The Half-Life of a Social Statistician

Andrew Gelman

15 Nov 2012

sta·tis·tics *noun plural but singular or plural in construction* \stə-'tis-tiks\

Definition of STATISTICS

1 : a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data

2 : a collection of quantitative data

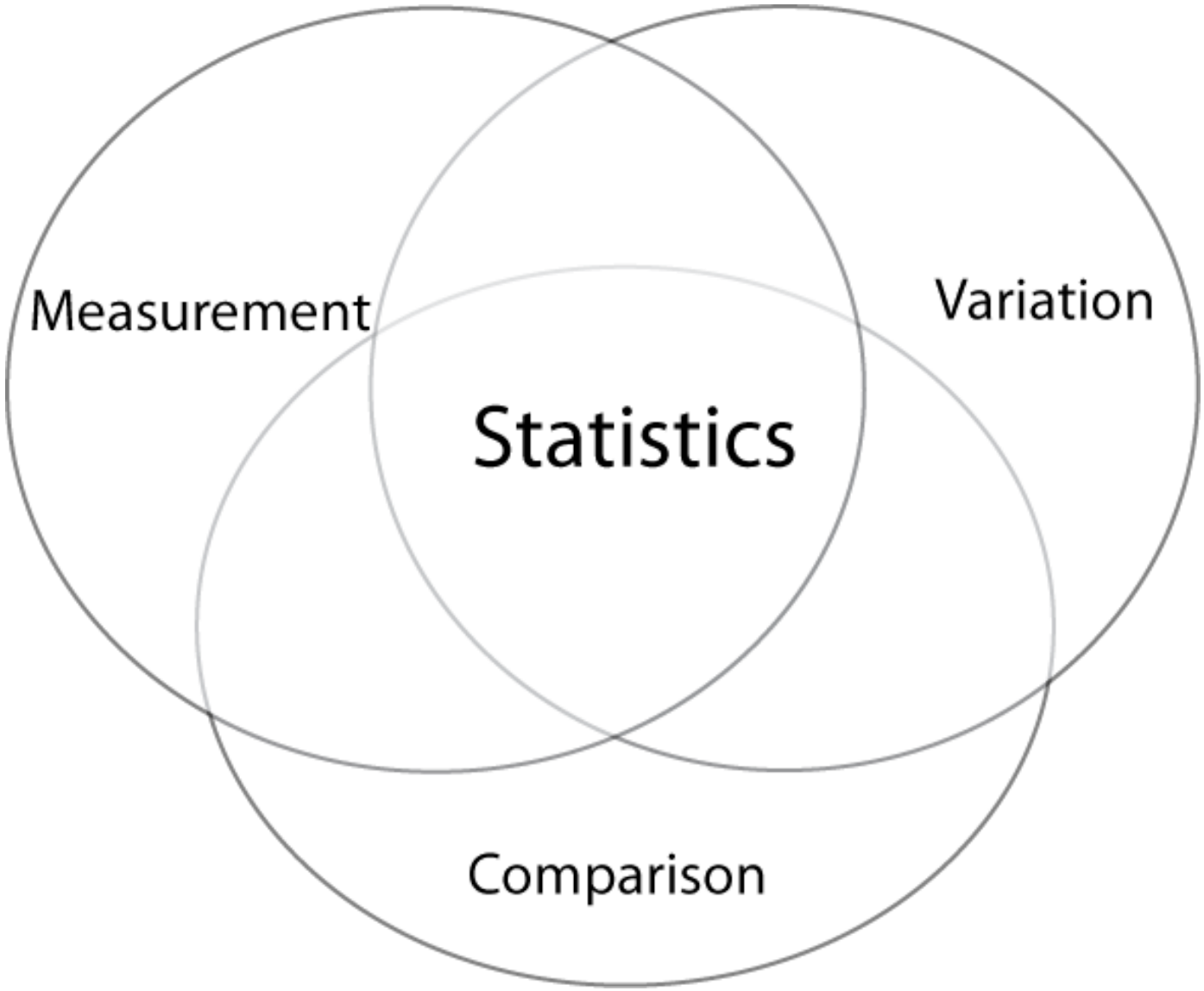
Origin of STATISTICS

German *Statistik* study of political facts and figures, from New Latin *statisticus* of politics, from Latin *status* state

