VARIABILITY: ONE STATISTICIAN'S VIEW

ROBERT GOULD
Department of Statistics, UCLA
rgould@stat.ucla.edu

SUMMARY

Although variability is of fundamental concern and interest to statisticians, often this does not get communicated to students who are taught instead to view variability as a nuisance parameter. A brief survey of a few case studies, as well as a recounting of some history, shows that variability is worthy of study in its own right, and examination of variability leads to insights that might have been missed had we focused all of our attention on the "trend" of the data. As one of the key components of statistical thinking, variability deserves more prominence in the classroom.

Keywords: Variability; Education; Statistical Thinking

1. INTRODUCTION

Many of the papers in this special issue discuss naive conceptions of variation. Much has been written on more experienced conceptions, particularly in the context of statistical thinking. Moore (1990), for example, puts variation at the heart of the process of statistical thinking and addresses the needs of statistical thinkers to acknowledge the omnipresence of variation, to consider variation in collecting data, to quantify variation, and to explain variation. Wild and Pfannkuch (1999) provide a thorough overview on the topic, placing variation in the context of a rather rich model of statistical thinking. When educators think about how to move students from their naive conceptions towards a more "professional" view, an understanding about how practitioners confront variation should be useful, and it is hoped that this paper provides some examples that will aid in the process. Statisticians are themselves a variable crew and these remarks should not be taken as a summary of the Profession, but instead as the thoughts of one practitioner.

It is fair to say that statisticians have a complex relationship with variability. Statisticians sometimes attempt to minimize variability, sometimes to maximize, sometimes to estimate or simply to "analyze" variance. Many statistics educators claim variability to be one of the fundamental concepts of statistics. (For example, Moore, (1990) and Snee, (1990).) Yet when most students first encounter statistics, they find that variability plays second fiddle to "central tendency." The conceptualization of data as "signal versus noise", which according to Pfannkuch (1997) some statisticians consider one of the major contributions of Statistics to Science, teaches students that the central tendency (however it's measured) is of primary importance and variability is simply a nuisance. A noisy one at that.

College level statistics does not completely ignore variability, of course. Many texts and one hopes many instructors discuss the importance of examining the shape of the distribution before making any conclusions about the data. DeVeux, Velleman and Bock (2004) write in their introductory statistics textbook that "the three rules of data analysis are 1) make a picture 2) make a picture and 3) make a picture." Most students learn, often in the first weeks of the course, that the mean by itself is not a sufficient summary of a distribution. But after that variability is brushed aside as attention focuses on estimating the mean, and students are

taught that standard deviation is a nuisance parameter that must be estimated if one is to do a proper hypothesis test on the mean or calculate a confidence interval for the mean or perform a comparison of means. Some introductory courses teach ANOVA, which although it pays tribute to variability in its name, is really about the simultaneous comparison of means from several populations.

The definition for variability used in this discussion is derived from Moore's definition of data analysis as "the examination of patterns and striking deviations from those patterns" (1997). Although Moore was describing the activity of data analysis, as a by-product he provides a wonderfully general definition of variation; variation is that which is not pattern. A case study in Section 2 will illustrate how useful and rich variability, using this broad definition, can be. A short examination of the history of statistics in Section 3 shows that Variation and Center have a long-standing rivalry. Finally, Section 4 illustrates the role variability plays in the analysis of three small data sets.

2. AN EXAMPLE OF REASONING WITH VARIABILITY

In 1982, Morton et. al. conducted a study to determine whether parents who worked around lead could expose their children to dangerous amounts of lead. Lead poisoning is particularly dangerous for children because excess levels of lead interferes with a child's development. For example, lead-based paint is no longer used in the interior of homes because children might ingest flakes of paint. But adults who work around lead might get sufficient amounts of lead dust on their skin or clothing to pose a hazard to their children.

The data consist of the blood lead levels (measured in micrograms per deciliter) of 33 children whose parents worked at a battery factory in Oklahoma and whose work exposed them to lead. We also have the lead levels for a control group consisting of an additional 33 children, matched for age and neighborhood, whose parents did not work around lead. (The data themselves are taken from Trumbo, (2001).)

The "research question" is phrased so as to invite a comparison of means. Is the typical lead level of the exposed children higher than the typical lead level of the control children? It is instructive to imagine a world in which means or medians are computed only as a last resort, and attempt to answer this question without relying on the concept of central tendency. Consider the histograms of the lead levels of the two groups of children (Fig. 1) below, which are presented with only frequency information. One displays the distribution of the control group, the other the exposed group. Can you tell which is which?

