R(x_i)|,$$

where $R(x)$ runs through all polynomials of degree not exceeding $n-2$. For instance, the approximation error of the oscillating function $f_i = s_i$ is independent of n,

$$1 = \min_R \max_i |s_i - R(x_i)|.$$

The barycentric form of the Lagrange interpolation formula,

$$L(x) = \frac{\sum_i w_i s_i f_i/(x-x_i)}{\sum_i w_i s_i/(x-x_i)},$$

provides numerous advantages [4].

Probabilities (1) originate in random matrix theory, where they offer alternative descriptions of a physical ensemble in terms of particles or holes. Many optimization problems involving the discriminant function through electrostatic equilibrium are underpinned by them [5]. From the mathematical perspective, these probabilities are self-dual under the duality definition given in [6] and developed in [7–9].

In mathematical statistics, probabilities (1) define the discrete posterior distribution for the location parameter against a non-informative prior for this parameter, mean, and independent variances [10]. The generalized Bayes estimator of the mean under the quadratic loss is

$$\delta = \sum_i w_i x_i = \sum_i \frac{x_i}{\prod_{j\neq i}|x_i - x_j|} \left[\sum_i \prod_{j\neq i} \frac{1}{|x_i - x_j|}\right]^{-1}. \tag{3}$$

This statistic serves as a semiparametric estimator of the symmetry center in a heterogeneous sample, meaning that it does not depend on the distribution of x's from a broad class.

Orthogonal polynomials against (1) exhibit striking symmetry. To see it, let $\delta_m = \sum_k w_k x_k^m, m = 0, 1, \ldots$ represent the moments of self-dual weights (1). The monic polynomials $T_m, m = -1, 0, 1, \ldots, n$, $T_{-1}(z) = 0, T_0(z) = 1, T_1(z) = z - \delta_1, \ldots, T_n(z) = \prod(z - x_k)$, are orthogonal in L_2,

$$\sum_i w_i T_m(x_i) T_p(x_i) = \delta_{mp}\sqrt{h_m h_p}. \tag{4}$$

With

$$W = \left[\sum_i \prod_{j\neq i} |x_i - x_j|^{-1}\right]^{-1},$$

the sequence $h_m = \sum_i [T_m(x_i)]^2 w_i, h_0 = 1, h_1 = \delta_2 - \delta_1^2, \ldots, h_{n-1} = W^2$, enjoys the mentioned symmetry as $h_m h_{n-m-1} = W^2$.

The orthogonal polynomials T_m are known to satisfy the three-term recurrence,

$$T_{m+1}(z) = (z - \alpha_m) T_m(z) - \beta_m T_{m-1}(z), \tag{5}$$

where the coefficients $\beta_m = h_m/h_{m-1}, \beta_m = \beta_{n-m}$, and $\alpha_m = \sum_i x_i [T_m(x_i)]^2 w_i / h_m$, also possess symmetry property: $\alpha_m = \alpha_{n-1-m}$.

Mathematical induction applied to (5) shows that for $i, m = 1, \ldots, n$,

$$W T_m(x_i) = s_i h_m T_{n-m-1}(x_i). \tag{6}$$

Identity (6) can be obtained from the original duality definition [6], according to which $T_m(x_i)$ is a multiple of $T_{n-1}(x_i)T_{n-m-1}(x_i)$.

The polynomial,

$$T_{n-1}(z) = \sum_i w_i \prod_{j\neq i}(z - x_j),$$

is less deviant from zero in L_∞: $T_{n-1}(x_i) = s_i W$, i.e., for any monic polynomial R of degree not exceeding $n-1$,

$$\max_i |T_{n-1}(x_i)| = W \leq \max_i |R(x_i)|. \tag{7}$$

The comparison of the extremes of $T_{n-1}(z)$ and those of the classical monic, degree $n-1$, Chebyshev polynomial on the interval, $[\min x_i, \max x_i]$, provides a sharp inequality,

$$W \leq \frac{(\max x_i - \min x_i)^{n-1}}{2^{n-2}} = 2\left(\frac{\max x_i - \min x_i}{2}\right)^{n-1}, \tag{8}$$

which holds for all distinct $x_1, \ldots, x_n$. The factor 2^{2-n} in (8) is given incorrectly in [10].

Equality in (8) is attained if and only if x's are extreme points of the mentioned Chebyshev polynomial on this interval, i.e., when for $j = 1, \ldots, n$,

$$x_j = \frac{\max x_i + \min x_i}{2} + \frac{(\max x_i - \min x_i)}{2} \cos\left(\frac{(j-1)\pi}{n-1}\right).$$

Then the probabilities (1) have a remarkably simple form,

$$w_j = \frac{1}{n-1},\ j = 2, \ldots, n-1, \quad w_1 = w_n = \frac{1}{2(n-1)}.$$

Since $w_1 < w_2, w_n < w_{n-1}$, this provides the closest resemblance of (1) to the uniform distribution.

In addition to $x_1, \ldots, x_n$, the polynomial $T_{n-1}^2(z) - W^2$ has $n-2$ real roots (which interlace those of T_{n-1}), so that, with R_{n-1} denoting the monic polynomial of degree $n-2$ having these roots,

$$T_{n-1}^2(z) - W^2 = T_n(z) R_{n-1}(z). \tag{9}$$

Therefore, R_{n-1} is the associated polynomial to T_{n-1},

$$R_{n-1}(z) = \sum_i \frac{[T_{n-1}(z) - T_{n-1}(x_i)]w_i}{z - x_i}. \tag{10}$$

The associated with T_m orthogonal monic polynomial R_m of degree $m-1, m = 1, \ldots, n-1$, satisfies the same recurrence (5), but the initial conditions are different: $R_0 = 0, R_1 = 1$, so that $R_2(z) = z - \alpha_1, \ldots, R_n(z) = T_{n-1}(z) = (z - \delta_1)R_{n-1}(z) - \beta_1 R_{n-2}(z)$.

The coefficients $\alpha_{n/2}, \beta_{n/2}$ admit an explicit form: for even n

$$\alpha_{(n-2)/2} = \alpha_{n/2} = \frac{\sum_i s_i x_i^2}{2\sum_i s_i x_i},\ \ \beta_{n/2} = \frac{(\sum_i s_i x_i)^2}{4}. \tag{11}$$

If n is odd, then

$$T_{(n-1)/2}(z) = P_o(z) = \prod_{j: s_j = -1} (z - x_j), \tag{12}$$

$h_{(n-1)/2} = W$, and

$$\alpha_{(n-1)/2} = \sum_i s_i x_i,\ \beta_{(n-1)/2} = \beta_{(n+1)/2} = \frac{\sum_i s_i x_i^2 - (\sum_i s_i x_i)^2}{4}. \tag{13}$$

With $P_e(z) = \prod_{j:s_j=1}(z - x_j)$, other central orthogonal polynomials are of the form

$$T_{n/2}(z) = \frac{P_o(z) + P_e(z)}{2}, \tag{14}$$

$$T_{(n+1)/2}(z) = \frac{(z - \sum_i s_i x_i)P_o(z) + P_e(z)}{2}, \tag{15}$$

and

$$T_{(n-2)/2}(z) = \frac{P_o(z) - P_e(z)}{\sum_i s_i x_i}. \tag{16}$$

One can represent $R_{n-1}(z) = R_e(z)R_o(z)$ as a product of two monic polynomials (with real roots) of degrees $n_e - 1$ and $n_o - 1$, respectively. Then with P_e and $T_{n-1}(z) - W = P_e(z)R_o(z)$ and $T_{n-1}(z) + W = P_o(z)R_e(z)$. If $s_i = 1$, $R_e(x_i) = 2W/P_o(x_i)$, when $s_j = -1$, $R_o(x_j) = -2W/P_e(x_j)$. Thus,

$$R_e(z) = 2W \sum_{s_i=1} \frac{\prod_{s_k=1,k\neq i}(z - x_k)}{P_o(x_i)P_e'(x_i)} = 2 \sum_{s_i=1} w_i \prod_{s_k=1,k\neq i} (z - x_k), \tag{17}$$

and

$$R_o(z) = 2 \sum_{s_j=-1} w_j \prod_{s_\ell=-1,\ell\neq j} (z - x_\ell). \tag{18}$$

Central associated polynomials are $R_{(n-1)/2} = R_o$,

$$R_{n/2}(z) = \frac{R_o(z) + R_e(z)}{2}, \tag{19}$$

$$R_{(n+1)/2}(z) = \frac{(z - \sum_i s_i x_i)R_o(z) + R_e(z)}{2}, \tag{20}$$

and

$$R_{(n-2)/2}(z) = \frac{R_o(z) - R_e(z)}{\sum_i s_i x_i}. \tag{21}$$

We summarize the main results as a theorem whose detailed proof can be found in in [10].

