THE MATCHING PROBLEM WITH RANDOM DECKS AND THE EXACT ERROR OF THE POISSON APPROXIMATION

by

Anirban DasGupta
Purdue University

Technical Report #99-01

Department of Statistics
Purdue University
West Lafayette, IN USA

April 1999

*

THE MATCHING PROBLEM WITH RANDOM DECKS
AND THE EXACT ERROR OF THE POISSON APPROXIMATION

by

Anirban DasGupta*
Purdue University

Abstract

The number of fixed points in the matching problem is well known to have a limiting Poisson distribution with mean 1. Diaconis (1987) gives the upper bound $\frac{2^m}{m!}$ for the total variation distance where m is the length of the permutation.

We present exact integral representations, valid for every m, for the total variation distance as well as the CDF itself of the number of fixed points Z_m. These integral representations lead to sharper analytical bounds, series expansions, and very sharp numerical approximations. Furthermore, they provide a direct proof that the first m moments of Z_m and the Poisson (1) distribution exactly coincide and they also show an interesting sign-changing property for the Poisson approximation of the distribution of the number of fixed points.

This sign-change result establishes zero to be the mode of Z_m for even m and 1 as the mode for odd m, also an interesting oscillation property.

Finally, we also study the case of random decks. In particular, we show that the number of fixed points has exactly a Poisson distribution if the deck size m has a geometric distribution, and we ask how to guess the size of the decks if we are told the observed number of matches.

*

* Research supported by a grant from The National Security Agency.

1. Introduction.

This article is about the matching problem. The matching problem may be described in many different ways and is a popular example in the theory of probability. Mathematically, it can be formulated in the following way:

Let $m > 1$ be a fixed positive integer and let $\{\pi(1), \pi(2), \ldots, \pi(m)\}$ denote a random permutation of $\{1, 2, \ldots, m\}$ in the sense of being selected from a uniform distribution on the set of $m!$ permutations of $\{1, 2, \ldots, m\}$. Let Z_m denote the number of fixed points of π, i.e.,

$$Z_m = \#i : \pi(i) = i. \tag{1}$$

The matching problem studies the random variable Z_m. It is well known that Z_m converges in law to a Poisson distribution with mean 1 as $m \to \infty$; furthermore, the Poisson approximation is very accurate even for small m. Evidence of this may be seen in Barbour et al, (1992), among a number of places. Somewhat more intriguing is the moment matching property; even the first m moments of Z_m are exactly equal to the corresponding moments of the Poisson (1) distribution. See Diaconis (1987). We show in Section 5 that even the exact distribution is Poisson if m is chosen according to a geometric distribution. We also ask how to guess the value of m if one is told how many matches were obtained. This seems like a fairly interesting question.

Bounds on the error of the Poisson approximation in the matching problem are readily available; they differ in their complexity. An appealing bound is given in Diaconis (1987): the total variation distance between the exact and the limiting distributions of Z_m is at most $\frac{2^m}{m!}$, for any m.

In this article, we first present an exact integral representation for $P(Z_m \leq k)$ for every m and $k < m$. Using this, we also present an exact integral representation for the total variation distance between Z_m and the Poisson distribution with mean 1. Some interesting consequences of these exact representations are the following:

a. For any $m, P(Z_m \leq k) - P(Poi(1) \leq k)$ is alternatively negative and positive for any two successive values of $k, 0 \leq k < m$;

b. The same is true also for $P(Z_m = k) - P(Poi(1) = k)$;

c. The Poisson approximation to the distribution of Z_m satisfies several bounds; for instance,

   i. The total variation distance satisfies the sharper bound

$$d_{TV}(Z_m, Poi(1)) \leq \frac{2^m}{(m+1)!}$$

   for all $m > 1$, and

   ii. The Kolmogorov distance satisfies the bound

$$\sup_{k \geq 0} \; |P(Z_m \leq k) - P(Poi(1) \leq k)| \; \leq c(m)\frac{2^m}{(m+1)!},$$

where the constant $c(m) < \frac{1}{2}$ for all $m \geq 4$;

