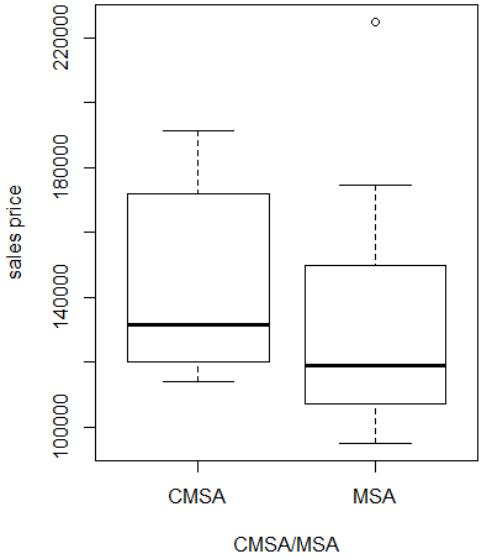
Kyle Beitz

STAT 4155 Homework 2

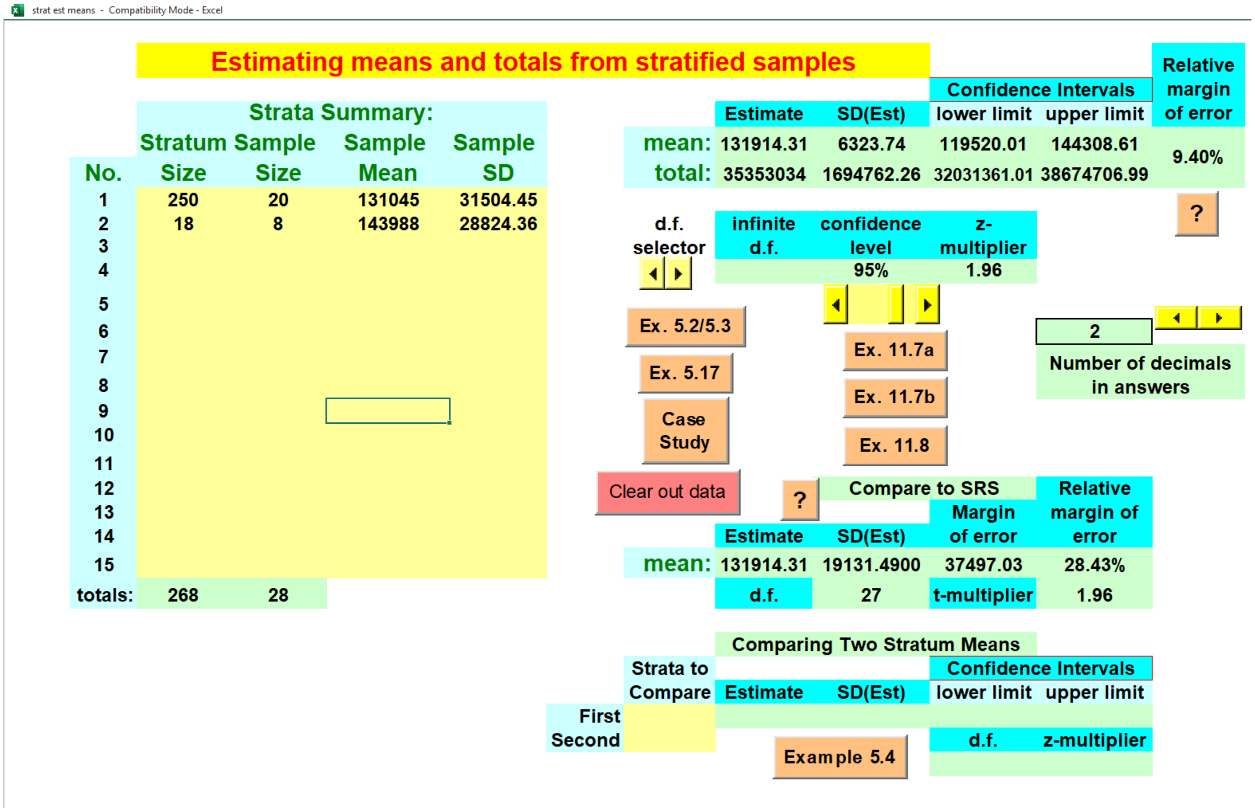
1. SMOG 5.4

a)



One unusual feature i can see from this boxplot is the outlier in the MSA strata. This outlier will impact the accuracy of our estimates, so it should be accounted for in some way when making inferences. If the outlier seems to be an undeniable error in the data (say a sales price of -99999) it can be removed or corrected safely by speaking to the person who collected the data and fixing the error. If the outlier is not an error (in this case a sales price of 22500 seems possible), the outlier should not be removed as it is important to the model and should be accounted for. If still unsure about the outlier, using robust statistics may be the way to go. The mean and variance are very sensitive to outliers, whereas the median and interquartile range are less so. The latter should be considered over the former in situations where you are worried outliers may impact estimates greatly.

b) (on following page)



Estimating means and totals from stratified samples

Strata Summary:

No.	Stratum Size	Sample Size	Sample Mean	Sample SD
1	250	20	131045	31504.45
2	18	8	143988	28824.36
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
totals:	268	28		

			Confidence Intervals		Relative margin of error
	Estimate	SD(Est)	lower limit	upper limit	
mean:	131914.31	6323.74	119520.01	144308.61	9.40%
total:	35353034	1694762.26	32031361.01	38674706.99	

infinite d.f.	confidence level	z-multiplier
	95%	1.96

2 Number of decimals in answers

	Compare to SRS			
	Estimate	SD(Est)	Margin of error	Relative margin of error
mean:	131914.31	19131.4900	37497.03	28.43%
	d.f.	27	t-multiplier	1.96

Comparing Two Stratum Means

Strata to Compare			Confidence Intervals	
	Estimate	SD(Est)	lower limit	upper limit
First				
Second			d.f.	z-multiplier

Mean Typical Sales Price Per House For All Metropolitan Areas of the U.S.:

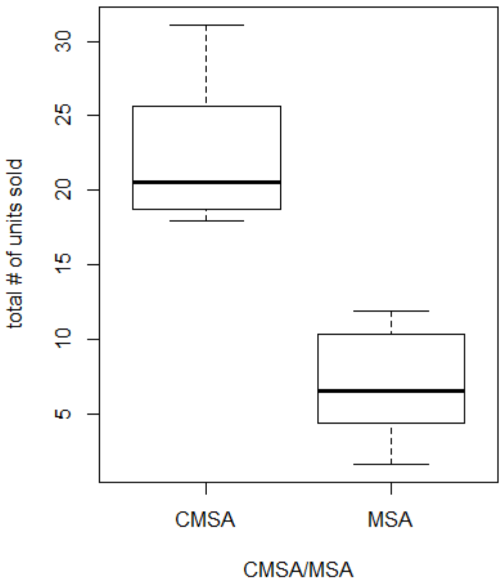
$$\underline{y}_{st} = 131914.31$$

Bound For Error of Estimation:

