Appendix 1

In this appendix, I explain why we use models that are hierarchical and Bayesian. To illustrate the approach I use in the main text, I build a series of models to estimate the probability that seedlings survive to become fruiting plants and fit them to a small dataset. First, I compare estimates from maximum likelihood and Bayesian approaches by fitting (1) a non-hierarchical model with a binomial likelihood and no prior and (2) a non-hierarchical model with a binomial likelihood with a beta prior. Second, I compare estimates from non-hierarchical and hierarchical approaches by developing (3) a hierarchical model with a binomial likelihood and year-level parameters on a beta distribution and (4) a hierarchical model with a binomial likelihood and year- and population-level parameters on a beta distribution. Finally, I compare four parameterizations for hierarchical models with a binomial likelihood in order to improve computational efficiency.

Maximum likelihood and Bayesian estimates

Maximum likelihood estimate

The following logic underlies how we calculate the maximum likelihood estimate for seedling survival to fruiting. For a single observation, the likelihood that we observe y fruiting plants in a plot if n seedlings were present in the plot can be written as a function of the probability of seedling survival p as $[y|p,n] = \binom{n}{y} p^y (1-p)^{n-y}$. For a set of N observations, each with a number of seedlings n_i and a number of fruiting plants y_i in the ith observation, then we can write the likelihood as

$$\mathcal{L} = [\boldsymbol{y}|p,\boldsymbol{n}] = \prod_{i=1}^{N} \binom{n_i}{y_i} p_i^y (1-p)^{n_i - y_i}. \quad (1)$$

which is often written as

$$\mathcal{L} = [\boldsymbol{y}|p, \boldsymbol{n}] = \prod_{i=1}^{N} \text{binomial}(n_i, p). \tag{2}$$

We can use the likelihood to obtain a maximum likelihood estimate (by minimizing the negative log-likelihood). The maximum likelihood estimate $\hat{p}$ is the overall proportion of seedlings that survive to become fruiting plants, summing all the observations. Specifically, it calculates a maximum likelihood estimate for each population in each year of the dataset. We thus give each site j and year k its own probability of success p_{jk} and obtain the MLEs $\hat{p}_{jk}$.

$$[\boldsymbol{y}|\boldsymbol{p}, \boldsymbol{n}] = \prod_{j=1}^{J} \prod_{k=1}^{K} \prod_{i=1}^{N} \text{binomial}(n_{ijk}, p_{jk}) \tag{3}$$

Binomial likelihood with a beta prior, complete pooling

We can turn this into a Bayesian model by adding a prior to our model. Because the beta is a conjugate prior for a binomial distribution, we use use a beta distribution for the prior. In other words, this choice of prior matches the likelihood in a way that the posterior has the same distribution as the prior (cf. Bolker p 177). A beta distribution with shape parameters $\alpha = \beta = 1$ corresponds to noninformative prior. For a set of N observations, each with a number of seedlings n_i and a number of fruiting plants y_i in the ith observation, we can write the joint posterior as

$$[\boldsymbol{y}|p, \boldsymbol{n}] = \prod_{i=1}^{N} \text{binomial}(n_i, p)\text{beta}(p|1, 1). \tag{4}$$

A single probability p represents the probability of seedling survival to fruiting for the all trials (a model with *complete pooling*). The opposite extreme is a model in which each trial

i has its own probability of seedling survival to fruiting p_i (a model with *no pooling*).

$$[\boldsymbol{y}|\boldsymbol{p},\boldsymbol{n}] = \prod_{i=1}^{N} \text{binomial}(n_i, p_i)\text{beta}(p_i|1,1). \tag{5}$$

To compare our site- and year-specific MLEs to estimates from Bayesian models, we give each site j and year k its own probability of success p_{jk}, and place a prior on each p_{jk}.

$$[\boldsymbol{y}|\boldsymbol{p},\boldsymbol{n}] = \prod_{j=1}^{J}\prod_{k=1}^{K}\prod_{i=1}^{N} \text{binomial}(n_{ijk}, p_{jk})\text{beta}(p_{jk}|1,1). \tag{6}$$

This is a model in which we are completely pooling observations from each site and year. Another thing we could say about this model is that it is Bayesian but non-hierarchical. This is extends what happens with the maximum likelihood estimates when we sum across all the plots at a site in a given year and calculate the proportion of seedlings that survive to become fruiting plants. One difference between the two approaches is that with the Bayesian model we account for the number of trials and counts; the data from one plot with a single seedling compromises with the prior to give us posterior probability of success (see **Comparison**). The Bayesian and frequentist estimates converge as the sample size approaches infinity.

Hierarchical models

Binomial model with a beta prior, partial pooling, parameterization via mean: one population, one year

Next, we'll consider adding pooling to our model. To explain this, we'll focus first on the data from one population in one year. We want a hierarchical model that estimates the probability of survivorship in each plot (θ_i) and simultaneously estimates the mean probability of survivorship in the year (ϕ). The probability of survivorship for each plot i is θ_i. We then assume that the probability of survivorship for the i plots is drawn from a

distribution of probabilities defined by the mean probability of survivorship in a given year ϕ and sample size κ. This effectively means that the prior on θ_i is itself a parameter; (i.e. we use hyperpriors rather than directly place priors on the probability of survivorship θ_i). We reparameterize the beta distribution with its mean ϕ and the parameter κ. The mean of a random variable distributed beta(α, β) is $\phi = \frac{\alpha}{\alpha+\beta}$. With a beta$(1, 1)$ prior, the parameter $\kappa = \alpha + \beta$ is roughly the sample size plus two.

$$\begin{aligned}[\boldsymbol{p}, \boldsymbol{\theta}, \phi, \kappa | \boldsymbol{y}, \boldsymbol{n}] = & \prod_{i=1}^{N} \text{binomial}(n_i, \theta_i) \\ & \times \text{beta}(\theta_i | \phi\kappa, (1-\phi)\kappa) \\ & \times \text{uniform}(\phi | 0, 1) \text{Pareto}(\kappa | 1.5, 1).\end{aligned} \tag{7}$$

We place a uniform prior on the mean probability of survivorship, ϕ, because it must lie between 0 and 1. We place a bounded, positive prior on κ with the Pareto(α, c) distribution, which is parameterized by a shape (α) and scale (c) parameter.

This parameterization is hierarchical because the estimates for ϕ and κ contribute to estimates for θ_i. [other effects?]

Binomial model with a beta prior, partial pooling, parameterization via mean: one population, multiple years

Next, we expand the scope of our analysis to include data from plots (i) in multiple years (j) of data for a single population $k = 1$. Here, we want a hierarchical model that estimates the probability of survivorship in each plot in each year (θ_{ij}). We want to simultaneously estimate the mean probability of survivorship in each year (ϕ_j) and the mean probability of survivorship in the population. The probability of survivorship for each plot i in a year j is θ_{ij}. We then assume that the probability of survivorship for the i plots in year j is drawn from a distribution of probabilities defined by the mean probability of survivorship

in a given year ϕ_j and sample size κ_j. Furthermore, we assume that the mean probability of survivorship in a given year ϕ_j is itself drawn from a distribution of probabilities defined by the mean probability of survivorship at the population ϕ_0 and the parameter κ_0.

$$\begin{aligned}[\boldsymbol{p}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\kappa}, \phi_0, \kappa_0 | \boldsymbol{y}, \boldsymbol{n}] = & \prod_{j=1}^{J} \prod_{i=1}^{N} \text{binomial}(n_{ij}, \theta_{ij}) \\ & \times \text{beta}(\theta_{ij} | \phi_j \kappa_j, (1-\phi_j)\kappa_j) \\ & \times \text{beta}(\phi_j | \phi_0 \kappa_0, (1-\phi_0)\kappa_0) \text{Pareto}(\kappa_j | 1.5, 1) \\ & \times \text{uniform}(\phi_0 | 0, 1) \text{Pareto}(\kappa_0 | 1.5, 1).\end{aligned} \tag{8}$$

This parameterization is hierarchical because the estimates for ϕ_j and κ_j contribute to estimates for θ_{ij}, and the estimates for ϕ_0 and κ_0 contribute to estimates for ϕ_j. [other effects?]

Binomial model with a beta prior, partial pooling, parameterization via mean: multiple populations, multiple years

Finally, we expand the scope of our analysis to include data from plots (i) in multiple years (j) of data for multiple populations k. Here, we want a hierarchical model that estimates the probability of survivorship in each plot in each year at each population (θ_{ijk}). We want to simultaneously estimate the mean probability of survivorship in each population ϕ_k, the mean probability of survivorship in each year ($\phi_j k$). The probability of survivorship for each plot i in a year j at population k is θ_{ijk}. We then assume that the probability of survivorship for the i plots in year j in population k is drawn from a distribution of probabilities defined by the mean probability of survivorship in a given year at a given population ϕ_{jk} and paramter κ_{jk}. Furthermore, we assume that the mean probability of survivorship in a given year in a given population ϕ_{jk} is itself drawn from a distribution of probabilities defined by the mean probability of survivorship for the population $\phi_{0,k}$ and the parameter $\kappa_{0,k}$. The model is

similar to the one for one population except the indexing has changed so that we estimate a mean probability of survivorship for each population.

$$\begin{aligned}[\boldsymbol{p}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\kappa}, \phi_0, \kappa_0 | \boldsymbol{y}, \boldsymbol{n}] = & \prod_{k=1}^{K} \prod_{j=1}^{J} \prod_{i=1}^{N} \text{binomial}(n_{ijk}, \theta_{ijk}) \\ & \times \text{beta}(\theta_{ijk} | \phi_{jk}\kappa_{jk}, (1-\phi_{jk})\kappa_{jk}) \\ & \times \text{beta}(\phi_{jk} | \phi_{0,k}\kappa_{0,k}, (1-\phi_{0,k})\kappa_{0,k}) \text{Pareto}(\kappa_j | 1.5, 1) \\ & \times \text{uniform}(\phi_{0,k} | 0, 1) \text{Pareto}(\kappa_{0,k} | 1.5, 1).\end{aligned} \tag{9}$$

This parameterization is hierarchical because the estimates for ϕ_{jk} and κ_{jk} contribute to estimates for θ_{ijk}, and the estimates for $\phi_{0,k}$ and $\kappa_{0,k}$ contribute to estimates for ϕ_{jk}. [other effects?]

