4.3 Statistical Power

Statistical power is a standard criterion by which statisticians evaluate and compare the effectiveness of test statistics at solving hypothesis-testing problems. We compare our test statistics on the basis of their observed statistical power at distinguishing various pairs of languages. In these comparisons we use the four critical levels 0.1, 0.01, 0.001, and 0.0001, which bound the type I error rate. Throughout we approximate observed power as explained in Section 3.5.

Figures 14–17 highlight our comparison for critical level 0.01 in graphs of power versus string length for distinguishing BCa English versus uniform noise, uniform noise versus BCa English, 0th-order BCa versus 1st-order BCa, and 1st-order BCa versus BCa English, respectively. Corresponding Tables 2–5 list power values at all four critical levels for three fixed string lengths. Figures 18–19 show how added noise affects statistical power when distinguishing BCa English from uniform noise, and uniform noise from BCa English, at critical level 0.01 for strings of length $\lg n = 7$. Corresponding Tables 6–7 list power values for three noise levels and all four critical levels. For additional similar power calculations, see supplemental Appendix F [17].

To interpret the power of a statistic at recognizing a language, keep in mind that 1 is perfect power (invalid strings are always rejected), and 0 is the worst possible power (invalid strings are never rejected). Although the practical implications of power depend crucially on the application, power less than 0.5 would typically be very poor, since few engineering applications could use any statistic that accepts so many invalid strings. An application might require power to be at least 0.7, and preferably greater than 0.9. Though minor differences in power (say for our data, less than 0.01) might not affect some engineering decisions, in typical cryptanalytic applications, even an apparently small increase in power from 0.90 to 0.91 would reduce by 10% the expected time spent processing spurious keys.

Throughout it is important to pay attention to the critical level, since for some tasks the relative power of the statistics changes with critical level. Cryptanalysts typically desire small critical levels (say, at most 0.01) since overlooking valid plaintext would be a serious error for many applications.

Although each test statistic is designed for a specific recognition task, we make a broad comparison by computing power for each test statistic at distinguishing each pair of languages considered. Results show there is a complicated interaction between statistical power, string length, noise level, and critical level: each graph and table has a leading statistic or statistics, but typically no statistic dominated all others at all string lengths and critical levels. Nevertheless, each statistic performed reasonably well at its designated task.

Power Analysis of Experiment 1

To understand the power graphs and tables intuitively, it is helpful to examine the histograms in Section 4.1. Consider Figure 14 and Table 2, for example, and examine the related histograms of the test statistics on BCa English and on uniform noise in Figures 1–6. For each critical level and for all sufficiently long strings, each test statistic attained a perfect power of 1, which reflects the fact that, for each statistic at such string lengths, the histogram for BCa English was perfectedly separated from the the histogram for uniform noise. The most interesting part of any power comparison corresponds to overlapping histograms.

In Figure 14 and Table 2, the statistic $\diamond X^2$ performed very well at distinguishing BCa English from uniform noise, dominating all other statistics at most string lengths. As expected, the statistics S and ML also performed well at this task, as they were designed to do. We were surprised, however,

that $\diamond X^2$ outperformed S given that S is an asymptotically most powerful test for distinguishing BCa English from uniform noise. Perhaps the relatively better performance of $\diamond X^2$ over S at $\lg n < 7$ is explained by the small string length. Even for the short string length $\lg n = 3$ at critical level 0.001, $\diamond X^2$ achieved a reasonably high power of over 0.79. Here and throughout, Anderson's statistic performed unimpressively—usually worse and never better than S (except for $\lg n = 1$). But for each critical level and for all sufficiently long strings ($\lg n \geq 10$), all statistics performed indistinguishably with perfect power.

In Figure 15 and Table 3, $\diamond ML$ dominated the other statistics at distinguishing uniform noise from BCa English, for string lengths $\lg n > 3$. The statistics IC, and $\diamond IND$, $\diamond X^2$, and S also performed relatively well. The strong performance of IC and $\diamond IND$ was expected: IC is designed to recognize uniform noise, and IND is designed to distinguish 0th-order language (which includes uniform noise) from 1st-order language (which we use to model BCa). We were surprised, however, that $\diamond ML$ outperformed IC since ML is designed to recognize BCa. We attribute this relatively better performance of $\diamond ML$ over IC to the fact that $\diamond ML$ depends on BCa transition probabilities whereas IC does not. As happened in Figure 14, for all critical levels and for all sufficiently long strings ($\lg n \geq 10$), all statistics performed indistinguishably with perfect power.

When distinguishing uniform noise from BCa English, the relative performance of the statistics varied depending on the critical level. For example, $\diamond X^2$ and S outperformed IC at critical level 0.1 but not at level 0.001, with the greatest difference in power between IC and S occurring at short string lengths and low critical levels. By contrast, when distinguishing BCa English from uniform noise, raising the critical level increased the power of all statistics but had little effect on their relative performance.

For test statistics $\diamond X^2$, $\diamond ML$, S, and $\ln A$, for the same critical level, as high or higher power was attained by distinguishing BCa English from uniform noise than by distinguishing uniform noise from BCa English. In particular, for short strings ($\lg n < 4$) and all critical levels, significantly higher power was so attained using the best statistic for each problem ($\diamond X^2$ and $\diamond ML$, respectively). The implication to the practitioner, however, depends on the application. For example, when recognizing valid plaintext in cryptanalysis, typically the cryptanalyst will set the threshold on the basis of minimizing the chance of overlooking valid plaintext. Thus, when distinguishing BCa English from uniform noise, the cryptanalyst will start with a desired critical level. But when distinguishing uniform noise from BCa English, the cryptanalyst will pick a critical level that achieves a desired power. Therefore, for this application, the choice of which of these two hypothesis testing problems to use cannot be decided from Figures 14 and 15 alone. In addition, this choice will also depend on other factors, such as the cryptanalyst's degree of belief in what plaintext language was used.

For test statistics IC and $\diamond IND$, however, higher power was attained when distinguishing uniform noise from BCa English. This behavior results from the fact that each of IC and $\diamond IND$ has a much smaller variance on uniform noise than on BCa English, as observed in the histograms.

In all of our power calculations, as expected, the equivalent statistics S, S/N, and $\hat{S}$ achieved nearly identical powers (typically through at least three decimal places). For this reason, Tables 2–7 do not list powers for S/N or $\hat{S}$. We attribute the minor differences in their powers to experimental approximations, possibly due to approximations in computing tail areas of the normal distribution and our method for estimating observed power. Similarly, the equivalent statistics IC and $\hat{IC}_\star$ also achieved very close statistical powers.

Figure 16 and Table 4 present observed powers for distinguishing 0th-order BCa from 1st-order

BCa. At all critical levels, the statistics S and $\diamond ML$ attained the highest power. Given that $\diamond IND$ is designed to distinguish unknown 0th-order noise from unknown 1st-order language, we were initially surprised that it did not perform better. We speculate that $\diamond X^2$, $\diamond ML$, S, and $\ln A$ outperformed $\diamond IND$ by exploiting knowledge of the BCa transition probabilities. Therefore, it would be interesting to compare S and $\diamond IND$ at distinguishing 0th-order and 1st-order languages unrelated to BCa. We did compute powers for distinguishing 0th-order BCa from WSJ1 (see supplemental Appendix F), and $\diamond IND$ achieved slightly higher power than it did in Table 4. But the overall results were similar due to the fact that BCa and WSJ1 are both types of English and have similar transition probabilities.

The purpose of Figure 17 and Table 5 is to assess and to compare the robustness of the test statistics with respect to our 1st-order model of BCa English. Whereas Figure 14 shows powers for real BCa English versus uniform noise, Figure 17 shows powers for simulated 1st-order BCa versus uniform noise. In both of these tables, $\diamond X^2$ attained the highest power in most situations, with $\diamond ML$ and S also performing well. Comparing the powers of all six statistics in Tables 5 and 2, we observe that the powers are almost identical, with S and $\diamond ML$ achieving slightly higher power in Table 5 than in Table 2. Thus, all six statistics are robust with respect to our 1st-order model of BCa.

Power Analysis of Experiment 2

Figure 18 shows the observed power of the test statistics at distinguishing BCa English from uniform noise, when various amounts of uniform noise are added to the candidate string. Figure 19 shows an analogous graph for distinguishing noise from BCa English. Both graphs are for string length $\lg n = 7$ and critical level 0.01. In Figures 18 and 19, $\diamond ML$ outperformed the other test statistics at most noise levels. It is striking that $\diamond ML$ performed so well at relatively high noise levels. For example, when distinguishing BCa English from uniform noise with 40% added noise, $\diamond ML$ achieved a power of over 0.8.

The statistics S, IC, and $\diamond IND$ also performed well under noisy conditions. As expected, however, IC and $\diamond IND$ attained relatively higher power at distinguishing uniform noise from BCa English than at distinguishing BCa English from uniform noise. By contrast, S attained relatively higher power at distinguishing from BCa English from uniform noise.

