

AGRODEP  
Household survey data course  
Dakar, 8-10 October 2012

# Weights



HarvestChoice  
BETTER CHOICES, BETTER LIVES



# Reasons for Weighting


Kish (1992) identifies six reasons for weighting survey data prior to analysis:

1. To reflect differential selection probabilities
2. To reduce biases introduced by errors in the sampling frame
3. To reduce bias introduced by non-response
4. To reduce sampling variance, by making use of auxiliary information (post-stratification)
5. To produce standardized estimates
6. To produce approximately unbiased estimates from a sample formed by combining other samples

# Sampling Weights

- Sampling weights are adjustment factors applied to each observation during the analysis to adjust for differences in probability of selection between cases in a sample, either due to design or randomness.
- In its simplest form, this is the reciprocal of the probability of selection.

# In many household surveys

- The sample is stratified by
  - Region
  - Urban / Rural
- Within each stratum
  - PSUs are selected with PPS 
  - Households are selected with equal probability within each PSU
- Then

$$p_{hij} = P_1 P_2 = \frac{k_h n_{hi}}{N_h} \frac{m_{hi}}{n'_{hi}}$$

- $p_{hij}$  = Probability of selecting household  $hij$  in PSU  $hi$  of stratum  $h$
- $P_1$  = Probability of selecting the PSU in stage 1
- $P_2$  = Probability of selecting the household stage 2
- $k_h$  = Number of PSUs selected in stratum  $h$
- $n_{hi}$  = Size of PSU  $hi$
- $N_h$  = Size of stratum  $h$
- $m_{hi}$  = Number of households selected in PSU  $hi$
- $n'_{hi}$  = Number of households listed in PSU  $hi$

$$p_{hij} = P_1 P_2 = \frac{k_h n_{hi}}{N_h} \frac{m_{hi}}{n_{hi}}$$

- If
  - The measure of size used in  $n_{hi}$  and  $N_h$  is the number of households
  - The same number of households are selected in all PSUs ( $m_{hi}=m$ , a constant)
  - $n_{hi}$  and  $n'_{hi}$  are the same in all PSUs
- Then
  - the formula simplifies to  $p_{hij} = k_h m / N_h$  (a constant in each stratum)
  - In other words, the sample is self-weighted within each stratum
- In practice, these conditions are seldom totally true. The sample is only approximately self-weighted.

# Sampling Weights

- Generally, weighting calculations are more complex and include many factors. A typical weight found in a household survey may be composed as follows:

$$W_{final} = W_{sel} \times W_{ps} \times W_{nr}$$

with *sel* represent selection, *ps* representing post – stratification, and *nr* representing non-response.

# Post Stratification

- Post stratification is generally the last step in the process and adjusts weighted totals to known population totals and has been shown in the literature to reduce overall variance.
- Example:

State	Weighted Total Population from Survey	Known Population	Adjustment Factor
Maryland	5,245,757	5,699,478	1.0865
Virginia	8,475,901	7,882,590	0.9300
DC	662,842	599,657	0.9047



# Non-Response

- Nearly every survey has some degree of non-response for which weights need to be adjusted.
- It is important to note that this is a non-response adjustment, not a “correction.” Without perfect information on all variables, it is not possible to completely correct for non-response.
- The best we can do is try to estimate the bias and make the best adjustment possible based on the information available.

# Non-Response

- **Simplest form:** If your cluster has 12 households, but one refuses, a simple calculation for the *nr* component of the weights would be  $\left(\frac{11}{12}\right)^{-1}$  or 1.09.
  - Each remaining household counts a bit more than 1 to make up for its missing neighbor.
- Other common methods of adjustment for non-response:
  - **Weighting class:** divides the data into cells (such as age x gender x geography) and assigns a correction factor based on the cell response.
  - **Propensity adjustment:** uses a basic regression to model non-response (based on propensity score)

# Raking

- Similar to post-stratification, most commonly used for non-response correction
- Example:

Let's analyze poverty in Washington. A baseline household survey is conducted to get a poverty measure. However, we know from the literature that men and minority populations have low response rates – what to do?

# Raking

- Weighted totals from your survey:

	White	African American	Latino/ Hispanic	Other	Total
Male	79,586	125,489	22,566	4,581	232,222
Female	97,089	185,057	22,689	5,422	310,757
Total	176,675	310,546	45,255	10,003	542,479

- Totals from census bureau

	White	African American	Latino/ Hispanic	Other	Total
Male					281,839
Female					317,818
Total	179,897	359,794	47,973	11,993	599,657

# Raking

- Rake across:



	White	African American	Latino/ Hispanic	Other	Total
Male	96,590	152,301	27,387	5,560	281,839
Female	99,295	189,262	23,205	5,545	317,818
Total	195,886	341,563	50,592	11,105	599,657

1.214

1.023

- Rake down:

	White	African American	Latino/ Hispanic	Other	Total
Male	0.902	0.909	0.895	0.901	
Female					
Total	179,897	359,794	47,973	11,993	599,657

- Repeat until factors converge to 1.

