"Politicians use statistics in the same way that a drunk uses lamp-posts—for support rather than illumination."

Andrew Lang

Appendix: Some simple distributions and techniques to uncover irregularities

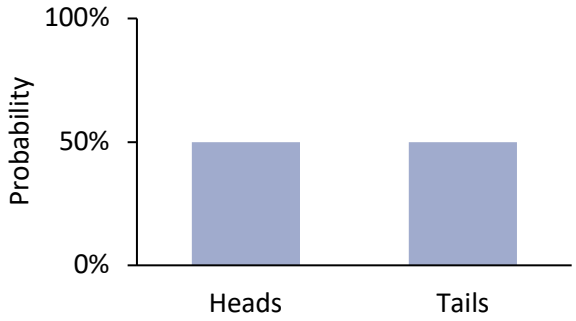
1. Cheating as irregularities

Cheating often reveals itself in some type of irregularity, such as a sudden improvement in performance. Thus, much of this book deals with the search for deviations from the norm. The norm can be defined by certain probability distributions or some other continuous pattern in the data, and deviations mean that the actual observations depart from the expected distribution or exhibit non-continuous patterns, like clustering. I elaborate in the following sections.

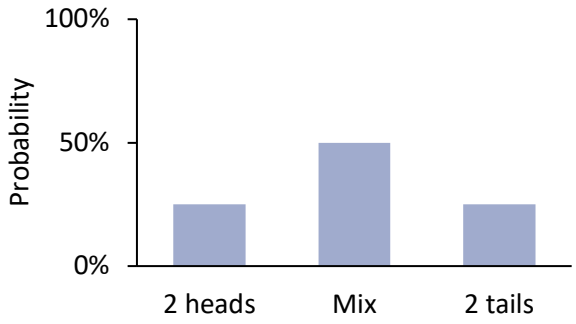
2. Binomial distribution

Many of the settings in this book are associated with two possible outcomes, such as winning or losing a game. In such situations, the binomial distribution is appropriate. If we have the probability of each of the outcomes, we can also use the binomial distribution formula to estimate probabilities of outcomes from repeated games.

Let me illustrate with some coin tosses. If you toss a fair (i.e., not loaded) coin once, the probability of heads is 50%. The probability distribution of that coin toss is simply:



If you toss the coin twice, the probability of two heads is 0.5 × 0.5 = 0.25, the probability of two tails is also 0.5 × 0.5 = 0.25, and the probability of one of each is 0.5. This is illustrated below.



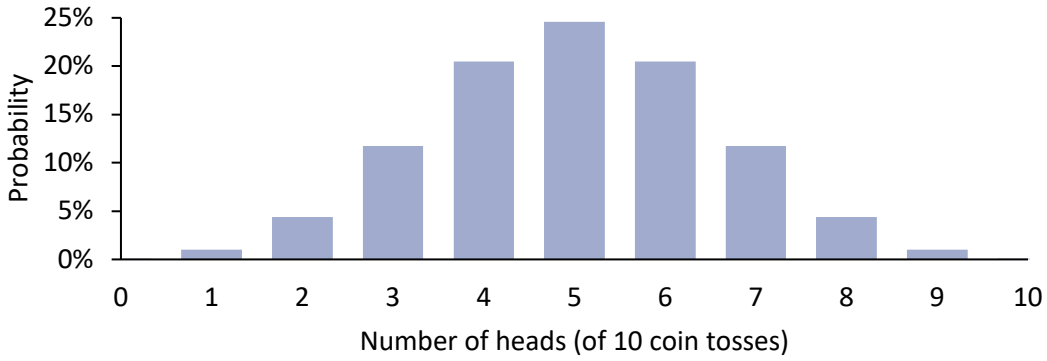
So far, we have managed with some simple calculations. But what is the probability of exactly six heads if you toss the coin ten times? This requires heavier artillery. The binomial distribution formula to the rescue:

$$\text{Probability of } k \text{ heads} = \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k}$$

where *n* is the number of coin tosses and p is the probability of heads in each toss.[114] Thus, we get:

$$\text{Probability of 6 heads} = \frac{10!}{6!\,(10-6)!} 0.5^6 (1-0.5)^{10-6} = 0.205$$

I also calculated the probability for other outcomes and made a probability distribution graph:

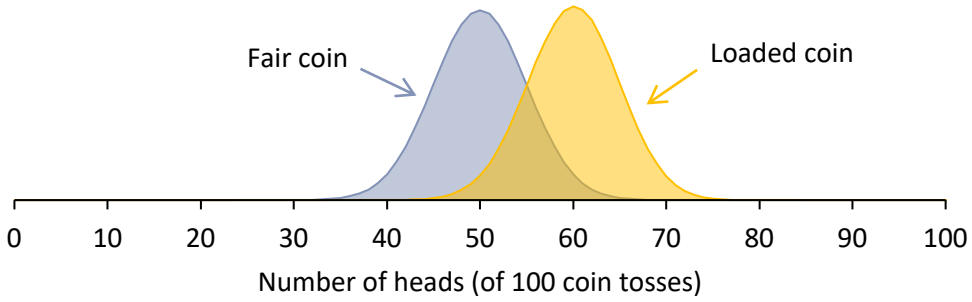


Let us ramp it up by using 100 coin-tosses and by comparing our fair coin to a loaded coin that has a 60% probability of heads. Then the distributions are given below. With the large number of coin tosses, I converted the columns into continuous lines, making them look like continuous distributions even if they are not. Unsurprisingly, the loaded coin has a distribution that is shifted to right relative to the one for the fair coin. But could we tell based on 100 coin-tosses whether a coin is loaded? That is not clear:

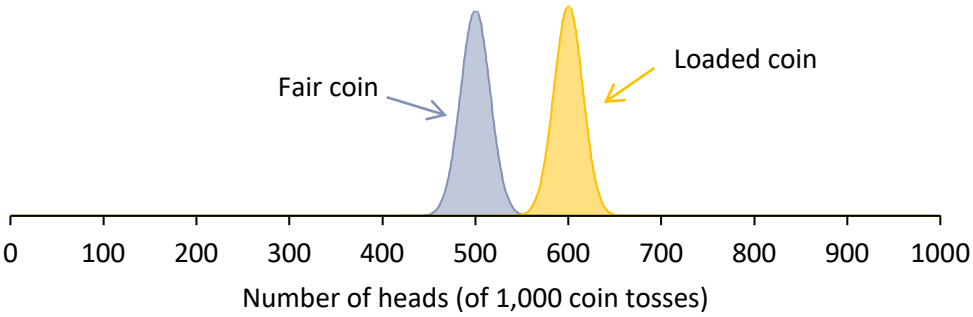
- Suppose that the loaded coin gives 56 heads, which is well within the probability distribution of the loaded coin. But that is also well within the probability distribution of the fair coin. Thus, we could not tell what distribution these 56 heads came from and whether a fair or loaded coin was used.

[114] You can find many binomial distribution calculators online via a simple search if the formula intimidates you. Or you can use the BINOMDIST function in Excel.

- Suppose instead that the loaded coin gives 64 heads, which is well within the probability distribution of the loaded coin. In contrast, it is unlikely to happen with a fair coin – 64 heads or more only happens 0.3% of the time when using a fair coin – so we could conclude with reasonable confidence that a loaded coin was used.

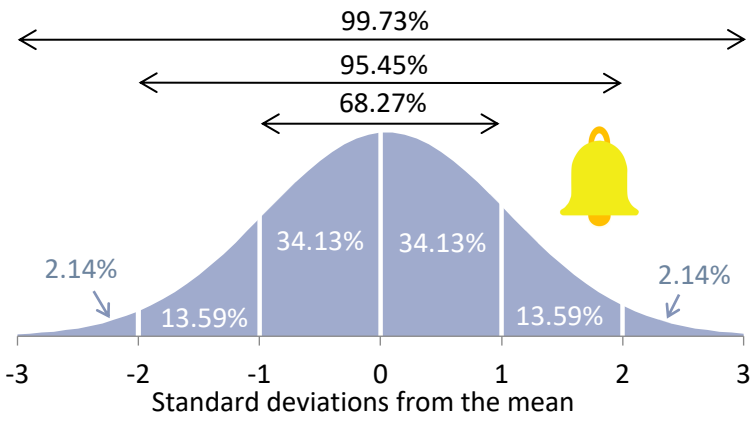


If we had a larger number of coin-tosses, we could tell with greater certainty whether a coin was fair. Consider the distributions below for 1,000 coin-tosses. Now there is barely any overlap in the distributions. If we were to use the loaded coin, it is almost certain (99.9%) that we would get more than 550 heads. Suppose then that we get 560 heads. The probability of that happening with a fair coin is less than 0.01%, so we could comfortably conclude that the coin was loaded. This illustrates the power of good statistical techniques combined with large samples to prove foul play.