$\underline{y}_{st} \pm 1.96(6323.74)$ <- Used 1.96 instead of 2 to match tool

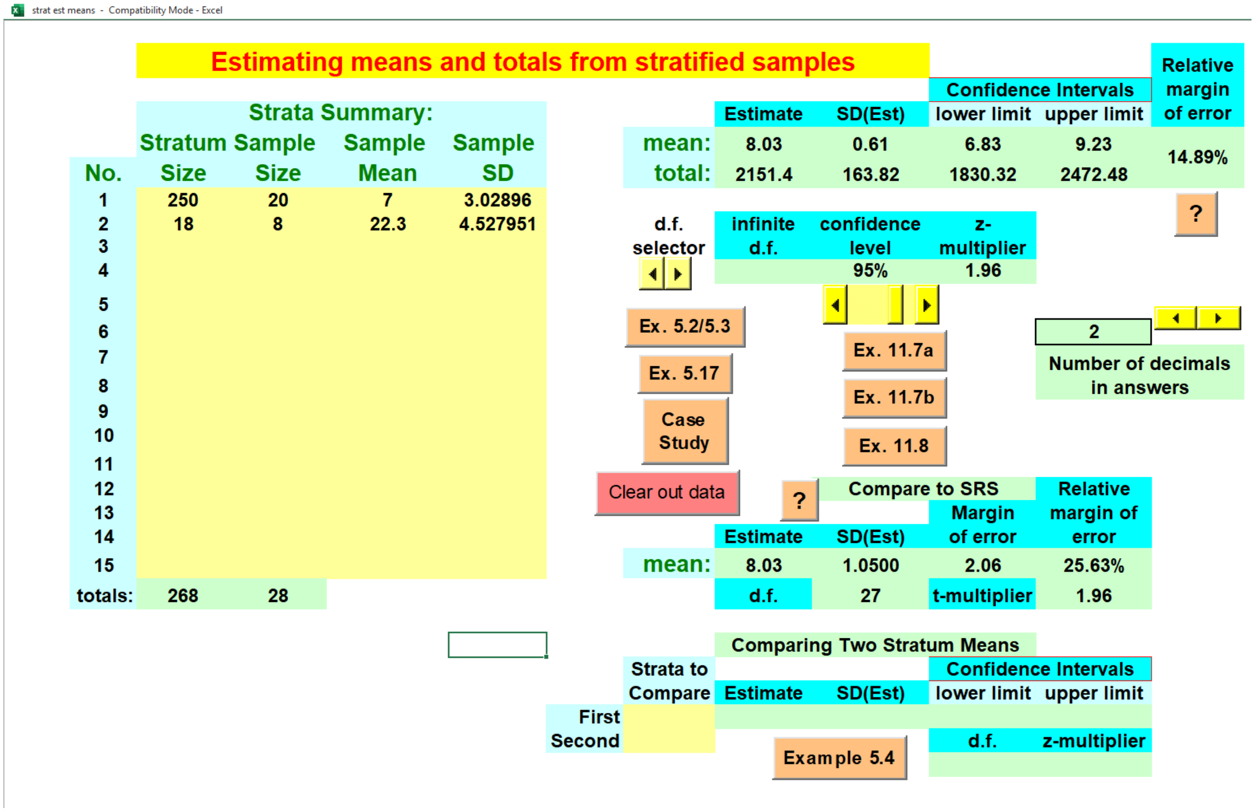
$$\underline{y}_{st} \pm 12394.5304$$

c) (on following page)



There are no outliers or unusual features of this boxplot to be concerned about.

d)



Estimating means and totals from stratified samples

Strata Summary:

No.	Stratum Size	Sample Size	Sample Mean	Sample SD
1	250	20	7	3.02896
2	18	8	22.3	4.527951
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
totals:	268	28		

			Confidence Intervals		Relative margin of error
	Estimate	SD(Est)	lower limit	upper limit	
mean:	8.03	0.61	6.83	9.23	14.89%
total:	2151.4	163.82	1830.32	2472.48	

d.f. selector	infinite d.f.	confidence level	z-multiplier
		95%	1.96

2
Number of decimals in answers

	Compare to SRS			
	Estimate	SD(Est)	Margin of error	Relative margin of error
mean:	8.03	1.0500	2.06	25.63%
	d.f.	27	t-multiplier	1.96

Comparing Two Stratum Means

Strata to Compare	Estimate	SD(Est)	Confidence Intervals lower limit	Confidence Intervals upper limit
First				
Second			d.f.	z-multiplier

Total Number Of Houses Sold In All Metropolitan Areas of the U.S. In 1993:

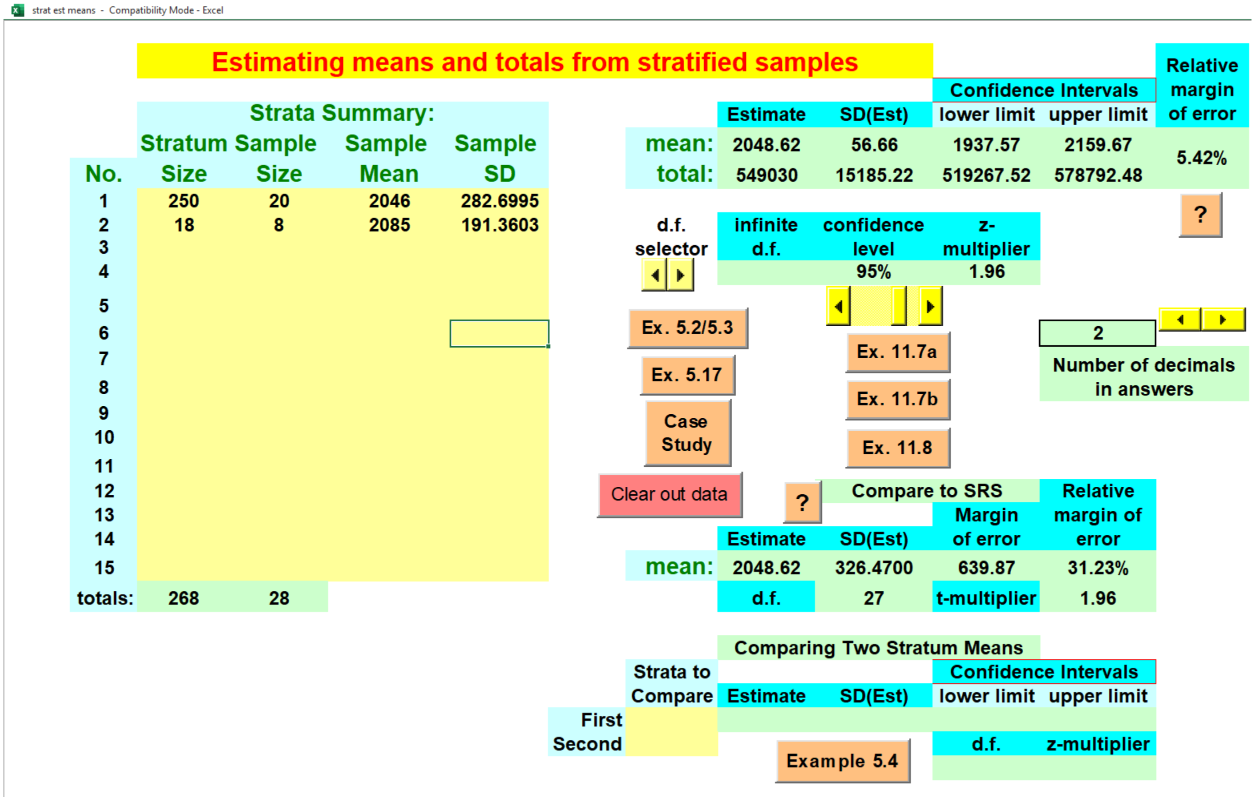
$$N\underline{y}_{st} = 2151.4$$

Bound For Error of Estimation:

$N\underline{y}_{st} \pm 1.96(163.82)$ <- Used 1.96 instead of 2 to match tool

$$N\underline{y}_{st} \pm 321.0872$$

e)



Estimating means and totals from stratified samples

Strata Summary:

No.	Stratum Size	Sample Size	Sample Mean	Sample SD
1	250	20	2046	282.6995
2	18	8	2085	191.3603
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
totals:	268	28		

			Confidence Intervals		Relative
	Estimate	SD(Est)	lower limit	upper limit	margin of error
mean:	2048.62	56.66	1937.57	2159.67	5.42%
total:	549030	15185.22	519267.52	578792.48	

d.f. selector	infinite d.f.	confidence level	z-multiplier
		95%	1.96

2
Number of decimals in answers

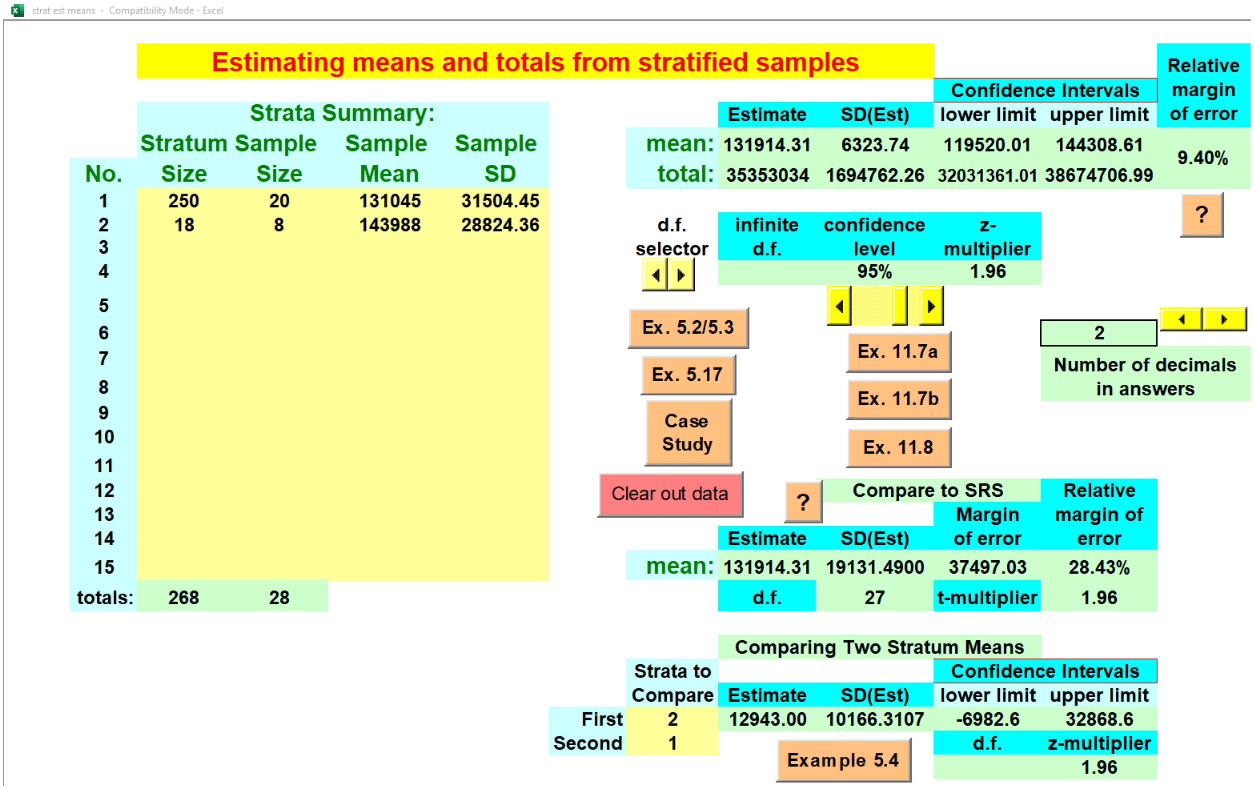
	Compare to SRS			Relative
			Margin	margin of
	Estimate	SD(Est)	of error	error
mean:	2048.62	326.4700	639.87	31.23%
	d.f.	27	t-multiplier	1.96

Comparing Two Stratum Means

Strata to			Confidence Intervals	
Compare	Estimate	SD(Est)	lower limit	upper limit
First				
Second				
			d.f.	z-multiplier

Part b shows price. The gain in precision for stratifying is approximately 19% (RMoE of SRS - RMoE of Stratified Random Sampling). Part d shows total units sold. The gain in precision for stratifying is approximately 11%. The above screenshot shows square footage. The gain in precision for stratifying is approximately 26%. Thus, stratification produces the least gain in precision for total units sold. This is because total units sold does not differ greatly between the two strata (MSA and CMSA). The difference between the two strata's SDs is only about 1.5, while the difference between the SDs of the two strata for price is approximately 2700 and the difference between the SDs of the two strata for square footage is approximately 90. Stratification produces large gains in precision when the two strata greatly differ from each other, but within each strata the measurements are similar. This is why stratifying within total units sold does not create large gains in precision.

