

A new test for the Zipf's law

Mikhail Chebunin

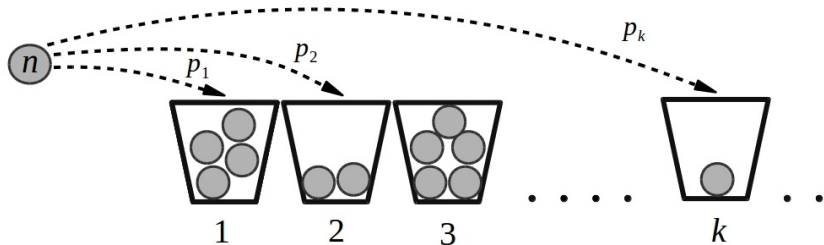
Sobolev Institute of Mathematics of SB RAS, Mathematical Center in
Akademgorodok and Novosibirsk State University, Russia.

Applied Probability Workshop, August 27, 2020.

Infinite urn models

We consider the classical infinite urn models with n balls. Each of n balls goes to urn $i \geq 1$ with probability $p_i > 0$, $\sum_i p_i = 1$, independently of other balls. We assume $p_1 \geq p_2 \geq \dots$.

We let $X_{n,j}$ be the number of balls in urn j , out of the first n balls, and let $X_n = (X_{n,1}, X_{n,2}, X_{n,3}, \dots)$.

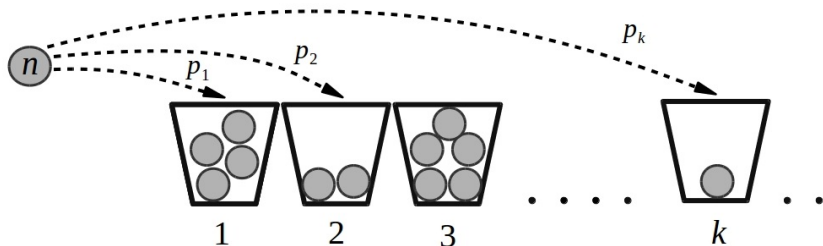


Infinite urn models

A functional of X_n which appears in many contexts is the number of nonempty urns

$$R_n = \#\{j : X_{n,j} > 0\}.$$

R_n is sometimes regarded as a measure of diversity of the sample.



Assumption

Let $\alpha(x) = \max\{j : p_j \geq 1/x\}$ and assume the function $\alpha(x)$ to be regularly varying at infinity,

$$\alpha(x) = x^\theta L(x) \quad \text{with } \theta \in [0, 1], \quad (1)$$

where $L(x)$ is a function slowly varying at infinity.

History

The first author, who considered this model was Bahadur (1960). He noticed that $\mathbf{E}R_n \uparrow \infty$, and proved the Law of Large Numbers for R_n .

Karlin (1967) has established asymptotic properties of random variables R_n , including the Strong Law of Large Numbers and the asymptotic normality in the range $\theta \in (0; 1]$.

Dutko (1989) proved asymptotic normality of R_n if $\mathbf{Var}R_n \rightarrow \infty$. This condition always holds if $\theta \in (0; 1]$ but can hold too for $\theta = 0$.

Gnedin, Hansen and Pitman (2007) focused on study of conditions for convergence $\mathbf{Var}R_n \rightarrow \infty$.

Hwang and Janson (2008) proved the local limit theorem.

C. and Kovalevskii (2016) have proved the Functional Central Limit Theorem for R_n if $\theta \in (0; 1]$.

Let for $t \in [0, 1]$

$$R_n(t) = \frac{R_{[nt]} - \mathbf{E}R_{[nt]}}{\sqrt{\mathbf{E}R_n}}.$$

Theorem

Let $\theta \in (0, 1]$, then process $(R_n(t), 0 \leq t \leq 1)$ converges weakly in the uniform metrics in $D([0, 1])$ to Gaussian process $R(t)$ with zero expectation and covariance function $K(t, \tau) = (t + \tau)^\theta - (\max(t, \tau))^\theta$.

Second interpretation

Let X_j be the urn that the ball j is thrown into.

Then the number of non-empty urns is the number of different elements of the sample size n from some distribution on positive integers.

We assume, that an (unobservable) sample

$$X_1, \dots, X_n$$

is taken from a one-parameter distribution

$$p_i(\theta) = \mathbf{P}\{X_1 = i\} > 0, \quad i \geq 1,$$

and just the number of its various elements is known.

Second interpretation

This model is typical for indexed text in the Internet. For example, Mandelbrot suggested that the probability of word's occurrence in the text is described by some one-parameter probabilistic law with a power-law decrease.

