Fundamental Statistical Concepts in Presenting Data

Principles for Constructing Better Graphics

Rafe M. J. Donahue, Ph.D.

Director of Statistics
Biomimetic Therapeutics, Inc.
Franklin, TN

Adjunct Associate Professor
Vanderbilt University Medical Center
Department of Biostatistics
Nashville, TN

Version 2.11
July 2011

This text was developed as the course notes for the course Fundamental Statistical Concepts in Presenting Data; Principles for Constructing Better Graphics, as presented by Rafe Donahue at the Joint Statistical Meetings (JSM) in Denver, Colorado in August 2008 and for a follow-up course as part of the American Statistical Association's LearnStat program in April 2009. It was also used as the course notes for the same course at the JSM in Vancouver, British Columbia in August 2010 and will be used for the JSM course in Miami in July 2011.

This document was prepared in color in Portable Document Format (pdf) with page sizes of 8.5in by 11in, in a deliberate spread format. As such, there are "left" pages and "right" pages. Odd pages are on the right; even pages are on the left.

Some elements of certain figures span opposing pages of a spread. Therefore, when printing, as printers have difficulty printing to the physical edge of the page, care must be taken to ensure that all the content makes it onto the printed page. The easiest way to do this, outside of taking this to a printing house and having them print on larger sheets and trim down to 8.5-by-11, is to print using the "Fit to Printable Area" option under Page Scaling, when printing from Adobe Acrobat. Duplex printing, with the binding location along the left long edge, at the highest possible level of resolution will allow the printed output to be closest to what was desired by the author during development.

Note that this is version 2.11. A large number of changes and enhancements have been made, many of them prompted by many kind readers (MD, HD, BH, WZ, PD, TK, KW, JF, and many others!) who have offered constructive criticism and feedback based on the original Version 0.9 that was used with the JSM class in July 2008 and on Version 1.0 that has been in place since early 2009. The author is aware, however, of the very high possibility of there still being any number of typos, grammatical errors, and misspellings. As always, gently alerting the author (via email at rafe.donahue@vanderbilt.edu or in person) will greatly improve upcoming versions. Thank you.

This book carries the rather promising, and perhaps over-achieving, subtitle "Principles for Constructing Better Graphics"; what does that imply about what you will find in the text that follows? What, specifically, are *better graphics?*

The risk of using a word like *better* is that there can be disagreements as to what that word actually means. In order to avoid these disagreements, let me state up front what I mean when I refer to *better graphics*.

I believe that a fundamental purpose of statistical graphics is to improve understanding of what the data mean since we collect data in an effort to make inference about some process about which we might not have perfect understanding. Consequently, we are seeking to be *inferentialists*: we seek to make inference and we want that inference to be valid; we want to avoid being duped or fooled or misled. We want our inference to allow us to make accurate and reproducible descriptions about the past and predictions about the future. We want legitimate inference, inference that will stand up to scrutiny and all manners of attack. We want justifiable, tenable, and defensible conclusions. We are not seeking to spin or manipulate or exaggerate.

We want understanding.

Thus, *better graphics* will promote understanding the data and the process from which they came. They will not necessarily have "pop" or "flash" or be "attention-grabbing." They will focus on the data and will let the data speak.

What is in the book?

This book is a collection of examples and descriptions of what to see and what works and what doesn't. It grew from a series of lectures and informal talks I gave while at Vanderbilt and reached its current form as the course notes for this class. At its heart, it is a collection of things I have learned, things I have ~~stolen~~ ~~lifted~~ ~~borrowed from~~ found that others have done, things I have discovered on my own. It is "Hey, look at this!" and "Here's why this works" and "This is the general principle that can be used elsewhere." It is based on nearly 20 years of post-graduate work as a statistician.

This book is a collection of principles that will help us determine what to do when trying to decide how to display our statistical data. The principles will help us discern better (and worse!) ways to show our data.

This book is about my experience and journey to discover understanding in the field of statistical graphics.

This book is sometimes an example of breaking the very principles that it lays out.

What is not in the book?

This book does not contain any sort of Absolute Truth that can be followed always, in every situation.

This book does not have answers to every question you can ask.

This book does not deliver edicts like “Never use a pie-chart” or “Always make the axes cover every conceivable datum”.

This book is not perfect. There are likely typos and errors. There is opportunity for improvement.

This book is not about pop and flash.

Displaying data, seeing individual data atoms and how they incorporate to synthesize a distribution, goes hand-in-hand with analysis and investigation of data. Pictures are worth more than a thousand words; they are priceless.

The two fundamental acts of science, description and comparison, are facilitated via models. By models, we refer to ideas and explanations that do two things: describe past observations and predict future outcomes.

Brad Efron▸, in an editorial in *AmStat News*, discusses statistics and the rules of science. In particular, he references Richard Proctor's 19th-century maps of future solar transits of Venus — those times in which it will be possible to see, with feet planted firmly on our planet, Venus pass between the earth and the sun. Two such occurrences exposed by Proctor in 1874 are transits predicted on June 8, 2004 and on June 6, 2012. The maps detail the locations on the earth where one can observe this celestial wonder.

▸Bradley Efron was President of the American Statistical Association (ASA) in 2004 and, as such, penned a series of monthly "President's Corner" essays in Amstat News, a membership magazine of the ASA. *Statistics and the Rules of Science* was published in the July 2004 issue of Amstat News. An online copy might be found with careful google searching; at time of this writing, a copy can be found at *http://www-stat.stanford.edu/~ckirby/brad/other/Article2004.pdf* [cited 26 February 2009].

The heliocentric model of our solar system is a grand example of a scientific model, one that allows us to compute the exact dates and times of transits of Venus, solar and lunar eclipses, and exact times of equinoxes and solstices. It allows us to compute dates of Easter and has enabled us to send spacecraft to orbit and land on other planets. In Efron's words, it exemplifies "the prestige and power of what science connotes in modern society." The Laws of Nature, as discovered by Newton, Galileo, Kepler, and their peers, are clockwork-like deterministic models that have helped pull our society, albeit kicking and screaming, into a modern scientific age.

The models of Newton and Galileo, those used by Proctor and still being used today, describe deterministic behavior; they have little room for statistical inventions like variation and random error. Statistical models, then, allow us to describe past observation and predict future outcome not with the certainty of Proctor, but within the confines of our understanding of probability and randomness. Statistical models become tools for understanding sources of variation.

The history of science over the centuries can be written in terms of improvements in resolution▸. Galileo's invention of the telescope and Leeuwenhoek's microscope improved our ability to see large worlds far away and small worlds close at hand. Each improvement in resolution has removed a layer of what was previously thought to be noise and replaced it with explanation. Statistical models serve as tools to understand sources of variation and allow us to investigate more intricate natural phenomena. Statistical models address the question: What is causing the differences in outcomes that I see?

▸Edward Tufte. 2008. "'The deepest photo ever taken' and the history of scientific discovery" <http://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=0000ZR&topic_id=1>, 16 June 2003 [cited 26 February 2009].

Statistical and probabilistic thinking centers on understanding distributions. Distributions consist of two components: first, a support set, a listing of what outcomes are possible, and second, a mass, or probability, function, which tells how likely the outcomes are. Scientific description and comparison, as we proceed to improve our resolution and advance science, will be done with data that contain variation. This description and comparison will be done through description and comparison of distributions. Although some of these distributions will be easily summarized via classic summary statistics like the mean and variance, more often than not, truly understanding the distribution requires more than just two summary statistics. As such, often the best thing is to try to show all the data; summary measures don't always tell the whole story.

Our quest for understanding distributions begins several years ago in the form of an elementary school math class homework assignment for one of my children. Paraphrasing, it asked, "Last year there were eight students in Mrs. Johnson's piano class. Their ages were 10, 11, 6, 11, 9, 48, 10, and 7. Calculate the mean

age. Calculate the median age." And then the "interpretation" component: "If you could only use one number to represent the ages in Mrs. Johnson's piano class, which one would it be?"

Of course, the fourth-graders are being taught something along the lines of "outliers can adversely skew the sample mean". The mean is conveniently simple to compute as 112/8 = 14. The median is found by putting the data in order (6, 7, 9, 10, 10, 11, 11, 48), noting that there are two 10's in the middle, and writing down "10" as the median.

The interpretation component is typically answered from remembering instructions from the teacher akin to "the median is a better number than the mean to represent data with outliers" so for the last part of the question the students will write "the median".

And all these thoughts are zooming around in my head as my son is showing me the homework problem and I realize that there is no gun to my head forcing me to use just one number to represent those data and I come to the realization that throughout this solution-finding process, the students are never being asked to *think*.

The data themselves form a distribution; it is this distribution that should be at the center of discussion of ages of students in Mrs. Johnson's piano class. While one can certainly always compute the mean and the median, these summary measures only tell part of the story; what we need is the distribution. So the Fundamental Principle of Statistical Data Displays, whether they be figures or summary statistics or whathaveyou, must be that **the exposition of the distribution is paramount.**

Different people might ask different questions of Mrs. Johnson's piano class data. *If I join that class, will there be any students my age? Will I be the youngest one? Might I be the oldest one?* Note that these questions are not necessarily answered by using the mean and the median. [Note that the mean is simply the total scaled by the number of observations. Thus, any question to which the mean is a relevant answer is a question to which the total is a relevant answer. Thus, *mean if, and only if, total*, meaning that if the mean is relevant, then the total must also be relevant, and if the total is relevant, then the mean must be relevant. As such, one could argue that since it seems unlikely that anyone would be interested in the total age of the students in the class, it seem unlikely that anyone would be interested in the mean age.]

If the number of data points being exposed is small, then a simple ordered list might suffice in our effort to expose the distribution, as we have shown above:

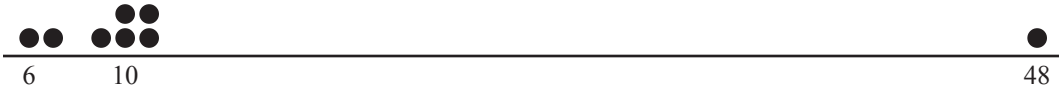
6, 7, 9, 10, 10, 11, 11, 48.

Note that this representation compresses outliers, as the 48 gets pulled in next to 11. Furthermore, the pairs of 10's and 11's actually spread out, so that duplicity in this case is illegitimately rewarded. We could physically space the values at their values, each number marking its own location:

```
          1011
6 7   9  1011                                                  48
```

At this point, we see that the current typography is failing us, as the reduced

spacing needed to accommodate the 48 forces the 10 and 11 to merge. Rather than reduce the size of the font, we may move to showing things pictorially; the distribution can be shown with dots replacing the numeral in what becomes, in essence, a histogram:



The individual dots sum up to produce what could be perceived as bars in a typical histogram but due to their uniqueness, no vertical scale is even necessary as a quick count points to areas where there are multiple individuals. So we build up the graphic from component elements; each atomic-level datum has its place in the graphic, building up the whole but maintaining its autonomy. Granted, some information is lost, in particular, without the original data list, we have to make the assumption the data points are supported on the integers. But what we get in return for giving up that precision is *a picture of the data set as a whole*; our purpose is not necessarily a visual element that carries all the value of a regurgitation of the data set, it is *a visual element that allows us to see each datum in relation to the others*, for this is how we make descriptions and comparisons, this is how we do science. Our visual element allows us to see the distribution, to see what is possible and how likely it is.

There is no real reason for this histogram to be all by itself and taking up all this space. This histogram shows us the distribution; it is a noun and can take its place as an element of a sentence. So we could say, "Mrs. Johnson's piano class last year had eight students of varying ages , seven of them young children and one older gentleman, Mr. Onaip, who provided the other students with an interesting perspective on music and life in general."

Tufte calls these small, in-line data displays *sparklines*, intense, simple, word-sized graphics▸. Their use appears to be ever increasing; however, most seem to take the form of time series plots.

▸Edward Tufte, *Beautiful Evidence* (Graphics Press, 2006), 47.

Like any good infomercial, "But wait! There's more!"; the fourth-grade math question continued: "This year there are eleven students (all eight have returned and there are three new students). The ages are 7, 8, 10, 11, 11, 12, 12, 49, 27, 47, and 49. Calculate the mean …". Three points were added to the distribution and it shifted to the right. Again, we expose the distribution, keeping things on the same scale, so as to facilitate comparisons between the two years: "This year, the age distribution is noticeably different, since Mr. Onaip has brought along two of his like-aged friends and his daughter, Allegro, who has just graduated from medical school."

We can now compare last year to this year and see a number of interesting features of the two distributions. We can see the younger children (and Mr. Onaip) all get a year older. We can see Mr. Onaip's two like-aged friends. We can see Allegro Onaip filling a place in the center of the distribution. Just for sake of comparison, we can compute the new mean, at slightly more than 22, and the new median, 12. Again, neither of our typical summary measures tells the story of the distribution, (either separately or collectively).

Another example: The following table shows the number of points little Johnny scored in each of his team's first 5 games:

Game	Points
1	6
2	0
3	6
4	23
5	25

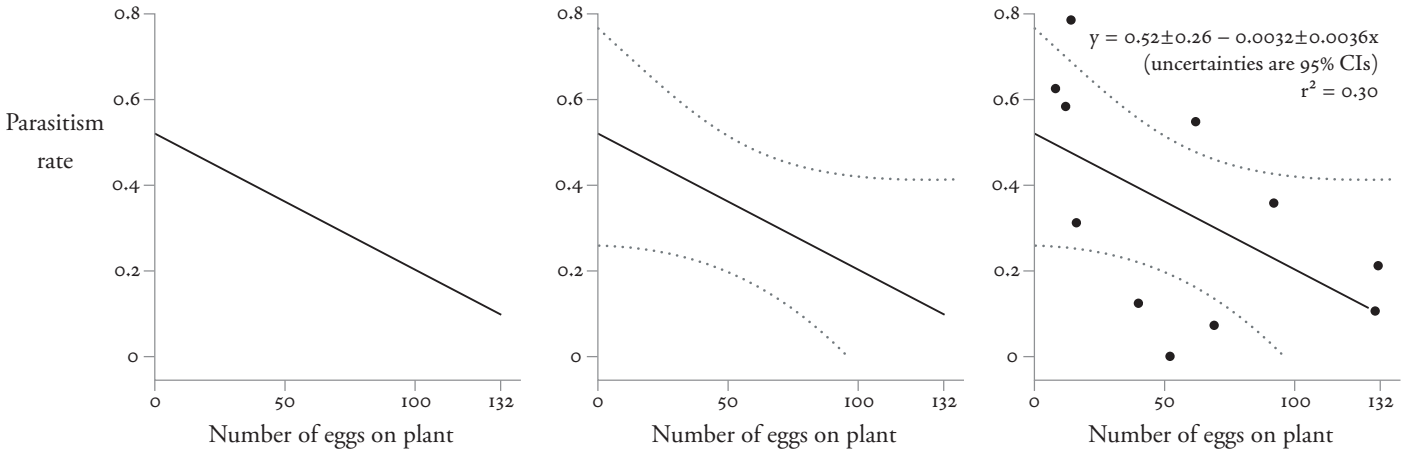
Note that mean points per game from little Johnny is 12, the median is 6, and the mode is also 6. But do any of these summaries tell the story of the data?

While the mean may be a relevant summary measure since we might be interested in little Johnny's season total as well, the really important question revolves around the stunning transformation that took place between the third and fourth games! What happened? Did little Johnny get new glasses? Did he get casts removed from both arms? Did the top scorer on the team break a leg? Did the team change levels of competition? Almost all the pertinent questions surrounding little Johnny and his basketball team are not answered by the mean, median, or mode of this data set; why are we teaching students that computations of summaries is how one does analysis?

We detect variation in the data set, not all of the data points are the same. We seek to understand and uncover the source or sources of that variation. Of note here is the fact that the source of variation is not exposed in this collection. All we have is a surrogate (the game number) for that source of variation. Statistical investigations involve finding those true sources of variation. Oftentimes, computing only summary statistics only serves to hide that telltale variation.

Distributions (and the data sets that create them) are difficult concepts. People often avoid working with distributions; to get around them, they like to summarize these distributions, instead of carrying around all the data points. All the common summary statistics, like means and medians, variances and standard deviations, minima and maxima, percentiles, even regression lines, chi-square values, and P values, are summary statistics of distributions of data.

A result then is that these summaries are shown graphically. We often say that we show them graphically so as to improve understanding. As such, we should make sure that the graphics that we produce actually improve our understanding. Consider the following plots.



Graphic courtesy of Anthony Darrouzet-Nardi, University of Colorado at Boulder.

The left plot shows a grand summary of the relationship between the number of eggs on a plant and the parasitism rate. This likely is an output from some sort of linear regression. In general it shows that as number of eggs increase, the parasitism rate decreases.

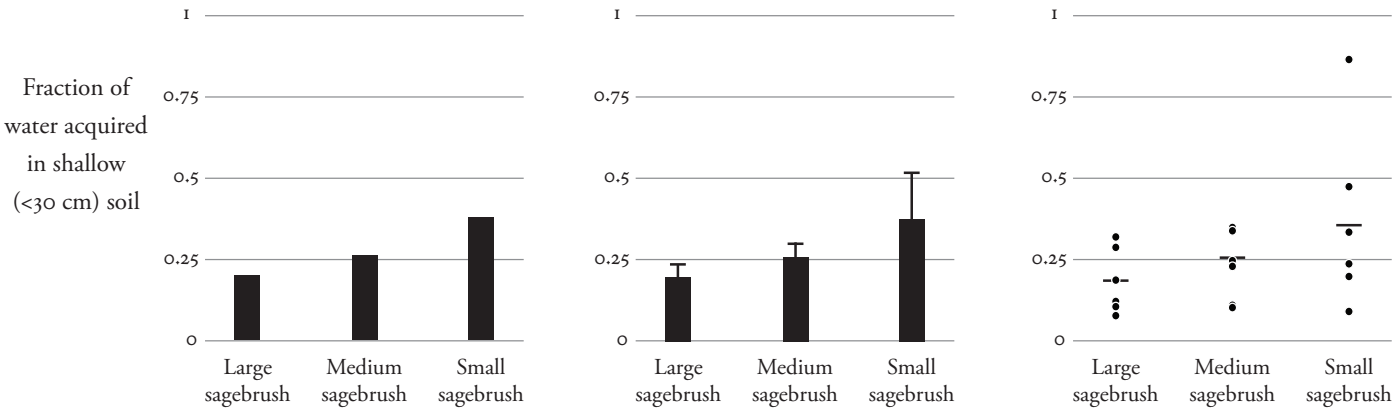
The middle plot adds some sort of error curves or prediction intervals that help us understand that the relationship is not exact. We are not sure if these relate to the uncertainty in the mean of the Y values for a given X or to the actual distribution of the Y values for a given X, but at least the author lets us know that he is aware of uncertainty in some sense.The third plot adds the actual data points to the plot, along with some explanatory notes telling us something about the equation of the regression line and the computed r-squared value.

So, we see in the last panel the data *and* the summaries that are being used to describe the model for the relationship between the number of eggs on plant and the parasitism rate.

The important question, however, is this: "Do you want these summaries to describe these data?" Moving from the first to the third panel, we might feel some sort of uneasiness. When we see the data and the summaries together, we might feel like we distrust the summaries. Regardless of your like or dislike of the summaries used below, what the designer of the figure has given you are the data themselves. So, if you agree or disagree with the presenter with regard to the interpretation, the data remain unchanged. **Show the atoms; show the data**.

Not all summaries are bad. Means and variances, for example, are well-used summaries since they are jointly minimally sufficient statistics for all the parameters in a normal (Gaussian, bellshaped) curve. That is, the mean and the variance in that case tell us all that the data can possibly tell us about a normal distribution. In the bivariate normal regression setting, as in the parasitism rate and number of eggs example, the regression line parameters, the estimated intercept and slope and the variance structure, act similarly. But all these conclusions of sufficiency hinge on the assumption that the underlying distribution is normal and that we have a random sample. What if we don't have that? Then what can we do? Is there a way to come up with summaries that can carry the information that we seek to carry?

The next plot shows another example of increasing resolution from summaries to data points. The example uses categorical classifications on the horizontal axis instead of a continuous predictor.



The three-part figure contains three steps from simple means of some outcome (left) to the entire distribution of responses (right). The figures here show the fraction of water acquired in shallow soil for three different sizes of sagebrush. As sagebrush grow, their root systems get water at different depths.

Graphic courtesy of Anthony Darrouzet-Nardi, University of Colorado at Boulder.

The first panel shows the means for each of the three sizes of plants in a traditional bar plot. Here we see that as the sagebrush gets smaller, the mean gets larger. Older, and hence bigger, plants get more water from deeper in the soil. [We might note that the ordering of the categories might be more intuitive if 'Small' were on the left and 'Large' were on the right. This figure, however, is actually a subset of a larger figure with more types of plants. In that context the ordering works well.]

The center panel shows the means with the addition of some measure of uncertainty, the dreaded 'dynamite plot', so named because of its resemblance to devices used to set off dynamite blasts. "But this shows the means and some measure of uncertainty; isn't this a good thing?"

At issue here is the visual perception of this plot; are we showing the distribution? Where are the data? If we are clever, we notice that the small sagebrush category on the right has a longer variability measure, so that group is more variable. But what can we conclude about the actual distribution of the data? Can we use the plot to make inference about the distribution?

If the variability stems are standard deviations, then we might use a statistical rule of thumb and guess that about 95% of the data fall within two such standard deviations of the mean. So, we need to visually double the length of the variability stem, add and subtract it from the top of the black bar, and then guess where the data might be.

If the variability stems are standard errors, then we only have a statement about where the mean is. If we recall that the standard error is estimated from the standard deviation divided by the square root of the sample size, then we can use the standard errors to estimate the standard deviation (if we know the sample size!) and then proceed as mentioned previously. Goodness; that's a lot of work to try to decipher a plot that was constructed to help us understand

the data.

Here's a simple rule of thumb: **each datum gets one glob of ink** and add extra ink only to increase understanding. If we apply such a rule, we see something like the third panel. Here we see each individual datum, the measurement from each plant. The small lines are the means that were shown with the bars in the other two panels. The bar plots, both with and without the error bars, don't show us where the data are and they present an impossible data mapping scheme: at the bottom of each bar in the bar plot (beneath the lowest datum), the ink there means that there are no data. Farther up, between the lowest datum and the mean, ink in this region implies that there are data there. Above the mean, but below the highest datum, the *lack* of ink is used to show data. And above the highest datum, the *lack* of ink is used to demonstrate the *lack* of data! That's just silly; no wonder kids think that graphs are hard to read.

A bad data-to-ink mapping:

Location	Data status	Ink status
Below lowest datum	Absent	Present
Between lowest datum and mean	Present	Present
Between mean and highest datum	Present	Absent
Above highest datum	Absent	Absent

A good data-to-ink mapping:

Location	Data status	Ink status
Everywhere	Absent	Absent
Everywhere	Present	Present

The individual dots on the plot help us to understand the distribution; the bar plots hide the distribution from us or even lie to us as to where the data are.

Note that standard error calculations tell us about our certainty in finding the mean, which may or may not even be important to the question at hand. Standard deviations tell us about uncertainty in the data themselves. If making inference about means, use standard errors. If making inference about individuals use standard deviations.

Analysis, from the Greek, implies breaking things down into component parts so as to understand the whole. Its opposite is *synthesis*, bringing together the parts to construct the whole. If we are going to use data displays to help us do data analysis, then we must make attempts to break the data down to their component parts, their atoms. Computing summary measures like means and medians and percentiles and standard deviations and even F and chi-square and t statistics and P values, is not analysis; it is synthesis! And, far worse than playing games with word meaning, data synthesis often obscures understanding the data.

Why, then, do we ever compute summary measures? The theory of statistics has a concept called sufficient statistics. The general idea is that if you know the data come from a distribution with a known form, then there are certain summaries of the data that tell you all you can possibly know about the distribution. In many of the nice theoretical distributions, the sum (and thus, by way of simply scaling by the sample size, the mean) of the data values is a sufficient statistic. And medians are close to means when you have well-behaved data, so people use medians too.

Often (more often than not?), however, the data are not well-behaved or they don't come from pretty distributions. Then what? Are there still sufficient statistics?

The answer is yes; there is always a set of sufficient statistics. That set of sufficient statistics, the order statistics, is essentially the data themselves.

That is why, when doing data analysis, we first plot the raw data. We show the atoms. We search for the fun stuff, like outliers. [The excitement is always found in the tails, or outliers, of the data.] We seek to understand their source. Remember that the goal is understanding of the distribution of the data; therefore, make every rational attempt to show all the data.

A little aside on means and summary statistics; show this page to your friends and colleagues who like dynamite plots, those who insist that they need to have them in their manuscripts and slideshows.

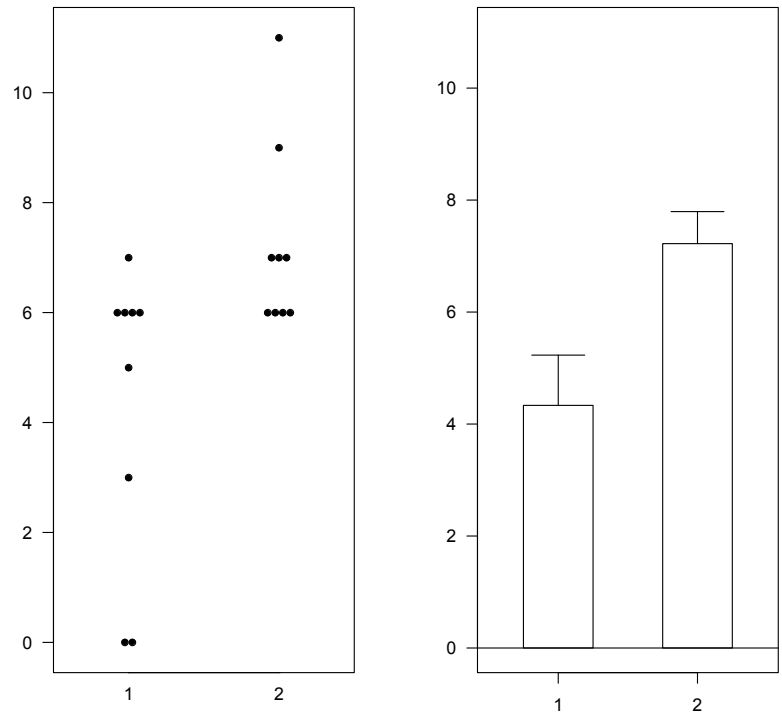
The very simple, quick and easy plots on these pages are courtesy of Tatsuki Koyama, Vanderbilt University Medical Center Department of Biostatistics.

On the left side of each plot are some raw data. Each plot is based on the same set of data. The data happen to be counts, so the support set is the integers.

The data in group 1 range from 0 to 7; those in group 2 range from 6 to 11. There are 9 data in each group. The mode for both groups is 6, perhaps alerting us to the fact that 6 may have some preference in selection. The atomic-level presentation on the left allows us to see the entire distribution.

The right side of each plot shows the mean and some measure of variation, based on the data on the left side of the plot. On this page, the measure of variation is the standard error; standard deviations are shown on the next page. We see that the means for the two groups are approximately 4 and 7. The standard error bars on this page tell us how certain we are in the location of the mean, if we know how to read them. The standard deviation bars on the next page tell us something about where the data are, if we know how to read them.

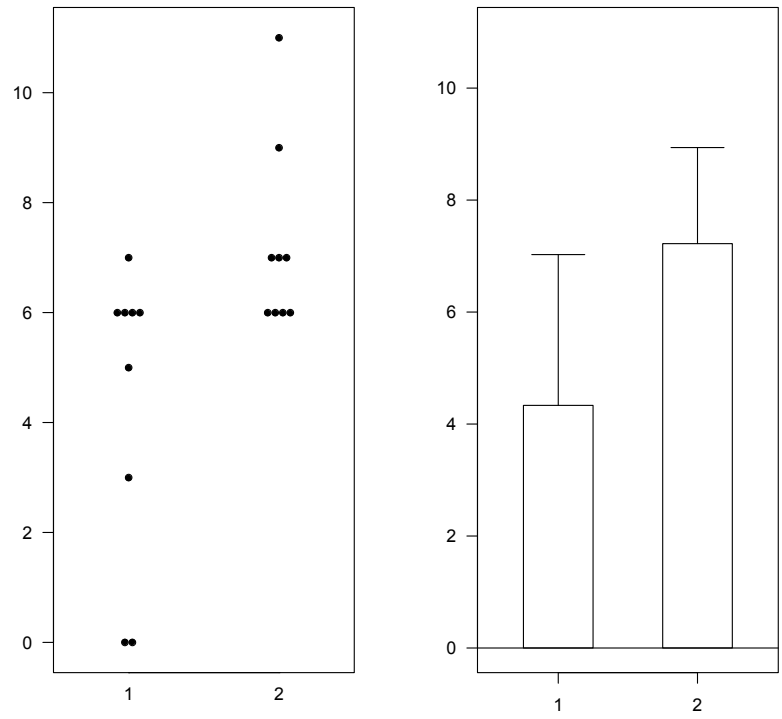
Here's the challenge for your friends. Ask them if they like the standard error bars or standard deviation bars. Based on their answers, show them the corre-



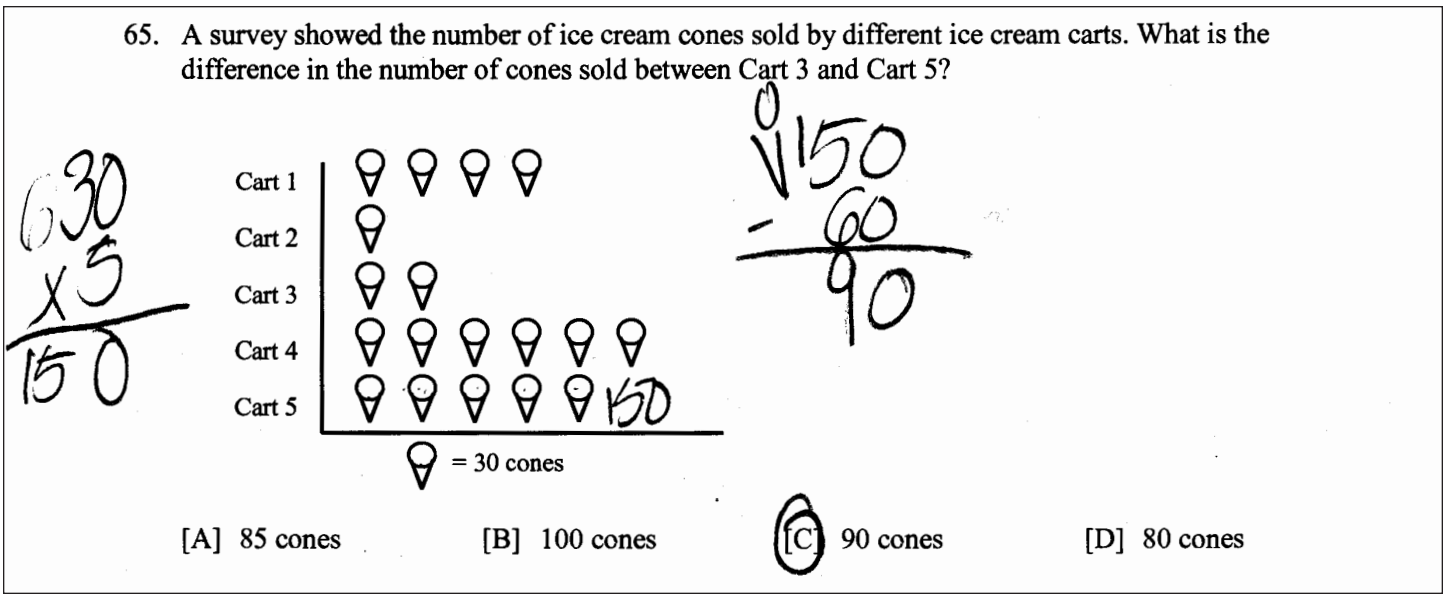
sponding plot but cover up the raw data plot. Then ask them to guess where the data are. If they are clever they might ask how many data point there are. Tell them. They might ask about the support set. Feel free to tell them that the support set is the non-negative integers. Ask lots of your friends to sketch the data. Look at the distribution of responses across the collection of your friends. Then have your friends look at all the distribution of responses across the collection of your friends. Have them all work together. Heck, buy them a pizza and drink as they work. Ask them how can it be that there are so many different interpretations from these plots, after all, they are showing the mean and some measure of variability, shouldn't that be enough? Work, work, work. Ask, ask, ask. Ask them if they think the distributions are the same or different and what does that mean and why do they think what they think.

When they are all done arguing over who is correct and who isn't, show them the data in the left plot. Tell them that each dot is one datum. Then ask them the same questions as before: ask them to now guess where the data are. Ask them whether they think the distributions are the same or different and what does that mean and why do they think what they think.

You see, the dynamite plots just obscure the data and hence obscure understanding of the data. Let's not provide plots that make it difficult to understand the data and what they are trying to say.



A scan of some more homework, complete with elementary student annotations:



This graphic is intended to help elementary school students learn about graphics and math. What it does is teach them that graphics are just brain-teaser puzzles: I encoded the data into this chart; your job is to get it out.

As you can see, the student here did exactly as she has been instructed to do: *Let's see here; cart 3 has sixt-..., oops, I mean, cart 5 has five little cone things so I take thirty times five; zero times five is zero; three times five gives me fifteen, so that gives me one-fifty for cart 5, I'll put that by cart 5; ok, now I subtract the sixty; so zero minus zero is zero; tens column: five minus six; ok, gotta borrow; cross off the one, make it zero, make the five a fifteen; fifteen minus six is nine; so I get ninety. Ok, it is [c]. Done.*

Of course, she could have noticed that cart 5 had three more cone thingys than cart 3 and then just multiplied three times thirty. Yikes.

The real question one ought to ask of these data is why one would go through all this trouble to write down five data? And why are all these carts selling multiples of 30 cones? Are they only selling by the case? And why is cart 4 selling six times as much ice cream as cart 2? What is the source of the variation among the carts? Why are we teaching children that graphics are an impediment to understanding instead of an aid? None of these questions is answered by this graphic puzzler.

Good graphics should reveal data. Tufte lists nine "shoulds" for graphical displays; they are fundamental and deserve mention here[1]. Graphical displays should

- show the data
- induce the viewer to think about the substance rather than about the methodology, graphic design, the technology production, or something else
- avoid distorting what the data have to say
- present many numbers in a small space
- make large data set coherent
- encourage the eye to compare different pieces of data
- reveal the data at several levels of detail, from a broad overview to fine structure
- serve a reasonably clear purpose: description, exploration, tabulation, or decoration
- be closely integrated with the statistical and verbal descriptions of a data set.

The ice cream cone data display probably fails at all of these, perhaps with the exception of distorting what the data have to say.

Again, however, "but wait, there's more"; the next question in the assignment also fails miserably. A question about ticket sales:

[1] Edward Tufte, *The Visual Display of Quantitative Information* (Graphics Press, 1983), 13. This book is a must-read and re-read for anyone trying to understand what to do with graphical displays of information. The book is one of four books on graphics and information design the Tufte has authored. He also offers a one-day course, during which the books are provided. See his website for details (www.edwardtufte.com).

66. The table shows the number of tickets sold for several school events. Which pictograph matches the table?

Ticket Sales

Fall Play	60 tickets
Winter Play	170 tickets
Spring Play	120 tickets
Music Program	80 tickets
Sock Hop	50 tickets

[A]

Ticket Sales

Fall Play	[][][][][][][][][][
Winter Play	[][][
Spring Play	[][][][][][]
Music Program	[][][]
Sock Hop	[][][][]
	[] = 20 tickets sold

[B]

Ticket Sales

Fall Play	[][][]
Winter Play	[][][][][][]
Spring Play	[][][][][][][][][
Music Program	[][][
Sock Hop	[][][][]
	[] = 20 tickets sold

[C]

Ticket Sales

Fall Play	[][][]
Winter Play	[][][][][][][][][]
Spring Play	[][][][][][]
Music Program	[][][][]
Sock Hop	[][][
	[] = 20 tickets sold

[D] None of these.

Goodness. Our student has diligently done the computations, converting the ticket stubs back into numbers, numbers that were adequately and precisely presented in the data table at the start of the problem. The secret encoding trick this time deals with the magical half-ticket, cleverly drawn to indicate 10 tickets.

Important reasoning about sources of variation relating to ticket sales, including why the school only sells tickets in multiples of 10 tickets, is completely obfuscated by the drudgery involved in turning perfectly good numbers into ticket stub images. Why are the tasks surrounding thinking about the distribution being ignored?

We do see in these little pictographs, however, inclinations of distributions. We are seeing evidence about what is possible and how often those things occurred. What we are not seeing, though, is any attempt to reason from the data, to make comparison or draw conclusions. The scientific component is being stifled. While we are seeing a distribution akin to the piano data, the presentation is treating the data as playthings and graphical presentation as a toy.

"*The graphical method has considerable superiority for the exposition of statistical facts over the tabular. A heavy bank of figures is grievously wearisome to the eye, and the popular mind is as incapable of drawing any useful lessons from it as of extracting sunbeams from cucumbers.*" — Arthur B. and Henry Farquhar▸

▸Arthur B. and Henry Farquhar, *Economic and Industrial Delusions, A Discussion of the Case for Protection* (G.P. Putnam's Sons, 1891), 55. I found this quote from the Farquhars in Howard Wainer's *Graphic Discovery, A Trout in the Milk and Other Visual Adventures*.

To paraphrase, I know there are useful lessons in this table, I just don't know how to get them out! One more display disaster and then suggestions for improvement:

Here we see a table of information on the planets (back when there were still nine planets). The table could use a bit of work, as inconsistencies and detours to information retrieval abound. Let's start at the top.

Planets	Diameters	Distance from the Sun	Length of One Year
Mercury	4,878 km	58 million km	88 days
Venus	12,104 km	108 million km	225 days
Earth	12,756 km	150 million km	365¼ days
Mars	6,794 km	228 million km	687 days
Jupiter	143,000 km	778 million km	12 years
Saturn	120,536 km	1,429 million km	30 years
Uranus	51,118 km	2,871 million km	84 years
Neptune	49,528 km	4,504 million km	165 years
Pluto	2,300 km	5,914 million km	249 years

20. Use Numbers Use the chart above. How much greater is Earth's diameter than Mercury's?

A 12,756 km
B 4,878 km
C 7,878 km
D 8,978 km

21. Interpret Data Use the chart above. How does distance from the Sun affect the length of a planet's year?

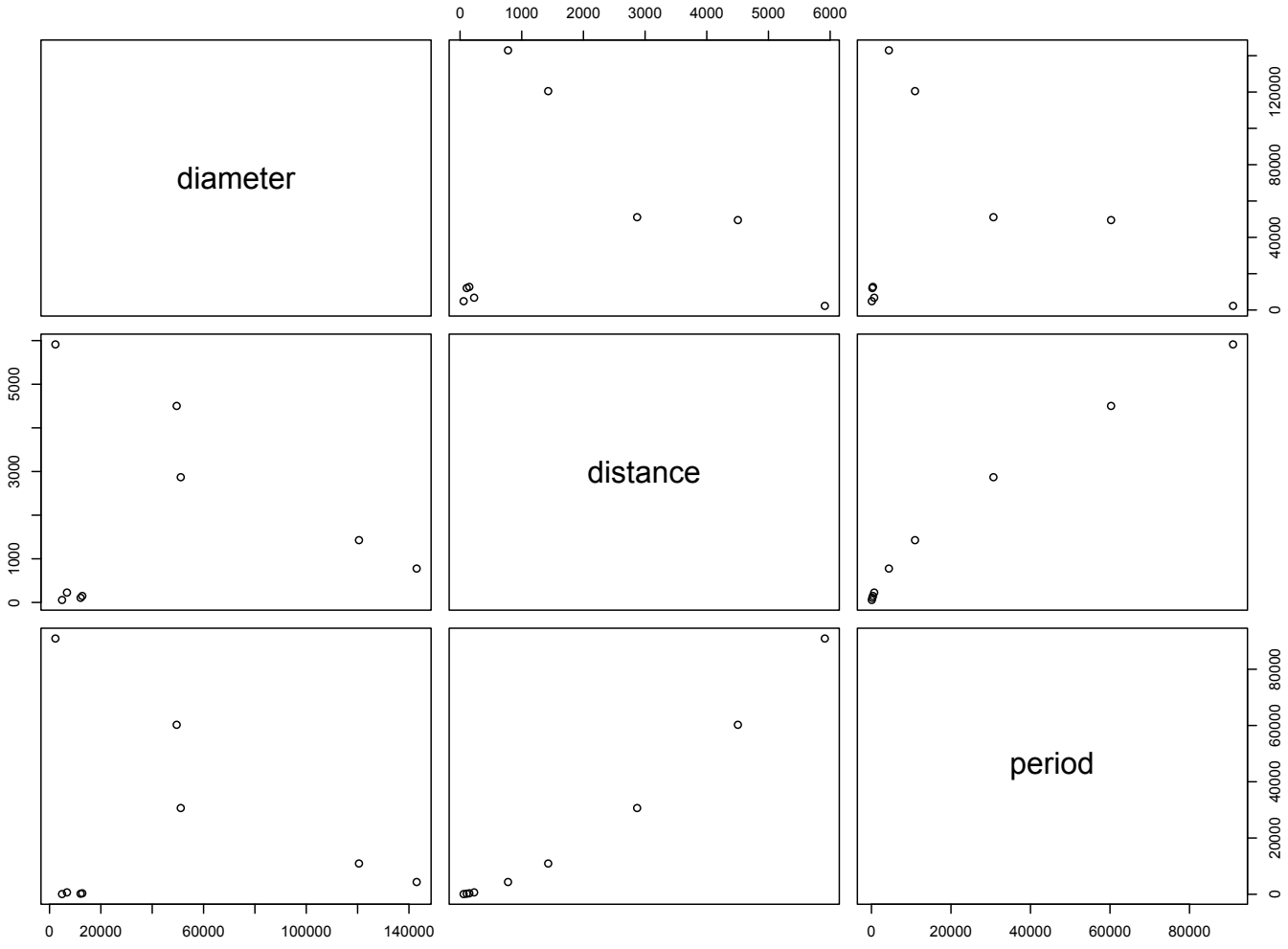
F the farther from the Sun, the shorter the year
G the closer to the Sun, the longer the year
H the farther from the Sun, the longer the year
J distance from the Sun has nothing to do with the length of a year

Note the subtle inconsistency in the column headers: planets and diameters are plural, distance and length are singular. The 'km' unit description for 'Diameters' could easily be placed in the header, as could the 'million km' descriptor in the Distance column. Lengths of One Year are presented at the top in days and toward the bottom in years. Yes, students should pay attention to units but improving understanding comes from maintaining consistency across entries. Note also that an Earth year is measured with precision of six hours but Jupiter through Pluto are measured to the nearest year, a drop in precision of 1400 times, over three orders of magnitude. No wonder the kids don't understand significant digits. Planet diameters share a similar fate, as Jupiter's is rounded to the nearest 1000 km (a kilo-kilometer, a megameter?) while the other planets enjoy precision to the nearest kilometer, excepting poor Pluto, who gets rounded to the nearest 100 km.