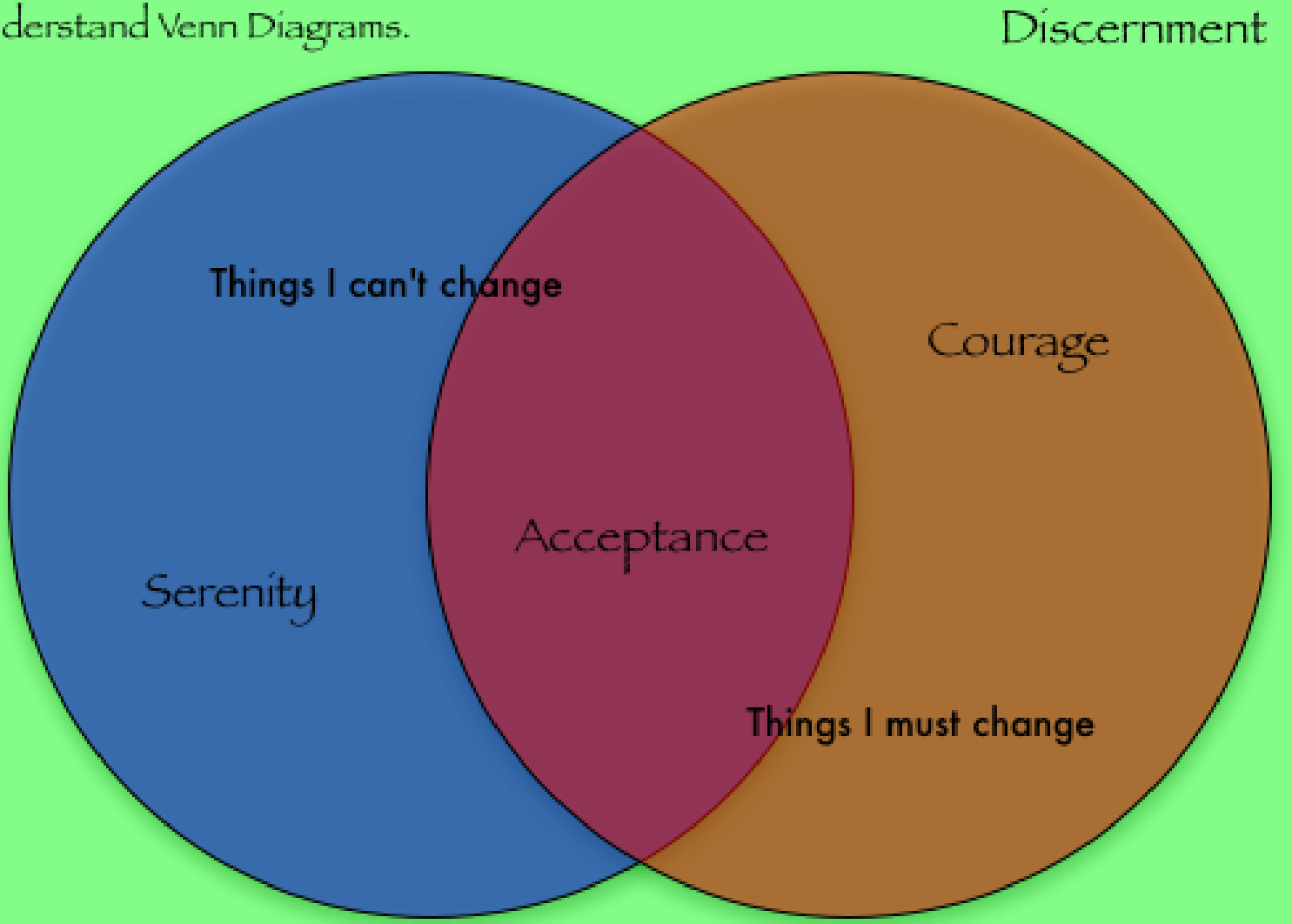
First Known Use: 1770

Andrew Gelman

The Half-Life of a Social Statistician

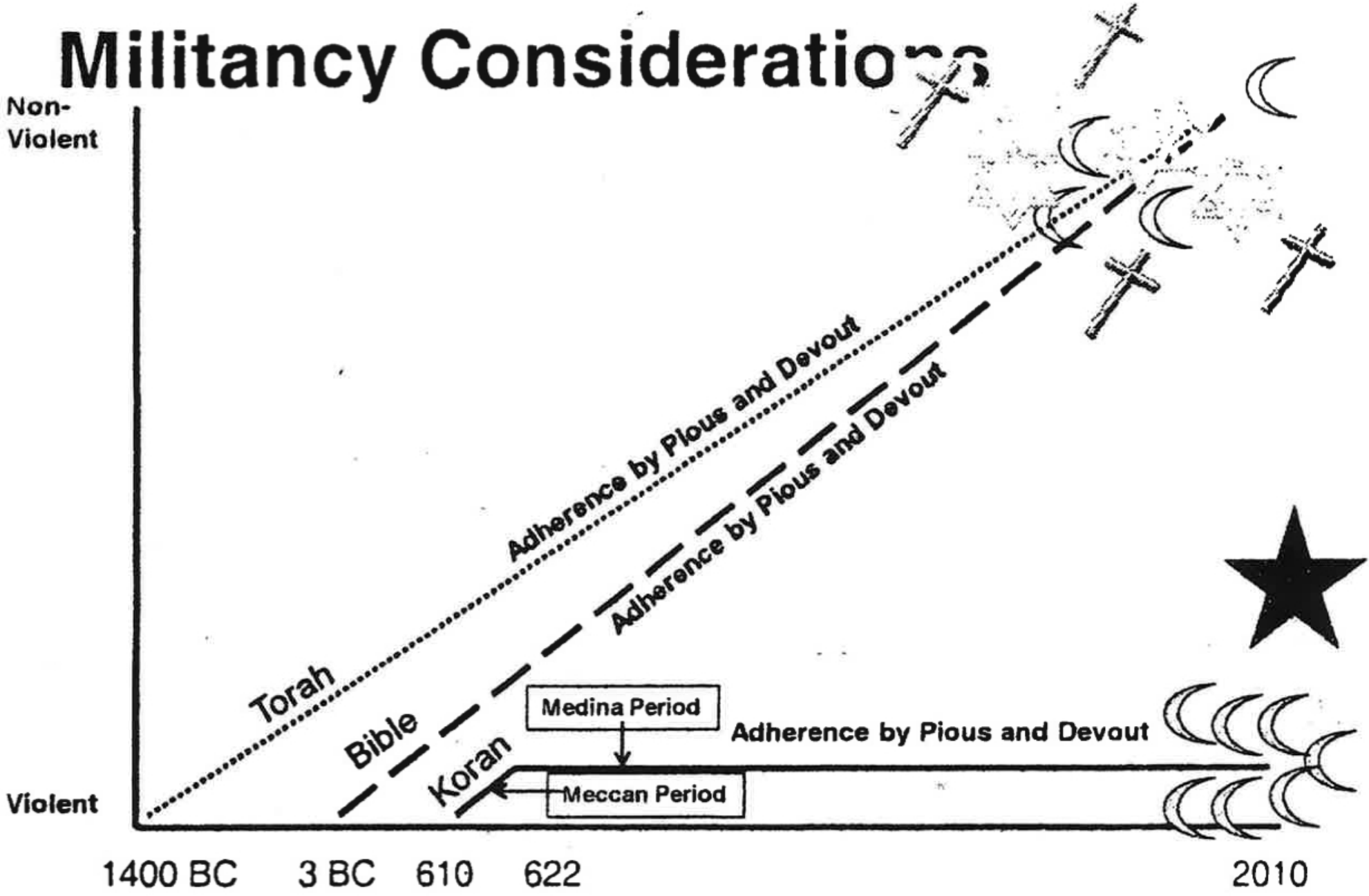

Measurement
Variation
Statistics
Comparison

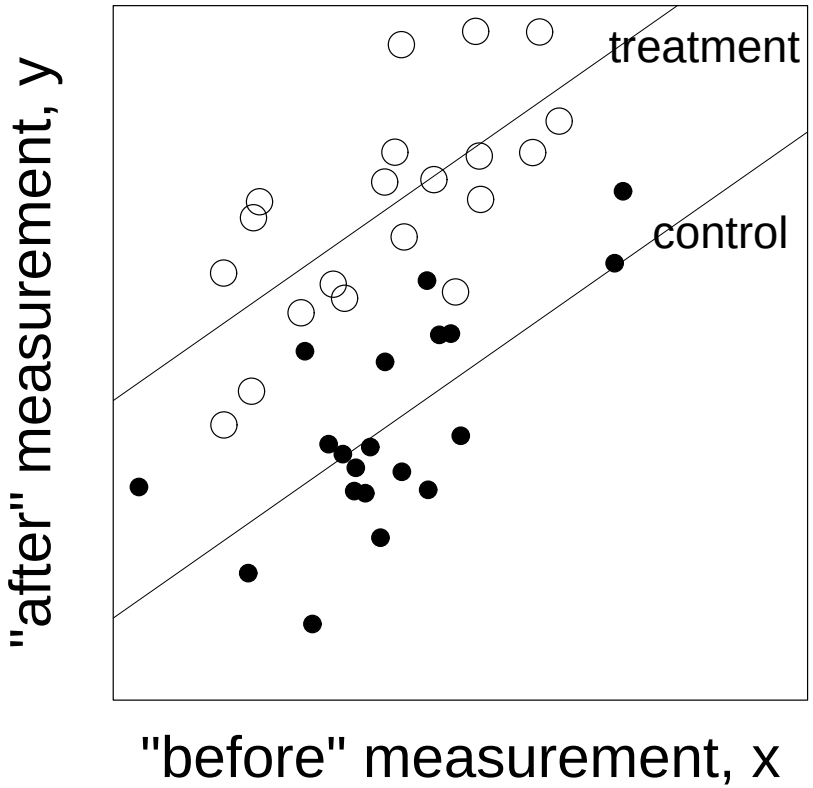
Lord, grant me the serenity to accept things that cannot be changed;
courage to change things that must be changed;
discernment to ignore things that don't need changing;
acceptance that some things I need to change, I can't,
And the wisdom to understand Venn Diagrams.

