1

One categorical predictor at k levels

Here we look at the regression model in the case where there are $k \geq 2$ groups. We have a single predictor $X \in \{1, 2, \ldots, k\}$. For observation ℓ we get X_ℓ which tells us which group and response $Y_\ell \in \mathbb{R}$. Instead of working with (X_ℓ, Y_ℓ) pairs it is more usual to use two indices getting $Y_{ij} \in \mathbb{R}$ for $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$. The total number of observations is denoted by $N = \sum_{i=1}^{k} n_i$.

We don't assume much about n_i. To rule out trivial cases take $n_i \geq 1$ and almost always we have $n_i \geq 2$ because otherwise there is no way to learn about the variance in group i. Things are simplest in the balanced case where n_i take a common value n.

The cell mean model for grouped data has

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \tag{1.1}$$

where $\varepsilon_{ij} \sim N(0, \sigma^2)$ are independent. Sometimes we re-write equation (1.1) as

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}. \tag{1.2}$$

Here μ is the 'grand mean' of the data and α_i is the 'effect' of level i. The effects model (1.2) is over-parametrized. The design matrix has rows like $(1, 0, \ldots, 1, \ldots, 0)$ in which the first and $i + 1$st elements are 1. The first column is the sum of the others and so the design matrix is singular. To break the tie, we impose a side condition $\sum_{i=1}^{k} n_i \alpha_i = 0$. That makes the average effect equal to zero. One could also use $\sum_{i=1}^{k} \alpha_i = 0$. This is more appealing because it makes the defining condition depend just on the groups we're studying, and not on the values n_i which could be random. But the weighted tie breaker condition makes some formulas simpler, so we use it. When the design is balanced then the two weightings are equivalent.

This setting is sometimes called the one way layout.

1.1 The analysis of variance

The first statistical problem is to decide whether the k groups really differ. This may be expressed through the null hypothesis

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

or equivalently

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0.$$

The cell mean model (1.1) written as a regression is $Y = Z\beta + \varepsilon$ where

$$Y = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{k1} \\ \vdots \\ Y_{kn_k} \end{pmatrix}, \quad Z = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}, \quad \beta = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix}$$

and ε is partitioned similarly to Y. We easily find that

$$\hat{\beta} = (Z'Z)^{-1}Z'Y = \begin{pmatrix} \bar{Y}_{1\cdot} \\ \bar{Y}_{2\cdot} \\ \vdots \\ \bar{Y}_{k\cdot} \end{pmatrix},$$

where

$$\bar{Y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij\cdot}$$

The least squares estimates for the cell mean model are $\hat{\mu}_i = \bar{Y}_{i\cdot}$.

For the effects model (1.2) we can fit the cell mean model first and then take $\hat{\alpha}_i = \bar{Y}_{i\cdot} - \hat{\mu}$ where

$$\hat{\mu} = \bar{Y}_{\cdot\cdot} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{N} \sum_{i=1}^{k} n_i \bar{Y}_{i\cdot}.$$

Another way to estimate the effects model is to use

$$Z = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{N \times k}$$

(dropping the first group from Z) and then solve $\hat{\mu} + \hat{\alpha}_1 = \hat{\beta}_1$ along with $\hat{\beta}_i = \hat{\alpha}_i - \hat{\alpha}_1$ for $i = 2, \ldots, k$ to get $\hat{\mu}$ and $\hat{\alpha}_i$. **[Exercise]**

The sum of squared errors from the cell mean model is

$$\text{SSE} = \text{SSW} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2,$$

where SSW is mnemonic for the sum of squares *within* groups. The sum of squared errors under the null hypothesis of a common group mean is simply

$$\text{SST} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{\cdot\cdot})^2.$$

The term SST refers to the *total* sum of squares.

Using the extra sum of squares principle leads us to testing H_0 with

$$F = \frac{\frac{1}{k-1}(\text{SST} - \text{SSW})}{\frac{1}{N-k}\,\text{SSW}}, \tag{1.3}$$

rejecting H_0 at level α if $F \geq F^{1-\alpha}_{k-1,N-k}$.