# Raking

- After convergence:

	White	African American	Latino/ Hispanic	Other	Total
Male	88,927	160,865	26,028	6,019	281,839
Female	90,970	198,929	21,945	5,974	317,818
Total	179,897	359,794	47,973	11,993	599,657

- Divide sample totals by raking totals to find adjustment factors

	White	African American	Latino/ Hispanic	Other
Male	88,927/79,586	160,865/125,489	26,028/22,566	6,019/4,581
Female	90,970/97,089	198,929/185,057	21,945/22,689	5,974/5,422

# Raking

- These adjustment factors would then be used as the  $w_{nr}$  factor in the weight calculations.

	White	African American	Latino/ Hispanic	Other
Male	1.117	1.282	1.153	1.314
Female	0.937	1.075	0.967	1.102

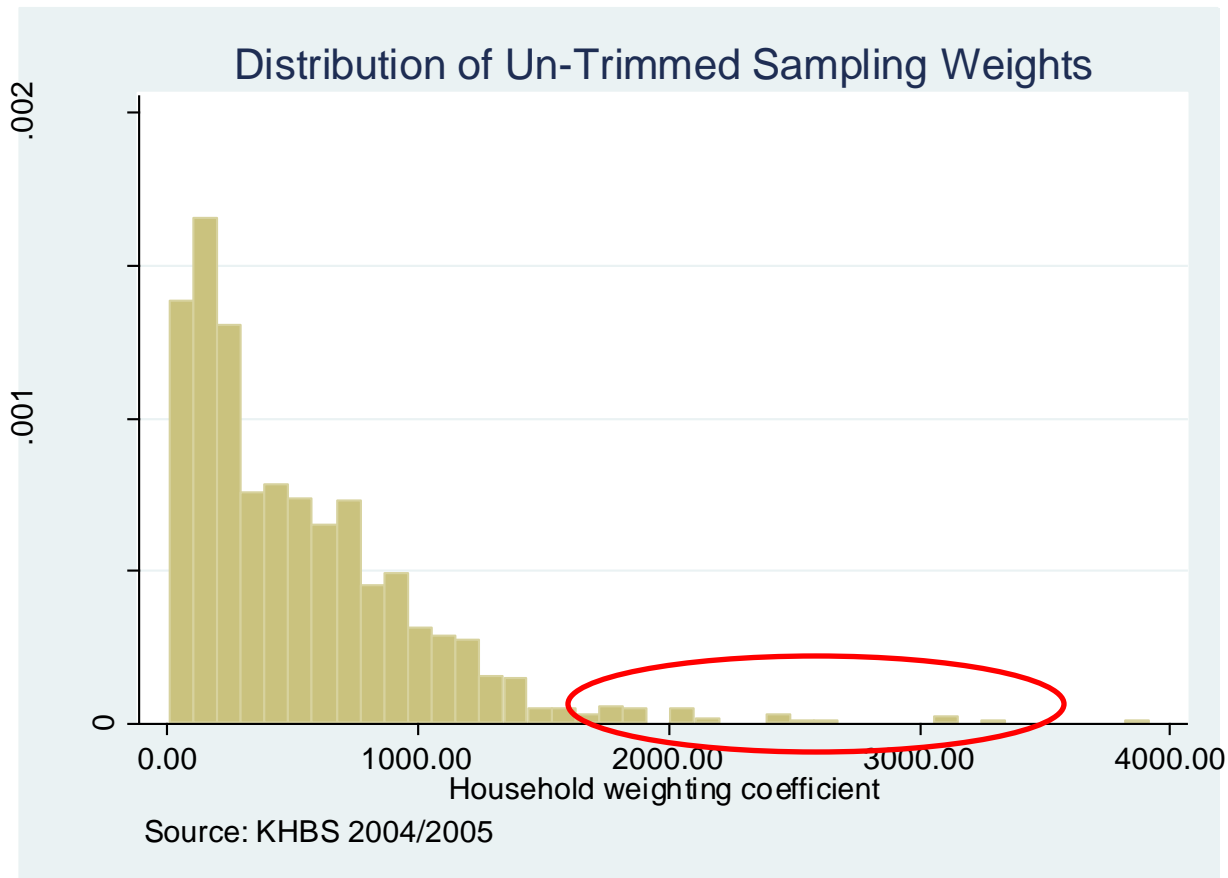
# Trimming

- Trimming weights replaces outlier weights to reduce the variance of the resulting estimations.
- This causes some bias in the estimates and needs to be carefully considered against gains in precision.
- **Note:** this is almost always done by the data provider and not the analyst.



# Trimming

- Like any other variable, weights have a distribution.



# Trimming

- Trimming the outlier weights decreases the standard errors but biases the estimate.

	Mean	Std. Err.	CV
untrimmed	6.899	0.434	0.878
99	6.882	0.429	0.830
95	6.875	0.421	0.760
90	6.895	0.422	0.713
75	6.964	0.386	0.544

# Final Thoughts

Coming back to the formula:

$$W_{final} = W_{sel} \times W_{ps} \times W_{nr}$$

The factors represent the importance given to different sources of evidence.

$W_{sel}$  is calculated by us – through our direct actions – we can be very confident of the quality of this calculation.

$W_{ps}$  is based on an external source – we have to decide how confident we are in the reliability of that source.

$W_{nr}$  is based on theory. Since there is no way to empirically prove the theory – our level of confidence is based on the literature - and how closely we feel that our situation matches this literature.