Soy Chip Data: Examination for Anomalies

Mark S. Kaiser and Alicia L. Carriquiry

Department of Statistics

Iowa State University

March 2009

1 Background

This report contains a statistical examination of data from a study titled "Efficacy of Soy Pasta Chips for Weight Loss", conducted in 2004 at the Fleming Heart and Health Institute of Omaha, Nebraska. Our understanding is that questions have been raised about the authenticity of the data produced by that study and, specifically, whether some of those data may have been fabricated. Statistical examination of a set of data cannot "prove" or "disprove" falsification of data records in an absolute sense, but it can determine whether certain types of anomalies exist that would not be expected in data from most scientific studies. The goal of this exercise was to uncover any such anomalies that might exist in the data from this study.

The data used in this analysis were taken from a final report signed by the principal investigator on 7 April 2004 and provided to us via electronic transmission by Dr. Richard Fleming. The data contain records for 60 individuals that consist of values for height, initial weight, weight at two weeks, weight at four weeks, and body mass index at the same time points as weight. Our examination of these data makes use of only the directly recorded variables of height and the three weight measurements.

2 Methods of Examination

Appropriate statistical methods for examination of data to detect potential fabrication depend on the characteristics of the study or studies of concern, including study design, objectives, and the analysis used to reach conclusions. Also important is the type of data fabrication suspected. The best methods for detection of one or a few fabricated data records differ from those more appropriate for the detection of wholesale fabrication of an entire or nearly an entire data set (e.g., Buyse *et al.* 1999). The study of concern here was of a very simple design with apparently self-selected subjects and lacking multiple medical centers or treatment groups, precluding the use of comparison of multiple centers or a suspect data set to an unsuspicious one (e.g., Al-Marzouki *et al.* 2005). The examination reported here focused on three aspects of the data records, *marginal and joint data structure*, *recorded data values*, and *influence on results*. The motivation for considering these aspects of the problem are described in this section.

Fabrication of data generally has a specific objective, either to influence the outcome of data analysis (e.g., show an effect of one or more treatments) or to avoid the effort needed to properly conduct data collection if a pattern seems clear from an analysis of some actual data. The former situation may result in alteration of one or more data records that have disproportionate influence on the outcome of statistical analysis for the study. Alternatively, if an entire data set is fabricated to exhibit an effect of some type (e.g., a difference in treatment group means), other characteristics of typical data sets that might also show such an effect (e.g., variance or covariance structure) are difficult to match. That is, most scientists cannot *preserve* higher-order structure in falsified data while achieving the desired first-order differences (Haldane 1948). The fabrication of data records as a matter of convenience may sometimes be detected based on either the number or distribution of digits in recorded data (e.g., Walter and Richards 2001). For example, the presence of "extra" digits in recorded data may indicate that other, possibly legitimate, records have been averaged to produce the falsified data, or a fabricated data set may contain a preference for certain digits in either the first or terminal places. This latter phenomenon is related to the fact that the human mind is a poor random number generator.

While a comparable data set from an undisputed study is not readily available for this analysis, it is possible to make use of theoretical probability distributions for comparison with the Fleming data set. Simulation of random values from theoretical probability distributions can be used to describe the expected behavior of actual data. Serious departures from such behavior are then a signal that something may be amiss in a given set of values. The Soy Chip study resulted in a four-dimensional multivariate observation for each subject, height, weight 0, weight 1, and weight 2. Assuming (which can be reasonably verified for the Fleming data) that a multivariate normal distribution provides a good model for the joint data characteristics, simulated values from this distribution can be used to examine what might be expected in terms of recorded data values (e.g., terminal digits) and whether or not averaging results should appear in randomly generated data.

3 The Effects of Self-Selected Samples and Study Design

Before proceeding to examination of the actual data values it is helpful to identify the effects of two inter-related aspects of the study under discussion, those being the self-selection of subjects who participated in the study, and the lack of a study design that would allow conclusions about the effect of soy chips on weight loss.

3.1 Self-Selection and "Expected" Data Values

A common-sense and, in fact, quite legitimate source of suspicion about a set of data values is if they appear "too good to be true". In some respects that seems to hold for the data under examination here. All but one of the subjects lost weight, at an average of over 3 pounds per two-week period. The "treatment" of soy chips appears, at first glance, to have produced remarkable results in terms of weight loss, perhaps as much as many of us could rationally attribute to some type of active weight loss regimen, yet alone simply an alternative snack food. The reason, in this case, that this aspect of the study fails to provide evidence that something is amiss in the data is that the subjects were self-selected for participation. There is little information in the final project report as to how subjects were recruited for the study, other than that being "motivated to lose weight" was one of the inclusion criteria. Put simply, it is not known what else these subjects were doing in attempts to lose weight. It is not known if they became aware of the study through social contact with other study participants (e.g., at health clubs or gyms or weight loss support groups), what characteristics they may have or may have not shared in common, what their overall diets consisted of, what they may have stopped eating if they ate soy chips instead, how much physical activity they engaged in, or any of a host of other factors that might have influenced the results of the study. It is not even known if they actually consumed the soy chips. The implication is that no results can be considered too "good" or too "bad" to be true, unless those results are outside of the realm of physical possibility.

An additional effect of the self-selection of subjects for the study is that comparisons of marginal distributions of, for example, height and initial weight, with what might be expected from a random sample of the population at large lose all meaning. It should not necessarily be expected that the distributions of height or initial weight are consistent with a random sample from the population, or that the relation between height and initial weight should be consistent with such a sample. In fact, it might be argued that, because the subjects in the study were "motivated to lose weight", the relation between height and initial weight should differ from the population at large, containing a greater proportion of individuals with weight on the higher end of the distributions of weight for a given height. There is, in this case, simply no appropriate "reference group" against which to compare the data values.

3.2 Study Design

Although the final study report described the study as a "randomized clinical trial", there exist no factors that could have been randomized. There was no control group (i.e., individuals not given soy chips), nor any alternative treatment group (e.g., apples as snacks); note that a true placebo (i.e., "fake" soy chips) or double-blind study design would be difficult to devise in this setting. Because it is not known what behaviors subjects were engaged in during the study, and there is no group that could be reasonably expected to have engaged in the same behaviors other than the consumption of soy chips, there is no scientific evidence that soy chips had any effect at all on the results. From a scientific viewpoint, the evidence in the data that being handed free soy chips promotes weight loss is exactly the same as the evidence that consuming soy chips promotes weight loss. In fact, the evidence that visiting the Health and Heart Institute to be weighed promotes weight loss is exactly the same as the evidence that consumption of soy chips promotes weight loss, which is to say there is no scientific evidence for any of these conclusions.

