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

Douglas Lovell

November, 2019

1 Aerobatic Competition

Aerobatics, as practiced by CIVA and the IAC, is a judged sport. In aerobatics, pilots fly figures made up of a basic flight path overlaid with rolls, in which the airplane moves about it's fuselage while it's center continues on the flight path.

The basic flight paths of the figures are made up of straight lines and loops. Straight lines may be horizontal, vertical, or along a forty-five degree line. Loops are segments of a circle of constant radius. They occur in transitions between lines. Their segments may form anything from one-eighth of a circle to an entire circle in one-eighth increments.

Rolls superimposed on the flight paths are subdivided into those in which the air maintains laminar flow over the surface of the airplane and those in which the flow is turbulent over part of the wing. The former are generally referred to as "aileron-rolls" or simply "rolls." The latter are referred to as "snap-rolls" or "flick-rolls", and "spins".

Aileron rolls and snap rolls may occur on any line, and at the top of full loops. Spins are always performed on a vertical line moving down. Rolls can occur in one-eighth increments up to two full rolls, and may be broken-up with "points", or pauses, in increments of one-eighth, one-quarter, and one-half.

The sport maintains a catalog of flight paths and rolls named the "Aresti catalog" after its developer, Spanish aviator Colonel José Luis Aresti Aguirre. [Reference to Aresti catalog] The Aresti catalog systematically assigns difficulty, a "K-factor" to each figure.

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 the sequence actually flown against the sequence commited to fly by the pilot.

1.1 Grading

The judges give each figure a grade from zero to ten in half point increments. They start from a perfect score of ten, and systematically deduct for flaws in the direction of the flight path, radius of loops, and degrees of roll. Thus, a flight program produces a three-dimensional matrix of grades consisting of Pilot X Judge X Figure. Each pilot receives one grade from each judge for each figure.

To complicate the grading slightly, a zero grade can take multiple forms. One form is that incremental flaws resulted in deductions that summed to ten or more. This is known as a "soft zero". A second form is that the figure flown did not match the figure that was supposed to be flown. It might have been missing a roll element or not have been flown along the prescribed flight path. This is known as a "hard zero". CIVA uses a form known as a "presentation zero" that a judge may use to indicate suspicion of a hard zero. A "conference zero" indicates that the judge changed their grade to a hard zero in review following the performance of the pilot.

A judge may also give a grade of "average". A grade of average indicates that the judge was either distracted or unsure of the figure that was prescribed to be flown and therefore unable to evaluate the figure performed against the figure prescribed. It is preferred that the judge give an average under these conditions, rather than make-up a grade. The effect of averages is a reduction in the number of judges evaluating the figure.

2 Clustering for Chi-Square

In order to apply the Chi-Square metric, we need to cluster grades which were given by the judge less than six times together with neighboring grades.

Example

The optimal method for doing this is the one that has least impact on the mean and variance of the data. The method described by Greenacre [2] and implemented by the greenclust [4] R package provides one; however, it reorders the clusters. We need to join only adjacent clusters, to maintain the order.

Another paper [3] from ten authors at NASA Ames and San Jose State University maintains order of partitions, but does not enable the constraint of minimum partition size. This and other algorithms for k-shape ordered partitioning require a cumulative function that measures the quality of each partition. Here, we find that the mean and variance do not always decrease or increase when joining two clusters.

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

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.

We report the mean, variance, and p value goodness of fit for both the clustered normal model and the normal model generated by FPS.

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.

2.1 Number of grades

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 six minus one, minus two for the two estimated values– mean and standard deviation. This means that, with six clusters we have three degrees of freedom. Three degrees of freedom is the minimum 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.

Finally, we produce a measure with all of the grades from a judge for a flight.

3 Shapiro-Wilk

The Shapiro-Wilk test for goodness of fit requires that all of the values are unique. In order to use the Shapiro-Wilk test, we cannot use the grades without perturbing them in order to make them unique. In one view, the grades given by the judges are severely rounded to increments of 0.5 from a continuum of performances by the pilots.

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.

The advantage of using a uniform distribution is that we can uniformly and reliably be perturbed within the range. The 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. The standard deviation must be such that the resulting perturbations are extremely rarely, with very low probability, outside of the range -2.5 to 2.5.

We chose to use for standard deviation a value that approximates the distribution of Bates [1]. The Bates distribution has the advantage of being distributed along a fixed interval around a mean, with decreasing probability for values further from the mean. A mean of zero and variance of $1/12n * 5$, where n is the number of instances of the grade, when used with the normal distribution, approximates the distribution of Bates between -2.5 and 2.5.

We perturbed the grades using random values chosen from a normal distribution with mean equal to zero and standard deviation equal to $\sqrt{5/12n}$. Choosing a standard deviation dependent on the number of instances of the grade causes the perturbations to remain closer to zero when there are more instances of a grade and within the -2.5 to 2.5 range. We report the p-value for the Shapiro-Wilk normality test after perturbing the grades.

