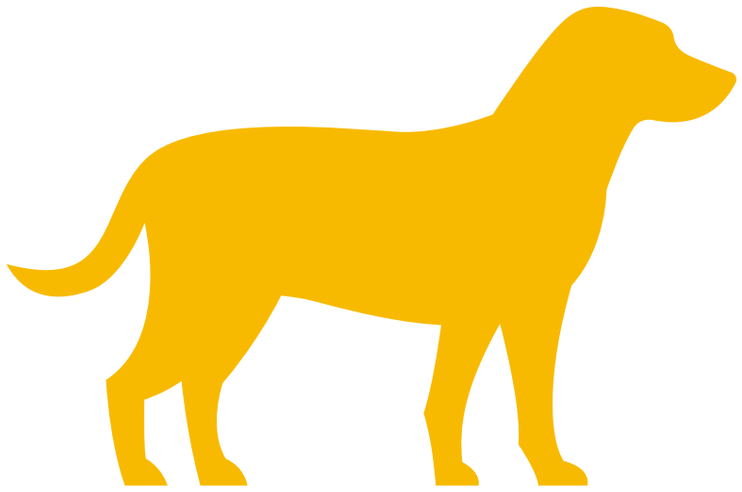
Bayesian Analysis of Correlated Proportions Using Stan in R

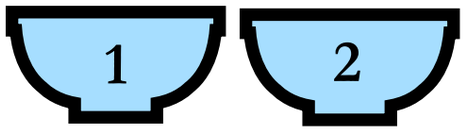
Jay Harrison
Jay.Harrison@ssmhealth.com

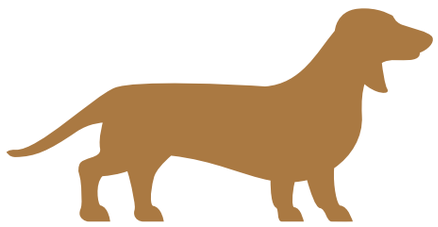
Acknowledgements

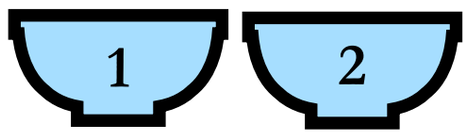
- Dan Ha
- Susan Jojola

What do dogs prefer to eat?

