خودآموز آمار و تحلیل کمّی در آزمون‌های

CFA و FRM

همراه با ارائه و بررسی ۱۵۰ مسأله

(قابل استفاده برای تمامی رشته‌های مدیریت، حسابداری و اقتصاد)

جلد دوم

تألیف: دکتر زهره طغرایی

انتشارات پندار پارس

سرشناسه	:	طغرایی، زهره، ۱۳۵۳-
عنوان و نام پدیدآور	:	خودآموز آمار و تحلیل کمی در آزمون‌های FRM و CFA همراه با حل تشریحی بیش از ۱۵۰ مسأله.../ تألیف زهره طغرایی .
شخصات نشر	:	تهران : پندار پارس ، ۱۴۰۴ -
شخصات ظاهری	:	ج.
نابک	:	ج۱. 978-622-7785-47-0 : ؛ ج۲. 978-622-7785-58-6
ضعیت فهرست نویسی	:	فیپا
وضوع	:	سرمایه‌گذاری -- تجزیه و تحلیل -- آزمون‌ها Investment analysis -- Examinations مدیریت سبد سرمایه‌گذاری -- آزمون‌ها Portfolio management-- Examinations
ده بندی کنگره	:	HG۴۵۲۹
ده بندی دیویی	:	۶/۳۳۲
شماره کتابشناسی ملی	:	۱۰۱۲۰۶۸۵
طلاعات رکورد کتابشناسی	:	فیپا

انتشارات پندارپارس

دفتر فروش: انقلاب، ابتدای کارگرجنوبی، کوی رشتچی، شماره ۱۴، واحد ۱۶
www.pendarepars.com

تلفن: ۶۶۵۷۲۳۳۵ - ۶۶۹۲۶۵۷۸ همراه: ۰۹۱۲۲۴۵۲۳۴۸

نام کتاب : **خودآموز آمار و تحلیل کمّی در آزمون‌های FRM و CFA همراه با ارائه و بررسی ۱۵۰ مسأله (قابل استفاده برای تمامی رشته‌های مدیریت، حسابداری و اقتصاد)- جلد۲**

ناشر : **انتشارات پندار پارس**

تالیف : **دکتر زهره طغرایی**

چاپ نخست : **بهمن ماه ۱۴۰۴**

شمارگان : **۲۰۰ نسخه**

ویراستار : **دکتر بهناز رضایی**

چاپ، صحافی : **وحید** **شابک : ۶-۵۸-۷۷۸۵-۶۲۲-۹۷۸**

قیمت : **۵۰۰.۰۰۰ تومان** **شابک دوره : ۳-۵۹-۷۷۸۵-۶۲۲-۹۷۸**

تقدیم به پدر و مادر عزیزم که
آموختن را از آنان آموختم

فهرست

پیش‌گفتار

در دنیای امروز، دانش مالی به یکی از ارکان اساسی تصمیم‌گیری‌های حرفه‌ای تبدیل شده است. آزمون‌های معتبر بین‌المللی همچون FRM[1] و CFA[2] به متخصصان مالی این امکان را می‌دهند که مهارت‌های خود را به سطحی بالاتر برده و جایگاهی برجسته در دنیای مالی کسب نمایند. یکی از کلیدی‌ترین مباحث این آزمون‌ها، آمار و تحلیل کمّی شامل مدل‌های مختلف آماری و یادگیری ماشین است؛ دانشی که نه تنها در امتحانات، بلکه در مسیر حرفه‌ای هر تحلیلگر مالی نقشی حیاتی ایفا می‌کند.

این کتاب به عنوان یک راهنمای خودآموز جامع طراحی شده است تا شرکت‌کنندگان برای شرکت در دو آزمون مذکور بتوانند با زبانی ساده و مثال‌های کاربردی، مفاهیم پیچیده مدل‌سازی‌های یادگیری ماشین و موضوعاتی همچون شبیه‌سازی و آزمون‌های ناپارامتری را درک کنند. همچنین، مثال‌های متعدد در حوزه مالی و بانکی در ارتباط با این موضوعات ارائه و بررسی شده‌اند.

به دلیل گستردگی ابعاد مختلف این روش‌ها بدیهی است، سعی شده تنها به جنبه‌های بنیادی پرداخته شود تا خواننده، فلسفه اصلی آنها را تا اندازه‌ای درک نموده و در ادامه مسیر یادگیری، علاوه بر حل مسائل دو آزمون مذکور، در صورتی که علاقمند به پیاده‌سازی این روش‌ها در محیط‌های برنامه‌نویسی همچون پایتون باشد، قادر به تحلیل خروجی‌های این پیاده‌سازی‌ها گردد. از این‌رو، همه دانشجویان رشته‌های مدیریت، حسابداری، اقتصاد و علاقمندان به یادگیری آمار و مدل‌سازی نیز می‌توانند از محتوای این کتاب بهره ببرند. هدف نویسنده این بوده که پیچیدگی‌های ریاضی را با توضیحات روشن و قابل فهم کاهش داده و ابزارهای تحلیل را به شیوه‌ای عملی ارائه نماید.

در جلد نخست این کتاب به مفاهیم پایه‌ای آمار و نیز مدل‌سازی رگرسیون پرداختیم و اکنون از آن مفاهیم در درک روش‌های پیشرفته‌تری که استفاده از آنها در دو دهه اخیر به طور گسترده و فزاینده‌ای در حوزه‌های مالی و اقتصادی رشد چشمگیری داشته بهره می‌بریم.

نویسنده با تکیه بر سال‌ها تجربه در تدریس و تحقیق در حوزه تحلیل داده در رشته‌های مختلف علمی، سعی بر آن نموده است که مباحث پایه‌ای در این کتاب را به شیوه‌ای ساده و کاربردی تشریح نماید. امیدواریم این کتاب بتواند راهگشای شما در مسیر موفقیت باشد و به شما کمک کند که با "اعتماد به نفس و دانش کافی" آزمون‌های پیش رو را پشت سر بگذارید. موفقیت شما، افتخار ماست.

در پایان از دوست عزیزم سرکار خانم دکتر بهناز رضایی که با دقت نظر قابل توجهی در ویراستاری این کتاب، کمال اهتمام را داشته‌اند سپاسگزارم.

---

[1] Financial Risk Management
[2] Chartered Financial Analyst

خواننده گرامی، دریافت نظرها، پیشنهادها، و انتقادهای شما در بهبود کیفیت این کتاب مهم و مؤثر است. با نشانی ایمیل زیر می‌توانید با نویسنده مکاتبه فرمایید:

zohreh_toghrayee@yahoo.com

https://www.linkedin.com/in/zohreh-toghrayee-3339-a668/

زهره طغرایی

پاییز ۱۴۰۴

فصل نخست

شبیه‌سازی آماری

مقدمه

برای توضیح شبیه‌سازی، ابتدا به واژه" شبیه‌سازی" دقت کنید. به نظر می‌رسد این کلمه می‌تواند معنیِ ساختن یک چیز، شبیه چیز دیگر را برساند. در این صورت، سؤالی که پیش می‌آید این است که چه چیزی، شبیه چه چیزی ساخته می‌شود؟ در دنیای بیرونی این کار به وفور انجام می‌شود. برای مثال، مجسمه بزرگان هنر و علم ساخته می‌شود. در واقع تصویر این مجسمه‌ها شبیه تصویر واقعی و یا تصویر تقریبی فرد مورد نظر است. یا اگر دیده باشید، از شهرهای تاریخی ماکت‌هایی در مقیاس خیلی کوچک می‌سازند. در حالت خاص‌تر، به منظور پیشگیری از هزینه‌های گزاف و صرف زمان غیر قابل توجیه، ابتدا ماکتی شبیه به شهری که قرار است ساخته شود با جزئیات ساخته می‌شود تا تمام ابعاد و چالش‌های احتمالیِ آن مشخص گردد. بنابراین، به نظر می‌رسد از یک چیز واقعی، تصویر و یا ویژگی‌های کلیدی در دسترس داریم و شبیهِ آن چیز را بر اساس آن تصویر و ویژگی‌ها در مقیاس کوچک و یا حتی مقیاس یکسان ولی در فضایی متفاوت که نسبت به فضای اصلی یک فضایی مصنوعی است می‌سازیم.
از نظر علمی، منظور از شبیه‌سازی، ساختن یک فرآیند یا یک سیستم در شرایط مصنوعی برای بررسی ابعاد مختلف ویژگی‌های آن سیستم و یا با هدف آموزش است. اما، شبیه‌سازی آماری چیست؟ اصولاً در آمار و احتمالات چه چیزی لازم است که شبیه‌سازی شود؟

