Paper CC-01

Smoothing Scaled Score Distributions from a Standardized Test using PROC GENMOD

Jonathan Steinberg, Educational Testing Service, Princeton, NJ
Tim Moses, Educational Testing Service, Princeton, NJ

ABSTRACT

The estimation of a scale score distribution for an educational assessment is usually done with respect to a sample of the complete testing population. The scale score range is typically broad enough to allow for a normal distribution of scores to occur within the testing population. However, the frequency distribution of scores obtained from the sample can often appear skewed and can exhibit large sampling fluctuations, particularly if the sample is not completely representative of the entire population. This can have significant implications if test-takers' scores are reported along with the corresponding percentile ranks. A potential solution to this issue is to employ loglinear smoothing of the observed frequency counts. Through this process, important statistical features of the underlying observed data (e.g. mean, variance, and skewness), known as moments, can be preserved using only a small number of parameters while ensuring that the resulting frequency distribution and percentile ranks are smooth and free of sampling fluctuations (Moses & von Davier, 2004). SAS® can employ loglinear smoothing using PROC GENMOD where the observed frequency is a function of the score, its square, and its cube. This paper will demonstrate how PROC GENMOD can be used in an applied educational setting to smooth the univariate observed frequency distributions of the three distinct sub-scales for a test within different sub-groups. These transformations allow for the reporting of the resulting percentile ranks in a reliable and meaningful way for the groups of interest. This paper is intended for those with a good working knowledge of loglinear smoothing models and a moderate level of SAS programming experience.

INTRODUCTION

Standardized testing in education is often conducted on large populations of test-takers. The tests themselves are designed such that the statistical properties from the resulting scores are appropriate for detailed analysis and reporting. Test-takers are accustomed to receiving their score(s) as well as a measure of how they have performed compared to the total testing population, usually in the form of a percentile rank or norm. According to general statistical principles, the distribution of scores within a test-taking group should be such that the sample and population score distributions are approximately similar.

However, in educational research practice, particularly in small-scale pilot studies, these principles cannot always be met. The reasons for this may be inadequate sample sizes, population score distributions that vary from expectations, or clustering of scores within a particular range of that score distribution due to test specifications, a mixture of participating sub-populations, or score scaling practices. When score reporting is desired, alternative statistical methods for handling the score distribution must be employed so that normative information, such as percentile ranks, can be considered reliable and meaningful to present to test-takers. This paper will discuss the use of loglinear smoothing using PROC GENMOD (SAS, 2002) to accomplish this task.

BACKGROUND ON LOGLINEAR SMOOTHING

Loglinear models that relate the log of the expected test score frequencies to a linear function of the test scores are often used as a smoothing technique for test score distributions (Holland & Thayer, 1987; Kolen, 1991). According to Rosenbaum and Thayer (1987), smoothing also provides more stability to a set of observed frequency data when sample sizes are small. The primary advantage of this method is that the statistical properties of the observed score distribution can be more readily preserved using a relatively small number of parameters (Moses & von Davier, 2004). A complete technical explanation of the properties of loglinear models that make these desirable for this situation can be found in Moses and von Davier (2006).

DESCRIPTION OF AN EXAMPLE

Consider a test that has three distinct sections, each of which has a unique score distribution. It is desired to report the percentile ranks for the scores on each section for several sub-groups, for example three grade levels (1, 2, and

3) by two native language status groups (A and B), for a total of six groups. Approximately 2600 students were tested, so the average group size was approximately 435 students. The scores on each test ranged from 0 to 30. If scores are uniformly distributed at each of the 31 score points, only about 15 students on average would be at each score point, which is quite small. However, given the test specifications, the potential distribution of scores at the extreme score points (e.g. 0, 1, 2, 28, 29, 30) may be such that fewer students, or no students at all, obtain such scores. The example in this paper will go through the smoothing process for one of the three test sections to demonstrate how the smoothing affects the six groups in different ways, while preserving some of the key statistical properties, known as moments, of the score distributions, namely the mean (first moment), variance (second moment), and skewness (third moment).

Table 1 displays the observed score distributions for the six groups on the particular section of the test and Table 2 displays the descriptive statistics (mean, variance, and skewness) for each group.

