Contents

4 Paired and blocked data, randomization inference 3

4.1 The ordinary two sample t-test 4

4.2 Randomization fixes assumptions 5

4.3 Paired analysis 7

4.4 Blocking 9

4.5 Basic ANOVA 9

4.6 ANOVA for blocks 12

4.7 Latin squares 13

4.8 Esoteric blocking 15

4

Paired and blocked data, randomization inference

In this lecture we begin to look at some more traditional areas of experimental design. Much of it is based on the work by George Box and co-authors. I quite like this book: Box et al. (1978). I'm citing the first edition which I prefer to the second. Wu and Hamada (2011) cover many of the same ideas with a rich collection of examples, mostly from industrial experimentation.

These basic experimental design ideas have been used to give us about a century of exponential growth in the quality and abundance of food and medicine and industrial products. Ideas and insights from domain experts get boosted by the efficiency with which well designed experiments can speed up learning of causal relationships.

In this work we take regular regression theory as a prerequisite. Things like normal theory regression, t-tests, p-values, confidence intervals and how to analyze such data are mostly assumed. This course is mostly about making data, while most other courses are about analyzing data. One exception: we will cover the analysis of variance (ANOVA) in more detail than usual statistics courses do. The ANOVA cannot be completely understood just in terms of adding binary predictors, sometimes called a one-hot encoding. There is a bit more going on.

The class web page has notes from Stat 305A on the one way ANOVA. Read up through Chapter 1.2 and then Chapter 1.7 on random effects which we will cover later. In between there is material on statistical power, interpretation of treatment contrasts, multiple comparisons for ANOVA and false discovery rates.

4.1 The ordinary two sample t-test

Let's recall how we would do a t-test for a treatment effect. We have data Y_{ij} for treatment groups $i = 1, 2$ and observations $j = 1, \dots, n_i$. Think of i as $W + 1$, where $W \in \{0, 1\}$ is the treatment variable in causal inference from Chapter 1. The goal is to learn about $\Delta = \mathbb{E}(Y_{1i}) - \mathbb{E}(Y_{2i})$. Defining this expectation will require a model, and unlike the science tables in potential outcomes, Δ here does not depend on i. The t statistic is

$$t_{\mathrm{obs}} = \frac{\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet} - \Delta}{s\sqrt{1/n_1 + 1/n_2}} \sim t_{(n_1+n_2-2)}.$$

This t_{obs} is the observed value of a t-distributed random variable. Here $\bar{Y}_{i\bullet} = (1/n_i)\sum_{j=1}^{n_i} Y_{ij}$ and s^2 is the pooled variance estimate. This result is a miracle. We have an algebraic expression t_{obs} involving our unknown Δ and some quantities that are known after a short computation. Arithmetic that combines knowns and unknowns ordinarily returns an unknown. This result is special, because while t_{obs} is unknown it has a known distribution. It is then called a ***pivotal*** quantity.

Using the pivotal quantity we can get a 99% confidence interval for Δ as

$$\Bigl\{\Delta \mid |t_{\mathrm{obs}}| \leqslant t^{0.995}_{(n_1+n_2-2)}\Bigr\}.$$

If a special value of Δ, call it Δ_0 is not in the confidence interval then we reject $H_0 : \Delta = \Delta_0$ at the 1% level. The usual Δ_0 is of course 0, corresponding to a null hypothesis of no treatment effect. We can get a p-value for $H_0 : \Delta = \Delta_0$ as

$$p = \Pr\bigl(|t_{(n_1+n_2-2)}| \geqslant |t_{\mathrm{obs}}|\bigr).$$

We can also get these results by pooling all $n_1 + n_2$ data into a regular regression model

$$Y_j = \beta_0 + \beta_1 W_i + \varepsilon_i, \quad i = 1, \dots, N \equiv n_1 + n_2 \tag{4.1}$$

where $W_i = 1$ if observation i is from treatment 1 and $W_i = 0$ if observation i is from treatment 2. Defining

$$X = \begin{pmatrix} 1 & W_1 \\ \vdots & \vdots \\ 1 & W_{n_1} \\ 1 & W_{n_1+1} \\ \vdots & \vdots \\ 1 & W_{n_1+n_2} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{pmatrix}$$

and similarly defining $Y \in \mathbb{R}^N$ with the treatment 1 data above the treatment 2 data, we can compute

$$\hat{\beta} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}Y \quad \text{and} \quad s^2 = \frac{1}{N-2}\sum_{i=1}^{N}(Y_j - \hat{\beta}_0 - \hat{\beta}_1 W_i)^2$$

and now

$$t_{\text{obs}} = \frac{\hat\beta_1 - \beta_1}{s\sqrt{((X^\mathsf{T}X)^{-1})_{22}}}.$$

In order to get these pivotal inferences we need to make 4 assumptions:

1) ε_i are normally distributed,
2) $\mathrm{var}(\varepsilon_i)$ does not depend on W_i,
3) ε_i are independent, and
4) there are no missing predictors.

