Economics 120A

2. Descriptive Statistics

Graham Elliott

Plan for this section of the course

We want to summarize data either visually or with summary statistics in order to understand what the data can tell us – either about the data in general or answer a specific question.

1. Graphs with a Single Variable

2. Graphs with Multiple Attributes

3. Summary Statistics for a Single Variable

4. Summary Statistics with Multiple Attributes

The Basic Problem – Too much data

For most situations with data, we have a lot of it (even before BIG data).

There is a lot of detail in the data, part of the job of statistics is to find a way to see what the data is telling us, either

(a) We have a theory about something and are trying to see if the data agrees, or

(b) More generally understand or learn about what is going on.

Example: Energy data

CAISO-supply-20190929.csv

The Basic Problem – Too much data

Other examples:

1. Scholastic Achievement.
Huge number of students to evaluate, can look at test scores, family background variables etc.

2. Income Inequality Evaluation
Could examine all tax records, but this is a huge volume of data.

3. Choosing a mutual fund.
There are thousands of mutual funds with returns histories as well as measures of diversification etc. How to choose one?

4. Marketing.
Analysts can access huge amounts of social media data. What to do with it?

The Basic Problem – What do we do?

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.

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. Think of them as 'estimates'.

(and when we later add the formality, we use the same or similar statistics).

Frequency Tables and Graphs

For a single variable we can consider a frequency graph or table, often as a variant of a histogram.

How we build this depends on the values the data can take on:

(a) Finite number of unordered values (often called qualitative data)

(b) Finite number of ordered values (often called ordinal data)

(c) An infinite number of ordered values (often called continuous data).

Frequency Tables and Graphs

Example: Finite number of ordered data.

I have 1599 observations on a quality ranking (1-10) for Spanish red wine.

Ordered because 10 is better than 2 for example.

Finite as there are only 10 possible values in the dataset.

Source: https://archive.ics.uci.edu/ml/datasets/Wine+Quality
Data from UCI machine learning repository

Frequency Graphs and Tables

Just to show a part of the data.

Quality is the last of the variables we are observing.

```
# import data

bd = pd.read_csv('winequality-red.csv', sep=";")
```

```
bd.head()
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

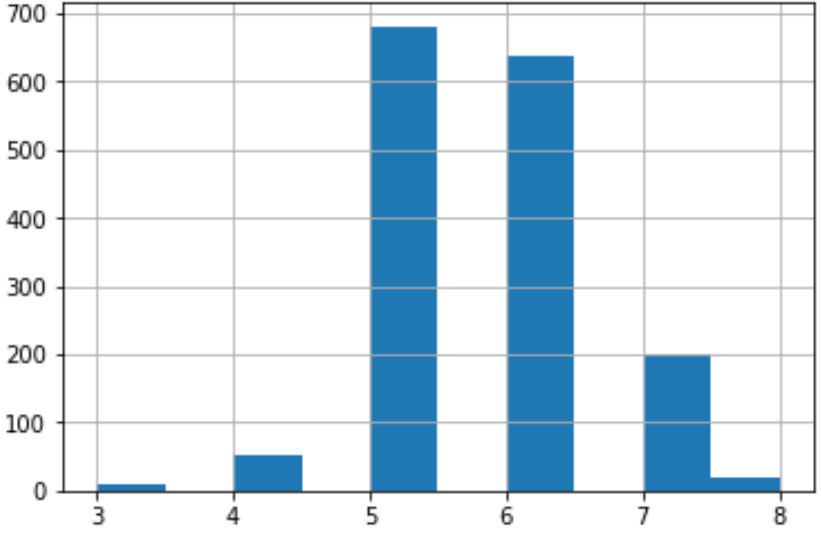
This is undertaken in python using Jupyter notebooks, the file and data are in Canvas.

Frequency Graphs and Tables

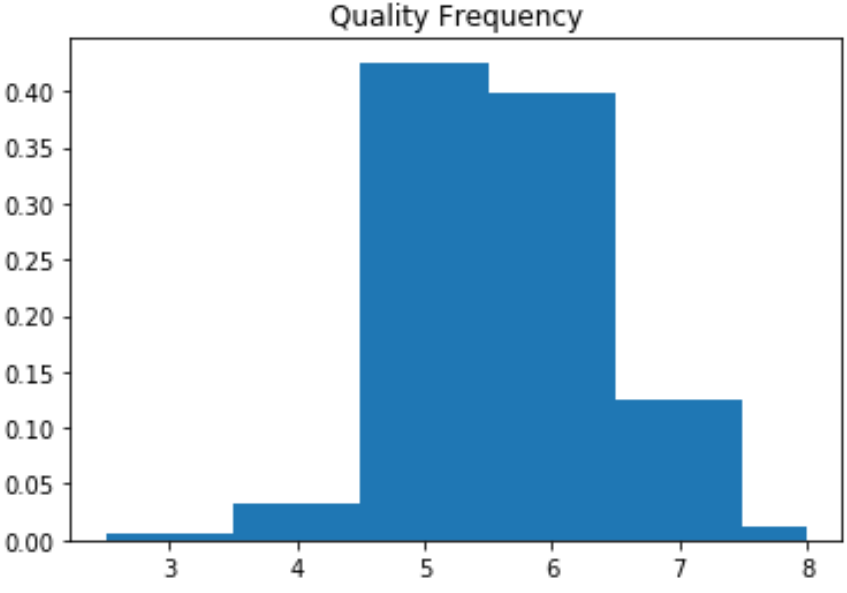
noun **STATISTICS**

a diagram consisting of rectangles whose area is proportional to the frequency of a variable and whose width is equal to the class interval.

