Probability

احتمال چیست ؟

ما در جهانی پر از عدم قطعیت زندگی می کنیم. برای مثال احتمالا فردا خونه هستم ، اگر شانس داشته باشیم در قرعه کشی برنده می شویم و... .

احتمال علم اندازه گیری عدم قطعیت است.

تابع احتمال به صورت ساده به شکل زیر بیان می شود :

$$P(A) = \frac{n_A}{n}$$

که P(A) احتمال وقوع پیشامد A در n بار آزمایش را نشان می دهد. n_A تعداد دفعاتی است که پیشامد A در n بار آزمایش رخ داده است.

برای مثال اگر یک سکه را ۱۰۰۰ دفعه پرتاب کنیم و ۸۰۰ مرتبه سکه شیر بیاید و ۲۰۰ بار خط بیاید ، می توانیم بگوییم احتمال شیر آمدن در پرتاب سکه برابر ۰.۸ است ، یا به عبارتی اگر این سکه را ۱۰ بار دیگر پرتاب کنیم انتظار داریم ۸ مرتبه به شیر برسیم و ۲ بار خط را ببینیم.

هر چه n را بزرگ تر کنیم به نتیجه دقیق تری خواهیم رسید. بر این اساس عده ای تعریف زیر را برای احتمال در نظر گرفتند :

$$Pr(A) = \lim_{n\to\infty} \frac{n(A)}{n}$$

اما این تعریف چند اشکال دارد :

۱- حد ممکن است وجود نداشته باشد.

۲- شاید حد مقدار ثابتی را ندهد ، یعنی در آزمایش های مختلف جواب های گوناگونی بگیریم.

۳- n را به سمت بی نهایت میل دادن در دنیای واقعی امکان ناپذیر است.

❖ فضای نمونه (Sample Space) :

به مجموعه کل نتایج ممکن فضای نمونه می گویند. نماد آن Ω است.

فضای نمونه به دو دسته کلی تقسیم می شود :

۱- گسسته : فضای نمونه شامل تعداد محدود یا نامحدود ولی قابل شمارش را فضای نمونه گسسته می گویند.

برای مثال فضای نمونه تاس.

۲- پیوسته : فضای نمونه شامل تعداد نامحدود و غیر قابل شمارش را فضای نمونه پیوسته می گویند.

برای مثال فضای نمونه بازه اعداد ۱ تا ۵.

❖ **واقعه (Event) :**

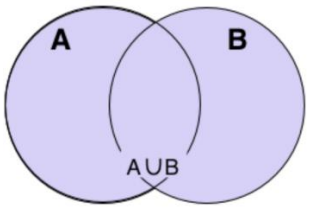
هر زیر مجموعه از فضای نمونه را واقعه یا رویداد و یا پیشامد می گویند.

برای مثال در آزمایش انداختن سکه داریم $\Omega=\{H,T\}$ که دارای ۲ عضو می باشد.

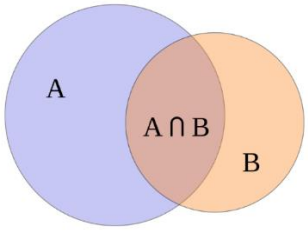
و یا در مثالی دیگر در آزمایش پرتاب دو سکه داریم$\Omega=\{(H,H),(H,T),(T,H),(T,T)\}$ که دارای ۴ عضو می باشد.

در آزمایش پرتاب دو سکه داریم $A=\{(H,H),(H,T),(T,H)\}$ که A واقعه این است که حداقل ۱ شیر بیاید.

نکته ! اجتماع دو مجموعه A و B مجموعه تمام عناصری است که در A یا در B یا در هر دو آن ها باشد و اجتماع دو مجموعه را با $A\cup B$ نمایش می دهیم.



نکته ! اشتراک دو مجموعه $A\cap B$ مجموعه عناصری است که هم در A و هم در B هستند.



❖ **Probability Function** :

P(A) را تابع احتمال تعریف می کنند که احتمال وقوع پیشامد A را بر اساس اصول موضوعه مشخص می کند.

برای مثال احتمال ۶ آمدن در یک بار پرتاب تاس برابر است با ۱/۶ و یا احتمال شیر آمدن در یک بار پرتاب سکه برابر است با ۰.۵ .

❖ **اصول موضوعه (Probability Axioms) :**

۱- $P(A) >= 0$

۲- $P(\Omega) = 1$

۳- اگر A و B ناسازگار باشند آنگاه $P(AUB)=P(A)+P(B)$

نکته ! دو مجموعه نسبت به هم ناسازگارند (Disjoint) اگر اشتراک دو مجموعه با هم تهی باشد. یعنی دو مجموعه هیچ اشتراکی با هم نداشته باشند.

برای مثال اگر A پیشامد زوج آمدن یک تاس و B پیشامد فرد آمدن یک تاس باشد آنگاه A و B نسبت به هم ناسازگارند.

و یا اگر A پیشامد آمدن ۴ در پرتاب تاس و B آمدن ۵ باشد ، دو پیشامد نسبت به هم ناسازگارند.

پس داریم :

$$P(AUB)=P(A)+P(B) = 1/3$$

حالت کلی تر اصل ۳ به شکل زیر می باشد :

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} A_i$$

در نتیجه تابع احتمال ، تابعی است که از اصول موضوعه پیروی می کند.

احتمال شرطی (Conditional Probability) :

احتمال شرطی یکی از مباحث مهم و اصلی در علم احتمال است. احتمال رخدادن واقعه A به شرط رخدادن واقعه B را به صورت زیر تعریف می‌کنیم :

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

به صورت P(A Given B) هم بیان می کنند.

✓ **مثال :** احتمال زوج آمدن در یک بار پرتاب تاس به شرط نیامدن ۶ چقدر است.

پیشامد B یعنی نیامدن ۶ در پرتاب تاس ، پس :

$$B = \{1, 2, 3, 4, 5\}$$

پیشامد A یعنی زوج آمدن در پرتاب تاس ، پس :

$$A = \{2, 4, 6\}$$

یعنی داریم :

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\{2, 4\})}{P(\{1, 2, 3, 4, 5\})} = \frac{2}{5}$$

❖ **قاعده بیز :**

قاعده بیز به صورت زیر بیان می شود :

$$P(A|B) = \frac{P(B|A) \, . P(A)}{P(B)}$$

✓ **مثال :** با استفاده از یک الگوریتم یادگیری ماشین ، احتمال آن که یک فرد سرطان داشته باشد و الگوریتم تشخیص دهد که سرطان دارد ، ۹۵ درصد است. احتمال آن که یک فرد سرطان نداشته باشد و الگوریتم تشخیص دهد که سرطان ندارد ، ۹۴ درصد است. احتمال سرطان داشتن یک فرد در جامعه 10^{-4} است. اگر نتیجه تشخیص الگوریتم برای یک فرد مثبت باشد ، با چه احتمالی فرد سرطان دارد.

برای حل مسئله فرض می کنیم :

احتمال سرطان داشتن یک فرد : A

نتیجه تشخیص الگوریتم مثبت باشد : B

احتمال سرطان نداشتن یک فرد : A'

نتیجه تشخیص الگوریتم منفی باشد : B'

داریم :

$$P(B|A) = 0.95$$
$$P(B'|A') = 0.94$$
$$P(A) = 0.0001$$

می خواهیم احتمال زیر را بدست بیاوریم :

$$P(A|B) = ?$$

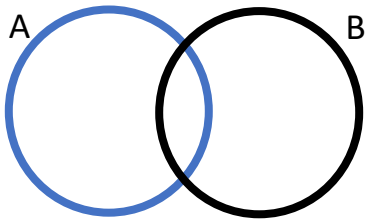
با استفاده از قاعده بیز داریم :

$$P(A|B) = \frac{P(B|A)\,.\,P(A)}{P(B)}$$

مقادیر $P(A)$, $P(B|A)$ جز فرضیات مسئله است ولی مقدار P(B) را باید بدست بیاوریم.

داریم :

$$P(B) = P(B \cap A) + P(B \cap A')$$



سپس :

$$P(B \cap A) = P(B|A)\,.\,P(A)$$
$$P(B \cap A') = P(B|A')\,.\,P(A')$$

و

$$P(A') = 1 - P(A)$$

به صورت منطقی داریم :

$$P(B|A') = 1 - P(B'|A')$$

و در نهایت :

$$P(A|B) = \frac{0.95 * 0.0001}{(0.95 * 0.0001) + ((1 - 0.94) * (1 - 0.0001))} = 0.00158$$

نکته ! حالت کلی قاعده بیز به شکل زیر است :

$$P(A_i|B) = \frac{P(B|A_i) \,.\, P(A_i)}{\sum P(B|A_i) \,.\, P(A_i)}$$

❖ **قانون احتمال کل :**

اگر فضای نمونه به n تا فضا تفکیک یا افراز شده باشد ، داریم :

$$P(B) = \sum_{i=1}^{n} P(B|A_i) \,.\, P(A_i)$$

✓ **مثال :** فرض کنید دو جعبه لامپ از کارخانه رسیده است. جعبه اول شامل ۲۱ لامپ است که ۱۲ تای آن سالم و ۹ تای آن خراب است. جعبه دوم شامل ۲۱ لامپ است که ۱۰ تای آن سالم و ۱۱ تای آن خراب است. لامپی را از جعبه بیرون می‌آوریم. احتمال این که آن لامپ خراب باشد چقدر است.

