برآورد مقدار e به کمک متغیرهای تصادفی زمان توقف

مهران نقی‌زاده قمی [۱] آزاده کیاپور [۲]

چکیده:

در این مقاله، به کمک متغیرهای تصادفی زمان توقف و انجام شبیه‌سازی، برآوردگرهای نااریب برای عدد نپر، e، به دست می‌آوریم.

واژه‌های کلیدی: توزیع یکنواخت، متغیرهای تصادفی زمان توقف، مقدار e.

۱ مقدمه

اولین مطالعه در مورد عدد نپر، e، در سال ۱۶۱۸ در جدول ضمیمهٔ کاری در مورد لگاریتم‌ها به وسیله‌ی جان نپر منتشر شد. البته کشف این عدد به یاکوب برنولی منتسب می‌شود که در تلاش برای یافتن بسط $\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n$ بود. یکی از مسأله‌های جالب در احتمال، مسأله‌ی جور کردن کلاه‌ها است: n نفر به مهمانی دعوت می‌شوند. در ورود به اتاق کلاه‌ها شماره‌گذاری شده و به هوا پرت می‌شوند. احتمال اینکه هیچ‌یک از مهمان‌ها کلاه خود را نیابد چقدر است؟ پاسخ این سؤال به صورت

$$p_n = 1 - \frac{1}{1!} + \frac{1}{2!} + \cdots + (-1)^n \frac{1}{n!}$$

است که برای n بزرگ به مقدار e^{-1} نزدیک می‌شود(راس (۱۹۹۷)، صفحه‌ی ۱۱۶). همچنین اگر شخصی که احتمال موفقیت او در آزمونی $\frac{1}{n}$ باشد و به طور مستقل n بار در آزمونی شرکت کند، به شرط بزرگ بودن n، احتمال اینکه در هیچ آزمونی موفق نباشد برابر e^{-1} است.

از کاربردهای دیگر e می‌توان به فرمول استرلینگ در تقریب مقدار $n!$ به فرم

$$n! \sim \sqrt{2\pi n}\ \left(\frac{n}{e}\right)^n$$

اشاره کرد.

راسل (۱۹۹۱) در یک مقاله‌ی آموزشی، با معرفی یک متغیر تصادفی زمان توقفو با انجام شبیه‌سازی به برآورد مقدار e پرداخت.

مقاله حاضر به این شکل ادامه می‌یابد. در بخش ۲، متغیرهای تصادفی زمان توقف معرفی می‌شوند. در بخش ۳، ویژگی‌های متغیر تصادفی بررسی شده به‌وسیله‌ی راسل در حالت کلی‌تر و یک متغیر تصادفی دیگر مورد بررسی قرار می‌گیرد. در بخش ٤ با انجام شبیه‌سازی، مقدار e برآورد می‌شود. برنامه‌های R در بخش ۵ آمده است.

۲ متغیرهای تصادفی زمان توقف

متغیرهای تصادفی هندسی و دوجمله‌ای منفی از جمله متغیرهای تصادفی زمان توقف هستند. برای مثال سکه‌ای که احتمال شیرآمدن با آن p باشد، را n بار پرتاب می‌کنیم. تعداد آزمایش‌های لازم برای رسیدن به اولین شیر، یک متغیر تصادفی زمان توقف و دارای توزیع هندسی با پارامتر p است. برای بررسی بیشتر، فرض کنید $X_1, X_2 \cdots \overset{iid}{\sim} Bin(1,p)$ باشند. متغیر تصادفی زیر

$$Y = \min\{n \geq 1 : X_n = 1\} \tag{۱}$$

را در نظر بگیرید. فرض کنید آزمایش (پرتاب سکه) را آن‌قدر ادامه دهیم تا در پرتاب ششم شیر بیاید، یعنی در ۵ پرتاب اول خط و در ششمی شیر آمده است و در نتیجه آزمایش متوقف می‌شود و بنابراین $Y = 6$. واضح است که متغیر تصادفی Y دارای توزیع هندسی با پارامتر p است، زیرا:

$$\begin{aligned}
P(Y=n) &= P(X_1=0,\cdots,X_{n-1}=0,X_n=1)\\
&= P(X_1=0)\cdots P(X_{n-1}=0)P(X_n=1)\\
&= p(1-p)^{n-1}, \quad n=1,2,\cdots.
\end{aligned}$$

