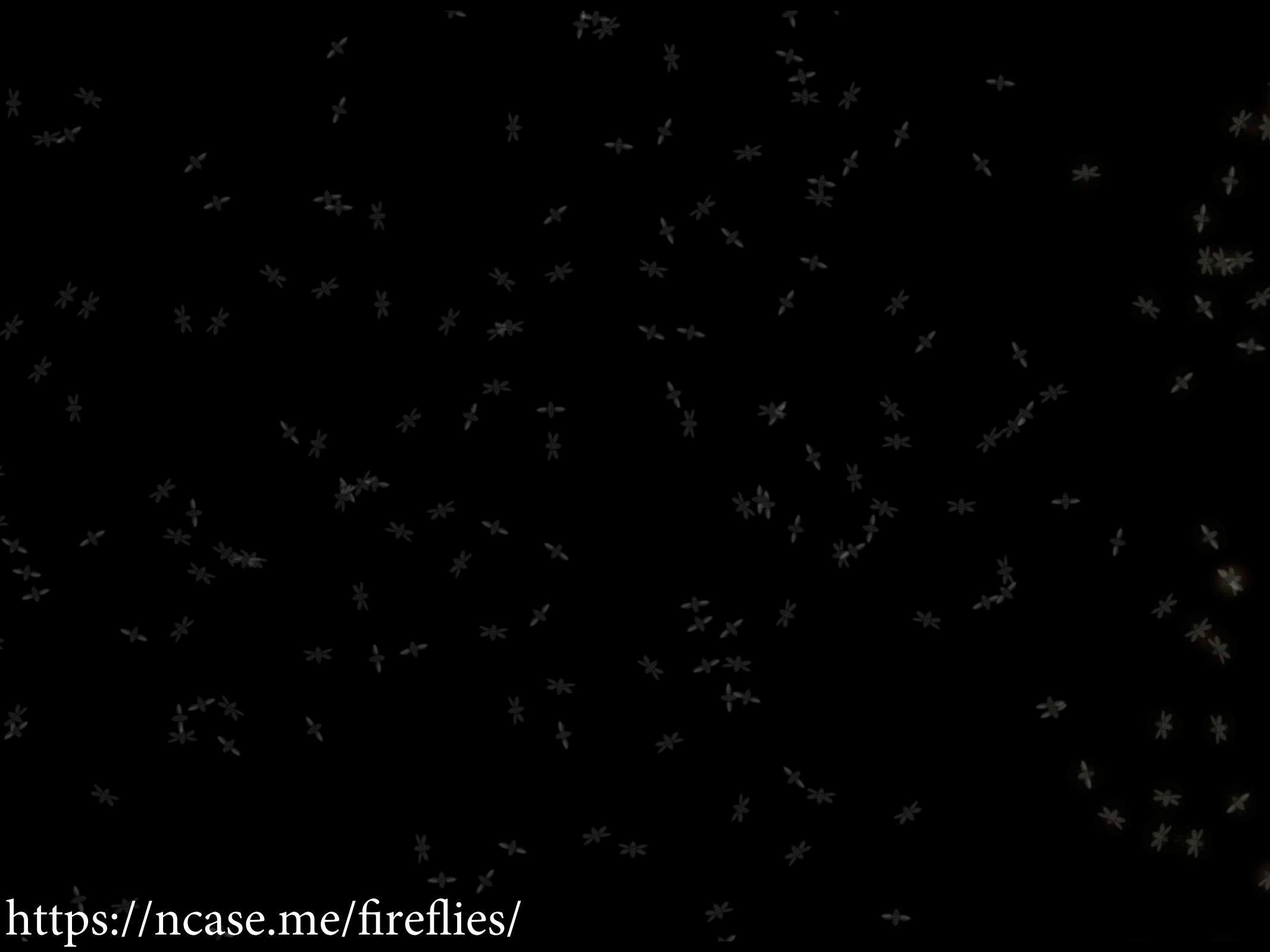


Statistical Rethinking

Winter 2019

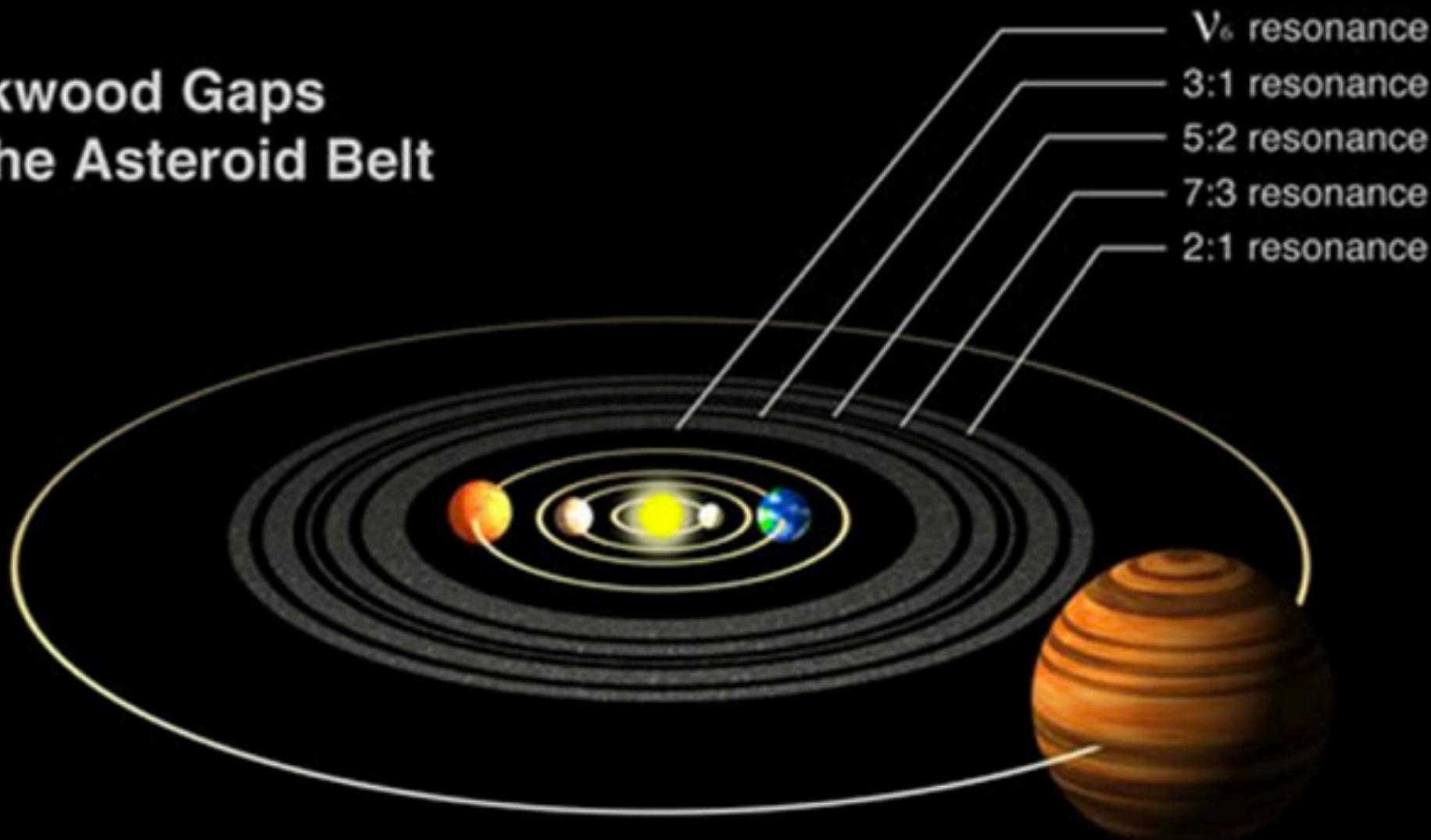
Lecture 12 / Week 6

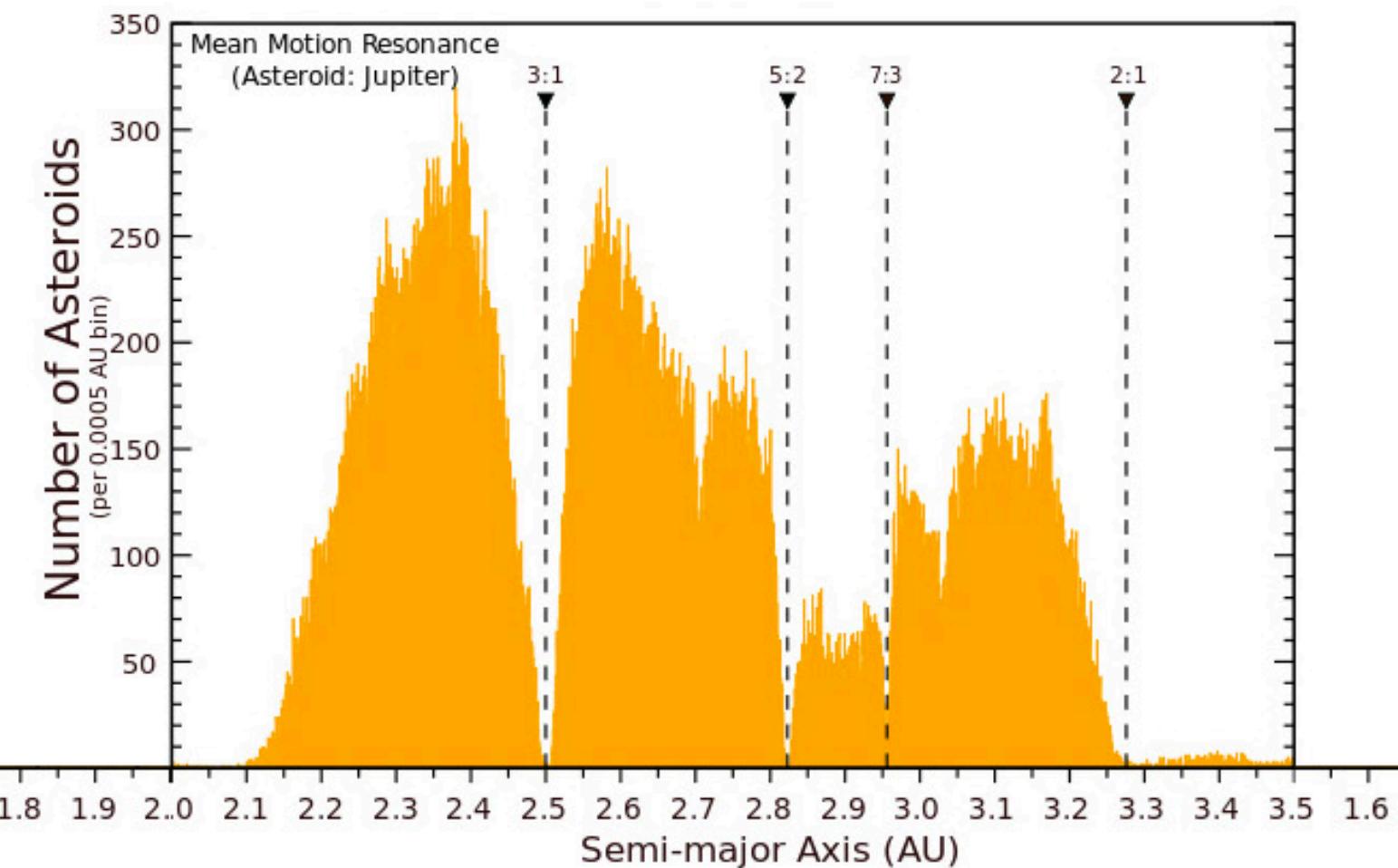
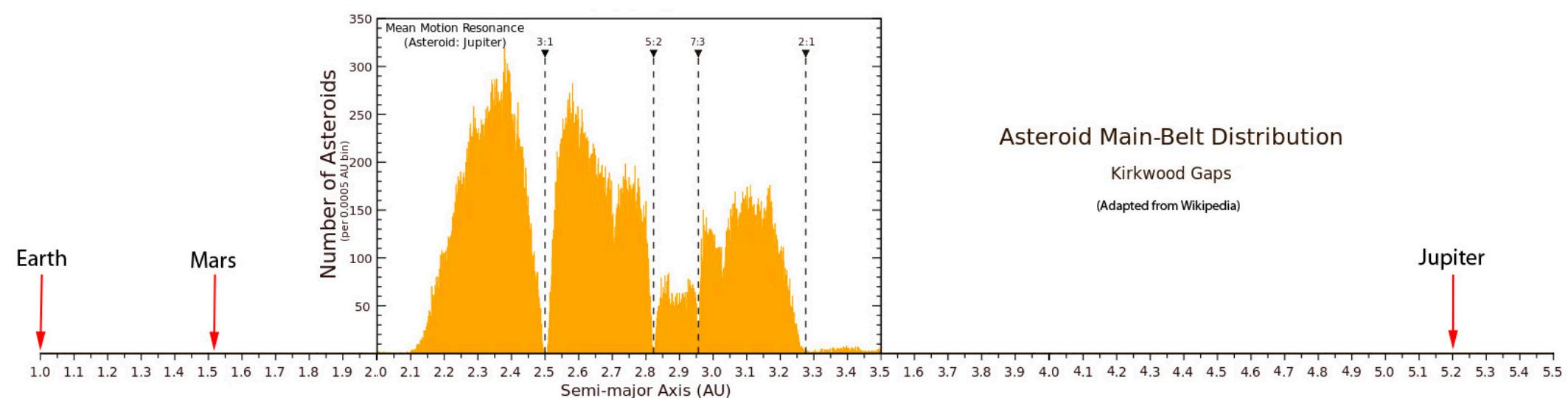
God Spiked
The Integers



<https://ncase.me/fireflies/>

Kirkwood Gaps in the Asteroid Belt





Logit link priors

- What about treatments?

