Contents

11 Some data analysis **3**
11.1 Contrasts 3
11.2 Normality assumption 5
11.3 Variance components 6
11.4 Unbalanced settings 7
11.5 Estimating or predicting the a_i 8
11.6 Missing data 9
11.7 Choice of response 10

11

Some data analysis

Most of these notes are about designing how to experimentally gather data with the assumption that they can be largely analyzed with methods familiar from linear regression. Here we look at some ways of analyzing data that are especially suited to designed experiments.

The course begin by grounding experimentation in causal inference. The notions of potential outcomes, randomization, the SUTVA assumption and external validity help us think about experimentation. Then A/B testing and bandit methods bridge us to problems of great current interest in industry. Then we began with more classical experimental design.

The chapters so far have included a number of categorical quantities. There are experimental units which may be plots or subjects and there are sub-units. There are experimental factors. A combination of factors comprises a treatment which may or may not involve important interactions. Those factors can be fixed or random, nested or crossed. We also saw blocks and replicates and repeated measures.

11.1 Contrasts

Beyond just rejecting H_0 or not rejecting it, we have an interest in the different expected values of Y. For a one way fixed effects model with

$$\mathbb{E}(Y_{ij}) = \mu + \alpha_i \equiv \mu_i$$

the comparisons of interest involve certain differences among the μ_i or α_i. We might want to compare two expected outcomes through $\mu_2 - \mu_7$. If we are compariing effectiveness of five soaps where the first three contain phosphates

and the other two do not then we might be interested in

$$(\mu_1+\mu_2+\mu_3)/3-(\mu_4+\mu_5)/2.$$

If we are comparing a new product to three old ones we might study

$$\mu_1-(\mu_2+\mu_3+\mu_4)/3.$$

These are all examples of ***contrasts***. Contrasts take the form $\sum_{i=1}^I \lambda_i\mu_i$ where $\sum_i \lambda_i = 0$ and, to remove an uninteresting case, $\sum_i \lambda_i^2 > 0$. A contrast also satisfies $\sum_{i=1}^I \lambda_i\alpha_i$. The reason why we have so much less interest in μ than α_i is that μ does not affect any comparisons of the levels of this factor and so does not affect many of our choices. Perhaps if μ is bad enough we might not want to use any of the levels of our factor, but when as usual we have to choose, μ plays no role in $\mathbb{E}(Y)$.

In the one way layout we can test a contrast with a t test, via

$$t=\frac{\sum_i \lambda_i\bar{Y}_{i\bullet}}{s\sqrt{\sum_i \lambda_i^2/n}}\sim t_{(N-k)} \qquad s=\sqrt{\mathrm{MSE}} \qquad N=\sum_{i=1}^I n_i.$$

We saw earlier that the presence of a random effect can complicate the inference on a fixed effect with which it is crossed. If A is fixed and B is random we can use

$$t=\frac{\sum_i \lambda_i\bar{Y}_{i\bullet\bullet}}{s\sqrt{\sum_i \lambda_i^2/(nB)}}\sim t_{((I-1)(J-1))} \qquad s=\sqrt{\mathrm{MSAB}}.$$

This formula is for a balanced setting. When MSAB is the appropriate denominator for our F test it provides the appropriate value of s for our t-test. The degrees of freedom to use are the number underlying the estimate s.

Suppose that we have a factor that represents I equispaced levels of a continuous variable. For instance 20kg, 40kg, 60kg and 80kg of fertilizer. It is then interesting to test the extent of a linear trend in the average responses Y. Centering these levels produces a contrast $\lambda=(-30,-10,10,30)$. A test of $\sum_i \lambda_i\alpha_i = 0$ is equivalent to one with $\lambda=(-3,-1,1,3)$. This is a ***linear contrast***. When there are an odd number of levels then the central element in the contrast has $\lambda_i = 0$. A test for curvature can be based on a quadratic contrast. If the levels are linearly related to i then we can take

$$\lambda_i=(i-\bar{i})^2-\frac{1}{I}\sum_i(i-\bar{i})^2$$

where $\bar{i}=(I+1)/2$.