Hierarchical model parameterization

I considered four parameterizations for the structure of the hierarchical model. The first is a parameterization of the mean probability of success and the sample size. The second is a parameterization in which I moment match the parameters of the beta distribution. The third is a parameterization in which I use a logit-link function and a centered parameterization. The fourth is a parameterization in which I use a logit-link function and a non-centered parameterization.

Hierarchical model with parameterization for beta distribution by mean probability of success, ϕ and sample size, κ.

$$\begin{aligned}[\boldsymbol{\theta}, \mu, \nu | \boldsymbol{y}, \boldsymbol{n}] = &\prod_{i=1}^{N} \text{binomial}(y_i | n_i, \theta_i) \\ &\times \text{beta}(\theta_i | \phi\kappa, (1-\phi)\kappa) \\ &\times \text{uniform}(\phi | 0, 1) \text{Pareto}(\kappa | 1.5, 1)\end{aligned} \tag{10}$$

Hierarchical model with parameterization of beta distribution by moment matching mean and variance.

$$\begin{aligned}[\boldsymbol{\theta}, \mu, \sigma | \boldsymbol{y}, \boldsymbol{n}] = &\prod_{i=1}^{N} \text{binomial}(y_i | n_i, \theta_i) \\ &\times \text{beta}(\theta_i | \frac{\mu^2 - \mu^3 - \mu \times \sigma^2}{\sigma^2}, \frac{\mu - 2 \times \mu^2 + \mu^3 - \sigma^2 + \mu \times \sigma^2}{\sigma^2}) \\ &\times \text{normal}(\mu | 0, 1) \text{Half} - \text{Cauchy}(\sigma | 0, 1)\end{aligned} \tag{11}$$

Hierarchical model with parameterization via log-odds probability of success, centered.

$$\begin{aligned}[\boldsymbol{\theta}, \boldsymbol{\alpha}, \mu, \sigma | \boldsymbol{y}, \boldsymbol{n}] = &\prod_{i=1}^{N} \text{binomial}(y_i | n_i, \text{logit}^{-}1(\alpha_i)) \\ &\times \text{normal}(\alpha_i | \mu, \sigma) \\ &\times \text{normal}(\mu | 0, 1) \text{Half} - \text{Cauchy}(\sigma | 0, 1)\end{aligned} \tag{12}$$

where

$$\theta_i = \text{logit}^{-}1(\alpha_i) \tag{13}$$

Hierarchical model with parameterization via log-odds probability of success, non-centered.

$$
\begin{aligned}
[\boldsymbol{\theta}, \boldsymbol{\alpha}, \mu, \sigma | \boldsymbol{y}, \boldsymbol{n}] = & \prod_{i=1}^{N} \text{binomial}(y_i | n_i, \text{logit}^{-}1(\alpha_i)) \\
& \times \text{normal}(\alpha_i^{std} | 0, 1) \\
& \times \text{normal}(\mu | 0, 1) \text{Half} - \text{Cauchy}(\sigma | 0, 1)
\end{aligned}
\tag{14}
$$

where

$$
\begin{aligned}
\theta_i &= \text{logit}^{-}1(\alpha_i) \\
\alpha_i &= \mu + \alpha_i^{std} \times \sigma
\end{aligned}
\tag{15}
$$

Comparison

First, we want to explore estimates from a Bayesian model compare to the maximum likelihood estimates. Second, we want to explore the effect of adding hierarchical structure to the parameter estimates in our models. Finally, we want to explore the effect of adding hierarchical structure on how appropriate our models are (compare posterior predictive checks).

Estimates from maximum likelihood versus Bayesian models

First, we want to explore estimates from a Bayesian model compare to the maximum likelihood estimates. The comparison I am interested in is the effect of adding a prior, rather than the effect of adding hierarchy to the model. I will compare the maximum likelihood estimates (equation (**??**)) and the non-hierarchical Bayeisan model (equation (**??**)). The first panel in Figure **??** shows that the maximum likelihood estimates are pretty similar to those from the Bayesian model with complete pooling per population and a beta-binomial parameterization. The major differences are where the maximum likelihood estimates ap-

proach 0 or 1. The second panel in Figure **??** shows that this is because those estimates come from year-population combinations with a small sample size. For example, the MLE for a year-population combination with one plot and 1 seedling that dies before fruiting would be 0. In the Bayesian model, the prior has a comparatively larger influence on the posterior in situations where there is little data. In this case, the posterior would be a compromise between our one data point and our prior. The estimates converge once we have 5 data points.

Hierarchical structure

Second, we want to explore the effect of adding hierarchical structure to the parameter estimates in our models. These comparisons will all be among models fit to data for one population ($k = 1$). I am interested in comparing a non-hierarchical model (equation (**??**)), a hierarchical model with year-level parameters (equation (**??**)), and a hierarchical model with year-level and population-level parameters (equation (**??**)). All these comparisons are made at a randomly selected site.

The first panel in figure **??** shows that the complete pooling model produces estimates for each plot that are identical; there is one estimate per year. The second and third panels show taht the posterior medians for θ_i and θ_{ij} from the hierarchical models are 'shrunk' from the point estimates y_{ij}/n_{ij} towards the year-level mean (dotted line). Figure **??** that the models shrink θ_i and θ_{ij} towards the year-level estimate, and that this effect varies across years.

In the non-hierarchical model, the posterior distribution for the probability of survivorship in each year is p_{jk} (where k=1) in equation (**??**)). In the hierarchical model with year-level parameters, the posterior distribution for the mean probability of survivorship in each year is ϕ in equation (**??**). In the hierarchical model with year- and population-level parameters, the posterior distribution for the mean probability of survivorship in each year is

ϕ_j, and the posterior distribution for the mean probability of survivorship in the population is ϕ_0 in equation (**??**). Figures **??**) and **??**) both show that the posterior distribution for the mean-probability of survivorship from the hierarchical models (red lines) is broader than for the probability of survivorship from the non-hierarchical model (blue lines). Figure **??**) also shows the posterior for the population-level probability of survivorship (black line).

Figure **??**) summarizes the posterior distributions for the year-level probability of survivorship by their median and 95% credible intervals. The hierarchical structure has the effect of broadening the credible intervals associated with the estimated mean probability of survivorship in each year.

Model parameterization

I compare a few key properties of these fits to provide a sense of which parameterization is most efficient here. First, look at the relationship of hyperparameters. Second, look at the number of effective samples. Third, look at the amount of shrinkage from each model fit. Fourth, compare the population-level estimates from each of the models.

Figure X shows how these different parameterizations sample hyperparameter space, and correlations between hyperparameters and probability of success. Uncorrelated sampling suggests the parameterization is exploring parameter space. 'Funnels' in correlations between probability of sucess and hyperparameters indicate that particular parameter combinations are undersampled. Figure X shows that the parameterizations can vary in how efficiently they sample. Each model was run for an identical number of iterations; in this figure, I compare the number of effective samples obtained for each θ_i. Figure X illustrates an important general feature of hierarchical models; pooling is greater when there are fewer trials. The right panel in the figure also demonstrates that the amount of shrinkage varies between parameterizations. The next figure shows the same points; the mean and sample size parameterization is the least shrunk towards the population mean (denoted by the dashed

line). Finally, Figure X shows that the median population-level estimates for the different parameterizations are similar but that they differ in how broad the credible intervals are. The log-odds parameterization imposes greater shrinkage around the population mean, leading to narrower credible intervals over the mean and sample size parameterization.

Posterior predictive checks

Finally, we want to explore the effect of adding hierarchical structure on how appropriate our models are (compare posterior predictive checks). The figures appended to this PDF (still need to figure out how to best integrate them) show graphical posterior predictive checks (cf. Gabry et al. 2019) for the non-hierarchical (blue), the hierarchical with year-level parameters (purple), and the hierarchical with year- and population-level parameters (gray). Gabry et al. (2019) describe how to perform PPCs: the posterior distribution is used to simulate data, and this data is summarized and compared to observed test statistics. For all three models, the mean of simulations from the posterior does a good job matching the observed mean from the data. This is perhaps not surprising because the models are parameterized via the mean probability of success. Posterior predictive checks for the variance may be more informative (variance is not a parameter in our models). The variance of simulations from the posterior of the non-hierarchical model under- or over-estimates the variance. The variance of simulations from both the hierarchical models do a reasonable job at matching the observed variance.

