Zero-inflated count models

ZIP vs. ZINB for the fishing data: MM, ML, Bayes (SIR) and `pscl`

Hedibert Freitas Lopes

Insper

These slides were prepared with the assistance of AI (Claude, Anthropic). The model, derivations and R code were reviewed by the author.

Roadmap: two models $\times$ three estimation strategies

	Method of moments	Maximum likelihood	Bayes (SIR)
ZIP(λ, π)	closed form (2 moments)	closed form (Lambert W)	uniform priors
ZINB(μ, θ, π)	closed form (3 factorial moments)	numerical (`pscl`)	equally vague prior

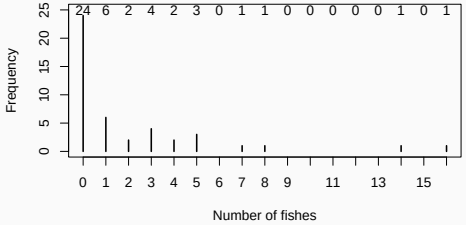
1. The data: fish caught by groups visiting a state park
2. ZIP: MLE, MM, SIR and posterior predictive checks
3. Why the ZIP is not enough: residual overdispersion
4. ZINB: MLE, MM, SIR and posterior predictive checks
5. Comparing all six fits; AIC, LR tests and Bayes factor

The tags [R: Ck] on each slide point to chunk Ck of the R script in the appendix.

The data

Fishing at a state park

- Wildlife biologists surveyed 250 groups: fish caught (`count`), number of children (`child`), group size (`persons`), camper or not (`camper`).
- Some visitors do not fish, but whether a group fished is not recorded. These groups add excess zeros to the zeros from groups that fished and caught nothing.
- Subset used: `camper==1 & child==1`: $n = 45$ groups, $n_0 = 24$ zeros.



$\bar{y} = 2.00, \quad s^2 = 12.32$

The variance is six times the mean. A plain Poisson cannot fit this.

53% of the groups caught nothing. [R: C1]

Sources: `https://stats.oarc.ucla.edu/r/dae/zip/`; `https://en.wikipedia.org/wiki/Zero-inflated_model`

Zero-inflated Poisson

The ZIP model

With probability π a group does not fish (structural zero). Otherwise its catch is Poisson(λ):

$$\Pr(y_i \mid \lambda, \pi) = \begin{cases} \pi + (1-\pi)e^{-\lambda}, & y_i = 0, \\ (1-\pi)\,\dfrac{\lambda^{y_i} e^{-\lambda}}{y_i!}, & y_i = 1, 2, \ldots \end{cases}$$

Moments:

$$E(y_i \mid \lambda, \pi) = (1-\pi)\lambda, \qquad V(y_i \mid \lambda, \pi) = (1-\pi)\lambda(1+\pi\lambda) \;\geq\; E(y_i).$$

- Zero inflation by itself already creates overdispersion: $V/E = 1 + \pi\lambda$.
- $\pi = 0$ gives back the Poisson.

Maximum likelihood

$$L(\lambda, \pi) \propto \left[\pi + (1-\pi)e^{-\lambda}\right]^{n_0} \left[(1-\pi)e^{-\lambda}\right]^{n-n_0} \lambda^{n\bar{y}}.$$

The MLEs have a closed form:

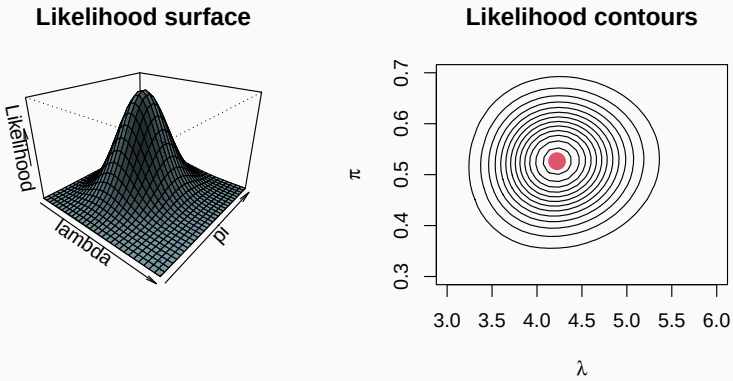
$$\hat{\lambda}_{mle} = W_0\left(-\bar{z}e^{-\bar{z}}\right) + \bar{z}, \qquad \hat{\pi}_{mle} = 1 - \frac{\bar{y}}{\hat{\lambda}_{mle}}, \qquad \bar{z} = \frac{\sum_i y_i}{n - n_0},$$

where W_0 is the principal branch of Lambert's W function. Here $\bar{z} = 90/21 = 4.286$, so

$$\hat{\lambda}_{mle} = 4.223, \qquad \hat{\pi}_{mle} = 0.526.$$

[R: C2]

Likelihood surface



The likelihood is unimodal and roughly elliptical around the MLE (red dot). [R: C2]

Method of moments

Set $\bar{y} = (1-\pi)\lambda$ and $s^2 = (1-\pi)\lambda(1+\pi\lambda)$ and solve:

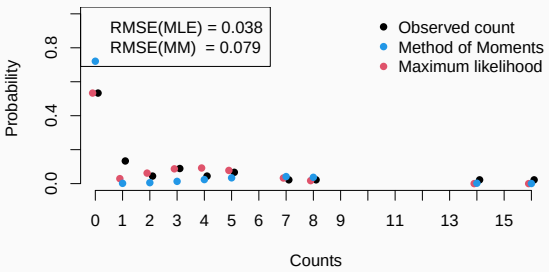
$$\hat{\lambda}_{mm} = \frac{s^2 + \bar{y}^2}{\bar{y}} - 1, \qquad \hat{\pi}_{mm} = \frac{s^2 - \bar{y}}{s^2 + \bar{y}^2 - \bar{y}}.$$

	λ	π
Method of moments	7.159	0.721
Maximum likelihood	4.223	0.526

- MM matches the variance exactly. Here that variance comes mostly from two groups (14 and 16 fish), so MM pushes λ up and π up with it.

[R: C3]

Comparing the fitted ZIPs



Expected counts ($n \times$ prob.)

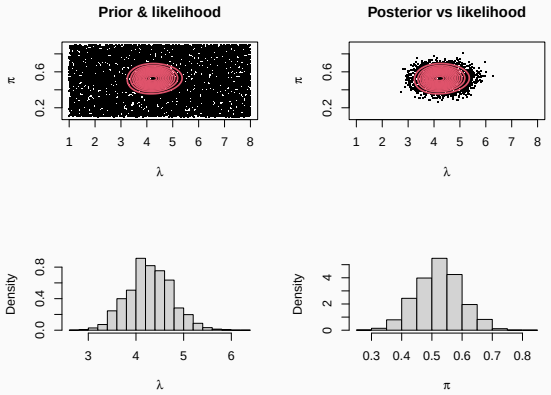
y	Obs.	MLE	MM
0	24	24	32
1	6	1	0
2	2	3	0
3	4	4	1
4	2	4	1
5	3	3	2
7	1	1	2
8	1	1	2
14	1	0	0
16	1	0	0

- The MLE wins (RMSE 0.038 vs 0.079) and reproduces n_0 exactly.
- Both versions miss the six 1's and give almost no mass to 14 or 16.

[R: C4]

Posterior inference via SIR

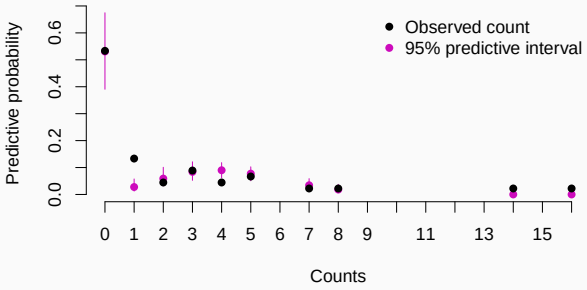
Priors: $\lambda \sim U(1, 8)$ and $\pi \sim U(0.1, 0.9)$. Draw $M = 10{,}000$ pairs, weight each by its likelihood, and resample.



	2.5%	50%	97.5%
λ	3.43	4.28	5.25
π	0.38	0.52	0.67

The posterior is centred on the MLE. π is quite uncertain: n is small, and a zero can come from either source. [R: C5]

Posterior predictive check



95% posterior predictive intervals for $\Pr(Y = k)$ compared with the observed frequencies.

The intervals miss $k = 1$ and $k = 4$, and they miss the large catches $(k = 14, 16)$ by orders of magnitude.

The Poisson component is too light-tailed. [R: C6]

Zero-inflated negative binomial

Overdispersion that zero inflation does not explain

At the MLE, the ZIP matches the mean but not the variance:

$$\hat{E}(y) = (1 - \hat{\pi})\hat{\lambda} = 2.00 = \bar{y}, \qquad \hat{V}(y) = (1 - \hat{\pi})\hat{\lambda}(1 + \hat{\pi}\hat{\lambda}) = 6.45 \ll s^2 = 12.32.$$

- Even among groups that do fish, catches vary much more than a Poisson allows: some groups are much better at fishing than others.
- Natural fix: a Gamma–Poisson mixture, i.e. the negative binomial, for the count part.

The negative binomial, (`mu`, `size`) parametrization

If $y \mid u \sim \text{Poisson}(\mu u)$ and $u \sim \text{Gamma}(\theta, \theta)$ (mean 1), then

$$\Pr(y \mid \mu, \theta) = \frac{\Gamma(y+\theta)}{\Gamma(\theta)\, y!}\, p^{\theta}(1-p)^{y}, \qquad p = \frac{\theta}{\theta+\mu},$$

$$E(y) = \mu, \qquad V(y) = \mu + \frac{\mu^2}{\theta}.$$

- θ (`size`) is the dispersion parameter. As $\theta \to \infty$ we get the Poisson(μ).
- In R: `dnbinom(y, size=theta, mu=mu)`, or equivalently `dnbinom(y, size=theta, prob=p)`.
- `pscl` reports $\log \mu$ (count intercept), $\log \theta$ and the logit of π.

