Week 2: Describing Probability Distributions and Probability Distributions with Multiple Variables:

Measures of Center:
- Mean
- Mode
- Median

* Expected Value: It can mean many things: likely
"E[X]"

→ Mean (Average) = $\frac{\text{summation weighted values}}{\text{Number of weighted values}}$
"μ" Weighted

Discrete: $E[X] = \sum_x x\, p_X(x)$ (PMF)

Continuous: $E[X] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$ (PDF)

Mean doesn't work when data contains outliers.

→ Median = middle value of ordered dataset.
If the number of values is even, then median = average of the two data in the middle

→ Mode = most frequent data in the dataset
Sometimes, we can have a multi-modal distribution; where many data have the same frequency that appears to be the highest.

→ Expected Value of a Function rather than random variables:

- $E[X] = x_1 p(x_1) + x_2 p(x_2) + \ldots + x_n p(x_n)$
- $E[g(X)] = g(x_1) p(x_1) + g(x_2) p(x_2) + \ldots + g(x_n) p(x_n)$
- $E[aX+b] = a\,E[X] + b$
- $E[X_1 + X_2] = E[X_1] + E[X_2]$

Measures of Spread:
- Variance
- Standard deviation

Variance:

Since deviation $(x - E[X])$ can give wronge indications because of negative and positive signs. To measure variance, we get the average of squared deviations of all values $(x - E[X])^2$.

average value or expected value

$$\text{Variance} = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

Same:
$$\text{Var}(X) = E[(X - \bar{X})^2]$$
$$\text{Var}(aX + b) = a^2 \text{Var}(X)$$

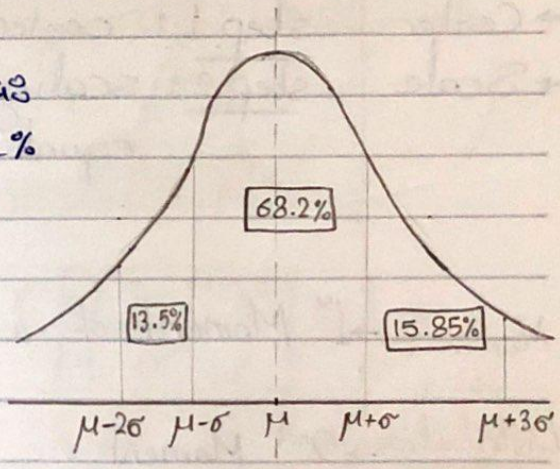
Standard Deviation:

To cover a drawback of the variance, which is having the units squared (not practical), we use standard deviation as the square root of the variance.

$$\text{std}(X) = \sqrt{\text{Var}(X)} = \sigma(X)$$

Gaussian Distribution for 2 variables looks like a bell.

For normally distributed datas:
- Between $[\mu - \sigma, \mu + \sigma]$ lies 68.2% of the data.
- Between $[\mu - 2\sigma, \mu + 2\sigma]$ lies 95% of the data.
- Between $[\mu - 3\sigma, \mu + 3\sigma]$ lies 99.7% of the data.



Normal or Gaussian Distribution $\mathcal{N}(\mu, \sigma)$

Total Response Time of a Computer System

$$R = T + L$$

- T: Processing Time
- L: Network (connection) Latency

we suppose that: $T \sim \mathcal{N}(10, 2^2)$ independent $L \sim \mathcal{N}(5, 1^2)$

$$\mu_R = E[R] = E[T] + E[L] = \mu_T + \mu_L = 10 + 5 = \underline{15}$$

Therefore, the summation of two gaussians is still a gaussian.

$$\sigma_R^2 = Var(R) = Var(T+L) = Var(T) + Var(L)$$
$$\sigma_R^2 = \sigma_T^2 + \sigma_L^2 = 4 + 1$$

$$R \sim \mathcal{N}(15, 5)$$

→ In general: $W = aX + bY$

independent $\begin{cases} X \sim \mathcal{N}(\mu_X, \sigma_X^2) \\ Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2) \end{cases}$

then $W \sim \mathcal{N}\left(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2\right)$

⊗ Standardize Distribution:

→ Center step 1: centering the dataset about zero. $(X \rightarrow X-\mu)$

→ Scale step 2: scaling the dataset to standard deviation equals to 1. $\left(X-\mu \rightarrow \frac{X-\mu}{\sigma}\right)$

