GLMS AND MODELLING

AFTER WEEK 5: EXPLANATION, CONSOLIDATION AND PREPARATION FOR POWER

Gavin Simpson
gavin@anivet.au.dk
Aarhus University

Wednesday, 21 October 2026

HOW TO USE THIS RECORDING

This is the follow-up to the simulation workshop. Keep your group's script and the after-class worksheet beside you.

Chapter	Focus	Recording time
1	The model behind the code	About 10 minutes
2	Four data generating processes	About 15 minutes
3	Fitting and repeated experiments	About 10 minutes

Pause to answer the questions. The worksheet adds about 25 minutes of practice.

THE MODEL BEHIND THE CODE

Chapter 1: about 10 minutes

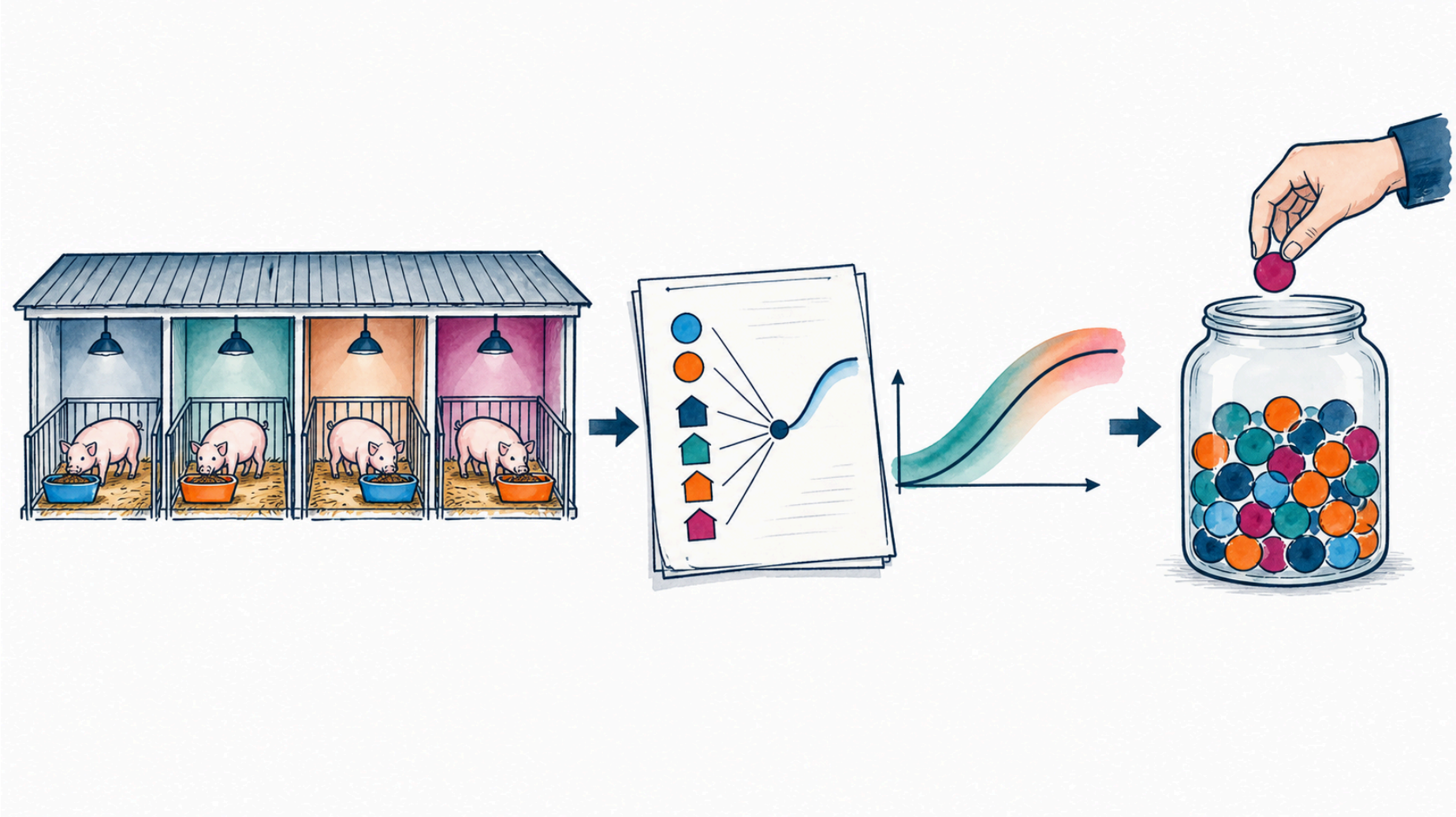
THE CLASS MODEL

We kept the treatment allocation and numerical assumptions fixed, then drew two sets of responses.

The GLM described the expected response for each pen and the variation remaining among comparable pens.

