Age-Period-Cohort Analyses of Public Health Data Collected from Independent Serial Cross-Sectional Complex Probability Sample Surveys

Philip J. Smith[1], Zhen Zhao[1], Kirk M. Wolter[2], James A. Singleton[1], J. Pekka Nuorti[1]

Centers for Disease Control and Prevention[1]
National Opinion Research Center and University of Chicago[2]

1. Introduction

Epidemiologic surveillance is the on-going and systematic collection, analysis, and interpretation of health data in the process of describing and monitoring a health event. Data from epidemiologic surveillance is used in planning, implementing, and evaluating public health interventions and programs. Surveillance data are used both to determine the need for public health action and to assess the effectiveness of programs.[1] Within this context surveillance methods based on tracking age, period, and cohort (APC) effects have been popular since their first use in the analysis of tuberculosis mortality rates.[2] Excellent summaries of the statistical literature on APC methods have been published elsewhere.[3,4,5,6,7,8] Each contribution to that literature evaluates APC effects using parametric statistical methods. Implicit in the parametric approach is the assumption that the data obtained represents one random sample among an infinite number of random samples that could have been obtained. This paper describes some of the challenges in evaluating APC effects when data is collected from independent serial cross-section complex probability sampling designs, and when the approach to statistical analysis follows the nonparametric sampling-design based approach to statistical inference[9] available in commercial[10,11] and open-source[12] software packages that account for the sample being drawn from a finite population according to a complex sampling design. In this paper, we let the health event of interest correspond to whether a child is up-to-date (UTD) on receiving recommended vaccines or not, and focus on evaluating how being UTD is affected by cohort effects among children at age *a*.

Keywords: Annual survey; age, period, and cohort effects; stratification, clustering, survey weights.

2. Survey Design-Based Age-Cohort Estimators

Consider a hypothetical surveillance program that consists of serial and independent cross-sectional survey that has been conducted for Y consecutive years, $y=1,\ldots,Y$ for the purpose of estimating the percentage of children who were UTD in each survey year. Let us assume that children who are within the scope of the survey are within a specified age range each survey, and that age range remains the same from year-to-year. Thus, in each annual survey, sampling is restricted to children who belong to specific birth cohorts.

2.1 Estimation using Data from 1 Survey Year

For birth cohorts that have members each of whom has achieved age $\geq a$, let W_{yi} denote the sampling weight for the i^{th} sampled child in year y; $X_{yi}^{(a)}=100$ if the i^{th} sampled child in year y is UTD at age a and =0, otherwise; and $\delta_{yi}^{(c)}=1$ if the i^{th} sampled child in year y belongs to birth cohort c and =0, otherwise. Then, using data from survey year y, an estimator for the percent of children in birth cohort c who are UTD by age a is:

$$\hat{p}_y^{(c,a)}=\sum_i W_{yi}\delta_{yi}^{(c)}X_{yi}^{(a)}\Big/\sum_i W_{yi}\delta_{yi}^{(c)}. \qquad (1)$$

If the sampling weights in (1) are post-stratified so that $\sum_i W_{yi}\delta_{yi}^{(c)}$ equals the number of children in cohort c in year y, then (1) is good estimator of the percentage of children in cohort c who are UTD by age a in the sense that the numerator of (1) estimates the number of children in cohort c who are UTD by age a and the denominator of (1) estimates the number of children in cohort c.

However, let N_{yc} denote the number of children in birth cohort c. To insure that $\sum_i W_{yi}\delta_{yi}^{(c)}$ equals N_{yc}, the survey weights for birth cohort c should be multiplied by $k_{yc}=N_{yc}\Big/\sum_i W_{yi}\delta_{yi}^{(c)}$ In this case,

$$\hat{p}_y^{(c,a)}=\sum_i k_{yc}W_{yi}\delta_{yi}^{(c)}X_{yi}^{(a)}\Big/\sum_i k_{yc}W_{yi}\delta_{yi}^{(c)}$$
$$=\sum_i W_{yi}\delta_{yi}^{(c)}X_{yi}^{(a)}\Big/\sum_i W_{yi}\delta_{yi}^{(c)}.$$

That is, regardless of whether the survey weights are post-stratified, (1) is a good estimate of the percentage of children who are UTD in birth cohort *c*.

