

نمونه گیری گیبس^۱

سعید زال زاده^۲

چکیده

در این مقاله تلاش شده است تا با ارائه مثالهای ساده و بدون پرداختن به مباحث نظری (نظیر اثبات همگرایی و یا مقایسه با روشهای دیگر)، روش نمونه گیری گیبس را توضیح داده و برخی از کاربردهای آن را معرفی کنیم.

۱. مقدمه

تکامل روزافزون رایانه و دسترسی آسان به رایانه های سریع، باعث شده است که روشهای الگوریتمی همچون نمونه گیری گیبس، ابزارهایی مفید و ضروری برای حل مسائل آماری باشند که حل آنها به صورت تحلیلی میسر نیست. نمونه گیری گیبس یک روش الگوریتمی (مثل الگوریتم EM) است که ایده اولیه آن توسط متروپلیس و تله^۳ در سال ۱۹۵۳ ارائه و سپس با مقاله ای که گیمان و گیمان^۴ [۴] در مورد مدلهای پردازش تصویر ارائه دادند، وارد مرحله جدیدی شد.

نمونه گیری گیبس اساساً مبتنی بر خاصیت مارکوفی فرایندهای تصادفی و روشهای مونت کارلوست، به طوری که گاهی نمونه های تولید شده با استفاده از این روش را نمونه مونت کارلو نیز می نامند.

۲. نمونه گیری گیبس چیست؟

نمونه گیری گیبس را می توان یک ابزار شبیه سازی برای به دست آوردن نمونه های تصادفی از تابع توزیع توأم نرمال نشده، دانست. این نمونه ها، نمونه هایی را نیز از توزیعهای حاشیه ای مربوط به توزیع توأم تولید می کنند.

در بسیاری از مسائل آماری با مجموعه ای از متغیرهای تصادفی سروکار داریم که علاقه مند به توزیعهای حاشیه ای آنها یا اطلاعاتی در مورد آنها هستیم. در چنین مواقعی اگر توزیع توأم نرمال شده این متغیرها موجود باشد، می توانیم چگالیهای حاشیه ای را به طور تحلیلی و دقیق به دست آوریم. غالباً در بسیاری از این مسائل، توزیع توأم متغیرها در عمل نامعلوم است، اما مجموعه ای از توزیعهای شرطی وجود دارد. اگر مجموعه مناسبی از چگالیهای شرطی موجود باشد، می توان با توسل به روشهای مونت کارلو و به کارگیری تکنیکهای نمونه گیری مجدد^۵ مانند نمونه گیری جانشین ساز (جایگشتی)^۶ و یا نمونه گیری گیبس، یک نمونه مؤثر از چگالی توأم و احتمالاً حاشیه ای مورد نظر را تولید کرده و سپس به وسیله نمونه تولید شده، تابع چگالی و یا خصیصه های مربوط به آن را با درجه اطمینان دلخواهی محاسبه کرد.

نمونه گیری گیبس در واقع راهی برای تولید نمونه تصادفی از یک تابع چگالی حاشیه ای است، بدون آن که نیازی به محاسبه خود تابع چگالی باشد. بیشترین کاربرد نمونه گیری گیبس، استفاده از آن برای به دست آوردن توزیعهای پسین در مدلهای بیزی است. علاوه بر این، در روشهای کلاسیک نیز (برای محاسبه توابع درستنمایی و تعیین

^۱ Geman, S. & Geman, D.
^۵ Resampling Technique
^۶ Substitution Sampling

^۱ Gibbs Sampling
^۲ گروه ریاضی، دانشکده علوم پایه، دانشگاه سمنان
^۳ Metropolis and Teller

$$Y'_0, X'_0, Y'_1, X'_1, Y'_2, X'_2, \dots, Y'_k, X'_k \quad (3)$$

که آن را دنباله گیس^۷ می نامند، انجام می شود. برای تولید دنباله گیس (۳) ابتدا مقدار اولیه $Y'_0 = y'_0$ را تعیین کرده و سپس (۳) را به طور مکرر و متناوب با استفاده از

