Contents

SRS Sample Selection

We used the *RANDBETWEEN(1,51)* function on Excel to draw 10 random id numbers to produce a simple random sample (SRS) of 10 states; however, some were repeated. As we wanted SRS without replacement, we repeated the function until we obtained 10 unique ids. We copied the values into another column to stop the random number change and have a fixed sample - using the ids to find the corresponding states.

We selected the following sample:

Figure 1: SRS Selected Sample

RANDBETWEEN Func		Fixed id Number	Corresponding State
40		11	Georgia
6		8	Delaware
6		7	Connecticut
24		17	Kansas
8		16	Iowa
48		44	Texas
32		36	Ohio
51		12	Hawaii
19		29	Nevada
13		31	New Jersey

Excel Page: a

SRS Estimation of Mean

Based on the SRS obtained, we estimated the US 1992 mean per capita education expenditure to be **998.3**.

To calculate this in Excel, we used:

= AVERAGE(number1, number2)

The standard error was calculated to be **44.62** with a 95% confidence interval of **(910.83, 1085.77)**, meaning that we are 95% confident that the interval contains the actual value of the US 1992 mean per capita education expenditure.

We calculated the variance using:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y_s})^2$$

Then to obtain S.E:

$$S.E. = \sqrt{\frac{s^2(1-f)}{n}}$$

From this, we calculated the lower and upper-bound using:

$$(\bar{y_s} - Z_{1-\frac{\propto}{2}} \times S.E\ ,\ \bar{y_s} + Z_{1-\frac{\propto}{2}} \times S.E)$$

The standard error, 44.62, is quite sizeable, consequently, the width of the confidence interval is relatively large. This suggests that the, US 1992 mean per capita education expenditure that we calculated is imprecise. The reason for the large standard error may be because at times, random selection can make the sample unrepresentative

Figure 2: Figures and Results for Mean Estimations using SRS

id	State	ExpPP90	ExcPC90	TeaSal90	%Compl	ExpPP92	ExpPC92	TeaSal92	%Dropout	Region	Popultn	Enroll	Teachers
8	Delaware	5848	835	33.4	80.7	6080	899	34.5	11.2	3	0.7	106	6.1
17	Kansas	4706	819	28.7	82.2	5131	941	30.7	8.4	2	2.5	437	29.3
44	Texas	4056	849	27.5	74.3	4651	988	29	12.5	3	17.6	3383	212.6
31	New Jersey	8439	1114	35.7	79.4	10219	1337	41	9.3	1	7.8	1090	80.5
7	Connecticut	7934	1129	40.5	80.6	8299	1227	47	9.2	1	3.3	469	34.8
16	Iowa	4590	777	26.7	83.4	4949	882	29.2	6.5	2	2.8	484	31.5
29	Nevada	4387	878	30.6	84	4910	1006	33.9	14.9	4	1.3	201	11.4
36	Ohio	4394	720	31.2	77.6	5451	921	33.3	8.8	2	11	1772	103.2
11	Georgia	4456	796	27.9	71.1	4720	884	29.5	14.1	3	6.8	1152	70.3
12	Hawaii	4504	710	32	82.3	5453	898	34.5	7	4	1.2	172	10

id	State	ExpPC92	y_i-mean	$(y_i\text{-mean})^2$
11	Georgia	884	-114.3	13064.49
8	Delaware	899	-99.3	9860.49
7	Connecticut	1227	228.7	52303.69
17	Kansas	941	-57.3	3283.29
16	Iowa	882	-116.3	13525.69
44	Texas	988	-10.3	106.09
36	Ohio	921	-77.3	5975.29
12	Hawaii	898	-100.3	10060.09
29	Nevada	1006	7.7	59.29
31	New Jersey	1337	338.7	114717.69

Sample Mean		998.3
Sum of $(y_i\text{-mean})^2$		222956.1
Variance of Sample		24772.9
Variance of Mean		1991.546863
Standard Error		44.62675053
95% CI:	Upper bound	1085.768431
	Lower bound	910.831569

Excel Page: b

Stratified Sampling vs Cluster Sampling

Stratified Sampling

Stratified sampling is the division of sampling units into strata (non-overlapping groups) according to specific characteristics; SRS are then drawn from the strata. Stratified sampling requires large differences between each stratum (heterogeneity) but small differences between units within the strata (homogeneity). The greater the difference between the strata, the greater the precision.

Here, we can stratify the population (states) by region or teacher salary. Then we can select an SRS from these strata using proportional allocation, as we do not know the variance of each stratum.

The benefits of using stratified sampling here are the following:

1. Precise estimation of subgroups
2. Avoids Bias
3. Expenditure per capita is highly dependent on factors such as
    a. Ratio between students and teachers
    b. Average teacher salary

    Etc.

The downsides of using stratified sampling here are the following:

1. Have limited information available to best create homogeneous strata
    a. Therefore, classifications between groups may not be clear
    b. Therefore, may lead to inefficient allocation
2. Size of sample must respect budget constraint

Cluster Sampling

Cluster sampling combines several population units into N groups, 'clusters'. An SRS of n clusters is then taken, and all elements within the chosen clusters are sampled. This method requires minor differences between clusters (homogeneity) and large differences within (heterogeneity). Here we would have to use multi-stage cluster sampling to select our sample. We instead take an SRS of the

clusters and then an SRS of states from the chosen clusters.

We could group each state into one of four clusters based on region - Northeast, Midwest, South and West; the clusters would be different sizes; as there are 9 states in the Northeast, 12 in the Midwest, 17 in the South, and 13 in the West. Our first stage would be selecting an SRS of size 2 through proportional allocation. In our second stage, we could again use SRS with proportional allocation to choose 10 states *within* the two selected clusters.

