INFINITE OCCUPANCY

A. D. BARBOUR

Uni. Zürich

CLASSICAL OCCUPANCY

n balls thrown independently and uniformly into m boxes: $N_j := \#$ balls in box j, $1 \le j \le m$

$$\sum_{j=1}^{m} N_j = n.$$

Questions: what can be said about

$K_{n0} := \#$ empty boxes;

$K_{nr} := \#$ boxes w/ exactly r balls;

Sequentially: how many balls thrown until the first time that $n-r$ boxes are occupied, T_r:

Coupon collector problem.

Unequal box probabilities:

box i chosen w.pr. p_i, $1 \le i \le m$: $\sum_{i=1}^{m} p_i = 1$

Sampling theory.

$N \sim MN(n : p_1, \ldots, p_m)$; various non-linear functions of N are of interest.

INFINITE OCCUPANCY

Bahadur 1960
Karlin 1967

$p_1 \geq p_2 \geq \cdots$, $\sum_{j\geq 1} p_j = 1$, all positive.

$K := K_n :=$ # occupied boxes

$$= \sum_{j\geq 1} I[N_j \geq 1] =: \sum_{j\geq 1} I_j ;$$

$K_{n,r}$, $r \geq 1$, as before.

Sequentially it is clear that $K_n \to \infty$ a.s.

Karlin showed: $\boxed{K_n / \mathbb{E}K_n \to 1 \text{ a.s.}}$

CLT for K_n?

Ex: $p_j = q^{j-1}(1-q)$, $j \geq 1$, some $0 < q < 1$:

Geometric probabilities. Then, since $N_j \sim Bi(n, p_j)$, we have

$$\mathbb{E}I_j = \mathbb{P}[N_j \geq 1] \sim 1 - e^{-np_j} ; \quad \operatorname{var} I_j \sim e^{-np_j}(1-e^{-np_j})$$

and, since $(N_j, j \geq 1)$ are negatively associated,

$$\operatorname{var} K_n \leq \sum_{j\geq 1}' \operatorname{var} I_j \sim \sum_{j\geq 1} e^{-np_j}(1-e^{-np_j}).$$

Define j_n to satisfy $nq^{j_n-1}(1-q) \geq 1 > nq^{j_n}(1-q)$.

Then

$$\sum_{j\geq 1} e^{-np_j}(1-e^{-np_j})$$

$$\leq np_{j_n+1} \sum_{s\geq 0} q^s + \sum_{s\geq 0} e^{-np_{j_n} q^{-s}}$$

$$\leq \sum_{s\geq 0} q^s + \sum_{s\geq 0} e^{-q^{-s}} < \infty.$$

So $\operatorname{var} K_n$ remains bounded, and the CLT does not hold.

REGULAR VARIATION.

Karlin defined a measure ν by

$$\nu(x,\infty) := \#\{j : p_j \geq x\}.$$

The sequence $(p_j, j\geq 1)$ is called regularly varying with exponent β, $0<\beta<1$, if

$$\nu(x,\infty) \sim \ell(1/x)\, x^{-\beta} \quad \text{as } x \downarrow 0,$$

for $\ell(\cdot)$ slowly varying. [e.g. $p_j \sim j^{-1/\beta}$].

Karlin's theorems: if regularly varying ξ then

(i) (sequentially) $K_n \sim \Gamma(1-\beta) n^\beta \ell(n)$ a.s.

(ii) $\mathbb{E}K_n \asymp \operatorname{var} K_n \asymp n^\beta \ell(n)$

(iii) $(K_n - \mathbb{E}K_n)/\sqrt{\operatorname{var} K_n} \to_d N(0,1)$

(iv) $(K_{nr} - \mathbb{E}K_{nr})/\sqrt{n^\beta \ell(n)}$ are asymptotically jointly ($r \geq 1$) normal as $n \to \infty$.

Dutko (1989) showed that

$$(K_n - \mathbb{E}K_n)/\sqrt{\operatorname{var} K_n} \to_d N(0,1) \text{ iff } \operatorname{var} K_n \to \infty.$$

WARNING.

for $V_r(n) := \operatorname{Var} K_{n,r}$, can have

- $\limsup_{n\to\infty} V_r(n) = \infty$ for all r;
- $V_r(n) \to \infty$ for all $r \leq r_0$;
- $\liminf_{n\to\infty} V_r(n) = 0$ for all $r > r_0$;

} Simultaneously

and

- joint multivariate normal convergence ONLY IF R.V.

B + Cnedin. 2008

POISSONIZATION

If instead $\boxed{Po(n)}$ balls are thrown, and $L_j := \#$ balls in box j, then $L_1, L_2, \ldots$ are independent, with $L_j \sim Po(np_j)$; so the limit theory should not be too surprising.

Cheap Poisson argument.

