Weldon's Dice, AUTOMATED

Zacariah Labby

Walter Frank Raphael Weldon's data on 26,306 rolls of 12 dice have been a source of fascination since their publication in Karl Pearson's seminal paper introducing the χ^2 goodness-of-fit statistic in 1900. A. W. Kemp and C. D. Kemp also wrote about the historical data in 1991, including methods of analysis beyond Pearson's goodness-of-fit test.

Although modern random number generators have come a long way in terms of periodicity and correlation, there is still a certain cachet in the apparent randomness of rolling dice, even if this appearance is ill-founded. As Pierre-Simon Laplace said, "The word 'chance' then expresses only our ignorance of the causes of the phenomena that we observe."

Weldon's Dice Data

In a letter to Francis Galton—dated February 2, 1894—Weldon reported the results of 26,306 rolls of 12 dice, where he considered five or six dots (pips) showing to be a success and all other pip counts as failures. The data were presented in tabular form, with the number of successes per roll tallied as in Table 1. Weldon was motivated to

Table 1—Weldon's Data on Dice: 26,306 Throws of 12 Dice

Number of Successes	Observed Frequency	Theoretical Frequency, p = 1/3	Deviation
0	185	203	-18
1	1149	1216	-67
2	3265	3345	-80
3	5475	5576	-101
4	6114	6273	-159
5	5194	5018	176
6	3067	2927	140
7	1331	1255	76
8	403	392	11
9	105	87	18
10	14	13	1
11	4	1	3
12	0	0	0
Total	**26,306**	**26,306**	$\chi^2_{[10]} = 32.7$

Note: A die was considered a success if five or six pips were showing.

Walter F. R. Weldon (1860–1906), an English biologist and biometrician

collect the data, in part, to "judge whether the differences between a series of group frequencies and a theoretical law, taken as a whole, were or were not more than might be attributed to the chance fluctuations of random sampling."

The simplest assumption about dice as random-number generators is that each face is equally likely, and therefore the event "five or six" will occur with probability 1/3 and the number of successes out of 12 will be distributed according to the binomial distribution. When the data are compared to this "fair binomial" hypothesis using Pearson's χ^2 test without any binning, Pearson found a p-value of 0.000016, or "the odds are 62,499 to 1 against such a system of deviations on a random selection."

The modern application of the goodness-of-fit test requires binning such that each theoretical bin has at least approximately four counts, and for the data in Table 1, this results in the bins 10, 11, and 12 grouped into one "10+" bin. With the appropriate binning, the p-value for the original data becomes 0.00030, a larger but still significant result.

The conclusion is that the dice show a clear bias toward fives and sixes, which Pearson estimated was probably due to the construction of the dice. Most inexpensive dice have hollowed-out pips, and since opposite sides add to seven, the face with six pips is lighter than its opposing face, which has only one pip.

While the dice may not follow the fair binomial hypothesis, they still may follow a binomial hypothesis with bias toward fives and sixes. The overall probability of a five or six is estimated as 0.3377 from the data, and Pearson outlines the comparison of the dice data to this alternate theoretical distribution in his illustration II of the 1900 paper. Correcting errors in his arithmetic, $\chi^2 = 17.0$ for the unbinned data and $\chi^2_{[9]} = 8.20$ for the binned data (with binning performed as outlined above).

As many university students learn in introductory statistics courses, the estimation of one variable by maximum likelihood must be 'repaid' by dropping one degree of freedom in the goodness-of-fit test, hence the nine degrees of freedom for the "biased binomial" test. The resulting p-value is 0.51, meaning there is not sufficient evidence to refute the claim that, although biased, the dice still follow the binomial distribution. These two applications of the original dice data have served as examples, introducing the χ^2 goodness-of-fit statistic to many.

