On Kangaroos and Cookies: Models for Paired Designs

by

David A Freedman

The discussion is keyed to the exercise on pp258–9 of FPP,[†] where pairs of kangaroos are timed as they run a maze. To see if vitamins help, one animal in the pair is chosen at random for treatment. (See also exercise 11, pp262-63, on the health effects of smoking, and exercise 11, p489, on introductory pricing for cookies.)

Experimental units are indexed by $i = 1, \ldots, n$. Each unit has a response a_i^T if in treatment, and a_i^C if in control. The "strict" null hypothesis holds that $a_i^T = a_i^C$ for all i; the "weak" null, that $\frac{1}{n}\sum_{i=1}^n a_i^T = \frac{1}{n}\sum_{i=1}^n a_i^C$. If the null is rejected, there will be some interest in estimating the average causal effect over the study population, $\big(\frac{1}{n}\sum_{i=1}^n a_i^T\big) - \big(\frac{1}{n}\sum_{i=1}^n a_i^C\big)$. Results are presented here to suggest:

(i) The sign test is appropriate for the strict null; the weak null is better handled via a t-test on paired differences, if pairing is to be done at all.

(ii) Pairing can help, or hurt. (In this respect, randomized experiments differ from stratified random random sampling.)

On the strict null, $a_i^T = a_i^C = a_i$. Suppose $n = 2m$ and we pair unit $i = 1, \ldots, m$ with unit $m + i$. The difference between the treatment unit and conrol unit is either $a_i - a_{m+i} = d_i$ or $a_{m+i} - a_i = -d_i$, at random. So, the sign test is in order. Turn now to the t-test.

Example 1. Suppose $a_i^T = 4$ and $a_i^C = 3$ for $i = 1, \ldots, m$, whilst $a_i^T = 2$ and $a_i^C = 1$ for $i = m + 1, \ldots, 2m$. Recall that i is paired with $m + i$ for $i = m + 1, \ldots, 2m$. Each pair of units can be represented graphically, as follows:

$$4{:}3 \leftrightarrow 2{:}1$$

For each pair, we see $4 - 1 = +3$ or $2 - 3 = -1$, at random. The sign test will indicate neutrality, although treatment helps every unit. The t-test gives the right answer here: the average causal effect would be estimated as $(3 - 1)/2 = +1$, with variances to be discussed below.

Example 2. Suppose $a_i^T = 4$ and $a_i^C = 1$ for $i = 1, \ldots, m$, whilst $a_i^T = 2$ and $a_i^C = 3$ for $i = m + 1, \ldots, 2m$. The pairing is

$$4{:}1 \leftrightarrow 2{:}3$$

For each pair, we see $4 - 3 = +1$ or $2 - 1 = +1$, at random: i.e., we always see $+1$. The sign test will indicate that treatment is uniformly successful, although treatment helps only half the units. Again, the t-test gives the right answer. The average causal effect is estimated as $+1$, with no variance.

We focus on Example 1, and consider three designs:

(A) Paired trials, as discussed above. The estimator is the average of the m sample differences.

[†] David Freedman, Robert Pisani, Roger Purves (2007). *Statistics*. 4th ed, Norton, New York.

(B) Choose m units at random without replacement for treatment, the remaining units being controls. The estimator is the sample average of the treatment units minus the sample average of the control units.

(C) Flip a coin for each unit; if the coin lands 1, put the unit in treatment; if 0, in control. With this design, we consider two estimators:

(C1) Take the sample sum of the treated units minus the sample sum of the control units, and divide by m.

(C2) Take the sample average of the treated units minus the sample average of the control units.

The estimators are all unbiased, with variances shown below; results for (C2) are asymptotic, denominators being random. The paired estimator (A) is inefficient. The inefficiency may be explained on the grounds that the pairing in Example 1 is perverse, since like and unlike are paired. On the other hand, the equally-perverse pairing in Example 2 gives an estimator whose efficiency is nonpareil. The estimator (C1) is very inefficient, since it does not adjust for difference in size between treatment group and control group. However, (C2) is about as good as the best, namely, (B)—which is based on a design with no pairing.

(A) $\dfrac{4}{m}$

(B) $\dfrac{2}{m}\Big(1 - \dfrac{1}{2m-1}\Big)$

(C1) $\dfrac{58}{4}\dfrac{1}{m}$

(C2) $\dfrac{2}{m}$

Proof of (A). Recall that the pairing is

$$4{:}3 \leftrightarrow 2{:}1$$

Let $X_i = 1$ or 0 with a 50–50 chance, independently for $i = 1, \ldots, m$. If $X_i = 1$, the difference between the treatment and control units is $4 - 1 = +3$. If $X_i = 0$, the difference between the treatment and control units is $2 - 3 = -1$. The difference between the treatment and control units in the ith pair is therefore $3X_i - (1 - X_i) = 4X_i - 1$. Our estimator is

$$\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m}(4X_i - 1)$$

whose expectation is 1 and whose variance is $4/m$, as required.