$$X'_i \sim f(x | Y'_i = y'_i) \quad (4)$$

$$Y'_{i+1} \sim f(y | X'_i = x'_i)$$

تولید می کنیم. تحت شرایط کلی، به طور منطقی، ثابت می شود که وقتی $k \rightarrow \infty$ توزیع X'_k به $f(x)$ ، توزیع حاشیه ای X ، همگرا می شود. بنابراین برای k به اندازه کافی بزرگ، مشاهدات نهایی $X'_k = x'_k$ در (۳) واقعا یک نمونه مؤثر از نقاط $f(x)$ است. برای مثال، اسمیت^۸ m دنباله گیس مستقل را با طول k تولید کرد و سپس مقادیر نهایی X'_k را از هر یک از دنباله ها به کار برد. در این روش اگر k به اندازه کافی بزرگ انتخاب شود، یک نمونه i.i.d. از $f(x)$ به دست می آید. برای روشن شدن مطلب، روش بالا را در مثالهای زیر توضیح می دهیم.

مثال ۱- فرض کنید X و Y دو متغیر تصادفی با تابع چگالی توأم زیر باشند:

$$f(x, y) \propto \binom{n}{x} y^{x+\alpha+1} (1-y)^{n-x+\beta-1} \quad (5)$$

$$; \quad x=0,1,\dots,n, \quad 0 \leq y \leq 1$$

همچنین فرض کنید علاقه مند به محاسبه برخی از مشخصات $f(x)$ (توزیع حاشیه ای X) باشیم. با استفاده از نمونه گیری گیس یک نمونه از توزیع حاشیه ای X را تولید می کنیم. از (۵) (بدون این که به α ، β و n بستگی داشته باشد) نتیجه می شود که

$$\begin{aligned} f(x|y) &\sim \text{Binomial}(n, y) \\ f(y|x) &\sim \text{Beta}(x+\alpha, n-x+\beta) \end{aligned} \quad (6)$$

با به کارگیری طرح مکرر (۴) برای توزیعهای (۶) می توانیم یک نمونه $x_1, x_2, \dots, x_m$ را از $f(x)$ تولید کرده و سپس با استفاده از آن مشخصات مورد نظر را برآورد کنیم. ممکن است خواننده عنوان کند

^۷ Gibbs sequence

^۸ Smith

برآوردهای (MLE)، استفاده از این روش بسیار سودمند است.

فرض کنید توزیع توأم $f(x, y_1, \dots, y_p)$ معلوم باشد و بخواهیم برخی از مشخصات توزیع حاشیه ای

$$f(x) = \int \dots \int f(x, y_1, \dots, y_p) dy_1 \dots dy_p \quad (1)$$

را مثل میانگین و واریانس به دست آوریم. احتمالا به طور معمول باید $f(x)$ را محاسبه کرده و با استفاده از آن، مشخصات مورد نظر را محاسبه کنیم.

در بسیاری از حالتها، محاسبه انتگرال (۱) مستقما یا با کمک آنالیز عددی بسیار مشکل است. در چنین حالتهایی نمونه گیری گیس راه حل دیگری برای به دست آوردن $f(x)$ ارائه می کند.

به جای محاسبه یا تقریب $f(x)$ به طور مستقیم، با استفاده از نمونه گیری گیس می توان یک نمونه مؤثر $x_1, \dots, x_m$ از توزیع $f(x)$ را بدون نیاز به $f(x)$ تولید کرد. با شبیه سازی یک نمونه به اندازه کافی بزرگ، میانگین و واریانس یا هر مشخصه دیگر $f(x)$ را می توان با درجه اطمینان دلخواهی به دست آورد. واقعیت مهم این است که نتایج نهایی محاسبات، هر چند که بر اساس شبیه سازی است ولی برآوردهایی برای کمیتهای جامعه هستند. به عنوان مثال، برای محاسبه میانگین $f(x)$ از

$$\frac{1}{m} \sum_{i=1}^m x_i$$

و این حقیقت که با احتمال یک داریم:

$$\frac{1}{m} \sum_{i=1}^m X_i \rightarrow E(X) \quad (2)$$

استفاده می کنیم. بنابراین وقتی که m به اندازه کافی بزرگ اختیار شود، هر مشخصه جامعه، حتی خود چگالی را می توان با درجه اطمینان دلخواه به دست آورد. برای درک نمونه گیری گیس ابتدا حالت دو متغیره را بررسی می کنیم.

