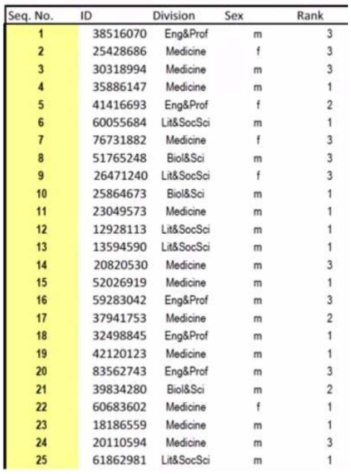
W4 Forming Groups

4.1 Auxiliary data to be more efficient

Forming Groups

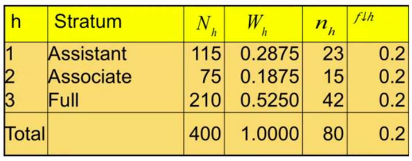
The procedure

- Before, we just used the “ID”s to random sample faculty members
- But now we are trying to use **“auxiliary” information such as division, sex and rank**



Seq. No.	ID	Division	Sex	Rank
1	38516070	Eng&Prof	m	3
2	25428686	Medicine	f	3
3	30318994	Medicine	m	3
4	35886147	Medicine	m	1
5	41416693	Eng&Prof	f	2
6	60055684	Lit&SocSci	m	1
7	76731882	Medicine	f	3
8	51765248	Biol&Sci	m	3
9	26471240	Lit&SocSci	f	3
10	25864673	Biol&Sci	m	1
11	23049573	Medicine	m	1
12	12928113	Lit&SocSci	m	1
13	13594590	Lit&SocSci	m	1
14	20820530	Medicine	m	3
15	52026919	Medicine	m	1
16	59283042	Eng&Prof	m	3
17	37941753	Medicine	m	2
18	32498845	Eng&Prof	m	1
19	42120123	Medicine	m	1
20	83562743	Eng&Prof	m	3
21	39834280	Biol&Sci	m	2
22	60683602	Medicine	f	1
23	18186559	Medicine	m	1
24	20110594	Medicine	m	3
25	61862981	Lit&SocSci	m	1

- Stratification procedure
  - population (faculty, step1)
  - frame (faculty list, step 2)
    - auxiliary variables: things known about each element in the population before the sample is drawn
    - sequence number, ID, rank, sex, division
  - divide list into groups based on the auxiliary variables
    - must be ‘discrete’ (categorical)
    - must be known for every element in the list
  - count up the number of elements in each group N_h
  - Compute the fraction of the population in each group W_h
  - Draw a sample from each group n_h (sample, step 3)
  - Keep track of sampling rates $f_h=n_h/N_h$
  - sampling fraction

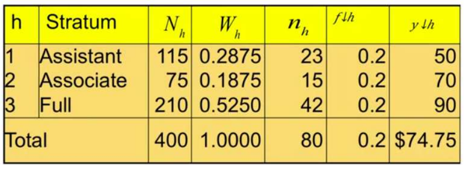


h	Stratum	N_h	W_h	n_h	f↓h
1	Assistant	115	0.2875	23	0.2
2	Associate	75	0.1875	15	0.2
3	Full	210	0.5250	42	0.2
Total		400	1.0000	80	0.2

- Stratification procedure – ESTIMATION
  - calculate estimate for each group (estimation, step 4a)
  - say means y1 = \$50, y2 = \$70, y3 = \$90
  - combine estimates across groups (step 4b)

$$\bar{y} = \sum_{h=1}^{H} W_h \bar{y}_h$$

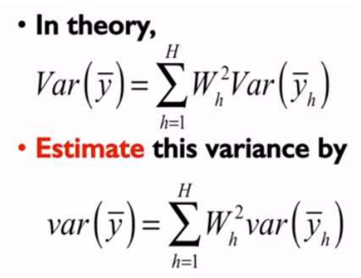
or here y_w = (0.2875) (\$50) + (0.1875)\$70 + (0.5250)(\$90) = \$74.75



h	Stratum	N_h	W_h	n_h	f↓h	y↓h
1	Assistant	115	0.2875	23	0.2	50
2	Associate	75	0.1875	15	0.2	70
3	Full	210	0.5250	42	0.2	90
Total		400	1.0000	80	0.2	\$74.75

