ASYMPTOTIC DISTRIBUTIONS FOR THE COUPON COLLECTOR'S PROBLEM

By Leonard E. Baum and Patrick Billingsley

The Institute for Defense Analyses and The University of Chicago

We sample with replacement from a population of size n, each population element having probability $1/n$ of being drawn. Let W_n be the drawing on which, for the first time, the number of distinct elements that have been sampled is $a_n + 1$, where $0 \leqq a_n < n$. The random variable W_n is the time a coupon collector must wait to fill out a given portion of the set. Here we work out the asymptotic distribution of W_n under various assumptions on the behavior of a_n. This seems to have been done only for special cases, which is surprising.

Let $b_n = n - a_n$, and let μ_n and σ_n^2 denote the mean and variance of W_n.

Theorem 1. *If $a_n/n^{\frac{1}{2}}$ converges to 0, then $W_n - a_n - 1$ converges in probability to 0.*

Theorem 2. *If $a_n/n^{\frac{1}{2}}$ converges to a positive constant λ, then $W_n - a_n - 1$ converges in law to the Poisson distribution with mean $\lambda^2/2$.*

Theorem 3. *If $a_n/n^{\frac{1}{2}}$ and b_n both go to infinity, then $(W_n - \mu_n)/\sigma_n$ converges in law to the normal distribution with mean 0 and variance 1.*

Theorem 4. *If $b_n = b$ is constant, then $\exp\{-(n^{-1}W_n - \log 2n)\}$ converges in law to the chi-square distribution with $2b$ degrees of freedom.*

Rényi [4], in another connection, has previously obtained a part of Theorem 3, and Erdös and Rényi [3] have obtained a special case of Theorem 4.

These theorems have to do with sums of independent random variables, because W_n has the same distribution as

$$X_{n/n} + X_{(n-1)/n} + \cdots + X_{b_n/n},$$

where the variables in the sum are all independent and X_p $(0 < p \leqq 1)$ represents a variable with distribution

$$P\{X_p = k\} = q^{k-1}p, \qquad k = 1, 2, \cdots, \quad (q = 1 - p).$$

The characteristic function of X_p is

$$E\{e^{itX_p}\} = pe^{it}/(1 - qe^{it}); \tag{1}$$

it follows that its mean and variance are $1/p$ and q/p^2, so that

$$\mu_n = n\sum_{k=b_n}^{n} 1/k \tag{2}$$

and

$$\sigma_n^2 = n\sum_{k=b_n}^{n} (n - k)/k^2. \tag{3}$$

Below we find the asymptotic value of σ_n^2 for the various cases considered.

Received 2 March 1965; revised 6 August 1965.

PROOF OF THEOREMS 1 AND 2. By (1), the characteristic function of $W_n - a_n$ is

$$\prod_{k=0}^{a_n}\{(1 - k/n)/(1 - (k/n)e^{it})\}. \tag{4}$$

We use the estimate

$$1 + z = e^{z+\theta z^2} \quad \text{if} \quad |z| \leqq \tfrac{1}{2}. \tag{5}$$

(Here and in what follows, θ is a real or complex number, not the same at each occurrence, satisfying $|\theta| \leqq 1$.) If t is fixed, then, for large n, this estimate can be applied to the numerator and denominator of each factor of (4), which yields

$$\exp \sum_{k=0}^{a_n}\{(k/n)(e^{it} - 1) + 2\theta k^2/n^2\}.$$

Now

$$\sum_{k=0}^{a_n} k/n = a_n(a_n + 1)/2n \to \lambda^2/2,$$

where $\lambda = 0$ under the hypothesis of Theorem 1, and $\lambda > 0$ under the hypothesis of Theorem 2. Moreover, $\sum_{k=0}^{a_n} k^2/n^2 \to 0$. It follows that if $\lambda = 0$, then (4) converges to 1 for all t, and that if $\lambda > 0$, then (4) converges to the characteristic function $\exp\{\frac{1}{2}\lambda^2(e^{it} - 1)\}$ of the correct Poisson distribution.

