WEIGHTING

Hector Maletta[1]
(Rev.) 12 March 2007

1. Weight matters

This paper deals with weighting, its function in statistical analysis, and its use in SPSS.[2] A typical problem involving weighting can be seen in the following example.

A survey has been carried out on the population of a certain region of the United States, to estimate voting preferences in the region for an upcoming national election. A sample of 1000 voters, including 500 African Americans (AA) and 500 Non African Americans (NAA) has been drawn at random, and the result was that 60% of the sample announced they would vote Democratic. The pollster in charge hastily made the prediction that Democrats would win by a near landslide margin. As it turned out, this was a wrong prediction: 55% of voters chose the Republican candidate, and just 45% cast the Democratic ballot.

What went wrong with the pollster's prediction is that it failed to take into account the fact that blacks were over-represented in the sample, and at the same time they were very strongly Democratic (about 80%) while whites and other non-blacks were under-sampled and they were also more evenly divided between the two parties. The prediction of 60% Democratic was based, in fact, on 40% of the NAA sample and 80% of the AA intending to vote for the Democratic Party. As the same number of people from each ethnic group was interviewed, the overall simple average was 60% Democratic. But in fact, blacks were not even close to represent 50% of the population. Their share in the region was about 18% of the population. So in fact giving each portion of the sample its proper demographic weight (82% non-blacks, 18% blacks) gives a very different prediction:

$$(0.82 \times 40) + (0.18 \times 80) = 32.8+14.4=47.2\% \text{ Democratic}$$

This is not far from the actual result (45% Democratic). If the pollster knew about weighting a better prediction could have been produced.

The error pertained only to the overall vote prediction. The sample results by race, reporting the strong Democratic preferences of AA (80%) and the Republican inclinations of 60% of non-blacks in the population of that region, were not tainted by any problem, and they were not expected to be, because they were based on a random sample of individuals belonging to each group. Only the overall percentage, resulting from the combination or aggregation of both groups, suffered from the gross over-sampling of African Americans in the sample. In fact, weighting appears as an issue whenever several entities (in this case people) are **aggregated** into a general figure like an overall percentage. Giving some parts of the sample more weight than they are entitled to may in that situation entail errors of estimation.

---

[1] Universidad del Salvador, Buenos Aires, Argentina. Email: hmaletta@fibertel.com.ar. Alternative email: hector.maletta@mail.salvador.edu.ar.

[2] I am grateful to King Douglas and Raynald Levesque for comments on a draft of this paper; to Raynald also for posting the paper on www.spsstools.net; and to Anton Balabanov who, in the process of translating this paper into Russian, discovered several inconsistencies and imprecise sentences that needed editing. I am also grateful to several members of the SPSSX-L forum, for asking questions about the use of weights in SPSS which prompted me to think about this matter,

2. Weighting cases and weighting variables

A data set is usually composed of **cases** for which several **variables** are predicated. Thus the data set is a rectangular array of *n* cases by *k* variables. Statistical analyses usually proceed by aggregating cases and/or variables in some meaningful way. For instance, one variable can be aggregated across cases to yield a total or a mean or average.[3] Also, several variables can be aggregated within each case to yield an overall score or scale, for instance an average grade for each student from the grades obtained in different school courses.

One common situation is having groups of **cases** of different size. For instance, when obtaining the average income of a population based on the average income of several population subgroups or geographical areas, a proper computation requires that each subgroup or area is given its proper weight in the total: suppose you have the average (per capita) income of Canada, the US and Mexico, and want to compute the average North American per capita income. The simple average of the three countries' per capita GDP would not do, since the US population is far larger than the population of the other two countries. Multiplying each national per capita income by the country's population, adding up the results, and dividing by the total population of North America, would give the appropriate answer. In an equivalent formulation, multiplying each national per capita income by the country's proportional share in North American population, and adding up the results, would yield the same result.

The same happens with weighting **variables** instead of weighting cases. Suppose you want to construct an overall scale based on several variables, but some of the variables are considered more important than others. For instance, some grades may come from short one-week courses or workshops while others represent full-semester courses involving much more study work (and more credits). Multiplying each grade by some measure of the course's length and importance (such as credits) may give a more adequate grade average than simply averaging all grades.

This paper is concerned mainly with weighting **cases,** not with weighting variables. In particular, it is especially concerned with the weighting of cases made necessary by **sampling** (as opposed to other factors such as the different sizes of the three North American countries alluded before). However, the general principles apply to any kind of case weighting.

Weighting variables, on the other hand, is a completely different matter. When adding several variables, for example in scale construction, some of them may be given more weight based on a variety of considerations. The simplest aggregation of variables with equal weight, such as a simple average of various grades in school to obtain a general grade average, may be regarded as unfair in some regards, because (for instance) longer or more difficult courses should be given more weight. These weights can be given exogenously (e.g. the courses' credit ratings), or may arise from the data themselves as in the case of various procedures such as factor analysis, which reduce a number of observed variables to a smaller number of underlying "factors", and each observed variable has a different weight towards each underlying factor. As said before, this matter of weighting the variables is outside the scope of this paper and will be ignored hereafter.

3. Sampling

Samples, as is well known, can be random or not random. Most statistical analyses are based on the assumption that a **random** sample is used, and most of this paper is concerned with weighting of random samples. Weighting, of course, cannot do the trick of converting a non-random sample into a random one (though it can somewhat improve the estimates derived from it). Random sampling from a given population usually involves one or more of the following devices:

- **Simple random sampling**: Cases are selected from a list containing all cases that belong to the population of interest. Example: the population is defined as all current members of the American Medical Association, which are fully listed by the Association, and a

[3] In this paper, "mean" and "average" are used interchangeably, and refer to the arithmetic mean.

sample is drawn by selecting one of every 100 members from the list, using some standard random sampling procedure. Another example: randomly select 20 US senators from the complete list of current members of the US Senate. In some instances the list does not actually exist but some other equivalent device is used. For instance, select one out of every 10 homes in a street, just by walking down the street and counting homes. You may also have to fix the starting point at random, e.g. using a table of random numbers to choose one of the first ten homes from the North (or South) end of the street.

