UNIT II

Normal Distribution and its types

1. Introduction

1. The normal distribution, also called the Gaussian distribution, is the most widely used probability distribution in statistics.
2. It is a continuous distribution that is symmetric about the mean.
3. The graph of the normal distribution has a bell-shaped curve.

Univariate Multivariate Normal Distribution

- Gaussian distribution is a synonym for normal distribution.
- S is a set of random values whose probability distribution looks like a bell-shaped curve.
- If a probability distribution plot forms a bell-shaped curve like above and the mean, median, and mode of the sample are the same, that distribution is called normal distribution or Gaussian distribution.

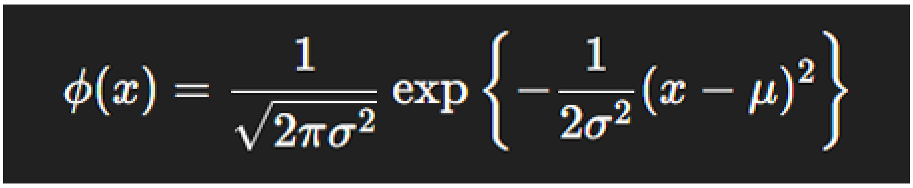
The Gaussian distribution is parameterized by two parameters:

1. **Mean (µ)**
   - It tells you where the center of the curve is.
   - It's the average value.
2. **Variance (σ^2)**
   - It tells you how spread out the values are.
   - A small variance = thin and tall curve (most values are close to the mean).
   - A large variance = wide and flat curve (values are more spread out).

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

Univariate Normal Distribution

A random variable X is normally distributed with mean µµµ and variance o^2 if it has the probability density function of X as:



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

Key Points

- This function represents the bell-shaped curve commonly seen in statistics.
- In the formula, the term (x−μ)^2 represents the squared difference between the variable xxx and its mean μ.
- This squared difference is minimized when x=μx .
- Therefore, the quantity reaches its maximum when x=μx .
- Since the exponential function is a monotonic function, the normal density also reaches its maximum value at x=μx

Role of Variance (σ^2)

- The variance σ^2 defines the spread (or width) of the curve.
- If σ^2 is large → the distribution is wider (more spread out).
- If σ^2 is small → the distribution is narrower (more concentrated around the mean).

Multivariate Gaussian Distribution

- Multivariate analysis is a branch of statistics concerned with the analysis of multiple measurements, made on one or several samples of individuals.
- Multivariate statistical analysis is concerned with data that consist of sets of measurements on a number of individuals or objects.

Why is the multivariate normal distribution so important?

There are three reasons:

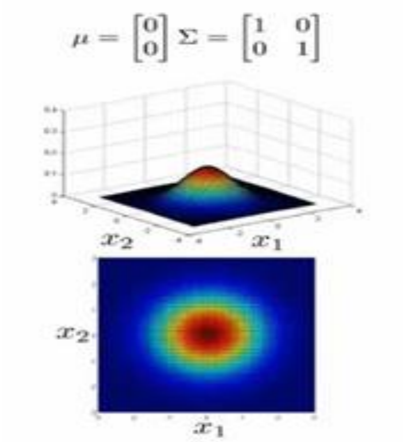
1. Mathematical Simplicity. It turns out that this distribution is relatively easy to work with, so it is easy to obtain multivariate methods based on this particular distribution.
2. Multivariate Central Limit Theorem: If X1,X2,…,Xn are independent and identically distributed random vectors, then for large n, the sample mean vector x¯ is approximately multivariate normally distributed.

Many natural phenomena may also be modeled using this distribution, just as in the univariate case.

$$p(x;\ \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right)$$

Visual Representation of Multivariate Gaussian Distribution

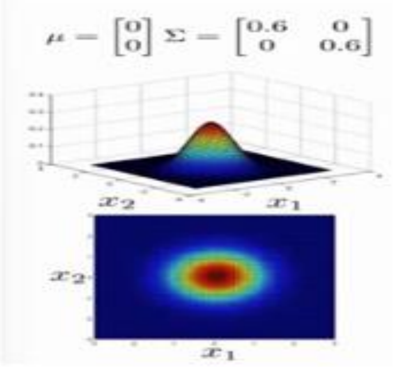
a. **Standard Normal Distribution**

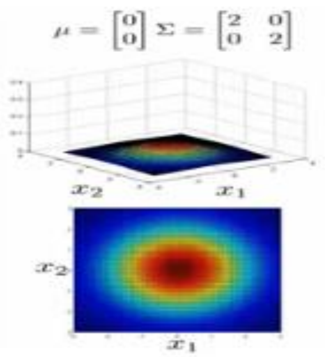


The dark red color area in the center shows the highest probability density area. The probability density keeps going lower in the lighter red, yellow, green, and cyan areas. It's the lowest in the dark blue color zone.

