Is It One or Two Groups of Data?

How to Determine It

BY DEAN V. NEUBAUER

Q: When I look at my data it appears that instead of one group of data I see two. How can I determine statistically if there are one or two groups of data present when I expect to see only one?

A: It is not uncommon for researchers to look at their data and see something they didn't expect. Often the person is expecting to see two sets of measurements look statistically similar when selected from what is believed to be a single population (measurement system) with mean μ and variance σ^2. Typically, one can test the difference between two sample means, say $\bar{x}_1$ and $\bar{x}_2$, with sample variances, s_1^2 and s_2^2, based on sample sizes, n_1 and n_2, respectively, using a student's t statistic as described in the basic statistics standard ASTM E2586, Practice for Calculating and Using Basic Statistics. However, in this case, we are interested in whether the two samples actually represent clusters of data from two different populations.

A set of observations $x_1, x_2, \ldots, x_n$ can be partitioned into two clusters $x_{11}, x_{12}, \ldots, x_{1n_1}$ and $x_{21}, x_{22}, \ldots, x_{2n_2}$. Fortunately, we don't need to consider all 2^n possible clustering of the data. If the observations are ordered, then we only need to consider $(n - 1)$ partitions of the data.

$\{x_1\},\{x_2, L, x_n\}$
$\{x_1, x_2\},\{x_3, L, x_n\}$
etc.
$\{x_1, L, x_{n-1}\}, \{x_n\}$

For the two clusters, Engelman and Hartigan[1] define a statistic between whose maximum value represents the maximum distance between the optimal clusters (partitions). This maximum value is denoted by C. If C exceeds a critical value, then it indicates that the clusters do represent two populations with means, μ_1 and μ_2, respectively, with the same variance, σ^2. The formula looks like this:

$$C = = \frac{n_1 n_2 \left(\bar{x}_1 - \bar{x}_2\right)^2}{\left\{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2\right\}(n_1 + n_2)}$$

The value of C measures the distance between clusters and tests for the following:

- Null hypothesis: $x_1, x_2, \ldots, x_n$ is a random sample from a single population with mean μ and variance σ^2, against the
- Alternate hypothesis: For some partition of $x_1, x_2, \ldots, x_n$ the cluster $x_{11}, x_{12}, \ldots, x_{1n_1}$ is a sample from a population with mean, μ_1, and variance, σ^2; and the cluster $x_{21}, x_{22}, \ldots, x_{2n_2}$ is a sample from a population with mean, μ_2, and same variance, σ^2.

We will let the critical value C_α be such that $P(C < C_\alpha) = \alpha$ under the null hypothesis. So, for a set of data, we will compute the value of C based on the sample statistics for both samples and then compare C to C_α. Table 1 presents the critical values for C_α for testing the most extreme grouping possible, so any value of C that exceeds C_α is solid evidence of clustering.

EXAMPLE

Suppose we have a dataset of eight lots that were tested for some characteristic and yielded the values 102, 95, 75, 201, 67, 194, 81 and 187. So, can these lots can be divided into two groups, e.g., did the lots come from a single process or not? Ordering the values and looking at the possible $(n - 1)$ partitions give

{67},{75,81,95,102,187,194,201}
{67,75},{81,95,102,187,194,201}
{67,75,81},{95,102,187,194,201}
{67,75,81,95},{102,187,194,201}
{67,75,81,95,102},{187,194,201}
{67,75,81,95,102,187},{194,201}

{67,75,81,95,102,187,194},{201}

The only reasonable possibility for two groups is the partition {67,75,81,95,102}, {187,194,201}. Using the equation for C we have

$$C = \frac{(5)(3)(84-194)^2}{\{(5-1)(206)+(3-1)(49)\}(5+3)} = 24.61$$

In this example, we have a total of $n = 8$ values so we choose C_α from Table 1 for an appropriate value of α. Here we see that $C = 24.61$ is more significant (larger) than the critical value at $\alpha = 0.01$ ($C_{0.01} = 15.1$), so we can safely conclude that the lots came from two different groups (processes) with at least 99 percent confidence. In this case, the means of the two groups are 84 and 194, respectively, and their common standard deviation is estimated to be

$$s = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} = \sqrt{\frac{(5-1)(206)+(3-1)(49)}{5+3-2}} = 12.4$$

with 6 degrees of freedom.

Table 1 – Critical Values for C for Testing for Two Clusters

n	α=.10	α=.05	α=.01
5	15.10	24.00	74.10
6	9.84	14.10	33.10
7	7.66	10.50	20.90
8	6.46	8.39	15.10
9	5.68	7.18	11.70
10	5.14	6.34	9.89
11	4.75	5.77	8.66
12	4.45	5.34	7.78
13	4.21	5.00	7.11
14	4.01	4.73	6.59
15	3.85	4.51	6.15
16	3.71	4.31	5.82
17	3.59	4.15	5.53
18	3.49	4.01	5.29
19	3.40	3.89	5.08
20	3.32	3.78	4.91
21	3.25	3.69	4.74
22	3.19	3.61	4.59
23	3.13	3.53	4.47
24	3.08	3.46	4.35
25	3.03	3.40	4.25
30	2.84	3.16	3.86
35	2.71	2.99	3.59
40	2.62	2.86	3.39
45	2.54	2.77	3.24
50	2.48	2.69	3.12

REFERENCE

1. Engelman, L., and Hartigan, J. A., "Percentage Points of a Test for Clusters," *Journal of the American Statistical Association*, Vol. 64, No. 328, Dec. 1969, pp. 1647-1648.

DEAN V. NEUBAUER, *Corning Inc., Corning, N.Y., is an ASTM fellow; he serves as chairman of Committee E11 on Quality and Statistics, chairman of E11.90.03 on Publications and coordinator of the DataPoints column.*

snonline

Get more tips for ASTM standards development at www.astm.org/sn-tips.

Find other DataPoints articles at www.astm.org/standardization-news/datapoints.