Lecture 5

Two sample setting

Disclaimer: *These lecture notes have not been subjected to the usual scrutiny reserved for formal publications. For instance, citations are mostly omitted or imprecisely made. The notes are meant as a memory aid for students of stat 305A at Stanford University. Copyright: A. B. Owen 2026.*

Here we consider data coming from two groups. They could be coded $\{0,1\}$ or $\{1,2\}$ or $\{-1,+1\}$ or $\{\text{Treatment}, \text{Control}\}$ or $\{\text{Old}, \text{New}\}$ or $\{A, B\}$. We will look at the linear model for this setting. We'll see that a new issue comes up about whether $\sigma^2 I$ is a good model for $\text{var}(Y)$.

5.1 The regression model

We can code the predictors into our design matrix via

$$X = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} 1 \\ \vdots \\ 1 \end{matrix}}\right\} n_0 \\ \left.\vphantom{\begin{matrix} 1 \\ \vdots \\ 1 \end{matrix}}\right\} n_1 \end{matrix}$$

so the first column is an intercept and the second column is the indicator of observations being in the second group. If we code it as $\boldsymbol{x}_i = (1, w_i)^{\mathsf{T}}$

for $w_i \in \{0,1\}$, then we have a column of 1s and a column W of w_i values. Taking $\beta = (\beta_0, \beta_1)^\mathsf{T}$ we have $\mathbb{E}(y\,|\,w=0) = \beta_0$ and $\mathbb{E}(y\,|\,w=1) = \beta_0 + \beta_1$. The second parameter β_1 is generally the more consequential of the two. It gives us the difference in group means $\beta_1 = \mathbb{E}(y\,|\,w=1) - \mathbb{E}(y\,|\,w=0)$.

If y is something we collect or incur or experience after w, then we get β_0 either way and β_1 is the extra we get (on average) from picking $w=1$ instead of $w=0$. Under the null hypothesis $H_0 : \beta_1 = 0$, the two groups have the same expected value of y. More generally we could want a confidence interval for β_1.

The t test for $H_0 : \beta_1 = \beta_{1,0}$ depends on

$$t = \frac{\hat\beta_1 - \beta_{1,0}}{s\sqrt{((X^\mathsf{T}X)^{-1})_{22}}}. \tag{5.1}$$

We do it this way because

$$\mathrm{var}(\hat\beta) = \begin{pmatrix} \mathrm{var}(\hat\beta_0) & \mathrm{cov}(\hat\beta_0,\hat\beta_1) \\ \mathrm{cov}(\hat\beta_0,\hat\beta_1) & \mathrm{var}(\hat\beta_1) \end{pmatrix} = (X^\mathsf{T}X)^{-1}\sigma^2$$

so we want the lower right hand element of $(X^\mathsf{T}X)^{-1}$ in the denominator of (5.1). In class we did the matrix algebra to show that $\mathrm{var}(\hat\beta_1) = n\sigma^2/(n_0 n_1)$. We also worked out that $\hat\beta_1 = \bar y_1 - \bar y_0$ where $\bar y_w$ is the average of the y_i with the given value of $w \in \{0,1\}$. It then follows that $\mathrm{var}(\hat\beta_1) = \sigma^2(1/n_0 + 1/n_1)$ which matches $n\sigma^2/(n_0 n_1)$.

We could also have coded the predictors as

$$X = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}1\\\vdots\\1\end{matrix}}\right\} n_0 \\ \left.\vphantom{\begin{matrix}1\\\vdots\\1\end{matrix}}\right\} n_1 \end{matrix}$$

and then taking $\beta = (\beta_0, \beta_1)^\mathsf{T}$ we get $\mathbb{E}(y\,|\,w) = \beta_w$ for $w \in \{0,1\}$. Then the parameter of greatest interest is the linear combination $\boldsymbol{c}\beta$ for $\boldsymbol{c} = (-1, 1)$. Exercise: show that $\hat\beta_w = \bar y_w$ where $\bar y_w$ is the average of the n_w values y_i for which $w_i = w$. In this parameterization it would be more intuitive to use $(\mu_0, \mu_1)^\mathsf{T}$ as the parameter instead of $(\beta_0, \beta_1)^\mathsf{T}$.

