دانشکده‌ی علوم ریاضی

احتمال و کاربردآن ۲۸ اردیبهشت ۹۳

جلسه‌ی ۲۲: قضیه‌ی حد مرکزی و توزیع نرمال دو متغیره

مدرّس: دکتر شهرام خزائی *نگارنده: سیّده‌بهناز حسینی‌سنو*

۱ مقدمه

در جلسات قبل، راجع به متغیرهای تصادفی مستقل با توزیع یکسان[۱] صحبت کردیم. گفتیم چنانچه تعداد n متغیر i.i.d را جمع و بر تعدادشان تقسیم کنیم، حد این عبارت به مقدار متوسط متغیرها میل می‌کند. از این اتفاق دو تفسیر ارائه کردیم:

تفسیر اول: همگرایی در احتمال

$$\forall \varepsilon > 0 : \lim_{n \to +\infty} \Pr\{|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mathrm{E}[X]| > \varepsilon\} = 0$$

تفسیر دوم: همگرایی almost sure

$$\lim_{n \to +\infty} \frac{X_1 + X_2 + \cdots + X_n}{n} = \mathrm{E}[X]$$

(توجه شود که این نوع همگرایی، قوی‌تر از همگرایی در احتمال است؛ اما در این درس نیاز به آن نمی‌پردازیم.)

نوع دیگری از همگرایی وجود دارد که همگرایی در توزیع[۲] نامیده می‌شود و مبنای قضیه‌ی بسیار مهم حدّ مرکزی[۳] است که در این جلسه به آن خواهیم پرداخت.

۲ همگرایی در توزیع و قضیه‌ی حدّ مرکزی

قضیه ۱ (قضیه حدّ مرکزی) *اگر $X_1, X_2, \ldots$ دنباله‌ای از متغیرهای تصادفی مستقل با توزیع یکسان با مقدار متوسط μ و واریانس σ^2 باشند، توزیع*

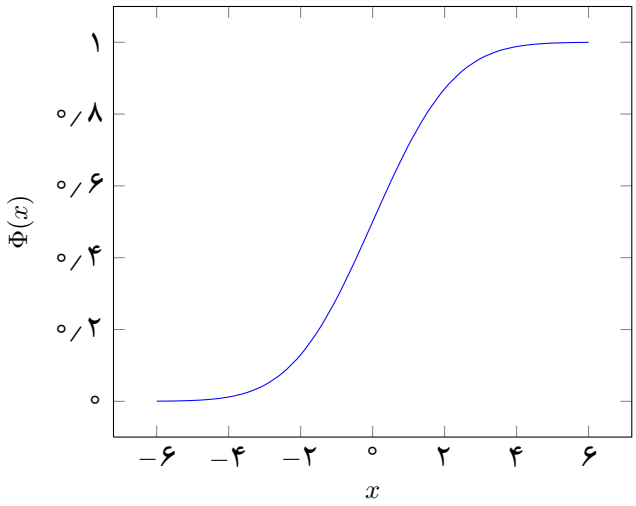
$$Z_n = \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

[۱] Independent and identically distributed random variables (i.i.d)
[۲] Convergence in distribution
[۳] Central limit theorem

به توزیع متغیر تصادفی نرمال استاندارد میل می‌کند. به بیان دیگر:

$$\lim_{n\to\infty} \Pr(Z_n \le t) = \lim_{n\to\infty} \Pr(\frac{X_۱ + X_۲ + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \le t)$$
$$= \frac{۱}{\sqrt{۲\pi}} \int_{-\infty}^{t} e^{-x^۲/۲}\, dx = \Phi(t).$$

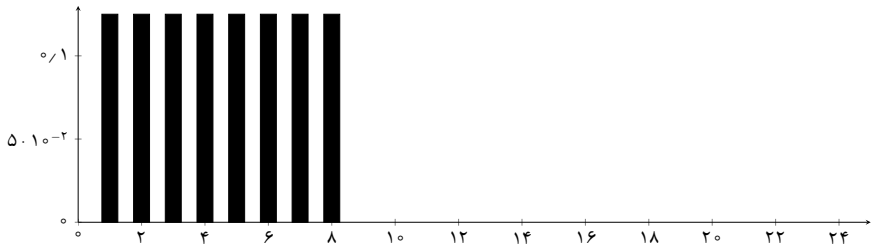
همان‌طور که در جلسات قبل گفته شد، تابع $\Phi(t)$، تابع توزیع تجمعی متغیر تصادفی نرمال (گوسی) استاندارد است که به شکل زیر است:



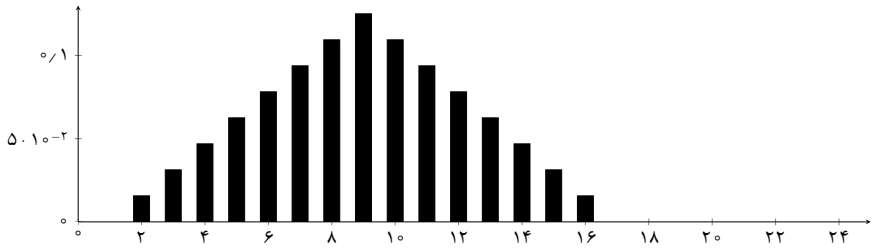
نکته ۱ *به طور کلی، توزیع‌های نامتقارن (مانند توزیع پوآسون یا توزیع نمایی)، به شرط موجود و محدود بودن مقدار متوسط و واریانس، زمان بیشتری برای همگرایی می‌برند.* ♦

قضیه‌ی حدّ مرکزی نشان می‌دهد که بسیاری از پدیده‌های طبیعی، با تقریب خوبی یک توزیع نرمال استاندارد هستند. تغییرات حاصل از فرایندهای طبیعی، اغلب نتیجه‌ی جمع شدن تعداد زیادی عامل کوچک و ناچیز است. اگرچه تأثیر هریک از این عوامل به‌تنهایی ممکن است نادیده گرفته شود، اما تأثیری که مجموع آن‌ها بر جای می‌گذارد قابل چشم‌پوشی نیست. از این رو، مطالعه‌ی توزیع مجموع تعداد زیادی از متغیرهای تصادفی مستقل ضروری است. به طور مثال، نویز موجود در مدارهای مخابراتی حاصل از چندین اثر است و جمع شدن تعداد زیادی نویز مختلف، در نهایت توزیع گاوسی به‌دست می‌دهد.

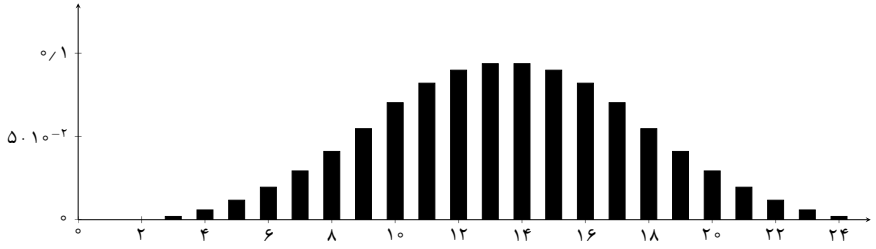
مثال ۱ *فرض کنید X یک متغیر تصادفی گسسته است که مقادیر ۱ تا ۸ را با احتمال $\frac{۱}{۸}$ می‌گیرد. تابع جرم احتمال مجموع دو نمونه از متغیر تصادفی X با محاسبه‌ی کانولوشن X با خودش به‌دست می‌آید. اگر این روند را برای دو، سه، چهار و یا تعداد بیشتری نمونه از X ادامه دهیم، با افزایش تعداد نمونه‌ها مشاهده می‌شود که فرم تابع جرم احتمال مجموع متغیرها به فرم تابع چگالی احتمال متغیر تصادفی نرمال نزدیک می‌شود. اما این به معنای میل کردن تابع جرم احتمال مجموع متغیرها به سمت تابع چگالی احتمال متغیر تصادفی نرمال نیست. در واقع، قضیه حدّ مرکزی به صورتی که بیان شد، صرفاً همگرایی تابع توزیع تجمعی مجموع متغیرهای تصادفی (که با کم کردن میانگین و تقسیم بر انحراف معیار نرمال نیز شده‌اند) را به سمت تابع توزیع تجمعی احتمال متغیر تصادفی نرمال استاندارد، تضمین می‌کند.*



تابع توزیع احتمال متغیر X



تابع توزیع تجمعی مجموع دو نمونه از X



تابع توزیع تجمعی مجموع سه نمونه از X

نکته ۲ *قضیه حدّ مرکزی درباره همگرایی تابع توزیع تجمعی دنباله متغیرهای تصادفی* $\frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$ *به سمت تابع توزیع تجمعی متغیر تصادفی نرمال استاندارد، مطرح می‌شود. با این وجود، در عمل برای* n*های نسبتاً بزرگ (مثلاً* $n > ۱۰$*) تابع توزیع تجمعی متغیر تصادفی* $X_1 + X_2 + \cdots + X_n$ *با تابع توزیع تجمعی متغیر تصادفی نرمال* $N(n\mu, n\sigma^2)$ *تقریب زده می‌شود.*

مثال زیر کاربردی از این تقریب را نشان می‌دهد.