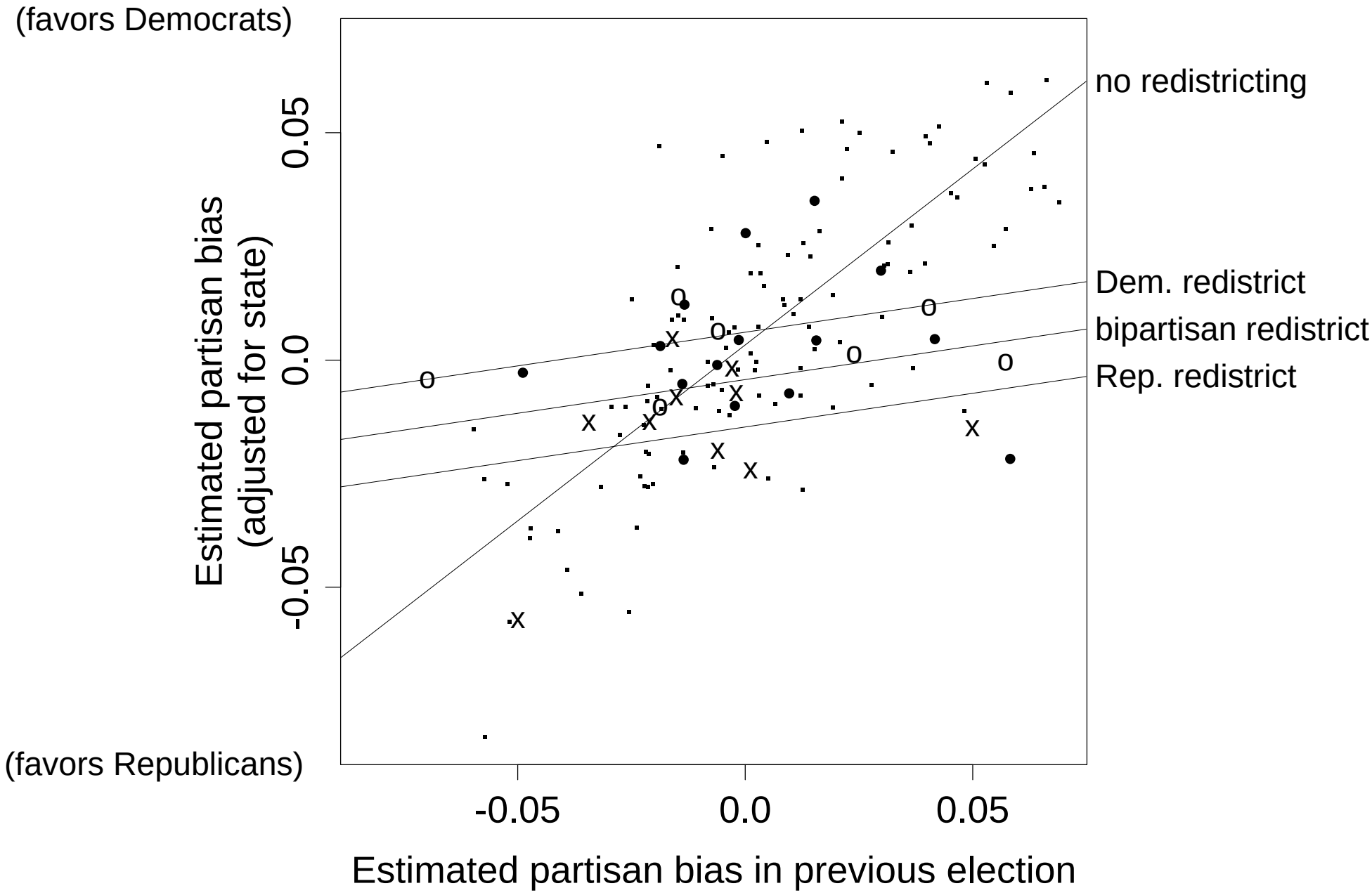


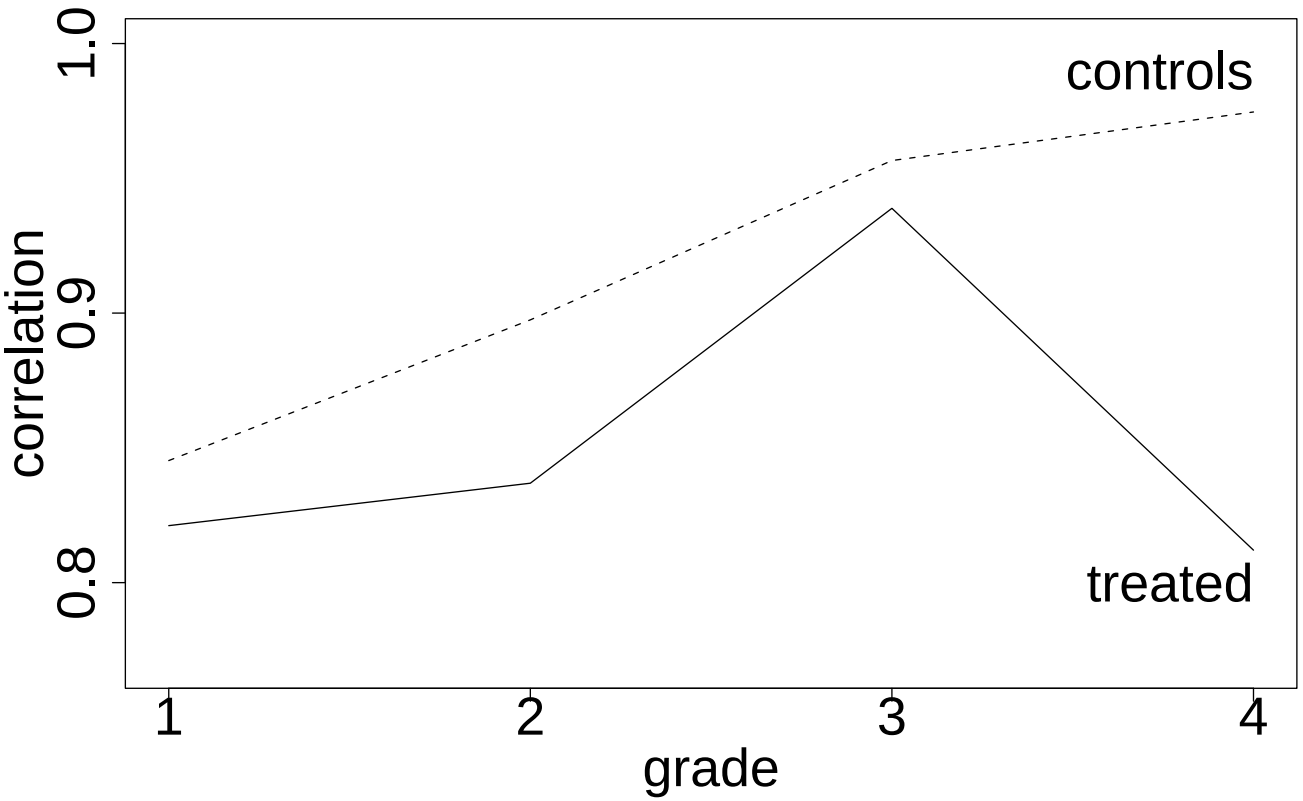
The Ser-Venn-ity Prayer

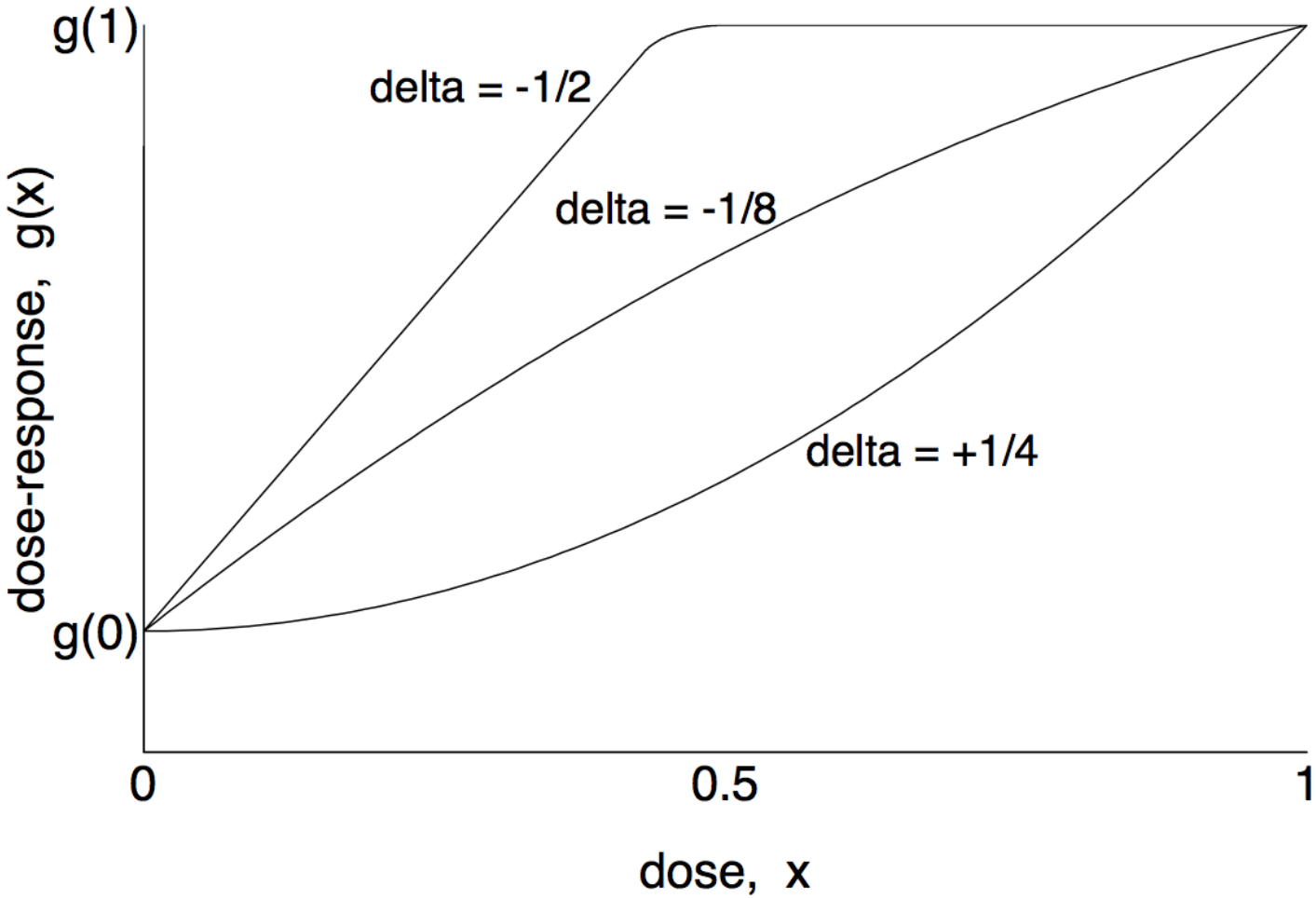
Militancy Considerations

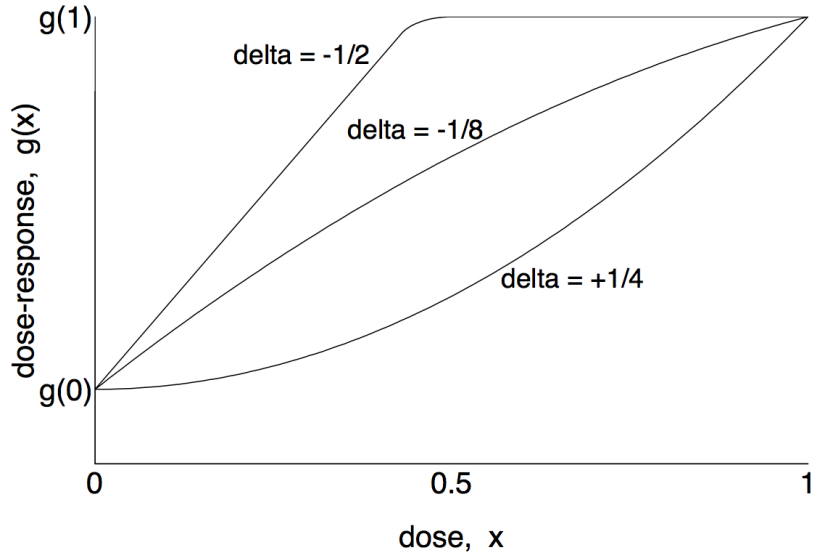


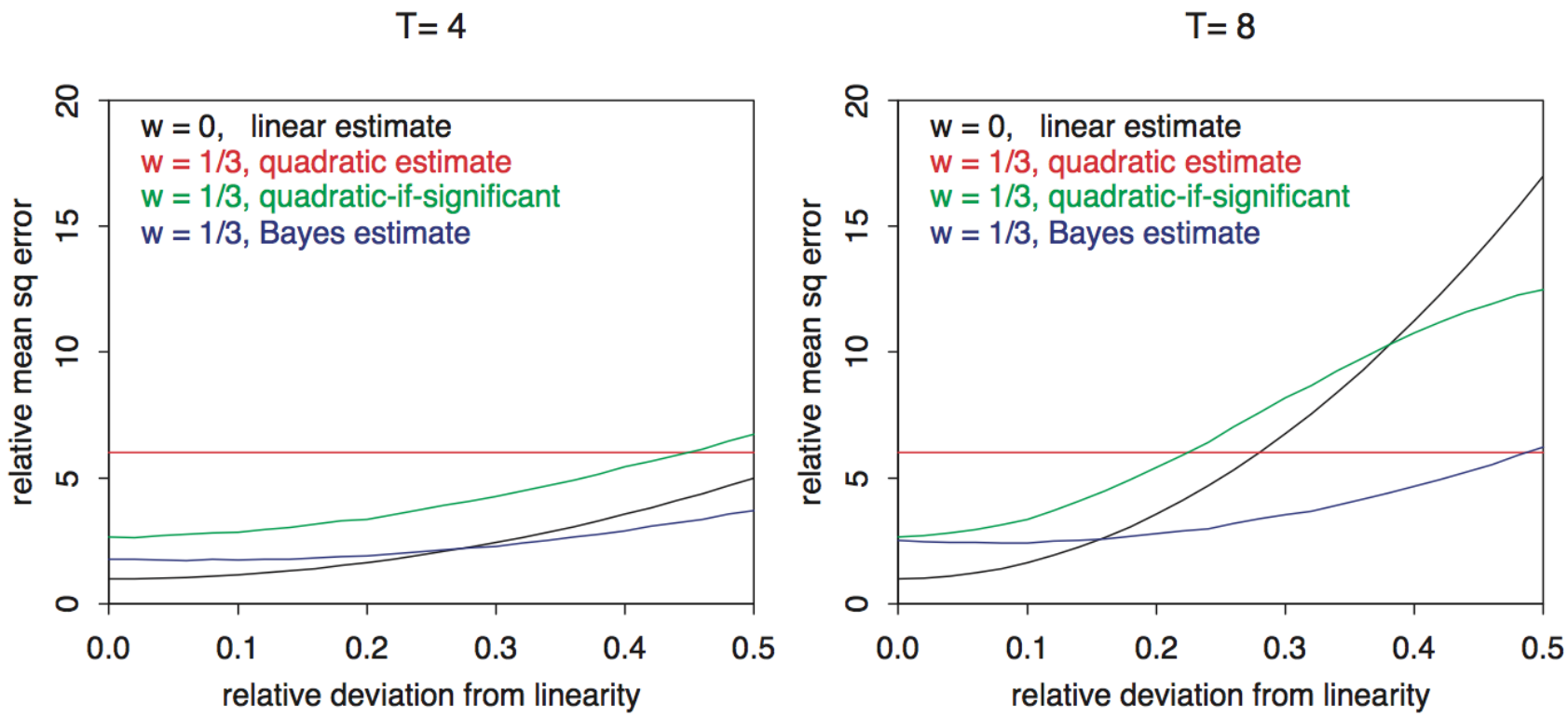

treatment
control
"after" measurement, y
"before" measurement, x

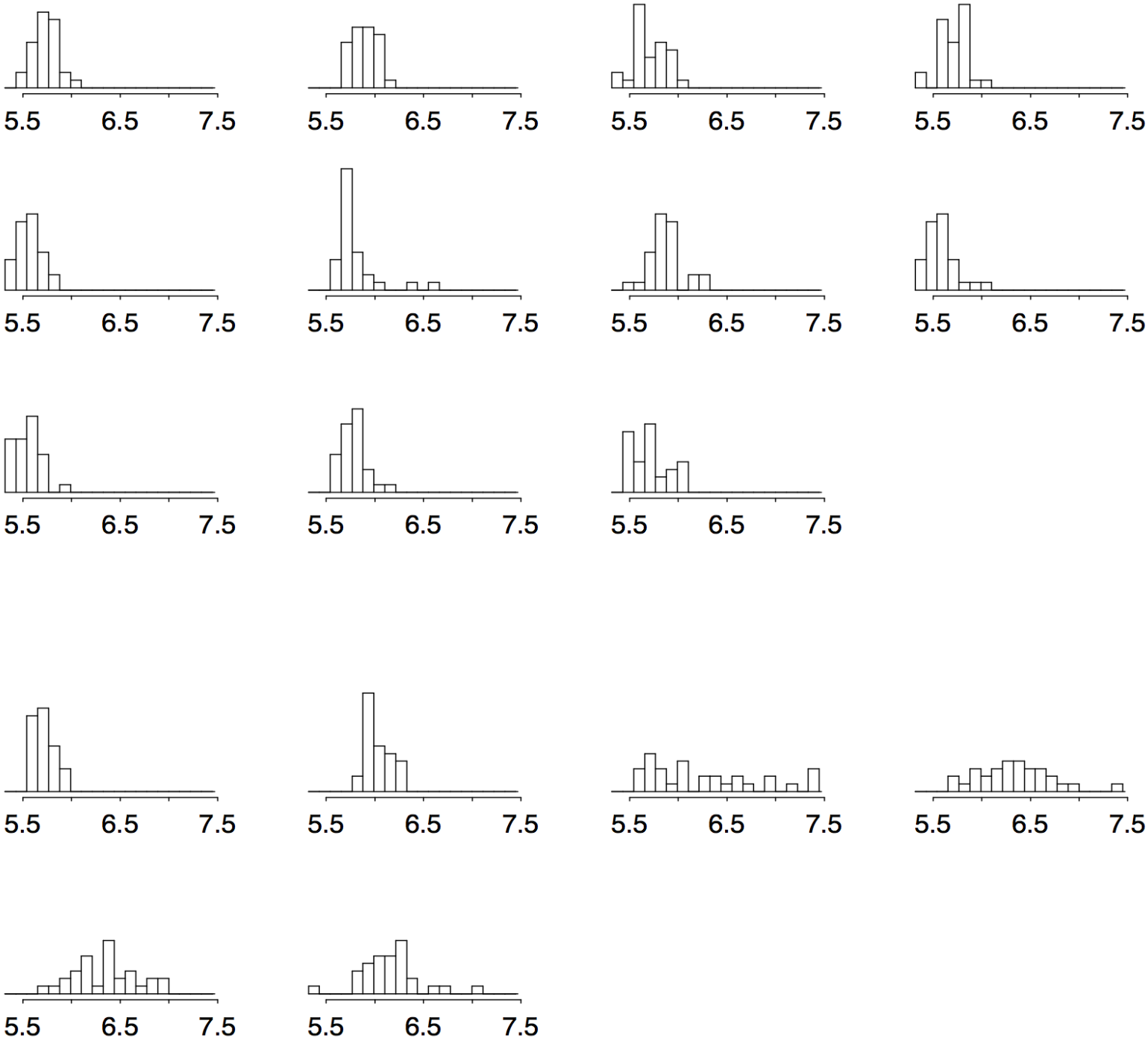

(favors Democrats)
(favors Republicans)
Estimated partisan bias
(adjusted for state)
0.05
0.0
-0.05
-0.05
0.0
0.05
Estimated partisan bias in previous election
no redistricting
Dem. redistrict
bipartisan redistrict
Rep. redistrict


1.0
0.9
0.8
correlation
controls
treated
1
2
3
4
grade


g(1)
delta = -1/2
delta = -1/8
dose-response, g(x)
delta = +1/4
g(0)
0
0.5
1
dose, x


g(1)
g(0)
dose-response, g(x)
delta = -1/2
delta = -1/8
delta = +1/4
0
0.5
1
dose, x


T= 4
T= 8
w = 0, linear estimate
w = 1/3, quadratic estimate
w = 1/3, quadratic-if-significant
w = 1/3, Bayes estimate
relative mean sq error
0
5
10
15
20
0.0
0.1
0.2
0.3
0.4
0.5
relative deviation from linearity


5.5
6.5
7.5

"Brief Thoughts on Maps"

Albert Szent-Gyorgyi, who knew a lot about maps
according to which life is on its way somewhere or other,
told us this story from the war
due to which history is on its way somewhere or other:

The young lieutenant of a small Hungarian detachment in the Alps
sent a reconnaissance unit out into the icy wasteland.
It began to snow
immediately, snowed for two days and the unit
did not return. The lieutenant suffered: he had dispatched
his own people to death.