The benefits of using cluster sampling here are the following:

1. Allows samples to be taken when sampling frame is difficult to form:
    a. But here, a frame listing the clusters is easily found
    b. Thus, it's practically convenient to select samples in clusters
2. Requires fewer resources:
    a. Therefore, would be easier and cheaper to conduct
3. Would be more precise in comparison to just SRS
    a. Particularly in using proportional allocation to select clusters and then sampling units

The downsides of using cluster sampling here are the following:

1. Could be prone to bias if clusters are formed with biased opinion
    a. However, this *could not* occur here as clusters are already formed by region
2. Tendency for higher sampling error than in other methods
3. In this case, we have too few clusters to appropriately use SRS for selecting them

Stratified Random Sampling (STS)

To further analyse the data, we turn to stratification sampling, which involves the separation of the population into different groups, strata. As aforementioned, to do this, we first divide the population into strata with the following requirements:

i. Small within-group variance

ii. Large between-group variance

These requirements are so that the samples selected from each stratum represent the whole stratum, thus providing more precise estimations for the entire population. For our data, we could sample 10 representative states to estimate all 51.

The procedure for selecting our sample was as follows:

1. Choose stratification factor
2. Build stratus
3. Calculate within and between-group variance to ensure strata are correctly formed
4. Determine sample size in each stratum
5. Select 10 samples

First, we chose teacher salary as the stratification factor - using the variable ExpPC92 to assess it. Intuitively, we assumed the relationship that: the higher the teacher's salary, the higher the 1992 expenditure per capita (ExpPC92). We then built three strata (as shown in *Figure 3* below) according to the teacher salary ranges.

Figure 3: H_1, H_2 & H_3 Stratas

	H1			H2			H3		
id	**State**	**TeaSal92**	**id**	**State**	**TeaSal92**	**id**	**State**	**TeaSal92**
42	S. Dacota	23.3	44	Texas	29	24	Minnesota	33.7
25	Mississippi	24.4	16	Iowa	29.2	29	Nevada	33.9
35	N. Dakota	24.5	34	N. Carolina	29.2	38	Oregon	34.1
37	Oklahoma	25.3	11	Georgia	29.5	8	Delaware	34.5
13	Idaho	26.3	20	Maine	30.1	12	Hawaii	34.5
45	Utah	26.5	51	Wyoming	30.4	15	Indiana	34.8
4	Arkansas	26.6	17	Kansas	30.7	48	Washington	34.8
32	New Mexico	26.7	18	Kentucky	30.9	50	Wisconsin	35.2
1	Alabama	27	10	Florida	31.1	40	Rhode Island	36
19	Louisiana	27	3	Arizona	31.2	41	Illinois	36.5
28	Nebraska	27.2	47	Virginia	31.9	42	Massachusett	37.3
49	West Virginia	27.4	6	Colorado	33.1	43	Pennsylvania	38.7
27	Montana	27.6	30	New Hampshi	33.2	44	Maryland	39.5
41	S. Carolina	28.3	36	Ohio	33.3	45	California	40.2
43	Tennessee	28.6	46	Vermont	33.6	46	New Jersey	41
26	Missouri	28.9				47	Michigan	41.1
						48	DC	41.3
						49	New York	43.3
						50	Alaska	44.7
						51	Connecticut	47

Excel Page: c

We did not have the data to calculate the within-strata and between-strata variance; therefore, we will only briefly introduce them by providing the formula below.

Within-strata variance:

$$S_w^2 = \sum_{H=1}^{3} \sum_{i=1}^{ni} (y_{ij} - \overline{y_i})^2$$

Between-strata variance:

$$S_b^2 = \sum_{H=1}^{3} n_i (\overline{y_i} - \overline{y})^2$$

$\overline{y_i}$ = sample mean

y_{ij} = sample unit

$\overline{y}$ = population mean

n_i = sample size

Following this, we determined the sample size of each stratum using proportional estimation as we did not know the variance for each stratum.

The formula for proportional allocation we used is:

$$n_h = n \times \frac{N_h}{N}$$

Figure 4 presents the calculated values and rounded sample sizes:

Figure 4: Strata Sample Sizes

Strata	Teacher Salary Range	Strat size(NH)	Sample size	Sample size(*)
H1	[23,29)	16	3.137254902	3
H2	[29,33.7)	15	2.941176471	3
H3	[33.7,47]	20	3.921568627	4

Excel Page: c

Finally, we selected samples from each stratum by applying SRS, obtaining the following sample:

Figure 5: STS Sample

id	State	ExpPP90	ExcPC90	TeaSal90	%Compl	ExpPP92	ExpPC92
28	Nebraska	3874	652	25.5	82.2	4676	866
4	Arkansas	3272	621	22	67.6	3770	758
37	Oklahoma	3484	683	23.1	75.4	3939	814
11	Georgia	4456	796	27.9	71.1	4720	884
17	Kansas	4706	819	28.7	82.2	5131	941
16	Iowa	4590	777	26.7	83.4	4949	882
50	Wisconsin	5703	887	31.9	81.1	5972	976
8	Delaware	5848	835	33.4	80.7	6080	899
48	DC	7407	945	38	72.9	8116	1059
50	Wisconsin	5703	887	31.9	81.1	5972	976

Excel Page: d

Although we could not use Neyman allocation here, we also considered it.

Neyman allocation formula:

$$n_h = n \times \frac{N_h S_h}{\sum N_h S_h}$$

S_h can be calculated by square rooting S_h^2:

$$S_h^2 = \frac{1}{(n_h - 1)} \sum_{j=1}^{N_h} (y_{hj} - \bar{y}_h)^2$$

As in proportional allocation, we use a sample to 'mirror' the population. If we had sample variances it would have been optimal to use Neyman allocation rather than proportional allocation, as it is designed to maximise the precision of the estimation within the given budget.

