Aaron Brown

Correlation Bingo

What is the correlation between two randomly chosen Bingo cards in the same game? The answer is not obvious. It turns out the question is not obvious either

Of all the elementary concepts taught in introductory statistics, none is as dangerous as correlation. Correlation is an unreliable friend outside a textbook, yet it forms the basis of a large part of statistical reasoning.

I do not deny that you can define correlation for real measurements, but it's a misleading statistic. Real things are related in complicated ways. Some of those ways will cause the things to move in the same direction, but to different degrees, and some will cause the things to move in the opposite direction. For example, two companies in the same industry will be affected in opposite ways by competitive events (such as one company introducing a successful new product that reduces sales by the other) but in the same way by most macro events (such as an increase in economic activity). The fact that two things have moved in a certain relation in the past is very weak grounds for projecting the same relation in the future, because in the future different types of causes will have different relative importances.

Sometimes people describe this as correlations being "unstable" or, perhaps, "hard to estimate." I think the problem is different. In most cases of interest, correlation as a concept does not exist.

There is no general tendency for two things to move in a particular relation, just an average of many tendencies observed from the past.

Crises

For example, you often hear the faux wisdom: "Correlations go to one in a crisis." If you try to make this quantitatively precise and test it, you will find it has little predictive value. When you point out obvious counterexamples, people respond with: "Well, correlations go to one or negative one." That allows them to explain everything that went up and everything that went down. What about the things that didn't move? Well, if there are a lot of those, then it isn't a crisis. The correct observation is: "Sometimes lots of things move a lot, some in unexpected directions given the other things. We call that a 'crisis.'"

A general observation is when people have well-worn catch phrases for explaining the breakdowns of their models, don't trust their models.

When the degree of association is high and the dimensionality of the problem is low and we are concerned only with the center of the distribution, regression and correlation are sometimes appropriate tools. But outside of this comfortable zone, a zone in which any reasonable methods work pretty well, the tools are more likely to mislead than to inform.

Correlation is particularly toxic when mixed with the multivariate Normal distribution. That distribution allows you to make strong predictions about large numbers of variables, based only on pairwise information. That is, if you know how A is related to B to Z, how B is related to A and C through Z, and so on up to how Z is related to A to Y, you can make strong predictions about the entire alphabet. In practical inference, knowing pairwise relations is of small use in predicting the joint behavior of three or more variables.

A simple example I use to illustrate that is 11 bonds, each with a 10 percent chance of default, all uncorrelated. The fact that two bonds are uncorrelated means only that their joint probability of default is the product of their individual probabilities, 10% x 10% = 1%.

Imagine we put 100 slips of paper into a hat, and draw out one to find which bonds default. We have

to put each bond on exactly ten slips of paper, so there is a 10 percent chance of a default. We have to put every pair of bonds on exactly one slip of paper so there is a 1 percent chance of joint default. But that's all we know. There are lots of ways to fill out the slips to meet these conditions. For example, we could put every pair of bonds on one slip of paper (55 in all) and leave 45 blank. That means there is a 55 percent chance of two defaults and a 45 percent chance of no defaults. Another choice is to write all ten bonds on one slip of paper, and each bond alone on nine slips. In that case, there is a 99 percent chance of one default and a 1 percent chance of ten defaults. These two situations are clearly quite different, but they are indistinguishable if you only know the individual default probabilities and all 55 pairwise correlations.

Bingo

Over the years, I have found that the game of Bingo is a more useful mental model than correlation for thinking about association among variables. Bingo is one of the most popular gambling games in the world since its evolution from similar games in the 1920s. An order of magnitude more people play it than visit casinos each year. I will describe the most common set of rules, but there are many variants.

Bingo is played with five-by-five cards that have numbers from 1 to 75 randomly assigned to the cells (not completely randomly, as will be explained). The numbers from 1 to 75 are called out in random order. The player marks any cell containing a called number. When she has five cells in a row marked, she calls out "bingo," and if she is the first to do so, she wins the prize. The five cells can be horizontal, vertical, or on either of the two main diagonals (12 ways in all).