```
import matplotlib.pyplot as plt
bd['quality'].hist(bins=10)
plt.show()
```



Frequency Graphs and Tables



Two changes are made here:

(a) Divide the height by the total number, so now the area under the histogram adds to one (frequencies not outcome number)

(b) Removed the spaces between, so the areas are easier to see.

Frequency Graphs and Tables

Mathematically what does this mean?

The frequency is the number of values at each value divided by the number of total values.

i.e. denote the data as x_i for i=1,..,n. So there are n data points. Then the frequency at any possible value x (any of the values x_i can take on) is

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} 1(x_i = x)$$

possible outcome

where $1(x_i = x)$ is equal to 1 if the condition is true and zero otherwise. Do this for each possible x.

$$1(A) = \begin{cases} 1 & \text{if } A \\ 0 & \text{if not} \end{cases}$$

estimate of distribution

Frequency Graphs and Tables

When the number of potential outcomes is large, or as is often the case could be infinite (although data never is of course) then we need to do a little more work. There are too many possible x's.

We can take the possible outcomes and 'bin' them, choosing ranges and treating all the values inside that range as the same. Then graph this as our frequency histogram (take the proportion of the data inside each bin and graph that).

How to choose the number of bins?
(a) Usually 5-20 is the suggested number
(b) There are automated rules if you really want
(c) Just try a few and look to see if the picture clearly does what you want!

Now for any bin from x^{j-1} to x^j we have the frequency calculated by

range { $x^1 - x^2$, $x^2 - x^3$

$$\mathrm{f}\left(x^{j-1} \le x < x^j\right) = \frac{1}{n}\sum_{i=1}^{n} 1\left(x^{j-1} \le x_i < x^j\right).$$

Frequency Graphs and Tables

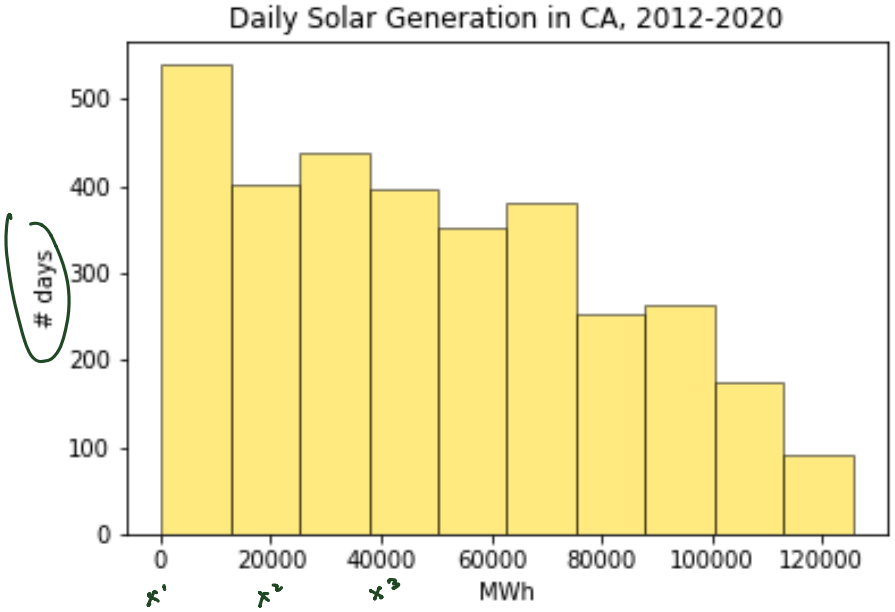
Solar Energy on the Grid in California

The data I have is Solar energy generated by solar energy plants (PV and thermal) and sells through CAISO markets. So it ignores rooftop solar.

I have daily data from 2012 until the end of last year, which generally covers the rise of this source of energy for electricity in California.

It is growing as a source but both faces challenges and creates challenges for the energy marketplace.

Frequency Graphs and Tables

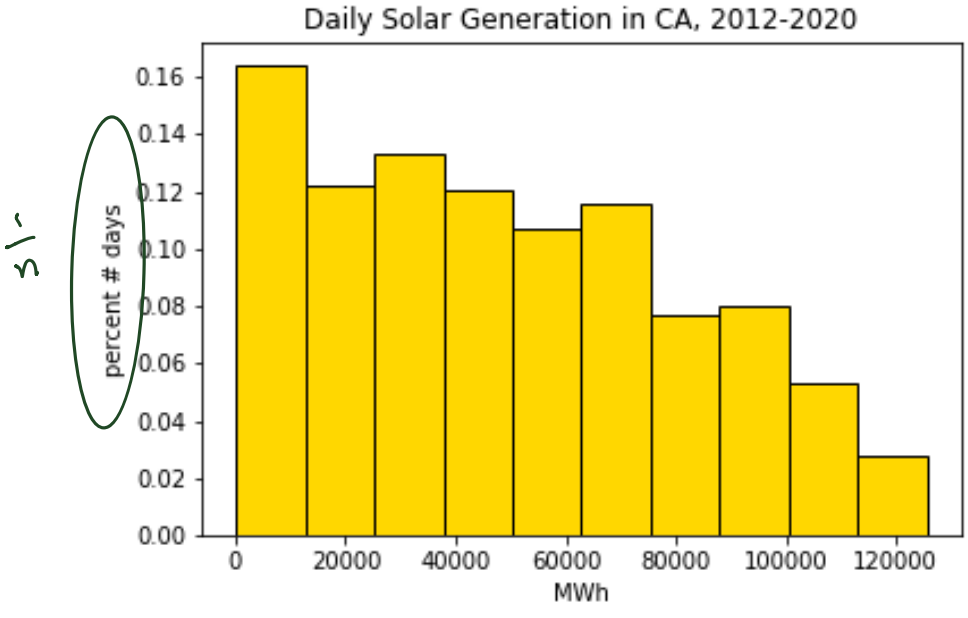


The data is daily solar generation from 2012-2020. (From CAISO)

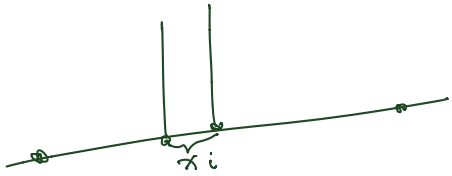
The whole data is unwieldly, but this makes it clear how the spread of solar energy production differs across the days. It makes a lot of things unclear as well, as we will see (information is lost).

Frequency Graphs and Tables

It is generally easier to see what is going on with a frequency histogram though.



Frequency Graphs and Tables

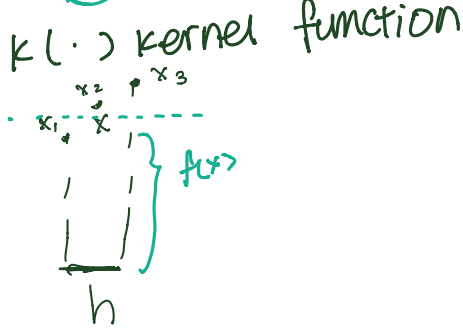


Why use bins? The histogram is

$$f\left(x^{j-1} \le x < x^{j}\right) = \frac{1}{n}\sum_{i=1}^{n} 1\left(x^{j-1} \le x_i < x^{j}\right).$$

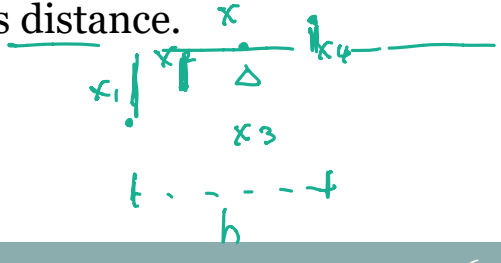
Why not have a value at each x? We do not have enough observations at each x, so how about taking a weighted average of the ones near x.

We could use



$$f(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right).$$

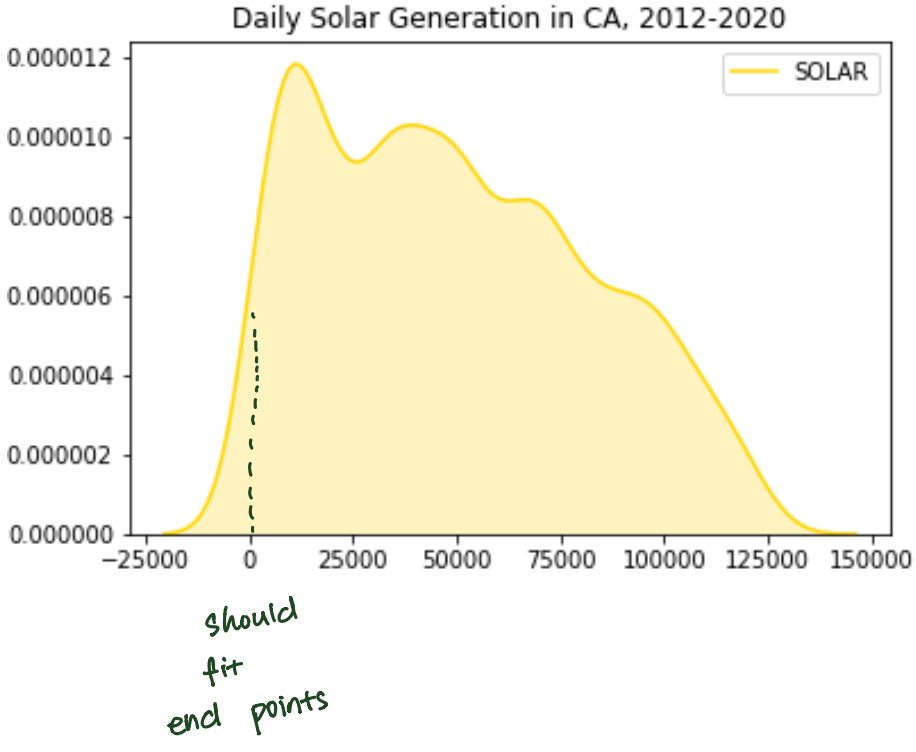
Here K(.) is a function (Kernel) that gives lower weights the further the data is from x and h is the bandwidth that stretches this distance.



We do this for every x (easy on a computer)

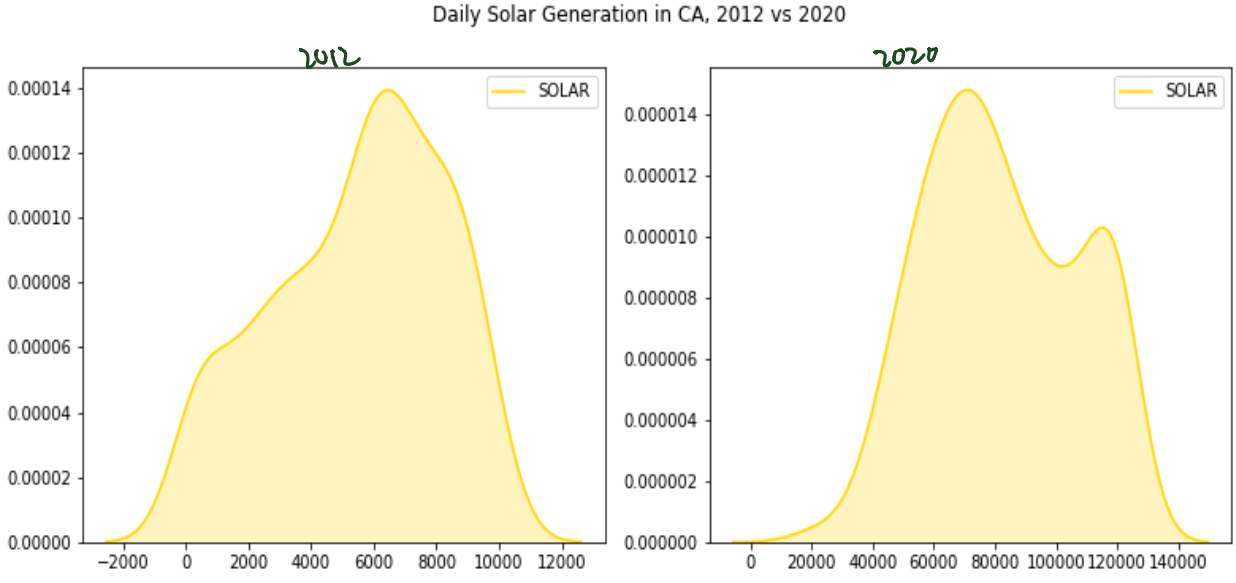
Frequency Graphs and Tables

Kernel Density Estimation (KDE).



Frequency Graphs and Tables

Looking at multiple pictures allows us to make comparisons.

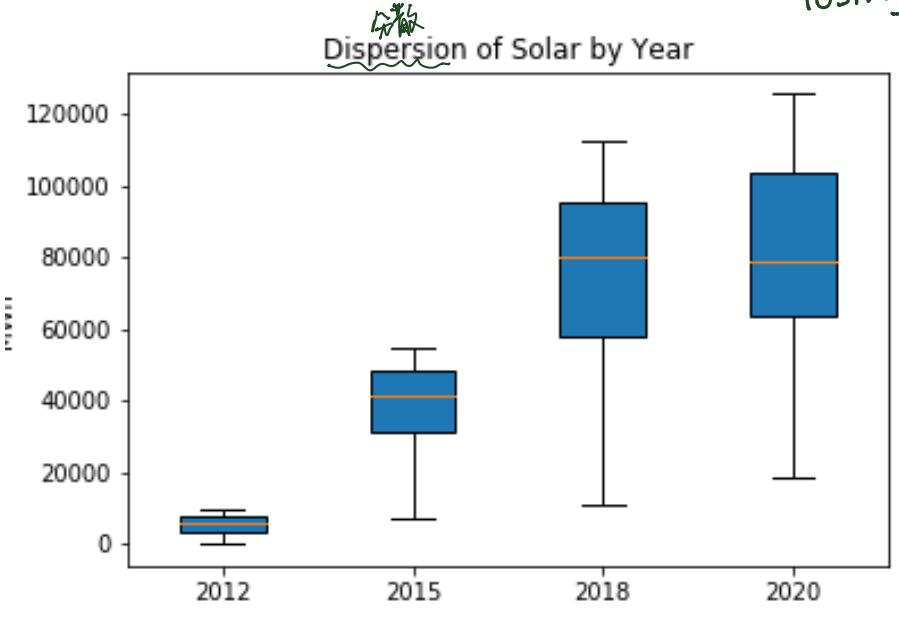


Frequency Graphs and Tables

This gets a bit tedious and over informative when there are many comparisons.

One could consider dropping more information and using boxplots.

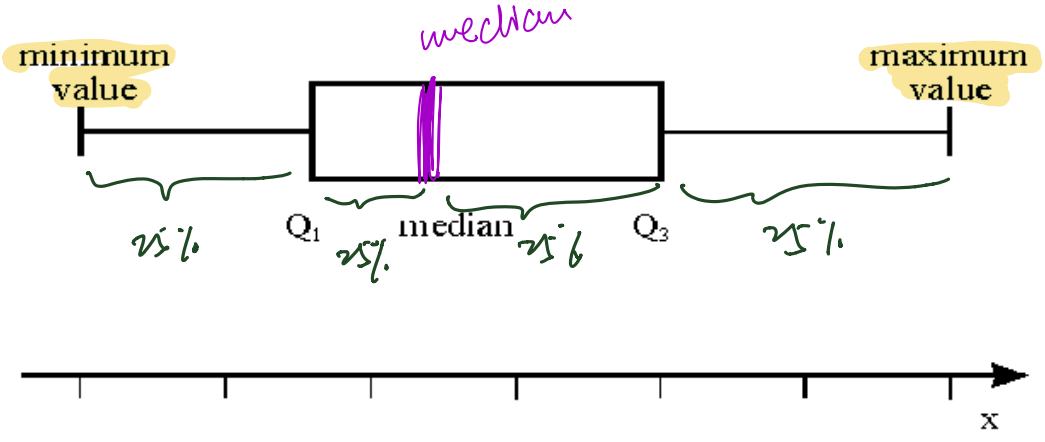
losing more info.



Box Plot

Box Plots

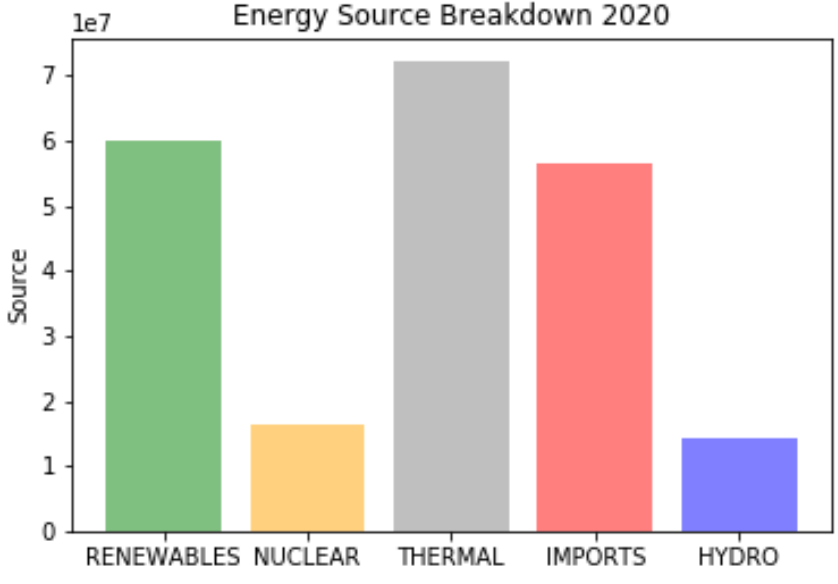
Much Simpler view of data. Useful for comparisons.



We will discuss calculation of the median and the interquartile ranges in the next subsection.

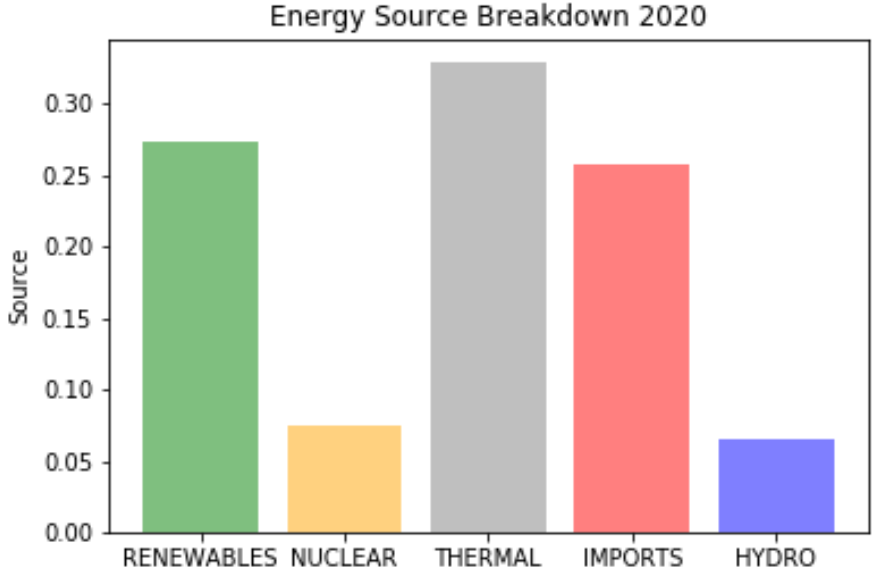