d. The total variation distance admits an explicit series expansion, the first few terms of which give a surprisingly accurate approximation essentially for any m:

$$d_{TV}(Z_m, Poi(1)) \approx \frac{2^m}{(m+1)!} \cdot \frac{m^4 + 12m^3 + 51m^2 + 88m + 56}{m^4 + 14m^3 + 71m^2 + 154m + 120}; \quad (2)$$

e. The most likely number of fixed points is alternatively 0 or 1. In fact, the mode of Z_m is 0 for even m and 1 for odd m;

f. The first m moments of Z_m and the Poisson (1) distribution coincide; this is known, but we give a direct proof using the integral representation for $P(Z_m \leq k)$ itself.

The integral representation for $P(Z_m \leq k)$ comes from a careful calculation. The calculation is heavy at times and uses certain facts about Hypergeometric and incomplete gamma functions; but at the end, things fall into place and a simple exact representation arises. We attempted to find a simpler derivation of the final answer avoiding the Hypergeometric functions, but did not succeed. The very special facts about the Hypergeometric and incomplete gamma functions used in this derivation are stated at the beginning for an easier flow of the derivation. The literature on the matching problem is rich and old; one may see Feller (1973) and Montmort (1980) for an entry and a historical perspective.

2. Poisson Approximation of the CDF.

We state the integral representation first.

2.1. Exact Expression

Theorem 1. Let $0 \leq k < m$ be a fixed integer. Then

$$P(Z_m \leq k) - P(Poi(1) \leq k) = \frac{(-1)^{m-k}}{k!(m-k-1)!} \int_0^1 t^{-m-1}(1-t)^{m-k-1}\gamma(m+1,t)dt, \quad (3)$$

where $\gamma(m+1,t)$ is the incomplete gamma function $\int_0^t e^{-v}v^m dv$.

2.2. Certain Technical Facts

The derivation of the integral representation (3) requires a number of technical facts. These are stated below.

Lemma 1.

Let p^{F_q} denote the general hypergeometric series and

$$\Gamma(\alpha,x) = \int_x^\infty e^{-t}t^{\alpha-1}dt, \gamma(\alpha,x) = \int_0^x e^{-t}t^{\alpha-1}dt$$

the incomplete gamma functions. Then

$$\frac{\Gamma(1+j,-1)}{ej!} = \sum_{i=0}^{j} \frac{(-1)^i}{i!} \tag{4}$$

$$\sum_{j=0}^{m-k-1} \binom{m}{j} t^j = (1+t)^m - t^{m-k} \binom{m}{k} \;\; {}_2F_1(1,-k;m-k+1;-t) \tag{5}$$

$$\int_0^\infty e^{-t} t^{m-k} \;\; {}_2F_1(1,-k;m-k+1;-t)dt = (m-k)! \int_0^\infty e^{-t}(1+t)^k dt \tag{6}$$

$$\int_0^1 e^t t^{m-k} \;\; {}_2F_1 \; (1,-k;m-k+1;t)dt = \frac{e}{m+1} \;\; {}_2F_2(1,m+1;m-k+1;m+2;-1) \tag{7}$$

$${}_2F_2(1,m+1;m-k+1;m+2;-1) = (m-k) \int_0^1 (1-t)^{m-k-1} \;\; {}_1F_1(m+1;m+2;-t)dt \tag{8}$$

$${}_1F_1(m+1;m+2;-t) = \frac{m+1}{2^{m+1}} e^{-\frac{t}{2}} \int_{-1}^{1} (1+u)^m e^{-\frac{t}{2}u} du \tag{9}$$

$$P(Poi(\lambda) > k) = \frac{\gamma(k+1,\lambda)}{k!} \tag{10}$$

Remark.

In the above, (4), (5), (6), (7) and (9) may be seen in Gradshteyn and Ryzhik (1980), (8) may be seen in Luke (1975), and (10) is well known.

2.3. Proof of the Integral Representation.

We will now present a derivation of the integral representation in (3).

Proof of Theorem 1:

Step 1.

