Cumulative frequency graphs

A cumulative frequency graph gives an additional statistical perspective on data presented in a *frequency table.* The cumulative frequency graph enables estimates to be made for the *median* and *upper and lower quartiles.* Combined with the data range and the *mean estimate*, these parameters can be used to construct a **box and whisker** plot. This is often a very useful summary of the key features of the data *distribution* revealed by a *histogram*.

Variable range	Frequency	Cumulative frequency	Cumulative frequency %
$x<20$	10	10	12
$x<30$	15	25	29
$x<60$	30	55	65
$x<120$	30	85	100

The cumulative frequency is the number of data values *that are less than a particular value*

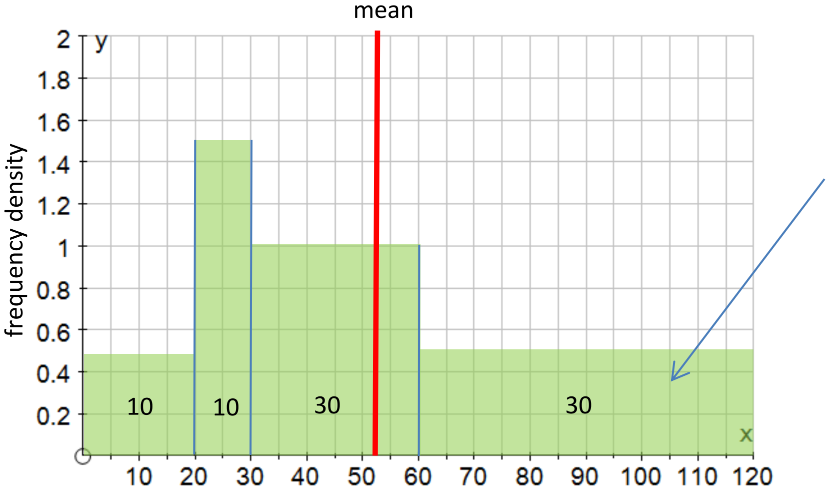
The **Lower Quartile** (LQ) corresponds to a cumulative frequency of **25%** of the total frequency.

The **Median** corresponds to **50%**

The **Upper Quartile** (UQ) corresponds to **75%**

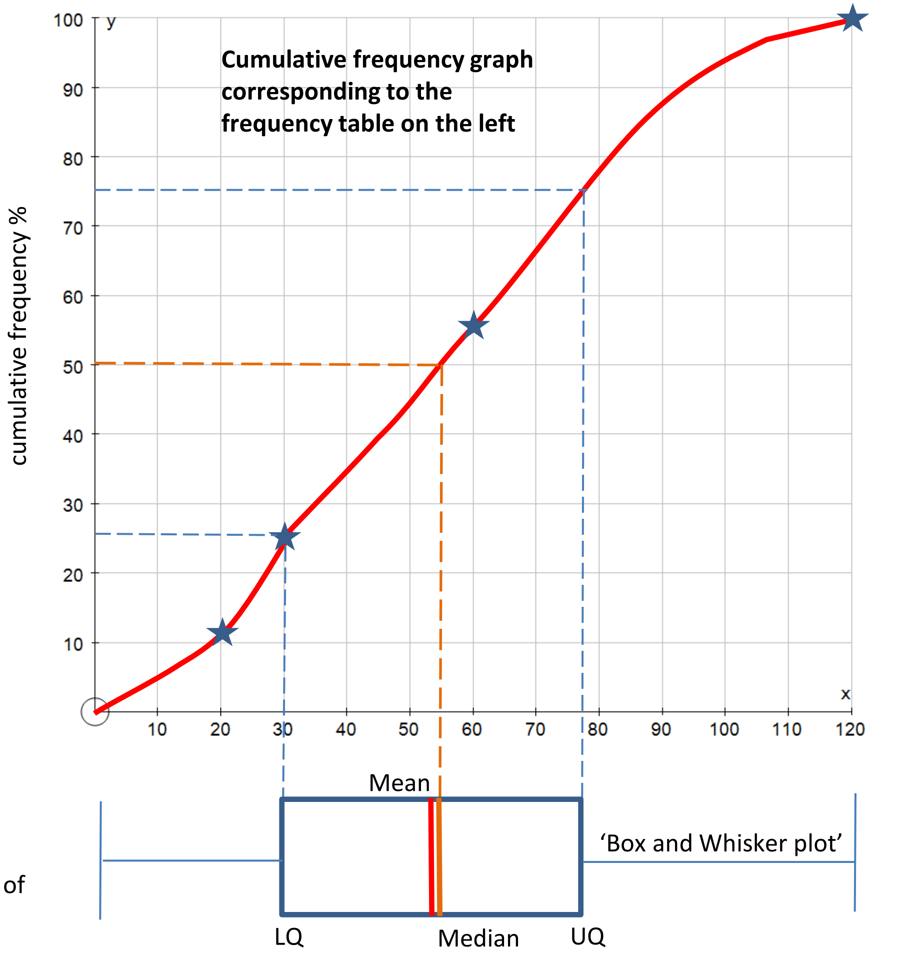
The **Inter-Quartile-Range** is the **UQ - LQ**

Histogram of the data in the table above. The mean estimate is 53.24



Total area of bars is the total frequency i.e. 85

Areas of each bar are given, which correspond to the frequency of measurements of the quantity x in the range associated with each bar.