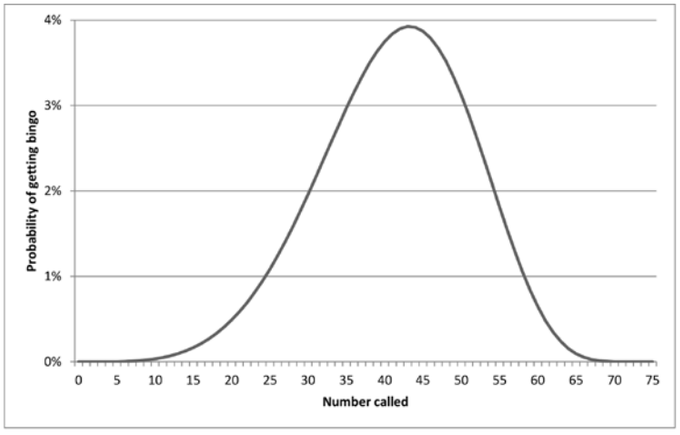
The center cell on all cards is marked "free" and is considered marked at the start of play. So four of the 12 possible bingos, both diagonals plus the middle horizontal and vertical ones, require only four numbers to complete, not five. The other non-randomness in number assignment is that the first column (the "B") gets only numbers from 1 to 15, the second ("I") column gets 16 to 30, and so on. The same number is never repeated on a card.

It takes a minimum of four marked squares to get bingo, and there are only four ways to do this out of the C(24,4) = 10,626 possible arrangements of four cells. The most cells that can be marked without getting bingo is 19, and there are only 24 arrangements that avoid bingo out of the C(24,19) = 42,504 possible arrangements of 19 cells. Table 1 shows the number of ways to get bingo for cards with between 4 and 19 cells marked.

Table 1: Bingo combinatorics: ways to get Bingo with from 4 to 19 markers on your card

Number of hits	Ways to get Bingo	Number of hits	Ways to get Bingo	Number of hits	Ways to get Bingo	Number of hits	Ways to get Bingo
4	4	8	27,102	12	841,100	16	644,445
5	88	9	92,520	13	1,113,360	17	331,056
6	912	10	244,092	14	1,174,620	18	133,428
7	5,928	11	507,696	15	981,424	19	42,480

Figure 1: Bingo probabilities: fraction of cards that will generate Bingo as the numbers from first to 75th are called



To compute the probability of getting bingo after k numbers are called, you have to sum terms from j = 4 to k, where j is the number of hits on the card, taking the number corresponding to j from table 1 and multiplying by $C(k,j)*2\text{-}k/C(24,j)$. Figure 1 shows the difference between those numbers, that is, the probability of getting your first bingo on a single card versus the number of the number called. There is zero probability of getting bingo on any of the first three numbers and less than one chance in 300,000 of getting it on the fourth number. There is a 0.03 percent chance of getting your first bingo on the 10th number, a 0.5 percent chance on the 20th, a 2 percent chance on the 30th, a 3.8 percent chance on the 40th, and a peak of a 3.92 percent chance of getting your first bingo on the 43rd number called. There is zero probability of getting your first bingo on the 72nd, 73rd, 74th, or 75th number called.

Bingo correlation

What is the correlation between two randomly chosen Bingo cards in the same game? The answer is not obvious. It turns out the question is not obvious either. There are a lot of different ways to define correlation, and one set of problems occurs when people use different definitions but assume the correlations are numerically the same. Some possible ways to define correlation are with respect to:

- The probability of the two cards both winning the game
- The change in the probabilities of the two cards winning the game when a number is called
- The probability of two cards getting bingo on the same call
- The number of calls before each card gets a bingo.

All four of these definitions are in use in credit modeling, and all are referred to as "credit correlation," although they are numerically and conceptually distinct.

Another ambiguity is whether you are talking about the unconditional correlation for two cards drawn at random with uniform probability of the

set of legal bingo cards or the correlation conditional on the numbers and layout of two specific cards, or some other specification.

Assuming we agree on definitions, there are still some non-intuitive facts about correlation. For example, consider that the chance is about 1 in 300,000 of a Bingo card winning on the first four numbers called. You might think that means there is one chance in $300{,}000^2 = 90$ billion of two randomly chosen cards both getting bingo on the first four numbers. That isn't true due to the restriction on the column that different numbers can be in (such as 1 to 15 having to be in the first column).

