Weight and See
Antony Unwin
Augsburg University


Weighting by population

Percentage black population in 437 MidWest counties

- Unweighted
- Weighted by black population
- Weighted by total population
- Weighted by white population

Weighting?

- Frequency weighting
- Probability weighting
- Importance weighting
- Analytic weighting
- (Other kinds of weighting)
Weighted graphics in R?

- Counts of “weight”, “wtd”
  - Paul Murrell’s book 0
  - Hadley Wickham’s Ph D 1
- Google within R
  - “weighted histogram” 1
  - “weights” 2350
  - “weighted” 4860

An illustrative example

There are $M$ groups with populations

$$N_j \quad \text{where} \quad j = 1, \ldots, M \quad \text{and} \quad \sum_{j=1}^{M} N_j = N.$$  

Simple random samples are taken from each group of sizes

$$n_j \quad \text{where} \quad j = 1, \ldots, M \quad \text{and} \quad \sum_{j=1}^{M} n_j = n.$$  

Each person $i$ sampled from group $j$ is checked to see if they are ill or healthy.

$$Z_{ij} = \begin{cases} 1 & \text{if ill}, \\ 0 & \text{if healthy}, \end{cases} \quad \text{with} \quad i = 1, \ldots, n_j \text{ and } j = 1, \ldots, M$$

Frequency weighting (FW)

- Multiple independent invididuals have the same values
- Typically multivariate categorical data
- e.g., the Berkeley discrimination dataset either 4526 cases (Gender, Admit, Dept) or 24 combinations G(2) x A(2) x D(6)

FW example

Let $m_{kj}$ denote the number of persons in group $j$ with health $k$, i.e.

$$m_{kj} = \sum_{i=1}^{n_j} (Z_{ij} = k)$$

then the dataset can be written in the form

$$(k, j, m_{kj}) \quad \text{where} \quad k \in \{0, 1\} \quad \text{and} \quad j = 1, \ldots, M.$$  

The values $m_{kj}$ are then frequency weights:

$$w_{kj}^{FW} = m_{kj} \quad \text{with} \quad m_{0j} + m_{1j} = n_j$$
Probability weighting (PW)

• When not all cases have the same probability of being sampled
  – stratified sampling
  – cluster sampling
  – adjusting for non-response
• Weighting is the inverse of the probability of being sampled

PW example

In the illustrative example, the dataset \( \{Z_{ij}\} \) with \( i = 1, \ldots, n_j \) and \( j = 1, \ldots, M \) has \( n \) entries and probability weights that are inversely proportional to their sampling probability:

\[
w_{ij}^{PW} \propto \frac{1}{n_j/N_j} = \frac{N_j}{n_j}
\]

If individual \( i \) has probability \( p_i \) of appearing in the sample, then his weight

\[
w_i^{PW} \propto 1/p_i
\]

The weights may be normalised so that they sum to 1 or to an ‘equivalent sample size’ \( n \).

Importance weighting (IW)

• Cases have different importance
  – companies weighted by turnover or market capitalisation
  – market share changes weighted by shares
  – regional morbidity rates weighted by population
  – medical side-effect rates weighted by seriousness of effect

IW example

In the illustrative example, the dataset of illness rates

\[
\{ r_j = \frac{\sum_{i=1}^{n_j} Z_{ij}}{n_j} \}
\]

has importance weights

\[
w^{IW}(r_j) = N_j
\]

In this case it is interesting to note that

\[
\sum_{i=1}^{n_j} w_{ij}^{PW} \propto w^{IW}(r_j)
\]
Analytic weighting (AW)

- Combining results from studies
  - e.g., Metanalysis
- Weighting is the inverse of the variance (precision) or of the standard deviation
- cf. Bayesian modelling

**AW example**

Consider again the dataset of illness rates

\[ r_j = \frac{\sum_{i=1}^{n_j} Z_{ij}}{n_j} \quad \text{for } j = 1, \ldots, M. \]

Assuming an underlying binomial model \( Z_{ij} \sim B(n_j, p_j) \), gives analytic weights that are inversely proportional to variances:

\[ w_j^{AW}(r_j) \propto \frac{n_j}{r_j(1-r_j)}. \]

(If \( N_j >> n_j \) there is no need to correct for finite sampling.)

Assuming \( r_j \approx r \) for all \( j \) leads to

\[ w_j^{AW} \propto n_j = w_j^{FW} \]

Properties of weights

- Additive
- Non-negative
- No missing values in the weighting variable
- Weights have to be interpretable

Types of graphic

- **Point graphics**
  - dotplots, glyphs, scatterplots, rotating plots, parallel coordinate plots, graphs
- **Area graphics**
  - histograms, barcharts, mosaicplots, missing value plots
    (And what is a boxplot?)
Weighted area plots

- Barcharts and spineplots
- Histograms and spinograms
- Mosaicplots and their variants
  - multiple barcharts
  - fluctuation diagrams
  - same binsize
  - doubledecker

Self-weighted histograms

- Posterior probabilities of 36 models in a Bayesian Model Averaging analysis

Graphics for metanalysis

- Study results should be weighted by precision
- Displaying CIs for results gives more space to less precise studies (forest plots)
- Funnel plots would benefit from including precision with weight represented by size

Traffic flows

- Augsburg city centre traffic flows superimposed on a map of the city.
- Flows are represented by triangles whose base is proportional to the flow.
Software

• Mondrian
  – all platforms
  – stats.math.uni-augsburg.de/Mondrian/

• MANET
  – (only for Macintosh)
  – stats.math.uni-augsburg.de/MANET/

More on graphics

Summary comments

• Weighted graphics should be considered more often
• There are different kinds of weighting with different interpretations
• Importance weighting offers most flexibility and requires most care
• Weighted plots provide additional insights