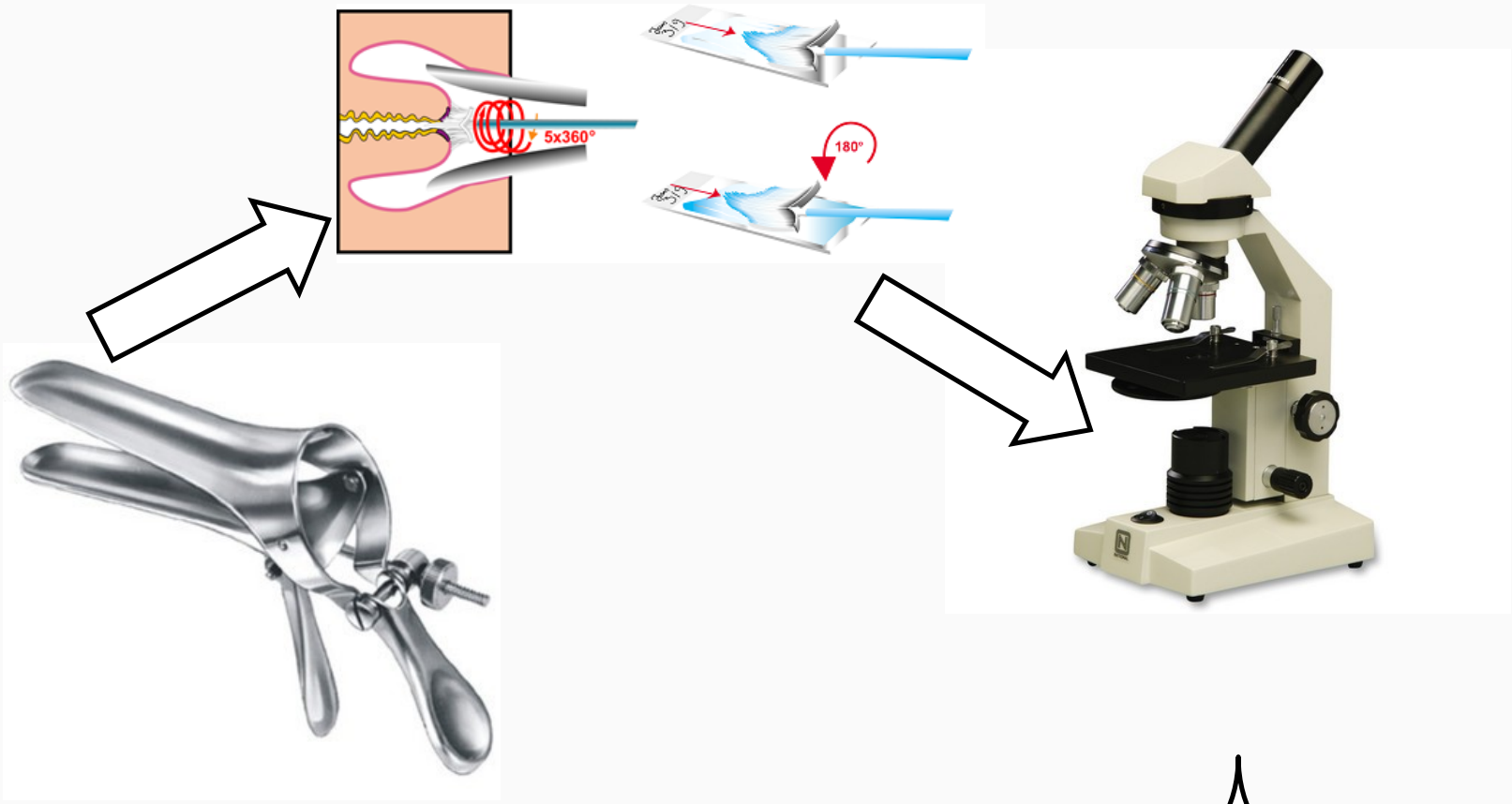
CopulaDTA: An R package for copula-based bivariate beta-binomial models for diagnostic test accuracy studies in Bayesian framework

