VISUAL ASSESSMENT VS. STATISTICAL GOODNESS OF FIT TESTS FOR IDENTIFYING PARENT POPULATION

Mari Berry, Brian Peacock*, Bobbie Foote, Lawrence Leemis
School of Industrial Engineering, The University of Oklahoma
Norman, OK 73019

* Advanced Vehicle Engineering
CPC Group, General Motors Corporation
Pontiac, MI 48058-1493

ABSTRACT

Statistical tests are used to identify the parent distribution corresponding to a data set. A human observer looking at a histogram can also identify a probability distribution that models the parent distribution. The accuracy of a human observer was compared to the chi-square test for discrete data and the Kolmogorov-Smirnov and chi-square tests for continuous data. The human observer proved more accurate in identifying continuous distributions and the chi-square test proved to be superior in identifying discrete distributions. The effect of sample size and number of intervals in the histogram was included in the experimental design.

BACKGROUND

Some optimization problems are very complex analytically. Numerical methods, for instance, still have a low probability of success in mixed-integer or integer optimization problems and combinatorial problems such as routing and scheduling. Gonen, Turen, and Foote (1982) reported that planners of electric power distribution systems (locating substations, routing power to customers, choosing line gauge) found feasible solutions from using geometrical maps of the problem when a commercial package failed to find a feasible solution in one hour on an IBM 370/158 computer. Hurst and Kohner (1981) reported enhanced success in their survey of human aided computation on routing problems when human pattern recognition was used. Brady and Rosenthal (1980), Brady, Rosenthal and Young (1983), and Elzinga and Hearn (1972) reported that nonlinear optimization problems with constraints can be solved by human pattern recognition based on geometric representations of the problem. Some of the problems solved easily by humans have not been solved by analytic methods.

There is a small but growing body of literature on the use of human pattern recognition capabilities and the use of this ability in man-machine systems to solve decision and optimization problems. This paper is a result of fundamental research into the basic abilities of humans in pattern recognition and an application.

INTRODUCTION

A sample does not always accurately depict the characteristics of the associated population distribution due to sampling variability. This fact is compounded if a relatively small sample is grouped by an inappropriate histogram interval width. It is of interest to investigate how these factors affect the ability of humans to discriminate in comparison with widely used statistical tests.

Researchers have long been interested in the human being as an intuitive statistician. A survey by Pollard (1984) shows a lengthy history of these inquiries. Experiments have centered around investigations of subjects' abilities to estimate means (central tendencies) and proportions, put confidence intervals around these quantities, and make probability estimates concerning problems that are the equivalent of tests of hypotheses about means and proportions. It has been clear that these intuitive statistical judgments in most circumstances are not normatively sound (not related to sample size and variability). Further, bias is almost always present, such as a tendency to use multiples of ten as answers, and to over or under estimate probabilities given certain experimental conditions.

Evans and Pollard (1985) continued this line of research involving experimental problems that are essentially tests of a hypothesis on the mean of a population. Their experiments demonstrated that subjects improved the accuracy of their judgments when data was displayed graphically. The results are consistent with the obvious fact that the visual and mental skills of humans are geared toward recognition of geometrical patterns and shapes in 2 and 3 dimensions and not as intuitive computers of arithmetic.

Downloaded from pro.sagepub.com at UNIV OF OKLAHOMA LIBRARIES on January 20, 2016

These results lead one to investigate a different kind of hypothesis testing problem, which is the determination of the parent distribution from a set of data. Probability distributions not only have distinct algebraic functional forms, but there is a 1-1 correspondence between a functional form and graphical pattern as a cumulative probability distribution function. The test of the hypothesis H_0: $f = f_0$ can be presented as an arithmetic logic problem or as a visual pattern recognition problem. The question here is: how well can humans perform as intuitive statisticians versus arithmetic logic when humans can bring to bear their strengths in pattern recognition.

Due to their pattern recognition and interpolation/extrapolation capability, it was postulated that human observers could perform statistical analyses better than statistical tests for small data sets. The internal rules that make the final determination possible are not of direct interest here; only that the ability to determine distribution models exists to some particular degree in humans. The major purpose of this research is to determine the relative sensitivity of the three discriminators, chi-square test, Kolmogorov-Smirnov (K-S) test and human observation, to the effects of modulating histogram interval width and sample size.