f)



Estimating means and totals from stratified samples

Strata Summary:

No.	Stratum Size	Sample Size	Sample Mean	Sample SD
1	250	20	131045	31504.45
2	18	8	143988	28824.36
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
totals:	268	28		

			Confidence Intervals		Relative margin of error
	Estimate	SD(Est)	lower limit	upper limit	
mean:	131914.31	6323.74	119520.01	144308.61	9.40%
total:	35353034	1694762.26	32031361.01	38674706.99	

infinite d.f.	confidence level	z-multiplier
	95%	1.96

2
Number of decimals in answers

	Compare to SRS			
			Margin of error	Relative margin of error
	Estimate	SD(Est)		
mean:	131914.31	19131.4900	37497.03	28.43%
	d.f.	27	t-multiplier	1.96

Comparing Two Stratum Means

	Strata to Compare			Confidence Intervals	
		Estimate	SD(Est)	lower limit	upper limit
First	2	12943.00	10166.3107	-6982.6	32868.6
Second	1			d.f.	z-multiplier
					1.96

Estimate The Difference In Average Typical Selling Price Between The Two Strata:

$$12943 \pm 1.96(10166.3107)$$

$$12943 \pm 19925.96897$$

CI: (-6982.6, 32868.6)

We cannot say that the houses in the CMSAs are, on average, higher priced than those in the MSAs because the above estimate and confidence interval includes 0. This means that there is a chance there is no real difference in average typical selling price between the two strata, even though the CMSAs sample mean is much larger.

2. SMOG 5.20

In this situation, we believe simple random sampling would work just as well because not much is changing between the months except for the month itself. There is no obvious reason why the batteries produced in one month would be much different than those produced in another month. The characteristics of the batteries produced month-to-month should be relatively homogenous (similar), so if simple random sampling is just as convenient, you might as well do that. If, however, there was reason to believe that batteries did vary greatly between months (say, for example, worker morale is higher in January than December so they therefore produce better working batteries in January), gains in precision would be made through stratifying by month.

3. SMOG 5.22

Stratification will produce large gains in precision over simple random sampling when measurements within each strata are homogenous. The more similar measurements are within each strata, the more payoff you get for using stratification (compared to simple random sampling). If each strata has a lot of variability within, stratification does not have as much of a positive impact on precision. To summarize, to gain the most precision with stratification, stratify the heterogeneous population into strata that have homogeneous measurements (measurements that are similar to each other). Putting all this into a more explicit framework, you gain the most precision using stratification when you have strata with wildly different means, small SD within each stratum, and large SD overall.

4. SMOG 5.24

If we used stratified random sampling we could make strata's through every 3 years of service and the firm and you could also make sub-strata within the above strata for male and female or job title. If we used simple random sampling we could do the sampling relatively easily and more cost-efficient. I would recommend stratified random sampling because it would give a more precise estimate because you are more likely to see a greater variance in simple random sampling between someone who is new and someone that has been with the firm for a long time. You could also see a greater variance between job titles as well.

5. SMOG 5.32 (part a only)

$$\underline{y_{st}} = \frac{1}{N} \sum_{i=1}^{L} N_i \underline{y_i}$$

= [(0.5*7.63)+(0.1*7.74)+(0.4*6.55)]

= 7.209

$$\hat{V}(\underline{y_{st}}) = \frac{1}{N^2} \sum_{i=i}^{L} N_i^2 (1 - \frac{n_i}{N_i})(\frac{s_i^2}{n_i})$$

$= \frac{1}{(1347+163+1095)^2} [(0.5)^2 * (\frac{.15^2}{1347}) + (0.1)^2 * (\frac{.35^2}{163}) + (0.4)^2 * (\frac{0.11^2}{1095})]$ ←- Ignoring fpc because the sample sizes are large enough to justify not using them.

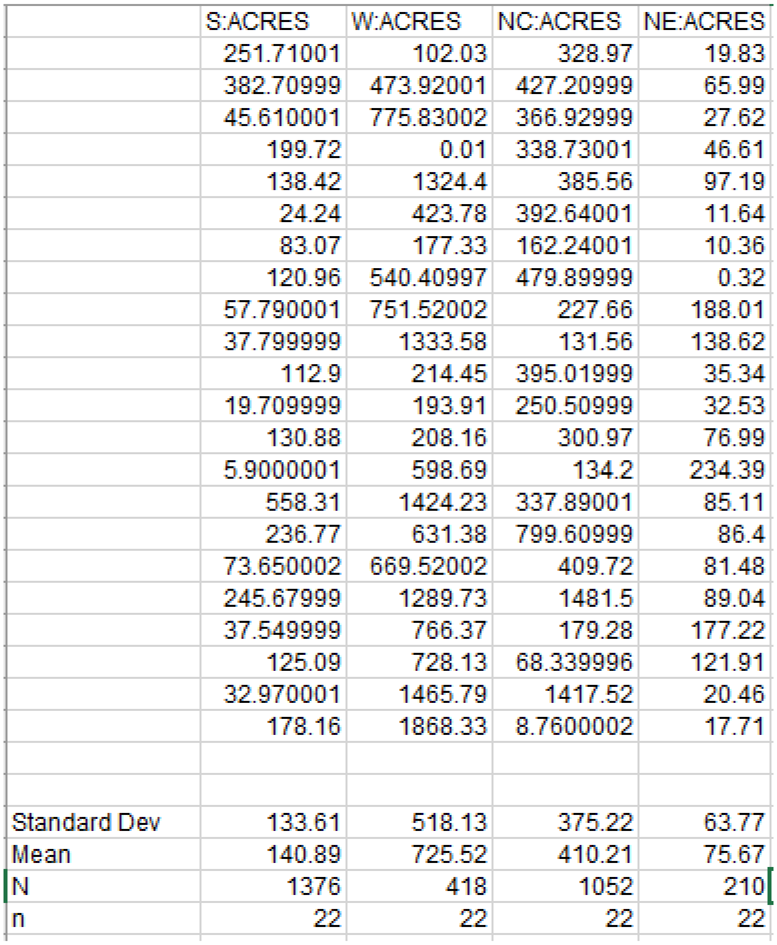
= 0.00000000517

$= \pm 2\sqrt{0.00000000517}$

= 0.00014

The mean time for the population of those giving anesthesia is 7.29 and the bound on error is 0.00014.

