Improving Quality and Productivity via Stratification: A Call Center Example for Forming Homogeneous Employee Groups

STEVE HILLMER AND CANAN KOCABASOGLU
UNIVERSITY OF KANSAS

In many organizations when multiple employees are doing similar jobs, oftentimes these employees are not performing at the same level. In these cases, there is an opportunity to identify the root causes for some employees' less-than-optimal performance and to initiate changes leading to improvements in these employees' execution of their responsibilities. One requirement to implement this strategy is to form homogeneous performance groups. In this study, the authors develop and introduce a simple statistical method to place employees performing similar tasks into homogeneous groups. The method is illustrated using data from a call center. The example illustrates the differences in performance that typically exist in organizations and is used to discuss improvement strategies.

Key words: performance improvement, stratification

INTRODUCTION

Discrepancies in employee performance are a challenge to companies in service and manufacturing industries alike. Considering the negative consequences of these discrepancies on quality and productivity in the short term and company performance over the long term, the challenges resulting from the disparity of performances are significant for any business.

There have been many inquiries into what individual traits and organizational attributes influence employee performance. Some of the individual factors considered are age, level of education, company tenure, and job satisfaction (for example, Judge et al. 2001; Tsui et al. 1997). In terms of organizational factors, technology (Papa and Tracy 1988), human resource practices (Huselid 1995), and incentive systems (Baker, Jensen, and Kevin 1988) are just a few of the drivers considered in past studies. While these inquiries have helped managers understand the challenges they generally need to be cognizant of in their efforts to improve employee performance, they do not necessarily help identify the specific issues that cause differences among employees of the same firm. One way to pinpoint these issues is to group employees according to their performance and compare these groups to understand the factors that cause the performance differential. There are few studies that take this approach. Moreover, even for research that has taken this perspective, the prevalent technique is to rank-order employees according to their performance and then group them as high versus low performers. This method has some

inherent limitations that can be avoided through the use of alternative statistical methods.

The purpose of this article is to develop and introduce a method that can be used as part of a stratification approach. Stratification involves identifying one or more relevant measures of employee performance and using these measures to classify employees into homogeneous groups (Kume 1985; and Kane 1989). The stratified groups have the property that while the performance of each employee within a group is approximately the same, there are considerable differences in performance between the groups. If the causes of the differential performance between groups can be identified and eliminated, then it is possible to increase the performance of the poorer performing groups to the level of the best performing group. While in practice it may not be possible to improve the performance of all employees to that of the best performing group, it is realistic to expect that several of the causes of the poorer performance can be identified and eliminated so that many employees' performance can be improved. The method discussed in this article significantly contributes to the first step in the implementation of a stratification strategy by introducing a process to establish homogeneous employee groups with respect to their performance.

The stratification strategy has the advantage that the best performing group can serve as a benchmark for the other groups. Once the employees are clustered according to their performance, managers can study the factors that can potentially lead to these differences. These factors may include but are not limited to work conditions, work methods, and training. With the appropriate changes, the quality of the work and the productivity of the organization can often be permanently increased without having to incur permanent increases in costs. In addition, the cost of these process changes is usually minimal compared to the quality and productivity gains.

STRATIFICATION OF AGENTS IN A CALL CENTER: AN ILLUSTRATIVE EXAMPLE

The method developed for this article is illustrated in a call center setting for the following reasons: 1) call centers have become an important part of many organizations and have established themselves as fundamental in sales and marketing and in providing customer support; 2) call centers can employ a significant portion of the workforce in some countries; and 3) most call centers have a history of routinely collecting a wealth of data on employee performance; thus, the method explained in this article can be easily applied in call centers.

While the abundance of data collected by call centers contains valuable information about how to improve the center's performance, the sheer amount of data also brings the challenge of identifying and interpreting the critical information in the data. For example, in a typical call center as many as 50 measurements each day may be collected for each rep. The specific data collected and monitored depend upon the particular call center; however, typical measurements include: the daily close rate, that is, the number of sales for the day divided by the number of calls taken for the day; the average daily sale amount for the customers who made a purchase; the average talk time per call each day; and the after-call work time, that is, the average time spent after the call is finished to record relevant customer information.

The following discussion is based on data collected on 51 customer service representatives. The data were provided by a call center in the Midwest. The data contained several performance measures for the call center representatives and were collected by the company over a one-year period.

