The binomial cumulative distribution function, or, is my system better than yours?

Barbara Di Eugenio*, Michael Glass*, Michael J. Scott†

*Department of Computer Science
University of Illinois, Chicago, IL USA
{bdieugen, mglass}@cs.uic.edu

†Department of Mechanical and Industrial Engineering
University of Illinois, Chicago, IL USA
mjscott@uic.edu

Abstract

In human language technology, it is becoming more and more common to run systematic evaluations in which two or more systems, or two or more versions of the same system, are pitted one against the other. We propose the binomial cumulative distribution function as a way to assess the cumulative effect of the measures collected in such evaluations. We present an application of this measure to the evaluation of the NL interface to an Intelligent Tutoring System. We conclude by discussing a few issues pertaining to this statistical measure.

1. Introduction

In human language technology, it is becoming more and more common to run systematic evaluations in which two or more systems, or two or more versions of the same system, are pitted one against the other (Young, 1997; Carenini and Moore, 2000; Reiter et al., 2001). Such evaluations are generally conducted by having each system run in the same condition: for example, different groups of users of comparable size interact with each system following a prepared script. During the experiments, a number of measures are collected. Measures may concern performance (e.g., time on task), or usability (i.e., answers to questions such as, was the system friendly?). These measures are then assessed in a pairwise fashion (Young, 1997; Carenini and Moore, 2000; Reiter et al., 2001). For example, to show that system B is better than system A, one could stipulate that there must be at least one statistically significant measure in favor of B and no significant measure in favor of A.

However, reality is often much murkier that the ideal result just described. A typical result of an evaluation may be that out of ten measures eight favor B and two favor A, but only two show statistical significance and those two point to opposite conclusions. In these situations, the evaluation does not support any conclusion on whether B is better than A. However, because the vast majority of measures is in favor of one of the evaluated systems, a legitimate question arises: does the cumulative effect of the measures in favor of system B warrant the conclusion that B is better than A?

The binomial cumulative distribution function (or sign test (Siegel and Castellan, 1988)) is the statistical measure that can answer this question. To our knowledge, it is not used in Human Language Technology. To use it, each measure must be labeled as a success for one of the evaluated systems. In the example above, we have 2 successes for A and 8 for B. The binomial cumulative distribution function (BCDF for short) answers the question: what is the probability that m successes out of n independent measures are due to chance (in our example, 8 successes out of 10 measures)?

We will illustrate the usage of the BCDF in an evaluation we ran to assess the improvement to the NL interface of an Intelligent Tutoring System (ITS). We pitted two versions of the same system one against the other; the two versions differ in that the first produces very repetitive feedback, the second more fluent feedback by using aggregation strategies. We collected 10 measures pertaining to the student's performance, the knowledge s/he acquired, and the usability of the system.

By using the conventional pairwise assessment of measures, only one measure approaches, but does not reach, statistical significance in favor of the second version of the system. However, all measures but one show a moderate preference for the second version. The BCDF confirms that the cumulative effect of these measures is not due to chance, i.e., it shows that the second version of the system outperforms the first.

In the last part of the paper, we will address a few issues pertaining to the usage of the BCDF. They include how to deal with ties and with apparently contradictory results. The latter situation arises when one or two statistically significant measures favor system A, but the cumulative effect favors system B.

2. The binomial cumulative distribution function

The BCDF is applied to the case of two related samples when the experimenter wishes to assess whether the two conditions are different. The null hypothesis tested by means of the BCDF is

$$P(X_i > Y_i) = P(X_1 < Y_i) = 0.5$$

In statistics books, the BCDF is usually applied to experiments in which a subject receives some treatment, and the experimenter is interested in the changes in the variable of interest before and after the treatment. For example, does the weight of a college freshman go up or down after the first semester? does the attitude of adults with respect to severity of punishment for juvenile delinquents change after seeing a certain documentary (Siegel and Castellan, 1988)? However, nothing in the assumptions underlying the BCDF prevents its application to different situations. The only assumption underlying the test is that the variable under consideration has a continuous distribution. It does not require that the subjects are all drawn from the same population, it only requires *matched pairs*, i.e., that within each pair the experimenter has achieved matching with respect to the variable of interest.

