PROCEEDINGS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 126, Number 4, April 1998, Pages 1177–1179
S 0002-9939(98)04524-9

VARIANCE AND CLUSTERING

YANNIS G. YATRACOS

(Communicated by Wei Y. Loh)

ABSTRACT. A measure of dissimilarity of real observations is introduced that is used to identify clusters and is based on "gaps", and also on averages of selected subgroups of the observations. This measure is surprisingly associated with the sample variance in a way that leads to a new identity and interpretation of the notion of variance.

INTRODUCTION. THE RESULTS

In cluster analysis the grouping of objects is done on the basis of their similarities (or dissimilarities). A group of objects is divided initially in two subgroups such that the objects in one subgroup are "far" from the objects in the other, with respect to a chosen dissimilarity measure. The subgroups may be further divided in the same way (Johnson and Wichern, 1988 [4]).

A dissimilarity measure used to identify *remote* clusters of real observations should also reflect the distance (the "gap") between these clusters. The use of "gaps" to determine clusters is not uncommon in practice. For example, to determine groups of students receiving the same letter grade one usually looks for large "gaps" in the list of the (ordered) numerical grades. Tukey (1949 [8]) proposed the examination of "gaps" between adjacent means as a first step to break up treatment means in groups, even though he later abandoned this significance-based method in favor of confidence-interval-based methods.

To divide real observations $X_1, ..., X_n$ into two clusters it is enough to examine the i smallest observations and the $(n-i)$ largest observations for $i = 1, ..., n-1$. The examination of such groups leads to the idea of examining similarity or divergence of separated groups of observations; that is, of groups of observations with convex hulls that do not intersect. Let $X_{(1)}, X_{(2)}, ..., X_{(n)}$ be the order statistic, that is, a permutation of $X_1, X_2, ..., X_n$ with $X_{(1)} \le X_{(2)} \le ... \le X_{(n)}$, and define $\bar{X}_{[k,m]} = \frac{1}{m-k+1}\sum_{j=k}^{m} X_{(j)},\ k \le m \le n$, for $\bar{X}_{[1,i]}$ and $\bar{X}_{[i+1,n]}$ to denote the averages of the i smallest observations and the $(n-i)$ largest observations respectively, for $i = 1, ..., n-1$; note that the average of the observations $\bar{X} = \bar{X}_{[1,n]}$. Classical measures of similarity (or divergence) of groups involve the difference $\bar{X}_{[i+1,n]} - \bar{X}_{[1,i]}$ (usually) squared. For separated groups this difference alone cannot describe similarity or divergence; for a given value of $\bar{X}_{[i+1,n]} - \bar{X}_{[1,i]}$ the distance $X_{(i+1)} - X_{(i)}$ between the two groups may vary. It is then rather natural that the difference $\bar{X}_{[i+1,n]} - \bar{X}_{[1,i]}$

Received by the editors October 4, 1996.
1991 *Mathematics Subject Classification.* Primary 62H30, 62J10.
Key words and phrases. Clustering, divergence of populations, analysis of variance.

as well as the "gap" $X_{(i+1)} - X_{(i)}$ (or other normalizing factors) enter in a measure of divergence of separated groups. A measure involving both these quantities is related with the sample variance as follows:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n-1}\frac{i(n-i)}{n^2}(\bar{X}_{[i+1,n]} - \bar{X}_{[1,i]})(X_{(i+1)} - X_{(i)}). \tag{1}$$

The sample variance is decomposed as the sum of the *divergence measures*

$$\frac{i(n-i)}{n^2}(\bar{X}_{[i+1,n]} - \bar{X}_{[1,i]})(X_{(i+1)} - X_{(i)})$$

of selected subsamples, thus leading to a new interpretation of the sample variance. The term that contributes the most in the sum determines the potential clusters. This decomposition holds also with $X_i, X_{i+1}, \bar{X}_{1,i} = \frac{1}{i}[X_1 + ... + X_i]$, and $\bar{X}_{i+1,n} = \frac{1}{n-i}[X_{i+1} + ... + X_n]$ instead of $X_{(i)}, X_{(i+1)}, \bar{X}_{[1,i]}, \bar{X}_{[i+1,n]}, 1 \le i \le n-1$, but the divergence measures are all positive when the observations are ordered. The coefficient $i(n-i)$ is used to balance the number of i "small" and $(n-i)$ "large" observations, and appears also with the term $(\bar{X}_{[i+1,n]} - \bar{X}_{[1,i]})^2$ in Karl Pearson's *coefficient of racial likeness* (Pearson, 1926 [5]). The divergence measures, standardized dividing by $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$, appear also in a new decomposition of the least squares estimate of the slope in simple regression, and may be used along with the residuals to identify high leverage points (Yatracos, 1997 [9]). Their maximum can also be used to break up treatment means into groups.