بخاطر بیاورید در دنیای واقعی وقتی برای بررسی یک ادّعا آزمایش تصادفی را به دفعات انجام داده و نمونه‌ای از مشاهدات به‌دست می‌آوریم، با فرض اینکه توزیع متغیر وابسته نرمال است، می‌توانیم یک رابطه رگرسیون خطی بین یک متغیر وابسته و چند متغیر مستقل برازش دهیم. نرمال بودنِ توزیع داده‌های متغیر وابسته، از فرض‌های اصلی یک مدل رگرسیون خطی است و زمانی که این مدل برای اولین بار ارائه شد از نظر تئوری اثبات شد که این فرض لازم است برقرار باشد تا برآوردهای به‌دست آمده در این مدل، ویژگی‌های مطلوب برای انجام استنباط آماری را داشته باشند. ارزیابی این فرض، پس از اجرای مدل با بررسی نرمال بودن توزیع باقیمانده‌های آن انجام می‌شود که در صورت نقض این فرض، راه حل‌هایی ارائه شده است[1].

اکنون، فرض کنید یک مدلِ آماری جدید پیشنهاد داده‌اید که فرض اصلی‌اش این است که داده‌های مورد نظر، از توزیع نرمال برخوردار باشد. ارزیابی این روش جدید، مانند ناااریب بودن برآوردگرهای پارامترهای مدل مذکور، در حالت معمول باید از طریق تئوری اثبات شود. ولی در بسیاری از مواقع، پیچیدگی مدل باعث می‌شود این اثبات دشوار و یا غیرممکن باشد. در این صورت، چگونه می‌توانید این روش را ارزیابی نمایید؟ آیا با اخذ نمونه‌ای در یک حوزه علمی مشخص (مانند پزشکی) به عنوان یک مجموعه داده می‌توانید این کار را انجام دهید؟ پاسخ خیر است. چرا که غالبا وقتی یک روش آماری جدید ارائه می‌شود، به یک حوزه خاص مثل بانکداری یا پزشکی اختصاص ندارد و به صورت عام تعریف شده و در هر رشته علمی قابل استفاده است. پس راه حل چیست؟ ما که به تمام داده‌های همه رشته‌های علمی با ویژگی نرمال بودن توزیع داده‌ها دسترسی نداریم. چه کار کنیم؟

[1] فصل هشتم جلد نخست همین کتاب

برای پاسخ به این پرسش می‌توانیم بگوییم درست است که به دنیای واقعی دسترسی نداریم، ولی می‌توانیم شبیهِ آن را بسازیم. برای این‌کار ابتدا از خود می‌پرسیم به چه چیزی نیاز داریم؟ پاسخ این است: داده‌هایی با توزیع نرمال به شکل زیر:

$$N(\mu, \sigma^2)$$

پس باید این داده‌ها را به حجم بسیار بزرگ تولید کنیم که حجم بسیار بزرگ به این دلیل است که بسیاری از روش‌های آماری برای برخوردار بودن از ویژگی‌های مناسب مانند نااریب بودن و دقت بالا، نیاز به فرض حجم نمونه بزرگ دارند. این یک مثال از شبیه‌سازی آماری است؛ یعنی تولید داده‌هایی از توزیع نرمال.

مثال‌هایی از شبیه‌سازی

در این قسمت، پیش از پرداختن به کاربردهای شبیه‌سازی آماری، ابتدا بدون استفاده از دانش آماری درباره شبیه‌سازی، دو مثال شرح داده و سپس به برخی از کاربردهای شبیه‌سازی پرداخته می‌شود.

مثال ۱-۱: فرض کنید نمی‌دانیم در پرتاب یک تاس، احتمالِ ظاهر شدن عدد ۶ چقدر است. این احتمال را به‌دست آورید.
پاسخ۱: به منظور محاسبه این احتمال، ۱۰۰۰ بار تاس را پرتاب نموده و تعداد دفعاتی که عدد ۶ ظاهر شده است را بر ۱۰۰۰ تقسیم می‌نماییم. نسبت به‌دست آمده را می‌توان به عنوان احتمال ظاهر شدن عدد ۶ در پرتاب تاس در نظر گرفت. انتظار داریم احتمال مورد نظر ۱/۶ به‌دست آید.

پاسخ ۲: شبیه‌سازی آماری: یکی از توزیع‌های گسسته، توزیع یکنواخت گسسته است. گوییم متغیر تصادفی X دارای توزیع یکنواخت گسسته است اگر فرمول توزیع احتمال آن به صورت زیر باشد:

$$P(X = x) = \frac{1}{b - a + 1} \quad ; a < x < b$$

شبیه‌سازی این توزیع آماری بسیار ساده است. گویی می‌خواهیم یک عدد تصادفی بین دو عدد a و b تولید کنیم. در مثال قبل، هدف از شبیه‌سازی، تولید یک عدد بین ۱ و ۶ است و بر اساس توزیع یکنواخت گسسته، احتمال انتخاب هر یک از این اعداد برابر است با:

$$\frac{1}{6 - 1 + 1} = \frac{1}{6}$$

پس اگر بتوان از توزیع یکنواخت در بازه $(1,6)$ به حجم ۱۰۰۰ عدد تولید کنیم مثل این است که تاسی را هزار بار پرتاب کرده‌ایم.

مثال ۱-۲: به یک بازیکن بیسبال هر بار فرصت داده می‌شود تا دو ضربه بزند. احتمال اینکه ضربه موفقیت آمیز بزند ۰.۱ است. به این بازیکن فرصت داده می‌شود تا دو ضربه بزند. می‌خواهیم احتمال این را به‌دست آوریم که هر دو ضربه را با موفقیت بزند. این کار را با استفاده از شبیه‌سازی انجام دهید.

برای شبیه‌سازی این مسئله، از دانش پایه آمار و ریاضی استفاده می‌کنیم. احتمال ضربه موفقیت آمیز ۰.۱ است و برای یافتن یک مورد بسیار ساده که بتوان با ۰.۱ تطابق داد، بخاطر می‌آوریم که در ریاضیات ۱۰ رقم داریم که تمام اعداد طبیعی با این ده رقم ساخته می‌شوند:

۰ و ۱ و ۲ و ۳ و ۴ و ۵ و ۶ و ۷ و ۸ و ۹

فرض می‌کنیم عدد ۲ متناظر با ضربه موفقیت آمیز است. یعنی اگر از بین این ۱۰ رقم، یک رقم به تصادف انتخاب کنیم، اگر ۲ ظاهر شود یعنی ضربه بازیکن موفقیت آمیز بوده است که احتمال رخ دادن آن ۰.۱ است. اگر دو رقم به تصادف انتخاب کنیم و هر دو رقم انتخاب شده ۲ باشند، یعنی هر دو ضربه بازیکن موفقیت آمیز بوده است. بنابراین، برای شبیه‌سازی این مسأله، دو بار انتخاب یک رقم از میان این ۱۰ رقم روش خوبی به نظر می‌رسد.

اکنون فرض کنید این کار توسط یک برنامه کامپیوتری انجام شود و اعداد دو رقمی مانند ۲۳ و ۵۴ و ۷۲ و .... تولید شود. بنابراین، اگر عدد ۲۲ انتخاب شده باشد، به این معنی است که بازیکن در بازی مورد نظر دو ضربه موفقیت‌آمیز زده است. با این فرض، ۵۰۰ بار اعداد دو رقمی را به تصادف انتخاب نماییم. نسبت دفعاتی که عدد ۲۲ تولید شده است را می‌توان به عنوان احتمال این در نظر بگیریم که بازیکن بیسبال در یک بازی هر دو ضربه را با موفقیت می‌زند.

