RESTORATION OF DATA WITH ROUNDING AND BOUNDING ERRORS

Jiahe Qian, ETS
Rosedale Road, Princeton, NJ 08541

KEY WORDS: Rounding and Bounding Errors, Adjustment, Ascertainment, Smoothing

In survey sampling, rounding and bounding errors often occur when reporting from memory. The errors could be the result of the gradual deterioration of memory or could also be influenced by the reporters' characteristics or social desires. A great deal of literature about the effects of memory related factors on response have appeared since Mahalanobis reported his findings (Mahalanobis, 1946; Neter & Waksberg, 1964; Sudman & Bradburn, 1973; Huttenlocher, Hedges, & Bradburn, 1990). Researchers in the field of Psychology are mainly concerned with how rounding and bounding errors occur when information was stated from memory. Many models about response and theories dealing with memory were proposed: the underlying form of representation, the pattern of information loss over time, the estimation processes of the events in response, etc. How data with rounding and bounding errors is used in statistical analysis is nevertheless an issue of interest. Some studies have been done inferring for individual cases via multiple imputation (Heitjan & Rubin, 1990); yet, this would not certify the yielding of an aggregate distribution that is consistent with that of population.

This paper focuses on obtaining unbiased information from data with rounding and bounding errors and providing empirical distribution of true populations. The goal is to adjust the frequency of data with small mean square errors (MSE). Since the true population is generally unknown, studies usually turn to examining the consistency of the statistics between the smoothed data and the assumptions of a true distribution.

For this purpose, it is necessary to first detect such errors, and then make the proper data adjustments to probabilities of the actual occurrences of events. Neglecting to make these adjustments could lead to the wrong conclusions; nonetheless, such errors in data are overlooked in many applications.

1. Bounding and Rounding Errors in Survey Data, Two Examples:

When reporting events from memory, rounding and bounding errors can occur depending upon the different types of events that are being reported, such as age, weight, elapsed time, etc.

a. One example is about reporting the amount of weight gained during a pregnancy in the National Maternal and Infant Health Survey (NMIHS), which studied the pregnancy experiences of women who had lost their infants. See Figure 1a. Similar problems can be found in the variable of the number of pounds lost during a pregnancy in NMIHS.

Medical knowledge shows that the change in weight of women after pregnancy observes a smooth continuous curve with one peak; however, the data shows extremely large counts at 5 and 10 pounds. These rounding errors, a systematic bias in response, are clearly shown in the data. Therefore, the distribution of the observed data does not represent the true distribution of weight gained during pregnancy. Also, due to the large counts at 5 and 10 pounds, the data may appear to have periodic circles of 5 and 10; but they are not real periodic intervals. So difference operator has no effect in eliminating them.

Such errors could be caused by memory flaws or lack of aided recall during the survey interview. Furthermore, we have reason to believe that, for cases with rounding errors, the systematic bias is also a forward bias. The reported number of pounds gained were rounded to the lower bound of 5's or 10's. One explanation for the forward bias is that being overweight is socially undesirable in America. This is supported by the fact of increased percentages of conventional arithmetic prototypes as the weights in the intervals increase, see Table 1.1.

Table 1.1. The proportions of observed responses for prototypes in observed data

Ranges in lbs	% of conventional arithmetic prototypes	% of unaltered response
0-9	0.395	0.605
10-19	0.590	0.410
20-29	0.633	0.367
30-39	0.717	0.283
40-49	0.818	0.182
50-59	0.842	0.158
60-69	0.867	0.133

b. Second example. In 1993 National Study of Post-secondary Faculty (NSOPF-93), the variables about the allocation of the total work time, X05C37-X08C37, have been found the rounding and bounding errors in response. Among them, X05C37 was about the percentage of time spent in teaching in the Fall of 1992.

c. Some results in study of response errors

Response bias could be caused by the effects of

memory deterioration or by the influence of the reporters' characteristics and the impact of social desirability about the questions asked. Many psychologists and statisticians have made interesting discoveries in this topic.

When reporting events from memory, the accuracy of reporting depends on what has been encoded in the recollection, and on the recall process that yields the image of recollection. The psychology of memory shows that the linkage of one event with others plays an important role in remembering and recalling events.