۱-۲ حالت دو متغیره

دو متغیر تصادفی X و Y را در نظر بگیرید. در نمونه گیری گیس برای تولید یک نمونه از چگالی $f(x)$ ، از چگالیهای شرطی $f(y|x)$ و $f(x|y)$ (که اغلب در مدلهای آماری معلوم اند) به جای خود $f(x)$ استفاده می شود. این کار با تولید یک دنباله از متغیرهای تصادفی

همچنین جزئیات مربوط به شرایط همگرایی، توسط کارلین و شرویش^۹ در سال ۱۹۹۰ مورد بررسی قرار گرفته است.

در این مثال توزیع حاشیه‌ای به سادگی قابل محاسبه نیست ولی با استفاده از نمونه گیری گیس برای شرطهای (۸) هر مشخصه $f(x)$ را می‌توان به دست آورد. شکل (۲) بافتنگار یک نمونه $m = 500$ تایی از $f(x)$ را که با استفاده از مقادیر نهایی دنباله گیس به طول $k = 15$ برای $B = 5$ به دست آمده، نشان می‌دهد.

۲-۲ برآورد چگالیهای حاشیه‌ای

نمونه گیری گیس را می‌توان برای برآورد خود چگالی حاشیه‌ای نیز به کار برد. در دنباله گیس (۳) $X'_k = x'_k$ یک تحقق از $f(x)$ یعنی $x_1, \dots, x_m$ از توزیع $f(x)$ و $Y'_k = y'_k$ یک تحقق از $f(y)$ یعنی $y_1, \dots, y_m$ از توزیع $f(y)$ هستند. علاوه بر این میانگین چگالیهای شرطی $f(x | Y_k = y_k)$ به $f(x)$ میل می‌کند. بنابراین برای محاسبه $f(x)$ کافی است میانگین مقادیر نهایی چگالیهای شرطی دنباله گیس را محاسبه کنیم، یعنی $f(x)$ را با استفاده از

$$f(x) = \frac{1}{m} \sum_{i=1}^m f(x | y_i) \quad (9)$$

برآورد نماییم. در رابطه (۹) $y_1, \dots, y_m$ مقادیر نهایی به دست آمده دنباله گیس هستند. پشتوانه فرمول (۹) این است که

$$E[f(x | y)] = \int f(x | y) f(y) dy = f(x) \quad (10)$$

روش برآورد چگالی (۹) را می‌توان برای توزیعهای گسسته نیز به کار برد. مثلاً اگر $y_1, \dots, y_m$ مقادیر نهایی y در m نمونه گیس برای چگالیهای شرطی مثال (۱) باشند، می‌توان متناظر با (۹) احتمالهای حاشیه‌ای X را با استفاده از

$$\hat{P}(X = x) = \frac{1}{m} \sum_{i=1}^m P(X = x | Y_i = y_i) \quad (11)$$

برآورد کرد. این روش را برای برآورد چگالی $f(x)$ در حالت سه متغیره در مثال ۳ به کار می‌بریم.