Even though one does not know the scale (actual values on the x-axis) of each group, nor their means or spread, it is fairly straightforward to identify the right-skewed histogram (on the left) as belonging to the exposed group. The reason is that the shape of the sample distribution is consistent with our theory of how the exposure happens; children receive varying amounts of additional exposure beyond the "normal" exposure represented by the control group. Thus, shape, even in the absence of information about absolute values, contributes important information that helps us in making sense of the data in light of the research question.

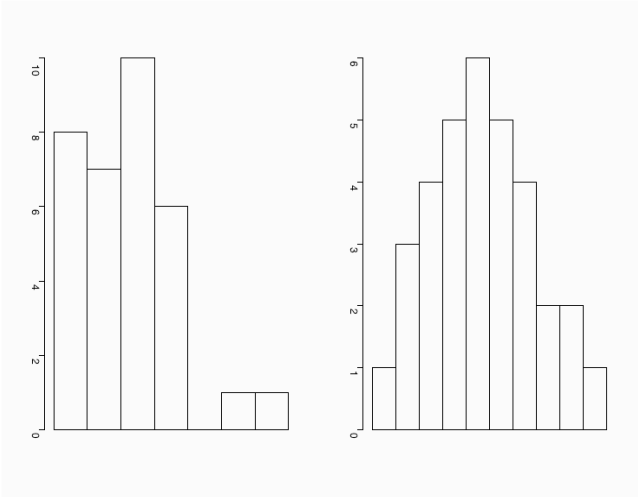

Figure 1. Lead levels in blood of 33 childred. Which histogram comes from the exposed group?

Of course, knowing the actual values (shown in Figure 2) and through them the means of the two groups contributes additional understanding. Indeed, the numeric values themselves carry quite a bit of the story. Most medical experts find lead levels of 40 mg/dl and above to be hazardous, and levels of 60 and above to require immediate hospitalization. Without any significance testing, it's quite apparent that the means are different and that the exposed group is dangerously higher.

At its heart, though, this study is concerned with a causal question: did the parents' exposure at work cause their children's elevated levels? While the difference in means establishes that the causal question is worth entertaining, it provides little evidence towards answering the question. Because this is an observational study, not a controlled experiment, definitive causal conclusions are impossible. However, I propose that the difference in shape between the two groups, while by no means confirming causality, by itself takes us closer to a causal conclusion than would a consideration of the means alone. The reason for this is that the proposed mechanism for the children's contamination had consequences for both the center and the shape of the distribution of lead levels. Had the centers differed but the shapes been similar we would not have been inclined to believe that the parents' exposure could have been the cause, but might instead suspected, perhaps, deliberate poisoning. On the other hand, had the centers for both groups been the same but the shapes appeared as they do here, then we would still have had strong reason for suspecting that parents' exposure to lead was a threat to children.

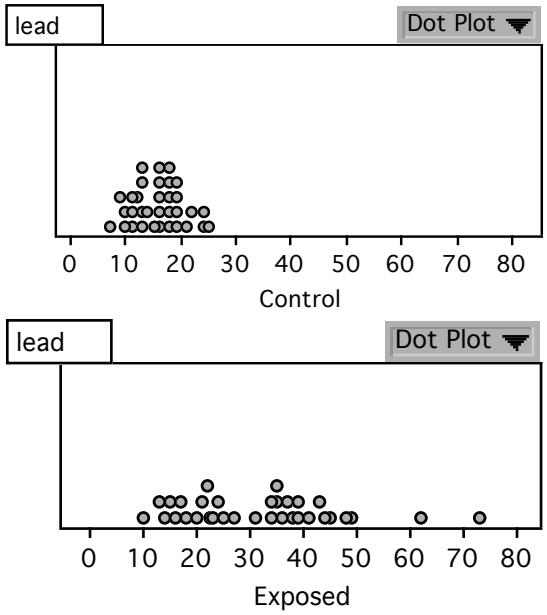

Figure 2. Blood lead levels for both groups of children.

The interplay between the center and the spread of distributions of data is, of course, not new. A brief survey of the history of modern statistics will show that there has long been some sort of "tension" among data analysts as to the proper roles for center and spread.