Two contrasts λ and λ' are orthogonal if $\sum_i \lambda_i\lambda_i' = 0$. Then $\sum_i \lambda_i\bar{Y}_{i\bullet}$ and $\sum_i \lambda_i'\bar{Y}_{i\bullet}$ are uncorrelated.

11.2 Normality assumption

We have used reference distributions like the t and F distributions derived from an assumption that the errors are normally distributed.

By the central limit theorem, $\bar{Y}_{i\bullet}$ is more nearly normally distributed than the Y_{ij} are. Similarly $\sum_i \lambda_i \bar{Y}_{i\bullet}$ involves more averaging than the individual $\bar{Y}_{i\bullet}$ do and so we anticipate that they more nearly follow a normal distribution. There is a special aspect of contrasts that helps. The main departure of an average from the normal distribution is typically from it's skewness $\mathbb{E}((Y-\mu)^3)/\sigma^3$ being nonzero. If $\bar{Y}_{1\bullet}$ and $\bar{Y}_{2\bullet}$ have similar skewness then much of it cancels in a difference like $\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}$. Since λ_i sum to zero, contrasts must also introduce some degree of cancellation in skewnesses.

Our t-test for a contrast is based on the approximation

$$\sum_i \lambda_i \bar{Y}_{i\bullet} \approx \mathcal{N}\Big(\sum_i \lambda_i \alpha_i, \sum_i \lambda_i^2 \frac{\sigma^2}{n}\Big)$$

or, for unbalanced samples,

$$\sum_i \lambda_i \bar{Y}_{i\bullet} \approx \mathcal{N}\Big(\sum_i \lambda_i \alpha_i, \sum_i \lambda_i^2 \frac{\sigma^2}{n_i}\Big).$$

The F test for H_0 is similarly robust to small departures from normality by the CLT because

$$\begin{pmatrix} \bar{Y}_{1\bullet} \\ \bar{Y}_{2\bullet} \\ \vdots \\ \bar{Y}_{I\bullet} \end{pmatrix} \doteq \mathcal{N}\left(\begin{pmatrix} \mu+\alpha_1 \\ \mu+\alpha_2 \\ \vdots \\ \mu+\alpha_I \end{pmatrix}, \operatorname{diag}\Big(\frac{\sigma^2}{n}, \frac{\sigma^2}{n}, \cdots, \frac{\sigma^2}{n}\Big) \right)$$

This is all we need for our usual derivation. The central limit theorem yields approximately Gaussian $\bar{Y}_{i\bullet}$ values and then sums of squares among them are approximately χ^2. The denominator in the F test uses a mean square such as MSE or MSAB as an estimate of σ^2. It commonly has many more degrees of freedom than the numerator. We do not need a central limit theorem for the denominator just that it yields a good approximation to σ^2.

Tests for a variance or ratio of variances are not robust to non-Gaussianity. A typical MSE is like

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

Under Gaussianity $s^2 \sim \sigma^2 \chi^2_{(n-1)}/(n-1)$. More generally

$$\operatorname{var}(s^2) = \Big(\frac{2}{n-1} + \frac{\kappa}{n}\Big)\sigma^4$$

(Miller, 1997, Chapter 7) where

$$\kappa = \frac{\mathbb{E}((Y-\mu)^4)}{\sigma^4} - 3$$

is the kurtosis of Y. The kurtosis is zero for Gaussian random variables but not necessarily for other variables. When Y has 'heavier tails' than the Gaussian distribution has, then $\kappa > 0$ and s^2 has higher variance than under a Gaussian assumption (and is not χ^2). When $\hat\sigma_1^2$ and $\hat\sigma_2^2$ are not scaled χ^2 random variables then we cannot expect their ratio $\hat\sigma_1^2/\hat\sigma_2^2$ to be approximately F distributed.

The situation is a better for

$$\frac{1}{n-1}\sum_{i=1}^{n}(\bar Y_{i\bullet}-\bar Y_{\bullet\bullet})^2$$

because the CLT is making each $\bar Y_{i\bullet}$ more nearly normally distributed than individual Y_{ij} are.