Theorem 1. *Polynomials T_m, which are orthogonal with regard to probabilities (1), satisfy (5) and (6), with their central versions in (12), (14)–(16). Their associate polynomials satisfy (9) and (10), with the central versions given in (19)–(21). Central coefficients are given in (11) and (13). Inequality (8) is valid.*

Coefficients $\alpha_{(n-1)/2}, \beta_{n/2}$ are completely determined by $\sum_i s_i x_i$. Two other coefficients involve quadratic forms involving $s_i x_i$. We refer to these functions of observations as *parity-based statistics* and embark on their study.

3. Parity-Based Distributions

The main object of interest in this section is the parity-based sums of the form, $\sum_i s_i x_i$.

The finite set $\{x_1, \ldots, x_n\}$ in Section 2 can be considered as the representative points of univariate statistical distribution [11]. Thus, we assume that it is a realization of n independent random variables with common continuous distribution function $F(x) = F_0(x)$, whose density $f(x)$ has all finite moments, $\mathbf{m}_p = Ex^p = \int x^p f(x)\, dx$, which determine F uniquely.

Denote by $s_1, \ldots, s_n$ the parity sequence corresponding to $x_1, \ldots, x_n$. To start exploring the behavior of the parity-based sums, notice that the distribution of a random parity s_i can be written as

$$P(s_i = \epsilon) = \frac{1}{2n} \sum_{r=1}^{n} [1 + \epsilon(-1)^{r+n}] = \frac{1}{2} \left[1 + \frac{\epsilon[1 - (-1)^n]}{2n} \right], \tag{22}$$

$\epsilon = \pm 1$, which does not depend on F.

The joint density of x_i and its parity s_i is

$$f(x_i, s_i = \epsilon) = \frac{f(x_i)}{2} \Big\{ 1 + \epsilon[2F(x_i) - 1]^{n-1} \Big\}. \tag{23}$$

Therefore, for $i = 1, \ldots, n$,

$$P(x_i \le x, s_i = \epsilon) = \frac{F(x)}{2} + \frac{\epsilon\{[2F(x) - 1]^n - (-1)^n\}}{4n}.$$

If f is symmetric, the distribution function of $y_i = s_i x_i$ is

$$\begin{aligned} G(y) &= P(x_i < y, s_i = 1) + P(x_i > -y, s_i = -1) \\ &= F(y) + \frac{[1 + (-1)^n]\{[2F(y) - 1]^n - 1\}}{4n}. \end{aligned} \tag{24}$$

These formulas can be derived from the distribution of order statistics whose rank has the same parity as the largest observation. For example, the conditional density of x_i and x_j for given $s_i = \epsilon_1, s_j = \epsilon_2$, has the form

$$\begin{aligned} f(x_i, x_j | s_i = \epsilon_1, s_j = \epsilon_2) = \frac{f(x_i) f(x_j)}{4P(s_i = \epsilon_1, s_j = \epsilon_2)} \Big\{ 1 + \epsilon(-1)^{n+r_1} [1 - 2F(x_i)]^{n-2} \\ + \epsilon_2 (-1)^{n+r_2} [1 - 2F(x_j)]^{n-2} + \epsilon_1 \epsilon_2 (-1)^{r_1 + r_2} [1 - 2|F(x_i) - F(x_j)|]^{n-2} \Big\}, \end{aligned} \tag{25}$$

where r_1 and r_2, $1 \le r_1 \ne r_2 \le 2$, are relative ranks for x_i and x_j, say, $r_1 = 1$ if $x_i < x_j$.

The next result provides the joint density of m-sub-vectors $x_{n_1}, \ldots, x_{n_m}$ and $s_{n_1}, \cdots, s_{n_m}$, the relevant conditional distributions, as well as the form of the moments of parity sums. Here, m is a fixed integer, $1 \le m \le n$. Thus, $M = \{1, \ldots, m\} \subset N = \{1, \ldots, n\}$.

Theorem 2. *The exchangeable distribution of $x_{n_1}, \ldots, x_{n_m}$ and $s_{n_1}, \cdots, s_{n_m}$, is provided by (28). The conditional density of $x_{n_j}, j \in M$, for a given value of the product $s_{n_1} \cdots s_{n_m}$ satisfies (33). The moments of the parity sum, $\sum_{i=1}^{n} s_i x_i$, can be found from (37).*

Proof. Let $z_1 < \ldots < z_m$ be the order statistics corresponding to $x_{n_j}, j \in M$. The joint distribution of $z_1, \ldots, z_m$ and the parities can be represented as a mixture of the conditional densities for given ranks. A particular density enters this mixture if and only if $s_{n_j} = (-1)^{n+R_j}$, where $R_j, 1 \le R_j \le n$, is the rank of x_{n_j} among the total sample.

Let r_i denote the rank of x_{n_i} within our subsample. Then $x_{n_j} = z_{r_j}$, and $s_j^M = (-1)^{m+r_j}$ is the parity of x_{nj} in this subsample.

Since the probability of any rank combination is $\binom{n}{m}^{-1}$, the classical formula for the distribution of several order statistics [12] implies that the joint density of $y_1, \ldots, y_m$ and the corresponding parities is

$$f(z_1, \ldots, z_m; s_{n_1}, \ldots, s_{n_m}) = m! \prod_{1}^{m} f(z_j) \tag{26}$$

$$\times \sum_{1\le p_1<\ldots<p_m\le n} \begin{pmatrix} n-m \\ p_1-1 \quad p_2-p_1-1 \quad \cdots \quad n-p_m \end{pmatrix} \prod_{j=1}^{m} \frac{[1+s_{n_j}(-1)^{n+p_j}]}{2}$$
$$\times \Delta_1^{p_1-1}\Delta_2^{p_2-p_1-1}\cdots\Delta_m^{p_m-p_{m-1}-1}\Delta_{m+1}^{n-p_m}.$$

Here, under the convention that $F(z_{m+1})=1, z(y_0)=0, \Delta_j = F(y_j)-F(y_{j-1})$ are familiar $m+1$ spacings, $\sum_{j=1}^{m+1}\Delta_j = 1$, which are known to have a Dirichlet distribution Dir_{m+1} with positive concentration parameters $r_1, r_2-r_1, \ldots, r_m-r_{m-1}, n+1-r_m$ [13]. Therefore,

$$\int\cdots\int_{z_1<\cdots<z_m} \prod_{j=1}^{m+1}[F(z_j)-F(z_{j-1})]^{r_j-r_{j-1}-1}\prod_{j=1}^{m} f(z_j)\,dz_j$$
$$= \frac{\Gamma(r_1)\Gamma(r_2-r_1)\cdots\Gamma(r_m-r_{m-1})\Gamma(n+1-r_m)}{m!\Gamma(n+1)}.$$

Integration of (26) over $z_1,\ldots,z_m$ gives our first combinatorial identity,

$$2^m\binom{n}{m}P(s_{n_j}=\epsilon_j, j=1,\ldots,m) = \sum_{1\le p_1<\ldots<p_m\le n}\prod_{j=1}^{m}[1+\epsilon_j(-1)^{n+p_j}]. \tag{27}$$