برای حل مسئله داریم :

$$P(Bad\ light\ bulb) = P(Bad\ light\ bulb \mid Box\ 1)P(Box\ 1) + P(Bad\ light\ bulb \mid Box\ 2)P(Box\ 2)$$

$$\Longrightarrow P(Bad\ light\ bulb) = \frac{1}{2} \cdot \frac{9}{21} + \frac{1}{2} \cdot \frac{11}{21} = \frac{10}{21}$$

✓ **مثال :** در مثال قبل ، احتمال این که لامپ خراب متعلق به جعبه دوم باشد چقدر است.

برای حل مسئله داریم :

$$P(Box\ 2 \mid Bad\ light\ bulb) = \frac{P(Bad\ light\ bulb \mid Box\ 2)P(Box\ 2)}{P(Bad\ light\ bulb)}$$

$$\Longrightarrow P(Box\ 2 \mid Bad\ light\ bulb) = \frac{\frac{1}{2} \cdot \frac{11}{21}}{\frac{10}{21}} = \frac{11}{20}$$

استقلال :

دو پیشامد نسبت به هم مستقل هستند اگر یکی از دو شرط زیر برقرار باشد :

۱- $P(A|B) = P(A), P(B|A) = P(B)$

۲- احتمال یکی از پیشامدها صفر باشد.

در واقع وقتی دو پیشامد نسبت به هم مستقل هستند ، دانستن اطلاعات یک پیشامد بر پیشامد دیگر تاثیری ندارد.

✓ **مثال :** فرض کنید می‌خواهیم احتمال follow back توسط یک فرد در instagram را بدست آوریم. اگر پیشامد A را follow back توسط آن فرد و پیشامد B را از دماغ فیل افتادن این شخص تعریف کنیم ، پر واضح است که این دو پیشامد مستقل نیستند و وابسته‌اند. زیرا اگر بدانیم که فرد مورد نظر از دماغ فیل افتاده ، احتمال follow back بسیار کم خواهد بود.

✓ **مثال :** در یک بار پرتاب سکه و تاس ، احتمال آن که شیر و ۶ بیاید چقدر است.

برای حل این مسئله فرض می کنیم که A پیشامد شیر آمدن در پرتاب سکه است و B پیشامد ۶ آمدن در پرتاب تاس است.

فضای نمونه به شکل زیر است :

Ω={(H,1), (H,2), (H,3), (H,4), (H,5), (H,6), (T,1), (T,2), (T,3), (T,4), (T,5), (T,6)}

داریم :

$$P(B) = \frac{1}{6}$$

همچنین داریم :

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{1}{6}$$

پس می توان نتیجه گرفت که :

$$P(B|A) = P(B)$$

✓ **مثال :** بر اساس جدول زیر بررسی کنید که آیا جنسیت فرزند اول بر روی جنسیت فرزند دوم اثر دارد یا خیر.

فرزند دوم / اول	پسر	دختر
پسر	۱/۴	۱/۴
دختر	۱/۴	۱/۴

جنسیت فرزند اول را پیشامد A و جنسیت فرزند دوم را B در نظر می گیریم.

برای بررسی این موضوع باید مستقل بودن پیشامدهای A و B را بررسی کنیم.

داریم :

فرزند اول پسر : A

فرزند اول دختر : A′

فرزند دوم پسر : B

فرزند دوم دختر : B′

همچنین داریم :

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{4}}{P(B \cap A) + P(B \cap A')} = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2}$$

و نیز داریم :

$$P(A) = \frac{1}{2}$$

پس با توجه به اینکه هر دو مقدار برابر است ، این دو پیشامد مستقل از هم هستند و اثری روی هم ندارند.

✓ **مثال :** بر اساس جدول زیر بررسی کنید که آیا جنسیت فرزند اول بر روی جنسیت فرزند دوم اثر دارد یا خیر.

فرزند دوم / اول	پسر	دختر
پسر	۱/۸	۱/۸
دختر	۳/۸	۳/۸

جنسیت فرزند اول را پیشامد A و جنسیت فرزند دوم را B در نظر می گیریم.

برای بررسی این موضوع باید مستقل بودن پیشامدهای A و B را بررسی کنیم.

داریم :

فرزند اول پسر : A

فرزند اول دختر : A'

فرزند دوم پسر : B

فرزند دوم دختر : B'

همچنین داریم :

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{8}}{P(B \cap A) + P(B \cap A')} = \frac{\frac{1}{8}}{\frac{4}{8}} = \frac{1}{4}$$

و نیز داریم :

$$P(A) = \frac{1}{4}$$

پس با توجه به اینکه هر دو مقدار برابر است ، این دو پیشامد مستقل از هم هستند و اثری روی هم ندارند.

✓ **مثال :** بر اساس جدول زیر بررسی کنید که آیا جنسیت فرزند اول بر روی جنسیت فرزند دوم اثر دارد یا خیر.

فرزند دوم / اول	پسر	دختر
پسر	۳/۱۲	۱/۱۲
دختر	۵/۱۲	۳/۱۲

جنسیت فرزند اول را پیشامد A و جنسیت فرزند دوم را B در نظر می گیریم.
برای بررسی این موضوع باید مستقل بودن پیشامدهای A و B را بررسی کنیم.
داریم :
فرزند اول پسر : A
فرزند اول دختر : A'
فرزند دوم پسر : B
فرزند دوم دختر : B'
همچنین داریم :

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{3}{12}}{P(B \cap A) + P(B \cap A')} = \frac{\frac{3}{12}}{\frac{8}{12}} = \frac{3}{8}$$

و نیز داریم :

$$P(A) = \frac{1}{3}$$

پس با توجه به اینکه هر دو مقدار برابر نیست ، این دو پیشامد نسبت به هم وابسته اند و روی هم اثر دارند.

نکته ! A و B مستقل هستند اگر و تنها اگر $P(A \cap B) = P(A).P(B)$.

❖ **احتمال حاشیه ای (Marginal Probability) :**

با توجه به مثال قبل داریم :

	دختر	پسر	فرزند دوم / اول
۴/۱۲	۱/۱۲	۳/۱۲	پسر
۸/۱۲	۳/۱۲	۵/۱۲	دختر
	۴/۱۲	۸/۱۲	

❖ **استقلال شرطی :**

پیشامدهای A و B با دانستن C مستقل هستند اگر

$$P(A|B,C) = P(A|C), P(B|A,C) = P(B|C)$$

و یا :

$$P(A|B \cap C) = P(A|C), P(B|A \cap C) = P(B|C)$$

✓ **مثال :** یک بازی را در نظر بگیرید که دو تاس دارد. اگر جمع دو تاس بیشتر از ۱۰ باشد ، برنده بازی می شویم و در غیر این صورت بازنده خواهیم شد. پیشامد A را ۵ آمدن تاس اول و B را ۵ آمدن تاس دوم در نظر بگیرید. همچنین پیشامد C را برنده شدن بازیکن فرض می‌کنیم. بررسی کنید که آیا پیشامد A و B مستقل هستند یا خیر.

برای حل سوال ابتدا پیشامد C را در نظر نمی گیریم ، داریم :

$$P(A \cap B) = \frac{1}{36}, P(A).P(B) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

پس مستقل هستند.

اما اگر پیشامد C را در نظر بگیریم ، داریم :

$$P(A|C) = \frac{1}{3}$$

بدلیل آن که برای رخ دادن پیشامد C ، یعنی برای آن که جمع دو تاس بیشتر از ۱۰ باشد حتما باید اعداد ۴ یا ۵ یا ۶ بیاید و چون احتمال ۵ آمدن را بررسی کردیم پس حاصل ۱/۳ می شود.

و همچنین داریم :

$$P(A|B,C) = 1$$

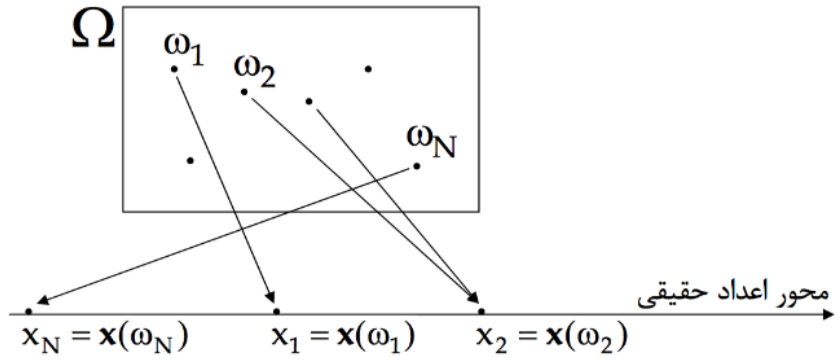
زیرا زمانی که مجموع تاس حتما ۱۰ می شود و حتما تاس دوم ۵ آمده است پس قطعا تاس اول نیز ۵ آمده است.