- **Stratified sampling**. The population is exhaustively divided into two or more strata, and then a simple random sample is drawn from each stratum, not leaving any stratum aside. This makes more likely that cases are evenly selected from all the strata, thus reducing the possibility that the sample is disproportionately concentrated on one part of the population. Also, if the stratification is somehow related to the variables of interest, then the variance of those variables **within each stratum** will be lower than the overall variance of the population at large. This **reduces overall sampling error**. For instance, suppose the study is on the entire population of New York City and the variable of interest is household income. If each borough of the city is regarded as a stratum, it is likely that average income differs across boroughs, and the variance of household incomes within each borough (relative to the borough mean) is smaller than it is for the entire city (relative to the city mean) because affluent or poor people tend to concentrate on certain zones (e.g. incomes may be more homogeneous within Queens or within Manhattan than they are for the entire city).
- **Cluster sampling**. The population, or each stratum, is subdivided into a number of clusters or subdivisions of any kind; then some of the clusters are selected, and a random sample of cases is chosen **within each selected cluster**. For instance, in each borough of NYC a number of ZIP code areas are randomly selected, and then a random sample of households is in turn selected within each chosen ZIP code area. Unlike stratification, clustering involves two stages of selection: first some clusters are selected, and only then some cases are selected within the selected clusters. Households in other ZIP code areas have no chance of being interviewed once the sample of ZIP codes has been already selected. Clustering, unlike stratification, tends to **increase overall sampling error.** This danger is highest when there are relatively few large clusters, quite different to each other, and very few are actually selected. Suppose for a sample of the US the country is split into three Regions, i.e. three strata, according to Standard Time Zones (East, Center, West) and then two States (clusters) are chosen within each region. For the East, results will be very different if New York and Maine are chosen than in case Georgia and Alabama are selected. If there are numerous small clusters, as in the case of ZIP code areas, and the sample of ZIP areas is relatively large itself, comprising a fairly generous number of clusters, the possibility of distortion is lower.

Stratification reduces sampling error for the same overall sample size, or minimizes sample size for a given level of expected error, and this makes stratification a must for realistic sampling design of heterogeneous populations. Clustering, instead, actually increases sampling error, and this would make it unadvisable, were it not for its big cost advantages. Having sample cases concentrated in small geographical areas saves a lot on field work costs. Also, a complete list of cases within a cluster, such as a listing of households in one ZIP code area, is more easily compiled than a complete list of cases for the whole population.

The three devices are usually combined in many real-life situations. In particular, large surveys in the Social Sciences usually combine the three. For instance, the country may be divided into regional strata, then clusters (communities, cities, neighborhoods, etc.) are chosen within each region, perhaps some smaller clusters (e.g. city blocks or ZIP areas) are selected within each selected larger cluster, and finally some persons or households are selected within each selected small cluster.

Sampling with and without replacement. Ordinary sampling in large populations is usually done without replacement (WOR). Once a unit has been selected, it cannot be selected again, and the rest of the sample is chosen among the remaining population. The opposite method is sometimes used, namely sampling with replacement (WR). With this latter approach, once a unit is selected it re-enters the population and can thus be selected again. All units, therefore, are selected from the same population, not excluding units already selected. In WR a given unit may thus be selected several times, and thus end up being represented in the sample by several copies of itself. This can be seen as a kind of weighting: a unit repeated three times in the sample would have a weight of 3, while those not repeated would have a weight of 1. This concept will be revisited below, but in general we will proceed in this paper as if the given sample is obtained without replacement.

Proportionate and disproportionate sampling. A sample is a fraction of the corresponding population, and this fraction is called a **sampling ratio**, equivalent to sample size divided population size (n/N). It represents the probability that a unit is selected from a given population. If one selects one out of every ten households, the sampling ratio is 1/10=0.10, and each household has a 10% probability of being selected. More generally, if n_k is the size of the sample from population group k and N_k is the size of population group k, the sampling ratio is $f_k=n_k/N_k$.

In stratified and clustered samples, sampling ratios may differ from one part of the sample to another, for different reasons. Then two kinds of sample can be identified, namely those with a uniform or a variable sampling ratio. In samples with variable sampling ratio, i.e. arising from disproportionate sampling, the chances of one individual to be included in the sample may differ from the chances of other individual in the population.

For instance, if a sample is taken from the lists of members of several scientific societies such as the American Economic Association and the American Psychological Association, it is possible that, say, 1:10 economists are chosen but only 1:200 psychologists, implying that sampling ratios (and the probability of a professional to be included in the sample) are 0.100 for members of AEA and 0.005 for members of APA.

Planned and realized sampling ratio. Sometimes the actual number of cases in the sample differs from the planned number, due to several possible reasons (interview refused by some contacted respondents, people not contacted because not found at home after repeated visits, questionnaires filled but rejected due to poor quality, etc.). Whereas the planned sampling ratio is embodied in careful selection procedures in the field, such as using random numbers or randomly choosing from a list, refusals and rejections may obey non-random reasons. For instance, perhaps upper income households are more likely to refuse responding, or lower-income households may be more likely to produce poor quality questionnaires (especially if the questionnaires are self-administered).

In some surveys, some failures are replaced, on the assumption that the replacement is similar to the case being replaced. This is more likely to be true, for instance, in cases apparently due to random reasons such as not finding anybody at home at various visits, but is likely to involve a bias if a household not willing to be interviewed is replaced by a willing one, because the latter may have different characteristics from the former.

The existence of replacements as well as unplanned changes in sample size may involve assumptions and decisions, which are to some degree arbitrary, affecting which sample size (planned or actual) is to be used to compute sampling ratios (and thus to compute weights, as will be explained below). This will vary from one survey to the next, and for the purpose of this paper, the question is ignored. Thus n_k is simply regarded as **the** sample size for population group k, without further elaboration. It is usually the **realized** sample, but in some cases it may include some correction for refusals and replacements.

4. Sample weighting

A sample is, by definition, smaller than the corresponding population. Besides, it could be a disproportionate representation of the population if some groups are over- or under-represented. Thus it differs from the population in scale, in proportions, or both. Applying weights to samples seeks the general goal of making the sample more like the population at large. Its two main possible specific purposes (not exclusive of each other) are:

a. **Correcting for scale**
b. **Correcting for proportions**

Scale weighting. For some purposes, estimating population totals based on a sample may be important, and practitioners may want that tables reflect the actual size of the population rather than the size of the sample. For instance, an announcement that the finale of the World Cup was viewed on TV by 2.5 billion people worldwide may be more interesting than learning that it was viewed by 15,365 persons interviewed in various countries (which may be actually where the other estimate came from).

Expanding the scale of the results, from sample to population scale, is usually done by means of weights of the general form $w=N/n$, whereby sample results are multiplied by the reciprocal of the sampling ratio n/N.

In a simple random sample there is only one sampling ratio throughout, so scale weighting is achieved by multiplying everything by N/n. This overall or uniform **scale factor** will be designated by v and defined by:

$$v = \frac{N}{n} \qquad \textbf{Eq. 1}$$

Suppose, for instance, you have a simple random sample of doctors who are members of the American Medical Association. You simply selected n doctors out of a list of N doctors. You want to know how many doctors smoke. Instead of reporting the actual number of smokers in your sample, you multiply that number by $v=N/n$, to obtain an estimate of the total number of smoking doctors in the US. There is only one scale factor for all cases, because the sample is a simple random sample.

Proportional weighting. In complex sampling designs, e.g. stratified or clustered sampling, sampling ratios vary. As a result, proportions in the overall sample may not coincide with proportions in the population. Correcting for this is done with proportional weights, which are specific to every subdivision of the sample for which the sampling ratio is homogeneous (e.g. each stratum), and can be designated as π_k with the general form π_k*=% of stratum in population / % of stratum in the sample*:

$$\pi_k = \frac{N_k / N}{n_k / n} \qquad \textbf{Eq. 2}$$

In this kind of weights, cases in stratum k receive a proportional weight $\pi = 1$ if that group is represented in the sample in the same proportion it appears in the total population. When π_k **<1** the group has been over-sampled (its proportion in the sample, n_k/n, is larger than its proportion in the population, N_k/N), and in the opposite case π_k **>1** means the group was under-sampled (its proportion in the sample is smaller than its proportion in the population). Proportional weights inflate under-sampled cases, and deflate the over-sampled ones.