مثال ۱-۳: فرض کنید شخصی از میان سهام‌های موجود به تصادف، یک سهام انتخاب می‌نماید. احتمال اینکه سهام انتخاب شده سهامی با بازده بسیار بالا باشد ۰.۱ است. در هر بار این شخص، اجازه دارد به تصادف دو سهام انتخاب نماید. با استفاده از شبیه‌سازی، احتمال اینکه هر دو سهام انتخابی وی، بازده بسیار بالا داشته باشد را محاسبه نمایید.

مانند مثال قبل، شبیه‌سازی را انجام می‌دهیم و احتمال مورد نظر را به‌دست می‌آوریم.

اکنون، فرض کنید بتوان پدیده‌های مطرح شده در مسائل مورد نظر را به صورت متغیرهای تصادفی تعریف نمود که از توزیع‌های آماری مشخصی برخوردار هستند. در این صورت، چگونه می‌توان از شبیه‌سازی آماری برای حل مسأله بهره برد؟ توزیع‌های آماری، هسته مرکزیِ شبیه‌سازی‌های آماری هستند؛ چرا که هدف از شبیه‌سازی‌های آماری در اغلب موارد تولید مقادیر از یک توزیع آماری شناخته شده است.

کلاس‌های شبیه‌سازی آماری

شبیه‌سازی آماری به معنی تولید داده است که می‌تواند به دو شکل کلی انجام شود. در هر دو شکل، به دفعات بسیار زیاد مثلا ۵۰۰۰ بار یا بیشتر، تولید داده انجام می‌شود که هر یک از این دفعات، منتهی به یک مجموعه داده با حجم نمونه مشخص می‌گردد. به این ترتیب، ۵۰۰۰ مجموعه داده با حجم یکسان در اختیار داریم. برای مثال ۵۰۰۰ مجموعه داده که حجم هر یک ۱۰۰ است، بر اساس یک شبیه‌سازی تولید می‌شود. این دو رویکردِ تولید داده به صورت دو کلاسِ رایج از شبیه‌سازی‌های آماری بوده و به شکل زیر است:

۱. شبیه‌سازی‌هایی که در آن، توزیع آماری مشخص است و داده‌ها از این توزیع آماری تولید می‌شوند. از جمله این روش‌ها می‌توان به روش شبیه‌سازی مونت کارلو اشاره داشت. در مورد توزیع‌های آماریِ پیچیده‌تر مانند توزیع‌های پسین در تحلیل بیز[1] نیز از روش‌های پیشرفته‌ای مانند *زنجیره مارکوف مونت کارلو*[2] بهره برد.

۲. توزیع آماری مشخص نیست و تنها یک نمونه از جامعه در دسترس داریم و می‌خواهیم درباره پارامتر جامعه برآورد انجام دهیم. روش نمونه‌گیری بوت استرپ[3] و جک نایف[4] از جمله روش‌های این گروه است.

پیش از پرداختن به معرفی دو روش شناخته شده از دو رویکرد فوق، ابتدا به معرفی مفاهیم مقدماتی در شبیه‌سازی آماری می‌پردازیم و در قسمت‌های بعد، با تفصیل بیشتر به معرفی مفاهیم مقدماتی شبیه‌سازی آماری و سپس روش مونت کارلو که از قدیمی‌ترین روش‌های شبیه‌سازی آماری رویکرد اول است و نیز نمونه‌گیری بوت استرپ که از روش‌های کاربردی از رویکرد دوم می‌باشد می‌پردازیم.

مفاهیم مقدماتی شبیه‌سازی

در ادامه این فصل، مبانی شبیه‌سازی و مراحل اصلی بررسی خواهد شد. این بخش شامل مباحث زیر است:
تولید اعداد تصادفی– روش شبیه‌سازی مونت کارلو– کاربردهای عمومی و مالی شبیه‌سازی مونت کارلو– تکنیک‌هایی برای کاهش خطای استاندارد برآوردهای حاصل از شبیه‌سازی.

تولید یک عدد تصادفی

در این قسمت، به معرفی یکی از ساده‌ترین روش‌های تولید اعداد تصادفی و نحوه سنجش کیفیت آنها و کاربرد آنها در شبیه‌سازی پرداخته خواهد شد.

---

[1] Bayesian Analysis
[2] Markov Chain Monte Carlo(MCMC)
[3] Bootstrap Resampling
[4] Jacknife Resampling

تولید اعداد تصادفی به فرآیند تولید دنباله‌ای از اعداد گفته می‌شود که الگوی مشخصی ندارند و به صورت غیرقابل پیش‌بینی ظاهر می‌شوند؛ به عبارتی، کاملا تصادفی هستند. یعنی ساختن عددی که از قبل قابل پیش‌بینی نیست. فرض کنید یک تاس معمولی داریم و پیش از اینکه این تاس را پرتاب کنیم، نمی‌دانیم چه عددی ظاهر می‌شود. به این‌گونه اعداد، اعداد تصادفی گفته و به دو گروه اعداد تصادفی حقیقی و اعداد تصادفی شبه حقیقی تقسیم می‌شوند.

الف) اعداد تصادفی حقیقی[1]

اعداد تصادفی حقیقی از پدیده‌های فیزیکی غیرقابل پیش‌بینی تولید می‌شوند. برخلاف اعداد شبه‌تصادفی که توسط الگوریتم‌های ریاضی تولید می‌شوند و قابل بازتولید هستند، اعداد تصادفی حقیقی کاملا غیرقطعی و غیرقابل پیش‌بینی‌اند. از جمله روش‌های تولید اعداد تصادفی حقیقی به موارد زیر می‌توان اشاره نمود:

۱. اعداد تصادفی حقیقی تولید شده از پدیده‌های فیزیکی مانند نویز الکتریکی. برای مثال، نویز به‌صورت سیگنال آنالوگ ثبت شده و سپس به بیت‌های دیجیتال تبدیل می‌شود.
۲. اعداد تصادفی حقیقی تولید شده از پدیده‌های کوانتومی مانند عبور یا عدم عبور فوتون از یک آینه نیمه‌شفاف که می‌تواند یک عدد تصادفی حقیقی تولید نماید که بسیار امن و غیرقابل پیش‌بینی و مناسب برای رمزنگاری کوانتومی است.

ب) اعداد شبه‌تصادفی[2]

تولید اعداد شبه‌تصادفی[3] یکی از پایه‌های اصلی در علوم کامپیوتر، آمار، یادگیری ماشین، رمزنگاری و شبیه‌سازی است. با استفاده از الگوریتم‌های ریاضی و یک مقدار اولیه به نام بذر اولیه[4] دنباله‌ای از اعداد تولید می‌شود که ظاهرا تصادفی هستند، ولی در واقع قابل پیش‌بینی و قابل بازتولید هستند؛ به این معنی که اگر همان مقدار اولیه را استفاده کنیم، همان دنباله از اعداد تولید خواهد شد.

تولید عدد شبه‌تصادفیِ بین صفر و یک، یکی از رایج‌ترین و بنیادی‌ترین ابزارها در شبیه‌سازی آماری است. تولید عدد شبه‌تصادفی بین صفر و یک به معنی تولید عددی از توزیع یکنواخت پیوسته در بازه $(1و0)$ با استفاده از الگوریتم‌های تولید اعداد شبه تصادفی می‌باشد. اعداد شبه‌تصادفی بین ۰ و ۱ پایه‌ای برای تولید نمونه از توزیع‌های آماری مختلف هستند. در واقع، با استفاده از این اعداد می‌توانیم بسیاری از توزیع‌های آماری دلخواه را شبیه‌سازی کنیم[5]. برای مثال، می‌خواهیم با تولید اعداد شبه تصادفی بین ۰ و ۱ به اعدادی با توزیع برنولی دست یابیم، یعنی

---

[1] Real Random Numbers
[2] Pseudo-Random Number
[3] Pseudo-Random Number Generation – PRNG
[4] Seed
[5] چگونگی انجام این کار در کتب شبیه‌سازی آماری به تفصیل توضیح داده شده است.

