Sampling Modenization: Substratification

Kolokium Statistik 2015
Bahagian Metodologi dan Penyelidikan

Tempat : Institut Latihan Statistik Malaysia
Tarikh : 27 & 28 Oktober 2015

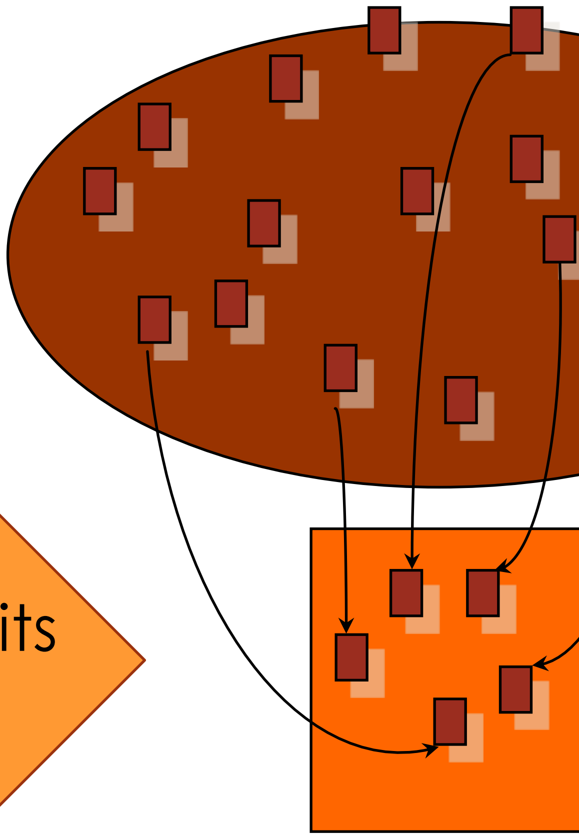
Manisah Othman (manisah@stats.gov.my)
Nor Shahida Jamil (shahida.jamil@stats.gov.my)
Nik Zawahir Nik Yusoff (zawahir@stats.gov.my)
Hafizah Tajudin (hafizah.tajudin@stats.gov.my)

Sampling

Population

Sampling frame

Element for:
- inference/conclusion



Sample

Sampling process
- selecting sampling units
- representative the population

Overview of the population for inference or generalization the population

Stratification

Population can be constructing into strata:

- homogeneous,
- the sampling unit selected from each strata proportionate to the size of the strata.

Sampling process

- Identify VOI
- Identify data distribution
- Designing stratification
- Calculate sample size

Stratification:

