

AD-A242 134



CRM 91-82 / June 1991

2

Comparison of Three Procedures for Smoothing Score Distributions

D. R. Divgi

DTIC
S
C

CNA

CENTER FOR NAVAL ANALYSES

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268

91-14034



91 TO 24 14 3

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

Work conducted under contract N00014-91-C-0002.

This Research Memorandum represents the best opinion of CNA at the time of issue.
It does not necessarily represent the opinion of the Department of the Navy.

REPORT DOCUMENTATION PAGE

Form Approved
OPM No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Information and Regulatory Affairs, Office of Management and Budget, Washington, DC 20503.

1. AGENCY USE ONLY (Leave Blank)		2. REPORT DATE June 1991	3. REPORT TYPE AND DATES COVERED Final	
4. TITLE AND SUBTITLE Comparison of Three Procedures for Smoothing Score Distributions			5. FUNDING NUMBERS C - N00014-91-C-0002 PE - 65153M PR - C0031	
6. AUTHOR(S) D.R. Divgi				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Center for Naval Analyses 4401 Ford Avenue Alexandria, Virginia 22302-0268			8. PERFORMING ORGANIZATION REPORT NUMBER CRM 91-82	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Commanding General Marine Corps Combat Development Command (WF 13F) Studies and Analyses Branch Quantico, Virginia 22134			10. SPONSORING/MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES				
12a. DISTRIBUTION/AVAILABILITY STATEMENT Approved for Public Release; Distribution Unlimited			12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 200 words) Distribution of test scores need to be smoothed in equating and/or norming. Popular parametric smoothing procedures are based on beta-binomial and loglinear models. A new approach has been developed using polynomials of the beta-binomial cumulative distribution function. The same approach was also applied to extend the beta-binomial family to more than four parameters. These methods were compared using cross-validation in two examinee samples who took the Armed Services Vocational Aptitude Battery. Results show that the loglinear and extended beta-binomial families fit the data about equally well.				
14. SUBJECT TERMS ASVAB (armed services vocational aptitude batter), Comparison, Normalizing (statistics), Polynomials, Probability distribution functions, Scoring, Statistical distributions, Statistical processes			15. NUMBER OF PAGES 20	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT CPR	18. SECURITY CLASSIFICATION OF THIS PAGE CPR	19. SECURITY CLASSIFICATION OF ABSTRACT CPR	20. LIMITATION OF ABSTRACT SAR	

NSN 7540-01-280-5500

Standard Form 298, (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
299-01

CNA

CENTER FOR NAVAL ANALYSES

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268 • (703) 824-2000

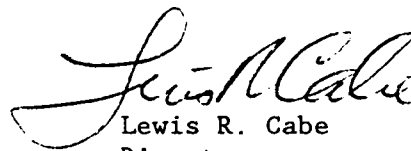
12 July 1991

MEMORANDUM FOR DISTRIBUTION LIST

Subj: Center for Naval Analyses Research Memorandum 91-82

Encl: (1) CNA Research Memorandum 91-82, *Comparison of Three Procedures for Smoothing Score Distributions*, by D.R. Divgi, Jun 1991

1. Enclosure (1) is forwarded as a matter of possible interest.
2. Distributions of test scores need to be smoothed in equating and/or norming. Popular parametric smoothing procedures are based on beta-binomial and loglinear models. A new approach has been developed using polynomials of the beta-binomial cumulative distribution function. The same approach was also applied to extend the beta-binomial family to more than four parameters. These methods were compared using cross-validation in two examinee samples who took the Armed Services Vocational Aptitude Battery. Results show that the loglinear and extended beta-binomial families fit the data about equally well.



Lewis R. Cabe
Director
Manpower and Training Program

Distribution List:
Reverse page



Approved for	
By	
Date	
Initials	
Signature	
Print Name	
Grade	
Branch	
Unit	
Address	
City	
State	
Zip	
Phone	
Fax	
E-mail	
Other	

A-1

Subj: Center for Naval Analyses Research Memorandum 91-82

Distribution List

SNDL

A1 DASN - MANPOWER
A1H ASSTSECNAV MRA
A2A CNR
A6 HQMC MPR & RA
Attn Code M
Attn Code MP
Attn Code MR
Attn Code MA
Attn Code MPP-54
FF38 USNA
Attn Nimitz Library
FF42 NAVPGSCOL
FF44 NAVWARCOL
Attn: E-111
FJA1 COMNAVMILPERSCOM
FJA13 NAVPERSRANDCEN
Attn Technical Director (Code 01)
Attn Dir, Testing Systems (Code 13)
Attn Technical Library
Attn Dir, Personnel Systems (Code 12)
Attn CAT/ASVAB PMO
Attn Manpower Systems (Code 11)
FJB1 COMNAVCRUITCOM
FT1 CNET
V12 MCCDC
Attn Training and Educations Center
Attn Warfighting Center (WF-13F)

