Quantitative Geographical Analysis
Week 3: Three Givens.
Background paper for February 5, 2001 seminar discussion on spatial data.
Elvin Wyly

1. Introduction.

In most scholarly, policy, and popular circles, the meaning and implications of "data" considerations are too often ignored, downplayed, or misinterpreted. What we want to explore this week is a deep sensitivity to considerations of data. Far more than a mundane detail ('get the data and run the numbers') or a naive, inductive empiricism ('what do the numbers say?'[1]), data represent the abstract codification of very specific decisions of what to observe, of how to measure and record it, at what time and what place, and what it ultimately means. That sounds like a lot of useless verbiage, but hopefully this background paper will make it a bit clearer, and it will become even more evident as you look through the contributions of Fotheringham, Anderson and Feinberg, Longley, Gould, and Cochrane.

2. Data.

First consider this. Crack open the dictionary, look up the word "data," and here's what you find: **data** *pl* of DATUM. Fine. Customer service ain't what it used to be. Go to DATUM. **datum** n. a known fact || the assumption which forms the basis for an inference or a conclusion || a starting point from which, e.g., a survey is made.

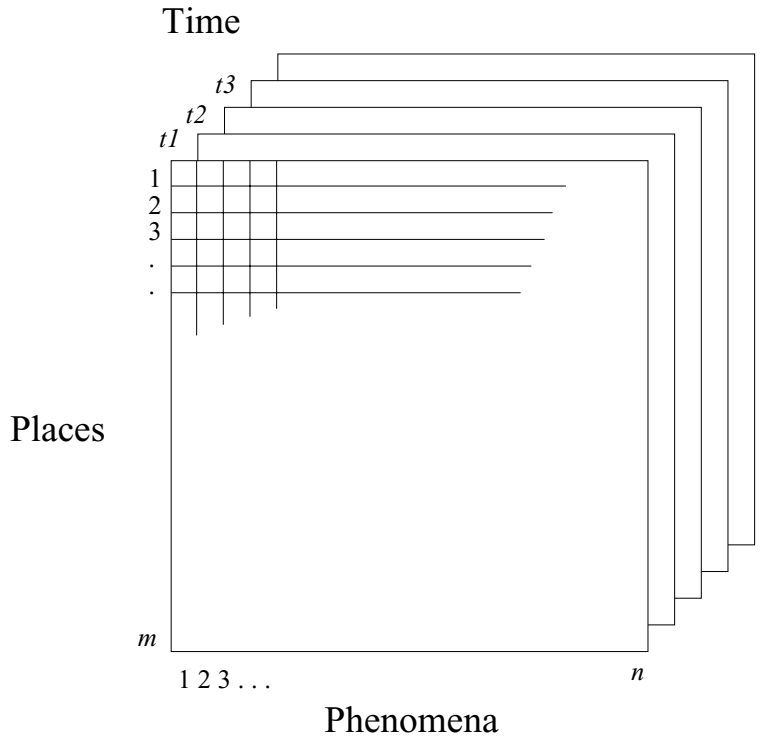
The word is Latin for "given."

What, then, are the implications of these three givens for the endeavor of scientific inquiry? It is all too clear from our engagement last week with the epistemological debates of the Quantitative Revolution that there is deep disagreement over what "known facts" count, which assumptions should be adopted, and what starting point we should use for our inquiry. How we think about gathering information reflects, and in turn, reinforces our understanding of the similarities and differences among phenomena in different places. In the mid-1960s, Brian Berry proposed the conceptual scheme of the "data cube" as a starting point for thinking about empirical observation (see next page):[2]