3. HISTORICAL OVERVIEW

I make no claims to being a historian, and I use no primary sources for this survey. Unless otherwise noted, the 19th century history is a paraphrase of Gigerenzer, Swijtink, Porter, Daston, Beatty, & Kruger, (1997) and the history of ANOVA is from Searle, Casella, & McCulloch, (1992).

Adolphe Quetelet, impressed by growing evidence of "statistical regularities" in practically every aspect of life that governments measured in the early 1800's, founded "social physics" to determine laws that governed society that would be analogous to those laws that governed the motion of the planets. Some examples of statistical regularities, such as the fairly constant ratio between male and female births, were well known at this time and not too surprising. But other regularities were more surprising because they seemed to suggest an implicit order arising from chaos. Examples of such regularities include Laplace's demonstration that the number of dead letters at the Paris postal system was fairly constant, as well as those later "discovered" in the homicide, crime, and marriage rates published in the 1827 volume of the French judicial statistics.

Quetelet, along with others, believed that these regularities were not just descriptions of societies, but instead represented some underlying "reality" about society. He invented *l'homme moyen*, the Average Man, to be an abstraction of the typical member of a society. While there might not be laws that governed individuals, there were laws, Quetelet believed, that governed the behavior of the Average Man. Hence the Average Man was not a mere description of a society, but something more real.

Quetelet's views were influential throughout Europe for much of the 19th century. Florence Nightengale corresponded with him and called him the founder of "the most important science in the whole world" (Coen, (1984)). Adolph Wagner, a German economist,

believed the power of statistical regularity was so strong that in 1864 he compared it to a ruler with power so great that it could decree how many suicides, murders, and crimes there would be each year in the kingdom. (But, of course, this ruler lacked the power to predict precisely who would commit, or fall victim to, these acts.)

The Average Man was a useful tool because he could be used to compare traits of different cultures and, presumably, his behavior (or more accurately, his propensity towards certain types of behavior) could be predicted. However, he had troubling implications for the concept of free will. If there were a quota for murders, by what mechanism were people compelled to fill it? Surely people could choose to not fill the quota, if they wished. Interestingly for our purposes, by arguing in support of the existence of free will in society, critics of the Average Man also argued in support of elevating the use of variability in statistical analyses of social data.

Gustav Rumelin claimed that variability was a characteristic of the higher life forms and reflected autonomy. Humans have free will and are therefore more variable, presumably, than single-celled organisms. If humans were homogenous, than Statistics would be unnecessary, and therefore social statistics should study variation rather than simply reporting the average. Wilhelm Lexis, a German social statistician, studied the annual dispersions of these so-called statistical regularities and compared them to chance models. In almost every case he found that the observed dispersions were greater than that predicted by his chance models, and used this to conclude that the existence of free-will prevented the existence of statistical regularity and, therefore, studying averages of populations was a waste of time.

A generation later, in 1925, R.A. Fisher invented ANOVA to cover the need for "a more exact analysis of the causes of human variability." Ironically, ANOVA really tends to treat variability as a nuisance and its main focus, once one is satisfied that the variance is behaving, is to concentrate on comparing means. Nonetheless, once ANOVA was later framed in the context of the linear model, it became possible for researchers to model and investigate variance components directly.

While this historical overview is a selective review, it is clear that discussions of the interplay between measures of variability and measures of center have long been a part of modern statistical development. Social scientists' struggle to understand how, when, and whether to assess variability has been complex and at times controversial, but valuable. Data analysts have learned that one must consider both the center and the variability in order to understand scientific and social phenomena.

4. CASE STUDIES

The following three case studies should illustrate how and what data analysts learn from variation. While the data presented here are real, the analyses are not. The analyses are meant to be instructive and are not necessarily what a data analyst would actually perform.

4.1 CASE STUDY 1: UCLA RAIN

Time-series analyses are notoriously signal-noise oriented and provide a good opportunity to examine the role of variability in a context where one might think that role is secondary. This series is monthly rainfall at the campus of the University of California, Los Angeles from January 1936 to the end of June 2003. A typical goal of an analysis of data of this type might be to model the trend so that predictions for the future could be made, but here we focus more on how our view of the trend changes as we re-organize the data.

The three graphs below show, respectively, the time-series with the overall average superimposed (top), rainfalls organized by month (middle) and a smoothed time-series showing total annual rainfall.

