Computational Statistical Experiments:

STAT 218 - 07S2 (C) Student Projects Report UCDMS 2008/5

This document was completed with the fiscal support from external grants to Dominic Lee. It was typeset by Zhu Sha. All the projects were supervised by Raazesh Sainudiin while he coordinated the second-year course called STAT 218 - 07S2 (C) during Semester Two of 2007 (16/07/2007-15/11/2007) at the

Department of Mathematics and Statistics,
University of Canterbury,
Private Bag 4800,
Christchurch, New Zealand.

Contents

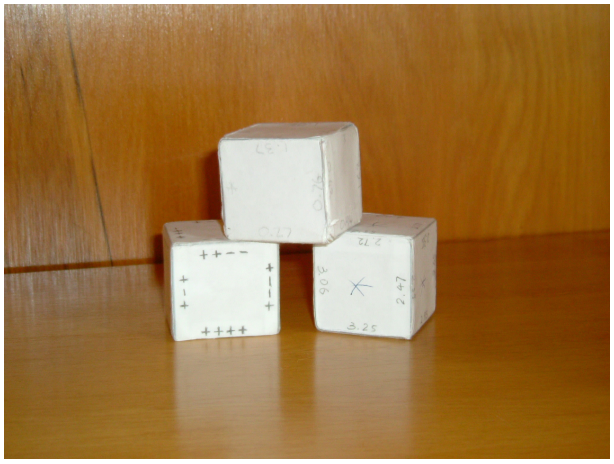
Chapter 1

Investigation of a Statistical Simulation from the 19th Century

Brett Versteegh
and ZHU Sha (Joe)

Abstract

This project is designed to investigate Sir Francis Galton's statistical dice experiment. We constructed Galton's dice according to his prescriptions and tested the null hypothesis that the outcomes from these dice do indeed follow a discrete approximation to the normal distribution with median error one. The inverse distribution function sampler and Chi Squared test are the statistical methodologies employed in this project.



Introduction & Motivation

The report will firstly cover the background and motivation of this project. Secondly, the methodologies used will be explained before outlining the results and subsequent conclusion found by undertaking this experiment. Finally, a potential modification to Galton's method will be examined as a means of sampling from a standard normal distribution.

Background - Francis Galton

Born in 1822, Francis Galton was considered by many, at an early stage, to be a child prodigy. By the age of two, he could read; at five, he already knew some Greek, Latin and long division.

After his cousin, Charles Darwin, published *The Origin of Species* in 1859, Galton became fascinated by it and thus devoted much of his life to exploring and researching aspects of human variation. Galton's studies of heredity lead him to introduce the statistical concepts of regression and correlation. In addition to his statistical research, Galton also pioneered new concepts and ideologies in the fields of meteorology, psychology and genetics.

Background - Statistical Dice

This experiment came about from Galton's need, as a statistician, to draw a series of values at random to suit various statistical purposes. Dice were chosen as he viewed them to be superior to any other randomisation device. Cards and marked balls were too tedious to be continually shuffled or mixed following each draw, especially if the required sample size was large.

The dice he created made use of every edge of each face which allowed for 24 equal possibilities as opposed to the six of a normal die.

For further details on Galton's experiment, please refer to his article "Dice for Statistical Experiments"; *Nature* (1890) No 1070, Vol 42 (This article is available free for download. Please refer to the references section for the website.)

Motivation

The motivation behind this project is to reconstruct Galton's dice using the methods outlined in his 1890 *Nature* article "Dice for Statistical Experiments"and then harness the power of modern computers to determine how effective this technique was for simulating random numbers from the following distribution.

Galton outlines that the samples were taken from a normal distribution with mean zero and median error one. We shall call this distribution Galton's Normal distribution or GN. However, for the experiment

to work, we must use a discrete approximation of the normal distribution, which we will define as Galton's Discrete Normal or GDN. Both will be formally explained in the Methodology section.

To determine the success of this experiment, we formulate the following question as a statistical hypothesis test: "Are our sampled values taken independently and identically from an appropriate discrete distribution which approximates Galton's normal distribution?"

Materials & Methods

Experiment Process

In order to recreate Galton's Dice Experiment, we have chosen to replicate the design he explains in his *Nature* article.