دنباله‌ای از مقادیر ۰ و یا ۱ تولید شوند. به این منظور، دنباله‌ای از اعداد بین ۰ و ۱ تولید شده و هر عددی که از ۰.۵ کوچکتر باشد را معادل ۰ و اگر از ۰.۵ بزرگتر باشد معادل ۱ قرار می‌دهیم.

از جمله کاربردهای تولید اعداد شبه تصادفی بین ۰ و ۱ می‌توان به موارد زیر اشاره نمود:

۱. **مدل‌سازی پدیده‌های تصادفی**: مانند صف، بیماری، بازار و ترافیک
۲. **محاسبه انتگرال‌ها با استفاده از روش مونت کارلو**: این روش در قسمت‌های بعدی معرفی می‌شود.
۳. **انجام آزمون فرض:** با استفاده از نمونه‌گیری بوت استرپ[1]
٤. **تحلیل ریسک**. فرض کنید می‌خواهیم بازده یک سرمایه‌گذاری را در طول یک سال پیش‌بینی کنیم. مدل بازده روزانه را می‌توان به‌صورت تصادفی از توزیع نرمال ساخت و برای تولید اعداد از توزیع نرمال، تولید اعداد شبه تصادفی بین ۰ و ۱ مورد نیاز می‌باشد.

مثال ۱–٤: یکی از مثال‌های رایج از کاربرد تولید اعداد شبه‌تصادفی بین ۰ و ۱ تقریب π است. از ریاضیات مدرسه می‌دانیم این عدد برابر است با نسبت محیط دایره به قطر آن و مقدارش تقریبا برابر با ۳.۱۴۱۵۹ است. به منظور تقریب آن، روش زیر را بکار می‌بریم:

مربعی به ضلع ۲ و سپس داخل آن، یک دایره به قطر ۲ (شعاع ۱) رسم می‌کنیم. مساحت مربع برابر است با:

$$\text{مساحت مربع} = 2 \times 2 = 4$$

مساحت دایره نیز برابر است با:

$$\text{مساحت دایره} = \pi \times 1 = \pi$$

به منظور تخمین π، نقاط تصادفی درون مربع تولید می‌کنیم. هر نقطه دارای مختصات x و y است که هریک از این نقاط، از اعداد شبه تصادفی بین ۰ و ۱ تولید می‌شوند؛ چرا که هر بار که دو عدد تصادفی بین ۰ و ۱ تولید می‌کنیم، این دو عدد را معادل مختصات نقاط تصادفی در نظر می‌گیریم. از این‌رو، درون مربع، مختصاتِ متناظر با این دو عدد را یافته و آن را به عنوان نقطه تصادفی رسم می‌کنیم. تمام این نقاط به طور یکنواخت درون مربع پخش می‌شوند. سپس، بررسی می‌کنیم که هر نقطه داخل دایره است یا خیر. یعنی اگر شرط زیر برقرار باشد، نقطه داخل دایره واقع شده است:

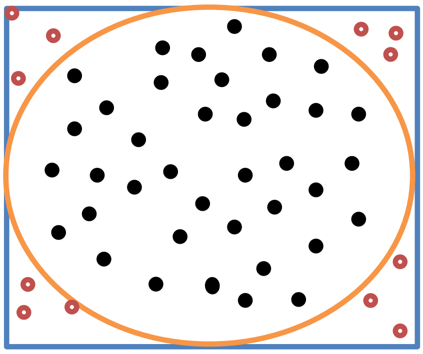
$$x^2 + y^2 \leq 1$$

نسبت تعداد نقاطی که داخل دایره هستند به کل نقاط تولید شده، نسبت مساحت دایره به مساحت مربع را نشان می‌دهد و بنابراین، می‌توان تقریبی برای π به‌دست آورد:

[1] Bootstrapping Method

$$\frac{\text{تعداد نقاط داخل دایره}}{\text{کل نقاط}} = \frac{\text{مساحت دایره}}{\text{مساحت مربع}} = \frac{\pi}{4} \Rightarrow \pi \approx 4 \times \frac{\text{تعداد نقاط داخل دایره}}{\text{کل نقاط}}$$

هرچقدر تعداد نقاط تولید شده بیشتر باشد، تقریب به‌دست آمده دقیق‌تر است.



ارتباط بین مثال عدد π و پیش‌بینی بازده سرمایه‌گذاری در این است که هر دو از یک روش استفاده می‌کنند: شبیه‌سازی. در تقریب عدد π ابتدا تعداد زیادی عدد تصادفی تولید می‌کنیم، سپس بررسی می‌کنیم هر نقطه داخل دایره یا خارج آن قرار می‌گیرد و در نهایت با نسبت دادن تعداد نقاط داخل دایره به کل نقاط، یک مقدار تقریبی از π به دست می‌آوریم. در پیش‌بینی بازده سرمایه‌گذاری نیز مراحل مشابهی انجام می‌شود: ابتدا بازده‌های تصادفی روزانه تولید می‌کنیم، سپس این بازده‌ها را برای یک سال جمع می‌کنیم، در گام بعد، صدها یا هزاران بار این فرآیند را تکرار می‌کنیم و در نهایت از نتایج به‌دست‌آمده، احتمال سود، احتمال ضرر و مقدار تقریبی بازده آینده را محاسبه می‌کنیم. بنابراین، همان‌طور که با کمک نقاط تصادفی می‌توان π را تقریب زد، با تولید مسیرهای تصادفی بازده نیز می‌توان آینده‌ی سرمایه‌گذاری و میزان ریسک آن را تخمین زد.

در قسمت بعد، یکی از شناخته‌شده‌ترین روش‌های شبیه‌سازی مرتبط با رویکرد اول (تولید داده از یک توزیع آماری) با نام مونت کارلو را معرفی می‌کنیم.

رویکرد اول: روش شبیه‌سازی مونت کارلو

یکی از رایج‌ترین رویکردهای شبیه‌سازی آماری، *تولید داده از یک توزیع آماری است*. شبیه‌سازی‌های آماری کاربردهای بسیار گسترده‌ای دارند که تکنیک مونت کارلو یکی از رایج‌ترین روش‌های شبیه‌سازی آماری است. اما این شبیه‌سازی در چه مواردی استفاده می‌شود؟ نکته مهمی که لازم است توجه کنیم این است که پس از تولید تعداد زیادی مجموعه داده، برای هر یک از این مجموعه داده‌ها حداقل یک برآورد یا آماره محاسبه می‌شود. برای مثال، ۵۰۰۰ مجموعه داده بر اساس یک روش شبیه‌سازی تولید می‌شود و میانگین متغیر تصادفی را در هر یک از

این مجموعه داده‌ها محاسبه می‌نماییم. به این ترتیب، تعداد ۵۰۰۰ میانگین در اختیار داریم. در این راستا، در متون شبیه‌سازی به دو موضوع پرداخته می‌شود:

- **روش تولید داده‌ها از یک توزیع آماری مشخص**: روش‌های آماری مختلفی برای تولید این داده‌ها وجود دارد که یادگیری آنها مستلزم دانش تئوری آماری قوی است که در حیطه این کتاب نمی‌گنجد.
- **محاسبه برآورد پارامتر مورد نظر پژوهشگر با توجه به داده‌های تولید شده از شبیه‌سازی:** یکی از کاربردهای رایج شبیه‌سازی آماری، به‌دست آوردن یک برآورد برای پارامتری مشخص در شرایطی است که محاسبه آن بر اساس فرمول‌های آماری، پیچیده و دشوار باشد. در روش مونت کارلو، از یک توزیع آماری مشخص، نمونه‌هایی با حجم معین به دفعات بسیار زیاد تولید، برای هر یک از نمونه‌ها آماره مورد نظر (مانند میانگین، انحراف معیار و...) محاسبه و در نهایت متوسط این آماره‌ها به عنوان برآورد پارامتر مورد بررسی استفاده می‌شود. توجه داشته باشیم روش مونت کارلو روشی برای محاسبه یک برآورد برای پارامتری مشخص با توجه به داده‌های تولید شده است. در حالیکه فرآیند تولید داده‌ها از توزیع آماری موردنظر به روش‌های مختلفی انجام می‌پذیرد. بنابراین، در اغلب متون آماری، از واژه "*برآوردگر مونت کارلو*" استفاده می‌شود که با عنوان "*شبیه‌سازی مونت کارلو*" شناخته شده است. از این‌رو، برخی از ویژگی‌های این برآوردگر مانند واریانس از اهمیت زیادی برخوردار است، چرا که کاهش آن باید مود توجه پژوهشگر قرار گیرد: کاهش واریانس برآوردگر، حاکی از بالابودن دقت آن است.

