

# Lecture 3: Data Description - Multiple Attributes

Graham Elliott

December 2008

# The Basic Objective

Most interesting problems relate not to means etc. but to relationships between variables.

The question then becomes how best to represent these relationships informally, either graphically or by the use of summary statistics. As in the single (univariate) case, we want to reduce the dimension without losing relevant information.

# Some Examples

- 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.

# Some Examples

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

# Some Examples

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

# Some Examples

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

# Multiple Attributes

In each of these cases, each datapoint has multiple attributes.

For example

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

# Multiple Attributes

In each of these cases, each datapoint has multiple attributes.

For example

- We observe for each mutual fund both returns and the number of different stocks held.
- For each person we might observe both the amount of violent TV seen as well as their history of being violent.

# Multiple Attributes

In each of these cases, each datapoint has multiple attributes.

For example

- We observe for each mutual fund both returns and the number of different stocks held.
- For each person we might observe both the amount of violent TV seen as well as their history of being violent.
- 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).

# Multiple Attributes

In each of these cases, each datapoint has multiple attributes.

For example

- We observe for each mutual fund both returns and the number of different stocks held.
- For each person we might observe both the amount of violent TV seen as well as their history of being violent.
- 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).
- We could observe not just the numbers of manatees that died but also the number of boat registrations.

# Raw Data

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

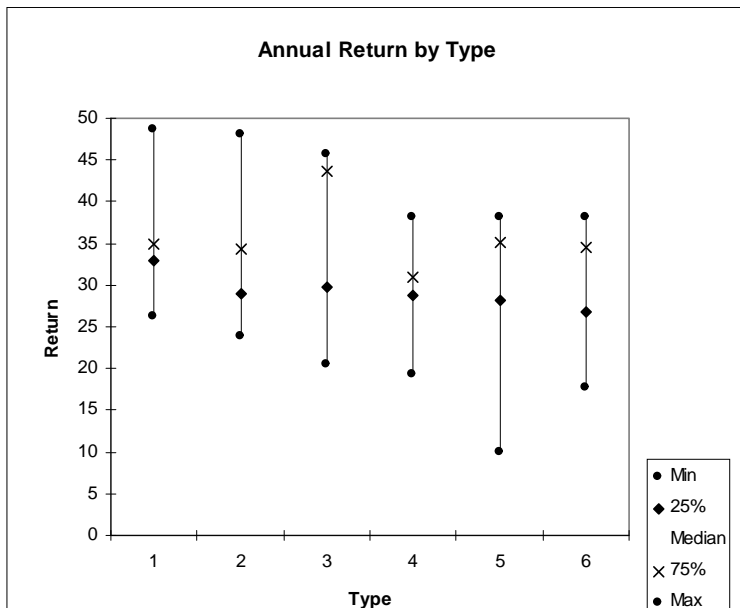
### III. Graphical Representations

If one of the variables is categorical (like college or not), we can simply use our univariate methods for each category (typically side by side).

e.g.1 Box Plots. Here one variable is mutual fund returns, the categorical data is type (large cap, medium cap, small cap, and growth or value).

e.g. 2 Histograms. Here the variable is average grade versus the categorical data by country.

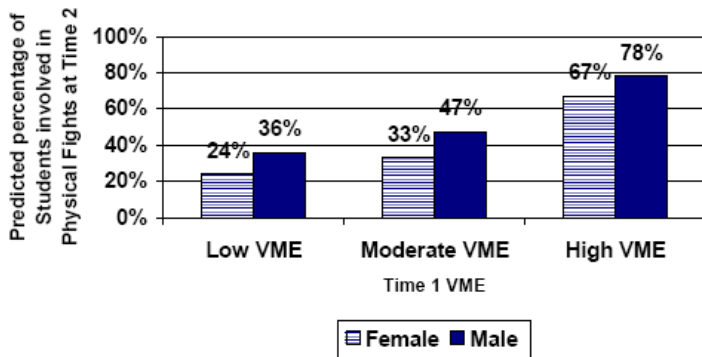
# Boxplots



# Side by side histograms

FIGURE 2

Predicted likelihood of physical fights at Time 2  
as a function of sex and media violence exposure

