Statistical Rethinking

Week 10:
Missing Data &
Other Opportunities

Richard McElreath

Missing data

- Missing values commonplace
- Usual approach: ***complete-case*** analysis
  - drop all cases with any missing values
- Discards a lot of information
- Alternatives
  - replace missing with mean of column: NEVER DO THIS
  - Multiple imputation
  - Bayesian imputation
  - others

im•pute | imˈpyo͞ot |

verb [with obj.]

represent (something, esp. something undesirable) as being done, caused, or possessed by someone; attribute: *the crimes* ***imputed to*** *Richard.*

- Finance assign (a value) to something by inference from the value of the products or processes to which it contributes: (as adj. **imputed**) *: recovering the initial outlay plus imputed interest.*
- Theology ascribe (righteousness, guilt, etc.) to someone by virtue of a similar quality in another: *Christ's righteousness has been imputed to us.*

Milk energy again

- `data(milk)`
- 12 missing values for neocortex
- Suppose values are ***Missing Completely At Random*** (**MCAR**)
  - MCAR: NAs sprinkled randomly
  - Distribution of observed values provides information
  - Can use to impute missing values
  - Must model the predictor

	kcal.per.g	mass	neocortex.perc
1	0.49	1.95	55.16
2	0.51	2.09	NA
3	0.46	2.51	NA
4	0.48	1.62	NA
5	0.60	2.19	NA
6	0.47	5.25	64.54
7	0.56	5.37	64.54
8	0.89	2.51	67.64
9	0.91	0.71	NA
10	0.92	0.68	68.85
11	0.80	0.12	58.85
12	0.46	0.47	61.69
13	0.71	0.32	60.32
14	0.71	0.60	NA
15	0.73	3.47	NA
16	0.68	1.55	69.97
17	0.72	7.08	NA
18	0.97	3.24	70.41
19	0.79	7.94	NA
20	0.84	12.30	73.40
21	0.48	7.59	NA
22	0.62	5.37	67.53
23	0.51	10.72	NA
24	0.54	35.48	71.26
25	0.49	79.43	72.60
26	0.53	97.72	NA
27	0.48	40.74	70.24
28	0.55	33.11	76.30
29	0.71	54.95	75.49

Milk energy MCAR

- Suppose your undergrad assistant lost those neocortex values
- Consider just neocortex variable:
  - Q: What is your best guess of each missing value?
  - A: Posterior distribution derived from remaining data

	neocortex.perc
1	55.16
2	NA
3	NA
4	NA
5	NA
6	64.54
7	64.54
8	67.64
9	NA
10	68.85
11	58.85
12	61.69
13	60.32
14	NA
15	NA
16	69.97
17	NA
18	70.41
19	NA
20	73.40
21	NA
22	67.53
23	NA
24	71.26
25	72.60
26	NA
27	70.24
28	76.30
29	75.49

Milk energy MCAR

- Place a unique parameter for each missing value
  - NC1 ... NC12
  - These are values to be imputed

	neocortex.perc
1	55.16
2	NC1
3	NC2
4	NC3
5	NC4
6	64.54
7	64.54
8	67.64
9	NC5
10	68.85
11	58.85
12	61.69
13	60.32
14	NC6
15	NC7
16	69.97
17	NC8
18	70.41
19	NC9
20	73.40
21	NC10
22	67.53
23	NC11
24	71.26
25	72.60
26	NC12
27	70.24
28	76.30
29	75.49

Milk energy MCAR: model

$$N = [0.55, N_2, N_3, N_4, 0.65, 0.65, \ldots, 0.76, 0.75].$$

$$\begin{aligned}
k_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta_N N_i + \beta_M \log M_i \\
N_i &\sim \text{Normal}(\nu, \sigma_N) \\
\alpha &\sim \text{Normal}(0, 10) \\
\beta_N &\sim \text{Normal}(0, 1) \\
\beta_M &\sim \text{Normal}(0, 1) \\
\sigma &\sim \text{Cauchy}(0, 1) \\
\nu &\sim \text{Normal}(0.5, 1) \\
\sigma_N &\sim \text{Cauchy}(0, 1)
\end{aligned}$$

Milk energy MCAR: model

$$N = [0.55, N_2, N_3, N_4, 0.65, 0.65, \ldots, 0.76, 0.75].$$

linear model using mix of observed and imputed values

$$\begin{aligned} k_i &\sim \text{Normal}(\mu_i, \sigma) \\ \mu_i &= \alpha + \beta_N N_i + \beta_M \log M_i \\ N_i &\sim \text{Normal}(\nu, \sigma_N) \end{aligned}$$

Milk energy MCAR: model

$$N = [0.55, N_2, N_3, N_4, 0.65, 0.65, \ldots, 0.76, 0.75].$$

$$\begin{aligned} k_i &\sim \text{Normal}(\mu_i, \sigma) \\ \mu_i &= \alpha + \beta_N N_i + \beta_M \log M_i \\ N_i &\sim \text{Normal}(\nu, \sigma_N) \end{aligned}$$

when obs, likelihood; when imputed, prior

mean neocortex (to be estimated)

std dev of neocortex (to be estimated)

Fitting

R code 14.7

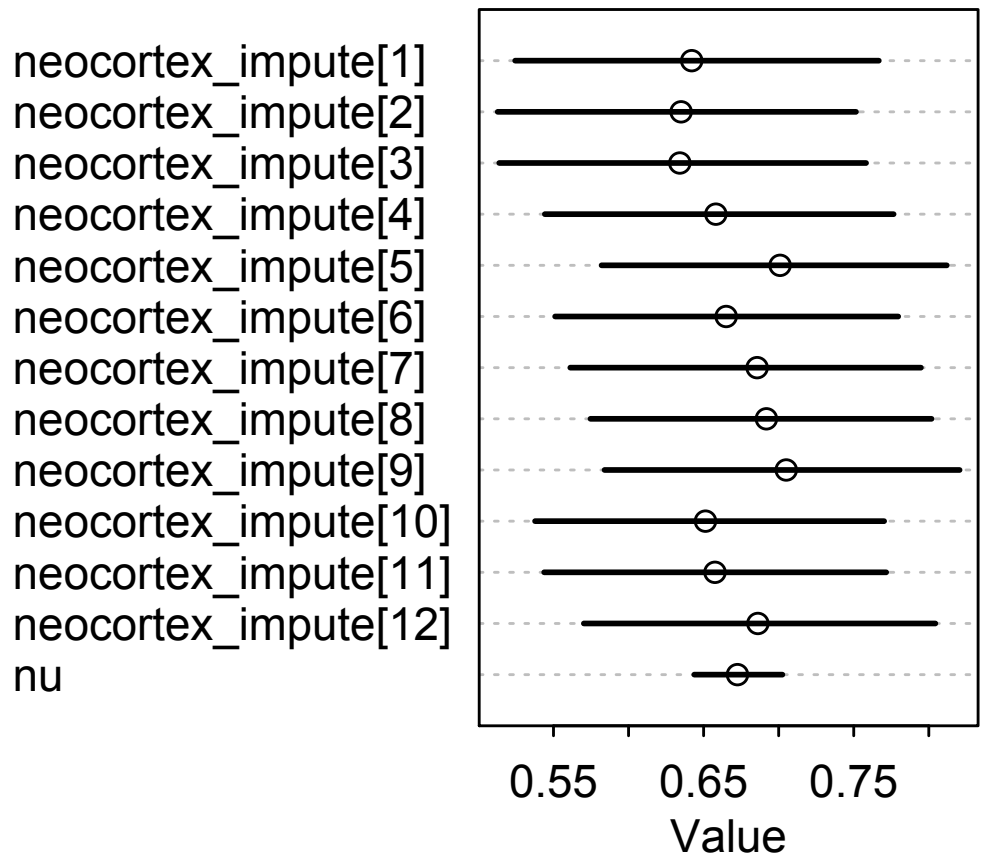
```
# prep data
data_list <- list(
    kcal = d$kcal.per.g,
    neocortex = d$neocortex.prop,
    logmass = d$logmass
)

# fit model
m14.3 <- map2stan(
    alist(
        kcal ~ dnorm(mu,sigma),
        mu <- a + bN*neocortex + bM*logmass,
        neocortex ~ dnorm(nu,sigma_N),
        a ~ dnorm(0,100),
        c(bN,bM) ~ dnorm(0,1),
        nu ~ dnorm(0.5,1),
        sigma_N ~ dcauchy(0,1),
        sigma ~ dcauchy(0,1)
    ) ,
    data=data_list , iter=1e4 , chains=2 )
```

