# Lecture 14: Size-biasing, regression effect and dust-to-dust phenomena.

David Aldous

October 19, 2017

- Why do lottery winners live longer than others (on average)?
- Why do your friends have more friends than you do (on average)?
- Why do sports teams that do very well one year tend to do less well next year (on average)?
- Why are movie sequels worse than the original (on average)?
- Why does the popularity of a particular birth name tend to rise and then fall?

The theme of this lecture is that these are all "general statistical effects". In any particular case there might also be relevant causal factors, but a specific causal explanation is not *required*; believing that causal explanation is necessary constitutes one of several *fallacies*.

In Fall 2014 my department taught four lower division courses, with student enrollments 409, 197, 414, 192. The average of these four numbers is **303**. Is this the average class size? Well, **from the Professors' viewpoint**, it is.

What about the students' viewpoint? There are 1,212 students; 409 of them are in a class of size 409, and so on. The average of these 1,212 numbers is **342**. So this is the average class size **from the students' viewpoint**.

A common example is family (number of children) size. Suppose each child is in exactly one family.
[board]

Mathematically, imagine individuals placed into groups.

$p(i)$ = proportion of groups with exactly $i$ individuals
$\mu$ = mean size of groups
$q(i)$ = proportion of individuals in size-$i$ groups.

The relationship is

$$q(i) = ip(i)/\mu, \ i = 1, 2, 3, \ldots.$$

Rewriting in terms of random variables

$X =$ size of uniform random group

$Y =$ size of group containing uniform random individual

The relationship is

$$\mathbb{P}(Y = i) = i\mathbb{P}(X = i)/\mathbb{E}X.$$

This leads to several formulas: [board]

$$\mathbb{E}Y = \mathbb{E}(X^2)/\mathbb{E}X; \quad \mathbb{E}X = 1/\mathbb{E}(1/Y).$$

And unless all groups are the same size, we always have

$$\mathbb{E}Y > \mathbb{E}X$$

U.S. 2000 census data for household size

| Household size | Number of households |
|---|---|
| 1 | 27,230,075 |
| 2 | 34,418,046 |
| 3 | 17,439,027 |
| 4 | 14,973,089 |
| 5 | 6,936,886 |
| 6 | 2,636,134 |
| 7 + | 1,846,844 |
| total | 105,480,101 |

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7+ | ave | |
|---|---|---|---|---|---|---|---|---|---|
| $p(i)$ | 25.8 | 32.6 | 16.5 | 14.2 | 6.6 | 2.5 | 1.7 | 2.6 | $= \mu = \mathbb{E}X$ |
| $q(i)$ | 10.0 | 25.3 | 19.2 | 22.0 | 12.7 | 5.8 | 5.1 | 3.4 | $= \mathbb{E}Y$ |

In many settings, both viewpoints are relevant for different purposes – for instance, the distribution of class size from the Professors' viewpoint is also relevant for the provision of classrooms.

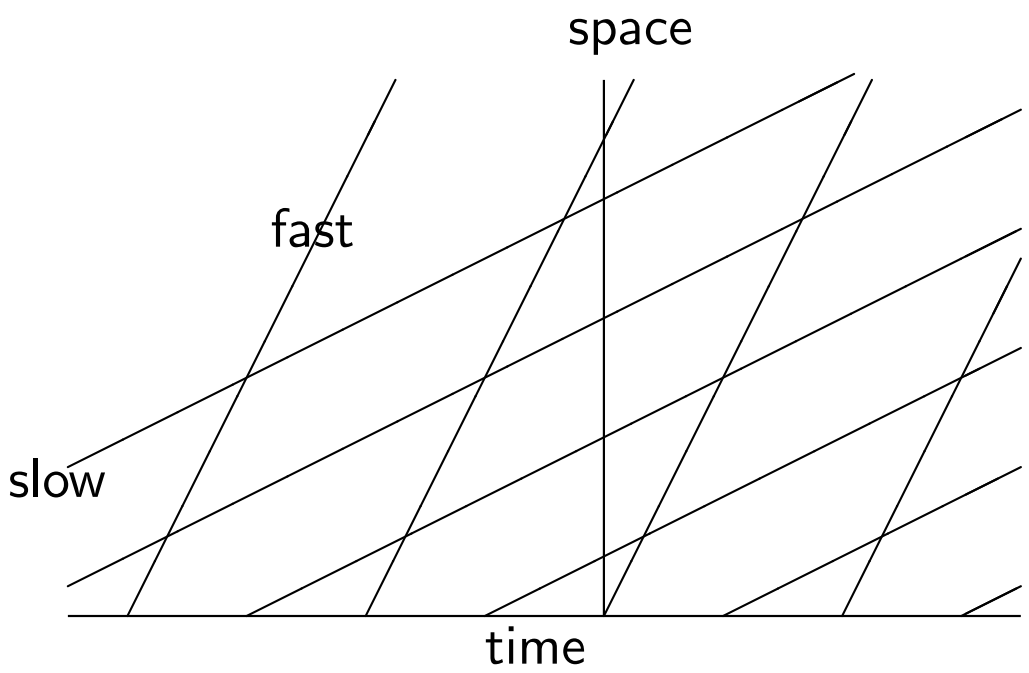
Another use of size-biasing appears in auditing financial accounts. Given a long list of bookkeeping entries, if you want to sample some to check that they match actual legitimate expenses, then it is sensible to sample with probability proportional to dollar amount, because what we are ultimately interested in is the overall dollar amount of any discrepancies.