R code
11.7

```
m11.2 <- quap(  
  alist(  
    pulled_left ~ dbinom( 1 , p ) ,  
    logit(p) <- a + b[treatment] ,  
    a ~ dnorm( 0 , 1.5 ) ,  
    b[treatment] ~ dnorm( 0 , 10 )  
  ) , data=d )
```

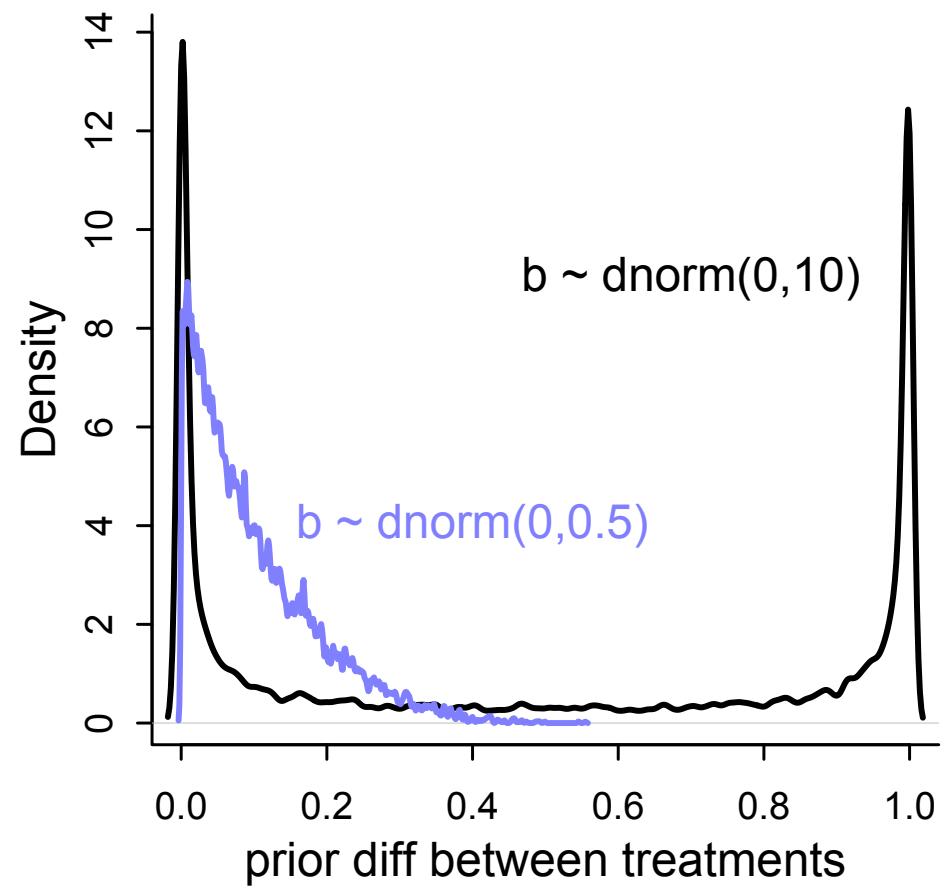
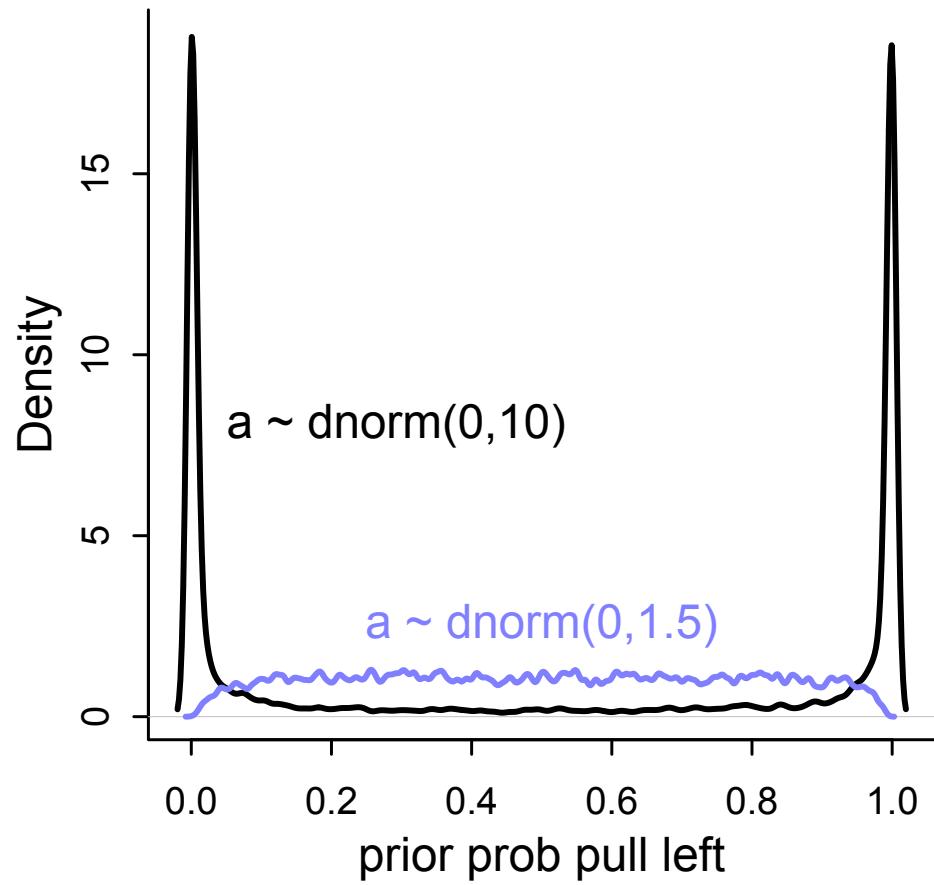


Figure 11.3

```

# particles in 11-dimensional space
m11.4 <- ulam(
  alist(
    pulled_left ~ dbinom( 1 , p ) ,
    logit(p) <- a[actor] + b[treatment] ,
    a[actor] ~ dnorm( 0 , 1.5 ) ,
    b[treatment] ~ dnorm( 0 , 0.5 )
  ) ,
  data=dat_list , chains=4 )
precis( m11.4 , depth=2 )

```

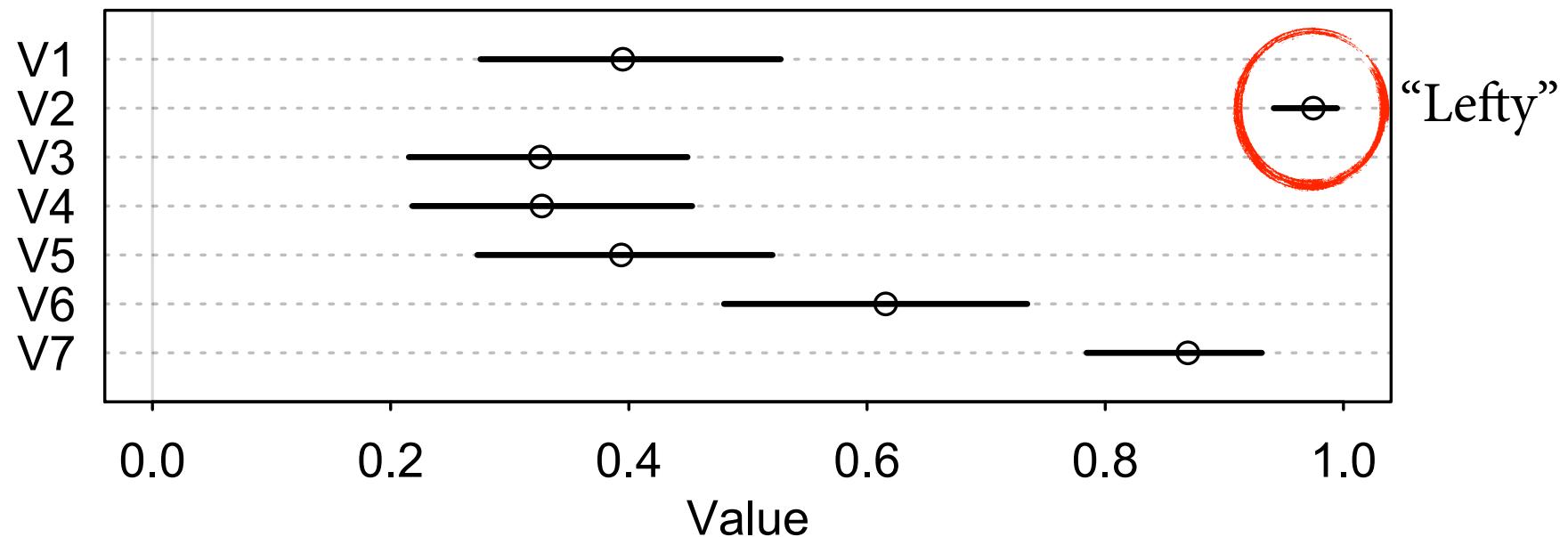
		mean	sd	5.5%	94.5%	n_eff	Rhat	
Chimpanzees	a[1]	-0.44	0.34	-0.97	0.11	736	1	
	a[2]	3.90	0.77	2.78	5.22	921	1	
	a[3]	-0.75	0.34	-1.29	-0.20	886	1	
	a[4]	-0.74	0.34	-1.28	-0.19	770	1	
	a[5]	-0.44	0.34	-0.98	0.08	832	1	
	a[6]	0.48	0.34	-0.08	1.02	854	1	
	a[7]	1.96	0.41	1.29	2.61	847	1	
Treatments	RN	b[1]	-0.05	0.29	-0.51	0.42	781	1
	LN	b[2]	0.48	0.29	0.03	0.94	657	1
	RP	b[3]	-0.39	0.28	-0.83	0.07	669	1
	LP	b[4]	0.36	0.29	-0.11	0.81	732	1

Individual differences

R code

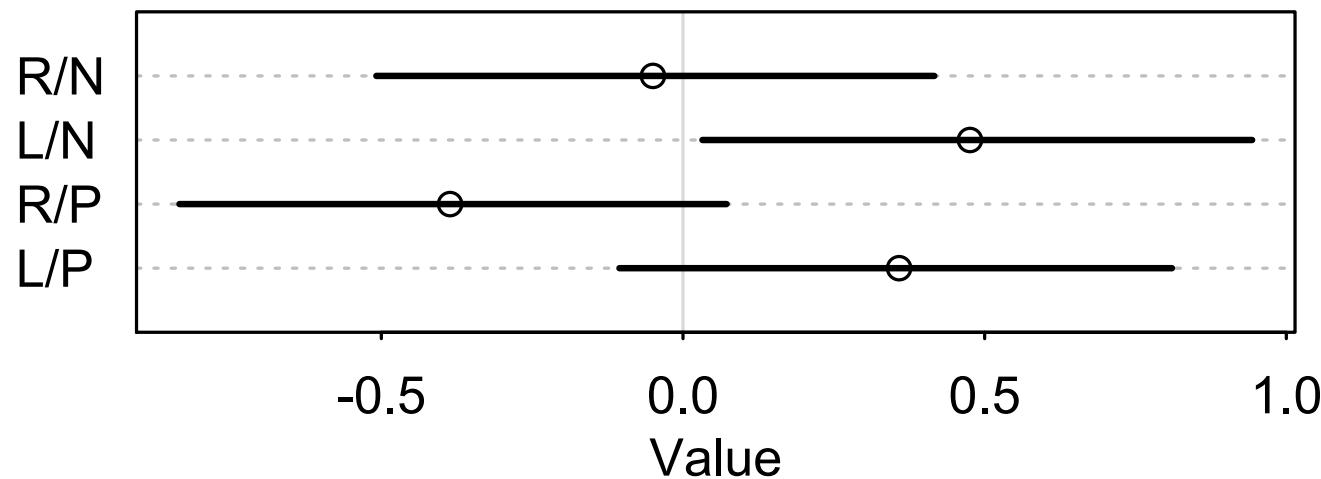
11.11

```
post <- extract.samples(m11.4)
p_left <- inv_logit( post$a )
plot( precis( as.data.frame(p_left) ) , xlim=c(0,1) )
```



Treatments

```
labs <- c("R/N","L/N","R/P","L/P")
plot( precis( m11.4 , depth=2 , pars="b" ) , labels=labs )
```



