Distribution of grades from aerobatic judges, goodness of fit to normal

Douglas Lovell

Draft, July, 2020

Abstract

In the sport of competition aerobatics, pilots fly combinations of loops, rolls, and other maneuvers in high performance airplanes, in front of judges observing from the ground. The judges quantify the quality of the loop and roll combinations using grades.

At the world level, a statistical scoring system treats judge grades as if they are normally distributed, in an attempt to detect and eliminate biases among the judges and produce scores that rank the pilots.

This paper reports goodness of fit to the normal distribution of grades given by judges in 2710 flight programs during six years from 2014 to 2019, and reports the results. The measurements show poor fit to the grades of the judges. The poor fit for the normal distribution challenges the validity of the statistical scoring system and the results produced by it.

1 Aerobatic Competition

Aerobatics, as practiced by the World Air Sports Federation (FAI) Aerobatics Comission (CIVA) and by the International Aerobatic Club (IAC), is a judged sport. The pilot competitors perform in front of judges, who give grades to the pilots. In this way, it is similar to figure skating, gymnastics, and diving.

CIVA flies a handful of European and World contests each year. Results are posted online [3] in human readable format. The largest contests are the World Aerobatic Championship [6] and the World Advanced Aerobatic Championship, which attract about sixty pilots, mostly from Europe and the United States.

The IAC flies about forty regional contests and one national contest per year in the United States. Results from IAC contests are posted online [5] in human and machine readable formats.

1.1 Grading

A “flight program” consists of each of the competing pilots flying a sequence of figures in front of the judges. The sequences of figures are predetermined for the flight program by any one of several methods. Judges receive the sequences that each pilot will fly in order to evaluate each figure flown by the pilot and give it a grade.

The grade is a value from zero to ten in half point increments. A flight program produces a grade $g_{j,p,f}$ given by each judge j for each figure f flown by each pilot p.

2 Tested hypothesis

CIVA contests use a statistical system of converting judge grades to scores, which system is documented in the FAI sporting code as the ”Fair Play System” (FPS) [4].

The CIVA FPS represents subsets G of the grades $g_{j,p,f}$ as the normal distribution derived from their mean and variance, $Normal(\mu_G, \sigma_G^2)$. It does this in order to compare and identify grades not in keeping with those from other judges. The validity of the normal distribution representation of the grades is fundamental to the correctness of the method. If it is not valid, the method is fundamentally flawed.

This paper measures goodness of fit to the normal distribution of grades given by judges in 2710 flight programs from IAC contests during six years from 2014 to 2019, and reports the results.

For a particular judge j we have a set $G = g_{j,p_1,f_1}, g_{j,p_2,f_1}, ..., g_{j,p_n,f_1}, ..., g_{j,p_n,f_m}$ as grades given n different pilots flying the same m figures in a flight program. The grades are discrete, measured in half increments from zero to ten. We test the null hypothesis $H0$ that the grades G fit the normal distribution derived from their mean and variance, $Normal(\mu_G, \sigma_G^2)$.

3 Measurements

We apply a number of standard measures of goodness of fit– Chi-Square, Anderson-Darling, Shapiro-Wilk, Lilifords, and Cramer-von Mises.

The data is rounded. A judge does not have a continuous range between zero and ten, but must round to the nearest half grade. This means that it must either be treated as discrete, using clusters as with Chi-Square, or given random perturbations in order to make it continuous.

The full data set, together with R code for retrieval and analysis, is available from [1].

3.1 Clustering for Chi-Square

In order to apply the Chi-Square metric, we need to cluster grades. The optimal method for doing this is the one that has least impact on the mean and variance of the data.

The grade clusters must satisfy two constraints in order to get a goodness of fit metric from the Chi-Square method.

First, there need to be at least six clusters. The number of degrees of freedom for the Chi-Square metric is then, at minimum three– six minus one, minus two for the estimated mean and standard deviation. Three is the least number we believe will give a meaningful goodness of fit test.

