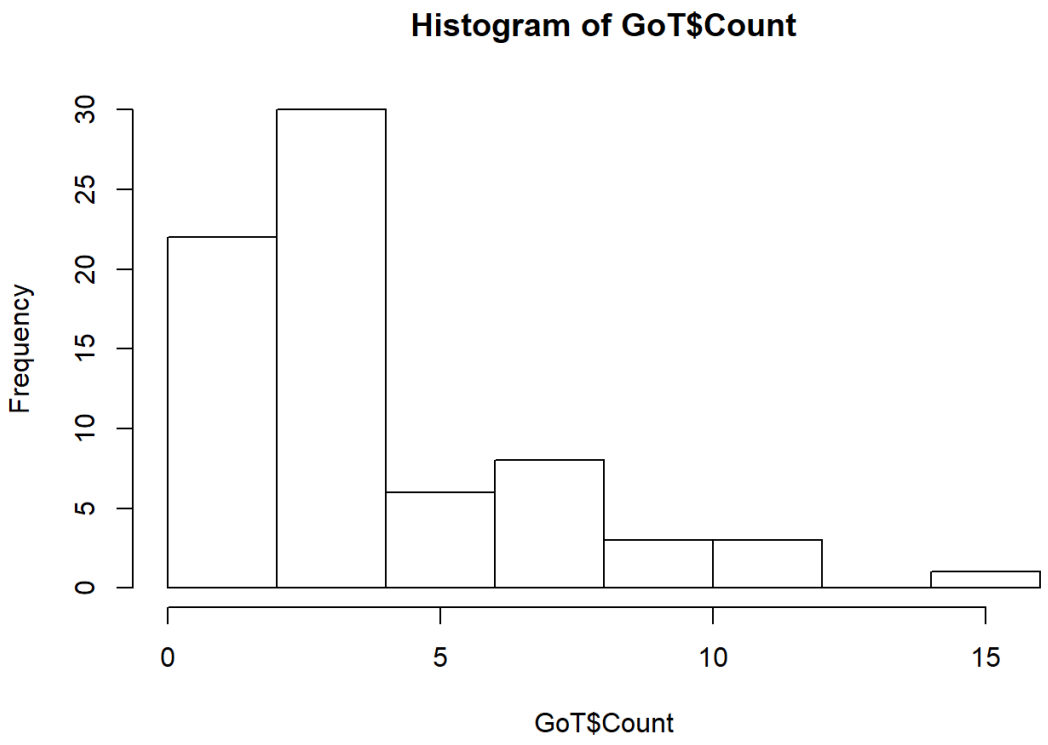
Lab3

Bingying Liu

2/3/2020

Exericise 1: Plot a histogram of the death counts.

```
GoT <- read_xlsx("GoT_Deaths.xlsx", col_names = T)
hist(GoT$Count)
```

