Design Matrices and Experimental Data

Bayesian Modeling for Socio-Environmental Data

N. Thompson Hobbs and Christian Che-Castaldo

April 16, 2019

Analyzing experimental data: Why Bayes?

KEY TO STATISTICAL METHODS

	Design or Purpose	Measurement Variables	Ranked Variables	Attributes
1 variable 1 sample	Examination of a single sample	Procedure for grouping a frequency distribution, Box 2.1; stem-and-leaf display, Section 2.5; testing for outliers, Section 13.4 Computing median of frequency distribution, Box 4.1 Computing arithmetic mean: unordered sample, Box 4.2; frequency distribution, Box 4.3 Computing standard deviation: unordered sample, Box 4.2; frequency distribution, Box 4.3 Setting confidence limits: mean, Box 7.2; variance, Box 7.3 Computing g_1 and g_2, Box 6.2		Confidence limits for a percentage, Section 17.1 Runs test for randomness in dichotomized data, Box 18.3
	Comparison of a single sample with an expected frequency distribution	Normal expected frequencies, Box 6.1 Goodness of fit tests: parameters from an extrinsic hypothesis, Box 17.1; from an intrinsic hypothesis, Box 17.2 Kolmogorov–Smirnov test of goodness of fit, Box 17.3 Graphic "tests" for normality: large sample sizes, Box 6.3; small sample sizes (rankit test), Box 6.4 Test of sample statistic against expected value, Box 7.4		Binomial expected frequencies, Box 5.1 Poisson expected frequencies, Box 5.2 Goodness of fit tests: parameters from an extrinsic hypothesis, Box 17.1; from an intrinsic hypothesis, Box 17.2
1 variable ≥2 samples	Single classification	Single classification anova: unequal sample sizes, Box 9.1; equal sample sizes, Box 9.4 Planned comparison of means in anova, Box 9.8; single degree of freedom comparisons of means, Box 14.10 Unplanned comparison of means: T-method, equal sample sizes, Box 9.9; T′, GT2, and Tukey–Kramer, unequal sample sizes, Box 9.10; Welsch step-up, Box 9.11; STP test, Section 9.7; contrasts using Scheffé, T, and GT2, Box 9.12; multiple confidence limits, Section 14.10 Estimate variance components: unequal sample sizes, Box 9.2; equal sample sizes, Box 9.3 Setting confidence limits to a variance component, Box 9.3 Tests of homogeneity of variances, Box 13.1 Tests of equality of means when variances are heterogeneous, Box 13.2	Kruskal–Wallis test, Box 13.5 Unplanned comparison of means by a nonparametric STP, Box 17.5	*G*-test for homogeneity of percentages, Boxes 17.5 and 17.8 Comparison of several samples with an expected frequency distribution, Box 17.4; unplanned analysis of replicated tests of goodness of fit, Box 17.5
	Nested classification	Two-level nested anova: equal sample sizes, Box 10.1; unequal sample sizes, Box 10.4 Three-level nested anova: equal sample sizes, Box 10.3; unequal sample sizes, Box 10.5		
	Two-way or multi-way classification	Two-way anova: with replication, Box 11.1; without replication, Box 11.2; unequal but proportional subclass sizes, Box 11.4; with a single missing observation, Box 11.5 Three-way anova, Box 12.1 More-than-three-way classification, Section 12.3 and Box 12.2 Test for nonadditivity in a two-way anova, Box 13.4	Friedman's method for randomized blocks, Box 13.9	Three-way log-linear model, Box 17.9 Randomized blocks for frequency data (repeated testing of the same individuals), Box 17.11

Experimental designs expressed as joint distributions

$$
\begin{aligned}
\left[\alpha, \beta, \sigma \mid \mathbf{y}\right] &\propto \prod_{i=1}^{N}\prod_{j=1}^{M} \text{normal}\left(y_{i,j} \mid g\left(\alpha, \beta, x_{i,j}, w_{i,j}\right), \sigma^2\right) \times \\
&\quad \text{normal}\left(\alpha \mid 0, 1000\right) \text{normal}\left(\beta_1 \mid 0, 1000\right) \times \\
&\quad \text{normal}\left(\beta_2 \mid 0, 1000\right) \text{uniform}\left(\sigma \mid 0, 100\right) \\
g\left(\alpha, \beta, x_{i,j}, w_{i,j}\right) &= \alpha + \beta_1 x_{i,j} + \beta_2 w_{i,j}
\end{aligned}
$$

Photo c/o of the Minnesota Agricultural Experiment Station at http://www.maes.umn.edu.