Second, there need to be at least six instances within each cluster. This is the stricter criterion that leads to reduction in number of clusters.

Needing at minimum six clusters with six grades, in a uniform distribution we would need thirty-six grades. The distributions are not uniform, so we nearly double that number. We combine figure grades following the FPS ordering of figures in order to produce a minimum of sixty grades per group.

The constraints at play here are:

1. A minimum number of six grades in any cluster
2. Only adjacent clusters may be combined
3. Combine the least number
4. Have the weighted mean and variance close to the original

The clustering method described by Greenacre [7] and implemented by the greenclust [10] R package provides reorders the clusters. We need to join only adjacent clusters, to maintain the order.

The algorithm of [8] and other algorithms for k-shape ordered partitioning require a cumulative function that measures the quality of each partitioning. Here, we find that the mean and variance do not always decrease or increase when joining two clusters.

Where n is the number of grade values in the range of grades, we explore the 2^{n-1} combinations of joins using the heuristic of choosing to first try joining clusters with the smallest number of grades, and a bound of minimum count of joins. The partition is complete when no cluster contains less than six grades. If two solutions have the same number of joins, we select the one with minimum of $(\mu' - \mu) + (\sigma' - \sigma)$, in which $\mu' - \mu$ is the difference in the weighted mean, and $\sigma' - \sigma$ is the difference in the weighted variance.

In order to reduce the number of combinations to explore, we apply a pre-processing step that combines strings of zero sized clusters into one together with the lower numbered cluster neighboring the string. Strings of zero sized clusters appear frequently in the data. This pre-processing step avoids having the algorithm try all combinations of zero sized clusters.

3.2 Shapiro-Wilk

The grades given by the judges are rounded to increments of 0.5 from a continuum of performances by the pilots. In order to use the Shapiro-Wilk test, we must perturb the grades in order to make them unique.

Two methods to perturb the grades are

- add values selected from a random uniform distribution between -2.5 and 2.5.
- add values selected from a random normal distribution.

Applying a uniform distribution is the Smirnov transformation described in [9]. The method uniformly and reliably perturbs the data within the range and has the advantage of prior use in practice. A disadvantage is that the mean of the perturbed values can end-up anywhere between -2.5 and 2.5 added to the actual grade.

Using the random normal distribution ensures that the mean of the perturbed values will remain close to the actual grade value. The disadvantage is that we must choose a standard deviation for the random normal distribution such that the resulting perturbations are extremely rarely, with very low probability, outside of the range -2.5 to 2.5. A second, potential disadvantage is that this is a variation of the Smirnov transformation that, so far as we know, we are inventing for this study

We perturbed the grades using random values chosen from a normal distribution with mean equal to zero and standard deviation equal to $\sqrt{5/12n}$, approximating the distribution of Bates [2] between -2.5 and 2.5. We report the p-value for the Shapiro-Wilk normality test after perturbing the grades.

4 Judge Grade Distributions

We can now look at the goodness of fit measured results. We measured in two ways. First, we measured fit to all of the grades given by a judge for a particular flight. This results in more grades; however, those grades are given to a large variety of figures of all basic types and difficulties within the category graded.

Second, we measured fit to grades from FPS figure groups. The FPS groups grades according to the figure graded. When there are fewer than a dozen pilots, FPS combines grades given to figures of similar difficulty such that there are a minimum of a dozen grades in a figure group.

In both cases, we look only at flights for which there were more than five pilots and more than eleven grades given. We also look only at flights in which all of the pilots flew the same figures. To do so, we select for the flight format, taking only Known, Unknown, second unknown, first flights.

4.1 All figures

We now look at distributions of all of the grades given by a judge for a given flight– all pilots and all figures. The distribution of flight formats is as in Table 1.

Known	Unknown	Unknown II	Flight 1
1488	1160	27	19

Table 1: Distribution of flight formats for all-figures measurements

In all, we look at 2,511 powered flights and 183 glider flights. The flights breakdown by category is as in table 2.

