STSCI 4780
Relationships between variables: Preliminaries (Conditional dependence & independence, graphical models, regression)

Tom Loredo, CCAPS & SDS, Cornell University

Agenda

1. **Relationships between variables**

2. **Joint distributions and graphical models**

3. **Example: Binomial prediction**

Agenda

1. **Relationships between variables**
2. **Joint distributions and graphical models**
3. **Example: Binomial prediction**

Relationships between variables

We're interested in settings where each case/item/object has *two or more properties* $(x, y, \ldots)$; we want to learn how they are related

Goals

- **Explanatory:** Seek to understand the processes/mechanisms linking x and y...
- **Predictive:** Seek to predict a future y value from observing or controlling a future x value

We will develop tools and terminology for building and describing explanatory and predictive models for multivariate data

Terminology

Types of studies

- **Correlation/dependence:** Learn about the *joint distribution*, $p(x, y)$, in settings where x and y are both potentially uncertain/random
- **Regression/conditional density estim'n:** Learn about the *conditional distribution*, $p(y|x)$, in settings where x is controllable/deterministic, or in settings where x is random but becomes known

Names of variables (conditional/regression setting)

- x: covariate, regressor, predictor, explanatory variable, input, independent variable
- y: response, prediction, output, dependent variable
- Either/both may be vectors

Conditional distribution properties

- **Regression function:** The *conditional expectation value* (conditional mean) of y *given* x is the regression function

$$f(x) = \mathbb{E}(y|x) \equiv \int dy\ y\ p(y|x)$$

- **Variance:**
  - $\mathrm{Var}(y|x) = \text{Const}$: *homoskedastic*
  - $\mathrm{Var}(y|x) \neq \text{Const}$: *heteroskedastic*

Regression = Learning a conditional expectation

Conditional density estimation = Learning a conditional distribution, $p(y|x, \cdots)$

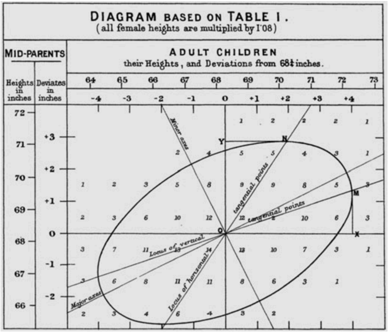
(Joint) Density estimation = Learning $p(x, y)$ (when x is also uncertain/random)

Examples with random x

Population studies

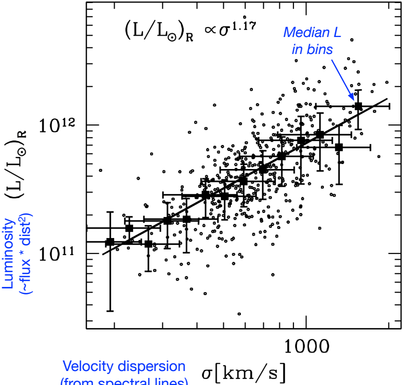
Heights of parents ("midparent") and children

Contour of counts in cells



Galton (1885) "Regression Towards Mediocrity in Hereditary Stature"

Faber-Jackson relation for elliptical galaxies

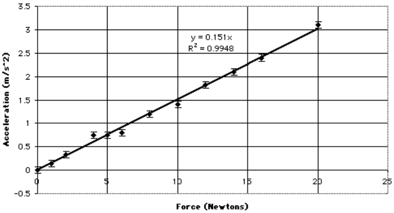


Bernardi+ (2002)

Examples with deterministic x
Curve fitting

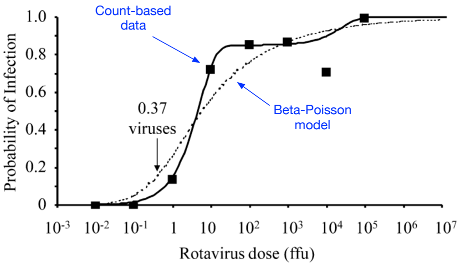
Newton's 2nd law: $a = \frac{F}{m}$

Apply different forces to a fixed mass



Batesville HS AP Physics Class

Dose-response curve



Gale (2003), "Developing risk assessments of waterborne microbial contaminations"

Agenda

1. Relationships between variables
2. **Joint distributions and graphical models**
3. Example: Binomial prediction

Joint, conditional, and marginal distributions

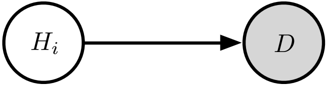
Bayesian inference is largely about the interplay between *joint*, *conditional*, and *marginal* distributions for related quantities

Ex: Bayes's theorem relating hypotheses and data ($||\mathcal{C}$):

$$P(H_i|D) = \frac{P(H_i)P(D|H_i)}{P(D)} = \frac{P(H_i, D)}{P(D)} = \frac{\text{joint for everything}}{\text{marginal for knowns}}$$

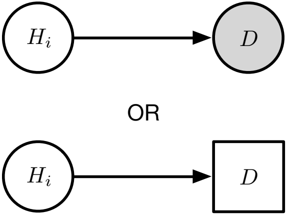
The usual form identifies an *available factorization* of the joint