Distribution on predictor signals map2stan to look for NAs.

If it finds any, replaces with parameters.

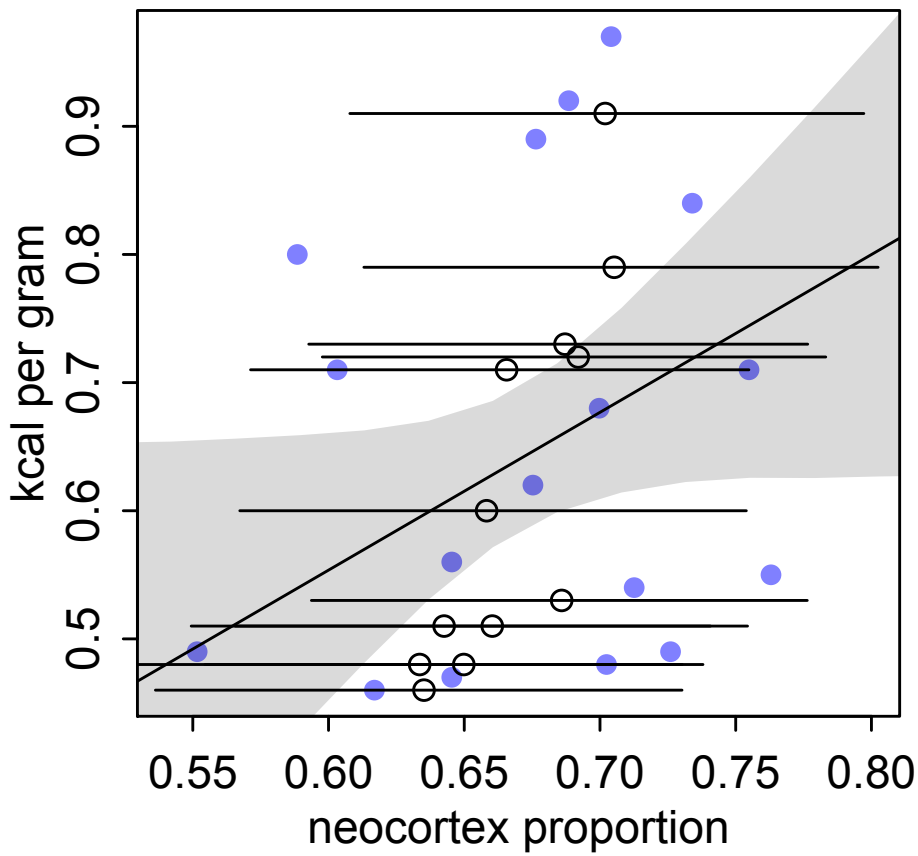
Results

- Reduced slopes compared to complete case analysis
  - bN: 2.8 => 1.2
  - bM: –0.10 => –0.05
- 12 imputed variables
  - wide confidence intervals
  - NOT same as prior
  - Why differ?



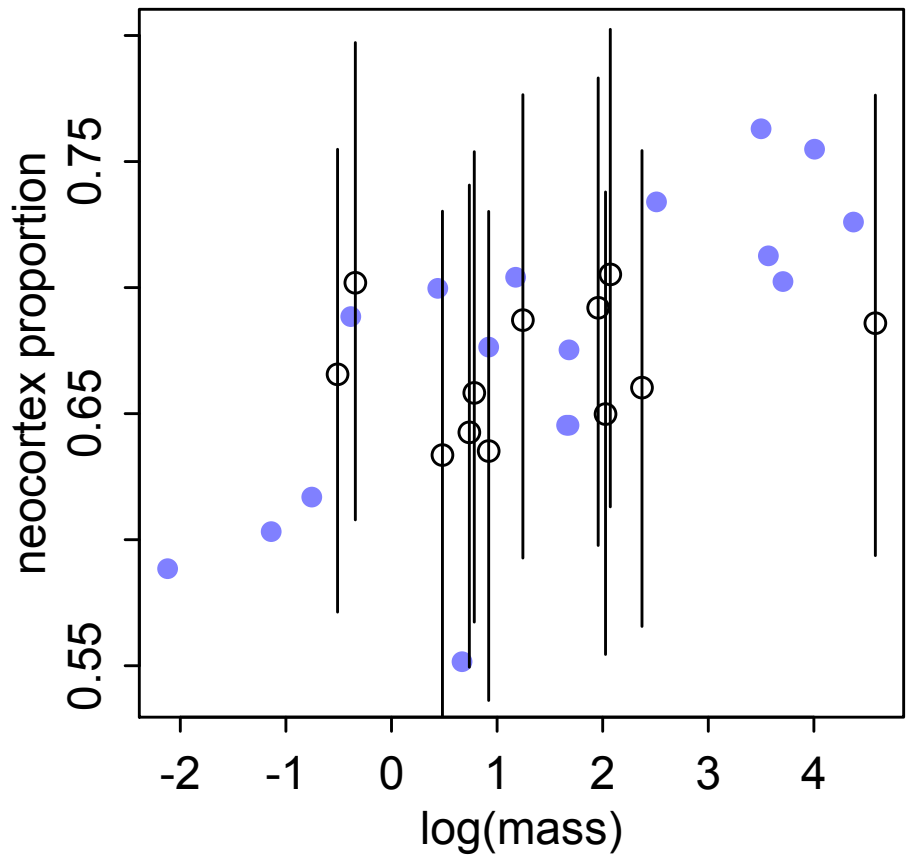
Results

- Imputed values weakly track regression
  - observed neocortex associated with milk energy
  - imputed values weakly associated with paired milk energy
  - this is logical, a consequence of the model definition



