EXTREME VALUE DISTRIBUTIONS FOR RANDOM COUPON COLLECTOR AND BIRTHDAY PROBLEMS

Lars Holst*
Royal Institute of Technology

September 6, 2000

Abstract

Take n independent copies of a strictly positive random variable X and divide each copy with the sum of the copies, thus obtaining n random probabilities summing to one. These probabilities are used in independent multinomial trials with n outcomes. Let N_n (N_n^*) be the number of trials needed until each (some) outcome has occurred at least c times. By embedding the sampling procedure in a Poisson point process the distributions of N_n and N_n^* can be expressed using extremes of independent identically distributed random variables. Using this, asymptotic distributions as $n \to \infty$ are obtained from classical extreme value theory. The limits are determined by the behaviour of the Laplace transform of X close to the origin or at infinity. Some examples are studied in detail.

Keywords: Poisson embedding; point process; Polya urn; inverse gaussian; lognormal; gamma distribution; repeat time

AMS 1991 SUBJECT CLASSIFICATION: PRIMARY 60G70
SECONDARY 60C05

1 Introduction

Consider a random experiment with n outcomes having probabilities $p_1, \dots, p_n$. Independent trials are performed until each outcome has occurred at least c times. Let N_n be the number of trials needed and let N_n^* ($< N_n$) be the number of trials when some unspecified outcome has occurred c times.

*Dept. Mathematics, KTH, SE–10044, Stockholm, Sweden. E-mail: lholst@math.kth.se

To find the distribution of N_n for $c = 1$ and $p_1 = \cdots = p_n = \frac{1}{n}$ is usually called the *coupon collector's problem.* The approach by embedding in Poisson point processes given in Section 2 below gives the relation

$$\sum_{i=1}^{N_n} Z_i = n \max(Y_1, \ldots, Y_n),$$

where the random variables $N_n, Z_1, Z_2, \ldots$ are independent, the Z's being $Exp(1)$ (density e^{-z} for $z > 0$) and the Y's are independent and $Exp(1)$. This implies

$$E(N_n) = nE(\max(Y_1, \ldots, Y_n)) = n \sum_{j=1}^{n} \frac{1}{j} \sim n \log n, \quad n \to \infty,$$

and the limit distribution

$$\lim_{n\to\infty} P(N_n/n - \log n \leq x) = e^{-e^{-x}},$$

see Section 4.1 below. To find the distribution of N_n^* for $c = 2$ and equal p's is the *birthday problem.* In this case the embedding approach gives

$$\sum_{i=1}^{N_n^*} Z_i = n \min(Y_1, \ldots, Y_n),$$

where $N_n^*, Z_1, Z_2, \ldots$ are independent, the Z's $Exp(1)$, and the Y's independent and $\Gamma(2, 1)$ (we denote by $\Gamma(c, 1)$ a gamma distribution with density $y^{c-1}e^{-y}/(c-1)!$ for $y > 0$). We have

$$E(N_n^*) = nE(\min(Y_1, \ldots, Y_n)) = n \int_0^\infty (1 + y)^n e^{-ny} dy \sim \sqrt{\pi n/2}, \quad n \to \infty,$$

and the limit distribution

$$\lim_{n\to\infty} P(N_n^*/\sqrt{2n} \leq x) = 1 - e^{-x^2}, \quad x > 0,$$

see Section 5.1 below. Combinatorial approaches to the coupon collector's problem or the birthday problem have a long history and can be found in many texts, see e.g. Feller (1968) or Blom, Holst and Sandell (1994).

In Holst (1995) it is proved that the distribution functions of N_n^* can be partially ordered in the p's by Schur-convexity and that N_n^* is stochastically largest in the symmetric case. A slight modification of the argument shows partial ordering in the collector problem and that N_n is stochastically smallest in the symmetric case.

Papanicolaou, Kokolakis and Boneh (1998) studied a "random" coupon collector problem for the case $c = 1$ by letting the p's be random and given by

$$\frac{X_1}{X_1 + \cdots + X_n}, \ldots, \frac{X_n}{X_1 + \cdots + X_n},$$

where $X_1, X_2, \ldots$ are independent identically distributed positive random variables. Applications of the model were given and asymptotic results for $E(N_n)$ as $n \to \infty$ were derived. Note that $X_1 = \cdots = X_n = 1$ gives the classical case.

In our paper asymptotic results are obtained for the random coupon collector problem both for the distribution and the mean of N_n for $c \geq 1$, generalizing those of Papanicolaou *et al* (1998) for the mean. We prove our results by embedding in Poisson point processes. A similar approach is used in Holst (1995) to study birthday problems. By this device distributional problems on N_n are transformed so that classical extreme value theory for independent identically distributed random variables can be applied, c.f. Resnick (1987). In a similar way we study N_n^* for random p's given as above. Other recent papers using Poisson embedding on problems of a similar flavour as ours are Steinsaltz (1999) and Camarri and Pitman (2000).