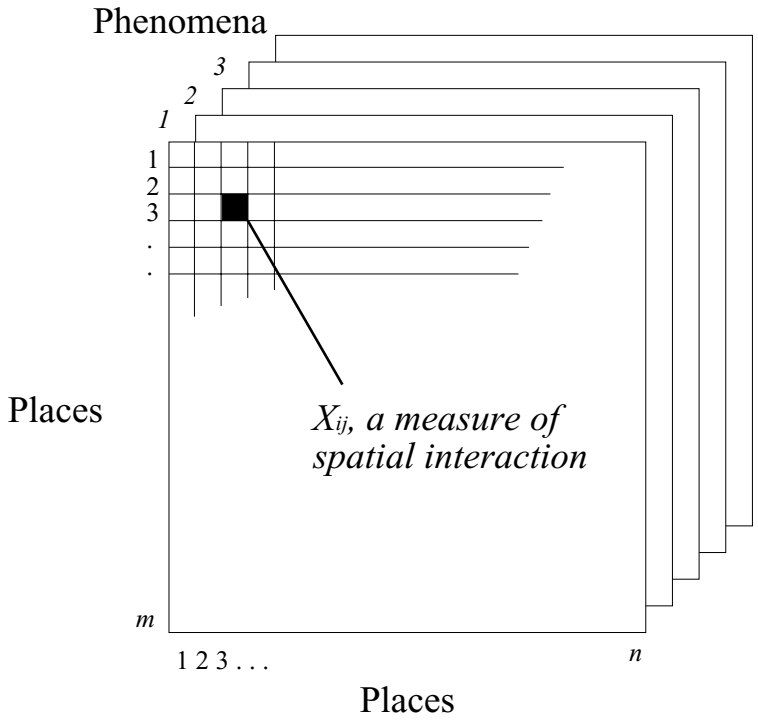
---

[1] Numbers really don't say anything at all, do they? They are imbued with meaning only when set in the context of human understanding and theoretical constructs. Peter Gould's "Letting the Data Speak for Themselves," which we read last week, was deliberately titled to evoke a reaction. It is not enough to come to 'the data' for a naive fishing expedition. But it is also risky to impose rigid, mechanistic understandings -- particularly of human interactions -- that sever the connections between data that are, after all, nothing more than limited, single-faceted abstractions of a complex, interconnected reality.

[2] Berry, Brian J.L. 1964. "Approaches to Regional Analysis: A Synthesis." *Annals of the Association of American Geographers* 54(1), 2-11. The general approach was used extensively in Chorley, Richard J., and Peter Haggett, editors. 1967. *Models in Geography*. London: Methuen.



Once empirical observation is conceptualized as one or more matrices, other possibilities immediately became obvious. First, we can imagine one of these cubes for each geographic scale at which "places" are defined (perhaps the data would percolate out into a well-organized fractal?). Second, we can tinker with the rows, columns, and "slices" of the cube. How about this :



Now we're able to view things just a bit more dynamically -- not just in terms of the static characteristics of places, but also in terms of the interdependencies among places. Several measures of spatial interaction are possible: flows of energy, people, commodities ... Unfortunately, the unbounded theoretical possibilities are often thwarted by the costs of gathering empirical data. In economics, though, there is a sufficiently well-developed infrastructure to support a data cube of flows of money and products between different industry groups (so that we can understand the broader implications, for example, of this past year's quadrupling of natural gas prices on different sectors). Analyzing this industry-by-industry flow matrix across different regions is the task of a large swath of regional science, which was born out of the links between Quantitative Revolution-style geography and neoclassical economics.

Regardless of how this conceptual scheme is drawn -- matrices, cubes, or four-dimensional object -- or how the axes are labeled, it is still necessary to decide what information to put in the cells. Indeed, this choice has considerable implications. Entering nominal data gives us the broad outlines of how a scholar thinks and organizes research on qualitative problems. Ordinal data may also be used. The vast majority of quantitative applications, of course, rely on interval/ratio data organized into some scheme resembling the data cube.

Still, there are additional complexities. Let's consider two deceptively simple questions, and unpack the apparently "known facts" to see what kinds of questions they raise.

i. **What is the population of Las Vegas?** First off, what is Las Vegas? The legally-defined municipality recognized under Nevada state law? The "functional" region of the local economy? The downtown and the strip? Or that unwieldy statistical beast crafted by the Census Bureau, the Metropolitan Statistical Area?[3] I crack open my 1999 *Statistical Abstract*: the Las Vegas NV-AZ MSA had an estimated population of 1,201,000 in 1996, a 40.9 percent increase over the estimate for 1990. We know this number is out of date. The recently-released estimates from the 2000 Census put the figure over 1.4 million, and growing at the rate of 6,000 per month.[4] But who are these 1.4 million? What does it mean to "live" in the place? Do we count part-year residents? How long must one live there to be considered a "permanent" resident?[5] What about those true statistical invisibles, the sunburned Vegas homeless population?

---

[3] Lengthy, if somewhat yawn-inducing, papers have been written on how metropolitan areas are, and should be, defined for alternative purposes. The question has assumed renewed urgency after each decade's census of population and housing, when the inherent ambiguity of a fluid, dispersed population plays havoc with the pigeon-holing necessities of federal data systems.

[4] Believe it or not, the language, if not the rhetoric, of "smart growth" has come to Vegas. The colorful mob lawyer-turned major Oscar Goodman glances out his office window at the subdivisions in the distance and reflects: "Look at all those gated communities, those big beautiful houses where people don't know their neighborhs and never talk to anyone...It's nothing but -- I'm not supposed to say this -- sprawl!" See Egan, Timothy. 2001. "Las Vegas Bet on Growth But Doesn't Love Payoff." *New York Times*, January 26, p. A1, A13.

[5] It should come as no surprise that this kind of question assumes primacy in college towns. Amidst the widespread attention to the undercount in the census, a back-page news item described the problems the Census Bureau faces with a 2.3 million person overcount -- a significant fraction of whom represent college students who filled out census forms at school while their parents also coded information for them at "home."