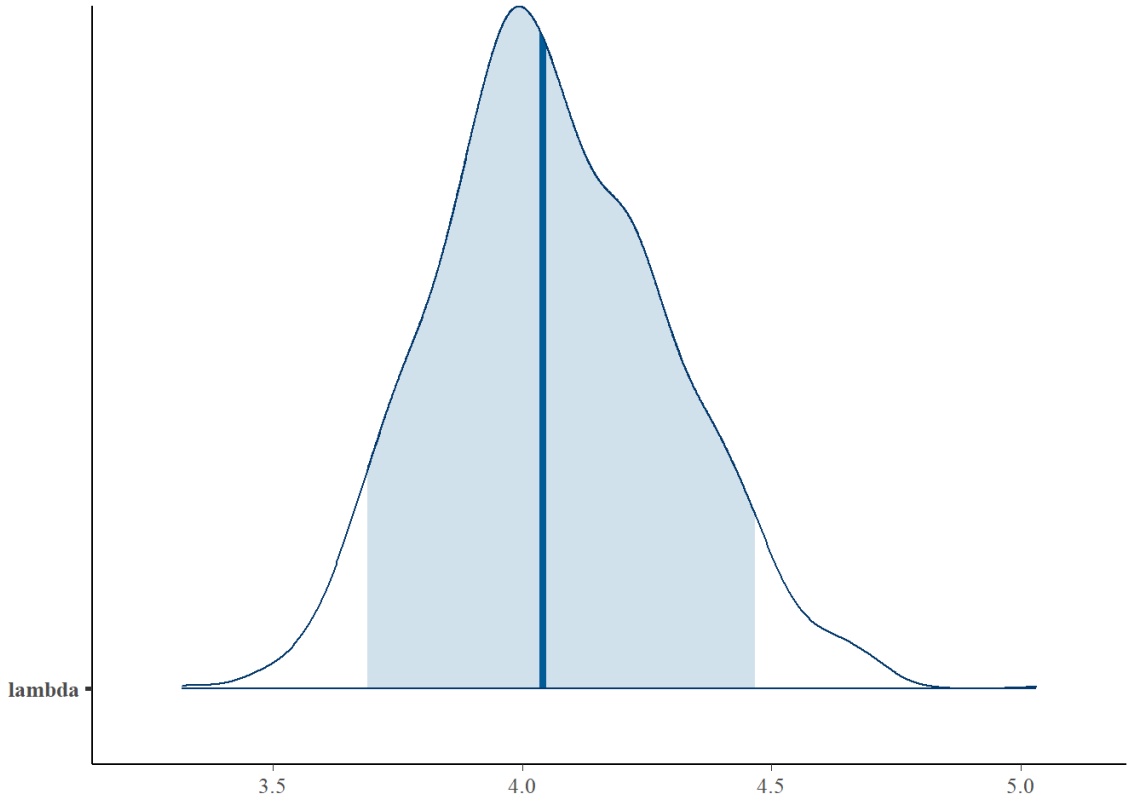


Exercise 2: Plot the smoothed posterior distribution for λ with a 90% Highest Posterior Density region.

```
y <- GoT$Count
n <- length(y)

# Poisson Model
stan_dat <- list(y = y, N = n)
fit <- stan("lab-03-poisson-simple.stan", data = stan_dat, refresh = 0, chains = 2)
```

```
## Warning in readLines(file, warn = TRUE): incomplete final line found on 'C:
## \Users\bl199\Desktop\Lab3\lab-03-poisson-simple.stan'
```

```
lambda_draws <- as.matrix(fit, pars = "lambda")

mcmc_areas(lambda_draws, prob = 0.9)
```



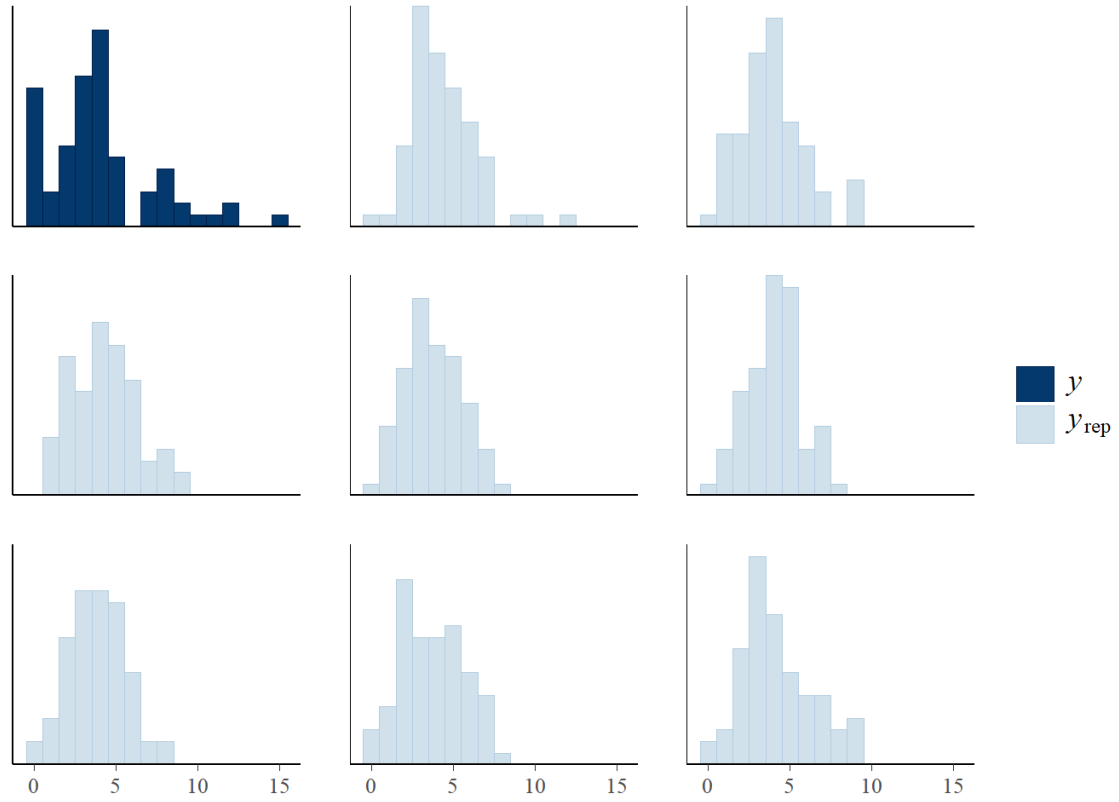

```
print(fit, pars = "lambda")
```

```
## Inference for Stan model: lab-03-poisson-simple.
## 2 chains, each with iter=2000; warmup=1000; thin=1;
## post-warmup draws per chain=1000, total post-warmup draws=2000.
##
##        mean se_mean   sd 2.5% 25%  50%  75% 97.5% n_eff Rhat
## lambda 4.06    0.01 0.24 3.63 3.9 4.04 4.22  4.55   732    1
##
## Samples were drawn using NUTS(diag_e) at Mon Feb 03 14:59:28 2020.
## 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).
```