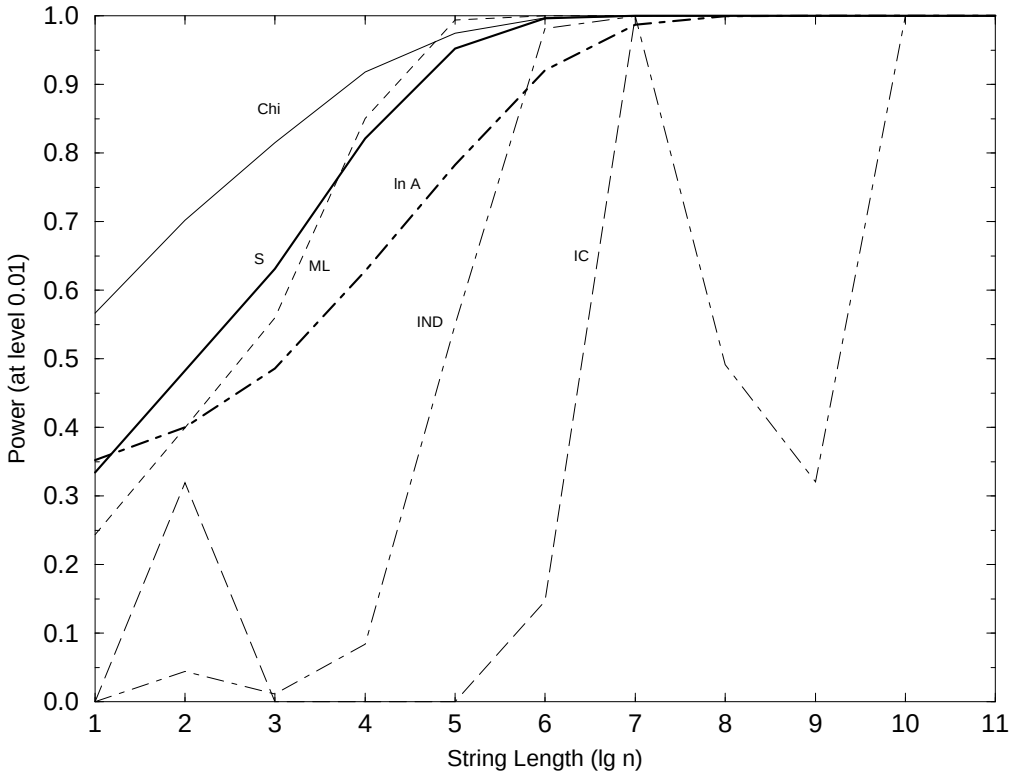

Figure 14: Power of the test statistics $\diamond X^2$, $\diamond ML$, $\diamond IND$, S, IC, and $\ln A$ at distinguishing BCa English from uniform noise at critical level 0.01 for strings of various lengths. Power was approximated from 10,000 randomly chosen strings at each length. Base language is BCa.

Table 2: Power of six test statistics at distinguishing BCa English from uniform noise at critical levels 0.1, 0.01, 0.001, 0.0001 for string lengths $\lg n = 4, 5, 6$. Base language is BCa.

$\lg n$	Critical Level	$\diamond X^2$	$\diamond ML$	$\diamond IND$	S	IC	$\ln A$
4	0.1000	0.9350	0.9462	0.3834	0.9009	0.0477	0.7452
4	0.0100	0.9183	0.8509	0.0839	0.8214	0.0000	0.6281
4	0.0010	0.9042	0.7338	0.0150	0.7433	0.0000	0.5333
4	0.0001	0.8914	0.6110	0.0024	0.6677	0.0000	0.4536
5	0.1000	0.9823	0.9991	0.8894	0.9786	0.3168	0.8667
5	0.0100	0.9747	0.9940	0.5501	0.9523	0.0002	0.7831
5	0.0010	0.9676	0.9804	0.2494	0.9200	0.0000	0.7062
5	0.0001	0.9607	0.9549	0.0906	0.8828	0.0000	0.6349
6	0.1000	0.9981	1.0000	0.9996	0.9989	0.9988	0.9588
6	0.0100	0.9969	1.0000	0.9815	0.9966	0.1466	0.9205
6	0.0010	0.9956	1.0000	0.8723	0.9927	0.0000	0.8787
6	0.0001	0.9942	1.0000	0.6390	0.9869	0.0000	0.8341

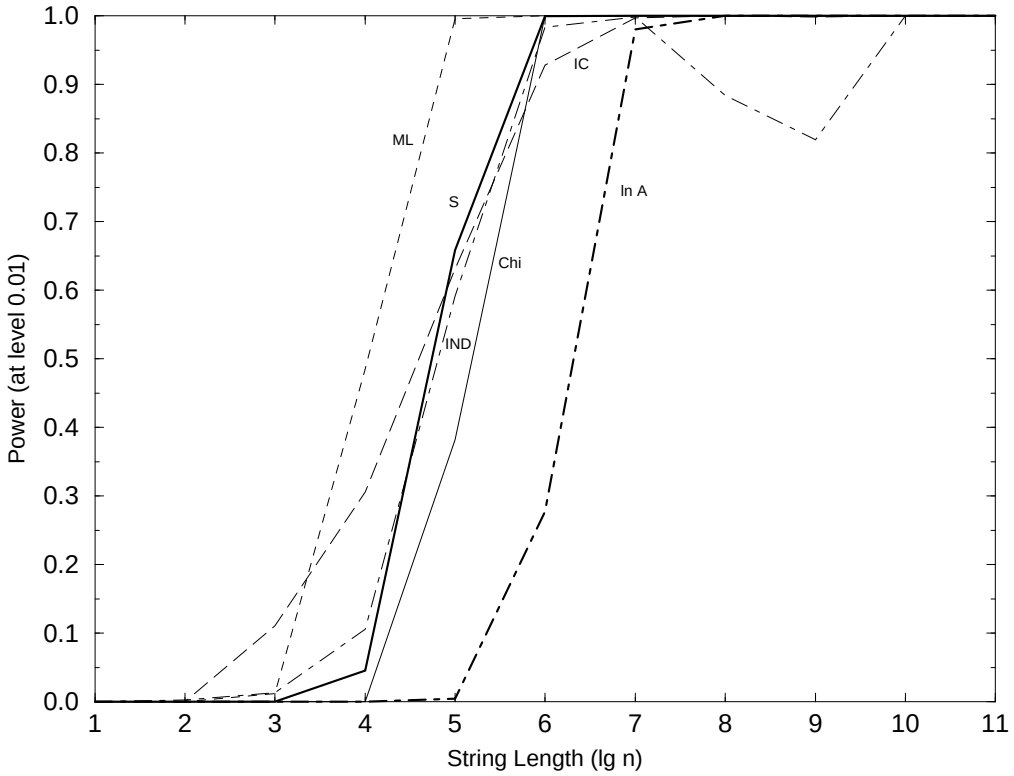

Figure 15: Power of the test statistics $\diamond X^2$, $\diamond ML$, $\diamond IND$, S, IC, and $\ln A$ at distinguishing uniform noise from BCa English at critical level 0.01 for strings of various lengths. Power was approximated from 10,000 randomly chosen strings at each length. Base language is BCa.

Table 3: Power of six test statistics at distinguishing uniform noise from BCa English at critical levels 0.1, 0.01, 0.001, 0.0001 for string lengths $\lg n = 4, 5, 6$. Base language is BCa.

$\lg n$	Critical Level	$\diamond X^2$	$\diamond ML$	$\diamond IND$	S	IC	$\ln A$
4	0.1000	0.9995	0.9702	0.4047	0.9024	0.4843	0.2506
4	0.0100	0.0000	0.4857	0.1057	0.0455	0.3060	0.0000
4	0.0010	0.0000	0.0751	0.0235	0.0001	0.1978	0.0000
4	0.0001	0.0000	0.0047	0.0048	0.0000	0.1290	0.0000
5	0.1000	1.0000	1.0000	0.8900	0.9997	0.7544	0.7701
5	0.0100	0.3820	0.9959	0.5916	0.6588	0.6314	0.0049
5	0.0010	0.0000	0.9091	0.3101	0.0348	0.5311	0.0000
5	0.0001	0.0000	0.6030	0.1370	0.0001	0.4466	0.0000
6	0.1000	1.0000	1.0000	0.9985	1.0000	0.9582	0.9968
6	0.0100	1.0000	1.0000	0.9836	0.9997	0.9282	0.2774
6	0.0010	0.4662	1.0000	0.9356	0.8913	0.8974	0.0013
6	0.0001	0.0000	1.0000	0.8445	0.2839	0.8657	0.0000

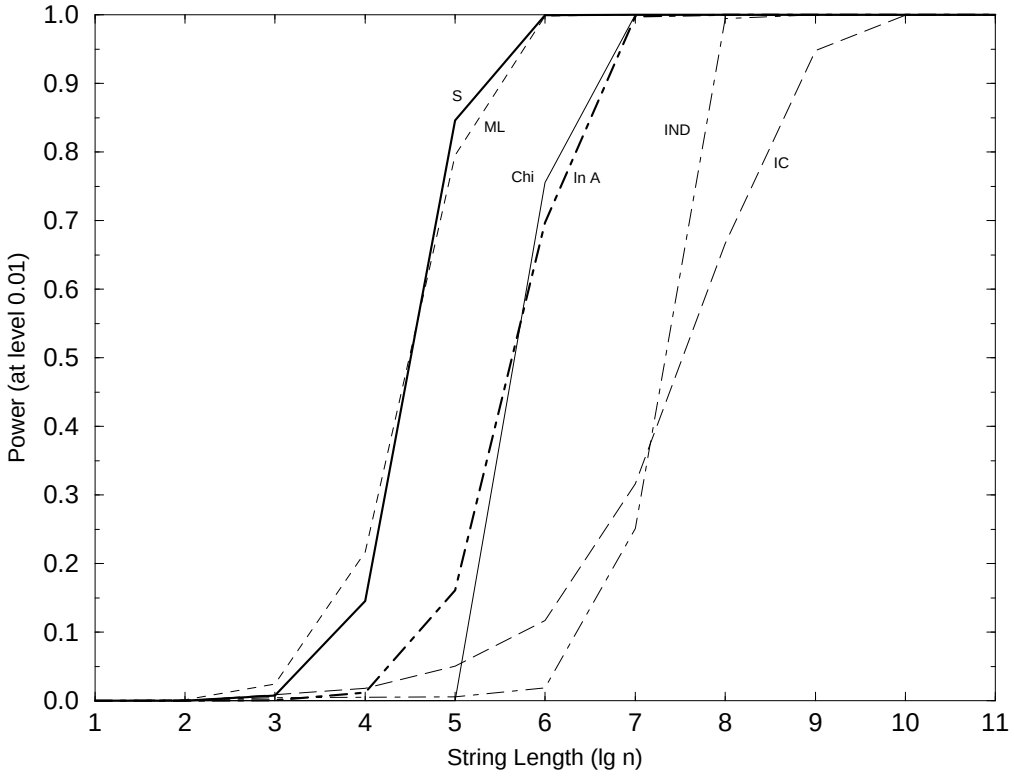

Figure 16: Power of the test statistics $\diamond X^2$, $\diamond ML$, $\diamond IND$, S, IC, and $\ln A$ at distinguishing 0th-order BCa from 1st-order BCa at critical level 0.01 for strings of various lengths. Power was approximated from 10,000 randomly chosen strings at each length. Base language is BCa.