In the following $X_1, X_2, \ldots$ denote independent copies of a strictly positive random variable X with mean $\mu = E(X) < \infty$. We will see that the limit behaviour of N_n as $n \to \infty$ is determined by the behaviour of the distribution function $F_X(x) = P(X \leq x)$ for small x, or equivalently by the behaviour of the Laplace transform $g_X(s) = E(e^{-sX})$ for large s. The limit behaviour of N_n^* is determined by the behaviour of $g_X(s)$ for small s.

The organization of the paper is as follows. In Section 2 the embedding of N_n in a Poisson point process is constructed. Using this an expression for $E(N_n)$ is derived. In Section 3 extreme value distributions of Fréchet type ($\exp(-y^{-\alpha})$) occur as limiting distributions of N_n and an example with the gamma distribution is analyzed. In Section 4 extremes of Gumbel type ($\exp(-e^{-y})$) are considered; examples discussed involve X having one-point, inverse gaussian and lognormal distributions. In Section 5 birthday problems are studied and limit distributions of Weibull type ($1 - \exp(-y^{\alpha})$) are obtained.

2 Embedding and $E(N_n)$

Let Π be a Poisson point process with intensity one in the first quadrant of the plane. Independent of Π let $X_1, X_2, \ldots$ be independent identically distributed strictly positive random variables with (finite) mean μ. Introduce random "strips" and set

$$I_{it} = \Pi \cap \{(x,s) : \sum_{j=1}^{i-1} X_j < x \leq \sum_{j=1}^{i} X_j,\ 0 < s \leq t\},$$

for $i = 1, 2, \ldots$ and $t > 0$. Here I_{it} is the set of points of Π in the i:th strip up to "time" t. Let $|I_{it}|$ denote the number of these points. As Π is a Poisson process with intensity one we have

$$\min\{t : |I_{it}| = c\} = Y_i / X_i,$$

where the independent random variables $Y_1, Y_2, \ldots$ are $\Gamma(c, 1)$ and independent of $X_1, X_2, \ldots$. The first time the first n strips all contain at least c points can be written

$$M_n = \max(Y_1/X_1, \ldots, Y_n/X_n).$$

Given $X_1, \ldots, X_n$, the projection on the s-axis of the points in these strips is a Poisson process with intensity $\sum_{j=1}^{n} X_j$. The total number of points in the n strips up to time M_n can be identified with N_n, because the probability that a point occurs in the i:th strip is $X_i / \sum_{j=1}^{n} X_j$ and the points are independent of each other. Thus with independent $Z_1, Z_2, \ldots$ all being $Exp(1)$ and independent of N_n, we have the basic relation:

$$\sum_{i=1}^{N_n} Z_i = M_n \sum_{j=1}^{n} X_j.$$

Using this different quantities of the distribution of N_n can be expressed in the random variables $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$.

Theorem 2.1 *With notation as above:*

$$E(N_n)/n = \mu E(M_{n-1}) + \sum_{j=0}^{c-1} \frac{c-j}{j!} E(X_n^j M_{n-1}^j e^{-X_n M_{n-1}}),$$

$$E(N_n)/n = \mu E(M_{n-1}) + o(1), \quad n \to \infty,$$

$$E(N_n) < \infty \Longleftrightarrow E(1/X) < \infty,$$

and for $c = 1$

$$E(N_n) = n\mu E(M_{n-1}) + 1 = n\mu \int_0^\infty [1 - (1 - g_X(s))^{n-1}] ds + 1.$$

Proof. The embedding implies $E(N_n) = E(M_n \sum_{j=1}^n X_j)$. Thus symmetry and independence give

$$E(N_n) = nE(M_n X_n) = nE\left(X_n \int_0^\infty \big[1 - P(M_{n-1} \le s) P(Y_n/X_n \le s|X_n)\big] ds\right)$$

$$= nE\left(X_n \int_0^\infty \big[1 - P(M_{n-1} \le s)\big] ds\right)$$

$$+ nE\left(X_n \int_0^\infty P(M_{n-1} \le s) P(Y_n/X_n > s|X_n) ds\right)$$

$$= n\mu E(M_{n-1}) + n \int_0^\infty P(M_{n-1} \le s)\, E(X_n P(Y_n > X_n s|X_n))\, ds.$$

As Y_n is $\Gamma(c,1)$ and independent of X_n we have

$$P(Y_n > X_n s|X_n) = \sum_{\ell=0}^{c-1} \frac{X_n^\ell s^\ell}{\ell!} e^{-X_n s}.$$

