بسمه تعالی

# موضوع : مبانی آمار و احتمال برای ورود به علم داده و تحلیل داده

## بخش چهاردهم: انواع تابع توزیع گسسته

- برنولی
- دوجمله‌ای
- هندسی
- پاسکال(دوجمله ای منفی)
- فوق هندسی
- پوآسون

در این پارت از اموزش به بررسی **انواع تابع توزیع گسسته** می پردازیم .

در اموزش های قبلی راجب **توابع احتمال گسسته و پیوسته و محاسبه میانگین و واریانس براساس آن** صحبت کردیم.

تا اینجا گفتیم که یک متغیر تصادفی گسسته می‌تواند مقادیر مشخص و قابل شمارشی داشته باشد و برای هر مقدار نیز یک احتمال تعریف می‌کنیم.

اما در مسائل مختلف، شکل این توزیع احتمال متفاوت است و داده ها از خود رفتار متفاوتی نشان می دهند در این صورت شاخصه های آماری متفاوتی خواهند داشت .

انواع توزیع های احتمالی رایج در این حوزه عبارتند از :

۱- توزیع برنولی(Bernoulli)
۲- توزیع دوجمله‌ای(Binomial)
۳- توزیع هندسی(Geometric)
۴- توزیع دوجمله‌ای منفی(Negative Binomial)
۵- توزیع پواسون(Poisson)
۶- توزیع فوق‌هندسی (Hypergeometric)

گفتیم برای بررسی جامعه آماری به بررسی نتیجه آزمایش تصادفی می پردازیم. حال اگر پیشامدی که برای آن آزمایش تصادفی بررسی می شود دو حالت موفقیت یا شکست داشته باشد می توان گفت متغیر تصادفی (مرتبط با پیشامد) دارای **توزیع برنولی** است:

**مثلاً:**

- قبولی / مردودی
- سالم / خراب
- خرید / عدم خرید

- پاسخ مثبت / پاسخ منفی
- خرابی / عدم خرابی

$P(X=1)=p\ ,\ P(X=0)=1-p$ ⇨ $P(X=x)=p^x(1-p)^{1-x}, x=0,1$

$$E(X)=p, \qquad V(X)=p(1-p)$$

حال اگر آن آزمایش تصادفی با نتیجه پیروزی و شکست را به تعداد n بار تکرار کنیم و در آن تکرار نتیجه پیروزی را بررسی کنیم . البته باید دقت داشت که تمامی نمونه گیری و بررسی های ما با جایگذاری است یعنی نمونه ها و نتایج مستقل از هم هستند. در این صورت می توان گفت متغیر تصادفی ما **توزیع دوجمله** دارد.

$$P(X=x)=\binom{n}{x}p^x(1-p)^{n-x}, x=0,1,2,\dots,n$$

$$E(X)=np, \qquad V(X)=np(1-p)$$

**مثال:** در یک فرآیند تولید ، احتمال معیوب بود قطعه ۰.۲ اگر یک نمونه ۵۰ تایی گرفته شود. احتمال بیرون امدن تعدادی مشخص از قطعه معیوب دارای توزیع دوجمله ای است.

حال اگر همان آزمایش برنولی را آنقدر تکرار کنیم تا به اولین موفقیت برسیم. می توان گفت که متغیر تصادفی مورد نظر از **توزیع هندسی** پیروی می کند:

$$P(X=i)=\ (1-p)^{i-1}p=q^{i-1}p \qquad i=1,2,3,\dots$$

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

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

**مثلاً:** احتمال اینکه یک مشتری پیشنهاد بانک را بپذیرد ۰،۲ است.

می‌خواهیم بدانیم احتمال اینکه **اولین مشتری پذیرنده، مشتری پنجم باشد** چقدر است.

این مسئله یک مثال از توزیع هندسی است.

در اینجا می گوییم اگر آن ازمایش برنولی اگر تا رسیدن به R امین موفقیت ادامه دهیم . در این صورت متغیر تصادفی مد نظر از **توزیع دوجمله ای منفی** پیروی می کند. در این صورت :

$$P(X=i)=\binom{i-1}{r-1}p^r(1-p)^{i-r} \qquad i=r,r+1,r+2,\dots$$

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

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

**مثال :** اگر یک تاس سالم را به طور مرتب پرتاپ کنیم ، احتمال این که: دومین نتیجه ۶ در پنجمین پرتاب حاصل شود دارای توزیع دوجمله ای منفی است.

حال اگر آزمایش دوجمله را به این فرض انجام دهیم که نمونه بررسی شده را بدون جایگذاری انجام دهیم یعنی : فرض کنید از جعبه ای که دارای m قطعه معیوب و N-m قطعه سالم است، n قطعه را به تصادف و بدون جایگذاری انتخاب کنیم. اگر X تعداد قطعات معیوب خارج شده باشد ،در این صورت متغیر تصادفی ما از **توزیع فوق هندسی پیروی** می کند. آنگاه داریم :

$$P(X=i)=\frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}} \qquad \max(0,n+m-N)\le i\le \min(m,n)$$

$$E(X)=np$$

$$Var\ (X)=np(1-p)\frac{N-n}{N-1}$$

در مورد مقادیری که i می تواند اختیار کند توجه داشته باشید که i نمی تواند از m و n بیشتر باشد.

**مثلاً:**

یک جعبه شامل:۸ قطعه سالم /۲ قطعه معیوب است.از جعبه ۳ قطعه را **بدون جایگذاری** انتخاب می‌کنیم.می‌خواهیم احتمال اینکه دقیقاً ۱ قطعه معیوب انتخاب شود را محاسبه کنیم. اینجا توزیع فوق‌هندسی مناسب است.

و در اخر می توان گفت اگر آزمایش دوجمله ای را با تکرار زیاد انجام دهیم براساس نتایج می توان نرخ رخ داد مورد نظر را حساب کرد. در این صورت برای پیشامد مد نظر با نرخ اتفاق مشخص می توان گفت این متغیر از **توزیع پوآسون پیروی** می کند:

$$f(x)=\frac{e^{-\lambda}\lambda^{x}}{x!},\lambda>0$$

$$E(X)=\lambda \qquad V(X)=\lambda$$

اگر در یک توزیع دوجمله ای n بزرگ و p کوچک باشد بطوری که p.n عددی معقول باشد ، این توزیع با توزیع پواسن با پارامتر $\lambda=n.p$ تقریب زد.

**مثلاً:**

- تعداد تماس‌های ورودی در یک ساعت
- تعداد خرابی ماشین‌آلات در یک ماه
- تعداد مشتریان واردشده به بانک در ۱۰ دقیقه
- تعداد تصادفات در یک روز
- تعداد خطاهای ثبت‌شده در یک صفحه
- تعداد درخواست‌های دریافت‌شده توسط یک سرور در یک دقیقه

در این مسائل معمولاً با تعداد وقوع یک رویداد سروکار داریم.