exhibit **8** of chapter 3: reference table

Break table for (negative) reciprocals (using −1000/number)

A) BREAK TABLES for SETTING DECIMAL POINT

Break	Start	Start	**Break**
1000	.x	a.	**1000**
10,000	.0x	ab.	**100**
100,000	.00x	abc.	**10**
1,000,000	.000x	abcd.	**1.**
10,000,000	.0000x	abcde.	**0.1**
100,000,000			**0.01**

Examples

Number	A	B	−1000/number
124.2	a.	−80	−8.0
.04739	abcde.	−212	−212**.
1242.	.x	−80	−.80

B) MAIN BREAK TABLE--digits of negative reciprocal

Break	Value	Break	Value	Break	Value	Break	Value	Break	Value
990	−100	**1639**	−60	**2469**	−40	**4115**	−240	**617**	−160
1010	−98	**1681**	−59	**2532**	−39	**4202**	−236	**633**	−156
1030	−96	**1709**	−58	**2597**	−38	**4274**	−232	**649**	−152
1053	−94	**1739**	−57	**2667**	−37	**4347**	−228	**666**	−148
1075	−92	**1770**	−56	**2740**	−36	**4425**	−224	**685**	−144
1099	−90	**1802**	−55	**2816**	−35	**4504**	−220	**704**	−140
1124	−88	**1835**	−54	**2899**	−34	**4587**	−216	**725**	−136
1149	−86	**1869**	−53	**2985**	−33	**4672**	−212	**746**	−132
1176	−84	**1905**	−52	**3077**	−32	**4762**	−208	**769**	−128
1205	−82	**1942**	−51	**3175**	−31	**4854**	−204	**793**	−124
1235	−80	**1980**	−50	**3287**	−30	**4950**	−200	**820**	−120
1266	−78	**2020**	−49	**3367**	−294	**505**	−196	**840**	−118
1299	−76	**2062**	−48	**3448**	−288	**515**	−192	**855**	−116
1333	−74	**2105**	−47	**3509**	−282	**526**	−188	**870**	−114
1370	−72	**2151**	−46	**3584**	−276	**538**	−184	**885**	−112
1408	−70	**2198**	−45	**3663**	−270	**549**	−180	**901**	−110
1449	−68	**2247**	−44	**3745**	−264	**562**	−176	**917**	−108
1493	−66	**2299**	−43	**3831**	−258	**575**	−172	**935**	−106
1538	−64	**2353**	−42	**3922**	−252	**588**	−168	**952**	−104
1587	−62	**2410**	−41	**4016**	−246	**602**	−164	**971**	−102
1639		**2469**		**4115**		**617**		**990**	

See text, page 78, for example of use.

JOHN W. TUKEY
Princeton University and
Bell Telephone Laboratories

Exploratory Data Analysis

ADDISON-WESLEY PUBLISHING COMPANY
Reading, Massachusetts • Menlo Park, California
London • Amsterdam • Don Mills, Ontario • Sydney

exhibit **22** of chapter 3: data and problems

(A) heights of "famous" waterfalls (both highest fall and total fall may appear for one waterfall) and (B) reservoir capacity of major world dams

A) Heights——in feet

0**	40,54,65,66,68,
0··	70,70,75,90,96,98
1**	01,09,15,20,25,25
1··	30,30,30,32,40,44,
1··	50,51,65,68,86
1··	93,95,98
2**	00,07,07,13,14,18
2··	20,30,40,45,51,51
2··	56,66,70,75,68
3**	00,00,08,11,15,17
3··	20,30,30,35,44,45
3··	55,60,70,70,94,94
4**	00,00,06,27
4··	50,59,59,70
5**	00,05,08,18,25
5··	40,42,90,94,97
6	00,20,26,30,40,50,56
7	00,26,41
8	20,30,48,80,89,90
9**	74,84,84
1***	000,100,170,218,250,312
1···	312,312,325,350,385
1···	430,535,600,612
1···	640,650,696,904
2***	000,425,600,648
3	110†,212†

† **3110 Tugela (5 falls), 3212 Angel**

B) CAPACITY——in 1000's of acre feet

0*	8	Speicheri
1*	·	
2	·	
3	2	Curnera
4	1	Zeuzier
5*	4	Alpe Gera
0**	61,70,70,81,81	
1**	14,37,39,46,48	
1··	52,61,86,95	
2	19,65	
3	24	
4	87	
5**	49	
0***	600,602,746,756,930	
1***	261,325,375,405,586,709	
2***	000,030,030,092,095	
2···	106,367,446,717	
3***	024,453,468,484,648,789	
4	413,493,500,500	
5***	·	
6	100,550,600	
7	055,060	
8	000,000,512	
9***	171,402,730,890	
1****	0945,2940,4755	
1····	9715,9000,9400	
2	3600,4500,4800,7000,7160	
3	1618,2471	
4	7020	Kuibyshev
5****	·	
0*****	62000	Portage Mt
1*****	15000‡,27281‡,45115‡	

‡ **115000 Manicouagan #5. 127281 Sadu-El-Aali (High Aswan) 145115 Bratsk**

P) PROBLEMS

22a) Panel A gives the heights of "famous" waterfalls. What expression is indicated? Make the corresponding stem-and-leaf. Comment.

22b) Panel B gives the reservoir capacity of major dams. What expression is indicated? Make the corresponding stem-and-leaf. Comment.

S) SOURCE

The World Almanac, 1966: (A) page 286; (B) page 260.

Effective comparison, including well-chosen expression

4

chapter index on next page

In dealing with batches, we have already given some attention to two reasons for choosing one form of expression rather than another:

- ◇ symmetry of spreading within each batch separately.
- ◇ agreement from batch to batch in amount of spreading.

Neither is a "big deal."

Symmetry of spread is, by itself, probably a "little deal."

Fortunately for all of us, however, conflict between these reasons is infrequent. A choice good for one is, much more often than not, a choice good for the other.

Agreement of spread is rather more important; it deserves to be a "middle deal".