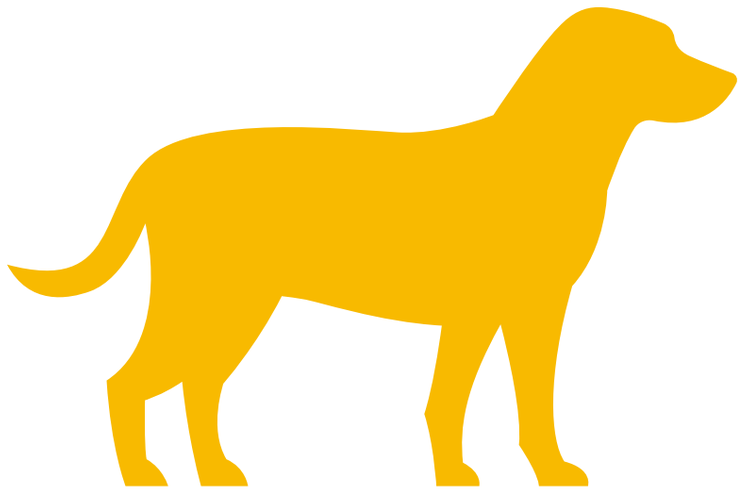


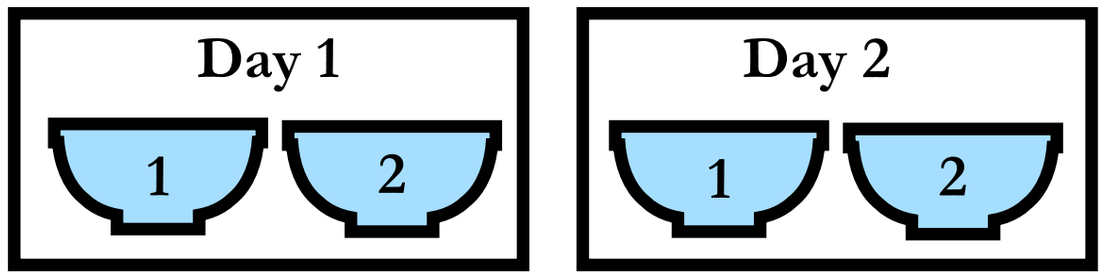


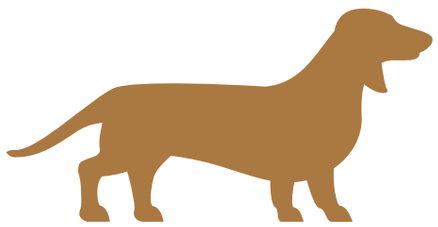


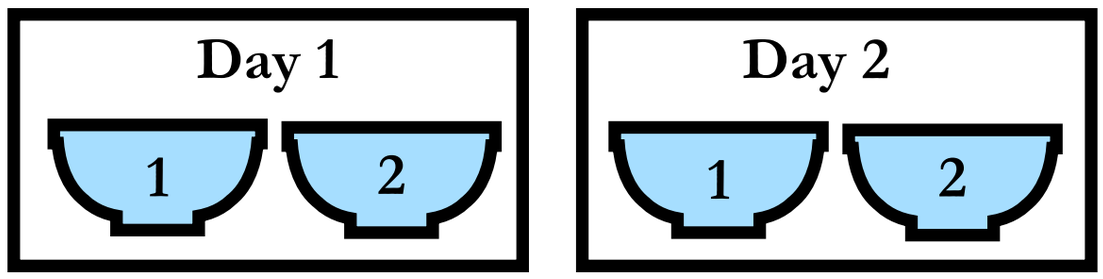


Consistency of choice





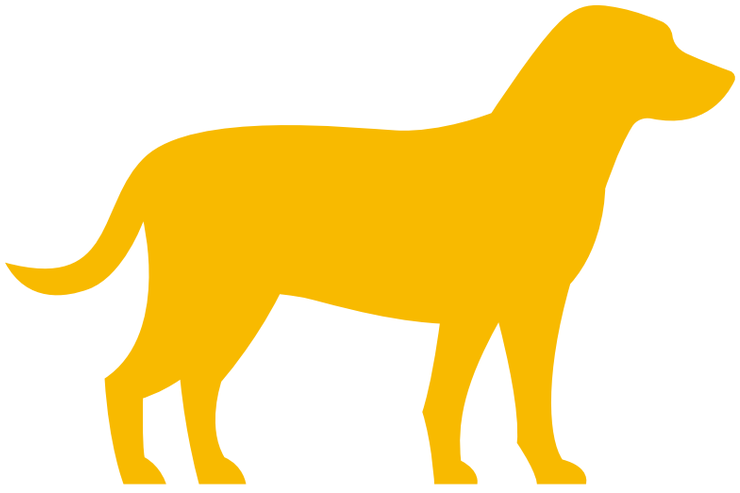


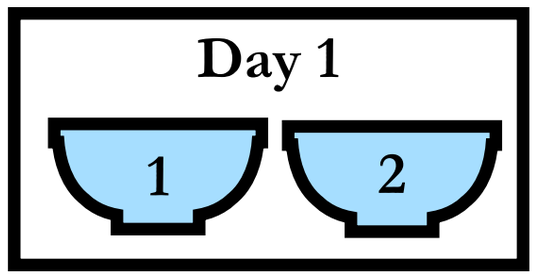


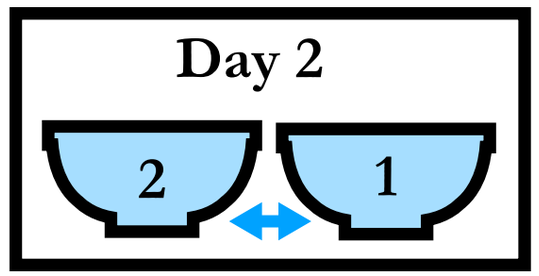
Could a dog prefer a side more than the food?

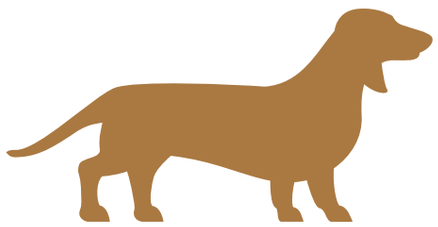
- **Side dominance ("handedness")**
- **Training**
- **Environmental factors**
- **Social factors**

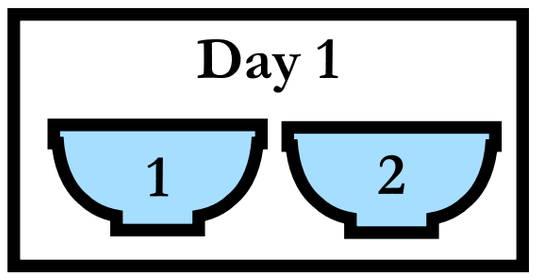
Adapting to potential side preference

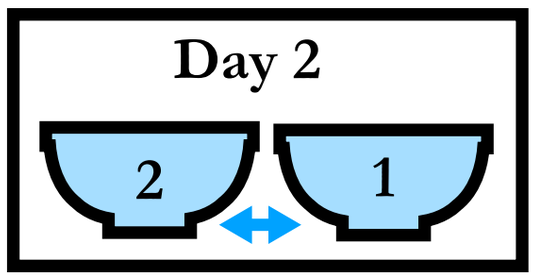




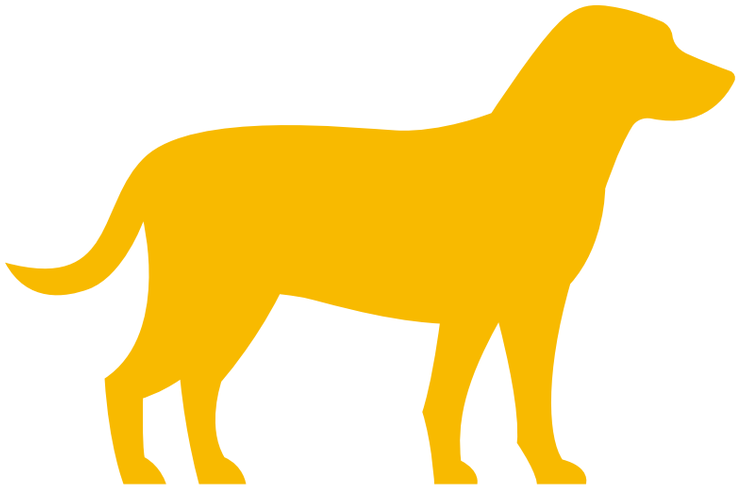


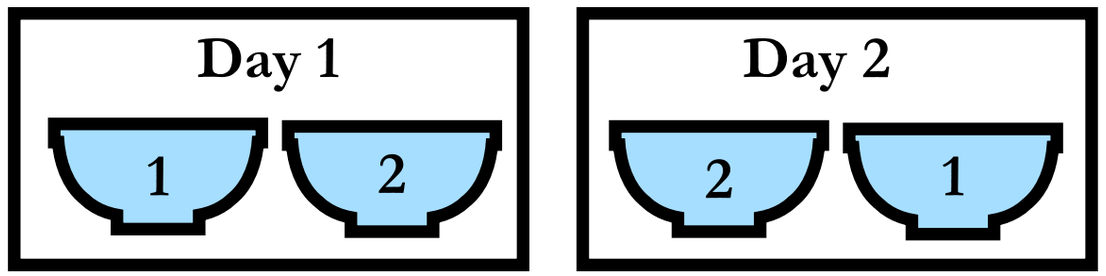


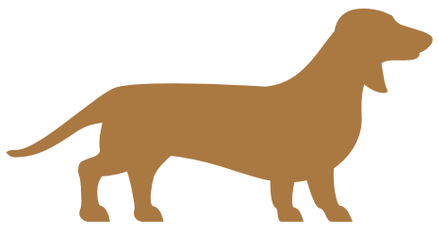


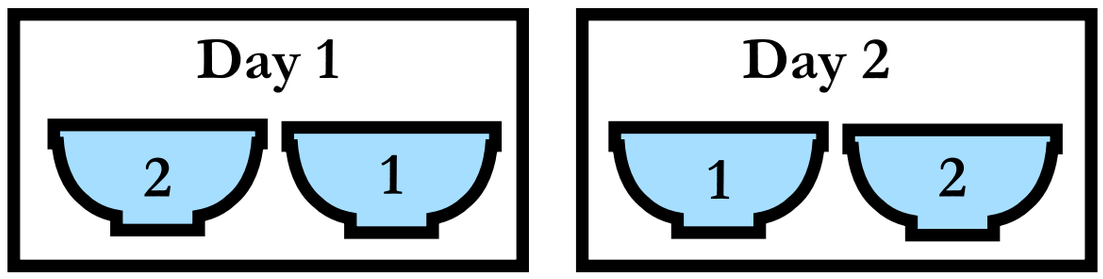


Why not perform a crossover study?