Design of Apparatus

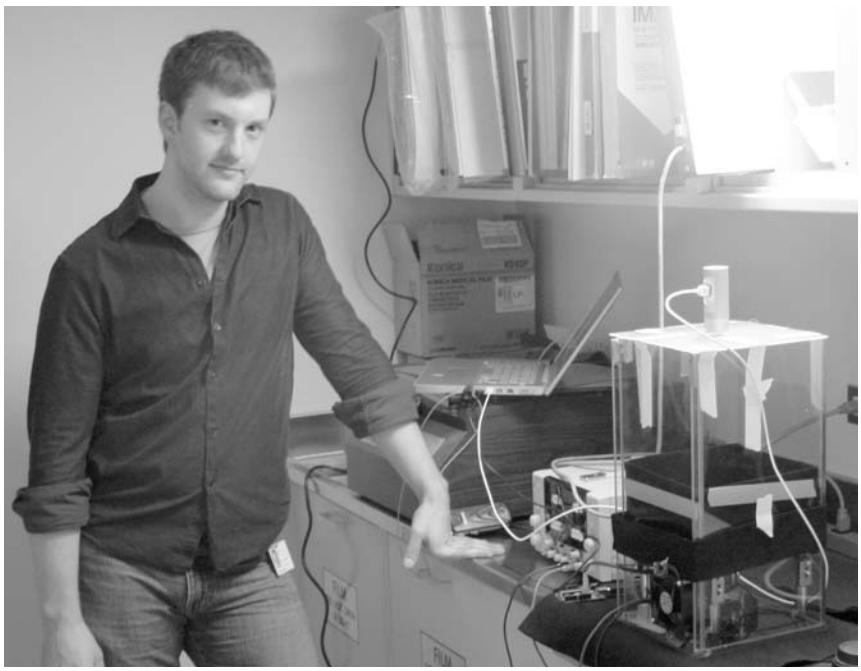
While it is possible to repeat Weldon's experiment by hand, such an endeavor would be dull and prone to error. Here, we will use an automatic process consisting of a physical box that rolls the dice, electronics that control the timing of the dice-rolling, a webcam that captures an image of the dice, and a laptop that coordinates the processes and analyzes the images.

The idea behind the dice-rolling is as follows: A thin plate of metal is covered in felt and placed between metal U-channel brackets on opposing faces of a plastic box. Solenoids with return springs (like electronic pinball plungers) are mounted under the four corners of the metal plate and the 12

(a) (b) (c)

Figure 1. Apparatus for the rolling of dice. (a) shows the locations of the solenoids in the box without the metal plate in place, along with the return spring plungers in place. (b) is a close-up view of the metal plate in bracket design. (c) shows the apparatus fully assembled in the midst of a rolling sequence.

dice are placed on top of it. The dice are inexpensive, standard white plastic dice with hollowed-out pips and a drop of black paint inside each pip. The dice have rounded corners and edges. The front panel of the plastic box is removable for easy access to the solenoids and dice. Once assembled, the inside of the box is lined with black felt to suppress reflections from the inside surface of the plastic (Figure 1).



Zacariah Labby beside his homemade dice-throwing, pip-counting machine

The solenoids are controlled through an Arduino USB board, which can be programmed using a computer. The USB board listens on the serial port, and when it receives an appropriate signal, it sends a series of digital on/off pulses to four independent relay switches, which control the solenoids' access to electricity. When the relay is placed in the on position, current flows to the solenoid, thereby depressing the plunger against the force of the return spring, due to magnetic inductance. When the relay returns to the off position, the plunger is allowed to freely accelerate under the force of the return spring until it hits the metal plate, transferring its momentum to the (limited) motion of the plate and the (unlimited) motion of the dice.