Stratified Sampling Estimation of Mean

Based on the stratified sample selected, we estimated the US 1992 mean per capita education expenditure to be **905.5**.

To calculate this, we used:

$$\overline{y_s} = \frac{\sum_{i=1}^{n} y_i}{n} = 905.5$$

To find S.E, we first found sample variances of each stratum, *h*:

$$S_h^2 = \frac{1}{(n_h - 1)} \sum_{j=1}^{nh} (y_{hj} - \overline{y_{hu}})^2$$

$$S_{h1}^2 = 2917.33,\ S_{h2}^2 = 1122.33,\ S_{h3}^2 = 4269.67$$

Figure 6: Stratum Mean and Variance

	Mean	Variance	Var(t str)
h1	812.67	2,917.33	202,268.44
h2	902.33	1,122.33	67340
h3	977.50	4,269.67	341,573.33

Excel Page: d

Thus, the variance of the estimated total is:

$$\hat{V}(\hat{t}_{str}) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h} = 611{,}181.78$$

The variance of the mean is equal to $1/N^2$ of the above:

$$var(\bar{y}_{str}) = \frac{1}{N^2} \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h} = 234.98$$

Finally, S.E is the square root of the variance of the mean:

$$S.E = \sqrt{s^2 y} = \mathbf{15.33}$$

Figure 7: Mean, Variance & S.E using STS

Mean							905.5
Variance							235.0
S.E							15.3

Excel Page: d

The figures calculated in Excel are shown below:

Figure 8: STS Figures & Calculations

id	State	ExpPP90	ExcPC90	TeaSal90	%Compl	ExpPP92	ExpPC92
28	Nebraska	3874	652	25.5	82.2	4676	866
4	Arkansas	3272	621	22	67.6	3770	758
37	Oklahoma	3484	683	23.1	75.4	3939	814
11	Georgia	4456	796	27.9	71.1	4720	884
17	Kansas	4706	819	28.7	82.2	5131	941
16	Iowa	4590	777	26.7	83.4	4949	882
50	Wisconsin	5703	887	31.9	81.1	5972	976
8	Delaware	5848	835	33.4	80.7	6080	899
48	DC	7407	945	38	72.9	8116	1059
50	Wisconsin	5703	887	31.9	81.1	5972	976
Mean							905.5
Variance							235.0
S.E							15.3
95% C.I (Z)							1.96
Lower bound							875.5
Upper bound							935.5

Excel Page: d

After reaching these values, we can see that the standard error for STS is much smaller than that for SRS, as 44.6 (SRS) is 2.9x greater than **15.3** (STS), indicating that the latter method is better and more precise.

Ratio and Regression Estimation of Mean

We can consider any variable provided to us in the survey; if there are two with which a specific correlation is implied, and we know all the information about one, then we can use this variable to estimate the other.

Here, our variable of interest *(y)* is the 1992 mean per capita education expenditure. It is assumed to be associated with the auxiliary variable *(x)*, the 1990 mean per capita education expenditure. Two kinds of estimations can be used: ratio and regression estimation.

We first used **ratio estimation**:

Figure 9: Ratio Estimation Figures & Calculations

id	State	ExpPP90	ExcPC90(X)	Xi-Xs(Mean)	TeaSal90	%Compl	ExpPP92	ExpPC92=(Y)	Yi-Ys(Mean)	E*J	TeaSal92	%Dropout	Region	Popultn	Enroll	Teachers
11	Georgia	4456	796	-66.7	27.9	71.1	4720	884	-114.3	7623.81	29.5	14.1	3	6.8	1152	70.3
8	Delaware	5848	835	-27.7	33.4	80.7	6080	899	-99.3	2750.61	34.5	11.2	3	0.7	106	6.1
7	Connecticut	7934	1129	266.3	40.5	80.6	8299	1227	228.7	60902.81	47	9.2	1	3.3	469	34.8
17	Kansas	4706	819	-43.7	28.7	82.2	5131	941	-57.3	2504.01	30.7	8.4	2	2.5	437	29.3
16	Iowa	4590	777	-85.7	26.7	83.4	4949	882	-116.3	9966.91	29.2	6.5	2	2.8	484	31.5
44	Texas	4056	849	-13.7	27.5	74.3	4651	988	-10.3	141.11	29	12.5	3	17.6	3383	212.6
36	Ohio	4394	720	-142.7	31.2	77.6	5451	921	-77.3	11030.71	33.3	8.8	2	11	1772	103.2
12	Hawaii	4504	710	-152.7	32	82.3	5453	898	-100.3	15315.81	34.5	7	4	1.2	172	10
29	Nevada	4387	878	15.3	30.6	84	4910	1006	7.7	117.81	33.9	14.9	4	1.3	201	11.4
31	New Jersey	8439	1114	251.3	35.7	79.4	10219	1337	338.7	85115.31	41	9.3	1	7.8	1090	80.5
f i ratio estimation																
Sum of Xi			8627													
Sum of Yi								9983								
B=Y/X																
B(r) = Sum of Yi/Sum of Xi		1.157181														
Mean of X(u)			818.4314													
Xs(Mean)			862.7													
Ys(Mean)								998.3								
sx^2		21404.46														
sx		146.3026														
sy^2				24772.9												
sy				157.3940914												
r		0.94318														
Mean of Y(r)		947.0732														
Var(mean of Y(r))		254.8295														
S.E(mean of Y(r))		15.96338														

Excel Page: e

We set the auxiliary variable *(x)*, the interest variable *(y)*, and the sample size - and assumed that *Y* is approximately proportional to *X*.