To get a bingo in four numbers, you need all of them to be in the middle column or else for exactly one of them in each of the first, second, fourth, and fifth columns. Only one set of four numbers in 25 qualifies. So, 96 percent of the time no card can possibly have a bingo in the first four numbers called, and 4 percent of the time the chance is 1 in 12,000, not one in 300,000. That means the chance of two specific cards getting bingo on the first four numbers is 1 in $25 \times 12{,}000^2 = 3.6$ billion, not 90 billion.

After ten numbers have been called, there is a 0.08 percent chance of getting a bingo with one card. But, again, the results are not independent. Twenty-three percent of the time, no card can have a bingo because:

1. No column has five numbers and the middle column does not have four, and
2. At least one of columns one, two, four, or five has no numbers in it.

At the other extreme, there is a 20 percent chance that you will have two to four times the unconditional probability of getting a bingo. The one thing that never happens is that you get a set of ten numbers such that the conditional probability of bingo is equal to the unconditional.

So far, it may be surprising that a seemingly small regularity in Bingo card construction causes such bunching of winners. The counterintuitive aspect arises when you observe only the bingos, not the numbers called. If you believe in correlation, you are apt to think that historical measured correlation predicts future correlation. But the opposite is true in Bingo. If you observe low correlation, it is likely because the called numbers are unfavorable for generating bingos. In that case, a single number – say, the first number called in column 1 – is likely to cause a large number of bingos, much larger than would ever be plausible under the historical observed correlation. On the other hand, if you observe a high correlation, it generally means the cards close to bingo have been largely eliminated, and going forward you expect low correlation of outcome.

All four of these definitions are in use in credit modeling, and all are referred to as "credit correlation," although they are numerically and conceptually distinct

Shared numbers

Another aspect of bingo correlation mathematics arises when we consider the conditional correlation of winning the game based on looking at the numbers on the cards. Two cards that share no numbers cannot both win the game, since a call that completes a bingo for one card cannot also complete a bingo for the other. Two randomly-chosen cards have a 2.8 percent chance of getting their first bingos on the same number. Two cards that share all 24 numbers, but in randomly different cells, have an 11 percent chance of getting their first bingos on the same number.

That is obvious; the more numbers that two cards share, the more chance of them both winning a bingo game. The cells that the matched numbers are in also matter. The eight cells on the two main diagonals are each part of three potential bingos, one of which is a four-number bingo. The eight cells (not counting the free cell) in the middle row and column are part of two potential bingos each, one of which is a four-number bingo. The remaining eight cells are also part of two potential bingos, but both are five-number bingos. The more of the overlapping numbers that are in the more favorable bingo cells, the higher the correlation. There are also higher-order effects; for example, if a set of shared numbers forms bingos on both cards, the chance of a joint win increases quite a bit.

Two randomly chosen cards share an average of 7.73 numbers (it would be 7.68 if the numbers were assigned randomly, but the column rules increase the overlap slightly). It is possible to arrange two cards with eight common numbers, suitably placed, such that their chance of joint winning is equal to the unconditional probability of two randomly chosen cards both winning. I can also select a set of cards, such that each pair of cards has eight overlaps, and they are arranged as described.

But correlation is only a pairwise concept, so the set of cards will have exactly the same correlation matrix whether I use the same eight numbers for all the cards, or different sets. This will not affect the chance of two specific cards sharing a win, but it makes a vast difference to the chance of larger numbers of multiple winners from this set of cards. This is a general point. Except in some situations, mainly found in textbooks, pairwise correlation is an unreliable guide to estimating the probability of unusual events involving more than two events.

These two examples only scratch the surface of the correlation lessons taught in the Bingo laboratory. I think there is a reason that analyzing popular games often results in more useful intuition than studying mathematically neater examples. Games survive because they both harmonize with and challenge our sense of probability. Introductory statistics is nearly universally regarded as deadly boring, because it is usually taught in a way that is both discordant and orthogonal to the risk and uncertainty that everyone faces. Playing around with play is both a better way to learn and a surer guide to wisdom.