To apply the BCDF to the evaluation of two systems, or of two versions of the same system, it is then necessary to collect the same (independent) measures under the same condition for each system. Each matched pair consists of the pair of values of measure X, one value for system A, the other for system B. Each pair is coded as a success for system A or B, as in Table 4.

To compute the probability that *m* successes out of *n* independent measures are due to chance, we start by computing the BCDF through $m-1$ for sample size n and probability $p = 0.5$, i.e., $bcdf(m-1, n, 0.5)$. The BCDF is computed as follows, with $p = q = 0.5$:

$$\sum_{i=0}^{m-1} \binom{n}{i} p^i q^{n-i}$$

It gives us the probability that out of n trials, the number of successes will fall between 0 and $m-1$, inclusive. Thus, $1 - bcdf(m-1, n, 0.5)$ will give us the probability that m or more successes out of n are due to chance.

The test based on the BCDF can be two-tailed or one-tailed. A two-tailed test simply measures whether the two conditions are different, regardless of which one is better. A one-tailed test measures which condition is better. The one-tailed test is appropriate for system evaluation of the sort we describe in this paper.

The BCDF is usually referred to as the Sign Test in statistics books (Siegel and Castellan, 1988). We keep the name BCDF because we are using it on slightly different kinds of data.

3. An illustrative example

We will illustrate the usage of the BCDF via an evaluation we ran of the NL interface to an ITS. We improved the feedback capability of an existing ITS, and we evaluated the two versions of the system via a user study. The ITS in question teaches troubleshooting of a home heating system. It is written within DIAG (Towne, 1997), an authoring system to develop ITSs to troubleshoot complex mechanical systems and circuitry.

A typical session with a DIAG application presents the student with a series of troubleshooting problems of increasing difficulty. To solve the problem, the student tests indicators and tries to infer which faulty part (RU) may cause the detected abnormal states. RU stands for *replaceable unit*, because the only course of action open to the student to fix the problem is to replace faulty components in the graphical simulation. Figure 1 shows the furnace system, one of the subsystems of the home heating system in our DIAG application. Figure 1 includes indicators (e.g., the gauges labeled Burner Motor RPM and Water Temperature), replaceable units, and other complex modules (e.g., the Oil Burner) that contain indicators and replaceable units. Complex components are zoomable.

At any point, the student can consult the built-in tutor via the Consult menu, activated by the Consult button (cf. Figure 1). For example, if the student has noted an abnormal reading of an indicator, s/he can ask the tutor for a hint regarding which RUs may cause the problem. After deciding which content to communicate, the original DIAG system (*DIAG-orig*) uses very simple templates to assemble the text to present to the student. The result is that the feedback that DIAG provides is repetitive, both inter- and intra-turn. In many cases, the feedback presents a single long list of many parts. The top part of Figure 2 shows the reply originally provided by DIAG to a request of information regarding the indicator named “Visual Combustion Check”.