Mixed or integrated weighting. The purely proportional weights of Eq.2 do not expand results to population size. Only the proportions are restored. On the other hand, we have seen that simply inflating sample results to population scale by a uniform factor N/n does not correct the

proportions in the sample. A mixed weight accomplishing the two functions can be obtained by multiplying the two:

$$w_k = v\pi_k \qquad \textbf{Eq. 3}$$

If the scale factor of Eq. 1 is multiplied by the proportional factor of Eq 2, the result is reduced to the simple expression of a reciprocal sampling ratio for the stratum in question:

$$w_k = \frac{N_k / N}{n_k / n} \times \frac{N}{n} = \frac{N_k}{n_k} \qquad \textbf{Eq. 4}$$

These weights (Eq. 4) accomplish the two functions at once: they adjust sample figures to population scale, **and** they also correct imbalances in sampling ratios from one stratum to the next. In view of this they may be called combined, mixed or integrated weights.

Weighting with sub-strata and with clusters. When the sampling design involves a hierarchical scheme of major strata divided into sub-strata, or some form of clustering is used, weights are in principle computed the same way. The final sampling ratio is the product of successive levels of sampling ratios, and the same for their reciprocals. Suppose for instance, that a first-order stratification divides the population into *k* strata (e.g. US States), then each stratum is subdivided into *h* sub-strata (e.g. counties), then a certain proportion p_{kh} of clusters (e.g. ZIP areas) is selected in each sub-strata, and finally a proportion *q* of final sampling units (e.g. households) is selected within each selected cluster (i.e. within each selected ZIP area). Assume for the sake of simplicity that no information exists as to the total number of households per ZIP zone, so it must be inferred on the basis of households counted in the selected ZIP areas. Each household in the m^{th} ZIP zone of the k^{th} county belonging to the i^{th} State will have a weight defined by:

$$w_{mki} = \frac{H_{mki}}{h_{mki}} \frac{Z_{ki}}{z_{ki}} \qquad \textbf{Eq. 5}$$

In plain words, the final weight equals the reciprocal of the probability of the household being selected within its ZIP zone, multiplied by the reciprocal of the probability of that ZIP zone being selected among all the ZIP zones in the county. Since all counties in each state are included in the sample, and also all states without exception, the probability of a county or state being selected is uniformly 1, and is therefore omitted from the equation in this particular example. In more general terms, the probability of a final unit being selected is the product of selection probabilities at all previous stages of the selection process, from top to bottom. If some primary units are selected with probability p_i, and within the selected units there is a selection of secondary units with probability $p_{(i)j}$, i.e. probability p_j within the i^{th} primary unit, and so on until the elementary units are selected with probability $p_{(ijkm\ldots)z}$ within the selected penultimate units, then the overall probability of a final unit to be selected is:

$$p_{ijkm\ldots z} = p_i p_{(i)j} p_{(ij)k} p_{(ijk)m} \cdots p_{(ijkm\ldots)z} \qquad \textbf{Eq. 6}$$

The overall mixed weight (performing both scale and proportion weighting) is simply the reciprocal of this probability:

$$w_{ijkm\ldots z} = \frac{1}{p_{ijkm\ldots z}} = \frac{1}{p_i p_{(i)j} p_{(ij)k} p_{(ijk)m} \cdots p_{(ijkm\ldots)z}} = \frac{1}{p_i} \frac{1}{p_{(i)j}} \frac{1}{p_{(ij)k}} \frac{1}{p_{(ijk)m}} \cdots \frac{1}{p_{(ijkm\ldots)z}} \qquad \textbf{Eq. 7}$$

Multiplying **these weights** $w_{ijkm\ldots z}$ by *n/N* (i.e. by the simple overall sampling ratio) will convert these mixed weights into proportional weights with no scale effect, previously designated as π_k (in this case they would be designated as $\pi_{ijkm\ldots z}$ because of the many stages involved in sample design).

5. Sample and population: The law of large numbers

Sample results are supposedly representative of the population from which the sample comes. This is formalized in the mathematical theory of sampling error, whose cornerstone is the so-

called Law of Large Numbers. Simply stated, that Law says that if repeated samples of the same size are drawn at random from the same population, and a certain measure (such as a mean) is taken in each sample, the estimate from each sample may differ from that of other samples, but the distribution of these results will have the shape of a normal curve, whose average is the true population mean

Suppose, for instance, that several samples of n male adults are drawn at random from a male population, and each man in each sample has his height measured in centimeters. All samples have the same size (say, $n = 100$ males), and for each sample the average height is computed. For the k^{th} sample this average height would be H_k. The sample mean would differ from one sample to the next. If many sample means are taken, they would tend to be normally distributed about a "mean of means" (the mean of all sample means), and this mean of means would tend to equal the mean height of all adult males in the population. We say "tend to be" and "tend to equal", because this would be a function of (a) the number of samples taken and (b) the size of the samples. The larger the samples, and the more numerous they are, more and more the distribution of their means would approximate a bell-shaped distribution and more and more their overall mean of means would approximate the true population mean.

Variable distribution and sampling distribution. A variable in a population (or in a sample) may have any frequency distribution. Individual heights in a given sample of men (or in the entire population) may include many different values, with some people shorter and some taller than average. In the case of heights, the distribution of individual heights within a sample or within the entire population is likely to be itself a normal distribution, because most biological variables have that kind of distribution, but other variables (e.g. income) may have a more skewed or otherwise non-normal distribution. Some variables may have a uniform distribution, or a U-shaped distribution, where cases are concentrated at the extremes with fewer in the middle, or they may have a decreasing distribution with many cases in the low range and fewer at higher values of the variable, or whatever other distribution one may possibly imagine. In other words, the distribution of the variable **among individuals**, at sample and population level, is indeterminate and has little to do with the present discussion.

Whatever the distribution of a variable within a sample, it will always have a mean. If many samples are drawn, we would have many **sample means.** The **sampling distribution** of a variable is not the distribution of individual values about the sample mean but the distribution of **sample means** about the population "**mean of means**". The first is "within the sample" and the second is "across samples".

Statistical inference requires that samples are random, because it is a proven mathematical theorem (and also an empirical finding) that the sampling distribution of sample means coming from many random samples of the same population tends to be a normal distribution, whose mean tends to coincide with the population mean.[4] Nothing of the sort is implied or required from the within-sample distribution of individuals around the sample mean. It is very important to stress this point, because many people confuse the normality of the sampling distribution with a requirement that the variables themselves have a normal distribution about their sample means.

