SIMULATION WORKSHOP

WEEK 5: ONE DESIGN, FOUR DATA GENERATING PROCESSES

Gavin Simpson
gavin@anivet.au.dk
Aarhus University

Wednesday, 21 October 2026

WORKSHOP ROUTE

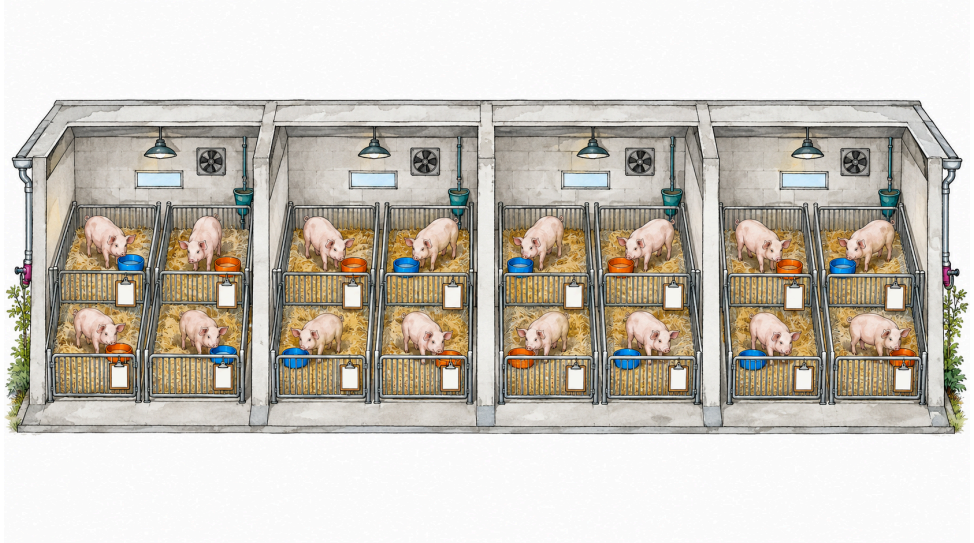
Time	Task
13:20–13:40	Design and assumption record
13:40–13:55	Break
13:55–14:25	First simulated dataset
14:25–14:55	Checks, plot and second response draw
14:55–15:10	Break
15:10–15:30	Supplied GLM and interpretation
15:30–15:45	Group audit and reports
15:45–16:00	Synthesis, after-class work and exit

THE SHARED EXPERIMENT

Four rooms, four pens per room and six pigs per pen. Every pen has its own shared feeder.

Within each room, two pens receive control and two receive high-fibre diet.

The observation period is the same for all pens. We simulate one response per pen.



FOUR GROUPS OF THREE

Driver: types and runs code.

Model translator: explains the score, mean and response.

Design auditor: checks units, allocation and assumptions.

Rotate at 14:25 and 15:10 so every student tries every role.



YOUR RESPONSE AND RECIPE

Group	Response	Family and link
1	Mean daily gain, g/day	Gaussian, identity
2	Any pig treated, 0/1	Binomial, logit
3	Treatment events, count	Poisson, log
4	Feed conversion ratio, kg/kg	Gamma, log

The starter supplies parameter values and the random generator. Complete the design code, the model score and the expected response.

PACKAGE CHOICE AND FILES

The scripts use **dplyr** and **tibble** for table operations, and **ggplot2** for the plot. R's supplied **stats** package provides the response generators and GLMs.

Open your assigned `group-N-student.R`. Run sections with Ctrl/Cmd+Enter and save your edited script.

`NA` marks a value to complete. The checks deliberately stop incomplete code and tell you which part needs attention.

TASK 1: RANDOMISATION

Fill the two marked operations in your script.

Incomplete student template (pseudocode):

```
labels <- rep(c("Control", "High fibre"), each = 2)
set.seed(505)

design <- pens |>
  group_by(____) |>
  mutate(diet = ____(labels)) |>
  ungroup()
```

Explain why the shuffle happens separately in each room and why labels must be sampled without replacement.

WORKED ALLOCATION: ONE COMMON DESIGN

After attempting the template, compare this completed allocation. `pens` has one row per pen; `design` shuffles diets within rooms.

```
library("dplyr")
library("tibble")
pens <- tibble(pen_id = sprintf("P%02d", 1:16),
  room = factor(rep(paste0("Room ", 1:4), each = 4)))
labels <- rep(c("Control", "High fibre"), each = 2)
set.seed(505)
design <- pens |> group_by(room) |>
  mutate(diet = sample(labels)) |> ungroup()
```

CHECKPOINT 1

This executed check uses the completed `design` above.

```
design |> count(room, diet)
```

```
# A tibble: 8 × 3
  room   diet           n
  <fct>  <chr>      <int>
1 Room 1 Control        2
2 Room 1 High fibre     2
3 Room 2 Control        2
4 Room 2 High fibre     2
5 Room 3 Control        2
6 Room 3 High fibre     2
7 Room 4 Control        2
8 Room 4 High fibre     2
```

```
nrow(design)
```

```
[1] 16
```

```
n_distinct(design$pen_id)
```

```
[1] 16
```

Expected: eight room-by-diet combinations with two pens each, 16 rows and 16 unique IDs.