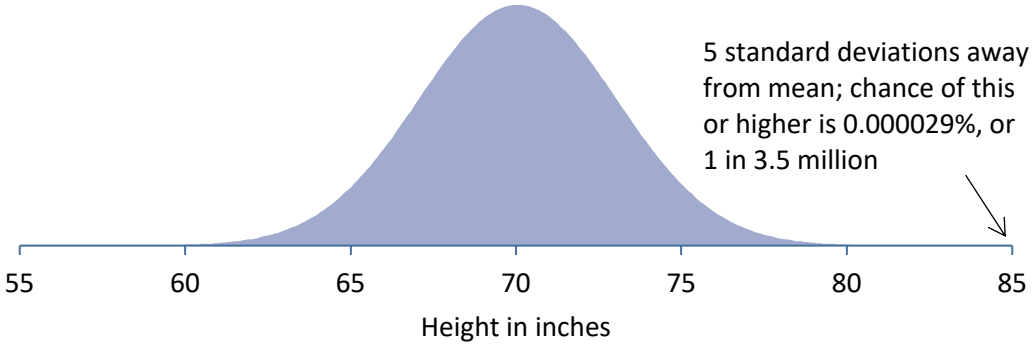


3. Normal distribution (bell curve)

The normal distribution is a symmetric and continuous distribution that happens to fit many data sets. It is often called a bell curve because of the resemblance of its shape to that of a bell, though based on the picture below, I think the nickname is a bit deceptive. While I will not state the formula for the normal distribution here, it is useful to know that one standard deviation in each direction from the average captures about 68% of the distribution and two standard deviations in each direction from the average capture about 95% of the distribution. That means that a 95% confidence interval for a variable that has a normal distribution can be found as two deviations below the average to two deviations above the average.



The heights of adults are distributed normally. In the US, the average height of adult men is about 70 inches (5' 10"), and the standard deviation is 3 inches. That means that 68% of US males are between 67 and 73 inches and 95% are between 64 and 76 inches. Suppose that your friend tells you that her uncle is 85 inches (7' 1"). Is that possible? Or does this mean that your friend's uncle is from a superhuman race with a different height distribution than the rest of us? Basketball fans know that such heights are not impossible. For example, Shaquille "Shaq" O'Neal is that height. If our estimates for the average height and standard deviation are correct, 85 inches is 15 inches away from the average of 70 inches, or 15/3 = 5 standard deviations. The chance of such an extreme height by pure chance is 0.000029%, or about 1 in 3.5 million.[115] Thus, I would be inclined to think that your friend exaggerated or was mistaken, but I could not rule out the possibility entirely.



You might have noticed in the previous section that the binomial distribution was flat with only one coin-toss, and it gradually morphed into a distribution that looked like the normal distribution as more coin-tosses were added. It turns out that it never quite turns into a normal distribution.[116] However, *the <u>average</u> for reasonably large samples is roughly normally distributed* around the population average, irrespective of what the distribution of the outcomes themselves looked like. Did you get that? This is

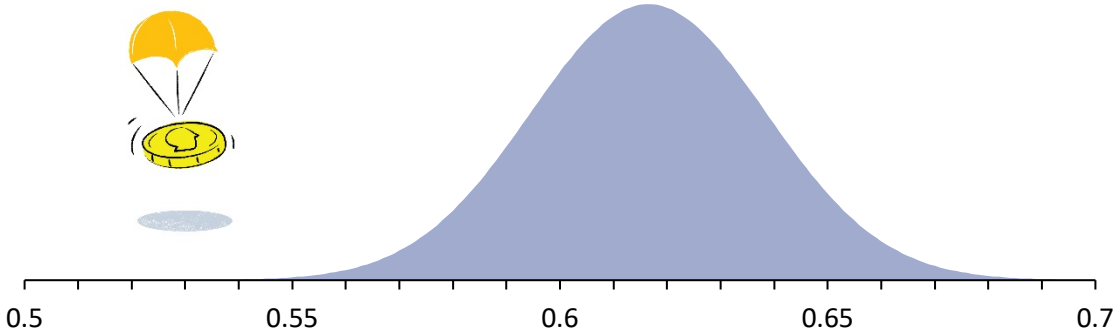
[115] You can use one of many normal distribution calculators available on the internet or the NORM.DIST function in Excel to verify.

[116] Even with many coin-tosses, e.g., a thousand, the distribution is still discrete. Also, with unequal probabilities, the distribution with many outcomes would be asymmetric, unlike the normal distribution.