Through conversations with management, it was determined that one of the key performance indicators for these representatives was the time spent on documentation after completing the call. This is referred to as the after-call work (ACW) time. The call center's data processing system routinely records the daily total time, in seconds, for each representative's ACW. This performance measure is used in the remainder of this section to illustrate how to use the method developed in this article to stratify the 51 representatives into groups exhibiting homogeneous performance. Each step in the methodology will be explained throughout. Readers who are interested in the technical aspects of the statistical method explained here should refer to the Appendix for an in-depth discussion of the method.

Step 1: Removing Outliers or Special Causes

When using this type of data, it is expected that there will be a few occasions where unusual circumstances influence daily measurements. Deming (1986) called these circumstances special causes. The measurements that have been influenced by *special causes* are not representative of an agent's overall performance; thus, they should be identified and eliminated from subsequent analysis. Accordingly, the first step in the analysis of the authors' example was to examine each representative's ACW measures to determine the outliers in the data. The tool that was used to identify these outliers was a control chart for individuals (Wheeler 1993). Thus, there was one control chart for each representative where all of his or her ACW measurements were plotted.

Figure 1 is a control chart for individual observations of the performance of one representative in the sample. Based upon the control chart, the 56th ACW measurement is indicative of a special cause because this point is outside the upper control limit. Since the observation at time 56 was much larger than the typical ACW, this point was removed so it did not distort the subsequent analysis. The control chart in Figure 2 shows that once the 56th observation was removed and the control limits were recalculated after the change, there were no other outliers. Control charts for all the representatives in the call center were developed in a similar manner and used to eliminate the outliers before proceeding to the next step.

Step 2: Stratify Agents by Long-Run Variances

The second step in the method proposed in this study was to group agents who exhibited homogeneous performance during the study period. One way to group the representatives is according to the mean time and variance of the ACW by using a sample of the ACW measures. To estimate a representative's mean and variance, a set of the most recent 40 observations from that representative's ACW measurements was selected. The representatives were first grouped according to

Figure 1 I Chart for Representative 1.

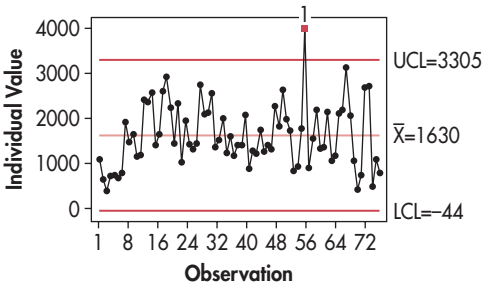
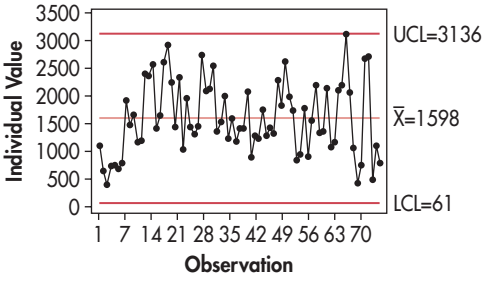


Figure 2 Revised I Chart for Representative 1.



the variance of the ACW, because the procedure for grouping by means assumes equal variances.

The method initially assumes that the variances of all the agents are equal. If this assumption is true, it is possible in the authors' example to estimate the common variance by computing the sample variance for each representative and then averaging the 51 sample variances to obtain the estimated pooled variance, $S_p^2 = 1{,}235{,}505$. Since the estimated S_p^2 is based on the results of 51 samples each of size 40, it is an accurate estimate of the assumed common variance. Next, a hypothesis testing procedure is used to determine the number of variance groups. This procedure begins with the assumption that all the representatives have the same variance equal to $S_p^2 = 1{,}235{,}505$. If this assumption is correct, then, as is explained in the Appendix, the ratio of the sample variance for any representative to the estimated pooled variance should fall between .366 and 2.069. If this ratio for a representative is less than .366 then it can be concluded that this representative's long-run variance is less than 1,235,505. If this ratio for a representative is larger than 2.069 then that representative's long-run variance is larger than 1,235,505. Both of these would

Table 1 Results of final cluster analysis based on the standard deviation.

Group	Number of agents	Estimate of pooled variance
Very small variance	3	177,062
Small variance	15	546,937
Middle variance	19	1,136,472
Large variance	14	2,334,468

indicate that there is more than one group of representatives. Once the number of groups is determined, a clustering algorithm, described in the Appendix, is used to assign each representative to an appropriate group. After two iterations of this procedure, four homogeneous groups of representatives with equal variances within each group were identified. The groupings for the 51 representatives are given in Table 1. Once the representatives have been grouped according to their variances, the next step is to form subgroups of representatives that have the same long-run mean.