Results

- Observed neocortex positively associated with observed body mass
- Imputed neocortex NOT associated with observed body mass
- Can do better
  - Imputation model should use body mass (at least)



Milk energy MCAR: Model 2

- Naive imputation model:

$$N_i \sim \text{Normal}(\nu, \sigma_N)$$

- Slightly less naive imputation model:

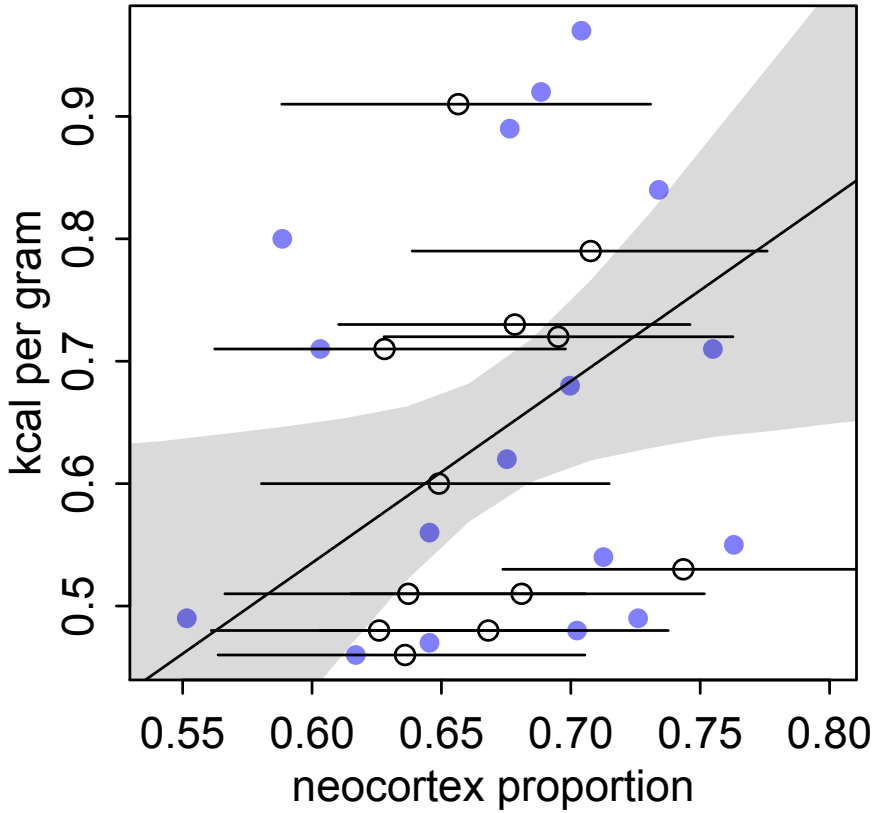
$$N_i \sim \text{Normal}(\nu_i, \sigma_N)$$
$$\nu_i = \alpha_N + \gamma_M \log M_i$$

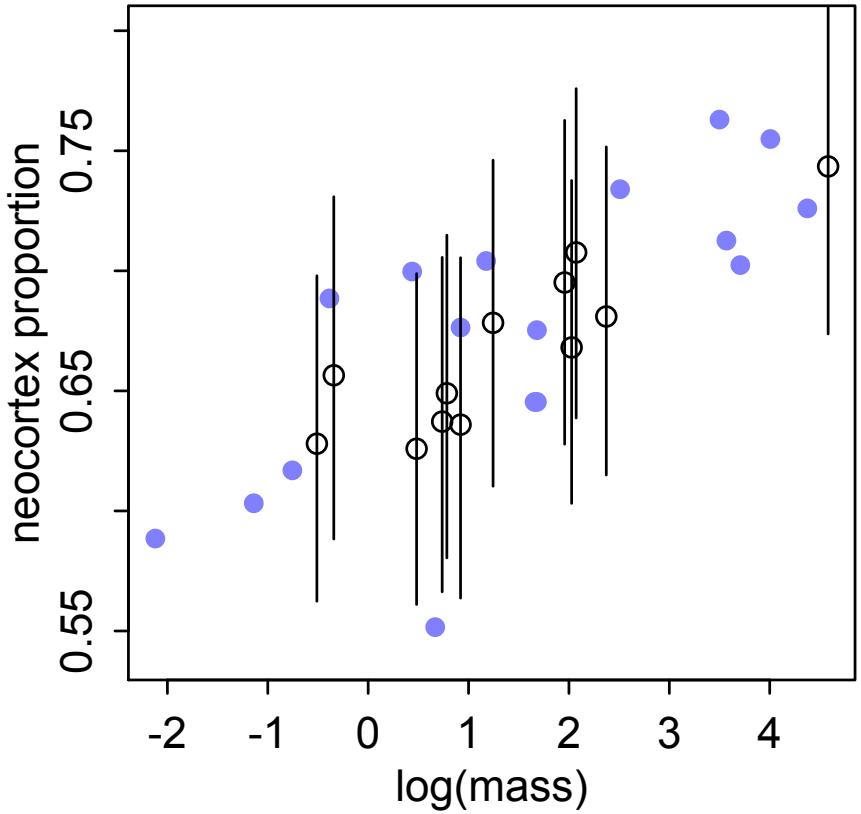
ordinary slope *body mass*

Milk energy MCAR: Model 2

- Slopes steeper now
- Confidence intervals on imputed values tighter
- Information used to update imputed values:
  - neocortex association with milk energy
  - neocortex association with log body mass

	Mean	StdDev
neocortex_impute[1]	0.64	0.04
neocortex_impute[2]	0.64	0.04
neocortex_impute[3]	0.63	0.04
neocortex_impute[4]	0.65	0.04
neocortex_impute[5]	0.66	0.04
neocortex_impute[6]	0.63	0.04
neocortex_impute[7]	0.68	0.04
neocortex_impute[8]	0.70	0.04
neocortex_impute[9]	0.71	0.04
neocortex_impute[10]	0.67	0.04
neocortex_impute[11]	0.68	0.04
neocortex_impute[12]	0.74	0.04
a	-0.29	0.44
bN	1.53	0.69
bM	-0.07	0.02
gM	0.02	0.01
a_N	0.64	0.01
sigma_N	0.04	0.01
sigma	0.14	0.02





- Range of imputed values still quite wide
- Bayes is not magic, just logic
- Imputation just logical consequence of defining full model for (1) outcome and (2) predictors
- Other methods illogical: Prevent feedback from regression to imputed values

The Golem of Prague

"Even the most perfect of Golem, risen to life to protect us, can easily change into a destructive force. Therefore let us treat carefully that which is strong, just as we bow kindly and patiently to that which is weak."