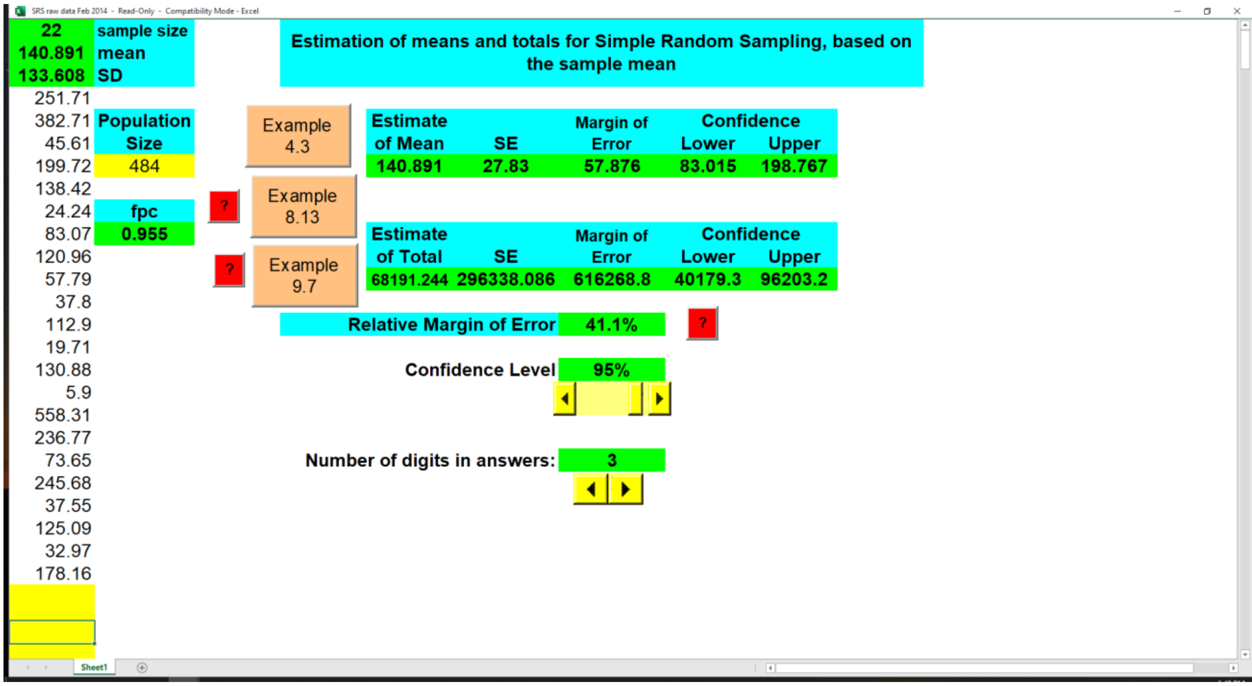
6. SMOG 5.35 (parts a, b, and e only)



	S:ACRES	W:ACRES	NC:ACRES	NE:ACRES
	251.71001	102.03	328.97	19.83
	382.70999	473.92001	427.20999	65.99
	45.610001	775.83002	366.92999	27.62
	199.72	0.01	338.73001	46.61
	138.42	1324.4	385.56	97.19
	24.24	423.78	392.64001	11.64
	83.07	177.33	162.24001	10.36
	120.96	540.40997	479.89999	0.32
	57.790001	751.52002	227.66	188.01
	37.799999	1333.58	131.56	138.62
	112.9	214.45	395.01999	35.34
	19.709999	193.91	250.50999	32.53
	130.88	208.16	300.97	76.99
	5.9000001	598.69	134.2	234.39
	558.31	1424.23	337.89001	85.11
	236.77	631.38	799.60999	86.4
	73.650002	669.52002	409.72	81.48
	245.67999	1289.73	1481.5	89.04
	37.549999	766.37	179.28	177.22
	125.09	728.13	68.339996	121.91
	32.970001	1465.79	1417.52	20.46
	178.16	1868.33	8.7600002	17.71
Standard Dev	133.61	518.13	375.22	63.77
Mean	140.89	725.52	410.21	75.67
N	1376	418	1052	210
n	22	22	22	22

Part A

a) South Region:


SRS raw data Feb 2014 - Read-Only - Compatibility Mode - Excel
Example 4.3
Example 8.13
Example 9.7
?
?
?
Sheet1

22	sample size
140.891	mean
133.608	SD

251.71
382.71
45.61
199.72
138.42
24.24
83.07
120.96
57.79
37.8
112.9
19.71
130.88
5.9
558.31
236.77
73.65
245.68
37.55
125.09
32.97
178.16

Population Size: 484

fpc: 0.955

Estimation of means and totals for Simple Random Sampling, based on the sample mean

Estimate of Mean	SE	Margin of Error	Confidence Lower	Confidence Upper
140.891	27.83	57.876	83.015	198.767

Estimate of Total	SE	Margin of Error	Confidence Lower	Confidence Upper
68191.244	296338.086	616268.8	40179.3	96203.2