Because of the physical structure of the brain, the linkage among events becomes weaker as time elapsed. Then reporting errors would occur in two possible situations: either (1) forgetting the event or (2) being influenced by reference episodes on the recall process.

In the latter case, the recollection process is also an estimation process. Whether or not the linkage is proper, rounding and bounding errors could still occur; however, when the proper linkage of the event is replaced by inaccurate reference, errors arise which could cause serious bias in the distribution of events, which are either temporal or non-time related.

For temporal events, research has found that the reported time of an event's occurrence tends to have a forward bias, so events are reported as occurring more recently than they when actually occurred (Huttenlocher, Hedges, and Bradburn, 1990). Such phenomena is called compression of time or telescoping. One explanation is that, when a recent event is taken as a reference by a respondent, whether consciously or not, the time interval between the present and when an event truly occurred would be remembered to be shorter than it actually was.

For the non-time related event, the respondent could inaccurately remember the scale of an event and place bounds as reference episodes on the recall process. A rounding error occurs when the scale of an event shrinks to the nearest bound. And a bounding error occurs when a bound is put as an upper bound or lower bound, then a border bias would occur.

The effects of rounding and bounding errors is the introduction of uncertainties in data along with the those associated with random error from sampling procedure. They introduce systematic bias in estimation. One of our interests is to detect such errors and make data adjustments to the distribution of actual occurrences of events in a population.

d. A statistical model for representation with rounding and bounding errors

By psychological analysis (Hedges & Bradburn, 1990), the model of responses with rounding and bounding errors consists of a bounding process and two response processes.

The bounding process can be expressed by a doubly truncated normal distribution:

$$f(X_i|\mu,\sigma) = \begin{cases} \dfrac{\varphi(X_i|\mu,\sigma)}{\phi(b)-\phi(a)}, & \text{if } 0 \le X_i \le 70, \\ 0, & \text{if } x<0 \vee x>70; \end{cases}$$

where $\varphi(X_i|\mu,\sigma) = \exp(-(X_i-\mu)^2/2\sigma^2)/\sqrt{2\pi}\,\sigma$ is a density function of normal distribution, the parameters in the normal distribution functions are defined as $b = (70-\mu)/\sigma$, and $a = -\mu/\sigma$.

The first response process yields unaltered data which is almost identical to the values in memory:

$$P(i|\alpha,\beta) = \frac{\phi((i+0.5-\mu)/\sigma) - \phi((i-0.5-\mu)\sigma)}{\phi(b)-\phi(a)}.$$

The second response process yields arithmetic prototypes:

$$P(5i|\alpha,\beta,c) = \frac{\phi((5i+5c-\mu)/\sigma) - \phi((5i-5c-\mu)/\sigma)}{\phi(b)-\phi(a)}.$$

The observed data yields from mixtures of two response processes:

$$y_i = \gamma\,\xi_1(\mu=w) + (1-\gamma)\,\xi_2(\mu=w)$$

where $\xi_1(\mu=w)$ forms the distribution defined in the first response process , and $\xi_2(\mu=w)$ forms the distribution defined by the second response process. The coefficient γ forms an exponential distribution (Sudman, 1973), and the empirical distribution for weight gained in NMIHS is showed in Table 1.1.

The statistical model of response processes is useful in explaining and analyzing the phenomena of rounding and bounding errors. However, the function of the model in restoration of true distribution is limited.

2. Approaches for the adjustment of data with rounding errors

The procedure of data adjustment consists two steps: the adjustment of outlier counts and smoothing approaches. We will demonstrate the effects of adjustments through the example of weight gained during a pregnancy in NMIHS.

In this example, based on medical knowledge, we assume the weight gained of pregnant women forms a smooth continuous curve with one peak. From medical literature (Ash, 1989), we find that the mean increase ranges from 23.6 to 33.5 pounds; generally a mean weight increase total of about 27.5 pounds is considered normal through out a pregnancy (Rossner, 1995).

a. The adjustment of outlier counts

Figure 1a shows that the count for respondents who gained 0 pounds is unusually large: 78, which looks like

an outlier. There are two explanations for it. First, there must be some respondents who lost weight yet put 0 for the question of weight gained. On the questionnaire, two questions: asking the amount of weight gained and weight lost were put together; but there was no instruction for women who lost weight to put a missing code on the question of weight gained. Second, zero is a special point which has dual rounding effects. Some women who gained less than 10 pounds would round the weight gained to 0, and some who lost less than 10 pounds would also round the weight gained to 0. Considering these factors, the estimated count of 0 for women who gained weight is put down as 17 in the adjustment.