EXPERIMENTAL DESIGN

Three probability mass functions and three probability density functions were investigated in this experiment (Hastings and Peacock, 1974). In order to reduce the probability that the subject could guess the correct answers, the discrete uniform and geometric distributions were added as distractors in the discrete case. In the continuous case, the normal distribution was used as a distractor. Two parameter sets, a and b, were created for each function as shown in Table 1.

The parameterizations of these distributions are given in the Hastings and Peacock (1974). These parameter combinations were chosen in such a way that the expected values of each distribution are very close, if not identical. Discrimination is easier, of course, if one is familiar with distribution characteristics.

Two major factors which influenced the outcome of this experiment were the sample size and the number of histogram intervals and interval width. Three sample sizes (12 as small, 30 as medium, and 86 as large) and two interval widths (1 and 2) were investigated in the discrete case. In the continuous case, 5, 10 and 15 histogram intervals were examined, with the same three sample sizes.

The inverse-transform method was used to generate random variables from the exponential and Weibull distributions. For the gamma distribution, a modified acceptance-rejection method was employed. A convolution algorithm was used to generate the negative binomial and binomial random variables. For the Poisson distribution, a method based on the relationship between the Poisson (λ) and exponential ($1/\lambda$), was exploited (Law and Kelton, 1982).

Students in the undergraduate "Applied Engineering Statistics" class participated in this experiment. The subjects had limited knowledge in probability and statistics; however, they all have had at least 2-3 years of engineering related courses and had studied the discrete and continuous distributions used in the experiment.

Since the subjects were likely to possess widely varying levels of motivation and understanding, the experiment was given on two occasions as a bonus examination. Prior to the examination, basic theories and applications of statistical distributions were covered in lectures.

At the beginning of the experiment, three trials were conducted for each run. Each histogram was presented for 10 seconds

Parent Distribution	SET a Parameters	E[X]	V[X]	SET b Parameters	E[X]	V[X]
Binomial	n=10; p=0.5	5	2.5	n=4; p=0.75	3	0.75
Negative Binomial	n=5; p=0.5	5	10	n=3; p=0.5	3	6
Poisson	λ=5	5	5	λ=3	3	3
Discrete Uniform *	a=0; b=10	5	10			
Geometric *				p=0.25	3	12
Exponential	b=3	3	9	b=1.5	1.5	2.25
Gamma	b=1; c=3	3	9	b=1;c=1.5	1.5	2.25
Weibull	b=1; c=0.42	2.99	73.94	b=1;c=0.6	1.5	6.96
Normal *	μ=3; σ=1	3	1	μ=1.5; σ=0.5	1.5	0.25

Table 1. Distributions and Parameter Values

* used as distractor

Downloaded from pro.sagepub.com at UNIV OF OKLAHOMA LIBRARIES on January 20, 2016

and students marked their answer on the answer sheet. Immediate feedback was provided at this point to insure that subjects fully understood the procedures and format of the experiment. The values of mean and standard deviation for each set were not provided to the subjects.

Two projectors and transparencies were used in both the experiments on discrete (runs Da and Db) and continuous (runs Ca and Cb) distributions. The experiment on discrete distributions was conducted on 28 March 86, and the experiment on continuous distributions on 11 April 86.

In run Da, the parameter set a of the discrete parent distributions was left on one projector while the associated histograms with varying intervals (21 in discrete and 24 in continuous) were displayed on the other projector for 10 seconds. After run Da, the parameter set b of parent distributions replaced the parameter set a. Corresponding histograms with different intervals (21 in discrete and 24 in continuous) were displayed on the other projector. The same procedure was used in the remaining runs. Subjects responded to each histogram by marking A, B, C or D on a multiple choice form corresponding to the four parent distributions. Runs Ca and Cb (continuous distributions) were administered in the same fashion. Examples of a histogram and the four parent distributions for parameter set Ca are shown in Figures 1 and 2 respectively.