Table 1: Observed Percentage of Test-Takers at Each Score Point

Score	Group 1A (n=452)	Group 1B (n=288)	Group 2A (n=547)	Group 2B (n=427)	Group 3A (n=597)	Group 3B (n=311)
0	0.00	0.00	0.00	0.00	0.00	0.00
1	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00	0.17	0.32
4	0.00	0.00	0.00	0.47	0.17	0.00
5	0.00	0.00	0.00	0.00	0.00	0.32
6	0.00	0.00	0.00	0.23	0.17	0.00
7	0.00	0.35	0.18	0.47	0.17	0.32
8	0.44	0.69	0.18	0.47	0.50	0.32
9	0.22	1.04	0.00	0.47	0.17	0.32
10	0.22	0.69	0.18	0.47	0.67	0.32
11	0.44	0.69	0.37	1.17	0.34	0.96
12	0.66	2.78	0.55	0.23	0.17	0.64
13	1.33	2.43	1.10	0.47	0.50	1.29
14	0.88	2.78	1.10	2.34	0.34	0.96
15	1.77	3.82	1.46	1.17	0.34	4.18
16	2.21	3.13	1.10	3.28	0.67	2.25
17	1.55	3.82	2.19	4.45	1.01	2.89
18	3.10	5.90	2.01	3.98	1.68	5.79
19	2.21	7.64	2.56	5.85	1.68	2.25
20	3.98	7.64	4.20	3.04	3.35	6.11
21	7.08	7.99	4.20	6.56	3.18	4.18
22	7.08	8.68	5.12	7.49	3.35	6.11
23	5.53	6.94	8.96	10.54	6.70	6.75
24	8.85	6.60	8.04	11.48	8.04	6.75
25	10.62	4.51	9.69	8.67	8.21	11.58
26	10.18	7.64	11.52	6.79	9.38	7.72
27	9.07	5.21	10.24	6.79	16.08	10.93
28	10.62	6.60	11.88	6.79	15.24	10.61
29	7.96	1.04	10.24	3.75	10.89	4.18
30	3.98	1.39	2.93	2.58	6.87	1.93

Table 2: Observed Descriptive Statistics by Group

Group	Mean	Variance	Skewness
1A	23.82743	19.61539	-0.97583
1B	20.97569	24.30951	-0.43630
2A	24.25960	17.81527	-1.11380
2B	22.36300	23.28342	-0.95299
3A	25.13735	18.67573	-1.88537
3B	22.80064	24.80529	-0.98925

It is worth noting that while the mean and variance can be computed using PROC FREQ with a WEIGHT statement, the skewness cannot be directly computed in SAS using the weighted frequency data. Therefore, the skewness was manually calculated according to Joanes and Gill (1998). As can be observed from the information in Tables 1 and 2, all groups have significant underrepresentation of test-takers at the lower end of the score distribution, which was expected given this was test where rights-only scoring was employed, and a score range of 0 to 30. The moments of the distribution directly reflect the inherent performance of the particular sample populations as the means are high within the score range and the skewness values are negative. This finding along with the sample size suggests how loglinear smoothing of the moments of the distribution may be useful in this case. The level of precision reported for the statistics in Table 2 is for comparative purposes to the statistics produced later from the loglinear smoothing.

CREATING THE DATA STEP

The DATA step is very straightforward, only requiring the score distribution and the frequency of scores at each score point to start. All 31 score points (0 to 30, inclusive) in this case need to be represented, so as shown in Figure 1, we insert frequencies of 0 for those particular values. The second step is to calculate the square and cube of the score points since these values will become part of the loglinear model.