The statistical significance of sample results. When only one sample is available, which is the usual situation in real life, the analysts knows that a sample mean is not likely to be very far away from the population value, because the means of the various samples have a normal distribution, and furthermore, the average of this normal distribution (the mean of means) is close to the mean of all individuals in the population. The actual sample could be very far from the population mean, but the probability of this happening is relatively small if the sample is large enough and has been obtained at random. For instance, there is about 95% probability that a sample mean

[4] Sample mean here refers to the average of continuous or interval-level variables, such as age or income, or to the average of binary variables (e.g. the percentage of females), which is equivalent to the average of the values 1 for women and 0 for males. So means or percentages are treated as equivalent.

height is within two standard deviations from the population mean height. One does not know whether this particular sample is not one of the unlucky samples in the other 5%, far below or far above the true value (perhaps our sample was almost entirely made of dwarfs or NBA players, even being a perfectly legitimate random sample).

However, analysts are willing to take the risk. They realistically do not claim or aspire to be totally sure. They content themselves with a certain degree of probability. They would be happy to say something like "From a sample of *n* cases, mean height is estimated to be 175 cm, and there is a 95% probability that the true mean is not beyond 3 cm either side of my estimate". The 95% confidence interval of ±3 cm is the range of heights where 95% of all possible samples are expected to fall. There is always the possibility of being stuck with a sample belonging to the other 5%, whose mean is situated outside the confidence interval around the population mean, but the probability is low enough to be accepted as one of the perils of working with samples.[5]

As said before, a confidence interval is delimited by a certain number of **standard deviations of the sampling distribution**. For a 95% level of confidence, the interval is delimited by 1.96 standard deviations each side of the sample mean. Of course stating that the sample mean is within 1.96 Std Dev from the population mean is equivalent to saying that the population mean is within 1.96 Std Dev from the sample mean. The standard deviations in question are **not** the standard deviations of the variable within the sample (variability of height among individuals) but the standard deviations **of the sampling distribution** (variability of average height across different samples from the same population).

The standard deviation of the sampling distribution of a variable *x* is also called the **standard error** of a sample estimate of *x*. It is defined by the following formula:

$$SE_x = \frac{\sigma_x}{\sqrt{n}} \qquad \textbf{Eq. 8}$$

In this formula, σ_x is the standard deviation of the variable in the entire population, and *n* is the size of the sample. The 95% confidence interval of a sample estimate is ±1.96 SE at each side of the sample estimate. Notice that SE is always lower (usually much lower) than the standard deviation of the variable in the population, since to obtain SE the standard deviation of the variable is divided by the square root of sample size. Thus for a sample of 100 the SE would be 10 times smaller than the standard deviation, because the square root of 100 is 10.

The quantity σ_x in the numerator of Eq. 8, i.e. the **population** standard deviation of variable X**,** is of course unknown, as is the population mean. Only the sample mean and the sample standard deviation are known. It is usually assumed, however, that a reasonable estimate of the population standard deviation σ_x is given by s_x, the standard deviation of X **in the sample**, i.e. in the one and only sample which has been actually observed. So in fact the standard error is estimated as:

$$SE_x = \frac{s_x}{\sqrt{n}} \qquad \textbf{Eq. 9}$$

Of course, the observed sample is only one of the many possible samples. Just as sample means vary across different samples around the true population mean, so sample standard deviations s_x vary across samples around the population standard deviation σ_x. There is no guarantee that the observed sample standard deviation s_x coincides with (or is even close to) the true population standard deviation σ_x, but there is usually no other way to go. One source of consolation is that the sampling variability of sample standard deviations is usually lower than the sampling variability of sample means, so the observed sample standard deviation could be a reasonably

[5] Changing the level of probability would result in a wider or narrower interval. Other usual confidence intervals are the 90% confidence interval (which is narrower) and the 99% confidence interval (which is wider). In general, the higher the confidence level you wish for your estimation, the wider would the confidence interval have to be.

good estimate of the true population variance (provided the sample is large enough and obtained by random sampling).

Now how is s_x computed? In a simple random sample, the unbiased estimate of the population standard deviation is just the sample standard deviation, only with *n-1* in the denominator instead of the usual *n*:

$$s_x = \sqrt{\left(\frac{\sum(x_i - \bar{x})^2}{n}\right)\left(\frac{n}{n-1}\right)} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} \qquad \textbf{Eq. 10}$$

In a **stratified** sample, the best estimate of the population variance is not given by this formula, which would correspond to the standard deviation of one stratum. What is needed is the **weighted average of the standard deviations of the various strata**, computed **relative to each stratum mean**, instead of the overall mean. Assuming each stratum represents a proportion f_k of the population, and a proportion g_k of the sample, the estimate of the population standard deviation is the weighted average of the standard deviation observed in the various strata, weighted by their share in the population (f_k).

$$\hat{\sigma}_x = \sum_k f_k \sqrt{\frac{\sum_i (x_{ik} - \bar{x}_k)^2}{n_k - 1}} \qquad \textbf{Eq. 11}$$

This estimate of the population standard deviation, which is itself computed on the basis of weighting the sample results, is the figure to use in Eq. 9 to estimate the standard error of a stratified sample mean. The sample mean itself is also computed as a weighted average of the means observed in the various strata. Therefore, for a stratified sample the estimation of sampling error and confidence intervals depends on the population share of all the strata and also on sample size (*n*). If the means of the various strata are different, the SE for a stratified sample of size *n* would be smaller than for a simple random sample of the same size.

When **clustering** is involved, the difficulty arises that only some clusters are chosen in each stratum, and this may disfigure the estimation of variance within each stratum. By choosing one cluster and disregarding others, one is betting that the selected clusters are representatives of the total, which may not be the case (imagine a random sample of households in some selected US States; the selected States could be mostly Western, or mostly Northeast Coast, or mostly Southern, with probably very different results). It is easily conceivable that variation within clusters is lower than variation across the entire stratum, since cluster members are geographically close to each other and may share many important characteristics, and be similar in others. For instance, income variation within a city block is likely to be lower than income variation across a whole city or borough. By observing several clusters, one may believe variability is low when in fact it is larger. Moreover, when only few large clusters exist and some of them are actually chosen, there is the plausible danger that the selected clusters are not representative of the average mean or variance in the population. Thus the average score in a scale of political orientation in a sample of predominantly Southern States may not coincide with the average of the country, and the **variability** of this variable in that sample may be much narrower than the variability existing in the whole country.

In other words, using cluster sampling increases sampling error, but it increases it by an unknown factor (unless we have some external source of information on the variance prevailing in the population or stratum from which the clusters were selected). Analysts often make the convenient assumption that no such distortion exists, i.e. that variation among members of selected clusters equals the variation that would be observed in the whole population or stratum, but this can only be reasonable if the total number of clusters is relatively large, and the number of selected clusters is also relatively large. Thus randomly choosing three counties in a State with nine

counties is more dangerous than randomly choosing 300 ZIP areas in a territory with 900 ZIP areas. In the first case there might be enormous heterogeneity between counties, and (for instance) choosing the county where the main city of the State is located would lead to quite different results than not choosing that particular county. With 300 out of 900 ZIP areas the risk is more widely spread and on the whole much less worrying.

