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

Fabio Grandi

Department of Computer Science and Engineering (DISI)
Alma Mater Studiorum – University of Bologna, Italy
fabio.grandi@unibo.it

ASM 2014 — Florence, Italy — November 22–24, 2014

Outline of the Talk

- Introduction and Motivation
- The γ-transform Theory
  - Definitions and properties
  - Probabilistic interpretation and physical meaning
  - Connection with probability generating function
- Examples
- Applications
- Conclusion

Introduction and Motivation (1)

A common method for studying a discrete r.v. X defined in $\{0, 1, 2, \ldots\}$ with p.d.f. $f(x)$ is through the *probability generating function*:

$$G(z) \;=\; \sum_{x \geq 0} z^x \, f(x)$$

In fact, being $G^{(r)}(z) = \sum_{k \geq r} k^{\underline{r}} \, z^{k-r} \, f(k)$ (where $k^{\underline{r}}$ is the r-th falling factorial power of k), all the factorial moments of X can easily be derived from $G(z)$ as:

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

and the p.d.f. can be reconstructed via the inversion formula:

$$f(x) \;=\; [z^x]G(z) \;=\; \frac{G^{(x)}(0)}{x!}$$

Introduction and Motivation (2)

In several cases of interest for data management:

- we are interested in the estimation of some characteristic values via the evaluation of the moments (e.g., $E[X]$ and σ_X^2) of a r.v.

Introduction and Motivation (2)

In several cases of interest for data management:

- we are interested in the estimation of some characteristic values via the evaluation of the moments (e.g., $\mathrm{E}[X]$ and σ_X^2) of a r.v.
- the r.v. under study is limited ($X \in \{0, 1, \dots, n\}$)

Introduction and Motivation (2)

In several cases of interest for data management:

- we are interested in the estimation of some characteristic values via the evaluation of the moments (e.g., $E[X]$ and σ_X^2) of a r.v.
- the r.v. under study is limited ($X \in \{0, 1, \ldots, n\}$)
- $f(x)$ has a complex expression that can be very difficult to determine

Introduction and Motivation (2)

In several cases of interest for data management:

- we are interested in the estimation of some characteristic values via the evaluation of the moments (e.g., $\mathrm{E}[X]$ and σ_X^2) of a r.v.
- the r.v. under study is limited ($X \in \{0, 1, \ldots, n\}$)
- $f(x)$ has a complex expression that can be very difficult to determine
- $G(z)$ has no "physical meaning" and, thus, cannot be directly derived from the nature of the problem

Introduction and Motivation (2)

In several cases of interest for data management:

- we are interested in the estimation of some characteristic values via the evaluation of the moments (e.g., $\mathrm{E}[X]$ and σ_X^2) of a r.v.
- the r.v. under study is limited ($X \in \{0, 1, \ldots, n\}$)
- $f(x)$ has a complex expression that can be very difficult to determine
- $G(z)$ has no "physical meaning" and, thus, cannot be directly derived from the nature of the problem
- the moments are usually not easy to compute from $f(x)$ or $G(z)$

Introduction and Motivation (2)

In several cases of interest for data management:

- we are interested in the estimation of some characteristic values via the evaluation of the moments (e.g., $\mathrm{E}[X]$ and σ_X^2) of a r.v.
- the r.v. under study is limited ($X \in \{0, 1, \ldots, n\}$)
- $f(x)$ has a complex expression that can be very difficult to determine
- $G(z)$ has no "physical meaning" and, thus, cannot be directly derived from the nature of the problem
- the moments are usually not easy to compute from $f(x)$ or $G(z)$

Introduction and Motivation (2)

In several cases of interest for data management:

- we are interested in the estimation of some characteristic values via the evaluation of the moments (e.g., $\mathrm{E}[X]$ and σ_X^2) of a r.v.
- the r.v. under study is limited ($X \in \{0, 1, \ldots, n\}$)
- $f(x)$ has a complex expression that can be very difficult to determine
- $G(z)$ has no "physical meaning" and, thus, cannot be directly derived from the nature of the problem
- the moments are usually not easy to compute from $f(x)$ or $G(z)$

Hence, we are looking for a more handy approach, better suited to a *finite* discrete r.v.

Introduction and Motivation (3)

In particular,

$$\mathrm{E}[X^{\underline{r}}] \;=\; G^{(r)}(1) \;=\; \sum_{i \geq 0} \frac{G^{(r+i)}(0)}{i!}$$

is formally an infinite Taylor (McLaurin) series involving derivatives.

