# Disadvantages of SRS

- Can't guarantee coverage of particular subpopulations. Comparison of regions (see stratification)
- Low-probability outcomes hard to measure (see stratification)
- Expensive (see multi-stage)

# Stratification - How does over sampling affect inference?

| | | | |
|---|---|---|---|
| **FRAME** | | | |
| **Income** | | **Stratum 1** | **Stratum 2** |
| 10, 10, 10, | | 10, 10, 10, | 3, 3, 3, (low) |
| 9, 9, 9, | STRATIFY SAMPLE TO | 9, 9, 9, | 2, 2, 2, (low) |
| 8, 8, 8, | ENSURE POOR | 8, 8, 8, | 1, 1, 1 (low) |
| 7, 7, 7, | ARE REPRESENTED | 7, 7, 7, | |
| 6, 6, 6, | | 6, 6, 6, | |
| 5, 5, 5, | | 5, 5, 5, | |
| 4, 4, 4, | | 4, 4, 4, | |
| 3, 3, 3, (low) | | | |
| 2, 2, 2, (low) | | P=3/21 | P=3/9 |
| 1, 1, 1 (low) | | W=7 | W=3 |

Choose 3 obs from each stratum. (why?) Low income stratum has 3 in 9 chance, high income 3 in 21. Therefore, we need to give smaller weights to low-income draws so that weighted sample estimates will reflect characteristics of the population.

# Sample weight example
# (We' ll look at stratification example later)

Nepal Living Standard Survey (NLSS)

weights_example.do

Help File

# Comment (Stata weights example)

- Variance and standard deviation of the distribution of y

$$\mathrm{Var}(y) = \frac{1}{(n-1)}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2, \quad \mathrm{Stnd.Dev.} = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

- (Variance &) Standard error of the mean of y, Assuming SRS

$$\mathrm{Var}(\bar{y}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2, \quad \mathrm{SE}(\bar{y}) = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

# Stratification - an algebraic statement regarding precision.

Let $\hat{\sigma}_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}\left(y_{i,h} - \bar{y}_h\right)^2$, or the sample variance of y in stratum h. Discuss the terms.

# Stratification - precision (cont.)

Because each sample is an independent, random draw within each stratum, the variance of $\bar{y}$ for the full sample can be expressed as ...

$$\hat{\sigma}_{SS}^2 = \frac{1}{n}\sum_{h=1}^{H}\omega_h \hat{\sigma}_h^2 \ ,$$

where SS abreviates 'stratified sample', n is the sample size,

H is the number of strata, and

$\omega_h$ is the proportion of the sample in stratum h $(n_h/n)$.

Or, in words, the sample variance of $\bar{y}$ is the weighted average of the variance in each stratum.

# Stratification - precision (cont.)

With some algebraic manipulation of terms, $\hat{\sigma}^2_{SS}$ can be written in terms of $\hat{\sigma}^2_{SRS}$ (i.e. the variance of $\bar{y}$ if $y_i$ had been drawn following a Simple Random Sample).

$$\hat{\sigma}^2_{SS} = \hat{\sigma}^2_{SRS} - \frac{1}{n}\sum_{h=1}^{H}\omega_h(\bar{y}_h - \bar{y})^2$$

Discuss terms, note that result => $\hat{\sigma}^2_{SS} < \hat{\sigma}^2_{SRS}$ .

Or, in words, stratifying the sample can improve the precision (reduce the variance) of the estimator.

# Stratification - implications (cont.)

$$\hat{\sigma}^2_{SS} = \hat{\sigma}^2_{SRS} - \frac{1}{n}\sum_{h=1}^{H}\omega_h(\bar{y}_h - \bar{y})^2$$

=> Precision is increased as ....

... heterogeneity across strata $\uparrow$.

Or in other words the more $\bar{y}_h$ differs from $\bar{y}$.