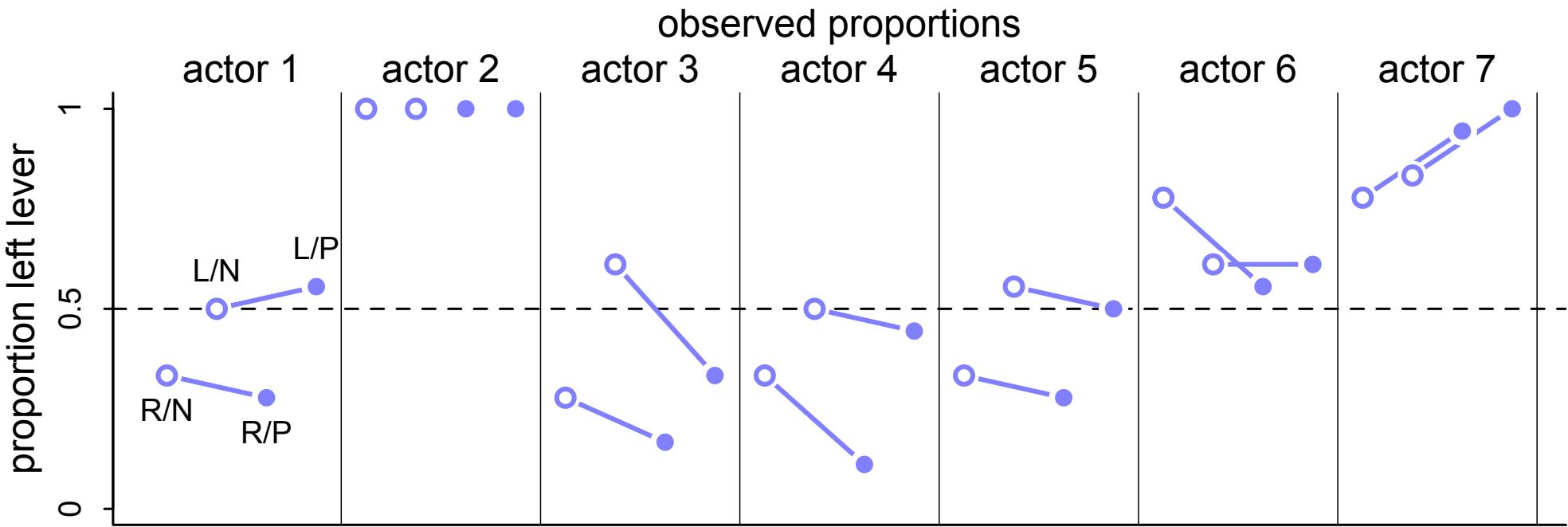


Figure 11.4

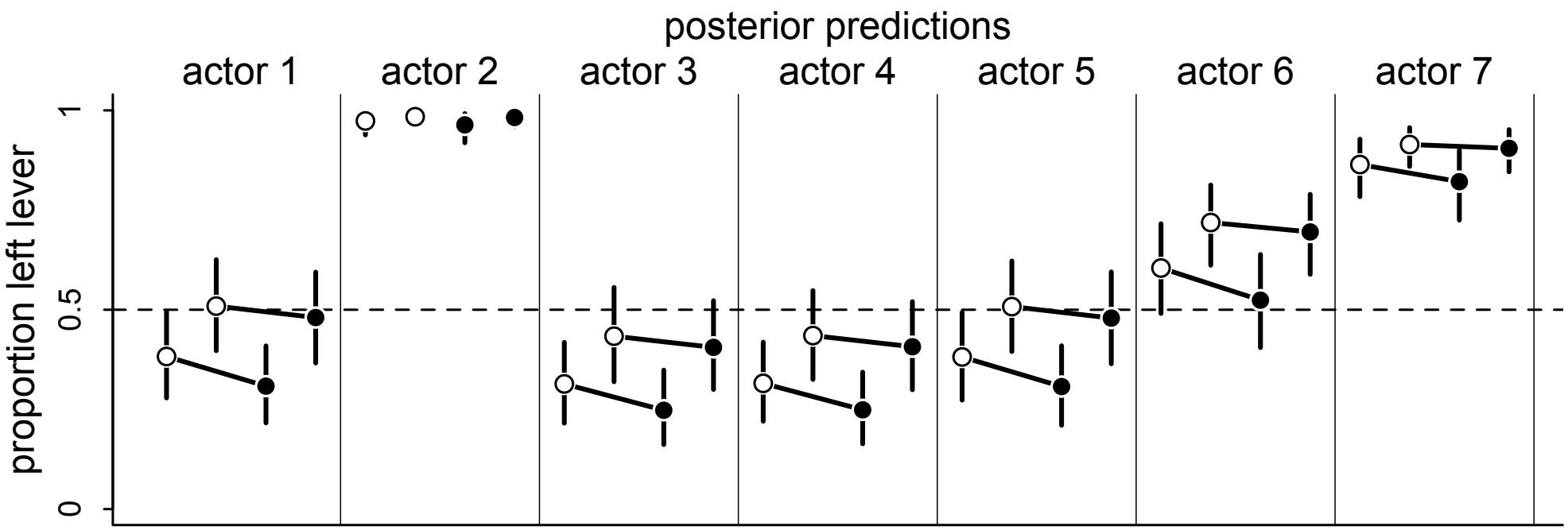
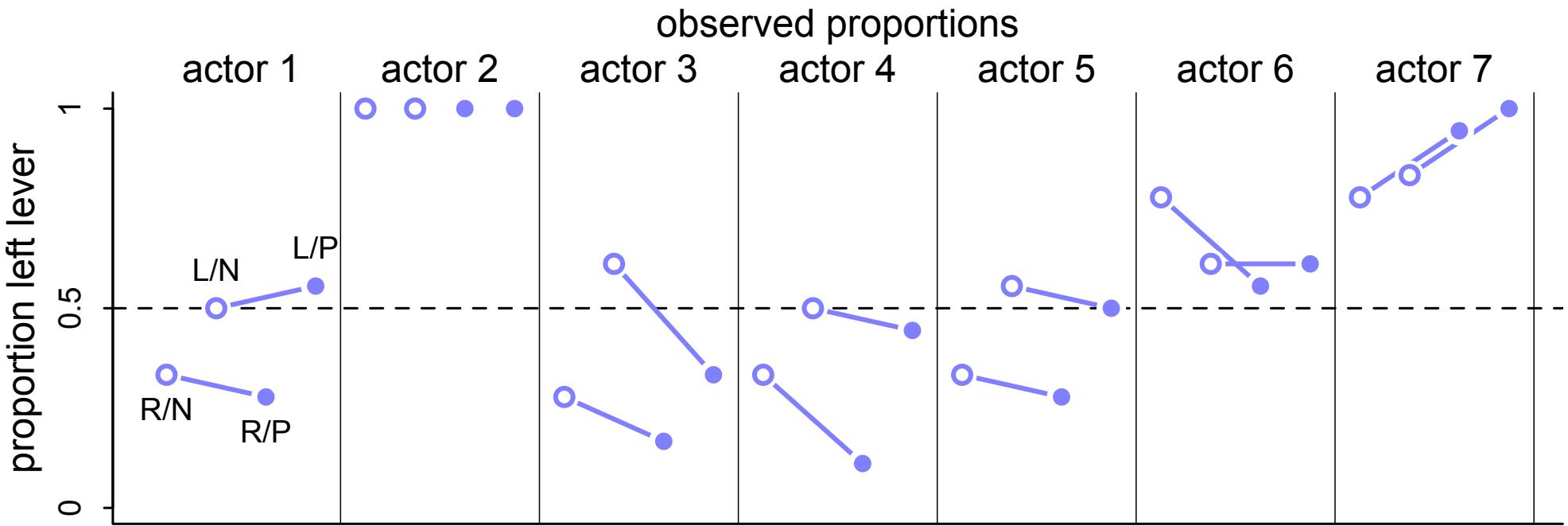


Figure 11.4

Comparing no-interaction

```
m11.5 <- ulam(  
  alist(  
    pulled_left ~ dbinom( 1 , p ) ,  
    logit(p) <- a[actor] + bs[side] + bc[cond] ,  
    a[actor] ~ dnorm( 0 , 1.5 ) ,  
    bs[side] ~ dnorm( 0 , 0.5 ) ,  
    bc[cond] ~ dnorm( 0 , 0.5 )  
  ) ,  
  data=dat_list2 , chains=4 , log_lik=TRUE )
```

R code
11.19

```
compare( m11.5 , m11.4 , func=L00 )
```

	L00	pL00	dL00	weight	SE	dSE
m11.5	531.2	7.9	0.0	0.66	19.17	NA
m11.4	532.6	8.7	1.4	0.34	19.01	1.28

Relative and absolute effects

- Parameters on *relative* effect scale
- Predictions on *absolute* effect scale
- Proportional odds: Relative effect measure

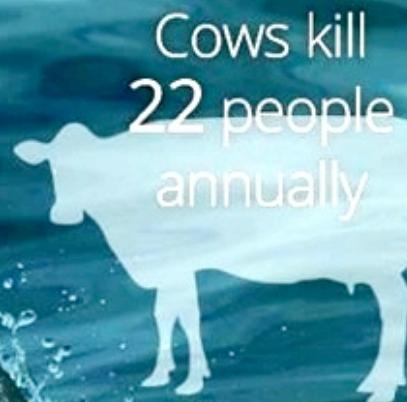
R code
11.22

```
post <- extract.samples(m11.4)
mean( exp(post$b[,4]-post$b[,2]) )
```

[1] 0.9206479



Deer kill
130 people
annually



Cows kill
22 people
annually



Jellyfish kill
40 people
annually

A white silhouette of a shark breaching the surface of a dark blue, wavy background.

Sharks kill
5 people
annually



Ants kill
30 people
annually

A white silhouette of a hippo standing in a dark blue, wavy background.

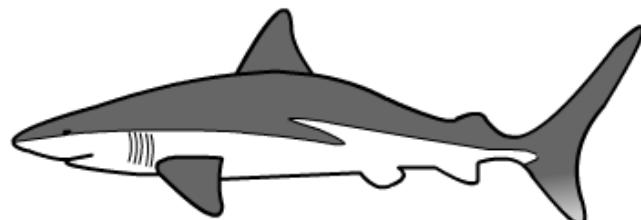
Hippos kill
2,900 people
annually

A white silhouette of a horse walking in a dark blue, wavy background.

Horses kill
20 people
annually

Relative and absolute effects

- Parameters on *relative* effect scale
- Predictions on *absolute* effect scale
- Using relative effects may exaggerate importance of predictor
 - Good for scaring people, getting published
 - Not so good for public health, scientific progress
 - But needed for causal inference



relative shark



absolute penguin

Risk communication

- Many people mistake relative risk for absolute risk
- Example:
 - 1/1000 women develop blood clots
 - 3/1000 women on birth control develop blood clots
 - => 200% increase in blood clots!
 - Change in probability is only 0.002
 - Pregnancy much more dangerous than blood clots

The screenshot shows the homepage of DailyMail.com. The main header reads "Daily Mail.com". Below it is a navigation bar with links: Home, U.K., News (which is highlighted in blue), Sports, U.S. Showbiz, Australia, Fem, News Home, Arts, Headlines, Pictures, Most read, News Board, and Wire. There are three main image thumbnails: one showing a group of women, another showing a man with a rifle, and a third showing Homer Simpson. Below these thumbnails are headlines: "EXCLUSIVE: Grandfather of 'the" (partially cut off), "ISIS attacks Iraqi base where 320 US" (partially cut off), and "Has Homer Simpson actually been i" (partially cut off).

Deadly risk of pill used by 1m GP in Britain told to warn about popular contraceptive

- Bestselling brands of birth control tablets linked to
- They are believed to double the risk compared to older
- 'Third-generation' contraceptives caused 14 deaths
- UK doctors have been ordered to alert women to the

Aggregated Binomial

R code
11.28

```
library(rethinking)
data(UCBadmit)
d <- UCBadmit
```

- Numbers accepted/rejected to 6 PhD programs at UC Berkeley (largest depts in 1973)
- Evidence of gender discrimination? Dean was afraid of lawsuit.
- Call in the statisticians!