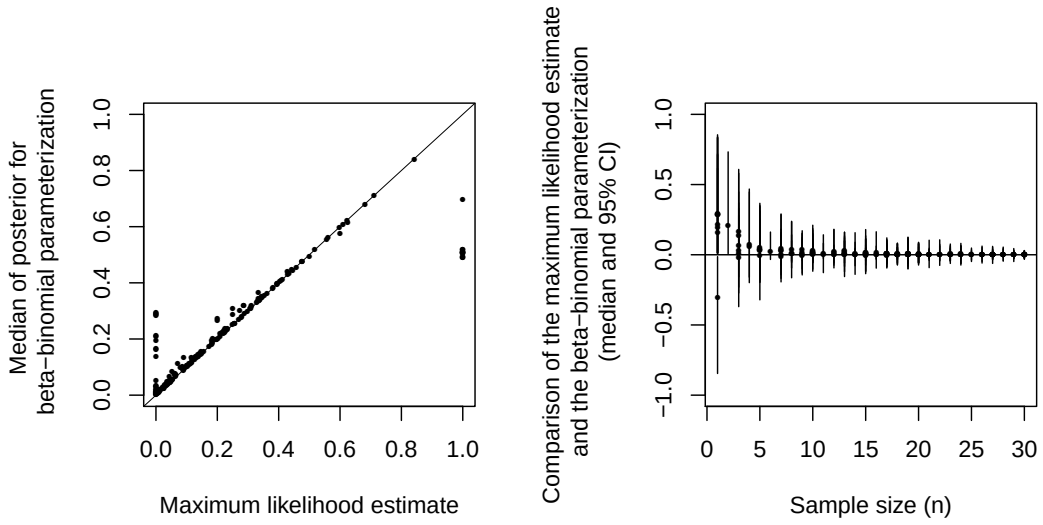


Figure 1: (A) This panel plots the median of the posterior from the beta-binomial with complete pooling per population against the maximum likelihood estimate. (B) This panel compares the full posterior distribution from the beta-binomial parameterization with the maximum likelihood estimate. The plot shows the median of the difference (with 95% CIs) against the sample size in the year-population combination for that estimate.

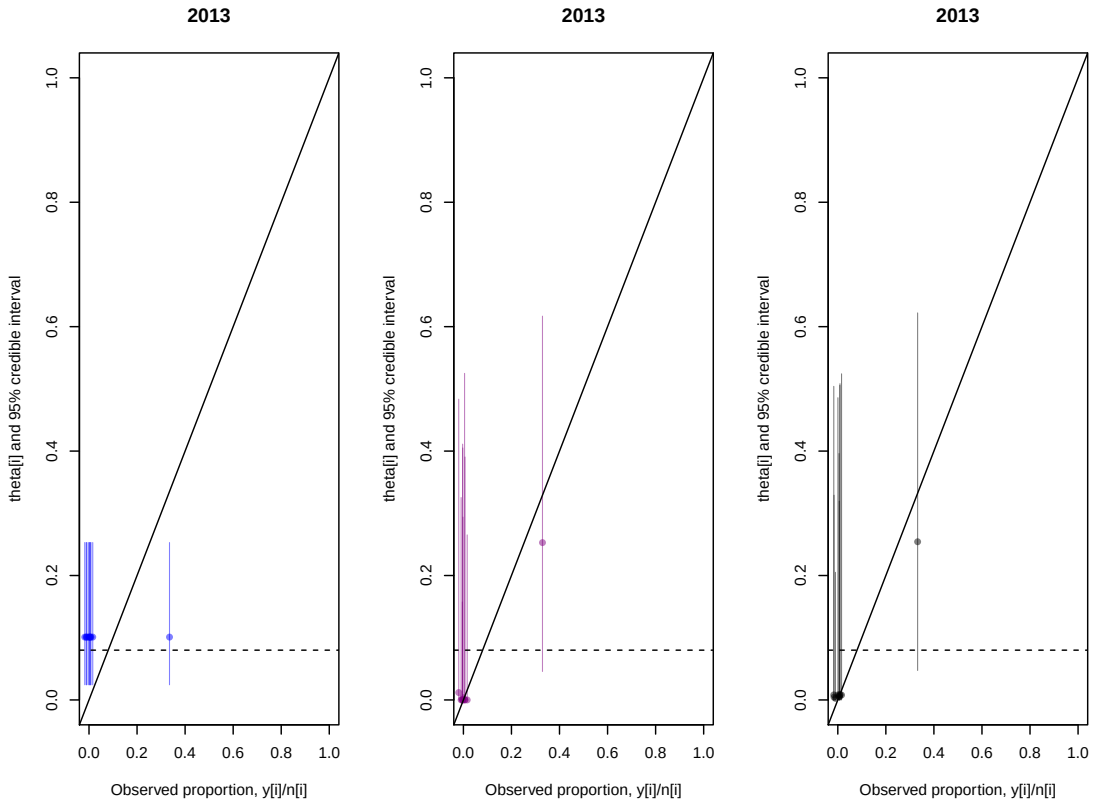


Figure 2: Posterior median and 95% credible intervals for a randomly selected year, based on simulations from the joint posterior distribution plotted against the unpooled estimates. The 1:1 line corresponds to unpooled estimates $\theta_i = y_i/n_i$. The horizontal, dashed line corresponds to the year-level maximum likelihood estimate. (A) Estimates from the complete pooling model. (B) Estimates from the hierarchical model with year-level parameters. (C) Estimates from the hierarchical model with year- and population-level parameters.

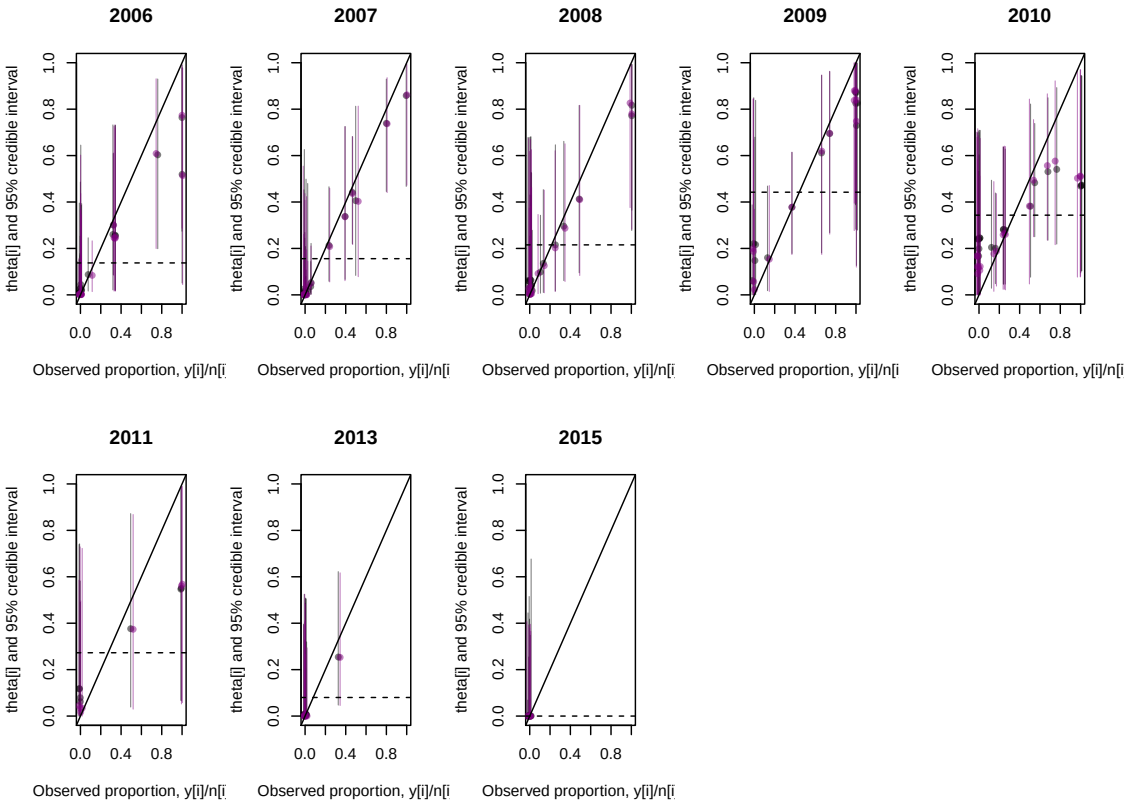


Figure 3: (A) Posterior median and 95% credible intervals for a randomly selected year, based on simulations from the joint posterior distribution plotted against the unpooled estimates. The 1:1 line corresponds to unpooled estimates $\theta_i = y_i/n_i$. The horizontal, dashed lines corresponds to the year-level maximum likelihood estimates.

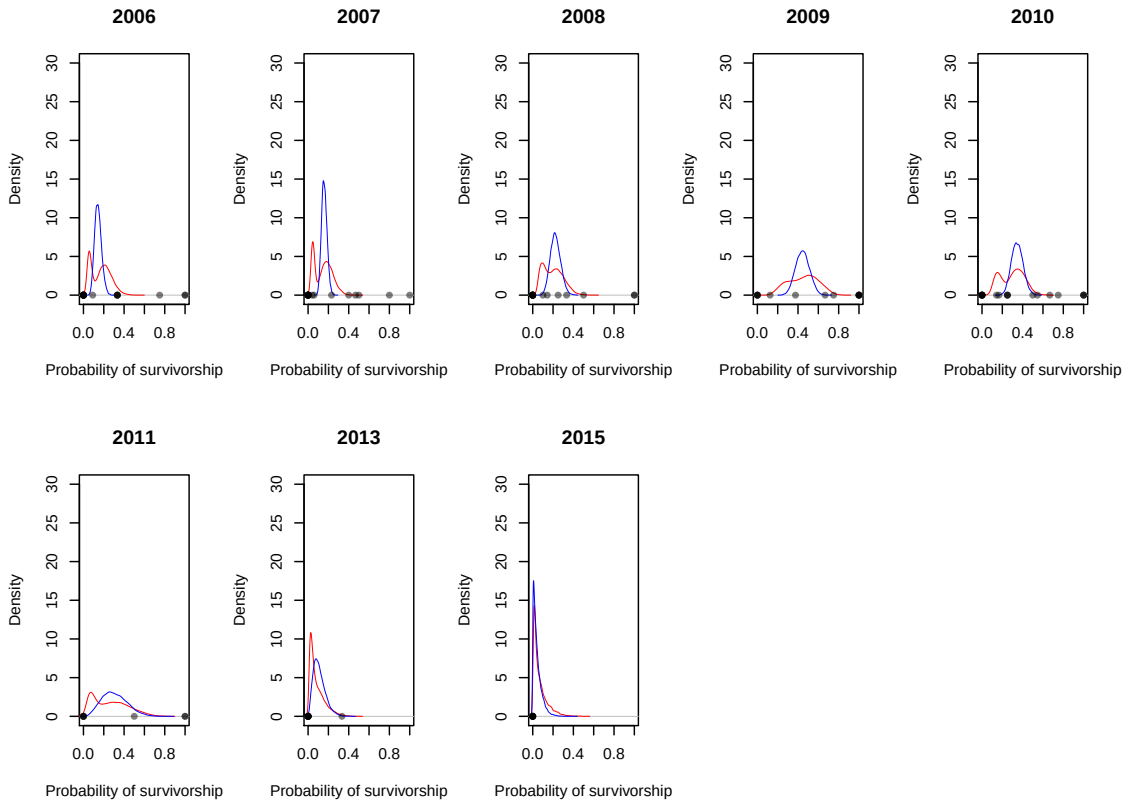


Figure 4: (A) Comparison of and nonhierarchical (blue) and year-level parameter hierarchical (red) model. The curves are posterior density distributions for the probability of survivorship p_{jk} (blue) and ϕ (red).