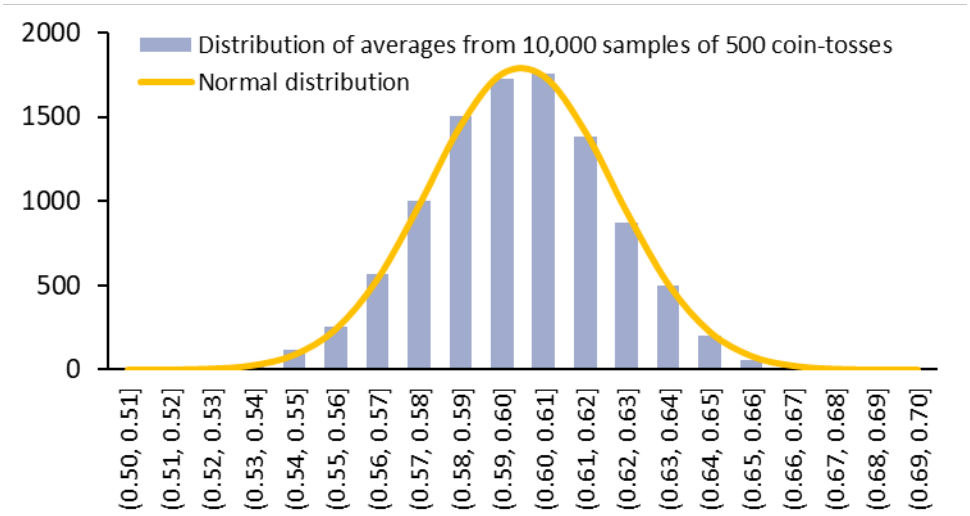
an important idea, so much so that it has its own name (the *Central Limit Theorem*) and explains why the normal distribution is so useful. Let me try to illustrate by expanding on the earlier coin tosses.

As a start, I tossed a coin 500 times and got 308 heads and 192 tails. (Secret to the reader: I used a loaded coin with a 60% probability of giving heads, and I want to see whether my statistical tests can tell that I used a loaded coin.) If I assign a value of one to the heads and zero to the tails, the average is 0.616 and the standard deviation is 0.487. Theory tells me that the average is normally distributed and that the standard deviation of the average depends on the size of the sample as follows:

$$Std.dev.of\ average = Std.dev.of\ tosses/\sqrt{No.of\ tosses} = 0.487/\sqrt{500} = 0.0218$$



Are you not convinced about the theory? Let me convince you then. To do that, I repeated the 500 coin-tosses 10,000 times.[117] After estimating the average for each of these 10,000 samples, I made a distribution of the averages below. It looks like it is normally distributed, right? It might be a little off, and had I used either more or larger samples, it would be closer still to the normal distribution. I also estimated the average of these averages to be 0.600 and the standard deviation of the averages to be 0.0221, very close to my earlier estimate of 0.0218 based on only one sample of 500 coin-tosses.



[117] Yes, that implies a total of five million coin-tosses, but I simulated them on my computer to save some time. To be honest, I do not even have a loaded coin.

My question then is whether I can show that the coin was not fair. What I know about a fair coin is that it has an average of 0.5. My initial sample had an average of 0.616. Having confirmed that this average comes from a normal distribution, I can test whether the true population average might have been 0.5. The estimated distance in standard deviations between 0.616 and 0.5 is $(0.616 - 0.5)/0.0218 = 5.3$, which in this context is an ocean apart; the probability that I would have drawn 0.616 from a normal distribution with an average of 0.5 is about one in 17 million. Thus, I can comfortably reject the possibility that the coin was fair.

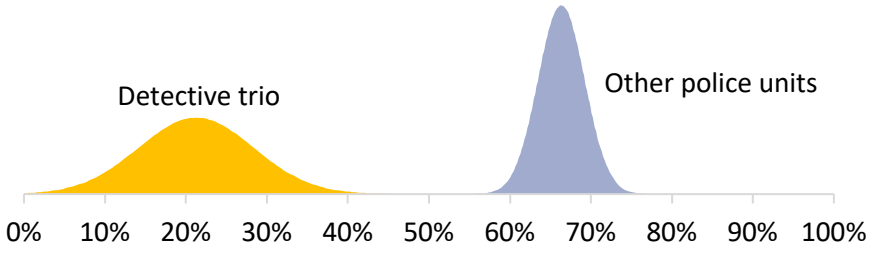
Enough coin tosses. Let us use another example with proportions based on an investigation by the Spotlight team of the Boston Globe.[118] In 1996, the Spotlight team investigated corruption in the Boston police. In one article, they focused on a trio of detectives accused of ripping off drug dealers during drug busts by stealing their drugs, cash, and other valuables. The article quoted a federal prosecutor who stated that "Once you kick the door, you almost always find cash." Yet, the Spotlight team found that the detective trio had reported to having found money in only seven of 33 searches (21%), compared to 179 of 270 searches (66%) for other police units.

