
Stem and Leaf

Technique

What's for breakfast? I'm in one of my fitness moods, it's 5 A.M. and the first question of the day is "What's for breakfast?" I want a high protein breakfast. And, more generally, I want to know what it is that makes one breakfast cereal different from another. My data are from *Food Values of Portions Commonly Used*, by Bowes and Church, a dietitian's handbook whose introduction is chock full of references that tell me where the data come from should I want to check for myself.¹ Here is their table of data on cereals, ready to serve, describing nutrients found in a standard portion, Figure .1

Reproduce page from Bowes and Church, page 11, on facing page.

Reading the top pair of lines, the table indicates that Gerber's barley cereal is usually served in a one cup portion weighing thirty-six grams. It provides 128 calories, 4.3 grams of protein, 27.3 grams of carbohydrates, .4 grams of fiber, .2 grams of fat, and negligible grams of polyunsaturated fat (FAP). Reading across, the data indicate the amounts of eight amino acids, in milligrams, of six minerals, in milligrams, and of six vitamins, in various units.

Focus on the grams of protein: I want to know how much protein these breakfast cereals tend to provide, what's high, what's low, and more generally what it is about "breakfast" cereal that leads some to

¹ The frequently updated edition is by Pennington and Church, published by Harper & Row. I'm using an out-of-date, 1975, edition, page 11, because it was less complete than more-current editions and therefore easier to use as an exercise. The up-to-date edition includes more information on more products.



From FOOD VALUES OF PORTIONS COMMONLY USED, BOWES & CHURCH, 1975, p. 11

Food	WT	CAL	CHO	FAT	TRP	LEU	LYS	MET	Na	Ca	P	THI	NIA	VIA
Measure	gm	PRO gm	FIB gm	FAP gm	PHA mg	ISL mg	VAL mg	THR mg	K mg	Mg mg	Fe mg	RIB mcg	ASC mg	VID iu
4.1 CEREALS (A) - READ TO SERVE														
BARLEY CEREAL, GERBER'S	36	128	27.3	.2	54	299	145	62	215	231	260	1015	5.1	(0)
1 cup		4.3	.4		222	183	216	145	149		18	763	0	
BRAN, ALL-, KELLOGG'S	28	95	21.4	.7					370	24	350	110	5.0	(0)
1/2 cup		3.1	2.3								2.9	90	0	400
BRAN FLAKES, 40%, KELLOGG'S	28	101	22.6	.6					340	16	170	100	2.4	(0)
3/4 cup		2.9	1.0								1.3	50	0	
BRAN FLAKES, 40%, POST'S	28	100	22.0	.5					(340)	-	110	130	1.5	(0)
3/4 cup		2.8	1.0								1.0	-	-	0
BRAN, RAISIN, KELLOGG'S	21	73	16.6	.4					280	13	105	63	1.5	
1/2 cup		1.8	-								1.0	-	0	0
BRAN, RAISIN, POST'S	28	99	22.0	.4							94	100	1.1	
2/3 cup		2.2	-								1.0	-	0	0
CHEERIOS, GENERAL MILLS'	25	102	17.7	1.8					275	42	100	302	.5	
1 cup		3.4	.3								1.1	49	0	0
CORN FETTI, POST'S	28	110	25.0	.1						-	-	110	.5	
3/4 cup		1.5									.4	-	0	0
CORN FLAKES (b)	25	95	21.0	.1	14	272	40	35	165	6	16	100	.5	(0)
1 cup		2.1	.2		92	80	100	72	40		.5	20		0
CORN SOYA SHREDS	28	103	21.0	.1					310	24	52	190	.6	(0)
3/4 cup		5.1	.3								1.2	40	0	0
GRAPE NUTS (c)	28	110	24.0	.2	-	196	45	39	17	-	-	130	1.5	(0)
1/4 cup		2.8			137	137	137	90			1.0	-	0	0
GRAPE NUT FLAKES	28	110	23.0	.4						-	-	130	1.6	(0)
3/4 cup		2.7									1.2	-	-	-
HIGH PROTEIN CEREAL, GERBER'S	29	102	14.5	.3					226	213	241	818	4.1	
3/4 cup		10.2	.4						313		14.	615		
HI PRO, GENERAL MILLS'	21	80	14.1	.3					294	65	84	345	3.1	
1 cup		4.8	.1								3.9	432		
KIX, GENERAL MILLS'	25	99	20.2	1.0					275	5	22	214	.7	(0)
1 cup		2.0	.1								1.5	46		0
KRUMBLES, KELLOGG'S	28	103	23.8	.3					170	11	110	10	2.0	(0)
3/4 cup		2.6									1.0	30	0	0
MIXED CEREAL, GERBER'S	21	76	15.4	.3					126	111	133	592	3.0	
1/2 cup		3.0	.2						72		10.5	445		
MUFFETS, QUAKER	23	80	18.2	.3					1.0	10	87	50	1.0	(0)
1 biscuit		2.2	.6								.9	20	0	0
OATMEAL (d)	27	98	18.6	.6	58	337	165	66	135	153	169	761	3.8	(0)
3/4 cup		4.5	.4		240	232	267	149	101	13.5	572			0
POST TOASTIES	28	100	24.0	.1						-	-	110	.5	(0)
1 1/4 cup		2.1	.2								.4	-	-	0
RICE CEREAL, GERBER'S	27	97	21.9	.4					205	179	171	761	3.8	
3/4 cup		1.5	.1						56		13.5	572		
RICE FLAKES	32	123	27.4	.1	16	-	20	-	311	11	53	110	1.4	(0)
1 cup		2.1	.3		102	-	-	-			.6	20		0
RICE KRISPIES, KELLOGG'S	28	107	25.1	.1					280	7	33	110	2.0	(0)
1 cup		1.6									.5	10		0
RICE, PUFFED, QUAKER	13	51	11.5	.1					.3	2	13	60	.6	(0)
1 cup		.8	.1								.2	10		0
SPECIAL K CEREAL, KELLOGG'S	16	60	12.5	.1					193	17	41	228	2.9	
1 cup		3.2									2.5	285	6	228
WHEAT FLAKES, QUAKER	36	125	28.0	.4	49	363	147	52	403	17	108	160	1.9	(0)
1 cup		4.4	.5		195	202	233	145			1.3	60	0	0
WHEATIES, GENERAL MILLS'	28	104	22.5	.6					392	11	78	167	1.6	(0)
1 cup		2.8	.5								1.7	47	0	0
WHEAT, PUFFED, QUAKER	12	43	9.5	.2					1	3	40	70	.9	(0)
1 cup		1.6	.2								.5	30		0
WHEAT, SHREDDED	22	84	18.3	.3	18	149	72	30	.5	11	93	65	1.0	(0)
1 biscuit		2.2	.5		105	98	126	88			.8	23		0
WHEAT CHEX, RALSTON	28	102	23.4	.3					225	11	105	40	1.5	
1/2 cup, 47 biscuits		2.8	.6								.9	60		

(a) All the cereals listed on this page may be served from the package without further preparation. When served with milk or cream and/or sugar the addenda should be consulted.