Relative Margin of Error: 41.1%

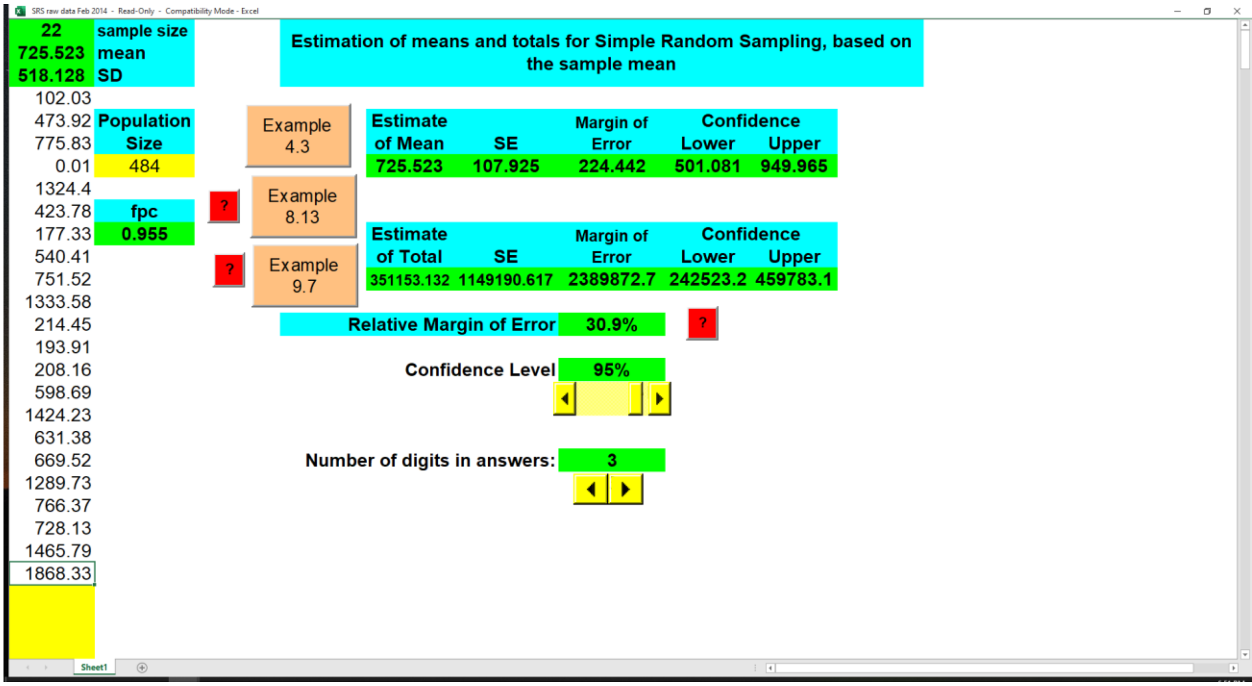
Confidence Level: 95%

Number of digits in answers: 3

i)

ii) Using a 95% confidence interval we get a mean of 140.891 and a margin of error of 57.876.

b) West Region:


SRS raw data Feb 2014 - Read-Only - Compatibility Mode - Excel
Example 4.3
Example 8.13
Example 9.7
?
?
?
Sheet1

22	sample size
725.523	mean
518.128	SD

102.03
473.92
775.83
0.01
1324.4
423.78
177.33
540.41
751.52
1333.58
214.45
193.91
208.16
598.69
1424.23
631.38
669.52
1289.73
766.37
728.13
1465.79
1868.33

Population Size: 484

fpc: 0.955

Estimation of means and totals for Simple Random Sampling, based on the sample mean

Estimate of Mean	SE	Margin of Error	Confidence Lower	Confidence Upper
725.523	107.925	224.442	501.081	949.965

Estimate of Total	SE	Margin of Error	Confidence Lower	Confidence Upper
351153.132	1149190.617	2389872.7	242523.2	459783.1

Relative Margin of Error: 30.9%

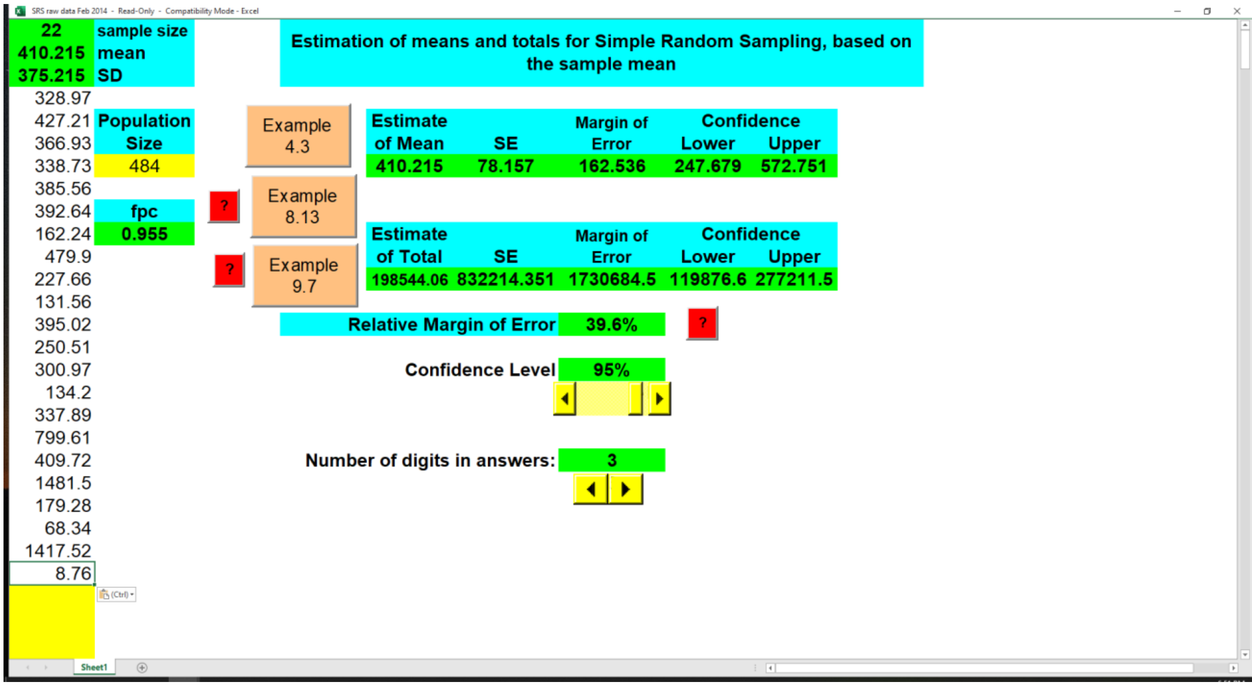
Confidence Level: 95%

Number of digits in answers: 3

i)

ii) Using a 95% confidence interval we get a mean of 725.523 and a margin of error at 224.442.

c) North Central Region:



SRS raw data Feb 2014 - Read-Only - Compatibility Mode - Excel

22	sample size
410.215	mean
375.215	SD

328.97
427.21
366.93
338.73
385.56
392.64
162.24
479.9
227.66
131.56
395.02
250.51
300.97
134.2
337.89
799.61
409.72
1481.5
179.28
68.34
1417.52
8.76

Population Size: 484

fpc: 0.955

Example 4.3

Example 8.13

Example 9.7

Estimation of means and totals for Simple Random Sampling, based on the sample mean

Estimate of Mean	SE	Margin of Error	Confidence Lower	Upper
410.215	78.157	162.536	247.679	572.751

Estimate of Total	SE	Margin of Error	Confidence Lower	Upper
198544.06	832214.351	1730684.5	119876.6	277211.5