The quantity SST − SSW in the numerator of the F statistic can be shown to be

$$\text{SSB} \equiv \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 = \sum_{i=1}^{k} n_i (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2, \tag{1.4}$$

the sum of squares *between* groups. Equation (1.4) is equivalent to the analysis of variance (ANOVA) identity:

$$\text{SST} = \text{SSB} + \text{SSW}. \tag{1.5}$$

Source	df	SS	MS	F
Groups	$k-1$	SSB	$\mathrm{MSB} = \mathrm{SSB}/(k-1)$	MSB/MSW
Error	$N-k$	SSW	$\mathrm{MSW} = \mathrm{SSW}/(N-k)$	
Total	$N-1$	SST		

Table 1.1: This is a generic ANOVA table for the one way layout.

The ANOVA identity (1.5) can be understood in several ways. Algebraicly we get it by expanding $\mathrm{SST} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot} + \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$, and watching the cross terms vanish. We can also divide the three terms by N and interpret it as a version of the identity $V(Y) = E(V(Y \mid X)) + V(E(Y \mid X))$. The name ANOVA derives from this partition, or analysis, of the variance of Y into parts. Finally we can view it as an example of the Pythagorean theorem using the right angle triangle with vertices

$$Y = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{k1} \\ \vdots \\ Y_{kn_k} \end{pmatrix}, \quad \widehat{Y}_{\mathrm{FULL}} = \begin{pmatrix} \bar{Y}_{1\cdot} \\ \vdots \\ \bar{Y}_{1\cdot} \\ \bar{Y}_{2\cdot} \\ \vdots \\ \bar{Y}_{2\cdot} \\ \vdots \\ \bar{Y}_{k\cdot} \\ \vdots \\ \bar{Y}_{k\cdot} \end{pmatrix}, \quad \text{and} \quad \widehat{Y}_{\mathrm{SUB}} = \begin{pmatrix} \bar{Y}_{\cdot\cdot} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \bar{Y}_{\cdot\cdot} \end{pmatrix}.$$

1.2 Mean squares and the ANOVA table

The numerator of the F statistic is $\mathrm{SSB}/(k-1)$. We divide the sum of squares between the k treatment means by $k-1$ which is the number of degrees of freedom among that many means. The denominator of the F statistic is $\mathrm{SSW}/(N-k)$. This is also a sum of squares divided by its degrees of freedom. The ANOVA commonly features such quantities. They are called mean squares. Here we write them MSB and MSW and the F ratio is MSB/MSW. The quantity MSW is the usual unbiased estimate s^2 of the variance σ^2 from the regression of Y in the cell mean model.

The quantities underlying the F test are commonly arranged in an ANOVA table, as shown in Table 1.1. The degrees of freedom are the dimensions of some model spaces.

The vector Y varies in one uninteresting direction, along $(\mu, \mu, \ldots, \mu) \in \mathbb{R}^N$ and in $N-1$ non-trivial directions orthogonal to this line. Of those $N-1$ non-trivial directions there are $k-1$ dimensions along which the group means

can vary about the overall mean. Within group i there are $n_i - 1$ dimensions along which the n_i observations can vary about their common mean. The total degrees of freedom for variation within groups is $\sum_{i=1}^{k}(n_i - 1) = N - k$.

Under H_0 we have $Y \sim N(\mu \mathbb{1}_N, \sigma^2 I_N)$ where $\mathbb{1}_N$ is a column vector of N ones and $\mu \in \mathbb{R}$. The error $\varepsilon = Y - \mu\mathbb{1}_N \sim N(0, \sigma_N^2)$ has a spherically symmetric distribution. It's squared projection on the $k-1$ dimensional space spanned by the cell mean model has a $\sigma^2\chi^2_{(k-1)}$ distribution. It's squared projection on the $N-k$ dimensional space orthogonal to the cell mean model has a $\sigma^2\chi^2_{(N-k)}$ distribution. These are independent of each other. Under H_0 the F statistic MSB/MSW $\sim F_{k-1,N-k}$.

1.3 Power

The distribution of the F statistic for the one way layout is

$$F'_{k-1,N-k}\left(\frac{1}{\sigma^2}\sum_{i=1}^{k} n_i\alpha_i^2\right).$$

We can write the noncentrality parameter as

$$N\frac{\sum_{i=1}^{k}(n_i/N)\alpha_i^2}{\sigma^2}.$$

The noncentrality and hence the power goes up with N, down with σ and it goes up with the variance $\sum_{i=1}^{k}(n_i/N)\alpha_i^2$, among the group means.