Table 4: Power of six test statistics at distinguishing 0th-order BCa English from 1st-order BCa English at critical levels 0.1, 0.01, 0.001, 0.0001 for string lengths $\lg n = 4, 5, 6$. Base language is BCa.

$\lg n$	Critical Level	$\diamond X^2$	$\diamond ML$	$\diamond IND$	S	IC	$\ln A$
4	0.1000	0.2699	0.8481	0.1030	0.9230	0.1428	0.4691
4	0.0100	0.0000	0.2169	0.0055	0.1458	0.0186	0.0120
4	0.0010	0.0000	0.0176	0.0003	0.0021	0.0023	0.0001
4	0.0001	0.0000	0.0007	0.0000	0.0000	0.0003	0.0000
5	0.1000	0.9963	0.9951	0.1133	0.9998	0.2449	0.8769
5	0.0100	0.0002	0.7954	0.0056	0.8462	0.0507	0.1612
5	0.0010	0.0000	0.3223	0.0002	0.2170	0.0099	0.0052
5	0.0001	0.0000	0.0642	0.0000	0.0117	0.0019	0.0001
6	0.1000	1.0000	1.0000	0.2283	1.0000	0.4028	0.9963
6	0.0100	0.7554	0.9984	0.0188	0.9999	0.1172	0.6970
6	0.0010	0.0004	0.9491	0.0011	0.9651	0.0302	0.1430
6	0.0001	0.0000	0.7096	0.0001	0.6239	0.0072	0.0089

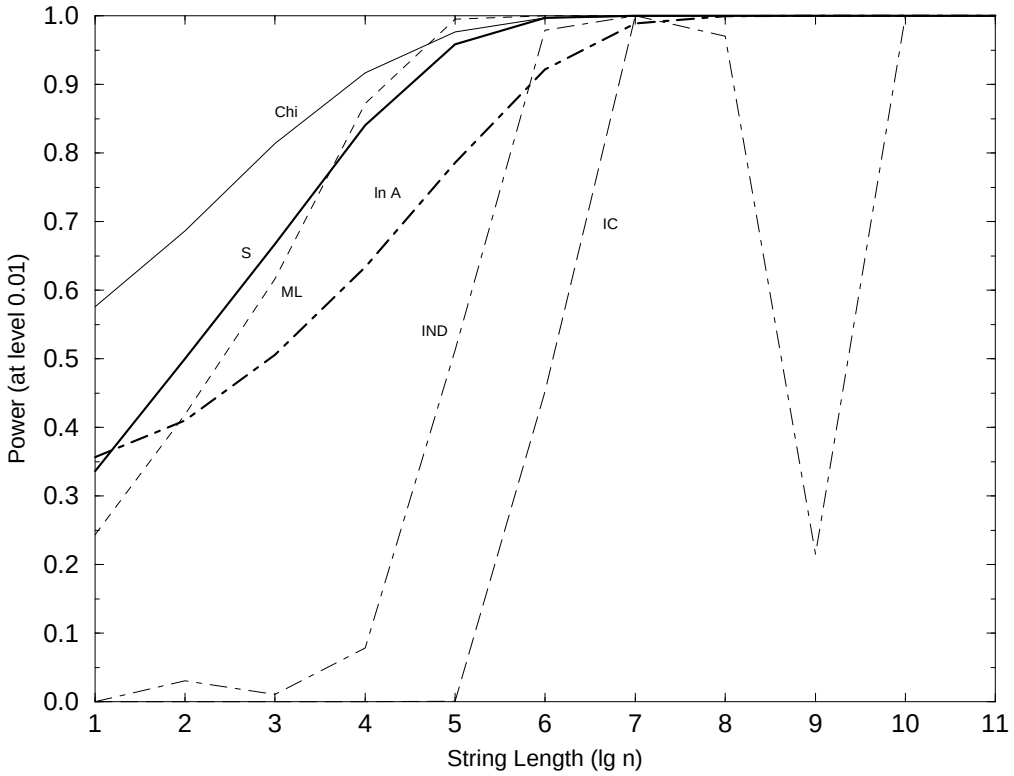

Figure 17: Power of the test statistics $\diamond X^2$, $\diamond ML$, $\diamond IND$, S, IC, and $\ln A$ at distinguishing 1st-order BCa English form uniform noise at critical level 0.01 for strings of various lengths. Power was approximated from 10,000 randomly chosen strings at each length. Base language is BCa.

Table 5: Power of six test statistics at distinguishing 1st-order BCa English from uniform noise at critical levels 0.1, 0.01, 0.001, 0.0001 for string lengths $\lg n = 4, 5, 6$. Base language is BCa.

$\lg n$	Critical Level	$\diamond X^2$	$\diamond ML$	$\diamond IND$	S	IC	$\ln A$
4	0.1000	0.9343	0.9533	0.4115	0.9082	0.0451	0.7483
4	0.0100	0.9168	0.8726	0.0787	0.8406	0.0000	0.6336
4	0.0010	0.9019	0.7717	0.0112	0.7745	0.0000	0.5405
4	0.0001	0.8883	0.6628	0.0013	0.7101	0.0000	0.4619
5	0.1000	0.9831	0.9992	0.8914	0.9807	0.5102	0.8684
5	0.0100	0.9766	0.9951	0.5126	0.9589	0.0010	0.7860
5	0.0010	0.9707	0.9845	0.1983	0.9328	0.0000	0.7103
5	0.0001	0.9648	0.9645	0.0581	0.9030	0.0000	0.6398
6	0.1000	0.9982	1.0000	0.9997	0.9991	0.9999	0.9598
6	0.0100	0.9971	1.0000	0.9791	0.9974	0.4523	0.9223
6	0.0010	0.9960	1.0000	0.8467	0.9945	0.0018	0.8811
6	0.0001	0.9948	1.0000	0.5748	0.9905	0.0000	0.8372

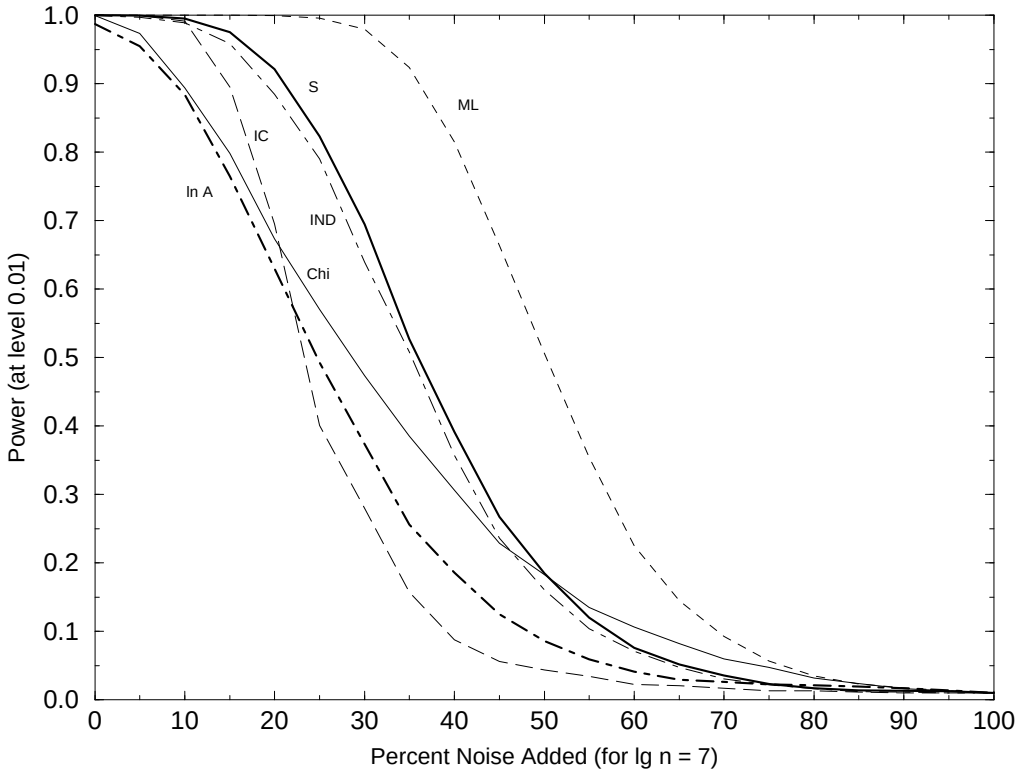

Figure 18: Power of the test statistics $\diamond X^2$, $\diamond ML$, $\diamond IND$, S, IC, and $\ln A$ at distinguishing BCa English from uniform noise at critical level 0.01 with various amounts of added uniform noise for strings of length $\lg n = 7$. Power was approximated from 10,000 randomly chosen strings at each noise level. Base language is BCa.