۳-۲ حالت سه متغیره

واقعیتی که نمونه گیری گیس از آن استفاده می‌کند این است که اطلاعات نهفته در توزیعهای شرطی، برای تعیین توزیع توأم (در صورت

که در این مثال نیازی به نمونه گیری گیس نیست، زیرا می‌توان $f(x)$ را از رابطه (۵) به صورت توزیع بتا-دوجمله‌ای به دست آورد و سپس مشخصات مورد نیاز $f(x)$ را مستقیماً با استفاده از آن و بدون آن که خود را با توزیعهای شرطی درگیر کنیم، محاسبه کرد. به هر حال، این مثال، مثالی ساده برای نشان دادن این موضوع است که:

$$f(x) = \binom{n}{x} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(x + \alpha)\Gamma(n - x + \beta)}{\Gamma(\alpha + \beta + n)} \quad (v)$$

$$; x = 0, 1, \dots, n$$

شکل (۱) بافتنگار دو نمونه $x_1, x_2, \dots, x_m$ را با $m = 500$ ، $\alpha = 2$ و $\beta = 4$ ، $n = 16$ که به دو روش از چگالی $f(x)$ تولید شده‌اند، نشان می‌دهد. در یک روش، این نمونه مستقیماً از چگالی (۷) تولید شده و در روش دیگر، این نمونه با استفاده از چگالیهای شرطی (۶) و طرح نمونه گیری گیس با $k = 15$ تولید شده است. تشابه دو بافتنگار در شکل (۱) تأییدی بر این ادعاست که طرح نمونه گیری گیس واقعاً متغیرهایی از توزیع حاشیه‌ای را تولید می‌کند. این دو نمونه به وسیله برنامه‌ای که با نرم‌افزار s-plus نوشته شده است، به دست آمده‌اند.

نکته‌ای که باید به آن توجه داشته باشیم این است که نمونه گیری گیس نیازی به توزیع توأم که با استفاده از رابطه

$$f(x) = \frac{f(x, y)}{f(y | x)}$$

$f(x)$ را محاسبه کند، ندارد. مثال بعد نشان می‌دهد که نمونه گیری گیس برای وقتی که $f(x)$ و $f(y)$ یا $f(x, y)$ قابل محاسبه نباشند، ضروری است.

مثال ۲- فرض کنید X و Y دارای توزیعهای شرطی نمایی محدود شده به بازه $(0, B)$ باشند، یعنی

$$f(x | y) \propto ye^{-yx}; 0 < x < B < \infty \quad (8)$$

$$f(y | x) \propto xe^{-xy}; 0 < y < B < \infty$$

که در آن B یک ثابت معلوم است. محدودیت به بازه $(0, B)$ ما را مطمئن می‌کند که توزیع حاشیه‌ای $f(x)$ وجود دارد. اگر B متناهی نباشد، آنگاه چگالیهای (۸) یک زوج معتبر از چگالیهای شرطی نیست. به این مفهوم که چگالی توأم $f(x, y)$ وجود ندارد و در این صورت دنباله گیس همگرا نخواهد شد. شرایط تعیین چگالیهای حاشیه‌ای و

روش مکرر (۱۲) یک نمونه

$$Y'_0, Z'_0, X'_0, Y'_1, Z'_1, X'_1, Y'_2, Z'_2 \quad (13)$$

را با این خاصیت که برای مقادیر بزرگ k ، $X'_k = x'_k$ یک نمونه مؤثر از $f(x)$ است، تولید می‌کند.

مثال ۳- در توزیع مثال (۱)، N را یک متغیر تصادفی بواسون با پارامتر λ در نظر می‌گیریم و لذا توزیع توأم به صورت

$$f(x, y, n) \propto \binom{n}{x} y^{x+\alpha-1} (1-y)^{n-x+\beta-1} e^{-\lambda} \frac{\lambda^n}{n!} \quad (14)$$