(b) The amino acid values are from reference 15. Sodium and potassium figures are calculated from reference 14 and 14a.

(c) These values for amino acids are derived from reference 15.

(d) The ready-to-serve product is indicated here. Amino acid data is from reference 15.

NOTE: A serving of cereal varies with individual taste, age, and activity level. A common size serving is 1 ounce.

be high in protein while others are low? What is the “mechanism” behind the facts?

O.K., that’s the agenda, but first some technique — the real agenda for this chapter. And the first technique, not my breakfast, is called the “Stem and Leaf.”

This is a technique that is both extremely useful and extremely modest — hard to take seriously until it “works” for you, time after time — too simple to pay off, but it does. It is also one of those techniques that falls on the unseen side of the data analysis — rarely seen in a final report, but often found on the scratch pad of the data analyst, usually in pencil, usually with notes and scratchings all over it: It is informal and extremely useful.

Let me begin simply and mechanically, without context, by simply extracting the protein numbers from Figure 1 and illustrating the technique.

4.3, 3.1, 2.9, 2.8, 1.8, 2.2, 3.4, 1.5, 2.1, 5.1, 2.8, 2.7, 10.2, 4.8, 2.0, 2.6, 3.0,
2.2, 4.5, 2.1, 1.5, 2.1, 1.6, .8, 3.2, 4.4, 2.8, 1.6, 2.2, 2.8

Figure 2

Numbers, for Practice, Extracted from Protein Values of Figure 1.

Mechanically, beginning with these numbers, a Stem and leaf is a new copy of the same numbers — but re-grouped by size and presented in a way that shows the “shape” of the data. It’s a first step in engaging your intuition and experience as allies in the process of data analysis.

Working just with the numbers, Figure 2 is a stem and leaf for the numbers in Figure 1.

0	8	↖ 0.8
1	85566	
2	98218706211828	
3	1402	
4	3854	
5	1	

Stem for numbers greater than or equal to 0 but less than 1
Stem for numbers greater than or equal to 1 but less than 2

10	2
----	---

Figure 2
Stem and Leaf, Integer Stems

The “stems” identify ranges, dividing the numbers into major divisions. The “leaves” identify each of the numbers within the division: Leaves are the numbers attached to the stems. Here, in this first example, stems divide the data using digits to the left of the decimal, with stems 0 through 10. The leaves mark each datum on the stem with an additional digit. So, the number 0.8 is represented in stem 0, identified by the leaf “8”. And the number 10.2 is represented in stem “10”, identified by the leaf “2”. All of the numbers greater than or equal to 0, but less than 1, are represented in the first stem. All of the numbers greater than or equal to 1, but less than 2, are represented in the second stem. The leading digits, “0”, “1”, “2”, and up to “10” label the stems. The final digits identify the leaves.

Altogether, the effect of the stem and leaf, when it is completed, is to put the data in rank order, roughly, and to show the shape of the distribution of the numbers. Much of the value of the stem and leaf lies in the process as much as the result: In the process you almost literally *feel* which numbers are typical as you record the leaves one by one, putting them in place. By the time you are done you “know” your data — which numbers are small, which numbers are large compared to the rest, which ones stand out from the others flagging that, for these few

numbers something is different — they don't belong with the rest. In the process, you begin to get a feel for the numbers.

When you are done you have a picture. And what you do with the picture is stand back and look at it for a moment: Here, for these thirty numbers, the shape is symmetrical with many numbers in the middle, and few at either end. Separately, one of the thirty numbers is out on its own. That symmetrical shape, many numbers in the middle and a few at either end, is what's known as a bell-shaped distribution — some data show it, some data don't. And the one suspicious case at the end is what's known as an outlier.

That's a stem and leaf for these numbers. But usually it's not quite that simple. Usually you have to try a couple of different stem and leaf drawings before you get one that looks right. What does it mean to "get one that looks right?" To show you what I mean, let me practice with these numbers.

For practice, I can use different stems that expand the stem and leaf as in Figure 3. Here I've expanded the whole thing twice as far physically: I've used 0.0 through 0.4 for the first stem, 0.5 through 0.9 for the second stem, and so forth, giving me twice as many stems.

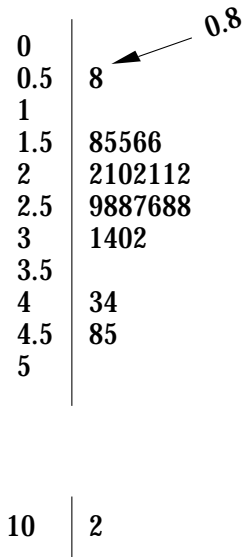


Figure 3
Stem and Leaf, Expanded

Or, I can compress the stem and leaf by dividing at 0, 2, 4, 6, 8 and 10, as in Figure 4.

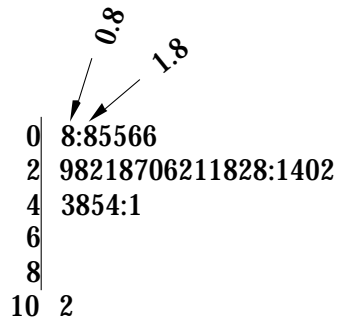


Figure 4
Stem and Leaf, Compressed

If you expand the data too much, then the shape gets irregular—showing gaps as in Figure 3. Alternatively I can try a third set of stems that compress the data. And if you compress the data too much then the shape disappears into a lump. How many stems should you use? Try for 5 to 10 but there is no fixed answer. Look for a shape that is fairly compact, like Figure 2. If you expand it and the shape begins to break up, as in Figure 3, then you've gone to far. Here, Figure 2 is good enough.

One not-so-minor detail of the stem and leaf that should be made explicit is the labeling. Labels are as much a part of the technique as the numbers. And the rule for labeling is: Make it Clear. Labels need to identify the stems, they need to identify the leaves, and— when we get to real data—they have to describe the unit of analysis (e.g., kilograms of protein or grams of protein or milligrams of protein). Briefly, the labels are an echo of the “Who, What, Where, Why, When, and How” They specify what the data are about.