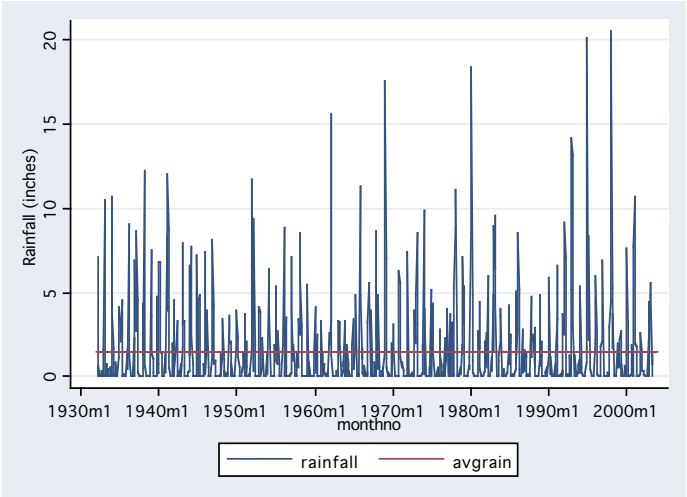

Figure 3. Rainfall in inches at UCLA

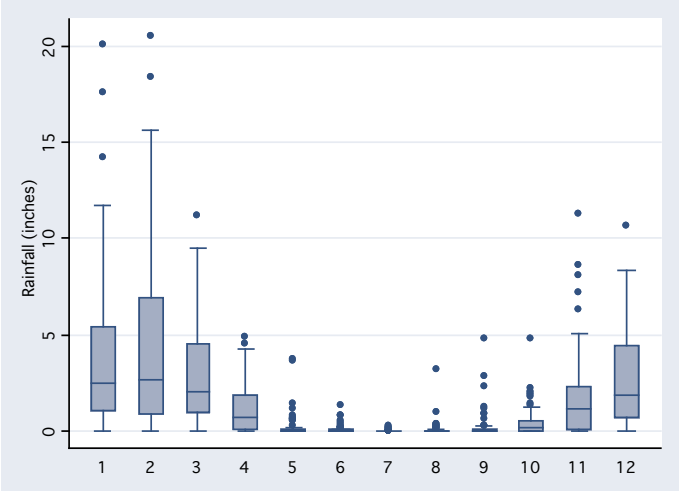

Figure 4. Rainfall by month.

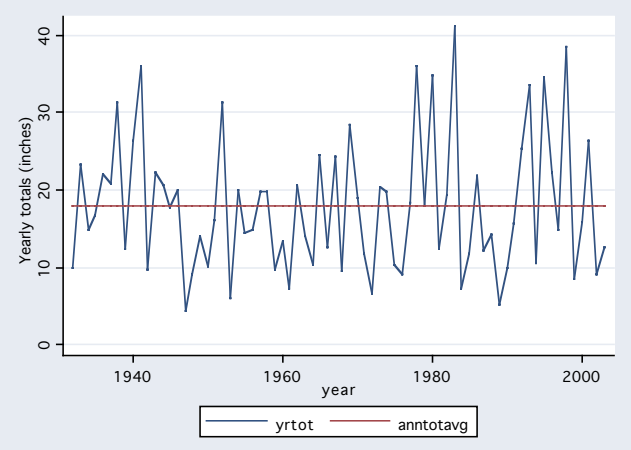

Figure 5. "Smoothed" rainfall, with average annual rainfall.

The unprocessed time-series (top) impresses mainly by its unruliness. The second graph shows a pattern any southern Californian would recognize: wet winters, dry summers. The third graph shows an historic trend and one can, for example, search for evidence of drought. These last two graphs illustrate the relationship between center and variation suggested within Moore's definition of data analysis. Variation is defined in contrast to pattern. In the second graph, we can see variation within a particular month; for example we can see that there was a particularly rainy September with almost 5 inches of rain. (We also notice that this amount is substantial for any month, but is particularly heavy for September.) We can also see variation across months which provides an understanding of seasonal fluctuation. The third graph shows us annual variation with respect to an overall mean, which might be of interest to farmers or climatologists.

4.2 CASE 2: LONGITUDINAL DRINKING PATTERNS

Do people drink less as they age? This was a question investigated by Moore, et. al. (2003). Drinking patterns are fairly complex in that drinking varies quite a bit from person to person, and individuals vary their drinking from year to year. One difficulty in such an analysis is in isolating the effects of cohort (people born at a certain historic time might share drinking patterns) and period (changes in price or supply might affect the entire population at a certain point in time.)