The estimate of σ^2 is

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat\beta_0 - \hat\beta_1 w_i)^2.$$

We can recode the response values for $w = 0$ as y_{0j} for $j = 1, \ldots, n_0$ and similarly for $w = 1$ as y_{1j} for $j = 1, \ldots, n_1$. Then after some algebra,

$$s^2 = \frac{1}{n-2}\left[\sum_{j=1}^{n_0}(y_{0j} - \bar{y}_0)^2 + \sum_{j=1}^{n_1}(y_{1j} - \bar{y}_1)^2\right]$$

where $\bar{y}_w = (1/n_w)\sum_{j=1}^{n_w} y_{wj}$. Then for s_w^2 equal to the variance of y_{wj}, we get

$$s^2 = \frac{1}{n-2}\Big[(n_0 - 1)s_0^2 + (n_1 - 1)s_1^2\Big].$$

5.2 Unequal variances

We ordinarily would expect that the y values in the two groups have different means, though maybe we don't have enough data to tell them apart. There may not be much reason to suppose that they have the same variance either. So then, using the second design matrix in the previous section, we suppose that

$$\mathrm{var}(\bar{y}_1 - \bar{y}_0) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}$$

with $\sigma_1 \neq \sigma_0$. Under this model,

$$\mathrm{var}(Y) = \begin{pmatrix} \sigma_0^2 I_{n_0} & 0 \\ 0 & \sigma_1^2 I_{n_1} \end{pmatrix}.$$

Not having $\mathrm{var}(Y) = \sigma^2 I$ causes a new problem.

Suppose now that we are testing $H_0 : \mu_1 - \mu_0 = \Delta$ where we write $\mu_w = \mathbb{E}(y \mid w)$, and Δ is some value of interest, such as $\Delta = 0$. Our t statistic is

$$t = \frac{\bar{y}_1 - \bar{y}_0 - \Delta}{s\sqrt{1/n_0 + 1/n_1}}. \tag{5.2}$$

The denominator should estimate the standard deviation of the numerator. With unequal variances this might not hold, even as $\min(n_0, n_1) \to \infty$. The variance of the numerator in (5.2) is

$$\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}$$

so it weights the variances in inverse proportion to the sample sizes n_w. The mean square of the denominator is

$$\Big(\frac{1}{n_0} + \frac{1}{n_1}\Big)\Big(\frac{(n_0 - 1)\sigma_0^2 + (n_1 - 1)\sigma_1^2}{n_0 + n_1 - 2}\Big) \approx \Big(\frac{1}{n_0} + \frac{1}{n_1}\Big)\Big(\frac{n_0\sigma_0^2 + n_1\sigma_1^2}{n_0 + n_1}\Big)$$

for large $\min(n_0, n_1)$. The variances are weighted roughly proportionally to sample sizes, not inversely proportionally. That is a mismatch.

The variance is too large, compared to its estimate, by a factor of roughly

$$f = \left[\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}\right] \Big/ \left[\Big(\frac{1}{n_0} + \frac{1}{n_1}\Big)\Big(\frac{n_0\sigma_0^2 + n_1\sigma_1^2}{n_0 + n_1}\Big)\right].$$

Writing $n_1 = Rn_0$ and $\sigma_1^2 = \theta\sigma_0^2$, we get

$$f = \frac{\theta + R}{R\theta + 1}.$$

This argument comes from Chapter 2.3 of Miller (1986).

For $f > 1$ we can consider the t test to be ***naive*** by a factor of f. The real variance is f times what the model says. For $f = 2$, your variance is twice what you think. Then when you think your interval is plus or minus 1.96 standard errors it is actually only $1.96/\sqrt{2} \doteq 1.39$ standard errors. Then the confidence interval coverage that you thought was 95% is only $\Phi(1.39) - \Phi(-1.39) \doteq 83\%$. Similarly, a test on $|t|$ that we naively think has $\alpha = 0.05$ would really have $\alpha = 0.17$.