2.2 Estimation using Data from All Survey Years

Using data from all survey years, an estimator for the percentage of children in birth cohort c who are UTD by age a is

$$\hat{p}^{(c,a)}=\sum_y\sum_i W_{yi}\delta_{yi}^{(c)}X_{yi}^{(a)}\Big/\sum_y\sum_i W_{yi}\delta_{yi}^{(c)}.$$

One condition that insures $\hat{p}^{(c,a)}$ to be a good estimator of the percentage of children in cohort *c* who are UTD by age *a* is that the survey weights are post-stratified so that $\sum_y \sum_i W_{yi}\delta_{yi}^{(c)}$ is equal to the number of children in cohort *c*. Again, the survey weights could be post-stratified to insure that condition. However, note that

$$\hat{p}^{(c,a)} = \frac{\sum_y \sum_i W_{yi}\delta_{yi}^{(c)}\hat{p}_y^{(c,a)}}{\sum_y \sum_i W_{yi}\delta_{yi}^{(c)}}$$

$$= \frac{\sum_y \left(\sum_i W_{yi}\delta_{yi}^{(c)}\right)\left[\frac{\sum_i W_{yi}\delta_{yi}^{(c)}X_{yi}^{(a)}}{\left(\sum_i W_{yi}\delta_{yi}^{(c)}\right)}\right]}{\sum_y \sum_i W_{yi}\delta_{yi}^{(c)}}$$

$$= \sum_y \omega_y^{(c)}\hat{p}_y^{(c,a)} \qquad (2)$$

where $\omega_y^{(c)} = \sum_i W_{yi}\delta_{yi}^{(c)} \Big/ \sum_y \sum_i W_{yi}\delta_{yi}^{(c)}$.

Specifically, $\hat{p}^{(c,a)}$ is a weighted combination of the annual estimates, $\left\{\hat{p}_y^{(c,a)}\right\}$, of the percentage of children who are UTD in birth cohort *c* by age *a* where the weights $\left\{\omega_y^{(c)}\right\}$ in that combination add to 1. When the sampling probabilities $\left\{1/W_{yi}\right\}$ in each annual survey do not vary greatly from the fraction of children sampled from the target population, $\omega_y^{(c)}$ is approximately equal to the percentage of sampled children in cohort *c* and survey year *y* among all sampled children in cohort *c* across all survey years.

Since $\hat{p}_y^{(c,a)}$ was determined to be a good estimate of the percentage of children who are UTD in birth cohort *c* at age *a* in survey year y, $\hat{p}^{(c,a)}$ is a sensible estimate of the percentage of children who are UTD when data are combined from serial cross-sectional surveys.

2.3 Computer code for $\hat{p}^{(c,a)}$

Estimates for $\hat{p}^{(c,a)}$ can be computed using any commercial or open-source software package that allows survey design-based estimators to be computed. To illustrate this, let "**dsn**" denote the data set name that contains the following variables

- **year**: *y* variable denoting the survey year,
- **stratum:** denotes the strata used by the sampling design (strata may change from year-to-year),
- **psu**: denotes the primary sampling units that are nested within strata (psus may change from year-to-year),
- **wt**: the sampling weight,
- **cohort**: *c* variable denoting the birth cohort,
- **utd_16**: $X_{yi}^{(a)}$ variable denoting whether the person is UTD by 16 months of age (*a*=16),
- **utd_19**: $X_{yi}^{(a)}$ variable denoting whether the person is UTD by 19 months of age (*a*=19).

2.3.1 SAS Code

```
proc surveymeans data=dsn mean stderr;
strata          year stratum ;
cluster         psu;
weight          wt;
class           cohort;
domain          cohort;
var             utd_16 utd_19 ;
run;
```

2.3.2 SUDAAN Code

```
proc descript data=dsn filetype=sas design=wr ;
nest            year stratum psu / psulev=3 ;
weight          wt;
subgroup         cohort;
levels          26;
var             utd_16 utd_19 ;
tables          cohort;
print           mean semean;
run;
```

2.3.3 R Code

```
library( survey )
dsn2<-svydesign ( data = dsn ,
      strata   =~ interaction(year , stratum) ,
      ids      =~psu ,
      nest     = TRUE ,
      weights =~wt )
svyby(~ utd_16 + utd_19,~ cohort,dsn2,svymean)
```

