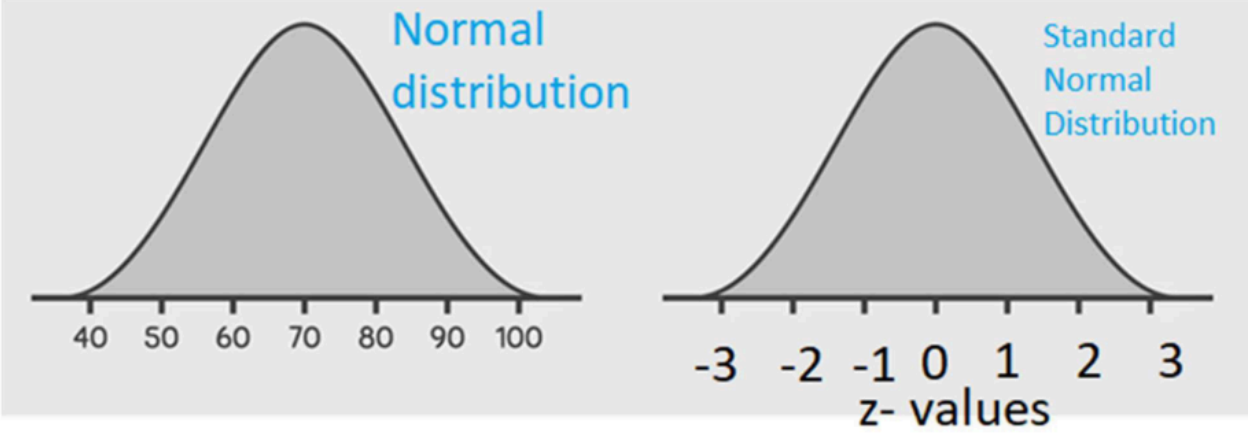
Z-Score

Z Score gives how many standard deviations away from mean a value is. However, to understand the probability associated with it, we need to refer to Z-Table.

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

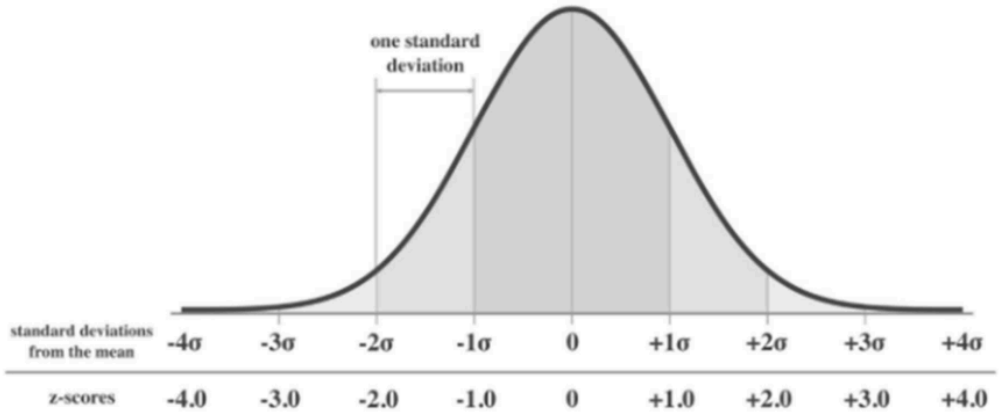
μ = Mean
σ = Standard Deviation

Q1. For a recent final written statistics exam for a “Data scientist” job selection process, the mean was 70 with a standard deviation of 10. If you scored 76 marks. What is your area in the Normal distribution?.



Mean value 70 Z score = (70-70)/10=0
80 marks Z score = (80-70)/10=1
60 marks Z score = (60-70)/10=-1
76 marks Z score =(76-70)/10=0.60
In the z table the value of 0.60 is 0.7257.
This is the value of area under curve or the percentile.

In Standard Normal Distribution mean = 0 and std = 1.
We can convert any normal distribution to standard normal distribution

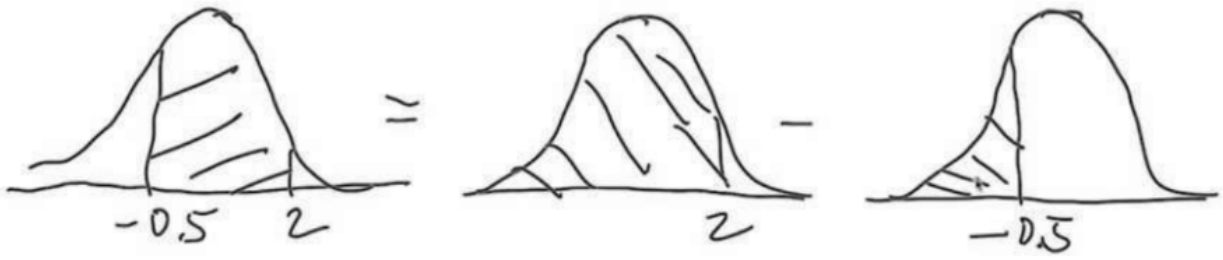


2 Types of problems :
Type 1: Comparison of 2 different Normally Distributed values (Z-Score is enough)
Type 2: Finding the probability or percentage of values. (Need Z-table)

Q1. Fathers height follows normal distribution with a mean of 68.3 inches and a SD of 1.8 inches. What percentage of fathers have heights between 67.4 and 71.9?