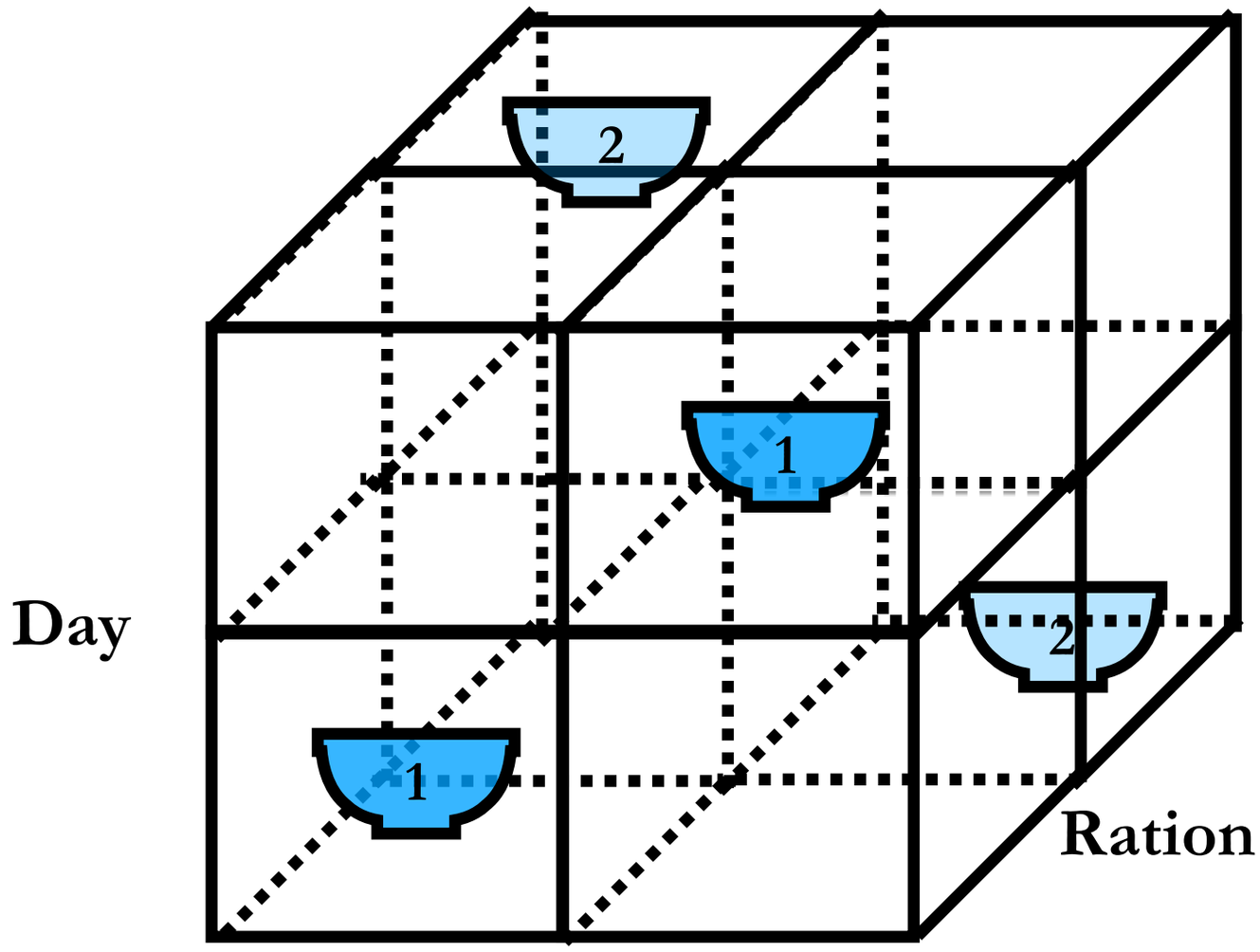




Experimental design

- Two rations in two bowls
- Two days
- Ration 1 on left on Day 1, on right on Day 2
- Typically 15-20 dogs
- Responses in proportions, not weights

A very incomplete block design



Typical analysis

- For each dog, calculate the intake ratio = (total weight of food consumed from Ration 1 over two days) / (total weight of food consumed over two days).
- Why not use the average of (fraction of food from Ration 1 on first day) and (fraction of food from Ration 2 on second day)?

Typical analysis

- Use a familiar statistical test (t-test, sign test, signed rank test) to determine whether the mean or median intake ratio is significantly different from 1/2.

Drawbacks of typical analysis

- Failure to reject H_o: Mean = 1/2 is often incorrectly interpreted as equality of preference.
- A dog with a strong side preference will pull the mean toward 1/2.
- Normality and symmetry are suspect.
- With small sample sizes, only strong preferences will be identified.

Alternative analysis strategy

Assume that Ration 1 was placed in the bowl on the left on Day 1 for each dog i.

p_{ijk} = Proportion of food eaten by Dog i on Day j from Ration k, with $p_{i11} + p_{i21} + p_{i12} + p_{i22} = 1$

Goal: To assess food preference, make inferences about the mean for $p_{i11} + p_{i21}$

Analyzing proportions as compositions

Aitchison, J., 1982. The statistical analysis of compositional data. Journal of the Royal Statistics Society, Series B 44, 139-177.

Aitchison, J., 1994. Principles of compositional data analysis. Multivariate Analysis and Its Applications 24, 73-81.

Main idea: To deal with the dependence among m proportions in a composition, transform the data using {ln (p_1/p_m), ln (p_2/p_m), …, ln (p_{m-1}/p_m)}, then apply ordinary multivariate procedures

Why not Dirichlet?

The Dirichlet distribution is an extension of the beta distribution for 3 or more categories.

Aitchison: "A major obstacle to its use in the statistical analysis of compositional data is that it seldom, if ever, provides an adequate description of actual variability of compositions."

Analyzing proportions as compositions

When transforming the proportions {ln (p_1/p_m), ln (p_2/p_m), …, ln (p_{m-1}/p_m)}, the choice of the "anchor" component p_m does not matter.

