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

FABIO GRANDI
Alma Mater Studiorum – Università di Bologna
Department of Computer Science and Engineering
Viale Risorgimento 2, I-40136 Bologna BO
ITALY
fabio.grandi@unibo.it

Abstract: A general method that can be used for the study of a discrete and finite random variable is presented. The method is based on the introduction of a transform of the probability density function, called γ-transform. A formula for computing the factorial moments directly from the γ-transform is derived. Moreover, it is shown how the γ-transform can be simply derived owing to its physical meaning for several combinatorial problems. Examples and applications relevant for computer science are provided.

Key–Words: Discrete probability, factorial moments, transforms, combinatorial analysis, estimation

1 Introduction

Several modeling problems relevant for performance evaluation of information processing and retrieval systems [1, 2, 3, 6, 7, 13, 14] 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 be handled 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* [12]. The determination of the moments from such distributions is not straightforward; even the evaluation of the variance may result in a very hard task.

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

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

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

$$\begin{aligned} f(x) &= [z^x]G(z) && (2)\\ \mathrm{E}[X^{\underline{r}}] &= G^{(r)}(1) && (3) \end{aligned}$$

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.

One way to prove formula (3) is through Taylor series expansions. Since

$$\begin{aligned} f(x) &= \frac{G^{(x)}(0)}{x!}\\ &= \frac{1}{x!}\sum_{j\geq 0}\frac{(-z)^j}{j!}G^{(x+j)}(z) && (4) \end{aligned}$$

we have:

$$\begin{aligned} \mathrm{E}[X^{\underline{r}}] &= \sum_{x\geq r}\frac{x^{\underline{r}}}{x!}\sum_{j\geq 0}\frac{(-z)^j}{j!}G^{(x+j)}(z)\\ &= \sum_{i\geq 0}\frac{1}{i!}\sum_{j\geq 0}\frac{(-z)^j}{j!}G^{(r+i+j)}(z)\\ &= \sum_{i\geq 0}\frac{G^{(r+i)}(0)}{i!} \;=\; G^{(r)}(1) && (5) \end{aligned}$$

Although the probability generating function approach is a very general methodology, we put forward the claim that it might not be the most convenient when dealing with a *finite* random variable, that 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 [10] (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 show the practical consequences that arise from it.

Our alternative approach is based on the introduction in Section 2 of a new transform, called γ-transform, which we defined in [7] 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 3: owing to a combinatorial identity demonstrated in Sec. 2, 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. Examples and outstanding applications are presented in Sections 4 and 5. Conclusions can finally be found in Section 6.

2 Preliminaries

2.1 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:

$$\gamma(y) \;=\; \sum_{x=0}^{n} \frac{\binom{y}{x}}{\binom{n}{x}}\, f(x) \qquad (6)$$

2.2 Antitransformation formula

The corresponding *inversion formula* is given by:

$$f(x) \;=\; \binom{n}{x} \sum_{j=0}^{x} (-1)^j \binom{x}{j} \gamma(x-j) \qquad (7)$$

and can be demonstrated as follows. It can be observed from (6) that $\gamma(y)$ is a polynomial function of degree n in y and, thus, it can be expressed as a finite Newton series:

$$\gamma(y) \;=\; \sum_{x=0}^{n} \binom{y}{x} \Delta^x \gamma(0) \qquad (8)$$

Comparing (6) with (8) yields:

$$f(x) \;=\; \binom{n}{x} \Delta^x \gamma(0) \qquad (9)$$

Eq. (7) can easily be obtained from (9) when expliciting the x-th difference.

2.3 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:*

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

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

Proof: Owing to the definition of the r-th difference, the right-hand side of (10) can be rewritten as:

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

In the above, the first equality is obtained by computing $\Delta^r \gamma(n-r)$ from Eq. (8). The final expression (11) equals the left-hand side of (10), thanks to Eq. (9). □

In order to support our claim, it can be noticed how (7) and (10) can actually represent finite calculus's counterpart of (4) and (5), respectively.

3 Probabilistic interpretation

3.1 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 2 *Given a discrete random variable X with values in $\{0, 1, \ldots, n\}$, its r-th factorial moment is provided by:*

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

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

Proof: It immediately follows from the definition of the 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:

$$\begin{aligned} \mathrm{E}[X] &= n\,[1-\gamma(n-1)] && (13)\\ \sigma_X^2 &= n^2\left[\gamma(n-2)-\gamma^2(n-1)\right] \\ &\quad + n\,[\gamma(n-1)-\gamma(n-2)] && (14) \end{aligned}$$

which are very simple formulae.