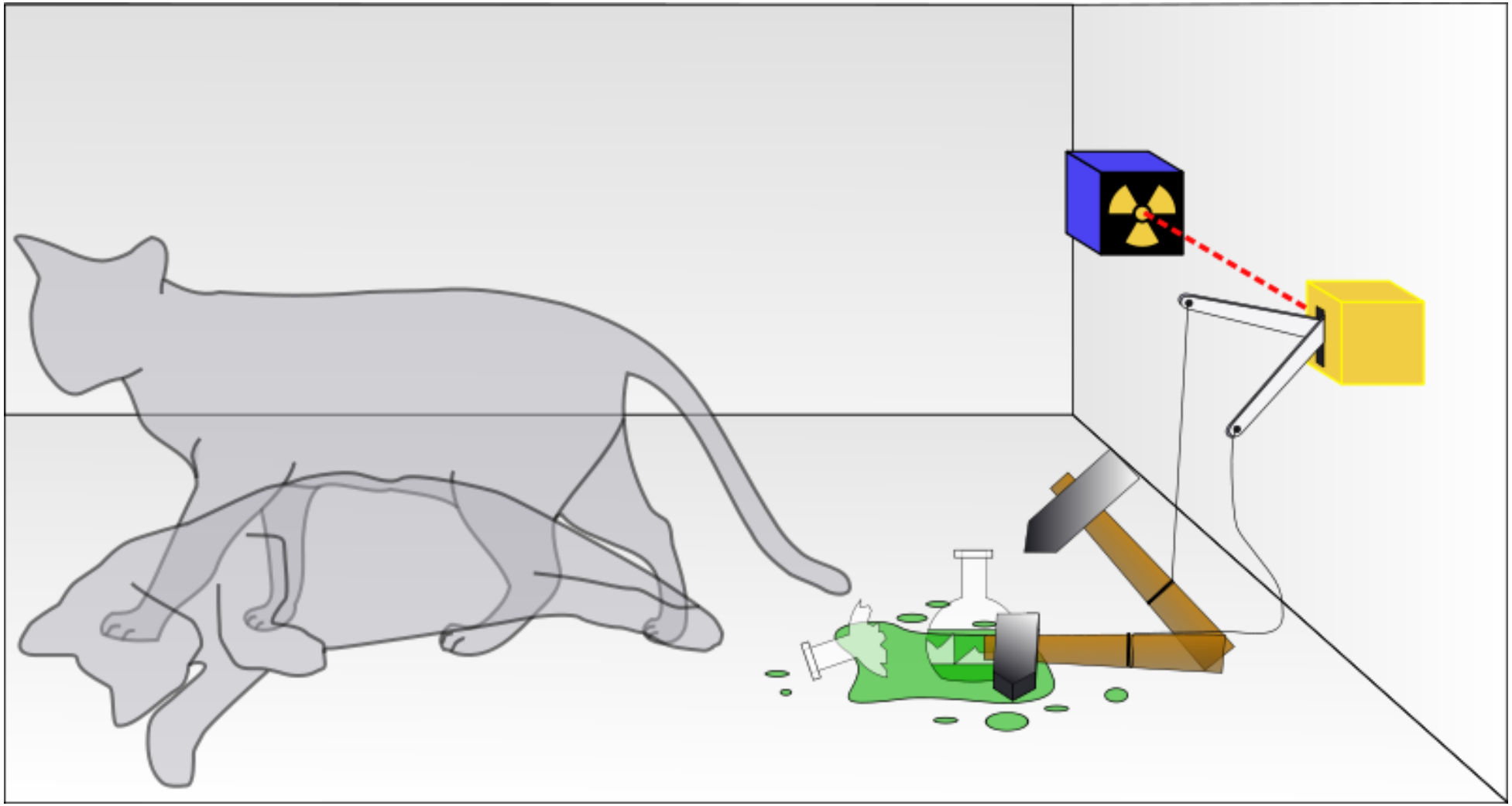
But the third day the unit came back.
Where had they been? How had they made their way?
Yes, they said, we considered ourselves
lost and waited for the end. And then one of us
found a map in his pocket. That calmed us down.
We pitched camp, lasted out the snowstorm and then with the map
we discovered our bearings.
And here we are.

The lieutenant borrowed this remarkable map
and had a good look at it. It was not a map of the Alps
but of the Pyrenees.

Goodbye now.

From Wikipedia:

In several published articles, Weick related a story that originally appeared in a poem by Miroslav Holub that was published in the Times Literary Supplement. Weick plagiarized Holub in that he republished the poem (with some minor differences, including removing line breaks and making small changes in a few words) without quotation or attribution. Some of Weick's articles included the material with no reference to Holub; others referred to Holub but without indicating that Weick had essentially done a direct copy of Holub's writing. The plagiarism was detailed in an article by Thomas Basbøll and Henrik Graham.[5] In a response, Weick disputed the claim of plagiarism, writing, "By the time I began to see the Alps story as an example of cognition in the path of the action, I had lost the original article containing Holub's poem and I was not even sure where I had read the story . . . I reconstructed the story as best I could."[6] Weick did not give a plausible explanation of how this reconstruction led to the appearance of a story with wording nearly identical to Holub's. Basbøll and Graham write, "The American Historical Association acknowledges the existence of this common defence in specific cases of plagiarism, tersely remarking that it "is plausible only in the context of a wider tolerance of shoddy work."



Hierarchy of stories:

1. Joke, or fiction such as Harry Potter
2. Urban legend
3. Dubious story (the lost soldiers in the Alps)
4. Plausible but undocumented story (e.g., Iraqs weapons of mass destruction)
5. Documented anecdote

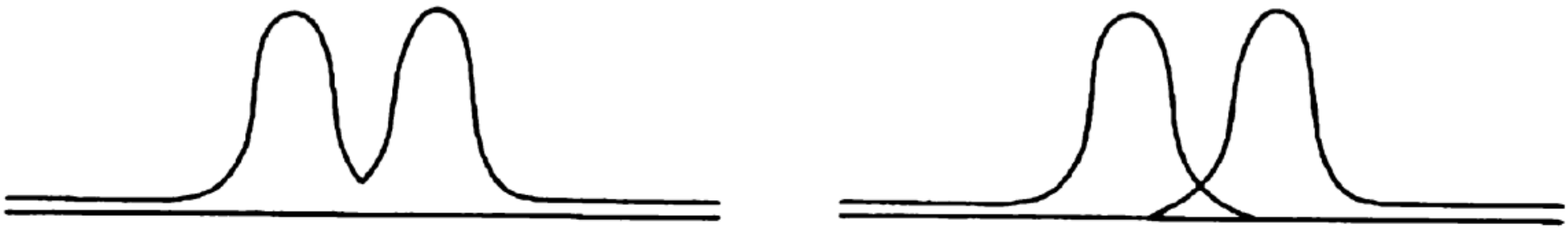


Fig. 2. Sketch from an example in Zeisel (1985), who writes, "When the frequency of [driving] accidents is plotted against the time of menstruation a surprisingly shaped curve arises [left graph]. Upon investigation, the curve turned out to be the composite of two easily identified separate curves [right graph]; one for parous women (those who had given birth) and one for nonparous women. The one group had the accident peak immediately after their period, the other immediately before it."

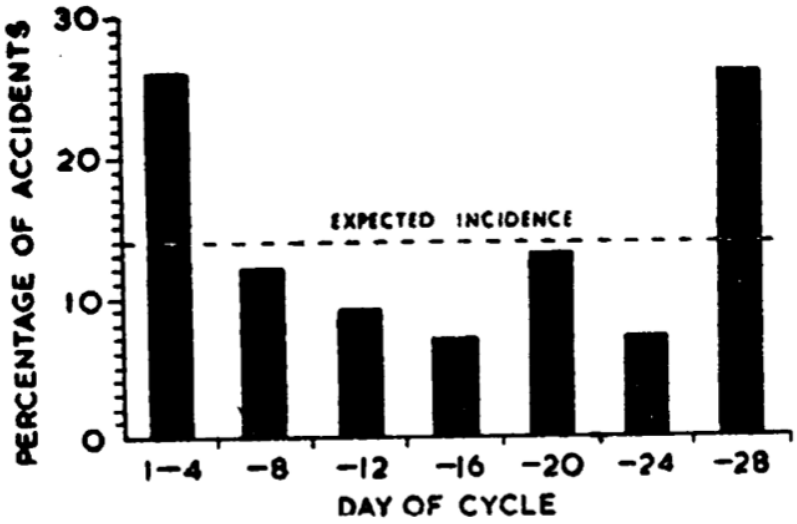


FIG. 1.—Distribution of 84 accidents in the menstrual cycle.

