Nepal Journal of Science and Technology 9 (2008) 179-186

Nonparametric Distribution Fitting of Age at First Marriage of Nepalese Women through Scatter Plot Smoother

Ganga Shrestha and Srijan Lal Shrestha
Central Department of Statistics, Tribhuvan University, Kirtipur, Kathmandu
e-mail: srijan_shrestha@yahoo.com

Abstract

Age at first marriage is one of the important aspects for studying fertility behavior of women since it is believed that the number of children ever born is inversely related to age at first marriage. An attempt has been made to fit the Nepal Demographic and Health Survey (NDHS)-2006 data for age at first marriage of Nepalese women. The lack of fitting of some parametric distributions like lognormal and weibull was detected when applied to population census data of Nepal-2001 and the NDHS data-2006. Even though there is possibility of fitting more complex theoretical distributions, this highlighted the need to explore for alternative methods. In the process, it was identified that nonparametric smoothing is perhaps one of the viable solutions to this problem apart from fitting more sophisticated parametric distributions. Consequently, two nonparametric smoothers were examined for their suitability for distributional fitting, namely locally weighted regression smoother (LOESS) and cubic smoothing splines. Finally, cubic smoothing splines was chosen and fitted to NDHS data for Nepal and also separately for different ecological belts of Nepal since it fitted better than LOESS.

Key words: cubic smoothing splines, locally weighted regression smoother, Nepal demographic and health survey, nonparametric smoothing, theoretical probability distributions

Introduction

Probability distributions related to waiting time for conceptions and birth intervals in women have been popular among researchers interested in mathematical demography. The study of women's age at first marriage is nevertheless an interesting aspect in this field due to its importance in present and future fertility pattern of a population. It is now realized that considerable attention is needed in the study of age factor at the time of marriage since it leads to child bearing period of women. Coale (1971) proposed a distribution for age at first marriage in a cohort determined by three parameter exponential risk function given by

$$n(x) = \beta e^{-r e^{-\delta x}} \qquad (1)$$

For $x > 0$, β, r, and δ are positive constants. For some standard values of these parameters, further evidence was presented that the distribution of age at first marriage is governed by a family of risk functions depending upon two more parameters, α and k which vary between populations. Here, α represents the earliest age of first marriage and k is the speed at which marriages take place. The distribution as well as its derivative corresponding to equation (1) was not expressible in a closed form and hence was a difficult task to use this parametric model. Coale and McNeil (1972) used the distribution of age at first marriage as the distribution of the convolution of number of exponentially distributed components. Their distribution as the convolution of normal distribution of age of entry into a marriageable and a few as three exponentially distributed delays received remarkable agreement between theoretical and observed values.

The broad similarity across geographic locations and historical periods in the distribution of ages at which people first marry has been reported in the studies by Coale (1971) and Coale and McNeil (1972). The bell shaped curve of age at first marriage of women shows a quick rise from a minimum marriage age to attain maximum frequency followed by a long tail there after. The existing demographic and sociological theories have broadly classified the timing of marriage across two dimensions. The first dimension can be

considered as a continuum covering the ways in which people decide when to marry. At one extreme, the timing of marriage is a response to social norms, while at the other extreme, marriages happen when rational individuals decide the time will maximize their utility function subject to constraints (Goldstein & Kenney 2001). The second dimension is concerned with the heterogeneity of individuals, something more specific to modeling age patterns of marriage. On one end of the dimension, it can be considered that all individuals belonging to the same cohort or population are subject to the same type of influence On the other end, it can be thought that individuals as being heterogeneous in their choice of when to marry, according to some unobservable characteristics (Bennet *et al.* 1989, Bloom & Bennet 1990).

In the current context, the distributional fitting of age at first marriage of women was tried with some parametric probability distributions such as lognormal and weibull on the available data of Nepal Population Census-2001 and Nepal Demographic and Health Survey (NDHS)-2006. Data were categorized with respect to eco-belts, development regions and rural / urban Nepal for distributional fitting. It was observed that none of these theoretical probability distributions fitted well to categorized as well as non-categorized data with regard to various standard statistical procedures such as chi-square test, Kolmogorov Smirnov test, and probability and quantile plots.