3.2 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 3 *Let X be a random variable with values in $\{0,1,\ldots,n\}$ and probability density function $f(x)$. X 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 $\mathcal{Y}$. Then it can be shown that:*

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

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:

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

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:

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

□

Moreover, also the inversion formula (7) 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 (7).

3.3 Relationship with $G(z)$

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

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

In order to prove it, it is sufficient to show that the density function (7) 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:

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

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:

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

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

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

we can write:

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

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 (6): since

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

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). All the other formulae concerning $G(z)$ can also be obtained from the corresponding ones concerning $\gamma(y)$ by taking the same limit. This is another point in favour of our initial claim.

4 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.

4.1 Uniform distribution

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

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

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 [10].

Applying Corollary 2 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 [10] can be used in the last step.

4.2 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:

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

owing to the binomial Theorem.

Applying Corollary 2 we easily obtain:

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

4.3 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 2 we obtain:

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

which is the value usually found in the literature.

4.4 Beta-binomial distribution

If X is a beta-binomial random variable:

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

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

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

In the above, the summation is a hypergeometric which can be evaluated as a Vandermonde' s convolution [10] (also the one in the next paragraph).

The factorial moments of X can be computed as:

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

5 Applications

The utility of the γ-transform approach lies in the fact that some estimation problems can be described by "complex" distributions which do in fact 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. This class of problems includes several modeling and estimation problems relevant for performance evaluation of information retrieval and database management systems, which are briefly referenced and analyzed in this Section.

In general, 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. Furthermore, if the experiment considered is composed of m independent subexperiments, $\gamma(y)$ 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 subexperiment be effected by selecting the successes only in a subset of y trials (which is also independent of k if the subexperiments are indistinguishable). In this case, 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 $\psi_k(\cdot)$.

5.1 Set union problem

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

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 subexperiments. The γ-transform of the probability density function of X can be derived 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:

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

Therefore, the probability density function of X is:

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

By means of Corollary 2, we can easily derive the expected value and the variance of X as:

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

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

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 word to be “or”-ed. 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 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 [9]. It was also noticed that the method sketched in [1] and developed in [11] through Markov chains and heavy matrix manipulations lead 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 have shown in [9].

An interesting case also arises when $s_k = s$ for each k (the subexperiments are indistinguishable), and X represents the number of “1” bits in the more “classical” superimposed codes adopted for information retrieval [13]. The density function and the expected value of X which can be derived in this way agree with those presented in [13].

Moreover, if $s = 1$ then X represents the number of distinct objects selected in sampling with replacement m objects from a population of n. For example, X may represent 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. The expected value which derives from (21) agrees with Cárdenas’ formula [3]. For an expression of the underlying density function see, for instance, [4, 6]. A comparison of the γ-transform approach to this simple problem with alternative methods (namely combinatorial calculus, the principle of inclusion and exclusion, generating functions and Markov chains) can be found in [7]. 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.

5.2 Group inclusion problem

An even more general problem with important applications to information processing is described in the following. Let $\mathcal{Q}$ be a set with cardinality q composed of n groups of objects, each of size g (namely $q = g\,n$). 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 $\mathcal{Q}$ with cardinality s_k. From another point of view, X is the number of distinct elements in a random subset of a *multiset* in which all the n distinct objects appear g times.

In this case, Eq. (19) can still be used with $\psi_k(y) = \binom{g\,y}{s_k}$, yielding:

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

and, thus,

$$f(x) = \binom{n}{x}\sum_{j=0}^{x}(-1)^j\binom{x}{j}\prod_{k=1}^{m}\frac{\binom{g(x-j)}{s_k}}{\binom{g\,n}{s_k}} \tag{23}$$

$$\mathrm{E}[X] = n\left[1-\prod_{k=1}^{m}\frac{\binom{q-g}{s_k}}{\binom{q}{s_k}}\right] \tag{24}$$

$$\begin{aligned}\sigma_X^2 = {}& n^2\left[\prod_{k=1}^{m}\frac{\binom{q-2g}{s_k}}{\binom{q}{s_k}}-\prod_{k=1}^{m}\frac{\binom{q-g}{s_k}^2}{\binom{q}{s_k}^2}\right]+\\ &n\left[\prod_{k=1}^{m}\frac{\binom{q-g}{s_k}}{\binom{q}{s_k}}-\prod_{k=1}^{m}\frac{\binom{q-2g}{s_k}}{\binom{q}{s_k}}\right]\end{aligned} \tag{25}$$

An interesting case takes place when $m = 1$ and X represents the number of blocks accessed in a file (with a total number of n blocks) during the retrieval of s_1 distinct records. The expected value agrees with Yao's formula [14]. Derivations of the distribution of X in this case can be found, for instance, in [2, 4, 6].