The chi-square and Kolmogorov-Smirnov goodness of fit tests are used to assess how well a parametric model approximates a data set. Both of these tests are given in Law and Kelton (1982). The test statistic for the chi-square test is:

$$\chi^2 = \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j}$$

where np_j is the expected number of the data values which would fall in the j-th of k

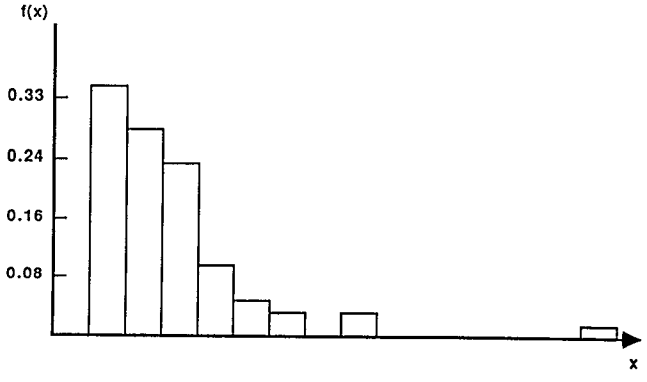

Figure 1. Histogram

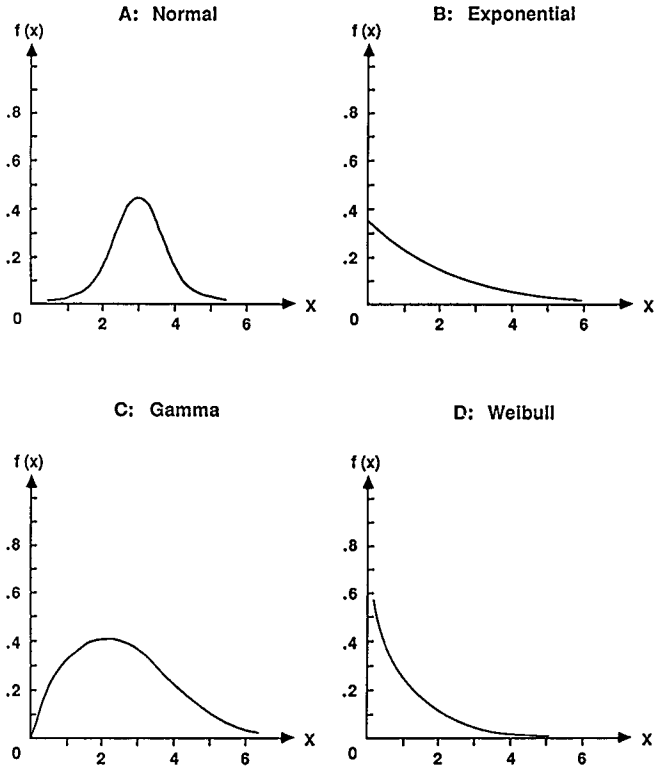

Figure 2. Continuous Distributions Parameter Set Ca

intervals if the hypothesized distribution were used, n is the size of the random sample and N_j is the actual number of data values in the j-th interval. The Kolmogorov-Smirnov test statistic is:

$$D_n = \max\{D_n^+, D_n^-\}$$

where

$$D_n^+ = \max_i \{\frac{i}{n} - \hat{F}(X_{(i)})\}$$

$$D_n^- = \max \{\hat{F}(X_{(i)}) - \frac{i-1}{n}\}$$

and $\hat{F}$ is the cumulative distribution function for the hypothesized model, n is the sample size, and $X_{(i)}$ is the i-th order statistic, i=1,2,...,n. For both tests, a smaller test statistic indicates a better fit of the model to the data. Other articles on goodness of fit tests are given by Lilliefors (1967, 1969), Massey (1951) and Williams (1950).

RESULTS

Human Observations

The student's score was recorded after each experiment. The performance curves for both experiments are negatively skewed. In addition, the two curves appear to correspond to a mixture of two populations, with approximately 30% with lower scores and 70% with higher scores. Figures 3, 4 and Table 2 summarize scores of human observers.

Downloaded from pro.sagepub.com at UNIV OF OKLAHOMA LIBRARIES on January 20, 2016

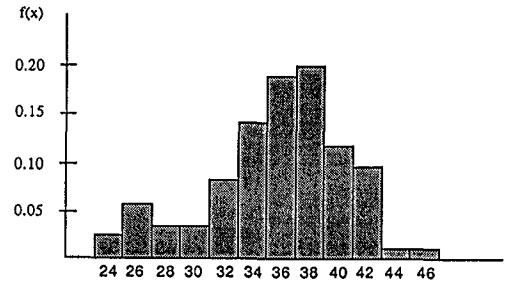

Figure 3. Histogram of Subject's Scores (discrete distributions)

