**W4 Forming Groups**

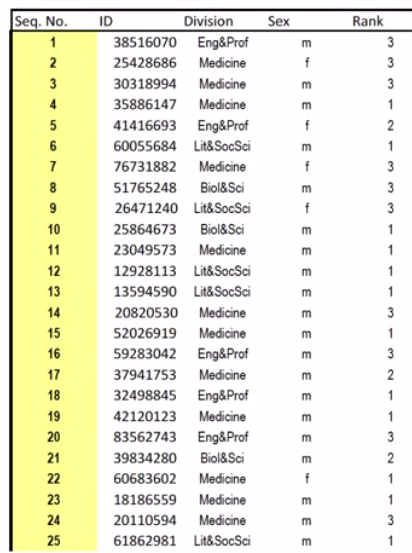
**4.1 Auxiliary data to be more efficient**

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**



• 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

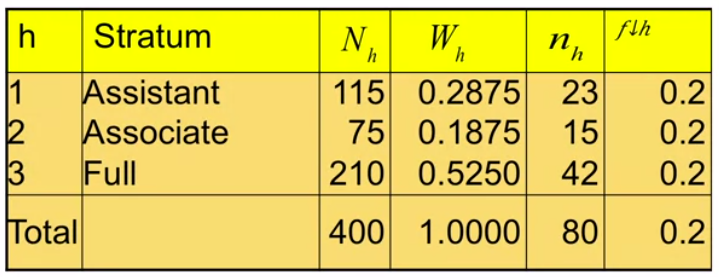
- count up the number of elements in each group Nh

- Compute the fraction of the population in each group Wh

- Draw a sample from each group nh (sample, step 3)

- Keep track of sampling rates fh=nh/Nh

- sampling fraction

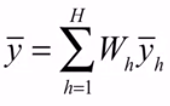


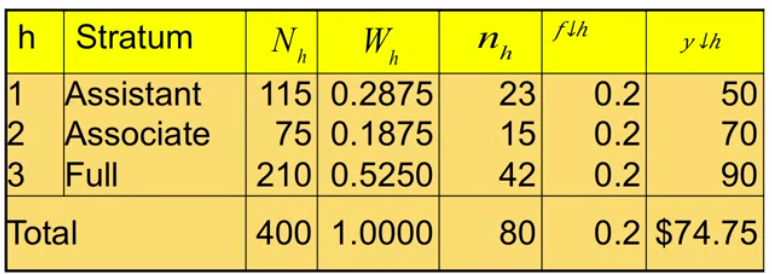
• Stratification procedure – ESTIMATION

- calculate estimate for each group (estimation, step 4a)

- say means y1 = $50, y2 = $70, y3 = $90

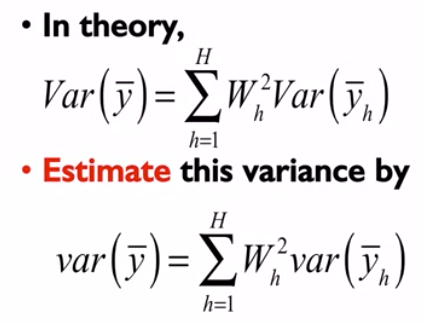
- combine estimates across groups (step 4b)

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



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

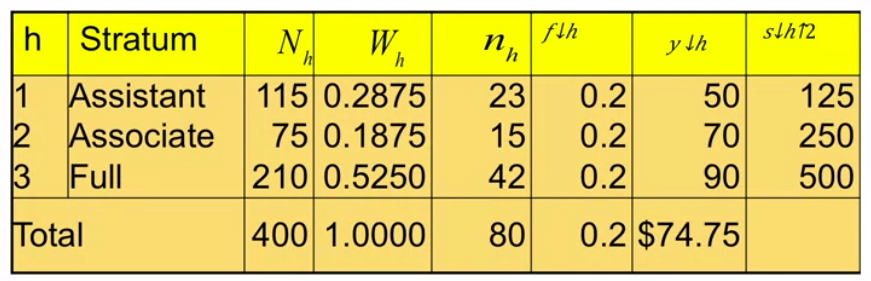
**4.2 Sampling Variances for Stratified samples**



• And what is var(yh)?