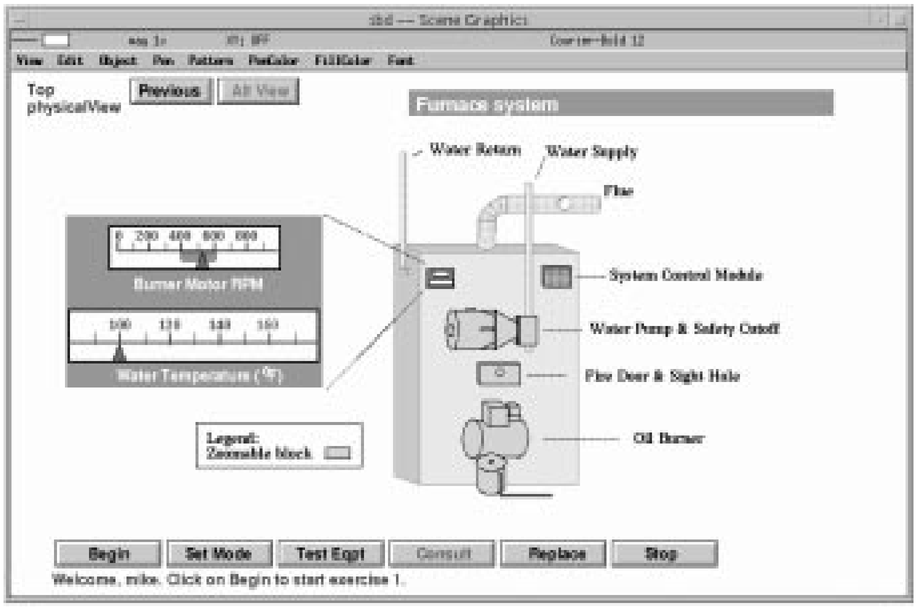
We set out to rapidly improve DIAG’s feedback mechanism. Our main goals were to to assess whether simple NLG techniques would lead to measurable improvements in the system’s output, and to conduct a systematic evaluation that would focus on language only. Thus, we did not change the tutoring strategy, or alter the interaction between student and system in any way. Rather, we concentrated on improving each single turn by avoiding excessive repetitions. We chose to achieve this by: introducing syntactic aggregation (Dalianis, 1996; Huang and Fiedler, 1996; Shaw, 1998; Reape and Mellish, 1998) and what we may call *functional aggregation*, namely, relating the parts mentioned to the structure of the system; and improving the format of the output.

To improve on *DIAG-orig*, we integrated the original system with EXEMPLARS (White and Caldwell, 1998), a surface generator from CoGenTex Inc. We call the second version of the system *DIAG-NLP*. EXEMPLARS is an object-oriented, rule based generator. It mixes template-style and more sophisticated types of text planning. The bottom part of Figure 2 shows our sentence planning component at work. The revised output groups the parts under discussion by the system modules that contain them (Oil Burner and Furnace System), and by the likelihood that a certain RU causes the observed symptoms. Notice how the *Ignitor Assembly* is singled out in the revised answer. Among all mentioned units, it is the only one that cannot cause the symptom. This fact is lost in the original answer.

3.1. Evaluation

We conducted an empirical evaluation designed as a between-subject study. Both groups interact with the same DIAG application that teaches them to troubleshoot a home-heating system. One group interacts with *DIAG-orig* and the other with *DIAG-NLP*.

Seventeen subjects were tested in each group. The 34 subjects were all science or engineering majors affiliated with our university. Each subject read some short material about home heating, went through the first problem as a



Welcome, mike. Click on Begin to start exercise 1.

Figure 1: A screen from a DIAG application on home heating

trial run, then continued through the curriculum on his/her own. The curriculum consists of three problems of increasing difficulty. As there was no time limit, every student solved every problem. At the end of the experiment, each subject was administered a questionnaire.

A detailed log was collected for each subject. It includes, for each problem: whether the problem was solved; total time, and time spent reading feedback; how many and which indicators and RUs the subject consults DIAG about; how many, and which RUs the subject replaces.

The questionnaire is divided into three parts. The first part tests the subject's understanding of the domain. Because the questions asked are fairly open ended, this part was scored as if grading an essay. The second part of the questionnaire asks the subject to rate the system's feedback along four dimensions on a scale from 1 to 5 (see Table 3). The third part concerns the subjects' remembering their actions, specifically, the RUs they replaced. We quantify the subjects' recollections in terms of precision and recall with respect to the log of the subject's actions that the system collects. We also compute the F-measure, $\frac{(\beta^2+1)PR}{\beta^2 P+R}$, that smooths precision and recall off, with $\beta = 1$. In Table 2, we report the F-measure (precision and recall are .74 and .73 respectively for *DIAG-orig*, and .65 and .63 for *DIAG-NLP*).