By using the usual formula for the probability of the occurrence of exactly j of m events,

$$P(Z_m = j) = \frac{1}{j!} \sum_{i=0}^{m-j} \frac{(-1)^i}{i!} \tag{11}$$

Step 2.

Hence, by a change of variable, and (4),

$$P(Z_m > k) = \sum_{j=0}^{m-k-1} \frac{1}{(m-j)!} \sum_{i=0}^{j} \frac{(-1)^i}{i!}$$

$$= \frac{1}{e} \sum_{j=0}^{m-k-1} \frac{1}{j!(m-j)!} \int_{-1}^{\infty} e^{-t} t^j \, dt \tag{12}$$

Step 3.

From (12) and (5),

$$P(Z_m > k) = \frac{1}{em!} \int_{-1}^{\infty} e^{-t} \{(1+t)^m - \binom{m}{k} t^{m-k} \;\; {}_2F_1(1, -k; m-k+1; -t)\} dt$$

$$= 1 - \frac{1}{ek!(m-k)!} \{ \int_{0}^{\infty} e^{-t} t^{m-k} \;\; {}_2F_1(1, -k; m-k+1; -t) dt$$

$$+ \int_{-1}^{0} e^{-t} t^{m-k} \;\; {}_2F_1(1, -k, m-k+1; -t) dt \} \tag{13}$$

Step 4.

By substituting successively (6), (7), and (8) into (13),

$$P(Z_m > k) = 1 - \frac{1}{ek!} \int_{0}^{\infty} e^{-t} (1+t)^k dt + \frac{(-1)^{m-k+1}}{k!(m-k-1)!(m+1)}$$

$$\int_{0}^{1} (1-t)^{m-k-1} \;\; {}_1F_1(m+1; m+2; -t) dt \tag{14}$$

Step 5.

By (9) and Fubini's theorem,

$$\int_{0}^{1} (1-t)^{m-k-1} \;\; {}_1F_1(m+1; m+2; -t) dt = \frac{m+1}{2^{m+1}} \int_{0}^{1} (1-t)^{m-k-1} e^{-\frac{t}{2}} (\int_{-1}^{1} (1+u)^m e^{-\frac{t}{2}u} \, du) dt$$

$$= \frac{m+1}{2^{m+1}} \int_0^1 (1-t)^{m-k-1} \left(\int_0^2 e^{-\frac{t}{2}x} x^m dx \right) dt$$

$$= (m+1) \int_0^1 t^{-m-1}(1-t)^{m-k-1} \gamma(m+1,t) dt \tag{15}$$

Step 6

Therefore, from (14) and (10),

$$P(Z_m > k) = P(Poi(1) > k) + \frac{(-1)^{m-k+1}}{k!(m-k-1)!} \int_0^1 t^{-m-1}(1-t)^{m-k-1} \gamma(m+1,t) dt,$$

on writing

$$\frac{1}{ek!} \int_0^\infty e^{-t}(1+t)^k dt = \frac{1}{k!} \int_1^\infty e^{-t} t^k dt \quad = 1 - \frac{\gamma(k+1,1)}{k!} \tag{16}$$

Formula (3) follows.

Corollary 1.

If m is even, $P(Z_m \leq k)$ is overestimated by $P(Poi(1) \leq k)$ for all odd $k < m$ and underestimated by $P(Poi(1) \leq k)$ for all even $k < m$. If m is odd, exactly the opposite is true.

Proof: Follows from (3).

2.4. Bounds on the Error: Kolmogorov Distance

The integral representation (3) in Theorem 1 is especially suitable for bounding the error of the Poisson approximation to $P(Z_m \leq k)$. A simple bound is presented below.

Theorem 2.

a.

$$\sup_{1 \leq k \leq m-1} |P(Z_m \leq k) - P(Poi(1) \leq k)| \leq c(m) \frac{2^m}{(m+1)!}, \tag{17}$$

where

$$c(m) = \frac{1}{\sqrt{2\pi}} \sqrt{\frac{m}{m-1}} \; e^{\frac{1}{12m}}. \tag{18}$$

b. The Kolmogorov distance between Z_m and the Poisson (1) satisfies the same bound

$$\sup_{k\geq 0} |P(Z_m \leq k) - P(Poi(1) \leq k)| \leq c(m)\frac{2^m}{(m+1)!},$$

for all $m \geq 2$.

