CS340 Machine learning
Causality

“Structure and strength in causal induction”, Griffiths and Tenenbaum, Cognitive Psychology, 51:334-384, 2005

Does C cause E?

- Consider the case of a single cause and a single effect.
- The data can be summarized as a contingency table.

	Effect absent $E = 0$	Effect present $E = 1$
Cause absent $C = 0$	$N(E = 0, C = 0)$	$N(C = 0, E = 1)$
Cause present $C = 1$	$N(E = 0, C = 1)$	$N(C = 1, E = 1)$

Which chemical causes the effect?

- Chemical 1 is injected into 60 mice, of which 36 show an effect; c1 is not injected into another 60 mice, of which 30 show an effect

	Effect absent $E = 0$	Effect present $E = 1$
Cause absent $C = 0$	$30/60 = 0.5$	$30/60 = 0.5$
Cause present $C = 1$	$24/60 = 0.4$	$36/60 = 0.6$

- Chemical 2 is injected into 60 mice, of which 60 show an effect; c2 is not injected into another 60 mice, of which 54 show an effect

	Effect absent $E = 0$	Effect present $E = 1$
Cause absent $C = 0$	$6/60 = 0.1$	$54/60 = 0.9$
Cause present $C = 1$	$0/60 = 0$	$60/60 = 1$

Measures of causal strength

- A χ^2 score or mutual information yields a measure of statistical dependency between C and E, but is symmetric, so cannot tell us about causality.
- We will see how a simple Bayesian model can capture people's intuitive notions of causality better than rival approaches.
- In psychology, 2 measures of causal strength are popular:
- Delta P: $\Delta P = p(e=1|c=1) - p(e=1|c=0)$
- Causal power: $CP = \frac{\Delta P}{1 - p(e=1|c=0)}$
- Intuitively, CP discounts cases in which the effect is already present (so masking any possible effect of C)