Once we answer this question for Las Vegas, we need to decide on some criteria for how we'll approach the same ambiguities in other cities in our data cube (which might, in this case, be defined by rows of cities, columns of variables such as population, income, etc., and slices for different time periods). What about college towns? Retirement communities with significant snowbird populations? The Outer Banks in winter, or in September when the barometer falls rapidly? (think of an asymptote approaching, but never reaching, zero).

ii. **What is the distribution of soil types in central New Jersey?** Suppose we are examining historical records of different land parcels, and trying to understand the distribution of agricultural activities in the late nineteenth century. Can we classify individual land parcels according to their soil type, and, by inference, potential productivity. First, what soil classification do we use? The first systematic taxonomy, developed by Russian scientists in the late 1800s? The genetic, formation-oriented classification developed in the U.S. in 1938? Or the latest version, the official "Soil Taxonomy" in use since 1965? The latter classification departs from earlier ones by virtue of its emphasis not on the process of formation, but on the characteristics of different soils; and soils modified by human activity are classified along with 'natural' ones. And then there is the question of scale -- both geographic, and taxonomic. At the geographic scale of the world map, the mid-Atlantic is part of the soil *order* of ultisols, highly chemically weathered soils that have developed under warm, moist climatic conditions. Looking closer at the region would reveal a finer pattern, roughly corresponding to the physiographic regions of New Jersey, of soil *suborders* and *great groups*. And if we chose a one-acre plot at random, we would likely find a mixture of several individual soil types, and, as anyone who's ever seen a soil survey knows, drawing the boundaries between soil types imposes a line on a messy continuum.

What do these random thoughts have to do with the enterprise of quantitative analysis? First, they convey some sort of appreciation for the inherent ambiguity -- or at least the constructed nature -- of the 'known facts.' They give us a 'starting point' for analysis and understanding. Or they alert us to the 'assumptions which form the bases for inferences or conclusions.' The seemingly neutral step of obtaining data has led us to the problem of implicit theories -- which are embedded in classifications of cities, suburbs, and metropolitan areas, or soil taxonomies developed according to formation processes or physical characteristics.

Or, to put it even more simply, consider what Friederich Nietzche had to say: "There are no facts as such. We must always begin by introducing a meaning in order for there to be a fact."

To be practical. Writing the 'data and methods' section of your scholarship need not be a laborious task of attempting to justify every single decision you've made on what data to use for a particular research question. Work carefully to conceptualize the 'data cube' that seems most relevant to the question you're exploring. Then search to find data that will, to the extent possible, satisfy these requirements. There will be no perfect data sets. In many applications,

you will have to design a social survey, or define a spatial sampling network, in order to obtain the specialized, "primary" data you need. For other questions, secondary data sources -- information originally gathered for other purposes -- will be sufficient. Indeed, one of the major consequences (both analytically, and in the policy and legal arenas) is the dramatic expansion of methods for linking databases assembled for unrelated purposes, by different institutions, according to different standards. Either way, whatever data sources you use, remain sensitive to the voices and silences of these 'givens' -- and use the 'data and methods' section of your research papers to convey some of these nuances.

3. Key Questions.

This week, we want to explore four main sets of questions. This list is suggestive, and should not be interpreted as some straight-jacket to narrow our field of vision. It's just a starting point for discussion. As you read through the articles slated for this week, consider these sets of issues:

i. What are the distinctive qualities of spatial data? What analytical tools are available to measure, control for, or visualize these special properties? These are the issues taken up by Fotheringham et al. Although we're not engaged in a hands-on GIScience experience in this seminar, the essential concepts and techniques outlined in Fotheringham's chapter provide nice background information for those first approaching any real GIS *analysis* (as opposed to the elementary mapping operations). The material should be only too familiar to those of you who know much more about contemporary GIS than I do.

ii. What happens when the "spatial data" describe people? What happens when we try to classify different people in different places? What happens when the methods of classification have a meaning that is, shall we say, a bit problematic? Margo Anderson is a well-known social and urban historian who has written extensively on the politics of the U.S. census (she got her Ph.D. from Rutgers in the late 1970s). Feinberg is a statistician. They take up the question of the racial classifications used in the census, and trace the history of the 'undercount' as well as of the categories themselves. The popular press always carries a steady stream of articles based on racial and ethnic breakdowns and population projections. You will never read these things in the same light after considering Anderson and Feinberg's analysis.

iii. What are the methodological and institutional considerations involved in changing certain kinds of "official" secondary data sources? Paul Longley, well known in GIScience and quantitative circles, wrote an editorial dealing with the complexities of adding a question of income to the British Census. He lays out, in a very concise essay, the kinds of errors and ambiguities that creep in to this kind of effort. And, to his questions, we can add others: what, after all, are we measuring when we ask about 'income'? What is the unit of measurement (person, household, family unit)? What is counted as income? Why do we always resort to the 'default' standard of *annual* income? What are the

alternatives to 'income' to measure social inequalities in opportunity, autonomy, and power?

iv. What happens when the data, those seemingly innocuous 'known facts,' have the following characteristics: they a) are extremely sensitive pieces of information that could be vulnerable to abuse, b) represent a condition that leads to horrible death, c) are difficult to obtain, d) are based on definitions that have changed over time, and e) are extraordinarily difficult to obtain at a spatially-disaggregated level. We cannot simply walk away, and resign ourselves to the impossibility of an appropriate spatial analysis. It is critically important to know how, when, and where HIV spread through the 1980s and 1990s, and to offer insight on what this means and what can be done. One of the best Masters' Theses I've ever seen was designed to expose the consequences of ignoring spatial data on HIV: federal funding to respond to HIV is based on state (and sometimes county) diagnosis records; but when people are diagnosed with this diagnosis, many of them make important decisions on where to spend the rest of their lives. There is a significant stream of urban-to-rural migration in some parts of the country, as people diagnosed in the city decide to go home to their families in smaller towns or rural areas -- where the health care system is not bolstered by the federal and state funding streams.