The fitted coefficients changed because each dataset contained different information about the same generating model.

THE GENERATIVE SEQUENCE

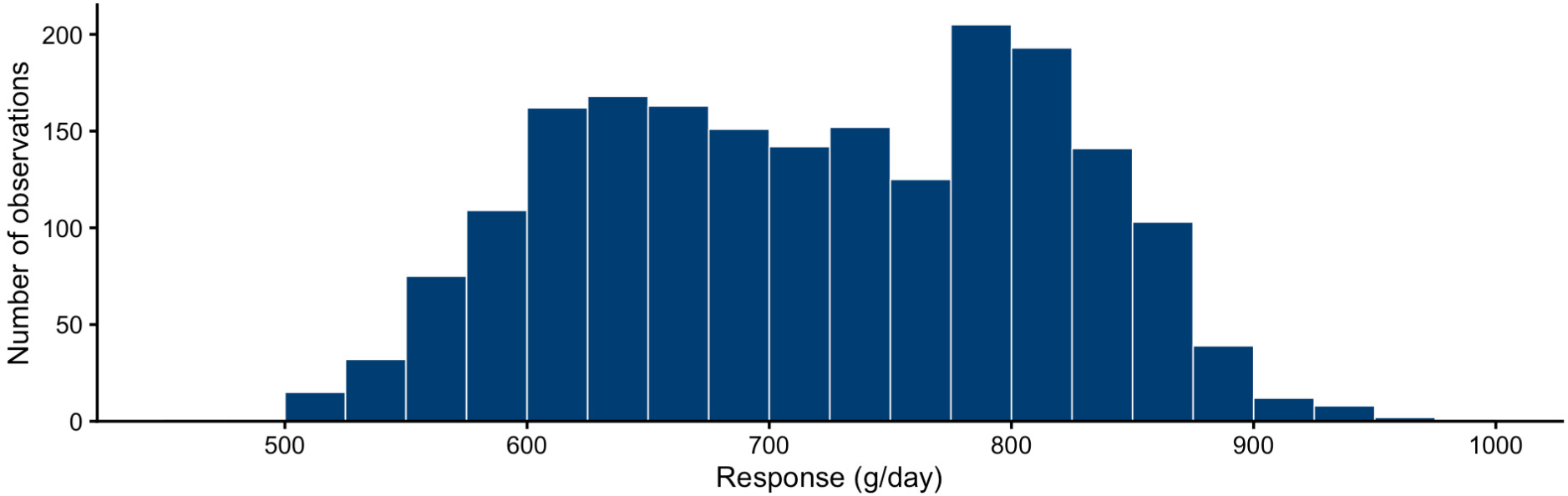


A CONDITIONAL MODEL

For a pen with a particular diet and room, write:

$$E(Y_i \mid \text{diet}_i, \text{room}_i) = \mu_i$$

The family describes responses **given those predictors**. Pooling pens with different means can produce a distribution with two peaks.



COMPONENTS OF A GLM

$$\eta_i = \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}, \qquad g(\mu_i) = \eta_i, \qquad Y_i \mid \boldsymbol{x}_i \sim F_i$$

The **linear predictor** combines treatment and room terms. The **link** connects the expected response to that predictor. The **family** determines response support and a mean–variance relationship.

