The γ-transform Approach: a New Method for the Study of a Discrete and Finite Random Variable

Fabio Grandi

***Abstract*—**In this paper, we present a new method for the study of a discrete random variable whose probability distribution has a finite support. The approach is based on the introduction of a transform of the probability density function, named γ-transform, which better suits the finite nature of the random variable than the traditional probability generating function. In particular, in addition to the transformation/anti-transformation pair, a simple formula is presented for computing the factorial moments of a random variable directly from the γ-transform of its probability density function. Moreover, it is shown how the γ-transform can be determined from the nature of the combinatorial problem under study thanks to its physical meaning. Examples and applications to estimation problems relevant for computer science are provided, in which the simple construction of a γ-transform gives immediate access to the complete characterization of the underlying probability distribution (density function and moments).

***Keywords*—**Discrete probability, factorial moments, transforms, combinatorial analysis, estimation

I. Introduction

SEVERAL modelling problems relevant for performance evaluation of information retrieval and database management systems [1]–[7], [9], [10], [12]–[16], [18]–[21], [23], [24] imply the study of a discrete and finite random variable. Although such problems may allow a simple determination of the expected value of the random variable involved, the probability density function is usually difficult to compute and handle for the evaluation of higher-order moments. As a matter of fact, even very simple problems yield complex probability distributions, involving alternating-sign summations with binomial coefficients, owing to their relationship with the *principle of inclusion and exclusion* [22]. Furthermore, the determination of the moments from such distributions is not straightforward; even the evaluation of the variance may result in a challenging task.

In general, a common method for the study of a (non negative) discrete random variable X consists in using the *probability generating function*, defined as

$$G(z) \;=\; \mathrm{E}[z^X] \;=\; \sum_{x\geq 0} z^x\, f(x) \qquad (1)$$

where $f(x)$ is the probability density function (mass function) of X, and which can also be regarded as a sort of *z-transform* of the function $f(\cdot)$. Using standard techniques, $G(z)$ can be formally derived from the nature of the problem under study (e.g., see Appendix D). Then, $f(x)$ and all the *factorial moments* of X can be computed from $G(z)$ thanks to:

$$f(x) \;=\; [z^x]G(z) \qquad (2)$$

$$\mathrm{E}[X^{\underline{r}}] \;=\; G^{(r)}(1) \qquad (3)$$

where the notations $[x^m]A$ and $x^{\underline{m}}$ stand for the coefficient of x^m in A and for m-th falling factorial power of x, respectively.

Notice that equation (3) can easily be derived from the expression of the Taylor (McLaurin) series expansion of the r-th derivative of G:

$$G^{(r)}(z) \;=\; \sum_{i\geq 0} \frac{G^{(r+i)}(0)}{i!} z^i \qquad (4)$$

and owing to the fact that $f(x) = G^{(x)}(0)/x!$.

Although the probability generating function approach is a very general methodology, we put forward the claim that it might not be the most convenient approach when dealing with a random variable having a distribution with *finite* support, that is which takes values only in a finite set and, thus, has only a finite number of nonnull moments. We would rather explore the possibility that a methodology based on a *finite Newton series* [17] —involving finite summations and differences— could be more appropriate than the above one based on a Taylor expansion —involving derivatives and formally infinite summations. Supporting such a claim has been the main motivation of this work, which will illustrate the practical consequences that arise from its assertion.

Aimed at fulfilling this aim, our alternative approach is based on the introduction in Section II of a new transform, called γ-transform, that we defined in [14] and that satisfies the above mentioned "finiteness" requirements. The adoption of the γ-transform as finite calculus's answer to the probability generating function is the subject of Section III: owing to a combinatorial identity demonstrated in Sec. II, we will show how the new transform allows a fast determination of all the factorial moments of a discrete and finite random variable; moreover, the physical meaning of the new transform is explained, which will allow a direct derivation of its expression in the context of a given combinatorial problem; finally, relationships between $G(z)$ and the γ-transform are discussed. Examples and outstanding applications are presented in Sections IV and V, respectively. Conclusions will eventually be found in Section VI, whereas a case study involving the

F. Grandi is with the Computer Science and Engineering Department (DISI), Alma Mater Studiorum – Università di Bologna, Viale Risorgimento 2, I-40136 Bologna BO, Italy (phone +39-051-2093555; fax: +39-051-2093953; e-mail: fabio.grandi@unibo.it).

comparison of the γ-transform with alternative approaches can be found in the Appendix.

II. Preliminaries

In this section, we introduce the definition of γ-transform for a generic function $f(\cdot)$ along with some fundamental combinatorial identities involving it.

A. The gamma-transform

Let $f(\cdot)$ be a fixed function defined in $\{0, 1, \ldots, n\}$, then its γ-*transform* is defined in $\{0, 1, \ldots, n\}$ by:

$$\gamma(y) = \sum_{x=0}^{n} \frac{\binom{y}{x}}{\binom{n}{x}} f(x) \tag{5}$$

B. Anti-transformation formula

The corresponding *inversion formula* is given by:

$$f(x) = \binom{n}{x} \sum_{j=0}^{x} (-1)^j \binom{x}{j} \gamma(x-j) \tag{6}$$

and can be demonstrated as follows. It can be observed from (5) that $\gamma(y)$ is a polynomial function of degree n in y and, thus, it can be expressed as a finite Newton series:

$$\gamma(y) = \sum_{x=0}^{n} \binom{y}{x} \Delta^x \gamma(0) \tag{7}$$

Comparing (5) with (7) yields:

$$f(x) = \binom{n}{x} \Delta^x \gamma(0) \tag{8}$$

From Eq. (7), thanks to the properties of Newton series, we can derive:

$$\Delta^r \gamma(y) = \sum_{x=0}^{n} \binom{y}{x-r} \Delta^x \gamma(0) \tag{9}$$

Hence, Eq. (6) can easily be obtained from (8) when expliciting the x-th difference using (9).

In order to support our claim, it can be noticed how (9) can actually represent the finite calculus's counterpart of (4).

C. A combinatorial identity

A fundamental identity involving the γ-transform is the subject of the next Theorem.

Theorem 1: If $f(\cdot)$ is a fixed function defined in $\{0, 1, \ldots, n\}$, then the following combinatorial identity holds:

$$\sum_{x=0}^{n} x^{\underline{r}} f(x) = n^{\underline{r}} \sum_{i=0}^{r} (-1)^i \binom{r}{i} \gamma(n-i) \tag{10}$$

where $\gamma(\cdot)$ is the γ-transform of $f(\cdot)$.

Proof: Owing to the definition of r-th difference, the summation in the right-hand side of (10) is $\Delta^r \gamma(n-r)$. Hence, thanks to (9), the right-hand side of (10) can be rewritten as:

$$\begin{aligned} & n^{\underline{r}} \Delta^r \gamma(n-r) \\ &= \sum_{x=0}^{n} n^{\underline{r}} \binom{n-r}{x-r} \Delta^x \gamma(0) \\ &= \sum_{x=0}^{n} x^{\underline{r}} \binom{n}{x} \Delta^x \gamma(0) \end{aligned} \tag{11}$$

The final expression (11) equals the left-hand side of (10), thanks to Eq. (8). ■

III. Probabilistic interpretation

In this section, we bring to light the role played by the γ-transform when $f(\cdot)$ represents the probability density function of a finite and discrete random variable.

A. Evaluation of the moments

Let X be a discrete random variable with values in $\{0, 1, \ldots, n\}$ and probability density function $f(x)$. All the moments of X can be computed from the γ-transform of $f(\cdot)$ as stated by the following Corollary of Theorem 1.

Corollary 1: Given a discrete random variable X with values in $\{0, 1, \ldots, n\}$, its r-th factorial moment is provided by:

$$\mathrm{E}[X^{\underline{r}}] = n^{\underline{r}} \sum_{i=0}^{r} (-1)^i \binom{r}{i} \gamma(n-i) \tag{12}$$

where $\gamma(\cdot)$ is the gamma-transform of the probability density function of X.