We saw that there are two, and only two, ways to get $f = 1$. We can have $\theta = 1$, so $\sigma_0^2 = \sigma_1^2$ as is commonly assumed for our regression model, or, we can have $n_0 = n_1$. The second point is that, when the sample sizes are equal, or nearly equal, the two sample t test becomes robust to unequal variances. That is there is an advantage to having balanced sample sizes.

5.3 Welch's t

We might think of running an hypothesis test on $H_0 : \sigma_0 = \sigma_1$ and only doing the t test if it is not rejected. A moment's reflection reveals some problems with that. Variances are generally harder to estimate than means, even if the data are Gaussian. Also, not rejecting H_0 does not by any means establish that H_0 holds. We would need something to do if H_0 were rejected. Also, inserting an ***if then else*** into the middle of your workflow, while sometimes unavoidable, complicates the analysis of what you do.

A better solution is to use Welch's t, denoted

$$t' = \frac{\bar{y}_1 - \bar{y}_0 - \Delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}}}.$$

Now the variance of the numerator, divided by the mean square of the denominator approaches 1 as $\min(n_0, n_1) \to \infty$. So the method is not naive like the usual t statistic. Under the null, $t' \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$.

The usual statistical tests for t' do not use quantiles of $\mathcal{N}(0,1)$. Instead they use $t_{(\nu)}$ for a judiciously chosen ν. The most common approach is due to Satterthwaite. I follow the explanation of Miller (1986, Chapter 2).

We suppose that Welch's t has a Gaussian numerator with variance $\sigma^2 = \sigma_0^2/n_0 + \sigma_1^2/n_1$ and then we want to approximate the square denominator by a distribution

$$\frac{\sigma^2 \chi^2_{(\nu)}}{\nu}$$

for some ν. For σ^2 we can plug in $s_0^2/n_0 + s_1^2/n_1$. For ν, the approach notes that

$$\mathrm{var}\Big(\frac{\sigma^2\chi^2_{(\nu)}}{\nu}\Big) = \frac{2\sigma^4}{\nu} \tag{5.3}$$

and

$$\mathrm{var}\Big(\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}\Big) = \frac{2\sigma_0^4}{(n_0-1)n_0^2} + \frac{2\sigma_1^4}{(n_1-1)n_1^2}. \tag{5.4}$$

The second equation leans strongly on the Gaussian assumption because it does not include any kurtoses. Satterthwaite equates (5.3) to (5.4), and gets

$$\nu = \frac{\Big(\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}\Big)^2}{\frac{1}{n_0-1}\big(\frac{\sigma_0^2}{n_0}\big)^2 + \frac{1}{n_1-1}\big(\frac{\sigma_1^2}{n_1}\big)^2}, \tag{5.5}$$

then plugs in s_w^2 for σ_w^2.

This is a surprising thing to do. It estimates ν using square sample variances s_w^4, so fourth powers of the data come in. It has the saving grace that

$$\min(n_0 - 1, n_1 - 1) \leqslant \nu \leqslant n_0 + n_1 - 2.$$

The right hand side above would be appropriate for t if $\sigma_0^2/\sigma_1^2 = 1$ but t' is not generally t. If $\sigma_0^2/\sigma_1^2 = 1$ and $n_0 = n_1$ then $n_0 + n_1 - 2$ degrees of freedom are appropriate. If $\sigma_0^2/\sigma_1^2 \to 0$ or ∞, then $\nu \to n_1 - 1$ or $n_0 - 1$ respectively which is reasonable if all the variance in $\bar{y}_1 - \bar{y}_0$ is coming from just one of the two samples. Finally for large n_w the t tables for the degree of freedom values bracketing ν lead to almost the same thresholds and hence almost the same confidence intervals.

It is a slight nuisance that inferences under Welch's t are not exact. It is only a slight nuisance because the original t test is not exact either when, as invariably holds, the errors are not Gaussian.