The reasons behind the lack of fitting of the selected theoretical distributions could be many: These may include inaccurate reporting of age at first marriage and occurrence of other non-statistical errors, large sample sizes, remarkable degree of non-smoothness of the data suitable for parametric distributions, inappropriate choice of distributions, and need of piece-wise fitting dependent upon neighborhood data rather than global fitting of a single distribution to the whole range of values of the variable. Due to these problems, alternative statistical tools were explored and examined for their appropriateness in distributional fitting. During the process, a solution was identified as the use of non-parametric functions in the form of scatter plot smoothers as widely used in generalized additive models (GAM).

Materials and Method

The distributional data fitting of age at first marriage for Nepalese women was carried out by the application of non-parametric smoothers namely LOESS and cubic smoothing splines. First of all, the observed frequency distribution of age at first marriage was constructed. Taking the frequency (number of females) for the particular age as the dependent variable and the age itself as the predictor variable, the analysis was carried out. Data were taken from the Nepal Demographic and Health Survey-2006. The survey was conducted in the year 2006 and the number of married females covered in the survey was around 8621 (NDHS 2006). Distributional fitting, categorized according to ecological regions of Nepal and the whole Nepal was estimated using statistical software, namely SPLUS 2000 and Statistical Analysis System (SAS) version 9.

A brief introductory discussion on smoother, cubic smoothing splines, locally weighted regression smoother, and generalized cross validation have been presented in the following sections.

Smoother

A smoother is a tool for summarizing the trend of a response variable, Y as a function of one or more predictor measurements, X_1, X_2,..., X_p. It produces an estimate of the trend that is less variable than Y itself; hence the name smoother. These smoothers are nonparametric in nature since they do not assume the rigid form of dependence of Y on the predictor variables. The single predictor case is called as scatter plot smoothing since the dependence can be visualized in a two dimensional scatter plot. There are two main decisions to be made in scatter plot smoothing namely, how to average the response values in each neighborhood, and how big to take the neighborhoods. The question of how to average within a neighborhood is really the question of which type of smoother to use

and therefore, depends upon the functional form of the smoothers. Some of the different types of smoothers are running line smoother, kernel smoother, regression splines, locally weighted regression smoother (Loess), and cubic smoothing splines. The size of the neighborhood is expressed in terms of an adjustable parameter called as smoothing parameter (λ). In general, large neighborhoods produce an estimate with low variance (high smooth) but potentially high bias. Conversely, small neighborhoods produce an estimate with low bias but potentially high variance (less smooth). Thus, there is a fundamental tradeoff between the bias and the variance, governed by the smoothing parameter with small λ resulting in small bias and large variance (equivalently, large degrees of freedom) and large λ resulting in small variance and large bias (equivalently, small degrees of freedom).

Cubic smoothing splines

A cubic smoothing spline minimizes the penalized least square given by

$$\sum_{i=1}^{n}\left[y_i - f(x_i)\right]^2 + \lambda\int_a^b \left\{f''(x_i)\right\}^2 dx \qquad (2)$$

where λ is a fixed constant, and $a \le x_1 \le x_2 \le \ldots \le x_n \le b$. The first term measures closeness to the data while the second term penalizes curvature in the function. There exists an explicit, unique minimizer which is a natural cubic spline with knots at the unique values of x_i. Linear functions have $\int\left\{f''(x_i)\right\}^2 dx = 0$, while non-linear functions produce values bigger than zero. It governs the tradeoff between the goodness of fit to the data and wiggleness of the function. The parameter λ is the smoothing parameter. Here, Y is a response or outcome variable, and X is a prognostic factor. We wish to fit a smooth curve *f*(*x*) that summarizes the dependence of Y on X. If we were to find the curve that simply minimizes $\sum\left[y_i - f(x_i)\right]^2$, the result would be an interpolating curve that would not be smooth at all. For any value of λ, the solution to (2) is a cubic spline, i.e., a piecewise cubic polynomial with pieces joined at the unique observed values of X in the dataset. Fast and stable numerical procedures are available for computation of the fitted curve (O'Sullivan 1985) as in backfitting, an iterative algorithm proposed by Friedman and Stuetzle (1981).

