A NOTE ON THE RELATION OF INVERSE-PROBABILITY-WEIGHTING AND MATCHING ESTIMATORS

Michael Lechner

Swiss Institute for Empirical Economic Research

University of St.Gallen

Revised version: July 2009

Date this version has been printed: 27 September 2012

Pre-print. Published Paper can be downloaded from
http://dx.doi.org/10.1080/03610920903453434
(Communications in Statistics- Theory and Methods).

Abstract

This paper compares the inverse-probability-of-selection-weighting estimation principle with the matching principle and derives conditions for weighting and matching to identify the same and the true distribution, respectively. This comparison improves the understanding of the relation of these estimation principles and allows constructing new estimators.

Keywords: Matching, inverse-of-selection-probability weighting, treatment evaluation, unconfoundedness.

JEL classification: C21, C13, C14.

Addresses for correspondence

Michael Lechner
Professor of Econometrics
Swiss Institute for Empirical Economic Research (SEW)
University of St. Gallen, Varnbüelstrasse 14, CH-9000 St. Gallen, Switzerland
Michael.Lechner@unisg.ch, www.sew.unisg.ch/lechner

1 Introduction*

This paper considers the estimation problem that occurs when the distribution of a random variable *Y* in population (*t*arget*)* cannot be directly observed but must be learned from its distribution in another population (*o*bserved*)*. Such a situation is typical for so-called observational studies in which the task is to the compare the outcome of some treatment (*S*) for the treated population (*S*=*t*) with the outcome (*Y*) that would occur without the treatment *(S*=*o)*. The former can be directly learned from the data. In an observational study, the latter has to be learnt from the observed non-treatment outcomes (*Y*) of the non-treated (*o*). By doing so, one has to take into account that exogenous factors (*X*) influencing the outcome may be differently distributed among treated and non-treated. Therefore, using the unweighted mean of the observed outcome for the non-treated to proxy the non-treatment outcomes of the treated leads to confounding. This confounding can however be eliminated by an appropriate reweighting of the data taking based on the distribution of *X* among the treated and the non-treated (see for example Rubin, 1974, 1979) if it comes from these observable variables only. Clearly, beyond so-called treatment evaluation studies, there are many other situations where such adjustments for exogenous factors are a helpful and frequently used empirical exercise.

Matching methods that are in many cases based on the so-called propensity score (i.e. the conditional on *X* probability of being observed in the treated instead of the non-treated population) and inverse-of-selection-probability weighting schemes (IPW) are very popular choices to perform these adjustments non- or semi-parametrically. Imbens (2004) provides an excellent re-

* I am also affiliated with ZEW, Mannheim, CEPR and PSI, London, IZA, Bonn, and IAB, Nuremberg. I thank Markus Frölich for helpful comments on a previous version of this paper. I also thank an anonymous referee of this journal for helpful comments that improved the readability of the paper considerably.

view of many aspects and variants of these classes of estimators and nicely summarizes the literature.

In this paper, we compare the IPW estimation *principle* with the matching estimation *principle* and show that they are closely related. Conditions on the properties of the respective weights are derived such that weighting and matching on those weights identify the same distribution. Further conditions are provided for the identified distribution being the distribution of interest. It is not the objective of this paper to analyse the small sample properties (e.g. Frölich, 2004), nor is it the goal to study the asymptotic properties (e.g. Abadie and Imbens, 2006, Firpo, 2007) of the many different estimators that are available in such a context. Instead, we investigate the conditions the weights have to fulfil in the population (or in an indefinitely large i.i.d. sample) to form the basis for a consistent estimator.

Such a comparison does not only deepen the understanding of the *relation* of those two frequently used estimation *principles*, but does also allow constructing new estimators. For example, if a weighting estimator is known to be consistent and its weights fulfil the conditions derived in this paper, then we show that a consistent matching-type estimator can be constructed solely based on such weights (such weights typically have a close connection to the propensity score).

