Common statistical tests as Bayesian models

- Most common statistical tests are linear models

t-test	mean of data	`stan_glm(y ~ 1)`
paired *t*-test	mean of diffs	`stan_glm((y1 - y2) ~ 1)`
Pearson correl.	linear model	`stan_glm(y ~ 1 + x)`
two-sample *t*-test	group means	`stan_glm(y ~ 1 + gid)`
ANOVA	hier. model	`stan_glm(y ~ 1 + (1 | gid))`

...

Common statistical tests as Bayesian models

- Most common statistical tests are linear models

t-test	mean of data	`stan_glm(y ~ 1)`
paired *t*-test	mean of diffs	`stan_glm((y1 - y2) ~ 1)`
Pearson correl.	linear model	`stan_glm(y ~ 1 + x)`
two-sample *t*-test	group means	`stan_glm(y ~ 1 + gid)`
ANOVA	hier. model	`stan_glm(y ~ 1 + (1 | gid))`

...

- Possible to extend, e.g., with group specific variances and and different distributions such *t*- or Poisson distribution
  - and go beyond named tests

Common statistical tests as Bayesian models

- Most common statistical tests are linear models

t-test	mean of data	`stan_glm(y ~ 1)`
paired *t*-test	mean of diffs	`stan_glm((y1 - y2) ~ 1)`
Pearson correl.	linear model	`stan_glm(y ~ 1 + x)`
two-sample *t*-test	group means	`stan_glm(y ~ 1 + gid)`
ANOVA	hier. model	`stan_glm(y ~ 1 + (1 | gid))`

...

- Possible to extend, e.g., with group specific variances and and different distributions such *t*- or Poisson distribution
  - and go beyond named tests
- See longer list and illustrations (with `lm`) at https://lindeloev.github.io/tests-as-linear/ and with rstanarm in Regression and other stories

Frequentist hypothesis testing

- Frequentist approach can be used to to make estimates and confidence intervals, but for some reason null hypothesis testing has had a very big role

Frequentist hypothesis testing

- Frequentist approach can be used to to make estimates and confidence intervals, but for some reason null hypothesis testing has had a very big role
  - reporting just the null hypothesis testing result throws away lot of useful information

Frequentist hypothesis testing

- Frequentist approach can be used to to make estimates and confidence intervals, but for some reason null hypothesis testing has had a very big role
  - reporting just the null hypothesis testing result throws away lot of useful information
  - some Bayesians are also into null hypothesis testing

Frequentist hypothesis testing

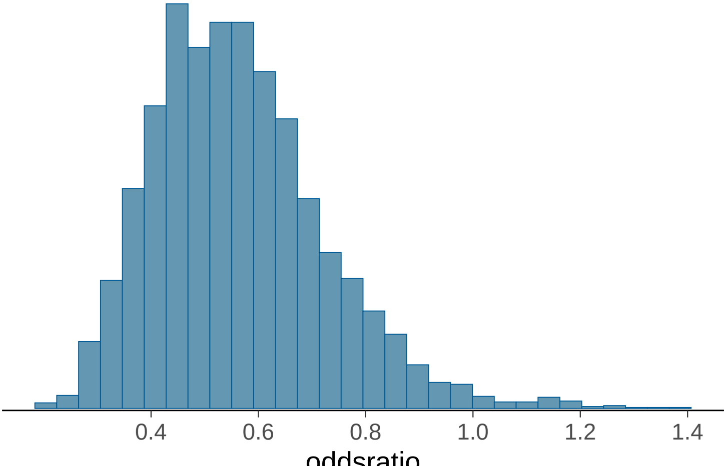
- Frequentist approach can be used to to make estimates and confidence intervals, but for some reason null hypothesis testing has had a very big role
  - reporting just the null hypothesis testing result throws away lot of useful information
  - some Bayesians are also into null hypothesis testing
- Frequentist null hypothesis testing
  - asks what if data is generated from the smaller model
  - doesn't tell whether the more complex model is good enough