OPNAV

OP-11B

OP-136

OTHER

Military Accession Policy Working Group (18 copies)

Defense Advisory Committee on Military Personnel Testing (8 copies)

Comparison of Three Procedures for Smoothing Score Distributions

D. R. Divgi

Operations and Support Division

CNA

CENTER FOR NAVAL ANALYSES

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268

ABSTRACT

Distributions of test scores need to be smoothed in equating and/or norming. Popular parametric smoothing procedures are based on beta-binomial and loglinear models. A new approach has been developed using polynomials of the beta-binomial cumulative distribution function. The same approach was also applied to extend the beta-binomial family to more than four parameters. These methods were compared using cross-validation in two examinee samples who took the Armed Services Vocational Aptitude Battery. Results show that the loglinear and extended beta-binomial families fit the data about equally well.

CONTENTS

	Page
Introduction	1
Data	2
Estimation of Parameters	2
Criterion for Evaluation	3
Results	4
Discussion	7
References	9

TABLES

	Page
1 Distribution of Family and of Number of Parameters with Smallest Chi-Square	3
2 Percentiles of Differences between Adjusted Chi-Squares, and of Smallest Adjusted Chi-Squares	6

INTRODUCTION

One major concern in norming and equating of test scores is the random error in score frequencies, especially at low scores where data are sparse. This error can be reduced by smoothing the frequencies. Two major approaches to parametric smoothing are in use at present. One is based on Lord's beta-binomial models [1]. The number of parameters in the beta distribution can be two, three, or four. The other approach is based on Holland and Thayer's use of the loglinear model [2].

In Lord's beta-binomial models, at any given proportion correct true score T , the number correct score X is assumed to have a binomial distribution with the probability parameter equal to T . The true score is assumed to have a beta distribution. In the two-parameter model, the true scores can vary from zero to one and the two parameters are determined by its mean and variance. In the three-parameter model, the third parameter is the smallest value of the true score in the examinee population. In the four-parameter model, the fourth parameter equals the largest true score in the population.

In the loglinear model, the logarithm of the probability of a score X is assumed to be a polynomial of X . If the polynomial contains p terms, the maximum likelihood estimates of the polynomial coefficients are such that the first p moments of X in the fitted distribution equal those in the sample [2]. The number of powers in the polynomial can be increased indefinitely until the fit of the model is considered satisfactory.

A third approach, developed by the author, uses polynomial families in which the fitted cumulative distribution function (cdf) is a constrained polynomial of some convenient parametric cdf, say of the negative hypergeometric (NH) distribution. Suppose F is the cdf of the NH distribution. Let G be the true cdf. We assume that G is a polynomial of F , subject to the end point constraints

$$G = 0 \text{ when } F = 0, \quad G = 1 \text{ when } F = 1.$$

The general form of a polynomial which obeys the above constraints is

$$G = F + F(1-F) [a_2 + a_3F + a_4F^2 \dots a_pF^{p-2}].$$

The $p+1$ parameters to be fitted are the mean and variance of the NH distribution plus the $p-1$ coefficients in the polynomial. The polynomial must be monotone increasing in $(0,1)$. In particular, its slopes at $F=0$ and $F=1$ must be nonnegative. Theoretical details and computational formulas will be presented in a forthcoming research memorandum.

The beta-binomial methods have two shortcomings. One is that, with three and four parameters for the beta, the density requires single and double summations. Moreover, each term in the sum contains quantities

that are computed by recursion. This makes it difficult to calculate derivatives analytically. More important, the number of free parameters cannot exceed four. In contrast, the loglinear and polynomial families can use as many parameters as are needed to obtain a satisfactory fit. Therefore, as described later, the beta-binomial family was extended into a polynomial family with the four-parameter beta-binomial as the base distribution F.

The objective of the study was to compare the beta-binomial, loglinear, and polynomial family approaches in terms of their fit to a large number of score distributions.

DATA

The Armed Services Vocational Aptitude Battery (ASVAB) is used in selection and classification of applicants to the military services. It yields scores on nine power tests and two speed tests. The number of items in the power tests ranges from 15 to 50. The speed tests, Numerical Operations and Coding Speed, contain 50 and 84 items, respectively.