We can use these numbers to estimate whether the proportions differ across the two groups. In doing so, we might interpret the proportion for other police units as a benchmark for how often cash is present during drug busts, and the test for the difference in proportion tells us whether the detective trio reports cash less often than it is present. Of course, it is possible that both the detective trio and other police units steal cash, in which case we can interpret the difference in proportion as a measure for how much worse the detective trio is compared to other police units. Either way, if we find the difference in proportions to be statistically significant, the detective trio got some "splainin" to do.

A proportion is like an average, so it is also distributed normally. In this case, we have two normal distributions with averages of 21% and 66%. Furthermore, the standard deviation of the averages can be estimated as $\sqrt{p(1-p)/n}$ where p is the proportion and n is the sample size.[119] Based on the averages, standard deviations, and the normality assumption, I made a figure of the distributions for the two proportions below.

[118] The Spotlight team is the oldest newspaper investigative journalist unit in the US and was featured in the 2015 Academy award-winning movie *Spotlight* about the investigation of child sex abuse among Roman Catholic priests in the Boston area.

[119] In fear of getting too technical here, let me quickly show why this simple formula works, or you can just trust me and skip this footnote. The standard deviation is the square root of the sum of the squared deviations from the proportion divided by the sample size. Thus, there would be $p \times n$ cases where the squared deviation is $(1-p)^2$ and $(1-p) \times n$ cases where the squared deviation is $(0-p)^2$. Adding up the squared deviations gives $n \times p \times (1-p)$. Dividing by n and taking the standard deviations gives $\sqrt{p \times (1-p)}$. To get the standard deviation for the average, we divide by the square root of n to get $\sqrt{p \times (1-p)/n}$.



There is no overlap in these distributions, proving that the proportions are truly different.[120] While the different does not prove that the detective trio has been stealing, a Boston police official said that it should have served as a "red flag for any decent supervisor." It turned out that these detectives also had a much larger consumption than what is normal for cops, e.g., having bought several expensive condominiums, and further interviews and investigation proved that the trio had indeed stolen cash, sometimes a portion of what they found and other times all of it.

Lastly, we will examine an example with continuous variables. Suppose that you just bought a new car, and after commuting on the interstate for a few weeks, you are skeptical about the manufacturer's claim that the car gets 30 miles per gallon on the highway. Thus, you decide to test for yourself. You start with a full tank, drive 462 miles, and then fill up 16.1 gallons to full tank again and estimate the miles per gallon to be 462 / 16.1 = 28.70. You repeat this cycle another four times until you have the data tabulated below:

Miles	Gallons	MPG
462	16.1	28.70
446	15.1	29.54
505	17.3	29.19
464	16.2	28.64
434	14.9	29.13

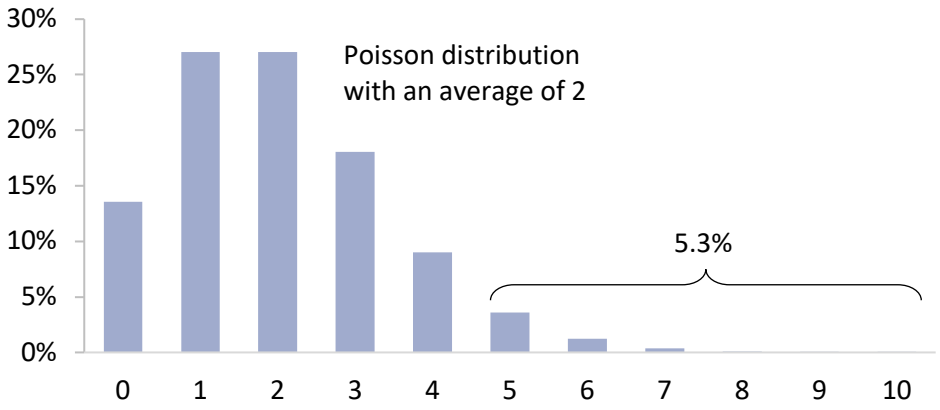
The sample average miles per gallon is 29.04, the sample standard deviation is 0.37, and the standard deviation of the sample average is $0.37/\sqrt{5} = 0.17$. The sample average of 29.04 is 0.96 from the manufacturer's claim, representing 0.96 / 0.17 = 5.8 standard deviations. A normal distribution calculator shows that if the manufacturer's claim is correct, the probability that such a deviation should occur is 0.000001%. This means that unless the driving conditions under your tests were unusual, the manufacturer's claim is incorrect.[121]