UCB admissions

R code
11.28

```
library(rethinking)
data(UCBadmit)
d <- UCBadmit
```

	dept	applicant.gender	admit	reject	applications
1	A	male	512	313	825
2	A	female	89	19	108
3	B	male	353	207	560
4	B	female	17	8	25
5	C	male	120	205	325
6	C	female	202	391	593
7	D	male	138	279	417
8	D	female	131	244	375
9	E	male	53	138	191
10	E	female	94	299	393
11	F	male	22	351	373
12	F	female	24	317	341

Trials vary by row

$$A_i \sim \text{Binomial}(N_i, p_i)$$

$$\text{logit}(p_i) = \alpha_{\text{GID}[i]}$$

$$\alpha_j \sim \text{Normal}(0, 1.5)$$

```
d$gid <- ifelse( d$applicant.gender=="male" , 1 , 2 )
m11.7 <- quap(
  alist(
    admit ~ dbinom( applications , p ) ,
    logit(p) <- a[gid] ,
    a[gid] ~ dnorm( 0 , 1.5 )
  ) , data=d )
precis( m11.7 , depth=2 )
```

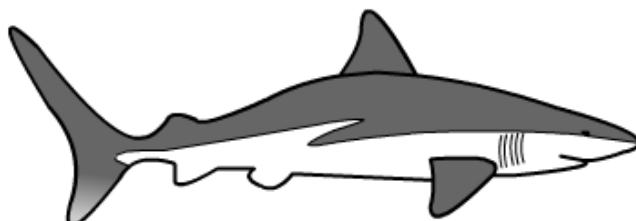
	mean	sd	5.5%	94.5%
a[1]	-0.22	0.04	-0.28	-0.16
a[2]	-0.83	0.05	-0.91	-0.75

Posterior contrast

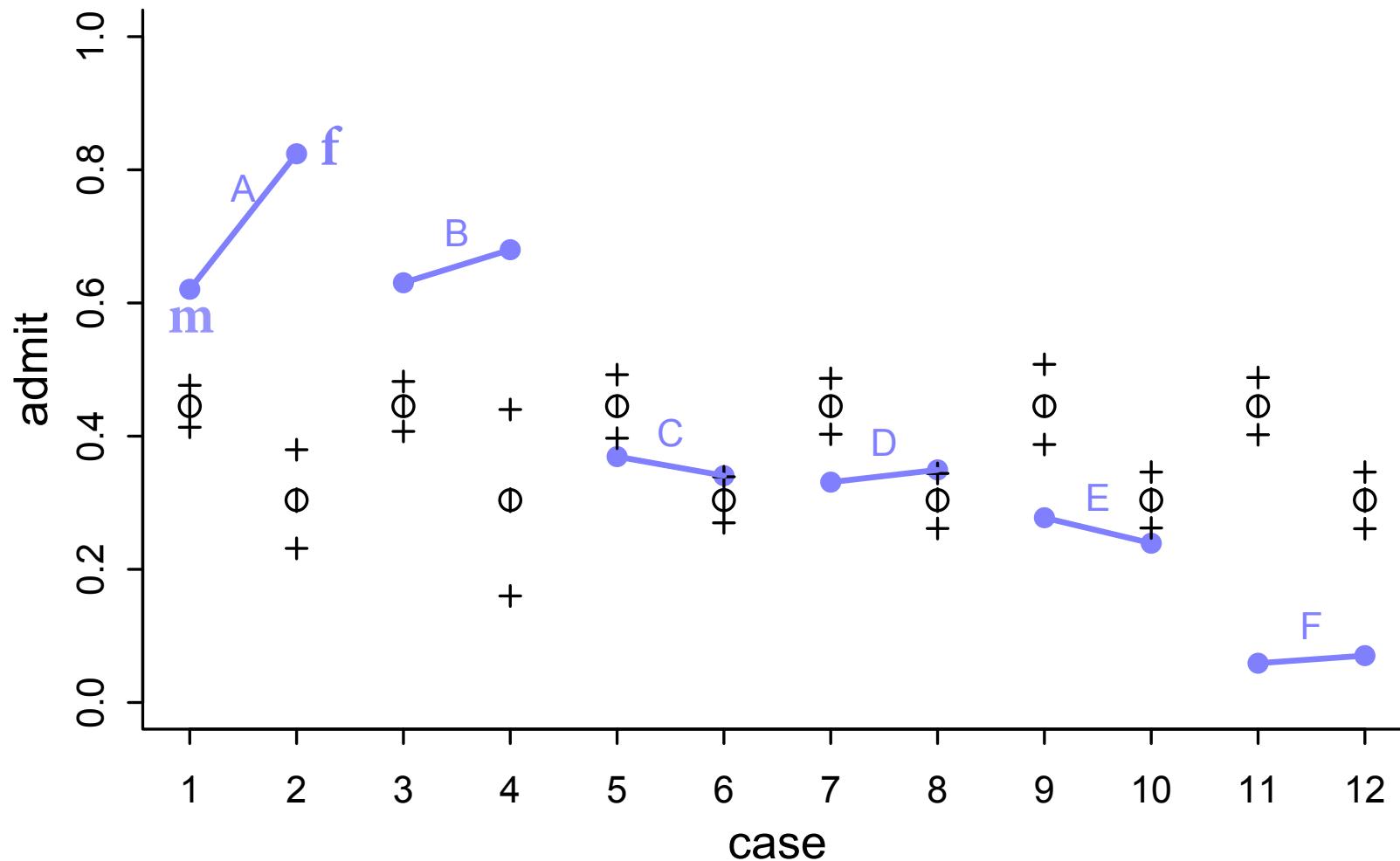
- Compute the contrast between genders
- On both logit (shark) and prob (penguin) scales

```
post <- extract.samples(m11.7)
diff_a <- post$a[,1] - post$a[,2]
diff_p <- inv_logit(post$a[,1]) - inv_logit(post$a[,2])
precis( list( diff_a=diff_a , diff_p=diff_p ) )
```

```
'data.frame': 10000 obs. of 2 variables:
  mean    sd  5.5% 94.5%      histogram
diff_a 0.61  0.06  0.51   0.71      [histogram]
diff_p 0.14  0.01  0.12   0.16      [histogram]
```



Posterior validation check

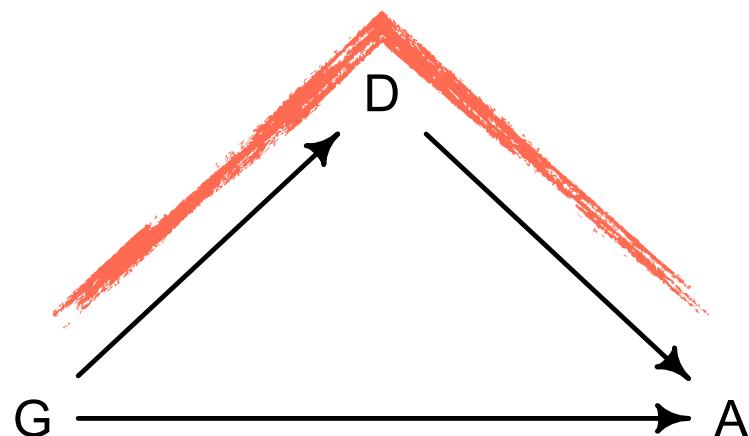


Females admitted more in all but 2 departments!

Figure 11.5

Backdoor admissions

- Backdoor path through department
- Use unique intercepts to adjust for that path



$$\begin{aligned}A_i &\sim \text{Binomial}(N_i, p_i) \\ \text{logit}(p_i) &= \alpha_{\text{GID}}[i] + \delta_{\text{DEPT}}[i] \\ \alpha_j &\sim \text{Normal}(0, 1.5) \\ \delta_k &\sim \text{Normal}(0, 1.5)\end{aligned}$$

Stratification by department

$$A_i \sim \text{Binomial}(N_i, p_i)$$

$$\text{logit}(p_i) = \alpha_{\text{GID}[i]}$$

$$\alpha_j \sim \text{Normal}(0, 1.5)$$

Stat Q: What are the average probabilities of admission for females and males across all departments?

Causal Q: What is the TOTAL influence of gender?

Stratification by department

$$A_i \sim \text{Binomial}(N_i, p_i)$$

$$\text{logit}(p_i) = \alpha_{\text{GID}[i]}$$

$$\alpha_j \sim \text{Normal}(0, 1.5)$$

$$A_i \sim \text{Binomial}(N_i, p_i)$$

$$\text{logit}(p_i) = \alpha_{\text{GID}[i]} + \delta_{\text{DEPT}[i]}$$

$$\alpha_j \sim \text{Normal}(0, 1.5)$$

$$\delta_k \sim \text{Normal}(0, 1.5)$$

Stat Q: What are the average probabilities of admission for females and males across all departments?

Causal Q: What is the TOTAL influence of gender?

Stat Q: What is the average difference in probability of admission for females and males within departments?

Causal Q: What is the DIRECT influence of gender?

Stratification by department

```
d$dept_id <- rep(1:6,each=2)
m11.8 <- quap(
  alist(
    admit ~ dbinom( applications , p ) ,
    logit(p) <- a[gid] + delta[dept_id] ,
    a[gid] ~ dnorm( 0 , 1.5 ) ,
    delta[dept_id] ~ dnorm( 0 , 1.5 )
  ) , data=d )
precis( m11.8 , depth=2 )
```

	mean	sd	5.5%	94.5%
a[1]	-0.53	0.53	-1.38	0.32
a[2]	-0.43	0.53	-1.28	0.42
delta[1]	1.11	0.54	0.25	1.96
delta[2]	1.06	0.54	0.20	1.92
delta[3]	-0.15	0.53	-1.00	0.70
delta[4]	-0.18	0.54	-1.04	0.67
delta[5]	-0.62	0.54	-1.48	0.23
delta[6]	-2.17	0.55	-3.05	-1.30

Again with shark & penguin

- Difference on logit and probability scales

R code
11.33

```
post <- extract.samples(m11.8)
diff_a <- post$a[,1] - post$a[,2]
diff_p <- inv_logit(post$a[,1]) - inv_logit(post$a[,2])
precis( list( diff_a=diff_a , diff_p=diff_p ) )
```

'data.frame': 10000 obs. of 2 variables:
 mean sd 5.5% 94.5% histogram
diff_a -0.10 0.08 -0.22 0.03 
diff_p -0.02 0.02 -0.05 0.01 

Backdoor admissions

- What happened? Females apply more to most selective departments. So overall rate of admission lower.
- Proportions of m/f applications by department:

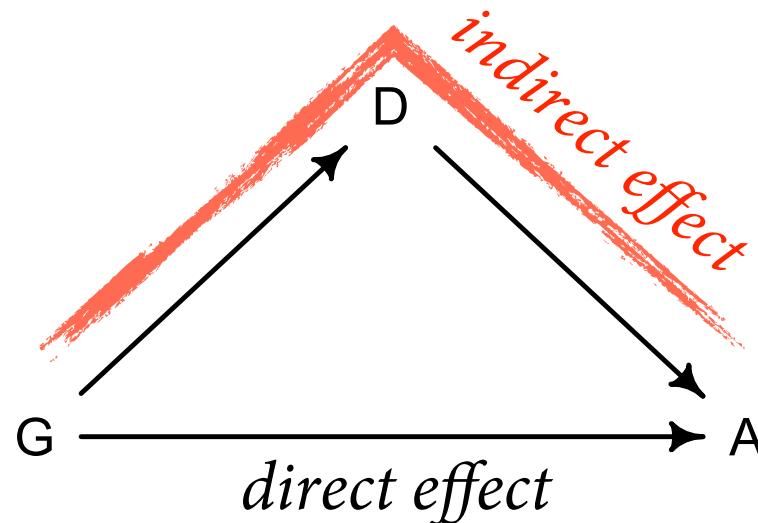
R code
11.34

```
pg <- sapply( 1:6 , function(k)
              d$applications[d$dept_id==k]/sum(d$applications[d$dept_id==k]) )
rownames(pg) <- c("male","female")
colnames(pg) <- unique(d$dept)
round( pg , 2 )
```

	A	B	C	D	E	F
male	0.88	0.96	0.35	0.53	0.33	0.52
female	0.12	0.04	0.65	0.47	0.67	0.48

Backdoor admissions

- Careful about causal interpretation
- No evidence for direct path $G \rightarrow A$
- Lots of evidence of indirect path $G \rightarrow D \rightarrow A$
- Total causal influence of G still strong
- Still results in disenfranchisement
- But effective intervention very different



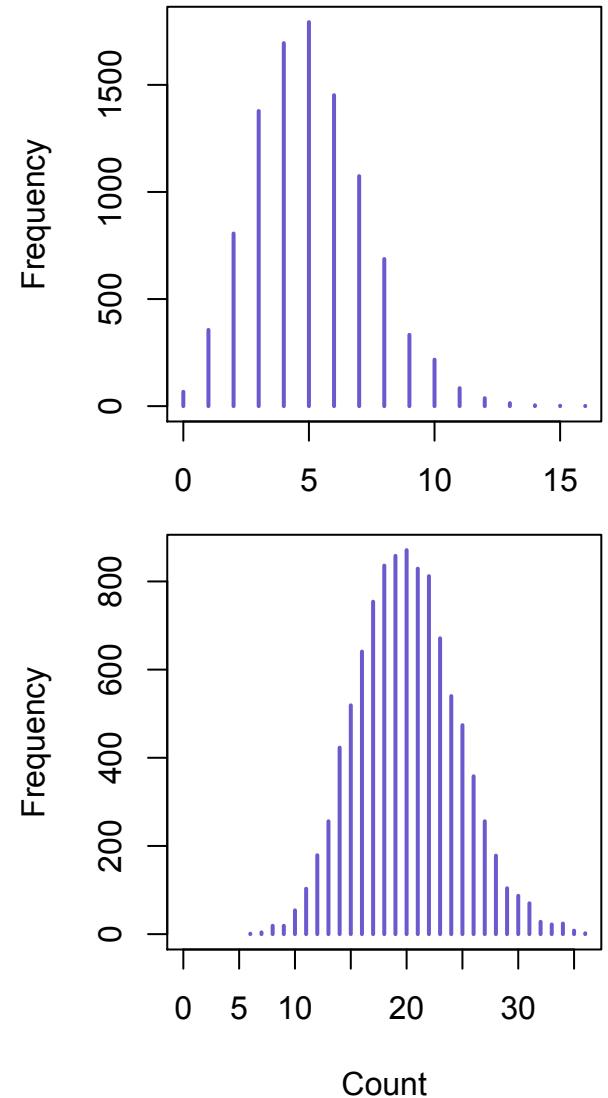
Poisson GLMs

$$y \sim \text{Poisson}(\lambda)$$

$$\text{E}(y) = \lambda$$

$$\text{var}(y) = \lambda$$

- Counts without upper limit, constant expected value
- Single parameter: events per unit time/distance
- Variance equal to mean