PROOF OF THEOREM 3. Since verification of Lindeberg's condition involves prohibitive calculations, we attack the problem directly.

Write $\alpha_n = a_n/n$ and $\beta_n = b_n/n$. It is enough to prove asymptotic normality in each of the following three cases:

(i) α_n and β_n bounded away from 0 and 1,

(ii) $\alpha_n \to 0$ (and $n\alpha_n^2 \to \infty$),

(iii) $\beta_n \to 0$ (and $n\beta_n \to \infty$).

For if asymptotic normality fails under the hypothesis of the theorem itself, then, passing to a subsequence, we see that it fails for one or the other of these three cases.

Let us estimate σ_n^2. Since

$$\sum_{k=1}^{m} 1/k = \log m + C + 1/2m - R_m, \qquad 0 < R_m < 1/8m^2, \tag{6}$$

where C is Euler's constant (see [1], p. 125), we have

$$\sum_{k=b_n}^{n} 1/k = -\log \beta_n + \theta/n\beta_n. \tag{7}$$

By the mean-value theorem, the integral of $1/x^2$ over $(k, k + 1)$ differs from $1/k^2$ by at most $2/k^3$, so that $\int_{b_n}^{n} x^{-2}\,dx = (1 - \beta_n)/n\beta_n$ differs from $\sum_{k=b_n}^{n-1} 1/k^2$ by at most $\sum_{k=b_n}^{\infty} 2/k^3 \leqq 4/b_n^2$. Thus

$$\sum_{k=b_n}^{n} 1/k^2 = (1 - \beta_n)/n\beta_n + 5\theta/n^2\beta_n^2, \tag{8}$$

which, together with (7) and (3), yields

$$\sigma_n^2 = (n/\beta_n)(1 - \beta_n + \beta_n \log \beta_n) + 6\theta/\beta_n^2. \tag{9}$$

In case (i) we have

$$\sigma_n^2 \sim (n/\beta_n)(1 - \beta_n + \beta_n \log \beta_n); \tag{10}$$

in case (ii), expanding $\log(1-\alpha_n)$, we have

$$\sigma_n^2 \sim \tfrac{1}{2}n\alpha_n^2; \tag{11}$$

and in case (iii) we have

$$\sigma_n^2 \sim n/\beta_n. \tag{12}$$

Consider for the moment only cases (i) and (ii). If $\phi_p(s)$ denotes the characteristic function of $X_p - p^{-1}$, then

$$\begin{aligned} 1/\phi_p(s) &= (1/p)(e^{isq/p} - qe^{is/p}) \\ &= 1 + \tfrac{1}{2}s^2q/p^2 + (\theta/3)s^3q/p^4 = 1 + \theta s^2q/p^3, \end{aligned}$$

as follows from (1) and the estimates

$$e^{ix} = 1 + ix - \tfrac{1}{2}x^2 + (\theta/6)x^3 = 1 + ix + (\theta/2)x^2. \tag{13}$$

If

$$s^2q/p^3 \leqq \tfrac{1}{2}, \tag{14}$$

then the estimate (5) gives

$$1/\phi_p(s) = \exp\{\tfrac{1}{2}s^2q/p^2 + (\theta/3)s^3q/p^4 + \theta s^4q^2/p^6\}. \tag{15}$$

Let $\psi_n(t)$ denote the characteristic function of $(W_n - \mu_n)/\sigma_n$. Then $\psi_n(t) = \prod^* \phi_p(t/\sigma_n)$, where, in the product, p assumes successively the values n/n, $(n-1)/n, \cdots, b_n/n$. In cases (i) and (ii), for each fixed t and for sufficiently large n, (14) with $s = t/\sigma_n$ holds for all $p \geqq \beta_n$, so that, by (15),

$$1/\psi_n(t) = \exp\{(t^2/2\sigma_n^2)\textstyle\sum^* q/p^2 + (\theta t^3/3\sigma_n^3)\sum^* q/p^4 + (\theta t^4/\sigma_n^4)\sum^* q^2/p^6\},$$

where the range of $\sum^*$ is the same as that of $\prod^*$. Now in case (i) we have, by (10),