[120] We could also test the difference in proportions more formally based on the test-statistic $= \frac{Difference\ in\ proportions}{\sqrt{p(1-p)(1/n_1+1/n_2)}}$, where p is the overall proportion across the two groups, n_1 and n_2 are two sample sizes, and the test-statistic is distributed normally. In this case, the test-statistic equals 5.02, which occurs by chance in about one in 750,000 cases. In other words, there is minimal likelihood that the trio of detectives reported cash in fewer cases than other police units by pure chance.

[121] Several car manufacturers have admitted to cheating on the mileage claims, for example by having the car shift differently on the treadmill during testing to optimize fuel economy at the expense of performance. More famously, VW cheated on its emission tests for its TDI diesel engines, leading to a scandal that cost it tens of

4. Poisson distribution

The Poisson distribution is a discrete distribution that gives the probability of a given number of events occurring during fixed period. The only input for the distribution is the average number of events per period. For example, suppose that, on average, two hurricanes hit the US every year based on data for the last few decades. Then, assuming that there is no change in the global climate, we can use the Poisson distribution to estimate the probability that, say, five or more hurricanes will hit the US next year. The Poisson distribution with an average of two events per period is pictured below. From this distribution, we see that the probability of an outcome of five or more is 5.3%.[122]



The chapter on lotteries uses the Poisson distribution to test whether it is possible to win as frequently as some players did without cheating.

5. Discontinuities and bunching

If you walk at the airport and see people with all kinds of colors, but then you see a cluster of people wearing hot pink, what is your conclusion? Clearly, the cluster is highly irregular and suggests collusion,

billions of dollars in 2014. In particular, VW employed software that could detect when its engine was being tested based on speed and steering wheel movement, at which point the software would engage the testing mode. The testing mode turned on a system that temporarily reduced emissions, including a nitrogen oxide trap, at the cost of reduced performance and fuel economy. When researchers at West Virginia University tested VW cars on the road, thus bypassing the testing mode, they discovered emission levels up to 40 times above VW's claims.

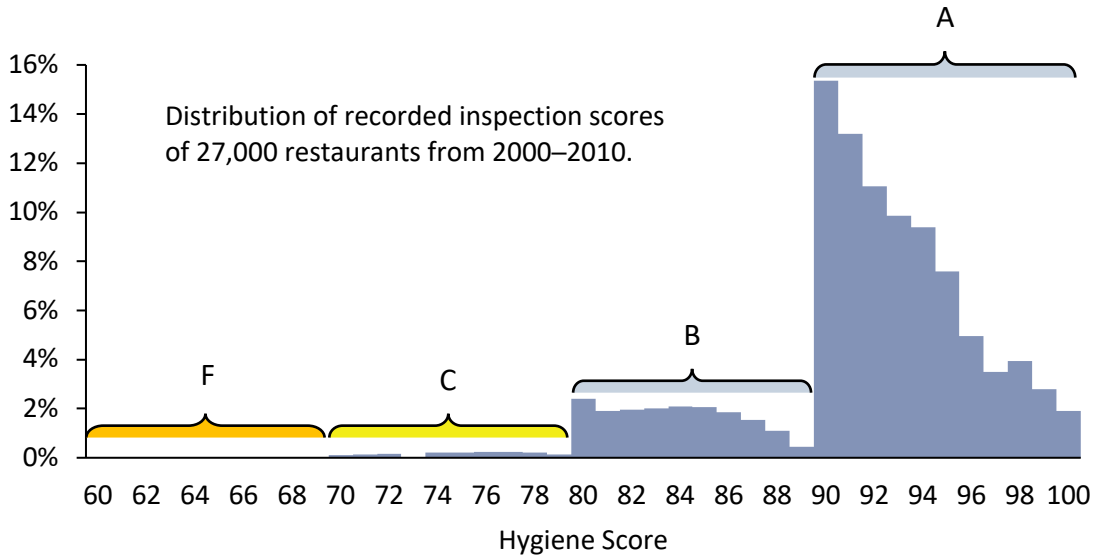
[122] Like with the other distributions, there are plenty of online Poisson distribution calculators that you can use for this.

most likely a group of individuals traveling together who have agreed beforehand to coordinate the outfit to better see each other during their travel.