Cut-off & Neyman method

and

Stratification package - R

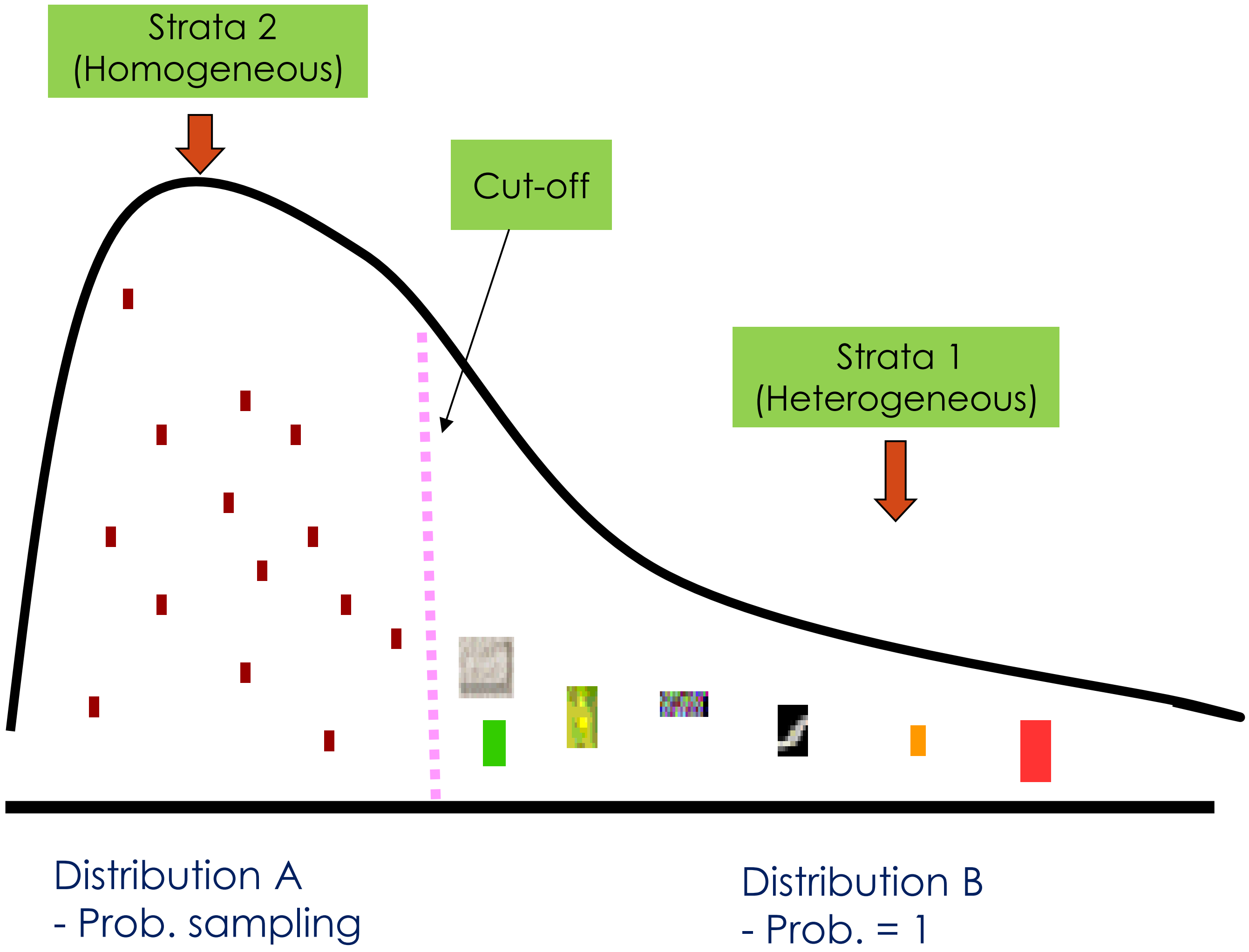
Annual Manufacturing Survey

Stratification – cut-off & Neyman method

10203 – Production of fish meals for human consumption or animal feed

Obs	ID	Output
1	000000436181	54,438,247
2	000000498672	8,616,263
3	000000463440	7,527,864
4	000000359019	6,950,000
5	000000589832	4,614,509
6	000000572256	4,562,237
7	000000587023	3,378,078
8	000000425143	3,236,343
9	000000498353	2,575,968
10	000000302835	2,516,669
11	000000584591	2,247,149
12	000001950522	1,922,023
13	000002040522	1,650,837
14	000000487492	1,318,515
15	000000508949	1,167,421
16	000000282960	821,046
17	000000575516	397,651
18	000000318290	311,602
19	000000283858	302,231
20	000000546500	235,110
⋮		
50	000002795269	114,303
51	000000598790	113,605
52	000000264785	111,592
53	000001847206	97,185
54	000000323659	95,986
55	000000284340	72,000
56	000002293349	56,394
57	000000598112	50,000
58	000000598316	45,000
59	000000281597	44,581
60	000000294158	36,558
61	000000266353	36,436
62	000000422545	36,000
63	000001275477	34,379
64	000000258423	33,600
65	000000267136	32,890
66	000000042586	26,220
67	000002461148	25,640
68	000002439686	25,600
69	000000281988	17,000
70	000000061722	17,000
71	000000267972	15,054
72	000001847680	15,000
73	000000077114	14,400
74	000000259702	10,208
75	000000599611	8,200
76	000001221736	4,500

Construct stratification:

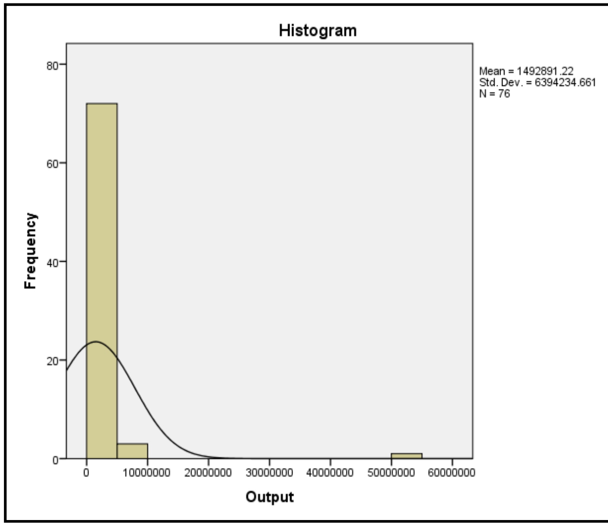


10203

Obs	ID	Output	Cum_output	% cut-off	Strata
1	000000436181	54,438,247	54,438,247	48	1
2	000000498672	8,616,263	63,054,510	56	1
3	000000463440	7,527,864	70,582,374	62	1
4	000000359019	6,950,000	77,532,374	68	1
5	000000589832	4,614,509	82,146,883	72	2
6	000000572256	4,562,237	86,709,120	76	2
7	000000587023	3,378,078	90,087,198	79	2
8	000000425143	3,236,343	93,323,541	82	2
10	000000302835	2,516,669	98,416,178	87	2
11	000000584591	2,247,149	100,663,327	89	2
12	000001950522	1,922,023	102,585,350	90	2
13	000002040522	1,650,837	104,236,187	92	2
14	000000487492	1,318,515	105,554,702	93	2
15	000000508949	1,167,421	106,722,123	94	2
16	000000282960	821,046	107,543,169	95	3
17	000000575516	397,651	107,940,820	95	3
18	000000318290	311,602	108,252,422	95	3
⋮					
43	000003046235	114,303	111,584,588	98	3
44	000003051555	114,303	111,698,890	98	3
45	000002040463	114,303	111,813,193	99	4
46	000002833225	114,303	111,927,495	99	4
47	000002740034	114,303	112,041,798	99	4
48	000002989361	114,303	112,156,100	99	4
49	000003038437	114,303	112,270,403	99	4
50	000002795269	114,303	112,384,705	99	4
51	000000598790	113,605	112,498,310	99	4
52	000000264785	111,592	112,609,902	99	4
53	000001847206	97,185	112,707,087	99	4
⋮					

Full coverage for strata 1

Cut-off	Strata	N
70	1	4
71-94	2	11
95-98	3	29
99-100	4	32
Total		76



10203

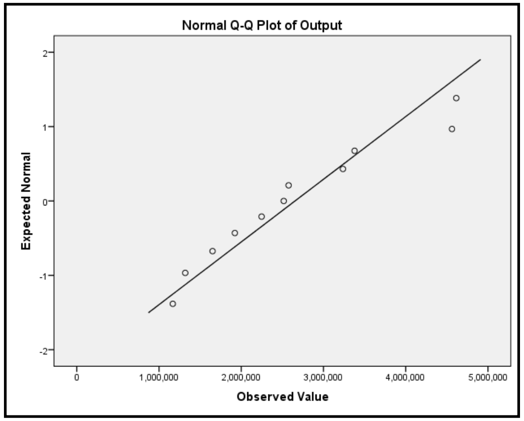
Strata 2

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Output	.162	11	.200*	.925	11	.359

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

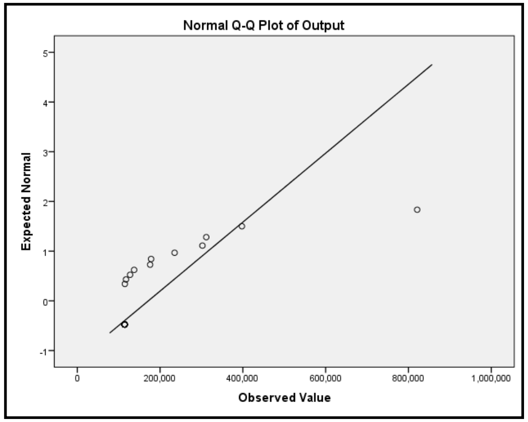


Strata 3

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Output	.353	29	.000	.464	29	.000

a. Lilliefors Significance Correction

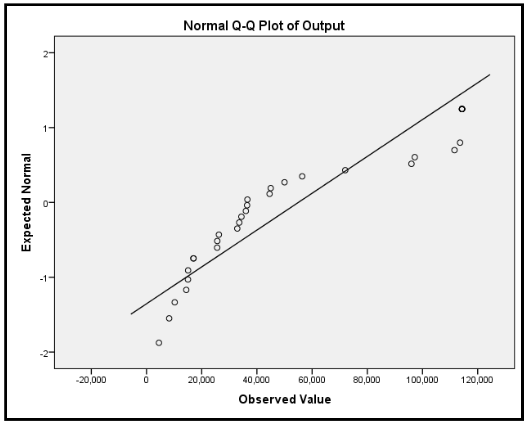


Strata 4

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Output	.207	32	.001	.832	32	.000

a. Lilliefors Significance Correction



10203

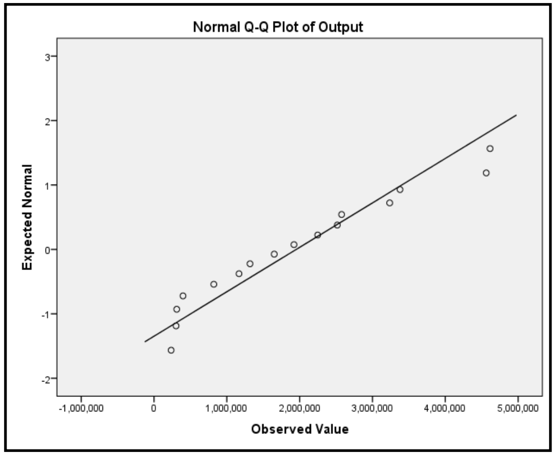
Strata 2

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Output	.118	16	.200*	.921	16	.174