In the opinions of the authors of this report, a manuscript based on this study would not be accepted for publication in any scientific journal. The potential relevance of this for the current exercise is in terms of the objective of data fabrication in this study, as mentioned in the previous section. Any individual with an advanced degree in a scientific discipline, or at the very least any individual who has successfully published a paper in a scientific journal, should understand the shortcomings of the study design. The objective of data fabrication in this situation would then almost necessarily be one of convenience rather than career development, that is, simply to avoid the time and energy needed to collect actual observations. Common sense, as well as formal logic, then suggests that the time and energy needed to falsify data records should be no more than (and, given the anathema with which data fabrication is viewed in the scientific community, perhaps even considerably less than) the time and energy needed to make actual observations. The only alternative I can think of regarding motivation for data falsification in this situation is pleasing a client from whom a substantial amount of additional funding is anticipated but, if it becomes clear to the client that the study was largely without scientific merit because results cannot be published, this would seem a remote possibility.

4 Marginal and Joint Data Structure

The first approach used in this exercise was to examine the marginal and joint data structures for the entire set of data. This examination might indicate the presence of records that were altered in a manner that failed to preserve the overall coherence (or general behavior) of the collection of data in a manner consistent with typical probabilistic rules. For example, if a number of records were falsified for a particular weight (e.g., weight2 at week 4) they might stand out as having a different relation with height than they did at an earlier stage (e.g., weight1 at week 2). If entire data records were falsified the relation among variables in those records (ht, wt0, wt1, wt2) may not follow the overall pattern of the set of data. In a sense, then, this examination is one of *data consistency*. An individual falsifying a few data records would need to take care that those records "fit" the general pattern in the entire data set. An individual falsifying the bulk of records or fabricating an entire data set would need to take care that those records were both biologically consistent and probabilistically consistent. Probabilistically consistent here means that there should exist some joint probability distribution that could have "generated" the observed data. While no theoretical probability distribution is "correct" in a real problem, real data tend to follow the patterns of data simulated from theoretical distributions and dictated by the rules of probability. Falsified data often fail to exhibit this same consistency (unless, of course, they were produced via simulation from theoretical probability distributions). Basic summary statistics for the Fleming data set are presented in Table 1.

There is a basic consistency in these summary values. Variances for the three

Variable	Min	Q1	Q2	Q3	Max	Mean	Variance
Height	60.50	63.94	66.00	68.44	76.00	66.32	10.439
Weight0	146.0	165.1	185.0	205.5	301.0	193.71	1409.587
Weight1	139.0	162.2	182.5	201.6	295.0	189.76	1370.250
Weight 2	128.5	159.5	179.0	199.0	293.0	186.41	1357.250

Table 1: Basic summary statistics for the Fleming data.

weight values are not dramatically different, and a decrease in weight is seen for various quantiles in a consistent manner. There is perhaps a surprising difference between the minimum weight at times 2 and 1, that difference being -10.5 pounds. The minimum weights at times 1 and 2 correspond to the same individual. To see if this should be considered an extreme value, inconsistent with the overall data structure, a set of data was simulated from a multivariate normal distribution with means and variances that match the values of Table 1. The minimum values for "weight1" and "weight2" in these simulated data also corresponded to the same data record (i.e., the same simulated "individual") and the difference was -9.7. While not demonstrating that the one actual data record could not have been fabricated, this does demonstrate that the occurrence would not be unexpected under a typical probabilistic structure of the kind used to model data.

Correlations among the variables of height, weight0, weight1 and weight2 are reported for the Fleming data in Table 2. Extremely high correlations (for which the values of correlations between weight0, weight1, and weight 2 would qualify) are sometimes taken as an indication of results "too good to be true" (e.g., Akhtar-

Danesh and Dehghan-Kooshkghazi 2003). But that is a weak argument against the Fleming data set in this case, partially because of the self-selection of study participants as discussed in the previous section of this report, and partially because of a combination of the ranges for weight measurements in Table 1 and the physiological realities of how much weight an individual can gain or lose in a period of several weeks. Correlation is a measure of linear association between two variables and this measure is affected by the range of values considered. A wide range of initial values (e.g., a range of 155 lbs. in weight0 for comparison with weight1 or a range of 156 lbs in weight1 for a comparison with weight2), coupled with the biological reality that any individual is unlikely to lose or gain more than a small fraction of their initial value *relative to the initial range* indicates that high correlations are to be expected in this situation. The Fleming data are also consistent with the anticipation that weights observed at more distant time points (i.e., weight0 and weight2) should still be correlated, but somewhat less highly correlated than weights observed at less distant time points (i.e., weight0 and weight1).

	ht	wt0	wt1	wt2
ht	1.0000000	0.5263469	0.5274059	0.5289093
wt0	0.5263469	1.0000000	0.9989028	0.9961254
wt1	0.5274059	0.9989028	1.0000000	0.9983947
wt2	0.5289093	0.9961254	0.9983947	1.0000000

Table 2: Correlations for the Fleming data.

Scatterplots of weights at times 0, 1 and 2 against height are presented for the Fleming data in Figure 1. The first thing to note here is the similarity of the three scatterplots. This should be expected, again because of the total range of weights contained in the data sets and the physiological realities of how much weight can change for humans over a period of several weeks. It appears that one could pick out individuals on these plots and that is, in fact, true. What would be disturbing would be to find individuals with radically different positions on one or more of the three plots, and that does not occur. One may also notice that there are more widely scattered points above the bulk of the data pattern than there are below. This is not necessarily unexpected, because the self-selected sample of participants were individuals who considered themselves overweight. Statistically, this data pattern suggests distributions of weight for given heights that are skew right rather than symmetric.

Overall, there is little in the set of data values examined to suggest that they could not be the result of a study with an absence of fabricated data. The data values may be considered as *internally consistent*. At this point we would have no justification for suggesting that the data have been manipulated in a manner consistent with the falsification of data. Examination of data in the manner of this section, however, is not a powerful approach for identification of anomalies because of the lack of a reference for comparison. As indicated previously, the population as a whole will not serve this purpose because subjects in the Fleming study were not intended to be a random sample from the population, and we lack data from a comparable undisputed study for comparison as well. What we can say is that the data set fails to contain obvious glaring inconsistencies that would suggest fabrication of data.