The data come from the 1971, '82, '87, and '92 waves of the NHANES study, a national, longitudinal random sample of about 18000 U.S. residents (NCHS 1971, 1987, 1990, 1992, 1994). Subjects were asked questions about the quantity and frequency of their drinking, and responses were converted into a "quantity/frequency index" (qfi) that corresponds approximately to the average number of drinks per week. Figure 6 shows what 1971 looked like. The story, once again, is in the variety of drinking. (Note the outlier above 80.)

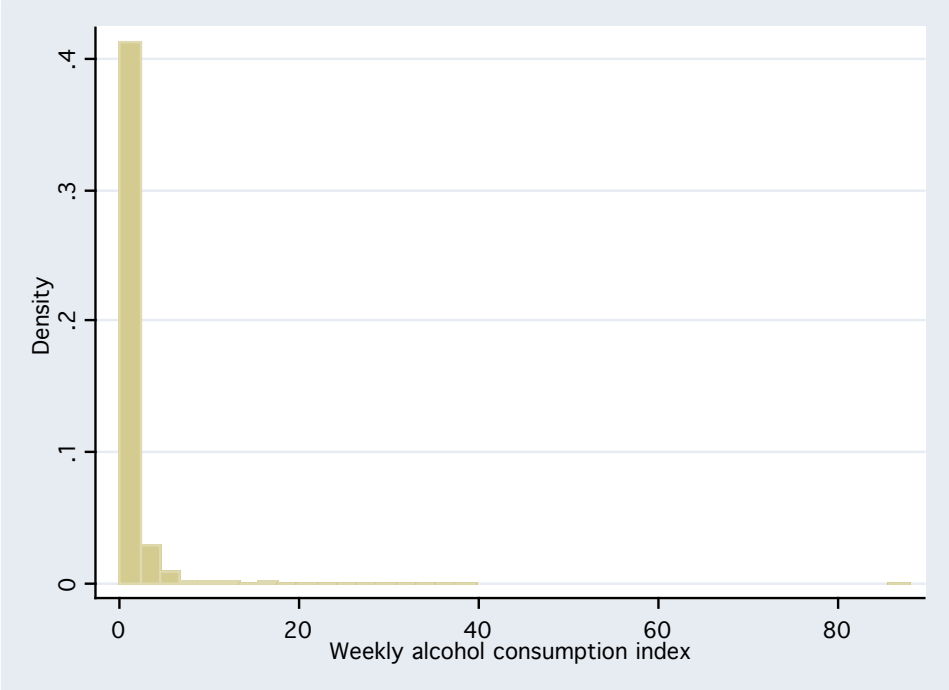

Figure 6. Drinks per week of a national sample in 1971.

We see that the vast majority drinks little, but a minority drinks very much. The shape of the distribution is interesting in that it tells us that a simple model, in which we look for

"typical" drinking with some people deviating from the norm, will be inappropriate. At the very least a log transformation of the data is necessary, and thus our consideration of variation has affected our conception of the model.

The purpose of this study was to examine drinking over time. Figure 7 shows that drinking did in fact vary at the different waves of the survey.