5.4 Discovering argon

There is a nice example of comparing the means of two samples here: `https://www.r-bloggers.com/2013/03/discovering-argon-with-the-2-sample-t-test/`

The story is that Rayleigh was measuring the weight of nitrogen obtained two ways. One was to take air and drive off the oxygen reasoning that what was left was nitrogen. The other way started with several compounds such as ammonia. There were $n_0 = 7$ samples taken from air and $n_1 = 8$ from what the source calls a 'chemical process'. The samples from the air were systematically a bit heavier than the others. This eventually led to the discovery of argon by Rayleigh and Ramsay in 1895.

They could see that the from-air samples had a higher mean but they did not actually use the t test. That was invented later by Gosset working at Guinness and writing as 'Student'.

5.5 Paired t tests

Sometimes the data we get are naturally paired. For example, in a pair of twins, one person might get a new medical treatment and the other might get the prior standard of care. Or when measuring health we might get grip strength from both left and right hands. Or we might have one measurement before some event and another one after it. Box, Hunter and Hunter consider giving kids two different kinds of shoe and then comparing wear on their soles.

We get values y_{1i} and y_{2i} for $i = 1, \dots, n$. Figure 5.1 shows what the joint distribution might look like for the examples in the prior paragraph. In particular, something like grip strength would vary considerably from person to person and, ordinarily, much less from hand to hand for a given person.

We analyze such data with a paired t test. First we form differences $d_i = y_{1i} - y_{2i}$ for $i = 1, \dots, n$. Then we test $\mathbb{E}(d) = \Delta$ (again such as $\Delta = 0$) by using the one sample t test we studied earlier, using

$$t = \sqrt{n}\frac{\bar{d} - \Delta}{s_D}$$

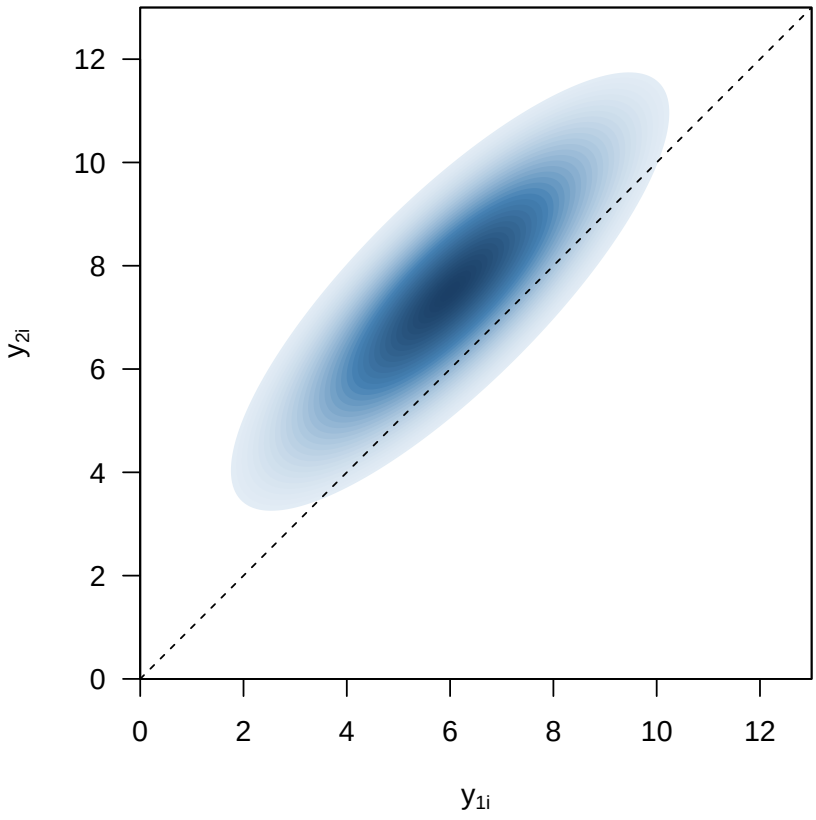


Figure 5.1: Hypothetical nearly Gaussian joint distribution for two variables in a paired context.

where s_D^2 is the sample variance of d_i.