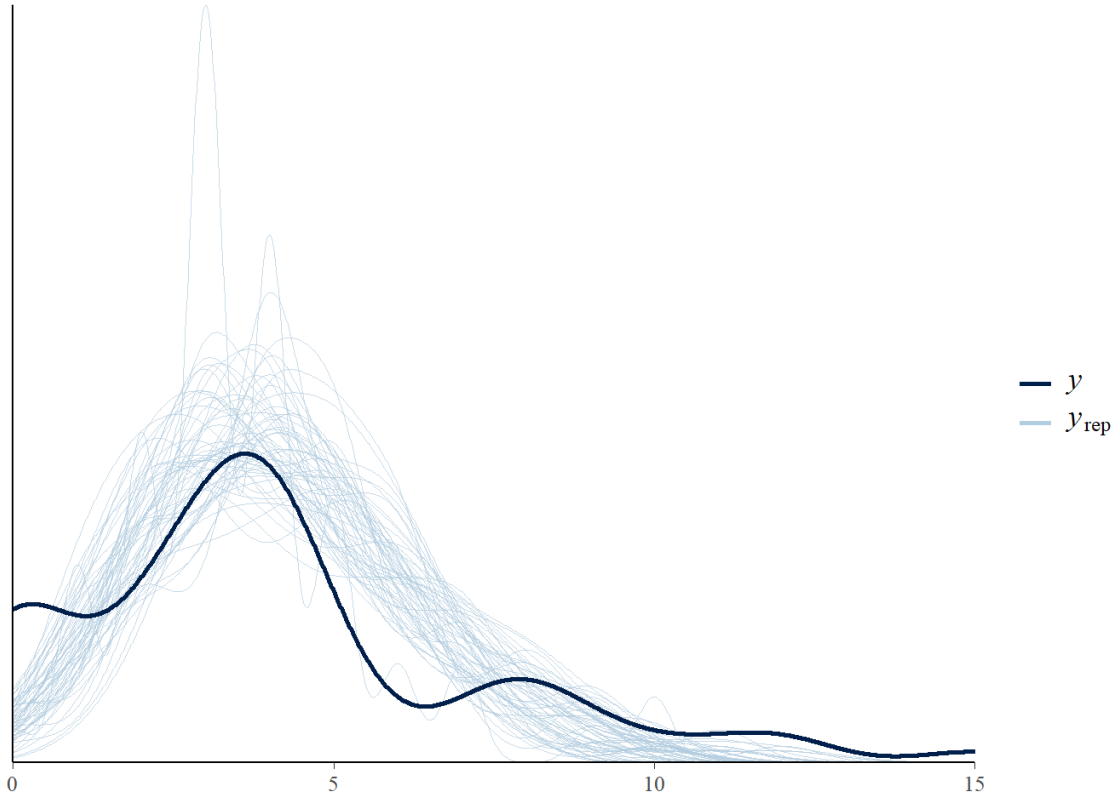
Exercise 3: Generate posterior predictive samples using the posterior values

```
y_rep <- apply(lambda_draws, 1, function(x){rpois(n = n, lambda = x)}) %>% t()
```

```
# compare the empirical distribution of data y to distribution of simulated date
ppc_hist(y, y_rep[1:8, ], binwidth = 1)
```




```
# compare density estimates
ppc_dens_overlay(y, y_rep[1:60, ])
```

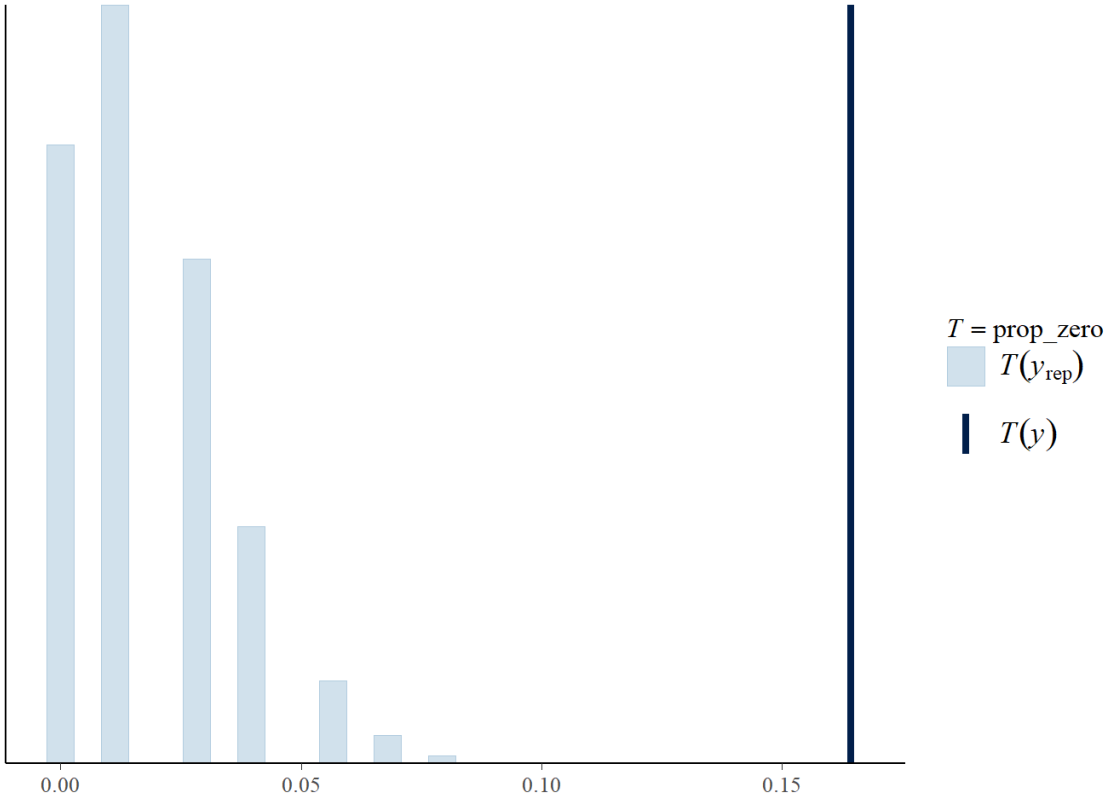


```
# compare proportion of zeros
prop_zero <- function(x){
  mean(x == 0)
}
prop_zero(y)
```