Seven forms of the ASVAB have been administered to randomly equivalent samples from two populations. One population consists of recent military recruits. Very low scores are rare in this population because people with low aptitude have already been rejected. The sample size by form ranges from 2,501 to 2,774. The second population consists of applicants for enlistment in the military services, and hence provides a wider range of scores. The sample sizes vary from 13,010 to 14,963. Both data sets were provided by the Air Force Human Resources Laboratory.

ESTIMATION OF PARAMETERS

In the beta-binomial family, the binomial error model was used because, according to Lord ([1], p. 253), the simple binomial works as well as the compound binomial for fitting univariate distributions. Lord (p. 265) provides formulas for computing moments of proportion-correct true scores from those of observed scores. From the first four moments of true scores and the formulas for moments of the beta distribution ([3], p. 40), one calculates the beta parameters that make the theoretical values equal to the empirical ones. This is easy for the three-parameter model; for the four-parameter model, a search procedure is needed to find the fourth parameter. Sometimes it is not possible to match the empirical moments; the four-parameter model may be reduced to the special case of the three-parameter model, and sometimes the latter to the two-parameter, i.e., NH, distribution.

The recursive procedure of Lord and Novick ([4], eqn. 23.6.4) was used for the negative hypergeometric distribution. For three- and four-parameter beta-binomial families, probabilities of scores can be calculated using equation 52 in [1].

Parameters of the loglinear model were estimated by maximum likelihood. Holland and Thayer ([2], p. 5) provide expressions for first and second derivatives of the logarithm of the likelihood function. A minimizing routine written by the author was used to fit the parameters.

Maximum likelihood estimation can be used for the polynomial family as well, but minimum chi-square is more convenient. The Pearson chi-square uses expected frequencies in the denominator. If one uses observed frequencies in the denominator, calculations become much simpler [5]. For given mean and variance of the NH distribution, the chi-square is a quadratic function of the coefficients of the polynomial. Hence one can obtain them by solving linear equations. For the same reason, it is easy to impose the constraints of nonnegative slopes at end points. Nonlinear minimization is needed only over the two parameters of the NH distribution. Also, once scores are grouped so as to have a specified minimum frequency in each group, there is no danger of small denominators. Formulas for chi-square and its derivatives will be included in a forthcoming technical report.

The concept of a polynomial family of distributions is very flexible. Any parametric distribution can be used as the base, subject only to the feasibility of computer programming. Preliminary results showed that fit of the four-parameter beta-binomial model was excellent in many cases but unsatisfactory in others. Therefore, the beta-binomial family was extended by using polynomials as follows: The four-parameter beta-binomial distribution was used as the base. The third and fourth parameters of the beta distribution, which represent minimum and maximum true scores, were held fixed. The first two parameters, which describe a standard beta distribution, were treated as free parameters of the base distribution, and a polynomial family was fitted in the same way as with the NH distribution. Derivatives with respect to the beta parameters were computed from finite differences.

Seven distributions were fitted in each family, with the total number of free parameters varying from two to eight.

CRITERION FOR EVALUATION

It is possible to evaluate goodness of fit in the same sample as the one used to estimate model parameters. Pearson chi-square is convenient and popular, but it may favor the polynomial family, for which the parameters were estimated by minimum chi-square. Similarly, chi-square using likelihood ratio may favor the loglinear model, for which maximum likelihood estimation was used.

Therefore, cross-validation was used to evaluate the families. For each form of each test, the available data were split into two random samples with each examinee having a probability of .5 of going into either sample. For each smoothing procedure, parameters were estimated

in sample 1 and the estimates were used to compute score probabilities and hence log likelihood in sample 2. In any comparison, the method with higher likelihood was considered to have fitted better.

The log likelihood was subtracted from the maximum possible value, which is obtained by treating the observed proportion of each score as its true probability in the population. The difference, multiplied by two, is similar to a chi-square statistic. (This quantity does in fact have a chi-square distribution if the log likelihood is computed in sample 1, scores are grouped so that expected frequency in each group is large enough, and the model being fitted is correct.) Therefore, the statistic will be referred to as "chi-square" even though its true distribution is not strictly chi-square. Interpretation of its absolute value remains subjective, but one adjustment is helpful. The number of items varied from 15 to 84. All else being equal, a longer test yields a higher chi-square because it provides more degrees of freedom. Therefore, for a test with n items, each chi-square was divided by $3n/4$, which is the number of distinct scores above chance. Thinking of $3n/4$ as the degrees of freedom in a large sample, one might say that the fit of a model is satisfactory if this "adjusted chi-square" is two or less.