For the stems there are at least three styles I could have used to label the expanded stems in Figure 3. In Figure 4 I used digits. Alternatively, as in Figure 5, I could have simply repeated the leading digit, clear enough in context, or repeated the leading digit and marked the second case with a “*”, to distinguish it from the other.

0	0	0
0	0	0*
1	1	1
1	1	1*
2	2	2
2	2	2*
3	3	3
3	3	3*
4	4	4
4	4	4*
5	5	5
5	5	5*

Figure 5

Stem and Leaf, Alternative Labels for Stems

Which style should you use? Actually I myself used all three of these styles as I worked with these data. First I used the labels on the left of Figure 5 — they're the easiest — using each label twice. But then, as I worked with this style, I didn't like it: For some reason, I found these stems easy to write down but hard to use: Using these stems I kept making errors by attaching the leaves to the wrong stems. So I changed to the "*" version in order to differentiate the stems. But then, as I used it, I kept making errors — still putting the leaves on the wrong stems. So I changed again, settling on the labels 0, 0.5, ... of Figure 2. With the limits of the stem visible on the page, built-in to the labels for the stems, I was able to work faster and with fewer errors.

And I give you this blow by blow summary of my thinking — first I did this and then I did that — just to make it clear how you decide, and how I decide, to do the stem and leaf: Don't look for the "one true way" of doing it. There is none. Instead, think of what you are trying to accomplish and feel free to change technique until it works.

Also note that I put a label on each of the completed stem and leaf diagrams. a "0.8" with an arrow attached. That was for my benefit, to help me read the stem and leaf when I have to go back to make sense of my own work — an hour, or a day or a month later. And you should use

it yourself so that your own stem and leaf drawings can be read: Labels are *not* an after-thought. They are part of good technique.

Exercises

1. Practice several forms of the stem and leaf on the numbers for fat content in grams

.2, .7, .6, .5, .4, .4, 1.8, .1, .1, .1, .2, .4, .3, .3, 1, .3, .3, .3, .6, .1, .4, .1, .1, .1, .1, .4, .6, .2, .3, .3

and, again, on the numbers of calories for the same breakfast cereals

128, 95, 101, 100, 73, 99, 102, 110, 95, 103, 110, 110, 102, 80, 99, 103, 76, 80, 98, 100, 97, 123, 107, 51, 60, 125, 104, 43, 84, 102

2. Use a pair of dice and construct the stem and leaf drawing for one hundred passes with the dice. Before you begin, ask questions: What do expect the diagram to look like? Why? Now, throw the dice. What do you get and why?

Application: Protein Content of Breakfast Cereals

Now, what's for breakfast? For *data* analysis — with emphasis on the word *data* — I have to place these numbers back in context: The numbers record the grams of protein in “commonly used portions” of breakfast cereals. Now, in context, there is a difference: Here is where

I get to assert the advantages of my human brain as compared to the mechanical “brain” of a computer. In context this human brain has expectations, and a little experience with breakfast, and I intend to use those advantages in the course of my analysis. I, the analyst with expectations, expect the quality of the food to depend on the ingredients and also upon the manufacturer and the process. In context I expect these cereals to divide into two batches, good stuff and bad stuff, or, health foods and junk foods.

Having thought about what I expect, I’m ready to begin my stem and leaf. What kind of stems? Well, as before, when I looked at the numbers, I see numbers like 4.3, 3.1, and 2.9, reading down. So, again, I’ll start with stems indicating grams of protein in one gram intervals.

0 |
1 |
2 |
3 |
4 |
5 |
6 |

Protein in Grams

O.K., now what about the leaves? You’ve seen these numbers before, as numbers. Now, in context, these are data: They have been identified with something real and that makes a difference. Bearing in mind that I’m using the numbers to get at something else, something about breakfast cereals, I’m going to use leaves that advance my purpose: I’m *expecting* the ingredients to tell part of the story so, somehow, the ingredients should be marked, in the leaves. And again, I’m *expecting* that some manufacturers make a better product. So, I want to keep track of manufacturers. And then, for bookkeeping purposes, I note that the data come in alphabetical order. That’s useful: I want to use that alphabetical order in order to be able to connect my summary of the data, in the stem and leaf, back to the full data that it comes from. All together, there’s a lot of information here, waiting to be organized. So, I’m going to use labels for the leaves, not numbers. And I’m going to

build into those labels whatever useful information I can manage (in addition to the grams of protein). So I begin

0			
1			
2		Bran Fl Kell;	<i>Stem and Leaf</i>
3		Bran Kell;	<i>First Three Leaves</i>
4		Barley Gerb;	
5			
6			

Protein in Grams

Those are the first three leaves and, pausing for the moment, let me note a few things: For one, the physical lengths of these three stems are not quite equal, so the visual shape of the stem and leaf will be a little distorted. True, but that’s not too important. For another, note that I changed my stem labels a little, on the fly. The second cereal was “Bran, All-, Kellogg’s,” in the data, and I wrote down “Bran Kell,” in the leaf. But then the third cereal entry was also “Bran Kell”, like the second, except that this third food is a flake and the second was not. Ah, that’s new information, at no extra cost — something about the process. So I’ll put that information into the third leaf, even though I didn’t use it in the second leaf: Consistency is nice, but not when it gets in the way. (And if this becomes important, the textual stems, unlike numerical stems, will make it easy to go back to the data for more detail.) Continuing

0			
1		Bran Rais Kell;	
2		Bran Fl Kell; Bran Fl Post;Bran Rais Post;	<i>Stem and Leaf</i>
3		Bran Kell; Cheerios GM	<i>First Six Leaves</i>
4		Barley Gerb;	
5			
6			

Protein in Grams

Those are the first seven leaves and I've got trouble again: the data themselves are not always labeled the same way and my seventh datum, for "Cheerios," doesn't specify ingredients. What do I do? I use it anyway. I use what I've got. And if the missing information becomes important, I can find out later. Continuing:

0			
1		Bran Rais Kell; Corn F Post;	
2		Bran Fl Kell; Bran Fl Post; Bran Rais Post; Corn Fl; Grape Nuts; Grape Nut Fl	
3		Bran Kell; Cheerios GM	
4		Barley Gerb;	<i>Stem and Leaf</i>
5		Corn Soya;	<i>First Eleven Leaves</i>
6			
7			
8			
9			
10		High Pro Gerb;	

Protein in Grams for Commonly Used Portions of Breakfast Cereal

Example: High Protein Gerbers 10.2 grams

Thirteen items into the procedure, and now there's a big one, out of line with the rest: High Pro Gerbers has several times more protein than the competition. I could stop now and think about it, but there's not much more data, so I'll continue.