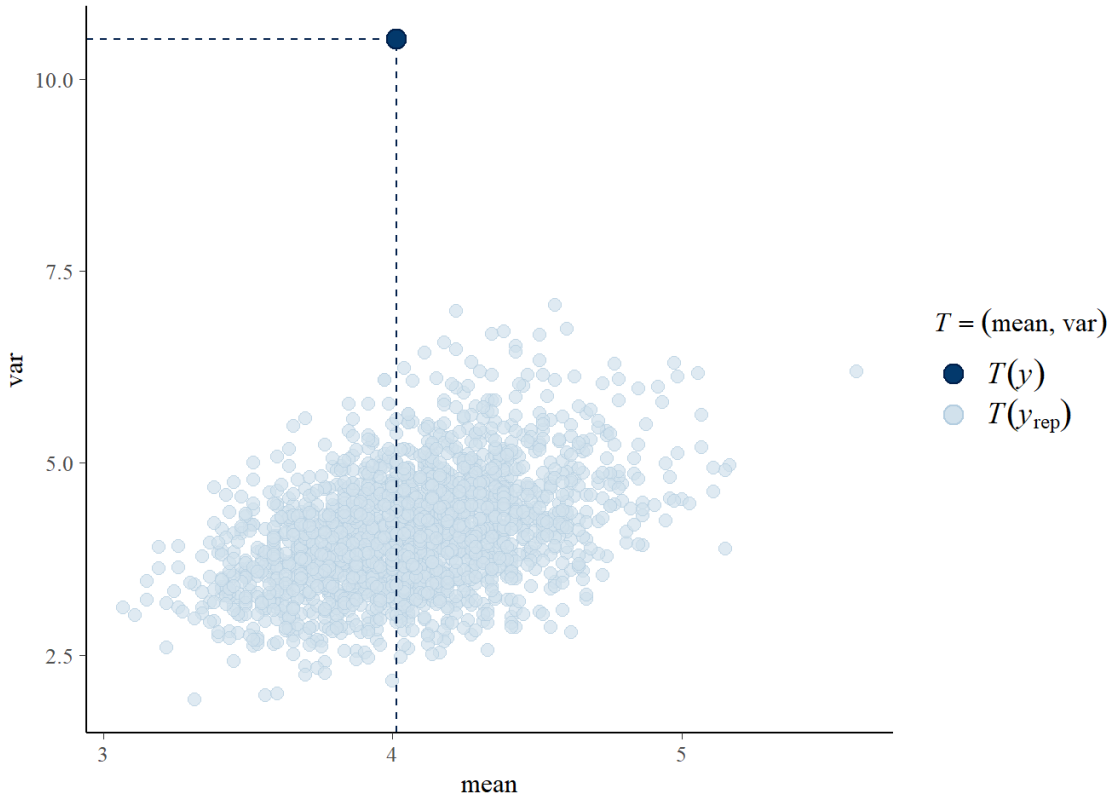
```
## [1] 0.1643836
```

```
ppc_stat(y, y_rep, stat = "prop_zero")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

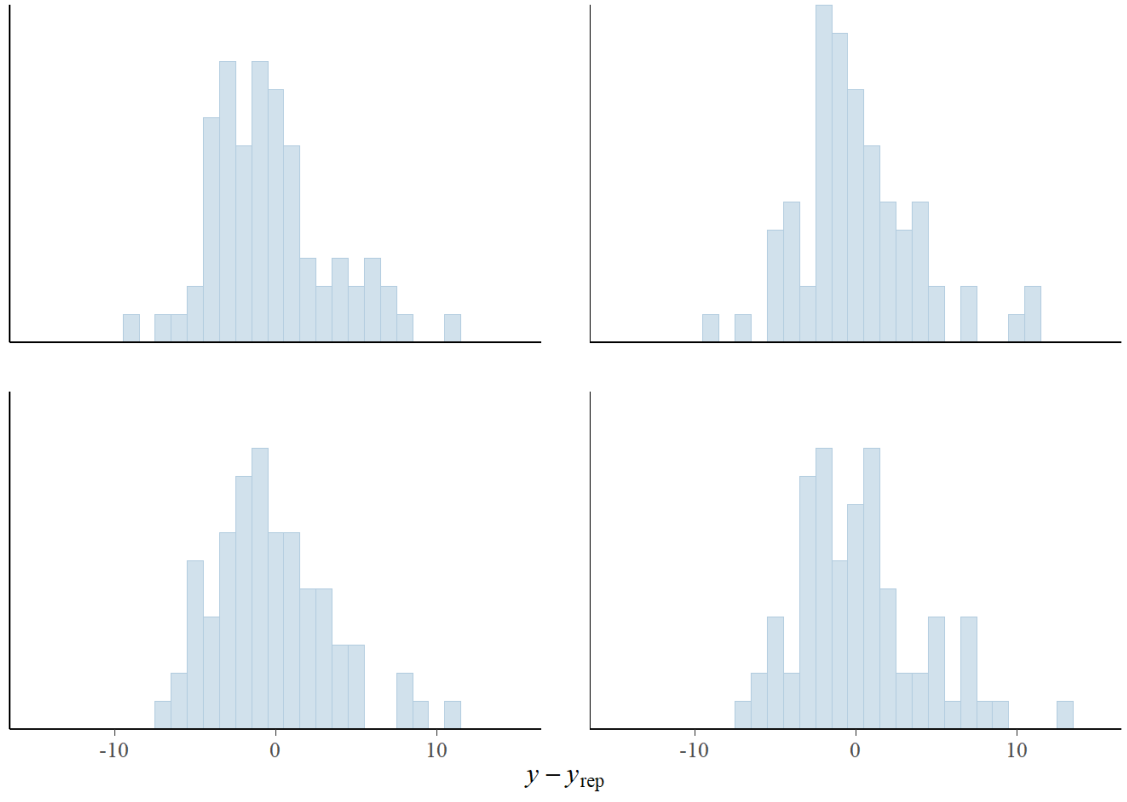