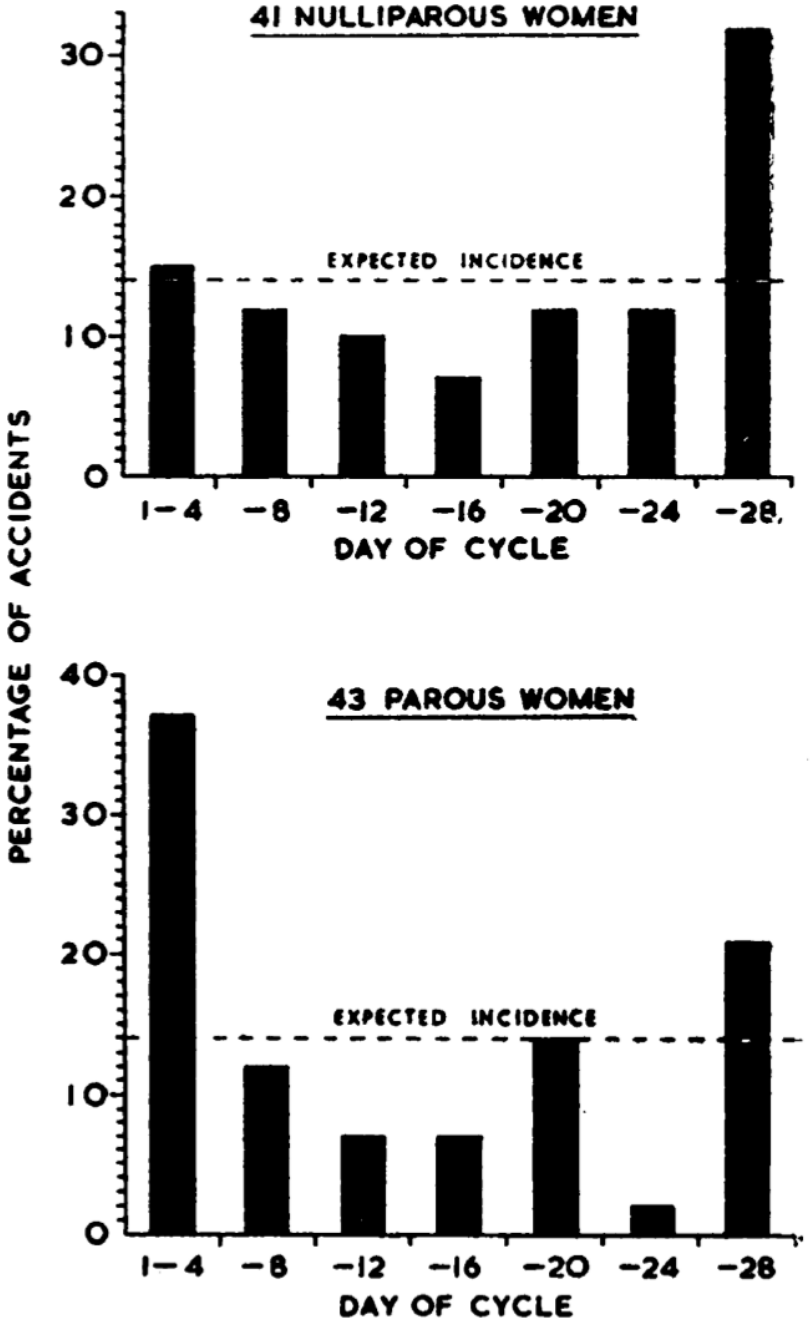


FIG. 2.—Distribution of accidents in nulliparous and parous women.

SEX DISTRIBUTION OF INFANTS

The relative number of male and female births has been a favourite subject of statistical investigation. In the course of the 24 months of the years 1908 and 1909, 93,661 infants were born in Vienna;[8] the average was therefore 3903 births per month. Of these infants, 48,172 were boys, corresponding to an average of 2007 male births per month. The proportion of male births for the whole period is

$$48{,}172/93{,}661 = 0.51432.$$

For the 24 single months, this proportion varied from 0.4990 in March 1909, to 0.5275 in August of the same year. The following table contains the 24 monthly values:

0.5223, 0.5125, 0.5141, 0.5246, 0.5126, 0.5136,
0.5187, 0.5213, 0.5105, 0.5203, 0.5124, 0.5141,
0.5143, 0.5093, 0.4990, 0.5097, 0.5140, 0.5089,
0.5129, 0.5275, 0.5178, 0.5130, 0.5177, 0.5027.

The average of these 24 values is 0.51433; the dispersion, calculated according to the rules given above, is 0.000 0533.

Now, according to Lexis's theory, we ask for the value of the expected variance; this value is calculated from the distribution in the dispersion is smaller than in that considered before. As it is very probable that the sex ratio of live births depends on race or social conditions or both, it must be expected that observations within a more or less mixed population will show a smaller dispersion than those within a perfectly homogeneous one. This hypothesis seems appropriate to explain the subnormal dispersion of the sex ratio in Vienna,