Δ P vs CP

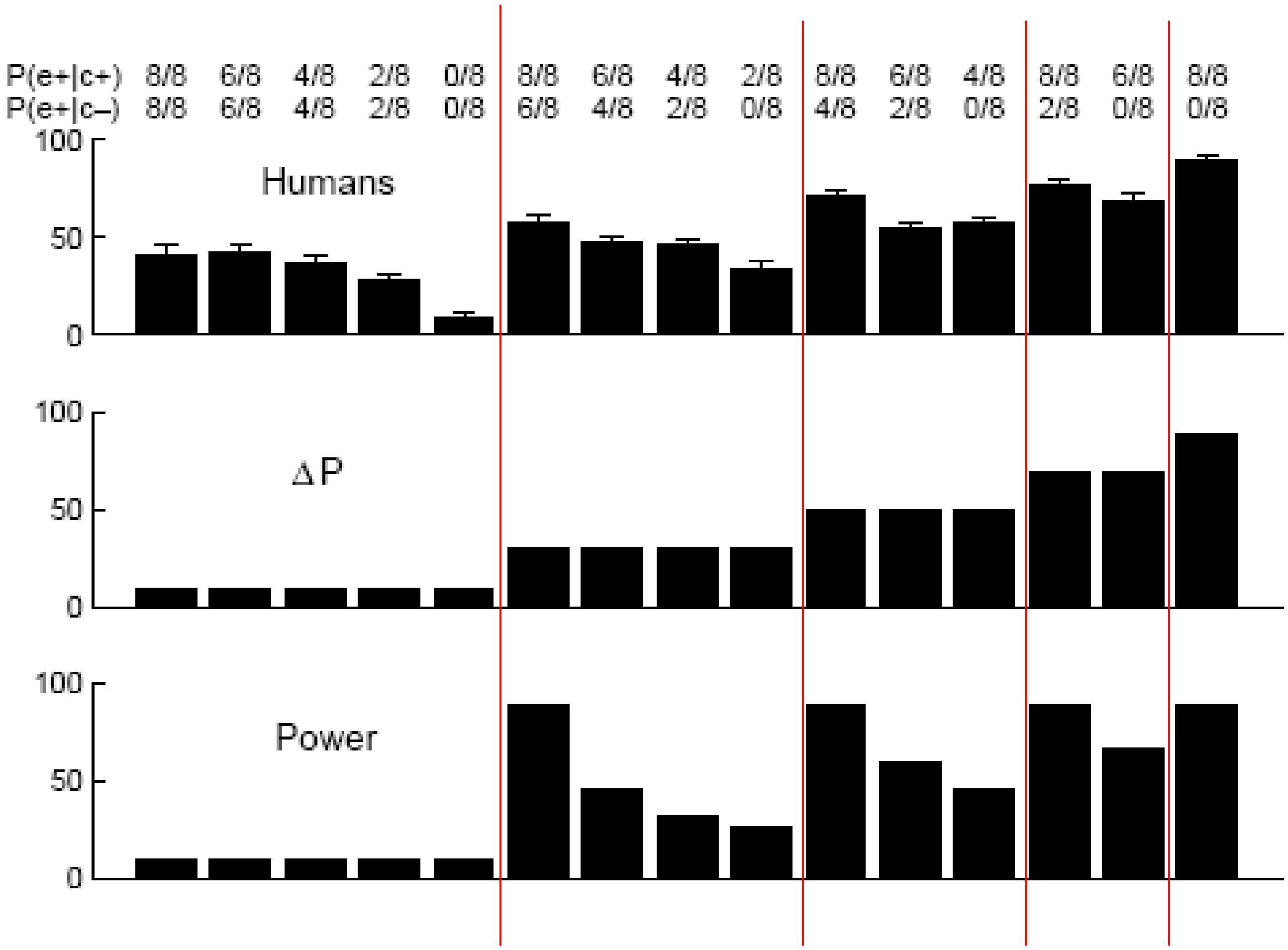
- Chemical 1: Δ P = 0.1, CP = 0.2

	Effect absent $E = 0$	Effect present $E = 1$
Cause absent $C = 0$	$30/60 = 0.5$	$30/60 = 0.5$
Cause present $C = 1$	$24/60 = 0.4$	$36/60 = 0.6$

- Chemical 2: Δ P = 0.1, CP = 1

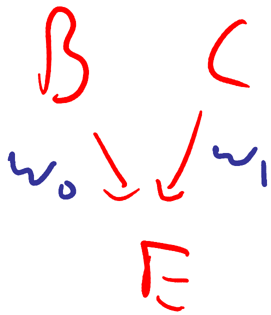
	Effect absent $E = 0$	Effect present $E = 1$
Cause absent $C = 0$	$6/60 = 0.1$	$54/60 = 0.9$
Cause present $C = 1$	$0/60 = 0$	$60/60 = 1$

Comparison with humans



Noisy-OR model

- Consider the case of a single cause and a single effect.

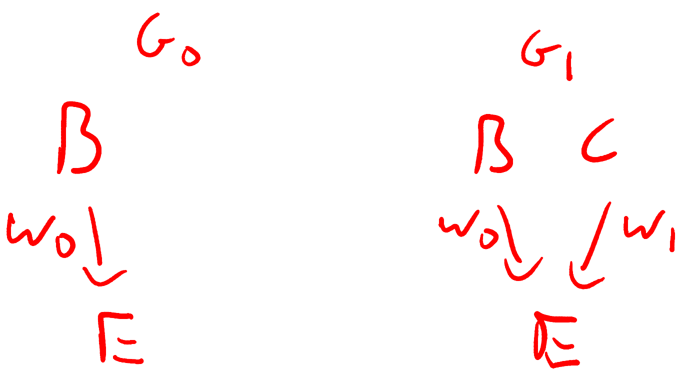


B	C	$P(E=0\|C,\mathbf{w})$	$P(E=1\|C,\mathbf{w})$
1	0	$1-w_0$	$1-(1-w_0)$
1	1	$(1-w_0)(1-w_1)$	$1-(1-w_0)(1-w_1)$

- Causal power is equivalent to the MLE for w_1.

Bayesian model selection

- "Causal Power" estimates the strength of the C->E edge.
- "Causal support" estimates the probability that there is any kind of C->E link, integrating out the strength