Table 6: Power of six test statistics at distinguishing BCa English from uniform noise at critical levels 0.1, 0.01, 0.001, 0.0001 for strings of length $\lg n = 7$ at noise levels 10%, 20%, and 30%. Base language is BCa.

% Added Noise	Critical Level	$\diamond X^2$	$\diamond ML$	$\diamond IND$	S	IC	$\ln A$
10	0.1000	0.9754	1.0000	1.0000	0.9993	1.0000	0.9557
10	0.0100	0.8938	1.0000	0.9893	0.9956	0.9919	0.8840
10	0.0010	0.7645	1.0000	0.8293	0.9859	0.1807	0.7951
10	0.0001	0.6132	1.0000	0.4372	0.9672	0.0001	0.6981
20	0.1000	0.8907	1.0000	0.9988	0.9836	1.0000	0.8276
20	0.0100	0.6735	0.9997	0.8845	0.9212	0.6952	0.6301
20	0.0010	0.4519	0.9950	0.4412	0.8122	0.0157	0.4540
20	0.0001	0.2774	0.9709	0.1047	0.6745	0.0000	0.3141
30	0.1000	0.7687	0.9993	0.9853	0.9005	0.9897	0.6367
30	0.0100	0.4733	0.9797	0.6396	0.6948	0.2796	0.3735
30	0.0010	0.2569	0.8872	0.1653	0.4772	0.0034	0.2078
30	0.0001	0.1281	0.6999	0.0193	0.3002	0.0000	0.1115

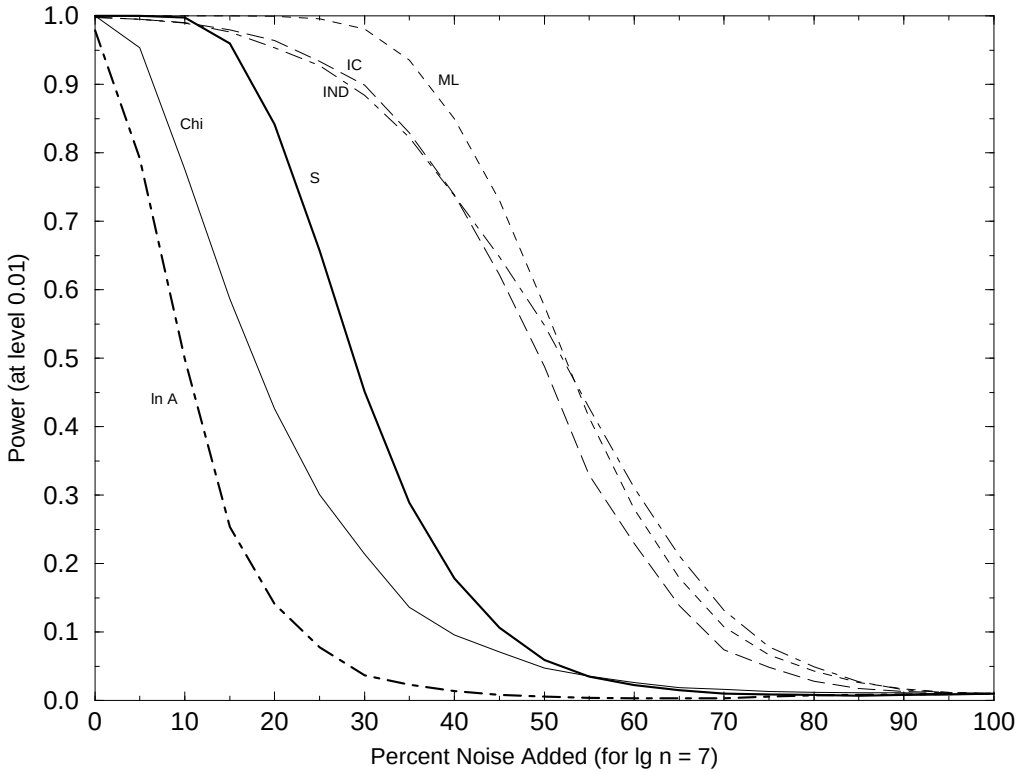

Figure 19: Power of the test statistics $\diamond X^2$, $\diamond ML$, $\diamond IND$, S, IC, and $\ln A$ at distinguishing uniform noise from BCa English noise at critical level 0.01 with various amounts of added uniform noise for strings of length $\lg n = 7$. Power was approximated from 10,000 randomly chosen strings at each noise level. Base language is BCa.

Table 7: Power of six test statistics at distinguishing uniform noise from BCa English at critical levels 0.1, 0.01, 0.001, 0.0001 for strings of length $\lg n = 7$ at noise levels 10%, 20%, and 30%. Base language is BCa.

% Added Noise	Critical Level	$\diamond X^2$	$\diamond ML$	$\diamond IND$	S	IC	$\ln A$
10	0.1000	0.9886	1.0000	0.9982	1.0000	0.9951	0.9842
10	0.0100	0.7762	1.0000	0.9896	0.9979	0.9905	0.4988
10	0.0010	0.3634	1.0000	0.9698	0.9326	0.9849	0.0575
10	0.0001	0.1034	1.0000	0.9360	0.6462	0.9785	0.0020
20	0.1000	0.8874	1.0000	0.9887	0.9941	0.9824	0.7601
20	0.0100	0.4262	0.9996	0.9540	0.8420	0.9645	0.1411
20	0.0010	0.1134	0.9956	0.8947	0.4584	0.9436	0.0087
20	0.0001	0.0202	0.9792	0.8145	0.1549	0.9201	0.0003
30	0.1000	0.7151	0.9988	0.9637	0.9007	0.9509	0.4337
30	0.0100	0.2137	0.9808	0.8841	0.4512	0.8993	0.0367
30	0.0010	0.0368	0.9150	0.7755	0.1246	0.8419	0.0015
30	0.0001	0.0045	0.7874	0.6540	0.0228	0.7810	0.0000

4.4 Discussion

We now discuss several issues raised by our power calculations.

• First, in each of Figures 14, 15, and 17, but not in Figure 16, the power curve for $\diamond IND$ has an unusual downward spike at string lengths $\lg n = 8, 9$. In addition, the IC power curve in Figure 14 has an upward spike at $\lg n = 2$. The $\diamond IND$ spikes are most prominent when distinguishing BCa from uniform noise (a task for which IND was not intended), and absent when distinguishing 0th-order BCa from 1st-order BCa (a task for which IND was explicitedly designed). When distinguishing uniform noise from BCa, the $\diamond IND$ spikes are present but less pronounced.

When we had first observed the $\diamond IND$ spikes from our initial run of Experiment 1, we had suspected that they might be statistical anomalies due to our small sampling size of $h = 100$. But the spikes reappeared when we repeated Experiment 1 with $h = 10,000$. Moreover, from the repeatability of the phenomenon, we cannot simply attribute the spikes to statistical anomalies or to weak pseudorandom number generators. Similarly, given the short and moderate lengths of the affected strings, we cannot attribute the spikes to overlapping strings. From the means of $\diamond IND$ in Table 10 of Appendix B, however, we find a partial explanation: For $3 < \lg n < 9$, the observed means for $\diamond IND$ are less on BCa than on uniform noise, but for $\lg n \geq 9$, the means are greater on BCa than on uniform noise. The $\diamond IND$ spikes correspond to this "crossing of the means" at $\lg n = 9$. We find this phenomenon puzzling, and we do not have a good explanation for the isolated IC spike.

• Second, it would be interesting to compare our power results with previous related power calculations from other researchers. Unfortunately, we are not aware of any such prior work that would permit a direct comparison. We can, however, make some informal comparisons with the work of Baldwin and Sherman [3], and with that of Davies and Ganesan [12].

In their solution of the Decipher Puzzle, Baldwin and Sherman recognized standard English versus 0th-order English with the $\hat{S}$ statistic using a 26-state 1st-order model of English with transition probabilities published by Beker and Piper [4]. Each of their input strings consisted of a sequence of approximately ten independent bigrams, and they rejected strings for which $|\hat{S}| > 4$. Thus, they worked with at a critical level less than 0.0001. Although Baldwin and Sherman did not compute power, they found that their test worked "very well" in practice for their application. By contrast, when distinguishing BCa from uniform noise at level 0.001 for strings of dependent bigrams, we observed the power of S to be 0.5256 for $\lg n = 3$ and 0.7433 at $\lg n = 4$. Thus, our power calculations seem somewhat pessimistic in comparison to the experience of Baldwin and Sherman.

In their BApasswd checker, Davies and Ganesan used an equivalent variation of the S statistic in a 27-state 2nd-order model, estimating their own transition probabilities with the Good-Turing method [19]. Davies and Ganesan rejected bad eight-letter passwords using experimentally-determined thresholds. Although they did not compute power, it is possible to estimate power from their reported data for distinguishing uniform noise from valid English. For example, consider their dictionary file BP6 (bad passwords) as valid English, and consider their file GP1 (good passwords) of randomly-generated passwords with the letters 'A'–'Z' as uniform noise. According to their Figure 7 [12], using their threshold, their test accepted 96.26% of the random GP1 passwords as noise (corresponding to a critical level of 0.0374), while accepting only 12.29% of the English words in BP6 as noise (corresponding to a power of 0.8781). By contrast, when distinguishing uniform noise from BCa at critical level 0.1, we observed the power of S to be 0.3676 at $\lg n = 3$ and 0.9997 at

$\lg n = 4$. Thus, our power calculations also seem pessimistic in comparison to the experimence of Davies and Ganesan, perhaps reflecting a strong advantage of using a trigram model and of using the Good-Turing parameter-estimation method.