Peter Gould and Michelle Cochrane both take up the issue of spatial data on HIV cases, in different ways. Peter deals with a number of technical and methodological concerns in a way that, I hope you agree, conveys the importance of understanding the meaning of the three givens. He also has important things to say on the politics of spatial data, of confidentiality and analytical goals, and the uses of cartography.[6] Michelle Cochrane, a recently-minted Ph.D., takes a radically different approach to the question of what HIV data mean. She explores the shifting definitions and assumptions used to chart the epidemic over the course of the last twenty years, and shows that those definitions of 'data' with which we began -- as known facts, as givens, as assumptions -- are anything but self-evident.

4. A Few Useful Basics.

Finally, a few basics. Some of you have emphasized that you have little statistical background, and so this is meant as a refresher. As you glance through it, it will look very familiar and elementary. Ignore it if you don't need it, but, in case it's of some use...

---

[6] I apologize for the crude black-and-white reproduction of the maps. I'll bring the color originals. They have an impact, or at least they did a number of years ago when HIV was still seen as the new menace, and not simply the a priori fact of epidemiological life that it is now. Geographers are accustomed thinking about spatial diffusion as a benign, wine-stain-on-a-tablecloth process, but when what was spreading killed so many young, vibrant, and creative people in such painful ways...

Statisticians draw a sharp distinction between *descriptive* techniques, which are designed to give a basic profile of a certain set of measurements, and *inferential* statistics -- where we have drawn a small sample from a large population, and our goal is to determine whether relations observed in our sample are representative of broader relations in the population.

For the moment, let's confine our attention to descriptive methods, as applied to interval-ratio data.[7] Consider a data set made famous by a scientist in the 1930s, Fisher, who made precise measures of four different species of flowers, and used the results to see what specific measures (variables) best helped to distinguish among the species.[8] Each row is a separate observation; the four variables are 1) sepal length, 2) sepal width, 3) petal length, and 4) petal width, all in centimeters.[9] Fisher measured fifty flowers from each of the three species, so we have 150 observations:

```
5.1     3.5     1.4     0.2
4.9     3       1.4     0.2
4.7     3.2     1.3     0.2
4.6     3.1     1.5     0.2
5       3.6     1.4     0.2
5.4     3.9     1.7     0.4
4.6     3.4     1.4     0.3
5       3.4     1.5     0.2
4.4     2.9     1.4     0.2
4.9     3.1     1.5     0.1
5.4     3.7     1.5     0.2
4.8     3.4     1.6     0.2
4.8     3       1.4     0.1
4.3     3       1.1     0.1
5.8     4       1.2     0.2
5.7     4.4     1.5     0.4
5.4     3.9     1.3     0.4
5.1     3.5     1.4     0.3
5.7     3.8     1.7     0.3
5.1     3.8     1.5     0.3
5.4     3.4     1.7     0.2
5.1     3.7     1.5     0.4
4.6     3.6     1       0.2
5.1     3.3     1.7     0.5
4.8     3.4     1.9     0.2
5       3       1.6     0.2
5       3.4     1.6     0.4
5.2     3.5     1.5     0.2
5.2     3.4     1.4     0.2
4.7     3.2     1.6     0.2
4.8     3.1     1.6     0.2
5.4     3.4     1.5     0.4
5.2     4.1     1.5     0.1
5.5     4.2     1.4     0.2
4.9     3.1     1.5     0.2
5       3.2     1.2     0.2
5.5     3.5     1.3     0.2
4.9     3.6     1.4     0.1
```

[7] Recall the four main types of data: nominal (categories or names that have no quantitative meaning), ordinal (ranks), interval (measurements where differences between consecutive numbers are of equal intervals), and ratio (interval measurements where the zero point is not arbitrary). Time is an interval ratio, but not ratio (or do you believe Monty Python?). Temperature is interval for Census or Fahrenheit, but ratio only for Kelvin. Most quantitative measures only work with interval or ratio data, although there are some useful statistics, known as "nonparametric" statistics, for ordinal data.

[8] See Fisher, R.A. 1936. "The Use of Multiple Measures in Taxonomic Problems." *Annals of Eugenics* 7, 179-188. The methods refined by Fisher came to be known as discriminant analysis. The flower species are Setosa, Versicolor, and Virginica.

[9] The sepal is the protective leaf-like sheath at the base of the flower petals.

