Randomization

Gary W. Oehlert

School of Statistics
University of Minnesota

January 17, 2016

Randomization

A randomization is a random method to assign treatments to units.

Or vice versa.

Haphazard is not random.

You need a clearly defined and understood probabilistic scheme. It does not need to be a complicated scheme.

Why randomize?

Randomization protects agains confounding.

Randomization can be a basis for inference.

Simple example

Treatments: three nano-tech coatings plus control (no coating).

Units: 20 swatches of fabric.

Randomization: randomly divide the 20 units into four groups of five units each. Give treatment 1 to group 1, treatment 2 to group 2, and so on.

The well-understood probabilistic model is equivalent to drawing cards from a deck of 20 cards (without replacement).

Randomization versus confounding

On average, randomization balances assignments of treatments to units.

- Each treatment gets approximately the same number of high-yielding units.
- Each treatment gets approximately the same number of low-yielding units.
- Each treatment gets approximately the same number of odd ball units.
- Each treatment gets approximately the same number of standard units.

And there's more!

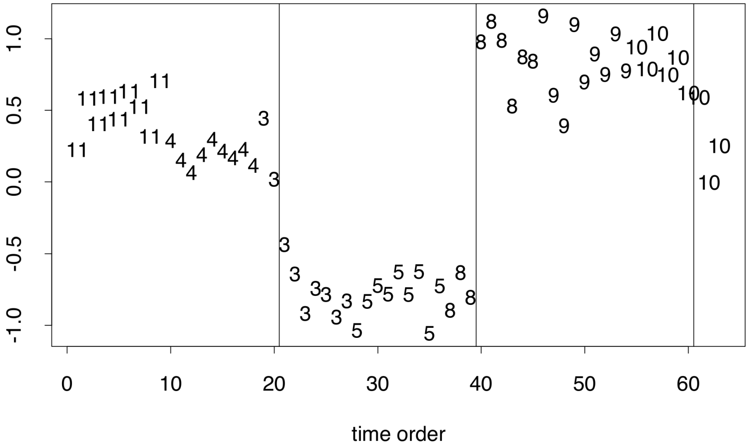
We don't even need to know what we should balance for; randomization does it (approximately) for us.

The deviation implied in the "approximately equal" assignments follows an understood probability mechanism that we can account for.

Without randomization, deviation from balancing can lead to confounding: the units are different in some unknown way, and we cannot account for it.

Lack of randomization can cause big trouble.

Cd in time order



Nitty gritty

To randomly assign N units into groups of size $n_1, n_2, \ldots, n_g$ with $n_1 + n_2 + \cdots + n_g = N$, first put the units in random order. Take first n_1 of the randomly ordered units for group 1, and so on.

To get a random order for, say, $N = 12$ units, give the R command

```
sample.int(12)
[1]  9  8  3  6  1 10  5  7  4  2 12 11
```

This will give you a random ordering of the case (unit) numbers. Then make your split.

There are other, fancier, sampling functions in R.

Randomization for inference

Ordinary data analysis (with normal distributions):

- Data are random samples from some distributions.
- Inference is about parameters (usually means) of the distributions.
- Sampling induces a distribution on the test statistic (under the null).
- P-value says something about how far the observed test statistic is into the tails of the distribution.

Groups are known, data are random samples.

Randomization testing turns that on its head:

- The null is that the treatment groupings are inert (just labels) and do not affect the responses. The data are the data; they are considered fixed.
- The only thing random is the assignment of the data to groups.
- Randomization induces a null distribution on a test statistic.
- P-value says something about how far the observed test statistic is into the tails of the distribution.

Data are known (constant); groups are random.

OK, why?

We will see that

1. Computing the randomization null can be a real pain in the neck.
2. The results of randomization tests and standard tests are often effectively identical.

Where is the value in randomization tests?

If you did the randomization, then the randomization test is valid.

No debates about assumptions, residuals, and so on. It just works. That can be important in legal settings.

Two-sample setting

Data $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_m$. We are testing for equal means. The (pooled) two-sample t-test is

$$t = \frac{\bar{x} - \bar{y}}{s_p \sqrt{1/m + 1/n}}$$

with

$$s_p^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{j=1}^{m} (y_j - \bar{y})^2}{n + m - 2}$$

Null distribution is t with n+m-2 df. This assumes normality, independence, and equal variances. (There is also an unpooled version of the test, but the same points apply.)

With the same data, the randomization test statistic is

$$t = \bar{x} - \bar{y}$$

The null distribution is found by computing the difference of means for all possible assignments of N units into groups of size m and n (i.e., all possible randomizations). The only assumption is the randomization.

These outcomes are equally likely. The p-value is fraction as extreme or more extreme than statistic in the data.

${}_NC_n = \frac{N!}{n!(N-n)!}$ grows very quickly with N, making an exact computation cumbersome.
${}_{10}C_5 = 252$ ${}_{20}C_{10} = 184,756$ ${}_{30}C_{15} = 15,511,752$. Don't want to do it by hand!

Paired setting

In paired data we have x_i and $y_i, i = 1, 2, \ldots, n$ measured on the same unit or units that are similar in some way. Inference is on the differences $d_i = x_i - y_i$.

The paired t-test is

$$t = \frac{\bar{d}}{s/\sqrt{n}}$$

with

$$s^2 = \sum_{i=1}^{n} (d_i - \bar{d})^2/(n-1)$$

Null distribution is t with n-1 df. This assumes normality, independence, and constant variance.

With the same data, the randomization test statistic is just $\bar{d}$.

Under the randomization, the two units in the pair either received treatments A and B (in that order), or B and A (in that order). Under the randomization null hypothesis, the sign of each difference plus or minus with probability one half (independently).

The null distribution is found by looking at $\bar{d}$ for all 2^n possible outcomes for the n different signs.

These outcomes are equally likely. The p-value is fraction as extreme or more extreme than statistic in the data.

Examples in R