January 29, 2018

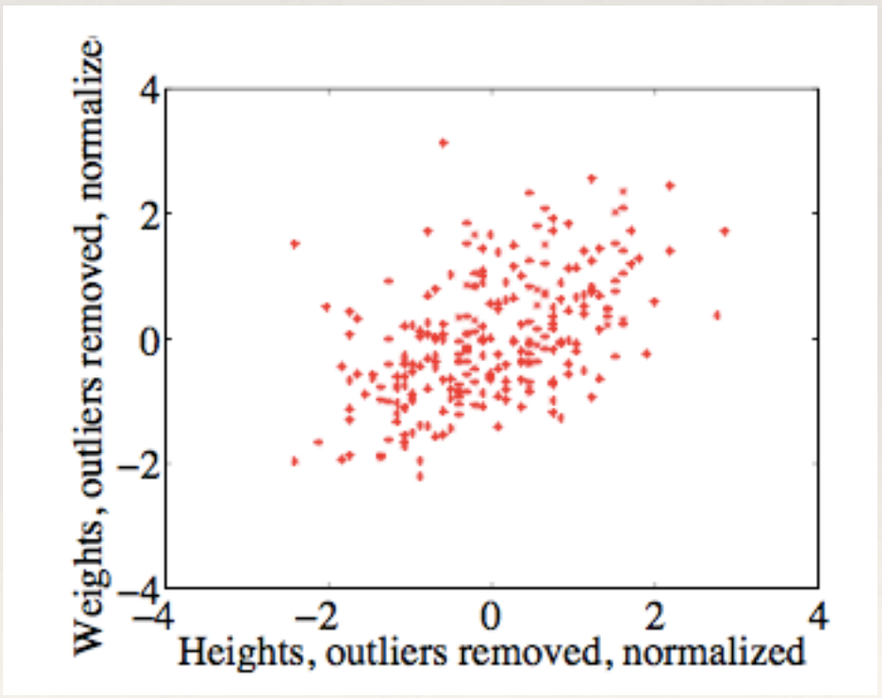
CS 361: Probability & Statistics

Relationships in data

Correlation review

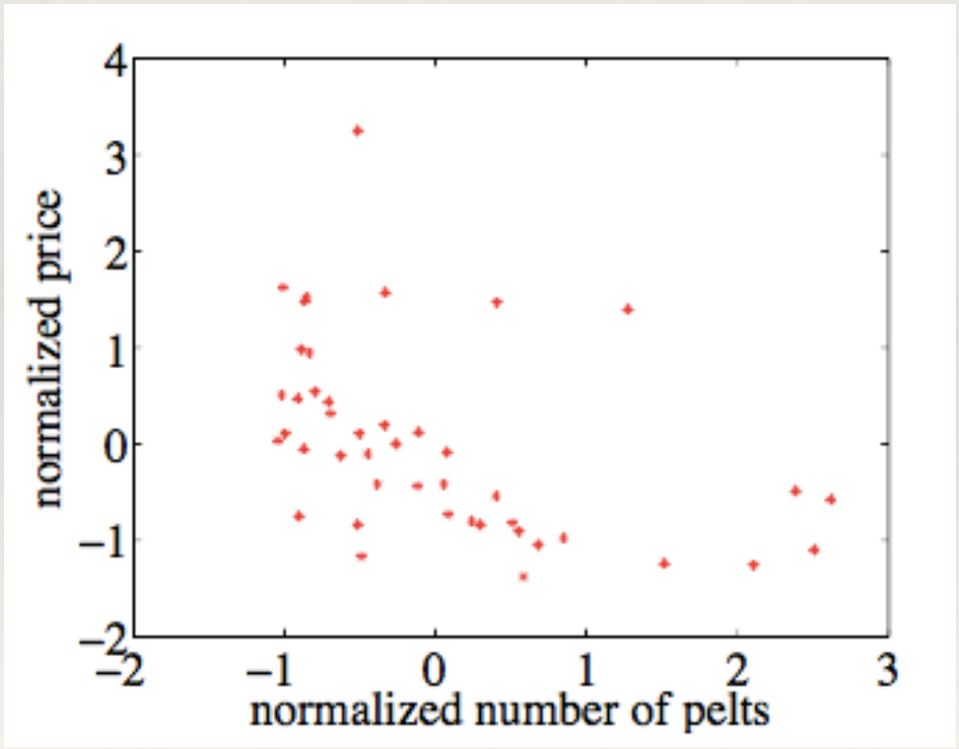
Correlation

- Broadly, if x changes, what does y do?
- If a small x and small y (respectively large x and large y) tend to occur together we say there is **positive correlation** between x and y



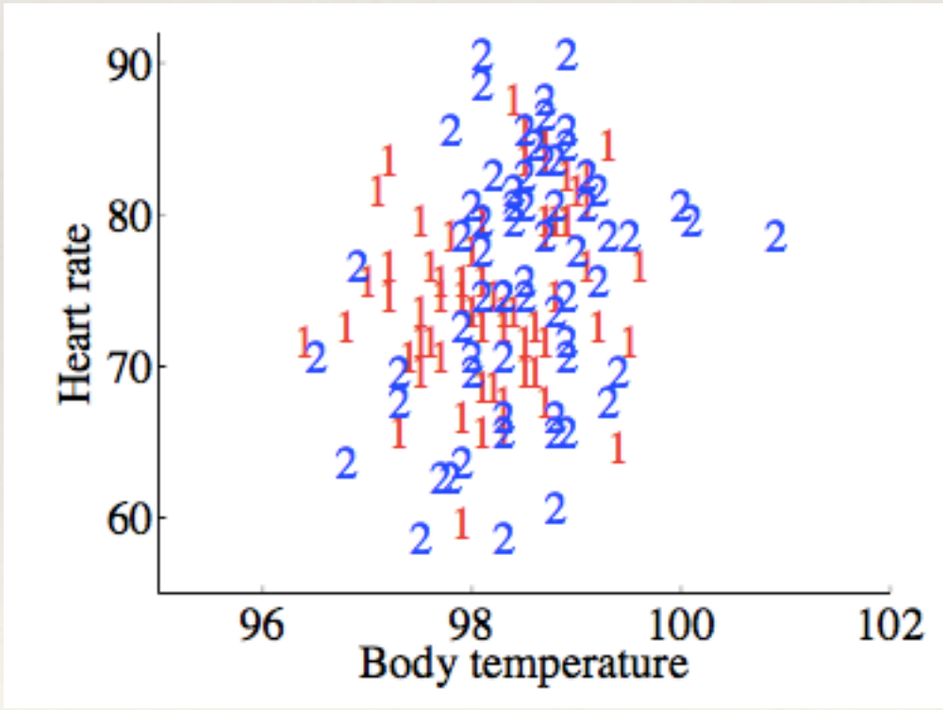
Correlation

- If small values of x tend to occur with large values of y and large values of x tend to occur with small values of y we say that x and y are **negatively correlated**



Correlation

- When there is no tendency for x and y to be either large or small together, we say there is **zero correlation**
- Our data will be more of a blob



Correlation coefficient

Suppose we have N data items that are each 2-vectors

$(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$

Normalize the data

$$\hat{x}_i = \frac{x_i - \text{mean}(\{x\})}{\text{std}(\{x\})} \qquad \hat{y}_i = \frac{y_i - \text{mean}(\{y\})}{\text{std}(\{y\})}$$

The correlation coefficient of x and y is the mean of the product $\hat{x}_i\hat{y}_i$, i.e.