Introduction and Motivation (3)

In particular,

$$E[X^{\underline{r}}] = G^{(r)}(1) = \sum_{i \geq 0} \frac{G^{(r+i)}(0)}{i!}$$

is formally an infinite Taylor (McLaurin) series involving derivatives.

We would rather use a different approach exploiting the finiteness of X, based indeed on a finite Newton series involving finite differences.

Introduction and Motivation (3)

In particular,

$$E[X^r] \;=\; G^{(r)}(1) \;=\; \sum_{i \geq 0} \frac{G^{(r+i)}(0)}{i!}$$

is formally an infinite Taylor (McLaurin) series involving derivatives.

We would rather use a different approach exploiting the finiteness of X, based indeed on a finite Newton series involving finite differences.

Claim

The γ-transform approach is our proposed solution of such a kind

The γ-transform — Transformation Formula

The γ-transform of a function is defined by the following transformation formula:

Definition

Let $f(\cdot)$ be a fixed function defined in the discrete domain $\{0, 1, \ldots, n\}$

The γ-transform *of $f(\cdot)$ can be defined in $\{0, 1, \ldots, n\}$ as:*

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

The γ-transform — Anti-transformation Formula

The *inversion formula* for the γ-*transform* is given by:

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

By definition, $\gamma(y)$ is a polynomial function of degree n in y and, thus, it can be expressed as a finite Newton series:

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

Hence, by comparison with the definition of $\gamma(y)$ we obtain:

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

The anti-transformation formula follows by expliciting the x-th difference.

The γ-transform — A Combinatorial Identity (1)

A fundamental identity involving the γ-transform is the subject of the following Theorem:

Theorem

If $f(\cdot)$ is a fixed function defined in $\{0, 1, \ldots, n\}$ and $\gamma(\cdot)$ is its γ-transform, then the following combinatorial identity holds:

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

The γ-transform — A Combinatorial Identity (2)

Proof Owing to the definition of the r-th difference, the right-hand side of the identity to be proved can be rewritten as:

$$n^{\underline{r}}\,\Delta^r\,\gamma(n-r)$$

Then we can compute $\Delta^r\,\gamma(n-r)$ from $\gamma(y)=\sum_{x=0}^{n}\binom{y}{x}\Delta^x\,\gamma(0)$ and, thus, $\Delta^r\,\gamma(y)=\sum_{x=0}^{n}\binom{y}{x-r}\Delta^x\,\gamma(0)$, yielding:

$$\sum_{x=0}^{n} n^{\underline{r}}\binom{n-r}{x-r}\Delta^x\,\gamma(0)$$

Since $n^{\underline{r}}\binom{n-r}{x-r}=x^{\underline{r}}\binom{n}{x}$ and $f(x)=\binom{n}{x}\Delta^x\,\gamma(0)$, this equals the left-hand side of the identity to be proved

The γ-transform — Probabilistic Interpretation

Corollary

Given a discrete r.v. X with values in $\{0, 1, \dots, n\}$ and probability density function $f(x)$, its r-th factorial moment is provided by:

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

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

Proof It immediately follows from the previous Theorem and from the definition of expected value

The γ-transform — Evaluation of the Moments

Thanks to the previous Corollary, and since

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

where $\left\{ {r \atop s} \right\}$ is a Stirling number of the second kind, all the standard moments of a discrete and finite r.v. can *easily* be computed from the γ-transform of the density function.

Example

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

The γ-transform — Physical Meaning (1)

Let X be a r.v. with values in $\{0, 1, \ldots, n\}$ and p.d.f. $f(x)$, representing 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 $\mathcal{N}$.

Theorem

If $\mathcal{Y} \subseteq \mathcal{N}$ is a subset of trials fixed before the experiment and $\Pr[\mathcal{Y}]$ is the probability that the experiment be effected as if the successes could only be selected from $\mathcal{Y}$, then

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

where $\gamma(\cdot)$ is the γ-transform of $f(\cdot)$ and $y = |\mathcal{Y}|$

The γ-transform — Physical Meaning (2)

Proof Since the experiment can provide any number $X \in \{0, 1, \ldots, n\}$ of successes, $\Pr[\mathcal{Y}]$ can be expressed via the total probability Theorem:

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

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

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

The γ-transform — Physical Meaning (3)

Also the inversion formula can be derived with probabilistic arguments.

