 doi: 10.14355/srl.2015.04.003

Gap-Measure Tests with Applications to Data Integrity Verification

Truc Le, Jeffrey Uhlmann
Department of Computer Science
University of Missouri - Columbia
tdlxqb@mail.missouri.edu, uhlmannj@missouri.edu

***Abstract*—In this paper we propose and examine gap statisics for assessing uniform distribution hypotheses. We provide examples relevant to data integrity testing for which max-gap statistics provide greater sensitivity than chi-square (χ^2), thus allowing the new test to be used in place of or as a complement to χ^2 testing for purposes of distinguishing a larger class of deviations from uniformity. We establish that the proposed max-gap test has the same sequential and parallel computational complexity as χ^2 and thus is applicable for big data analytics and integrity verification.**

***Index Terms*—Hypothesis testing, distribution testing, chi-square testing, data integrity, big data, gap statistics, max gap, data integrity, Gonzalez algorithm**

I. Introduction

Distribution testing is a fundamental statistical problem that arises in a wide range of practical applications. At its core the problem is to assess whether a dataset that is assumed to comprise samples from a known probability distribution is in fact consistent with that assumption. For example, if the end state of a computer simulation of a physical system is a set of points with a physics-prescribed distribution, then any detected deviation from that expected distribution could undermine confidence in the results obtained and possibly in the integrity of the simulation system itself. Data integrity verification is a related application for distribution testing in which the objective is to detect evidence of tampering, e.g., human-altered data. For example, many sources of numerical data conform to the Benford-Newcomb first-digit distribution, and identified deviations from this distribution have been used to uncover acts of scientific misconduct and accounting fraud.

There is of course no way to make an unequivocal binary assessment of whether a dataset of samples conform to a given distribution assumption, but it is possible to devise statistical tests which can assign a rigorous likelihood estimate to the hypothesis that the dataset does (or does not) represent samples from the assumed distribution. In this paper we briefly review the most widely-used method for distribution testing, the chi-square (χ^2) test, and then develop alternative tests based on the statistics of gap-widths between data items of consecutive rank. Our principal contribution is a max-gap test which is shown to provide superior sensitivity to specific deviations from a uniform distribution that are relevant to data integrity testing. We show that this test can be evaluated with the same optimal computational complexity (serial and parallel) as the conventional χ^2 test and is therefore suitable for extremely large-scale datasets.

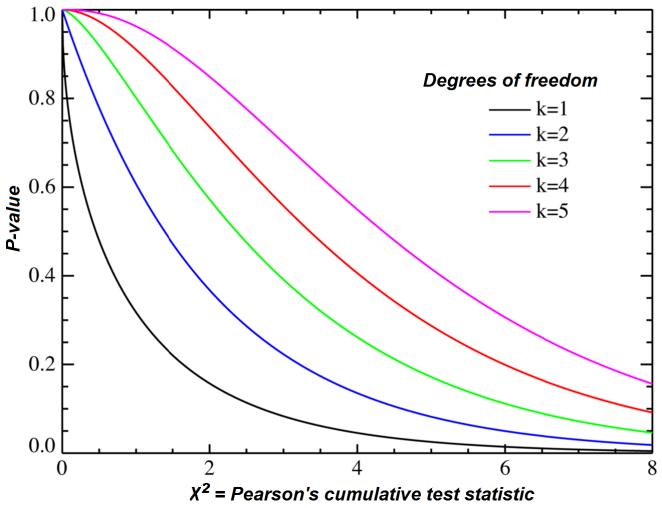

Fig. 1. Complement of the cumulative distribution function of the χ^2 distribution, showing χ^2 on the x-axis and p-value on the y-axis [1].

II. Chi-square Test

The (χ^2) test is a statistical measure that can be applied to a discrete dataset to assess the hypothesis that its elements were sampled from a particular distribution. More specifically, it is a histogram-based method to measure the goodness-of-fit between the observed frequency distribution and the expected (theoretical) frequency distribution. The general procedure of the test includes the following steps:

1) Calculate the chi-square statistics χ^2, which resembles a normalized sum of squared differences (deviations) between observed and expected frequencies.
2) Determine the degrees of freedom, df, of that statistic, which is essentially the number of frequencies reduced by the number of parameters of the fitted distribution.
3) Compare χ^2 to the critical value computed from the chi-square distribution with df degrees of freedom, which in many cases gives a good approximation of the distribution of χ^2.