Frequency Graphs and Tables

Often the 'X axis' data is categorical. This creates no problem, we can still use a histogram with the same rules but now there is no obvious ordering along the X axis variables.



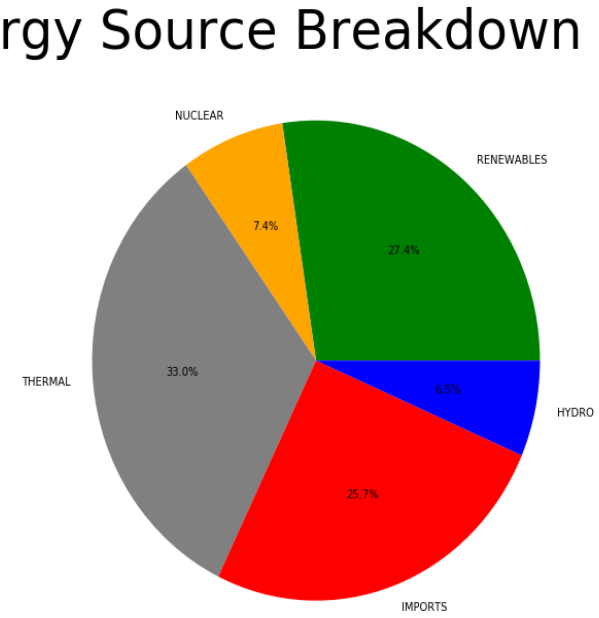
Frequency Graphs and Tables

If we do a 'frequency' plot, it is again so that the total area of the graph is equal to one. This gives the percentage for each source.



Frequency Graphs and Tables

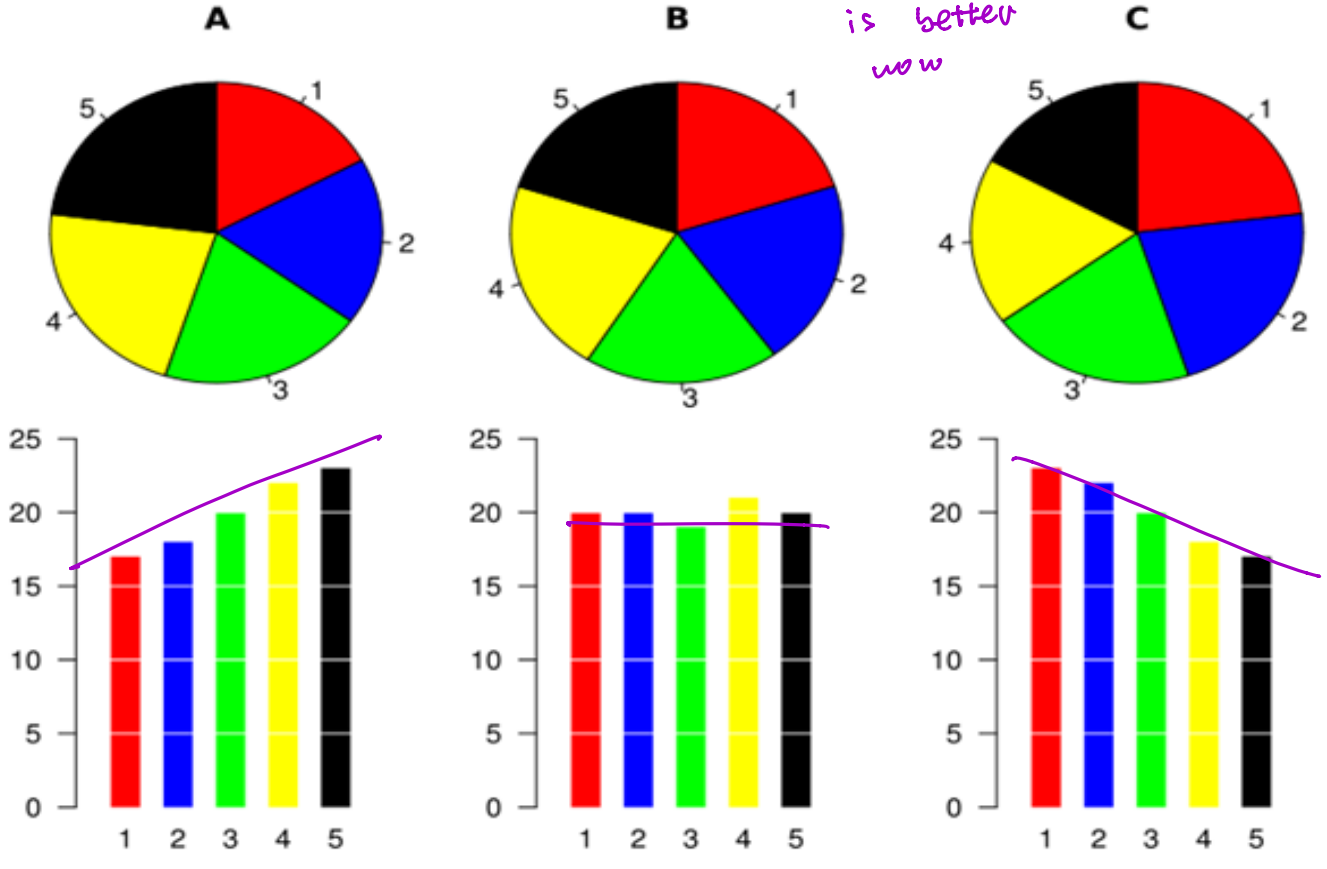
In this case often a pie chart is used, which is really a variation of a frequency histogram.



Frequency Graphs and Tables

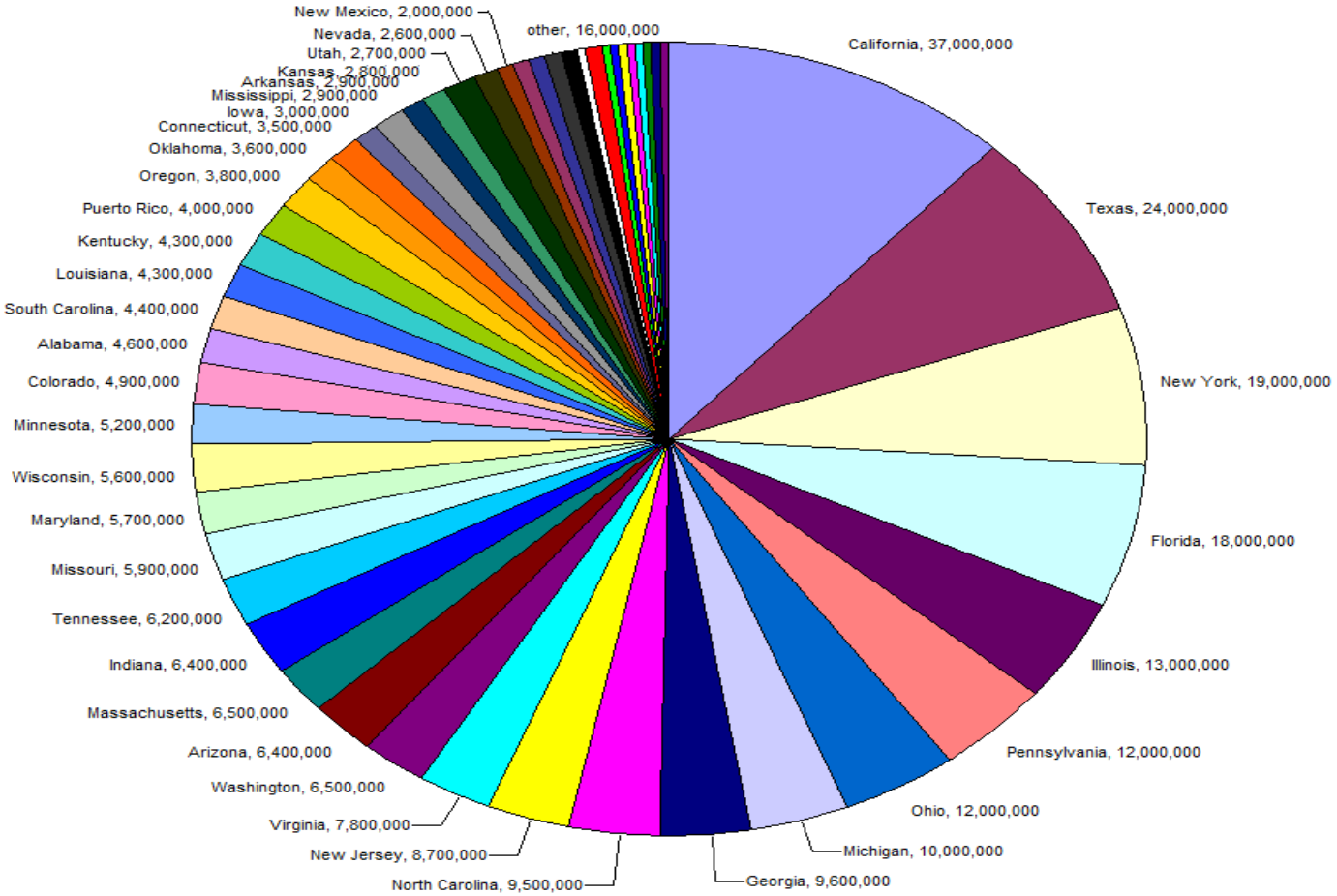
Pie Charts are fine but bar charts are often more informative when the differences are subtle.

histogram is better now



Frequency Graphs and Tables

Pie charts need to be relatively simple



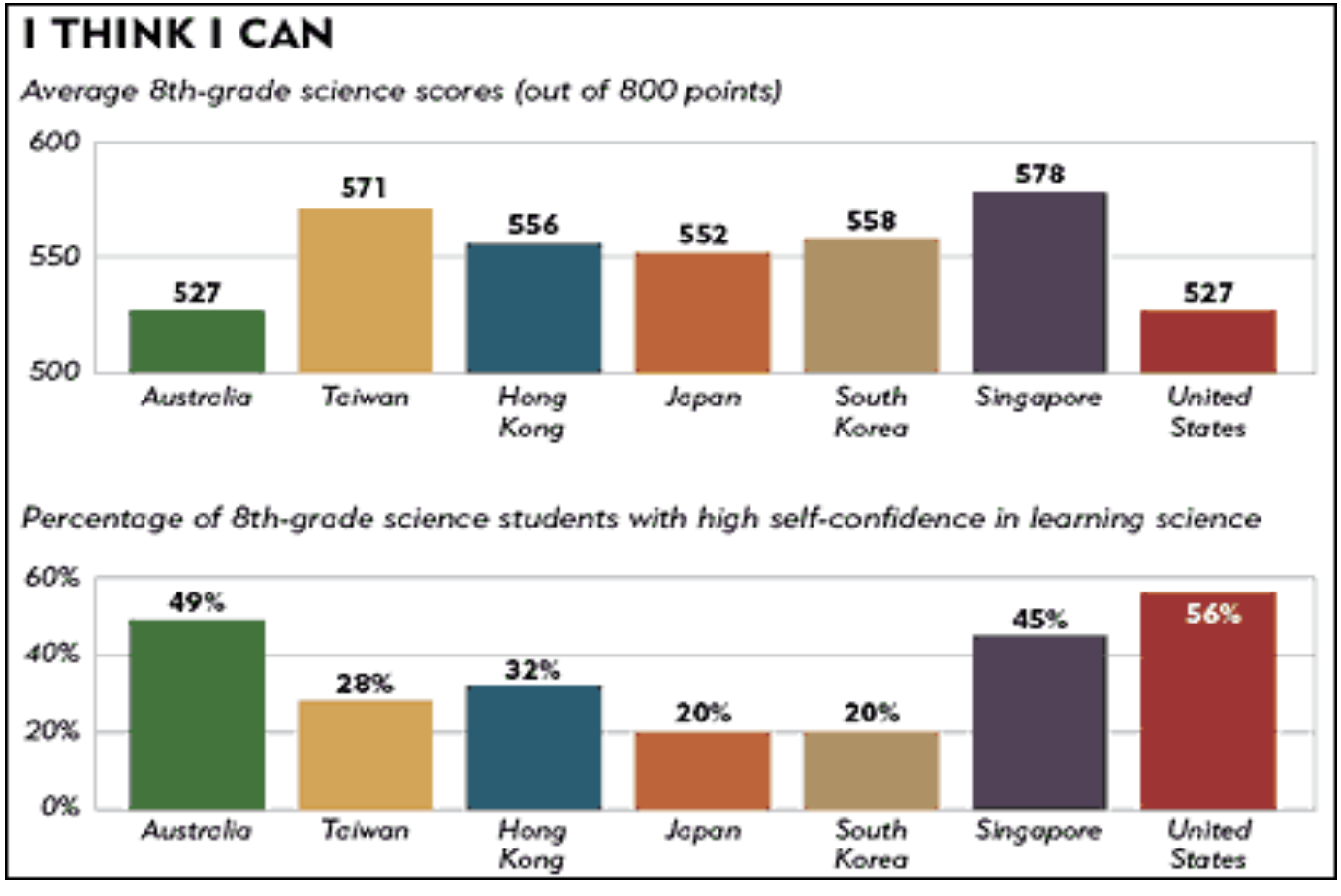
Misrepresenting Data with Histograms

It is important for much of the 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.

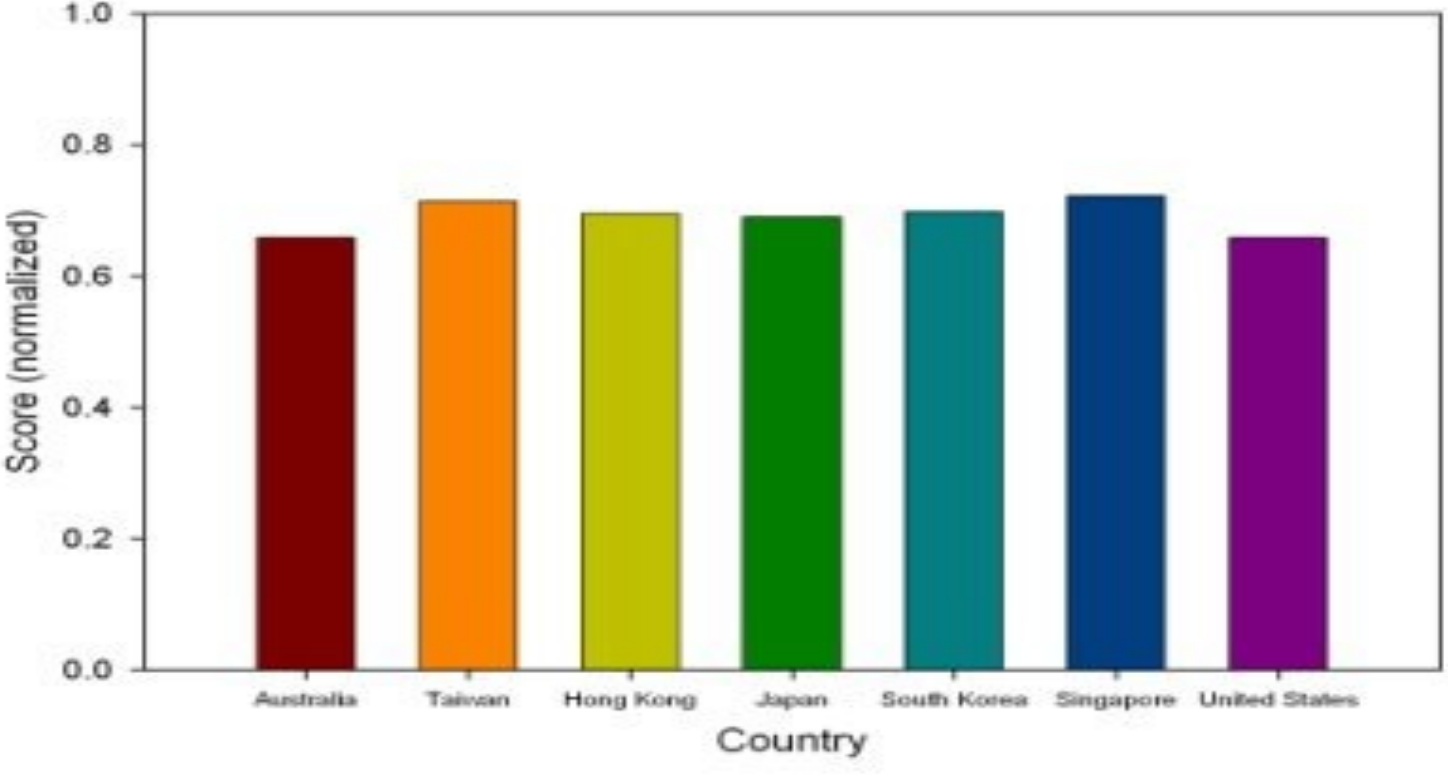
Misrepresenting Data with Histograms

This summary appeared in the `Atlantic Monthly'



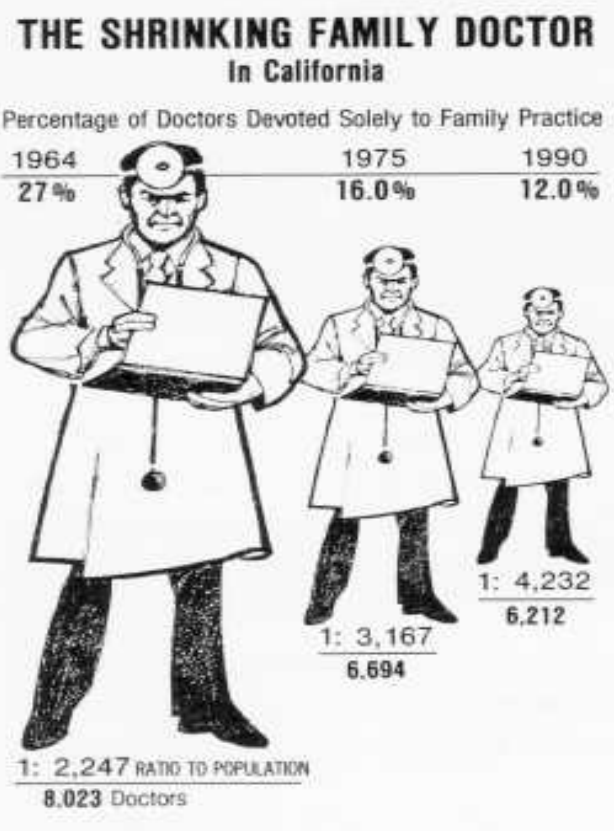
Misrepresenting Data with Histograms

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



Misrepresenting Data with Histograms

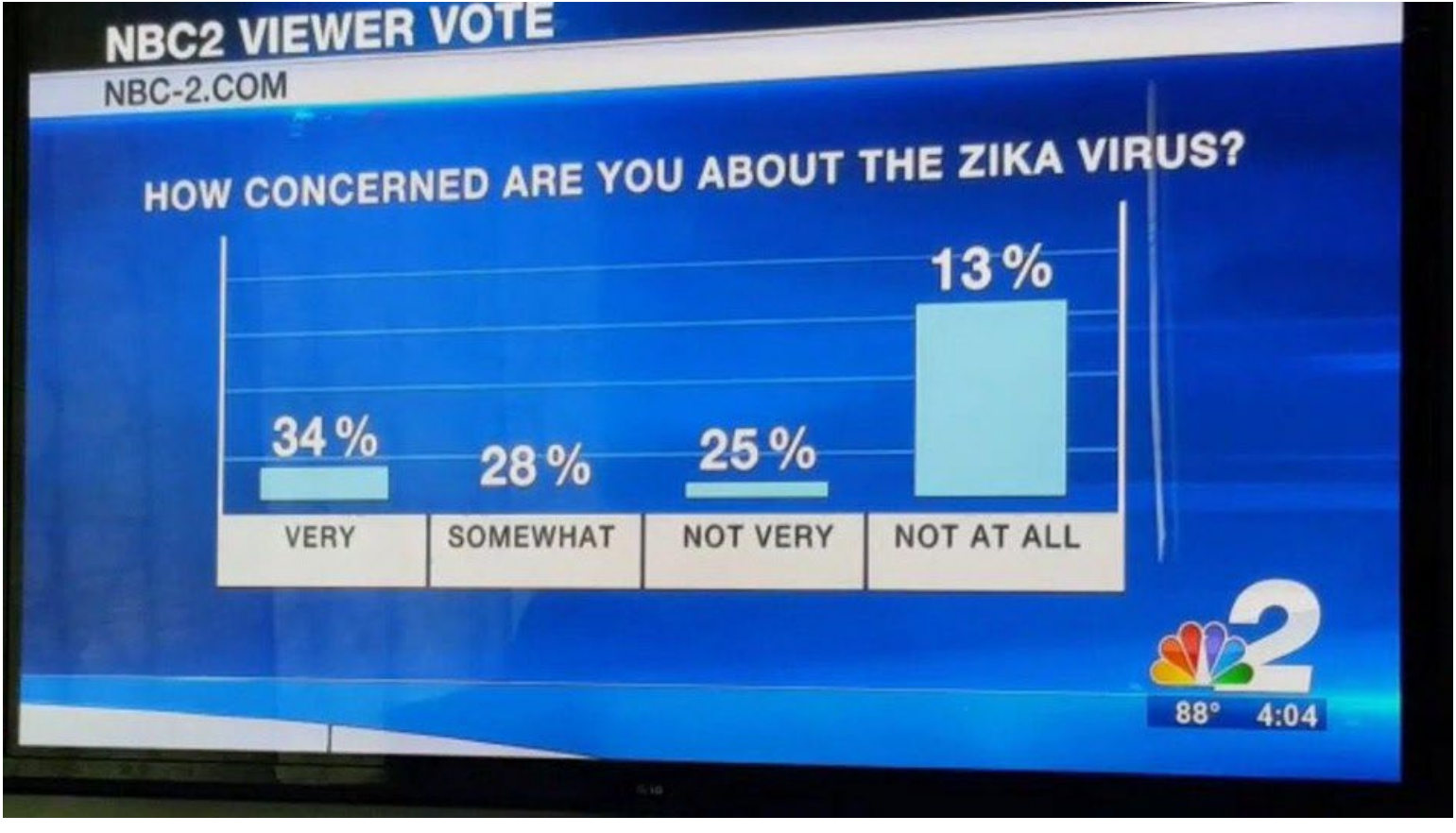
The areas of the histogram must be kept proportional to the numbers reported.



Los Angeles Times, August 5, 1979, p. 3.

Misrepresenting Data with Histograms

Of course you can also just screw up completely



Time Series Plots

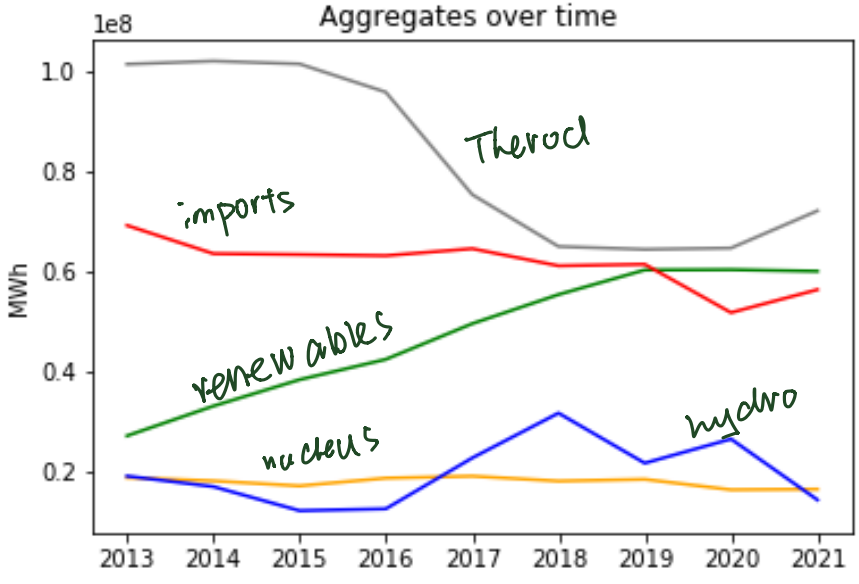
When data is ordered by time, it is often much more insightful to graph it against time.

For example consider how we looked at solar over time – we could see it increasing using the boxplots. We could do this with all the different energy sources at once, but this would be difficult to interpret.

Instead we can use a time series plot. We lose information here as we will see, but gain some perspective because we can look at how everything varies with time.

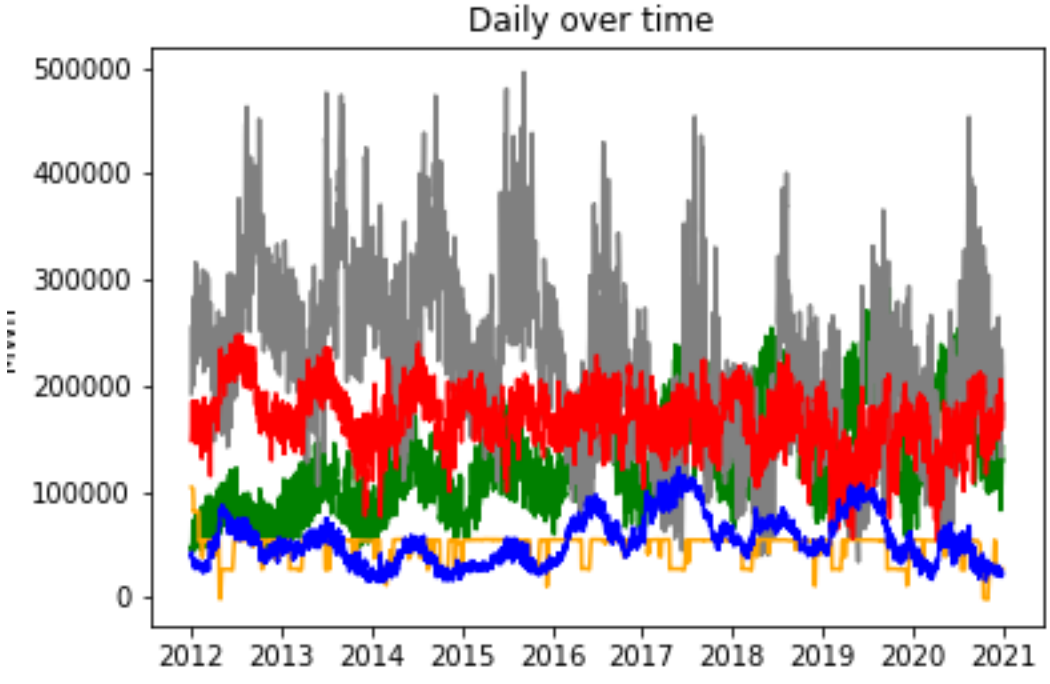
Time Series Plots

This is annual aggregates plotted with time on the x axis.



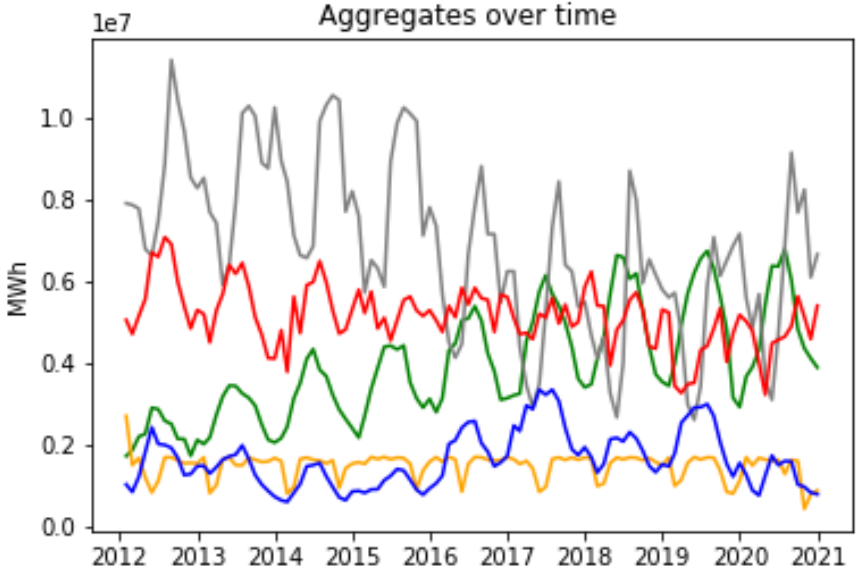
Time Series Plots

This is daily aggregates plotted with time on the x axis.



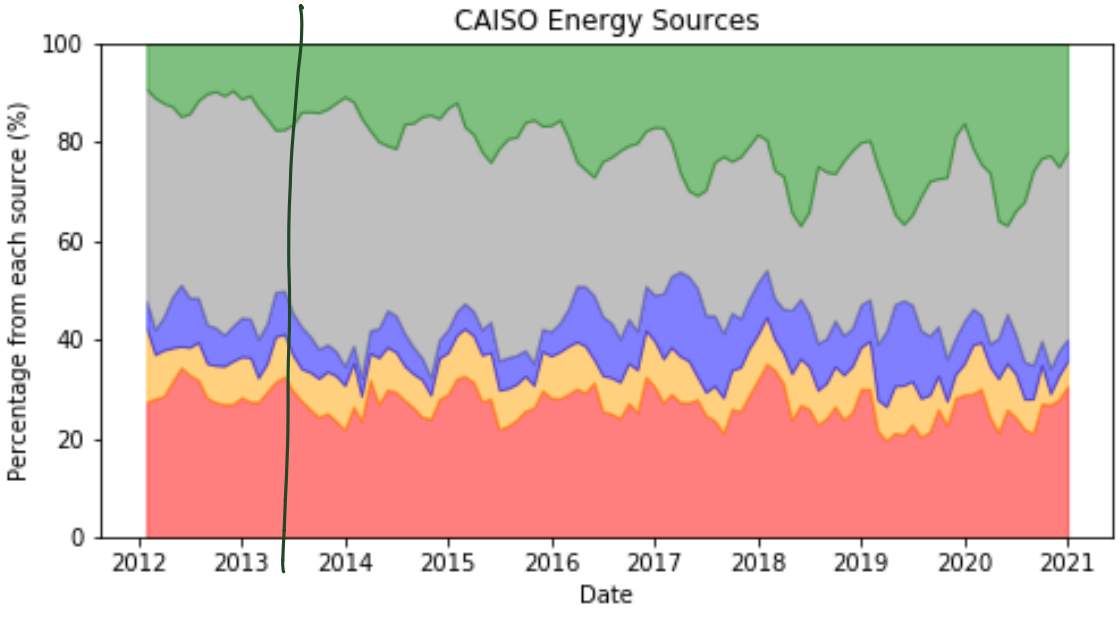
Time Series Plots

This is monthly aggregates plotted with time on the x axis.



Time Series Plots

This is monthly percentages plotted with time on the x axis, made pretty.

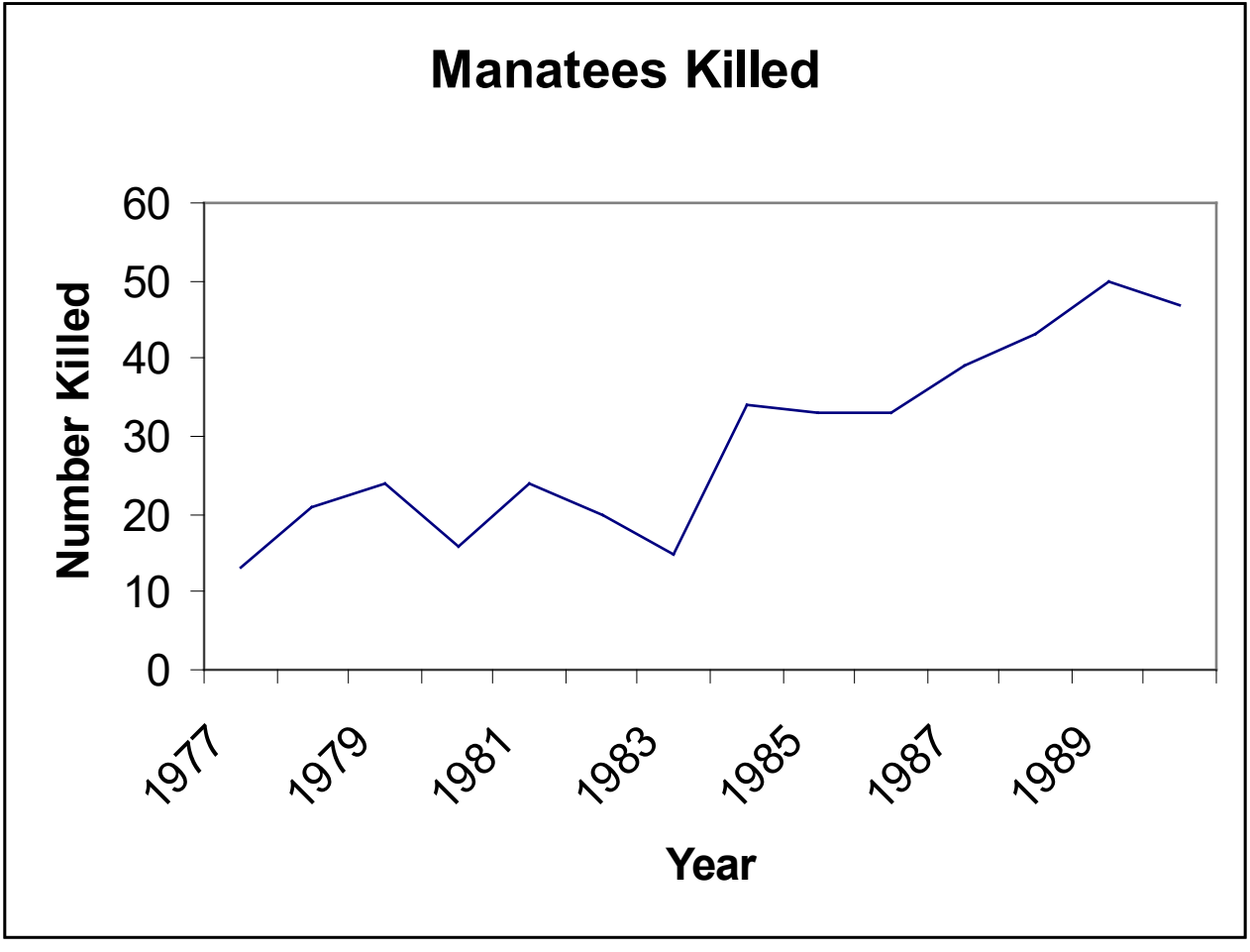


Time Series Plots

Time series plots are straightforward, however it can still lead to misleading results if you play around with the scaling.

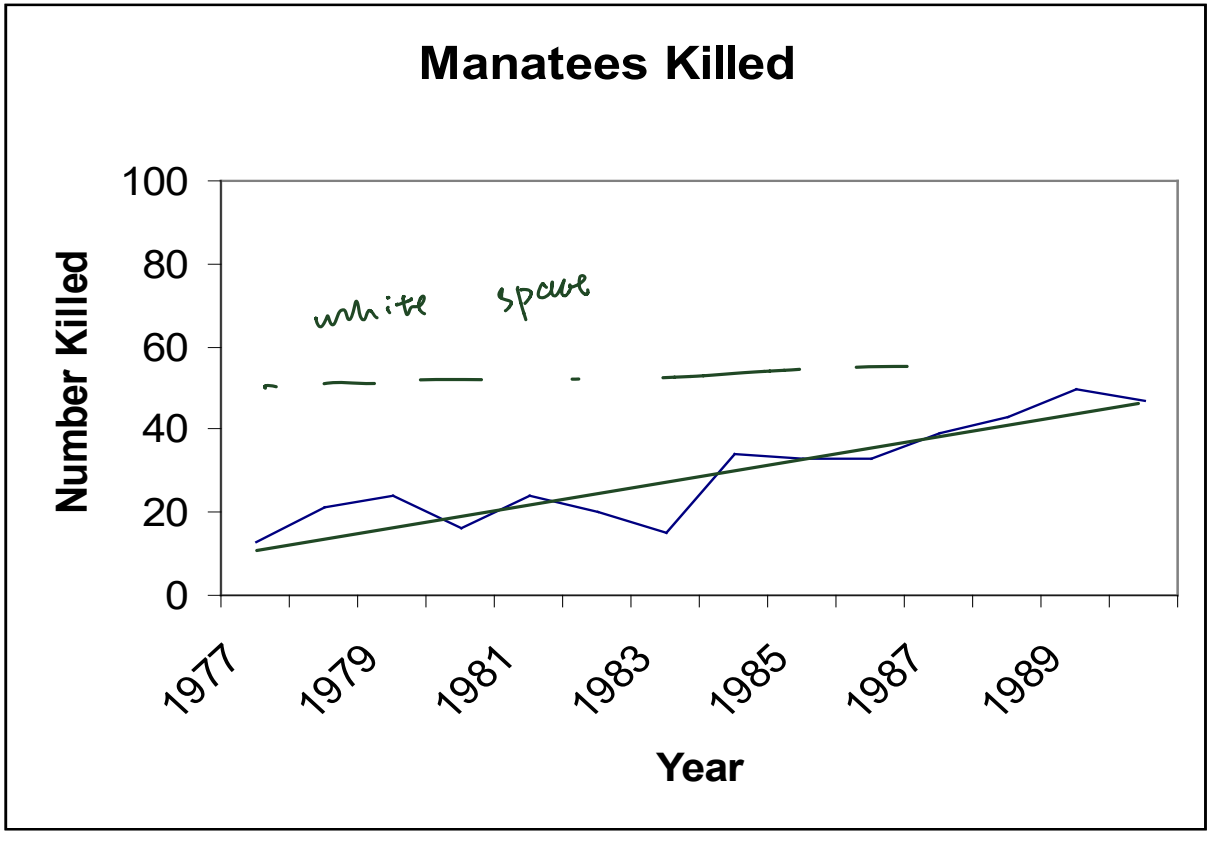
e.g. Manatee deaths in Florida. I have data for each year on the number of Manatees killed.

Time Series Plots



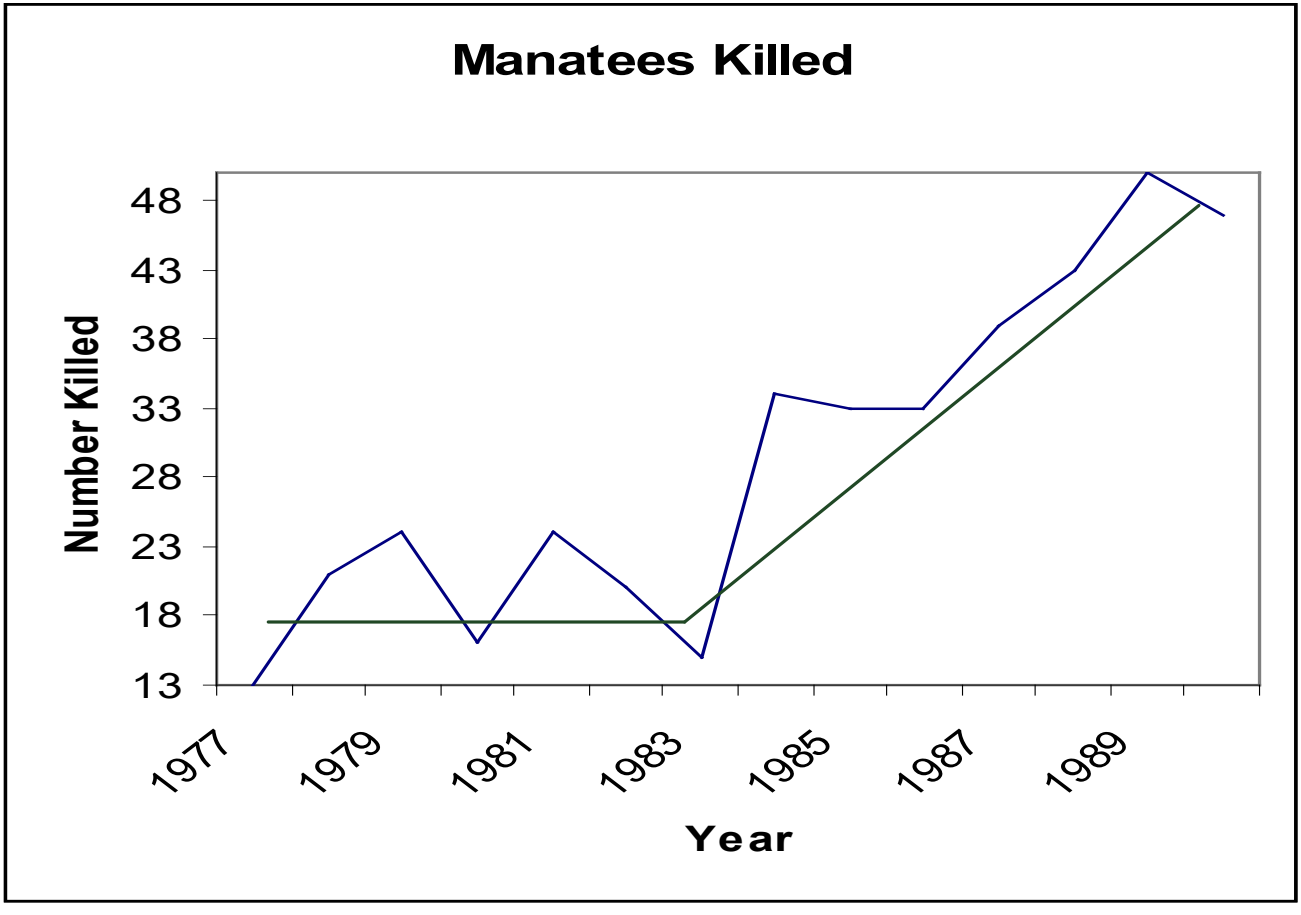
Time Series Plots

Power Boat Enthusiast Graph

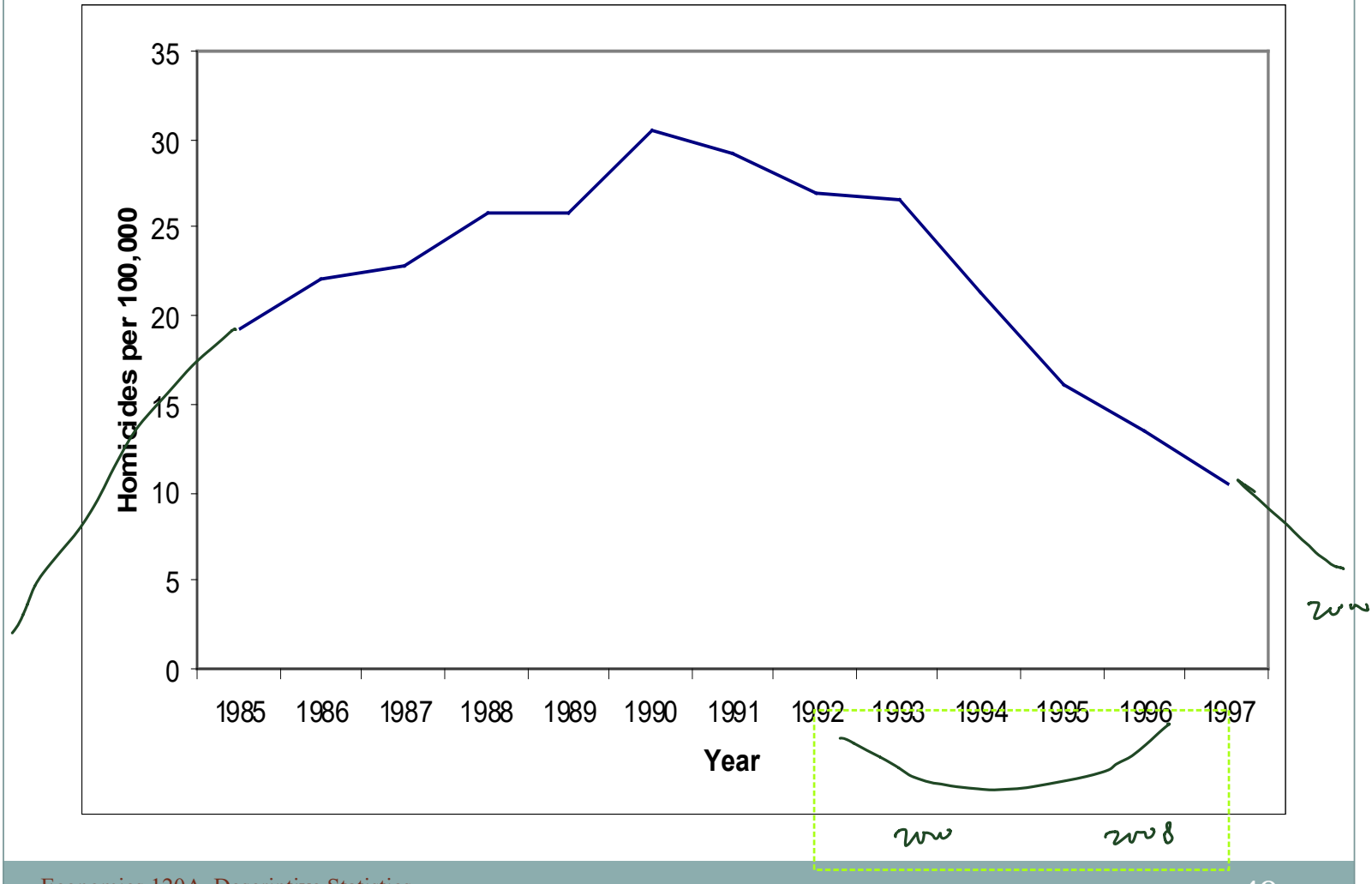


Time Series Plots

Environmentalist Graph



New Policing and Crime



New Policing and Crime

Why the drop?

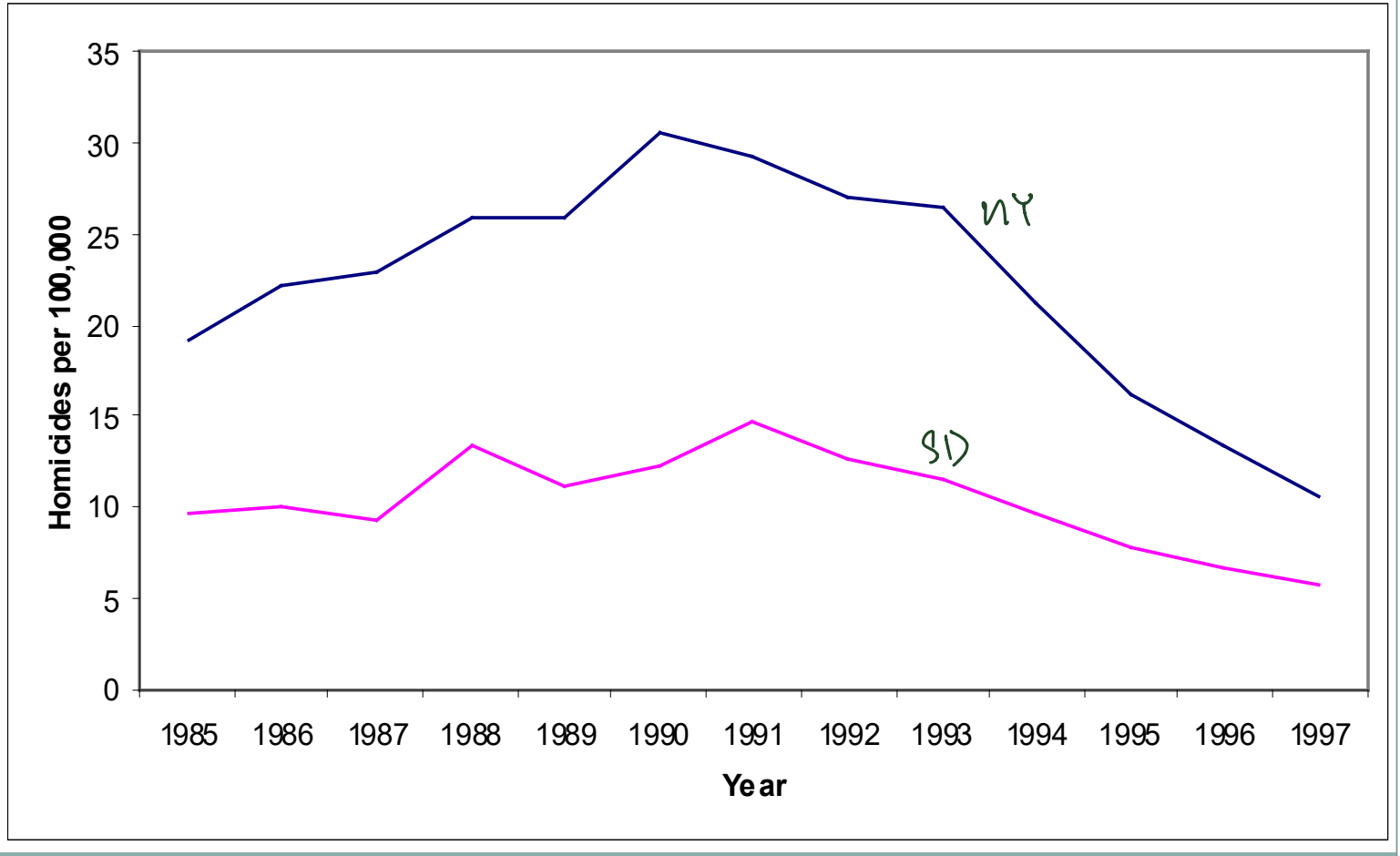
Giuliani argued it is their (with his chief of police Bratton) community policing

- aggressively enforcing misdemeanor crimes ('broken windows')
- better technology to follow up on problems and target problem areas.

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

New Policing and Crime

Far reaching effects?



New Policing and Crime

Really far reaching effects?

Data with Multiple Attributes

Renewables vs Thermal Generation

Are renewables in California really replacing generation from burning fossil fuels?

For questions like this, there are a number of issues.

First – does it seem in the data that higher levels of solar are related to lower levels of thermal generation? This question is one of correlation.

Second – Is it the case that higher levels of solar are resulting in lower levels of thermal generation? This question is one of causation.

To answer either of these questions, we need data with multiple attributes – in this case we need to measure both solar and thermal generation at the same time.