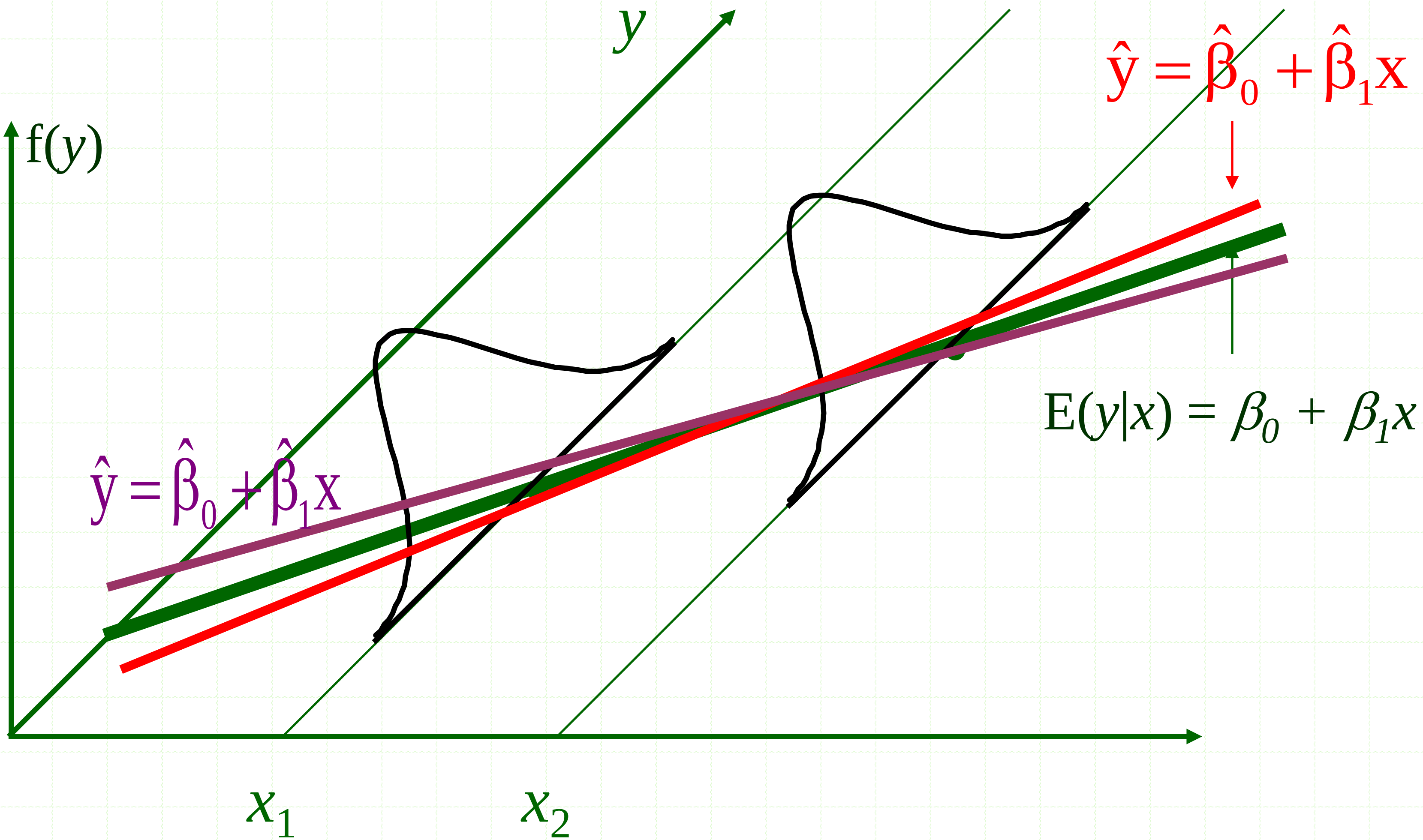
... homogeneity within strata $\uparrow$.

Decreased variance within each stratum ($\hat{\sigma}^2_h$) reduces the sum of the strata variances, or reduces $\hat{\sigma}^2_{SS}$.

Caveats: $\omega_h$ and relative magnitudes in practice.

# Comment: N or n?

# How precisely parameters are estimated is determined by sample not population size



# Stratification - Illustration of increased precision

Overhead Slides to provide some intuition

$$\hat{\sigma}^2_{SS} < \hat{\sigma}^2_{SRS}$$