Three Statistical Models for Estimating Length of Stay

By Steve Selvin

The probability density functions implied by three methods of collecting data on the length of stay in an institution are derived. The expected values associated with these density functions are used to calculate unbiased estimates of the expected length of stay. Two of the methods require an assumption about the form of the underlying distribution of length of stay; the third method does not. The three methods are illustrated with hypothetical data exhibiting the Poisson distribution, and the third (distribution-independent) method is used to estimate the length of stay in a skilled nursing facility and in an intermediate care facility for patients enrolled in California's MediCal program.

A fundamental measure used in health services research is the mean length of stay for a defined set of patients in a specific institution. In many cases the mean length of stay is employed to compare different health facilities or changes within a single facility and plays a central role in the evaluation of utilization.

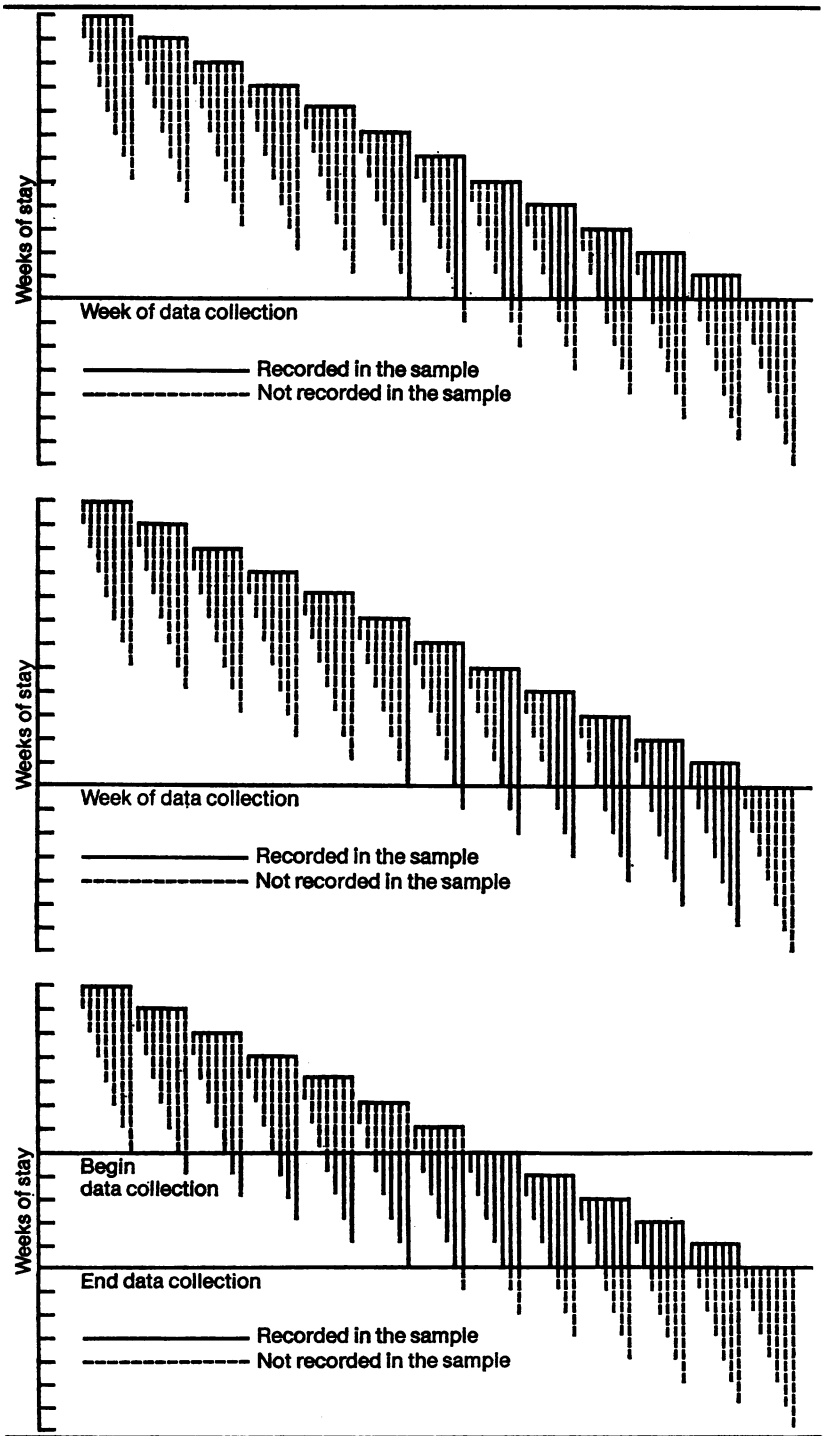
The ideal estimate of the expected length of stay is made by selecting a set of patients from those newly admitted to an institution and then following these patients until each of them is discharged (i.e., complete follow-up). The amount of time these patients remain in the institution divided by the total number of those who are followed is an estimate of the expected length of stay. However, this type of complete follow-up is expensive, time-consuming, and not always possible, especially in long-term institutions. If certain statistical assumptions are realistic, an estimate of mean length of stay can be made from more efficient sampling patterns that do not involve following each patient to discharge. This topic has been dealt with by other authors from a variety of points of view [1–4].

