Causal Inference from Experimental Data

Philip Dawid

30th Fisher Memorial Lecture

10 November 2011

Decision problem

I have a headache. Should I take aspirin?

- Two possible treatments:
  - t: take 2 aspirin
  - c: do nothing
- Outcome measure Y: time it takes my headache to go away
- Loss: $L(y)$ if $Y = y$.
- Whichever treatment applied, Y is uncertain
  - model as random

Which treatment should I choose?

Decision tree (hypothetical approach)

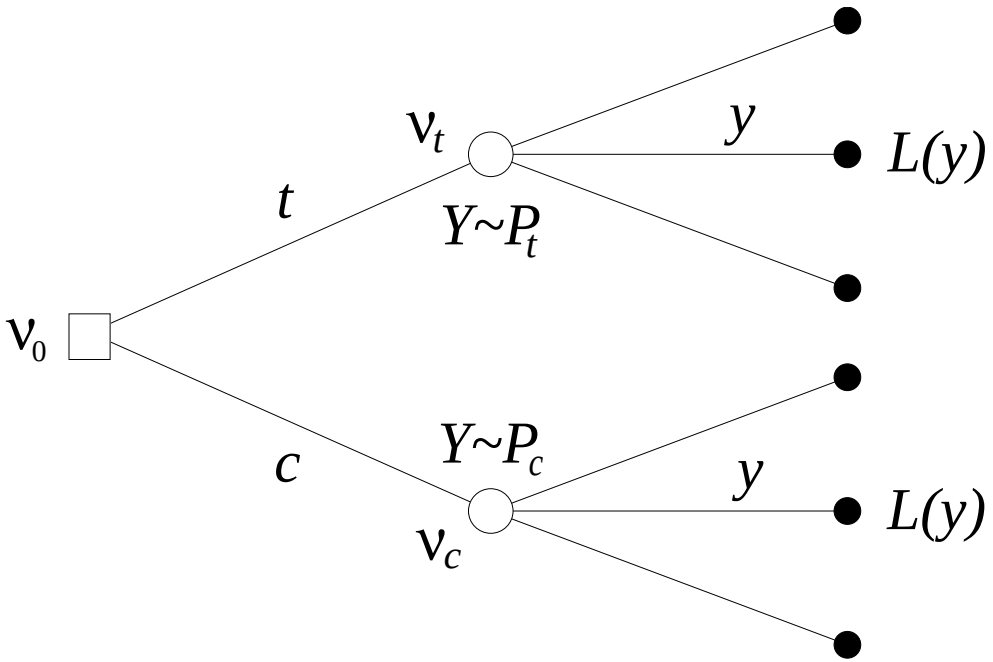

Figure: Decision tree

Choose t if $\mathrm{E}_{P_t}\{L(Y)\} \leq \mathrm{E}_{P_c}\{L(Y)\}$

Potential responses (counterfactual approach)

- Y_t if t applied
- Y_c if c applied

- Both pre-existing: joint distribution for (Y_t, Y_c)
- Application of t [c] uncovers value of Y_t [Y_c]

$$\text{Choose } t \text{ if } \mathrm{E}\{L(Y_t)\} \leq \mathrm{E}\{L(Y_c)\}$$

In either approach, relevant data might be gathered to:

- reduce uncertainty
- estimate "objective" distributions
  - **OBSERVATIONAL STUDIES**
  - **EXPERIMENTAL STUDIES**

Causal inference:
Use such data to help predict my outcome (under any hypothesised treatment) and so guide my decision

Predictive causal inference

Suppose I have done a study to measure the responses Y for two groups of individuals:

Treatment group: Given aspirin (t), responses $Y_{t1}, \ldots, Y_{tn}$

Control group: No aspirin (c), responses $Y_{c1}, \ldots, Y_{cm}$

For my own decision problem, I need to predict my own response Y under two hypothesised scenarios:

- I will take aspirin, t: $Y \sim P_t$ (or, potential response Y_t)
- I will not take aspirin, c: $Y \sim P_c$ (or, potential response Y_c)

Predictive causal inference

IF I can consider myself exchangeable with the members of the treatment group (on all relevant pre-treatment variables), then I can

- estimate my P_t
- predict Y, given t
- predict Y_t

from their responses.

Likewise, IF I can consider myself exchangeable with the members of the control group, then I can estimate my P_c from their responses.

Predictive causal inference

For both to be valid, it is necessary (though not sufficient) for the two groups to be exchangeable with each other.

This is the reason why we must take care that in our study we are comparing like with like:

- randomisation
- blocking
- covariates
- ...