نوشته می‌شود. فرض کنید علاقه‌مند به توزیع حاشیه‌ای X باشیم. برخلاف مثال (۱)، در این جا نمی‌توان توزیع حاشیه‌ای X را به سرعت محاسبه کرد. با توجه به (۱۲) سه چگالی شرطی کامل را نیاز داریم. این چگالی‌ها به صورت زیر به دست می‌آیند:

$$\begin{aligned} f(x|y, n) &\sim \text{Binomial}(n, y) \\ f(y|x, n) &\sim \text{beta}(x+\alpha, n-x+\beta) \quad (15) \\ f(n|x, y) &\sim e^{-(1-y)} \frac{[(1-y)\lambda]^{n-x}}{(n-x)!} \lambda \end{aligned}$$

$$; n = x, x+1, \dots$$

اگر m نمونه گیس (۱۳) را تولید کنیم، می‌توانیم با استفاده از رابطه‌ای شبیه به رابطه (۱۱) چگالی X را به صورت زیر برآورد کنیم:

$$\begin{aligned} \hat{P}(X=x) = \\ \frac{1}{m} \sum_{i=1}^n \sum_{j=1}^m P(X=x | Y_i = y_i, N = n_j) \quad (16) \end{aligned}$$

در رابطه (۱۶) $x_1, \dots, x_m$ و $y_1, \dots, y_m$ مقادیر به دست آمده از m نمونه گیس اند. شکل (۳) برآورد احتمالات توزیع حاشیه‌ای X را که با استفاده از (۱۶) بر اساس یک نمونه $m=500$ تایی با دنباله‌هایی به طول $k=10$ برای سه چگالی (۱۵) با $n=16$ ، $\alpha=2$ و $\beta=4$ به دست آمده است، نشان می‌دهد.

۲-۴ حالت چند متغیره

برای سادگی کار، توزیعها را به طور کلی با براکت (II) نشان می‌دهیم. مثلاً توزیعهای توأم، شرطی و حاشیه‌ای را به صورت $[X, Y]$ ، $[X|Y]$ و $[X]$ نشان می‌دهیم. همچنین علامت * را برای حاصلضرب

وجود کافی است. در حالت دو متغیره نمونه‌گیری گیس برای به دست آوردن چگالیهای حاشیه‌ای از چگالیهای شرطی به سرعت پیش می‌رود. اما در حالت چند متغیره، روابط میان چگالیهای شرطی، حاشیه‌ای و توأم بسیار پیچیده می‌شوند. به عنوان مثال رابطه

$$\text{حاشیه‌ای} \times \text{شرطی} = \text{توأم}$$

برای تمام چگالیهای شرطی و چگالیهای حاشیه‌ای به دست نمی‌آید. بدین معنی که ممکن است راههای متعددی برای ایجاد یک چگالی حاشیه‌ای از چگالیهای شرطی وجود داشته باشد و بتوان مجموعه‌های مختلفی از توزیعهای شرطی را برای محاسبه توزیع حاشیه‌ای مورد نظر به کار برد.

تعریف - فرض کنید $U_1, \dots, U_k$ ، k متغیر تصادفی باشند. چگالیهای شرطی هر متغیر به شرط $(k-1)$ متغیر دیگر را یک چگالی شرطی کامل^{۱۱} یا شرطی یک متغیره می‌نامیم. بنابراین برای این k متغیر، k چگالی شرطی کامل

$$\begin{aligned} f_{U_i|U_1, \dots, U_{k-1}}(u_i | u_1, \dots, u_{k-1}); i=1, 2, \dots, k \\ i \neq j, t \in \{1, \dots, k-1\} \end{aligned}$$

وجود دارند.

در نمونه‌گیری گیس همیشه مجموعه تمام چگالیهای شرطی کامل برای تعریف تکرارها به کار برده می‌شوند. بساگ^{۱۱} در سال ۱۹۷۴ نشان داد که این مجموعه از چگالیهای شرطی برای تعیین توزیع توأم (و در نتیجه برای تعیین هر توزیع حاشیه‌ای) به طور یکتا کافی است و این کوچکترین مجموعه لازم از چگالیهای شرطی برای تعریف تکرارهاست [۲].

