COMBINING ESTIMATES FROM SUBPOPULATIONS TO FORM AN ESTIMATE FOR THE ENTIRE POPULATION

Suppose we have several (say k) sub-populations or "strata" of sizes N_1, N_2, ... N_k, which form one entire population of size $\sum N_k = N$. Suppose we are interested in some quantitative or qualitative characteristic of this entire population. We denote this numerical or binary characteristic in each individual by Y, and an aggregate or summary (across all individuals in the population) by θ, which could stand for an average (μ), a total quantity ($T_{amount} = N\mu$), a proportion (π), a percentage (% = 100π) or a total count ($T_c = N\pi$). Examples of θ include:

<u>If Y is a measured variable (i.e. "numerical")</u>

μ:	the annual (per capita) consumption of cigarettes
T_{amount}:	the total undeclared yearly income
	($T_{amount} = N\mu$ and conversely that $\mu = T_{amount} \div N$)

<u>If Y is a binary variable (i.e. "yes/no")</u>

π:	the proportion of persons who exercise regularly
100π %:	the percentage of children who have been fully vaccinated
$N\pi$:	the total number of persons who need R_x for hypertension
	($T_c = N\pi$; $\pi = T_c \div N$)

The sub-populations might be age groups, the two sexes, language groups, occupations, provinces, etc. There is a corresponding θ for each of the K sub-populations, but one needs subscripts to distinguish one subpopulation from another. Rather than study every individual, one might instead measure Y in a *sample* from each subpopulation.

• To estimate the overall μ, π, or $\pi_{\%}$, combine the estimates as follows:

Sub Popln	Size	Relative Size $W_i = N_i \div N$	Sample Size	Estimate of θ_i	SE of estimate
1	N_1	W_1	n_1	e_1	$SE(e_1)$
2	N_2	W_2	n_2	e_2	$SE(e_2)$
...	...	...	...	...	......
k	N_k	W_k	n_k	e_k	$SE(e_k)$
Total	$\sum N = N$	$\sum W=1$	$\sum n=n$	$\sum W_i e_i$	$\sqrt{\sum W_i^2 [SE(e_i)]^2}$

• To estimate the overall T_{amount} or T_c , use weights of $W_i = N_i$

Note1- If any sampling fraction $f_i = n_i \div N_i$ is sizable, the SE of the e_i should be scaled down i.e. it should be multiplied by $\sqrt{(1-f_i)}$

Note2- If an unstratified sample of size n is taken, but later stratified into k substrata, each of the n_i will be approximately the same fraction of its corresponding N_i. Thus, the estimate from the single overall sample will, by virtue of its self-weighting nature, be relatively unbiased and will be close to the weighted estimate. However, if one ignores the strata, the SE calculated from the single unstratified sample of n may be too large: If the variability in Y within a stratum is smaller than across strata, the smaller SE obtained from the SE's of the individual stratum specific estimates more accurately reflects the uncertainty in the overall estimate.