11.3 Variance components

Our model for a one way layout with random effects is

$$Y_{ij}=\mu+a_i+\varepsilon_{ij},\quad i=1,\dots,I\quad j=1,\dots,n$$

where $a_i\sim\mathcal{N}(0,\sigma_A^2)$ and $\varepsilon_{ij}\sim\mathcal{N}(0,\sigma_E^2)$ are all independent. We have renamed σ^2 to be σ_E^2 here. The variances σ_A^2 and σ_E^2 are called ***variance components***. For a thorough treatment of variance components see Searle et al. (1992).

In this simple variance components model we have

$$\mathbb{E}(\mathrm{MSA})=n\sigma_A^2+\sigma_E^2\quad\text{and}\quad\mathbb{E}(\mathrm{MSE})=\sigma_E^2.$$

It is quite common in more complicated variance components settings to have σ_E^2 in every expected mean square. The reason is that the errors ε_{ij} contribute variance to every observation and there is no way to cancel them out.

We are often most interested in estimating σ_A^2 and σ_E^2 and related ratios such as $\sigma_A^2/(\sigma_A^2+\sigma_E^2)$. We can get unbiased estimates of them by taking

$$\hat\sigma_E^2=\mathrm{MSE}\quad\text{and}\quad\hat\sigma_A^2=\frac{\mathrm{MSA}-\mathrm{MSE}}{n}.$$

The estimate of σ_A^2 is potentially awkward because it can be negative. It is then common to take

$$\hat\sigma_A^2=\max\Big(\frac{\mathrm{MSA}-\mathrm{MSE}}{n},0\Big).$$

This estimate is no longer unbiased. It satisfies $\mathbb{E}(\hat\sigma_A^2)>\sigma_A^2$ because we sometimes increase an unbiased estimate to zero, but never decrease it. If we are averaging estimates like this over many data sets we might prefer to use any negative values we get so as not to get a biased average.

We can also look at this setting through the correlation patterns in the data. If $j\neq j'$ then

$$\mathrm{cov}(Y_{ij},Y_{ij'})=\mathrm{cov}(a_i+\varepsilon_{ij},a_i+\varepsilon_{ij'})=\mathrm{cov}(a_i,a_i)=\sigma_A^2$$

and so

$$\rho \equiv \mathrm{corr}(Y_{ij}, Y_{ij'}) = \frac{\sigma_A^2}{\mathrm{var}(Y_{ij})} = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2}.$$

We can interpret $\hat{\sigma}_A^2$ as an indication that $\rho < 0$. Negative correlations are impossible under our random effects model but distributions with those negative correlations do exist. For instance the correlation matrix for Y_{ij} for $j = 1, \ldots, n$ could be

$$\begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{n \times n}$$

for $-1/(n-1) \leqslant \rho \leqslant 1$. The lower limit on ρ is there to keep the correlation matrix positive semi-definite.

In this correlation model

$$\mathrm{var}\left(\sum_{j=1}^{n} Y_{ij} \right) = n\sigma^2 + n(n-1)\rho\sigma^2$$

or equivalently

$$\mathrm{var}(\bar{Y}_{i\bullet}) = \frac{\sigma^2}{n}(1 + (n-1)\rho).$$

Here $1 + (n-1)\rho$ is the design effect we saw in cluster randomized trials. What we see with $\rho < 0$ is that the variance of $\bar{Y}_{i\bullet}$ or of $\sum_j Y_{ij}$ is less than what it would be for independent observations. If there is some mechanism keeping the total more constant than under independence that could explain negative correlations. Cox (1958) considers animals that share a pen into which some constant amount of food is placed. That could introduce negative correlations in their weights. In a ride hailing setting with a fixed number of passengers we might see negative correlations among the number of rides per driver. In both of those settings we could get positive correlations too. The quantity of food or of passengers could fluctuate up and down generating positive correlations.

If negative correlations are statistically convincing then we can move away from the ANOVA and model the covariance matrix of the data instead.

11.4 Unbalanced settings

In most of these notes we look at balanced data settings. In a few cases the unbalanced sample sizes cause no difficulty. For instance this is true for the one way layout with fixed effects. In other settings unbalanced sample sizes cause severe complications that we do not delve into in a first course on experimental design.

