Normal Distribution and Z Score: Math and statistics for data science

In [4]:

```
import pandas as pd
import seaborn as sn
```

We are going to use heights dataset from kaggle.com. Dataset has heights and weights both but I have removed weights to make it simple

https://www.kaggle.com/mustafaali96/weight-height

In [5]:

```
df = pd.read_csv("heights.csv")
df.head()
```

Out[5]:

	gender	height
0	Male	73.847017
1	Male	68.781904
2	Male	74.110105
3	Male	71.730978
4	Male	69.881796

(1) Outlier detection and removal using Standard Deviation

In [6]:

```
df.height.describe()
```

Out[6]:

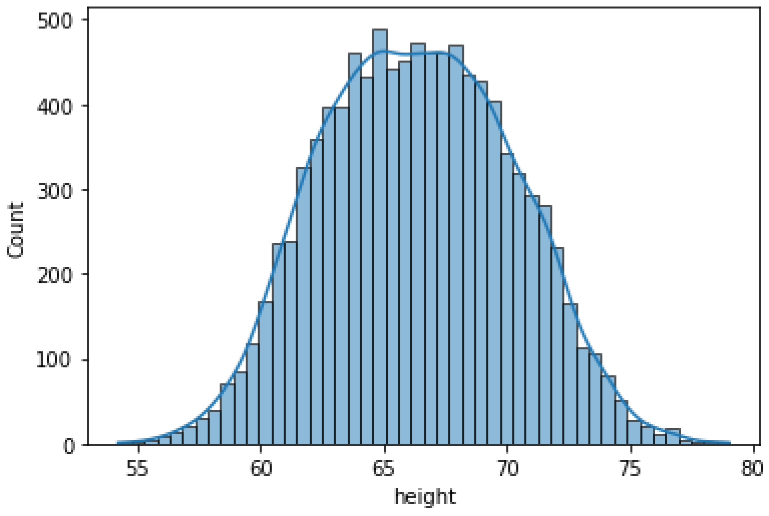
```
count    10000.000000
mean        66.367560
std          3.847528
min         54.263133
25%         63.505620
50%         66.318070
75%         69.174262
max         78.998742
Name: height, dtype: float64
```

In [7]:

```
sn.histplot(df.height, kde=True)
```

Out[7]:

```
<AxesSubplot:xlabel='height', ylabel='Count'>
```



```
In [8]:  mean = df.height.mean()
          mean
```

```
Out[8]:  66.3675597548656
```

```
In [9]:  std_deviation = df.height.std()
          std_deviation
```

```
Out[9]:  3.847528120795573
```

```
In [10]:  mean-3*std_deviation
```

```
Out[10]:  54.824975392478876
```

```
In [11]:  mean+3*std_deviation
```

```
Out[11]:  77.91014411725232
```

```
In [12]:  df[(df.height < 54.82) | (df.height > 77.91)]
```

Out[12]:

	gender	height
994	Male	78.095867
1317	Male	78.462053
2014	Male	78.998742
3285	Male	78.528210
3757	Male	78.621374
6624	Female	54.616858
9285	Female	54.263133

```
In [13]:  df_no_outlier = df[(df.height<77.91) & (df.height>54.82)]
          df_no_outlier.shape
```

```
Out[13]:  (9993, 2)
```

```
In [21]:  df_no_outlier
          df_no_outlier.describe()
```

Out[21]:

	height
count	9993.000000
mean	66.363856
std	3.835511
min	54.873728
25%	63.505894
50%	66.317755
75%	69.169353
max	77.547186

(2) Outlier detection and removal using Z Score

Z score is a way to achieve same thing that we did above in part (1)

Z score indicates how many standard deviation away a data point is.

For example in our case mean is 66.37 and standard deviation is 3.84.

If a value of a data point is 77.91 then Z score for that is 3 because it is 3 standard deviation away (77.91 = 66.37 + 3 * 3.84)

Calculate the Z Score

Let's add a new column in our dataframe for this Z score

In [15]:
```
df['zscore'] = ( df.height - df.height.mean() ) / df.height.std()
df.head(5)
```

Out[15]:

	gender	height	zscore
0	Male	73.847017	1.943964
1	Male	68.781904	0.627505
2	Male	74.110105	2.012343
3	Male	71.730978	1.393991
4	Male	69.881796	0.913375

Above for first record with height 73.84, z score is 1.94. This means 73.84 is 1.94 standard deviation away from mean

In [16]:
```
df.height.mean()
```

Out[16]:
```
66.3675597548656
```

In [17]:
```
df.height.std()
```

Out[17]:
```
3.847528120795573
```

In [18]:
```
(73.84-66.37)/3.84
```

Out[18]:
```
1.9453124999999998
```

In [19]:
```
df[df['zscore']>3]
```

Out[19]:

	gender	height	zscore
994	Male	78.095867	3.048271

	gender	height	zscore
1317	Male	78.462053	3.143445
2014	Male	78.998742	3.282934
3285	Male	78.528210	3.160640
3757	Male	78.621374	3.184854

In [20]:

```
df[df['zscore']<-3]
```

Out[20]:

	gender	height	zscore
6624	Female	54.616858	-3.054091
9285	Female	54.263133	-3.146027

Exercise

You are given bhp.csv which contains property prices in the city of banglore, India. You need to examine price_per_sqft column and do following,

(1) Remove outliers using percentile technique first. Use [0.001, 0.999] for lower and upper bound percentiles

(2) After removing outliers in step 1, you get a new dataframe.

(3) On step(2) dataframe, use 4 standard deviation to remove outliers

(4) Plot histogram for new dataframe that is generated after step (3). Also plot bell curve on same histogram

(5) On step(2) dataframe, use zscore of 4 to remove outliers. This is quite similar to step (3) and you will get exact same result

In [ ]:

```
```