Creating the Dice

We chose to use rimu as it was readily available and inexpensive, unlike the mahogany that Galton had access to. As per his specifications, the wood was cut into six cubes of 1.25 inches (3.2 cm) wide, high and deep, before being covered in a paper template that was designed to fit tightly around the wood. The paper was adhered using general PVA glue.

The only change to Galton's original specification was that we chose to write the values to two decimal places on the faces, as opposed to one decimal place. This was to ensure a higher level of precision when plotting the results.

Collecting the Data

The experiment was carried out by shaking all of the first three dice (dice 1) at once and rolling them across the flat surface of a table top. We interpreted Galton's terminology of the values that "front the eye"to be the results that one can see by looking directly down on top of the dice. The three dice were then lined up into a row and the values called out and entered onto a Notepad document. We used the following formula to calculate the optimal number of trials needed for our investigation:$f(x)_{min} * sample\ size \approx 5$, where $f(x)_{min}$ is the smallest probability for the discrete distribution.

The same rolling process was then performed for dice 2 (two dice at once) and 3 (only one die) with the single exception that we did not need to roll these dice as many times as dice 1.

Statistical Methodology

Firstly, we will define Galton's Normal distribution. As derived from an article published in *Statstical Science*[1], Galton's Normal Distribution has a mean of zero but the variance is not one. Instead, Galton's sample is taken from a half-normal distribution with a "probable error"(median error) of one. This implies that the probability between zero and one is a quarter, allowing us to solve the following equation to determine the variance:

$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{x^2}{2\sigma^2})$$
$$\frac{1}{4} = \int_0^1 \phi(x)dx$$
$$\frac{1}{4} = \int_0^1 \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{x^2}{2\sigma^2})dx$$
$$\sigma = 1.4826$$

$\therefore$ GN $\sim$ N (0, 1.48262)

Secondly, we must determine how Galton calculated the values[2] to use on his dice. It was our assumption that he used the midpoints of a set of intervals that partition [0, 1] and we undertook the following processes to confirm this.

We divided the interval [0.5 1] equally into 24, with the last 3 intervals further divided into 24 subintervals. In total, this gave us 21 + 24 intervals to allocate along the y-axis. The midpoint of each interval was taken in order to compute its corresponding x value under the inverse CDF map.

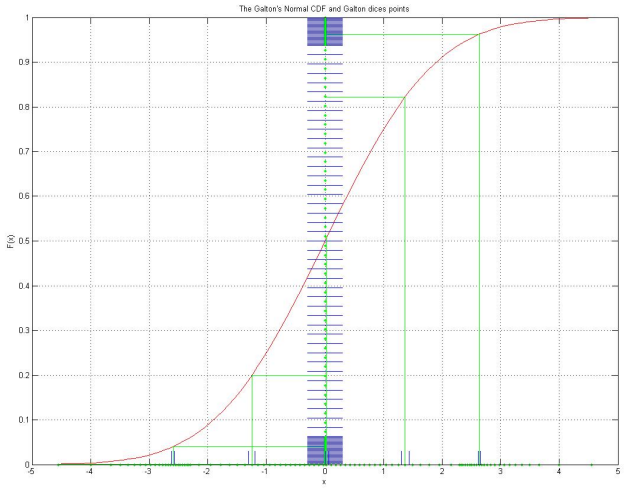


Figure 1.1: Plot showing the midpoints mapping back to specific values on the x axis.

The easiest way to do this would have been to evaluate the inverse CDF function at the midpoints.

[1]Stochastic Simulation in the Nineteenth Century. *Statistical Science* (1991) Vol 6, No 1, pg 94.
[2]See Appendix B.

However, a closed form expression for the inverse CDF does not exist for a Normal distribution. Thus, we applied numerical methods to solve for x (Newton's method).

We believe the midpoint assumption was correct, as the mapped values are very close to Galton's actual figures and the differences can be attributed to an imprecise value for the standard deviation.

Thirdly, we can now determine Galton's discrete approximation to the Normal. This is necessary as the values drawn from throwing Galton's dice come from a discrete distribution, not the continuous Galton Normal. In doing this, we are also able to define our null hypothesis formally: $H_0 : x_1, x_2, \ldots, x_n IID \sim GDN$ Galton's Discrete Normal (GDN) is an approximation to Galton Normal (GN).