Figure 10: Variable Setting

Excpc90=X	Exppc92=Y	n=10

Excel Page: e

Then we estimated the ratio $\hat{B}$, of characteristics *Y/X*, using:

$$\frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i} \text{ x}$$

under SRS

Figure 11: B_r Calculation

Sum of Xi	8627
Sum of Yi	9983
B(r) = Sum of Yi/Sum of Xi	1.1571809

Excel Page: e

We estimated the ratio of the population mean with:

$$\widehat{\overline{Y_r}} = \widehat{B_r} * \overline{X_u}$$

where $\overline{X_u}$ is the population mean of *X* calculated by:

$$\frac{\sum_{i=1}^{N} X_i}{N}$$

Figure 12: Ratio Estimator of Population Mean

Mean of X(u)			818.43137
Mean of Y(r)		947.07319	

Excel Page: e

Then we calculated the estimated variance of $\overline{Y_r}$ and its S.E:

$$Var(\widehat{\overline{Y_r}}) = \frac{1-f}{n} \frac{\sum_{i=1}^{n} (Y_i - \widehat{B_r} * X_i)^2}{n-1} = \frac{1-f}{n} \frac{\sum {Y_i}^2 - 2\widehat{B_r} \sum Yi * Xi + \widehat{B_r}^2 \sum {X_i}^2}{n-1}.$$

where $f = \frac{n}{N}$, and *B(r)* is the ratio estimator of *B* obtained in the previous step.

It can also be written as:

$$Var(\widehat{\overline{Y_r}}) = \frac{1-f}{n} ({S_y}^2 - 2BRS_xS_y + B^2{S_x}^2)$$

where *S*, *R* and *B* are population variables.

As we only have sample information, we estimated the variance of $\widehat{Var}(\widehat{\overline{Y}_r})$ by:

$$\frac{1-f}{n}\left(s_y{}^2 - 2\widehat{B_r}s_x s_y + \widehat{B_r}s_x{}^2\right)$$

Substituting *S*, *R* and *B* with the sample variance *s*, sample correlation coefficient *r*, and ratio estimator *B(r),* respectively, using:

$$s_y{}^2 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y_s})^2}{n-1}$$

$$s_x{}^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X_s})^2}{n-1}$$

$$r = \frac{S_{xy}}{\sqrt{S_x S_y}} = \frac{\sum_{i=1}^{n}(X_i - \overline{X_s})(Y_i - \overline{Y_s})}{\sqrt{\{\sum_{i=1}^{n}(X_i - \overline{X_s})^2\}\{\sum_{i=1}^{n}(Y_i - \overline{Y_s})^2\}}}$$

Below are the pre-calculations used:

Figure 13: Ratio Estimation Pre-calculations

	Xi-Xs(mean)	(Xi-Xs(mean))^2	Yi-Ys(mean)	(Yi-Ys(mean))^2	(Xi-Xs(mean))*(Yi-Ys(mean))
	-66.7	4448.89	-114.3	13064.49	7623.81
	-27.7	767.29	-99.3	9860.49	2750.61
	266.3	70915.69	228.7	52303.69	60902.81
	-43.7	1909.69	-57.3	3283.29	2504.01
	-85.7	7344.49	-116.3	13525.69	9966.91
	-13.7	187.69	-10.3	106.09	141.11
	-142.7	20363.29	-77.3	5975.29	11030.71
	-152.7	23317.29	-100.3	10060.09	15315.81
	15.3	234.09	7.7	59.29	117.81
	251.3	63151.69	338.7	114717.69	85115.31
sx^2		21404.45556			
sx		146.3026164			
sy^2				24772.9	
sy				157.3940914	
sum(Yi-Ys(mean))*(Xi-Xs(mean))					195468.9
r		0.943180128			

Excel Page: e

Finally, we can see that the mean of ExpPC92 is **947.07**, and its variance is 254.83. The standard error is the square root of the variance, **15.96**.

Figure 14: Ratio Estimation Mean, Variance & S.E

Mean of Y(r)	947.0732
Var(mean of Y(r))	254.8295
S.E(mean of Y(r))	15.96338

Excel Page: e

We then used **regression estimation**:

Figure 15: Regression Estimation Figures & Calculations

id	State	ExpPP90	ExcPC90(X)	Xi-Xs(mean)	TeaSal90	%Compl	ExpPP92	ExpPC92=(Y)	Yi-Ys(mean)	E*J	TeaSal92	%Dropout	Region	Popultn	Enroll	Teachers
11	Georgia	4456	796	-66.7	27.9	71.1	4720	884	-114.3	7623.81	29.5	14.1	3	6.8	1152	70.3
8	Delaware	5848	835	-27.7	33.4	80.7	6080	899	-99.3	2750.61	34.5	11.2	3	0.7	106	6.1
7	Connecticut	7934	1129	266.3	40.5	80.6	8299	1227	228.7	60902.81	47	9.2	1	3.3	469	34.8
17	Kansas	4706	819	-43.7	28.7	82.2	5131	941	-57.3	2504.01	30.7	8.4	2	2.5	437	29.3
16	Iowa	4590	777	-85.7	26.7	83.4	4949	882	-116.3	9966.91	29.2	6.5	2	2.8	484	31.5
44	Texas	4056	849	-13.7	27.5	74.3	4651	988	-10.3	141.11	29	12.5	3	17.6	3383	212.6
36	Ohio	4394	720	-142.7	31.2	77.6	5451	921	-77.3	11030.71	33.3	8.8	2	11	1772	103.2
12	Hawaii	4504	710	-152.7	32	82.3	5453	898	-100.3	15315.81	34.5	7	4	1.2	172	10
29	Nevada	4387	878	15.3	30.6	84	4910	1006	7.7	117.81	33.9	14.9	4	1.3	201	11.4
31	New Jersey	8439	1114	251.3	35.7	79.4	10219	1337	338.7	85115.31	41	9.3	1	7.8	1090	80.5
f ii regresssion estimation																
B1(reg)=sum(Yi-Ys(mean))*(Xi-Xs(mean))/sum(Xi-Xs(mean)^2																
B0(reg)=Ys(mean)-B1(reg)*Xs(mean)																
Xs(mean)			862.7													
Ys(mean)								998.3								
B1(reg)		1.0146844														
B0(reg)		122.93179														
Yi=122.93+1.01*Xi																
Calculate Mean of Y(u) by Regresssion Estimation																
Mean of Y(reg)								953.3813153								
Sy^2								24772.9								
Correlation coefficient(from f(i))								0.943180128								
Var(mean of Y(reg))								219.8891708								
S.E(mean of Y(reg))								14.82866045								

