9/13/2014

Randomness

Research question: How are the results of the chi-square test and runs tests, affected by a change in the string length of repeating decimals?

Name: Tiffany McKenzie

Word Count: 3958

Candidate Number: 000512 -0082

Supervisor: Miss Suzanne Doering

Subject: Mathematics

May 2015 Session

Abstract

Randomness has been a topic of frequent mathematical and philosophical discussion for ages. However, in today's modern world, randomness has become especially important in numerous fields, such as in the natural sciences and engineering.

In this essay, the effect of the length of the repeated pattern (the string length) on the results of the chi-square and runs tests for randomness were investigated .This research was carried out by testing arbitrarily chosen repeating decimals to 1000 decimal places against the results of the randomness control, pi. After the repeating decimals were tested, their results were used to construct two hypotheses.

The first hypothesis is that according to the chi-square test, repeating decimals become more random with longer string lengths. The results of the investigation from the chi-square test, showed that as the string length of a decimal approaches infinity, the chi-square value approaches zero, indicating perfect randomness and uniform distribution. On the other hand, the results also showed that a string length approaching infinity can cause the chi-square value to show infinity-perfect non-randomness. These results could therefore help mathematicians to better understand numbers with infinite string lengths.

The second hypothesis is that according to the runs test, a repeating decimal with either too short or too long a string length is non-random. This hypothesis has been confirmed in the general solution. Apart from the hypothesis, it was also discovered that a repeating decimal with string length one, has an undefined runs test result. This was therefore acknowledged in the conclusion as one of the weaknesses of this test.

However, despite their weaknesses, each test allows for randomness to be viewed from different perspectives and therefore facilitates a better understanding of the characteristics of random distributions.

(285 words)

Table of Contents

Introduction

The world is governed by chance. Randomness stalks us every day of our lives.
—*Paul Auster*

So much of life, it seems to me, is determined by pure randomness.
—*Sidney Poitier*

In this extended essay, the weaknesses and the effect of string length in repeating decimals on the results of the chi-square and runs tests for randomness are investigated.

Random processes, processes seemingly uninfluenced by anything, occur every day. The use of random numbers in science, technology, mathematics and engineering is not as new as one may think. For ages, man has used randomness to make important decisions. For example, making a decision with a coin toss. A few of the biggest areas in which randomness is important are in gaming/lotteries, computer science (specifically encryption) and in engineering for simulating real-world phenomena. It is therefore, evident that the importance of random numbers extends past academia, into our everyday lives.

In order to make these calculations as accurate as possible, the random numbers used in the calculations need to be of very high quality. For this, there are random number generators. In fact, many can be found online and even programmed in scientific calculators. However, many of these 'randomly' generated numbers are not that 'random'. Flaws in their generation can be detected by randomness tests and test suites. These tests are very important for testing the quality of random numbers and therefore, calculations in science, engineering etc. are very dependent on the thoroughness of these tests. In this essay, two of the most commonly used tests for randomness will be investigated, making the results conclusions of this paper relevant to scientists, mathematicians and engineers.

Approaching the Problem

Non-randomness can be detected by theoretical statistical tests such as the chi-square test and the runs test. These two tests were chosen based their abilities to detect different types of non-randomness. This will help to give better coverage, making non-randomness less difficult to identify and will also be helpful in deciding which test is best suited for detecting non-randomness in repeating decimals.

In order to observe patterns and quantitatively investigate the apparent difference in randomness of different repeating decimals, the tests will be carried out on the same list of decimals and also to make the number of digits tested constant. This will help me to make fairer comparisons with the results. The irrational number, pi to 1000 decimal places, will be used as a control or standard for randomness.

The repeating decimals which will be investigated have been chosen based on their variations in the length of their repeated pattern (which in this essay will be called a 'string'). For example, the repeating decimal 0.3333 has a string length of 1, while 0.141414 has a string length of 2. Also, for the sake of simplification, only digits after the decimal point will be considered in the tests for randomness.

Repeating Decimals to be Tested: A string is the repeated pattern in the decimal. The string length is therefore, the length of the repeated pattern. The decimals which will be tested (see table below) have been arbitrarily chosen and will all be tested to 10000 digits after the decimal point, so that they may be comparable to the control results of π to 1000 decimal places.

Repeating Decimal	String Length
$0.\overline{3}$	1
$0.\overline{12}$	2
$0.\overline{456}$	3
$0.\overline{6789}$	4
$0.\overline{87654}$	5
$0.\overline{765432}$	6
$0.\overline{0123456}$	7
$0.\overline{01234567}$	8
$0.\overline{012345678}$	9
$0.\overline{0123456789}$	10
Π	∞

Test One: The Chi-Square Test

The chi-square test aims to detect non-randomness by estimating how closely an observed distribution matches an expected distribution.

The null hypothesis, H_0: There is no difference between the observed frequency (O) and the expected frequency (E)

The alternate hypothesis, H_1: There is a difference between the observed frequency (O) and the expected frequency (E).

The level of significance, α, determines the sensitivity of the test. For this essay the level of significance, α, will be equal to 0.05. This choice was made arbitrarily, although it also one of the most commonly used values of α. A 0.05 level of significance means that the null hypothesis is rejected 5% percent of the time when it is actually true.

The degree of freedom is the number of outcomes that are free to vary. For example, if a person has gets a result of two heads and a tail in three coin tosses, then the result of the first toss could be a head, the second-a tail and then the third would then have to be a head. Only two of the tosses are open to variation, while the third is dependent on the previous results. There are 2 degrees of freedom in the triple coin toss and the same logic follows in this example. The degree of freedom is equal to the number of possible outcomes minus one. Therefore, the degree of freedom in this example is equal to $10 - 1 = 9$.

Since, the chi-square test is based on the concept of randomness being associated with uniformed distribution, the expected number of each digit, 0 to 9, is equal to the number of digits, 1000, divided by the number of possible outcomes, 10. That is $\frac{1000}{10} = 100$.

The observed number of each digit is simply the number of times digit is found in the decimal tested. In the first decimal, $0.\overline{3}$ to 1000 decimal places, the observed number of threes is 1000.