Pilot count is the number of pilots who participated in a flight. Grade count is the number of pilots times the number of figures, equals the number of grades given by a judge for all of the figures and pilots in a flight. Their distributions in this data set are as in Table 3.

The correlations between the four measures are strong, with the exception of Lilifords to Shapioro-Wilk. Find their values in Table 4.

Table 5 shows summary results from the various goodness of fit tests. The numbers for measures found to be valid and invalid cover all of the results. The other two columns contain counts of p-value results from valid measures greater than and less than or equal to the traditional cutoff of 0.05.

The null hypothesis for these tests is that the distributions fit a normal distribution. Using the tradi-

primary	sportsman	intermediate	advanced	unlimited
177	1160	854	444	59

Table 2: Distribution of categories for all-figures measurements

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Pilot Count	6	7	8	9.277	11	26
Grade Count	35	65	81	94.55	110	359

Table 3: Distribution of pilot and grade counts for all-figures measurements

	sw	lf	ad	cvm
sw	1.000	0.432	0.792	0.616
lf	0.432	1.000	0.709	0.814
ad	0.793	0.709	1.000	0.937
cvm	0.616	0.814	0.937	1.000

Table 4: Correlation of all-figures measurements

	Invalid	Valid	> 0.05	<= 0.05
chiSq.t.p	1149	1545	934	611
chiSq.d.p	1149	1545	821	724
sw.p.value	0	2694	106	2588
lf.p.value	0	2694	36	2658
ad.p.value	0	2694	32	2662
cvm.p.value	0	2694	47	2647

Table 5: All figure GOF measure p-value summary

tional cutoff, the Chi-Squared test finds no support for fit to normal in roughly half of the cases. The other tests almost never find support for fit to normal.

We can go deeper by looking at the distributions of p-values and then looking at some specific examples within the various quantiles.

Table 6 shows the spread of p-values from the various tests. The numbers are from only those tests reported as valid. The minimum values from the data are always zero, and so omitted from the table.

Note that except for Chi-Squared, the mean value is greater than the third quantile value, demonstating that more than three-quarters of the values fall below the mean.

We can look at plots of the distributions to find more insights about the fits. One power study [11] finds that the Anderson-Darling test is among three most powerful. It has high correspondence with the other tests. In the following, we show two plots side-by-side for representative p-values from the Anderson-Darling measure. The left plot is a histogram of the judge grades with derived normal curve superimposed. The right plot is a standard Q-Q plot using the derived normal curve.

At the smallest p-value seen in Figure 1 we find a skew toward higher grades. The judge graded a majority of 9.0 with a mean grade of 8.5. The scarce 4.0 will always be seen as an outlier with respect to the normal curve. The upper tail of the normal curve exceeds the highest grade.

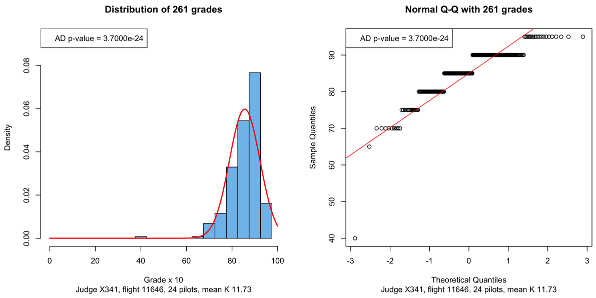

Figure 1: All figures example at minimum Anderson-Darling p-value

At the first quantile upper limit seen in Figure 2 we also find a skew toward the higher grades, but also a longer tail to the lower grades. The upper tail of the normal curve exceeds the highest grade.