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

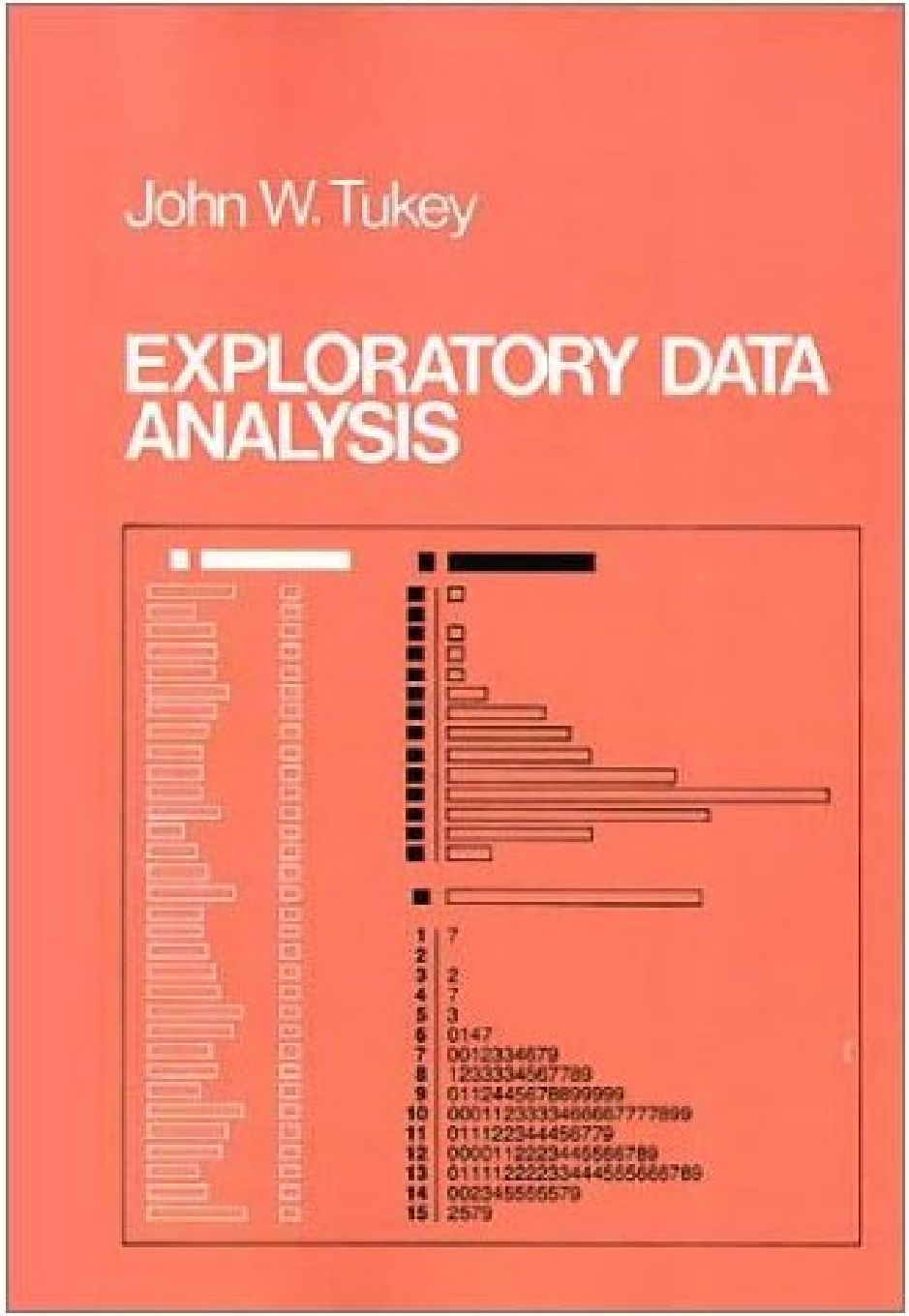
View ALL Data Sets

Housing Data Set

Download: Data Folder, Data Set Description

Abstract: Taken from StatLib library

Data Set Characteristics:	Multivariate	**Number of Instances:**	506	**Area:**	N/A
Attribute Characteristics:	Categorical, Integer, Real	**Number of Attributes:**	14	**Date Donated**	1993-07-07
Associated Tasks:	Regression	**Missing Values?**	No	**Number of Web Hits:**	62398


John W. Tukey
EXPLORATORY DATA ANALYSIS

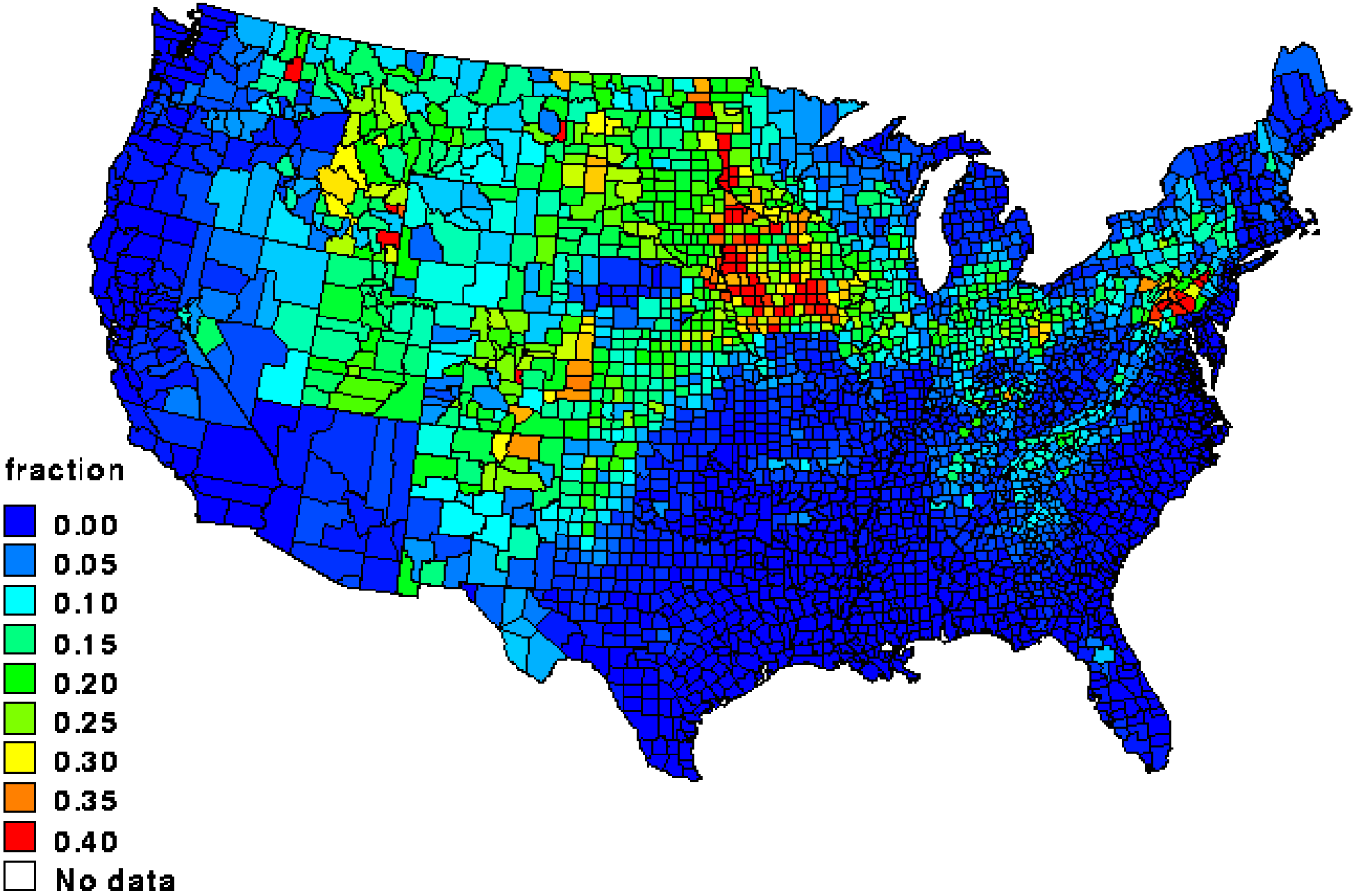
A) The MEAN MONTHLY TEMPERATURES--in °F

	Flagstaff	Phoenix	Yuma
July	65.2	90.1	94.6
Aug	63.4	88.3	93.7
Sept	57.0	82.7	88.3
Oct	46.1	70.8	76.4
Nov	35.8	58.4	64.2
Dec	28.4	52.1	57.1
Jan	25.3	49.7	55.3

B) FITTED PLACE VALUES, their RESIDUALS--and some MONTH MEDIANS

fit	46.1	70.8	76.4	70.8
July	19.1	19.3	18.2	19.1
Aug	17.3	17.5	17.3	17.3
Sept	10.9	11.9	11.9	11.9
Oct	0	0	0	0
Nov	−10.3	−12.4	−12.2	−12.2
Dec	−17.7	−18.7	−19.3	−18.7
Jan	−20.8	−21.1	−21.1	−21.1

Predicted fraction of homes over 4 pCi/L



Some reasons to believe:

- Mathematical theory (e.g., asymptotic convergence)
- Computer simulations (e.g., approximate coverage of intervals)
- Solutions to toy problems (e.g., 8 schools)
- Improved performance on benchmark problems
- Cross-validation and external validation
- Success as recognized in a field of application
- Success in the marketplace