For the last one, we need to know that $\mathbb{E}(Y_j)$ is not really $\beta_0 + \beta_1 W_i + \beta_2 U_i$ for some other variable U_i.

Assumption 1 is hard to believe, but the central limit theorem reduces the damage it causes. Assumption 2 can be serious but does little damage if $n_1 \doteq n_2$. We can also just avoid pooling the variances and use $\sqrt{s_1^2/n_1 + s_2^2/n_2}$ in place of $s\sqrt{1/n_1 + 1/n_2}$.

Assumption 3 is critical and violations can be hard to detect. Assumption 4 is even more critical and harder to detect. We almost don't even notice we are making an assumption about U_i because U_i is missing from equation (4.1).

4.2 Randomization fixes assumptions

Box et al. (1978) consider a hypothetical neighbor with two fertilizers and 11 tomato plants. Let's go with 10 plants. We could plant them in a row like this:

A	A	A	A	A	B	B	B	B	B

That would not be a good design. Maybe there's a hidden trend variable $U_i = i$ where the plots correspond to $i = 1, \dots, 10$ from left to right.

We could instead try:

A	B	A	B	A	B	A	B	A	B

That is better but could still be problematic. For instance there could be correlations between the yield of adjacent plants. Those would be positive if nearby locations had similar favorability. Or they could be negative if one plants roots or shade adversely affected its neighbors.

We could then try randomizing the run order perhaps getting this:

A	B	B	B	A	A	B	A	A	B

A random order cannot correlate with any trend.

Under our model, the t statistic numerator $\hat\Delta - \Delta = \bar Y_{\bullet A} - \bar Y_{\bullet B} - \Delta$ equals

$$\begin{aligned}
&\bigl(\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4 + \varepsilon_5 - \varepsilon_6 - \varepsilon_7 - \varepsilon_8 - \varepsilon_9 - \varepsilon_{10}\bigr)/5, &&\text{A's first}\\
&\bigl(\varepsilon_1 - \varepsilon_2 + \varepsilon_3 - \varepsilon_4 + \varepsilon_5 - \varepsilon_6 + \varepsilon_7 - \varepsilon_8 + \varepsilon_9 - \varepsilon_{10}\bigr)/5, &&\text{alternate}\\
&\bigl(\varepsilon_1 - \varepsilon_2 - \varepsilon_3 - \varepsilon_4 + \varepsilon_5 + \varepsilon_6 - \varepsilon_7 + \varepsilon_8 + \varepsilon_9 - \varepsilon_{10}\bigr)/5, &&\text{random}
\end{aligned}$$

in our three allocations.

If there is an unknown U_i then it is within the ε_i. If $U_i = c \times (i - 5.5)$ then our model has put that U_i inside ε_i and we get a bias of

$$\mathbb{E}(\hat{\Delta}) - \Delta = \begin{cases} -5c, & \text{A's first} \\ -c, & \text{alternate} \\ 0.6c, & \text{random.} \end{cases}$$

Putting A's first gave the worst bias. The alternating plan improved a lot, but could have done very badly with some high frequency bias. The random plan came out best. The bias will be $O_p(1/\sqrt{N})$ under randomization, whether the U_i constitute a trend or an oscillation or something else.

Next, let's consider what happens if there are correlations in the ε_i. We will consider local correlations

$$\mathrm{corr}(Y_i, Y_{i'}) = \begin{cases} 1, & i = i' \\ \rho, & |i - i'| = 1 \\ 0, & \text{else.} \end{cases}$$

Now

$$\mathrm{var}(\hat{\Delta}) = \frac{1}{25} v^{\mathsf{T}} \mathrm{cov}(\varepsilon) v$$

where $v_i = 1$ for $W_i = 1$ and $v_i = -1$ for $W_i = 0$. Using σ^2 for $\mathrm{var}(\varepsilon_i)$, we get

$$\mathrm{var}(\hat{\Delta}) = \frac{2}{5}\sigma^2 + \frac{\sigma^2}{25} \times \begin{cases} 14\rho, & \text{A's first} \\ -18\rho, & \text{alternate} \\ 2\rho, & \text{random.} \end{cases}$$

The data analyst will ordinarily proceed as if $\rho = 0$ especially in small data sets where we cannot estimate ρ very well. For the plants ρ could well be positive or negative making $\mathrm{var}(\hat{\Delta})$ quite different from $2\sigma^2/5$.