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

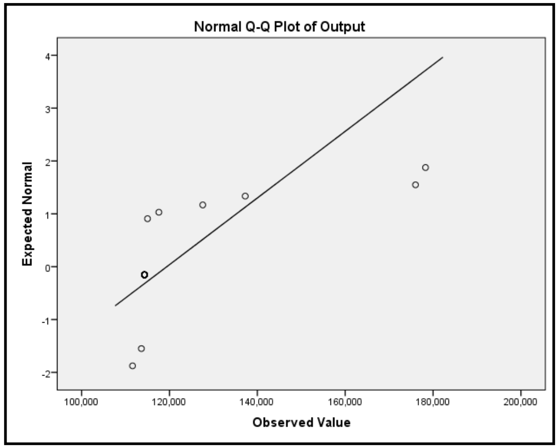


Strata 3

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Output	.453	32	.000	.388	32	.000

a. Lilliefors Significance Correction

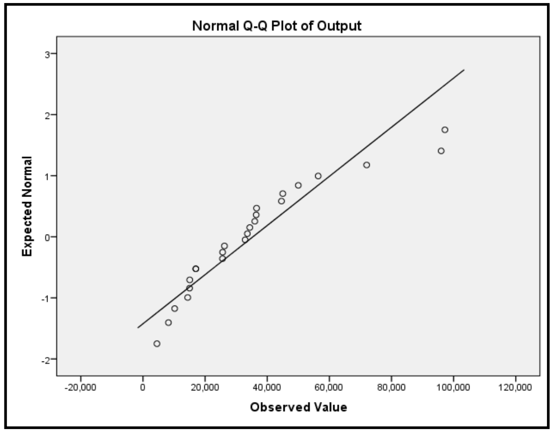


Strata 4

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Output	.190	24	.025	.876	24	.007

a. Lilliefors Significance Correction



10203

Neyman allocation; $$\frac{(\sum N_h S_h)^2}{N^2 \frac{d^2}{z^2} + \sum N_h S^2{}_h}$$

Strata	N	Std. Dev.	Sh2	$N_h S_h$	$N_h S_h^2$	d	V	n
2	11	1186356.39	1.40744E+12	13049920	1.54819E+13			3
3	29	144169.5661	20784863781	4180917	6.02761E+11			8
4	32	40629.36484	1650745288	1300140	52823849205			9
	72			**18530977**	**1.61374E+13**	**1,796,368**	**8.39998E+11**	**20**

Construct stratification using stratification package - R

Literature review

- Univariate Stratification of Survey Populations.
  Author: Sophie Baillargeon and Louis-Paul Rivest

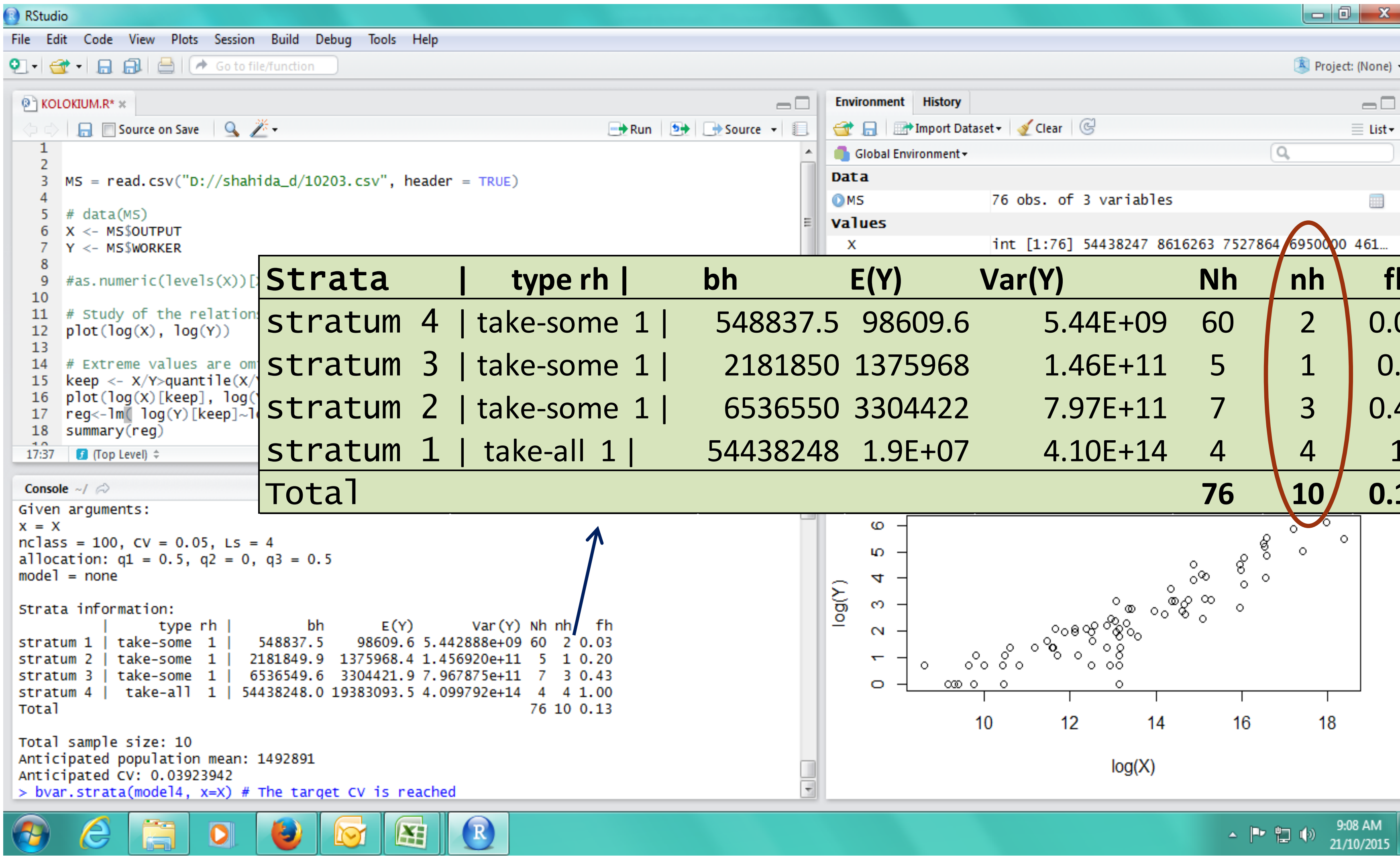
  Univariate stratification of survey populations with a generalization of the Lavallee-Hidiroglou method of stratum construction. The generalized method takes into account a discrepancy between the stratification variable and the survey variable. The determination of the optimal boundaries also incorporate, if desired, an anticipated non-response, **a take-all stratum for large units, a take-none stratum for small units, and a certainty stratum to ensure that some specific units are in the sample**. The well known cumulative root frequency rule of Dalenius and Hodges and the geometric rule of Gunning and Horgan are also implemented.