Here is a more subtle hypothetical example. Suppose vehicles on a freeway move at different speeds, but each speed does not change in time. What is the average speed of the traffic? Here are two ways you might gather data.

(i) A police officer stands at a particular point with a *radar gun* and measures the speed of each passing vehicle for an interval of time. Take the average of those measured speeds.

(ii) Imagine an airplane that can see an long section of the freeway, and imagine a device that at one time instant can measure the speeds of all the vehicles in that section at that instant. Take the average of those measured speeds.

These will give different answers!




[board]
We get the same relationship for density functions

$$f_Y(v) = v f_X(v)/\mathbb{E}X$$

$Y$ = speed measured by police officer
$X$ = speed measured by plane.

*Assuming that winning the lottery (winning a large sum) has no effect on your future lifespan, what do we expect is the relationship between lifetime of lottery winners compared to lifetime of the general population?*

As an (unrealistic) starting model, suppose that at age 18 people decide how many lottery tickets to buy per week, do not change this number as they age, and that the choice of number has no connection with life expectancy. Then a person who lives to 78 has twice the chance to win as does a person who lives to 48, simply because they buy twice as many tickets. So in this scenario the distribution of lifetime-after-age-18 of lottery winners will be the lifetime-biased version of the distribution for the general population, and in particular the mean lifetime will be noticeably longer.

So it is a **fallacy** to argue

*we observe that lottery winners live longer than others on average, so this must be due to some cause – they become richer and happier and that makes them live longer.*

– it's just a statistical effect.

Of course our assumptions are unrealistic in detail. The age-at-winning must match the age-profile of lottery ticket buyers, which is somewhat tilted toward older adults. (see e.g. Kaplan *Lottery winners: the myth and reality*).

The statistical effect here has nothing to do with lotteries in particular. For instance if you compare

- actors who have won an Oscar
- actors who have been nominated for an Oscar but never won

then you expect the average lifetime of the former to be longer.

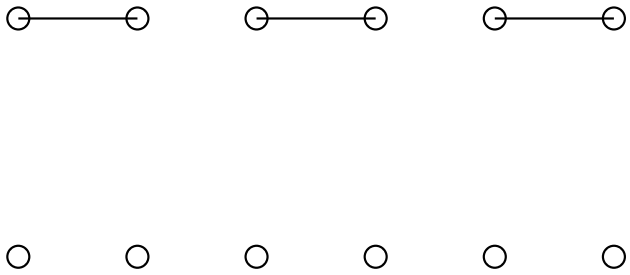
**Size biasing in social networks.**
In the simplest version, a *social network* is a graph where the vertices are individual people and the edges indicate some specific type of relationship, which for concreteness we'll call *friends*. In such a network there is a distribution

$$p_i = \text{ proportion of people with } i \text{ friends } = P(J \text{ has } i \text{ friends })$$