Figure 1: Example DATA Step for Loglinear Smoothing of Observed Score Frequencies

```
data llin;
       input score freq;
       cards;
0      0
1      0
2      0
3      0
4      0
5      0
6      0
7      1
8      2
9      3
10     2
11     2
12     8
13     7
14     8
15     11
16     9
17     11
18     17
19     22
20     22
21     23
22     25
23     20
24     19
25     13
26     22
27     15
28     19
29     3
30     4
;

data llin_example;
       set llin;
       score2 = score**2;
       score3 = score**3;
run;
```

THE PROC GENMOD PROCEDURE

Finally, the loglinear model is implemented using PROC GENMOD where the set of observed frequencies is the dependent variable and the individual score point, its square, and its cube as independent variables. The link function is **log** and the distribution is **Poisson**, as is typical of models with count data. The output data set is specified and the variable "p" represents the smoothed frequency at each score point. The variable is formally named **p3** to represent that the smoothing is to be done based on the first three moments of the distribution. Figure 2 displays the PROC GENMOD syntax.

Figure 2: Complete Example of a PROC GENMOD Statement

```
proc genmod data=llin_example;
       output out=work.example p=p3;
       model freq=score score2 score3/link=log dist=p type3;
       title 'Loglinear Smoothing of the First 3 Moments';
run;
```

EVALUATING MODEL RESULTS

Table 3 displays the output data set from the PROC GENMOD syntax.

Table 3: Smoothed Percentage of Test-Takers at Each Score Point

Score	Group 1A (n=452)	Group 1B (n=288)	Group 2A (n=547)	Group 2B (n=427)	Group 3A (n=597)	Group 3B (n=311)
0	0.02	0.01	0.03	0.09	0.19	0.08
1	0.02	0.02	0.03	0.09	0.14	0.08
2	0.03	0.03	0.03	0.10	0.11	0.09
3	0.04	0.05	0.04	0.11	0.09	0.11
4	0.04	0.07	0.04	0.12	0.08	0.13
5	0.06	0.11	0.05	0.14	0.08	0.15
6	0.08	0.18	0.06	0.18	0.08	0.19
7	0.10	0.27	0.08	0.22	0.09	0.23
8	0.14	0.40	0.10	0.28	0.10	0.30
9	0.19	0.58	0.14	0.36	0.12	0.38
10	0.26	0.84	0.19	0.48	0.15	0.50
11	0.36	1.17	0.26	0.64	0.18	0.65
12	0.50	1.61	0.36	0.85	0.24	0.86
13	0.69	2.15	0.50	1.14	0.32	1.12
14	0.94	2.80	0.70	1.52	0.43	1.46
15	1.29	3.54	0.97	2.02	0.60	1.89
16	1.73	4.36	1.35	2.64	0.82	2.43
17	2.31	5.22	1.85	3.41	1.14	3.08
18	3.01	6.06	2.50	4.30	1.58	3.83
19	3.86	6.81	3.32	5.29	2.17	4.68
20	4.82	7.40	4.31	6.33	2.94	5.58
21	5.87	7.78	5.45	7.32	3.92	6.49
22	6.94	7.88	6.67	8.16	5.10	7.32
23	7.93	7.70	7.88	8.73	6.47	8.01
24	8.75	7.24	8.96	8.92	7.93	8.44
25	9.27	6.55	9.75	8.66	9.38	8.57
26	9.43	5.68	10.12	7.97	10.62	8.34
27	9.17	4.73	9.99	6.92	11.47	7.75
28	8.49	3.77	9.32	5.64	11.75	6.87
29	7.46	2.88	8.19	4.30	11.37	5.79
30	6.22	2.10	6.76	3.05	10.34	4.61

It is evident from examining Table 3 that the random fluctuations, or so-called “jaggedness” (Moses & von Davier, 2004) in the observed score distribution have been remediated through the loglinear smoothing model. This can be confirmed visually through the use of PROC GPLOT in displaying the observed and smoothed frequencies on a single graph. This can be accomplished through the following code shown in Figure 3, produced for Group 1A with the resultant plot displayed in Figure 4:

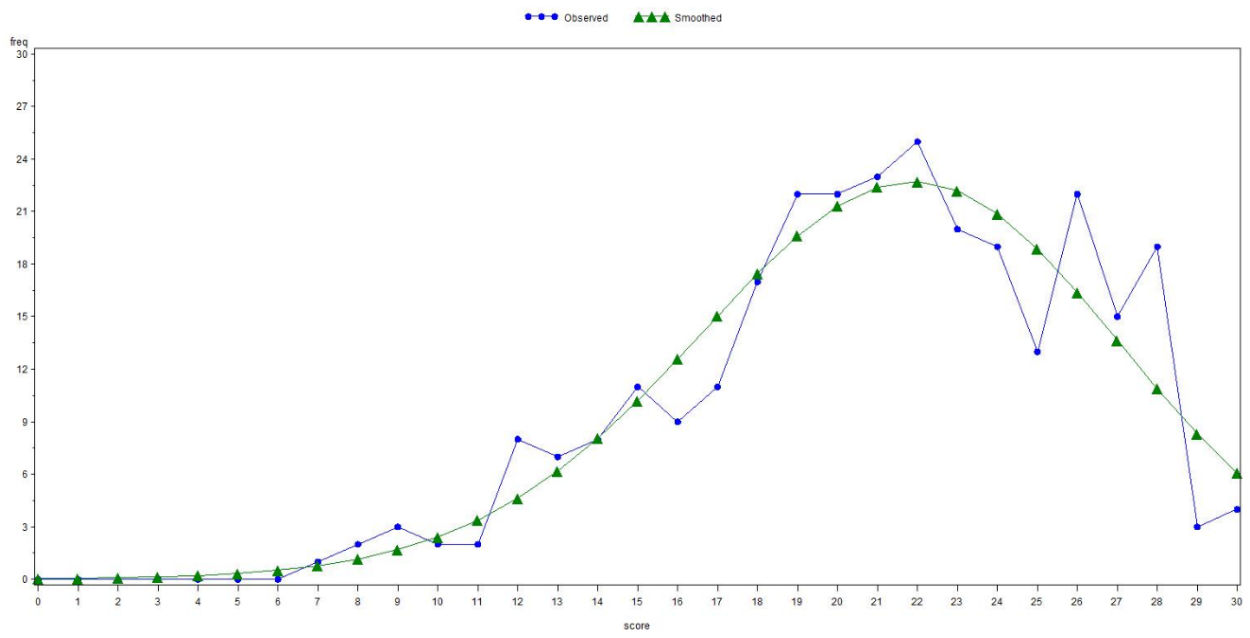
Figure 3: Syntax for PROC GPLOT to Display Observed and Smoothed Frequencies for Group 1A

```
/* Define titles and legend */
title1 'Observed and Smoothed Frequencies for Group 1A';
legend1 label=(height=1 position=top justify=center '') value=('Observed'  'Smoothed')
position=(top center);

 /* Define symbol characteristics */
symbol1 color=blue interpol=join value=dot height=1;
symbol2 color=green interpol=join font=marker value=C height=1;

 /* Generate plot of two variables */
ods rtf file='C:\Users\jsteinberg\Documents\gplot.doc';
proc gplot data=work.example;
   plot freq*score p3*score / overlay haxis=0 to 30 by 1
                    vaxis=0 to 30 by 3
                                legend=legend1
                                hminor=0
                                vminor=1
                    ;
run;
quit;
ods rtf close;
```

Figure 4: Graph Produced by PROC GPLOT of Observed and Smoothed Frequencies for Group 1A



Similar graphs can be produced for other groups. Additionally, one could combine the plots together for all six groups using PROC GREPLAY.

While formal tests of model fit are often employed, as discussed in Moses and von Davier (2004, 2006), the primary goal of the smoothing is to preserve the underlying statistical properties of the observed score distributions as shown above in Table 2. This then facilitates the production of meaningful and reliable percentile rank scores to test-takers. As shown in Table 4, the mean, variance, and skewness of the smoothed frequency distributions are virtually identical to the observed frequency distributions up to at least the third decimal place for the variance and skewness and the fifth decimal place for the mean.

Table 4: Smoothed Descriptive Statistics by Group

Group	Mean	Variance	Skewness
1A	23.82743	19.61543	-0.97584
1B	20.97569	24.30954	-0.43630
2A	24.25960	17.81528	-1.11381
2B	22.36300	23.28342	-0.95299
3A	25.13735	18.67573	-1.88537
3B	22.80064	24.80529	-0.98925

IMPLICATIONS FOR REPORTING

Since percentile ranks are desired in the reporting phase of the analysis, it is worth comparing for example, the cumulative percentages between the observed and smoothed frequency distributions. Table 5 indicates key percentiles and the corresponding score points by group.