Show these checks to the instructor before adding a response.

TASK 2: THE ASSUMPTION RECORD

Record the response unit and its possible values, then the supplied values for:

- the control expectation in **Room 1**
- the diet effect and its scale
- room differences on the model score scale
- remaining variation and any equal-exposure assumption

The zero room effect belongs to Room 1. These are scenario assumptions that we can question.

FIRST SIMULATION

13:55–14:25

TASK 3: SCORE, MEAN AND RESPONSE

Complete the two model lines in your script:

Incomplete student template (pseudocode):

```
eta = beta_0 + beta_diet * (diet == "High fibre") +
  room_effect
mu = YOUR_INVERSE_LINK
```

Choose `eta`, `plogis(eta)` or `exp(eta)` from your recipe. The supplied generator draws a response for each pen using seed 1501, 1502, 1503 or 1504.

WORKED GAUSSIAN DRAW

This is Group 1's illustration: Room 1 control mean 720 g/day, diet difference 45 g/day, residual SD 60. Other groups use their own family and inverse link.

```
set.seed(1501)
first <- design |> mutate(
  room_effect = c(0, 25, 45, 35)[as.integer(room)],
  eta = 720 + 45 * (diet == "High fibre") + room_effect,
  mu = eta, response = rnorm(n(), mean = mu, sd = 60))
head(first[c("pen_id", "diet", "eta", "mu", "response")], 2)
```

```
# A tibble: 2 × 5
  pen_id diet         eta    mu response
  <chr>  <chr>      <dbl> <dbl>    <dbl>
1 P01    Control      720   720     698.
2 P02    High fibre   765   765     847.
```

CHECKPOINT 2

Inspect `pen_id`, `diet`, `room_effect`, `eta`, `mu` and `response`.

Group	Required response check
1	Finite continuous observations
2	Every observation is 0 or 1
3	Every observation is a non-negative integer
4	Every observation is positive

Check the mean separately. For example, a Poisson mean of 1.95 is valid even though a count of 1.95 is impossible.