Zakrevskaya and Kovalevskii (2001) proved the consistency for one parametric family of an estimator of $\theta \in (0; 1)$ which is an implicit function of R_n . C. (2014) constructed an R_n -based explicit parameter estimator for $\theta \in (0; 1)$ and proved its consistency.

Asymptotically normal estimator

Let

$$p_i = c \cdot i^{-1/\theta} (1 + o(1/\sqrt{i})), \quad i \geq 1, \quad \theta \in (0, 1),$$

and $c = c(\theta) > 0$ is continuously differentiable by θ .

Theorem (C. and Kovalevskii (2018))

Let θ_n^* be solution of equation

$$R_n = \Gamma(1 - \theta_n^*) (cn)^{\theta_n^*}.$$

Then

$$\ln n \sqrt{R_n} (\theta_n^* - \theta)$$

converges in distribution to normal law with zero expectation and variance $2^\theta - 1$.

But inspite of numerous results in this field some problems remain unresolved. For example, the construction of trajectory estimates of the unknown distribution parameter and test of hypothesis based on information of number of different sample elements.

Trajectory estimators

We may propose the following estimator for parameter θ :

$$\theta_n = \int_0^1 \log^+ R_{[nt]} dA(t)$$

with function $A(\cdot)$ such that

$$\int_0^1 \log t dA(t) = 1, \lim_{x \searrow 0} \log x \int_0^x |dA(t)| = 0, A(0) = A(1) = 0,$$

here $\log^+ x = \max(\log x, 0)$. We assume $A(\cdot)$ to be the sum of a step function and a piecewise continuously differentiable function on $[0, 1]$.

Theorem (C. and Kovalevskii (2019))

If $p_i = i^{-1/\theta} l(i, \theta)$, $\theta \in [0, 1]$, and if $l(x, \theta)$ is a slowly varying function as $x \rightarrow \infty$, then the estimator θ_n is strongly consistent.

We need extra conditions to obtain the asymptotic normality of θ_n .

Theorem (C. and Kovalevskii (2019))

Let $A(t) = 0$, $t \in [0, \delta]$ for some $\delta \in (0, 1)$, and $p_i = ci^{-1/\theta}(1 + o(1/\sqrt{i}))$, $\theta \in (0, 1)$. Then

$$\sqrt{\mathbf{E}R_n}(\theta_n - \theta) - \int_0^1 t^{-\theta} R_n(t) dA(t) \rightarrow_p 0.$$

From Theorem, it follows that θ_n converges to θ with rate $(\mathbf{E}R_n)^{-1/2}$, and normal random variable $\int_0^1 t^{-\theta} R(t) dA(t)$ has variance $\int_0^1 \int_0^1 (st)^{-\theta} K(s, t) dA(s) dA(t)$.

Example

Take

$$A(t) = \begin{cases} 0, & 0 \leq t \leq 1/2; \\ -(\log 2)^{-1}, & 1/2 < t < 1; \\ 0, & t = 1. \end{cases}$$

Then

$$\hat{\theta} = \log_2(R_n/R_{[n/2]}), \quad n \geq 2.$$

The back sequences

Let $Y_{i,n} = X_{n-i+1}$, where X_{n-i+1} is the urn that the ball $n - i + 1$ is thrown into.

Let R'_i , $1 \leq i \leq n$ be the number of different elements (non-empty urns) of the sample $Y_{1,n}, Y_{2,n}, \dots, Y_{i,n}$ size i .

We introduce a new class of estimates that is based on the sequences (R_1, \dots, R_n) and (R'_1, \dots, R'_n) , where (R'_1, \dots, R'_n) are the sequences calculated from the back.

Let for $t \in [0, 1]$

$$R_n(t) = \frac{R_{[nt]} - \mathbf{E}R_{[nt]}}{\sqrt{\mathbf{E}R_n}}, \quad R'_n(t) = \frac{R'_{[nt]} - \mathbf{E}R'_{[nt]}}{\sqrt{\mathbf{E}R_n}}.$$

Theorem (C. (2020))

Let $\theta \in (0, 1]$, then process $(R_n, R'_n, 0 \leq t \leq 1)$ converges weakly in the uniform metrics in $D([0, 1]^2)$ to 2-dimensional Gaussian process (R, R') with zero expectation and covariance function

$$c_{R,R}(t, \tau) = c_{R',R'}(t, \tau) = K(t, \tau), \quad c_{R,R'}(t, \tau) = k(t, \tau),$$

where $K(t, \tau) = (t + \tau)^\theta - (\max(t, \tau))^\theta$, and $k(t, \tau) = ((t + \tau)^\theta - 1)\mathbf{1}(t + \tau > 1)$.