2.4 An Alternative Estimator – And a Discussion

We refer to $\hat{p}^{\left(c,a,\left\{w_y^{(c)}\right\}\right)} = \sum_y w_y^{(c)}\hat{p}_y^{(c,a)}$ as a "composite estimator" for the percentage of children who are UTD in cohort *c* by age *a*, where $\left\{\hat{p}_y^{(c,a)}\right\}$ are the "yearly component estimators" and the $\left\{w_y^{(c)}\right\}$ are weights that add to 1. One of the infinite number of choices for the weights $\left\{w_y^{(c)}\right\}$ is

$$w_y^{(c)} = 1\Big/\hat{v}\left(\hat{p}_y^{(c,a)}\right), \qquad (3)$$

where $\hat{v}\left(\hat{p}_y^{(c,a)}\right)$ denotes the estimated variance of $\hat{p}_y^{(c,a)}$. That choice (3) has the property that it approximately minimizes the variance of $\hat{p}^{\left(c,a,\left\{w_y^{(c)}\right\}\right)}$. The choice (3) puts greater weight on yearly component estimators, $\hat{p}_y^{(c,a)}$, that have the greatest estimated precision.

The choice of weights associated with the estimator (2) does not have this desirable feature. In particular, if the number of children in cohort *c* is not changing appreciably from year-to-year, then the weights $\left\{\omega_y^{(c)}\right\}$ in (2) are approximately equal, and the yearly component estimates receive the same weight, regardless of their estimated precision. Consequently the resulting composite estimate (2) will be less precise, and potentially less accurate. SAS code for the alternative estimator is given in the Appendix.

However, that choice of weights (3) is not without its drawbacks. In particular, this choice excludes the possibility of using commercial or open-source software to conduct multivariate statistical analyses.

2.5 Multivariate Statistical Analyses – By Example

In spite of the potential imprecision that estimates obtained from (2) may have, there is another worth-while convenience available with the estimator (2). Specifically, if analysts decide that it is appropriate to use the weights $\omega_y^{(c)}$ described in Section 2.2, conducting multivariate analyses are straightforward using software packages that account for a complex sampling design and survey weights.

To illustrate these methods we use data from the National Immunization Survey (NIS). Methods used by the NIS are described by Smith et al.[13] Table 1 gives the estimated percentage of children UTD for receiving the 7-valent pneumococcal conjugate vaccine (PCV7) by 16 months of age. A vaccine shortage that began in December 2001 and ended in May 2003, affected the uptake of birth cohorts born after 2000, quarter 4. The low point of the percentage of children UTD by 16 months occurred among the birth cohort born in 2001, quarter 2.

A scientific question is: "Did the shortage affect children who received all vaccine doses from public facilities more than children who received all vaccine doses from all private facilities?" Moreover, "Was the drop in vaccination coverage in children by 16 months of age greater for children who received all vaccine doses from public facilities greater than for children who received all vaccine doses from all private facilities?"

Table 1: Estimated percentages using equation (2) of children UTD for receiving pneumococcal conjugate vaccine by 16 months of age by selected birth cohorts and facility type. National Immunization Survey, 2001-2005.

QUARTERLY BIRTH COHORT YEAR, QUARTER	PERCENTAGE UTD AT 16 MONTHS	
	ALL PUBLIC FACILITIES	ALL PRIVATE FACILITIES
	% (95%CI)	% (95%CI)
1999,4	0.1±0.2	4.6±1.1
2000,1	1.0±0.8	17.1±2.0
2000,2	8.3±3.3	28.1±2.4
2000,3	15.3±4.4	33.3±2.5
2000,4	**16.8±6.3**	**30.9±2.5**
2001,1	14.5±5.2	26.2±2.4
2001,2	**9.1±3.5**	**21.8±2.4**
2001,3	11.9±3.6	22.9±2.2
2001,4	11.3±3.8	23.1±2.1
2002,1	15.3±4.2	33.9±2.9
2002,2	21.8±5.6	39.8±3.1
2002,3	32.5±6.7	42.9±3.2
2002,4	39.8±8.7	43.5±3.5
2003,1	15.2±8.0	20.8±2.9
2003,2	9.1±4.7	14.7±2.9
2003,3	20.9±7.8	31.3±3.7
2003,4	30.0±9.3	41.7±4.8

Note that the difference in the decline in vaccination coverage (between children born in 2000 quarter 4 to children born in 2001 quarter 2) between children who receive all vaccine doses from public providers and those who receive all doses from all private providers is an interaction term in the regression of vaccination status on birth cohort, facility type and the interaction between cohort and facility.