In short, unless some estimation is available of the possible increase in sampling error due to clustering, applying the above formulas to samples that include clustering in some of its stages may lead to underestimation of sampling error. The resulting estimate of population standard deviation, if based on Eq.10-11, would likely be on the low side, and consequently the estimate of the standard error of estimates would also be lower than it should be. When no clustering is involved, instead, the formulas for random stratified sampling (Eq.10-11) would lead to unbiased estimates of the standard error.

Weighted and unweighted data for significance tests. For samples designed other than by simple random sampling designs, computing significance tests on the basis of unweighted data would create bias in significance tests, as it would in any other statistical estimation (means, standard deviations, percentages or whatever).

Sampling ratios, and thus weights, must be taken into account to compute valid significance tests for sample data. However, what is crucial in this regard is the **proportional** aspect of weighting (giving each sample observation its proper relative weight), and not the **scale** effect (expanding sample results to population size).

6. Weighting in SPSS

By default SPSS does not apply weights to cases in a data set. If the data set comprises *n* cases, totals in a table will show a total of *n.* This can be seen as if every case has a weight of 1. However, the user can use the individual values of any variable as case weights. For instance, the command WEIGHT BY X would use the values of variable X as case weights.

Weights can be uniform (all cases have the same weight) or variable (cases have possibly different weights). If all cases are assigned a uniform weight, for instance 100, then all the data would be multiplied by 100, and table totals for example would be 100 times the sample size. Using such a uniform weight would increase the **scale** of the totals without altering the **proportions** between them. If the sample comprises, for instance, 50 males and 40 females, the weighted totals would be 5000 males and 4000 females. The weighted total is, in fact, the sum of all weights (in this case, 50 x 100 = 5000 and 40 x 100 = 4000).

The user can also assign specific weights to each case, which is the most usual situation. Suppose the sample was stratified, and the three strata were randomly sampled with sampling ratios of 0.05, 0.005 and 0.0005. The weights (responding to the formula N_k/n_k) would be the reciprocal of these fractions, i.e. respectively 20, 200 and 2000. The user can assign these values to all cases by creating a variable (call it X) which equals 20 for every case in stratum one; 200 for every case in stratum two, and 2000 for cases in stratum three. Then the command WEIGHT BY could be used to establish the X variable as the weighting variable. From the moment this WEIGHT BY command is issued, SPSS would multiply each case by its weight whenever asked to do some statistical analysis. Suppose a data file with seven cases, pertaining to three strata. The file is initially not weighted.

The average age of the subjects would be computed by multiplying each age by the corresponding weight (which is 1 by default), adding up these products, and dividing by the **sum of weights**, which in this case equals the number of cases (7). This average would be 44.57 years for the seven cases included in the example.

Case ID	Weight	Stratum	X	Age	Age x Weight
1	1	1	20	18	18
2	1	1	20	21	21
3	1	2	200	42	42
4	1	2	200	38	38
5	1	2	200	37	37
6	1	3	2,000	72	72
7	1	3	2,000	84	84
Totals					
n=7	**sum=7**			**312**	**312**

Average = 312/7 = 44.57

This is, strictly speaking, a **weighted** average, but the weights are all 1 and thus trivial. For this reason this is called the **unweighted** average of the variable, not because it is really not weighted, but because all the weights equal 1.

After issuing the command WEIGHT BY X, the values of X are used as weights.[6] The average now would respond to the same formula: multiply individual ages by the weights, add up, and divide by the sum of weights, as shown in the following table.

Weight	Stratum	X	Case ID	Age	Age x Weight
20	1	20	1	18	360
20	1	20	2	21	420
200	2	200	3	42	8,400
200	2	200	4	38	7,600
200	2	200	5	37	7,400
2,000	3	2,000	6	72	144,000
2,000	3	2,000	7	84	168,000
Totals					
4,640			**n=7**	**312**	**336,180**

Average = 336,180 / 4,640 =72.45

The sum of weights is now 4,640, and the average is now 72.45 years of age as can be easily checked. The average age is now higher, because the older people in the sample (cases 6 and 7) are given much more weight than the rest of the sample. The "Totals" row shows not only the sum of ages (312) and the number of cases in the sample (7) but also the sum of weights and the sum of the products of age by weight. Their ratio 336,180 / 4,640 = 72.45 is the **weighted mean** of age for these seven subjects. This is a properly weighted estimate of mean age for the population represented by this sample. Notice that in both cases the mean is obtained by dividing the sum of Age x Weight by the sum of weights, though in the first case all weights were 1.

The weights used in the second example are **mixed** or **integrated** weights. They increase the scale of the figures and at the same time correct the proportions in the sample. For this reason, the total "number of cases" (equivalent to the sum of weights) appears to be 4,640 instead of 7. The weights here are generally the reciprocal of sampling ratios, N_k/n_k.

Any frequency tabulation coming from this weighted file will show a total of 4,640 cases, i.e. the sum of weights, because the seven cases "represent" 4,640 cases in the population.

For some purposes, this scale effect, augmenting table totals to population scale, may not be desirable. Purely proportional weighting with no scale effect can easily be substituted. To achieve this, we create a new variable Y which is the product of X by $n/N = 7/4640 = 0.00091623$.

[6] What is specified for weighting is the weighting variable itself, not its specific *values* at the time it is designated. If the values of X are subsequently modified, e.g. by a transformation command or by importing new values from some outside source, the *new* values of X would be automatically used for weighting, without requiring that the X variable is designated again as the weighting variable.

COMPUTE Y=X*7/4640.
FORMAT Y (F12.8).

[The FORMAT command instructs SPSS to show a maximum of 12 places for Y, including the decimal point, and 8 decimals, in the visible display of the new variable. This does not affect the internal precision of the variable (SPSS always has about 15 decimals of precision for all variables) but just its visible appearance.**]** The resulting weights are as follows:

Case ID	X	Y = Xn/N
1	20	0.03017241
2	20	0.03017241
3	200	0.30172414
4	200	0.30172414
5	200	0.30172414
6	2,000	3.01724138
7	2,000	3.01724138
Total	4,640	7.00000000

Since Y equals the old variable X divided by *N* and multiplied by *n*, the sum of Y equals 7, the sample size (except for rounding error). Now the weighted number of cases (i.e. the sum of weights) is again 7, with no change of scale. Any frequency table produced by this file, if weighted by Y, would give a total of 7 cases. The weights, however, are no longer all equal to 1. Some are lower than 1, some are larger. The first two cases are counted (roughly) as 0.03 cases each, the third and fourth as 0.3 cases each, and the last two cases as equivalent to 3 cases each.

The estimation of the mean of Age for these seven cases is done based on the weighted cases, i.e. multiplying each age by the corresponding weight. The mean, which is independent of scale, is always 72.45, as was when the mixed (and inflationary) weights **X** were used. The following table shows how the average age is now computed in the fictional sample of just seven cases. Notice that the sum of weights (first column) equals the number of cases in the sample.