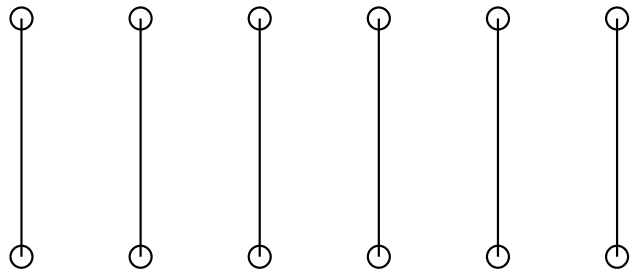
where $J$ denotes a uniform random person. Now consider a two-stage procedure; first pick a uniform random person $J$, then pick a uniform random friend $J^*$ of $J$. What can we say about

$$p_i^* = P(J^* \text{ has } i \text{ friends })?$$

This turns out to be conceptually similar to size-biasing, in that on average $J^*$ will have more friends than does $J$ – **the friendship paradox**. Let's look at two hypothetical examples.



all friends



4 out of 5 friends

[draw other edges on board]

The point of the example is that each network has $p_1 = p_5 = \frac{1}{2}$. But the values $(p_i^*)$ are different;

$$p_1^* = p_5^* = \tfrac{1}{2} \text{ (left network)}\ , \quad p_1^* = \tfrac{1}{10}, p_5^* = \tfrac{9}{10} \text{ (right network)}.$$

Thus in contrast to the basic size-biasing context, there isn't a general formula for $(p_i^*)$; it depends on the structure of the network. But a math argument [board] shows

$$\mathbb{E}(\text{number of friends of } J) \le \mathbb{E}(\text{number of friends of } J^*).$$

In words,

*your friends have more friends than you do, on average.*

*(*) your friends have more friends than you do, on average.*

Seeing this effect in data, one might be inclined to look for causal explanations. Presumably there is some measurable aspect $f$ of personality which is correlated with number of friends – so maybe you tend to have friends with higher values of $f$ than you do. But the point is that no such detailed explanation is needed; (*) is a purely statistical effect, a logical consequence of the **fact** that different people have different numbers of friends, not requiring a causal explanation of that fact.

**Math aside.** If our original choice of random person $J$ is size-biased by "number of friends", then for the random friend $J^*$ we do indeed have the property that the distribution of number of friends is the same for $J^*$ as for $J$.

**The regression effect and the regression fallacy.** This is a textbook topic . As a simple example, take a sport where teams play in leagues and have a "final standing" each year, given by the proportion of games won, in which case the average over all teams must be 0.5. The **regression effect** predicts that

- for a team with above average performance this year, say a final standing of 0.6, its final standing *next year* is likely to be *less than* this year's 0.6.
- Analogously, for a team with below average performance this year, say a final standing of 0.4, its final standing *next year* is likely to be *more than* this year's 0.4.

This effect will be more noticeable for the best and worst teams.
[show page]
The prediction is correct substantially more than 50% of the time.

Another textbook example where one would confidently expect to see the regression effect are midterm and final exams (with scores measured in "standard units", that is SDs above or below average).

The **regression fallacy** is to presume the regression effect must have some cause specific to the given context, and then to ascribe it to the intuitively most plausible potential cause. In the sports example above, for the bottom teams one might speculate that new players or coaches are hired after a bad year and that this causes the subsequent improvement. In the exams example one might speculate that students doing well on the midterm might slacken off working so hard later.

To see why such "causal" explanations are not necessary, consider a different example: two people rate the same set of movies (or wines or restaurants . . . ) on a (subjective) numerical scale, which is then converted to standard units. For the subset of movies that the first person rates around $+1$ (one SD above average) their average rating by the other person will typically be some number $\rho$ between 0 and 1. It doesn't matter which person is deemed "first", so there can be no notion here of some underlying cause of changes in time.

The phrase **regression to the mean** is often misinterpreted; here is my suggestion for how to think about it. The key point is that it is an assertion about **averages**, not about **individuals**.

[board: example with sports teams]

Why are movie sequels worse than the original (on average)?
[board: this is also *survivorship bias*]

## Dust to dust models

Here is a topic not treated in basic textbooks or popular science accounts. The type of data under consideration can be illustrated by three examples. For the first two there is extensive data and for the third there is extensive math theory.

**Given names.** The percentage of U.S. babies each year with a particular given name.

**Stocks.** One can measure the "size" of a corporation, e.g. as market capitalization, and then take its size as a percentage of the total market capitalization.

**Alleles.** The relative frequency of each *allele* of a gene.

This is **categorical** data. In the first two examples we know there are many categories (names; corporations). In the third example we are interested in the case where there are in fact many alleles;.