Proof:

a **Step 1.**

By (3), for $1 \leq k \leq m-1$ (and for $k = 0$ also),

$$\begin{aligned}|P(Z_m \leq k) - P(Poi(1) \leq k)| &= \frac{1}{k!(m-k-1)!}\int_0^1 (1-t)^{m-k-1} t^{-m-1}\gamma(m+1,t)dt. \\ &\leq \frac{1}{k!(m-k-1)!(m+1)}\int_0^1 (1-t)^{m-k-1}dt \\ &= \frac{1}{(m+1)k!(m-k)!} \\ &= \frac{m!}{k!(m-k)!}\frac{1}{(m+1)!}, \qquad (19)\end{aligned}$$

as $t^{-m-1}\gamma(m+1,t) \leq \frac{1}{m+1}$ for $0 < t < 1$.

Step 2.

By Stirling's approximation, for any integer j,

$$e^{-j} j^{j+\frac{1}{2}}\sqrt{2\pi} \leq j! \leq e^{-j} j^{j+\frac{1}{2}}\sqrt{2\pi}\;\; e^{\frac{1}{12j}}. \qquad (20)$$

From (19) and (20),

$$\begin{aligned}&|P(Z_m \leq k) - P(Poi(1) \leq k)| \\ &\leq \frac{\frac{1}{(m+1)!}\sqrt{m}\;\; e^{\frac{1}{12m}}}{\sqrt{2\pi}\;\; (\frac{m-k}{m})^{m-k}\;\; (\frac{k}{m})^k\;\; \sqrt{k(m-k)}} \qquad (21)\end{aligned}$$

By calculus, for $0 \leq x \leq 1, x^x(1-x)^{1-x} \geq \frac{1}{2}$, and hence $(\frac{m-k}{m})^{m-k}(\frac{k}{m})^k \geq 2^{-m}$. Substituting this into (21), one has

$$|P(Z_m \leq k) - P(Poi(1) \leq k)| \;\; \leq \frac{\sqrt{m}\;\; e^{\frac{1}{12m}}}{\sqrt{2\pi}\sqrt{k(m-k)}}\frac{2^m}{(m+1)!}, \qquad (22)$$

a local bound.

The uniform bound stated in the theorem follows from (22).

b We use part a to obtain part b

Step 1.

Clearly,

$$\sup_{k\geq 0} |P(Z_m \leq k) - P(Poi(1) \leq k)| = \max\{|P(Z_m = 0) - P(Poi(1) = 0)| \\ \sup_{1\leq k\leq m-1} |P(Z_m \leq k) - P(Poi(1) \leq k)| P(Poi(1) > m)\}. \quad (23)$$

Step 2.

By using (10), one obtains easily that

$$P(Poi(1) > m) \leq \frac{1}{(m+1)!}. \quad (24)$$

Step 3.

$P(Z_m = 0)$ equals $\sum_{j=0}^{m} \frac{(-1)^j}{j!}$, which may be written as $\frac{1}{e}(1 - \frac{\gamma(m+1,-1)}{m!})$ (pp 940 in Gradshteyn and Ryzhik (1980)).

Hence,

$$\begin{aligned} &|P(Z_m = 0) - P(Poi(1) = 0)| \\ &= \frac{1}{em!} \int_0^1 e^v v^m dv \\ &\leq \frac{1}{(m+1)!}. \end{aligned} \quad (25)$$

Step 4.

Combining (23), (24), and (25) with the result (17) in part a, part b now follows as $c(m)2^m \geq 1$ for all $m \geq 2$.

We present below a brief computation on the sharpness of the bound of Theorem 2 on the Kolmogorov distance. We denote the Kolmogorov distance by d_k.