Step 3: Stratify Agents by Long-Run Means

The next step in the authors' method was to take each group of agents with equal variances and create subgroups, where agents in each subgroup had the same long-run mean. For the call center example, this procedure is described using the second variance group in Table 1. The method to form the subgroups with equal long-run means is based upon constructing confidence intervals for the differences in the means between all possible pairs of representatives. Since there were 15 representatives, the method required the estimation of a total of $\binom{15}{2} = 105$ confidence intervals for the difference between all possible representatives. The method maintains an overall simultaneous confidence level of 95 percent for the 105 confidence intervals. Since there were 15 representatives in this group, the common standard deviation for this group was 739.6 and the sample size was 40. It follows that if the difference in the sample means between any two representatives is

Table 2 Sample means for the small variance group.

Agent	Sample mean
Agent 38	675.0
Agent 39	1144.0
Agent 28	1241.1
Agent 24	1344.0
Agent 30	1425.0
Agent 1	1631.0
Agent 13	1744.0
Agent 47	2142.0
Agent 9	2283.0
Agent 32	2327.0
Agent 2	2523.0
Agent 27	2991.0
Agent 26	3451.0
Agent 33	4264.0
Agent 35	7647.0

larger than 637.3, then these representatives have different long-run means. The Appendix provides details on how the value 637.3 was derived. Thus, the basic idea was to form as many subgroups as necessary so that the difference in the sample means reported in Table 2 within each subgroup was less than 637.3.

In view of this, the first step was to determine the number of subgroups needed. As seen in Table 2, the 15 representatives in the group under consideration were sorted in ascending order according to their sample mean. Considering the values in Table 2, the largest sample mean minus the smallest sample mean divided by the highest difference in means allowed within a group was $(7{,}647 - 675)/637.3 = 10.94$. This suggested that it would require 11 subgroups to ensure that, within each group, the distance between any two sample means was smaller than 637.3. A closer look at Table 2, however, revealed that the difference between the sample mean for representative 35 and that of representative 33 was $7647 - 4264 = 3{,}383$, which is several times larger than

Table 3 Groups based on means for small variance group.

Group number	Agents in group	Centroid of group
Group 1	A38, A39, A28	1020
Group 2	A24, A30, A1, A13	1536
Group 3	A47, A9, A32, A2	2319
Group 4	A27	2991
Group 5	A26	3451
Group 6	A33	4264
Group 7	A35	7647

the highest difference in means allowed in a subgroup. This suggested that it was not necessary to have 11 subgroups since it was clear that representative 35 would form one group with a much larger mean than the other representatives. The number of groups required, beginning with representative 38 and ending with representative 33, was estimated to be (4,264 – 675)/637.3 = 5.63. Thus, six groups were required for the first 14 representatives, and representative 35 was a seventh group. Once it was determined that seven groups were required, a clustering procedure was used to put the representatives into groups. Each representative was placed in the group for which his or her sample mean was closest to the average of the sample means for the representatives in the group. Table 3 reports the final subgroupings and average of the sample means for the representatives in the second variance group. It can be verified, for each subgroup, that the difference in each pair of sample means is less than the 637.3.

A similar analysis was done to determine subgroups with the same long-run variance and the same long-run mean for each of the four variance groupings. These groups are reported in Table 4.

Table 4 Final groupings of all agents.

Variance grouping			
Group 1	**Std. deviation = 420.8**		
	Mean subgrouping	**Subgroup mean**	**No. of agents**
	Subgroup 1	583	2
	Subgroup 2	1014	1
Group 2	**Std. deviation = 739.6**		
	Mean subgrouping	**Subgroup mean**	**No. of agents**
	Subgroup 1	1020	3
	Subgroup 2	1536	4
	Subgroup 3	2319	4
	Subgroup 4	2819	1
	Subgroup 5	3451	1
	Subgroup 6	4264	1
	Subgroup 7	7647	1
Group 3	**Std. deviation = 1066.1**		
	Mean subgrouping	**Subgroup mean**	**No. of agents**
	Subgroup 1	2225	4
	Subgroup 2	3226	7
	Subgroup 3	3989	5
	Subgroup 4	5282	3
Group 4	**Std. deviation = 1527.9**		
	Mean subgrouping	**Subgroup mean**	**No. of agents**
	Subgroup 1	2292	2
	Subgroup 2	4198	5
	Subgroup 3	4780	1
	Subgroup 4	5973	5
	Subgroup 5	8954	1

