# Model Checking

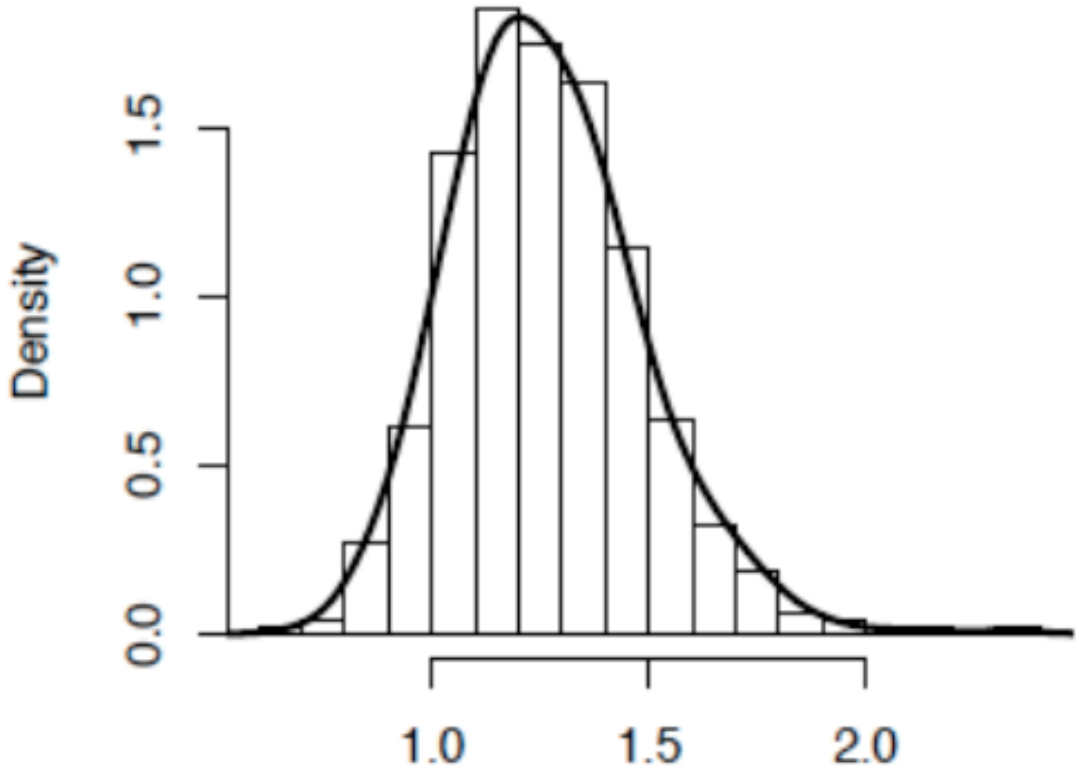
## Bayesian Modeling for Socio-Environmental Data

Chris Che-Castaldo, Mary B. Collins, N. Thompson Hobbs

August 2017

# What is the first question you should ask after fitting a model?

- *Are the predictions of the model consistent with the data?*
- Is the deterministic model a reasonable representation of the process?
- Have you made the right choices for distributions to represent uncertainties?



