STATISTICS FOR DATA SCIENCE

Normal Distribution

Prof. Uma D
Prof. Suganthi S
Prof. Silviya Nancy J

Department of Computer Science and Engineering

STATISTICS FOR DATA SCIENCE

Normal Distribution

Prof. Uma D
Prof. Suganthi S
Prof. Silviya Nancy J

STATISTICS FOR DATA SCIENCE

Topics to be covered...

- Continuous Probability Distribution
- Normal Distribution
- Probability Density Function
- Standard Normal Distribution
- Linear Function of a Normal Random Variable
- Linear Function of a Independent Normal Random Variable
- Two independent normally distributed random variables

Continuous Probability Distribution

- There are many different types of continuous random variables

- We try to pick a model that
  - Fits the data well
  - Allows us to make the best possible inferences using the data.

- One important continuous random variable is the normal random variable

STATISTICS FOR DATA SCIENCE

Normal Distribution

Carl Friedrich Gauss
(1777–1855)

Source:https://en.wikipedia.org/wiki/Carl_Friedrich_Gauss

STATISTICS FOR DATA SCIENCE

Normal Distribution

- German mathematician and scientist
- Contributions in many fields of mathematics and science
- Referred to as the “Prince of Mathematicians”
- Credited with the use of the probability distribution now known as the normal or Gaussian distribution (bell curve)

Source:https://en.wikipedia.org/wiki/Carl_Friedrich_Gauss

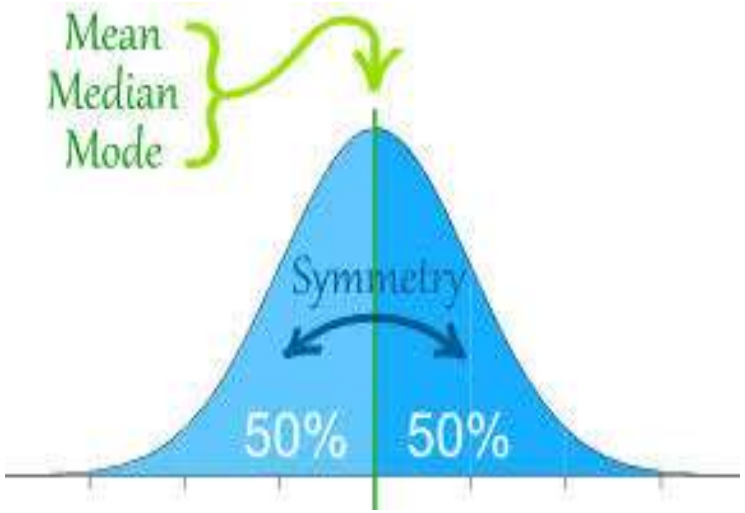
STATISTICS FOR DATA SCIENCE

Normal Distribution

Extremely important continuous probability distribution.

Rises frequently in theory and practice.

We say that a random variable is normally distributed with mean μ and standard deviation σ if the **probability density function** is given by



Source Image: dietassessmentprimer.cancer.gov

Probability density function

Extremely important continuous probability distribution.

Rises frequently in theory and practice.

We say that a random variable is normally distributed with mean μ and standard deviation σ if the **probability density function** is given by

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{for} \quad -\infty<x<\infty$$

Examples of Normal Distribution

- Heights of People
- Test Scores
- Errors in measurements
- Blood Pressure
- Size of things produced by machines