بدلیل آن که حاصل دو عبارت برابر نیستند پس دو پیشامد نسبت به هم مستقل نیستند.

نکته ! $P(A|B,C) = P(A|C) \rightarrow \frac{P(A\cap B\cap C)}{P(B\cap C)} = \frac{P(A\cap C)}{P(C)} \rightarrow \frac{P(A\cap B\cap C)}{P(C)} = \frac{P(A\cap C).P(B\cap C)}{P(C).P(C)} \rightarrow P(A\cap B|C) = P(A|C).P(B|C)$

متغیر تصادفی (Random Variable) :

به تابعی می گویند که خروجی های یک آزمایش تصادفی را به اعداد حقیقی نسبت می دهد.



برای مثال در آزمایش تصادفی پرتاب سکه ، شیر آمدن یا Head را به عدد حقیقی صفر و خط آمدن یا Tail را به عدد حقیقی ۱ نسبت می دهیم.

یا در مثالی دیگر در آزمایش تصادفی پرتاب تاس ، پیشامد هر یک از اعداد روی تاس را به یک عدد حقیقی نسبت می دهیم.

نکته ! به دلیل آن که با این تابع مانند متغیرهای ریاضی رفتار خواهیم کرد ، یعنی از عملیات های جمع ، ضرب و... استفاده خواهیم کرد ، به آن متغیر تصادفی می گویند.

نکته ! متغیرهای تصادفی را معمولا با حروف بزرگ و مقادیر آن ها را با حروف کوچک نشان می دهند.

متغیر تصادفی گسسته :

متغیرهای تصادفی گسسته مربوط به آزمایش هایی هستند که خروجی آن ها گسسته است.

برای مثال در پرتاب سکه خروجی آزمایش مقدار گسسته دارد ، یعنی یا صفر می شود و یا ۱.

❖ تابع جرم احتمال (Probability Mass Function) :

تابع جرم احتمال ، احتمال مقادیر مختلف متغیر تصادفی را نشان می دهد.

تابع جرم احتمال به صورت زیر نوشته می شود :

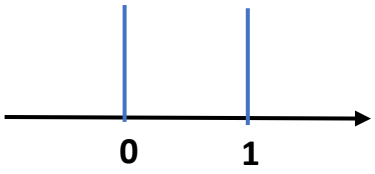
$$p_X(x) = P(\{X = x\})$$

که در آن X متغیر تصادفی و x مقدار آن است.

برای مثال پرتاب سکه ، تابع جرم احتمال به شکل زیر است :

$$p_X(0) = P(\{X = 0\}) = \frac{1}{2}$$

$$p_X(1) = P(\{X = 1\}) = \frac{1}{2}$$



نکته ! $R_X = Range(X)$ محدوده ای است که متغیر تصادفی X می تواند داشته باشد.

ویژگی های تابع جرم احتمال به شرح زیر است :

۱- $\forall x \in R_x : 0 \leq p(x) \leq 1$

۲- $\sum_{x \in R_x} p(x) = 1$

۳- اگر x در محدوده Range(X) نباشد ، مقدار احتمال آن صفر در نظر گرفته می شود.

✓ **مثال :** اگر در یک آزمایش متغیر تصادفی ، مقدار ماکزیمم خروجی پرتاب دو تاس سالم باشد ، تابع جرم احتمال را برای این آزمایش بدست بیاورید.

مقدار X	۱	۲	۳	۴	۵	۶
PMF	۱/۳۶	۳/۳۶	۵/۳۶	۷/۳۶	۹/۳۶	۱۱/۳۶

تمام ویژگی های تابع جرم احتمال به درستی صدق می کند.

✓ **مثال :** اگر در یک آزمایش متغیر تصادفی ، تعداد شیر آمدن در پرتاب سه سکه باشد تابع جرم احتمال را برای این آزمایش بدست بیاورید.

فضای نمونه به شکل زیر است :

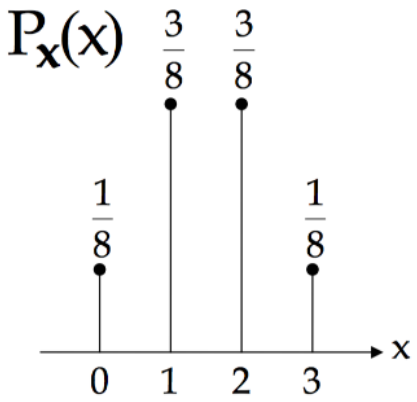
$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

داریم :

$$\begin{cases} i = 0 : \omega \in \{TTT\} \rightarrow p(0) = \frac{1}{8} \\ i = 1 : \omega \in \{HTT, THT, TTH\} \rightarrow p(1) = \frac{3}{8} \\ i = 2 : \omega \in \{HHT, HTH, THH\} \rightarrow p(2) = \frac{3}{8} \\ i = 3 : \omega \in \{HHH\} \rightarrow p(3) = \frac{1}{8} \end{cases}$$

که i متغیر تصادفی است.

تابع احتمال این آزمایش به شکل زیر است :



❖ **تابع توزیع انباشته یا تجمعی (Cumulative Distribution Function) :**

تابع توزیع تجمعی ، احتمال کمتر بودن مقدار متغیر تصادفی از یک مقدار به خصوص را نشان می دهد.

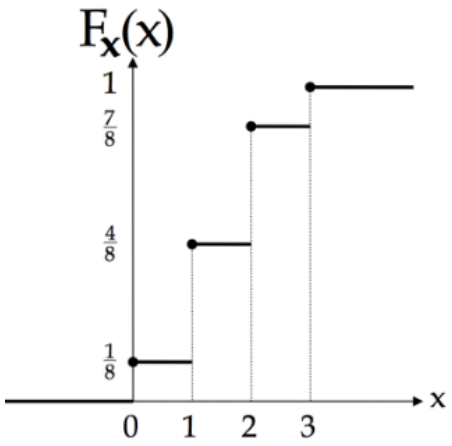
$$F(x) = P(X \leq x)$$

✓ **مثال :** اگر در یک آزمایش متغیر تصادفی ، مقدار ماکزیمم خروجی پرتاب دو تاس سالم باشد ، تابع جرم احتمال و تابع توزیع تجمعی را برای این آزمایش بدست بیاورید.

مقدار X	۱	۲	۳	۴	۵	۶
PMF	۱/۳۶	۳/۳۶	۵/۳۶	۷/۳۶	۹/۳۶	۱۱/۳۶
CDF	۱/۳۶	۴/۳۶	۹/۳۶	۱۶/۳۶	۲۵/۳۶	۳۶/۳۶

در واقع تابع توزیع تجمعی ، جمع مقادیر متغیرهای تصادفی تا قبل از متغیر تصادفی مد نظر است.

✓ **مثال :** اگر در یک آزمایش متغیر تصادفی ، تعداد شیر آمدن در پرتاب سه سکه باشد تابع توزیع تجمعی را برای این آزمایش رسم کنید.



نکته ! برای متغیر تصادفی گسسته ، تابع توزیع تجمعی به صورت پلکانی است.

خواص تابع توزیع تجمعی به شرح زیر است :

۱- $F(-\infty) = 0, F(\infty) = 1$

۲- $x_1 \leq x_2 \rightarrow F(x_1) \leq F(x_2)$

۳- $P(\{X > x\}) = 1 - F(x)$

امید ریاضی (Expected Value) :

یکی از شاخصه های مهم برای متغیرهای تصادفی ، امید ریاضی یا Expected Value است که به واسطه آن می توان شهود و درک بهتری نسبت به یک آزمایش تصادفی بدست آورد.

Expected Value به عنوان میانگین وزن دار بر روی متغیر تصادفی محسوب می شود. به عبارت دیگر داریم :

$$E(X) = \sum_{x \in R_X} xp(X = x)$$

✓ **مثال :** Expected Value آزمایش پرتاب سکه را بدست بیاورید.

داریم :

$$E(X) = \frac{1}{2} \times 1 + \frac{1}{2} \times 0 = \frac{1}{2}$$

✓ **مثال :** Expected Value آزمایش پرتاب تاس را بدست بیاورید.

داریم :

$$E(X) = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = \frac{7}{2}$$

با استفاده از پایتون داریم :

```python
import numpy as np

X = np.array([1, 2, 3, 4, 5, 6])

x = np.random.choice(X, size = 1000)

print("E(X):", np.average(x))
```

نکته ! Expected Value عدد ثابت برابر با همان عدد ثابت می شود.