مثال ۲ *کسر* f *از افراد جامعه نوشابه را به دلستر ترجیح می‌دهند. مقدار* f *را تخمین بزنید.*
مطلوب هر مسئله‌ی تخمین، نزدیک بودن جواب به‌دست آمده به جواب واقعی است. از n *نفر به طور تصادفی نمونه‌گیری می‌کنیم. می‌خواهیم با احتمال بیش از ۹۵ درصد، خطای محاسبه‌ی* f *کمتر از ۰٫۰۱ شود. به بیان دیگر می‌خواهیم:*

$$\Pr(|M_n - f| > ۰/۰۱) \leq ۰/۰۵ \qquad (۱)$$

که

$$M_n = \frac{X_1 + X_2 + \cdots + X_n}{n} \ ,$$

و متغیر تصادفی X_i *را به این صورت تعریف می‌کنیم:*

$$X_i = \begin{cases} ۱ & \text{اگر شخص } i\text{ام بگوید بله} \\ ۰ & oth. \end{cases}$$

که X_i دارای توزیع برنولی با پارامتر f بوده و متوسط و انحراف معیار آن به صورت زیر است:

$$\mathrm{E}[X_i] = f\ ,\ \mathrm{Var} = \sqrt{f(۱-f)} \le \frac{۱}{۴} \qquad (۲)$$

در جلسه‌ی قبل، بدون آگاهی از توزیع M_n و تنها با استفاده از قضیه‌ی چبیشف، تعداد نمونه‌ها را ۵۰۰۰۰ تخمین زدیم. با استفاده از قضیه‌ی حد مرکزی نشان می‌دهیم تعداد نمونه‌هایی که در عمل بدان نیاز داریم کمتر از کرانی است که قضیه‌ی چبیشف به‌دست می‌دهد. داریم:

$$\begin{aligned}\Pr(|M_n - f| > ۰/۰۱) &= \Pr(\frac{X_۱ + \cdots + X_n - nf}{n} > ۰/۰۱) \\ &= \Pr(\frac{X_۱ + \cdots + X_n - nf}{\sigma\sqrt{n}} > \frac{۰/۰۱\sqrt{n}}{\sigma}) \\ &\approx \Pr(|Z| > \frac{۰/۰۱\sqrt{n}}{\sigma})\end{aligned}$$

که Z یک متغیر تصادفی با توزیع نرمال استاندارد است و به عنوان تخمینی از متغیر تصادفی اولیه در نظر گرفته شده است.
از (۱) و (۲) داریم:

$$\Pr(|Z| > \frac{۰/۰۱\sqrt{n}}{\sigma}) \le \Pr(|Z| > \frac{۰/۰۱\sqrt{n}}{\frac{۱}{۲}}) = ۰/۰۵ \qquad (۳)$$

می‌دانیم:

$$\begin{aligned}\Pr(|Z| \ge a) &= ۲\Pr(Z > a) \\ &= ۲(۱ - \Pr(Z \le a)) \\ &= ۲(۱ - \Phi(a)) \qquad (۴)\end{aligned}$$

پس از (۳) و (۴) می‌توان نوشت:

$$۲(۱ - \Phi(۰/۰۲\sqrt{n})) = ۰/۰۵$$

$$\Phi(۰/۰۲\sqrt{n}) \cong ۰/۹۷۵ \Longrightarrow n \simeq ۹۶۰۰. \quad \blacklozenge$$

مثال ۳ (Random walks) شخص مستی در مبدأ مکان قرار دارد و در هر مرحله یه قدم به سمت چپ یا راست برمی‌دارد. مطلوب است متوسط فاصله‌ی شخص از مبدأ پس از N مرحله.
متغیر تصادفی X_i را به صورت زیر تعریف می‌کنیم:

$$X_i = \begin{cases} ۱ & \text{با احتمال } \frac{۱}{۲} \\ -۱ & \text{با احتمال } \frac{۱}{۲} \end{cases}$$