0	Rice Puff Q;
1	Bran Rais Kell; Corn F Post; Rice Gerb; Rice Kr Kell; Wheat Puff Q
2	Bran Fl Kell; Bran Fl Post; Bran Rais Post; Corn Fl; Grape Nuts; Grape Nut Fl; Kix GM; Krumb Kell; Muff Quak; Post T; Rice Fl; Wheats GM; Whet Shred; Wheat Chex Ralst
3	Bran Kell; Cheerios GM; Mixed Gerb; Spec K Kell
4	Barley Gerb; High Pro GM; Oatml; Wheat Fl Quak
5	Corn Soya;
6	
7	<i>Stem and Leaf</i>
8	<i>Breakfast Cereals</i>
9	
10	High Pro Gerb;

Protein in Grams for Commonly Used Portions of Breakfast Cereal
Example: High Protein Gerbers 10.2 grams

O.K., that's it for the moment: I ran out of room and ruined the shape on stem "2". I've still got one real stand out, "High Pro Gerb". And I've been a little inconsistent, paying more attention to the process: flaked, puffed, or shredded, than I had intended. This is probably good enough. But I probably could have figured out pretty early that I should have been using different stems: I used eleven stems because eleven gave me a convenient division of the range between zero and ten. But with these stems things are bunching up in part of the range, while almost half of the stems, between five and ten are nearly empty. I might have been better off expanding the stem and leaf within the range from zero to five, just leaving the one very high protein cereal, Gerbers High Pro, out there on its own.

0		
0.5	Rice Puff Q;	
1		
1.5	Bran Rais Kell; Corn F Post; Rice Gerb; Rice Kr Kell; Wheat Puff Q	
2	Bran Rais Post; Corn Fl; Kix GM; Muff Quak; Post T; Rice Fl; Whet Shred;	
2.5	Bran Fl Kell; Bran Fl Post; Grape Nuts; Grape Nut Fl; Krumb Kell; Wheats GM; Wheat Chex ; Ralst	
3	Bran Kell; Cheerios GM; Mixed Gerb; Spec K Kell	
3.5		
4	Barley Gerb; Wheat Fl Quak	
4.5	High Pro GM; Oatml;	
5	Corn Soya;	<i>Stem and Leaf</i>
4.5		<i>Expanded</i>

10 | High Pro Gerb;

Protein in Grams for Commonly Used Portions of Breakfast Cereal

Example: High Protein Gerbers 10.2 grams

Trying out the expanded stem and leaf, there's not much difference. Either way, by grams or expanded to half-gram intervals, it's clear that the most frequent values are in the neighborhood of 2.5 grams of protein. And, still, the prominent event is the outlier, High Pro Gerber's, with protein that is four times the typical value.

Now, stepping back to look at these things, looking at either one of the stem and leaf diagrams, what can I learn about cereal? For one thing, I learn that one of my expectation was wrong: I expected something simple, good food versus bad food. Wrong: It's not that simple. With the exception of Gerber's, most of the foods form one nice batch, one nice distribution not two batches, good and bad, but one. And while "bran cereals" get lots of good public relations as "health foods", they are not notably high in protein. "Good" versus "bad" is too simple.

However, while the twenty-nine cereals (other than Gerbers) are not divided into two types, good versus bad, the range of values shown

in the stem and leaf diagram shows that there is a very large variation within this group. How large? The protein values range from one cereal in the 0-gram stem to one cereal in the 5-gram stem. Precisely, how large? Using the alphabetical cues in the stem and leaf, I can quickly fill-in this detail by going back to the data for more information: The stem and leaf diagram displays extremes “Rice Puff Q” and “Corn Soya”, which allows me to glance back to the data for the full numbers, which are 0.8 and 5.1. So, how large is the variation? With the exception of one outlier, the protein values range from 0.8 to 5.1 grams of protein per serving, with the high protein cereals (at the end of the range) providing six times the protein content of the lowest value — a big contrast.

Now, I wanted to know “Why?” Why do some cereals have high protein. What about High Pro Gerbers? If I’m looking for high protein this is the first place to begin. It has unusually high protein, extremely high. Why? Perhaps it is the manufacturer. Is there something about Gerber that I should favor, as a brand? That’s one hypothesis — it’s the manufacturer. And I can check that hypothesis by going back through the leaves, marking the manufacturer’s names: (Here I’ve marked off Gerber with bolding. Ordinarily, I’d circle the Gerber’s in my existing stem and leaf, or mark them with a bright color.) Looking at the marked-up stem and leaf, I see Gerber all over, low, medium, and high. So, it’s not that simple. Cross-off that hypothesis.

0		
0.5	Rice Puff Q;	
1		
1.5	Bran Rais Kell; Corn F Post; Rice Gerb ; Rice Kr Kell; Wheat Puff Q	
2	Bran Rais Post; Corn Fl; Kix GM; Muff Quak; Post T; Rice Fl; Whet Shred;	
2.5	Bran Fl Kell; Bran Fl Post; Grape Nuts; Grape Nut Fl; Krumb Kell; Wheats GM; Wheat Chex Ralst	
3	Bran Kell; Cheerios GM; Mixed Gerb ; Spec K Kell	
3.5		
4	Barley Gerb; Wheat Fl Quak	
4.5	High Pro GM; Oatml;	
5	Corn Soya;	
4.5		<i>Stem and Leaf</i>
		<i>Highlighting Gerber</i>
10	High Pro Gerb;	

Protein in Grams for Commonly Used Portions of Breakfast Cereal

Example: High Protein Gerbers 10.2 grams

Let me try again. Maybe it is Gerber, but complicated by the choice of ingredients. Maybe Gerber's Rice is higher protein than other people's rice. Highlighting again, to check my hunch: No, Gerber's rice is right in there among the other rice cereals. But note, before I go on, how easy it was to do these checks of my hunches, or "hypotheses", and how hard it would have been if I had just recorded the digits (of Figure __). And note the rudiments of scientific reason built in to these last few steps: I used my expectations to form testable hypotheses. I formulated the display in order to test the hypotheses. I tested them. And, so far at least, both appear to be false.