Thus

$$\int_0^\infty P(M_{n-1} \le s) E(X_n P(Y_n > X_n s|X_n)) ds$$

$$= \sum_{\ell=0}^{c-1} E\left(\int_0^\infty P(M_{n-1} \le s) X_n \frac{X_n^\ell s^\ell}{\ell!} e^{-X_n s} ds\right)$$

$$= \sum_{\ell=0}^{c-1} \int_0^\infty P(X_n M_{n-1} \le s) \frac{s^\ell e^{-s}}{\ell!} ds = \sum_{\ell=0}^{c-1} P(X_n M_{n-1} \le V_\ell),$$

where V_ℓ is $\Gamma(\ell+1, 1)$. Hence

$$\sum_{\ell=0}^{c-1} P(X_n M_{n-1} \le V_\ell) = \sum_{\ell=0}^{c-1} E\left(\sum_{j=0}^{\ell} \frac{X_n^j M_{n-1}^j}{j!} e^{-X_n M_{n-1}}\right)$$

$$= \sum_{j=0}^{c-1} \frac{c-j}{j!} E(X_n^j M_{n-1}^j e^{-X_n M_{n-1}}).$$

Combining the results above proves the first assertion. As $M_n \to \infty$ a.s. as $n \to \infty$ the second assertion follows. It is readily seen that the third assertion holds for any c if it holds for $c = 1$.

Let $c = 1$. Then the Y's are $Exp(1)$ and we have

$$P(Y/X > s) = E(P(Y > sX|X)) = E(e^{-sX}) = g_X(s).$$

Thus $P(M_{n-1} \le s) = (1 - g_X(s))^{n-1}$, and therefore

$$E(e^{-X_n M_{n-1}}) = E(g_X(M_{n-1})) = -\int_0^\infty g_X(s)(n-1)(1-g_X(s))^{n-2} g'_X(s)ds = \frac{1}{n},$$

proving the last formula in the assertion. Furthermore,

$$E(M_n) = \int_0^\infty [1 - (1 - g_X(s))^n]ds = \int_0^\infty g_X(s) \sum_{k=0}^{n-1} (1 - g_X(s))^k ds.$$

Therefore $E(M_n) < \infty$ if and only if

$$\int_0^\infty g_X(s)ds = \int_0^\infty E(e^{-sX})ds = E\left(\int_0^\infty e^{-sX}ds\right) = E\left(\frac{1}{X}\right) < \infty.$$

Proving the third assertion for $c = 1$ and therefore for all positive integers c. □

The distribution function $F_c(s) = P(Y/X \le s)$ is important for studying N_n. The following result will be useful later on.

Proposition 2.1 *Let X and Y be independent positive random variables, X with distribution function F_X and Laplace transform g_X, and Y being $\Gamma(c, 1)$. Then for $s > 0$:*

$$g_X^{(k)}(s) = (-1)^k E(X^k e^{-sX}),$$

$$1 - F_c(s) = P(Y/X > s) = \sum_{k=0}^{c-1} (-1)^k \frac{s^k}{k!} g_X^{(k)}(s) = \int_0^\infty F_X(x/s) \frac{x^{c-1}e^{-x}}{(c-1)!} dx,$$

$$\begin{aligned}
F'_c(s) &= (-1)^c \frac{s^{c-1}}{(c-1)!} g_X^{(c)}(s) = \frac{c}{s}(F_c(s) - F_{c+1}(s)) \\
&= \frac{1}{s} \int_0^\infty F_X(x/s)(x-c) \frac{x^{c-1}e^{-x}}{(c-1)!} dx, \\
F''_c(s) &= -\frac{1}{s^2}\left[(-1)^{c+1} \frac{s^{c+1}}{(c-1)!} g_X^{(c+1)}(s) - (-1)^c \frac{s^c}{(c-2)!} g_X^{(c)}(s)\right] \\
&= -\frac{1}{s^2} \int_0^\infty F_X(x/s)((x-c)^2 - c) \frac{x^{c-1}e^{-x}}{(c-1)!} dx.
\end{aligned}$$

Proof. As Y is $\Gamma(c,1)$ we have

$$P(Y/X > s) = E(P(Y > sX|X)) = E\left(\sum_{k=0}^{c-1} \frac{s^k X^k}{k!} e^{-sX}\right),$$

and also

$$P(Y/X > s) = E(P(X < Y/s|Y)) = \int_0^\infty F_X(x/s) \frac{x^{c-1}e^{-x}}{(c-1)!} dx.$$

By differentiation the other formulas follows by straightforward calculations. $\square$

3 Extremes of Fréchet type for N_n

In this section we consider X such that for some $\alpha > 0$ and for some slowly varying function L

$$P(X \le x) = x^\alpha L(x), \quad x \downarrow 0.$$

Recall that L is slowly varying at 0 if $L(tx)/L(x) \to 1$ as $x \downarrow 0$ for every fixed $t > 0$. A special case is the gamma distribution. The limiting distributions of N_n are extreme value distributions of Fréchet (or Φ_α) type, c.f. Resnick (1987).