In the data, we also found more counts clustered at 10's than 5's, so it seems that there were two periods in the data. One explanation is that some women rounded the weight gained to the lower 10's and omitted the lower 5's. To clarify the procedure of smoothing, we presume that very few cases skipped the nearest lower 10's and were rounded the value to next lower 10's (e.g. 33 is rounded to 20 instead of 30); the same assumption goes for the 5's as well. Before smoothing, we adjusted some counts at 10's to the 5's above them. We assume the chance for respondents to round to 10's declines when the single digit of the pounds gained increases. Therefore, as the trend is increasing, the adjusted counts at 10d and 10d+5 are set $\hat{x}_{10d} \doteq x_{10d} - .75\,\delta_{10d}$ and $\hat{x}_{10d+5} \doteq x_{10d+5} + 0.75\,\delta_{10d}$. Whereas, $\hat{x}_{10d} \doteq x_{10d} - .25\,\delta_{10d}$ and $\hat{x}_{10d+5} \doteq x_{10d+5} + 0.25\,\delta_{10d}$ as the trend decreasing. For the results of ascertainment, see Figure 1b.

b. The smoothing approaches

In smoothing, several approaches are used to adjust the rounding and bounding errors in the variable of weight gained in NMIHS.

i. Weighted average

$$\hat{x}_i = \sum_{k=i-2}^{i+2} w_k x_k, \text{ for } i = 0, 1, 2, \ldots, n,$$

where $\mathbf{w}_i = (w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2})$ is the vector of weights for moving average. In Figure 2a, $\mathbf{w}_i = (0.2, 0.2, 0.2, 0.2, 0.2)$. In NMIHS example, the normal weights, $\mathbf{w}_i = (0.094, 0.234, 0.344, 0.234, 0.094)$, do not demonstrate advantages.

Weighted average is usually the first step of smoothing. Other smoothing approaches are applied to the output data of the moving average approach.

ii. Probability model

For the weight gained example in NMIHS, the distribution of the number of pounds gained is skewed to the left, like Gamma distribution. If Gamma distribution,

$$f(x) = \begin{cases} \dfrac{\theta^r x^{r-1} e^{-\theta x}}{\Gamma(r)}, & \text{if } x \geq 0, \\ 0, & \text{if } x < 0, \end{cases}$$

is taken as the theoretical frequency distribution, we shall use moment estimates, $\hat{\theta} = \bar{\mu}/\hat{\sigma}^2$ and $\hat{r} = \bar{\mu}^2/\hat{\sigma}^2$, based on observational curves to estimate the parameters. The fitted probability distribution does not appear suitable, especially at lower end. However, truncated Gamma distribution gives a reasonably smooth curve; see Figure 5a.

iii. Local regression model

Instead of fitting a curve throughout the range of the independent variable's values by regular regression, local regression calculates the best fitting linear regression model to those observations in a neighborhood of each selected value of independent variable. The fitted curve is spanned by the a series of such fitted values.

In the weight gained example in NMIHS, each value of the number of pounds gained will have a fitted count by a series of quadratic polynomials:

$$y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i.$$

In fitting local regression models, the smoothing parameter equals 0.53 and degree is 2. And observations are assigned neighborhood weights which are normally distributed.

The curve, in Figure 6a, shows a good fit.

iv. Wavelets smooth approach

One of the applications of Wavelet expansions is smoothing data. In the smooth procedure of noisy sampled data (Donoho, 1995), the noise added to signals is filtered through wavelet shrinkage with small MSE. Wavelet expansions have similar convergence properties as Fourier series. The bases of wavelet transformation are orthonormal bases of various space. Additionally, wavelet bases are designed to give simultaneous time frequency localization information, and nonlinear spatial adaptive methods for noisy data.