Frequentist hypothesis testing

- Frequentist approach can be used to to make estimates and confidence intervals, but for some reason null hypothesis testing has had a very big role
  - reporting just the null hypothesis testing result throws away lot of useful information
  - some Bayesians are also into null hypothesis testing
- Frequentist null hypothesis testing
  - asks what if data is generated from the smaller model
  - doesn't tell whether the more complex model is good enough
- Some frequentists are now advocating looking at intervals and equivalence testing

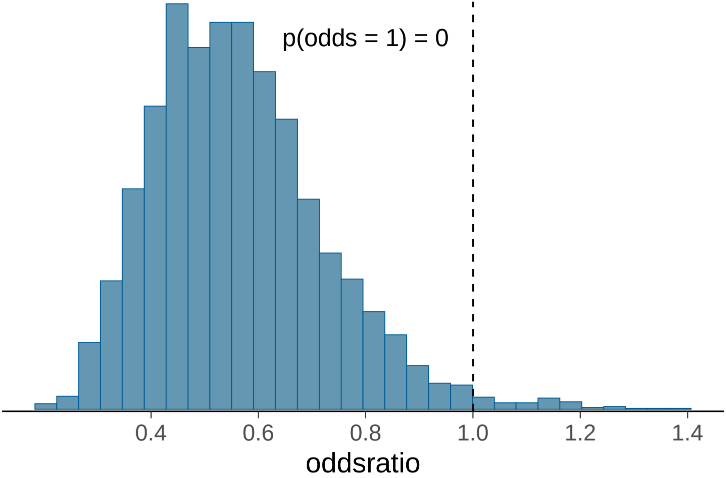
Bayesian hypothesis testing

- Instead of hypothesis testing, report full posterior and
  - compare to expert information
  - combine with utility/cost function



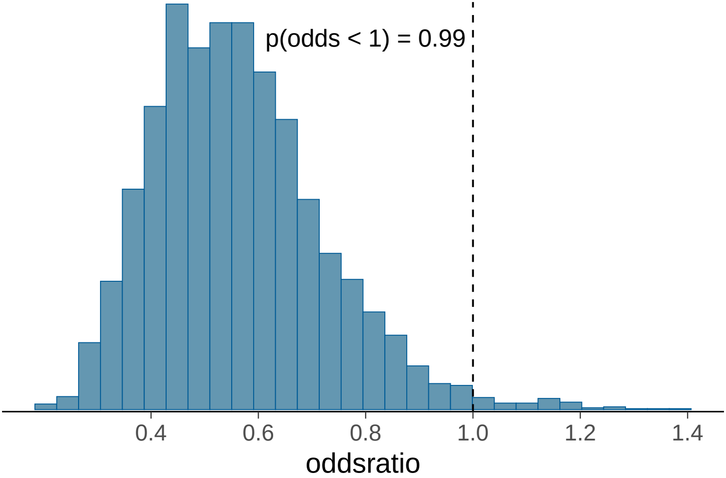
Bayesian hypothesis testing

- Instead of hypothesis testing, report full posterior
  - for continuous posterior there is zero probability that e.g. treatment effect is exactly zero



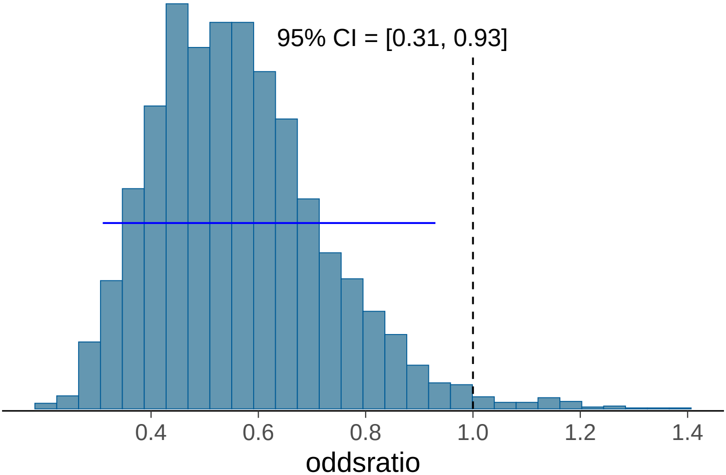
Bayesian hypothesis testing

- Instead of hypothesis testing, report full posterior
  - for continuous posterior we could compute the probability that we know the sign of the effect