The ZINB model

$$\Pr(y_i \mid \mu, \theta, \pi) = \begin{cases} \pi + (1-\pi)\left(\dfrac{\theta}{\theta+\mu}\right)^{\theta}, & y_i = 0, \\ (1-\pi)\,\mathrm{NB}(y_i \mid \mu, \theta), & y_i = 1, 2, \ldots \end{cases}$$

$$E(y_i) = (1-\pi)\mu, \qquad V(y_i) = (1-\pi)\mu\left(1 + \frac{\mu}{\theta} + \pi\mu\right).$$

- Overdispersion now has two sources: zero inflation ($\pi\mu$) and heterogeneity among groups that fish (μ/θ).
- Nested cases: $\theta \to \infty$ gives the ZIP; $\pi = 0$ gives the NB.

ZINB: maximum likelihood with `pscl::zeroinfl`

Intercept-only model for both parts: `zeroinfl(y ~ 1 | 1, dist = ...)`.

ZIP check [R: C7]: $e^{1.4405} = 4.223 = \hat{\lambda}_{mle}$ and $\text{logit}^{-1}(0.1057) = 0.526 = \hat{\pi}_{mle}$, which reproduces the Lambert-W closed form.

ZINB [R: C8] (no closed form; the likelihood is maximized numerically):

	estimate (s.e.)	natural scale	
count intercept	1.062 (0.458)	$\hat{\mu} = 2.892$	NB mean for groups that fish
$\log \theta$	-0.394 (0.956)	$\hat{\theta} = 0.674$	strong heterogeneity
zero intercept	-0.808 (1.301)	$\hat{\pi} = 0.308$	non-fishing probability

$\hat{p} = \hat{\theta}/(\hat{\theta} + \hat{\mu}) = 0.189$. Implied moments: $\hat{E}(y) = 2.00$, $\hat{V}(y) = 12.36 \approx s^2 = 12.32$.

ZINB: method of moments

Three parameters require three moments. Factorial moments $f_k = E[Y(Y-1)\cdots(Y-k+1)]$ are easiest:

$$f_k = (1-\pi)\,\mu^k \prod_{j=0}^{k-1}\Big(1+\frac{j}{\theta}\Big) \quad\Rightarrow\quad a \equiv \frac{f_2}{f_1} = \mu\Big(1+\frac{1}{\theta}\Big), \quad b \equiv \frac{f_3}{f_2} = \mu\Big(1+\frac{2}{\theta}\Big).$$

Solving gives closed-form estimators:

$$\hat\mu_{mm} = 2a - b, \qquad \hat\theta_{mm} = \frac{\hat\mu_{mm}}{b-a}, \qquad \hat\pi_{mm} = 1 - \frac{f_1}{\hat\mu_{mm}}.$$

With $f_1 = 2$, $f_2 = 14.04$, $f_3 = 140.93$: $\hat\mu_{mm} = 4.010$, $\hat\theta_{mm} = 1.331$, $\hat\pi_{mm} = 0.501$.

- f_3 depends heavily on the two large catches (14 and 16 contribute 87% of it), so the MM estimates move far from the MLE, just as in the ZIP case.
- Matching $(\bar y, s^2, \hat p_0)$ instead gives (2.906, 0.683, 0.312), which is almost the MLE.

[R: C9]

ZINB: priors as vague as the ZIP's

ZIP

- $\lambda \sim U(1, 8)$
- $\pi \sim U(0.1, 0.9)$

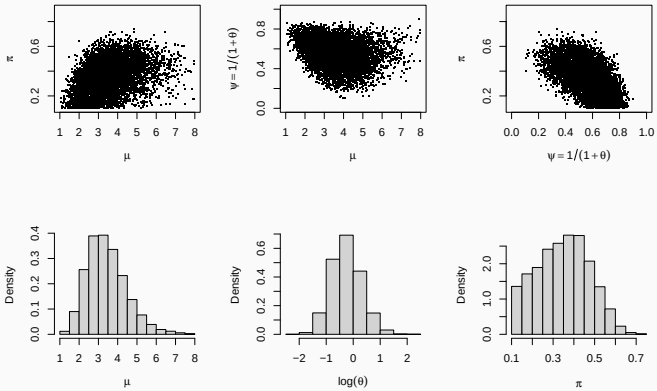
ZINB

- $\mu \sim U(1, 8)$ (same role as λ)
- $\pi \sim U(0.1, 0.9)$ (same)
- $\psi = \dfrac{1}{1+\theta} \sim U(0, 1)$

- $\theta \in (0, \infty)$ has no flat proper prior, so we put the uniform on a bounded transformation. $\psi \to 0$ gives the Poisson (ZIP) limit and $\psi \to 1$ gives extreme overdispersion.
- Implied prior: $p(\theta) = (1+\theta)^{-2}$, $\Pr(\theta \le t) = t/(1+t)$, prior median $\theta = 1$.
- SIR as before: $M = 100{,}000$ prior draws are weighted by the likelihood and 10,000 are resampled. Effective sample size $\approx$ 8,300.