An example of the complement of the cumulative distribution function of the χ^2 distribution is shown in Fig. 1 with different degrees-of-freedom values. For our uniformity test, the procedure is re-written as follows:

1) Given N observations, construct an N-bin histogram. Let b_i be the bin count for the i^{th} bin ($i = 1, \ldots, N$), which is the observed frequency distribution. As we are testing for uniformity, the expected frequency distribution $e_i = 1, \forall i = 1, \ldots, N$.
2) Compute the chi-square test statistics by

$$\chi^2 = \sum_{i=1}^{N} \frac{(b_i - e_i)^2}{e_i} = \sum_{i=1}^{N} (b_i - 1)^2 \tag{1}$$

3) The degrees of freedom of the statistics is $df = N-1$ for this case because if we know the counts for $N-1$ bins, the count for the remaining bin is uniquely determined.
4) Compute the complement of the cumulative distribution function of the χ^2 distribution with χ^2 and df obtained from the previous steps. Compare this value with the significance level α for the test result.

Despite being the de facto standard for assessing dataset consistency with respect to a given distribution assumption, the χ^2 test is not optimally sensitive to the types of deviation from uniformity that arise in many data integrity applications. One example involves narrow-band missing data resulting from a corrupted sensor or measurement process. Another example involves data that is generated from a non-random process and exhibits a higher degree of data regularity than is expected for a uniform distribution [2]. Datasets of the latter kind are typical of artificial and human-generated data, e.g., as in a forged dataset that has been tailored to include deviations that qualitatively resemble (to humans) uniform random deviates. In the following section we demonstrate the advantage of the proposed max-gap test over χ^2 for narrow-band and high-regularity deviations from uniformity.

III. Max-gap Test

The maximum gap, or max-gap, for a dataset of real values is defined as the maximum difference between elements of consecutive rank, which can be determined from a sorted ordering of the dataset. The distribution of spacings between consecutive-rank items in a dataset has been examined in the literature [3], [4], [5], [6], and we summarize here some of the results relevant to gap analysis. Assume we are given $N-1$ observations on the open unit interval $(0, 1)$ which divide the interval into N intervals whose lengths in ascending order are denoted by $S_{(1)} < S_{(2)} < \cdots < S_{(N)}$. For uniformity testing we are interested in $S_{(N)}$, as it is the *max-gap* of the observations. The exact distribution of $S_{(N)}$ is [6]:

$$P(S_{(N)} \leq x) = \sum_{\nu=0}^{N} (-1)^{\nu} \binom{n}{\nu} (1 - nx)_{+}^{N-1} \tag{2}$$

where $a_+ = max(a, 0)$.

From the p-value of the max-gap $S_{(N)}$, denoted by p, we can perform a max-gap test for uniformity by checking the condition $p \geq \alpha$ for *one-sided* test or $1 - \frac{\alpha}{2} \geq p \geq \frac{\alpha}{2}$ for *two-sided* test where α is the significance level. When N is large we may replace computation of the exact cumulative distribution of the max-gap in (2) with the following asymptotic result [6]:

$$P\left(S_{(N)} \leq x\right) \overset{N\to\infty}{=} e^{-e^{\ln N - Nx}}, \tag{3}$$

and the expected value of $S_{(N)}$ is

$$E\left(S_{(N)}\right) \overset{N\to\infty}{=} \frac{\gamma + \ln N}{N} \tag{4}$$

where γ is the Euler's constant.