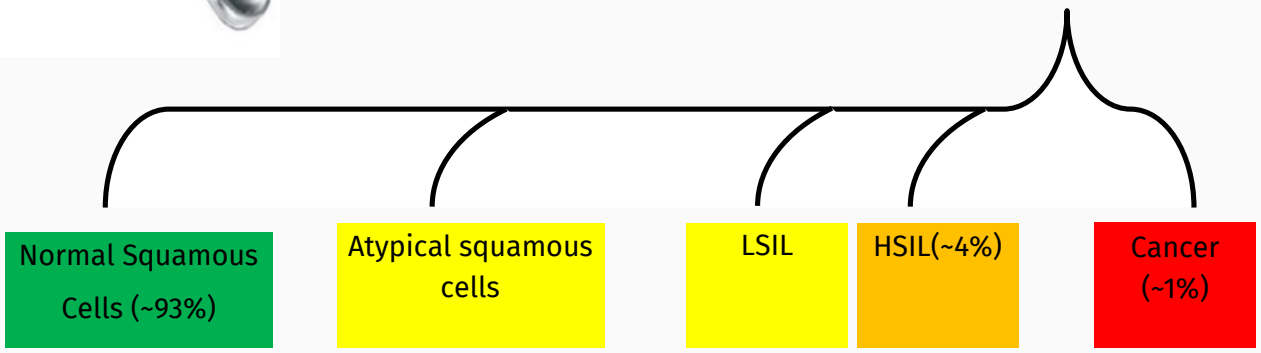
Victoria Nyaga[1,2], Marc Arbyn[2], Marc Aerts[1].

May 18, 2016

1. Universiteit Hasselt.
2. Scientific Institute of Public Health.

Introduction





Introduction

Triage group	Management	Implication(s)
Normal Squamous Cells (~93%)	Routine check-up every 3 years	
Atypical squamous cells	?Repeat pap smear ? HPV DNA Testing ?HPV Mrna testing ?...	♦ Prediction on natural history difficult ♦ Spontaneous regression ◊ Anxiety ◊ Invasive treatment ◊ Financial burden ◊ Discomfort ◊ Overdiagnosis ◊ Overtreatment
LSIL		
HSIL(~4%)	Further exploration and treatment	Prevention
Cancer (~1%)	Surgery Radiotherapy ...	Increased survival

Research interest

Diagnostic tests identifying women whose cervical lesions confer increased risk of cervical cancer.

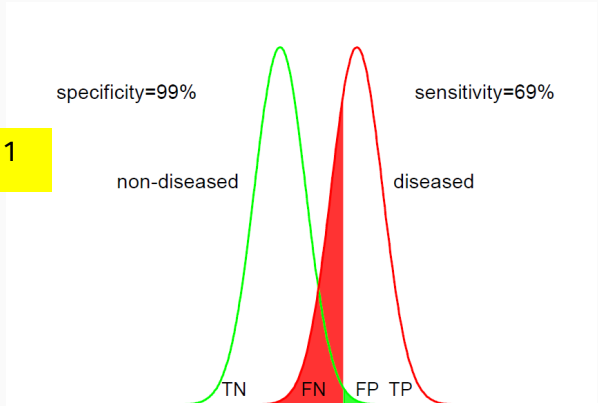
Arbyn et al. (2013)

Among ASCUS triage group determine;

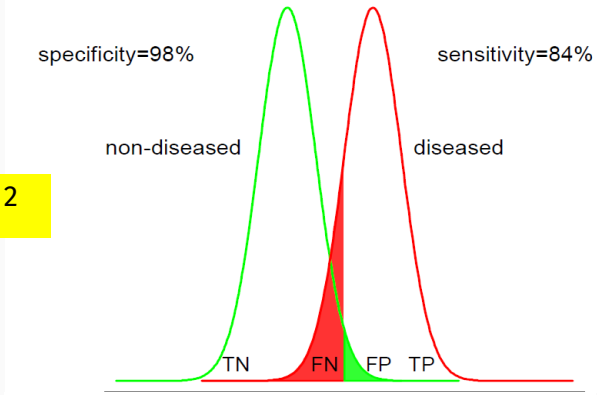
- Accuracy of
  1. HPV DNA Testing using HC2, and
  2. Repeat cytology in detecting CIN2+.
- Differences in accuracy between the two triage tests.

