

Lecture 2: Data Description

Graham Elliott

December 2008

The Basic Objective

The basic problem facing any analysis of data or presentation of results of some study - formal or informal - is a tradeoff between

A. Being able to get all the information out of a set of data, which one can potentially do if they have all the data, and

B. Being able to actually see the information in the data, which is quite hard if you have large sets of data.

Some Examples

- Investing in a Mutual Fund.
Barrons lists past returns (1,5 yr) but it runs many pages.

Some Examples

- Investing in a Mutual Fund.
Barrons lists past returns (1,5 yr) but it runs many pages.
- Scholastic Achievement
Thousands of students in each cohort, how do you report how well they do vs other countries etc?

Some Examples

- Investing in a Mutual Fund.
Barrons lists past returns (1,5 yr) but it runs many pages.
- Scholastic Achievement
Thousands of students in each cohort, how do you report how well they do vs other countries etc?
- Typical income of a US resident
Are income gaps widening? Cannot tell easily from a full printout of IRS tax forms

What do we do?

Descriptive Statistics:

We examine informal methods of reducing and clarifying information in data.

We regard these methods as informal as we do not give precise probabilistic answers to questions, but we move towards answers just the same.
(and when we later add the formality, we use the same or similar statistics).

- Graphs with a Single Variable

- Graphs with a Single Variable
- Summary Statistics

- Graphs with a Single Variable
- Summary Statistics
- Graphs with Multiple Attributes

- Graphs with a Single Variable
- Summary Statistics
- Graphs with Multiple Attributes
- Summary Statistics with Multiple Attributes

1. Frequency Tables and Graphs

e.g. Mutual Funds problem.

The easiest way to reduce the information is to consider not the exact returns but classes of returns,

i.e. if Firm 1 has a return of 20.1% and firm 2 has 21.2% we may lump them together and consider this as both being between 20 and 22%.

We can then count up how many companies enter each class, and graph the classes versus the number of mutual funds in each class.

1. Frequency Tables and Graphs

I have data on 194 mutual funds.

The returns range from 10% to 38.6%.

Take 'bins' of every 5 from 9.99 to 49.99.

All the observations below 14.99 into the '14.99' bin etc, and all the ones over 49.99 into the 'More' bin (so the '14.99' bin is 9.99-14.99 etc).

I get the following summary of the data in more manageable form.

Frequency Table

Bin	Frequency
$x < 10$	0
$10 \leq x < 15$	1
$15 \leq x < 20$	3
$20 \leq x < 25$	11
$25 \leq x < 30$	54
$30 \leq x < 35$	91
$35 \leq x < 40$	18
$40 \leq x < 45$	11
$45 \leq x < 50$	5
$x \geq 50$	0

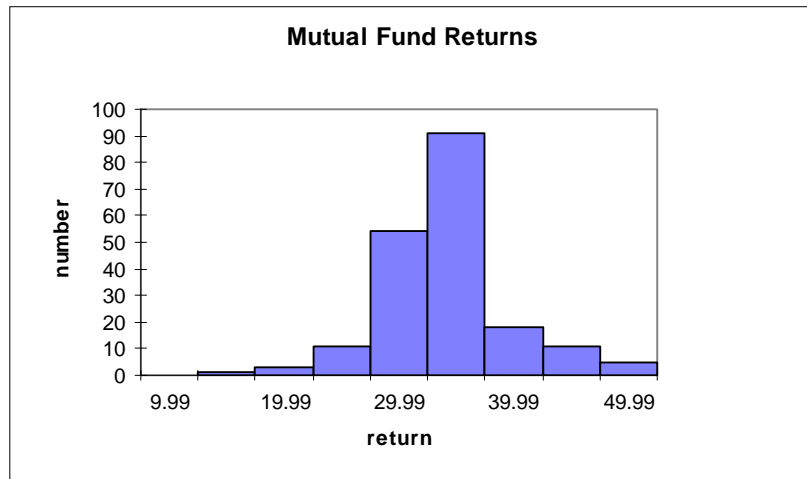
Frequency Table

We can read off quite quickly most funds are around the 25-40% return range (over half in fact), there are a few lousy performers and a few quite good performers. From the alphabetical listing in the newspaper such results are just not obvious.