Indeed, in situations where more important reasons apply, what is good for the present minor reasons is, again much more often than not, good for the major reasons. There are exceptions, and when we recognize one, the major reason will have to take control, but such instances are not common. Choice of expression is rarely a balancing of conflicting reasons. Much more often it is a matter of stretching the information the data offers about alternative choices far enough to make a choice.

Most batches of data fail to tell us exactly how they should be analyzed. In making our most careful choices, we typically have to depend upon other bodies of data on similar subjects and upon experience.

Choice of expression is only one of such choices.

This does not mean that we cannot do relatively well on the basis of the data before us; often, indeed we can do tremendously better than if we used the data in a raw form, or made other choices without thought.

We will not meet "big deals" until we come to deal with more structured data——data with more "handles" for us to work with.

review questions

What are two reasons for choice of expression? How important are they? Do they usually agree with each other? With major reasons about which we have not yet learned? Does a body of data usually indicate clearly how it should be analyzed?

4A. Alternative forms of display of summaries

Suppose we know how we want to analyze our data and what sorts of summaries we want to suggest or present. There is always still a question of presenting them. There are choices among alternatives, sometimes among more alternatives than we think of at first.

Rainfall (including snow converted to rain) at New York City offers a simple and convenient example. Exhibit **1** shows the data for the first six years of each of seven decades. Exhibit **2** shows four ways to display the data in more or less summarized form:

- as stem-and-leaf displays.
- as medians (unadorned).
- as slightly graphic medians.
- as schematic plots, modified to emphasize medians.

exhibit **1** of chapter 4: New York City rainfall

Precipitation* in New York City (to nearest inch) in years ending with 0, 1, 2, 3, 4, and 5

A) DATA

	189-	190-	191-	192-	193-	194-	195-
−0	52	42	36	49	35	45	45
−1	41	47	40	34	36	36	47
−2	39	47	38	43	39	50	46
−3	53	49	44	37	50	40	38
−4	44	42	34	38	45	52	43
−5	36	44	41	37	33	46	41
median	**42h**	**45h**	**39**	**37h**	**37h**	**45h**	**44**

* Rain plus rain equivalent of snow.

P) PROBLEM

1a) Find the data for 1960 to 1965, and use it to extend exhibit 2, below, to this additional decade.

S) SOURCE

Available from a variety of sources, including The World Almanac.

exhibit **2** of chapter 4: New York City rainfall

Four kinds of summaries of New York City precipitation compared

A) STEM-and LEAF COMPARISON––unit 1 inch

	1890–95	1900–05	1910–15	1920–25	1930–35	1940–45	1950–55
5·					0	02	
5*	23				5	56	567
4·		779		9		0	13
4*	14	224	014	3		6	8
3·	69		68	778	569		
3*			4	4	3		

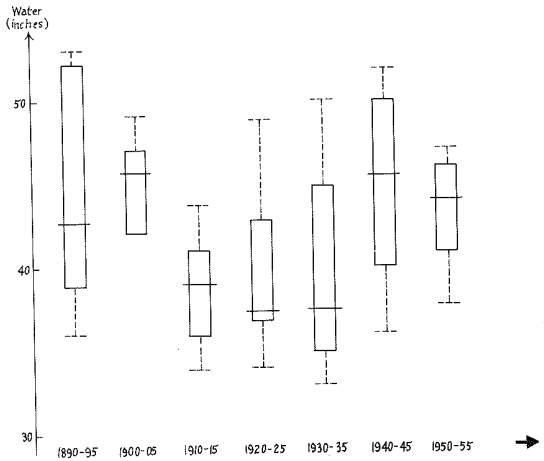
B) MEDIANS––unadorned

1890–95	1900–05	1910–15	1920–25	1930–35	1940–45	1950–55
42.5	45.5	39.0	37.5	37.5	45.5	44.0

C) PARTLY GRAPHIC MEDIANS

44:45		45h			45h	44
42:43	42h					
40:41						
38:39			39			
36:37				37h	37h	

D) SCHEMATIC PLOTS––modified to stress medians



All four displays tell the same story:

- in the 1890's, 1900's, 1940's, and 1950's, high typical annual rainfalls, near 44 inches;
- in the three intervening decades, lower typical rainfalls, near 38 inches.

There is now an understandable message––1910 to 1935 seems to have been a period of lowered rainfall. Whether or not this ought to be regarded as accidental is a confirmatory question, a question we have not tried to answer––and will not try to answer here. However, it should be regarded––since anything that happens once may happen twice––this earlier low-rainfall period might well have been a warning to prepare for the "drought" of the 1960's. Our concern is that use of any one of these displays uncovers a message in this instance––and that their use in other instances may also uncover messages.

How well does each display do? There will undoubtedly be some differences of opinion, as well there may be. If we compare the stem-and-leaf (panel A) with the schematic plot (panel D), I find little to choose. (If there were 600 values in each batch, rather than 6, we would almost surely prefer the schematic plot.) It is easier to compare the other pair of panels, in which the information about each batch is pared down to one number––the median––with each other. The expenditure of a few extra lines to make the medians partly graphic has shown us directly––with one *coup d'oeil*––most of what is going on, yet there has been no loss of detail in panel C as compared with panel B. We should be on a constant lookout for opportunities to use partly graphic displays.

Comparison of one pair of displays with the other is––and should be––harder. Sometimes we need the additional detail about the spread of the batches. Sometimes it merely confuses us slightly. We have to be guided by what we know or feel about the users, be they us or be they someone else.

review questions

What were the four kinds of display? What story did each tell? How did they pair off? How do the two displays in each pair compare? How did one pair compare with the other?

exhibit **2** of chapter 4 (continued)

P) PROBLEM

2a) Make a display like panel C, showing values for hinges (light) as well as for medians (bold), omitting hinges when they would overlap. How well do you like the result? Compare its effectiveness with panel C.

4B. Comparing several batches (continued)

When medians alone do not tell us enough, we want at least to look at schematic summaries, probably in the form of schematic plots. We want these summaries to tell us the story as clearly and as simply as they can. Both symmetry of spread within batches and, especially, balance of spread between batches, will help.