The purpose of this article is to examine three sampling patterns for collecting length-of-stay data and to show how, if certain assumptions are met, unbiased estimates of expected mean lengths of stay can be derived from the data so collected. With two of these sampling patterns, deriving unbiased estimates requires an additional assumption about the form of the underlying theoretical distribution of lengths of stay; with the third, however—the interval pattern, to be described presently—no such assumption is necessary.

The research reported herein was supported by grant no. 003-P30-74/01 from the Department of Health, Education, and Welfare. The opinions and conclusions expressed herein are solely those of the author and should not be construed as representing the opinions or policy of any agency of the U.S. government.

Address communications and requests for reprints to Steve Selvin, Associate Professor, Department of Biomedical and Environmental Health Sciences, School of Public Health, University of California, Berkeley, CA 94720.

Schematic representation of data collected by three sampling patterns: top, retrospective pattern; center, partial follow-up pattern; bottom, interval pattern. Solid lines indicate portions of stays included in sample.



The three sampling patterns are the retrospective, partial follow-up, and interval patterns. A retrospective sample is collected by identifying a series of patients at a specific time and ascertaining the number of completed weeks each patient has been institutionalized. The portion of each patient's stay that would be included in data thus collected is shown schematically in the top segment of the accompanying figure. (The time unit could be days, weeks, months, or any other measure, but for simplicity it will be referred to here as weeks.) Partial follow-up sampling involves identifying a series of patients at a specific time and observing each patient's entire stay from admission to discharge, as shown schematically in the center segment of the figure. Interval sampling consists of identifying a number of patients at a specific time and following these and any newly admitted patients over a specified period. At the end of this interval, some patients will have been discharged and others will still be institutionalized, as shown in the bottom segment of the figure.

Data collected by any of these three methods are "length biased." Retrospective and interval samples contain an excess of patients recorded for short stays; a partial follow-up sample, on the other hand, contains an excess of patients with long stays. The incomplete nature of the data resulting from these three sampling methods is the reason for the length bias in any direct computation of the mean length of stay. This type of bias is encountered elsewhere. For example, length bias often occurs when data are collected to estimate the duration of disease.

The Models

If two assumptions are met, an unbiased estimate of expected length of stay may be obtained from any of the three sampling methods; these assumptions are (1) that the probability distribution of lengths of stay is discrete (i.e., each patient's stay is an integral number of time units) and (2) that the probability distribution of lengths of stay is the same for each cohort entering the institution. That is, the distribution associated with the institutionalized population is assumed to be stationary with respect to time.

The data obtained by the three methods can be related to their respective probability density functions, which permit estimation of the expected stay without length bias. Let x denote the number of complete weeks of stay recorded for each patient, with a mean for all patients $\bar{x}$; X is a random variable from the distribution of which x is sampled and which has an associated theoretical probability density function $f(x)$. The variable X is the number of whole weeks of a patient's total stay. If a patient is discharged during the fourth week, $X = 3$ for each of the sampling schemes.