[R: C10]

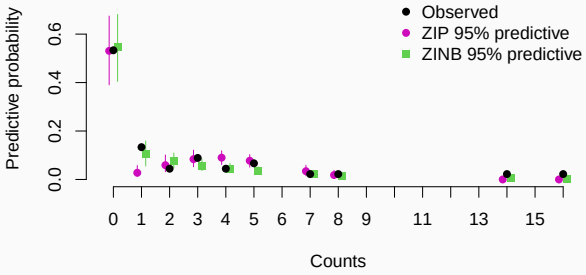
ZINB: posterior via SIR



	2.5%	50%	97.5%
μ	1.79	3.31	5.84
θ	0.29	0.75	2.41
π	0.12	0.35	0.58

μ, θ and π are strongly dependent a posteriori: more non-fishers ($\pi \uparrow$) goes with a larger mean among fishers ($\mu \uparrow$) and less heterogeneity ($\psi \downarrow$). [R: C10]

Posterior predictive: ZIP vs ZINB



95% posterior intervals for $\Pr(Y = k)$.

The ZINB interval covers the six 1's, which the ZIP interval misses, and it puts non-negligible mass on 14 and 16.

Caveat: these intervals describe the model probabilities. They do not include the sampling noise in the observed frequencies ($n = 45$), so a single observed frequency can fall outside even when the fit is good. [R: C11]

Two models, three methods

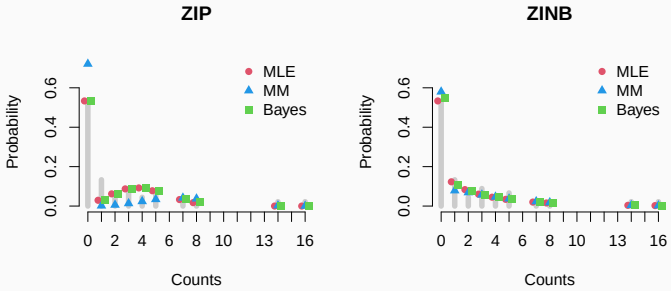
All estimates

Model	Parameter	MM	MLE	Bayes: median (95% CI)
ZIP	λ	7.159	4.223	4.28 (3.43, 5.25)
	π	0.721	0.526	0.52 (0.38, 0.67)
ZINB	μ	4.010	2.892	3.31 (1.79, 5.84)
	θ	1.331	0.674	0.75 (0.29, 2.41)
	π	0.501	0.308	0.35 (0.12, 0.58)

- Under the vague priors, Bayes and ML agree closely. The posterior median sits slightly on the side of the right-skewed posterior.
- MM is the outlier in both models: it is driven by the high-order moments, which here depend on two observations.
- Under the ZINB, π is much smaller and much more uncertain, because a zero can now also come from a group that fishes and catches nothing.

[R: C2, C3, C5, C8–C10]

Fitted probabilities: 2 models $\times$ 3 methods



RMSE	MM	MLE	Bayes
ZIP	0.079	0.038	0.038
ZINB	0.029	0.021	0.022

Bayes = posterior mean of $\Pr(Y = k)$. Any ZINB fit beats the best ZIP fit, and the choice of model matters more than the choice of method.

[R: C12]

Expected counts

		ZIP			ZINB		
y	Obs.	MM	MLE	Bayes	MM	MLE	Bayes
0	24	32.4	24.0	23.9	26.1	24.0	24.6
1	6	0.1	1.3	1.3	3.5	5.5	4.8
2	2	0.3	2.8	2.8	3.1	3.8	3.5
3	4	0.6	3.9	3.8	2.6	2.7	2.6
4	2	1.1	4.1	4.1	2.1	2.0	2.0
5	3	1.5	3.5	3.5	1.7	1.5	1.6
7	1	1.9	1.5	1.6	1.0	0.9	0.9
8	1	1.7	0.8	0.9	0.8	0.7	0.7
14	1	0.1	0.0	0.0	0.2	0.2	0.2
16	1	0.0	0.0	0.0	0.1	0.1	0.1

The ML fits reproduce the 24 zeros exactly; this is a property of intercept-only zero-inflated MLEs. [R: C12]

Who are the zeros?

Share of the zeros attributed to non-fishing groups, $\pi / \Pr(Y = 0)$:

	MM	MLE	Bayes (post. mean)
ZIP	1.000	0.987	0.986
ZINB	0.865	0.578	0.630

- In the ZIP, a Poisson(4.2) almost never produces a 0, so essentially every zero is attributed to a group that did not fish.
- In the ZINB, heterogeneity ($\theta < 1$) means many groups that fish still catch nothing. Only about 60% of the zeros are attributed to non-fishing.
- How many visitors don't fish depends far more on the count distribution than on the estimation method.

[R: C12]

Model comparison: frequentist and Bayesian

Model	log-lik	# par.	AIC
Poisson	−131.88	1	265.76
NB	−80.08	2	**164.16**
ZIP	−94.08	2	192.16
ZINB	−79.88	3	165.77

Bayes factor (SIR priors): $p(y) = E_{\text{prior}}[L]$, estimated by averaging the likelihood over the prior draws:

$$\log p(y \mid \text{ZIP}) = -97.35, \quad \log p(y \mid \text{ZINB}) = -83.25,$$

$$\text{BF}_{\text{ZINB,ZIP}} \approx 10^{6.1}.$$

- ZIP vs ZINB: $LR = 28.4$, $p \approx 5 \times 10^{-8}$ (boundary, $\frac{1}{2}\chi^2_1$). The likelihood ratio and the Bayes factor both reject the ZIP decisively.
- NB vs ZINB: $LR = 0.39$, $p = 0.27$. Once the counts are NB, the data show no evidence of zero inflation. Covariates (e.g. `persons`, `livebait`) would be needed to identify π.
- The Bayes factor depends on the prior ranges, unlike posterior summaries. Widening $U(1, 8)$, for example, would change it.