4.4	3	1.3	0.2
5.1	3.4	1.5	0.2
5	3.5	1.3	0.3
4.5	2.3	1.3	0.3
4.4	3.2	1.3	0.2
5	3.5	1.6	0.6
5.1	3.8	1.9	0.4
4.8	3	1.4	0.3
5.1	3.8	1.6	0.2
4.6	3.2	1.4	0.2
5.3	3.7	1.5	0.2
5	3.3	1.4	0.2
7	3.2	4.7	1.4
6.4	3.2	4.5	1.5
6.9	3.1	4.9	1.5
5.5	2.3	4	1.3
6.5	2.8	4.6	1.5
5.7	2.8	4.5	1.3
6.3	3.3	4.7	1.6
4.9	2.4	3.3	1
6.6	2.9	4.6	1.3
5.2	2.7	3.9	1.4
5	2	3.5	1
5.9	3	4.2	1.5
6	2.2	4	1
6.1	2.9	4.7	1.4
5.6	2.9	3.6	1.3
6.7	3.1	4.4	1.4
5.6	3	4.5	1.5
5.8	2.7	4.1	1
6.2	2.2	4.5	1.5
5.6	2.5	3.9	1.1
5.9	3.2	4.8	1.8
6.1	2.8	4	1.3
6.3	2.5	4.9	1.5
6.1	2.8	4.7	1.2
6.4	2.9	4.3	1.3
6.6	3	4.4	1.4
6.8	2.8	4.8	1.4
6.7	3	5	1.7
6	2.9	4.5	1.5
5.7	2.6	3.5	1
5.5	2.4	3.8	1.1
5.5	2.4	3.7	1
5.8	2.7	3.9	1.2
6	2.7	5.1	1.6
5.4	3	4.5	1.5
6	3.4	4.5	1.6
6.7	3.1	4.7	1.5
6.3	2.3	4.4	1.3
5.6	3	4.1	1.3
5.5	2.5	4	1.3
5.5	2.6	4.4	1.2
6.1	3	4.6	1.4
5.8	2.6	4	1.2
5	2.3	3.3	1
5.6	2.7	4.2	1.3
5.7	3	4.2	1.2
5.7	2.9	4.2	1.3
6.2	2.9	4.3	1.3
5.1	2.5	3	1.1
5.7	2.8	4.1	1.3
6.3	3.3	6	2.5
5.8	2.7	5.1	1.9
7.1	3	5.9	2.1
6.3	2.9	5.6	1.8
6.5	3	5.8	2.2
7.6	3	6.6	2.1
4.9	2.5	4.5	1.7
7.3	2.9	6.3	1.8
6.7	2.5	5.8	1.8
7.2	3.6	6.1	2.5
6.5	3.2	5.1	2

6.4	2.7	5.3	1.9
6.8	3	5.5	2.1
5.7	2.5	5	2
5.8	2.8	5.1	2.4
6.4	3.2	5.3	2.3
6.5	3	5.5	1.8
7.7	3.8	6.7	2.2
7.7	2.6	6.9	2.3
6	2.2	5	1.5
6.9	3.2	5.7	2.3
5.6	2.8	4.9	2
7.7	2.8	6.7	2
6.3	2.7	4.9	1.8
6.7	3.3	5.7	2.1
7.2	3.2	6	1.8
6.2	2.8	4.8	1.8
6.1	3	4.9	1.8
6.4	2.8	5.6	2.1
7.2	3	5.8	1.6
7.4	2.8	6.1	1.9
7.9	3.8	6.4	2
6.4	2.8	5.6	2.2
6.3	2.8	5.1	1.5
6.1	2.6	5.6	1.4
7.7	3	6.1	2.3
6.3	3.4	5.6	2.4
6.4	3.1	5.5	1.8
6	3	4.8	1.8
6.9	3.1	5.4	2.1
6.7	3.1	5.6	2.4
6.9	3.1	5.1	2.3
5.8	2.7	5.1	1.9
6.8	3.2	5.9	2.3
6.7	3.3	5.7	2.5
6.7	3	5.2	2.3
6.3	2.5	5	1.9
6.5	3	5.2	2
6.2	3.4	5.4	2.3
5.9	3	5.1	1.8

Let's focus just on the first column -- the sepal length variable. Its average or mean, which we'll call μ (Greek letter mu) is defined as the sum, divided by the number of observations:

$$\mu = \frac{\sum X}{N}$$

So in the case of our sepal length variable, we sum the values, obtaining 876.5, and divide by 150, giving us 5.8433. An alternative measure of central tendency is the median, which is the value at which one-half of the observations are less, and one-half are more. To obtain this, we rank the observations, and then count halfway down. For the sepal length variable, we sort the list, count down to the half-point, and ... oops, the half-point is between 75 and 76. By convention, when we have an even number of observations, the two 'tied' middle values are added together and divided by two. In our case, when we sort the list, both the 75th and 76th observations are 5.8, so this is our median.[10] In the case of this data set, therefore, there is a very close correspondence between the mean (5.84) and the median (5.8). The mean is more sensitive to extremes, and so for many uses it is preferable to use the median. A set of scores is

[10] A third common measure of central tendency is the mode, which is the value that occurs most frequently in a list of observations.

said to be "skewed" when the median and the mean diverge from one another; our case study variable of sepal width has almost no skewness.