Retrospective Sample

For retrospectively collected data, the theoretical density function $f(x)$ reflects the probability $P(X \geq x)$ that a patient's actual stay is greater than or equal to x weeks; EX is the theoretical expected length

of stay, free of length bias, associated with the (as yet unspecified) density function $f(x)$.

In collecting a retrospective sample, the number of weeks since admission is taken from admission records for all patients present at the time of sampling, and the longest stay is k whole weeks; thus $x = 0, \ldots, k$. For each value of x the number of patients discharged during week $x + 1$ is recorded, which allows derivation of the sample density, $f_r(x)$, in terms of the theoretical density function $f(x)$. For each value of x, $P(X \geqslant x)$ is related to $f(x)$ as follows [5,6]:

$$\begin{aligned} P(X \geqslant 0) &= f(0) + f(1) + \ldots + f(k) + \ldots = \sum_{x=0}^{\infty} f(x) \qquad x = 0 \\ P(X \geqslant 1) &= f(1) + \ldots + f(k) + \ldots = \sum_{x=1}^{\infty} f(x) \qquad x = 1 \\ &\ldots \qquad\qquad \ldots \\ P(X \geqslant k) &= f(k) + \ldots = \sum_{x=k}^{\infty} f(x) \qquad x = k \end{aligned}$$

Necessarily, the sample density function $f_r(x)$ must sum to one over all values of x; i.e.,

$$\sum_{x=0}^{k} f_r(x) = 1 \tag{1}$$

and this allows $f_r(x)$ to be expressed for each value of x in terms of $P(X \geqslant x)$, normalized so that Eq. 1 is fulfilled:

$$\begin{aligned} f_r(0) &= P(X \geqslant 0) \Big/ \sum_{x=0}^{k} P(X \geqslant x) \qquad x = 0 \\ f_r(1) &= P(X \geqslant 1) \Big/ \sum_{x=0}^{k} P(X \geqslant x) \qquad x = 1 \\ &\ldots \qquad\qquad \ldots \\ f_r(k) &= P(X \geqslant k) \Big/ \sum_{x=0}^{k} P(X \geqslant x) \qquad x = k \end{aligned}$$

Thus one can define $f_r(x)$ as

$$f_r(x) = P(X \geqslant x) \Big/ \sum_{x=0}^{k} P(X \geqslant x) \qquad x = 0, 1, \ldots, k \tag{2}$$

By the properties of discrete density functions [6], it can be shown that

$$\sum_{x=0}^{k} P(X \geqslant x) = \sum_{x=0}^{\infty} (x+1) f(x) \tag{3}$$

Since the right side of Eq. 3 is equal to $EX + 1$ [5,6], Eq. 2 can be expressed as $f_r(x) = P(X \geqslant x)/(EX + 1)$.

From these relationships it follows that the expected mean stay in the sample is

$$EX_r = \sum_{x=0}^{k} x f_r(x) = \sum_{x=0}^{k} x P(X \geqslant x)/(EX+1) = \sum_{x=0}^{\infty} \frac{x(x+1)}{2} f(x)/(EX+1)$$

and this yields [6]:

$$EX_r = (\mu_2' + EX)/2(EX+1) \tag{4}$$

where μ_2' is the second noncentral moment of the theoretical density function $f(x)$ and the other symbols are as defined before. To evaluate Eq. 4 one must assume some form for $f(x)$ so that μ_2' may be explicitly expressed.

Partial Follow-up Sample

In collecting a partial follow-up sample one takes the entire stay from admission to discharge for those patients present at the start of sampling, using their admission records for the elapsed portion of their stays and observing them until they are discharged. Again k is the longest stay and $x = 0, \ldots, k$. In this situation it is convenient to consider the probability $P(X = x)$ that a patient's stay is equal to x whole weeks. For $x = 0$, $P(X = 0)$ is proportional to $f(0)$; for $x = 1$, $P(X = 1)$ is proportional to $2f(1)$; for $x = 2$, $P(X = 2)$ is proportional to $3f(2)$; and so on to $x = k$, for which $P(X = k)$ is proportional to $(k + 1)f(k)$. Thus, for all x, $P(X = x)$ is proportional to $(x + 1)f(x)$.