Proof: It immediately follows from the definition of expected value and Theorem 1. ■

Obviously, all the standard moments can be computed from (12), thanks to:

$$\mathrm{E}[X^r] = \sum_{s=0}^{r} \left\{ {r \atop s} \right\} \mathrm{E}[X^{\underline{s}}]$$

where $\left\{ {r \atop s} \right\}$ is a Stirling number of the second kind. For instance, this entails:

$$\mathrm{E}[X] = n\,[1 - \gamma(n-1)] \tag{13}$$

$$\begin{aligned} \sigma_X^2 &= n^2 \left[\gamma(n-2) - \gamma^2(n-1)\right] \\ &\quad + n\,[\gamma(n-1) - \gamma(n-2)] \end{aligned} \tag{14}$$

which are really simple formulae.

B. Physical meaning

An important physical meaning can be given to the γ-transform of the probability density function of a discrete and finite random variable, as stated by the following Theorem.

Theorem 2: Let X be a random variable, with values in $\{0, 1, \ldots, n\}$ and probability density function $f(x)$, which can be regarded as the number of successes occurring in an experiment composed of a set $\mathcal{N}$ of n indistinguishable trials,

effected as if the successful trials were randomly selected in $\mathcal{N}$. Let $\mathcal{Y} \subseteq \mathcal{N}$ be a subset of trials fixed before the experiment and let $\Pr[\mathcal{Y}]$ be the probability that the experiment be effected as if the successes could only be selected from $\mathcal{Y}$ rather than from the whole $\mathcal{N}$. Then it can be shown that:

$$\Pr[\mathcal{Y}] \quad = \quad \gamma(y)$$

where $\gamma(\cdot)$ is the γ-transform of the probability density function of X and y is the cardinality of the set $\mathcal{Y}$.

Proof: Since in general the experiment can provide any number $X \in \{0, 1, \ldots, n\}$ of successes, $\Pr[\mathcal{Y}]$ can be determined by means of the total probability Theorem as follows:

$$\Pr[\mathcal{Y}] \quad = \quad \sum_{x=0}^{n} \Pr[\mathcal{Y}|X = x] \Pr[X = x]$$

Since all the trials are indistinguishable and, thus, $\binom{m}{x}$ is the number of ways of choosing the x successes in a set of m trials, we have:

$$\Pr[\mathcal{Y}] \quad = \quad \sum_{x=0}^{n} \frac{\binom{y}{x}}{\binom{n}{x}} f(x)$$

■

Moreover, also the inversion formula (6) can be proved with only probabilistic arguments, as shown in the following. Let $\Pr[\mathcal{X}']$ be the probability that the successful trials only be selected in the set $\mathcal{X}'$, then as a consequence of the principle of inclusion and exclusion we have:

$$\begin{aligned}
&\Pr[X = x] \\
&= \sum_{\substack{\mathcal{X} \subseteq \mathcal{N} \\ |\mathcal{X}| = x}} \left(\Pr[\mathcal{X}] - \sum_{\substack{\mathcal{X}' \subseteq \mathcal{X} \\ |\mathcal{X}'| = x-1}} \Pr[\mathcal{X}'] + \cdots \right. \\
&\qquad \left. \cdots + (-1)^{x-1} \sum_{\substack{\mathcal{X}' \subseteq \mathcal{X} \\ |\mathcal{X}'| = 1}} \Pr[\mathcal{X}'] + (-1)^x \Pr[\emptyset] \right) \\
&= \sum_{\substack{\mathcal{X} \subseteq \mathcal{N} \\ |\mathcal{X}| = x}} \sum_{j=0}^{x} (-1)^j \sum_{\substack{\mathcal{J} \subseteq \mathcal{X} \\ |\mathcal{J}| = j}} \Pr[\mathcal{X} \setminus \mathcal{J}] \qquad (15)
\end{aligned}$$

Owing to the physical meaning of $\gamma(\cdot)$, the probability $\Pr[\mathcal{X} \setminus \mathcal{J}]$ is exactly $\gamma(x-j)$. Hence, thanks to the indistinguishability of trials (summations reduce to counts of equal quantities), it can easily be verified that (15) equals the right-hand side of (6).

C. Relationship with $G(z)$

The following relationship between the γ-transform and the probability generating function $G(z)$ can also be shown:

$$G(z) \quad = \quad \sum_{j=0}^{n} \binom{n}{j} z^j (1-z)^{n-j} \gamma(j) \qquad (16)$$

In order to prove it, it is sufficient to show that the density function (6) can be derived from (16) as $f(x) = [z^x]G(z)$. By means of the binomial Theorem and with simple manipulations, Eq. (16) can be rewritten as:

$$G(z) \quad = \quad \sum_{i=0}^{n} z^i \binom{n}{i} \sum_{j=0}^{i} (-1)^{i-j} \binom{i}{j} \gamma(j)$$

which evidences the $[z^i]G(z)$ term.

An inverse relationship can be derived as follows. Since $\gamma(y)$ is a non-decreasing function (with $\gamma(0) = f(0)$ and $\gamma(n) = 1$) and since from (16) we have:

$$\sum_{j=0}^{n} \binom{n}{j} \gamma(j) \quad = \quad \sum_{j=0}^{n} \binom{n}{j} \gamma(n-j) \quad = \quad 2^n G(1/2)$$

where also $G(1/2)$ is usually a function of n, letting

$$g(x) \quad = \quad \Delta^x \left[2^n G(1/2) \right](0)$$

we can write:

$$\gamma(y) \quad = \quad \begin{cases} g(y) & \text{if } g(n) = 1 \\ g(n-y) & \text{if } g(0) = 1 \end{cases} .$$

Moreover, it can also be shown that the probability generating function approach can be derived as a limit of the γ-transform theory when the discrete random variable involved is *not limited*. For instance, consider the γ-transform definition (5): since

$$\frac{\binom{y}{x}}{\binom{n}{x}} \quad = \quad \prod_{i=0}^{x-1} \frac{y/n - i/n}{1 - i/n}$$

we can let $n, y \to \infty$ (maintaining constant the ratio $y/n = z$) obtaining:

$$\lim_{n,y \to \infty} \gamma(y) = G(z)$$

owing to definition (1). Also other formulae concerning $G(z)$ can be obtained from the corresponding ones concerning $\gamma(y)$ by taking the same limit. This is the final argument in favor of our initial claim. Formal similarities between the approaches based on the probability generating function and on the γ-transform can be eventually appreciated in Table I.

IV. Examples

Examples of application of the γ-transform approach are provided in this Section. Its use is shown here in evaluating the factorial moments of a random variable with well-known distributions.

A. Uniform distribution

Let X be uniformly distributed in $\{0, 1, \ldots, n\}$:

$$f(x) \quad = \quad \frac{1}{n+1}$$

TABLE I
A SUMMARY COMPARISON BETWEEN THE APPROACHES BASED ON $G(z)$ AND $\gamma(y)$.

probability generating function	**γ-transform**
X discrete and infinite	X discrete and finite
$G(z) = \sum_{x\geq 0} z^x\, f(x)$	$\gamma(y) = \sum_{x=0}^{n} \binom{y}{x} \Big/ \binom{n}{x} f(x)$
$f(x) = \frac{1}{x!} G^{(x)}(0)$	$f(x) = \binom{n}{x} \Delta^x \gamma(0)$
$\mathrm{E}[X^{\underline{r}}] = G^{(r)}(1)$	$\mathrm{E}[X^{\underline{r}}] = n^{\underline{r}}\, \Delta^r \gamma(n-r)$

The γ-transform of the density function can be evaluated as:

$$\begin{aligned}\gamma(y) &= \frac{1}{n+1} \sum_{x=0}^{n} \frac{\binom{y}{x}}{\binom{n}{x}} \\ &= \frac{1}{n+1-y}\end{aligned}$$

owing to identity (5.33) of [17].

