Nordhausen University
of Applied Sciences

Randomization does not help much, comparability does

Uwe Saint-Mont

Hochschule Nordhausen

Abstract

Following Fisher (1935/71), it is widely believed that randomization "relieves the experimenter from the anxiety of considering innumerable causes by which the data may be disturbed." In particular, it is said to control for known and unknown nuisance factors that may challenge considerably the validity of a result. Looking for quantitative advice, I studied a number of straightforward, mathematically simple models. (Only one of them is presented here.)

However, they all demonstrate that the optimism with respect to randomization is wishful thinking rather than based on fact. In small to medium-sized samples, **random allocation of units to treatments typically *yields* a considerable imbalance between the groups**, i.e., confounding due to randomization is the rule rather than the exception. For this, and further reasons, classical experimentation based on sound background theory and the systematic construction of comparable groups, seems to be preferable.

Savage's Example

Our conclusion, based on an explicit quantitative analysis, coincides with the qualitative argument given by Savage (1962):

> Suppose we had, say, thirty fur-bearing animals of which some were junior and some senior, some black and some brown, some fat and some thin, some of one variety and some of another, some born wild and some in captivity, some sluggish and some energetic, and some long-haired and some short-haired.
>
> It might be hard to base a convincing assay of a pelt-conditioning vitamin on an experiment with these animals, for every subset of fifteen might well contain nearly all of the animals from one side or another of one of the important dichotomies [. . .]
>
> **Thus contrary to what I think I was taught, and certainly used to believe, it does not seem possible to base a meaningful experiment on a small heterogenous group.**

The Logic of the Experiment

If one compares two groups of subjects (Treatment T versus Control C, say) and observes a substantial difference in the end (e.g. $\bar{Y}_T > \bar{Y}_C$), that difference must be due to the experimental manipulation - IF the groups were equivalent at the very beginning of the experiment:

Start of Experiment	T	$=$	C	T	$\neq$	C
Intervention	Yes		No	Yes		No
End of Experiment						
(Observed Effect)	$\bar{Y}_T$	$>$	$\bar{Y}_C$	$\bar{Y}_T$	$>$	$\bar{Y}_C$
Conclusion	Intervention caused the effect			Intervention OR prior difference caused the effect		

Worrall (2007): "It is entirely possible that **any particular randomization may have produced a division** into experimental and control groups **that is unbalanced** with respect to 'unknown' factor X[. . .]"

Random Confounding (Binary Model)

Suppose there is a nuisance factor X taking the value 1 if present and 0 if absent. One may think of X as a genetic aberration, a psychological disposition or a social habit. Assume that the factor occurs with probability p in a certain person (independent of anything else). Given this, $2n$ persons are randomized into two groups of equal size by a chance mechanism independent of X.

Let S_1 and S_2 count the number of persons with the trait in the first and the second group respectively. S_1 and S_2 are independent random variables, each having a binomial distribution with parameters n and p. A natural way to measure the extent of imbalance between the groups is $D = S_1 - S_2$. Obviously, $ED = 0$ and

$$\sigma^2(D) = \sigma^2(S_1) + \sigma^2(-S_2) = 2\sigma^2(S_1) = 2np(1-p).$$

Iff $D = 0$, the two groups are perfectly balanced with respect to factor X. In the worst case $|D| = n$, that is, in one group all units possess the characteristic, whereas it is completely absent in the other. For fixed n, let the two groups be comparable if $|D| \leq n/i$ with some $i \in \{1, \ldots, n\}$. Iff $i = 1$, the groups will always be considered comparable. However, the larger i, the smaller the number of cases we classify as comparable. In general, n/i defines a proportion of the range of $|D|$ that seems to be acceptable. Since n/i is a positive number, and $S_1 = S_2 \Leftrightarrow |D| = 0$, the set of comparable groups is never empty.

Given some constant $i (< n)$, the value n/i grows at a linear rate in n, whereas $\sigma(D) = \sqrt{2np(1-p)}$ grows much more slowly. Due to continuity, there is a single point $n(i,k)$, where the line intersects with k times the standard deviation of D. Beyond this point, i.e. for all $n \geq n(i,k)$, at least as many realizations of $|D|$ will be within the acceptable range $[0, n/i]$. Straightforward algebra gives,

$$n_p(i,k) = 2p(1-p)i^2k^2.$$

A typical choice could be $i = 10$ and $k = 3$, which specifies the requirement that most samples be located within a rather tight acceptable range. Relaxing the criterion of comparability (i.e. a smaller value of i) decreases the number of subjects necessary. The same happens if one decreases the number of standard deviations k. Depending on p, the following numbers of subjects are needed per group (and twice this number altogether):

p	i	k	$n_p(i,k)$	i	k	$n_p(i,k)$	i	k	$n_p(i,k)$
1/2	10	3	450	5	3	113	10	2	200
1/5	10	3	288	5	3	72	10	2	128
1/10	10	3	162	5	3	41	10	2	72
1/100	10	3	18	5	3	5	10	2	8

This shows that **randomization works, if the number of subjects ranges in the hundreds** or if the probability p is rather low. (By symmetry, the same conclusion holds if p is close to one.) Otherwise there is hardly any guarantee that the two groups will be comparable. Rather, they will differ considerably due to random fluctuations.