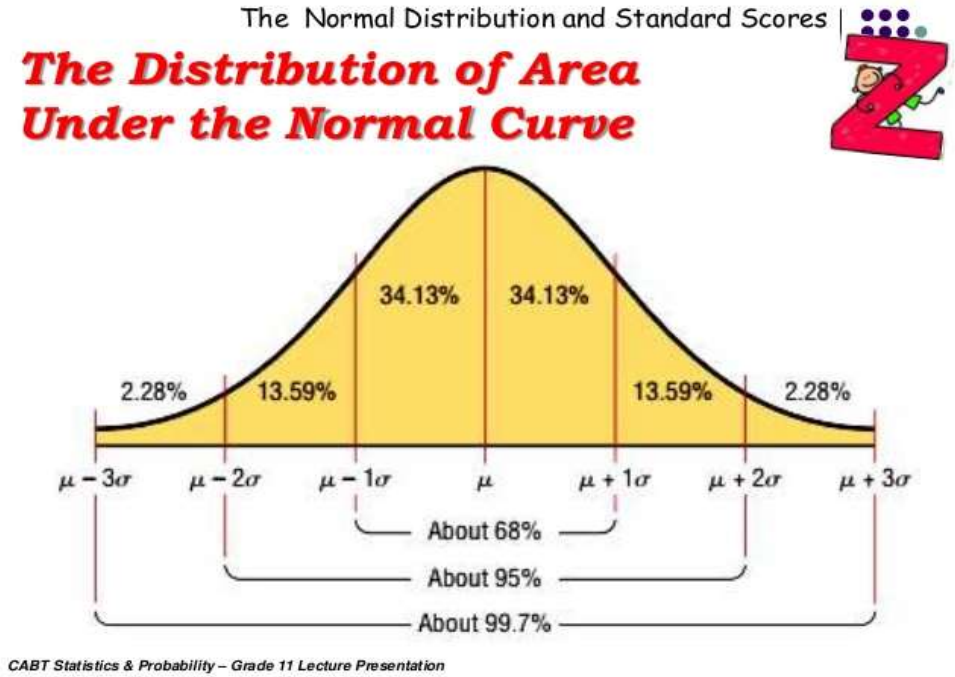
STATISTICS FOR DATA SCIENCE

Normal Distribution

Why to know Standard Deviation?

Any value is

- **Likely** to be within **1** standard deviation of the mean.
- **Very Likely** to be within **2** standard deviations.
- **Almost certainly** within **3** standard deviations.

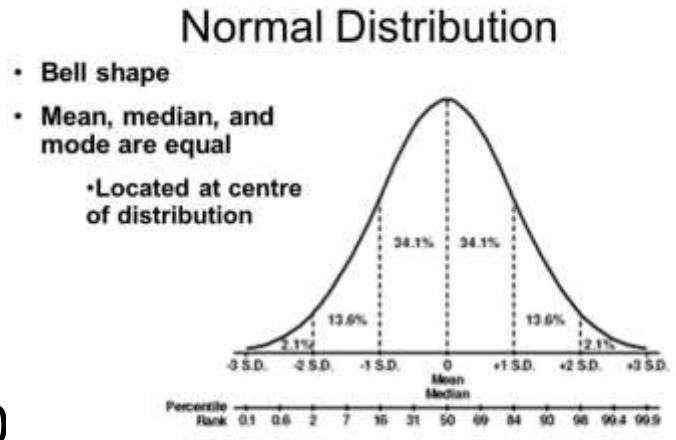


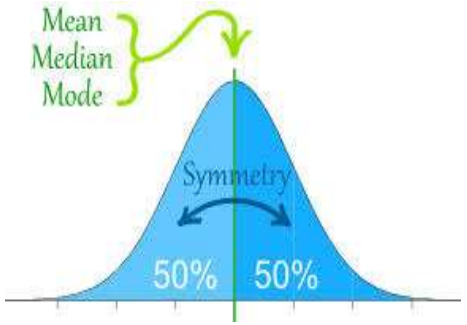
Source Image: www.slideshare.net

Normal Distribution

The Properties of a Normal Distribution

- Mean=Median=Mode
- Symmetry about the center(mean) μ.
- 50% of the values less than the mean and 50 greater than the mean.
- Changing **μ *shifts*** the ***distribution*** **left or right**.
- Changing **σ *increases or decreases*** the **spread.**