Box et al. (1978) take the view that randomization makes it reasonably safe to use our usual statistical models. A forthcoming book by Tirthankar Dasgupta and Donal Rubin will, I expect, advocate for using the actual randomization that was done to drive the inferences.

4.2.1 About permutation testing

The original motivation for the t-test by Fisher was based on the asymptotic equivalence between a t-test and a permutation test. As a result we do not expect permutation tests to repair any problems that would have affected the t-test.

A t-test tests the 'small' null hypothesis $H_0 : \mathbb{E}(Y \mid A) = \mathbb{E}(Y \mid B)$ that the mean of Y is the same for $W = 0$ and $W = 1$. A permutation test addresses the 'large' null hypothesis $\mathcal{H}_0 : \mathcal{L}(Y \mid A) = \mathcal{L}(Y \mid B)$. Here $\mathcal{L}(\cdot)$ refers to the law or distribution of contents, so this hypothesis makes the strong assumption that the distribution of Y is exactly the same for $W = 0$ and $W = 1$. It is a test of $\mathcal{H}_0$ designed to have power versus H_0.

Permutation tests have the advantage that they are very easy to explain to non-statistician users and they appear to have very clear validity.

Permutation tests can be cumbersome. In an observational setting where we get (X_i, Y_i, W_i) for a $W_i \in \{0, 1\}$ and $X_i \in \mathbb{R}$ it is tricky to use permutations to study whether $Y \perp\!\!\!\perp W$. We could permute (W, X) versus Y or we could permute W versus (X, Y). Neither gives an exact test. [This was studied by David Freedman.] Losing exactness loses a lot of the motivation behind permutations.

One of the best analyses of permutation tests is in the statistical theory book by Lehmann and Romano. They show how it comes from a group symmetry argument.

We will take the BHH view that if our experiment was randomized then we are reasonably safe to use the usual regression models.

4.3 Paired analysis

The next (hypothetical) example from BHH involves 10 kids and running shoes. There were two different materials for the soles of those shoes. Each kid gets one material on the right shoe and the other one on the left. We can diagram the situation as follows, deciding randomly whether to use left or right for material A:

A	B	B	A	B	A	B	A	A	B
A	B	B	A	A	B	A	B	B	A

BHH contemplate very big differences between the kids. Suppose that some are in the chess club while others prefer skateboarding. Figure 4.1 shows an exaggerated simulated example of how this might come out. The left panel shows that tread wear varies greatly over the 30 subjects there but just barely between the treatments. The right panel shows a consistent tendency for tread B to show more wear than tread A, though with a few exceptions.

The way to handle it is is via a paired t-test. Let $D_i = Y_{1i} - Y_{2i}$ for $i = 1, \dots, n$ (so there are $N = 2n$ measurements). Then do a one-sample t-test for whether $\mathbb{E}(D) = \Delta$ where Δ is ordinarily 0.

The output from a paired t-test on this data is

```
t = -2.7569, df = 29, p-value = 0.009989
95 percent confidence interval: -0.59845150 -0.08868513
```

with of course more digits than we actually want. The difference is barely significant at the 1% level. An unpaired t-test on this data yields:

```
t = -0.2766, df = 57.943, p-value = 0.7831
95 percent confidence interval:  -2.829992  2.142856
```