Excel Page: e

If the variable of interest Y and the auxiliary variable X are linearly related, but the line does pass through the origin, they are not perfectly proportional; in this case, ratio estimation is not appropriate. Therefore, regression estimation must be introduced. Here, our auxiliary variable, X, is ExpPC90, and our interest variable, Y, is ExpPC92.

The equation showing that Y_i and X_i are linearly related is:

$$Y_i = B_{0(reg)} + B_{1(reg)} \times X_i$$

when $B_0 = 0$

To calculate the mean of y, we went through the following steps:

1. Calculated $\widehat{B_{0(reg)}}$ and $\widehat{B_{1(reg)}}$ with:

$$\widehat{B_{0(reg)}} = \overline{Y_s} + \widehat{B_{1(reg)}} \times \overline{X_s}$$

$$\widehat{B_{1(reg)}} = \frac{\sum_{i=1}^{n}(X_i - \overline{X_s})(Y_i - \overline{Y_s})}{\sum_{i=1}^{n}(X_i - \overline{X_s})^2}$$

To do this, we needed the following pre-calculations:

Figure 17: Regression Estimation Pre-calculations

	Xi-Xs(mean)	(Xi-Xs(mean))^2	Yi-Ys(mean)	(Yi-Ys(mean))^2	Xi-Xs(mean)*Yi-Ys(mean)
	-66.7	4448.89	-114.3	13064.49	7623.81
	-27.7	767.29	-99.3	9860.49	2750.61
	266.3	70915.69	228.7	52303.69	60902.81
	-43.7	1909.69	-57.3	3283.29	2504.01
	-85.7	7344.49	-116.3	13525.69	9966.91
	-13.7	187.69	-10.3	106.09	141.11
	-142.7	20363.29	-77.3	5975.29	11030.71
	-152.7	23317.29	-100.3	10060.09	15315.81
	15.3	234.09	7.7	59.29	117.81
	251.3	63151.69	338.7	114717.69	85115.31
sum(Yi-Ys(mean))*(Xi-Xs(mean))					195468.9
sum(Xi-Xs(mean)^2		192640.1			

Excel Page: e

2. Estimated $\widehat{B_{0(reg)}}$ and $\widehat{B_{1(reg)}}$ and formed the Y_i function

Figure 18: B_1 & B_0 Values

B1(reg)		1.0146844
B0(reg)		122.93179
Yi=122.93+1.01*Xi		

Excel Page: e

3. Estimated regression estimator of population mean and found its variance and S.E, using:

$$\widehat{\overline{Y_{reg}}} = \overline{Y_s} + \widehat{B_{1(reg)}}(\overline{X_u} - \overline{X_s})$$

$$\widehat{Var}(\widehat{\overline{Y_{reg}}}) = \frac{1-f}{n} \times s_y{}^2 \times (1\text{-}r^2)$$

Figure 19: Mean, Variance & S.E using Regression Estimation

Mean of Y(reg)	953.3813153
Sy^2	24772.9
Correlation coefficient(from f(i))	0.943180128
Var(mean of Y(reg))	219.8891708
S.E(mean of Y(reg))	14.82866045