# What is model checking?

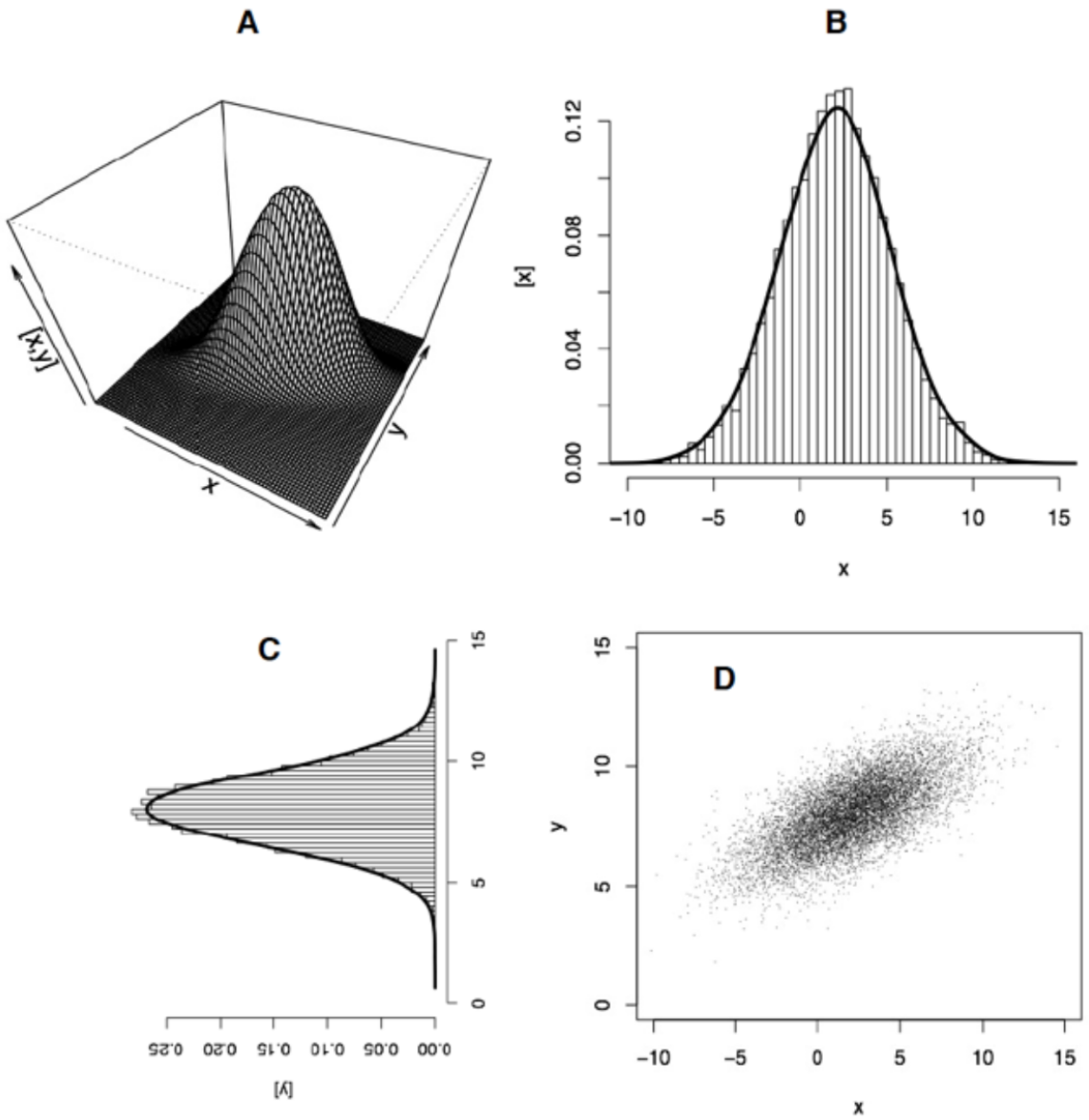
**Statistical inference relies on the assumption that your model can give rise to the data**. *Model checking* is the process of evaluating whether this is true.

# Usefulness of the marginal distribution

Recall, if function $f(A, B)$ specifies the joint probability of the continuous random variables $A$ and $B$ then,

- $\int f(A, B)dB$ is the marginal probabilty of $A$ and
- $\int f(A, B)dA$ is the marginal probability of $B$.
- This idea applies to any number of jointly distributed random variables. We simply integrate over all but one.

# Marginal distributions



# Posterior predictive checks

$$[y^{new} \mid y] = \underbrace{\int_{\theta_1} \ldots \int_{\theta_n} [y^{new} \mid \theta_1 \ldots \theta_n, y][\theta_1 \ldots \theta_n \mid y] d\theta_1 \ldots d\theta_n}_{\textbf{Posterior Predictive Distribution}}$$

- It is called *posterior* because it is conditional on the observed $y$ and *predictive* because it is a simulation of observable $y^{new}$, given modeled parameter estimates.
- Posterior predictive checks show the probability of a new simulated $y$ conditional on $\theta$, conditional on the data in hand, $y$.
- This is a marginal distribution because we are integrating over the $\theta$.