$$(1/\sigma_n^3)\textstyle\sum^* 1/p^4 \leqq (n^4/\sigma_n^3)\sum_{k=b_n}^{\infty} 1/k^4 = O(n/\sigma_n^3\beta_n^3) = O(n/n^{\frac{3}{2}}) \to 0$$

and

$$(1/\sigma_n^4)\textstyle\sum^* 1/p^6 \leqq (n^6/\sigma_n^4)\sum_{k=b_n}^{\infty} 1/k^6 = O(n/\sigma_n^4\beta_n^5) = O(n/n^2) \to 0.$$

And in case (ii) we have, by (11),

$$\begin{aligned} (1/\sigma_n^3)\textstyle\sum^* q/p^4 &\leqq [1/\sigma_n^3(1-\sigma_n)^4]\sum_{k=0}^{a_n} k/n \\ &= O((1/n^{\frac{3}{2}}\alpha_n^3)\alpha_n^2n^2/n) = O(1/\alpha_n n^{\frac{1}{2}}) \to 0 \end{aligned}$$

and

$$\begin{aligned} (1/\sigma_n^4)\textstyle\sum^* q^2/p^6 &\leqq [1/\sigma_n^4(1-\alpha_n)^6]\sum_{k=0}^{a_n} k^2/n^2 \\ &= O((1/n^2\alpha_n^4)\alpha_n^3n^3/n^2) = O(1/\alpha_n n) \to 0. \end{aligned}$$

In both of the cases (i) and (ii), the equality $\sum^* q/p^2 = \sigma_n^2$ now implies $\psi_n(t) \to e^{-(\frac{1}{2})t^2}$, which proves asymptotic normality.

It remains to settle case (iii), which we do in effect by replacing X_p with an

exponentially distributed variable having the same mean—that is, by comparing the characteristic function

$$\prod^* \{pe^{it/\sigma_n}e^{-it/\sigma_n p}/(1 - qe^{it/\sigma_n})\} \tag{16}$$

of $(W_n - \mu_n)/\sigma_n$ with the function

$$\prod^* \{e^{-it/\sigma_n p}/(1 - it/\sigma_n p)\}. \tag{17}$$

The ratio of (17) to (16) is

$$\prod^* \{(e^{-it/\sigma_n} - q)/p(1 - it/\sigma_n p)\} = \prod^* \{1 + \theta(t^2/2{\sigma_n}^2 p)/(1 - it/\sigma_n p)\} = \prod^* \{1 + \theta t^2/2{\sigma_n}^2 p\}, \tag{18}$$

where we have used (13) and the inequality $|1 - it/\sigma_n p| \geqq 1$. For n large (and t fixed), the estimate $1 + z = e^{2\theta z}$, valid for $|z| \leqq \frac{1}{2}$, can be applied to each factor on the right, by (12). Therefore the ratio is

$$\exp\{(\theta t^2/{\sigma_n}^2)\textstyle\sum^* 1/p\} = \exp\{\theta t^2(n/{\sigma_n}^2)\textstyle\sum_{k=b_n}^n 1/k\}.$$

By (12) and (6), the ratio goes to 1.

It suffices then, to prove that (17) approaches $e^{-(\frac{1}{2})t^2}$. Its reciprocal is, by (13),

$$\prod^* \{e^{it/\sigma_n p}(1 - it/\sigma_n p)\} = \prod^* \{1 + \tfrac{1}{2}(t^2/{\sigma_n}^2)1/p^2 + \theta(t^3/{\sigma_n}^3)1/p^3\}.$$

By (12), we can, for large n, apply (5) to each factor, which gives

$$\exp\{\tfrac{1}{2}(t^2/{\sigma_n}^2)\textstyle\sum^* 1/p^2 + \theta(t^3/{\sigma_n}^3)\sum^* 1/p^3 + \theta(t^4/{\sigma_n}^4)\sum^* 1/p^4\}. \tag{19}$$

By (12), the second and third terms in the exponential are respectively $O(n/{\sigma_n}^3{\beta_n}^2) = O(1/(n\beta_n)^{\frac{1}{2}})$ and $O(n/{\sigma_n}^4{\beta_n}^3) = O(1/n\beta_n)$, and these both converge to 0 by hypothesis in case (iii).