Ans:
Step 1: Standardize
Z(67.4) = 67.4-68.3/1.8 = -0.5
Z(71.9) = 71.9-68.3/1.8 = 2

Step 2:
Area to left of the point z ie -0.5 = 0.3085
Area to left of the point z ie 2 = 0.9772



Percentage of fathers have heights between 67.4 and 71.9 =
0.9772-0.3085 = 0.668

Q2. What is the 30 percentile of the Father's height?

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

μ = Mean
σ = Standard Deviation

Z = -0.5217
Mean = 68.3, sd = 1.8 ,x = (-0.5217*1.8) + 68.3

Q3. One student score 80 marks in Mathematics and 75 Marks in English. At this point we can say he has performed excellent in Math as compare to English.
Consider the average class score of Mathematics is 90, SD is 5. Average class score of English is 60 , SD is 5. Verify the performance?

Ans:- lets calculate the Z values
Zm=(80-90)/5= -2
Ze=(75-60)/5=3

3 is the higher value and it's closer to +3

Q4. What proportion of students are between 5.81 feet & 6.3 feet height. Given Mean=5.5, sd=0.5 feet.
Step1: Convert to Z scores
Z1=(5.81-5.5)/0.5= 0.619
Z2=(6.3-5.5)/0.5= 1.599
Step 2: Find probability / area from the z table
Area to left of the point z ie 0.619 = 0.7291
Area to left of the point z ie 1.599 = 0.9772

Percentage of students heights between 5.81 and 6.3 = 0.9772-0.7291
= 0.2481 / 24.81%

Q5. Mean height of Gurkhas is 146 cm with Sd of 3 cm . What is the probability of
(a)Height greater than 152 cm.
(b)Height between 140 and 150 cm.

Ans:

(a) Z score (152) = (152-146)/3 = 2
Area to the right = 1 - 0.9772 = 0.0228

Calculate Z scores
(b) Z score (140) = (140 -146)/3 = -2
Z score (150) = (150 -146)/3 = 1.333

Step 2:
Area to left of the point z ie -2 = 0.0228
Area to left of the point z ie 1.333 = 0.9082

Height between 140 and 150 cm = 0.9082 - 0.0228 = 0.8860 / 88.6%

Q6. According to the Center for Disease Control, heights for U.S. adult females and males are approximately normal.
Females: mean of 64 inches and SD of 2 inches
Males: mean of 69 inches and SD of 3 inches
Find the probability of a randomly selected U.S. adult female being taller than 65 inches.

Inferential Statistics

Hypothesis Testing:

Hypothesis - Specific prediction based on some previous research. Testing this prediction is called Hypothesis testing.

Point Estimate: A point estimate is a single value (such as a sample mean) that serves as the best guess or approximation of a population parameter(such as population mean) based on sample data.

Confidence Interval: A confidence interval is a range of values that is likely to contain the true population parameter(like mean) with a certain level of confidence. It provides a range of plausible values for the parameter of interest.

Confidence Level: The confidence level is the probability that the confidence interval contains the true population parameter. It is often expressed as a percentage (e.g., 95%, 90%, etc.).

Significance Level: The significance level, denoted by α, is the probability of rejecting the null hypothesis when it is actually true. It sets the threshold for determining whether the results of a hypothesis test are statistically significant.

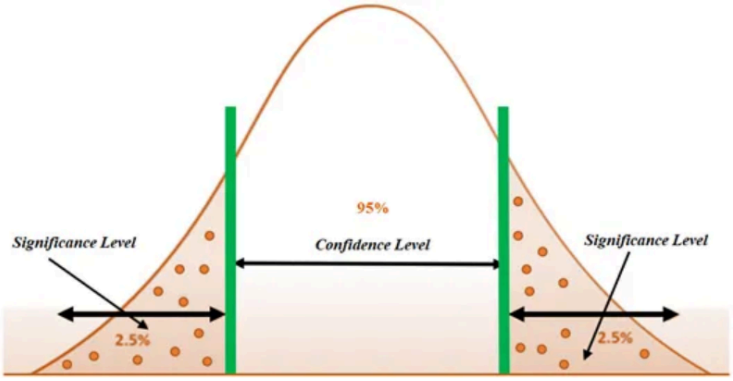
Eg: **Hypothesis**: Researchers have observed Avg Salary of Data Scientist around the world is changed after recession in 2023. Before recession in 2023 avg salary of data scientist was 50000

How to test this hypothesis?

