Analyzing the MARCH data accounting for clustering and weighting

Michael R. Elliott

Department of Biostatistics

Institute for Social Research

University of Michigan

mrelliot@umich.edu

INSTITUTE FOR SOCIAL RESEARCH
MICHIGAN PROGRAM IN SURVEY METHODOLOGY
UNIVERSITY OF MICHIGAN

BIOSTATISTICS

Birth Certificates as a Sampling Frame

Use birth certificates as a sampling frame

- All birth certificates in the state include both the location of the birth and the attending physician, if one.
- 98+% of Lower Peninsula MI births are at one of 95 birthing hospitals in the with an attending physician.
  - Birth certificates can be used to construct a sampling frame consisting of clusters of births grouped by hospital and provider.
- Sample hospitals using probability proportional to size

yields (approximately) equal probability of selection:

$$p(i \in S) = p(\alpha \in S)\, p(i \in S \mid \alpha \in S) = \frac{aM_\alpha}{\sum_\alpha M_\alpha}\frac{b}{M_\alpha} = \frac{ab}{\sum_\alpha M_\alpha} = \frac{n}{N}$$

where α indexes the hospital, i the mother in the hospital α, M_α the number of births in hospital α, a and b the number of hospitals sampled and the number of mothers sampled.

Birth Certificates as a Sampling Frame

First and second-stage sampling:

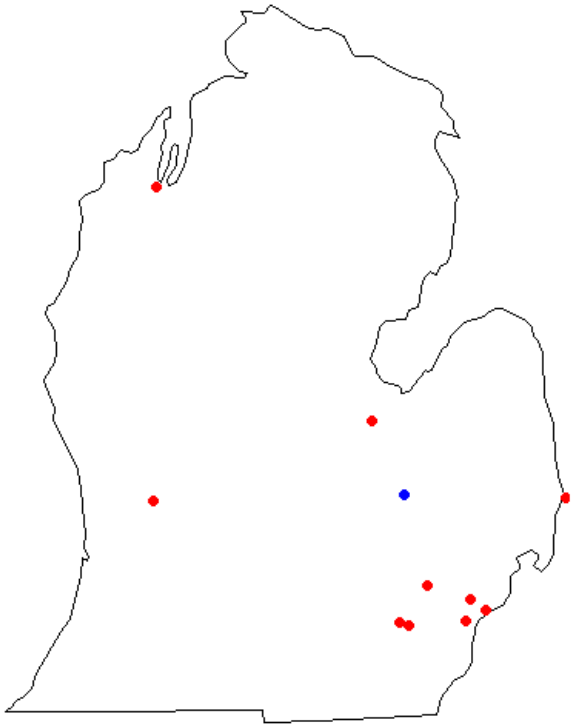
- Based on (guess)timates of available resources and costs, set a=10 and b=100.
- Stratified into 5 strata based on quintiles of Black deliveries.
  - Implicit stratification on urban-rural and to a somewhat lesser extent on socio-economic status.
  - Used birth certificate data from 215,486 live births in 2012-2013.

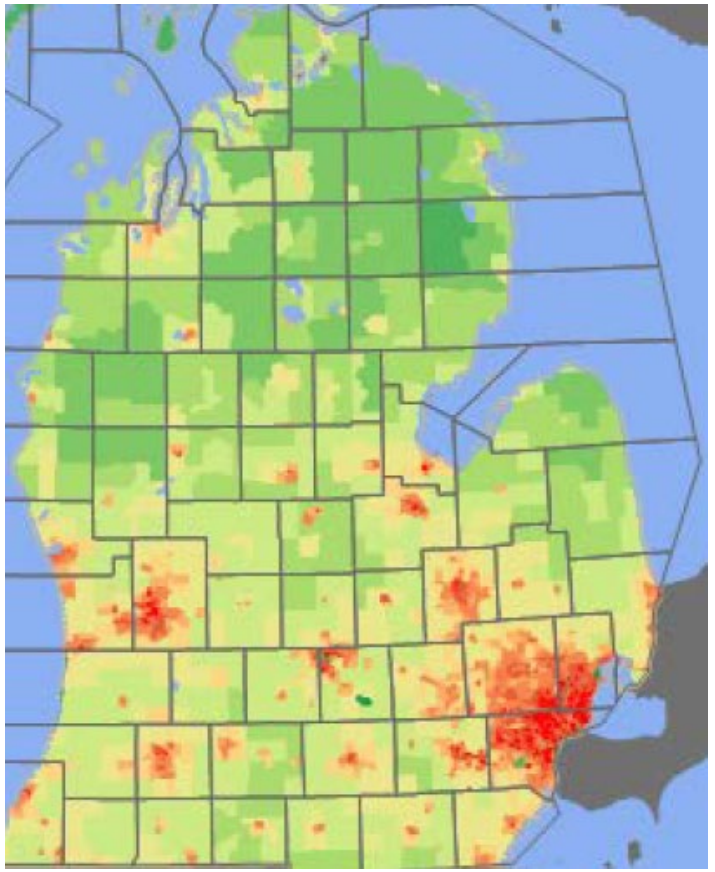
Recruitment actually takes place at provider clinics:

- Once hospitals sampled, attending physicians were grouped into provider clinics and a second stage of PPS sampling was undertaken to obtain 2 practices per hospital, with 50 sampled per clinic.

Also included Hurley Hospital at Flint with certainty: total sample of 1,100

MARCH Hospitals





2010 Population Density

(red=dense, green=sparse)

Recruitment

- Clinic personnel required to introduce the study to patients. If a woman was interested, the researcher then proceeded with the consent process.
  - Some variability in enthusiasm among clinic staff affects response rates.
- All women are recruited at the first prenatal visit.
- To avoid oversampling women who switch providers, we drop and then replace those who leave the sampled clinic.

Data Collection

- Prenatal study visit at first, second (if the gestational age at first visit allowed) and third trimesters,
  - Phone and online surveys.
  - Biological specimens were collected at the clinic, coordinated with routine clinical care:
    - urine samples in all three trimesters
    - blood samples in the first and third trimesters.
- Placenta at childbirth.
- Phone and online surveys at three months, nine months, two years, and annually thereafter.
- In-person child assessments occurred between ages four and five.
- Self-collected biological specimens (urine, stool, blood spot, hair, saliva and maternal saliva and urine).

Response Rate

- All 10 hospitals sampled were recruited. (One had to be replaced after it closed before recruitment began.)
- 7 of 20 clinics refused (RR=65%); 5 were replaced, while recruitment at 2 hospitals used a single clinic (100 each).
- 1,130 women were recruited and gave birth between 12/17 and 8/23.
- Because recruitment occurred at the convenience of the clinic, obtaining a well-defined third-stage response rate is challenging.
  - An upper bound can be established by the 306 individuals who declined to participate but were willing to complete a brief questionnaire about their demographics and reasons for declining.
  - This yields a third-stage response rate of 1130/(1130+306)=78.6%.
- Combining these all three stages yields an overall (upper bound) response rate of 1 x 0.650 x 0.786=51.1%.

Weighting

- In theory, the sample design means that each birth had an equal chance of being recruited. But two eventualities prevent this from occurring in practice:
  - The actual proportion of MI births in a hospital during 2017-2023 doesn't necessarily match the proportion from 2012-2013.
  - Not all recruited women responded.

Weighting

- First-stage hospital-level weighting correction for the 10 PPS sampled hospitals is given by

$$wt_{1i} = c\frac{\%\text{ of 2017-2023 births in hospital } i}{\%\text{ of 2012-2013 births in hospital } i}$$

  where c is a normalizing constant designed so that the sum of the weights adds to 581,485 -- the total number of Lower Peninsula Michigan hospital births between 12/1/17 and 8/31/23 minus the number of births during that time in Hurley hospital.
  - For the Flint hospital first stage weight matched weighted fraction of sample Hurley births to the proportion of Hurley births in the population.
- Weighted sum across all 11 hospitals equals 589,957, the total number of Lower Peninsula Michigan hospital births from December 2017 through August 2023.

Weighting

- Second-stage weight uses birth certificate data to develop raking (calibration) weights.
  - Can adjust for non-response as well as sampling error.
- Raking, or iterative proportional fitting, uses the known marginal population of selected variables to develop weights so that the weighted marginal distributions of these variables match their population marginal distribution.
  - The IPF algorithm begins with the hospital-weighted data so that the design weight information is retained.
- From mother: age, race/ethnicity, education, marital status gestational age, smoking, alcohol, gestational diabetes.
- From child: sex, birthweight, Apgar score.

