Some Heuristics for Analysis of Variance

BY K. RUBEN GABRIEL*

University of Rochester

Definition of the variance of a sample as one half of the average squared difference is more intuitive than the common definition as an average squared deviation from the mean. Similarly, the one-way ANOVA F-statistic has intuitively appealing definitions as the average of squared t-statistics, either of the t's for all pairwise comparisons of means, or of the t's for comparing each mean with the average of all the others. These are distinct from Scheffé's definition of F as the square of the maximum of t's for all contrasts. Analogous definitions extend to two-way ANOVA and to MANOVA.

It is more intuitive to define overall test statistics as averages of statistics for simple components of the overall hypothesis than by the classical definitions, and there is much to be said for that in terms of heuristic appeal and didactic usefulness.

1. Introduction

In statistical inference, the choice and definition of a statistic is determined by its optimality properties for estimation and testing. For example, the center of a sample is usually defined as the mean because that definition is in many ways optimal for Gaussian data. In descriptive statistics, on the other hand, the purpose of a statistic is to reveal interesting features of the data, so the choice of a statistic becomes heuristic. Thus, the median may be chosen to describe a sample's center because of its intuitive appeal as bisecting the observations into an equal number of larger and smaller ones; the mean is much less appealing as it is defined by a quite non-intuitive algorithm involving addition and division. Of course, some statistics may be intuitively motivated as well as optimally inferential in certain contexts, and in non-parametric inference heuristic criteria are often used because optimality is difficult to define.

Heuristics are also important for intuitively motivating the use of particular statistics. Thus, it is didactically preferable to introduce the median as having the simple property of bisecting by size, rather than as having the more abstract property of minimizing the sum of absolute deviations. Several other well known statistics are shown in this paper to have simple definitions with heuristic appeal, and these may be used to advantage in teaching instead of the classical less intuitive definitions.

2. The sample variance

The variance of a sample of observations x_i, $i = 1, \dots, n$, is usually defined by means of deviations from the mean $\bar{x} = \sum_{i=1}^{n} x_i/n$

*K. Ruben Gabriel died 25 May 2003. Ruben was a good friend to the editors and many of the authors whose work is represented in this volume, and we will miss him.

as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{2.1}$$

An alternative, but equivalent, definition based on the differences $x_i - x_e$ between all pairs of observations, due to Gini [7], is

$$s^2 = \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{e=1}^{n}(x_i - x_e)^2. \tag{2.2}$$

We are more familiar with (2.1) but (2.2) is intuitively more appealing. Common sense would measure variability by averaging the (squared) *differences between observations* rather than the (squared) *deviations from the mean.* Why involve the mean in defining a measure of variability? – The definition of the inter-quartile range does not involve the median or any other measure of the center. Of course, the measurement of variability by means of squares rather than of absolute values still needs to be motivated for both (2.1) and (2.2), and heuristics are not going to help there. Note, by the way, that (2.2) may be seen as a U-statistic of spread in the sense of Bickel and Lehmann [1]: this is an instance where the non-parametric approach uses a more heuristic definition.

The proof of the equivalence of (2.1) and (2.2) is quite elementary. The following proves the more general, weighted, equality

$$\sum_{i=1}^{n}\sum_{e=1}^{n} w_i w_e (x_i - x_e)^2 = \sum_{e=1}^{n} w_e \sum_{i=1}^{n} w_i (x_i - \bar{x}_w)^2 \tag{2.3}$$

where $w_1, \ldots, w_n$ are the weights and $\bar{x}_w = \sum_{i=1}^{n} w_i x_i / \sum_{i=1}^{n} w_i$ is the weighted mean. First, $\sum_{i=1}^{n}\sum_{e=1}^{n} w_i w_e (x_i - x_e)^2$ is rewritten as $\sum_{i,e=1}^{n} w_i w_e \{(x_i - \bar{x}_w) - (x_e - \bar{x}_w)\}^2$ which becomes $\sum_{i=1}^{n} w_i (x_i - \bar{x}_w)^2 \sum_{e=1}^{n} w_e - 2\sum_{i=1}^{n}\sum_{e=1}^{n} w_i (x_i - \bar{x}_w) w_e (x_e - \bar{x}_w) + \sum_{i=1}^{n} w_i \sum_{e=1}^{n} w_e (x_e - \bar{x}_w)^2$. The middle term equals zero and the other two terms are equal, and that proves the equality.

The prevalence of (2.1) in teaching, and the common ignorance of (2.2), is partly due to the former being easier to compute, for it involves fewer terms. But computational convenience should not be a didactic criterion. Indeed, in the days of pencil and paper calculation, we used

$$s^2 = \frac{1}{n-1}\left\{\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n\right\} \tag{2.4}$$

and looked up the squares in Barlow's Tables (Comrie, [4]). But no one suggested then that this formula gave insight into the idea of a variance, nor does anyone now propose to introduce the variance by means of algorithms used by calculators or computer software. So why continue to introduce the variance by (2.1)? – Surely the more heuristic (2.2) would make it easier to motivate the concept of measuring variability.