At the second quantile upper limit seen in Figure

	1st Qu.	Median	Mean	3rd Qu.	Max.
chiSq.t.p	0.006	0.126	0.272	0.491	0.999
chiSq.d.p	0.002	0.067	0.216	0.358	0.993
sw.p.value	0.012e-04	0.896e-04	0.785e-02	0.209e-02	0.452
lf.p.value	0.0	9.200e-07	2.462e-03	1.121e-04	2.208e-01
ad.p.value	2.000e-08	7.280e-06	2.612e-03	3.129e-04	1.912e-01
cvm.p.value	4.400e-07	2.942e-05	3.327e-03	5.980e-04	2.781e-01

Table 6: All figure GOF measure p-value distributions

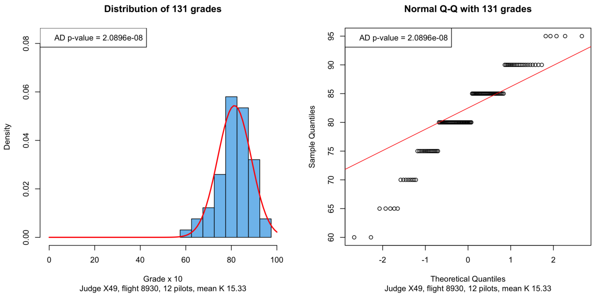

Figure 2: All figures example at first quantile Anderson-Darling p-value

3 we have a bimodal distribution in which the judge gives a large number of 7.0 and 9.0 with fewer grades given with values 7.5, 8.0, 8.5 and then a lesser number given with values 5.0, 6.0, 6.5, 9.5, and 10.0. That the upper tail of the normal curve exceeds the highest grade is beginning to look like a pattern.

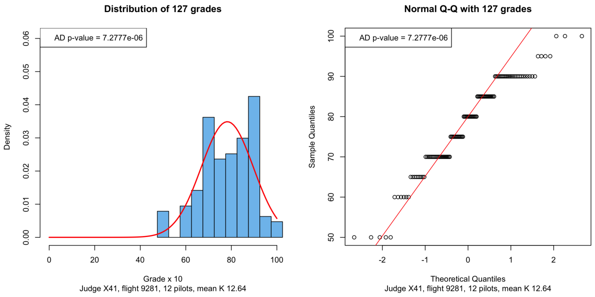

Figure 3: All figures example at second quantile Anderson-Darling p-value

At the third quantile upper limit seen in Figure 4 the picture has improved somewhat. There are a few too many grades with value 6.0 and 8.0, too few with grades 7.5 and 10.0. To fit the curve, the judge should have given a dash of grades with value 10.5, which is not a valid grade.

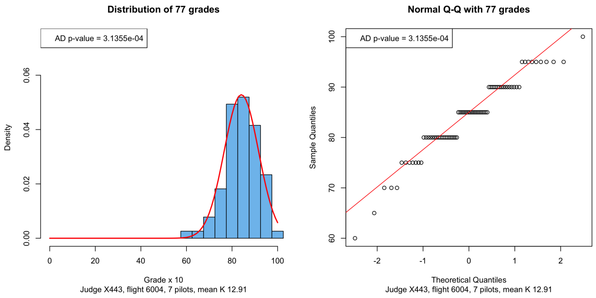

Figure 4: All figures example at third quantile Anderson-Darling p-value

At the maximum value seen in Figure 5 we have a judge who spreads their grades out more than in the other examples. The spread does show a strong, though excessive peak at the mean grade of 7.0. There are a few too many grades with value 4.0, 5.0, 8.0, and 9.5. There are too few with values 5.5, 7.5, and 8.5. This judge almost manages to get the entire normal curve within the range of grades; however, the upper tail still exceeds the maximum grade of 10.0.

