Session 06

Machine Learning | Zahra Amini

Telegram: @zahraamini_ai & Instagram:@zahraamini_ai & LinkedIn: @zahraamini-ai

https://zil.ink/zahraamini

1. تعریف توزیع (Distribution)

توزیع در آمار به نحوه پراکندگی یا چگونگی انتشار مقادیر یک مجموعه داده اشاره دارد. به عبارت دیگر، توزیع نشان می‌دهد که چگونه مقادیر داده‌ها در طول یک محدوده تغییر می‌کنند و چند بار هر مقدار خاص رخ می‌دهد. این مفهوم به ما کمک می‌کند که بدانیم داده‌ها چگونه در یک مجموعه پراکنده شده‌اند.

ویژگی‌های توزیع:

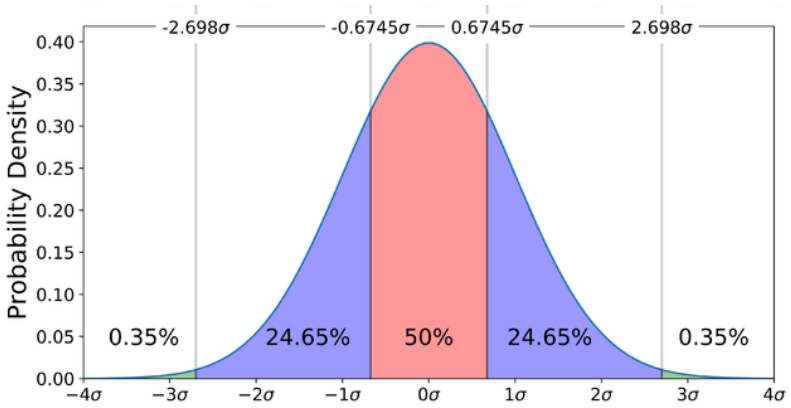
- **میانگین (Mean)**: میانگین یا مقدار متوسط داده‌ها.
- **میانه (Median)**: مقداری که نصف داده‌ها کمتر از آن و نصف بیشتر از آن است.
- **واریانس (Variance)**: اندازه‌گیری پراکندگی داده‌ها حول میانگین.
- **چولگی (Skewness)**: میزانی که توزیع به سمت چپ یا راست کشیده شده است.
- **کشیدگی (Kurtosis)**: مقدار پخ بودن یا کشیدگی توزیع.

2. توزیع نرمال (Normal Distribution)

توزیع نرمال یا **توزیع گاوسی** یکی از پرکاربردترین توزیع‌ها در آمار است. این توزیع به خاطر شکل **زنگوله‌ای** خود که تقارن کامل دارد، بسیار مشهور است. توزیع نرمال بسیاری از پدیده‌های طبیعی مانند قد، وزن، و نمرات آزمون‌ها را مدل می‌کند.

ویژگی‌های توزیع نرمال:

- **تقارن**: توزیع نرمال به صورت کاملاً متقارن است، یعنی نیمی از داده‌ها در سمت چپ و نیمی در سمت راست میانگین قرار دارند.
- **میانگین، میانه و نما یکسان هستند**: در توزیع نرمال، میانگین (Mean)، میانه (Median) و نما (Mode) همگی برابرند و در مرکز توزیع قرار دارند.
- **پهنای منحنی**: پهنای منحنی توزیع نرمال با استفاده از واریانس (یا انحراف استاندارد) تعیین می‌شود. انحراف استاندارد میزان پراکندگی داده‌ها در اطراف میانگین را نشان می‌دهد.
- **قانون 68-95-99.7**: این قانون بیان می‌کند که در توزیع نرمال:
  - 68٪ داده‌ها در فاصله یک انحراف استاندارد از میانگین قرار می‌گیرند.
  - 95٪ داده‌ها در فاصله دو انحراف استاندارد از میانگین قرار می‌گیرند.
  - 99.7٪ داده‌ها در فاصله سه انحراف استاندارد از میانگین قرار می‌گیرند.