Design matrix: What is this?

E[y] Design Matrix

$$\begin{bmatrix} \widehat{y_1} \\ \widehat{y_2} \\ \widehat{y_3} \\ \widehat{y_4} \\ \vdots \\ \widehat{y_N} \end{bmatrix} = \begin{bmatrix} 1 & 1.2 \\ 1 & 3.4 \\ 1 & 1.7 \\ 1 & 7.9 \\ & \vdots \\ 1 & 4.3 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \alpha + \beta \times 1.2 \\ \alpha + \beta \times 3.4 \\ \alpha + \beta \times 1.7 \\ \alpha + \beta \times 7.9 \\ \vdots \\ \alpha + \beta \times 4.3 \end{bmatrix}$$

- Great! But how do we handle categorical experimental treatments?

Parameterize a model with categorical predictors

unique solution

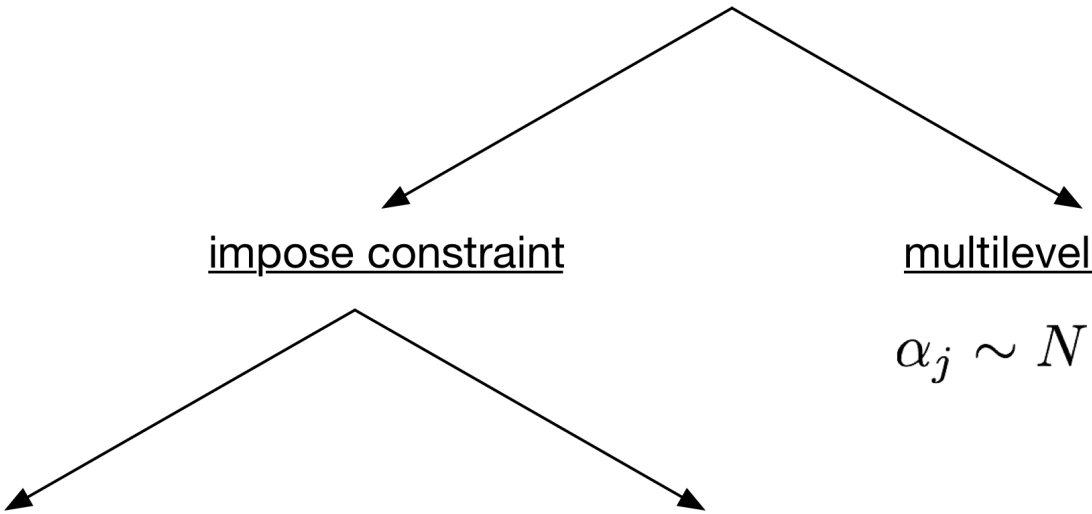
cell means model

$$g(\boldsymbol{\alpha}, \mathbf{x}) = \alpha_j x_{i,j}^{(j)}$$

over-parameterized

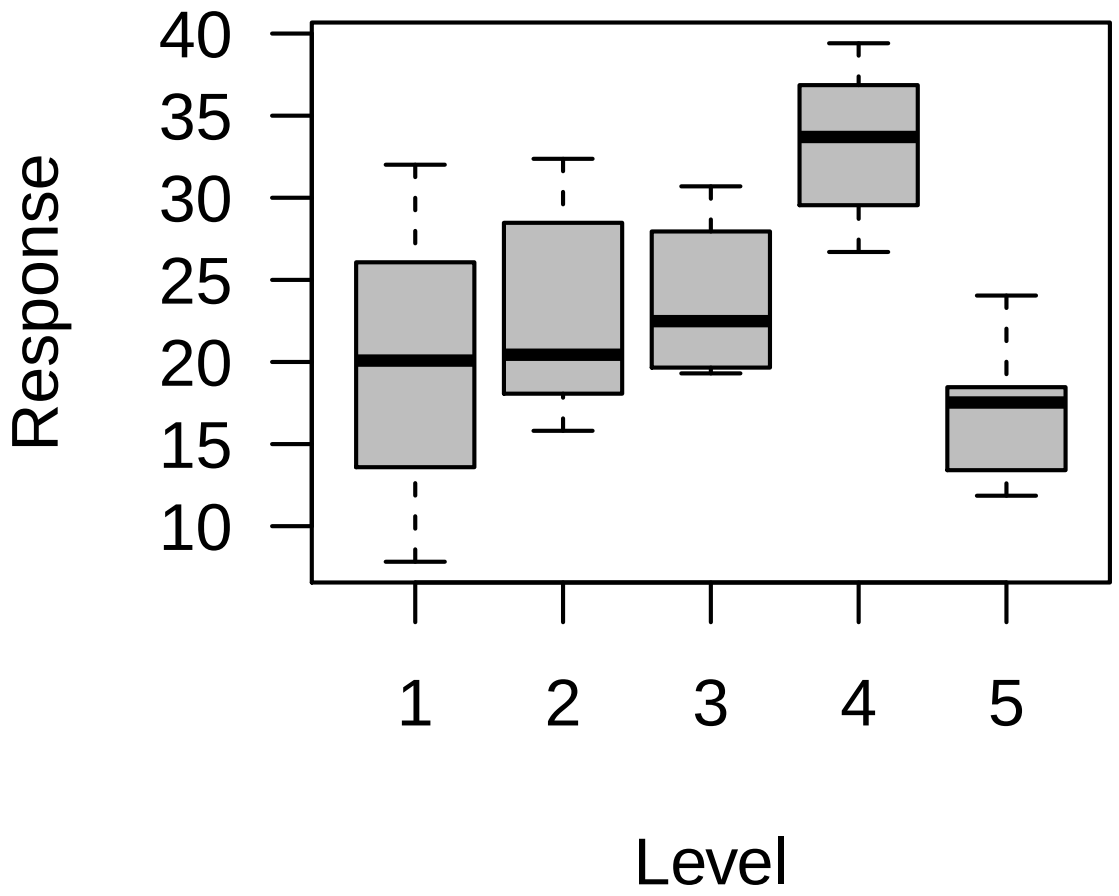
effects model

$$g(\mu, \boldsymbol{\alpha}, \mathbf{x}) = \mu + \sum_{j=1}^{M} \alpha_j x_{i,j}^{(j)}$$



Simulate data for CRD Design

- Completely randomized design (CRD) with 1 factor and 5 levels
- Simulate data for a factor with 5 levels
- 10 replicates per level, 50 replicates overall