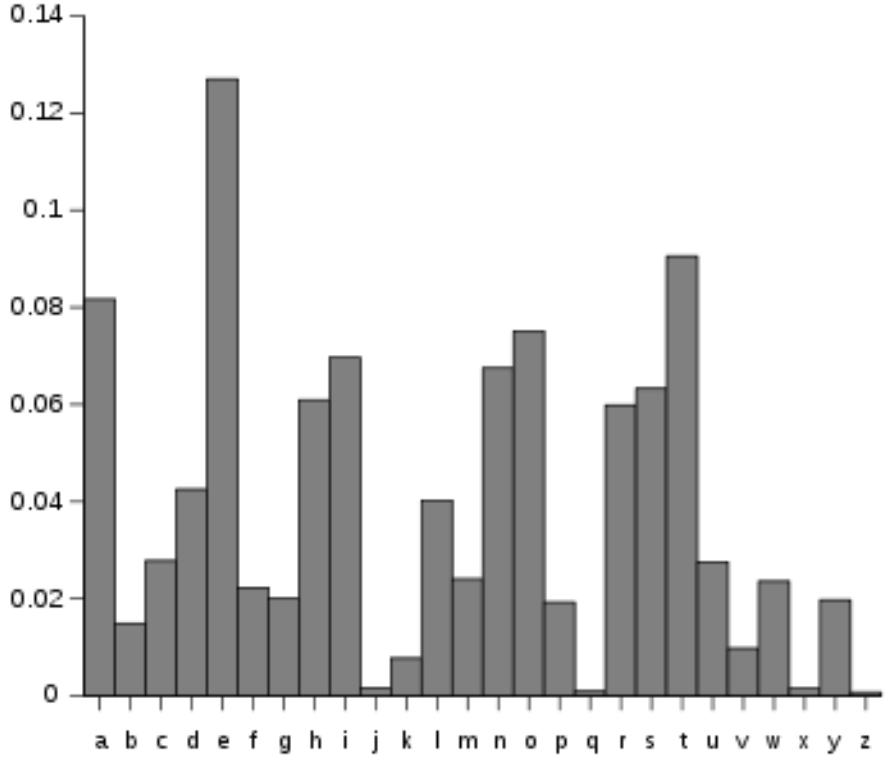
We saw categorical data in a previous lecture. Here are some copied slides.

For a probability distribution over numbers – Binomial or Poisson, Normal or Exponential – the mean or standard distribution are examples of "statistics" – numbers that provide partial information about the distribution.

Consider instead a probability distribution over an arbitrary finite set $S$

$$\mathbf{p} = (p_s, s \in S)$$

Examples we have in mind for $S$ are
Relative frequencies of letters in the English language



Relative frequencies of words in the English language
Relative frequencies of phrases or sentences in the English language
[show Google Ngram]
Relative frequencies of given names [show]

For such $S$ *mean* does not make sense. But statistics such as

$$\sum_s p_s^2$$

and

$$-\sum_s p_s \log p_s$$

do make sense.

For categorical data it is often most natural to list the categories in ranked order; largest, second largest, etc as with baby names.

What can we say about changes in frequencies, as time goes by? We can observe the changes for any particular category, but how do we pick the category to study? Three possible ways

- Pick the largest category.
- Pick a category uniformly at random.
- Pick a random size-biased category, i.e. the category of a uniform random individual.

The final scheme is the best way to see a "typical" category.
[explain]