'Frequency table' since each number is an estimate/calculation of the frequency of members of that group.

Histogram

An even clearer way to present these numbers is to graph them.



A key to drawing these carefully is to ensure that the areas of each block are compatible.

Here I did this by making each category the same width (since they are for equal blocks of 5).

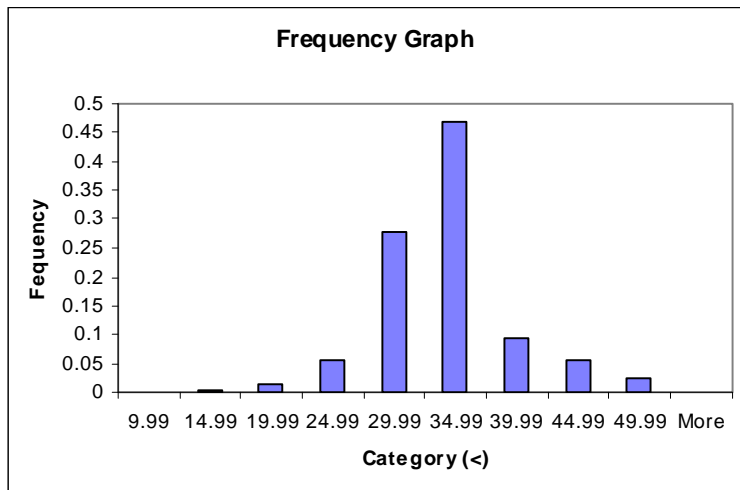
Heights are also constructed so 50 is twice 25 etc.

Not doing this leads to a visual misrepresentation of the data.

Consider changing this to percentages, i.e. divide the height by $n=194$. We can make the widths equal to one, which makes the area under the curve add up to one. (the reason is that since proportions add to one, and base is one, then base times height is one).

Histogram

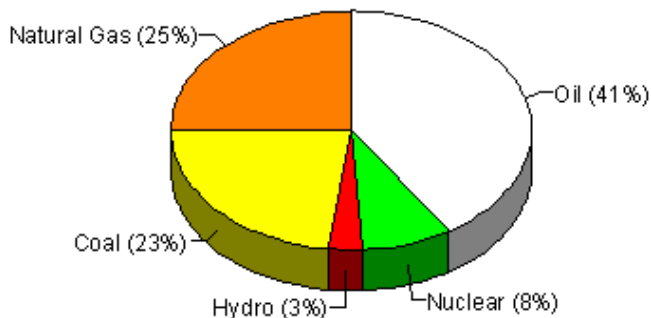
We can redraw the graph



Categorical data is when the x value has no obvious ordering. Frequency graphs are still useful, either as we did them or alternatively as pie charts. The same ideas in terms of making areas proportional still matter, although it is basically impossible to fail at this for a pie chart.

US Fuel Consumption

Figure 1.1
U.S. Energy Consumption by Source
1992



Source: EIA

Misrepresenting Data with Histograms

It is important to mainstream media to trash younger generations (older readers love it!).

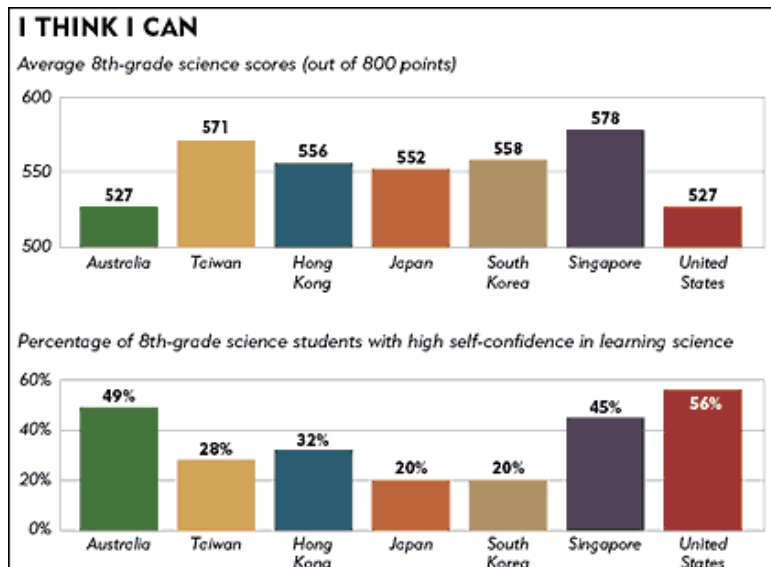
The National Center for Education Statistics reports “Trends in International Mathematics and Science Study (TIMMS).