By replacing the summation variables in (26) with $p_1-1, p_2-p_1-1, \cdots, n-p_m$ and using multinomial theorem, we arrive at the form of the joint distribution of $x_{n_1},\ldots,x_{n_m}$ and $s_{n_1},\cdots,s_{n_m}$,

$$f(x_{n_1},\ldots,x_{n_m};s_{n_1},\ldots,s_{n_m}) \tag{28}$$
$$= \frac{\prod_j f(x_{n_j})}{2^m}\sum_{k=0}^{m}(-1)^{nk}\sum_{K\subset M,|K|=k}\left[\prod_{i\in K}(-1)^{r_i}[1-2\sum_{i\in K}s_i^K F(x_{n_i})]^{n-m}\right]\prod_{i\in K}s_{n_i}$$
$$= \frac{\prod_j f(x_{n_j})}{2^m}\sum_{k=0}^{m}(-1)^{(n-m)k}\sum_{K\subset M,|K|=k}\prod_{i\in K}s_{n_i}s_i^M[1-2\sum_{i\in K}s_i^K F(x_{n_i})]^{n-m}.$$

Here, for $i\in K$, $s_i^K = \mathrm{sign}(\prod_{j\neq i, j\in K}(x_{n_i}-x_{n_j})$, is the parity of x_{n_i} in the subsample $x_{n_j}, j\in K$, where k is the cardinality of K. Thus, this joint density is a linear function of parity products, $\prod_{i\in K}s_{n_i}, K\subset M$.

The joint Dirichlet distribution of spacings implies that $\sum_{j\in K}s_j^K F(x_j)$ is beta-distributed with parameters $(q, m+1-q), q=q_k$, where for any subset K of M,

$$q_k = j_k - j_{k-1} + \ldots - (-1)^k j_1 = \sum_{i\in K}(-1)^{j_i+k}j_i,\ q_k\ge 0 = q_0. \tag{29}$$

Clearly, $(-1)^{q_k} = (-1)^{\sum_{i\in K} i} = (-1)^{\sum_{j\in \mathrm{rK}} r_j}$. If k is even, $k/2\le q_k\le m-k/2$; for odd k, $(k+1)/2\le q_k\le m-(k-1)/2$, so that

$$q_m = \sum_{j=1}^{m}(-1)^{j+m} = \lceil m/2\rceil = m/2 + [1-(-1)^m]/4, \tag{30}$$

and $(-1)^{q_m} = (-1)^{m(m+1)/2}$.

Thus, density (28) is related to the classical hypergeometric function, ${}_2F_1(m-n, q; m+1; z) = F(m-n, q; m+1; z)$ (actually a polynomial in z of degree $n-m$). Our argument shows the special role of the specific value $z=2$, which appears as the factor at $\sum s_j^K F(x_j)$ in (28).

Indeed, by using the fundamental integral representation of the hypergeometric function [14], and setting $q = q_k$, one obtains the following expression:

$$E[1 - 2\sum_{j\in K} s_j^K F(x_j)]^{n-m} = \frac{\int_0^1 (1-2u)^{n-m} u^{q-1}(1-u)^{m-q}\, du}{B(q, m+1-q)} \tag{31}$$

$$= F(m-n, q; m+1; 2).$$

Observe that for $\epsilon_j = \pm 1, j \in M$, 2^m functions, $\prod_K \epsilon_i, K \subset M$, are linearly independent. Indeed, they are orthogonal under the natural inner product,

$$\sum_{\epsilon_j = \pm 1, j\in M} \prod_K \epsilon_i \prod_L \epsilon_i = 2^m \delta_{L,K},$$

where K and L are fixed subsets of M. Also,

$$\sum_{\epsilon_j = \pm 1, j\in M, \prod_K \epsilon_i = \epsilon} \prod_L \epsilon_\ell = 2^{m-1}[\delta_{\emptyset,K} + \epsilon\delta_{L,K}].$$

Simplifying the notation from x_{n_j} to x_j, we see that the joint density of $x_j, j \in M$, and of the product, $\prod_K s_j$, $K \subset M$, has the form

$$f(x_j, j \in M; \prod_{i\in K} s_i = \epsilon) \tag{32}$$

$$= \frac{\prod_M f(x_j)}{2}\Big\{1 + \epsilon(-1)^{nk + \sum_K r_j}[1 - 2\sum_{j\in K} s_j^K F(x_j)]^{n-m}\Big\}.$$

When $K = M$, one obtains

$$E(\prod_M s_j | x_j, j \in M) = (-1)^{nm + m(m+1)/2}[1 - 2\sum_M s_j^M F(x_j)]^{n-m}, \tag{33}$$

so that by (31) with q_m given in (30),

$$E\prod_M s_j = (-1)^{nm + m(m+1)/2} F(m-n, q_m; m+1; 2). \tag{34}$$

To derive formulas for the moments of the parity-based sum, $\sum_{i=1}^n s_i x_i$, we need an extension of (32) for positive integers $\nu_j, j = 1, \ldots, m$. For this purpose, the form of the joint density of $x_j, j \in M$, and the product of the corresponding parities, $\prod_1^m s_j^{\nu_j}$, is desired.

This density can be obtained from (32), since $\prod_1^m s_j^{\nu_j} = \epsilon$, if and only if with $D = \{j, j = 1, \ldots, m, \nu_j \text{ odd }\}$, one has $\prod_{j,\nu_j\in D} s_j = \epsilon$. Thus, with d denoting the cardinality of D and r_j still denoting the rank of x_j within the subsample,

$$E(\prod_1^m s_j^{\nu_j} | x_j, j \in M) = (-1)^{nd + \sum_D r_j}[1 - 2\sum_{j\in D} s_j^D F(x_j)]^{n-m},$$

and

$$E\prod_{j=1}^m (s_j x_j)^{\nu_j} = (-1)^{nd} E(-1)^{\sum_{j\in D} r_j}[1 - 2\sum_{j\in D} s_j^D F(x_j)]^{n-m} \prod_{j=1}^m x_j^{\nu_j} \tag{35}$$

$$= (-1)^{nd + d(d+1)/2} E \prod_{j\in D} x_j^{\nu_j}[1 - 2\sum_{j\in D} s_j^D F(x_j)]^{n-m} \prod_{i\notin D} \mathbf{m}_{\nu_i}[1 - 2\sum_{j\in D} s_j^D F_{\nu_i}(x_j)],$$

where $F_\nu(x) = \int_{-\infty}^x u^\nu f(u)\, du/\mathbf{m}_\nu, \nu = 0, 2, 4, \ldots, s_i^D = \text{sign}(\prod_{j\neq i, j\in D}(x_i - x_j)$.

To prove (35), we evaluate the following conditional expectation:

$$E\left(\prod_{i:\nu_i \text{ even}}(-1)^{r_i}x_i^{\nu_i}|x_j, j\in D\right) = (-1)^{m(m-d)}[\prod_{i\neq\ell:i,\ell\notin D}\text{sign}(x_i-x_\ell)]$$

$$\times\prod_{i\notin D}\mathbf{m}_{\nu_i}E\left(\prod_{i:\nu_i \text{ even}}\prod_{k:\nu_k \text{ odd}}\text{sign}[\prod_{k\neq i}(x_i-x_k)]x_i^{\nu_i}|x_j, j\in D\right)$$

$$= (-1)^{m(m-d)+(m-d)(m-d-1)/2}\prod_{i\notin D}\mathbf{m}_{\nu_i}[1-2\sum_D s_j^D F_{\nu_i}(x_j)].$$

Identity (35) is valid when some $\nu_i, i\notin D$ vanish. Thus, for any non-negative integers $\nu_j, j=1,\ldots,n$,

$$E\prod_{j=1}^n(s_jx_j)^{\nu_j} = (-1)^{nd+d(d+1)/2}E\prod_{j\in D}x_j^{\nu_j}\prod_{i\notin D}\mathbf{m}_{\nu_i}[1-2\sum_{j\in D}s_j^D F_{\nu_i}(x_j)], \tag{36}$$