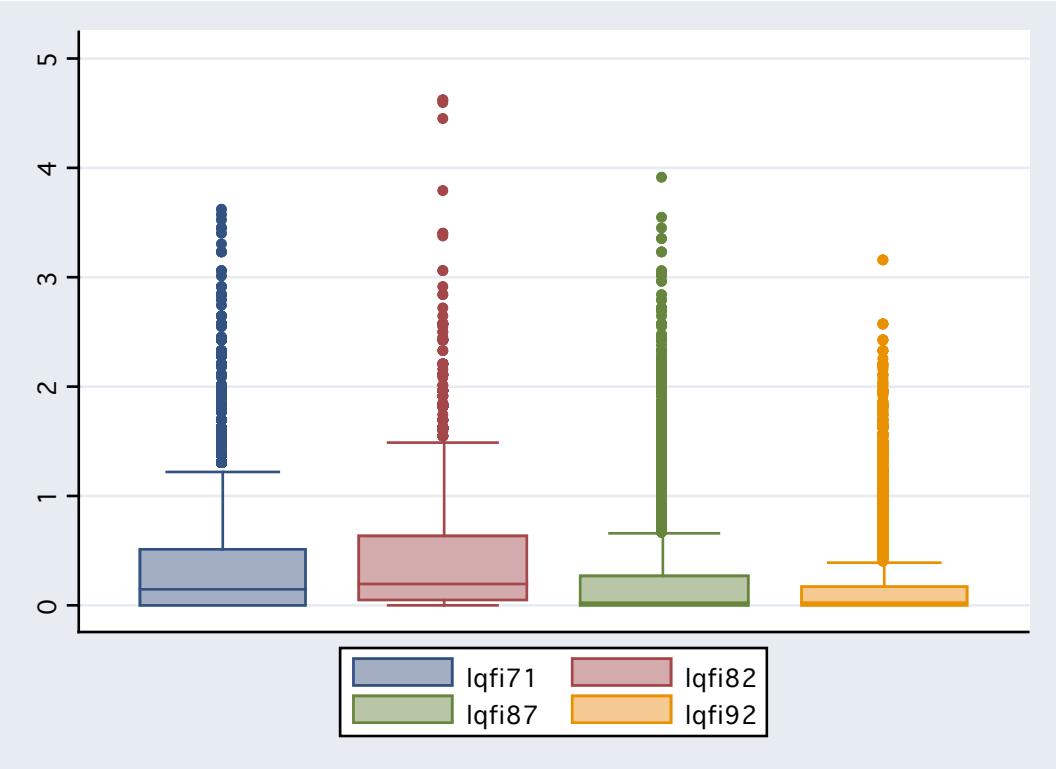

Figure 7. Drinking index at each wave of the survey.

One is compelled, at this point, to attempt to explain the variation with some sort of model. As a first cut, many find it useful to talk about two types of variation: explained and unexplained, or if you prefer, deterministic and stochastic. (Wild and Pfannkuck (1997)) Deterministic variation is that which we believe will have a regular structure, a structure that can be defined by a model. What is left over is then stochastic variation.

Let's consider a very simple model in which the only explanation for variation is time. We would then have log(qfi+1) = .97 - .017*year as one model for the deterministic variation; we could conclude that drinking amounts decreased slightly over time, while acknowledging that there was still quite a bit of variance left unexplained. A more complex model would then chip away at the unexplained variation bit by bit. For example, another source of variation is the individual; different people drink different amounts and will change differently over time. We could then fit a mixed linear model in which each individual is allowed his or her own slope and intercept (Laird & Ware, 1982). The variation is now much more complicated; we have variation with respect to each individual's path as well as variation between individuals. Examination of these different sources of variation might lead to further refinements of the model and force us to consider such questions as whether observations within individuals are independent and whether slopes are correlated with intercepts.

One potential deterministic model for these data that includes age, cohort, and period explanations for variation is log(qfi+1) = .4 - .13*(age in decades) + .18*(per capita consumption in alcohol) + .035*(birthyear times age in decades). This model, if valid, suggests that drinking declines with age, across all generations and periods of (recent)

history, but the decline depends on when a person was born. Those born more recently decline more slowly. (We used per capita alcohol consumption to control for historic variations in drinking. So this model says that even in times in which the country as a whole drank more (or less), individuals on the average declined as they aged.)

Although the end result of this analysis is a model for the trend, the model has been shaped and refined by our conceptualization of the variation.

4.3 CASE 3: CHIPMUNKS

About 15000 years ago, during the Wisconsin glacial stage, chipmunks that lived in the pinyon-juniper woodland in the Mojave Desert region were able to move from mountain to mountain, since the cooler temperatures allowed their pinyon and juniper trees to grow in the basins between mountains. Later, these woodland areas retreated to higher elevations and with them the chipmunks. This resulted in isolated chipmunk communities. Kelly Thomas, a graduate student at UCLA, wanted to study morphological differences in separated chipmunk populations. She captured several chipmunks at six different sites, took five morphological measurements of each chipmunk, and wanted to compare them to see if there were differences in size and shape at different sites. This is a fairly common activity for population biologists. They seek to quantify the shape of animals using, ideally, a small set of numbers.

Principal Components Analysis (PCA) is a data reduction technique that focuses on the co-variation of multi-variable samples. We use it here to attempt to reduce the number of variables we'll use to compare the chipmunks from five down to two. PCA does this by creating a set of linear combinations of the original data that maximize the variation and are orthogonal to one another. The reason for maximizing the variance is so that the resulting set of measurements will have the greatest possible dispersion. This, in turn, will make it easier to distinguish between populations of chipmunks. This is analogous to writing an exam to distinguish those who learned the material from those who did not. If everyone receives approximately the same score, it is difficult to distinguish those who really understood the material.