Weight	Stratum	X	Y	Case ID	Age	Age x Weight
0.03017241	1	20	0.03017241	1	18	0.5431034
0.03017241	1	20	0.03017241	2	21	0.6336207
0.30172414	2	200	0.30172414	3	42	12.6724138
0.30172414	2	200	0.30172414	4	38	11.4655172
0.30172414	2	200	0.30172414	5	37	11.1637931
3.01724138	3	2000	3.01724138	6	72	217.2413793
3.01724138	3	2000	3.01724138	7	84	253.4482759
Totals						
Sum=7				**n=7**		**507.1681034**

Average = 507.1681034 / 7 = 72.45

To summarize, SPSS assumes by default that all cases have a unit weight, but users may use any variable to assign weight to cases. Usually, weights are a variable in the file, like X or Y in the above examples. Any [numerical] variable may be used for weighting through the syntax command WEIGHT BY… An alternative way is through the graphic interface menu, choosing Data / Weight data / Weight by… If a file is weighted by some variable, it remains so until the order is cancelled by WEIGHT OFF or another weight variable is established. If a weighted file is saved, it will still be weighted when re-opened. Also, it is worth noticing that the weighting facility uses the **current** values of the weighting variable: if the weighting variable is recalculated and given new values it is automatically re-weighted by the new values of the variable. There is no need to instruct SPSS again to use that variable just because it has new values.

If the sum of weights is not equal to the number of cases, i.e. if the weights are **scale weights**, the totals computed from the sample would be scaled up to the scale represented by the sum of weights, usually the scale of the population. This affects all frequencies and totals, including the number of cases in a table, the grand total of incomes for all households in a table, and so on.
Weights affect the **scale** of the cases and also the **proportion** of each case relative to others. Even for inflationary weights, **relative** quantities such as means or percentages are only affected by the **proportional** aspect of weights. If weights are different for different cases, even if the sum of weights equals the number of cases in the sample, weighted averages will differ from simple, non-weighted averages.
If a single scale weight is applied to all cases (for example, multiplying all cases by N/n), the scale of totals will be increased but percentages and means will not be affected, because proportions between cases have not been altered.
Weights are only taken into account when they are positive. Any non-positive value for the weight variable (negative, zero or missing) is not used. If a case has a negative, zero or missing weight, the case is not included in the analysis, and SPSS issues a warning in the output.

7. Significance tests and weighting in SPSS

As said before, SPSS always takes data as weighted, even if by default the weight of all cases is uniformly 1 and thus irrelevant. When a statistical procedure requires taking the total number of cases, SPSS invariably takes the sum of weights, not the actual number of cases in the file, even if often both amount to the same figure.
This means, for instance, that in the calculation of standard errors, which involve the square root of sample size (see Eq.8 and Eq. 9), the standard deviation is divided by the square root of the sum of weights. If the sum of weights is different from the actual number of cases in the file, as with inflationary weights, then SPSS may divide the standard deviation by a number that is far larger than the square root of sample size. In doing so, SPSS would estimate a standard error for its estimates that is misleadingly small.
The weights available for many sample surveys are of the mixed or integral type, combining the two functions of re-scaling and re-proportioning the sample. If so, the analyst is faced with a stark choice:

- If significance tests are computed with a weighted file using inflationary weights, proportions between strata will be right but the standard errors would be under-estimated because inflationary weights fool SPSS into believing that the sample is as large as the population. If a sample of, say, 1000 cases is taken from a population of 20 million, many small differences, which are in fact not significant with the given sample size, will appear as significant just because SPSS believes the "sample size" equals the estimated population size of 20 million.
- If weights are discarded, and significance tests applied on the unweighted data file, the tests may give wrong results when sampling ratios vary between groups of cases. The scale of the sample will be right but the proportions of the strata may be distorted. Sample cases from the under-sampled strata would receive the same weight as cases from the over-sampled strata, causing errors of estimation for both averages and standard errors (mean age in the above example would be estimated at 44 years instead of 72, and also standard errors may be wrong).

The point is that a correct analysis in this case requires using purely proportional weights, eliminating the scale effect. This could be achieved, as seen before, by multiplying all existing (inflationary) weights by n/N, whereby the sum of weights will be n, and the proportions between strata will be right.
This solution would get unbiased estimates and correct significance tests, **insofar as the effect of clustering in sample design is not important**. In fact, the real constraint is the intra-cluster

correlation of variables. A suitable summary of this constraint is as follows: "The degree of underestimation [of standard errors] depends upon the size of the intra-cluster correlation coefficient for the variables being analyzed. The higher the intra-cluster correlation, the more serious the underestimation of the variability".[7]

As a rule of thumb, whenever clustering is "populous" (meaning that a stratum is subdivided into a **large** number of relatively **small** clusters, like ZIP areas in a State, and then a **large** number of those numerous small clusters is actually selected), the error involved is likely to be slight. On the other hand, in the case of "sparse clustering" (meaning that the stratum is divided into a **small** number of relatively **large** clusters, as counties within a State) and then a **small** number of these large clusters is chosen, then the possibility of under-estimating the standard error would be greater. The rationale for this rule of thumb is that in the case of many small clusters intra-cluster correlation coefficients are likely to be relatively small. When no clustering is involved, or intra-cluster correlation coefficients are all small and possibly near zero, a purely proportional weight yields correct estimates of data variability and standard errors, and thus provides accurate estimates of significance.

For complex samples in general, and especially those involving clustering, SPSS provides a special module, called **Complex Samples**, which computes standard errors for a variety of sample designs.[8] A similar feature exists in other statistical software packages such as SUDAAN. Whenever possible, significance tests for complex sample data involving clustering should be conducted using those software packages. Proportional weights in the presence of large clusters and nonzero intra-cluster correlation are likely to underestimate error and overestimate significance.

The variance of the sampling distribution of a sample measure (mean, percentage, regression coefficient, etc.) can also be estimated through so-called *bootstrapping* methods. Essentially, this computational device involves taking a large number of random samples of the same size, based on the observed sample. The relevant measure (e.g. a mean) is taken in all these samples, and the sampling distribution of this measure is estimated based on the results of all samples.[9]

---

[7] Donna Brogan, "Pitfalls of using standard statistical software packages for sample survey data", Rollins School of Public Health, Emory University, Atlanta, 1997, appearing as a chapter in **Encyclopedia of Biostatistics**, edited by P. Armitage and T. Colton, Wiley, 1998, and also in the second edition, 2005.

[8] Complex Samples was originally distributed by SPSS as an independent piece of software, produced by another company, WesVar. In more recent times, starting with version 13, SPSS Inc. includes its own version of Complex Samples as an integrated module of SPSS.