Rabbi Judah Loew ben Bezalel (1512–1609)

From *Breath of Bones: A Tale of the Golem*

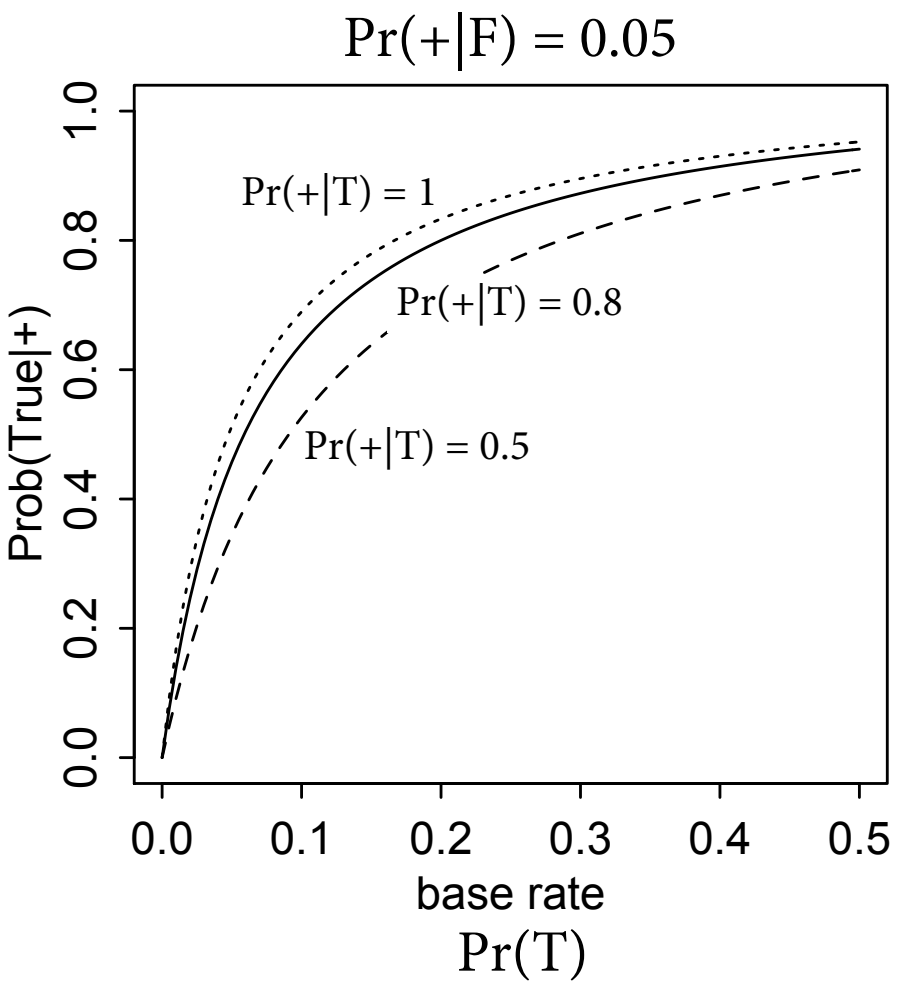
Stats not substitute for science

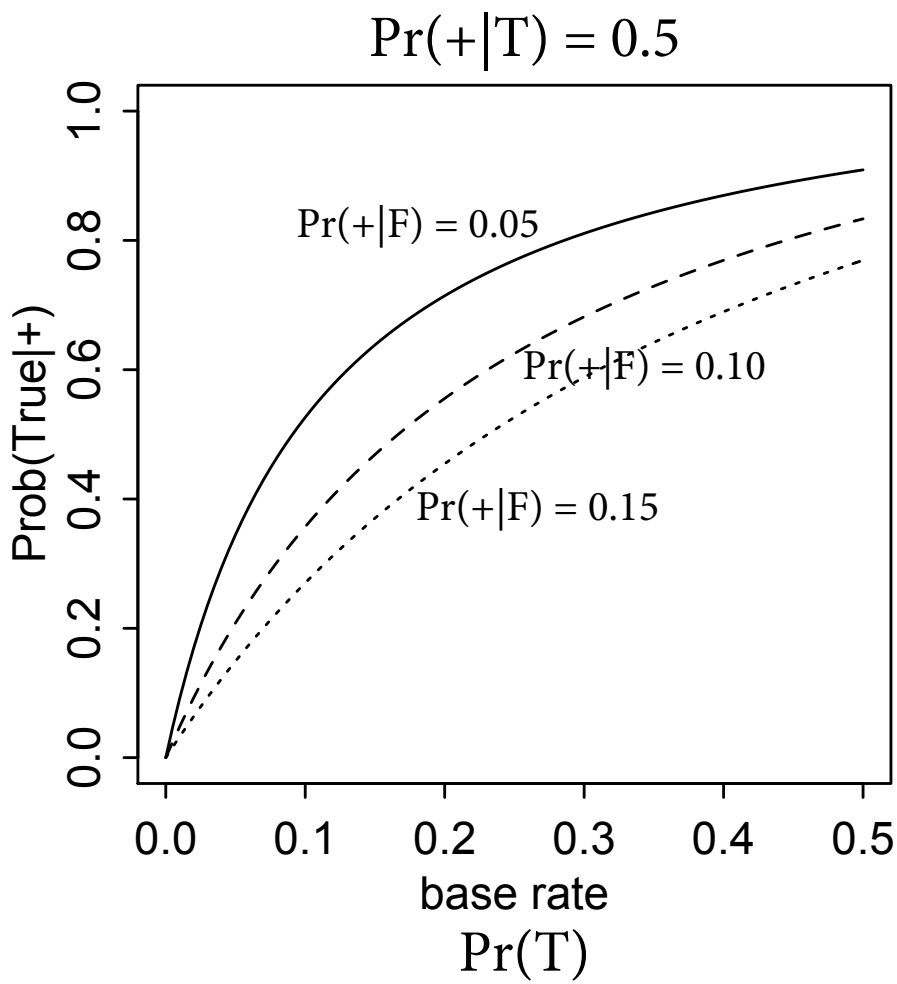
- Assume
  - Probability false positive finding is 5%
  - Probability true positive finding is 80% (power)
  - Conditional on positive finding, what is probability finding is true?

Stats not substitute for science

- Assume
  - Probability false positive finding is 5%
  - Probability true positive finding is 80% (power)
  - Conditional on positive finding, what is probability finding is true?