Proof of (B). With this design, there is no pairing, Consider a finite population consisting of m 1's and m 0's. We implement random assignment by taking a random permutation of the 1's and 0's, with 1 standing for assignment to treatment and 0 for assignment to control. Let X be the number of 1's among the first m draws, i.e., the number of units $i = 1, \ldots, m$ assigned to treatment. The number of such units assigned to control is $m - X$. The corresponding numbers for units $i = m+1, \ldots, 2m$ are $m - X$ and X, respectively. The sample sum of responses for treatment units is $4X + 2(m - X) = 2X + 2m$. The sample sum of responses for control units is $3(m - X) + X = 3m - 2X$. The difference is $4X - m$, and our estimator is

$$\hat{\theta} = (4X - m)/m.$$

The expected value is 1. The variance is as required, because X is hypergeometric with

$$\mathrm{var}(X) = \frac{m-1}{2m-1}\frac{m}{4} = \Big(1 - \frac{1}{2m-1}\Big)\frac{m}{8}$$

Proof of (C1). Let $X_i = 1$ or 0 with a 50–50 chance, independently for $i = 1, \ldots, 2m$. Let $X = \sum_{i=1}^{m} X_i$, the number of treatment units among $i = 1, \ldots, m$. Let $Y = \sum_{i=m+1}^{2m} X_i$, the number of treatment units among $i = m+1, \ldots, 2m$. Now X and Y are independent $\mathrm{Bin}(n, 1/2)$. The sample sum of treatment responses is $4X + 2Y$. The sample sum of control responses is $3(m - X) + (m - Y)$. The difference is $(7X - 3m) + (3Y - m) = 7X + 3Y - 4m$. Our estimator is

$$\hat{\theta} = (7X + 3Y - 4m)/m$$

The expectation is 1 and the variance is $58/(4m)$, as required. This estimator is particularly inefficient, because it does not adjust for the difference in sizes between treatment and control groups. The next estimator fixes that problem; the calculation is asymptotic, because group sizes are random.

Proof of (C2). We use the same notation as in (C1). Our estimator is

$$\hat{\theta} = \hat{\theta}_1 - \hat{\theta}_2$$

where

$$\hat{\theta}_1 = \frac{4X + 2Y}{X + Y}, \quad \hat{\theta}_2 = \frac{3(m - X) + (m - Y)}{(m - X) + (m - Y)}$$

We use the delta method. Dependence on m in suppressed in the notation: $X = m/2 + \sqrt{m/4}\,\xi$, where $\xi \to N(0, 1)$ as $m \to \infty$. Likewise, $Y = m/2 + \sqrt{m/4}\,\eta$, where $\eta \to N(0, 1)$ as $m \to \infty$. The variables ξ and η are independent. The numerator of $\hat{\theta}_1$ is $3m + \sqrt{m/4}\,(4\xi + 2\eta)$. The denominator is $m + \sqrt{m/4}\,(\xi + \eta) = m\big(1 + \sqrt{1/(4m)}\,(\xi + \eta)\big)$. Since $1/(1 + \delta) = 1 - \delta + O(\delta^2)$ for small δ—and $(\xi + \eta)/\sqrt{m}$ is small—the inverse of the denominator is

$$\frac{1}{m}\left(1 - \frac{1}{2}\sqrt{\frac{1}{m}}(\xi + \eta) + O\Big(\frac{1}{m}\Big)\right)$$

Thus,

$$\begin{aligned}\hat{\theta}_1 &= \left(3 + \frac{1}{2}\sqrt{\frac{1}{m}}(4\xi + 2\eta)\right)\left(1 - \frac{1}{2}\sqrt{\frac{1}{m}}(\xi + \eta) + O\Big(\frac{1}{m}\Big)\right) \\ &= 3 + \frac{1}{2}\sqrt{\frac{1}{m}}(4\xi + 2\eta - 3\xi - 3\eta) + O\Big(\frac{1}{m}\Big) \\ &= 3 + \frac{1}{2}\sqrt{\frac{1}{m}}(\xi - \eta) + O\Big(\frac{1}{m}\Big)\end{aligned}$$

Similarly,

$$\hat{\theta}_2 = 2 + \frac{1}{2}\sqrt{\frac{1}{m}}(\eta - \xi) + O\Big(\frac{1}{m}\Big)$$

So,

$$\hat{\theta}_1 - \hat{\theta}_2 = 1 + \sqrt{\frac{1}{m}}(\xi - \zeta) + O\Big(\frac{1}{m}\Big)$$

as required.

Why are randomized experiments different from stratified random samples?

Consider Example 1 from the sampling perspective. Each of the m pairs seems to be a stratum of size two; we choose at random one of two numbers, $+3$ or -1. The usual comparison to a simple random sample would involve drawing m times at random without replacement from a population consisting of m tickets marked $+3$ and m tickets marked -1. That, however, is not a good model for an experiment where m units are chosen at random for treatment, leaving the other m units for controls. In the experiment, $2m$ responses are observed. The treatment group provides m responses, about half being $+4$ and the others $+2$. Similarly, the control group provides m responses, about half being $+3$ and the others $+1$. The usual comparison between stratified and unstratified sampling designs is not directly relevant when comparing paired and unpaired experimental designs.

Nominal variances give the right answers in example 1

(A) Our sample comprises m observations, the ith one being $4(X_i - 1)$, where the X_i are independent coin tosses. Plainly, the sample variance converges to $16/4 = 4$, so $\widehat{\text{var}}(\hat{\theta}) \approx 4/m$. A similar argument works for (C1).

(B) The usual estimator $\widehat{\text{var}}(\hat{\theta})$ is $\hat{v}/m$, where $\hat{v} = \hat{v}_1 + \hat{v}_2$, with $\hat{v}_1$ being the sample variance of the treatment units and $\hat{v}_2$ being the sample variance of the control units. We use previous notation. The treatment sample comprises X responses that are $+4$ and $m - X$ responses that are $+2$, where X is hypergeometric. Now

$$\hat{v}_1 = (4-2)^2 \Big(\frac{X}{m}\Big)\Big(\frac{m-X}{m}\Big) \to 1$$

Likewise, $\hat{v}_2 \to 1$. Thus, $\hat{v} \to 2$. A similar argument works for (C2).[‡]

Notes for Statistics 215
Department of Statistics
UC Berkeley, CA 94720-3860
July, 2008

[‡] But see note 12 to chapter 27, ppA32–33 in FPP. Example 1 is special, because the (unobservable) correlation between responses across units equals 1.