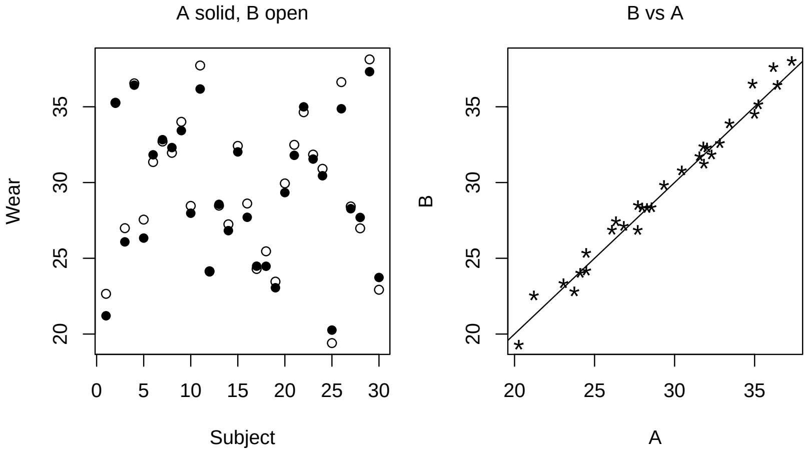

Figure 4.1: Hypothetical shoe wear numbers for 30 subject and soles A versus B.

and the difference is not statistically significant, with a much wider confidence interval.

In this setting the paired analysis is correct or at least less wrong and that is not because of the smaller p-value. It is because the unpaired analysis ignores correlations between measurements for the left and right shoe of a given kid.

In class somebody asked what would be missing from the science table for this example. We get both the A and B numbers. What we don't get is what would have happened if a kid who got $\boxed{A}\boxed{B}$ had gotten $\boxed{B}\boxed{A}$ instead. The science table would have had a row like $\boxed{LA}\boxed{LB}\boxed{RA}\boxed{RB}$ for each kid and we would only see two of those four numbers. We would never get $\boxed{LA}\boxed{LB}$ for any of the kids. It is certainly possible that there are trends where left shoes get a different wear pattern than right shows. Randomization protects against that possibility.

If we model the (Y_{1j}, Y_{2j}) pairs as random with a correlation of ρ and equal variance σ^2 then our model gives

$$\mathrm{var}(D_j) = \mathrm{var}(Y_{1j} - Y_{2j}) = 2\sigma^2(1-\rho)$$

and we see that the higher the correlation, the more variance reduction we get. Experimental design offers possibilities to reduce the variance of your data and this is perhaps the simplest such example.

The regression model for this paired data is

$$Y_{ij} = \mu + b_j + \Delta W_{ij} + \varepsilon_{ij}$$

where b_j is a common effect from the j'th pair, Δ is the treatment effect and $W_{ij} \in \{0, 1\}$ is the treatment variable. This model forces the treatment difference to be the same in every pair. Then

$$D_j = Y_{1j} - Y_{2j} = (\mu + b_j + \Delta W_{1j} + \varepsilon_{1j}) - (\mu + b_j + \Delta W_{2j} + \varepsilon_{2j}) = \Delta + \varepsilon_{1j} - \varepsilon_{2j}.$$

4.4 Blocking

Pairs are blocks of size 2. We can use blocks of any size $k \geqslant 2$. They are very suitable when there are $k \geqslant 2$ treatments to compare. Perhaps the oven can hold $k = 3$ cakes at a time. Or the car has $k = 4$ wheels on it at a time.

If we have $k = 3$ treatments and block size of 3 we can arrange the treatments as follows:

$$\underbrace{\begin{array}{|c|c|c|}\hline B & A & C \\ \hline\end{array}}_{\text{block 1}} \quad \underbrace{\begin{array}{|c|c|c|}\hline C & B & A \\ \hline\end{array}}_{\text{block 2}} \quad \underbrace{\begin{array}{|c|c|c|}\hline B & A & C \\ \hline\end{array}}_{\text{block 3}} \quad \bullet\bullet\bullet \quad \underbrace{\begin{array}{|c|c|c|}\hline A & B & C \\ \hline\end{array}}_{\text{block B}}$$

with independent random assigments within each of B blocks.

Suppose that there are positive correlations for measurements within blocks but independence between blocks. Then differences of averages $\bar{Y}_{A\bullet} - \bar{Y}_{B\bullet}$, $\bar{Y}_{A\bullet} - \bar{Y}_{C\bullet}$, and $\bar{Y}_{B\bullet} - \bar{Y}_{C\bullet}$ should cancel out block effects just like we saw with paired tests and be more accurate than unblocked experiment:

A	C	B	B	C	B	A	A	···	B

with $N = kB$ cells. This latter design would be randomized completely in one of $N!/(B!)^k$ ways.

There are lots of use cases for blocked experiments in agriculture and a few from medicine and industry. In each of the settings below we might have B blocks that each can have k experimental runs.