The deviation is how far the observed values are from the expected values. That is therefore equal to the expected value of each digit subtracted from the observed value of the respective digit. For the digit 3, the deviation is equal to $1000 - 100 = 900$. For all the other digits (1, 2,4,5,6,7,8,9, 0), the deviation is equal to -100.

Although the calculations for the deviations are accurate, the sum of deviations may not give a clear picture of how close or far the decimal is from having a uniform distribution of digits. This is because if a negative deviance is added to a positive deviance, then the sum of the deviances will be less than the actual deviance of the decimal from uniform distribution. In order to address this problem, the deviance can be squared in order for all the measures of deviances to be positive. The deviances squared can then be summed, to get a quantitative measure of how much the decimal deviated.

However, in order to fairly measure how much it deviated, the sum of the deviances squared needs to be compared to the sum of expected values. This can be done by dividing this sum of deviations by the sum of the expected values.

In following this reasoning, the formula for the chi-square value is: $\chi^2 = \sum \frac{(O-E)^2}{E}$. For the decimal $0.\overline{3}$, $\chi^2 = 9 \cdot 100 + 8100 = 9000$.

Using a chi-square reference table, the critical value of χ^2 can be found with the level of significance, 0.05 and the degree of freedom, 9. According to following reference table, the critical value is equal to 16.919.

Table X The χ^2 Distribution

df	.25	.10	.05	.025	.01	.005
8	10.219	13.362	15.507	17.535	20.090	21.955
9	11.389	14.684	16.919	19.023	21.666	23.589

Source: From *The Handbook of Statistical Tables*, by D. B. Owen, p. 50. Copyright © 1962 Addison Wesley Longman Publishing Co. Reprinted by permission of Addison Wesley Longman.

[1]

[1] Taken form Palomar College, https://www2.palomar.edu/users/rmorrissette/Lectures/Stats/ChiSquare/TableX.jpg

The calculations for the chi-square value of $0.\overline{3}$ to 1000 decimal places, are tabulated below.

Digits	0	1	2	3	4	5	6	7	8	9
Expected Frequency , E	100	100	100	100	100	100	100	100	100	100
Observed Frequency , O	0	0	0	1000	0	0	0	0	0	0
Deviation $(\mathrm{O}-\mathrm{E})$	-100	-100	-100	900	-100	-100	-100	-100	-100	-100
Deviation Squared $(\mathrm{O}-\mathrm{E})^2$	10000	10000	10000	810000	10000	10000	10000	10000	10000	10000
Deviation Squared divided by the Expected Frequency $\frac{(\mathrm{O}-\mathrm{E})^2}{\mathrm{E}}$	100	100	100	8100	100	100	100	100	100	100

The probability, p, of obtaining the observed sample results when the null hypothesis is actually true can be calculated using an excel chi-square function.

p	0
Reject H_0 Hypothesis	Yes

If the $\chi^2 > CV$ and $p < \alpha$, then the null hypothesis can be rejected. In this example, 9000 is greater than the critical value, 16.919 and the probability, 0, is less than the level of significance, 0.05. These results reject the null hypothesis and it can therefore be interpreted that according to the chi-square test, decimal $0.\overline{3}$, is not random.

(See appendix one for calculation tables for other repeating decimals of varying string lengths.)

Constructing Hypothesis for the Chi-Square Test

Table Showing the Results of the Chi-Square Test

Decimal	String Length, n	H_0	p Value
$0.\overline{3}$	1	×	0
$0.\overline{12}$	2	×	0
$0.\overline{456}$	3	×	0
$0.\overline{6789}$	4	×	0
$0.\overline{87654}$	5	×	1.7241×10^{-209}
$0.\overline{765432}$	6	×	1.003×10^{-137}
$0.\overline{0123456}$	7	×	1.08397×10^{-86}
$0.\overline{01234567}$	8	×	9.97615×10^{-49}
$0.\overline{012345678}$	9	×	8.69644×10^{-20}
$0.\overline{0123456789}$	10	✓	1
Control: Π	∞	✓	0.853049013

Key:

- × Means that the null hypothesis was rejected.
- ✓ Means that the null hypothesis was accepted.

χ^2. Hypothesis

It can be observed that as the string length increases, the probability of the decimal being random increases. However, the control variable, π, which should represent show represent a number which has passed the chi-square test for randomness, has a surprisingly lower probability of being random than the repeating decimal, $0.\overline{0123456789}$'s probability. This could be attributed to the fact that the value of pi which was tested is only a decimal rational approximate of the true value of pi. Therefore, not all the digits of pi are represented in the pi decimal tested, whereas all of the digits in a string are present in the repeating decimal $0.\overline{0123456789}$. However, the jump from 8.69644×10^{-20} to 1, is quite big and therefore, $0.\overline{0123456789}$, could be a special case for the chi-square test.

Apart from this anomaly with the control, π, in general, it can be observed that the p value increases with an increase in the string length, n. It can therefore, be hypothesized that, according to the chi-square test, as the string length increases, the decimal becomes more random.

General Solution One- For the Chi Square Test Hypothesis

Let N be the number of digits tested in the repeating decimal.

Let n be the number of digits in a string.

Let c_i be how many of each respective digit in the in a string. For example, in the repeating decimal, $0.\overline{456}$, $c_0 = 0$, $c_1 = 0$, $c_2 = 0$, $c_3 = 0$, $c_4 = 1$, $c_5 = 1$, $c_6 = 1$, $c_7 = 0$, $c_8 = 0$, and $c_9 = 0$

Recall [2] $\chi^2 = \frac{(O-E)^2}{E}$

The expected number of each digit is equal to the number of digits tested divided by the number of possible digits or outcomes (0 to 9). $E = \frac{N}{10}$

The observed number of each digit is equal to the number of times each digit appears in a string multiplied by the number of string repeats. $O = \frac{c_i N}{n}, \in \mathbb{Z}^+$

Therefore, $\chi^2 = \frac{\left(c_0N/n - N/10\right)^2 + \left(c_1N/n - N/10\right)^2 + \left(c_2N/n - N/10\right)^2 + \cdots \left(c_9N/n - N/10\right)^2}{N/10} = \frac{10\sum_{i=0}^{9}\left(c_iN/n - N/10\right)^2}{N}$

$\left(\frac{c_iN}{n} - \frac{N}{10}\right)^2 = \frac{c_0^2N^2}{n^2} - \frac{2c_0^2N^2}{10n} + \frac{N^2}{100}$

Since there are ten digits for which this expansion will be done, there will be ten of each term from the above expansion.