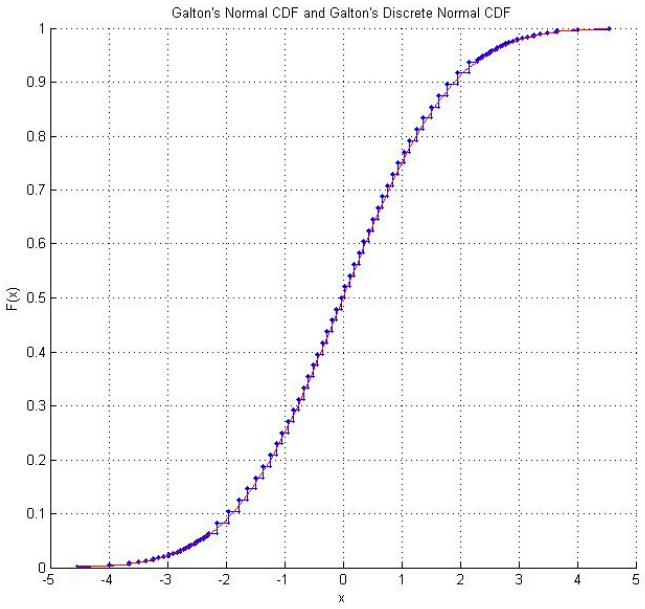


Figure 1.2: Plot showing both the GN and GDN CDFs. They are very similar.

Fourthly, as the distribution is now discrete, we can apply the Chi Squared Test to evaluate our null hypothesis. The test used had the following parameters: Degrees of Freedom: $90 - 1 = 89$ $\alpha = 0.05$; Critical Value = 112.

Results

Once the experiment was complete and the results collated, they were run through a methodological tester to ensure all values were correct. Testing the data involved running all our sampled values through a `Matlab` function which checked each number against Galton's 45 possible values. Any values that did not match were outputted as ones and the erroneous data were removed before a graph was plotted to measure how well our experiment sampled from GDN.

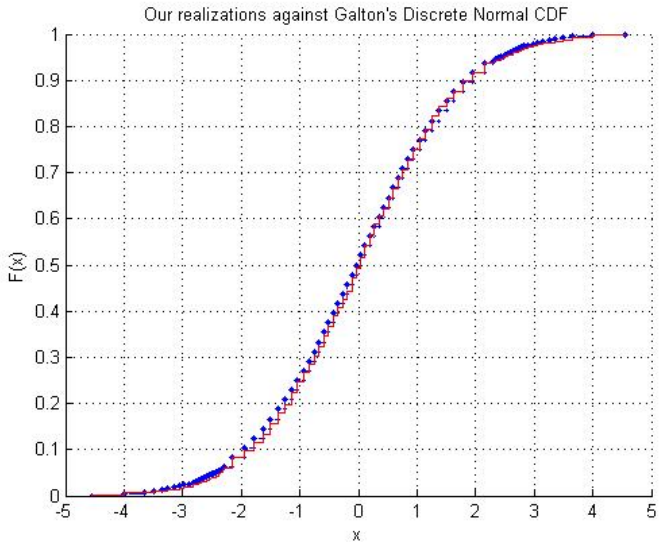


Figure 1.3: Plot showing the empirical DF of our results against GDN. Our values take on a stair case appearance and are very close to GDN. The main deviations occur mostly in the tails.

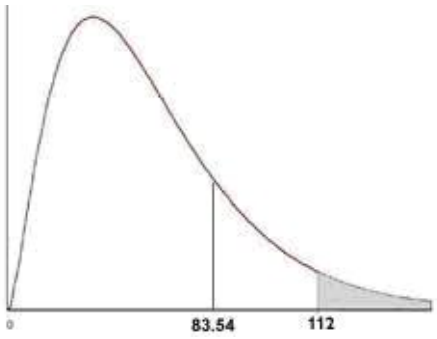
Chi Squared Test

A Chi Squared test was then performed on the data and the results[1] are summarised below.

[1]For the full table, please see Appendix A.