The solenoids operate independently, and their power is supplied from a standard DC power supply. If the four solenoids are numbered clockwise one through four, three solenoids at a time are depressed, initially leaving out number four. Then, the solenoids spring back, and 0.25 seconds later,

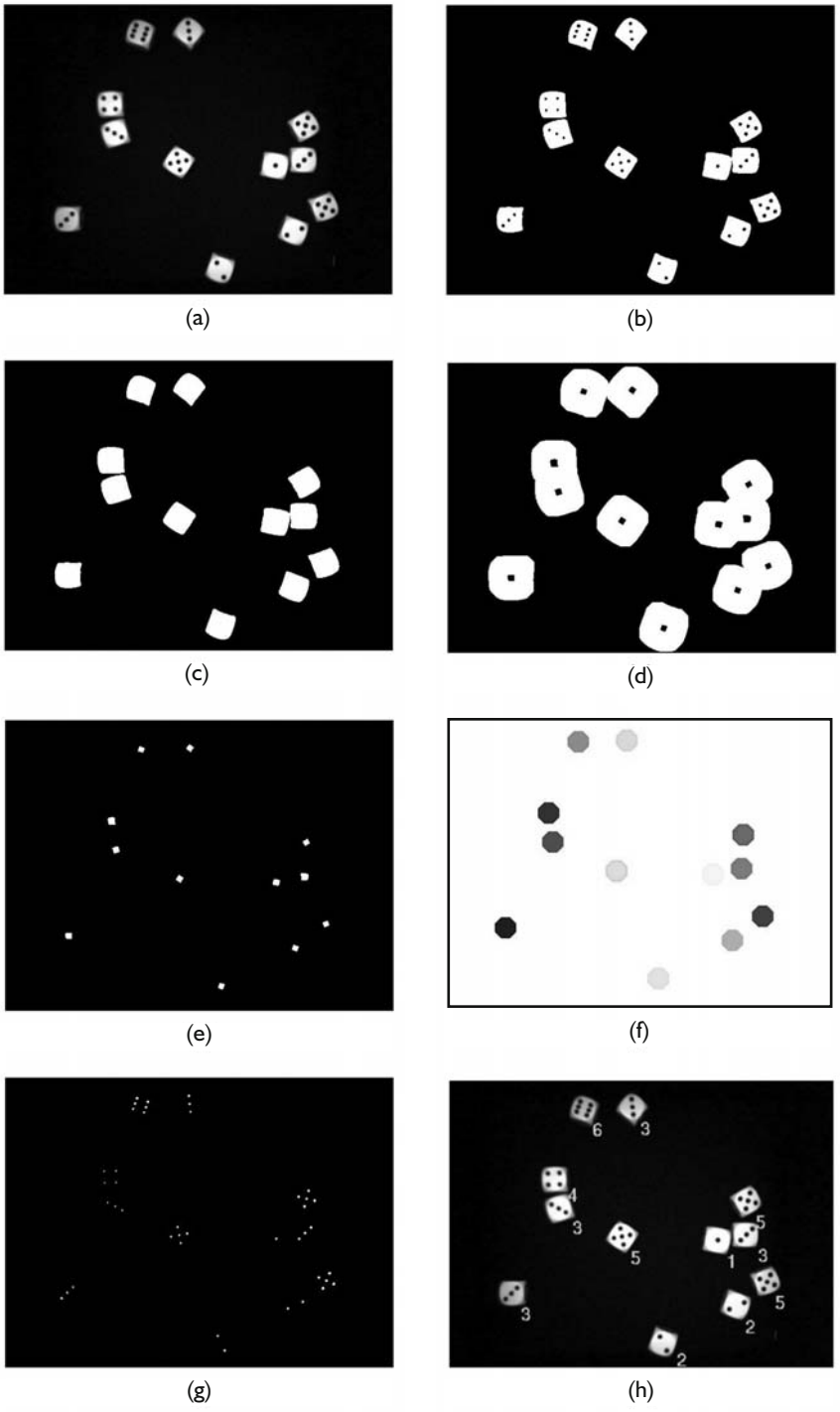


Figure 2. Image analysis procedure. The raw grayscale image in (a) is thresholded in a consistent manner with the aid of controlled lighting, leading to (b). After the holes in the image have been filled, (c), the edges are identified and dilated to ensure complete separation of touching dice (d). Once the edges have been removed from (c), the centers of the dice, (e), are identified and a unique mask is placed over the center of each die (f). The pips are identified from the subtraction of (c) and (b), and the results are eroded to ensure the pips on a six-face do not bleed together (g). Finally, the number of unique pips beneath each mask is counted and stored in memory. The results are displayed for the user to monitor (h).