For data like that in Figure 5.1 it is clear that studying the differences is more precise. Each of y_{1i} and y_{2i} has a very large variance while d_i has tiny variance. If

$$\mathrm{var}\left(\begin{pmatrix} y_{1i} \\ y_{2i} \end{pmatrix}\right) = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \sigma^2 \tag{5.6}$$

then $\mathrm{var}(\bar{d}) = 2\sigma^2(1-\rho)/n$. That is the same variance we would get in a least squares regression where we used $\bar{y}_1 - \bar{y}_2$ as our estimate. So $\cdots$ what actually goes wrong with least squares? Hint: look at the usual s^2 in least squares. It is not s_D^2 and it does not correctly capture the variance of the numerator.

The reason we do not use the usual two sample least squares confidence intervals here is that $\mathrm{cov}(y_{1i}, y_{2i}) \neq 0$. A better model is

$$\mathrm{var}\left(\begin{pmatrix} y_{1i} \\ y_{2i} \end{pmatrix}\right) = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \tag{5.7}$$

which includes the case (5.6) from before. Just like in the unequal variance setting, the problem with a least squares solution is not about $\mathbb{E}(Y) = X\beta$. It is about $\mathrm{var}(Y) \neq \sigma^2 I$. Here

$$\mathrm{var}(Y) = \begin{pmatrix} \blacksquare & & & \\ & \blacksquare & & \\ & & \blacksquare & \\ & & & \ddots \\ & & & & \blacksquare \end{pmatrix}$$

with one gray block for each of the n data pairs.

In our paired analysis $\mathrm{var}(\bar{y}_1 - \bar{y}_2) = (\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)/n$ and for $\rho > 0$ this is less than we would have gotten had we collected n observations with $w = 1$ and n unrelated observations with $w = 2$. So pairing, where appropriate is another design choice that we should make. We can model the paired setting as

$$\begin{aligned} y_{1i} &= \mu_1 + a_i + \varepsilon_{1i} \\ y_{2i} &= \mu_2 + a_i + \varepsilon_{2i} \end{aligned}$$

where the shared mean zero random variables a_i induce the correlation. Then $d_i = \mu_1 - \mu_2 + (\varepsilon_{1i} - \varepsilon_{2i})$ has canceled the common within pair effect a_i.

How can we tell if our analysis should be paired or not paired? If there is some real world connection between responses within each pair (e.g., same subject ID), then a paired analysis is appropriate. If there is no such connection then an unpaired analysis is appropriate. There is a slight risk that we get data and were not told about a connection. If we see correlations between y_{2r-1} and y_{2r} or between y_r and $y_{r+n/2}$ for $r = 1, \ldots, n/2$, we might suspect a correlation and then investigate, usually asking the source of our data about this.

It is possible that $\rho < 0$. For example, two plants in the same pot might compete for water or light or fertilizer and then have negatively correlated growth. We cannot get $\rho < 0$ out of a model with independent random a_i as above. That common source of variation can induce positive but not negative correlations.

5.6 Permutation tests

The two sample setting is one where we can introduce the permutation test. Suppose that $y_{1i} \overset{\text{iid}}{\sim} F_1$ for $i = 1, \ldots, n_1$ independently of $y_{2i} \overset{\text{iid}}{\sim} F_2$ for $i = 1, \ldots, n_2$. Here we consider two different null hypotheses. These are

$$H_0 : \mathbb{E}(y_{1i}) = \mathbb{E}(y_{2i}) \quad \text{and} \quad \mathcal{H}_0 : F_1 = F_2.$$

Clearly $\mathcal{H}_0 \Rightarrow H_0$ (when both F_1 and F_2 have means).

Suppose that $\mathcal{H}_0$ holds and let $F = F_1 = F_2$. Under $\mathcal{H}_0$ our n data points could have been generated by sampling n times independently from F and then randomly selecting n_1 of the values to be the y_{1i}, letting the other $n_2 = n - n_1$ of them be the y_{2i}.