For an illustration consider the one way random effects model

$$Y_{ij} = \mu + a_i + \varepsilon_{ij}, \qquad i = 1, \ldots, I, \quad j = 1, \ldots, n_i$$

and suppose that we want to estimate the grand mean μ. Two natural estimates are

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij} \quad \text{and} \quad \hat{\mu} = \frac{1}{I}\sum_{i=1}^{I}\bar{Y}_{i\bullet}.$$

That is we can average all of the data or average all of the group means. If we actually knew σ_A^2 and σ_E^2 then we could compute the minimum variance unbiased linear combination of $\bar{Y}_{i\bullet}$ as

$$\begin{aligned}\hat{\mu} &= \sum_{i=1}^{I} \bar{Y}_{i\bullet}/\mathrm{var}(\bar{Y}_{i\bullet}) \Big/ \sum_{i=1}^{I} 1/\mathrm{var}(\bar{Y}_{i\bullet}) \\ &= \sum_{i=1}^{I} \bar{Y}_{i\bullet}/(\sigma_A^2 + \sigma_E^2/n_i) \Big/ \sum_{i=1}^{I} 1/(\sigma_A^2 + \sigma_E^2/n_i).\end{aligned}$$

Now if $\sigma_A^2 \gg \sigma_E^2/n_i$ for all i then averaging the $\bar{Y}_{i\bullet}$ would be nearly optimal. If instead, $\sigma_A^2 \ll \sigma_E^2/n_i$ for all i then averaging all the data would be nearly optimal. In practice we don't ordinarily know these variance components but this analysis would let us make a reasonable choice between the two natural estimates above given a guess or assumption on the variance components.

For much more about variance components and unbalanced data, see Searle et al. (1992).

11.5 Estimating or predicting the a_i

There are settings where we actually want to know something about a_i for a specific experimental unit i, even though a_i are thought to be sampled from some distribution.

Searle et al. (1992) give an example from dairy science. Suppose that i represents a bull and j represents a cow that is a daughter of bull i. The setting is nested because cow $j = 1$ for bull i' has nothing to do with cow $j = 1$ for bull i. Now let

$$Y_{ij} = \mu + a_i + \varepsilon_{ij}$$

be some measure of milk yield or quality from cow $j \in \{1, 2, \ldots, \ldots, n_i\}$ of bull i. We might want to estimate a_i in order to judge whether to keep using bull i. Sometimes this problem is described as ***predicting*** a_i because a_i is random. The term "predicting" seems unnatural here because a random effect is not necessarily a quantity defined through the future.

To see how this works we once again make a simplifying assumption that we know μ and σ_A^2 and σ_E^2. If we want to estimate $\mu + a_i$ we can do better than just using $\bar{Y}_{i\bullet}$. Following Searle et al. (1992, Chapter 3.4) suppose that

$$\begin{pmatrix} a_i \\ \bar{Y}_{i\bullet} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ \mu \end{pmatrix}, \begin{pmatrix} \sigma_A^2 & \sigma_A^2 \\ \sigma_A^2 & \sigma_A^2 + \sigma_E^2/n_i \end{pmatrix} \right).$$

Our best estimate of a_i is $\mathbb{E}(a_i \,|\, \bar{Y}_{i\bullet})$ (after arguing that observations from $i' \neq i$ don't help and that only the sufficient statistic $\bar{Y}_{i\bullet}$ is useful). Using properties of the bivariate Gaussian distribution

$$\begin{aligned}\mathbb{E}(a_i \,|\, \bar{Y}_{i\bullet}) &= \mathbb{E}(a_i) + \mathrm{cov}(a_i, \bar{Y}_{i\bullet})\mathrm{var}(\bar{Y}_{i\bullet})^{-1}(\bar{Y}_{i\bullet} - \mu)\\ &= \frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2/n_i}\big(\bar{Y}_{i\bullet} - \mu\big).\end{aligned}$$

Under very strong assumptions of normality and knowing μ, σ_A^2 and σ_E^2 we would estimate (predict) a_i by

$$\frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2/n_i}\big(\bar{Y}_{i\bullet} - \mu\big).$$

