Statistical Measures

Theme

Part 1: Overview

In[13]:=
```
data = {5, 4, 10, 1, 5, 25};
```

Mode value (most common value in the set), nominal

In[14]:=
```
mode = Commonest[data]
```

Out[14]=
```
{5}
```

Average value, interval

In[15]:=
```
avg = Mean[data] // N
```

Out[15]=
```
8.33333
```

Outer quartiles (25 % and 75 %), ordinal or interval (different definitions exist)

In[16]:=
```
(* This uses the average between two values when the list size is even *)
quantile[data_, q_] := Quantile[data, q, {{1/2, 0}, {0, 1}}];
```

In[17]:=
```
{q25, q75} = quantile[data, {0.25, 0.75}]
```

Out[17]=
```
{4, 10}
```

Median value (50 % quantile), ordinal or interval

In[18]:=
```
q50 = quantile[data, 0.5]
```

Out[18]=
```
5.
```

Range, interval

In[19]:=
```
range = Max[data] - Min[data]
```

Out[19]=
```
24
```

Interquartile range (75 % quantile - 25 % quantile), interval

In[20]:=
```
IQR = InterquartileRange[data]
```

Out[20]=
```
6
```

Variance with $\frac{1}{n-1}$ scaling, interval

In[21]:=
```
var = Variance[data] // N
```

Out[21]=
```
75.0667
```

The skewness is a bit special in the way that multiple definitions exist. According to the script, we get

In[22]:=

$$\text{skewness} = \frac{1}{\text{Length[data]}} \sum_{i=1}^{\text{Length[data]}} \left(\frac{\text{data[[i]]} - \text{Mean[data]}}{\text{StandardDeviation[data]}} \right)^3 \text{ // N}$$

Out[22]=
```
1.04667
```

where the standard deviation is bias-corrected with $\frac{1}{n-1}$. Using the non-biased corrected standard deviation results in

In[23]:=
```
Skewness[data] // N
```
Out[23]=
```
1.37589
```

There is yet another definition based on the previous formula which applies a different bias correction

In[24]:=

$$\frac{\sqrt{\text{Length[data]} * (\text{Length[data]} - 1)}}{\text{Length[data]} - 2} \text{Skewness[data] // N}$$

Out[24]=
```
1.88401
```

However, in all cases we can say that the data is right-skewed (for more information see the Wikipedia article) and requires ratio-scaled data

Quartile skewness, ratio (more information)

In[25]:=
```
qSkewness =
  ((quantile[data, 0.75] - quantile[data, 0.5]) - (quantile[data, 0.5] - quantile[data, 0.25])) /
   (quantile[data, 0.75] - quantile[data, 0.25])
```
Out[25]=
```
0.666667
```

In summary, we get:

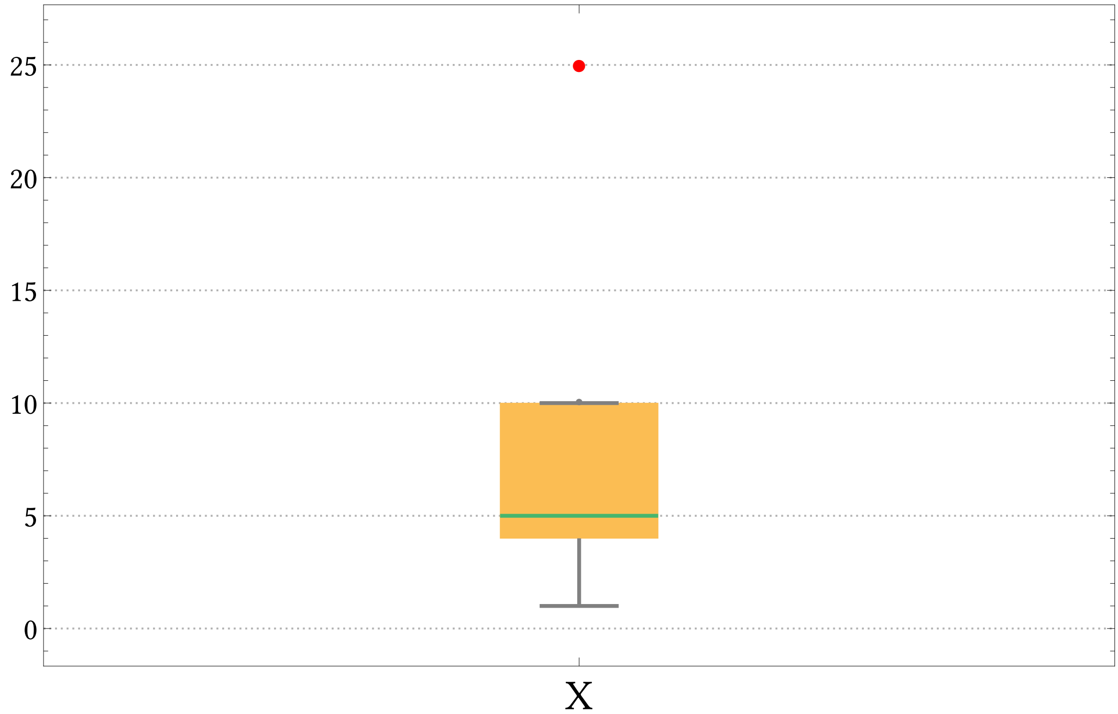
In[26]:=
```
TableForm[{
   {"Mode", "nominal", mode},
   {"Arithmetic mean", "interval", avg},
   {"Quantile 25 %", "ordinal or interval", q25},
   {"Median", "ordinal or interval", q50},
   {"Range", "interval", range},
   {"Interquartile range", "interval", IQR},
   {"Variance", "interval", var},
   {"Skewness", "ratio", skewness},
   {"Quartile skewness", "rato", qSkewness}
  }, TableHeadings → {None, {"Measure", "Requred scaling", "Value of measure for X"}}]
```
Out[26]//TableForm=