Express this via a *directed acyclic graph* (DAG):



Joint distribution structure as a graph

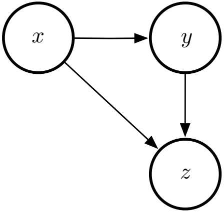
- Graph = *nodes/vertices* connected by *edges/links*
- Circular/square nodes/vertices = a priori uncertain/random quantities
    - Gray or square = quantity becomes known as data
- Directed edges specify conditional dependence
- Absence of an edge indicates conditional *in*dependence
  $\rightarrow$ a variable can be *dropped* in a factor in the joint
  $\rightarrow$ *the most important edges are the missing ones*



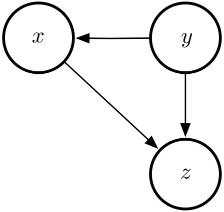
$$P(H_i, D) = P(H_i) \times P(D|H_i)$$

$$p(x, y, z)$$

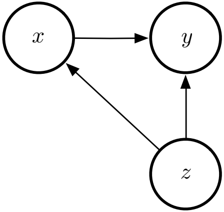
$p(x)p(y|x)p(z|x,y)$



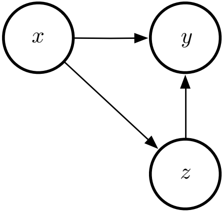
$p(y)p(x|y)p(z|y,x)$



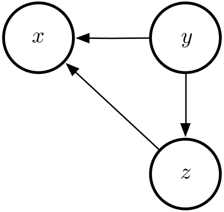
$p(z)p(x|z)p(y|z,x)$



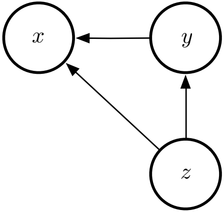
$p(x)p(z|x)p(y|x,z)$



$p(y)p(z|y)p(x|y,z)$

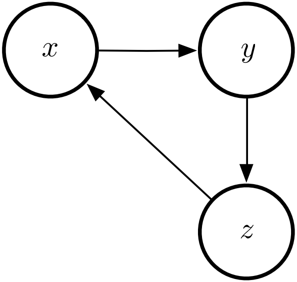


$p(z)p(y|z)p(x|z,y)$



Cycles not allowed

$$p(x|z) \times p(y|x) \times p(z|y)?$$



We can focus on *directed acyclic graphs* (DAGs)

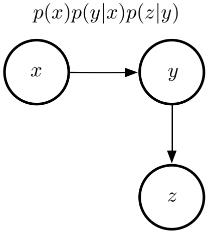
Conditional independence

Suppose for the problem at hand z is independent of of x when y is known:

$$p(z|x,y) = p(z|y)$$

"z is *conditionally independent* of x, given y"

$$z \perp\!\!\!\perp x \mid y$$



Absence of an edge indicates conditional *in*dependence

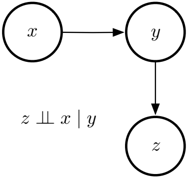
Missing edges indicate simplification in structure

$\rightarrow$ *the most important edges are the missing ones*

DAGs with missing edges

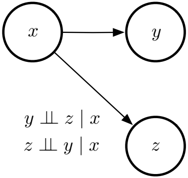
Conditional independence

$p(x)\, p(y|x)\, p(z|y)$



“Causal chain”

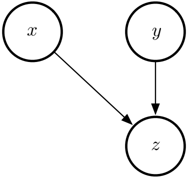
$p(x)\, p(y|x)\, p(z|x)$



“Common cause”

Conditional dependence

$p(x)\, p(y)\, p(z|x, y)$



“Common effects”

Conditional vs. complete independence

"z is *conditionally* independent of x, given y"

$$\neq$$

"z is independent of x"

(Complete) independence would imply:

$$p(z|x) = p(z) \quad \text{(i.e., not a function of } x\text{)}$$

Conditional independence is weaker:

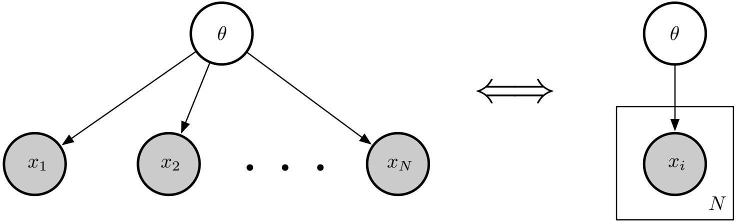
$$\begin{aligned} p(z|x) &= \int dy\ p(z, y|x) \\ &= \int dy\ p(y|x)\, p(z|x, y) \\ &= \int dy\ p(y|x)\, p(z|y) \quad \text{since } z \perp\!\!\!\perp x \mid y \end{aligned}$$

Although x drops out of the last factor, x dependence remains in $p(y|x)$

x *does* provide information about z, but it only does so through the information it provides about y (which directly influences z)

