These notes are to be considered only a preliminary draft. For this reason they may contain typos, errors, any inaccuracies.

1 Stratified Sample: Proportional Allocation

We want to estimate:

- The proportion of students who passed an exam;
- The average number of daily hours spent by each student in front of the television.

Our population is composed of $N = 1872$ students. We select a sample of $n = 250$ students.
The course of study has a regular length of 4 years.
We assume that our sample is divided into 4 subsamples corresponding to the 4 years in which the course of study is divided.
The years of study represent the strata from which four independent samples are selected with a constant sampling fraction.
The costant sampling fraction, considering the dimensions of our population and of our sample, is

$$\frac{250}{1872}$$

.

Data and our calculus are reported in Table 1. In the table it is possible to read:

Table 1: Table 1: Proportional Allocation

Strata	(1) N_h	(2) W_h	(3) n_h	(4) $\sum y_{hi}$	(5) $\bar{y}_h$	(6) s_h^2	(7) r_h	(8) p_h
year 1	524	0.28	70	168	2.40	0.941	35	50%
year 2	487	0.26	65	169	2.60	1.088	39	60%
year 3	449	0.24	60	123	2.05	0.804	45	75%
year 4	412	0.22	55	88	1.60	0.643	44	80%
Total	1872	1.00	250	548			163	

1. Population size for each strata (absolute frequencies)
2. Population size for each strata (relative frequencies)
3. Sample size for each strata (absolute frequencies)
4. Total of hours spent in front of the television in the sample for each strata
5. Sample mean of hours spent in front of the television for each strata
6. Sample variances for the number of hours spent in fron of the television
7. Students who passed the text (in the sample)

8. Proportion of students who passed the text (in the sample)

For the sample, the average number of daily hours devoted by each student to television (request 1) can be calculated in two ways:

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

or, due to the fact that we are considering a proportional allocation, the sample mean:

$$\bar{y}_{st} = \frac{1}{n} \sum_h \sum_i y_{hi} = \frac{548}{250} = 2.192$$

It is possible to compute the rate (p_{st}) of students who passed the test in the same way (request 2):

$$p_{st} = 100 \sum_h \frac{r_h}{n} = 100 \frac{163}{250} = 65.2\%$$

Every estimates of the mean or of the total or rate must be accompanied by a measure of variability of the estimates. The estimated variance is computed using the formula:

$$v(\bar{y}_{st}) = \frac{(1-f)}{n} \sum_{h=1}^{H} W_h s_h^2 = \Big(1 - \frac{250}{1872}\Big) \frac{0.8808}{250} = 0.003053$$

and

$$se(\bar{y}_{st}) = \sqrt{0.003053} = 0.0553$$

The estimate of the variance of a proportion can be find in the same way but it is necessary to remember that:

$$s_h^2 = \frac{n_h p_h q_h}{(n_h - 1)}$$

(s_h^2 is the sample variance (divided by $(n_h - 1)$)).

As a consequence:

$$v(p_{st}) = \Big(1 - \frac{250}{1872}\Big) \frac{2160}{250} = 7.486$$

$$se(p_{st}) = \sqrt{7.486} = 0.02736$$

Now suppose that we want to estimate the *Deff* (i.e., design effect). The simple random sample is a useful term of comparison to evaluate the goodness of an alternative design. To judge an estimator relative to a particular design, it is appropriate to relate its variance to that of the analogous estimator in simple random sampling. This is the deff.

$$Deff^2 = \frac{s_w^2}{s^2}$$

$s^2 = 1.088$(belive me! you do not have the data to compute it).

As a consequence:

$$deff^2(\bar{y}_{st}) = \frac{0.8808}{1.008} = 0.87$$

This number means that (with proportional allocation stratified sampling) we have a 13% reduction in variance compared to simple random sampling. This is due to the fact that the means change significantly from strata. The deff can also be interpreted in another way: with a simple random sampling we have to extract a number of units (students) equal to

$$n^{star} = \frac{n}{deff} = \frac{250}{0.87} = 286$$

to obtain an accuracy equal to that of the estimator in the example.