Treatments	Block	Response
Potato variety	Farm split into k plots	Yield
Cake recipe	Bake event, oven holds k cakes	Moisture
Diets	Litters of k animals	Weight gain
Cholesterol meds	Volunteer	Chol. levels
Sunscreen	Volunteer	Damage
Technician	Shift	Production
Ways to teach reading	School	Comprehension
Ion injection	Cassette of Si wafers	Yield or speed

A block is usually about the same size as our number of treatments. If the problem is to compare a control treatment to $k - 1$ alternatives and the block has size $k + 1$ then we might apply the control treatment twice within each block, especially if comparisons to the control of greatest importance.

4.5 Basic ANOVA

The class web page has Stat 305A notes on how to use regression to analyze this ANOVA.

The statistical model for a most basic ANOVA comparing $k \geqslant 2$ treatments is

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, \dots, k \quad j = 1, \dots, n_i. \tag{4.2}$$

This is called the one-way ANOVA because it has only one treatment factor. We will later consider multiple treatment factors. This model is not identified, because we could replace μ by $\mu - \eta$ and α_i by $\alpha_i + \eta$ for any $\eta \in \mathbb{R}$ without changing Y_{ij}. One way to handle that problem is to impose the constraint $\sum_{i=1}^k n_i \alpha_i = 0$. Many regression packages would force $\alpha_1 = 0$. This model can be written

$$Y_{ij} = \mu_i + \varepsilon_{ij} \tag{4.3}$$

which is known as the ***cell mean model***. We can think of a grid of boxes or cells $\begin{array}{|c|c|c|c|}\hline \mu_1 & \mu_2 & \cdots & \mu_k \\ \hline \end{array}$ and we want to learn the mean response in each of them.

The null hypothesis is that the treatments all have the same mean. That can be written as

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

or as

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

The 'big null' is that $\mathcal{L}(Y_{i1}) = \mathcal{L}(Y_{i2}) = \cdots = \mathcal{L}(Y_{il})$ and that is what permutations test.

We can test H_0 by standard regression methods. Under H_0 the linear model is just

$$Y_{ij} = \mu + \varepsilon_{ij}. \tag{4.4}$$

We could reject H_0 by a likelihood ratio test if the 'full model' (4.3) has a much higher likelihood than the 'sub model' (4.4). When the likelihoods involve Gaussian models, log likelihoods become sums of squares and the results simplify.

Here are the results in the balanced setting where $n_i = n$ is the same for all $i = 1, \dots, k$. The full model has MLE

$$\hat\mu_i = \bar Y_{i\bullet} = \frac{1}{n} \sum_{j=1}^n Y_{ij}$$

and sum of squares

$$\sum_{i=1}^k \sum_{j=1}^n (Y_{ij} - \hat\mu_i)^2 = \sum_{i=1}^k \sum_{j=1}^n (Y_{ij} - \bar Y_{i\bullet})^2.$$

The sub-model from the null hypothesis has MLE

$$\hat\mu = \bar Y_{\bullet\bullet} = \frac{1}{k} \sum_{k=1}^k \bar Y_{i\bullet} = \frac{1}{nk} \sum_{i=1}^k \sum_{j=1}^n Y_{ij}.$$

Source	df	SS	MS	F
Treatments	$k-1$	SSB	MSB = SSB/$(k-1)$	MSB/MSW
Error	$N-k$	SSW	MSW = SSW/$(N-k)$	
Total	$N-1$	SST		

Table 4.1: This is the ANOVA table for a one way analysis of variance.

These sums of squared errors are connected by the ANOVA identity

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij}-\bar{Y}_{\bullet\bullet})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{Y}_{i\bullet}-\bar{Y}_{\bullet\bullet})^2}_{\text{SSB}} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij}-\bar{Y}_{i\bullet})^2}_{\text{SSW}}.$$

The total sum of squares is equal to the sum of squares between treatment groups plus the sum of squares within treatment groups. This can be seen algebraicly by expanding $\sum_{i=1}^{k}\sum_{j=1}^{n}(Y_{ij}-\bar{Y}_{i\bullet}+\bar{Y}_{i\bullet}-\bar{Y}_{\bullet\bullet})^2$. It is also just Pythagoras (orthogonality of the space of fits and residuals) from a first course in regression.