  - But there are two more steps to go … standard error and confidence interval computation

4.2 Sampling Variances for Stratified samples



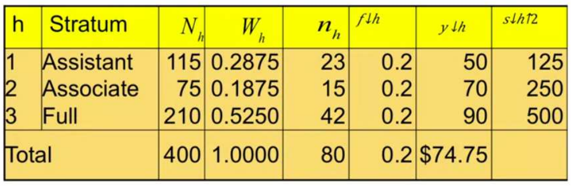
- **In theory,**

$$Var(\bar{y}) = \sum_{h=1}^{H} W_h^2 Var(\bar{y}_h)$$

- **Estimate this variance by**

$$var(\bar{y}) = \sum_{h=1}^{H} W_h^2 var(\bar{y}_h)$$

- And what is var(y_h)?
- For SRS within strata, var(y_h) = (1-f_h) * s_h^2/(n_h)
- We thus need the within stratum variances



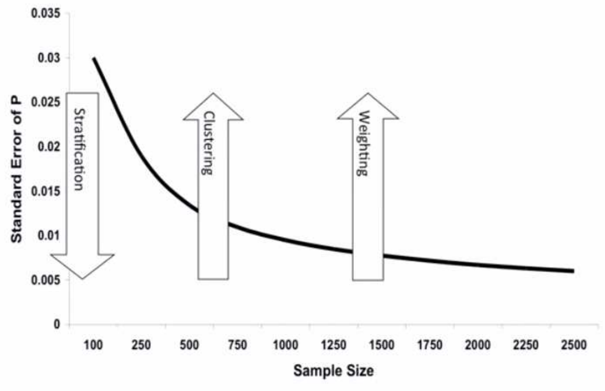
h	Stratum	N_h	W_h	n_h	f↓h	y↓h	s↓h↑2
1	Assistant	115	0.2875	23	0.2	50	125
2	Associate	75	0.1875	15	0.2	70	250
3	Full	210	0.5250	42	0.2	90	500
Total		400	1.0000	80	0.2	$74.75	

- This simplifies a little, since here 1-f_h = 0.8
- var(y) = (0.2875)^2 (0.8)(125)/23 + (0.1875)^2 (0.8)250/15 + (0.5250)^2 (0.8)500/42 = 3.453
- And se(y) = 1.858
- Thus, we have completed **step 6a & 6b** – within stratum sampling variances and combining across strata
- The last step, **step 7**, is the confidence interval
- Here let’s use the t-distribution
- We have n_h – 1 degrees of freedom for each stratum and n – H for combining across strata
- For a 95% confidence interval, then $t_{(0.975, 80-3)}$ = 1.991
- The 95% confidence interval is then:

(y – $t_{(0.975,77)}$ * se(y), y + $t_{(0.975,77)}$ * se(y)) = (74.75 – 1.991 x 1.858, 74.75 + 1.991 x 1.858) = (71.05, 78.45)

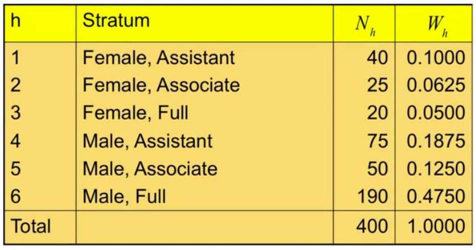
- here, as in cluster sampling, there is another issue to be addressed – how does the sampling variance from stratified sampling compare to simple random sampling?
- That is, what is the design effect, deff?
- For SRS, we need to calculate var(y) = (1-f) * s^2 / n
- From a separate calculation, s^2 = 647.8
- For our sample of size n = 80, from a population of size N = 400, or a sampling fraction f = 0.2, var(y) = (1-0.2) * 647.8 / 80 = 6.478
- Compared to the stratified sample, deff(y) = var(y) / var_{SRS}(y) = 3.453 / 6.478 = 0.5331
- As we did for cluster samples, this design effect can be thought of in several ways

