**Methods for Deriving Cutoffs to form Strata**

**Contents**

1. Cumulative Method of forming strata (Equal Sum of Values in Strata) p. 1
2. The method of forming strata p.2
3. Equal *WhSh* per stratum method p.4

**Cumulative Method of forming strata (Equal Sum of Values in Strata)**

Do not include elements in certainty strata. Sort the data in ascending order. Calculate the sum of all the elements. Divide this sum by the number of strata to get the length of the strata interval. Create the cumulative distribution of values of the elements. Assign the elements to strata based on the strata interval and cumulative distribution.



**The 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 method. It is an algorithm that approximates achieving the minimum variance of the estimated mean by minimizing where H is the number of strata. It is easier to implement then the Equal *WhSh* per stratum method.

The population units selected with certainty are not included in this process since they have their own stratum.

The description of the method below follows Cochran (1977) *Sampling Techniques*, p 129 – 130.

To form strata for a population *yi*, one needs to know the range of values in each strata. For example, the values in stratum 1 are greater than or equal to 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: *y0, y1, y2, y3, y4, y5*  where *y0* = the smallest value in the population and *y5* = the largest value in the population. For example, the values in stratum 1 are greater than or equal to the lowest value in the population, *y0,* and less than or equal to the largest value in the stratum, *y1.*

To implement the 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.

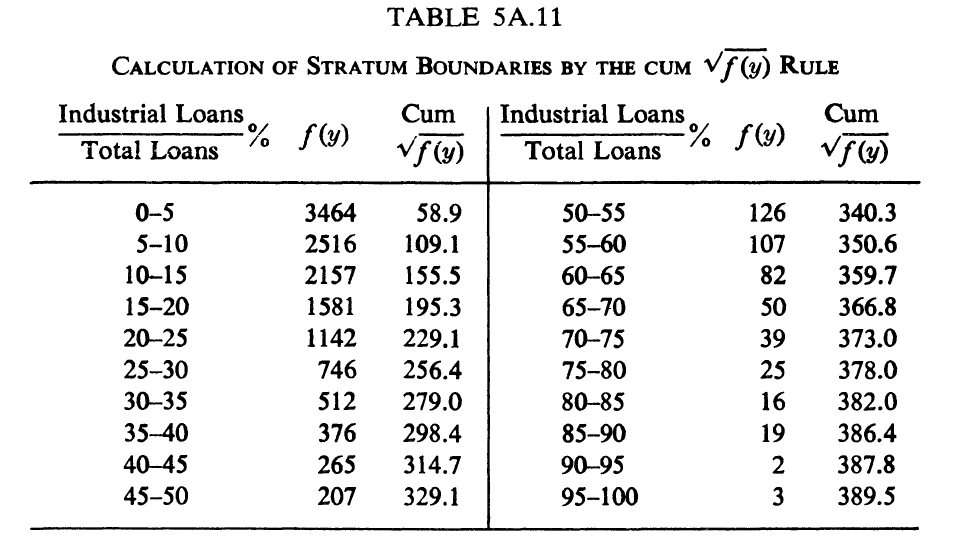
Then for a range in values = the frequency, or the number, of the population units in the range.

Note, = 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:

.

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



Once Table 5A.11 is completed, then one finds the value that divides the population into equal intervals based on .

Suppose we want 5 strata. Since the total of = 389.5, we divide 389.5/5 = 77.9 to determine the length of the ranges of . 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 the1st stratum. The stratum intervals are shown below:

The first two strata with the intervals have ranges of that are substantially different from the ranges for the other three strata. However, a more equal distribution of to the five strata is not possible without a finer subdivision of the original 5% intervals in Table 5A.11. [[1]](#footnote-1) Do not include elements in certainty strata in the analysis.

**Equal *WhSh* per stratum method**

**Notation:**

*N* = number of elements in the population

*Nh* = number of elements in the population from strata h

*n* = number of elements in the entire sample

*nh* = number of elements in the sample from strata h

*H* = number of strata

*Sn* = standard deviation of the elements in strata h from the population

*Wh = Nh/N*

The objective is to find samples sizes for strata such that *k = WhSh* is constant for all strata.

Note: *k = (Nh/N)(Sh)* (1)

Such that: *Nh=(k/ Sh)(N)* (2)

Taking sums we obtain: (3)

Solving for k gives: *k =* (4)

Substitute for k in (2) to derive *Nh*

But, from (4) we do not know *Sh*. We can initialize *Sh*by picking strata with another method such as the cumulative method. We then find *k* and calculate *Nh* for all the strata. From these strata we calculate a new set of *Sh*. We keep iterating by finding a new *k* as shown in equation (4) until the change in the number of elements in each strata is less than some tolerance. A stopping rule could be, stop iterating if the change in the number of elements in each stratum is less than 5%.

1. If the initial class intervals in the original distribution of the data are of unequal length, then a slight change is needed. When the interval changes from one of length d to one of length *ud* the value of for the second interval is multiplied by when forming cum . In the example shown here the initial class intervals are of the same length. [↑](#footnote-ref-1)