This will give ten of the first term: $\frac{N^2}{n^2} \cdot \sum_{i=0}^{9} c_i^2$,

Ten of the second term: $\frac{-2N^2}{10n} \cdot (\sum_{i=0}^{9} c_i)$. Since the digits make up a string, then it can be deduced that the sum of the number of each digit in the string is equal to the length of the string.

$\sum_{i=0}^{9} c_i = c_0 + c_1 + c_2 + \cdots c_9 = n$. Therefore, ten of the second term is $\frac{-2N^2}{10n} \cdot n = \frac{-2N^2}{10}$

And ten of the last term: $\frac{N^2}{100} \cdot 10 = \frac{N^2}{10}$

Therefore, $\chi^2 = \frac{\frac{N^2}{n^2}\sum_{i=0}^{9} c_i^2 - \frac{2N^2}{10} + \frac{N^2}{10}}{\frac{N}{10}} = \frac{10N^2\sum_{i=0}^{9} c_i^2}{Nn^2} - \frac{10N^2}{10N}$

$$\chi^2 = \frac{10N\sum_{i=0}^{9} c_i^2}{n^2} - N$$

[2] Taken from http://www.itl.nist.gov/div898/handbook/eda/section3/eda35f.htm

In order for the p value to be as high as possible (greater than the 0.05 level of significance), the χ^2 value needs to be at a minimum. To get χ^2 at a minimum, the denominator has be at a maximum. Since the denominator is n, the string length, it could be deduced that as the string length increases, so does the apparent randomness of the repeating decimal.

However, a new problem then arises. In order for n to increase, either c_0, c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8, or c_9 increases or any combination of the terms increases. An increase in any of the c_i values, also implies that there will an increase in the value of $\sum_{i=0}^{9} c_i^2$, and therefore an increase in the numerator, counteracting the minimizing effect of increasing the denominator.

With both the numerator and the denominator increasing with a longer string length, the χ^2 value, then becomes dependent on which of the two (numerator or denominator) increases at a faster rate. It is however, very difficult to tell which increases faster from the information given in the deduced value of χ^2.

In order eliminate one of the unknowns in the formula, n will be written in terms of c_i.

$n = \sum_{i=0}^{9} c_1$, and hence

$$n^2 = \sum_{i=0}^{9} {c_i}^2 + c_0 \sum_{i=1}^{9} c_i + c_1\left(c_0 + \sum_{i=2}^{9} c_i\right) + c_2\left(\sum_{i=1}^{1} c_i + \sum_{i=3}^{9} c_i\right) + c_3\left(\sum_{i=1}^{2} c_i + \sum_{i=4}^{9} c_i\right) + c_4\left(\sum_{i=1}^{3} c_i + \sum_{i=5}^{9} c_i\right) + c_5\left(\sum_{i=1}^{4} c_i + \sum_{i=6}^{9} c_i\right) + c_6\left(\sum_{i=1}^{5} c_i + \sum_{i=7}^{9} c_i\right) + c_7\left(\sum_{i=1}^{6} c_i + \sum_{i=8}^{9} c_i\right) + c_8\left(\sum_{i=0}^{9} c_i + c_9\right) + c_9 \sum_{i=0}^{8} c_i\ .$$

Therefore, $n^2 > \sum_{i=0}^{9} c_i^2$.

Although, this implies that the denominator increases faster than the numerator, the coefficients of $\sum_{i=0}^{9} c_i^2$, in the numerator also have to be considered. In order to more concretely prove or disprove the hypothesis, the limit of χ^2, as $n \to \infty$, can be evaluated.

Since $n = \sum_{i=0}^{8} c_i$, when $n \to \infty$, either c_0, c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8, or c_9 approaches infinity or any combination of the terms are approaching infinity.

CASE #1:

In this case, only one of the terms, for example c_0, approaches infinity while the other terms remain fixed. The limit of χ^2 can be found by $\lim_{c_0\to\infty} \frac{10N\sum_{i=0}^{9} c_i^2}{(\sum_{i=0}^{9} c_1)^2} - N$

When χ^2 is multiplied by $\frac{1/c_0^2}{1/c_0^2}$, $\lim_{c_0\to\infty} \frac{10N\sum_{i=0}^{9} c_i^2 / c_0^2}{(\sum_{i=0}^{9} c_i)^2 / c_0^2} - N = \lim_{c_0\to\infty} \frac{10N\sum_{i=0}^{9} c_i^2 / c_0^2}{(\sum_{i=0}^{9} c_i)^2 / c_0^2} - N.$

Evaluate $\lim_{c_0\to\infty} \frac{\sum_{i=0}^{9} c_i^2 / c_0^2}{(\sum_{i=0}^{9} c_i)^2 / c_0^2}$.

The numerator will be equal to:

$({c_0}^2/{c_0}^2) + ({c_1}^2/{c_0}^2) + ({c_2}^2/{c_0}^2) + ({c_3}^2/{c_0}^2) + ({c_4}^2/{c_0}^2) + ({c_5}^2/{c_0}^2) + ({c_6}^2/{c_0}^2) + ({c_7}^2/{c_0}^2) + ({c_8}^2/{c_0}^2) + ({c_9}^2/{c_0}^2)$

Since c_0→∞, while c_1 through to c_9 remain fixed, the value of, for example ${c_1}^2/{c_0}^2$ will converge to zero. This makes the numerator equal to 1+0+0+0+0+0+0+0+0+0=1

The denominator will be equal to $\sum_{i=0}^{9} c_i^2 / {c_0}^2 + c_0\sum_{i=1}^{9} c_i + \cdots c_9 \sum_{i=0}^{8} c_i / {c_0}^2$. That is

$1 + c_0\sum_{i=1}^{9} c_i + \cdots c_9 \sum_{i=0}^{8} c_i / {c_0}^2$

$c_0\sum_{i=1}^{9} c_i / {c_0}^2 = \sum_{i=1}^{9} c_i / c_0$

Sincec_1 through to c_9 are fixed, then their sum is fixed and $\lim_{c_0\to\infty} \sum_{i=1}^{9} c_i / c_0 = 0$

This can be applied to remaining parts of the denominator to give a denominator that is equal to

1+0+0+0+0+0+0+0+0+0+0=1

Therefore, as $c_0 \to \infty$, $\chi^2 = 10N \cdot 1 - N = 10N - N = 9N$. Since $n = \sum_{i=0}^{9} c_i$, as $c_0 \to \infty$, $n \to \infty$ and since $N = nk$ where $k \in \mathbb{Z}^+$ and where $k > 1$, N also approaches infinity, $N \to \infty$. Therefore, $9N \to \infty$.