Evidence of our student's computations are present, showing that she is able to find the difference in planetary diameters, for whatever reason there might be to do such a thing. Oh, yes, there is a reason; it is right there at the header for Question 20: **Use Numbers**. This is not science, it is a subtraction problem encased

in a riddle. And the Interpret Data question is the attempt to extract sunbeams from a cucumber. Assuming the student doesn't notice that choices F and G are logically equivalent and thus must both be considered incorrect, and assuming further that choice J follows choice H, directly bypassing I, this is precisely the time to learn to reason from graphics.

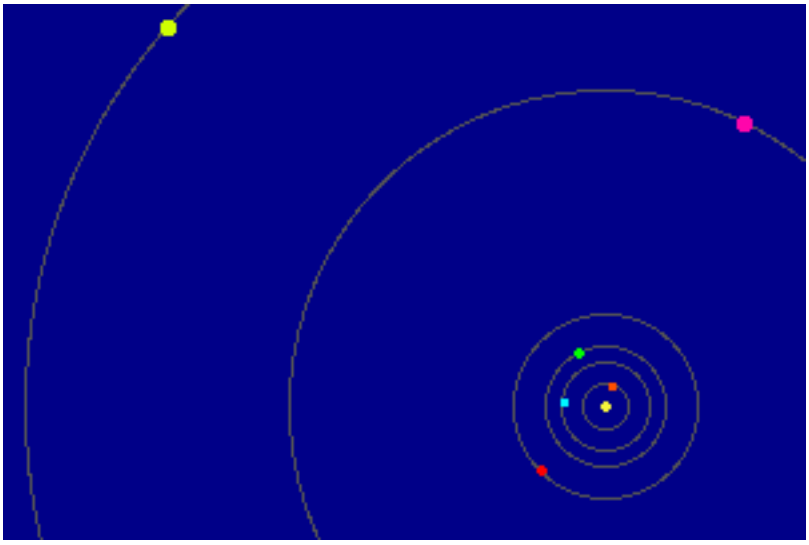
These nine rows represent nine multivariate data. A collection of even hastily-drawn scatter plots (using only default graphics settings in R) reveals the relationship that bears Kepler's name:



Here we see the multivariate data in projections onto the three bivariate planes. We can examine the relationship between distance and diameter and detect the four small inner planets and then see that Pluto is small too. The gas giants Jupiter and Saturn are truly giant compared to the small ones, while Neptune and Uranus fill out a niche in the middle. Planetary period ("Length of One Year") as a function of distance from the sun is seen in the middle plot in the bottom row. The smooth relationship is driven by the fact that the period increases with the 3/2 power of the semimajor axis, or, approximately, the average distance to the sun. The plot also shows the inverse relation (right, middle), that the distance is related to the 2/3 power of the period.

Yet, we can do even better by showing a bird's-eye view of the solar system with

some crude animation. The web site http://www.edstephan.org/Astronomy/planets.html presents just such an animation, a still version of which is here:



The applet is a joy to watch. The inner planets whirl around the sun while the outer planets lumber along. The students could actually use computers to *learn something* aside from updating Facebook and making PowerPoint slides. A site from which this applet is referenced, http://csep10.phys.utk.edu/astr161/lect/history/kepler.html, has a straightforward presentation of Kepler's laws.

The remade data table could look like this:

Planet	Diameter (thousand km)	Distance from the Sun (million km)	Period (Earth years)
Mercury	4.9	58	0.24
Venus	12	110	0.62
Earth	13	150	1.0
Mars	6.8	230	1.9
Jupiter	140	780	12
Saturn	120	1400	30
Uranus	51	2900	84
Neptune	50	4500	170
Pluto	2.3	5900	250

Now we can see Kepler's law in action. Look at distances for Mercury and Pluto. Pluto is at 5900, Mercury is at 58, a ratio of approximately 100:1. Now look at their periods. Pluto is at 250, Mercury is as 0.24, a ratio of approximately 1000:1. What is the 2/3 power of 1000? It is 100! We can see this in the table.

The lesson from remaking the table and clearing out all the non-data items is **stay out of the way of the data**. This principle shows up within Tufte's five principles‣ (above all else show the data; maximize the data-ink ratio; erase non-data ink; erase redundant data-ink; revise and edit), as the antithesis of some of Wainer's How To Display Data Poorly rules‣ (show as little data as possible;

‣Tufte presents these "five principles in the theory of data graphics" at the conclusion of Chapter 4 of *The Visual Display of Quantitative Information*, page 105.

‣Wainer's first chapter in *Visual Revelations* is entitled "How to Display Data Poorly". This chapter presents a dozen rules for *poor* data display. From these he develops three rules for *good* display: 1. Examine the data carefully enough to know what they have to say, and then let them say it

hide what data you do show; emphasize the trivial, label (a) illegibly, (b) incompletely, (c) incorrectly, and (d) ambiguously; more is murkier: (a) more decimal places and (b) more dimensions), and in Cleveland's Clear Vision▸ (make the data stand out; avoid superfluity; use visually prominent graphical elements to show the data). In redrawing the planet table, we took out the clutter and kept the focus on the data. The ice cream cones and ticket stubs get in the way of the data. Our plots of the piano class focus on the distribution of the data, keeping mind of the data atoms themselves.

with a minimum of adornment; 2. In depicting scale, follow practices of "reasonable regularity"; and 3. Label clearly and fully.

▸The second chapter of Cleveland's The *Elements of Graphing Data* is "Principles of Graph Construction". Section 2.2 of this chapter is Clear Vision and points to the importance of the viewer being able to "disentangle the many different items that appear on a graph". Cleveland devotes thirty pages to this topic.

Here is another table, in the format in which it came from the researchers:

Incidence of ROP by Weight Class and Year

	No ROP		< Prethreshold		Prethreshold		Threshold		Total # infants	
	1995-1996	1986-1987	1995-1996	1986-1987	1995-1996	1986-1987	1995-1996	1986-1987	1995-1996	1986-1987
< 750 gm	7 (29.2%)	1 (3.7%)	7 (29.2%)	11 (40.7%)	9 (37.5%)	8 (29.6%)	1 (4.2%)	7 (25.9%)	24	27
750-999 gm	20 (55.5%)	3 (9.4%)	9 (25.0%)	9 (28.1%)	4 (11.1%)	12 (37.5%)	3 (8.3%)	8 (25.0%)	36	32
1000-1250 gm	27 (84.4%)	14 (46.7%)	4 (12.5%)	12 (20.0%)	1 (3.1%)	2 (6.7%)	0 (0%)	2 (6.7%)	32	30
Total (all wt)	54 (58.7%)	18 (20.2%)	20 (21.7%)	32 (36.0%)	14 (15.2%)	22 (24.7%)	4 (4.3%)	17 (19.1%)	92	89

ROP is retinopathy of prematurity, a disease of the eye that affects prematurely born babies. In the table above, infants can be classified as having no ROP, less-than-prethreshold ROP, prethreshold ROP, or threshold ROP. These are ordered categories of disease, making the data ordered categorical data, sometimes called ordinal data.

Birth weight is an index of amount of prematurity; children born early typically weigh less than those at full term.

I received the above table from two researchers who were interested in comparing ROP "then" (10 years ago, in 1986-1987) and "now" (1995-1996, which was the present time as of the collection of these data). They wanted to know if ROP was more or less common now compared to then, since they had been influential in changing how and when children were screened for ROP. *Had their methods made any improvement?*

Attempting to answer this question from the table they provided is difficult, if not impossible. The overbearing gridlines aside, the table prohibits the investigation of the fundamental comparison of interest.

The fundamental comparison here is a comparison of distributions, in particular, the distribution of ROP ratings for children *then* and the distribution of ROP ratings for children *now*. As such, we need to expose those distributions of ROP ratings; we can do better than the draft table I received.

If incidence of ROP is the outcome, there are essentially three comparisons that could be made. The first is between levels of ROP, the comparison between the rates in the ordinal categories. The second is between the three birth weight classes. The third is between the time frames, then and now.

Note that this first comparison, between levels of ROP, compares the levels within the distribution; we want to compare the distributions as a whole. Since question of interest is between time frames, let's start with just comparing then with now, and ignore, for the time being, the fact that we know the birth weights. The bottom row of the table gives us the numbers we need, but the presentation transposes the ROP and time frame. Just using the bottom row of numbers and rearranging the numbers, we get

Birth weight class	Time frame	No ROP	Less than prethreshold	Prethreshold	Threshold	Total number of infants
Total (all weight classes)	1986-1987	18 (20.2%)	32 (36.0%)	22 (24.7%)	17 (19.1%)	89
	1995-1996	54 (58.7%)	20 (21.7%)	14 (15.2%)	4 (4.3%)	92

Now we can compare the distributions (the percentages that fall in each of the categories) between the two time frames but we are still letting all manner of extra typographic paraphernalia interfere with seeing those numbers.

We want the percentages so as to understand the distribution; we can reduce or eliminate the heavy and daunting partition cage; we can scrap the parentheses and percent signs; we can use more-meaningful and understandable integer percentages; we can drop the 19's and use 2-digit years, perhaps to the chagrin of our friends in IT; but we will need to add some explanatory notes at least in the form of a more descriptive table title. Notice we have already re-ordered the time frames so that 1986-1987 comes before 1995-1996 and we have removed some symbols ('<' and '#') from the column headers.

Percentage of infants with each level of retinopathy of prematurity (ROP) in 1986-87 and 1995-96

Birth weight class	Time frame	No ROP	Less than prethreshold	Prethreshold	Threshold	Total number of infants
Total (all weight classes)	86-87	20	36	25	19	89
	95-96	59	22	15	4	92

Now we see the *distributions* laid out horizontally. We see that the percentage of no ROP has increased from 20 percent in 1986-1987 to 59 percent in 1995-1996. Over time there has been a shift to the left in the distribution of ROP severity.. More children now are having no ROP and also now there are fewer children at each of the graduated levels of the disease than there were in 1986-1987.

But perhaps the change is due to birth weight class changes; do these differences still show up when adjusting for weight class? The statistician will say that this now looks like a job for a Cochran-Mantel-Haenszel test, a way to compare the distributions of ordered categorical data while controlling for a possibly-explanatory stratification variable. Using our table set-up, it is straightforward to include the individual birth weight classes that make up the total. We will add the strata and some additional notes to the title and footnotes.

Fewer babies are developing ROP now compared to a decade ago: comparing the percentage of infants with each level of retinopathy of prematurity (ROP) in 1986-87 and 1995-96 by birth weight class

Birth weight class	Time frame	No ROP	Less than prethreshold	Prethreshold	Threshold	Total number of infants
Less than 750 gm	86-87	4	41	30	26	27
	95-96	29	29	38	4	24
750-999 gm	86-87	9	28	38	25	32
	95-96	56	25	11	8	36
1000-1250 gm	86-87	47	40	7	7	30
	95-96	84	13	3	0	32
Total (all weight classes)	86-87	20	36	25	19	89
	95-96	59	22	15	4	92

Data from Allegro Onaip, M.D. and Sophie Bistro, M.D., Very Scientific Eye Center, Highly Sophisticated University Medical Center; analysis and table by Rafe Donahue, Ph.D., Trustworthy and Unbiased University. Substantially fewer low-birth-weight babies are developing ROP in the mid-90s compared to the mid-80s; Cochran-Mantel-Haenszel test for differences between time frames with respect to ROP category controlling for birth weight class is significant at p=0.001, chi-square value=29.2 with 1 degree of freedom.

What we have now is a visual display of the CMH test. The strata have become, quite literally, strata, or layers, in the table. We see how the components make up the whole and can now make comparisons on multiple levels. We see in the higher birth weights that nearly all the children (over 80 percent) are in the no ROP group, compared with fewer than half a decade ago. Within each stratum, the distribution shows a shift toward better outcome with respect to ROP when comparing now to then. While the numbers are small in each individual stratum, combined they form a solid case in support of the work done by Drs. Onaip and Bistro.

The lessons learned from the ROP data table and from the pictographs and planet data can be found in Tufte's five principles: **erase non-data ink; eliminate redundant ink**. These are part of the general lesson of staying out of the way of the data. Expose the distribution.

A bibliography, only 1/5 of the way through the text!

Cleveland, William S. (1993) *Visualizing Data*. Summit, NJ: Hobart Press.

Cleveland, William S. (1994) *The Elements of Graphing Data*. Summit, NJ: Hobart Press.

Farquhar, Arthur B. and Henry. (1891) *Economic and Industrial Delusions: A Discussion of the Case for Protection*. New York: G. P. Putnam's Sons.

Knuth, Donald E. (1984) *The TeXbook*. Reading, MA: Addison-Wesley Publishing.

Playfair, William. (2005, 1801) *Playfair's Commercial and Political Atlas and Statistical Breviary*. Edited by Howard Wainer and Ian Spence. New York: Cambridge University Press.

Tufte, Edward R. (1983) *The Visual Display of Quantitative Information*. Chesire, CT: Graphics Press.

Tufte, Edward R. (1990) *Envisioning Information*. Chesire, CT: Graphics Press.

Tufte, Edward R. (1997) *Visual Explanations*. Chesire, CT: Graphics Press.

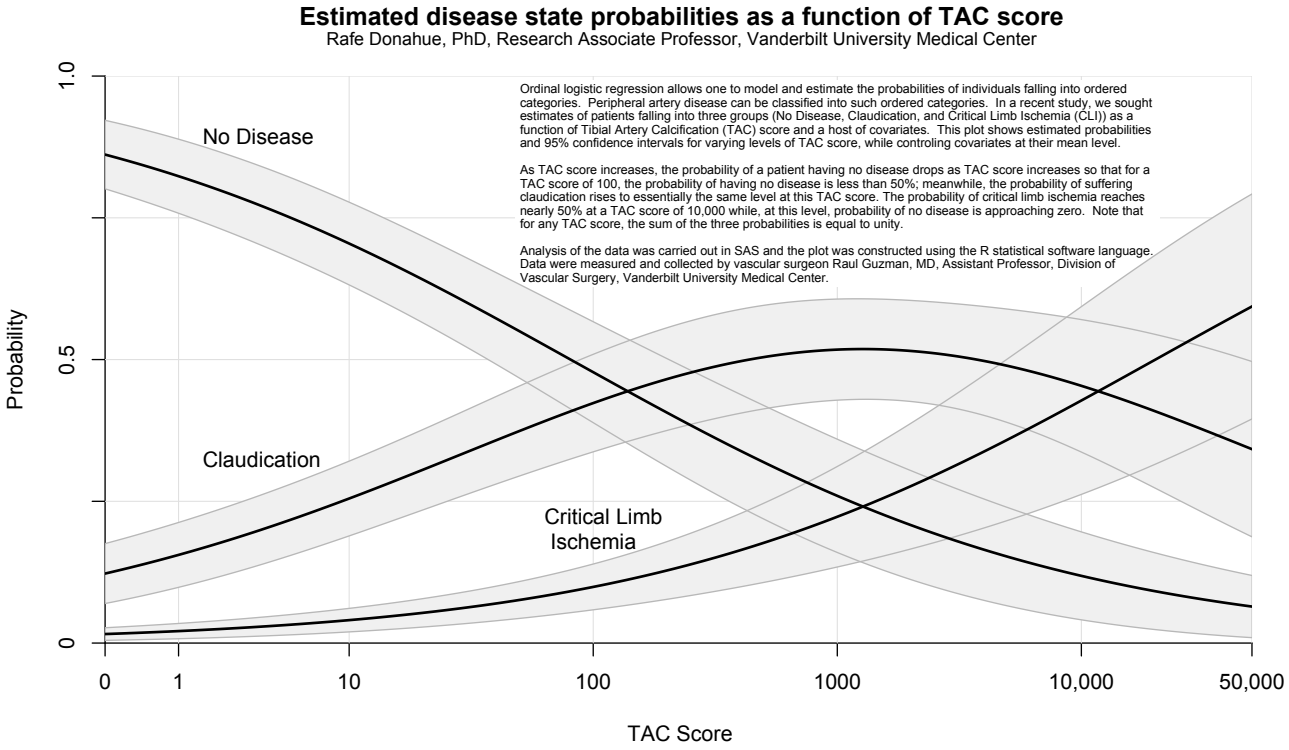
Tufte, Edward R. (2006) *Beautiful Evidence*. Chesire, CT: Graphics Press.

Wainer, Howard. (1997) *Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaprte to Ross Perot*. New York: Copernicus.

Wainer, Howard. (2005) *Graphic Discovery: A Trout in the Milk and Other Visual Adventures*. Princeton, NJ: Princeton University Press.

Wainer, Howard. (2008) Improving Graphic Displays by Controlling Creativity, *AmStat News*, **21(2)**, 46–52.

Wilkinson, Leland. (1999) *The Grammar of Graphics*. New York: Springer.



The original of this plot is 10.5 inches wide and 6 inches tall and rendered in the deeply rich and flexible portable document format (pdf). As a vector-based plot, it is fully scalable and maintains its clarity and resolution when magnified or reduced. An electronic copy is available from the author and makes a great gift.

The buildup of fatty deposits in artery walls is called peripheral artery disease. Patients can be classified, using an ordinal scale, based on the level of peripheral artery disease as having either no disease, claudication (pain in the legs due to poor circulation), or critical limb ischemia (dangerously restricted blood supply). Patients in a study were assessed relative to their peripheral artery disease state, a tibial artery calcification (TAC) score, and a host of covariates (age, race, smoking status, and the like). The goal of the data examination was to understand the relationship between peripheral artery disease category (the outcome) and TAC score (a measure of calcium in the leg arteries) and the covariates; in particular, how do changes in TAC score impact the probability that a patient will be in a particular peripheral artery disease state?

The plot above shows model probabilities for a collection of patients, estimating probabilities of being in each of the three disease states as a function of tibial artery calcification score, with each of the covariates held at its mean level. The model is an ordinal logistic regression model, an extension of a logistic regression model where the outcome, instead of having only two levels, say 0 and 1, as would be the case in a logistic regression, can have many different levels (in this case, three levels), with the caveat that the levels are ordinal — they carry a natural ordering.

The plot, as presented above, shows a number of pertinent features. At the top is a title which describes what the plot represents. Don't underestimate the value of titles in graphics that might need to stand on their own. Also, prominently displayed is the author of the plot and his affiliation, so as to lend some level of authority and credibility to the information being presented. The three estimated probabilities from the ordinal logistic regression model are presented with measures of uncertainty (the light grey 95% confidence intervals), demonstrating that the author and analyst understand that the probabilities are estimated imprecisely.

The estimated probability curves are labeled on the plot itself, providing instant access to understanding which curve stands for which disease level. Rather than a legend floating nearby and using different, visually-distinct lines to show the curves, the on-field labeling allows the viewer quick and easy access to the categories without having to withdraw attention from the plot in order to determine what is what. The labels require no pointy arrows or connectors; the words' proximity to the curves they label provide all the connection necessary.

The text on the plot itself provides a description of the method and the model, grounds the viewer with a simple but comfortable introduction to the problem, and then tells us how to read the plot and what to see. We are told to move left to right and see that as TAC score increases, the chance of having no disease decreases. At the same time, claudication and critical limb ischemia chances rise. We are told that a TAC score of approximately 100 assigns no disease and claudication essentially equal probabilities and that, for purposes of this model, these three disease classifications represent a partition of the patients, in that the three probabilities always sum to unity.

We are also told how the analysis was carried out and who is responsible for the data measurements and collection.

Does this plot tell us all we need to know about TAC score and peripheral artery disease? Certainly not. But it does tell us quite a bit and it tells us who is responsible. It shows (estimated) distributions of patients' peripheral artery disease at any particular level of TAC and allows us to make comparisons across those levels. This data display reveals the model that is used to *describe* the data set and can be used to help *predict* future observations.

And it shows us that these models can be quite pretty, even in grey scale.

Note that all the relevant data-related elements are prominent and that supporting information has taken a visually subordinate role. The point estimates for the probability curves are in bold black curves, at the front, on top of the other elements, in particular, the ranges of grey that depict the 95% confidence intervals. These, along with the labels and textual narrative, sit on top of the grid of reference lines. The support elements are humble and unassuming, yielding visual prominence to the data-related elements. Like good type-setting or a good wait-staff, the support elements are out of the way until they are needed; they never interfere, they only support.

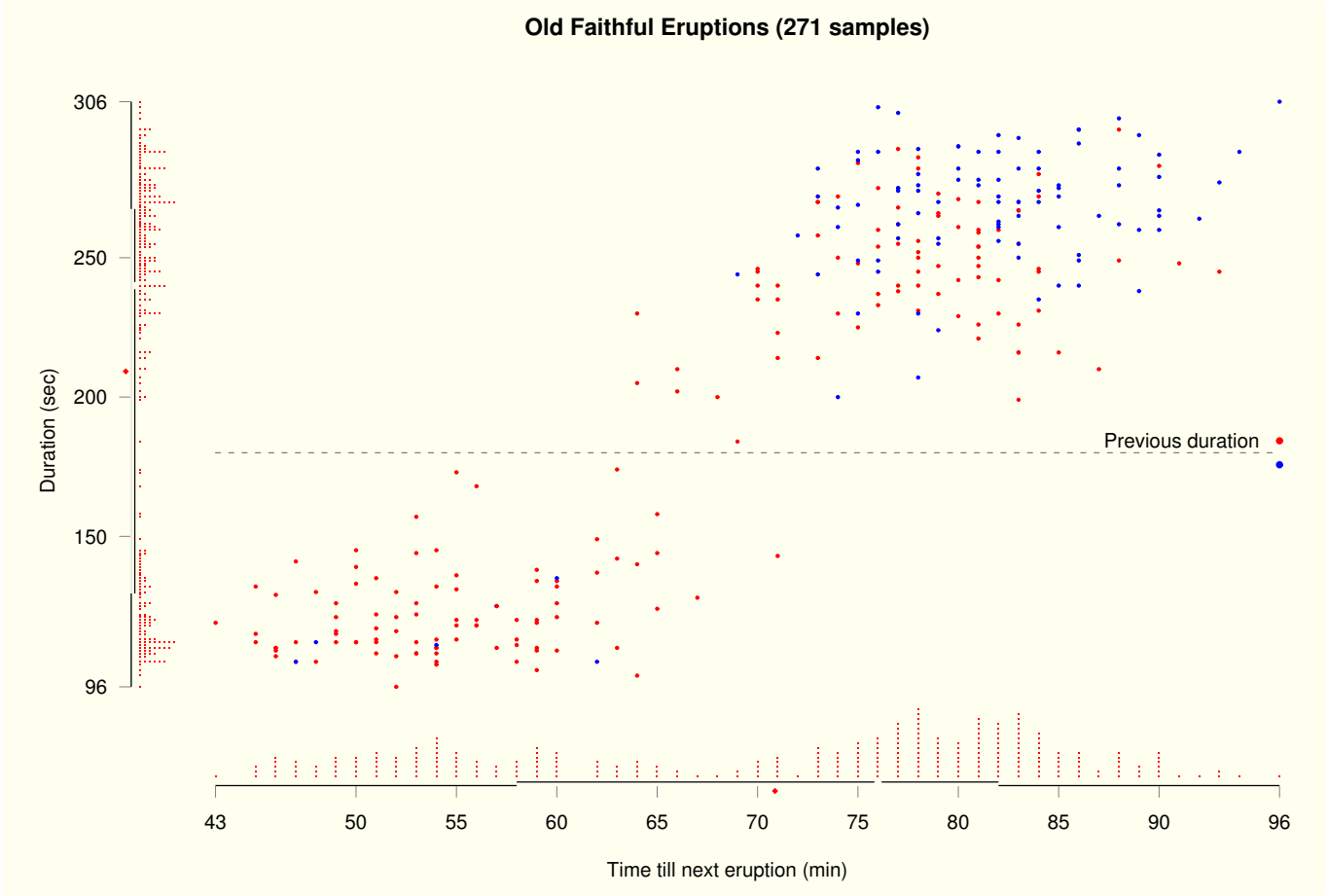
All of these features require time and effort on the part of the author, but the result is a rich, nearly stand-alone graphic that tells a story.

From this plot: **take time to document and explain; integrate picture and text; put words on your displays; tell the viewer what the plot is and what to see; engage the viewer**.

The data in the study from which this graphic derives are complex multivariate data in patients with a complex disease using a complex data analysis method. Why should we expect any figure attempting to enlighten us to this relationship to be without explanation? Should we expect it to be simple? Should we expect it to be small? Yes, a picture may be worth a thousand words but good description and attention to detail never hinder understanding.

Old Faithful geyser is an American landmark, a geological wonder located in Yellowstone National Park in Wyoming. It takes its name from its predictable eruptions.

Data concerning over 250 eruptions of the Old Faithful geyser are included with the R software environment. The relationship between time until next eruption and eruption duration are shown in the plot below.



Graphic courtesy of Steven J. Murdoch, University of Cambridge. Code used to generate this plot is available from the author at www.cl.cam.ac.uk/~sjm217/projects/graphics/. I was fortunate to come across Dr. Murdoch's work through postings on Tufte's *Ask E.T.* forum, which can be found on Tufte's website.

The design of the plot employs some automated techniques for computing summary statistics for the marginal distributions. For both marginals, summaries (the extrema, the three quartiles, and the mean) are shown. The minima and maxima are depicted as the labels at the ends of the axes; the first and third quartiles are shown with the shifts in the axis bars; the medians are shown by the breaks in the bars, and the means are show by the dots just below the axes. Adjacent to each axis is a histogram, similar to what we created for the piano class data, showing each datum where it falls in the distribtuion.

The data are (at least) bivariate and show how eruption duration is related to how long one waited to see the eruption. In general, we see a positive correlation across the range of the data points. The coloration of the dots shows whether or not the duration of the previous eruption was more (red) or less (blue) than an arbitrarily-selected 180 seconds, indicated by the horizontal dotted line. This is showing a negative autocorrelation: short-duration eruptions are followed by long-duration eruptions and long-duration eruptions are followed by short-dura-

tion eruptions; the eruption durations are not independent!

The automatic axes are triumphs of programming and subtle visual information encoding, providing summaries without having to compute and plot them by hand; yet, the summaries, in conjunction with the histogram, demonstrate exactly why such summaries do not tell the whole story. The data are bimodal; they form two distinct groups, one up and to the right and one down and to the left. The marginal distributions also expose this bimodality with two humps or peaks on each axis.

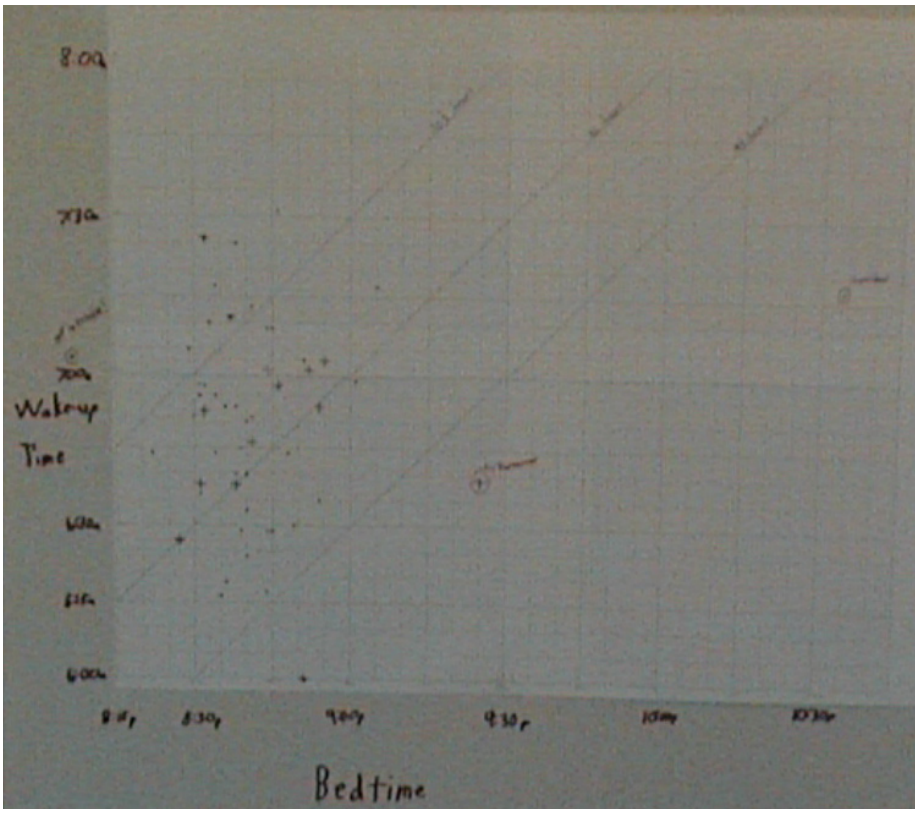
When I was young, I always heard that Old Faithful erupts every hour. (After all, it is Old and Faithful.) Obviously, this is not the case. Examining the data summaries, the argument can be adjusted to say that "on average, Old Faithful erupts approximately every 70 minutes." This is certainly true (truer?), but still a corrupting statement nonetheless. The issue is the bimodality, only revealed by examining the atomic-level data. There are really two peaks. Very rarely are the eruptions an hour apart, in fact, if it has been exactly 60 minutes and you forgot to put new film in the camera, you are more likely to wait for more than 15 minutes than less than 15 minutes; you might have time to reload your film! This is great information. Examining only the summaries would cause one to completely misunderstand the data; showing the atomic-level data reveals the nature of the data themselves. Were we to report only the mean of the time between eruptions (71 minutes) or the median (76 minutes), we would not and could not understand the distribution. All we would know would be the location of the center point, regardless of the amount of mass living there.

The data also show a peculiarity, also revealed by looking at the atomic-level data. The time-till-next-eruption data has no values at 61 minutes. Every other integer from 44 minutes through 94 minutes is represented; where is 61? My personal conjecture is that these data show human digit preference: the desire to conform the geyser's eruptions to man's arbitrary clock may have enticed a data collector to move a "61" down to "60". Yes, it is certainly possible for there to be no 61 minutes observations in the data set but experience with human nature has made me more cynical. Regardless, this anomaly is only revealed through an examination of the data atoms themselves.

Attempt to show each datum; show the atomic-level data. Avoid arbitrary summarization, particularly across sources of variation.

Elementary school science fairs produce wonderful arrays of elementary science lessons, from tadpoles-to-frog development to baking-soda-and-vinegar volcanoes. Occasionally, there are also grand data displays, and sometimes, these are actually instructive.

The figure below is a figure from such a science fair. The resolution problems are the result of poor photography with a weak camera on the part of the photographer, not poor scholarship on the part of the student. A zoomed-in version of the same plot, showing more detail, is on the facing page.



The plot shows 54 multivariate data. Each night, this student recorded bedtime and the subsequent waketime; these values are plotted on the horizontal and vertical axes. The individual points are also coded as to whether that night's sleep represented a weeknight (dot) or a weekend night (cross). Diagonal lines across the chart show nights of equal amounts of sleep. The bottom line, passing through the point (8:30p, 6:00a) shows all points with 9.5 hours sleep. The middle line shows 10 hours sleep. The top line shows 10.5 hours sleep. Furthermore, there are three data points specially annotated; we will take these up shortly.

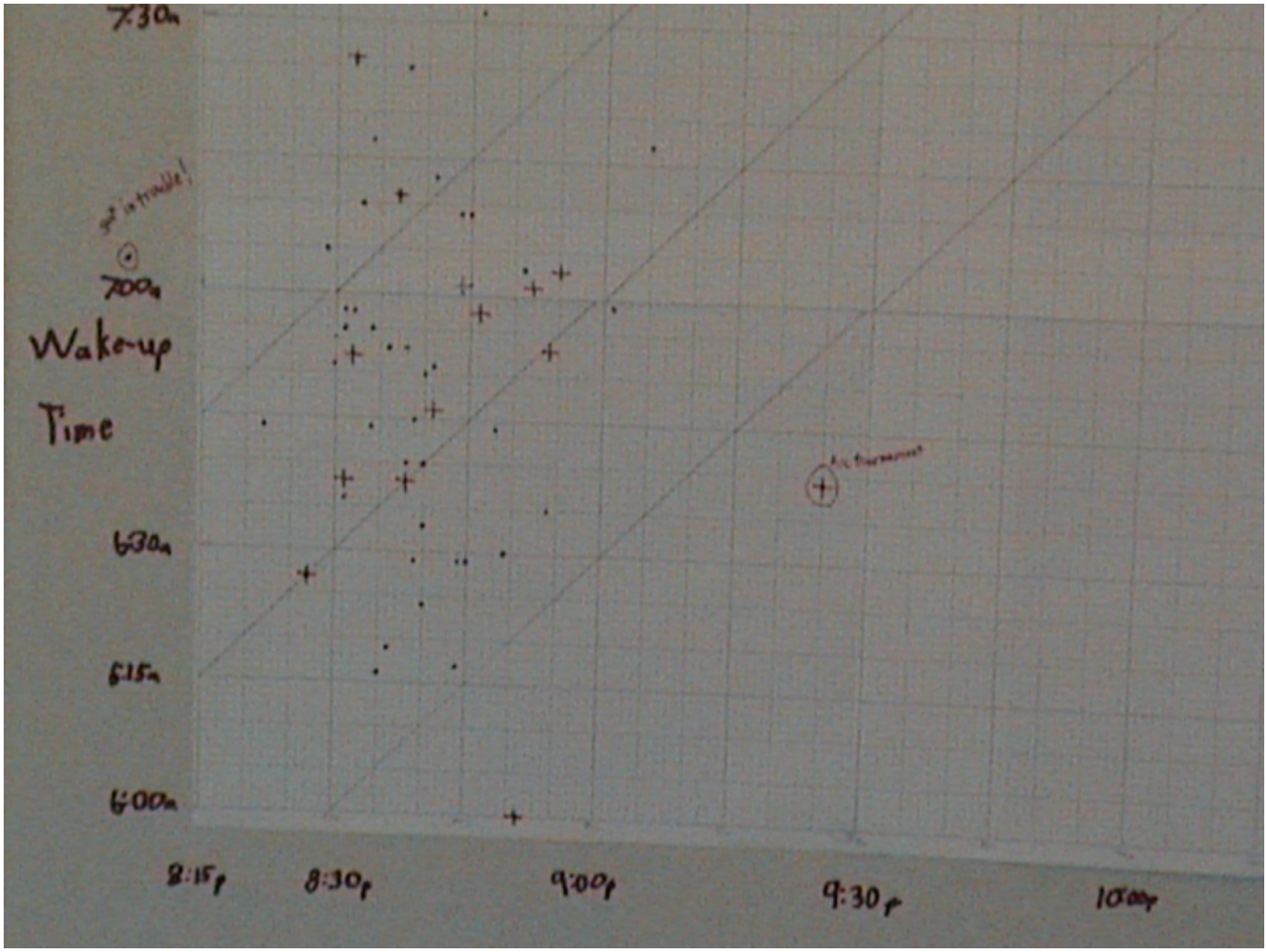
The plot, and the study from which it derives, is designed to answer the question of whether or not early bedtimes yield early waketimes and late bedtimes yield late waketimes. Does this student "sleep in" following a late bedtime?

A solid answer to these questions is not evident but a lesson in understanding data and the processes that generate them, along with reinforcement of data display principles, is available. All the data are plotted: we see the bivariate nature of the bed- and waketimes. The variation in the bedtimes is tight; we must have

parents here who are sticklers for getting to bed on time. Nearly all the bedtimes are between 8:30p and 9:00p, regardless of weeknight or weekend night status. Of the 54 data, all but 8 of them have their abscissa between 8:30p and 9:00p.

The waketimes, however, are much more variable, with no such tight 30 minute boundaries. The parents are obviously letting this child get up on his own volition. A range that includes a similar proportion of the ordinates is nearly an hour in width.

Three annotated points describe the outliers in the data. At the far right, only visible in the zoomed-out version is a bedtime after 10:30p on a weeknight. The description? "Super Bowl". Another anomaly: nearly 9:30p. "Acc Tournament". The last outlier, so far to the left, it falls off the plotting region. Bedtime of barely 8:00p. Explanation? "got in trouble!" Note that we see outliers in bedtime but no such explanations for early or late awakenings. Apparently, such phenomena are not considered note-worthy.



Overall, we see no consistent relationship between bedtime and waketime; we see a child that sleeps until he is no longer tired, and then awakes on his own. I know this to be true; he is my eldest son.

Staining cells consists of slicing up little bits of tissue and then putting staining solutions onto the bits of tissue. Cells with certain characteristics react differentially to different stains. Under a microscope, trained individuals can determine what type of cells are in the tissue by examining whether or not the cells in the tissue react to different stains.

Preparing the tissue specimens takes time and, since the tissue samples are biological entities, the lab researchers who had contacted me were interested in examining the impact of the amount of time used in preparation of the specimens on the reactions of certain types of tumors to certain stains. So, an experiment had been concocted and run and data were collected. Seven tumor types were sampled, each with eight different stains, each at four time points (0, 5, 15, and 60 minutes). At each time point, three replicated readings of percent of cells stained were obtained; the outcome values range from 0 to 100.

There were 7x8x4x3=672 data, less three missing observations. How do we show these data? How do we expose the distribution? Can we show each datum? What are the sources of variation? On the surface, this problem looks like a classic analysis of variance model, with sources of variation being tumor, stain, time, replicates, and appropriate interactions. We could easily drop these 669 data into any competent statistical package and generate the ANOVA. But what do the atomic-level data say?