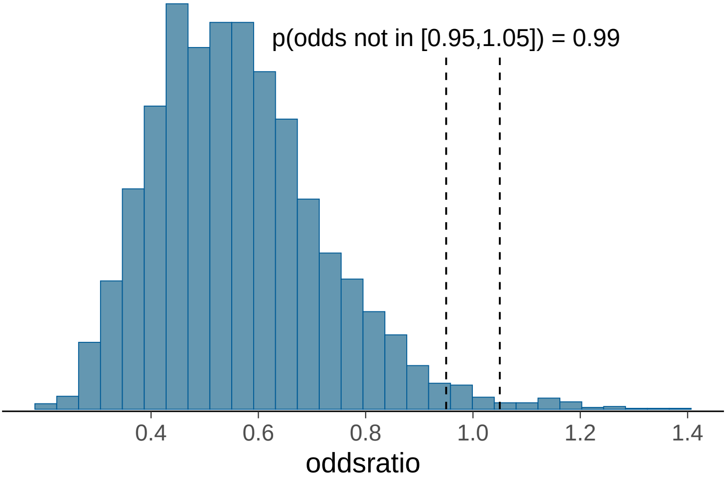
Bayesian hypothesis testing

- Instead of hypothesis testing, report full posterior
  - for continuous posterior some people compare whether posterior interval includes null case



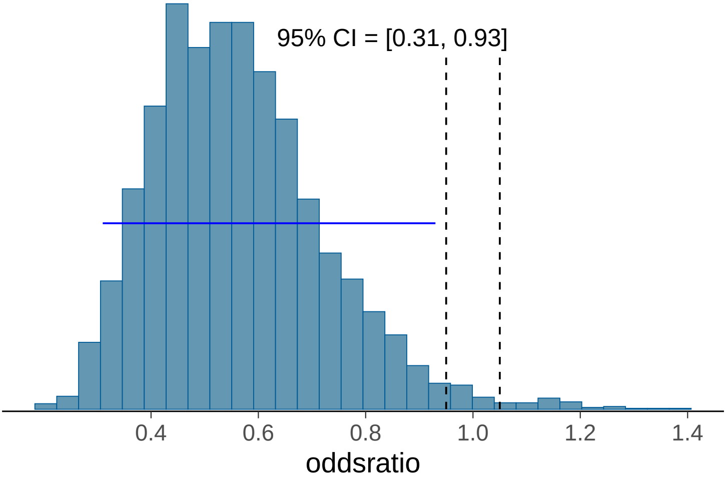
Bayesian hypothesis testing

- Equivalence testing (region of practical equivalence)
  - what is the probability that the effect is closer than ϵ to null, where ϵ is based on what is practically useful effect size



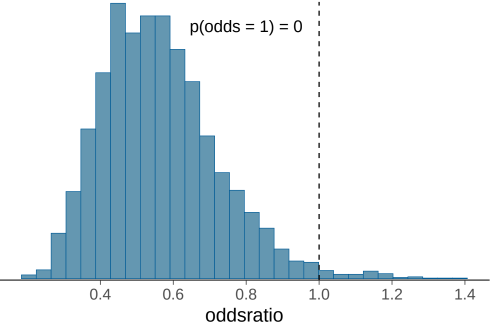
Bayesian hypothesis testing

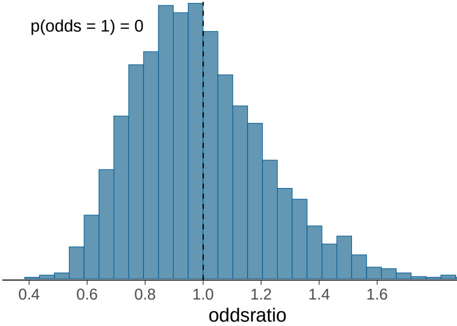
- Equivalence testing (region of practical equivalence)
  - some people combine posterior interval and region of practical equivalence