Data with Multiple Attributes

More Examples

1. The mutual fund data had a single attribute per observation, that of the return. We could similarly have collected not only the return for each mutual fund, but also the number of different stocks that each mutual fund invested in.

2. Arguments are often made that violence on TV or in video games is contributing to increased violence in society. Is it in the data?

3. You are here getting a college degree, presumably in part at least to improve your future financial situation. Do more educated people earn more on average?

4. We saw the data on annual manatee deaths - is it likely that it is rising due to greater numbers of boaters on the swamps and rivers?

Data with Multiple Attributes

In most interesting data analyses, each datapoint has multiple attributes. For example;

1. We observe for each mutual fund both returns and the number of different stocks held.

2. For each person we might observe both the amount of violent TV seen as well as their history of being violent.

3. For each person we could observe both the number of years of education and their subsequent income (or alternatively whether or not they have a college degree and their income).

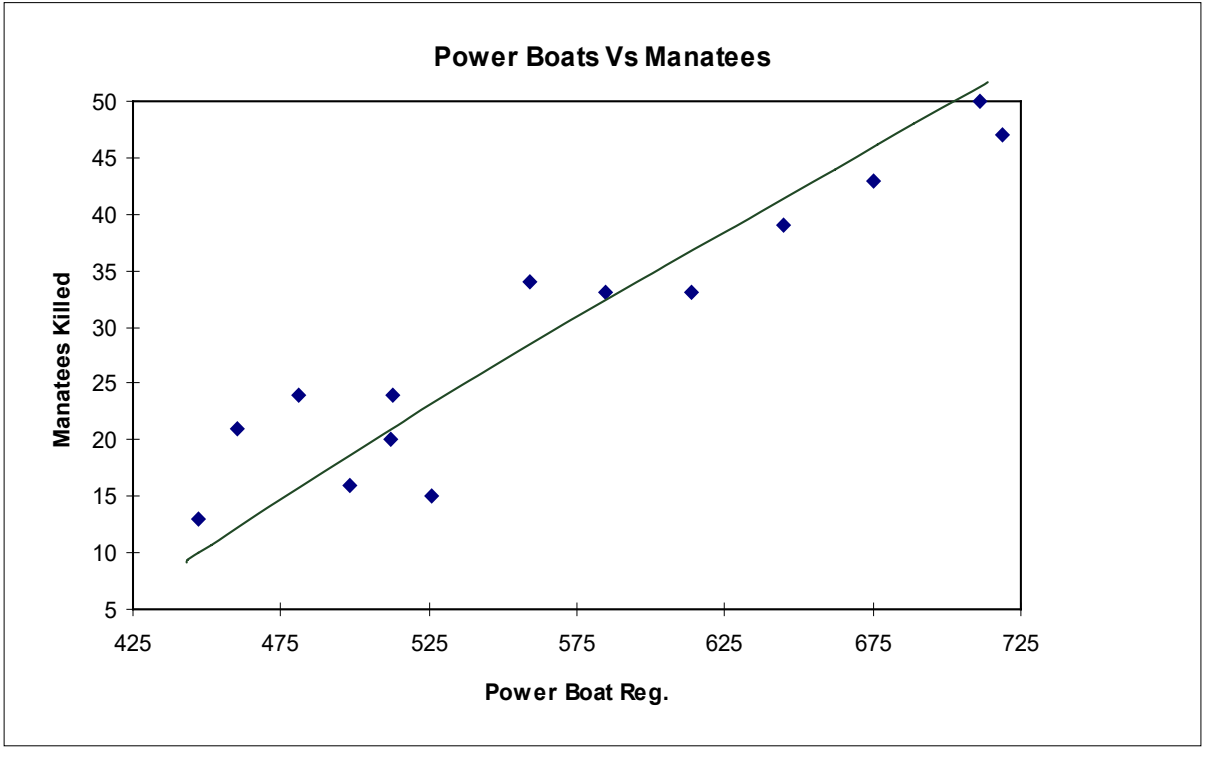
4. We could observe not just the numbers of manatees that died but also the number of boat registrations.

Manatee Example

Year	Power Boat Registrations	Manatees killed
1977	447	13
1978	460	21
1979	481	24
1980	498	16
1981	513	24
1982	512	20
1983	526	15
1984	559	34
1985	585	33
1986	614	33
1987	645	39
1988	675	43
1989	711	50
1990	719	47

Scatterplots

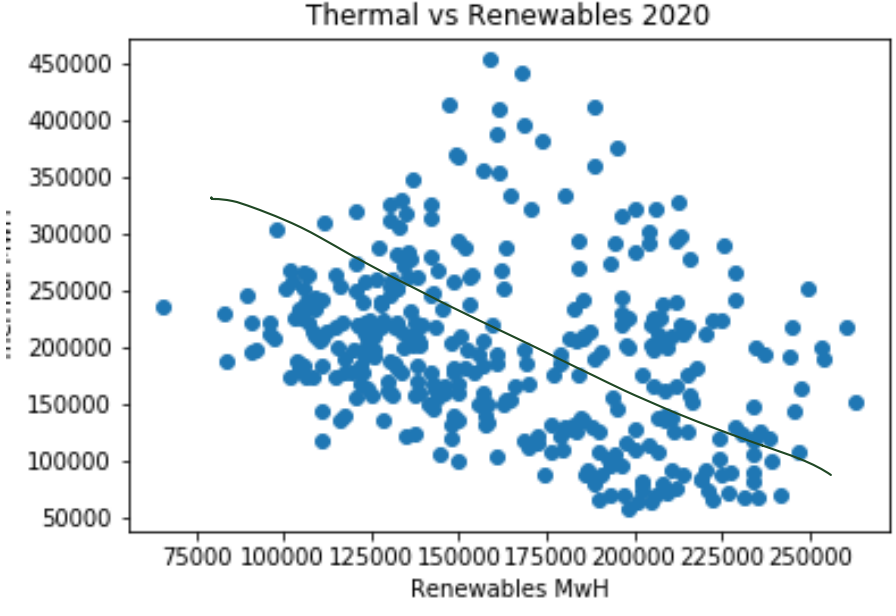
If both of the variables can take on many different values, the typical approach is known as a 'scatter plot', i.e. In this method we use a Cartesian diagram and let one of the variables be measured on the x axis and the other on the y axis



Scatterplots

For renewables and thermal, we can look at a scatterplot of the daily pairs of observations.

What do we see?



Scatterplots

In looking at scatterplots there are three things that we can determine in most cases,

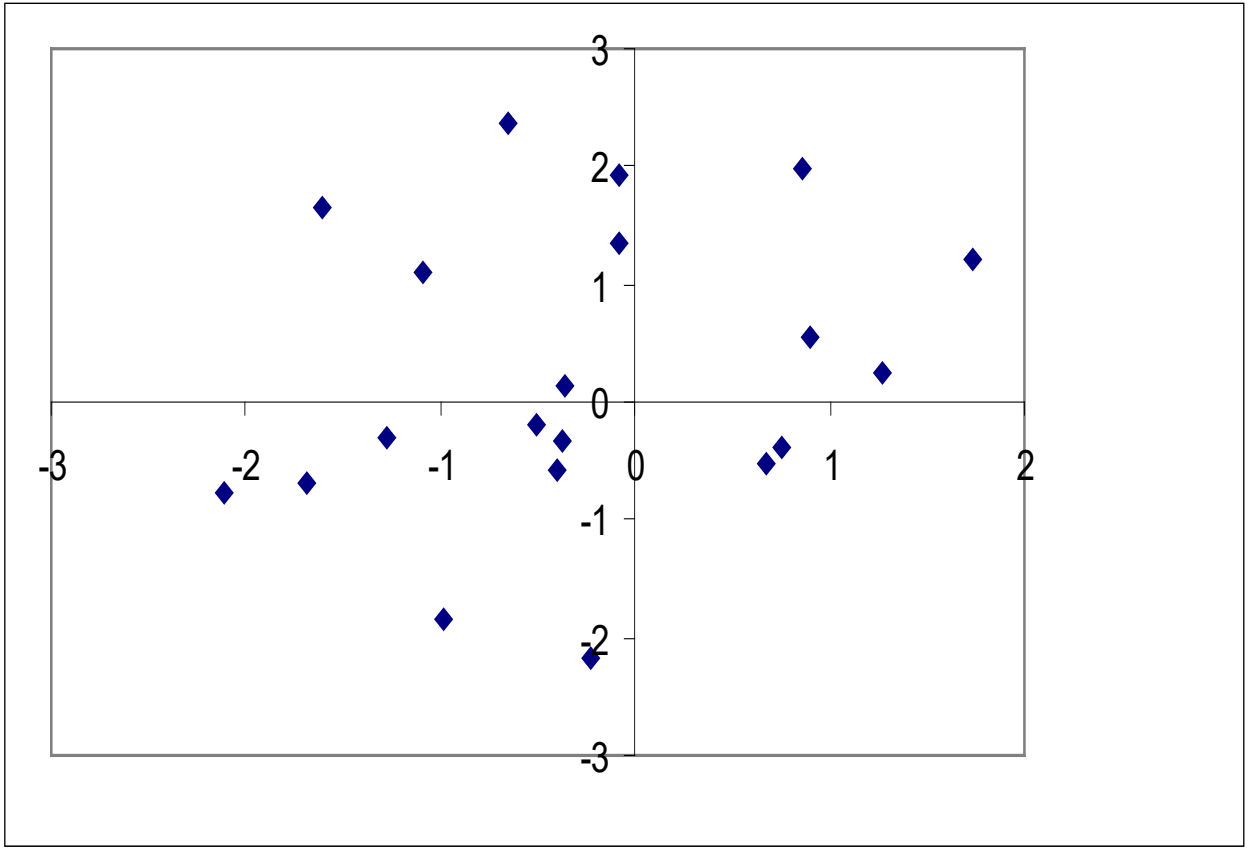
1. Is the relationship positive or negative? For this problem it is quite clear that there is a positive relationship between the number of manatees killed and the number of power boats out there. This is indicated by the low number of deaths when there were few boats, and more deaths when there are many.

2. What is the form of the relationship? i.e. is it linear, exponential?, rising then falling etc. Here the form appears linear (be sure to recall that there is randomness).

3. What is the strength of the relationship? Is the relationship really clear? Or are the observations all over the place. Here the relationship looks fairly clear, the observations lie within a tight band.

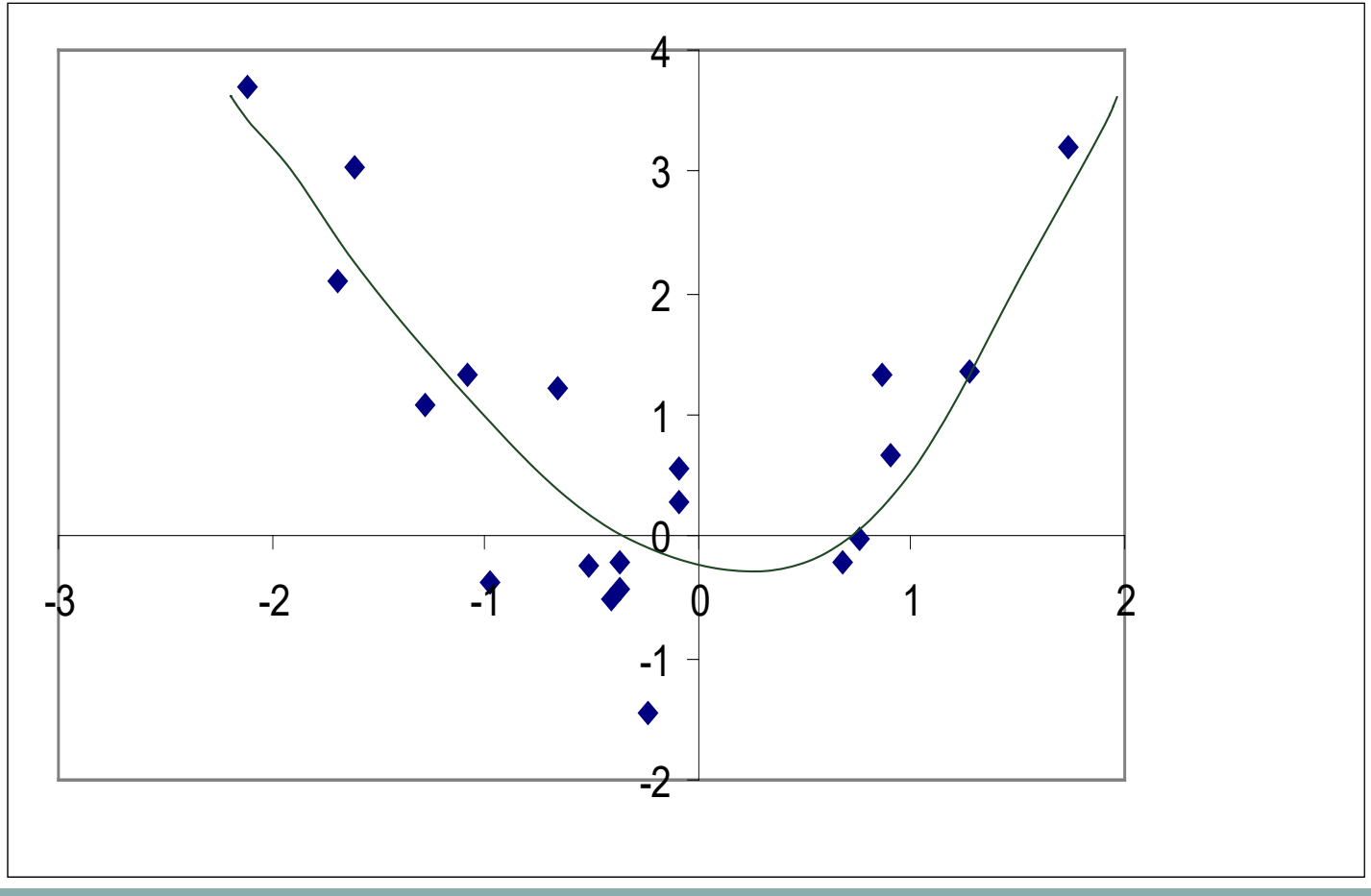
Scatterplots

There may be little or no relationship that is easy to see.



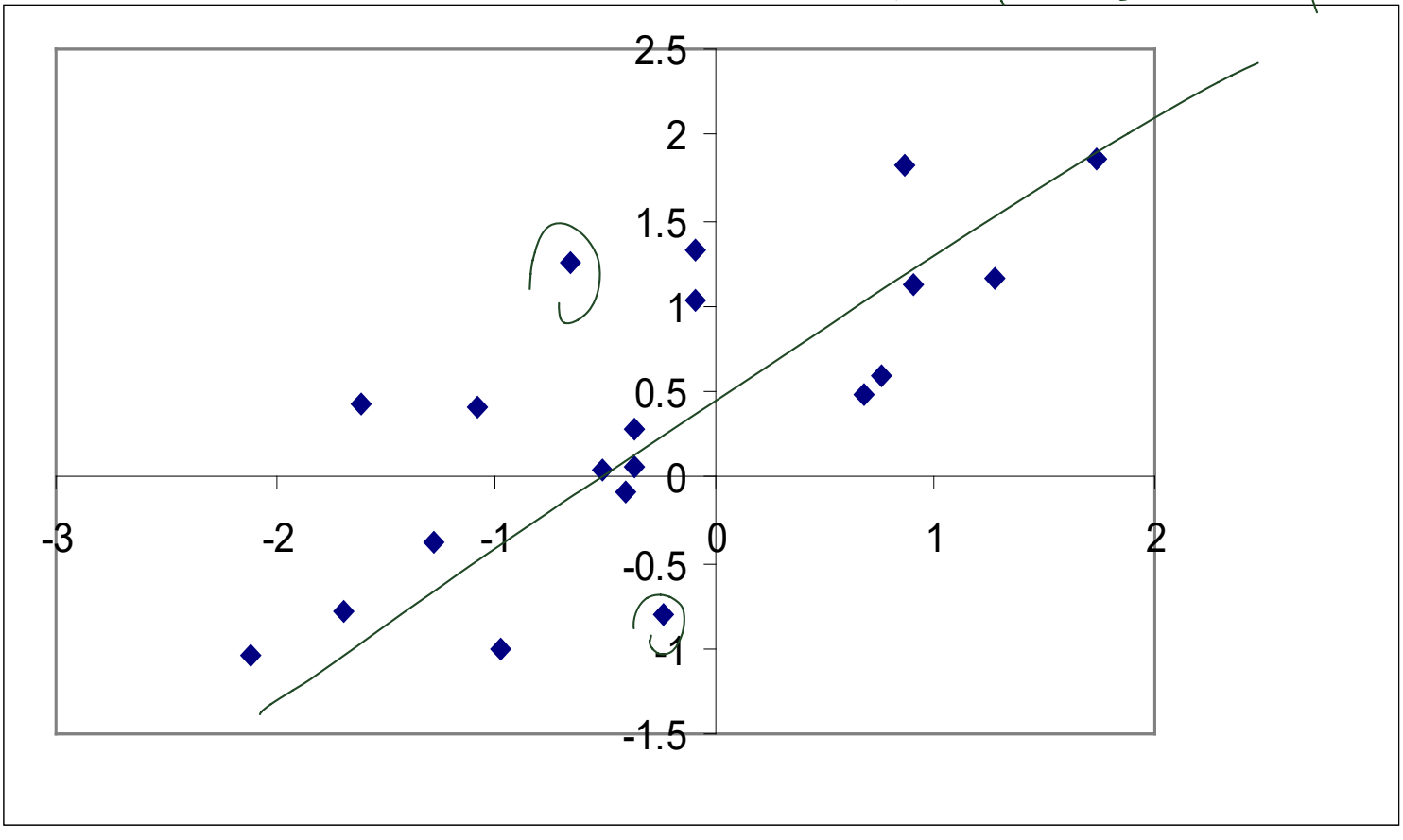
Scatterplots

It may not be linear.



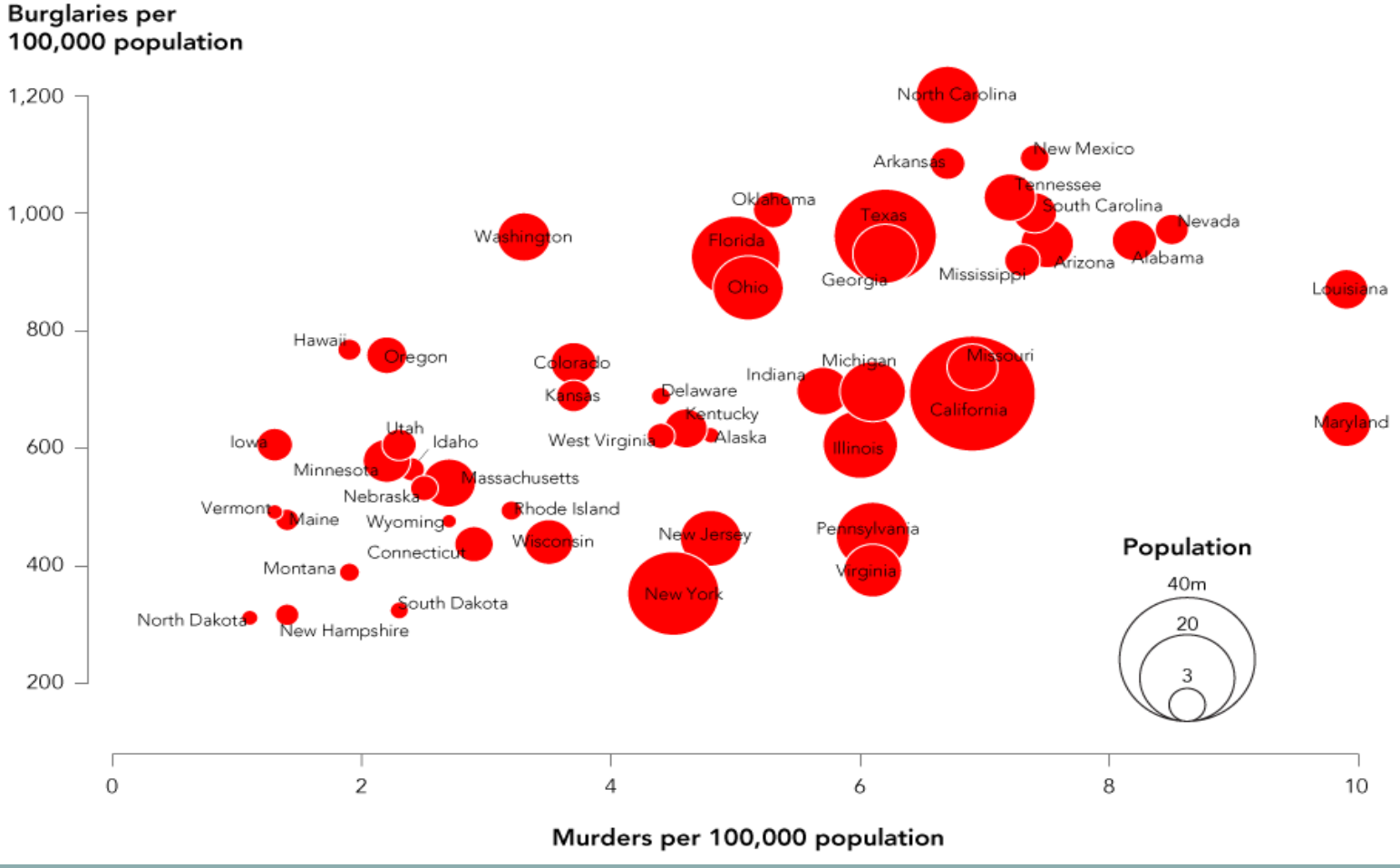
Scatterplots

It may be linear looking but not very clear.



Bubble Plots

Useful way of incorporating a third variable into a two dimensional



Bubble Plots

This is just a scatterplot with the additional variable (in the crime example this is the state population) for the bubble.

The rules for the bubble are the same as for the histogram --- to obtain an accurate idea of relative sizes the areas of the bubble should match up with the value being graphed, so in the example a state with double the population of another state should have a bubble with twice the area (and not twice the diameter) of the first state. The graph on the previous page is accurate.

Excel does this as a default.

Usefulness of Graphical Methods

For some problems, graphical methods can illuminate

- which theory might be correct
- what types of things need explaining
- what types of explanations might be overreaching.

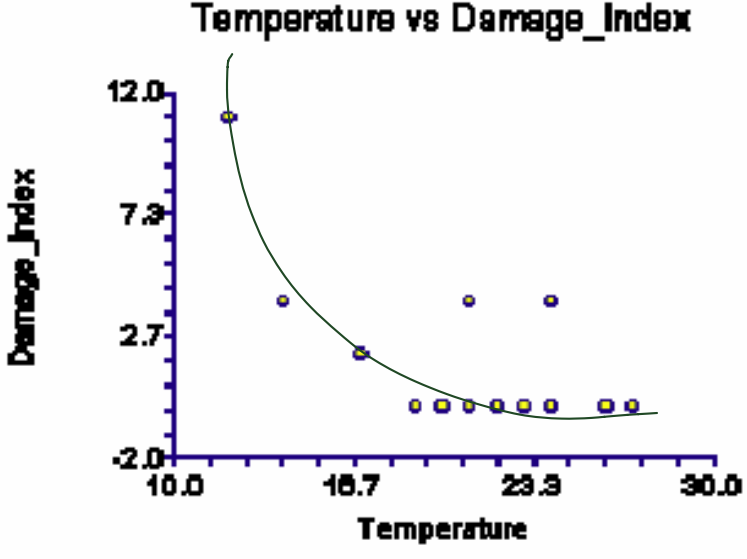
For example, 1986 Challenger Shuttle Disaster

The shuttle exploded during ascent due to the failure of an o-ring seal on the right rocket booster.

Problem: failure due to the cold (it was around freezing at the time of liftoff).