Parameterize a model with categorical predictors

unique solution

cell means model

$$g(\boldsymbol{\alpha}, \mathbf{x}) = \alpha_j x_{i,j}^{(j)}$$

over-parameterized

effects model

$$g(\mu, \boldsymbol{\alpha}, \mathbf{x}) = \mu + \sum_{j=1}^{M} \alpha_j x_{i,j}^{(j)}$$

What is this superscript?

...pose constraint

multilevel model

$$\alpha_j \sim N\left(0, \sigma^2\right)$$

set to zero

$$\alpha_M = 0$$

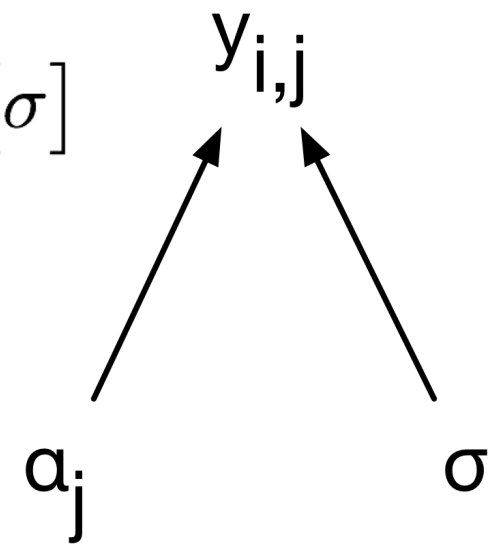
sum to zero

$$\alpha_M = -(\alpha_1 + \alpha_2 + \cdots + \alpha_{M-1})$$

Cell Means Model: Joint and DAG

$$[\boldsymbol{\alpha}, \sigma \mid \mathbf{y}] \propto \prod_{i=1}^{N} \prod_{j=1}^{M} \left[y_{i,j} \mid g(\boldsymbol{\alpha}, \mathbf{x}), \sigma^2 \right] [\alpha_j] [\sigma]$$

$$g(\boldsymbol{\alpha}, \mathbf{x}) = \alpha_j x_{i,j}^{(j)}$$



- Interest in group means and not effects
- Have prior information for group means
- Lack prior information for group means - use vague priors
- Number of parameters = number of unknowns
- Recover effects or grand mean as derived quantities

