کاربردهایی از اعداد تصادفی و شبیه‌سازی

هادی علیزاده نوقابی [۱] ، رضا علیزاده نوقابی [۲]

چکیده:

شبیه‌سازی از روش‌هایی است که امروزه در بسیاری از علوم مورد توجه قرار گرفته است. اعداد تصادفی نقش مهمی در شبیه‌سازی دارند و اکثر شبیه‌سازی‌ها بر پایه‌ی اعداد تصادفی انجام می‌شوند. در این مقاله به‌عنوان کاربردهایی از اعداد تصادفی و شبیه سازی به محاسبه انتگرال‌ها و براورد عدد π می‌پردازیم. از آنجا که زیربنای یک مطالعه شبیه‌سازی این است که بتوان اعداد تصادفی را تولید کرد که نمایانگر مقدار متغیر تصادفی موردنظر باشد و اینکه توزیع نرمال یکی از توزیع‌های مهم آماری می‌باشد، به عنوان کاربرد سوم از اعداد تصادفی روش‌های مختلف تولید داده از توزیع نرمال را بیان می‌کنیم. نقاط ضعف هر یک از این روش‌ها را بیان و سپس یک روش جدید برای تولید داده از توزیع نرمال را معرفی می‌کنیم.

واژه‌های کلیدی: عدد تصادفی، شبیه سازی، تولید داده از توزیع نرمال، روش تبدیل معکوس، تبدیلات باکس—مولر، روش قطبی، روش عدم پذیرش.

۱ مقدمه

امروزه شبیه‌سازی در بسیاری از علوم مورد استفاده قرار گرفته است. کاربرد شبیه‌سازی در مدل سازی مسایل به ما امکان می‌دهد که سیستم‌های پیچده‌ای را که در عمل بررسی آن‌ها بسیار مشکل است مطالعه کنیم. حل معادلات و محاسبه انتگرال‌های پیچیده که از راه‌های کلاسیک قابل محاسبه نیستند و مسایل بسیار پیچیده دیگر، نمونه‌هایی از کاربرد شبیه‌سازی هستند.

یافتن توزیع تجربی آماره ها و برآوردگرها، مقایسه دو برآوردگر، مقایسه آزمون‌های نیکویی برازش و محاسبه مسایل احتمالی پیچیده مثال‌هایی از کاربرد شبیه سازی در مسایل آماری هستند. برای مثال دیاز—امپرانز[۵] برای تعیین توان و سطح آزمون‌ها از شبیه سازی استفاده کرد. علیزاده نوقابی[۱] برآوردگرهای مختلف آنتروپی را با برآوردگر جدیدش به کمک شبیه‌سازی مقایسه کرد. همچنین علیزاده نوقابی و ارقامی [۲] هفت آزمون مختلف نرمال بودن را به کمک شبیه سازی مونت کارلو با یکدیگر مقایسه کردند. روبرت و کسلا [۱۵] و هادگسن و بروک [۹] برخی روش‌های آماری و هلبی و سینگ[۸]، کنا و شینر[۱۱] و

[۱] دانشجوی دکتری آمار، دانشگاه فردوسی مشهد، گروه آمار
[۲] دانشجوی کارشناسی آمار، دانشگاه فردوسی مشهد، گروه آمار

ویلیام و همکاران[۱۲] برخی مسایل پزشکی را به کمک شبیه‌سازی بررسی کردند.
برای شبیه سازی هر پیشامد، می‌توان از اعداد تصادفی استفاده کرد. یک عدد تصادفی نمایشگر مقدار یک متغیر تصادفی که در (۱ ,۰) دارای توزیع یکنواخت است، می باشد.
در قدیم اعداد تصادفی به‌وسیله دست یا به‌طور مکانیکی با استفاده از تکنیک‌هایی نظیر چرخاندن چرخها، پرتاب تاس یا برزدن کارت‌ها تولید می‌شده است. روش مدرن برای تولید اعداد تصادفی این است که از کامپیوتر استفاده کنیم. یکی از معمولترین روش‌ها برای تولید اعداد تصادفی روش همنهشتی ضربی است که با یک مقدار صحیح اولیه x_0 شروع می‌شود و سپس مقادیر x_n، $n \geq 1$ را به‌طور متوالی از

$$x_n = ax_{n-1}$$

به پیمانه m، محاسبه می‌کند که a و m اعداد صحیح مثبت هستند. بنا بر این هر x_n برابر ۰ یا ۱ یا ... یا $m - 1$ خواهد بود و کمیت $\frac{x_n}{m}$ را به عنوان یک عدد تصادفی در نظر می‌گیرند.
روش دیگر تولید اعداد تصادفی به‌صورت