A comparison of these estimation principles from this population perspective has not yet received much attention in the literature. There are some estimator specific asymptotic and small sample results of estimators based on IPW and stratification, which is very similar to matching, by Lunceford and Davidian (2004). However, their comparison is for a case when both estimation principles necessarily identify the same density. Furthermore, there are formal comparisons (e.g. Ichimura and Taber, 2001, Frölich and Lechner, 2006), as well as more informal comparisons (e.g. Hogan and Lancaster, 2004), of instrumental variable methods with weighting and matching approaches. They use a different perspective, though, and the method of instrumental variables is not the focus of this paper.

The plan of this note is as follows: In the next section same notation is introduced, the objects of interest are defined, and regularity conditions are imposed to simplify the subsequent analysis. Section 3 defines the estimation principles. In Section 4, matching and weighting estimators are related. Section 5 concludes.

2 Notation, targets for the estimation, and sampling frame

There are two subpopulations depending on the value of the random variable S, namely the *target* population ($S=t$) and the *observed* population ($S=o$).[1] Realisations of Y, which is the random variable of interest, are only available for the observed population, whereas realisations of a set of confounding variables X (and weights W to be introduced below) are observed in both subpopulations. Using the language of Rubin (1974, 1979), Y is a potential outcome and would be indexed as Y^o, i.e. the outcome a unit would experience if subject to regime $S=o$. In this literature, there is usually a corresponding potential outcome for the second treatment state ($S=t$), Y^t. As there is no conceptional different between the two potential outcomes, we concentrate on the first case only. Thus, in this paper, Y refers to Y^o only and the indexing for potential outcomes is not necessary.

---

[1] As a convention, capital letters denote random variables and small letters denote either realisations or specific values of those random variables. Furthermore, $F_{A|B}(a,b)$ denotes the cumulative distribution functions of A conditional on B evaluated at the value a for B = b. $f_{A|B}(a,b)$ is the corresponding density. If A is discrete, $f_{A|B}(a,b)$ is the corresponding mass function.

We are interested in learning the ('counterfactual') distribution of *Y* in the subpopulation in which it is not observed, i.e. $F_{Y|S}(y,t)$. To do so, the information in *X* and *W* will be combined with the corresponding distribution of *Y* in the population (*S*=*o*) in which realisation of *Y* are observable.

In Assumption 1, the sampling scheme is formalized and enough regularity and support is assumed to concentrate on the key issues of comparing the two estimation principles.

Assumption 1 (sampling frame, common support, regularity)

(a) Sampling: There is a random sample of size *N* ($\{s_i, x_i, w_i, y_i\}_{i=1}^{N}$) from the joint distribution of the random variables (*S, X, W, Y*). In the subsample with $s_i = t$, the values of y_i are not observable.

(b) Support and regularity: Assume that all probabilities, densities, and moments that are of interest exist, are finite, and have sufficient support. Densities and probabilities are nonzero if required.

Two remarks are in order concerning Assumption 1: First, the i.i.d. sampling assumption is not important for what follows. What is important is that y_i is observed in one subsample, but not in the other. Second, below we will consider the case that the weights are deterministic functions of *S* and *X*, but this restriction of generality is not necessary so far. Third, part b) is obviously overly restrictive (and somewhat imprecise), but it allows concentrating on the main issues in this paper without additional, inessential, notation.

Having defined the general model, the following section presents the different population problems that matching and direct weighting estimators solve as estimation primciples.

3 Principles of direct weighting and a matching estimation

It is the idea of the weighting estimator to take the empirical mean in the observable population over an individual specific weight times a function of the observed random variables of interest in the subpopulation in which *y* is observed. This idea is formalized in Definition 1.

Definition 1 (estimation principle of weighting estimator)

The estimation principle of the weighting estimator is defined for a one dimensional random variable *W such that:* $F^{W}_{Y|S}(y,t) = \underset{W|S=o}{E}\left[wF_{Y|W,S}(y,w,o)\right]$.