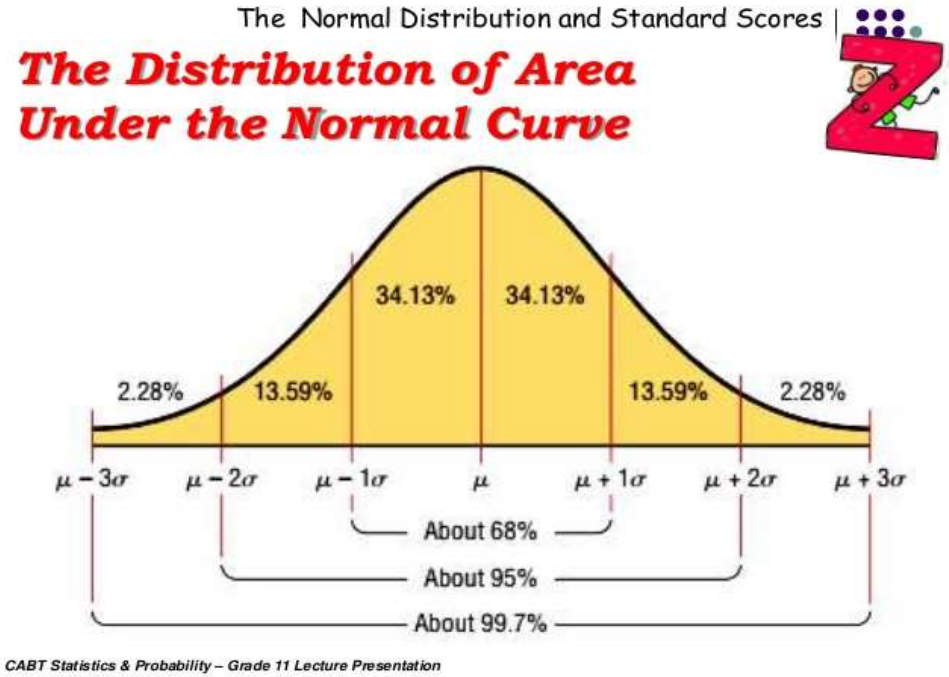


Source Image: dietassessmentprimer.cancer.gov, www.google.com

Normal Distribution

The Properties of a Normal Distribution

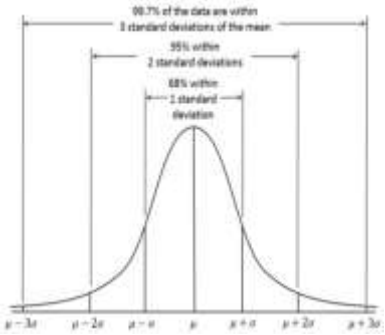
- Approximately **68%** of the area is within **1** standard deviation.
- Approximately **95%** of the area is within **2** standard deviations.
- Approximately **99.7%** of the area is within **3** standard deviations.



Source Image: www.slideshare.net

Normal Distribution

How good is rule for real data?



Suppose SAT scores roughly follow a normal distribution(with range restricted to 200-800), and the average math SAT is 500 with a standard deviation of 50, then:

- 68% of students will have scores between 450 and 550.
- 95% will be between 400 and 600.
- 99.7% will be between 350 and 650.

What if you wanted to know the math SAT score corresponding to the 95^{th} percentile(=95% of students are lower)?

Standard Normal Distribution

- Is a Normal distribution with mean 0 and variance 1.
- Random Variable that has standard normal distribution is referred using letter Z.

$$Z \sim N(0, 1)$$

- Probabilities associated with Normal Variates can be calculated by using transformations to the Standard Normal Variate (z) – using z-table.

Standardizing Normally Distributed Random Variables

We can convert a Random Variable X having a Normal distribution with any mean and Standard deviation in to the Random variable that has a Standard Normal Distribution.

$$X \sim N(\mu, \sigma^2)$$

Standardizing X : using a basic linear transformation:

$$z = (x - \mu) / \sigma$$

Example

Find area under the normal curve:

a) To the left of z = -0.49
b) To the left of z = 0.49
c) To the right of z = 0.49
d) Between z = 0.40 and z = 1.30
e) Between z = -1.50 and z = 0.90

Solution

a) To the **left of z = -0.49** = **0.3121**
b) To the **left of z = 0.49** = **0.6879**
c) To the **right of z = 0.49**
= 1 – 0.6879 = **0.3121**

d) Between z = 0.40 and z = 1.30
= Area to left of 1.30 – Area to left of0.40
= 0.9032 – 0.6554
= **0.2478**

e) Between z = -1.50 and z = 0.90
= Area to left of 0.90 – Area to left of-1.50
= 0.8159 – 0.0668
= **0.7491**