As in Eq. 1, it is necessary that $\sum_{x=0}^{k} f_p(x) = 1$, where $f_p(x)$ is the sample density for the partial follow-up sample; and, as before, one can express $f_p(x)$ for each value of x, normalized so that the above relation holds:

$$\begin{aligned}
f_p(0) &= f(0)/\sum_{x=0}^{k} (x+1)\, f(x) && x = 0 \\
f_p(1) &= 2\, f(1)/\sum_{x=0}^{k} (x+1)\, f(x) && x = 1 \\
&\cdots && \cdots \\
f_p(k) &= (k+1)\, f(k)/\sum_{x=0}^{k} (x+1)\, f(x) && x = k
\end{aligned}$$

Thus the sample density function is defined as

$$f_p(x) = (x+1)\, f(x)/\sum_{x=0}^{k} (x+1)\, f(x) \qquad x = 0, \ldots, k$$

The expected theoretical length of stay EX is related [5,6] to the theoretical density function as

$$\sum_{x=0}^{\infty} (x+1)\, f(x) = EX + 1$$

and thus one may express the sample density function as

$$f_p(x) = (x+1)\, f(x)/(EX+1)$$

The expected stay for the sample, EX_p, is derived as

$$EX_p = \sum_{x=0}^{k} x\, f_p(x) = \sum_{x=0}^{k} x\,(x+1)\, f(x)/(EX+1)$$

which yields [6]

$$EX_p = (\mu_2' + EX)/(EX+1) \tag{5}$$

Again μ_2' is the second noncentral moment of the density function

$f(x)$, and, as with the retrospective sample, the form of this function must be assumed before Eq. 5 can be evaluated.

Interval Sample

For an interval sample one specifies a period of k weeks over which the sample will be observed; portions of any stay that extend before or after this period are ignored. For each value of $x = 0, \ldots, k$ one records the number of patients who are discharged after staying through x whole weeks of the observation period. In this situation the probability $P(X = x)$ that a patient's total stay is x whole weeks is proportional to the theoretical density function $f(x)$ plus the probability $P(X \geqslant x)$ that his total stay will be equal to or longer than x weeks. Thus for $x = 0$, $P(X = 0)$ is proportional to $2P(X \geqslant 0) + (k-2)$. $f(0)$; for $x = 1$, $P(X = 1)$ is proportional to $2P(X \geqslant 1) + (k-3)f(1)$; for $x = 2$, $P(X = 2)$ is proportional to $2P(X \geqslant 2) + (k-4)f(2)$; for $x = k-2$, $P(X = k-2)$ is proportional to $2P(X \geqslant k-2)$; and for $x = k-1$, $P(X = k-1)$ is proportional to $\sum_{x=k-1}^{\infty} P(X \geqslant x)$.

Reasoning on lines similar to those in the two previous situations, one can show that the sum of these terms defines the relation between the expected length of stay recorded during the sampling interval and the theoretical density function:

$$\sum_{x=0}^{k-2} [2P(X \geqslant x) + (k-2-x)\, f(x)] + \sum_{x=k-1}^{\infty} P(X \geqslant x) = EX + k$$

Since Eq. 1 holds also for $f_i(x)$, the probability density function may be written as:

$$f_i(x) = [2P(X \geqslant x) + (k-2-x)]/(EX + k) \qquad x = 0, \ldots, k-2$$

$$f_i(x) = \sum_{x=k-1}^{\infty} P(X \geqslant x)/(EX + k) \qquad x = k-1$$

Then the expected stay EX_i may be derived as

$$EX_i = \sum_{x=0}^{k-1} x\, f_i(x)$$

From these relationships it can be shown that

$$EX_i = \frac{\sum_{x=0}^{k-2} x[2P(X \geqslant x) + (k-2-x)\, f(x)] + (k-1) \sum_{x=k-1}^{\infty} P(X \geqslant x)}{EX + k}$$