• Third, when recognizing BCa English, there is some bias resulting from the fact that we draw our candidate strings from exactly the same base sample file from which we computed the BCa transition probabilities. For example, one consequence of doing so is that unadulterated candidate strings from BCa will never contain impossible bigrams that do not appear in the base sample. For the strings lengths for which we computed power, we do not expect this bias to make much difference; however, for extremely long strings, this bias may have a significant effect.

• Fourth, our method for approximating statistical introduces some error. For example, skew in any histogram would diminish the accuracy of our power calculations involving that histogram. Although there is some skew in our histograms (*e.g.* S/N on BCa), most of the histograms from which power calculations were made appear reasonably symmetrical.

• Finally, we would like to identify additional possible sources of errors that may have affected our results. These possible sources include the sampling sizes of 100 and 10,000 strings, overlapping candidate strings, roundoff error in the computer arithmetic, approximation errors in computing tail areas of the Gaussian distribution, and weaknesses of the pseudorandom number generator. To check the accuracy of our tail area calculations, we recomputed many power calculations using the SAS statistical package [34]. Both calculations agreed through at least four decimal places.

5 Conclusion

In this experimental study we have empirically characterized the distributions of the test statistics X^2, ML, IND, S, and IC when applied to nine types of real and simulated language, including four types of everyday American English. Moreover, to compare the effectiveness of these statistics, we approximated their powers at distinguishing various pairs of the test languages at several critical levels. To the best of our knowledge, this study is the first of its kind to examine the actual behavior of these important test statistics on human languages.

Our results are descriptions of the distributions of the test statistics (given in histograms, graphs, and tables) and comparisons of the effectiveness of these test statistics at distinguishing language pairs (given in graphs and tables of statistical power). As presented and discussed in Section 4, these results show a complex interaction among observed power, string length, and critical level; therefore, practitioners should carefully choose the test statistics that best match the demands of their applications. The following nine observations highlight our findings.

1. Each test statistic performed well at its designated task. Thus, X^2, ML, and S performed well at recognizing English; IC performed well at detecting nonuniform language; and for the statistics that do not depend on BCa parameters, IND performed reasonably well at distinguishing 0th-order BCa from 1st-order BCa. The statistic ML, however, outperformed the theoretically optimal S statistic at distinguishing BCa English from uniform noise for string lengths $\lg n < 7$.

2. For distinguishing BCa English from uniform noise, X^2 had the best overall performance, with ML and S also performing well. For string lengths $\lg n < 7$, we recommend using X^2; for

longer strings, all three of these statistics work perfectly at critical levels 0.1 through 0.0001. Even at critical level 0.0001 and short string length $\lg n = 3$, X^2 attained a reasonable power of approximately 0.77.

3. For distinguishing uniform noise from BCa English, *ML* had the overall best performance, with *IC*, X^2, *S*, and *IND* also performing well. At critical levels 0.01 through 0.0001 *IC* attained higher power than did X^2 (and than did *ML* for $\lg n \leq 3$), but at critical level 0.1 X^2 performed better than did *IC*. For string lengths $\lg n \leq 7$, we recommend using *ML* for critical levels 0.1 through 0.0001. For $\lg n > 7$, each of *ML*, *IC*, X^2, *S* worked perfectly at these critical levels. At $\lg n = 4$ and critical level 0.1, *ML* achieved a power of over 0.99.

4. For distinguishing BCa English from uniform noise using strings of length $\lg n = 7$ corrupted with uniform noise, *ML* outperformed the other statistics at all critical levels. At critical levels 0.01 through 0.0001, *S* had the overall second-best performance. We recommend using *ML* for this problem. At $\lg n = 7$ and critical level 0.01, the power of *ML* remained above 0.8 through noise level 40%. For noise levels 0%–15%, *ML* had attained power greater than 0.99 at critical levels 0.1 through 0.0001; for noise levels 70%–100%, all statistics had power less than 0.5 at these critical levels.

5. For distinguishing uniform noise from BCa English using strings of length $\lg n = 7$ corrupted with uniform noise, *ML* had the overall best performance, with *IC* and *IND* also performing well. We recommend using *ML*. The performance of *ML* on this problem was very similar to, and slightly better than, its performance at distinguishing uniform noise from BCa English under noisy conditions.

6. The *S* statistic outperformed Anderson's variation of it, except when distinguishing BCa English from uniform noise at string length $\lg n = 1$. Therefore, we do not recommend Anderson's variation.

7. For $\lg n < 7$, our four types of real English (BCa, BCf, BCg, and WSJ1) had similar means for all statistics. For longer strings, the statistics could distinguish BCa, BCf, and WSJ1 on the basis of their means.

8. Strict standard normal interpretations of the normalized statistics $\diamond X^2$, $\diamond ML$, and $\diamond IND$ do not apply, except when recognizing long strings of a known simulated 1st-order language. For best results, use experimentally-determined thresholds for all statistics, including $\hat{S}$ and $\hat{IC}_\star$. These thresholds can be computed as explained in Section 3.5.

9. As expected, the performance of all statistics on BCa English closely matched their performance on simulated 1st-order BCa English. In this sense, our 1st-order statistics are robust with respect to our 1st-order model. Nevertheless, minor differences can be seen in their histograms (*e.g.* 1st-order language produced more symmetrical and bell-shaped distributions). In addition, to human observers, our 1st-order BCa strings do not closely resemble real English.

Complete experimental data, including additional descriptive graphs and power calculations, are given in the supplemental appendices of our technical report [17].

Using a broad, exploratory, descriptive approach, we have exposed general trends and uncovered interesting phenomena. We consider these trends and phenomena more important than our particular numerical results, since models and languages vary with the application, and since our power calculations are approximate. Also, we believe that more powerful results can be achieved with higher-order models and with fancier methods for estimating the transition probabilities.

Our study is an initial investigation of several standard statistical techniques. We tried several basic approaches; additional careful experimentation is needed to assess the effectiveness of variations of these approaches. Such variations include more sophisticated ways to estimate the transition probabilities (*e.g.* flattening techniques and the Good-Turing method [19, 9]), alternate test statistics (*e.g.* Good's G statistic [20, 10, 11] and Knuth's spectral test [25]), and more elaborate models (*e.g.* higher-order models and hidden Markov models [32]). For example, it would be interesting to explore the following questions: How accurate is an rth-order model of English for $1 < r \leq 10$? Given a maximum number of states to use in the model, what is the best way to use these states? And how do the behaviors of the statistics change when the statistics are applied to real languages as opposed to idealized languages of the model?

We compared the performance of the test statistics at distinguishing specific pairs of languages, focusing on BCa English versus uniform noise, and uniform noise versus BCa English. Therefore, our experiments say little about the performance of the test statistics on language-recognition problems with compound hypotheses (*e.g.* recognizing a single known language, and detecting unknown nonuniform language, and distinguishing unknown 0th-order noise from unknown 1st-order language), especially when the compound hypotheses are unrestrictive.

A significant weakness of our 1st-order Markov model of English is that many of its transition probabilities are zero (*e.g.* 168 out of 729 for BCa). This situation becomes worse for higher-order models. Not only do these zeroes make inefficient use of the model's parameters, they also create additional problems. Namely, as the number of zeroes increases, the our statistics perform worse using straightforward maximum-likelihood estimates of the transition probabilities, thereby increasing the need for fancier methods for estimating the transition probabilities. By contrast, hidden Markov models make more efficient of their parameters. In addition, hidden Markov models can more accurately model the behavior of characters that have multiple roles. For example, a hidden Markov model can capture the fact that, in English, the letter 'y' sometimes functions as a consonant and sometimes as a vowel.

In addition to their many applications in cryptanalysis, the test statistics are useful in variety of language-processing tasks—for example, recognizing the language of documents sent electronically. Also, although we have worked exclusively with strings of characters, it is sometimes possible to apply the techniques to sequences of words. For example, working at the word level is useful when decrypting codes (as opposed to ciphers) and when parsing and translating natural languages. Nevertheless, we urge caution in generalizing our findings to other languages, to higher-order models, and to other methods of estimating transition probabilities.

Throughout we made extensive use of our toolkit and of several standard data manipulation programs, including the Awk text-processing language and the ACE/gr interactive graphing program. Although these tools provided an adequate environment in which to carry out our project, this paper would be well-suited for presentation as a dynamic paper in which the reader could interactively modify data and experiments and explore the results.

Whereas some formal definitions of secure cryptosystems state precise conditions under which the cryptanalyst cannot distinguish valid messages from invalid messages (*e.g.*, see Goldwasser and

Micali [18]), we study techniques for so distinguishing messages produced by imperfect ciphers. We hope that our empirical study will be helpful to a variety of practitioners who wish to solve language-recognition problems with statistical techniques.

Acknowledgments

We are grateful to Peter Matthews for helpful remarks and suggestions. In addition, we thank James Mayfield for providing us with a copy of the Brown Corpus, and we thank Rita M. Doerr for answering some of our questions about this corpus. Thanks also to Robert Baldwin, Thomas Cain, James Mayfield, Bryan Olson, Raymond Pyle (Bell Atlantic), and James Reeds for comments. We are also grateful to Tomoko Shimakawa for computing several tail areas using SAS. Computer work was carried out on a DECstation 5000/200 and on a Silicon Graphics Iris/Indigo at the University of Maryland Baltimore County, and on a SUN Sparcstation at Bell Atlantic.