For simulation, we must also specify any remaining scale or dispersion parameters.

- η — "eta", the values on the *linear predictor*
- μ — "mu", the expected value or mean of the response on the response scale
- *g — the **link** function

LINK AND INVERSE LINK

Family used in class	Link, from mean to score	Inverse link, from score to mean
Gaussian	$\eta = \mu$	$\mu = \eta$
Binomial	$\eta = \log\{p/(1-p)\}$	$p = 1/(1+\exp(-\eta))$
Poisson and Gamma	$\eta = \log(\mu)$	$\mu = \exp(\eta)$

R FUNCTIONS FOR THE LOGIT AND INVERSE LOGIT

R uses

- `qlogis(p)` for the logit, i.e. $\eta = \log\{p/(1-p)\}$, and
- `plogis(eta)` for its inverse, i.e. $p = 1/(1+\exp(-\eta))$

These are the links chosen for our scenarios, but other links are available.

Notice that we never transform the data themselves!

ROOM EFFECTS AND THE INTERCEPT

Our generator uses:

$$\eta_i = b_0 + b_{\text{diet}} I(\text{high fibre}) + \gamma_{\text{room}(i)}$$

Room 1 has $\gamma_1 = 0$, so b_0 describes its control mean on the link scale.

Room 1 is also the first factor level in our fitted model. With treatment contrasts, the fitted intercept estimates b_0, and each room coefficient compares that room with Room 1.

FIXED ROOM MEANS AND INDEPENDENCE

In the workshop, each room effect is a fixed number. We draw responses independently after setting those means.

Shared room conditions could create additional dependence, and diet effects could differ among rooms. Either would require a richer generator and an appropriate analysis.

Simulation forces us to state which variation we include and which questions remain open.

PAUSE: MODEL ANATOMY

In your script, locate:

1. the allocation
2. the numerical model parameters
3. the linear predictor and inverse link
4. the random response generator

Explain which would change if you changed only the response seed. Then explain what would change if you increased the treatment effect.

FOUR DATA GENERATING PROCESSES

Chapter 2: about 15 minutes

GAUSSIAN PEN MEANS

$$Y_i \mid \boldsymbol{x}_i \sim N(\mu_i, \sigma^2), \qquad \mu_i = \eta_i$$

In our scenario, $\sigma = 60$ g/day and the diet changes expected daily gain by +45 g/day.

The Gaussian model assigns probability across the real line. Negative daily gain can make sense, but extreme tails may still be biologically implausible.

The assumed SD describes variation among pen means after accounting for diet and room.

BINARY PENS AND BASELINE PROBABILITY

$$Y_i \sim \text{Bernoulli}(p_i), \qquad \text{Var}(Y_i \mid \boldsymbol{x}_i) = p_i(1 - p_i)$$

The response records whether **any** pig in a pen required treatment. There is one binary trial per pen, even though each pen contains six pigs.

A fixed odds ratio can imply different absolute probability changes at different baseline probabilities.