5 Recorded Data Values

Any numerical data value consists of a sequence of digits. For example, the value of 156 for an initial weight in this study has the digits 1, 5 and 6, in that order. There are two common approaches for examination of recorded digits in data records – investigation of recorded values that contain "extra digits", and comparison of distributions of the values 0 through 9 in various places in the data (e.g., first digit or last digit). We consider these two approaches in turn.

5.1 Records with Extra Digits

The majority of the data contained in the Fleming data set are recorded to the nearest whole number (i.e., height to the nearest inch, weight to the nearest pound) but there are a number of records that contain extra digits of either 0.25, 0.5 or 0.75. Table 3 presents the frequencies of these extra digits for the four observed variables.

Extra Digits	Height	Weight0	Weight1	Weight2
0.25	5	0	0	0
0.50	9	11	9	3
0.75	4	0	0	0

Table 3: Frequency of extra digits in the Fleming data.

Data records with extra digits relative to the majority of the data may indicate that other data records were averaged to produce the suspect record (e.g., Walter and Richards 2001). For example, if two records with weights of 174 and 177 are averaged the result is 175.5, and the extra digit is easily recorded by an individual falsifying data. Of course, the mere presence of extra digits in some records does not necessarily indicate the record was constructed, but in the absence of falsification it would be unusual for one (entire) record to be the average of two others, even more unusual for this to be true of two records, and so forth. In the Fleming data there are four variables, giving rise to four possible places where data averaging may have occurred to produce false data. A computer function was written (see Appendix 1) that took each record with extra digits for height and compared values of the four variables to averages of all other unique pairs of records (of which there are $59(58)/2 = 1711$). Each instance in which any of the variables in the "suspect" record with extra digits was found to correspond to the average of two other records was saved. Of the 18 suspect records in the Fleming data, pairs of other subjects were found such that the average of exactly one variable in those records matched the value in the suspect record in 17 cases. For 12 of the suspect records pairs of other subjects could be found that, when averaged, produced the values in the suspect record for exactly 2 variables. But for none of the suspect records was it possible to locate a pair of other subjects that when averaged produced 3 or all 4 of the variables in the suspect record. The results for suspect records having at least two variables equal to the average of other records are presented in Table 4. In this table, the column labeled "suspect" gives the subject number from the original data corresponding to a data record having extra digits for height. The columns labeled "other 1" and "other 2" give subject numbers from two other records that were found to average to the suspect record value for two or more of the variables. The column labeled "nflags" gives the number of variables (out of the 4 possible but at least 2) for which the two other records produced averages equal to what was reported for the suspect record, and the columns labeled "flag1" through "flag4" give the specific variables for which averages matched the value of the suspect record (flag1=height, flag2=weight0, flag3=weight1 and flag4=weight2).

There are several aspects of the results in Table 4 that are of interest.

1. Note first that there are quite a few of the records with extra digits for height (12 out of 18 to be exact) that have at least two variables equal to the averages of two other records in the data set, but there are none that have all four variables equal to the average of two other records.

2. Curiously, many of the suspect records in Table 4 contain variables that have values equal to the average of more than one pair of other records (e.g., suspect

suspect	other1	other2	nflags	flag1	flag2	flag3	flag4
1	17	28	2	1	0	1	0
1	17	33	2	0	1	1	0
1	28	55	2	0	1	0	1
1	34	36	2	0	1	1	0
2	12	28	2	0	1	0	1
2	27	30	2	0	0	1	1
2	27	58	2	0	0	1	1
6	24	48	2	0	1	1	0
6	42	48	2	0	1	1	0
8	6	10	2	0	1	1	0
8	9	28	2	0	1	0	1
8	38	48	2	0	1	1	0
8	50	59	2	0	1	1	0
10	34	55	2	1	0	0	1
11	53	55	2	0	1	1	0
13	25	40	2	0	1	0	1
22	44	55	2	0	1	1	0
26	17	29	2	0	0	1	1
28	3	33	2	0	1	1	0
28	27	56	2	0	0	1	1
28	27	59	2	0	1	1	0
28	41	60	2	0	0	1	1
28	50	59	2	0	1	0	1
28	53	58	2	1	0	0	1
34	25	60	2	0	1	1	0
34	26	39	2	0	1	1	0
34	39	49	2	1	0	0	1
35	12	43	2	1	0	1	0
35	12	59	2	1	1	0	0

Table 4: Suspected data records with

record 1, 2, 6, 8).

3. The number of suspect records that have values equal to averages of other records seems more prevalent for weight variables than for the variable of height.

4. There are no suspect records that are are the same in total (i.e., for all four variables) to averages of other records. In fact, there does not appear to be a simple pattern for which variables are averages of other records. For example, subject numbers 17 and 28 as well as subject numbers 17 and 33 average to the value of weight1 for subject number 1. Subject numbers 17 and 28 also average to the height value for subject 1, but subject numbers 17 and 33 do not, while subject numbers 17 and 33 average to the value of weight0 for subject 1 but subject numbers 17 and 28 do not.

Overall, the results of Table 4 indicate that, if the suspect records with extra digits for height in the Fleming data were constructed using a process of averaging other data records, this was done according to some complex system that is difficult to uncover. For example, subject 1 had matches (i.e., flags) that involved subject numbers 17, 28, 33, 55, 34 and 36. The record for subject 1 was not a match for the average of any 3 of these other records (of which there are 20), any 4 of these records (of which there are 15), any 5 of these records (of which there are 6) or all 6 of the records. The number of instances in which some variables in the records for which height contained extra digits turn out to be equal to averages of other records is, however, curious.

To examine whether or not the phenomena of Table 4 should be considered “out of the ordinary”, we compared the results given in that table with data generated randomly from a coherent probabilistic structure. To accomplish this, 60 records were simulated from a four-dimensional multivariate normal distribution with means, variances, and covariances equal to the realized values from the Fleming data set. This data set, then, was simulated to match the marginal and joint data structures of the Fleming data set, but to be a case in which other aspects of the data followed a typical probabilistic structure difficult for humans to duplicate if asked to purposely falsify data (this entire simulated data set is contained in Appendix 2). The four variables in the simulated data will be called height, weight0, weight1 and weight2, in analogy with the actual problem. Each simulated record was then rounded to the nearest whole number. Following the frequencies of Table 3, 18 values for the variable height were randomly selected to have an extra digit added to their values; to 5 records the value of 0.25 was added, to 9 records the value of 0.50 was added, and to 4 records the value of 0.75 was added. In addition, 11 records were randomly selected to have a value of 0.50 added to weight0, another 9 records randomly selected to have a value of 0.50 added to weight1, and 3 records were randomly selected to have a value of 0.50 added to weight2. Running these simulated data through the same computer function used to produce Table 4 from the Fleming data gave the results presented in Table 5.