Experimentation

Treatments $t \in \mathcal{T}$, units $u \in \mathcal{U}$

Each set may be structured, *e.g.*:

- factorial/nested/. . . treatments
- row-by-column, split-plot . . . layout
- blocking, covariates,. . .

Can apply any t to any u and measure outcome variable Y

Experimentation

- Select a set $\mathcal{U}_0$ of experimental units
- Assign treatments to them
  - map $\tau : \mathcal{U}_0 \to \mathcal{T}$

— each possibly with random element

Want to assess/compare the effects of the treatments

- ideally for new unit u_0 (= me?)

What is the *causal effect* of a treatment?

Hypothetical approach (Fisher)

Consider distribution P_t of response Y_u of unit u if assigned treatment t

e.g. $Y_u \sim N(\alpha_t, \phi_Y)$

Causal contrast is *e.g.* $\mathrm{E}_t(Y_u) - \mathrm{E}_{t'}(Y_u)$

- How much longer/shorter I expect my headache to last, if I were to take aspirin rather than nothing
- $\alpha_t - \alpha_{t'}$
- Average Causal Effect, ACE_u

What is the *causal effect* of a treatment?

Counterfactual approach (Neyman, Rubin)

Consider values $(\eta_{tu} : t \in \mathcal{T})$ of the responses of each unit u to each treatment t

- POTENTIAL RESPONSES

Causal contrast is *e.g.* $\eta_{tu} - \eta_{t'u}$

- How much longer/shorter my headache would last, if I were to take aspirin rather than nothing
- Individual Causal Effect, ICE_u

Counterfactual approach needs to posit the simultaneous existence (and joint distribution) of η_{tu} and $\eta_{t'u}$.

- We observe η_{tu} for $u \in \mathcal{U}_0$, $t = \tau(u)$
- We can never observe both η_{tu} and $\eta_{t'u}$
- So we can never observe ICE_u

Fundamental Problem of Causal Inference (FPCI) (Holland)

The physical array

t_1	Y_{11}	Y_{12}	$\cdots$	$\cdots$
t_2	$\cdots$	$\cdots$	$\cdots$	Y_{2n}
t_3	Y_{31}	$\cdots$	$\cdots$	Y_{3n}

The metaphysical array

	u_1	u_2	u_3	$\cdots$	$\cdots$	u_{N-1}	u_N
t_1	η_{11}	η_{12}	η_{13}	$\cdots$	$\cdots$	$\eta_{1,N-1}$	η_{1N}
t_2	η_{21}	η_{22}	η_{23}	$\cdots$	$\cdots$	$\eta_{2,N-1}$	η_{2N}
t_3	η_{31}	η_{32}	η_{33}	$\cdots$	$\cdots$	$\eta_{3,N-1}$	η_{3N}

Treatment allocation and observation

	u_1	u_2	u_3	$\cdots$	$\cdots$	u_{N-1}	u_N
t_1	η_{11}	η_{12}	η_{13}	$\cdots$	$\cdots$	$\eta_{1,N-1}$	η_{1N}
t_2	η_{21}	η_{22}	η_{23}	$\cdots$	$\cdots$	$\eta_{2,N-1}$	η_{2N}
t_3	η_{31}	η_{32}	η_{33}	$\cdots$	$\cdots$	$\eta_{3,N-1}$	η_{3N}

From metaphysical to physical array

Define

- $u_{ti} :=$ ith unit to which treatment t is applied
- $Y_{ti} := \eta_{t,u_{ti}}$

t_1	η_{11}	η_{13}	$\cdots$	$\cdots$
t_2	$\cdots$	$\cdots$	$\cdots$	η_{2N}
t_3	η_{32}	$\cdots$	$\cdots$	$\cdots$

$=$

t_1	Y_{11}	Y_{12}	$\cdots$	$\cdots$
t_2	$\cdots$	$\cdots$	$\cdots$	Y_{2n}
t_3	Y_{31}	$\cdots$	$\cdots$	Y_{3n}

Connexions

Any model for the (unobservable) metaphysical array induces a model for any observable physical array

But this map is many-to-one

- because any within-unit dependence of the (η_{tu}) is not reflected in physical array

Flexibility

Metaphysical modelling appears more flexible

- Is this a good thing?

Different metaphysical models may be observationally indistinguishable

- What if we use that flexibility to make different inferences in such cases?