Cell Means Model: Design Matrix

E[y] Design Matrix Parameters

$$\begin{bmatrix} \widehat{y_{n,1}} \\ \widehat{y_{n,2}} \\ \widehat{y_{n,3}} \\ \widehat{y_{n,4}} \\ \widehat{y_{n,5}} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \end{bmatrix}$$

Cell Means Model: JAGS

```
#priors
for (i in 1:5) {
  alpha[i] ~ dnorm(0, 0.001)
}
sigma ~ dunif(0, 100)
tau <- 1 / ( sigma * sigma)

# Likelihood
for (i in 1:50) {
  y[i] ~ dnorm(alpha[x[i]], tau)
}

# Derived quantities
effect.2.1 <- mu[2] - mu[1]
effect.3.1 <- mu[3] - mu[1]
grandMean <- mean(mu[])
```

- Use the index trick!
- Compute effects and grand mean as derived quantities

Parameterize a model with categorical predictors

unique solution

cell means model

$$g(\boldsymbol{\alpha}, \mathbf{x}) = \alpha_j x_{i,j}^{(j)}$$

over-parameterized

effects model

$$g(\mu, \boldsymbol{\alpha}, \mathbf{x}) = \mu + \sum_{j=1}^{M} \alpha_j x_{i,j}^{(j)}$$

impose constraint

multilevel model

$$\alpha_j \sim N\left(0, \sigma^2\right)$$

set to zero

$$\alpha_M = 0$$

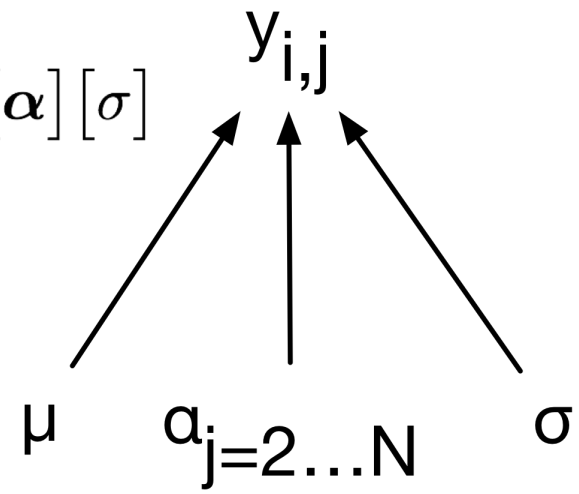
sum to zero

$$\alpha_M = -(\alpha_1 + \alpha_2 + \cdots + \alpha_{M-1})$$

Effects Models - Set to Zero: Joint and DAG

$$[\mu, \boldsymbol{\alpha}, \sigma \mid \mathbf{y}] \propto \prod_{i=1}^{N} \prod_{j=1}^{M} \left[y_{i,j} \mid g(\mu, \boldsymbol{\alpha}, \mathbf{x}), \sigma^2 \right] [\mu] [\boldsymbol{\alpha}] [\sigma]$$

$$g(\mu, \boldsymbol{\alpha}, \mathbf{x}) = \mu + \sum_{j=2}^{M} \alpha_j x_{i,j}^{(j)}$$



- Interest in effects and not means
- Have prior information for effect sizes
- Lack prior information for effect sizes - can estimate conservatively
- Number of parameters $>$ number of unknowns requires constraint
- Recover group means as derived quantities

Effects Model- Set to Zero: Design Matrix

$$\underbrace{\begin{bmatrix} \widehat{y_{n,5}} \\ \widehat{y_{n,1}} \\ \widehat{y_{n,2}} \\ \widehat{y_{n,3}} \\ \widehat{y_{n,4}} \end{bmatrix}}_{\text{E[y]}} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix}}_{\text{Design Matrix}} \underbrace{\begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix}}_{\text{Parameters}} = \begin{bmatrix} \mu \\ \mu + \alpha_1 \\ \mu + \alpha_2 \\ \mu + \alpha_3 \\ \mu + \alpha_4 \end{bmatrix}$$