خواص Expected Value به شرح زیر است :

۱- $E(aX + b) = aE(X) + b$

۲- $E(X + Y) = E(X) + E(Y)$

✓ **مثال :** اگر مقادیر متغیر تصادفی X بدین شکل باشد ، $X = \{1, 2, 3, 4, 5, 6\}$ و اگر رابطه $Y = 2X + 3$ بین متغیرهای تصادفی X و Y برقرار باشد. مقدار $E(Y)$ را بدست بیاورید.

روش اول :

$$Y = \{5, 7, 9, 11, 13, 15\}$$

پس داریم :

$$E(Y) = 5 \times \frac{1}{6} + 7 \times \frac{1}{6} + 9 \times \frac{1}{6} + 11 \times \frac{1}{6} + 13 \times \frac{1}{6} + 15 \times \frac{1}{6}$$
$$= \frac{60}{6} = 10$$

روش دوم :

$$E(X) = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = \frac{7}{2}$$

پس داریم :

$$E(Y) = 2E(X) + 3 = 2 \times \frac{7}{2} + 3 = 10$$

واریانس (Variance) :

یکی دیگر از شاخصه های مهم برای متغیرهای تصادفی ، واریانس است.
Expected Value با وجود این‌که در مورد مقدار میانگین اطلاع می‌داد ، در مورد میزان پراکندگی خروجی‌های یک آزمایش تصادفی یا متغیر تصادفی اطلاع خاصی نمی‌داد.

✓ **مثال :** فشار خون در سه گروه مجزا پنج نفره از افراد را مورد بررسی قرار دادیم که به شرح جدول زیر است :

گروه x	گروه y	گروه z
۱۴	۱۱	۱۰
۱۱	۱۰	۱۰
۱۰	۱۰	۱۰
۹	۱۰	۱۰
۶	۹	۱۰

پراکندگی متغیر تصادفی را بررسی کنید.

متغیر تصادفی در این مثال فشار خون است. Expected Value برای هر سه گروه برابر ۱۰ است. اگر از گروه z یک فرد را انتخاب کنیم ، قطعا فشار خون این فرد برابر ۱۰ است زیرا همه نمونه های این گروه برابر ۱۰ است اما اگر از گروه y یک فرد را انتخاب کنیم ، فشار خون این فرد حدود ۱۰ است چرا که اکثر نمونه های این گروه برابر ۱۰ هستند و مابقی نمونه های این گروه نزدیک به ۱۰ است. این بدین معنی است که پراکندگی نمونه های این گروه کم است. حال اگر از گروه x یک فرد را انتخاب کنیم دیگر نمی توان به طور قطع در مورد مقدار فشار خون نظر داد زیرا پراکندگی نمونه های این گروه زیاد است.

می توان گفت مفهوم واریانس بدین شکل است که نمونه ها چقدر حول Expected Value هستند.

❖ **تعریف واریانس :**

به میانگین ، فاصله متغیر تصادفی از Expected Value واریانس می گویند.

داریم :

$$Var(X) = E\big(X - E(X)\big)^2$$

و یا داریم :

$$Var(X) = E(X - \mu)^2$$

همچنین با ساده کردن عبارت بالا داریم :

$$Var(X) = E\big(X - E(X)\big)^2 = E(X^2) - (E(X))^2$$

نکته ! به حاصل جذر واریانس ، انحراف معیار می گویند. داریم :

$$SD(X) = \sqrt{Var(X)}$$

نکته ! واریانس عدد ثابت برابر با صفر می شود. داریم :

$$Var(C) = E(C - E(C))^2 = E(C - C)^2 = 0$$

خواص واریانس به شرح زیر است :

۱- $Var(aX + b) = a^2 Var(X)$

۲- اگر $X_1, X_2, \dots, X_n$ متغیرهای تصادفی دو به دو مستقل از هم باشند و $X = X_1 + X_2 + \cdots + X_n$ باشد ، داریم :

$$Var(X) = Var(X_1) + Var(X_2) + \cdots + Var(X_n)$$

✓ **مثال :** واریانس آزمایش پرتاب تاس را محاسبه کنید.

داریم :

X	X – E(X)	$(X - E(X))^2$
1	1 – 7/2	25/4
2	2 – 7/2	9/4
3	3 – 7/2	¼
4	4 – 7/2	¼

5	5 – 7/2	9/4
6	6 – 7/2	25/4

سپس :

$$Var(X) = E\big(X - E(X)\big)^2 = 2 \times \left(\frac{25}{4} \times \frac{1}{6} + \frac{9}{4} \times \frac{1}{6} + \frac{1}{4} \times \frac{1}{6}\right)$$
$$= \frac{35}{12} = 2.92$$

برای محاسبه واریانس و انحراف معیار با استفاده از پایتون داریم :

```python
import numpy as np

X = [1, 2, 3, 4, 5, 6]

print("Var(X):", np.var(X))
print("SD(X):", np.std(X))
```

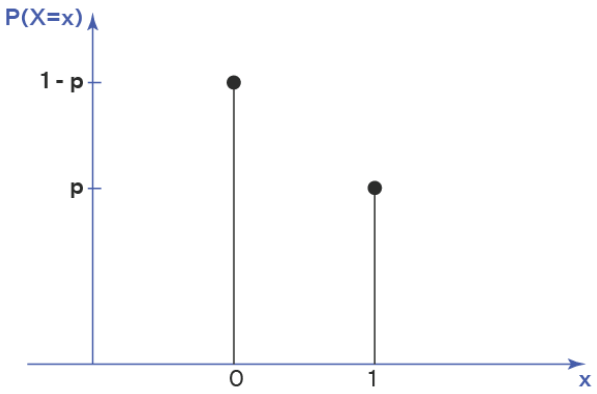
توزیع های احتمالاتی گسسته :

نمودارهای PMF برخی از آزمایش های تصادفی بسیار پر کاربرد هستند که چند نمونه از این توزیع های احتمالاتی گسسته را بررسی خواهیم کرد.

❖ توزیع برنولی (Bernoulli Distribution) :

توزیع برنولی نتیجه آزمایش برنولی است. در این آزمایش تنها دو حالت وجود دارد که یکی از این حالت ها پیروزی و حالت دوم شکست است. به عبارت دیگر خروجی این آزمایش یا پیروزی است یا شکست. برای حالت پیروزی متغیر تصادفی را ۱ و برای حالت شکست متغیر تصادفی را صفر در نظر می گیرند. احتمال پیروزی را برابر با p و احتمال شکست را برابر 1-p یا q در نظر می گیریم.

نمودار PMF این توزیع به شکل زیر است :



Expected Value توزیع برنولی به شکل زیر است :

$$E(X) = \sum xp(x) = 1 \times p + 0 \times (1 - p) = p$$

واریانس توزیع برنولی نیز به شکل زیر است :

$$Var(X) = E(X^2) - E(X)^2 = p - p^2 = p(1 - p) = pq$$

- ✓ **مثال :** در آزمایش پرتاب سکه اگر شیر آمدن پیروزی و خط آمدن شکست باشد و اگر فرض کنیم با احتمال 0.6 در این آزمایش پیروز خواهیم شد ، Expected Value و واریانس این آزمایش را حساب کنید.

این آزمایش از توزیع برنولی پیروی می کند و مقدار p آن برابر 0.6 است.

پس داریم :

$$E(X) = p = 0.6$$

همچنین داریم :

$$Var(X) = p(1 - p) = 0.6 \times 0.4 = 0.24$$

برای پیاده سازی توزیع برنولی در پایتون داریم :

```python
import matplotlib.pyplot as plt
from scipy.stats import bernoulli

b = bernoulli(0.6)

X = [0, 1]

plt.bar(X, b.pmf(X))
plt.xlabel("Random Variable")
plt.ylabel("Probability")
plt.title("Bernoulli Distribution")
plt.xlim(-2, 3)
plt.show()
```

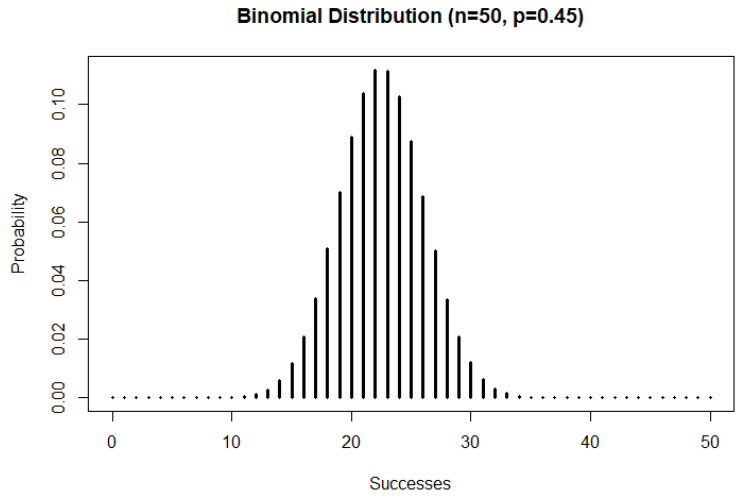
❖ **توزیع دوجمله ای (Binomial Distribution) :**

در آزمایش برنولی فقط یکبار آزمایش انجام می شد و نتیجه آن پیروزی یا شکست بود اما اگر آزمایش برنولی را چند بار انجام دهیم که مجدد نتیجه آن یا پیروزی و یا شکست باشد ، توزیع دوجمله ای را خواهیم داشت. اگر تعداد تکرار آزمایش را n در نظر بگیریم و احتمال پیروزی را p و احتمال شکست را 1-p در نظر بگیریم آنگاه داریم :

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{(n-k)!\,k!} p^k (1-p)^{n-k}$$

در رابطه بالا k تعداد پیروزی در n بار انجام آزمایش است.