At this stage of generality, $F^{W}_{Y|S}(y,t)$ may or may not be equal to the true distribution $F_{Y|S}(y,t)$ and *w* may or may not correspond to some function of selection probabilities. The idea of using a monotone functions of the inverse selection probability (the so-called propensity score, defined as $P(S=t \mid X=x, S\in\{t,o\})$ by Rosenbaum and Rubin, 1983) for weighting is probably due to Horowitz and Thomson (1952), but has since then been analysed by many others.

A second set of weights define a matching-type estimator, which, like all estimators, has an interpretation as a weighting estimator as well. The difference between direct weighting and matching is how the weights are defined. For direct weighting, the weights are directly defined such that their expectation over the observable population fulfils a certain restriction. For matching, the weights are implicitly determined from the distribution of some variables in the population for which *Y* is not observable. We call those variables *X*.

Definition 2a (matching principle based on covariates)

The vector of random variables *X* defines the principle of matching estimating such that:

$$F^{M(x)}_{Y|S}(y,t) = \underset{X|S=t}{E}\left[F_{Y|X,S}(y,x,o)\right].$$

As for weighting, at this level of generality there is no need to assume that $F_{Y|S}^{M(x)}(y,t)$ = $F_{Y|S}(y,t)$. The idea of adjusting observed (confounding) variables directly could be traced back to Fechner (1860). It is more formally discussed in Wilks (1932) and Rubin (1974, 1979). The conditions required for $F_{Y|S}^{M(x)}(y,t)$ = $F_{Y|S}(y,t)$ come under the heading of the conditional independence or the no confounding assumptions (e.g., see Cochrane and Chambers, 1965, Rubin, 1974). These assumptions imply that there is no confounding once the exogenous characteristics X are controlled for. Practical estimation methods for matching estimation as well as weighting estimators are extensively discussed by Imbens (2004). Based on the observation by Rosenbaum and Rubin (1983) that controlling for the propensity score $P(S=t \mid X=x, S \in \{t,o\})$ removes confounding as well if it is removed by controlling for X directly, many of the frequently used estimators are based on controlling / matching / weighting by some function of the propensity score.

Analogously to the weighting principle, Definition 2b) defines a matching estimator that is based on a one-dimensional covariate, which we call W as before. Later on, we show that W may be directly related to that propensity score.

Definition 2b (matching principle based on weights)

The random variable W (weight) defines the principle of weight-based matching estimating such that: $F_{Y|S}^{M(w)}(y,t) = \underset{W|S=t}{E}\left[F_{Y|W,S}(y,w,o)\right]$.

The key distinction between these three definitions is the different dimension of the random variables W and X and how they are used to adjust the distribution. A particular example in which a matching estimator is consistent for a one-dimensional conditioning variable is of course the propensity score (see Rosenbaum und Rubin, 1983).

The following analysis sheds more light on the relation between these estimation principles.

4 The relation of the estimation principles

Do weights exist for which both estimation principles identify the same distribution? Theorem 1 establishes the existence of such weights and shows that they have a well-known form.

Theorem 1 (weights that lead to equivalence of weighting and matching)

a) If Assumption 1 holds, then the following equation (explicitly) defines the weights that lead to $F^{W}_{Y|S}(y,t) = F^{M(x)}_{Y|S}(y,t)$:

$$w = w(x) = \frac{P(S=t \mid X=x)}{P(S=o \mid X=x)} \frac{P(S=o)}{P(S=t)} \equiv \frac{f_{X|S}(x,t)}{f_{X|S}(x,o)}.$$

b) If Assumption 1 holds, then the following equation (implicitly) defines the weights that lead to $F^{W}_{Y|S}(y,t) = F^{M(w)}_{Y|S}(y,t)$:

$$w = \frac{P(S=t \mid W=w)}{P(S=o \mid W=w)} \frac{P(S=o)}{P(S=t)} \equiv \frac{f_{W|S}(w,t)}{f_{W|S}(w,o)}.$$

The proof of this theorem is relegated to Appendix A.1.