Moments of a Distribution

→ 1st Moment of a random variable distribution is:
$$E[X] = p_1 x_1 + p_2 x_2 + \cdots + p_n x_n$$

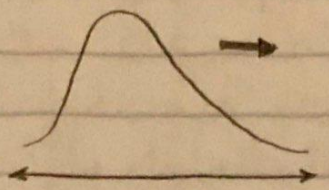
→ 2nd Moment:
$$E[X^2] = p_1 x_1^2 + p_2 x_2^2 + \cdots + p_n x_n^2$$

→ K^{th} Moment:
$$E[X^k] = p_1 x_1^k + p_2 x_2^k + \cdots + p_n x_n^k$$

skewness: is detected using the third moment of standardized event or distribution

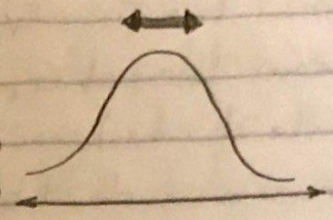
$$E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \text{Skewness}$$

- Positively Skewed
- Right Skewed



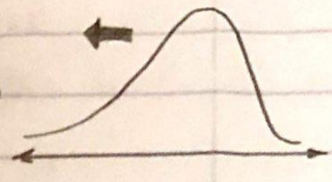
$$E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] > 0$$

- Not Skewed
- zero Skewed



$$E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = 0$$

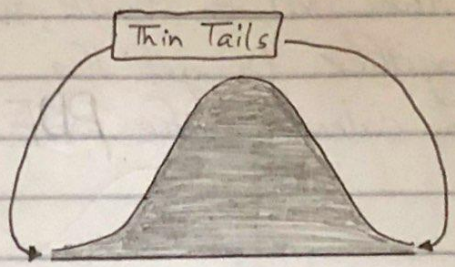
- Negatively Skewed
- Left Skewed



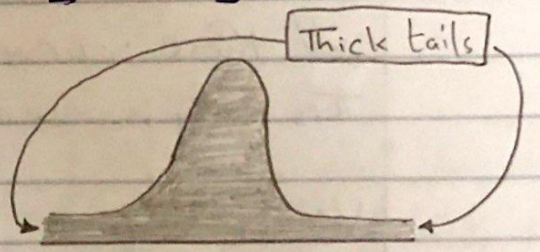
$$E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] < 0$$

Sometimes, there are different curves, which have similar variance, standard deviation, expected value and even skewness. Therefore, we use Kurtosis; it's the fourth moment of expected value of standardized distribution.

$$\text{Kurtosis} = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$$



$$E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \text{small}$$



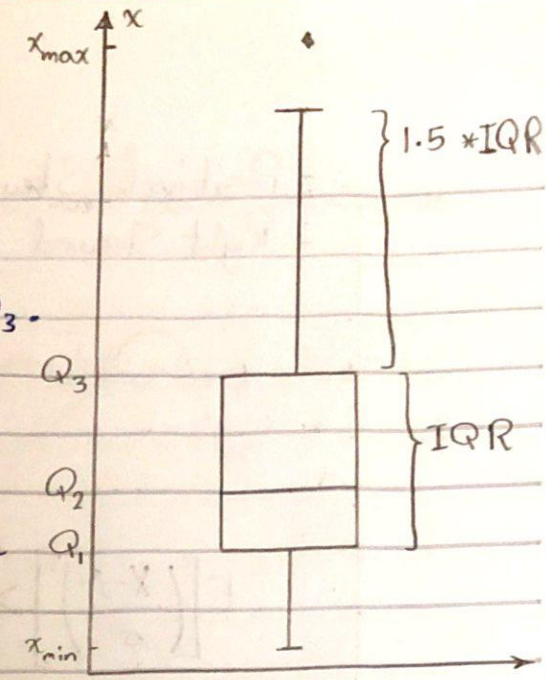
$$E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \text{Large}$$

The k^{th} quantile ($q_{k/100}$) is the value that leaves k% of the values to the left and (100−k)% to the right.