Follow-Up to Determine the Root Causes of the Differences

The value of establishing these groups is that they can be used to identify the causes of the differential performance among agents and to develop improvement strategies. It is usually easier to identify and remove the causes of differences in mean performance levels; therefore, a good way to target the initial efforts is to

concentrate on the agents within the subgroupings that have the same variance. For example, consider the group of representatives whose long-run standard deviation is 739.6 in Table 4. There are 15 representatives in this group that are divided into seven subgroups based upon the representatives' long-run mean ACW. ACW involves recording relevant information about the call into the call center database so that this information is available in case some form of follow-up is required. Thus, representatives should record all relevant information, yet at the same time be efficient. This means that the desired ACW should be neither too short nor too long. Assume that for the purposes of this discussion that the subgroup with the average of 2319 seconds per day (subgroup 3) has the most desirable performance and, therefore, is selected as the benchmark.

The improvement strategy involves comparing the working styles and working conditions of representatives in the benchmark group to those of representatives in groups with a significantly shorter or longer average ACW. The first subgroup in Table 2 contains representatives with a significantly shorter ACW than that of the representatives in the benchmark group. The causes for such a difference in performance will vary depending upon the particular call center. One possibility is that representatives with the lower-than-desired average ACW are spending less time on ACW because they are not recording all of the required information about their calls. This pattern of behavior will be costly to the call center in the long run, because of the rework required to obtain the information that should have been recorded originally. A second possibility is that representatives with a lower-than-desired average ACW are recording the information required *during* the call. While this reduces the time required for the ACW, this behavior increases the risk that these representatives are not giving the customer adequate attention during the call. If the result is that the customer's request is misunderstood, then the customer will be dissatisfied and will probably need to call a second time, or even worse, will consider taking his or her business to the competition. Thus, the net result of keeping the ACW short will be an increase in rework or loss of goodwill or even loss of business, all of which are costly.

A similar inquiry should be carried out for representatives who are in groups with a significantly longer mean ACW than that of the benchmark. According to Table 2, these are the agents in subgroups 5, 6, and 7. One possible reason for a longer-than-desired ACW is that representatives are recording unnecessary information. This behavior increases labor costs. A second possibility is that agents in the benchmark group have learned ways to efficiently record the information that is required, while representatives in groups with larger long-run means are not as efficient. Such a discrepancy in work methods would suggest a need in training.

After determining and eliminating the causes of differences in long-run average ACW, comparisons of the groups with different long-run variances can be made. As in the case with means, statistically significant differences in sample variances are primarily caused by the different working styles or working conditions of different groups. For example, one possible cause of representatives with a large variance in ACW is that the data collection process is unstructured. As a result, the amount of time spent recording information on a call can show considerable variation. Once this is understood, improvements in data collection can streamline the recording process. A second possibility is that representatives with a very small long-run variance record nearly the same information for every call. This behavior is problematic because it implies that some unique aspects of the call are not recorded. This will lead to rework if the information is needed in the future.

The examples discussed in this section are not exhaustive, but focus on some prevailing causes of employee performance discrepancies in a call center setting. Close observation and comparison of the homogeneous groups can reveal a variety of issues not considered here.

Advantages of Statistical Grouping versus Ranking

A common way to evaluate the performance of employees doing similar jobs is by ranking them based upon a chosen performance measure (for example, Kalra and

Mengze 2001). Thus, in the call center case, representatives might be ranked according to how close their average ACW time for a given period is to the desired level. For example, if the desired daily ACW in seconds was 2300, then each agent might be ranked according to the distance of their sample mean from 2300, with those closer ranked as better performing. These rankings could be used to form rough groupings like those proposed in this article and be used in an attempt to find the causes of the differences in performance reflected in the ranks. Yet, for the reasons explained later, the method proposed in this article has several advantages over the more informal ranking-based method.

