Cumulative frequency graphs

When you only know how many values fall in each class, a rising graph of running totals tells you what share lies below any value, and which value has a given share below it.

WORDS TO KNOW

Cumulative frequency: a running total. The count of a class plus the counts of every class below it.

Cumulative relative frequency graph: the same graph with each height divided by n, so its scale runs from 0% to 100%.

pth percentile: the value with p percent of the data at or below it. The median and the quartiles Q_1 and Q_3 are the 50th, 25th and 75th percentiles.

Cumulative frequency graph (ogive): a graph that shows, above each value, how many data values are less than it. Say "OH-jive".

Percentile rank: the percent of the data at or below a value. For measured data such as times, "at or below" and "less than" agree, so you read it off the graph.

KEY IDEAS

1. Plot each running total at the upper boundary of its class

A running total counts every value in its class and below, and all of them are less than the class's upper boundary. So that is where the point goes. A value on a boundary goes in the class to its right.

Plot a point at height 0 above the lowest boundary. Then plot each running total above its class's upper boundary, and join the points with straight segments.

A class with f values and width w climbs with slope $\frac{f}{w}$. A steep segment is a crowded class, a flat one is an empty class. The graph never goes down.

2. Read across a class as if its values were spread evenly

A straight segment assumes the values in a class are spread evenly across it. Take a class with lower boundary L, width w and f values, with C_{below} values below L.

Value to percent: the count below x is about

$$C_{\text{below}} + \frac{x - L}{w} \cdot f$$

Divide by n to get the percentile rank.

Percent to value: the target count is $\frac{p}{100} \cdot n$. Find the class where the running totals pass it, then

$$x \approx L + \frac{\text{target} - C_{\text{below}}}{f} \cdot w$$

3. Exact at a boundary, an estimate in between

At a boundary the height is a count straight from the table, so it is exact. Between boundaries every reading is an estimate. Read backward too: the climb across a class is its count, or on a percent scale its share of n.

Free sample from math.wiki for classroom use. Do not post. math.wiki

WATCH OUT

- **Do not plot at the midpoint.** Not all the values of a class are below its middle, so the running total is only true at the upper boundary.
- **The target count is a share of n.** For the pth percentile with n values, go across from $\frac{p}{100} \cdot n$ on a count scale, not from p.
- **Steepest is not always biggest.** With classes of different widths, the steepest segment is the most crowded class per unit of width, which may not hold the most values.
- **Wide classes give rough answers.** Merging classes leaves the evenly spread assumption more room to be wrong.

WORKED EXAMPLES

Example 1. A library timed 60 study visits, in minutes. Find the running totals, the points of the cumulative frequency graph and the steepest segment.

Visit (minutes)	0 to 30	30 to 60	60 to 90	90 to 120	120 to 150
Count	3	15	24	12	6

1. Add each count to the total before it: 3, then 3 + 15 = 18, 18 + 24 = 42, 42 + 12 = 54 and 54 + 6 = 60. The last total is all 60 visits, which checks the adding.
2. Start at $(0, 0)$, since no visit is shorter than 0 minutes, then use the upper boundaries. All widths are 30, so the steepest segment has the largest count, 24.

Answer: points $(0, 0)$, $(30, 3)$, $(60, 18)$, $(90, 42)$, $(120, 54)$, $(150, 60)$. The steepest segment is 60 to 90 minutes, slope $\frac{24}{30} = 0.8$ visits per minute.

Example 2. For the same visits, estimate the percentile rank of a 135 minute visit, and how many visits lasted more than 100 minutes.

1. 135 is halfway through the 120 to 150 class. Spread evenly, half of its 6 visits, 3, are below 135. With the 54 below 120, that is 54 + 3 = 57 visits, and 57 out of 60 is 95%.
2. 100 is a third of the way through the 90 to 120 class, so a third of its 12 visits, 4, are below 100. That makes 42 + 4 = 46 visits below 100, so 60 − 46 = 14 are longer.

Answer: about the 95th percentile, and about 14 visits longer than 100 minutes. Both are estimates.

Example 3. For the same visits, estimate the median, Q_1, Q_3 and the interquartile range.

1. Targets: $0.25 \cdot 60 = 15$, $0.50 \cdot 60 = 30$ and $0.75 \cdot 60 = 45$ visits.
2. Place each target with the running totals. 15 lies between 3 and 18: 12 visits into the 15 of 30 to 60. 30 lies between 18 and 42: 12 visits into the 24 of 60 to 90. 45 lies between 42 and 54: 3 visits into the 12 of 90 to 120.
3. Split each class evenly.

$$Q_1 \approx 30 + \frac{12}{15} \cdot 30 = 54 \qquad M \approx 60 + \frac{12}{24} \cdot 30 = 75 \qquad Q_3 \approx 90 + \frac{3}{12} \cdot 30 = 97.5$$

Answer: median about 75 minutes, $Q_1 \approx 54$, $Q_3 \approx 97.5$, interquartile range about 97.5 − 54 = 43.5 minutes.
Check: the median lies between the quartiles, 54 < 75 < 97.5.

Free sample from math.wiki for classroom use. Do not post. math.wiki