ODDS AND PROBABILITIES

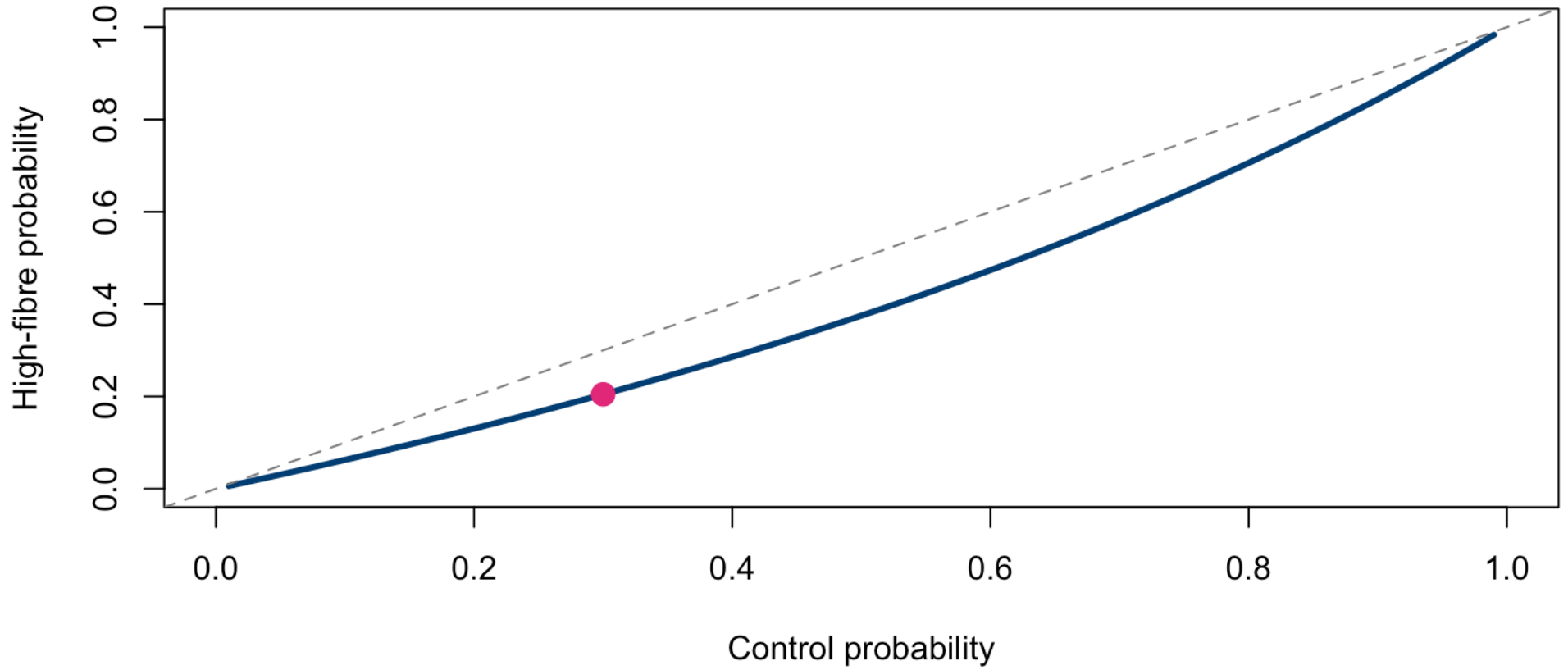
In Room 1, control probability is 0.30, giving odds $0.30/0.70 = 0.4286$.

Multiplying those odds by 0.60 gives high-fibre odds of 0.2571. Converting back gives:

$$p_{\text{high}} = \frac{0.2571}{1 + 0.2571} \approx 0.2045$$

This is about a **9.55 percentage-point decrease**, not a probability of 0.18 (0.3 / 0.6). The odds ratio and probability ratio are different quantities.

A CONSTANT ODDS RATIO



The curve uses odds ratio 0.60 throughout. The dashed line marks equal probabilities.