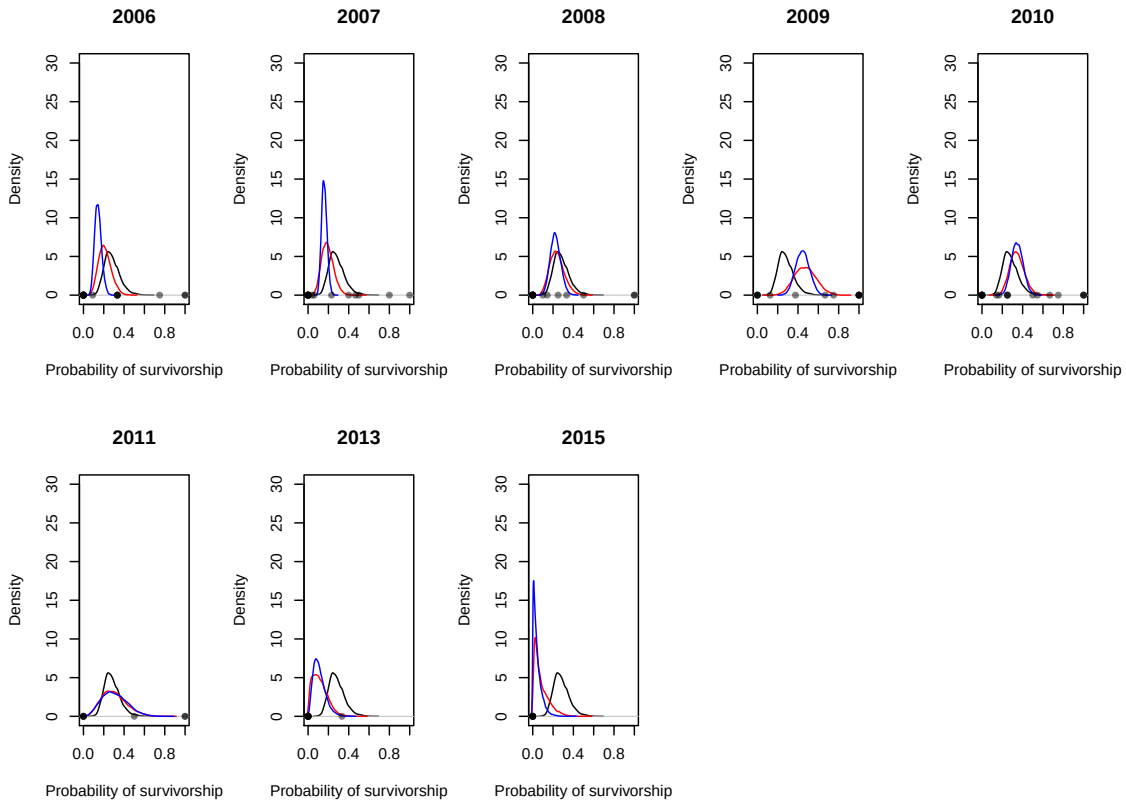


Figure 5: (A) Comparison of and nonhierarchical and year- and population-level parameter hierarchical model. The curves are posterior density distributions for the probability of survivorship p_{jk} (blue), ϕ_j (red), and ϕ_0 (black).

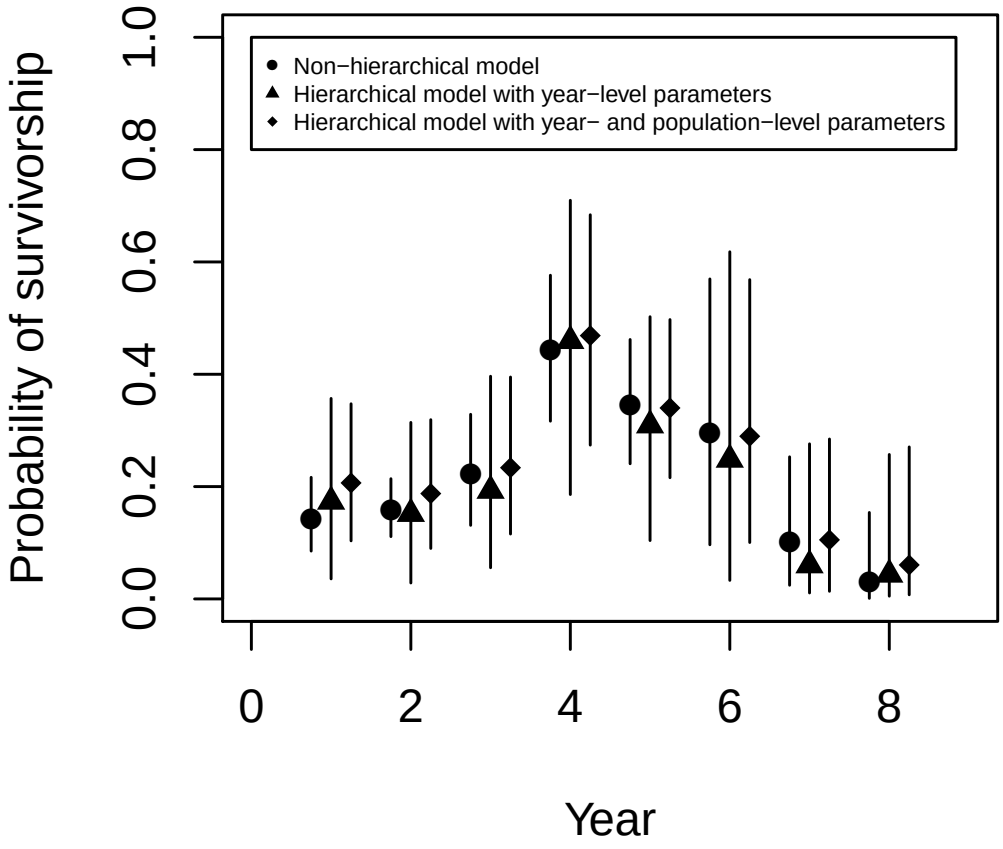
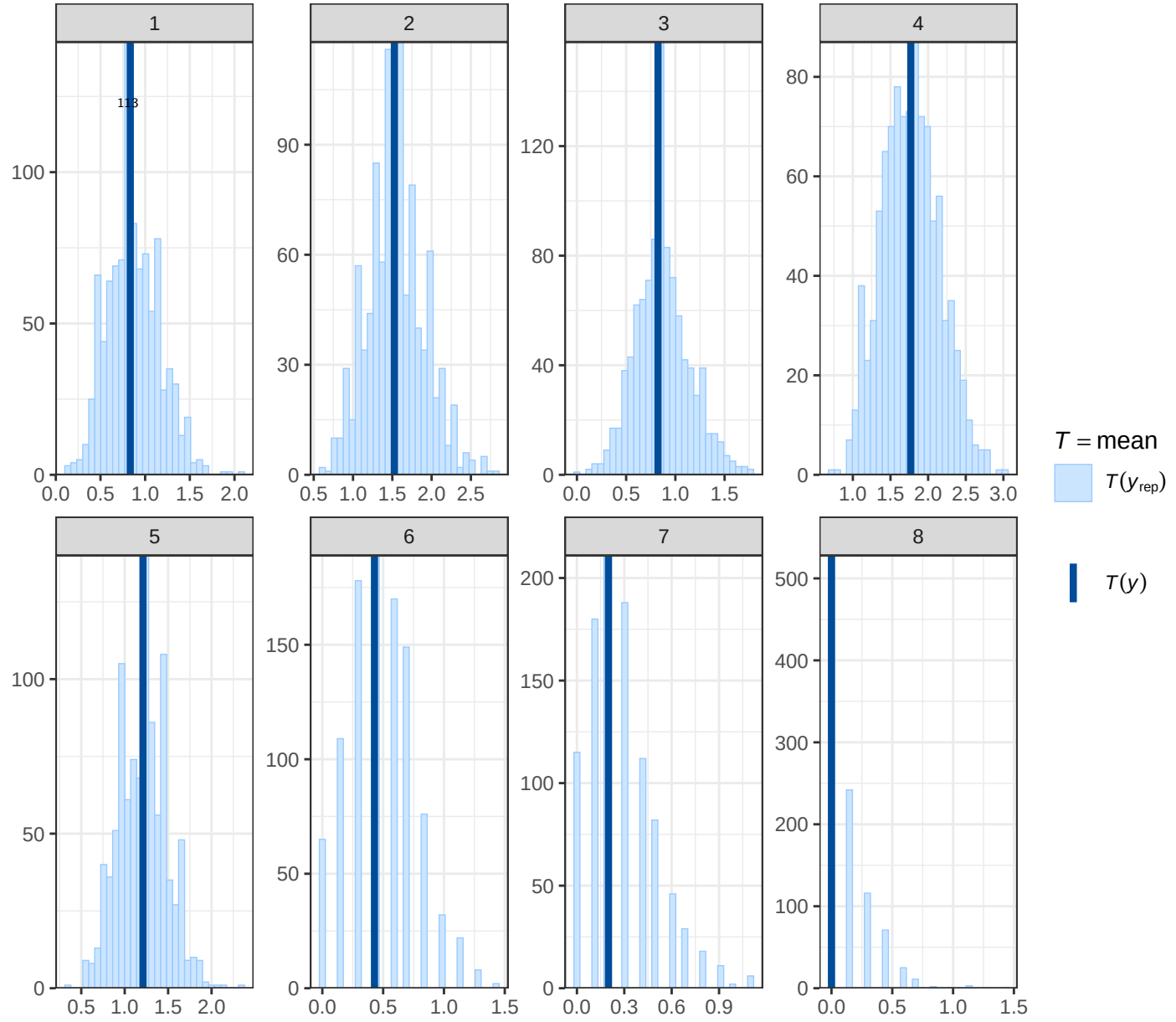


