September 12, 2017

CS 361: Probability & Statistics

Correlation

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 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

- So here is our process for predicting y_0 from x_0

- Transform the data set into standard coordinates, to get

$$\hat{x}_i = \frac{1}{\mathsf{std}(x)}(x_i - \mathsf{mean}(\{x\}))$$
$$\hat{y}_i = \frac{1}{\mathsf{std}(y)}(y_i - \mathsf{mean}(\{y\}))$$
$$\hat{x}_0 = \frac{1}{\mathsf{std}(x)}(x_0 - \mathsf{mean}(\{x\})).$$

- Compute the correlation

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

- Predict $\hat{y}_0 = r\hat{x}_0$.

- Transform this prediction into the original coordinate system, to get

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

Prediction

- So 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

- Substituting in our predictor, we get

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

- Expanding and simplifying

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

- And then substituting, we get

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

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

Prediction error

- How can we interpret the error of our predictor?

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

- It depends on how correlated the data are, a strong negative or positive correlation gives us better prediction performance
- No 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

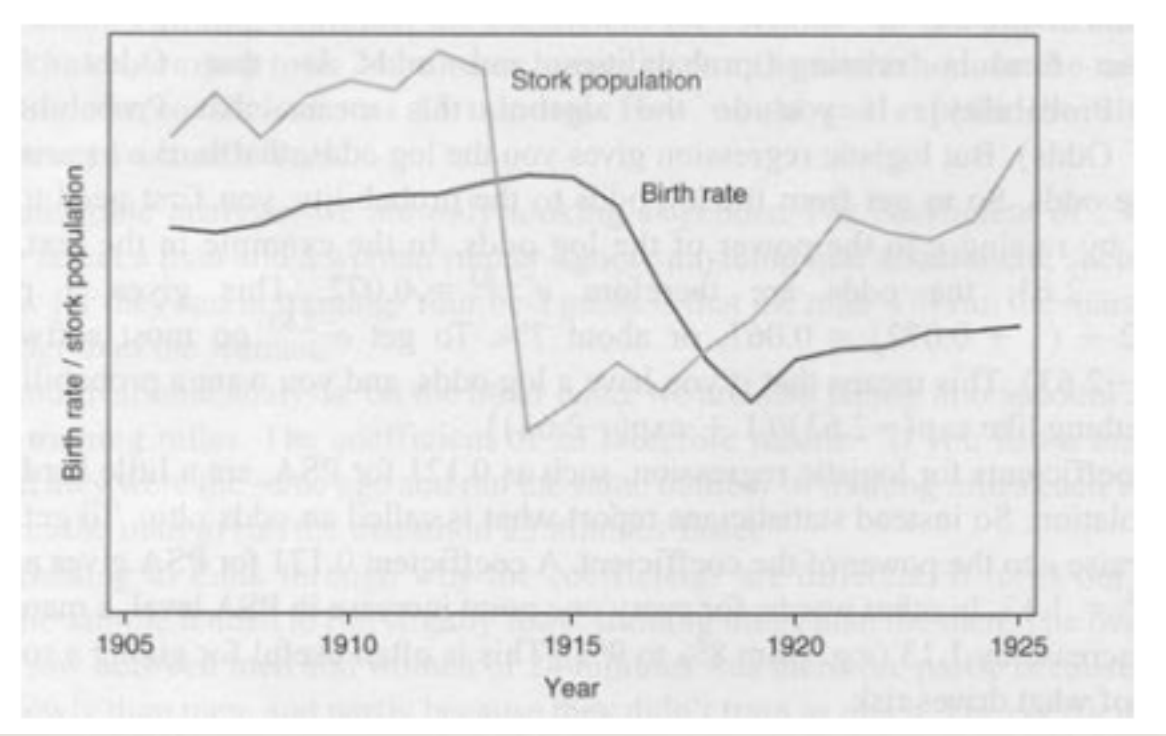
- We can observe or calculate when data tend to vary positively or negatively with one another
- If we look at enough pairs of variables, this can happen by chance