$$\text{corr}(\{(x,y)\}) = \frac{\sum_i \hat{x}_i \hat{y}_i}{N}$$

we will sometimes write *r* for the correlation coefficient

Correlation

The largest possible correlation is 1 and happens when $\hat{x} = \hat{y}$

The smallest possible correlation is -1 and happens when $\hat{x} = -\hat{y}$

Using correlation for prediction

Using correlation to predict

- One useful task is to take what we know about the data we have and make predictions about data we don't yet have or measurements we have that are incomplete
- Example: we might like to go into the fur pelt business and have a bunch of historical data on supply and prices. We know the price today and would like to guess as to the total supply
- That is we have a bunch of pairs *(x,y)* for prices and supply. But our state of knowledge today might be *(x_0, ???)*
- Correlation will be useful for this task

Prediction

- We want a predictor that we can apply to any x
- We want it to behave well on our existing data
- We can choose the predictor by considering the error the predictor will have

Prediction

Since it's possible to convert to and from standard coordinates and we know standard coordinates have nice properties like 0 mean and 1 standard deviation, we will first convert

We will write $\hat{y}_i^p$ to indicate our predicted value of $\hat{y}_i$ for the point $\hat{x}_i$

Predicting a y from an x

We will look at a simple predictor: a linear predictor

So our prediction function will have the form

$$\hat{y}_i^p = a\hat{x}_i + b$$

We will come up with a single prediction function that is informed by the dataset in question. Which means we will find an *a* and *b* for the equation above

We will do this by considering the **error** of the prediction function. The error that the function would make on data item *i* is given by

$$u_i = \hat{y}_i - \hat{y}_i^p$$

{u} then is a dataset and we can perhaps use its mean and variance

Prediction

The mean of our errors should be 0 or else we could reduce our prediction error by subtracting a constant

Let's use this assumption to get a value for b in $\hat{y}_i^p = a\hat{x}_i + b$

Recalling that $u_i = \hat{y}_i - \hat{y}_i^p$ we have $\text{mean}(\{u\}) = \text{mean}(\{\hat{y} - \hat{y}^p\})$

Which we rewrite as $\text{mean}(\{u\}) = \text{mean}(\{\hat{y}\}) - \text{mean}(\{\hat{y}^p\})$

Prediction

Since $\hat{y}_i^p = a\hat{x}_i + b$ we rewrite

$$\text{mean}(\{u\}) = \text{mean}(\{\hat{y}\}) - \text{mean}(\{\hat{y}^p\})$$

as

$$\text{mean}(\{u\}) = \text{mean}(\{\hat{y}\}) - \text{mean}(\{a\hat{x} + b\})$$

Using properties of the mean we rewrite as

$$\text{mean}(\{u\}) = \text{mean}(\{\hat{y}\}) - a\,\text{mean}(\{\hat{x}\}) - b$$

Prediction

Since x and y are both in standard normal coordinates, we can simplify this

$$\text{mean}(\{u\}) = \underbrace{\text{mean}(\{\hat{y}\})}_{0} - a\underbrace{\text{mean}(\{\hat{x}\})}_{0} - b$$

getting

$$\text{mean}(\{u\}) = 0 - a0 - b$$

Recall that we said we wanted the mean of our error to be 0, so we can solve for b above and get $b=0$

Prediction

So now we have a predictor with the form $\hat{y}^p = a\hat{x}$

Ideally our error would have 0 mean and 0 standard deviation which means we always predict exactly correctly

Let's try and choose an a that minimizes standard deviation

But since we want to keep our math simple, let's equivalently find an a that minimizes the variance of the error

In order to have a shorthand available to us, let's write r for the correlation of x and y

Prediction

We want to minimize $\text{var}(\{u\}) = \text{var}(\{\hat{y} - \hat{y}^p\})$

Using the form of our prediction function, we rewrite this as

$$\text{var}(\{u\}) = \text{var}(\{\hat{y} - a\hat{x}\})$$

Remember that we can write the variance as the mean of some quantity: the squared distances from the mean of the data