[R: C13]

Take-aways

1. Excess zeros and overdispersion are different problems. The ZIP handles the first. Here the second remains: ZIP variance 6.45 vs $s^2 = 12.32$.
2. **Methods.** MM gives closed forms in both models (2 moments for the ZIP, 3 factorial moments for the ZINB), but it depends on high-order moments and so on a few observations. ML and Bayes with vague priors agree, and Bayes also quantifies the large uncertainty in π and θ.
3. **Models.** The ZINB beats the ZIP under every method (RMSE, LR, BF $\approx 10^6$). It also changes the substantive conclusion: about 99% vs about 60% of the zeros attributed to non-fishing.
4. Without covariates, π and θ are weakly separated, and AIC prefers the NB alone.

Notes on the source materials

Corrections made in going from the original HTML and script to these slides:

- The subset has $n = 45$ observations (the frequency table sums to 45), not 42. The MLE/MM numbers on the HTML page already use $n = 45$.
- ZIP pmf: $k! \to y_i!$. The SIR predictive quantiles used `0.925`; corrected to `0.975`.
- ZINB script: `theta/(theta+mu)` was computed with `mu` set to the *zero-inflation* probability. The NB probability uses the count mean, $\mu = e^{\text{count intercept}}$, giving $p = 0.189$.
- Script: `my` was used before it was defined and `avg_probs` was never defined. Both are now defined in chunk C12.
- The subset is `camper==1 & child==1` (exactly one child). There is no live-bait filter.

R code

R code: map of chunks to slides

Chunk	What it does	Used on slide(s)
C1	read data, subset, frequency plot	3
C2	ZIP likelihood, Lambert-W MLE, surface/contours	5–6, 19
C3	ZIP method of moments	7, 19
C4	fitted ZIPs (MLE vs MM), RMSE, expected counts	8
C5	ZIP SIR posterior for (λ, π)	9, 19
C6	ZIP posterior predictive intervals	10
C7	`zeroinfl` ZIP (checks C2)	14
C8	`zeroinfl` ZINB: MLE of (μ, θ, π), implied moments	14, 19
C9	ZINB method of moments (factorial moments)	15, 19
C10	ZINB SIR posterior for (μ, θ, π)	16–17, 19
C11	ZIP vs ZINB posterior predictive intervals	18
C12	6 fits: probabilities, RMSE, expected counts, structural zeros	20–22
C13	log-lik, AIC, LR tests, marginal likelihoods / Bayes factor	23

Packages: `pscl`, `MASS`, `LambertW` (`lamW::lambertW0` also works). The full script is `zip_zinb.R`.

```r
## ---- C1: data ----
data <- read.csv("https://stats.idre.ucla.edu/stat/data/fish.csv")
head(data)
y = data$count[data$camper==1 & data$child==1]

par(mfrow=c(1,1))
plot(table(y),xlab="Number of fishes",ylab="Frequency",axes=FALSE,ylim=c(0,25))
axis(2);axis(1,at=min(y):max(y))
for (i in min(y):max(y))
  text(i,25,sum(y==i))
```

C2 – ZIP likelihood and closed-form MLE

(slides 5–6)

```r
## ---- C2: zip_mle ----
#install.packages("LambertW")
library("LambertW")
likelihood = function(y,lambda,pi){
  ifelse(y==0,pi+(1-pi)*exp(-lambda),(1-pi)*dpois(y,
      lambda))
}
n           = length(y)
n0          = sum(y==0)
ybar        = mean(y)
s2          = var(y)
zbar        = sum(y)/(n-n0)
lambda.mle = W(-zbar*exp(-zbar))+zbar
pi.mle      = 1-ybar/lambda.mle
c(n=n,n0=n0,ybar=ybar,s2=s2,zbar=zbar)
c(lambda.mle=lambda.mle,pi.mle=pi.mle)

N = 30
```

```r
lambdas = seq(3,6,length=N)
pis = seq(0.3,0.7,length=N)
like = matrix(0,N,N)
for (i in 1:N)
  for (j in 1:N)
     like[i,j] = prod(likelihood(y,lambdas[i],pis[j]))

par(mfrow=c(1,2))
persp(x=lambdas,y=pis,z=like,theta=40,phi=25,col="
      lightblue",shade=0.5,
       expand=0.7,xlab="lambda",ylab="pi",
       zlab="Likelihood",main="Likelihood surface")
contour(lambdas,pis,like,drawlabels=FALSE,xlab=
       expression(lambda),
         ylab=expression(pi),main="Likelihood contours")
points(lambda.mle,pi.mle,col=2,pch=16,cex=2)
```