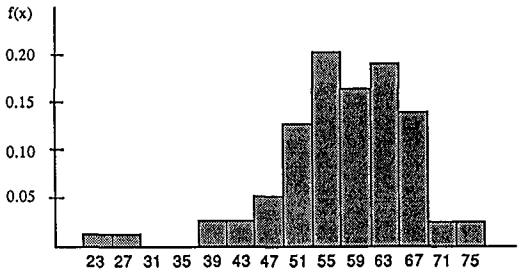

Figure 4. Histogram of Subject's Scores (continuous distributions)

Discrete Distributions	
Total number of participants:	86
Total number of problems:	82
Range	
Lowest Score:	23 (28.0%)
Highest Score:	47 (57.3%)
Mean Score:	35.26 (43.0%)
Standard deviation:	4.91 (6.0%)
Continuous Distributions	
Total number of participants:	79
Total number of problems:	96
Range	
Lowest Score:	21 (21.9%)
Highest Score:	74 (77.1%)
Mean Score:	56.82 (59.2%)
Standard deviation:	9.09 (9.5%)

Table 2. Experiment Summary

The mean score for continuous distributions was considerably higher than the mean score for discrete distributions, although a larger variance was observed in the continuous distributions experiment.

Statistical Tests vs. Human Observers

In order to compare the performance of statistical tests against the performance of human observers, one point was assigned to the statistical tests when the correct parent distribution was identified. The statistical tests always selected the chi-square or Kolmogorov-Smirnov statistic which was the smallest when choosing a distribution. Figures 5 and 6 illustrate the comparison between the subjects and the statistical tests for various histogram interval widths and sample sizes.

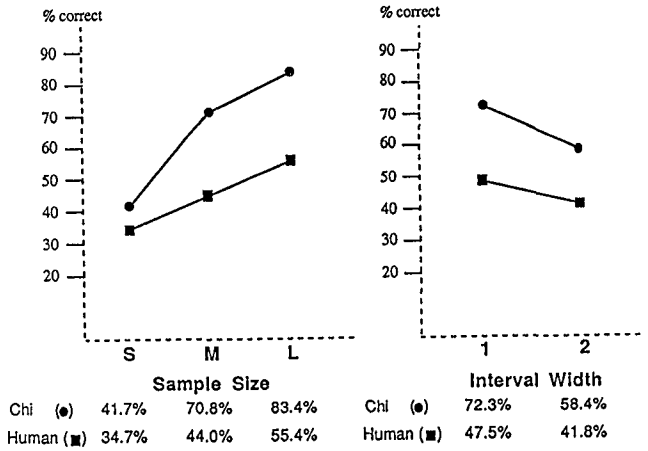

Figure 5. Discrete Distributions
Human Observers vs. Chi-square Test

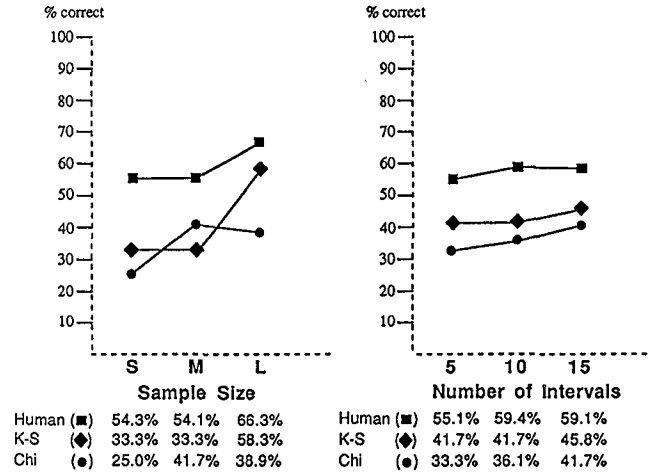

Figure 6. Continuous Distributions
Human Observers vs. Chi-square/K-S Tests

An exceptionally poor performance for the human observers in the discrete case was attributed to confusion between the distractor and negative binomial distribution. In the parameter set Da, only 10.8% of the subjects identified the negative binomial in the small sample size, while 81.4% of the subjects gave the answer as the discrete uniform. In the parameter set Db, only 28.5% of the subjects identified the negative binomial in the small sample size, while 60% of the subjects marked the answer as the geometric.