  - for example, the actual size of 80 in the stratified proportionately allocated sample is the equivalent of having a simple random sample of n_{eff} = 80 / 0.5331 = 150
  - That is, the gains in precision are the equivalent of adding 70 cases to the sample
  - Alternatively, the standard errors are smaller and the confidence intervals are narrower by a factor of 1 – sqrt(0.5331) = 1 – 0.7301 = 0.2699, 27%



4.3 More on Grouping

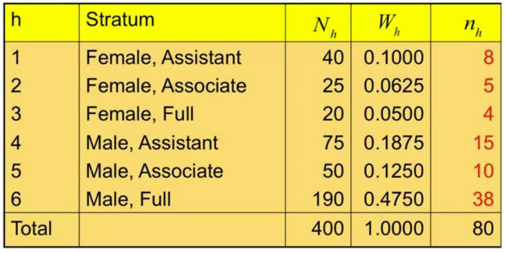
- We can also use more than one variable in the stratification
- For example, we can use, in addition to rank, sex:



h	Stratum	N_h	W_h
1	Female, Assistant	40	0.1000
2	Female, Associate	25	0.0625
3	Female, Full	20	0.0500
4	Male, Assistant	75	0.1875
5	Male, Associate	50	0.1250
6	Male, Full	190	0.4750
Total		400	1.0000

By adding extra variables (auxiliary info) to stratification, we attain greater homogeneity within each group, smaller sampling variance (under the same sample size), and thus smaller design effects and eventually even more gains in precision.

- The sample size, or allocation needs to be determined for each stratum again
  - e.g. if we again select 20% of the elements in the population the sample, ….



h	Stratum	N_h	W_h	n_h
1	Female, Assistant	40	0.1000	8
2	Female, Associate	25	0.0625	5
3	Female, Full	20	0.0500	4
4	Male, Assistant	75	0.1875	15
5	Male, Associate	50	0.1250	10
6	Male, Full	190	0.4750	38
Total		400	1.0000	80

- But in general, then, how should we form strata?
- The best advice is to make the strata internally homogeneous
- That means that the strata should differ as much as possible from each other – have big differences between the means of the strata
- Another way to say this is to find background or auxiliary variables that explain as much of the variance of the variable on interest as possible

- Availability of data
  - census
  - administrative reports
  - other surveys
- Multipurpose surveys
  - survey of households in Qatar
  - fixed assets, buildings, use of expatriate labor, expenditures, income, health, health care use, psychological well-being, social integration
- Domains of study
  - subpopulations for which separate estimates are required
  - geographic subdivisions such as provinces, districts, subdistricts
  - socio-demographic characteristics, such as age groups, occupation, income, education
- When one has multiple potential stratifying variables how does one choose which to use?
  - one consideration: how large are the stratum sizes?
  - if we are able to use only one, or a subset of variables, choose those that are going to have bigger differences in outcomes across categories
  - are there bigger differences in income between categories of rank or between categories of sex?

4.4 Allocate Sample

- The stratified sampling approach has several disadvantages:
  - **gains in precision (depending on allocations)**
  - administrative convenience
  - guaranteed representation of important domains
  - acceptability/**credibility**

- flexibility

Allocations

- How should we determine sample sizes across groups?
- Consider the basic parts of stratified sampling:

- Many allocations are possible
- For our H = 6, n = 80, for example
- $(n_1, n_2, n_3, n_4, n_5, n_6)$
- (1,1,1,1,1,75), (2,1,1,1,1,74), (2,2,1,1,1,73) ….

