The $cum\sqrt{f}$ method of forming strata

The goal of stratification is to reduce the variance of the estimated population mean as much as possible. A method often used for forming strata is called the $cum\sqrt{f}$ method. It is an algorithm that approximates achieving the minimum variance of the estimated mean $\bar{y}_{st}$ by minimizing $\sum_{h=1}^{H} W_h S_h$ where H is the number of strata. And, it is easier to implement. The population units selected with certainty are not included in this process since they have their own stratum.

The description of the $cum\sqrt{f}$ method below follows Cochran (1977) *Sampling Techniques*, p 129 – 130.

To form strata for a population y_i, one needs to know the range of values in each strata. For example, the values in stratum 1 are greater than the lowest value in the population and less than or equal to the largest value in the stratum. If we have 5 strata (excluding certainty units), we can define the range in each stratum by their boundaries: $y_0, y_1, y_2, y_3, y_4, y_5$ where y_0 = the smallest value in the population and y_5 = the largest value in the population. For example, the values in stratum 1 are greater than the lowest value in the population, y_0, and less than or equal to the largest value in the stratum, y_1.

To implement the $cum\sqrt{f}$ algorithm, one first needs to divide the values of the population into equal ranges. The easiest way to do this is in terms of percentages of the population total of the variable used in designing the strata, for example 0 to 5% of the total, 5% to 10% of the total and so on until the last range is 95% to 100% of the total.

Then f for a range in values = the frequency, or the number, of the population units in the range.

And, $cum\sqrt{f}$ = the cumulative value of the square root of f.

Example. The data in Table 5A.11 below show the frequency distribution of the percentage of bank loans devoted to industrial loans in a population of 13,435 banks in the U.S. (McEvoy 1956). The distribution is skewed with its mode at the lower end. The 1st row on the left has the range 0 to 5% of the total loan amount and shows that 3464 banks have 0 – 5% of their loans devoted to industrial loans so

$$cum\sqrt{f} = \sqrt{3464} = 58.9.$$

The 2nd row on the left has the range 5% to 10% of the total loan amount and shows that 2516 banks have 5 – 10% of their loans devoted to industrial loans so

$$cum\sqrt{f} = \sqrt{3464} + \sqrt{2516} = 58.9 + 50.2 = 109.1$$

TABLE 5A.11

CALCULATION OF STRATUM BOUNDARIES BY THE CUM $\sqrt{f(y)}$ RULE

Industrial Loans / Total Loans %	$f(y)$	Cum $\sqrt{f(y)}$	Industrial Loans / Total Loans %	$f(y)$	Cum $\sqrt{f(y)}$
0–5	3464	58.9	50–55	126	340.3
5–10	2516	109.1	55–60	107	350.6
10–15	2157	155.5	60–65	82	359.7
15–20	1581	195.3	65–70	50	366.8
20–25	1142	229.1	70–75	39	373.0
25–30	746	256.4	75–80	25	378.0
30–35	512	279.0	80–85	16	382.0
35–40	376	298.4	85–90	19	386.4
40–45	265	314.7	90–95	2	387.8
45–50	207	329.1	95–100	3	389.5

Once Table 5A.11 is completed, then one finds the value that divides the population into equal intervals based on $cum\sqrt{f}$.

Suppose we want 5 strata. Since the total of $cum\sqrt{f}$ = 389.5, we divide 389.5/5 = 77.9 to determine the length of the ranges of $cum\sqrt{f}$. Therefore, the division points that determine the stratum ranges are

0, 77.9, 155.8 (2*77.9), 233.7 (3*77.9), 311.6 (4*77.9), and 389.5.

Then we choose groupings from Table 5A.11 that match these breakpoints as close as possible. For example, 77.9 is between 58.9 and 109.1 but closer to 58.9 so we choose 58.9 as the largest value for the 1st stratum.

The stratum intervals are shown below:

	Stratum				
	1	2	3	4	5
Boundaries	0–5%	5–15%	15–25%	25–45%	45–100%
Interval on cum $\sqrt{f}$	58.9	96.6	73.6	85.6	74.8

The first two strata with the intervals have ranges of $cum\sqrt{f}$ that are different from the ranges for the other three strata. However, a more equal distribution of $cum\sqrt{f}$ to the five strata is not possible without a finer subdivision of the original 5% intervals in Table 5A.11.