and (by a long process) this equation can be simplified to

$$EX_i = (k-1)EX/(EX + k) \tag{6}$$

This case differs from the previous two since EX_i is directly related to EX. Therefore an estimate of EX can be made without knowledge of μ_2', which would require at least partial knowledge of the form of the distribution $f(x)$. Instead, one can estimate EX by substituting for EX_i in Eq. 6 the sample mean $\bar{x}_i$ from the interval data, which gives an estimated length of stay for the sample as

$$\hat{EX} = k\bar{x}_i/(k-1-\bar{x}_i) \tag{7}$$

As the number of weeks k becomes large in comparison to $\bar{x}_i$, the value of $\hat{E}X$ approaches EX. Another way of viewing this relationship is that as k becomes large $f(x)$ approaches a continuous distribution. For large values of k, the mean length of stay for the sample is an almost unbiased estimate of the expected length of stay EX. The magnitude of the bias will vary directly with $(1 + \bar{x}_i)/k$, so it can be easily investigated for any specific situation.

Examples

To estimate the mean length of stay for the retrospective and partial follow-up cases, a parametric assumption is necessary. The Poisson distribution is often assumed to represent the distribution of hospital stays; for the present purpose it will provide a convenient illustration. The Poisson density function is

$$f(x) = e^{-\lambda}\lambda^x/x$$

where λ represents the expected length of stay; that is, $\lambda = EX$. The second moment of $f(x)$ is $\mu_2' = \lambda^2 + \lambda$ [5]. When the Poisson density function describes the underlying probability density of a sample of retrospectively collected data, then, from Eq. 4,

$$EX_r = (\lambda^2 + 2\lambda)/2(\lambda + 1)$$

An estimate λ' of λ, the unbiased expected length of stay, is obtained by replacing EX_r by the observed mean value $\bar{x}_r$ and solving for λ', which yields

$$\lambda_r' = (\bar{x}_r - 1) + (\bar{x}_r^2 + 1)^{\frac{1}{2}} \tag{8}$$

Similarly, if the Poisson density function is assumed, one can derive the estimated length of stay for partial follow-up data by substituting λ for EX and $\bar{x}_p$ for EX_p in Eq. 5, which gives

$$\lambda_p' = [\bar{x}_p - 2 + (\bar{x}_p^2 + 4)^{\frac{1}{2}}]/2 \tag{9}$$

When lengths of stay are collected by interval sampling over an interval of length k, a parametric assumption is not necessary, as was mentioned. Given the assumption of the Poisson distribution, however, $\lambda_i' = \hat{E}X$ in Eq. 6 and the estimated length of stay is

$$\lambda_i' = k\bar{x}_i/(k - 1 - \bar{x}_i) \tag{10}$$

Table 1 presents a set of illustrative (fictional) data: values of $f(x)$, $f_r(x)$, $f_p(x)$, and $f_i(x)$ (for $k = 5$) that would result from the three sampling patterns applied to a group of patients whose stays follow a Poisson distribution with a true mean of $\lambda = 2$.

Application of the Interval Sampling Method

Data collected from Alameda County (CA) over the period from July 1, 1972 to April 3, 1974 were used to estimate the length of stay for two types of patients enrolled in the MediCal insurance program. The two groups are patients in a skilled nursing facility and patients in an intermediate care facility as defined by the Medicare guidelines.

Table 1. Values of Probability Density Functions from Three Sampling Patterns Applied to Hypothetical Patients with Poisson-distributed Lengths of Stay

Weeks completed, x	Poisson density, $f(x)$	Retrospective, $f_r(x)$	Partial follow-up, $f_p(x)$	Interval ($k=5$), $f_i(x)$
0	0.135	0.333	0.045	0.344
1	0.271	0.288	0.180	0.325
2	0.271	0.198	0.271	0.208
3	0.180	0.108	0.240	0.092
4	0.090	0.048	0.150	0.031
5	0.036	0.018	0.072	...
6	0.012	0.006	0.028	...
7	0.004	0.002	0.011	...
8–∞	0.001	0.000	0.003	...
Observed mean, $\bar{x}$ ...	2.00	1.33	2.67	1.14
Estimated* mean, λ' ...	2.0	2.0	2.0	2.0

* From Eqs. 8, 9, and 10, respectively.