References

[1] Anderson, Roland. April 1989. Recognizing complete and partial plaintext. *Cryptologia*, 13(2): 161–166.

[2] Anderson, T. W. and Leo A. Goodman. 1957. Statistical inference about Markov chains. *Annals of Mathematical Statistics*, 28: 89–110.

[3] Baldwin, Robert W. and Alan T. Sherman. 1990. How we solved the $100,000 Decipher Puzzle: (16 hours too late). *Cryptologia*, 14(3): 258–284.

[4] Beker, Henry and Fred Piper. 1982. *Cipher Systems*. New York: John Wiley.

[5] Bhat, U. Narayan. 1984. *Elements of Applied Stochastic Processes*. New York: John Wiley.

[6] Billingsley, Patrick. 1961. *Statistical Inference for Markov Processes*. Chicago: University of Chicago Press.

[7] Billingsley, Patrick. 1961. Statistical methods in Markov chains. *Annals of Mathematical Statistics*, 32(1): 12–40.

[8] Callimahos, Lambros D. and William F. Friedman. 1959. *Military Cryptanalytics Part II*, Volume 1. Washington, DC: United States Government. [Available through Aegean Park Press, Laguna Hills, CA.]

[9] Church, Kenneth W. and William A. Gale. 1991. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computers, Speech, and Language*, 5(1).

[10] Crook, J. F. and I. J. Good. 1980. On the application of symmetric Dirichlet distributions and their mixtures to contingency tables, Part II. *Annals of Statistics*, 8(6): 1198–1218.

[11] Crook, James Flinn and Irving John Good. 1982. The powers and strengths of tests for multinomials and contingency tables. *Journal of the American Statistical Association*, 77(380): 793–802.

[12] Davies, Chris I. and Ravi Ganesan. 1993. BApasswd: A new proactive password checker. *Proceedings of the 16th National Computer Security Conference*. To appear.

[13] Francis, W. Nelson and Henry Kučera; with the assistance of Andrew W. Mackie. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton-Mifflin.

[14] Friedman, William F. 1925. The index of coincidence and its applications in cryptanalysis. Technical Paper, War Department, Office of the Chief Signal Officer. Washington, DC: United States Government Printing Office. [Available through Aegean Park Press, Laguna Hills, CA.]

[15] Friedman, William F. 1941. *Military Cryptanalysis Part IV: Transposition and Fractionating Systems.* Washington, DC: United States Government. [Available through Aegean Park Press, Laguna Hills, CA.]

[16] Ganesan, Ravi and Alan T. Sherman. Statistical techniques for language recognition: An introduction and guide for cryptanalysts. *Cryptologia.* To appear. [Preliminary versions are available as Technical Report TR CS-93-02, Computer Science Department, University of Maryland Baltimore County (February 1993), and as Technical Report CS-TR-3036/UMIACS-TR-93-16, University of Maryland College Park (February 1993).]

[17] Ganesan, Ravi and Alan T. Sherman. 1993. Statistical techniques for language recognition: An empirical study using real and simulated English (with supplement). Technical Report TR CS-93-03, Computer Science Department, University of Maryland Baltimore County.

[18] Goldwasser, Shafi and Silvio Micali. 1984 Probabilistic encryption. *JCSS*, 28(2): 270–299.

[19] Good, Irving John. 1965. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods.* Cambridge MA: MIT Press.

[20] Good, I. J.. 1967. A Bayesian significance test for multinomial distributions. *Journal of the Royal Statistical Society*, B 29(3): 399–431.

[21] Good, I. J. 1981. The fast calculation of the exact distribution of Pearson's Chi-squared and of the number of repeats within cells of a multinomial by using a fast Fourier transform. *Journal of Statistical Computation and Simulation*, 14(1): 71–78.

[22] Good, I. J. 1982. Comment on Patil and Taillie's paper on diversity, *Journal of the American Statistical Association*, 77(379): 561–563.

[23] Good, Irving John and James Flinn Crook. 1974. The Bayes/non-Bayes compromise and the multinomial distribution. *Journal of the American Statistical Association*, 69(347): 711–720.

[24] Good, I. J., T. N. Gover and G. J. Mitchell. 1970. Exact distributions for X^2 and for the likelihood-ratio statistic for the equiprobable multinomial distribution. *Journal of the American Statistical Association*, 65(329): 267–283.

[25] Knuth, Donald E. 1981. *Seminumerical Algorithms.* In *The Art of Computer Programming.* Vol. 2, Reading, MA: Addison-Wesley.

[26] Kučera, Henry and W. Nelson Francis W. 1967. *Computational Analysis of Present-Day American English.* Providence RI: Brown University Press.

[27] Kullback, Solomon. 1959. *Statistical Methods in Cryptanalysis.* Washington, DC: United States Government. [Available through Aegean Park Press, Laguna Hills, CA.]

[28] Kullback, Solomon. 1959. *Information Theory and Statistics.* New York: John Wiley.

[29] Kullback, S., M. Kupperman and H. H. Ku. 1962. Tests for contingency tables and Markov chains. *Technometrics*, 4(4): 573–608.

[30] Larsen, Richard J. and Morris L. Marx. 1990. *Statistics*, Englewood Cliffs, NJ: Prentice Hall.

[31] Lehmann, E. L. 1991. *Testing Statistical Hypotheses*, Pacific Grove, CA: Wadsworth.

[32] Rabiner, Lawrence R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2): 257–286.

[33] Levin, Bruce and James Reeds. 1977. Compound multinomial likelihood functions are unimodal: Proof of a conjecture of I. J. Good. *Annals of Statistics*, 5(1): 79–87.

[34] 1985. *SAS User's Guide: Basics*, Version 5 Edition. Cary, North Carolina: SAS Institute.

[35] Schrift, A. W. and A. Shamir. 1993. Universal tests for nonuniform distributions. *Journal of Cryptology*, 6(3): 119–133.

[36] Sinkov, Abraham. 1966. *Elementary Cryptanalysis: A Mathematical Approach*, New Mathematical Library No. 22. Washington D.C.: The Mathematical Association of America.

[37] Trivedi, Kishor Shridharbhai. 1982. *Probability and Statistics with Reliability, Queuing, and Computer Science Applications*, Englewood Cliffs, NJ: Prentice Hall.

[38] West, Eric N. and Oscar Kempthorne. 1972. A comparison of the Chi^2 and likelihood ratio tests for composite alternatives. *Journal of Statistical Computation and Simulation*, 1: 1–33.

Appendix A: A Statistical Approach to Language Recognition

We adopt a statistical approach to language recognition, where each language is modeled by a finite stationary Markov chain. Typically, two steps are required: an off-line step and an on-line step. The off-line step determines the parameters of the Markov model. For example, for natural languages we estimate the parameters from a large base sample. The on-line step compares the features of a candidate string with those of the known model. We make this comparison by computing a test statistic and by applying a decision procedure to the resulting value of the test statistic. Each off-line and on-line language sample consists of a string of characters; the extracted features are the k-gram frequency counts for some k. Although our methods apply for any k, for simplicity we state all formulas and carry out all experimental work with $k = 2$; that is, we work within 1st-order (bigram) models.

This appendix briefly reviews our language model and four important language-recognition problems. For more details, see our companion guide [16].

A.1 A Markov Model of Language

We represent a Markov model of language as a four-tuple $M = (m, \mathcal{A}, \{Y_t\}, r)$, where $m \in \mathbb{Z}^+$ is the number of *states*, $\mathcal{A}$ is the set of states, and $r \in \mathbb{Z}^+$ is the *order* of the Markov chain. For each time $t \in \mathbb{Z}^+$, the random variable Y_t takes on a value in $\mathcal{A}$. When run for n steps, we think of the chain as generating the string $Y_1 Y_2 \cdots Y_n$. In our experimental work we arbitrarily take $m = 27$, $r = 1$, and $\mathcal{A} = \{$'A', 'B', ..., 'Z', '␣'$\}$, where '␣' denotes the blank character.

For all $t > 1$, the conditional distribution of the random variable Y_t given Y_{t-1} is defined by an $m \times m$ matrix $P_M = [p_{ij}]$ of *transition probabilities*. For each $1 \le i, j \le m$, the constant p_{ij} is the *a priori* probability that the chain will move from state i to state j, given that the chain is in state i. In our Markov chain simulator, for the initial distribution of Y_1 we use the distribution given by the steady-state state (*i.e.* unigram) probabilities of M.

A.2 Four Language-Recognition Problems

Our experiments empirically determine how well five test statistics solve the following four language-recognition problems. In each problem we test some null hypothesis H_0 versus some alternative hypothesis H_1, given some candidate string. In each problem we assume that the candidate string was produced by some 1st-order Markov chain with transition probabilities P_M.

1. Recognizing a single known language.
   Test $H_0 : P_M = P_B$ versus $H_1 : P_M \neq P_B$, for some known matrix of transition probabilities P_B.
2. Distinguishing a known language from uniform noise.
   Test $H_0 : P_M = P_B$ versus $H_1 : P_M = P_U$, where $P_U = [1/m]$ is an $m \times m$ matrix of uniform transition probabilities.
3. Distinguishing unknown 0th-order noise from unknown 1st-order language.
   Test $H_0 : \text{order}(M) = 0$ versus $H_1 : \text{order}(M) = 1$.
4. Detecting non-uniform unknown language.
   Test $H_0 : P_M = P_U$ versus $H_1 : P_M \neq P_U$.