Locally weighted regression smoother (LOESS)

Local regression was proposed by Cleveland *et al.* (1988). The idea of local regression is that at a predictor *x*, the regression function η(x) can be locally approximated by the value of a function in some specified parametric class. A weighted least squares algorithm is used to fit linear or quadratic functions of the predictors at the centers of neighborhoods. The radius of each neighborhood is chosen so that the neighborhood contains a specified percentage of the data points. The smoothing parameter for the local regression procedure, which controls the smoothness of the estimated curve, is the fraction of the data in each local neighborhood. The procedure for obtaining the LOESS is given below.

The smooth function s(x) is built pointwise as follows (equation 3 -6):

1. A point, say x_0 is taken. The k nearest neighbors of x_0 are detected, which constitute a neighborhood $N(x_0)$. The number of neighbors, k is the percentage of the total number of points called the span.
2. The largest distance between x_0 and another point in the neighborhood is calculated:

$$\Delta(x_0) = \text{Max}_{N(x_o)}\left|x_o - x_i\right| \qquad (3)$$

3. Weights to each point in $N(x_0)$ are assigned using the tri-cube weight function:

$$W\left(\frac{\left|x_o - x_i\right|}{\Delta(x_o)}\right) \qquad (4)$$

where

$$W(u) = \begin{cases}(1-u^3)^3 & \text{for } 0 \le u < 1\\ 0 & \text{otherwise}\end{cases} \qquad (5)$$

4. The weighted least squares fit of y on the neighborhood $N(x_0)$ is taken as:

$$\hat{y}_o = S(x_o) \qquad (6)$$

5. The procedure is repeated for each predictor value.

Generalized cross validation

A general criterion for selection of the smoothing parameter and subsequently, the degrees of freedom is the generalized cross validation (GCV) (equations 7 to 10). In choosing the smoothing parameter, cross validation can be used. Cross validation works by leaving points (x_i, y_i) out one at a time, estimating the squared residual for smooth function at x_i based on the remaining n-1 data points, and choosing the smoother to minimize the sum of those squared residuals. The cross validation function is defined as

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \hat{\eta}_{\lambda}^{-1}(x_i)\right]^2 \qquad (7)$$

where $\hat{\eta}_{\lambda}^{-1}(x_i)$ indicates the fit at x_i, computed by leaving out the i^{th} data point.

All of the smoothers fit by the GAM can be formulated as a linear combination of the sample responses

$$\hat{\eta}(x) = A(\lambda)Y \qquad (8)$$

for some matrix $A(\lambda)$, which depends on λ. Let a_{ii} be the diagonal elements of the $A(\lambda)$,. Then the CV function can be expressed as

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \hat{\eta}_{\lambda}^{-1}(x_i)}{1 - a_{ii}}\right]^2 \qquad (9)$$

In most cases, it is very time consuming to compute the quantity a_{ii}. To solve this computational problem, Wahba (1990)[11] proposed the generalized cross validation function (GCV) that can be used to solve a wide variety of problems involving selection of a parameter to minimize the prediction risk. The GCV function is defined as

$$GCV(\lambda) = \frac{n\sum_{i=1}^{n}\left[y_i - \hat{\eta}_{\lambda}^{-1}(x_i)\right]^2}{\left[n - tr(A(\lambda))\right]^2} \qquad (10)$$

The GCV formula simply replaces the a_{ii} with $\frac{tr(A(\lambda))}{n}$. Therefore, it can be viewed as a weighted version of CV.

Results and Discussion

The results of descriptive analysis and nonparametric smoothing are presented in the following sections.

General

The observed mean age of women at first marriage for Nepal is low and around 17 with standard deviation of 3 years. It is lower in the Terai region (16.5) as compared to hill (17.4) and mountain (17.6). The variation of age at marriage is highest in mountain with standard deviation (SD) of 3.4 years. Regarding CEB, mean for Nepal, mountain, hill, and Terai are 3.04, 3.23, 2.97, and 3.03, respectively. Considering median value of age at first marriage, it is 16 in the terai region and 17 in the other regions of Nepal. Similarly, median value for CEB is found to be same (3) for all the regions of Nepal. As expected and believed, negative and significant correlations ($p < 0.01$) were observed between age at first marriage and CEB for Nepal and separately for three ecological regions of Nepal. However, in Terai region despite with lowest mean age at first marriage, CEB is not highest in the region but observed in the mountain region, instead for all married women and for age group 45-49. This indicates that CEB is not dependent upon age at first marriage alone and can be affected by other factors as well (Table1).