$D=\{j, j=1,\ldots,n, \nu_j \text{ odd}\}$,

More general formula involves functions $\phi_j, j=1,\ldots,n$,

$$E\prod_{j=1}^n s_j^{\nu_j}\phi_j(x_j) = (-1)^{nd+d(d+1)/2}E\prod_{j\in D}\phi_j(x_j)$$

$$\times\prod_{i\notin D}\left[\int_{-\infty}^{\infty}\phi_i(u)f(u)\,du - 2\sum_{j\in D}s_j^D\int_{-\infty}^{x_j}\phi_i(u)f(u)\,du\right].$$

According to (36),

$$E\left[\sum_1^n s_jx_j\right]^p = \sum_{\sum\nu_j=p}\binom{p}{\nu_1\ \cdots\ \nu_n}(-1)^{nd+d(d+1)/2} \tag{37}$$

$$\times E\prod_{j:\nu_j \text{ odd}}x_j^{\nu_j}\prod_{i:\nu_i \text{ even}}\mathbf{m}_{\nu_i}[1-2\sum_D s_j^D F_{\nu_i}(x_j)]$$

where d is as above and $p-d$ is even; (37) presents the correct version of Formula (34) in [10]. □

By using (28), one can find the joint (symmetric) density of $y_{n_1},\ldots,y_{n_m}$. For example, when f is assumed symmetric, the distribution of $y_1=s_1x_1$ and $y_2=s_2x_2$ is exchangeable, so that it suffices to determine its density when $|y_1|<|y_2|$. Then $\epsilon_1y_1\wedge\epsilon_2y_2=-|y_2|$, if $\epsilon_2=-\text{sign}(y_2)$; $=\epsilon_1y_1$, otherwise. Similarly, $\epsilon_1y_1\vee\epsilon_2y_2=|y_2|$, when $\epsilon_2=\text{sign}(y_2)$; $=\epsilon_1y_1$, otherwise. Thus,

$$\sum_{\epsilon_1y_1<\epsilon_2y_2}\epsilon_1[1-2F(\epsilon_1y_1)]^{n-2} + \sum_{\epsilon_2y_2<\epsilon_1y_1}\epsilon_2[1-2F(\epsilon_2y_2)]^{n-2}$$

$$= [1-(-1)^n][1-2F(y_1)]^{n-2} - 2\text{sign}(y_2)[1-2F(-|y_2|)]^{n-2},$$

and

$$\sum_{\epsilon_1y_1<\epsilon_2y_2}\epsilon_2[1-2F(\epsilon_2y_2)]^{n-2} + \sum_{\epsilon_2y_2<\epsilon_1y_1}\epsilon_1[1-2F(\epsilon_1y_1)]^{n-2}$$

$$= [1-(-1)^n][1-2F(y_1)]^{n-2} + 2\text{sign}(y_2)[1-2F(|y_2|)]^{n-2}.$$

Since

$$\sum_{\epsilon_1,\epsilon_2} \epsilon_1\epsilon_2[1-2|F(\epsilon_1 y_1)-F(\epsilon_2 y_2)|]^{n-2}$$

$$= 2\text{sign}(y_2)\Big\{[1-2|F(y_1)-F(|y_2|)|]^{n-2}-[1-2|F(y_1)+F(|y_2|)-1|]^{n-2}\Big\}$$

$$= 2\Big\{[1-2|F(y_1)-F(y_2)|]^{n-2}-[1-2|F(y_1)+F(y_2)-1|]^{n-2}\Big\},$$

it follows that the joint density of y_1, y_2 when $|y_1| < |y_2|$ is

$$g(y_1,y_2) = \sum_{\epsilon_1,\epsilon_2=\pm 1} f(\epsilon_1 y_1, \epsilon_2 y_2, s_i=\epsilon_1, s_j=\epsilon_2) \tag{38}$$

$$= f(y_1)f(y_2)\Big\{1+\frac{\text{sign}(y_2)[1+(-1)^n]}{2}[1-2F(|y_2|)]^{n-2}$$

$$-\frac{1}{2}\Big\{[1-2|F(y_1)-F(y_2)|]^{n-2}-[1-2|F(y_1)+F(y_2)-1|]^{n-2}\Big\}.$$

In the symmetric case, $E(\sum_{i=1}^n s_i x_i)^p = 0$, for odd n and p. Indeed, $\sum_{i=1}^n s_i x_i$ up to multiple $(-1)^n$ coincides with the parity sum derived from the sample $(-x_1,\ldots,-x_n)$, which is equidistributed with x's.

If $p=1$, the first moment can be derived from (23); the form of the second moment follows from (25),

$$E(\sum_{i=1}^n s_i x_i)^2 = n\mathbf{m}_2 - n(n-1)Ex_1x_2[1-2|F(x_2)-F(x_1)|]^{n-2}.$$

When $p=3\le n$,

$$E(\sum_{i=1}^n s_i x_i)^3 = nEs_1x_1^3 + 3n(n-1)Es_1x_1x_2^2 + n(n-1)(n-2)E\prod_1^3 s_i x_i$$

$$= n\int x^3 f(x)[2F(x)-1]^{n-1}\,dx$$

$$-3n(n-1)\int xf(x)[2F(x)-1]^{n-2}\left[\mathbf{m}_2 - 2\int_{-\infty}^x y^2 f(y)\,dy\right]dx$$

$$+6n(n-1)(n-2)\iint_{x<y<z} xyzf(x)f(y)f(z)$$

$$\times[1-2F(z)+2F(y)-2F(x)]^{n-3}\,dxdydz.$$

When $p=4\le n$,

$$E(\sum_{i=1}^n s_i x_i)^4 = n\mathbf{m}_4 + 3n(n-1)\mathbf{m}_2^2 Ex_1x_2^3[1-2|F(x_2)-F(x_1)|]^{n-2}$$

$$-6n(n-1)(n-2)\mathbf{m}_2 Ex_1x_2x_3^2[1-2|F_2(x_2)-F_2(x_1)|][1-2|F(x_2)-F(x_1)|]^{n-3}$$

$$+E\prod_1^4 x_i[1-2\sum_1^4 s_jF(x_j)]^{n-4}.$$

4. Parities and Hypergeometric Function

For fixed $m, 1 \le m \le n$, the joint distribution of parities $s_{n_1}, \ldots, s_{n_m}$ is obtained in Theorem 2, whose notation we follow. According to (27) and (28), if $1 \le m \le n$,

$$P(s_{n_j} = \epsilon_j, j = 1, \ldots, m) = \frac{\sum_{1 \le p_1 < \ldots < p_m \le n} \prod_{j=1}^{m} [1 + \epsilon_j (-1)^{n+p_j}]}{2^m \binom{n}{m}} \tag{39}$$

$$= \frac{1}{2^m} \sum_{k=0}^{m} (-1)^{nk+k(k+1)/2} F(k-n, q_k; k+1; 2) k! E_k(\mathbf{ffl}).$$

Here, for $\mathbf{ffl} = (\epsilon_1, \ldots, \epsilon_m)$, $E_k(\mathbf{ffl}) = \sum_{K \subset M, |K|=k} \prod_K \epsilon_i$, is the k-th elementary symmetric function whose values depend only on d, the number of ϵ's equal to -1. Indeed

$$\prod_{j=1}^{m} (z - s_j) = (z+1)^d (z-1)^{m-d} = \sum_{k=0}^{m} (-1)^{m-k} E_k(\mathbf{ffl}) z^{m-k},$$

so that

$$E_k(\mathbf{ffl}) = (-1)^{k-d} \sum_{\ell=0}^{k} (-1)^{\ell} \binom{m-d}{\ell} \binom{d}{k-\ell}.$$

Now we give explicit formulas for $F(m-n, q_k; k+1; 2)$ involving double factorials. See [15] for a survey of related combinatorial identities, and [16] for further instances of closed-form expressions for this function at specific arguments.