- Sampling Strata: An R package for the Optimization of Stratified Sampling.
  Author: Guilio Barcaroli, Italian National Institute of Statistics (Istat)

- Survey Methodology – The construction of stratified designs in R with the package stratification.
  Author: Louis-Paul Rivest and Sophie Baillargeon, Statistics Canada

R Package

10203



Strata	type rh	bh	E(Y)	Var(Y)	Nh	nh	fh
stratum 4	take-some 1	548837.5	98609.6	5.44E+09	60	2	0.03
stratum 3	take-some 1	2181850	1375968	1.46E+11	5	1	0.2
stratum 2	take-some 1	6536550	3304422	7.97E+11	7	3	0.43
stratum 1	take-all 1	54438248	1.9E+07	4.10E+14	4	4	1
Total					76	10	0.13

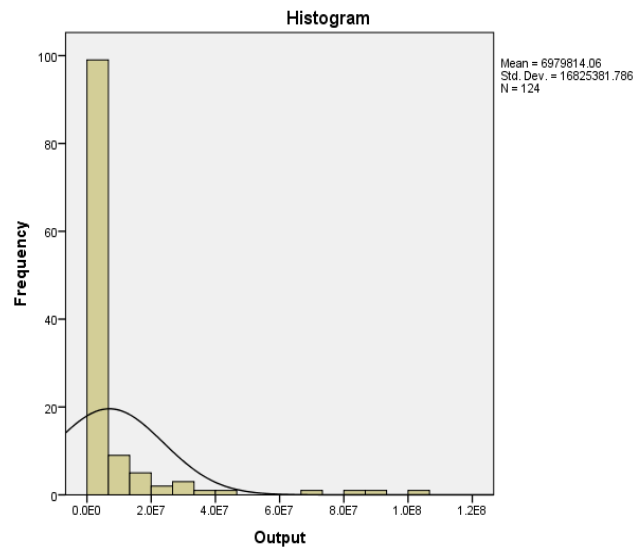
29200 – Manufacture of bodies (coachwork) for motor vehicles; manufacture of trailers and semi-trailers

Obs	ID	Output
1	000000489548	102,093,255
2	000000575360	88,424,188
3	000000478592	86,611,415
4	000000530322	73,033,450
5	000000552591	40,332,221
6	000000486889	37,186,823
7	000000483165	32,032,329
8	000000483711	30,802,394
9	000000556648	30,760,448
10	000000520099	21,102,203
11	000000519615	20,083,603
12	000000457541	19,749,767
13	000000450329	19,702,174
14	000000504333	18,647,518
15	000000567158	18,540,910
16	000000504890	18,436,354
17	000000448089	12,343,123
18	000000466248	11,500,742
19	000000542906	10,193,977
20	000000552936	9,120,269
.		
.		
.		
105	000000324784	205,458
106	000000497353	186,456
107	000001879482	174,330
108	000000137766	149,802
109	000000723698	138,072
110	000000125014	123,848
111	000000529685	117,532
112	000000121689	111,538
113	000000125351	110,963
114	000000407875	110,500
115	000000335733	107,487
116	000000127018	102,832
117	000000270475	72,854
118	000000319315	63,005
119	000002276785	61,065
120	000000270426	59,470
121	000000293296	57,600
122	000000285334	53,241
123	000000287332	52,527
124	000000119912	37,160

29200

Obs	ID	Output	Cum_output	Cut-off	Strata Cut-Off
1	000000489548	102,093,255	102,093,255	12	1
2	000000575360	88,424,188	190,517,443	22	1
3	000000478592	86,611,415	277,128,858	32	1
4	000000530322	73,033,450	350,162,308	40	1
5	000000552591	40,332,221	390,494,529	45	1
6	000000486889	37,186,823	427,681,352	49	1
7	000000483165	32,032,329	459,713,681	53	1
8	000000483711	30,802,394	490,516,075	57	1
9	000000556648	30,760,448	521,276,523	60	1
10	000000520099	21,102,203	542,378,726	63	1
11	000000519615	20,083,603	562,462,329	65	1
12	000000457541	19,749,767	582,212,096	67	2
13	000000450329	19,702,174	601,914,270	70	2
14	000000504333	18,647,518	620,561,788	72	2
⋮					
70	000002755898	1,009,920	847,858,024	98	3
71	000000319789	970,662	848,828,686	98	3
72	000000527775	968,662	849,797,348	98	3
73	000000263275	947,786	850,745,134	98	3
74	000000123287	880,816	851,625,950	98	3
75	000000212728	829,061	852,455,011	98	3
76	000000237589	683,491	853,138,502	99	4
77	000000174283	667,847	853,806,349	99	4
78	000000189461	592,970	854,399,319	99	4
79	000000128078	551,686	854,951,005	99	4
80	000000740600	526,450	855,477,455	99	4
81	000000454856	516,612	855,994,067	99	4
82	000000133086	460,587	856,454,654	99	4
83	000000760398	453,279	856,907,933	99	4
84	000000127918	442,880	857,350,813	99	4
85	000000319464	438,890	857,789,703	99	4
⋮					