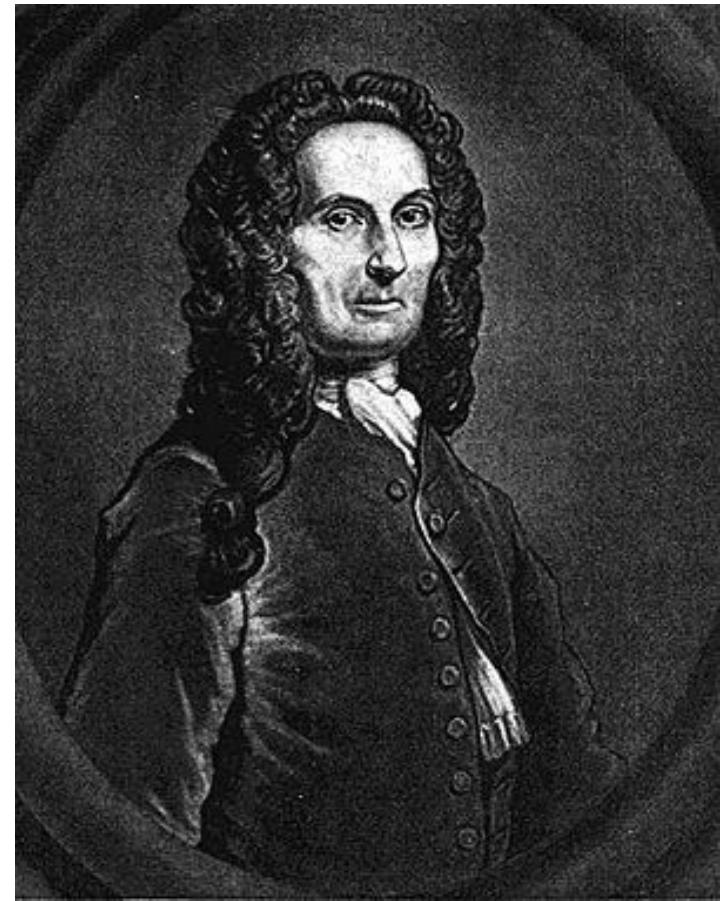


Poisson GLMs

- Examples: Soccer goals, fission events, photons striking a detector, DNA mutations, soldiers killed by horses



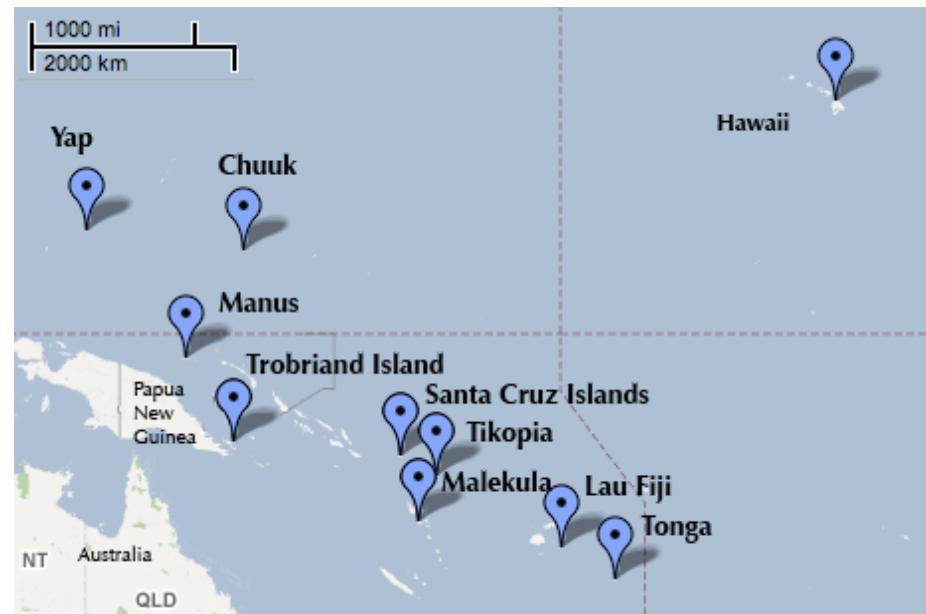
Siméon Denis Poisson (1781–1840)



Abraham de Moivre (1667–1754)

Oceanic tool complexity

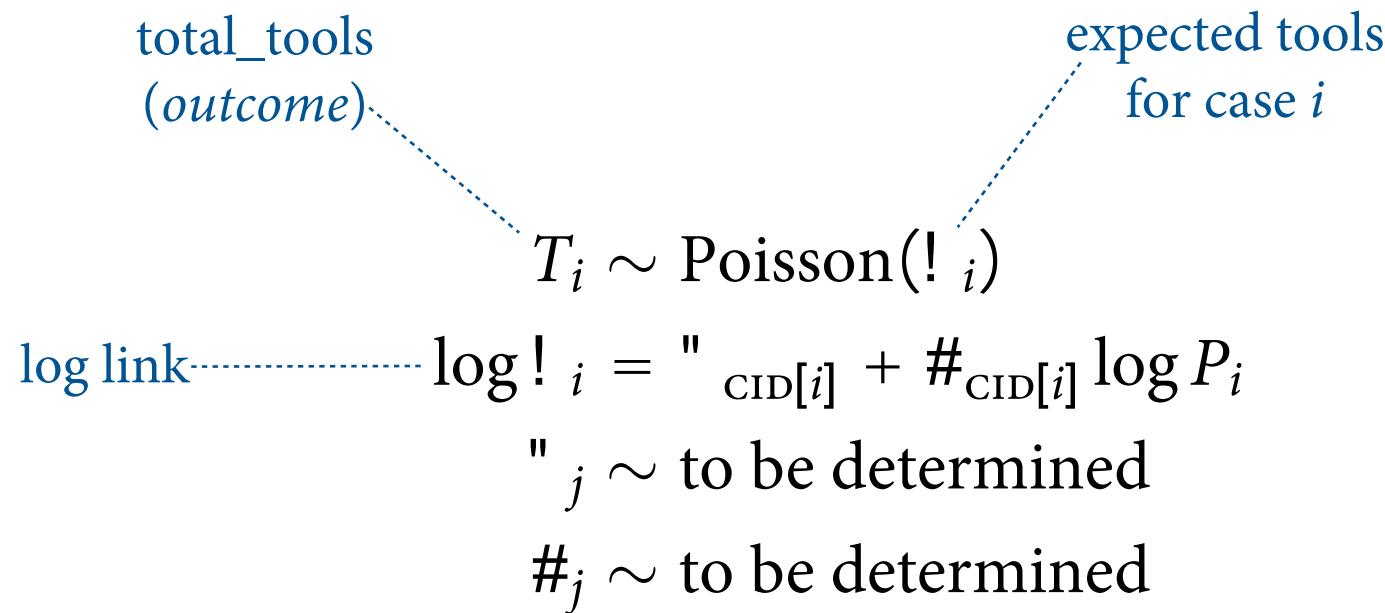
culture	population	contact	total_tools	mean_TU
Malekula	1100	low	13	3.2
Tikopia	1500	low	22	4.7
Santa Cruz	3600	low	24	4.0
Yap	4791	high	43	5.0
Lau Fiji	7400	high	33	5.0
Trobriand	8000	high	19	4.0
Chuuk	9200	high	40	3.8
Manus	13000	low	28	6.6
Tonga	17500	high	55	5.4
Hawaii	275000	low	71	6.6



- (1) Complexity of toolkit proportional to magnitude of population?
- (2) Contact with other islands moderates impact?

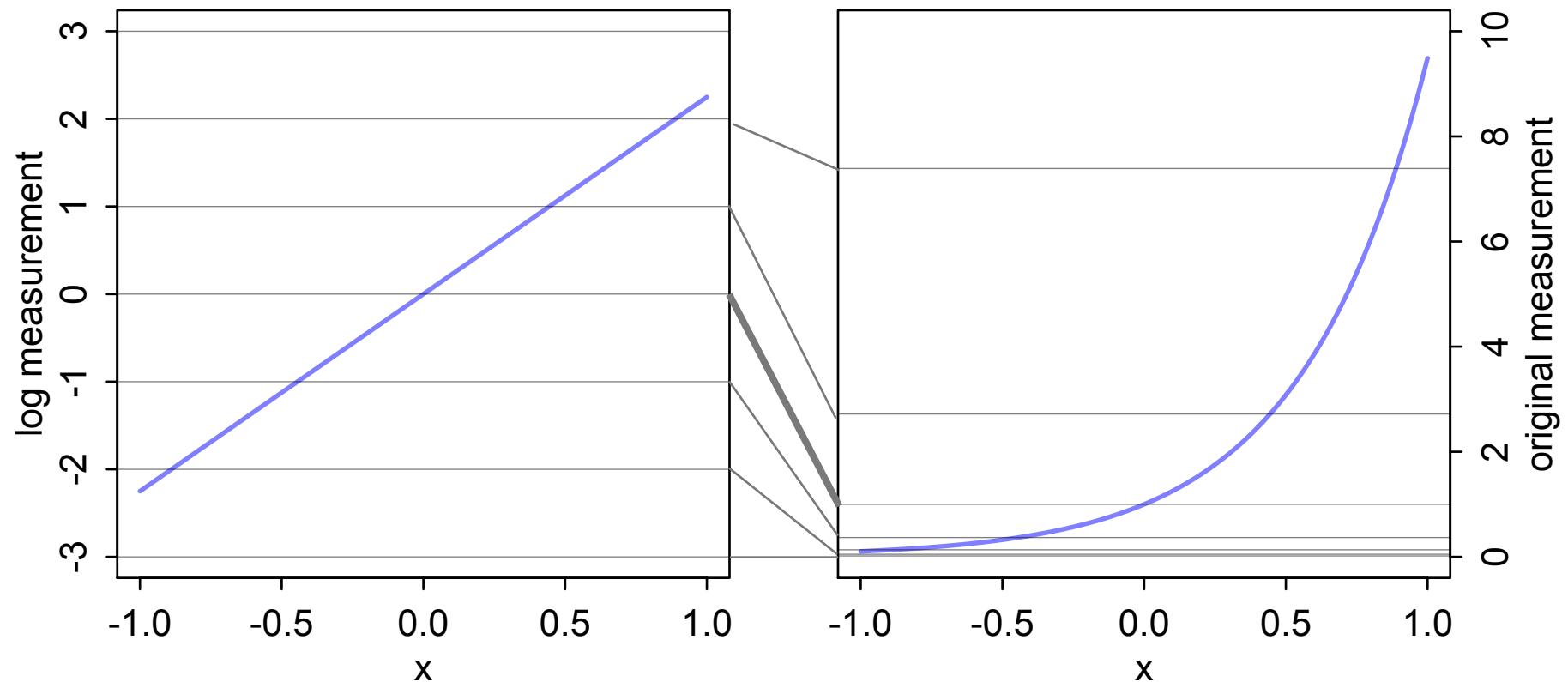


Anatomy of Poisson GLM



Log link

- Goal: Map linear model to positive reals
- Log link maps all negative numbers to $[0,1]$
- All positive numbers to $[1,\infty]$



