2. Measure of Variability

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
dataset = pd.read_csv('titanic.csv')
```

In [4]:
```
dataset.head(3)
```

Out[4]:

	Survived	Pclass	Name	Sex	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare
0	0	3	Mr. Owen Harris Braund	male	22.0	1	0	7.2500
1	1	1	Mrs. John Bradley (Florence Briggs Thayer) Cum...	female	38.0	1	0	71.2833
2	1	3	Miss. Laina Heikkinen	female	26.0	0	0	7.9250

2.1 Range

In [8]:
```
min_r = dataset['Age'].min()
max_r = dataset['Age'].max()
```

In [9]:
```
min_r, max_r
```

Out[9]:
```
(0.42, 80.0)
```

In [10]:
```
range = max_r - min_r
```

In [11]:
```
range
```

Out[11]:
```
79.58
```

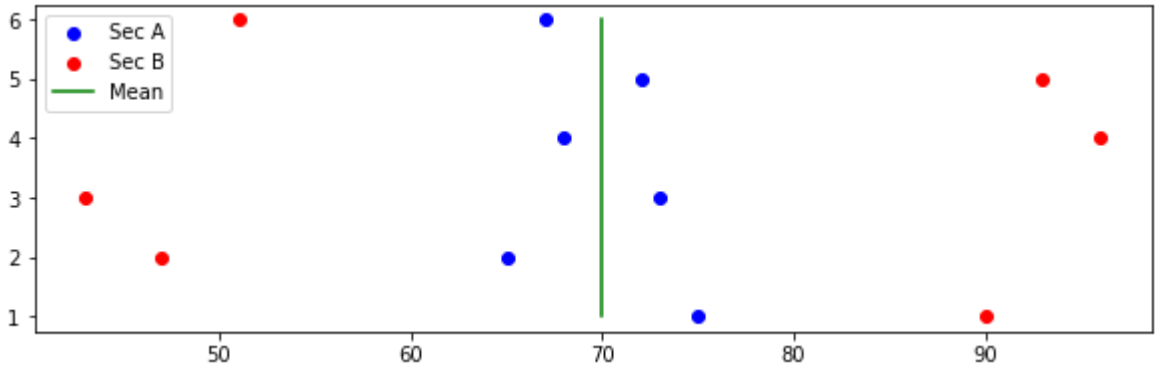
2.2 Mean Absolute Division

To simply print graph

In [23]:
```
sec_a = np.array([75,65,73,68,72,67])
sec_b = np.array([90,47,43,96,93,51])
ne = np.array([1,2,3,4,5,6])
```

In [36]:
```
mean = np.mean(sec_a)
```

In [42]:
```
plt.figure(figsize=(10,3))
plt.scatter(sec_a, ne, color="blue", label="Sec A")
plt.scatter(sec_b, ne, color="red", label="Sec B")
plt.plot([70,70,70,70,70,70], ne, c="green", label="Mean")
#plt.plot([mean for i in range(1,7)], ne, c="green", label="Mean")
plt.legend()
plt.show()
```



To use MAD formula

In [44]:
```
# To calculate xi-x
sec_b - mean
```

Out[44]:
```
array([ 5., -5.,  3., -2.,  2., -3.])
```

In [48]:
```
# To calculate |xi-x|
np.abs(sec_a -mean)
```

Out[48]:
```
array([5., 5., 3., 2., 2., 3.])
```

In [49]:
```
# To calculate sigma|xi-x|
np.sum(np.abs(sec_a - mean))
```

Out[49]:
```
20.0
```

In [51]:
```
# To calculate sigma|xi-x|/n
mad_sec_a = np.sum(np.abs(sec_a - mean))/len(sec_a)
```

In [52]:
```
# Likewise we will calculat mean absolute division of sec_b
mad_sec_b = np.sum(np.abs(sec_b - mean))/len(sec_b)
```

In [53]:
```
mad_sec_a, mad_sec_b
```

Out[53]:
```
(3.3333333333333335, 23.0)
```

So you will take sec_a for machine learning model as it contains low mean absolute division

2.3 Calculate Standard Deviation and Variance

In [55]:
```
# To calculate standard deviation of data of section A and section B
np.std(sec_a), np.std(sec_b)
```

Out[55]:
```
(3.559026084010437, 23.18045153428495)
```

In [56]:
```
# To calculate variance of data of section A and section B
np.var(sec_a), np.var(sec_b)
```

Out[56]:
```
(12.666666666666666, 537.3333333333334)
```

So We will take data of section A becuase it has low variance as well as less standard deviation

To calculate std and var on real world data

In [58]:
```
dataset = pd.read_csv('titanic.csv')
```

In [60]:
```
dataset.head(3)
```

Out[60]:

	Survived	Pclass	Name	Sex	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare
0	0	3	Mr. Owen Harris Braund	male	22.0	1	0	7.2500
1	1	1	Mrs. John Bradley (Florence Briggs Thayer) Cum...	female	38.0	1	0	71.2833
2	1	3	Miss. Laina Heikkinen	female	26.0	0	0	7.9250

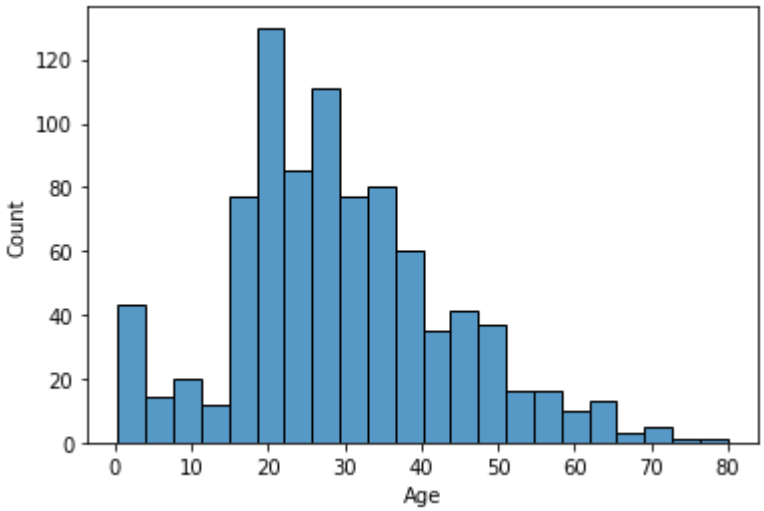
In [61]:
```
dataset['Age'].var()
```

Out[61]:
```
199.42829701227413
```

In [64]:
```
dataset['Age'].std()
```

Out[64]:
```
14.12190840546256
```

In [63]:
```
sns.histplot(x='Age', data=dataset)
plt.show()
```



In [65]:
```
dataset.describe()
```

Out[65]:

	Survived	Pclass	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare
count	887.000000	887.000000	887.000000	887.000000	887.000000	887.00000
mean	0.385569	2.305524	29.471443	0.525366	0.383315	32.30542
std	0.487004	0.836662	14.121908	1.104669	0.807466	49.78204
min	0.000000	1.000000	0.420000	0.000000	0.000000	0.00000
25%	0.000000	2.000000	20.250000	0.000000	0.000000	7.92500
50%	0.000000	3.000000	28.000000	0.000000	0.000000	14.45420
75%	1.000000	3.000000	38.000000	1.000000	0.000000	31.13750
max	1.000000	3.000000	80.000000	8.000000	6.000000	512.32920

In [ ]:
```

```