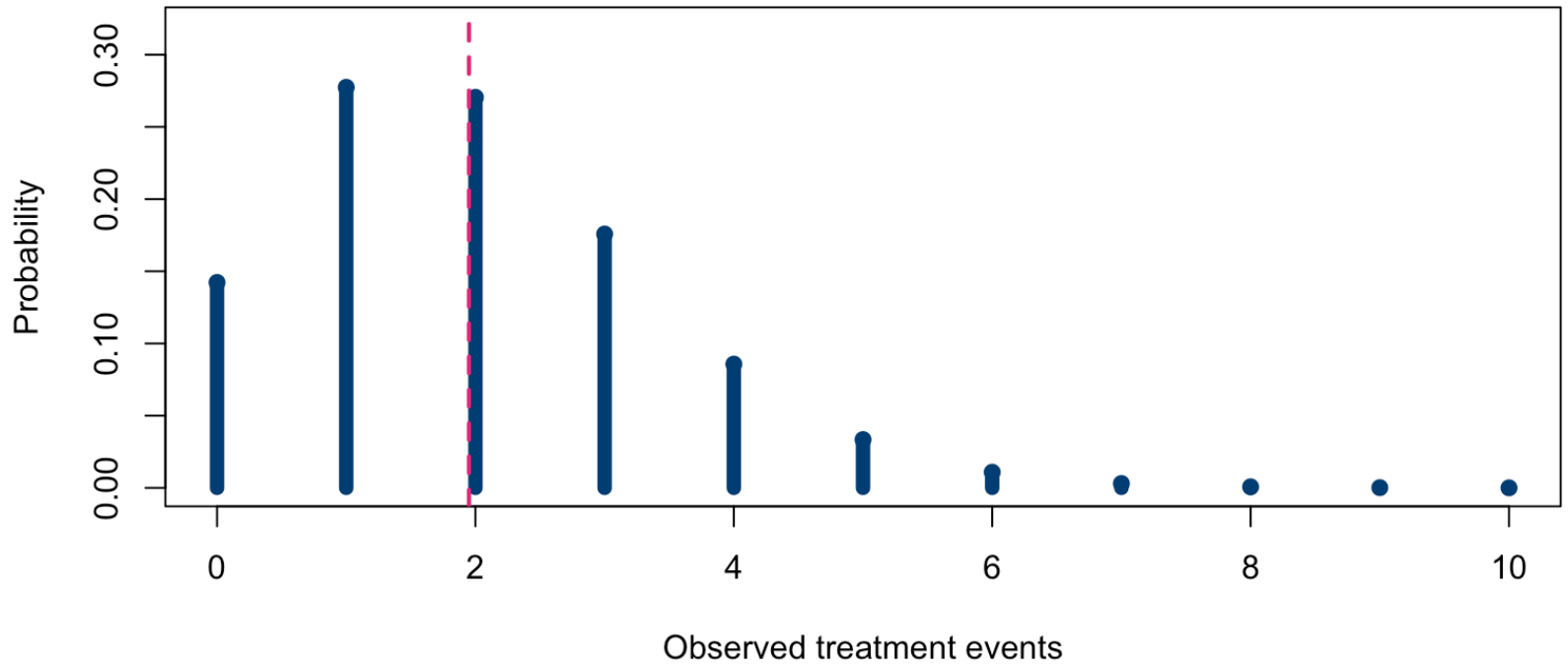
POISSON COUNTS AND EXPOSURE

$$Y_i \sim \text{Poisson}(\mu_i), \qquad \text{Var}(Y_i \mid \boldsymbol{x}_i) = \mu_i$$

Our expected-count ratio is 0.65. It is also a rate ratio because observation time and pen size are equal.

If exposure differed, a count model could use $\log(\text{exposure}_i)$ as an offset when the scientific assumptions justify a rate model.

FRACTIONAL MEANS, INTEGER OBSERVATIONS



The dashed line marks the mean, 1.95. The response places probability only on integer counts.

EXTRA VARIATION IN COUNTS

Pens with the same predicted mean may vary more than a Poisson model allows.

Possible causes include unmodelled heterogeneity, dependent events or a different biological mechanism. A negative-binomial generator is one possible sensitivity scenario.

Choosing a more variable generator also requires checking the planned analysis. A Poisson fit may understate uncertainty under extra variation.