Results

Variable	% Unweighted	% Selection weighted	% Fully Weighted	% Population
Male	48.2	47.9	51.1	51.1
Low birth weight				
<2500 g	10.2	9.0	1.5	1.5
>=2500 g	90.8	91.0	91.1	91.1
Mother Age (year)				
18-24	19.5	17.7	21.3	21.3
25-34	62.5	63.7	62.1	62.1
35+	17.0	18.6	16.6	16.6
Race				
White	65.6	69.9	72.2	72.2
Black	30.3	25.3	19.2	19.2
Asian	1.2	1.4	2.4	2.4
Other	2.9	3.3	6.0	6.0
Hispanic	3.5	3.4	6.8	6.8
Education				
High School or less	37.6	33.4	36.2	36.2
Some College	29.1	27.7	30.6	30.6
College Degree	33.3	38.9	33.2	33.2
Married	52.8	57.2	59.8	59.8
Gestation Age				
Less than 32 weeks	2.4	2.4	1.6	1.6
32-37 week	22.5	22.1	19.7	19.7
38 weeks or more	75.1	75.	78.7	78.7
Reported smoking	25.8	23.3	11.5	11.5
Reported alcohol use	0.9	0.9	0.8	0.8
Gestational diabetes	7.7	8.0	6.9	6.9
Apgar score less than 8	4.2	4.3	4.6	4.6

Accounting for Complex Sample Design in Analysis

- Standard "Stat 101" approaches for analyzing data do not account for the weights, or the multi-stage sampling of the hospitals rather than sampling of individuals.
  - Weights: "count" those that had lower chance of being selected proportional to the inverse of the probability of being selected.
    - Impacts both point estimation and variance estimation.
  - Clusters: women who give birth at a given hospital may be more alike each other than women at a different hospital.
    - Impacts variance estimation only.

Accounting for Complex Sample Design in Analysis: Weights

- Weighted estimators replace unweighted sums in statistics with weighted sums:

$$\bar{Y} = \sum_{i=1}^{n} w_i y_i \Big/ \sum_{i=1}^{n} w_i \qquad \hat{\beta}^w = \frac{\sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i y_i - \sum_{i=1}^{n} w_i x_i \sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i^2 - \left(\sum_{i=1}^{n} w_i \sum_{i=1}^{n} w_i x_i \right)^2},$$

- Also impacts variance, depending on the association between w_i and y_i
  - No association: variance inflated by $1 + cv_w^2$ over simple random sample assumption
  - Stronger association → less inflation, or even deflation (improved efficiency)
  - In MARCH, $cv_w^2 = \frac{\bar{w}^2}{sd(w)^2} = \frac{522.1^2}{393.8^2} = 1.76$

Accounting for Complex Sample Design in Analysis: Clustering

- Generally *increases variance*, with increase being a function of how similar subjects within a cluster are relative to the whole population ("intra-cluster correlation" or ρ). Assume a clusters with b observations per cluster:
  - Variance is inflated by $1+\rho(b-1)$.
  - ρ =0 → independence within clusters → variance same as SRS.
  - ρ =1 → all observations within identical → variance inflated by factor of b → effective sample size is $n/b = ab/b = a$.
  - For MARCH, $b \approx 103$, so $\rho = 0.01$ would double variance, $\rho = 0.02$ triple variance, etc.
    - Fortunately, most within-hospital correlations are small.

Accounting for Complex Sample Design in Analysis: Examples

- Key variables in dataset:
  - Weight is fwt
  - Cluster in hospital
- Most statistical analysis packages have specialized components that will deal with complex sample design (though not Excel)
  - R: survey package
  - SAS: PROC SURVxxx procedures
  - Stata: svyxxx procedures
  - SPSS: Complex Samples module

Accounting for Complex Sample Design in Analysis: Examples

- Let's look at a (toy) example: marijuana use by mothers.
- Mothers are coded as having used marijuana if they either self-reported use or if cannabinoids were detected in their urine.
  - 15 women are excluded from the analysis due to the lack of self-report and urinalysis data, leaving 1,115 available for analysis.
- Note that the design is focused on births, not mothers: there are 50 twins in the sample and thus 1105 mothers.
- Either focus analysis on births – the proportion of children born with mothers using marijuana, etc. – or use child weights in lieu of mother weights.
- The former is preferred, but the latter won't be off by too much.