```
# plot the means and variances for all of simulated datasets
ppc_stat_2d(y, y_rep, stat = c("mean", "var"))
```




```
# plot predicted errors
ppc_error_hist(y, y_rep[1:4, ], binwidth = 1) + xlim(-15, 15)
```

```
## Warning: Removed 8 rows containing missing values (geom_bar).
```



Exercise 4: Based on these PPCs, does this model appear to be a good fit for the data?

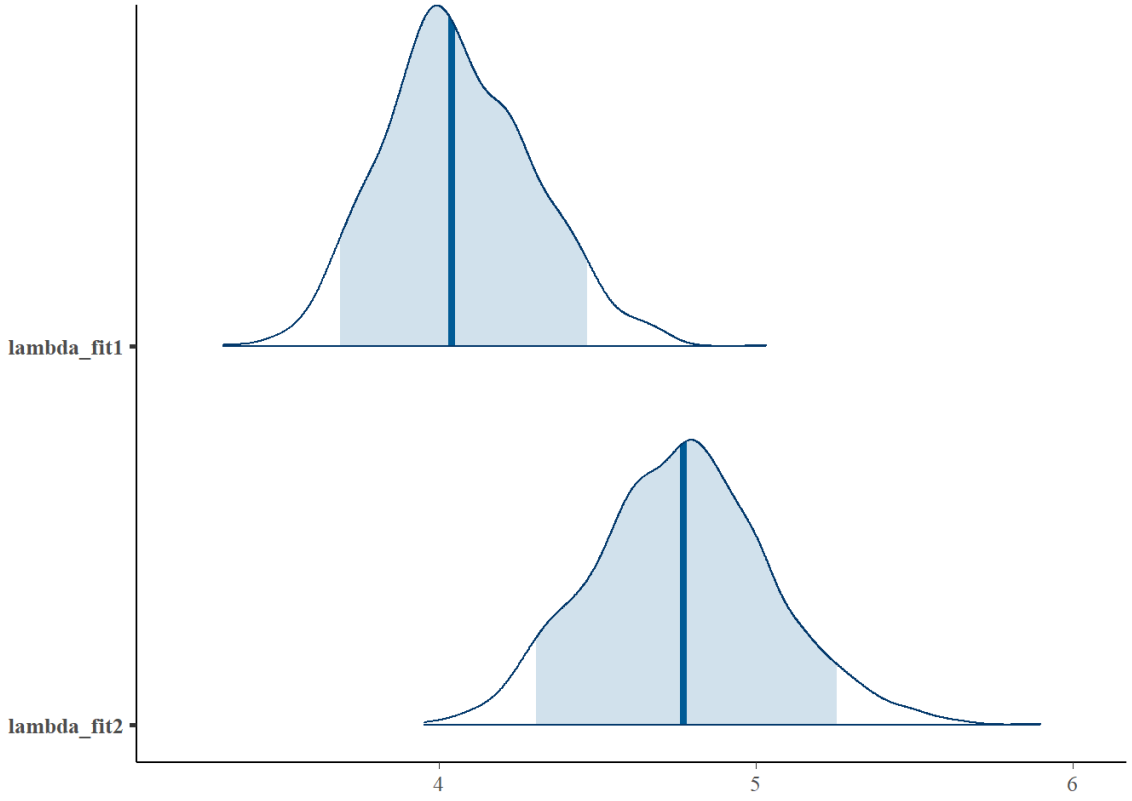
No, thie model doesn't model 0s as in oberserved data. From histogram, density estimate, proportion of zeros plots, we can all see that there are many more 0-valued observations in the observed data than there are in the simulated data. And the error plot shows significantly larger difference between y and y_rep around 0.

Exercise 5: Using the code provided for the simple Poisson model, simulate draws from the posterior density of λ with the "Poisson Hurdle" model.

```
fit2 <- stan("lab-03-poisson-hurdle.stan", data = stan_dat, refresh = 0, chains = 2)
```

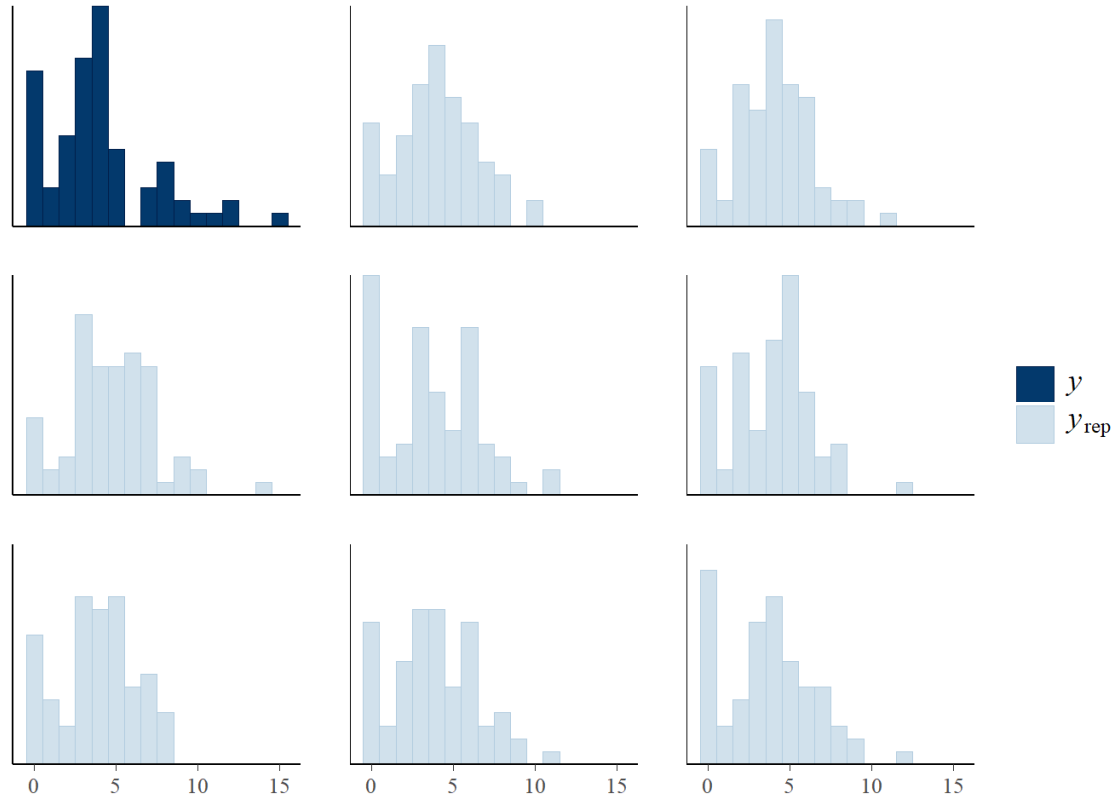
```
## Warning in readLines(file, warn = TRUE): incomplete final line found on 'C:
## \Users\bl199\Desktop\Lab3\lab-03-poisson-hurdle.stan'
```