Priors & the log link

- Log link not intuitive — simulate

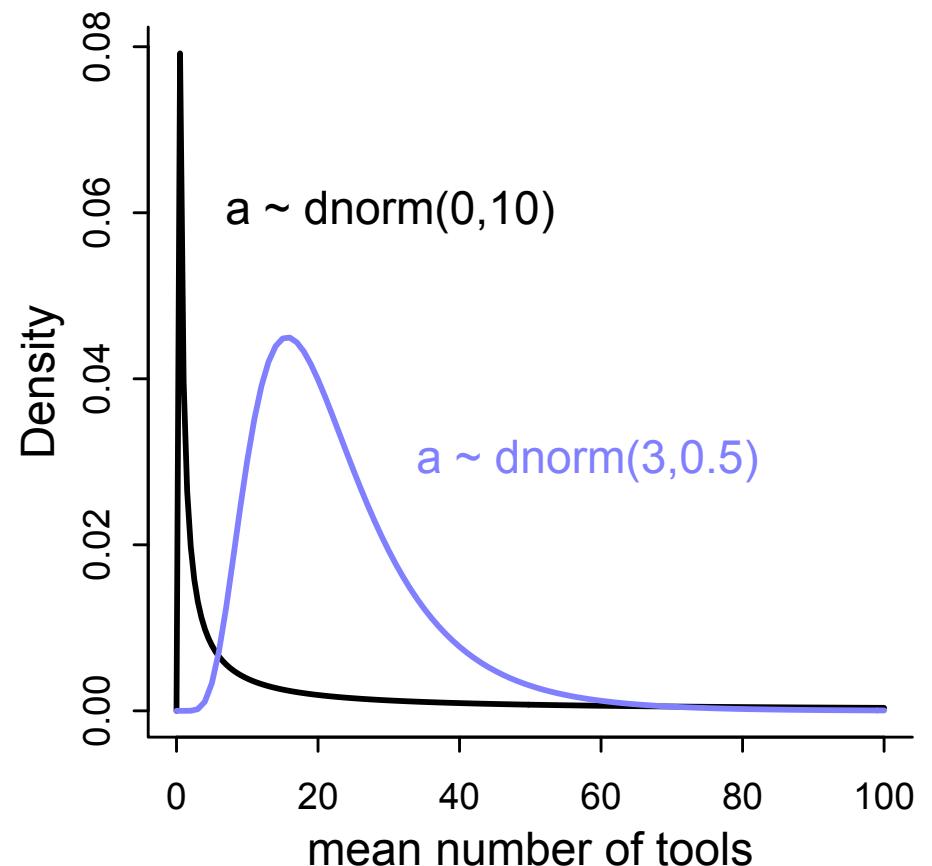
$$T_i \sim \text{Poisson}(\lambda_i)$$

$$\log \lambda_i = \alpha$$

$$\alpha \sim \text{Normal}(0, 10)$$

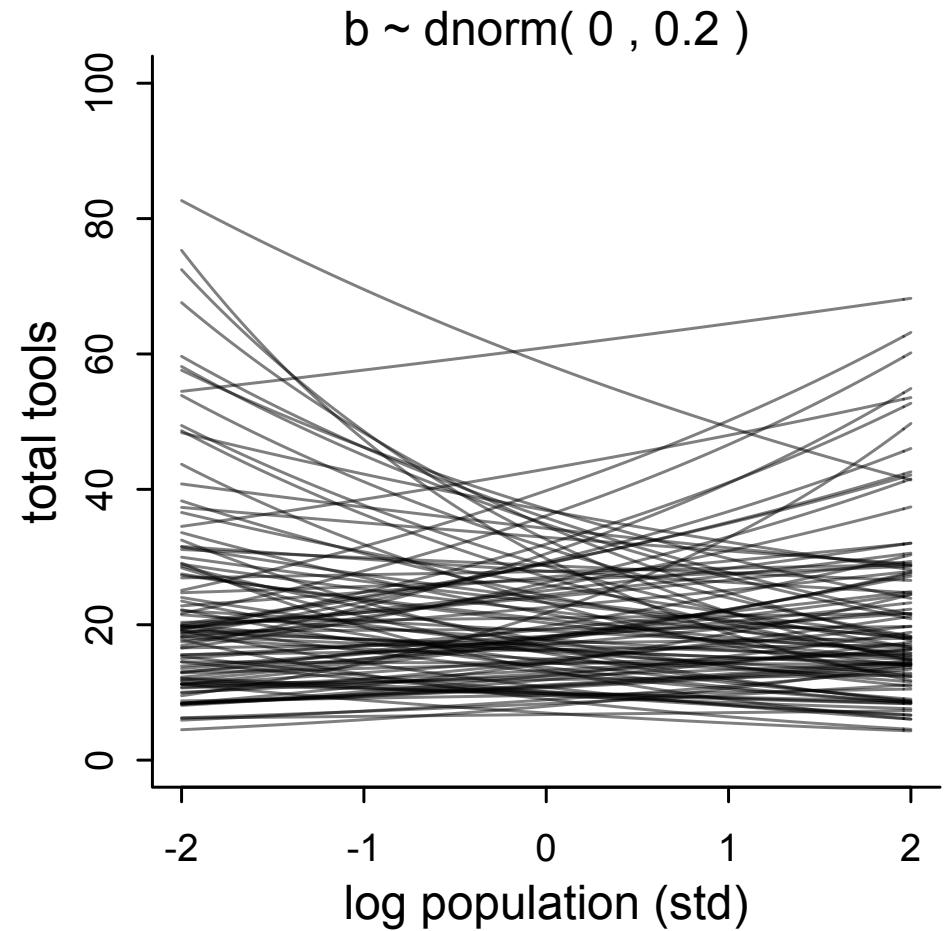
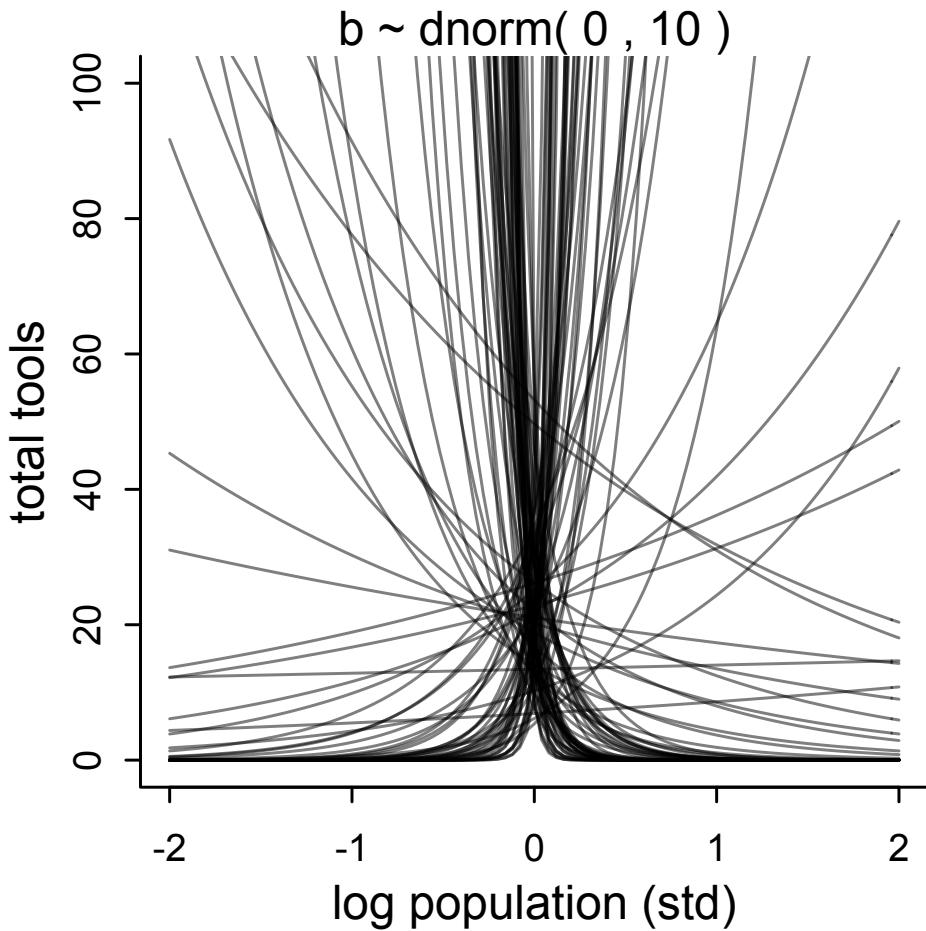
```
a <- rnorm(1e4, 0, 10)
lambda <- exp(a)
mean( lambda )
```

```
[1] 9.622994e+12
```



Priors & the log link

- Slopes equally unintuitive



Tools models

R code
11.48

```
dat <- list(
  T = d$total_tools ,
  P = d$P ,
  cid = d$contact_id )

# intercept only
m11.9 <- ulam(
  alist(
    T ~ dpois( lambda ),
    log(lambda) <- a,
    a ~ dnorm(3,0.5)
  ), data=dat , chains=4 , log_lik=TRUE )

# interaction model
m11.10 <- ulam(
  alist(
    T ~ dpois( lambda ),
    log(lambda) <- a[cid] + b[cid]*P,
    a[cid] ~ dnorm( 3 , 0.5 ),
    b[cid] ~ dnorm( 0 , 0.2 )
  ), data=dat , chains=4 , log_lik=TRUE )
```

Compare using PSIS-LOO

```
compare( m11.9 , m11.10 , func=LOO )
```

R code
11.49

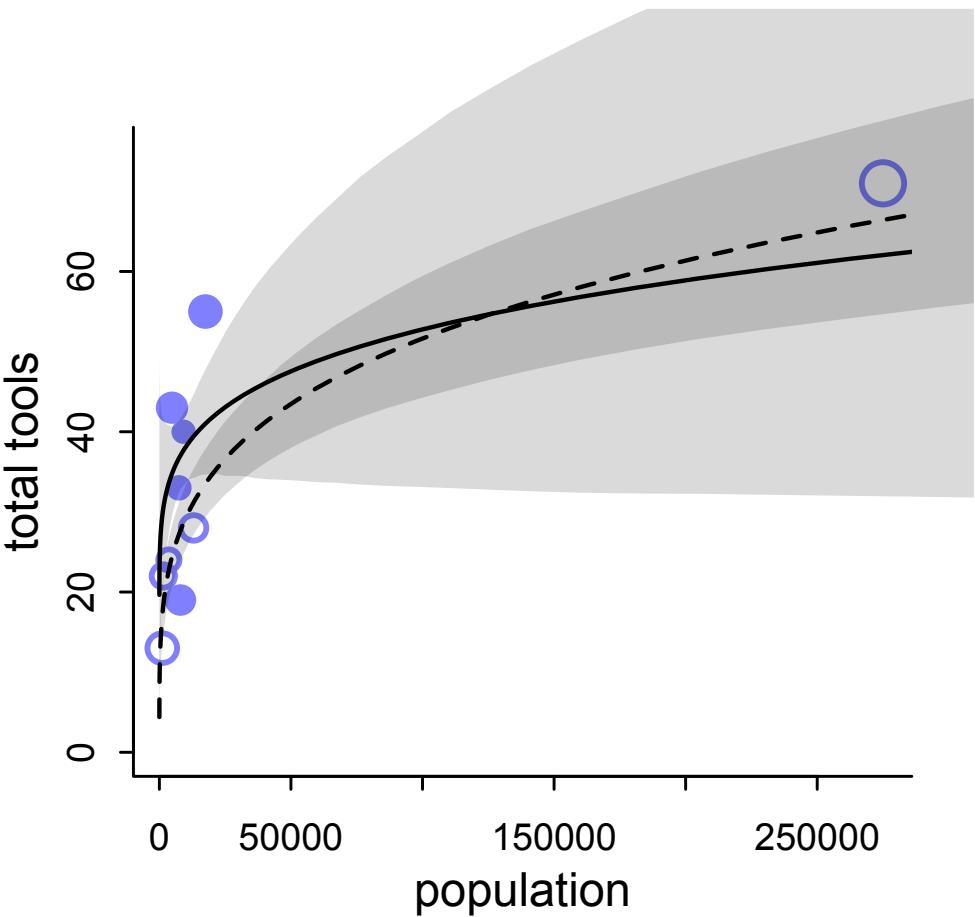
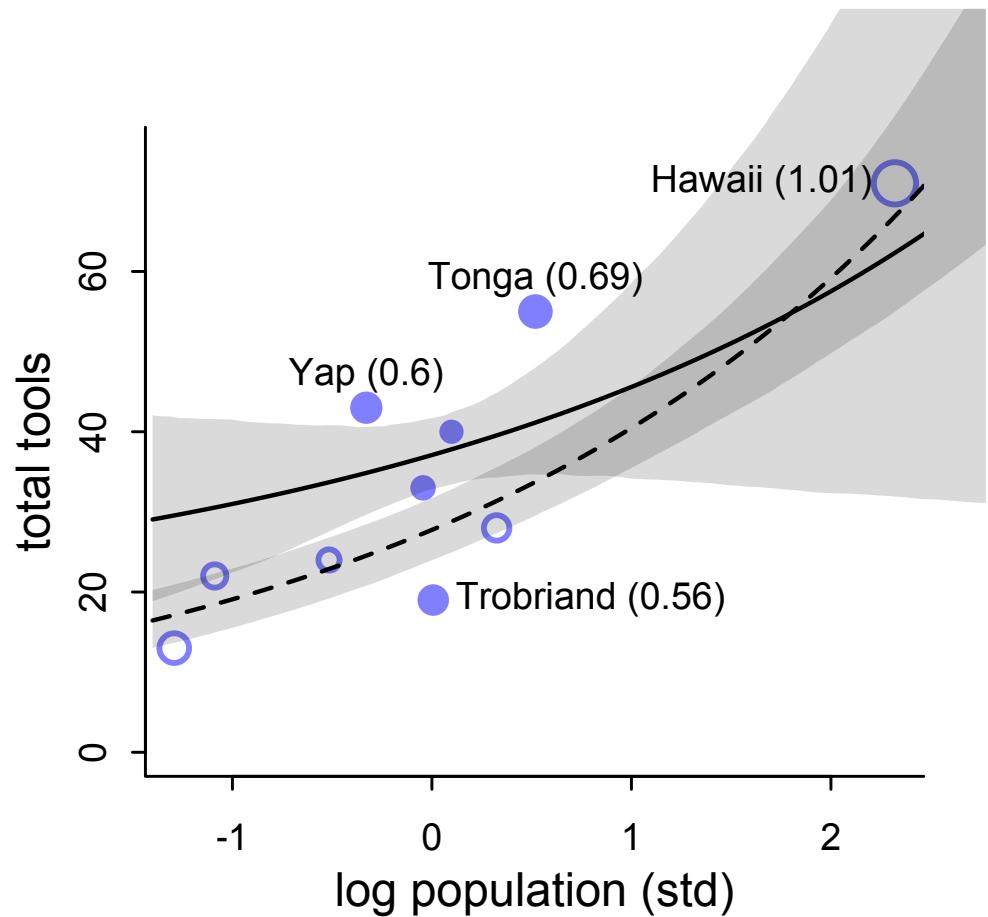
```
    LOO  pLOO dLOO weight     SE   dSE
m11.10 85.5  7.1  0.0      1 13.22    NA
m11.9 141.1  8.0 55.5      0 33.33 32.78
```