نمودار PMF این توزیع برای مثال خاص به شکل زیر است :



Expected Value توزیع دوجمله ای به شکل زیر است :

$$E(X) = np$$

واریانس توزیع دوجمله ای نیز به شکل زیر است :

$$Var(X) = Var(X_1) + Var(X_2) + \cdots + Var(X_n) = np(1-p) \\ = npq$$

- ✓ **مثال :** اگر آزمایش پرتاب سکه را ۱۰۰ بار انجام دهیم و اگر پیروزی را شیر آمدن سکه و شکست را خط آمدن آن در نظر بگیریم. چنانچه احتمال پیروزی در هر پرتاب برابر 0.6 باشد ، احتمال ۴۵ بار پیروز شدن در ۱۰۰ بار انجام آزمایش و مقدار Expected Value و مقدار واریانس را محاسبه کنید.

در این آزمایش مقدار n برابر ۱۰۰ و مقدار k برابر ۴۵ است.

پس داریم :

$$P(45) = \frac{100!}{(100-45)!\,45!} 0.6^{45}(1-0.6)^{100-45}$$

مقدار Expected Value نیز برابر است با :

$$E(X) = np = 100 \times 0.6 = 60$$

و مقدار واریانس برابر است با :

$$Var(X) = npq = 100 \times 0.6 \times 0.4 = 24$$

برای پیاده سازی توزیع دوجمله ای در پایتون داریم :

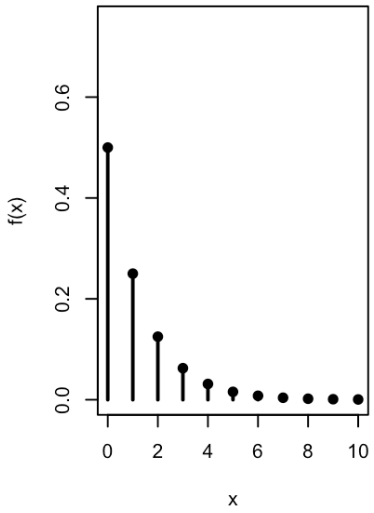
```
from scipy.stats import binom
import matplotlib.pyplot as plt

p = 0.6
n = 100

b = binom(n, p)

plt.bar(range(1, 101), b.pmf(range(1, 101)))
plt.title("Binomial Distribution")
plt.xlim(30, 90)
plt.show()
```

❖ **توزیع هندسی (Geometric Distribution) :**

در توزیع دوجمله ای هدف بدست آوردن احتمال وقوع k بار پیروزی در n بار انجام آزمایش است. اما در توزیع هندسی هدف بدست آوردن پیروزی است. یعنی اینقدر آزمایش را تکرار می کنیم تا به پیروزی یا نتیجه دلخواه برسیم. اگر p را احتمال پیروزی در هر بار انجام آزمایش در نظر بگیریم و k را تعداد انجام آزمایش تا رسیدن به پیروزی در نظر بگیریم داریم :

$$P(X = k) = (1 - p)^{k-1}p$$

نمودار PMF این توزیع برای مثال خاص به شکل زیر است :



Expected Value توزیع هندسی به شکل زیر است :

$$E(X) = \frac{1}{p}$$

واریانس توزیع هندسی نیز به شکل زیر است :

$$Var(X) = \frac{1-p}{p^2}$$

- ✓ **مثال :** علی به دنبال فرد خاصی می گردد. او شروع به پرسیدن نام افراد حاضر در سالن می کند تا فرد مد نظرش را پیدا کند. اگر پیدا کردن فرد مد نظر را پیروزی در نظر بگیریم و اگر با احتمال 0.7 در پرسیدن از افراد شکست بخوریم یعنی فرد مد نظر ما نباشند. احتمال آن که ۴ نفر قبل از پیدا کردن فرد مد نظر مورد پرسش قرار بگیرند چقدر است.

داریم :

$$P(X = 4) = (1 - 0.3)^{4-1} \times 0.3 = 0.1029$$

برای پیاده سازی توزیع هندسی در پایتون داریم :

```
from scipy.stats import geom
import matplotlib.pyplot as plt

p = 0.3
X = [1, 2, 3, 4, 5, 6, 7, 8]

g = geom.pmf(X, p)
plt.plot(X, g, "bo")
plt.title("Geometric Distribution")
plt.vlines(X, 0, g, lw = 8)
plt.show()
```

متغیر تصادفی پیوسته :

متغیرهای تصادفی پیوسته مربوط به آزمایش هایی هستند که خروجی آن ها پیوسته است.

برای مثال مقادیر مربوط به قد افراد که پیوسته است.

❖ تابع چگالی احتمال (Probability Density Function) :

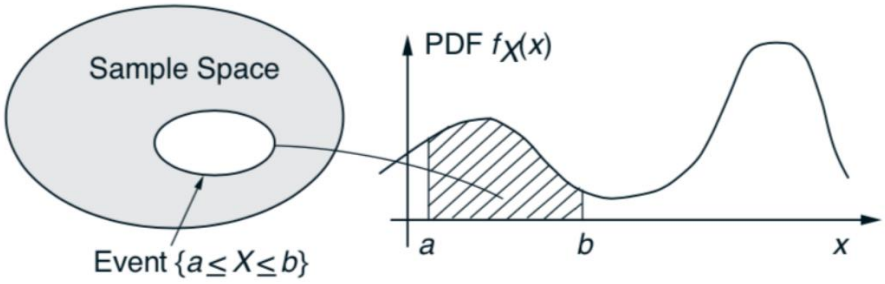
تابع چگالی احتمال ، احتمال مقادیر مختلف متغیر تصادفی پیوسته را نشان می دهد.

تابع چگالی احتمال به شکل زیر تعریف می شود :

$$P(X \in A) = \int f_X(x)dx$$

همچنین داریم :

$$P(X \in I) = P(a \leq X \leq b) = \int_a^b f_X(x)dx$$



با توجه به رابطه بالا می توان با بدست آوردن انتگرال در بازه مشخص ، مقدار احتمال رخ دادن متغیر تصادفی با مقادیر مشخص در آن بازه را بدست آورد.

نکته ! به ازای مقدار خاص برای متغیر تصادفی داریم :

$$P(X = a) = \int_a^a f_X(x)dx = 0$$

با توجه به نکته بالا باز یا بسته بودن بازه اهمیتی ندارد.

$$P(a \leq X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a < X \leq b)$$

نکته ! تابع چگالی احتمال ، تابعی نامنفی است یعنی :

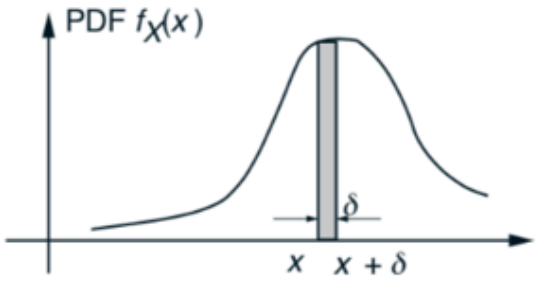
$$f_X(x) \geq 0$$

نکته ! برای تابع چگالی احتمال داریم :

$$P(-\infty \leq X \leq \infty) = \int_{-\infty}^{\infty} f_X(x)dx = 1$$

نکته ! با توجه به نکته اول ، نمی توان احتمال در یک نقطه را بدست آورد. اما اگر بخواهیم تقریبی از احتمال را در آن نقطه داشته باشیم ، می توان بازه ای بسیار کوچک حول آن نقطه در نظر گرفت و احتمال آن بازه را بدست آورد. یعنی داریم :

$$P(x \leq X \leq x + \delta) = \int_x^{x+\delta} f_X(t)dt \approx f_X(x).\delta$$



✓ **مثال :** تابع زیر را به عنوان تابع توزیع چگالی متغیر تصادفی X تعریف کرده ایم.
بررسی کنید که آیا این تابع معتبر است یا خیر.

$$f_X = \begin{cases} \frac{1}{2\sqrt{x}} & 0 < x \leq 1 \\ 0 & otherwise \end{cases}$$

داریم :

$$P(-\infty \leq X \leq \infty) = 1 \rightarrow \int_{-\infty}^{\infty} f_X(x)dx = \int_0^1 \frac{1}{2\sqrt{x}}dx$$
$$= \sqrt{x}\Big|_0^1 = 1$$

پس معتبر است.