Table 1. Descriptive statistics on age at first marriage and children ever born

Region	Age at first marriage			Children ever born(CEB)			Correlation	CEB45-49	
	Mean	Median	SD	Mean	Median	SD		Mean	Median
Nepal	16.97	16.0	2.994	3.04	3.0	2.187	-0.207	5.3	5
Mountain	17.60	17.0	3.446	3.23	3.0	2.322	-0.143	5.7	6
Hill	17.36	17.0	3.255	2.97	3.0	2.188	-0.228	5.2	5
Terai	16.48	16.0	2.525	3.03	3.0	2.145	-0.221	5.2	5

Nonparametric smoothing

Results of smoothing presented in Tables 2 and 3 show that cubic smoothing spline is a better fit than LOESS with regard to various criteria. For instance, mean square errors are much lower, R square values are higher, and generalized gross validation values are lower for cubic smoothing splines as compared to LOESS. Consequently, cubic smoothing splines is selected for distribution fitting in the current analysis. Distributional fittings are presented separately for whole of Nepal and different ecological belts. In order to test whether the observed and expected frequencies are statistically insignificant, chi-square tests were conducted. The computed values of chi-square are found to be: 24.02 at 21 df (insignificant with p value = 0.29), 8 at 15 df (insignificant with p value = 0.92), 25.86 at 18 df (just insignificant with p value = 0.103), and 5.56 at 19 df (insignificant with p value = 0.9987) for Nepal, mountain, hill, and Terai regions of Nepal, respectively.

The smoothed curves appeared like a bell shape arising quickly from a minimum marriage age with elongated tails capturing people who marry late in life, more or less like the shape of log-normal distribution. Studies conducted before also found similar shaped fitted curves (Coale 1971, Coale & McNeil 1972). If we observe the smoothed curves more carefully, it can be noticed that the curve for Terai region is thinner and elongated more vertically than observed for hill and mountain regions of Nepal. This can be interpreted as expected frequencies of age at marriage rise and fall more sharply around the modal value for Terai women as compared to the women of other regions. The smoothed frequency distributions for Nepal show that 20% of the total expected frequency lies below age 15, 40% of the total frequency lies below 16, 60% of the total frequency lies below 17, and 80% of the total frequency lies below 19 years of age at first marriage. The figures demonstrate that overwhelmingly large proportion of Nepalese women marry in their teen age. This has resulted and will result in many negative consequences to Nepalese women, community and to Nepal.

It is to be noted that during the process of smoothing observations with frequency less than 4 were treated as outliers and excluded from analysis since this can lead to negative expected frequencies and retained only if the next lower or higher age at marriage had higher frequency (>4) and did not lie in the boundaries. In the process, 5, 15, 13, and 6 extreme observations were deleted in the current analysis for Nepal, mountain, hill, and Terai regions, respectively.

Table 2. Estimates for cubic smoothing splines

Estimate	Nepal	Mountain	Hill	Terai
Smoothing parameter	0.008	0.0375	0.0501	0.0036
Degrees of freedom	14.7	8.4	8.9	15.4
Mean square error	1978.6	72.02	537.9	457.47
R square	0.991	0.991	0.998	0.99
GCV	5931.9	151.77	1013.3	2026.3

Table 3. Estimates for LOESS

Estimate	Nepal	Mountain	Hill	Terai
Smoothing parameter	0.2165	0.3361	0.2938	0.2832
Number of neighbors(k)	4	5	5	5
Degrees of freedom	10.6	7.3	8.5	9.0
Mean square error	4870.67	86.54	670.20	2193.2
R square	0.989	0.987	0.988	0.984
GCV	9375.21	159.19	1218.31	3977.3

The observed and expected frequency distributions (Table 4) and corresponding figures (Figures 1 to 4) are presented below.