Propose the following estimator for parameter θ :

$$\theta'_n = \int_0^1 \log^+ R'_{[nt]} dA(t).$$

It is easy to see that the estimator θ'_n has similar properties as the estimator θ_n . Consider the properties of the following estimator $\hat{\theta} = (\theta_n + \theta'_n)/2$.

Corollary

Let $A(t) = 0$, $t \in [0, \delta]$ for some $\delta \in (0, 1)$, and $p_i = ci^{-1/\theta}(1 + o(1/\sqrt{i}))$, $\theta \in (0, 1)$. Then

$$\sqrt{\mathbf{E}R_n}(\hat{\theta} - \theta) - \frac{1}{2} \int_0^1 t^{-\theta}(R_n(t) + R'_n(t)) dA(t) \rightarrow_p 0.$$

From the corollary, it follows that $\hat{\theta}$ converges to θ with rate $(\mathbf{E}R_n)^{-1/2}$, and normal random variable

$$\frac{1}{2} \int_0^1 t^{-\theta}(R(t) + R'(t)) dA(t)$$

has variance

$$\frac{1}{2} \int_0^1 \int_0^1 (st)^{-\theta}(K(s, t) + k(s, t)) dA(s) dA(t).$$

Statistical test for a known rate

Let $0 < \theta < 1$ be known. We introduce an *empirical bridge* $\overset{\circ}{R}_n$ as follows.

$$\overset{\circ}{R}_n(k/n) = (R_k - (k/n)^\theta R_n) / \sqrt{R_n},$$

$0 \leq k \leq n$, where $R_0 = 0$. We construct a piecewise linear approximation: for $0 \leq t \leq 1/n$ and $0 \leq k \leq n - 1$,

$$\overset{\circ}{R}_n\left(\frac{k}{n} + t\right) = \overset{\circ}{R}_n(k/n) + nt \left(\overset{\circ}{R}_n((k+1)/n) - \overset{\circ}{R}_n(k/n) \right).$$

Theorem (C. and Kovalevskii (2019))

Under the assumptions of previous Theorem,

$$\sup_{0 \leq t \leq 1} |\overset{\circ}{R}_n(t) - (R_n(t) - t^\theta R_n(1))| \rightarrow 0 \text{ a.s.}$$

Statistical test for a known rate

A similar result is also true for the *empirical bridge* $\overset{\circ}{R}'_n$ which built according to the back process

$$\overset{\circ}{R}'_n(k/n) = (R'_k - (k/n)^\theta R'_n) / \sqrt{R'_n},$$

$0 \leq k \leq n$, where $R_0 = 0$. We construct a piecewise linear approximation: for $0 \leq t \leq 1/n$ and $0 \leq k \leq n - 1$,

$$\overset{\circ}{R}'_n\left(\frac{k}{n} + t\right) = \overset{\circ}{R}'_n(k/n) + nt \left(\overset{\circ}{R}'_n((k+1)/n) - \overset{\circ}{R}'_n(k/n) \right).$$

Theorem

Under the assumptions of previous Theorem,

$$\sup_{0 \leq t \leq 1} |\overset{\circ}{R}'_n(t) - (R'_n(t) - t^\theta R'_n(1))| \rightarrow 0 \text{ a.s.}$$

Corollary (C. (2020))

Under the previous assumptions, $(\overset{\circ}{R}_n, \overset{\circ}{R}'_n)$ converges weakly to 2-dimensional Gaussian process $(\overset{\circ}{R}, \overset{\circ}{R}')$ that can be represented as

$(\overset{\circ}{R}(t), \overset{\circ}{R}'(t)) = (R(t) - t^\theta R(1), R'(t) - t^\theta R'(1)), 0 \leq t \leq 1.$
Its correlation function is given by covariance function

$$c_{R,R}(s, t) = c_{R',R'}(s, t) = \overset{\circ}{K}(s, t), \quad c_{R,R'}(t, \tau) = \overset{\circ}{k}(t, \tau),$$

where

$$\overset{\circ}{K}(s, t) = K(s, t) - s^\theta K(1, t) - t^\theta K(s, 1) + s^\theta t^\theta K(1, 1),$$

$$\overset{\circ}{k}(s, t) = k(s, t) - s^\theta k(1, t) - t^\theta k(s, 1) + s^\theta t^\theta k(1, 1).$$

Statistical test for a known rate

Now we show how to implement the goodness-of-fit test in

this case. Let $W_n^2 = \int_0^1 \left(\overset{\circ}{R}_n(t) \right)^2 + \left(\overset{\circ}{R}'_n(t) \right)^2 dt$.