Given a hypothetical pattern in the group means as well as an idea of σ we can pick sample sizes to get the desired power.

1.4 Contrasts

When $k = 2$, then MSB/MSW $\sim F_{1,N-2}$. This F statistic is just the square of the usual t statistic. If we reject H_0 then we conclude that the two groups have different means. The reason that $k = 2$ is so simple is that there is just one interesting comparison $\mu_2 - \mu_1$ among the cell means. For larger k there is a whole $k-1$ dimensional family of differences.

When $k \geq 3$ then rejecting H_0 tells us that there are some differences among the cell means μ_i but it does not say which of them are different.

If we're interested in the differences between groups i and i' then we could use a t test based on

$$t = \frac{\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}}{s\sqrt{1/n_i + 1/n_{i'}}} \tag{1.6}$$

and rejecting when $|t| \geq t^{1-\alpha/2}_{(N-k)}$. The degrees of freedom in the t test are the same as in the estimate s^2 of σ^2. For the moment, we'll not worry about whether the variances are equal in the groups.

It is possible that the test in (1.6) rejects the hypothesis $\mu_i = \mu_{i'}$ on the same data for which the F test based on (1.3) fails to reject $\mu_1 = \mu_2 = \cdots = \mu_k$. It seems illogical that two groups could appear to differ while a superset of k groups appear not to differ. But there are $k(k-1)/2$ different pairs that could be tested via (1.6). Under H_0 we expect $k(k-1)\alpha/2$ false rejections from doing all pairwise t tests, but only 1 false rejection from the F test. So it is quite reasonable that (1.6) could be significantly large when the F test is not, despite how hard it might be to explain to a non-statistician.

Less intuitively, we could find that the F test rejects while none of the individual t tests reject. For example we might find that both $\bar{Y}_{2\cdot} - \bar{Y}_{1\cdot}$ and $\bar{Y}_{3\cdot} - \bar{Y}_{1\cdot}$ are almost but not quite significantly large. Perhaps they get a p-value of 0.06. Either of $\mu_2 - \mu_1$ or $\mu_3 - \mu_1$ could be zero. But it may be really unlikely that both of them are zero. In a case like this we would be confident that some difference exists but not sure which of the possible differences is real.

The two group t test above is just one way to look at group differences. More generally we could look at contrasts. A contrast is a linear combination $\sum_{i=1}^k \lambda_i \mu_i$ of the cell means where $\sum_{i=1}^k \lambda_i = 0$, and to rule out trivialities, $\sum_{i=1}^k \lambda_i^2 > 0$. We easily find that $\sum_{i=1}^k \lambda_i \mu_i = \sum_{i=1}^k \lambda_i \alpha_i$. The t test for a contrast is

$$t_\lambda = \frac{\sum_{i=1}^k \lambda_i \bar{Y}_{i\cdot}}{s\sqrt{\sum_{i=1}^k \lambda_i^2/n_i}}. \tag{1.7}$$

The meaningful contrasts depend on the context of the group means. Consider the contrast vectors

$$\begin{aligned}
\lambda_1 &= (4, 4, 4, 0, 0, -3, -3, -3, -3)\\
\lambda_2 &= (-2, -1, 0, 1, 2), \quad \text{and}\\
\lambda_3 &= (1, -1, -1, 1).
\end{aligned}$$

Contrast vector λ_1 tests for a difference between the average of the first 3 groups and the average of the last 4 groups, ignoring the middle two. Dividing it by 12 gives three elements of $1/3$ two elements of 0 and four elements of $-1/4$. Multiplying a contrast vector by a positive value does not change its meaning, and may make it easier to work with. The second contrast λ_2 is sensitive to a linear trend in the group means.

The third contrast could be used to test for a synnergy or interaction when the four groups have a 2×2 structure. In class we had an example where groups 1 and 2 had a fungicide while groups 3 and 4 did not. In that example groups 1 and 3 got a fertilizer and groups 2 and 4 did not. If the two changes give additive results then the mean for group 1 should be $\mu_4 + (\mu_3 - \mu_4) + (\mu_2 - \mu_4) = \mu_2 + \mu_3 - \mu_4$. We test for this synnergy via $\mu_1 - (\mu_2 + \mu_3 - \mu_4) = \mu_1 - \mu_2 - \mu_3 + \mu_4$ which is contrast λ_3.