Bayesian hypothesis testing

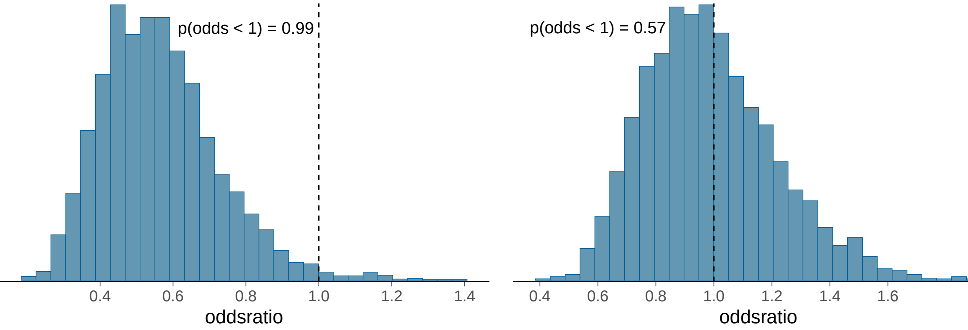
- Instead of hypothesis testing, report full posterior
  - for continuous posterior there is zero probability that e.g. treatment effect is exactly zero





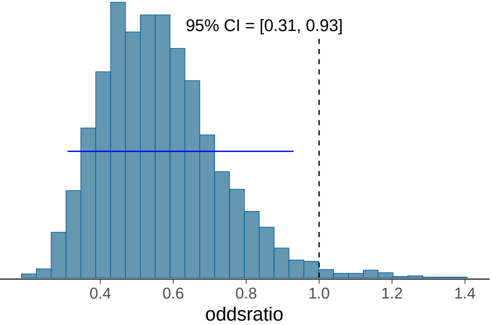
Bayesian hypothesis testing

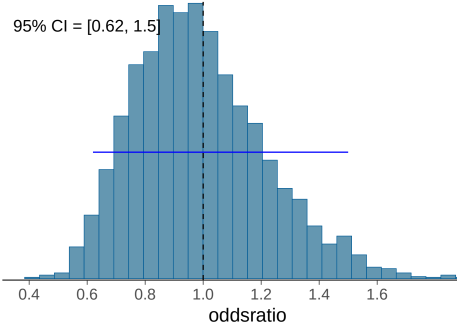
- Instead of hypothesis testing, report full posterior
  - for continuous posterior we could compute the probability that we know the sign of the effect



Bayesian hypothesis testing

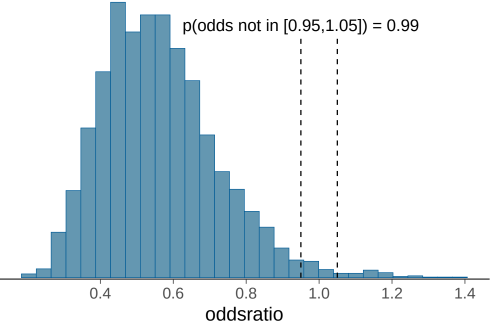
- Instead of hypothesis testing, report full posterior
  - for continuous posterior some people compare whether posterior interval includes null case

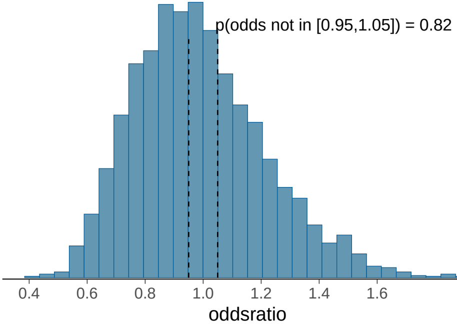




Bayesian hypothesis testing

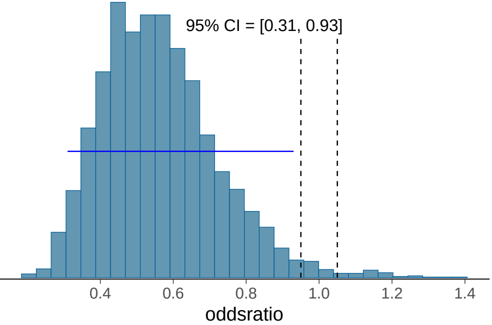
- Instead of hypothesis testing, report full posterior
  - region of practical equivalence (ROPE)