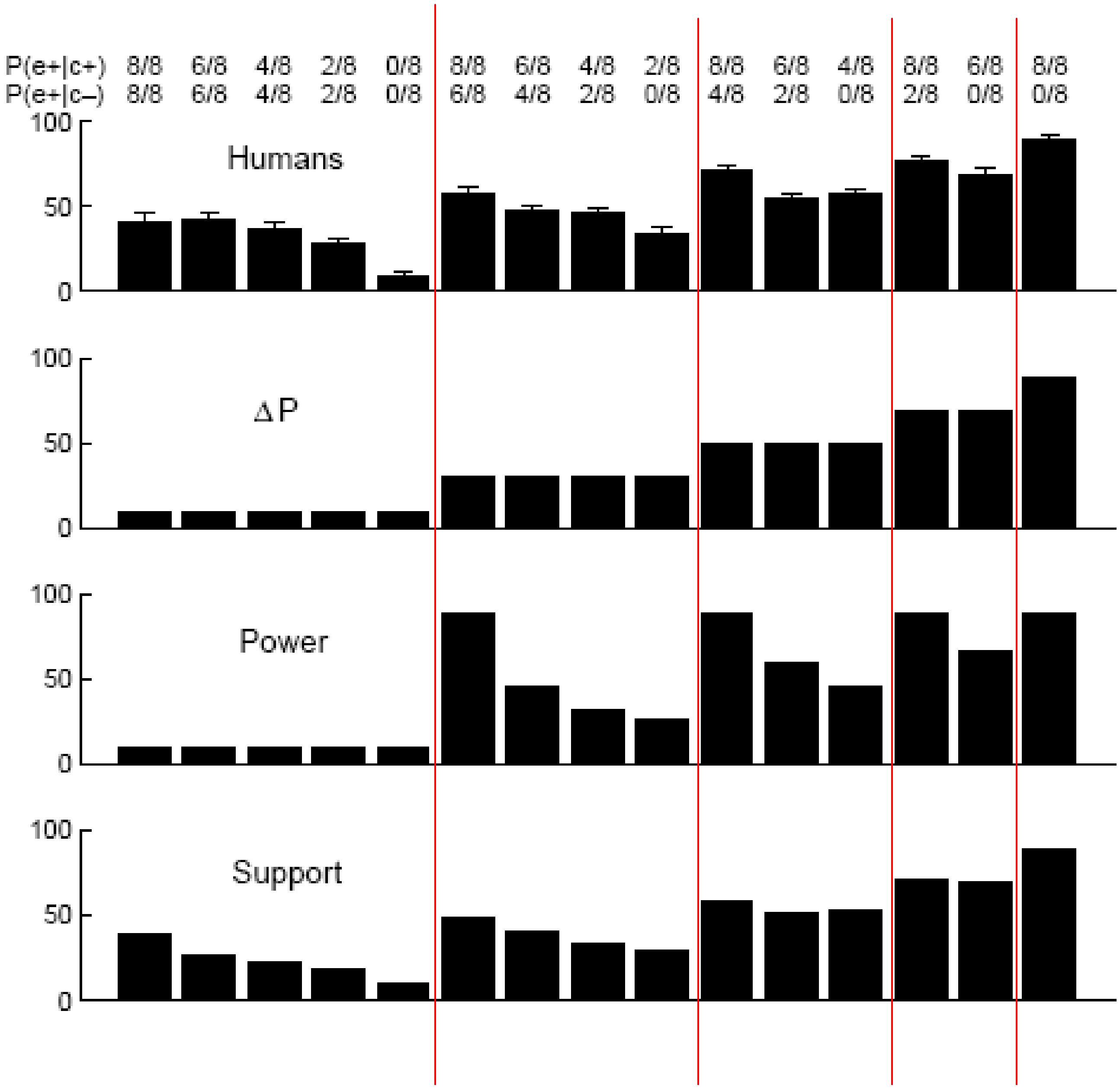
$$
\begin{aligned}
\text{causal support} &= p(G_1|D) \\
&= \frac{p(D|G_1)}{p(D|G_1) + p(D|G_2)} = \frac{1}{1 + BF(1,0)}
\end{aligned}
$$

Humans are Bayesian

Heuristic

MLE

Bayes



Computing p(D|G0)

- The CPD is

$$\begin{array}{cc} P(E=0|C,\mathbf{w}) & P(E=1|C,\mathbf{w}) \\ \hline 1-w_0 & 1-(1-w_0) \end{array}$$

- The evidence for G0 is

$$\begin{aligned} p(D|G_0) &= \int_0^1 w_0^{N(e=1)}(1-w_0)^{N(e=0)} Beta(w_0|a,b) dw_0 \\ &= \frac{B(a+N(e=1), b+N(e=0))}{B(a,b)} \end{aligned}$$

- For a uniform prior, we get

$$p(D|G_0) = B(N(e=1)+1, N(e=0)+1)$$

Computing p(D|G1)

- The CPD is

B	C	$P(E=0\|C,\mathbf{w})$	$P(E=1\|C,\mathbf{w})$
1	0	$1-w_0$	$1-(1-w_0)$
1	1	$(1-w_0)(1-w_1)$	$1-(1-w_0)(1-w_1)$

- There is no conjugate prior for this.
- So we will use Monte Carlo integration to compute

$$p(D|G_1) = \int\int p(D|w_0,w_1)p(w_0,w_1)dw_0dw_1 = E[p(D|w_0,w_1)]$$

Monte Carlo integration

- Suppose we want to evaluate the integral

$$E[h(X)] = I = \int h(x)p(x)dx$$

- In low dimensions, we can use numerical integration (eg. quadrature: in matlab, quad, dblquad, triplequad).
- In higher dimensions, a better approach is to sample S values x^s from p(x) and then use the law of large numbers

$$\hat{I} = \frac{1}{S}\sum_{s=1}^{S} h(x^s)$$

which has standard error

$$se = \sqrt{\frac{\hat{\sigma}^2}{S}}, \quad \hat{\sigma}^2 = \frac{1}{S-1}\sum_{s=1}^{S}(h(x_s) - \hat{I})^2$$