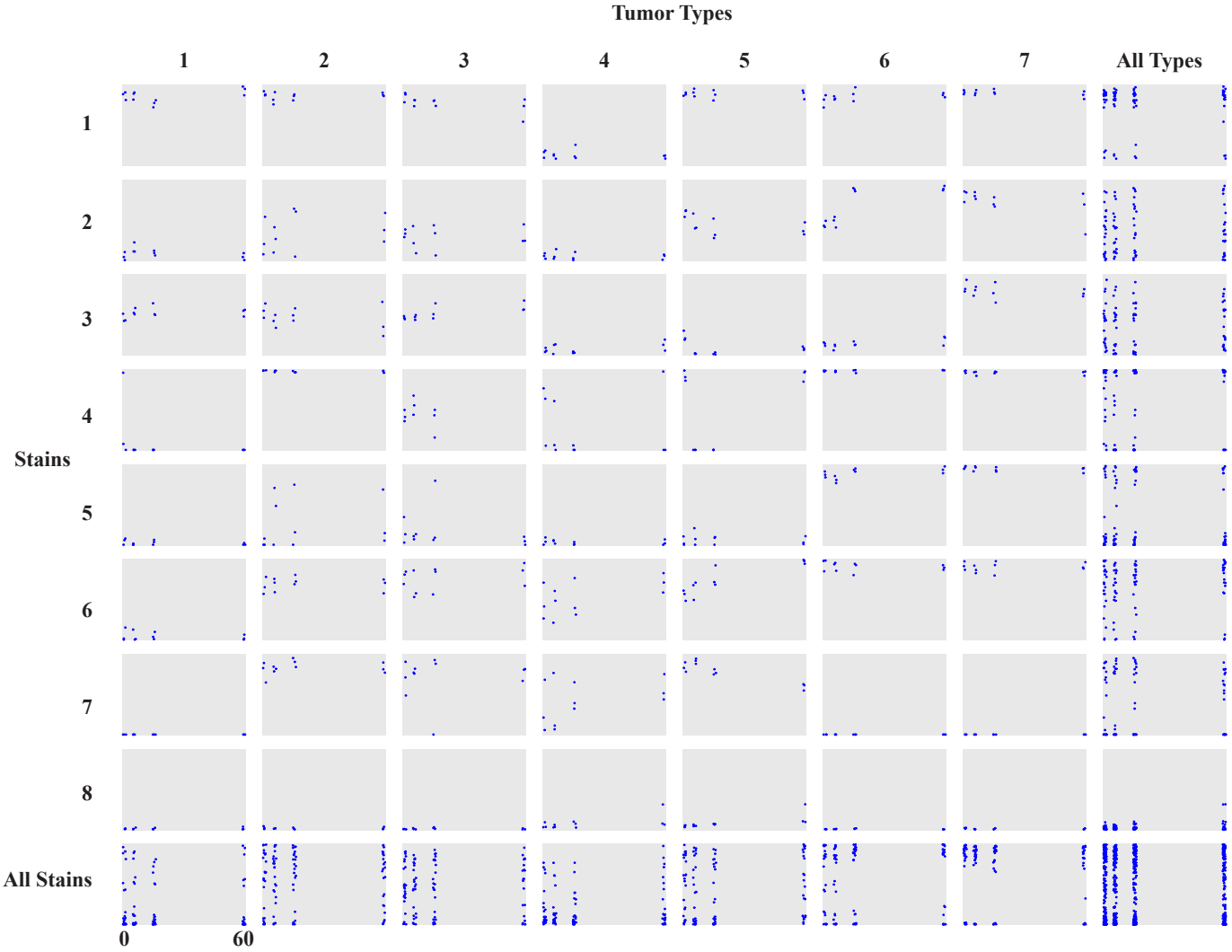
A plot of the 669 data is shown in the following graphic. The seven types of tumors are the columns; the eight different stains are the rows. In each little box, time flows from left to right from 0 to 60 minutes while percent of cells stained is shown vertically from 0% to 100%. We can see the three replicates at each time point. The main effects of tumor are shown in the bottom row; the main effects of the stains are shown in the rightmost column. The intersections of the rows and the columns then represent the interactions between tumor and stain. **Our data display is the model**: we see the distributions as functions of the levels of the sources of variation.

Look at the top row: for this stain, all but the fourth tumor type show high levels of staining. What is the overall effect of this top-row stain? Since it depends on the level of tumor, there is interaction here. Are there other interactions? All but the last stain show interaction with tumor type; main effects in the presence of interactions are of little value. To see why, compare the marginal distributions to the interaction plots: what does the main effect mean when it depends on the level of a second source of variation?

Note the little clusters of the three replicates at each time point in the plot at the top left. These three points show us something about the residual error: it is small — but it is not small everywhere! While the residual, between-replicate error is small for stain 1 on tumor 1, for some combinations, the error is gigantic. Check, for example, stain 4 on tumor type 1. Note the variation at time 0, with one reading near 100% and two readings near 0%. These outliers are rampant: stain 5, tumors 2 and 3; stain 4, tumor 4; stain 7, tumor 3; and more. These outliers are output from the same process that generated the nice data for stain 1, tumor 1; ought we not make certain we understand the source of their variation before even trying to understand the effects due to time?

And what of time? Is there an effect? Whatever it is, relative to the differences seen between tumor and stains and their interactions, the time effect is most

certainly minimal. Furthermore, if there is any evidence of a time effect, it most certainly interacts with stain and tumor type: for example, stain 1, tumor 3 shows a drop over time; stain 2, tumor 6 shows a rise over time; stain 4, tumor 5 shows a U shape.



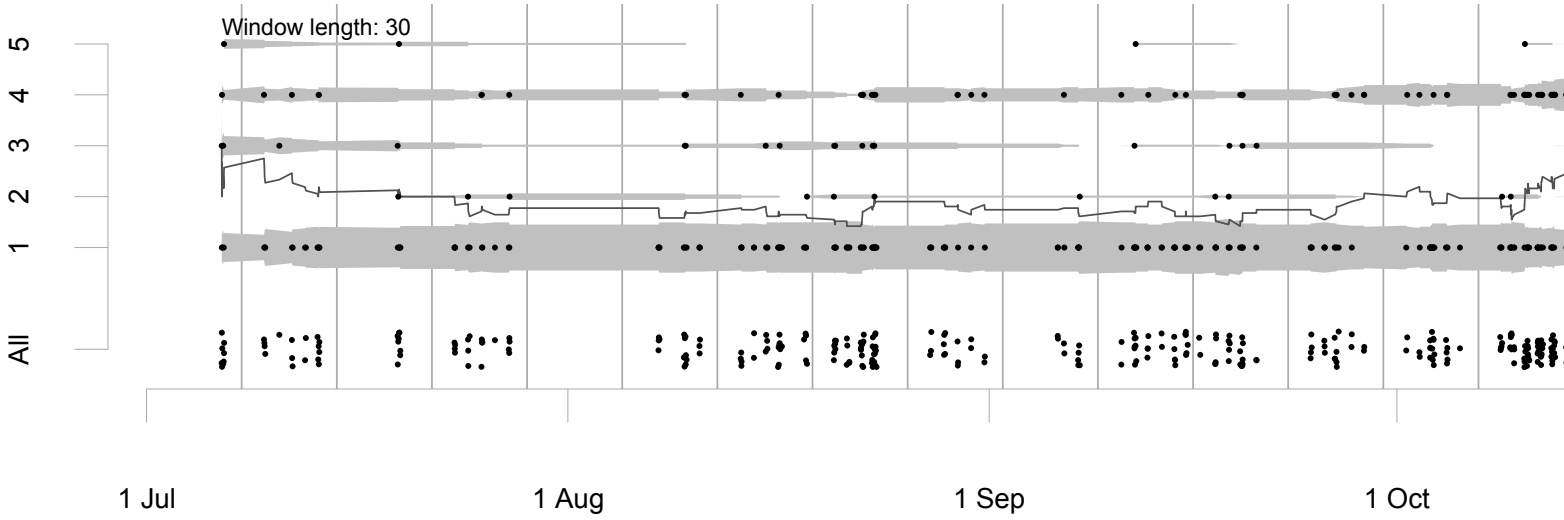
The overall grand mean is in the neighborhood of 50%; in light of many drastic "all or none" experiences in the data, what does such a mean represent? Is reporting the mean even a rational thing to do?

The initial description of the problem sounds like a classic three-way ANOVA with replicates, but examination of the data reveals a number of issues that must be addressed: interactions between tumor type and stain and time, inconsistent variances in the distributions of the replication process, highly nonnormal distributions. The data are complicated from a complicated process. Ignoring interactions and consistencies will not make them go away. The atomic-level data compel us to work harder to understand the sources of variation in the process that generated these data.

People involved with surgery are interested in measuring performances of the people involved in the surgery. Eight measures, and a Global Rating, were collected by an observer for each of 549 surgeries done, with each measure corresponding to something someone is supposed to be doing with (or to) the patient or with (or to) some other member of the staff. The measures were scored on a discrete, five-point scale, from 1 to 5.

Data were collected from early July 2007 through January 2008. The statistician was given the data, which consisted of dates and times of the observed surgeries, identifiers for people involved, and the scores on each of the measures. Summaries (means and such) of the scores had been calculated within each month. Could the statistician please do a further analysis of the scores?

The plot shows the 549 data for one of the eight measures. The horizontal axis shows time, progressing from July 2007 through January 2008. Vertical grey bars mark the start of each week; they are placed at midnight between Saturday and Sunday. The ticks marking the starts of each of the months mark midnight as well. Each datum is shown in two locations: once to mark the day and time of the surgery and once to mark the score given (1, 2, 3, 4, or 5) for that surgery.



A backwards-looking moving-average window of length 30 observations is used to compute both the mean score, as marked by the wandering black line, and the density estimate at each surgery, as marked by the varying grey bands. Hence, computations at a particular observation are based upon the current observation and the 29 that precede it. The vertical thickness of the polygon that is the grey background shows the estimated probability of each score across time, based on the moving average. At any point in time, the total vertical widths of the grey polygons is the same, summing to 100%.

The data row labeled "All" clearly shows that not all days are equal. July 1, 2007 fell on a Sunday, so the first surgeries that were included in the set must have occurred on a Friday. (We can tell they aren't on Saturday by looking farther to the right and seeing some Saturday observations during the second full week in August.) Sunday surgeries first occur during the third full week in September, amidst a streak of eleven consecutive days with observations.

The week that contains August 1 shows no data at all, as does the week of Thanksgiving. The last full week of the year has only one observation, taking

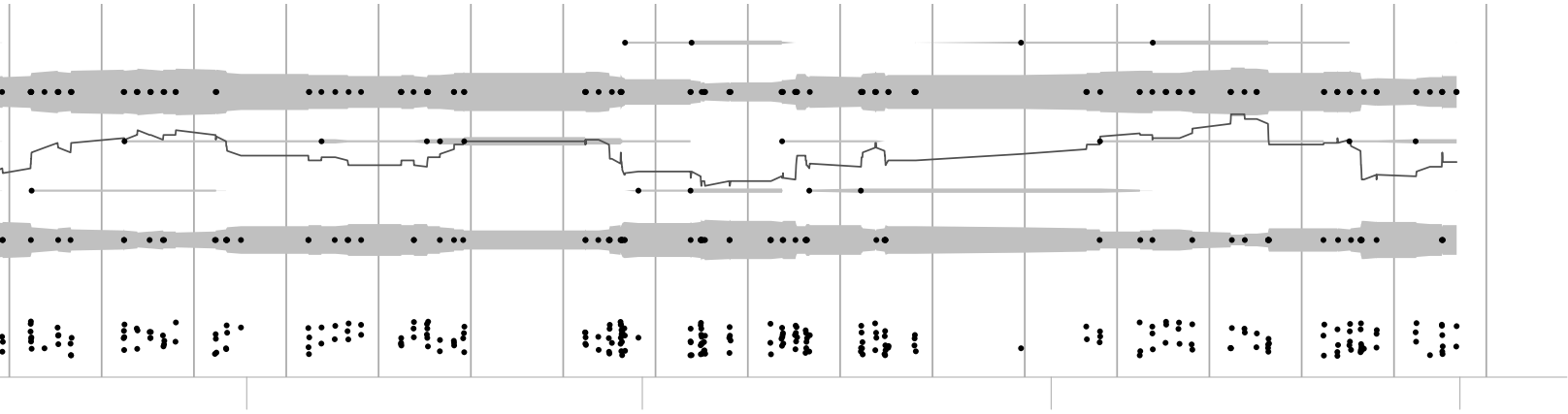
place late on Saturday, the only observation in a twelve-day stretch of otherwise empty days.

Some weeks have a full complement of workday surgeries while others show conspicuous differences. Labor Day is exposed by the absence of observations early in September. Many weeks, too numerous to list, show particular days that are anomalies.

The first full week of October has oodles of surgeries, more than the resolution of the display will allow us to count.

The scores are predominantly 1's, at least until October, when a shift occurs. Starting on October 8, the predominant score (the mode!) moves from 1 to 4; consequently, the mean value then jumps from approximately 2 to over 3. This trend is essentially maintained until the end of the time under study, except during early December and late January, when the 1's return for a brief encore.

Evidence of the temporal intensity in number of observations can be seen not only in the density of the dots in the "All" data row, but also in the length of the



lone data points that score 5. Note that the score of 5 that occurs on September 11 influences the summary statistics for approximately a week, indicating that it took approximately a week to record 30 observations. In contrast, the 5 that was witnessed on Wednesday, October 10, burns out by Friday, October 12, due to the high number of observations during that time.

The vertical positioning of the "All" data row has been jittered, creating some vertical separation between the points. The first draft of this plot routine simply placed the dots at the same vertical location, resulting in reduction in resolution. We do place, however, individual score data at their exact score position, only allowing differences in time to discriminate between equal scores on equal days. An example of the overlap can be seen on many dates, perhaps most noticeably in mid-October where the overlapping dots are wider than a single dot.

Further notes are also in order. One thing that we are *not* doing in this graphic is *accounting*; we are not computing total number of observations in any arbitrary time interval, say, one month, as had been done when the data were originally handed to the statistician. Accounting has a very real purpose and value: it *ac-*

counts. In the sense of these data, accounting is literally counting, adding one to each count in each month for each observation in that month. In order to do such accounting, one must perform some sort of arbitrary binning along the time axis. The account will tell us how many 1's or 2's or whatevers occur in any given period. It is dry and precise, clean and sterile, rigid and exact. Reporting of counts, aka, *accounting*, is certainly a necessary process, most certainly in any sort of regulated environment, in any environment where people need to show the counts, for whatever reason they may possess.

What we are doing by exposing the distribution and emphasizing each individual datum is pushing us more to do *analysis*, the understanding of the total by understanding the component parts. The first thing we see is that the observers of the surgeries, those people who are actually generating the data, are not observing *all* the surgeries, at least, we wouldn't believe this to be the case; how could a surgery center go twelve days with only one surgery? So the observations in our data set are not the complete record of all the surgeries, they are only a sample of what really happened. Is it rational to assume they constitute a random sample? What impact is there on the ability to generate valid inference if they do or do not constitute a random sample? More investigation into the nature of the data collection scheme is warranted.

We also see that we need to be cautious with how we treat each datum, as arbitrary time boundaries and summaries thereof can change the weighting of the data in the summaries. One reason we used a window length based on number of surgeries instead of on absolute time was to follow a "one datum, one vote" policy. By setting up our boundaries relative to the number of surgeries observed instead of setting up boundaries relative to arbitrary time points, we allow the data that fall in the high density areas to carry as much weight as those data that fall in the low density regions. As such, our computations (estimates? of what?) in mid-October weight the data with the same weighting scheme as our estimates in late December. An accounting scheme that partitions time into week- or month-based boundaries weights those data falling in high-intensity regions less than those in low-intensity regions.

Note that day of the week has an impact on the data and whether or not they exist. Surgeries on weekends are much less likely to be in our data set. The impact of this selection bias is not known in these data but any cautious analyst of these data should investigate this sort of concern.

Our data display shows the varying distributions in score value as a function of time. This implies that, on some level, we are viewing time as a source of variation.

Of course, this is pure folly. Time is *not* a source of variation; time is a surrogate, an index that marks locations where other sources of variation have had their impact. A change in score did not occur *because* the calendar said mid-October; rather, *something happened in mid-October that resulted in a change* in score. Any rational investigation of these data will need to examine ancillary or auxiliary pieces of information and determine what happened in mid-October to produce what we have seen: an increase in the intensity of surgery observations and a corresponding (consequential? subsequential?) rise in scores.

Times series plots are always a victim of this pseudo-variation source. Time series plots are very good at telling us what happened and when it happened but

they aren't very good at telling us why they happened. **Time series make fine accounting but poor scientific models.**

We should be wary then of accounting when it is being passed off as analysis. Arbitrary summarization across a source of variation (whatever is also indexed by time) can actually mask variation that we are trying observe. Further examination of these data needs to include those sources of variation (who was involved, what kinds of patients, what kinds of procedures, etc.) that are being surrogated by time. If the slow and steady, unrelenting, dispassionate drumbeat of time is the only source of variation in these data, how can we expect to change the state of Nature and improve the surgery scores? So, unless specifically doing accounting, **avoid arbitrary summarization, particularly across sources of variation**.

When attempting to show large numbers of data points using ink, a graphic will often be problematic because of large areas of overlap of the data points. The simplest solution to this problem is to use smaller dots; after all, in mathematics a "point" truly has no dimension, so why would we want one to be represented by a gigantic dot? Most often, very small dots, on the order of the size of the period at the end of this sentence will suffice, since we are interested in exposing the distribution of the data. We saw small dots in the exposition of the piano class data. Even dots this small, however, are not always sufficient to show the depth of complexity in the data, most certainly when there are repeats in the data or when the dots, when at the limit of visual perception, still overlap.

As we have seen, one way to avoid this overlap issue is to *jitter* the data points. Jittering adds small amounts of noise to each datum so as to allow the individual data points to not fall directly on top of one another. This jittering technique can be seen, for example, in the plot of the cancer staining on page 31, where the horizontal values, the time readings, were slightly jittered. This allows the three readings at, say, time 0 to all be visible, even if they have the same percent of cells stained. A note-worthy example can be found at time 0 for stain 7, tumor 1, where all the response values are 0%. Instead of one dot that is three layers of ink thick, we see a glob of ink representing more than one observation. It isn't a perfect solution, but it is in keeping with a "one datum, one ink-dot" goal. Perhaps some users of our graphics, those with heightened tactile sensory capabilities, could detect a double- or triple-thick ink dot as opposed to a singlet, but most ordinary viewers cannot. Thus, jittering helps to provide a way around the two-data-at-one-location issue.

This graphic from a medical journal‣ is another example of jittering. It exposes some other attributes of a fine data graphic, along with some elements that are unnecessary. We see on a logarithmic scale the APC/MYOD1 ratios from three different groups of samples. Each datum is depicted with its own small dot. The shadow behind each datum creates the effect that each dot is hovering above the grid, which, although fun-looking, is certainly unnecessary. The hover-distance could be used to encode more information. If there is no need to do such additional encoding, then simple flat dots are sufficient. These data are two-dimensional with "group" as one dimension and "APC/MYOD1 ratio" as the other; this is simple "cause" and "effect". Why encode them with three dimensional images?

‣Usadel, H., *et al*. Detection of APC Promoter Methylation in Lung Cancer. Cancer Research, *62*: 371–375, 2002.

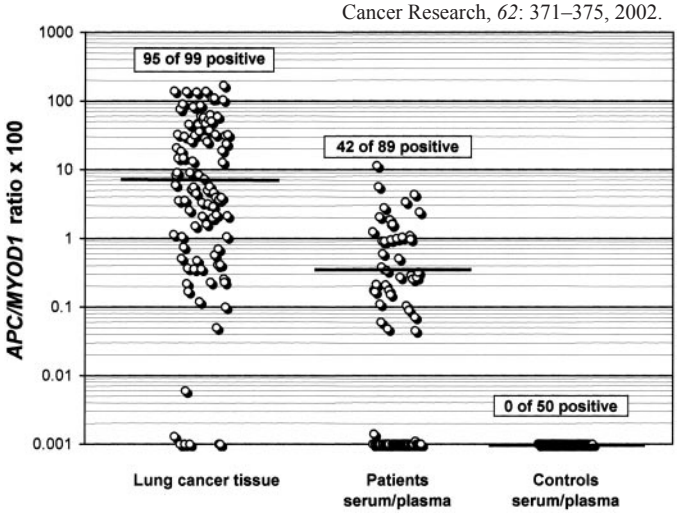


Fig. 2. *APC:MYOD1* ratios × 100 on a log scale. Lung cancer tissue and paired serum and plasma samples from patients with lung cancer positive for methylated *APC* Promoter 1 A DNA. *APC:MYOD1* ratios of serum from healthy control individuals are negative. *Boxes,* the number of samples positive for methylated *APC* promoter DNA. *Bars,* the median *APC*-methylation level within a sample type. Values diagrammed at 0.001 are zero values, which cannot be plotted correctly on a log scale.

The horizontal position of each datum within its category is jittered to allow discrimination among the individual data. Two summary measures are provided. The boxes list the number of positive samples in that group and provide an aid to the obviously difficult task of trying to count all the little balloons. The horizontal bar depicts the median level of the response so as to provide an objective numerical summary measure of the location of the distribution. Furthermore, the description of the plot, including a notation about how zero values are presented on this logarithmic scale, is included in the caption.

This graphic operates successfully on several levels and it is through this multi-layering that it provides its value. At the lowest level are the data themselves, the atoms that underlie the experiment. But these atoms work together to demonstrate the distribution, which is depicted by the aggregate collection and summarized by the median and counts of non-zeros. Good graphics operate on multiple levels, like a good map. The viewer is rewarded for making the effort to look for detail. If I look carefully at the individual data, I can see the four lung cancer tissue data that are not considered positive. I also see a fifth that is nearly at the same level, even though the zero values are artificially placed at 0.001, as is described in the caption. Investment by the viewer results in increased understanding. Getting closer improves resolution.

The same thing happens with a good map; zooming in improves resolution. My road atlas has subtle little blue buildings showing the rest-stops, if you look closely. At a distance, the major roads are visible. As one gets closer, more and more detail becomes apparent.

Tufte calls this concept "layering and separation"‣. Good maps have it. Good graphics have it. **Reward the viewer's investment in the data display** is the principle at work here.

‣Edward Tufte, *Envisioning Information* (Graphics Press, 1990), 53.

Another way around the high data concentration, or overlap, issue, and some other issues, as well, is to plot, instead of the data points themselves, the *empirical cumulative distribution function* of the data. For any data point in the support set, the empirical cumulative distribution function shows (a function of) the number of data points less than or equal to that number. Thus, we can show the cumulative number of points less than or equal to a given point. If we seek to compare multiple such plots, we might seek to scale these plots by the number of data points under consideration and hence plot the proportion or percentage of the data less than or equal to each value in the data set.

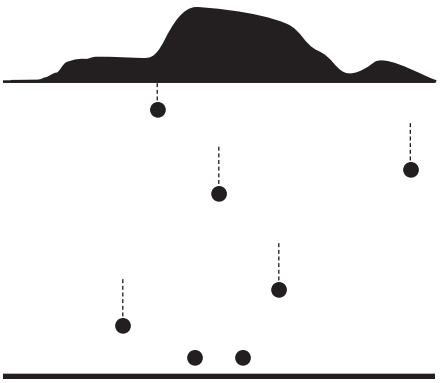
A helpful reviewer has reminded me that some people make an effort to work around the high data concentration by using a plot symbol that is a small open circle. In fact, the default plot symbol in at least one recent release of R is the open circle. An example can be found in the planet plot we examined on page 18.

The plot from the medical journal on the preceding page hints of using the small open circle symbol but the symbols used there differ in two aspects. First, they carry shadows, as we discussed previously. Secondly, they are not transparent; they tend to stack up upon each other: examine the collection of points at the value of 0.001.

Using transparent circle symbols, ie, open circles that would allow a viewer to see the circles cross each other, can aid in viewing data that are closely concentrated but even this trick runs into problems when the data get tightly congested.

Empirical cumulative distribution functions, sometimes called ECDFs, are approximations of their corresponding theoretical cumulative distribution functions, or CDFs. While lay people typically are comfortable with histograms as approximations of density functions, ECDFs and CDFs seem to produce more confusion and discomfort. Regardless, there are times when they prove extremely valuable. Let us examine the path toward the development of an ECDF.

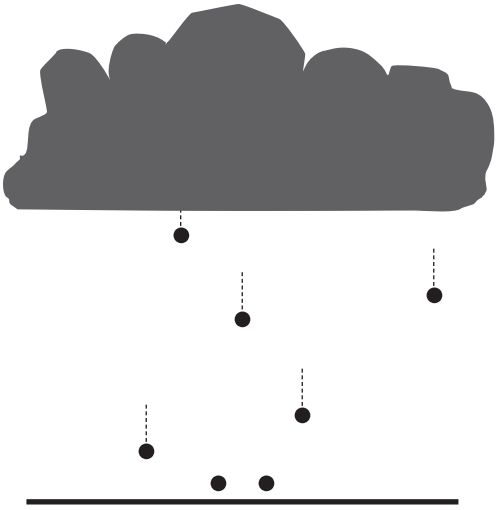
Recall that we seek to expose the distribution of the data. To do so, think of the data as raindrops (datadrops?) that fall from a distribution (a density) onto a line that houses the support set. The datadrops will be most dense at the highest part of the distribution. The drops pile up to estimate the distribution from which they came; the datadrops collectively will estimate the density of the process that created them.



Our little figure here shows the generating density function, the distribution that is generating the datadrops. In reality, this distribution is not know. Some of its attributes may be assumed or conjectured — it may be assumed to be Gaussian or *t*

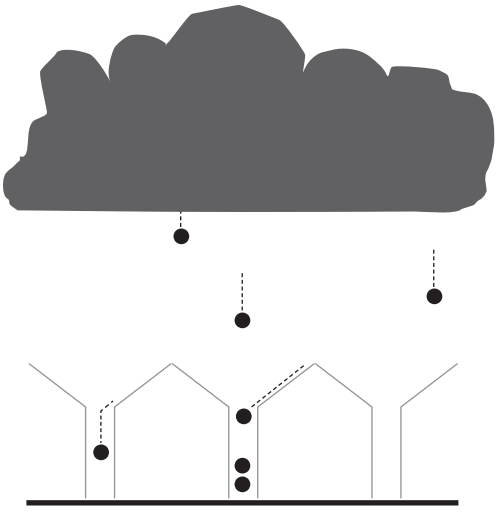
There is no hard and fast rule to use here other than to try a method and examine the result. If a method can adequately expose the distribution (that is, it exposes the support set points and the masses associated with those points), then that method must be sufficient!

or chi-square or a mixture or symmetric or whatever — and we can use these assumptions to estimate relevant components of the distribution. Regardless of assumptions that might be made concerning the distribution that is generating the data, what is most certain is that the distribution is not known. So, in reality, we do not see the distribution generating the data, we see data falling from a cloudy sky, with the clouds obscuring our view of the distribution. Our objective is to make inference about this hidden distribution, based on the distribution of the datadrops that we have seen fall from the clouds.



All of this is fine in the case of a discrete density. In this case, the datadrops will form perfect little stacks on each support point, like the piano class data, with the height of each stack proportional to the estimate of the generating density's height at that point. In the case of continuous data, however, the picture is not as rosy. Since the datadrops have width that is infinitesimally small, the datadrops in this case will not pile up. They will not form stacks or piles. They will only form an infinitely thin layer of data, one whose depth we cannot measure.

One method used as a work-around is to bin the data. Think of a data-raingauge: the datadrops are collected into a partition of bins that cover the support set; each datadrop falls into one, and only one, bin. This then forces the datadrops to stack up. The binning algorithm maps the data to a discrete support set, allowing us to see the distribution.

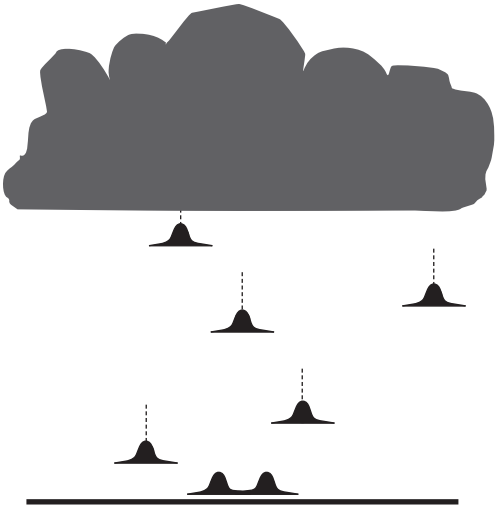


As with any work-around, binning can produce problems. How many bins? Should they be equally spaced? If not equally spaced, how do we scale back the counts in each bin? If the original, hidden, data-generating distribution is skewed, what do I do with the resultant bias? What do I do at the endpoints of the bins? How do I reconcile the fact that two datadrops that start to fall near to each other but are captured by different bins can end up farther apart in the binned representation than two datadrops that start far apart but happen to be in the same bin?

All of these questions are answerable and the statistical literature is rife with answers that produce optimal results in special situations. Unfortunately, however, there is no one binning solution that can be endorsed for all situations.

This does not mean that binning should never be used. It only means that we must realize the issues and biases that arise from the decisions we make to display our datadrops.

Another method for dealing with continuous support sets is kernel estimation, where, instead of infinitely small dots that fail to stack, the data are thought to be miniature densities in their own right. So, instead of viewing the data as raindrops, think of them like pancakes that are high in the middle and low on the outside that will build depth as they overlap each other and form an estimate of the density. Alternatively, think of the datadrops as being made of something akin to statistical silly putty, whereas when they strike the ground, they will flatten and squish outward to make a little mound. These little densities or pancakes or squishes are called kernels and the literature is full of different kernels one can use to help estimate the density.



Of course, kernels have problems that are similar to binning. How wide should my kernel be? What should be its shape? What should I do with the edges of my data support set, where the kernel method places nonzero estimates outside the range of the data?

Again, statistical practice has suggestions for optimal kernels in a bevy of different situations; yet, wider kernels tend to oversmooth the density, attenuating the bumps, and narrower kernels tend to be too focused and overamplify bumps. And, like binning, many of the choices of kernels are arbitrary and there is no one-size-fits-all. Kernel estimates can be helpful but we still need to be cautious of the shy consequences of our sometimes-arbitrary decisions.

(E)CDFs become more and more useful as the distribution function generating the data becomes more pathological. Displays of large data sets, data sets with both discrete and continuous components, and data sets with uneven clusters of points in the support set can be aided by employing ECDFs.

The ECDF can be viewed in our falling datadrops metaphor, but instead of points or pancakes, we visualize each datum as a horizontal stick of thickness $1/n$ and infinite length extending from the datum value out to infinity. Furthermore, we will re-sort the data so that they are in order from smallest to largest. This way, the sticks stack themselves up to make a staircase.

What results from this maneuvering is a non-decreasing function as one moves from left to right across the support set of the data. At each datum, the ECDF grows by an amount equal to $1/n$, where n is the sample size. This way, there need not be any binning nor any arbitrary decisions to get the methods to work. All the data at each point in the support set get to contribute to the estimate of the distribution function. If there is point mass at a single point, say in the case of zero-point inflation, or if the distribution is continuous, the ECDF will allow the

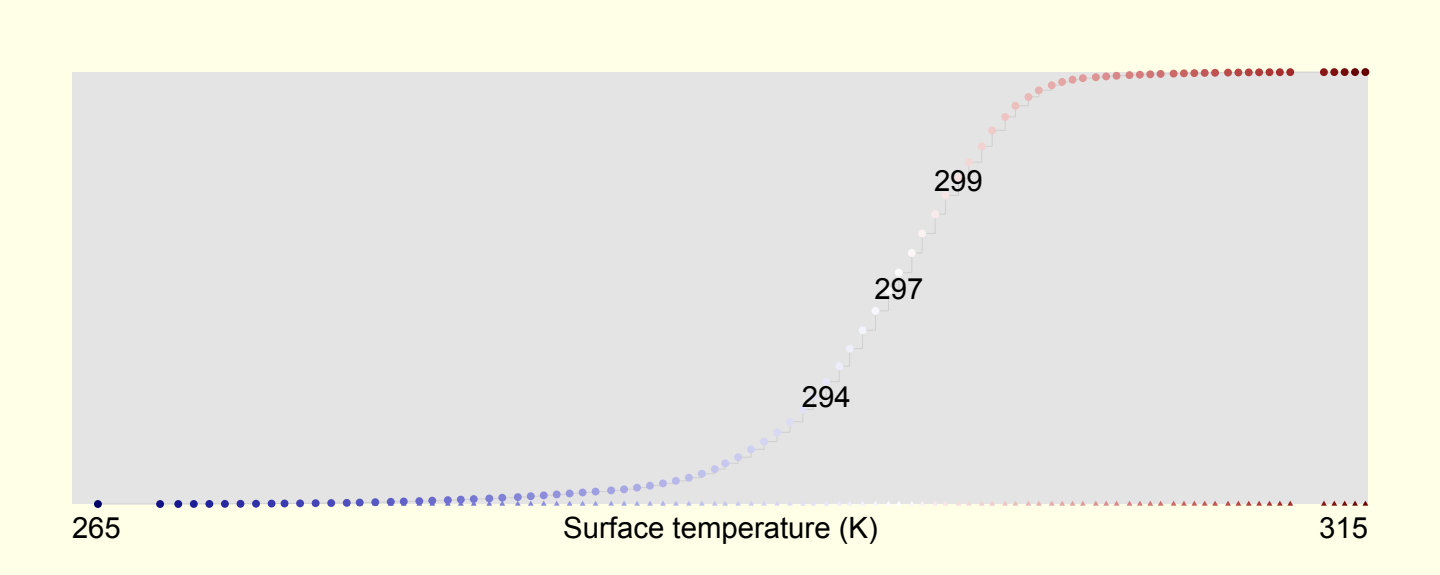
distribution to be seen. Every time the ECDF moves up, there is a datum. Points with more data produce bigger jumps.

The trick to reading an ECDF is to remember that the steepness of the ECDF translates into data density; uniform steepness means uniformly distributed data, flatness means no data.

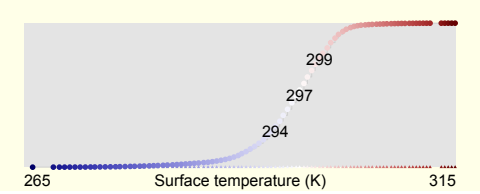
Steepness equals dataness.

Typically we do not keep the long bars but only show the steps and the ECDF then looks like a function, where the action is focused on the support set. We can do subtle modifications to the endpoints and to the summation element, like make it $1/(n+1)$ instead of $1/n$, but the fundament remains the same: each element in the data set causes the ECDF to rise by one unit (or one weighted unit, as in the case of the mean cumulative function, where the weighting can be based on the number of units under consideration at that time point). If we don't want to scale the ECDF to 100%, then the height at each location is an indication of the total number of points, instead of the proportion. Such a scaling is sometimes useful when counts, instead of proportions, are of interest.

Simple modifications to ECDFs can encode extra information. This ECDF shows the distribution of surface temperature data representing 576 geographic locations (a 24-by-24 grid) over 72 consecutive months, for a total of 41,472 data points. Temperatures are shown in Kelvin; conversion to possibly-more-user-friendly units is left as an exercise to the reader. This ECDF adds summary information to the plot, to allow for more information to be transmitted. We have added the value of the three quartiles, those points that show the 25th, 50th, and 75th percentile points (294 K, 297 K, and 299 K, respectively). The range of the grey data field (from 265 to 315) is shown by the labels below the horizontal axis. The support set is shown as the projection of the points onto the horizontal axis. The color scheme maps blue to the lower temperatures and red to the higher temperatures, with white at the median.

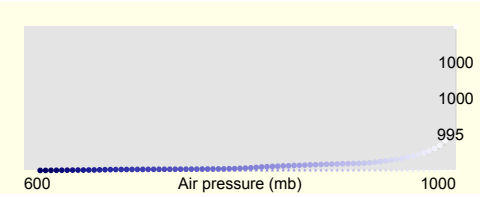


Of course, this plot need not be so large and can be shrunk with essentially no loss of information. Following the "steepness equals dataness" principle, we see the majority of the data, surface temperatures in the this case, are in the neighborhood of 297 K. In fact, 50% of the data are tucked within the two extreme quartiles, within 294 and 299. We can also see relatively uniform spread between the two extreme quartiles and some left skewness in the distribution as evidenced by the lower curve being less drastic than the top curve. It appears that the *left* edge has been dragged out, hence the *left* skewness.



The surface temperature data are only one variable from a collection of weather data provided by NASA to the ASA for the Data Expo at the Joint Statistical Meetings in Seattle in August 2006. In addition to the surface temperature data, each of 24-by-24=576 longitude-latitude pairs over a 72 month period contained monthly averages for temperature, air pressure, ozone, and low, medium, and high cloud cover. The goal of the Data Expo was the visual presentation of these data so as to detect interesting particulars in the data set. More information can be found at http://statcomputing.org/dataexpo/2006

The data set from which these surface temperatures came also includes a number of other variables, among them is air pressure. With these air pressure data we see the value that we get from an ECDF, as opposed to what we would have seen with a histogram. Note the mass of data at the 1000 mb point. This would have produced issues with binning in the typical histogram since over 50% of all the data are at the value of 1000 mb. In fact, 75% of the data are at 995 or above, yet the support set extends all the way down to nearly 600 mb.

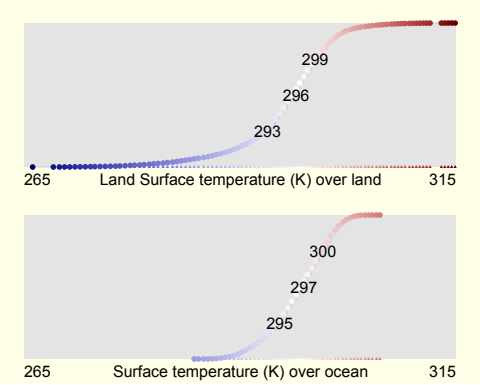


Learning more about the data set reveals why these anomalies exist in the data: the 576 locations where these readings were collected exist over a range of longitudes running from 56.25W to 113.75W, essentially from the East coast of South America to the West coast of North America, and over a range of latitudes extending from 21.25S to 36.25N, from central South America to central North America. As such, there are readings over a wide range of land formations, from deserts to mountaintops to tropical islands and over broad expanses of ocean. The air pressures at 1000 mb reflect readings that are at or near elevations of zero; they are the readings that were done at or near sea level.

These air pressure data are highly skewed left; yet, no rational transformation of them would make them Guassian or even close to it. The mixture distribution, that of point mass at 1000 mb and a greater spread across the other values, provides that thorn.

We see with the air pressure data the reasoning behind the beige background to the ECDF outside of the grey data field: the white data points that show the median are thus visible when they lean over the edge of the grey frame. (In general, another principle arises here: **Colors show off better against off-whites than against bright white. Use off-whites and beiges and light greys for backgrounds to frame displays — or let the data do it!**)

Mixture distributions caused by the influence of land and sea effects can also be seen in the surface temperature data. The ECDFs of the partitioned data collections show the land and sea components. Most striking is the difference in the overall range of the data over the two geological elements. Over land, the range of the data is drastically greater than over the water. While the center points of



the distribution are essentially unchanged, the water temperatures are much less disparate.

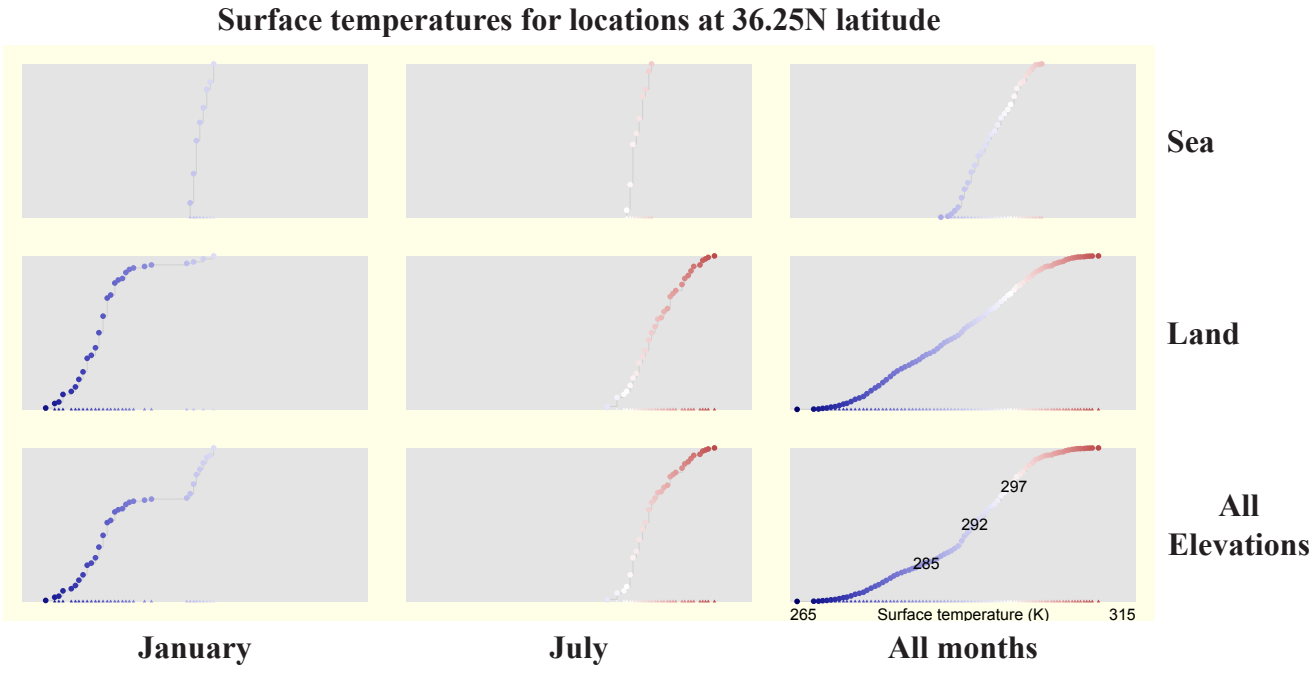
Of course, these data represent the entire constellation of locations from the original data set. Zooming in on a single latitude and examining effects due to time can be carried out via these ECDFs as well. At the bottom of this page we have the surface temperature data at latitude 36.25N, a line the runs roughly from Big Sur, California through North Carolina.