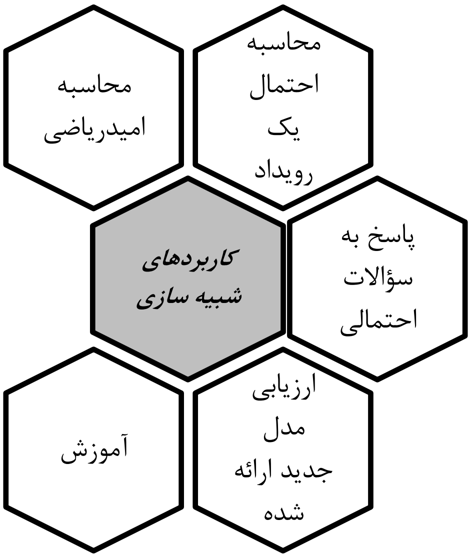
موضوع دیگری که لازم است به آن اشاره شود این است که در بررسی‌های رایج آماری در مورد یک موضوع مشخص، ابتدا لازم است داده‌های مورد نیاز جمع‌آوری شود و سپس بر اساس یک آماره آزمون با یک توزیع معین می‌توان در مورد آن موضوع با استفاده از آزمون فرض تصمیم‌گیری نمود[۱]. در بسیاری از این بررسی‌ها، برای اطمینان از شکل توزیع احتمال این داده‌ها از آزمون‌های آماری استفاده می‌شود. برای مثال، آزمون کولموگروف اسمیرنوف برای بررسی اینکه آیا مجموعه داده مورد بررسی دارای توزیع نرمال است یا خیر مورد استفاده قرار می‌گیرد. مثلا، در زمانِ جمع‌آوری داده برای برازش مدل رگرسیون خطی ساده، مطمئن نیستیم که توزیع داده‌ها نرمال است، به همین دلیل است که پس از برازش مدل، این فرض را ارزیابی می‌نماییم. اما، در شبیه‌سازی آماری از ابتدا، یک توزیع آماری برای پدیده مورد نظر، در نظر گرفته و مجموعه داده‌هایی از آن توزیع تولید می‌کنیم. به این ترتیب، مطمئن هستیم این داده‌ها از توزیع احتمال مذکور برخوردار هستند.

در قسمت بعد به کاربردهای مختلف روش شبیه‌سازی مونت کارلو پرداخته می‌شود.

[۱] جلد اول همین کتاب.

کاربردهای شبیه‌سازی مونت کارلو

شبیه‌سازی‌های آماری مانند روش مونت کارلو با اهداف مختلفی انجام می‌شود. به عبارتی، روش شبیه‌سازی مونت کارلو به سؤالات مختلفی پاسخ می‌دهد و همین موضوع باعث جذابیت آن شده است. در جلد ۱ این کتاب ملاحظه شد محاسبه امید ریاضی یک متغیر تصادفی کمّی پیوسته، مستلزم استفاده از قواعد انتگرال‌گیری است. گاهی به‌دست آوردن امید ریاضی این متغیر تصادفی مستلزم استفاده از محاسبات ریاضی دشوار و پیچیده‌ای است، ولی با استفاده از روش مونت کارلو به سهولت انجام می‌پذیرد. همچنین، به‌جز متغیرهای تصادفی که دارای توزیع‌های مشخصی همچون نرمال، تی –استودنت و ... که دارای جداول آماری آماده هستند و برای محاسبه احتمالات مختلف مربوط به متغیر تصادفی مورد استفاده قرار می‌گیرند، محاسبه چنین احتمالاتی نیز با استفاده از انتگرال‌گیری انجام می‌شود. در اینجا نیز، محاسبه احتمالات مختلف در برخی مسائل با دشواری‌هایی مواجه می‌شود که به کمک روش مونت کارلو به‌راحتی قابل محاسبه است. به دلیل اهمیت کاربردهای شبیه‌سازی مونت کارلو، در قسمت‌های بعدی به ۵ کاربرد رایج این روش پرداخته می‌شود که در گراف زیر نیز نمایش داده شده است.



توجه داشته باشید، در هر یک از کاربردهای روش مونت کارلو، فرض می‌شود روش تولید داده معلوم بوده و قادر هستیم با استفاده از این روش، داده‌های مورد نظر خورد را تولید نماییم و از روش مونت کارلو تنها برای به‌دست آوردن برآوردهای مورد نظر بهره می‌بریم.

کاربردهای شبیه‌سازی – حالت اول: محاسبه امید ریاضی

همان‌گونه که می‌دانید میانگین یک متغیر تصادفی که همان امیدریاضی آن متغیر است، در مورد متغیر تصادفی گسسته بر اساس جمع و برای متغیر تصادفی پیوسته کمّی، بر مبنای انتگرال به‌دست می‌آید:

متغیر تصادفی گسسته $E(X) = \sum_x xP(x)$

متغیر تصادفی کمی پیوسته $E(X) = \int xP(x)dx$

در برخی موارد، به‌دست آوردن این میانگین برای متغیر تصادفی کمّی پیوسته به دلیل اینکه انتگرال مذکور قابل حل نیست، امکان‌پذیر نمی‌باشد. شبیه‌سازی مونت کارلو برای رفع این مشکل به این صورت عمل می‌نماید که ابتدا از توزیع احتمال $P(x)$ مجموعه‌ای از داده‌ها با حجم بسیار بزرگ (n) تولید نموده و سپس برآوردِ میانگین متغیر تصادفی را به شکل زیر به‌دست می‌آورد:

$$\hat{E}(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

این حالت، یکی از رایج‌ترین کاربردهای شبیه‌سازی مونت کارلو است؛ چرا که بسیاری از مسائل مطرح شده می‌تواند به صورت "امیدریاضی یک متغیر تصادفی" بیان شود. به منظور درک بیشتر این موضوع، انتگرال زیر را در نظر بگیرید که در آن دو تابع $g(x)$ و $f(x)$ در هم ضرب شده‌اند:

$$\int g(x)f(x)dx$$

برای حل این انتگرال، فرض می‌کنیم X یک متغیر تصادفی با تابع چگالی احتمال $f(x)$ است و بنابراین، می‌توان این انتگرال را به صورت امید ریاضی $g(x)$ در نظر گرفت که:

$$E\big(g(X)\big) = \int g(x)f(x)dx$$

به عبارتی، روش مونت کارلو یک انتگرال را به جمع، تبدیل و محاسبه امیدریاضی را آسان می‌نماید.

به طور کلی، برای حل این انتگرال، سه روش وجود دارد:

الف) روش‌های ریاضی انتگرال‌گیری: قابل کاربرد برای توابع ریاضی انتگرال‌پذیر.