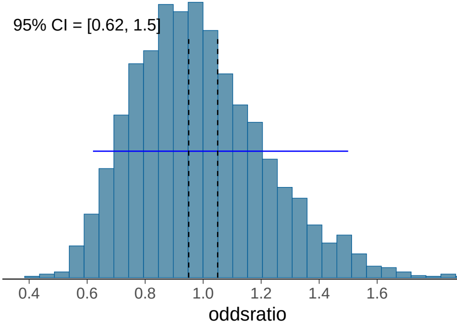




Bayesian hypothesis testing

- Instead of hypothesis testing, report full posterior
  - region of practical equivalence (ROPE)



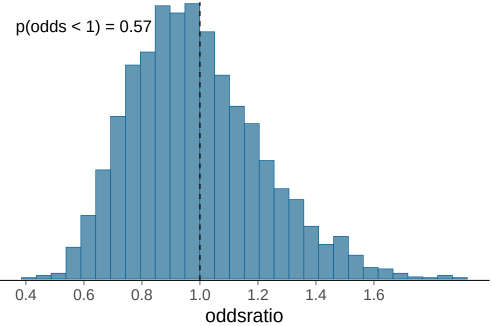


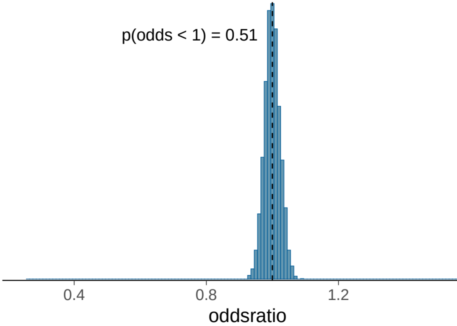
Bayesian hypothesis testing

- Instead of hypothesis testing, report full posterior
  - for continuous posterior there is zero probability that e.g. treatment effect is exactly zero

Bayesian hypothesis testing

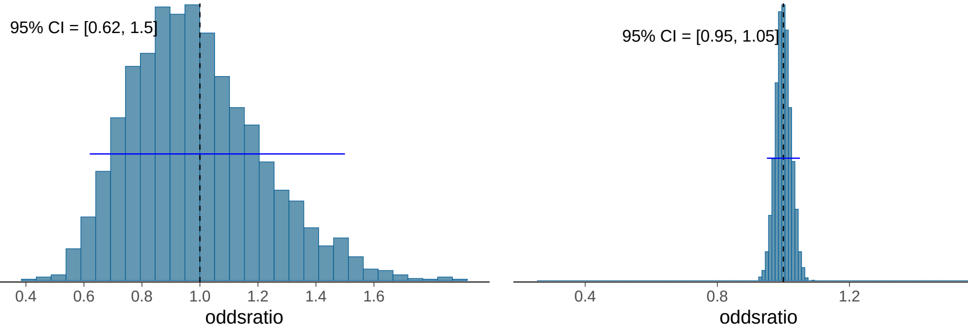
- Instead of hypothesis testing, report full posterior
  - for continuous posterior we could compute the probability that we know the sign of the effect





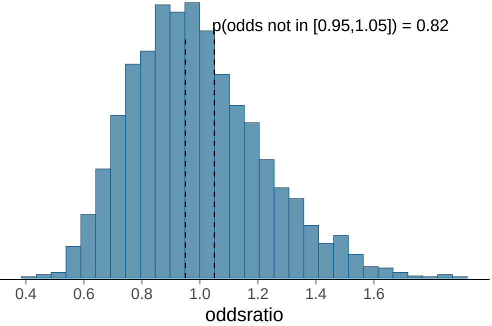
Bayesian hypothesis testing

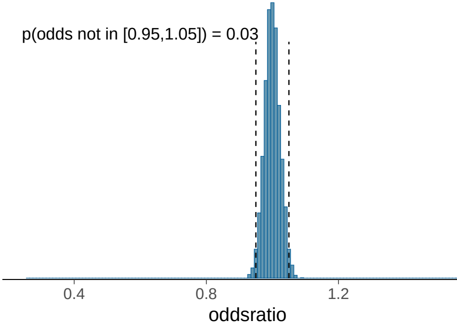
- Instead of hypothesis testing, report full posterior
  - for continuous posterior some people compare whether posterior interval includes null case



