ESTIMATION PROCEDURES USED TO PRODUCE WEEKLY FLU STATISTICS FROM THE HEALTH INTERVIEW SURVEY

James T. Massey, Gail S. Poe, Walt R. Simmons
National Center for Health Statistics

1. INTRODUCTION

In April 1976, the United States Congress appropriated $135 million for a national immunization program against the A/New Jersey or "Swine Flu." The Center for Disease Control (CDC) was charged with the responsibility of developing a comprehensive immunization delivery system and with the assessment of the coverage of the vaccination program, as well as the surveillance of flu cases. Missing from CDC's surveillance systems was a system through which national estimates could be made from a national probability sample or a full census. Although CDC's basic systems could provide partial information for the entire country, they could not provide sufficient data for production of estimates that could be assessed for precision. CDC therefore, requested the National Center for Health Statistics (NCHS) to collect influenza activity data in the Health Interview Survey (HIS).

In the Health Interview Survey a probability sample of households representing the civilian, noninstitutionalized U.S. population is interviewed each week. Interviewing is done continuously on a weekly sample of about 800 households. In response to CDC's need a supplemental set of questions on influenza and influenza vaccinations was added to the regular HIS interview questionnaire in the last week of September 1976.

In the regular HIS processing procedures, the time between the data collection and publication of the results is generally at least one year. Because of the demand for timely data on influenza cases and vaccinations the HIS implemented a rapid reporting system in which estimates of influenza-like illnesses; bed days due to such illnesses; and all types of influenza, including swine flu, vaccinations were published weekly three weeks after the week for which the estimates were made and only one week after the data were collected.

The HIS sample is designed so that tabulations can be provided for each of four major geographic regions, for large metropolitan areas, and for urban and rural sectors of the United States. The sample is also designed so that households interviewed each week represent those in the target population and that the weekly samples are additive over time. A rapid reporting system was used one other time in the history of the HIS, and that was during the 1957-58 influenza epidemic. At that time, weekly reports also were issued.

The weekly reports for 1977 were continued through April, and the estimates presented were provisional. Final estimates will be published after several months of extensive data processing in which medical coding is completed and many error and consistency checks are made on the data. The HIS weekly estimates were not part of CDC's systems for detecting early outbreaks of influenza. Because of the national scope of the data, local outbreaks of influenza-like illness possibly were undetected. However, when used in conjunction with other sources of information within the CDC surveillance system, the HIS data could confirm or deny early inferences regarding the spread of this disease and its effect.

2. STATISTICAL METHODS

Several estimation procedures were considered by NCHS for estimating the weekly number of flu cases, the number of bed-days due to flu, and the number of all types of flu shots and swine flu shots. The two most prominent estimators are described below along with some of their properties.

Since the HIS uses a two-week reference period to collect data on the incidence of acute conditions a two-week reference period was also chosen for the influenza supplement. That is, during each week of interviewing a flu case is enumerated if its onset occurred during the two weeks preceding the interview week, a bed-day is enumerated if it occurred during the two weeks prior to the interview week, and a flu shot is enumerated if it were received in the two weeks prior to the interview week. Thus, for each week i of interest the following two independent estimates can be made for the number, say X_i, of flu cases, bed-days, or flu shots.

α_i' - the estimate for the "last week" obtained from interview week (i+1) and,

β_i' - the estimate for "week before" obtained from interview week (i+2).

The first estimator of X_i considered was used during the 1957-58 flu epidemic to estimate the incidence of acute upper respiratory conditions and is given by

$$X_i'' = \frac{1}{2}(\alpha_i' + \beta_i').$$

The estimator, X_i'', is unbiased and, if one assumes that the variance of X_i'' is constant from week to week, then the variance of X_i'' is given by

$$\sigma^2_{X_i''} = \frac{1}{2}\sigma^2_{\alpha_i'} .$$

The second estimator to be considered is given by

$$X_i' = \frac{1}{4}\left(\beta_{i-1}' + \alpha_i' + \beta_i' + \alpha_{i+1}'\right)$$

$$= \frac{1}{2}\left(U_i' + U_{i+1}'\right)$$

where $U_i' = \frac{1}{2}\left(\beta_{i-1}' + \alpha_i'\right)$ is the average weekly estimate obtained from interview week (i+1).