```
lambda_draws2 <- as.matrix(fit2, pars = "lambda")

lambdas <- cbind(lambda_fit1 = lambda_draws[, 1],
                 lambda_fit2 = lambda_draws2[, 1])

# Shade 90% interval
mcmc_areas(lambdas, prob = 0.9)
```

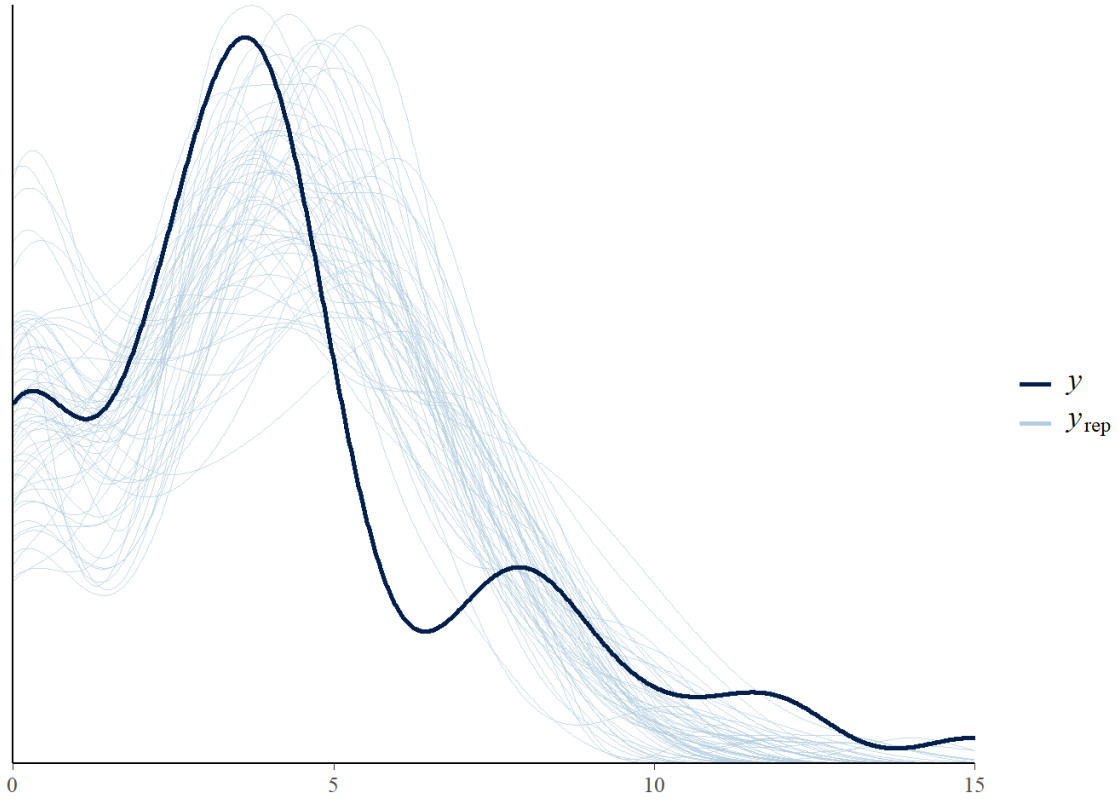


Exercise 6: Produce the same PPC vizs as before for the new results. Comment on how this second model compares to both the observed data and to the simple Poisson model.