C3 – ZIP method of moments


```
## ---- C3: zip_mm ----
lambda.mm  = (s2+ybar^2)/ybar-1
pi.mm      = (s2-ybar)/(s2+ybar^2-ybar)

tab = rbind(c(lambda.mm,pi.mm),c(lambda.mle,pi.mle))
colnames(tab) = c("lambda","pi")
rownames(tab) = c("Method of moments","Maximum likelihood")
round(tab,3)
```

C4 – Comparing fitted ZIPs


```r
## ---- C4: zip_fit ----
yy = sort(unique(y))
nn = length(yy)
freq = rep(0,nn)
for (i in 1:nn)
  freq[i] = sum(y==yy[i])/n
fit.mle  = likelihood(yy,lambda.mle,pi.mle)
fit.mm   = likelihood(yy,lambda.mm,pi.mm)
rmse.mle = sqrt(mean((freq-fit.mle)^2))
rmse.mm  = sqrt(mean((freq-fit.mm)^2))

par(mfrow=c(1,1))
plot(yy+0.1,freq,ylim=c(0,1),xlab="Counts",ylab="
     Probability",pch=16,axes=FALSE)
axis(2); axis(1,at=0:max(yy))
```

```r
points(yy-0.1,fit.mle,col=2,pch=16)
points(yy,fit.mm,col=4,pch=16)
legend("topright",legend=c("Observed count","Method of
      Moments","Maximum likelihood"),
       col=c(1,4,2),bty="n",pch=16)
legend("topleft",legend=c(paste("RMSE(MLE) = ",round(
      rmse.mle,3),sep=""),
                          paste("RMSE(MM)  = ",round(
      rmse.mm,3),sep="")))

tab = round(cbind(yy,n*freq,n*fit.mle,n*fit.mm))
colnames(tab) = c("counts","Frequency","MLE","MM")
tab
```

C5 – ZIP posterior via SIR


```r
## ---- C5: zip_sir ----
set.seed(2025)
M = 10000
lambda.d = runif(M,1,8)
pi.d     = runif(M,0.1,0.9)
w = rep(0,M)
for (i in 1:M)
 w[i] = prod(likelihood(y,lambda.d[i],pi.d[i]))

ind = sample(1:M,size=M,prob=w,replace=TRUE)
lambda.p = lambda.d[ind]
pi.p     = pi.d[ind]

par(mfrow=c(2,2))
plot(lambda.d,pi.d,xlab=expression(lambda),ylab=
      expression(pi),pch=".")
contour(lambdas,pis,like,col=2,add=TRUE,drawlabels=FALSE
      ,lwd=2)
title("Prior & likelihood")
plot(lambda.p,pi.p,xlim=range(lambda.d),ylim=range(pi.d)
      ,
     xlab=expression(lambda),ylab=expression(pi),pch="."
      )
contour(lambdas,pis,like,col=2,add=TRUE,drawlabels=FALSE
      ,lwd=2)
title("Posterior vs likelihood")
hist(lambda.p,prob=TRUE,main="",xlab=expression(lambda))
hist(pi.p,prob=TRUE,main="",xlab=expression(pi))

round(rbind(lambda=quantile(lambda.p,c(0.025,0.5,0.975))
      ,
            pi    =quantile(pi.p,c(0.025,0.5,0.975))),3)
```

C6 – ZIP posterior predictive


```
## ---- C6: zip_predictive ----
pred = matrix(0,M,nn)
for (i in 1:M)
  for (j in 1:nn)
    pred[i,j] = likelihood(yy[j],lambda.p[i],pi.p[i])

qpred = t(apply(pred,2,quantile,c(0.025,0.5,0.975)))
par(mfrow=c(1,1))
plot(yy,qpred[,2],ylim=range(qpred,freq),pch=16,xlab="Counts",
     ylab="Predictive probability",col=6,axes=FALSE)
axis(2); axis(1,at=0:max(yy))
segments(yy,qpred[,1],yy,qpred[,3],col=6)
points(yy,freq,pch=16)
legend("topright",legend=c("Observed count","95% predictive interval"),
       col=c(1,6),bty="n",pch=16)
```

C7–C8 – pscl: ZIP check and ZINB MLE


```
## ---- C7: zip_pscl ----
library(pscl)

# Zero-Inflated Poisson (ZIP): intercept-only count and zero parts
fit_zip <- zeroinfl(y~1|1,dist="poisson")
summary(fit_zip)
exp(fit_zip$coefficients$count)          # lambda  (should match lambda.mle)
1/(1+exp(-fit_zip$coefficients$zero))    # pi      (should match pi.mle)
```

```
## ---- C8: zinb_mle ----
# Zero-Inflated Negative Binomial (ZINB): maximum likelihood via pscl
fit_zinb <- zeroinfl(y~1|1,dist = "negbin")
summary(fit_zinb)
mu    = exp(fit_zinb$coefficients$count)          # NB mean (non-structural part)
pi_nb = 1/(1+exp(-fit_zinb$coefficients$zero))    # zero-inflation probability
theta = fit_zinb$theta                            # NB size (dispersion)
p     = theta/(theta+mu)                          # NB "success" probability
c(mu=mu,theta=theta,p=p,pi=pi_nb)

# implied moments of the ZINB
c(mean=(1-pi_nb)*mu, var=(1-pi_nb)*mu*(1+mu/theta+pi_nb*mu), ybar=ybar, s2=s2)
```