Exhibit **3** gives--in numbers--summaries of three batches from Winsor and Clarke's (1940) analysis of the catch of plankton using different kinds of nets. Three choices of expression are compared in panels B and C--raw count, root count, and log count. Whether we concentrate on symmetry--looking for midsummaries that do not drift within a batch--or on balanced spread--looking for spreads that agree from batch to batch--we come to the same position:

- ◇ raw counts are unacceptable.
- ◇ we are torn between root counts and log counts, and might even like to find a compromise.

exhibit **3** of chapter 4: Plankton hauls

Estimated number of three kinds of plankton caught (in six hauls of each of two nets)

A) ESTIMATED COUNTS

Kind I: 387, 428, 470, 497, 537, 540, 620, 760, 845, 895, 1020, 1050
Kind IV: 6060, 7600, 7900, 8260, 8600, 8900, 9250, 9830, 10200, 11000, 15500
Kind III: 189, 223, 278, 281, 288, 290, 314, 328, 328, 346, 395, 433 (hundreds)

B) RAW COUNTS--summaries (rounded) in units

	Kind I				Kind IV				Kind III			
M6h	580	580			9075	9075			302**	302**		
H3h	677	484	870	386	9048	8080	10,015	1935	308**	280**	337**	57**
1	718	387	1050		10,780	6060	15,500		311**	189**	433**	

H-spreads INcrease rapidly from left to right; mids INcrease from M toward 1.

C) ROOT COUNTS--summaries (rounded) in units

	Kind I				Kind IV				Kind III			
M6h	24	24			96	96			172	172		
H3h	26	22	29h	7h	95	90	100	10	176	168	184	16
1	26	20	32		101	78	124		172	136	208	

H-spreads INcrease slowly from left to right; mids INcrease very slightly from M towards 1. →

exhibit **3** of chapter 4 (continued)

D) LOG COUNTS--summaries in 0.01 (rounded)

	Kind I				Kind IV				Kind III			
M6h	276	276			396	396			448	448		
H3h	281	268	294	26	396	391	400	9	449	445	453	8
1	280	259	302		398	378	419		446	428	464	

H-spreads DEcrease moderately slowly from left to right; mids nearly neutral.

E) HOW MUCH RE-EXPRESSION for SIMILAR SPREAD?

Kind	**log M**	**log Hspr**
I	2.76	2.59
IV	3.96	3.29
III	4.48	3.76

Diff		**Ratio**
III – I	1.72	1.17	.7
IV – I	1.20	.70	.6
III – IV	.52	.47	.9

Ratios between .5 and 1.0; hence look at $\sqrt{\ }$ and log.

P) PROBLEMS

3a) Extend panel A to include eighths and mideighths. Any change in conclusions?

3b) Extend panel B to include EH-spreads. Any change in conclusions?

3c) Extend panel C to include log EHspr's and corresponding differences and ratios. Any change in conclusions?

3d) Invent a set of batches where logs will make spreads reasonably constant. Do the calculations of panel C. Are the ratios what you would expect?

3e) Invent a set of batches where roots will make spreads reasonably constant. Do the same.

S) SOURCE

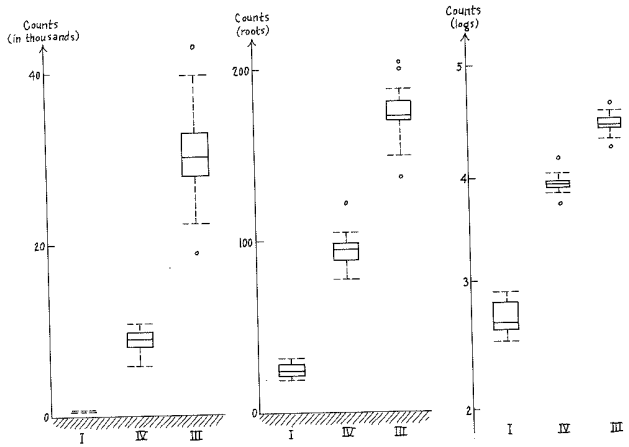
C. P. Winsor and G. L. Clarke (1940). Journal of Marine Research (Sears Foundation) 3: 1.
Also used by: G. W. Snedecor (1946). *Statistical Methods,* 4th ed., page 451.

Exhibit **4** shows the schematic plots graphically. The leftmost, using raw counts, is of little use as a set of schematic plots--we can see hardly more than if the medians were shown. The schematic plot for kind I is too small to see any detail. That for kind III is so large as to claim much undeserved attention. Either of the other two expressions gives about the same information, makes about the same impression. We cannot say that one is clearly better. We can say that either will clearly do.

We ought not to have to calculate through all three cases in order to find which one or two are likely to be satisfactory. Panel E of exhibit 3 introduces a way of learning a fair amount from much less work. If we look at the relation between median and H-spread--also, in suitably large batches, between median and E-spread, etc.--we can learn roughly what change of expression is likely to help. This is easier to do in terms of logs.

exhibit **4** of chapter 4: plankton hauls

Schematic plots for the data of exhibit 3, according to three choices of expression



Accordingly, panel E gives first log median and log H-spread of each batch of raw values, and then the differences in these logs as we step from one batch to another. As the ratios show, the change in log H-spread is rather more than 1/2 the change in log median and rather less than 1 times that change.

Had these ratios been close to 1/2, we would expect roots to help us. Had they been close to 1, we would expect logs to help us. Since they are in the wishy-washy in-between, we may expect either to help us, usefully but not perfectly.

review questions

What do we ask of schematic plots? Why? Where in exhibit 3 do we look to see how well what we ask is provided? Does exhibit 4 convey the same message about choice of expression for these particular data? How can we use the results of analyzing one choice of expression to guide us in what expression is likely to do well for us?

4C. A more extensive example

In 1950, Bruner, Postman, and Mosteller reported the detailed results of a simple experiment in which subjects looked at a flat picture--the Schroeder staircase--which can be easily seen in perspective in two ways. They were given different instructions as to how they were to try to manage their perspective, and the number of changes of perspective were counted in each of 10 successive minutes under each instruction. Setting aside the first two minutes, to protect somewhat against "starting-up" effects, there are 19 batches (one per subject) of 8 counts under the "alternate" instruction.