```
y_rep2 <- as.matrix(fit2, pars = "y_rep")

ppc_hist(y, y_rep2[1:8, ], binwidth = 1)
```




```
ppc_dens_overlay(y, y_rep2[1:60, ])
```

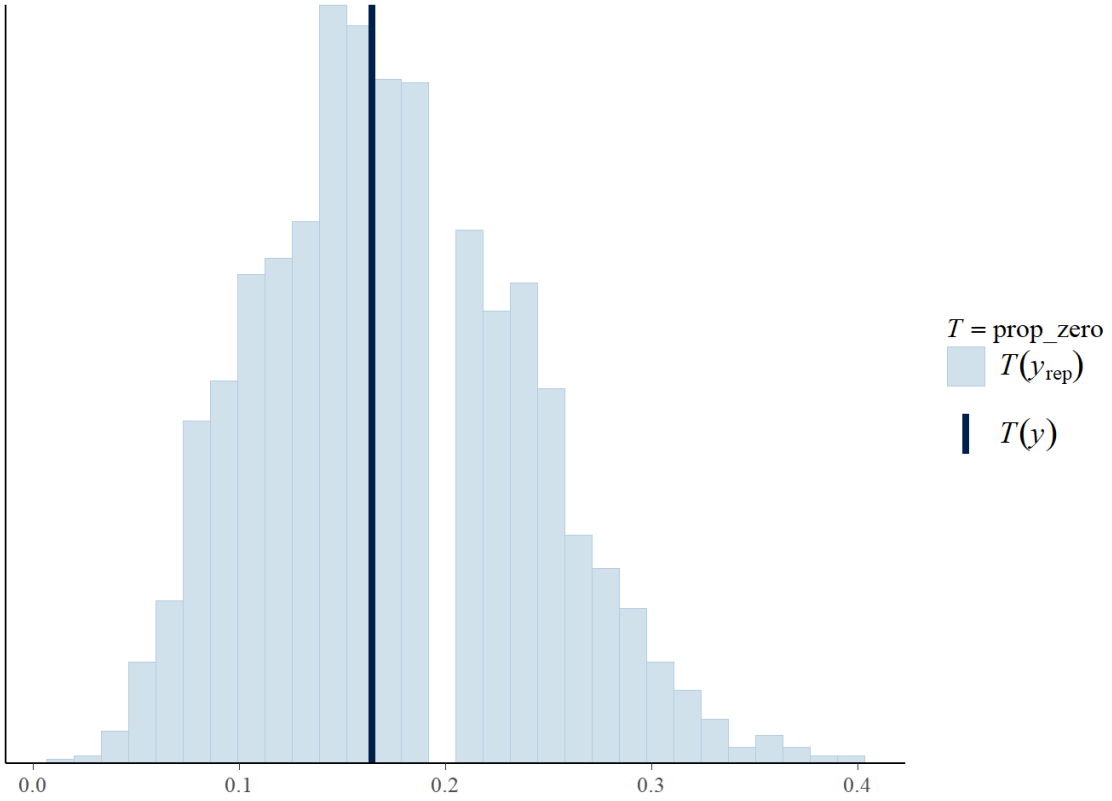



```
prop_zero <- function(x){
  mean(x == 0)
}
prop_zero(y)
```

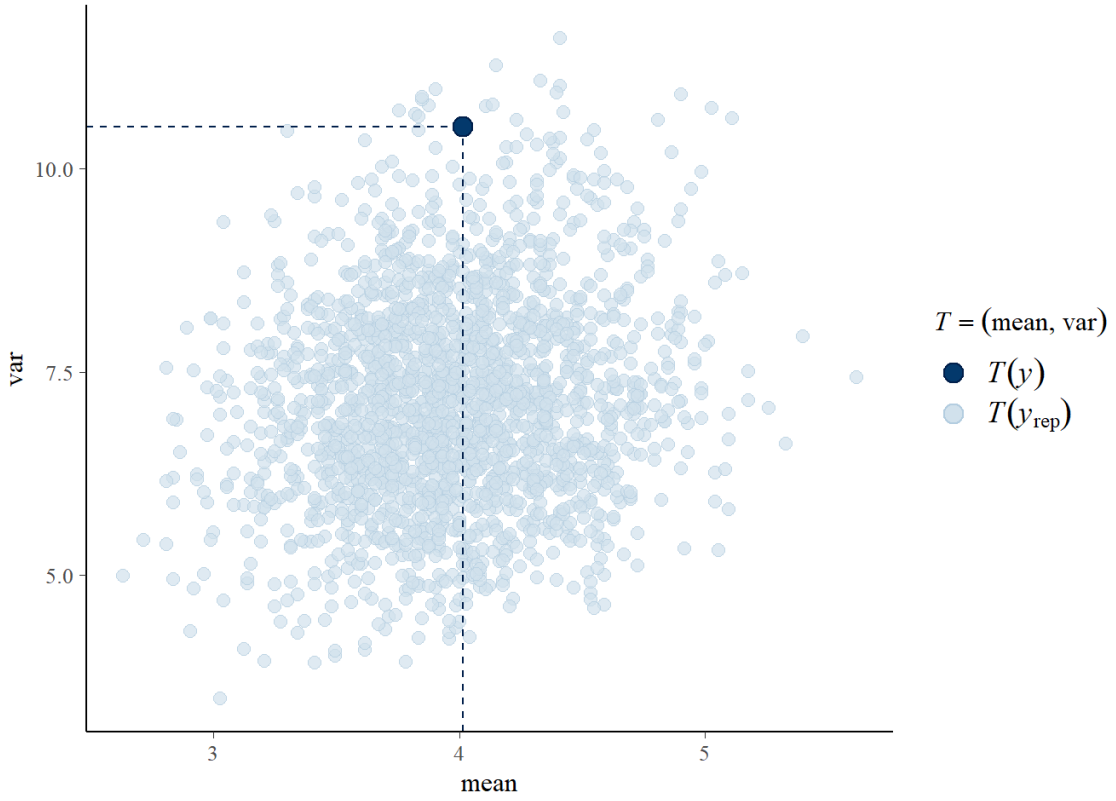
```
## [1] 0.1643836
```

```
ppc_stat(y, y_rep2, stat = "prop_zero")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