Applying Corollary 1 to compute the factorial moments, we obtain:

$$\begin{aligned}\mathrm{E}[X^{\underline{r}}] &= n^{\underline{r}} \sum_{i=0}^{r} (-1)^i \binom{r}{i} \frac{1}{i+1} \\ &= \frac{n^{\underline{r}}}{r+1}\end{aligned}$$

as identity (5.41) of [17] can be used in the last step.

B. Binomial distribution

If we consider a random variable X following a binomial distribution:

$$f(x) = \binom{n}{x} p^x q^{n-x}$$

(with $p+q=1$), we can easily obtain the γ-transform as:

$$\begin{aligned}\gamma(y) &= \sum_{x=0}^{n} \binom{y}{x} p^x q^{n-x} \\ &= q^{n-y}\end{aligned}$$

thanks to the binomial Theorem.

Applying Corollary 1 we easily obtain:

$$\begin{aligned}\mathrm{E}[X^{\underline{r}}] &= n^{\underline{r}} \sum_{i=0}^{r} \binom{r}{i} (-q)^i \\ &= n^{\underline{r}}\, p^r\end{aligned}$$

Notice how the γ-transform could also be directly derived from its physical meaning in a simple way, as X counts the number of Bernoulli trials underlying the binomial experiment that produce a successful outcome. If the successful trials could only be selected in a subset $\mathcal{Y} \subseteq \mathcal{N}$, then all the (independent) trials in $\mathcal{N} \setminus \mathcal{Y}$ lead to a failure, which has a probability $q^{|\mathcal{N}\setminus\mathcal{Y}|}$.

C. Hypergeometric distribution

If X has a hypergeometric distribution:

$$f(x) = \frac{\binom{n}{x}\binom{N-n}{k-x}}{\binom{N}{k}}$$

we can easily compute the γ-transform:

$$\begin{aligned}\gamma(y) &= \frac{\sum_{x=0}^{n} \binom{y}{x}\binom{N-n}{k-x}}{\binom{N}{k}} \\ &= \frac{\binom{y+N-n}{k}}{\binom{N}{k}}\end{aligned}$$

owing to Vandermonde's convolution formula.

By applying Corollary 1 we obtain:

$$\begin{aligned}\mathrm{E}[X^{\underline{r}}] &= n^{\underline{r}}\, \frac{\sum_{i=0}^{r} (-1)^i \binom{r}{i}\binom{N-i}{k}}{\binom{N}{k}} \\ &= n^{\underline{r}}\, \frac{\binom{N-r}{N-k}}{\binom{N}{k}} = r!\, \frac{\binom{n}{r}\binom{k}{r}}{\binom{N}{r}}\end{aligned}$$

which is the value usually found in the literature.

Notice how also in this case the γ-transform could be directly derived from its physical meaning, as X counts the number of successful trials in a sample of size k extracted (without replacement) from a population $\mathcal{N}$ of N trials, n of which are successful. If the successful trials could only be selected in a subset of $\mathcal{N}$ with size y, then $\gamma(y)$ can be computed as the probability that the sample is actually extracted, not from a set of $N-n$ failures and n successes, but from a set of $N-n$ failures and y successes.

D. Beta-binomial distribution

If X is a random variable with beta-binomial distribution:

$$f(x) = \binom{n}{x} \frac{\mathrm{B}(x+\alpha, n+\beta-x)}{\mathrm{B}(\alpha,\beta)}$$

we can compute the γ-transform of the density function as follows:

$$\begin{aligned}\gamma(y) &= \frac{1}{\mathrm{B}(\alpha,\beta)}\sum_{x=0}^{n}\binom{y}{x}\frac{\Gamma(x+\alpha)\Gamma(n+\beta-x)}{\Gamma(n+\alpha+\beta)} \\ &= \frac{\mathrm{B}(\alpha,n+\beta-y)}{\mathrm{B}(\alpha,\beta)}\end{aligned}$$

The summation above (and also the one in the next paragraph) is a hypergeometric which can be evaluated as a Vandermonde's convolution [17].

The factorial moments of X can be computed as:

$$\begin{aligned}\mathrm{E}[X^{\underline{r}}] &= n^{\underline{r}}\frac{\Gamma(\alpha+\beta)}{\Gamma(\beta)}\sum_{i=0}^{r}(-1)^i\binom{r}{i}\frac{\Gamma(\beta+i)}{\Gamma(\alpha+\beta+i)} \\ &= n^{\underline{r}}\frac{\mathrm{B}(\alpha+r,\beta)}{\mathrm{B}(\alpha,\beta)}\end{aligned}$$

V. Applications

The utility of the γ-transform approach lies in the fact that some estimation problems can be described by means of complex distributions which do indeed have a simple γ-transform. Not only are the moments easy to compute from the γ-transform in these cases, but also the γ-transform can be directly and easily derived from the nature of the problem. In particular, expressing the γ-transform from its physical meaning, results particularly simple when the random variable X represents a count of trials with a successful outcome and the order in which successes occur is irrelevant (as seen in the previous section for the Binomial and hypergeometric distributions). This class of problems includes several modelling and estimation problems relevant for performance evaluation of information retrieval and database management systems, which are briefly referenced and analyzed in this Section.

In such cases, since $\gamma(y)$ is a probability, it can be noticed that it could also be expressed as:

$$\gamma(y) = \frac{\psi(y)}{\psi(n)} \tag{17}$$

where $\psi(y)$ represents the *number of ways* in which the experiment considered could be effected by selecting the successes only in a subset of y trials.

In general, if the experiment considered is composed of m independent sub-experiments, $\gamma(y)$ can conveniently be expressed as:

$$\gamma(y) = \prod_{k=1}^{m}\gamma_k(y) \tag{18}$$

where $\gamma_k(y)$ is the probability that the k-th sub-experiment be effected by selecting the successes only in a subset of y trials (which is also independent of k if the sub-experiments are indistinguishable).

When both assumptions hold (i.e., the underlying experiment is composed of independent sub-experiments and sub-experiments can be modelled as counting of trials), Eq. (17) and (18) can be combined yielding:

$$\gamma(y) = \prod_{k=1}^{m}\frac{\psi_k(y)}{\psi_k(n)} \tag{19}$$

with an obvious meaning of $\psi_k(\cdot)$.

A. Set union problem

Let $\mathcal{N}$ be a set with cardinality n, let $\mathcal{S}_k \quad (1 \le k \le m)$ be a random subset of $\mathcal{N}$ with cardinality s_k, and X the random variable denoting the cardinality of the union set $\mathcal{U} = \bigcup_{k=1}^{m}\mathcal{S}_k$.

The set union problem corresponds to the execution of the experiment schematized in Fig. 1, where the k-th sub-experiment effects a sampling of s_k objects from $\mathcal{N}$ and X counts the number of distinct objects globally selected. Sampling is without replacement within each sub-experiment and with replacement between different sub-experiments.

