Asymptotics for k-fold repeats in the birthday problem with unequal probabilities. *

by Michael Camarri

Technical Report No. 524

Department of Statistics
University of California
367 Evans Hall # 3860
Berkeley, CA 94720-3860

July 10, 1998

Abstract

In a previous paper Camarri and Pitman studied the asymptotics for repeat times in random sampling by a method of Poisson embedding. Here we extend these results to k-fold repeats and also indicate the relationships between the repeat processes of various orders.

1 Introduction

The *birthday problem* in its classical form asks for the minimum sized group required so that the probability of at least one repeated birthday within the

*Research supported in part by N.S.F. Grants FD92-24857, FD94-04345, FD92-24868 and FD97-03691

group is greater than one half. The assumptions are a year of length 365 days, with each day equally likely as a birthday, and birthdays independent from person to person. It is well known that the answer is 23.

A number of generalisations of this problem have been studied. (See Diaconis and Mosteller [6] for a conversational summary of some of these approaches.) Here we are interested in the *sequential birthday problem* as stated by Camarri and Pitman [4]. They consider an iid sequence with a common discrete distribution p, on a finite or countable set S, and calculate the distributions of the times at which repeated values occur. Of interest also is the *repeat process*, that is, the point process (with the same indexing as the original iid sequence) that counts the number of repeats. Asymptotics for the (joint) distributions of the repeat times are calculated by considering the limit of the sequence of repeat processes generated by an underlying sequence of discrete distributions.

In the first part of this paper we generalise the results of [4] to k-fold repeats, that is, we are interested in the times by which values have occurred k or more times. The proofs have the same design as the corresponding results in [4] and we provide only a sketch of the changes.

Note that this k-fold birthday problem is a specialised *quota problem*. In this more general setting, each value j is assigned a quota v_j and we are interested in the times at which quotas are met (or exceeded.) Holst [7, 8] studied this problem using the same Poisson embedding and derived expressions for the moments of these general quota fulfilment times. Clearly the techniques of [4] can also be used to derive asymptotic distributions for the quota problem. Note further that martingale dynamics techniques such as those of Brown [2] and Barbour and Brown [1] can be used to find rates of convergence in these limit theorems. This in turn allows us to use the asymptotic distributions as approximations to the exact repeat process distributions and supplies total variation bound error estimates. See [3] for more comments on this.

Finally we investigate the structure between the various k-fold repeat processes. That is, we show how the form of one particular K-fold repeat process can determine the forms of the other k-fold repeat processes. We conclude with some examples that show that all possible limiting regimes and structures can be realised.

2 Results

Consider the following extension of the sequential birthday problem;
For each $n \geq 1$ let $(p_{ni}, i \geq 1)$ be a *ranked discrete distribution*, let $Y_{n0}, Y_{n1}, \ldots$ be an iid sequence with this common distribution and let

$$s_{nk} \equiv \left(\sum_i p_{ni}^k\right)^{1/k} \quad \text{and} \quad \theta_{ni}^{(k)} \equiv p_{ni}/s_{nk}.$$

Say that the sequence $(Y_{nj}, j \geq 0)$ has a *k-fold repeat* at time t if the value Y_t occurs k or more times in $(Y_{n0}, \ldots Y_{nt})$ and define $(R_{nm}^{(k)}, m \geq 1)$ to be the times of those k-fold repeats (in increasing order.)

We study the asymptotics of the $R_{nm}^{(k)}$ by embedding in a Poisson process (see Construction 7). The set of all possible limiting distributions is characterised by the following theorem.

