INVARIANCE AND DESCRIPTIVE STATISTICS

Guido del Pino
Universidad Católica de Chile, Chile
delpino499@gmail.com

The main focus of teaching statistics at school is statistical literacy, which uses as little mathematics as possible. This fact together with an unappealing training emphasizing scattered definitions and computational recipes, makes mathematics teachers to strongly dislike statistics. Although the main issue is statistical literacy, providing an underlying mathematical structure for the statistical concept should help making the subject more interesting to mathematically trained people. It is shown that the mathematical concepts of invariance and equivariance under a family of transformations, including some concrete and intuitive interpretations, provide an insight on frequency distributions, graphical displays, and summary measures. The latter can then be rigorously defined, which makes it possible to construct new measures.

INTRODUCTION

It is known that mathematics teachers and their students tend to dislike statistics. One explanation for this behavior is that their training includes mainly scattered definitions and computational details. Another is that statistical literacy uses very little mathematics. To make statistics more attractive for mathematics teachers we provide an underlying mathematical structure for the different statistical concepts, which is based on invariance, maximal equivariant, and equivariance under some family of transformations. Though these concepts are normally treated in advanced courses of mathematics and statistics, we only deal here with elementary aspects and also provide intuitive interpretations. This structure gives a useful insight into frequency distributions, graphical displays, and summary measures.

We deal with the rectangular $n \times k$ data sets generated by n units and k variables, the nature of which is unrestricted, so that nominal, ordinal, discrete, and continuous variables are included. Thus the row associated with unit i can be identified with an array z_i of length k and the whole data set corresponds to the array $z = (z_1, \ldots, z_n)$. Introducing additional variables if necessary, it is possible to assign a non-informative label to each row, and these labels can be dropped if one so wishes. The relevant information is said to be *invariant under permutations* of the rows or *permutation invariant*. The intuitive idea of invariance is that if the data is transformed in a certain way, some statistics (which may include graphical displays) remain unchanged. A *maximal invariant* is a function of any any invariant statistic and represents the maximal possible reduction of the data using invariant arguments. Maximal invariants are not unique but they are equivalent in the sense that we one can get one from the other. An statistic is *equivariant* if it transforms in the same way as the data do.

A crucial result is that a frequency distribution is a maximal invariant under permutations, which leads to all usual ways of describing univariate and bivariate distributions, as well as their corresponding displays. To study summary measures we need to discuss invariance and equivariance under other families of transformations, like change of location, change of scale, and change of sign. To motivate the concepts we first recall some well known properties of the of the mean, of the standard deviations, and of the quantiles. We then turn around the argument and use some of these properties to rigorously define each type of measure.

METHOD

We start by applying permutation invariance to show that the vector of frequencies is a maximal invariant. For a fixed number of data values, the vector of frequencies is equivalent to the vector of relative frequencies, which is just the frequency distribution. We look for equivalent maximal invariants depending on the type of variable. For a continuous variable this is satisfied by the vector of order statistics and by the dot plot. For two categorical variables, contingency table is the maximal invariant. For two continuous variables it is shown that the scatter plot is a maximal invariant. When the variable X takes r values and Y is a continuous variable, finding a dot plot for Y for each of the r different values of x leads to parallel dot plots. To find rigorous definitions of

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org [© 2014 ISI/IASE]

the different types of summary measure we start by recalling the properties of the mean and the standard deviations, and frame them in the language of invariance of equivariance. The quantiles satisfy some properties of the mean, but not all. We use the properties that are satisfied to define a measure of position. The case of the quantiles is quite special since they satisfies *reflection equivariance* and a *time reversal property*. A measure of central tendency is then a particular measure of position satisfying reflection equivariance and a condition that essentially states thet if we compute the summary measure for *symmetric data* we get the center of the distribution. Finally we define a measure of dispersion as that that satisfies some invariance and equivariance conditions plus positivity. Finally, we provide three general methods that can be used to create measures of dispersion.