The overall data, across all 72 months of data are shown in the lower right corner, along with the scale range for the grey data field and the three quantiles. We see, via the values of the quantiles and the overall coloring of the points, that these data are pushed in general left of the overall data set. We should expect as much since they are at the northern-most boundary of the complete data set.

The data for all the January readings are shown in the lower left, with the July data in the lower center. Here we see classic northern hemisphere climate: cool in winter (note the preponderance of blue) and warm in the summer (red). We also see the mixture distribution in the January data with the cool water but the very cold land. The July data are also a mixture, although the mixture is less pronounced. Sea and land values across all months shown are in the right-most column. Again, we see increased range of temperatures over land compared to over sea.

[Note that there are ten months whose data are not shown; only the extremes in January and July have been selected. All twelve months data are shown in the 'All months' plots.]

The interactions between time and elevation are shown in the four plots in the upper left. We see the consistent cool in January at sea, the even-cooler land temperatures on land, the consistent warm at sea in July, and the extreme temperatures in July over land. These distributions (along with the ten months not

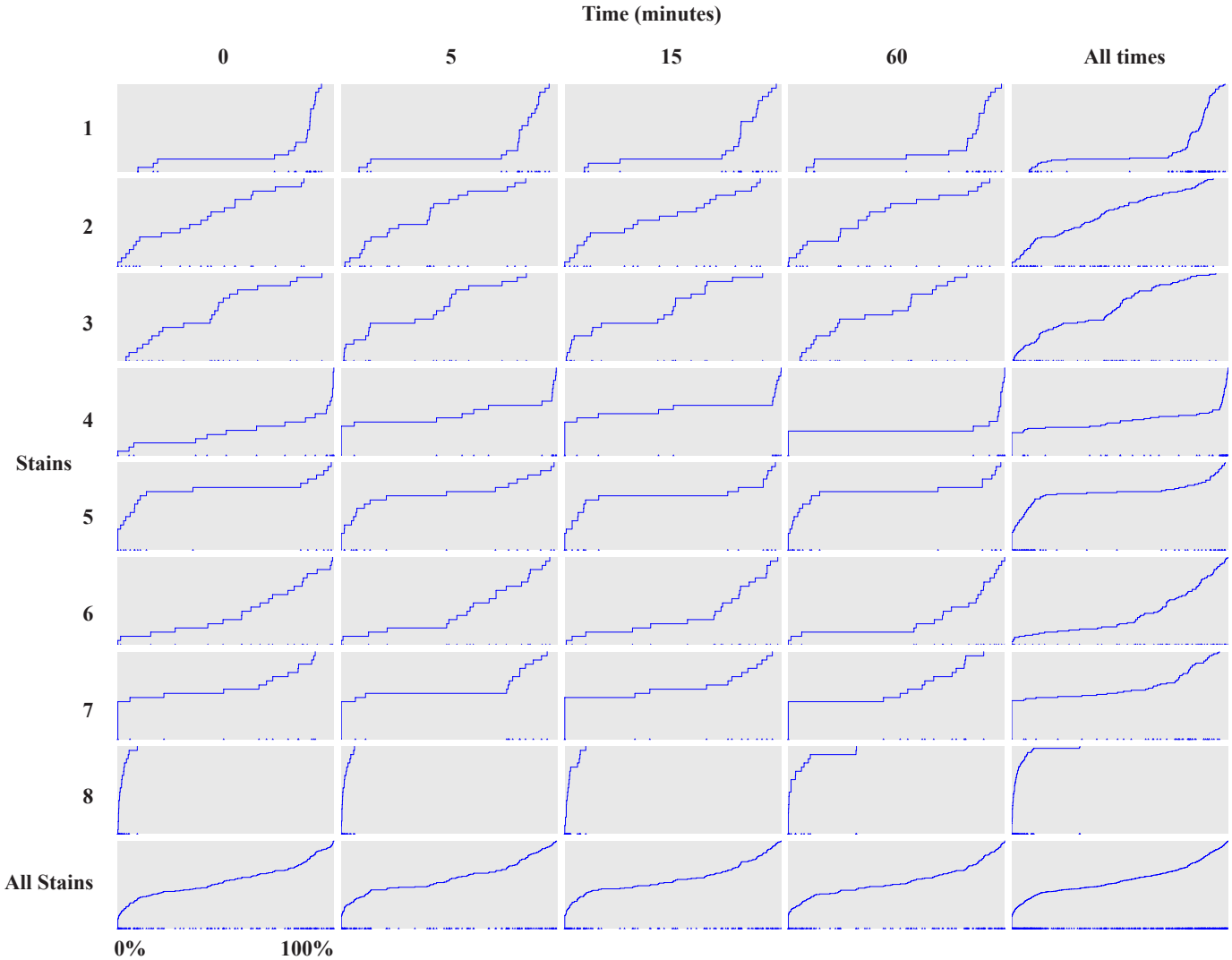


shown) combine to produce the overall distribution seen in the lower-right corner. The component parts are synthesized to produce the whole; the whole must be analyzed to its component parts for us to understand the distribution of the data.

With this new understanding of ECDFs, we can return now to the cell staining data. Recall that there is interest in the effects due to time. As such, we can decompose the data into a collection of ECDFs, again splitting on stain, but this time splitting out the time component and drawing the ECDF across the tumor types and the replicates.

Again, the stains are the rows but now we have the columns as the times. The distribution of all the time 0 readings is in the first column. In the first row, we see the small clump of data on the low end of the outcome distribution (tumor type 4) while the remaining values are clustered near the top. The other elements of the first row are eerily similar to the first column. The four time points combine to produce the smooth ECDF in the far right column.

The main effects of time are shown in the bottom row. The four curves are dead-ringers for the combined ECDF in the lower-right corner, implying that



the distribution at each is much like that overall, and hence the distributions are all similar; there is no noticeable effect due to time. The ECDFs allow us to see points with large amounts of mass; check out stain 5, time 0, or any of the stain 7 plots. The stain 8 plots contain most of their mass near zero.

Steepness equals dataness.

With the original cell staining plot and the cell staining ECDF plot, and also with the date and land/sea ECDF plot, we are beginning to see the value of a concept Tufte refers to as small multiples▸. We have one type of plot, either the scatter plot in the cell staining data, or the ECDF plot in the other data presentations, that is repeated over and over again. The layout of the small multiples acts to index the distributions being exposed, typically with the indices being sources of variation that are relevant in the mental model for understanding the data.

Again, **the data display is the model**, as it exposes the sources of variation and describes each source's contribution to the total.

The small multiple concept is also exposed in the ROP table, where the strata are the multiples and the birth weight category is the source of variation that indexes them.

The idea of using small multiples allows us to make multiple comparisons within our field of vision. Once we learn one of the plots, we can allow our eyes to range over the display to look for comparisons and contrasts. Tufte says in *Envisioning Information* that *Comparisons must be enforced within the scope of the eyespan*. [This is one reason that paper, with its remarkably high resolution, provides such a superior medium for graphic presentation in comparison to the computer screen. While the typical computer screen may present a million pixels, run-of-the-mill ink-jet printers can function at well over a million pixels per sqaure inch, allowing typical increases in resolution, and hence understanding, on the order of 100 times. Large format paper provides even larger gains.]

Small multiples should be arranged so as to facilitate the comparison of most interest to the viewer. In the ROP data table, the fundamental comparison was then versus now. With the temperature data, we can compare January with July and also land with sea.

When more than two sources of variation are present, critically thinking about the comparison task that the viewer is going to make will allow the author of the graphic to make the fundamental comparison the tightest or closest comparison.

If we were to look at the temperature data at a latitude other than the 36.25N location we looked at previously, we would have three sources of variation: time, land/sea, and latitude. We would then need to decide how to arrange the small multiples so as to put the elements of the primary comparison in close proximity. Note that there is no one right way to do this other than to understand what comparisons are of interest to the audience. It also may be necessary to construct multiple versions of the same plot (as we did with the cell staining data) to allow different questions of the data to be answered.

▸Tufte, *VDQI*, 42 and elsewhere. Also, Tufte devotes a full chapter to the small multiples concept in *Envisioning Information*.

William Cleveland, without using the small multiples term explicitly, uses the concept extensively throughout *Visualizing Data* (1993) and *The Elements of Graphing Data* (1994), both published by Hobart Press. Cleveland refers to small multiples as 'multiway plots' and 'panels'.

Leland Wilkinson, in *The Grammar of Graphics* (Springer, 1999), uses the concept of a 'frame' to describe "a coordinate system applied to tuples whose entries are limited to intervals". This allows one to partition the data to demonstrate differences in the distributions based on the levels of the framed variables.

Wainer, in "Improving Graphic Displays by Controlling Creativity" in the Spring 2008 issue of *Chance*, discusses the value in repeating presentation form: "This makes it easier on the reader who only has to master the design once…".

A list of the principles, found on the indicated page in bold

6. The exposition of the distribution is paramount.

9. Show the atoms; show the data.

11. Each datum gets one glob of ink.

22. Erase non-data ink; eliminate redundant ink.

25. Take time to document and explain; integrate picture and text; put words on your displays; tell the viewer what the plot is and what to see; engage the viewer.

27. Attempt to show each datum; show the atomic-level data. Avoid arbitrary summarization, particularly across sources of variation.

30. The data display is the model.

35. Time series make fine accounting but poor scientific models.

35. Avoid arbitrary summarization, particularly across sources of variation.

37. Reward the viewer's investment in the data display.

40. In viewing CDFs, steepness equals dataness.

42. Colors show off better against off-whites than against bright white. Use off-whites and beiges and light greys for backgrounds to frame displays — or let the data do it!

45. The data display is the model.

54. The data display is the model.

54. First Think. Then Plot. Then Think Again.

55. Plot cause versus effect.

57. Typically, color ought be used for response variables, not design variables — but not always.

63. We understand the individual responses by comparing them to a distribution of like individuals.

64. Avoid red and green as comparators. Colorblindness is too common to justify the red-green contrasts. And avoid garish color schemes; try to use subtler colors.

66. Data presentation layouts and designs should be driven by intended use.

66. Time series make fine accounting but poor scientific models.

72. Design first on paper with a pencil and eraser; use a white board. Think about what you want to do and show, not just what you can do with the standard charting icons in the graphing instant wizard machine. If the tool doesn't let you do it, get a different tool.

86. The data display is the model.

—. If everything is bold, then nothing is bold. **Colors matter. Type faces matter. Type sizes matter. Presentation order matters. Have a reason for the choices you make.**

—. How often do we really need to report more than 2 significant digits?

—. Use bigger sheets of paper. 8.5in-by-11in is too small!

Companies with sales forces send information to these sales forces. And they hope that the sales forces read the information that they send and then act accordingly.

One pharmaceutical company with whom I consulted spent considerable sums of time and money developing a set of monthly reports of prescriber habits (which prescribers wrote prescriptions for which products) that were sent to the sales force monthly. Six reports had been developed. Four of them represented a cross between two methods of summarization, "Us" versus "Them", and "Prescriber" versus "Territory": "Our" sales at the Prescriber level, "Competitive" sales at the Prescriber level, Our sales at the Territory level, and Competitive sales at the Territory level. There was also a report of Competitive Sales to Managed Care Organizations and a report of District Sales for Us. A territory was a collection of prescribers; a district was a conglomeration of territories.

The purpose of these reports to the field was to provide the field agents with valuable information that would help determine which prescribers to visit and what to say to them. This, in the end, would improve sales of the company's products and result in higher bonuses and commissions for the field agents.

The data in the reports were consolidated on a monthly basis and presented as monthly totals. In essence, they represented an accounting of the sales within and across prescribers and territories. As with most accounting, the counting component took some time. Hence, data for, say, March were collected throughout March and cleaned and checked and counted in April and released in, typically, early May. April data were processed in May and released around the first of June. May data were processed in June and released around the first of July. And so on. There was a lag time of approximately one month between when the books closed on a "data month" and when the data were available to be released to the field.

When the company released the data to the field agents, the data were sent electronically and the company decided to check to see who looked at what reports and when. So the field agents' laptops were equipped with software that tracked when they looked at each report, every time they looked at a report. As the consulting statistician, I was given these "who-looked-at-what-and-when" data and was asked to determine if the field agents were looking at the reports and using them. The company sought "80% utilization", whatever that means.

There were literally hundreds of thousands of data records, each one indicating a "who" (the id of a looker), a "what" (which report the looker looked at), and a "when" (the precise date and time when the looker looked at what was looked at). The report that had been accessed was indexed by the type of report (one of the six reports listed above) and the data month it represented.

How to report such data?

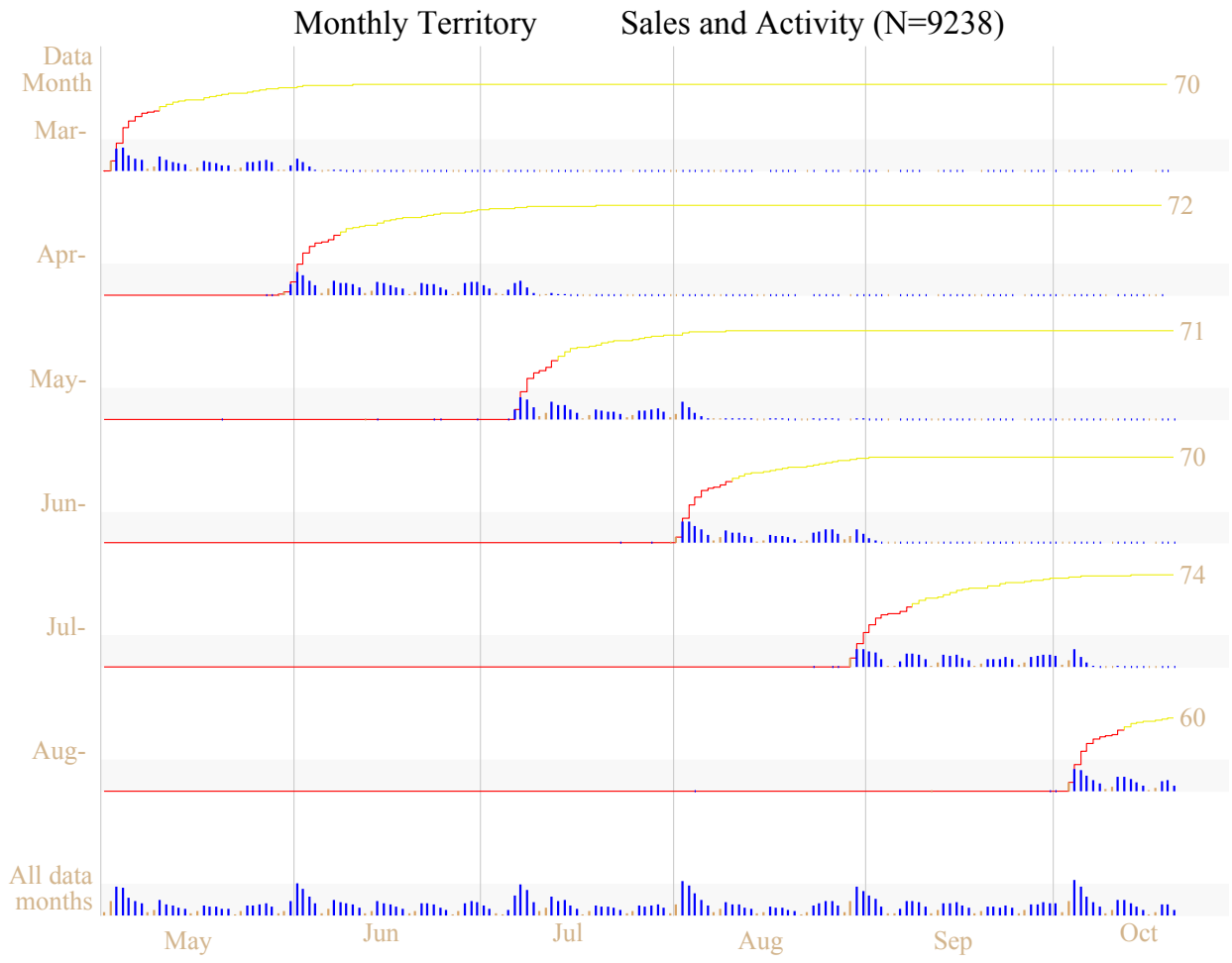
The atomic-level data here are impulse data, little blips of action that occur at a specific point in time. The IT group had been able to easily account the data, counting how many blips of each type occurred during an arbitrary interval of time. As such, they could compute "numbers of hits" during any given day or week or month and plot pretty 3-D spinning hexagonal bar-graphs and show them over time, a nearly content-free presentation that was accompanied with plenty of pomp. Was it possible to show more detail, a closer image of the data

atoms? Could we understand more about the data than just an accounting?

The following plot shows the report utilization for the time frame from 1 May through 21 October for the Monthly Prescriber "Us" Sales and Activity Report, one of the six reports being examined. (The precise year and company name have been obscured on confidentiality grounds and the precise data have been tweaked.) The plot shows seven time series, one for each of the six most-recent data months and one for the entirety across all data months, based on the 9238 field agents in the data set during that time frame.

The time axis depicts the individual days, with weekdays in blue and weekends in dusty sandy color. Month boundaries are shown with the light grey vertical lines; this time axis is consistent throughout the six data months and the composite. Each little vertical bar on a given day shows the proportion of field agents who looked at that report on that day, with no line meaning zero and a line all the way to the top of that month's range being 100%. The shaded region directly behind the vertical bars marks 25% utilization.

Rising from each skyline of little bars is a smoke-trail that represents the cumulative utilization; it shows the cumulative proportion of field agents utilizing that report from that data month at that time point. This ECDF is color-coded to



change from red to yellow at 50% and from yellow to green at 80% utilization; the people investigating the utilization data had "dashboards" and "metrics" that sought certain levels of usage — red is bad, yellow is caution, green is good. The most recent cumulative percentage of utilization is labelled at the far right, so that the graphic also contains some functionality as a look-up table as we can tell exactly what the utilization levels were for the data months in the report.

Moving from left to right, and generally from top to bottom, a wonderful richness in the utilization data emerges.

The March data month reports were released likely on a Sunday, 2 May. The Monday and Tuesday that followed, 3 and 4 May, both saw approximately 20% of the field agents use the fresh reports, as the bars there look to be approximately 80% of the 25% band. The jump in cumulative utilization seems slightly higher on Monday than on Tuesday, even though the individual days' utilizations seems slightly higher on Tuesday than on Monday, implying that some people who looked at the reports on Tuesday had already viewed them on Monday. Utilization of this report drops over the course of this week, with days later in the week showing lower utilization than days earlier in the week, a pattern that is repeated across many of the weeks under consideration.

Within a week of the apparent release (I was never told *exactly* when the reports were made available, as this was a source of some contention between different groups within the company), the overall utilization of this report passes 50%, as the line changes from red to yellow. But the rate of new looks, as revealed by the slope of the cumulative line, and the total number of lookers, as revealed by the little blue and sandy bars, is declining now that these data are more than a week old.

Starting with the third week in May, usage is reasonably stable, but there are few new users. The new users seem to show up more often on Mondays, as this uptick in the cumulative line seems steepest. Weekends are drastically different from weekdays.

But the end of May brings the advent of the April data month data, apparently a day or so early (!) as there is a spike in utilization of the April data month data on the last day of May, a Monday. Utilization of April's data looks quite like that for March's data, with the exception of the first week, with the first Monday's utilization of April's data lower than the first Monday's utilization of March's data. Perhaps the early arrival of the reports caught the users off-guard, or perhaps Memorial Day observances took priority over viewing Monthly Territory Sales and Activity reports.

Regardless, the cumulative line grows to at least 50% in a week and then tapers off, much like what happened to March's data. We see, however, six weeks of non-trivial usage of the April data before it becomes obsolete (due to the arrival of May's data) instead of the five we saw with the March data. This is a consequence of the early release of the April data month data, along with what appears to be late release of the May data month data.

The May data month data appear to have been released on the Tuesday after the 4th of July holiday break. With the 4th landing on a Sunday, the following Monday was an off day. The May data month data are pushed to 50% cumulative utilization in about a week as well, with no individual day have more than

the typical 20% usage.

We might now glance down to the cumulative row and note the dramatic spikes when new data were released and the within-week declines and the dismal weekends, although Sunday does seem to trump Saturday.

We might also glance to the top and notice the curious case of the March data month data still being used by at least someone after the April data month data had been released, and after the May data month data, and so on. Someone was still using the March data month data in early October, after five updated versions of this report had been issued! Why?

Looking at the May data month, the eagle-eye might notice some curiosities in the May data month utilization *before* the May data were released. Note the very, very small tick on approximately 19 May, and again on (approximately) 14, 23, 24, and 30 April: the May data month data had been viewed *nearly two months before they were released!?!* Furthermore, now note with some horror that *all* of the cumulative lines are red and *show some utilization prior to the reports actually being issued!*

This glaring error must certainly point to faulty programming on the author's part, right? On the contrary, an investigation into the utilization database, sorting and then displaying the raw data based on date of report usage, revealed that hidden within the hundreds of thousands of raw data records were some reports that were accessed prior to the Pilgrims arriving at Plymouth Rock in 1620! How could such a thing be so?

The answer is that I lied when I told you that the database held a record of the *date and time* each report was accessed. I told you that because that is what I had been told. The truth is that the database held a record of the *data and time that was present on the field agent's laptop* each time the report was accessed. Some of the field agents' laptops, therefore, held incorrect date and time values. Viewing all the atomic-level data reveals this anomaly. Simply accounting the number of impulses in a certain time interval does nothing to reveal this issue with data.

But I know that it is not possible for those usages to take place before the reports were released, so why not just delete all the records that have dates prior to the release date, since they must be wrong? True, deleting all the records prior to the release would eliminate that problem, but such a solution is akin to turning up the car radio to mask the sounds of engine trouble. The relevant analysis question that has been exposed by actually looking at the data is *why are these records pre-dated?*

We see variation in the data; we need to understand these sources of variation. We have already exposed day-of-the-week, weekday/weekend, and data-month variation. *What is the source of the early dates? Are the field agents resetting the calendars in their computers? Why would they do that? Is there a rogue group of field agents who are trying to game the system? Is there some reward they are perceiving for altering the calendar? Is there a time-zone issue going on? Are the reports released very early in the morning on the east coast, when it is still the day before on the west coast?* And, most vitally, *how do we know that the data that* ***appear*** *to be correct, actually* ***are*** *correct???*

The process that generated the early usage data was the same process that generated the correct-looking data; why hold suspect only those data that *appear* out of line? How do we know that a usage record that appeared two days after release really occurred two days after release and not two weeks after release?

The original accounting of these impulse data counted the raw number of hits on each report during each calendar month. Looking at the bottom "All data months" row, we see the daily usages, summed across all the data months, that would be used to generate such counts.

The accounting time series plots showed a spike in usage of this report in August, with a dips in July and September. Such a dramatic change caused much consternation in the company. *Why were these changes appearing? Do we need more training? Was there an important announcement? Were new agents coming on line?*

We can see now, using the daily data and splitting on the data months, why these blips occurred. The May data month data were released later than might be typical in July, after the July 4th holiday, on a Tuesday, 6 July. This left some residual usage four weeks later, during the first week in August. Furthermore, July data month data were released *early*, in late August, stealing some of September's usage and assigning it to August.

So, August got increased hits due to differences in release dates and calendar quirks. There were no problems with training or announcements or new agents. Accounting processes (using arbitrary calendar month cut-offs) out-of-step with the data release process (subject to whims due to holidays and weekends) produced what appeared to be variation between data months when no such differences existed! The arbitrary binning served only to cover and obfuscate the variation sources inherent in the data as they were being generated.

But this is only one of the six reports upon which we could report usage data. The complete constellation of the six reports is also available.

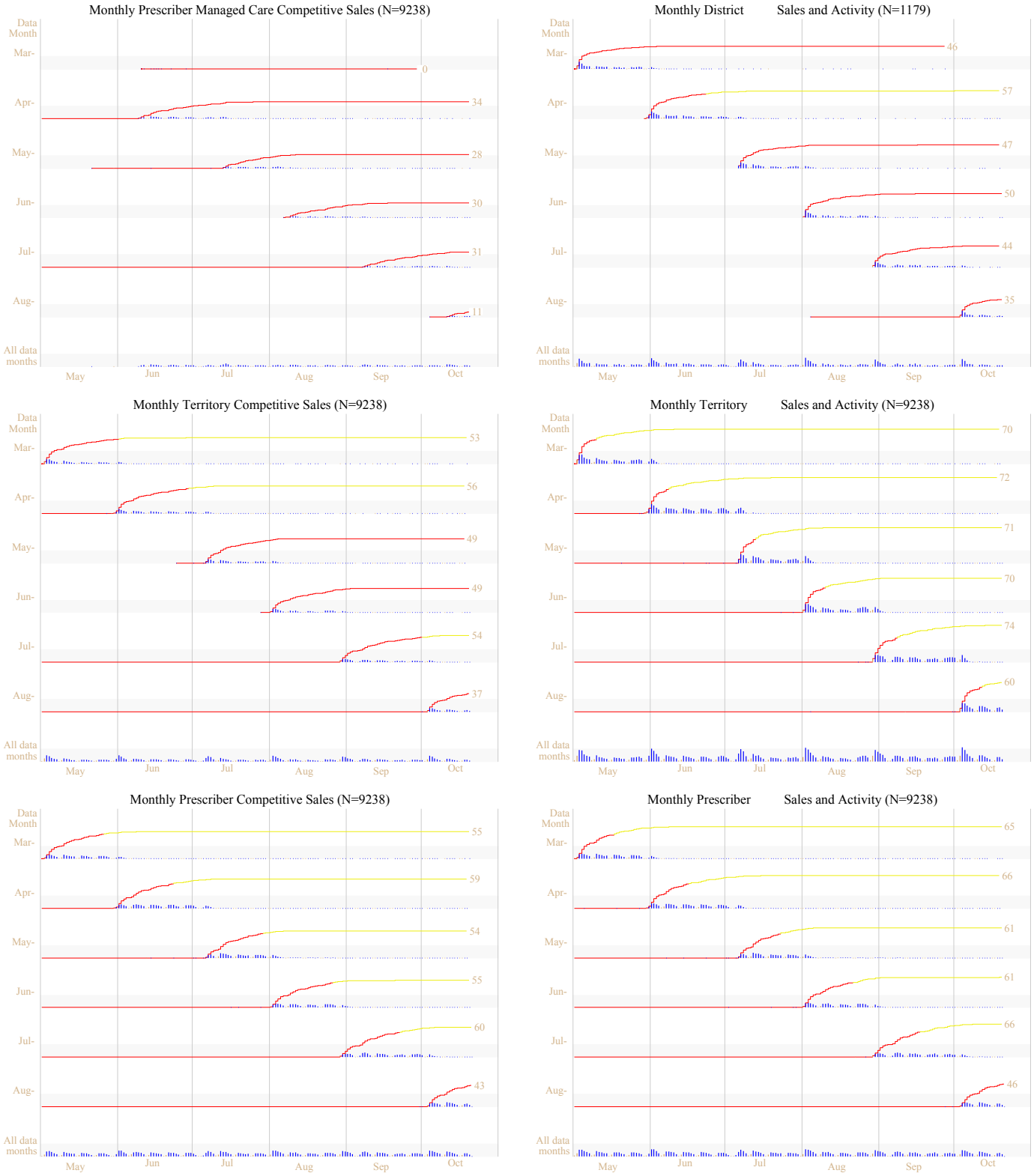
The next page shows all six reports, organized onto a grid. For the bottom four reports, the right column is Us and the left column is Them while the bottom row shows data concerning the Prescriber reports and the middle row shows the Territory-based data. The top two reports show the two other report data: the Managed Care data and the District data. The latter is of special interest to District Managers. At the top of the report is a (partially blanked) title, along with contact information with regard to the author. The bottom of the page provides a quick summary of the plots (one that was constant across all runs of these reports on reports), some comments regarding what to see or note ("The managed care report shows substantially lower use than the other five reports."), some credit to others involved in making the report possible, and a date/time stamp.

Each of the six plots is identical in form to the others and differs only in content. Once we have learned to read the Monthly Territory Sales and Activity report, we can now read the five others.

Being able to see all the plots on one page facilitates comparisons and con-

report utilization: 1 May through 21 October

Rafe Donahue, PhD



This collection of six plots shows report utilization on each day from 1 May through 21 October for the six monthly reports. On each day, for each of the data months shown and across all data months (including those not shown), the proportion of users who used the report is shown. The denominator used (N) is presented in the title for that plot. Weekdays are shown in blue and weekends in orange. Horizontal reference lines depict 25% utilization. Each cumulative proportion line, labeled with the most recent value, changes from red to yellow at 50% utilization and to green at 80% utilization. The managed care report (upper left) shows substantially lower use than the other five report. Note that the cumulative percentage numbers for the same months differ from (are lower than) what they were on the last report. This is due to inaccuracies in the computation of denominators for computing percentages. This will be addressed in future reports. While the actual percentages are suspect, the overall shape of the usage patterns is still valid. The raw utilization data that were used to generate this display were extracted on 21 October , courtesy of and , Data Delivery Systems. Run date: 13:56 01DEC

trasts between the usage patterns of the different reports. First, note that not all data months in all the reports show the dramatic, off-the-chart early usage. The Monthly Territory Competitive Sales report (middle left) usage has two data months (May and June) that show early, but not pre-May early, usage. The Monthly Prescriber Managed Care Competitive Sales report (upper right) and the Monthly District Sales and Activity report (upper left, used exclusively by District Managers, those employees overseeing the field agents), show reduced early usage too. Perhaps the District Managers, those in charge, are the more grown-up ones, playing fewer games with the calendars. The reduced early usage in the Monthly Prescriber Managed Care Competitive Sales report (note that the cumulative usage proportions are all less than 35%) may be explained by its lower overall usage; with fewer people using the report, we are less likely to have people with wayward calendars using the report.

We also note that something is odd with the March data month for that report, as its usage is nil. Later investigation revealed that this report was released in conjunction with the April data month version of the same report and was only released as a shell.

The take-up rates of the five reports we have not examined are considerably lower than the take-up rates for the Monthly Territory Sales and Activity report (middle right) that we examined in detail earlier. Whereas the first report we examined reached 50% usage in approximately one week, the report below it (lower right) takes approximately two weeks to reach 50% saturation, with three weeks for the lower left, and nearly a month (or longer!) for the Monthly Territory Competitive Sales report (middle left).

Why should we see such an interaction between Us/Them (column) and summarization level (Prescriber/Territory) when it comes to report usage? The answer can be found by following the money.

While the company sought to improve the company's bottom line, the field agents were typically just as centrally focused. Field agents' bonuses were based on their sales in their territory. As such, the Monthly Territory Sales and Activity report (middle right) carried the information that was most closely a surrogate for the field agents' bonus check amount. Next closest to home was the data based on the individual prescribers, as these numbers could point to which prescribers were doing what. Further down were the data in the lower left report, those data that showed what competitors' products the prescribers were writing. And last was how the competitors did on the Territory level; sometimes this report failed to generate 50% utilization.

Note the anomaly with the upper left report and the start of usage with the June and July data months. Where the other reports suffered from the aberrant accounting / arbitrary binning problem we discussed earlier, this report, due to the fact that it was produced through different means by a different group and released at a different time, suffered no similar problem. The time series report of hits in a month showed no spike in August and dip in July and September.

The note at the bottom of the report discusses issues with the denominators: field agents come and go — getting a count is a more difficult endeavor than one might imagine. As such, the number listed in the report (9238) was an estimate that was often contested and whose refinement was a constant work-in-progress. And the stated goal of 80% utilization was in constant flux.

This report, then, *shows a model* for varying utilization proportions. **The data display is the model.** In general, in a data display, the display design parameters show the sources of variability: rows, columns, sort orders, subrows, subcolumns, locations, shape, etc. Distributions of responses (variability) can be connoted by location, color, item size, shape, etc. The layout of such a model should be driven by its intended use and such a layout should facilitate the comparison of interest, typically a comparison of distributions.

Our display here shows that our model supposes the differences in utilization proportions are functions of report, data-month, and date of usage. Furthermore, we see that the report component can be viewed as having further sub-components of Us versus Them, Prescriber versus Territory, and the interaction of those components. Utilization also depends on field agents versus managers.

And hiding deep inside all of this is a strange and funky interaction between release date and day of the week and weekday/weekend and month boundaries.

We cannot pretend that the usage data can be understood by just accounting; *it's more complicated than that*‣.

‣Edward Tufte, 2008. "Grand truths about human behavior" <www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=0002XS> [cited 11 June 2008]. Among the other Grand Truths are All the world is Multivariate, Much of the world is distributed lognormal, Rehearsal improves performance, and Richard Feynman's "For a successful technology, reality must take precedence over public relations, for Nature cannot be fooled."

Wanting it to be so, does not necessarily make it so.

This report of the usage of the six reports can be adapted easily to use at varying levels of the organization. It can be rerun on, say, a District Manager level, so such a manager can see what his or her field agents are doing.

Or it can be split based on the products being addressed by the field agents. Perhaps the field agents representing different products behave differently.

At the lowest level of detail, this report can be run all the way down to the individual field agents and we could find out which ones are doing untoward things with their computers. Zooming in on the atomic-level data while preserving the form will allow us truly to do *analysis* of the data. Zooming all the way out and blurring the message in the data by computing arbitrary monthly summaries only covers the variation sources we are trying to expose by collecting the data in the first place. The most obvious feature of the data display that the canned program had produced, the surge of use in August and the dips in July and September, had really nothing to do with the actual data themselves. This feature wasn't even a feature of the *data* at all; it was a consequence of the arbitrary binning that came from a desire to produce monthly totals in a canned database reporting tool.

Just because the reporting tool that comes with the database program can cut the data easily by monthly boundaries and produce pretty 3-D spinning hexagonal bar-graphs doesn't mean that that is the right model for the data.

I have a t-shirt that I got free from Springer. The shirt promotes the R statistical language and environment and contains the slogan "First plot; then model." I'm thinking that even this mantra needs to be rewritten because it ignores the most fundamental component of data analysis, one that is fostered and improved by graphic presentations of data: *thinking.* The slogan should read

First <u>Think</u>. Then Plot. Then <u>Think</u> again.

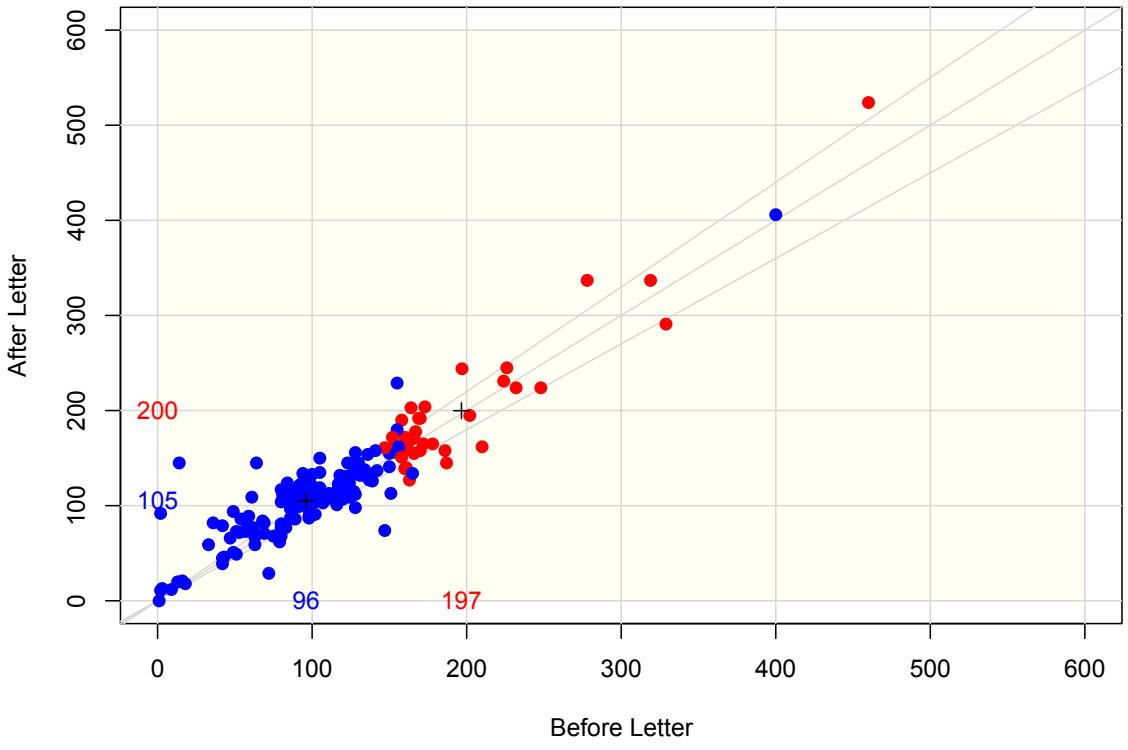
When making comparisons, we often seek to make a comparison of distributions. As such, we need to worry about more than just the mean or the median. Plotting the raw data, the data atoms, allows us to understand the entire distribution.

A researcher approached the Biostatistics Clinic with a question regarding the number of narcotics prescriptions certain prescribers were writing. In the past, some of the prescribers had received letters from the State telling them that Big Brother knew that they had been writing lots of narcotics scripts. Some prescribers got these letters, some did not. What was the effect of this intervention (the letter) on the writing of narcotics prescriptions?

The data consisted of a prescriber id, a number representing the number of narcotics prescriptions *before* the letter went out, a number of representing the number of narcotics prescriptions *after* the letter went out, and a flag to indicate whether or not that prescriber was on the list of those who were to get a letter.

The determiner for getting a letter was based on the rate of narcotics prescriptions, not on a random assignment. Therefore, getting a letter was a consequence of number of narcotics prescriptions: more prescriptions means you got a letter.

The data then are bivariate in nature, with a 'before' component and an 'after' component. The first plot of the data was simply that scatter plot.