حال متغیر تصادفی

$$Z = \min\left\{n \geq 5 : \sum_{i=1}^{n} X_i = 5\right\}$$

[۱] گروه آمار، دانشگاه مازندران، بابلسر

[۲] گروه آمار، دانشگاه آزاد، واحد بابل، بابل

را در نظر بگیرید که بیان کنندهٔ اولین آزمایشی است که منجر به مشاهدهٔ پنجمین شیر می‌شود و بعد از آن آزمایش متوقف می‌شود. متغیر تصادفی Z دارای توزیع دوجمله‌ای منفی با پارامترهای ۵ و n است.

برای مثال دیگر فرض کنید X_1، X_2 و $\cdots$ متغیرهای تصادفی پیوسته با تابع چگالی f باشند. تعریف کنید

$$N = \min\{n \geq 2 : X_n = X_{2:n}\},$$

که در آن $X_{2:n}$ دومین بزرگترین مشاهده در بین X_1، X_2 و $\cdots$ و X_n است. متغیر تصادفی N یک متغیر تصادفی زمان توقف است. برای مثال فرض کنید اولین مقدار تولید شده از توزیع نمایی $x_1 = 5$ و مقدار دوم $x_2 = 3$ باشد. آیا x_2 در بین x_1 و x_2 دومین بزرگترین مشاهده است؟ پاسخ منفی است. پس مقدار دیگری مثلاً $x_3 = 4$ تولید می‌کنیم. آیا x_3 در بین x_1، x_2 و x_3 دومین بزرگترین مشاهده است؟ پاسخ مثبت است، بنابراین آزمایش متوقف و $N = 3$ می‌شود. نکته‌ی جالب در مورد متغیر تصادفی N این است که متغیر تصادفی X_N یعنی اولین متغیر تصادفی مشاهده شده که دومین بزرگترین مقدار مشاهده شده در بین آنها است، هم‌توزیع با X_1 و دارای تابع چگالی f است (راس (۱۹۹۷)، صفحه‌ی ۱۲۱).

متغیر تصادفی پواسون را نیز می‌توان به عنوان یک متغیر تصادفی زمان توقف در نظر گرفت. فرض کنید $U_1, U_2, \cdots \overset{iid}{\sim} U(0,1)$ باشند. می‌توان نشان داد که متغیر تصادفی زیر

$$N = \min\{n \geq 1 : \prod_{i=1}^{n} U_i < e^{-\lambda}\} - 1$$

دارای توزیع پواسون با پارامتر λ است(روهاتگی و صالح (۲۰۰۱)، صفحه‌ی ۲۱۷).

۳ تقریب مقدار e

فرض کنید $U_1, U_2 \cdots$ دنباله‌ای از متغیرهای تصادفی مستقل از توزیع $U(0,1)$ باشند. متغیرهای تصادفی زمان توقف N و M را به‌صورت

$$N = \min\left\{n \geq 2 : U_n > U_{n-1}\right\},$$

$$M = \min\{n \geq 2 : \sum_{i=1}^{n} U_i > 1\}$$

در نظر بگیرید. نشان می‌دهیم $E(N) = E(M) = e$.

N، شاخص اولین متغیر تصادفی یکنواخت است که از مقدار بلافاصله ماقبل خود بزرگتر است. برای توجیه متغیر تصادفی N فرض کنید اعداد تصادفی فرضی زیر از توزیع $U(0,1)$ تولید شده باشند

$$u_1 = 0/64, u_2 = 0/38, u_3 = 0/41$$

اولین مقداری که از مقدار بلافاصله ماقبل خود بزرگتر است u_3 است. بنابراین آزمایش متوقف می‌شود و $N = 3$ است.