❖ **تابع توزیع انباشته یا تجمعی (Cumulative Distribution Function) :**

تابع توزیع تجمعی ، احتمال کمتر بودن مقدار متغیر تصادفی از یک مقدار به خصوص را نشان می دهد.

داریم :

$$F(x) = P(X \leq x)$$

پس :

$$F(a) = \int_{-\infty}^{a} f_X(x)dx$$

خواص تابع توزیع تجمعی به شرح زیر است :

۱- $\lim_{x\to-\infty} F(x) = 0$ زیرا $F(-\infty) = P(\{X \leq -\infty\}) = 0$

۲- $\lim_{x\to\infty} F(x) = 1$ زیرا $F(\infty) = P(\{X \leq \infty\}) = 1$

۳- $x_1 \leq x_2 \rightarrow F(x_1) \leq F(x_2)$

۴- $P(\{X > x\}) = 1 - F(x)$

۵- $P(\{x_1 < x \leq x_2\}) = F(x_2) - F(x_1)$

امید ریاضی پیوسته (Expected Value) :

Expected Value برای متغیر تصادفی پیوسته به صورت زیر بدست می آید.

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

✓ **مثال :** اگر تابع چگالی احتمال متغیر تصادفی x به شکل زیر تعریف شده باشد.

$$f(x) = \begin{cases} 2x & x \in [0,1] \\ 0 & otherwise \end{cases}$$

Expected Value آن را بدست بیاورید.

داریم :

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_{0}^{1} x(2x)dx = \frac{2}{3}$$

خواص Expected Value به شرح زیر است :

۱- $E(aX + b) = aE(X) + b$

۲- $E(X + Y) = E(X) + E(Y)$

واریانس پیوسته (Variance) :

واریانس برای متغیر تصادفی پیوسته همانند متغیر تصادفی گسسته است.

به میانگین ، فاصله متغیر تصادفی از Expected Value واریانس می گویند.

داریم :

$$Var(X) = E\big(X - E(X)\big)^2$$

و یا داریم :

$$Var(X) = E(X - \mu)^2$$

همچنین با ساده کردن عبارت بالا داریم :

$$Var(X) = E\big(X - E(X)\big)^2 = E(X^2) - (E(X))^2$$

نکته ! به حاصل جذر واریانس ، انحراف معیار می گویند. داریم :

$$SD(X) = \sqrt{Var(X)}$$

خواص واریانس به شرح زیر است :

۱- $Var(aX + b) = a^2 Var(X)$

توزیع های احتمالاتی پیوسته :

نمودارهای PDF برخی از آزمایش های تصادفی بسیار پر کاربرد هستند که چند نمونه از این توزیع های احتمالاتی پیوسته را بررسی خواهیم کرد.

❖ توزیع یکنواخت (Uniform Distribution) :

در توزیع یکنواخت ، احتمال رخدادن همه ی نقاط در بازه مثلا [a, b] برابر و یکسان است.

داریم :

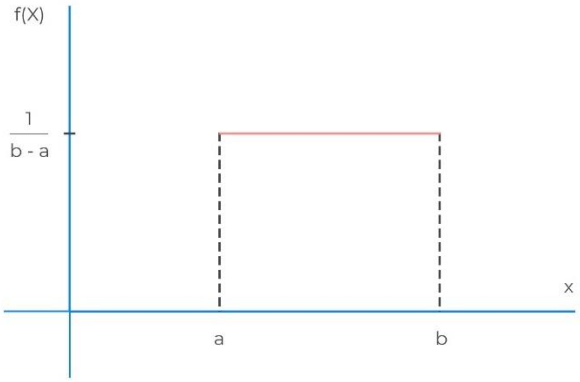
$$f_X(x) = c \qquad x \in [a, b]$$

اگر احتمال رخدادن نقاط در بازه مد نظر برابر با c باشد.

و یا داریم :

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & otherwise \end{cases}$$

نمودار PDF این توزیع به شکل زیر است :



Expected Value توزیع یکنواخت به شکل زیر است :

$$E(X) = \frac{a+b}{2}$$

واریانس توزیع یکنواخت نیز به شکل زیر است :

$$Var(X) = \frac{(b-a)^2}{12}$$

✓ **مثال :** با استفاده از پایتون توزیع یکنواختی با کران پایین ۲ و کران بالا ۳ رسم کنید.

برای پیاده سازی توزیع یکنواخت در پایتون داریم :

```
import numpy as np
import seaborn as sns

sns.distplot(np.random.uniform(low = 2, high = 3, size = (100)), hist = False)

plt.show()
```

❖ **توزیع نرمال (Normal Distribution) :**

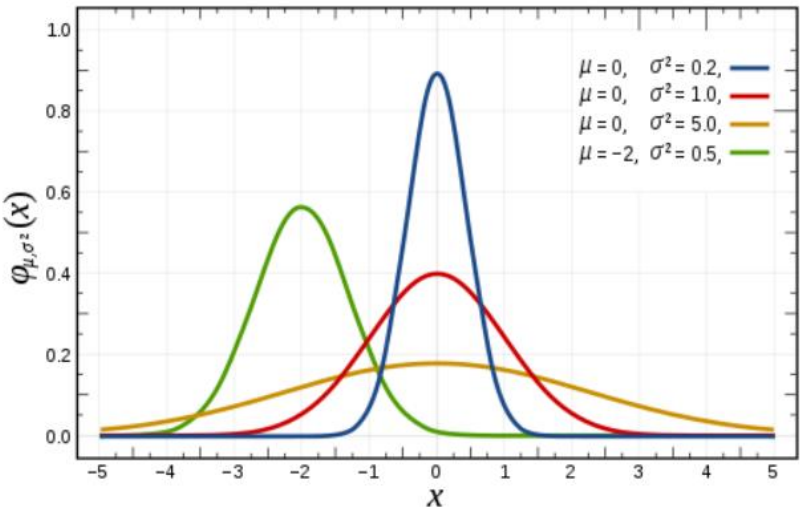
توزیع نرمال یا گاوسی مهم ترین توزیع احتمالاتی در آمار است زیرا این توزیع با بسیاری از پدیده های طبیعی تطابق دارد. برای مثال قد افراد ، فشار خون ، میزان IQ افراد و... از مواردی هستند که از این توزیع پیروی می کنند.

تابع چگالی احتمال توزیع نرمال به صورت زیر است :

$$f(x;\ \mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\,\sigma^2}}$$

که μ ، میانگین و σ^2 ، واریانس است که اگر از آن جذر بگیریم انحراف معیار بدست می آید.

نمودار PDF این توزیع به شکل زیر است :



Expected Value توزیع نرمال به شکل زیر است :

$$E(X) = \mu$$

واریانس توزیع نرمال نیز به شکل زیر است :

$$Var(X) = \sigma^2$$

توزیع نرمال با مقدار میانگین صفر و واریانس ۱ را توزیع نرمال استاندارد می نامند که آن را با Z نشان می دهند.

در توزیع نرمال استاندارد ، تابع توزیع تجمعی را φ می نامند.

یعنی داریم :

$$F(z) = \varphi(z)$$

برای تابع توزیع تجمعی توزیع نرمال فرمول خاصی وجود ندارد به همین دلیل برای بدست آوردن مقادیر از جدول مربوط به φ استفاده می کنند.

لینک دسترسی به جدول مربوط به φ :

https://www.mathsisfun.com/data/standard-normal-distribution-table.html

برای تبدیل توزیع نرمال به نرمال استاندارد داریم :

$$Z = \frac{X - \mu}{\sigma}$$

نکته ! در بعضی از پروژه های ماشین لرنینگ از نرمالایز کردن دیتاها استفاده می شود. یعنی دیتاهایی که از توزیع نرمال پیروی می کنند را طوری تغییر داد که از توزیع نرمال استاندارد پیروی کنند.

نکته ! $\varphi(-z) = 1 - \varphi(z)$

نکته ! توزیع نرمال استاندارد کاملا متقارن است.