Considering the inclusion of an element of $\mathcal{N}$ in $\mathcal{U}$ to be a successful trial, the selections of the subsets $\mathcal{S}_1,\ldots,\mathcal{S}_m$ can be regarded as mutually independent sub-experiments. The γ-transform of the probability density function of X can be expressed according to Eq. (19), since $\psi_k(y) = \binom{y}{s_k}$ is the number of ways in which the elements of $\mathcal{S}_k$ can be selected only in a subset of $\mathcal{N}$ with cardinality y, yielding:

$$\gamma(y) = \prod_{k=1}^{m}\frac{\binom{y}{s_k}}{\binom{n}{s_k}}$$

Therefore, the probability density function of X is:

$$f(x) = \binom{n}{x}\sum_{j=0}^{x}(-1)^j\binom{x}{j}\prod_{k=1}^{m}\frac{\binom{x-j}{s_k}}{\binom{n}{s_k}} \tag{20}$$

By means of Corollary 1, we can easily derive the expected value and the variance of X as:

$$\mathrm{E}[X] = n\left[1-\prod_{k=1}^{m}\left(1-\frac{s_k}{n}\right)\right] \tag{21}$$

$$\begin{aligned}\sigma_X^2 = n^2&\left[\prod_{k=1}^{m}\left(1-\frac{s_k}{n}\right)\left(1-\frac{s_k}{n-1}\right)-\prod_{k=1}^{m}\left(1-\frac{s_k}{n}\right)^2\right]+ \\ &n\left[\prod_{k=1}^{m}\left(1-\frac{s_k}{n}\right)-\prod_{k=1}^{m}\left(1-\frac{s_k}{n}\right)\left(1-\frac{s_k}{n-1}\right)\right]\end{aligned} \tag{22}$$

Set union problems of interest for computer science are numerous. For instance, X can be regarded as the number of "1" bits in a binary word of n bits resulting from the inclusive or of m words, where s_k is the number of "1" bits in the k-th operand word. Thus, the set union problem is equivalent to the estimation of the signature weight as generated by the superimposed coding technique adopted in "multiple" m signature files [1]. The estimation is needed

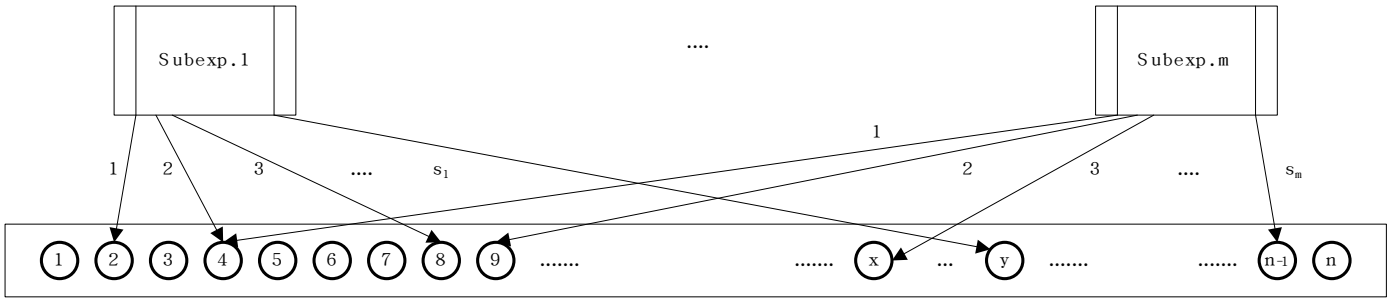

Fig. 1. The "set union problem" experiment

for performance evaluation of such organizations used for information retrieval applications. The equivalence of (20) with the density function published in [1] was shown in [16]. It was also noticed that the method sketched in [1] and developed in [21] through Markov chains and heavy matrix manipulations leads to a slightly less handy formula than (20). Moreover, as far as we know, no other authors derived a closed formula like Eq. (22) for the evaluation of the variance of X, which is indeed necessary, for instance, for an accurate evaluation of the false drop probability as we showed in [16]. The set union problem has been also studied in [2, Sec. 3.1.2] to derive the statistics for the maintenance of a distributed document classifier: considering that the size of the intersection between two subsets follows a hypergeometric distribution, an iterative formula (with a subset added to the union at each iteration) has been proposed for the incremental estimation of the union size. Although no closed formula has been provided, the resulting expected value agrees with (21). No expression for the probability density function or higher moments has been derived in that study.

An interesting case also arises when $s_k = s$ for each k (the sub-experiments are indistinguishable), and X represents the number of "1" bits in the more "classical" superimposed codes [23] adopted for information retrieval with signature files [9]. In such a case, Equations (20)–(22) reduce to:

$$f(x) = \binom{n}{x} \sum_{j=0}^{x} (-1)^j \binom{x}{j} \left[\frac{\binom{x-j}{s}}{\binom{n}{s}} \right]^m \quad (23)$$

$$\mathrm{E}[X] = n \left[1 - \left(1 - \frac{s}{n} \right)^m \right] \quad (24)$$

$$\sigma_X^2 = n \left(1 - \frac{s}{n} \right)^m \cdot \left[1 - \frac{(n-s)^m}{n^{m-1}} + \frac{(n-s-1)^m}{(n-1)^{m-1}} \right] \quad (25)$$

The density function (23) and the expected value (24) agree with those presented in [23].

Moreover, if $s = 1$ then X represents the number of distinct objects selected in sampling with replacement m objects from a population of n. Equations (23)–(25) become:

$$f(x) = \binom{n}{x} \sum_{j=0}^{x} (-1)^j \binom{x}{j} \left(\frac{x-j}{n} \right)^m \quad (26)$$

$$\mathrm{E}[X] = n \left[1 - \left(1 - \frac{1}{n} \right)^m \right] \quad (27)$$

$$\sigma_X^2 = n \left(1 - \frac{1}{n} \right)^m \cdot \left[1 - \frac{(n-1)^m}{n^{m-1}} + \frac{(n-2)^m}{(n-1)^{m-1}} \right] \quad (28)$$

For example, X represents the number of blocks accessed in a file (with a total number of n blocks) during the retrieval of m records that are not necessarily distinct, under the *total uniformity* [15] assumption (i.e., each record has the same probability to be selected and blocks contain the same number of records). The estimation of such value is necessary for cost-based query optimization [19]and database physical design [10]. The expected value (27) agrees with the formula of Cárdenas [4]. For an expression of the underlying density function see, for instance, [5], [13]. The variance (28), which we derived via the γ-transform approach for the first time in [14], agrees with the value computed (for a number of empty urns equivalent to *non* selected blocks) in [8] using a bivariate generating function approach (see Appendix E).

A comparison of the γ-transform approach with alternative methods (namely, combinatorial calculus, the principle of inclusion and exclusion, generating functions and Markov chains) in the application to this simple problem can be found in the Appendix. Such a comparison highlights the valuability of the new approach from a practical point of view, as it saves heavy computations which are otherwise needed for the evaluation of the probability density function and of higher-order moments.

B. Group inclusion problem

An even more general problem with important applications to database management and design is described in the following. Let $\mathcal{N}$ be a set with cardinality N composed of n groups of objects, each of size g (namely, $N = g\,n$). We now define X as the number of distinct groups represented by the elements included in the union $\mathcal{U} = \bigcup_{k=1}^{m} \mathcal{S}_k$, where each $\mathcal{S}_k$ is a random subset of $\mathcal{N}$ with cardinality s_k. From another point of view, X is the number of distinct elements in the union of random subsets of a *multiset* in which all the n distinct objects appear g times.

The group inclusion problem corresponds to the execution of the experiment schematized in Fig. 2. Sampling of objects

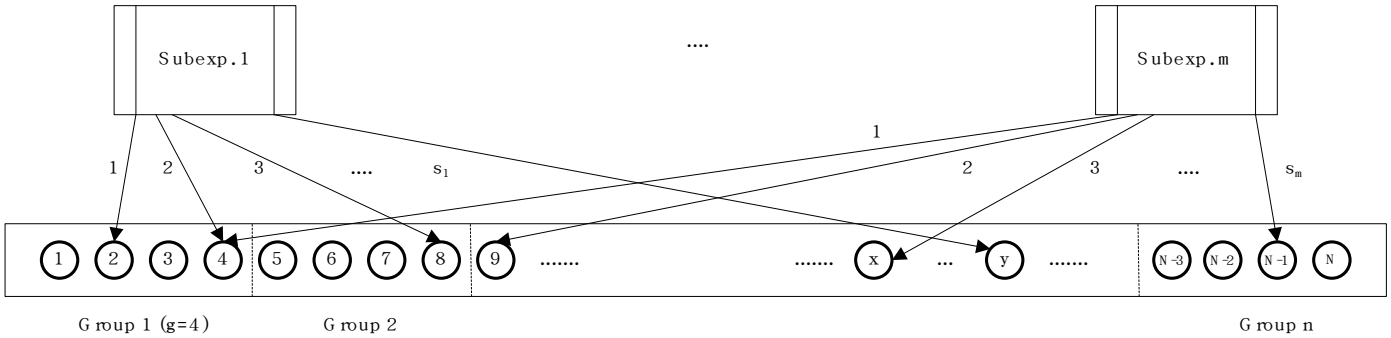

Fig. 2. The "group inclusion problem" experiment

from $\mathcal{N}$ is effected in the same way as in the set union problem, but X counts the number of distinct groups from which objects are globally selected.