Accounting for Complex Sample Design in Analysis: Examples

- The estimated proportion of children born to mothers using cannabis, ignoring the sample design is 24.0% (95% CI 21.5%-26.5%).
- When accounting for the sample design, the estimate is considerably lower: 17.2% (95% CI 14.9%-19.5%).

Accounting for Complex Sample Design in Analysis: Examples

- The estimated proportion of mothers using cannabis, ignoring the sample design is 24.0% (95% CI 21.5%-26.5%).
- When accounting for the sample design, the estimate is considerably lower: 17.2% (95% CI 14.9%-19.5%).

R code:

```
> mydata<-data.frame(Cannabis_Overall,hospital,fwt)
> mydesign<-svydesign(data=mydata, ids=hospital, weights=fwt)
> meancannabis<-svymean(~Cannabis_Overall,design=mydesign,na.rm=TRUE)
> meancannabis
 mean     SE
Cannabis_Overall 0.17185 0.0103
> confint(meancannabis,df=degf(mydesign))
                    2.5 %    97.5 %
Cannabis_Overall 0.1488119 0.1948887
```

Accounting for Complex Sample Design in Analysis: Examples

- The estimated proportion of mothers using cannabis, ignoring the sample design is 24.0% (95% CI 21.5%-26.5%).
- When accounting for the sample design, the estimate is considerably lower: 17.2% (95% CI 14.9%-19.5%).

R code:

```
> mydata<-data.frame(Cannabis_Overall,hospital,fwt)
> mydesign<-svydesign(data=mydata, ids=hospital, weights=fwt)
> meancannabis<-svymean(~Cannabis_Overall,design=mydesign,na.rm=TRUE)
> meancannabis
                  mean     SE
Cannabis_Overall 0.17185 0.0103
> confint(meancannabis,df=degf(mydesign))
                     2.5 %    97.5 %
Cannabis_Overall 0.1488119 0.1948887
```

Accounting for Complex Sample Design in Analysis: Examples

- The estimated proportion of mothers using cannabis, ignoring the sample design is 24.0% (95% CI 21.5%-26.5%).
- When accounting for the sample design, the estimate is considerably lower: 17.2% (95% CI 14.9%-19.5%).

R code:

```
> mydata<-data.frame(Cannabis_Overall,hospital,fwt)
> mydesign<-svydesign(data=mydata, ids=hospital, weights=fwt)
> meancannabis<-svymean(~Cannabis_Overall,design=mydesign,na.rm=TRUE)
> meancannabis
                  mean     SE
Cannabis_Overall 0.17185 0.0103
> confint(meancannabis,df=degf(mydesign))
                     2.5 %    97.5 %
Cannabis_Overall 0.1488119 0.1948887
```

Accounting for Complex Sample Design in Analysis: Examples

SAS code:

```
PROC SURVEYMEANS;
CLUSTER hospital;
VAR Cannabis_Overall;
WEIGHT fwt;
```

Accounting for Complex Sample Design in Analysis: Examples

- The "design effect" of a sample design is given by the ratio of the variance of an estimator assuming a simple random sample and the variance of that estimator that accounts for the design:

$$deff = \frac{v(\bar{y}_w)}{v(\bar{y}_{SRS})}$$

- Easy to compute here: $v(\bar{y}_{SRS}) = \frac{\hat{p}(1-\hat{p})}{n} = \frac{.240 \times .760}{1105} = 1.638 \times 10^{-4}$

  $v(\bar{y}_w) = 1.061 \times 10^{-4}$, so *deff*=0.65

- Equivalently, an effective simple random sample size of 1115/0.65=1722.
- "Superefficiency" may be due to the unintentional oversampling of (tobacco) smokers who have higher rates of marijuana use.
  - Do not expect this to hold for all estimates from MARCH.

Accounting for Complex Sample Design in Analysis: Examples

- Consider logistic regression of marijuana use by mother's education (less than high school, high school only, some college, and college graduate).