✓ **مثال :** توزیع نرمالی با میانگین 4- و واریانس 9 داریم. مطلوب است محاسبه احتمالات زیر :

۱- $P(X < 0)$

۲- $P(-7 < X < -1)$

۱- داریم :

$$Z = \frac{X+4}{3} \rightarrow X = 3Z - 4$$

پس :

$$P(X < 0) = P(3Z - 4 < 0) = P(3Z < 4) = P\left(Z < \frac{4}{3}\right)$$
$$= P(Z < 1.333) \rightarrow \varphi(1.333) = 0.4082$$

۲- داریم :

$$P(-7 < X < -1) = P(-7 < 3Z - 4 < -1)$$

پس :

$$P(-7 < 3Z - 4) = P(-1 < Z)$$

و :

$$P(3Z - 4 < -1) = P(Z < 1)$$

با توجه به تقارن توزیع نرمال استاندارد داریم :

$$2\varphi(1) = 2 \times 0.3413 = 68.2$$

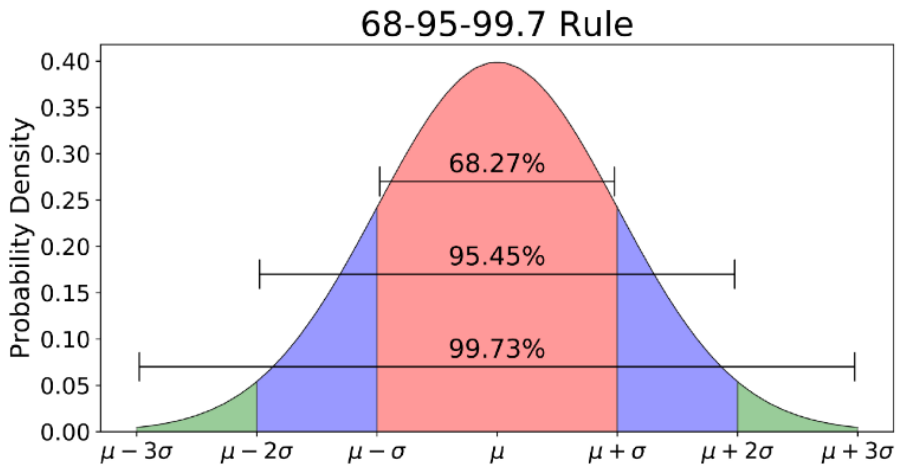
نکته ! به میزان جا به جایی قله توزیع نرمال به سمت چپ یا راست چولگی (Skewness) می گویند.

نکته ! به میزان کشیده شیدن قله توزیع نرمال به سمت بالا یا پایین کشیدگی (Kurtosis) می گویند.

برای پیاده سازی توزیع نرمال در پایتون داریم :

```
import numpy as np
import seaborn as sns

sns.distplot(np.random.normal(loc = -4, scale = 3, size = (100)), hist = False)

plt.show()
```

نکته ! در توزیع های نرمال ۶۸ درصد داده ها در بازه بین اختلاف میانگین و انحراف معیار و جمع این دو قرار دارد. همچنین ۹۵ درصد داده ها در بازه بین اختلاف میانگین و دو برابر انحراف معیار و جمع این دو قرار دارد و نیز ۹۹.۷ درصد داده ها در بازه بین اختلاف میانگین داده ها و سه برابر انحراف معیار و جمع این دو قرار دارد.



در ادامه با چند موضوع کاربردی در علم آمار و احتمال آشنا خواهیم شد.

نمونه و جامعه (Population and Sampling) :

تصور کنید می خواهیم از دانش آموزان کلاس سوم آزمایش خون بگیریم. تمامی دانش آموزان مدرسه به عنوان جامعه یا Population محسوب خواهند شد.

دانش آموزان کلاس سوم که مورد آزمایش قرار خواهند گرفت به عنوان جامعه هدف یا Target Population نامیده می شوند.

طبیعتا از دانش آموزان غایب نمی توان آزمایش گرفت پس تعدادی از دانش آموزان کلاس سوم در دسترس خواهند بود که به آن ها accessible می گویند.

و در نهایت به آزمایش های گرفته شده Sample می گویند.

❖ روش های نمونه گیری :

در نمونه گیری چندین روش وجود دارد اما باید به این نکته توجه داشت که حتما نمونه گیری باید از کل جامعه باشد برای مثال اگر بخواهیم از جمعیت ایران نمونه گیری کنیم نباید فقط از چند شهر بخصوص نمونه گیری کرد ، بلکه باید از تمامی شهرها متناسب با جمعیت آن شهر نمونه گیری کرد.

۱- Simple Random Sampling : در این روش نمونه گیری کاملا به صورت تصادفی است.

۲- Systematic Sampling : در این روش بر اساس یک قاعده خاص نمونه گیری رخ می دهد. برای مثال می خواهیم از دانش آموزان یک کلاس پرسش داشته باشیم. اگر بر اساس قاعده ای خاص مثلا دانش آموزانی که شماره کلاسی آن ها از رابطه $n = 2s + 3$ پیروی می کند ، را انتخاب کنیم ، به صورت سیستماتیک نمونه گیری کرده ایم.

۳- Stratified Sampling : در این روش ابتدا نمونه ها طبقه بندی می شوند و سپس از طبقه های مختلف نمونه گیری می شود. برای مثال اگر دانش آموزان یک کلاس را به دو طبقه بلند قد و کوتاه قد تقسیم بندی کنیم و سپس ۲ نفر از طبقه بلند قد و ۳ نفر از طبقه کوتاه قد برای نمونه گیری انتخاب کنیم ، به صورت Stratified نمونه گیری کرده ایم.

۴- Clustered Sampling : در این روش نمونه ها را خوشه بندی می کنیم و سپس از بین خوشه ها چند خوشه را برای نمونه گیری انتخاب می کنیم. برای مثال اگر هر دو نفر از دانش آموزان یک کلاس را بر اساس ترتیب لیست کلاسی در یک خوشه قرار دهیم و سپس از بین خوشه ها سه خوشه را انتخاب کنیم ، به روش Clustered نمونه گیری کرده ایم.

❖ **اندازه نمونه (Sample Size) :**

تعداد نمونه هایی که برای تحلیل آماری جامعه انتخاب می کنیم ، مهم است.
برای دانستن تعداد نمونه مورد نظر در یک تحلیل آماری می توان از فرمول استفاده کرد اما چون هدف محاسبات نیست و می خواهیم که از این مطالب به صورت کاربردی استفاده کنیم می توانیم با استفاده از سایت زیر به صورت آنلاین تعداد نمونه لازم برای تحلیل آماری را بدست بیاوریم.
https://www.calculator.net/sample-size-calculator.html

منظور از Confidence Level درصد اطمینان است.
منظور از Margin of Error بازه اطمینان است.

مرکزیت داده ها :

❖ **میانگین (Mean) :**

به جمع تمامی نمونه ها تقسیم بر تعداد نمونه ها میانگین می گویند.

❖ **میانه (Median) :**

میانه مقداری در وسط نمونه ها است برای مثال اگر ۱۰ داده داشته باشیم ، ۵ داده سمت چپ میانه و ۵ داده سمت راست میانه خواهد بود.

❖ **مد (Mode) :**

در بین داده ها ، نمونه ای که بیشترین تکرار را داشته باشد به عنوان مد شناخته می شود.

❖ میانگین اصلاح شده :

برای مثال اگر بخواهیم میانگین تعدادی داده را حساب کنیم که ۱۰ درصد داده ها ، به صورت داده های پرت هستند ، یعنی نسبت به مابقی داده ها مقادیر خیلی دورتری دارند. می توانیم ۱۰ درصد داده پرت را کنار گذاشته و میانگین مابقی داده ها که داده های پرت نیستند را حساب کنیم ، به این میانگین ، میانگین اصلاح شده می گویند.

با استفاده از پایتون می توان میانگین اصلاح شده را به صورت زیر محاسبه کرد :

```python
import numpy as np
from scipy import stats

arr = np.array([1, 4, 4.25, 4.25, 4.5, 4.75, 4.75, 5, 5.25, 10])

print(stats.trim_mean(arr, 0.1))
print(arr.mean())
```

ضریب تغییر (Coefficient of Variation) :

فرض کنید می خواهید میزان تغییر قیمت یک سهم را با سهم دیگر مقایسه کنید. ممکن است سهم اول یکی یکی قیمتش تغییر کند اما سهم دوم ده تا ده تا تغییر کند. برای آن که بتوان این دو سهم را با هم مقایسه کرد از مفهوم جدیدی به اسم ضریب تغییر استفاده می کنند.

داریم :

$$CoV = \frac{STD}{Mean}$$

توزیع فراوانی داده ها (Histogram) :

یکی دیگر از توزیع های مهم آماری ، توزیع هیستوگرام است. در این توزیع میزان فراوانی داده ها در مقادیر مختلف نشان داده می شود.

با استفاده از پایتون داریم :

```
import numpy as np
import matplotlib.pyplot as plt

x = np.random.randint(0, 50, size = 100)

plt.hist(x, bins = 30)
plt.show()
```

چارک (Quartile) :

اگر داده ها را مرتب کنیم ، سپس داده ها را به چهار قسمت تقسیم بندی کنیم (یعنی هر قسمت ۲۵ درصد داده ها را شامل می شود). داده انتهایی قسمت اول را با علامت Q_1 نشان می دهند. داده انتهایی قسمت دوم را با علامت Q_2 نشان می دهند. داده انتهایی قسمت سوم را با علامت Q_3 نشان می دهند.

نکته ! Q_2 همان میانه است زیرا نیمی از داده ها سمت چپ آن و نیمی از داده ها سمت راست آن قرار دارند.

برای محاسبه مقدار Q ها داریم :

$$Q_1 = \frac{n+1}{4}$$

و

$$Q_2 = \frac{(n+1)\times 2}{4}$$

و نیز

$$Q_3 = \frac{(n+1)\times 3}{4}$$

که در آن n تعداد داده ها است.