For example, X represents the number of blocks accessed in a file (with a total number of n blocks) during the execution of a *batch reading*, composed of m independent queries, the k-th thereof retrieves s_k distinct records, under the total uniformity assumption. Such estimation is necessary to build accurate access cost models to be used for multi-query optimization [18], [20].

In this case, Eq. (19) can still be used, with $\psi_k(y) = \binom{g\,y}{s_k}$, that counts the number of ways the s_k objects can be selected from y groups only, yielding:

$$\gamma(y) = \prod_{k=1}^{m} \frac{\binom{g\,y}{s_k}}{\binom{N}{s_k}}$$

and, thus,

$$f(x) = \binom{n}{x} \sum_{j=0}^{x} (-1)^j \binom{x}{j} \prod_{k=1}^{m} \frac{\binom{g(x-j)}{s_k}}{\binom{N}{s_k}} \tag{29}$$

$$\mathrm{E}[X] = n \left[1 - \prod_{k=1}^{m} \frac{\binom{N-g}{s_k}}{\binom{N}{s_k}} \right] \tag{30}$$

$$\sigma_X^2 = n^2 \left[\prod_{k=1}^{m} \frac{\binom{N-2g}{s_k}}{\binom{N}{s_k}} - \prod_{k=1}^{m} \frac{\binom{N-g}{s_k}^2}{\binom{N}{s_k}^2} \right] + n \left[\prod_{k=1}^{m} \frac{\binom{N-g}{s_k}}{\binom{N}{s_k}} - \prod_{k=1}^{m} \frac{\binom{N-2g}{s_k}}{\binom{N}{s_k}} \right] \tag{31}$$

To the best of our knowledge, no estimation formulae have been presented before the introduction of the γ-transform approach for the probabilistic characterization of this problem.

An also interesting case takes place when $m = 1$, that is a single query is considered and, thus, X represents the number of blocks accessed during the retrieval of $s_1 = s$ distinct records (n is the total number of blocks, g is the number of records per block and, thus, $g\,n = N$ is the total number of records). Equations (29)–(31) become:

$$f(x) = \binom{n}{x} \sum_{j=0}^{x} (-1)^j \binom{x}{j} \frac{\binom{g(x-j)}{s}}{\binom{g\,n}{s}} \tag{32}$$

$$\mathrm{E}[X] = n \left[1 - \frac{\binom{N-g}{s}}{\binom{N}{s}} \right] \tag{33}$$

$$\sigma_X^2 = n^2 \left[\frac{\binom{N-2g}{s}}{\binom{N}{s}} - \frac{\binom{N-g}{s}^2}{\binom{N}{s}^2} \right] + n \left[\frac{\binom{N-g}{s}}{\binom{N}{s}} - \frac{\binom{N-2g}{s}}{\binom{N}{s}} \right] \tag{34}$$

The expected value (33) agrees with the formula of Yao [24]. Derivations of the distribution (32) can be found, for instance, in [3], [5], [12], [13]. The variance (34), which we derived via the γ-transform approach for the first time in [14], agrees with the value computed in [12] by means of a bivariate generating function approach (see Appendix E).

C. Yet another cell visit problem

Let us finally consider another application that can effectively be described in terms of the γ-transform. Assume we have N objects with D distinct types distributed into n cells, with the constraint that each cell contains representatives of exactly d distinct object types ($N \geq d\,n$). Then consider m sub-experiments, in the k-th of which all the cells containing at least one representative of s_k out of D distinct object types are visited (e.g., to retrieve all the representatives of that type). If $\mathcal{S}_k$ is the set of cells visited in the k-th sub-experiment, then we define X as the random variable counting the number of distinct cells globally visited in the whole experiment (i.e., equal to the cardinality of the union set $\bigcup_{k=1}^{m} \mathcal{S}_k$).

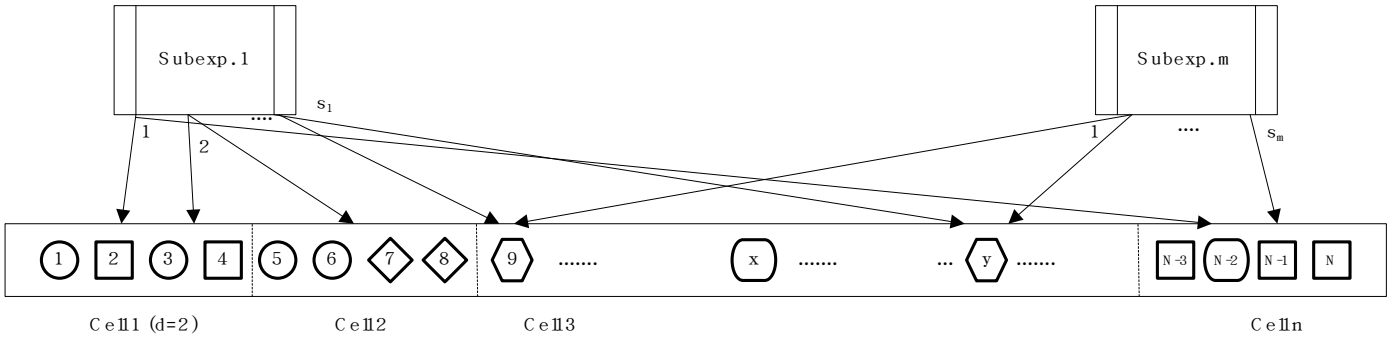

Fig. 3. The "yet another cell visit problem" experiment

This formalization corresponds to the execution of the experiment schematized in Fig. 3, where object types are represented with different shapes. Notice that the constraint on cell occupancy is that each cell contains exactly the same number of object types, possibly with a different number of representatives (a different number of objects could even be contained in each cell). In sub-experiment 1, the first object type selected is square (objects 2, 4, N-3, N-1, N), the second is round (objects 1, 3, 5, 6), ... and the s_1-th is hexagon (objects 9, y) and so on. X counts the number of distinct cells from which objects are globally selected.