Usefulness of Graphical Methods

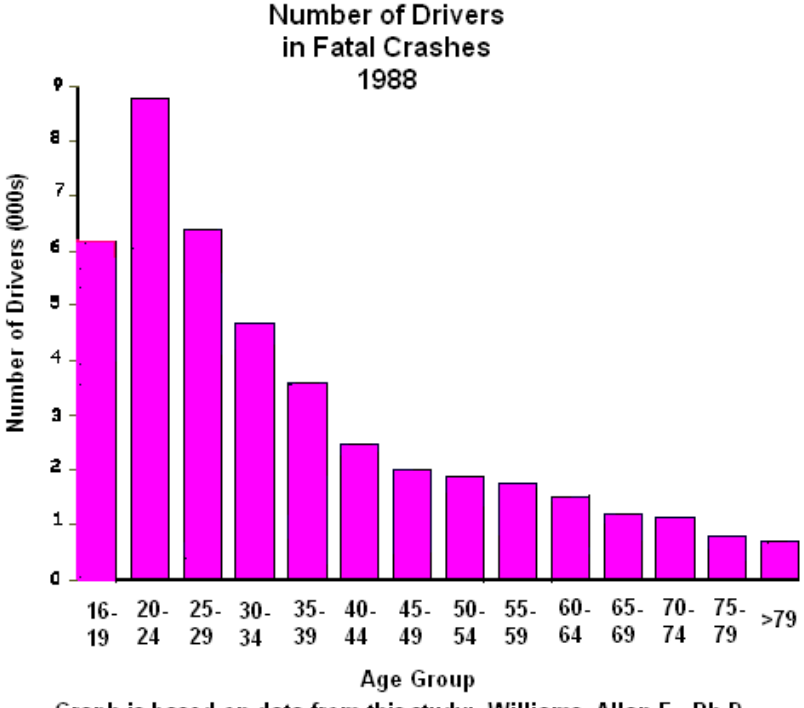
A simple scatterplot showing the link between O-ring damage and ambient temperature during previous launches may have changed the decision about launching. How much damage would you have expected at 0^{0} Celsius?



adapted from: Tufte, E.R., *Visual Explanations*

Usefulness of Graphical Methods

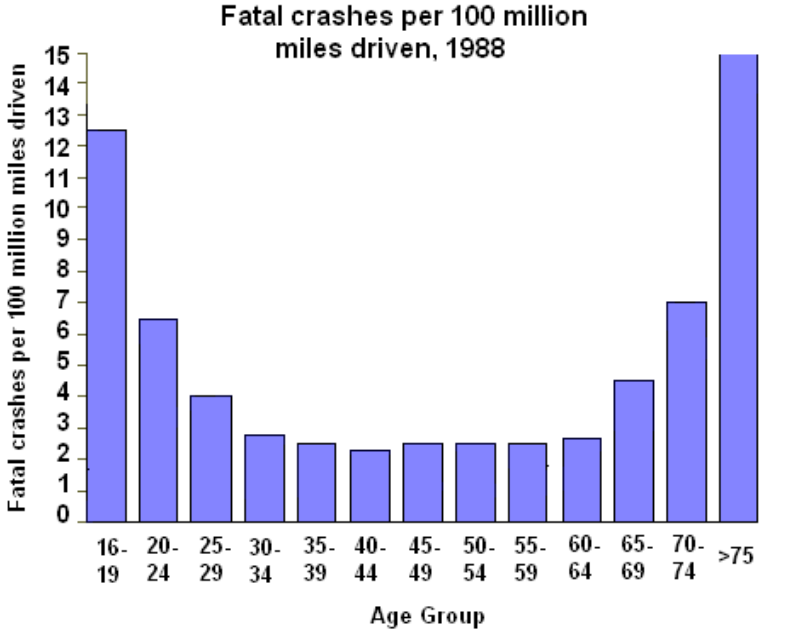
We still need to think carefully about what we are doing



Graph is based on data from this study: Williams, Allan F., Ph.D., and Oliver Carston, Ph.D., "Driver Age and Crash Involvement," Am J Public Health 1989; 79: 326-327.

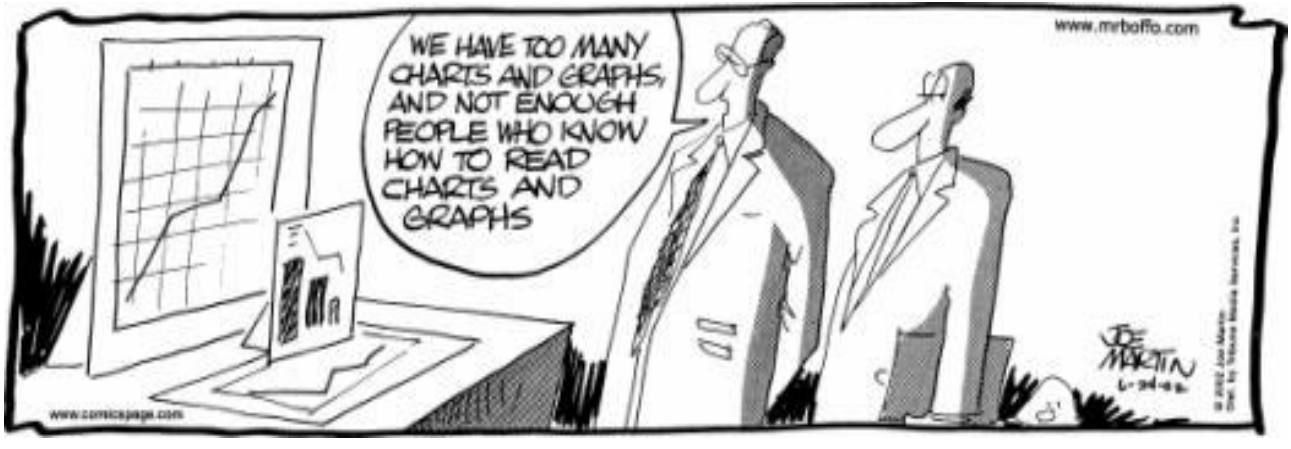
Usefulness of Graphical Methods

Need to really compare 'apples to apples' and divide by miles driven



Graph is based on data from this study: Williams, Allan F., Ph.D., and Oliver Carston, Ph.D., "Driver Age and Crash Involvement," Am J Public Health 1989; 79: 326-327.

Analytical Skills



The Basic Idea

Rather than draw pictures, we often want to report numbers that summarize the information in the data without presenting all of it.

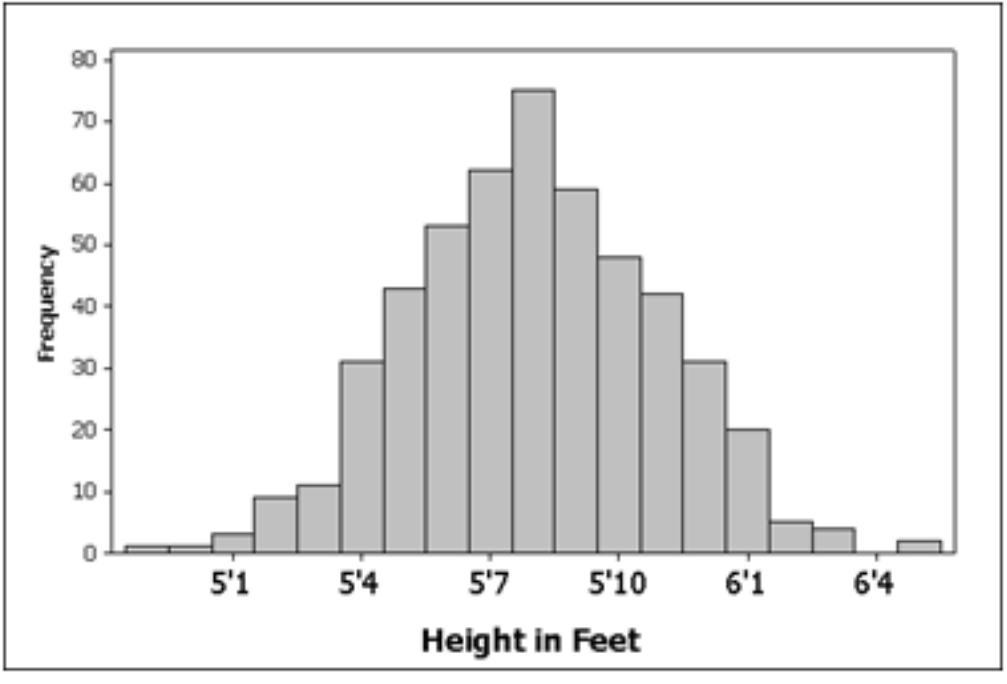
As with graphical approaches, we want to retain the salient information for our problem but at the same time ignore the less useful information and detail.

Our approach is basically to characterize features of the data.

Suppose that you needed to summarize the height of college students. The usual way to do this might be to answer `students have an average height of ...' This is an attempt to render the great diversity of heights into a simple number which tries to give the idea of the `usual' student. In most cases the first number you might report is like this, some type of measure of central tendency of the distribution of heights.

A second measure might be to talk about how spread out they are, etc.

The Basic Idea



Measures of Central Tendancy

There are three common measures of central tendency of a set of data. These are

1. Mean
2. Median
3. Mode

We will go through each, but ultimately focus on the mean.

Some notation.
We will denote observed values as $x_1, x_2, x_3, \ldots, x_n$ for n observations.

Notice that the book uses capital letters, I use small letters. The distinction is important because later we will use capital letters for a very related concept, and it is extremely important to keep the capital letters concept and the data separate.

The Sample Mean

We have data $x_1, x_2, x_3, \ldots, x_n$.

The sample mean is simply the sum of the values divided by the number of values.

So mathematically we have

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$ sample mean

This is often referred to as the average or sample average.

We are using the modifier 'sample' here because later we will have a concept for the mean that is not from data, and we need to be clear unless it is clear from context.

The Sample Mean

Insurance Example: When you write off your car, and the insurance adjuster decides how much to pay, they look at similar cars for sale, adjust for features, and typically use this as a basis for the `worth' of your smashed up car. Suppose the adjuster obtains the following numbers:

x_1	2500
x_2	2100
x_3	2300
x_4	1000
x_5	2000
	9900

$\frac{9900}{5} = 1980$

The answer is in \$ (if we add \$ to \$ we get \$). You always need to know the units of measurement to give a meaningful answer.

The Sample Mean

Suppose we were to think of each observation as being a weight (say lb) and that we spread them out on a beam according to their actual value (on the x axis = beam).

i.e. Suppose that we have just one observation, the beam balances at that observation = mean.

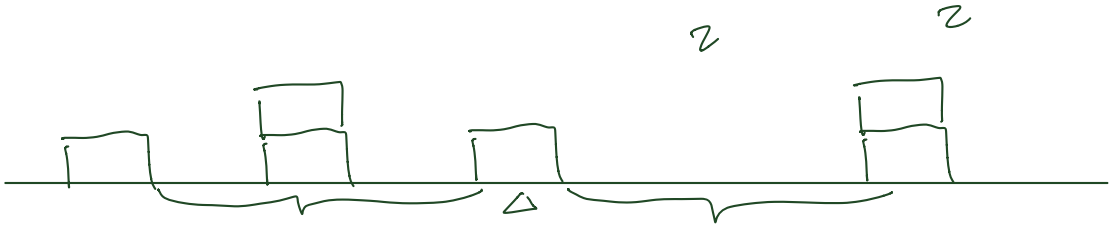
fulcrum is where the mean is

The Sample Mean

Now suppose we have two more observations, one below the mean and another equidistant from the mean but above the mean.

The Sample Mean

In each case, the mean does not change. But what we see clearly is that each observation is able to affect the mean.



$1 \cdot 2 + 2 \cdot 1$

The Sample Mean for Binary Data

It is often useful to code non-numerical data using numbers.