Let $\Pr[\mathcal{X}']$ be the probability that the successful trials only be selected in $\mathcal{X}'$, then by the principle of inclusion and exclusion we have:

$$\Pr[X = x] = \sum_{\substack{\mathcal{X} \subseteq \mathcal{N} \\ |\mathcal{X}| = x}} \left(\Pr[\mathcal{X}] - \sum_{\substack{\mathcal{X}' \subseteq \mathcal{X} \\ |\mathcal{X}'| = x-1}} \Pr[\mathcal{X}'] + \cdots \right.$$

$$\left. \cdots + (-1)^{x-1} \sum_{\substack{\mathcal{X}' \subseteq \mathcal{X} \\ |\mathcal{X}'| = 1}} \Pr[\mathcal{X}'] + (-1)^x \Pr[\emptyset] \right)$$

$$= \sum_{\substack{\mathcal{X} \subseteq \mathcal{N} \\ |\mathcal{X}| = x}} \sum_{j=0}^{x} (-1)^j \sum_{\substack{\mathcal{J} \subseteq \mathcal{X} \\ |\mathcal{J}| = j}} \Pr[\mathcal{X} \setminus \mathcal{J}]$$

The γ-transform — Physical Meaning (4)

Owing to the physical meaning of $\gamma(\cdot)$, $\Pr[\mathcal{X} \setminus \mathcal{J}] = \gamma(x-j)$ and, thus

$$\begin{aligned} \Pr[X = x] &= \sum_{\substack{\mathcal{X} \subseteq \mathcal{N} \\ |\mathcal{X}| = x}} \sum_{j=0}^{x} (-1)^j \sum_{\substack{\mathcal{J} \subseteq \mathcal{X} \\ |\mathcal{J}| = j}} \gamma(x-j) \\ &= \binom{n}{x} \sum_{j=0}^{x} (-1)^j \binom{x}{j} \gamma(x-j) \end{aligned}$$

(since trials are indistinguishable, summations reduce to counts of equal quantities)

The γ-transform — Relationship with $G(z)$ (1)

The probability generating function $G(z) = \mathrm{E}[z^X]$ can be expressed in terms of the γ-transform as follows

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

To prove it, we can show that the p.d.f. can be derived from the expression above as $f(x) = [z^x]G(z)$. By means of the binomial Theorem and with simple manipulations, it 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.

The γ-transform — Relationship with $G(z)$ (2)

Also an inverse relationship can be derived as follows. From:

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

we can extract $\gamma(y)$ or $\gamma(n-y)$ as

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

(the choice depends on the constraint $\gamma(n) = 1$)

The γ-transform — Relationship with $G(z)$ (3)

The approach based on $G(z)$ can be derived as a limit of the γ-transform theory when the discrete r.v. involved becomes *unlimited*. For instance, in the $\gamma(y)$ definition, since

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

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

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

Also other formulae concerning $G(z)$ can be obtained from the corresponding ones concerning $\gamma(y)$ by taking the same limit.

Summary Comparison Between the Approaches

p.g.f.	**γ-transform**

Summary Comparison Between the Approaches

p.g.f.	**γ-transform**
X discrete and infinite	X discrete and finite

Summary Comparison Between the Approaches

p.g.f.	**γ-transform**
X discrete and infinite	X discrete and finite
$G(z) \;=\; \sum_{x\geq 0} z^x \, f(x)$	$\gamma(y) \;=\; \sum_{x=0}^{n} \binom{y}{x} / \binom{n}{x} \, f(x)$

Summary Comparison Between the Approaches

p.g.f.	**γ-transform**
X discrete and infinite	X discrete and finite
$G(z) = \sum_{x \geq 0} z^x \, f(x)$	$\gamma(y) = \sum_{x=0}^{n} \binom{y}{x} / \binom{n}{x} \, f(x)$
$f(x) = \frac{1}{x!} G^{(x)}(0)$	$f(x) = \binom{n}{x} \Delta^x \, \gamma(0)$

Summary Comparison Between the Approaches

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

Summary Comparison Between the Approaches

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

Remark

We can say that the γ-transform plays the role of a "finite counterpart" of the probability generating function

Examples — Uniform Distribution

Let X be a discrete r.v. uniformly distributed in $\{0, 1, \ldots, n\}$:

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

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

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

Hence, factorial moments can be computed as:

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

Examples — Binomial Distribution

Let X be a discrete r.v. following a binomial distribution in $\{0, 1, \ldots, n\}$:

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

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

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

Hence, factorial moments can be computed as:

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

Examples — Hypergeometric Distribution

Let X be a discrete r.v. with a hypergeometric distribution in $\{0, 1, \ldots, n\}$:

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

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

$$\gamma(y) = \sum_{x=0}^{n} \binom{y}{x}\binom{N-n}{k-x} \Big/ \binom{N}{k} = \binom{y+N-n}{k} \Big/ \binom{N}{k}$$

Hence, factorial moments can be computed as:

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

Examples — Beta-binomial Distribution (1)

Let X be a discrete r.v. with a beta-binomial distribution in $\{0, 1, \ldots, n\}$:

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

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

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

Examples — Beta-binomial Distribution (2)

Hence, factorial moments can be computed as:

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

Application to Estimation Problems (1)

Some estimation problems involving a "complex" p.d.f. in fact have a simple γ-transform

If the underlying experiment is composed of m independent subexperiments, $\gamma(y)$ can be expressed as:

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

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

$\gamma_k(y)$ is also independent of k if the subexperiments are indistinguishable

Application to Estimation Problems (2)

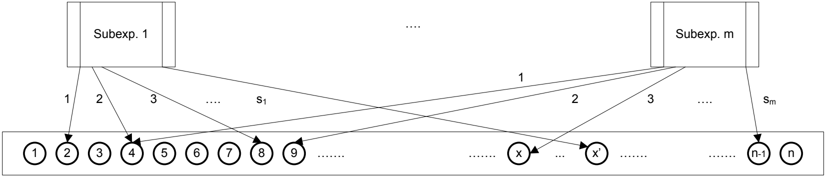
Being $\psi_k(y)$ the number of ways in which the k-th subexperiment can be effected by selecting the successes only in a subset of y trials, $\gamma(y)$ can conveniently be expressed as:

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

Hence, the solution of estimation problems involving the probabilistic characterization of some experiment (i.e., determination of the p.d.f. and moments of a r.v. X measuring the experiment results) reduces to the determination of the counting of events $\psi_k(y)$

Applications — Set Union Problem (1)

Let $\mathcal{N}$ be a set with cardinality n, let $\mathcal{S}_k$ $(1 \leq k \leq m)$ 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 k-th subexperiment does random sampling without replacement of s_k objects from $\mathcal{N}$ into $\mathcal{S}_k$. Sampling is with replacement between different subexperiments. X is the number of distinct objects altogether selected during the m subexperiments.

Applications — Set Union Problem (1)

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

Being the inclusion in $\mathcal{U}$ of an element of $\mathcal{N}$ a successful trial, the selections of the subsets $\mathcal{S}_1, \ldots, \mathcal{S}_m$ can be regarded as mutually independent subexperiments. Hence $\psi_k(y) = \binom{y}{s_k}$ 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:

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

Applications — Set Union Problem (2)

Hence, the p.d.f., expected value and variance of X can easily be computed from $\gamma(y)$:

$$
\begin{aligned}
f(x) &= \binom{n}{x} \sum_{j=0}^{x} (-1)^j \binom{x}{j} \prod_{k=1}^{m} \binom{x-j}{s_k} \Big/ \binom{n}{s_k} \\
\mathrm{E}[X] &= n \left[1 - \prod_{k=1}^{m} \left(1 - \frac{s_k}{n}\right) \right] \\
\sigma_X^2 &= n^2 \left[\prod_{k=1}^{m} \left(1 - \frac{s_k}{n}\right) \left(1 - \frac{s_k}{n-1}\right) - \prod_{k=1}^{m} \left(1 - \frac{s_k}{n}\right)^2 \right] + \\
& \quad n \left[\prod_{k=1}^{m} \left(1 - \frac{s_k}{n}\right) - \prod_{k=1}^{m} \left(1 - \frac{s_k}{n}\right) \left(1 - \frac{s_k}{n-1}\right) \right]
\end{aligned}
$$

Applications — Set Union Problem (3)

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 used for information retrieval. The p.d.f. and E[X] agree with those found by Aktug & Kan [1993] (as we showed in 1995).

If $s_k = s$ for each k (the subexperiments are indistinguishable), X represents the signature weight as generated by the more "classical" superimposed coding. The p.d.f. and E[X] agree with those found by Roberts [1979].