Hypothetical data

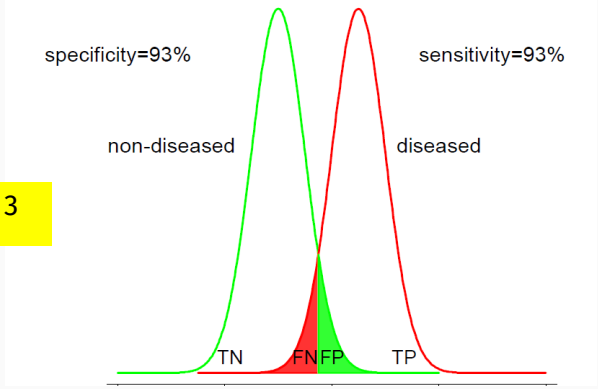
Study 1



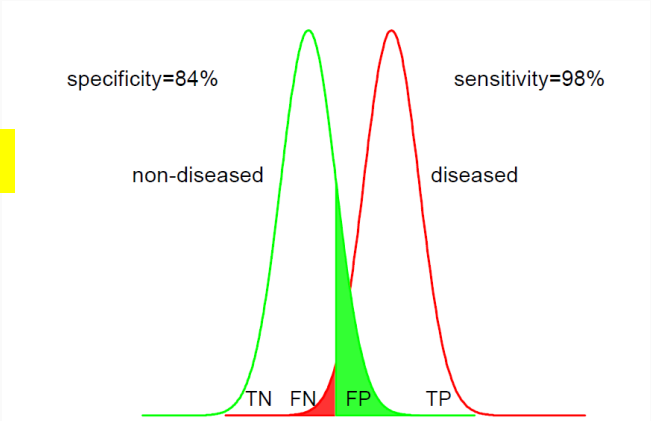
Study 2



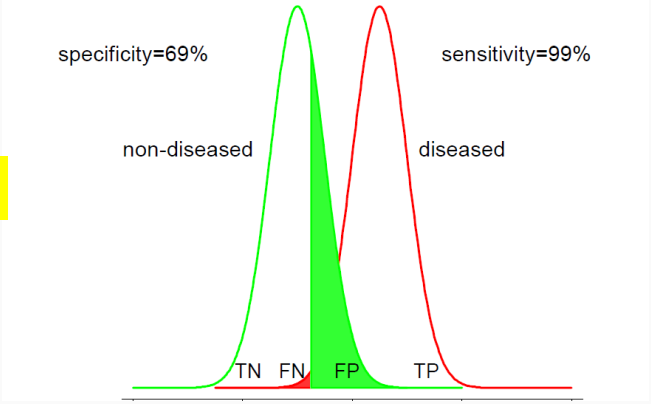
Study 3



Study 4



Study 5



... N

Goal

- Obtain population-averaged estimates.
- Account for the present intrinsic correlation between sensitivity and specificity.
- Obtain forest plots.

Outline

1. Bivariate normal vs. beta random-effects
2. The CopulaDTA package
3. Application

Bivariate normal vs. beta random-effects

$$TP_i \mid se_i,\ \mathbf{x}_i \ \sim\ bin(se_i,\ Dis_i), i\ =\ 1,\ \ldots\ N,$$

$$TN_i \mid sp_i,\ \mathbf{x}_i \ \sim\ bin(sp_i,\ NonDis_i), i\ =\ 1,\ \ldots\ N.$$

Normal distribution

$$\begin{pmatrix} logit(se_i) \\ logit(sp_i) \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{logitse} \\ \mu_{logitsp} \end{pmatrix}, \mathbf{\Sigma} \right)$$

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

Beta distribution

$$\begin{pmatrix} se_i \\ sp_i \end{pmatrix} \sim f(se_i)f(sp_i)c(F(se_i), F(sp_i), \theta)$$