Relative Margin of Error: 39.6%

Confidence Level: 95%

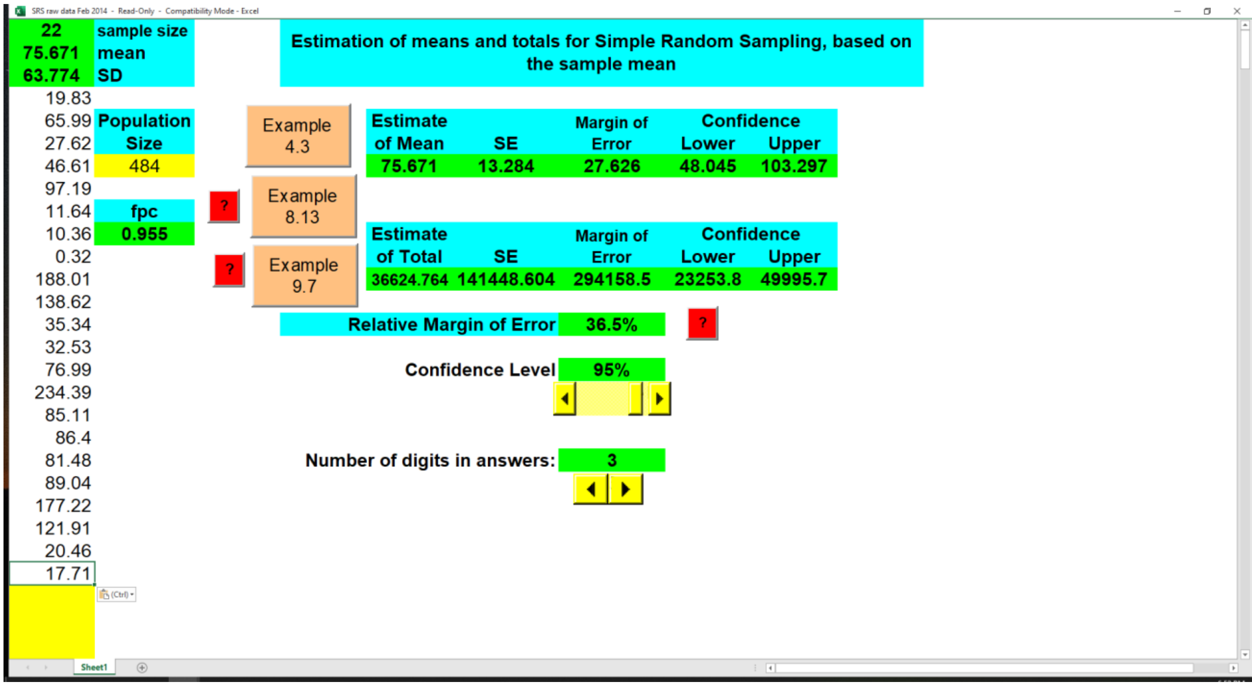
Number of digits in answers: 3

Sheet1

i)

ii) Using a 95% confidence interval we get a mean of 410.215 and a margin of error at 162.536.

d) North East Region:



SRS raw data Feb 2014 - Read-Only - Compatibility Mode - Excel

22	sample size
75.671	mean
63.774	SD

19.83
65.99
27.62
46.61
97.19
11.64
10.36
0.32
188.01
138.62
35.34
32.53
76.99
234.39
85.11
86.4
81.48
89.04
177.22
121.91
20.46
17.71

Population Size: 484

fpc: 0.955

Example 4.3

Example 8.13

Example 9.7

Estimation of means and totals for Simple Random Sampling, based on the sample mean

Estimate of Mean	SE	Margin of Error	Confidence Lower	Upper
75.671	13.284	27.626	48.045	103.297

Estimate of Total	SE	Margin of Error	Confidence Lower	Upper
36624.764	141448.604	294158.5	23253.8	49995.7

Relative Margin of Error: 36.5%

Confidence Level: 95%

Number of digits in answers: 3

Sheet1

i)

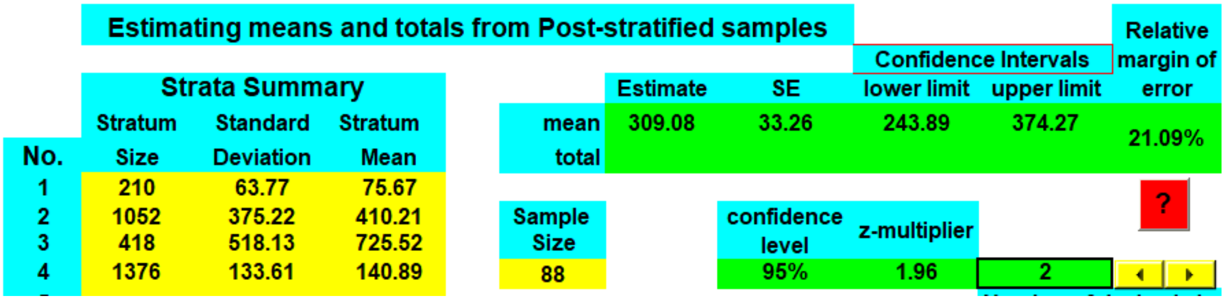
ii) Using a 95% confidence interval we get a mean of 75.671 and a margin of error at 27.626.

Part B

To find the total region we take the sample size times the mean:

- South Region:
  - Total acreage: 193865.89
  - Bound on error: $2\sqrt{(1376)^2(1-\frac{22}{1376})\frac{(133.608)}{22}} = 77762.4$
- West Region:
  - Total acreage: 303268.5
  - Bound on error:$2\sqrt{(418)^2(1-\frac{22}{418})\frac{(518.128)}{22}} = 89885.9$
- North Central Region:
  - Total acreage: 431545.7
  - Bound on error:$2\sqrt{(1052)^2(1-\frac{22}{1052})\frac{(375.215)}{22}}$ =166542.6
- North East Region:
  - Total acreage: 15890.99
  - Bound on error: $2\sqrt{(210)^2(1-\frac{22}{210})\frac{(63.774)^2}{22}}$ =5403.2

Part E



Estimating means and totals from Post-stratified samples

Strata Summary

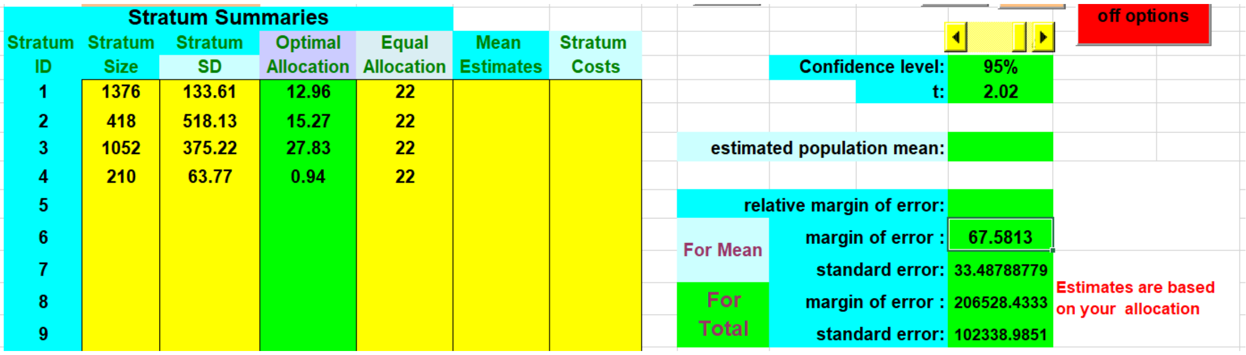
No.	Stratum Size	Standard Deviation	Stratum Mean
1	210	63.77	75.67
2	1052	375.22	410.21
3	418	518.13	725.52
4	1376	133.61	140.89

	Estimate	SE	Confidence Intervals lower limit	 upper limit	Relative margin of error
mean	309.08	33.26	243.89	374.27	21.09%
total					

Sample Size
88

confidence level	z-multiplier	
95%	1.96	2

The mean acreage per county across the United States is 309.08.