In both x1 and x2 direction, the highest probability density is at 0 as the μ is zero.

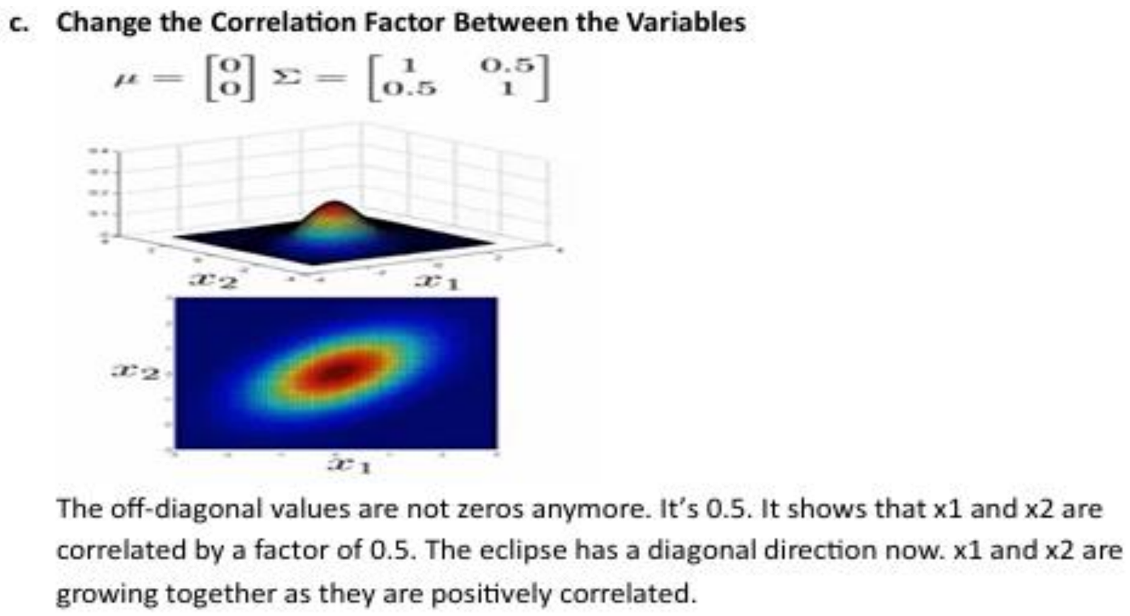
b. **Changing the Standard Deviation – Sigma**





when the standard deviation sigma shrinks, the range also shrinks. At the same time, the height of the curve becomes higher to adjust the area.