Probability (T and C comparable)

The distribution of D is well known, thus it is possible to compute the probability $q = q(i,n,p)$ that two groups, constructed by randomization, will be comparable. If $i = 5$, i.e., if one fifth of the range of $|D|$ is judged to be comparable, we obtain:

p	n	$q(i,n,p)$	p	n	$q(i,n,p)$	p	n	$q(i,n,p)$
1/2	5	0.66	1/10	5	0.898	1/100	5	0.998
1/2	10	0.74	1/10	10	0.94	1/100	10	0.9997
1/2	25	0.88	1/10	25	0.98999	1/100	25	0.999999
1/2	50	0.96	1/10	50	0.999	1/100	50	1

Thus, it is **rather difficult to control a factor** that has a probability of about 1/2 in the population. However, even if the probability of occurrence is only about 1/10, one needs more than 25 people per group to have reasonable confidence that this nuisance factor has not produced a substantial imbalance.

Several factors

The situation becomes worse if one takes more than one nuisance factor into account. Given m independent binary factors, each of them occurring with probability p, the probability that the groups will be balanced with respect to all nuisance variables is q^m. Numerically, the above results yield:

p	n	q	q^2	q^5	q^{10}
1/2	5	0.66	0.43	0.12	0.015
1/2	10	0.74	0.54	0.217	0.047
1/2	25	0.88	0.78	0.53	0.28
1/2	50	0.96	0.93	0.84	0.699

p	n	q	q^2	q^5	q^{10}
1/10	5	0.898	0.807	0.58	0.34
1/10	10	0.94	0.88	0.74	0.54
1/10	25	0.98999	0.98	0.95	0.90
1/10	50	0.999	0.9989	0.997	0.995

Accordingly, given m independent binary factors, each occurring with probability p_j (and corresponding $q_j = q(i,n,p_j)$), the probabilities closest to 1/2 will dominate $1 - q_1 \cdots q_m$ which is the probability that the two groups are not comparable due to an imbalance in at least one variable. In a typical study with $2n = 100$ persons, for example, it does not matter if there are one, two, five or even ten factors, if each of them occurs with probability of 1/100. However, if some of the factors are rather common (e.g. $1/5 < p_j < 4/5$), this changes considerably. In a smaller study with fewer than $2n = 50$ participants, a few such factors suffice to increase the probability that the groups constructed by randomization won't be comparable to 50%. **With a few units per group, one can be reasonably sure that some undetected, but rather common, nuisance factor(s) will make the groups non-comparable** which is the crux of Savage's example.

Interactions

The situation deteriorates considerably if there are interactions between the variables that may yield convincing alternative explanations for an observed effect. It is possible that all factors considered in isolation are reasonably balanced (which is often checked in practice), but that a certain combination of them affects the observed treatment effect. Given m factors, there are $m(m-1)/2$ possible interactions between just two of the factors, and $\binom{m}{\nu}$ possible interactions between ν of them. Thus, **there is a high probability that some considerable imbalance occurs in at least one of these numerous interactions**, in small groups in particular. Detected or undetected, such imbalances provide excellent alternative explanations of an observed effect.

Conclusions

Deliberately, the above model has been kept as simple as possible. Therefore, its results are straightforward and they agree with other natural models: If n is small, it is almost impossible to control for a trait that occurs frequently at the individual level or for a larger number of confounders via randomization. It is of paramount importance to understand that **random fluctuations lead to considerable differences between** small or medium-sized **groups, making them very often non-comparable, thus undermining the basic logic of experimentation.** That is, 'blind' randomization does not create equivalent groups, but rather *provokes* imbalances and thus artifacts: Even in larger samples one needs considerable luck to succeed in creating equivalent groups (p close to 0 or 1, a small number of nuisance factors m or a favourable dependence structure that balances all factors, including their relevant interactions, if only some crucial factors are to be balanced by chance).

Therefore, it seems much more advisable to use background knowledge in order to minimize the difference between groups with respect to known factors or specific threats to experimental validity. At the end of such a conscious construction process, randomization finds its proper place. Only if no reliable context information exists, is unrestricted randomization the method of choice. It must be clear, however, that it is a weak guard against confounding, yet the only one available in such inconvenient situations.

In a nutshell, **the above analysis strongly recommends traditional experimentation, thoroughly selecting, balancing and controlling factors and subjects with respect to known relevant variables,** thereby using broader context information. For more details see Saint-Mont (2015).

References

[1] Fisher, R.A. (1935/71), *The Design of Experiments*, Macmillan.

[2] Saint-Mont, U. (2015), "Randomization does not help much, comparability does", PLoS ONE 10(7), doi:10.1371/journal.pone.0132102.

[3] Savage, L.J. in Cox, D.R.; and Barnard, G.A. (eds., 1962), *"The foundations of statistical inference. A discussion"*, Methuen, London.

[4] Worrall, J. (2007). "Why There's No Cause to Randomize", Brit. J. Phil. Sci. 58, 451-488.