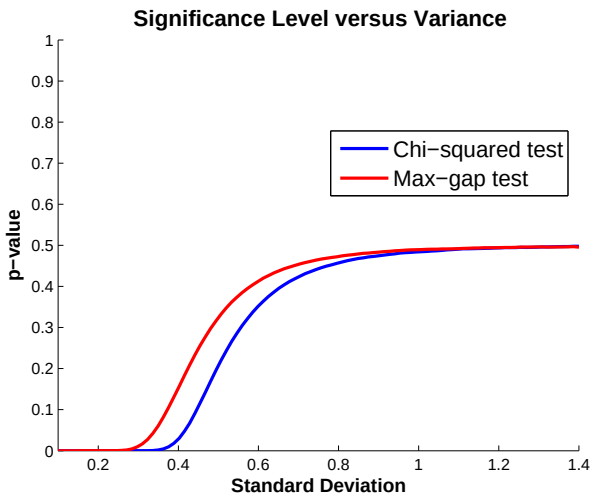

Fig. 2. p-value of the χ^2 test and the max-gap test of a normal distribution sampled within a fixed interval. When the standard deviation (σ) is small, both tests easily identify the data's non-uniformity. As σ increases, the data distribution approaches uniformity within the sample interval and hence the p-values converge to 0.5. This is an example in which the χ^2 test is more sensitive than the max-gap test.

An efficient max-gap test for uniformity can then be formalized as follows: Given $N-1$ observations x_i, and a significance level α, we first compute the max-gap $S_{(N)}$ of $\{0, 1\} \cup \{x_i\}$. Next, the p-value of the statistics is calculated by

$$p = 1 - e^{-e^{\ln N - NS_{(N)}}} \tag{5}$$

If the p-value satisfies $p \geq \alpha$ for the one-sided test, or $1 - \frac{\alpha}{2} \geq p \geq \frac{\alpha}{2}$ for the two-sided test, the observations are deemed to pass the test. Otherwise the set of observations is assessed to be inconsistent with a uniform-sampling hypothesis and fails the test.

In the next section, we present some experiments to compare the sensitivities between the χ^2 test and the max-gap test with emphasis on narrow-band missing data and highly regularized data mentioned in section I.

IV. Experiments

We conducted four experiments involving datasets of $N = 10,000$ samples, with the result for each experiment obtained as an average of one million independent tests. Sensitivity is assessed by comparing the respective p-values for the one-sided forms of the two tests, where smaller values indicate greater sensitivity. The first experiment was performed using a dataset of samples from a true uniform distribution. As

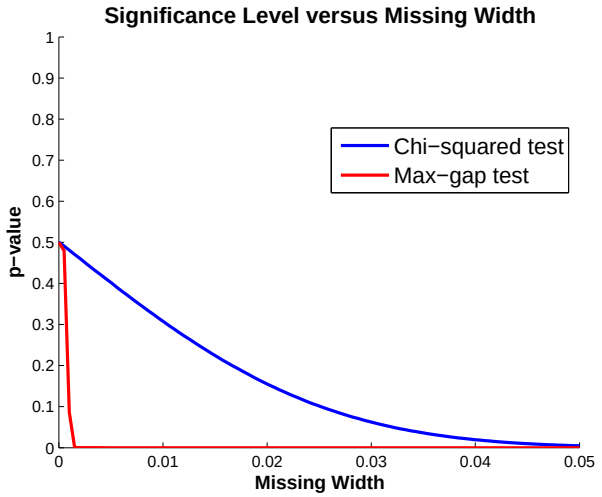

Fig. 3. p-value of the χ^2 test and the max-gap test for narrow-band missing data. In this case the max-gap test exhibits much greater sensitivity.