three solenoids are depressed, leaving out number three. The pattern continues, leaving out two, then leaving out one, then repeating the entire pattern. When only one solenoid is left unaltered at a time, the metal plate tilts away from that solenoid, and the dice roll on the plate. When the other three solenoids return to their standard positions, the dice pop into the air.

After the dice have come to rest for a few seconds, a laptop connected to a webcam captures a grayscale image of the dice. The acquisition and processing of the image, along with the entire automation process, is controlled from the computer. The lighting in the experiment room is carefully controlled using radiography light boxes (for viewing X-ray films), which provide uniform and diffuse lighting to reduce glare on the surfaces of the dice.

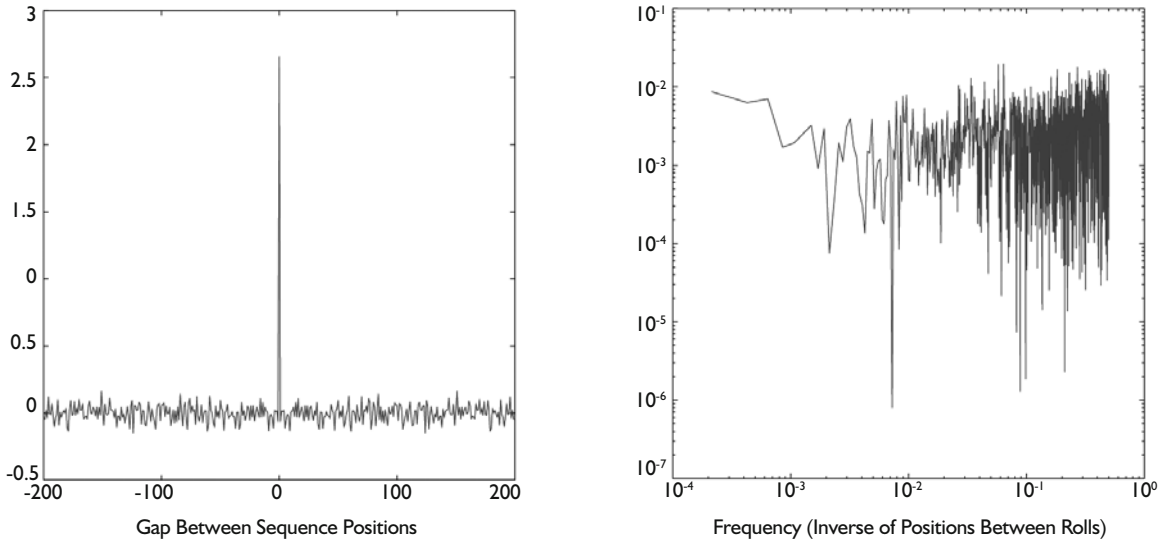
Because the lighting is carefully controlled, a simple threshold can be applied to the grayscale image, resulting in an array of ones and zeros. The 'holes' in the image, which represent the pips of the dice, are filled in, and the resulting image shows the dice as white squares on a black background. When two dice are in close contact, the thresholding process often fails to completely separate the dice and an edge-finding algorithm is used to find any transition regions in the original image.

The edges are dilated (inflated) to ensure the shared border between any two touching dice is entirely encompassed within the detected edges, and any regions labeled as edges are subsequently removed from the thresholded image. This leaves the center of each die as an independent region. The uniquely connected components are identified and a mask is placed at the centroid of each to serve as a search space for pips.