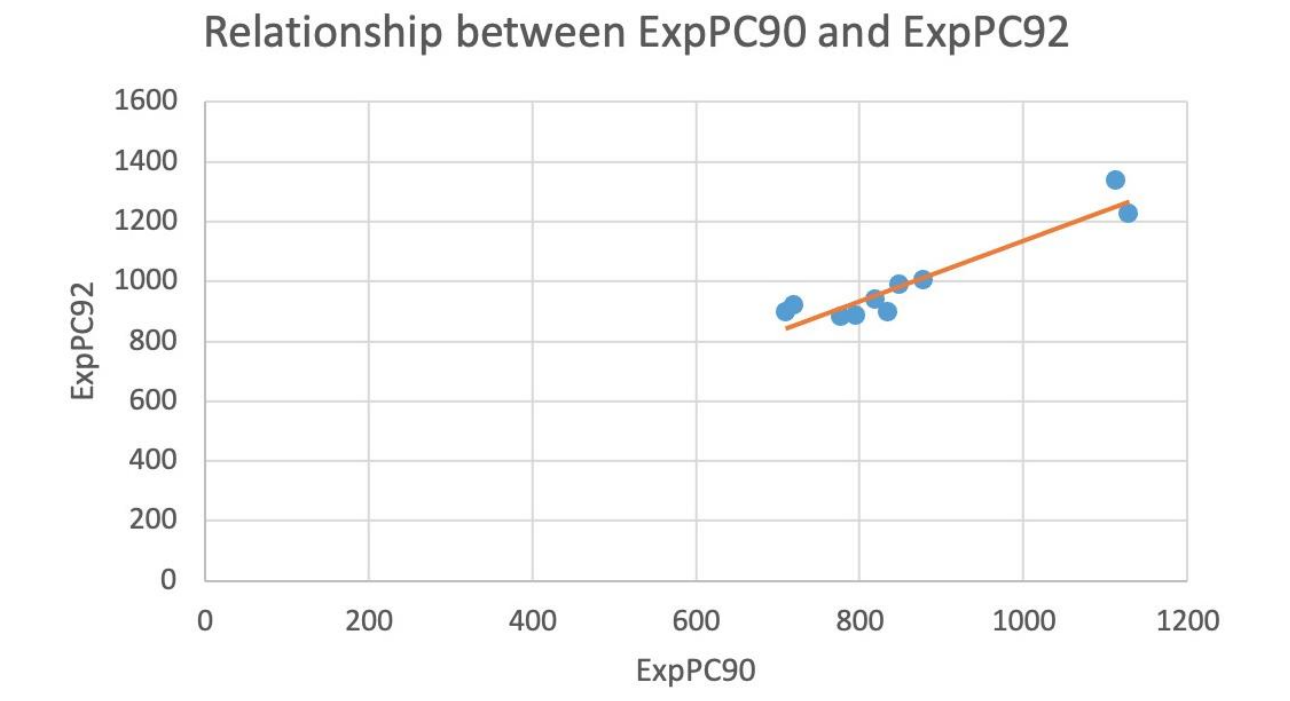
Excel Page: e

4. Finally, calculated that the regression estimate of the mean of ExpPC92 was **953.38,** and its variance was 219.89; the S.E is the square root of the variance, **14.83**.

In comparing ratio and regression estimates, If *Y* and *X* are perfectly (and positively) proportional and the best-fit line passes through the origin, ratio estimation is more appropriate. Regression estimation would instead be better suited if they are linearly related and the best-fit line does not pass through the origin but intercepts the *y*-axis.

Using the data, we can draw a graph between ExpPC90 and ExpPC92:

Figure 20: Expenditure Linear-Regression Scatter Graph



Excel Page: e

We can see the line does not pass through (800,800), so regression estimation is better suited here. We can also see from that $\widehat{Var}(\widehat{\overline{Y}_r})$ is greater than $\widehat{Var}(\widehat{\overline{Y}_{reg}})$, meaning the regression estimator is more accurate. However, the sample size, *n*, of 10, can be considered small. $\hat{B}$ is always a biased estimator of *B*, but the bias becomes smaller as *n* becomes larger. Therefore, a larger *n* is better to minimise bias.

Comparing Estimates

As we do not know the actual population mean of *Y*, we cannot directly compare the estimates (SRS, ratio, and regression estimates) we previously calculated to the actual value. So, we instead use the standard error and confidence intervals. A confidence interval shows an *x*% probability that the actual value of what you are estimating lies within it; the interval obtained for the regression estimate is the narrowest (so it has the smallest S.E), suggesting it is the most precise method.

We can also compare the ratio estimator with the unbiased estimate previously obtained using SRS by using the covariance, *CV*, and the correlation coefficient between *x* and *y*, *R*. When *n* is large, if $R > \frac{1}{2}\frac{CV_X}{CV_y}$, then the ratio estimator will be superior to the SRS estimator. But in this case, *n* is only 10; accordingly, we cannot calculate CV_y and *R*, so we use the sample to estimate it:

$$r > \frac{1}{2}\frac{cv_x}{cv_y}$$

where:

$$r = \frac{\sum_{i=1}^{n}(X_i-\overline{X_s})(Y_i-\overline{Y_s})}{\sqrt{\{\sum_{i=1}^{n}(X_i-\overline{X_s})^2\}\{\sum_{i=1}^{n}(Y_i-\overline{Y_s})^2\}}} \qquad cv_x = \frac{s_x}{\overline{x_s}} \qquad cv_y = \frac{s_y}{\overline{y_s}}$$

Figure 21: Comparing Ratio & SRS Estimators

CVx=sx/xs (mean)		0.169586897		
CVy=sy/ys (mean)				0.157662117
r		0.943180128		
Cvx/CVy		1.075635039		
0.5*CVx/CVy		0.53781752		
r > 0.5*CVx/CVy	so the ratio estimator is more accurate than the unbiased estimator			