First, the statistical method in this article takes into consideration the possible differences in variance for different agents, while a method based upon ranks implicitly assumes that the variance in performance for all agents is equal. The implications of such an assumption are twofold: First, large differences in variances can have a substantial impact upon the appropriateness of statistical methods used to do the grouping (Ramsey and Schafer 2002). In the call center example for which the authors used real data, there were substantial differences in the variances of the performances of the representatives and, therefore, any method that ignores these substantial differences in variance would be inferior to the method suggested in this article. Also, while the traditional focus has been on the differences of sample means, an additional goal of process improvement should be to improve upon or eliminate the root causes of excessive variance in agents' performance, which implies the additional need to group representatives based upon differences in their sample variances.

Second, the statistical method builds in a procedure to find and eliminate data points that are impacted by special causes. These points should be removed from the analysis because they will seriously distort the results and can lead managers to reach erroneous conclusions. While the removal of the outliers is not unique to the method offered in this article, it is usually not considered when ranking agents.

Third, the statistical method developed in this article forms the groups by taking into account the inherent common cause variation in performance, whereas a ranking approach does not. While common cause variation leads agents to have different sample means and sample variances (Deming 1986; Joiner 1994), the differences are not caused by factors that can be easily identified. In other words, managers' efforts are usually futile when they attempt to find the reasons for performance differences due to common causes. This can be explained as follows: A critical part of any process improvement strategy adopted by a manager is to identify and remove the root causes of differences in performance by comparing the working styles and working conditions of two agents. Yet if the difference in performance between two agents is only caused by random variation, then there is no identifiable single root cause for that difference. In contrast, when the difference in performances of separate agents is so large as to be statistically significant, past practice suggests that there is a good chance the root cause of the difference in performance can be isolated. The advantage of the statistical method proposed here is that it ensures that agents will be compared only when the chance of finding the cause of the difference is very likely.

CONCLUSIONS

This article has outlined and illustrated a general approach that can be used to increase the consistency among employee performance. This approach is based upon the use of measurements that can be routinely gathered in operations. The method accounts for inherent common cause variation in employee performance and uses homogeneous employee groups to identify causes of differential performance. This method was demonstrated using a case study of customer service representatives in a call center. The methods in this article can be used to identify causes of less-than-optimal performance, which, in turn, can lead to action to improve the performance of employees and thereby improve the overall performance of the call center. While the specific call center application is important, the methods developed in this article can also be applied in other situations where multiple employees perform the same job tasks.

REFERENCES

Baker, G. P., M. C. Jensen, and J. M. Kevin. 1988. Compensation and incentives: practice vs. theory. *The Journal of Finance* 43, no. 3: 593-616.

Deming, W. E. 1975. On probability as a basis for action. *The American Statistician* 29: 146-152.

Deming, W. E. 1986. *Out of the Crisis*. Cambridge, Mass.: MIT Center for Advanced Engineering Study.

Huselid, M. A. 1995. The impact of human resource management practices on turnover, productivity, and corporate financial performance. *The Academy of Management Journal* 38, no. 3: 635-672.

Joiner, B. 1994. *Fourth generation management*. New York: McGraw-Hill.

Judge, T. A., C. J. Thoresen, J. E. Bono, and G. K. Patton. 2001. The job satisfaction-job performance relationship: A qualitative and quantitative review. *Psychological Bulletin* 127, no. 3: 376-407.

Kane, V. E. 1989. *Defect prevention*. New York: Marcel Dekker.

Kalra, A., and S. Mengze. 2001. Designing optimal sales contests: A theoretical perspective. *Marketing Science* 20, no. 2: 170-193.

Kume, H. 1985. *Statistical methods for quality improvement*. Tokyo, Japan: AOTS Chosakai Ltd.

Larson, R. G., and L. Marx. 2001. *An introduction to mathematical statistics and its applications*. Upper Saddle River, N.J.: Prentice Hall.

MacQueen, J. B. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley symposium on mathematical statistics and probability 1*. Berkeley, Calif., University of California Press: 281-297.

Miller, R. G. 1966. *Simultaneous statistical inference*. New York: McGraw Hill.

Papa, M. J., and K. Tracy. 1988. Communicative indices of employee performance with new technology. *Communication Research* 15, no. 5: 524-544.

Ramsey, F. L., and D. W. Schafer. 2002. *The statistical sleuth*. Pacific Grove, Calif.: Duxbury.

Tsui, A. S., J. L. Pearce, L. W. Porter, and A. M. Tripoli. 1997. Alternative approaches to the employee-organization relationship: Does investment in employees pay off? *The Academy of Management Journal* 40, no. 5: 1089-1121.

Wheeler, D. J. 1993. *Understanding variation*. Knoxville, Tenn.: SPC Press.