Measures of central tendency tell only half the story. We also need a measure of the "spread" of the observations. One logical approach would be to compare each of the observations to the mean, obtaining a deviation for each, which we will call little *x*:

$$X - \mu = x$$

	x
5.1-5.8433	-.7433
4.9-5.8433	-.9433
4.7-5.8433	-1.1433
.	
.	
.	
6.3-5.8433	.4567
6.5-5.8433	.6567
6.2-5.8433	.3567
5.9-5.8433	.0567

If we need a single summary measure of the "spread" in sepal length, however, we encounter one of the nasty habits of the deviation measures: adding them up always yields zero. The best way out if this problem is to square the deviations:

	x	x^2
5.1-5.8433	-.7433	.5525
4.9-5.8433	-.9433	.8898
4.7-5.8433	-1.1433	1.3071
.		
.		
.		
6.3-5.8433	.4567	.2086
6.5-5.8433	.6567	.4312
6.2-5.8433	.3567	.1272
5.9-5.8433	.0567	.0032

Now when we add up the squared deviations, we get a rough measure of the variability of the observations around the mean. This is called the *sum of squares*, and for our dataset it's 102.168. If we divide the sum of squares by the number of observations, we can find the average of the squared deviations. This is the *variance*:

$$\frac{\sum(x^2)}{N}$$

Where little *x*, recall, is the difference between each value and the mean. The variance appears in all sorts of statistical applications where we are comparing the 'spread' of different variables. For our dataset, it's 0.6811. It's useful on its own, but it's even more relevant when we take one more step, and go back out of squares. Taking the square root of the variance gives us the *standard deviation*, typically denoted by lowercase sigma, σ:

$$\sigma = \sqrt{\frac{\sum(x^2)}{N}}$$

The standard deviation tells us how much, on average, each observation differs from the mean. For our dataset, σ is 0.8253. The sepal length of each of the flowers Fisher measured could be expected, on average, to be 0.8253 centimeters *longer* or *shorter* than the mean of 5.8433.

The standard deviation, in turn, allows us to express each of the observations in our dataset on a new, 'generic' measurement scale, which we will denote with *z*:

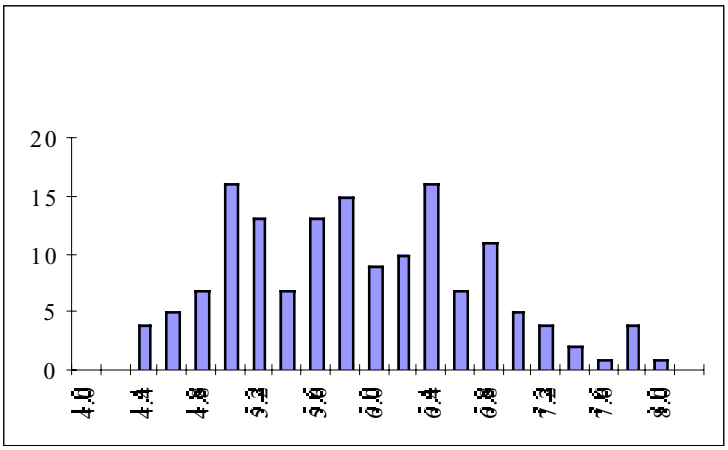
$$z = \frac{X - \mu}{\sigma}$$

The utility of z scores should be clear: we are now able to express each variable in terms that do not depend on the measurement scale; so it becomes possible to compare variables that are expressed on wildly different scales. Z scores measure observations *in terms of how many standard deviations away from the mean they are*.

Description or Inference?

Descriptive statistics only take us so far. The more important task is to understand if a limited sample of observations can help us to understand, with a given level of confidence, broader relations in an entire population. Inferential statistics refers to the standards, norms, and assumptions used to accomplish this task.

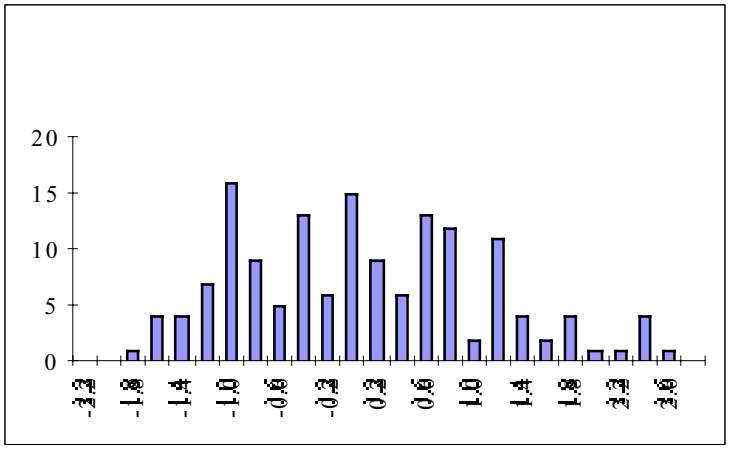
If we draw a frequency histogram of Fisher's measures, we get this:



The more observations added to the sample (which is now 150) would not alter the fundamental shape of this curve, but would smooth it out. The larger the number of observations, the closer this distribution will appear to the familiar, smooth bell curve of the *normal distribution*. The normal distribution is a purely theoretical function, based on the unlimited number of surveys, measures, and experiments conducted by scientists and statisticians over the course of the past

several hundred years. Many, but not all phenomena exhibit a normal distribution. As a consequence, classical inferential statistics is based on the assumption that the data are normally distributed: the normal bell curve became the gold standard by which all other distributions were judged. So when analysts collect data today, and use off-the-shelf computer statistics package for inferential purposes, they are implicitly accepting this assumption. Not surprisingly, it's often a mistake to accept the assumption.