ب) روش‌های عددی: در صورتی که نتوان از روش‌های بند الف استفاده نمود و تعداد متغیرها کم باشد. اما وقتی تعداد متغیرهای تصادفی زیاد باشد، روش‌های عددی کارا نیستند. لازم به ذکر است در تحلیل‌های آماری وقتی نمونه تصادفی به حجم n استخراج می‌شود، عملا تعداد متغیرهای تصادفی نیز n است که تعداد زیادی محسوب می‌شود:

$$X_1, X_2, \dots, X_n$$

ج) روش شبیه‌سازی مونت کارلو: وقتی تابع ریاضی مود نظر پیچیده باشد و یا تعداد متغیرها زیاد باشد، این روش به‌کار می‌رود که با متوسط‌گیری از تعداد زیادی از مقادیر متغیر تصادفی انجام می‌شود. به این ترتیب که از تابع چگالی احتمال $f(x)$ تعداد بسیار زیاد (n) داده تولید می‌کنیم و سپس، امید ریاضی فوق را به شکل زیر به‌دست می‌آوریم:

$$\hat{E}(g(X)) = \frac{1}{n}\sum_{i=1}^{n} g(x_i)$$

گفتنی است پایه و اساس روش مونت کارلو همین متوسط‌گیری است و هر مسأله‌ای را که بتوان به شکل انتگرال مذکور تعریف نمود، به‌وسیله روش مونت کارلو قابل حل است.

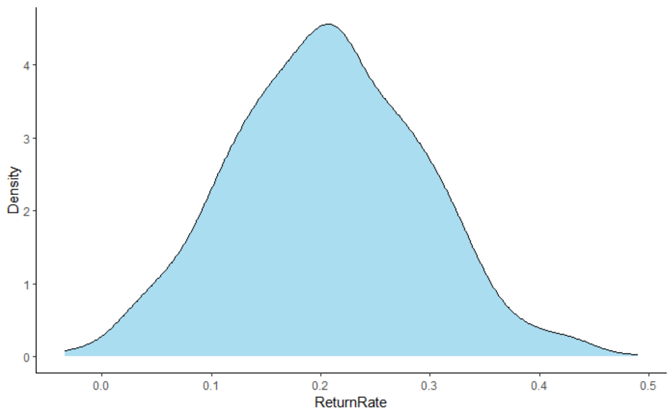
کاربردهای شبیه‌سازی – حالت دوم: پاسخ به سؤالات احتمالی

در این حالت، پدیده‌ای مورد بررسی قرار می‌گیرد که می‌توان متناظر با آن پدیده، متغیر تصادفی با یک توزیع آماری مشخص تعریف نمود. از این‌رو، اگر داده‌هایی با حجم بالا از توزیع مورد نظر داشته باشیم، می‌توانیم به سؤالات احتمالی درباره پدیده مذکور پاسخ دهیم.

مثال ۱-۵: فرض کنید نرخ بازده روزانه سهام یک شرکت تولیدی برای کارشناسان بانکی به منظور اعطاء وام از یک سو و کارشناسان ریسک سایر شرکت‌ها در راستای مدیریت ریسک پورتفوهای خود از سوی دیگر، اهمیت قابل توجهی دارد. از طرفی، این کارشناسان به داده‌های نرخ بازده شرکت مذکور دسترسی ندارند، ولی می‌دانند این نرخ بازده دارای توزیع نرمال به شکل زیر است:

$$RR \sim N(0.21\ , 0.09)$$

علیرغم عدم دسترسی به داده‌های نرخ بازده این شرکت، می‌توان با استفاده از رویکرد شبیه‌سازی، تصمیم‌های بهتری در مورد مدیریت سهام این شرکت و یا وضعیت تسهیلات اعطایی به این شرکت اخذ نمود. با استفاده از تکنیک‌های شبیه‌سازی، مجموعه داده‌ای از توزیع نرمال فوق تولید می‌کنیم و تصویر اولیه‌ای از آن را به‌دست می‌آوریم (شکل زیر).



سپس، با استفاده از روش استانداردسازی که در جلد نخست این کتاب آموختیم، به سؤالات احتمالی پاسخ می‌دهیم. برای مثال، ممکن است برای کارشناس ریسک یکی از شرکت‌های طرف مقابل این سؤال مطرح شود که چقدر احتمال دارد نرخ بازده روزانه این شرکت، صفر و یا منفی شود. به این سؤال به صورت زیر پاسخ داده می‌شود:

$$P(RR \leq 0) = P\left(Z \leq \frac{0 - 0.21}{0.09}\right) = P(Z \leq -2.33) = 0.0099$$

همان‌گونه که مشاهده می‌شود احتمال فوق بسیار کوچک است؛ گویی در ۱۰۰۰ روزِ پیشِ رو، انتظار داریم ۹ روز با نرخ بازده صفر و یا منفی مواجه شویم. با استفاده از داده‌های تولید شده نیز می‌توان برآوردهایی از صدک‌ها و شاخص‌های آماری به‌دست آورد.

مثال ۱-٦: در مثال فوق، در حالت سخت‌گیرانه‌تر فرض کنید تنها اطلاعی که داریم این است که می‌دانیم نرخ بازده روزانه این شرکت، یک توزیع نرمال با انحراف معیار ۰.۰۹ است (به عبارتی، میانگین این توزیع نرمال در دسترس نیست). چگونه می‌توانیم به سؤالات احتمالی که در اخذ تصمیماتشان نقش مهمی دارند پاسخ دهیم؟

برای پاسخ به این سؤال می‌توان گفت در این حالت، چندین سناریو طراحی می‌کنیم که هر یک از این سناریوها شامل در نظر گرفتن یک توزیع نرمال با مقداری برای میانگین و نیز انحراف معیار ۰.۰۹ است. در واقع، تفاوت سناریوها، از نظر مقدار میانگین توزیع نرمال است؛ یعنی، در سناریوهای مختلف، مقادیر مختلفی برای میانگین توزیع نرمال در نظر گرفته شده است. سپس، برای هر یک از این سناریوها شبیه‌سازی انجام داده و بر اساس هر یک از داده‌های شبیه‌سازی شده به سؤالات احتمالی پاسخ می‌دهیم. بنابراین، برای سناریوهای مختلف اطلاعاتی به‌دست می‌آوریم که چشم‌انداز بهتری برای کارشناسان ارائه نموده و به آنها در تحلیل‌ها و تصمیم‌گیری‌هایشان کمک می‌نماید. توجه داشته باشید سناریوهای مختلف باید طیف متنوعی از مقادیر میانگین این توزیع نرمال را در بر داشته باشد. برای مثال، تحت ۴ سناریو، مقدار پایین، صفر، متوسط و بسیار بالا برای میانگینِ این توزیع نرمال در نظر گرفته می‌شود تا بتوان در شرایط مختلف، درباره سهام این شرکت اطلاعات مفید به‌دست آورد.

کاربردهای شبیه‌سازی - حالت سوم: محاسبه احتمال یک رویداد

اگر هدف از بررسی، برآورد احتمالِ رخداد خاصی باشد، نمونه‌های زیادی از توزیع احتمال آن رخداد، تولید می‌نماییم. در اینجا، ابتدا رخداد مورد نظر را به صورت یک متغیر تصادفی تعریف می‌کنیم که توزیع احتمال آن را نیز به‌دست می‌آوریم. سپس، نسبت نمونه‌هایی که در آنها این رخداد اتفاق افتاده را به عنوان برآوردی برای این احتمال در نظر می‌گیریم. توجه داشته باشید تفاوت این حالت با حالت قبلی در این است که در حالت قبلی، سؤالات احتمالی به صورت بازه‌هایی مطرح می‌شوند که طیف متنوعی از این بازه‌ها می‌تواند در نظر گرفته شود. در صورتی که، در این

حالت، پژوهشگر در صددِ محاسبه احتمال *یک رویداد مشخص* است که از طریق محاسبه میانگین نمونه‌های تولیدشده به‌دست می‌آید.