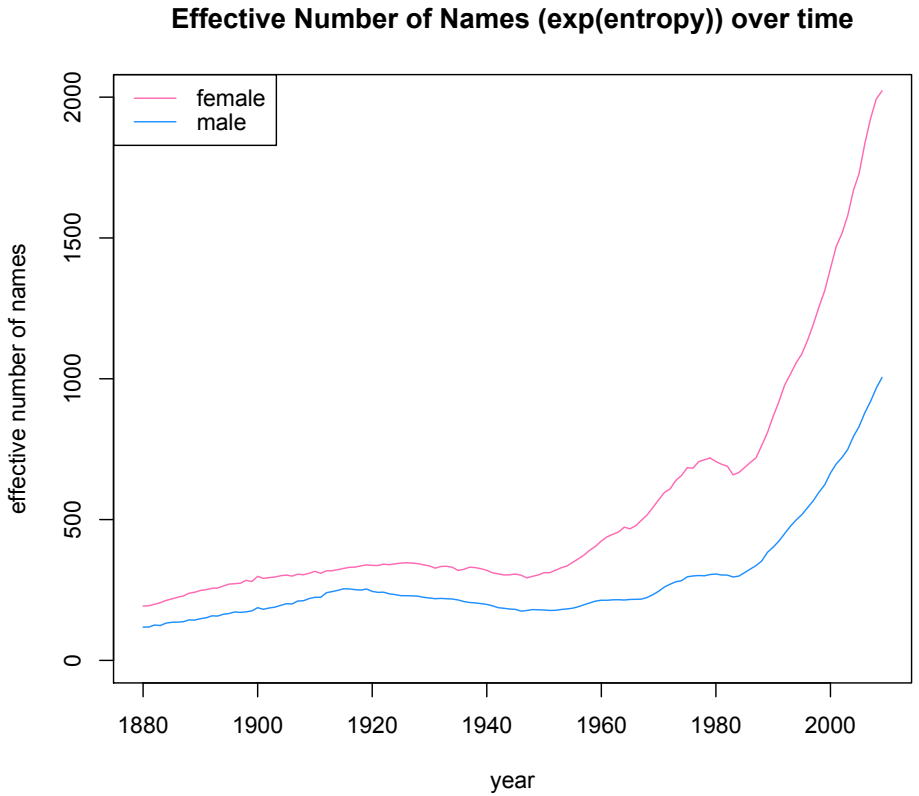
Now imagine that the category frequencies change in some unpredictable way such that
(i) the statistical properties of the distribution of frequencies of category do not change in time
(ii) there is no intrinsic reason why a particular category should have a larger or smaller frequency.
We can make simple "toy" probability models of such contexts (details not important) and in any such model, theory predicts

> *Take a size-biased pick of a category at a particular past time $t_0$. That category size will have tended to increase from a low level in the far past to a maximum at some time near $t_0$ and then to decrease toward a very low level in the far future (relative to $t_0$), if enough time has passed since $t_0$ for us to observe the latter.*

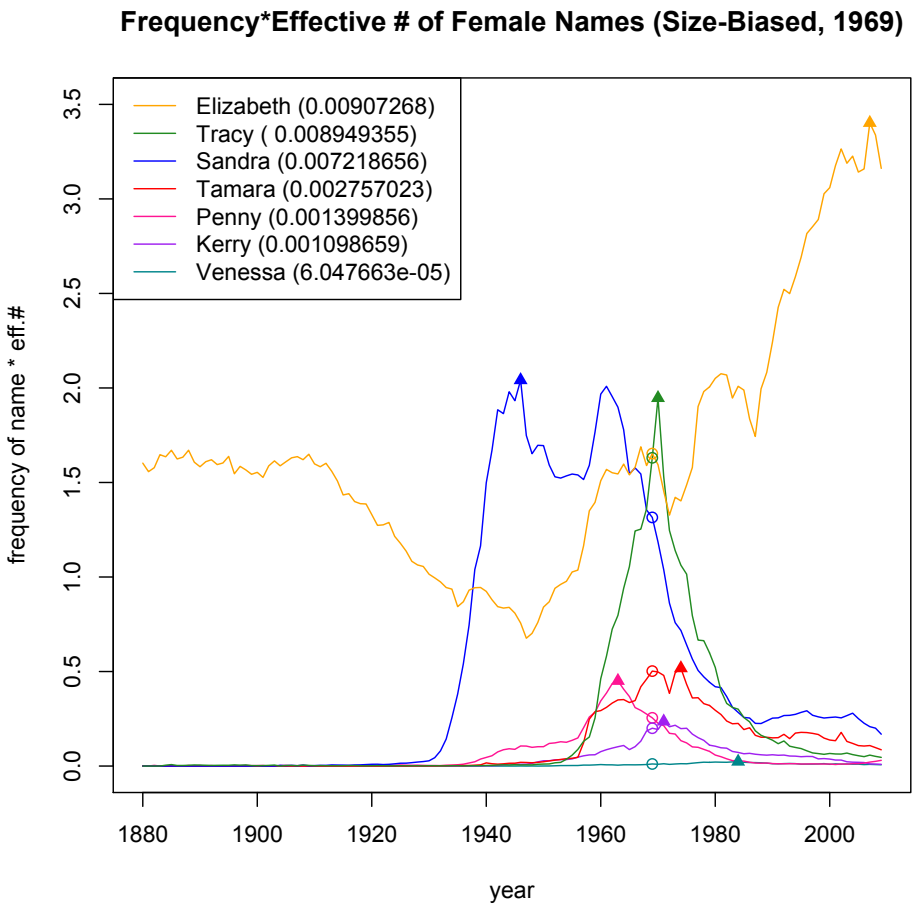
Let me call this the "dust to dust" property (no standard name). In some sense it's a variant of regression effect – saying that our initial size-biased pick "regresses" toward behaving like a uniform pick over categories, which would have a very small frequency.