BIOGRAPHIES

Steve Hillmer is a professor for the School of Business at the University of Kansas. He has a doctorate from the University of Wisconsin-Madison. His research interests immediately after finishing his doctorate were in the area of time-series analysis. He spent the 1979-1980 academic year at the United States Census Bureau modeling U.S. economic times series and in modifying the seasonal adjustment methods that were first developed in his dissertation. Most of his research over the next several years was a result of his experiences in the analysis of data at the Census Bureau. In the late 1980s he became interested in the role statistics played in quality management and his intellectual interests shifted. He has published several papers on the theoretical foundations of W. E. Deming's principles of management. His current research interest related to statistics is involved in the use of statistics in applied problems. He can be reached by e-mail at hillmer@ku.edu.

Canan Kocabasoglu is an assistant professor in decision sciences for the School of Business at the University of Kansas. She holds a doctorate in operations management from University at Buffalo, State University of New York. Her research interests are in collaboration in forward and reverse supply chains, competitive priorities, purchasing practices, buyer-seller relationships and the implications of information technology on operations management. She has published in *Journal of Operations Management, OMEGA—The International Journal of Management Science,* and *Journal of Supply Chain Management.*

APPENDIX
Statistical Method to Group Employees Based Upon Sample Data

The statistical procedure explained in this Appendix was implemented in the Minitab statistical program to perform the case study computations. The authors will provide interested readers a Minitab macro to perform the methods in the Appendix upon request.

Assume, for the purposes of this article, that there is one measure that captures an important part of agents' performance in their jobs and that this measure is taken each day the agent works. Let $X_{k,i}$ denote the value for agent k on day i. Employee performance data will sometimes be affected by one-time unusual events (special causes) that have a large impact upon the measurement $X_{k,i}$. Data affected by special causes are not representative of an agent's long-run performance; thus, these data should be identified and eliminated from subsequent analysis (Joiner 1994). Control charts can be used to identify the timing of special causes for each agent (Wheeler 1993). After special causes have been removed from the performance data for each agent, the next task is to form homogeneous groups of

agents based on the performance measurements. Both the long-run mean and variance of the performance should be the same within each group. The agents' sample means and sample variances will be used to form the groups. Of course, even if two agents' performance comes from the same theoretical distribution, their sample means and variances will not be the same due to random variation. Therefore, any method used in forming homogeneous groups must take into consideration the impact of random variation.

In forming the groups, one can make the error of putting agents with the same performance in different groups or putting two or more agents into the same group when they have different long-run means or variances. These two errors are analogous to a type I error and type II error in a traditional hypothesis testing scenario (Larson and Marx 2001). The stratification strategy will only be useful if the agents in different groups have an easily identifiable cause for the differential performance. Thus, when forming the groups it is more problematic to make a type I error, since this mistake may lead to a search for a cause in differential performance when in truth there is no such cause. On the other hand, if a type II error is made, the only problem is that some agents with slightly different performance will be put in the same group so that the cause of the actual differential performance will not be considered. Thus, it is desirable to make the probability of a type I error—the significance level—small. The methods suggested in this article will involve multiple comparisons. It is important to maintain an overall small level of significance for the entire set of comparisons, since this will help ensure that the chance of a type I error is small.

When comparing the performance of two different agents, there is at least a slight difference in their long-run performance. For instance, the theoretical long-run means of any two agents are always at least slightly different. Suppose there are two agents with only a small performance difference. This small difference can be found using traditional hypothesis testing methods by taking the sample sizes that are sufficiently large (Deming 1975). Since, in practice, very large samples are available in call center data, in theory it is possible to find statistically significant differences between any two agents. The practical significance of the discovery of a very small difference is another matter. When using a stratification strategy, a very small difference in the performance of two agents will have no practical significance because it will be difficult or impossible to isolate the underlying causes of the difference. Thus, to ensure that the statistically significant differences are also large enough to be of practical significance, the sample sizes must be kept at a moderate value. In addition, equal sample sizes for all agents will be chosen to simplify the analysis.