For example, X represents the number of blocks accessed in a file (composed of n blocks) during the execution of a batch reading, composed of m independent queries, the k-th thereof retrieves all the records matching s_k distinct key values in the presence of data duplication and of *uniform clustering* [15] of the data (i.e., each value has the same probability to be selected and blocks contain the same number of distinct values). The parameter d represents the number of distinct key values contained in any block under the uniform clustering assumption. This problem can be described as an experiment in which trials correspond to cells, and successes to cells to be visited. Therefore, in order to express the γ-transform, we can use Eq. (18): for the k-th sub-experiment, $\gamma_k(y)$ represents represents the probability that $n-y$ of the cells have been excluded *a priori* from the result. Once these (indistinguishable) cells have been fixed, each of them has the same probability of being excluded from the result, which is independent on y and can be evaluated as:

$$\gamma_k = \frac{\binom{D-d}{s_k}}{\binom{D}{s_k}} \tag{35}$$

if the s_k object types are distinct, and:

$$\gamma_k = \left(1-\frac{d}{D}\right)^{s_k} \tag{36}$$

if they are not. In fact, (35) and (36) represent the probability that the d objects types contained in the cell are not involved in the sub-experiment. In both cases, the γ-transform of the density function has the form:

$$\gamma(y) = \prod_{k=1}^{m} \gamma_k^{n-y}$$

In particular, if the s_k objects selected in a sub-experiment are distinct, the probability density function for our cell visit problem from (35) becomes:

$$f(x) = \binom{n}{x} \sum_{j=0}^{x} (-1)^j \binom{x}{j} \prod_{k=1}^{m} \left[\frac{\binom{D-d}{s_k}}{\binom{D}{s_k}} \right]^{n-x+j} \tag{37}$$

The expected value and variance of X can then be computed as:

$$\mathrm{E}[X] = n \left[1 - \prod_{k=1}^{m} \frac{\binom{D-d}{s_k}}{\binom{D}{s_k}} \right] \tag{38}$$

$$\sigma_X^2 = n \prod_{k=1}^{m} \frac{\binom{D-d}{s_k}}{\binom{D}{s_k}} \left[1 - \prod_{k=1}^{m} \frac{\binom{D-d}{s_k}}{\binom{D}{s_k}} \right] \tag{39}$$

Else, if the s_k objects are not distinct, the probability density function from (36) becomes:

$$f(x) = \binom{n}{x} \sum_{j=0}^{x} (-1)^j \binom{x}{j} \prod_{k=1}^{m} \left(1-\frac{d}{D}\right)^{s_k(n-x+j)} \tag{40}$$

The expected value and variance of X can then be computed as:

$$\mathrm{E}[X] = n \left[1 - \prod_{k=1}^{m} \left(1-\frac{d}{D}\right)^{s_k} \right] \tag{41}$$

$$\sigma_X^2 = n \prod_{k=1}^{m} \left(1-\frac{d}{D}\right)^{s_k} \cdot \left[1 - \prod_{k=1}^{m} \left(1-\frac{d}{D}\right)^{s_k} \right] \tag{42}$$

Such estimation formulae are needed to build accurate access cost models to be used for multi-query optimization in the presence of uniform clustering of the data (by the way, notice that a large fraction of the data columns in a relational database fits the uniform clustering model). Notice that none of these results have been derived before and they would be quite hard to derive without the help of the γ-transform theory.

An also interesting case is when a single sub-experiment is considered ($m = 1$). Hence, X represents the number of blocks accessed in a file (composed of n blocks) during the retrieval of $s_1 = s$ distinct key values in the presence of data duplication and of uniform clustering of the data, whose estimation is needed for cost-based single query optimization [6], [15] and database physical design [7].

In case the s data values are distinct, we can use equations (37)–(39) which reduce to:

$$\begin{aligned} f(x) &= \binom{n}{x} \sum_{j=0}^{x} (-1)^j \binom{x}{j} \left[\frac{\binom{D-d}{s}}{\binom{D}{s}} \right]^{n-x+j} \\ &= \binom{n}{x} \left[1 - \frac{\binom{D-d}{s}}{\binom{D}{s}} \right]^{x} \left[\frac{\binom{D-d}{s}}{\binom{D}{s}} \right]^{n-x} \end{aligned} \quad (43)$$

and:

$$\mathrm{E}[X] = n \left[1 - \frac{\binom{D-d}{m}}{\binom{D}{s}} \right] \quad (44)$$

$$\sigma_X^2 = n \frac{\binom{D-d}{s}}{\binom{D}{s}} \left[1 - \frac{\binom{D-d}{s}}{\binom{D}{s}} \right] \quad (45)$$

In case the s data values are not distinct, from equations (40)–(42) we can derive:

$$\begin{aligned} f(x) &= \binom{n}{x} \sum_{j=0}^{x} (-1)^j \binom{x}{j} \left(1 - \frac{d}{D} \right)^{s(n-x+j)} \\ &= \binom{n}{x} \left[1 - \left(1 - \frac{d}{D} \right)^{s} \right]^{x} \left(1 - \frac{d}{D} \right)^{s(n-x)} \end{aligned} \quad (46)$$

and:

$$\mathrm{E}[X] = n \left[1 - \left(1 - \frac{d}{D} \right)^{s} \right] \quad (47)$$

$$\sigma_X^2 = n \left(1 - \frac{d}{D} \right)^{s} \left[1 - \left(1 - \frac{d}{D} \right)^{s} \right] \quad (48)$$

In both cases, which can also be regarded as particular cases of binomial distributions, the expected values agree with those derived in [6], [15]. Even for these simpler problems, to the best of our knowledge, no derivation of the probability density function and of σ_X^2 has been done before.

VI. Conclusion

In this paper, we have put forward the claim that the classical approach for the study of a discrete random variable based on the probability generating function could not be the most appropriate when the distribution of the random variable has a finite support. In such a case, in order to better suit the finiteness property, we proposed an alternative approach based on the introduction of a new transform, named γ-transform, of the probability density function. We have shown how, substituting an approach based on Taylor expansions with an approach based on finite Newton series, the probability density function and all the factorial moments of a finite random variable can easily be computed from the γ-transform. We have also shown how the probability generating function approach can be obtained back as a limit of the γ-transform theory when the domain of the discrete random variable becomes unlimited, which completes the support of our claim.

Moreover, we also showed how the expression of the γ-transform can be derived in an easy way thanks to its physical meaning for several combinatorial problems. Several examples of its useful application to modelling problems relevant for performance evaluation and physical design of database management or information retrieval systems were provided, showing how the γ-transform approach looks really attractive in such domains. All the ready-to-use formulae presented in this work are general and simple to handle. They hide the most difficult computations involved in the probabilistic description and characterization of a problem, which have been embedded in their derivations.

Appendix

In this Appendix, a comparison of the proposed approach with alternative methods can be appreciated. This aims at an evaluation of the γ-transform approach from a practical point of view. Let us make use of one of the most simple problems in the family we considered in Section V: the sampling with replacement of m objects out of a population of n (e.g., set union of random subsets each containing only one element). Notice that such a problem can be considered a particular case, with $s_1 = s_2 = \cdots = s_m = 1$, either of the set union problem, or of the group inclusion problem, or of the cell visit problem (with $D = N$ and $d = N/n$) studied in Sec. V.

The only simple thing to determine for this problem is the expected value (27) of X, since $(1 - 1/n)^m$ represents the probability that one of the n objects is not included in the result of the m selections [4]. The evaluation of the variance or of the complete distribution (26) of X cannot be effected in an elementary way. The main methods to be used are exposed in the following and compared with the γ-transform approach.

A. Combinatorial calculus

The density function can be directly computed as the ratio between the counts of favorable and total events N_F/N_T. In this case we simply have $N_T = n^m$ which represents the number of ways of putting m different objects into n different

cells (more objects can fit into the same cell). The number N_F is not as simple: it represents the number of ways of putting m different objects into n different cells so that exactly x cells are occupied. Skilled readers can evaluate this number as [22, Ch. 5, p. 92]:

$$N_F = n^{\underline{x}} \left\{ {m \atop x} \right\} \tag{49}$$

and derive:

$$\begin{aligned} \Pr[X = x] &= \frac{n^{\underline{x}} \left\{ {m \atop x} \right\}}{n^m} = \frac{\binom{n}{x}}{n^m} x! \left\{ {m \atop x} \right\} \\ &= \frac{\binom{n}{x}}{n^m} \sum_{j=0}^{x} (-1)^{x-j} \binom{x}{j} j^m \end{aligned} \tag{50}$$

where (6.19) of [17, Sec. 6.1] has been used in the last step. Clearly, (50) equals the density function (26). The direct determination of the moments from (50) is not straightforward, since it requires computations equivalent to those involved in the proof of Theorem 1. Explicit passages, without resorting to the notion of finite difference and exploiting its properties, are as shown in [14, Lemma1, Lemma 2 and Th. 1].