فرض کنید برای سه متغیر تصادفی X ، Y و Z چگالیهای شرطی کامل یعنی $f_{X|YZ}$ ، $f_{Y|XZ}$ و $f_{Z|XY}$ وجود داشته باشند. نمونه‌گیری گیس در هر تکرار از این چگالیهای شرطی نمونه‌گیری می‌کند، یعنی ۱- امین تکرار در نمونه‌گیری گیس به صورت زیر است:

$$\begin{aligned} X'_j &\sim f(x | Y'_j = y'_j, Z'_j = z'_j) \\ Y'_{j+1} &\sim f(y | X'_j = x'_j, Z'_j = z'_j) \\ Z'_{j+1} &\sim f(z | X'_j = x'_j, Y'_{j+1} = y'_{j+1}) \end{aligned} \quad (17)$$

^{۱۱} Full Conditional

^{۱۱} Besag

تولید می کنیم. با این روند هر یک از متغیرها به نوبت در یک چرخه ظاهر شده و هر چرخه شامل k متغیر تصادفی تولید شده است. پس از i تکرار از این چرخه ها، $(U_1^{(i)}, \dots, U_k^{(i)})$ تولید می شود. گیمان و گیمان در سال ۱۹۸۴ ثابت کردند که وقتی $i \rightarrow \infty$ داریم: [۴]

$$(U_1^{(i)}, \dots, U_k^{(i)}) \xrightarrow{d} [U_1, \dots, U_k]$$

و همچنین

$$U_s^{(i)} \xrightarrow{d} U_s \sim [U_s]$$

چگالیها به کار می بریم. حال با استفاده از این نماد گذاری، روند نمونه گیری گیس را می توان برای k متغیر به صورت زیر بیان کرد.

فرض کنید $U_1, U_2, \dots, U_k$ ، k متغیر تصادفی باشند که چگالیهای شرطی کامل آنها وجود دارند. همچنین فرض کنید $U_1^{(0)}, \dots, U_k^{(0)}$ یک مجموعه آغازی دلخواه باشد. در این صورت $U_1^{(1)}, \dots, U_k^{(1)}$ را به ترتیب از

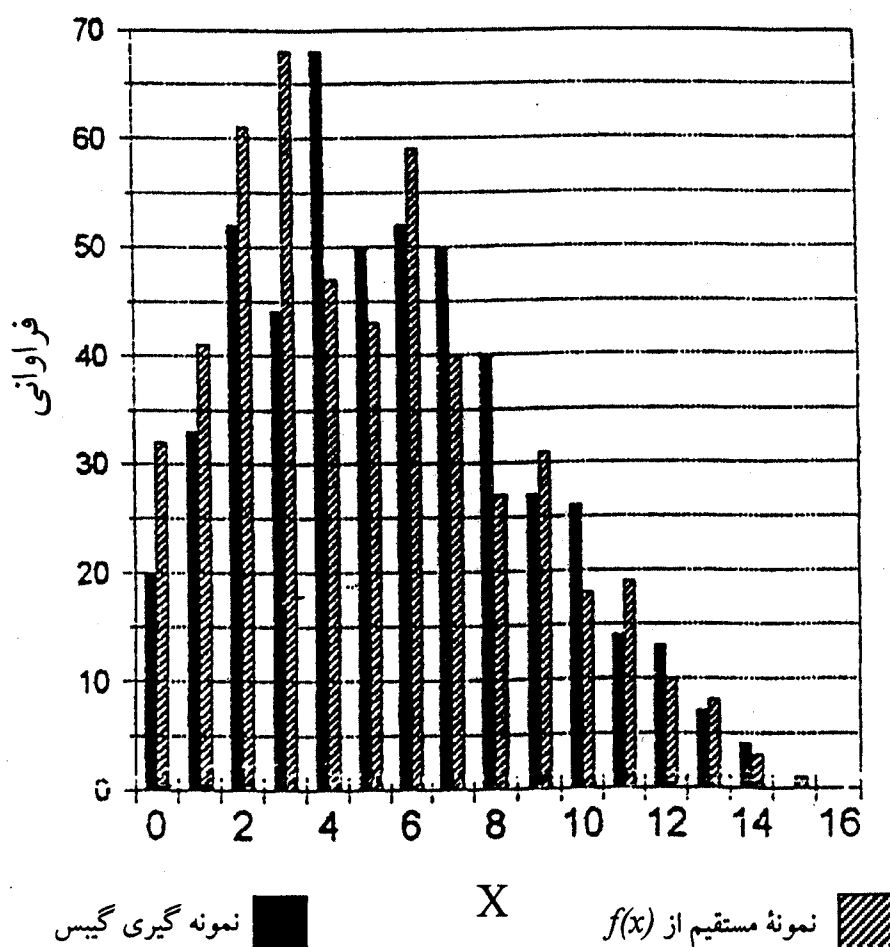
$$U_1^{(1)} \sim [U_1 | U_2^{(0)}, U_3^{(0)}, \dots, U_k^{(0)}]$$