This result is consistent with the theory of the chi-square test. If the number of only one of the digits approaches infinity then there will be bias towards that digit and the absolute value of the deviation of expected from observed frequency will be greater, and therefore, also approach infinity.

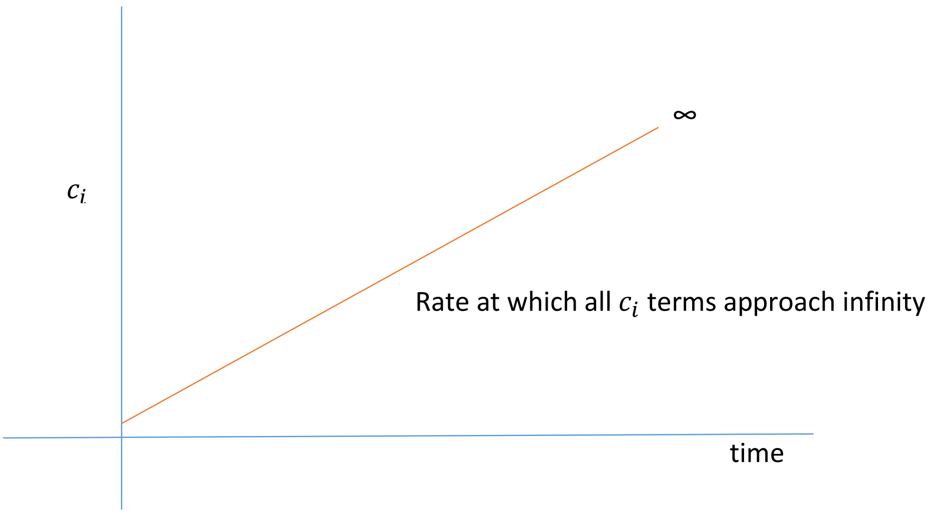
CASE#2

In this case a combination of c_i terms (excluding the combination which includes all of them) independently approaching infinity causes the string length, n, to approach infinity. Finding the exact limit of this will be very difficult. However, the result will be similar to that of CASE#1. In the case where some terms approach infinity while others do not, the value of chi-square is most likely cause the rejection of the null hypothesis. This is because the having some terms go to infinity will cause the data to have greater densities of certain digits and smaller densities of others. This will therefore cause the distribution of the data to not be uniform and in turn will most likely lead to the null hypothesis being rejected. Although, it is not impossible for the null hypothesis to be accepted in this case, it is highly unlikely.

CASE#3

In this case all of the c_i terms approach infinity. This can happen in two ways.

The first way is when the c_i terms approach infinity at the same rates, so that all c_i terms simultaneously 'meet' at infinity. In order for two lines to 'meet' at infinity they need to be either parallel or collinear. This is illustrated in the graph.



Since all the c_i terms approach infinity at the same rate, then the value of c_0 as $c_0 \to \infty$, is equal to the value of all the other c_i terms as they also approach infinity. With this information limit can be evaluated.

$$\lim_{c_i \to \infty} \frac{10N \sum_{i=0}^{9} c_i^2 / c_0^2}{(\sum_{i=0}^{9} c_i)^2 / c_0^2} - N$$

Finding the limit of $\frac{\sum_{i=0}^{9} c_i^2 / c_0^2}{(\sum_{i=0}^{9} c_i)^2 / c_0^2}$:

- Numerator: Since $c_0 = c_1$, ${c_1}^2 / {c_0}^2 = 1$. Therefore, $\lim_{c_0 \to \infty} \sum_{i=0}^{9} {c_i}^2 / {c_0}^2 = 10$
- Denominator:

  $\lim_{c_0 \to \infty} \sum_{i=0}^{9} {c_i}^2 / {c_0}^2 + c_0 \sum_{i=1}^{9} c_i + \cdots c_9 \sum_{i=0}^{8} c_i / {c_0}^2 = 10 +$
  $\lim_{c_0 \to \infty} c_0 \sum_{i=1}^{9} c_i + \cdots c_9 \sum_{i=0}^{8} c_i / {c_0}^2$

  For $c_0 \sum_{i=1}^{9} c_i / {c_0}^2 = c_0(c_1 + c_2 + c_3 + c_4 + c_5 + c_6 + c_7 + c_8 + c_9) / {c_0}^2 = 9$. This then repeats 10 times, making the limit of the denominator equal to $10 + 9 \cdot 10 = 100$

The limit of χ^2 for the first way of CASE#3, is therefore equal to $\frac{10N \times 10}{100} - N = N - N = 0$.

$$\lim_{c_i \to \infty} \frac{10N \sum_{i=0}^{9} c_i^2 / c_0^2}{(\sum_{i=0}^{9} c_i)^2 / c_0^2} - N = 0.$$

This explains why $0.\overline{0123456789}$ has a calculated χ^2 value of zero.

The second way in which all c_i terms can approach infinity, is if they all approach infinity but at different rates. It may or may not be possible to evaluate the limit of this χ^2 as $n \to \infty$.

Chi-Square Test Conclusion

Hypothesis one states that according to the chi-square test, as the string length increases, the decimal becomes more random. According to the reasoning and calculations made in the general solution, this hypothesis can be partially confirmed and partially disproved. The hypothesis is false when string length increases due to an increase in the number of a specific digit. This would cause a reveal a bias toward that digit, and the decimal would therefore not have a uniform distribution among its digits. An example of this can be found in the repeating decimal, $0.\overline{123456780000000000000000000000009}$. Although the string length in this decimal is long, it would be deemed non-random by the chi-square test.