	Data Values	Observed Count	Expected Count	$(O-E)^2/E$
	-4.55	5	5.046875	0.000435372
	-4	7	5.046875	0.755853328
	-3.65	5	5.046875	0.000435372
	3.49	6	5.046875	0.180001935
	...	...	...	...
	3.65	8	5.046875	1.727989551
	4	11	5.046875	7.022107198
	4.55	5	5.046875	0.000435372
Total		1938	1938	
Chi Test Result				83.54798762

$$T = \Sigma_{i=1}^{90} \frac{(Obseverd - Expected)^2}{Expected} = 83.548$$



Conclusion

We cannot reject H0 at = 0.05 because the observed test statistic is outside the rejection region. In relation to our statistical question, this means that there is insufficient evidence to suggest that our sample is not from GDN.

Potential Modification

Since the standard normal distribution is more common in all areas, we wanted to convert Galton's Dice into a new set which can be used for simulating the standard normal distribution.

In his experiment, Galton took the mid-point of each probability interval, and then found the corresponding x- values. Instead of applying a tedious calculation to find the x-values, we took a z-value table, and found

the corresponding z-values to the upper bound of those intervals. This enables the creation of two new dice[2]:

Dice (1)	0.05	0.10	0.15	0.21	0.27	0.32
	0.37	0.43	0.49	0.55	0.61	0.67
	0.74	0.81	0.89	0.97	1.05	1.15
	1.26	1.38	1.53	*	*	*
Dice (2)	1.56	1.58	1.60	1.62	1.65	1.68
	1.70	1.73	1.76	1.79	1.83	1.86
	1.90	1.94	1.99	2.04	2.09	2.15
	2.23	2.31	2.42	2.56	2.80	4.00

Through `Matlab`, we were able to map the data gathered during our original experiment into the values shown in previous table, corresponding to the standard Normal, and develop the following plot:

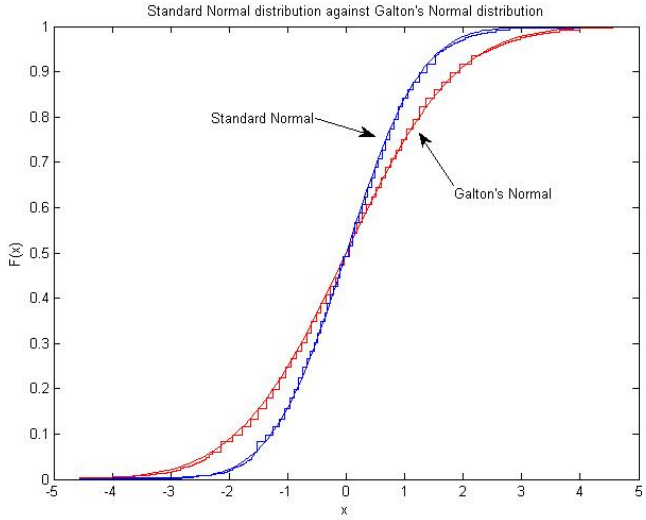


Figure 1.4: Plot showing the Standard Normal Distribution against Galton's Normal Distribution.

[2]Tables showing the new values for dice 1 & 2. The third dice can remain the same as Galton's.

Author Contributions

Brett - Constructed dice, gathered majority of the data results, constructed report, conducted spell/grammar check.

Joe - Wrote up `Matlab` code to analyse and plot data, entered in data results, constructed presentation and discovered a modification to Galton's experiment.

References

Dice for Statistical Experiments. *Nature* (1890) Vol 42, No 1070

Stochastic Simulation in the Nineteenth Century. *StatisticalScience* (1991) Vol 6, No 1

http://www.galton.org

http://www.anu.edu.au/nceph/surfstat/surfstat-home/tables/normal.php

Appendix A

	Data Values	Observed Count	Expected Count	$(O-E)^2/E$
	-4.55	5	5.046875	0.000435372
	-4	7	5.046875	0.755853328
	-3.65	5	5.046875	0.000435372
	-3.49	1	5.046875	3.245017415
	-3.36	3	5.046875	0.830156734
	-3.25	3	5.046875	0.830156734
	-3.15	3	5.046875	0.830156734
	-3.06	4	5.046875	0.217153638
	-2.98	4	5.046875	0.217153638
	-2.9	3	5.046875	0.830156734
	-2.83	8	5.046875	1.727989551
	-2.77	5	5.046875	0.000435372
	-2.72	3	5.046875	0.830156734
	-2.68	3	5.046875	0.830156734
	-2.64	3	5.046875	0.830156734
	-2.59	6	5.046875	0.180001935
	-2.55	4	5.046875	0.217153638
	-2.51	5	5.046875	0.000435372
	-2.47	4	5.046875	0.217153638
	-2.43	6	5.046875	0.180001935
	-2.39	10	5.046875	4.861116486
	-2.35	6	5.046875	0.180001935
	-2.32	8	5.046875	1.727989551
	-2.29	6	5.046875	0.180001935
	-2.15	47	40.375	1.087074303
	-1.95	28	40.375	3.792956656
	-1.78	34	40.375	1.006578947
	-1.63	33	40.375	1.347136223
	-1.5	45	40.375	0.529798762