M، کمترین مقداری از n است که برای آن $\sum_{i=1}^{n} U_i > 1$ (راسل (۱۹۹۱)، صفحه‌ی ۶۶). دوباره اعداد تصادفی تولیدشده را درنظربگیرید. ملاحظه می‌کنید که $u_1 + u_2 = 1/02 > 1$. بنابراین $M = 2$ است.

ابتدا توزیع N را می‌یابیم. تعداد $n!$ ترتیب ممکن یکسان از $U_1, \cdots, U_n$ وجود دارد. بنابراین داریم:

$$P(N > n) = P(U_1 > \cdots > U_n) = \frac{1}{n!}.$$

حال نشان می‌دهیم $P(M > n) = \frac{1}{n!}$. برای این کار در حالت کلی‌تر نشان می‌دهیم

$$P(M(x) > n) = \frac{x^n}{n!}, \qquad n \geq 1 \tag{۲}$$

که در آن

$$M(x) = \min\left\{n \geq 1 : \sum_{i=1}^{n} U_i > x\right\}.$$

رابطه‌ی (۲) برای $n = 1$ درست است، زیرا

$$P(M(x) > 1) = P(U_1 \leq x) = x.$$

فرض می‌کنیم برای هر $0 < x \leq 1$، فرض استقراء یعنی $P(M(x) > n) = \frac{x^n}{n!}$ برقرار باشد. در این‌صورت داریم

$$\begin{aligned} P(M(x) > n+1) &= \int_0^1 P(M(x) > n+1 | U_1 = y) dy \\ &= \int_0^x P(M(x) > n+1 | U_1 = y) dy \\ &= \int_0^x P(M(x-y) > n) dy \\ &= \int_0^x \frac{(x-y)^n}{n!} dy \\ &= \int_0^x \frac{u^n}{n!} du = \frac{x^{n+1}}{(n+1)!}, \end{aligned}$$

و در نتیجه

$$\sigma_N = \sqrt{E(N^2) - E^2(N)} = \sqrt{3e - e^2} \qquad (٦)$$

فرض کنید $N_1, N_2, \cdots, N_n$ و $M_1, M_2, \cdots, M_n$ نمونه‌هایی تصادفی به ترتیب از توزیع‌های N و M باشند. در این صورت طبق قانون قوی اعداد بزرگ (کاسلا و برگر، صفحه‌ی ۲۳۵) برای n بزرگ داریم:

$$E(N) \approx \bar{N}, \quad E(M) \approx \bar{M}$$

که در آن

$$\bar{N} = \frac{1}{n}\sum_{i=1}^{n} N_i, \quad \bar{M} = \frac{1}{n}\sum_{i=1}^{n} M_i$$

با توجه به رابطه‌های (۳) و (٦) برآوردهای e به‌صورت

$$\hat{e}_N = \bar{N} \quad , \quad \hat{e}_M = \bar{M},$$

و برآوردهای σ به‌صورت

$$\hat{\sigma}_N = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(N_i - \bar{N})^2}, \quad \hat{\sigma}_M = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(M_i - \bar{M})^2}$$

هستند.

٤ شبیه‌سازی

در این بخش، با انجام شبیه‌سازی از توزیع N و M، به برآورد مقدار e می‌پردازیم. مقدار e در نرم‌افزار R با عبارت exp(1) برابر $2/718282$ به دست می‌آید. با توجه به رابطه‌ی (٦) مقدار $\sigma = 0/875094$ به دست می‌آید. برآوردهای $\hat{e}$ و بازه‌های $\hat{e} \pm \frac{\hat{\sigma}}{\sqrt{n}}$ برای تکرارهای مختلف در جدول (۱) آمده است. نتایج نشان می‌دهد که با استفاده از متغیرهای تصادفی زمان توقف N و M به‌خوبی می‌توان مقدار e را برآورد نمود.