Theorem 1 (i) *If $p_{n1} \to 0$ as $n \to \infty$ and $\theta_i^{(k)} \equiv \lim_n \theta_{ni}^{(k)}$ exists for each i, then for each $r \geq 0$*

$$\lim_{n\to\infty} P[s_n R_{n1}^{(k)} > r] = \prod_{i=1}^{\infty} \left(e^{-\theta_i^{(k)} r} \sum_{m=0}^{k-1} \frac{(\theta_i^{(k)})^m}{m!} \right) e^{-\left(1-\sum_{i=1}^{\infty} (\theta_i^{(k)})^k\right) r^k/k!} \tag{1}$$

(ii) *Conversely, if there are positive constants $c_n^{(k)} \to 0$ and $d_n^{(k)}$ such that the distribution of $c_n^{(k)}(R_{n1}^{(k)} - d_n^{(k)})$ has a non-degenerate weak limit as $n \to \infty$, then $p_{n1} \to 0$ and limits $\theta_i^{(k)}$ exist as in* (i)*, so the weak limit is just a rescaling of that described in* (i)*, with $c_n^{(k)}/s_n^{(k)} \to \alpha$ for some $0 < \alpha < \infty$, and $c_n^{(k)} d_n^{(k)} \to 0$.*

If $(s_n, n \in \mathbb{Z}^+)$ and $(t_n, n \in \mathbb{Z}^+)$ are two sequences of constants say that they are *strictly different* if $\lim_n s_n/t_n \in \{0, \infty\}$ and *similar* if $\lim_n s_n/t_n \in (0, \infty)$. We call the sequence $(s_{nk}, n \geq 1)$ the *natural scaling* for the process of k-fold repeats and define the *natural process of k-fold repeats* to be the counting process $W_n^{(k)} \equiv (W_n^{(k)}(t), t \geq 0)$ where

$$W_n^{(k)}(t) \equiv \sum_i 1(s_{nk} R_{nm}^{(k)} \leq t).$$

In the special case $\theta_1^{(k)} = 0$ the distribution in (1) simplifies to that of the first point of a Poisson process on $[0, \infty)$ of rate $t^{k-1}/(k-1)!$ at time t. More generally it can be seen that (1) is the distribution of the first point of the limiting process in the following theorem.

Theorem 2 *Under hypothesis (i) of Theorem 1, for all $m \geq 1$ as $n \to \infty$*

$$(s_{nk} R_{n1}^{(k)}, \ldots, s_{nk} R_{nm}^{(k)}) \xrightarrow{d} (S_1, \ldots S_m)$$

where $S_1 < S_2 < \ldots$ are the arrival times of the superposition of independent processes $\{M^*, M_1^{(-k)}, M_2^{(-k)}, \ldots\}$ where M^* is a Poisson process on $[0, \infty)$ of rate $(1 - \sum_{i=1}^{\infty} (\theta_{ki})^k) t^{k-1}/(k-1)!$ at time t, M_i is a Poisson process on $[0, \infty)$ of rate θ_{ki} and $M_i^{(-k)}$ denotes the process M_i with its first $k-1$ points removed.

If $\theta_1^{(k)}$ exists and is positive we say that the limit distribution is *mixed* and if it is zero we say that it is *atomless*. We refer to non-zero $\theta_i^{(k)}$ as *atoms* and say that a limit is *purely atomic* if all repeats correspond to atoms. (That is if $\sum_i (\theta_i^{(k)})^k = 1$.) Say that a k-fold repeat is *genuine* if it is not also a $(k+1)$-fold repeat and say that the natural asymptotic k-fold repeat process is *genuine* if (almost surely) it is composed entirely of genuine repeats. Our first structural result is the following.

Lemma 3 *If the natural limiting k-fold repeat process is atomless, then it is genuine.*

Thus we can think of our limit distributions as composed of two parts; the atoms which produce one genuine repeat each (at a time distributed as a gamma random variable) and then a string of non-genuine repeats (which form a homogeneous Poisson process); and an atomless section which produces only genuine repeats. Interestingly (and perhaps counter-intuitively), given a sequence of underlying discrete distributions at most one of the limiting k-fold repeat processes can be mixed, and this particular process determines the distributions of the others.