In our companion paper we explain these problems further, give examples, and point out a likelihood ratio test for each problem. For example, a cryptanalyst might use Problem 2 when distinguishing valid English plaintext from invalid plaintext, and Problem 4 is especially useful when the plaintext language is unknown.

Appendix B: Detailed Tabular Results of Experiment 1

Table 8: Sample means and standard deviations of the statistic $\diamond X^2$ computed on 100 randomly chosen strings of various lengths drawn from nine test languages. Values are mean ± standard deviation. Base language is BCa.

	English				Simulated English		Not English		
lg n	BCa	BCf	BCg	WSJ1	1st-order BCa	1st-order BCf	0th-order BCa	uniform noise	er repeated
1	−28.45±2.40	−28.22±3.61	−26.89±5.25	−26.95±4.33	−28.07±3.63	−27.90±2.80	−12.96±40.36	−6.74±47.57	−29.81±0.45
2	−23.78±4.41	−23.68±5.71	−23.92±3.96	−23.73±3.97	−23.48±5.20	−23.82±4.76	−5.23±31.78	13.79±44.76	−27.55±0.25
3	−17.56±9.30	−19.28±5.58	−19.03±10.11	−16.65±8.84	−19.15±4.64	−17.82±10.21	29.83±57.99	66.59±76.12	−24.91±0.17
4	−12.02±12.47	−13.01±9.68	−14.08±17.54	−15.20±5.60	−14.15±9.69	−13.64±8.78	60.15±68.36	111.15±75.68	−21.32±0.11
5	−11.06±7.62	−9.75±9.81	−9.87±8.93	−9.65±7.25	−11.18±5.96	−10.23±9.81	88.39±53.46	166.52±78.15	−16.36±0.08
6	−8.67±6.22	−7.58±10.17	−5.13±13.53	−8.03±7.77	−6.89±8.02	−7.77±8.12	116.99±55.50	190.15±63.49	−9.42±0.06
7	−3.87±10.54	−6.04±8.30	−2.50±16.45	−4.53±8.57	−4.94±7.67	−3.92±15.82	148.35±38.20	226.32±57.19	0.34±0.04
8	−4.25±5.79	−2.20±14.27	−2.77±9.46	0.17±11.26	−5.55±6.28	−3.80±10.91	155.53±33.91	254.43±47.46	14.11±0.03
9	−0.97±7.23	−0.08±9.10	2.10±12.61	0.24±9.46	−4.65±5.91	−2.05±9.45	188.72±30.54	282.49±40.44	33.56±0.02
10	2.39±7.09	3.76±9.10	7.70±22.68	8.59±11.18	−3.40±5.43	−1.45±7.14	231.14±28.19	351.35±35.86	61.03±0.01
11	5.97±6.85	8.57±9.47	12.52±19.36	16.45±16.97	−3.11±3.39	0.69±5.42	298.98±30.45	445.77±36.43	99.87±0.01
12	11.82±7.73	16.83±11.44	17.48±15.90	19.45±11.20	−2.51±2.97	1.24±4.26	412.90±31.63	615.91±43.94	154.79±0.01
13	16.52±8.41	27.95±22.65	24.07±18.88	26.26±10.42	−1.87±2.37	4.66±4.01	570.08±36.99	836.60±39.88	232.45±0.00
14	20.59±8.13	37.50±24.62	34.35±15.80	36.93±10.73	−1.82±1.84	9.40±4.19	804.18±29.58	1171.67±34.19	342.28±0.00
15	24.95±8.81	45.92±25.73	48.13±19.65	45.17±11.11	−1.43±1.75	16.67±3.73	1137.11±30.47	1648.94±44.91	497.60±0.00
16	24.97±7.28	51.82±23.20	59.37±16.39	58.75±6.76	−1.12±1.30	29.99±4.42	1616.53±26.94	2339.87±35.78	717.24±0.00
17	25.18±7.92	67.39±20.78	80.02±15.26	77.22±5.96	−0.35±1.59	49.01±4.68	2283.82±22.16	3311.60±27.99	1027.89±0.16

Table 9: Sample means and standard deviations of the statistic $\diamond ML$ computed on 100 randomly chosen strings of various lengths drawn from nine test languages. Values are mean ± standard deviation. Base language is BCa.

	English				Simulated English		Not English		
lg n	BCa	BCf	BCg	WSJ1	1st-order BCa	1st-order BCf	0th-order BCa	uniform noise	er repeated
1	−29.75±0.59	−29.77±0.72	−29.58±0.80	−29.48±0.65	−29.71±0.66	−29.63±0.65	−28.97±1.05	−29.52±2.08	−30.01±0.17
2	−27.53±0.73	−27.49±0.62	−27.50±0.66	−27.46±0.70	−27.43±0.67	−27.51±0.57	−26.45±1.03	−26.56±1.83	−28.02±0.09
3	−24.79±0.76	−24.91±0.73	−24.94±0.70	−24.66±0.71	−24.95±0.68	−24.94±0.72	−23.29±1.21	−22.98±1.91	−25.59±0.06
4	−21.88±0.83	−21.99±0.90	−22.12±0.80	−21.92±0.93	−22.11±0.82	−21.98±0.80	−19.59±1.57	−18.60±1.64	−22.31±0.04
5	−18.88±0.98	−18.78±0.85	−18.82±0.90	−18.59±0.78	−18.66±0.86	−18.79±0.79	−14.71±1.37	−13.00±1.58	−17.78±0.03
6	−15.59±0.88	−15.51±0.89	−15.39±0.93	−15.26±0.83	−15.30±0.99	−15.35±0.83	−9.08±1.63	−6.14±1.61	−11.44±0.02
7	−11.77±1.42	−12.24±1.07	−12.26±0.98	−11.84±1.09	−12.16±0.92	−12.27±0.93	−2.65±1.81	2.67±1.43	−2.54±0.01
8	−8.50±1.34	−8.94±1.20	−8.66±1.23	−8.23±1.60	−9.53±0.90	−9.48±0.88	5.19±1.80	12.75±1.72	10.03±0.01
9	−5.44±1.76	−5.26±1.31	−5.06±1.70	−4.75±1.70	−7.22±1.00	−7.06±0.87	15.42±1.66	25.81±1.72	27.78±0.01
10	−1.26±2.16	−1.64±1.51	−1.61±2.17	−0.44±2.32	−5.46±0.83	−5.23±0.95	29.43±1.84	43.40±1.89	52.85±0.01
11	2.67±2.63	3.51±2.79	3.55±2.95	4.41±2.85	−4.21±0.86	−3.10±1.01	49.37±1.91	68.01±2.10	88.30±0.00
12	7.12±3.64	9.53±3.43	8.72±3.86	9.80±3.60	−3.30±1.01	−1.11±0.95	79.20±2.11	105.99±2.21	138.43±0.00
13	12.40±4.84	17.03±5.27	15.54±4.94	16.39±3.75	−2.48±0.92	1.96±0.93	121.91±2.48	160.03±2.52	209.32±0.00
14	17.26±5.89	25.70±6.68	25.07±5.59	24.37±5.34	−1.45±1.05	6.88±1.27	183.50±1.95	237.78±2.72	309.56±0.00
15	22.89±7.29	34.10±6.92	35.85±6.10	33.32±6.21	−0.69±1.13	14.48±1.42	271.49±2.05	348.30±2.78	451.32±0.00
16	24.52±6.47	41.61±6.18	49.44±8.25	47.66±7.76	−0.04±0.99	26.34±1.70	396.41±1.75	505.44±2.42	651.80±0.00
17	25.68±7.17	57.11±6.79	69.65±9.68	66.61±8.48	0.93±1.27	44.47±1.91	573.09±1.40	726.99±1.94	935.32±0.00

Table 10: Sample means and standard deviations of the statistic $\diamond IND$ computed on 100 randomly chosen strings of various lengths drawn from nine test languages. Values are mean ± standard deviation.

