Appendix G

Optimal Stratification

Contents:

CSRF- Equal Cell Width (pages 2-4)

CSRF- Unequal Cell Width (pages 5-10)

In using the invoice or recorded amounts to stratify, there is a method that provides "optimal" strata boundaries that is most often used in practice. There are two methods for optimal stratification. One method divides the frame into cells of equal size or "width". Calculations are made to combine these cells into strata by establishing break values, or strata boundaries. The State of New York, in their sampling manual, has an example of these calculations which is reprinted here.

Others take the basic method used by New York, and add an additional variable, requiring an additional calculation. The size, or width, of the cells is varied throughout the sampling frame. The Washington State DOR has an example of this in their Statistical Sampling Manual, which is also reprinted here.

Either method is often referred to as Cumulative Square Root of the Frequency (CSRF) or some similar name. In some cases, it is referred to as "optimal" stratification.

CSRF- Equal Cell Width [1]

Of the various methods available to determine stratum boundaries, New York State has opted to use the *cumulative of the square root of the frequency* method. This method is one that is well documented in many statistics textbooks and one that is well suited for determining stratum boundaries for accounting populations. This method allows for greater efficiency in the sampling process compared with setting the boundaries using judgment only. Using judgment in setting up stratum boundaries is a perfectly acceptable method, but may not be the most efficient in many cases. The department is committed to performing audits using the most efficient and accurate methods available.

Calculation of stratum boundaries using the *cumulative of the square root of the frequency* method

The table below illustrates how stratum boundaries are determined using the cumulative of the square root of the frequency method.

Step 1: Evaluate the population and determine if transactions below a certain dollar level should be eliminated from being sampled based on the dollar significance of those transactions. In this example, no transactions under $100 will be reviewed. This is a decision that is made using auditor judgment and has no statistical significance.

Step 2 Evaluate the population and determine what the high dollar cutoff will be for detailed review. In this example (on page 18), all items in the amount of $10,000 and up will be reviewed in detail.

Step 3 Stratify the remaining transactions within the population of items to be sampled ($100 - $9999.99) using a set dollar interval between the strata ($500). We use $500 for the current example to limit the number of strata which need to be displayed. In actuality, $100 intervals are used.

Step 4 Determine the frequency $f(y)$ for each dollar range. This is nothing more than the number of transactions within each dollar range.

[1] Pages 2-4 Taken from the New York State Department of Taxation and Finance's publication 132; *Computer-Assisted Audits – Guidelines and Procedures for Sales Tax Audits*.

Step 5 Calculate the square root of the frequency for the first sampled range (100 - 499.99), which calculates out to be 146.5. Now calculate the square root for the next sampled range (500 - 999.99), which calculates out to be 94.1. Since we are accumulating this column, each result gets added to the number before it. The cumulative of the first two strata equals the 146.5 + 94.1 which equals the 240.6. Continue this process for each of the sampled strata.

Step 6 Once the *cumulative square root of the frequency* for the last sampled stratum (9500 - 9999.99 in this case) is calculated, the strata can be determined. For the final result, 770.1 is divided by the number of sampled strata desired (6 in this case, but can vary between 3-9 sampled strata). This result is the test interval of 128.4.

Step 7 The first strata range is determined by comparing the test interval of 128.4 against the calculated cumulative square root of the frequency for each dollar range and selecting the range that is closest to this test interval amount. In the immediate instance, 128.4 is closest to the 146.5 for the first sampled strata (100 - 499.99). The second stratum is determined by taking 2 X 128.4 and finding the closest range. Strata 2 will encompass the range of 500 - 999.99 since 240.6 is the closest interval to 256.8. The third stratum is determined by taking 3 X 128.4 and finding the closest range. Strata 3 will encompass the range of 1000 - 2499.99 since 401 is the closest interval to 385.2. This process will continue until all 6 sampled strata are determined.