The estimator X_i' is a weighted average of four weekly estimates obtained from interview weeks (i+1) and (i+2). Since the estimator contains information from the week on either side of the week of interest a smoothing effect results. The expected value of X_i' is given by

$$E(X_i') = X_i + \frac{1}{4}\left(X_{i-1} + X_{i+1} - 2X_i\right).$$

The bias of X_i' is given by the second term on the right hand side of the above equation. In most situations (especially if a linear trend is present) this bias will be small. The only time when this bias might be more than a few percentage points is at the maxima or minima points of a trend.

Again, assuming that the variance of the weekly statistics remains constant from week to week, the variance of X_i' can be expressed as

$$\sigma_{X_i'}^{2} = \frac{1}{16}\left[4\sigma_{\alpha_i'}^{2} + 4\ \mathrm{COV}\left(\beta_{i-1}', \alpha_i'\right)\right]$$

$$= \frac{1}{4}\sigma_{\alpha_i'}^{2}\left[1 + r_{\beta_{i-1}',\alpha_i'}\right]$$

where $r_{\beta_{i-1}',\alpha_i'}$ is the correlation between the incidence of "last week" and the "week before" from a single week's sample. For most acute conditions the correlation is assumed to be small, although the correlation will be higher for very contagious diseases. The assumption of equal weekly variances should hold unless X_i changes considerably from week to week. It should also be noted that the weekly statistics U_i' and U_{i+1}' are independent since they are obtained from independent weekly samples.

Comparing the variances of X_i'' and X_i',

$$\sigma_{X_i'}^{2} < \sigma_{X_i''}^{2} \quad \text{for } r_{\beta_{i-1}',\alpha_i'} < 1.$$

Using the data on the number of flu cases for the 1975-76 flu season (the last quarter of 1975 plus the first quarter of 1976) the correlation coefficient, $r_{\beta_{i-1}',\alpha_i'}$, was estimated to be approximately 0.75. Thus, for flu cases the variance of X_i' is approximately 13 percent smaller than the variance of X_i''. Based on a mean squared error criteria there is little to choose between X_i' and X_i''.

Another important feature of the estimator X_i', however, is that it can be formed using the two-week average estimates U_i' and U_{i+1}' and doesn't require the formation of two separate weekly estimates for each week of interviewing. Operationally, this feature reduces the number of weekly tabulations in half. For this reason the estimator X_i' was selected for making our weekly estimates.

The difference between incidence for adjacent weeks is estimated by

$$d_i' = X_i' - X_{i-1}'$$

and the variance of d_i' can be shown to be

$$\sigma_{d_i'}^{2} = \frac{1}{4}\sigma_{\alpha_i'}^{2}\left[1 + r_{\beta_{i-1}',\alpha_i'}\right]$$

$$= \sigma_{X_i'}^{2}.$$

The variance of $\sigma_{d_i'}^{2}$ can also be shown to be equal to $2\sigma_{X_i'}^{2}\left(1 - r_{X_i', X_{i-1}'}\right)$ and, thus, $r_{X_i', X_{i-1}'} = \frac{1}{2}$. This correlation is intuitively obvious since U_i' is used to form one half of both the estimators X_i' and X_{i-1}'.

The weekly estimates can be summed to form aggregates such that for N weeks

$$X_s' = X_1' + X_2' + \ldots\ldots + X_N'.$$

The variance of X_s', assuming equal weekly variances, is given by

$$\sigma_{X_s'}^{2} = N\,\sigma_{X_i'}^{2} + 2\,r_{X_1'X_2'}\,\sigma_{X_i'}^{2} + 2\,r_{X_2'X_3'}\,\sigma_{X_i'}^{2} + \ldots + 2\,r_{X_{N-1}'X_N'}\,\sigma_{X_i'}^{2}$$

$$= (2N-1)\,\sigma_{X_i'}^{2}.$$

The relative standard error of X_s' can be written as

$$V_{X_s'} = \frac{\sqrt{(2N-1)}\,\sigma_{X_i'}}{X_s'}$$

$$\doteq \sqrt{\frac{2N-1}{N^2}}\ V_{X_i'}.$$

For NCHS's weekly publications $V_{X_s'}$ is further approximated by

$$\sqrt{2/N}\ V_{X_i'}$$

where $V_{X_i'}$ is the relative standard error of X_i'

and is defined as $\sigma_{X_i'}/X_i'$.