Treatment-Unit Additivity (TUA) ?

$$\eta_{tu} = \alpha_t + \beta_u + \epsilon_{tu}$$

$$\beta_u \sim N(0, \phi_\beta), \quad \epsilon_{tu} \sim N(0, \phi_\epsilon), \quad \text{indep.}$$

Then $Y_{ti} \sim N(\alpha_t, \phi_Y)$, indep., with $\phi_Y = \phi_\beta + \phi_\epsilon$

TUA: $\phi_\epsilon = 0$

- No observable (physical) consequences!
- Hypothesis of TUA untestable

So we should make identical inferences, whether or not TUA assumed (??)

Neyman's null hypothesis

Statistical problems in agricultural experimentation (1935)

H_0: $\overline{\eta}_t$ does not depend on t,

where

$$\overline{\eta}_t := n_0^{-1} \sum_{u \in \mathcal{U}_0} \eta_{tu}$$

– average outcome over all experimental units, if they were all to receive treatment t

Remarks:

- refers to metaphysical array, not to any population parameter
- depends on scale of measurement
- depends on units $\mathcal{U}_0$ used in experiment
- we can never observe both $\overline{\eta}_t$ and $\overline{\eta}_{t'}$ (FPCI)

TUA: $\eta_{tu} = \alpha_t + \beta_u$

With TUA, Neyman's null hypothesis becomes

α_t $[\eta_{tu}]$ does not depend on t

Remarks:

- No longer depends on $\mathcal{U}_0$
- If $\beta_u \overset{\text{i.i.d.}}{\sim} N(0, \phi)$ (say), we have observations

$$Y_{ti} \sim N(\alpha_t, \phi) \quad \text{indep.}$$

- Standard (normal theory or randomisation) analysis

No TUA

Neyman considers the validity of standard ANOVA tests based on expected mean squares under randomisation

Under his own "average null" hypothesis, usual mean squares for treatments and for error need not have the same expectation if we can not assume TUA

- OK for randomised blocks
- but not for Latin square

So: Is the Latin square design biased?

Wilk and Kempthorne (1950s)

W & K extend (and correct) Neyman's analysis

- more complex layouts
- null hypotheses relating to "metaphysical averages" over larger — but still finite — set $\mathcal{U}_1$ of plots
- also possibly over larger set $\mathcal{T}_1$ of treatments, from which experimental treatments $\mathcal{T}$ chosen at random
- analysis depends on $\mathcal{U}_1$, $\mathcal{T}_1$, and varies according as TUA assumed or not

As we let $\mathcal{U}_1$ become infinite in all directions, we recover standard results, and TUA becomes irrelevant

Fisher on Neyman

"... apparent inability to grasp the very simple argument by which the unbiased character of the test of significance might be demonstrated"

"the sole source of his statement ... appears to lie in the peculiarity of his own notation"

Rubin on Fisher

COPSS R. A. Fisher Lecture 2004

"Fisher never used the potential outcomes framework, originally proposed by Neyman in the context of randomized experiments, and as a result he provided generally flawed advice concerning the use of analysis of covariance to adjust for posttreatment concomitants in randomized trials"

Fisher's null hypothesis

"the treatments are wholly without effect"

Metaphysical: Values of (η_{tu}) do not depend on t

Physical: Joint distribution of (Y_{ti}) takes no account of treatment asignments

Cox (1958): Variation between observed treatment means is consistent with sampling from a homogeneous group of observations

All lead back to standard (ANOVA, permutation. . .) tests

Symmetry Modelling (APD 1988)

Develop implications of invariance of attitudes to the data under suitable permutations

- *randomisation*
- *exchangeability*

Independent of scale of measurement, TUA, ...

Group invariance $\Rightarrow$

- Structure for covariances between observations
- Synthetic "population linear model" (PLM)
- Decomposition of data into uncorrelated strata

Null hypothesis = larger symmetry group

One-way layout (fully random)

Physical array (Y_{ti})

t_1	Y_{11}	Y_{12}	$\cdots$	Y_{1n}
t_2	Y_{21}	Y_{22}	$\cdots$	Y_{2n}
t_3	Y_{31}	Y_{32}	$\cdots$	Y_{3n}

Full model: can permute t, and i within each level of t $\quad \widetilde{T}/\widetilde{I}$

Null model: can permute all values arbitrarily $\quad \widetilde{T \cdot I}$

Covariance structure (size-independent)

Full model: $$\operatorname{cov}(Y_{ti}, Y_{t'i'}) = \begin{cases} \gamma_U & (t \neq t') \\ \gamma_T & (t = t', i \neq i') \\ \gamma_0 & (t = t', i = i') \end{cases}$$