# Consider,

$$\mu_i = g(\theta_1, \theta_2, \theta_3, x_i) \tag{1}$$

$$y_i \sim \text{normal}(\mu_i, \sigma^2) \tag{2}$$

# Also see box 8.1 in Hobbs and Hooten

A new data set at each iteration

| $k$ | $\theta_1$ | $\theta_1$ | $\theta_3$ | $i=1$ | $i=2$ | $i=3$ | ... | $i=Y$ |
|---|---|---|---|---|---|---|---|---|
| 1 | .42 | 3.3 | 20.3 | $y^{new}_{1,1}$ | $y^{new}_{1,2}$ | $y^{new}_{1,3}$ | $\cdots$ | $y^{new}_{1,Y}$ |
| 2 | .41 | 2.3 | 18.5 | $y^{new}_{2,1}$ | $y^{new}_{1,2}$ | $y^{new}_{1,3}$ | $\cdots$ | $y^{new}_{1,Y}$ |
| 3 | .46 | 3.1 | 16.6 | $y^{new}_{3,1}$ | $y^{new}_{1,2}$ | $y^{new}_{1,3}$ | $\cdots$ | $y^{new}_{1,Y}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ |
| $K$ | .39 | 3.4 | 22.1 | $y^{new}_{n,1}$ | $y^{new}_{n,2}$ | $y^{new}_{n,3}$ | $\cdots$ | $y^{new}_{1,Y}$ |

# This is easier done than said.

- We have a model $g(\theta, x)$ that predicts a response $y$. We estimate the posterior distribtuion, $[\theta \mid y]$.
- For any given value of $x_i$, we can simulate the posterior predictive distribution $y^{new}$ by making a draw from $[y^{new} \mid g(\theta, x), \sigma^2]$.
- In MCMC this means making draws from the data model at each iteration; each draw is conditional on the current parameter values.
- We can simulate a new dataset by repeating these draws for all values of the $x$.
- Accumlating many of these draws defines the *posterior predictive distribution* in exactly the same way that many draws allow us to define the posterior distribution of the parameters.

$$g(b_0, b_1, x_i) = b_0 + b_1 x_i$$

$$[b_0, b_1, \tau \mid \mathbf{y}] \propto \prod_{i=1}^{n} \text{normal}\left(y_i \mid g(b_0, b_1, x_i)_i, \tau\right) \times$$

$$\text{normal}\left(b_0 \mid 0.0001\right) \text{normal}(b_1 \mid 0, .0001) \text{gamma}\left(\tau \mid .01 .01\right)$$

```
model{
b0 ~ dnorm(0,.0001)
b1 ~ dnorm(0,.0001)
tau ~ dgamma(.01,.01)
sigma<-1/sqrt(tau)
for(i in 1:length(y)){
  mu[i] <- b0 + b1*x[i]
  y[i] ~ dnorm(mu[i],tau)
  #posterior predictive distribution of y.new[i]
  y.new[i] ~ dnorm(mu[i],tau)
  }
}
```

# The Checking Part

- $T(y, \theta)$ is a test statistic (e.g., mean, standard deviation, CV, quantile, or sums of squares discrepancy) calcuated from the observed data.
- $T(y^{new}, \theta)$ is the corresponding statistic from the simulated, which is generated from the posterior predictive distribution.
- We calcuate:

$$P_b = \Pr(T(y^{new}, \theta) \geq T(y, \theta) \mid y)$$

- If $P_B$ is very large or very small, then the difference between the observed data and the simulated data cannot be attributed to chance. **This indicates lack of fit**.

# Candidates for test statistics

- mean
- variance
- coefficient of variation
- quantiles
- maximum, minimum
- discrepancy
- chi-square
- deviance

# R. A. Fischer's Ticks *A simple example*

We want to know (for some reason) the average number of ticks on sheep.

- We round up 60 sheep and count ticks on each one.
- Does a Poisson distribution fit the distribution of the data?