Pips are identified from the original thresholded image, again through the identification of uniquely connected components. These initial components are eroded (i.e., a few pixels along the identified perimeters are removed) to prevent any pips from 'bleeding' together, which is common with dice showing six pips. The number of pips under each die mask is counted and the results are both stored in an array in the computer and displayed to the user to ensure proper operation. The image analysis steps are shown in Figure 2.

(a) (b) (c)

Figure 3. Images leading to errors during analysis. In (a), a die has landed perfectly on one corner. In (b), one die has landed atop another, leading to the software only identifying 11 dice. Finally, in (c), one die has come to rest against the felt-covered wall of the apparatus, leading to an improper lighting situation identified as an error.



(a) (b)

Figure 4. Lack of correlation between dice-rolling iterations. (a) shows the autocorrelation of the sequence of the number of successes per roll (central portion). The only point above the noise floor is at zero lag. (b) shows the corresponding power spectrum, which has the characteristic appearance of white noise.

There are numerous opportunities for error in this sequence, and error-catching steps are taken during the image analysis. For instance, if any uniquely labeled die is not within a tight size range—as is possible if two dice are perfectly flush and not separated—the image is considered to be an error. If 12 dice are not found in the image, which happens when one die lands on top of another during the tossing sequence, the image is an error. If any die does not land on a face, but rather on an edge or corner, the lighting is such that the whole die will not be found and the image is an error. Also, whenever any pip is too oblong, as is the case when pips bleed together, the image is an error. A few images that lead to errors are shown in Figure 3.

Any time an error occurs (approximately 4% of the time), the image is saved externally and, when possible, the numbers on the 12 dice are entered manually. Unfortunately, some images are impossible to count manually (see Figure 3a). Those that are possible to count manually have a bias toward showing a large number of sixes, as the pips on sixes can bleed together, and therefore

Table 2—Current Data on Dice: 26,306 Throws of 12 Dice

Number of Successes	Observed Frequency	Theoretical Frequency $p = 1/3$	Theoretical Frequency $p = 0.3343$
0	216	203	199
1	1194	1216	1201
2	3292	3345	3316
3	5624	5576	5551
4	6186	6273	6272
5	5047	5018	5040
6	2953	2927	2953
7	1288	1255	1271
8	406	392	399
9	85	87	89
10	13	13	13
11	2	1	1
12	0	0	0
Total	**26,306**	$\chi^2_{[10]} = 5.62$	$\chi^2_{[9]} = 4.32$

Note: A die was considered a success if five or six pips were showing.

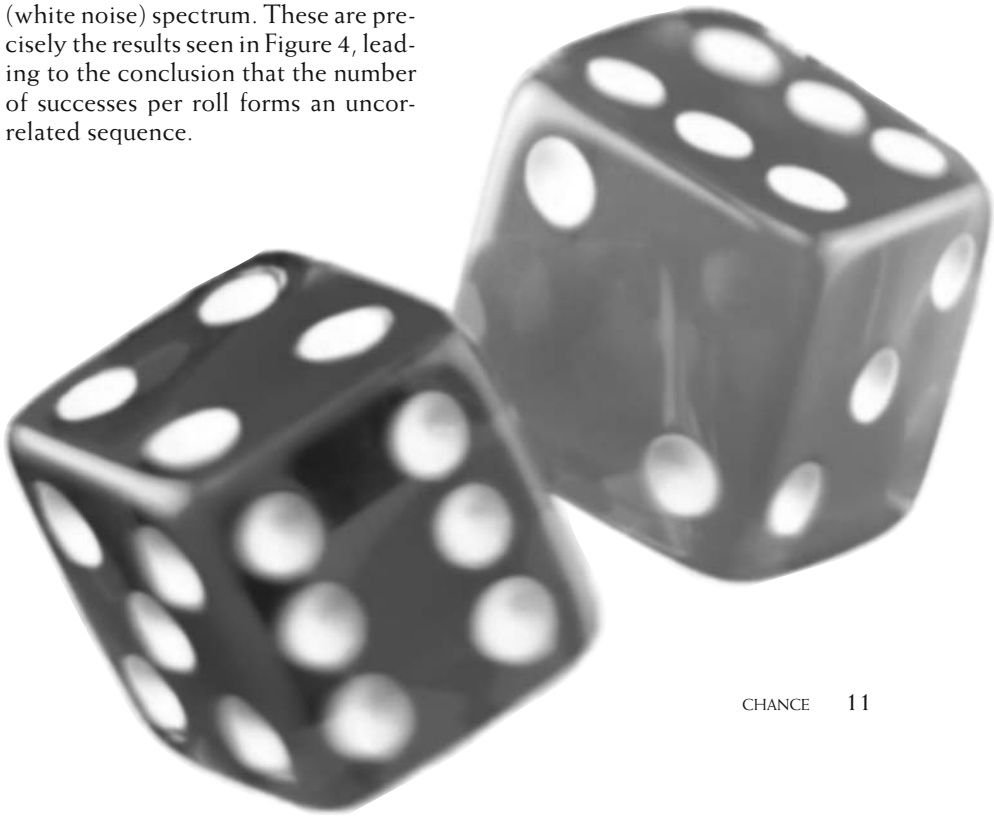
ignoring error images would possibly lead to bias in the results. With the entire rolling-imaging process repeating every 20 seconds, there are just more than 150 error images to process manually each complete day of operation. At the previously mentioned rate of operation, Weldon's experiment can be repeated in a little more than six full days.