The square of the t test for a contrast is

$$t_\lambda^2 = \frac{(\sum_{i=1}^k \lambda_i \bar{Y}_{i\cdot})^2}{s^2 \sum_{i=1}^k \lambda_i^2/n_i} \sim F'_{1,N-k}\left(\frac{(\sum_{i=1}^n \lambda_i \mu_i)^2}{\sigma^2 \sum_{i=1}^k \lambda_i^2/n_i}\right).$$

We can get a simpler expression for the balanced case. Then the noncentrality is n/σ^2 times

$$\frac{(\sum_{i=1}^k \lambda_i \mu_i)^2}{\sum_{i=1}^k \lambda_i^2} = \frac{(\sum_{i=1}^k \lambda_i (\mu_i - \mu))^2}{\sum_{i=1}^k \lambda_i^2}. \tag{1.8}$$

The noncentrality does not change if we normalize λ to make it a unit vector. We maximize it by taking the contrast λ to be parallel to the vector $(\mu_1 - \mu, \mu_2 - \mu, \ldots, \mu_k - \mu) = (\alpha_1, \ldots, \alpha_k)$. It might look like we could maximize it by taking a vector λ parallel to $(\mu_1, \ldots, \mu_k)$ but such a vector might not sum to zero, so it would not be a contrast.

If there is a linear trend among the levels, so that $\mu_i = \alpha + \beta i$, then the most powerful contrast for testing it has $\lambda_i \propto (i - (k+1)/2)$. This can have surprising consequences. When $k = 3$ we get $\lambda \propto (-1, 0, 1)$. The contrast does not even use the middle group! The middle group is not exactly useless. It contributes to s^2 and we can use it to test whether the true trend might be nonlinear. Whenever k is odd, the middle group in a linear contrast will get weight 0. When k is even, then all groups get positive weight. For $k = 4$ the contrast comes out proportional to $(-3, -1, 1, 3)$.

Finally we can look at the contrast that gets the most significant t statistic. Again assuming a balanced anova, we can maximize t^2 by taking $\lambda_i = \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}$. This is the *cheater's contrast* because it looks at the actual pattern in the group means and then tests for it.

After some algebra we find that

$$t^2_{\text{cheat}} = \frac{\text{SSB}}{s^2} \sim (k-1) F_{k-1, N-k}. \tag{1.9}$$

This distribution takes larger values than the $F_{1,N-k}$ distribution appropriate to contrasts that are specified in advance. When k is large it can be far larger.

1.5 Multiple comparisons

If we run all $k(k-1)/2$ two group hypothesis tests at level α, then we have much greater than α probability of rejecting H_0 when it holds. Our test-wise type I error rate is α but our experiment-wise error rate could be much higher. If we run all possible contrasts, then we're essentially using the cheater's contrast and have a much higher experiment-wise type I error rate.

It is possible to adjust the test statistics to get control of the experiment-wise type I error rate. The simplest method is Bonferroni's method. We test each of the $k(k-1)/2$ pairs at level $2\alpha/(k(k-1))$ instead of at level α. More generally, if we have a list of m contrasts to test, we can test them at level α/m. Clearly, the higher m is, the more stringent our tests become.

Bonferroni's test gets too conservative when m is large. The cheater's contrast is the most significant of the infinite collection of contrasts that we could

possibly make. But we don't have to divide α by ∞. Scheffé's procedure is to declare the contrast $\sum_i \lambda_i \mu_i$ significantly different from zero if

$$|t_\lambda| \geq \sqrt{(k-1)F^{1-\alpha}_{k-1,N-k}}.$$

If something is significant by this criterion, then no matter how much we scoured the data to come up with our hypothesis, we still have an experiment-wise type I error that is at most α.

There is no free lunch in using Scheffé's criterion. It gives a large threshold. Tukey introduced a test that is more appropriate than Scheffé's when we only want to do the $k(k-1)/2$ pairwise tests, and not all possible contrasts. Suppose that we sort the groups by their group means, so that $\bar{Y}_{(1)\cdot} \leq \bar{Y}_{(2)\cdot} \leq \cdots \leq \bar{Y}_{(k)\cdot}$. For balanced data sets, we will get at least one significant pairwise t-test whenever $\bar{Y}_{(k)\cdot} - \bar{Y}_{(1)\cdot}$ is large enough. Tukey worked out the distribution of

$$Q_{n,k} \equiv \frac{\max_i \bar{Y}_{i\cdot} - \min_i \bar{Y}_{i\cdot}}{s\sqrt{2/n}},$$

(for normal data) and his test declares groups i and i' significantly different when the t test to compare those groups has absolute value larger than the $1-\alpha$ quantile of $Q_{n,k}$.