In the following analysis, each chipmunk provided two scores. Each score was a different linear combination of its mass, body length, tail length, hind-foot size and ear length. The first score emphasized overall mass and the second corresponded roughly to shape. Although the procedure was not entirely successful with these data (the two scores accounted for only 50% of the total variation), we gained some insight into the data. First, ear measurements were not strongly correlated with any of the other measurements. On the other hand, chipmunks with long bodies tended to have long tails, and those with big hind feet tended to have the greatest mass. More interestingly, by plotting the two scores for all chipmunks, we were able to discern that chipmunks from the same sites tended to have similar scores, which provided evidence that these scores could be useful for distinguishing chipmunks from different regions.

This procedure is often used for exploratory analysis. This example used PCA to informally assess similarities among chipmunks in similar sites. Interestingly, we did this by dealing directly with the variation and covariation among the variables.

5. CONCLUSION

This paper has presented several examples illustrating how a statistician thinks about variability. It falls on educators to consider how conceptions of variation aid or hinder how students learn statistical thinking.

Wild & Pfannkuch mention imagination as one of eight "dispositions" that statistical thinkers possess. My belief is that variation is the fuel to statistical imagination. The case studies presented above illustrate, to various degrees, how consideration of variation drives the analysis by provoking the statistical imagination to explore alternative models.

Statistical imagination begins when variation is observed. When confronted with a time-series, one intuitively seeks for some sense of order out of the chaos. Might some of the "noise" be abated by removing seasonal effects? Monthly effects? We can aggregate the data in different ways to get different pictures of the "trend", but in all ways, we are exploring different models of the variation in order to better observe the trend.

By defining variation, we define trend. The process of modeling variation is made explicit in time-series analysis, in which analysts consider explicit structures for variation to take into account local correlations between observations. Modeling variation and covariation is also done explicitly when analyzing longitudinal models. The alcohol consumption case study discusses implications of including different sources of variation in the model. For example, if one thinks (quite reasonably) that a source of variation is due to different drinking behaviors between individuals, then one has a more complex model than if one simply models the population as a homogeneous mass. Analysts must explicitly build other assumptions about variability directly into the model. How do subsequent observations within a person correlate? How do people within a subgroup co-vary with each other? How do basic parameters of the model -- for example the rate at which people change their alcohol consumption as they age -- correlate among individuals? The statistical imagination, guided by expert substantive opinion, shapes the model through consideration of variation.

Models play an important role in statistics, and one hopes that introductory statistics students, at least at the college level, learn not only to interpret basic statistical models, but also to develop an understanding of the sometimes tenuous relationship between the model and reality. There is however a danger, I believe, in over-emphasizing models to beginning students.

Chatfield (1988) distinguishes between "confirmatory" and "exploratory" analyses, and it is models used in confirmatory analyses that I think should be de-emphasized. Confirmatory models harden the boundaries between "signal" and "noise". The conceptualization of data as "signal and noise" is of course very important, but perhaps the value-laden language creates too much of a sense of finality and inhibits statistical imagination in students. *This* is signal, but *this* is noise, and one should not pay attention to noise. In practice, one *does* pay attention to statistical noise. Confirmatory models do provide the all-important p-value, but students who possess (or are taught to possess) statistical imagination will see that models can be used for exploration and insight without necessarily needing that final confirmation step. Principal components analysis is one example in which the variation is the primary focus of the analysis, and, at least in the chipmunk case study above, no model is produced or required. (I don't mean to overstate my case here. There is a model, in a general sense, underlying the analysis. But it is used as just one step in an exploration, not as a final statement about the structure of the data or reality.) The lead case study is a good example of how the "signal" by itself (or at least a narrowly defined signal) provides an incomplete analysis and, in fact, isn't sufficient for answering basic research questions concerning the effects of lead exposure on children.

An important reason for focusing students' attention on variation is to encourage them to think not in terms of procedures ("Which test do I apply here?") but instead to exercise their

statistical imaginations in order to understand the real issues behind the data. Often this translates to a search for causes of variability. Statistics courses perhaps give short shrift to the problem of inferring causality and are known for merely cautioning students against making conclusions about causality based on association. But students want, and maybe even need, causal explanations.