GAMMA MEANS AND SPREAD

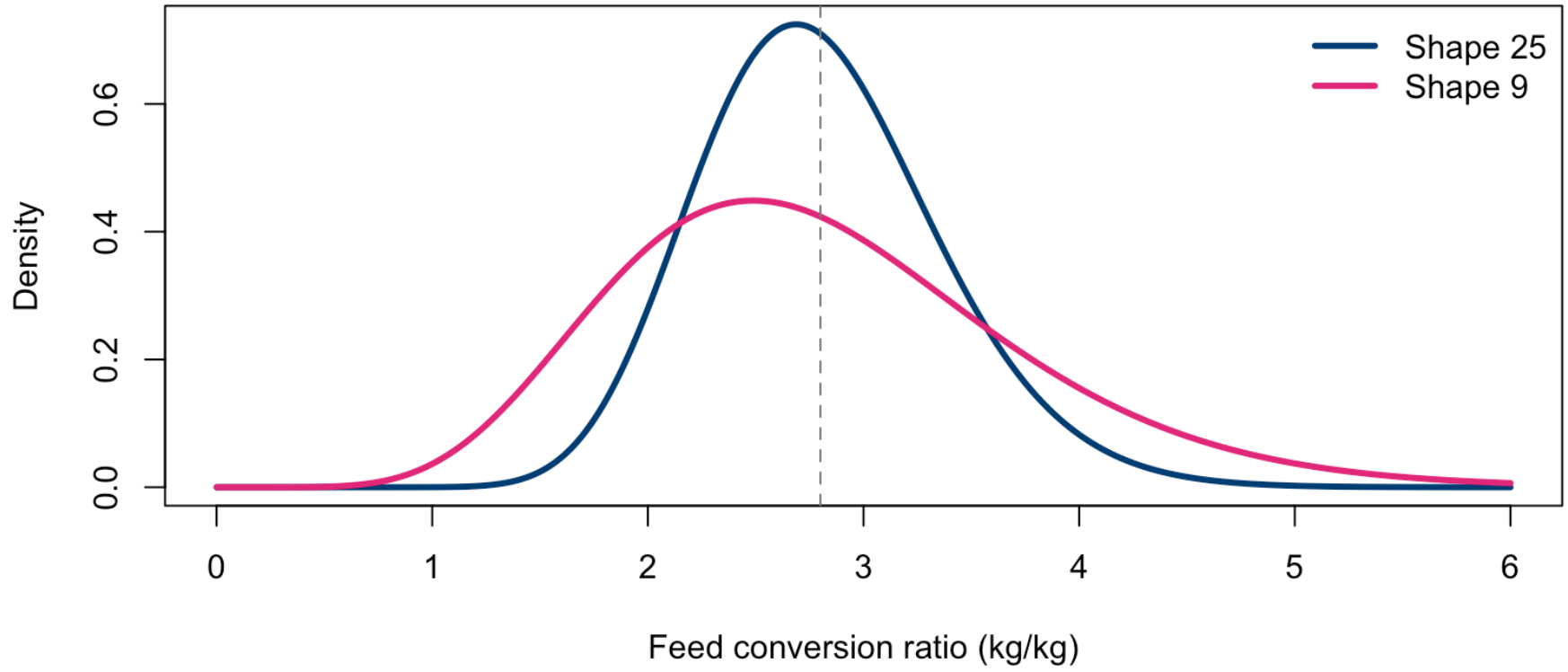
R's `rgamma()` uses a shape k and scale θ parameterisation for a Gamma draw:

$$E(Y) = k\theta, \qquad \mathrm{Var}(Y) = k\theta^2$$

Setting $\theta = \mu/k$ gives mean μ, variance μ^2/k and coefficient of variation $1/\sqrt{k}$.

Our shape is $k = 25$, so the coefficient of variation is 0.20: $1/\sqrt{25}$. The log-link treatment coefficient gives a ratio of expected feed-conversion-ratio values.

SAME GAMMA MEAN, DIFFERENT SPREAD



Both means are 2.80. Reducing shape increases relative variation.

CONDITIONAL VARIATION SUMMARY

Response	Conditional variance	Extra simulation parameter
Gaussian	σ^2	SD $\sigma = 60$
Binary	$p(1-p)$	None beyond p
Poisson	μ	None beyond μ
Gamma	μ^2/k	Shape $k = 25$

For the last three models, changing the mean changes the variance. Increasing sample size and changing the response variation are different design questions.

PAUSE: ASSUMPTION AUDIT

Which assumption in your model most needs external evidence?

Explain how you could change it while preserving valid response values. State what else must change in the code to keep the intended mean.

Use the sensitivity task in the after-class worksheet.



FITTING AND REPEATED EXPERIMENTS

Chapter 3: about 10 minutes

GENERATING AND ESTIMATING

During simulation, we choose the parameters and draw possible responses.

During fitting, we give R the responses and model structure, and it estimates the unknown coefficients. It does not receive our generating values.