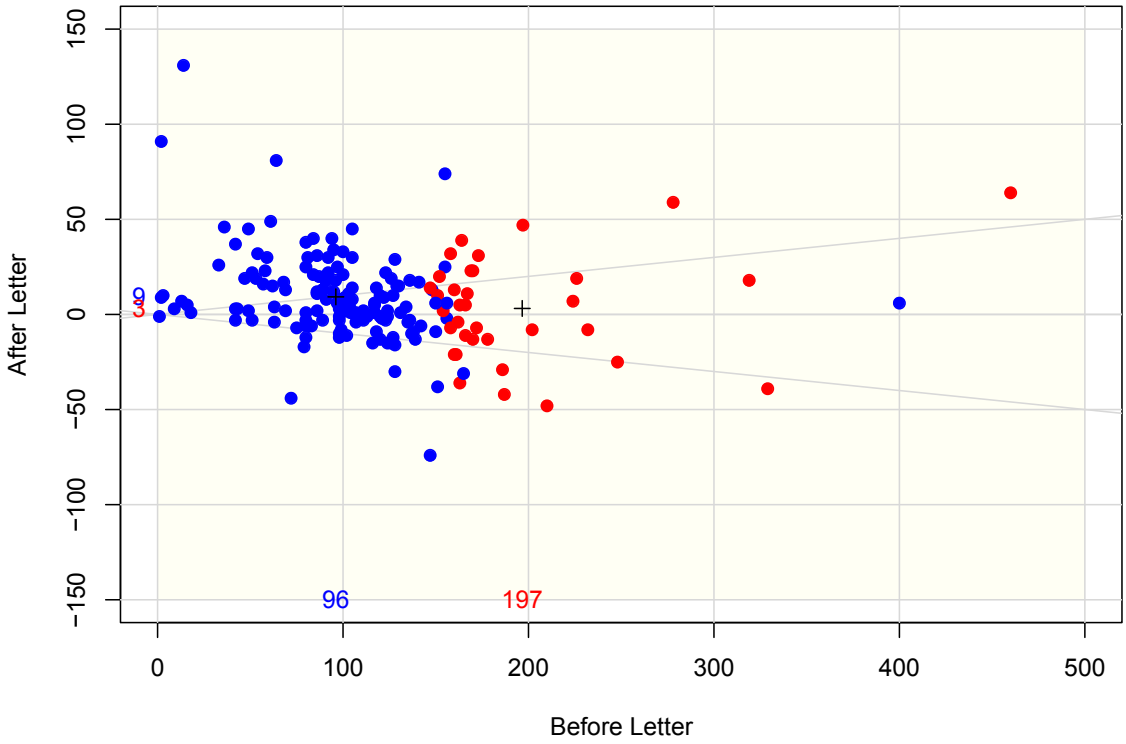
We have here the 'before' and 'after' values as shown by the colored dots. The slanty lines show equality and 10% increases and decreases. If our data display here is the model, then we are implying that the number of narcotics scripts written after is a function of number of narcotics scripts written before, since we put cause on the horizontal axis and effect on the vertical axis. **Plot cause versus effect.**

We have also lightly shaded the plotting region so as to soften the data against the background. The red and blue dots contain an additional piece of information. The red dots are those prescribers who were on the list to get the letter, while the blue dots did not get a letter. The marginal means are shown labelled in red and blue and their intersections are marked with the cross-hairs. Immediately, then, we can see that the threshold for getting a letter appears to be in the

neighborhood of 150, although that threshold is not absolute. We also see a 'no letter' blue dot prescriber at 400.

But the question of interest, at least to those who were sending out the letters, was related to *change* in number of scripts, so we can change the plot to look at 'change' as a function of 'before'.

Here we have change as a function of before. This allows us to see the smaller differences better across the range of the before values. The mean changes are shown now along the vertical axis, as 3 for the red group and 9 for the blue group. The cross-hairs still mark their intersections.



The slanty lines have become more horizontal, with the plus and minus 10% lines meeting at, and diverging from, the origin.

The outliers near zero (increases in the neighborhood of 100 prescriptions!) show some monstrous gains that may or may not have been as evident in the previous plot. A drop of nearly 75 prescriptions in the blue group is visible too.

Statisticians are certainly tempted in situations such as these to fit regression models, to estimate the mean value of change for a given value of before. We can see from the plot that the mean change is an increase of 9 for the blues and an increase of 3 for the reds. Did the letters have an effect? Any competent statistician can get his or her computer to fit the regression and determine if 9 is significantly greater than 3. But is that the question that we want to ask here?

The mean, as we have mentioned earlier, is a sufficient statistic for lots of well-behaved distributions, hence its appeal as a theoretical quantity for examination. But real life looks like the data above. Is the mean relevant here? Again, *mean if, and only if, total* must carry the day here, implying that we should be interested in comparing, between the red and blue groups, the mean increase in narcotics prescriptions if and only if we are interested in the comparing the total increase in narcotics prescriptions between the red and blue groups.

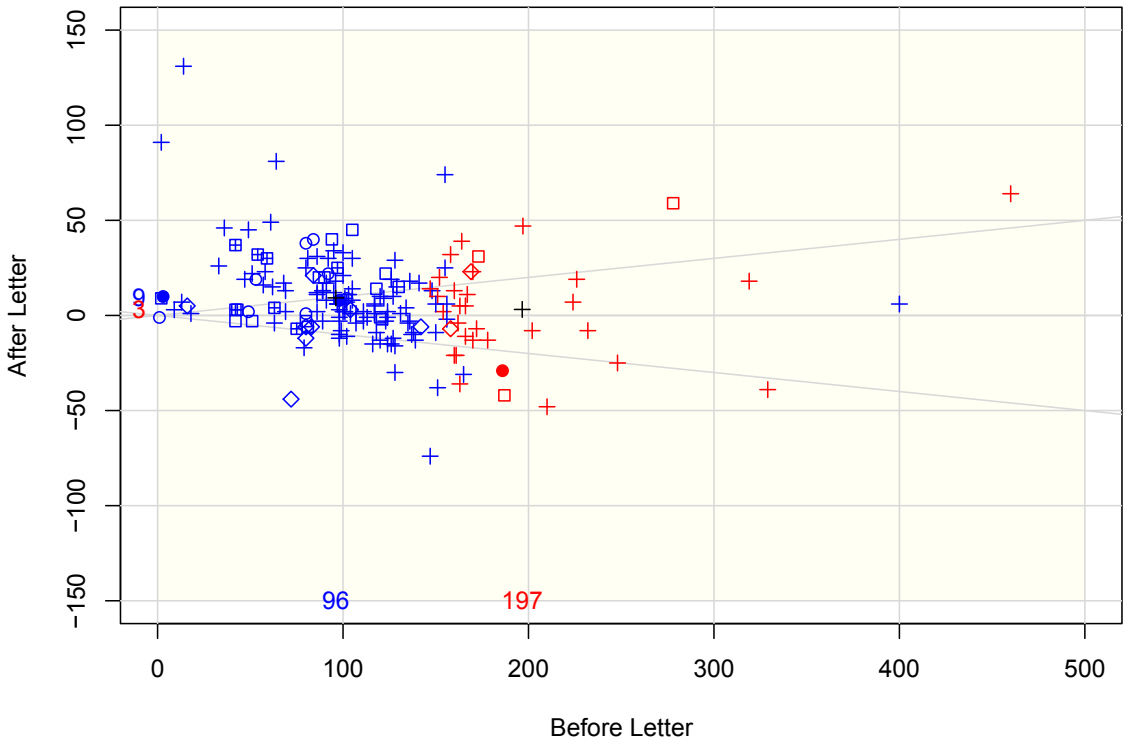
So herein lies the rub with a study such as this one: the red group (the treated group) is different from the blue group (the control group) on the basis of the before level; those prescribers with higher before values tended to get the letter.

Did the letter keep some red prescribers from writing illicit prescriptions and flying off the charts, like those blue prescribers who went from near zero to 100? Can we be interested in the mean or total value for those who got letters when we know that letter-writing was not randomized?

Look at the data; what do you *think*?!?

But the point, as regards the presentation of data, is that there are more data in this data set, a variable I have not discussed yet. This additional variable maps each prescriber into a group of like prescribers. These different prescriber types, as presented in the data set, are: MD, PA, RN, LPN, NP, and DO.

The tendency is often to create six different symbols, say, crosses, hashes, open dots, closed dots, and whatnot, and then use those symbols to depict the different levels of presciber type.



Well, this certainly doesn't make things any easier. Regardless of the fact that I failed to provide a legend to tell the viewer which symbol shows which group, the busyness of the overlapping symbols makes the plot no more valuable than the previous plot where all the symbols are the same. So we will need to do things differently.

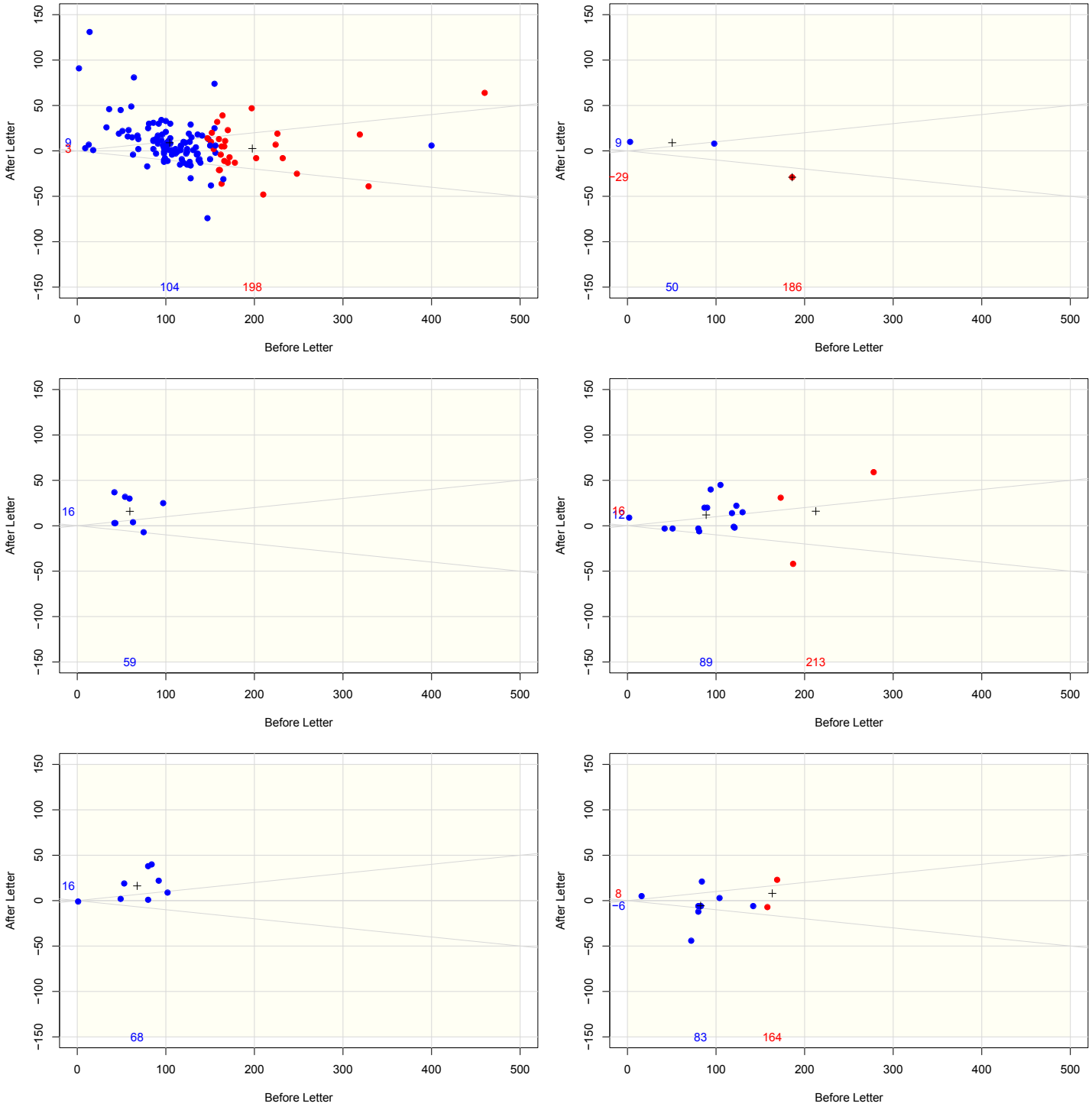
Generally, using different plotting symbols, whether they be based on shape or color or both, to denote sources of variation does not work well unless there is dramatic spatial separation between the groups under consideration. That is why the color-coding used to reflect the letter-getting status worked but the symbol-shape-coding used to reflect prescriber group failed. **Typically, color ought be used for response variables, not design variables — but not always!**

Recall that the prescriber type is a source of variation, a partition of the data set, a way to generate component parts that contribute to the whole. The overall plot then is a plot of a mixture distribution, where the different prescriber types each contribute to the distribution.

We should be able then to break out these components into small multiples and see the individual components. If we assign a consistent scale across the components and revert to the simple red and blue dots, we get the following six

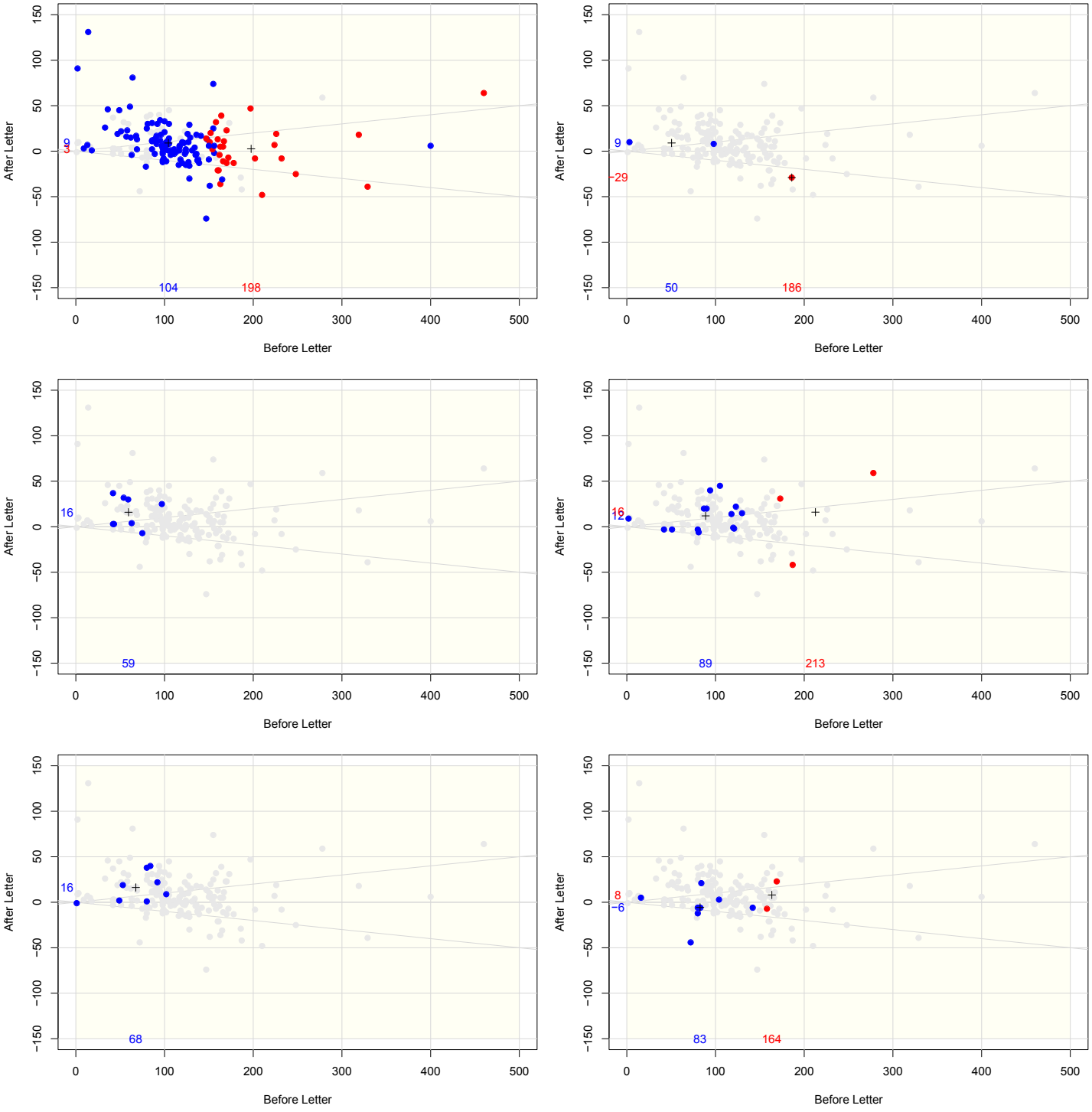
components.

Here the small multiples concept reveals its value, allowing us to see all the prescriber groups in comparison to one another; still, it would be nice to be able to compare them all to the composite mixture distribution. We could do this by adding another panel to the plot or, better yet, by embedding the mixture in the individual panels. We will add the missing points to each panel in a light grey, essentially highlighting each prescriber group and sending the comparators to the background, thus allowing us to make comparisons to the distribution as a whole.



The point here is that we understand individuals by way of reference to a distribution of like individuals. What appears at the outset to be simply a distribution of prescribers is, upon closer examination, a mixture distribution. Keeping the relevant background distribution in play aids in make comparisons.

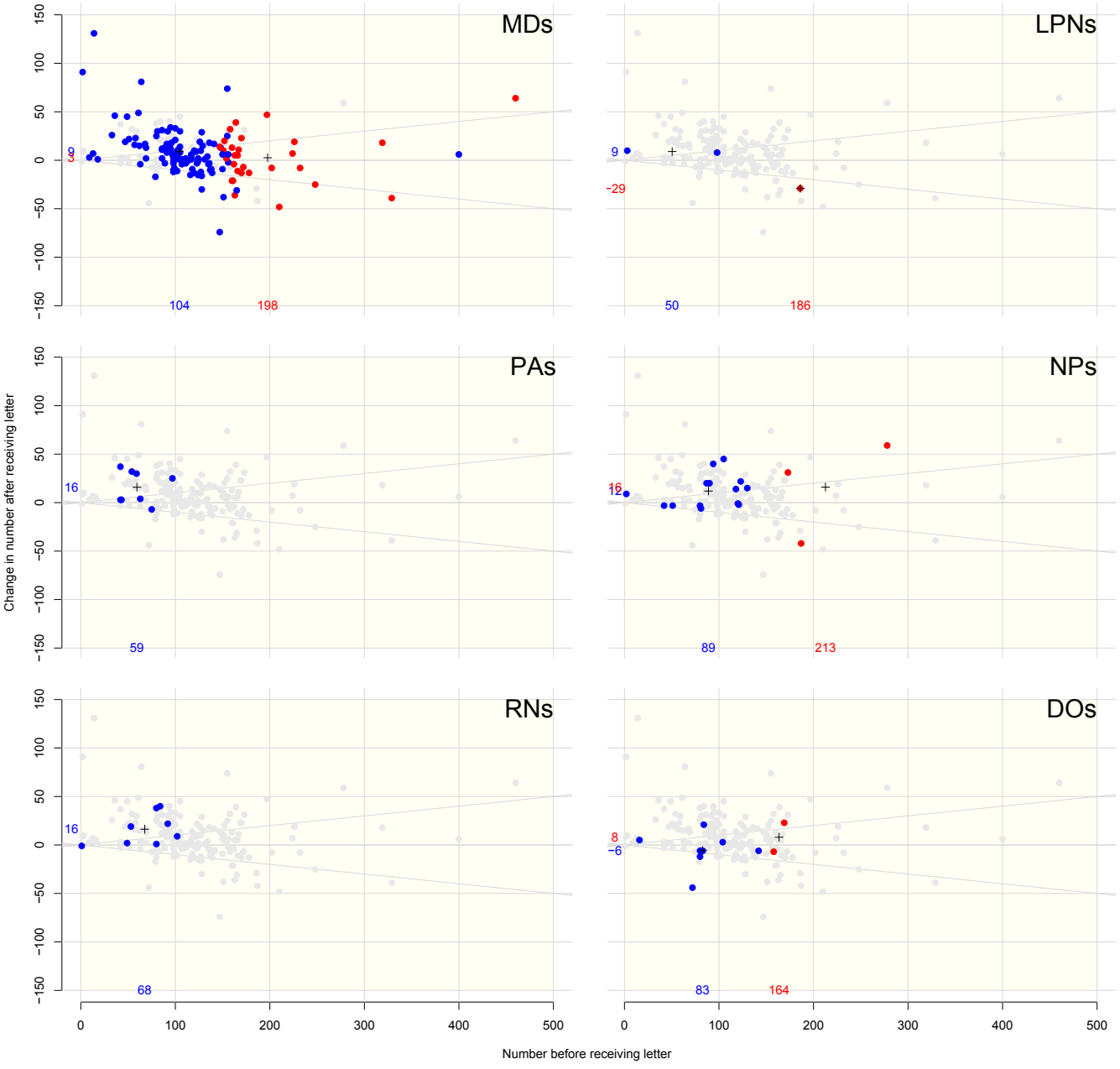
We might call this concept "you are here", with the idea being that we show a distribution of responses, of which one point, or collection of points, is highlighted. The improvement in understanding comes from showing these collections in a small multiples format and showing the individuals under consideration against a background of all others, thereby highlighting additional sources of variation.



This layering and separation, in conjuction with the small multiples splits, gives us more than either technique alone.

We can clean up the panels and integrate them into one complete plot. Removing redundant labels on both the horizontal and vertical axes, creating a proper label for the vertical 'change' axis, and adding labels to the six panels so as to be able to tell which group is which leads to the plot below.

Of course, this is not a finished product; some work remains undone. The labels are simply the labels that were found in the original data set. We must contact the researcher if we are going to get the precise definitions although some rational guesses might yield Medical Doctor, Physician's Assistant, Registered Nurse, Licensed Practical Nurse, Nurse Practitioner, and Doctor of Osteopathic Medicine. A proper title will need to be generated and we will need to get exact



details about what these counts actually are. Are they totals over a year or averages or what? What exactly constitutes a narcotic prescription? And who are these prescribers and how did they get into this data set?

Further, the grid lines and the slanty reference lines may or may not be overbearing, depending upon the final printing or screen view of these plots. An issue with variation in printing is that what one sees on the screen is not always what comes out on the paper. Finding a solution based on the final medium of presentation is never easy, certainly when the final medium is not known.

The small multiples you-are-here plot shows us that the MDs are the dominant group in the types of prescribers. They also include almost all the prescribers who were on the list to get a letter. One LPN, three NPs, and two DOs fill out the collection of letter-getters. Notice that there are no PAs nor RNs amongst those being tracked by Big Brother.

Our grid of prescriber groups is artificial in its 3-by-2 construction; perhaps a rearrangement to 1-by-6 or 6-by-1, with a sensible ordering across those groups, would help the understanding. (Our 3-by-2 grid is one of convenience to fit the plots onto the page.) More discussions with the eventual users of the plot will be needed to answer these questions.

You-are-here can be used whenever responses are complicated by issues with overlapping plots. Sharon Phillips, a colleague at Vanderbilt University Medical Center, offers up a draft you-are-here plot showing the relationship between resting metabolic rate (RMR) in kilocalories per day and body mass in 22 obese patients in a weight loss study as another example.

The data are serial readings across time. There are three time points (baseline, 6 months, 12 months) for each patient, each of which indexes a bivariate reading of resting metabolic rate and body mass. The researchers are interested in looking at how body mass, which is dropping over time in a weight loss study, impacts resting metabolic rate.

Body mass can be decomposed (analyzed!) into fat mass and fat-free, or lean, mass, so that the sum of the two is the overall body mass▸.

▸Technically, this isn't true. Body mass is composed of fat mass, fat-free mass, and the mass of the head, so technically, fat mass plus fat-free mass equals body mass minus head mass. For purposes of this example, however, we will work as if fat mass and fat-free mass sum to the body mass. Essentially, we are working with the assumptions that all these patients do not have heads!

Each patient, then, lives somewhere in the bivariate plane, and follows a path across the relationship between the two variables. As a group, the patients constitute a distribution of such paths. A spaghetti plot of the data would do little to help us understand the nuances of the data atoms, although such a plot might demonstrate some overall patterns present in the data.

Change from baseline to 6 months is shown with the red arrow; change from 6 months to 12 months is shown with the blue arrow. Patients with missing data points will be missing some changes; as such, some patients only show raw data points: baseline in red, 6 months in blue, 12 months in black.

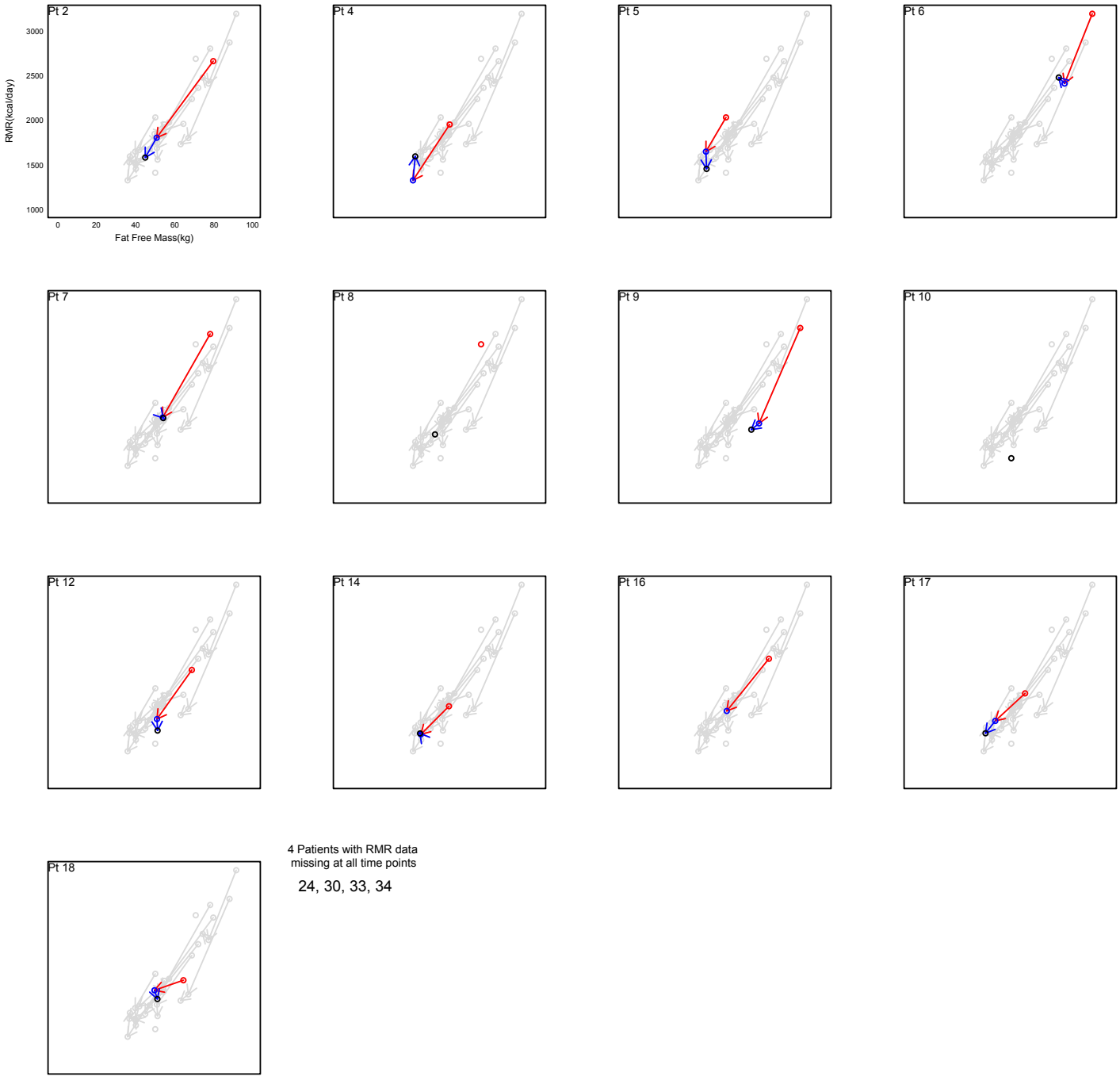
There are two sets of plots. The first (on the left) shows resting metabolic rate as a function of the fat-free mass; the second (on the right) shows RMR as a function of the fat mass. The two sets of plots then show two different models for RMR, one using lean mass as a covariate, one using fat mass as a covariate.

The two models show different relationships. Looking at the left set of plots, the

lean mass provides a consistent, positive correlation between mass and RMR. Patients in general show much more dramatic changes between baseline and 6 months than they do from 6 to 12 months, as evidenced by the typically longer red arrows compared to the blue. Patient 4 (top row, second from left) is somewhat of an outlier, as this patient's blue arrow shows a dramatic uptick in the second half of the six months. Other than that, the left plot shows little variation in the relationship between mass and RMR.

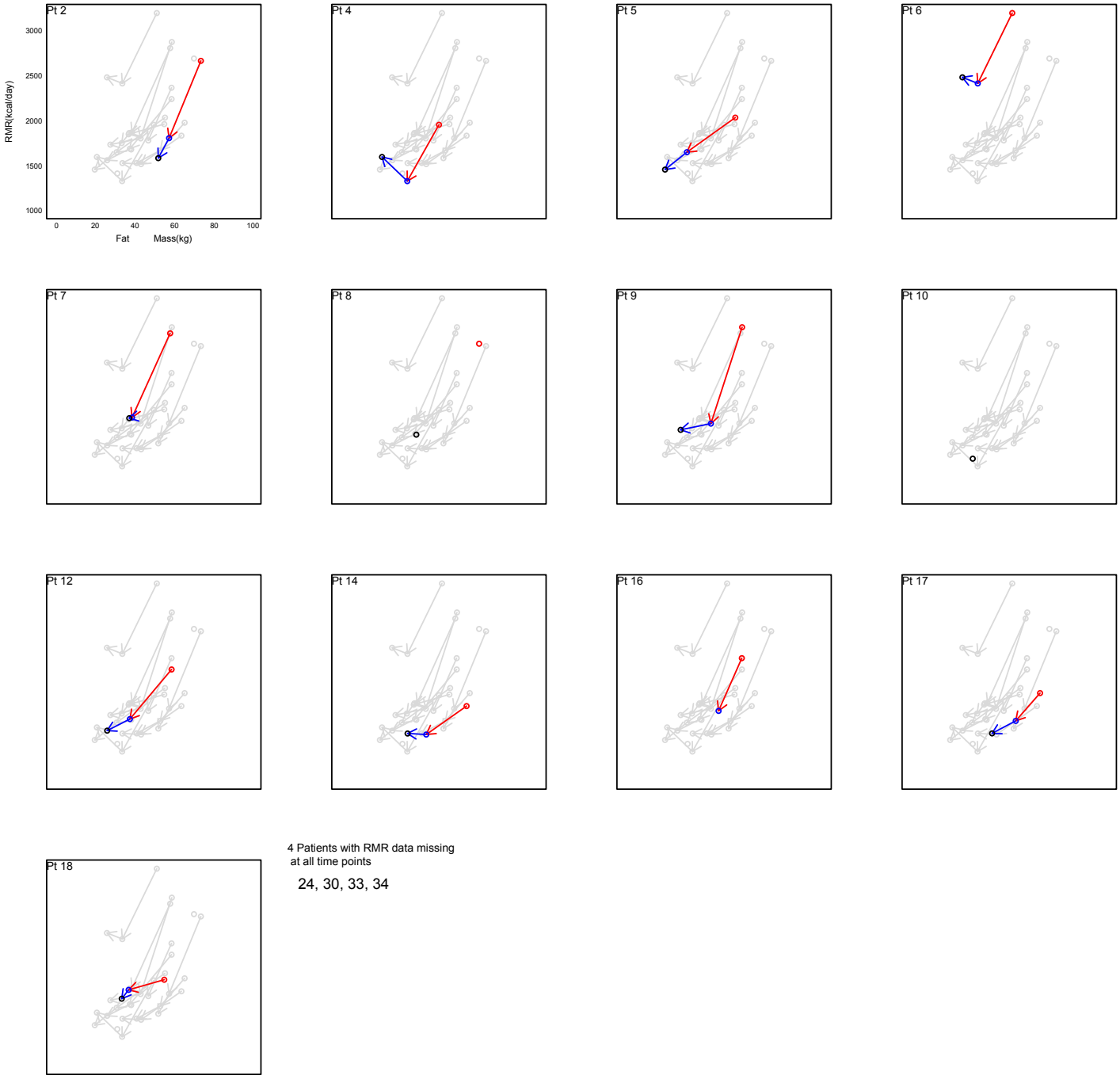
There are four patients missing all their values, as listed at the bottom of the plot. Patients 8, 10, and 16 are missing at least one time point.

If this were a final version of the plot, we might soften the boundaries on all the



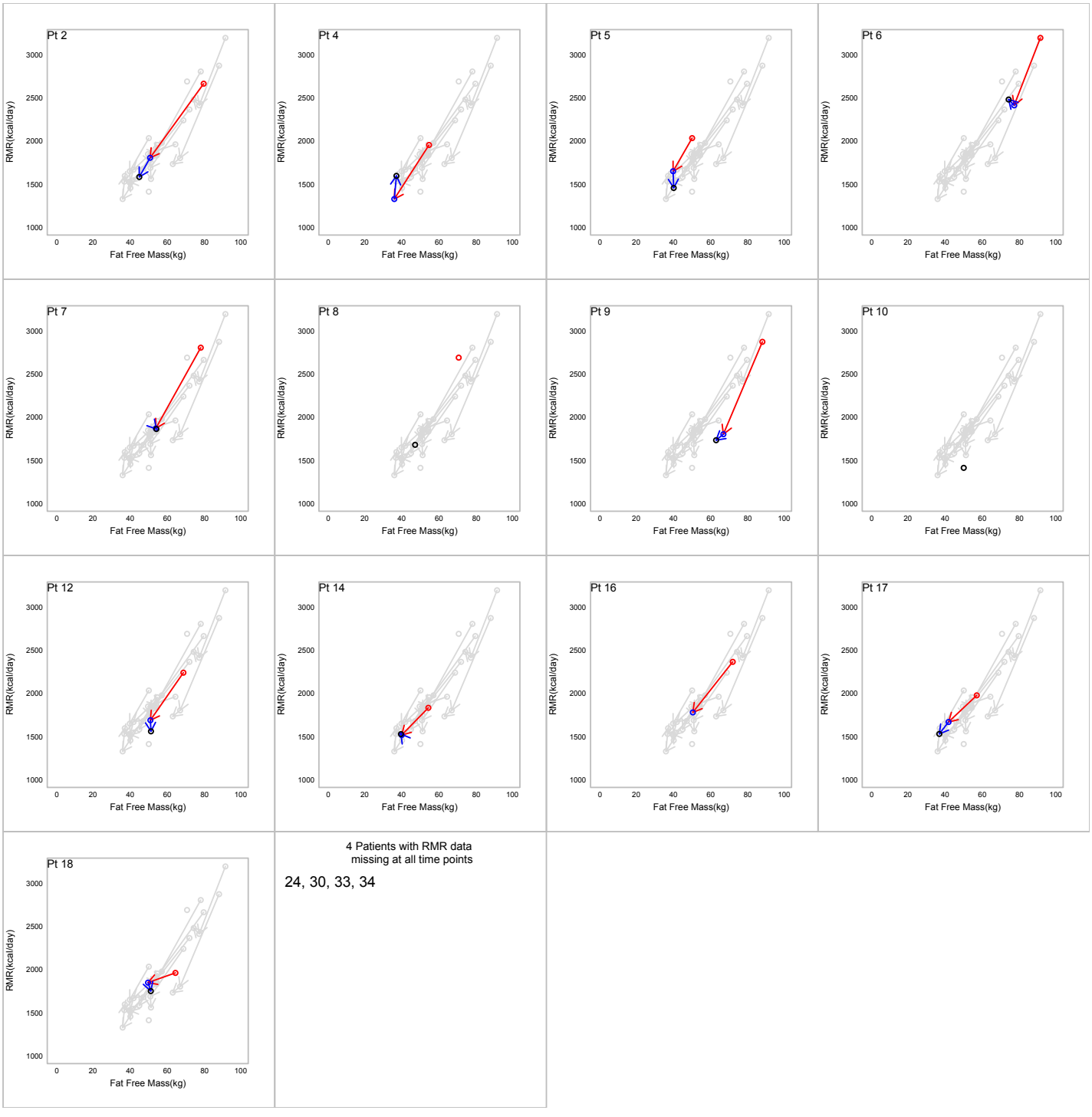
plots, or put them on a soft beige or off-white, and change the fonts to something more readable. But the lesson here focuses on the nature of the you-are-here mentality: **we understand the individual responses by comparing them to a distribution of like individuals**.

The plot on the right shows fat mass as a predictor of RMR. This plot shows dramatically greater variation both across patients, as exposed by the disparity in the grey background distribution, and within patients, as exposed by the stronger changes between the pre- and post-6-month paths. Putting the collection of plots on facing pages allows us to compare individual patients' mass-type/RMR relationships.



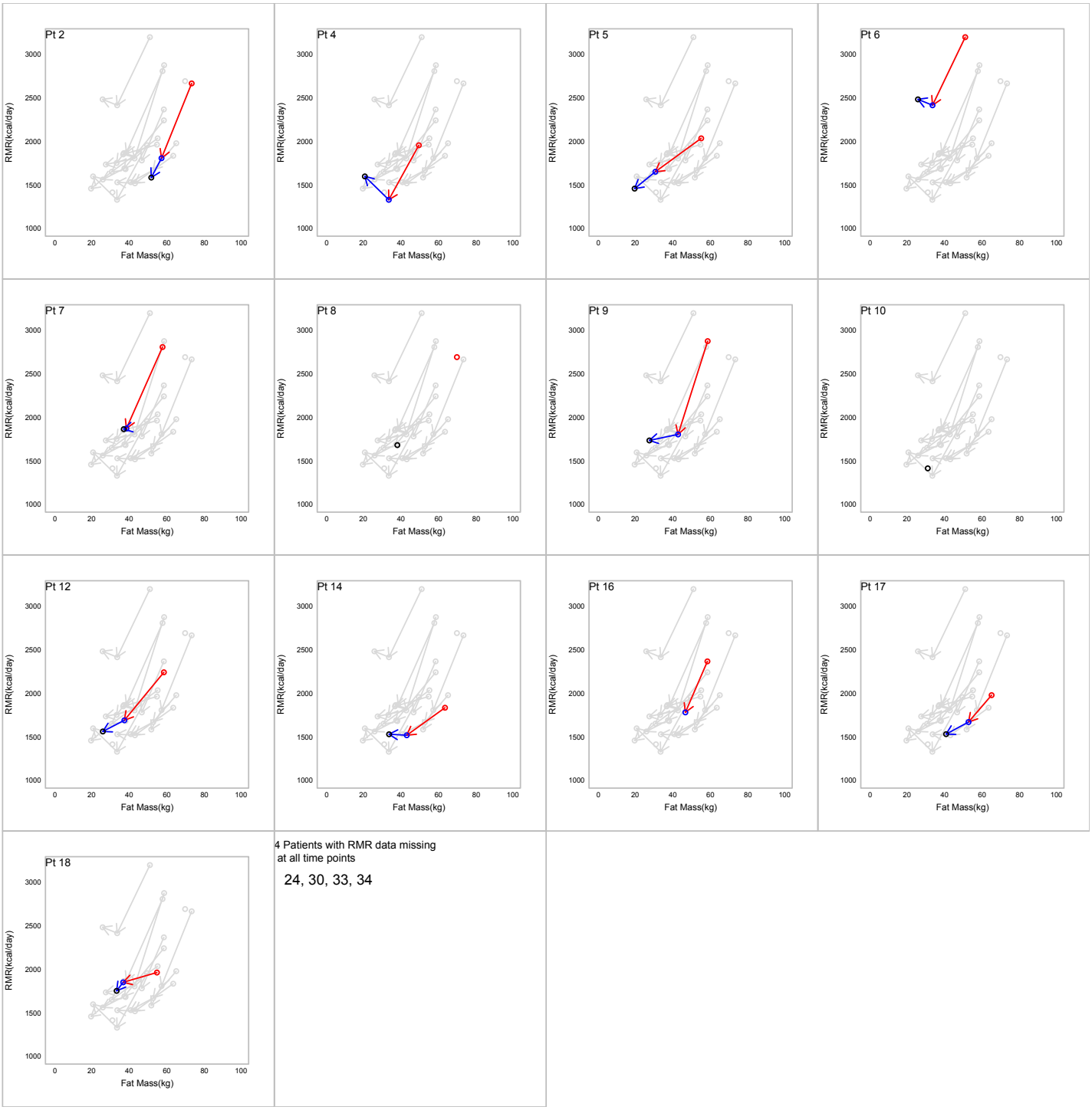
The explanation as to why these patients show such disparity when examined on the basis of fat mass and so little variation when using fat-free mass as a covariate centers on the understanding that the patients lean mass structures were highly consistent; it is when looking at these patinets' fat masses that we note the differences. The lean mass patients individually enclosed within the fat mass patients were very much the same!

And a note of color choices: **Avoid red and green as comparators. Color-blindness is too common to justify the red-green contrasts. And avoid garish color schemes; try to use subtler colors.** Remember, the goal is the exposition of the distribution of the data. Don't drown out the data with colors that are too loud.