Strata	cut-off	N
1	65%	11
2	66-94	33
3	95-98	31
4	99-100	49
Total		124



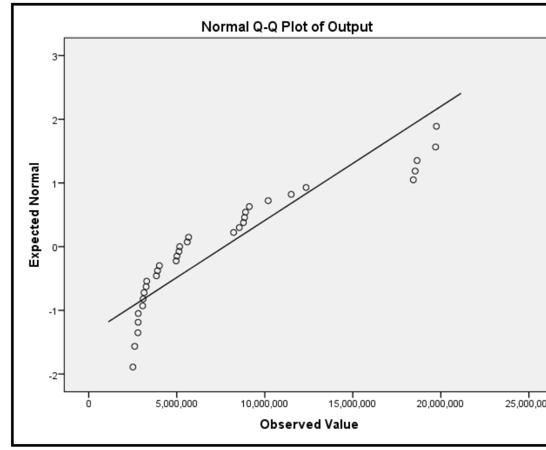
29200

Strata 2

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Output	.218	33	.000	.806	33	.000

a. Lilliefors Significance Correction

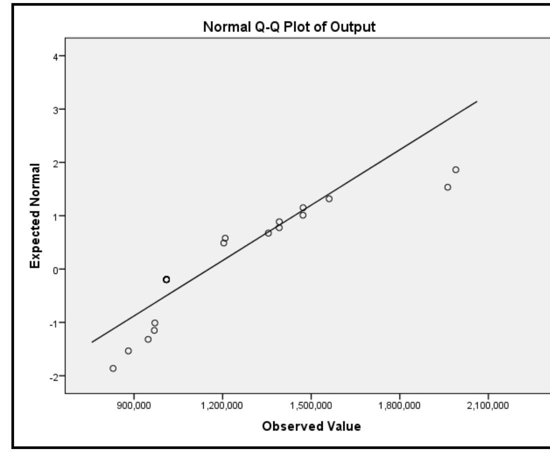


Strata 3

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Output	.368	31	.000	.738	31	.000

a. Lilliefors Significance Correction



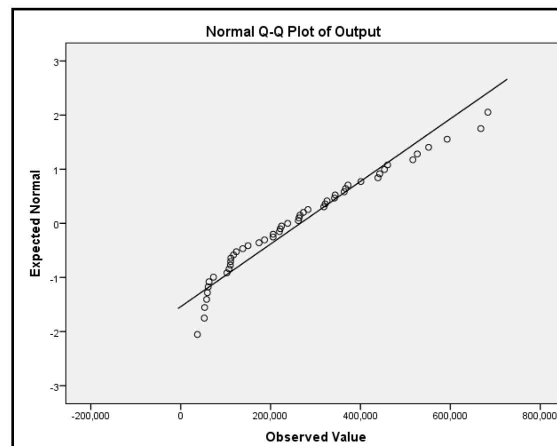
Strata 4

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Output	.101	49	.200*	.937	49	.011

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction



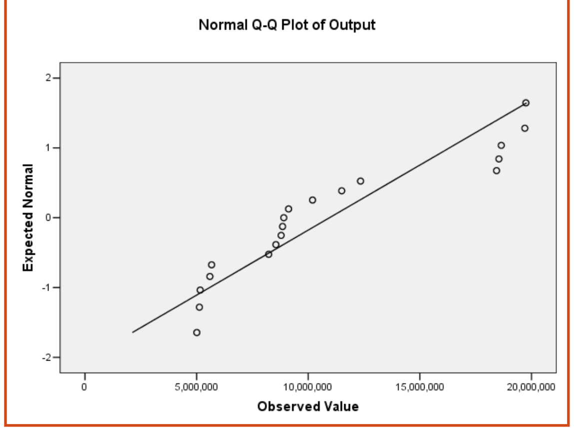
29200

Strata 2

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Output	.213	19	.024	.842	19	.005

a. Lilliefors Significance Correction

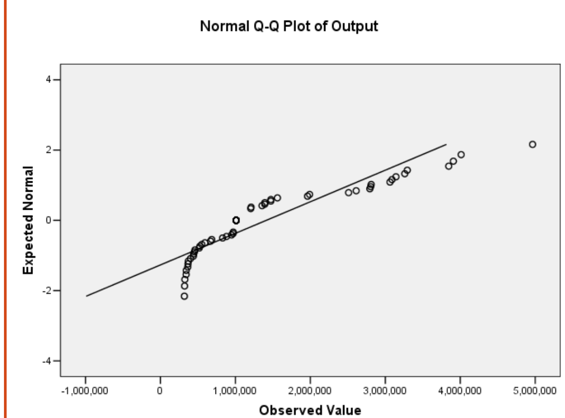


Strata 3

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Output	.266	64	.000	.820	64	.000

a. Lilliefors Significance Correction

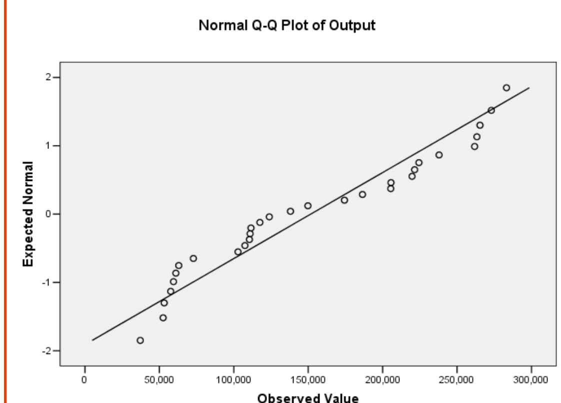


Strata 4

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Output	.137	30	.156	.918	30	.023