2.1 Estimating Sampling Variance

There are several alternative methods for approximating the sampling variance of X_i' and three such methods which are described below were compared.

If the weekly statistics are reasonably stable from week to week with no apparent trend, then a simple estimate of $\sigma_{X_i'}{}^2$ can be made using any two consecutive weekly estimates of X_i.

For a single week i, the estimated sampling variance is given by

$$s_{X_i'}{}^2 = \frac{1}{4}(U_i' - U_{i+1}')^2$$

and the relative standard error is given by

$$v_{X_i'} = \frac{U_i' - U_{i+1}'}{U_i' + U_{i+1}'} .$$

By summing over k weeks a more stable estimate of $\sigma_{X_i'}{}^2$ can be obtained such that

$$s^2_{X_i'} = \frac{1}{4(k-1)} \sum_{j=1}^{k-1} (U_j' - U_{j+1}')^2$$

and

$$v^2_{X_i'} = \frac{s^2_{X_i'}}{\overline{U}}$$

where

$$\overline{U} = \frac{1}{k}\sum_{j=1}^{k} U_j' .$$

A second method of approximating $\sigma_{X_i'}{}^2$ uses least squares regression to fit three consequtive values of U_i' and looks at the deviation about the regression line to estimate $\sigma_{X_i'}{}^2$. A more stable estimate is again obtained by averaging a series of estimates. This method is satisfactory for linear trends but tends to over estimate the sampling variance when there are changes in direction in the trend. The approximation is given by

$$s^2_{U_i'} = \frac{1}{6(k-2)} \sum_{j=2}^{k-1} (U_{j-1} - 2U_j + U_{j+1})^2$$

and

$$s^2_{X_i'} = \frac{1}{2} s^2_{U_i'} .$$

Yet another method of estimating the variance of the weekly statistics is to compute a simple random sample variance estimate and inflate by a design factor. The design factor represents the increase or decrease in precision due to deviations from a simple random sample design such as stratification and clustering. For the HIS the design factor has been shown to be around two. Using this assumption estimates of variances were calculated for the number of flu cases and compared to the estimates obtained using the second method of approximation presented above. The results are presented below in terms of percent relative standard error (PRSE) which is the standard error of an estimate divided by the estimate itself multiplied by 100.

Size of Estimate (In thousands)	Method 3 (Simple Random Sample)	Method 2 (Regression)
1500	19	21
2000	16	18
2500	14	15
3000	13	13
4000	11	10
5000	10	9
6000	9	8

The overall average values of the percent relative standard errors for the number of flu cases and bed days shown in NCHS's weekly flu publication were obtained using data from the 1975 flu season (September to March) and then reverified using the first 14 weeks of the 1976 flu season. The PRSE for the weekly estimate of flu shots was not available in 1975 and was approximated using the 1976 data. The first two methods presented in this section were used to estimate the PRSE's and the methods were found to be quite comparable. The average weekly PRSE is about 16 for flu cases and 20 for bed days and flu shots.

2.2 Weighting and Post-Stratification

Up to this point it has been assumed that the average weekly estimates, U_i', have already been calculated. The first step in the weekly estimation procedure is to calculate U_i'. This is done by weighting the weekly sample data. Except for minor adjustments due to nonresponse and subsampling the HIS sample is self-weighting (each sample person has the same probability of selection in the national sample). For weekly estimation each sample person is assigned the same probability of selection. One final post-stratification adjustment is required, however, to adjust each week's sample to the same national population. Since each week's sample is a random sample, the distribution of sample persons will vary from week to week by age and race and an adjustment to the population distribution will improve the precision of the weekly estimates. The population distribution is obtained from the Bureau of the Census and adjustments are made each week for ten age-race groups. If

y_{jk} = total number of sample persons in the jkth age-race cell reported during interview week (i+1),

z_{jk} = total number of flu cases, bed days, or flu shots in the jkth age-race cell reported during interview week (i+1), and

Y_{jk} = population control (Census value) for jk^{th} age-race cell for week i,

then the average weekly estimate U_i' obtained from interview week (i+1) is given by

$$U_i' = \frac{1}{2} \sum_{jk} z_{jk} Y_{jk}/y_{jk}.$$

The U_i' are then used to calculate the X_i's and their sampling errors.