Root of the mean of the squared distance

Consider the squared distance from point i to the mean

$$d_i = (x_i - \text{mean}(\{x\}))^2$$

These N distances form a dataset {d}, with mean

$$\text{mean}(\{d\}) = \frac{1}{N}\sum_{i=1}^{N} d_i$$

So $\text{std}(\{x\}) = \sqrt{\text{mean}(\{d\})}$

And $\text{var}(\{x\}) = \text{mean}(\{d\})$

Reminder

$$\text{std}(\{x_i\}) = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}(x_i - \text{mean}(\{x\}))^2}$$

Prediction

We want to minimize $\text{var}(\{u\}) = \text{var}(\{\hat{y} - \hat{y}^p\})$

Using the form of our prediction function, we rewrite this as

$$\text{var}(\{u\}) = \text{var}(\{\hat{y} - a\hat{x}\})$$

Remember that we can write the variance as the mean of some quantity: the squared distances from the mean of the data

$$\text{var}(\{u\}) = \text{mean}(\{(\hat{y} - a\hat{x} - \underline{0})^2\})$$

mean({u})

Prediction

Expanding $\text{var}(\{u\}) = \text{mean}(\{(\hat{y} - a\hat{x})^2\})$ we get

$$\text{var}(\{u\}) = \text{mean}(\{(\hat{y})^2 - 2a\hat{x}\hat{y} + a^2(\hat{x})^2\})$$

Which we rewrite as

$$\text{mean}(\{(\hat{y})^2\}) - 2a\text{mean}(\{\hat{x}\hat{y}\}) + a^2\text{mean}(\{(\hat{x})^2\})$$

Each of these teams has something we recognize

$\text{mean}(\{(\hat{y})^2\}) = (\text{std}(\{\hat{y}\}))^2$ $\quad$ $\text{mean}(\{\hat{x}\hat{y}\}) = \text{corr}(\{(x, y)\})$

$\text{mean}(\{(\hat{x})^2\}) = (\text{std}(\{\hat{x}\}))^2$ $\quad$ $\text{mean}(\{\hat{x}\hat{y}\}) = r$

Prediction

So we can simplify

$$\text{mean}(\{(\hat{y})^2\}) - 2a\,\text{mean}(\{\hat{x}\hat{y}\}) + a^2\,\text{mean}(\{(\hat{x})^2\})$$

As $\text{var}(\{u\}) = 1 - 2ar + a^2$

Since we are searching for an *a* that minimizes variance, we take the derivative and set equal to 0 and get

$$-2r + 2a = 0$$

So we will use *a*=*r* in our predictor

Summary of what we proved

We wanted a way of predicting y from x

We chose to think in standard coordinates and to use a linear predictor of the form

$$\hat{y}_i^p = a\hat{x}_i + b$$

Assuming the mean of the predictor error was 0 gave us $b=0$

Minimizing the variance of the error gave us $a=r$

So our final predictor is

$$\hat{y}_i^p = r\hat{x}_i$$

Prediction

- Here is our process for predicting y_0 from x_0

- Transform the data set into standard coordinates, to get

$$\hat{x}_i = \frac{1}{\text{std}(x)}(x_i - \text{mean}(\{x\}))$$

$$\hat{y}_i = \frac{1}{\text{std}(y)}(y_i - \text{mean}(\{y\}))$$

$$\hat{x}_0 = \frac{1}{\text{std}(x)}(x_0 - \text{mean}(\{x\})).$$

- Compute the correlation

$$r = \text{corr}(\{(x, y)\}) = \text{mean}(\{\hat{x}\hat{y}\}).$$

- Predict $\hat{y}_0 = r\hat{x}_0$.
- Transform this prediction into the original coordinate system, to get

$$y_0 = \text{std}(y) r \hat{x}_0 + \text{mean}(\{y\})$$

Prediction

We have $\hat{y}^p = r\hat{x}_0$

Or

$$\frac{y^p - \text{mean}(\{y\})}{\text{std}(y)} = r\frac{x_0 - \text{mean}(\{x\})}{\text{std}(x)}$$

Another way of reading this is if x_0 is *k* standard deviations from its mean, predict a *y* that is *kr* standard deviations from its mean