Results

After all 26,306 runs were completed, all error images were processed to remove potential bias from the following analysis. (There were 27 'uncountable' error images.) Initially, a "success" will hold the same meaning as it did for Weldon: a five or six showing on the up-face of a die. To assess any correlation (or inadequacy) in the dice-rolling procedure, it is useful to look at the autocorrelation of the sequence of successes per iteration.

If a high number of successes in one iteration leads, for example, to a correlated number of successes in the next iteration or the iteration following, this will appear as an identifiable peak in the sequence autocorrelation. However, if the sequence is largely uncorrelated, the only identifiable peak in the autocorrelation will be at zero lag between iterations, and the Fourier transform of the autocorrelation will be a uniform (white noise) spectrum. These are precisely the results seen in Figure 4, leading to the conclusion that the number of successes per roll forms an uncorrelated sequence.

The distribution of successes per iteration is shown in Table 2, where it can be seen that the χ^2 values are not large enough to reliably reject either the fair or biased binomial hypotheses.



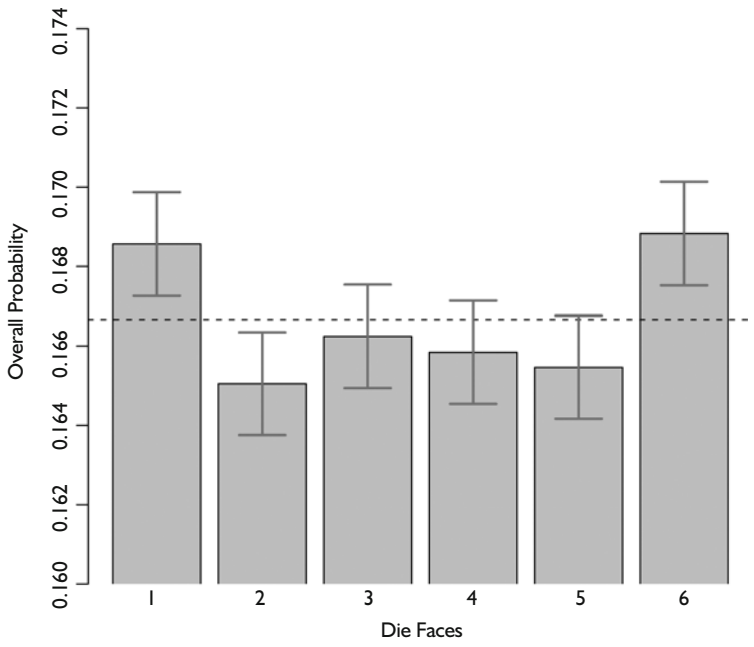


Figure 5. Probability of observing each number of pips out of 12 times 26,306 total rolls. The error bars are 95% confidence intervals according to binomial sampling, where $\sigma^2 = p(1-p)/315672$ and the dashed line shows the fair probability of 1/6 for each face.