	DIAG-orig	DIAG-NLP
Total Time	29.8'	28.0'
Feedback Time	6.9'	5.4'
Indicator consultations	11.4	5.9
RU consultations	19.2	18.1
Parts replaced	3.85	3.33

Table 1: Performance measures

Results. Tables 1, 2, and 3 show the results for the cumulative measures across the three problems (individual problems show the same trends).

	DIAG-orig	DIAG-NLP
Essay score	81/100	83/100
RU recollection	.72	.63

Table 2: Learning and recollection measures

	DIAG-orig	DIAG-NLP
Usefulness	4.35	4.47
Helped stay on right track	4.35	4.35
Not misleading	4.00	4.12
Conciseness	3.47	3.76

Table 3: Usability measures

Differences on individual measures are not statistically significant; *indicator consultations* comes closest to statistical significance, as it exhibits a non-significant trend in favor of *DIAG-NLP* (Mann-Whitney test, U=98, p=0.11). However, given that almost all individual measures are in favor of *DIAG-NLP*, we use the BCDF to assess whether *cumulatively* these measures show that *DIAG-NLP* outperforms *DIAG-orig*.

We consider only independent measures (total time and feedback time in Table 1 are not independent). For each measure, we decide for which system its value indicates a success — the magnitude of the difference is irrelevant.

Table 4 combines the independent measures from Tables 1, 2 and 3 and shows whether they represent a success for *DIAG-orig* or *DIAG-NLP*. Because *Helped stay on right track* is a tie and can therefore be considered a success for either system, we will report two sets of statistics (see discussion of ties below). The probability of 9 successes out of 10 measures is p = 0.011, of 8 successes out of 10 measures is p = 0.0545 (in the former case, we consider *Helped stay on right track* a success for *DIAG-NLP*, in the latter, for *DIAG-orig*). The former is significant, the latter marginally significant, and in fact, very close to significance (we fol-

	DIAG-orig	DIAG-NLP
Total Time		√
Indicator consultations		√
RU consultations		√
Parts replaced		√
Essay score		√
RU recollection	√	
Usefulness		√
Helped stay on right track	√	√
Not misleading		√
Conciseness		√

Table 4: Successes for each system

low standard practice and consider p $\leq$ 0.05 significant, and 0.05< p $\leq$0.1 marginally significant). It may be questioned whether *Total Time* is an independent measure, as total time may have decreased in *DIAG-NLP* because of fewer consultations. If we leave it out, the probability of 8 successes out of 9 is p = 0.02 , and of 7 successes out of 9 is p = 0.09 (respectively significant and marginally significant). Note that if we eliminate *Helped stay on right track* altogether, as suggested by (Siegel and Castellan, 1988), we obtain p=0.02 and p=0.035, respectively, according to whether *Total Time* is included or not (both significant). We can then conclude that the better measures for *DIAG-NLP*, albeit individually not statistically significant, cumulatively show that *DIAG-NLP* outperforms *DIAG-orig*.

4. Further discussion of the BCDF

4.1. Ties

In our discussion of the DIAG application, we provided three sets of measures, according to how the tied measure is considered. In one case we consider it as a success for DIAG-orig, in the second a success for DIAG-NLP, in the third case we throw it out altogether.

Statistics books such as (Siegel and Castellan, 1988) do in fact suggest that ties should be disregarded when applying the Sign Test. There are two justifications for such an approach. First, ties are in theory impossible and in practice extremely unlikely because the variable of interest is continuous (Walpole et al., 1998). Second, disregarding ties will not appreciably affect the results, if the number of ties is small with respect to the sample size n.[1]

However, it seems to us that in general this approach cannot be correct. It amounts to disregarding evidence that there is no difference between the two conditions, when in fact the BCDF is computed to try to reject the null hypothesis, i.e., to prove they are different. Such an approach boosts success in the case of a high number of ties. Suppose that in our evaluation of *DIAG-NLP* we had 10 more measures, all of them ties, with a resulting sample size of $n = 20$. If we disregarded the 11 ties, we would conclude that 9 measures out of 9 are in favor of *DIAG-NLP*, i.e., that *DIAG-NLP* is much better than it really is.