Dollar Range (interval)	Number of items frequency f(y)	Cumulative Square root of the frequency $\sum\sqrt{f(y)}$	Strata #	Strata range
0 - 99.99	20090			***Eliminated***
100 - 499.99	21472	***146.5***	***1***	***100 - 499.99***
500 - 999.99	8850	***240.6***	***2***	***500 - 999.99***
1000 -1499.99	4149	305.0	***3***	***1000 - 2499.99***
1500 - 1999.99	2637	356.4		
2000 - 2499.99	1990	***401.0***		
2500 - 2999.99	1620	441.2	***4***	***2500 - 3999.99***
3000 - 3499.99	1815	483.8		
3500 - 3999.99	1056	***516.3***		
4000 - 4499.99	879	546.0	***5***	***4000 - 6499.99***
4500 - 4999.99	627	571.0		
5000 - 5499.99	598	595.5		
5500 - 5999.99	520	618.3		
6000 - 6499.99	425	***638.9***		
6500 - 6999.99	471	660.6	***6***	***6500 - 9999.99***
7000 - 7499.99	458	682.0		
7500 - 7999.99	429	702.7		
8000 - 8499.99	358	721.6		
8500 - 8999.99	241	737.2		
9000 - 9499.99	275	753.7		
9500 - 9999.99	269	***770.1***		
10000 and up	234		***7***	***Detail***

The population profile below summarizes the results from using the square root of the frequency method to determine the strata boundaries.

Population Profile

Strata #	Strata range	Number of items	Population $'s	Standard deviation
1	0 - 99.99	20,090	738,807	-
2	100 - 499.99	21,472	5,432,668	109.89
3	500 - 999.99	8,850	6,368,355	144.67
4	1000 - 2499.99	8,776	14,041,520	431.59
5	2500 - 3999.99	4,491	14,103,433	411.27
6	4000 - 6499.99	3,049	15,469,958	715.53
7	6500 - 9999.00	2,501	20,025,062	1010.26
8	10000 and up	234	6,319,170	-

CSRF- Unequal Cell Width[2]

Once the sampling frame is established, there are several ways to determine stratum breaks. In CAAP[3] audits, the "Cumulative Frequency of the Square Root" is the preferred method to establish strata boundaries.

Background

Stratification boundaries can be determined using several approaches. The most common method is CSRF. This methodology has universal acceptance in sampling. There are many references for this methodology.[4] The CSRF methodology has been attributed to an article appearing in the Journal of American Statistical Association in 1959.[5]

It would be optimal to establish strata boundaries based on the error amounts. Since the error amount is unknown, populations are stratified using the invoice amounts.[6] CSRF will distribute the variation of the invoice amounts in an "optimal" manner to the strata.

CSRF methodology breaks down the population into intervals. These intervals are sometimes referred to as "cells" which can be of equal or unequal width. In most references, the method is described using cells of equal width. When broken into unequal widths, an extra step

---

[2] Pages 4-10 taken from the Washington State Department of Revenue publication; *Statistical Sampling Manual*, Appendix F. Footnotes 4-7 are also directly from the publication.

[3] In Washington, their computer assisted audit unit is referred to as "CAAP".

[4] The following are some of the references:

1. *Statistical Auditing*, Donald Roberts, AICPA, 1978, pages 97 to 98.
2. *Sampling Techniques*, William Cochran, John Wiley and Sons, 1977, pages 127-131.
3. *Applications of Statistical Sampling to Auditing*, Alvin Arens & James K Loebbecke, Prentice Hall, Englewood CA 1981, pages 256 – 258.
4. IRS training publication: Advanced *Statistical Sampling, Training* 3174-002 Rev (05-92) TPDS 87030A in Lesson 10, pages 6 – 10.
5. *Survey Sampling*, Leslie Kish, John Wiley & Sons, 1995, pages 104 –106.
6. *Sampling of Populations: Methods and Applications*, 3rd Edition, Paul Levy and Stanley Lemeshow, John Wiley & Sons, New York, 1999, pages 179 - 183.
7. *Elementary Survey Sampling*, 5th edition, Richard Scheaffer, William Mendenhall, R. Lyman Ott, Duxbury Press, New York, 1996, page 165.

[5] According to *Elementary Survey Sampling*, 5th edition, Richard Scheaffer et al. footnote13, page 189, CFSR was developed by T. Dalenius and J.L. Hodges Jr.