Similarly, if a combination of digits (excluding a combination of all the digits) increases then there would be biases towards these digits and the decimal would therefore, most likely, not have a uniform distribution of digits. This can be seen in the repeating decimal, $0.\overline{123333333344444444444567}$. Therefore, although the string is long, the χ^2 can still be high.

On the other hand, in the case where all digits approach infinity at the same rate, the χ^2 value approaches zero. A χ^2 of zero, indicates perfect randomness. An example of this is seen in test results of the repeating decimal $0.\overline{0123456789}$.

Further Investigation:

Finding the limits to CASE#2 (a combination of c_i terms approach infinity) and CASE#3-part 2 (all c_i terms approach infinity but at different rates), would be extremely helpful in gaining a better understanding of how the chi-square test operates. It would also be interesting to link this data from rational repeating decimals to understanding irrational numbers. Irrational numbers are unpredictable and have no repeating patterns. They could therefore, be defined as numbers with infinite string length. Having a complete understanding of the effect of n approaching infinity on the value of the chi-square, would therefore, help mathematicians to better understand the distribution (uniform nor not uniform) of the digits of irrational numbers.

Test Two: Test for Runs Above and Below Mean:

The runs test is a statistical test which detects non-randomness based on binomial distribution.

The threshold value is an arbitrarily chosen reference value. The threshold which will be used in these calculations is the mean.

According the runs test, in a random data set, the probability that a value is greater than or less than the threshold value, follows a binomial distribution.

If a digit is greater than the threshold, then it is assigned the label '1'. On the other hand, if a digit is less than or equal to the threshold then it is assigned the label, '2'.

A run is series of increasing or decreasing values. It can also be defined as a series of successive zeroes or ones. For example, in the binary representation, 00 11 0 1 0000 1 00 1111 00, there are 9 observed runs.

The level of significance, α, of 0.05, will be used in these calculations.

The null hypothesis, H_0 states that the sequence was produced in a random manner.

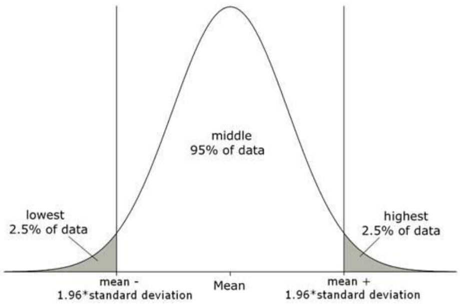
The alternate hypothesis, H_1 states that the sequence was not produced in a random manner.

The null hypothesis is rejected if the number of runs is two few and also if the number of runs is too many. For example, in sequence 1, 0000011111, there are two few runs, and as a result the null hypothesis is rejected. Another example can be found in sequence 2, 1010101010. In this data set, the number of runs is too many and therefore, the null hypothesis is rejected.

The conditions for the rejection of the null hypothesis are:

- Test statistic ≤ lower critical value
- Test statistic ≥ upper critical value

The critical values for large data sets can be found using a normal distribution curve.



[3]

The level of significance, α, is on the y axis. The critical values which correspond to 0.05 are ± 1.96

N_0 is the number of zeroes, N_1 is the number of ones, and N is the total number of digits.

The expected number of runs, [4]$\mathrm{E(R)} = 1 + \frac{2N_0N_1}{N} = 1 + \frac{2N_0N_1}{N_1+N_0}$. This gives a number of runs which is near half the number of digits, while considering the number of ones and zeroes available to form runs.

The variance of the data is a measure of how far each value within a data set is from the mean. It is equal to [5] $\frac{2N_1N_0(2N_1N_0-N)}{N^2(N-1)}$.

The standard deviation is equal to the square root of the variance.

The test statistic, Z can be calculated by comparing the deviation between the number of observed runs and the expected number of runs, with the standard deviation of the series of binary labels. That is, the test statistic,[6] $Z = \frac{R-E(R)}{sR}$.

A sample calculation can be performed with the repeating decimal $0.\overline{12}$ to 1000 decimal places.

The values of N_0, N_1, and R, are substituted into the equations to give te following tabulated results.

---

[3] Taken from http://msor.rsscse.org.uk/leaflets/ssim/SDandCI.php
[4] Taken from http://www.itl.nist.gov/div898/handbook/eda/section3/eda35d.htm
[5] Taken from http://www.itl.nist.gov/div898/handbook/eda/section3/eda35d.htm
[6] Taken fromhttp://www.itl.nist.gov/div898/handbook/eda/section3/eda35d.htm

Sample Calculations for $0.\overline{12}$:

Mean	1.5
R (Number of observed runs)	1000
N_0 (Number of 0s)	500
N_1 (Number of 1s)	500
N (Number of digits)	1000
$E(R)$	501.000
$Var(R)$	249.750
$St.Dev(R)$	15.803
Z	31.575

Since, $Z > 1.96$, the null hypothesis is rejected and the repeating decimal is therefore, non-random.

(See appendix two for calculation tables for other repeating decimals of varying string lengths)

Constructing a Hypothesis for the Runs Test

Table Showing the Results of the Chi-Square Test

Decimal	String Length, n	H_0	Z
$0.\overline{3}$	1	Z undefined	Undefined
$0.\overline{12}$	2	×	31.575
$0.\overline{456}$	3	×	15.727
$0.\overline{6789}$	4	✓	-0.063
$0.\overline{87654}$	5	✓	1.252
$0.\overline{765432}$	6	×	-10.567
$0.\overline{0123456}$	7	×	-13.234
$0.\overline{01234567}$	8	×	-15.883
$0.\overline{012345678}$	9	×	--17.355
$0.\overline{0123456789}$	10	×	-19.046
Control: Π	∞	✓	1.211

Key:

× Means that the null hypothesis was rejected.

✓ Means that the null hypothesis was accepted.

Hypothesis Two:

As seen in the results above, it seems as though a decimal is deemed non-random if the string is too long, as well as if the string is too short. This can be seen specifically where the string length increases from 2 to 3 after which the null hypothesis is accepted. It can also be seen where the string length increases from 4 to 5 and the null hypothesis goes from being accepted, to being rejected. It can therefore be hypothesized that according to the runs test, a number is non-random whenever the string is too short or too long. However, the contrary is true for the irrational number, pi. It seems as though, although its string length is very long, that it is still random. However, this is the only piece of data which is not congruent with the others, so the stated hypothesis will be maintained.