[9] We have only one sample of size *n*, and cannot know the sampling distribution for all possible samples of that size. Bootstrapping estimates the sampling distribution with a large number of samples drawn **from the same single sample** we have. This is accomplished by using the given sample as a population of size *n*, and drawing many secondary samples of size *n* through sampling **with replacement**. Thereby a certain case A in the actual sample, once selected for a secondary sample, may possibly be selected again and again in the same secondary sample; thus sample size *n* may be reached before some other case (say B) is selected at all. Therefore a new sample of *n* cases may contain two or more copies of case A, and perhaps no copy of case B, thus yielding a different mean for the variable of interest. This is equivalent to having many samples composed of the same cases but giving varying weights to each case (represented by the number of replications of each case in the bootstrapped sample). In the above example, if case A is selected three times and case B is not selected, the resulting sample would have case A with a weight of 3, and case B with a weight of 0. After repeating this procedure many times (perhaps many thousands of times), obtaining as many secondary samples of size *n*, the distribution of the many means thus obtained will have a variance, which is an estimate of the unknown population variance. This procedure cannot possibly include cases in the population which were not included in the given sample, but it has been shown that the estimates obtained are usually quite good. See C. Z. Mooney and R. D. Duval, **Bootstrapping: A Nonparametric Approach to Statistical Inference** (Sage Publications, Quantitative Applications in the Social Sciences Series No. 95, Newbury Park, CA, 1993) for a short introduction.

To summarize, computing significance tests with SPSS without Complex Samples in the presence of unequal sampling ratios requires that the sample is weighted, and furthermore requires that weights are purely proportional (not altering the scale of results but keeping the sum of weights equal to sample size). If the sample design includes clustering, and especially if the clustering is "sparse" in the sense defined above (few large clusters instead of numerous small clusters), and/or many or all intra-cluster correlation coefficients are large, then SPSS Complex Samples (or SUDAAN) should be used.

8. Other issues with weighting in SPSS

8.1. The perils of fractional weights

When one out of every 20 cases has been sampled, and then the sampling ratio is 0.05, the scale weight of each case is 20. It means that it is one case in the sample, but it counts as 20 cases in the population. But it is often the case that the reciprocal of sampling ratios is not a nice integer but a fractional number. For instance, if one stratum has a population of 125,467 and a sample of 1,283 cases is used, the sampling ratio is 1,283/125,467 = 0.0102258 (rounded to 6 decimals). Its reciprocal is 97.791894, and therefore each case in the sample counts as more than 97 cases but less than 98 cases in the population.

SPSS in general accepts fractional weights, but they nonetheless involve some complications worth noticing. One problem is that there are certain statistical procedures which do **not** accept fractional weights, notably Cox Regression (the main instrument of Survival Analysis, which estimates the effects of independent variables on the chances of subjects to survive for longer or shorter periods before some event hits them). If fractional weights are being used, they should be reduced to integers before applying Cox Regression, or the procedure will fail.

Another problem with fractional weights is **rounding**. SPSS automatically rounds weighted **frequencies** to the nearest integer.[10] This rounding is done by default on the total weighted frequency, not on individual weights. In this case, table results may show some small inconsistencies. In the following fictional example, the table crosses two variables, approval of a proposition before and after reading some information. Six people disapproved, three of them before and three after reading. They correctly appear as six cases, 3 before and 3 after, in the unweighted table. In the weighted table they appear as 293 before and 293 after, but their total does not appear as 293+293=586, but as 587.

Unweighted

	Before	After	Total
Approve	…	…	…
Disapprove	3	3	6 = 3 + 3

Weighted

	Before	After	Total
Approve	…	…	…
Disapprove	293	293	587 ≠ 293+293

These small discrepancies in table totals are almost invariably the result of rounding in the presence of fractional weights. When weights are "large", rounding off the decimals does not change them by much. Between 98 and 97.79, for instance, there is a relative difference of only 0.2 percent, not likely to cause much trouble. But in the case of "small" weights, not very different from 1, rounding off the decimals may cause a large relative jump in the value of the weights. This is often the case with "purely proportional" weights which in general are defined by $(N_k/N)/(n_k/n)$. Typically, these weights may be numbers below or above 1, such as 0.80 or 1.45.

[10] Recent versions of SPSS let the user control how rounding is done in procedures like CROSSTABS.

Rounding off these decimals is not an indifferent matter. Both 0.80 and 1.45 will be rounded to 1, which kills the very purpose of weighting. Suppose the weight of some cases is 0.30. If one case with weight 0.30 appears in a cell, the weighted (and rounded) total will be zero cases in that cell. If a second cell includes two cases from the same stratum, their weighted total will be 0.60, rounded to one case. If two more cells have respectively 3 and 4 cases from the same stratum, their sum would be 0.90 and 1.20, but both will be again rounded to just one case, meaning that (with those numerical weights) one case in the sample turns into an estimate of no case at all, while two, three or four cases in the sample translate into an estimate of only one case; therefore passing from no case to one case would have no visible impact as the weighted cell would be empty in both cases (one truly empty, the other with 0.30 cases equivalent to none); doubling the number of cases from 1 to 2 would turn an empty cell into one with one inhabitant, but doubling the sample frequency again (from 2 to 4) may cause no effect at all.
All this does not sound right indeed. When large samples are involved, one or two cases added or subtracted may cause not much trouble, but in disaggregated tables with numerous cells there may be many low-frequency cells suffering the above problems, e.g. appearing empty although in fact populated, or appearing to have the same number of cases when in fact one of them harbors more cases than the other. Besides the statistical effect, readers may be disconcerted by these anomalies.
A parallel, but different problem, one much more troubling indeed, is the fact that SPSS rounds off the **frequencies**, but not the **totals of interval variables** in weighted cells. Suppose a cell in a table shows the frequency, i.e. the number of people in the cell, and the next cell shows their total income. Suppose a cell contains only one case with weight 0.30, representing a person who happens to have an income of $10,000. When computing the frequency SPSS would report zero cases (1 case x 0.30 = 0.30 weighted cases, rounded to zero), but the neighboring cell for total income would report $10,000 x 0.30= $3,000. This income of $3,000 would appear as earned by nobody, since the number of people in that cell is reportedly zero.
To avoid this in case like this one, SPSS in recent versions does not show any total or average when the sum of weights in a cell is below 1, but this in turn may cause inconsistent sums for rows or columns. Also, if the sum of weights exceeds 1, SPSS will show non-rounded totals and averages, and this may cause some problems and small discrepancies. When using fractional weights, a weighted frequency of 14.05 and another of 14.49 would both be rounded to 14, while average or total incomes for these cases are not rounded at all. Suppose two cells have an aggregate weighted income of $4,214.53 and their weighted frequencies are 14.02 and 14.49 people. The frequencies would both appear to be 14, the aggregate income would be $4,214.53 in both cases, but the mean income would be different: 300.6 for the first cell and 290.85 for the second. For large frequencies, say in the hundreds or thousands, this is hardly a problem, a few decimals could hardly make a difference, but for cells containing one-digit or two-digit frequencies these discrepancies might cause some confusion. Some warnings in the tables about figures not adding up due to rounding are advisable.