[6] Generally, there is a relationship between error and invoice amounts. We know that the variation in the errors has something to do with the range of values in the non-zero error amounts. However, the biggest contributor to the variation of errors is the number of zero errors as percentage of the total (the more zero errors, the more variable the population). The number of zero errors in the population will not have anything to do with variation in values of invoice amounts. CFSR will not consider the effect of the zeros.

is involved.[7] CAAP generally uses cells of unequal width because the stratum breaks become more exact. Note that CSRF does not dictate the number of strata, only the breaks between the strata.

Mechanics

The following is a sample of the procedures used to determine stratum breaks assuming cells of unequal width.

Step 1 Evaluate the population and determine if transactions below a certain dollar level should not be sampled based on the dollar significance of those transactions.

In this example, the taxpayer and the auditor agreed on the accounts of interest and the sample population. No transactions equal to or less than $100.00 will be reviewed.

Step 2 Evaluate the population and determine the dollar amounts that will be reviewed on an actual basis.

All items greater than $25,000.00 will be reviewed in detail (actual basis).

Step 3 Stratify the remaining transactions within the population of items to be sampled ($100.01 - $25,000.00) into dollar ranges.

Step 4 Determine the frequency for each dollar range. This is the number of transactions within each range.

Step 5 Calculate the square root of the frequency (number of transactions) for the first range ($100.01 – $150.00). This value is 129.82. Then calculate the square root of the next range ($150.01 – $200.00). This value equals 101.01. Continue this process for each of the ranges.

Step 6 Calculate the square root of the range width for the first range ($150.00 – $100.01). This value is 7.07. Then calculate the square root of the next range

[7] In *Statistical Auditing*, Roberts describes the methodology of unequal lengths in a footnote on page 98. In the book by Cochran, *Sampling Techniques*, he describes this modification on page 130.

($200.00 – $150.00). Again, this value equals 7.07. Continue this process for each of the ranges.

Step 7 For each range, multiply the square root of the frequency (step 5) times the square root of the range (step 6).

Step 8 Accumulate the values (as calculated in step 7) for each range. For the first range, the cumulative square root value is 917.87. The cumulative value for the second range is 1,632.15 (714.28 + 917.87 - accumulated value for the preceding range). For each range, add the square root value (step 7) to the preceding square root value (step 7) to arrive at the cumulative square root value. Continue this process for each of the ranges.

Step 9 Divide the total cumulative square root value (18,422.66) by the number of strata (4 in this audit) to arrive at cumulative square root value for each stratum. In this audit the value is 4,605.67.

Step 10 Multiply the result of step 9 by 1 to calculate the first stratum range, or 4,605.67. Locate the cumulative square root value closest to this amount. In this case, it is 4,623.62. Locate the ending range associated with this value ($500.01 - $550.00). The first stratum range (boundary) is $100.01 – $550.00.

Step 11 Multiply the result of step 9 by 2 to calculate the second stratum range, or 9,211.33. Locate the cumulative square root value closest to this amount. In this case, it is 9,213.44. Locate the ending range associated with this value ($2,100.01 – $2,200.00). The second stratum range (boundary) is $550.01 – $2,200.00.

Step 12 Multiply the result of step 9 by 3 to calculate the third stratum range, or 13,817.00. Locate the cumulative square root value closest to this amount. In this case, it is 13,871.18. Locate the ending range associated with this value ($8,250.01 – $8,500.00). The third stratum range (boundary) is $2,200.01 – $8,500.00.

Step 13 Multiply the result of step 9 by 4 to calculate the fourth stratum range, or 18,422.66. Locate the cumulative square root value closest to this amount. In this case, it is 18,422.66. Locate the ending range associated with this value ($20,000.01 – $25,000.00). The fourth stratum range (boundary) is $8,500.01–$25,000.00.

Step 14 Calculate the coefficient of variation (standard deviation divided by the mean) for each of the strata (4 strata in this audit). Compute the average coefficient of variation for all the strata. An average coefficient of variation of less than 50% is the target range. In this example, the combined coefficient of variation is 40.20%.