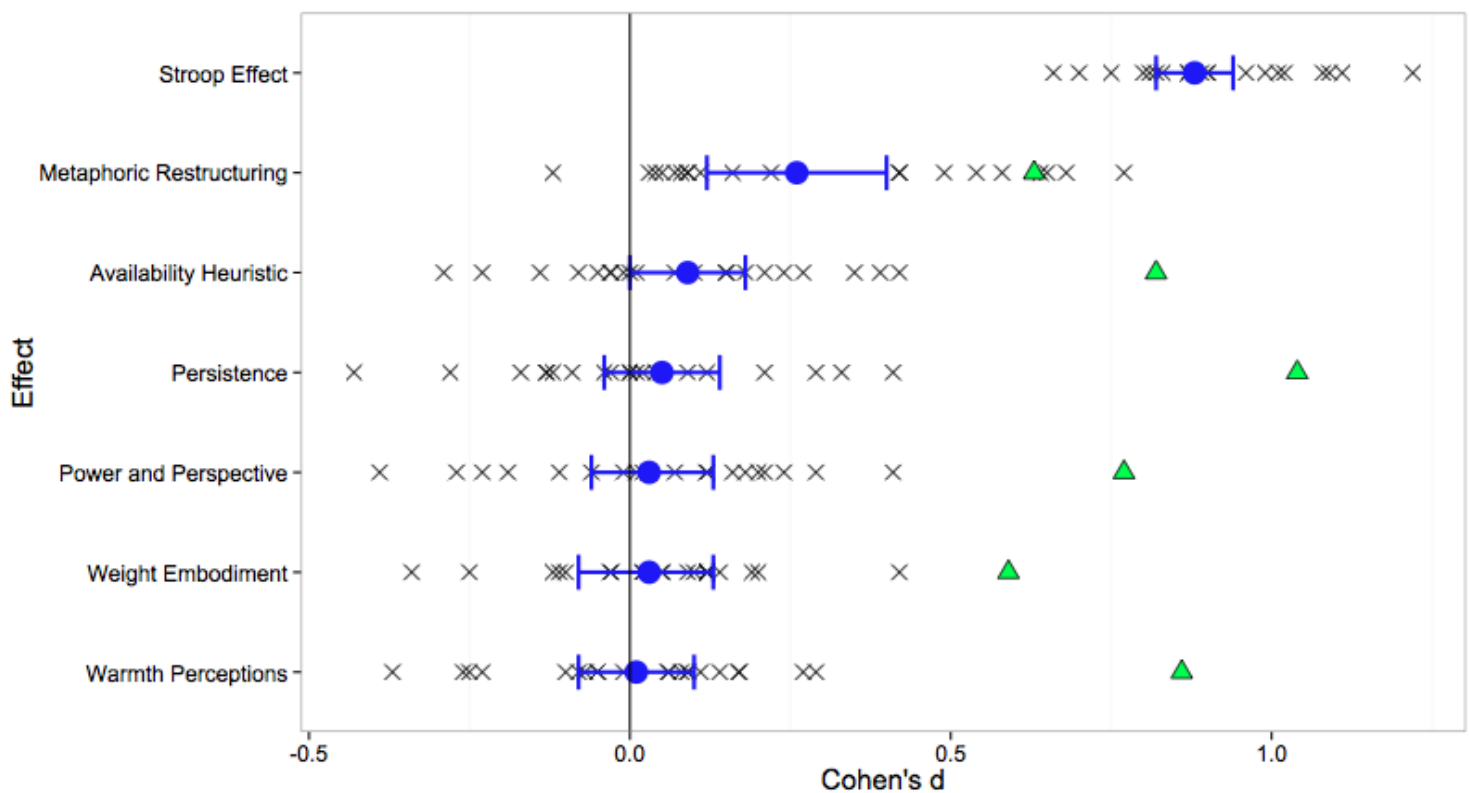
$$\begin{aligned}\Pr(T|+) &= \frac{\Pr(+|T)\Pr(T)}{\Pr(+)} \\ &= \frac{\Pr(+|T)\Pr(T)}{\Pr(+|T)\Pr(T) + \Pr(+|F)\Pr(F)}\end{aligned}$$