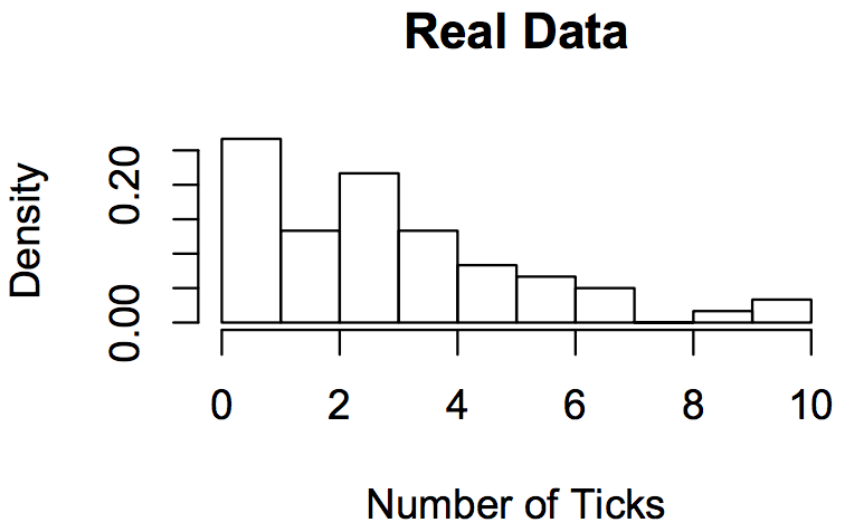
$$[\lambda \mid \mathbf{y}] \propto \prod_{i=1}^{60} \text{Poisson}[y_i \mid \lambda][\lambda]$$

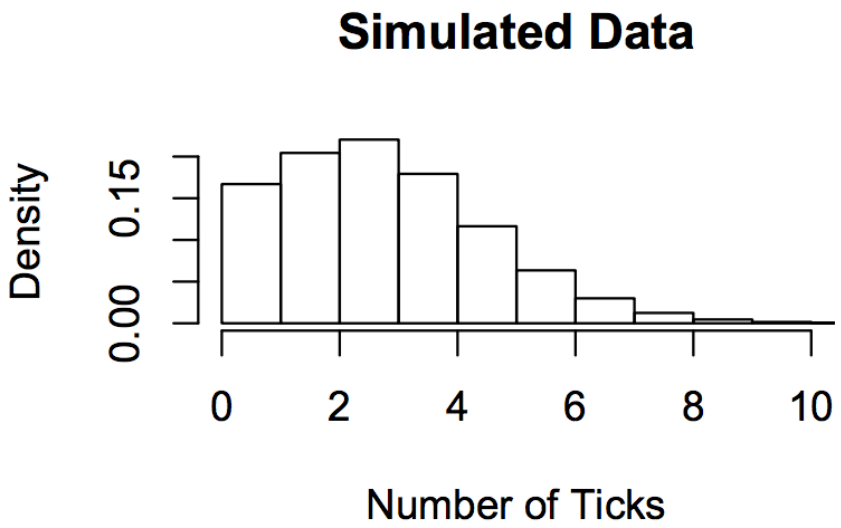
- For each value of $\lambda$ in the MCMC chain, we generate a new data set, $y^{new}$, by sampling from:

$$y_i^{new} \sim \text{Poisson}(\lambda)$$

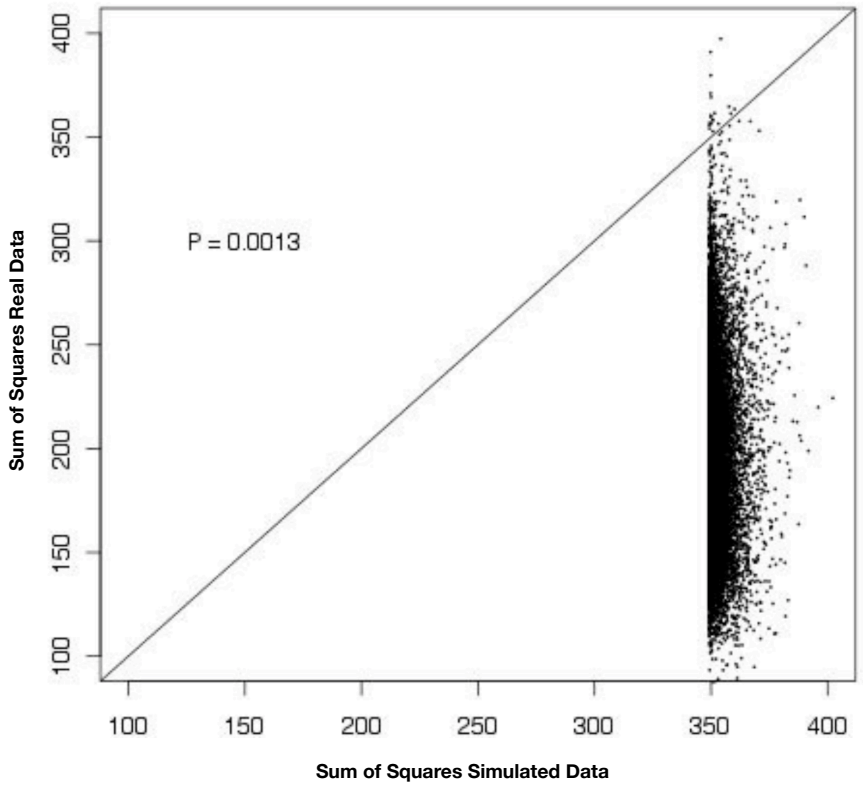
```
model{
lambda ~ dgamma(0.001,0.001)
for(i in 1:60){
     y[i] ~ dpois(lambda)
     y.new[i] ~ dpois(lambda)  #simulate a new data set of 60 points
}
cv.y <- sd(y[ ])/mean(y[ ])
cv.y.new <- sd(y.new[])/mean(y.new[ ])
pvalue.cv <- step(cv.y.new-cv.y)  # find Bayesian P value--the mean of
many 0's and 1's returned by the step function, one for each iteration in
the chain. The function step(z) returns a 1 if z > 0, returns 0
otherwise.
mean.y <-mean(y[])
mean.y.new <-mean(y.new[])
pvalue.mean <-step(mean.y.new - mean.y)
for(j in 1:60){
     sq[j] <- (y[j]-lambda)^2
     sq.new[j] <- (y.new[j]-lambda)^2
}
fit <- sum(sq[])
fit.new <- sum(sq.new[])
pvalue.fit <- step(fit.new-fit)
} #end of model
```

# Simple Model



# Posterior Predictive Checks



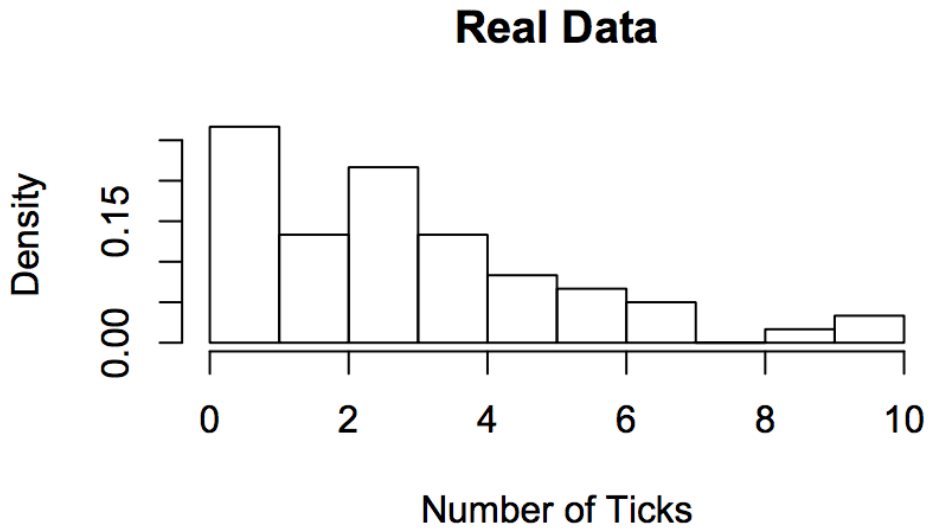

- P value for CV= .0013
- P value for mean = .51
- This is a two-tailed probability, *values close to 0 and 1 indicate lack of fit*.