Null model:

$$\gamma_T = \gamma_U$$

Population linear model (size-independent)

Full model:

$$Y_{ti} = \underset{(m,\phi_\mu)}{\mu} + \underset{(0,\phi_\alpha)}{\alpha_t} + \underset{(0,\phi_\epsilon)}{\epsilon_{ti}}$$

(all uncorrelated)

Null model:

$$\phi_\alpha = 0$$

Data decomposition (size-dependent)

Full model:

$$Y_{ti} = Y_{..} + (Y_{t.} - Y_{..}) + (Y_{ti} - Y_{t.})$$

Null model:

$$Y_{ti} = Y_{..} + (Y_{ti} - Y_{..})$$

equality of expected mean-squares

$\Rightarrow$ standard F-test

Prediction

Use PLM for prediction/ causal inference/ decision-making:

$$\widetilde{\text{Machine}} * \widetilde{\text{Worker}} \,/\, \widetilde{\text{Run}}$$

$$\underset{}{Y_{mwr}} = \underset{(m,\phi_\mu)}{\mu} + \underset{(0,\phi_\alpha)}{\alpha_m} + \underset{(0,\phi_\beta)}{\beta_w} + \underset{(0,\phi_\gamma)}{\gamma_{mw}} + \underset{(0,\phi_\epsilon)}{\epsilon_{mwr}}$$

Prediction

old machine m, old worker w, new run r

$$\widetilde{\text{Machine}} * \widetilde{\text{Worker}} \,/\, \widetilde{\text{Run}}$$

$$Y_{mwr} = \underset{(m,\phi_\mu)}{\mu} + \underset{(0,\phi_\alpha)}{\alpha_m} + \underset{(0,\phi_\beta)}{\beta_w} + \underset{(0,\phi_\gamma)}{\gamma_{mw}} + \underset{(0,\phi_\epsilon)}{\epsilon_{mwr}}$$

Prediction

old machine m, new worker w

$$\widetilde{\text{Machine}} * \widetilde{\text{Worker}} \,/\, \widetilde{\text{Run}}$$

$$Y_{mwr} = \underset{(m,\phi_\mu)}{\mu} + \underset{(0,\phi_\alpha)}{\alpha_m} + \underset{(0,\phi_\beta)}{\beta_w} + \underset{(0,\phi_\gamma)}{\gamma_{mw}} + \underset{(0,\phi_\epsilon)}{\epsilon_{mwr}}$$

Prediction

new machine m, new worker w

$$\widetilde{\text{Machine}} * \widetilde{\text{Worker}} \,/\, \widetilde{\text{Run}}$$

$$\underset{}{Y_{mwr}} = \underset{(m,\phi_\mu)}{\mu} + \underset{(0,\phi_\alpha)}{\alpha_m} + \underset{(0,\phi_\beta)}{\beta_w} + \underset{(0,\phi_\gamma)}{\gamma_{mw}} + \underset{(0,\phi_\epsilon)}{\epsilon_{mwr}}$$

Extensions

Straightforward extensions to balanced poset/distributive block structure designs with full symmetry

Can apply general approach to problems with restricted symmetry:

- more complex group theory
- invariant subspaces need not be irreducible
- null hypothesis (larger symmetry group) may lead to splitting of a reducible subspace

Mixed model

$$\text{Machine} * \widetilde{\text{Worker}} / \widetilde{\text{Run}}$$

PLM: $Y_{mwr} = (\boldsymbol{\mu})_m + (\boldsymbol{\alpha}_w)_m + \epsilon_{mwr}$

- $\boldsymbol{\mu}$ unstructured: fixed effects
- $\boldsymbol{\alpha}_w \sim (\mathbf{0}, \boldsymbol{\Phi})$, uncorrelated vectors across w
- $\epsilon_{mwr} \sim (0, \phi_m)$, all uncorrelated

Null hypotheses

No systematic differences between workers:

$$\text{Machine} \,/\, \widetilde{\text{Shift}} \,/\, \widetilde{\text{Run}}$$

No differences between workers:

$$\text{Machine} \,/\, \widetilde{\text{Run}}$$

"Fixed" effects are actually "random":

$$\widetilde{\text{Machine}} * \widetilde{\text{Worker}} \,/\, \widetilde{\text{Run}}$$

Relationships between data decompositions determine tests

Conclusion

For causal inference and experimentation:

- Think about decision and prediction
- Think hypothetical (Fisher) rather than counterfactual (Neyman)

For building models and hypotheses:

- Think about structure of attitudes to the data

THANK YOU!