It is a cross country comparison.



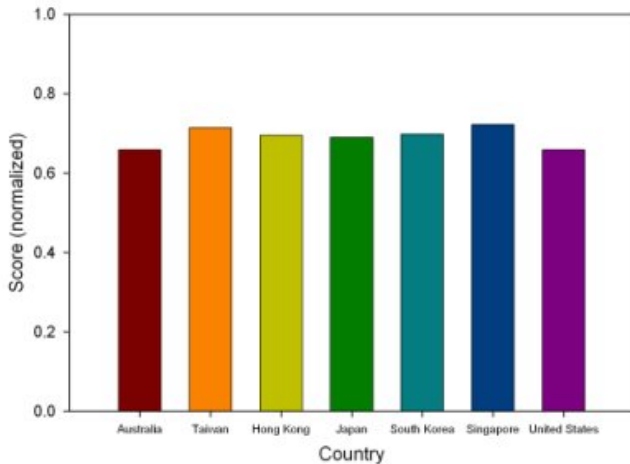
International Science and Mathematics

This summary appeared in the 'Atlantic Monthly'.



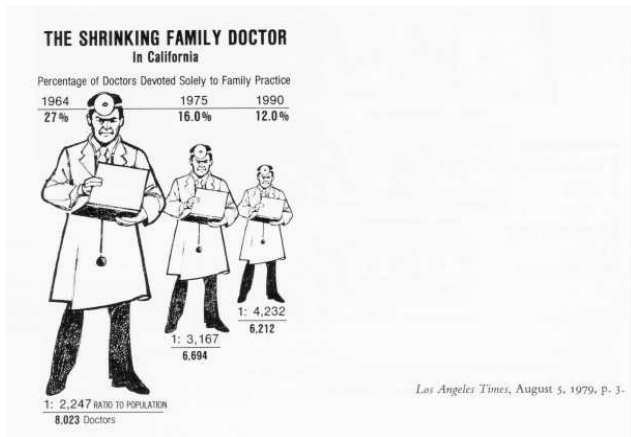
International Science and Mathematics

When we get the scale correct, a different picture emerges.



Picture Histograms

The problem here is that the areas are not kept proportional (From 'Tufté (1983, p. 69)').



Cumulative Frequency Tables

Rather than compute for each bin, i.e. compute numbers where $10 \leq x < 15$, could compute the cumulative number of observations below any value.

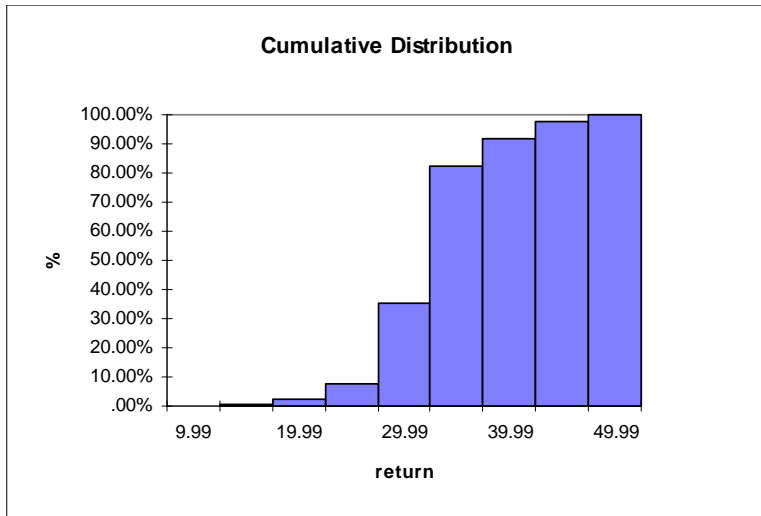
In this case we have entries like $x < 15$, which simply sums all the values up to that point in the regular frequency table.