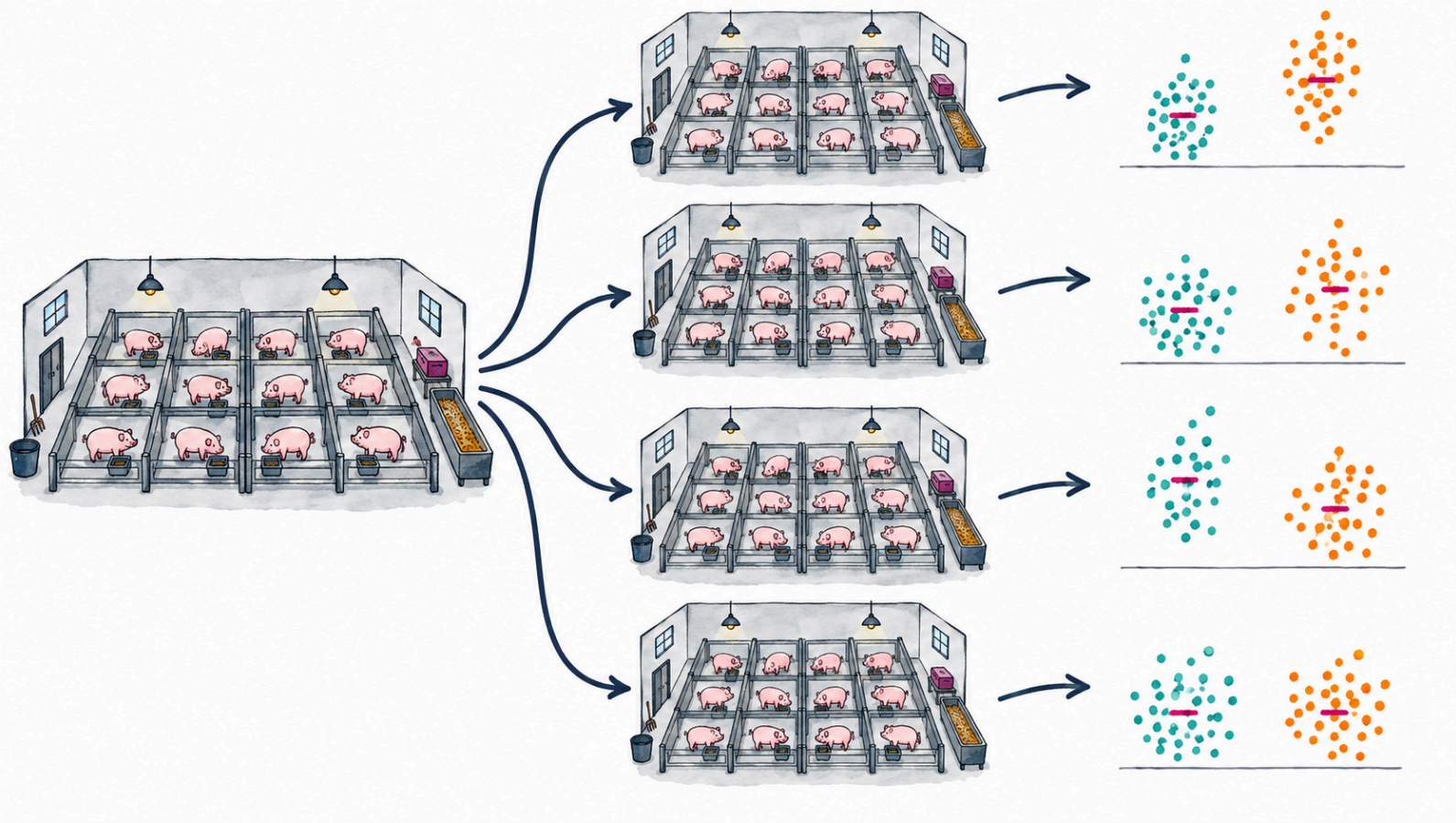
TASK 4: THE SECOND POSSIBLE DATASET

Rotate roles. Predict what will change before you run a new response seed.

Keep the allocation and all model parameters fixed. Change only the response seed to 2501, 2502, 2503 or 2504.

Compare the two response columns and the supplied plots. Keep both datasets, even if they suggest different treatment patterns.

ONE DESIGN, SEVERAL POSSIBLE OUTCOMES



ANALYSIS AND AUDIT

15:10–15:45

TASK 5: THE SUPPLIED GLM

This executed illustration continues the Gaussian `first` dataset above.

```
model_family <- gaussian("identity")
fit <- glm(response ~ diet + room,
           family = model_family, data = first)
coef(summary(fit))[1:2, c("Estimate", "Std. Error")]
```

```
               Estimate Std. Error
(Intercept)    747.160427   34.39116
dietHigh fibre   1.326009   30.76039
```

```
fit$converged
```

```
[1] TRUE
```

The formula compares diets while allowing room differences. `model_family` is already set for your group.

Record the generating coefficient, fitted coefficient, and any warnings. Compare the first and second datasets.

INTERPRETATION CARD

Group	Generating value	What to report
1	+45	Expected gain difference in g/day
2	0.60	Odds ratio, `exp(beta_diet)`
3	0.65	Expected-count ratio, `exp(beta_diet)`
4	0.92	Expected FCR ratio, `exp(beta_diet)`

For Groups 2–4, the generating coefficient is the logarithm of the value shown. The fitted coefficient usually differs because the observations vary.

WHEN A FIT IS UNSTABLE

For a binary response, 16 pens can leave very few observed events in some rooms or diets.

A warning, extreme coefficient or huge standard error belongs in your record. Ask the instructor before interpreting that estimate.

Keep the dataset. Explain its valid generator and the difficulty in estimating its treatment effect.

TASK 6: ANOTHER GROUP'S ASSUMPTIONS

At 15:30, pass your record clockwise. Check:

1. Do the unit and response support agree with the code?
2. Does the mean stay in its valid range?
3. Does the effect interpretation match the link?
4. Which assumption should we change in a sensitivity analysis?

Each group reports one defended choice and one assumption to investigate.

AFTER-CLASS WORK

Watch the GLM consolidation recording **after today's activity**. It explains the modelling choices you have now used.

Complete the short reflection and R checks in the after-class activity. Bring your corrected script, one model assumption to vary, and a proposal for what "detecting the effect" should mean to Week 6.

Allow about 60 minutes including the recording.

INDIVIDUAL EXIT QUESTION

A simulated experiment gives a clear difference between diets.

Why does that single result fail to establish adequate power? State what you would keep fixed and what you would repeat.

Also name one value your group's response can take and one it cannot take.

SOURCES

- R Core Team. GLM documentation and distribution families.
- Wickham et al. (2019). Welcome to the Tidyverse.
- Colegrave and Ruxton (2020). *Power Analysis: An Introduction for the Life Sciences*. Oxford University Press. Instructor background.