If we draw a histogram of Fisher's data, this time using z-scores, we obtain this:



with most of the observations clustering not too far from the zero point. This is, with a few warts and blemishes, the *standard normal distribution*. A "perfect," theoretical standard normal distribution has a mean of zero, a standard deviation of 1.0, and a median of 0, and displays the nice, smooth bell curve (the shape of the bell curve, by the way, is defined as a function of the natural logs, which have as their base the 'universal constant'; this will reappear when we dive into logistic regression). Clearly, Fisher's dataset is far from perfect -- if you look at the left half, it's got a bit of 'bimodal' flavor to it, but on the right this simple pattern gets even more messy. The standard normal distribution is the underlying assumption of a vast body of statistical techniques known as the "general linear model." It is often not satisfied in practice, but remains widely used because it provides a yardstick by which to compare observed distributions. It is also a judgement call as to whether a particular variable has a normal distribution. Should we assume that Fisher's measurement of sepal length exhibits a normal distribution?

There is one additional complication with inferential techniques. The formulas shown above for the variance and standard deviation are, in fact, slightly biased if we are attempting to use this sample to infer to a larger population. The sample variance provides an *underestimate* of the actual variability in the population, in part because we are using information from the sample in order to calculate formulas to describe the sample. This problem of circularity is known as the "degrees of freedom." Degrees of freedom is measured as a number expressing the number of observations that provide genuinely new information. So the formulas shown above

must be adjusted for df. In our example, when we calculate the mean of our 150 observations, as soon as I tell you 149 values you know the last one. Degrees of freedom in this case is 149, or n-1, and so the formula for variance becomes:

$$\frac{\sum(x^2)}{N-1}$$

and the sample standard deviation is:

$$\sigma = \sqrt{\frac{\sum(x^2)}{N-1}}$$

Purists sometimes use 's' to represent the sample standard deviation, and "σ" to denote the true, population standard deviation. Obviously, the difference between s and σ is small with large samples, but can be quite considerable with small samples from populations that have a high degree of variability. In our example, the sample standard deviation is 0.6811 if we assume that the world only has 150 flowers, and .6857 if we are using this sample to infer to the broader population.

The standard normal distribution is part of a family of functions called 'probability density functions.' These allow us to compare observed phenomena with theoretical expectations, under the assumption of a population that exhibits a normal distribution. So, for Fisher's data, we can ask such questions as: what is the likelihood that one of his flowers has a z-score of 2.5 -- two and one half standard deviations above the mean? This would be a whopper of an iris, perhaps a genetically-modified FrankenFlower", with a sepal length of more than seven and one-half centimeters. The standard normal distribution can be expressed as a table of probabilities, as illustrated in Table A.1 (attached). There is only a probability of 0.0062 that a flower's sepal length will exceed the mean by 2.5 standard deviations.

That's an extremely brief recap of the most common descriptive statistics, and a tiny, tiny bit of statistical inference. There's much more of use when you need to provide readers with a basic profile of your data, or to compare different samples from the same population, or when you need to compare samples from different populations, and so on. Since our primary focus in this class is multivariate applications, we're not going to get a systematic review of inferential tests and assumptions; I will bring some of this material in at relevant points in the class, but if you need further guidance, I recommend Perry R. Hinton, 1995, *Statistics Explained*. London and New York: Routledge. ISBN 0-415-10286-3, paperback. Also included as an attachment is a handy chart from Hinton's volume that provides a short guide to different statistical tests used for different purposes.

Now for the push-button answer. Start up SAS, and put the following lines in the program editor:

```
libname qga "d:\qga";
*note this is a comment - begins with asterisk, ends, like all lines, with;
data iris;
title 'Fisher (1936) taxonomy data';
input sepallen sepalwid petallen petalwid;
        label sepallen='sepal length in cm';
        label sepalwid='sepal width in cm';
        label petallen='petal length in cm';
        label petalwid='petal width in cm';
cards;
[here is where you paste the data lines from above]
;
run;

proc univariate data=iris;
*note if we don't use a 'var' statement, by default all numeric;
*variables are analyzed with proc univariate;
run;
```