Theorem 4 *If the natural K-fold repeat process $W_n^{(K)}$ has a mixed limit (that is not purely atomic) then*
(i) $W_n^{(k)}$ converges weakly for all k, and has an atomless limit for $2 \leq k < K$ and a purely atomic limit for $k > K$.
(ii) The natural timescales of the k-fold repeat processes are all strictly different for $2 \leq k \leq K$ and for $k > K$ are all similar to s_{nK}.

Corollary 5 *If $W_n^{(K)}$ has an atomless limit then*
(i) $W_n^{(k)}$ converges weakly for all $2 \leq k \leq K$ and has an atomless limit.
(ii) The natural timescales of these processes are all strictly different.

Corollary 6 *If $W_n^{(K)}$ has a purely atomic limit then*
(i) $W_n^{(k)}$ converges weakly for all $k \geq K$ and has a purely atomic limit.
(ii) The natural timescales of these processes are all similar.

3 Proofs

3.1 Limit Theorems for k-fold repeats

Our framework is the following:
For each $n = 1, 2, \ldots$ let $(p_{ni}, i = 1, 2, \ldots)$ be a *ranked discrete distribution*, that is $p_{n1} \geq p_{n2} \geq \cdots \geq 0$ and $\sum_i p_{ni} = 1$, and suppose that $p_{n1} \to 0$ as $n \to \infty$. (Note: we do not assume that the supports of these distributions are finite.)

We embed our iid sequences in a Poisson process as follows

Construction 7 *Let N be a homogeneous Poisson process on $[0, \infty) \times [0, 1]$ with rate 1, with points $\{S_0, S_1, \ldots\}$ (ordered (a.s.) by their first co-ordinate) and define*

$$N(t) \equiv N([0, t] \times [0, 1]) \quad \textit{and} \quad N(t^-) \equiv N([0, t) \times [0, 1]).$$

For $n > 0$ partition $[0, 1]$ into intervals $I_{n1}, I_{n2}, \ldots$ such that the length of I_{ni} is p_{ni}. Let N_{ni} be N restricted to $[0, \infty) \times I_{ni}$. (Note that the N_{ni} are independent Poisson processes with rates p_{ni} respectively.) For $n > 0$, $i \geq 0$ define

$$Y_{ni} = \sum_j j 1(S_i \in [0, \infty) \times I_{nj})$$

Clearly the sequence $(Y_{n0}, Y_{n1}, \ldots)$ is iid with distribution $(p_{ni}, i = 1, 2, \ldots)$. For $k \geq 2$ let $(R_{nm}^{(k)}, m \geq 1)$ mark the k-fold repeats in this sequence and let $(T_{nm}^{(k)}, m \geq 1)$ be the corresponding times within N, that is

$$T_{nm}^{(k)} \equiv \inf\{t : N(t) > R_{nm}^{(k)}\}$$

or equivalently

$$N(T_{nm}^{(k)-}) = R_{nm}^{(k)}. \tag{2}$$

The use of this Poissonisation is justified by the following lemma.

Lemma 8 *If $p_{n1} \to 0$ as $n \to \infty$ then for all $m \geq 1, k \geq 2$*

$$\frac{R_{nm}^{(k)}}{T_{nm}^{(k)}} \xrightarrow{p} 1 \quad \text{as } n \to \infty.$$

Proof. By the Strong Law of Large Numbers, $N(t^-)/t \xrightarrow{p} 1$ as $t \to \infty$. It is enough to show that $T_{nm}^{(k)}$ converges in probability to infinity. However this follows from $T_{nm}^{(k)} \geq T_{n1}^{(2)}$ and the corresponding result in [4]. □

One main advantage of switching to this Poissonised timescale is that the arrival processes for each value are now independent of each other. In particular the distribution of $T_{n1}^{(k)}$ can be easily written down.

$$P[s_{nk} T_{n1}^{(k)} > r] = \prod_i \left(1 - e^{-\theta_{ni}^{(k)} r} \sum_{m=k}^{\infty} \frac{(\theta_{ni}^{(k)} r)^m}{m!} \right) \tag{3}$$

Simple Taylor series estimates show

$$\log P[s_{nk} T_{n1}^{(k)} > r] = r^k / k! + O(\theta_{n1}^{(k)})$$

and hence we have established

Lemma 9 *If $\theta_{n1}^{(k)} \to 0$ as $n \to \infty$ then*

$$\lim_{n \to \infty} P[s_{nk} R_{n1}^{(k)} > r] = e^{-r^k/k!} \text{ for } r \geq 0. \tag{4}$$

Note: this also follows from Lemma 10 below.