Theorem 3.1 *Let $a_n \to \infty$ such that $na_n^{-\alpha} L(1/a_n)\,\Gamma(\alpha + c)/(c-1)! \to 1$. Then*

$$P(N_n/na_n\mu \le y) \to e^{-y^{-\alpha}}, \quad y > 0,$$

$$E(N_n)/na_n\mu \to \Gamma(1 - 1/\alpha), \quad \alpha > 1, \quad \textit{and} \quad E(N_n) = +\infty, \quad \alpha < 1.$$

Proof. Using Proposition 2.1 we get as $s \to \infty$

$$P(Y/X > s) = \int_0^\infty (x/s)^\alpha L(x/s) \frac{x^{c-1}}{(c-1)!} e^{-x} dx$$

$$\sim s^{-\alpha} L(1/s) \int_0^\infty \frac{x^{\alpha+c-1}e^{-x}}{(c-1)!} dx = s^{-\alpha} L(1/s)\,\Gamma(\alpha + c)/(c-1)!.$$

Hence for $y > 0$

$$nP(Y/X > a_n y) \sim ny^{-\alpha} a_n^{-\alpha} L(1/a_n)\,\Gamma(\alpha + c)/(c-1)! \sim y^{-\alpha}.$$

Poisson convergence gives

$$\sum_{j=1}^n I(Y_j/X_j > a_n y) \to Poisson(y^{-\alpha}).$$

Thus for $y > 0$

$$P(M_n/a_n \le y) = P\left(\sum_{j=1}^{n} I(Y_j/X_j > a_n y) = 0\right) \to e^{-y^{-\alpha}}.$$

From the behaviour of $P(Y/X > s)$ as $s \to \infty$ we have for any integer $0 < k < \alpha$ that $E((Y/X)^k) < \infty$. Hence by Resnick (1987, p. 77) $E((M_n/a_n)^k) \to \Gamma(1 - k/\alpha)$ and Theorem 2.1 gives for $\alpha > 1$

$$E(N_n)/na_n\mu = E(M_{n-1})/a_n + o(1/a_n) \to \Gamma(1 - 1/\alpha), \quad n \to \infty.$$

If $\alpha < 1$ then $E(1/X) = +\infty$ implying $E(N_n) = +\infty$. Thus the second and third assertions are proved.

By the embedding we have

$$E\left(e^{-tM_n \sum_{j=1}^{n} X_j}\right) = E\left((e^{-t\sum_{j=1}^{N_n} Z_i}|N_n)\right) = E\left((1+t)^{-N_n}\right).$$

Therefore for $s \ge 0$ and $t = e^{s/na_n\mu} - 1$ we get

$$E\left(e^{-sN_n/na_n\mu}\right) = E\left(\exp\left(-s \cdot \frac{e^{s/na_n\mu} - 1}{s/na_n\mu} \cdot \frac{M_n}{a_n} \cdot \frac{\sum_{j=1}^{n} X_j}{n\mu}\right)\right).$$

As

$$\frac{e^{s/na_n\mu} - 1}{s/na_n\mu} \to 1, \quad P(M_n/a_n \le y) \to e^{-y^{-\alpha}}, \quad \frac{\sum_{j=1}^{n} X_j}{n\mu} \to 1 \quad \text{in probability,}$$

it follows that

$$P\left(\frac{e^{s/na_n\mu} - 1}{s/na_n\mu} \cdot \frac{M_n}{a_n} \cdot \frac{\sum_{j=1}^{n} X_j}{n\mu} \le y\right) \sim P\left(\frac{M_n}{a_n} \le y\right) \to e^{-y^{-\alpha}}, \quad n \to \infty.$$

Thus, by the continuity theorem for Laplace transforms we have for $s \ge 0$ that

$$E(e^{-sN_n/na_n\mu}) \to \int_0^\infty e^{-sy} d(e^{-y^{-\alpha}}),$$

from which the first assertion of the theorem follows. $\square$

3.1 Example: gamma distribution

Let X be $\Gamma(\alpha, 1)$. Then

$$g_X(s) = E(e^{-sX}) = (1+s)^{-\alpha},\ s > -1, \quad P(X \le x) \sim x^{\alpha}/\Gamma(\alpha+1),\ x \downarrow 0.$$