Accounting for Complex Sample Design in Analysis: Examples

- R code

```
> mjage<-svyglm(Cannabis_Overall~as.factor(educ),family=binomial,design=mydesign)
Warning message:
In eval(family$initialize) : non-integer #successes in a binomial glm!
> summary(mjage)
Call:
svyglm(formula = Cannabis_Overall ~ as.factor(educ), design = mydesign,
   family = binomial)
Survey design:
svydesign(data = mydata, ids = hospital, weights = fwt)
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -4.4285     0.5442  -8.138 8.17e-05 ***
as.factor(educ)2  2.9381     0.6891   4.263 0.003731 **
as.factor(educ)3  3.5115     0.5637   6.229 0.000433 ***
as.factor(educ)4  3.8326     0.6191   6.191 0.000449 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1.002761)
Number of Fisher Scoring iterations: 7
```

Accounting for Complex Sample Design in Analysis: Examples

```
> confint(mjage,df=degf(mydesign))
                   2.5 %    97.5 %
(Intercept)      -5.715306 -3.141683
as.factor(educ)2  1.308502  4.567655
as.factor(educ)3  2.178497  4.844595
as.factor(educ)4  2.368708  5.296548
```

Accounting for Complex Sample Design in Analysis: Examples

- Less than high school education vs. college graduate: OR=44.2 (95% CI 10.7-199.6)
- High school education vs. college graduate: OR=33.5 (95% CI 8.8-127.0)
- Some college vs. college graduate: OR=18.9 (95% CI 3.7-96.3)
- Treating MARCH as a simple random sample:
  - Less than high school education vs. college graduate: OR=57.3 (95% CI 21.8-150.5)
  - High school education vs. college graduate: OR=39.4 (95% CI 15.8-98.2)
  - Some college vs. college graduate: OR=23.2 (95% CI 9.2-58.2)
- Design effect:
  - Less than high school education 1.47
  - High school education 1.21
  - Some college 1.26

Accounting for Complex Sample Design in Analysis: Examples

SAS code:

```
PROC SURVEYLOGISTIC;
CLASS EDUC;
CLUSTER hospital;
MODEL Cannabis_Overall=EDUC;
WEIGHT fwt;
```

Accounting for Complex Sample Design in Analysis: Degrees of Freedom

- You might remember that when you have a small sample size, you have to account for the degrees of freedom available to estimate the variance and construct confidence intervals
  - *n*-1 for mean: *t*-statistic for sample of 30 means that you multiply the standard error by 2.05 instead of 1.96
- Here we have 1130: t-statistic you multiply the standard error by 1.962 instead of 1.960.

Accounting for Complex Sample Design in Analysis: Degrees of Freedom

- You might remember that when you have a small sample size, you have to account for the degrees of freedom available to estimate the variance and construct confidence intervals
  - n-1 for mean: t-statistic for sample of 30 means that you multiply the standard error by 2.05 instead of 1.96
- ~~Here we have 1130: t-statistic you multiply the standard error by 1.962 instead of 1.960.~~ WRONG!

Accounting for Complex Sample Design in Analysis: Degrees of Freedom

- For clustered designs, you have *a*-1 degrees of freedom to estimate a mean: 9.
- Use 2.262 instead of 1.960 to compute confidence intervals
  - In marijuana use example, standard error is 0.0103, but confidence interval is (0.172-2.262*.0103,0.172+2.262*.0103) or (0.149,0.195), not (0.172-1.960*.0103,0.172+1.960*.0103) or (0.152,0.192)

Accounting for Complex Sample Design in Analysis: Degrees of Freedom

- Situation is worse for regression modes, where the degrees of freedom is *a*-*p*-1, where p is the number of predictors in the model.
  - 10-3-1=6 in the logistic regression model example
- Created an alternative cluster variable that separates hospitals into 2 subsets at random: pseudohospital
  - Use instead of hospital
  - Allows for regression models of up to 15 predictors
  - Somewhat anticonservative: overestimate precision of variance estimates, but properly accounts for overall hospital clustering.

Accounting for Complex Sample Design in Analysis

Questions?

Feel free to contact me at mrelliot@umich.edu