Note that this theorem does not establish that either a matching or a weighting estimator based on these weights identifies the true distribution of interest. However, if one of the estimation principles leads to the true distribution using such weights, then Theorem 1 implies that the other estimation principle recovers the true distribution as well.

The weights appearing in Theorem 1a) and 1b) are so-called inverse-probability-of-selection (IPW) weights either as function of X or of the summary measure W. One such summary measure that fulfils this criterion in the binary treatment model under the unconfoundedness

(Rubin, 1974) assumption is an appropriate transformation of the already mentioned propensity score (see Rosenbaum and Rubin, 1983). It is also obvious that the weights are monotone functions of these selection probabilities.[2]

Next, Theorem 2 provides conditions under which the matching and weighting principles identify the true distribution. As Y is unobservable if $S=t$, these conditions are identifying, i.e. they have no empirical counterpart and can thus not be tested.

Theorem 2 (weights leading to consistent matching estimation)

a) If the vector X fulfils the following condition, then the matching principle identifies $F_{Y|S}(y,t)$:

$$\underset{X|S=t}{E}\left[F_{Y|X,S}(y,x,t)-F_{Y|X,S}(y,x,o)\right]\overset{!}{=}0 \text{, or}$$

$$\underset{X|S=o}{E}\left\{\frac{P(S=t\,|\,X=x)}{P(S=o\,|\,X=x)}\frac{P(S=o)}{P(S=t)}\left[F_{Y|X,S}(y,x,t)-F_{Y|X,S}(y,x,o)\right]\right\}\overset{!}{=}0 .$$

b) If the weights W fulfil the following condition, then the matching principle identifies $F_{Y|S}(y,t)$:

$$\underset{W|S=t}{E}\left[F_{Y|W,S}(y,w,t)-F_{Y|W,S}(y,w,o)\right]\overset{!}{=}0 \text{, or}$$

$$\underset{W|S=o}{E}\left\{\frac{P(S=t\,|\,W=w)P(S=o)}{P(S=o\,|\,W=w)P(S=t)}\left[F_{Y|W,S}(y,w,t)-F_{Y|W,S}(y,w,o)\right]\right\}\overset{!}{=}0 .$$

---

[2] Note that $P(S=t\,|\,X=x)=P(S=t\,|\,X=x,S\in\{t,o\})\,P(S\in\{t,o\}\,|\,X=x)$ and $P(S=o\,|\,X=x)=P(S=o\,|\,X=x,S\in\{t,o\})\;P(S\in\{t,o\}\,|\,X=x)$. Thus, the second term cancels in the expression above to yield: $\frac{P(S=t\,|\,X=x)}{P(S=o\,|\,X=x)}=\frac{P(S=t\,|\,X=x,S\in\{t,o\})}{1-P(S=t\,|\,X=x,S\in\{t,o\})}$.

Noting that $f_{W|S}(w,t)$ = $\frac{P(S=t\,|\,W=w)}{P(S=o\,|\,W=w)}\ \frac{P(S=o)}{P(S=t)}\ f_{W|S}(w,o)$ and $f_{X|S}(x,t)$ = $\frac{P(S=t\,|\,X=x)}{P(S=o\,|\,X=x)}\ \frac{P(S=o)}{P(S=t)}\ f_{X|S}(x,o)$,[3] the proof is direct. It is therefore omitted for the sake of brevity. These conditions are less restrictive than directly requiring $F_{Y|X,S}(y,x,t)$ = $F_{Y|X,S}(y,x,o)$ or $F_{Y|W,S}(y,w,t)$ = $F_{Y|W,S}(y,w,o)$, as the standard formulation of the 'no-confounding' assumption would do. However, any difference between $F_{Y|X,S}(y,x,t)$ and $F_{Y|X,S}(y,x,o)$, or between $F_{Y|W,S}(y,w,t)$ and $F_{Y|W,S}(y,w,o)$, has to be averaged away in the respective distribution of X or W.