In Theorem 3.1 we have $\mu = \alpha$ and take

$$a_n = [n(\alpha + c - 1)\cdots(\alpha+1)/(c-1)!]^{\frac{1}{\alpha}},$$

where $a_n = n^{\frac{1}{\alpha}}$ for $c = 1$. For $X_1, \ldots, X_n$ independent and $\Gamma(\alpha, 1)$ the sum $X_1 + \cdots + X_n$ is $\Gamma(n\alpha, 1)$ and independent of $(X_1, \ldots, X_n)/(X_1 + \cdots + X_n)$, which has the symmetric Dirichlet distribution $D(\alpha, \ldots, \alpha)$. Hence for $\alpha > 1$ it follows by the embedding that

$$E(N_n) = E(M_n \sum_{j=1}^{n} X_j) = E(M_n) \cdot \left[E\left(\frac{1}{\sum_{j=1}^{n} X_j} \right) \right]^{-1}$$

$$= (n\alpha - 1)E(M_n) \sim n^{1+\frac{1}{\alpha}} \alpha [(\alpha + c - 1)\cdots(\alpha+1)/(c-1)!]^{\frac{1}{\alpha}} \Gamma(1 - 1/\alpha).$$

The mean is infinite for $\alpha \le 1$. Note that $(Y/c)/(X/\alpha)$ has an F-distribution.

For the exponential case $\alpha = 1$, we take $a_n = cn$ and get the limit

$$P(N_n/cn^2 \le y) \to e^{-1/y}, \quad n \to \infty.$$

The "probabilities" $X_k/(X_1 + \cdots + X_n)$ for $k = 1, \ldots, n$ can be interpreted as the spacings in a random sample of size $n-1$ from a uniform distribution on the unit interval. This corresponds to a $D(1, \ldots, 1)$ prior distribution on the drawing probabilities. Unconditionally the drawing procedure is a Polya urn scheme with n balls of different colours at start and replacing each drawn ball together with one new of the same coulour. A general Polya scheme corresponds to having some $\alpha > 0$.

4 Extremes of Gumbel type for N_n

In this section we consider distributions such that $P(X \le x) \to 0$ faster than any power as $x \downarrow 0$. Extreme value distributions will be of Gumbel (or Λ) type, see Resnick (1987).

Assume for the Laplace transform $g_X(s) = E(e^{-sX})$ and its derivatives that for $k = 0, 1, 2, \ldots$ and as $s \to \infty$

$$h_k(s) := \frac{sE(X^{k+1}e^{-sX})}{E(X^k e^{-sX})} \to \infty, \quad \frac{h_{k+1}(s)}{h_k(s)} = \frac{E(X^{k+2}e^{-sX})E(X^k e^{-sX})}{(E(X^{k+1}e^{-sX}))^2} \to 1.$$

Using Proposition 2.1 this implies for Y being $\Gamma(c, 1)$ that

$$P(Y/X > s) = 1 - F_c(s) \sim s^{c-1}E(X^{c-1}e^{-sX})/(c-1)!,$$

$$F_c'(s) = s^{c-1}E(X^c e^{-sX})/(c-1)!, \quad F_c''(s) \sim -s^{c-1}E(X^{c+1}e^{-sX})/(c-1)!.$$

Thus

$$\frac{(1 - F_c(s))F_c''(s)}{(F_c'(s))^2} \to -1,$$

and $F_c''(s) < 0$ for s sufficiently large. Then from classical extreme value theory, see Resnick (1987, Prop. 1.1 and 2.1),

$$P((M_n - b_n)/a_n \le y) \to e^{-e^{-y}}, \quad (E(M_n) - b_n)/a_n \to \gamma, \quad n \to \infty,$$

where $M_n = \max(Y_1/X_1, \ldots, Y_n/X_n)$ and γ is Euler's constant, and with the norming constants given from

$$\frac{1}{n} = 1 - F_c(b_n), \quad a_n = (1 - F_c(b_n))/F_c'(b_n).$$

The limit behaviour of N_n will now be obtained by the embedding.

Theorem 4.1 *Let X satisfy the conditions above and $E(X^2) < \infty$. Then with a_n and b_n as above*

$$P\left(\left(N_n/n\mu - b_n\right)/a_n \le y\right) \to e^{-e^{-y}}, \quad \left(E\left(N_n/n\mu\right) - b_n\right)/a_n \to \gamma.$$