In smoothing data with rounding and bounding errors, a multi-level shrink procedure is used to reduce the fluctuation in signals. The coefficients of wavelets in both smooth and un-smooth vectors are shrunk to the means of the vectors separately at different levels of discrete orthogonal wavelet transforms. The discrete wavelet transform operates upon a data vector with length of an integer power of two. In NMIHS, the

range, of gained weight, considered was from 0 to 63, i.e. N=64=2^6. This constraint could cause a loss of effective range of data, though there was no such problem in NMIHS example.

Multi-level shrink procedure aims to restore a population distribution through the adjustment of the frequency of data with small MSE. Usually, the true distribution is unknown, then the criteria becomes examining whether the data after wavelet smoothing is consistent with the assumptions of the true distribution.

The approach of multi-level shrinkage:

a) Discrete wavelet transform data with rounding and bounding errors: $\mathbf{w} = \boldsymbol{\omega}\mathbf{y}$, where vector $\mathbf{y}$ consists of raw data, vector $\mathbf{w}$, wavelet coefficients, has 2^J elements, and for $w_{j,k}$, j=0, 1,..., J-1; k=0, 1, ..., 2^j-1;the remaining element is labeled as $w_{-1,0}$; also in $\boldsymbol{\omega} = \prod_{j=0}^{J-1} \boldsymbol{A}_j \boldsymbol{\omega}_j$, the $\boldsymbol{\omega}_j$ is the wavelet transform matrix at level j plus the $\mathbf{A}_j$ is the position transform matrix at level j. Since $\boldsymbol{\omega}$ is non-singular, $\mathbf{y} = \boldsymbol{\omega}^{-1}\boldsymbol{w}$.

b) Shrink the wavelet coefficients of the original data and the 'mother-function' to their means of the coefficients separately.

The wavelet coefficients shrinkage and the reverse transform can be expressed in matrix form $\hat{\mathbf{y}} = \boldsymbol{T}\boldsymbol{w}$, where $\mathbf{T} = \prod_{j=0}^{J-1} \boldsymbol{\omega}_{J-1-j}^{-1} \boldsymbol{A}_{J-1-j}^{-1} \boldsymbol{T}_{J-1-j}$.

Let $\mathbf{w}_{J-p} = (\prod_{j=p}^{J-1} \boldsymbol{\omega}_{J-1-j}^{-1} \boldsymbol{A}_{J-1-j}^{-1} \boldsymbol{T}_{J-1-p})\boldsymbol{w}$, and

$\mathbf{w}_{J-p} = \boldsymbol{\omega}_{J-1-p}^{-1} \boldsymbol{A}_{J-1-p}^{-1} \boldsymbol{T}_{J-1-p} \boldsymbol{w}_{J-1-p}$, p=1, 2,..., J-1.

Shrinkage matrix $\mathbf{T}_{J-1-p}$ is defined as

$$\mathbf{T}_{J-1-p} \triangleq \begin{pmatrix} \mathbf{T}_{J-1-p,1} & 0 & 0 \\ 0 & \mathbf{T}_{J-1-p,2} & 0 \\ 0 & 0 & \mathbf{I}_{J-1-p,3} \end{pmatrix}, \text{ where}$$

$\mathbf{T}_{J-1-p,1}$, $\mathbf{T}_{J-1-p,2}$, and identity $\mathbf{I}_{J-1-p,3}$ have orders of 2^{J-1-p}, 2^{J-1-p}, and $\sum_{s=1}^{p} 2^{J-s}$. If shrinkage is necessary, $\mathbf{T}_{J-1-p,1}$ equals $\boldsymbol{\alpha}_{J-1-p,1} + (\mathbf{I}_{J-1-p,1} - \boldsymbol{\alpha}_{J-1-p,1})\,\mathrm{diag}(\boldsymbol{\mu}_{J-1-p,1})\mathrm{diag}(\mathbf{w}_{J-1-p,1})^{-1}$, otherwise $\mathbf{I}$. Similarly, if shrinkage is needed, $\mathbf{T}_{J-1-p,2}$ equals $\boldsymbol{\alpha}_{J-1-p,2} + (\mathbf{I}_{J-1-p,2} - \boldsymbol{\alpha}_{J-1-p,2})\,\mathrm{diag}(\boldsymbol{\mu}_{J-1-p,2})\mathrm{diag}(\mathbf{w}_{J-1-p,2})^{-1}$, otherwise $\mathbf{I}$. Diagonal matrix $\boldsymbol{\alpha}_{J-1-p,1}$ has elements between 0 and 1 , vector $\mathbf{w}_{J-1-p,1}$ has elements $w_{j,k}$ (j= 0, 1,..., J-1-p; k=0, 1, ..., 2^j-1), the elements of vector $\boldsymbol{\mu}_{J-1-p,1}$ equal the mean of $\mathbf{1}'\boldsymbol{w}_{J-1-p,1}$. Similarly for the definitions of $\boldsymbol{\alpha}_{J-1-p,2}$, $\mathbf{w}_{J-1-p,2}$, $\boldsymbol{\mu}_{J-1-p,2}$.