Theorem 3. *If $1 \le k \le m \le n, n \ge 2$, are positive integers, the following identities hold for the hypergeometric function:*

$$F(m-n, (k+1)/2; k+1; 2) = \frac{(n-m-1)!!k!!}{(n-m+k)!!}, \quad n-m \text{ even}, \ k \text{ odd}, \tag{40}$$

$$F(m-n, (k+1)/2; k+1; 2) = 0, \quad n-m \text{ odd}, \ k \text{ odd}, \tag{41}$$

$$F(m-n, k/2; k+1; 2) = F(m-n, (k+2)/2; k+1; 2) \tag{42}$$

$$= \frac{(n-m-1)!!(k-1)!!}{(n-m+k-1)!!}, \quad n-m \text{ even}, \ k \text{ even},$$

$$F(m-n, k/2; k+1; 2) = -F(m-n, (k+2)/2; k+1; 2) \tag{43}$$

$$= \frac{(n-m)!!(k-1)!!}{(n-m+k)!!}, \quad n-m \text{ odd}, \ k \text{ even}.$$

For any positive integers $q, 1 \le q \le k$, and p,

$$F(-p, q; m+1; 2) = (-1)^{n-m} F(-p, m+1-q; m+1; 2), \tag{44}$$

and (46) is valid.

Proof. To prove Formulas (40)–(43), we use the well-known facts about the hypergeometric function. According to 15.8.13 in [14]

$$F(m-n, (k+1)/2; k+1; z)$$

$$= \frac{(2-z)^{n-m}}{2^{n-m}} F\left(\frac{m-n}{2}, \frac{m-n+1}{2}; \frac{k+2}{2}; \frac{z^2}{(2-z)^2}\right).$$

If $n-m$ is even, this identity means that

$$F(m-n,(k+1)/2;k+1;z)=\frac{(2-z)^{n-m}}{2^{n-m}} \tag{45}$$

$$\times\sum_{0\le j\le(n-m)/2}\frac{[(m-n)/2]^{\bar{j}}[(m-n-1)/2]^{\bar{j}}}{[(j+2)/2]^{\bar{j}}j!}z^{2j}(2-z)^{n-m-2j}.$$

Here, for any real a and non-negative integer j, $[a]^{\bar{j}}=a(a+1)\cdots(a+j-1)=\Gamma(a+j)/\Gamma(a)$ is the ascending factorial.

The only term of the finite series in the right-hand side of (45) without a positive power of $(2-z)$ corresponds to $j=j_0=(n-m)/2$. Its coefficient equals

$$\frac{[(m-n)/2)]^{\bar{j}_0}[(m-n-1)/2]^{\bar{j}_0}}{[(k+2)/2]^{\bar{j}_0}j_0!}=\frac{\Gamma\left(\frac{n-m+1}{2}\right)\Gamma\left(\frac{k+2}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-m+k+1}{2}\right)},$$

which is seen to coincide with (40). This coefficient vanishes when $n-m$ is odd, implying (41).

One has

$$\frac{kz}{2}F\left(m-n,\frac{k}{2};k+1;z\right)$$

$$=kF\left(m-n,\frac{k}{2};k;z\right)+k(z-1)F\left(m-n-1,\frac{k}{2};k;z\right),$$

15.5.16 in [14] leading to (42) and (43).

Identity (39) means that for any positive integers, $k\le m\le n$,

$$\sum_{K\subset M,|K|=k}\sum_{1\le p_1<\ldots<p_m\le n}(-1)^{\sum_K p_j}=(-1)^{k(k+1)/2}\binom{n}{m}F(k-n,q_k;k+1;2). \tag{46}$$

$$=(-1)^{nk}\binom{n}{m}E\prod_K s_{n_j}=(-1)^{nk}\binom{n}{m}[2P(\prod_K s_{n_j}=1)-1].$$

In the last two formulas, K is any k-element subset of M. □

The values of the hypergeometric function entering (39) and other formulas with $0\le m\le n$, $q_m=\lceil m/2\rceil$, can be summarized as follows:

$$F(m-n,q_m;m+1;2)=\begin{cases}\frac{(m-1)!![1+(-1)^m]}{2(n-1)(n-3)\cdots(n-m+1)} & n \text{ even}\\ \frac{(m-1)!!}{(n-2)\cdots(n-m)} & n \text{ odd}, m \text{ even}\\ \frac{m!!}{(n-2)\cdots(n-m-1)} & n \text{ odd}, m \text{ odd}.\end{cases} \tag{47}$$

When $m=0$, this function takes its largest value, $F(-n,0;1;2)=1$. If $q=m+1$, $F(m-n,m+1;m+1;2)=(-1)^{n-m}$.

It is immediate that when $n\to\infty$ and m is fixed,

$$F(m-n,q_m;m+1;2)\sim\begin{cases}\frac{(m-1)!![1+(-1)^m]}{2n^{m/2}} & n \text{ even}\\ \frac{(m-1)!!}{n^{m/2}} & n \text{ odd}, m \text{ even}\\ \frac{m!!}{n^{(m+1)/2}} & n \text{ odd}, m \text{ odd}.\end{cases}$$

Now by using (39), one gets the first-order approximation for fixed $m \geq 2$,

$$2^m P(s_{n_j} = \epsilon_j, j = 1, \ldots, m) = 1 + \frac{m}{2n}\left[\epsilon_m - (-1)^n \epsilon_1 - \sum_{j=1}^{m-1} \epsilon_j \epsilon_{j+1}\right] + O\left(\frac{1}{n^2}\right), \quad (48)$$

which indicates the deviation from uniformity of the distribution of $(s_{n_1}, \ldots, s_{n_m})$.

Another proof of Formulas (40)–(43) in Theorem 3 uses the following generating function:

$$\sum_{k=0}^{\infty} F(-k, (p+1)/2; k+1; 2) z^k$$

$$= \frac{1}{B((p+1)/2, (p+1)/2)} \int_0^1 \frac{u^{(p-1)/2}(1-u)^{(p-1)/2}\, du}{1 - z^2(1-2u)} = F\left(1, \frac{1}{2}; \frac{p+2}{2}; z^2\right),$$

which is an even function of z when p is odd. If p is even,

$$\sum_{k \text{ even}} F(-k, p/2; p+1; 2) z^k + \sum_{k \text{ odd}} F(-k, p/2; p+1; 2) \frac{k z^k}{p+k}$$

$$= \sum_{k \text{ even}} F(-k, p/2; p+1; 2)(z+1) z^k$$

$$= \frac{1}{B(p/2, (p+2)/2)} \int_0^1 \frac{u^{(p-2)/2}(1-u)^{p/2}\, du}{1 - z^2(1-2u)} = F\left(1, \frac{1}{2}; \frac{p+1}{2}; z^2\right).$$

These facts can be found in 15.15.1 [14].

It is well known that the probabilities defining the classical hypergeometric distribution with parameters $m, n_e, n_o = n - n_e$, can be determined from its probability generating function, which is the (finite) hypergeometric series $F(m-n, -n_e; n_o - m + 1; z)$.

Therefore, the probability that such a random variable takes an even value (under any positive integers n_e and $n_o, n_e + n_o = n$) is

$$\sum_{\max(0, m-n_e) \leq k \leq \min(n_o, m)} \frac{[1 + (-1)^k] \binom{n_o}{k} \binom{n_e}{m-k}}{2\binom{n}{m}}$$

$$= \frac{1}{2}\left[1 + \frac{(n-m)! n_o!}{n!(n_o - m)!} F(-m, -n_e; n_o - m + 1; -1)\right].$$

If $n_e - n_o = [1 - (-1)^n]/2$, this probability coincides with $P(s_1 \cdots s_m = 1)$ whose expression through $F(m-n, q_m; m+1; 2)$ is given in Theorem 3.

The joint distribution of $s_1, \ldots, s_m$, which define traditional hypergeometric random variable $(s_1 + \cdots + s_m + m)/2$,

$$\frac{1}{2^m} \prod_{k=1}^{m} \left[1 + \frac{\epsilon_k (n_e - n_o - \epsilon_{k-1} - \cdots - \epsilon_1)}{n - k + 1}\right], \quad (49)$$

differs from (39). Indeed, the parities $s_1, \ldots, s_m, 1 \leq m \leq n$, in (39) are special because of their association with ranks of the subsample.