General Solution Two- For the Runs Test Hypothesis

Let a be the number of 1s in the binary representation of the string and let b be the number of 0s in the binary representation of the string.

Let N be the total number of digits in the binary representation.

The total number of 1s in the decimal, $N_1 = \frac{n_1 N}{n}$ ……. Equation1

The total number of 0s in a decimal, $N_0 = \frac{n_0 N}{n}$………..Equation 2

The expected number of runs, $E(R) = \frac{2N_1N_0}{N} + 1$

Substitute equations 1 and 2 in the above equation to get:

$$E(R) = \frac{2n_1n_0N + n^2}{n^2}$$

The variance, $\mathrm{Var} = \frac{2\mathrm{AB}(2\mathrm{AB}-\mathrm{N})}{N^2(N-1)}$. Substitute of equations 1 and 2 in this equation to get:

$$Var = \frac{2n_0n_1N^3(2n_0n_1N - n^2)}{n^4N^2(N-1)}$$

The standard deviation is the square root of the variance $= \sqrt{\frac{2n_0n_1N^3(2n_0n_1N-n^2)}{n^4N^2(N-1)}}$

$$St.Dev. = \frac{1}{n^2N} \cdot \sqrt{\frac{2n_0n_1N^3(2n_0n_1N - n^2)}{(N-1)}}$$

The general solution to the total number of runs in any decimal, can be written for 2 different cases..

<u>In CASE#1: </u>The binary representation of the string ends and starts in the same label-'1' or '0'.

Let the number of runs in a string be r. If the string starts and ends in the same level (0 or 1), then a new string will start where the last one ended, which is at the same label. This means that the run at the end of a string will continue into the beginning of a new string, rather than having a new run at the beginning of the new string. For example,

Let's say that there is the binary string, 01010. This string has 5 runs. When a new string starts the result is: 0101<u>00</u>1010. However, where the two strings meet, the number of runs is the same. This means that for CASE#1, every new string, will mean one less run than expected. Subtract all of these ones to get:

$r \cdot \frac{N}{n} - (\frac{N}{n} - 1)$ total number of runs. When simplified:

$$CASE\#1: Total\ number\ of\ runs, R = \frac{N(r-1)}{n} + 1$$

The test statistic, $Z = \frac{R-E(R)}{St.\ Dev}$

$$CASE\#1: TEST\ STATISTIC, Z = \frac{N^2\sqrt{N-1}\ \cdot [Nn(r-1) - 2n_0n_1]}{\sqrt{2n_0n_1N^3(2n_0n_1N - n^2)}}$$

According to the deduced formula for the test statistic of CASE#1, as n^2 increases, the dominator should decrease and the value of the test statistic should increase. However because $n = a + b$, it must also be considered that as n increases, the value ab also increases.

At the minimum value, $n_1 = 1, n_0 = 1, r = 2$

$$Z = \frac{N^2\sqrt{N-1}\ \cdot [Nn(r-1) - 2n_0n_1]}{\sqrt{2n_0n_1N^3(2n_0n_1N - n^2)}} = \frac{N^2\sqrt{N-1}\ \cdot [2N-2]}{\sqrt{2N^3(2N-4)}} = \frac{N(N-1)^3}{N-2}$$

$N(N-1)^3 > N - 2$.Therefore, at shorter string lengths, the CASE#1 test statistic, is relatively high.

For a longer string length, where $a = 5, b = 5, r = 2,$

$z = \frac{N^2\sqrt{N-1}\cdot[10N-50]}{\sqrt{50N^3(50N-100)}}$. Considering that N is a fairly high value (since tests are being done on repeating decimals), the value of the CASE#1 test statistic is still relatively high, even with a longer string length,

<u>In CASE#2:</u> The first string starts and ends in different binary digits.

In this case, the total number of runs will be the number of times the string is repeated, $\frac{N}{n}$, multiplied by the number of runs per string, *r*.

$$CASE\#2: Total\ number\ of\ runs, R = \frac{rN}{n}$$

$$CASE\#2: TEST\ STATISTIC, Z = \frac{n^2N\sqrt{N-1}(rNn - 2n_0n_1N + n^2)}{\sqrt{2n_0n_1N^3(2n_0n_1N - n^2)}}$$

At the minimum value, $Z = \frac{4N\sqrt{N-1}(2N+4)}{\sqrt{2N^3(2N-4)}}$, the value of the CASE#2 test statistic will also be relatively high.

However, at maximum value tested, where $a = 5$, $b = 5$, $r = 2$,

$Z = \frac{100N\sqrt{N-1}(-30N+100)}{\sqrt{50N^3(50N-100)}}$. Therefore, the value of the CASE#2 test statistic gets very low, even negative, as $n_0 n_1$ and n increase.

Runs Test Conclusion:

Hypothesis two states that according to the runs test, a number is non-random whenever the string is too short or too long. This hypothesis was however, true to different extents, depending on the CASE. The reason the data and the test results pointed to the hypothesis made is because all of the values tested fit the CASE#2 criteria and therefore, had the CASE#2 test statistic. On the other hand, when a decimal fits the CASE#1 criteria, which is that the binary representation starts and ends with the same label-'1' or '0', the test statistic differs greatly from the CASE#2 test statistic. In CASE#2, as the string length increases, the test statistic decreases to point where it is no longer between the upper and lower critical boundaries. However, in CASE#1, the test statistic will take much longer to decrease, or even to become negative. According to the calculations made, the value of the CASE#1(string starts and ends in same label-eg. 1001) is $Z = \frac{N^2\sqrt{N-1}\cdot[Nn(r-1)-2n_0n_1]}{\sqrt{2n_0n_1N^3(2n_0n_1N-n^2)}}$ and the value of the CASE#2 test statistic is,

$$Z = \frac{n^2N\sqrt{N-1}(rNn-2n_0n_1N+n^2)}{\sqrt{2n_0n_1N^3(2n_0n_1N-n^2)}}.$$

Overall, the hypothesis was proven true for both, however, CASE#2's test statistic, decreases faster than CASE#1's.