Figure 7b shows a curve smoothed by local regression and wavelet aapproach.

3. The comparison of different approaches

Although the true distribution of weight gained in NMIHS is unknown, medical knowledge shows that the change in weight of women after pregnancy forms a smooth continuous curve with one peak. Table 3.1. shows some of the statistics between the adjusted data and the conclusions from some medical literatures.

Table 3.1. The comparison of different approaches in NMIHS data
(By medical literature: mean = 27.5)

	Bias	Smoothness
(1)	-1.52	N/A
(2)	-1.05	poor
(3)	-0.57	good
(4)	-2.63	good
(5)	1.46	good
(6)	2.33	middle
(7)	1.63	good

Table 3.2. The comparison of different approaches in simulation
(Population mean=24.57)

	Bias	χ^2
(1)	-1.56	1.21
(2)	0.44	1.37
(3)	-0.09	143.9
(4)	-1.22	14.61
(5)	-0.09	18.47
(6)	1.30	45.48
(7)	1.07	59.43

Computer simulation also provides a way to assess different approaches of adjustment. First, we create a data set of known distributions; a Gamma distribution in our simulation. Then based on the statistical model for representation with rounding and bounding errors in l.d., we generate a set of data with rounding and bounding errors. Finally, we apply different smoothing approaches to the data. Table 3.2. shows statistics of bias and χ^2 when compared with the original

distribution. In Tables 3.1 and 3.2, the different approaches of adjustment are: (2) moving average, (3) Gamma distribution, (4) truncated Gamma distribution, (5) local regression, (6) wavelet, (7) wavelet+local regression; and (1) raw data with rounding and bounding errors.

Consistent with the conclusions in Table 3.1., the moving average is the basic step of smoothing. Local regression gives the best results while wavelet shrinkage improves the adjustment from moving average results. The probability model could work very well, but it depends on specific data.

References

Antoniadis, A. and Oppenheim, G., 1995, Wavelets and Statistics, Springer, NY.

Ash, S., et al., 1989, Maternal Weight Gain, Smoking, and Other Factors in Pregnancy as Predictors of Infant Birth weight in Sydney Women, Aust NZJ Obstet Gynaecol, 29: 3(1), 212-219

Donoho, D. L., and Johnstone, I. M., 1995, Adapting to Unknown Smoothness Via Wavelet shrinkage, JASA, 90, 1200-1224.

Fisher, R. A., 1934, The Effect of Methods of Ascertainment upon the Estimation of Frequencies, Ann. Eugen., 6, 12-25.

Heitjan, D. F., and Rubin, D. B., 1990, Inference from Coarse Data via Multiple Imputation with Application to Age Heaping, JASA, 85, 304-314.

Huttenlocher, J., Hedges, L., and Bradburn, N., 1990, Reports of Elapsed Time: Bounding and Rounding Processes in Estimation, Journal of Experimental Psychology: Learning, Memory, and Cognition, Vol 16, No. 2, 196-213.

Lehmann, E. L., 1975, Nonparametrics, Holden-Day, Inc. San Francisco.