Relation (1) may be derived as a corollary of Proposition 1 (the population variance identity) using the empirical cumulative distribution function and the Lebesgue-Stieltjes theory (see for example Ross, 1984, [7], p. 299). Proposition 1 may also be derived (Pitman, 1996 [6]) from Hoeffding's covariance formula (Hoeffding, 1940 [3]), which may be found in Block and Fang (1988 [2]).

Proposition 1. *For a random variable X with distribution function F and with $E|X| < +\infty$,*

$$Var(X) = \int_{-\infty}^{+\infty} P(X > y)P(X \le y)[E(X|X > y) - E(X|X \le y)]dy. \tag{2}$$

Proof. The integrand in (2) is non-negative; therefore the integral can be broken into two parts, each computed separately, with domains of integration from 0 to $+\infty$ and from $-\infty$ to 0. Denote $X^+ = max(X, 0), X^- = max(-X, 0), I_A(x) = 1$ if $x \in A$ and 0 otherwise; note that $X^+ \ge 0, X^- \ge 0, EX = EX^+ - EX^-$. The integral from 0 to $+\infty$ can be written as follows:

$$\int_0^{+\infty} P(X > y)P(X \le y)[E(X|X > y) - E(X|X \le y)]dy$$

$$= \int_0^{+\infty}\int_0^{+\infty} xI_{(y,+\infty)}(x)dF(x)dy - \int_0^{+\infty} P(X > y)\int_{-\infty}^{+\infty} xdF(x)dy. \tag{3}$$

By Tonelli's theorem (Billingsley, 1979, [1], p. 200) the order of integration is reversed in the first integral in (3). Since for any non-negative random variable W, $EW = \int_0^{+\infty} P(W > w)dw = \int_0^{+\infty} P(W \ge w)dw$ (Billingsley, 1979, [1], p. 239), (3) becomes

$$\begin{aligned}\int_0^{+\infty} x \int_0^x I_{(y,+\infty)}(x)dydF(x) - EX\int_0^{+\infty} P(X>y)dy \\ = \int_0^{+\infty} x^2dF(x) - EXEX^+.\end{aligned} \tag{4}$$

The integral from $-\infty$ to 0 can be written as

$$\int_{-\infty}^0 P(X \le y)EXdy - \int_{-\infty}^0\int_{-\infty}^0 xI_{(-\infty,y]}(x)dF(x)dy. \tag{5}$$

With similar arguments as for (3) the integrals in (5) become:

$$EX\int_{-\infty}^0 P(-X^- \le y)dy = EX\int_0^{+\infty} P(X^- \ge z)dz = EXEX^-, \tag{6}$$

$$\begin{aligned}\int_{-\infty}^0\int_{-\infty}^0 xI_{(-\infty,y]}(x)dF(x)dy &= \int_{-\infty}^0 x\int_{-\infty}^0 I_{(-\infty,y]}(x)dydF(x) \\ &= -\int_{-\infty}^0 x^2dF(x).\end{aligned} \tag{7}$$

Using (3)-(7) and the relation $VarX = E(X^2) - (EX)^2$, we obtain (2). □

Acknowledgements

Thanks are due to Professor Steve Evans and Professor Jim Pitman for providing references related to Hoeffding's covariance formula. Thanks are also due to a referee and the editor, Professor Wei-Yin Loh, for suggesting Lebesgue-Stieltjes integration to unify the proofs of (1) and (2) and for the prompt processing of this paper. This work was partially supported by the Mathematics Research Center (CRM) in Montreal.

References

[1] Billingsley, P.(1979) *Probability and Measure.* Wiley, New York. MR **80h:**60001

[2] Block, H. and Fang, Z.(1988) A multivariate extension of Hoeffding's Lemma. *Ann. Probab.* **16**, 1803-1821. MR **90a:**62133

[3] Hoeffding, W.(1940) Masstabinvariante Korrelationstheorie. *Schr. Math. Inst. Univ. Berlin* **5**, 181-223.

[4] Johnson, R.A. and Wichern, D.W.(1988) *Applied Multivariate Statistical Analysis.* 2nd ed., Prentice-Hall, New Jersey. MR **84a:**62086 (1st ed.)

[5] Pearson, K.(1926) On the coefficient of racial likeness. *Biometrika* **18**, 105-117.

[6] Pitman, J.(1996) Personal communication.

[7] Ross, S.(1984) *A First Course in Probability.* 2nd ed., Macmillan, New York. MR **85a:**60005

[8] Tukey, J. W.(1949) Comparing individual means in the analysis of variance. *Biometrics* **5**, 99-114. MR **11:**43c

[9] Yatracos, Y. G.(1997) Variance, clustering and identification of high leverage points. Tech. Report STT 97-1, Dept. of Mathematics and Statistics, Univ. of Montreal.

Département de mathématiques et de statistique, Université de Montréal, CP 6128, Suc. Centre Ville, Montréal, Canada H3C 3J7

E-mail address: yatracos@dms.umontreal.ca