Table 5: Key Percentiles and Corresponding Observed and Smoothed Score Points by Group

	25th Percentile		50th Percentile		75th Percentile	
Group	**Observed**	**Smoothed**	**Observed**	**Smoothed**	**Observed**	**Smoothed**
1A	22	22	25	25	28	28
1B	18	18	22	22	25	25
2A	23	22	26	25	28	28
2B	20	20	24	23	26	26
3A	24	24	27	27	29	29
3B	20	20	25	24	27	27

There is a high level of consistency between the observed and smoothed scores at the key percentiles presented in Table 5. Therefore, one can conclude that while the score reporting to test-takers may not have been adversely impacted if smoothing had not been employed, smoothing is particularly helpful for providing results that are more interpretable and usually more accurate for score ranges with sparse data (Moses & von Davier, 2004). However, the differences between the observed and smoothed scores would be greater at lower percentiles where observed data is sparse, namely that where the observed percentage of scores is zero (see score points 0 through 10 for Group 3B in Table 1). Yet, as noted, this example is not ideal in the world of testing since the 25^{th} and 50^{th} percentiles almost all occurred within the top third of the score scale (21-30).

CONCLUSION

This paper demonstrated the use of loglinear smoothing models in an educational testing setting when sample sizes were not as large as for most standardized tests. As noted, the statistical properties of the observed score distributions were preserved, thus the model fit the data well. For score reporting purposes, loglinear smoothing provides some additional necessary stability to the score points attached to key percentiles in the distribution. However, careful interpretation is needed when test-takers are provided normative information about their performance on tests that may be easier or harder than expected, or where norms are produced using small samples, as is the case here.

REFERENCES

Daniel, W. W. (1995). Biostatistics: a foundation for analysis in the health sciences (6th edition). New York: John Wiley & Sons, Inc.

Holland, P. W. & Thayer, D. T. (1987). *Notes on the use of log-linear models for fitting discrete probability distributions*. Technical Report 87-79. Princeton, NJ: Educational Testing Service.

Joanes, D. N. & Gill, C. A. (1998). Comparing measures of sample skewness and kurtosis. *The Statistician, 47*(1), 183-189.

Kolen, M. J. (1991). Smoothing methods for estimating test score distributions. *Journal of Educational Measurement*, *28*(3), 257–282.

Moses, T., & von Davier, A. A. (2004). *Using PROC GENMOD for loglinear smoothing*. In W. Stinson & E. Westerlund (Co-chairs), NorthEast SAS Users Group, Inc. seventeenth annual conference proceedings. Retrieved March 10, 2011, from http://www.nesug.org/html/Proceedings/nesug04.pdf

Moses, T. P. & von Davier, A. A. (2006). *A SAS macro for loglinear smoothing: applications and implications*. ETS Research Report RR-06-05. Princeton, NJ: Educational Testing Service.

Rosenbaum, P. R., & Thayer, D. (1987). Smoothing the joint and marginal distributions of scored two-way contingency tables in test equating. *British Journal of Mathematical and Statistical Psychology*, *40*, 43–49.

SAS Institute. SAS/STAT Software (2002). The GENMOD procedure, Version 9. Cary, NC: SAS Institute.

SAS Institute. SAS/STAT Software (2009): SAS Version 9.2 Online Help and Documentation, Cary, NC: SAS Institute.

ACKNOWLEDGMENTS

The author would like to thank Michaela Arzt, Ted Blew, Jennifer Brown, Amy Cellini, Scott Davis, Dan Eignor, Kim Fryer, Bruce Kaplan, Ed Kulick, Shuhong Li, Rick Morgan, Cindy Nguyen, and Mikyung Wolf for their support and assistance in proofreading this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Jonathan Steinberg
Principal Research Data Analyst
Data Analysis and Research Technologies
Educational Testing Service
Rosedale Road – Mail Stop 20-T
Princeton, NJ 08541
(609) 734-5324
jsteinberg@ets.org

TRADEMARKS