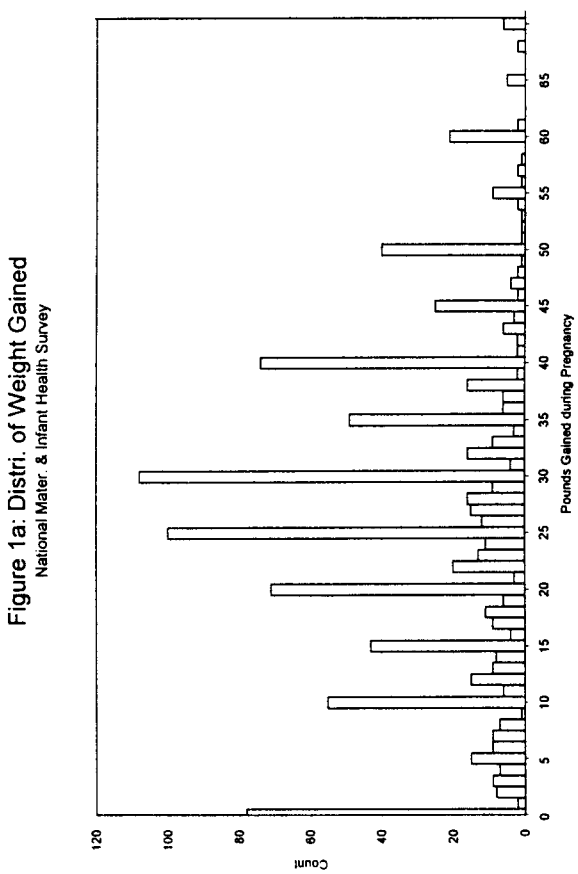
Mahalanobis, P.C., 1946, Recent Experiments in Statistical Sampling in the Indian Statistical Institute, Journal of the Royal Statistical Society, 109, 326-70.

Neter, J., & Waksberg, J., 1964, A study of Response Errors in Expenditures Data from Household Interview, JASA, 59, 18-55.

Rao, C.R., 1985, A Weighted Distributions Arising Out of Methods of Ascertainment: What Population Does a Sample Represent? In Atkinson and Fienberg, S.E., A Celebration of Statistics, The ISI Century Volume, Springer, NY.

Rossner, S. & Ohlin, A., 1995, Pregnancy as a Risk Factor for Obesity: Lessons from the Stockholm Pregnancy and Weight Development Study, Obesity Research, Vol 3 Suppl 2, 267-275.

Sudman, S., and Bradburn, N., 1973, Effects of Time and Memory Factors on Response in Surveys, JASA, 63, 805-815.



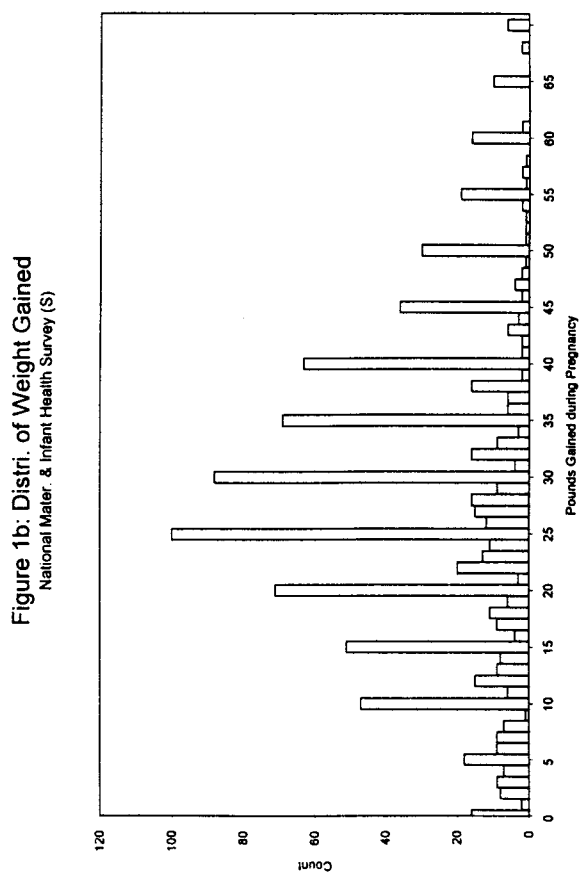


Figure 6a: Distri. of Weight Gained
Approach: Local Regression (S+m5x1+loc)