$\mathrm{f}(\mathrm{se}_i) = Beta(\mu_{se}, \psi_{se})$,
$\mathrm{f}(\mathrm{sp}_i) = Beta(\mu_{sp}, \psi_{sp})$,
$\mathrm{c}(.)$ = copula density: frank, gaussian, ...
$\mu_. = logit^{-1}(\nu_.) = \frac{\alpha_.}{(\alpha_.+\beta_.)}$
$\psi_. = (\alpha_. + \beta_.)$ *or* $\frac{1}{1+\alpha_.+\beta_.}$

Bivariate normal vs. beta random-effects

Normal distribution

- Availability of software; SAS, R,...
- Approximative distribution:
  - Large sample size.
- $$\begin{pmatrix} \mu_{logitse} \\ \mu_{logitsp} \end{pmatrix}$$
  - Expit transformation
  - $$\begin{pmatrix} \mu_{se} | \epsilon_{logitse} = 0 \\ \mu_{sp} | \epsilon_{logitsp} = 0 \end{pmatrix}$$
  - $$\begin{pmatrix} \mu_{se} \\ \mu_{sp} \end{pmatrix} = \begin{pmatrix} E(logit^{-1}(\mu_{logitse} + \epsilon_{logitse})) \\ E(logit^{-1}(\mu_{logitsp} + \epsilon_{logitsp})) \end{pmatrix}$$

Beta bistribution

- Programming skills needed: R, JAGS, **STAN**, ...
- Natural choice
- $$\begin{pmatrix} \mu_{se} \\ \mu_{sp} \end{pmatrix}$$

CopulaDTA package

MCMC sampling engine

Stan

1. Easy to extend compared to JAGS(.dll).
   functions {}
   data {}
   transformed data {}
   parameters {}
   transformed parameters {}
   model {}
   generated quantities {}
2. Faster convergence with fewer iteration even with poor initial values than with JAGS.

CopulaDTA package

Functions, objects and methods

- **cdtamodel function → cdtamodel object**

  R > model < - cdtamodel(copula = 'fgm',
  + **modelargs**=list(formula.se = StudyID ∼ Test - 1,
  + formula.sp = StudyID ∼ Test - 1,
  + formula.omega = StudyID ∼ Test - 1,
  + param=2,
  + prior.lse='normal', par.lse1=0, par.lse2=5,
  + prior.lsp='normal', par.lsp1=0, par.lsp2=5,...))

- **fit function → cdtafit object**

  R >fitmodel <- fit(cdtamodel object, data, SID, cores=3,chains=3,
  + iter=6000, warmup=1000,thin=10, ...)

- **Methods for cdtafit object: print, summary, plot, str**

Application

Package Installation

```
R> install.packages("CopulaDTA", dependencies = TRUE)
R> library(CopulaDTA)
```

Data

```
R> data(ascus)
R> ascus
```

Test	StudyID	TP	FP	TN	FN
RepC	Anderson 2005	6	14	28	4
RepC	Bergeron 2000	8	28	71	4
RepC	Del Mistro 2010	20	191	483	7
.	.	.	.	.	.
HC2	Silverloo 2009	34	65	81	2
HC2	Solomon 2001	256	1050	984	11

Application

Model fitting

```
R> frank <- cdtamodel(copula = "frank",
+ modelargs = list(formula.se = StudyID ~ Test + 0))
R> fitfrank <- fit(frank,
+ data = ascus,
+ SID = "StudyID",
+ iter = 19000,
+ warmup = 1000,
+ thin = 20,
+ seed = 3)
```