LQ = 30 Mean = 53.24 Median = 54.5 UQ = 78 IQR = 48

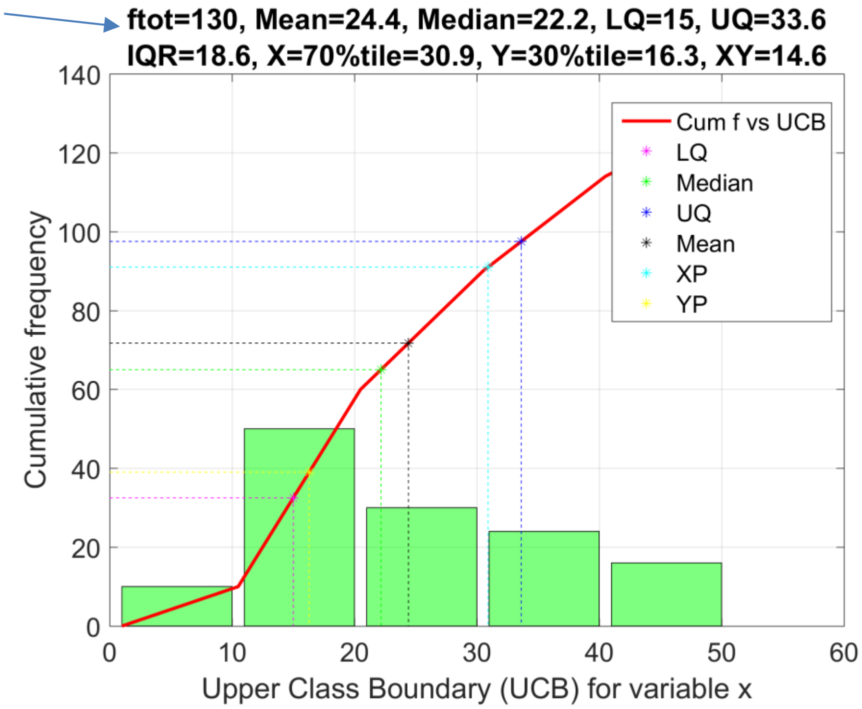
A high IQR implies a high degree of spread in the data. A significant difference between mean and median gives clues to the symmetry of the distribution, and can be a useful guide interpreting the histogram.

Cumulative frequency graphs with rounded variable ranges

Sometimes a frequency table will have *gaps*, due to a **variable being rounded to the nearest integer**. In this case, define the ***Upper Class Limits*** **(UCB)** to be half-way in the gaps, to enable a cumulative frequency graph to be plotted. Unless some underlying smooth-curve model is proposed for such a graph, *linear sections* are an appropriate model between data points of cumulative frequency vs UCBs. This allows for **LQ**, **median**, **UQ** to be estimated via ***linear interpolation.***

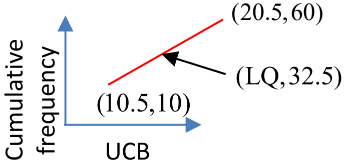
Variable range	Frequency	Cumulative frequency	UCB
$1 \le x < 10$	10	10	10.5
$11 \le x \le 20$	50	60	20.5
$21 \le x \le 30$	30	90	30.5
$31 \le x \le 40$	24	114	40.5
$41 \le x \le 50$	16	130	50.5

i.e. total frequency



Linear interpolation to find quartiles and percentiles

LQ when cumulative frequency = 0.25 x 130 = 32.5. This is in the second line segment of the cumulative frequency vs UCB *piecewise-linear* graph.



Since the *gradient* is constant in this line segment

$$\frac{\text{LQ}-10.5}{32.5-10} = \frac{20.5-10.5}{60-10}$$

$$\therefore \boxed{\text{LQ}=15}$$

Repeat the same method to find the **median** (when cumulative frequency = 0.5 x 130 = 65), **UQ** (when cumulative frequency = 0.75 x 130 = 97.5) and indeed any other 'percentile', such as 30% and 70% as shown in the graph.

To estimate the **mean**, use the middle of the UCBs, *with exception of the highest one*. (i.e. *don't add 0.5 to this*).

Example using the frequency table on this page:

i.e. not 50.5

$$\bar{x} \approx \frac{\tfrac{1}{2}(1+10.5)(10)+\tfrac{1}{2}(10.5+20.5)(50)+\tfrac{1}{2}(20.5+30.5)(30)+\tfrac{1}{2}(30.5+40.5)(24)+\tfrac{1}{2}(40.5+50)(16)}{130} = \boxed{24.4}$$