Example

Let Z ~ N(0, 1) . Find a constant c for which

a) $P(Z \geq c) = 0.1587$

b) $P(c \leq Z \leq 0) = 0.4772$

c) $P(-c \leq Z \leq c) = 0.8664$

Solution

a) **P(Z ≥c) = 0.1587**

=> Area to left of c = 1 – 0.1587 = 0.8413

=> **c = 1.00**

b) **P(c ≤ Z ≤ 0) = 0.4772**

Area to left of 0 = 0.5

=>Area to left of c = 0.5 – 0.4772 = 0.0228

=> **c = -2.00**

c) **P(-c ≤ Z ≤ c) = 0.8664**

P(0 ≤ Z ≤ c) = 0.8664/2 = 0.4332

Area to right of c = 0.5 – 0.4332 = 0.0668 => Area to left of -c = 0.0668 => c=-1.50

Area to left of c = 1 – 0.0668 = 0.9332 => **c = 1.50**

Example

If $X \sim N(2, 9)$, compute:

a) $P(X \geq 2)$

b) $P(1 \leq X < 7)$

c) Find the median of X.

d) Find 75^{th} percentile of X.

Text-Book-Chapter 4.5- Page.No:252-Q.No:4

Solution

a) P(X ≥ 2)

$Z = (x - 2)/3 = (2 - 2)/3 = 0$

$P(Z \geq 0) = 0.5$

b) P(1 ≤ X < 7)

$P((1 - 2)/3 \leq Z < (7 - 2)/3) = P(-1/3 \leq Z < 5/3) = P(-0.33 \leq Z < 1.67)$

= Area to left of 1.67 – Area to left of -0.33

= 0.9525 – 0.3707

= 0.5818

Solution

c) Find the median of X.

$P(Z \leq c) = 0.5 \Rightarrow c = 0.0$

$\Rightarrow 0.0 = (x - 2)/3 \Rightarrow x = 2$

d) Find 75^{th} percentile of X.

$P(Z \leq c) = 0.75$ => closest area to 0.7500 is 0.7486 corresponding to a z-score of 0.67

$\Rightarrow 0.67 = (x - 2)/3 \Rightarrow x = 4.01$

Example

The lifetime of a battery is in a certain application is normally distributed with mean 16 hours, standard deviation 2 hours.

a) What is the probability that a battery will last more than 19 hours?

b) Find the 10^{th} percentile of the lifetimes.

c) A particular battery lasts 14.5 hours. What percentile is its lifetime on?

Text-Book-Chapter 4.5- Page.No:252-Q.No:5

Solution

a) What is the probability that a battery will last more than 19 hours?

To find : $P(X \geq 19)$

$Z = (19 - 16)/2$

$=> 1.5$

$P(Z \geq 1.5) = 1 - P(Z \leq 1.5) = 1 - 0.9332 = 0.0668$

b) Find the 10^{th} percentile of the lifetimes.

$P(Z \leq c) = 0.1000$

Closest area to 0.1000 is 0.1003 corresponding to a z-score of -1.28

$=> -1.28 = (x - 16)/2 => x = 13.44$

c) A particular battery lasts 14.5 hours. What percentile is its lifetime on?

$Z = (14.5 - 16) / 2 = -0.75$

$P(Z \leq -0.75) = 0.2266$

=> its lifetime is approximately on 23^{rd} percentile.

STATISTICS FOR DATA SCIENCE

Linear Function of a Normal Random Variable

If $X \sim N(\mu_1, \sigma_1^2)$ and a and b are constants, and

$$Y = aX + b$$

then,

$$Y \sim N(a\mu_X + b, a^2\sigma_Y^2)$$

Linear Function of a Independent Normal Random Variable

Let $X_1, X_2, \ldots, X_n$ be independent and normally distributed with means $\mu_1, \mu_2, \ldots, \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$. Let $c_1, c_2, \ldots, c_n$ be constants, and $c_1X_1 + c_2X_2 + \ldots + c_nX_n$ be a linear combination. Then

$$c_1X_1 + c_2X_2 + \ldots + c_nX_n$$
$$\sim N(c_1\mu_1 + c_2\mu_2 + \ldots + c_n\mu_n, c_1^2\sigma_1^2 + c_2^2\sigma_2^2 + \ldots + c_n^2\sigma_n^2)$$

Two independent normally distributed random variables

Sum/ Difference of two independent normally distributed random variables is normal.