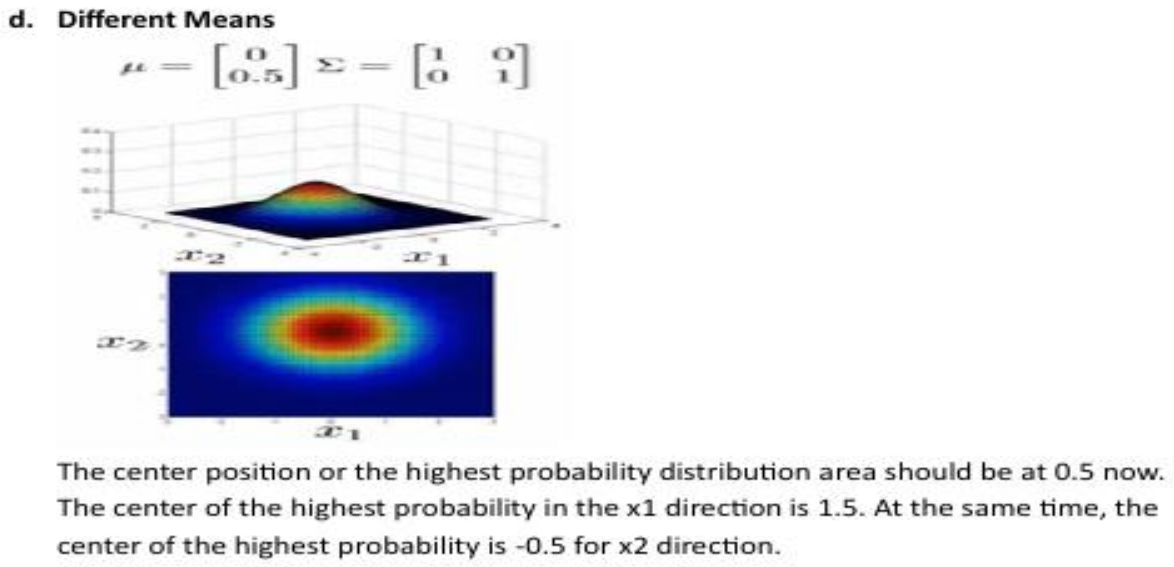
In the contrast, when sigma is larger, the variability becomes wider. So, the height of the curve gets lower.



c. Change the Correlation Factor Between the Variables

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

The off-diagonal values are not zeros anymore. It's 0.5. It shows that x1 and x2 are correlated by a factor of 0.5. The eclipse has a diagonal direction now. x1 and x2 are growing together as they are positively correlated.

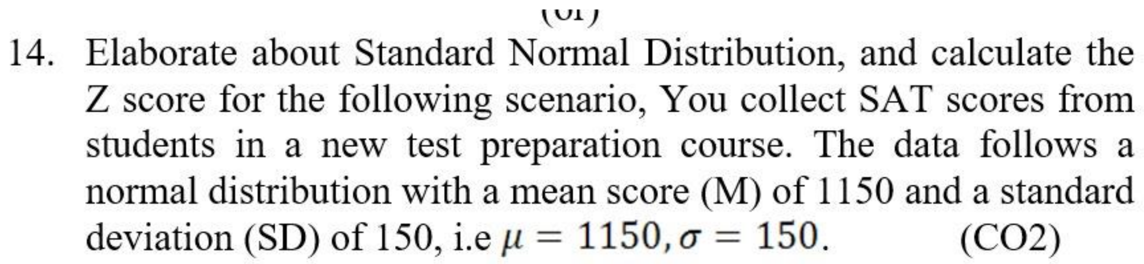


d. Different Means

$$\mu = \begin{bmatrix} 0 \\ 0.5 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

The center position or the highest probability distribution area should be at 0.5 now. The center of the highest probability in the x1 direction is 1.5. At the same time, the center of the highest probability is -0.5 for x2 direction.

NUMERICAL PROBLEM :



(or)

14. Elaborate about Standard Normal Distribution, and calculate the Z score for the following scenario, You collect SAT scores from students in a new test preparation course. The data follows a normal distribution with a mean score (M) of 1150 and a standard deviation (SD) of 150, i.e $\mu = 1150, \sigma = 150.$ (CO2)

Question Breakdown

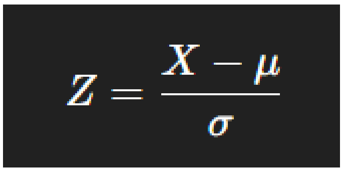
Q: Elaborate about **Standard Normal Distribution**, and calculate the **Z score** for the given scenario.

- SAT scores follow N(μ=1150, σ=150)
- We need to explain standard normal distribution and calculate the Z-score.

1. Standard Normal Distribution

1. The **Standard Normal Distribution** is a special case of the normal distribution.
2. It has a **mean (μ) = 0** and a **standard deviation (σ) = 1**.
3. Any normal distribution N(μ,σ)N(\mu, \sigma)N(μ,σ) can be converted into the standard normal distribution using the **Z-score transformation**.
4. The Z-score tells us how many standard deviations a particular value XXX is away from the mean.