Step 15 If the average coefficient of variation is greater than 50%; consideration should be given to adding another stratum. On the other hand, if the percentage is significantly less than 50%, then the total number of strata may be reduced.

Example

Attached is an example of a workpaper applying the "Cumulative Square Root of the Frequency" method:

Cumulative Square Root of the Frequency - Strata Boundary Determination
CSRF-Unequal

Range Beginning Amount	Range Ending Amount	Range Width Value	Range Count (#)	Square Root of the Count (#)	Square Root of the Width (Value)	Sum of the Square Root Count (#) * Width (Value)	Cumulative Square Root Values Count (#) * Width (Value)	Stratum
100.01	150.00	49.99	16853	129.82	7.07	917.87	917.87	1
150.01	200.00	50.00	10204	101.01	7.07	714.28	1,632.15	1
200.01	250.00	50.00	7046	83.94	7.07	593.55	2,225.70	1
250.01	300.00	50.00	5476	74.00	7.07	523.26	2,748.96	1
300.01	350.00	50.00	4339	65.87	7.07	465.78	3,214.74	1
350.01	400.00	50.00	4016	63.37	7.07	448.11	3,662.84	1
400.01	450.00	50.00	2352	48.50	7.07	342.93	4,005.77	1
450.01	500.00	50.00	2010	44.83	7.07	317.02	4,322.79	1
500.01	550.00	50.00	1810	42.54	7.07	300.83	4,623.62	1