Proof. By the embedding we have

$$\sum_{i=1}^{N_n} Z_i = M_n \sum_{j=1}^{n} X_j.$$

Using the estimates above it follows that $b_n \to \infty$ and

$$\frac{b_n^2}{na_n^2} = b_n^2 \cdot \frac{(1 - F_c(b_n))(F_c'(b_n))^2}{(1 - F_c(b_n))^2} = b_n^2 \cdot \frac{(F_c'(b_n))^2}{(1 - F_c(b_n))F_c''(b_n)} \cdot F_c''(b_n)$$

$$\sim -b_n^2 F_c''(b_n) \sim -\frac{1}{(c-1)!}E((b_nX)^{c+1}e^{-b_nX}) \to 0.$$

Thus

$$\mathrm{Var}\left(\frac{b_n}{a_n} \cdot \frac{\sum_{j=1}^n X_j}{n}\right) = \frac{b_n^2}{a_n^2} \cdot \frac{\mathrm{Var}(X)}{n} \to 0,$$

and we get that

$$\frac{M_n \sum_{j=1}^n X_j}{n a_n \mu} - \frac{b_n}{a_n} = \frac{M_n - b_n}{a_n} \cdot \frac{\sum_{j=1}^n X_j}{n\mu} + \frac{b_n}{a_n} \cdot \left(\frac{\sum_{j=1}^n X_j}{n\mu} - 1 \right)$$

has the same asymptotic behaviour as $(M_n - b_n)/a_n$. Furthermore by Theorem 2.1 and the estimates above

$$\mathrm{Var}\left(\sum_{i=1}^{N_n} (Z_i - 1)/n a_n \right) = E(N_n)/(n a_n)^2 = (\mu E(M_{n-1}) + o(1))/(n a_n)^2 \to 0.$$

Hence

$$\frac{M_n \sum_{j=1}^n X_j}{n a_n \mu} - \frac{b_n}{a_n} = \frac{\sum_{i=1}^{N_n} Z_i}{n a_n \mu} - \frac{b_n}{a_n} = \frac{\sum_{i=1}^{N_n} (Z_i - 1)}{n a_n \mu} + \frac{1}{a_n} \left(\frac{N_n}{n\mu} - b_n \right)$$

has the same asymptotic distribution as

$$\frac{1}{a_n} \left(\frac{N_n}{n\mu} - b_n \right),$$

that is the same as that of $(M_n - b_n)/a_n$. The convergence of the mean also follows from Resnick (1987, Prop. 2.1). □

4.1 Example: constant probabilities

For $X \equiv \mu$ we have $E(e^{-sX}) = e^{-s\mu}$, $h_k(s) = \mu s$ and $h_{k+1}(s)/h_k(s) = 1$. Furthermore

$$\frac{1}{n} = \sum_{k=0}^{c-1} \frac{(b_n \mu)^k}{k!} e^{-b_n \mu}, \quad \frac{1}{a_n} = n\mu \frac{(b_n \mu)^{c-1}}{(c-1)!} e^{-b_n \mu},$$

implies

$$b_n \mu = \log n + (c-1) \log\log n - \log(c-1)! + o(1), \quad a_n \mu = 1 + o(1).$$

Hence

$$P(N_n/n - \log n - (c-1)\log\log n + \log(c-1)! \le y) \to e^{-e^{-y}},$$

$$E(N_n)/n = \log n + (c-1)\log\log n - \log(c-1)! + \gamma + o(1).$$

For $c = 1$ this is the result for the classical coupon collector's problem given in the Introduction. Recall that N_n is stochastically smallest among all positive distributions of X when X is constant.

4.2 Example: inverse gaussian distribution

Let X be inverse gaussian with mean $\mu = 1$ and variance $\sigma^2 = 1/2\psi$, that is

$$E(e^{-sX}) = e^{2\psi - 2\psi\sqrt{1+s/\psi}},$$

$$P(X \leq x) = \Phi\left(\sqrt{2\psi}\left(\sqrt{x} - \frac{1}{\sqrt{x}}\right)\right) + e^{4\psi}\Phi\left(-\sqrt{2\psi}\left(\sqrt{x} + \frac{1}{\sqrt{x}}\right)\right),$$

where Φ is the standard normal distribution function. For $s \to \infty$ we have

$$s^k E(X^k e^{-sX}) \sim (\psi s)^{k/2} E(e^{-sX}),$$

$$1 - F_c(s) = \frac{(\psi s)^{(c-1)/2}}{(c-1)!}\left(1 + O\left(\frac{1}{\sqrt{s}}\right)\right) E(e^{-sX}),$$

$$F_c'(s) = \frac{(\psi s)^{c/2}}{s\,(c-1)!}\left(1 + O\left(\frac{1}{\sqrt{s}}\right)\right) E(e^{-sX}).$$

Thus, the assumptions of Theorem 4.1 are satisfied. From $1 - F_c(b_n) = 1/n$ we get

$$2\sqrt{\psi b_n} = \log n + (c-1)\log\log n - (c-1)\log 2 - \log(c-1)! + 2\psi + o(1),$$

$$a_n = \frac{1 - F_c(b_n)}{F_c'(b_n)} \sim \frac{\log n}{2\psi},$$

and

$$b_n/a_n = \frac{1}{2}\log n + (c-1)\log\log n - \log((c-1)!2^{c-1}) + 2\psi + o(1).$$