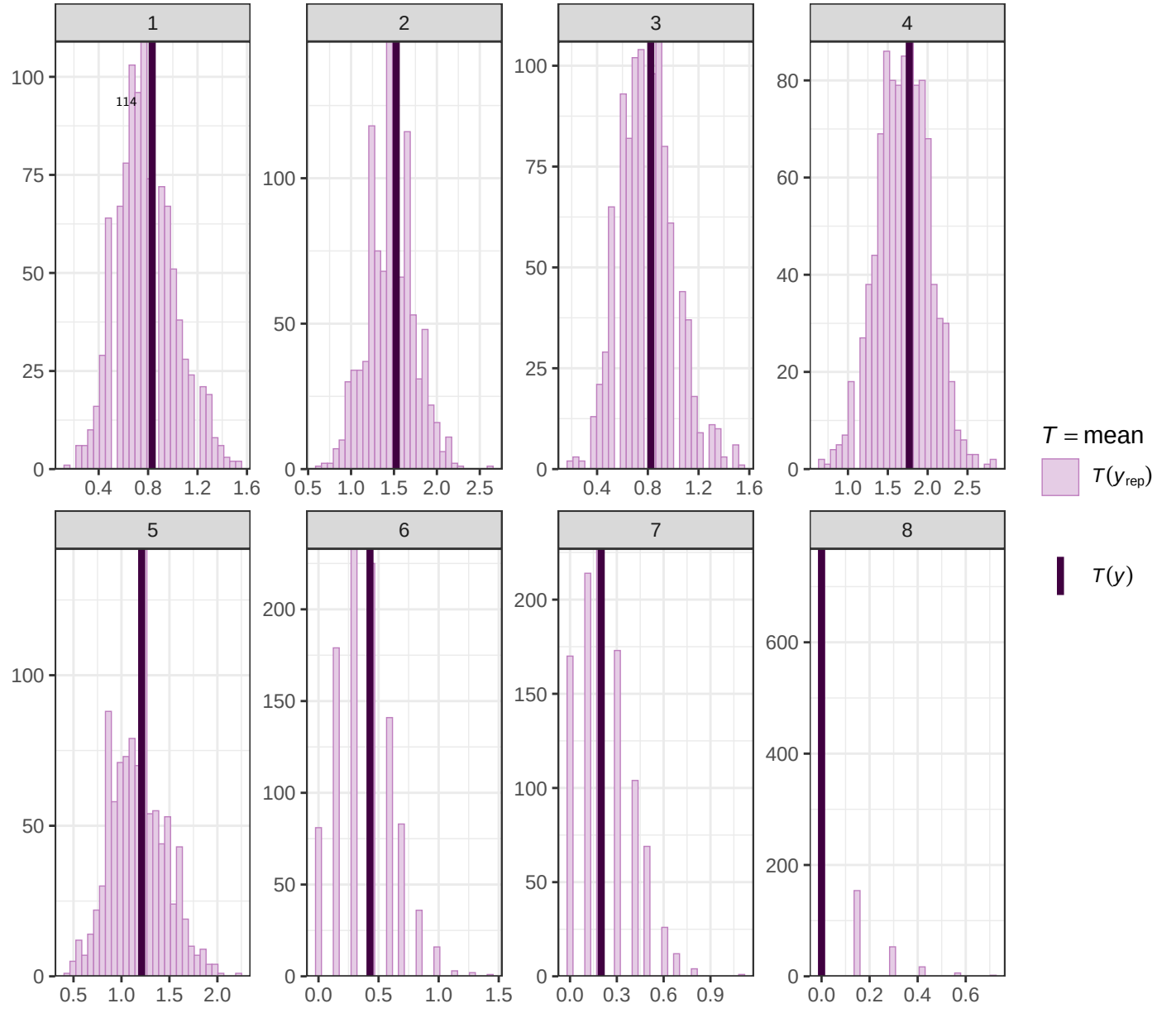
Figure 6: (A) Comparison of median and 95% confidence intervals for models with three levels of structure, fit to the same dataset.

Posterior predictive checks for number of fruiting plants; non−hierarchical model



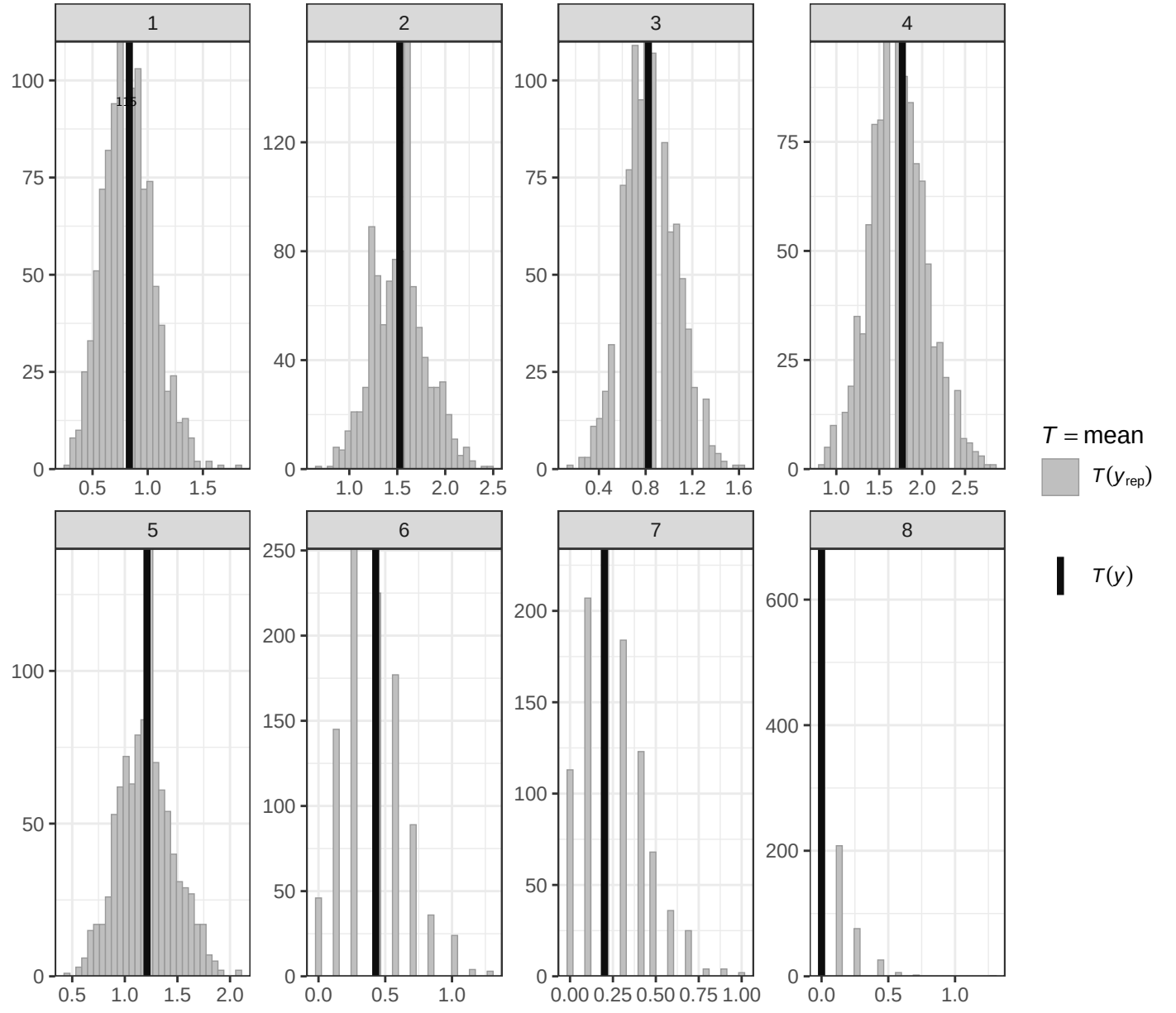
the density of observed data (y) and the lighter lines show the densities of Y_rep from 1000 draws of the posterior

Posterior predictive checks for number of fruiting plants; hierarchical model with year−lev