Sometimes we don't need to look at all pairwise comparisons. In a medical setting group 1 might get a placebo while groups 2 through k each get one of $k-1$ experimental treatments. If all of our comparisons are between the placebo and one of the experimental treatments, then our inferences can be based on the distribution of

$$D \equiv \max_{2\leq i\leq k} \frac{|\bar{Y}_{i\cdot} - \bar{Y}_{1\cdot}|}{s\sqrt{1/n_1 + 1/n_i}}.$$

Dunnett worked out the distribution of this quantity (for the balanced Gaussian setting) and his test declares treatment i different from placebo when the absolute value of the t statistic between these groups exceeds the $1-\alpha$ quantile of D.

Modern computer power now lets do our own home-cooking for multiple comparisons. If we can make a finite list $\mathcal{L} = \{\lambda_1, \ldots, \lambda_m\}$ of the contrasts we're interested in, we can then simulate the distribution of $\max_{1\leq\ell\leq m} |t_{\lambda_\ell}|$ by sampling millions of normally distributed data sets with the same values of n_i as we have in the real data. We can also simulate from non-normal models if we have some that we would prefer.

Multiple comparison testing is a special case of multiple hypothesis testing. We formulate a family of null hypotheses, and make up a test with a family-wise type I error rate of at most α. The hard part is deciding what constitutes a family of tests. At one extreme, each test is its own family. At the other extreme, you could make up a family using every test that you ever make.

Whether some tests belong together as a family of tests depends on the context, and in particular on what actions you will take based on the outcomes

	H_0 not rejected	H_0 rejected	Total
H_0 True	U	V	m_0
H_0 False	T	S	m_1
Total	$m-R$	R	m

of the tests, and on what the costs are from errors. It is an area where reasonable people can disagree.

Suppose for example that you are about to test $k-1$ unproven new treatments against a placebo in a medical context. Whichever one looks best will be followed up on, assuming that it is statistically significant, and developed. If you're wrong, and none of the new treatments are any good, then you'll lose a lot of time and money following up. Here one clearly wants the family to include all of the new versus placebo tests. You lose the same if you make two wrong rejections, following up one, as you would if you made just one wrong rejection.

In other cases the testing situations are independent and what you win or lose is additive (and small compared to your budget or total tolerance for loss). Then each test ought to be it's own family.

1.6 False discovery rates

Having to choose between the family-wise error rate and the experiment-wise error rate is a false dichotomy. After making m hypothesis tests each test has either been rejected or not. Unknown to use, each of those m null hypotheses is either true or not. The possibilities can be counted in a 2×2 grid illustrated in Table 1.6.

Walking through this table, we have made m hypothesis tests of which m_0 really are null and m_1 are not. The total number of tests rejected is R of which V were true nulls and S were false nulls. Of the $m-R$ tests we do not reject, there are U true and T false. So we have made V type I errors (rejecting true H_0) as well as T type II errors (failing to reject a false H_0).

Put another way, we have made R discoverise of which V are false discoveries and S. So our proportion of false discoveries is $\mathrm{FDP}=V/R$ where by convention $0/0$ is interpreted as 0. The false discovery rate is $\mathrm{FDR}=E(V/R)$.

Benjamini and Hochberg have developed a way of controlling the false discovery rate. The BH method, and a good many others, are developed in the book by Dudoit and van der Laan. Suppose that we control the false discovery rate to say 20%. Then on average our rejected hypotheses are at least 80% valid and of course up to 20% false discoveries. FDR methods are often used to screen a large list of hypotheses to identify a smaller list worthy of follow up experiments. It may well be acceptable to have a list of leads to follow up on that are mostly, but not completely, reliable.

The derivation of FDR control methods is outside the scope of this work. But operationally it is easy to do. The BH method works as follows. Take m p-values and sort them, getting $p_{(1)}\le p_{(2)}\le p_{(3)}\le\cdots\le p_{(m)}$. Let $L_i=i\alpha/m$.