For two disjoint subsets K and K' of M and any L,

$$\sum_{\epsilon_i = \pm 1, \prod_K \epsilon_j = \mathbf{ffl}_1, \prod_{K'} \epsilon_j = \mathbf{ffl}_2} \prod_L \epsilon_\ell$$

$$= 2^{m-2}[\delta_{\emptyset, L} + \mathbf{ffl}_1 \delta_{K, L} + \mathbf{ffl}_2 \delta_{K', L} + \mathbf{ffl}_1 \mathbf{ffl}_2 \delta_{K \cup K', L}],$$

so that one obtains

$$4P(\prod_{i\in K} s_i=\epsilon_1,\prod_{i\in L} s_i=\epsilon_2|x_j, j\in M)=1+\epsilon_1(-1)^{nk+\sum_K r_i}[1-2\sum_K s_i^K F(x_i)]^{n-m} \quad (50)$$

$$+\epsilon_2(-1)^{n\ell+\sum_L r_i}[1-2\sum_L s_i^L F(x_i)]^{n-m}$$

$$+\epsilon_1\epsilon_2(-1)^{n(k+\ell)+\sum_{K\bigcup L} r_i}[1-2\sum_{K\bigcup L} s_i^{K\bigcup L} F(x_i)]^{n-m}.$$

It follows that

$$P(\prod_{i\in K} s_i=\epsilon_1,\prod_{i\in Ll} s_i=\epsilon_2)=\frac{1}{4}\Big[1+(-1)^{nk+q_k}F(m-n,q_k;m+1;2)\epsilon_1$$

$$+(-1)^{n\ell+q_\ell}F(m-n,q_\ell;m+1;2)\epsilon_2+(-1)^{n(k+\ell)+q_{k+\ell}}F(m-n,q_{k+\ell};m+1;2)\epsilon_1\epsilon_2\Big],$$

with similar formulas for the joint distribution of products of s's over several disjoint subsets.

As in Theorem 2, all joint probabilities are linear functions of ϵ_1, ϵ_2 and $\epsilon_1\epsilon_2$,

$$4P(s_i=\epsilon_1, s_j=\epsilon_2)=1+\epsilon_1\alpha+\epsilon_2\beta+\epsilon_1\epsilon_2\chi$$

with some coefficients α, β, χ. The degree of dependence of s's (or of their Bernoulli versions $(s_i+1)/2$) can be measured via the correlation coefficient between s_i and s_j,

$$\rho=\frac{2P(s_is_j=1)-1-[2P(s_i=1)-1][2P(s_j=1)-1]}{4\sqrt{P(s_i=1)P(s_i=-1)P(s_j=1)P(s_j=-1)}}=\frac{\chi-\alpha\beta}{\sqrt{(1-\alpha^2)(1-\beta^2)}}.$$

In our examples, $\chi<\alpha\beta$, and the dependence is negative.

In (39), if n is even, $\beta=-\alpha=-\chi=1/(n-1)$, when n is odd, $\alpha=\beta=-\chi=1/n$. More generally, in (50) $\alpha=(-1)^{nk+q_k+q_{m-k}}F(k-n,q_k;k+1;2), \beta=(-1)^{n(m-k)+q_k+q_{m-k}}F(m-k-n,q_{m-k};m-k+1;2), \delta=(-1)^{nm+q_m}F(m-n,q_m;m+1;2)$. Then if n is even, and m,k are odd, $\rho=0$, which means that $\prod_K s_{n_j}$ and $\prod_{K'} s_{n_j}$ are independent, $P(\prod_M s_{n_j}=\pm1)=1/2$.

When $n_e-n_o=[1-(-1)^n]/2$, one achieves in (49), $\alpha=\beta=[1-(-1)^n]/(2n)$, $\chi=-[2n+(-1)^n-1]/[2n(n-1)]$.

These formulas may find further use in probability modeling and estimation of entropy of binary sequences [17].

5. Parity-Based Sums and Dirichlet Distribution

We start here with the identity, which is similar to the formulas (40)–(43) in Theorem 3. Namely, for any positive integers p and n with $\nu_i\geq 0, i=1,\dots n$, forming a weak decomposition of p into n (non-negative) parts, one has

$$\sum_{\nu_1+\cdots+\nu_n=p}\prod_i s_i^{\nu_i}=\begin{cases}\binom{(p+n-2)/2}{p/2} & n \text{ even}, p \text{ even}\\ 0 & n \text{ even}, p \text{ odd}\\ \binom{(p+n-1)/2}{p/2} & n \text{ odd}, p \text{ even}\\ \binom{(p+n-2)/2}{(p-1)/2} & n \text{ odd}, p \text{ odd}.\end{cases} \quad (51)$$

Indeed, $\sum_{\nu_1+\cdots+\nu_m=p} 1 = \binom{p+n-1}{n-1}$, so that with $q = q_n$ defined by (30), the sum in the left-hand side of (51) can be written as

$$f_{m,p}(q) = \sum_{k=0}^{p} (-1)^k \binom{m-q+k-1}{m-q-1} \binom{q-1+p-k}{q-1}, \tag{52}$$

which is the coefficient at z^p in the series expansion of $(1-z)^{-q}(1+z)^{q-m}$, cf. Section 1.3 in [18].

One has, $f_{m,p}(m-q) = (-1)^p f_{m,p}(q), f_{m,0}(q) = 1, f_{0,p}(q) = 0, p \geq 1$. A formula similar to (52) for the generating function, $z^m(1-z)^{-q}(1+z)^{q-m}$, obtains for the partition of p (strictly positive ν's, $\sum_{k=1}^{m} \nu_k = p$).

Theorem 4. *For $f_{m.p}(q)$ given in (52), $q = q_m = \lceil m/2 \rceil$, one has*

$$f_{m+1,p}(q) = \binom{m+p}{p} F(-p, q; m+1; 2) \tag{53}$$

as well as (54) and (55).

Proof. The equality (53) holds because of (40)–(43). Indeed, its left-hand part satisfies (51), which corresponds to the weak decomposition of p. By comparing this equality with the partition of p into m (positive size) blocks, one obtains a representation of $f_{m,p}(q), q = q_m$,

$$(-1)^{mp} f_{m,p}(q) = \sum_{\nu_1+\cdots+\nu_m=p} \prod_i (-1)^{i\nu_i}$$

$$= \sum_{K\subset M} \sum_{\sum_K \nu_i = p, \nu_i > 0} \prod_{i\in K} (-1)^{i\nu_i} = \sum_{K\subset M} (-1)^k \sum_{\sum_K \nu_i = p-k, \nu_i \geq 0} \prod_{i\in K} (-1)^{i\nu_i}$$

$$= \sum_{K\subset M} (-1)^{k(p-k+1)} f_{k,p-k}\left(\frac{k + \sum_K (-1)^i}{2}\right).$$

Here, $k, 1 \leq k \leq m \wedge p$, denotes the cardinality of $K = \{i : \nu_i > 0\}$, so that $[k + \sum_K (-1)^i)]/2$ is the multiplicity of 1 among $(-1)^i, i \in K$.

The number of possible choices of the set K with the given multiplicity j of 1 in the set $\{(-1)^i, i \in K\}$, is $\binom{q}{j}\binom{m-q}{k-j}$. Therefore,

$$(-1)^{mp} f_{m,p}(q) = \sum_{k=1}^{m\wedge p} (-1)^{k(p-k+1)} \sum_{j=0}^{k} \binom{q}{j} \binom{m-q}{k-j} f_{k,p-k}(j). \tag{54}$$

Alternative representation of $f_{m,p}(q), p \geq 1$, results from the binomial theorem applied to $(1-s)^{-q}(1+s)^{q-m}, q = q_m$,

$$f_{m,p}(q) = \begin{cases} \frac{1}{2^q} \sum_{j=0}^{q} \binom{q}{j} f_{q,p}(j) & m \text{ even} \\ \frac{1}{2^q} \sum_{j=0}^{q} \binom{q}{j} [f_{q,p}(j) - f_{q,p-1}(j)] & m \text{ odd.} \end{cases} \tag{55}$$

□

The coefficients $f_{m,p}(q_m)$ also appear in the conditional cumulative distribution function corresponding to (32). For positive N put,

$$U_m^N(y_1, \ldots, y_m) = \int \cdots \int_{-\infty < x_j < y_j, j=1,\ldots,m} [(-1)^m - 2\sum_j (-1)^{r_j} F(x_j)]^N \prod_j f(x_j)\, dx_j \tag{56}$$

$$= (-1)^{mN} \int \cdots \int_{0<u_j<F(y_j),j=1,\ldots,m} [1 - 2\sum_j s_j^M u_j]^N \prod_j du_j,$$

where r_j denotes the rank of x_j, $s_j^M = \text{sign}(\prod_{k\neq j, j,k\in M}(x_j - x_k))$.