Procedure for Grouping by Variances

The first step in forming homogeneous groups is to divide the agents into groups that have the same variance within each group. The procedure makes the initial assumption that all the agents have the same long-run variance so that data from all agents can be pooled to estimate the common variance (Larson and Marx 2001). The data from each agent are then used to test the hypothesis that the agent's variance is equal to the common pooled variance estimate. If this hypothesis cannot be rejected for all the agents, it is concluded that all the agents have the same variance; otherwise, there are at least two groups. In the latter case, the number of variance groups is determined and the clustering procedure described later is used to form the groups. This procedure may be repeated several times until the hypothesis that all agents in a group have the same variance cannot be rejected. The steps to form groups of equal long-run variances follow. The hypothesis testing procedure is based upon the standard generalized likelihood ratio methodology (Larson and Marx 2001).

1. Compute the sample variances for each agent; denote the sample variance for agent k by $S_k^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_{k,j} - \overline{X}_k)^2$ where $\overline{X}_k$ is the sample mean of agent k.
2. Compute the pooled variance $S_p^2 = \frac{1}{M}(S_1^2 + \ldots + S_M^2)$.
3. Assume that all agents have the same variance estimated by S_p^2. For each of the M agents test $H_0 : \sigma_i^2 = S_p^2$ versus $H_a : \sigma_k^2 \neq S_p^2$ for k = 1, …, M. In order to hold the overall joint significance level

at .01 for the set of M tests, the significance level for each individual test is $\bar{\alpha} = .01/M$. The individual significance level is based upon the Bonferroni method (Miller 1966). The test statistic for agent k is $\frac{(n-1)S_k^2}{S_p^2}$ (Larson and Marx 2001).

4. Determine the initial number of groups characterized by the variance. If χ^2_{n-1} denotes a chi-squared random variable with n-1 degrees of freedom, let χ^2_{β} be the constant for which $P(\chi^2_{n-1} < \chi^2_{\beta}) = \beta$. There is one group if $\chi^2_{\bar{\alpha}/2} < \frac{(n-1)S_k^2}{S_p^2} < \chi^2_{1-\bar{\alpha}/2}$ for all k. There are two initial groups if $\frac{(n-1)S_k^2}{S_p^2} < \chi^2_{\bar{\alpha}/2}$ for at least one k and $\frac{(n-1)S_k^2}{S_p^2} < \chi^2_{1-\bar{\alpha}/2}$ for all k or if $\chi^2_{1-\bar{\alpha}/2} < \frac{(n-1)S_k^2}{S_p^2}$ for at least one k and $\chi^2_{\bar{\alpha}/2} < \frac{(n-1)S_k^2}{S_p^2}$ for all k. There are three initial groups if $\frac{(n-1)S_k^2}{S_p^2} < \chi^2_{\bar{\alpha}/2}$ for at least one k and if $\chi^2_{1-\bar{\alpha}/2} < \frac{(n-1)S_k^2}{S_p^2}$ for at least one k.

5. If there is only one group based upon the variances, go to the procedure to group the agents based upon the means. If there are two or more variance groups, based upon the number of groups, use the cluster analysis method described later to determine the groupings.

6. Let the number of agents in group j be M_j, then for group j compute the pooled variance for all the agents in this group: $S^2_{pj} = \frac{1}{M}(S_1^2 + \ldots + S^2_{Mj})$. For each of the agents in the jth group test $H_0 : \sigma_k^2 = S^2_{Pj}$ versus $H_a : \sigma_k^2 \neq S^2_{pj}$ with $\bar{\alpha} = .01/M_j$. Use the same rules as in step 4 to determine the number of subgroups in group j. If each of the j groups has exactly one subgroup, then go to the procedure to group based upon the means. If there is more than one subgroup in any of the j groups, then appropriately modify the total number of groups and go to step 5.

Procedure to Form Groups Based Upon the Sample Means

After having completed the procedure to group the agents based upon their variance, suppose there are J groups. This procedure should be performed on each of these J groups, since it is assumed that the variance of all agents within each group is the same. This procedure is based upon making multiple comparisons for the family of all possible differences in means for the individuals in each common variance group. In order to maintain the overall significance level at a specified small value, Tukey's multiple comparison method (Larson and Marx 2001) is used because it is the standard method to make multiple comparisons of the differences in all possible pairs of means for a large number of groups.

1. Let there be M_j agents in group j, then the Tukey half width (THW) for comparisons of the differences in means is THW = $S_{Pj}\,[q_{Mj,\nu}(\alpha)/\sqrt{n}\,]$ (Larson and Marx 2001). S_{Pj} is the pooled standard deviation for the M_j agents in group j. If $Q_{Mj,\nu}$ denotes a random variable with a studentized range distribution for M_j groups and $\nu = (n-1)M_j$ degrees of freedom then $q_{Mj,\nu}(\alpha)$ as the constant for which $P[Q_{Mj,\nu} > q_{Mj,\nu}(\alpha)] = \alpha$.