Application

Posterior Estimates

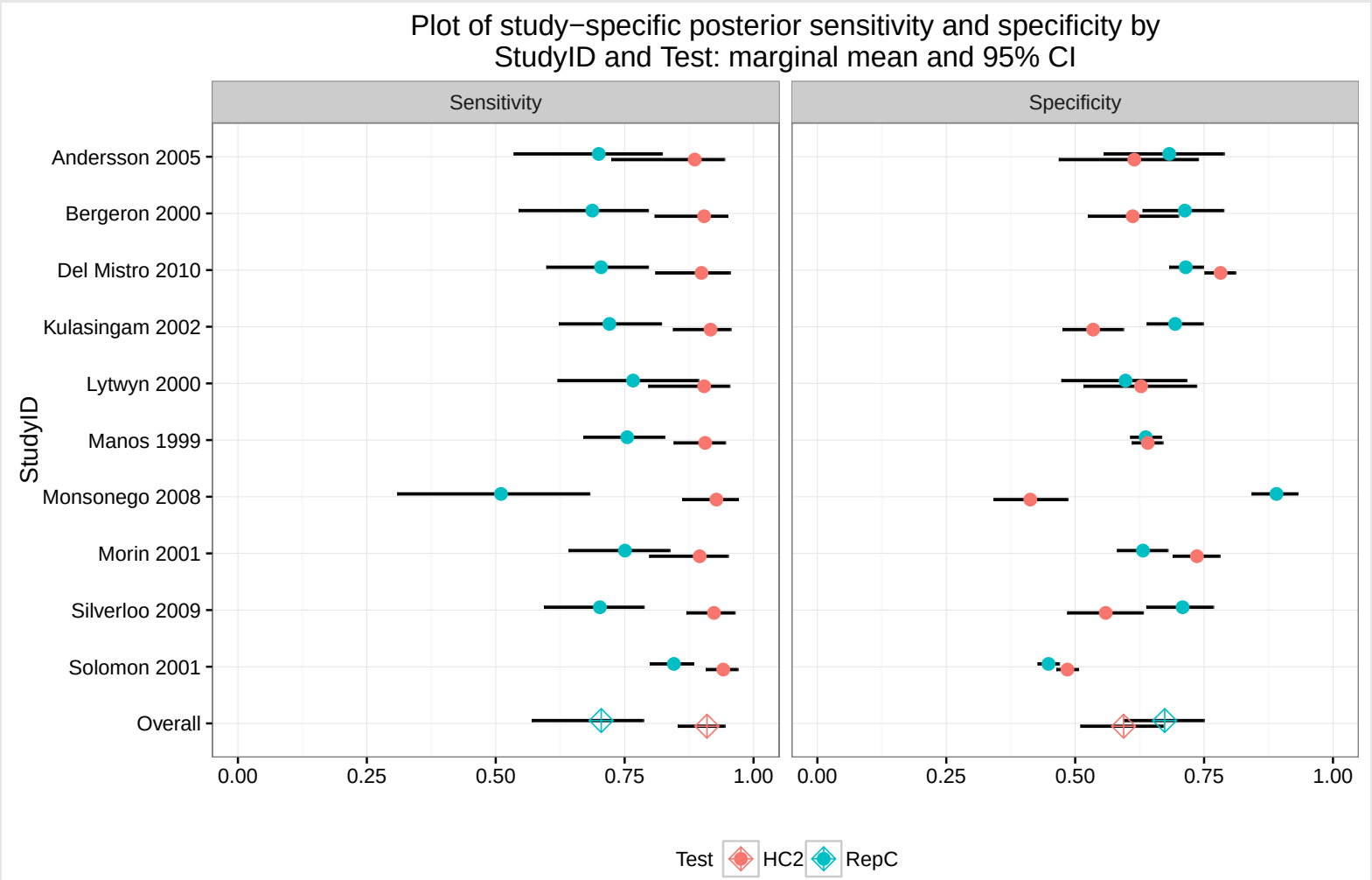
R > print(fitfrank, digits=4)

```
Posterior marginal mean sensitivity and specificity
        with 95% credible intervals

          Parameter    Mean   Lower   Upper   n_eff  Rhat
MUse[1] Sensitivity  0.9097  0.8530  0.9441  384.26 1.000
MUse[2] Sensitivity  0.7046  0.5701  0.7863   94.37 1.016
MUsp[1] Specificity  0.5939  0.5102  0.6711 1187.25 1.000
MUsp[2] Specificity  0.6736  0.5925  0.7497  680.34 1.007
RRse[1] Sensitivity  1.0000  1.0000  1.0000 2700.00   NaN
RRse[2] Sensitivity  0.7748  0.6312  0.8687  126.35 1.014
RRsp[1] Specificity  1.0000  1.0000  1.0000 2700.00   NaN
RRsp[2] Specificity  1.1392  0.9535  1.3558 1037.83 1.005
ktau[1] Correlation -0.5238 -0.8201  0.6067 2700.00   NaN
ktau[2] Correlation -0.7410 -0.8627 -0.3253 2700.00   NaN


Model characteristics

Copula function: frank, sampling algorithm: NUTS(diag_e)

Formula(1):  MUse ~ Test + 0
Formula(2):  MUsp ~ Test + 0
Formula(3):  Omega ~ Test + 0
3 chain(s)each with iter=19000; warm-up=1000; thin=20.
post-warmup draws per chain=900;total post-warmup draws=2700.

Predictive accuracy of the model

Log point-wise predictive density (LPPD): -2660.0999
Effective number of parameters: 16.4699
Watanabe-Akaike information Criterion (WAIC): 5353.1396
```

