Five Number Summary in Python

In [58]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

In [59]:
```
## Define our dataset
dataset= [11,10,12,14,12,15,14,13,15,102,12,14,17,19,107, 10,13,12,14,12,108,12,11,14,13,15,10,15,12,10,14,13,15,10]
```

In [60]:
```
df = pd.DataFrame(dataset)
desc = df.describe()
print(desc)
```

```
               0
count   34.000000
mean    21.176471
std     26.768899
min     10.000000
25%     12.000000
50%     13.000000
75%     15.000000
max    108.000000
```

In [61]:
```
Q1, Q2, Q3 = np.percentile(dataset, [25, 50, 75])
print(Q1, Q2, Q3)
```

```
12.0 13.0 15.0
```

In [62]:
```
data_min, data_max = np.min(dataset), np.max(dataset)
print(data_min,data_max)
```

```
10 108
```

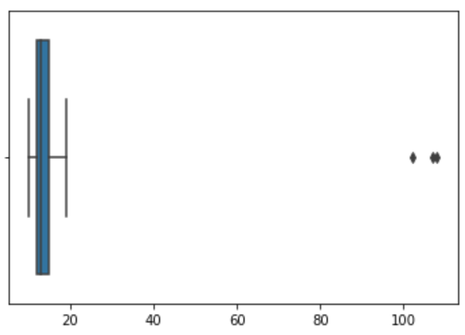
Visual Representation of Five Number Summary - Boxplot

A boxplot is a visual representation of a five number summary

In [63]:
```
sns.boxplot(dataset)
```

```
C:\Apps\Python37\lib\site-packages\seaborn\_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in an error or misinterpretation.
  FutureWarning
```

Out[63]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x22237e8bf98>
```



Detecting outliers in datsaet

1) Outliers are clearly seen in the boxplot 2) When it comes to outliers, we remove everything that isn't between a lower fence and an upper fence 3) That is: a) Sort the data b) Calculate Q1 aqnd Q3 c) IQR = (Q3-Q1) d) Find the Lower fence = Q1-1.5(IQR) e) Find the upper fence = Q3+1.5(IQR)

In [64]:
```
IQR = Q3 - Q1
lower_fence = Q1 - (1.5 * IQR)
upper_fence = Q3 + (1.5 * IQR)
print(lower_fence, upper_fence)
```

```
7.5 19.5
```

Clearly 102 and 107 are the outliers as shown in the boxplot and they do not lie between lower and upper fence