0	
0.5	Rice Puff Q;
1	
1.5	Bran Rais Kell; Corn F Post; Rice Gerb; Rice Kr Kell; Wheat Puff Q
2	Bran Rais Post; Corn Fl; Kix GM; Muff Quak; Post T; Rice Fl; Whet Shred;
2.5	Bran Fl Kell; Bran Fl Post; Grape Nuts; Grape Nut Fl; Krumb Kell; Wheats GM; Wheat Chex Ralst
3	Bran Kell; Cheerios GM; Mixed Gerb; Spec K Kell
3.5	
4	Barley Gerb; Wheat Fl Quak
4.5	High Pro GM; Oatml;
5	Corn Soya;
4.5	<i>Stem and Leaf</i> <i>Highlighting Rice</i>
10	High Pro Gerb;

Protein in Grams for Commonly Used Portions of Breakfast Cereal
 Example: High Protein Gerbers 10.2 grams

Now I've used up everything I know about Gerbers High Pro, the extreme case, trying to figure out the "mechanism" that makes it special. No luck. So, having used up what I know about Gerbers, I'll have to look elsewhere. I'll go to the next largest value and see what this one might tell me: It says "Corn Soya", labeled by ingredients. Does this tell me anything? Comparing this second highest protein cereal to the lowest protein cereal my leaves show me "Corn Soya" versus "Rice Puff". That suggests a possibility, perhaps the important feature is the combination of ingredients, corn and soy, at one extreme versus rice at the other? To check that out I'll go back to the stem and leaf, making it up again. (Ordinarily I'd go back through the same stem and leaf with more colored pens or mark it with some fancy symbol, marking each ingredient where it is known. Here, I'll use different type fonts.)

0	
0.5	Rice Puff Q;
1	
1.5	Bran Rais Kell; <u>Corn</u> F Post; Rice Gerb; Rice Kr Kell; <u>Wheat</u> Puff Q
2	Bran Rais Post; <u>Corn</u> Fl; Kix GM; Muff Quak; Post T; Rice Fl; <u>Wheat</u> Shred;
2.5	Bran Fl Kell; Bran Fl Post; Grape Nuts; Grape Nut Fl; Krumb Kell; <u>Wheats</u> GM; <u>Wheat</u> Chex Ralst
3	Bran Kell; Cheerios GM; Mixed Gerb; Spec K Kell
3.5	
4	Barley Gerb; <u>Wheat</u> Fl Quak
4.5	High Pro GM; Oatml;
5	<u>Corn</u> Soya;
4.5	<i>Stem and Leaf</i>
	<i>Highlighting Ingredient</i>
10	High Pro Gerb;

Protein in Grams for Commonly Used Portions of Breakfast Cereal

Example: High Protein Gerbers 10.2 grams

That wasn't what I expected. Generalizing from two cases, I expected, or hoped to find that corn (underlined) is high in protein and notably separate from rice (bold). But the corn and rice are roughly in the same range — except in the one case where the corn was mixed with soy. *Other* contrasts, other than corn versus rice do look promising: There is a suggestion that the wheat cereals are higher protein than rice cereals. So I'll follow up on the "ingredients hypothesis", with a new set of stem and leaf drawings. For this hypothesis I can use a pair of stem and leaf drawings, back to back, separated by ingredient. For example, wheat versus rice:

	0	
	0.5	Rice Puff Q;
	1	
Wheat Puff Q	1.5	Rice Gerb; Rice Kr Kell
Whet Shred	2	Rice Fl;
Wheats GM; Wheat Chex Ralst	2.5	
	3	
	3.5	
Wheat Fl Quak	4	
	4.5	Stem and Leaf
	5	Wheat Versus Rice
	4.5	

That looks good and I'm feeling more certain of my hypothesis. I'll venture a guess now: Watch the ingredient. Guess that wheat cereals are higher protein than rice cereals. And, noting that corn cereal is pretty much like rice cereal *except for the one case in which the corn is mixed with soy* — I'll guess that soy beans are the key to the high protein content of corn soya. And that in turn leads me to a guess about Gerbers High Protein.

O.K. now, stepping aside from the data analysis, using the stem and leaf drawings, what have I got: I have a hunch that the main ingredient is the best predictor of protein content (not the manufacturer), plus a reasonable hypothesis explaining the anomalous Corn Soya (it has too much protein for a corn cereal), plus a guess that could explain Gerber's with its remarkably high protein. Have I proven anything? Have I proved, statistically, that the wheat distribution is different from the rice distribution — beyond a statistical doubt? No, nor do I need to. I've combed these data for ideas, detective style. I've eliminated some bright ideas that turned out to be poor and I've moved toward one idea that looks good, so far. That's my data analysis. And, since I'm squeezing these data to learn something about the breakfast cereals — remembering my goal — there's an easy way to check my

hunch: Find Gerbers in the store and read the box. Sure enough, moving my data analysis laboratory to the nearest grocery store, the *only* ingredient in High Protein Gerbers is finely sliced soybeans. Q.E.D.

Suggested reading: "They Should have Used A Shovel", David Freedman, *Sociological Methodology??*, _____

Exercises

1. Using the examination of protein content as a model, examine the distribution of fat content and calories for these cereals.

2. The "Dow Jones Industrial Average" is a weighted average of the prices for shares of stock for thirty companies. The data below show the change in price for each of the thirty companies. Construct a stem and leaf diagram for the change in price during the week March 18 to March 25 (column 5). Discuss. (Check your hand outs for more recent prices)

	<i>Dates:</i>	<i>March 18, 1994</i>	<i>March 25, 1994</i>	<i>Change</i>	<i>Change as Per Cent of March 18th Value</i>
	<i>Dow Jones Average of Thirty Industrials</i>	<i>3,895.65</i>	<i>3,774.73</i>	<i>- 120.92</i>	<i>-3.10%</i>
		<i>Price per Share of Stock</i>	<i>Price per Share of Stock</i>		
<i>1</i>	<i>Alcoa Aluminum</i>	<i>\$77.13</i>	<i>\$76.25</i>	<i>- \$0.88</i>	<i>- 1.13%</i>
<i>2</i>	<i>Allied Signal</i>	<i>\$77.00</i>	<i>\$76.50</i>	<i>- \$0.50</i>	<i>- 0.65%</i>
<i>3</i>	<i>American Express</i>	<i>\$30.13</i>	<i>\$29.63</i>	<i>- \$0.50</i>	<i>- 1.66%</i>