Analyzing proportions as compositions

Instead of $\{\ln(p_1/p_m), \ln(p_2/p_m), \ldots, \ln(p_{m-1}/p_m)\}$, transform the feeding experiment data using:

$X_{i1} = \ln[(p_{i11} + p_{i21}) / (1 - (p_{i11} + p_{i21}))]$ for intake ratio in favor of Ration 1,

$X_{i2} = \ln[(p_{i11} + p_{i22}) / (1 - (p_{i11} + p_{i22}))]$ for left side preference, and

$X_{i3} = \ln[(p_{i11} + p_{i12}) / (1 - (p_{i11} + p_{i12}))]$ for Day 1 appetite,

using p_{i11} as the anchor quantity in all three transformations.

Analyzing proportions as compositions

Why use $X_{i3} = \ln [(p_{i11} + p_{i12}) / (1- (p_{i11} + p_{i12}))]$ for Day 1 appetite?

- **Needed to complete the upcoming model**
- **Consider impact of consecutive feeding trials**

Analyzing proportions as compositions

Ideally:

Mean(X_{i1}) is different from 0 (indicating preference),

Mean(X_{i2}) is close to 0 (indicating no severe impacts of environment on side preference), and

Mean(X_{i3}) is close to 0 (indicating consistency of appetite)

Roadblock

$$X_{i1} = \ln [(p_{i11} + p_{i21}) / (1 - (p_{i11} + p_{i21}))]$$

$$X_{i2} = \ln [(p_{i11} + p_{i22}) / (1 - (p_{i11} + p_{i22}))]$$

$$X_{i3} = \ln [(p_{i11} + p_{i12}) / (1 - (p_{i11} + p_{i12}))]$$

$$p_{i11} + p_{i21} + p_{i12} + p_{i22} = 1$$

Can we back-transform and express each p_{ijk} as a unique function of X_{i1}, X_{i2}, and X_{i3}?

Roadblock

$p_{ijk} = f(X_{i1}, X_{i2}, X_{i3})$

Why would this be useful?

- **Other summaries, e.g. acquired taste**

Roadblock

Ryacas - R package interface to yacas (Yet Another Computer Algebra System)

```
> library(Ryacas)
> exp1 <- expression(x^2 + sin(x))
> exp2 <- expression(exp1/(log(x)))
> print(yacas(exp2))
expression((x^2 + sin(x))/log(x))
```

and lots of warning messages

Solution!

Use wolframalpha.com

Reduce[{x == Log[(p + q)/(1 - p - q)], y == Log[(p + r)/(1 - p - r)], z == Log[(p + s)/(1 - p - s)], p + q + r + s == 1}, {p, q, r, s}]

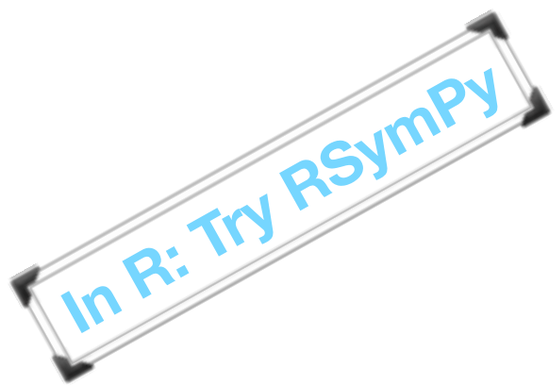
Solution over the reals:

$$p = \frac{2\,e^{x+y+z} + e^{x+y} + e^{x+z} + e^{y+z} - 1}{2\,(e^x + 1)\,(e^y + 1)\,(e^z + 1)} \wedge q = \frac{e^{x+y} + e^{x+z} + 2\,e^x - e^{y+z} + 1}{2\,(e^x + 1)\,(e^y + 1)\,(e^z + 1)}$$

$$r = \frac{e^{x+y} - e^{x+z} + e^{y+z} + 2\,e^y + 1}{2\,(e^x + 1)\,(e^y + 1)\,(e^z + 1)} \wedge s = -\frac{e^{x+y} - e^{x+z} - e^{y+z} - 2\,e^z - 1}{2\,(e^x + 1)\,(e^y + 1)\,(e^z + 1)}$$

Another solution

```
from sympy import symbols, log
pi11, pi12, pi21, pi22 = symbols('pi11 pi12 pi21 pi22')
Xi1,Xi2,Xi3 = symbols('Xi1 Xi2 Xi3')
from sympy import solve
#Equations are written as expressions that should be equal to zero
Eq1=Xi1 - log ((pi11 + pi21) / (1- (pi11 + pi21)))
Eq2=Xi2 - log ((pi11 + pi22) / (1- (pi11 + pi22)))
Eq3=Xi3 - log ((pi11 + pi12) / (1- (pi11 + pi12)))
Eq4=1-pi11-pi12-pi21-pi22
sols=solve([Eq1,Eq2,Eq3,Eq4],[pi11, pi12, pi21, pi22])
print(sols)

sols[pi11]
print("pi11 = ", sols[pi11])
print("pi12 = ", sols[pi12])
print("pi21 = ", sols[pi21])
print("pi22 = ", sols[pi22])
```



Another solution

```
print("pi11 = ", sols[pi11])
pi11 =  (exp(Xi1 + Xi2)/2 + exp(Xi1 + Xi3)/2 + exp(Xi2 + Xi3)/2 +
exp(Xi1 + Xi2 + Xi3) - 1/2)/(exp(Xi1) + exp(Xi2) + exp(Xi3) + exp(Xi1
+ Xi2) + exp(Xi1 + Xi3) + exp(Xi2 + Xi3) + exp(Xi1 + Xi2 + Xi3) + 1)

print("pi12 = ", sols[pi12])
pi12 =  (exp(Xi3) - exp(Xi1 + Xi2)/2 + exp(Xi1 + Xi3)/2 + exp(Xi2 +
Xi3)/2 + 1/2)/(exp(Xi1) + exp(Xi2) + exp(Xi3) + exp(Xi1 + Xi2) +
exp(Xi1 + Xi3) + exp(Xi2 + Xi3) + exp(Xi1 + Xi2 + Xi3) + 1)

print("pi21 = ", sols[pi21])
pi21 =  (exp(Xi1) + exp(Xi1 + Xi2)/2 + exp(Xi1 + Xi3)/2 - exp(Xi2 +
Xi3)/2 + 1/2)/(exp(Xi1) + exp(Xi2) + exp(Xi3) + exp(Xi1 + Xi2) +
exp(Xi1 + Xi3) + exp(Xi2 + Xi3) + exp(Xi1 + Xi2 + Xi3) + 1)

print("pi22 = ", sols[pi22])
pi22 =  (exp(Xi2) + exp(Xi1 + Xi2)/2 - exp(Xi1 + Xi3)/2 + exp(Xi2 +
Xi3)/2 + 1/2)/(exp(Xi1) + exp(Xi2) + exp(Xi3) + exp(Xi1 + Xi2) +
exp(Xi1 + Xi3) + exp(Xi2 + Xi3) + exp(Xi1 + Xi2 + Xi3) + 1)
```