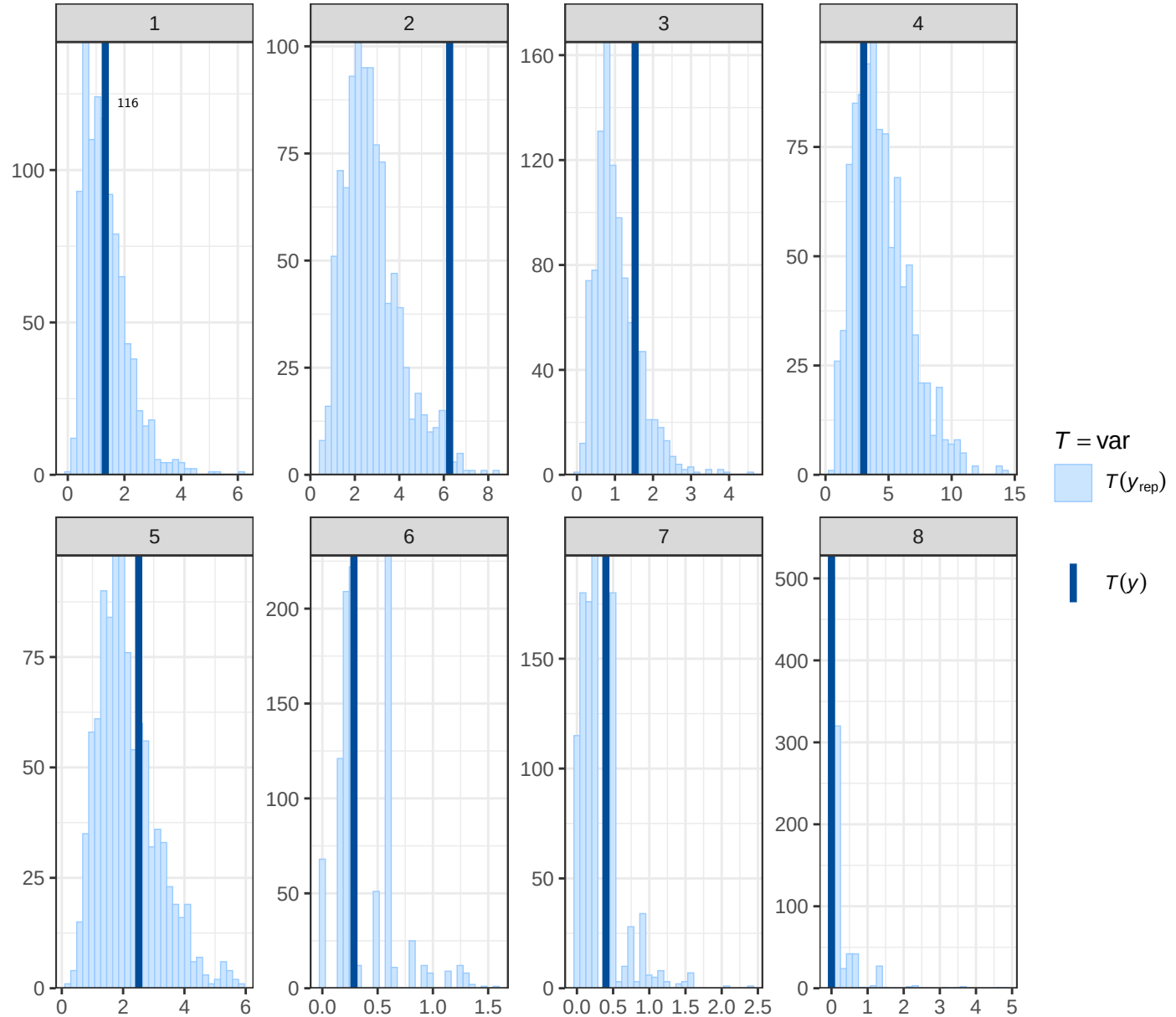
the density of observed data (y) and the lighter lines show the densities of Y_rep from 1000 draws of the posterior

Posterior predictive checks for number of fruiting plants; hierarchical model with year− ar

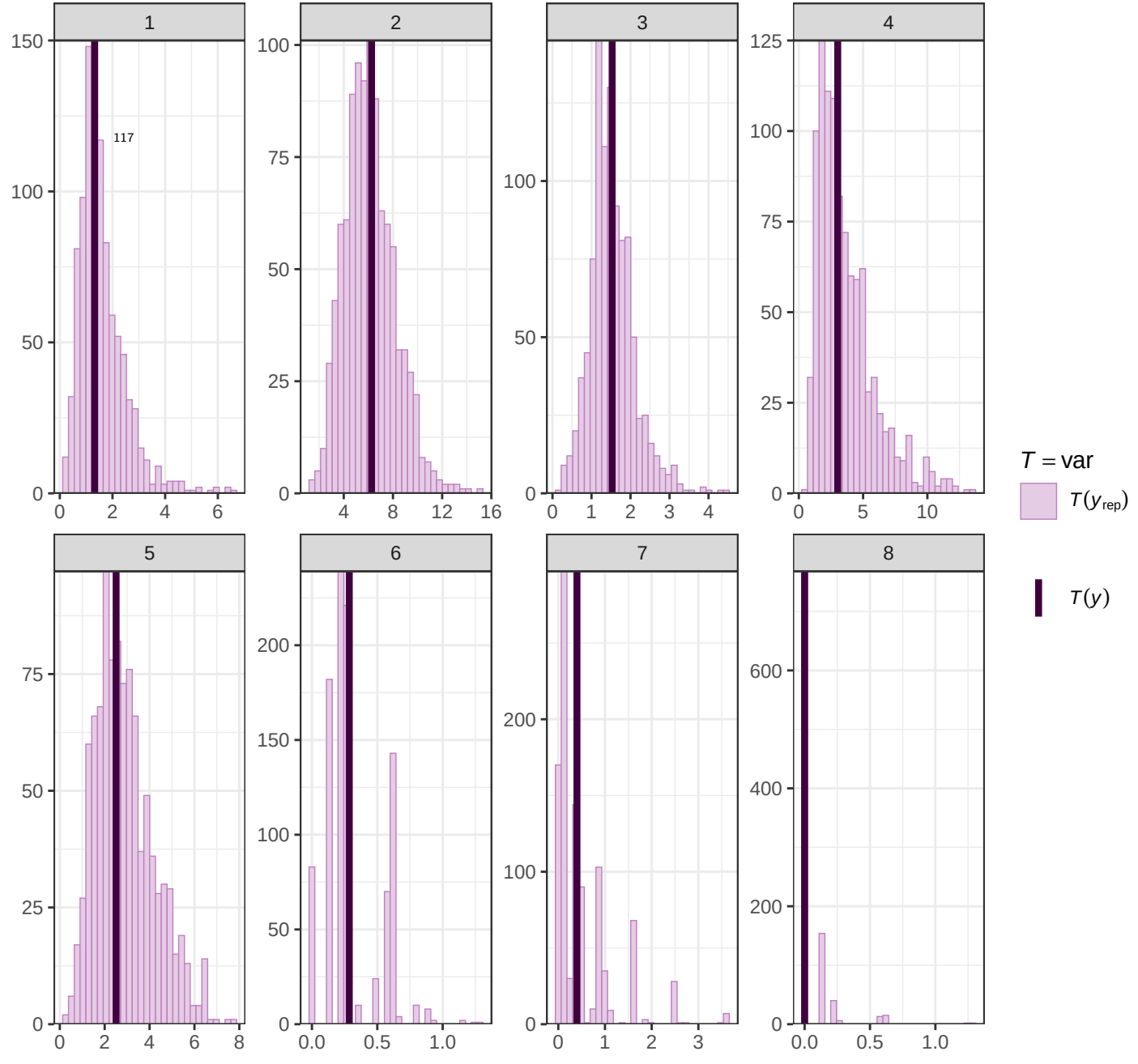


the density of observed data (y) and the lighter lines show the densities of Y_rep from 1000 draws of the posterior