4.2 FPS groups

The distribution of grades given by a judge over all figures in a flight provides measures of a larger number of grades. The FPS however uses figure groups that divide the judges' grades into subsets grouped

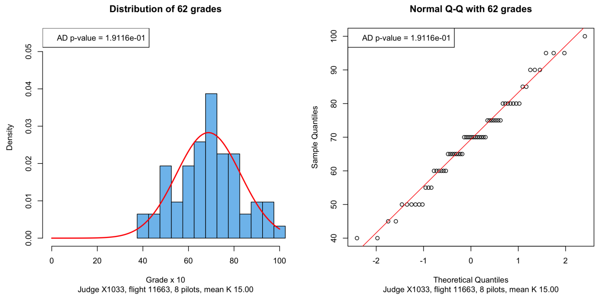

Figure 5: All figures example at maximum Anderson-Darling p-value

by same figure flown. We now look at distributions of grades given by a judge within the FPS groups for a given flight.

The distribution of flight formats is as in Table 7, also selecting for formats in which all of the pilots fly the same figures.

Known	Unknown	Unknown II	Flight 1
7989	5734	175	49

Table 7: Distribution of flight formats for FPS group measurements

In all, we look at 13,298 powered figure groups and 649 glider. The figure groups breakdown by category is as in table 8.

The distributions of pilot and grade count in this data set are as in Table 9.

The correlations between the four measures are stronger than for the all-figures data. Find their values in Table 10.

Table 11 shows summary results from the various goodness of fit tests. The numbers for measures found to be valid and invalid cover all of the results. The other two columns contain counts of p-value results from valid measures greater than and less than or equal to the traditional cutoff of 0.05. There was only one valid result for the Chi-Squared test due to the reduced number of grade data points within the FPS groups. For that reason, we omit the Chi-Squared result.

The tests find support for fit to normal, using the traditional cutoff of 0.05, in a little more than half, roughly 55% of the cases.

Table 6 shows the spread of p-values from the various tests. The numbers are from only those tests reported as valid. The minimum values from the data are always zero, and so omitted from the table.

At the smallest p-value seen in Figure 6 we find a judge who, looking at three figures from seven pilots, almost always gave a grade of 9.0 with the few exceptions being a grade of 8.0.

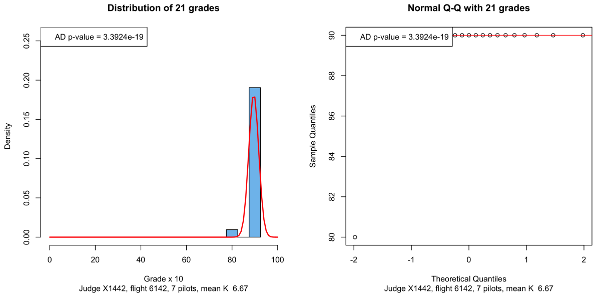

Figure 6: FPS groups example at minimum Anderson-Darling p-value

At the first quantile upper limit seen in Figure 7 we find a judge who most often gives a grade of seven, rarely lower and sometimes higher. This results in a skew toward the lower values.

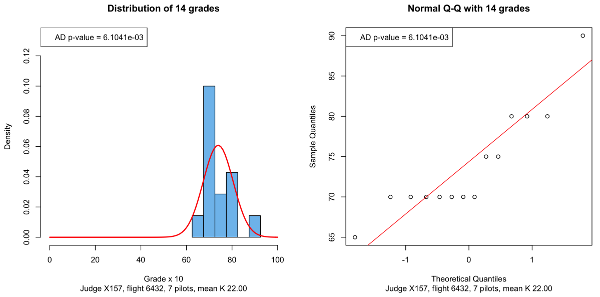

Figure 7: FPS groups example at first quantile Anderson-Darling p-value

At the second quantile upper limit we can look at distributions from three judges looking at the same figure group. This is instructive for seeing the further

primary	sportsman	intermediate	advanced	unlimited
531	6345	4439	2313	319

Table 8: Distribution of categories for FPS group measurements

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Pilot Count	6	7	9	10.82	13	26
Grade Count	12	13	15	16.20	18	48

Table 9: Distribution of pilot and grade counts for FPS group measurements

	sw	lf	ad	cvm
sw	1.000	0.711	0.897	0.813
lf	0.711	1.000	0.870	0.926
ad	0.897	0.870	1.000	0.975
cvm	0.813	0.926	0.975	1.000

Table 10: Correlation of FPS figure group measurements

	Invalid	Valid	> 0.05	<= 0.05
sw.p.value	0	13947	8311	5636
lf.p.value	6	13941	6351	7590
ad.p.value	6	13941	6458	7483
cvm.p.value	6	13941	6640	7301

Table 11: FPS groups GOF measure p-value summary

variety of distributions of grades that judges generate.