5.3 Yet another cell visit problem

Let us finally consider an application that cannot be reduced to an inclusion-exclusion problem but that can effectively be described though in terms of the γ-transform. Assume we have D distinct objects distributed into n cells, with the constraint that each cell contains exactly d distinct objects ($D \leq d\, n$). Let X be the random variable counting the number of all the cells which contain at least one of m distinct objects randomly selected out of D. For example, X represents the number of blocks accessed in a file (composed of n blocks) during the retrieval of m distinct data values in the presence of data duplication and of *uniform clustering* of the data [8]. Under these hypotheses, d represents the number of distinct data values contained in any block. This problem can be described as an experiment in which trials correspond to cells, and successes to cells to be visited. Therefore, $\gamma(y)$ represents the probability that $n - y$ of the cells have been excluded *a priori* from the result. Once these cells have been fixed, each of them has the same probability of being excluded from the result, which can be evaluated as

$$q = \frac{\binom{D-d}{m}}{\binom{D}{m}} \tag{26}$$

if the m objects are distinct, and

$$q = \left(1 - \frac{d}{D}\right)^m \tag{27}$$

if they are not. In both cases, the γ-transform of the density function has the form

$$\gamma(y) = q^{n-y}$$

and represents a particular case of binomial distribution. Notice that, in such a case, the experiment can be considered to be composed of n (indistinguishable) Bernoulli trials, each of which produces a success or a failure. The outcomes are in this case the successes of the repeated trials. Thus, $\gamma(y)$ can be expressed as the probability that a fixed subset (with cardinality $n - y$) of the n trials always lead to a failure (being q the probability of a failure). Hence, for a binomial distribution, the transform $\gamma(y)$ has a feasible expression of the form $\psi(y)/\psi(n)$ only if $p = q = 1/2$. In such a case we have $\psi(y) = 2^y$, since in each of the y independent selections of an outcome to be included in the result we have exactly two choices with the same probability: to choose a success (outcome included) or to choose a failure (outcome excluded). When $p \neq q$, we could formally define $\psi(y)$ as $(1/q)^y$, but this represents a non-feasible number of events ($1/q$ can be an irrational number as well). On the other hand, if we consider an experiment which can be described by means of the principle of inclusion and exclusion, it is always possible to find out a physical meaning of $\psi(y)$, by virtue of Theorem 3.

In particular, if the m objects are distinct, the probability density function for our cell visit problem from (26) becomes:

$$\begin{aligned}
&\Pr[X = x] \\
&= \binom{n}{x} \sum_{j=0}^{x} (-1)^j \binom{x}{j} \left[\frac{\binom{D-d}{m}}{\binom{D}{m}}\right]^{n-x+j} \\
&= \binom{n}{x} \left[1 - \frac{\binom{D-d}{m}}{\binom{D}{m}}\right]^{x} \left[\frac{\binom{D-d}{m}}{\binom{D}{m}}\right]^{n-x}
\end{aligned} \tag{28}$$

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

$$\mathrm{E}[X] = n \left[1 - \frac{\binom{D-d}{m}}{\binom{D}{m}}\right] \tag{29}$$

$$\sigma_X^2 = n \frac{\binom{D-d}{m}}{\binom{D}{m}} \left[1 - \frac{\binom{D-d}{m}}{\binom{D}{m}}\right] \tag{30}$$

Else, if the m objects are not distinct, the probability density function from (27) becomes:

$$\begin{aligned}
&\Pr[X = x] \\
&= \binom{n}{x} \sum_{j=0}^{x} (-1)^j \binom{x}{j} \left(1 - \frac{d}{D}\right)^{m(n-x+j)} \\
&= \binom{n}{x} \left[1 - \left(1 - \frac{d}{D}\right)^m\right]^x \left(1 - \frac{d}{D}\right)^{m(n-x)}
\end{aligned} \tag{31}$$

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

$$\mathrm{E}[X] = n\left[\left(1-\frac{d}{D}\right)^m\right] \tag{32}$$

$$\sigma_X^2 = n\left(1-\frac{d}{D}\right)^m\left[1-\left(1-\frac{d}{D}\right)^m\right] \tag{33}$$

In both cases, the expected values agree with those derived in [5, 8].