Then W_n^2 converges weakly to

$$W^2 = \int_0^1 \left(\overset{\circ}{R}(t) \right)^2 + \left(\overset{\circ}{R}'(t) \right)^2 dt.$$

So the test rejects the basic hypothesis if $W_n^2 \geq C$. The p-value of the test is $1 - F_\theta(W_{n,obs}^2)$. Here F_θ is the cumulative distribution function of W^2 and $W_{n,obs}^2$ is a concrete value of W_n^2 for observations under consideration.

Statistical test for an unknown rate

Let us introduce the process (\hat{R}_n, \hat{R}'_n) :

$$\hat{R}_n(k/n) = \frac{R_k - (k/n)^{\hat{\theta}} R_n}{\sqrt{R_n}}, \quad \hat{R}'_n(k/n) = \frac{R'_k - (k/n)^{\hat{\theta}} R'_n}{\sqrt{R'_n}},$$

$0 \leq k \leq n$. As for \hat{R}_n , let for $0 \leq t \leq 1/n$ and $0 \leq k \leq n-1$

$$\hat{R}_n\left(\frac{k}{n} + t\right) = \hat{R}_n(k/n) + nt \left(\hat{R}_n((k+1)/n) - \hat{R}_n(k/n) \right),$$

$$\hat{R}'_n\left(\frac{k}{n} + t\right) = \hat{R}'_n(k/n) + nt \left(\hat{R}'_n((k+1)/n) - \hat{R}'_n(k/n) \right).$$

Theorem (C. (2020))

Under the previous assumptions, $(\widehat{R}_n, \widehat{R}'_n)$ converges weakly to 2-dimensional Gaussian process $(\widehat{R}, \widehat{R}')$ that can be represented as $(\widehat{R}(t), \widehat{R}'(t))$, $0 \leq t \leq 1$, where

$$\widehat{R}(t) = \overset{\circ}{R}(t) - \frac{t^\theta \log t}{2} \int_0^1 u^{-\theta} (R(u) + R'(u)) dA(u),$$

$$\widehat{R}'(t) = \overset{\circ}{R}'(t) - \frac{t^\theta \log t}{2} \int_0^1 u^{-\theta} (R(u) + R'(u)) dA(u).$$





Statistical test for an unknown rate




Corollary

Let $\widehat{W}_n^2 = \int_0^1 \left(\widehat{R}_n(t) \right)^2 + \left(\widehat{R}'_n(t) \right)^2 dt$. Then \widehat{W}_n^2 converges weakly to $\widehat{W}^2 = \int_0^1 \left(\widehat{R}(t) \right)^2 + \left(\widehat{R}'(t) \right)^2 dt$.

So the test rejects the basic hypothesis if $\widehat{W}_n^2 \geq C$. The p-value of the test is $1 - \widehat{F}_\theta(\widehat{W}_{n,obs}^2)$. Here \widehat{F}_θ is the cumulative distribution function of \widehat{W}^2 and $\widehat{W}_{n,obs}^2$ is a concrete value of \widehat{W}_n^2 for observations under consideration.

Thank you for your attention!

-  *Bahadur R. R.* On the number of distinct values in a large sample from an infinite discrete distribution // Proceedings of the National Institute of Sciences of India. – 1960. – V. 26A. – № 2. – P. 67-75.
-  *Chebunin M., Kovalevskii A.* Functional central limit theorems for certain statistics in an infinite urn scheme // Statistics and Probability Letters. – 2016. – V. 119. – P. 344-348.
-  *Chebunin M.* Estimation of parameters of probabilistic models which is based on the number of different elements in a sample // Sib. Zh. Ind. Mat., 17:3 (2014), 135–147.
-  *Dutko M.* Central limit theorems for infinite urn models // Ann. Probab. – 1989. – V. 17. – P. 1255-1263.

-  *Gnedin A., Hansen B., Pitman J.* Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws // *Probability Surveys*. – 2007. – V. 4. – P. 146-171.
-  *Hwang H.-K., Janson S.* Local Limit Theorems for Finite and Infinite Urn Models // *The Annals of Probability*. – 2008. – V. 36. – № 3. – P. 992-1022.
-  *Karlin S.* Central Limit Theorems for Certain Infinite Urn Schemes // *Journal of Mathematics and Mechanics*. – 1967. – V. 17. – № 4. – P. 373-401.