C9 – ZINB method of moments


```r
## ---- C9: zinb_mm ----
# ZINB pmf (same role as likelihood() for the ZIP)
likelihood.nb = function(y,mu,theta,pi){
  ifelse(y==0,pi+(1-pi)*dnbinom(0,size=theta,mu=mu),
              (1-pi)*dnbinom(y,size=theta,mu=mu))
}
# Factorial moments: E[Y(Y-1)...(Y-k+1)] = (1-pi) mu^k
      prod_{j<k} (1+j/theta)
f1 = mean(y)
f2 = mean(y*(y-1))
f3 = mean(y*(y-1)*(y-2))
a  = f2/f1                    # = mu(1+1/theta)
b  = f3/f2                    # = mu(1+2/theta)
mu.mm    = 2*a-b
theta.mm = mu.mm/(b-a)
pi.mm.nb = 1-f1/mu.mm
c(f1=f1,f2=f2,f3=f3)
c(mu.mm=mu.mm,theta.mm=theta.mm,pi.mm=pi.mm.nb)
```

```r

# Remark: matching mean, variance and P(Y=0) instead of
      the 3rd moment
p0 = mean(y==0)
g  = function(par){ mu=exp(par[1]); th=exp(par[2]); pi=
      plogis(par[3])
  c((1-pi)*mu-ybar, (1-pi)*mu*(1+mu/th+pi*mu)-s2,
    pi+(1-pi)*(th/(th+mu))^th-p0) }
o  = optim(c(1,0,0),function(p) sum(g(p)^2),control=list
      (reltol=1e-14,maxit=1e4))
c(mu=exp(o$par[1]),theta=exp(o$par[2]),pi=plogis(o$par
      [3]))
```

C10 – ZINB posterior via SIR

(slides 16–17)

```r
## ---- C10: zinb_sir ----
# Priors as vague as for the ZIP:
#   mu ~ U(1,8) (same as lambda), pi ~ U(0.1,0.9) (same as ZIP),
#   psi = 1/(1+theta) ~ U(0,1)   (psi -> 0: Poisson/ZIP; psi -> 1: extreme overdispersion)
set.seed(2025)
M.nb    = 100000
mu.d    = runif(M.nb,1,8)
pinb.d  = runif(M.nb,0.1,0.9)
psi.d   = runif(M.nb,0,1)
theta.d = (1-psi.d)/psi.d

ytab = table(y); yv = as.numeric(names(ytab)); ny = as.vector(ytab)
lw.nb = rep(0,M.nb)
for (i in 1:M.nb)
  lw.nb[i] = sum(ny*log(likelihood.nb(yv,mu.d[i],theta.d[i],pinb.d[i])))
w.nb = exp(lw.nb-max(lw.nb))
ess  = 1/sum((w.nb/sum(w.nb))^2)          # effective sample size
ess
```

```r
ind.nb  = sample(1:M.nb,size=10000,prob=w.nb,replace=TRUE)
mu.p    = mu.d[ind.nb]
theta.p = theta.d[ind.nb]
psi.p   = psi.d[ind.nb]
pinb.p  = pinb.d[ind.nb]

par(mfrow=c(2,3))
plot(mu.p,pinb.p,xlim=c(1,8),ylim=c(0.1,0.9),pch=".",xlab=expression(mu),ylab=expression(pi))
plot(mu.p,psi.p,xlim=c(1,8),ylim=c(0,1),pch=".",xlab=expression(mu),ylab=expression(psi==1/(1+theta)))
plot(psi.p,pinb.p,xlim=c(0,1),ylim=c(0.1,0.9),pch=".",xlab=expression(psi==1/(1+theta)),ylab=expression(pi))
hist(mu.p,prob=TRUE,main="",xlab=expression(mu))
hist(log(theta.p),prob=TRUE,main="",xlab=expression(log(theta)))
hist(pinb.p,prob=TRUE,main="",xlab=expression(pi))

round(rbind(mu   =quantile(mu.p,   c(0.025,0.5,0.975)),
            theta=quantile(theta.p,c(0.025,0.5,0.975)),
            pi   =quantile(pinb.p, c(0.025,0.5,0.975))),3)
```

C11 – ZIP vs ZINB posterior predictive


```r
## ---- C11: zinb_predictive ----
pred.nb = matrix(0,10000,nn)
for (j in 1:nn)
  pred.nb[,j] = likelihood.nb(rep(yy[j],10000),mu.p,
      theta.p,pinb.p)

qpred.nb = t(apply(pred.nb,2,quantile,c(0.025,0.5,0.975)
      ))
par(mfrow=c(1,1))
plot(yy-0.15,qpred[,2],ylim=c(0,0.75),xlim=c(-0.3,max(yy
      )+0.3),pch=16,col=6,
     xlab="Counts",ylab="Predictive probability",axes=
      FALSE)
axis(2); axis(1,at=0:max(yy))
segments(yy-0.15,qpred[,1],yy-0.15,qpred[,3],col=6)
points(yy+0.15,qpred.nb[,2],pch=15,col=3)
segments(yy+0.15,qpred.nb[,1],yy+0.15,qpred.nb[,3],col
      =3)
points(yy,freq,pch=16)
legend("topright",legend=c("Observed","ZIP 95%
      predictive","ZINB 95% predictive"),
       col=c(1,6,3),pch=c(16,16,15),bty="n")

# does the 95% interval cover the observed frequency?
rbind(count=yy, ZIP=(freq>=qpred[,1] & freq<=qpred[,3]),
      ZINB=(freq>=qpred.nb[,1] & freq<=qpred.nb[,3]))
```