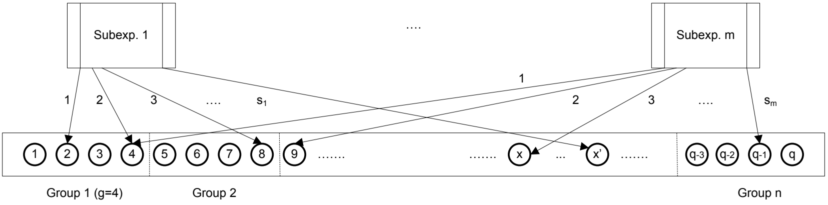
Applications — Set Union Problem (4)

If $s = 1$ then 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. $\mathrm{E}[X]$ agree with Cárdenas' formula and the p.d.f. with the expression derived by Gardy & Puech [1984] and Ciaccia, Maio & Tiberio [1988].

As far as we know, no expression had been derived for σ_X^2 before the introduction of the γ-transform theory.

Applications — Group Inclusion Problem (1)

Let $\mathcal{Q}$ be a set with cardinality q composed of n groups of objects, each of size g (namely $q = g\,n$), and X a r.v. denoting 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.



The k-th subexperiment does random sampling without replacement of s_k objects from $\mathcal{N}$ into $\mathcal{S}_k$. Sampling is with replacement between different subexperiments. X is the number of distinct groups from which objects are altogether selected during the m subexperiments.

Applications — Group Inclusion Problem (1)

Let $\mathcal{Q}$ be a set with cardinality q composed of n groups of objects, each of size g (namely $q = g\,n$), and X a r.v. denoting 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.

Being the inclusion in $\mathcal{U}$ of elements of a given group a successful trial, the selections of the subsets $\mathcal{S}_1, \ldots, \mathcal{S}_m$ can be regarded as mutually independent subexperiments. Hence $\psi_k(y) = \binom{g\,y}{s_k}$ is the number of ways in which the elements of $\mathcal{S}_k$ can be selected only from y groups, yielding:

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

Applications — Group Inclusion Problem (2)

Hence, the p.d.f., expected value and variance of X can easily be computed from $\gamma(y)$:

$$
\begin{aligned}
f(x) &= \binom{n}{x} \sum_{j=0}^{x} (-1)^j \binom{x}{j} \prod_{k=1}^{m} \binom{g(x-j)}{s_k} \Big/ \binom{g\,n}{s_k} \\
\mathrm{E}[X] &= n \left[1 - \prod_{k=1}^{m} \binom{q-g}{s_k} \Big/ \binom{q}{s_k} \right] \\
\sigma_X^2 &= n^2 \left[\prod_{k=1}^{m} \binom{q-2g}{s_k} \Big/ \binom{q}{s_k} - \prod_{k=1}^{m} \binom{q-g}{s_k}^2 \Big/ \binom{q}{s_k}^2 \right] + \\
& \quad n \left[\prod_{k=1}^{m} \binom{q-g}{s_k} \Big/ \binom{q}{s_k} - \prod_{k=1}^{m} \binom{q-2g}{s_k} \Big/ \binom{q}{s_k} \right]
\end{aligned}
$$

Applications — Group Inclusion Problem (3)

If $m = 1$, 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 p.d.f. agrees with expressions derived by Bitton & DeWitt [1983], Gardy & Puech [1984] and Ciaccia, Maio & Tiberio [1988]. E[X] agrees with Yao's formula [1977].

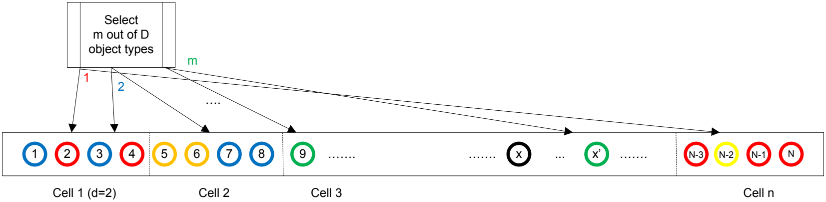
As far as we know, no expression had been derived for σ_X^2 before the introduction of the γ-transform theory.

In general, the Group Inclusion Problem is equivalent to the estimation of data access costs via an (unclustered) index scan for the retrieval of all the records matching m distinct values, if pointers are unioned before accessing data.

As far as we know, no exact models for the general problem have been proposed before the introduction of the γ-transform theory.

Applications — Another Cell Visit Problem (1)

Assume we have D distinct object types distributed into n cells, with the constraint that each cell contains representatives of exactly d distinct object types. A cell can contain more objects of the same type (total number of objects $N \geq d\, n$). Let X be the r.v. counting the number of cells which contain at least one representative of m distinct object types randomly selected out of D.