We ***shrink*** $\bar{Y}_{i\bullet} - \mu$ towards zero, shirking it a lot if $\sigma_A^2 \ll \sigma_E^2/n_i$. So we estimate $\mu + a_i$ by

$$\frac{\sigma_E^2/n_i}{\sigma_A^2 + \sigma_E^2/n_i}\mu + \frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2/n_i}\bar{Y}_{i\bullet}.$$

This is a linear combination of the population mean μ and the average for unit i. As n_i increases we trust $\bar{Y}_{i\bullet}$ more. This estimate is the ***BLUP***, for best linear unbiased predictor. It minimizes variance among linear combinations of data. With our simplifying assumptions here, the data is just $\bar{Y}_{i\bullet}$. The approach generalizes but becomes complex to depict.

11.6 Missing data

Suppose we have randomized blocks

$$Y_{ij} = \mu + \alpha_i + b_j + \varepsilon_{ij}$$

viewing the block as a random effect, and the observation $Y_{i'j'}$ is missing. We could replace it with whatever minimizes

$$\sum_i \sum_j (Y_{ij}^* - \mu - \alpha_i - b_j)^2$$

where Y_{ij}^* is Y_{ij} if we have it and a parameter if we don't.

This amounts to running a regression on the row and column indicators with a special variable X with $X_{ij} = 1$ if $i = i'$ and $j = j'$ and $X_{ij} = 0$ otherwise. Because we have fit one more parameter we subtract one from the error df getting $(I-1)(J-1)-1$. We can adjust for a small number of missing responses this way. For more details see Montgomery (1997).

11.7 Choice of response

Suppose that $\mathbb{E}(Y_{ijk}) \doteq e^{\mu+\alpha_i+\beta_j+\gamma_k}$. We can still write

$$\mathbb{E}(Y_{ijk}) = \tilde{\mu} + \tilde{\alpha}_i + \tilde{\beta}_j + \widetilde{\alpha\beta}_{ij} + \tilde{\gamma}_k + \widetilde{\alpha\gamma}_{ik} + \widetilde{\beta\gamma}_{jk} + \widetilde{\alpha\beta\gamma}_{ijk}$$

for some new parameters. But we may have made the problem much harder by introducing high order interactions.

In a setting like this, $\log(Y)$ may have a more nearly additive model than Y does. If $\log(Y)$ is nearly additive then Y may not be. The expression above has $\log(\mathbb{E}(Y_{ijk}))$ additive which is not the same has having $\mathbb{E}(\log(Y_{ijk}))$ additive. Conversely, sometimes Y is more nearly additive than $\log(Y)$.

There is a strong simplification from modeling on a nearly additive scale because interactions bring in so many more parameters. Also many of our models and methods use aliasing of the interactions and that is less harmful when they are much smaller. It may then require some after thought to translate a model for transformed Y to get conclusions for $\mathbb{E}(Y)$. A very difficult situation arises when Y is measured in dollars and the model works with $\log(Y)$¿

As a second example, suppose that

$$\mathbb{E}(Y_{ijk}) \doteq \mu + \alpha_i + \beta_j + \gamma_k,$$

and let

$$\tilde{Y} = \begin{cases} 0, & |Y-\tau| > \delta \\ 1, & |Y-\tau| \leqslant \delta. \end{cases}$$

I.e. $\tilde{Y}$ = "Y is ok". Even if we ultimately care about $\tilde{Y}$ it can be much simpler to model Y because $\tilde{Y}$ can have lots of interactions. For instance, suppose that larger i implies larger α_i. Then $\tilde{Y}$ increases with i when $\beta_j + \gamma_k$ is small but decreases with i when $\beta_j + \gamma_k$ is large. That translates into greater impact from interactions. It is better to model Y statistically and then derive consequences for $\tilde{Y}$ from the model for Y.

Bibliography

Cox, D. R. (1958). *Planning of experiments.* Wiley.

Miller, R. G. (1997). *Beyond ANOVA: basics of applied statistics.* CRC press, Boca Raton, FL.

Montgomery, D. C. (1997). *Design and analysis of experiments.* John wiley & sons, 4 edition.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance components.* John Wiley & Sons, New York.