We propose the following ways to deal with ties. In case of a single tie, two sets of measures can be provided, one in which the tie is turned into a success for system A, one in which it is turned into a success for system B, as we have done in this paper. This has the advantage of leaving the sample size n unchanged. However, even if the single tie were disregarded, we expect the results not to change much. If there are two or more ties, we propose that half of the ties are turned into successes for system A, and half for system B. In the case of $2k + 1$ tie, the remaining tie can be disregarded. In this way, we don't change the sample size n, or only change it minimally.

4.2. Strength of results, and contradictory conclusions

Two other issues related to the BCDF may be addressed:

1. If a large number of observations favor one side at relatively strong levels of significance, none of which are statistically significant, then the BCDF seems to be an underestimate of the significance of the difference.

2. What should be done if a large number of measures favor one side without statistical significance for any one measure, but a small number favor the other side at statistical significance? These two results are apparently contradictory.

Consider the following example, with ten measures (we don't use the DIAG example because there is no statistically significant measure). For each measure we have a p-value, i.e., a significance level. Suppose two measures favor system A, with p-values

$$(0.02, 0.17)$$

and eight measures favor system B, with p-values

$$(0.06, 0.10, 0.20, 0.30, 0.33, 0.35, 0.4, 0.4)$$

This example illustrates both situations described above. The BCDF gives a significance level of 0.0547 for system B, calculated with $1 - bcdf(7, 10, 0.5)$. However, the BCDF only estimates the probability that eight of ten measures will favor B randomly, and thus overestimates the probability that eight of ten measures will favor B at a significance level no greater than 0.4. The proposed test is to consider the probability that, if system B is truly equivalent to system A, eight of ten measures will have p-values less than or equal to 0.4. This probability is 0.0123, calculated by $1 - bcdf(7, 10, 0.4)$. The new test is more accurate, and gives a stronger indication of significance.

It is, however, necessary to also calculate the significance level for A over B, based on the two measures in A's favor. Using the same method, $1 - bcdf(1, 10, 0.17)$ gives 0.5270 as the probability that at least two measures will favor A at the 0.17 significance level or better. In this case, we can also consider the chance that one measure out of ten will yield a p-value of 0.02 (the one significant measure in A's favor): $1 - bcdf(0, 10, 0.02)$ gives 0.0861 as the level of significance in A's favor. Recall that 0.0123 is the level of significance in B's favor. This is fairly strong evidence

[1] It is questionable whether the number of ties in the two examples in (Siegel and Castellan, 1988) is really negligible: in one, there are 3 ties out of $n = 17$, in the other, 15 out of $n = 100$.

Measure	p
1-bcdf(7,10,0.4)	0.0123
1-bcdf(5,10,0.35)	0.0949
1-bcdf(4,10,0.33)	0.2064
1-bcdf(3,10,0.3)	0.3504
1-bcdf(2,10,0.2)	0.3222
1-bcdf(1,10,0.1)	0.2639
1-bcdf(0,10,0.06)	0.4614

Table 5: p for different subsets of measures

that B outperforms A overall; still, in this case it seems that it is worth considering individual performance measures.

The calculation of the p-value in favor of A shows that it will not always be the case that the strongest significance will be obtained by considering the probability that all measures in favor of one system exceed the weakest measure in favor of that system. For each system, one probability calculation can be made for each subset of measures in favor of that system, and the strongest significance should be considered. We saw above which of two measures in favor of A was the stronger. For B, there are seven possible measures (not eight, because two measures have the same p-value, $p = 0.4$), as illustrated in Table 5. Note that the significance level is not monotonic in the number of measures considered.