It is often unclear what the underlying distribution of the data is. Regardless, we generally expect the distribution for a large sample to be smooth. Researchers often use interruptions or bunching in the data as clues of irregularities. I find those types of studies to be particularly insightful.

Balázs Kovácsa, David Lehman, and Glenn Carroll examined the distribution of hygiene scores from 336,208 inspections of 27,119 restaurants conducted by 493 inspectors from 2000 to 2010 by the Los Angeles Department of Public Health (LADPH). Such inspections are conducted to reduce cases of foodborne illnesses and improve public health. The numerical hygiene scores are translated into letter grades like those used in the school system (with A being the best, etc.), which are printed on a placard to be displayed in a window of the restaurant. Customers perceive a letter grade of A to be a certification of food safety and are more likely to visit such restaurants.

The figure below shows the distribution of the hygiene scores. Impartial inspections should yield a continuous distribution. That is clearly not what we see here. Rather, we see discontinuous jumps between 79 and 80 points and between 89 and 90 points, precisely the places that separate the letter grades C and B and the grades B and A, respectively. Thus, a suspiciously high number of restaurants on the borderlines received scores that were just enough to secure either an A or a B.

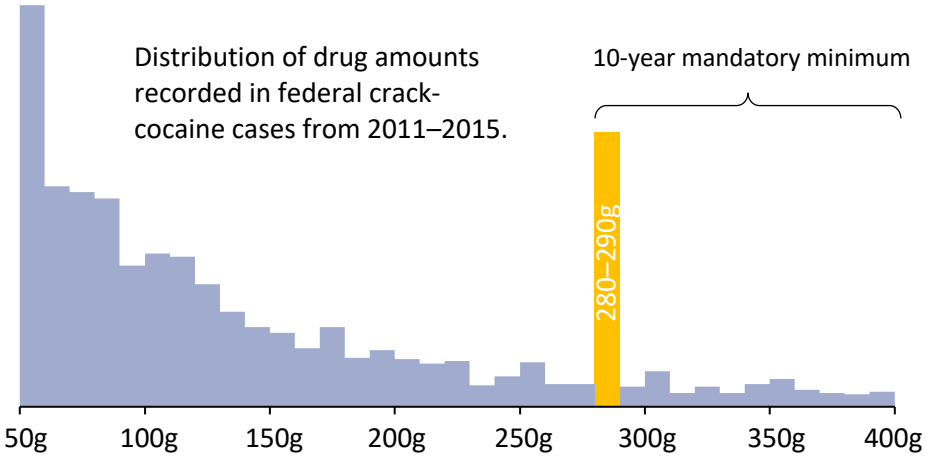


The authors further found evidence that grade inflation was more likely for repeated interactions, i.e., when an inspector had visited the same restaurant before. Perhaps the inspectors became friendly with the restaurant-owners over time and wished to help in small ways when they could. Or perhaps there was plain old bribery or corruption involved. I lean toward the former, and less perturbing, possibility.

Cody Tuttle examined the distribution of drug amounts recorded in federal crack-cocaine cases. The Fair Sentencing Act of 2010 increased the ten-year mandatory minimum threshold for crack-cocaine from 50g to 280g. The figure below shows the distribution of drug amounts during the five years after the legislative change. One cannot help but notice the unusual bunching for the category just above 280g. (Coloring it orange also contributes, I suppose.)

What is going on here? Why would so many drug-dealers and traffickers carry just enough crack-cocaine to subject themselves to a minimum ten-year prison sentence? It would be more rational for them to carry just *below* the 280g threshold. Well, it is probably not drug-dealers being irrational. Rather, the legal system gives the police and prosecutors flexibility in the amount used for sentencing; they can change the amount to the 280g threshold or right above to increase the chance of harsh sentencing. Further evidence suggests that this happens at the prosecutor level: the bunching was absent from the data on state-level convictions and drug seizures, but it was present in the prosecutor case management data. Also, 20–30% of the prosecutors were responsible for the bunching!

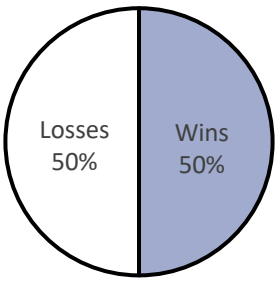
Even more troubling, Tuttle showed that the bunching disproportionately affects minorities. That is, the bunching was much more prevalent among blacks and Hispanics than among whites. Thus, racial discrimination appears embedded in a subset of prosecutors.