$$x_n = (ax_{n-1} + c)$$

به پیمانه m، می‌باشد که c مقداری صحیح می‌باشد. این روش همنهشتی مرکب نامیده می‌شود. برای جزییات بیشتر درباره روش‌ها، مارساگلیا[۱۸] و دمیرتاس[۴] را ببینید.
ممکن است بعد از تولید کردن تعدادی داده، دنباله اعداد تولید شده دوباره تکرار شوند. بنابراین باید ثابت‌های a و m را طوری انتخاب کنیم که برای هر مقدار اولیه x_0، احتمال وقوع این تکرار کوچک شود و یا اینکه تعداد داده‌هایی که می‌توانند تولید شوند قبل از شروع تکرار، زیاد باشند. شرایطی که ثابت‌های انتخابی a و m باید داشته باشند در راس[۱٦] بحث شده است.
اعداد تصادفی اغلب براساس الگویتم های خاصی تولید می شوند. البته آنچه که امروزه به عنوان اعداد تصادفی هستند، اعداد شبه تصادفی‌اند که توسط الگوریتم‌های ریاضی خاصی توسط رایانه تولید می‌شوند. همه الگوریتم‌های موجود مناسب نیستند. برای مثال یک الگوریتم ممکن است در بیشتر موارد نوعی همبستگی در دنباله‌ی اعدادش مشاهده شود. برای تشخیص کیفیت تصادفی بودن اعداد تولید شده توسط یک الگوریتم، می‌توان زوج‌های تصادفی (X, Y) از این اعداد را در دستگاه محور مختصات دکارتی رسم کرد. اگر تجمع نقاط در یک قسمت از صفحه مختصات بیشتر باشد، نشان دهنده‌ی کیفیت پایین این اعداد است.

در یک مطالعه شبیه سازی مهم این است که بتوانیم داده هایی را تولید کنیم که نمایانگر مقدار متغیر تصادفی مورد نظر باشد، زیرا داده ها نقش مهمی دارند و استنباط ها بر اساس داده‌ها انجام می‌شوند. برخی از مزایای داده‌های شبیه‌سازی شده نسبت به داده‌های واقعی را می‌توان به‌صورت زیر

بیان کرد.

الف) تولید آن‌ها بسیار ساده و ارزان است.

ب) به دفعات بسیار زیاد می‌توان آنها را تولید کرد.(تکرار پذیری آسان)

ج) توزیع داده‌های شبیه سازی شده مشخص است.

د) میانگین و سایر پارامترهای توزیع داده های شبیه سازی شده مشخص است.

ه) در داده های شبیه سازی شده ناخالصی مانند خطای اندازه گیری، مقادیر دور افتاده، بی‌پاسخی و غیره وجود ندارد.

و) شرایط و مفروضات مدل مانند توزیع، استقلال، همگنی واریانس کاملاً مشخص است.

در بخش های ۲، ۳ و ۴ به بیان برخی از کاربردهای اعداد تصادفی می‌پردازیم.

۲ محاسبه انتگرال‌ها

یکی از کاربردهای اولیه اعداد تصادفی در محاسبه انتگرال‌ها بوده است. فرض کنید $g(x)$ یک تابع است و می‌خواهیم θ را محاسبه کنیم که

$$\theta = \int_{0}^{1} g(x)dx.$$

برای محاسبه مقدار θ توجه داریم که اگر توزیع U از یکنواخت $(0,1)$ باشد آنگاه θ را به‌صورت زیر می‌توان بیان نمود.

$$\theta = E[g(U)]$$

اگر $U_1, U_2, ..., U_k$ متغیرهای تصادفی مستقل یکنواخت در $(0,1)$ باشند در این‌صورت متغیرهای تصادفی $g(U_1), g(U_2), ..., g(U_k)$ مستقل و همتوزیع با میانگین θ خواهند بود. بنابراین طبق قانون قوی اعداد بزرگ با احتمال یک داریم:

$$\frac{1}{k}\sum_{i=1}^{k} g(U_i) \longrightarrow E[g(U)] = \theta \quad , \quad k \longrightarrow \infty.$$

پس اگر تعداد زیادی از اعداد تصادفی u_i را تولید نموده و متوسط مقدار $g(u_i)$ را به عنوان تقریب در نظر بگیریم می توانیم θ را تقریب کنیم. این طریقه تقریب نمودن انتگرالها را روش مونت کارلو می نامند.

اگر بخواهیم

$$\theta = \int_{a}^{b} g(x)dx.$$

را محاسبه کنیم تغییر متغیر $y = \frac{x-a}{b-a}$ را انجام می‌دهیم بنا بر این داریم:

$$\theta = \int_{0}^{1} g(a+(b-a)y)(b-a)dy = \int_{0}^{1} h(y)dy$$

که $h(y) = (b-a)g(a+(b-a)y)$. بنا بر این اگر تعداد زیادی عدد تصادفی را تولید نموده و متوسط مقدار h را که در این اعداد تصادفی بدست می آید، در نظر بگیریم تقریبی برای θ خواهد بود.

بطور مشابه برای محاسبه

$$\theta = \int_{0}^{\infty} g(x)dx.$$

از تغییر متغیر $y = \frac{1}{x+1}$ استفاده می‌کنیم و بنابراین:

$$\theta = \int_{0}^{1} h(y)dy.$$

که $h(y)=\frac{g(\frac{۱}{y}-۱)}{y^۲}$.