For further comparisons with the weighting principle to be discussed below it is useful to reformulate $F_{Y|W,S}(y,w,t)$ in terms of $F_{Y|W,S}(y,w,o)$, i.e. $F_{Y|W,S}(y,w,t)$ = $\frac{P(S=t\,|\,Y\le y,W=w)}{P(S=o\,|\,Y\le y,W=w)}\ \frac{P(S=o\,|\,W=w)}{P(S=t\,|\,W=w)}\ F_{Y|W,S}(y,w,o)$. From this property, we restate the conditions of Theorem 2b):

Theorem 2b'

If the weights *W* fulfil the following condition, then the matching principle identifies $F_{Y|S}(y,t)$:

$$\underset{W|S=o}{E}\left\{\frac{P(S=o)}{P(S=t)}\left[\frac{P(S=t\,|\,Y\le y,W=w)}{P(S=o\,|\,Y\le y,W=w)}-\frac{P(S=t\,|\,W=w)}{P(S=o\,|\,W=w)}\right]F_{Y|W,S}(y,w,o)\right\}=$$
$$=\underset{W|S=t}{E}\left\{\left[\frac{P(S=o\,|\,W=w)}{P(S=t\,|\,W=w)}\,\frac{P(S=t\,|\,Y\le y,W=w)}{P(S=o\,|\,Y\le y,W=w)}-1\right]F_{Y|W,S}(y,w,o)\right\}\overset{!}{=}0.$$

[3] This result follows directly from $\frac{f_{A,B}(a,b)}{f_{A|B}(a,b)}=\frac{f_{B|A}(b,a)\,f_A(a)}{f_{A|B}(a,b)}=f_B(b)=\frac{f_{A,B}(a',b)}{f_{A'|B}(a',b)}=\frac{f_{B|A}(b,a')\,f_A(a')}{f_{A|B}(a',b)}$.

If the classical 'no confounding' assumption is fulfilled conditional on the weights (i.e. $P(S=t \mid Y \leq y, W=w)$ = $P(S=t \mid W=w)$ and $P(S=o \mid Y \leq y, W=w)$ = $P(S=o \mid W=w)$), or if $\frac{P(S=o \mid W=w)}{P(S=t \mid W=w)}$ = $\frac{P(S=o \mid Y \leq y, W=w)}{P(S=t \mid Y \leq y, W=w)}$, then the true distribution of interest is identified. Note that if the conditions of Theorem 2b) or 2b') are satisfied and if the weights have the form as in Theorem 1b), then the weighting principle based on such weights identifies the true distribution of interest as well.

Next, we consider the condition such that weighting identifies the true distribution.

Theorem 3 (weights leading to consistent weighting estimation)

If the weights *W* fulfil the following condition, then the weighting principle identifies $F_{Y|S}(y,t)$:

$$\underset{W|S=o}{E}\left\{\left[\frac{P(S=t \mid Y \leq y, W=w)}{P(S=o \mid Y \leq y, W=w)} \frac{P(S=o)}{P(S=t)} - w\right] F_{Y|W,S}(y,w,o)\right\} =$$

$$= \underset{W|S=t}{E}\left\{\frac{P(S=o \mid W=w)}{P(S=t \mid W=w)}\left[\frac{P(S=t \mid Y \leq y, W=w)}{P(S=o \mid Y \leq y, W=w)} - w\frac{P(S=t)}{P(S=o)}\right] F_{Y|W,S}(y,w,o)\right\} \overset{!}{=} 0.$$

The proof of Theorem 3 is contained in Appendix A.2. Clearly, if w = $\frac{P(S=t \mid Y \leq y, W=w)}{P(S=o \mid Y \leq y, W=w)} \quad \frac{P(S=o)}{P(S=t)}$, this criterion is fulfilled. However, since $P(S=t \mid Y \leq y, W=w)$ cannot be directly learned from the data,[4] more assumptions are needed. They could be of the different types extensively discussed in the literature (e.g. Heckman, LaLonde, and Smith, 1999). For example, if $\frac{P(S=t \mid Y \leq y, W=w)}{P(S=o \mid Y \leq y, W=w)}$ is equal to

[4] Obviously if $P(S=t \mid Y \leq y, W=w)$ cannot be consistently estimated, $P(S=o \mid Y \leq y, W=w)$ cannot be learned from the data either.