Use 'Locals/Submit' to submit the statements, and what appears in the output buffer gives you the information you need. There's a lot of information here, and what is relevant depends on what you're trying to understand with your particular dataset. The mean, median, standard deviation, and variance are there, but there's also USS (uncorrected sum of squares), CSS ([mean-]corrected sum of squares), and a bunch of other things. Consult the SAS manual for further information on decoding the other stuff if it's of use to you.

```
Fisher (1936) taxonomy data
                                                13:05 Monday, January 15, 2001

                            Univariate Procedure

Variable=SEPALLEN      sepal length in cm

                Moments                                 Quantiles(Def=5)

 N              150  Sum Wgts        150     100% Max       7.9       99%       7.7
 Mean       5.843333  Sum           876.5      75% Q3        6.4       95%       7.3
 Std Dev    0.828066  Variance   0.685694      50% Med       5.8       90%       6.9
 Skewness   0.314911  Kurtosis   -0.55206      25% Q1        5.1       10%       4.8
 USS         5223.85  CSS        102.1683       0% Min       4.3        5%       4.6
 CV         14.17113  Std Mean   0.067611                              1%       4.4
 T:Mean=0   86.42537  Pr>|T|       0.0001     Range          3.6
 Num ^= 0        150  Num > 0         150     Q3-Q1          1.3
 M(Sign)          75  Pr>=|M|      0.0001     Mode             5
 Sgn Rank     5662.5  Pr>=|S|      0.0001


                                  Extremes

                         Lowest    Obs     Highest    Obs
                            4.3(      14)      7.7(     118)
                            4.4(      43)      7.7(     119)
                            4.4(      39)      7.7(     123)
                            4.4(       9)      7.7(     136)
                            4.5(      42)      7.9(     132)


                            Fisher (1936) taxonomy data                      2
                                                13:05 Monday, January 15, 2001

                            Univariate Procedure

Variable=SEPALWID      sepal width in cm

                Moments                                 Quantiles(Def=5)

 N              150  Sum Wgts        150     100% Max       4.4       99%       4.2
 Mean       3.057333  Sum           458.6      75% Q3        3.3       95%       3.8
 Std Dev    0.435866  Variance   0.189979      50% Med         3       90%      3.65
```

Skewness	0.318966	Kurtosis	0.228249	25% Q1	2.8	10%	2.5
USS	1430.4	CSS	28.30693	0% Min	2	5%	2.3
CV	14.25642	Std Mean	0.035588			1%	2.2
T:Mean=0	85.9083	Pr>|T|	0.0001	Range	2.4		
Num ^= 0	150	Num > 0	150	Q3-Q1	0.5		
M(Sign)	75	Pr>=|M|	0.0001	Mode	3		
Sgn Rank	5662.5	Pr>=|S|	0.0001				

Extremes

Lowest	Obs	Highest	Obs
2(	61)	3.9(	17)
2.2(	120)	4(	15)
2.2(	69)	4.1(	33)
2.2(	63)	4.2(	34)
2.3(	94)	4.4(	16)

Univariate Procedure

Variable=PETALLEN petal length in cm

Moments

Quantiles(Def=5)

N	150	Sum Wgts	150	100% Max	6.9	99%	6.7
Mean	3.758	Sum	563.7	75% Q3	5.1	95%	6.1
Std Dev	1.765298	Variance	3.116278	50% Med	4.35	90%	5.8
Skewness	-0.27488	Kurtosis	-1.4021	25% Q1	1.6	10%	1.4
USS	2582.71	CSS	464.3254	0% Min	1	5%	1.3
CV	46.97441	Std Mean	0.144136			1%	1.1
T:Mean=0	26.0726	Pr>|T|	0.0001	Range	5.9		
Num ^= 0	150	Num > 0	150	Q3-Q1	3.5		
M(Sign)	75	Pr>=|M|	0.0001	Mode	1.4		
Sgn Rank	5662.5	Pr>=|S|	0.0001				

Extremes

Lowest	Obs	Highest	Obs
1(	23)	6.4(	132)
1.1(	14)	6.6(	106)
1.2(	36)	6.7(	118)
1.2(	15)	6.7(	123)
1.3(	43)	6.9(	119)

Univariate Procedure

Variable=PETALWID petal width in cm

Moments

Quantiles(Def=5)

N	150	Sum Wgts	150	100% Max	2.5	99%	2.5
Mean	1.199333	Sum	179.9	75% Q3	1.8	95%	2.3
Std Dev	0.762238	Variance	0.581006	50% Med	1.3	90%	2.2
Skewness	-0.10297	Kurtosis	-1.3406	25% Q1	0.3	10%	0.2
USS	302.33	CSS	86.56993	0% Min	0.1	5%	0.2
CV	63.55511	Std Mean	0.062236			1%	0.1
T:Mean=0	19.2706	Pr>|T|	0.0001	Range	2.4		
Num ^= 0	150	Num > 0	150	Q3-Q1	1.5		
M(Sign)	75	Pr>=|M|	0.0001	Mode	0.2		
Sgn Rank	5662.5	Pr>=|S|	0.0001				

Extremes

Lowest	Obs	Highest	Obs
0.1(	38)	2.4(	137)
0.1(	33)	2.4(	141)
0.1(	14)	2.5(	101)
0.1(	13)	2.5(	110)
0.1(	10)	2.5(	145)