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

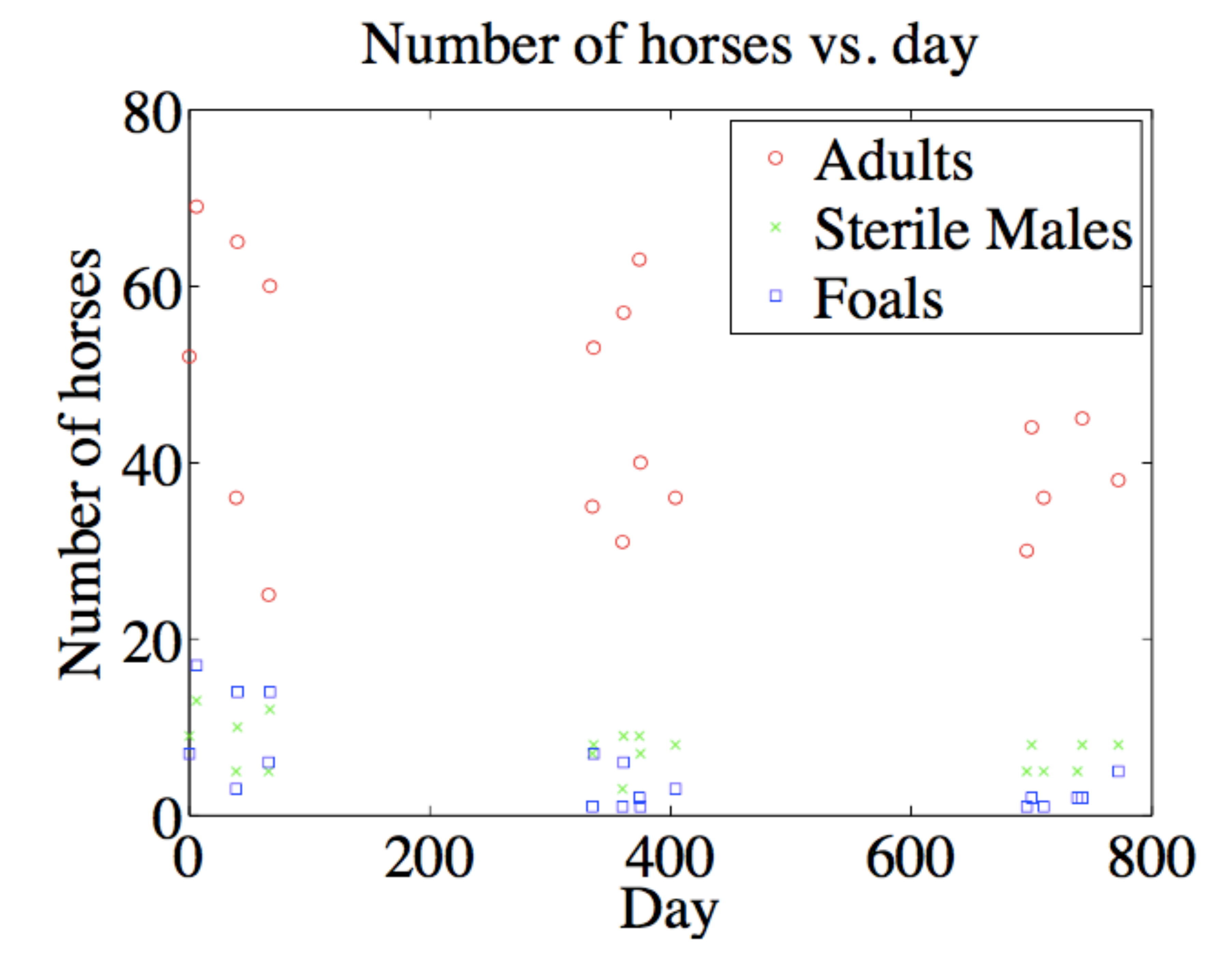


- Shoe size and reading comprehension

Example

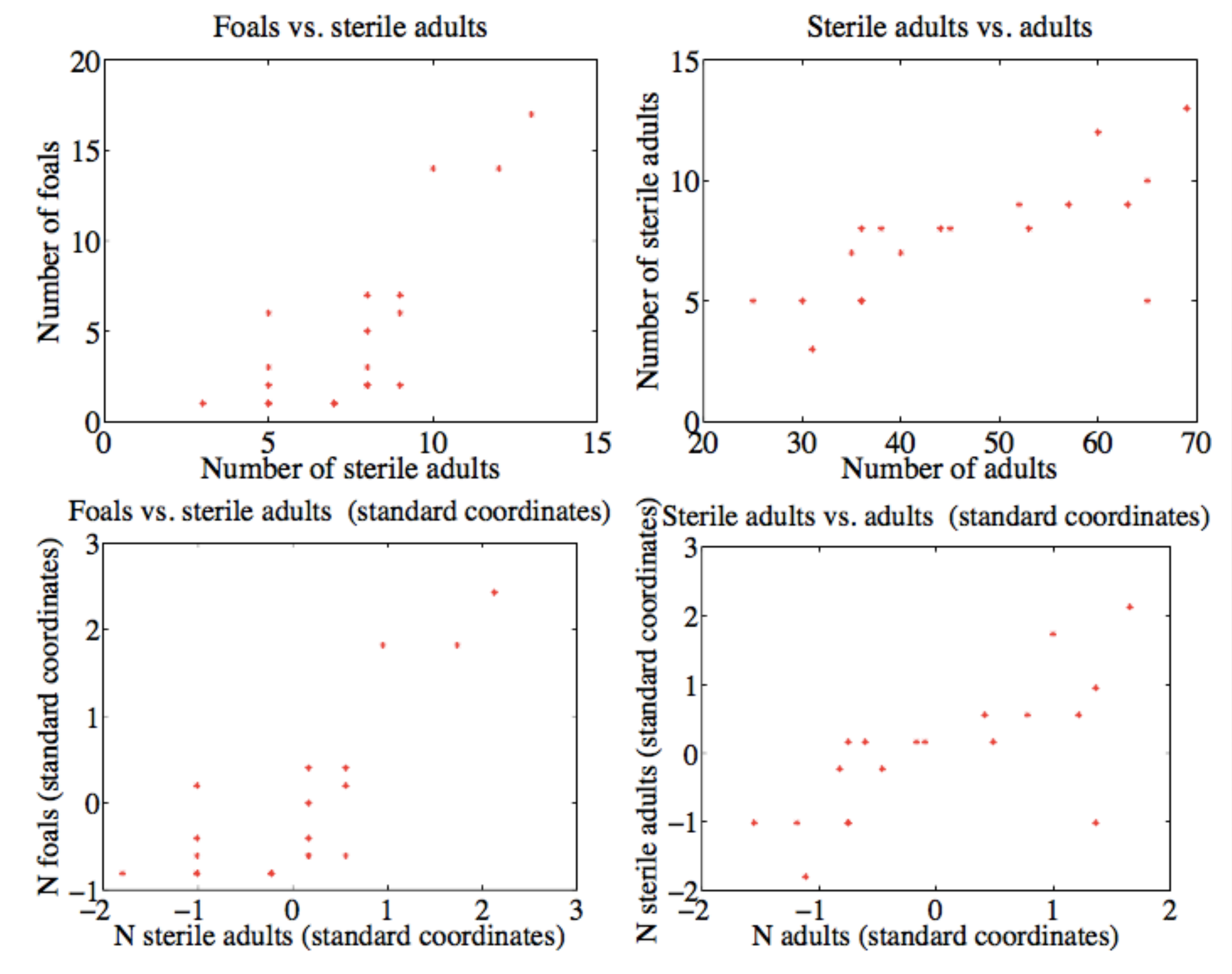
- Controlling wild horse populations
- Hypothesis is that sterilizing some males will cause the number of new births to go down
- What should we expect in terms of correlation and what might our scatterplots look like if we are right or wrong?