Let our test statistic be $\bar{y}_1 - \bar{y}_2$. Now suppose that we select n_1 of the pooled values in Y to be in a new group 1 and let the other n_2 be in a new group 2. We write the test statistic here as $\bar{y}_1^* - \bar{y}_2^*$. One way to do this is to take data $(w_i^*, y_i^*) = (w_{\pi(i)}, y_i)$ where π is a uniformly distributed random permutation of $(1, 2, \ldots, n)$. Uniform here means that all $n!$ of those permutations have probability $1/n!$. Although there are $n!$ permutations there are only

$$N = \binom{n}{n_1} = \frac{n!}{n_1! n_2!}$$

distinct assignments.

In a permutation test, we take all of those assignments and report a p-value

$$p = \frac{\sum_{j=1}^{N} 1\{|\bar{y}_1^{*j} - \bar{y}_2^{*j}| \geqslant |\bar{y}_1 - \bar{y}_2|\}}{N}$$

where $*j$ indicates the j'th random assignment of all n labels. We will not prove thtat this test is valid. While it is intuitively believable the proof is subtle. It's in the book by Lehmann and Romano.

One advantage of the permutation approach is that we could also run it using a difference of medians or 90th percentiles between the two groups.

For even modest data set sizes, N can be extremely large. What we can do in that case is take a random sample of M permutations. We do that by repeatedly randomly permuting the sample labels with respect to the response values. Then we use

$$\tilde{p} = \frac{1 + \sum_{j=1}^{M} 1\{|\bar{y}_1^{*j} - \bar{y}_2^{*j}| \geqslant |\bar{y}_1 - \bar{y}_2|\}}{1 + M}.$$

By putting that 1 in the numerator and denominator we arrange to never get $\tilde{p} = 0$, which we know is wrong because $p \geqslant 1/N$. That 1 in numerator and denominator corresponds to the original 'unpermuted' data. When we sample our permutations it makes sense to take M to be a round number minus one, such as 99,999.

Notice that we can never get $\tilde{p} < 1/(1+M)$. For this reason permutations are not a good choice when we need $\tilde{p} < 5 \times 10^{-8}$ or a similarly small value.

A difficulty with the permutation approach is that, while it may be sensitive to H_0 being wrong it is actually designed to test $\mathcal{H}_0$. It is possible for H_0 to hold while $\mathcal{H}_0$ fails. One quite possible way for this to happen is to have $\sigma_1^2 \neq \sigma_2^2$. The reason is that, Fisher's justification of the t-test involved showing that it asymptotically matches the permutation test based on $\bar{y}_1 - \bar{y}_2$. The two tests agree asymptotically. So we should not expect the permutation test to be better (asymptotically). In particular the permutation test goes wrong if both the variances are unequal and the sample sizes are unequal.

To test that $\mathbb{E}(y_{1i}) - \mathbb{E}(y_{2i}) = \Delta$, we can run a permutation test comparing values of $y_{1i} - \Delta$ to y_{2i}. The unrejected Δ form a confidence interval for the mean difference when $F_1(y) = F_2(y - \Delta)$ for all $y \in \mathbb{R}$. That is, $y_{1i} - \Delta$ has the same distribution as y_{2i}.

Chung and Romano (2013) showed that we can permute

$$t' = \frac{\bar{y}_1 - \bar{y}_2 - \Delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

The test is exact for the case where $y_{1i} - \Delta$ has the same distribution as y_{2i} and it is asymptotically valid when $\min(n_1, n_2) \to \infty$ with n_1/n bounded away from 0 and 1, and both variances are finite.

5.7 Potential outcomes

The two sample setting can also be studied in the potential outcomes framework used in causal inference. It is also called the Neyman–Rubin model. Here it is standard to use $w \in \{0, 1\}$. In this model we have $n = n_0 + n_1$ observations y_i indexed by $i = 1, \ldots, n$. For each i, there is the value $y_i(0)$ that we would have gotten with $w_i = 0$ and a value $y_i(1)$ that we would have gotten with $w_i = 1$. These are potential outcomes. We only get to see whichever one corresponds to the actual w_i in our data.