The leading cause of missed points by the chi-square test was attributed to the confusion between the negative binomial and Poisson (36.0%). The shapes of the probability mass functions for these two distributions are nearly identical.

A poor performance by the chi-square test in the continuous case was largely due to the fact that the chi-square test could not distinguish the exponential from the

Downloaded from pro.sagepub.com at UNIV OF OKLAHOMA LIBRARIES on January 20, 2016

gamma and vice versa. The loss of 7 points (7/24=29.2%) in the parameter set Ca and 8 points (8/24=33.3%) in the parameter set Cb were the result of these comparisons.

The test statistics for these two distributions were so close (not a single statistic could be rejected at α =0.01) that if a half point was assigned to the second choice of the chi-square test, a dramatic improvement of the score for the statistical tests was observed.

CONCLUSIONS

Because the subjects were not provided with the mean values or standard deviations with the histogram, they had to rely solely on their pattern recognition abilities. However, when a few extreme data values existed which represented less than say, 2% of the total data elements, they were not plotted on the graph. This modification to the histogram meant that the subjects would not be distracted by extreme distribution tail values, allowed for a shorter horizontal axis and did not preclude the uniform distribution when there was a value in the right hand tail of a distribution.

This experiment also indicated that an increase in the sample size did not benefit the chi-square test. This was due to the fact that the maximum number of intervals of 15 failed to take advantage of the larger sample size. As expected, the K-S test, on the other hand, showed a consistent improvement as the sample size increased. Despite the fact that some students did poorly, their average performance on the continuous case exceeded that of the statistical tests.

The K-S test, as expected, made good use of the available information and performed well. The chi-square test is hampered by small sample sizes and small numbers of intervals. The human observer, however, was able to intuitively smooth the data which was an advantage in the case of small samples, and is a surprisingly sound judge of the type of parent population a sample comes from given a proper visual display of the data.

A review of the results compared to the visual data admits a very straight forward explanation of the mental processes used by students to make decisions. The students can be hypothesized to carry a "template" of the shape of the probability density function mentally. This template is mentally superimposed on the histogram. If there is a good fit in the "center" a decision is made. Misfits at the tails are ignored. This research shows that the ability humans have in comparing geometric patterns as to congruence is very powerful.

REFERENCES

Brady, Stephen D. and Rosenthal, Richard E, 1980. "Interactive Computer Graphical Solutions of Constrained Minimax Location Problems", AIIE (IIE) Transactions, September, 241-248.

Brady, Stephen D., Rosenthal, Richard E., and Young, Donovan, 1983. IIE Transactions, 15, 3, September, 242-253.

Elzinga, Jack and Hearn, Donald W., 1972. "Geometrical Solutions for some Minimax Location Problems", Transportation Science, 6, 379-394.

Evans, J. and Pollard, P., 1985. "Intuitive Statistical Inferences about Normally Distributed Data". Acta Psychologica, 60, 57-71.

Gonen, Turen and Foote, B.L., 1982. "Mathematical Dynamic Optimization Model for Electrical Distribution System Planning", Electrical Power and Energy Systems, 4, 2, April, 129-136.

Hastings, N.A.J. and Peacock, J. B., 1974. Statistical Distributions. London: Butterworth & Company.

Hurst, E. Gerald and Kohner, Michael, 1981. "Optimization in Interactive Planning Systems", Department of Decision Sciences, The Wharton School, University of Pennsylvania, Philadelphia, PA, 19104, 81-05-05, presented

Law, Averill M. and Kelton, W. David, 1982. Simulation Modeling and Analysis. New York: McGraw-Hill.

Lilliefors, Hubert W., 1967. "On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown". Journal of American Statistical Association, 62, 399-402.

Lilliefors, Hubert W., 1969. "On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean Unknown". Journal of American Statistical Association, 64, 387-389.

Massey, Frank J., 1951. "The Kolmogorov-Smirnov Test for Goodness of Fit". Journal of American Statistical Association, 46, 68-78.

Pollard, P., 1984. "Intuitive Judgements of Proportions, Means and Variances: A Review". Current Psychological Research and Reviews, 3, 1, 5-18.

Williams, C. Arthur, 1950. "On the Choice of the Number and Width of Classes for the Chi-square Test of Goodness of Fit". Journal of American Statistical Association, 45, 77-86.

Downloaded from pro.sagepub.com at UNIV OF OKLAHOMA LIBRARIES on January 20, 2016