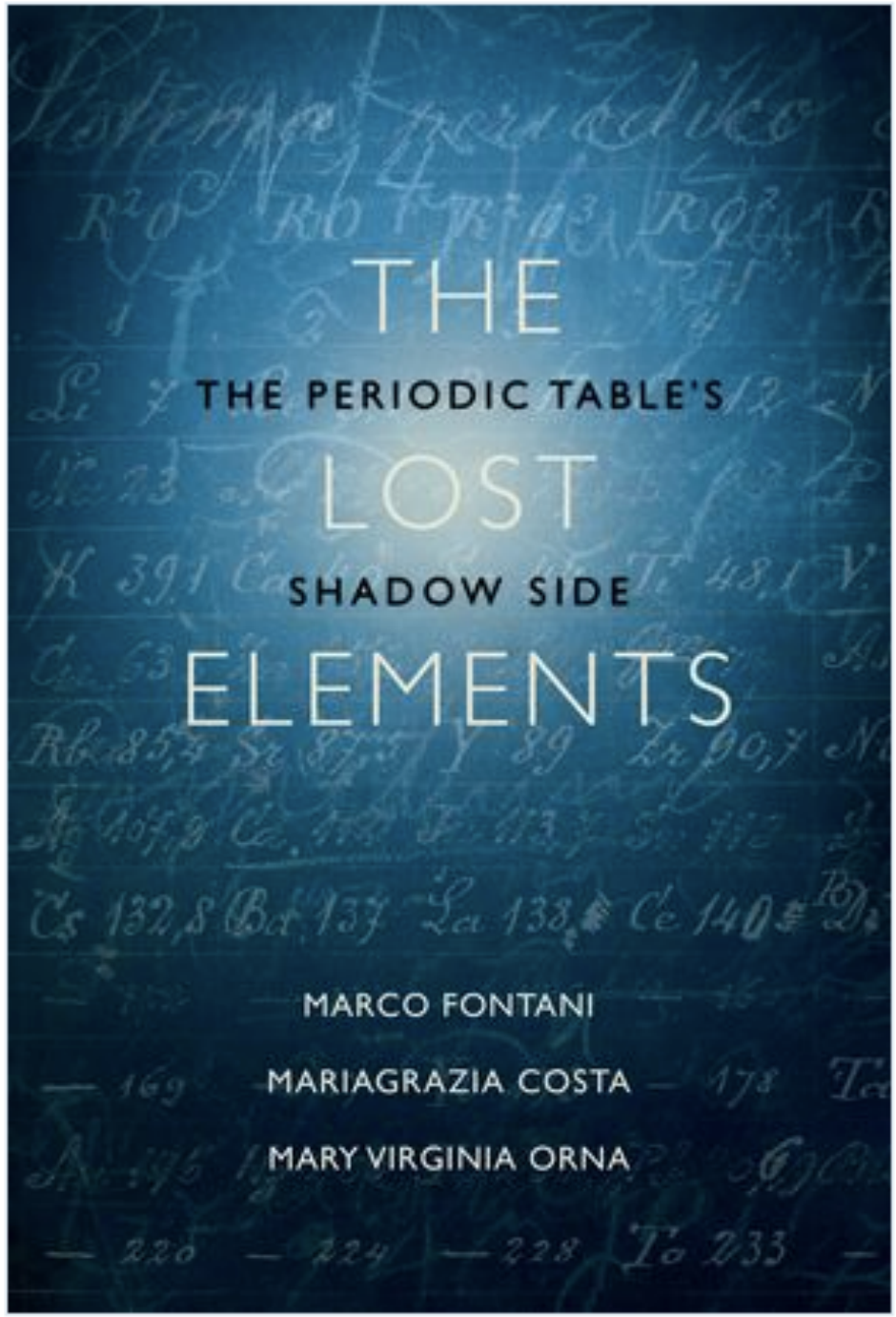

Pr(+|F) = 0.05
Pr(+|T) = 1
Pr(+|T) = 0.8
Pr(+|T) = 0.5
Prob(True|+)
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.1
0.2
0.3
0.4
0.5
base rate
Pr(T)


Pr(+|T) = 0.5
Pr(+|F) = 0.05
Pr(+|F) = 0.10
Pr(+|F) = 0.15
Prob(True|+)
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.1
0.2
0.3
0.4
0.5
base rate
Pr(T)

What's the base rate?

- No one knows the base rate
  - except for GWAS: $Pr(T) < 10^{-5}$
- Frighteningly low, judging by replication results



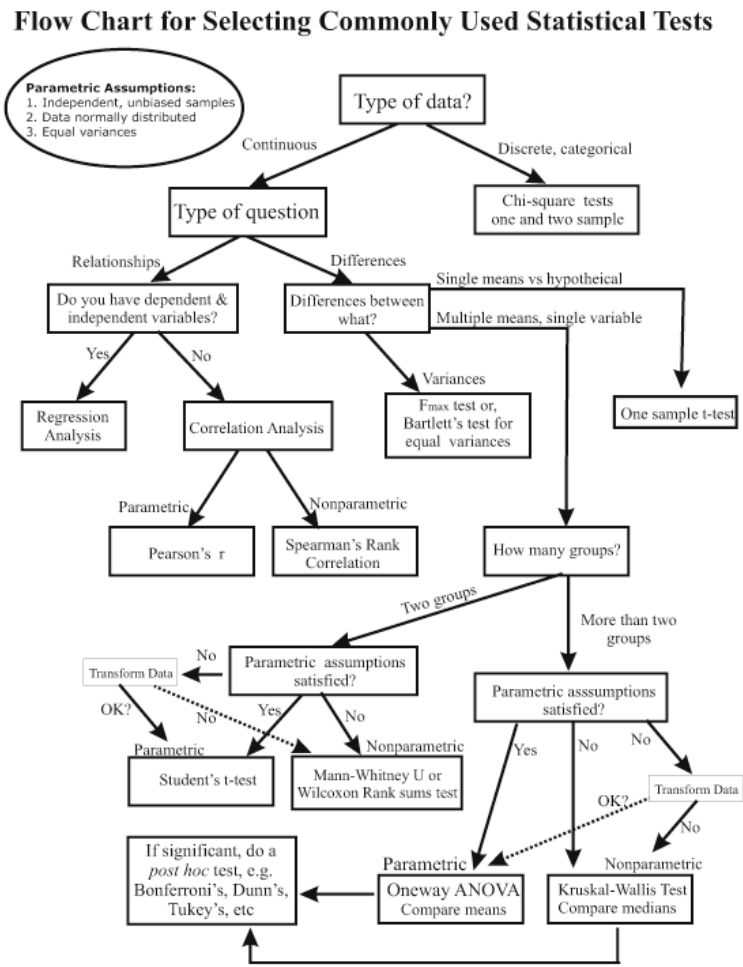


Replication always necessary

Communication always suspect

Recipes and mantras

- Anxiety => statistical compulsive hand washing
- Made worse by field of Statistics being autonomous
- *Objective*: Everyone does it the same way => safe
- *Subjective*: Expertise matters
- But if we must have recipes and mantras...



Recipes and mantras

- Recipe for Bayesian data analysis
  - Define model(s)
  - Fit model(s)
  - Check fit(s)
  - Critique model(s)
  - Repeat
- Details always depend upon context, purpose

Recipes and mantras

- Recipe for choosing likelihood functions
  - What constraints do you know, before you see the data?
  - What aspects of the data do you care about?
  - What can you actually calculate and understand?
  - Nothing forces you to choose only one
- Recipe for choosing priors
  - Guard against overfitting (flat never best)
  - Meaningful parameter: What do you already know? Exploit maximum entropy again.
  - No ideas? Try different priors and see how sensitive