For a Yes/No or Success/Fail type problem where there are only two outcomes, we can code responses as

$$x_i = \begin{cases} 0 & No/Fail \\ 1 & Yes/Success \end{cases}$$

Some examples:

1. Political poll. "Did you vote for the incumbent?" has a yes/no answer.
2. Employed/unemployed. In a labor study you might want to measure such occurences. Obviously this is of interest for research into the effects of unemployment.
3. Male/Female. Often in studies such attributes are employed as 'causes' and examined.

The Sample Mean for Binary Data

Notice that the sample mean is constructed via the same formula as above, i.e.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

but this sum is just the number of yes/successes in the sample, since otherwise we get zeros in the sum, hence

$$\bar{x} = \frac{number\ of\ successes}{total\ number\ of\ observations}$$

so is equal to the proportion of successes in the sample.

Since it is just a special case of a sample mean, we will examine it in the same way as we shall see later in the course. Often this special case is called a 'sample proportion' .

The Sample Median

The second most used concept of the `center' of a distribution is the median. This is the very middle observation in the data.

To compute this, simply rank all of the observations from the smallest to the largest, and report the value of the middle observation.

e.g. Car insurance example

x_1	2500
x_2	2100
x_3	2300
x_4	1000
x_5	2000

Sort

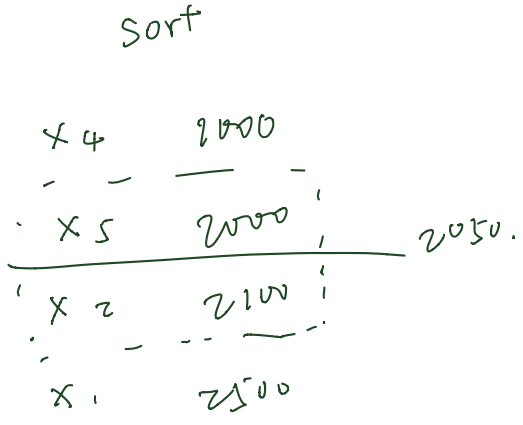
x_4	1000
x_5	2000
x_2	2100
x_3	2300
x_1	2500

The Sample Median

The above strategy works for an odd number of observations. If I remove one of the observations (x_3) from our example we see what to do for an even number of observations.

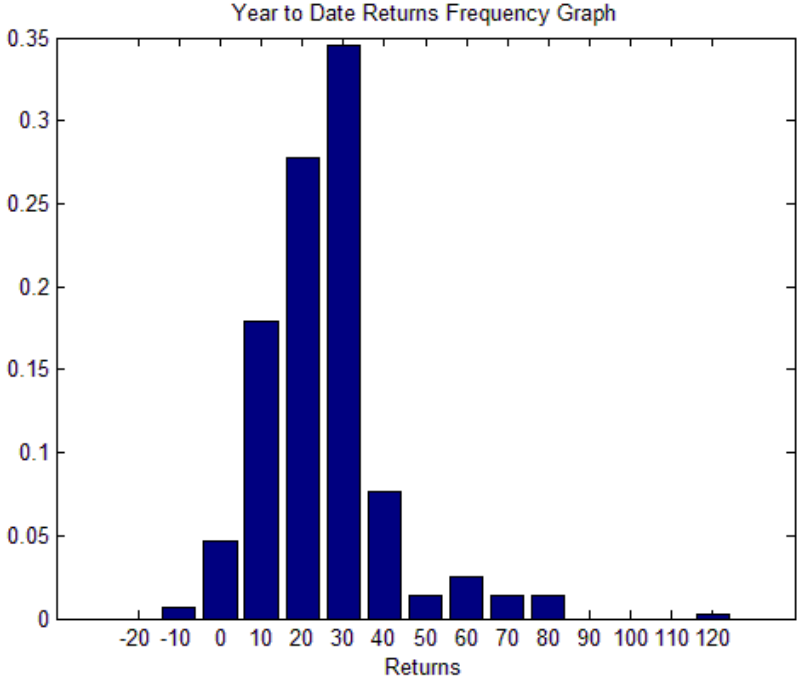
e.g. Car insurance example

x_1	2500
x_2	2100
x_4	1000
x_5	2000



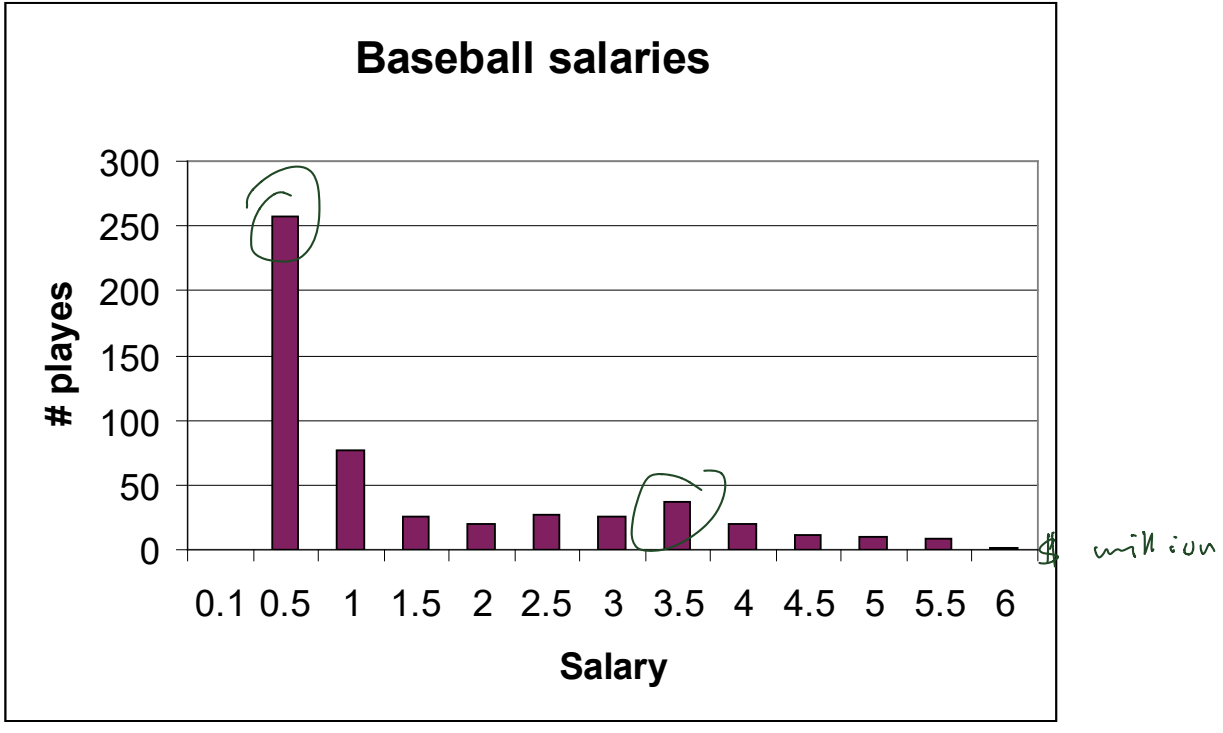
The Sample Mode

The mode of a distribution of data (or set of data) is the most frequently observed value. This is easiest to see from a frequency distribution.



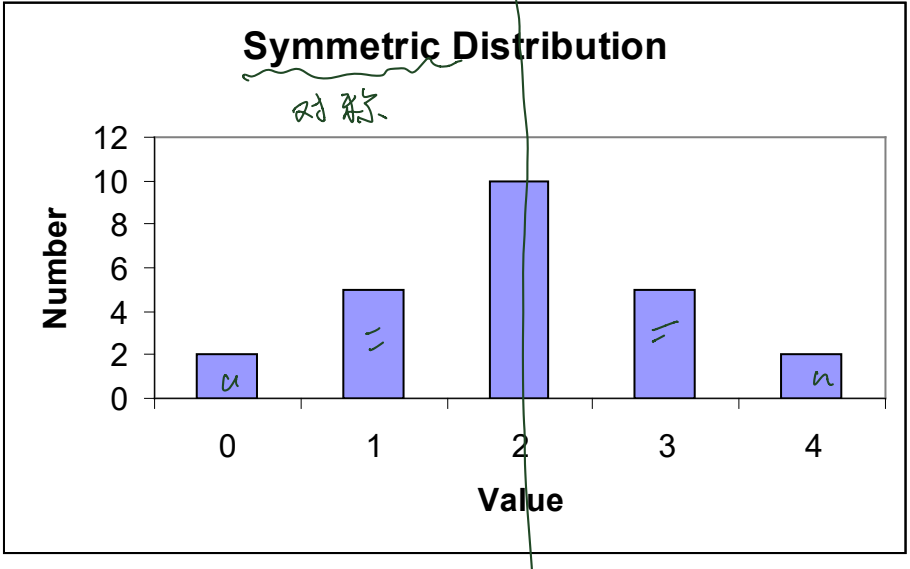
The Sample Mode

Unlike the mean or median, it is possible to have more than one mode. The following gives 1994 baseball salaries (source: USA today) where salaries include pro-rated signing bonuses. The salaries are given in millions.



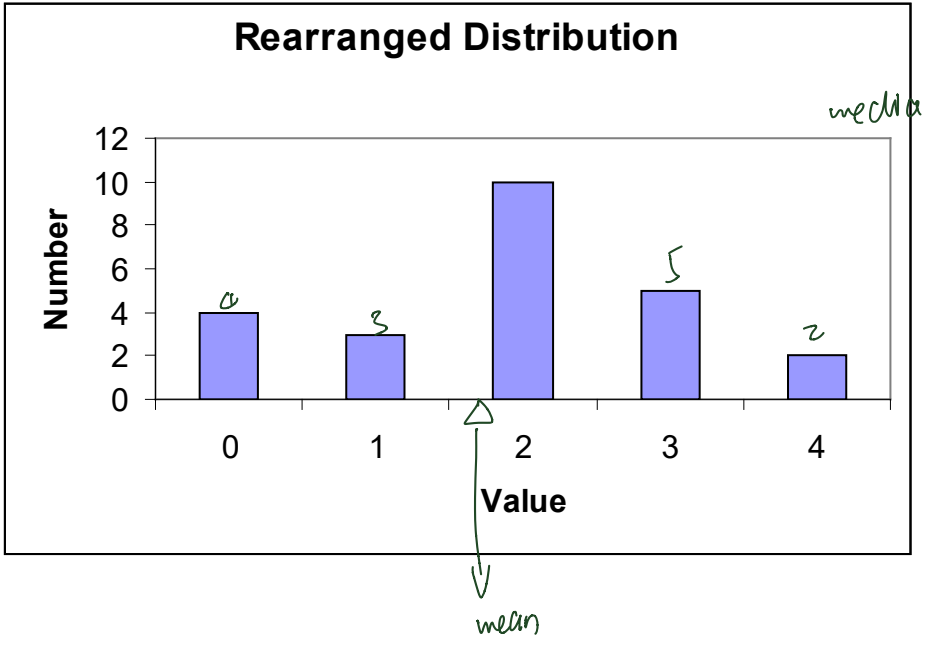
The relationship between the measures

Consider the following distribution



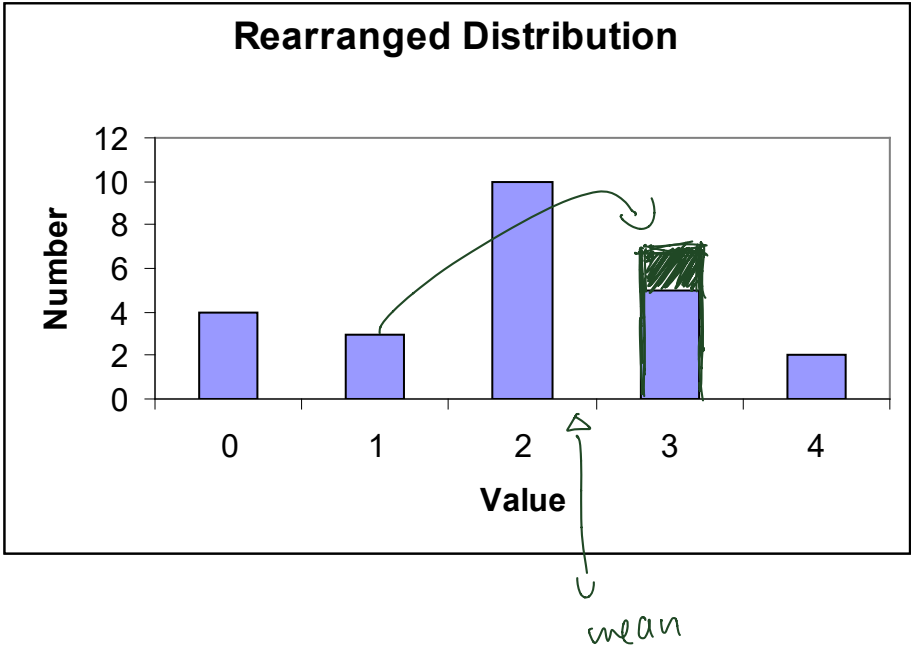
The relationship between the measures

Suppose now that we take from the highest two columns and put these observations at the lowest two columns. What is affected now?



The relationship between the measures

Starting from the original symmetric histogram, suppose that we take some of the observations from the second column and place in the fourth, what happens to the different measures?



median does not change bc #s of observations at x = 2

The relationship between the measures

What we see here is that the mean is sensitive to small changes. This sensitivity is very apparent when the observations are `outliers', or observations far from the rest of the observations.

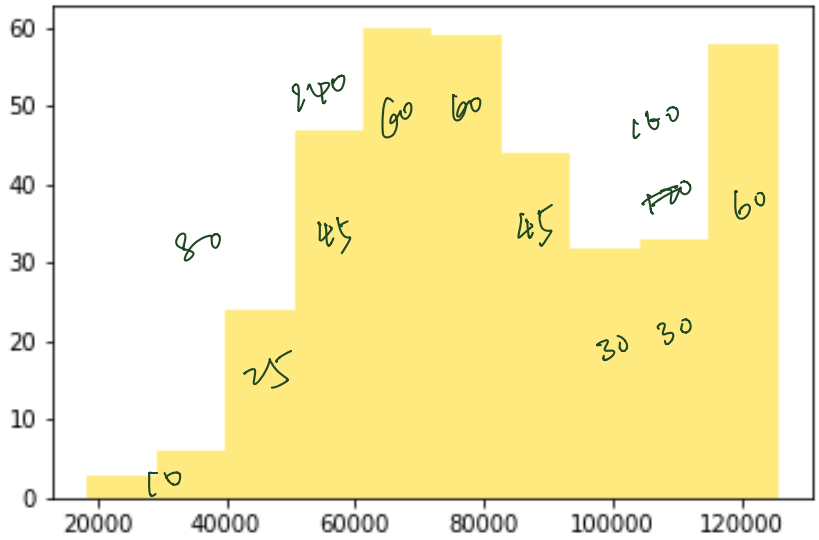
Consider the car insurance example. The value of 1000 is clearly far from the other observations, it is an outlier. The median here is $1980, much lower than the median of $2100.

Is the insurance adjuster going to use the median of the mean to pay you back? Notice that $1980 is not in any way a typical observation, it is lower than four of the five observations.

The Sample Mean - Energy Example

This is solar for 2020 in CA.

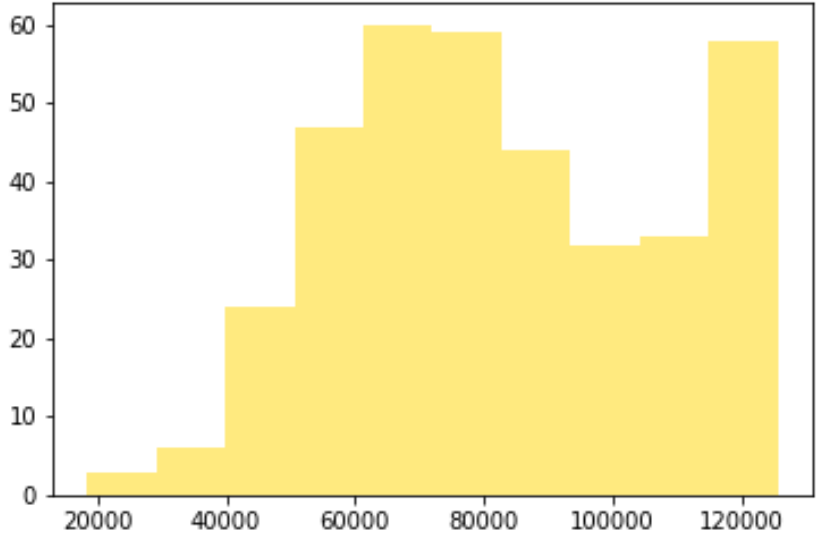
Where are the mean and median?



The Sample Mean - Energy Example

This is solar for 2020 in CA.

Where are the mean and median?



Mean = 82256
Median = 78810

The relationship between the measures

Which is the best measure?

This really depends on the problem. The mean is often the most easily understood, however is subject to the problem that we have seen in that it is the most sensitive to the actual data counted. i.e.

Suppose you just happen to mismeasure one observation, or perhaps there is a `weird' observation, then the mean can be very misleading.