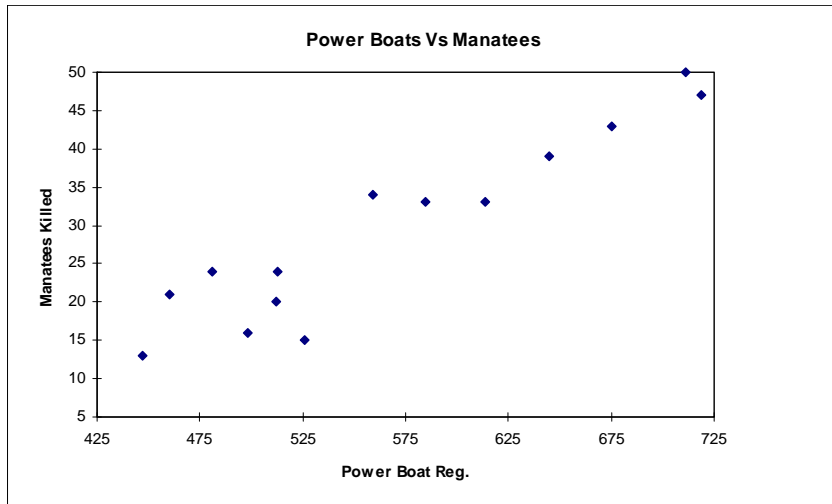


Gentile et. al. (2004), manuscript., Iowa State Univ.

If both of the variables can take on many different values, this becomes more difficult and arbitrary (we can always make categories, but this is not often the best way to go).

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

# Power Boats and Manatees



# Scatterplots

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

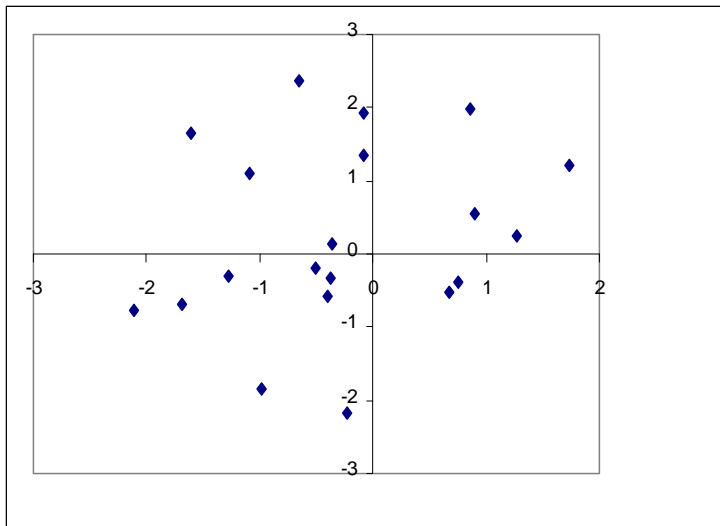
Is the relationship positive or negative? For this problem it is quite clear that there is a positive relationship between the number of manatees killed and the number of power boats out there. This is indicated by the low number of deaths when there were few boats, and more deaths when there are many.

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).

Finally, we can ask 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.