4	<i>AT&T</i>	\$53.50	\$52.63	- \$0.88	- 1.64%
5	<i>Bethlehem Steel</i>	\$21.50	\$20.88	- \$0.63	- 2.91%
6	<i>Boeing</i>	\$47.00	\$45.75	- \$1.25	- 2.66%
7	<i>Caterpillar</i>	\$119.13	\$116.63	- \$2.50	- 2.10%
8	<i>Chevron</i>	\$92.75	\$89.50	- \$3.25	- 3.50%
9	<i>Coca Cola</i>	\$41.88	\$41.75	- \$0.13	- 0.31%
10	<i>Disney</i>	\$47.00	\$45.00	- \$2.00	- 4.26%
11	<i>duPont</i>	\$58.75	\$56.25	- \$2.50	- 4.26%
12	<i>Kodak</i>	\$45.13	\$44.75	- \$0.38	- 0.83%
13	<i>Exxon</i>	\$65.88	\$65.50	- \$0.38	- 0.57%
14	<i>General Electric</i>	\$104.50	\$102.13	- \$2.38	- 2.27%
15	<i>General Motors</i>	\$59.88	\$56.88	- \$3.00	- 5.01%
16	<i>Goodyear</i>	\$45.25	\$41.75	- \$3.50	- 7.73%
17	<i>IBM</i>	\$57.13	\$54.00	- \$3.13	- 5.48%
18	<i>International Paper</i>	\$70.25	\$66.38	- \$3.88	- 5.52%
19	<i>MacDonalds</i>	\$60.88	\$58.38	- \$2.50	- 4.11%
20	<i>Merck</i>	\$31.75	\$30.13	- \$1.63	- 5.12%
21	<i>MMM</i>	\$103.00	\$100.00	- \$3.00	- 2.91%
22	<i>Morgan</i>	\$64.75	\$63.88	- \$0.88	- 1.35%
23	<i>Philip Morris</i>	\$55.00	\$51.50	- \$3.50	- 6.36%
24	<i>Proctor & Gamble</i>	\$56.63	\$53.75	- \$2.88	- 5.08%
25	<i>Sears</i>	\$48.13	\$46.00	- \$2.13	- 4.42%
26	<i>Texaco</i>	\$66.25	\$65.88	- \$0.38	- 0.57%
27	<i>Union Carbide</i>	\$25.88	\$25.00	- \$0.88	- 3.38%
28	<i>United Technologies</i>	\$68.25	\$66.50	- \$1.75	- 2.56%
29	<i>Woolworth</i>	\$19.88	\$19.13	- \$0.75	- 3.77%
30	<i>Westinghouse</i>	\$13.25	\$13.00	- \$0.25	- 1.89%

	<i>Dates:</i>	<i>March 3, 1995</i>	<i>March 10, 1995</i>	Change	Change as a Per Cent of the March3rd Value
	<i>Dow Jones Average of Thirty Industrials</i>	3989.61	4035.61	46	1.15%

		Price per Share of Stock	Price per Share of Stock		
1	<i>Alcoa Aluminum</i>	\$78.50	\$76.50	-2.00	-2.55%
2	<i>Allied Signal</i>	\$37.75	\$38.00	0.25	0.66%
3	<i>American Express</i>	\$33.38	\$32.63	-0.75	-2.25%
4	<i>AT&T</i>	\$51.25	\$52.13	0.88	1.72%
5	<i>Bethlehem Steel</i>	\$15.38	\$15.13	-0.25	-1.63%
6	<i>Boeing</i>	\$46.25	\$46.88	0.63	1.36%
7	<i>Caterpillar</i>	\$49.00	\$49.88	0.88	1.80%
8	<i>Chevron</i>	\$46.88	\$48.00	1.12	2.39%
9	<i>Coca Cola</i>	\$55.13	\$56.75	1.62	2.94%
10	<i>Disney</i>	\$53.75	\$56.13	2.38	4.43%
11	<i>duPont</i>	\$55.38	\$55.38	0.00	0.00%
12	<i>Kodak</i>	\$51.38	\$51.88	0.50	0.97%
13	<i>Exxon</i>	\$63.25	\$65.00	1.75	2.77%
14	<i>General Electric</i>	\$53.00	\$54.75	1.75	3.30%
15	<i>General Motors</i>	\$39.88	\$41.63	1.75	4.39%
16	<i>Goodyear</i>	\$37.13	\$36.00	-1.13	-3.04%
17	<i>IBM</i>	\$79.88	\$81.13	1.25	1.56%
18	<i>International Paper</i>	\$73.50	\$73.00	-0.50	-0.68%
19	<i>MacDonalds</i>	\$33.00	\$33.88	0.88	2.67%
20	<i>Merck</i>	\$41.63	\$41.88	0.25	0.60%
21	<i>MMM</i>	\$54.38	\$56.13	1.75	3.22%
22	<i>Morgan</i>	\$65.38	\$63.25	-2.13	-3.26%
23	<i>Philip Morris</i>	\$62.00	\$63.38	1.38	2.23%
24	<i>Proctor & Gamble</i>	\$66.13	\$67.25	1.12	1.69%
25	<i>Sears</i>	\$50.38	\$50.75	0.37	0.73%
26	<i>Texaco</i>	\$63.75	\$65.13	1.38	2.16%
27	<i>Union Carbide</i>	\$28.00	\$27.75	-0.25	-0.89%
28	<i>United Technologies</i>	\$66.00	\$66.13	0.13	0.20%
29	<i>Woolworth</i>	\$15.63	\$15.88	0.25	1.60%
30	<i>Westinghouse</i>	\$14.75	\$14.63	-0.12	-0.81%

3. Ditto for 1995, March 3 to March 10 (column 5).

4. Ditto for change during the one year, approximately, between March 18, 1994 and March 10, 1995. Before you begin: What do you expect? Why. And then when you see the Stem and Leaf: What did you find? What questions did it raise? Why? Explain — “What’s going on here?”

5. Construct a stem and leaf diagram for the states of the United States with respect to their rates of infant mortality.

	<i>Division and State</i>	<i>Region</i>	<i>Total Infant Mortality Rate, 1988</i>
1		U.S.	10.0
2		N.E.	8.1
3	Maine	N.E.	7.9
4	New Hampshire	N.E.	8.3
5	Vermont	N.E.	6.8
6	Massachusetts	N.E.	7.9
7	Rhode Island	N.E.	8.2
8	Connecticut	N.E.	8.9
9		M.A.	10.3
10	New York	M.A.	10.8
11	New Jersey	M.A.	9.9
12	Pennsylvania	M.A.	9.9
13		E.N.C.	10.5
14	Ohio	E.N.C.	9.7
15	Indiana	E.N.C.	11.0
16	Illinois	E.N.C.	11.3
17	Michigan	E.N.C.	11.1
18	Wisconsin	E.N.C.	8.4