Table 4. Observed and smoothed frequency distributions for the whole of Nepal and eco-belts Age at

first marriage	Mountain		Hill		Terai		Nepal	
	Observed	Smoothed	Observed	Smoothed	Observed	Smoothed	Observed	Smoothed
10	-	-	-	-	5	4	8	3
11	-	-	-	-	15	14	26	23
12	11	10	46	39	64	60	121	125
13	51	53	153	180	184	198	388	412
14	91	100	325	326	482	483	898	902
15	152	145	455	442	778	782	1385	1385
16	180	169	531	489	922	884	1633	1554
17	156	165	421	459	624	652	1201	1270
18	151	146	417	393	416	412	984	958
19	116	115	285	307	239	244	640	662
20	81	80	229	226	159	154	469	455
21	37	50	159	154	89	94	285	294
22	43	35	87	96	69	65	199	192
23	23	24	59	61	35	37	117	121
24	17	17	45	43	21	21	83	83
25	12	12	37	32	15	14	64	60
26	8	8	18	21	9	9	35	39
27	6	6	17	14	8	8	31	28
28	-	-	3	9	9	8	15	19
29	-	-	13	8	5	5	20	16
30	-	-	4	5	-	-	7	9
31	-	-	-	-	-	-	7	6
Total	1135	1135	3304	3304	4148	4148	8616	8616

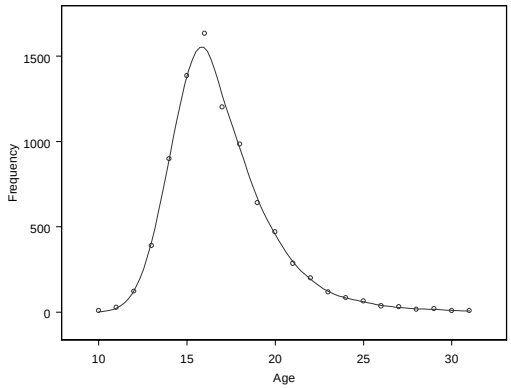

Fig. 1 Fitted frequency distribution of age at first marriage for Nepal

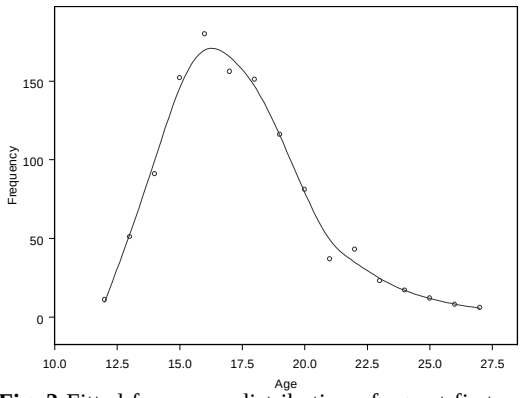

Fig. 2 Fitted frequency distribution of age at first marriage for mountain region of Nepal

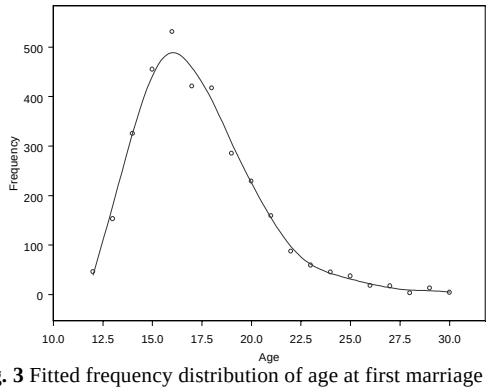

Fig. 3 Fitted frequency distribution of age at first marriage for hill region of Nepal

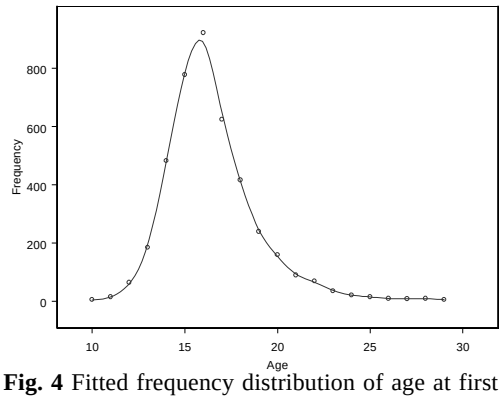

Fig. 4 Fitted frequency distribution of age at first marriage for Terai region of Nepal

Age at first marriage is undoubtedly an important demographic characteristic of a population. Early marriage is directly associated with the early initiation of child bearing and high fertility which may have adverse effects on the health of mothers. In a developing country like Nepal, low mean age at first marriage suggests higher children ever born in Nepalese women and ultimately high fertility rate in the country. Age at first marriage has a major effect on childbearing because women who marry early have, on average, a longer period of exposure to the risk of becoming pregnant and a greater number of lifetime births. There are many other social, economical, and health concerns related to early marriage. It may cause not only health hazards to women but also affect their mental development due to increased domestic responsibility, deprivation of freedom and education as well. In this context, the paper aims to highlight this issue by studying the distributional pattern of age at first marriage and attempts to fit a nonparametric function in the form of scatter plot smoother.