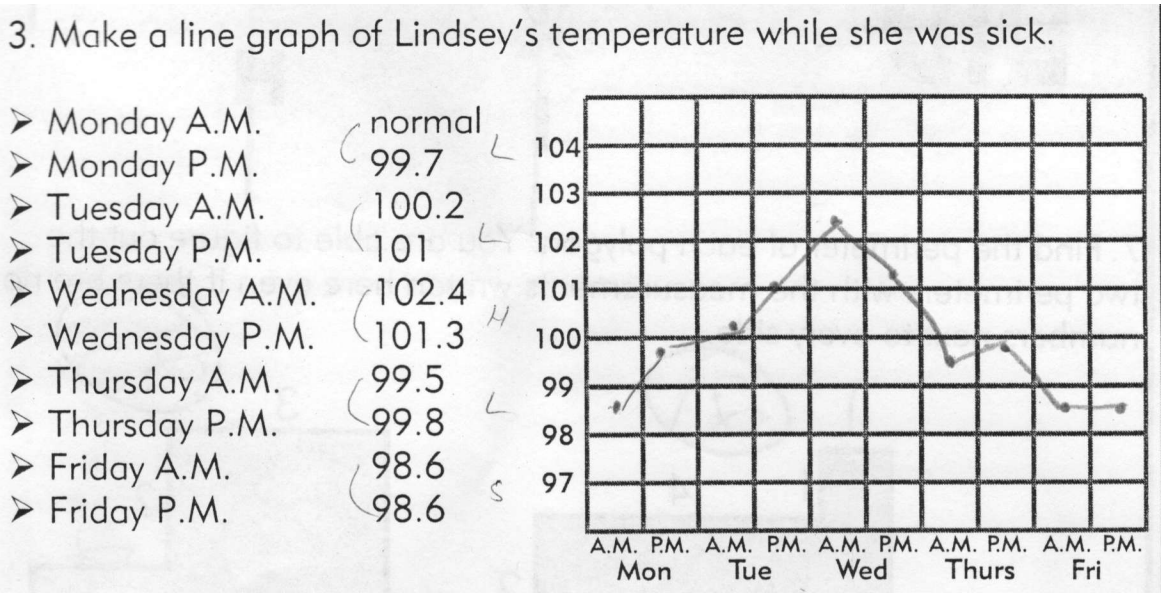
So, why are these two plots here? After Sharon provided me with the plots on the previous spread, she sent me these with the explanation that her colleagues agreed that softening the boundaries was a good idea but they also wanted the scales placed on each of the small multiples; they were worried that the readers would have difficulty making the comparisons without them. <sigh.>

Compare these pages to the previous ones Sharon produced: are these focused on the data elements or on the residual clutter of the redundant design? Note how the data are now visually crowded out, how the repetitive scale labels take center stage, and how our ability to *understand the individual responses by comparing them to a distribution of like individuals* is markedly limited.



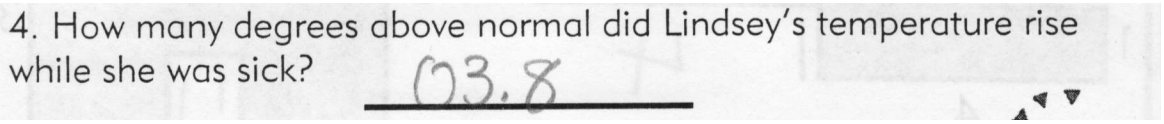
On drawing the right plot to answer the question, and asking the right question: We should not expect every plot, regardless of how fancifully drawn, to answer all the questions we can pose to the data it presents. **Data presentation layouts and designs should be driven by intended use.** (Note the differences between a list of phone numbers sorted by name and a list of phone numbers sorted by number.)

Here are some more wonderful homework data, and some thoughts on what the graphs are and are not. The original assignment:



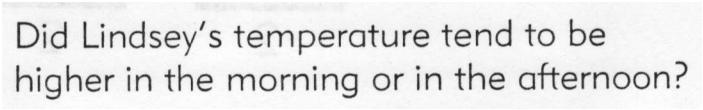
So our student again has dutifully produced the "line graph" and we see a time course of Lindsey's illness. Note that Lindsey's temperature peaked on Wednesday A.M. at 102.4. In the interest of time, we will assume that the units are degrees Fahrenheit. Note also that this time series (even with all the foibles surrounding the heavy gridlines, and labeling *between* the lines (in the spaces) instead of *on* the lines, and the temperature value of "normal" on Monday A.M., whatever that means) provides simply an accounting of Lindsey's fever trends. To do any analysis of these data, we will need additional information, such as when she went to a doctor, if at all, what medications or treatments she received, if any, and so on. There is variation in the data; what caused that variation? Our model cannot be "time caused the variation": **time series make fine accounting but poor scientific models**.

The question continues, however:



Of course, here we see the issue with the undefined "normal" reading on Monday A.M.: how can we compute distance from normal if normal is not defined? This is where parents become valuable in homework, as most parents can help and reply that normal is 98.6. Simple subtraction yields the distance between 102.4 (the maximum value during the sickness; *why should we be using the maximum???*) and 'normal': 3.8.

The final question demonstrates that those posing the questions do not understand that the line graph does not answer all the questions:



Daddy, what's the answer to this one?

Goodness, I'm not sure. What do you think?

I don't know. All I can see is that it is highest on Wednesday A.M. Should I put "in the morning"?

So the answer has to do with defining "tend to be higher". We could compute the means for morning and for afternoon. Or we could use the maximum, or the minimum, or any other summary measure.

What we have, though, is a distribution of temperatures in the morning and a distribution of temperatures in the afternoon. We are asked to compare these two distributions.

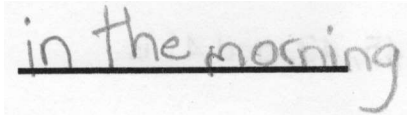
We might be tempted, as is evidenced by the doodling on the original question, to reduce the data to matched pairs. Note the penciled-in hooks connecting the morning and afternoon data on each day and the Lower (in the morning), Higher, and Same designations next to each pair: three L's, one H, and one S. So the temperature tended to be higher in the afternoon (lower in the morning)?

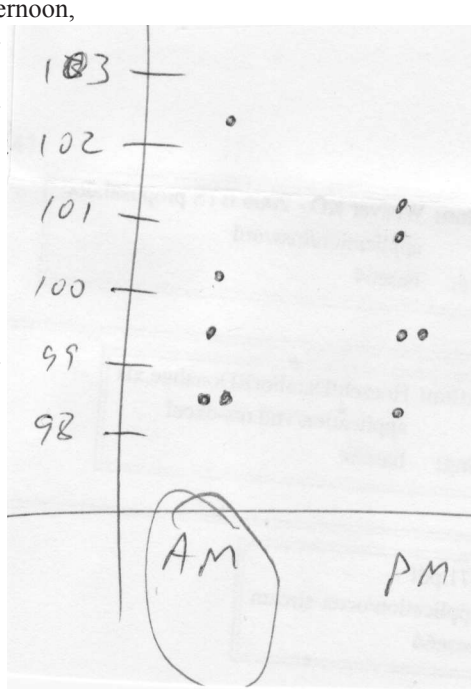
But why the arbitrary grouping within named days? Why not match Monday afternoon with Tuesday morning, and Tuesday afternoon with Wednesday morning, and so on? If we do things that way, we get two mornings higher and two afternoons higher …

What we really need is a straight up comparison between morning and afternoon, since that's what the question asked. Using highly technologically advanced tools known as paper and pencil, we get a plot with morning and afternoon as the sources of variation; time has been taken out of consideration. (For the sticklers, we will also leave out autocorrelation.)

So I asked her which group looks higher and she circled the AM group. Why? Because it had the highest point. I thought, so be it, I'm not going to bias her understanding of data by telling her that people like the mean. Why mess up a good thing?

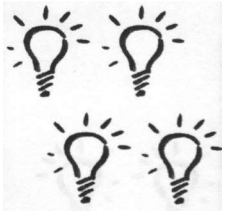
We could argue and discuss the merits and perils of using different summaries to describe those distributions but at least we have a picture that compares the distributions, at least we can see about what we would be arguing. She wrote:





Some of the clip-art on the original homework page. How this helps the students understand graphics, or mathematics, or medicine, or data analysis, or statistics is beyond me. What it tells me is that the writers of this homework piece are not particularly serious about what they are teaching our students.

Readers might be interested to note that the question on drawing the line graph was considered to be "four light bulbs" in difficulty:



while the interpretation question that so vexed our student and her father was considered to be "two light bulbs" in difficulty:

Go figure.

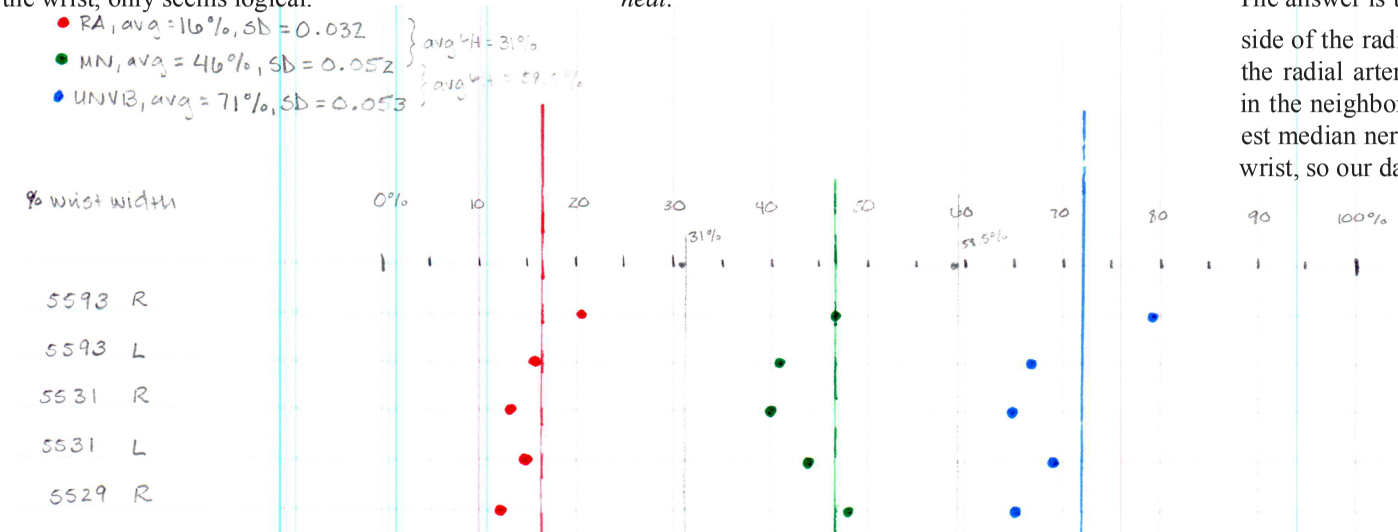
Your wrist is a miracle of biological development but modern societal practices, namely computer keyboard use, have produced an ailment that can create painful problems in the wrist. The National Institutes of Health describes carpal tunnel syndrome "as a compression of the median nerve within the carpal tunnel, a narrow, rigid passageway of ligament and bones at the base of the hand". Treatment of carpal tunnel syndrome sometimes involves injection of steroids directly into the carpal tunnel in an effort to reduce inflammation and hence compression of the median nerve.

A medical student working with a hand surgeon came to me with data. She and the surgeon had investigated the location of several key landmarks (the radial artery, the median nerve, and the ulnar nerve) in cadaver's wrists. They said they wanted to find the average location of these anatomical components so they could do a better job with injection treatments for carpal tunnel syndrome: if they could make a good guess about where these three components were located, they could avoid them when injecting the steroids. One seeks to avoid injections directly into the radial artery, an artery on the lateral (thumb) side of the wrist where one takes a pulse. Avoiding directly injecting into the median nerve, which runs down the center of the wrist, and the ulnar nerve, located on the pinky (ulnar, medial) side of the wrist, only seems logical.

After some brief anatomy training and a description of the procedure, we were able to determine that we did not seek to find the places where these landmarks *were* located; rather, we sought to be able to find a place in the wrist where these landmarks were *not* located. We were not looking for a measure of central tendency, we were seeking the edges of distributions.

The researchers had examined the wrists of 18 cadavers, so they had readings on 36 wrists. We surmised that there ought be correlation within a subject with regard to the location of the landmarks. The locations of the landmarks were given in percent distance across the wrist. Zero was set to be the radial side of the wrist, one-hundred was set to be the ulnar side of the wrist. If the wrist was, say, 17.5 cm wide and the radial artery, median nerve, and ulnar nerve were found at 2.8, 8.4, and 14.0 cm from the radial side of the wrist, then the data for that wrist would be the triple (16%, 48%, 80%). They had already computed means of these locations.

The medical student was eager to understand the data, so we developed a plan for her to investigate the data, not just the summaries. We discussed that she was interested in exposing the *distribution* of these landmarks. With her having only paper and pencil, I asked her to plot each wrist as simple dots on a line, and label each line with the wrist identifier. *Take your time; nice and neat.*



Within 48 hours I received an estatic phone call: could I see her right away? She had completed the plot and needed to show me!

A scan of her plot, complete with the mean locations of the radial artery, median nerve, and ulnar nerve, is shown on this page. What she was so eager to show me was the distribution of the locations of the landmarks including, most importantly, the minimum and maximum values for each location. We can see the strong correlation of the three components: when the radial artery is close to the thumb, the median and ulnar nerve tend to be close too. The mean lines she drew show the center, but if you look closely at the scale along the top, you will see the minimum value (31%) and the maximum value (59.5%) for the median nerve, the most crucial component when trying to miss the valuable landmarks when injecting.

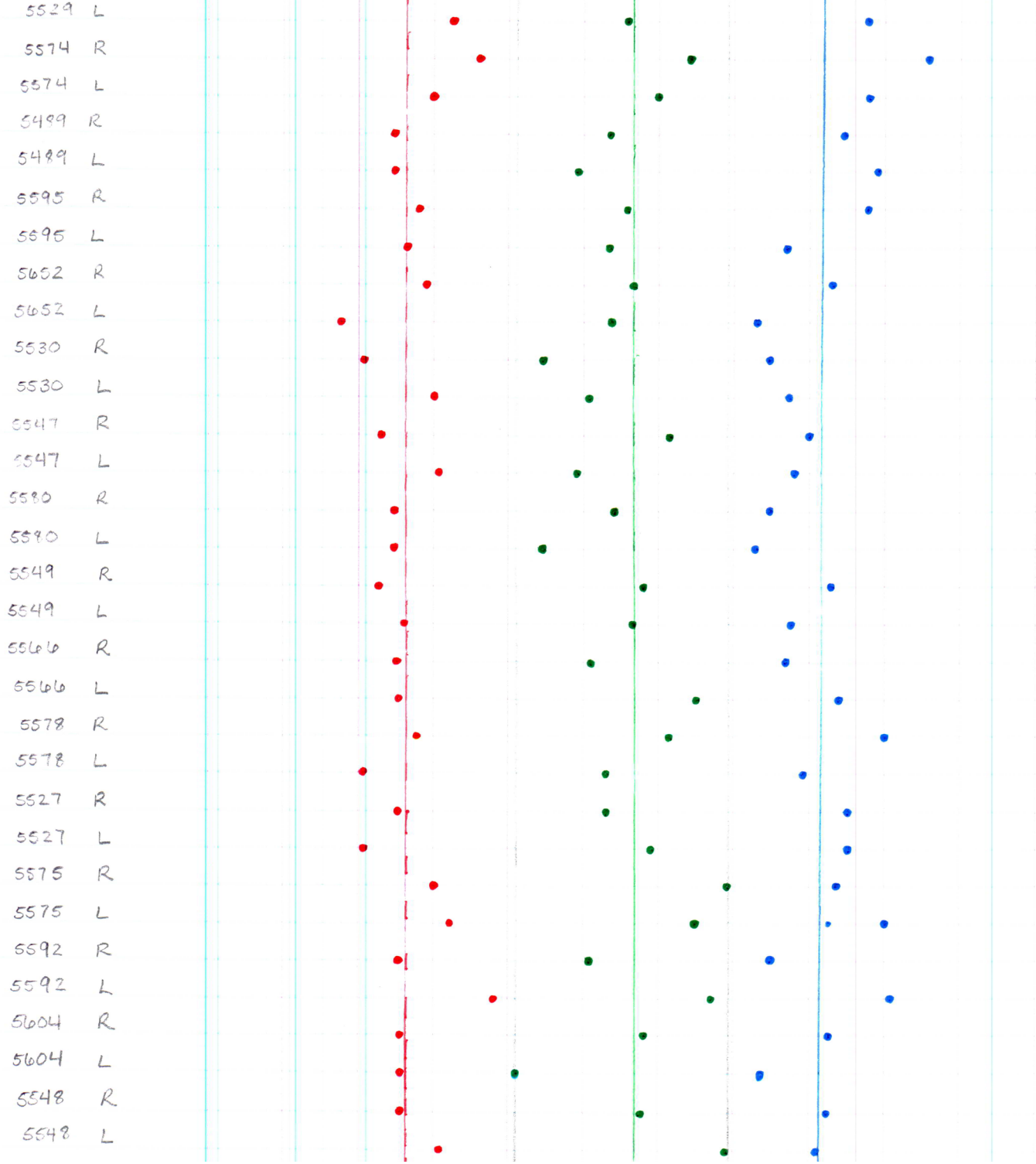
We see that the variation of the radial artery (in red, on the left) is drastically smaller than that for the median or ulnar nerves.

Based on her plot, where can we do the injections so as to minimize the chance of hitting something important? Surely we don't want to inject at the means, as we are likely to strike something bloody or painful.

The answer is that we ought to inject just to the medial side of the radial artery. The most medial reading for the radial artery is cadaver 5592's left wrist, a value in the neighbor hood of just less than 30%. The lowest median nerve value is 31% for cadaver 5604's left wrist, so our data show a gap here.

There is another gap in the region between the median and ulnar nerves. The largest median nerve value is just shy of 60%, the lowest ulnar nerve value appears to be just above 60%. Perhaps we can use this gap too. What a wonderful plot of the variation of anatomy across individuals!

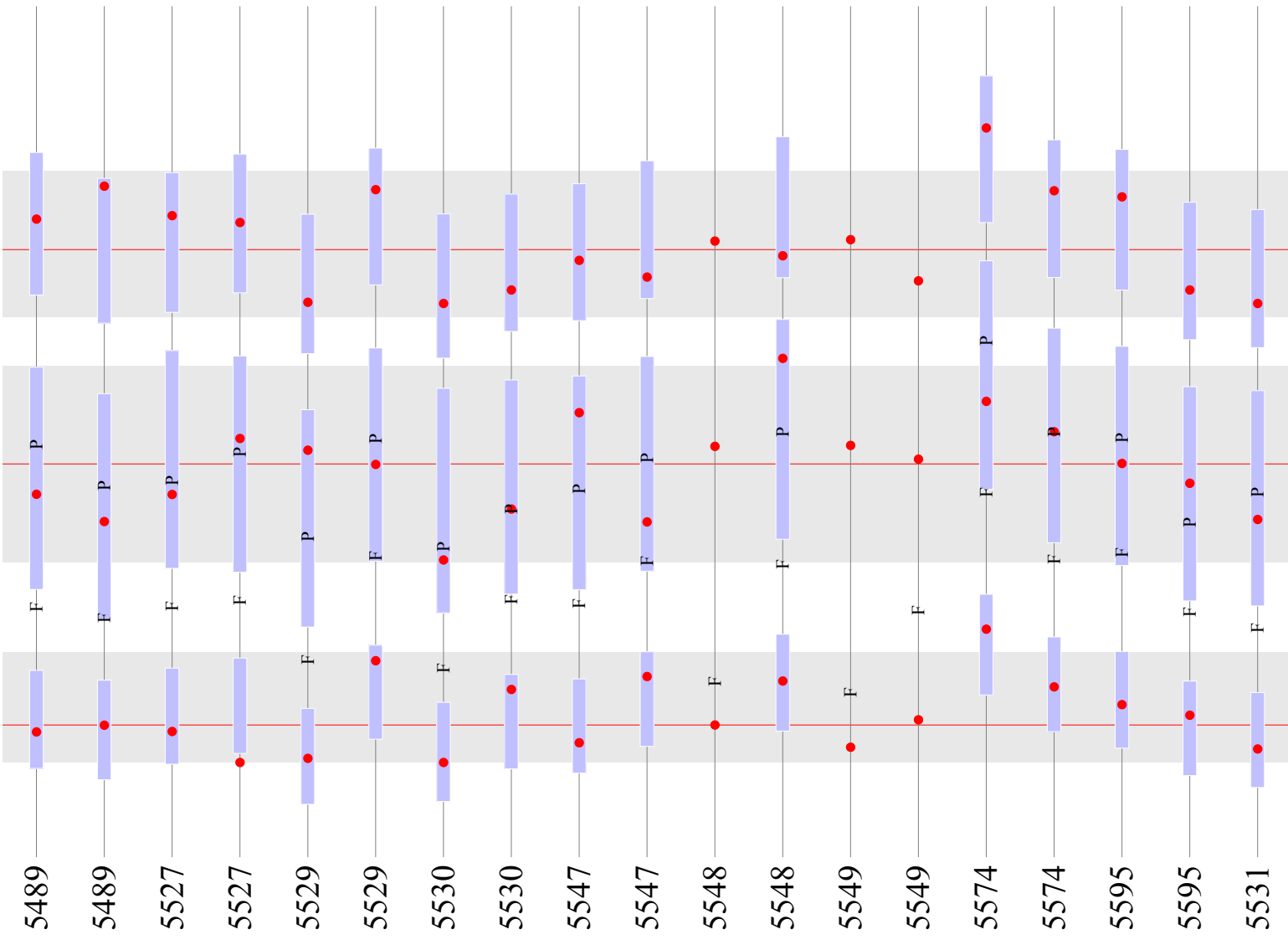
But the researchers needed something



fancier, they said. They couldn't just put this plot in a paper or on a poster, they said. (Why not?, I countered.) And we need statistics, they said. Arguing against them, I didn't carry the day so I agreed to make a computer version of this hand-drawn plot. I agreed to add some fancy statistics.

The data set actually contained a few other measurements of value. The plamaris longus and flexor carpi radialis tendons are tendons that are easily identified on external examination. That is, they can easily be spotted just by looking at the wrist. Perhaps determining the locations of these easily observable landmarks could help locate the unseen landmarks that needed to be avoided.

I agreed to model the locations of the three points of interest and use the model estimates to help identify the places to avoid. We would then augment the pen-and-ink plot with a computer drawn image that would show us the locations of the Big Three (radial artery, median nerve, ulnar nerve), the point estimates of these three based on the model that included cadaver sex, hand (left or right), wrist width, and locations of the palmaris longus and flexor carpi radialis tendons, and estimated upper and lower 95-percent confidence limits for those estimates. Note that all these positions and classifications are determinable with external examination so, if one were so inclined, one could build a fancy look-up table or fancy on-line or handheld scheme using this model to output the places to avoid. I also agreed to plot the naive 5th, 50th, and 95th percentile points as computed from the raw values of the Big Three. This plot appears on the next spread.



This plot shows the individual data, along with all the summary statistics described on the previous page. The radial artery is on the left, so this plot is correct for the left wrist, as viewed with your hand up and your palm facing you, and a mirror for your right wrist.

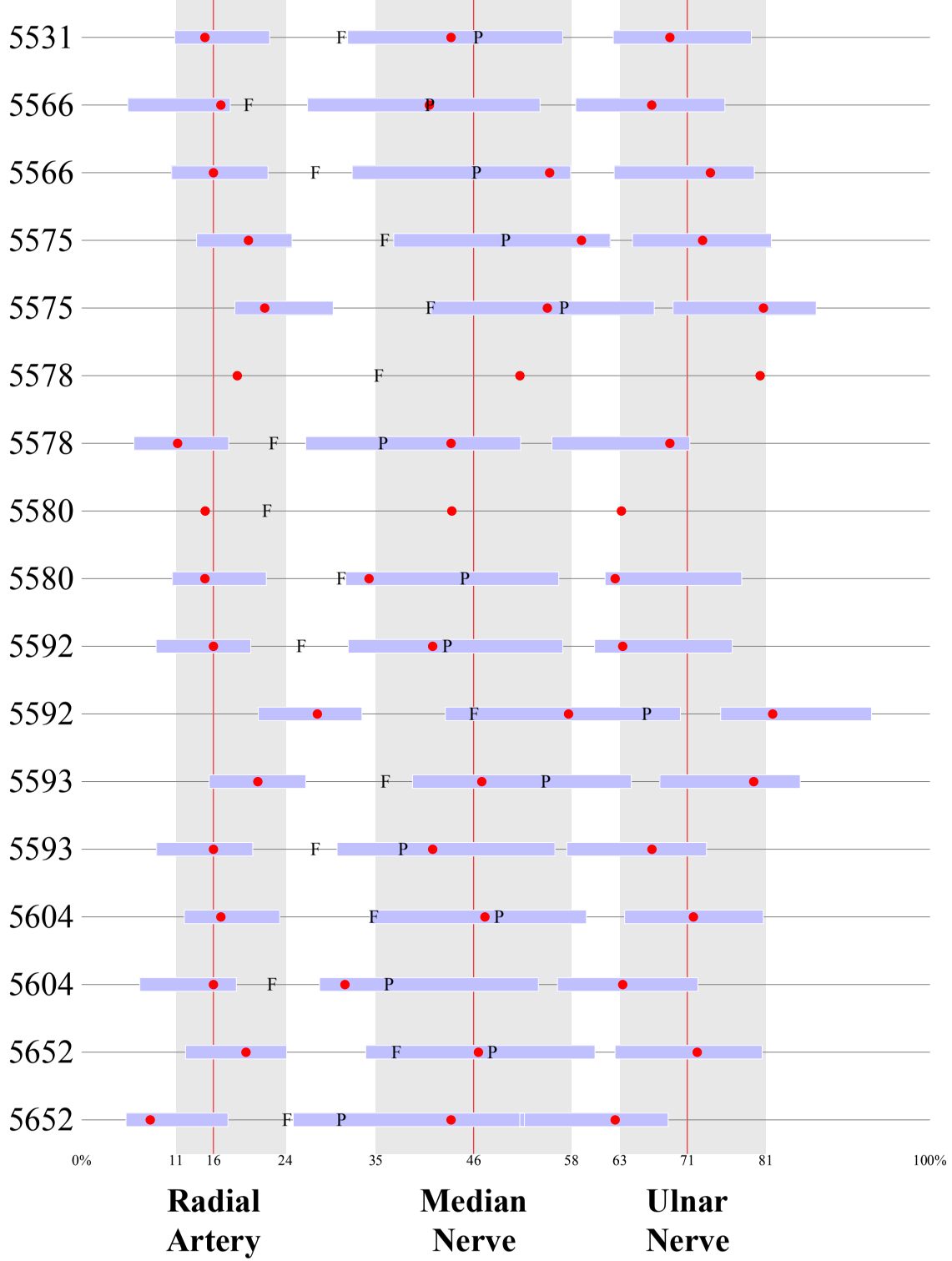
As with the hand-drawn plot, the specimens are presented in right/left pairs, with right first, then left. I chose to sort the specimen numbers within the sexes, as males typically have wider wrists than females and, as such, have greater distances in general between the radial artery and the median nerve, thinking this might impact the outcome. Males run from 5489 through 5595; females run from 5531 through 5652.

The very thin horizontal lines associated with each wrist marks the range from 0% to 100%. The Big Three locations are shown by the red dots. These are the raw, atomic-level data.

For each of the Big Three, the vertical grey band shows the range from the 5th to the 95th percentiles of the distribution. The red line inside of each grey band depicts the median value. The actual values of these quantiles is shown at the bottom of the plot, so, for example, 90% of the radial artery measurements fall in the range from 11% to 24% of the way across the wrist. While the median and ulnar nerves show relatively symmetric distributions, the radial artery shows more skew, with a heavy tail in the direction of the median nerve. We also see the greatest variation in the median nerve location, compared to the other two landmarks, as evidenced by the much wider percentile range. As with the hand-drawn plot, we see the smallest variation across the the radial artery measurements.

The "F" and "P" markers locate the flexor carpi radialis and palmaris longus tendons, the additional data points that were not included before. In every instance, the flexor carpi radialis falls between the radial artery and the median nerve and, therefore, acts as an anatomical barrier between these two landmarks. The palmaris longus, however, shows greater variation, at least relative to our landmarks, as it is sometimes lateral to, sometimes medial to, and sometimes directly over the median nerve.

The light blue rectangles show the 95% confidence limits on the point estimates of the Big Three locations computed



from the observable measurements; these rectangles mark the locations where the model thinks the Big Three are located. Note that there are several wrists for whom there are no such estimates. Note also that these wrists have no palmaris longus tendon measurements. Without this value in the model, we chose not to compute the estimates. [Of course, we could have conjectured a palmaris longus location through any number of such imputation techniques. We chose not to do so, however, so as to draw attention to the missing data.]

The estimates nearly always cover the true value. The counter-example can be found for the left wrist for specimen 5527, where the model misses the radial artery. The over-coverage is likely due to standard errors that are probably over-estimated: the models for estimating the locations fail to take into account the correlation of the Big Three locations, as the estimates are computed one landmark at a time.

The confidence intervals for the last wrist specimen (5652, at the bottom) show overlap between the estimates of the median and ulnar nerves.

Where does one inject?

The distribution-based percentiles shown by the grey boxes, and which mimic the hand-drawn plot, show us pretty much what we need: in the absence of other information, inject somewhere approximately one-fourth to one-third of the way across the wrist.

But how often do we have no other information? When looking at the atomic-level data, we can see that a strategy for finding the injection spot might be: Find the radial artery by either looking or feeling; move just medial to the radial artery until you find the flexor carpi radialis and then inject at, or just lateral to, the flexor carpi radialis, obviously avoiding the radial artery.

Of course, a test of this strategy ought to be carried out on an additional set of cadaver wrists and patients, as a model always fits its own model-fitting data the best. This would validate (or refute!) our strategy.

Our plot of the wrist data is an example of a number of our principles. We have small multiples. We have atomic-

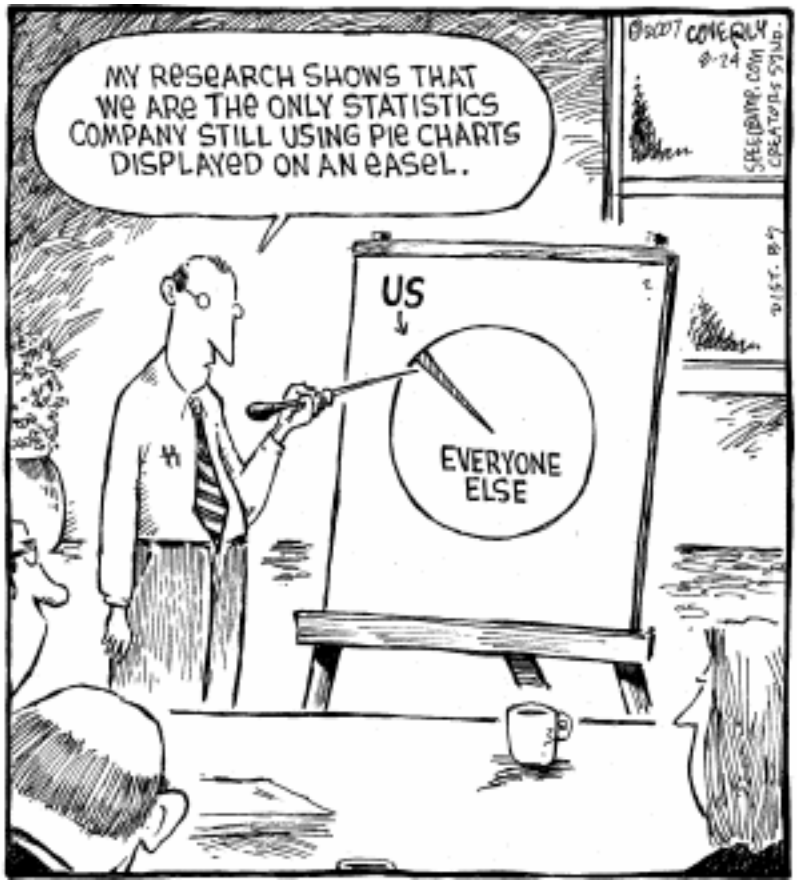
level data. We have layering and separation; our graphic works at several levels from the individual datum level, to distributions (with the inclusion of the percentiles), to a grand overview. The model-based predictors show us what is gained or lost when taking a modelling approach and reinforce that the exposition of the *distribution,* not the summaries of the distribution, is paramount.

And there is one other feature of our graphic that is either fortuitous or a consequence of good design or further evidence that the author is not working with a full deck: when the graphic is put together, overall it becomes a metaphor for the data themselves. Is it only coincidence that the long thin bands on the plot resemble an arm, the original source of the data themselves? Should the final design invoke an image of the original source of the data?

There is a lesson about design and planning here as well: **Design first on paper with a pencil and eraser; use a white board. Think about what you *want* to do and show, not just what you *can* do with the standard charting icons in the graphing instant wizard machine. If the tool you have doesn't let you do it, get a different tool.**

Most statistical software allows the user access to more fundamental drawing components and functions and have the facility to allow the user to operate in a "go here, do this" framework. All the plots we have seen have been composed of simple dots, lines, boxes, and connected curves. Once one does the computations in a statistics package, getting the information onto the page is typically just a matter of some simple algebra, cleverness, and programming. More often than not, the static routines built into statistical software will only take you so far; more work will be required to get the information onto the page.

Complicated data presented as serious evidence require serious tools. We often present complicated multivariate data; why do we expect that a simple pre-fabricated graphing instant wizard machine will be able to give us what we want? Learn to use the primitive tools in the software.



speedbump.com, 2007-08-24

Calls to the call center are databased; that is, every call that comes into the call center has its relevant information stored: where did the call originate, who answered, in what category was the question, when did the call start, when did it stop, and so on. Some of these calls generate "cases"; there is action that needs to be done after the call. Call centers are interested in measuring their capabilities and oftentimes, as in this instance, the time until a case can be called "closed" is a metric that these centers use to grade themselves.

As the consulting statistician, I was told that the leaders of the call center were interested in reducing the time to closure for the incoming calls. Of course, I was told the mean time to closure was some number of minutes, either 2 or 20 or 200 or something, I forget; it really doesn't matter for this discussion. They told me the mean, so naturally I asked for the raw, atomic-level data.

They gave me the data: a printout from an SQL routine that told me, accurate to twenty decimal places (I am not making this up!), the mean time to closure.

No, I need the data that you used to get these means; do you have that data?

After several weeks, I was given a data set with hundreds of call durations.

Do you have the start and stop times from which you calculated these durations, the actual times the calls came in and when the cases were opened and closed?

After several more weeks, I finally got the data: among other things, start and stop times for each of the calls. A plot of these data is at right.

The horizontal axis shows the day and time during the week when the call "came in" and a case was created. Note that no calls came in on Saturday, as the call center was closed. The vertical axis shows the time until the case that was generated from that call was closed. Note that the time until closure ranges from "negative" to "> 90 days". From this distribution, the call center had been calculating means as a summary measure for that distribution. [Remember: mean if, and only if, total.]

The horizontal axis has been marked with two relevant time points, namely 8am and 8pm, the times when the permanent and contract call staff answered calls. From 8pm until midnight, only contract staff handled all the incoming calls.

The vertical axis is not your typical linear or logarithmic axis. It shows continuous time but the mapping becomes more compressed as time increases in an effort to avoid bunching up the atomic data ink dots in this highly, highly skewed distribution. Within each band, time is uniformly spaced; however, these even-length bands do not contain the same amount of total time.

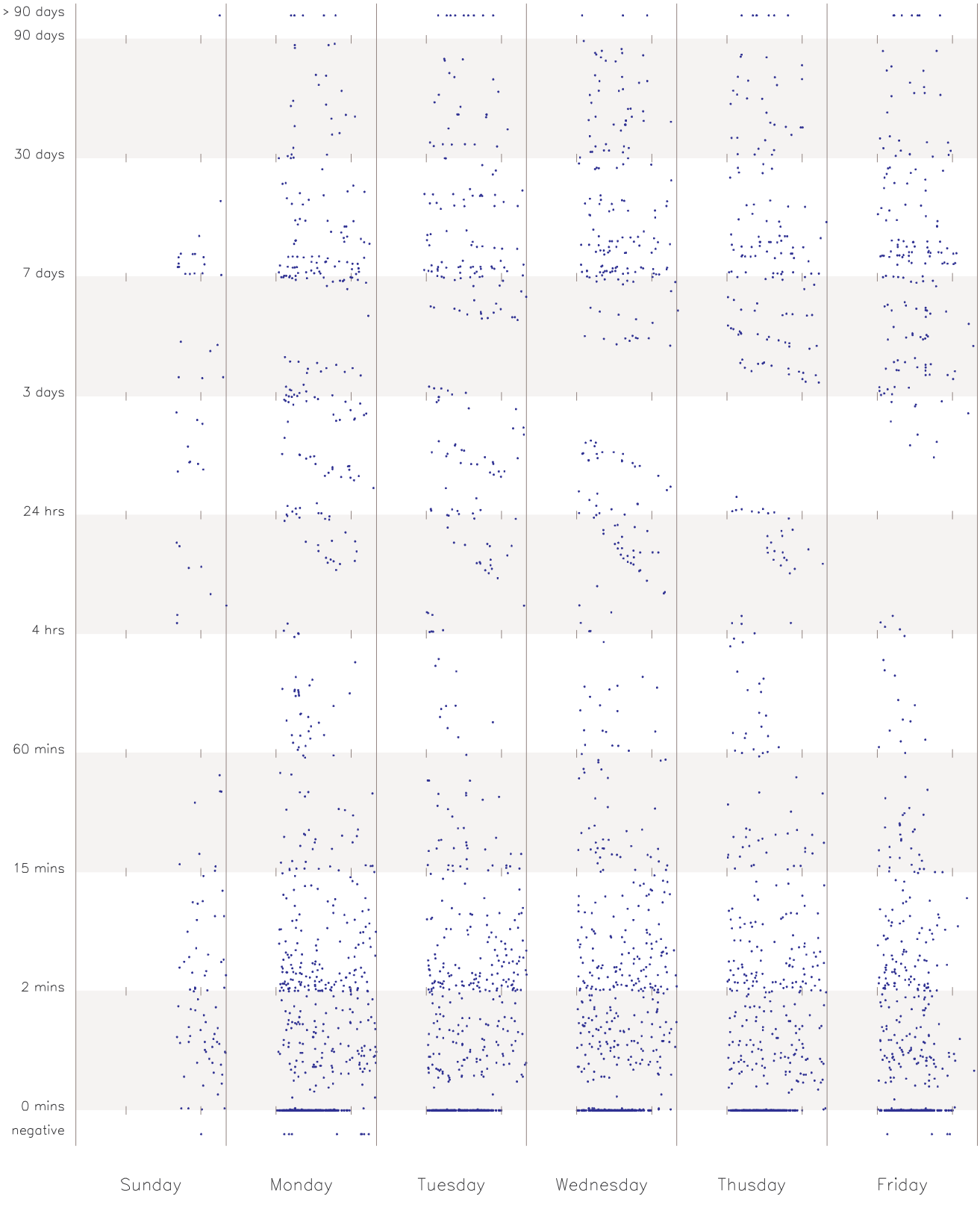
Some observations are in order, once we can see the atomic-level data:

There is point mass at zero. Note that this is not approximately zero, but actually zero. A look at the raw data revealed that the time of closure was identical to the time of opening these cases. They are not *rounded* to zero; they are *exactly* zero, down to the precision of the data collection device, the timeclock on the computers where these cases were logged. Note that this point mass only occurs, however, between 8am and 8pm.

Time to closure of cases by day of week and time

Subset of cases: all [ ] cases, April through September [ ]

Day and time of case creation is shown on the horizontal axis. Time to closure is shown on the vertical axis.
Small tick marks show 8am and 8pm. Time on the vertical axis is uniformly spaced between demarcations.



Rafe Donahue, PhD [ ]

There are values in excess of 90 days. All these values, except the one that was generated late on a Sunday evening, occur between 8am and 8pm.

The non-uniform scaling of the vertical axis produces a non-uniform window that demonstrates the absence of any closures taking place on weekends. Note that for cases that were opened late on Sunday, there are no closures in the time window that spans 5 to 7 days into the future. This is because such closures would necessarily occur on Saturday or early on Sunday, at times when the call center is not in operation.

This sliding weekend effect appears to get wider as the week goes on, a consequence of the unequal spacing on the vertical axis. Cases being opened on Friday get closed within slightly more than 4 hours, or they don't get closed until after the weekend. As the time scale becomes more compressed, we can see day/night differences; look, for example, at Thursday and the nearly parallel curves showing the case closures 4, 5, 6, and 7 days after the cases were opened.