Bayes's theorem with IID samples

For model with parameters θ predicting data $D = \{x_i\}$ that are IID given θ:



$$p(\theta, D) = p(\theta)p(\{x_i\}|\theta) \;=\; p(\theta)\prod_{i=1}^{N} p(x_i|\theta)$$

"IID" means each datum is *conditionally independent* of others, *given* θ

To find the posterior for the unknowns (θ), divide the joint by the marginal for the knowns ($\{x_i\}$):

$$p(\theta|\{x_i\}) = \frac{p(\theta)\prod_{i=1}^{N} p(x_i|\theta)}{p(\{x_i\})} \quad \text{with} \quad p(\{x_i\}) = \int d\theta\, p(\theta)\prod_{i=1}^{N} p(x_i|\theta)$$

Agenda

Binomial counts

... n_1 heads in N flips

... n_2 heads in N flips

Suppose we know n_1 and want to predict n_2

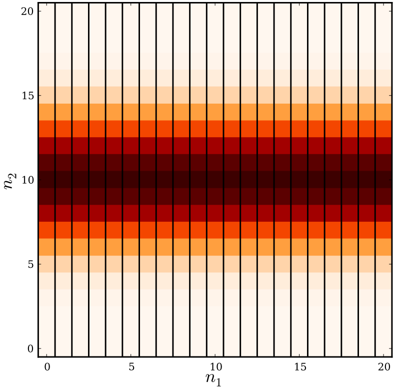
Predicting binomial counts — known α

Success probability $\alpha \rightarrow p(n|\alpha) = \frac{N!}{n!(N-n)!}\alpha^n(1-\alpha)^{N-n} \qquad || \, N$

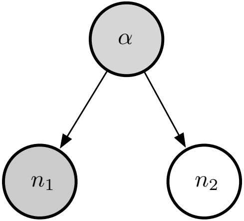
Consider two successive runs of $N = 20$ trials, *known* $\alpha = 0.5$

$$p(n_2|n_1, \alpha) = p(n_2|\alpha) \qquad || \, N$$

n_1 and n_2 are *conditionally independent*



DAG for binomial prediction — known α



$$p(\alpha, n_1, n_2) = p(\alpha)p(n_1|\alpha)p(n_2|\alpha)$$

$$\begin{aligned}
p(n_2|\alpha, n_1) &= \frac{p(\alpha, n_1, n_2)}{p(\alpha, n_1)} \\
&= \frac{p(\alpha)p(n_1|\alpha)p(n_2|\alpha)}{p(\alpha)p(n_1|\alpha)\sum_{n_2} p(n_2|\alpha)} \\
&= p(n_2|\alpha)
\end{aligned}$$

Knowing α lets you predict each n_i, independently

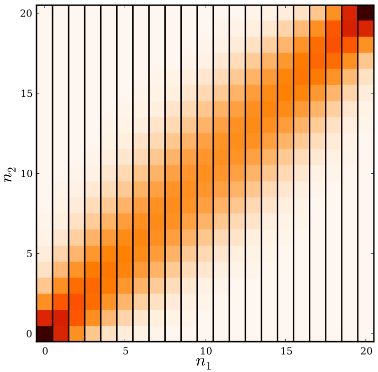
Predicting binomial counts — uncertain α

Consider the same setting, but with α *uncertain*

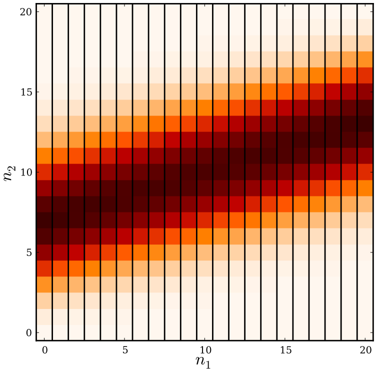
Outcomes are *physically* independent, but n_1 tells us about $\alpha \rightarrow$ outcomes are *marginally dependent* (see Lec 12 for calculation):

$$p(n_2|n_1, N) = \int d\alpha\ p(\alpha, n_2|n_1, N) \ = \ \int d\alpha\ p(\alpha|n_1, N)\, p(n_2|\alpha, N)$$

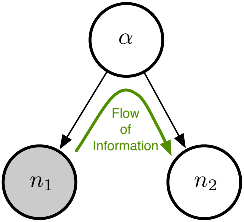
Flat prior on α



Prior: $\alpha = 0.5 \pm 0.1$



DAG for binomial prediction



$$p(\alpha, n_1, n_2) = p(\alpha)p(n_1|\alpha)p(n_2|\alpha)$$

From joint to conditionals:

$$p(\alpha|n_1, n_2) = \frac{p(\alpha, n_1, n_2)}{p(n_1, n_2)} = \frac{p(\alpha)p(n_1|\alpha)p(n_2|\alpha)}{\int d\alpha\ p(\alpha)p(n_1|\alpha)p(n_2|\alpha)}$$

$$p(n_2|n_1) = \frac{\int d\alpha\, p(\alpha, n_1, n_2)}{p(n_1)}$$

Observing n_1 lets you learn about α
Knowledge of α affects predictions for n_2 $\rightarrow$ dependence on n_1