The early columns of exhibit **5** show the median M and hinges H for each of the 19 batches. To make both looking at and plotting easier, the subjects (rows in the exhibit) have been arranged in increasing order of median counts. These early columns also give log M, H-spread, and log Hspr. The logs are plotted in exhibit **6** and clearly trend upward.

If we ask how fast they trend upward, the easiest way to get a rough answer is to pick out some "typical" points near each end, and then to hold your clear plastic triangle or ruler across the picture, turning and slipping the edge to what seems to you to be a good fit. In the upper panel of exhibit 6, it seems natural to take the two points at the far lower left and the one point at the far upper right. The lower portion of panel A of exhibit 5 calculates the tilts of the two lines thus fixed as

$$\frac{134-48}{216-108}=\frac{86}{108}=.8 \quad \text{and} \quad \frac{134-30}{216-122}=\frac{104}{94}=1.1.$$

exhibit **5** of chapter 4: Schroeder staircase

Calculations for perspective reversals (all logs in 0.01)

A) BASIC CALCULATIONS for THREE EXPRESSIONS

			raw counts			
					H spread	
#	M	log M	H	H	raw	log
2	12	108	11	14	3	48
(*) 9	16h	122	16	18	2	30
13	22	134	21	23	2	30
19	22	134	20	22	2	30
12	28	145	22	48	26	142
4	29h	147	24	32	8	90
16	33h	152	32	35	3	48
15	34	153	33	36	3	48
11	34	153	30	38	8	90
6	36	156	30	41	11	104
3	36h	156	34	38	4	60
18	36h	156	34	44	10	100
5	38h	159	37	41	4	60
14	44	164	43	48	5	70
7	45	165	42	46	4	60
8	64	181	54	67	13	111
1	74h	187	64	98	34	153
10	92	196	86	95	9	95
(*) 17	144h	216	132	154	22	134
17 – 2		108			19	86
diffs 17 – 9		94			20	104
(ratios) (17 – 2)						(0.8)
(tilts) (17 – 9)						(1.1)

P) PROBLEMS

5a) Plot the raw H-spread for raw counts against raw medians (also for raw counts). Is there a trend? Which points near each end of the point cloud would you choose as reasonably typical?

5b) Do the same for the raw H-spread for root counts.

5c) Do the same for the raw H-spread for log counts.

5d) Look up the source and apply a similar analysis to the last 8 minutes under the "natural" instruction.

5e) Do the same for the "hold" instruction.

5f) Find logs of median root counts, and plot against logs of median raw counts. What do you see? Should you have expected it?

5g) Find logs of median log counts and plot against logs and median row counts. What do you see? Should you have expected it?

exhibit **5** of chapter 4 (continued)

Panel A continued

root counts (*p)				log counts			
		H spread				H spread	
H	H	raw	log	H	H	raw	log
33	37	4	60	104	115	11	104
40	42	2	30	120	126	6	78
46	48	2	30	132	136	4	60
45	47	2	30	130	134	4	60
47	69	22	134	134	168	34	153
49	57	8	90	138	151	13	111
57	59	2	30	151	154	3	48
57	60	3	48	152	156	4	60
55	62	7	85	148	158	10	100
55	64	9	95	148	161	13	111
58	62	4	60	153	158	5	70
58	66	8	90	153	184	11	104
61	64	3	48	157	161	4	60
66	69	3	48	163	168	5	70
65	68	3	48	162	166	4	60
74	82	8	90	173	183	10	100
80	99	19	128	181	200	19	128
93	98	5	70	193	198	5	70
115	124	9	95	212	219	7	85
		5	35			–4	–19
		7	65			+1	7
			(.3)				(–.2)
			(.7)				(.1)

PROBLEMS (continued)

5h) In view of (5f) and (5g), do you think it would have mattered if the center and lower panels of exhibit 6 had used logs of median root counts and log counts, respectively, in place of logs of median row counts? Check your answer for the lower panel.

5i) Do you think such a choice will ever matter? If not, why not? If yes, give an example.

5j) Look up the Bruner, Postman, and Mosteller paper, and apply the condensed approach (see text) to the data for the "natural" instruction.

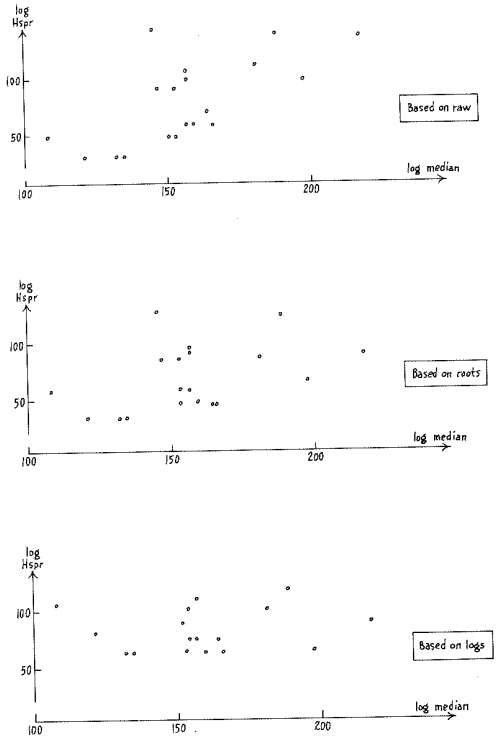
5k) Find a set of data of about the same size that interests you and carry through a similar analysis.

S) SOURCE

J. S. Bruner, L. Postman, and F. Mosteller (1950). "A note on the measurement of reversals of perspective." Psychometrika 15: 63–72. Table 1 on page 65.

exhibit **6** of chapter 4: Schroeder staircase

Plots of log H-spread against log median for the perspective reversal examples



Clearly, the slope is about 1. If the rule of thumb mentioned above works, this would lead us to expect that logs will do well as an expression that keeps the H-spread from trending.