Note that those cases that were opened late in the day on Friday (after approximately 5pm) were either closed quickly, within about 15 minutes, or were delayed until Sunday. Very few cases were opened after 8pm on Friday.

Once again, our data display is the model so our model here states that case-creation time is a source of variation in time to closure. A lesson one can learn from these data is that the call center might be able to reduce its average time to closure simply by tinkering with the days and hours that the center is staffed.

But from a data point of view, one must be concerned with the presence of the negative times to closure and the point masses at zero. I investigated these anomalies by tracing the data back through its sources until I found the database programmers huddling in the basement of the building. After showing them the plot and explaining the problem, they told me why such an instance should not be viewed as a problem at all: the server that collected the case-open times and the server that collected the case-close times weren't necessarily the same server. As such, because the servers weren't necessarily sychronized, it was possible to have negative values sometimes. *That is ok*, they said. *We usually just delete the negative values from the database, since we know they cannot be right.*

How can one know, then, I asked, *that values that are at, say, 74 seconds shouldn't really be at 98 seconds because the clocks are out of sync by 24 seconds?*

There is no way to tell, I was informed, because it is not possible to know which server is making the time stamp. But I was assured that this wasn't a problem because values of 74 seconds are possible but values of -20 seconds are not!

And what of the point mass at zero? Surely the server synchronization problem didn't explain that too?

I went to talk to the people who actually took the calls and opened and closed the cases. The call reps took calls over headsets and were talking while looking up various pieces of information. If a case was to be opened, that is, if the call reps couldn't derive the answer in real time, then they filled out an on-screen form and sent it into "the system".

But some of the senior permanent staff had figured out a trick. Because calls

came fast and furiously during the day and time-on-hold and number-of-rings-before-answering were also gradable metrics, the call reps learned to do what they could to keep their non-talk time to a minimum. And this meant making sure the computer was ready when the calls came in and setting things up so as to not have to wait for the computer to respond when they needed information.

What all this meant is that the experienced permanent staff had learned to open up oodles of the on-screen case-open forms at the start of their shift, *before* their session of call-taking began, filling their large computer screens with empty case screens. By doing so, they were avoiding a certain downtime while fielding calls by having these forms open, instead of waiting for the case form to open when they first answered the call.

They also found a way to close a case immediately upon opening it (or open it as it closed), so that they would get credit for resolving issues, and resolving them quickly, in the event that they were able to answer the question without actually creating a case. So, in essence, the zero-length cases were a different kettle of fish altogether, an artifact not of the way cases were perceived by management but as a way to keep the system running, and not getting bogged down in waiting for the computers.

Learning that there were two different types of operators of the system, I wondered if operator type was a source of variation. We can split the data, then, on the basis of contract or permanent staff and look at the resulting distribution of times to closure for each group.

For simplicity's sake, I simply re-ran the previous plot twice, once subsetting on only those cases taken by the contracted staff and once subsetting on only those cases taken by the permanent staff. The two re-runs are shown on the following spread (reduced to fit and allow for some text).

Cases for the contract staff are on the left and cases for the permanent staff are on the right and the side-by-side differences are astounding. The compelling distribution we saw in the last plot is really a mixture distribution from two highly different processes. The process on the left shows the variance that is naturally in the process when one fills out the on-screen forms according to plan while the process on the right shows the use of the trickery of data entry on the part of the permanent call reps.

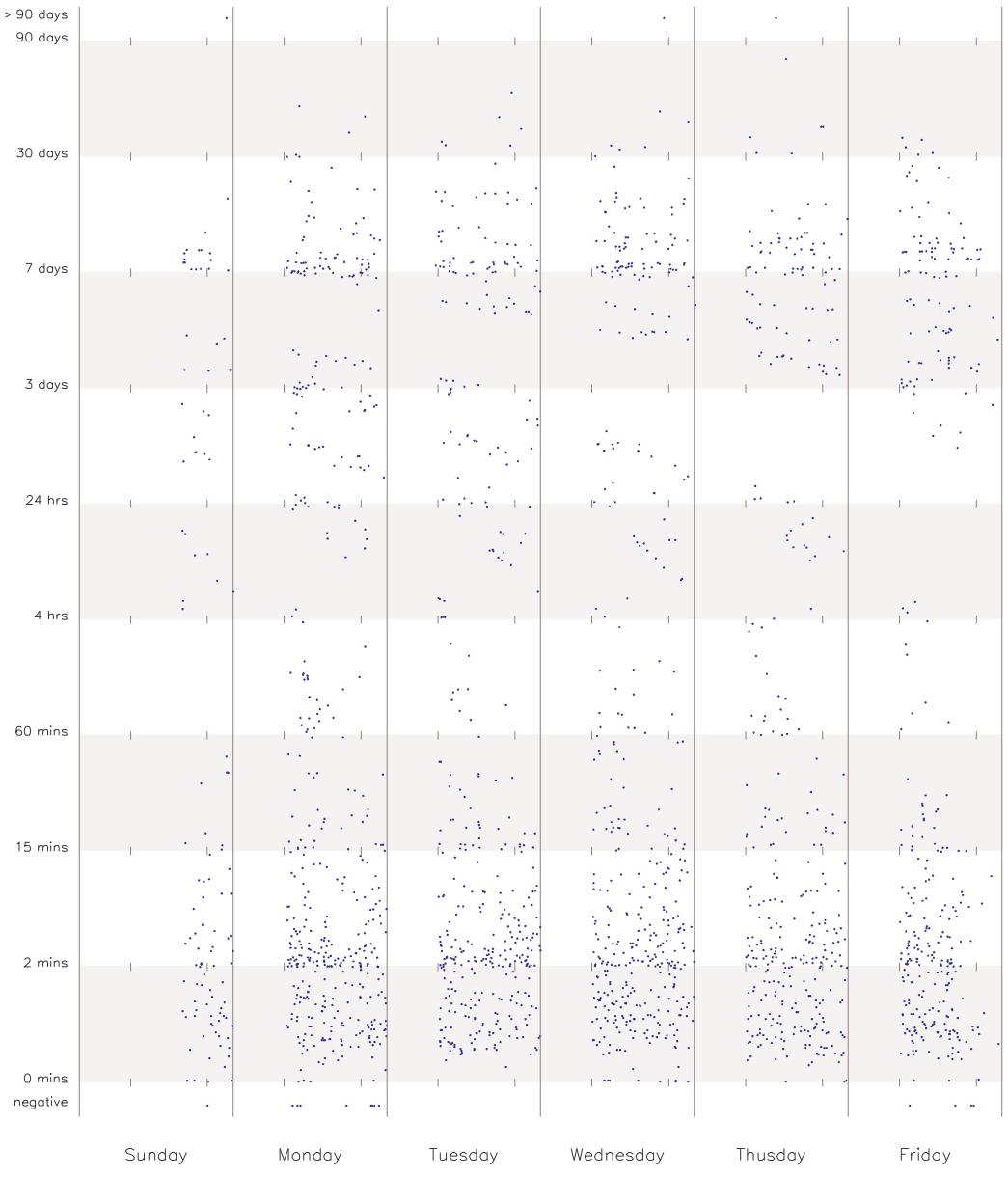
The left distribution shows essentially none of the zeros but all of the negatives. Obviously, then, there is something about the way the permanent staff are tricking the system that allows them to avoid the obvious server synchronization problem; whether or not their non-negative times are biased by synchronization issues cannot be assessed.

The distribution of times to closure for the permanent staff is stunning in its uniformity, once the point mass is taken out of consideration. These permanent staff created all but three of the cases that have dragged on for over 90 days.

Are these two staffs behaving differently or are they getting calls of different natures? Are the "tough" calls being sent more often to the permanent staff? Is there any selection bias in whom gets what type of call?

Are the staffs even doing the same thing? The presence of the point mass at zero for the permanent staff and the negatives for the contract staff and the disparities in the distributions lead one to wonder if these are really the same process. Should both groups really be combined in an effort to examine times to closure?

And what of the mean time for the total mixture distribution? Does the trick employed by the permanent staff carry enough weight to offset the impact caused by the over-90 days cases? Could we lower the mean significantly (whatever that means) by just teaching the contract staff the trick with the on-screen windows? If that were enough to impact the mean time to closure, what does that say about the process as a whole and the use of the mean as a summary statistic for measuring performance?

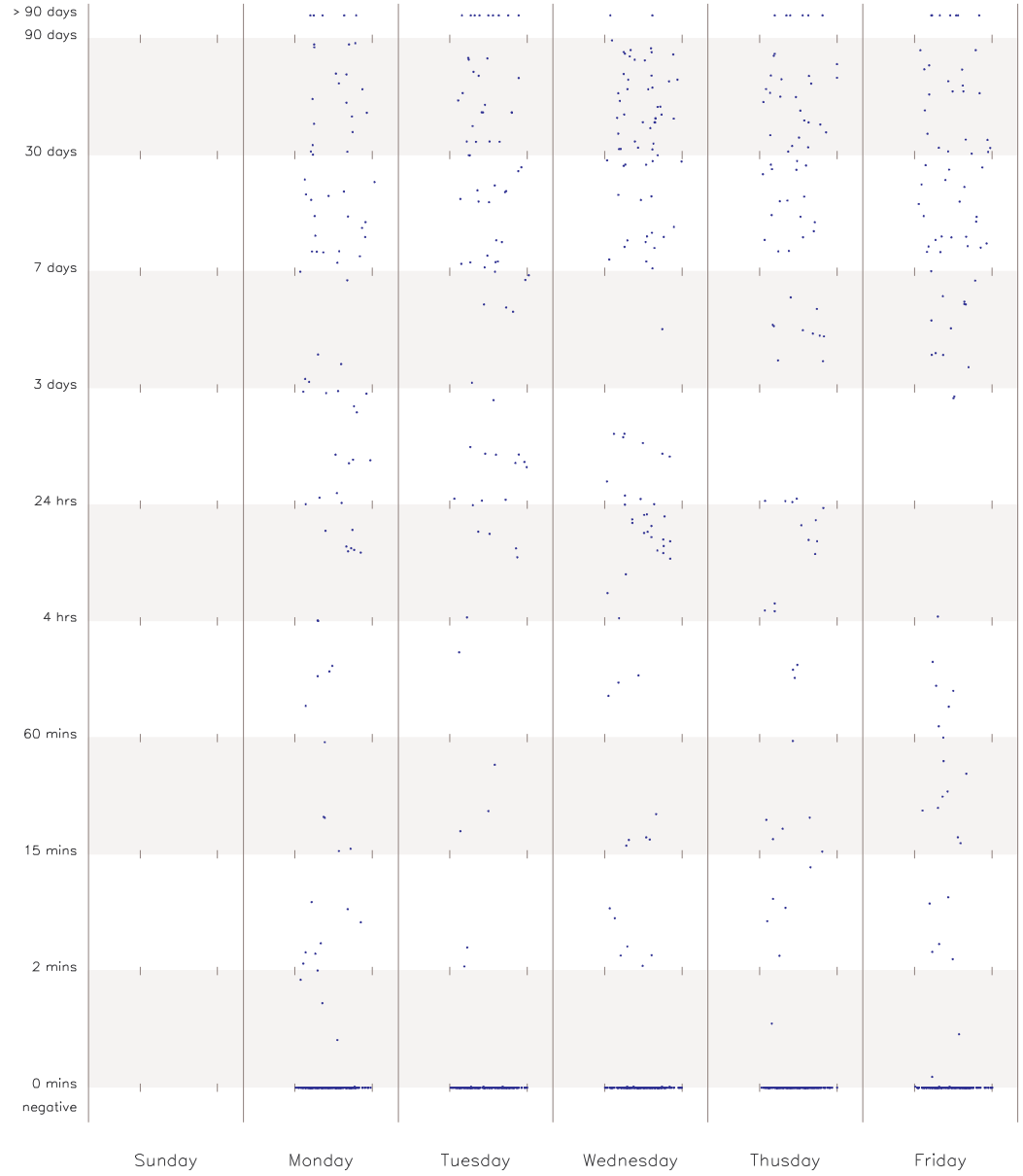


A few notes on features of the plots themselves. Note that we could have employed a you-are-here method to these plots by putting the complementary data points in the background of the plots. Unfortunately, when these plots were made, I had not yet envisioned that idea.

The time demarcations are alternating, light-weight stripes in the background, improving our ability to focus on the data. Again, like a good wait-staff, they are there when you need them, but essentially invisible when you don't.

The original plots, subsets of which are shown here, included titles and credits and explanations like the mixture distribution, so as to provide background information that supports the data and their interpretation.

The vertical axis is neither linear nor logarithmic but a hybrid that allows us to deal with the point mass, the anomalies (negatives and over-90s), and the extreme skewness in the data. The linear scaling between connection points allows us the ability to easily interpolate between the connection points if we need to do so.



Following trauma, one reaction of the body is hyperglycemia, the elevation of the amount of sugar in the blood. Exactly why this is so is the subject of some debate; yet, physicians work to avoid high blood sugar values in their patients, as these high blood sugar values can have long-term negative consequences such as retinopathy and nephropathy.

In an effort to keep glucose under control, physicians in the intensive care unit (ICU) have implemented a protocol, or standard course of treatment, that is given to all patients who are in the ICU. This protocol states that all patients are to have their blood glucose levels measured every two hours. Based upon the results of this blood glucose reading, a certain amount of insulin is given. This has the general effect of reducing blood glucose because that is what insulin does.

The protocol also contains provisions to increase blood sugar should hypoglycemia be found. This takes the form of reducing the insulin level given (sometimes to zero but not always) and sometimes actually administering dextrose.

So the goal of the protocol is to keep the patients' blood glucose levels relatively constant and within a safe range. Big drops are bad, and big increases are also bad, unless, of course, we are deliberately trying to drop or raise the patient's blood glucose by some large amount.

Because this protocol is computerized (the blood glucose readings are fed into a computer and this computer tracks the insulin that has been given in past intervals and computes what the new value should be), all the blood glucose values and the henceforth derived insulin values are all in a database.

What do the data tell us about the protocol?

The plot below shows over 53,000 changes in blood glucose across 876 patients in the ICU. Each red dot shows the time in hours between successive readings on the horizontal axis and the change in glucose level in mg/dl on the vertical axis.

The grey shaded region shows moving 5^{th} and 95^{th} percentile estimates. The black and grey lines across the center of the plot show, respectively, the moving mean and median glucose levels.

While the mean and median show little to no relevant change, the variation in the data highlights the difficulty in maintaining glucose control in these patients.

While the time between readings is in the neighborhood of 2 hours, there is marked deviation from that desired time interval. There are readings in the data set that are 2, 3, 4, and more times the desired value of 2 hours. There are also readings that take place minutes apart and a cluster of readings at three-quarters of an hour. This 45 minute cluster is a result of a special provision in the protocol that calls for early re-readings in the event of hypoglycemia. As such, there is a selection-bias-induced increase in the change from the previous reading when that reading falls 45 minutes after its predecessor: values that follow low values tend to be higher and produce positive changes. This is both regression to the mean and selection bias and the physiological effects of reducing insulin and increasing dextrose.

The plot shows us layering and separation, with the individual data lying atop the 5^{th} and 95^{th} percentile points. The mean and median are nearly identical, only wavering when the data count gets low, in the neighborhood of 5 hours or so, attesting to the symmetry in the distribution. The grid lines hide quietly behind the shaded quantiles, except where we added dashed lines atop the blood splatter at 2 hours and at zero change. Note also the redrawing of the quantile boundary lines atop the places where the atomic-level data density obscures the edge of the shading.

The value of the atomic-level data is seen when one thinks about changes of blood glucose of 150 mg/dl in either direction.

What is most astonishing, however, is the fact that the edict of the protocol, the "every 2 hours" requirement, is not even remotely followed. *On average*, patients are being measured at 2 hours but the effects of insulin and glucose given intravenously are wonderfully rapid; is checking at 1 hour and checking at 3 hours really the same thing?

The intent of regulating blood glucose looks to have, on average, little effect, but the variation present in the data leads one to wonder about the impact of changes of 50 mg/dl or more. Even though only approximately 10 percent of the data are changes in excess of 50 mg/dl in either direction, that still points to over 5,000 such readings.

But changes are made from start points and end points, so we could just as well plot the starting glucose values as a function of time between readings and the ending glucose readings as a function of time. [We note, here, that plotting starting glucose values as a function of time between a starting value and an ending value in an interval violates a premise concerning cause and effect: how can the starting glucose value be impacted by the time between that value and one that succeeds it? We likely should plot time between readings as a function of the start-


Change in glucose reading
150
100
50
0
−50
−100
−150
0
1
2
3
4
5
6
7
8
Time between readings (hours)

ing reading, as we know that time to the next reading is at least partially dictated by the starting reading, certainly in the case of hypoglycemia. Regardless, we will plot the glucose readings as functions of the time between readings, so as to maintain design consistency.]

The plots of starting reading against time between and ending reading against time between are shown at right.

The plot of the starting glucose readings against time follows the same general format as the change versus time plot. The points and lines and shading still mean the same thing as before with the exception of two additional reference lines marking 80 mg/dl and 110 mg/dl on the vertical axis. These are considered by the investigators to be normal ranges; the investigators sought to keep patients within this range.

Of note now is the difference between the mean and median values. The skewness in the data pulls the mean above the median. This skew is also present in the greater distance between the 95th percentile value and the median as compared to the distance between the 5th percentile value and the median.

We also see the reduced value of starting glucose values at approximately 45 minutes, a consequence of the protocol seeking to retest hypoglycemic patients at an earlier time point. The extra mass at that time point has a sharp cutoff at the 80 mg/dl line, indicating the threshold below which quick retests for hypoglycemia were mandated.

Careful examination of the atomic-level data reveals what is likely a detection limit phenomenon in the neighborhood of 10 mg/dl. Both the start and end readings show some readings all at the same distance from the horizontal axis. What is the impact of such a threshold?

Note the number of individual ending data points below the hypoglycemic threshold value of 80 mg/dl. While one might argue that it is only approximately 5% of the total, with over 50,000 data points, 5% is over 2500 individual data. So that means we have records of 2500 hypoglycemic events.

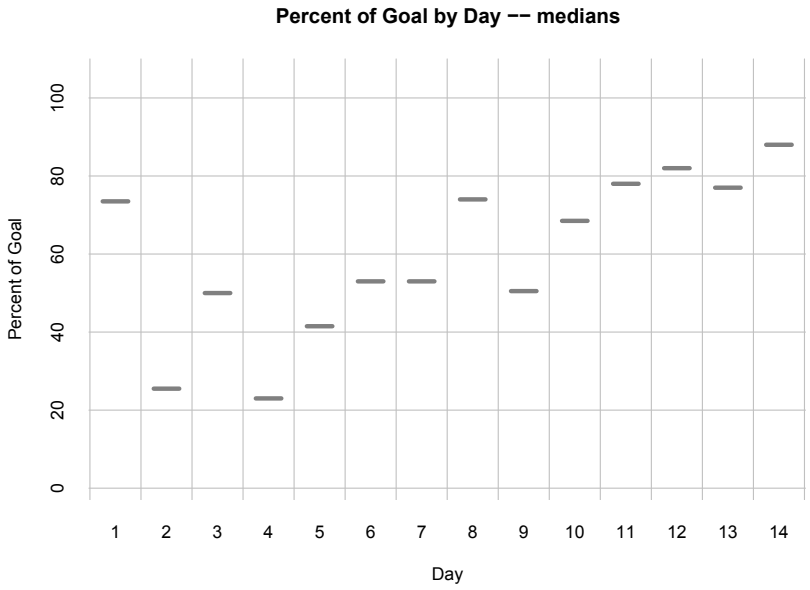
While the mean, if not within the normal 80-110 range, is certainly close to being considered normal, the variation is dramatic. The raw number of hypo- and hyperglycemic events is staggering. What we see, as we have seen before, is that *the mean is not the distribution*. As statisticians, we are called to understand the distribution, not just the mean. Nothing makes this more evident than actually seeing the distribution. While this protocol keeps the *average* near the normal range, keeping the *distribution* within the normal range is a more difficult task. I can put my feet in the oven and my head in the freezer and *on average* have a normal body temperature, but that hardly implies that I am comfortable.


Starting glucose reading
0
50
100
150
200
250
300
0
1
2
3
4
5
6
7
8
Time between readings (hours)


Ending glucose reading
0
50
100
150
200
250
300
0
1
2
3
4
5
6
7
8
Time between readings (hours)

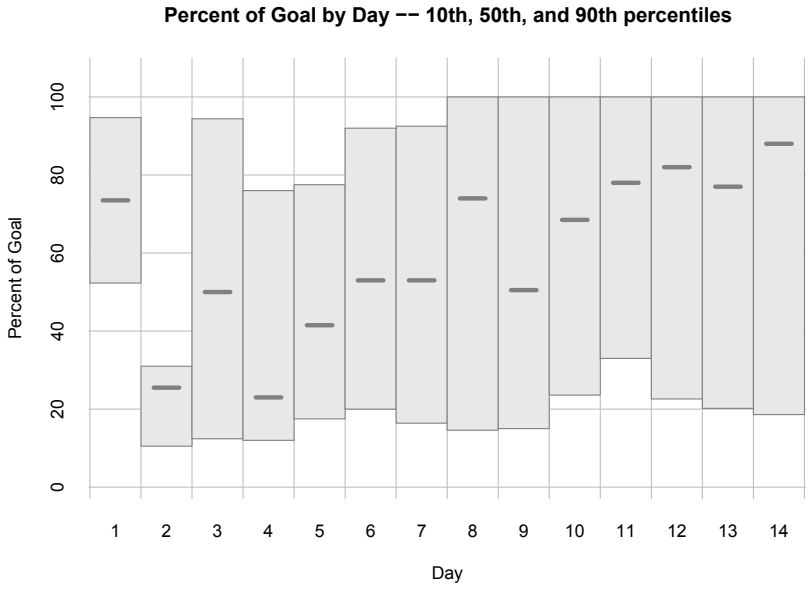
In the surgical intensive care unit (SICU), one issue is the amount of nutrition each patient there receives relative to the amount that the caregivers desire for the patient. This "amount of nutrition" is typically expressed as a "percent of goal". Ideally, the physicians would like all the patients at some minimal amount of nutrition but other factors, such as swelling or infection or whatnot, can hamper this effort.

I was presented with a collection of data that gave, for each of the first 14 days since start of stay in the SICU, the "percent of goal" for each patient. Several dozen patients have data in this collection. Some have data on just one day; some have data on almost all the days; there is not a consistent number of patients on any given day. The goal here is to talk about the presentation of data and the impact of summary measures.



This first plot shows the median percent of goal for each of the first 14 days. One can see rather variable median percents of goal over the first 4 or 5 days, followed by a general, but not monotonic, rise in the percent of goal over the last week. On day 14, the median percent of goal is 90%, perhaps an indication that all is well. After all, the median percent of goal is nearly 100%, so we must be doing well, right?

The second plot adds to the medians the 10th and 90th percentile points. Now we see evidence of the distribution of the percents of goal. None of us would admit now that we didn't realize that there just had to be spread in the data; we probably just didn't realize how big it could be.

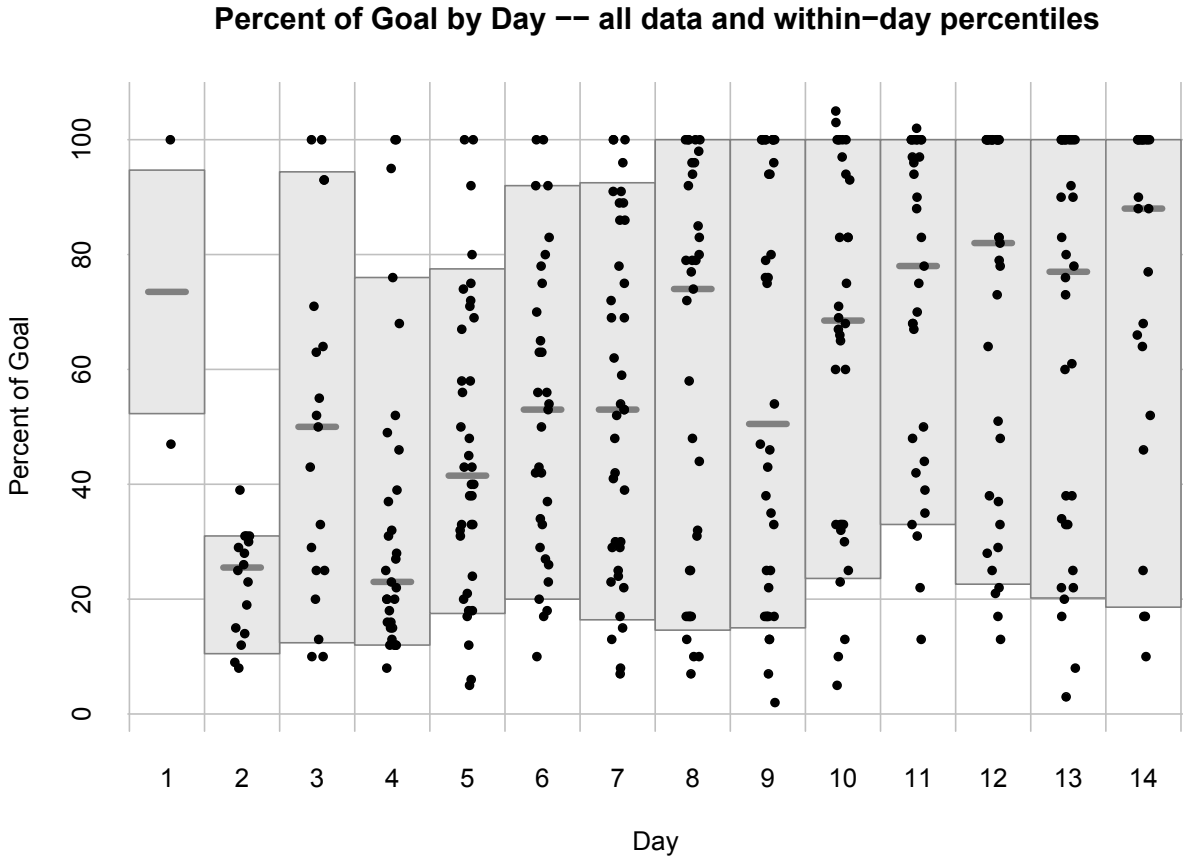


But recall that the median is the point where 50% of the data lie on each side; as such, having, even at day 14, a median of 90% implies that 50% of the patients who have data that day have a percent of goal that is less than 90%. How much less than 90%? Well, the 10th percentile is about 20% of goal. Yikes. Even though the median is 90%, 1 in 10 patients is getting less than 20% of what he or she should receive on day 14.

And compare days 1 and 2: real high then real low. That seems perhaps like odd behavior. Perhaps

more digging is needed.

The third plot adds all the actual data points to the summaries that have just been discussed. The points have been horizontally "jittered" slightly so that identical values don't obscure each other. All sorts of discoveries are available now to those who will take the effort to look:



Day 1 has only two data. Perhaps the median and the 10th and 90th percentiles (however they may be estimated from a sample of size 2) aren't very good summary measures. In fact, on day 1 we have three summary measures to show only two data. Goodness. That's just silly.

Day 2 has a whole bundle of data from 10% to 40% of goal.

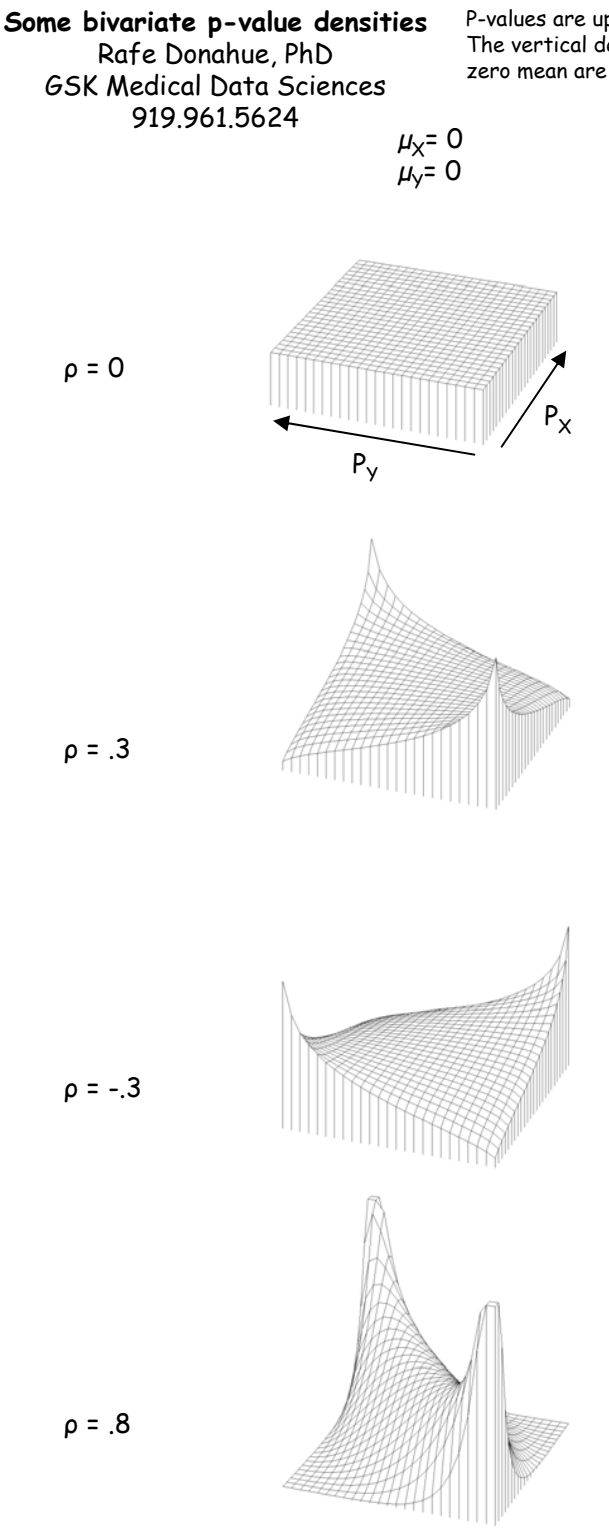
The minimum on day 14 is a lowly 10% while more than 50% of the points are above 90% of goal.

On day 10, two patients actually have values above 100%, and there is also a datum over 100% on day 11.

There is a growing proportion of patients who have 100% of goal achieved as the study goes on.

All of these discoveries are available once we view the data on the atomic level; we have done data analysis. Investigation of these peculiarities of the data is perhaps the fundamental component of data analysis. We reveal the anomalies in the data through analysis, the study of the component parts.

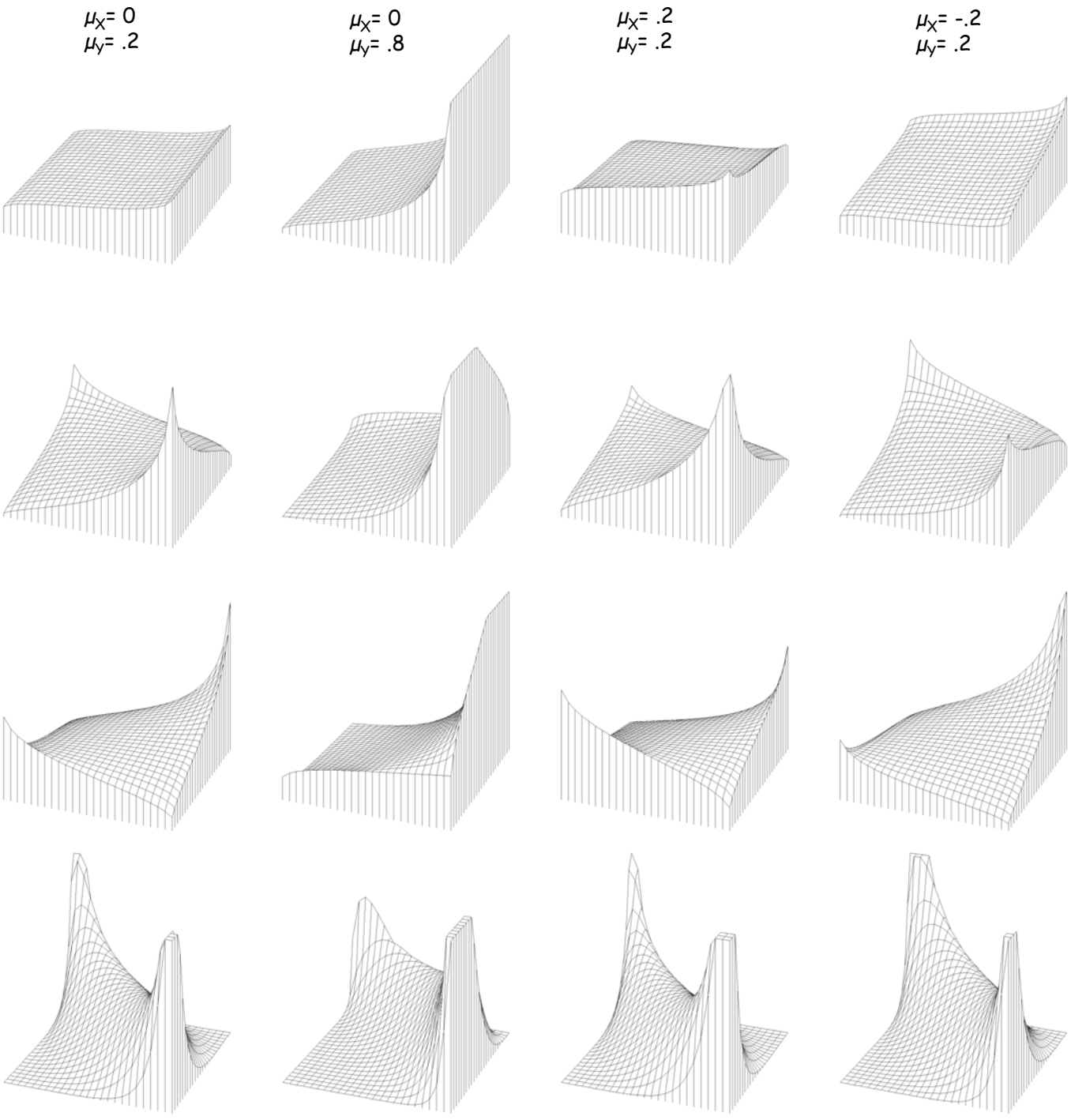
Some other miscellaneous plots, with various features that can be described:



The data display is the model. Note that the shape of the bivariate density changes with changing values of the parameters. What we see in the plot is a consequence of the changing levels of the parameters. The model for understanding the density then is described by the layout.

On the next spread, the data themselves show us a Cochran-Mantel-Haenszel test. The differing baseline levels are the strata, summing up to produce the distribution at the top. The distribution at the top is decomposed into sources of variation due to treatment, baseline, and remission status.

pper-tailed from null hypotheses of zero mean for bivariate normally distributed X and Y with unit variances and the specified parameters.
ensity scale is truncated at 4 units, leading to some flat-top peaks. The p-value scale runs from 0 to 1. Note that marginals in cases with
uniform although conditional densities need not be so. Power to detect means of 0.2 and 0.8 is approximately 7% and 20%, respectively.



HAMD hypersomnia items total change score distribut

Rafe Donahue, PhD, GSK

All patients

Patients wi

These 1806 patients are from studies 30926, 30927, 4001–4003, and 209. Each dot shows one patient. The plot shows the distribution for each dose group as a w
Solid boxes depict the median in each row; cornered boxes show the 1st and 3rd quartiles. Negative change (to the left) shows reduction in hypersomnia total score
to smaller sample size?). The bupropion group shows median hypersomnia levels better than, or no worse than, the SSRIs at all baseline levels for all patient groups.

ion at 8 weeks by baseline level and remission status

Biomedical Data Sciences

thout remission

Patients with remission

rom baseline value −−>

<−− Change from baseline value −−>

Placebo

Bupropion

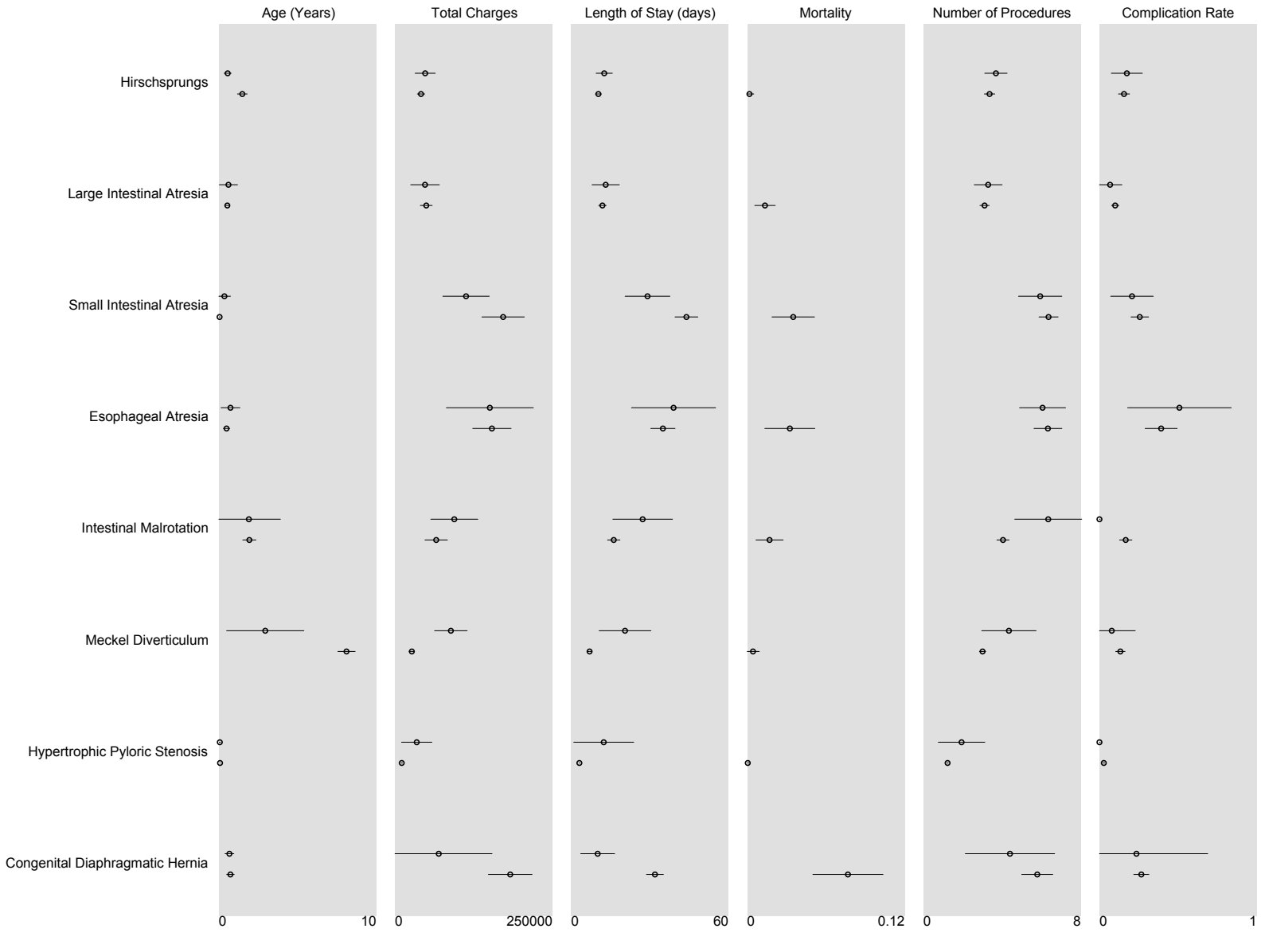
SSRIs

0 1 2 3 4 5 6

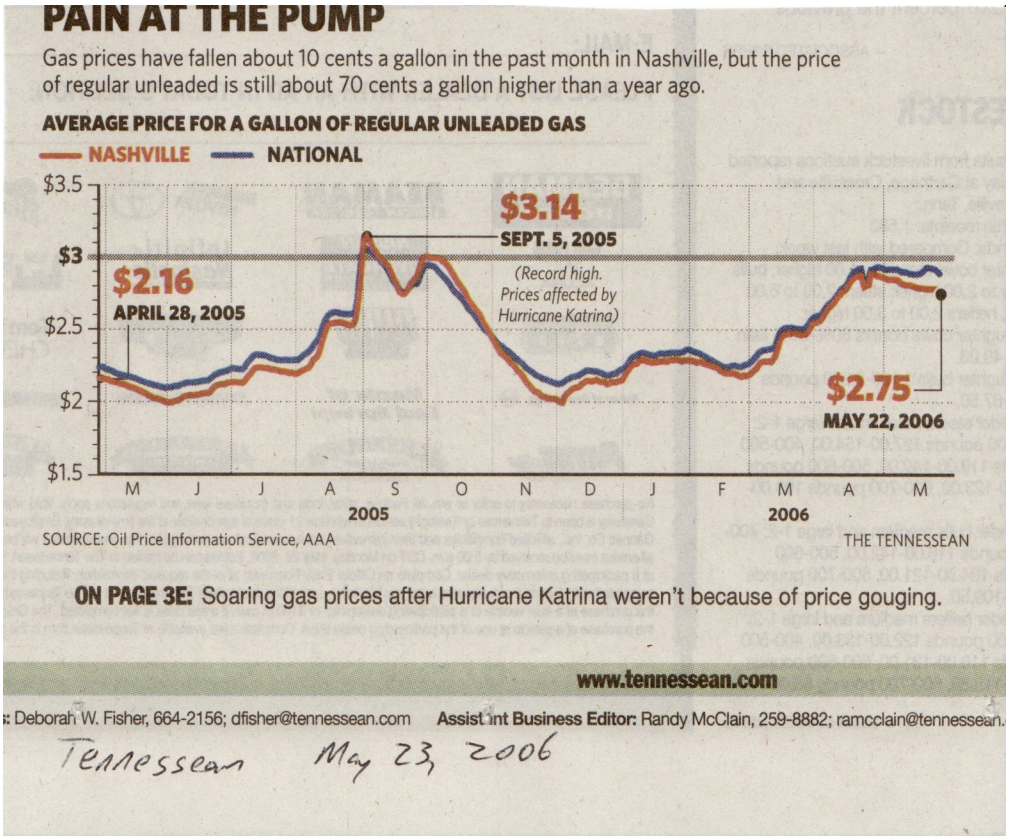
−6 −5 −4 −3 −2 −1 0 1 2 3 4 5 6

whole (top left), and then split out by baseline level (rows), and then remission status (center and right columns). Remission is HAMD17 of at most 7.
. Each box holds potentially 256 patients. Narrower support is seen for the remitters compared to nonremitters (perhaps due
Those with no median difference show advantage for bupropion or similarity with SSRIs at one or more of the quartiles in 13 of 14 cases. 12:25 17MAY2005

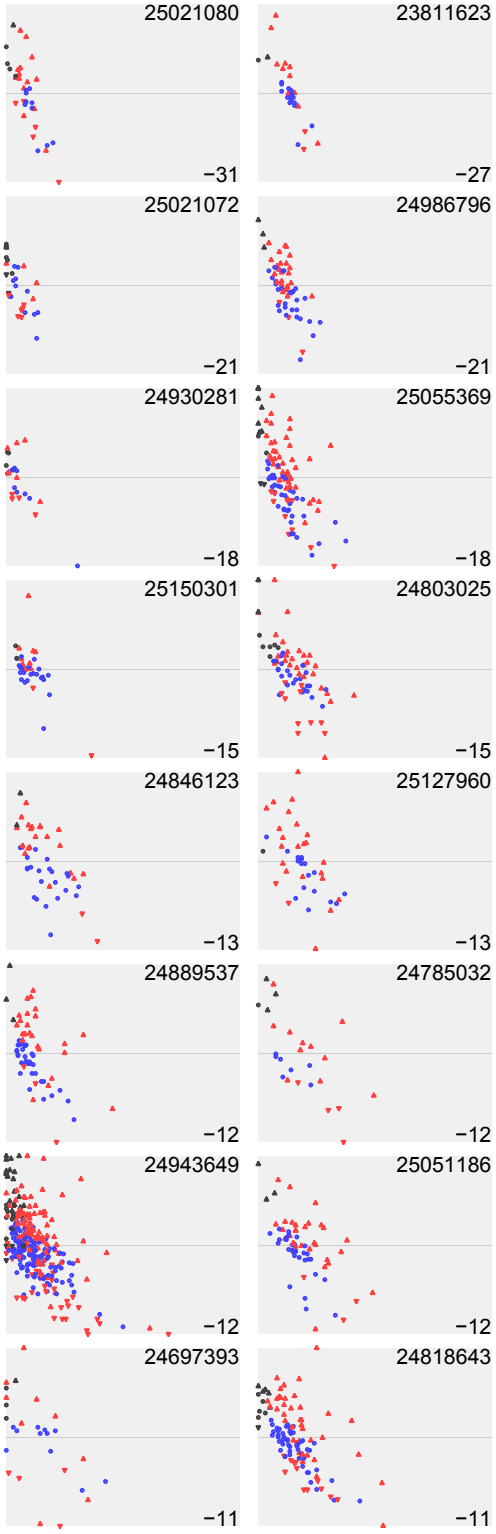
Estimates and confidence intervals for patients grouped by reason for treatment (Hirshsprungs, Large Intestinal Atresia, etc.) and two different time frames (Then and Now). Note the missing mortality data for one of the time frames.