$$U_2^{(1)} \sim [U_2 | U_1^{(1)}, U_3^{(0)}, \dots, U_k^{(0)}]$$

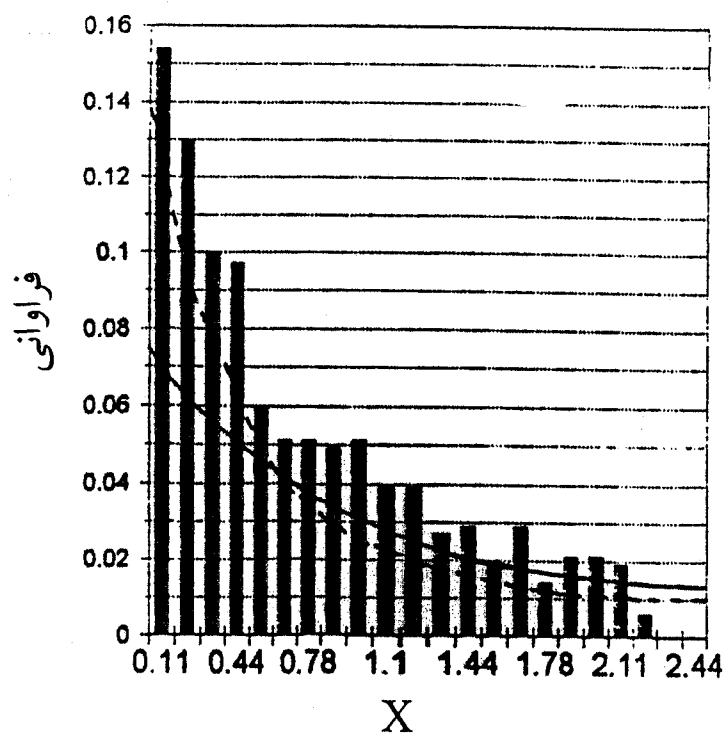
$$U_3^{(1)} \sim [U_3 | U_1^{(1)}, U_2^{(1)}, U_4^{(0)}, \dots, U_k^{(0)}]$$

:

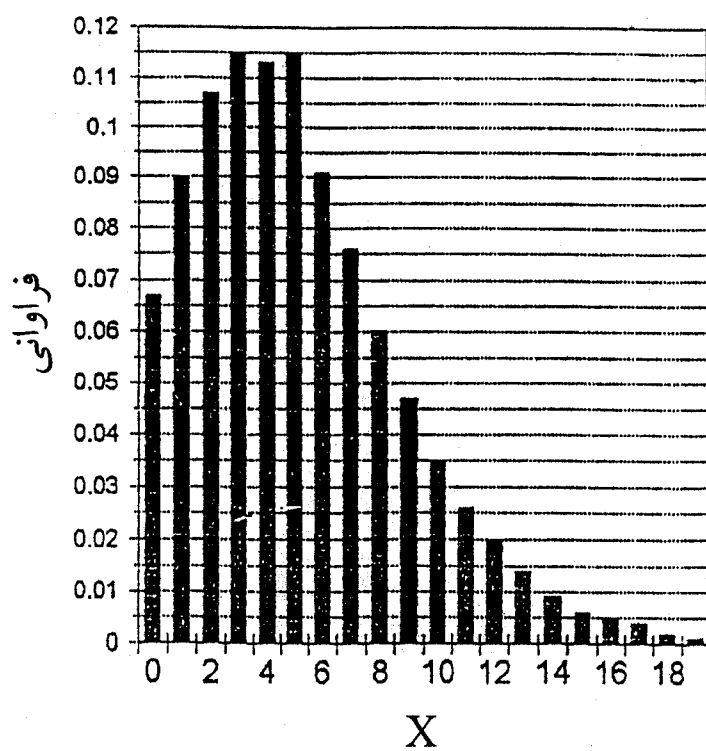
$$U_k^{(1)} \sim [U_k | U_1^{(1)}, \dots, U_{k-1}^{(1)}]$$