As a sketch of the proof of Theorem 1, (i) follows from Lemma 9 exactly as the corresponding result followed in [4] and (ii) again follows from a convergence of types argument and by noting that the righthand side of (3) (considered as a function of r) uniquely determines the constants $(\theta_{n1}^{(k)}, \theta_{n2}^{(k)}, \ldots)$.

For Theorem 2 we first prove Lemma 10 below and note that it is then straightforward to modify the arguments in [4] to complete the proof.

Lemma 10 *Let M_k be an inhomogeneous Poisson process on $[0, \infty)$ of rate $t^{k-1}/(k-1)!$ at time t and let $S_1 < S_2 < \ldots$ be its arrival times. If $\theta_{n1}^{(k)} \to 0$ as $n \to \infty$ then for all m as $n \to \infty$*

$$(s_{nk} R_{n1}^{(k)}, \ldots, s_{nk} R_{nm}^{(k)}) \xrightarrow{d} (S_1, \ldots S_m).$$

Proof. Consider counting processes $X_n^{(k)} \equiv (X_n^{(k)}(t), t \geq 0)$ where $X_n^{(k)}(t)$ tallies the number of repeats up to time t/s_{nk} , precisely

$$X_n^{(k)}(t) = \sum_{m=1}^{\infty} 1(T_{nm}^{(k)} \leq t/s_{nk}).$$

From the general theory of point processes (see for example Daley and Vere-Jones [5]) it is enough to show that the processes $X_n^{(k)}$ converge weakly to M_k. Further it is sufficient to show that the compensators of $X_n^{(k)}$ converge pointwise in probability to the compensator of M_k, namely the process $(t^k/k!, t \geq 0)$.

Let $N_{ni}^{(-k)}$ denote the process N_{ni} with its first $k-1$ points removed and let $N_{ni}^{(-k)}(t) \equiv N_{ni}^{(-k)}([0,t])$. Clearly

$$X_n^{(k)}(t) = \sum_i N_{ni}^{(-k)}(t/s_{nk}).$$

If we define our filtrations to be those generated by the natural filtrations of the processes $(N_{ni}(t/s_{nk}), t \geq 0)$ then $C_n^{(k)} \equiv (C_n^{(k)}(t), t \geq 0)$, the compensator of $X_n^{(k)}$, is given by

$$C_n^{(k)}(t) = \sum_i \theta_{ni}^{(k)}(t - s_{nk}T_{ni1}^{(k)})^+ \tag{5}$$

where $T_{ni1}^{(k)}$ is the time of the kth point of N_{ni} and thus $s_{nk}T_{ni1}^{(k)}$ has a Gamma distribution with parameters $(k-1, \theta_{ni}^{(k)})$. We complete the proof by showing

$$\left| EC_n^{(k)}(t) - t^k/k! \right| \leq \theta_{n1}^{(k)} t^{k+1}/(k-1)! \tag{6}$$

$$\operatorname{Var} C_n^{(k)}(t) \leq \frac{2\theta_{n1}^{(k)} t^{k+1}}{(k+1)!}[1 + k^2\theta_{n1}^{(k)}t]. \tag{7}$$

Let $Q_\theta(k)$ be the right tail of a Poisson(θ) random variable, that is

$$Q_\theta(k) = \sum_{m=k}^{\infty} \frac{e^{-\theta}\theta^m}{m!}.$$

We make use of the following bound

$$(1-\theta)\theta^k/k! \leq Q_\theta(k) \leq \theta^k/k! \tag{8}$$

The sum is bounded below by its first term so $e^{-\theta} \geq 1-\theta$ supplies the lower bound. Straightforward estimates of Taylor series remainders give

$$e^{\theta} - (1+\theta+\ldots+\theta^{k-1}/(k-1)!) \leq e^{\theta}\theta^{k}/k!$$

and multiplication by $e^{-\theta}$ establishes the upper bound.