We can represent all of the potential outcomes in a matrix $\mathcal{Y} \in \mathbb{R}^{n \times 2}$. We only get to see one value in each row, so our observed values $\mathcal{Y}_{\text{obs}}$ include

missing values as follows:

$$\mathcal{Y} = \begin{array}{c} \\ i{=}1 \\ 2 \\ 3 \\ \vdots \\ i \\ \vdots \\ n \end{array} \begin{array}{c} \begin{array}{cc} w_i{=}0 & w_i{=}1 \end{array} \\ \left[\begin{array}{cc} y_1(0) & y_1(1) \\ y_2(0) & y_2(1) \\ y_3(0) & y_3(1) \\ \vdots & \vdots \\ y_i(0) & y_i(1) \\ \vdots & \vdots \\ y_n(0) & y_n(1) \end{array}\right] \end{array} \qquad \mathcal{Y}_{\text{obs}} = \begin{array}{c} \\ i{=}1 \\ 2 \\ 3 \\ \vdots \\ i \\ \vdots \\ n \end{array} \begin{array}{c} \begin{array}{cc} w_i{=}0 & w_i{=}1 \end{array} \\ \left[\begin{array}{cc} y_1(0) & ? \\ y_2(0) & ? \\ ? & y_3(1) \\ \vdots & \vdots \\ y_i(0) & ? \\ \vdots & \vdots \\ ? & y_n(1) \end{array}\right] \end{array}$$

Writing the potential outcomes this way is making an assumption that $(y_i(0), y_i(1))$ does not depend on $w_{i'}$ for any $i' \neq i$. When that dependence happens it is called ***interference***. As an example, suppose that LinkedIn makes some experimental interventions on some of their users. User i could be affected directly by their own w_i but also indirectly by $w_{i'}$ if that affects the behavior of user i' who is linked to user i.

Defining the table $\mathcal{Y}$ requires another assumption, namely that there is only one possible value for y_i when $w_i = 0$ and only one when $w_i = 1$. That could be violated if, for example, 'treatment' means taking aspirin but the dose could be 200mg or 400mg. Those are called 'hidden versions' of the treatment. Non-interference and a lack of hidden versions taken together are known as ***SUTVA***. That stands for Stable Unit Treatment Value Assumption.

The value of y_i that we get is

$$y_i = y_i(w_i) = w_i y_i(1) + (1 - w_i) y_i(0).$$

The values in $\mathcal{Y}$ can be fixed (non-random) entries. In a randomized controlled trial, we select w_i at random and the randomness in those values of w_i drives our inferences.

The main quantities of interest are

$$\bar{y}(0) = \frac{1}{n}\sum_{i=1}^{n} y_i(0) \quad \text{and} \quad \bar{y}(1) = \frac{1}{n}\sum_{i=1}^{n} y_i(1),$$

the averages in the two columns of $\mathcal{Y}$ and we are most interested in $\Delta = \bar{y}(1) - \bar{y}(0)$ which is the average of the individual treatment effects $d_i = y_i(1) - y_i(0)$. If we randomize the w_i then we can learn something about the average treatment effect without actually seeing any of the individual treatment effects.

In the simplest randomization to study, the n_1 observations with $w_i = 1$ are a simple random sample of the n observations depicted as rows in our table $\mathcal{Y}$. That means that all $\binom{n}{n_1}$ ways of picking those observations to experiment on have equal probability, which must obviously be $1/\binom{n}{n_1}$. We can estimate Δ by

$$\begin{aligned}\hat{\Delta} = \bar{y}_1 - \bar{y}_0 &= \frac{1}{n_1}\sum_{i=1}^{n} w_i y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-w_i)y_i \\ &= \frac{1}{n_1}\sum_{i=1}^{n} w_i y_i(1) - \frac{1}{n_0}\sum_{i=1}^{n}(1-w_i)y_i(0),\end{aligned}$$

and then remembering that w_i are random while $y_i(0)$ and $y_i(1)$ are not

$$\begin{aligned}\mathbb{E}(\hat{\Delta}) &= \mathbb{E}\left(\frac{1}{n_1}\sum_{i=1}^{n} w_i y_i(1) - \frac{1}{n_0}\sum_{i=1}^{n}(1-w_i)y_i(0)\right) \\ &= \frac{1}{n_1}\sum_{i=1}^{n} \mathbb{E}(w_i) y_i(1) - \frac{1}{n_0}\sum_{i=1}^{n}\mathbb{E}(1-w_i)y_i(0).\end{aligned}$$