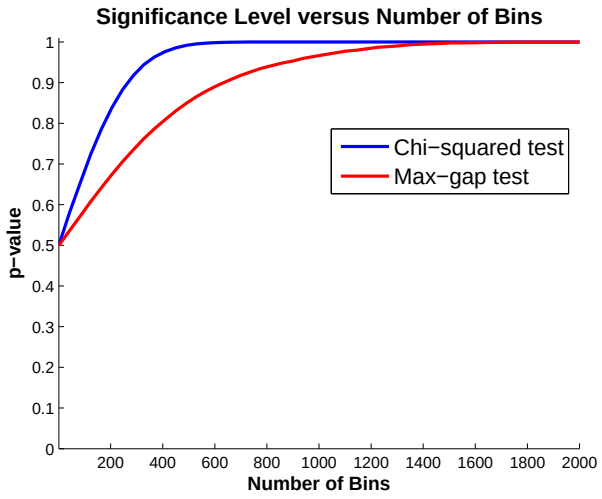

Fig. 4. p-value of the χ^2 test and the max-gap test for high-regularity data. Regularity for a dataset of size N is parameterized by a number of bins k with N/k uniform samples within each of k equal-width bins. (Thus $k = 1$ generates a uniform distribution and increasing k approaches regular spacing.) The max-gap test again demonstrates greater sensitivity.

expected, the dataset passed both tests for uniformity with $p = 0.5$.

The second experiment examined sensitivity to the difference between a uniform distribution and a normal distribution with standard deviation σ sampled within a fixed interval $(0, 1)$. The distinctive shape of the normal distribution is realized within the interval when σ is small but flattens with increasing values and approaches uniformity. Both tests are equally sensitive for small σ, and both approach $p = 0.5$ for large σ, but the χ^2 test exhibits higher sensitivity for intermediate values (see Fig. 2). The latter is not surprising because the χ^2 test is ideally suited for discriminating deviations from normality.

The third experiment examined sensitivity of the two tests to a uniform distribution with a narrow-band exclusion (Fig. 3). As suggested by (4), the max-gap test provides much greater sensitivity in this case. In fact, the χ^2 test provides only modest sensitivity as the exclusion width approaches one percent of the distribution window.

The fourth experiment examined sensitivity to regularity in sample spacing. In this case N/k samples were distributed uniformly within each of k equal-width subdivisions of the distribution interval. Thus $k = 1$ represents a uniform sampling over the entire interval and produces a uniform distribution. As k increases to N the spacing between samples becomes increasingly regular. Although uniform and high-regularity distributions are difficult for humans to distinguish visually, Fig. 4 shows that the max-gap test provides high sensitivity – significantly greater than χ^2.

V. Practical Considerations

The one-sided variants of the max-gap and χ^2 tests were used because they provide a practical balance between high sensitivity and low false alarm rates, but the one-sided or two-sided of either test may provide the optimal trade-off for the needs of a particular given application. In some applications the optimal trade-off might be obtained from a *min-gap*, $S_{(1)}$, test. The min-gap approximated distribution is given by [6]

$$P\left(S_{(1)} \le x\right) \overset{N\to\infty}{=} e^{-e^{\ln N - Nx}} \sum_{\nu=0}^{N-1} \frac{\left(e^{\ln N - Nx}\right)^{\nu}}{\nu!}, \quad (6)$$

and its expected value is [6]

$$E\left(S_{(1)}\right) \overset{N\to\infty}{=} \frac{\gamma + \ln N + \sum_{i=1}^{N-1} \frac{1}{i}}{N} \quad (7)$$

where γ is the Euler's constant.

The min-gap test can be defined and performed analogously to the max-gap test, and additional experiments show that it provides competitive sensitivity in the case of high data regularity but impractically low sensitivity to narrow-band missing data. For purposes of data integrity verification the principal value of the min-gap test may be for identifying spuriously replicated data items.

In terms of computational complexity, both χ^2 and max-gap tests can be evaluated in optimal $O(N)$ time and $O(N)$ space. This complexity is achieved for max-gap by the use of the Gonzalez algorithm [7], [8] to determine the max-gap in linear time without sorting. The Gonzalez algorithm performs a special type of binning and it guarantees an empty bin which, by the pigeonhole principle, makes sure that the max-gap data items will be found as the maximum and minimum values, respectively, in consecutive non-empty bins. This algorithm allows the max-gap test to be as efficient and parallelizable[1] as the χ^2 test. The min-gap test can be computed in randomized optimal expected $O(N)$ time and space using Rabin's algorithm[9], [10] to find the closest pair of data items, though the computational overhead is somewhat larger than that of χ^2 and max-gap. Unlike the Gonzalez algorithm for max-gap, Rabin's algorithm generalizes efficiently to higher dimensions.