Later columns of exhibit 5 show calculations of hinges, H-spread, and log H-spread for both expression in roots and expression in logs. The center and lower panels of exhibit 6 show the corresponding plots. The trend is clearly smaller for roots and negligible for logs.

For reasons pretty well explained by the answers to problems (5f) to (5i), we have not bothered to change the horizontal scales in the center and lower panels of exhibit 6. Instead we have used log median (raw count) throughout. (This makes the three pictures easier to compare; but this is a little thing, since our next step is to learn how to make one picture do.)

a condensed approach

If we had wanted to hold our calculation to a reasonable minimum, we could have avoided most of exhibits 5 and 6. Since we are dealing with counts, we expect to use either roots or logs. So we may go ahead as follows:

- ◇ find raw median and hinges; find log raw median.
- ◇ find root hinges; find H-spread for roots; find logs of same.
- ◇ make picture of log H-spread (for roots) vs. log M (for raws).
- ◇ select typical points; find tilts for same.
- ◇ working with the typical points only, find hinges for other expressions; find H-spreads; find logs of same.
- ◇ compare tilts of log H-spread for typical points.
- ◇ choose expression. (If we're lucky, it will be the one we started with.)

In this example, the tilt (against log median of raw counts) drops by almost 1/2 at each step of the re-expression scale (1.1 for raws, 0.7 for roots, 0.1 for logs). Dropping by about 1/2 happens very often.

Note also the unusually high position (on any of the three plots) of one subject who turns out to be #12. Turning back to the original data, we find that his ten counts, in order of occurrence were:

30, 22, 14, 22, 24, 32, 18, 48, 52, 53.

It would appear that after seven minutes this subject suddenly learned how to make his perspective reverse rapidly.

review questions

To speed our calculations what kind of points should we select from our first picture? How can we use them to find a tilt? What did the rule of thumb say about tilts? How well did it apply in this example? What expressions should we try first with counted data? Was any one of the 19 subjects used by Bruner, Postman, and Mosteller unusual? How? Are you surprised? Why/ why not?

4D. The meaning of comparison

We have said much about comparison. We have used many pictures and many numbers to make comparisons. We have not yet said enough about comparison itself. It is time we did.

Two kinds of comparisons come up in the simplest of common language: "Bill is a head taller than Jim." "George weighs twice as much as his brother Jack." Each of these statements says what you must do to one person to make him match the other. The first statement bases itself on a plus sign, and says how much we would have to add to Jim's height to match Bill's. The second bases itself on a times sign, and says what we would have to multiply Jack's weight by to match George's.

There are many ways in which addition is simpler than multiplication; doing hand arithmetic and sliding things bodily along graphs are only two. Logarithms were invented centuries ago to reduce multiplication to addition, thus making hand arithmetic much easier. We can, do, and should use them to avoid comparisons by multiplication. When multiplication would otherwise seem needed, taking logs will let us deal with comparison by addition.

When we think of comparison, then, we want to think of what needs to be added or subtracted to the data--as carefully expressed for analysis--in order to make one thing match another. If the data is originally, or commonly, expressed in another way, we will try to translate back to that expression, too.

Thus, if boys' weights are really better thought of as "twice" or "three-quarters" of one another, we will do our analysis in log weight, finding, perhaps, that we have to add 0.30 to Jack's to match George's. Having found this, we would also convert back to raw weight, and say that Jack's had to be doubled in order to match George's. Both forms of answer help.

review questions

What kinds of comparison come up in common language? How ought we think of such comparisons? How many forms of answer help?

4E. Adjustments, rough and exact

rough adjustments

If we want to put several batches under the same microscope at the same time, we need to have them centered at about the same level. If they are not

naturally somewhat similarly centered, we will have to adjust them so that they will be.

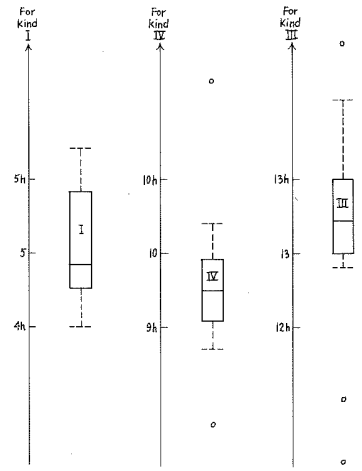
Thanks to the flexibility of the human eye, it is not important--if we are to look at a picture--that the centering be exactly the same. Rough adjustments will serve us essentially as well as precise ones.

Exhibit **7** returns to the catch-of-plankton example. For variety, we use an expression that compromises between roots and logs. (Recall that logs play the

exhibit **7** of chapter 4: plankton hauls

Rough residuals for the plankton counts plotted in terms of $\sqrt[4]{\text{count}} = \sqrt{\sqrt{\text{count}}}$ (see exhibit 1 for data)

A) PLOT of ROUGH RESIDUALS--of plankton counts



B) PROBLEMS

The fourth roots used to make the plots above were NOT obtained by two applications of exhibit 7 of chapter 3.

7a) Obtain them by two such applications, when the values will differ somewhat, and make the plots analogous to the above.

7b) Do you think the difference is important?

role of a zero power; clearly 1/4 is a natural compromise between 1/2 and 0.) Thus expression is

$$\sqrt[4]{\text{count}} = \sqrt{\sqrt{\text{count}}}$$

As exhibit 7 shows, when expressed in this way, the catches are in the vicinity of 5, 10, and 13 for Kinds I, IV and III, respectively.

In exhibit 7, then, we have shifted each schematic plot by a different amount. We have chosen round amounts, so that the three schematic plots do not exactly line up. We do have them under a common microscope, however, and can see clearly their general similarity in spread, both from hinge to hinge and from extreme to extreme.

exact adjustment

Round values and approximate adjustment save arithmetic if we are going to make a plot. If we want to present the same information numerically, we need to do subtractions to get small numbers. Since it is about as easy to subtract one number as another, it usually pays to line something up exactly. Exhibit **8** shows what happens in the catch-of-plankton example if we line up the medians exactly. Clearly the results are rather easy to compare, or even to combine.

exhibit **8** of chapter 4: plankton hauls

Results of lining medians up by adjustment (subtracting a constant) in the catch-of-plankton example expressed in log counts (all logs in .01's)

A) LETTER-VALUE DISPLAYS--for logs in 0.01

Kind I		Kind IV		Kind III	
	276		396		448
268	294	391	400	445	453
259	302	378	419	428	464

B) SAME for SHIFTED LOGS

Kind I down 276		Kind IV down 396		Kind III down 448	
	0		0		0
− 8	18	− 5	4	− 3	5
−17	26	−18	23	−20	16

P) PROBLEMS

8a) Imitate panel B for the catches of plankton expressed in roots (see exhibit 3 for original form). What seems to be going on?