Another roadblock

Data from a feeding trial to compare two commercial dog food brands

	Day 1 Ration 1 Left	Day 1 Ration 2 Right	Day 2 Ration 1 Right	Day 2 Ration 2 Left
Dog 1	0.0 g	100.7 g	0.0 g	100.2 g

$p_{111} + p_{121} = 0$, but $\ln(0) = ???$

Dealing with zeroes in composition data

Aitchison: "If there are only a few zeros of the no-trace type then replacement by positive values smaller than the traceable amounts will allow an analysis."

The dog food scales can record weights to 0.1 g. Is it reasonable to substitute a small number for a 0?

Dealing with zeroes in composition data

Aitchison: "In such circumstances it will always be wise to perform a sensitivity analysis to determine the effect that different zero replacement values have on the conclusions of the analysis."

R package zCompositions

zCompositions (Palarea-Albaladejo & Martin-Fernández) can substitute for zeroes and adjust remaining proportions downward to preserve the sum of 1.

Zero replacement methods include:

- Substitution of a fraction of the detection limit
- Multiple imputation
- Replacement using an interval-censoring model

Bayesian multivariate model

For each dog i, assume that the three logits (X_{i1}, X_{i2}, X_{i3}) follow a multivariate normal distribution with 3 mean parameters, 3 variance parameters, and 3 correlation parameters. Assume that the dogs are independent.

Why Bayes?

- **Dealing with small sample sizes**
- **Opportunity for exposure in another field**

Informative prior distributions

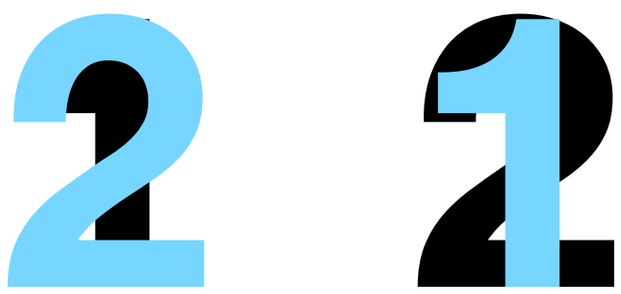
Data from 789 previous two-day, two-bowl tests with dogs conducted within the previous year were available.

1. For each test, transform each dog's data to (X_{i1}, X_{i2}, X_{i3}), replacing zeroes with simple replacement of 0.05 grams.
2. Calculate the sample means, variances, and correlations of (X_{i1}, X_{i2}, X_{i3}) in each test.
3. Use histograms of the 789 summary statistics for each parameter to decide on a prior distribution.

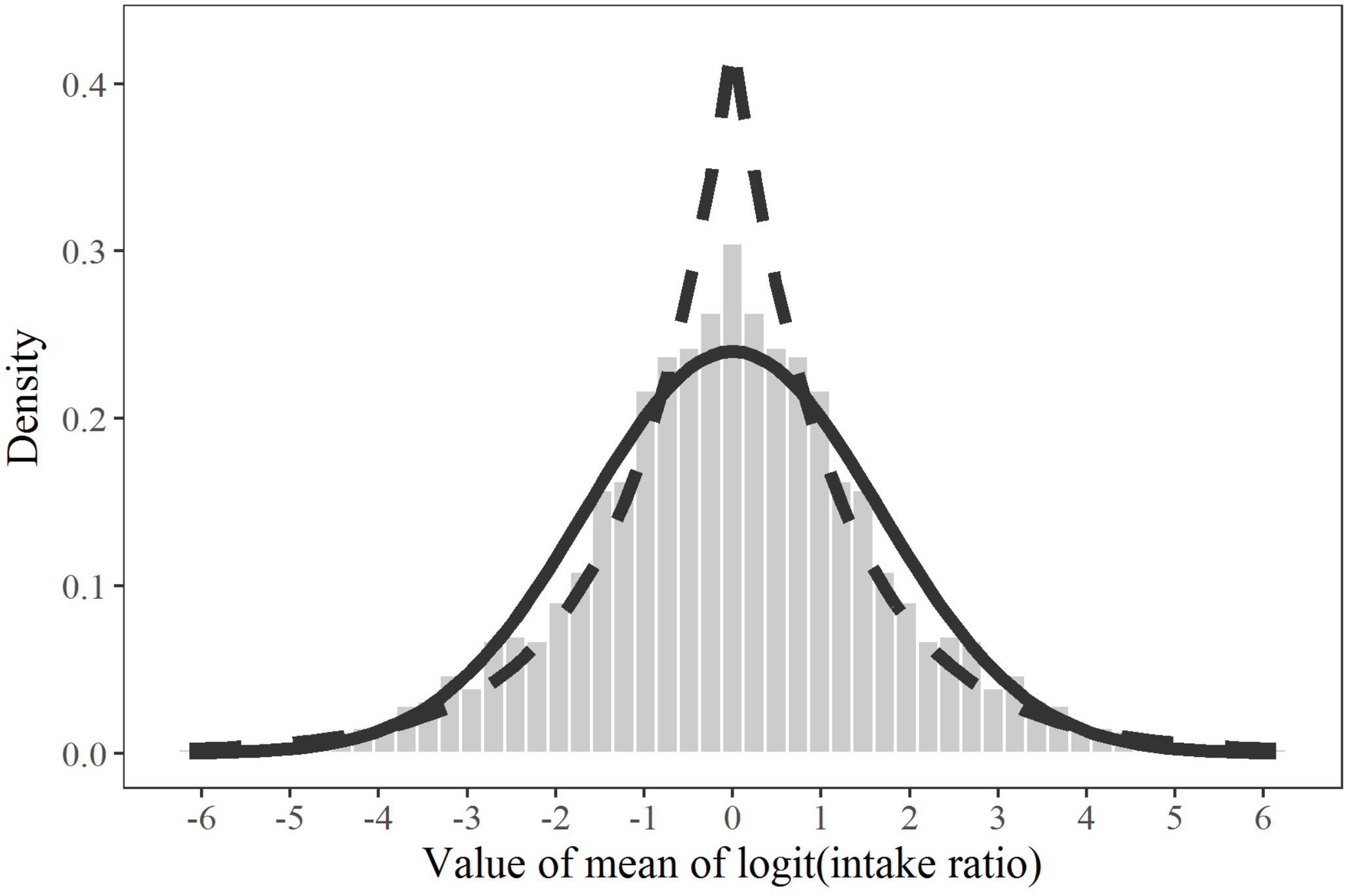
Roadblock

Could these summary statistics from previous experiments be biased?