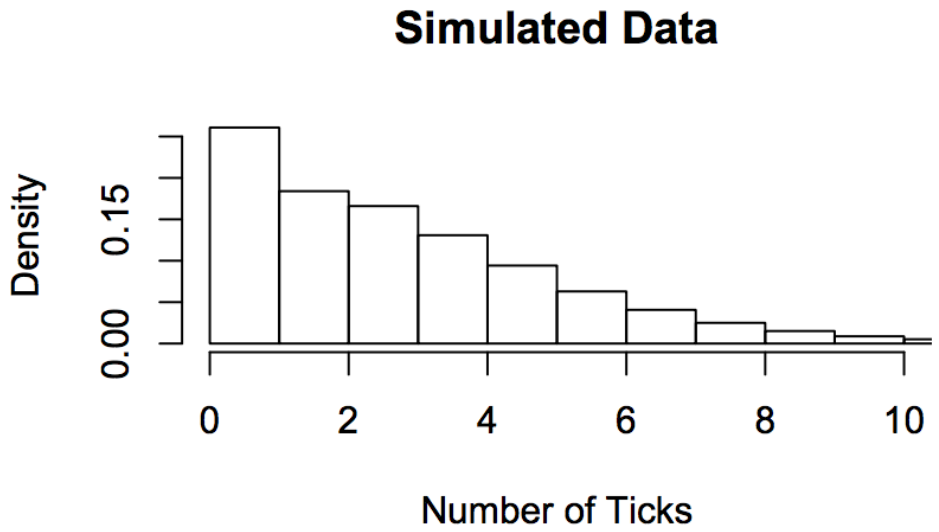
# How could you modify this model to allow "extra" variance?

- Draw a Bayesian network and write out the posterior and joint distributions.
- Don't use the negative binomial, please.

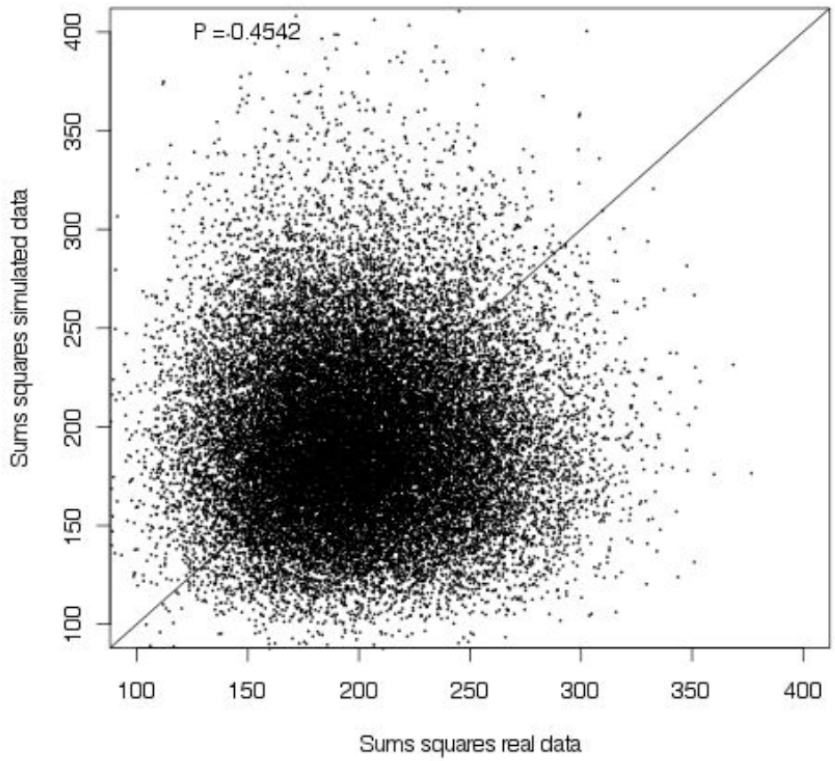
# Hierarchical model

$$[a, b, \lambda \mid \mathbf{y}] \propto \prod_{i=1}^{60} [y_i \mid \lambda_i][\lambda_i \mid a, b][a][b]$$


Real Data
Density
0.00
0.15
0
2
4
6
8
10
Number of Ticks


Simulated Data
Density
0.00
0.15
0
2
4
6
8
10
Number of Ticks

# Posterior Predictive Checks




- P value for CV= .45
- P value for mean = .5