8b) Do the same for the raw counts.

8c) Can you find the median of the three displays in panel B? Why/why not?

review questions

Can rough adjustments serve as well as exact ones? What light does exhibit 7 shed on this question? If we are not making pictures, should we use rough adjustments? Why?

4F. Residuals

If we wanted to know about the variations in plankton catches in rather general terms, we might like to combine all the knowledge that we can gain from the 3 batches of 12 at hand. If we define

$$\text{residual} = \text{given value} - \text{summary value}$$

we can turn each given value into a residual, for example by using the median of the corresponding batch as the summary value.

For the catch-of-plankton data, we dare not combine residuals based on raw counts, but we might try doing this for either root counts or log counts.

Exhibit **9** gives the log counts, their residuals from the median, and--in stem-and-leaf form--the result of pooling the three sets together. (A tendency to straying may be suggested by the E-spread being considerably more than twice the H-spread).

While residuals are useful as an aid to pulling together information from several batches (a rudimentary form of what might be called mustering and borrowing strength), **we will soon learn much more important uses for them:**

- ◇ as keys to the successive step-by-step improvement of our analyses.
- ◇ as keys to turning an investigator's eye toward the adequacy of our current analysis.

exhibit **9** of chapter 4: plankton hauls

Combining residuals for the catches of plankton expressed in logs (data and source, exhibit 3)

A) LOGS of COUNTS in 0.01's--two in median bold

Kind I: 259, 263, 267, 270, 273, **273, 279,** 288, 293, 295, 301, 302

Kind IV: 378, 388, 390, 392, 393, **395, 397,** 398, 399, 401, 404, 419

Kind III: 428, 435, 444, 445, 446, **446, 450,** 452, 453, 454, 460, 464

B) RESIDUALS from MEDIAN--for logs of counts in 0.01's.

Kind I: −17, −13, −9, −6, −3, −3, 3, 12, 17, 19, 25, 26

Kind IV: −18, −8, −6, −4, −3, −1, 1, 2, 3, 5, 8, 23

Kind III: −20, −13, −4, −3, −2, −2, 2, 4, 4, 6, 12, 16 →

review questions

What is a residual? Can there be more than one set of residuals for a given set of data? What are three ways in which residuals can serve us? Which are the more important?

exhibit **9** of chapter 4 (continued)

C) RESIDUALS POOLED

```
 2    2·  56
 3    2*  3
 6    1·  796
 8    1*  22
11    0·  586
18    0*  3123244
18   −0*  334314322
 9   −0·  9686
 5   −1*  33
 3   −1·  78
 1   −2*  0
```

36 pooled residuals

adj: −20(III), 23(IV)

out: 26(I)

	(mid)		(spr)
M18h	0	0	
H 9h	1	−5 7	12
E 5	2	−13 17	30
1	3	−20 26	46

	18
f	−23 25
	xxx one
F	43
	xxx

P) PROBLEMS

9a) Make a comparative plot of 4 batches, the 3 separate batches of residuals above and the pooled batch.

9b) Pool the residuals for the data and mode of expression of (5j) above.

9c) Same for (5k) above.

9d) Same for a new set of self-selected data.

4H. How far have we come?

In this chapter we have begun to take seriously one of the most important things we can do with data––comparison. Most of our attention has been devoted to learning how to strengthen comparisons––so that more nearly the full value of what the data offers us is put to use. We had already learned many of the techniques we need to do this––re-expression, useful display, subtraction. We now put them more carefully to work.

We are now ready:

- ◇ to make some kinds of slightly graphic summaries––displays that show forth the main issue without loss of numerical detail.
- ◇ to use plots of log spread against log level to help us judge what re-expression to try next.
- ◇ to calculate residuals by subtracting summary values from individual values. (We will learn about more general kinds of residuals in later chapters.)

Both **symmetry** of spreading-out and **similar extent** of spread make for useful comparison. While these are little and middle deals, respectively, the re-expression they point to is likely to be one that a BIG DEAL (when we have one) will ALSO prefer. We can seek such an expression by trial and error. With more data, though, it helps to use a systematic approach, plotting the log of a simple measure of spread against the log of a simple measure of level, and using the apparent tilt as a guide to both direction and amount of change.

If we want to compare in detail––rather than looking at the broadest questions––we want to adjust what is being compared so that the gross differences are taken out of the way.

Comparison, in common language, is either a matter of **difference** or a matter of **ratio**. Ratios almost always are of amounts or counts, which are never negative. Logs were invented to bring such ratios to differences. (Having values that are not fenced in, in particular are not kept from being negative, generally is likely to be helpful.)

Most importantly, we have just begun to learn that NO BODY OF DATA TELLS US ALL we need to know ABOUT ITS OWN ANALYSIS. It always takes information and insight gained from other, parallel bodies to let us analyze our body of data as well as we can. (If we don't have it, we do as well as we can.)

We are now well started on the analysis of the simplest kinds of data. We have a good chance of making an effective analysis: expressing the given values usefully, summarizing batches well enough for exploration, calculating residuals so that we can try to look deeper.