Formula for the Probability Density Function (PDF) of a Normal Distribution:

The probability density function (PDF) of a normal distribution is given by the following equation:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Where:

- x is the value for which you want to calculate the probability density.
- μ is the mean of the distribution.
- σ is the standard deviation.
- $\exp$ is the exponential function.

3. توزیع نرمال استاندارد

توزیع نرمال استاندارد یک حالت خاص از توزیع نرمال است که میانگین آن صفر و انحراف استاندارد آن یک است. به این معنا که تمام داده‌ها به گونه‌ای نرمال‌سازی شده‌اند که $\mu = 0$ و $\sigma = 1$ باشد. داده‌ها در این حالت به صورت واحدهای استاندارد (z-scores) بیان می‌شوند.

محاسبه z-score:

مقدار z (یا امتیاز استاندارد) برای یک مقدار خاص x به صورت زیر محاسبه می‌شود:

$$\frac{x - \mu}{\sigma} = z$$

این فرمول نشان می‌دهد که مقدار x چند انحراف استاندارد از میانگین فاصله دارد.

4. کاربرد توزیع نرمال

- **آمار استنباطی**: در بسیاری از آزمون‌های آماری فرض می‌شود که داده‌ها از یک توزیع نرمال پیروی می‌کنند. این فرضیه به خصوص در آزمون‌های t و تحلیل واریانس (ANOVA) اهمیت دارد.
- **مدل‌سازی طبیعی**: بسیاری از پدیده‌های طبیعی (مانند قد، وزن، نمرات آزمون‌ها و غیره) به طور تقریبی از توزیع نرمال پیروی می‌کنند.
- **تخمین‌ها**: در مواردی که نمی‌دانیم توزیع داده‌ها چگونه است، فرض نرمال بودن داده‌ها می‌تواند تقریب خوبی برای تحلیل و استنباط باشد.

1. تعریف همبستگی

همبستگی یک معیار آماری است که قدرت و جهت رابطه خطی بین دو متغیر را نشان می‌دهد. این مفهوم به ما کمک می‌کند که بفهمیم آیا تغییرات یک متغیر با تغییرات متغیر دیگر همسو است یا نه.

- اگر **همبستگی مثبت** باشد، به این معناست که با افزایش مقدار یکی از متغیرها، مقدار متغیر دیگر نیز افزایش می‌یابد.
- اگر **همبستگی منفی** باشد، نشان می‌دهد که با افزایش مقدار یکی از متغیرها، مقدار متغیر دیگر کاهش می‌یابد.
- اگر **همبستگی صفر** باشد، یعنی هیچ رابطه‌ای بین دو متغیر وجود ندارد.

2. ضریب همبستگی

برای اندازه‌گیری همبستگی، از ضریب همبستگی پیرسون (Pearson Correlation Coefficient) استفاده می‌شود که با r نشان داده می‌شود. مقدار این ضریب بین -1 و $+1$ قرار دارد:

- $r = +1$: همبستگی کامل مثبت (رابطه خطی مثبت کامل).
- $r = -1$: همبستگی کامل منفی (رابطه خطی منفی کامل).
- $r = 0$: هیچ همبستگی (بدون رابطه خطی بین متغیرها).

3. فرمول ضریب همبستگی پیرسون

فرمول محاسبه ضریب همبستگی بین دو متغیر X و Y به این شکل است:

$$r = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 \sum (Y_i - \overline{Y})^2}}$$

- X_i and Y_i: Observed values of variables X and Y.
- $\overline{X}$ and $\overline{Y}$: Means of X and Y.