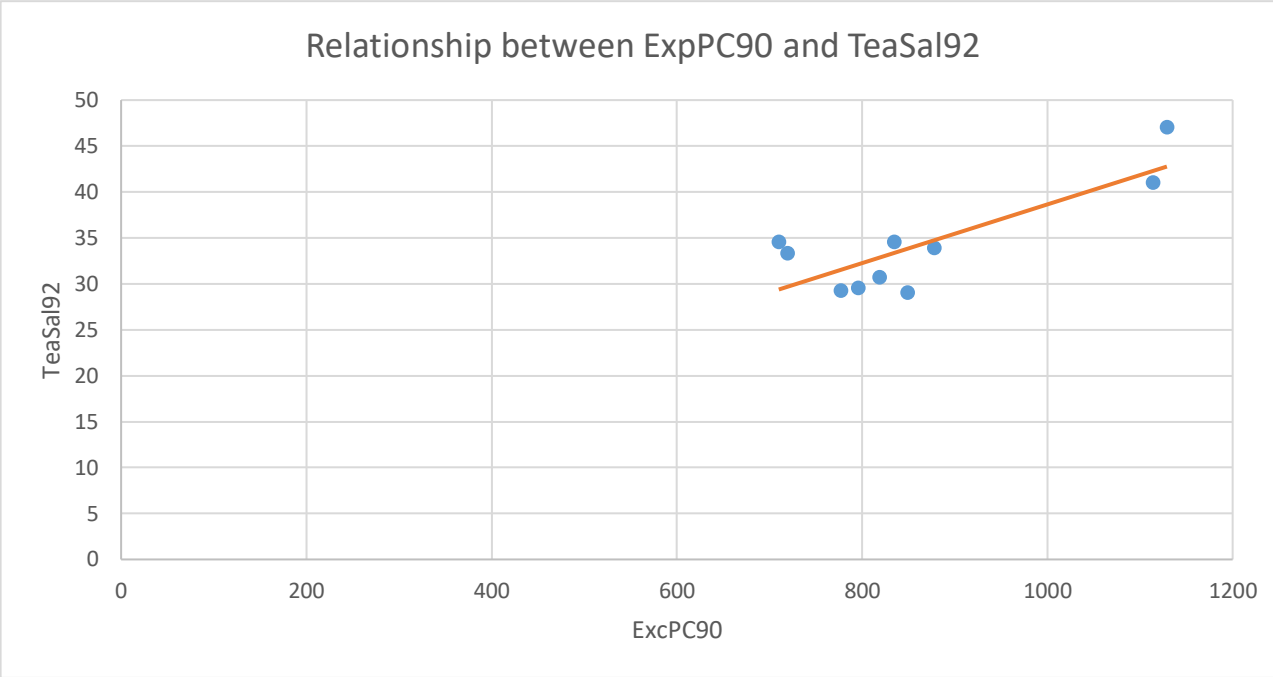
Excel Page: e

Figure 21 shows that the ratio estimator is more accurate than the SRS estimator. Overall, in this case, the regression estimator is the best of the three - SRS, regression, and ratio estimators - we have so far considered.

Mean 1992 Teacher Salary Estimation

To estimate the 1992 mean teacher salary in the US, we used the 1990 per capita expenditure (ExcPC90) using the original sample obtained from SRS. To choose the best method to estimate, we produced a diagram (*Figure 22*) to visualise the relationship between the variables (ExcPC90 (*x*) and TeaSal92 *(y)*).

Figure 22: Teacher Salary Linear-Regression Scatter Graph



Excel Page: f

As the graph exhibits linearity between the variables but does not pass (0,0), regression estimation, is best suited.

To do this, we first calculated $\overline{y_s}$ and $\overline{x_s}$.

$$\overline{y_s} = \frac{\sum_{i=1}^{n} y_i}{n} = 34.26$$

$$\overline{x_s} = \frac{\sum_{i=1}^{n} x_i}{n} = 862.7$$

Then we calculated $\widehat{B_{0(reg)}}$ and $\widehat{B_{1(reg)}}$ using:

$$\widehat{B_{0(reg)}} = \overline{Y_s} + \widehat{B_{1(reg)}} \times \overline{X_s}$$

$$\widehat{B_{1(reg)}} = \frac{\sum_{i=1}^{n}(X_i-\overline{X_s})(Y_i-\overline{Y_s})}{\sum_{i=1}^{n}(X_i-\overline{X_s})^2}$$

Figure 23: B1 & B0 Values

B1	0.031942363
B0	6.703323451

Excel Page: f

We then applied the regression estimation formula to get $\widehat{\overline{Y_{reg}}} = \mathbf{32.8}$

$$\widehat{\overline{Y_{reg}}} = \overline{Y_s} + \widehat{B_{1(reg)}}(\overline{X_u} - \overline{X_s}) = 32.8$$

Figure 24 presents all the other values:

Figure 24: Mean Teacher Salary Figures & Calculations

id	State	ExcPC90(x)	TeaSal92(y)
8	Delaware	835	34.5
17	Kansas	819	30.7
44	Texas	849	29
31	New Jersey	1114	41
7	Connecticut	1129	47
16	Iowa	777	29.2
29	Nevada	878	33.9
36	Ohio	720	33.3
11	Georgia	796	29.5
12	Hawaii	710	34.5
Mean		862.7	34.26
xu	818.4313725		
B1	0.031942363		
B0	6.703323451		
Correlation Coeffient	0.812819256		
yreg	32.84595543		

Excel Page: f

Student-to-Teacher Ratio Estimation

We calculated the student-to-teacher ratio of our original sample (selected through SRS) to be **15.90** with low variability. We can also say that our estimate of the whole population is relatively precise, as we obtained a very small standard error of **0.44**.

To calculate this, we first found s^2 using:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y_s})^2$$

Then to obtain S.E we used:

$$S.E. = \sqrt{\frac{s^2(1-f)}{n}}$$

Then we calculated the confidence interval by:

$$(\bar{y_s} + Z_{1-\frac{\propto}{2}} \times S.E.,\ \bar{y_s} - Z_{1-\frac{\propto}{2}} \times S.E.)$$

Thereby establishing with 95% confidence that the student-teacher ratio in the US is between **15.04 and 16.75**. Therefore, most states are estimated to have a similar number of teachers to students.

Figure 25: Teacher-to-Student Ratio Estimation Figures & Calculations

id	State	%Dropout	Enroll	Teachers	Student - Teacher Ratio		y_i-mean	$(y_i\text{-mean})^2$
8	Delaware	11.2	106	6.1	17.37704918		1.479475675	2.188848274
17	Kansas	8.4	437	29.3	14.91467577		-0.982897737	0.966087961
44	Texas	12.5	3383	212.6	15.91251176		0.014938254	0.000223151
31	New Jersey	9.3	1090	80.5	13.54037267		-2.357200834	5.556395773
7	Connecticut	9.2	469	34.8	13.47701149		-2.420562011	5.859120448
16	Iowa	6.5	484	31.5	15.36507937		-0.53249414	0.283550009
29	Nevada	14.9	201	11.4	17.63157895		1.734005442	3.006774874
36	Ohio	8.8	1772	103.2	17.17054264		1.272969131	1.620450408
11	Georgia	14.1	1152	70.3	16.38691323		0.489339724	0.239453366
12	Hawaii	7	172	10	17.2		1.302426495	1.696314775
Sample Mean		15.8975735						
Sum of $(y_i\text{-mean})^2$		21.41721904						
s^2		2.379691004						
var		0.191308493						
Standard Error		0.437388263						
	Upper bound	16.7548545						
95% CI:	Lower bound	15.04029251						