2. For each of the M_j agents compute the sample average, $\overline{Y}_k$ for k = 1 to M_j. Order these sample means from the smallest to the largest values and denote the ordered sample means by $\overline{Y}_{(k)}$ so that $\overline{Y}_{(1)} \leq \ldots \leq \overline{Y}_{(Mj)}$. Compute the tentative number of subgroups for this variance group as NG = $\frac{[\overline{Y}_{(Mj)} - \overline{Y}_{(1)}]}{\text{THW}}$. Round NG up to the next whole integer.

3. If there is only one subgroup based upon the mean groupings, go to the next variance grouping. If there are two or more subgroups, based upon the value of NG use the cluster analysis method described later to determine the groupings.

4. Verify that all the sample means in the final subgrouping satisfy the condition that $|\overline{Y}_k - \overline{Y}_I| \leq$ THW. If this condition is not satisfied for all pairs of means in the subgroup, then increase the NG so that the condition will be met and go to step 3 in this procedure.

Cluster Analysis to Form Final Groups

The procedure to form the groups is based on a K-means cluster algorithm originally suggested by MacQueen (1967).

1. Begin with K initial clusters based upon either the grouping from the variance procedure or from the mean procedure. When grouping the variances, each agent's classification metric will be characterized by the sample standard deviation. When grouping the means, each agent's classification metric will be characterized by the sample mean.
2. For each of the initial clusters, find the average of the classification metrics for all the agents in the cluster. This is the cluster's centroid.
3. Proceed through all the agents and assign each agent to the cluster whose centroid is nearest to the agent's classification measure. The distance from the classification measure to the centroid is the Euclidean distance. Recalculate the centroid for the cluster receiving the new agent and for the cluster losing the agent.
4. Repeat step 3 until no more reassignments take place.

Guidance for Selecting the Sample Size

In many situations there will be so much data that getting large samples for each agent will not be a problem. Having too large a sample, however, needs to be avoided since if the sample size is too large, a statistically significant difference in sample means may not be large enough to be of practical significance. Guidance in determining the sample size will be based upon ensuring that a statistically significant difference in two means will also be large enough to be practically significant.

The formula for the Tukey half-width (THW) for determining the confidence interval for the difference in two long-run means is THW = $S_p\,[q_{M,\nu}(\alpha)/\sqrt{n}\,]$ where S_p is the pooled sample standard deviation, $q_{M,\nu}(\alpha)$ is the tail value from the studentized range tables with M groups and ν degrees of freedom, and n is the sample size. It is desired to have the error for the difference in two means to be no larger than PD = the smallest difference in means that is of practical significance. It should be likely that the causes of a difference of at least PD can be found. The desired sample size can be found by setting THW equal to PD and solving the resulting equation for n. It follows that $n = \dfrac{S_p^2 q_{M,\nu}^2(\alpha)}{\mathrm{PD}^2}$. This should be the largest sample selected because it is desired to ensure that the resulting THW is not smaller than the PD.

In order to use the formula for determining n, the three values PD, S_p, and $q_{M,\nu}(\alpha)$ must be specified. PD is the difference between means for which it is believed is large enough so that the main causes of the difference can be found. The value of S_p can be estimated from the available data after the data have been screened and any special causes have been removed. The value of $q_{M,\nu}(\alpha)$ can be determined from tables for the studentized range. While the degrees of freedom will not be known exactly, they can be roughly estimated so that an approximate value for $q_{M,\nu}(\alpha)$ can be determined.

In the case study in this article, suppose that it is believed that it should be possible to find the causes of a difference in long-run means of at least 900 = PD. In order to estimate the value of S_p, data for each agent can be used to estimate that agent's sample standard deviation S_k. The values of S_k can be averaged to estimate the value of S_p. If it appears that the variance may change for different agents, the extreme values can be removed before computing the average. In the example the average standard deviation was of S_p = 1050. It is desired to have the value of α = .01. The values of $q_{M,\nu}(.01)$ for a large degree of freedom and a moderate number of groups are approximately 5.50. Substituting these values into the formula for the sample size yields $n = \dfrac{(1050)^2(5.50)^2}{(900)^2} = 41.17$. Thus, a sample size of about 40 would be appropriate.