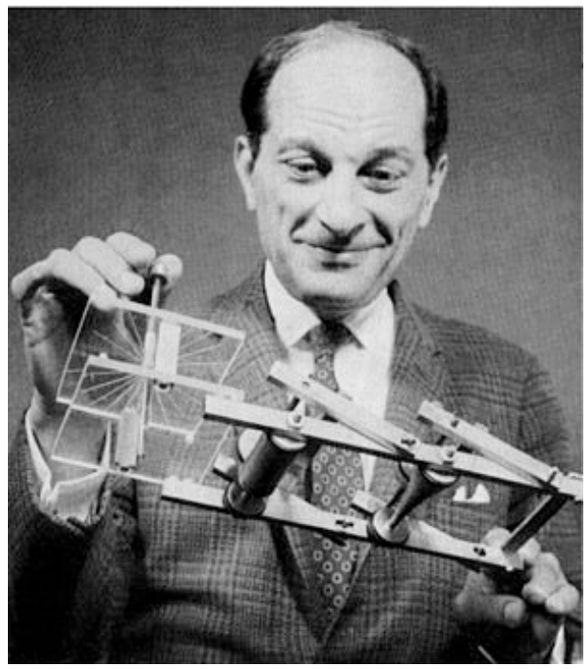
Solution:



Example: Prior distribution for mean of logit(intake ratio)



Stan



See https://mc-stan.org

- "Sampling Through Adaptive Neighborhoods"
- Stanislaw Ulam

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

Number 247 SEPTEMBER 1949 *Volume 44*

THE MONTE CARLO METHOD

NICHOLAS METROPOLIS AND S. ULAM
Los Alamos Laboratory

- Based on Hamiltonian Monte Carlo
- Uses No U-Turn Sampling (NUTS)

Compared to WinBUGS, Stan …

- Seems faster
- Seems to converge more quickly
- Seems easier to program
- Allows direct specification of log-likelihood
- Has default prior distributions for parameters
- Has associated packages in R that simplify model construction and posterior evaluation
- Is integrated into the tidyverse (tidybayes) and Shiny (shinystan)

But the best thing is…

- In Stan, normal distributions are specified like this:

y ~ normal (mean, std. deviation)

R script

```
library(pacman)

p_load(tidyverse, zCompositions, HDInterval,
rstan, shiny)

#Shiny must be Version 1.1 or later
 p_load(shinystan)
```

But I wanna play Pac Man!

https://github.com/RLesur/Rcade

R script

```
#Import the raw consumption data.
#Values represent grams of food consumed.
bothdays <- read.table(
  header = TRUE,
  text = “
   Dog  Day1Ration1  Day1Ration2  Day2Ration1  Day2Ration2
   1          0        100.7           0        100.2
  (etc.)
   24       135.7        150.3         34.7        150.3
   “
)
```

R script

Convert to proportions with a sum of 1

```
bothdays <- bothdays %>% mutate(
  PDay1Ration1 = Day1Ration1 / (Day1Ration1 +
Day1Ration2 + Day2Ration1 + Day2Ration2),
  PDay1Ration2 = Day1Ration2 / (Day1Ration1 +
Day1Ration2 + Day2Ration1 +
    Day2Ration2),
  PDay2Ration1 = Day2Ration1 / (Day1Ration1 +
Day1Ration2 + Day2Ration1 +
    Day2Ration2),
  PDay2Ration2 = 1 - PDay1Ration1 - PDay1Ration2
- PDay2Ration1
)
```

R script

```
forzcomp <- bothdays %>% dplyr::select(PDay1Ration1:PDay2Ration2)
checkfor0 <-
  forzcomp %>% summarise(
    Z11 = sum(PDay1Ration1 == 0),
    Z12 = sum(PDay1Ration2 == 0),
    Z21 = sum(PDay2Ration1 == 0),
    Z22 = sum(PDay2Ration2 == 0)
  )

if (checkfor0$Z11 == 0 &
    checkfor0$Z12 == 0 &
    checkfor0$Z21 == 0 & checkfor0$Z22 == 0)
  imputed <- forzcomp
if (checkfor0$Z11 > 0 |
    checkfor0$Z12 > 0 | checkfor0$Z21 > 0 | checkfor0$Z22 > 0) {
  bothdays <-
    bothdays %>% mutate(Total = Day1Ration1 + Day1Ration2 +
Day2Ration1 + Day2Ration2)
```

Check to see if there are any zeroes that need to be replaced

If not, let the "imputed" data frame contain the original proportion data

R script

```
#Use 0.1 grams as the lowest detectable scale weight.
detectionlimits <-
  matrix(rep(.1 / bothdays$Total, 4),
    ncol = 4,
    byrow = FALSE)
#Use the maximum likelihood estimates with a multiplicative
#lognormal model for left-censoring.
 imputed <-
   multLN(
    forzcomp,
    label = 0,
    detectionlimits,
    rob = FALSE,
    random = FALSE
   )
#Alternatively, use simple replacement based on one-half of the
#lowest detectable scale weight.
  #imputed <- multRepl(forzcomp,label=0,detectionlimits,delta=0.5)
}
```

Each dog has different limits based on its total amount of food eaten

Substitute proportions of 0 with small, reasonable values using zCompositions

Comparison of original data (forzcomp) and imputed data (imputed)

	Original				Imputed			
Dog	D1R1	D1R2	D2R1	D2R2	D1R1	D1R2	D2R1	D2R2
1	0.000	0.501	0.000	0.499	0.0002	0.5011	0.0001	0.4986
etc.								
24	0.288	0.319	0.074	0.319	0.288	0.319	0.074	0.319