سودمندی استفاده از اعداد تصادفی در تقریب نمودن انتگرل‌ها در مورد انتگرال‌های چند بعدی آشکارتر می شود. فرض کنید g تابعی $n-$بعدی بوده و محاسبه

$$\theta=\int_۰^۱\int_۰^۱\dots\int_۰^۱ g(x_۱,x_۲,...,x_n)dx_۱dx_۲...dx_n$$

مورد توجه باشد. راه حل روش مونت کارلو جهت براورد θ در این است که θ را بتوان به‌صورت امید ریاضی زیر بیان کرد.

$$\theta=E[g(U_۱,...,U_n)]$$

که در آن $U_۱,U_۲,...,U_n$ متغیر های تصادفی مستقل یکنواخت در $(۰,۱)$ می‌باشند. بنابراین اگر k مجموعه مستقل

$$\begin{array}{c} U_۱^۱,...,U_n^۱ \\ U_۱^۲,...,U_n^۲ \\ \vdots \\ U_۱^k,...,U_n^k \end{array}$$

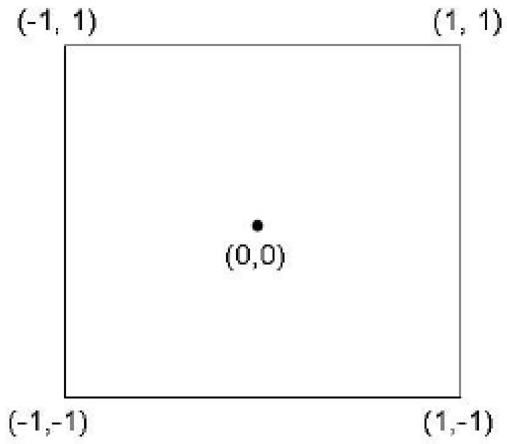
را که هریک شامل n متغیر تصادفی مستقل یکنواخت در $(۰,۱)$ می‌باشند تولید کنیم و چون تمام متغیرهای تصادفی $g(U_i,...,U_i)$، $i=۱,۲,...,k$ مستقل و همتوزیع با میانگین θ هستند لذا می‌توانیم θ را با

$$\frac{۱}{k}\sum_{i=۱}^{k} g(U_۱^i,...,U_n^i)$$

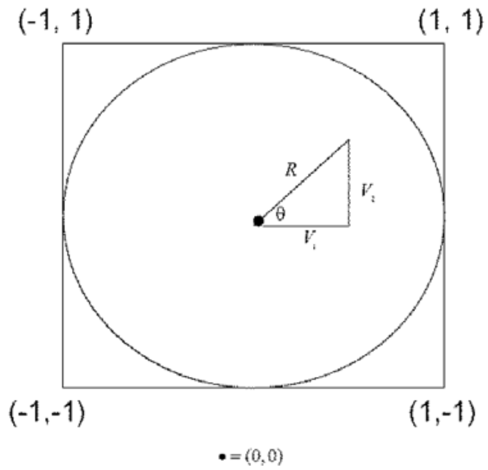
براورد کنیم.

۳ براورد π

فرض کنید بردار تصادفی (X,Y) در مربعی با مساحت ۴ و به مرکز مبدا مختصات دارای توزیع یکنواخت باشد (شکل ۱ را ببینید). اکنون احتمال واقع شدن یک نقطه تصادفی در داخل دایره‌ای به شعاع یک را که در مربع محاط شده است در نظر می‌گیریم (شکل ۲).



شکل ۱.



شکل ۲.

چون (X,Y) در مربع بطور یکنواخت توزیع شده است بنا بر این احتمال واقع شدن (X,Y) در دایره برابر است با

$$P(X^{۲}+Y^{۲}\leq ۱)=\frac{\pi}{۴}$$

بنابراین اگر نقاط تصادفی زیادی را در مربع تولید کنیم نسبت نقاطی که در داخل دایره واقع می شوند تقریبا $\frac{\pi}{۴}$ است. حال اگر X و Y مستقل بوده و هر دو در $(-۱, ۱)$ دارای توزیع یکنواخت باشند چگالی توام آن‌ها به‌صورت زیر خواهد بود:

$$f(x,y)=f(x)f(y)=\frac{۱}{۲}\frac{۱}{۲}=\frac{۱}{۴}$$
$$-۱\leq x\leq ۱\ \ ,-۱\leq y\leq ۱$$

چون تابع چگالی (X,Y) در مربع ثابت است لذا توزیع (X,Y) در مربع یکنواخت است. اکنون اگر U در $(۰, ۱)$ یکنواخت باشد آنگاه $۲U$ در $(۰, ۲)$ یکنواخت بوده و بنابراین $۲U-۱$ در $(-۱, ۱)$ یکنواخت است. بنابراین اگر اعداد تصادفی $U_۱$ و $U_۲$ را تولید نموده و قرار دهیم $X=۲U_۱-۱$، و $Y=۲U_۲-۱$

$$I=\begin{cases} ۱ & X^{۲}+Y^{۲}\leq ۱ \\ ۰ & o.w. \end{cases}$$

آنگاه خواهیم داشت:

$$E(I)=P(X^{۲}+Y^{۲}\leq ۱)=\frac{\pi}{۴}$$

بنابراین با تولید نمودن زوجهای زیادی از اعداد تصادفی $U_۱$ و $U_۲$ و برآورد $\frac{\pi}{۴}$ بوسیله کسری از این زو ها که

$$(۲U_۱-۱)^{۲}+(۲U_۲-۱)^{۲}\leq ۱$$

می‌توانیم π را براورد کنیم.