Now $\mathbb{E}(w_i) = \Pr(w_i = 1)$ is the probability that observation i is one of the n_1 observations selected to have $w = 1$. There are $\binom{n-1}{n_1-1}$ equally probable ways to complete that sample in a way that includes observation i. So

$$\mathbb{E}(w_i) = \frac{\binom{n-1}{n_1-1}}{\binom{n}{n_1}} = \frac{\frac{(n-1)!}{(n_1-1)!(n-n_1)!}}{\frac{n!}{n_1!(n-n_1)!}} = \frac{n_1}{n},$$

which we would almost certainly have guessed on intuitive grounds. Then $\mathbb{E}(1-w_i) = 1 - n_1/n = n_0/n$ and

$$\mathbb{E}(\hat{\Delta}) = \frac{1}{n_1}\sum_{i=1}^{n}\frac{n_1}{n} y_i(1) - \frac{1}{n_0}\sum_{i=1}^{n}\frac{n_0}{n} y_i(0) = \Delta.$$

This is a nice example of statistical thinking. By setting up the problem this way and sampling appropriately, we can get an unbiased estimate the average of the d_i without actually observing ***any*** of the individual d_i.

The usual confidence intervals for Δ are of the Welch's t form such as

$$\hat{\Delta} \pm \Phi^{-1}(1-\alpha/2)\sqrt{\frac{s_0^2}{n_0} + \frac{s_1^2}{n_1}} \tag{5.8}$$

based on a central limit theorem. Their justification is asymptotic as $\min(n_0, n_1) \to \infty$. Theorem 6.2 of Imbens and Rubin (2015) shows that

$$\mathrm{var}(\hat{\Delta}) = \frac{S_0^2}{n_0} + \frac{S_1^2}{n_1} - \frac{S_d^2}{n}$$

where $S_w^2 = \sum_{i=1}^n (y_i(w) - \bar{y}(w))^2/(n-1)$, and

$$S_d^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i(1) - y_i(0) - \Delta)^2.$$

As a result, a confidence interval like (5.8) is based on an overestimate of $\mathrm{var}(\hat{\Delta})$. We might also use a Satterthwaite based $t_{(\nu)}^{1-\alpha/2}$ cutoff. The justification would still be asymptotic. In the potential outcomes setting we don't get any values of $d_i = y_i(1) - y_i(0)$ and then we have no sample estimate of S_d^2.

Randomizing treatments as above is sometimes called the gold standard causal inference approach. The assignment cannot depend on unmeasured covariates like $\boldsymbol{x}_i$ because the randomization is done without any reference to such values, if we even have them. Findings on our n study cases may or may not generalize to subjects other than those n. That is an issue of ***external validity*** while the random assignments provide ***internal validity*** for our n study subjects.

Causal inference is a deep and specialized topic that considers what we can infer about the causal effect on y_i if we were to intervene to select w_i. The hardest problems arise when we want a causal conclusion from purely observational (w_i, y_i) pairs (plus covariates $\boldsymbol{x}_i$). That generally requires some form of causal assumption (in addition to SUTVA). We will look at some causal issues that can be addressed within our regression model, but statistical education really requires taking an entire course centered on that topic. For more on causal inference see Imbens and Rubin (2015), Ding (2024) or Wager (2024). For a less technical first read to get intuition I recommend Rosenbaum (2023).

Bibliography

P. Ding. *A First Course in Causal Inference*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, Boca Raton, FL, 2024.

G. W. Imbens and D. B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, Cambridge, 2015.

R. G. Miller. *Beyond ANOVA, Basics of Applied Statistics*. Wiley, 1986.

P. R. Rosenbaum. *Causal inference*. MIT Press, Cambridge, MA, 2023.

S. Wager. Causal inference: A statistical learning approach. `https://web.stanford.edu/~swager/causal_inf_book.pdf`, 2024.