Warning messages:

```
1: Some Pareto k diagnostic values are too high. See help('pareto-k-diagnostic') for details.
```

- Warning indicates strongly influential points
- Look at those pLOO values: No relationship to parameter count — not unusual & not a mistake

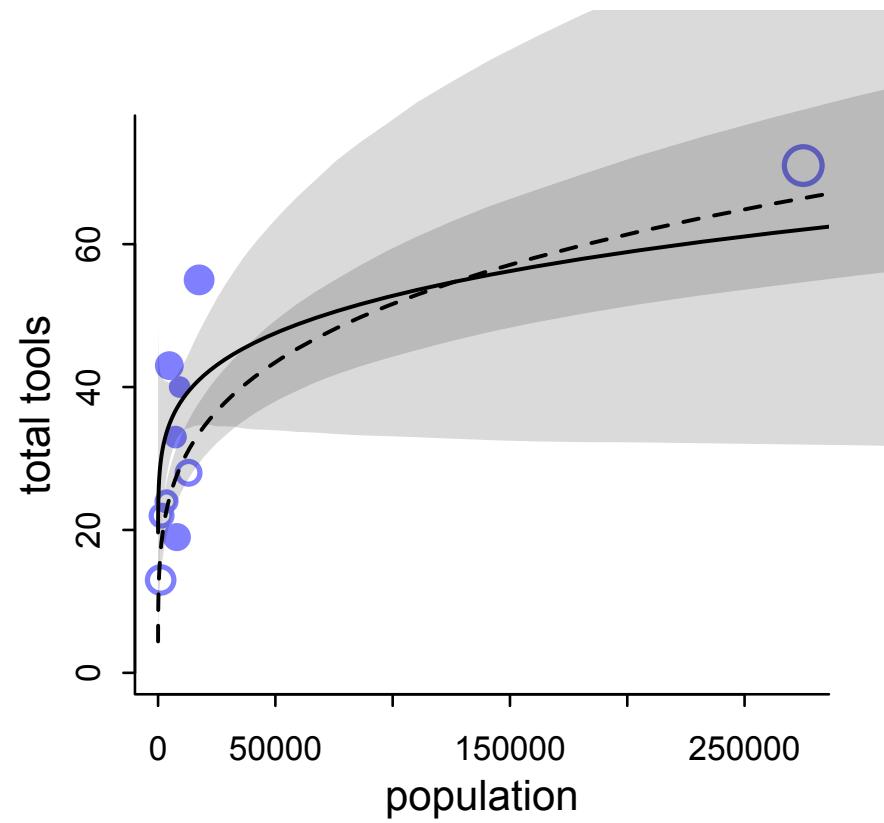
Hawaii has leverage



- Point size proportional to Pareto- k diagnostic value

Generalized Linear Madness

- This model is terrible:
- Intercepts don't pass through origin
- Zero population = zero tools
- We can do better by thinking *scientifically* instead of *statistically*



Scientific model

- Change in tools per unit time:

$$\Delta T =$$

Scientific model

- Change in tools per unit time:

Diminishing returns
("elasticity")

Innovation rate

$$\Delta T = \alpha P^\beta$$

Population

The diagram illustrates a scientific model. At the top right, the text "Diminishing returns ("elasticity")" is displayed. Below it, the term "Innovation rate" is written. A dashed blue line connects this text to the left side of the central equation. The central equation is $\Delta T = \alpha P^\beta$. To the left of the equation, the word "Population" is written, and a dashed blue line connects it to the right side of the equation. The entire equation is centered under the heading "Diminishing returns ("elasticity")".

Scientific model

- Change in tools per unit time:

$$\Delta T = \alpha P^\beta - \gamma T$$

Diminishing returns
("elasticity")

The diagram illustrates the components of the equation $\Delta T = \alpha P^\beta - \gamma T$. It features a central equation with three dashed arrows pointing to its terms. The first arrow, from the left, points to the term αP^β and is labeled "Innovation rate". The second arrow, from the bottom, points to the term T and is labeled "Population". The third arrow, from the right, points to the term γT and is labeled "Loss rate".

Scientific model

- Solve for steady state expected number of tools
- Where $\Delta T = 0$

$$\hat{T} = \frac{\alpha P^\beta}{\gamma}$$

$$T_i \sim \text{Poisson}(\lambda_i)$$

$$\lambda_i = \alpha P_i^\beta / \gamma$$

No *ad hoc* link function!

Scientific model

$$T_i \sim \text{Poisson}(\lambda_i)$$

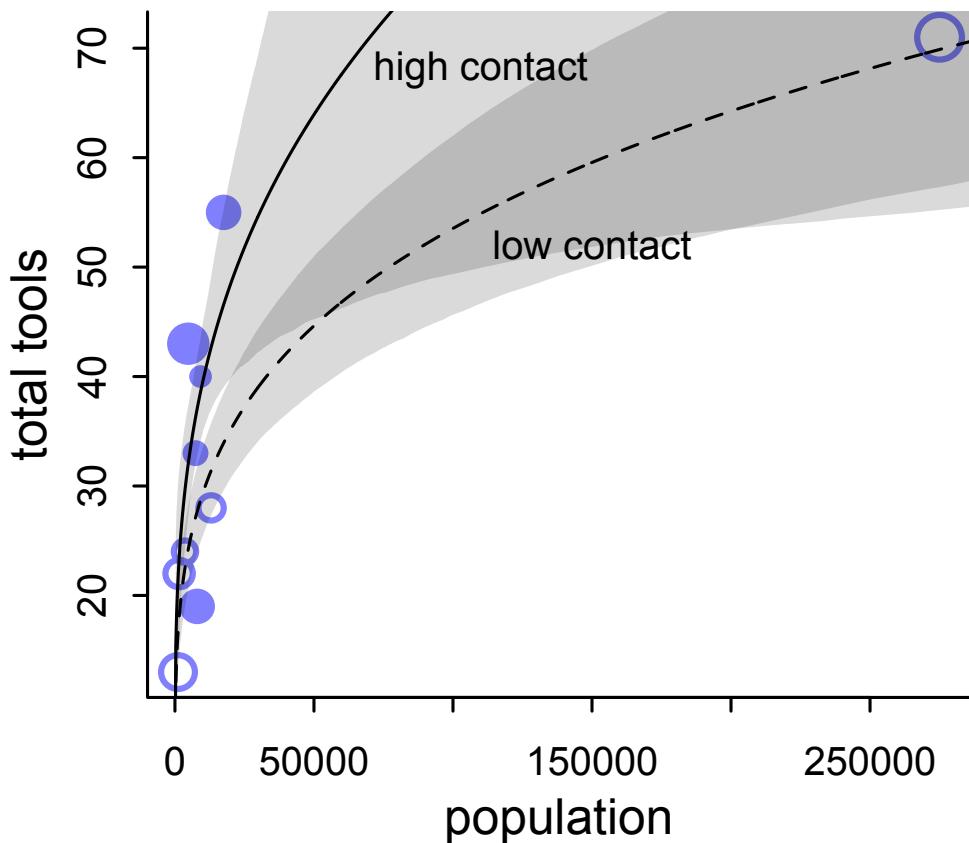
$$\lambda_i = \alpha P_i^\beta / \gamma$$

```
dat2 <- list( T=d$total_tools, P=d$population, cid=d$contact_id )
m11.11 <- ulam(
  alist(
    T ~ dpois( lambda ),
    lambda <- exp(a[cid])*P^b[cid]/g,
    a[cid] ~ dnorm(1,1),
    b[cid] ~ dexp(1),
    g ~ dexp(1)
  ), data=dat2 , chains=4 , log_lik=TRUE )
```

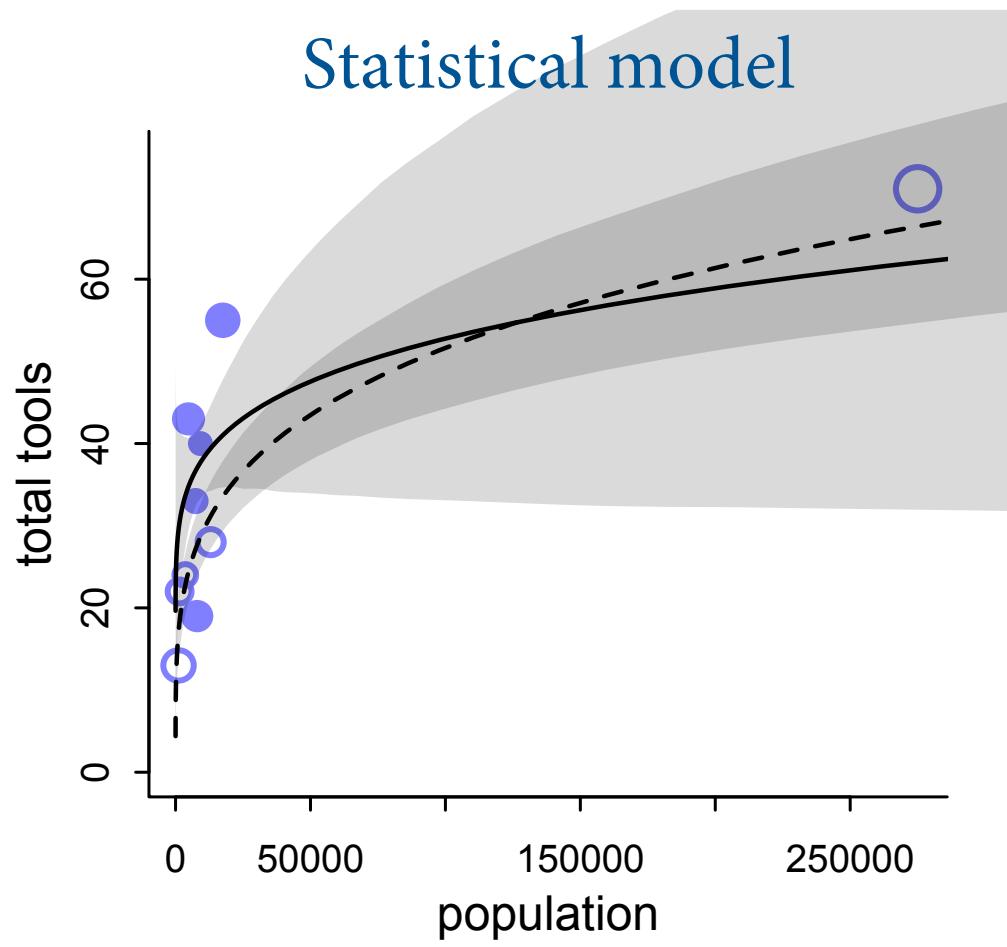
R code
11.52

Science pays

Scientific model



Statistical model



Model violations now mean something.
Parameters now mean something.