R script

```
#Use a logarithmic transformation for the
#three quantities of interest
logimputed <- imputed %>% mutate(
  logit.ir = log((PDay1Ration1 + PDay2Ration1) /
(PDay1Ration2 + PDay2Ration2)),
  logit.left = log((PDay1Ration1 + PDay2Ration2) /
(PDay2Ration1 + PDay2Ration1)),
  logit.app = log((PDay1Ration1 + PDay1Ration2) /
(PDay2Ration1 + PDay2Ration2))) %>%
  dplyr::select(logit.ir, logit.left, logit.app)

#Begin preparing data for analysis with Stan.
X <- as.matrix(logimputed)
sbdata <- list(X = X,
  N = dim(X)[1],
  K = dim(X)[2])
```

R script

```
rstan_options(auto_write = TRUE)
options(mc.cores = parallel::detectCores())
```

R script

Following Stan code is based on:

https://github.com/stan-dev/example-models/blob/master/Bayesian_Cognitive_Modeling/ParameterEstimation/DataAnalysis/Correlation_1_Stan.R

R script

```
#Stan programming code to specify model
#and prior distributions.
stancode <- "
data {
int<lower=0> N;
int<lower=0> K;
vector[K] X[N];
}
parameters {
vector[K] mu;
vector<lower=0>[K] sigma;
real<lower=0,upper=1> beta12;
real<lower=0,upper=1> beta13;
real<lower=0,upper=1> beta23;
}
```

Sections of the code are separated with { }.

This specifies the raw data X as a collection of N=24 vectors, each with K=3 elements

These betas will have beta distributions in (0,1). They will be transformed to correlations in (-1,1).

R script

```
transformed parameters{
cov_matrix[K] T;
for (h in 1:K){
T[h,h]=square(sigma[h]);
}
T[1,2]=(2*beta12-1)*sigma[1]*sigma[2];
T[2,1]=(2*beta12-1)*sigma[1]*sigma[2];
T[1,3]=(2*beta13-1)*sigma[1]*sigma[3];
T[3,1]=(2*beta13-1)*sigma[1]*sigma[3];
T[2,3]=(2*beta23-1)*sigma[2]*sigma[3];
T[3,2]=(2*beta23-1)*sigma[2]*sigma[3];
}
```

R script

```
model {
mu[1] ~ normal(0,1.7);
mu[2] ~ double_exponential(0,0.28);
mu[3] ~ double_exponential(0,0.13);
sigma[1] ~ gamma(9.4, 4.5);
sigma[2] ~ gamma(3.1, 2.3);
sigma[3] ~ gamma(1.1, 2.8);
beta12 ~ beta(12.1,12.1);
beta13 ~ beta(7.0,7.0);
beta23 ~ beta(3.7,3.7);
X~multi_normal(mu,T);
}"
```

Double exponential (Laplace) distributions instead of normal distributions for "peaked" priors

R script

```
fit <- stan(
  model_code = stancode,
  data = sbdata,
  control = list(adapt_delta = 0.99),
  iter = 7250,
  warmup = 1000,
  chains = 4,
  seed = 20180104
)

m <- data.frame(as.matrix(fit))
```

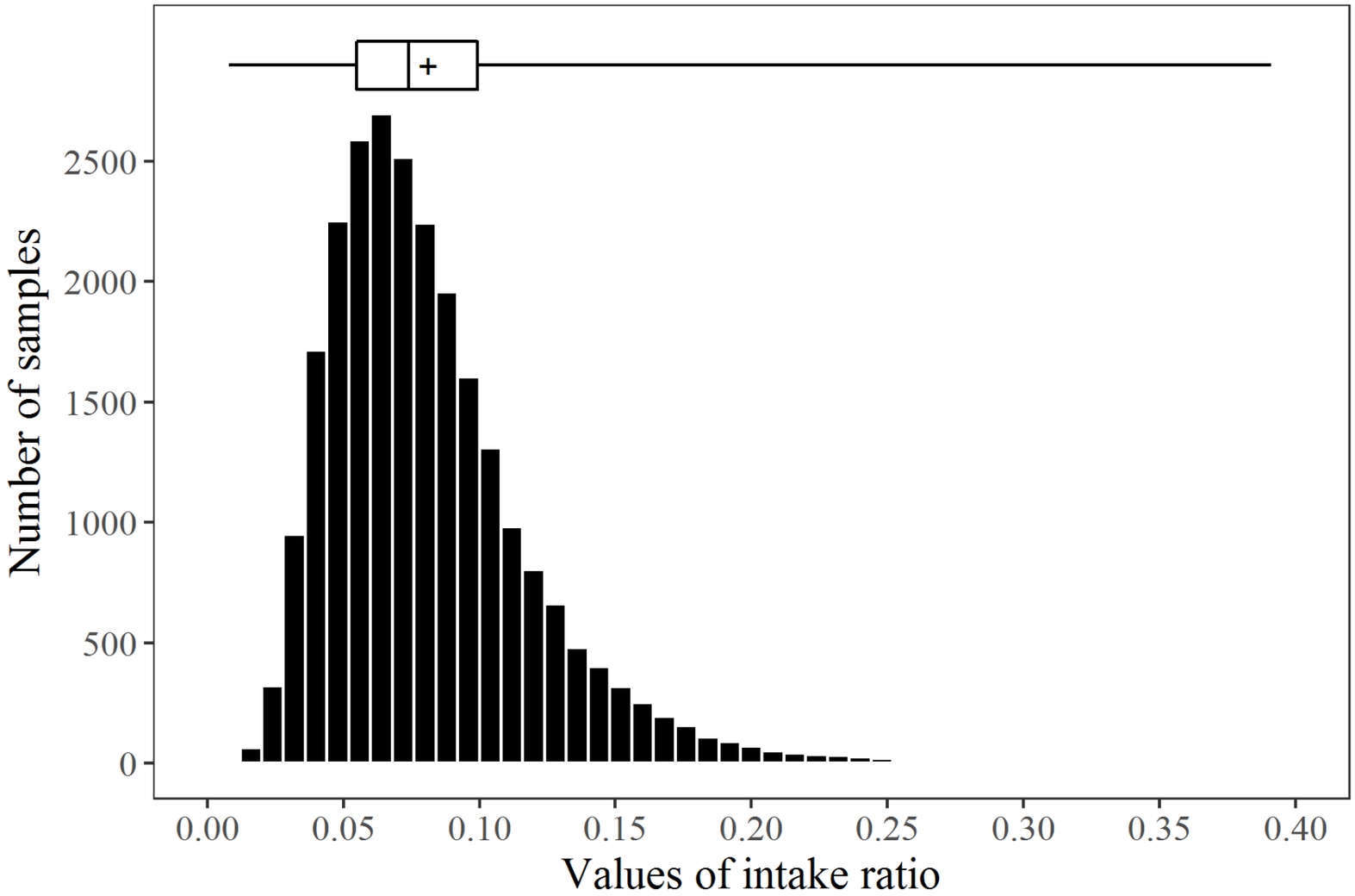
R script

```
#The data frame m has the samples from the
#posterior distributions for each parameter.

#Make back-transformations for proportions

invlogitfunc <- function(mu) {
  y <- (exp(mu)) / (1 + exp(mu))
  return(y)
}
m <- m %>% mutate(
  intakeratio = invlogitfunc(mu.1.),
  leftpreference = invlogitfunc(mu.2.),
  day1appetite = invlogitfunc(mu.3.)
)
```