In particular, letting children born in 2000, quarter 4 (level 5 of **cohort**) and who received all doses at public facilities (level 1 of **facility**) be the reference category for this analysis, the following SUDAAN code conducts the relevant analysis.

```
proc regress data=dsn filetype=sas design=wr ;
nest        year stratum psu / psulev=3 ;
weight      wt ;
subgroup  cohort  facility;
levels      26    2     ;
reflevel    cohort = 5          facility =1 ;
model       utd_16=cohort  facility  cohort* facility;
run;
```

The relevant estimated interaction term from this regression is -1.4% (±8.0%), with a two-sided p-value of 0.72. Evaluating the significance of that interaction term is identical to conducting the significance test that compares the decline in vaccination coverage for children receiving all doses from public facilities (7.7%=**16.8%-9.1%**) to the decline in vaccination coverage for children receiving all doses from private facilities (9.1%=**30.9%-21.8%**): the difference of those differences is -1.4%. Also, the standard errors and p-values from these two analyses are identical, because the computations in these two analyses are identical.

It is sometimes thought that the use of the SUDAAN **proc regress** procedure to analyze binary data violates the "normality assumption" and an assumption about the homoscedasticity of variance for the regression. However, in survey design approach to the analysis of data from complex probability sample surveys, no parametric distributional assumptions are made.[9, 14] Analysis of complex survey data using the survey design-based approach provided by commercial and open-source software is non-parametric. For example, in the SUDAAN **proc regress** analysis, no distributional assumptions are made beyond those specified directly in the **proc regress** code, namely that the data were drawn randomly with replacement from a finite population with probabilities equal to the inverse of the variable listed on the **weight** variable, and with restrictions on the randomization distribution defined by the stratification and primary sampling units listed on the **nest** statement that randomized units from the finite population as either into the sample or not in the sample. The estimated finite population regression coefficients from **proc regress** are estimated percentage differences from the reference category that have estimated standard errors that are correct with respect to those non-parametric assumptions.

The **proc regress** code listed in this section is a multivariate analysis. If analysts want to control for variation in the **utd_16** dependent variable attributable to other factors in this analysis, those variables would be specified in the **subgroup** and **reflevel** statements (if they are categorical predictors) and as independent variables in the **model** statement.

3. Conclusions and Discussion

As suggested in Section 2.3, there is an infinite number of ways to obtain prevalence estimates that depend on age and cohort effects. In this paper we have presented 2 methods that are useful. If an analyst wishes to report point estimates only, then the method described in Section 2.4 may be most appropriate because it yields estimates that approximately have minimum variance. On the other hand, analysts who want to conduct multivariate analyses may use the methods described in Sections 2.3 and 2.5. When analyses are conducted with those methods, it is essential to know the extent to which the estimates obtained using those methods differ from those obtained using the methods in Section 2.4. When membership in birth cohorts are defined by wide ranges of birth dates, the effect of sudden and strong period effects on those birth cohorts may be attenuated because of the lack of resolution caused by the width of ranges of birth dates. By using narrower ranges, resolution and sensitivity may be increased provided the sample size is sufficiently large to yield precise estimates. Age-period and cohort-period methods require other special considerations when data is obtained from independent serial cross-sectional surveys. In subsequent work, we will describe how those analyses can be conducted.