• For SRS within strata, var(yh) = (1-fh) \* sh^2/(nh)

• We thus need the within stratum variances



• This simplifies a little, since here 1-fh = 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 nh – 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

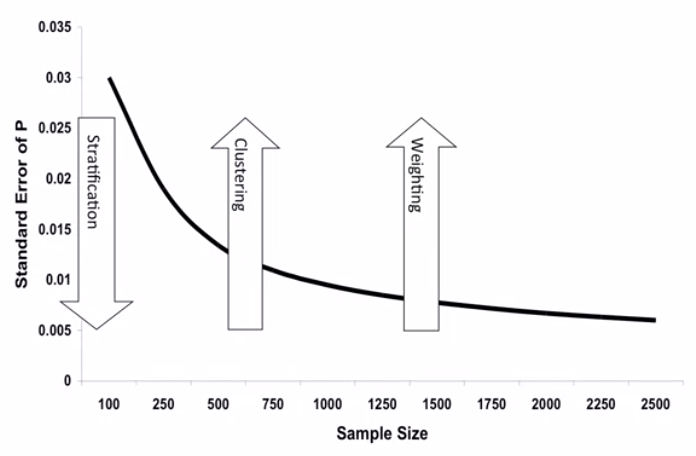
• Compared to the stratified sample, deff(y) = var(y) / varSRS(y) = 3.453 / 6.478 = 0.5331

• 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 neff = 80 / 0.5331 = 150

- That is, the gains in precision are the equivalent of adding 70 cases to the sample

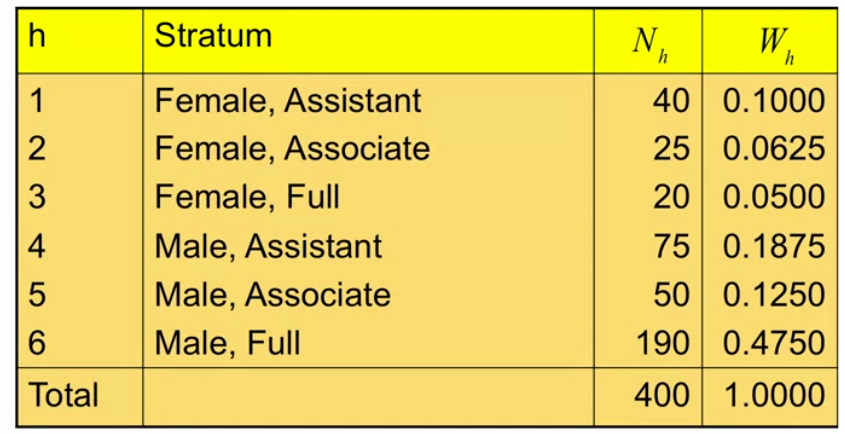
- Alternatively, the standard errors are smaller and the confidence intervals are narrower by a factor of 1 – sqrt(0.5331) = 1 – 0.7301 = 0.2699, 27%



**4.3 More on Grouping**

• We can also use more than one variable in the stratification

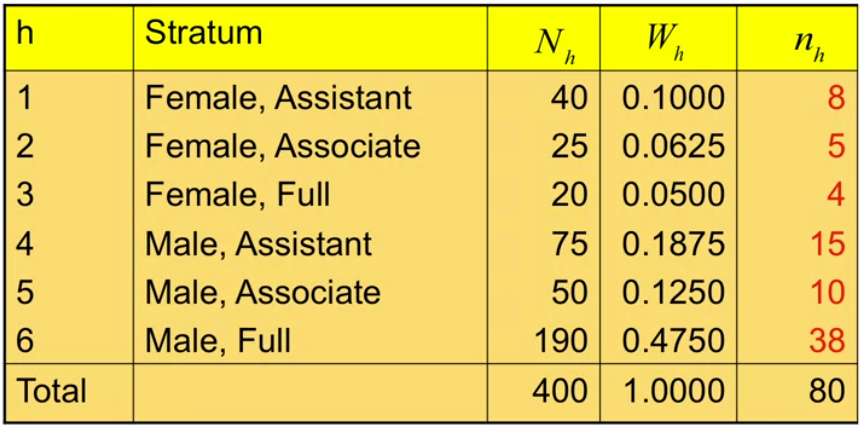
• For example, we can use, in addition to rank, sex:



**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, ….



• 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?