Bayesian hypothesis testing

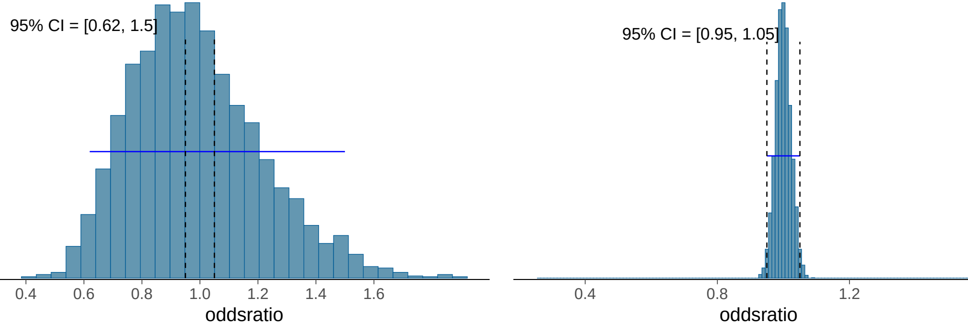
- Instead of hypothesis testing, report full posterior
  - region of practical equivalence (ROPE)





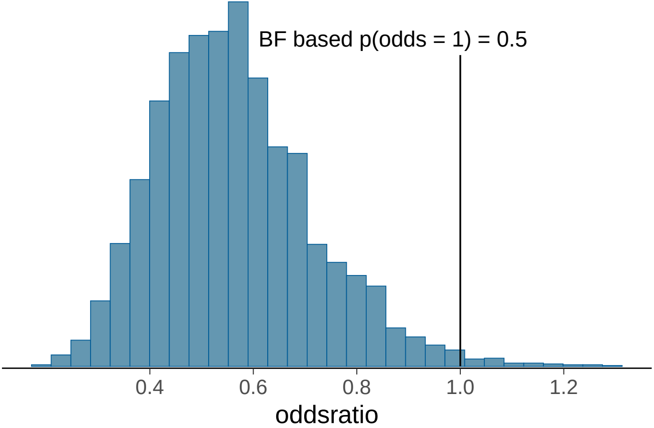
Bayesian hypothesis testing

- Instead of hypothesis testing, report full posterior
  - region of practical equivalence (ROPE)



Bayesian hypothesis testing

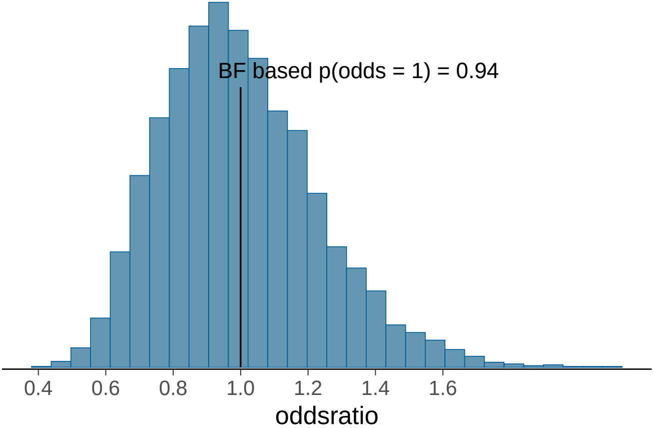
- Bayes factor
  - null model has, e.g., the treatment effect fixed to 0
  - assumes that there is non-zero probability that the treatment effect can be exactly zero (point mass)
  - requires posterior inference for the null model, too