3. The one-way analysis of variance F-statistic The usual formulation of one-way ANOVA focuses on the F -statistic defined as the ratio

$$(3.1)\qquad \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x})^2/[(k-1)s_p^2],$$

with obvious notation, s_p^2 being the pooled "within samples" variance estimate. This definition is commonly motivated as a "between to within " comparison of alternative estimates of the variance, a definition that implicitly assumes understanding of the distribution of the sum of squares of sample means. More intuitive explanations can be provided by using one or another of the following alternative, but equivalent, definitions of this statistic. All of these definitions are based entirely on t-statistics which are more intuitive in their structure since each one of them involves only a single comparison.

The first alternative is based on the pairwise t-statistics $t_{j,g} = (\bar{x}_j - \bar{x}_g)/[s_p\sqrt{n_j^{-1} + n_g^{-1}}]$ that compare the means $\bar{x}_j$ and $\bar{x}_g$ of pairs of samples j, g, but use the within samples estimate s_p^2, rather than the more commonly used separate estimates based only on samples j and g. This allows the definition of the F-statistic as

$$(3.2)\qquad F = \frac{1}{2(k-1)n}\sum_{j=1}^{k}\sum_{g=1}^{k}(n_j + n_g)t_{j,g}^2$$

where $n = \sum_{j=1}^{k} n_j$. Since $\sum_{j=1}^{k}\sum_{g=1}^{k}(n_j + n_g) = 2(k-1)n$, this is an average of the $t_{j,g}^2$'s, i.e., of the two-sided pairwise test statistics, each weighted by the number of observations in that pair of samples. The proof of (3.2) follows the same lines as that of (2.2).

Definition (3.2), though computationally more cumbersome than (3.1), is heuristically more attractive: It uses the collection of sample-pair test statistics to define the all-sample statistic, just as the collection of hypotheses of pairwise equalities is equivalent to the hypothesis of equality of all expectations. It is intuitive to reject the overall hypothesis if, and only if, the statistics testing the pairwise hypotheses are large.

The second alternative is based on the one-vs-the-rest statistics $t_{j,\backslash j} = (\bar{x}_j - \bar{x}_{\backslash j})/[s_p\sqrt{n_j^{-1} + (n - n_j)^{-1}}]$ that compare one sample mean $\bar{x}_j$ to the mean $\bar{x}_{\backslash j} = (n\bar{x} - n_j\bar{x}_j)/(n - n_j)$ of all $n_{\backslash j} = n - n_j$ observations outside that sample, again using the within samples estimate s_p^2. The subscript $\backslash j$ reads "all except". This definition of the F-statistic is

$$(3.3)\qquad F = \frac{1}{(k-1)n}\sum_{j=1}^{k} n_{\backslash j} t_{j,\backslash j}^2 .$$

Since $\sum_{j=1}^{k} n_{\backslash j} = (k-1)n$, this is a weighted average of the $t_{j,\backslash j}^2$'s, i.e., of the two-sided one-vs-the- rest test statistics weighted by the numbers of observations in the rest of the samples. To prove (3.3), introduce the definition of $t_{j,\backslash j}$ and rearrange terms.

Definition (3.3) does not involve much more computation than (3.1), and yet is heuristically more attractive: It defines the all-sample statistic as an average of the k statistics for comparing any one sample with the average of all other samples, just as the hypothesis of equality of all expectations is equivalent to the collection of hypotheses that each individual expectation is equal to the average of all the other expectations. And again it is intuitive to reject the overall hypothesis if, and only if, the statistics for the one-vs-the-rest hypotheses are large.

A third alternative is based on the t -statistics $t_c = (\sum_j c_j\bar{x}_j)/[s_p\sqrt{\sum_j(c_j^2/n_j)}]$. for all contrasts $\sum_j c_j\bar{x}_j$. That formula is less intuitive in the way it accommodates the different coefficients $(c_1,\ldots,c_k)$, but it is easy to see that it simplifies to the above two t's when the contrasts have, respectively, $c_j = 1, c_g = -1$ and all other coefficients zero, or $c_j = 1$ and $c_g = -1/(k-1)$ for all $g \neq j$. These t's for contrasts allow the third alternative definition

$$F = \max_{c_1,\ldots,c_k:\sum c_i=0} t_c^2/(k-1) \tag{3.4}$$

as a constant multiple of the maximum of the squares of the t's for all contrasts (Scheffé, [9], Appendix III; Gabriel and Peritz, [5]).