- CRD with 1 factor and 5 levels - Remove parameter by setting $\alpha_5 = 0$ - Group 5 is now represented by intercept μ - α_j represent deviations from this baseline/control group

Effects Models - Set to Zero: JAGS

```
# Priors
for (i in 1:4){
  alpha[i] ~ dnorm(0, 0.001)
}
mu ~ dnorm(0, 0.001)
sigma ~ dunif(0, 100)
tau <- 1 / ( sigma * sigma)

# Likelihood
for (i in 1:50) {
  y[i] ~ dnorm(yhat[i], tau)
  yhat[i] <- mu + alpha[1]*treatment1[i] + alpha[2]*treatment2[i] + alpha[3]*treatment3[i] +
    alpha[4]*treatment4[i]
}

# Derived quantities
cell[5] <- mu
for (i in 1:4){
  cell[i] <- mu + alpha[i]
}
grandMean <- mean(cell[])
```

- Compute cell and grand means as derived quantities

Parameterize a model with categorical predictors

unique solution

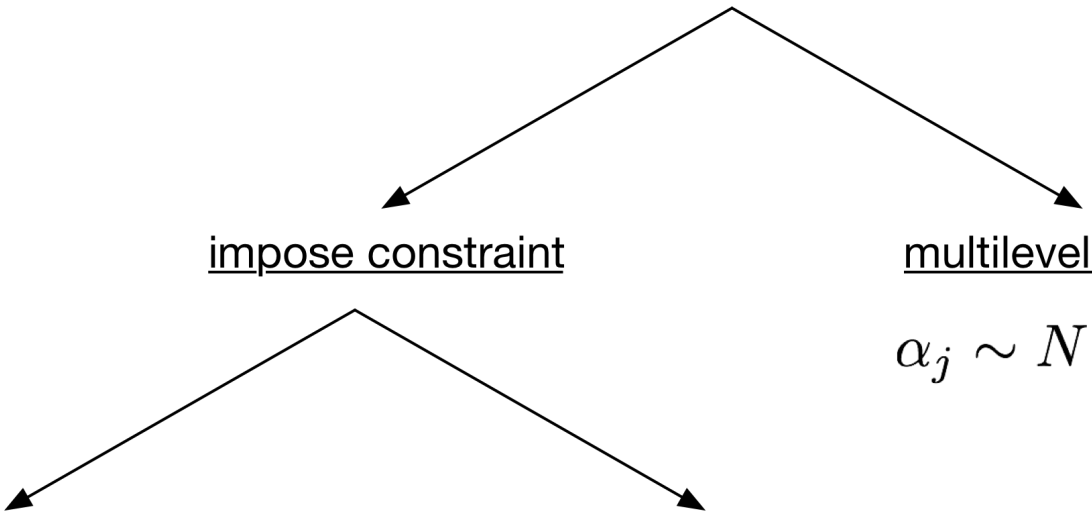
cell means model

$$g(\boldsymbol{\alpha}, \mathbf{x}) = \alpha_j x_{i,j}^{(j)}$$

over-parameterized

effects model

$$g(\mu, \boldsymbol{\alpha}, \mathbf{x}) = \mu + \sum_{j=1}^{M} \alpha_j x_{i,j}^{(j)}$$