4 Judge grade distributions for 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.

Known	Unknown	Unknown II	Flight 1
1488	1160	27	19

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

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, second and third known flights. The distribution of flight formats is as in Table 1:

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 traditional 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

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	9.348	11	26
Grade Count	35	64	81	94.97	110	359

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

	sw	lf	ad	cvm
sw	1.000	0.417	0.795	0.613
lf	0.417	1.000	0.696	0.809
ad	0.795	0.696	1.000	0.935
cvm	0.613	0.809	0.935	1.000

Table 4: Correlation of all-figures measurements

	Invalid	Valid	> 0.05	<= 0.05
chiSq.t.p	1243	1607	966	641
chiSq.d.p	1243	1607	851	756
sw.p.value	0	2850	122	2728
lf.p.value	0	2850	36	2814
ad.p.value	0	2850	35	2815
cvm.p.value	0	2850	51	2799

Table 5: All figure GOF measure p-value summary

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. 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.

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 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.

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.

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.

References

[1] Bates distribution. Wikipedia.

[2] M.J. Greenacre, 1988. Greenacre, M.J. Journal of Classification (1988) 5: 39. https://doi.org/10.1007/BF01901670.

	1st Qu.	Median	Mean	3rd Qu.	Max.
chiSq.t.p	0.006	0.124	0.272	0.492	0.999
chiSq.d.p	0.002	0.067	0.216	0.358	0.993
sw.p.value	0.012e-04	0.914e-04	0.837e-02	0.229e-02	0.452
lf.p.value	0.0	9.700e-07	2.453e-03	1.246e-04	2.208e-01
ad.p.value	2.000e-08	7.350e-06	2.694e-03	3.441e-04	1.912e-01
cvm.p.value	4.400e-07	2.989e-05	3.383e-03	6.480e-04	2.781e-01

Table 6: All figure GOF measure p-value distributions

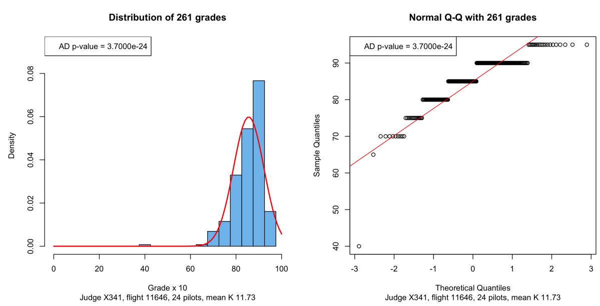

Figure 1: All figures example at minimum Anderson-Darling p-value

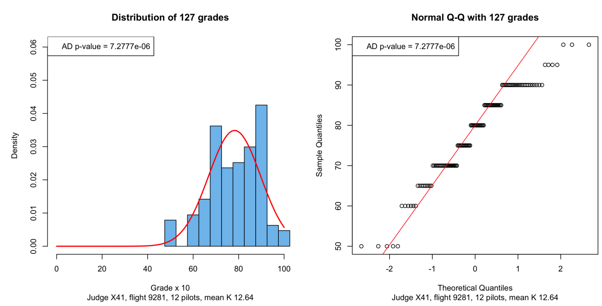

Figure 3: All figures example at second quantile Anderson-Darling p-value

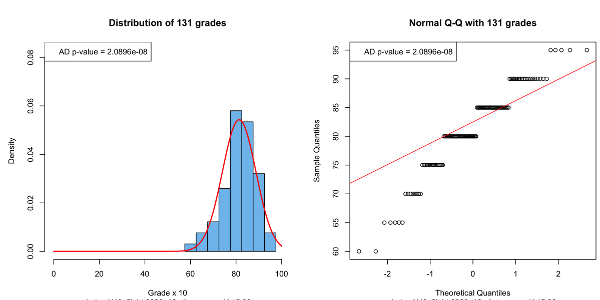

Figure 2: All figures example at first quantile Anderson-Darling p-value

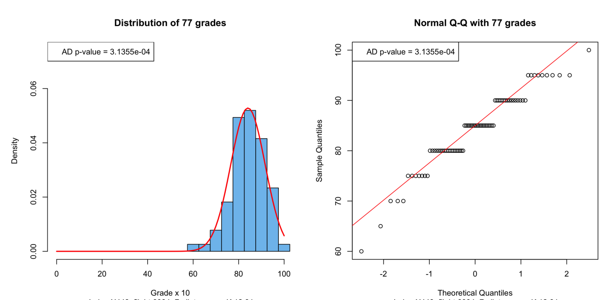

Figure 4: All figures example at third quantile Anderson-Darling p-value

[3] et. al. Jackson, B. An algorithm for optimal partitioning of data on an interval, 2005.

[4] Jeff Letton. Combine categories using greenacre's method. CRAN. GitHub:https://github.com/JeffJetton/greenclust.

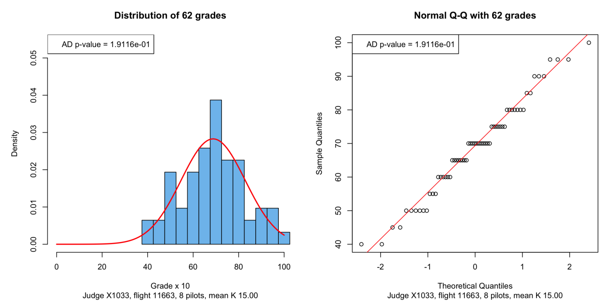

Figure 5: All figures example at maximum Anderson-Darling p-value