Cumulative Square Root of the Frequency - Strata Boundary Determination

CSRF-Unequal

Range Beginning Amount	Range Ending Amount	Range Width Value	Range Count (#)	Square Root of the Count (#)	Square Root of the Width (Value)	Sum of the Square Root Count (#) * Width (Value)	Cumulative Square Root Values Count (#) * Width (Value)	Stratum
550.01	600.00	50.00	1639	40.48	7.07	286.27	4,909.89	2
600.01	650.00	50.00	1198	34.61	7.07	244.74	5,154.64	2
650.01	700.00	50.00	1174	34.26	7.07	242.28	5,396.92	2
700.01	750.00	50.00	1008	31.75	7.07	224.50	5,621.42	2
750.01	800.00	50.00	1255	35.43	7.07	250.50	5,871.92	2
800.01	850.00	50.00	717	26.78	7.07	189.34	6,061.26	2
850.01	900.00	50.00	873	29.55	7.07	208.93	6,270.18	2
900.01	950.00	50.00	615	24.80	7.07	175.36	6,445.54	2
950.01	1,000.00	50.00	611	24.72	7.07	174.79	6,620.33	2
1,000.01	1,050.00	50.00	574	23.96	7.07	169.41	6,789.74	2
1,050.01	1,100.00	50.00	424	20.59	7.07	145.60	6,935.34	2
1,100.01	1,150.00	50.00	388	19.70	7.07	139.28	7,074.62	2
1,150.01	1,200.00	50.00	347	18.63	7.07	131.72	7,206.34	2
1,200.01	1,250.00	50.00	417	20.42	7.07	144.40	7,350.74	2
1,250.01	1,300.00	50.00	493	22.20	7.07	157.00	7,507.74	2
1,300.01	1,350.00	50.00	312	17.66	7.07	124.90	7,632.64	2
1,350.01	1,400.00	50.00	271	16.46	7.07	116.40	7,749.04	2
1,400.01	1,450.00	50.00	230	15.17	7.07	107.24	7,856.28	2
1,450.01	1,500.00	50.00	316	17.78	7.07	125.70	7,981.98	2
1,500.01	1,600.00	100.00	312	17.66	10.00	176.64	8,158.62	2
1,600.01	1,700.00	100.00	421	20.52	10.00	205.18	8,363.80	2
1,700.01	1,800.00	100.00	457	21.38	10.00	213.78	8,577.57	2
1,800.01	1,900.00	100.00	288	16.97	10.00	169.71	8,747.28	2
1,900.01	2,000.00	100.00	350	18.71	10.00	187.08	8,934.36	2
2,000.01	2,100.00	100.00	210	14.49	10.00	144.91	9,079.28	2
2,100.01	2,200.00	100.00	180	13.42	10.00	134.16	9,213.44	2
2,200.01	2,300.00	100.00	193	13.89	10.00	138.92	9,352.37	3
2,300.01	2,400.00	100.00	117	10.82	10.00	108.17	9,460.53	3
2,400.01	2,500.00	100.00	199	14.11	10.00	141.07	9,601.60	3
2,500.01	2,600.00	100.00	98	9.90	10.00	98.99	9,700.59	3
2,600.01	2,700.00	100.00	202	14.21	10.00	142.13	9,842.72	3
2,700.01	2,800.00	100.00	133	11.53	10.00	115.33	9,958.05	3
2,800.01	2,900.00	100.00	169	13.00	10.00	130.00	10,088.05	3
2,900.01	3,000.00	100.00	203	14.25	10.00	142.48	10,230.52	3
3,000.01	3,100.00	100.00	114	10.68	10.00	106.77	10,337.30	3
3,100.01	3,200.00	100.00	94	9.70	10.00	96.95	10,434.25	3
3,200.01	3,300.00	100.00	99	9.95	10.00	99.50	10,533.75	3
3,300.01	3,400.00	100.00	71	8.43	10.00	84.26	10,618.01	3
3,400.01	3,500.00	100.00	82	9.06	10.00	90.55	10,708.56	3
3,500.01	3,600.00	100.00	91	9.54	10.00	95.39	10,803.96	3
3,600.01	3,700.00	100.00	63	7.94	10.00	79.37	10,883.33	3
3,700.01	3,800.00	100.00	142	11.92	10.00	119.16	11,002.49	3
3,800.01	3,900.00	100.00	80	8.94	10.00	89.44	11,091.94	3
3,900.01	4,000.00	100.00	93	9.64	10.00	96.44	11,188.37	3
4,000.01	4,100.00	100.00	120	10.95	10.00	109.54	11,297.92	3
4,100.01	4,200.00	100.00	117	10.82	10.00	108.17	11,406.08	3
4,200.01	4,300.00	100.00	96	9.80	10.00	97.98	11,504.06	3
4,300.01	4,400.00	100.00	45	6.71	10.00	67.08	11,571.14	3
4,400.01	4,500.00	100.00	59	7.68	10.00	76.81	11,647.96	3
4,500.01	4,600.00	100.00	30	5.48	10.00	54.77	11,702.73	3
4,600.01	4,700.00	100.00	49	7.00	10.00	70.00	11,772.73	3
4,700.01	4,800.00	100.00	38	6.16	10.00	61.64	11,834.37	3
4,800.01	4,900.00	100.00	23	4.80	10.00	47.96	11,882.33	3
4,900.01	5,000.00	100.00	93	9.64	10.00	96.44	11,978.77	3
5,000.01	5,250.00	250.00	66	8.12	15.81	128.45	12,107.22	3
5,250.01	5,500.00	250.00	64	8.00	15.81	126.49	12,233.71	3
5,500.01	5,750.00	250.00	99	9.95	15.81	157.32	12,391.03	3
5,750.01	6,000.00	250.00	121	11.00	15.81	173.93	12,564.96	3
6,000.01	6,250.00	250.00	63	7.94	15.81	125.50	12,690.46	3
6,250.01	6,500.00	250.00	91	9.54	15.81	150.83	12,841.29	3
6,500.01	6,750.00	250.00	168	12.96	15.81	204.94	13,046.23	3
6,750.01	7,000.00	250.00	64	8.00	15.81	126.49	13,172.72	3
7,000.01	7,250.00	250.00	63	7.94	15.81	125.50	13,298.22	3
7,250.01	7,500.00	250.00	47	6.86	15.81	108.40	13,406.61	3
7,500.01	7,750.00	250.00	48	6.93	15.81	109.54	13,516.16	3
7,750.01	8,000.00	250.00	82	9.06	15.81	143.18	13,659.34	3
8,000.01	8,250.00	250.00	23	4.80	15.81	75.83	13,735.17	3
8,250.01	8,500.00	250.00	74	8.60	15.81	136.01	13,871.18	3