Since the divisor is the same for all methods and for all forms of any given test, this adjustment has no effect on the comparison of any two models. Also, because model fitting and evaluation are done in different samples, it is possible for the chi-square to increase when another free parameter is added, even within the same family. An additional parameter reduces the bias of a model, i.e., the difference between the true distribution and its best approximation using the model. However, the extra parameter also adds to the random error in the estimated distribution. It is possible for the increase in random error to exceed the reduction in systematic error, and hence for the chi-square in sample 2 to increase when the extra parameter is added. As will be seen below, this happened quite frequently.

RESULTS

Power and speed tests were analyzed separately. Only the power test results will be reported in detail. Multiple choice tests of speed are used much less than power tests. Also, the number of distinct forms was only 14 for speed tests compared to 63 for power tests. The two samples differ not only by a factor of five in size but also in the length of the lower tail. Therefore, their results will be presented separately.

With three families and two to eight parameters, 21 different models were used for each form of each test. The smallest of the 21 chi-squares was found, and the corresponding family and number of parameters were noted. Table 1 shows the distributions of these best-fitting models. The beta-binomial family is the best one overall in the recruit sample, and the loglinear family is the best overall in

the applicant sample. The number of parameters has a mode of 4 in the recruit sample, showing that the fit often got worse on adding another parameter.

Table 1. Distribution of family and of number of parameters with smallest chi-square

	Number of free parameters							Total
Family	2	3	4	5	6	7	8	
Recruit sample								
Loglinear	0	2	9	2	3	3	1	20
Beta-binomial	0	6	4	2	7	7	3	29
Polynomial of NH	1	2	4	1	2	0	4	14
Total	1	10	17	5	12	10	8	63
Applicant sample								
Loglinear	0	0	2	3	4	17	12	38
Beta-binomial	1	0	0	6	4	3	5	19
Polynomial of NH	0	0	1	1	2	0	2	6
Total	1	0	3	10	10	20	19	63

For more detailed comparisons, the loglinear family was treated as the reference. For each test, form, and number of parameters, adjusted chi-squares for the beta-binomial and the NH polynomial family were subtracted from that for loglinear. Thus, a negative result for beta-binomial indicated that the loglinear model fitted better than beta-binomial for that form using that number of parameters. For a given pair of families and a specific number of parameters, the available number of differences was 63, which is large enough for meaningful calculation of percentiles. Seven such percentiles are reported in table 2. "L - B, 3" means loglinear minus beta-binomial using three parameters, and so on. "P" stands for polynomial of the NH distribution. The medians are the easiest to interpret: for example, the value of 0.59 in the "L - B, 3" line for the recruit sample indicates that, with three parameters, the beta-binomial fitted better on the whole than the loglinear model did.

To evaluate the best fit that could be achieved using the available models, the lowest adjusted chi-square (over all three families and two to eight parameters) was identified for each test*form combination. Sixty-three of these minimum values were available in each sample. Their percentiles are also shown in table 2.

The results in table 2 yield the same conclusion as those in table 1: the beta-binomial model worked best in the recruit sample, loglinear in the applicant sample, and both were clearly superior to the polynomial of NH cdf. The smallest adjusted chi-square indicates that, if a value of two or less represents satisfactory fit, such fit was achieved in over 90 percent of the test*form combinations.

Table 2. Percentiles of differences between adjusted chi-squares, and of smallest adjusted chi-squares

Chi-square, no. of par.	95	90	Percentile				
			75	50	25	10	5
Recruit sample							
L - B, 3	2.43	2.43	1.24	0.59	0.03	-0.63	-6.18
L - B, 4	0.58	0.41	0.18	0.01	-0.09	-0.21	-0.63
L - B, 5	0.46	0.23	0.14	-0.06	-0.14	-0.28	-0.35
L - B, 6	0.72	0.50	0.15	0.03	-0.09	-0.21	-0.35
L - B, 7	0.61	0.49	0.23	0.06	-0.07	-0.16	-0.23
L - B, 8	0.92	0.58	0.30	0.14	-0.10	-0.30	-0.38
L - P, 3	1.52	1.36	0.62	0.12	-0.84	-1.91	-2.40
L - P, 4	0.57	0.40	0.19	-0.10	-0.44	-0.77	-0.93
L - P, 5	0.33	0.26	0.05	-0.10	-0.48	-0.65	-0.83
L - P, 6	0.89	0.47	0.17	-0.10	-0.21	-0.48	-0.63
L - P, 7	0.60	0.42	0.14	-0.05	-0.30	-0.50	-0.60
L - P, 8	0.81	0.58	0.23	-0.06	-0.19	-0.45	-0.94
Smallest	2.04	1.72	1.50	1.21	1.00	0.79	0.71
Applicant sample							
L - B, 4	0.90	0.69	0.30	-0.07	-1.08	-1.94	-2.54
L - B, 5	0.79	0.51	0.28	-0.32	-1.06	-1.83	-2.08
L - B, 6	0.58	0.36	0.20	-0.04	-0.77	-1.63	-2.15
L - B, 7	0.43	0.30	0.02	-0.21	-0.94	-1.74	-2.24
L - B, 8	0.40	0.34	0.15	-0.08	-0.62	-1.64	-2.01
L - P, 4	1.17	0.87	0.55	-0.27	-4.06	-7.82	-8.51
L - P, 5	0.70	0.59	0.09	-0.45	-3.59	-7.05	-8.15
L - P, 6	0.86	0.55	0.17	-0.22	-1.20	-2.83	-3.59
L - P, 7	0.62	0.44	-0.10	-0.37	-1.44	-2.77	-4.10
L - P, 8	0.42	0.31	-0.03	-0.27	-1.06	-1.63	-2.14
Smallest	2.53	2.28	1.96	1.55	1.30	1.01	0.89