R script

```
#Show a summary of the posterior
#distribution for the intake ratio.
hist(m$intakeratio)
summary(m$intakeratio)
sd(m$intakeratio)

#Calculate the highest density 95%
#credible interval.
hdi(m$intakeratio, credMass = 0.95)
```

Conclusions



The posterior mean for the (back-transformed) intake ratio was .081, with a 95% credible interval of (.023, .155).

Dogs strongly preferred Ration 2.

Conclusions

The posterior distributions for the mean left preference and mean Day 1 appetite had 0.5 near the center, indicating no significant side preference or difference in consumption between days.

R script

```
launch_shinystan(fit)
```

Not shown here

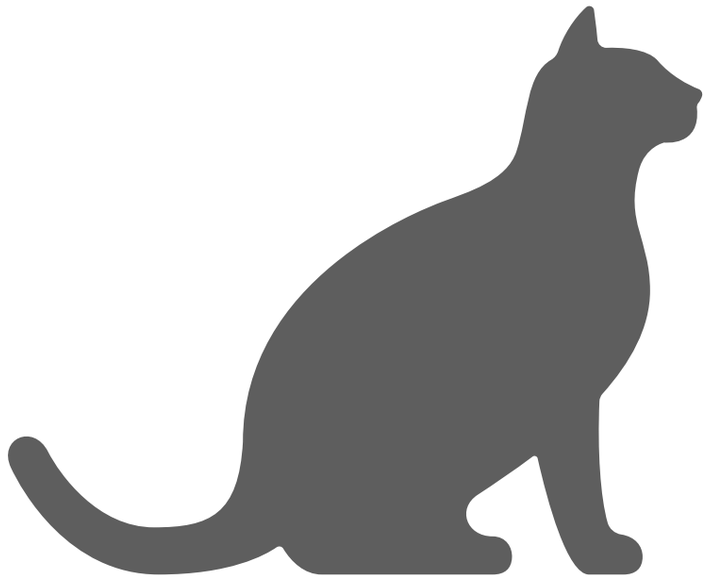
Posterior distributions for all 9 parameters

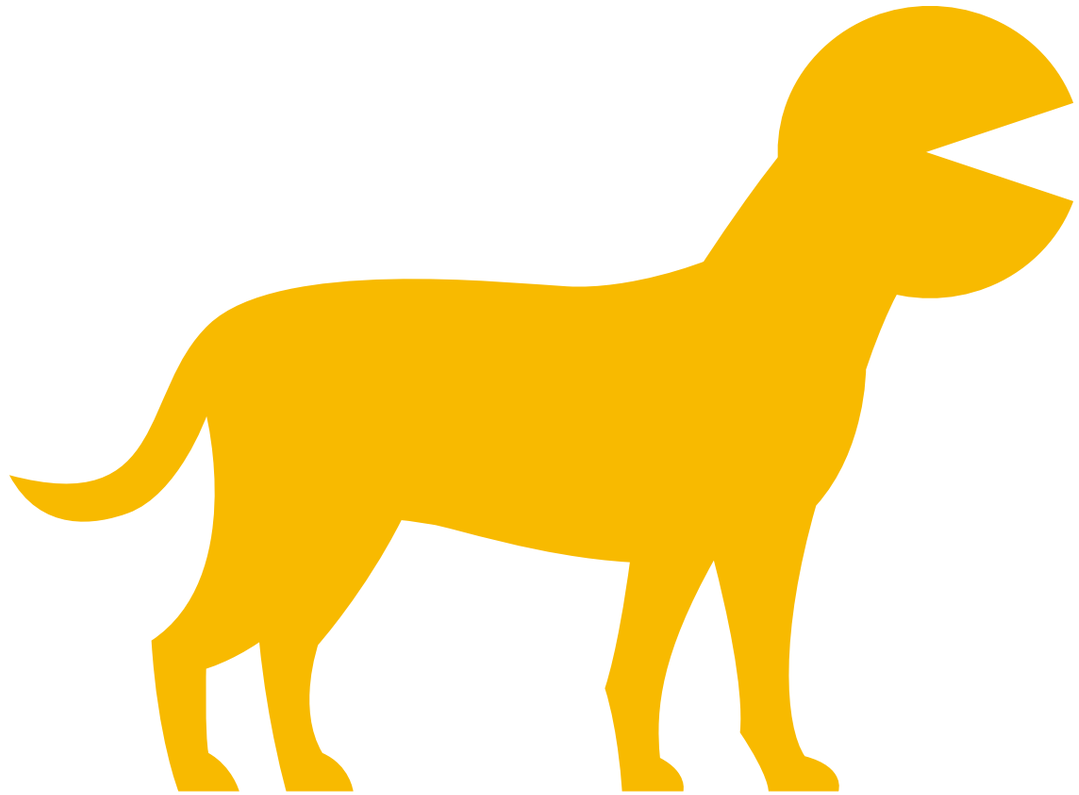
Sensitivity analyses using:

- **Informative priors**
- **Default priors (improper uniform over (-Inf, Inf) or (0,Inf) or uniform over (-1,1))**
- **Cauchy and half-Cauchy priors**
- **Simple zero replacement**
- **Censored lognormal zero replacement**

Things someone else can do

- Model zero proportions as interval-censored values within Stan, bypassing zCompositions
- Try a more rigorous hierarchical model
- Figure out what to do with cats





Thank you!

Jay.Harrison@ssmhealth.com