Pearl (2000) makes the point that causal explanations make an early appearance in the Bible. "Did you eat that apple?" God asks Adam. "Eve made me" is Adam's answer (greatly paraphrased). Causality makes for good story-telling, and links data to reality. The search for causal explanations can lead to heightened statistical imagination, and so students should be given the opportunity to reason, for example, about how an intervention will affect the shape of a distribution. Assuming that exposure to lead does lead to higher lead levels, why are the shapes of the lead distributions in Figures 1 and 2 natural?

If our primary goal is to teach statistical thinking (rather than statistical techniques), then we should look to the noise, and not the signal.

ACKNOWLEDGEMENTS

Many thanks to the direction of Joan Garfield and encouragement of Dani Ben-Zvi. Special thanks to J. Murakami, Department of Atmospheric Sciences, UCLA for the rain data. I'd like to dedicate this paper to the memory of Winifred Adams Lyon.

REFERENCES

Chatfield, Christopher (1988). *Problem Solving: A statistician's guide.* Chapman and Hall.

Coen, I. B. (1984). Florence Nightingale, *Scientific American 250*, 128-137.

DeVeaux, R., Velleman, P., Bock, D., (2004). *Intro Stats*. Addison Wesley.

Gigerenzer, G., Switjtink, Z., Porter, T., Daston, L., Beatty, J. & Kruger, L. (1997). *The Empire of Chance: How probability changed science and every day life.* Cambridge University Press.

Laird, N.M., Ware, H. (1982). Random Effects Models for Longitudinal Data, *Biometrics, 38,* 963-974.

Moore, A., Gould, R., Reuben, D., Greendale, G. , Carter, K., Zhou, & K., Karlamangla, A. (2003). Do Adults Drink Less as They Age? Longitudinal Patterns of Alcohol Consumption in the U.S., to be published in *American Journal of Public Health.*

Moore, D. (1997) Probability and Statistics in the Core Curriculum. In J. Dossey (Ed.), *Confronting the Core Curriculum*. USA: Mathematical Association of America.

Moore, D. (1990). Uncertainty. In L. Steen (Ed.), *On the shoulders of giants: new approaches to numeracy*. USA: National Academy Press.

Morton, D. (1982). Lead absorption in children of employees in a lead-related industry, *American Journal of Epidemiology, 155*, 549-555.

National Center for Health Statistics. Plan and operation of the National Health and Nutrition Examination Survey, United States. 1971-73. *Vital Health Stat [1].* 19773;10a. DHEW publication PHS 79-1310.

National Center for Health Statistics. Plan and operation of the National Health and Nutrition Examination Survey, United States. 1971-73. *Vital Health Stat [1].* 19773;10b. DHEW publication PHS 79-1310.

National Center for Health Statistics. Plan and operation of the NHANES I Epidemiologic Follow-up Study 1982-84. *Vital Health Stat [1]*. 1987;22. DHHS publication PHS 87-1324.

National Center for Health Statistics. Plan and operation of the NHANES I Epidemiologic Follow-up Study 1986. *Vital Health Stat [1]*. 1990;25. DHHS publication PHS 90-1307.

National Center for Health Statistics. Plan and operation of the NHANES I Epidemiologic Follow-up Study 1987. *Vital Health Stat [1]*. 1992;27. DHHS publication PHS 92-1303.

National Center for Health Statistics. Statistical Issues in Analyzing the NHANES I Epidemiologic Followup Study. *Vital and Health Statistics, Series 2: Data Evaluation and Methods Research No. 121*. Hyattsville MD: National Center for Health Statistics, 1994. DHHS Publication No. (PHS) 94-1395.

Pearl, Judea (2000). *Causality: Models, Reasoning, and Inference.* Cambridge University Press.

Searle, S.R., Casella, G., & McCulloch, C.E. (1992). *Variance Components*, New York: John Wiley & Sons.

Snee, R.D. (1990). Statistical Thinking and Its Contribution to Quality. *The American Statistician, 44*, 116-121.

Thomas, K. (2002). Effects of habitat fragmentation on montane small mammal populations in the Mojave National Preserve (Unpublished report for UCLA OBEE 297A, UCLA, USA, 2002).

Trumbo, B. (2001). *Learning Statistics with Real Data*, San Francisco: Duxbury Press.

Wild, C.J., Phannkuch, M. (1999). Statistical Thinking in Empirical Enquiry, *International Statistical Review 67*, 3, 223-265.

ROBERT GOULD
Dept. of Statistics
8130 Math Science Building
MC 155404
UCLA
Los Angeles, CA 90095-1554