C12 – Six fits compared (1/2)


```
## ---- C12: fits_compared ----
# Fitted P(Y=k) at the observed support yy, for the 2 models x 3 methods
probs = cbind(
  ZIP.MLE   = likelihood(yy,lambda.mle,pi.mle),
  ZIP.MM    = likelihood(yy,lambda.mm,pi.mm),
  ZIP.Bayes = colMeans(pred),
  ZINB.MLE  = likelihood.nb(yy,mu,theta,pi_nb),
  ZINB.MM   = likelihood.nb(yy,mu.mm,theta.mm,pi.mm.nb),
  ZINB.Bayes= colMeans(pred.nb))
rmse = sqrt(colMeans((probs-freq)^2))
round(matrix(rmse,2,3,byrow=TRUE,dimnames=list(c("ZIP","ZINB"),c("MLE","MM","Bayes"))),3)
round(cbind(count=yy,Observed=n*freq,n*probs),1)
```

```
par(mfrow=c(1,2))
for (m in 1:2){
  plot(yy,freq,type="h",lwd=6,col=grey(0.8),ylim=c(0,0.75),axes=FALSE,
       xlab="Counts",ylab="Probability",main=c("ZIP","ZINB")[m])
  axis(2); axis(1,at=0:max(yy))
  for (k in 1:3) points(yy+(k-2)*0.25,probs[,3*(m-1)+k],pch=c(16,17,15)[k],col=c(2,4,3)[k])
  legend("topright",legend=c("MLE","MM","Bayes"),col=c(2,4,3),pch=c(16,17,15),bty="n")
}

# share of the zeros that are structural (non-fishing groups): pi / P(Y=0)
P0 = probs[1,]
share = c(pi.mle/P0[1], pi.mm/P0[2], mean(pi.p/likelihood(rep(0,M),lambda.p,pi.p)),
          pi_nb/P0[4], pi.mm.nb/P0[5], mean(pinb.p/likelihood.nb(rep(0,10000),mu.p,theta.p,pinb.p)))
round(matrix(share,2,3,byrow=TRUE,dimnames=list(c("ZIP","ZINB"),c("MLE","MM","Bayes"))),3)
```

C12 – Six fits compared (2/2)

(original script's zip_prob check)

```r

# the ZIP probabilities "by hand" from pscl's predictions (original script)
my = max(y)
lambda <- predict(fit_zip, type = "count")
pi_zip <- predict(fit_zip, type = "zero")
zip_prob <- function(y, pi, lambda) {
  pois_part <- (1 - pi) * dpois(y, lambda = lambda)
  ifelse(y == 0, pi + pois_part, pois_part)
}
fit1 = sapply(0:my, function(y) zip_prob(y, pi_zip[1], lambda[1]))
avg_probs = colMeans(predict(fit_zip, type="prob", at=0:my))
max(abs(fit1-avg_probs))     # ~ 0
```

C13 – Model comparison


```
## ---- C13: model_comparison ----
library(MASS)
fit_pois <- glm(y~1, family=poisson)
fit_nb   <- glm.nb(y~1)
mods = list(Poisson=fit_pois, NB=fit_nb, ZIP=fit_zip,
      ZINB=fit_zinb)
cmp  = t(sapply(mods, function(m)
          c(logLik=as.numeric(logLik(m)), df=attr(logLik
      (m),"df"), AIC=AIC(m))))
round(cmp,2)

# LR test ZIP vs ZINB (theta -> infinity is on the
      boundary: halve the p-value)
LR = 2*(logLik(fit_zinb)-logLik(fit_zip))
c(LR=LR, p=0.5*pchisq(LR,1,lower.tail=FALSE))
# LR test NB vs ZINB (pi = 0 is on the boundary: halve
      the p-value)
LR2 = 2*(logLik(fit_zinb)-logLik(fit_nb))
c(LR=LR2, p=0.5*pchisq(LR2,1,lower.tail=FALSE))
```

```
# Bayes: marginal likelihood p(y) = E_prior[ L ]
      estimated by the SIR prior draws
lw.zip  = log(w)                                  # ZIP:
       w = likelihood at prior draws
logml.zip  = max(lw.zip) + log(mean(exp(lw.zip-max(lw.
      zip))))
logml.zinb = max(lw.nb)  + log(mean(exp(lw.nb-max(lw.nb)
      )))
c(logml.ZIP=logml.zip, logml.ZINB=logml.zinb, log10.BF=(
      logml.zinb-logml.zip)/log(10))
```