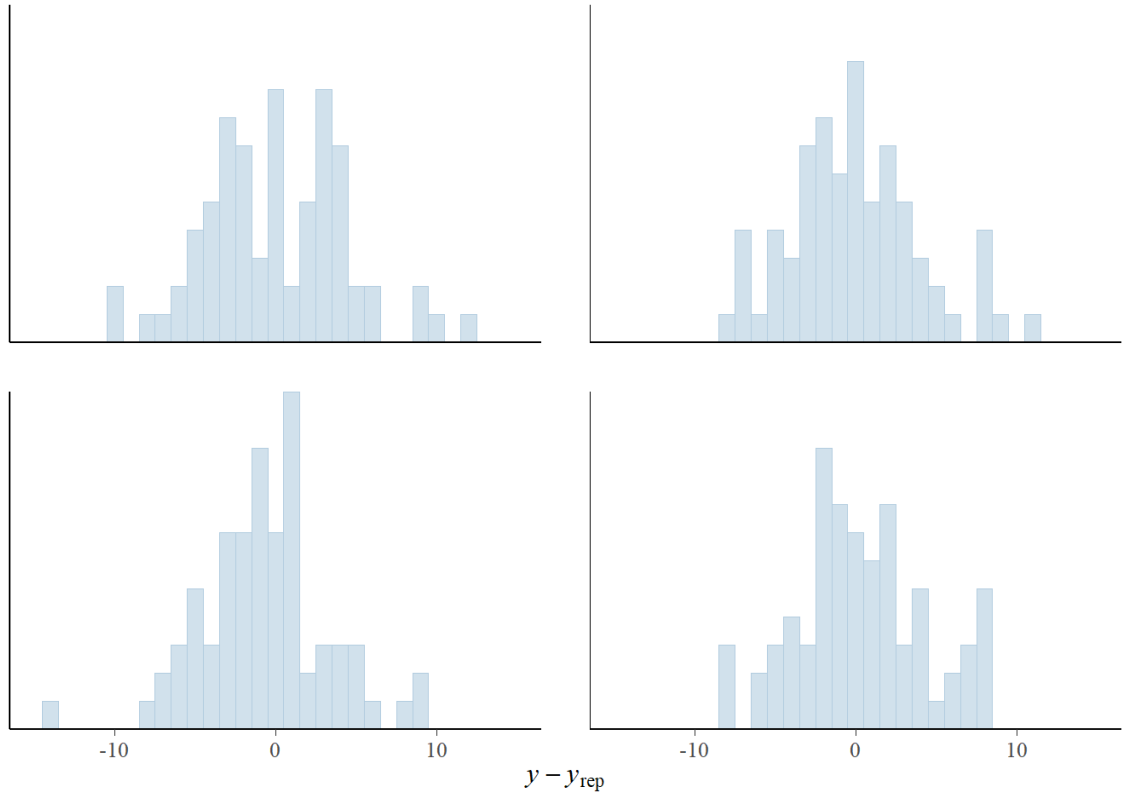



```
ppc_stat_2d(y, y_rep2, stat = c("mean", "var"))
```




```
ppc_error_hist(y, y_rep2[1:4, ], binwidth = 1) + xlim(-15, 15)
```

```
## Warning: Removed 8 rows containing missing values (geom_bar).
```



The second model is better at modelling 0-valued observations than the simple Poisson model. Distribution plot of both empirical and posterior predictive data shows peak at 0; density estimate for several y_rep has a similar curvature around 0; empirical proportion of 0 lies in the all of the posterior predictive proportions of 0 with similar height; mean and variance point of the observed statistics also lies

in the posterior predictive point cloud; difference between y and yrep decreases at 0 in error plot.

Exercise 7: Which model performs better in terms of prediction?

```
## Leave-one-out cross-validation
log_lik1 <- extract_log_lik(fit, merge_chains = FALSE)
r_eff1 <- relative_eff(exp(log_lik1))
(loo1 <- loo(log_lik1, r_eff = r_eff1))
```

```
##
## Computed from 2000 by 73 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo   -203.2 14.5
## p_loo         2.6  0.5
## looic       406.4 28.9
## ------
## Monte Carlo SE of elpd_loo is 0.1.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

```
log_lik2 <- extract_log_lik(fit2, merge_chains = FALSE)
r_eff2 <- relative_eff(exp(log_lik2))
(loo2 <- loo(log_lik2, r_eff = r_eff2))
```

```
##
## Computed from 2000 by 73 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo   -183.3 10.8
## p_loo         2.8  0.5
## looic       366.6 21.7
## ------
## Monte Carlo SE of elpd_loo is 0.0.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

```
compare(loo1, loo2)
```

```
## Warning: 'compare' is deprecated.
## Use 'loo_compare' instead.
## See help("Deprecated")
```

```
## elpd_diff        se
##      19.9       8.6
```

Second model performs better since its point estimate of elpd_loo is lower and standard error is lower than first model as well.

Exercise 8: Why are PPCs important?

We use posterior predictice checks to look for systematic discrepancies between real and simulated data. Therefore, PPCs are important.

Exercise 9: Was the second model a good fit for the data? Why or why not?

The second model is a good fit for the data although it's not a perfect fit. From the density estimate plot, we can see that from value 5-15, density estimate of the observed has some wave-like characteristics while posterior predictive density estimate doesn't have. Therefore, there's potential for improvement.

Exercise 10: If someone reported a single LOOCV error to you, would that be useful? Why or why not?

The result will be kind of biased, because predictive error is completely determined by the quality of one validation point, and the performance of one point couldn't determine performance of the model on other unseen data. So this will not be as useful.