Applications — Another Cell Visit Problem (1)

Assume we have D distinct object types distributed into n cells, with the constraint that each cell contains representatives of exactly d distinct object types. A cell can contain more objects of the same type (total number of objects $N \geq d\,n$). Let X be the r.v. counting the number of cells which contain at least one representative of m distinct object types randomly selected out of D.

Thus, $\gamma(y)$ represents the probability that $n - y$ fixed cells have been excluded *a priori* from the result. Each of them has the same probability of being excluded from the result, which can be evaluated as $\binom{D-d}{m}/\binom{D}{m}$ yielding:

$$\gamma(y) \;=\; \left[\frac{\binom{D-d}{m}}{\binom{D}{m}}\right]^{n-y}$$

Applications — Another Cell Visit Problem (2)

Hence, the p.d.f., expected value and variance of X can easily be computed from $\gamma(y)$:

$$
\begin{aligned}
f(x) &= \binom{n}{x} \sum_{j=0}^{x} (-1)^j \binom{x}{j} \left[\binom{D-d}{m} \Big/ \binom{D}{m} \right]^{n-x+j} \\
\mathrm{E}[X] &= n \left[1 - \binom{D-d}{m} \Big/ \binom{D}{m} \right] \\
\sigma_X^2 &= n \binom{D-d}{m} \Big/ \binom{D}{m} \left[1 - \binom{D-d}{m} \Big/ \binom{D}{m} \right]
\end{aligned}
$$

Applications — Another Cell Visit Problem (3)

In case the m object types randomly selected out of D might be *non distinct* (i.e., sampling is with replacement), the probability of a cell to be excluded from the result can be evaluated as $(1 - d/D)^m$ yielding:

$$\gamma(y) \;=\; \left(1 - \frac{d}{D}\right)^{m(n-y)}$$

Applications — Another Cell Visit Problem (4)

Hence, the p.d.f., expected value and variance of X can easily be computed from $\gamma(y)$:

$$
\begin{aligned}
f(x) &= \binom{n}{x} \sum_{j=0}^{x} (-1)^j \binom{x}{j} \left(1 - \frac{d}{D}\right)^{m(n-x+j)} \\
\mathrm{E}[X] &= n \left[1 - \left(1 - \frac{d}{D}\right)^m\right] \\
\sigma_X^2 &= n \left(1 - \frac{d}{D}\right)^m \left[1 - \left(1 - \frac{d}{D}\right)^m\right]
\end{aligned}
$$

Applications — Another Cell Visit Problem (5)

X may represent 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, where d represents the number of distinct values contained in any block.

Both in the case of distinct and non distinct values, $\mathrm{E}[X]$ agree with those derived by Ciaccia [1993] and Grandi & Scalas [1993].

No expressions for the p.d.f. and σ_X^2 have been proposed before the introduction of the γ-transform theory (but can be determined in a simple way as a particular case of Binomial distribution).

Conclusion

- We presented the γ-transform approach as a tool for the study of a discrete and finite r.v.

Conclusion

- We presented the γ-transform approach as a tool for the study of a discrete and finite r.v.
- The approach is more handy than using the p.g.f. which does not exploit the finiteness of the r.v.

Conclusion

- We presented the γ-transform approach as a tool for the study of a discrete and finite r.v.
- The approach is more handy than using the p.g.f. which does not exploit the finiteness of the r.v.
- Owing to its physical meaning, the γ-transform can be directly derived from the nature of the problem

Conclusion

- We presented the γ-transform approach as a tool for the study of a discrete and finite r.v.
- The approach is more handy than using the p.g.f. which does not exploit the finiteness of the r.v.
- Owing to its physical meaning, the γ-transform can be directly derived from the nature of the problem
- Several modeling problems of interest for performance evaluation of database management and information retrieval systems involve a very complex p.d.f. but with a simple γ-transform

Conclusion

- We presented the γ-transform approach as a tool for the study of a discrete and finite r.v.
- The approach is more handy than using the p.g.f. which does not exploit the finiteness of the r.v.
- Owing to its physical meaning, the γ-transform can be directly derived from the nature of the problem
- Several modeling problems of interest for performance evaluation of database management and information retrieval systems involve a very complex p.d.f. but with a simple γ-transform
- In such cases, the γ-transform allows immediate determination of $\mathrm{E}[X]$ and σ_X^2 which are the most relevant modeling parameters