e.g. - the insurance example.
e.g. - Life Expectancy

The relationship between the measures

Reported life expectancy has changed dramatically over the last few hundred years. In the late middle ages/renaissance times life expectancy was in the 30-40 years range, now it is closer to 70.

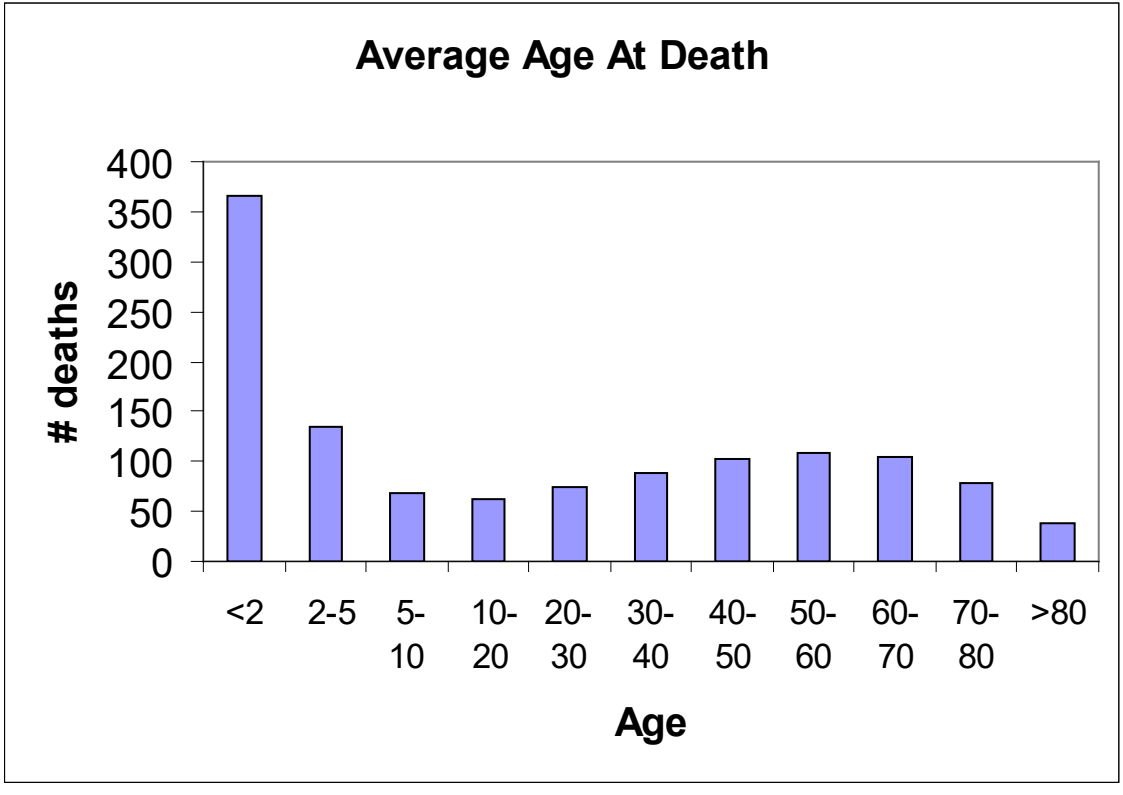
This statistic is usually reported as the mean lifetime. It is no surprise, given advances in medicine over the years.

I have data for 1740-1792 for Edinburgh, the data was extracted from gravestones in the city. The average life expectancy was just over 27 years. What impression does this give you?

To me it gives the idea that back in those days one should not have procrastinated too long on anything, otherwise you would never get anything done.

But is this right? Were there hardly any old people?

The relationship between the measures



Measures of Spread

The concept of the center of a distribution only takes us so far in understanding what a typical observation might look like.

Suppose that we are trying to decide between two investment advisors. Both have the same average return. The first advisor has returns that are always positive, the second has some returns that are higher than the other investor but also some negative, i.e. the distributions are as follows.

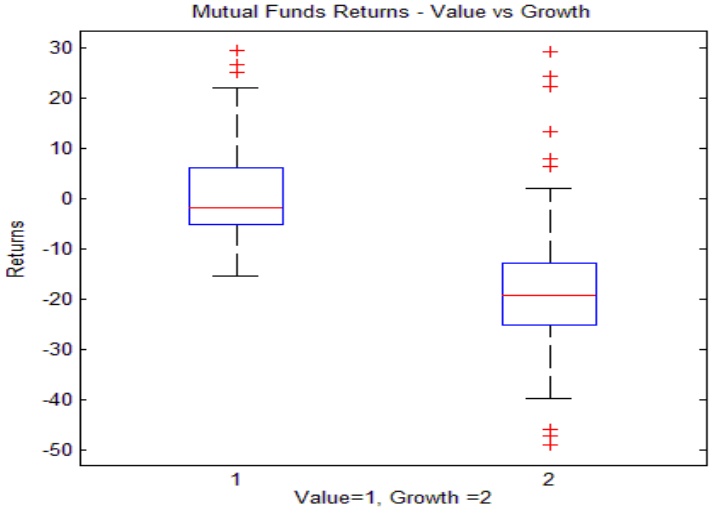
(put in two pictures here of the histogram of returns).

The mean is clearly not enough to describe either of these distributions, we need more. It is indeed helpful in understanding what a `typical' observation looks like to know how spread out the observations are. As before we have many different measures.

Range

This is just the minimum and maximum. These are the end points on the box plot.

The problem with this is that they are very sensitive to outliers, e.g. all the observations might be the same except for a large and small outlier. This really does not tell us a whole lot, but it is useful.



Interquartile Range

Quartiles are just each successive 25% of the data (bottom 25%, second 25% - which takes us to the median, third 25% and final 25% - which takes us to the high value).

The IQR is the endpoints of the box in a boxplot, i.e. the IQR contains the middle 50% of the observations.

This is related to the median conceptually, and has similar properties (the lower quartile is the median of the `bottom' half of the observations, the upper quartile is the median of the upper set of observations).

Variance/Standard Deviation

The variance looks at the spread using all of the observations and how far they are from the mean:

$$x_i - \bar{x}$$

It may seem sensible to take the average deviation, but this is always zero.

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) = \frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n} \bar{x}$$

$$= \bar{x} - \frac{n\bar{x}}{n}$$

$$= \bar{x} - \bar{x} = 0$$

$$\frac{x_i - \bar{x}}{n}$$

We do not want observations higher and lower than the mean to cancel, so we want to take a function of the difference, say the squared difference. This is what the variance does.

Variance/Standard Deviation

The formulas are

Variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

The standard deviation is the (positive) square root of the Variance.

A note on units. If x has some units (e.g. $), then the variance is $\$^2$, which are hard to interpret. But the standard deviation is back in $, the same units as x.

Notice that we divide by (n-1) rather than n, which seems a little unusual. The reason for this will be clear much later in the course.

Variance/Standard Deviation

Car Insurance Example

$\bar{x} = 1980$

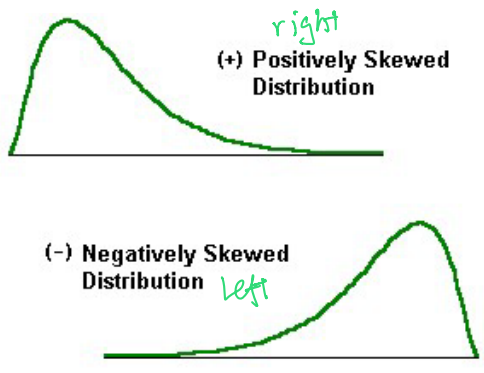
x_1	2500		270400
x_2	2100	$(2100-1980)^2$	$=14400$
x_3	2300		102400
x_4	1000		960400
x_5	2000	$(2000-1980)^2=$	400

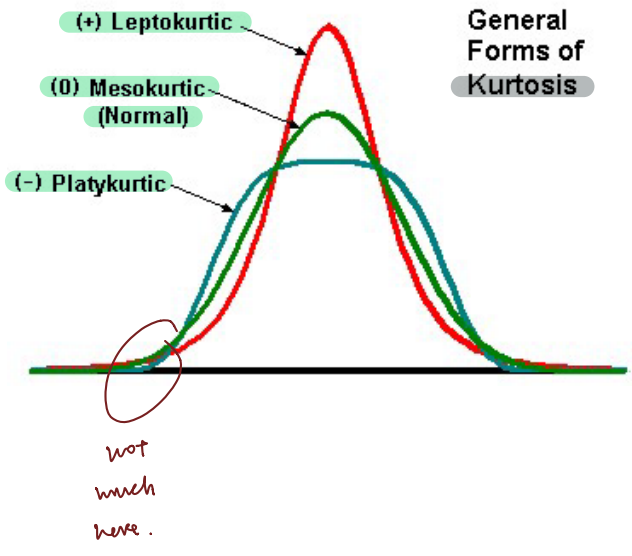
We have that s^2 =337000 and so s=581.

Measures of Skewness and Kurtosis

Skewness is whether or not the distribution leans to the left (negative or left skewness) or right.

Kurtosis is whether the distribution is has 'fat' tails.





Basic Issue

We saw in the graphical examination of data with multiple attributes that we could examine a scatterplot to see if there was some relationship between the variables.