In Figure 8 we have simply a judge who favors the grade of 8.0 for thirteen pilots all flying the same figure. The normal distribution would have more of those eights as 7.5 or 7.0 grades, resulting in a little bit of a skew toward the higher grades, mostly due to too many in the middle.

In figure 9 the judge gives an almost uniform distribution from 7.0 to 9.0, although with a complete absence of grade 7.5.

In figure 10 the judge uses a much larger range of grades, from 4.0 to 9.0. Due to there being only thirteen grades in all, the distribution contains zero, one, two, or three instances of each possible grade in the range.

At the third quantile upper limit seen in Figure 11

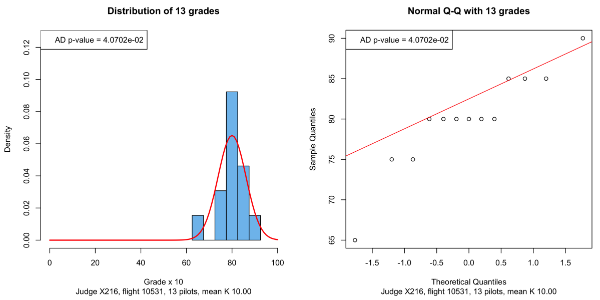

Figure 8: FPS groups first example at second quantile Anderson-Darling p-value

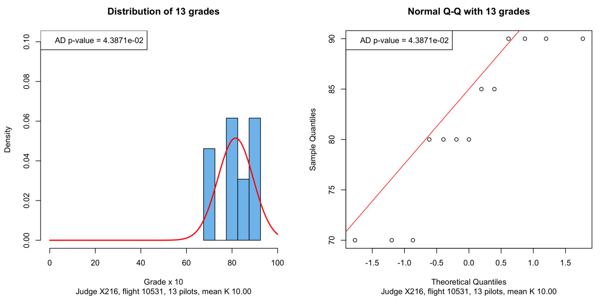

Figure 9: FPS groups second example at second quantile Anderson-Darling p-value

	1st Qu.	Median	Mean	3rd Qu.	Max.
sw.p.value	0.165e-01	0.854e-01	0.180	0.267	0.999
lf.p.value	0.435e-02	0.371e-01	0.115	0.153	0.997
ad.p.value	0.610e-02	0.407e-01	0.107	0.144	0.953
cvm.p.value	0.696e-02	0.446e-01	0.109	0.152	0.925

Table 12: FPS group GOF measure p-value distributions

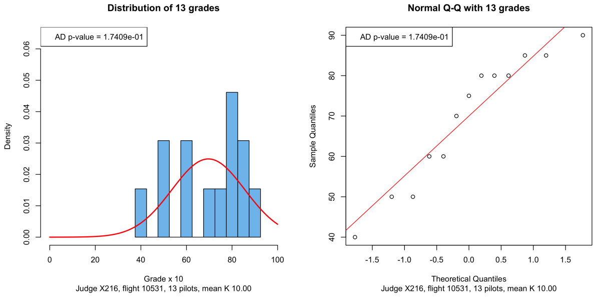

Figure 10: FPS groups third example at second quantile Anderson-Darling p-value

the judge gives most of their grades in the middle of their range, with a couple of instances above or below. More in the middle and fewer at the tails makes for a better fit to normal, but few of the individual judge grade counts match the density that would be expected from the normal.