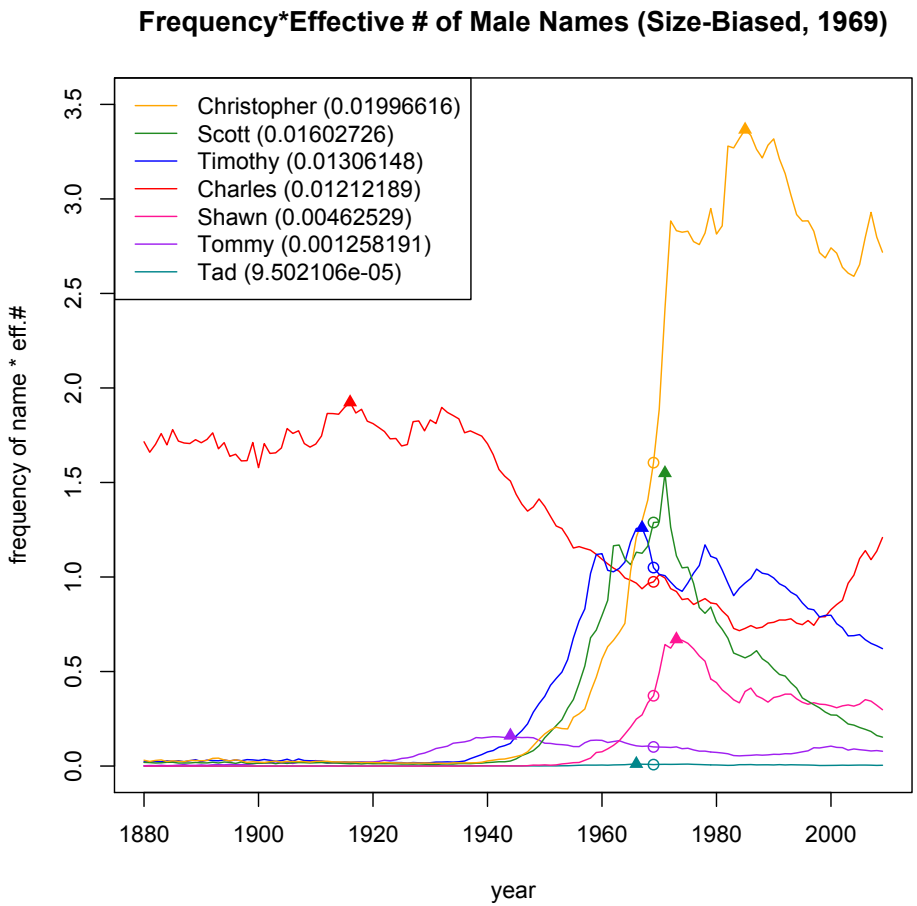
## The given names data-set

This fascinating data-set was introduced in a previous lecture where we observed a dramatic increase in diversity over the last 30 years.




To examine our "dust to dust" prediction we need to adjust for this increasing diversity effect (by simply multiplying observed frequency of a specific name by effective number of names).




The graphics show adjusted frequencies of seven names, chosen size-biased by 1969 births (i.e. the names of uniform random picks of babies born in 1969). The triangles show the year of maximum adjusted popularity of each name; the median difference from 1969 is around 10 years, demonstrating our dust-to-dust phenomenon.

## Projects.

- I surmise that when Colleges state their "average class size" they are using the Professor's viewpoint rather than the (more honest) student viewpoint. Can you find data to check this?
- Find stock market data to examine the qualitative "dust-to-dust" property.
- Find data on the $t$-year correlation for sports team winning percentage.
- The paper *You Name It – How Memory and Delay Govern First Name Dynamics* by David A. Kessler et al. contains different math analysis – how does it compare to ours?

The 2000 book *A Matter of Taste: How Names, Fashions and Culture Change* by Stanley Lieberson provides a fascinating sociological analysis.