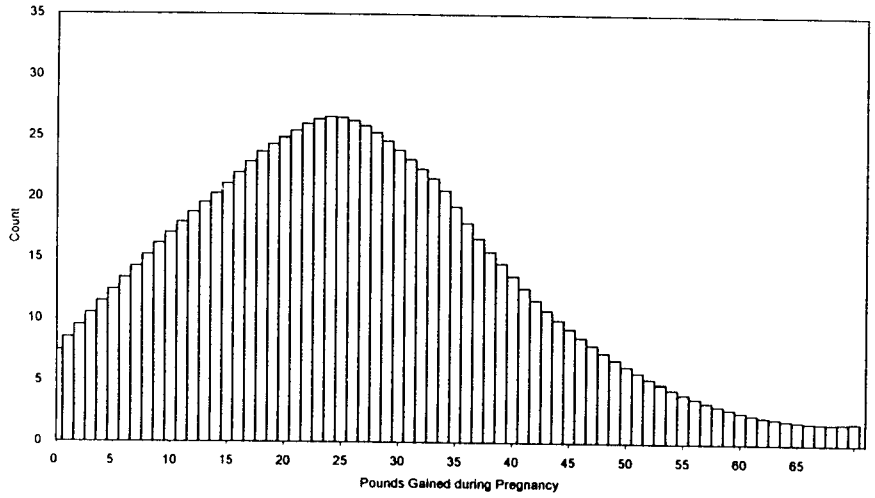


Figure 2a: Distri. of Weight Gained
Approach: Moving Average (S+m5x1)

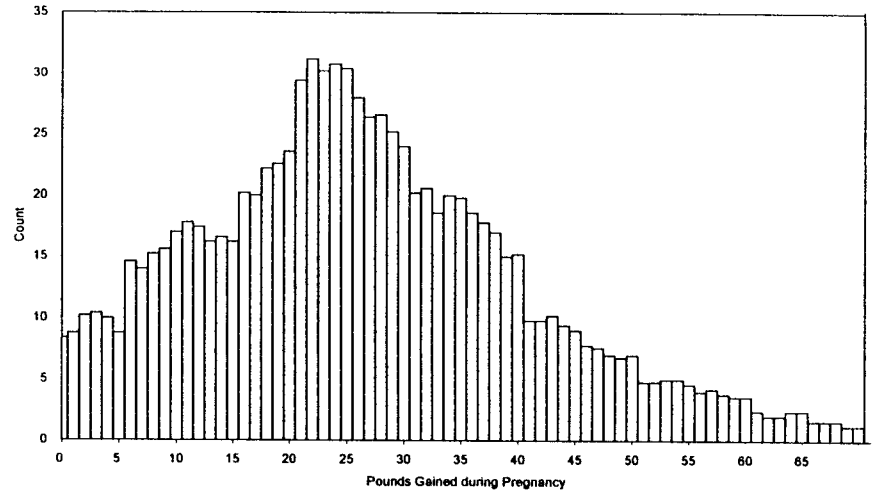


Figure 7b: Distri. of Weight Gained
Approach: Wavelet+Local Regression

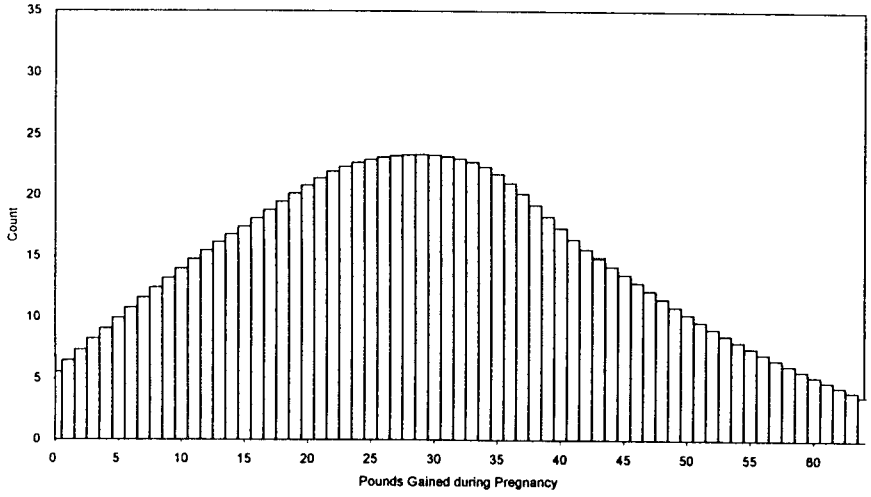


Figure 5a: Distri. of Weight Gained
Approach: Truncated Gamma distribution