If all $p_{(i)} > L_i$ then reject no hypotheses. Otherwise let $r = \max\{i \mid p_{(i)} \le L_i\}$. Finally, reject any hypothesis with $p_i \le p_{(r)}$.

If the m p-values are statistically independent, then the BH procedure gives FDR $\le \alpha\, m_0/m \le \alpha$. When the p-values are dependent, then we can adjust the BH procedure taking $L_i = i\alpha/(mC_m)$ instead where $C_m = \sum_{i=1}^{m} 1/i \doteq \log(m)$. The result controls FDR for any possible dependence pattern among the p_i so it may be conservative for a particular dependence pattern.

1.7 Random effects

The foregoing analysis covers treatments that are known as fixed effects: we want to learn about the actual k levels used. We also often have treatment levels that are random effects. There, the k levels we got are thought of as a sample from a larger population of levels.

For example, if one group got vitamin C and the other did not, then this two group variable is a fixed effect. Similarly, if we have three types of bandage to use that would be a fixed effect.

If instead we have 24 patients in a clinical trial, then we are almost surely looking at a random effect. We may want to learn something about those 24 patients, but our main interest is in the population of patients as a whole. The 24 we got are a sample representing that paopulation. Similarly 8 rolls of ethylene vinyl acetate film in a manufacturing plant will constitute a random effect. Those 8 rolls will soon be used up and irrelevant, but we want to learn something about the process from which they (and future rolls) come.

Some variables are not clearly a fixed or a random effect. Perhaps they're both. For example, in a factory with four lathes, we might be interested in the differences among those four specific machines, or we might be interested more generally in how much the some product measurements vary from lathe to lathe.

The model for random effects is

$$Y_{ij} = \mu + a_i + \varepsilon_{ij}, \quad i = 1, \dots, k,\ j = 1, \dots, n_i, \tag{1.10}$$

where now a_i are random and independent of ε_{ij}. The standard assumptions are that $a_i \sim N(0, \sigma_A^2)$ while $\varepsilon_{ij} \sim N(0, \sigma_E^2)$.

In this model all of the observations have mean μ and variance $\sigma_A^2 + \sigma_E^2$. But they are not all independent. For $j \ne j'$ we have $\mathrm{cor}(Y_{ij}, Y_{ij'}) = \sigma_A^2/(\sigma_A^2 + \sigma_E^2)$.

In this simplest random effects setting we're interested in σ_A^2, often through the ratio σ_A^2/σ_E^2. That way we can tell what fraction of the variation in Y_{ij} comes from the group variable and what fraction comes from measurement error.

We assume that learning about σ_A^2 will tell us something about the other levels of a_i that were not sampled. This is most straightforward if the levels $i = 1, \dots, k$ in the data set really are a sample from the possible levels. In practice it is a little more complicated. The patients in a clinical trial may be something like a sample from the kind of patients that a given clinic sees. They won't usually be a random sample from the population (they're probably sicker

than average). They might or might not be similar to the patients seen at other clinics.

The statistical analysis of the random effects model still uses the ANOVA decomposition:

$$\begin{aligned}
\mathrm{SST} &= \mathrm{SSA} + \mathrm{SSE} \\
\mathrm{SST} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2 \\
\mathrm{SSA} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \\
\mathrm{SSE} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2,
\end{aligned}$$

where now we use SSA for the sum of squares for factor A instead of the SSB for sum of squares between groups.

The fact that the treatment levels are random has not changed the algebra or geometry of the ANOVA decomposition but it does change the distributions of these sums of squares. To get an intuitive idea of the issues we look at the balanced case where $n_i = n$.

First SSE is the sum of k independent terms that are $\sigma_E^2\,\chi^2_{(n-1)}$ for $i = 1, \ldots, k$. Therefore

$$\mathrm{SSE} \sim \sigma_E^2\,\chi^2_{(N-k)}.$$

Next the group means are IID:

$$\bar{Y}_{i\cdot} = \mu + a_i + \frac{1}{n}\sum_{j=1}^{n}\varepsilon_{ij} \sim N\Big(\mu, \sigma_A^2 + \frac{1}{n}\,\sigma_E^2\Big).$$

Then

$$\frac{1}{n}\mathrm{SSA} = \sum_{i=1}^{k}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \sim \Big(\sigma_A^2 + \frac{1}{n}\sigma_E^2\Big)\chi^2_{(k-1)},$$

because it is the sample variance formula from k IID normal random variables.