3. RESULTS

Figure 1 below shows the weekly estimates of flu-like illness for September 20, 1976 through April 17, 1977 and Figure 2 contains the estimates of bed days due to flu for the same time period. The curves are similar to ones for the last several years. The relative smoothness of the curves further indicates the stability of the estimators which were employed in the rapid reporting system. The table below gives the actual weekly estimates for the variables of interest.

4. CONCLUSION

Although there was, fortunately, no epidemic of Swine Flu this year we felt that the HIS rapid reporting system was a success. Reliable weekly estimates were published only one week after the data were collected and only three weeks after the reference week. It would have been impossible to implement such a system after the detection of an epidemic because of the rapidity with which such a virus spreads throughout the entire country.

The influenza supplement will provide health data analysts and planners with extensive information on the correlates of influenza. Among the many areas that may be examined are the relationships among other health and demographic characteristics (obtained in the main interview) and influenza, as well as the effect of influenza symptoms on limitation of usual activity (such as work loss). Additionally, the characteristics of persons who obtained flu vaccinations as opposed to those who did not and the timing of the vaccination in relation to the contraction of an upper respiratory illness may be studied.

REFERENCES

Poe, Gail S. and Massey, James T., "Estimating Influenza Cases and Vaccinations by Means of Weekly Rapid Reporting System," Public Health Reports, 92 No. 4 (1977), 299-306.

Figure 1. Weekly Estimates of Flu-Like Illnesses in the United States

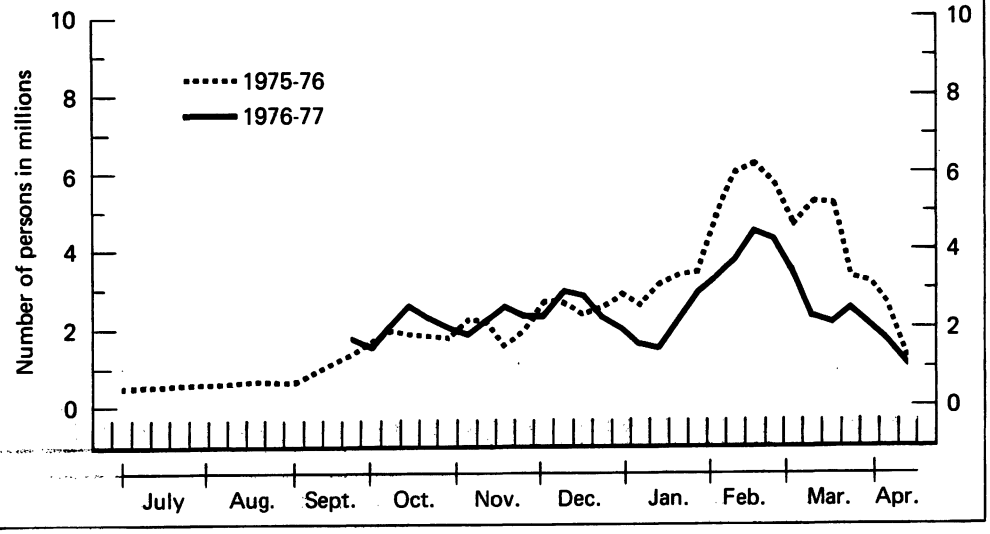
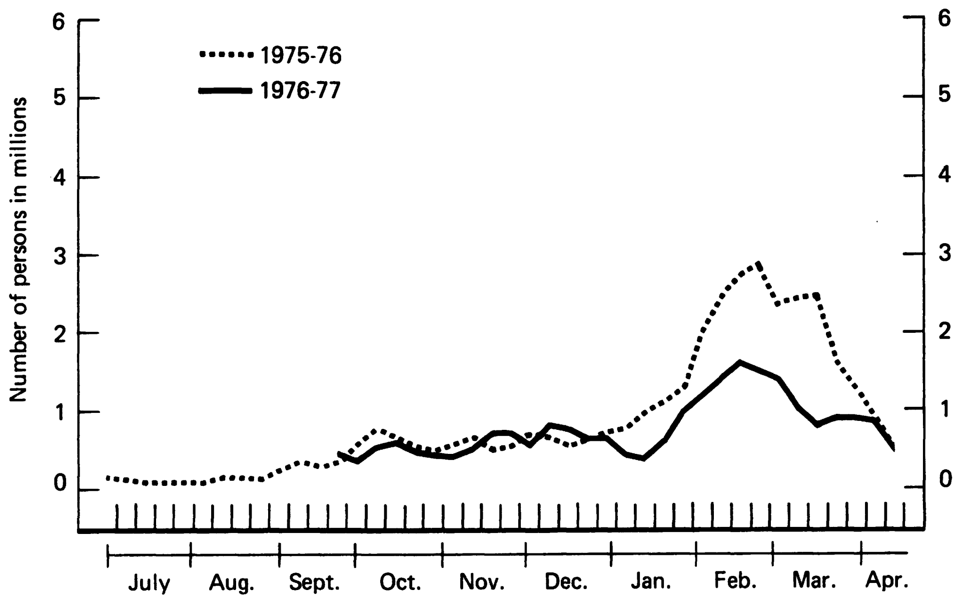