در نتیجه استقراء کامل و رابطه‌ی (۲) اثبات می‌شود. با قرار دادن $x = 1$ در رابطه‌ی (۲) داریم

$$P(M > n) = \frac{1}{n!}.$$

بنابراین متغیرهای تصادفی N و M هم‌توزیع هستند. حال به محاسبهٔ امید ریاضی متغیرهای تصادفی N و M می‌پردازیم. داریم

$$E(M) = E(N) = \sum_{n=0}^{\infty} P(N > n) = \sum_{n=0}^{\infty}\frac{1}{n!} = e,$$

و بنابراین:

$$E(M) = E(N) = e. \qquad (۳)$$

تابع جرم احتمال را به‌صورت زیر

$$\begin{aligned} P(N = n) &= P(N > n-1) - P(N > n) \\ &= \frac{1}{(n-1)!} - \frac{1}{n!} = \frac{n-1}{n!}, \quad n > 1 \end{aligned}$$

به دست می‌آوریم. برای یافتن انحراف معیار، از تابع مولد احتمال بهره می‌گیریم. تابع مولد احتمال متغیر تصادفی N برابر مقدار زیر است:

$$\begin{aligned} \phi_N(s) &= E(s^N) = \sum_{n=2}^{\infty} s^n P(N = n) \\ &= \sum_{n=2}^{\infty} s^n \, \frac{n-1}{n!} \\ &= \sum_{n=2}^{\infty} s^n \, \frac{1}{(n-1)!} - \sum_{n=2}^{\infty} s^n \frac{1}{n!} \\ &= s\sum_{n=1}^{\infty}\frac{s^n}{n!} - \sum_{n=2}^{\infty}\frac{s^n}{n!} \\ &= s[e^s - 1] - [e^s - s - 1] \\ &= 1 + e^s(s-1) \end{aligned}$$

.

با توجه به ویژگی‌های تابع مولد احتمال داریم

$$\phi''_N(1) = E(N^2) - E(N). \qquad (٤)$$

مشتق دوم تابع مولد احتمال عبارت از

$$\phi''_N(s) = e^s + se^s \qquad (٥)$$

است. با استفاده از رابطه‌های (۳)، (٤) و (٥) داریم

$$E(N^2) = \phi''_N(1) + E(N) = 3e.$$

۵ برنامه‌های R

```
### approximation of e based on N ###

N<-c()
for(j in 1:10000){
i<-1 ; u<-runif(1)
repeat{
u2<-runif(1)
if (u2>u) break
i<-i+1 ; u<-u2
}
N[j]<-i+1
}
mean(N)
sd(N)/10000
```

```
### approximation of e based on  M ###

M<-c()
for(j in 1:10000){
i<-0 ;s<-0
repeat{
u<-runif(1) ; s<-s+u
if (s>1) break
i<-i+1
}
M[j]<-i+1
mean(M)
sd(M)/10000
```

جدول ۱. میانگین و انحراف استاندارد e

n	$\hat{e}_N \pm \frac{\hat{\sigma}_N}{\sqrt{n}}$	$\hat{e}_M \pm \frac{\hat{\sigma}_M}{\sqrt{n}}$
۱۰۰۰	$2/7231 \pm 0/83 \times 10^{-3}$	$2/716 \pm 0/88 \times 10^{-3}$
۱۰۰۰۰	$2/7275 \pm 0/89 \times 10^{-4}$	$2/7006 \pm 0/86 \times 10^{-4}$
۵۰۰۰۰	$2/7197 \pm 0/175 \times 10^{-4}$	$2/7194 \pm 0/174 \times 10^{-4}$
۱۰۰۰۰۰	$2/7275 \pm 0/8734 \times 10^{-5}$	$2/7184 \pm 0/8735 \times 10^{-5}$

مراجع

[1] Cassela, G. and Berger, R. (2002), *Statistical Inference*, 2nd Edn., Thomson Learning.

[2] Rohatgi, V. and Saleh A. (2001), *An Introduction to Probability and Statistics*, 2nd Edn, John Wiley and Sons.

[3] Ross, S. M. (1997), *Introduction to Probability Models*, Academic.

[4] Russell, K. G. (1991), *Estimating Value of e by Simulation*, the American Statistisian, 45(1), 66-68.