1. **Define what is null hypothesis and alternate hypothesis**
   There are two types of hypothesis:
   Null Hypothesis (H0) - The null hypothesis is a statement of no effect or no difference. It represents the assumption that there is no change, no effect, or no relationship between variables.
   Example: The average salary of data scientists around the world did not change after the recession in 2023.

   Alternate Hypothesis (H1/Ha) - The average salary of data scientists around the world changed after the recession in 2023.
   Example: The average salary of data scientists around the world changed after the recession in 2023.

Case 1: Avg Salary of a sample $\overline{X}$ = 52000 (No Significant difference from mean)
Case 2: Avg Salary of a sample $\overline{X}$ = 48000 (No Significant difference from mean)
Case 3: Avg Salary of a sample $\overline{X}$ = 60000 (Significant difference from mean)
Case 4: Avg Salary of a sample $\overline{X}$ = 40000 (Significant difference from mean)



Significance Level and Confidence Level

Define the confidence interval:
Confidence interval = Point estimate +/- margin of error.
In this example: 50000 + margin of error , 50000 - margin of error

Margin of Error
The margin of Error used to determine the by which the sample result

Mean will differ from the value of the entire population mean.
A higher margin of error indicates a high chance that the result of the sample may not be the true reflection of the whole population.

Margin of Error = $Z * \sigma / \sqrt{n}$

Z value - Point on the x axis where my significance level lies.
We can get the Z value based on significance level from Z table

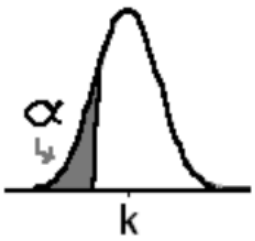
10% α ---- z value = 1.65

5% α ---- z value = 1.96

2% α---- z value = 2.33

1% α---- z value = 2.58

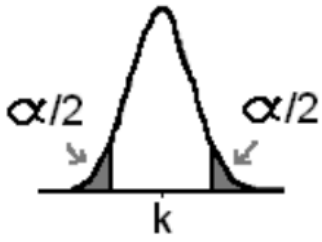
Z value also helps me understand whether we need to perform 1 tailed test / 2 tailed test



$H_0: \mu = k$

$H_1: \mu < k$

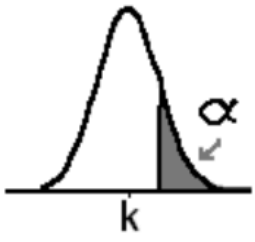
α	z critical
0.10	-1.28
0.05	-1.65
0.01	-2.33



$H_0: \mu = k$

$H_1: \mu \neq k$

α	z critical
0.10	± 1.65
0.05	± 1.96
0.01	± 2.58



$H_0: \mu = k$

$H_1: \mu > k$

α	z critical
0.10	1.28
0.05	1.65
0.01	2.33

Outcome from my Hypothesis testing:
Case 1: Reject null hypothesis hence proving the alternate hypothesis
Case 2: Failed to reject null hypothesis hence the null hypothesis remains the same

Q1. A Researcher has agreed upon data on a daily return of a portfolio of call options over a recent 250 days period. The mean daily return has been 0.1% and the standard deviation of daily return portfolio is 0.25%. The researcher's believe that the mean daily portfolio is not 0. Construct a hypothesis of research belief at 95% confidence interval.

Step 1: Define Hypothesis
Ho - Mean = 0
Ha - Mean != 0

Step 2: Check the tails of the Test
Mean > 0 or Mean < 0 then the null hypothesis will be rejected, hence it is two tailed test

Step 3: Construct the confidence intervals

Confidence interval with 95% confidence level

$\alpha = 5\%$, 2.5% and 2.5%

Z value for 2.5 ie 0.025 is 1.96

Confidence interval = Point estimate +/- margin of error.

= 0 +|- (1.96 * 0.0025)/sqrt(250)

= (-0.0003, +0.0003)

Step 4: Check if mean lies in between the confidence interval
Mean daily return was 0.1% ie 0.001. Since it lies outside the region, we reject the null hypothesis. The claim of the researcher is correct.

Q2. A principal at a school claims that the students in his school are above average in terms of intelligence. A random sample of 30 students IQ scores has been taken which has a mean of 112.5. The mean population IQ is 100 with a std of 15.

Critical Method

Step 1: Define Hypothesis

Ho : Mean <= 100

Ha : Mean > 100

Step 2: Check if it's one tailed / two tailed. -> one tailed, right tailed

We don't have significance level, assuming that its 5%

Zcritical = 1.65

Zcal = X-Mean/std/sqrt(n)

= 112.5-100/15/sqrt(30)

= 4.57

Zcal > Zcritical , reject the null hypothesis the principals claim is correct

P value method - used in python

P value is the measure of strength of evidence against a null hypothesis. Ie towards alternate hypothesis

Large p value suggests weaker evidence, small p value suggests stronger evidence.

If p value > α -> Failed to reject null hypothesis

If p value < α -> Reject the null hypothesis