with `bridgesampling` package, see also BDA3 13.10

Bayesian hypothesis testing

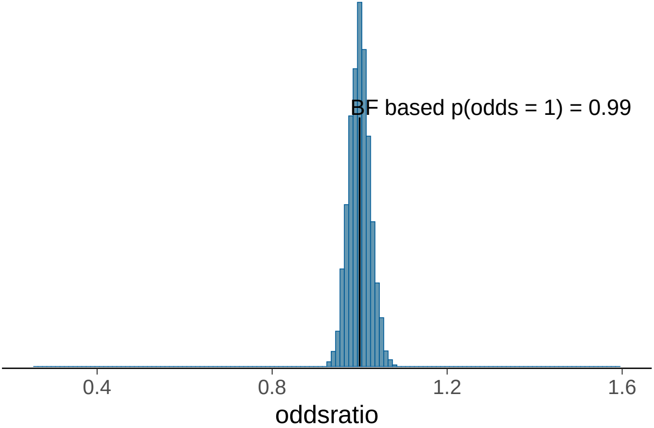
- Bayes factor
  - null model has, e.g., the treatment effect fixed to 0
  - assumes that there is non-zero probability that the treatment effect can be exactly zero (point mass)
  - requires posterior inference for the null model, too



with `bridgesampling` package, see also BDA3 13.10

Bayesian hypothesis testing

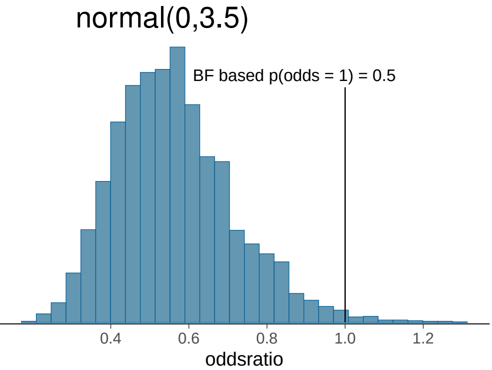
- Bayes factor
  - null model has, e.g., the treatment effect fixed to 0
  - assumes that there is non-zero probability that the treatment effect can be exactly zero (point mass)
  - requires posterior inference for the null model, too

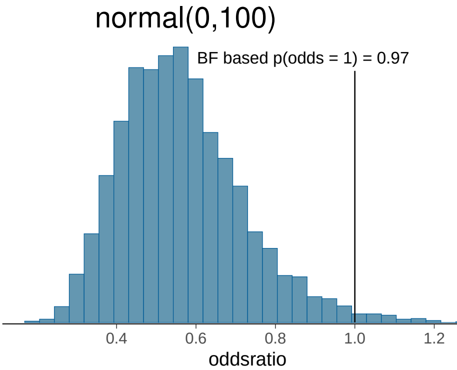


with `bridgesampling` package, see also BDA3 13.10

Bayesian hypothesis testing

- Bayes factor
  - sensitive to the prior choice even when the posterior is not





with `bridgesampling` package, see also BDA3 13.10

Bayesian hypothesis testing

- Predictive performance
  - is there difference in predictive performance with, e.g., treatment effect fixed to zero or unknown treatment effect
  - requires posterior inference for the null model or projection from the full to null
  - looking at the posterior is better if parameters are independent

Bayesian hypothesis testing

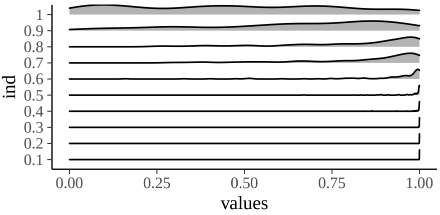
- Predictive performance
  - is there difference in predictive performance with, e.g., treatment effect fixed to zero or unknown treatment effect
  - requires posterior inference for the null model or projection from the full to null
  - looking at the posterior is better if parameters are independent

In the beta blockers example

- Leave-one-person-out works, but is less efficient than looking at the posterior (see https://avehtari.github.io/modelselection/betablockers.html)

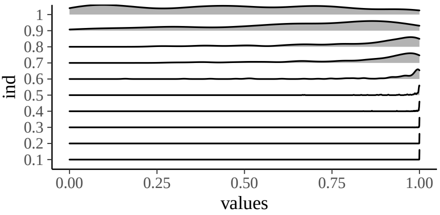
Simulation experiment

p(odds < 1)