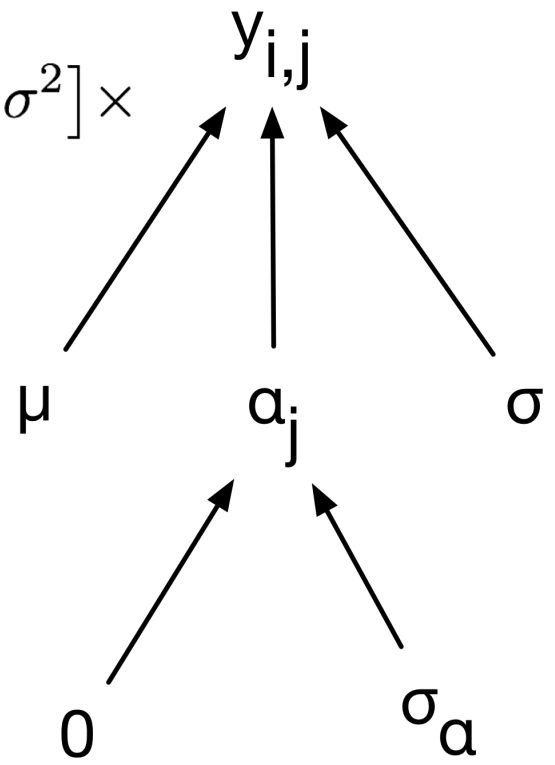


Effects Model - Multi-level: Joint and DAG

$$[\mu, \boldsymbol{\alpha}, \sigma, \sigma_\alpha \mid \mathbf{y}] \propto \prod_{i=1}^{N} \prod_{j=1}^{M} \left[y_{i,j} \mid g(\mu, \boldsymbol{\alpha}, \mathbf{x}), \sigma^2 \right] \times$$

$$[\alpha_j \mid 0, \sigma_\alpha][\mu][\sigma][\sigma_\alpha]$$

$$g(\mu, \boldsymbol{\alpha}, \mathbf{x}) = \mu + \sum_{j=1}^{M} \alpha_j x_{i,j}^{(j)}$$



- Interest in effects and not means
- Have prior information for effect sizes
- Lack prior information for effect sizes - can estimate conservatively
- Number of parameters $>$ number of unknowns is ok! Why?
- Recover group means as derived quantities

Effects Model - Multi-Level: Design Matrix

E[y] Design Matrix Parameters

$$\begin{bmatrix} \widehat{y_{n,1}} \\ \widehat{y_{n,2}} \\ \widehat{y_{n,3}} \\ \widehat{y_{n,4}} \\ \widehat{y_{n,5}} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \end{bmatrix} = \begin{bmatrix} \mu + \alpha_1 \\ \mu + \alpha_2 \\ \mu + \alpha_3 \\ \mu + \alpha_4 \\ \mu + \alpha_5 \end{bmatrix}$$

- CRD with 1 factor and 5 levels
- Intercept, μ, is the grand mean
- α_j represent deviations from the grand mean
- α_j are partially pooled allowing us to estimate all of them directly

Effects Models - Multi-level: JAGS

```
# Priors
mu ~ dnorm(0, 0.001)
for (i in 1:2){
  sigma[i] ~ dunif(0, 100)
  tau[i] <- 1 / ( sigma[i] * sigma[i])
}

# Likelihood
for (i in 1:5){
  alpha[i] ~ dnorm (0, tau[2])
}
for (i in 1:50) {
  y[i] ~ dnorm(y.hat[i], tau[1])
  y.hat[i] <- mu + alpha[x[i]]
}

# Derived quantities
for (i in 1:5){
  cell[i] <- mu + mean(alpha[i])
}
```

- Use index trick!
- Compute cell means as derived quantities

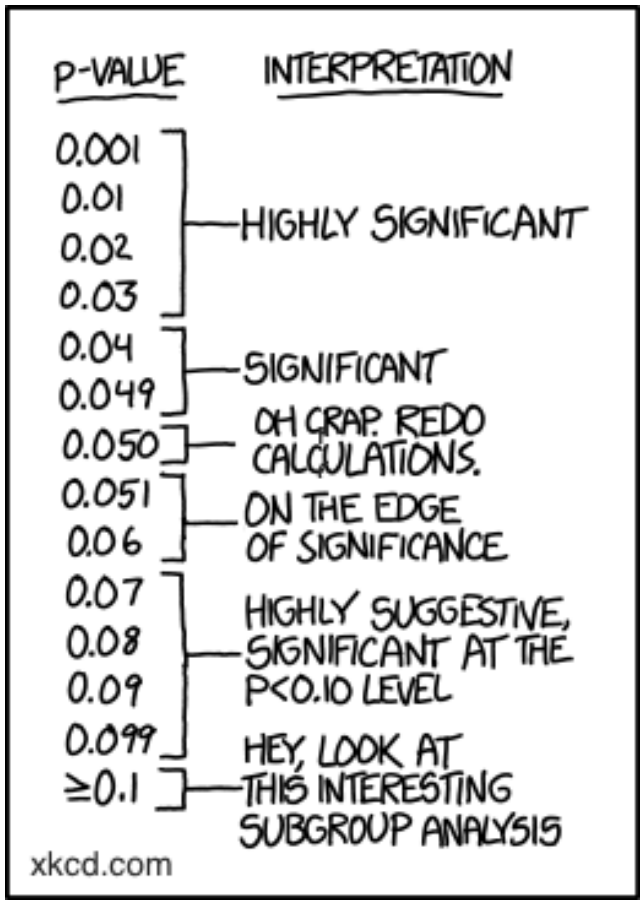
Bayesian Approach to Experimental Analysis