Definition (3.4) is intuitively appealing since it shows the F-statistic to be large if and only if at least one contrast has a large t_c^2-statistic. It corresponds to the equivalence of the overall equality hypothesis with the intersection of the hypotheses that individual contrasts have zero expectation. Its intuitive appeal has led Brown and Hollander [3] to use it to introduce the F-statistic in their textbook.

Each one of the above alternative definitions uses a decomposition of the overall hypothesis into components such that the former is true if, and only if, *all* of the latter are true, i.e., respectively, all the pairwise contrasts are null, all the one-vs-the-rest contrasts are null, and all contrasts are null. It is therefore intuitive to reject the overall hypothesis if, and only if, at least one component hypothesis is rejected, and therefore to define the overall statistic as the *maximum* of the component statistics. That, indeed, is the Union-Intersection principle (Roy, [8]). Definition (3.4) is the only one which satisfies that principle, whereas (3.2), (3.3) define the overall statistic as an *average* of the component statistics, a heuristically less compelling choice. (Indeed, if one applied the Union-Intersection principle to the pairwise components one would obtain the Studentized Range statistic instead of the F-statistic). Definitions (3.2), (3.3), on the other hand, are based on simpler t-statistics which compare two means, whereas definition (3.4) is based on t-statistics for all contrasts, a more complicated notion.

4. F-statistics in balanced two-way ANOVA The usual analysis of a two-factor design involves three F-ratios, one for each main effect and one for interaction. Each of these F-statistics can be expressed as a mean of appropriate t^2's, as follows. Denote the A and B factor levels as $i(= 1,\ldots,k)$ and $j(= 1,\ldots,q)$, respectively, the number of replications in the (i,j) -th cell as $n_{i,j} = n_{i+}n_{+j}/n$ (this expression being possible since the layout is assumed to be balanced), where $n_{i+} = \sum_{j=1}^q n_{i,j}$, $n_{+j} = \sum_{i=1}^k n_{i,j}$, and $n = \sum_{i=1}^k n_{i+} = \sum_{j=1}^q n_{+j}$. The mean in cell (i,j) is denoted by $\bar{x}_{ij}$, and the marginal means by $\bar{x}_{i\cdot} = \sum_{j=1}^q n_{+j}\bar{x}_{ij}/n$, $\bar{x}_{\cdot j} = \sum_{i=1}^k n_{i+}\bar{x}_{ij}/n$ and

$\bar{x}_{\cdot\cdot} = \sum_{i=1}^{k} \sum_{j=1}^{q} n_{i+}n_{+j}\bar{x}_{ij}/n$. Also, s^2 denotes the independent error variance estimate.

The A main effect statistic can be defined as the average

$$F_A = \frac{1}{2(k-1)n} \sum_{i=1}^{k} \sum_{e=1}^{k} (n_{i+} + n_{e+}) t_{i,e\in A}^2 \tag{4.1}$$

of the squares of the pairwise A levels statistics $t_{i,e\in A} = (\bar{x}_{i\cdot} - \bar{x}_{e\cdot})/[s\sqrt{n_{i+}^{-1} + n_{e+}^{-1}}]$. This is analogous to definition (3.2). The B main effect statistic can similarly be defined as the average

$$F_B = \frac{1}{2(q-1)n} \sum_{j=1}^{q} \sum_{g=1}^{q} (n_{+j} + n_{+g}) t_{j,g\in B}^2 \tag{4.2}$$

of the squares of the pairwise B levels statistics $t_{j,g\in B} = (x_{\cdot j} - \bar{x}_{\cdot g})/[s\sqrt{n_{+j}^{-1} + n_{+g}^{-1}}]$. The AB interaction statistic can also be defined as an average

$$F_{AB} = \frac{1}{4n(k-1)(q-1)} \sum_{i=1}^{k} \sum_{e=1}^{k} \sum_{j=1}^{q} \sum_{g=1}^{q} (n_{i+} + n_{e+})(n_{+j} + n_{+g}) t_{i,e\in A,j,g\in B}^2 \tag{4.3}$$

of the squares of the tetrad difference statistics

$$t_{i,e\in A,j,g\in B} = (\bar{x}_{ij} - \bar{x}_{ej} - \bar{x}_{ig} + \bar{x}_{eg})/[s\sqrt{\frac{n}{n_{i+}n_{+j}} + \frac{n}{n_{e+}n_{+j}} + \frac{n}{n_{i+}n_{+g}} + \frac{n}{n_{e+}n_{+g}}}],$$

which test AB interaction at A levels i and e and B levels j and g (Gabriel, Wax and Putter, [6]). This can be proved by applying the earlier arguments twice, once for j and g and once for i and e. Again, (4.3) provides a heuristic basis for explaining F_{AB} since null AB interaction is equivalent to the lack of interaction between any two levels of A and any two levels of B.