B. Principle of inclusion and exclusion

The density function can be evaluated by means of the principle of inclusion and exclusion. To this end, computations similar to those leading to the derivation of Eq. (15) presented in Sec. III-B must be effected [14, Th. 4]. Once the distribution has been evaluated, the determination of the moments has to be done as in the previous case.

C. Markov chain

The repeated selection of objects can be viewed as a Markov process, where each step corresponds to the selection with replacement of one object. Being $X^{(r)}$ the random variable representing the number of distinct objects selected after the first r choices, the one-step transition probabilities can be computed as:

$$\begin{aligned} &\Pr\left[X^{(r)} = k | X^{(r-1)} = i\right] \\ &= \begin{cases} i/n & \text{if } k = i \\ 1 - i/n & \text{if } k = i+1 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

This is the generic entry P_{ik} (with $0 \le i, k \le n$) of the probability matrix $P = P^{(1)}$ which is the same at every step of the chain. It can be noticed that P is triangular (upper bidiagonal indeed), thus its diagonal element $P_{ii} = i/n$ is also its eigenvalue λ_i. If E is a matrix whose columns are distinct eigenvectors of P, then we can write $P = E\Lambda E^{-1}$, where $\Lambda = \mathrm{diag}(\lambda_0, \ldots, \lambda_n)$. This similarity transformation allows the m-step probability matrix to be rewritten for the whole process as:

$$P^T = (E\Lambda E^{-1})^m = E\Lambda^m E^{-1}$$

The k-th eigenvector $\boldsymbol{e}^k$ can be computed via the equation $P\boldsymbol{e}^k = \lambda_k \boldsymbol{e}^k$. Imposing that $(\boldsymbol{e}^k)_k = 1$, the simultaneous equations yield the solution:

$$E_{ik} = (\boldsymbol{e}^k)_i = \binom{n-i}{n-k}$$

After some algebraic manipulations, matrix E can be inverted providing:

$$(E^{-1})_{ik} = (-1)^{k-i} \binom{n-i}{n-k}$$

Therefore, being $\Lambda_{ik} = i/n\ \delta_{ik}$ and $(\Lambda^m)_{ik} = (i/n)^m\ \delta_{ik}$, where δ_{ik} is a Kronecker's delta, we can finally evaluate the probability matrix for the complete Markov chain as follows:

$$\begin{aligned} &(E\Lambda^m)_{ik} \\ &= \sum_{j=0}^{n} \binom{n-i}{n-j} \left(\frac{k}{n}\right)^m \delta_{jk} \\ &= \binom{n-i}{n-k} \left(\frac{k}{n}\right)^m \\ &(P^m)_{ik} = (E\Lambda^m E^{-1})_{ik} \\ &= \sum_{j=0}^{n} \binom{n-i}{n-j} \left(\frac{j}{n}\right)^m (-1)^{k-j} \binom{n-j}{n-k} \\ &= \binom{n-i}{n-k} \sum_{j=0}^{n} \binom{k-i}{k-j} (-1)^{k-j} \left(\frac{j}{n}\right)^m \end{aligned}$$

Hence, the desired density function can be finally determined[1] as:

$$\begin{aligned} &\Pr[X = x] \\ &= \Pr[X^{(m)} = x | X^{(0)} = 0] = (P^m)_{0x} \\ &= \binom{n}{n-x} \sum_{j=0}^{n} \binom{x}{x-j} (-1)^{x-j} \left(\frac{j}{n}\right)^m \end{aligned}$$

which is equivalent to (26) (in order to show it, the proof in [16, Sec. 3] can easily be adapted). Yet the moments have to be computed from the density function as in the previous cases.

D. Generating function

We can compute the density function by means of a generating function approach. The enumerator for the m choices can easily be written as the formal polynomial:

$$P(\bar{x}) = \left(\frac{x_1}{n} + \cdots + \frac{x_n}{n}\right)^m \tag{51}$$

[1]This is the method used in [21] to determine the probability distribution of the signature weight in "multiple" m signature files. However, it can be applied to every experiment that can be decomposed into m independent sub-experiments and, thus, Eq. (18) holds. In this general case, it can be shown that the transition probabilities for the r-th step are given by:

$$(P_r)_{ik} = \binom{n-i}{k-i} \sum_{j=0}^{k-i} \binom{k-i}{j} (-1)^j \gamma_r(k-j)$$

This can be shown by directly computing $\Pr[X^{(r)} = k | X^{(r-1)} = i]$ using the principle of inclusion and exclusion as outlined in the previous section. Moreover, $P_r = E\Lambda_r E^{-1}$, where E is still the eigenvector matrix of the example and $\Lambda_r = \mathrm{diag}(\gamma_r(0), \ldots, \gamma_r(n))$.

where the variable x_i represent the choice of the i-th object and $1/n$ is the constant probability for the choice of an object. The probability generating function $G(z)$ for the cardinality of the result can be computed as:

$$G(z) = \Theta_{x_1^j \to z} \cdots \Theta_{x_f^j \to z} P(\bar{x})$$

where the operator $\Theta_{x_i^j \to z}$ entails the substitution of the variable z for the term x_i^j. It can be shown [13, Lemma 1] that $G(z)$ can also be computed as:

$$G(z) = \sum_{j=0}^{n} G_j \, z^j (1-z)^{n-j} \tag{52}$$

where the term G_j is obtained by summing over every j-combination $\{i_1, \ldots, i_j\}$ of the indexes $\{1, \ldots, n\}$ the value obtained from $P(\bar{x})$ by putting $x_i = 1$ if index i is in the j-combination, or $x_i = 0$ if is not:

$$\begin{aligned} G_j &= \sum_{1 \le i_1 < \cdots < i_j \le n} \left(\sum_{i \in \{i_1, \ldots, i_j\}} \frac{1}{n} \right)^m \\ &= \binom{n}{j} \left(\frac{j}{n} \right)^m \end{aligned}$$

and, thus, (52) results:

$$G(z) = \sum_{j=0}^{n} \binom{n}{j} \left(\frac{j}{n} \right)^m z^j (1-z)^{n-j}$$

Hence, the probability density function of X can be evaluated as:

$$\Pr[X = x] = [z^x] G(z)$$

which can be done as shown in Sec. III-C. The determination of the moments of X requires the evaluation of the derivatives of $G(\cdot)$, owing to Eq. (3). For instance, this yields:

$$\begin{aligned} \mathrm{E}[X^{\underline{1}}] &= n\, G_n - G_{n-1} \\ \mathrm{E}[X^{\underline{2}}] &= n(n-1)\, G_n - 2(n-1)\, G_{n-1} + 2\, G_{n-2} \end{aligned}$$

The higher the moment, the more its evaluation is complicated: the dependence of the moments on coefficients G_j looks less simple than their dependence (12) on the γ-transform values $\gamma(y)$. On the other hand, we can use our γ-transform theory to easily explicit such a dependence as follows. Owing to the relationship between $G(z)$ and $\gamma(y)$ shown in Sec. III-C, from the comparison between (16) and (52) we can derive the γ-transform as:

$$\gamma(y) = \frac{G_y}{\binom{n}{y}} \tag{53}$$

By substituting (53) in (12) we eventually obtain:

$$\begin{aligned} \mathrm{E}[X^{\underline{r}}] &= n^{\underline{r}} \sum_{l=0}^{r} (-1)^l \frac{\binom{r}{l}}{\binom{n}{l}} G_{n-l} \\ &= r! \sum_{l=0}^{r} (-1)^l \binom{n-l}{r-l} G_{n-l} \end{aligned} \tag{54}$$

which holds for any distribution with the probability generating function put in the form (52).