$\frac{P(S=t|W=w)}{P(S=o|W=w)}$, which follows from the 'no confounding' assumption, then the weights fulfil the conditions for the equality of matching and weighting and matching leads to consistent estimation as well. Note again that the matching and weighting conditions are identical for IPW-selection on observables weights: $w = \frac{P(S=t|W=w)}{P(S=o|W=w)}\frac{P(S=o)}{P(S=t)}$.

5 Conclusion

We compare the inverse-probability-of-selection-weighting principle with the matching principle. Conditions on the properties of the respective weights are derived such that weighting and matching on those weights identify the same distribution. Further conditions are provided such that this distribution is the true one. Under suitable regularity conditions, estimators using consistent estimators of such weights appropriately will usually be consistent. It is also shown that those weights can be directly linked to the propensity score.

This comparison deepens the understanding of the relation of these estimation principles and allows constructing new consistent estimators. For example, if a weighting estimator is consistent and its weights fulfil the conditions derived in this paper, then a consistent matching-type estimator can be constructed based solely on such weights.

References

Abadie, A., and G. W. Imbens (2006): "Large Sample Properties of Matching Estimators for Average Treatment Effects", *Econometrica*, 74, 235-267.

Cochran, W. J., and S. P. Chambers (1965): "The Planning of Observational Studies of Human Populations", *Journal of the Royal Statistical Society, Series* A, 128 (2), 234-266.

Fechner, G. T. (1860), *Elemente der Psychophysik*, Leipzig: Breitkopf und Härtel.

Firpo, S. (2007): "Efficient Semiparametric Estimation of Quantile Treatment Effects", *Econometrica*, 75, 259-276.

Frölich, M. (2004): "Finite sample properties of propensity-score matching and weighting estimators," *Review of Economics and Statistics*, 86, 77-90.

Frölich, M., and M. Lechner (2006): "Exploiting regional treatment intensity for the evaluation of labour market policies", 5728, C.E.P.R. Discussion Papers.

Heckman, J. J., R. LaLonde, and J. Smith (1999): "The Economics and Econometrics of Active Labor Market Programs", in: O. Ashenfelter and D. Card (eds.), *Handbook of Labour Economics*, Vol. 3, 1865-2097, Amsterdam: North-Holland.

Hogan, J., and T. Lancaster (2004): "Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies", *Statistical Methods in Medical Research*, 13, 17-48.

Horvitz, J. L., and D. Thomson (1952): "A Generalisation of Sampling without Replacement from a Finite Population", *Journal of the American Statistical Association*, 47, 663-685.

Ichimura, H., and C. Taber (2001): "Propensity-Score Matching with Instrumental Variables", *American Economic Review - Papers and Proceedings*, 91, 2, 119-124.

Imbens, G. W. (2004): "Nonparametric Estimation of Average Treatment Effects under Exogeneity: a Review", *Review of Economic and Statistics*, 86, 4-29.

Lunceford, J. K., and M. Davidian (2004): "Stratification and weighting via the propensity score in estimation causal treatment effects: a comparative study", *Statistics in Medicine*, 23, 2937-2960.

Rosenbaum, P. R., and D. B. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrica*, 70, 41-50.

Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology*, 66, 688-701.

Rubin, D. B. (1979): "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies", *Journal of the American Statistical Association*, 74, 318-328.

Wilks, S. S. (1932): "On the distribution of statistics in samples from a normal population of two variables with matched sampling of one variable," *Metron*, 9, 87-126.