Table 1

m	$d_k(Z_m, Poi(1))$	Bound in Theorem 2
5	.0113	.0201
6	.0032	.0056
8	.0002	.0003
10	5.46×10^{-6}	1.09×10^{-5}
20	3.22×10^{-15}	8.44×10^{-15}

3. The Variational Distance

3.1. Exact Expression

Akin to the representation (3), the variational distance

$$d_{TV}(Z_m, Poi(1)) = \sup_{B \subseteq Z_+} |P(Z_m \in B) - P(Poi(1) \in B)|$$

also admits an integral representation. In fact, formula (3) is the building block of this representation for the variational distance which we present below.

Theorem 3.

$$d_{TV}(Z_m, Poi(1)) = \frac{1}{m!} \int_0^1 e^{-t} t^m dt + \frac{1}{(m-1)!} \int_0^1 (2-t)^{m-1} t^{-m-1} \gamma(m+1, t) dt \tag{26}$$

Proof:

Step 1.

$d_{TV}(Z_m, Poi(1))$

$$= \frac{1}{2} \sum_{k=0}^{\infty} |P(Z_m = k) - P(Poi(1) = k)|$$
$$= \frac{1}{2} \sum_{k=0}^{m} |P(Z_m = k) - P(Poi(1) = k)|$$

$$+\frac{1}{2}P(Poi(1)>m)$$

$$=\frac{1}{2}\sum_{k=0}^{m-1}|P(Z_m=k)-P(Poi(1)=k)|$$

$$+\quad \frac{1}{2}\left(\frac{1}{m!}-\frac{e^{-1}}{m!}\right)+\frac{1}{2}\frac{\int_0^1 e^{-t}t^m dt}{m!} \tag{27}$$

Step 2.

By formula (3), for $0\le k\le m-1$,

$$P(Z_m=k)-P(Poi(1)=k)$$
$$=\{P(Z_m\le k)-P(Poi(1)\le k)\}$$
$$-\{P(Z_m\le k-1)-P(Poi(1)\le k-1)\}$$
$$=\frac{(-1)^{m-k}}{k!(m-k-1)!}\int_0^1 t^{-m-1}(1-t)^{m-k-1}\gamma(m+1,t)dt$$
$$+\frac{(-1)^{m-k}}{(k-1)!(m-k)!}\int_0^1 t^{-m-1}(1-t)^{m-k}\gamma(m+1,t)dt$$
$$=\frac{(-1)^{m-k}}{k!(m-k)!}\int_0^1 (m-kt)t^{-m-1}(1-t)^{m-k-1}\gamma(m+1,t)dt, \tag{28}$$

on a few lines of algebra.

Step 3.

Therefore,

$$\sum_{k=0}^{m-1}|P(Z_m=k)-P(Poi(1)=k)|$$
$$=\int_0^1\{mt^{-m-1}\gamma(m+1,t)\sum_{k=0}^{m-1}\frac{1}{k!(m-k)!}(1-t)^{m-k-1}$$
$$-t^{-m}\gamma(m+1,t)\sum_{k=0}^{m-1}\frac{k(1-t)^{m-k-1}}{k!(m-k)!}\}dt$$
$$=\frac{1}{(m-1)!}\int_0^1 t^{-m-1}\gamma(m+1,t)\{2(2-t)^{m-1}-1\}dt, \tag{29}$$

on some more algebra.

Step 4.

In (29), the term

$$\int_0^1 t^{-m-1}\gamma(m+1,t)dt = \frac{1}{m}\int_0^1 e^{-t}(1-t^m)dt \tag{30}$$

by Fubini's theorem.

Step 5.

If one now combines (27), (29), and (30), one gets the integral representation (26) of Theorem 3 on some cancellation.

Corollary 2.

For all $m \geq 2$,

$$d_{TV}(Z_m, Poi(1)) \leq \frac{2^m}{(m+1)!}. \tag{31}$$

Proof:

From (26), $d_{TV}(Z_m, Poi(1))$

$$\leq \frac{1}{m!}\int_0^1 t^m dt + \frac{1}{(m-1)!(m+1)}\int_0^1 (2-t)^{m-1}dt = \frac{2^m}{(m+1)!}.$$

The proof of Theorem 3 shows that the phenomenon of alternative under and overestimation for the CDF that was described in Corollary 1 is true for the individual probabilities as well. Here is the statement.