Then, the mentioned distribution function can be expressed through U_m^{n-m} as

$$F_\epsilon(x_{n_1}, \ldots, x_{n_m}) = \frac{1}{2P(\prod_j s_{n_j} = \epsilon)}$$

$$\times \left[\prod_M F(x_{n_j}) + \epsilon(-1)^{m(m-1)/2} U_m^{n-m}(x_{n_1}, \ldots, x_{n_m}) \right],$$

which implies that $U_m^N(\infty, \ldots, \infty) = (-1)^{mN} F(-N, q_m; m+1, 2)$.

Thus, under the Dirichlet distribution Dir_{m+1} of $\Delta_1, \ldots, \Delta_{m+1}$, with all concentration parameters 1, one has for even m,

$$U_m^N(x, \infty, \ldots, \infty) = \int_0^{F(x)} \int_0^1 \cdots \int_0^1 [1 - 2(\Delta_m + \Delta_{m-2} + \cdots + \Delta_2)]^N d\text{Dir}_{m+1}.$$

Now we use the facts that the marginal distribution of Δ_1 has a beta-density $\beta(\Delta_1)$ with parameters $(1, m)$, and that the conditional distribution of $(\Delta_2', \ldots, \Delta_{m+1}') = (\Delta_2/(1-\Delta_1), \ldots, \Delta_{m+1}/(1-\Delta_1))$ is Dir_m.

Thus, when m is even,

$$(-1)^{mN} U_m^N(x, \infty, \ldots, \infty) \tag{57}$$

$$= \int_0^{F(x)} \int_0^1 \cdots \int_0^1 \left[\frac{v_1}{1-v_1} + 1 - 2 \sum_{i=2,\ldots,m} \frac{\Delta_i}{1-\Delta_1} \right]^N (1-v_1)^N \beta(v_1)\, dv_1 d\text{Dir}_m$$

$$= \sum_{k=0}^N \binom{N}{k} F(-k, q_m; m, 2) \int_0^{F(x)} v_1^{N-k} (1-v_1)^k \beta(v_1)\, dv_1$$

$$= \binom{m+N}{N}^{-1} \sum_{k=0}^N \frac{f_{m+1,k}(q_m)}{m+k} I_{F(x)}(N-k+1, m+k).$$

Here we took advantage of the binomial theorem, (31) and (53); in the last equality $I_p(N-k+1, m+k) = \int_0^p v^{N-k}(1-v)^{m+k-1}\, dv / B(N-k+1, m+k), 0 \le p \le 1$, denotes the incomplete beta function.

Identity (57) also holds for odd values of m. For example,

$$U_1^N(x) = \frac{(-1)^N \{1 - [1 - 2F(x)]^{N+1}\}}{2(N+1)},$$

and

$$U_2^N(x, \infty) = \frac{F(x)}{N+1} - \frac{[1-(-1)^N]\{1-[1-2F(x)]^{N+2}\}}{4(N+1)(N+2)}.$$

Interest in U is due to the multivariate integration by parts, which also motivates the study of its derivatives in the next section.

6. Partial Derivatives

For $1 \le m \le n$, we look at the properties of the symmetric function $G = G_m$, which so far is defined almost everywhere (for pairwise different x's),

$$G(x_1, \ldots, x_m) = [1 - 2\sum_{j=1}^m s_j F(x_j)]^n,$$

$s_j = \text{sign}(\prod_{j;x_j \neq x_i}(x_i - x_j)$.

If the data consists of clusters, $C_j = \{i : x_i = x_j\}, j = 1, \ldots, k$, then using the definition by continuity put

$$G(x_1, \ldots, x_m) = [1 - 2 \sum_{j:|C_j| \text{ odd}} s_j F(x_j)]^n,$$

which allows possibly equal x's. Notice that functions s_j are discontinuous.

If all x's coincide, then under this definition, $G(x, \ldots, x) = 1$, if m is even; $= [1 - 2F(x)]^n$, if m is odd. If there are $m - k$ points equal to $+\infty$, then $G = [1 - 2\sum_{j=1}^k s_j F(x_j)]^n, m - k$ even;$= [-1 - 2\sum_{j=1}^k s_j F(x_j)]^n, m - k$ odd. If $m - k$ of x's are equal to $-\infty$, then $G = [1 - 2\sum_{j=1}^k s_j F(x_j)]^n$.

Thus, $G(x_1, \ldots, x_m)$ becomes a continuous function whose absolute value is bounded by 1. Actually, it possesses the Lipschitz property if the density f is bounded.

Our goal here is to determine the generalized derivative of order m, $\partial^m G(x)/\prod_j \partial x_j, x = (x_1, \ldots, x_m)$, so that for any smooth compactly supported φ, the multivariate integration by parts formula holds

$$\iint \frac{\partial^m G(x)}{\prod_j \partial x_j} \varphi(x)\, dx = (-1)^m \iint \frac{\partial^m \varphi(x)}{\prod_j \partial x_j} G(x)\, dx.$$

Theorem 5. *With* $(n)_k = n(n-1)\cdots(n-k+1) = \Gamma(n+1)/\Gamma(n+1-k)$, *denoting the descending factorial, one has*

$$\frac{\partial^m}{\prod_j \partial x_j} G(x_1, \ldots, x_m) = \sum_{k=1}^m (-2)^k (n)_k [1 - 2\sum_{j=1}^m s_j F(x_j)]^{n-k} \tag{58}$$

$$\times \left[\binom{m}{k}\right]^{-1} \sum_{K \subset M, |K|=k} \prod_{j \in K} f(x_j) \frac{\partial^{m-k}}{\prod_{i \notin K} \partial x_i} \prod_{j \in K} s_j.$$

Proof. For $1 \leq k \leq m$, differentiation over $x_1, \ldots, x_k$ shows that

$$\frac{\partial^k}{\partial x_1 \cdots \partial x_k} [1 - 2\sum_{j=1}^m s_j F(x_j)]^n = \sum_{p=1}^k (-2)^p (n)_p [1 - 2\sum_{j=1}^m s_j F(x_j)]^{n-p} A_p^{(k)},$$

where

$$A_p^{(k)} = \prod_{j=1}^p f(x_j) \frac{\partial^{k-p}}{\prod_{\ell=p+1}^k \partial x_\ell} \prod_{j=1}^p s_j.$$

For $k \geq 2$, $A_p^{(k)} = A_p^{(k)}(x_1, \ldots, x_k)$ is a symmetric function of its arguments,

$$A_p^{(k)} = \left[\binom{k}{p}\right]^{-1} \sum_{1 \leq j_1 < \ldots < j_p \leq k} \prod_{i=1}^p f(x_{j_i}) \frac{\partial^{k-p}}{\prod_{\ell \neq j_1, \ldots, j_p} \partial x_\ell} \prod_{i=1}^p s_{j_i}.$$

For $1 \leq p \leq k$, these functions can be characterized by the following recursion with $A_0^{(k)} = 0, A_0^{(0)} = 1, A_p^{(k)} = 0, p > k$,

$$A_p^{(k)} = s_k f(x_k) A_{p-1}^{(k-1)} + \frac{\partial}{\partial x_k} A_p^{(k-1)}.$$

The proof is by induction. When $k = 1$,

$$\frac{\partial}{\partial x_i} [1 - 2\sum_{j=1}^m s_j F(x_j)]^n = -2n s_i f(x_i) [1 - 2\sum_{j=1}^m s_j F(x_j)]^{n-1}.$$

Indeed, shift invariance of s_j means that for any i, $\sum_j \frac{\partial s_j}{\partial x_i} = 0$. Thus, $A_1^{(1)} = s_1 f(x_1)$, and for $k \geq 2$,

$$A_1^{(k)} = f(x_1)\frac{\partial^{k-1} s_1}{\partial x_2 \cdots \partial x_k} = (-2)^{k-1} f(x_1) \prod_{j=2}^{k} \delta(x_1 - x_j) \prod_{i=k+1}^{m} \text{sign}(x_1 - x_i),$$

which indeed is a symmetric function of $x_1, \ldots, x_k$.