4P. Additional problems

See exhibits **10** through **17**.

exhibit **10** of chapter 4: data and problems

Percent of major party vote for Republican presidential candidates in eight southwestern states

A) DATA

	% for Nixon in 1960	% for Goldwater in 1964
Arizona	55.6%	50.5%
California	50.3%	40.8%
Colorado	54.9%	38.4%
Nevada	48.8%	36.1%
New Mexico	49.6%	40.6%
Oklahoma	59.0%	44.3%
Texas	50.5%	36.6%
Utah	54.8%	45.3%
(All U.S.)	(49.9%)	(38.7%)

P) PROBLEMS

10a) The snowfalls in New York City for 20 consecutive winters from 1918–19 to 1937–38 were (in inches) 3.5, 55.4, 18.2, 29.7, 55.2, 26.3, 27.9, 35.8, 21.9, 14.3, 13.3, 13.5, 9.7, 5.1, 24.5, 53.1, 29.0, 32.8, 11.9, 13.9. Those from 1938–39 to 1957–58 were: 31.9, 22.2, 35.0, 10.2, 27.6, 26.0, 26.7, 26.6, 33.2, 61.5, 43.0, 10.4, 10.9, 14.4, 9.1, 17.1, 10.9, 29.8, 19.1, 37.9. Compare these two batches in all the ways used in exhibit 2.

10b) Panel A gives, for two elections, percent of the major party vote for Republican presidential candates in eight southwestern states. Make schematic plots for:

- the 1960 data.
- the 1964 data.
- the change (swing) from 1960 to 1964.

Discuss your results.

10c) Choose another group of eight states, find the data, and make the same plots as for (10b).

S) SOURCE

Richard M. Scammon. America at the Polls: A Handbook of Presidential Election Statistics, 1920–1964. Pittsburgh: University of Pittsburgh Press (1965) 521 pp.

exhibit **11** of chapter 4: data and problems

Seasonal snowfall in Buffalo, New York and Cairo, Illinois, from 1918–19 to 1937–38 (inches)

A) DATA

	Buffalo	Cairo
1918–19	25.0	1.8
1919–20	69.4	4.5
1920–21	53.5	13.9
1921–22	39.8	4.0
1922–23	63.6	1.2
1923–24	46.7	6.8
1924–25	72.9	7.2
1925–26	79.6	11.5
1926–27	83.6	6.2
1927–28	80.7	0.4
1928–29	60.3	11.5
1929–30	79.0	12.4
1930–31	64.8	11.3
1931–32	49.6	2.9
1932–33	54.7	7.4
1933–34	71.8	2.7
1934–35	49.1	1.6
1935–36	103.9	14.1
1936–37	51.6	5.4
1937–38	81.6	3.0

P) PROBLEMS

11a) The snowfalls in Buffalo, New York and Cairo, Illinois for the 20 consecutive winters from 1918–19 to 1937–38 are given in Panel A. Make schematic plots for both places, for this time period.

11b) What light do these two batches throw on how they should be expressed?

S) SOURCE

Report of the Chief of the Weather Bureau, 1918–1919 to 1934–1935, and U.S. Meteorological Yearbook 1935 to 1938.

exhibit **12** of chapter 4: data and problems

Consistency of drape measurements

A) DRAPE, measured by AREA in SQUARE INCHES

Sample A	Sample B
28.92	50.04
28.82	49.94
28.96	40.08
28.89	50.27
28.96	50.03
28.85	50.06
28.97	50.00
28.89	49.99
29.00	49.75
28.99	49.94

P) PROBLEMS

12a) In 1950, Chu, Cummings, and Teixeira gave the results of drape tests in panel A. Adjust schematic summaries to equal medians. Does expression in square inches seem to give equal spread?

12b) Compare log H-spreads with log medians. What expression is suggested? Repeat (12a) for this expression.

S) SOURCE

C. C. Chu, C. L. Cummings, and N. A. Teixeira, 1950. "Mechanics of elastic performance of textile materials. Part V: A study of factors affecting the drape of fabrics—The development of a drape meter," Textile Research Journal 20: 539–548.

exhibit **13** of chapter 4: data and problem

Microdetermination of carbon monoxide in air

A) PARTS PER MILLION of CARBON MONOXIDE

Sample	Standard*	Trial results †
A	90.5	95,96,92,102,103,93,101,92,95,90
B	184.6	184,202,215,204,195,201,201,169,182,192,
C	44.8	40,54,42,49,64,62,50,67,64,43,
D	320	261,279,281,278,269,264,266,261,266,276,
E	244.7	215,214,197,216,215,208,226.208,216,214,
F	25.8	26,23,25,25,21,22,27,27,21,25,
G	66.2	56,55,61,57,60,57,65,55,60,61,
H	137.8	128,119,119,123,117,122,127,121,122,119,
I	137.8	155,142,146,149,149,146,152,159,

*** I_2O_5 method**
† Beckman–McCullough method, in order of analysis

P) PROBLEM

13a) In 1948, Beckman, McCullough, and Crane gave the results in panel A of microdeterminations of carbon monoxide in air. Adjust schematic summaries to equal medians.

S) SOURCE

A. O. Beckman, J. D. McCullough, and R. A. Crane, 1948. "Microdetermination of carbon monoxide in air," Analytical Chemistry 20: 674–677.

exhibit **14** of chapter 4: data and problem

Polarographic determination of aluminum

A) % Al_2O_3--oxygen-containing samples--in 0.01%

Certified value	Sample	Determinations
416	Limestone	**3**** | 92,94 **4** | 16,36
196	Silica brick	**19*** | 02336 **20*** | 04
191	Magnetite ore	171,175,177,190,210
189	Soda-lime glass	159,161,214,214
103	Iron ore	**10*** | 336799 **11*** | 236678
6.7	Dolomite	56,56 (in 0.001%)

B) % Al_2O_3--alloys--in 0.01%

113	Manganese bronze 62	**11*** | 444468
097	Manganese bronze 62b	**8*** | 9 **9*** | 4799
106	Nitralloy steel 106	**9*** | 5 **10*** 88 **11*** 01
107	Nitralloy steel 106a	**10*** | 08 **11*** | 022
026	High silicon steel	**25*** | 18 (in 0.001%) **26*** | 009 **27*** | 0

P) PROBLEM

14a) In 1950, Willard and Dean gave the results in panel A on the measurements of aluminum content. Find residuals. How should they be pooled? Combine them and discuss the results.