Or that the predicted value of *y* goes up by *r* standard deviations for every 1 standard deviation that *x* increases by

Predictor error

- We constructed our predictor so that the mean of the error was 0
- This does not mean that we will make zero errors or even make a small number of errors, though. Why?
- It is useful to look at the RMS of our errors

$$\sqrt{\text{mean}(\{(y^p - \hat{y})^2\})} = \sqrt{\text{mean}(\{u^2\})}$$

Prediction error

Let's see if we can simplify this RMS error. First we make a substitution based on our prediction function

$$\text{mean}(\{(y^p - \hat{y})^2\}) = \text{mean}(\{(r\hat{x} - \hat{y})^2\})$$

Expanding

$$\text{mean}(\{(u)^2\}) = r^2\,\text{mean}(\{(\hat{x})^2\}) - 2r\,\text{mean}(\{\hat{x}\hat{y}\}) + \text{mean}(\{(\hat{y})^2\})$$

var({x-hat})=1 r var({y-hat})=1

$$\text{mean}(\{(u)^2\}) = r^2 - 2r^2 + 1$$

$$\text{RMS error} = \sqrt{1 - r^2}$$

Prediction error

- How can we interpret the error of our predictor?

$$\text{RMS error} = \sqrt{1 - r^2}$$

- Error depends on how correlated the data are, a strong negative or positive correlation gives us better prediction performance
- Low correlation makes for a bad predictor

Takeaway

- We are able to make predictions and have more or less confidence in them depending on our data
- We can spot correlation visually

Correlation confusion

Correlation confusion

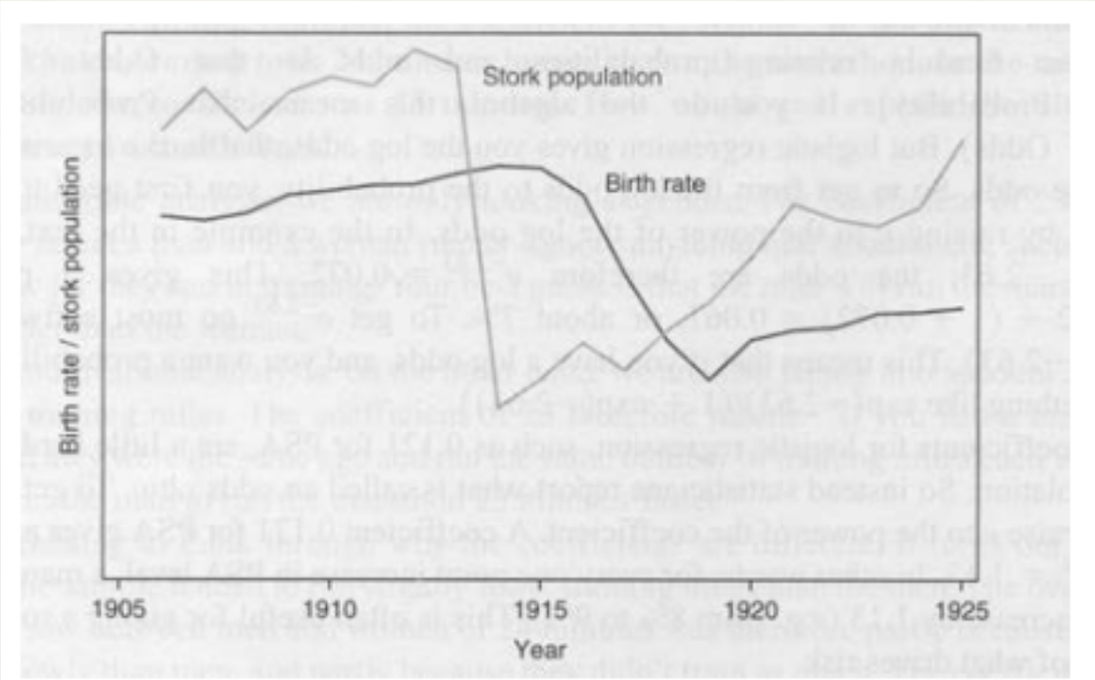
- We can observe or calculate when data tend to vary positively or negatively with one another
- If we look at enough pairs of variables, this can happen by chance