	Data Values	Observed Count	Expected Count	$(O-E)^2/E$
	-1.37	46	40.375	0.783668731
	-1.25	37	40.375	0.282120743
	-1.14	48	40.375	1.44001548
	-1.04	48	40.375	1.44001548
	-0.94	35	40.375	0.715557276
	-0.85	34	40.375	1.006578947
	-0.76	34	40.375	1.006578947
	-0.67	41	40.375	0.009674923
	-0.59	49	40.375	1.84249226
	-0.51	37	40.375	0.282120743
	-0.43	44	40.375	0.325464396
	-0.35	33	40.375	1.347136223
	-0.27	36	40.375	0.474071207
	-0.19	36	40.375	0.474071207
	-0.11	55	40.375	5.297600619
	-0.03	38	40.375	0.139705882
	0.03	45	40.375	0.529798762
	0.11	53	40.375	3.947755418
	0.19	48	40.375	1.44001548
	0.27	40	40.375	0.003482972
	0.35	35	40.375	0.715557276
	0.43	32	40.375	1.737229102
	0.51	42	40.375	0.065402477
	0.59	41	40.375	0.009674923
	0.67	46	40.375	0.783668731
	0.76	35	40.375	0.715557276
	0.85	38	40.375	0.139705882
	0.94	45	40.375	0.529798762
	1.04	44	40.375	0.325464396
	1.14	43	40.375	0.170665635

	Data Values	Observed Count	Expected Count	$(O-E)^2/E$
	1.25	55	40.375	5.297600619
	1.37	38	40.375	0.139705882
	1.5	35	40.375	0.715557276
	1.63	32	40.375	1.737229102
	1.78	42	40.375	0.065402477
	1.95	33	40.375	1.347136223
	2.15	40	40.375	0.003482972
	2.29	3	5.046875	0.830156734
	2.32	4	5.046875	0.217153638
	2.35	3	5.046875	0.830156734
	2.39	4	5.046875	0.217153638
	2.43	6	5.046875	0.180001935
	2.47	5	5.046875	0.000435372
	2.51	3	5.046875	0.830156734
	2.55	8	5.046875	1.727989551
	2.59	1	5.046875	3.245017415
	2.64	7	5.046875	0.755853328
	2.68	5	5.046875	0.000435372
	2.72	4	5.046875	0.217153638
	2.77	4	5.046875	0.217153638
	2.83	6	5.046875	0.180001935
	2.9	5	5.046875	0.000435372
	2.98	5	5.046875	0.000435372
	3.06	6	5.046875	0.180001935
	3.15	5	5.046875	0.000435372
	3.25	4	5.046875	0.217153638
	3.36	5	5.046875	0.000435372
	3.49	6	5.046875	0.180001935
	3.65	8	5.046875	1.727989551
	4	11	5.046875	7.022107198

	Data Values	Observed Count	Expected Count	$(O-E)^2/E$
	4.55	5	5.046875	0.000435372
Total		1938	1938	
Chi Test Result				83.54798762

Appendix B

Table 1

0.03	0.51	1.04	1.78
0.11	0.59	1.14	1.95
0.19	0.67	1.25	2.15
0.27	0.76	1.37	*
0.35	0.85	1.50	*
0.43	0.94	1.63	*

Table 2

2.29	2.51	2.77	3.25
2.32	2.55	2.83	3.36
2.35	2.59	2.90	3.49
2.59	2.64	2.98	3.65
2.43	2.68	3.06	4.00
2.47	2.72	3.15	4.55

Table 3

++++	+–+	–++	+-+
+++-	+—	–+-	+–
++-+	-+++	—+	-++
++–	-++-	—-	-+-
+-++	-+-+	+++	–+
+-+-	-+–	++-	—