Binning is performed on the data in Table 2 as outlined under "Weldon's Dice Data." The overall probability of a five or six showing is estimated to be 0.3343. From these results, the dice seem to be in accordance with the fair binomial hypothesis, unlike Weldon's dice. This is as far as Weldon (or Pearson) could have gone with the original 1894 data, but this is by no means the end of the story.

Besides the automation, which is a time-saving step, the unique aspect of this experiment is that the individual number of pips on each die is recorded with each iteration, and not just whether the die was a success or failure. This allows a much deeper analysis of the data. For instance, instead of jointly analyzing fives and sixes as a success, we find some interesting results if a success is considered to be only one face. First, the probabilities for the individual faces are estimated to be $Pr_1 = 0.1686$, $Pr_2 = 0.1651$, $Pr_3 = 0.1662$, $Pr_4 = 0.1658$, $Pr_5 = 0.1655$, and $Pr_6 = 0.1688$, whereas the fair hypothesis would indicate that each face should have probability $Pr_i = 1/6 = 0.1667$.

These probabilities, along with their uncertainties in a binomial model, are shown in Figure 5. The tossing apparatus does not seem to alter the probabilities of the individual die faces over time. Comparing the first third and final thirds of dice rolls, and adjusting for multiple comparisons, reveals that none of the face probabilities significantly changed over time.

When comparing the observed number of counts for each pip face with the expected fair value (12 times 26,306/6 = 52,612) in a χ^2 test, the resulting $\chi^2_{[5]} = 25.0$ and $p = 0.00014$ leave little doubt that the dice results are biased. If the dice were biased in the manner Pearson assumed—namely due to pip-weight imbalance—we would expect the probabilities of the individual faces to follow a linear trend of -5, -3, -1, 1, 3, 5 for faces one through six, respectively. Fitting and testing for this pattern yields a p-value of 0.00005, allowing reliable rejection of the pip-weight-trend hypothesis.

Discussion and Conclusion

One interesting point from the data obtained here is the revelation of a non sequitur by Pearson in his biography of Galton. In a footnote, he writes:

> Ordinary dice do not follow the rules usually laid down for them in treatises on probability, because the pips are cut out on the faces, and the fives and sixes are thus more frequent than aces or deuces. This point was demonstrated by W. F. R. Weldon in 25,000 throws of 12 ordinary dice. Galton had true cubes of hard ebony made as accurate dice, and these still exist in the Galtoniana.

Weldon's dice were most likely made from wood, ivory, or bone with carved pips, but that fives and sixes jointly occurred more often than one would expect under the fair hypothesis