Appendix A: Proof of Lemmas and Theorems

A.1 Proof of Theorem 1

Start the proof of Theorem 1a) by rewriting both estimation principles using iterated expectations (*I.E.*) to make them more easily comparable:

$$F_{Y|S}^{M(x)}(y,t) = \underset{X|S=t}{E}\left[F_{Y|X,S}(y,x,o)\right] \overset{I.E.}{=} \underset{X|S=t}{E}\left[\underset{W|X,S=o}{E} F_{Y|W,X,S}(y,w,x,o)\right].$$

Applying Bayes' Law to the densities of X conditional on S, we get the following equality:

$$f_{X|S}(x,t) = \frac{P(S=t \mid X=x)P(S=o)}{P(S=o \mid X=x)P(S=t)} f_{X|S}(x,o).$$

Therefore, we can change expectation in the first part of the previous equation to ease comparison between the estimation principles:

$$\begin{aligned}
\underset{X|S=t}{E}\left[\underset{W|X,S=o}{E} F_{Y|W,X,S}(y,w,x,o)\right] &\overset{\text{Bayes Law}}{=} \underset{X|S=o}{E}\left[\frac{P(S=t \mid X=x)P(S=o)}{P(S=o \mid X=x)P(S=t)} \underset{W|X,S=o}{E} F_{Y|W,X,S}(y,w,x,o)\right] = \\
&= \underset{X|S=o}{E}\left[\underset{W|X,S=o}{E} \frac{P(S=t \mid X=x)P(S=o)}{P(S=o \mid X=x)P(S=t)} F_{Y|W,X,S}(y,w,x,o)\right] \\
&= \underset{W,X|S=o}{E}\left[\frac{P(S=t \mid X=x)P(S=o)}{P(S=o \mid X=x)P(S=t)} F_{Y|W,X,S}(y,w,x,o)\right].
\end{aligned}$$

Next, iterated expectations are applied to the weighting principle:

$$\begin{aligned}
F_{Y|S}^{W}(y,t) &= \underset{W|S=o}{E}\left[wF_{Y|W,S}(y,w,o)\right] \overset{I.E.}{=} \underset{W|S=o}{E}\left[\underset{X|W=w,S=o}{E} wF_{Y|W,X,S}(y,w,x,o)\right] = \\
&= \underset{W,X|S=o}{E}\left[wF_{Y|W,X,S}(y,w,x,o)\right].
\end{aligned}$$

Therefore, the difference between the distributions identified by the two estimation principles is:

$$F^{M(x)}_{Y|S}(y,t) - F^{W}_{Y|S}(y,t) = \underset{W,X|S=o}{E}\left\{\left[\frac{P(S=t\,|\,X=x)P(S=o)}{P(S=o\,|\,X=x)P(S=t)} - w\right]F_{Y|W,X,S}(y,w,x,o)\right\}.$$

Therefore, if $w = w(x) = \dfrac{P(S=t\,|\,X=x)P(S=o)}{P(S=o\,|\,X=x)P(S=t)}$, both estimation principles have the same limit (sufficient condition). Since $P(S=t\,|\,X=x) = \dfrac{f_{X,S}(x,t)}{f_X(x)} = \dfrac{f_{X|S}(x,t)P(S=t)}{f_X(x)}$ and $P(S=o\,|\,X=x) = \dfrac{f_{X|S}(x,o)P(S=o)}{f_X(x)}$, we obtain the second representation of the weights as $w = w(x) = \dfrac{f_{X|S}(x,t)}{f_{X|S}(x,o)}$ shown in Theorem 1a). q.e.d.

The proof of Theorem 1b) proceeds along the same lines as the previous one, but without the explicit conditioning on *X* that is not necessary in this case. Thus, we get:

$$F^{M(w)}_{Y|S}(y,t) = \underset{W|S=t}{E}\left[F_{Y|W,S}(y,w,o)\right] \overset{\text{Bayes Law}}{=} \underset{W|S=o}{E}\left[\frac{P(S=t\,|\,W=w)P(S=o)}{P(S=o\,|\,W=w)P(S=t)}F_{Y|W,S}(y,w,o)\right].$$

Therefore, if $w = \dfrac{P(S=t\,|\,W=w)P(S=o)}{P(S=o\,|\,W=w)P(S=t)} = \dfrac{f_{W|S}(w,t)}{f_{W|S}(w,o)}$, both estimation principles have the same limit. q.e.d.