To avoid this problem altogether, it is preferable that **integer** weights are used for all work involving frequencies, but for calculations involving standard errors and significance, as explained before, sample size has to be preserved and thus purely proportional weights are advisable, and they are inherently fractional. Proportional weights are normally small and inevitably involve significant decimals. Thus users may hold two sets of weights in their data set, one inflationary set of **integer** weights and one purely proportional set, which may have decimals. If a user needs to produce a table **and** a significance test on its contents (for example a contingency table and the corresponding chi square statistic) it is best that the table itself is computed with integer weights, perhaps inflationary ones, to yield consistent frequencies, and the chi square with purely proportional weights. The SPSS procedure CROSSTABS, in this case,

would have to be run twice, once with each type of weight, using the table resulting from one and the chi square resulting from the other. The following syntax shows a simple example.

```
VAR LABEL X 'Mixed or integral weights' Y 'Purely proportional weights'.
WEIGHT BY X.
CROSSTABS OPINION BY SEX.
WEIGHT BY Y.
CROSSTABS OPINION BY SEX / STATISTICS=CHISQ.
WEIGHT OFF.
```

In this sequence of commands, first the data are weighted by X, which is a set of integrated or mixed weights with only integer values, producing both scale and proportion effects. A cross tabulation is ordered to show the distribution of certain opinions by sex. The resulting table (whose totals correspond to population totals because the weights inflate the scale to population size) will be used for presenting the data in tabular form. Then the data set are re-weighted by Y, a purely proportional set of weights preserving sample size. The second CROSSTABS specifies that the chi square statistic is produced alongside the table. The table produced in this second run would be disregarded, but the chi square from this run will be used to test the null hypothesis that opinion and sex are unrelated. For procedures not reporting frequencies, such as REGRESSION, proportional weights are recommended in order to preserve sample size and thus produce realistic significance test results.
However, if the sampling design involves clustering, however, even using proportional weights may produce biased significance tests, and therefore the COMPLEX SAMPLES module should be used.

8.2. Empty categories in tables and charts

There is another use of weighting in SPSS which comes very handy on occasion. In SPSS versions prior to version 12, categories without any cases are not shown in tables and charts. For example, if a question asks for the most recent country visited by respondents, and nobody responds "Switzerland", then Switzerland will not appear in the table or chart produced with that question. It is often the case that users want all categories shown, even the "empty" ones. There is no apparent option to that effect in SPSS, except a trick using weighting. The trick consists of creating a **fictitious additional case**, whose last visited country is Switzerland, and assigning this case a very small weight such as 0.000001. This weight will be rounded to zero, and thus the case would not appear in the frequency tables or charts, but the Switzerland category will be shown.
To show how to do this, suppose the new case is manually added at the end of the data set. Suppose you need Switzerland to appear in table showing last country visited by sex of the travelers. Only three variables must have valid values in the new case: ID, LASTTRIP and SEX. Assume the fictitious case has an ID=99999, and the variable LASTTRIP is marked for that case with the code for Switzerland. The sex of this imaginary traveler may be male or female; suppose we make it female (SEX=2 in the codification used for that survey). Suppose that the file, previously, had been weighted by a variable called OLDWGT. The following commands may be issued:

```
COMPUTE NEWGT=OLDWGT.
IF (ID=99999)NEWGT=0.000001.
WEIGHT BY NEWGT.
CROSSTABS LASTTRIP BY SEX.
```

In the resulting tabulation of last destination by sex a line for Switzerland will be shown, but the entire line will be empty because the only case there (a female) has been weighted down to zero.

More precisely, the cell for females will show a zero, and the cell for males will be blank (if you want both to show zeros, you may create **two** fictitious cases, one for each sex). The weighted number of females and the weighted total number of cases in the table will not be affected by this procedure. Notice that the fictitious case needs not only the data about having visited Switzerland, but also the information about sex: if sex for the new case is blank, it would be a system-missing-value case and would not be included in the table anyway. Of course, once its purpose is fulfilled the fictitious case may be deleted from the file to avoid error or confusion in the future. But before that, by introducing valid values of other variables into the phantom case, the user may be able to cause other empty categories to appear in tables, as with the Switzerland destination used as an example above.

Notice that even if the weighted frequency is rounded to zero, the fictitious case still has values in the variables. What is rounded to zero is the frequency of that case, but not the value of any interval variable characterizing it. Suppose the table crosses the destination of latest trip with annual income. For Switzerland to appear in the table you must create a case with valid values both for country visited and for income. In recent versions of SPSS, as said before, totals or means are not reported for cells where the sum of weights is lower than 1, but this is unclear for older versions. If the figure assigned for income is sufficiently large, and the SPSS version not recent, some total income may possibly appear in the cell. For instance, if the fictitious total income assigned to the phantom case is $100,000, and its weight is 0.0001, the weighted and rounded number of cases would be zero but the weighted and rounded total income would be $10, which is inconsistent. To be on the safe side, therefore, just make sure that the weighted income of the phantom case is sufficiently low (less than $0.50) as to be rounded down to zero. For instance, if the phantom case is assigned an income of $100, and its weight is 0.0001, the weighted income would be $0.01, which would be rounded to $0 anyway. An equivalent approach is making the weight sufficiently small (e.g. changing it from 0.0001 to 0.000000000001). This would ensure that (if you leave the phantom case in the file) the income of the phantom case does not get added to the income of real cases in other procedures.

Using weights for empty categories, however, is becoming obsolete, as it could be avoided in recent versions of SPSS. In versions 12 and later, the CTABLES procedure allows users to decide whether to include or exclude empty categories, through the /EMPTY=INCLUDE statement within the /CATEGORIES subcommand. Likewise, in version 14 the same effect can be obtained for charts.

8.3. Other people's tables

Another special use of weighting with SPSS is related to analysis of tables taken from books, articles or any other sources. Sometimes in the course of research one finds a table in some publication, and wishes to ascertain, say, the significance of a percentage difference or the degree of association between the concerned variables. To achieve this, the user should create a small SPSS data file representing the **cells** in the table. Each cell in a contingency table corresponds to a combination of values of variables in the table, and the frequency in the table represents the number of cases having that particular combination of values. Suppose the table in question is as follows, showing type of occupation by region of residence.

	1. Blue collar	2. Clerical	3. Professional	4. Self-employed
1. East	332	418	125	62
2. West	465	328	211	87
3. South	225	248	152	112

The user wants to know whether any association exists between the two variables, but the source does not give any association measure. To achieve this, the user should assign codes to the values

of each variable, and prepare an SPSS file with the following structure, in which each cell is converted into a "case":

Region	Occup	Freq
1	1	332
1	2	418
1	3	125
1	4	62
2	1	465
2	2	328
2	3	211
2	4	87
3	1	225
3	2	248
3	3	152
3	4	112

Of course region 1 is East, 2 is West and 3 is South, and likewise for occupations. If the file with this structure is in the data editor of SPSS, and variable FREQ is used for weighting, a cross tabulation would produce the same original table with the desired association statistics. The required syntax would be:

```
WEIGHT BY FREQ.
CROSSTABS REGION BY OCCUP/STATISTICS=ALL.
```

The STATISTICS keyword may be specified to produce all the available measures, as in this example, or just the ones the researcher wants, for instance chi square (replacing keyword ALL by CHISQ in the above CROSSTABS command). In this example, the Pearson chi square value estimated by SPSS is 101.03436, while the alternative estimation by likelihood ratio is 98.03069, both of which are significant ($p<0.00001$).
Therefore the WEIGHT command in SPSS does not only fulfill the function of giving data their proper proportional and/or absolute weight, but may also be used for other purposes as shown above.