The F-test statistic based on the extra sum of squares principal is

$$F = \frac{\frac{1}{k-1}\bigl(\text{SSE}_{\text{null}}-\text{SSE}_{\text{full}}\bigr)}{\frac{1}{N-k}\text{SSE}_{\text{full}}} = \frac{\frac{1}{k-1}(\text{SST}-\text{SSW})}{\frac{1}{N-k}\text{SSW}} = \frac{\frac{1}{k-1}\text{SSB}}{\frac{1}{N-k}\text{SSW}} \equiv \frac{\text{MSB}}{\text{MSW}}.$$

Here, $N=\sum_i n_k = nk$ is the total sample size. When we divide a sum of squares by its degrees of freedom the ratio is called a mean square. We should reject the null hypothesis if MSB is large. The question 'how large?' is answered by requiring it to be a large enough multiple of MSW. We reject H_0 if $p = \Pr(F_{k-1,N-k} \geqslant F; H_0)$ is small.

These notes assume familiarity with the simple ANOVA tables for regression and the one way analysis of variance. Table 4.1 contains the ANOVA table for this design. There are two sources of variation in this data: treatment groups and error. Because there are k treatments there are $k-1$ degrees of freedom. There are n_i-1 degrees of freedom for error in each of the k treatment groups for a total of $\sum_i(n_i-1)=N-k$. There is often another column for the p-value.

The mean square column provides information on statistical significance. The sum of squares column is about practical significance. For instance $R^2 =$ SSB/SST is the fraction of variation explained by the model terms.

To see why we care about mean squares consider $\varepsilon \sim \mathcal{N}(0,\sigma^2 I_N)$. This is a vector of noise that can be projected onto a one dimensional space parallel to $(1,1,\dots,1)$ where it affects $\bar{Y}_{\bullet\bullet}=\hat{\mu}$, a $k-1$ dimensional space spanned by between treatment differences $\bar{Y}_{i\bullet}-\bar{Y}_{\bullet\bullet}$ where it affects SSB and an $N-k$ dimensional space of within treatment differences $Y_{ij}-\bar{Y}_{i\bullet}$. If Y_{ij} would be just noise ε_{ij} then we would have $\hat{\mu}^2 \sim \sigma^2\chi^2_{(1)}$, SSB $\sim \chi^2_{(k-1)}$ and SSW $\sim \chi^2_{(N-k)}$, all independent. The χ^2 mean equals its degrees of freedom and so we normalize sums of squares into mean squares.

4.6 ANOVA for blocks

The model for a blocked analysis is

$$Y_{ij} = \mu + \alpha_i + b_j + \varepsilon_{ij} \qquad i = 1, \dots, k \quad j = 1, \dots, n.$$

Note that this model does not include an interaction. The treatment differences $\alpha_i - \alpha_{i'}$ are the same in every block j. All values in block j are adjusted up or down by the same constant b_j. We denote it by b_j instead of β_j because we may not be very interested in block j per se. A block might be a litter of animals or one specific run through of our laboratory equipment. In a surfing competition it might be about one wave with three athletes on it. That wave is never coming back so we are only interested in α_i, and maybe how that wave helps us compare α_i for different i, but not b_j.

The parameter estimates here are $\hat\mu = \bar Y_{\bullet\bullet}$, $\hat\alpha_i = \bar Y_{i\bullet} - \bar Y_{\bullet\bullet}$, $\hat b_j = \bar Y_{\bullet j} - \bar Y_{\bullet\bullet}$, and

$$\hat\varepsilon_{ij} = Y_{ij} - \hat\mu - \hat\alpha_i - \hat b_j = Y_{ij} - \bar Y_{i\bullet} - \bar Y_{\bullet j} + \bar Y_{\bullet\bullet} = (Y_{ij} - \bar Y_{i\bullet}) - (\bar Y_{\bullet j} - \bar Y_{\bullet\bullet}).$$

We should get used to seeing these alternating sign and difference of differences patterns.

The ANOVA decomposition is

$$\mathrm{SST} = \mathrm{SSA} + \mathrm{SSB} + \mathrm{SSE}$$

where

$$\begin{aligned}
\mathrm{SST} &= \sum_{i=1}^{k}\sum_{j=1}^{B}(Y_{ij} - \bar Y_{\bullet\bullet})^2,\\
\mathrm{SSA} &= \sum_{i=1}^{k}\sum_{j=1}^{B}(\bar Y_{i\bullet} - \bar Y_{\bullet\bullet})^2,\\
\mathrm{SSB} &= \sum_{i=1}^{k}\sum_{j=1}^{B}(\bar Y_{\bullet j} - \bar Y_{\bullet\bullet})^2, \quad \text{and}\\
\mathrm{SSE} &= \sum_{i=1}^{k}\sum_{j=1}^{B}(Y_{ij} - \bar Y_{i\bullet} - \bar Y_{\bullet j} + \bar Y_{\bullet\bullet})^2.
\end{aligned}$$

The ANOVA table for it is in Table 4.2. You could write SSA as $\sum_i B(\bar Y_{i\bullet} - \bar Y_{\bullet\bullet})^2$ and that is definitely what you would do in a hand calculation. The way it is written is more intuitive. All the sums of squares are sums over all data points.