اکنون می‌خواهیم واریانس براورد π را با استفاده از مشروط سازی کاهش دهیم. همان‌طور که می‌دانیم فرمول واریانس شرطی به‌صورت زیر است.

$$Var(X)=E[Var(X|Y)]+Var[E(X|Y)]$$

با توجه به اینکه واریانس همواره نامنفی است جملات سمت راست تساوی فوق نامنفی‌اند و داریم:

$$Var(X)\geq Var[E(X|Y)] \qquad (۱)$$

اکنون فرض کنید می‌خواهیم با انجام یک مطالعه شبیه سازی مقدار $\theta=E(X)$ را معین کنیم. همچنین فرض کنید یک متغیر دوم مانند Y وجود دارد طوریکه $E(X|Y)$ معلوم بوده و مقداری را اختیار می‌کند که از شبیه سازی می‌توان آن را تعیین نمود. چون

$$E\{E(X|Y)\}=E(X)=\theta$$

نتیجه می‌گیریم که $E(X|Y)$ نیز یک براورد کننده‌ی نااریب θ بوده و با توجه به (۱) نتیجه می‌شود که $E(X|Y)$ براوردکننده‌ای بهتر از براوردکننده‌ی X است.

اکنون با استفاده از مشروط سازی براورد π را بدست می آوریم. دیدیم که برای براورد π از تعیین دفعاتی که نقطه بطور تصادفی انتخاب شده در داخل مربعی به مساحت ۴ و به مرکز مبدا، در

داخل دایره‌ی محاطی به شعاع یک واقع می‌شود، استفاده کردیم. با فرض $V_i = 2U_i - 1$، $i = 1, 2$ که U_i عدد تصادفی از توزیع یکنواخت $(0, 1)$ است و

$$I = \begin{cases} 1 & V_1^2 + V_2^2 \leq 1 \\ 0 & o.w. \end{cases}$$

آنگاه مشاهده کردیم که $E(I) = \frac{\pi}{4}$. با استفاده از متوسط مقادیر I و با بکار بردن $E(I|V_1)$ به جای I می‌توان براورد $\frac{\pi}{4}$ را بهتر نمود. با توجه به استقلال V_1 و V_2 و اینکه V_2 در $(-1, 1)$ یکنواخت است داریم:

$$\begin{aligned} E[I|V_1 = v] &= P\{V_1^2 + V_2^2 \leq 1 | V_1 = v\} \\ &= P\{v^2 + V_2^2 \leq 1 | V_1 = v\} \\ &= P\{V_2^2 \leq 1 - v^2\} \\ &= P\{-(1 - v^2)^{(1/2)} \\ &\leq V_2 \leq (1 - v^2)^{(1/2)}\} \\ &= \int_{-(1-v^2)^{(1/2)}}^{(1-v^2)^{(1/2)}} (\tfrac{1}{2}) dx \\ &= (1 - v^2)^{(1/2)} \end{aligned}$$

بنابراین

$$E(I|V_1) = (1 - V_1^2)^{(1/2)}$$

و لذا براوردکننده $(1 - V_1^2)^{(1/2)}$ نیز دارای میانگین $\frac{\pi}{4}$ بوده و واریانسی کمتر از I دارد. چون

$$\begin{aligned} E[(1 - V_1^2)^{(1/2)}] &= \int_{-1}^{1} (1 - x^2)^{(1/2)} (\tfrac{1}{2}) dx \\ &= \int_{0}^{1} (1 - x^2)^{(1/2)} dx \\ &= E[(1 - U^2)^{(1/2)}] \end{aligned}$$

بنابراین آن را با استفاده از براوردکننده‌ی $(1 - U^2)^{(1/2)}$ که U عددی تصادفی است، تا اندازه‌ای می‌توان ساده کرد. داریم:

$$\begin{aligned} Var[(1 - U^2)^{(1/2)}] &= E[1 - U^2] - (\tfrac{\pi}{4})^2 \\ &= \frac{2}{3} - (\tfrac{\pi}{4})^2 = 0/0498 \end{aligned}$$

که در تساوی اول از $Var(W) = E[W^2] - (E[W])^2$ استفاده می‌شود. از طرف دیگر چون I یک متغیر تصادفی برنولی با میانگین $\frac{\pi}{4}$ است داریم:

$$Var(I) = (\tfrac{\pi}{4})(1 - \tfrac{\pi}{4}) = 0/1686$$

که نشان می‌دهد مشروط‌سازی واریانس را به اندازه ۷۰/۴۴ درصد کاهش می‌دهد. همان‌طور که می‌دانیم واریانس برآوردکننده‌ها دارای اهمیت می‌باشند و به عنوان نشانی از کارایی براوردگر می‌باشد. همچنین توجه داریم که هرچه واریانس کمتر باشد مقدار شبیه سازی لازم برای حصول دقتی ثابت کمتر می‌شود.

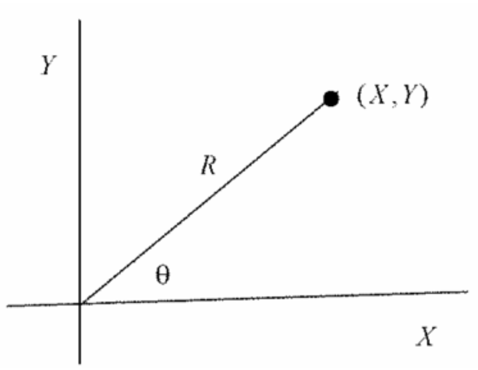
۴ تولید داده از توزیع نرمال

در یک مطالعه شبیه سازی مهم این است که بتوانیم داده‌هایی را تولید کنیم که نمایانگر مقدار متغیر تصادفی مورد نظر باشد. محققان بسیاری در زمینه تولید داده از متغیرهای تصادفی، روش‌هایی را پیشنهاد کرده اند.(ببینید داگ پونار [۳]، دوری [۶]، اکویر [۱۰] و اسچمیسر[۱۷]) در این بخش

روش‌های مختلف تولید داده از توزیع نرمال را که یکی از توزیع‌های مهم آماری می‌باشد، بیان می‌کنیم.