The issue with having string lengths that are too short can be seen in the test results of the repeating decimal of string length 1. This only allows for one label-'1' or'0' to be used and therefore, most calculations will equate zero, causing the value of the test statistic to be undefined. Also, as seen in the solution, a small value for n, causes the test statistic to be greater and therefore, the shorter the string length, more likely it will be that the test statistic will be greater than or equal to the upper critical value.

Further Investigation:

It would be better to have a more generalized proof of the effects of changing the string length, rather than just showing specific, but relevant examples. Another investigation which would also be interesting is the effect on the increase in string length on the increase or decrease in the number of runs.

Appendix One-Chi-Square Test Calculations

(ii) 0.12

Digits	0	1	2	3	4	5	6	7	8	9
Expected Frequency , E	100	100	100	100	100	100	100	100	100	100
Observed Frequency , O	0	500	500	0	0	0	0	0	0	0
Deviation $(\mathrm{O}-\mathrm{E})$	-100	400	400	-100	-100	-100	-100	-100	-100	-100
Deviation Squared $(\mathrm{O}-\mathrm{E})^2$	10000	160000	160000	10000	10000	10000	10000	10000	10000	10000
Deviation Squared divided by the Expected Frequency $\frac{(\mathrm{O}-\mathrm{E})^2}{\mathrm{E}}$	100	1600	1600	100	100	100	100	100	100	100

α	0.05
DF	9
χ^2	4000
Critical Value	16.9
Reject H_0 Hypothesis	Yes

p	0
Reject H_0 Hypothesis	yes

(iii) 0.456

Digits	0	1	2	3	4	5	6	7	8	9
Expected Frequency , E	100	100	100	100	100	100	100	100	100	100
Observed Frequency , O	0	0	0	0	334	333	333	0	0	0
Deviation $(\mathrm{O}-\mathrm{E})$	-100	-100	-100	-100	234	233	233	-100	-100	-100

Deviation Squared $(O-E)^2$	10000	10000	10000	10000	54756	54289	54289	10000	10000	10000
Deviation Squared divided by the Expected Frequency $\frac{(O-E)^2}{E}$	100	100	100	100	547.56	542.89	542.89	100	100	100

α	0.05
DF	9
χ^2	2333.34
Critical Value	16.9
Reject H_0 Hypothesis	Yes

p	0
Reject H_0 Hypothesis	yes

(iv) 0.6789

Digits	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Expected Frequency, E	100	100	100	100	100	100	100	100	100	100
Observed Frequency, O	0	0	0	0	0	0	250	250	250	250
Deviation $(O-E)$	-100	-100	-100	-100	-100	-100	150	150	150	150
Deviation Squared $(O-E)^2$	10000	10000	10000	10000	10000	10000	22500	22500	22500	22500
Deviation Squared divided by the Expected Frequency $\frac{(O-E)^2}{E}$	100	100	100	100	100	100	225	225	225	225

α	0.05
DF	9

χ^2	1500
Critical Value	16.9
Reject H_0 Hypothesis	yes

p	0
Reject H_0 Hypothesis	yes

(v) 0.87654

Digits	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Expected Frequency, E	100	100	100	100	100	100	100	100	100	100
Observed Frequency, O	0	0	0	0	200	200	200	200	200	0
Deviation $(\mathrm{O}-\mathrm{E})$	-100	-100	-100	-100	100	100	100	100	100	-100
Deviation Squared $(\mathrm{O}-\mathrm{E})^2$	10000	10000	10000	10000	10000	10000	10000	10000	10000	10000
Deviation Squared divided by the Expected Frequency $\frac{(\mathrm{O}-\mathrm{E})^2}{\mathrm{E}}$	100	100	100	100	100	100	100	100	100	100

α	0.05
DF	9
χ^2	1000
Critical Value	16.9
Reject H_0 Hypothesis	yes

p	1.7241E-209
Reject H_0 Hypothesis	yes

(vi) 0.765432

Digits	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**

Expected Frequency, E	100	100	100	100	100	100	100	100	100	100
Observed Frequency, O	0	0	166	166	167	167	167	167	0	0
Deviation $(\mathrm{O}-\mathrm{E})$	-100	-100	66	66	67	67	67	67	-100	-100
Deviation Squared $(\mathrm{O}-\mathrm{E})^2$	10000	10000	4356	4356	4489	4489	4489	4489	10000	10000
Deviation Squared divided by the Expected Frequency $\frac{(\mathrm{O}-\mathrm{E})^2}{\mathrm{E}}$	100	100	43.56	43.56	44.89	44.89	44.89	44.89	100	100

α	0.05
DF	9
χ^2	666.68
Critical Value	16.9
Reject H_0 Hypothesis	yes

p	1.003E-137
Reject H_0 Hypothesis	yes

(vii) 0.0123456

Digits	0	1	2	3	4	5	6	7	8	9
Expected Frequency, E	100	100	100	100	100	100	100	100	100	100
Observed Frequency, O	143	143	143	143	143	143	142	0	0	0
Deviation $(\mathrm{O}-\mathrm{E})$	43	43	43	43	43	43	42	-100	-100	-100
Deviation Squared $(\mathrm{O}-\mathrm{E})^2$	1849	1849	1849	1849	1849	1849	1764	10000	10000	10000
Deviation Squared divided by the Expected Frequency $\frac{(\mathrm{O}-\mathrm{E})^2}{\mathrm{E}}$	18.49	18.49	18.49	18.49	18.49	18.49	17.64	100	100	100

α	0.05

DF	9
χ^2	428.58
Critical Value	16.9
Reject H_0 Hypothesis	yes

p	1.08397E-86
Reject H_0 Hypothesis	yes

(viii) 0.01234567

Digits	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Expected Frequency, E	100	100	100	100	100	100	100	100	100	100
Observed Frequency, O	125	125	125	125	125	125	125	125	0	0
Deviation $(\mathrm{O}-\mathrm{E})$	25	25	25	25	25	25	25	25	-100	-100
Deviation Squared $(\mathrm{O}-\mathrm{E})^2$	625	625	625	625	625	625	625	625	10000	10000
Deviation Squared divided by the Expected Frequency $\frac{(\mathrm{O}-\mathrm{E})^2}{\mathrm{E}}$	6.25	6.25	6.25	6.25	6.25	6.25	6.25	6.25	100	100