RESULTS

For a categorical variable the maximal invariant under permutations is the frequency array $n \ = (n_1, \dots, n_r)$, where n_j is the number of units associated category c_j. When n is known the n_i can be replaced by the *relative frequencies* $f_i = \frac{n_i}{n}$, and we call the corresponding vector (or function) the *frequency distribution*. The use of relative frequencies can be also justified when the statistc of interest does not change when each data value is replicated k times, where k is a positive integer. We say that this statistic is *invariant under replications*.

The same holds for ordinal, discrete, and continuous variables, although in general the values *are not known a priori*. From the point of view of the observed data, discrete and continuous variables behave alike, since the observed values belong to a finite set. A conceptual difference between continuous and discrete variables is that in the first all values should be different if written with enough decimal places and so there will be no ties. In this case $n_i = 1$, for all i, and by permutation invariance the only relevant information is the set of observed values. This implies the ordered data $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$. Since the set of these *order statistics* is equivalent to the frequency distribution, it is also a maximal invariant. The case of two variables X and Y can be analyzed similarly by using the artificial variable $Z = (X, Y)$; the joint frequency distribution X and Y is then a maximal invariant. When X and Y have r and s possible values the contingency table is also maximal invariant. When n is fixed the absolute frequencies can be replaced by relative frequencies without any loss of information. When the frequency of each x is fixed by design, the collection of *conditional frequency distributions* of Y for given $X = x$ is then a maximal invariant. When X and Y are both continuous ties seldom arise and the set of n vectors (x_i, y_i), or the corresponding scatter plot are maximal invariants. Assume that X is categorical and that we want to compare the distribution of a continuous variable Y between the different categories x. In this case the collection of conditional distributions or the parallel dot plot are maximal invariants.

MEASURES OF POSITION AND CENTRAL TENDENCY

We start by reviewing the elementary properties of the mean. It is easily checked that the mean is (a) invariant under permutations and (b) invariant under replications, so that it is determined by the relative frequencies. For a given function $g, y_i = g(x_i)$ induces a transformation of x into y . Then the mean m is (c) *location equivariant:* $y_i = x_i + c$, implies $m(y) = m(x) + c$; (d) *scale equivariant.* $y_i = dx_i, d > 0$ implies $m(y) = dm(y)$; (e) *reflection equivariant:* $y_i = -x_i$ implies $m(y) = -m(x)$. Furthermore, (c, d) is equivalent to (f) *invariance under changes of location and scale:* $y_i = a + bx_i, b > 0$ implies $m(y) = a + bm(x)$, while (c,d,e) implies (g) *invariance under affine linear transformations* $y_i = a + bx_i$ implies $m(y) = a + bm(x)$, where b is allowed to be negative.

A summary measure is a *measure of position* if (a)-(d) hold; a measure of position is a measure of central tendency when (e), as well as a symmetry condition (g) to be defined below holds. According to this definition, the mean, the median, the minimum, the maximum, and the p-quantiles can be checked to be measures of position. The p-quantiles must be defined separately according to np being an integer or not.

The data vector $z \ = (z_1, \dots, z_n)$ is said to be symmetric about 0 if $z_{(n-i+1)} = z_{(i)}$ for $i = 1, \dots, n;$ x is said to be symmetric about c if the vector y generated by $y_i = x_i - c$ is symmetric about 0. c is called the center of the distribution, A measure of position S is said to be of central tendency if (h) For x symmetric about c, $S(\,x\,) = c$, but when n is even the middle observation must be excluded before computing the center. The mean, the median, and the midrange are measures of central tendency. A more general family is

$$S_n = \sum_{i=1}^{n} c_i X_{(i)} \; where \; c_i \; = \; c_{n+1-i}, for \; all \; i \; and \; \sum_{i=1}^{n} \; c_i = 1.$$

Some additional properties of the quantiles are:

- *Equivariance under continuous strictly increasing transformations* g. If $y_i = g(x_i)$, $S(\,y\,) = g(S(\,x\,))$. Location and scale equivariance is a particular case.
- *Time reversal property*. Let r be an integer: (i) If $r - 1 < np < r$, the p-quantile is the $r - th$ observation counting from the first and the $(1 - p)$-quantile is the $r - th$ observation counting from the last; (ii) If $r = np$, the p-quantile is the average of the $r - th$ and the $(r + 1) - th$ observation counting from the first, while the $(1 - p)$-quantile is the average of the $r - th$ and the $(r + 1) - th$ observation counting from the last.