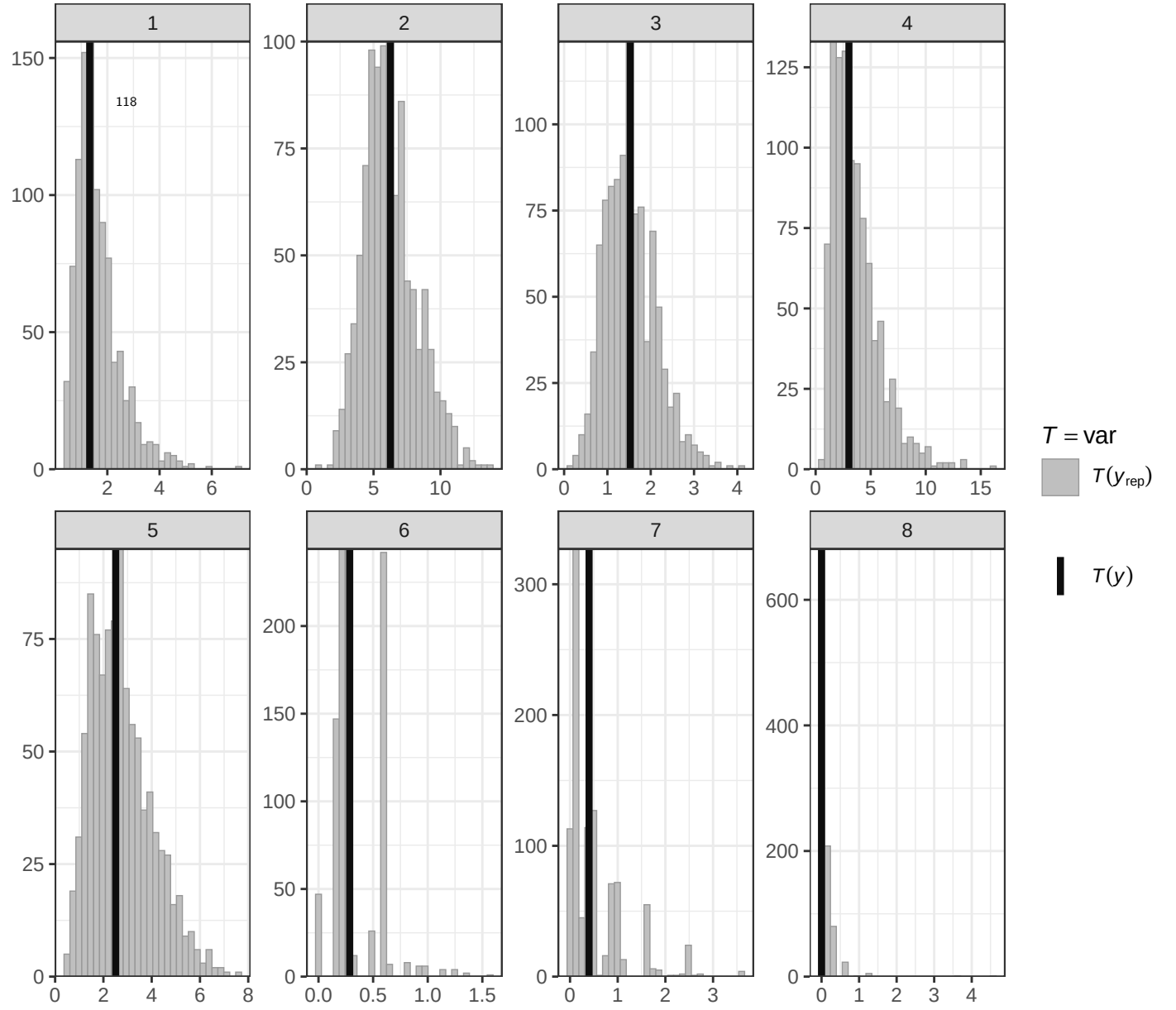
Posterior predictive checks for number of fruiting plants; non−hierarchical model



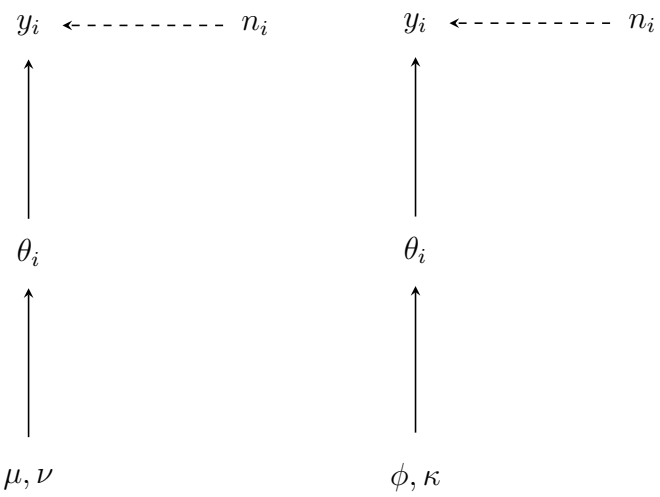
the density of observed data (y) and the lighter lines show the densities of Y_rep from 1000 draws of the posterior

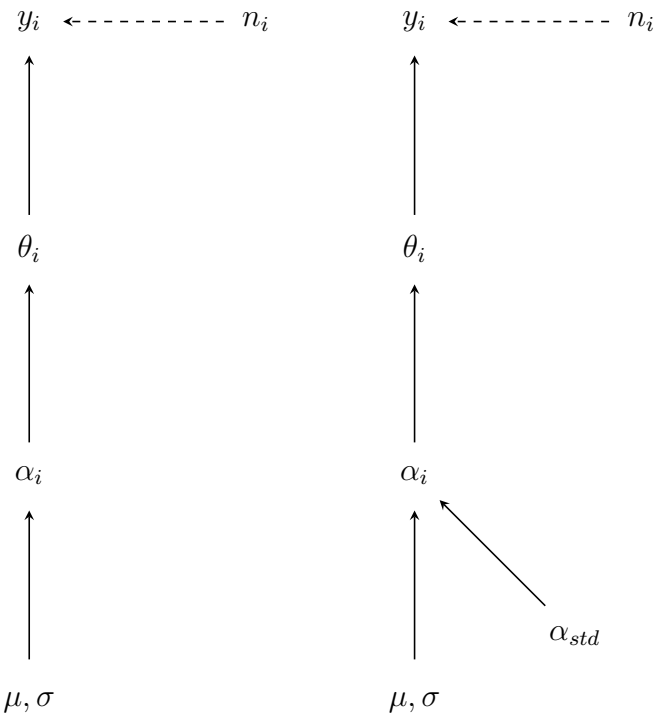

1
2
3
4
5
6
7
8
117
$T = \text{var}$
$T(y_{rep})$
$T(y)$

Posterior predictive checks for number of fruiting plants; hierarchical model with year− ar



the density of observed data (y) and the lighter lines show the densities of Y_rep from 1000 draws of the posterior


y_i
n_i
θ_i
μ, ν
y_i
n_i
θ_i
ϕ, κ


y_i
n_i
θ_i
α_i
μ, σ
y_i
n_i
θ_i
α_i
α_{std}
μ, σ

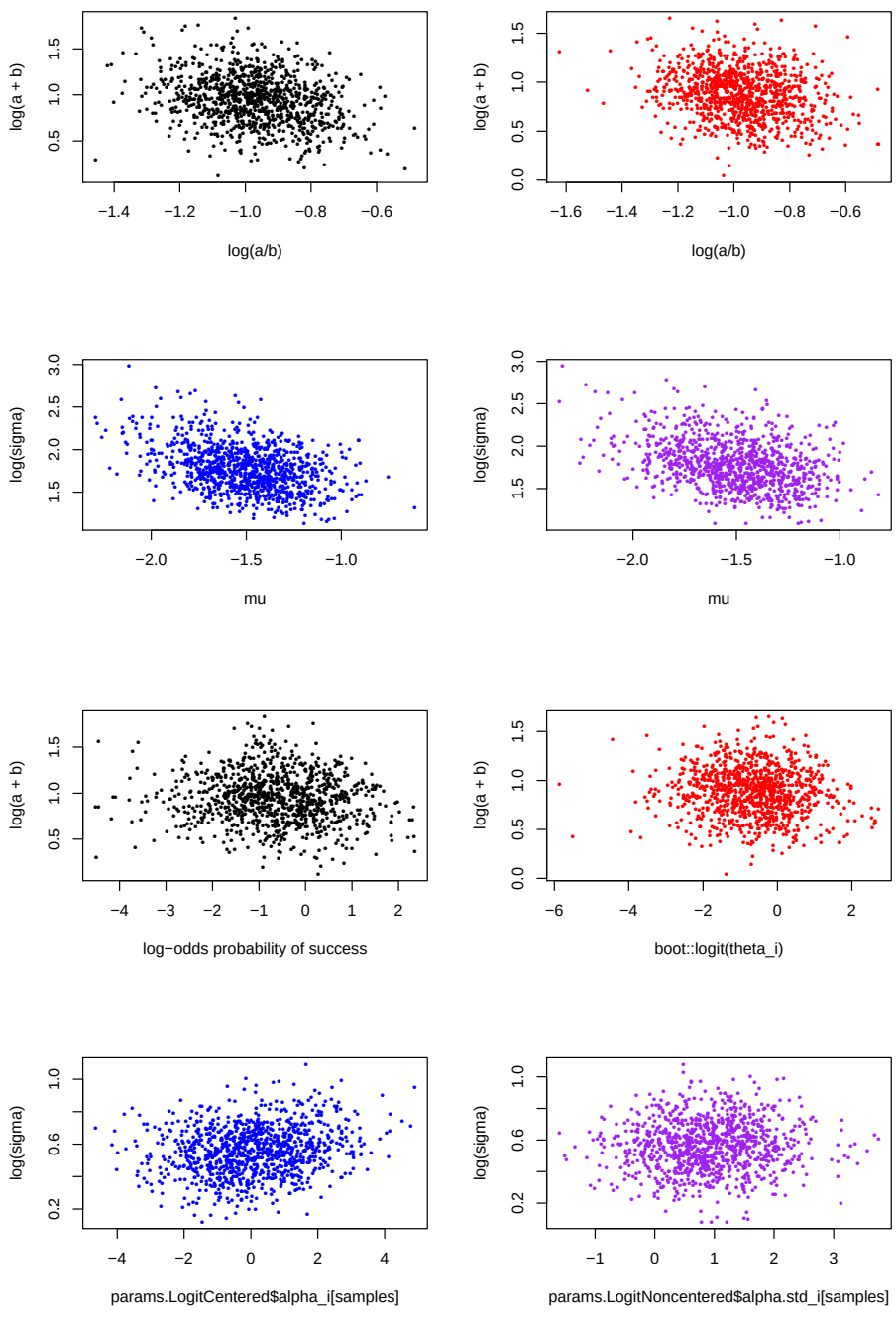

log(a + b)
log(a/b)
log(a + b)
log(a/b)
log(sigma)
mu
log(sigma)
mu
log(a + b)
log−odds probability of success
log(a + b)
boot::logit(theta_i)
log(sigma)
params.LogitCentered$alpha_i[samples]
log(sigma)
params.LogitNoncentered$alpha.std_i[samples]