Correlation confusion

- Correlation can happen because there is a causal relationship
- The percentage you have your accelerator pressed down where 100% is all the way to the floor and your actual acceleration will be correlated

Correlation confusion

- Latent variables

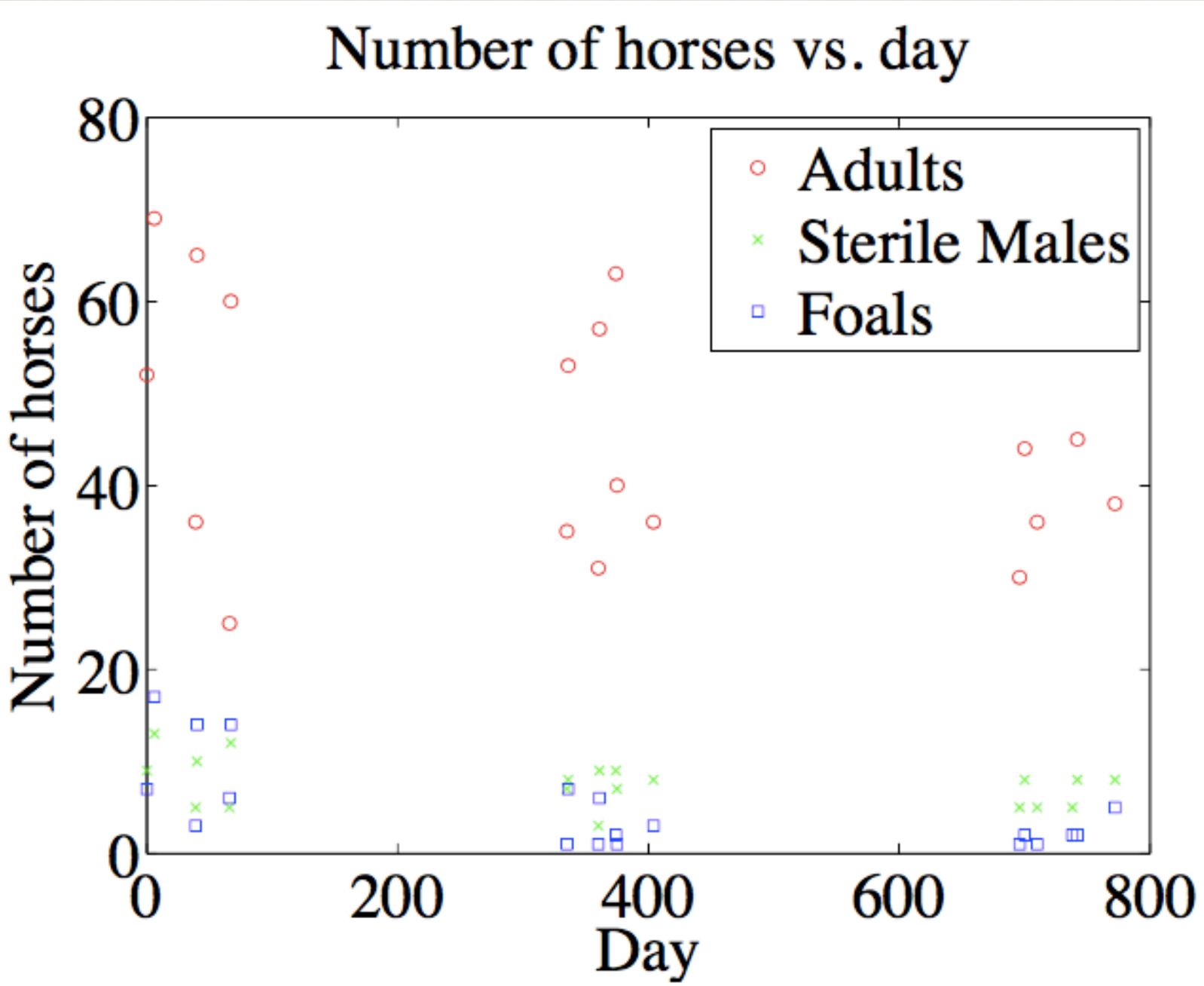