6 Conclusion

In this paper, we 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 random variable is finite. In such a case, 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 (involving derivatives and formally infinite summations) with an approach based on finite Newton series (involving finite summations and differences), 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 shown how the expression of the γ-transform can be simply derived owing to its physical meaning for several combinatorial problems. Several examples of its useful application to modeling problems relevant for performance evaluation of information processing and retrieval systems were provided, showing how the γ-transform approach looks really attractive in such domains.

Acknowledgements: We thank Paolo Ciaccia for all the helpful discussions had during the development of this work.

References:

[1] D. Aktug and F. Can, Analysis of Multiterm Queries in a Dynamic Signature File Organization, Proc. of 16th ACM-SIGIR Intl' Conf., 1993, pp. 96–105.

[2] 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.

[3] A.F. Cárdenas, Analysis and Performance of Inverted Database Structures, *Communications of the ACM*, Vol.18, No.5, 1975, pp. 253–263.

[4] 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.

[5] P. Ciaccia, Block Access Estimation for Clustered Data, *IEEE Transactions on Knowledge and Data Engineering*, Vol.5, No.4, 1993, pp. 712–718.

[6] 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.

[7] F. Grandi, *Advanced Access Cost Models for Databases*, Ph. D. Thesis, University of Bologna, Italy, 1994, *in Italian*.

[8] 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.

[9] F. Grandi, On the Signature Weight in "multiple" m Signature Files, *ACM SIGIR Forum*, Vol.29, No.1, 1995, pp. 20–25.

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

[11] E. S. Murphree and D. Aktug, Derivation of Probability Distribution of the Weight of the Query Signature, preprint, Department of Mathematics and Statistics, Miami University, Oxford, Ohio, 1992.

[12] J. Riordan, *An Introduction to Combinatorial Analysis*, John Wiley & Sons, 1958.

[13] C.S. Roberts, Partial Match Retrieval via the Method of Superimposed Codes, *Proceedings of the IEEE*, Vol.67, No.12, 1979, pp. 1624–1642.

[14] S.B. Yao, Approximating Block Accesses in Database Organizations, *Communications of the ACM*, Vol.20, No.4, 1977, pp. 260–261.

Errata Corrigenda

Erratum (at page 19)

$$\begin{aligned} f(x) &= \frac{G^{(x)}(0)}{x!} \\ &= \frac{1}{x!}\sum_{j\geq 0}\frac{(-z)^j}{j!}G^{(x+j)}(z) \end{aligned} \qquad (4)$$

we have:

$$\begin{aligned} \mathrm{E}[X^{\underline{r}}] &= \sum_{x\geq r}\frac{x^{\underline{r}}}{x!}\sum_{j\geq 0}\frac{(-z)^j}{j!}G^{(x+j)}(z) \\ &= \sum_{i\geq 0}\frac{1}{i!}\sum_{j\geq 0}\frac{(-z)^j}{j!}G^{(r+i+j)}(z) \\ &= \sum_{i\geq 0}\frac{G^{(r+i)}(0)}{i!} \;=\; G^{(r)}(1) \end{aligned} \qquad (5)$$

Correction

$$\begin{aligned} f(x) &= \frac{G^{(x)}(0)}{x!} \\ &= \frac{1}{x!}\sum_{j\geq 0}\frac{z^j}{j!}G^{(x+j)}(z) \end{aligned} \qquad (4)$$

we have:

$$\begin{aligned} \mathrm{E}[X^{\underline{r}}] &= \sum_{x\geq r}\frac{x^{\underline{r}}}{x!}\sum_{j\geq 0}\frac{z^j}{j!}G^{(x+j)}(z) \\ &= \sum_{i\geq 0}\frac{1}{i!}\sum_{j\geq 0}\frac{z^j}{j!}G^{(r+i+j)}(z) \\ &= \sum_{i\geq 0}\frac{G^{(r+i)}(0)}{i!} \;=\; G^{(r)}(1) \end{aligned} \qquad (5)$$

Erratum (at page 25)

Assume we have D distinct objects distributed into n cells, with the constraint that each cell contains exactly d distinct objects ($D \leq d\ n$). Let X be the random variable counting the number of all the cells which contain at least one of m distinct objects randomly selected out of D.

Correction

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 ($N \geq dn$). Let X be the random variable counting the number of all the cells which contain at least one representative of m distinct object types randomly selected out of D.

Erratum (at page 26)

$$\mathrm{E}[X] = n\left[\left(1-\frac{d}{D}\right)^m\right] \qquad (32)$$

Correction

$$\mathrm{E}[X] = n\left[1-\left(1-\frac{d}{D}\right)^m\right] \qquad (32)$$