Cumulative Frequency Tables

Bin	Frequency	Cumulative Frequency
$x < 10$	0	0
$10 \leq x < 15$	1	1
$15 \leq x < 20$	3	4
$20 \leq x < 25$	11	15
$25 \leq x < 30$	54	69
$30 \leq x < 35$	91	160
$35 \leq x < 40$	18	178
$40 \leq x < 45$	11	189
$45 \leq x < 50$	5	194
$x \geq 50$	0	194

Cumulative Frequency Graph

We can obviously graph this as well



Box Plots

These are useful when we have Numerical data that is comparably across a few categories.

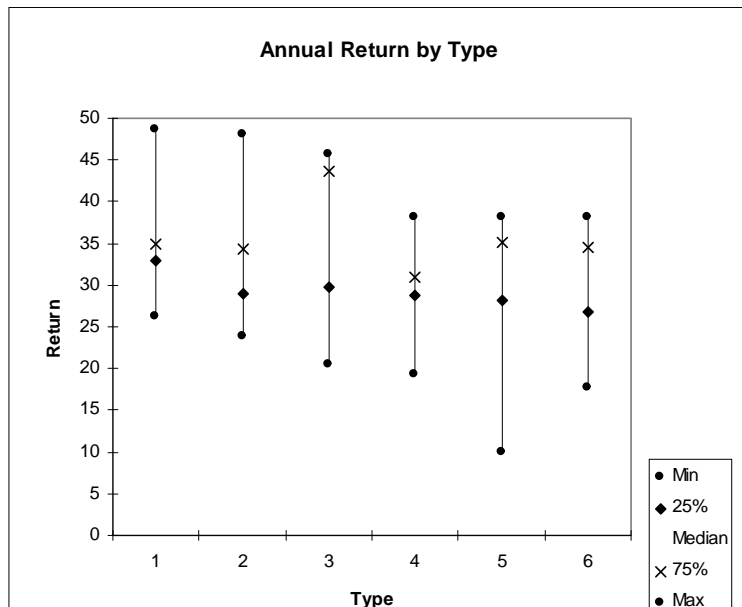
e.g. Mutual Funds investing problem.

Large cap, medium or small cap?

Growth or Value orientation?

Here we have six categories of numerical return data.