۱.۴ تولید داده از توزیع نرمال با استفاده از تبدیلات باکس—مولر

فرض کنید X و Y متغیرهای تصادفی نرمال استاندارد بوده و R و θ نشان‌دهنده مختصات قطبی بردار (X,Y) باشد. (شکل ۳ را ببینید)



شکل ۳

یعنی

$$R^2 = X^2 + Y^2 \qquad , \quad tan(\theta) = \frac{Y}{X}$$

چون X و Y مستقل هستند لذا چگالی توام آن‌ها برابر حصلضرب چگالی‌های آن‌ها خواهد بود که به‌صورت زیر می‌باشد.

$$f(x,y) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \ \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} = \frac{1}{2\pi}e^{-\frac{x^2+y^2}{2}} \quad (2)$$

برای تعیین چگالی توام R^2 و θ که آن را $f_{R^2,\theta}(d,\theta)$ نشان می‌دهیم از تغییر متغیر زیر استفاده می‌کنیم:

$$d = x^2 + y^2 \qquad , \quad \theta = tan^{-1}(\frac{y}{x}).$$

به‌سادگی می‌توان دید که ژاکوبین این تبدیل برابر ۲ است بنابراین با توجه به معادله (۲) تابع چگالی توام R^2 و θ به‌صورت زیر است:

$$f_{R^2,\theta}(d,\theta) = \frac{1}{2} \ \frac{1}{\sqrt{2\pi}}e^{-\frac{d}{2}}$$
$$0 < d < \infty \ , 0 < \theta < 2\pi$$

همانطور که مشاهده می کنیم تابع چگالی توام برابر حاصلضرب یک چگالی نمایی با میانگین ۲ (یعنی $\frac{1}{2}e^{-\frac{d}{2}}$) و یک چگالی یکنواخت در $(0, 2\pi)$ (یعنی $\frac{1}{2\pi}$) است بنابراین R^2 و θ که R^2 نمایی با میانگین ۲ و θ دارای توزیع یکنواخت در $(0, 2\pi)$ می باشد مستقل‌اند.

حال می‌توانیم یک جفت متغیر تصادفی نرمال مستقل X و Y را تولید کنیم بدین ترتیب که ابتدا مختصات قطبی آن‌ها را تولید نموده و سپس آن‌ها را به مختصات مستطیلی برمی‌گردانیم که این کار به‌صورت زیر انجام می‌شود.

مرحله۱) اعداد تصادفی U_1 و U_2 را تولید می‌کنیم.

مرحله۲) قرار دهید $R^2 = -2log(U_1)$ و $\theta = 2\pi U_2$. (بدین ترتیب R^2 دارای توزیع نمایی با میانگین ۲ و θ دارای توزیع یکنواخت بین ۰ و 2π است)

مرحله۳) اکنون قرار دهید

$$X = R\ cos(\theta) = \sqrt{-۲log(U_۱)}\ cos(۲\pi U_۲)$$
$$Y = R\ sin(\theta) = \sqrt{-۲log(U_۱)}\ sin(۲\pi U_۲)$$

این تبدیلات به تبدیلات باکس—مولر[۳] معروف است.

متاسفانه استفاده از تبدیلهای باکس—مولر برای تولید یک جفت نرمال واحد از نظر محاسبه چندان کارا نیست که دلیل آن لزوم محاسبه‌ی توابع مثلثاتی سینوس و کسینوس است. با محاسبه غیر مستقیم سینوس و کسینوس یک زاویه تصادفی راهی برای رهایی از مشکل اتلاف وقت وجود دارد، که این روش را در زیربخش زیر بیان می کنیم.

۲.۴ روش قطبی برای تولید داده از توزیع نرمال

برای شروع توجه داریم که اگر U در $(۰,۱)$ یکنواخت باشد آنگاه $۲U$ در $(۰,۲)$ و بنابراین $۲U - ۱$ در $(-۱,۱)$ یکنواخت است. بنابراین اگر اعداد تصادفی $U_۱$ و $U_۲$ را تولید نموده و قرار دهیم

$$V_۱ = ۲U_۱ - ۱ \quad , \quad V_۲ = ۲U_۲ - ۱$$

آنگاه $(V_۱, V_۲)$ در مربعی به مساحت ۴ و به مرکز $(۰,۰)$ دارای توزیع یکنواخت می باشد(شکل ۲ را ببینید). اکنون فرض کنید یک چنین جفتهای $(V_۱, V_۲)$ را دایما تولید کنیم تا یکی را که در دایره‌ای به شعاع یک و مرکز $(۰,۰)$ قرار دارد پیدا کنیم یعنی $(V_۱, V_۲)$ به طوری که $V_۱^۲ + V_۲^۲ \leq ۱$. حال نتیجه می گیریم که یک چنین جفت $(V_۱, V_۲)$ دارای توزیع یکنواخت در دایره است. اگر R و θ نشان دهنده‌ی مختصات قطبی این جفت باشد آنگاه اثبات استقلال R و θ که توزیع $R^۲$ در $(۰,۱)$ یکنواخت است مشکل نمی باشد. بنابراین چون θ زاویه ای است تصادفی لذا با تولید کردن یک نقطه تصادفی $(V_۱, V_۲)$ در دایره و قرار دادن

$$sin\theta = \frac{V_۲}{R} = \frac{V_۲}{(V_۱^۲ + V_۲^۲)^{۱/۲}}$$
$$cos\theta = \frac{V_۱}{R} = \frac{V_۱}{(V_۱^۲ + V_۲^۲)^{۱/۲}}$$