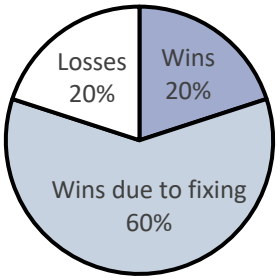
6. Estimating manipulation when the distribution of outcomes in the absence of manipulation is known

Many outcomes are split about evenly. For example, the distribution of coin-tosses with a fair coin should be 50% heads and 50% tails, and the gender at birth should be close to 50% boys and 50% girls. When we observe deviations from these expected outcomes, we can estimate the fraction of outcomes that have been tampered with using some simple formulas. Indeed, the chapter on match-fixing uses this in a couple of cases, while the chapter on backdating uses it once.

Suppose that the chance that a random tennis player wins a game is 50%, but that this jumps to 100% if the game is fixed in her favor. Further, suppose that a fraction p of her games are fixed and the rest $(1-p)$ are not. Then she should win a fraction $q = 100\% \times p + 50\% \times (1-p)$ of her games. We can rearrange this to solve for the fraction of manipulated games as $p = 2q - 1$. For example, if we see that she wins 80 of 100 games, we can estimate that fraction as $p = 2q - 1 = 2 \times 80\% - 1 = 60\%$. That is what the pie charts below depict.



No game-fixing



60% of games are fixed

We could also generalize to allow for imperfect match fixing and a fraction of wins in the absence of game fixing that differs from 50%. Suppose that game-fixing inflates the chance of winning to f, which is between 50% and 100%, and that the fraction of wins in the absence of game fixing is n. Then the tennis player should win a fraction $q = f \times p + n \times (1 - p)$ of her games, which we can rearrange to $p = (q - n)/(f - n)$. For example, if she ordinarily wins 50% of her games without game-fixing (so n is 50% just like above), game-fixing increases the winning probability to 90% (meaning that f is 90%), and she wins 80 of 100 games, the estimated fraction of fixed games is $p = (0.8 - 0.5)/(0.9 - 0.5) = 75\%$. This is higher than the estimate of 60% above, showing that the estimated fraction of fixed games would be higher if the match fixing is imperfect.

Let us use the formula for gender at birth. In both China and India, the sex ratio at birth is about 1.11, meaning that 1.11 boys are born for every girl, which translates to $1.11/(1.11 + 1) = 52.6\%$ boys. This seems to be high and suggests tampering, e.g., sex-selective abortions. In comparison, the sex ratio at birth is 1.05 in the US, translating to $1.05/(1.05 + 1) = 51.2\%$ boys. If the sex ratio should be 1.00 in the absence of human tampering, what fraction of births have been tampered with? Assuming that any tampering alters the probability of a boy from 50% to 100% (i.e., f is 100%), we get $p = \frac{q-n}{f-n} = \frac{52.6\%-50\%}{100\%-50\%}$ =5.2%. How should we interpret this? A possible interpretation is that the parents examined the gender of the fetus in 5.2% of the cases with the intention of abortion if it was a girl, and, consequently, about 5.2% of the female fetuses were aborted.

Suppose instead that we are quite confident that more boys are born for natural reasons, and that we can use the sex ratio in the US as an estimate for what that the sex ratio should be by nature. In other words, we are assuming that no tampering occurs in the US. If so, we can re-estimate the fraction of births that have been tampered with in China and India as $p = \frac{q-n}{f-n} = \frac{52.6\%-51.2\%}{100\%-51.2\%}$ =2.9%.

Sex-selective abortions are controversial, so I am very reluctant to draw any firm conclusions based on these estimates. My goal was merely to show that if we start with a set of assumptions for (i) what the distribution of outcomes should be in the absence of manipulation and (ii) how effective the manipulation is, we can use samples of outcomes to arrive at reasonable estimates for the fraction of those outcomes that were manipulated.

Bibliography:

Tuttle, Cody, 2019, Racial Disparities in Federal Sentencing: Evidence from Drug Mandatory Minimums, Working paper.

Kovácsa, Balázs, David W. Lehman, and Glenn R. Carroll, 2020, Grade inflation in restaurant hygiene inspections: Repeated interactions between inspectors and restaurateurs, *Food Policy* 97.

O'Neill, Gerard, Mitchell Zuckoff, and Dick Lehr, 1996, Corruption probe shakes up Boston police detective unit: The case of the disappearing money, *The Boston Globe*.