Formula:



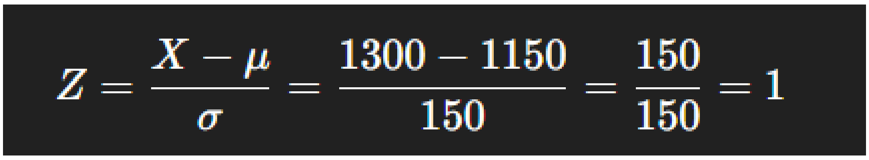
$$Z = \frac{X - \mu}{\sigma}$$

2. Given Data

- Mean (μ) = 1150
- Standard deviation (σ) = 150
- Random variable (X) = not specified → we usually compute for a score, say X=1300 (you can substitute any value).

3. Example Calculation (for X=1300X = 1300X=1300)

The Z-score is **1**, which means a score of 1300 is **1 standard deviation above the mean**.



$$Z = \frac{X - \mu}{\sigma} = \frac{1300 - 1150}{150} = \frac{150}{150} = 1$$

4. Interpretation

- If Z>0 the score is above the mean.
- If Z<0 , the score is below the mean.
- The Z-score allows us to use standard normal distribution tables to find probabilities.

Hypothesis Testing

Hypothesis testing is a part of statistical analysis, where we test the assumptions made regarding a population parameter.

It is generally used when we were to compare:

- a single group with an external standard
- two or more groups with each other

A Parameter is a number that describes the data from the population whereas, a Statistic is a number that describes the data from a sample.

Terminologies

Null Hypothesis:

Null hypothesis is a statistical theory that suggests there is no statistical significance exists between the populations.

It is denoted by $\mathbf{H_0}$ and read as **H-naught**.

Alternative Hypothesis:

An Alternative hypothesis suggests there is a significant difference between the population parameters. It could be greater or smaller. Basically, it is the contrast of the Null Hypothesis.

It is denoted by $\mathbf{H_a}$ or $\mathbf{H_1}$.

- $\mathbf{H_0}$ must always contain **equality (=)**.
- $\mathbf{H_a}$ always contains **difference (≠, >, <)**.

Types of Hypothesis Tests

Two-tailed test: Tests for any difference

H_0: $\mu_1 = \mu_2$

H_a: $\mu_1 \neq \mu_2$

One-tailed test:

To test if one mean is greater than or less than the other:

$\mathbf{H_0}$**:** $\boldsymbol{\mu_1 = \mu_2}$

$\mathbf{H_a}$**:** $\boldsymbol{\mu_1 > \mu_2}$ *(right-tailed)*

or

$\mathbf{H_a}$**:** $\boldsymbol{\mu_1 < \mu_2}$ *(left-tailed)*

Level of Significance:

Denoted by **alpha (α)**. It is a fixed probability of wrongly rejecting a true Null Hypothesis.

For example, if **α = 5%**, that means we are okay to take a 5% risk and conclude there exists a difference when there is **no actual difference**.

Steps of Hypothesis Testing

1. Define the null hypothesis (**H_0**) and alternative hypothesis (**H_a**)
2. **Set the Level of Significance (α)**

   Choose a significance level (e.g., **α = 0.05**) that determines the acceptable risk of rejecting a true null hypothesis.
3. **Collect Sample Data and Calculate Test Statistic & P-value**

   Perform the hypothesis test appropriate for your data.

   Compute the **test statistic** (e.g., *t*, *z*) and the corresponding **p-value**.
4. **Make a Conclusion**
   - If p-value ≤ α → **Reject H_0**
   - If p-value > α → **Fail to reject H_0**
5. **Confusion Matrix in Hypothesis Testing**

	H_0 is True	**H_0 is False**
Reject H_0	Type I Error (α)	☑ Correct Decision
Fail to Reject H_0	☑ Correct Decision	Type II Error (β)

$$\text{Accuracy} = \frac{\#\ \text{correct predictions}}{\#\ \text{total cases}}$$

Confidence:

The **probability of accepting a True Null Hypothesis**.

It is denoted as **(1 - α)**.

Power of Test:

The **probability of rejecting a False Null Hypothesis**, i.e., the ability of the test to detect a true difference.

It is denoted as **(1 - β)** and its value lies between **0 and 1**.

Type I Error (False Positive):

Occurs when we **reject a True Null Hypothesis**.

It is denoted as **α.**

Type II Error (False Negative):

Occurs when we **accept a False Null Hypothesis**.

It is denoted as **β**.