- 25% Quantile → 1st quartile Q_1
- 50% Quantile → 2nd quartile (median) Q_2
- 75% Quantile → 3rd quartile Q_3

$$P(X \leq q_{(K/100)}) = \frac{k}{100}$$

Box Plot:

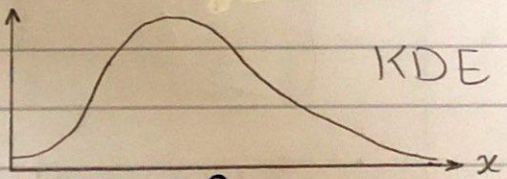


- steps: 1- draw box from Q_1 to Q_3.
  2- draw line at Q_2.
  3- Extend whiskers to maximum of 1.5 times the interquartile range (IQR*1.5).
  4- Represent the outliers with rhombus dots.

- insights: 1- Skewness 2- Presence of outliers
  3- Spread (Dispersion) Analysis

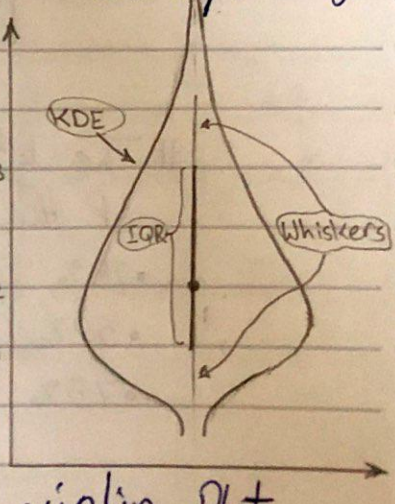
Kernel Density Estimation:

It is done by representing values with gaussian curves centered on the value itself, then add all these curves together to get the KDE graph. This way is used to calculate the PDF of datasets.



Violin Plot:

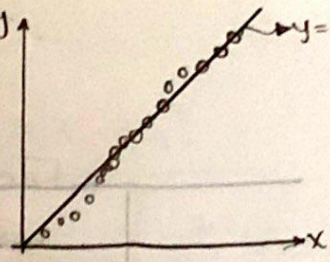
It is simply putting KDE with a box plot together



violin Plot

Many model assumes that data follows gaussian (normal) distribution:
- → Linear Regression
- → Logistic Regression
- → Gaussian Naive Bayes
- → Others

, and so does some data science tests.



Quantile-Quantile (QQ) Plot:

It is a graph of Theoretical - Sample quantiles as scatter plot with a line ($y=x$). The more the distribution of points is closer to the line the more normally distributed the data is.

Join Distribution (Discrete):

It is easy to plot a feature, but when we have two features to plot on the same graph it gets harder.

Example:

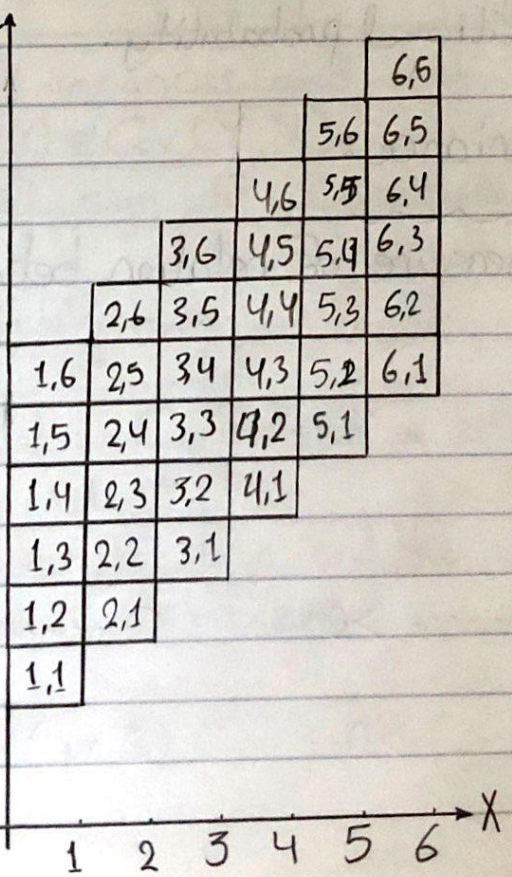
Get the joint distribution graph of features X and Y for event E; where:

E → 2 6-sided dices were rolled

X → Number on dice 1

Y → Sum of two numbers appeared

Solution:



- each box has probability of $\frac{1}{36}$
- If we replace each box with the value of its probability, we will get the joint distribution of X & Y.