Measure	Requred scaling	Value of measure for X
Mode	nominal	5
Arithmetic mean	interval	8.33333
Quantile 25 %	ordinal or interval	4
Median	ordinal or interval	5.
Range	interval	24
Interquartile range	interval	6
Variance	interval	75.0667
Skewness	ratio	1.04667
Quartile skewness	rato	0.666667

Part 2: Box-and-Whisker Plot

In[27]:=
```
BoxWhiskerChart[data // N,
 {"Outliers", {"Outliers", Style["●", Red, FontSlant → Plain]},
  {"FarOutliers", "○"}, {"MedianMarker", ■}},
 Method → {"BoxRange" → (Flatten[{Min[#], quantile[#, {0.25, 0.5, 0.75}], Max[#]}] &)},
 PlotTheme → {"myTheme", "Detailed"},
 ChartLabels → {"X"},
 ChartStyle → Directive[Thick],
 FrameTicksStyle → {Directive[FontSlant → Plain, FontSize → 16], Automatic}
]
```

Out[27]=



What do the numbers show us?

- 1: 75 % quantile
- 2: median (50 % quantile)
- 3: 25 % quantile
- 6: interquartile range

Lower whisker (more information)

In[28]:=
```
Min[Cases[data, x_ /; x ≥ quantile[data, 0.25] - 1.5 * IQR]]
```

Out[28]= 1

Upper whisker

In[29]:=
```
Max[Cases[data, x_ /; x ≤ quantile[data, 0.75] + 1.5 * IQR]]
```

Out[29]= 10

The sign of the skewness g can be inferred from the plot by looking at the median value (green line). If it is below the centre line, the data is right-skewed (like here).

Part 3: Quartile Skewness

- The quartile skewness is bounded to [−1; 1]
- Only the sign matters (like for the normal skewness)

        - [−1; 0[left-skewed
        - 0 not-skewed at all (symmetric)
        -] 0; 1] right-skewed
- We can think of the median value sliding between $\tilde{x}_{0.25}$ and $\tilde{x}_{0.75}$, thus leaving
    - $a = -1$ if $\tilde{x}_{0.75} = \tilde{x}_{0.5}$
    - $b = 1$ if $\tilde{x}_{0.25} = \tilde{x}_{0.5}$

To achieve $g_Q = -1$, we must ensure that the 75 % quantile is the same as the median, e.g.

```
In[30]:= QuartileSkewness[{1, 2, 2}]
Out[30]= -1
```

Analog for +1, the 25 % quantile and the median must be identical, e.g.

```
In[31]:= QuartileSkewness[{2, 1, 1}]
Out[31]= 1
```

Part 4: Cartoons

Just some thoughts on the cartoons:

- Mean: strong influence of extreme values
- Median: says nothing about the range of the data or how it is distributed
- Mode: the most common value may be totally meaningless if there are lots of other values. Also, it does not say how often a value occurs. Does not take into account the skewness or range of the data.
- Range: no information how often a value occurs
- Correlation coefficient: strong influence of outliers (quadratic!)
- Variance: influence of outliers, different distributions can lead to the same variance