Although there is a minor difference between the values of Table 5 and those from the Fleming data of Table 4 (i.e., 7 of the 18 "suspect" records in the simulated data matched averages of other records in 2 or more variables, while 12 of 18 did for the Fleming data) the patterns are remarkably similar. In fact, the second, third, and fourth characteristics of the data in Table 4 listed previously, which may have seemed suspicious, were reproduced nearly identically in the simulated data results of Table 5.

Neither Table 4 nor Table 5 report the number of "suspicious" records matching averages in only 1 of the four variables. A table of frequencies for the number of suspicious records (out of 18 for both the Fleming and simulated data) that had 1, 2, 3, or 4 of the variables height, weight0, weight1, and weight2 matching averages

suspect	other1	other2	nflags	flag1	flag2	flag3	flag4
25	16	58	2	0	1	1	0
33	11	58	2	1	1	0	0
34	15	57	2	0	1	1	0
34	17	57	2	1	1	0	0
34	49	58	2	0	1	0	1
39	1	50	3	0	1	1	1
39	2	57	2	0	1	1	0
39	32	35	2	0	0	1	1
42	5	24	2	0	1	1	0
42	22	35	2	0	0	1	1
42	28	49	2	0	1	0	1
42	37	38	2	0	0	1	1
50	1	30	2	0	1	0	1
59	25	34	2	0	1	0	1

Table 5: Data records in a simulated data set with heights recorded with extra digits for which variables were found to equal averages from two other records.

of pairs of other data records is presented in Table 6. An ordinary Chi-squared test of differences for these frequencies is not appropriate here as the entries in Table 6 are not independent (i.e., a given suspicious data record could have matches with multiple pairs of other records, some pairs matching 1 of the variables and other pairs matching 2 of the four variables). In addition, only one simulated data set is presented and other simulated data sets would vary from this one to some degree. The point of Table 6, however, is that it does not appear that the Fleming data are at all unusual compared to what might result from a completely random probabilistic mechanism with the same marginal and joint data characteristics. The only conclusion that seems plausible is that the patterns exhibited in the Fleming data and reported in Table 4 are entirely in concert with what might occur from a completely probabilistic structure matched to the marginal and joint structures of those data.

	No. of Variables			
Data Set	1	2	3	4
Fleming	17	12	0	0
Simulated	14	7	1	0

Table 6: Frequency of matches for "suspicious" data records with averages of other pairs of records for the Fleming and simulated data sets.

5.2 Distributions of Digits

There exist demonstrated distributions for the frequencies with which different digits (0 through 9) appear in data from various sources. There is a result known as *Benford's law* that indicates the relative frequencies of leading digits in data should follow an approximate logarithmic distribution (e.g., Buyse *et al.* 1999, Hill 1998).

This approximation often applies to financial data and other data consisting of an aggregation of various sources but does not typically apply to scientific data from a single data source (e.g., Hill 1998). In fact, a demonstration that Benford's law corresponds to a coherent probabilistic structure made use of random digits selected from random distributions (Hill 1996), a context that does not apply to most scientific investigations. The emphasis put on Benford's law by, for example, Buyse *et al.* 1999 seems misplaced, except perhaps in the examination of financial records for medical facilities.

The other use of distributions of digits in data to detect anomalies rests on the assumption that recorded data values may contain meaningful and nonmeaningful digits. The leading (first) digits of data values are often meaningful in indicating the magnitude of responses. The trailing (last) digit or digits are often nonmeaningful in this regard. For example, in a weight difference of 190.3 and 185.6 pounds, the first three digits of 190 and 185 are more meaningful than are the trailing decimal digits of 3 and 6. It is often assumed then that the meaningless digits should follow a uniform distribution on the discrete integer values from 0 to 9 (e.g., Walter and Richards 2001). Because the human mind appears to be a poor random number generator, fabricated data may often show a distribution of meaningless digits substantially different from a uniform distribution. But, as pointed out by O'Kelly (2004), data with non-meaningful trailing digits do not occur in many clinical trials, and that is the case here, except for perhaps the data records with extra recorded digits which have already been examined in the previous subsection. Hill and Richards (2002) also point out a number of potential pitfalls in testing digits, particularly in the absence of an unquestioned reference data set. It may remain true, however, that the last digits in a fabricated data set (even with most records recorded to the nearest whole number) would be difficult for a human to match to a probabilistic structure. Thus, if there is a use to be made of examining the distributions of digits in the Fleming data it would involve the final whole digit.

In order to demonstrate what an examination of trailing digits would suggest about the Fleming and simulated data sets, a computer function was written to give the frequency of final digits (as whole numbers – data records containing extra digits first had those digits removed) for each of the variables of height, weight0, weight1, and weight2, and to test the resultant empirical distributions against a theoretical uniform distribution. The results for the Fleming data are presented in Tables 7 and 8.

Digit	ht	wt0	wt1	wt2
0	6	8	7	8
1	5	4	2	5
2	7	4	3	5
3	6	5	6	6
4	4	7	8	6
5	8	6	7	9
6	7	4	9	3
7	6	5	7	7
8	6	10	4	5
9	5	7	7	6

Table 7: Observed frequencies of final digits in the Fleming data.

Under an assumption that the relative frequencies of final digits (0 through 9) should follow a uniform distribution, the expected frequency for each digit is, with 60 observations $60/10 = 6.0$. Standard Chi-squared tests of goodness of fit for such a uniform distribution to the values in Table 7 yields the results of Table 8. Clearly, none of the variables contain distributions of final digits having evidence of departure from a uniform distribution.

Variable	Test Statistic	$p-$value
Height	2.00	0.9915
Weight0	6.00	0.7399
Weight1	7.67	0.5680
Weight2	4.33	0.8881

Table 8: Test statistics and associated $p-$values for testing that the frequencies of final digits in the Fleming data differ from a uniform distribution.

Repeating this exercise with the data simulated from a multivariate normal distribution yields the observed frequencies of Table 9 and the associated test statistics and $p-$values of Table 10. These simulated data, as they should, also offer no evidence of a departure from a uniform distribution of final digits for any of the four variables.

Digit	ht	wt0	wt1	wt2
0	5	2	7	7
1	6	12	4	4
2	5	7	7	9
3	6	5	4	3
4	4	4	5	11
5	8	6	5	5
6	9	5	8	8
7	8	7	8	8
8	4	5	7	3
9	5	7	5	2

Table 9: Observed frequencies of final digits in the simulated data.