Application

Forest Plot

R > plot(fitfrank)



Summary

- Repeat cytology less sensitive than HC2 in diagnosing cervical precancer in women with equivocal pap smear.
- Marginal as well as study-specific estimates.
- Parameters with natural interpretation.

References

Arbyn M, Roelens J, Simoens C, Buntinx F, Paraskevaidis E, Martin-Hirsch PP, Prendiville WJ (2013).
"Human Papillomavirus Testing Versus Repeat Cytology for Triage of Minor Cytological Cervical Lesions."
Cochrane Database of Systematic Reviews, 31–201.

Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH (2005).
"Bivariate Analysis of Sensitivity And Specificity Produces Informative Summary Measures in Diagnostic Reviews."
Journal of Clinical Epidemiology, **58**(10), 982–990.

Nyaga VN (2015).
CopulaDTA: Copula Based Bivariate Beta-Binomial Model for Diagnostic Test Accuracy Studies.
R package version 0.0.2,
URL `http://CRAN.R-project.org/package=CopulaDTA`.

Good evening

&

Enlightening conference

Comparison

Parameter	Arbyn et al.(2013)*	Bayesian BRMA	BB (frank)
Sensitivity(HC2)	0.91[0.86, 0.94]	0.90[0.84, 0.95]	0.91[0.85, 0.94]
Specificity(HC2)	0.61[0.54, 0.68]	0.60[0.47, 0.72]	0.59[0.51, 0.67]
Sensitivity(RepC)	0.72[0.63, 0.79]	0.71[0.59, 0.80]	0.70[0.57, 0.79]
Specificity(RepC)	0.68[0.60, 0.76]	0.66[0.53, 0.78]	0.67[0.59, 0.75]

Posterior Estimates

```
Inference for Stan model: a502b880df243c0c4d60042c0ff01f04.
3 chains, each with iter=5000; warmup=1000; thin=10;
post-warmup draws per chain=400, total post-warmup draws=1200.

               mean se_mean     sd   2.5%  97.5% n_eff   Rhat
betamul[1,1] 2.3892  0.0090 0.2509 1.8846 2.8868   770 1.0014
betamul[1,2] 0.4404  0.0070 0.2063 0.0128 0.8608   865 1.0011
betamul[2,1] 0.9399  0.0073 0.2179 0.5022 1.3593   882 1.0013
betamul[2,2] 0.7649  0.0069 0.2088 0.3555 1.1983   914 0.9987
mu[1,1]      0.9140  0.0007 0.0199 0.8681 0.9472   764 1.0018
mu[1,2]      0.6073  0.0016 0.0487 0.5032 0.7028   871 1.0011
mu[2,1]      0.7170  0.0015 0.0441 0.6230 0.7957   884 1.0013
mu[2,2]      0.6807  0.0015 0.0449 0.5880 0.7682   925 0.9986
MU[1,1]      0.9033  0.0010 0.0273 0.8374 0.9446   723 1.0021
MU[1,2]      0.6008  0.0023 0.0619 0.4722 0.7193   756 1.0029
MU[2,1]      0.7076  0.0019 0.0544 0.5884 0.8017   837 0.9980
MU[2,2]      0.6644  0.0024 0.0618 0.5302 0.7753   683 1.0017

Samples were drawn using NUTS(diag_e) at Sat Jan 16 14:16:42 2016.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```