Let T have a Gamma$(k-1,\theta)$ distribution. Note that

$$P[T \leq t] = Q_{\theta t}(k-1).$$

Now

$$\begin{aligned} E\left[\theta(t-T)^{+}\right] &= \theta \int_0^t \frac{(t-r)e^{-\theta r}\theta^{k-1}r^{k-2}}{(k-2)!}dr \\ &= \theta t Q_{\theta t}(k-1) - (k-1)Q_{\theta t}(k) \end{aligned}$$

and similarly

$$E\left[\theta^2\left((t-T)^{+}\right)^2\right] = \theta^2 t^2 Q_{\theta t}(k-1) - 2\theta t(k-1)Q_{\theta t}(k) + k(k-1)Q_{\theta t}(k+1).$$

The bounds in (8) and simple algebra yield

$$\frac{(\theta t)^k}{k!} - \frac{(\theta t)^{k+1}}{(k-1)!} \leq E\left[\theta(t-T)^{+}\right] \leq \frac{(\theta t)^k}{k!} + \frac{(k-1)(\theta t)^{k+1}}{k!}$$

$$\mathrm{Var}\left[\theta(t-T)^{+}\right] \leq \frac{2(\theta t)^{k+1}}{(k+1)!}[1+(k^2-1)\theta t].$$

Applying these to (5) and using independence gives (6) and (7). □

3.2 Structural results

Before proving the structural result that relate the k-fold repeat processes of different orders we show that atomless repeats are genuine.

Proof (of Lemma 3.) To show that atomless processes are genuine it is enough to show that the first non-genuine repeat is not among the first m repeats almost surely, that is $P_n[T_{nm}^{(k)} \leq T_{n1}^{(k+1)}] \to 1$ as $n \to \infty$.

Since $s_{nk}/s_{n,k+1} \to \infty$ we can find a sequence t_n such that $s_{nk}t_n \to \infty$ and $s_{n,k+1}t_n \to 0$. Then

$$\begin{aligned}
P_n[T^k_{nm} \le T^{(k+1)}_{n1}] &\ge P_n[T^{(k)}_{nm} \le t_n \le T^{(k+1)}_{n1}] \\
&\ge P_n[T^{(k)}_{nm} \le t_n] + P_n[T^{(k+1)}_{n1} \ge t_n] - 1 \\
&= P_n[s_{nk}T^{(k)}_{nm} \le s_{nk}t_n] - P_n[s_{n,k+1}T^{(k+1)}_{n1} \le s_{n,k+1}t_n] \\
&\to 1
\end{aligned}$$

□

Proof (of Theorem 4.) Assume that $(X^{(n)}_K, n \ge 1)$ has a mixed limit. As shown above, the atomless part of this limiting distribution produces (almost surely) no $(K+1)$-fold repeats whilst the atoms each produce an infinite string of $(K+1)$-fold repeats. Hence there exists a purely atomic $(K+1)$-fold repeat process that uses the scaling $(s_{nK}, n \ge 1)$. Theorem 1 (ii) implies that the natural scaling is similar to this. By induction we can extend to all $k \ge K$.