Corollary 3.

If m is even, $P(Z_m = k)$ is overestimated by $P(Poi(1) = k)$ for all odd $k < m$ and underestimated by $P(Poi(1) = k)$ for all even $k < m$. If m is odd, exactly the opposite is true.

Proof:

Follows from (28).

Corollary 3 and the analytical bound (28) lead to the following result on the most likely number of matchings, i.e., the mode of Z_m.

Corollary 4.

If m is even and ≥ 4, the mode of Z_m is 0; if m is odd and ≥ 5, the mode of Z_m is 1.

Proof:

Let m be even. Then, by Corollary 3,

$$P(Z_m = 0) > P(Poi(1) = 0) = e^{-1} = P(Poi(1) = 1) > P(Z_m = 1).$$

For $k \geq 2$,

$$P(Z_m = k) \leq |P(Z_m = k) - P(Poi(1) = k)| + P(Poi(1) = k) \leq \frac{2^m}{(m+1)!} + \frac{e^{-1}}{k!},$$

by bound (28).

But $\frac{e^{-1}}{k!} \leq \frac{e^{-1}}{2}$ for $k \geq 2$ and $\frac{2^m}{(m+1)!} < \frac{e^{-1}}{2}$ for $m \geq 4$, and so $P(Z_m = k) < e^{-1}$ for all $k \geq 2$. This shows that 0 is the mode of Z_m for all even $m \geq 4$. The argument for odd $m \geq 5$ is similar and so we omit it.

3.2. Series Expansion

Formula (26) leads to a useful series expansion for the total variation distance. The first term of the expansion is $\frac{2^m}{(m+1)!}$, and the first five terms provide an excellent approximation to the total variation distance essentially for all m. The expansion is given below.

Proposition 1.

$$d_{TV}(Z_m, Poi(1)) = \frac{2^m}{(m+1)!} - \frac{1}{m!}\sum_{j=0}^{\infty} \frac{(-1)^j}{j!(m+j+2)} c_j, \tag{32}$$

where

$$c_j = \int_1^2 x^m (2-x)^j dx.$$

Proof: Step 1.

In the exact formula (26) for the total variation distance, write e^{-t} as $\sum_{j=0}^{\infty} (-1)^j \frac{t^j}{j!}$ and $\gamma(m+1,t)$ as $\sum_{j=0}^{\infty} \frac{(-1)^j t^{m+1+j}}{j!(m+1+j)}$ (see pp 941 in Gradshteyn and Ryzhik (1980)).

Step 2.

Integrate each term in formula (26) term by term. This results in

$$d_{TV}(Z_m, Poi(1)) = \sum_{j=0}^{\infty} \frac{(-1)^j}{j!(m+j+1)(m-1)!} \left(\frac{1}{m} + \int_0^1 t^j (2-t)^{m-1} dt\right). \tag{33}$$

Step 3.

In the integral $\int_0^1 t^j(2-t)^{m-1}dt$, make the change of variable $x = 2-t$, and then do an integration by parts. This leads to (29) with the first term as $\frac{2^m}{(m+1)!}$; the calculation is omitted.

If one computes the first five terms in the series expansion (32), then the following highly accurate approximation obtains:

$$d_{TV}(Z_m, Poi(1)) \approx \frac{2^m}{(m+1)!} \frac{m^4 + 12m^3 + 51m^2 + 88m + 56}{m^4 + 14m^3 + 71m^2 + 154m + 120}. \tag{34}$$

We are now in a position to compare the exact variational distance in (26), the analytical bound $\frac{2^m}{(m+1)!}$ in (31), the approximation in (34), and the previously known bound $\frac{2^m}{m!}$ (e.g., stated in Diaconis (1987)). A short summary is presented below. The approximation (34) is very accurate even for $m = 3$.