Results for speed tests showed that the beta-binomial family never yielded the lowest chi-square in either sample. (Recall that, while any distribution may be used for fitting data, theoretically the beta-binomial models do not apply to speed tests.) Loglinear and NH polynomial families provided the best fit almost equally often. The smallest adjusted chi-square was less than two for all forms in the recruit sample. In the applicant sample, the smallest adjusted chi-square was less than two for about half the forms, and was never more than 2.6. This may be considered a satisfactory fit since the size of sample 2 was over 6,000.

DISCUSSION

The results, based on two samples from different populations and 63 forms of power tests, show that both the loglinear and the beta-binomial families should be tried while smoothing score distributions. The polynomial family using negative hypergeometric as the base does not appear promising.

Maximum likelihood estimation for the loglinear family is much simpler, both in theory and in computer programs, than constrained minimum chi-square for the polynomials that extend the beta-binomial family beyond four parameters. However, once the computer programs have been written, this aspect does not matter. Another difference is that asymptotic standard errors are available in the loglinear family, for parameter estimates and hence also for the fitted probabilities [2]. However, the practical value of these standard errors is open to question. Maximum likelihood estimates have their theoretical asymptotic properties when the model, including the number of free parameters, is fully specified. Standard asymptotic theory using Fisher's information matrix is not applicable when stepwise fitting is used to determine the number of parameters in the model.

In the asymptotic limit, minimum chi-square and maximum likelihood yield the same estimates. Hence theoretical standard errors (subject to the same criticisms as above) can be derived for the polynomial family with a negative hypergeometric base. They cannot be computed for the beta-binomial family, as implemented in this study, because the third and fourth parameters of the beta distribution are estimated using moments and then held fixed while other parameters are optimized by minimum chi-square.

In an operational testing program, in contrast to research, one cannot stop with a general conclusion that two families are worth trying. For each form of each test, one member of one family must be chosen to calculate the smoothed distribution. The rule for making this choice depends on various factors including the sample size, the relative importance of random versus systematic error in the testing program, and the effect of rounding (which makes small changes in the fitted distributions irrelevant even if they are statistically significant).

The cross-validated chi-square used in this study provides one way of choosing a model: For any form of any test, one should use the combination of model and number of parameters that yields the smallest chi-square in the validation sample. However, this approach shows what works best with *half* the available sample size. It is likely that a larger number of parameters will be optimal when the entire sample is used to fit the same model. (This effect can be seen in table 1: the number of parameters needed to minimize the adjusted chi-square tends to be larger in the applicant sample than in the recruit sample.) Because of the variety of considerations involved, a thorough discussion of possible decision rules is beyond the scope of this study.

REFERENCES

- [1] Frederic M. Lord. "A Strong True-Score Theory, with Applications." *Psychometrika* (Jun 1965): 239-270
- [2] Paul W. Holland and Dorothy T. Thayer. *Notes on the Use of Log-linear Models for Fitting Discrete Probability Distributions*. RR-87-31. Princeton, NJ: Educational Testing Service, 1987
- [3] Norman L. Johnson and Samuel Kotz. *Continuous Univariate Distributions-2*. New York: John Wiley and Sons, 1970
- [4] Frederic M. Lord and Melvin R. Novick. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison Wesley, 1968
- [5] Jerzy Neyman. "Contributions to the Theory of the Chi-Square Statistic." *Proceedings of the Berkeley Symposium in Mathematical Statistics and Probability* (1949): 239-273