Variable	Test Statistic	$p-$value
Height	4.67	0.8623
Weight0	10.33	0.3242
Weight1	3.67	0.9320
Weight2	13.67	0.1345

Table 10: Test statistics and associated $p-$values for testing that the frequencies of final digits in the simulated data differ from a uniform distribution.

The upshot of this subsection is that, in the first place, the examination of of the Fleming data for assumed distributions of digit values in leading digits is problematic on theoretical grounds, although it is less so for trailing digits. Weights should not have leading digits less than 1 for overweight individuals (i.e., less than 100 pounds) and would be unlikely to have leading digits greater than 3. The distribution of final or trailing digits should not follow Benford's law because they do not correspond to observations from multiple sources. An examination of the Fleming data demonstrates that the distribution of trailing digits appears entirely consistent with what would be expected from a coherent probabilistic structure, which the simulated data are know to follow.

6 Could the Fleming Data Be Simulated?

The agreement of the Fleming data with values simulated from a multivariate normal distribution in terms of the averaging phenomena discussed in section 5.1, and the distribution of trailing digits in Section 5.2, raises the question of whether the data could have been produced wholesale (i.e., in entirety) from the use of a random number generator. The most likely candidate for such simulation would be a multivariate normal distribution with marginal and joint characteristics equal to the means, variances, and covariances reported for the Fleming data and described in Section 4 of this report. Given a moderate amount of statistical sophistication, many individuals could produce such a data set. That this is unlikely to be the case in the current situation is evidenced by the failure of marginal distributions of weight0, weight1, and weight2 to follow univariate normal distributions. A known property of multivariate normal distributions is that the marginal distributions corresponding to individual variables are univariate normal in form. Figure 2 presents histograms of the marginal distributions of weight0 for the simulated data set in the upper panel and the Fleming data set in the lower panel. The simulated data (upper panel) exhibit a distribution consistent with a normal theoretical distribution, which they should. The Fleming data (lower panel) exhibit a distinct skew right distribution, consistent with the observation of the scatterplots of weight versus height in Figure 1 (see Section 4 of this report). Is it possible to simulate data that have the characteristics of the Fleming data set? The answer is yes, it is possible, but doing so would require the ability to preserve means, variances, and correlations as described in Section 4 of this report, preserve the averaging property described in Section 5 of this report, **and** produce the difference in marginal distribution of weights at time 0 given in Figure 2. There exist ways to achieve all of this but they require a relatively high level of statistical knowledge, including the time and ability to write computer functions for tasks that are not readily available in pre-packaged routines.

7 Influence on Results

Falsification of data often has the objective of producing certain results in a data analysis. Quantification of the *influence* of each observation on the resultant analysis can then sometimes highlight one or a group of observations that played a large role in determining the outcome and conclusions of a study. While not in any manner evidence of falsified values by themselves, the occurrence of high influences can suggest cases worthy of additional examination. In the report on results of the Fleming study provided to us, the analysis consisted of two paired t-tests, one conducted on the difference in weight0 and weight1 values and the other conducted on the differences in weight1 and weight2 values. To examine the influence of recorded data values on these tests observations were deleted one at a time from the data, the test statistic recomputed without that value, and the difference (absolute value) of that deleted-case statistic with the test statistic computed using the entire data set was calculated. This value then provides an indication of the influence of individual observations on the test conducted with the entire set of values. A summary of the influence values produced using the Fleming and simulated data for the comparison of weight0 and weight1 values is presented in Table 11, and the same is reported for the comparison of weight1 and weight2 values in Table 12.

Data Set	Min	Q1	Q2	Q3	Max
Fleming	0.0223	0.1758	0.2461	0.3079	2.8390
Simulated	0.0211	0.1309	0.2784	0.3265	0.9403

Table 11: Summary of influence values for comparison of weight0 and weight1 records.

Data Set	Min	Q1	Q2	Q3	Max
Fleming	0.0111	0.1564	0.1833	0.2376	1.306
Simulated	0.062	0.1794	0.2491	0.2818	0.5538

Table 12: Summary of influence values for comparison of weight1 and weight2 records.

The most notable feature of both Table 11 and Table 12 is the extreme distance between the third quartile (or 75%−tile, denoted Q3) of influence values and the maximum influence value for the Fleming data. Stem and leaf plots demonstrate that this is due to only one extreme value that is hugely separated from the reamainder of the data. For example, the influence values for the Fleming data of Table 11 have the following stem-and-leaf plot:

```
The decimal point is at the |

0 | 0000001111222222222222222222222233333333333333333333333333
0 |
1 |
1 |
2 |
2 | 8
```

The data record that corresponds to the single observation with influence value 2.8 (which is just over 9 times larger than the next larges value) corresponds to subject 52 having height= 66, weight0= 186, weight1= 189 and weight2= 192. This subject gained weight between each weighing. The result is that, while highly influential relative to any of the other data records, the results for this subject decreased the size of the test statistic and hence the significance of the overall findings of the study. If this record was falsified the only reasonable objective would have been to purposely introduce one outlier into the data to make it look more "real", not to produce a desired result in the analysis of the study. This same observation is also the one extreme influence value for the Fleming data from Table 12.

8 Conclusions

As stated in the opening paragraph of this report, a statistical examination of data cannot definitively prove or disprove the falsification of data records. The analysis conducted in this report, however, does allow the following conclusions to be reached.

1. If the Fleming data were falsified it would appear that they were fabricated in a nearly wholesale fashion, that is, more-or-less in total. These data are internally consistent, consistent with the behavior of values simulated from a theoretical probability distribution, and there is only one data record with undue influence on the results of the study (and this influence was in the "wrong" direction).

2. Because of the properties listed in conclusion 1 and, in particular, the averaging behavior described in Section 5 that the Fleming data shared with simulated data , the most likely mechanism for fabrication in this study must be considered simulation from some theoretical probability model.

3. Because of the multivariate nature of the four recorded data values for each subject, maintaining internal consistency would require, or at least strongly suggest, that a multivariate probability distribution would need to have been employed to simulate data values. The candidate most readily available to non-statisticians (and even to statisticians without extensive experience in the construction of multivariate distributions from other probability structures) is the multivariate normal distribution.

4. The marginal moments (means, variances) and joint moments (covariance or correlation) of the Fleming data could easily be maintained through simulation from a multivariate normal distribution. However, the skew shape of marginal weight distributions (e.g., Figure 2) could not.