If $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ are independent random variables that are normally distributed, then their sum/difference is also normally distributed. i.e., if

If,

$$X \sim N(\mu_X, \sigma_X^2)$$

$$Y \sim N(\mu_Y, \sigma_Y^2)$$

$$X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

Then,

$$X - Y \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

Example

Let X be a normal random variable with mean 2 and variance 3,and

let X be a normal random variable with mean 1 and variance 4.

Assume that X1 and X2 are independent.

What is the distribution of the linear combination Y = 2X1+ 3X2 ?

Example

Y is normally distributed with mean 7 and variance 48 as the following calculation illustrates:

$$(2X_1 + 3X_2) \sim N(2(2)+3(1),\ 2^2(3)+3^2(4)) = N(7,48)$$

Example

A light fixture holds two lightbulbs.Bulb A is a type whose lifetime is normally distibuted with mean 800 hours and standard deviation 100 hours. Bulb B has a lifetime that is normally distibuted with mean 900 hours and standard deviation 150 hours.Assume the lifetimes of the bulbs are independent.

1) What is the probability Bulb B lasts longer than bulb A?

2) What is the probability Bulb B lasts 200 hours more than bulb A?

3)Another light fixture holds only one bulb. A bulb of type A is installed, and when it burns out, a bulb of type B is installed.

What is the probability that the total lifetime of the two bulbs is more than 2000 hours?

Text-Book-Chapter 4.5- Page.No:255-Q.No.21

Solution

Bulb A : $X \sim N(800, 100^2)$
Bulb B : $Y \sim N(900, 150^2)$

Assume the lifetimes of the bulbs are independent.

1) What is the probability Bulb B lasts longer than bulb A?

D = Y – X. The event B > A is the event D > 0

$$D \sim N(\mu_Y - \mu_X, \sigma_Y^2 + \sigma_X^2)$$

$$D \sim N(100, 180.28^2)$$

Since D is a linear combination of independent normal random variables,
D is normally distributed.P(D > 0)

Z = (0 – 100)/180.28 => z = -0.55

P(Z > -0.55) = 1 – P(Z < -0.55) = 1 – 0.2912 = 0.7088

Example

Bulb A : $X \sim N(800, 100^2)$
Bulb B : $Y \sim N(900, 150^2)$

Assume the lifetimes of the bulbs are independent.

2) What is the probability Bulb B lasts 200 hours more than bulb A?

D = Y – X. The event B > A is the event D > 0

$$D \sim N(\mu_Y - \mu_X, \sigma_Y^2 + \sigma_X^2)$$

$$D \sim N(100, 180.28^2)$$

Since D is a linear combination of independent normal random variables,
D is normally distributed.P(D > 200)

$Z = (200 - 100)/180.28 \Rightarrow z = 0.55$

$P(Z > 0.55) = 1 - P(Z < 0.55) = 1 - 0.7088 = 0.2912$

Example

3) Another light fixture holds only one bulb. A bulb of type A is installed, and when it burns out, a bulb of type B is installed.

What is the probability that the total lifetime of the two bulbs is more than 2000 hours?

$$X + Y \sim N(\mu_Y + \mu_X, \sigma_X^2 + \sigma_Y^2)$$

$$X + Y \sim N(1700, 180.28^2)$$

Since X + Y is a linear combination of independent normal random variables, X + Y is normally distributed.

P(X + Y > 2000)

Z = (2000 – 1700)/180.28 => z = 1.66

P(Z > 1.66) = 1 – P(Z < 1.66) = 1 – 0.9515 = 0.0485

THANK YOU

Prof. Uma D
Prof. Suganthi S
Prof. Silviya Nancy J

Department of Computer Science and Engineering