S) SOURCE

H. H. Willard and J. A. Dean (1950). "Polarographic determination of aluminum: Use of an organic reagent." Analytical Chemistry 22: 1264–1267. Table II on p. 1266.

exhibit **15** of chapter 4: data and problem

Kjeldahl ultramicrodetermination of nitrogen

A) The BATCHES--hundredths of microliters of 0.01N Hcl for a 10.42 microliter sample

Sample	Material	Determinations
A	Acetanilide	**6**** | 60,78,81, **7**** | 03,57,60,72,
B	Acetanilide	**6**** | 13,37,39,46, **7**** | 47,59,75,
C	$C_{14}H_{22}N_2O_2S$	**6**** | 25,63,67 **7**** | 32,43,63,
D	$C_{13}H_{20}N_2O_2S$	**6*** | 56,59,65,

P) PROBLEM

15a) In 1950, Kuck, Kingsley, Kinsey, Sheehan, and Swigert gave the results in exhibit 15 on the ultramicromeasurement of nitrogen. Calculate and pool residuals. Can these data show how they should best be expressed? If not, why? If yes, show how.

S) SOURCE

J. A. Kuck, A. Kingsley, D. Kinsey, F. Sheehan, and G. F. Swigert 1950. Kjeldahl ultramicrodetermination of nitrogen: Applications in the industrial laboratory. Analytical Chemistry 22: 604–611.

exhibit **16** of chapter 4: data and problems

Precipitation of platinum sulfide by various methods

A) AMOUNTS of PRECIPITATE--calculated as milligrams of platinum

Method	Platinum recovered from 10.12 mg Pt
Treadwell and Hall	**10.2*** 89
	10.3* 13334
Gilchrist and Wichers	**10.2*** 2334789
Hillebrand and Lundell	**10.2*** 25
	10.3* 03688

P) PROBLEMS

16a) In 1942, Geoffrey Beall ("The transformation of data from entomological field experiments so that analysis of variance becomes applicable," *Biometrika* **32**: 243–262) counted the number of *Phlegethontius quinquemaculata* (an insect) per plot after various insect control treatments. Slight simplifications of his results include (letters are treatments): A; 10, 7, 20, 14, 14, 12, 10, 23, 17, 20, 14, 13. B; 11, 12, 21, 11, 16, 14, 17, 17, 19, 21, 7, 13. C; 0, 1, 7, 2, 3, 1, 2, 1, 3, 0, 1, 4. D; 3, 5, 12, 6, 4, 3, 5, 5, 5, 5, 2, 4. E; 3, 5, 3, 5, 3, 6, 1, 1, 3, 2, 6, 4. F; 11, 9, 15, 22, 15, 16, 13, 10, 26, 26, 24, 13. What choice of expression is indicated if we take the counts for each treatment as a batch?

16b) In 1950, Jackson and Beamish gave the results in panel A on the precipitation of platinum sulfide. Compare the quality of the three methods. Find the residuals. Should they be pooled?

16c) In 1952, S. P. Hersh and D. J. Montgomery ("Electrical resistance measurements on fibers and fiber assemblies," *Textile Research Journal* **22**; 805–818) gave these resistivities (in 10^9 ohm-cm^2/cm) for the fibers indicated (at the relative humidities indicated in parentheses): Nylon monofilament--340 denier (60%): 12, 13, 12, 12, 16, 13, 12, 13. Nylon monofilament--30 denier 64%): 32, 26, 26, 29, 29, 41, 32. Nylon two--3 denier (85%): 1.00 (six times), 0.95 (three times), 1.10, 0.90. Human hair ~0.001 diameter (85%): 5.7, 6.8, 6.2, 6.2. Wool--Columbia 58's (85%): 0.095, 0.092, 0.089, 0.075, 0.050, 0.087. Make the best graphical presentation and comparison you can. (What do you think "resistivity," "denier," and "relative humidity" mean?)

S) SOURCE

D. S. Jackson and F. E. Beamish, 1950. "Critical examination of platinum sulfide precipitation," Analytical Chemistry 22: 813–817.

exhibit **17** of chapter 4: data and problems

Head breadths of termites (in hundredths of mm)

A) The DATA

	Nest 668	Nest 670	Nest 672	Nest 674	Nest 675
Small soldiers	227.3	247.9	249.4	244.7	245.6
	233.2	260.3	245.7	238.8	262.6
	237.5	261.3	245.2	251.5	263.3
	237.3	255.7	239.6	244.5	248.7
	231.8	237.7	277.9	231.2	241.0
Large workers	214.2	231.5	235.9	234.9	236.0
	213.9	246.3	234.4	230.0	247.8
	222.5	248.5	239.4	243.6	253.9
	235.2	247.7	236.4	242.3	246.1
	226.9	233.4	226.9	230.2	231.1

P) PROBLEMS

17a) In 1948, an anonymous correspondent (Query 60, *Biometrics* **4**: 213–214) gave weights (in grams) of ducks (at age six weeks) after various protein supplements as follows (protein source in parentheses): (Horsebean) 179, 160, 136, 227, 217, 168, 108, 124, 143, 140. (Linseed oil meal) 309, 229, 181, 141, 260, 203, 148, 169, 213, 257, 244, 271. (Soybean oil meal) 243, 230, 248, 327, 329, 250, 193, 271, 316, 267, 199, 171, 158, 248. (Sunflower seed oil meal) 423, 340, 392, 339, 231, 226, 320, 295, 334, 322, 297, 318. (Meat meal) 325, 257, 303, 315, 380, 153, 262, 242, 206, 344, 258. (Casein) 368, 390, 379, 260, 404, 318, 352, 359, 216, 222, 283, 332. What expression is indicated?

17b) Make an effective graphical presentation and comparison of the data given in (17a). (What do you think "sunflower seed oil meal" is?)

17c) In 1909, Warren gave the head measurements on termites in panel A. Find residuals. Should they be combined? If not, why? If yes, show how.

S) SOURCE

Ernest Warren, 1909. "Some statistical observations on termites, mainly based on the work of the late Mr. G. D. Haviland." Biometrika: 6: 329–347.