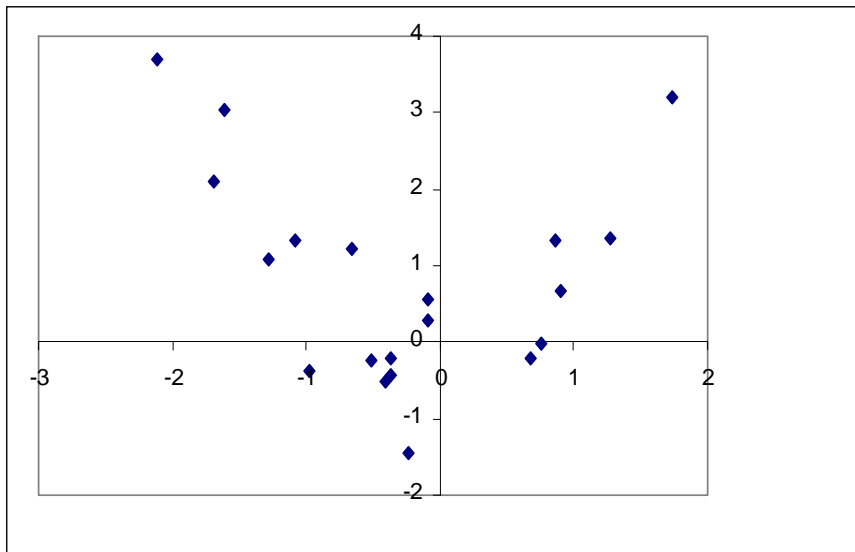
# Scatterplots

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



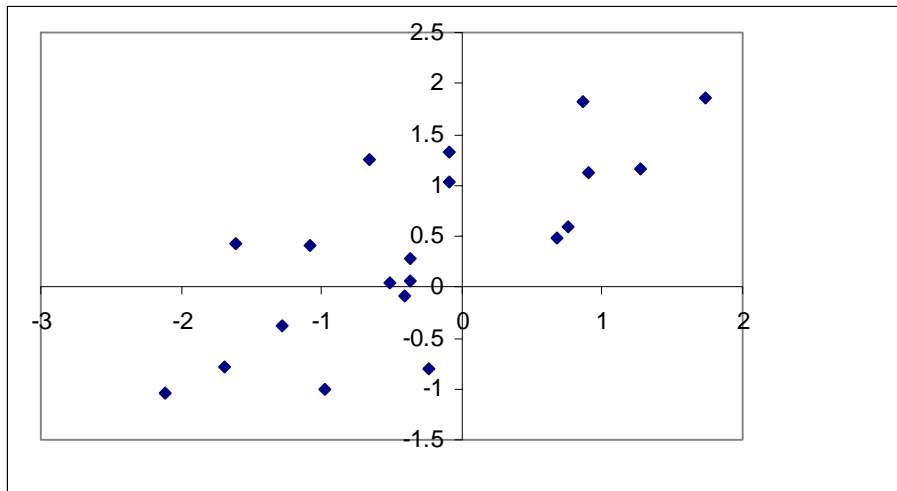
# Scatterplots

It may not be linear



# Scatterplots

It may be linear looking but not very clear



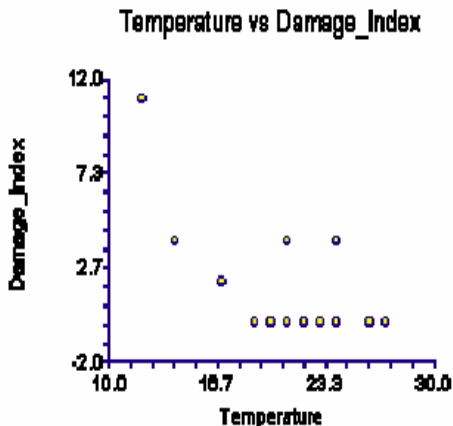
## 1986 Challenger Shuttle Disaster

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).

# Space Shuttle

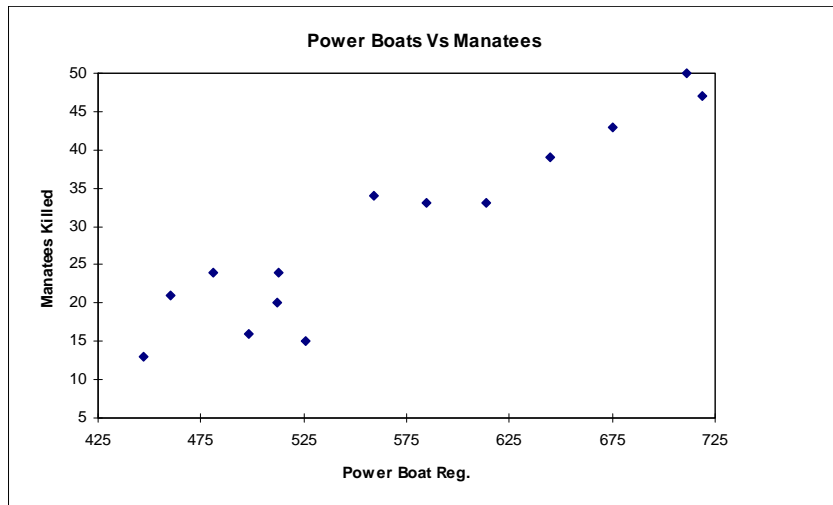
A simple scatterplot showing the link between O-ring damage and ambient temperature during previous launches may have changed the decision about launching. How much damage would you have expected at 0<sup>0</sup> Celsius?