Excel Page: g

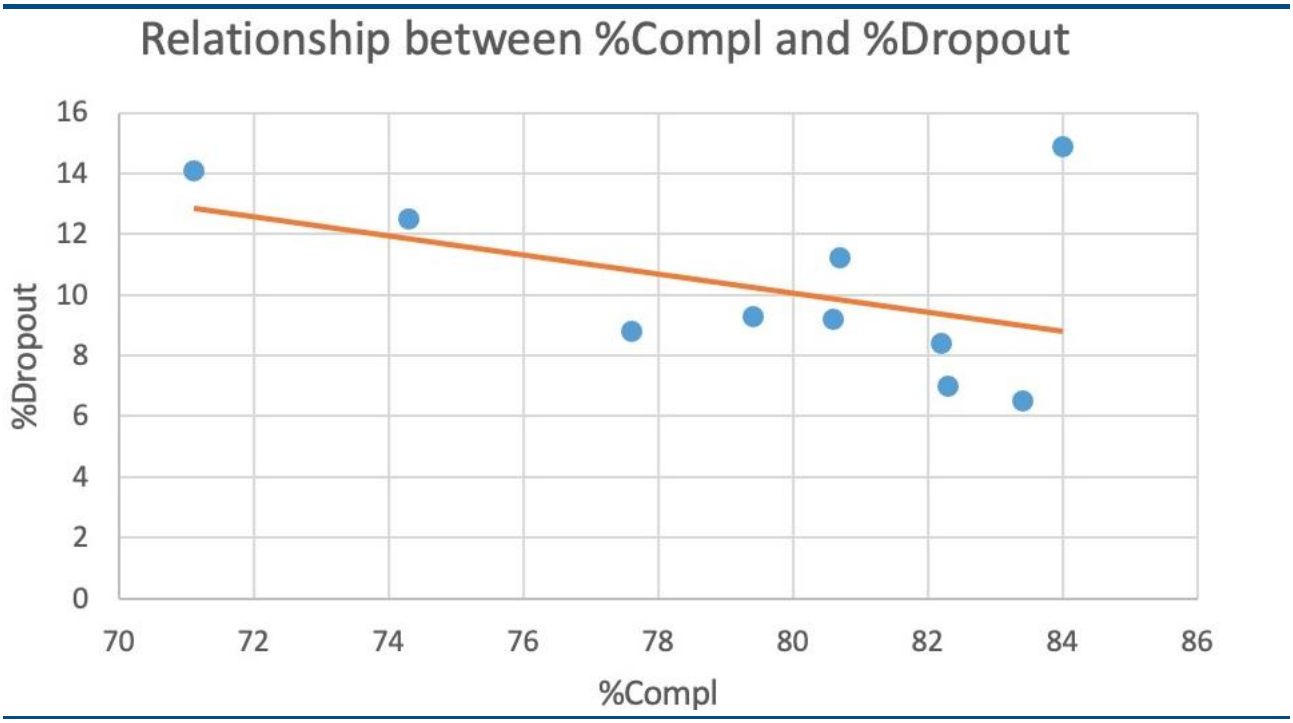
Average Dropout Rate Discussion

We initially wanted to use ratio estimation to estimate the average dropout rate by using the dropout rate (%Dropout), y, and the percentage of residents over 25 who completed high school (%Compl), x, as the auxiliary variable.

Using Excel to draw the graph of %Compl and %Dropout, we can see a negative linear relationship between the two variables and that the line does not pass through the origin. Therefore, we realised that we could not use ratio estimation.

We then instead turned to regression estimation. We believe our sample size of 10 is large enough, compared to our population size of 51, to avoid the bias of the estimate being too large to conduct a regression estimate. Nevertheless, it would be better to have a larger sample to avoid bias as much as possible.

Figure 26: %Compl & %Dropout Linear-Regression Scatter Graph



Excel Page: g

Conclusion

Simple Random Sampling

The benefits of using SRS here are the following:

1. Estimator for mean is unbiased
2. Simple to draw sample
3. Avoids bias

The downsides of using SRS here are the following:

1. Is imprecise as randomness in samples can lead to very variable outcomes
2. Difficult to gain access to larger population
    a. May be expensive and time-consuming

Stratified Sampling

The benefits of using STS here are the following:

1. Improved precision of estimator
2. Avoids bias
3. Allows us to study specific subpopulations
4. Can establish sample size within each stratum - to achieve desired precision level for subpopulation estimates
5. Can use different sampling and data collection methods for different stratums

The downsides of using STS here are the following:

1. We have limited information available to best create homogenous strata
    a. Therefore, classifications between groups may be unclear
    b. Therefore, may lead to an inefficient allocation
2. Size of sample must be chosen to respect budget constraint

We calculated the following mean estimates for each method:

SRS: 998.3

STS: 905.5

Ratio: 947.07

Regression: 953.38

Corresponding to these mean estimates, we obtained standard errors for each:

SRS: 44.6

STS: 15.3

Ratio: 15.96

Regression: 14.83

These results show us that the regression estimator is the *most precise* as it has the lowest standard error and, therefore, the narrowest confidence interval at 95% confidence. Although SRS was the most straightforward and fastest method among all those conducted - by comparing the standard errors, we can see that it is the *least precise* design.

We found that stratified sampling with proportional allocation was, *overall*, the best sampling design for our data. This is because we can easily group states into strata with minimal within-group variance. Using the regression estimator here would not work as well as the sample size, 10, is relatively small - the estimator in general for the ratio is biased, however, having a larger n would reduce the bias size.

To conclude, although stratified sampling has a slightly larger standard error than the regression estimator, overall, we believe it is the best method for this data.