X	y
1	2
2	3
3	5
4	7
5	8

Step 1: Calculate the means

$$\overline{X} = \frac{1+2+3+4+5}{5} = 3$$

$$\overline{Y} = \frac{2+3+5+7+8}{5} = 5$$

Step 2: Calculate $(X_i - \overline{X})$ and $(Y_i - \overline{Y})$

X	Y	$X_i - \overline{X}$	$Y_i - \overline{Y}$	$(X_i - \overline{X})(Y_i - \overline{Y})$	$(X_i - \overline{X})^2$	$(Y_i - \overline{Y})^2$
1	2	-2	-3	6	4	9
2	3	-1	-2	2	1	4
3	5	0	0	0	0	0
4	7	1	2	2	1	4
5	8	2	3	6	4	9

Step 3: Calculate the correlation coefficient

$$r = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2 \sum (Y_i - \overline{Y})^2}}$$

Summing the values:

$$\sum (X_i - \overline{X})(Y_i - \overline{Y}) = 6 + 2 + 0 + 2 + 6 = 16$$

$$\sum (X_i - \overline{X})^2 = 4 + 1 + 0 + 1 + 4 = 10$$

$$\sum (Y_i - \overline{Y})^2 = 9 + 4 + 0 + 4 + 9 = 26$$

Now, calculate r:

$$r = \frac{16}{\sqrt{10 \times 26}} = \frac{16}{\sqrt{260}} = \frac{16}{16.12} \approx 0.99$$

ضریب همبستگی r = 0.99 نشان می‌دهد که بین متغیرهای X, y یک رابطه خطی قوی و مثبت وجود دارد. یعنی با افزایش X، مقدار y نیز تقریباً به همان نسبت افزایش می‌یابد

- **همبستگی علیت را نشان نمی‌دهد**: همبستگی فقط به رابطه بین دو متغیر اشاره دارد و نمی‌گوید که یکی باعث دیگری می‌شود.
- **همبستگی‌های غیرخطی**: ضریب همبستگی پیرسون تنها برای روابط خطی مناسب است. برای روابط غیرخطی از روش‌های دیگر استفاده می‌شود.

ویژگی	Correlation (همبستگی)	Regression (رگرسیون)
هدف	بررسی قدرت و جهت رابطه بین دو متغیر	پیش‌بینی یا توضیح رابطه بین متغیر وابسته و متغیر مستقل
رابطه علیت	علیت را نشان نمی‌دهد	فرض می‌کند که متغیر مستقل علت تغییرات در متغیر وابسته است
خروجی	ضریب همبستگی (r)	معادله رگرسیون $(bX + a = Y)$
نوع رابطه	فقط قدرت رابطه را اندازه‌گیری می‌کند	رابطه علت و معلولی را مدل‌سازی می‌کند
کاربرد	تعیین رابطه بین متغیرها	پیش‌بینی و توضیح متغیر وابسته با استفاده از متغیرهای مستقل

ویژگی	نویز (Noise)	داده پرت (Outlier)
ماهیت	انحرافات کوچک و تصادفی که در تمام داده‌ها پراکنده‌اند.	نقاط دور از بقیه داده‌ها که به شدت از روند کلی جدا هستند.
علت	خطاهای اندازه‌گیری، نوسانات محیطی، یا عوامل تصادفی.	رویدادهای نادر، خطاهای جمع‌آوری داده یا پدیده‌های غیرمعمول.
اثر بر تحلیل	به صورت کلی می‌تواند روی دقت نتایج تاثیر بگذارد، اما نه به شکل عمده.	می‌تواند نتایج تحلیل را به شدت تغییر دهد و گاهی منجر به نتایج گمراه‌کننده شود.
نحوه مدیریت	کاهش یا فیلتر کردن از طریق روش‌های آماری یا فیلترهای نرم‌افزاری.	شناسایی و تصمیم‌گیری برای حذف یا نگهداری بر اساس تحلیل خاص.
توزیع	به طور تصادفی در سراسر مجموعه داده پخش می‌شود.	به طور خاص و قابل مشاهده در نقاط خاص از داده ظاهر می‌شود.