می توانیم سینوس و کسینوس یک زاویه تصادفی θ را تولید کنیم. حال از تبدیلات باکس—مولر نتیجه می‌شود که با تولید کردن یک عدد تصادفی U و قرار دادن

$$X = (-۲logU)^{۱/۲}\ \frac{V_۱}{(V_۱^۲ + V_۲^۲)^{۱/۲}}$$
$$Y = (-۲logU)^{۱/۲}\ \frac{V_۲}{(V_۱^۲ + V_۲^۲)^{۱/۲}}$$

می‌توانیم نرمالهای استاندارد مستقل را تولید کنیم. در حقیقت چون $R^۲ = V_۱^۲ + V_۲^۲$ در $(۰,۱)$ دارای توزیع یکنواخت و مستقل از زاویه تصادفی θ است، از آن مانند عدد تصادفی U که در معادلات بالا لازم بود می‌توان استفاده کرد. بنابراین با فرض $S = R^۲$، بدست می آوریم که

$$X = (-۲logS)^{۱/۲}\ \frac{V_۱}{S^{۱/۲}} = V_۱\ (\frac{-۲logS}{S})^{۱/۲}$$
$$Y = (-۲logS)^{۱/۲}\ \frac{V_۲}{S^{۱/۲}} = V_۲\ (\frac{-۲logS}{S})^{۱/۲}$$

[۳] Box-Muller

نرمال های استاندارد مستقلند هرگاه $(V_۱, V_۲)$ نقطه‌ای باشد که در دایره ای به شعاع یک و مرکز مبدا مختصات به تصادف انتخاب می‌شود و $S = V_۱^۲ + V_۲^۲$.

بنابراین بطور خلاصه مراحل زیر را برای تولید یک جفت نرمال استاندارد مستقل داریم.

مرحله۱) اعداد تصادفی $U_۱$ و $U_۲$ را تولید کنید.

مرحله۲) قرار دهید $V_۱ = ۲U_۱ - ۱$، $V_۲ = ۲U_۲ - ۱$ و $S = V_۱^۲ + V_۲^۲$.

مرحله۳) اگر $S > ۱$ است به مرحله ۱ برگردید.

مرحله۴) قرار دهید

$$X = \sqrt{\frac{-۲logS}{S}}\, V_۱ \quad , \quad Y = \sqrt{\frac{-۲logS}{S}}\, V_۲$$

این روش را روش قطبی [۴] می نامند. چون احتمال اینکه یک نقطه تصادفی واقع در مربع در داخل دایره قرار گیرد برابر $\frac{\pi}{۴}$ (یعنی مساحت دایره بخش بر مساحت مربع) است بنابراین روش قطبی بطور متوسط به $\frac{۴}{\pi} = ۱/۲۷۳$ تکرار مرحله نیاز دارد.

۳.۴ روش عدم پذیرش

فرض کنید برای تولید یک متغیر تصادفی که دارای تابع چگالی $g(x)$ است روشی داریم. از این امر می توان به عنوان مبنایی برای تولید از توزیع پیوسته‌ای که دارای تابع چگالی $f(x)$ است با تولید Y از g و سپس پذیرش این مقدار تولید شده با احتمالی مناسب با $\frac{f(y)}{g(y)}$ استفاده کرد. بویژه فرض کنید c ثابتی باشد که

$$\frac{f(y)}{g(y)} \leq c$$

برای تمام y ها، آنگاه روش عدم پذیرش [۵] برای تولید یک متغیر تصادفی که دارای چگالی f است بصورت زیر می‌باشد.

مرحله۱) Y را که دارای چگالی g می باشد تولید کنید.

مرحله۲) یک عدد تصادفی U را تولید کنید.

مرحله۳) اگر $U \leq \frac{f(y)}{cg(y)}$ است قرار دهید $X = Y$ در غیر اینصورت به مرحله ۱ برگردید.

قضیه زیر را که اثبات آن در راس [۱۶] آمده است، در مورد روش عدم پذیرش داریم.

قضیه ۱ ۱) متغیر تصادفی که با روش عدم پذیرش تولید می کنیم دارای چگالی f است.
۲) تعداد تکرارهای مورد نیاز الگوریتم یک متغیر تصادفی هندسی با میانگین c است.