The following induction steps are straightforward, so that (58) follows. □

For $m = 2, s_1 = -s_2$, so that according to (58),

$$\frac{\partial^2}{\partial x \partial y}[1 - 2|F(x) - F(y)|]^n = -4n(n-1)f(x)f(y)[1 - 2|F(x) - F(y)|]^{n-2}$$

$$+4nf(x)\delta(x - y).$$

For $m = 3$,

$$\frac{\partial^3}{\partial x \partial y \partial z}[1 - 2s_1F(x) - 2s_2F(y) - 2s_3F(z)]^n$$

$$= 8n(n-1)(n-2)[1 - 2s_1F(x) - 2s_2F(y) - 2s_3F(z)]^{n-3} f(x)f(y)f(z)$$

$$+8n(n-1)[1 - 2s_1F(x) - 2s_2F(y) - 2s_3F(z)]^{n-2}[s_1s_2 f(x)\delta(x - y) + s_1s_3 f(z)\delta(x - z)$$

$$+s_2s_3 f(y)\delta(y - z)] - 8nf(x)\delta(y - x)\delta(z - x).$$

Let

$$\widetilde{G}_m(x_1, \ldots, x_m) = \sum_{k=0}^{m-1} \sum_{K, |K|=k} (-1)^{m-k} G_k(x_j, j \in K),$$

where summation is over all proper subsets K of M, $\widetilde{G}_0 = 1$. Then $\widetilde{G}_m$ is *grounded*, i.e., it vanishes if $x_j = -\infty$ for at least one j. If one of the x's is equal to $+\infty$, then $\widetilde{G}_m(x_1, \ldots, x_{m-1}, \infty) = [(-1)^n - 1]\widetilde{G}_{m-1}(x_1, \ldots, x_{m-1})$, so that for even n, $\widetilde{G}_m$ vanishes if $x_j = +\infty$ for at least one j. When $m \geq 2$, $\widetilde{G}_m(x, \infty, \ldots, \infty) = (-1)^m 2^{m-2}[1 - (-1)^n]\{1 - [1 - 2F(x)]^n\}$, $\widetilde{G}_p(+\infty, \ldots, +\infty) = (-1)^m 2^{m-1}[1 - (-1)^n]$.

Thus,

$$\frac{\partial^m}{\prod_j \partial x_j} G(x_1, \ldots, x_m) = \frac{\partial^m}{\prod_j \partial x_j} \widetilde{G}_m(x_1, \ldots, x_m).$$

For $m = 2$ and even n it follows that for any integrable function $\phi(x, y) = \phi(y, x)$,

$$\int\int \phi(x,y) f(x) f(y) [1 - 2|F(x) - F(y)|]^{n-2} \, dxdy$$

$$= \frac{1}{n-1} \int \phi(x,x)) f(x) \Big\{ 1 - \frac{[1 - (-1)^n]}{2} [2F(x) - 1]^{n-1} \Big\} \, dx$$

$$- \frac{1}{4n(n-1)} \int\int \frac{\partial^2 \phi(x,y)}{\partial x \partial y} \Big\{ [1 - 2F(x \vee y) + 2F(x \wedge y)]^n$$

$$+ (-1)^n - [1 - 2F(x \vee y)]^n - [2F(x \wedge y) - 1]^n \Big\} \, dxdy,$$

which is useful to determine the covariance structure.

7. Conclusions

Interesting properties of self-dual probabilities demonstrate their potential in statistical estimation without any additional variance information. The polynomial approximation

over a finite set, as well as the Gauss hypergeometric function, are intimately related to data parities and parity-based distributions.

Several presented combinatorial identities may find wider use in probability theory applications, in particular, in random matrices and in statistical physics.

Funding: This research received no external funding.

Data Availability Statement: Data sharing is not applicable to this article as no new data were created or analyzed. All conclusions are based on mathematical reasoning, and the author is solely responsible for them.

Acknowledgments: Many thanks are due to anonymous referees for their very careful reading of the original version and all their critical comments. The role of *Mathematics* editorial board in securing such referees is also acknowledged.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Rukhin, A.L. Estimating common mean in a heteroscedastic variances model. *Mathematics* **2025**, *13*, 1290. [CrossRef]
2. Rukhin, A.L. Estimation of the common mean from heterogeneous normal observations with unknown variances. *J. R. Stat. Soc. Ser. B* **2017**, *79*, 1601–1618. [CrossRef]
3. Rivlin, T.J. *An Introduction to Approximation of Functions*; Dover: New York, NY, USA, 1969.
4. Trefethen, L.N. *Approximation Theory and Approximation Practice*; SIAM: Philadelphia, PA, USA, 2013.
5. Karlin, S.; Studden, W.J. *Tchebysheff Systems: With Applications in Analysis and Statistics*; Wiley: New York, NY, USA, 1966.
6. de Boor, C.; Saff, E.B. Finite sequences of orthogonal polynomials connected by a Jacobi matrix. *Linear Algebra Appl.* **1986**, *75*, 43–56. [CrossRef]
7. Borodin, A. Duality of orthogonal polynomials on a finite set. *J. Stat. Phys.* **2002**, *109*, 1109–1120. [CrossRef]
8. Vinet, L.; Zhedanov, A. The characterization of classical and semiclassical orthogonal polynomials from their dual polynomials. *J. Comp. Appl. Math.* **2004**, *172*, 41–48. [CrossRef]
9. Genest, V.; Tsujimoto, S.; Vinet, L.; Zhedanov, A. Persymmetric Jacobi matrices, isospectral deformations and orthogonal polynomials. *J. Math. Anal. Appl.* **2017**, *450*, 915–928. [CrossRef]
10. Rukhin, A.L. Parities and hypergeometric function. *Theory Probab. Appl.* **2025**, *70*, 355–374. [CrossRef]
11. Fang, K.-T.; Pan, J. A review: Representative points of statistical distributions and their applications. *Mathematics* **2023**, *11*, 2930. [CrossRef]
12. David, H.A.; Nagarajah, N.H. *Order Statistics*, 3rd ed.; Wiley: New York, NY, USA, 2003.
13. Ng, K.W.; Tian, G.-L.; Tang, M.-L. *Dirichlet and Related Distributions: Theory, Methods and Applications*; Wiley: New York, NY, USA, 2011.
14. Olver, F.W.J.; Lozier, D.W.; Boisvert, R.W.; Clark, C.W. *NIST Handbook of Mathematical Functions*; NIST: Gaithersburg, MD, USA; U.S. Department of Commerce: Washington, DC, USA; Cambridge University Press: Cambridge, UK, 2010.
15. Callan, D. A combinatorial survey of identities for the double factorials. *arXiv* **2009**, arXiv:0906.1317.
16. Li, Y.-W.; Qi, F. A new closed-form formula of the Gauss hypergeometric function at specific arguments. *Axioms* **2024**, *13*, 317. [CrossRef]
17. De Gregorio, J.; Sanchez, D.; Toral, R. Entropy estimators for Markovian sequences: A comparative analysis. *Entropy* **2024**, *26*, 79. [CrossRef] [PubMed]
18. Riordan, H.J. *Combinatorial Identities*; Wiley: New York, NY, USA, 1968.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.