Stratum ID	Stratum Size	Stratum SD	Optimal Allocation	Equal Allocation	Mean Estimates	Stratum Costs
1	1376	133.61	12.96	22		
2	418	518.13	15.27	22		
3	1052	375.22	27.83	22		
4	210	63.77	0.94	22		
5						
6						
7						
8						
9						

Stratum Summaries

	Confidence level:	95%
	t:	2.02
	estimated population mean:	
	relative margin of error:	
For Mean	margin of error :	67.5813
	standard error:	33.48788779
For Total	margin of error :	206528.4333
	standard error:	102338.9851

Estimates are based on your allocation

Comparing spreadsheet with hand calculations:

$$= \frac{1}{(3056)^2}[(1376)^2(1-\frac{22}{1376})(\frac{(133.61)^2}{22}) + (418)^2(1-\frac{22}{418})(\frac{(518.13)^2}{22}) + (1052)^2(1 - \frac{22}{1052})(\frac{(375.22)^2}{22}) + (210)^2(1-\frac{22}{210})(\frac{(63.77)^2}{22})$$

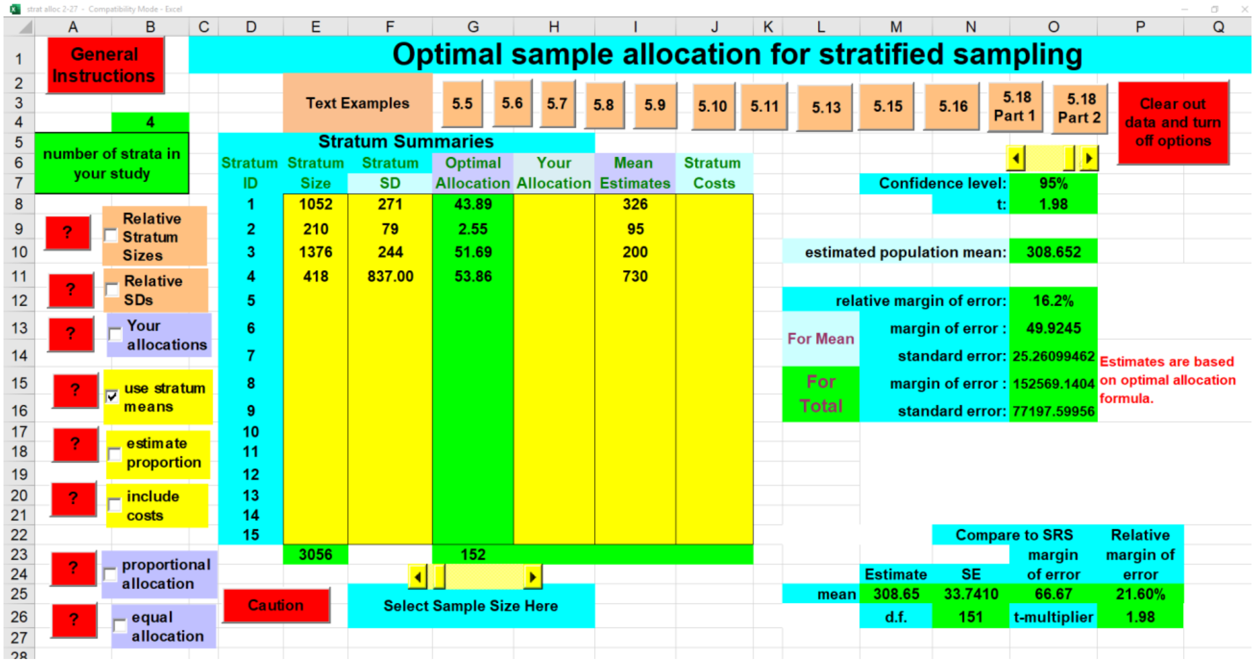
= 1121.4

= $2\sqrt{1121.4}$

= 66.97

The mean acreage per county across the United States is 309.08 and the bound on error is 66.97 or 309.08 ± 66.97

7. SMOG 5.36



Optimal sample allocation for stratified sampling

number of strata in your study: 4

□ Relative Stratum Sizes
□ Relative SDs
□ Your allocations
☑ use stratum means
□ estimate proportion
□ include costs
□ proportional allocation
□ equal allocation

Stratum Summaries

Stratum ID	Stratum Size	Stratum SD	Optimal Allocation	Your Allocation	Mean Estimates	Stratum Costs
1	1052	271	43.89		326	
2	210	79	2.55		95	
3	1376	244	51.69		200	
4	418	837.00	53.86		730	
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
	3056		152			

Confidence level: 95%
t: 1.98

estimated population mean: 308.652

relative margin of error: 16.2%

For Mean
margin of error : 49.9245
standard error: 25.26099462

For Total
margin of error : 152569.1404
standard error: 77197.59956

Estimates are based on optimal allocation formula.

	Compare to SRS			Relative
			margin	margin of
	Estimate	SE	of error	error
mean	308.65	33.7410	66.67	21.60%
	d.f.	151	t-multiplier	1.98

Using the allocation tool and plugging in all the known data, we can actually use this tool as a sample size calculator. Estimating mean acreage per county across the U.S., to get a margin of error of approximately 50,000 acres, we would use a sample size of 152. This gets us a margin of error of 49,924.5 acres (49.9245 in the tool because the data is in thousands).

The optimal allocation of these strata will be rounded to 44 for stratum 1, 2 for stratum 2, 52 for stratum 3, and 54 for stratum 4. The reason I rounded stratum 2 down to 2 instead of up is because I needed the four strata to add to 152 and the SD of stratum 2 is much lower than the other strata so it does not need as much attention as the other strata. To summarize, when estimating mean acreage, to get a margin of error of approximately 50,000 acres I would use a sample size of 152, with stratum 1 (North Central) being of size 44, stratum 2 (North East) being of size 2, stratum 3 (South) being of size 52, and stratum 4 (West) being of size 54.