a. Lilliefors Significance Correction



29200

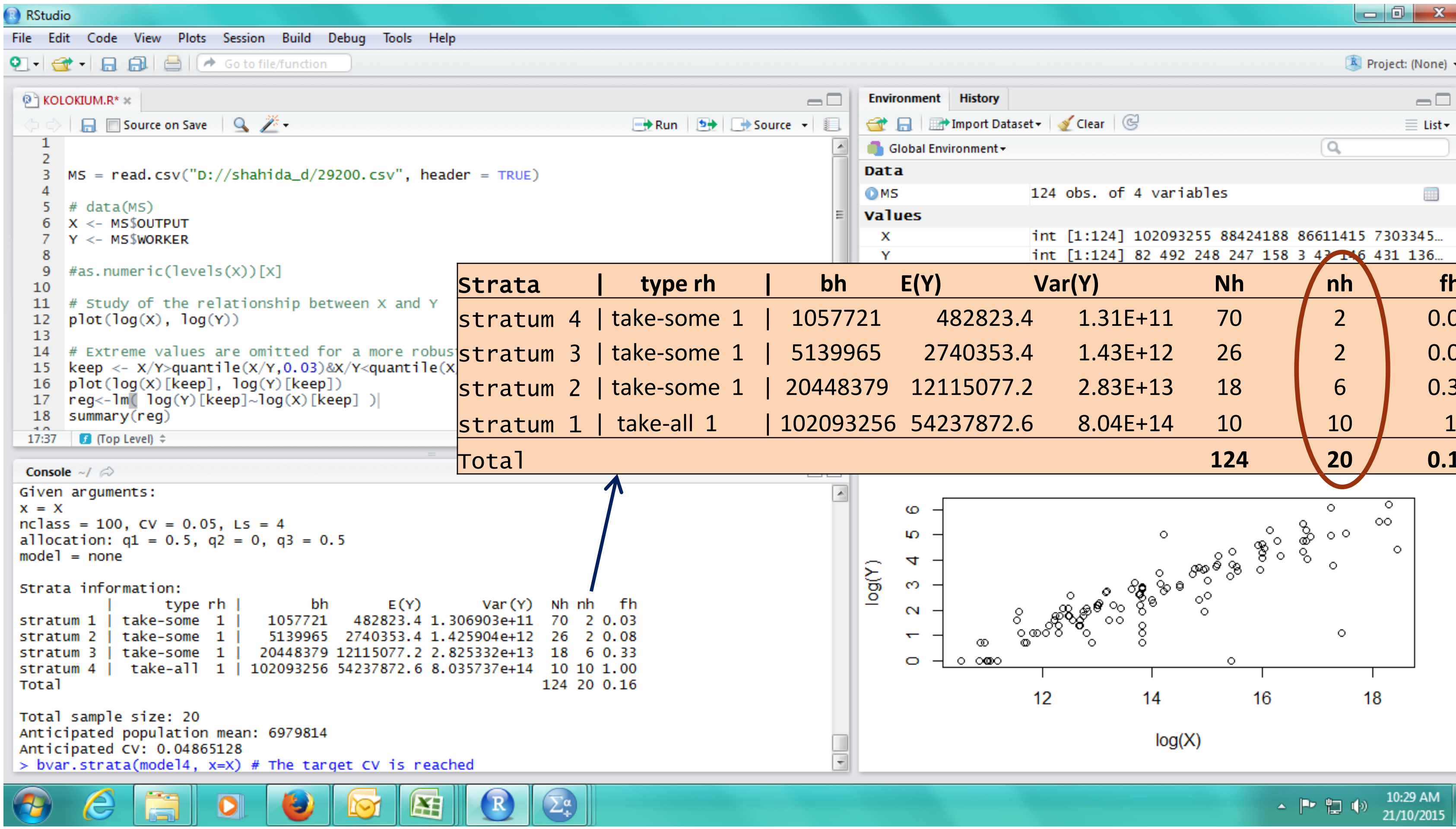
Neyman allocation;

$$\frac{(\sum N_h S_h)^2}{\frac{N^2 d^2}{z^2} + \sum N_h S^2_h}$$

Strata	N	Std Dev	Sh2	$N_h S_h$	$N_h S_h^2$	d	V	n
2	33	5583825.161	3.11791E+13	184266230	1.02891E+15			11
3	31	288754.6646	83379256329	8951394.6	2.58476E+12			10
4	49	172865.7785	29882577364	8470423.1	1.46425E+12			16
	113			201688048	1.03296E+15	15,151,731	5.97602E+13	**37**

R Package

29200



Strata	type rh	bh	E(Y)	Var(Y)	Nh	nh	fh
stratum 4	take-some 1	1057721	482823.4	1.31E+11	70	2	0.03
stratum 3	take-some 1	5139965	2740353.4	1.43E+12	26	2	0.08
stratum 2	take-some 1	20448379	12115077.2	2.83E+13	18	6	0.33
stratum 1	take-all 1	102093256	54237872.6	8.04E+14	10	10	1
Total					**124**	**20**	**0.16**

32901 – Manufacture of stationery

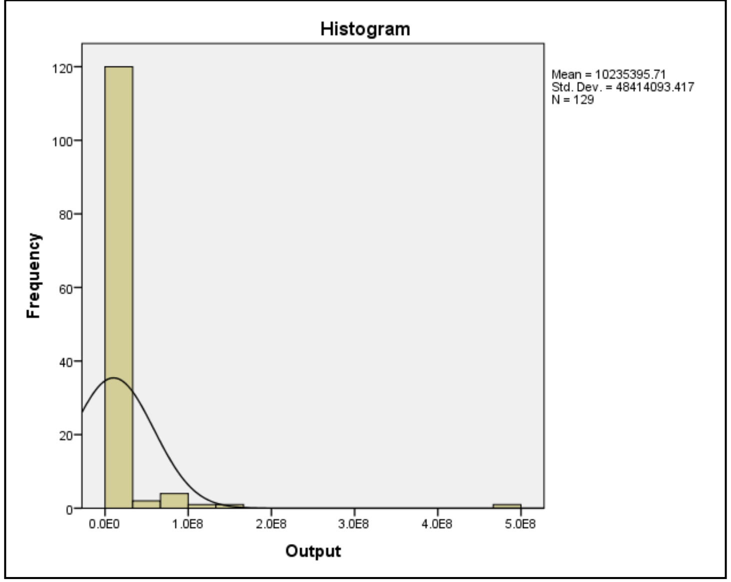
Obs	ID	Output	Cum_output
1	000000476155	499,123,382	499,123,382
2	000000442745	150,662,867	649,786,249
3	000000499732	106,211,595	755,997,844
4	000000541100	89,969,122	845,966,966
5	000000511676	78,759,056	924,726,022
6	000000457458	72,017,811	996,743,833
7	000000545058	68,645,645	1,065,389,478
8	000000557699	47,584,417	1,112,973,895
9	000000433272	39,119,732	1,152,093,627
10	000000556541	30,020,540	1,182,114,167
11	000000475926	21,221,943	1,203,336,110
12	000000531963	16,962,253	1,220,298,363
13	000001427670	16,447,803	1,236,746,166
14	000000446670	14,818,091	1,251,564,257
15	000000571832	11,760,017	1,263,324,274
16	000000502282	7,968,154	1,271,292,428
17	000000476285	7,510,737	1,278,803,165
18	000000536742	3,455,585	1,282,258,750
19	000000523060	2,118,374	1,284,377,124
20	000000135032	2,103,661	1,286,480,785
⋮			

⋮		
109	000000041318	47,433
110	000000237611	46,898
111	000000135762	46,105
112	000000317850	44,980
113	000000374049	38,880
114	000000421419	36,850
115	000001172832	36,000
116	000000428964	35,750
117	000000608818	34,329
118	000000237801	29,578
119	000000252416	29,062
120	000000371417	25,150
121	000000370266	24,780
122	000000423257	24,600
123	000000128797	19,890
124	000002112310	19,440
125	000000061984	18,385
126	000000362581	15,704
127	000000319725	14,401
128	000002238547	6,000
129	000002112115	4,160