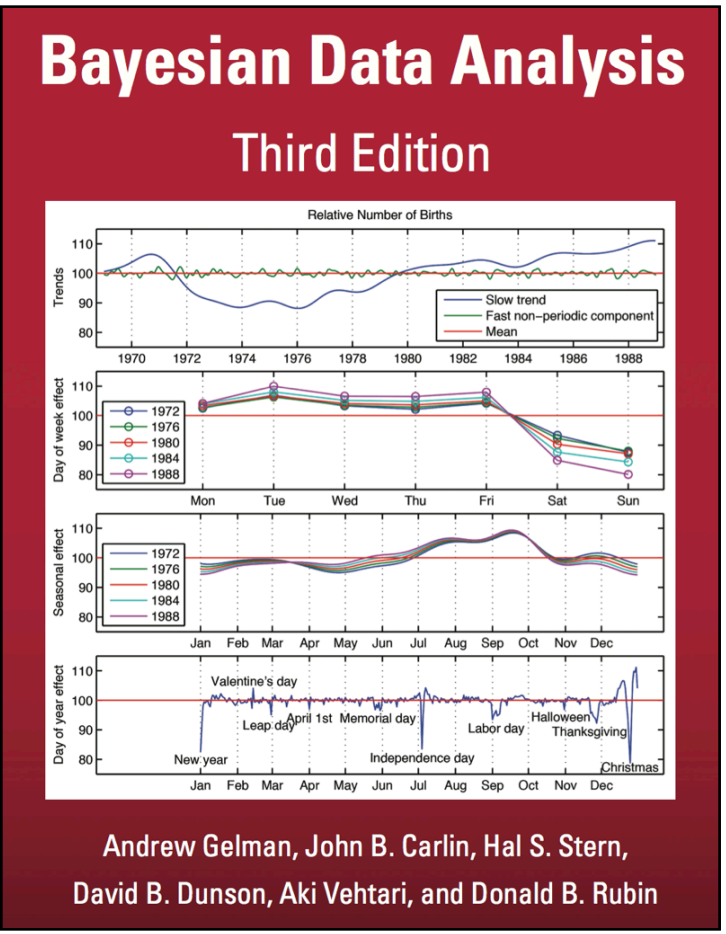
Recipes and mantras

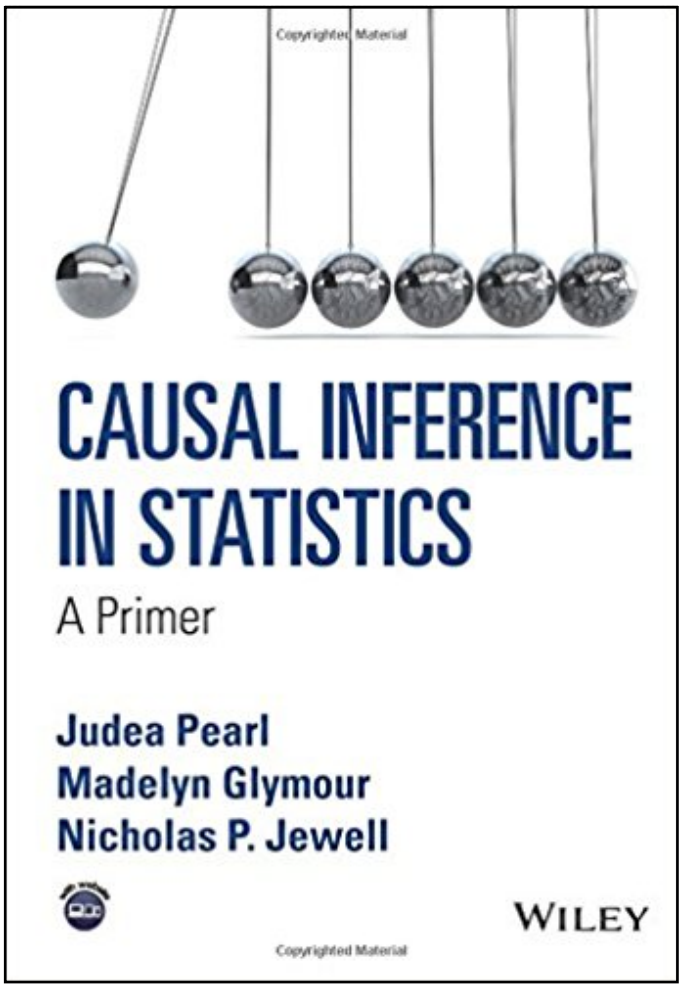
- Mantras:
  - Assume an effect and estimate it
  - Embrace and propagate uncertainty
  - Fitting is easy; prediction is hard
  - There is no right, only less wrong
  - Math is not real; only then can it be real

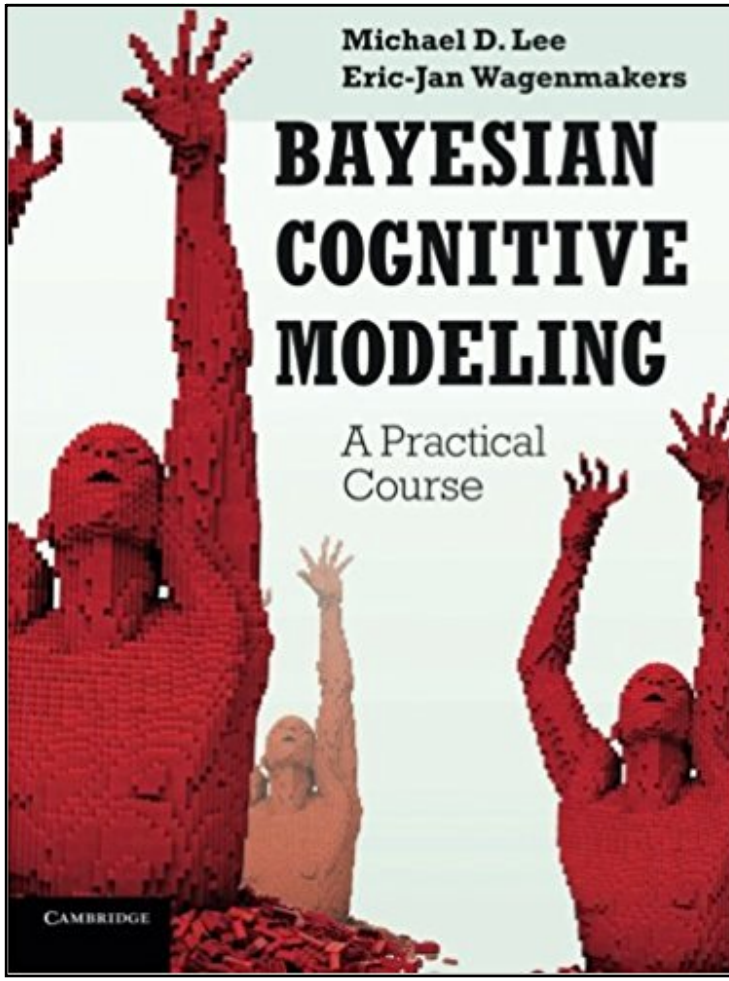
Writing Statistics

- Don't just describe; justify
- Describe all the models you tried
- Data snooping is okay WHEN HONEST
- Don't say "no effect of X"
- Do say "conditional on model, small/large association between X and Y"
- Estimates provide plausible values for these data and these models
- Don't rely on tables. Plot, plot, plot.
- Don't rely on parameters. Predictions!
- Document the analysis with a script.
- Publish/share the data. *Nullius in verba*
- Cite: Gelman et al 2014 (*Bayesian Data Analysis* 3rd ed.)

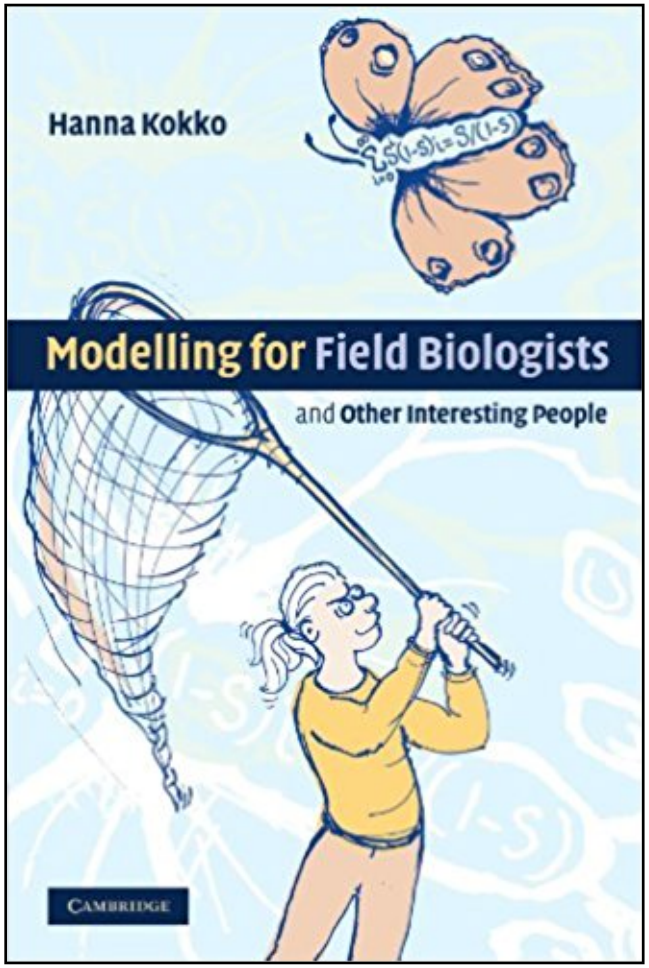
Where to now?