To test the null hypothesis that $\sigma_A^2 = 0$ we use $F = \mathrm{MSA}/\mathrm{MSE}$ as before. But now

$$\begin{aligned}
F &= \frac{\frac{1}{k-1}\mathrm{SSA}}{\frac{1}{N-k}\mathrm{SSE}} = \frac{\frac{n}{k-1}(\sigma_A^2 + \sigma_E^2/n)\chi^2_{(k-1)}}{\frac{1}{N-k}\sigma_E^2\chi^2_{(N-k)}} \\
&\sim \frac{n\sigma_A^2 + \sigma_E^2}{\sigma_E^2}\,F_{k-1,N-k} \\
&= \Big(1 + n\frac{\sigma_A^2}{\sigma_E^2}\Big)\,F_{k-1,N-k}.
\end{aligned}$$

So under H_0 where $\sigma_A^2 = 0$, we still get the $F_{k-1,N-k}$ distribution. We really had to, because H_0 for the random effects model leaves data with the same distribution as H_0 does for the fixed effects model. But when H_0 does not hold, instead of a noncentral F distribution the F statistic is distributed as a multiple of a central F distribution.

If we reject H_0 we don't have the problem of deciding which means differ, because our model has no mean parameters, except for μ. We may however want to get estimates for σ_A^2 and σ_E^2. The usual way is to work with expected mean squares

$$\begin{aligned} E(\text{MSE}) &= \sigma_E^2, \quad \text{and,} \\ E(\text{MSA}) &= n\sigma_A^2 + \sigma_E^2. \end{aligned}$$

Notice how the residual variance σ_E^2 contributes to the mean square for A. Unbiased estimators are then obtained as

$$\begin{aligned} \hat{\sigma}_E^2 &= \text{MSE}, \quad \text{and} \\ \hat{\sigma}_A^2 &= \frac{1}{n}\Big(\text{MSA} - \text{MSE}\Big). \end{aligned}$$

We can get $\hat{\sigma}_A^2 < 0$. This may happen if the true σ_A^2 is very small. Sometimes one just takes $\hat{\sigma}_A^2 = 0$ instead of a negative value. That will leave an estimate with some bias, because it arises as an unbiased estimate that we sometimes increase.

It is also possible that we can find $\sigma_A^2 < 0$ if observations from the same treatment group are negatively correlated with each other. That could never happen in the model (1.10). But it could happen in real problems. For example if the treatment group were plants in the same planter box (getting the same fertilizer) then competition among the plants in the same box could produce negative correlations among their yields.

Sometimes we want to get estimates of the actual random effect values a_i. Because a_i are random variables and not parameters, many authors prefer to speak of predicting a_i instead of estimating it. If we knew μ, σ_A^2, and σ_E^2, then we could work with the joint normal distribution of $(a_1, \ldots, a_k, Y_{11}, \ldots, Y_{kn_k})'$ and predict a_i by $E(a_i \mid Y_{11}, \ldots, Y_{kn_k})$. That is we predict the unmeasured value by it's conditional expectation given all of the measured ones. The result is

$$\hat{a}_i = \frac{n\sigma_A^2}{\sigma_E^2 + n\sigma_A^2}\Big(\bar{Y}_{i\cdot} - \mu\Big).$$

Then the corresponding estimator of $E(Y_{ij})$ is

$$\mu + \hat{a}_i = \frac{n\sigma_A^2 \bar{Y}_{i\cdot}}{\sigma_E^2 + n\sigma_A^2} + \frac{\sigma_E^2 \mu}{\sigma_E^2 + n\sigma_A^2}. \tag{1.11}$$

Equation (1.11) can be seen as shrinking the observed group average towards the overall mean μ. The amount of shrinkage increases with the ratio $\sigma_E^2/(n\sigma_A^2)$.

The book "Variance Components" by Searle, Casella and McCulloch gives an extensive treatment of random effects models. It includes coverage of maximum likelihood estimation of variance components in models like the one considered here as well as much more general models. It also looks at predictions of random effect components using estimated values of the parameters.