E. Bivariate generating function

An approach based on a *bivariate generating function* (BGF) [11, Sec. III.2], enumerating the set of possible allocations of objects into cells, can suitably be adopted to this problem. Using the variable z to "mark" the (undistinguishable) objects and the variable y to "mark" the (distinguishable) non-empty cells, we can use a BGF exponential with respect to z and ordinary with respect to x as follows:

$$\Phi(z, y) = (1 + y(e^z - 1))^n \tag{55}$$

Then, the probability density function and the factorial moments can then be computed from the BGF as follows:

$$\Pr[X = x] = \frac{[z^m y^x] \Phi(z, y)}{[z^m] \Phi(z, 1)} \tag{56}$$

$$\mathrm{E}[X^{\underline{r}}] = \frac{[z^m] \left. \frac{\partial^r \Phi(z, y)}{\partial y^r} \right|_{y=1}}{[z^m] \Phi(z, 1)} \tag{57}$$

leading, for instance, to:

$$\mathrm{E}[X] = \frac{[z^m] \left. \frac{\partial \Phi(z, y)}{\partial y} \right|_{y=1}}{[z^m] \Phi(z, 1)} \tag{58}$$

$$\sigma_X^2 = \frac{[z^m] \left. \frac{\partial^2 \Phi(z, y)}{\partial y^2} \right|_{y=1}}{[z^m] \Phi(z, 1)} + \mathrm{E}[X] - \mathrm{E}[X]^2 \tag{59}$$

Although this method is as general as the γ-transform approach, it is apparently more complex to apply even to a simple problem like that considered in this Appendix. First, the abstraction required to correctly write the underlying BGF (55) is more sophisticated than the derivation of the γ-transform based on its physical interpretation. Second, the evaluation of the derivatives of Φ and the rewritings required to evidence the coefficients of z^m and y^x are more complex operations than the application of the formulae made available by the γ-transform approach, as recalled in the section that follows.

F. γ-transform

After identifying the problem as a (finite) counting of trials, we can easily follow the γ-transform approach. We can formalize the problem as composed of m independent assignments of objects to cells, where $\psi_k(y) = y$ trivially counts the number of ways to select the cell for the assignment of an object, assuming that only y cells are available. Hence, (19) yields $\gamma(y) = (y/n)^m$. Expressing the probability density function is then straightforward using (6). The determination of the moments is also straightforward, since (12) can be used. Ready-to-use formulae (13)–(14) can also be employed for $\mathrm{E}[X]$ and σ_X^2.

ACKNOWLEDGMENT

The author wish to thank Paolo Ciaccia for all the helpful discussions had during the early development of this work.

REFERENCES

[1] D. Aktug and F. Can, "Analysis of multiterm queries in a dynamic signature file organization," in 'em Proc. of 16th ACM-SIGIR Intl' Conf., Pittsburgh, PA, 1993, pp. 96–105.

[2] Z. Berényi and I. Vajk, "A probabilistic model for compact document topic representation," in *Proc. of 9th WSEAS-SMO Intl' Conf.*, Budapest, Hungary, 2009, pp. 322–327.

[3] D. Bitton and D.J. DeWitt, "Duplicate record elimination in large data files," *ACM Transactions on Database Systems*, Vol. 8, No. 2, 1983, pp. 255–265.

[4] A.F. Cárdenas, "Analysis and performance of inverted database structures," *Communications of the ACM*, Vol. 18, No. 5, 1975, pp. 253–263.

[5] P. Ciaccia, D. Maio and P. Tiberio, "A unifying approach to evaluating block accesses in database organizations," *Information Processing Letters*, Vol. 28, No. 5, 1988, pp. 253–257.

[6] P. Ciaccia, "Block access estimation for clustered data," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 5, No. 4, 1993, pp. 712–718.

[7] P. Ciaccia and D. Maio, "Access cost estimation for physical database design," *Data & Knowledge Engineering*, Vol. 11, No. 2, 1993, pp. 125–150.

[8] M. Drmota, D. Gardy and B. Gittenberger, "A unified presentation of some urn models," *Algorithmica*, Vol. 29, No. 1/2, 2001, pp. 120–147.

[9] C. Faloutsos and S. Christodoulakis, "Signature files: an access method for documents and its analytical performance evaluation," *ACM Transactions on Information Systems*, Vol. 2, No. 4, 1984, pp. 267-288.

[10] S. Finkelstein, M. Schkolnick and P. Tiberio, "Physical database design for relational databases," *ACM Transactions on Database Systems*, Vol. 13, No. 1, 1988, pp. 91–128.

[11] P. Flajolet and R. Sedgewick, *Analytic Combinatorics*, Cambridge, UK: Cambridge University Press, 2009.

[12] D. Gardy and L. Némirovski, "Urn models and Yao's formula," in *Proc. of 7th ICDT Intl' Conf.*, Jerusalem, Israel, 1999, pp. 100–112.

[13] D. Gardy and C. Puech, "On the sizes of projections: a generating function approach," *Information Systems*, Vol. 9, No. 3/4, 1984, pp. 231–235.

[14] F. Grandi, "*Advanced access cost models for databases*," Ph.D. Dissertation, DEIS, University of Bologna, Italy, 1994, *in Italian*.

[15] F. Grandi and M.R. Scalas, "Block access estimation for clustered data using a finite LRU buffer," *IEEE Transactions on Software Engineering*, Vol. 19, No. 5, 1993, pp. 641–660.

[16] F. Grandi, On the signature weight in "multiple" m signature files, *ACM SIGIR Forum*, Vol. 29, No. 1, 1995, pp. 20–25.

[17] R.L. Graham, D.E. Knuth and O. Patashnik, *Concrete Mathematics*, Addison-Wesley, 1990.

[18] E. Hudnott, J. Sinclair and H. Darwen, "Rethinking database updates using a multiple assignment-based approach," *WSEAS Transactions on Computers*, Vol. 4, No. 9, 2010, pp. 392–405.

[19] I. Ioannidis, "Query optimization," *ACM Computing Surveys*, Vol. 28, No. 1, 1996, pp. 121-123.

[20] A. S. Mamaghani, K. Asghari, F. Mahmoudi and M. R. Meybodi, "A novel hybrid algorithm for join ordering problem in database queries," in *Proc. of 6th WSEAS-CIMMACS Intl' Conf.*, Tenerife, Spain, 2007, pp. 104–109.

[21] E. S. Murphree and D. Aktug, "Derivation of probability distribution of the weight of the query signature," Department of Mathematics and Statistics, Miami University, Oxford, Ohio, Tech. Rep., 1992.

[22] J. Riordan, *An Introduction to Combinatorial Analysis*, New York: John Wiley & Sons, 1958.

[23] C.S. Roberts, "Partial match retrieval via the method of superimposed codes," *Proceedings of the IEEE*, Vol. 67, No. 12, 1979, pp. 1624–1642.

[24] S.B. Yao, "Approximating block accesses in database organizations," *Communications of the ACM*, Vol. 20, No. 4, 1977, pp. 260–261.

Fabio Grandi received from the University of Bologna, Italy, a Laurea degree cum Laude in electronics engineering in 1988 and a Ph.D. in electronics engineering and computer science in 1994.

From 1989 to 2002 he has worked at the CSITE center of the Italian National Research Council (CNR) in Bologna in the field of neural networks and temporal databases, initially supported by a CNR fellowship. In 1993 and 1994 he was an Adjunct Professor at the Universities of Ferrara, Italy, and Bologna. In the University of Bologna, he was with the Dept. of Electronics, Computers and Systems from 1994 to 1998 as Research Associate and as Associate Professor from 1998 to 2012, when he joined the Dept. of Computer Science and Engineering. He is currently an Associate Professor of Information Systems in the School of Engineering and Architecture of the University of Bologna. His scientific interests include temporal, evolution and versioning aspects in data management, WWW and Semantic Web, knowledge representation, storage structures and access cost models.