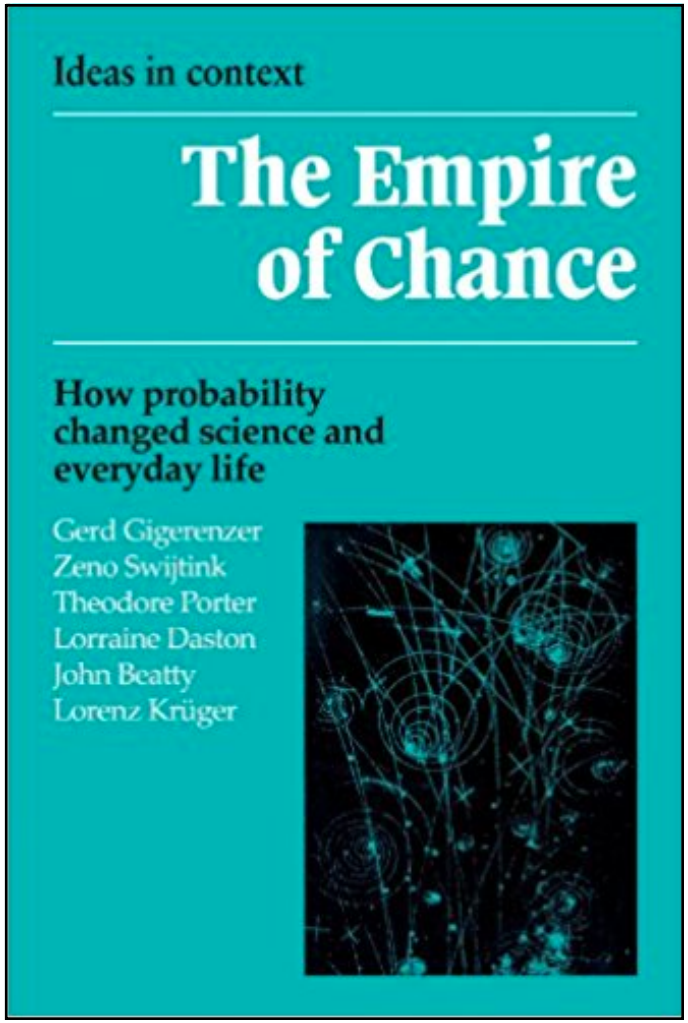

Bayesian Data Analysis
Third Edition
Relative Number of Births
Trends
Slow trend
Fast non−periodic component
Mean
Day of week effect
Seasonal effect
Day of year effect
Valentine's day
Leap day
April 1st
Memorial day
Labor day
Halloween
Thanksgiving
New year
Independence day
Christmas
Andrew Gelman, John B. Carlin, Hal S. Stern,
David B. Dunson, Aki Vehtari, and Donald B. Rubin

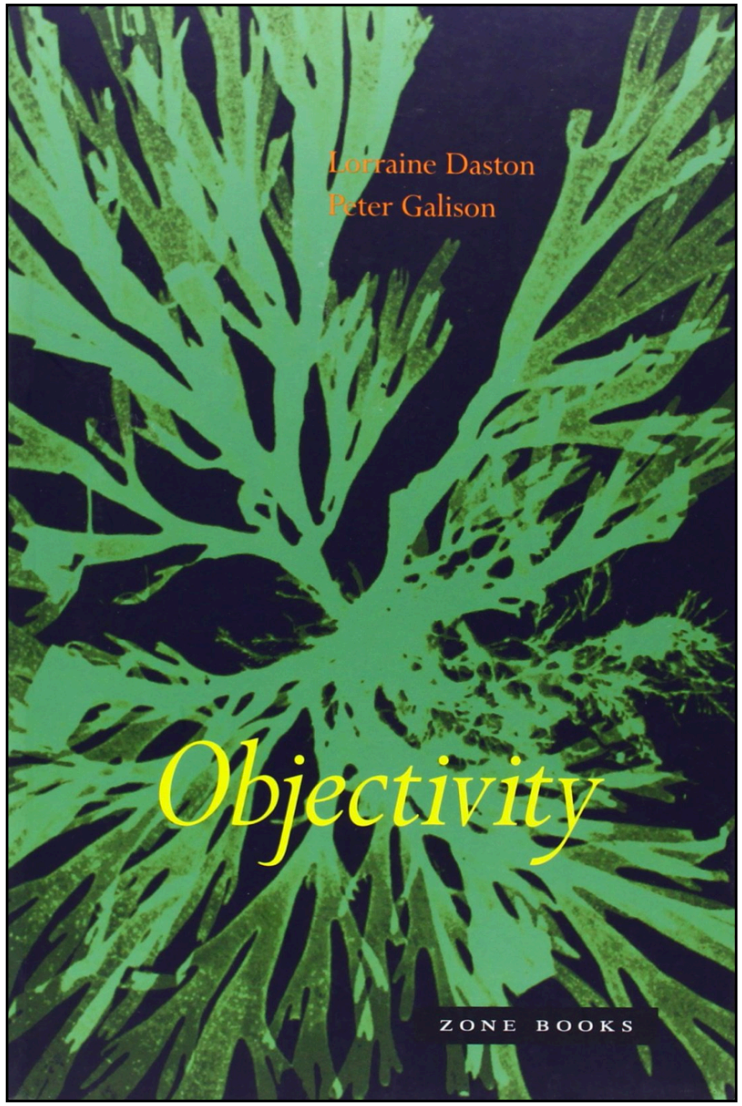

Copyrighted Material
CAUSAL INFERENCE
IN STATISTICS
A Primer
Judea Pearl
Madelyn Glymour
Nicholas P. Jewell
WILEY
Copyrighted Material


Michael D. Lee
Eric-Jan Wagenmakers
BAYESIAN
COGNITIVE
MODELING
A Practical
Course
CAMBRIDGE

Where to now?


Hanna Kokko
Modelling for Field Biologists
and Other Interesting People
CAMBRIDGE


Ideas in context
The Empire of Chance
How probability changed science and everyday life
Gerd Gigerenzer
Zeno Swijtink
Theodore Porter
Lorraine Daston
John Beatty
Lorenz Krüger


Lorraine Daston
Peter Galison
Objectivity
ZONE BOOKS