Definite integrals

- We can evaluate a definite integral by sampling uniformly within the range

$$
\begin{aligned}
I = \int_a^b h(x) &= (b-a) \int h(x) p(x) dx \\
p(x) &= U(b-a) = \frac{1}{(b-a)} I(a < x < b) \\
I &\approx \frac{1}{S} \sum_{s=1}^{S} h(x^s)
\end{aligned}
$$

- Thus the method can also be applied in non-statistical settings.

Estimating π

- Area of circle is

$$I = \int_{-r}^{r} \int_{-r}^{r} I(x^2 + y^2 \le r^2) dx dy$$

so π = I/r^2. Let $h(x, y) = I(x^2 + y^2 \le r^2)$

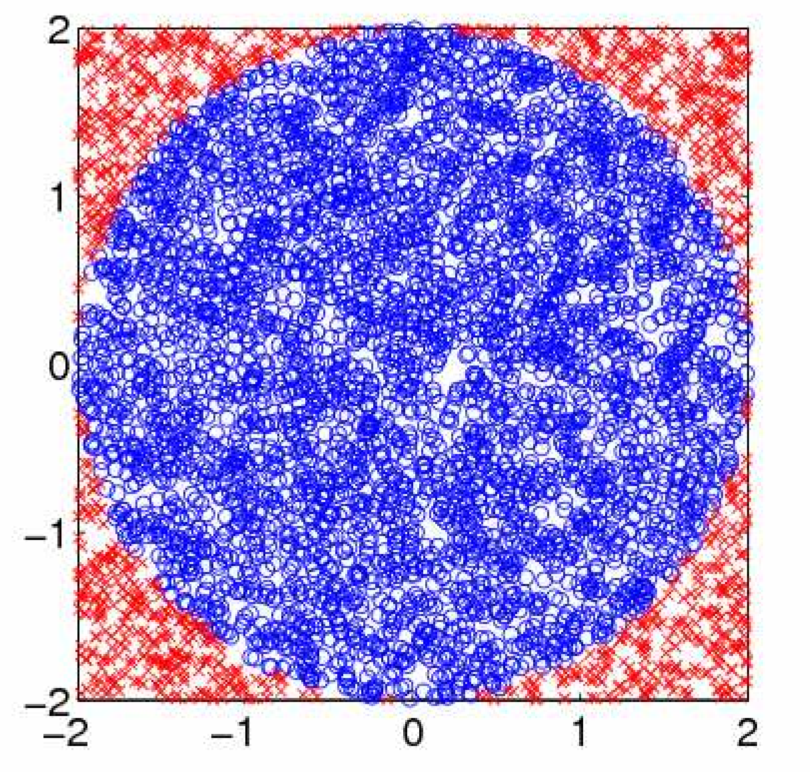
$$\begin{aligned} I &= (b_x - a_x)(b_y - a_y) \int \int h(x, y) p(x) p(y) dx dy \\ &= (2r)(2r) \int \int h(x, y) p(x) p(y) dx dy \\ &= 4r^2 \int \int h(x, y) p(x) p(y) dx dy \\ &\approx 4r^2 \frac{1}{S} \sum_s h(x^s, y^s) \end{aligned}$$

Estimating π

- Matlab

```
r=2;
S=5000;
xs = unifrnd(-r,r,S,1);
ys = unifrnd(-r,r,S,1);
rs = xs.^2 + ys.^2;
inside = (rs <= r^2);
samples = 4*(r^2)*inside;
Ihat = mean(samples)
piHat = Ihat/(r^2)
se = sqrt(var(samples)/S)
```

$$\hat{\pi} = 3.1416, \quad se = 0.09$$



Computing p(D|G1)

- If we use a uniform prior on w_0, w_1, we have

$$\begin{aligned} p(D|G_1) &= \int_0^1 \int_0^1 p(D|w_0, w_1) dw_0 dw_1 \\ &= \frac{1}{S} \sum_{s=1}^{S} p(D|w_0^s, w_1^s, G_1) \end{aligned}$$

where w_0^s, $w_1^s \sim U(0,1)$

Extensions

- It is easy to replace the noisy-OR model with others, e.g.,
- noisy AND-NOT: E will occur if B AND-NOT C. Use this if C is a preventive cause of E, rather than generative.

$$p(e = 1|c) = w_0(1 - w_1)^{I(c=1)}$$

- Use a Poisson model if C affects the rate of E.