نکته ! چنانچه با استفاده از فرمول های بالا مقادیری بدست آمد که در بین داده ها نبود باید بین دو داده نزدیک آن مقدار میانگین گرفت.

❖ Inter Quartile Range (IQR) :

به فاصله بین چارکی IQR می گویند که از رابطه زیر بدست می آید.

$$IQR = Q_3 - Q_1$$

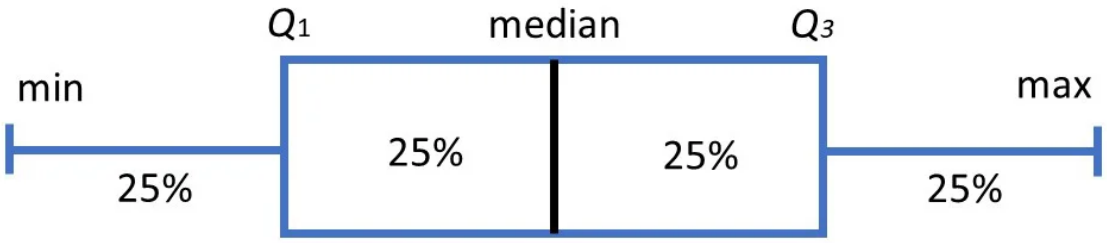
نکته ! می توان از مقایسه تفاضل چارک دوم از سوم و چارک اول از دوم ، در مورد چولگی اطلاعات کسب کرد.

در پایتون داریم :

```
import numpy as np
import pandas as pd

x = np.random.normal(loc = 4, scale = 2, size = 1000)
df = pd.DataFrame(x)
df.describe()
```

Boxplot :

نمودار جعبه ای به شکل زیر است :



چندک (Quantile) :

پیش تر با چارک آشنا شدیم که داده ها را به چهار قسمت تقسیم می کرد. چندک ها چندین مدل را شامل می شوند که یکی از آن ها چارک است. انواع دیگر چندک ، صدک یا Percentile است که داده ها را به صد قسمت تقسیم می کند ، دهک نیز نوع دیگری است که داده ها را به ده قسمت تقسیم می کند.

نکته ! می توان برای بررسی آن که آیا توزیع از توزیع نرمال استاندارد پیروی می کند یا خیر از پایتون استفاده کرد. برای مثال :

```
import numpy as np
import seaborn as sns

n = np.random.normal(loc = 4, scale = 3, size = 1000)
u = np.random.uniform(size = 1000)

sns.distplot(n, kde = True)
```

داریم :

```
from statsmodels.api import qqplot

qqplot(n, line = "45")
qqplot(u)
```

بدلیل آن که توزیع نرمال به صورت استاندارد نبود پس منطبق بر خط ۴۵ درجه نیست و همچنین توزیع یکنواخت نیز به شکل خط راست نیست.

همچنین می توان برای بررسی آن که آیا توزیع از توزیع دیگری پیروی می کند یا خیر از پایتون استفاده کرد. برای مثال :

```
from statsmodels.api import qqplot_2samples

qqplot_2samples(n, u)
```

کواریانس و همبستگی (Covariance and Correlation) :

به ارتباط بین جفت متغیرهای تصادفی زیر بیاندیشید :

۱- نمره‌ی ریاضی و نمره‌ی علوم یک سری دانش‌آموز

۲- قد انسان‌ها و اندازه‌ی شعاع دور سرشان

۳- جمعیت یک شهر و میزان آلودگی آن

۴- ارزش سهام شرکت اپل و تعداد سیب‌های فروش رفته

۵- میزان شارژ باتری لپتاپ و تعداد اپلیکیشن‌های در حال اجرا

۶- گل های زده شده توسط یک بازیکن فوتبال در فصل‌های مختلف و سن او در آن فصل‌ها

۷- دمای هوا و تعداد بستنی های فروش رفته

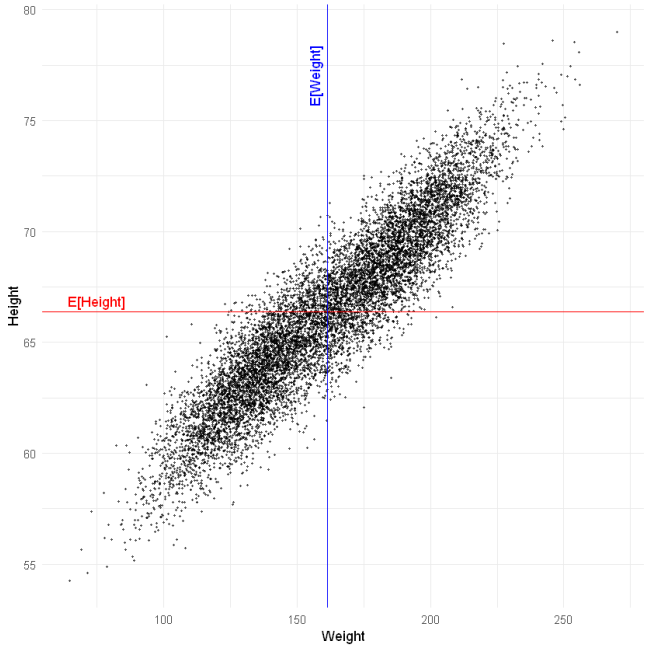
همانطور که احتمالا به صورت شهودی متوجه شدید ، بعضی از این موارد ارتباط مثبت ، بعضی منفی و بعضی دیگر هیچ ارتباطی با هم ندارند. correlation و covariance معیاری ریاضی از این ارتباط است.

❖ **کواریانس :**

کواریانس میزان تغییر دو متغیر با همدیگر نسبت به میانگین شان است. داریم :

$$Cov(X,Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

برای درک بهتر ارتباط بین قد و وزن در نمودار زیر نشان داده شده است.



نکته ! صعودی یا نزولی بودن نمودار بالا نشان دهنده کواریانس مثبت یا منفی خواهد بود.

نکته ! کواریانس صفر به معنی آن است که دو متغیر نسبت به هم مستقل هستند. کواریانس مثبت به معنی آن است که دو متغیر ارتباط مستقیم و منفی به معنی آن است که دو متغیر ارتباط معکوس نسبت به هم دارند.

❖ **همبستگی :**

مقدار کواریانس اطلاعات خاصی به ما نمی دهد و فقط علامت آن است که نشان دهنده ارتباط بین دو متغیر است. واحد کواریانس ، واحد متغیر X در واحد متغیر Y است. بنابراین اگر واحد متغیر ها را تغییر دهیم ، عدد کواریانس تغییر خواهد کرد.

برای مثال اگر بخواهیم ارتباط بین وزن و قد را بررسی کنیم ، اگر یکبار قد را برحسب سانتی متر در نظر بگیریم و بار دوم برحسب متر ، کواریانس تفاوت صد برابری خواهد داشت در صورتی که ارتباط بین قد و وزن تغییری نکرده است.

برای رفع این مشکل همبستگی را تعریف می کنند.

داریم :

$$Cov\left(\frac{X-\mu_X}{\sigma_X},\frac{Y-\mu_Y}{\sigma_Y}\right)=E\left[\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right]$$
$$=\frac{1}{\sigma_X\sigma_Y}E[(X-\mu_X)(Y-\mu_Y)]=\frac{Cov(X,Y)}{\sigma_X\sigma_Y}$$

در بالا متغیر های تصادفی استاندارد شده اند.

خواص همبستگی به شرح زیر است :

۱- همبستگی مقیاس ندارد. یعنی اگر عددی در متغیر تصادفی ضرب شود ، مقدار همبستگی تغییر نخواهد کرد.
۲- همبستگی دو متغیر تصادفی مستقل از هم برابر صفر خواهد بود.
۳- مقدار همبستگی بین ۱- تا ۱ خواهد بود. هر چه این مقدار به ۱- یا ۱ نزدیک تر باشد نشان دهنده وابستگی بیشتر متغیر خواهد بود (به صورت معکوس یا مستقیم).

نکته ! همبستگی به معنی علیت نیست. برای مثال فرض کنید یکی از متغیر های تصادفی میزان فروش بستنی است و متغیر دوم تعداد حمله کوسه در هر ماه باشد. اگر همبستگی این دو متغیر مثبت باشد دلیل بر آن نیست که فروش بیشتر بستنی باعث حمله بیشتر کوسه ها می شود. احتمالا در ماه های گرم سال بستنی فروش بیشتری دارد ، علاوه بر این در این ماه ها افراد بیشتری در کنار ساحل هستند که باعث حمله بیشتر کوسه ها می شود.

با استفاده از پایتون می توان مقدار همبستگی را بدست آورد :

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

data = pd.read_csv("C:/Users/Asus/Desktop/AI01/company_sales_data.csv")

data.corr()
```

```
plt.figure(figsize = (9, 9))
sns.heatmap(data.corr(), annot = True)
```