Recalling $\sigma^2 = 1/2\psi$ we obtain

$$P(N_n/\sigma^2 n\log n - b_n/a_n \leq y) \to e^{-e^{-y}}, \quad E(N_n)/n\log n = \sigma^2(b_n/a_n + \gamma) + o(1).$$

4.3 Example: lognormal distribution

Let X have a lognormal distribution. Without loss of generality let $X = e^{\sigma Z}$, where Z is standard normal and $\mu = E(X) = e^{\sigma^2/2}$. For $s > 0$ we have

$$E(X^k e^{-sX}) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{k\sigma z - s e^{\sigma z} - z^2/2} dz.$$

With z_s such that

$$\frac{z_s}{\sigma}\, e^{\sigma z_s} = s,$$

we find after some calculations of saddlepoint type that

$$E(X^k e^{-sX}) \sim \frac{1}{\sqrt{\sigma z_s}}\, e^{-k\sigma z_s - z_s/\sigma - z_s^2/2}, \quad s \to \infty.$$

With $y_n \to \infty$ such that

$$\frac{y_n^{c-3/2}}{\sigma^{c-1/2}(c-1)!}\, e^{-y_n^2/2 - y_n/\sigma} \sim \frac{1}{n},$$

that is roughly $y_n \sim \sqrt{2\log n}$, and

$$b_n = \frac{y_n}{\sigma}\, e^{\sigma y_n},$$

we obtain

$$1 - F_c(b_n) \sim \frac{1}{n}, \quad a_n = e^{\sigma y_n} \sim (1 - F_c(b_n))/F_c'(b_n).$$

This gives the limit

$$P\left(N_n/e^{\sigma^2/2} n e^{\sigma y_n} - y_n/\sigma \le y\right) \to e^{-e^{-y}}, \quad n \to \infty.$$

4.4 Example: strictly positive support

Let $X \ge d > 0$ where $d = \inf\{x : P(X \le x) > 0\}$. Set $X_d = X - d$. Then

$$E(X^k e^{-sX}) \sim e^{-sd} d^k \int_0^\infty P(X_d \le x/s) e^{-x} dx = d^k E(e^{-sX}), \quad s \to \infty.$$

Hence $h_k(s) \sim sd$ and $h_{k+1}(s)/h_k(s) \sim 1$ implying

$$1 - F_c(s) \sim \frac{(sd)^{c-1} e^{-sd}}{(c-1)!} E(e^{-sX_d}), \quad F_c'(s) \sim \frac{d(sd)^{c-1} e^{-sd}}{(c-1)!} E(e^{-sX_d}).$$

The assumptions of Theorem 4.1 are fullfilled and the norming constants can be determined from

$$\frac{1}{n} = \sum_{k=0}^{c-1} \frac{(b_n d)^k}{k!} e^{-b_n d} E(e^{-b_n X_d}), \quad a_n d = 1.$$

For $X \equiv d$ we get the example with constant probabilities. Other cases are modifications of it. For example let X_d be $\Gamma(\alpha, 1)$, then $\mu = d + \alpha$, $E(e^{-sX_d}) = (1+s)^{-\alpha}$ and the norming constants can be choosen as

$$b_n d = \log n + (c - 1 - \alpha) \log\log n + \log(d^\alpha/(c-1)!), \quad a_n d = 1,$$

giving the limit

$$P\left(\frac{d}{d+\alpha}\frac{N_n}{n} - b_n d \le y\right) \to e^{-e^{-y}}, \quad n \to \infty.$$

5 Extremes of Weibull type for N_n^*

In this section we consider N_n^* equals the number of trials until some (unspecified) outcome has occurred $c \geq 2$ times. As in Section 2 we get by embedding

$$\sum_{i=1}^{N_n^*} Z_i = M_n^* \sum_{j=1}^{n} X_j,$$

where

$$M_n^* = \min(Y_1/X_1, \ldots, Y_n/X_n).$$

With a proof similar to that of Theorem 2.1 we obtain:

Theorem 5.1 *We have*

$$E(N_n^*)/n = \mu E(M_{n-1}^*) - \sum_{j=c+1}^{\infty} \frac{j-c}{j!} E(X_n^j M_{n-1}^{*j} e^{-X_n M_{n-1}^*}).$$

In a similar way as before we get asymptotic results for N_n^* from extreme value theory. A crucial quantity is

$$F_c(s) = P(Y/X < s) = \sum_{k=c}^{\infty} \frac{s^k}{k!} E(X^k e^{-sX}), \quad s \geq 0.$$

If

$$\frac{sF_c'(s)}{F_c(s)} = \frac{s^c E(X^c e^{-sX})/(c-1)!}{\sum_{k=c}^{\infty} s^k E(X^k e^{-sX})/k!} \to c, \quad s \downarrow 0,$$

and $a_n \to 0$ such that $nF_c(a_n) \to 1$, then for $y > 0$

$$P(M_n^*/a_n \leq y) \to 1 - e^{-y^c}, \quad n \to \infty,$$

see Resnick (1987, Prop. 1.13, 1.16). Now small modifications of the proof of Theorem 3.1 give limits of Weibull type.