The first term in the exponential is $\frac{1}{2}t^2(n^2/{\sigma_n}^2)\sum_{k=b_n}^n 1/k^2$, which converges to $\frac{1}{2}t^2$, by (8) and (12). Therefore (19) converges to $e^{(\frac{1}{2})t^2}$, so that (16) converges to $e^{-(\frac{1}{2})t^2}$.

Rényi [4] proved asymptotic normality in cases (i) and (ii).

Proof of Theorem 4. The characteristic function of $(W_n - \mu_n)/n$ is just (16) with σ_n replaced by n. We compare it with (17), with σ_n again replaced by n. The ratio (see (18)) is

$$\prod^* \{1 + \theta t^2/2n^2 p\} = \prod_{k=b}^n \{1 + \theta t^2/2nk\}, \tag{20}$$

which converges to 1 for each t.

Let Y_k be an exponentially distributed random variable with mean $1/k$: $P\{Y_k \geqq x\} = e^{-kx}$. Now (17) is the characteristic function of $\sum_{k=b}^n (Y_k - 1/k)$, where the summands are assumed independent. By Kolmogorov's convergence theorem for random series (see [2], p. 108), this sum converges in law (even with probability 1) to the random variable

$$Z_b = \textstyle\sum_{k=b}^\infty (Y_k - 1/k). \tag{21}$$

From this and the fact that (20) converges to 1 for each t, we conclude that $(W_n - \mu_n)/n$ converges in law to (21).

From the formula for convolving densities, we see inductively that $\sum_{k=b}^{m} Y_k$ has density

$$d_{b,m}(x) = m\binom{m-1}{b-1}e^{-bx}(1 - e^{-x})^{m-b}, \quad x > 0,$$
$$= 0, \qquad x \leqq 0.$$

By (6) we have $\sum_{k=b}^{m} 1/k = \log m + A_b + \theta/m$, where $A_b = C - \sum_{k=1}^{b-1} 1/k$. Therefore the density $d_{b,m}(x + \sum_{k=b}^{m} 1/k)$ of $\sum_{k=b}^{m} (Y_k - 1/k)$ converges for each x to the limit

$$[1/(b-1)!]e^{-b(x+A_b)} \exp[-e^{-(x+A_b)}] \tag{22}$$

as m tends to infinity. By Scheffé's theorem [5], the integral laws converge as well, so that (22) is the density for Z_b.

A change of variable shows that $2e^{-(Z_b+A_b)}$ has on the positive half-line the density $[1/2^b\Gamma(b)]x^{b-1}e^{-x/2}$, which belongs to the chi-square distribution with $2b$ degrees of freedom. Since $(W_n - \mu_n)/n$ converges in law to Z_b,

$$2 \exp\{-[(W_n - \mu_n)/n + A_b]\} \tag{23}$$

converges in law to a chi-square distribution with $2b$ degrees of freedom. By the definition of A_b, (23) reduces to

$$\exp\{-[(1/n)W_n - \log 2n]\}e^{\theta/n},$$

which completes the proof.

It is easy to show that, under the conditions of Theorem 4, $\sigma_n^2 \sim n^2 \sum_{k=b}^{n} 1/k^2$.

Erdös and Rényi [3] proved Theorem 4 for the special case $b = 1$.

REFERENCES

[1] Cramér, Harald (1946). *Mathematical Methods of Statistics*. Princeton Univ. Press.

[2] Doob, J. L. (1953). *Stochastic Processes*. Wiley, New York.

[3] Erdös, P. and Rényi, A. (1961). On a classical problem of probability theory. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **6** 215–220.

[4] Rényi, A. (1962). Three new proofs and a generalization of a theorem of Irving Weiss. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **7** 203–214.

[5] Scheffé, Henry (1947). A useful convergence theorem for probability distributions. *Ann. Math. Statist.* **18** 434–438.