مثال ۱-۷: احتمال این را به‌دست آورید که در یک‌بار پرتاب دو تاس، جمع دو وجه ظاهر شده ۷ شود؟ به این منظور، ابتدا آزمایش تصادفی به صورت پرتاب دو تاس و سپس، متغیر تصادفی به صورت "جمع اعداد ظاهر شده وجوه دو تاس" تعریف می‌شوند. در مرحله بعد، این آزمایش تصادفی ۱۰۰۰۰۰ بار انجام می‌شود. تعداد دفعاتی که جمع دو عدد ظاهر شده ۷ است را می‌شماریم. این تعداد را بر ۱۰۰۰۰۰ بار تقسیم نموده و این نسبت را به‌عنوان برآوردِ احتمال مورد نظر می‌پذیریم[۱].

مثال ۱-۸: یکی از مهمترین کاربردهای شبیه‌سازی مونت کارلو، محاسبه احتمالِ یک رویداد مشخص بر اساس مفهوم امید ریاضی است. در این حالت، یک متغیر تصادفی به شکل زیر برای این رویداد تعریف می‌کنیم:

$$X = \begin{cases} 1 & \text{رویداد رخ دهد} \\ 0 & \text{رویداد رخ ندهد} \end{cases}$$

در این صورت، می‌توان گفت این متغیر تصادفی دارای توزیع برنولی است:

$$X \sim Bernoulli(p) \qquad ; p = P(X = 1)$$

همان‌گونه که می‌دانیم:

$$E(X) = p$$

بنابراین، برای محاسبه احتمالِ رخ دادنِ رویداد مذکور، کافی است امیدریاضی این متغیر تصادفی را به روش مونت کارلو محاسبه نماییم:

$$P(X = 1) = p = E(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

کاربردهای شبیه‌سازی - حالت چهارم: ارزیابی یک مدل آماری جدید پیشنهاد شده

در این حالت، یک مدل جدیدِ آماری با مجموعه‌ای از فرض‌های مشخص *برای اولین بار* ارائه می‌شود که اثبات مناسب بودنِ ویژگی‌های برآوردهای آن از نظر تئوری به دلیل پیچیدگی‌های ریاضیِ مدل، دشوار بوده و یا امکان‌پذیر

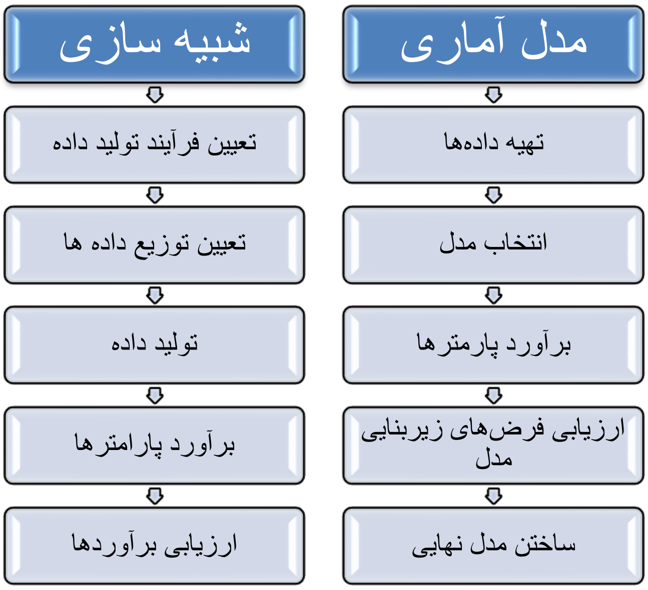
---

[۱] لازم به ذکر است انجام این آزمایش به تعداد دفعاتِ دلخواه و شمارشِ مذکور، به سهولت توسط محیط‌های برنامه‌نویسی مانند R و یا پایتون و حتی نرم‌افزار Excel قابل انجام است.

نیست[۱]. به همین دلیل، بر اساس فرض‌های مذکور، چندین مجموعه داده با سناریوهای مختلف و با حجم بالا تولید می‌شوند. تولید داده تحت سناریوهای مختلف باعث می‌شود که بتوان از ابعاد مختلف، توانمندیِ مدل را ارزیابی نمود.

در مرحله بعد، ویژگی‌های برآوردهای مدل مانند نااریب بودن، دقت، سازگاری و غیره به تنهایی و یا با هدف مقایسه با سایر مدل‌های موجود با استفاده از شاخص‌های آماری مورد ارزیابی قرار می‌گیرند که یکی از متداول‌ترین روش‌های به‌دست آوردن این شاخص‌ها، روش مونت کارلو است.

گفتنی است می‌توان فرآیند برازش یک مدل آماری شناخته شده مانند رگرسیون را با فرآیند شبیه‌سازی مقایسه نمود. با توجه به شکل زیر مشاهده می‌شود فرآیند شبیه‌سازی، تقریبا معکوس فرآیند برآورد یک مدل آماری موجود مانند مدل رگرسیون خطی است. این مدل را به‌خاطر آورید. به منظور شناسایی ارتباط بین دو متغیر وابسته Y و مستقل X با استفاده از مدل مذکور، یک فرآیند چند مرحله‌ای انجام می‌شود که در شکل زیر نشان داده شده است (قسمت مدل آماری):



[۱] زمانی که یک مدل جدید آماری توسط یک پژوهشگر ارائه می‌شود، این مدل شامل پارامترهایی است که باید برآورد شوند. یکی از ارزیابی‌های مهم این روش جدید، بررسی دو ویژگی مهم این برآوردگر شامل نااریبی و واریانس (دقت) است.

به عبارتی، برازش یک مدل رگرسیون خطی ساده از جمع‌آوری داده شروع شده و به ساختن مدل نهایی منتهی می‌شود. در این فرآیند، شکل کلی مدل از قبل، توسط آماردانان ارائه شده، ویژگی‌های برآوردهای پارامترهای آن بررسی و اثبات شده و قرار است برای یک مجموعه داده مشخص مورد استفاده قرار گیرد تا میزان ارتباط بین دو متغیر وابسته و مستقل به‌دست آید. اما، همان‌گونه که ملاحظه شد شبیه‌سازی، تقریبا معکوس این فرآیند است. یعنی، از ابتدا باید توزیع آماری داده‌ها انتخاب و بر اساس فرآیندی داده‌های مذکور تولید شوند. بنابراین، در انتهای کار نیازی به ارزیابی فرض توزیعی داده‌ها نیست.
کاربرد شبیه‌سازی در ارزیابی مدل آماری جدید معرفی شده، در بسیاری از تحقیقات آماری در رشته‌های مختلف مورد استفاده قرار می‌گیرد.

کاربردهای شبیه‌سازی – –حالت پنجم: آموزش

در این حالت می‌خواهیم بیاموزیم وقتی فرض‌های یک مدل آماری برقرار نباشد چه اتفاقی می‌افتد؟ برای مثال، چرا در یک مدل رگرسیون خطی باید جملات خطا دارای توزیع نرمال باشند؟ اگر توزیع نرمال نداشته باشند چه ایرادی دارد؟ به این منظور، برای یک مدل آماریِ موجود، مجموعه داده‌ای با برخورداری از فرض‌های اصلی مدل مذکور تولید می‌شود. از طرفی، مجموعه داده دیگری با ویژگی‌هایی تولید می‌شود که از فرضیات اصلی این مدل انحراف دارند. سپس، دو مدل مذکور باید مقایسه شوند که این مقایسه و ارزیابی، بر اساس شاخص‌هایی مانند خطای نوع اول انجام می‌شود.
به منظور شرح این روش، مثال زیر را در نظر بگیرید.