Table 2

m	Approximation (34)	Exact Value (26)	Bound (31)	$\frac{2^m}{m!}$
3	.23492	.23747	.33333	1.3333
5	.03436	.03379	.04444	.26667
6	.01011	.00951	.01270	.08889
8	.00059	.00042	.00071	.00635
10	.00002	.000008	.00003	.00028

4. The Property of Matching Moments

The number of fixed points Z_m has the striking property that its first m moments coincide with those of the Poisson distribution with mean 1; the earliest reference we know of is Frobenius (1904). It is a positive feature of our integral representation (3) that it provides a direct proof of this moment matching property without requiring group theoretic methods. We present this below.

Theorem 4.

Let $X \sim Poi(1)$. Then for all $0 < r \leq m, E(Z_m^r) = E(X^r)$.

Step 1.

To prove that the first moments of X and Z_m match, it will be enough to show that

$$\sum_{k=0}^{\infty} P(Z_m > k) = \sum_{k=0}^{\infty} P(X > k);$$

likewise, to prove that the second moments match, it will be enough to show that

$$\sum_{k=0}^{\infty} kP(Z_m > k) = \sum_{k=0}^{\infty} kP(X > k).$$

Most generally, we shall need to show that for all $r \leq m$,

$$\sum_{k=0}^{\infty} k(k-1)\dots(k-r+2)P(Z_m > k) = \sum_{k=0}^{\infty} k(k-1)\dots(k-r+2)P(X > k). \tag{35}$$

Step 2.

Since $P(Z_m > k) = 0$ for $k \geq m$, (35) is the same as

$$\sum_{k=0}^{m-1} k(k-1)\dots(k-r+2)\{P(X > k) - P(Z_m > k)\}$$
$$= -\sum_{k=m}^{\infty} k(k-1)\dots(k-r+2)P(X > k). \tag{36}$$

Step 3.

By representation (3) in Theorem 1, the left side of (36) equals

$$\sum_{k=0}^{m-1} k(k-1)\dots(k-r+2)\frac{(-1)^{m-k}}{k!(m-k-1)!} \int_0^1 (1-t)^{m-k-1} \ t^{-m-1}\gamma(m+1,t)dt$$

$$= \sum_{k=0}^{m-1} \frac{(-1)^{m-k}}{(k-r+1)!(m-k-1)!} \int_0^1 (1-t)^{m-k-1} \ t^{-m-1}\gamma(m+1,t)dt$$

$$= \int_0^1 \{\sum_{k=0}^{m-1} \frac{(-1)^{m-k}}{(k-r+1)!(m-k-1)!}(1-t)^{m-k-1}\}t^{-m-1}\gamma(m+1,t)dt \tag{37}$$

$$= -\int_0^1 \frac{t^{m-r}}{(m-r)!}t^{-m-1}\gamma(m+1,t)dt$$

$$= -\frac{1}{(m-r)!}\int_0^1 \frac{\gamma(m+1,t)}{t^{r+1}}dt, \tag{38}$$

as the quantity within {} in (37) equals $-\frac{t^{m-r}}{(m-r)!}$.

Step 4.

On the other hand, by virtue of identity (10), the right side of (36) equals

$$-\sum_{k=m}^{\infty} k(k-1)\dots(k-r+2)\frac{\gamma(k+1,1)}{k!}$$

$$= -\sum_{k=m}^{\infty} \frac{1}{(k-r+1)!}\int_0^1 e^{-v}v^k dv$$

$$= -\sum_{k=m-r+1}^{\infty} \frac{1}{k!}\int_0^1 e^{-v}v^{k+r-1}dv$$

$$= -\int_0^1 v^{r-1}e^{-v}P(Poi(v) > m-r)dv$$

$$= -\frac{1}{(m-r)!}\int_0^1 v^{r-1}\gamma(m-r+1,v)dv, \tag{39}$$

again, due to the general identity (10) that $P(Poi(v) > s) = \frac{\gamma(s+1,v)}{s!}$ for any v, any s.

Step 5.

The expressions in (38) and (39) are equal by integration by parts and induction.

This establishes the moment matching property stated in Theorem 4.

5. The Case of Random Decks

Feller (1973, pp 112) gives some interesting variations of the matching problem, one of which considers the target deck (which would be $\{1, 2, \dots, m\}$ in our notation) as being constructed at random. Below we provide a collection of results for the case of random decks; a particular amusing result is that if in the standard matching problem m is chosen according to a Geometric distribution, then the number of matches has exactly a Poisson distribution.