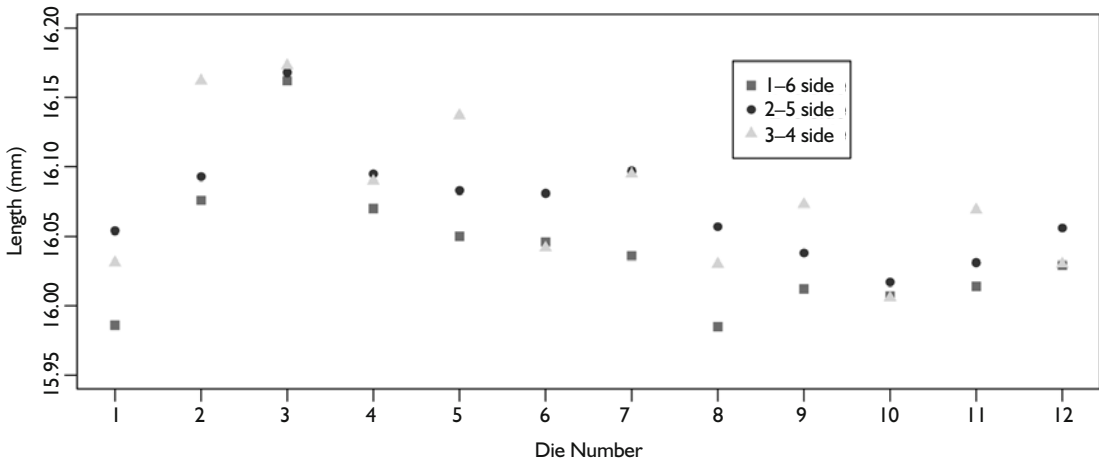


Figure 6. Measurement of axis length, in millimeters, for all 12 dice on all three axes. (The main axes of a standard die are the 1–6 axis, the 2–5 axis, and the 3–4 axis.) The 1–6 axis is consistently and significantly shorter than the other two axes.

does not automatically imply the cause Pearson suggests. The new data presented here show that, even though fives and sixes jointly appear slightly more often than would be expected under the fair hypothesis, fives and sixes do not both have an individual probability larger than 1/6.

The number of throws needed to observe these probability departures from fair is high and a testament to Weldon's perseverance. In Weldon's original data, the observed probability of a five or six was 33.77%, and at least 100,073 throws (or 8,340 throws of 12 dice) are needed to detect this departure from fair with 90% power at the 5% significance level. Here, the probability of throwing a six was determined to be 16.88%. At 90% power and the 5% significance level, 270,939 throws (or 22,579 throws of 12 dice) were needed to detect a departure as extreme.

The estimated probabilities for the six faces seen in "Results" might be explained by a mold for the plastic dice that is not perfectly cubic, with the one- and six-pip faces slightly larger than the faces with two and five pips. To further investigate this possibility, the dimensions of each of the axes of all 12 dice (i.e., the 1–6 axis, the 2–5 axis, and the 3–4 axis) were measured with an accurate digital micrometer. The results are shown in Figure 6, where it is seen that the 1–6 axis is consistently shorter than the other two, thereby supporting the hypothesis that the faces with one and six pips are larger than the other faces. A two-way ANOVA model (axis length modeled on axis number and die number) adjusted for multiple comparisons also showed that the 1–6 axis was significantly shorter than both of the other axes (by around 0.2%).

Pearson's suggestion for the cause of biased dice would also indicate that if a die were considered a success when four, five, or six pips were showing, that event should have a measurably higher probability than the complementary event (one, two, or three pips showing). However, the data obtained here indicate almost perfect balance, with $\hat{p}_{4,5,6} = 0.5001$. Perhaps with further investigation, Pearson may have unearthed evidence to support his claim that the pip-weight imbalance led to Weldon's data, but current observations suggest minor imperfections in the individual cubes may overshadow any effect due to carved-out pips. Interestingly, dice used in casinos have flush faces, where the pips are filled in with a plastic of the same density as the surrounding material and are precisely balanced. It would be reasonable to assume these dice would produce results in accordance with all fair hypotheses.

Further Reading

Iversen, Gudman R., Willard H. Longcor, Frederick Mosteller, John P. Gilbert, and Cleo Youtz. 1971. Bias and runs in dice throwing and recording: A few million throws. *Psychometrika* 36(1):1–19.

Kemp, A.W., and C.D. Kemp. 1991. Weldon's dice data revisited. *The American Statistician* 45(3):216–222.

Nagler, J., and P. Richter. 2008. How random is dice tossing? *Physical Review E* 78(3):036207.

Pearson, Karl. 1900. On the criterion that a given system of derivations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine* 5(50): 157–175.

van der Heijdt, L. 2003. *Face to face with dice: 5,000 years of dice and dicing*. Groningen, The Netherlands: Gopher Publishers.

Weldon's Dice, Automated, *www.youtube.com/watch?v=95EErdouO2w*.