[1]The max-gap and χ^2 tests are highly amenable to parallelization to achieve $O(N/P)$ time complexity on P processors.

VI. Discussion and Future Work

We have defined and developed a max-gap test for distinguishing deviations from uniformity in a 1D dataset of size N. By using Gonzalez's algorithm we have shown that this test can be performed with commensurate efficiency, both serial and in parallel, with the conventional χ^2 test. Our experiments demonstrate that the max-gap test provides improved sensitivity in two particular applications of relevance to data integrity verification. More generally, the proposed max-gap and min-gap tests are of potential value as alternatives or to complement the use of χ^2 for distribution testing and discrimination.

Potential future work could consider tests which jointly combine gap and χ^2 statistics into a more sophisticated single test which allows greater flexibility to optimize the sensitivity and false alarm trade-off for problems of high practical interest, e.g., big data analytics and integrity verification. On the algorithmic side, we have pointed out that the Gonzalez algorithm does not generalize to higher dimensions; however, relatively efficient subquadratic algorithms do exist for solving the largest empty circle and largest empty rectangle problems in two dimensions such as the algorithms in [11], [12]. In $d > 2$ dimensions it may be possible to devise gap-related statistical tests based on results from efficient algorithms for identifying approximations to the largest empty d-sphere or d-rectangle, but this is purely speculative. In higher dimensions it may be better to abandon gap-type statistics and focus on statistics gleaned from efficiently-computable k-d and orthant (quad, octant, etc.) tree decompositions of point sets.

If computational efficiency is less of a concern, a perhaps more fruitful direction for highly-sensitive distribution testing in high dimensions is to examine the length of the Euclidean minimum spanning tree (EMST) for a dataset. The expected length of the EMST of uniformly-distributed points can be determined using analysis similar to what has been described in this paper for estimating the expected values for the max and min gaps in 1D, and we conjecture that EMST length is likely to be more sensitive to many practically important types of deviations from uniformity than the conventional χ^2 test. Such an EMST test would be computationally expensive (though subquadratic), but this cost could be justified in applications for which subtle deviations are critically important, e.g., high-fidelity physics simulations.

References

[1] M. Haggstrom. (2010) Complement of chi-square cumulative distribution. [Online]. Available: http://en.wikipedia.org/wiki/File:Chi-square_distributionCDF-English.png

[2] J. H. Pitt and H. Z. Hill, "Statistical detection of potentially fabricated data: A case study," November 2013.

[3] D. A. Darling, "On a class of problems related to the random division of an interval," *The Annals of Mathematical Statistics*, vol. 24, no. 2, pp. 239–253, June 1953.

[4] R. Pyke, "Spacings," *Journal of the Royal Statistical Society*, pp. 395–449, 1965.

[5] ——, "Spacings revisited," pp. 417–427, 1972.

[6] L. Holst, "On the lengths of the pieces of a stick broken at random," *Journal of Applied Probability*, pp. 623–634, September 1980.

[7] T. Gonzalez, "Algorithms on sets and related problems," Department of Computer Science, University of Oklahoma, Norman, OK, Tech. Rep., 1975.

[8] ——, "Clustering to minimize the maximum intercluster distance," *Theoretical Computer Science*, vol. 38, pp. 293–306, 1985.

[9] M. Golin, R. Raman, C. Schwarz, and M. Smid, "Simple randomized algorithms for closest pair problems," *Nordic Journal of Computing*, vol. 2, no. 1, pp. 3–27, March 1995.

[10] R. Lipton, "Rabin flips a coin," in *The P=NP Question and Gdels Lost Letter*. Springer US, 2010, pp. 77–80.

[11] B. Chazelle, R. L. Drysdale, and D. T. Lee, "Computing the largest empty rectangle," *SIAM Journal of Computing*, vol. 15, no. 1, pp. 300–315, February 1986.

[12] A. Naamad, D. Lee, and W. Hsu, "On the maximum empty rectangle problem," *Discrete Applied Mathematics*, vol. 8, no. 3, pp. 267–277, 1984.