5.1. Distribution of Fixed Points

Theorem 5.

a. Suppose, in the matching problem, the size of the decks m, i.e., the length of the permutation, is chosen according to a geometric (p) distribution with mass function $p(1-p)^m, m \geq 0$. Then the number of fixed points has a Poisson $(1-p)$ distribution;

b. Suppose the size of the decks m is chosen according to a uniform $\{1, 2, \dots, M\}$ distribution. Then the number of fixed points Z has the distribution

$$\begin{cases} P(Z = k) = \frac{M-k+2}{M} \cdot P(Z_{M+1} = k) + \frac{(-1)^{M-k}}{Mk!(M-k+1)!}, 1 \le k \le M; \\ P(Z = 0) = 1 - \sum\limits_{k=1}^{M} P(Z = k). \end{cases} \tag{40}$$

c. Suppose the size of the decks m is chosen according to a Poisson (λ) distribution. Then the number of fixed points Z has the distribution

$$P(Z = k) = \frac{1}{k!} \int_0^{\lambda} e^{-t} t^{\frac{k-1}{2}} J(k-1, 2\sqrt{t}) \ dt, \tag{41}$$

where J denotes the usual Bessel J function;

The special value $P(Z = 0)$ is tabulated below for geometric, Poisson, and Uniform decks with the same expected deck size $= 5, 10, 25, 50$. The limiting value is $e^{-1} = .36788$. Note how interestingly, the convergence to the limiting value is from opposite directions for uniform and geometric decks and that the Poisson deck converges the fastest. Also, note that the fixed points of a uniform deck behave, asymptotically, like the fixed points of a nonrandom deck with deck size $M + 1$; that is what equation (40) says. The proof of Theorem 5 uses the results in Section 2; we are going to omit the proof.

Table 3: P(# fixed points = 0)

Expected Deck Size	Uniform Decks	Poisson Decks	Geometric Decks
5	.315	.36708	.43460
10	.34146	.36788	.40289
25	.35731	.36788	.38230
50	.36260	.36788	.37516

5.2. Guessing the Deck Size from Number of Matches

An intellectually interesting question is to ask how the unknown size of the decks can be guessed if we are only told the observed number of matches. It is intuitively clear that it is not easy to accurately estimate the deck size given the number of matches. Still it would be interesting to see some explicit formulas and computations. For the case of geometric decks, one can obtain the following neat formula from part a of Theorem 5.

Corollary 5.

Suppose m, the size of the decks, is chosen according to the geometric (p) distribution with mass function $p(1-p)^m, m \geq 0$. Then the conditional expectation of m given that k matches were observed equals $k + \frac{(1-p)^2}{p}$.

For uniform or Poisson decks, there is no such simple formula. But still it would be interesting to see some numbers. We took the pre-data expected deck size to be 50 in the table below.

Table 4

E (Deck Size$|k$ matches were observed)

k	Uniform Decks	Geometric Decks
0	50.36	49.02
1	49.51	50.02
2	50.01	51.02
10	54.01	59.02
20	59.01	69.02

References

Barbour, A.D., Holst, Lars, and Janson, Svante (1992). Poisson Approximation, Clarendon Press, Oxford.

Diaconis, Persi (1987). Application of the methods of moments in probability and statistics, Symp. in Applied Mathematics, AMS, Vol. 37, 125–142.

Feller, William (1973). An Introduction of Probability Theory and Its Applications, Vol. 1, John Wiley, New York.

Frobenius, G. (1904). Uver die charaktere der metrfach transitiven gruppen, Berliner Bericht, 558–571.

Gradshteyn, I.S. and Ryzhik, I.M. (1980). Table of Integrals, Series, and Products, Academic Press, New York.

Luke, Yudell (1975). Mathematical Functions and Their Approximations, Academic Press, New York.

Montmort, P.R. (1980). Essay d'analyse sur les jeux de hazard, Chelsea, New York.