We can also think of summary statistics to capture these (much of 120b and 120c does this).

Unfortunately, just as in the case of the center of a distribution, any choice of a summary statistic can lead to missing the real relationship, i.e. there is no ideal method.

Covariance/Correlation

The notions of correlation and covariance basically fall under the idea that there is some form to the pattern of the scatterplot. It might sweep upward from left to right, giving the suggestion that there is the likelihood that the larger one variable is the larger the other is likely to be.

Examples:

-We expect incomes to be positively correlated with education
-We saw a positive correlation between manatee deaths and boat registrations.

Our casual use of the word correlation is roughly correct, but we need to understand exactly how it is calculated and what the properties are.

Covariance/Correlation

Since we have two variables, we will denote the first as x, and the second as y.

Observations are indexed by i, and the data comes in pairs.

This is to say that for an examination of the correlation between income and education, we must observe both of these for the same person, do this for many people, to estimate the correlation.

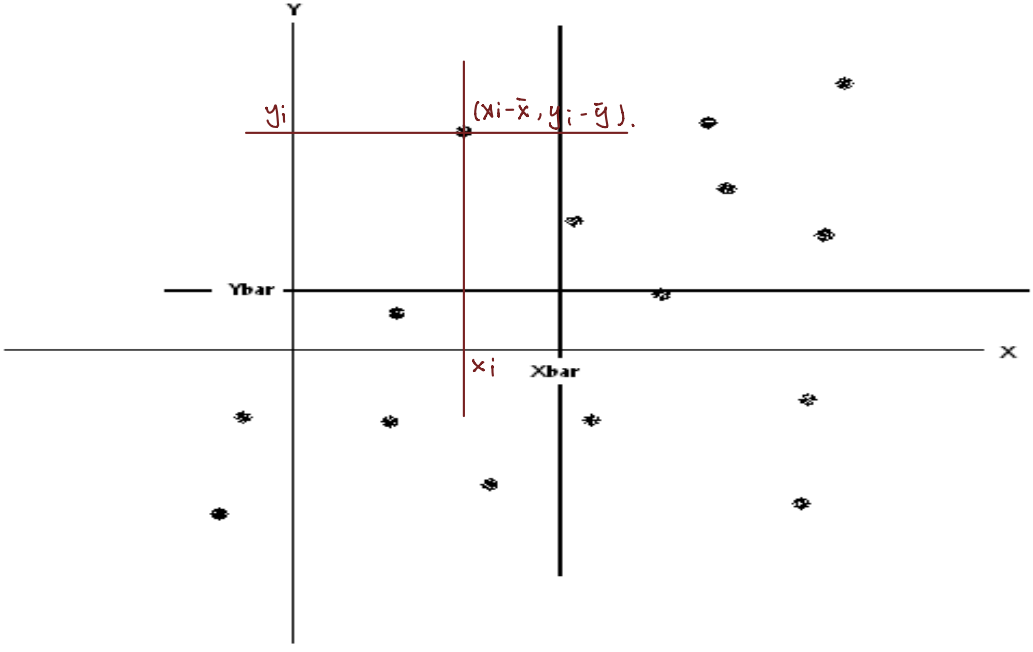
The covariance is defined as

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

where the sum is over n pairs of observations.

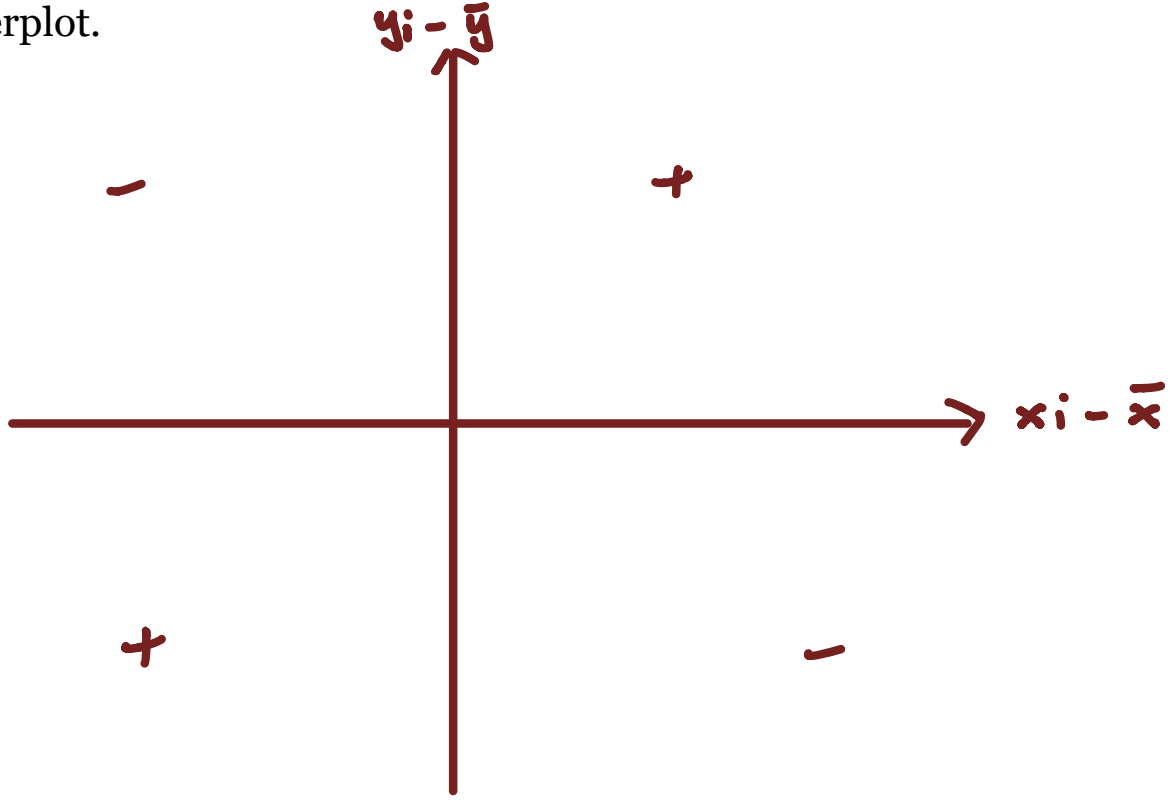
Covariance/Correlation

We can see this easily in a scatterplot



Covariance/Correlation

Notice that with the lines for the means, we have four quadrants in the scatterplot.



Covariance/Correlation

So we can see what happens to the covariance under different possibilities

(a) if there is a broad upward sweep of the data from the left to the right, which we have been calling a positive correlation, we see that most of the data points will be in the Lower Left or Upper Right quadrants. So most of the terms in the sum are positive, and we have a positive covariance.

(b) if there is a broad downward sweep of the data from the left to the right, which we would call a negative correlation, we see that most of the data points will be in the Upper Left or Lower Right quadrants. So most of the terms in the sum are negative, and we have a negative covariance.

(c) if there is no relationship, we would expect that the data would be evenly spread in all the quadrants. The positive terms are offset by the negative terms, and we have a zero covariance.

Covariance/Correlation

Notice the units of the covariance
- for income and education, it is something like dollar years
- for manatees and registrations it is something like manatee registrations.

This is troubling - just like the variance the actual number seems meaningless by itself (even though we learn something from its sign).

Instead we might want to remove the units, this is what the correlation does. The correlation is

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

unitless

不重要，只是一种关系（比例）。

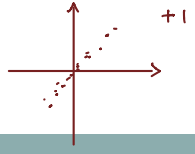
s_X standard deviation of x

s_Y standard deviation of y

This statistic has the same sign properties as the covariance (divide by a positive number always) but is bounded between -1 and 1.

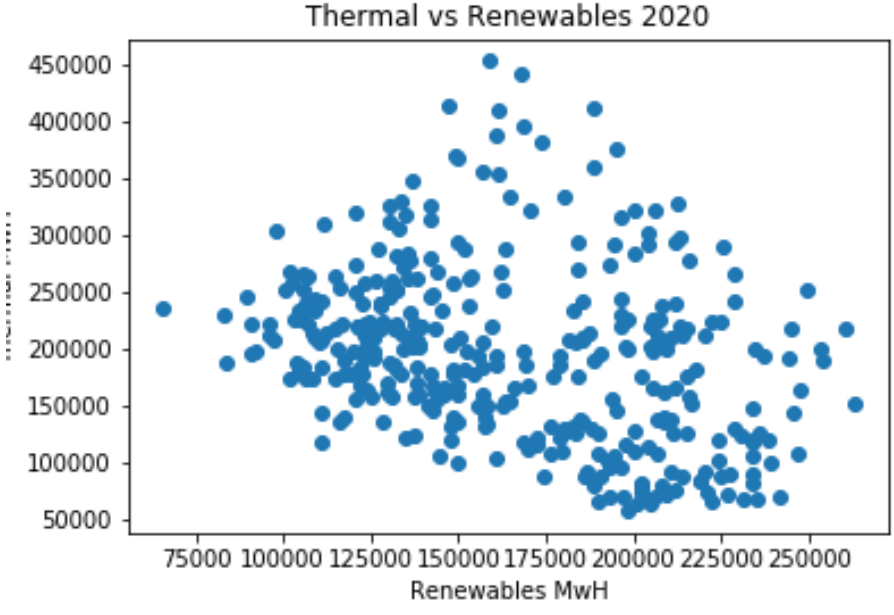
The stronger (+1, -1) the relationship, the closer is the correlation to plus or minus one.

0 is no covariance



Covariance/Correlation

Recall the relationship between renewables and thermal energy



The covariance is -983300930.

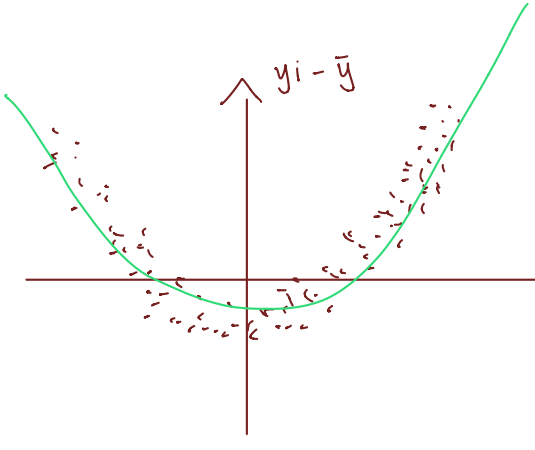
The correlation is -0.31.

Covariance/Correlation

Some issues with the correlation/covariance.

(a) measures linear relationship, i.e. could be a U shaped distribution but the positives and the negatives still offset

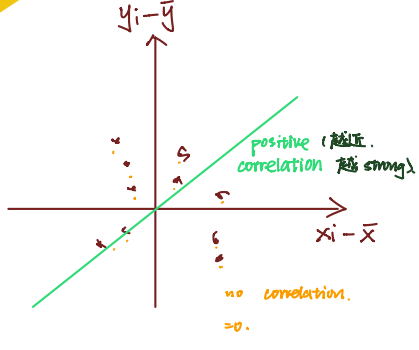
(b) as with the mean, can be seriously affected by outliers..



U shaped relationship.

=0. (there still got relationship).

[for linear line. if =0. no relationship].



Regression

The idea is to fit a line to the data summarizing the relationship.

Here it is 28667 – 0.545 * Renewables

根据线来估算图

未来Thermal

$y = \alpha + \beta x$

$\beta = \frac{S_{xy}}{S_x^2}$