Choose any sequence $j_n \to \infty$. Then

$$P_n := \sum_{j \geq j_n} p_j \to 0.$$

Thus $N_n^+ := \sum_{j \geq j_n} N_j$ has mean nP_n, and

$$d_{TV}(\mathcal{L}(N_n^+), Po(nP_n)) \leq P_n \to 0 \quad (n \to \infty).$$

Le Cam – Michel observation:

$$d_{TV}(\mathcal{L}(N_j, j \geq j_n), \mathcal{L}(L_j, j \geq j_n)) \leq P_n \to 0$$

also, for $(L_j, j \geq j_n)$ independent $Po(np_j)$ – distributed r.v.'s

Now choose $j_n \to \infty$ so slowly that

$$p_{j_n - 1} \geq \frac{4 \log n}{n} > p_{j_n} .$$

Then $$\mathbb{P}\Big[\bigcup_{j < j_n} \{N_j = 0\} \Big] \leq (j_n - 1)\Big(1 - \frac{4 \log n}{n}\Big)^n$$

$$\leq \frac{n}{4 \log n} \cdot n^{-4} \leq n^{-3} \to 0$$

as $n \to \infty$. Hence, defining

$$\tilde{K}_n := j_n - 1 + \sum_{j \geq j_n} I[L_j \geq 1],$$

we have

$$d_{TV}\big(\mathcal{L}(K_n), \mathcal{L}(\tilde{K}_n)\big) \leq n^{-3} + P_n ,$$

and K_n satisfies the CLT because $\tilde{K}_n$ does, if

$$\operatorname{var} \tilde{K}_n = \sum_{j \geq 1} e^{-np_j}(1 - e^{-np_j}) \to 0 ;$$

error rate $n^{-3} + P_n + 1/\sqrt{\operatorname{var} \tilde{K}_n}$.

LOCAL LIMIT THEOREMS

Hwang and Janson (Ann. Prob. 2008):

local limit approximation for K_n with error of order $O(1/\mathrm{var}K_n)$ if $\mathrm{var}K_n \to \infty$.

E.g. if regularly varying β, then

$$\frac{1}{\mathrm{var}K_n} \asymp \frac{1}{n^\beta \ell(n)}, \text{ whereas } P_n \asymp \left(\frac{\log n}{n}\right)^{1-\beta}.$$

So the cheap argument would give their result (in R.V. case) ONLY for $\beta < 1/2$.

RÖLLIN'S THEOREM for translated Poisson approximation.

Let W be an integer valued random variable, X some random element. Define

$$\mu_X := \mathbb{E}(W|X); \quad \mu = \mathbb{E}W = \mathbb{E}\mu_X;$$

$$\sigma_X^2 := \mathrm{var}(W|X); \quad \rho^2 = \mathbb{E}\sigma_X^2;$$

$$\tau^2 := \mathrm{var}\,\mu_X; \quad \nu^2 = \mathrm{var}\,\sigma_X^2.$$

Suppose that

$$U := (\mu_X - \mu)/\tau$$

satisfies

$$|\mathbb{E}\{f'(U) - Uf(U)\}| \leq \varepsilon \|f''\| \quad (*)$$

for nice f. Then

$$d_{loc}(\mathcal{L}(W), TP(\mu, \sigma^2))$$

$$\leq \mathbb{E}\, d_{loc}(\mathcal{L}(W|X), TP(\mu_X, \sigma_X^2))$$

$$+ O\left(\frac{1}{\rho^2}\left(1 + \frac{\nu^2}{\rho^2} + \frac{\varepsilon\tau^3}{\sigma^2}\right)\right),$$

where $\sigma^2 := \tau^2 + \rho^2 = \mathrm{var}W$.

So if, for instance, $\mathcal{L}(W|X)$ is the law of a sum of independent indicators, and if $\rho^2 \asymp \sigma^2$, $V^2 = O(\rho^2)$ and $\mathbb{E}\tau^3 = O(\sigma^2)$, then a local limit approximation to $\mathcal{L}(W)$ of ideal order $O(1/\sigma^2)$ is obtained.

How do we use this here?

- j_n as before, so that $\mathbb{P}[N_j \geq 1, \text{all } j < j_n] \geq 1 - n^{-3}$.
- j_0 fixed, with $P_0 := \sum_{j \geq j_0} p_j$ not too big.

Realize $(N_j, j \geq j_0) \sim MN(n; \underbrace{(p_j, j \geq j_0)}_{\text{defective dist}^n!})$

by taking $(M_j, j \geq j_0) \sim MN(n; \underbrace{([p_j/P_0], j \geq j_0)}_{\text{probability dist}^n})$

and then 'thinning' independently, with retention probability P_0.

Then $N_j \sim Bi(M_j, P_0)$ are <u>conditionally independent</u> given $M := (M_j, j \geq j_0)$, and hence so are the indicators $(\mathbb{I}[N_j \geq 1], j \geq j_n)$.

$$\mu_M = \sum_{j \geq j_n} \{1 - (1-P_0)^{M_j}\} =: \sum_{j \geq j_n}' q(M_j)$$

$$\sigma_M^2 = \sum_{j \geq j_n}' q(M_j)(1 - q(M_j))$$

$$\rho^2 = \mathbb{E}\sigma_M^2 = \sum_{j \geq j_n}' \mathbb{E}\{(1-P_0)^{M_j} - (1-P_0)^{2M_j}\}.$$

$M_j \sim Bi(n, p_j/P_0)$ gives explicit formulae:

$$\rho^2 \sim \sum_{j \geq j_n}' e^{-np_j}(1 - e^{-np_j(1-P_0)})$$

$$\asymp \sigma^2 \sim \sum_{j \geq j_n}' e^{-np_j}(1 - e^{-np_j}).$$

Then $\nu^2 = \mathrm{var}(\sigma_M^2) = O(\sigma^2)$ also (similar computation).

It thus remains to show that, for $U := (\mu_M - \mu)/\tau$, we have

$$| E\{ f'(U) - U f(U) \} | \leq \varepsilon \| f'' \| ,$$

with $\varepsilon \tau^3 = O(\sigma^2)$, and we're done.

- extensions.