Example



Number of horses vs. day
Adults
Sterile Males
Foals
Number of horses
0
20
40
60
80
0
200
400
600
800
Day

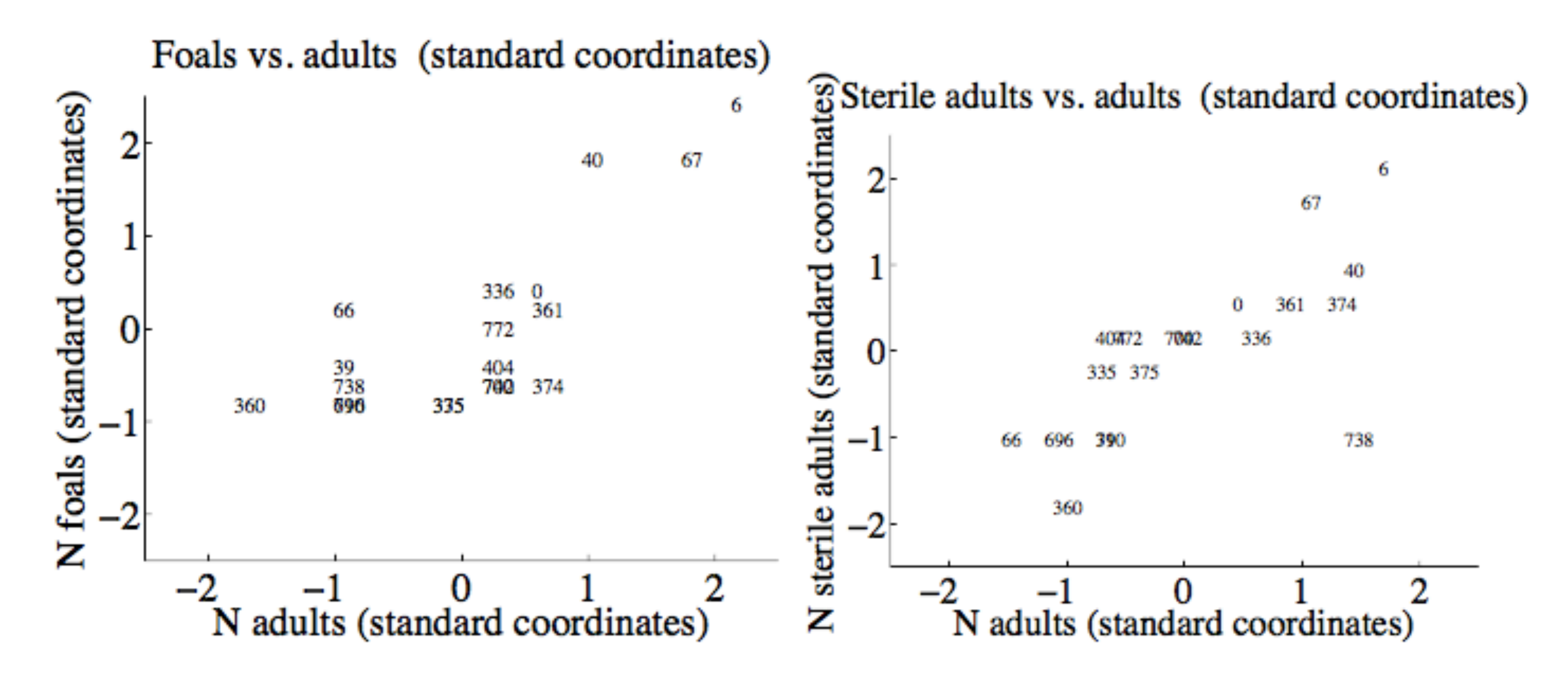
Sterile males



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



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

Probability theory

Probability

- Reasoning about uncertain outcomes with formal models
- Allows us to compute probabilities
- Experiments will be our data generating process