Figure 2. Weekly Estimates of Average Number of Persons in Bed Each Day Because of Flu-Like Illness in the United States



Weekly Estimates of Flu-Like Illnesses, Average Number of Persons in Bed Each Day, and Flu Shots: United States, 1976-77

Week	Flu-like illness	Average number of persons in bed each day because of flu-like illness	All types of flu shots		Swine flu shots	
			Each week	Cumulative since September 20	Each week	Cumulative since September 20
	Number in thousands					
September 20-26, 1976	1,710	419	638	638	*	*
September 27-October 3, 1976	1,490	369	1,130	1,768	*	*
October 4-10, 1976	2,115	517	1,846	3,614	1,200	1,668
October 11-17, 1976	2,567	591	3,196	6,810	2,378	4,046
October 18-24, 1976	2,239	489	4,737	11,547	4,014	8,060
October 25-31, 1976	2,062	413	5,122	16,669	4,634	12,694
November 1-7, 1976	1,880	410	5,580	22,249	5,019	17,713
November 8-14, 1976	2,319	511	6,749	28,998	6,391	24,104
November 15-21, 1976	2,493	704	5,379	34,377	5,154	29,258
November 22-28, 1976	2,277	700	4,101	38,478	3,921	33,179
November 29-December 5, 1976	2,276	573	4,128	42,606	3,972	37,151
December 6-12, 1976	2,857	789	2,616	45,222	2,361	39,512
December 13-19, 1976	2,831	753	1,235	46,457	1,063	40,575
December 20-26, 1976	2,207	631				
December 27,1976-January 2,1977	2,023	636				
January 3-9, 1977	1,646	465				
January 10-16, 1977	1,457	411				
January 17-23, 1977	2,127	562				
January 24-30, 1977	2,938	945				
January 31-February 6, 1977	3,242	1,205				
February 7-13, 1977	3,766	1,381				
February 14-20, 1977	4,540	1,597				
February 21-27, 1977	4,333	1,517				
February 28-March 6, 1977	3,290	1,299				
March 7-13, 1977	2,308	969				
March 14-20, 1977	2,190	778				
March 21-27, 1977	2,525	882				
March 28-April 3, 1977	2,123	896				
April 4-10, 1977	1,614	772				
April 11-17, 1977	1,138	516				

*Figure does not meet standards of precision.

NOTE: Even though the suspension of the Public Health Service immunization program was lifted on February 7, 1977, estimates of flu shots are not shown after the week ending December 19, 1976.