اگر موقعیت شخص را در لحظه‌ی Nام با S_N نشان دهیم، داریم:

$$S_N = X_۱ + X_۲ + \cdots + X_N$$

در جلسات قبل، برای سنجش متوسط فاصله‌ی شخص پس از N مرحله نشان دادیم:

$$\mathrm{E}[S_N^۲] = N\,,\ \mathrm{E}[|S_N|] \le \sqrt{N}$$

حال می‌خواهیم ثابت کنیم که جواب مسئله ضریبی از $\sqrt{N}$ است.
در اینجا تعدادی متغیر تصادفی $i.i.d$ با $\mathrm{E}[X_i] = ۰$ و $\mathrm{Var}(X_i) = ۱$ داریم. باید یک متغیر تصادفی بیابیم که در حد به سمت توزیع نرمال میل کند. پس:

$$\begin{aligned}
\mathrm{E}[|S_N|] &= \mathrm{E}[|X_۱ + X_۲ + \cdots + X_n|] \\
&= \sqrt{N}\mathrm{E}[\Big|\frac{X_۱ + \cdots + X_N}{\sqrt{N}}\Big|] \\
&\simeq \sqrt{N}\mathrm{E}[|Z|] \\
&= \sqrt{N}\int_{-\infty}^{\infty} |z| \frac{۱}{\sqrt{۲\pi}} e^{-z^۲/۲}\, dz \\
&= \sqrt{\frac{۲N}{\pi}} \qquad \blacklozenge
\end{aligned}$$

۳ توزیع نرمال دو متغیره

در جلسات قبل راجع به همبستگی صحبت کردیم. گفتیم متغیرهای تصادفی مستقل، ناهمبسته‌اند. یعنی کوواریانس و در نتیجه ضریب همبستگی آنها برابر صفر است که ضریب همبستگی را چنین تعریف می‌کنیم:

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X\,\sigma_Y} = \frac{\mathrm{E}[XY] - \mu_X\mu_Y}{\sigma_X\,\sigma_Y}$$

توجه شود که اگر دو متغیر تصادفی همبسته نباشند، لزوما مستقل نیستند؛ یعنی ربط سیستماتیک بین این دو متغیر وجود ندارد، امّا ممکن است ربط نهانی داشته باشند. می‌توان متغیرهای نرمالی تعریف کرد که از ناهمبستگی آنها، بتوان مستقل بودنشان را نتیجه گرفت.
متغیرهای تصادفی X و Y را که دارای تابع چگالی مشترک $f_{X,Y}(x,y)$ هستند مشترکاً نرمال گوییم، اگر دارای ویژگی‌های زیر باشند:

- توزیع کناری X نرمال باشد (پس خود X باید نرمال باشد).
- توزیع Y به شرط $X = x$ به ازای هر x حقیقی نرمال باشد.
- متوسط Y به شرط $X = x$ تابعی خطی از x باشد؛ یعنی $\mathrm{E}[Y|X = x] = ax + b$.
- واریانس Y به شرط $X = x$ مقداری ثابت باشد.

می‌توان نشان داد تنها تابع توزیع مشترکی که در شرایط بالا صدق می‌کند بدین شکل است:

$$f_{X,Y}(x,y) = \frac{۱}{۲\pi\sigma_X\,\sigma_Y\sqrt{۱-\rho}}\exp\{-\frac{۱}{۲(۱-\rho^۲}\,Q(x,y)\}\ ,$$

که

$$Q(x,y) = (\frac{x-\mu_X}{\sigma_X})^۲ - ۲\rho(\frac{x-\mu_X}{\sigma_X})(\frac{y-\mu_Y}{\sigma_Y}) + (\frac{y-\mu_Y}{\sigma_Y})^۲$$

و $-۱ \le \rho \le ۱$ ضریب همبستگی X و Y است. با قرار دادن $\rho = ۰$ در این تابع، $f_{X,Y}(x,y)$ حاصل‌ضرب دو توزیع نرمال X و Y خواهد شد. پس می‌توان نتیجه گرفت که متغیرهای تصادفی مشترکاً نرمال ناهمبسته، مستقل‌اند.

مثال ۴ *متغیر تصادفی X را معدل سال اول یک دانشجو و Y را معدل سال آخر وی در نظر بگیرید. معمولاً هر دو متغیر دارای توزیع نرمال هستند و توزیعی مشترکاً نرمال دارند.*

مثال ۵ *اگر قد پدری را X و قد دخترش را Y بگیریم، این دو متغیر نیز توزیع مشترکاً نرمال خواهند داشت.*

توزیع‌های نرمال چند متغیره نیز تعریف می‌شوند که در درس آمار به آن پرداخته خواهد شد.