Cumulative Square Root of the Frequency - Strata Boundary Determination

CSRF-Unequal

Range Beginning Amount	Range Ending Amount	Range Width Value	Range Count (#)	Square Root of the Count (#)	Square Root of the Width (Value)	Sum of the Square Root Count (#) * Width (Value)	Cumulative Square Root Values Count (#) * Width (Value)	Stratum
8,500.01	8,750.00	250.00	64	8.00	15.81	126.49	13,997.67	4
8,750.01	9,000.00	250.00	34	5.83	15.81	92.20	14,089.87	4
9,000.01	9,250.00	250.00	24	4.90	15.81	77.46	14,167.33	4
9,250.01	9,500.00	250.00	27	5.20	15.81	82.16	14,249.48	4
9,500.01	9,750.00	250.00	29	5.39	15.81	85.15	14,334.63	4
9,750.01	10,000.00	250.00	45	6.71	15.81	106.07	14,440.70	4
10,000.01	10,500.00	500.00	42	6.48	22.36	144.91	14,585.61	4
10,500.01	11,000.00	500.00	42	6.48	22.36	144.91	14,730.53	4
11,000.01	11,500.00	500.00	62	7.87	22.36	176.07	14,906.59	4
11,500.01	12,000.00	500.00	77	8.77	22.36	196.21	15,102.81	4
12,000.01	12,500.00	500.00	72	8.49	22.36	189.74	15,292.54	4
12,500.01	13,000.00	500.00	51	7.14	22.36	159.69	15,452.23	4
13,000.01	13,500.00	500.00	48	6.93	22.36	154.92	15,607.15	4
13,500.01	14,000.00	500.00	41	6.40	22.36	143.18	15,750.33	4
14,000.01	14,500.00	500.00	67	8.19	22.36	183.03	15,933.36	4
14,500.01	15,000.00	500.00	51	7.14	22.36	159.69	16,093.05	4
15,000.01	15,500.00	500.00	38	6.16	22.36	137.84	16,230.89	4
15,500.01	16,000.00	500.00	53	7.28	22.36	162.79	16,393.68	4
16,000.01	16,500.00	500.00	23	4.80	22.36	107.24	16,500.91	4
16,500.01	17,000.00	500.00	58	7.62	22.36	170.29	16,671.21	4
17,000.01	17,500.00	500.00	29	5.39	22.36	120.42	16,791.62	4
17,500.01	18,000.00	500.00	21	4.58	22.36	102.47	16,894.09	4
18,000.01	18,500.00	500.00	26	5.10	22.36	114.02	17,008.11	4
18,500.01	19,000.00	500.00	20	4.47	22.36	100.00	17,108.11	4
19,000.01	19,500.00	500.00	24	4.90	22.36	109.54	17,217.65	4
19,500.01	20,000.00	500.00	39	6.24	22.36	139.64	17,357.30	4
20,000.01	25,000.00	5,000.00	227	15.07	70.71	1065.36	18,422.66	4
			74506					

Cumulative Value	18,422.66
# of Strata	4
Amt per Stratum	4,605.67

STRATUM	AMT PER STRATUM	APPROX BREAK	COUNT	STD DEV	AVE	TOTAL VALUE	CV (SEE NOTE)	CV * TOTAL
1	4,605.67	4,605.67	54,106	116.58	236.44	12,792,823	49%	6,307,677
2	4,605.67	9,211.33	15,080	412.88	1,020.02	15,381,902	40%	6,226,230
3	4,605.67	13,817.00	3,986	1,679.82	4,203.43	16,754,872	40%	6,695,763
4	4,605.67	18,422.66	1,334	4,608.09	14,884.44	19,855,843	31%	6,147,192
			74,506			64,785,439		

Coefficient of Variation

AVE =>	40%

STRATUM	Range Beginning Amount	Range Ending Amount
1	100.01	550.00
2	550.01	2,200.00
3	2,200.01	8,500.00
4	8,500.01	25,000.00