مثال ۱–۹: همان‌گونه که می‌دانیم انجام آزمون فرض در یک مدل رگرسیون خطی ساده، مستلزم نرمال بودن توزیع جملات خطای مدل است. به منظور بررسی چراییِ این موضوع، دو مدل رگرسیونی زیر را در نظر بگیرید که در مدل اول، خطاهای مدل از توزیع نرمال پیروی کرده و توزیع خطاهای مدل دومی یک توزیع نمایی است.

$$Y_{1i} = 0.5 + 0.02X_i + \varepsilon_{1i} \quad ; \quad \varepsilon_{1i} \sim N(0,\sigma^2)\ , X \sim Uniform(5,23)$$

$$Y_{2i} = 0.5 + 0.02X_i + \varepsilon_{2i} \quad ; \quad \varepsilon_{2i} \sim Exp(1)\ \ , X \sim Uniform(5,23)$$

توجه داشته باشید که مقدار متغیر مستقل در هر دو مدل یکسان است و در هر دو حالت از یک توزیع یکنواخت تولید شده است. انتخاب توزیع یکنواخت کاملا دلخواه انجام شده است. استفاده از شبیه‌سازی برای انجام این بررسی و مقایسه دو مدل به این شکل است که دو مجموعه داده به شرح زیر تولید می‌شود:
الف) داده‌هایی به حجم ۲۰ که در آن توزیع جملات خطا نرمال است. برای مثال، این داده‌ها ۱۰۰۰۰ بار تولید می‌شوند. به عبارتی، ۱۰۰۰۰ مجموعه داده داریم که حجم هر یک از آنها ۲۰ است. توجه داشته باشید مقادیر متغیرهای تصادفی در این مجموعه‌ها با یکدیگر یکسان نیستند، ولی توزیع آماری آنها یکسان است. گویی به یک

دستگاه تولید کننده، توزیع آماری مشخصی مانند نرمال را وارد کرده و این دستگاه هر بار بسته‌های ۲۰ تایی از مقادیر مختلف تولید می‌نماید که همه این دسته‌ها تویع نرمال یکسانی دارند.

ب) داده‌هایی به حجم ۲۰ که در آن توزیع جملات خطا نمایی است. مانند حالت قبل، چنین داده‌هایی ۱۰۰۰۰ بار تولید می‌شوند.

ساده‌ترین روش تحلیل این داده‌های شبیه‌سازی شده به این صورت است که ابتدا یک مدل رگرسیونی روی هر یک از مجموعه داده‌ها در هر دو حالت اجرا کرده و برآورد پارامترها در هر دو حالت به‌دست می‌آیند. بنابراین، برای هر یک از این دو مجموعه داده، ۱۰۰۰۰ برآوردِ شیب، ۱۰۰۰۰ برآوردِ عرض از مبدأ، ۱۰۰۰۰ آماره آزمون و ۱۰۰۰۰ P-value متناظر با آن داریم.

اکنون، می‌خواهیم دو مدل را بر اساس خطای نوع اول[1] با یکدیگر مقایسه کنیم که به این منظور لازم است در هر یک از مدل‌های رگرسیونیِ برآورد شده، آزمون فرض انجام دهیم و انجام آزمون فرض در گام اول مستلزم محاسبه آماره آزمون است. همان‌گونه که از جلد نخست این کتاب آموختیم، در ساختن آماره آزمون و به دنبال آن P-value ، نیاز به برآوردِ شیب خط رگرسیونی و نیز انحراف معیار این برآورد داریم. غالبا فرض صفر در یک مدل رگرسیون خطی این است که آیا متغیر مستقل روی متغیر وابسته تأثیر معنی‌دار ندارد و به این ترتیب آماره آزمون برای هر یک از مجموعه داده‌ها به صورت زیر محاسبه می‌شود.

$$t_{ij}=\frac{b_{ij}}{S(ij)} \quad ; \quad i=1,2,\ldots,10000 \quad ; j=1,2$$

اندیس j مربوط به دو مدل رگرسیون برازش داده شده است و اندیس i مربوط به تعداد تکرار شبیه‌سازی می‌باشد. واضح است که هرچقدر مخرج این کسر که انحراف معیارِ برآوردگرِ شیب رگرسیونی است کوچکتر باشد، مقدارِ آماره آزمون بزرگتر بوده و بنابراین، مقدار P-value کوچکتر می‌شود. در چنین حالتی فرض صفر، بسیار قدرتمندتر رد می‌شود.

در مرحله بعد، میانگین و انحراف معیارِ ۱۰۰۰۰ برآورد شیب به‌دست آمده در هر یک از دو مدل را محاسبه می‌نماییم. ما این کار را در R انجام داده و نتایج به صورت زیر می‌باشند:

	مدل ۱	مدل ۲
میانگین برآوردگر شیب خط رگرسیونی	۱.۰۹۹	۱.۰۹۹
انحراف معیار برآوردگر شیب خط رگرسیونی	۰.۲۲	۰.۰۴

[1] یکی از مهمترین شاخص‌های آماری برای مقایسه دو مدل، خطای نوع اول است که در متون یادگیری ماشین معادل مثبت‌های کاذب می‌باشد.

با توجه به این نتایج، به نظر می‌رسد مدل ۲ با قدرت بیشتری فرض صفر را رد می‌نماید. ولی در چنین شرایطی که خطای مدل دوم دارای توزیع نرمال نیست، به منظور بررسی بیشتر، باید خطای نوع اول هم مطالعه شود تا مطمئن باشیم که این رد کردنِ فرض صفر تا چه حد می‌تواند به اشتباه انجام شده باشد.

در ادامه، خطای نوع اول هر مدل را محاسبه می‌نماییم که نسبت مواردی است که این آزمون فرض در ۱۰۰۰۰ مجموعه داده در هر یک از دو مدل رد شده است. نتایج در جدول زیر ارائه شده است.

	مدل ۱	**مدل ۲**
خطای نوع اول	۰.۰۵	۰.۰۸

با توجه نتایج فوق می‌توان دریافت خطای نوع اول در مدلی که جملات خطا توزیع نمایی دارند بیشتر از مدل با خطاهای نرمال است. به عبارتی، وقتی فرض صفر صحیح است، مدل دوم (که توزیع خطاها در آن نمایی است)، با احتمال بیشتری آن را به اشتباه رد می‌کند. به این ترتیب می‌توان با استفاده از مطالعات شبیه‌سازی، مدل‌های آماری موجود را تحت سناریوهای مختلف بررسی و مقایسه نمود.

در موضوعات مختلف حوزه مالی شامل مدیریت ریسک پورتفو، شبیه‌سازی آماری این امکان را فراهم می‌نماید که نوسانات داده‌های مالی نیز در نظر گرفته شده و نیز سناریوهای مختلفی طراحی شود؛ به این ترتیب که با فرض یک توزیع آماری، گستره‌ای از مقادیر متغیر مورد بررسی و احتمالاتِ بازه‌های مختلفی از این مقادیر در نظر گرفته می‌شود.

توجه: کاربرد شبیه‌سازی مونت کارلو در آمار به‌ویژه آمار بیزی بسیار وسیع است که در حیطه این کتاب نمی‌گنجد.

از نقطه نظر عملیاتی، یک شبیه‌سازی آماری را طبق مراحل زیر انجام می‌دهیم:

گام اول: با فرض اینکه متغیر مورد نظر یک توزیع آماری مشخص دارد، یک مجموعه داده با حجم مشخص از این توزیع آماری تولید می‌کنیم[1].

گام دوم: برآوردِ شاخص و یا آماره مورد نظر (مانند میانگین یا انحراف معیار) را برای این مجموعه داده محاسبه می‌کنیم.

گام سوم: گام اول و دوم به تعداد دفعات بسیار زیاد، برای مثال ۱۰۰۰۰ بار تکرار می‌شود و به این ترتیب ۱۰۰۰۰ شاخص یا آماره به‌دست می‌آوریم.

گام چهارم: خلاصه‌ای از ۱۰۰۰۰ شاخص یا آماره به‌دست آمده در گام سوم محاسبه می‌شود. تکنیک مونت کارلو به طور ویژه یکی از روش‌های شبیه‌سازی است که متوسط این ۱۰۰۰۰ شاخص را به عنوان برآوردی برای آماره مذکور در نظر می‌گیرد.

[1] چگونگیِ تولید داده از یک توزیع آماری مشخص به روش‌های مختلفی انجام می‌شود که خارج از حیطه این کتاب است و یادگیری این روش‌ها مستلزم کافی بودن دانش تئوری آماری است.