α	0.05
DF	9
χ^2	250
Critical Value	16.9
Reject H_0 Hypothesis	yes

p	9.97615E-49
Reject H_0 Hypothesis	Yes

(ix) 0.012345678

Digits	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Expected Frequency, E	100	100	100	100	100	100	100	100	100	100
Observed Frequency, O	112	111	111	111	111	111	111	111	111	0
Deviation $(\mathrm{O}-\mathrm{E})$	12	11	11	11	11	11	11	11	11	-100
Deviation Squared $(\mathrm{O}-\mathrm{E})^2$	144	121	121	121	121	121	121	121	121	10000
Deviation Squared divided by the	1.44	1.21	1.21	1.21	1.21	1.21	1.21	1.21	1.21	100

Expected Frequency $\frac{(O-E)^2}{E}$										

α	0.05
DF	9
χ^2	111.12
Critical Value	16.9
Reject H_0 Hypothesis	yes

p	8.69644E-20
Reject H_0 Hypothesis	yes

(x) 0.0123456789

Digits	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Expected Frequency, E	100	100	100	100	100	100	100	100	100	100
Observed Frequency, O	100	100	100	100	100	100	100	100	100	100
Deviation $(O - E)$	0	0	0	0	0	0	0	0	0	0
Deviation Squared $(O - E)^2$	0	0	0	0	0	0	0	0	0	0
Deviation Squared divided by the Expected Frequency $\frac{(O-E)^2}{E}$	0	0	0	0	0	0	0	0	0	0

α	0.05
DF	9
χ^2	0
Critical Value	16.9
Reject H_0 Hypothesis	No

p	1
Reject H_0 Hypothesis	No

Control: π

Digits	0	1	2	3	4	5	6	7	8	9
Observed Value	93	116	103	103	93	97	94	95	100	106
Expected Value	100	100	100	100	100	100	100	100	100	100
p	0.853049013									
Reject H_0 Hypothesis	No									

Appendix Two-Runs Test Calculations

(i) 0.3

Mean	3
R (Number of observed runs)	1
N_0 (Number of 0s)	1000
N_1 (Number of 1s)	0
N (Number of digits)	1000
$E(R)$	1.000
$Var(R)$	0.000
$St.Dev(R)$	0.000
Z	Undefined

(iii) 0.456

Mean	4.999
R (Number of observed runs)	667
N_0 (Number of 0s)	334
N_1 (Number of 1s)	666
N (Number of digits)	1000
$E(R)$	445.888
$Var(R)$	197.678
$St.Dev(R)$	14.060
Z	15.727

Null hypothesis is rejected.

(iv) 0.6789

Mean	7.5
R (Number of observed runs)	500
N_0 (Number of 0s)	500
N_1 (Number of 1s)	500
N (Number of digits)	1000
$E(R)$	501.000
$Var(R)$	249.750
$St.Dev(R)$	15.803
Z	-0.063

Null hypothesis is accepted.

(v) 0.87654

Mean	6

R (Number of observed runs)	500
N_0 (Number of 0s)	600
N_1 (Number of 1s)	400
N (Number of digits)	1000
$E(R)$	481.000
$Var(R)$	230.150
$St.Dev(R)$	15.171
Z	1.252

Null hypothesis is accepted.

(vi) 0.765432

Mean	4.504
R (Number of observed runs)	334
N_0 (Number of 0s)	499
N_1 (Number of 1s)	501
N (Number of digits)	1000
$E(R)$	500.998
$Var(R)$	249.748
$St.Dev(R)$	15.803
Z	-10.567

Null hypothesis is rejected.

(vii) 0.0123456

Mean	2.997
R (Number of observed runs)	286
N_0 (Number of 0s)	429
N_1 (Number of 1s)	571
N (Number of digits)	1000
$E(R)$	490.918
$Var(R)$	239.769
$St.Dev(R)$	15.484
Z	-13.234

Null hypothesis is rejected.

(viii) 0.01234567

Mean	3.5
R (Number of observed runs)	250
N_0 (Number of 0s)	500
N_1 (Number of 1s)	500
N (Number of digits)	1000
$E(R)$	501.000
$Var(R)$	249.750
$St.Dev(R)$	15.803

Z	-15.883

Null hypothesis is rejected.

(ix) 0.012345678

Mean	3.108
R (Number of observed runs)	224
N_0 (Number of 0s)	445
N_1 (Number of 1s)	555
N (Number of digits)	1000
$E(R)$	494.950
$Var(R)$	243.736
$St.Dev(R)$	15.612
Z	-17.355

Null hypothesis is rejected.

(x) 0.0123456789

Mean	4.5
R (Number of observed runs)	200
N_0 (Number of 0s)	500
N_1 (Number of 1s)	500
N (Number of digits)	1000
$E(R)$	501.000
$Var(R)$	249.750
$St.Dev(R)$	15.803
Z	-19.046

Null hypothesis is rejected.

Control: π

Mean	4.471
R (Number of observed runs)	520
N_0 (Number of 0s)	508
N_1 (Number of 1s)	492
N (Number of digits)	1000
$E(R)$	500.872
$Var(R)$	249.622
$St.Dev(R)$	15.799
Z	1.211
	0.887
	1.96

Null hypothesis is accepted.

References

1. http://www.random.org/analysis/Analysis2005.pdf, research paper by Charmaine Kenny, April 2005
2. http://www.itl.nist.gov/div898/handbook/, NIST Statistical Engineering e-Handbook, April 2012
3. http://csrc.nist.gov/publications/nistpubs/800-22-rev1a/SP800-22rev1a.pdf , NIST research paper, AndrewRukhin JuanSoto JamesNechvatal MilesSmid ElaineBarker Stefan Leigh MarkLevenson MarkVangel DavidBanks AlanHeckert JamesDray SanVo, April 2010
4. https://www2.palomar.edu/users/rmorrissette/Lectures/Stats/ChiSquare/TableX.jpg, Chi-Square Test Refernce Table
5. http://webspace.ship.edu/pgmarr/Geo441/Tables/Runs%20Table.pdf, http://faculty.tarleton.edu/crawford/documents/NormalTable.png, Critical Values for Runs Test

Quotations from:

6. Paul Auster, Author and director
7. Sidney Poitier, Filmwriter