The observed mean and median age at first marriage for Nepal, according to NDHS-2006 survey data is low and around 16.9 with SD 2.8 and 16, respectively. The value is lower in the Terai region (16.5) as compared to hill (17.3) and mountain (17.4). The variation of age at marriage is highest in hill with SD of 3.0 years. Considering median value of age at first marriage, it is 16 in the Terai region and 17 in the other regions of Nepal. The median age when compared to some other countries is much lower since it is 25.9 in USA (US Bureau of Census 2006), 28.5 in UK (Office for National Statistics 2005), 22.3 in India, 19.7 in Pakistan and 29.6 in Germany (Wikipedia encyclopedia 2008).

Regarding CEB, mean for Nepal, mountain, hill, and Terai are 3.04, 3.23, 2.97, and 3.03 respectively. Similarly, median value for CEB is found to be same (3) for all the regions of Nepal. As expected, negative and significant correlations ($p < 0.01$) were observed between age at first marriage and CEB for Nepal and separately for three ecological regions. However, in Terai region despite with lowest mean age at first marriage, CEB is not highest in the region but observed in the mountain region, instead. This indicates that CEB is not dependent upon age at first marriage alone and can be affected by other factors as well.

The nonparametric way of distribution fitting of age at first marriage on NDHS-2006 data with cubic smoothing splines showed remarkable agreement between observed values and the smoothed values. This has demonstrated that the traditional mode of distribution fitting through parametric theoretical distributions which fit data globally over the entire range has an alternative and dependable way out in the form of nonparametric smoothing which essentially fits data locally and piecewise. In the paper, this has been presented for the whole Nepal and also for different ecological regions.

References

Coale, A. J. 1971. Age patterns of marriage. *Population studies* 25:193-214.

Coale A. J. and D. R. McNeil. 1972. The distribution by age of the frequency of first marriage in a female cohort. *Journal of the American Statistical Society* 67:743-749.

Goldstein, J.R. and C. T. Kenney. 2001. Marriage delayed or marriage forgone? new cohort forecasts of first marriage for U.S. women. *American Sociological Review* 66:506–19.

Bennet, N.G., D. E. Bloom and P. H. Craig. 1989. The divergence of black and white marriage patterns. *American Journal of Sociology* 95:692–722.

Bloom, D.E. and N. G. Bennet. 1990. Modeling American marriage patterns. *Journal of the American Statistical Association* 85:1009–17.

Nepal Ministry of Health and Population. 2007. Nepal demographic and health survey 2006. Nepal Ministry of Health and Population (MOHP), New ERA and Macro International Inc., Kathmandu, Nepal.

Hastie, T. J. and R. J. Tibshirani. 1990. *Generalized Additive Models*. Chapman and Hall, Inc., New York.

O'Sullivan, F. 1985. Discussion of some aspects of the spline smoothing approach to nonparametric regression curve fitting by B. W. Silverman. *Journal of Royal Statistical Society B.* 36:111 – 147.

Friedman, J. H. and W. Stuetzle. 1981. Projection pursuit regression. *Journal of the American Statistical Association* 76:.817–823.

Cleveland, W. S., S. J. Devlin and E. H. Grosse. 1988. Regression by local fitting: methods, properties and computational algorithms. *Journal of Econometrics* 37: 87-114.

Wahba, G. 1990. *Spline functions for observational data*. CBMS-NSF Regional Conference Series, SIAM, Philadelphia, USA.

U. S. Bureau of Census. 2006. American Community Survey, USA. US Bureau of Census, USA. [electronic resource]. http:// www.infoplease.com/ipa/ A005061.html

Office of National Statistics. 2005. Marriage and divorce statistics historical series. Office of National Statistics, UK. [electronic resource]. www. statistics.gov.uk/ statbase/

Wikipedia (2008) Wikipedia free encyclopedia [electronic resource]. http://en.wikipedia.org/wiki/wikipedia