The data collection period was divided into four 24-week intervals; $k = 24$. Table 2 shows the number of patents discharged each week, the total number of patients, the total number of patient weeks completed, the length-biased mean stay $\bar{x}_i$, and the estimated length of stay $\hat{E}X$ calculated from Eq. 7. The direct computation of $\bar{x}_i$ yields a fairly consistent three-week underestimate as compared with $\hat{E}X$. For both types of facilities, the average pattern of stays is similar: discharges rise for the first three weeks and then decrease through the rest of the period, even though the number of patients in the skilled nursing facility increases markedly.

Discussion

As with all statistical models, the validity of any application depends on how well the underlying assumptions are met. Parametric assumptions about $f(x)$, required by retrospective and partial follow-up sampling, will depend on a variety of conditions. For example, the type of institution, the type of service within the institution, and perhaps the type of patient receiving a particular service are conditions that will affect the choice of any probability density function to describe the pattern of length of stay.

In many situations the interval sampling pattern may be feasible, and in such cases the only assumptions necessary are that the lengths of stay are discrete and that the underlying probability density is stable. Estimates derived as described from interval data may be called distribution-independent, and from this point of view the interval method is clearly superior to the retrospective and partial follow-up

Table 2. Number of Patients Completing x Weeks of Stay During 24-week Intervals in Two Facilities

Weeks completed, x	Skilled nursing facility; interval*				Intermediate care facility; interval*			
	1	2	3	4	1	2	3	4
0	86	82	75	162	10	21	15	11
1	69	80	75	119	14	8	5	12
2	73	122	81	102	12	13	9	5
3	63	73	75	78	7	6	10	8
4	100	45	51	107	24	4	9	10
5	27	29	42	27	9	5	4	1
6	19	38	30	53	9	6	5	4
7	15	15	11	30	1	3	3	0
8	26	17	22	35	6	1	6	6
9	6	6	10	20	0	1	4	1
10	8	12	9	139	1	0	1	28
11	3	6	7	13	1	2	2	1
12	12	12	13	11	0	1	0	4
13	3	8	7	106	1	0	2	15
14	3	5	6	4	0	5	0	0
15	4	1	3	1	1	0	0	0
16	4	4	6	0	1	0	0	1
17	6	6	6	0	2	1	1	0
18	5	1	3	0	0	1	0	0
19	1	10	3	0	3	2	0	0
20	1	2	2	0	1	0	0	0
21	5	4	2	0	0	3	1	0
22	3	2	0	0	0	0	0	0
23	1	4	2	0	0	0	0	0
24–∞	75	78	89	3	6	12	16	0
Total patients	618	662	630	1010	109	95	93	107
Total weeks completed	4107	4416	4515	5294	629	695	724	722
$\bar{x}_i$	6.64	6.67	7.17	5.24	5.77	7.32	7.78	6.75
$\hat{E}X$	9.25	9.80	10.87	7.08	8.03	11.19	12.27	9.96

* Interval 1 = 7/1/72 to 12/15/72; 2 = 12/16/72 to 6/1/73; 3 = 6/2/73 to 11/16/73; 4 = 11/17/73 to 5/3/74.

methods. In some situations the interval method will be more efficient than complete follow-up sampling.

REFERENCES

1. Falk, S. Average length of stay in long-term institutions. *Health Serv Res* 6:251 Fall 1971.
2. Gustafson, D.H. Length of stay: Prediction and explanation. *Health Serv Res* 3:12 Spring 1968.
3. Hanson, B.L. A statistical model for length of stay in a mental hospital. *Health Serv Res* 8:37 Spring 1973.
4. Whitmore, G.A. The inverse Gaussian distribution as a model of hospital stay. *Health Serv Res* 10:297 Fall 1975.
5. Kendall, M.G. and A. Stuart. *The Advanced Theory of Statistics.* London: Griffin, 1967.
6. Mood, A.M. and F.A. Graybill. *Introduction to the Theory of Statistics.* New York: McGraw-Hill, 1963.