حال برای تولید یک متغیر تصادفی نرمال استاندارد ابتدا توجه داریم که قدرمطلق Z دارای تابع چگالی احتمال زیر است.

$$f(x) = \frac{۲}{۲\pi} e^{-\frac{x^۲}{۲}} \quad x > ۰ \qquad (۳)$$

حال از این تابع چگالی با استفاده از روش عدم پذیرش، که g تابع چگالی نمایی با میانگین یک در نظر می‌گیریم یعنی

$$g(x) = e^{-x} \quad x > ۰$$

---
[۴] Polar Method
[۵] Rejection Method

داده تولید می کنیم. داریم

$$\frac{f(x)}{g(x)} = \sqrt{\frac{۲}{\pi}} e^{-x-\frac{x^{۲}}{۲}}$$

و بنابراین مقدار ماکزیمم $\frac{f(x)}{g(x)}$ در مقدار x ی که $-x-\frac{x^{۲}}{۲}$ را ماکزیمم می کند پیش می آید. محاسبه نشان می دهد که این در $x = ۱$ پیش می دهد و بنابراین

$$c = max\{\frac{f(x)}{g(x)}\} = \frac{f(۱)}{g(۱)} = \sqrt{\frac{۲e}{\pi}}$$

$$\longrightarrow \frac{f(x)}{cg(x)} = e^{-x-\frac{x^{۲}}{۲}} = e^{-\frac{(x-۱)^{۲}}{۲}}$$

بنابراین مقدار قدرمطلق یک متغیر تصادفی نرمال استاندارد را به طریق زیر می‌توان تولید نمود.
مرحله۱) Y یعنی یک متغیر تصادفی نمایی با میانگین یک را تولید کنید.
مرحله۲) یک عدد تصادفی U را تولید کنید.
مرحله۳) اگر $U \leq exp\{\frac{(Y-۱)^{۲}}{۲}\}$ قرار دهید $X = Y$، در غیر اینصورت به مرحله ۱ برگردید.
به محض اینکه یک متغیر تصادفی X که تابع چگالی آن مانند (۳) است شبیه سازی کردیم و از آن جا که توزیع چنین متغیر تصادفی مانند مقدار قدرمطلق یک نرمال استاندارد است می‌توانیم یک نرمال استاندارد Z را با فرض اینکه با احتمال مساوی X و $-X$ است بدست آوریم.
در مرحله ۳ اگر $U \leq exp\{-\frac{(Y-۱)^{۲}}{۲}\}$ یا معادل آن $-logU \geq \frac{(Y-۱)^{۲}}{۲}$ باشد مقدار Y را می‌پذیریم. اما می‌دانیم که $-logU$ دارای توزیع نمایی با میانگین یک است. بنابراین مراحل زیر را می‌توان جایگزین مراحل بالا کرد.
مرحله۱) متغیرهای نمایی مستقل $Y_۱$ و $Y_۲$ را با میانگین یک تولید کنید.
مرحله۲) اگر $Y_۲ \geq \frac{(Y_۱-۱)^{۲}}{۲}$ باشد قرار دهید $X = Y_۱$، در غیر اینصورت به مرحله ۱ برگردید.
چون $c = \sqrt{\frac{۲e}{\pi}} = ۱/۳۲۲$ بنابراین طبق قضیه ۱ تعداد تکرارهای مرحله ۲ از یک توزیع هندسی با میانگین ۱/۳۲ پیروی می‌کند.

۴.۴ روش جدید تولید داده از توزیع نرمال

این روش براساس تقریب پولیا[۶] [۱۴]، که این تقریب در جانسون و کاتز [۷] نیز آمده است، می‌باشد که تابع توزیع نرمال استاندارد را به‌صورت زیر تقریب می‌زند.

$$\Phi(x) \simeq \frac{۱}{۲}[۱ + \{۱ - exp(-\frac{۲x^{۲}}{\pi})\}^{۱/۲}]$$

که Φ تابع توزیع نرمال استاندارد است. بیشترین خطای این تقریب ۰/۰۰۳ است وقتی که $x = ۱/۶$ باشد.
از طرفی یک روش کلی برای تولید یک متغیر تصادفی پیوسته با تابع توزیع F، که روش تبدیل وارون نامیده می‌شود مبتنی بر قضیه زیر است.

قضیه ۲ فرض کنید U یک متغیر تصادفی یکنواخت در $(۰, ۱)$ باشد. برای هر تابع توزیع

[۶] Polya

پیوسته F متغیر تصادفی X که با $X = F^{-1}(U)$ تعریف می‌شود دارای تابع توزیع F است. ($F^{-1}(u)$ آن مقدار x تعریف می شود که $F(x) = u$).

قضیه بالا نشان می‌دهد که با تولید کردن متغیر تصادفی U و سپس قرار دادن $X = F^{-1}(U)$ می‌توانیم متغیر تصادفی X را از تابع توزیع پیوسته F تولید نماییم.

حال با استفاده از این تقریب و استفاده از روش تبدیل وارون برای تولید داده از توزیع نرمال داریم.

$$U = \Phi(x) = \frac{۱}{۲}[۱ + \{۱ - exp(-\frac{۲x^۲}{\pi})\}^{۱/۲}]$$
$$\longrightarrow x^۲ = -\frac{\pi}{۲} log\{۱ - (۲U - ۱)^۲\}$$
$$\longrightarrow x = \pm\sqrt{-\frac{\pi}{۲} log\{۱ - (۲U - ۱)^۲\}}$$

بنابراین بطور خلاصه مراحل زیر را برای تولید داده از توزیع نرمال استاندارد داریم.

مرحله ۱) عداد تصادفی U را از توزیع یکنواخت در $(۰, ۱)$ تولید کنید.