Theorem 5.2 *If* $sF_c'(s)/F_c(s) \to c$ *as* $s \downarrow 0$, $nF_c(a_n) \to 1$ *and* $na_n \to \infty$ *as* $n \to \infty$, *then*

$$P(N_n^*/na_n\mu \leq y) \to 1 - e^{-y^c},\ y > 0, \quad \text{and} \quad E(N_n^*)/na_n\mu \to c\,\Gamma(2 - 1/c).$$

5.1 Example: exponential moments

Suppose that the Laplace transform $g_X(s) = E(e^{-sX})$ is finite in a neighborhood of the origin. Then

$$\sum_{k=c+1}^{\infty} \frac{s^k}{k!} E(X^k e^{-sX}) = O(s^{c+1}).$$

Hence

$$F_c(s) = \frac{s^c}{c!} E(X^c e^{-sX}) + O(s^{c+1}) \sim \frac{s^c}{c!} E(X^c), \quad s \downarrow 0,$$

and therefore we can take

$$a_n = (c!/nE(X^c))^{1/c}.$$

$X \equiv \mu$ and $c = 2$ give the limit in the Introduction for the birthday problem

$$P(N_n^*/\sqrt{2n} \leq y) \to 1 - e^{-y^2}.$$

Recall that N_n^* is stochastically largest when X is constant.

If X is $Exp(1)$, then $\mu = 1$, $a_n = n^{-1/c}$ and we get the limit

$$P(N_n^*/n^{1-1/c} \leq y) \to 1 - e^{-y^c},$$

cf. Subsection 3.1 and the Polya urn scheme.

5.2 Example: lognormal distribution

Let $X = e^{\sigma Z}$ where Z is standard normal. Then $E(X^k) = e^{k^2\sigma^2/2}$ and we have

$$\sum_{k=c+1}^{\infty} \frac{s^{k-c}}{k!} E(X^k e^{-sX}) = \sum_{k=c+1}^{\infty} \frac{s^{k-c}}{k!} e^{k^2\sigma^2/2} E(e^{-se^{k\sigma^2} X})$$

$$= \sum_{k=c+1}^{\infty} \frac{e^{k^2\sigma^2/2} e^{-k(k-c)\sigma^2}}{k!} (se^{k\sigma^2})^{k-c} E(e^{-se^{k\sigma^2} X}) \to 0, \quad s \downarrow 0.$$

Hence

$$F_c(s) = \frac{s^c}{c!} E(X^c e^{-sX}) + o(s^c) \sim \frac{s^c}{c!} E(X^c) = \frac{s^c}{c!} e^{c^2\sigma^2/2}, \quad s \downarrow 0,$$

which gives

$$a_n \sim \left(c! e^{-c^2\sigma^2/2}/n\right)^{1/c}.$$

and the limit

$$P\big(N_n^*/n^{1-1/c}(c!)^{1/c} e^{(1-c)\sigma^2/2} \leq y\big) \to 1 - e^{-y^c}.$$

References

[1] Blom, G., Holst, S. and Sandell, D. (1994). *Problems and Snapshots from the World of Probability.* Springer, New York.

[2] Camarri, M. and Pitman, J. (2000). Limit distributions and random trees derived from the birthday problem with unequal probabilities. *Electronic Journal of Probability.* **5**, Paper no. 1, 1–19.

[3] Feller, W. (1968). *An Introduction to Probability Theory and Its Applications,* Vol. I, 3rd edn. John Wiley, New York.

[4] Holst, L. (1995). The general birthday problem. *Random Structures and Algorithms.* **6**, 201–207.

[5] Papanicolaou, V.G., Kokolakis, G.E. and Boneh, S. (1998). Asymptotics for the random coupon collector problem. *Journal of Computational and Applied Mathematics.* **93**, 95–105.

[6] Resnick, S. (1987). *Extreme Values, Regular Variation, and Point Processes.* Springer, New York.

[7] Steinsaltz, D. (1999). Random time changes for sock-sorting and other stochastic process limit theorems. *Electronic Journal of Probability.* **4**, Paper no. 14, 1–25.