5. Combining items 1 through 4 immediately above suggests that, if the Fleming data were fabricated, the procedure used to arrive at the reported values was necessarily complex, requiring considerable statistical expertise and time to conduct. If it were supposed that the most likely motivation for data fabrication in this situation was to save time and effort relative to actually performing the observational process, this would seem at odds with what would have been needed for fabrication of the data.

Overall, there is simply no data-driven evidence that the Fleming data set is other than would be expected under a legitimate study. Haphazard falsification would be unlikely to result in the consistency of behavior these data show with with a coherent (probabilistically consistent) probability distribution. One would need to simultaneously preserve the proper behavior of 6 correlations (ht-wt0, ht-wt1, ht-wt2, wt0-wt1, wt0-wt2, wt1-wt2), averaging properties, and distribution of trailing digits. That all of this could be produced by a human with a pencil and paper would be unusual, although it could result from simulation using the proper computer package. Simulation from a multivariate normal distribution could control all of these characteristics, which to statisticians depend on what are called the *first two moments* of distributions, in a consistent manner. Achieving this while at the same time producing marginal distributions that depart from those associated with a multivariate normal distribution in shape, which depends on what statisticians call *higher-order moments* would require considerable statistical ability as well as considerable time and effort.

9 Literature Cited

Akhtar-Danesh, N. and Dehghan-Kooshkghazi, M. (2003), How does correlation structure differ between real and fabricated data-sets? *BioMed Central Medical Research Methodology* **3**, 18-26.

Al-Marzouki, S., Evans, S., Marshall, T. and Roberts, I. (2005), *British Medical Journal* **331**, 267-270.

Buyse, M., George, S.L., Evans, S., Geller, N.L., Ranstam, J., Scherrer, B., Lesafre, E., Murray, G., Edler, L., Hutton, J., Colton, T., Lachenbruch, P. and Verma, B.L. (1999), The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. *Statistics in Medicine* **18**, 3435-3451.

Hill, T.P. (1998), The first digit phenomenon. *American Scientist* **86**, 358-363.

Hill, T.P. (1996), A statistical derivation of the significant-digit law. *Statistics in Science* **10**, 354-363. Mosimann, J.E. and Ratnaparkhi, M.V. (1996), Uniform occurrence of digits for folded and mixture distributions on finite intervals. *Communications in Statistics – Simulation and Computation* **25**, 481-506.

O'Kelly, M. (2004), Using statistical techniques to detect fraud: a test case. *Pharmaceutical Statistics* **3**, 237-246.

Walter, C.F. and Richards III, E.P. (2002), Legal defenses to the use of data digits to identify fabricated data. IEEE Engineering in Medicine and Biology **21**, 77-80.

Walter, C.F. and Richards III, E.P. (2001), Using data digits to identify fabricated data. IEEE Engineering in Medicine and Biology **20**, 96-100.

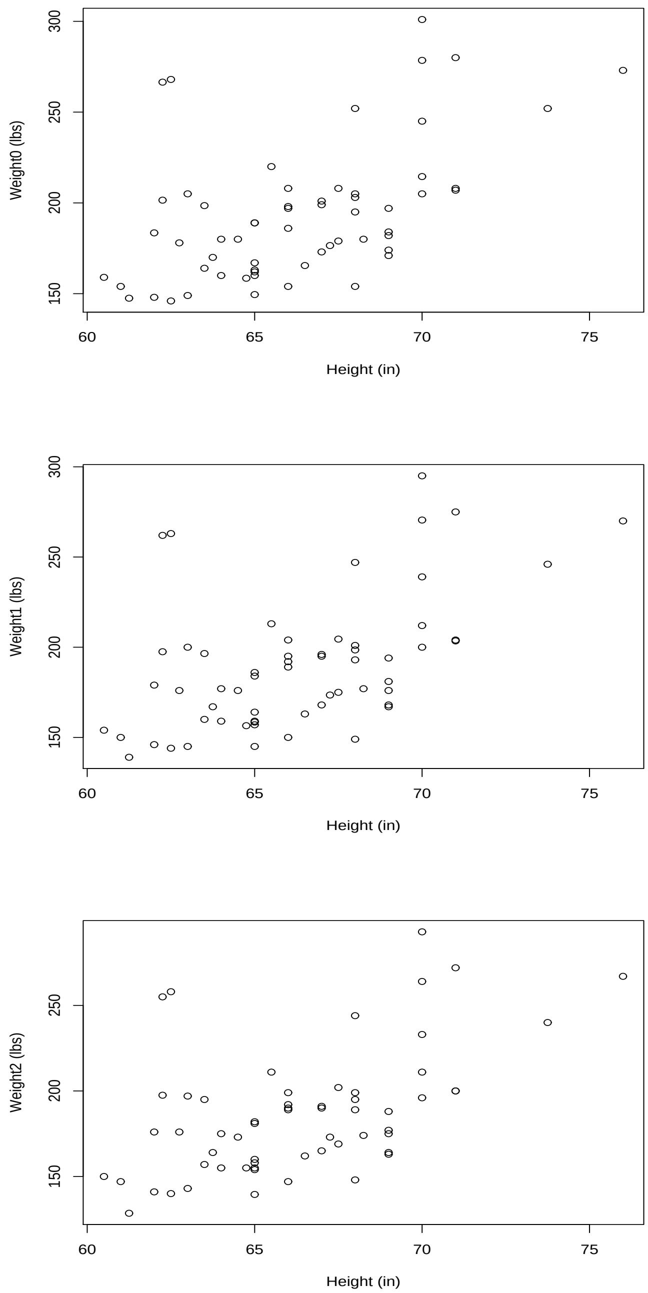

Figure 1: Scatterplots of weights against heights for the Fleming data.