A.2 Proof of Theorem 3

In the proof of this theorem Bayes' law will be frequently applied, i.e. $\dfrac{f_{A,B}(a,b)}{f_{A|B}(a,b)} = f_B(b) = \dfrac{f_{A,B}(a',b)}{f_{A'|B}(a',b)}$, with *A*, *B* being random variables. From this property, we get three conditions that are helpful in the proof of this theorem (as well as of Theorem 2b'):

$$F_{Y|W,S}(y,w,t) = \frac{P(S=t\,|\,Y\le y, W=w)}{P(S=o\,|\,Y\le y, W=w)}\frac{P(S=o\,|\,W=w)}{P(S=t\,|\,W=w)}F_{Y|W,S}(y,w,o); \qquad \text{(A.1)}$$

$$f_{W|S}(w,t)=\frac{P(S=t\,|\,W=w)P(S=o)}{P(S=o\,|\,W=w)P(S=t)}\,f_{W|S}(w,o)\,; \tag{A.2}$$

$$f_{W|S}(w,o)=\frac{P(S=o\,|\,W=w)P(S=t)}{P(S=t\,|\,W=w)P(S=o)}\,f_{W|S}(w,t)\,. \tag{A.3}$$

Using those properties, the proof of the first part of Theorem 3 is direct:

$$\begin{aligned}
&\underset{W|S=t}{E}F_{Y|W,S}(y,w,t)-\underset{W|S=o}{E}wF_{Y|W,S}(y,w,o)=\\
&\overset{(A.3)}{=}\underset{W|S=o}{E}\left[\frac{P(S=t\,|\,W=w)P(S=o)}{P(S=o\,|\,W=w)P(S=t)}F_{Y|W,S}(y,w,t)\right]-\underset{W|S=o}{E}wF_{Y|W,S}(y,w,o)=\\
&\overset{(A.1)}{=}\underset{W|S=o}{E}\left[\frac{P(S=o)}{P(S=t)}\frac{P(S=t\,|\,Y\le y,W=w)}{P(S=o\,|\,Y\le y,W=w)}F_{Y|W,S}(y,w,o)\right]-\underset{W|S=o}{E}wF_{Y|W,S}(y,w,o)=\\
&=\underset{W|S=o}{E}\left\{\left[\frac{P(S=o)}{P(S=t)}\frac{P(S=t\,|\,Y\le y,W=w)}{P(S=o\,|\,Y\le y,W=w)}-w\right]F_{Y|W,S}(y,w,o)\right\}\overset{!}{=}0.
\end{aligned}$$

We get the second part of Theorem 3 by combining the last line above with property (A.2):

$$\begin{aligned}
&\underset{W|S=o}{E}\left\{\left[\frac{P(S=o)}{P(S=t)}\frac{P(S=t\,|\,Y\le y,W=w)}{P(S=o\,|\,Y\le y,W=w)}-w\right]F_{Y|W,S}(y,w,o)\right\}=\\
&\overset{(A.2)}{=}\underset{W|S=t}{E}\left\{\frac{P(S=o\,|\,W=w)P(S=t)}{P(S=t\,|\,W=w)P(S=o)}\left[\frac{P(S=o)}{P(S=t)}\frac{P(S=t\,|\,Y\le y,W=w)}{P(S=o\,|\,Y\le y,W=w)}-w\right]F_{Y|W,S}(y,w,o)\right\}=\\
&=\underset{W|S=t}{E}\left\{\frac{P(S=o\,|\,W=w)}{P(S=t\,|\,W=w)}\left[\frac{P(S=t\,|\,Y\le y,W=w)}{P(S=o\,|\,Y\le y,W=w)}-w\frac{P(S=t)}{P(S=o)}\right]F_{Y|W,S}(y,w,o)\right\}\overset{!}{=}0.
\end{aligned}$$

q.e.d.