32901

Obs	ID	Output	Cum_output	Cut-off	Strata
1	000000476155	499,123,382	499,123,382	38	1
2	000000442745	150,662,867	649,786,249	49	1
3	000000499732	106,211,595	755,997,844	57	1
4	000000541100	89,969,122	845,966,966	64	2
5	000000511676	78,759,056	924,726,022	70	2
6	000000457458	72,017,811	996,743,833	75	2
7	000000545058	68,645,645	1,065,389,478	81	2
8	000000557699	47,584,417	1,112,973,895	84	2
9	000000433272	39,119,732	1,152,093,627	87	2
10	000000556541	30,020,540	1,182,114,167	90	2
11	000000475926	21,221,943	1,203,336,110	91	2
12	000000531963	16,962,253	1,220,298,363	92	2
13	000001427670	16,447,803	1,236,746,166	94	2
14	000000446670	14,818,091	1,251,564,257	95	2
15	000000571832	11,760,017	1,263,324,274	96	2
16	000000502282	7,968,154	1,271,292,428	96	2
17	000000476285	7,510,737	1,278,803,165	97	3
18	000000536742	3,455,585	1,282,258,750	97	3
⋮					
25	000000454411	1,712,907	1,295,946,200	98	3
26	000000464412	1,612,921	1,297,559,121	98	3
27	000000557721	1,291,041	1,298,850,162	98	3
28	000000438812	1,222,750	1,300,072,912	98	3
29	000000570786	1,035,881	1,301,108,793	99	4
30	000002050105	1,031,366	1,302,140,159	99	4
31	000000496299	899,385	1,303,039,544	99	4
32	000000237603	866,050	1,303,905,594	99	4
33	000000501953	790,400	1,304,695,994	99	4
34	000000463150	738,558	1,305,434,552	99	4
35	000000500214	685,401	1,306,119,953	99	4
⋮					

Strata	cut-off	N
1	60%	3
2	61-96	13
3	97-98	12
4	99-100	101
		129



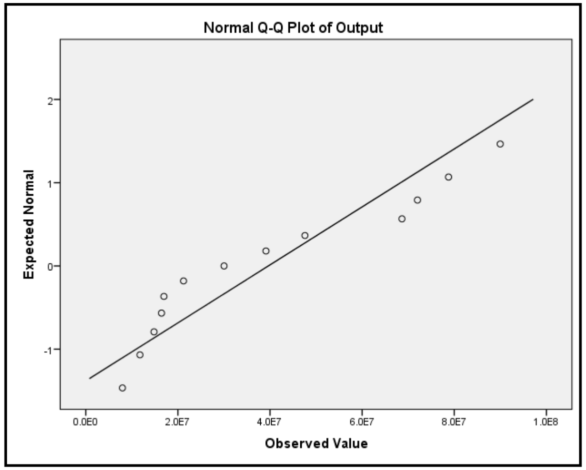
32901

Strata 2

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Output	.201	13	.156	.879	13	.069

a. Lilliefors Significance Correction

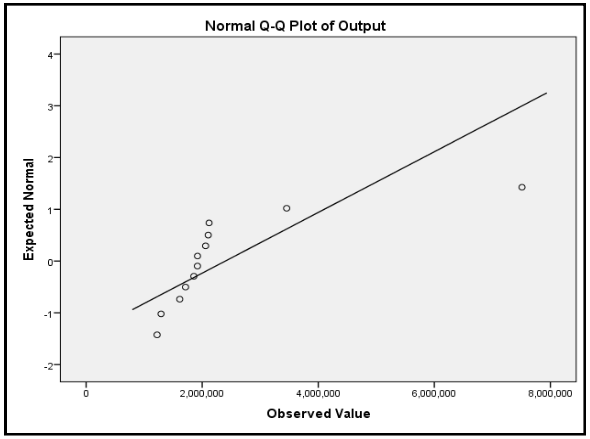


Strata 3

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Output	.399	12	.000	.595	12	.000

a. Lilliefors Significance Correction

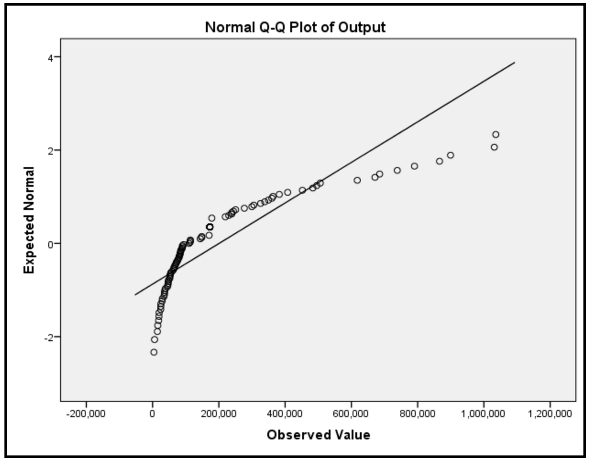


Strata 4

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Output	.252	101	.000	.738	101	.000

a. Lilliefors Significance Correction



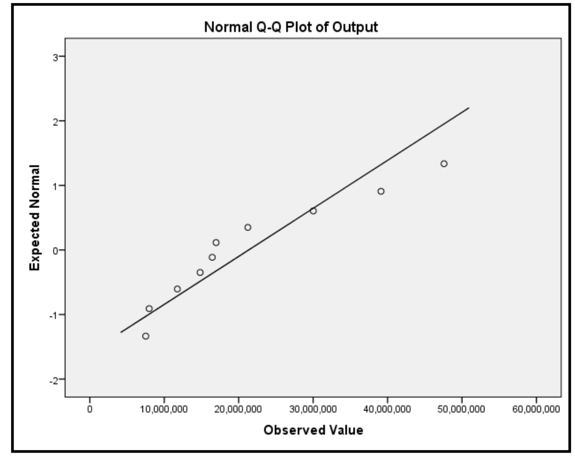
32901

Strata 2

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Output	.228	10	.152	.885	10	.150

a. Lilliefors Significance Correction

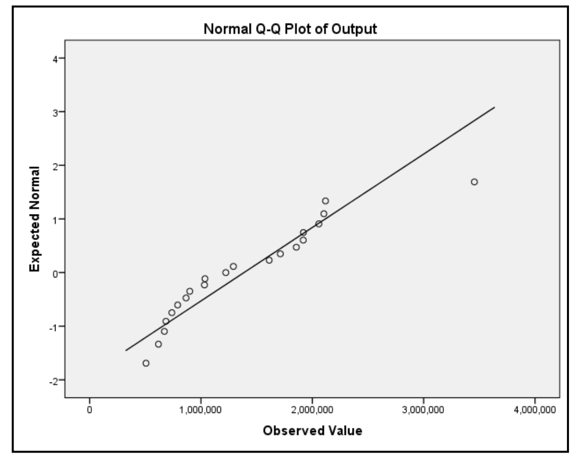


Strata 3

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Output	.161	21	.166	.890	21	.022

