Introduction to Inverse Probability Weighting

James Carpenter & Mike Kenward

Department of Medical Statistics
London School of Hygiene & Tropical Medicine

James.Carpenter@lshtm.ac.uk
https://missingdata.lshtm.ac.uk

June 2005

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Table of Contents

LONDON SCHOOL *of* HYGIENE &TROPICAL MEDICINE

Introduction

The aim of this presentation is to

- give an intuitive justification for Inverse Probability Weighting (IPW);
- look at a simple example;
- discuss methods to improve efficiency, and
- contrast with multiple imputation.

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Idea behind inverse probability weighting

Suppose the full data are

Group:		A			B			C	
Response:	1	1	1	2	2	2	3	3	3

The average response is 2. However, we observe:

Group:		A			B			C	
Response:	1	?	?	2	2	2	?	3	3

From the observed data, the average response is 13/6, biased.

Notice the probability of response is 1/3 in group A, 1 in group B and 2/3 in group C.

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Idea behind inverse probability weighting

Calculate weighted average, where each observation is weighted by $1/\{\text{Probability of response}\}$:

$$\frac{1 \times \frac{3}{1} + (2+2+2) \times 1 + (3+3) \times \frac{3}{2}}{\frac{3}{1} + 1 + 1 + 1 + \frac{3}{2} + \frac{3}{2}} = 2.$$

IPW has eliminated the bias in this case; more generally it will give estimators the property they 'home in' on the truth as the sample size increases (i.e. they are *consistent*).

A more mathematical view

Most estimators are the solution of an equation like

$$\sum_{i=1}^{n} U(x_i, \theta) = 0.$$

For example, if $U_i(x_i, \theta) = (x_i - \theta)$, solving

$$\sum_{i=1}^{n} (x_i - \theta) = 0$$

gives $\theta = \sum_{i=1}^{n} x_i / n$.

A more mathematical view

Theory says that if the average of $U_i(X_i, \theta_{true})$ over samples from the population is zero, our estimate will 'home in' on the truth as the sample gets large (this is called *consistency*).

If some of the observations x_i are unobserved, then it follows the corresponding $U's$ are missing from the above sum. Thus the average of $U_i(\theta_{true})$ is not zero, so estimates won't 'home in' on the truth.

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

A more mathematical view

However, now let

$$R_i = \begin{cases} 1 & \text{if } x_i \text{ is observed, with probability } p_i \\ 0 & \text{if } x_i \text{ is not observed} \end{cases}$$

Then, the average (over repeated samples from the population) of

$$\sum_{i=1}^{n} \frac{R_i}{p_i} U(x_i, \theta_{true}) = 0,$$

so parameter estimates will 'home in' on the truth as the sample size gets large.

In general, inverse probability weighting recovers consistent estimates when data are missing at random.