Alternative definitions are based on one-vs-the-rest t-statistics analogous to (3.3). Other alternatives are those based on maxima of the appropriate linear sets of contrasts (Bradu and Gabriel, [2]) and are analogous to those of (3.4).

Similar results could be derived for balanced higher-way ANOVAs.

5. The Hotelling-Lawley Trace of one-way MANOVA

One of the statistics for testing one-way MANOVA is the Hotelling-Lawley trace

$$tr\{\mathbf{S}_b\mathbf{S}_p^{-1}\} = tr\{\sum_{j=1}^{k} n_j(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})^T\mathbf{S}_p^{-1}\}, \tag{5.1}$$

where $\bar{\mathbf{x}}_j$ is the multivariate (column) vector of means of the j-th sample whose size is n_j, the overall mean vector is $\bar{\mathbf{x}} = \sum_{j=}^{k} n_j\bar{\mathbf{x}}_j/n$, where $n = \sum_{j=1}^{k} n_j$, and $\mathbf{S}_p$ is the pooled "within samples" variance estimate. This is commonly presented as a multivariate generalization of the between-to-within F-ratio of sums of squares in univariate one-way ANOVA. More intuitive motivations can be provided by using one or another of the following alternative, but equivalent, definitions of this trace statistic, all of which are based entirely on Hotelling T^2-statistics which are more intuitive in their structure.

The first alternative is based on the pairwise T^2-statistics

$$T^2_{j,g} = (\bar{\mathbf{x}}_j - \bar{\mathbf{x}}_g)^T \mathbf{S}_p^{-1} (\bar{\mathbf{x}}_j - \bar{\mathbf{x}}_g)/(n_j^{-1} + n_g^{-1})$$

that compare the mean vectors $\bar{\mathbf{x}}_j$ and $\bar{\mathbf{x}}_g$ of pairs of samples j, g . Thus, the trace in (5.1) can be defined as

$$(5.2) \qquad tr\{\mathbf{S}_b \mathbf{S}_p^{-1}\} = \frac{1}{2(k-1)n} \sum_{j=1}^{k} \sum_{g=1}^{k} T^2_{j,g},$$

an average of the $T^2_{j,g}$'s, i.e., of the pairwise test statistics. The proof of (5.2) follows the same lines as that of (3.2).

The second alternative definition is based on one-vs-the-rest T^2-statistics analogous to (3.3). A third alternative is that based on maxima of the appropriate linear sets of contrasts and is analogous to those of (3.4). Again, similar definitions can be derived for higher order MANOVA.

6. Concluding remarks Alternative definitions of the variance, and of ANOVA and MANOVA statistics, have been pointed out as being more heuristic, that is, as having more intuitive appeal than the definitions that are almost invariably used in classes and in texts. The classical definitions are conceptually more difficult, but had presumably been introduced because of their analogies with moments in physics. The beauty of their generalizations into lengths of projections onto subspaces still makes them attractive for the mathematically sophisticated. Also, they were computationally simpler than the heuristic definitions, and that justified their use when pencil and paper computations were used but is irrelevant today. Better intuitive understanding and more effective teaching should now be sought by using the more heuristic definitions, either instead of, or in addition to the classical ones.

REFERENCES

[1] P.J. Bickel and E.L. Lehmann. Descriptive statistics for nonparametric models iv. spread. In J. Jureckova, editor, *Contributions to Statistics*, pages 33–40, Dodrecht, Holland, 1979. D. Reidel.

[2] D. Bradu and K.R. Gabriel. Simultaneous statistical inference on interactions in two-way analysis of variance. *Journal of the American Statistical Association*, 69:428–436, 1974.

[3] B.W. Brown and M. Hollander. *Statistics: A Biomedical Approach.* Wiley, New York, 1977.

[4] L. J. Comrie. *Barlow's Tables of Squares, Cubes, Square Roots, Cube Roots and Reciprocals of all Integers up to 12,500.* Spon, London, 1941.

[5] K.R. Gabriel and E. Peritz. Least squares and maximal contrasts. *International Statistical Review*, 41:155–164, 1973.

[6] K.R. Gabriel, Y. Wax, and J. Putter. Simultaneous confidence intervals for product-type interaction contrasts. *Journal of the Royal Statistical Society Series B*, 35:234–244, 1973.

[7] C. Gini. *Memorie di metodologica statistica.* Veschi, Rome, 1912.

[8] S.N. Roy. *Some Aspects of Multivariate Analysis.* Wiley, New York, 1957.

[9] H. Scheffé. *The Analysis of Variance.* Wiley, New York, 1959.