Boxplots



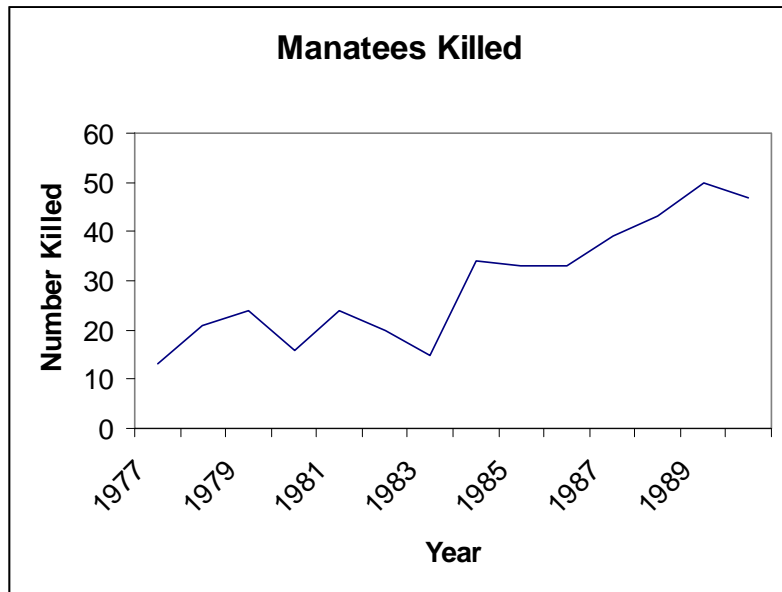
Time Series Plots

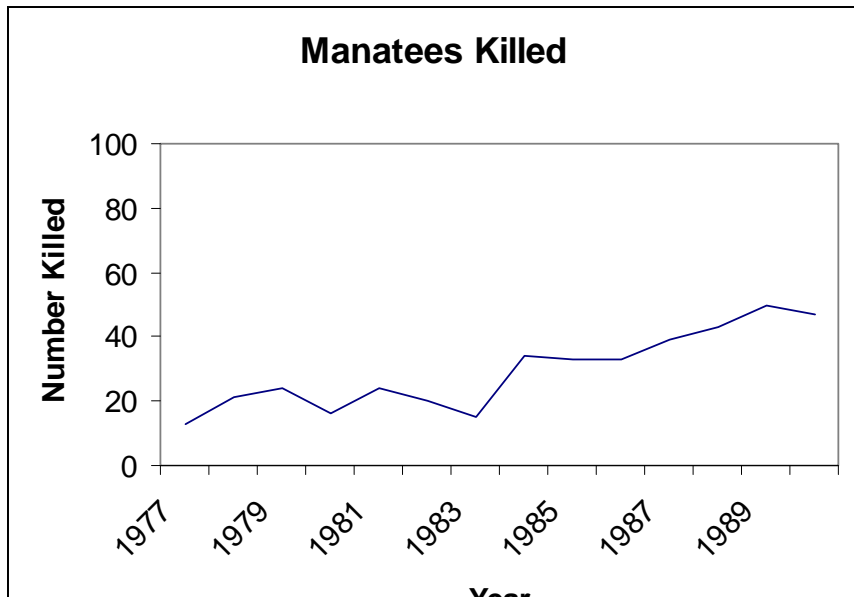
When data is ordered by time, it is often much more insightful to graph it against time.

This is straightforward, however it can still lead to misleading results if you play around with the scaling.