5. Conclusions

We have proposed that the binomial cumulative distribution function (or sign test) can be used to assess the cumulative effect of the measures collected in systematic evaluations that pit two systems, or two versions of the same system, one against the other. We have presented an application of the BCDF to the evaluation of the NL interface to an Intelligent Tutoring System. We have also discussed a few issues pertaining to the usage of the BCDF. They include how to deal with ties, and with apparently contradictory results. The latter situation arises when one or two statistically significant measures favor system A, but the cumulative effect favors system B.

Acknowledgements. This work is supported by grants N00014-99-1-0930 and N00014-00-1-0640 from the Office of Naval Research, Cognitive, Neural and Biomolecular S&T Division. We are grateful to CoGenTex Inc., in particular to Mike White, for making EXEMPLARS available to us.

6. References

Giuseppe Carenini and Johanna D. Moore. 2000. An empirical study of the influence of argument conciseness on argument effectiveness. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong.

Hercules Dalianis. 1996. *Concise Natural Language Generation from Formal Specifications*. Ph.D. thesis, Department of Computer and Systems Science, Stocholm UNiversity. Technical Report 96-008.

Xiaoron Huang and Armin Fiedler. 1996. Paraphrasing and aggregating argumentative text using text structure. In *Proceedings of the 8th International Workshop on Natural Language Generation*, pages 21–30, Sussex, UK.

Mike Reape and Chris Mellish. 1998. Just what *is* aggregation anyway? In *Proceedings of the European Workshop on Natural Language Generation*, Toulouse, France.

Ehud Reiter, Roma Robertson, A. Scott Lennox, and Liesl Osman. 2001. Using a Randomised Controlled Clinical Trial to Evaluate an NLG System. In *ACL-2001, Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 434–441, Toulouse, France.

James Shaw. 1998. Segregatory coordination and ellipsis in text generation. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pages 1220–1226, Montreal, Canada.

Sidney Siegel and N. John Castellan, Jr. 1988. *Nonparametric statistics for the behavioral sciences*. McGraw Hill.

Douglas M. Towne. 1997. Approximate reasoning techniques for intelligent diagnostic instruction. *International Journal of Artificial Intelligence in Education*.

Ronald E. Walpole, Raymond E. Myers, and Sharon L. Myers. 1998. *Probability and Statistics for Engineers and Scientists*. Prentice Hall, sixth edition.

Michael White and Ted Caldwell. 1998. Exemplars: A practical, extensible framework for dynamic text generation. In *Proceedings of the Ninth International Workshop on Natural Language Generation*, pages 266–275, Niagara-on-the-Lake, Canada.

R. Michael Young. 1997. *Generating Descriptions of Complex Activities*. Ph.D. thesis, Intelligent Systems Program, University of Pittsburgh.

```
The visual combustion check is igniting which is abnormal in this startup mode (normal is
combusting)
Oil Nozzle always
  produces this abnormality when it fails.
Oil Supply Valve always
  produces this abnormality when it fails.
Oil pump always
  produces this abnormality when it fails.
Oil Filter always
  produces this abnormality when it fails.
System Control Module sometimes
  produces this abnormality when it fails.
Ignitor Assembly never
  produces this abnormality when it fails.
Burner Motor always
  produces this abnormality when it fails.
and, maybe others affect this test.
```

```
The visual combustion check indicator is igniting which is abnormal in startup mode.
Normal in this mode is combusting.

Within the Oil Burner
   These replaceable units always produce this abnormal indication when they fail:
     Oil Nozzle;
     Oil Supply Valve;
     Oil pump;
     Oil Filter;
     Burner Motor.

     The Ignitor assembly replaceable unit never produces this abnormal indication when
it fails.

Within the furnace system,
  The System Control Module replaceable unit replaceable unit sometimes produces this
abnormal indication when it fails.

Also, other parts may affect this indicator.
```

Figure 2: Original (top) and enhanced (bottom) answers to the same *Consult Indicator* query