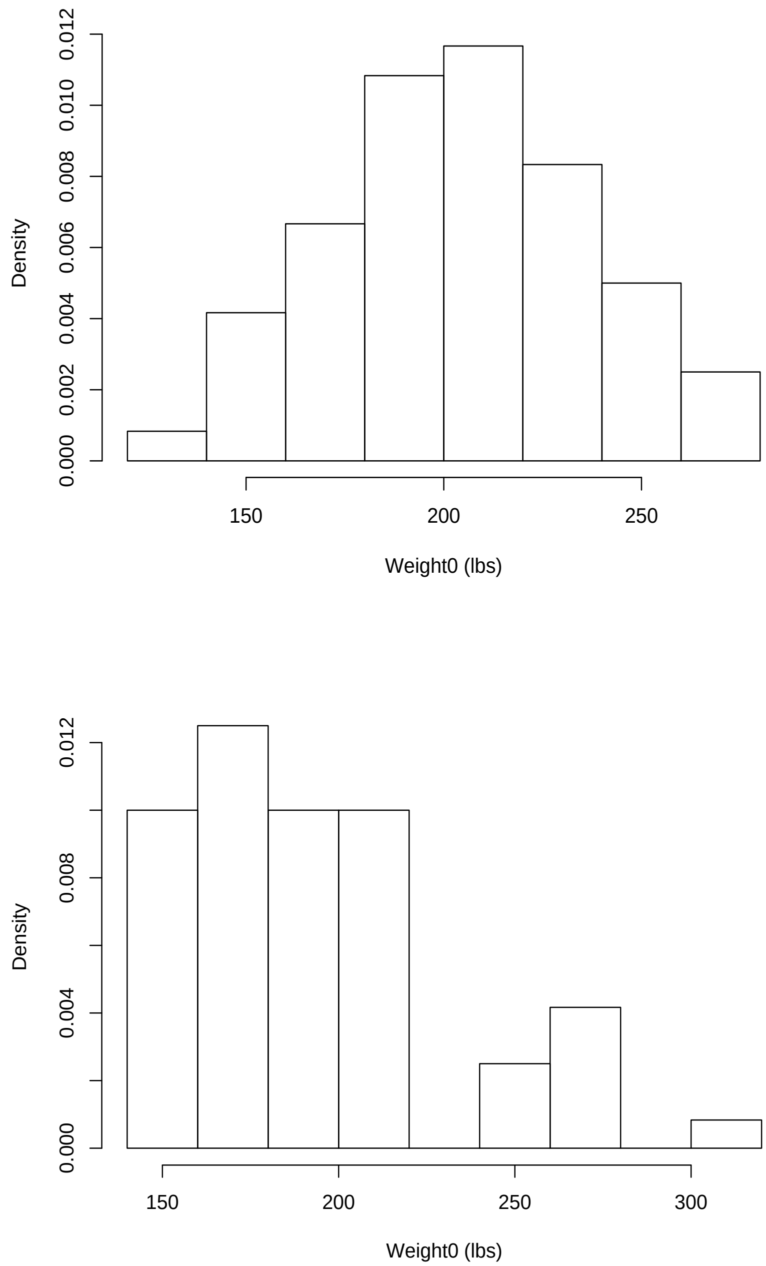

Figure 2: Histograms of weight at time 0 for the simulated data set (upper panel) and the Fleming data set (lower panel).

Appendix 1: R Functions Used in the Analysis of the Report.

1. Simulation of Values from a Multivariate Normal Distribution.

```
randdat<-function(muvect,Sigmat,n){
# requires package bayesurv
#
rawdat<-rMVNorm(n,muvect,Sigmat)
roundat<-round(rawdat,0)
orig<-1:60
ind1<-sample(orig,5)
ind2<-sample(orig[-ind1],9)
ind3<-sample(orig[-c(ind1,ind2)],4)
roundat[ind1,1]<-roundat[ind1,1]+0.25
roundat[ind2,1]<-roundat[ind2,1]+0.5
roundat[ind3,1]<-roundat[ind3,1]+0.75
ind11<-sample(orig,11)
roundat[ind11,2]<-roundat[ind11,2]+0.5
ind21<-sample(orig,9)
roundat[ind21,3]<-roundat[ind21,3]+0.5
ind31<-sample(orig,3)
roundat[ind31,4]<-roundat[ind31,4]+0.5
roundat<-cbind(1:n,roundat)
dat<-as.data.frame(roundat)
names(dat)<-c("subject","ht","wt0","wt1","wt2")
return(dat)
}
```

2. Compare “suspect” data records to averages of other pairs.

```
checkavging<-function(dat,suspectno){
 suspect<-dat[dat$subject==suspectno,]
 rdat<-dat[-suspectno,]
 rn<-dim(rdat)[1]
 npairs<-rn*(rn-1)/2
 res<-c(rep(0,7))
 cnt1<-0
 repeat{
  cnt1<-cnt1+1
  t1<-rdat[cnt1,]
  cnt2<-cnt1
  repeat{
   cnt2<-cnt2+1
   t2<-rdat[cnt2,]
   tsubs<-c(rdat$subject[cnt1],rdat$subject[cnt2])
#cat("tsubs: ",tsubs,fill=T)
   tavg<-0.5*(t1+t2)
   flag1<-(tavg$ht==suspect$ht)
   flag2<-(tavg$wt0==suspect$wt0)
   flag3<-(tavg$wt1==suspect$wt1)
   flag4<-(tavg$wt2==suspect$wt2)
   nflags<-flag1+flag2+flag3+flag4
   if(nflags>0){
   tres<-c(tsubs,nflags,flag1,flag2,flag3,flag4)
   res<-rbind(res,tres)}
   if(cnt2==rn) break
```

```
   }
  if(cnt1==rn-1) break
  }
return(res)
}
#----------------------------------------------------------------------------
summarycheckavg<-function(dat,suspectnos){
 sk<-length(suspectnos)
 res1<-NULL; res2<-NULL; res3<-NULL; res4<-NULL; res5<-NULL
 res6<-NULL; res7<-NULL; res8<-NULL
 cnt<-0
 repeat{
  cnt<-cnt+1
  tsus<-suspectnos[cnt]
  tres<-checkavging(dat,tsus)
  rs<-dim(tres)[1]
 if(is.null(rs)==FALSE){
  if(rs==1){
  res1<-c(res1,tsus)
  res2<-c(res2,tres[1])
  res3<-c(res3,tres[2])
  res4<-c(res4,tres[3])
  res5<-c(res5,tres[4])
  res6<-c(res6,tres[5])
  res7<-c(res7,tres[6])
  res8<-c(res8,tres[7])
  }
```

```
  if(rs>1){
   cnt2<-0
   repeat{
    cnt2<-cnt2+1
    ttres<-tres[cnt2,]
    res1<-c(res1,tsus)
    res2<-c(res2,ttres[1])
    res3<-c(res3,ttres[2])
    res4<-c(res4,ttres[3])
    res5<-c(res5,ttres[4])
    res6<-c(res6,ttres[5])
    res7<-c(res7,ttres[6])
    res8<-c(res8,ttres[7])
    if(cnt2==rs) break
    } } }
 if(cnt==sk) break
 }
res<-data.frame(suspect=res1,other1=res2,other2=res3,nflags=res4,
                flag1=res5,flag2=res6,flag3=res7,flag4=res8)
res2<-res[res$other1!=0,]
return(res2)
}
```