- Flexible framework
- Ease of interpreting effects

Make statements like:

$\Pr(\text{Browsed} > \text{Unbrowsed}) = .8$

CI95: effect of browse = -4.0



Bayesian ANOVA

A way to summarize the "relative importance of different sources of variation in a dataset." (Gelman and Hill, 2007)

- Uses the finite-population SD and not the superpopulation SD
- Can show variation decomposition across multiple levels
- Unbalanced data and complex or incomplete designs easily handled
- Can still be done with "fixed" effects

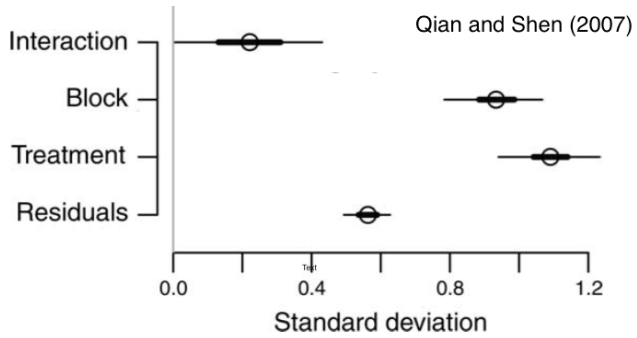


FIG. 1. Seaweed example, with ANOVA display of the estimated standard deviation of the estimated variance components showing a general pattern similar to that of the conventional ANOVA. Circles are estimated posterior means, short thick lines are the 50% posterior credible intervals, and the long thin lines are the 95% posterior credible intervals.

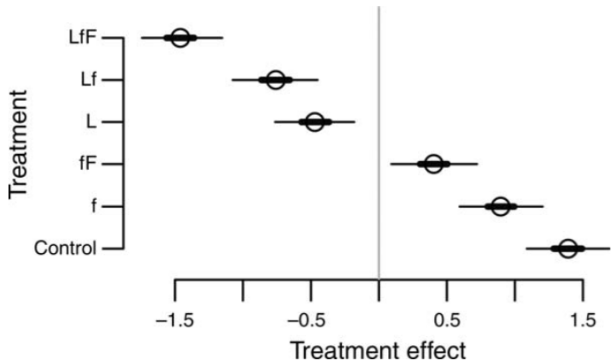


FIG. 2. Estimated treatment main effect of the seaweed grazer example shows that the regeneration rate decreases as grazing pressure increases. The six treatments are: C, control, no grazers allowed; L, only limpets allowed; f, only small fish allowed; Lf, large fish excluded; fF, limpets excluded; and LfF, all grazers allowed. The largest difference between treatments is

Hector et al. 2011, Qian and Shen 2007, Gelman 2005

Bayesian ANOVA: JAGS

- Compute finite-population SDs computation as derived quantities

```
# Priors
mu ~ dnorm(0, 0.001)
for (i in 1:2){
  sigma[i] ~ dunif(0, 100)
  tau[i] <- 1 / ( sigma[i] * sigma[i])
}

# Likelihood
for (i in 1:5){
  alpha[i] ~ dnorm (0, tau[2])
}
for (i in 1:50) {
  y[i] ~ dnorm(y.hat[i], tau[1])
  y.hat[i] <- mu + alpha[x[i]]
  s.yerr[i] <- y[i] - y.hat[i]
}

# Derived quantities
for (i in 1:5){
  cell[i] <- mu + mean(alpha[i])
}
s.alpha <- sd(alpha[])
s.y <- sd(y.err[])
```

Mutiple Mean Comparison

- Fundamentally different approach to mean comparisons
- Shrinkage and/or informed priors