Poisson exposure (offsets)

- Poisson outcome: events per unit time/distance
- Q: What if time/distance varies across cases?
- A: Use an *exposure*, aka *offset*

$$y_i \sim \text{Poisson}(\lambda_i)$$

$$\log \lambda_i = \log \frac{\mu_i}{\tau_i} = \alpha + \beta x_i$$

exposure *expected count*

$$y_i \sim \text{Poisson}(\mu_i)$$

$$\log \mu_i = \log \tau_i + \alpha + \beta x_i$$

Additional count distributions

- Multinomial/categorical: generalized binomial, more than 2 un-ordered outcomes
- Geometric: number of trials until specific event
- Mixtures, coping with heterogeneity:
 - Beta-binomial: varying probabilities
 - gamma-Poisson: aka negative-Binomial, varying rates
 - others (e.g. Dirichlet-multinomial)

Survival Analysis

- Count models are fundamentally about rates
 - Rate of heads per coin toss
 - Rate of tools per person
- Can also estimate rates by modeling time-to-event
- Tricky, because cannot ignore *censored* cases
 - Left-censored: Don't know when time started
 - Right-censored: Something cut observation off before event occurred
- Ignoring censored cases leads to inferential error
 - Imagine estimating time-to-PhD but ignoring people who drop out
 - Time in program before dropping out is info about rate

Survival Analysis

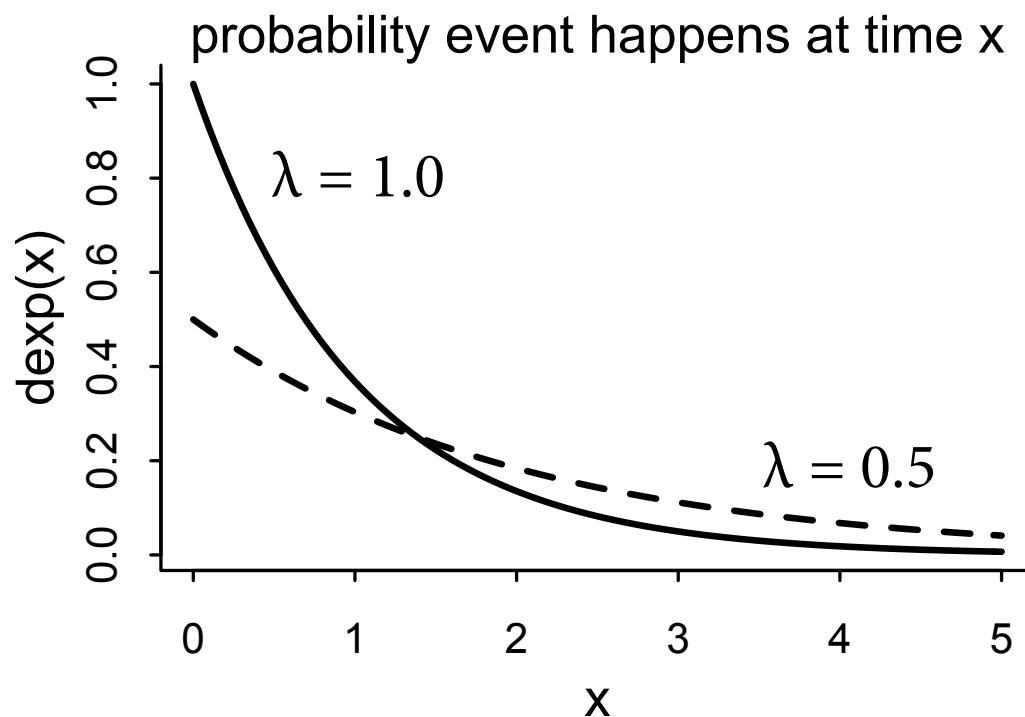
- Example: Cat adoptions
- `data(AustinCats)`
 - 20-thousand cats
 - time-to-event
 - Event either: (1) adopted or (2) something else
 - Something else could be: death, escape, **censored**



Un-censored observations

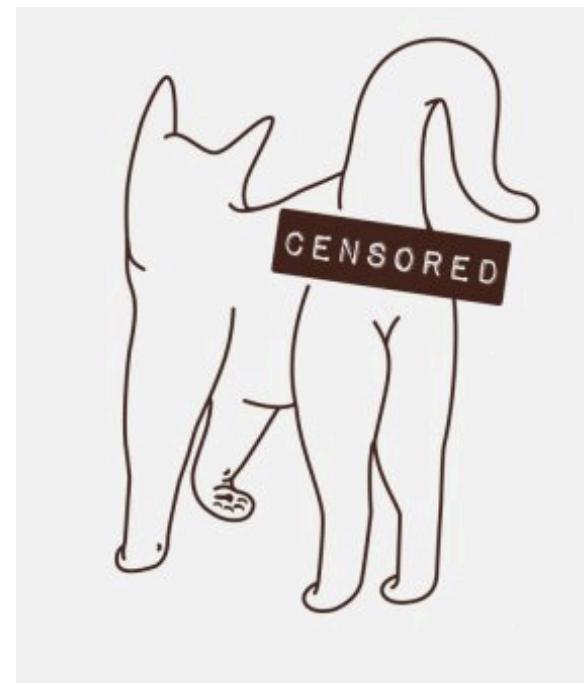
- For observed adoptions, just need:

$$D_i \sim \text{Exponential}(\lambda_i) \quad p(D_i | \lambda_i) = \lambda_i \exp(-\lambda_i D_i)$$

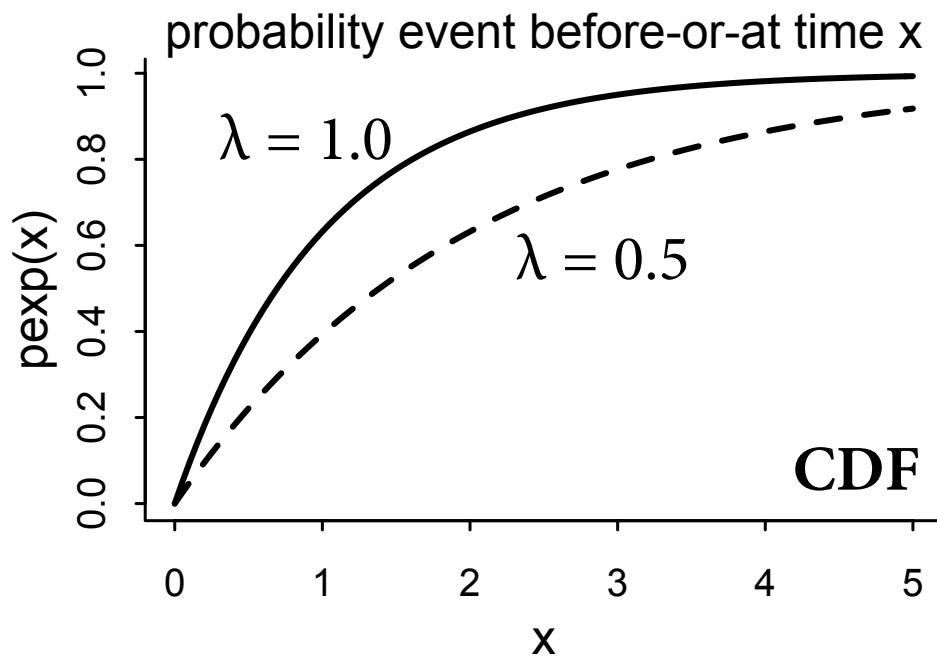


Censored cats

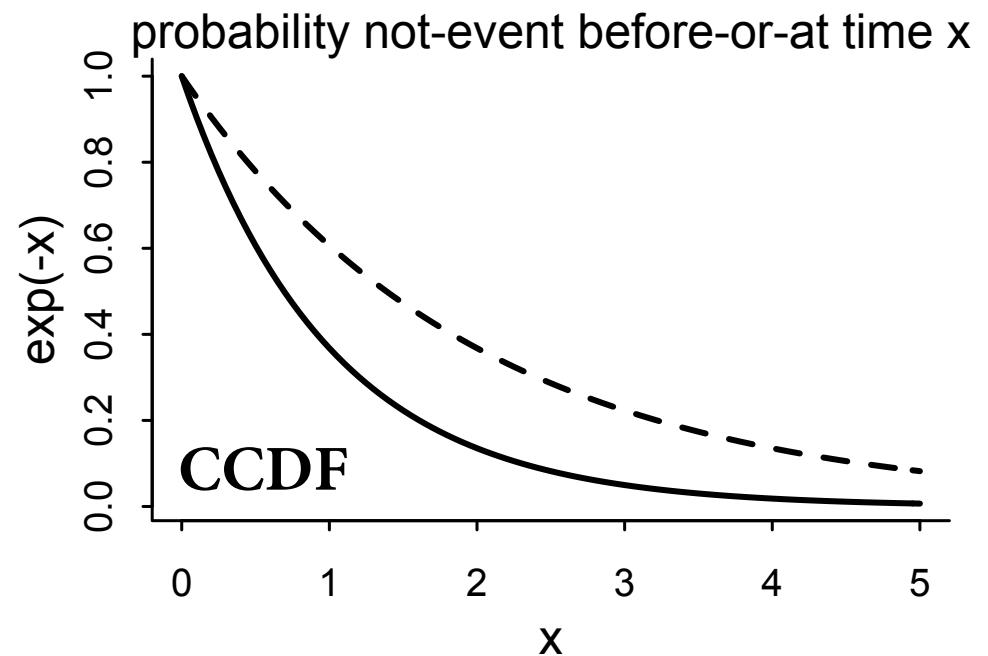
- Cumulative distribution (CDF): Probability event before-or-at time x
- Complementary cumulative distribution (CCDF): Probability not-event-yet



$$\Pr(D_i|\lambda_i) = 1 - \exp(-\lambda_i D_i)$$



$$\Pr(D_i|\lambda_i) = \exp(-\lambda_i D_i)$$



Cat code

$$D_i | A_i = 1 \sim \text{Exponential}(\lambda_i)$$

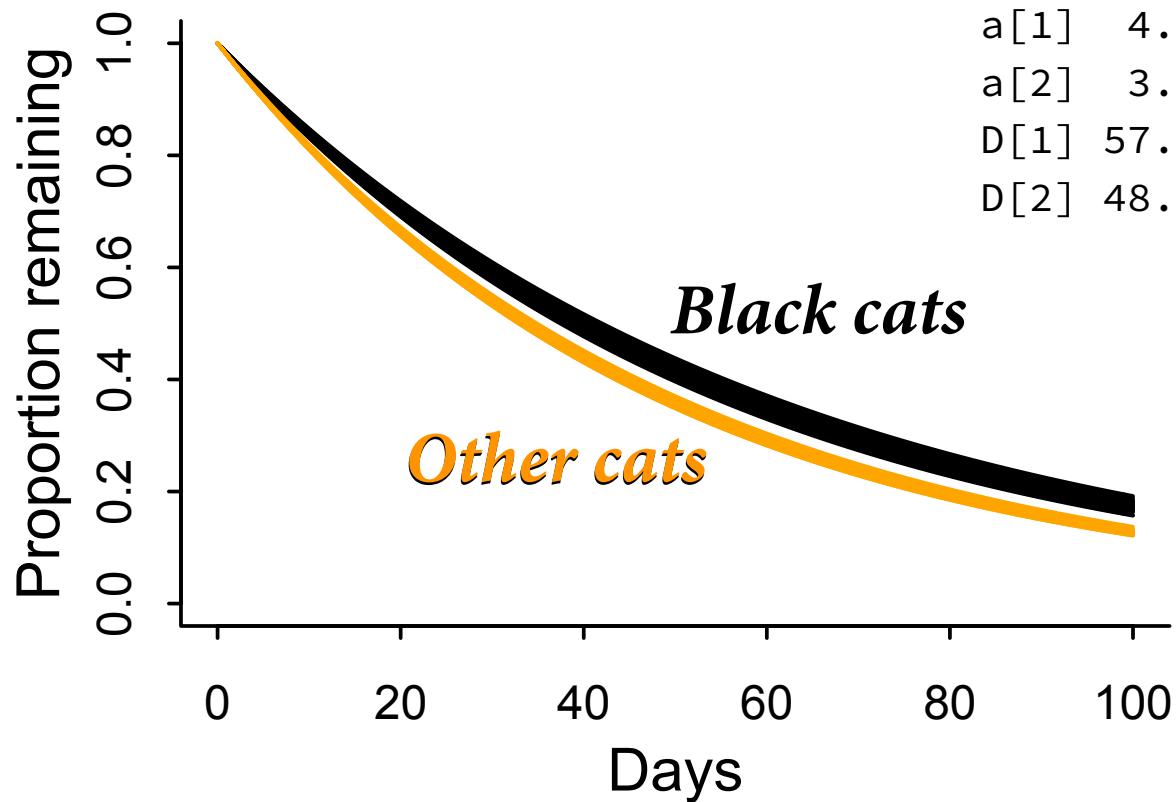
$$D_i | A_i = 0 \sim \text{Exponential-CCDF}(\lambda_i)$$

$$\lambda_i = 1/\mu_i$$

$$\log \mu_i = \alpha_{\text{CID}[i]}$$

```
m11.14 <- ulam(  
  alist(  
    days_to_event|adopted==1 ~ exponential( lambda ),  
    days_to_event|adopted==0 ~ custom(exponential_lccdf( !Y | lambda )),  
    lambda <- 1.0/mu,  
    log(mu) <- a[color_id],  
    a[color_id] ~ normal(0,1)  
, data=dat , chains=4 , cores=4 )
```

Posterior survival curves



```
post <- extract.samples( m11.14 )
post$D <- exp(post$a)
precis( post , 2 )
```

```
'data.frame': 2000 obs. of 4 variables:
               mean     sd   5.5% 94.5% histogram
a[1]    4.05 0.03  4.01  4.09
a[2]    3.88 0.01  3.87  3.90
D[1]   57.44 1.47 55.11 59.77
D[2]   48.44 0.49 47.71 49.22
```

Homework

- 3 problems, 2 data sets, multiple good DAGs
- One of the data sets (NWOGrants) is new in rethinking 1.83, so update
- Next week: More adventures with integers