Appendix: SAS code for the Alternative Estimator

```
/* Components for the composite estimate. */
proc surveymeans data = dsn nobs mean stderr;
class      cohort   year;
domain   cohort * year;
var        utd4_16 ;
cluster    psu;
strata     year stratum ;
weight     wt;
ods output statistics=stat
       domain=mystat2; run;
data raw ;
        set mystat2;
        mean              = mean;
        precision         = 1 / stderr**2;
        keep  cohort year mean precision; run;
/* For each birth cohort, sum the precisions across
survey years to obtain totals (tot). */
proc sort data = raw; by cohort; run;
proc means data = raw sum noprint;
        var      precision;
```

```
        by      cohort;
        output  out = totals sum=tot;
        run;
/*    Merge those totals into each birth cohort. */
proc sort data = totals; by cohort; run;
proc sort data = raw;  ; by cohort; run;
data raw2;
        merge  raw totals;
        by      cohort;
        run;
/* Compute the relative precision,  and the
    contributions to the composite estimates
    of the mean and variance */
data relative;
     set raw2;
     relative     = precision  /  tot;
    mean_part = relative * mean;
    variance     = 1 / precision;
    var_part     = ( relative ** 2 ) *  variance;
    run;
/* Add the components  of the composite estimate
   of the mean */
proc sort data = relative; by cohort; run;
proc means data = relative sum noprint;
        var mean_part;
        by cohort;
        output out = means  sum=mean;
        run;
data means;
        set means;
        keep cohort mean;
        run;
/* Add the components  of the composite estimate
   of the variance. */
proc means data = relative sum noprint;
        var var_part;
        by cohort;
        output out = variance sum=var;
        run;
/* Compute the standard errors  and
    95% ci half widths */
data ci;
        set variance;
        ci = 1.96 * sqrt(var);
        keep cohort  ci;
        run;
/* Merge the means and 95% ci half widths. */
proc sort data = means; by cohort; run;
proc sort data = ci      ; by cohort; run;
data birth_cohort;
        merge means (in=f1) ci (in=f2);
        by      cohort;
        if      f1 and f2;
        run;
/*  Produce a readable report. */
data birth_cohort2;
set  birth_cohort;
pct = put( round( mean , .1 ) , 5.1 );
temp  = put( round( ci, .1) , 5.1 );
plsmin   = 'b1'x;
ci_95 = put( compress( plsmin || temp ) , 5.1 );
utd= pct ||  ci_95 ;
if pct eq . then utd = "";
keep      cohort    utd ; run;
proc print data = birth_cohort2;  run;
```

References

[1] Centers for Disease Control and Prevention. Guidelines for evaluating surveillance systems. *Morb Mort Wkly Rep* 1988;37:1-18.

[2] Frost WH. The age selection of mortality from tuberculosis in successive decades. *Am J Hyg* 1939;30:91-96.

[3] Fienberg SE, Mason WM. Identification and estimation of age-period-cohort models in the analysis of discrete archival data. In Sociological Methodology 1979, Schuessler KF (ed.). Jossey-Bass, Inc: San Francisco, 1978; 1–67.

[4] Holford TR. The estimation of age, period and cohort effects for vital rates. *Biometrics* 1983; 39:311– 324

[5] Kupper LL, Janis JM, Salama IA, Yoshizawa CN, Greenberg BG. Age-period-cohort analysis: an illustration of the problems in assessing interaction in one observation per cell data. *Comm Stat—Theory Meth* 1983; 12:2779–2807.

[6] Holford TR. Age-period-cohort analysis. In Encyclopedia of Biostatistics, Armitage P, Colton T (eds). Wiley: Chichester, 1998; 82–99.

[7] Clayton D, Schiffers E. Models for temporal variation in cancer rates. I: age-period and age-cohort models. *Stat Med* 1987; 6:449– 467.

[8] Smith PJ. Power and sample size considerations for detecting deviations from trends in surveillance surveys. *J Royal Stat Soc, D: The Statistician.* 1997;46(3), 423-432.

[9] Cochran WG. Sampling Techniques – Third Edition. John Wiley and Sons, New York. 1977.

[10] SAS Institute Inc. SAS/STAT® Users Guide, Version 9, Cary, NC: SAS Institute Inc., 2006.

[11] Research Triangle Institute. SUDAAN User's Manual, Release 9.0. Research Triangle Park, NC: Research Triangle Institute, 2006.

[12] Lumley T. Analysis of complex survey samples. *J Stat Software* 2000;9(1):1-19.

[13] Smith PJ, Hoaglin DC, Battaglia MP, et al. Statistical Methodology of the National Immunization Survey: 1994-2002. National Center for Health Statistics. *Vital Health Stat Series 2*, 2005,2(138).

[14] Binder DA. On the Variances of Asymptotically Normal Estimators from Complex Surveys. *Int Stat Rev* 1983;51(3);279-292.