### i) Correlation/Covariance.

Above, we talked about a positive or negative relationship between the data. We can formalize this a little. Look at the scatterplot. Suppose that we placed a vertical line at the average number of power boat registrations, and a horizontal line through the mean of the y axis. Now consider what a positive relationship would be.

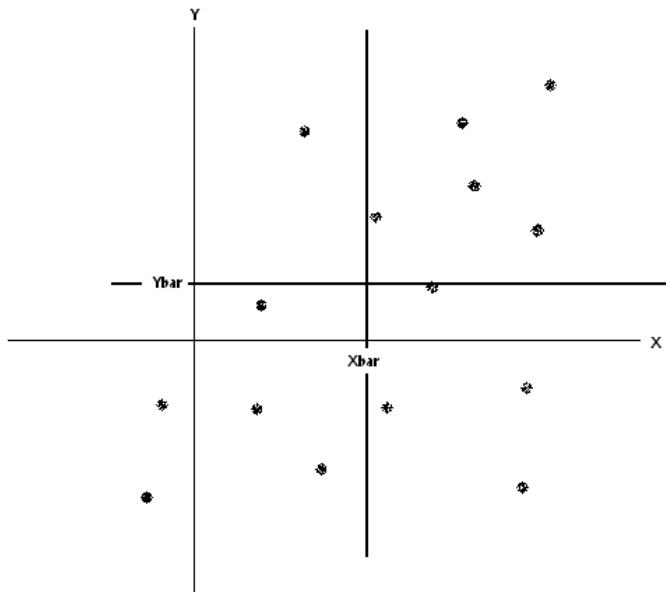
# Correlation/Covariance



Consider the following formula  
The covariance is defined as

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

# Correlation/Covariance



# Correlation/Covariance

Consider the signs of the components of the formula

$$(x_i - \bar{x})(y_i - \bar{y})$$

If we divide the Cartesian diagram into quadrants around the means of both  $x$  and  $y$

Then we have

Quadrant	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
Upper left	negative	positive	negative
Upper right	positive	positive	positive
Lower right	positive	negative	negative
Lower left	negative	negative	positive

We would prefer also to have an idea of how strong the relationship is, and also some idea of what the numbers mean.

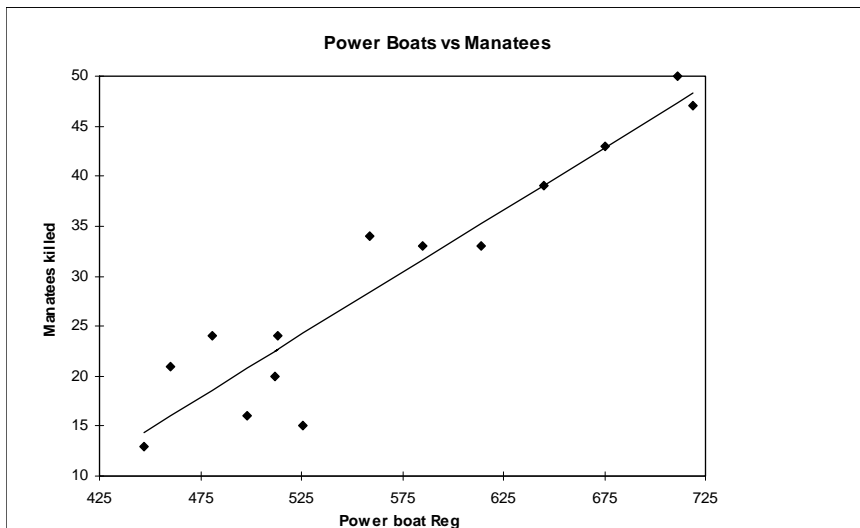
The correlation is

$$\rho_{xy} = \frac{s_{xy}}{s_x s_y}$$

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 the relationship, the closer is the correlation to plus or minus one.

# Regression

We could also think of putting a line through the data and calling this the relationship between the variables.



This is really the topic of 120b and 120c. But for this simple case the usual estimate of the regression line is

$$\hat{b} = \frac{s_{xy}}{s_x^2}$$

and

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

so involves the statistics we will be using in this course.

# Analytic Skill