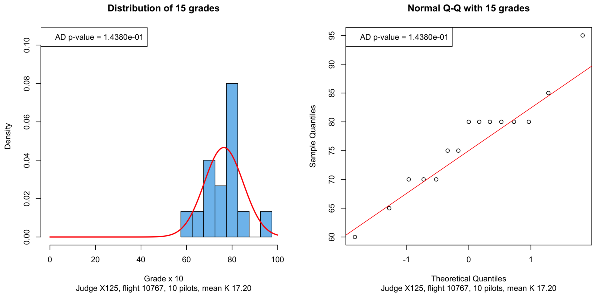

Figure 11: FPS groups example at third quantile Anderson-Darling p-value

At the maximum value seen in Figure 12 we have a better filled spread of grades with, as for the third quantile, more grades near the mean.

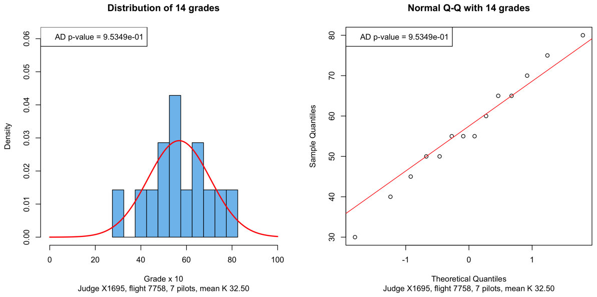

Figure 12: FPS groups example at maximum Anderson-Darling p-value

5 Conclusions

The judge grade data is rounded data because judges are limited to giving grades in increments of 0.5. The literature provides precedent for applying random perturbations to make the data continuous, then using the continuous tests for goodness of fit to normal.

The Chi-Squared test used with discrete data is shown in the literature to be less powerful than the continuous tests applied to rounded data with perterbations. However, in any case, although considerable effort was applied, there are frequently too few data values in these sets of grades with which to provide a valid Chi-Squared result.

The goodness of fit measures looking at all of the grades given by a judge during a flight fail to support the null hypothesis– that the grades fit a normal distribution derived from their mean and standard deviation –more than half of the time. The measures looking at FPS figure groups, that encompass fewer grades but for a single or small number of figures, fail to support the null hypothesis slightly less than half the time.

Looking at histograms of the grade frequencies overlaid with the normal model curves reveals a large variation in grade distributions given by judges and illustrates the lack of conformity to the normal distribution.

The lack of conformity of judge grade distributions to the normal distribution suggests reconsideration of the FPS scoring method, that uses at its foundation a normal distribution model of the judge grades.

References

[1] Acd stats. `https://github.com/wbreeze/acd_stats/`.

[2] Bates distribution. `https://en.wikipedia.org/wiki/Bates_distribution`.

[3] Civa results. `https://civa-results.com/`.

[4] Fai sporting code, section 6 international aerobatic events statistical method for processing scores version 2018-1. `https://www.fai.org/sites/default/files/documents/section6_part1_appendixfps_v2018_1.pdf`.

[5] Iac contests. `https://iaccdb.iac.org/`.

[6] World aerobatic championships. `https://en.wikipedia.org/wiki/FAI_World_Aerobatic_Championships`.

[7] M.J. Greenacre. Clustering the rows and columns of a contingency table. *Journal of Classification*, 5, 1988. Greenacre, M.J. Journal of Classification (1988) 5: 39. https://doi.org/10.1007/BF01901670.

[8] et. al. Jackson, B. An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters*, 12, 2005.

[9] Boris Lemeshko, Chimitova E.V, and Kolesnikov S.S. Nonparametric goodness-of-fit tests for discrete, grouped or censored data. 05 2007.

[10] Jeff Letton. Combine categories using greenacre's method. `https://cran.r-project.org/package=greenclust`. GitHub:https://github.com/JeffJetton/greenclust.

[11] Janet Chaseling Michael Steel and Cameron Hurst. Comparing the simulated power of discrete goodness-of-fit tests for small sample sizes. In *ASIMMOD*, pages 210–216, 2007.