Outcomes

- If we toss a fair coin a bunch of times we expect about the same number of heads and tails
- If we roll a die, we don't expect to see one number more often than any other
- We can formally state the set of **outcomes** (heads, tails) we expect from an **experiment** (flipping the coin)

Outcomes

- Tossing a fair coin once: {H, T}
- Tossing a die: {1, 2, 3, 4, 5, 6}
- Tossing two coins: {HH, HT, TH, TT}

Sample space

- The **sample space** is the set of all possible outcomes of an experiment, written Ω

Example

- Three playing cards: King (K), Queen (Q), Knight (N). One is turned over randomly
- What is the sample space?
- {K, Q, N}

Example

- Suppose we flip a card, turn it back over, rearrange the cards and flip another. What is the sample space?
- {KK, KQ, KN, QQ, QK, QN, NN, NK, NQ}

Example

- A couple decides to have children until they have both a boy and a girl or until they have three children
- What is the sample space?
- {BG, GB, BBG, GGB, BBB, GGG}

Example: Monty Hall

- There are three doors: door #1, door #2, door #3. Behind two of them are goats, behind one is a car. The goats are indistinguishable
- If we open the doors and note what we observe the sample space is {CGG, GCG, GGC}

Monty hall #2

- Consider the Monty Hall scenario with distinguishable goats, one male and one female
- {CFM, CMF, FCM, FMC, MCF, MCF}
- Notice how there are more outcomes here

More family planning

- A couple decides to have children
- They decide to have children until a girl and then a boy is born
- What is the sample space here?
- The set of all strings that end in GB and contain no other GBs
- As a regular expression: B*G+B
- Somewhere between two and infinite children

Sample spaces

- Each run of an experiment has exactly one outcome
- The set of all possible outcomes is the sample space
- We will need to think of sample spaces to think rigorously about probability
- Sample spaces can be finite or infinite

Probability

- We might like to think of how often we will see each particular outcome A if we repeat an experiment over and over

$$\lim_{N \to \infty} \frac{\#\{A\}}{N}$$

- If our experiment if flipping a coin and we repeat it a large number of times
- We probably want the relative frequencies of heads and tails to be non-negative and we want the frequency of either outcome to be at most 1
- Finally we want the relative frequencies of heads and tails to add up to 1

Probability

- More generally, for an experiment with sample space Ω intuition tells us that we want each outcome *A* to satisfy

$$0 \leq P(A) \leq 1$$

- And we also want the following

$$\sum_{A_i \in \Omega} P(A_i) = 1$$

Example

- If we have a biased coin where $P(H) = 1/3$ and $P(T) = 2/3$ and we toss it three million times, how many times will we expect to see heads?
- We will see close to a million heads and 2 million tails

Example

- We often look at experiments where each outcome is equally likely
- In the example earlier with 3 cards, a King, Queen, and Knight where we turn one over at random. If each is equally likely, what probability should we assign to each card?

Example

- Recall the Monty Hall setup: 3 doors, 2 with a goat, and one with a car
- If we open the first door, what is the probability that we see a goat? What is the probability we see the car?
- P(car) = 1/3, P(goat) = 2/3

Example

- Recall the Monty Hall setup with distinguishable goats
- What is the probability we find a female goat behind door #1?
- P(female goat) = 1/3

Events

- We are often interested in sets of outcomes
- For example, we might flip a coin three times
- Our sample space, the set of all outcomes, is {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}
- "Getting two tails" might be something we are interested int he probability of and is the set of outcomes {HTT, THT, TTH}
- We give sets of outcomes a special name, **events**, and the theory we develop will concern the probabilities of events