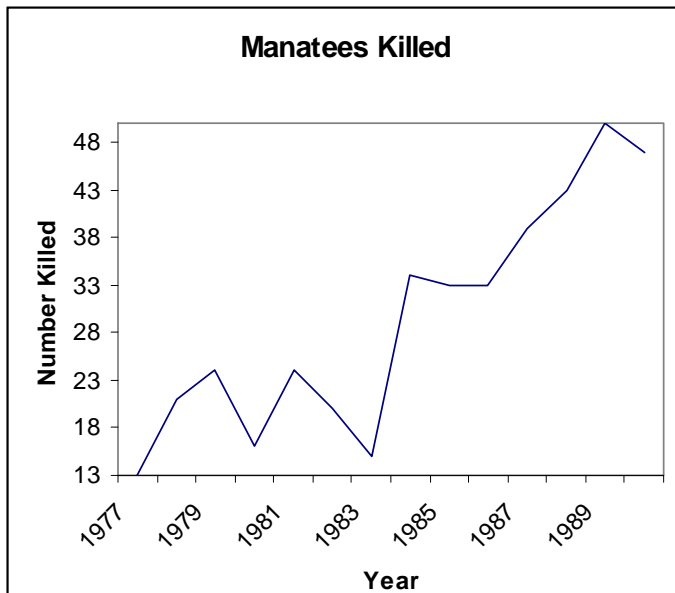
e.g. Manatee deaths in Florida

Manatee Deaths in Florida

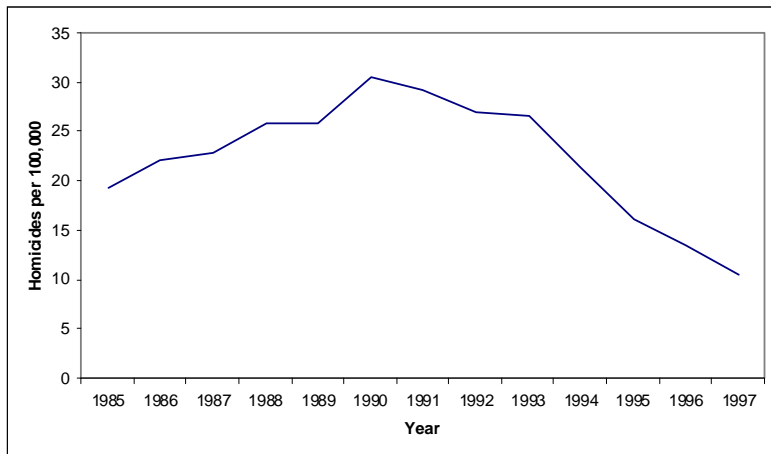




Environmentalist Graph



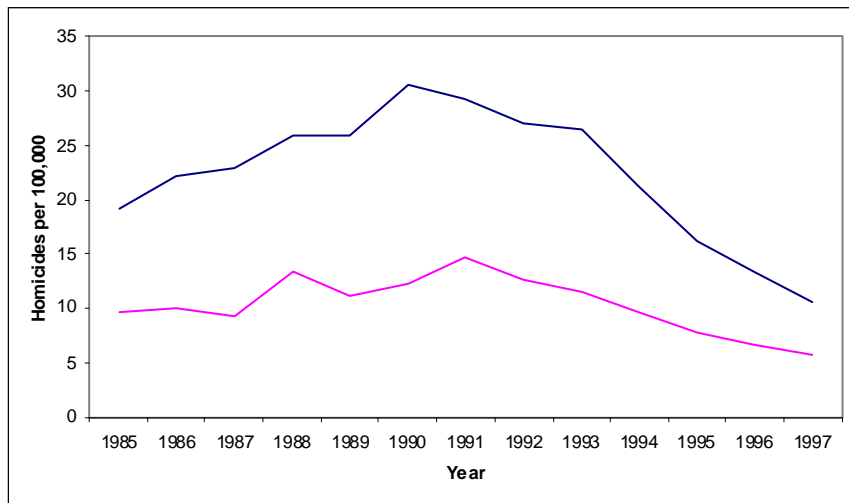
New Policing and Crime



Why the drop?

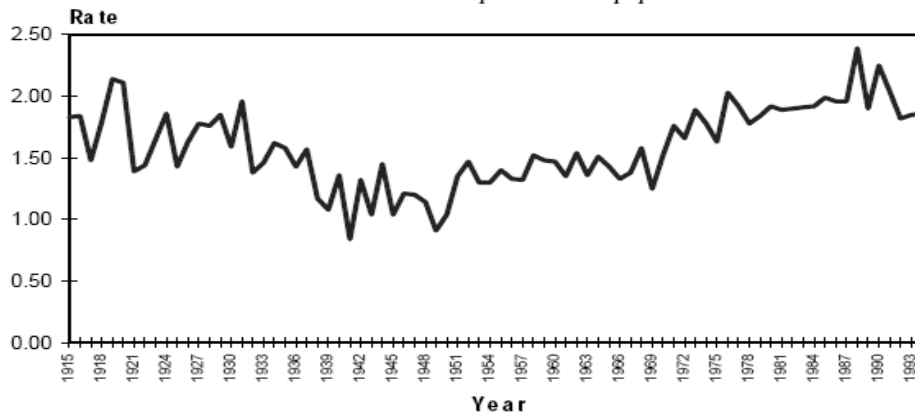
“The men and women of the NYPD are principally responsible for the dramatic crime decline that continues today ...” Bratton (1998)

Far reaching effects



Very far reaching effects

Figure 1. Trends in the homicide rate, Australia 1915-1994,
Rate = Number of homicides per 100 000 population



Source: Australian Bureau of Statistics, *Causes of Death*, compiled in Mukherjee, Scandia, Dagger & Matthews (1989 with updates).