- Shoe size and reading comprehension

Example

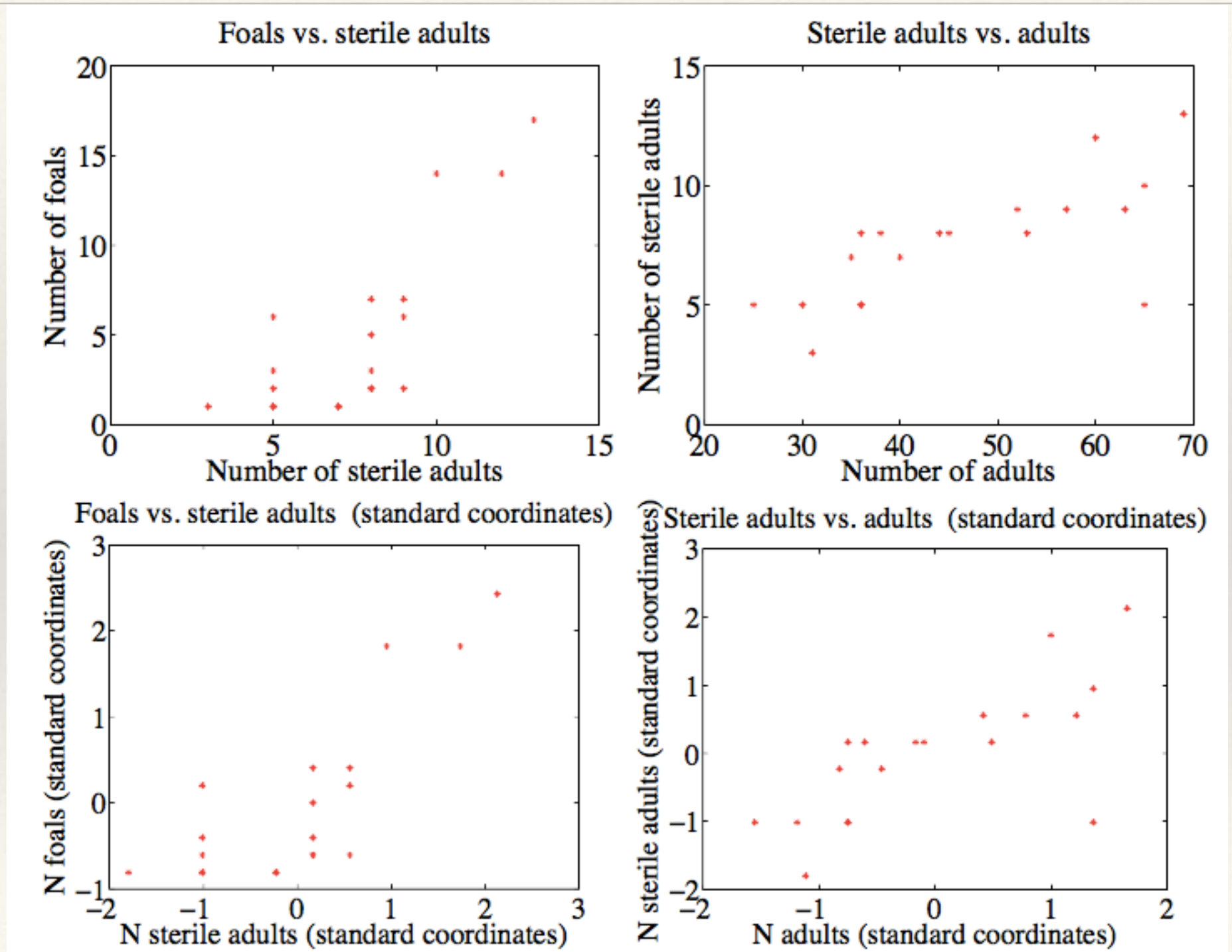
- Controlling wild horse populations
- Hypothesis is that sterilizing some males will cause the number of new births to go down
- What should we expect in terms of correlation and what might our scatterplots look like if we are right or wrong?