19		W.N.C.	8.9
20	Minnesota	W.N.C.	7.8
21	Iowa	W.N.C.	8.7
22	Missouri	W.N.C.	10.1
23	North Dakota	W.N.C.	10.5
24	South Dakota	W.N.C.	10.1
25	Nebraska	W.N.C.	9.0
26	Kansas	W.N.C.	8.0
27		S.A.	11.6
28	Delaware	S.A.	11.8
29	Maryland	S.A.	11.3
30	Dist Columbia	S.A.	23.2
31	Virginia	S.A.	10.4
32	West Virginia	S.A.	9.0
33	North Carolina	S.A.	12.5
34	South Carolina	S.A.	12.3
35	Georgia	S.A.	12.6
36	Florida	S.A.	10.6
37		E.S.C.	11.4
38	Kentucky	E.S.C.	10.7
39	Tennessee	E.S.C.	10.8
40	Alabama	E.S.C.	12.1
41	Mississippi	E.S.C.	12.3
42		W.S.C.	9.4
43	Arkansas	W.S.C.	10.7
44	Louisiana	W.S.C.	11.0
45	Oklahoma	W.S.C.	9.0
46	Texas	W.S.C.	9.0
47		Mt.	9.2
48	Montana	Mt.	8.7
49	Idaho	Mt.	8.8
50	Wyoming	Mt.	8.9

51	Colorado	Mt.	9.6
52	New Mexico	Mt.	10.0
53	Arizona	Mt.	9.7
54	Utah	Mt.	8.0
55	Nevada	Mt.	8.4
56		Pac	8.6
57	Washington	Pac	9.0
58	Oregon	Pac	8.6
59	California	Pac	8.6
60	Alaska	Pac	11.6
61	Hawaii	Pac	7.2

Total Infant Mortality Rate, 1988, Measured in Deaths of infants under 1 year old per 1,000 Live Births (Excluding fetal mortality).
Source: *Statistical Abstract of the United States, 1993*, p. 81 Table 112)

The Report: Protein Content of Breakfast Cereals

Common sense would expect data analysis to be cool and logical — with a clear plan and a clear execution. And, indeed, making data analysis look that way, at the end, shows good style: It leads to writing that is clear and to the point. But the straight forward logic of the public report you see in the evening news or read in a scientific journal has, often, little to do with the erratic and roundabout path by which “simple” truth is actually discovered. Data analysis has two phases — *doing* the analysis is one phase, *presenting* the analysis is another. And it would be hard to reverse engineer the rules of data analysis, hard to figure out how it was done, if all you were allowed to see were the final presentation. If nothing else, the presentation, slick, simple, and compelling, usually hides the number of hours of thinking that can lie behind a single graph, not to speak of hiding the bright ideas that the analyst followed to a conclusion only to find, at the conclusion, that the ideas led nowhere. I’m told that John von Neumann, an innovator in mathematics and computer science, once compared the discovery of a mathematical proof to the construction of a great cathedral: Like a cathedral, a mathematical proof is not complete until the scaffolding has been removed. So too with data analysis.

Even the methods I present, in a report, may be different from the methods I used in the act. For example, consider the simple “method” of computing an average: “Everyone knows” what an average is. And it is hard to present a report without writing down a few averages — average income, average age, and so forth. People feel comfortable with this sort of thing and, since you want them to understand your work, you have to accommodate their expectations. But the truth is that the median (the middle value in terms of size) and simple mathematics, which are rarely used in a presentation, are often used during the analysis — we often use one kind of technique when we are doing an analysis, and use other techniques when we are trying to get the idea

across — one kind of technique when we are in the act, another when we communicate..

There is also a difference in attitude toward error, during and after —treating it one way in the report, but treating it another way, and much more aggressively during the pursuit. In the report, error is error, the luck of the draw, random deviation — a real world event subject to the mathematical uncertainties of probability. But in the act, during the pursuit, there is no such thing as error: Every strange event is searched for meaning: Why is this number low while that number is high — and neither of them is average? Is there information here? Is there pattern to the these events?

For homework, give me the “works”. We want to see how you handle the messy process of data analysis. And we want you to clean it up for a report.

Protein Content of Breakfast Cereals

The protein component of common breakfast cereals varies widely from essentially negligible quantities of protein to a large fraction of an adult's entire daily requirement. The purpose of this analysis was to document this variation and to examine the source of this variation. Results from examination of thirty common breakfast cereals confirm that protein content varies from negligible amounts of protein to a substantial fraction of the daily requirement for protein (30-40% grams).

The analysis examined protein content among thirty common breakfast cereals in portions as they are commonly consumed, using data from Bowes and Church, Food Values of Portions Commonly Used, Harper and Row, 1975. The figure below demonstrates the range and variation of the portions: Typically these cereals provide two to three grams of protein per serving, about 5 to 10 percent of the minimum daily requirement. And careful selection of the cereal can find protein content that is two fold, three fold, or even greater than that of low protein cereals. At 10.2 grams of protein per serving one cereal, Gerber's High Protein Cereal, is in a class by itself while Quaker Puffed Rice, at 0.8 grams of protein provided only a negligible fraction of the daily requirement.

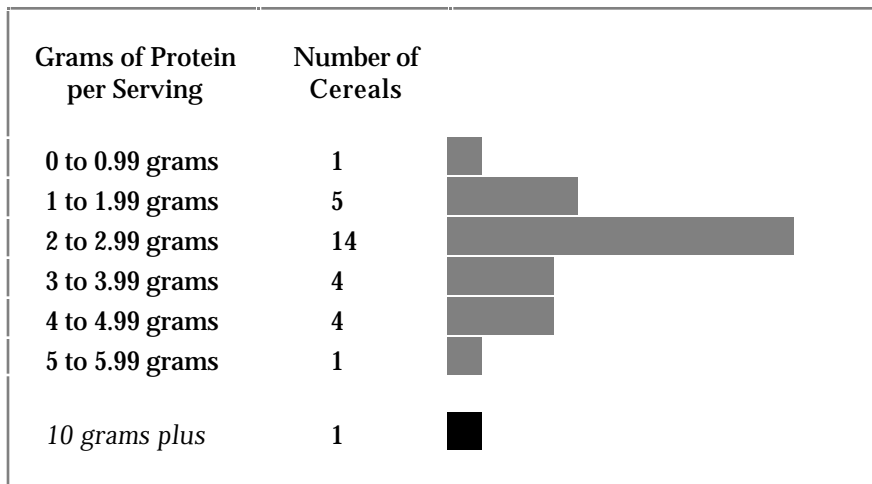
Context: Who, What, Where, ...

Newspaper style: Telegraph important results in the first paragraph. In professional papers, this would be an "abstract"

I looked up the minimum daily requirement to add context.

Source

Brief description: Typical, Extreme High, and Low



Brief Pictorial
 Overview of the Data
 (Less detail than the Stem
 Leaf, but appropriate for this
 report
 Change style to mark the
 break in the distribution of
 protein.