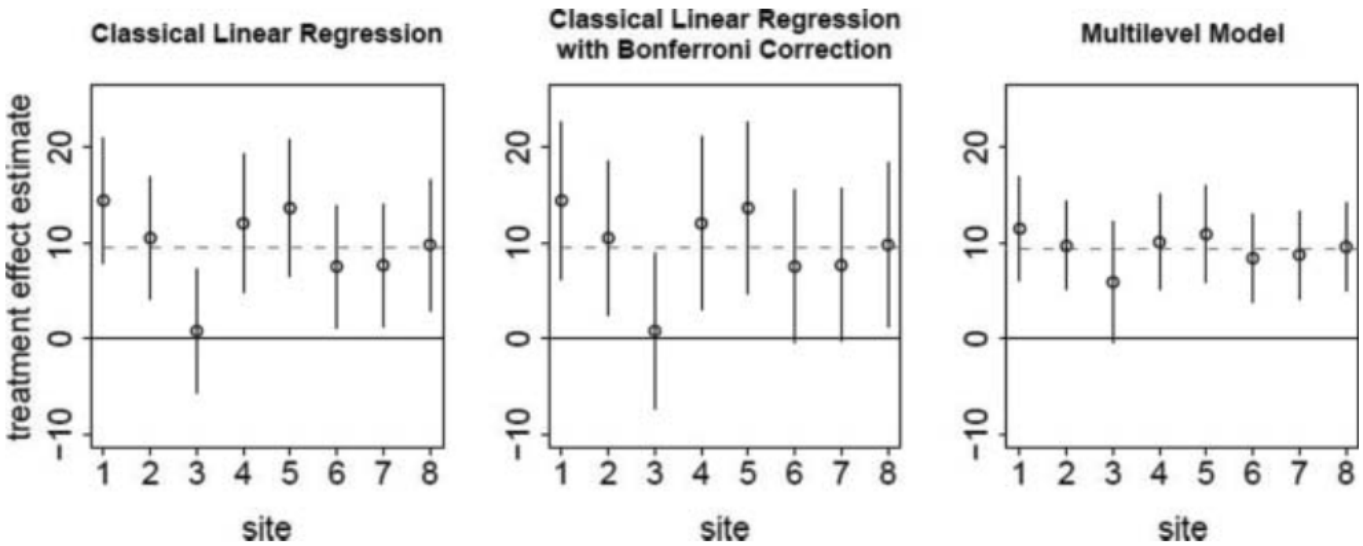


Figure 1. Treatment effect point estimates and 95% intervals across the eight Infant Health and Development Program sites. *Note.* The left panel display classical estimates from a linear regression. The middle panel displays the same point estimates as in the left panel but with confidence intervals adjusted to account for a Bonferroni correction. The right panel displays posterior means and 95% intervals for each of the eight site-specific treatment effects from a fitted multilevel model.

Gelman et al. 2012

Gelman 2013, Gelman et al. 2012

References

[1] A. Gelman. Analysis of variance – why it is more important than ever. Annals of Statistics, 33(1):1–31, 2005.

[2] A. Gelman and J. Hill. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, Boston, MA, USA, 2007.

[3] A. Gelman, J. Hill, and M. Yajima. Why we (usually) don't have to worry about multiple comparisons. Journal of Research on Educational Effectiveness, 5(2):189–211, 2012.

[4] A. Hector, T. Bell, Y. Hautier, F. Isbell, M. Kéry, P. B. Reich, J. van Ruijven, and B. Schmid. BUGS in the analysis of biodiversity experiments: Species richness and composition are of similar importance for grassland productivity. PLoS ONE, 6(3):e17434, 2011.

[5] S. S. Qian and Z. Shen. Ecological applications of multilevel analysis of variance. Ecology, 88(10):2489– 2495, 2007.

[6] A. Gelman and E. Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Department of Statistics, Columbia University, 2013.

Exercise

You have 60 plots spread out over a large area. For each plot, you apply one of three possible treatments (treatment A, B, or C). In addition, you are concerned about the effect of rainfall on these plots so you measure total rainfall on each plot during the course of the experiment.

- What experimental design is this?
- What would the design matrix look like?
- Write the DAG and joint for this experiment.

Let's say you thought the effects of rainfall varied by treatment.

- Modify your DAG and joint distribution to measure these effects.

Exercise

You have 60 plots organized into groups of three. For each group, you apply all three treatments, one treatment per plot. The plots are in a small area so you ignore rainfall this time.

- What experimental design is this?
- What would the design matrix look like?
- Write the DAG and joint for this experiment.

Now assume the number of treatments is 10 instead of 3.

- How would you model the effects hierarchically?