Example



Should expect a correlation between the number of sterile males and the number of foals?

Sterile males

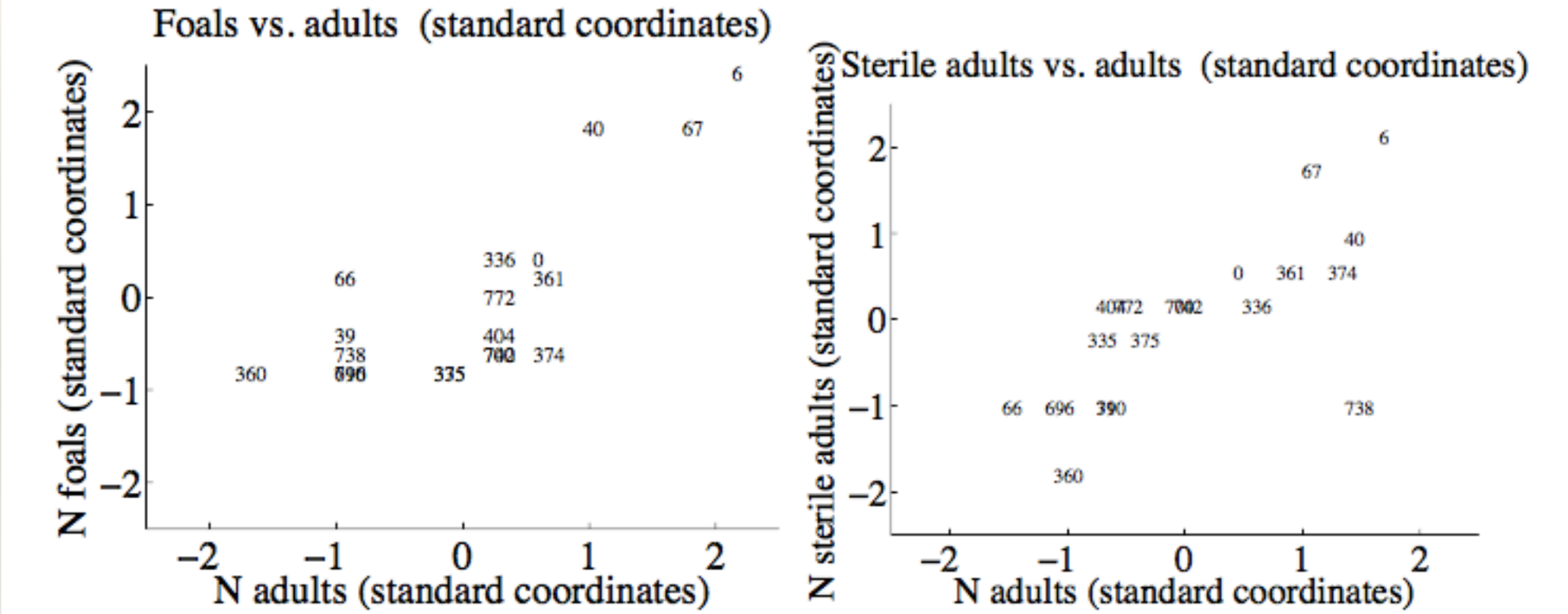


Correlation

- Correlation between sterile males and foals is 0.74
- Between sterile males and adults is 0.68
- What's going on?

Day of observation plot

The whole herd is shrinking over time



Correlation again

- Correlation between # of adults and day is -0.24
- Correlation between # of foals and day is -0.61
- Takeaway: you might need to think beyond just what correlation is telling you and plot things several ways