شکل ۱



شکل ۲



شکل ۳

مراجع

- [1] Casella, G. and George, E., 1992, *Explaining the Gibbs Sampler*, American Statistical Association, 46, 167-174.
- [2] Gelfand, A. E. and Smith, A. F. M., 1990, *Sampling-Based Approaches to Calculating Marginal Densities*, Journal of Statistical Association, 85, 398-409.
- [3] Gelfand, A. E., 2000, *Gibbs Sampling*, American Statistical Association, 95, 1300-1304.
- [4] Geman, S. and Geman, D., 1984, *Stochastic Relaxation, Gibbs Distribution and the Bayesian Restoration of Images*, IEEE Trans. Pattern Anal. Mach. Intell., 6, 721-741.

طرح یک مسئله احتمال

در هر کاری هر چقدر هم که احتمال موفقیت ناچیز باشد، با تلاشهای مکرر می توان احتمال موفقیت را تا حدی بسیار زیاد بالا برد. به عنوان مثال اگر احتمال موفقیت در کاری یک ششم باشد، با شش بار تلاش، احتمال حداقل یک بار موفقیت، بیشتر از ۶۳ درصد است. همچنین اگر احتمال موفقیت یک دهم باشد، با ده بار تلاش، احتمال حداقل یک بار موفق شدن نیز بیشتر از ۶۳ درصد است. حال اگر احتمال موفقیت در کاری بسیار کم مثلا معادل یک درصد باشد، آیا باید از تلاش نمودن ناامید شویم؟ هرگز؛ چون این بار نیز با ۱۰۰ بار تلاش، احتمال حداقل یک بار موفقیت ما به میزان بیشتر از ۶۳ درصد است و این شیوه برای کلیه احتمالات بسیار ناچیز، صادق است. به عنوان مثال به راحتی ثابت می شود که احتمال آنکه یک سکه را یک میلیون بار پرتاب کنیم و در تمامی پرتابها شیر بیاید، معادل مقدار $1/2^{1000000}$ است. حال اگر فرصت داشته باشیم و به تعداد $2^{1000000}$ بار تلاش کنیم، احتمال اینکه حداقل یک بار موفق شویم که یک میلیون بار پشت سر هم شیر بیاوریم، بیشتر از ۶۳ درصد است. در خصوص کارهای غیر قانونی نیز این مسأله باعث می شود که در نهایت در اثر تکرار جرم، احتمال دستگیری مجرمان بسیار زیاد باشد. به عنوان مثال، اگر احتمال دستگیری در یک سرقت، معادل $1/20$ باشد و این عمل مستقلا ۲۰ بار تکرار شود، احتمال حداقل یک دستگیری بیشتر از ۶۳ درصد است.

می توان نشان داد که اگر احتمال موفقیت در کاری مستقل و معادل $1/N$ باشد و N بار تلاش کنیم، احتمال حداقل یک بار موفق شدن بیشتر از ۶۳ درصد است. N عدد طبیعی بزرگتر از یک است. (این عدد در واقع معادل 0.632100 است که تعداد ارقام بعد از ممیز آن نامحدود است.)