MEASURES OF DISPERSION.

The standard deviation is the most common "measure of dispersion". As in the mean (a)and (b) hold. Furthermore, it is (1) positive; (2) location invariant; (3) scale equivariant; and (4) reflection invariant. We formally define a measure of dispersion as a summary measure satisfying (a)-(b) and (1)-(4). Three methods to construct dispersion measures are given below. The first two use deviations $d_i = |x_i - t|, i = 1, \dots, n$ from a measure of central tendency t.

Method I For a given measure of central tendency, say s, compute $s(d_1, d_2, \dots, d_n) > 0$ where s is a measure of central tendency (which may be eventually coincide with t.) Using the mean and the median as the measures of central tendency involved, we get four combinations. When both t and s correspond to the mean we get MAD.

Method II Choose an invertible function g from IR into IR, compute $w_i = g(x_i)$, and its average w. then the resulting measure of dispersion is $h(w)$, where h is the inverse of g. It can be proved that scale invariance implies that, except for a trivial constant, the only admissible transformations are $g(z) = z^r$ with positive r and $g(z) = \log z$. The arithmetic mean, the quadratic mean and the harmonic mean Correspond to $r = 1,2,-1$ respectively. The geometric mean corresponds to the logarithmic transformation. For $r = 2$ we get the standard deviation.

Method III Linear combinations of order statistics.

$$S_n = \sum_{i=1}^{n} c_i X_{(i)} \; where \; c_i \; = \; -c_{n+1-i}, for \; all \; i, and \; \sum_{i=1}^{n} \; c_i = 0.1.$$

Note that this measures are clearly location invariant and scale equivariant. If the measure is negative one needs only change the signs of the coefficients. Some examples are the range and the interquartile range. Note that $Q_2 - Q_1$ satisfies all properties, but the symmetry condition

ACTIVITIES

- *Activity 1* The teacher asks each student choose a color and constructs a tally diagram. Change the order of the choices and check that the tally diagram does not change. For very young students ask them to paint a block with his favorite color. Rearrange the blocks at random and check that the tally diagram remains unchanged.use painting of blocks instead.
- *Activity 2* The teacher asks each student separately to complete a survey, where each question must be answered with the choice: 1.- disagrees 2.- not sure 3.- agrees. After that process the data, find frequency distributions, and construct bar charts.
- *Activity 3:* (a) Given the raw data construct a 2×2 contingency table and the marginal distributions. (b) For these fixed marginal distributions make up several contingency tables satisfying them. (c) Make up questions that can or cannot be answered by the marginal distributions, but can be answered from the contingency table.
- *Activity 4:* For the ordered data 1, 3, 5,7, 9, 11,13, 15,17, 19 compute the p-quantiles for $p = 0.20, 0.25, 0.75$, and 0.80. Verify the time reversal property.
- *Activity 5:* Verify directly that $\frac{Q_1+Q_3}{2}$ and $\frac{Q_1+2Q_2+Q_3}{4}$ are measures of central tendency.

- *Activity 6:* Show directly that the quadratic mean and the harmonic means of the absolute deviations are measures of dispersion. Try this with small data sets.
- *Activity 7:* Verify directly that $Q_3 - Q_1$ and $Q_1 + Q_3 - 2Q_2$ are measures of dispersion

REFERENCES

Casella, G. C. (2001). *Statistical Inference* (2nd ed.). Cengage Learning.

Fraleigh, J. (1976). *A First Course In Abstract Algebra* (2nd ed.). Reading, MA: Addison-Wesley.

Herstein, I. N. (1964). *Topics In Algebra*. Waltham: Blaisdell Publishing Company.

Kay, D. C. (1969). *College Geometry*. New York: Holt, Rinehart and Winston.

Lehmann, E., & Casella, G. (1998). *Theory of Point Estimation* (2nd ed.). New York: Springer.

Lehmann, E. L., & Romano, J. P. (2008). *Testing Statistical Hypotheses* (3rd ed.). New York: Springer.

McCoy, N. H. (1968). *Introduction To Modern Algebra.* (Rev. ed.). Boston: Allyn and Bacon.