a. Lilliefors Significance Correction

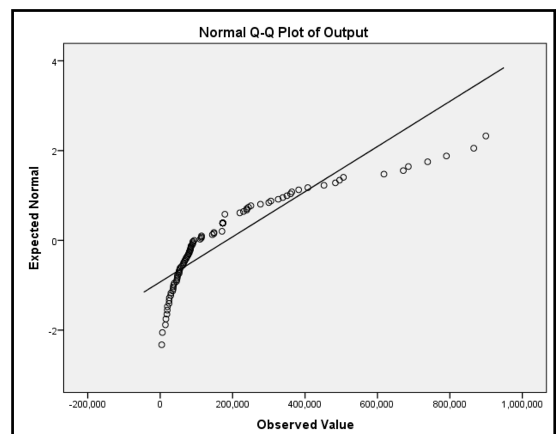


Strata 4

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Output	.240	99	.000	.762	99	.000

a. Lilliefors Significance Correction



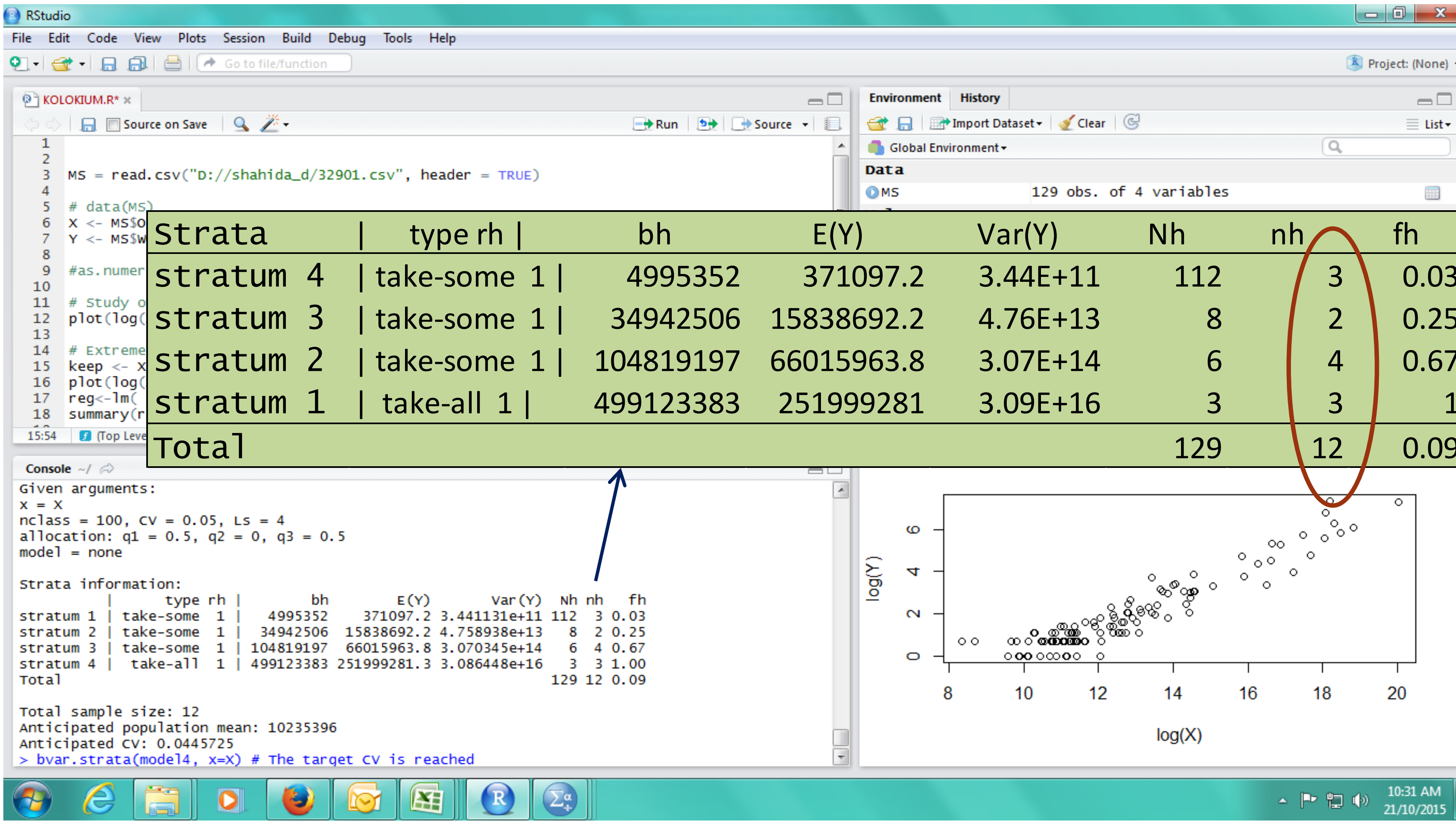
32901

Neyman allocation; $$\frac{(\sum N_h S_h)^2}{\frac{N^2 d^2}{z^2} + \sum N_h S^2{}_h}$$

Strata	N	Std Dev	Sh2	$N_h S_h$	$N_h S_h{}^2$	d	V	n
2	13	28718169	8.24733E+14	373336202	1.07215E+16			2
3	12	1705168	2.9076E+12	20462020	3.48912E+13			2
4	101	229965.7	52884207381	23226532	5.3413E+12			13
	126			417024754	1.07618E+16	28,218,410	2.07278E+14	**16**

R Package

32901



Strata	type rh	bh	E(Y)	Var(Y)	Nh	nh	fh
stratum 4	take-some 1	4995352	371097.2	3.44E+11	112	3	0.03
stratum 3	take-some 1	34942506	15838692.2	4.76E+13	8	2	0.25
stratum 2	take-some 1	104819197	66015963.8	3.07E+14	6	4	0.67
stratum 1	take-all 1	499123383	251999281	3.09E+16	3	3	1
Total					129	12	0.09

Conclusion

	Cut-off			R package		
MSIC	**10203**	**29200**	**32901**	**10203**	**29200**	**32901**
N	76	124	129	76	124	129
	n=24	n=48	n=20	n=10	n=20	n=12
Strata 1	4	11	3	4	10	3
Strata 2	3	11	2	3	6	4
Strata 3	8	10	2	1	2	2
Strata 4	9	16	13	2	2	3
RSE						
Strata 1	60.31	17.78	49.30	60.31	17.42	49.30
Strata 2	13.48	12.62	20.09	11.02	10.64	11.87
Strata 3	15.60	4.50	20.52	13.87	8.72	16.46
Strata 4	13.05	9.28	29.18	9.74	9.01	15.00

Thank you.