	English				Simulated English		Not English		
lg n	BCa	BCf	BCg	WSJ1	1st-order BCa	1st-order BCf	0th-order BCa	uniform noise	er repeated
1	−31.86±0.00	−31.86±0.00	−31.86±0.00	−31.86±0.00	−31.86±0.00	−31.86±0.00	−31.86±0.00	−31.86±0.00	−31.86±0.00
2	−28.90±0.47	−28.98±0.56	−28.90±0.47	−28.98±0.52	−28.86±0.44	−28.97±0.62	−29.23±0.96	−29.08±0.69	−28.63±0.00
3	−25.42±0.65	−25.61±0.61	−25.51±0.71	−25.43±0.63	−25.67±0.79	−25.69±0.75	−25.80±0.81	−25.43±0.61	−27.36±0.00
4	−21.63±0.72	−21.65±0.79	−21.77±0.76	−21.77±0.81	−22.00±0.91	−21.85±0.95	−22.22±1.00	−21.03±0.80	−25.38±0.00
5	−17.55±0.89	−17.33±0.90	−17.42±0.80	−17.20±0.73	−17.44±0.92	−17.54±0.92	−17.87±1.09	−15.43±0.73	−22.58±0.00
6	−12.64±0.80	−12.43±0.82	−12.43±0.82	−12.54±0.88	−12.53±0.87	−12.52±0.83	−13.20±1.14	−9.07±0.66	−18.64±0.00
7	−6.84±0.91	−6.67±0.98	−6.83±0.84	−6.68±0.97	−6.86±0.87	−6.94±0.88	−8.91±1.17	−2.96±0.50	−13.09±0.00
8	0.35±1.15	0.32±1.08	0.29±0.97	0.26±1.10	−0.33±0.85	−0.05±0.89	−4.79±0.95	2.74±0.47	−5.27±0.00
9	9.29±1.34	9.82±1.15	9.67±1.12	9.34±1.38	8.22±0.78	8.62±0.80	−1.71±0.92	6.74±0.72	5.78±0.00
10	22.29±1.91	22.64±1.36	22.41±1.39	22.15±1.83	20.40±1.13	20.96±1.04	0.86±0.99	8.02±0.91	21.40±0.00
11	40.47±1.92	41.50±1.72	40.95±1.79	40.27±1.80	37.50±0.97	38.82±0.95	2.41±0.92	6.93±0.95	43.48±0.00
12	65.89±2.14	67.54±2.13	67.19±1.88	66.00±2.24	62.72±1.07	64.39±1.02	3.42±1.15	5.73±0.97	74.69±0.00
13	102.61±2.83	105.68±2.51	104.69±2.10	102.09±2.37	99.05±0.95	101.56±1.01	4.18±0.97	5.26±0.93	118.84±0.00
14	155.00±2.77	158.48±2.74	158.36±2.82	153.72±3.20	151.07±1.09	154.54±1.25	5.03±1.09	5.01±0.99	181.27±0.00
15	228.36±2.41	233.75±2.70	233.16±2.81	226.63±3.15	225.21±0.96	230.23±1.26	5.40±0.93	4.79±0.99	269.55±0.00
16	332.31±2.50	340.10±2.61	340.13±2.86	331.32±3.78	330.48±0.90	337.69±0.98	5.43±0.80	4.89±0.79	394.41±0.00
17	480.13±2.76	491.29±2.44	491.19±2.31	478.96±3.82	479.62±0.77	489.85±0.84	5.96±1.30	4.86±1.10	570.97±0.01

Table 11: Sample means and standard deviations of the statistic $\hat{S}$ computed on 100 randomly chosen strings of various lengths drawn from nine test languages. Values are mean ± standard deviation. Base language is BCa.

	English				Simulated English		Not English		
lg n	BCa	BCf	BCg	WSJ1	1st-order BCa	1st-order BCf	0th-order BCa	uniform noise	er repeated
1	0.13±0.83	0.12±0.97	−0.17±1.21	−0.27±1.00	0.07±0.95	−0.04±0.88	−1.27±2.04	−1.13±2.89	0.56±0.21
2	0.14±1.00	0.06±0.90	0.11±0.95	0.02±0.94	0.01±1.00	0.08±0.81	−1.72±1.66	−1.77±2.65	0.92±0.12
3	−0.09±1.01	−0.02±0.99	0.09±0.90	−0.29±0.97	0.03±0.83	−0.02±0.92	−2.64±2.09	−3.09±2.95	1.44±0.08
4	−0.13±1.00	0.03±0.99	0.13±0.93	−0.13±0.95	0.04±0.90	−0.04±0.85	−3.83±2.43	−4.60±2.88	2.11±0.05
5	0.04±1.00	0.02±0.91	0.01±0.95	−0.14±0.88	−0.13±0.87	−0.04±0.86	−5.78±1.96	−6.68±2.93	3.03±0.04
6	0.10±0.84	0.17±0.95	0.05±0.94	−0.20±0.84	−0.15±1.00	−0.08±0.81	−8.14±2.16	−9.22±2.94	4.32±0.03
7	−0.33±1.18	0.48±1.05	0.27±0.90	−0.15±0.96	0.03±0.93	0.19±0.91	−11.44±2.32	−13.94±2.51	6.14±0.02
8	−0.01±1.08	0.45±1.04	0.18±1.01	−0.52±1.02	0.10±0.99	0.37±0.89	−16.22±2.17	−19.05±2.88	8.69±0.01
9	0.01±1.27	0.45±1.03	0.35±1.05	−0.50±1.10	−0.10±0.89	0.31±0.80	−23.20±1.92	−27.48±2.77	12.30±0.01
10	−0.14±1.39	0.56±1.12	0.50±1.15	−0.78±1.18	0.04±1.01	0.66±0.93	−32.78±2.09	−39.57±2.90	17.41±0.01
11	−0.07±1.46	0.85±1.38	0.61±1.47	−1.18±1.63	−0.15±0.91	0.84±0.84	−46.04±2.06	−55.13±3.10	24.63±0.00
12	−0.10±1.76	0.98±1.59	1.05±1.83	−1.42±1.41	−0.04±0.82	1.16±0.88	−65.19±2.19	−78.79±2.84	34.83±0.00
13	−0.07±2.11	1.79±1.91	1.55±1.92	−2.25±1.83	0.02±0.76	1.68±0.90	−92.14±2.50	−111.21±3.31	49.27±0.00
14	0.15±2.24	1.80±2.25	2.56±2.51	−3.15±1.97	0.06±0.84	2.20±0.99	−130.64±1.83	−157.19±3.65	69.68±0.00
15	−0.35±2.50	3.14±2.17	2.72±2.50	−4.53±1.81	0.17±0.67	3.26±1.04	−184.85±1.98	−222.19±3.23	98.54±0.00
16	−0.62±2.59	4.67±2.13	4.40±2.24	−6.33±1.91	0.17±0.65	4.52±1.07	−261.74±1.81	−314.54±2.89	139.35±0.00
17	−1.09±3.07	6.89±2.28	6.10±1.77	−8.80±1.44	0.33±0.45	6.41±1.16	−369.98±1.23	−444.31±2.03	197.08±0.00

Table 12: Sample means and standard deviations of the statistic IC computed on 100 randomly chosen strings of various lengths drawn from nine test languages. Values are mean ± standard deviation.

	English				Simulated English		Not English		
lg n	BCa	BCf	BCg	WSJ1	1st-order BCa	1st-order BCf	0th-order BCa	uniform noise	er repeated
1	0.0000±0.0000	0.0000±0.0000	0.0000±0.0000	0.0000±0.0000	0.0000±0.0000	0.0000±0.0000	0.0000±0.0000	0.0000±0.0000	0.0000±0.0000
2	0.0000±0.0000	0.0033±0.0333	0.0000±0.0000	0.0000±0.0000	0.0067±0.0469	0.0067±0.0469	0.0000±0.0000	0.0000±0.0000	0.3333±0.0000
3	0.0024±0.0124	0.0014±0.0082	0.0029±0.0132	0.0038±0.0130	0.0081±0.0225	0.0076±0.0222	0.0095±0.0352	0.0014±0.0082	0.4286±0.0000
4	0.0042±0.0064	0.0055±0.0079	0.0070±0.0113	0.0059±0.0070	0.0076±0.0108	0.0082±0.0108	0.0038±0.0059	0.0009±0.0031	0.4667±0.0000
5	0.0074±0.0048	0.0072±0.0054	0.0072±0.0056	0.0066±0.0048	0.0082±0.0055	0.0084±0.0052	0.0054±0.0041	0.0014±0.0016	0.4839±0.0000
6	0.0085±0.0033	0.0088±0.0033	0.0087±0.0034	0.0081±0.0033	0.0085±0.0035	0.0085±0.0032	0.0060±0.0033	0.0014±0.0010	0.4921±0.0000
7	0.0089±0.0021	0.0101±0.0027	0.0094±0.0021	0.0086±0.0020	0.0089±0.0021	0.0094±0.0021	0.0060±0.0018	0.0014±0.0004	0.4961±0.0000
8	0.0094±0.0019	0.0099±0.0017	0.0096±0.0018	0.0088±0.0014	0.0090±0.0014	0.0094±0.0015	0.0056±0.0012	0.0014±0.0002	0.4980±0.0000
9	0.0095±0.0012	0.0099±0.0012	0.0100±0.0013	0.0091±0.0011	0.0088±0.0008	0.0092±0.0008	0.0056±0.0007	0.0014±0.0001	0.4990±0.0000
10	0.0094±0.0008	0.0098±0.0008	0.0099±0.0008	0.0091±0.0008	0.0090±0.0007	0.0094±0.0006	0.0056±0.0006	0.0014±0.0001	0.4995±0.0000
11	0.0093±0.0008	0.0099±0.0007	0.0098±0.0008	0.0090±0.0006	0.0089±0.0004	0.0094±0.0004	0.0056±0.0004	0.0014±0.0000	0.4998±0.0000
12	0.0093±0.0005	0.0098±0.0006	0.0099±0.0007	0.0089±0.0005	0.0090±0.0003	0.0094±0.0003	0.0057±0.0003	0.0014±0.0000	0.4999±0.0000
13	0.0092±0.0005	0.0098±0.0005	0.0097±0.0006	0.0087±0.0004	0.0090±0.0002	0.0094±0.0002	0.0057±0.0002	0.0014±0.0000	0.4999±0.0000
14	0.0092±0.0003	0.0096±0.0004	0.0098±0.0006	0.0087±0.0003	0.0090±0.0001	0.0094±0.0002	0.0057±0.0001	0.0014±0.0000	0.5000±0.0000
15	0.0091±0.0002	0.0095±0.0003	0.0095±0.0004	0.0086±0.0002	0.0090±0.0001	0.0094±0.0001	0.0057±0.0001	0.0014±0.0000	0.5000±0.0000
16	0.0090±0.0002	0.0094±0.0001	0.0095±0.0003	0.0086±0.0001	0.0090±0.0000	0.0094±0.0001	0.0057±0.0000	0.0014±0.0000	0.5000±0.0000
17	0.0090±0.0002	0.0095±0.0001	0.0095±0.0002	0.0086±0.0001	0.0090±0.0000	0.0094±0.0001	0.0057±0.0000	0.0014±0.0000	0.5000±0.0000