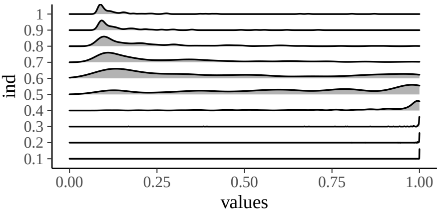


Simulation experiment

p(odds < 1)

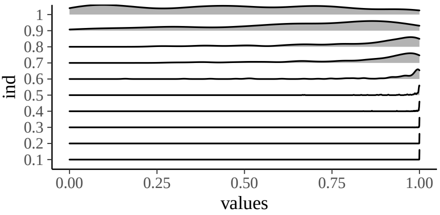


Marginal likelihood comparison

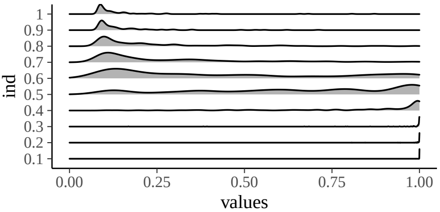


Simulation experiment

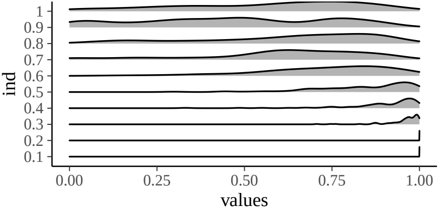
p(odds < 1)



Marginal likelihood comparison



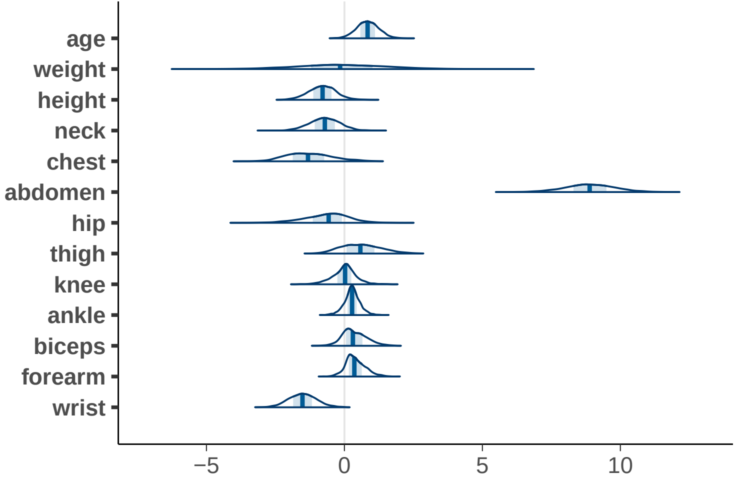
LOO comparison



Hypothesis testing and posterior dependencies

Looking at the marginal posterior $p(\beta < 0)$ can be misleading when there are many parameters

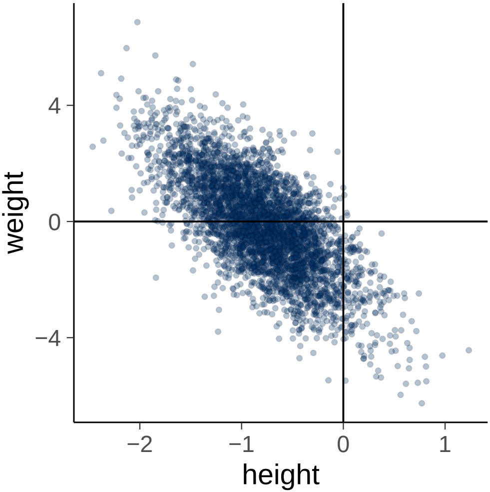
Marginal posteriors of coefficients



Hypothesis testing and posterior dependencies

Looking at the marginal posterior(s) can be misleading when there are many parameters

Bivariate marginal of weight and height



Hypothesis testing and posterior dependencies

In bodyfat example, starting from full model

- BF in favor of removing weight (p=0.92)
- LOO in favor of removing weight (p=0.99)

In bodyfat example, starting from model y $\sim$ abdomen

- BF in favor of adding weight (p=1.0)
- LOO in favor of adding weight (p=1.0)

Variable selection

More elaborate approaches are needed for variable selection

Projection predictive variable selection selects the minimal set of variables with similar predictive performance as the full model