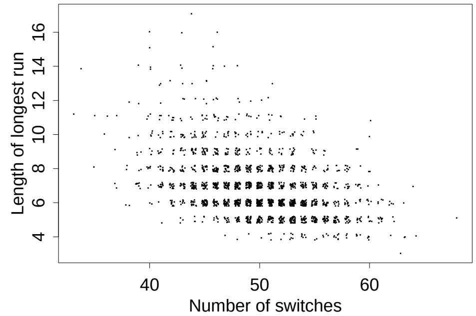

Figure B.2 *Length of longest run (sequence of successive heads or successive tails) versus number of runs (sequences of heads or tails) in each of 2000 independent simulations of 100 coin flips. Each dot on the graph represents a sequence of 100 coin flips; the points are jittered so they do not overlap. When plotted on this graph, the results from an actual sequence of 100 coin flips will most likely fall on a square with a large number of dots. In contrast, a sequence of heads and tails that is artificially created to look "random" will probably have too many runs that are not long enough, and hence will fall on the lower right of this graph.*

a regression on random numbers. An example that is not *necessarily* bad is using, as the x variable, the order of entry of units into the study. This can make sense if one expects or fears time trends (but it would probably be better to plot versus time itself rather than merely order). If there are no major time patterns, however, the choice of x variable might better be spent elsewhere.

You can make as many plots as you want (or as your paper budget allows), but it is useful to think a bit about each plot, just as it is useful to think a bit about each regression you run. This is as good a time as any to recommend that along with every regression you run, you should make a scatterplot. And, in addition, you should be making residual plots where necessary. We'll get to that later.

Jittering

If several data points have the same data values, add a small random number to each so that they do not fall on top of each other. This is called jittering. Jitter just enough so that the discrete nature of the data is still clear. For example, if data points are integers, we might add a random uniform number between -0.3 and $+0.3$ to each x and y value (see Figure B.2). Methods such as plotting 2's, 3's, or cute symbols for multiple data points can be misleading visually, and from a theoretical perspective are unsatisfying in that the display of any unit then depends too strongly on the other data values.

Symbols and auxiliary lines

The symbols of a scatterplot are important because they correspond to the units of analysis in your studies. It can be appropriate to use more than one scatterplot for multilevel data structures. At least in theory you can display five variables easily with a scatterplot: x, y, symbol, symbol size, and symbol color.

Symbols are best for discrete variables, and it's worth putting a little effort into making these symbols distinguishable and also appropriate. For example, we used open circles to indicate open seats in Figure 7.4. In plotting data from an experiment or observational study, you can use different large symbols for treated units and