Likelihood compares how well different parameter values account for the observed data. `glm()` carries out the numerical fitting.

THE MODEL CALL

```
fit <- glm(response ~ diet + room,
           family = poisson(link = "log"),
           data = first)
coef(fit)["dietHigh fibre"]
```

```
dietHigh fibre
    -0.4700036
```

The formula allows a diet contrast and room differences. The family call states both the data generating process and link.

The fitted diet coefficient estimates a log expected-count ratio. Its exponential estimates that ratio under this model.

PREDICTIONS ON THE RESPONSE SCALE

```
new_pens <- data.frame(
  diet = factor(c("Control", "High fibre"),
                levels = levels(first$diet)),
  room = factor(rep("Room 1", 2),
                levels = levels(first$room))
)
predict(fit, newdata = new_pens, type = "response")
```

```
       1        2
2.769231 1.730769
```

These are estimated **mean event counts** for Control and High fibre pens in Room 1, respectively. They do not include the random variation of a new observation.

COEFFICIENTS AND UNCERTAINTY

A coefficient estimate summarises one dataset. Its standard error describes estimated sampling uncertainty under the model.ß

A confidence interval adds that uncertainty to the report. For a log-link coefficient, exponentiating interval endpoints gives a ratio interval.

With small or sparse datasets, familiar large-sample intervals can behave poorly.

SPARSE BINARY DATA

Some simulated rooms may contain only zeros. Diet and room patterns can create separation or near-separation, leaving ordinary logistic estimates poorly determined.

Inspect warnings, fitted probabilities and coefficient standard errors. A convergence flag alone does not establish reliable estimation.

Retain these outcomes. Week 6 needs a predeclared strategy for unreliable fits and must report how often they occur.

WHAT POWER ASKS

For a stated nonzero effect, model, design and analysis, power is the probability that the chosen test rejects its null hypothesis.

Simulation estimates this by drawing many possible experiments and applying the same analysis and rejection rule each time.

A single estimate close to the generating effect, or a single small p-value, cannot establish that probability.

A WEEK 6 SIMULATION RUN

At each candidate sample size:

1. Construct the proposed design and specify the generating parameters.
2. Draw a response dataset.
3. Fit the planned model and apply a predeclared rejection rule.
4. Repeat and summarise rejection and analysis-failure rates.

We will first condition on a balanced allocation. Re-randomising within each repetition answers a related question averaged over treatment assignments.

MONTE CARLO UNCERTAINTY

Suppose 800 of 1,000 simulated experiments reject the null under an assumed effect. The estimated power is 0.80.

Its approximate Monte Carlo standard error is:

$$\sqrt{\frac{0.80(1-0.80)}{1000}} \approx 0.013$$

More repetitions reduce simulation noise. They do not make unrealistic biological assumptions more credible.

SAMPLE SIZE REFERS TO INDEPENDENT PENS

Pens per room	Rooms	Total pens	Pens per diet
4	4	16	8
8	4	32	16
12	4	48	24
16	4	64	32

This candidate grid increases independent pens while retaining four fixed room effects and balanced diets. Feasibility requires separate consideration.

SENSITIVITY AND FALSE POSITIVES

A useful power curve should state its effect, baseline, variation, allocation and analysis assumptions.

Compare it with at least one plausible alternative scenario. Also set the treatment effect to its null value and check the rejection rate.

An inflated false-positive rate can make an apparent gain in power misleading.

YOUR WEEK 6 PREPARATION

Complete the after-class worksheet and bring:

- a corrected and reproducible group script
- one explained effect on the response scale
- one assumption you would vary
- a draft detection rule and a question about it

You will learn to repeat the full simulation and analysis in Week 6. You do not need to write the power loop beforehand.

SOURCES AND FURTHER REFERENCE

- McCullagh and Nelder (1989). *Generalized Linear Models*, 2nd ed.
- R Core Team. Families, GLMs, predictions, and Gamma distributions.
- Colegrave and Ruxton (2020). *Power Analysis: An Introduction for the Life Sciences*. Oxford University Press.