Figure 7

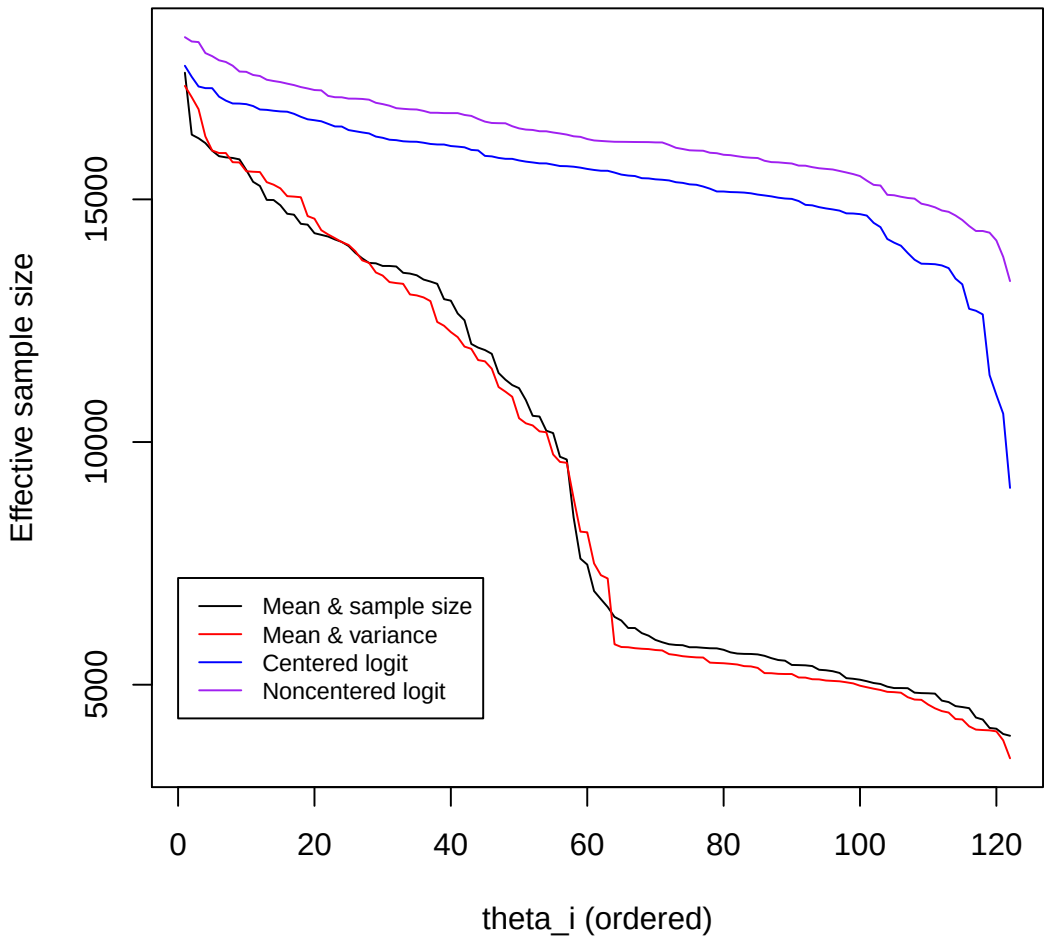

Effective sample size
5000
10000
15000
0
20
40
60
80
100
120
theta_i (ordered)
Mean & sample size
Mean & variance
Centered logit
Noncentered logit

Figure 8

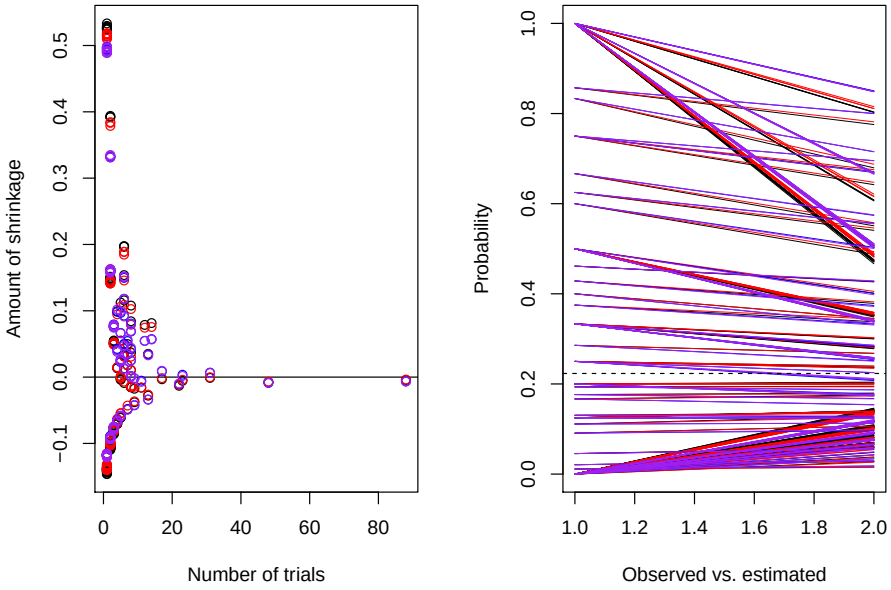
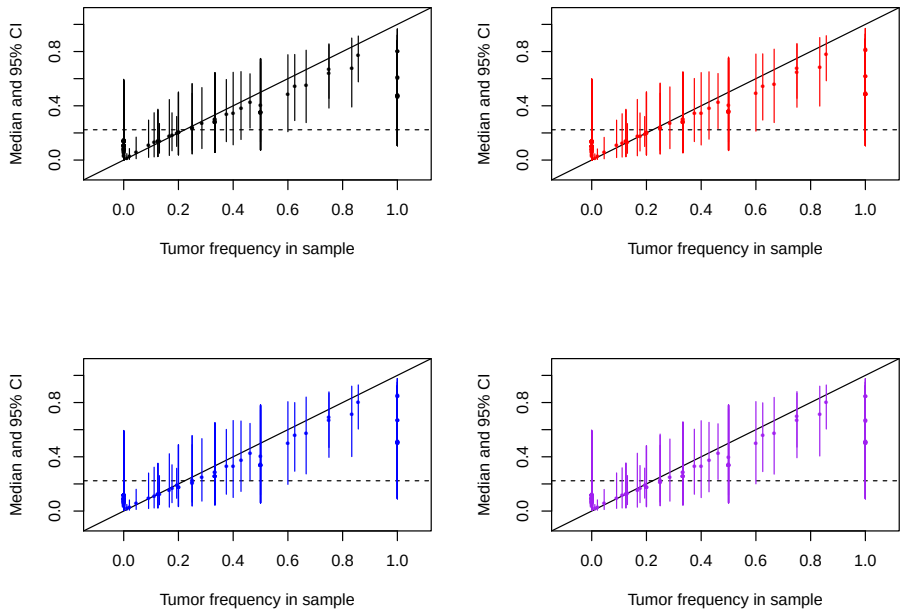

Amount of shrinkage
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0
20
40
60
80
Number of trials
Probability
0.0
0.2
0.4
0.6
0.8
1.0
1.0
1.2
1.4
1.6
1.8
2.0
Observed vs. estimated

Figure 9


Median and 95% CI
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
Tumor frequency in sample

Median and 95% CI
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
Tumor frequency in sample

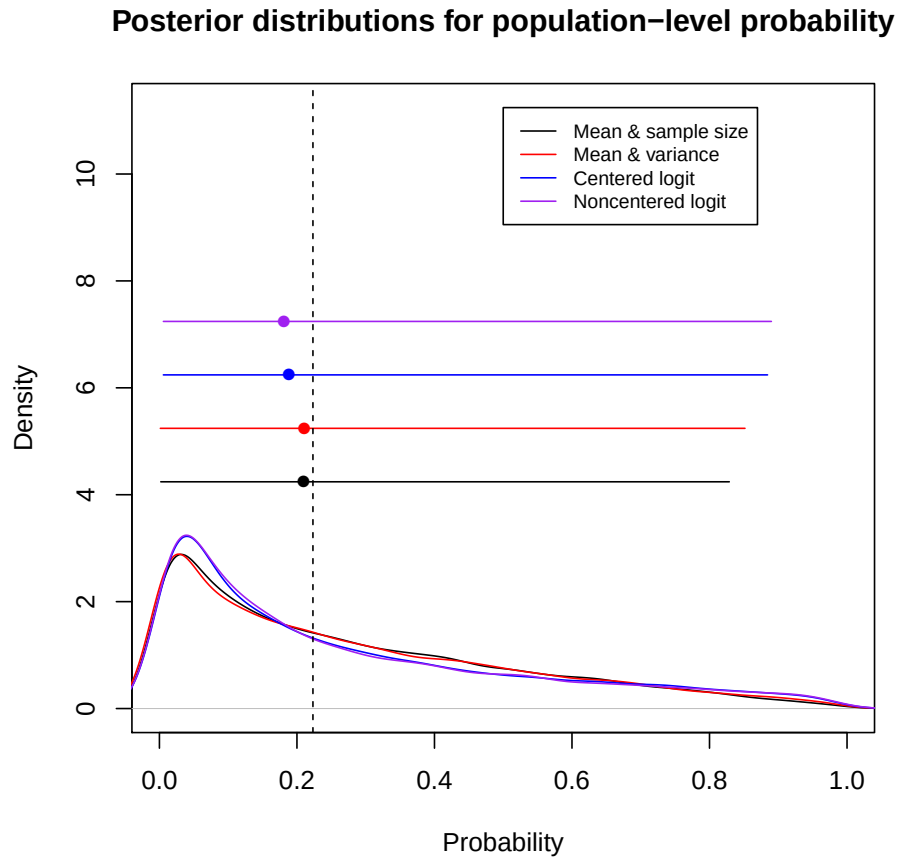

Posterior distributions for population−level probability
Mean & sample size
Mean & variance
Centered logit
Noncentered logit
Density
0
2
4
6
8
10
0.0
0.2
0.4
0.6
0.8
1.0
Probability

Figure 10