Proportionate

- We actually used one of these allocations in the sample of 80 from 400 faculty … **(8,5,4,15,10,38)**
- Why this allocation?
- Recall that this happens to be an allocation we got by taking the same percent or fraction of the elements in each of the six strata
- That is, we selected the sample using the same sampling rate, $n_h / N_h = f_h = n / N = f$
- But when $n_h / N_h = f_h = n / N = f$ something else happens
- The percent of the sample in each stratum is the same as the percent of the population in each stratum
- For example, for stratum 1, where in the population there are 40 of the 400, or 10% (W_h)
- But when we sampled at the same rate across strata, the number in the sample from stratum 1 is 10% of the sample, 8 out of 80, or 10%
- In other words, if we make the sample look like the population

$W_h = N_h / N = n_h / n$

- But that's the same as having the sampling fraction in all strata

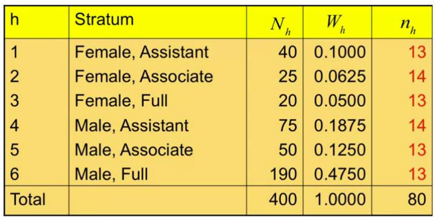
$f = n / N = n_h / N_h$

- And as select samples proportionately, we get design effects that are less than 1: deff(y) < 1

4.5 Other allocations

Equal Sample Size

- Another allocation that may make sense in other situations is to take the same or about the same number in each stratum:



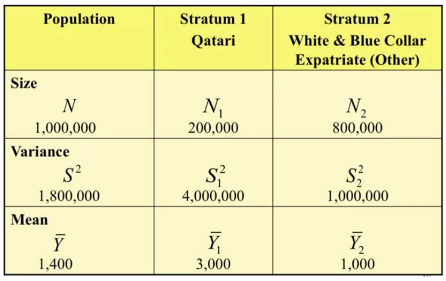
h	Stratum	N_h	W_h	n_h
1	Female, Assistant	40	0.1000	13
2	Female, Associate	25	0.0625	14
3	Female, Full	20	0.0500	13
4	Male, Assistant	75	0.1875	14
5	Male, Associate	50	0.1250	13
6	Male, Full	190	0.4750	13
Total		400	1.0000	80

- Here we end up with 27 assistant, 27 associate, and 26 full professors
  - compared to the proportionate allocation of 23, 15 and 38
- Or we have 40 females and 40 males
- These are of course quite different than the proportionate allocation (17 females and 63 males)
- Purposes of the 'equal' allocation
  - comparisons of different sized subgroups
  - better estimates for small sized groups
- Weighting the sample in necessary if we are going to combine across subgroups to get back to conclusions about the total population

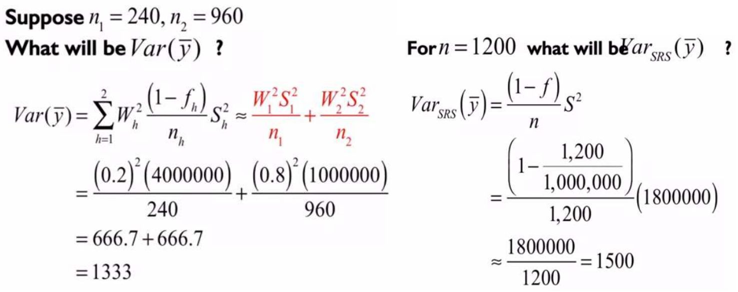
Domain estimation

Canada: Ignoring population distribution and just allocated about the same sample size to each district regardless of their pop size ➔ but we need weighting to allow pop distribution to be represented (e.g. Ontario is the biggest district in terms of pop size, so if we sample just about the same number as other districts, then Ontario is underrepresented, so we will need to x3 or x4 later (weighting) for more accurate pop distribution representation)

• On occasion, use an allocation that gives us the smallest sampling variance of all allocations – optimum allocation
• These purposes require us to look at the size of strata, and the variability within strata and sometimes even the cos within strata
• This kind of minimum variance estimation is beyond the scope of this course
• It arises in studies of a single variable, and when there is a lot of variance in the data – things like income, or expenditures, or wealth and so on
• Minimum variance allocation does not arise often in much of the social, public health, medical or other sciences
• Let's consider one other simple example to see how these allocation can affect the sampling variance:



Population	Stratum 1 Qatari	Stratum 2 White & Blue Collar Expatriate (Other)
Size N 1,000,000	N_1 200,000	N_2 800,000
Variance S^2 1,800,000	S_1^2 4,000,000	S_2^2 1,000,000
Mean $\bar{Y}$ 1,400	$\bar{Y}_1$ 3,000	$\bar{Y}_2$ 1,000



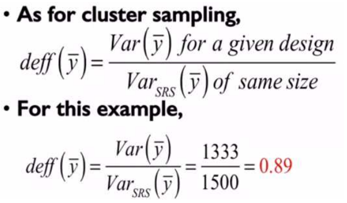
Suppose $n_1 = 240, n_2 = 960$

What will be $Var(\bar{y})$ **?**

$$Var(\bar{y}) = \sum_{h=1}^{2} W_h^2 \frac{(1-f_h)}{n_h} S_h^2 \approx \frac{W_1^2 S_1^2}{n_1} + \frac{W_2^2 S_2^2}{n_2}$$
$$= \frac{(0.2)^2(4000000)}{240} + \frac{(0.8)^2(1000000)}{960}$$
$$= 666.7 + 666.7$$
$$= 1333$$

For $n = 1200$ **what will be** $Var_{SRS}(\bar{y})$ **?**

$$Var_{SRS}(\bar{y}) = \frac{(1-f)}{n} S^2$$
$$= \frac{\left(1 - \frac{1,200}{1,000,000}\right)}{1,200}(1800000)$$
$$\approx \frac{1800000}{1200} = 1500$$

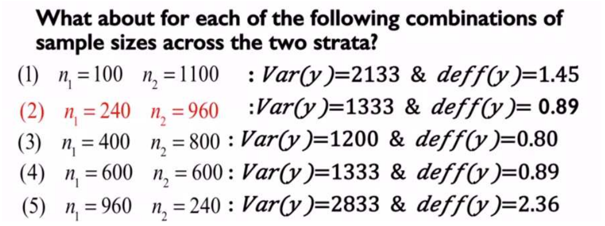


• **As for cluster sampling,**

$$deff(\bar{y}) = \frac{Var(\bar{y}) \text{ for a given design}}{Var_{SRS}(\bar{y}) \text{ of same size}}$$

• **For this example,**

$$deff(\bar{y}) = \frac{Var(\bar{y})}{Var_{SRS}(\bar{y})} = \frac{1333}{1500} = 0.89$$



What about for each of the following combinations of sample sizes across the two strata?

(1) $n_1 = 100 \quad n_2 = 1100$ **: *Var(y)*=2133 & *deff(y)*=1.45**
(2) $n_1 = 240 \quad n_2 = 960$ **:*Var(y)*=1333 & *deff(y)*= 0.89**
(3) $n_1 = 400 \quad n_2 = 800$ **: *Var(y)*=1200 & *deff(y)*=0.80**
(4) $n_1 = 600 \quad n_2 = 600$ **: *Var(y)*=1333 & *deff(y)*=0.89**
(5) $n_1 = 960 \quad n_2 = 240$ **: *Var(y)*=2833 & *deff(y)*=2.36**

4.6 Stratum or element weights?

• As mentioned, weighting the sample is necessary if we are going to combine across subgroups to get back to conclusions about the total population
• Weighting can be done in principle in two ways
• In practice, using statistical software, it is done in only one of these two ways

• One weighting method is to weight the stratum estimates by the size of the strata: y = sigma(h=1 to H)W$_h$ y$_h$
• Unfortunately, this is kind of weighting is not done is software
• Software only weights by element: y = sigma(i=1 to n)w$_h$ y$_i$
• is there a way to get the same result, the same y, using element weights as with stratum weights?
• yes, if w$_i$ = N$_h$ / n$_h$
• That is, if w$_i$ = N$_h$ / n$_h$ = 1/f$_h$
e.g. If for a stratum the sampling fraction is f$_h$ = 2/11, then the element weight for each element in the sample from the stratum is:
When the element level weight in disproportionate, stratified random sampling is the inverse of the sampling rate or sampling fraction for a given element, or the inverse of f$_h$. Here, 1/f$_h$ = 1/(2/11) = 11/2 = 5.5