We test for treatment effects via

$$p = \Pr\Bigl(F_{k-1,(k-1)(B-1)} \geqslant \frac{\mathrm{MSA}}{\mathrm{MSE}}\Bigr).$$

It is sometimes argued that one ought not to test for block effects. I don't quite understand that. If it turns out that blocking is not effective, then we

Source	df	SS	MS	F
Treatments	$k-1$	SSA	$\text{MSA} = \dfrac{\text{SSA}}{k-1}$	$\dfrac{\text{MSA}}{\text{MSE}}$
Blocks	$B-1$	SSB	$\text{MSB} = \dfrac{\text{SSB}}{B-1}$	$(*)$
Error	$(k-1)(B-1)$	SSE	$\text{MSE} = \dfrac{\text{SSE}}{N-k}$	
Total	$N-1$	SST		

Table 4.2: This is the ANOVA table for a blocked design.

could just not do it in the next experiment which might then be simpler to run and have more degrees of freedom for errror. A test can be based on MSB/MSE $\sim F_{B-1,(k-1)(B-1)}$.

The very old text books going back to 1930s place a lot of emphasis on getting sufficiently many degrees of freedom for error. That concern is very relevant when the error degrees of freedom are small, say under 10. The reason can be seen by looking at quantiles of $F_{\text{num,den}}$ such as $F^{.995}_{\text{num,den}}$ and $F^{.005}_{\text{num,den}}$ when the denominator degrees of freedom den is small. Check out `qf` in R, or it's counterpart in python or matlab. It is not a concern in A/B testing with thousands or millions of observations.

4.7 Latin squares

Latin squares let you block on two sources of unwanted variation at once. Suppose that you are testing 4 battery chemistries: A, B, C, D. You have 4 different drivers and 4 different cars. The following diagram has each of A, B, C and D exactly once per row and exactly once per column.

$$\begin{array}{r c} & \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{array}{r} \text{Car 1} \\ 2 \\ 3 \\ 4 \end{array} & \begin{bmatrix} A & B & C & D \\ C & D & A & B \\ B & C & D & A \\ D & A & B & C \end{bmatrix} \end{array}$$

You could have driver 1 (column) test car 1 one with treatment A. Then driver 2 takes car 2 with B and so on through all 16 cases ending up with driver 4 taking car 4 with treatment C. Now if there are car to car differences they are balanced out with respect to treatments. Driver to driver differences are also balanced out. This design only lets one car and one driver be on the track at once.

The model for this design is

$$Y_{ijt} = \mu + \underbrace{a_i}_{\text{row}} + \underbrace{b_j}_{\text{col}} + \underbrace{\tau_k}_{\text{trt}} + \underbrace{\varepsilon_{ijk}}_{\text{err}}\,.$$

k:	1	2	3	4	5	6	7
#:	1	1	1	4	56	9,408	16,942,080

Table 4.3: This is integer sequence number A000315 in the online encyclopedia of integer sequences by Neil J. A. Sloane: `https://oeis.org/A000315`.

It does not allow for any interactions between cars and drivers, cars and batteries or drivers and batteries. Later when we take a closer study of interactions we will see that an interaction between cars and drivers could look like an effect of batteries. If there are no significant interactions like this then a Latin square can be an extremely efficient way to gather information. Otherwise it is risky. Sometimes a risky strategy pays off better than a cautious one. Other times not.

To use a Latin square we start with a basic Latin square, perhaps like this one

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

and then randomly permute the rows and columns. We might as well also permute the symbols in it. Even if that is not necessary, it is easy to do, and simpler to just do it than think about whether you should. The above Latin square is called a ***cyclic*** Latin square because the rows after the first are simply their predecessor shifted left one space with wraparound.