If $\theta^{(K-1)}_{n1}$ does not tend to zero then (by the usual arguments) we can find a subsequence $(m_n, n \ge 1)$ such that

$$\lim_{n\to\infty} \theta^{(K-1)}_{n_m i} \quad \text{exists for all } i$$

and so along this subsequence the natural $(K-1)$-fold repeat process has a mixed limit. However, by the above argument this implies that the K-fold repeat process has a purely atomic limit contradicting the assumptions of the theorem. Thus the natural $(K-1)$-fold processes have an atomless limit. We can extend this backwards to all $k < K$ by noting that

$$(\theta^{(k+1)}_{n1})^{k+1} \ge (\theta^{(k)}_{n1})^k.$$

That the natural timescales are strictly different follows from

$$(s_{n,k+1}/s_{nk})^k \ge 1/\theta^{(k-1)}_{n1}.$$

□

3.3 Examples

We finish with some examples that show that all possible limits and structures can be obtained. Note that in these examples the distributions do not

necessarily sum to 1, but this can be rectified by perturbing p_{n1} appropriately.
(i) $W_n^{(k)}$ *atomless for all* k
Take $(p_{ni}, i \geq 1)$ to be uniform on $[n]$ for all n.
(ii) $W_n^{(k)}$ *has a (non-purely atomic) mixed limit with a finite number of atoms*
To obtain atoms $\theta_1^{(k)}, \ldots \theta_j^{(k)}$ define

$$a_i \equiv \frac{\theta_j^{(k)}}{\left(1 - \sum_i (\theta_i^{(k)})^k\right)^{1/k}}$$

and let

$$p_{ni} = \frac{a_i}{n^{(k-1)/k}} \text{ for } 1 \leq i \leq j \text{ and } p_{ni} = \frac{1}{n} \text{ for } j+1 \leq i \leq \#_n$$

where

$$\#_n = \lfloor n(1 - \sum_{i=1}^{j} a_i n^{-(k-1)/k}) \rfloor.$$

(iii) $W_n^{(k)}$ *has a mixed limit with summable atoms*
If $\sum_i a_i < \infty$ then let

$$p_{ni} = \frac{a_i}{n^{(k-1)/k}} \text{ for } 1 \leq i \leq n \text{ and } p_{ni} = \frac{1}{n} \text{ for } n+1 \leq i \leq \#_n$$

where

$$\#_n = \lfloor n(1 - \sum_{i=1}^{n} a_i n^{-(k-1)/k}) \rfloor.$$

(iv) $W_n^{(k)}$ *has a mixed limit with non-summable atoms*
If $\sum_i a_i = \infty$ then let $(r_n, r \geq 1)$ be such that $r_n \uparrow \infty$ and

$$\sum_{i=1}^{r_n} a_i/n \to 0 \quad \text{as } n \to \infty.$$

Then take

$$p_{ni} = \frac{a_i}{n} \text{ for } 1 \leq i \leq r_n \text{ and } p_{ni} = \frac{1}{n^{k/(k-1)}} \text{ for } r_n + 1 \leq i \leq \#_n$$

where

$$\#_n = \lfloor n^{k/(k-1)}(1 - \sum_i a_i n^{-1}) \rfloor.$$

References

[1] A. D. Barbour and T. C. Brown. The Stein-Chen method, point processes and compensators. *Ann. Probab.*, 20(3):1504–1527, 1992.

[2] T. Brown. A martingale approach to the Poisson convergence of simple point processes. *Ann. Probab.*, 6(4):615–628, 1978.

[3] M. Camarri. *Asymptotics for repeat times in random sampling.* PhD thesis, U.C. Berkeley, May 1998.

[4] M. Camarri and J. Pitman. Limit distributions and random trees derived from the birthday problem with unequal probabilities. Technical Report 523, Dept. Statistics, U.C. Berkeley, 1998. Available via http://www.stat.berkeley.edu/users/pitman.

[5] D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes.* Springer Series in Statistics. Springer-Verlag, New York, 1988.

[6] P. Diaconis and F. Mosteller. Methods for studying coincidences. *The Journal of the American Statistical Association*, 84(408):853 – 861, December 1989.

[7] L. Holst. On birthday, collectors', occupancy and other classical urn problems. *International Statistical Review*, 54:15 – 27, 1986.

[8] L. Holst. The general birthday problem. In *Proceedings of the Sixth International Seminar on Random Graphs and Probabilistic Methods in Combinatorics and Computer Science, "Random Graphs '93" (Poznań, 1993)*, volume 6, pages 201–208, 1995.