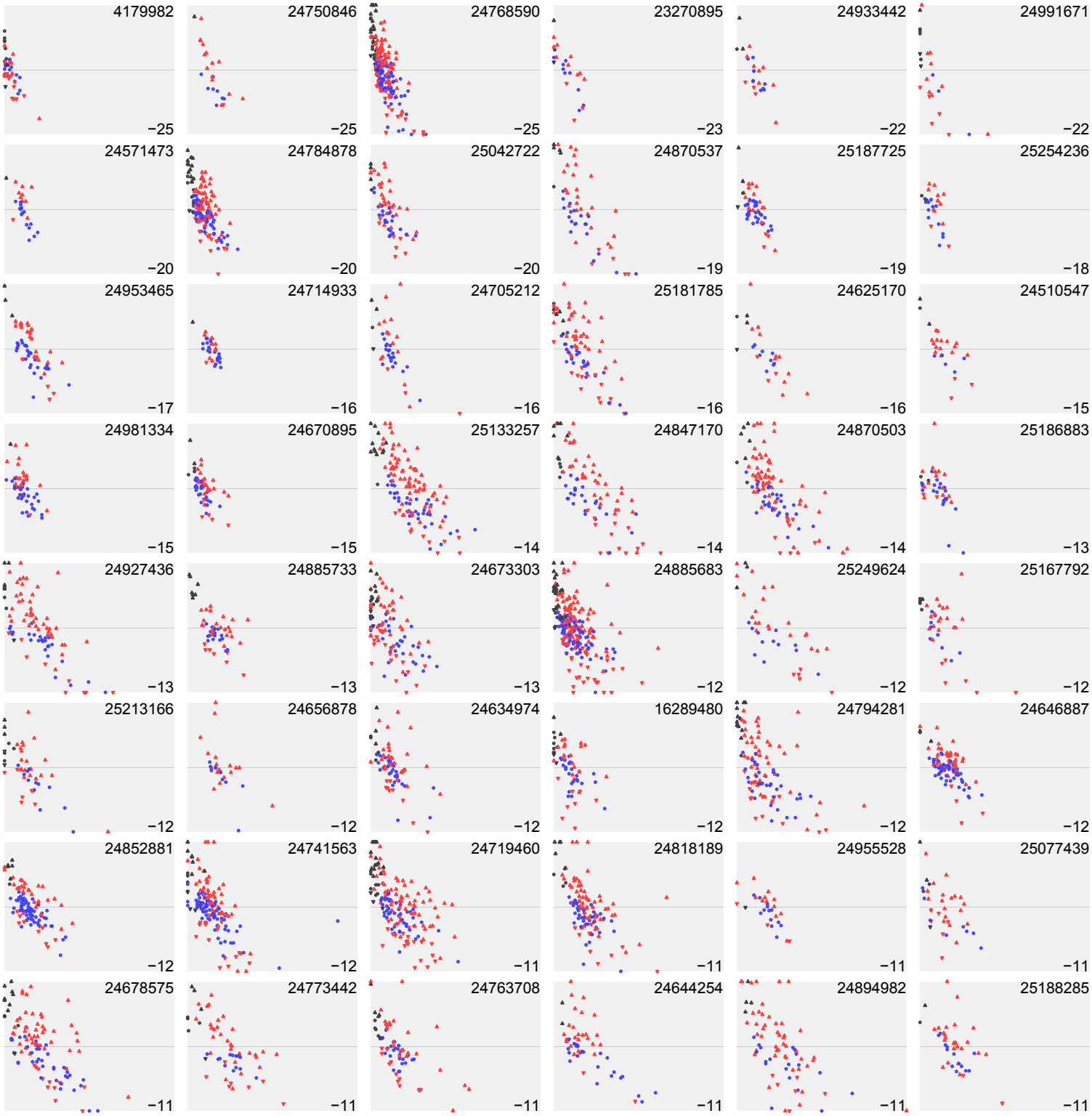


Times series make for fine accounting but are not models that improve understanding unless auxiliary data are added.



Insulin rate versus change in blood glucose for 64 ICU patients. Red dots are values that fell out of range, arrow direction shows high or low. Black values are values that were augmented with dextrose. Blue values are values that fell within normal range. Eight-digit number is patient identifier, two digit number is slope of simple linear regression (mg/dl/insulin-unit), showing how far blood glucose changed per unit dose of insulin. Patients are sorted by increasing slope. Horizontal reference line shows zero change.




4179982
−25
24750846
−25
24768590
−25
23270895
−23
24933442
−22
24991671
−22
24571473
−20
24784878
−20
25042722
−20
24870537
−19
25187725
−19
25254236
−18
24953465
−17
24714933
−16
24705212
−16
25181785
−16
24625170
−16
24510547
−15
24981334
−15
24670895
−15
25133257
−14
24847170
−14
24870503
−14
25186883
−13
24927436
−13
24885733
−13
24673303
−13
24885683
−12
25249624
−12
25167792
−12
25213166
−12
24656878
−12
24634974
−12
16289480
−12
24794281
−12
24646887
−12
24852881
−12
24741563
−12
24719460
−11
24818189
−11
24955528
−11
25077439
−11
24678575
−11
24773442
−11
24763708
−11
24644254
−11
24894982
−11
25188285
−11

Distribution of number of prescriptions per prescriber over a two-year period for four vendors of prescription data. Note how Vendor C looks different and has no data on odd values of the support set; only even numbers are present. Further inspection of the data on an atomic level revealed duplicated data in the database.

Number of prescriptions per patient. Note the double data for "AdvancePCS".

Table 4. Summary measures for distribution of transactions per patient based on a random sample of 100,00 patients

Market	Vendor	10 %ile	25 %ile	50 %ile	mean	75 %ile	90 %ile	95 %ile	99 %ile	max
depression	ArcLight	1	1	2	5.6	7	15	21	34	209
	Little Dendrite	1	1	3	6.5	8	17	24	44	171
	Big Dendrite	1	1	3	6.5	8	17	24	44	734
	Verispan	1	1	4	8.3	11	21	29	51	781
hiv	AdvancePCS	2	2	4	7.5	4	12	24	100	580
	ArcLight	1	1	1	2.9	2	3	8	48	192
	Little Dendrite	1	1	1	3.4	2	4	13	54	214
	Big Dendrite	1	1	1	3.4	2	4	13	53	204
	Verispan	1	1	1	3.5	2	5	15	55	221

What summary statistic should be used to index these data? At least the authors have allowed us to agree or disagree with their choice of the Pearson correlation p-value coefficient and have shown us the data.

Figure 2. Urine PGE-M distributions of healthy patients and those with Crohn's disease.

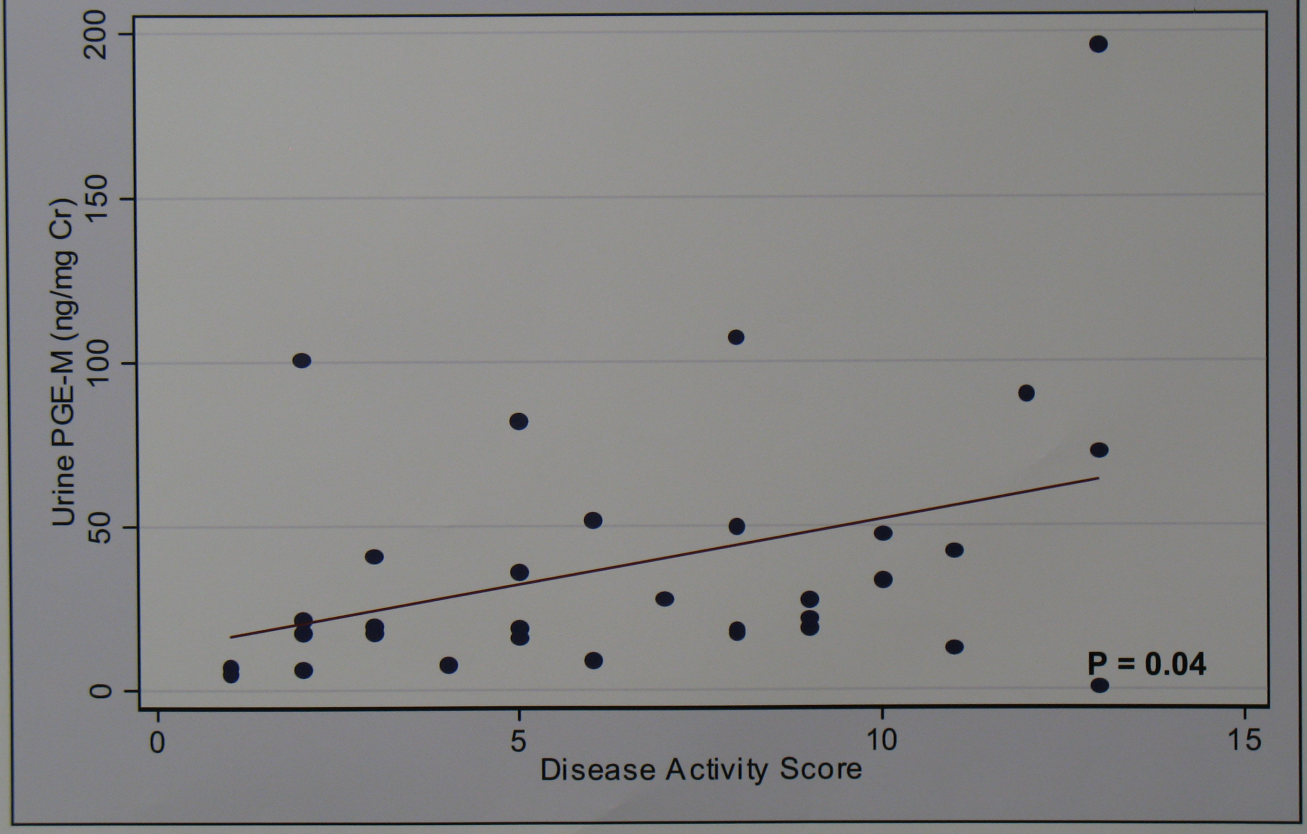
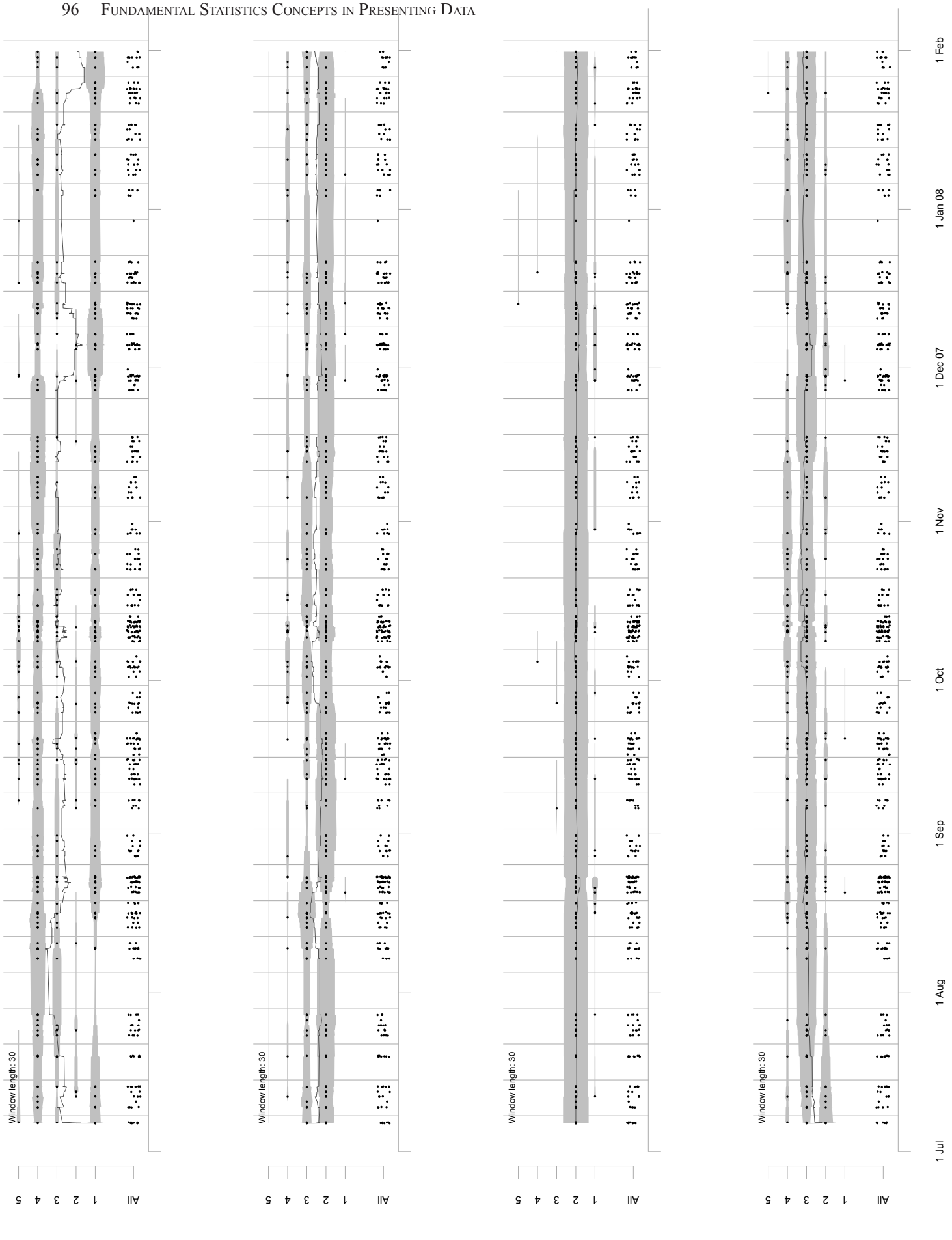


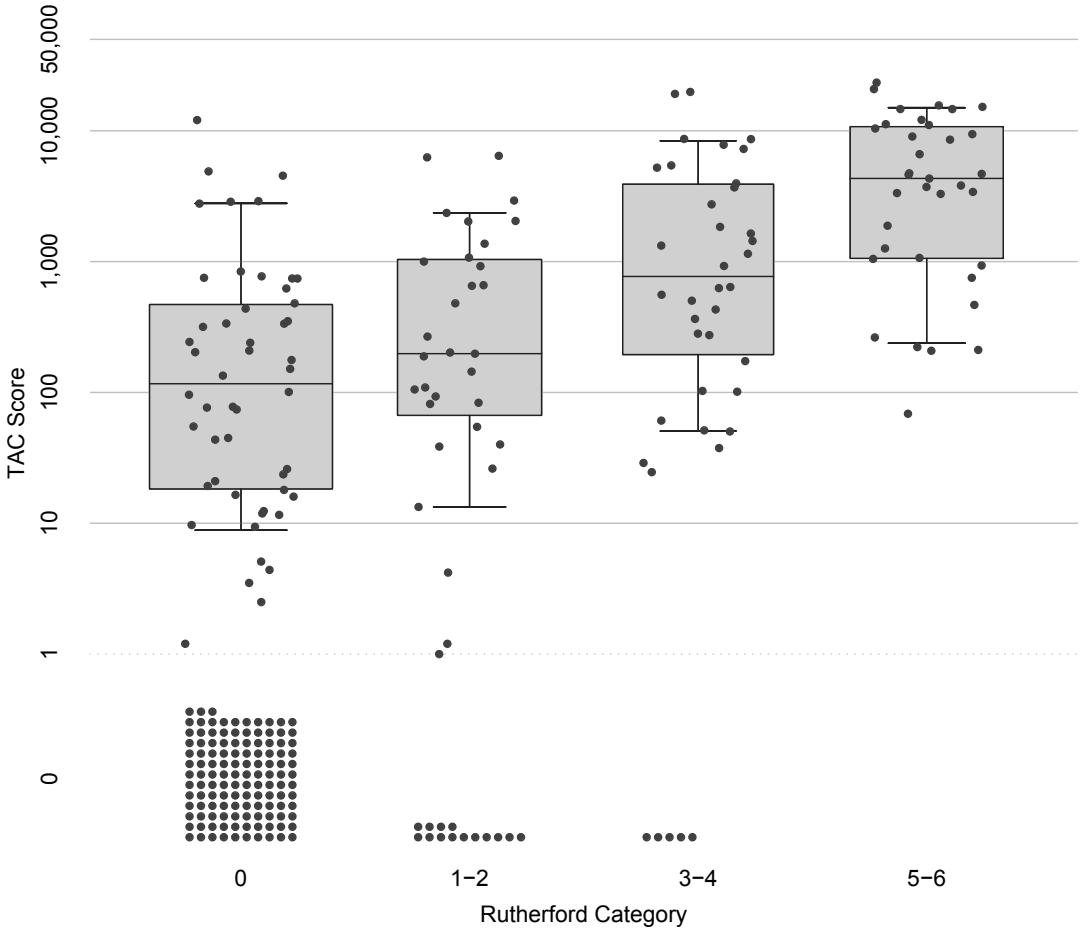
Figure 3. Correlation between urine PGE-M and disease activity as measured by the HBI.


Readiness
Organization
Completeness.of.Content
Comprehension
Window length: 30
All 1 2 3 4 5
1 Jul
1 Aug
1 Sep
1 Oct
1 Nov
1 Dec 07
1 Jan 08
1 Feb

Global.Rating
Degree.of.use.of.artifacts
Engagement
Coordination.and.Conflict.Resolution
Window length: 30
All 1 2 3 4 5
1 Jul
1 Aug
1 Sep
1 Oct
1 Nov
1 Dec 07
1 Jan 08
1 Feb

This plot shows the distribution of tibial arterial calcification (TAC) scores for patients in differing Rutherford categories, a measure of limb ischemia. Patients with no limb ischemia are in group 0 while those with the greatest level of ischemia are in group 6. Individual patients are shown as individual dots, with box and whisker plots that display the median, 25th, 75th, 10th, and 90th percentile estimates, summarizing the distribution of the nonzero TAC scores. One hundred twenty-three patients have a zero TAC score in Rutherford category 0, with fourteen in category 1-2, five in category 3-4, and none in category 5-6. As the Rutherford category increases, two effects are seen: the proportion of patients with nonzero values increases *and* the overall level of the TAC score increases. This mixture distribution of the TAC scores has an impact on the eventual analysis and modeling of the TAC scores, as standard methodologies must be modified to deal with the complexities in the distributions. Note the violation of summarization across a source of variation principle with the collapsing of the Rutherford categories from 7 levels to 4 levels. The presentation using the non-collapsed data was not quite as pretty or simple, as it show increases that were not as uniform. "It will be difficult for the doctors who read it to understand," said the investigator, himself an M.D. Why should we expect data from highly complicated biological processes to be simple and easy to understand?

On the opposing spread, a similar plot with a different rendering of the box-and-whisker plots, probably preferred due to its quieter nature, softer presentation, and stronger focus on the data themselves, except that the author accidentally hid median lines in the boxes!



Tufte presents six Principles of Analytical Design in a so-named chapter in *Beautiful Evidence*. They are worthy of exposition here, if only to entice the reader to obtain Tufte's fourth book and read of them fully. He discusses and describes them in detail when discussing Minard's map of Napolean's Russian campaign during 1812.

Here are his Principles of Analytical Design, which may or may not be Fundamental Statistical Concepts in Presenting Data. Regardless, the idea of analysis, as we have seen, is to understand the whole by decomposing into component parts. As such, these analytical design principles should be in play when we consider what statistical principles we seek to employ.

Principle 1: Show comparisons, contrasts, differences.

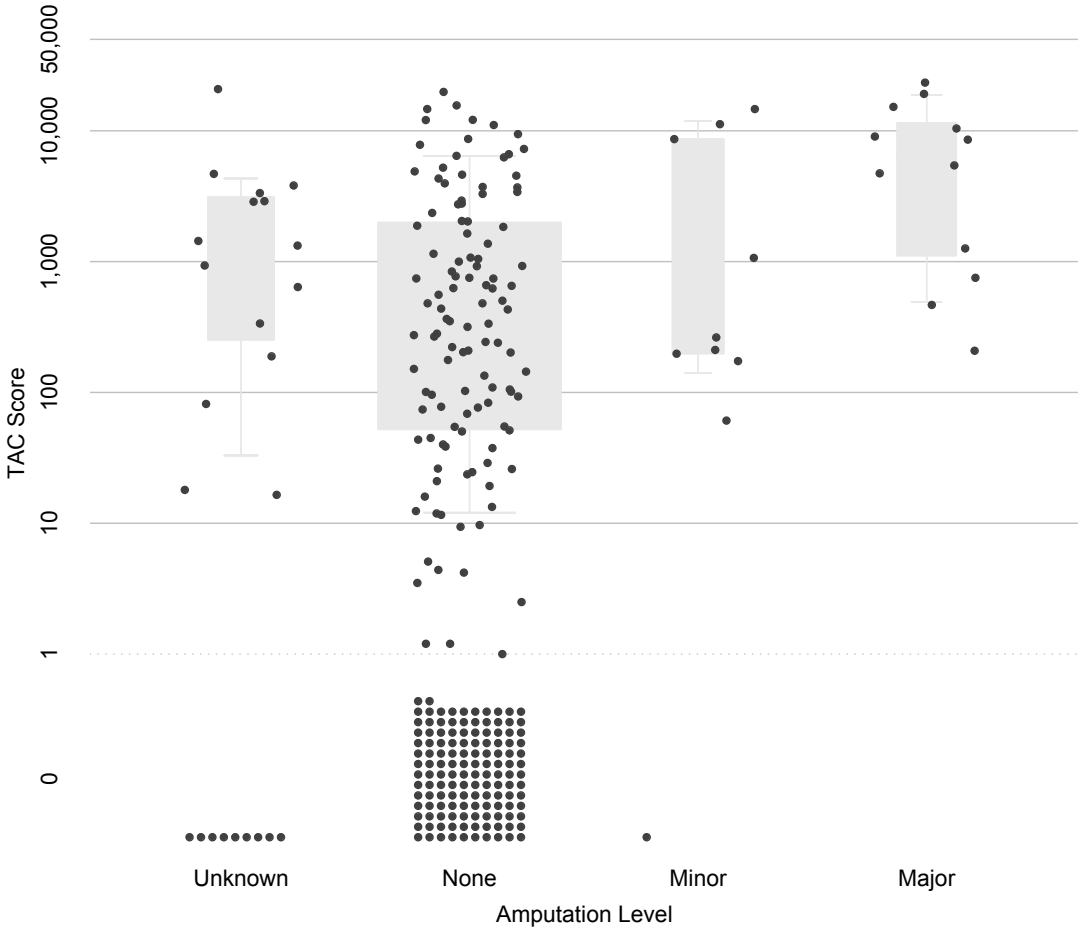
Principle 2: Show causality, mechanism, structure, explanation.

Principle 3: Show multivariate data; that is, show more than 1 or 2 variables.

Principle 4: Completely integrate words, numbers, images, diagrams.

Principle 5: Thoroughly describe the evidence. Provide a detailed title, indicate the authors and sponsors, document the data sources, show complete measurement scales, point out relevant issues.

Principle 6: Analytical presentations ultimately stand or fall depending on the quality, relevance, and integrity of their content.

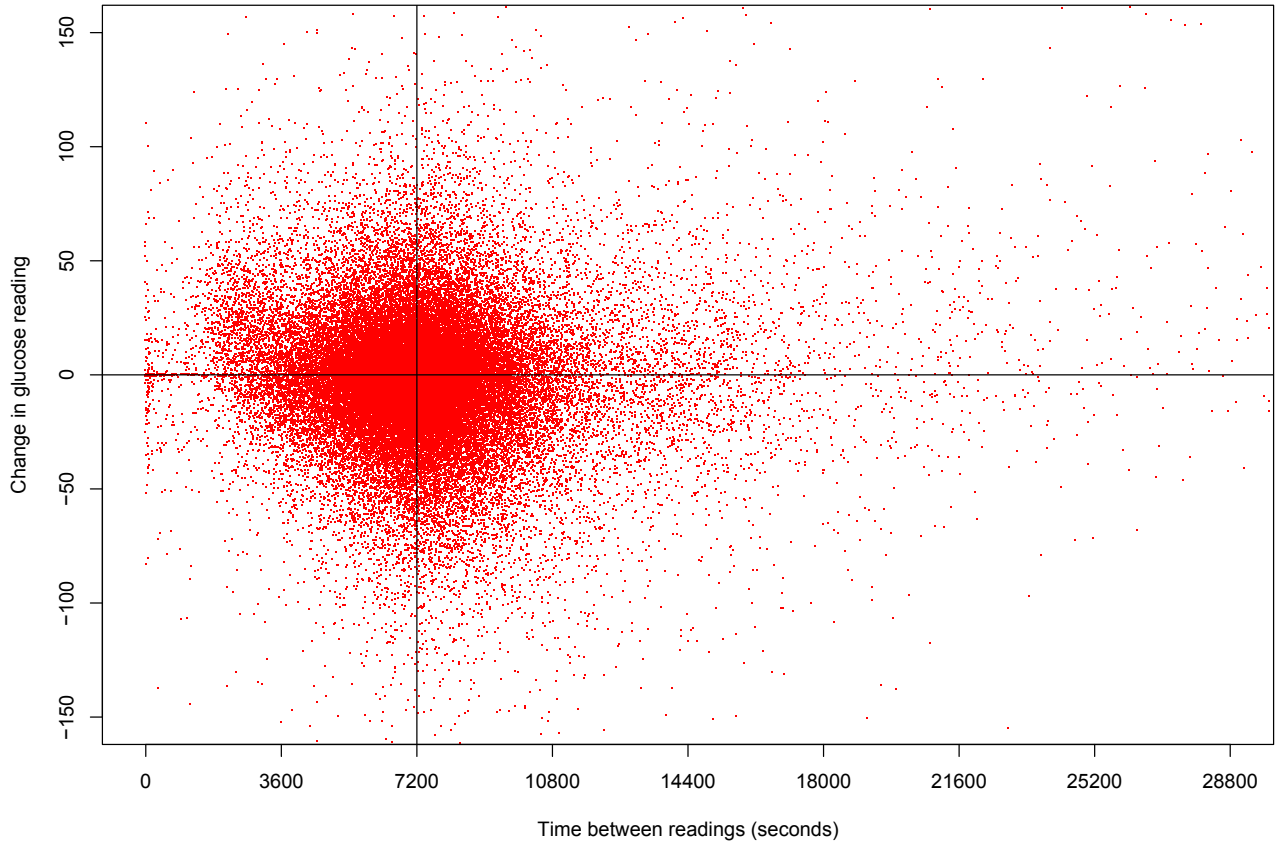


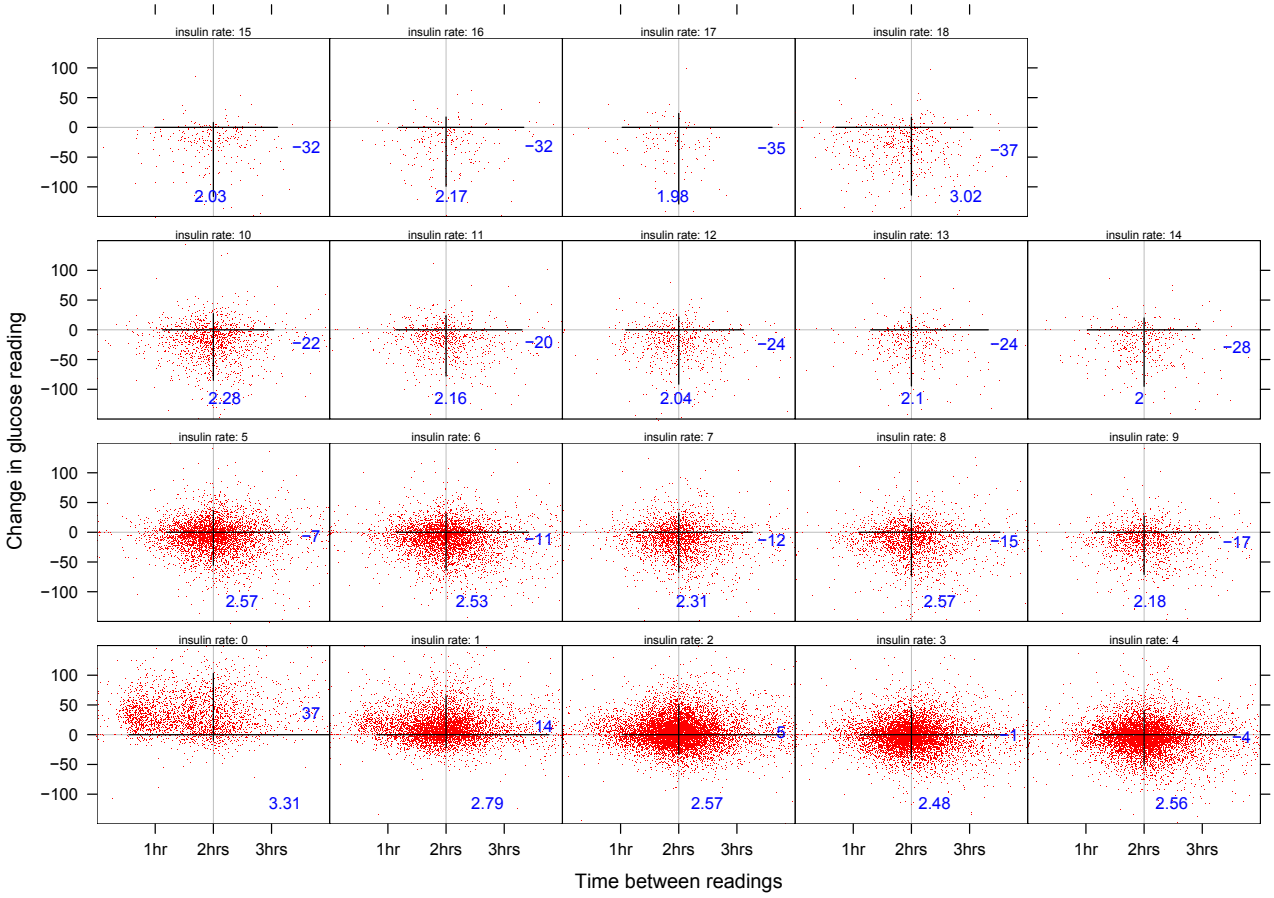
More on the blood-splatter plot. Below is the original change in glucose by time between readings plot, just the raw data atoms, no summaries. At right are two collections of small multiples, partitioning the grand plot below into a number of groups based on a covariate. Each panel in the small multiples plot shows the distribution along with the mean time and change value in blue (numbers are located at their value) and percentiles of the distribution shown by the dark lines overlaying the red dots.

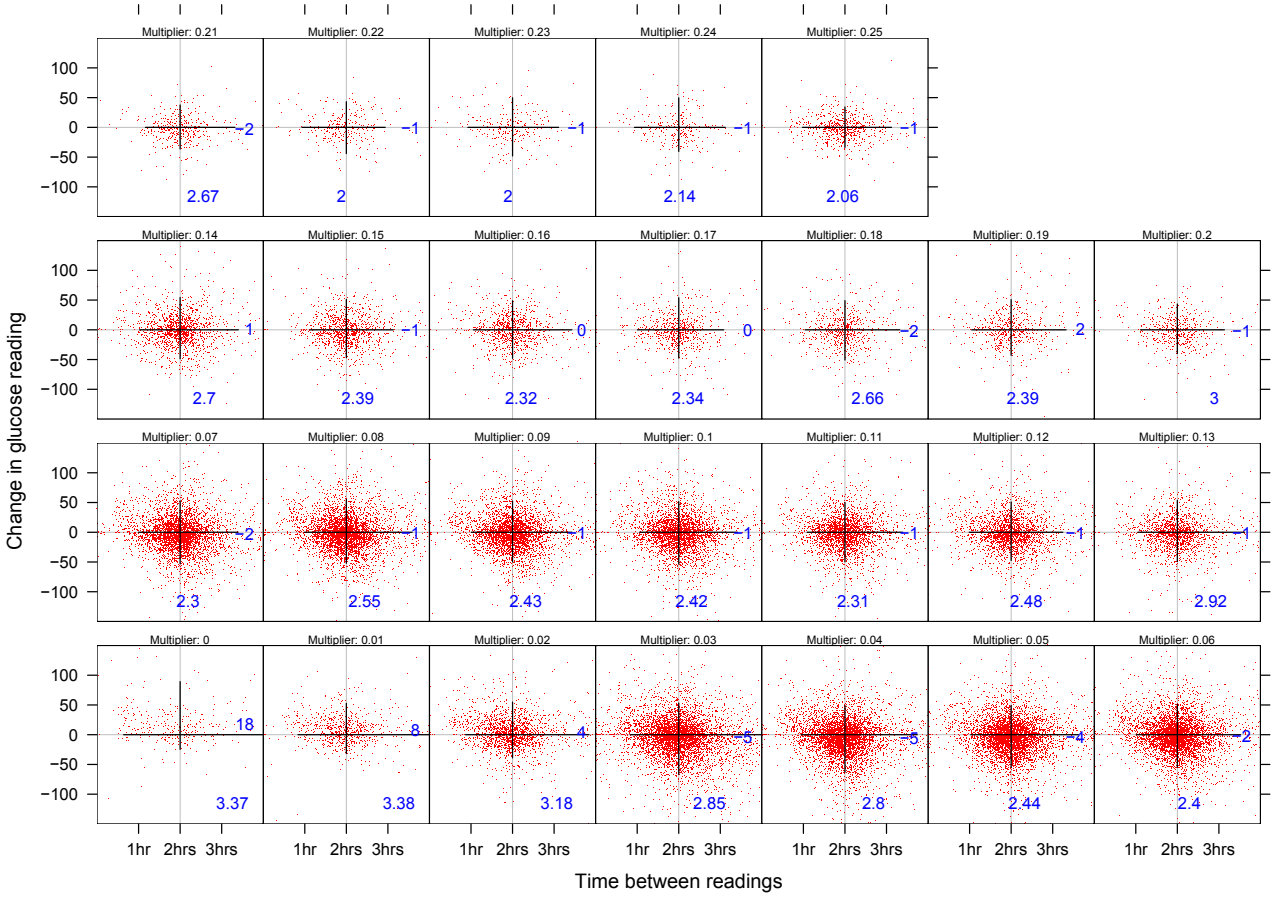
The collection on top splits on the basis of the amount of insulin given during the interval, rounded to the nearest unit. The collection on the bottom splits on the "multiplier", a parameter computed at each interval as part of the updating algorithm.

Note that the top plots show that insulin level is a source of variation in the change value, as the distribution moves down as the values of insulin move up. In contrast, the lower plots show little to no difference in change values as the multiplier changes.

Note also the covariance between the covariate and time: as the covariate value increases, the time between readings decreases.




insulin rate: 15
insulin rate: 16
insulin rate: 17
insulin rate: 18
insulin rate: 10
insulin rate: 11
insulin rate: 12
insulin rate: 13
insulin rate: 14
insulin rate: 5
insulin rate: 6
insulin rate: 7
insulin rate: 8
insulin rate: 9
insulin rate: 0
insulin rate: 1
insulin rate: 2
insulin rate: 3
insulin rate: 4
Change in glucose reading
Time between readings
1hr
2hrs
3hrs


Multiplier: 0.21
Multiplier: 0.22
Multiplier: 0.23
Multiplier: 0.24
Multiplier: 0.25
Multiplier: 0.14
Multiplier: 0.15
Multiplier: 0.16
Multiplier: 0.17
Multiplier: 0.18
Multiplier: 0.19
Multiplier: 0.2
Multiplier: 0.07
Multiplier: 0.08
Multiplier: 0.09
Multiplier: 0.1
Multiplier: 0.11
Multiplier: 0.12
Multiplier: 0.13
Multiplier: 0
Multiplier: 0.01
Multiplier: 0.02
Multiplier: 0.03
Multiplier: 0.04
Multiplier: 0.05
Multiplier: 0.06
Change in glucose reading
Time between readings
1hr
2hrs
3hrs

Notes