For a continuous dataset:
→ we use a scatter plot
→ Expected Value ($E[X], E[Y]$) → coordinate
→ We can calculate the variances of each feature on it's own: $Var(X) = E[X^2] - E[X]^2$

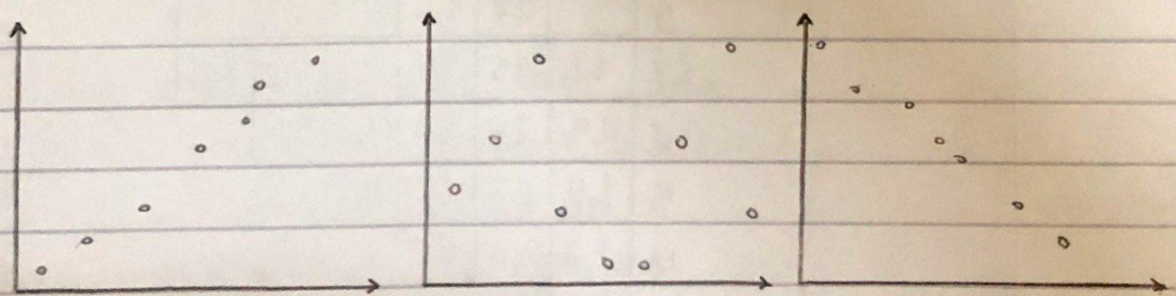
Discrete Joint Distribution	Continuous Joint Distribution
- Assign Probability to individual outcomes. - Countable possible outcomes.	- Assign Probability to ranges or intervals. - Infinite number of possible outcomes.

Marginal distribution is getting the probability of values of one feature (variable) from multi-variable graph or joint distribution.

Marginal distribution is useful in calculating conditional probability.

⊛ Covariance:

A measure of relation between two variables:



Covariance >0 | Covariance ≈ 0 | Covariance <0

in standardized Graph → $(\Sigma xy) > 0$ | $(\Sigma xy) \approx 0$ | $(\Sigma xy) < 0$

For equal Probabilities ⇒ $\underline{\text{Covariance}(X,Y)} = \frac{\sum_{i=1}^{n} (x_i - \mu_x)\cdot(y_i - \mu_y)}{n}$

$\underline{\text{Cov}(X,Y)}$

$Y = y$ Missused!!

General ⇒ $\text{Cov}(X,Y) = \sum_{i=1}^{n} P_{XY}(x_i, Y_i)\cdot(x_i - \mu_x)\cdot(y_i - \mu_y)$

$= E[XY] - E[X]\,[Y]$

→ In case, we have more than 2 variables: A, B, C, D and E for example. We use something called "Covariance Matrix"; where:

	A	B	C	D	E
A	Var(A)	Cov(B,A)	Cov(C,A)	Cov(D,A)	Cov(E,A)
B	Cov(A,B)	Var(B)	Cov(C,B)	Cov(D,B)	Cov(E,B)
C	Cov(A,C)	Cov(B,C)	Var(C)	Cov(D,C)	Cov(E,C)
D	Cov(A,D)	Cov(B,D)	Cov(C,D)	Var(D)	Cov(E,D)
E	Cov(A,E)	Cov(B,E)	Cov(C,E)	Cov(D,E)	Var(E)

Diagonal of variances and rest of covariances, hence, $\text{Cov}(X,Y) \equiv \text{Cov}(Y,X)$.

We use the sign of Covariance only.

⊛ Correlation Coefficient

Standardized Covariance

$$\text{Correlation Coefficient} = \frac{\text{Cov}(X,Y)}{\sigma_x \cdot \sigma_y}$$

range: $(-1, 1)$

we can use the sign and the value of the correlation coefficient.

Multivariable Gaussian Distribution:

For two variables X & Y that are independent and normally distributed, the graph would look like a bell shape. The equation would be:

$$f_{XY}(x,y) = f_X(x)\, f_Y(y)$$

$$f_{XY}(x,y) = \frac{1}{2\pi \sigma_X \sigma_Y} e^{-0.5\left(\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right)}$$

$$f_{XY}(x,y) = \frac{1}{2\pi \det\Sigma^{1/2}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix}x\\y\end{bmatrix} - M\right)^T \Sigma^{-1} \left(\begin{bmatrix}x\\y\end{bmatrix} - M\right)\right)$$

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

- $\Sigma \rightarrow$ covariance matrix
- $\mathrm{Cov}(X,Y) = 0$ if X & Y are independent
- $\sigma = \sqrt{\text{Variance}}$