For p_{st}, $s^2 = 2278$ and as a consequence $deff^2(p_st) = 0.95$

The gain is lower than what we would have expected looking at the percentage differences between the various layers.

Attention: The effect of the differences between the percentages is lower than that between the absolute values. Therefore small gains like the one observed are the norm in reality.

2 Stratified Sample: Non Proportional Allocation

Look at the previous example. We want to equally divide the sample of 250 students in 4 strata as if these strata were study domains and assumed equal variances and costs within them.

Table 2: Table 2: Equal Allocation

Strata	(1) N_h	(2) W_h	(3) n_h	(4) $\sum y_{hi}$	(5) $\bar{y}_h$	(6) s_h^2
year 1	524	0.28	63	151.2	2.40	0.941
year 2	487	0.26	63	163.8	2.60	1.088
year 3	449	0.24	62	127.1	2.05	0.804
year 4	412	0.22	62	99.2	1.60	0.643
Total	1872	1.00	250			

In Table 2 we have data with a stratified sampling with allocation equal in the four strata. We have the same values of Table 1 except that for n_h and $\sum y_{hi}$ (the sample sizes and the totals in each stratum). The estimate of the mean is:

$$\bar{y}_{st} = \sum W_h \bar{y}_h = 2.192$$

This is the same value that we obtained before but this is a weighted mean. The *simple* mean is 2.165 and it is wrong. The students of the last two years of the course, who turn out to dedicate television to an average time lower than the others, are overrepresented in the sample and consequently, by calculating a simple average, they would assume a weight greater than they should. The purpose of the weighted average estimator is to correct these sample imbalances. To estimate the standard error we use the formula:

$$v(\bar{y}_{st}) = \sum W_h^2 (1 - f_h) \frac{s_h^2}{n_h} = 0.0588$$

Bigger than 0.0553 obtained with proportional allocation.

3 Systematic sampling

Systematic sampling is a **probability sampling** method where the **elements are chosen from a target population by selecting a random starting point and selecting other members after a fixed sampling interval**. Sampling interval is calculated by dividing the entire population size by the desired sample size.

When you're sampling from a population, you want to make sure you're getting a fair representation of that population. Otherwise, your statistics will be biased or skewed and perhaps meaningless. One way to get a fair and random sample is to assign a number to every population member and then choose the nth member from that population. For example, you could choose every 10th member, or every 100th member. This method of choosing the nth member is called systematic sampling.

Systematic sampling is quick and convenient when you have a complete list of the members of your population (for example, this one of the members of Congress). **However, if there's some kind of pattern to the original list, then bias may creep in to your statistics.**

For example, if a list of people is ordered as MFMFMFMF, then choosing every 10th number will give you a sample consisting entirely of females.

Suppose you need to extract a sample of 100 students from a list of 1500 students. The reciprocal of the sampling fraction is equal to $\frac{1500}{100} = 15$.

To create the sample, we only need to select a random number between 1 and 15. This number identifies the first unit extracted and then we must proceed by selecting the other units with a 15 step arithmetic progression (I select one unit every $k = 15$) until the end of the list. For example: Firt numeber selected: 6. Than 6+15=21. Than 21+15=36 and so on. The last selected unit is: 6+99×15=1491. k can be a decimal number (a real number)

Attention: sample size can be random.

In the systematic sampling, as in the simple random sampling, **each unit has the same probability of be in the sample**. The population mean can be estimated using the arithmetic mean

$$\bar{y} = \frac{1}{n} \sum_{j=1}^{n} y_j$$

However, unlike what happens with simple random sampling, in systematic sampling not all combinations of n elements have the same probability of entering the sample.

Once the list order has been set and the first unit is selected from the first k units, only k combinations can be sampled, each with probability $\frac{1}{k}$. We can return to the situation of simple random sampling if we select the list randomly. However this is not a good practice because the aim of systematic sampling is to include in the sample units that are characterised by their position in the list.

There are some similarities between the systematic and the simple random sampling. Suppose that the sublist $k = \frac{N}{n}$ are the strata

...and about the variability? This is not measurable sampling design: No design-unbiased estimator of variance (because only one random draw)