The number of distinct $k \times k$ Latin squares to start with is given in Table 4.3. Two Latin squares are distinct if you cannot change one into the other by permuting the rows and columns and symbols. The number grows quickly with k. Be sure to permute the Latin square, especially if your starting pattern is cyclic. The cyclic pattern will be very bad if there is a diagonal trend in the layout. In many of the original uses the Latin square was made up of k^2 plots of land for agriculture.

Not only are Latin squares prone to trouble with interactions, they also have only a few degrees of freedom. With k^2 data points there are $k^2 - 1$ degrees of freedom about the mean. We use up $k - 1$ of them for each of rows, columns and treatments. That leaves $k^2 - 1 - 3(k-1) = (k-1)(k-2)$ degrees of freedom for error.

Box et al. (1978, Chapter 8) provide a good description of how to analyze Latin squares. I changed their car and driver example to have electric cars. They give ANOVA tables for Latin squares and describe how to replicate them in order to get more degrees of freedom for error. In a short course like this one, we will not have time to go into those analyses.

4.8 Esoteric blocking

There are a lot of more complicated and intricate ways to design experiments in blocks. I describe a few of them below. I consider them things to “know about”. If you ever find that you need them, then being able to connect the problem they solve to their name will help you search for designs and analysis strategies. They’re interesting to contemplate and we can really admire them from an aesthetic point of view. We will return to one of them later when we do space filling designs for computer experiments. For the rest, we don’t have time to study them carefully in a short course like this one.

In the tableaux below:

A α	B β	C γ	D δ
B δ	A γ	D β	C α
C β	D α	A δ	B γ
D γ	C δ	B α	A β

the Latin letters (A, B, C, D) form a Latin square. So do the Greek letters $(\alpha, \beta, \gamma, \delta)$. These two Latin squares are mutually orthogonal meaning that every combination of one Latin letter with one Greek letter appears the same number of times (actually once). From two mutually orthogonal Latin squares ***MOLS*** we get a Graeco-Latin square like the one shown.

We could use a Graeco-Latin square with treatments A, B, C and D blocked out against three factors: one for rows, one for columns and one for Greek letters. We are now in the setting of combinatoric existence and non-existence results. For instance, no Graeco-Latin square exists for $k = 6$. Euler thought there would be none for $k = 10$ but that was proved wrong in the 1950s.

The operational difficulties of arranging a real-world Graeco-Latin square experiment are daunting. It is easy to do in software on the computer. You can even do hyper-Graeco-Latin square experiments with three or more MOLS. For instance if k is a prime number you can have $k - 1$ MOLS and then block out k factors at k levels in addition to a treatment factor at k levels. Or you can embed k^2 points into $[0, 1]^{k+1}$ and have every pairwise scatterplot be a $k \times k$ grid. We will see this later for computer experiments and space-filling designs. Be sure to randomize!

Sometimes the number of levels in a block is less than the number of treatments we have in mind. For instance, consider a club of people are tasting 12 different wines and we don’t want anybody to taste more than 6 of them. Then we would like to arrange our tastings so that each person tastes 6 wines. Those people then represent ***incomplete blocks***.

In an ideal world, each pair of wines would be tasted together by the same number of tasters. That would give us ***balanced incomplete blocks***. This makes sense because the best comparisons between wines A and B will come from people who tasted both A and B. That is, from within block comparisons. There will also be between block comparisons. For instance if many people found A better than B and many found B better than C that provides evidence (through a regression model) that A is better than C. But the within block evidence from having A and C compared by the same people is more informative if the block effects (people) are large.

In sporting leagues we have k teams and we compare then in games that are (ordinarily) blocks of size $B = 2$. A tournament in which each pair of teams played together the same number of times would be a balanced incomplete block design.

There are also ***partially balanced incomplete block*** designs where the number of blocks where two treatments are together is either λ or $\lambda + 1$. So, while not equal, they are close to equal.

We will not consider how to analyze incomplete block designs. If you use one in your project, the other topics from this course will prepare you to read about them and adopt them.

There are even design strategies where one blocking factor has k levels and another has fewer than k levels. So the design is incomplete in that second factor. If you find yourself facing a situation like this, look for ***Youden squares***.

Bibliography

Box, G. E., Hunter, W. H., and Hunter, S. (1978). *Statistics for experimenters*, volume 664. John Wiley and sons New York.

Wu, C. J. and Hamada, M. S. (2011). *Experiments: planning, analysis, and optimization*, volume 552. John Wiley & Sons.