3. Examine distributions of trailing digits.

```
digitdist<-function(dat){
 ht<-dat$ht
 wt0<-dat$wt0
```

```
wt1<-dat$wt1
wt2<-dat$wt2
ht<-floor(ht)
wt0<-floor(wt0)
wt1<-floor(wt1)
wt2<-floor(wt2)
ldht<-ht-10*floor(ht/10)
ldwt0<-wt0-10*floor(wt0/10)
ldwt1<-wt1-10*floor(wt1/10)
ldwt2<-wt2-10*floor(wt2/10)
htfs<-NULL; wt0fs<-NULL; wt1fs<-NULL; wt2fs<-NULL
cnt<--1
repeat{
 cnt<-cnt+1
 thtf<-sum(ldht==cnt)
 twt0f<-sum(ldwt0==cnt)
 twt1f<-sum(ldwt1==cnt)
 twt2f<-sum(ldwt2==cnt)
 htfs<-c(htfs,thtf)
 wt0fs<-c(wt0fs,twt0f)
 wt1fs<-c(wt1fs,twt1f)
 wt2fs<-c(wt2fs,twt2f)
 if(cnt==9) break
 }
res1<-data.frame(digit=0:9,ht=htfs,wt0=wt0fs,wt1=wt1fs,wt2=wt2fs)
tstht<-sum((res1$ht-6)^2/6)
tstwt0<-sum((res1$wt0-6)^2/6)
```

```
 tstwt1<-sum((res1$wt1-6)^2/6)
 tstwt2<-sum((res1$wt2-6)^2/6)
 pht<-1-pchisq(tstht,9)
 pwt0<-1-pchisq(tstwt0,9)
 pwt1<-1-pchisq(tstwt1,9)
 pwt2<-1-pchisq(tstwt2,9)
 res2<-data.frame(var=c("ht","wt0","wt1","wt2"),
                  tst=c(tstht,tstwt0,tstwt1,tstwt2),
                  pval=c(pht,pwt0,pwt1,pwt2))
 res<-list(res1,res2)
 return(res)
}
```

4. Compute influence values.

```
influencefctn<-function(dat){
 wt2<-dat$wt2
 wt1<-dat$wt1
 wtdif<-wt1-wt2
 mn<-mean(wtdif)
 v2<-var(wtdif)
 n<-length(wtdif)
 realt<-mn/sqrt(v2/n)
 subs<-NULL; infls<-NULL
 cnt<-0
 repeat{
  cnt<-cnt+1
  tsub<-dat$subject[cnt]
```

```
  tvals<-wtdif[-cnt]
  tt<-mean(tvals)/sqrt(var(tvals)/(n-1))
  tinfl<-abs(tt-realt)
  subs<-c(subs,tsub)
  infls<-c(infls,tinfl)
  if(cnt==n) break
 }
res<-data.frame(subject=subs,influence=infls)
return(res)
}
```

Appendix 2: Data Sets Used in This Report.

1. The Fleming Data.

```
subject ht wt0 wt1 wt2
1 63.5 164 160 157
2 63.75 170 167 164
3 62.75 178 176 176
4 65 160 158.5 158
5 65 149.5 145 139.5
6 62.25 201.5 197.5 197.5
7 70 214.5 212 211
8 68.25 180 177 174
9 64 180 177 175
10 64.75 158.5 156.5 155
11 67.25 176.5 173.5 173
12 64 160 159 155
13 65.5 220 213 211
```

14 76 273 270 267

15 62 183.5 179 176

16 71 208 203.5 200

17 62.5 146 144 140

18 62.25 266.5 262 255

19 70 278.5 270.5 264

20 63.5 198.5 196.5 195

21 73.75 252 246 240

22 67.5 208 204.5 202

23 61.25 147.5 139 128.5

24 63 205 200 197

25 68 195 193 189

26 60.5 159 154 150

27 65 189 184 181

28 64.5 180 176 173

29 65 167 164 160

30 66 154 150 147

31 68 203 198.5 195

32 71 207 204 200

33 69 182 176 175

34 67.5 179 175 169

35 66.5 165.5 163 162

36 63 149 145 143

37 69 184 181 177

38 65 162 159 154

39 67 199 196 190

40 70 245 239 233

```
41 67 201 195 191
42 70 205 200 196
43 69 174 167 163
44 62.5 268 263 258
45 71 280 275 272
46 66 208 204 199
47 68 252 247 244
48 66 198 195 189
49 68 154 149 148
50 65 189 186 182
51 69 197 194 188
52 66 186 189 192
53 68 205 201 199
54 70 301 295 293
55 62 148 146 141
56 67 173 168 165
57 66 197 192 190
58 61 154 150 147
59 69 171 168 164
60 65 163 157 155
```

2. The Simulated Data.

```
subject ht wt0 wt1 wt2
1 67.5 207 202 200
2 62 161 161 161
3 70 269 263.5 254
4 65 188 184 181
```

5 69 249 244 237

6 67 166.5 162 157

7 75 211 208 204

8 66 208 205 202

9 65 205.5 200 196

10 66 206 200 197

11 65 181 178.5 174

12 66 200.5 196 192

13 66 171 168.5 167

14 71 235 232 231

15 66 179 173 170

16 61 161 157 155

17 63 179 175.5 174

18 72 147 145 143

19 70 231 225 220

20 63 136 132.5 125

21 63.25 217.5 213 212

22 69 236 231 226

23 67 171 166 162

24 71 193 188 186

25 67.5 174 169.5 166.5

26 72.5 265.5 258 254

27 65 214 211 207

28 65 185 180.5 180

29 63 192.5 189 184

30 67 231 227.5 224

31 65 192 188 185.5

32 67 218 217 216

33 63.5 184 177 168

34 65.75 222 215 209.5

35 67 207 201 196

36 66.5 257 256 254

37 72 223 218 212

38 71 221 214 210

39 66.25 213 209 206

40 66 239.5 236 233

41 67 143 140 137

42 64.25 221 216 211

43 66 209 203 198

44 68.25 181.5 179 177

45 69.5 243 234 229

46 70 252 247 242

47 64 158 156 155

48 68 222 220.5 215

49 70.5 257 249 242

50 69.75 219 216 212

51 69.25 156.5 154 150

52 68 191 187 184

53 64.5 182 180 174

54 73.75 252 247 242

55 70 194.5 190 186

56 61 210.5 206 204

57 68.5 265 257 253

58 62 187 182 177

59 71.75 198 192 188

60 64 145 142 140