Figure 1
 Range and Distribution of Grams of Protein per Serving for 29 Common
 Breakfast Cereals

While the data do not clearly indicate the principle ingredient for all of these cereals, fragmentary evidence suggests that the principle ingredient is the most reliable predictor of protein content. With considerable variation within groups, the averages range from 1.5 grams of protein for the four rice cereals to approximately 4 grams per serving for the oat and barley cereals, to 5.1 for the mixed corn/soy cereal, to a high of 10 grams of protein for the Gerbers soy cereal, summarizing the data shown in Figure 2.

Rice	Corn	Bran (Grain not Specified)	Wheat
Rice, Puffed, Quaker .8	Corn Fetti, Post's 1.5	Bran Raisin, Kellogg's 1.8	Wheat, Puffed Quaker 1.6
Rice Cereal, Gerbers 1.5	Corn Flakes 2.1	Bran, Raisin, Post's 2.2	Wheat, Shredded 2.2
Rice Krispies, Kellogg's 1.6		Bran Flakes, 40%, Post's 2.8	Wheaties, General Mills' 2.8
Rice Flakes 2.1		Bran Flakes 40% Kellogg's 2.9	Wheat Chex, Ralston 2.8
		Bran, All-Kellogg's 3.1	Wheat Flakes, Quaker 4.4
Average Grams per Serving	Average Grams per Serving	Average Grams per Serving	Average Grams per Serving
1.52	1.8	2.54	2.76
Oats	Barley	Corn Soya	Soy
Cheerios, General Mills' 3.4			
Oatmeal 4.5			
Average Grams per Serving	Barley Cereal, Gerbers	Corn Soya Shreds	High Protein Cereal, Gerber's
3.95	4.3	5.1	5.1

Figure 2

Grams of Protein per Serving and Average Grams of Protein per Serving, organized by Principal Ingredient

Exercise:

Expand your analysis of fat in breakfast cereal, or the stem and leaf of the Dow industrials, or of U.S. infant mortality rates to a complete write up.

Dates:	3/15/96	3/22/96	Change
Dow Jones Industrial Average	5,584.97	5636.64	51.67
	Price per Share of Stock	Price per Share of Stock	
Alcoa Aluminum	\$61.00	\$62.38	\$1.38
Allied Signal	\$56.25	\$57.13	\$.88
American Express	\$48.38	\$48.75	\$.37
AT&T	\$61.38	\$61.25	(-\$0.13)
Bethlehem Steel	\$13.75	\$13.75	\$0.00
Boeing	\$80.88	\$88.88	\$0.00
Caterpillar	\$72.00	\$69.25	(-\$2.75)
Chevron	\$54.88	\$55.25	\$.37
Disney	\$69.25	\$64.88	(-\$4.38)
duPont	\$81.25	\$83.00	\$1.75
Kodak	\$73.13	\$73.13	\$0.00
Exxon	\$79.00	\$81.38	\$2.38
General Electric	\$75.63	\$78.38	\$2.63
General Motors	\$52.25	\$53.38	\$1.12
Goodyear	\$51.25	\$52.00	\$.75
IBM	\$119.88	\$114.25	(-\$5.62)
International Paper	\$39.25	\$38.38	(-\$0.88)
MacDonalds	\$51.25	\$50.75	(-\$0.50)
Merck	\$62.13	\$63.50	\$1.38
Minnesota Mining & Mfg	\$63.50	\$64.50	\$1.00
Morgan	\$80.00	\$83.63	\$3.63
Philip Morris	\$95.38	\$86.25	(-\$9.25)
Proctor & Gamble	\$83.13	\$87.88	\$4.75
Sears	\$50.13	\$51.00	\$.87
Texaco	\$82.88	\$84.75	\$1.88
Union Carbide	\$47.63	\$48.25	\$.63
United Technologies	\$111.00	\$115.38	\$4.38
Woolworth	\$15.88	\$15.75	(-\$0.13)
Westinghouse	\$19.00	\$19.00	\$0.00

Prices of "Dow Jones 30 Industrials", March 15 – March 22, 1996

	<i>Division and State</i>	<i>Region</i>	<i>Total Infant Mortality Rate, 1988</i>
1		U.S.	10.0
2		N.E.	8.1
3	Maine	N.E.	7.9
4	New Hampshire	N.E.	8.3
5	Vermont	N.E.	6.8
6	Massachusetts	N.E.	7.9
7	Rhode Island	N.E.	8.2
8	Connecticut	N.E.	8.9
9		M.A.	10.3
10	New York	M.A.	10.8
11	New Jersey	M.A.	9.9
12	Pennsylvania	M.A.	9.9
13		E.N.C.	10.5
14	Ohio	E.N.C.	9.7
15	Indiana	E.N.C.	11.0
16	Illinois	E.N.C.	11.3
17	Michigan	E.N.C.	11.1
18	Wisconsin	E.N.C.	8.4
19		W.N.C.	8.9
20	Minnesota	W.N.C.	7.8
21	Iowa	W.N.C.	8.7
22	Missouri	W.N.C.	10.1
23	North Dakota	W.N.C.	10.5
24	South Dakota	W.N.C.	10.1
25	Nebraska	W.N.C.	9.0
26	Kansas	W.N.C.	8.0
27		S.A.	11.6
28	Delaware	S.A.	11.8
29	Maryland	S.A.	11.3
30	Dist Columbia	S.A.	23.2
31	Virginia	S.A.	10.4
32	West Virginia	S.A.	9.0
33	North Carolina	S.A.	12.5

Infant Mortality, U.S., by State, 1988e

34	South Carolina	S.A.	12.3
35	Georgia	S.A.	12.6
36	Florida	S.A.	10.6
37		E.S.C.	11.4
38	Kentucky	E.S.C.	10.7
39	Tennessee	E.S.C.	10.8
40	Alabama	E.S.C.	12.1
41	Mississippi	E.S.C.	12.3
42		W.S.C.	9.4
43	Arkansas	W.S.C.	10.7
44	Louisiana	W.S.C.	11.0
45	Oklahoma	W.S.C.	9.0
46	Texas	W.S.C.	9.0
47		Mt.	9.2
48	Montana	Mt.	8.7
49	Idaho	Mt.	8.8
50	Wyoming	Mt.	8.9
51	Colorado	Mt.	9.6
52	New Mexico	Mt.	10.0
53	Arizona	Mt.	9.7
54	Utah	Mt.	8.0
55	Nevada	Mt.	8.4
56		Pac	8.6
57	Washington	Pac	9.0
58	Oregon	Pac	8.6
59	California	Pac	8.6
60	Alaska	Pac	11.6
61	Hawaii	Pac	7.2

Total Infant Mortality Rate, 1988, Measured in Deaths of infants under 1 year old per 1,000 Live Births (Excluding fetal mortality).
Source: *Statistical Abstract of the United States, 1993, p. 81 Table 112*