مرحله ۲) اگر $U < ۰/۵$ قرار دهید

$$X = +\sqrt{-\frac{\pi}{۲} log\{۱ - (۲U - ۱)^۲\}}$$

در غیر این‌صورت

$$X = -\sqrt{-\frac{\pi}{۲} log\{۱ - (۲U - ۱)^۲\}}$$

توجه داریم که تقریب پولیا را می‌توان به‌صورت زیر نوشت.

$$(۲\Phi(x) - ۱)^۲ \simeq F(x^۲) \qquad (۴)$$

که F تابع توزیع نمایی با میانگین $\frac{\pi}{۲}$ می باشد. بنابراین (۴) رابطه تقریبی بین توابع توزیع نرمال و نمایی را بیان می کند.

برای تولید داده از توزیع نرمال با میانگین μ و واریانس $\sigma^۲$، $N(\mu, \sigma^۲)$، ابتدا با استفاده از روشهای موجود از توزیع نرمال استاندارد داده تولید می‌کنیم و سپس با استفاده از قضیه زیر که اثبات آن ساده می باشد از توزیع نرمال $N(\mu, \sigma^۲)$ داده تولید می‌کنیم.

قضیه ۳ اگر $Z \sim N(۰, ۱)$ آنگاه $X = \sigma Z + \mu \sim N(\mu, \sigma^۲)$.

۵ بحث و نتیجه گیری

در این مقاله کاربرد هایی از اعداد تصادفی را بیان کردیم و دیدیم که مسایل را می توان بطور ساده‌ای به کمک اعداد تصادفی و شبیه سازی حل کرد. همچنین مسایلی را که راه حل نظری مشکلی یا غیر قابل حل می باشند، از طریق شبیه سازی می توان آنها را حل کرد. مانند یافتن توزیع تجربی آماره‌ها یا برآوردگرهایی که از طریق نظری بدست آوردن توزیع آن‌ها مشکل می‌باشد. مقایسه برآوردگرهای مختلف، مقایسه توان آزمونهای مخلتف با یکدیگر و غیره را می‌توان به کمک شبیه سازی بررسی کرد. در این مقاله علاوه بر بررسی سه روش موجود برای تولید داده از توزیع نرمال، با استفاده از تقریب پولیا روشی جدید برای تولید داده از توزیع نرمال معرفی کردیم.

لازم به ذکر است که در سالهای اخیر با پیشرفت رایانه ها دیگر نباید نگران زمان مصرف شده برای محاسبه بود، زیرا با سرعت فوق العاده‌ای که این رایانه ها دارند در مدتی بسیار محدود می‌توانند هزاران عدد تصادفی تولید کنند.

مراجع

[1] Alizadeh Noughabi, H. (2010), A new estimator of entropy and its application in testing normality, to appear in *Journal of Statistical Computation and Simulation*.

[2] Alizadeh Noughabi, H. and Arghami N.R. (2010), Monte Carlo comparison of seven normality tests, to appear in *Journal of Statistical Computation and Simulation*.

[3] Dagpunar, T. (1988), *Principles of Random Variate Generation*, Clarendon Press, Oxford.

[4] Demirtas H. (2005), Pseudo-random number generation in R for some univariate distributions, *Journal of Modern Applied Statistical Methods*, 3, 300-311.

[5] Diaz-Emparanza I. (2002), Is a small Monte Carlo analysis a good analysis? Checking the size, power and consistency of a simulation-based test, *Statistical Papers*, 43, 567-577.

[6] Devroye, L. (1986), *Random Variate Generation*, Springer-Verlag, New York.

[7] Johnson, N. L. and Kotz, S. (1970), *Continuouse Univariate distributions.* , John Wiely, New York.

[8] Halabi S, Singh B. (2004), Sample size determination for comparing several survival curves with unequal allocations, *Statistics in Medicine*, 23, 1793-1815.

[9] Hodgson T, Burke M. (2000), On simulation and the teaching of statistics, *Teaching Statistics*, 22, 91-96.

[10] L'Ecuyer, P. (1988), *Random Numbers for Simulation*, Commun. Assoc. Comput. Math. 33, 22-25.

[11] Kenna LA, Sheiner LB. (2004), Estimating treatment effect in the presence of non-compliance measured with error: precision and robustness of data analysis methods, *Statistics in Medicine*, 23, 3561-3580.

[12] Williams, M.S., Ebel, E.D. and Wagner, B.A. (2007), Monte Carlo approaches for determining power and sample size in low-prevalence applications, *Preventive Veterinary Medicine*, 82, 151-158

[13] Marsaglia, G. (1972), *The Structure of Linear Congruential Sequences*, in Applications of Number Theory to Numerical Analaysis, S. K. Zaremba, ed., Academic Press, London, 249-255.

[14] Ploya, G. (1945), Remarks on computing the probability integral in one the two dimensions, *Proceedings of the 1st Berkely Symposium on Mathematical Statistics and Probability*, 63-78.

[15] Robert, C. P. and Casella, G. (1999), *Monte Carlo Statistical Methods*, New York: Springer.

[16] Ross, S. M. (2002), *Simulation*, 3rd ed., Academic Press.

[17] Schmeiser, B. W. (1980), Random Variate Generation, *Simulation Conf.*, Orlando, 79-104.