From a lottery participant:

“I realize I don’t have a chance, but nobody’s got a chance. So the way I look at it, I have a 50-50 chance – either I win it or someone else wins it,” reasoned Barrie Green, 60, after buying a single ticket Monday afternoon at the Merritt Restaurant and Bakery near his home in Oakland.

-From the San Francisco Chronicle, March 5, 2007

Perhaps we could convince Mr. Green that since $2 \times 50\% = 100\%$ *he would be certain to win if he bought two tickets.*

Comparing Two or More Data Sets

Often, we want to compare two or more components of a process, as in the following example.

Comparing Two or More Data Sets

Often, we want to compare two or more components of a process, as in the following example.

A bakery finds too much variation in the weights of its 1 lb loaves of white bread. This bread is produced on 3 machines. Which machine(s) is responsible for the problem?

Comparing Two or More Data Sets

Often, we want to compare two or more components of a process, as in the following example.

A bakery finds too much variation in the weights of its 1 lb loaves of white bread. This bread is produced on 3 machines. Which machine(s) is responsible for the problem?

To find out, 8 loaves are randomly sampled from each machine's production on a given day and weighed. For each machine, the weights are plotted versus the time order of production and no evidence of nonstationarity is found.

Comparing Two or More Data Sets

Often, we want to compare two or more components of a process, as in the following example.

A bakery finds too much variation in the weights of its 1 lb loaves of white bread. This bread is produced on 3 machines. Which machine(s) is responsible for the problem?

To find out, 8 loaves are randomly sampled from each machine's production on a given day and weighed. For each machine, the weights are plotted versus the time order of production and no evidence of nonstationarity is found.

A plot appropriate for comparing data from stationary processes is a **stratified plot**. The stratified plot for the bakery data is shown in Figure 5.

$\approx$

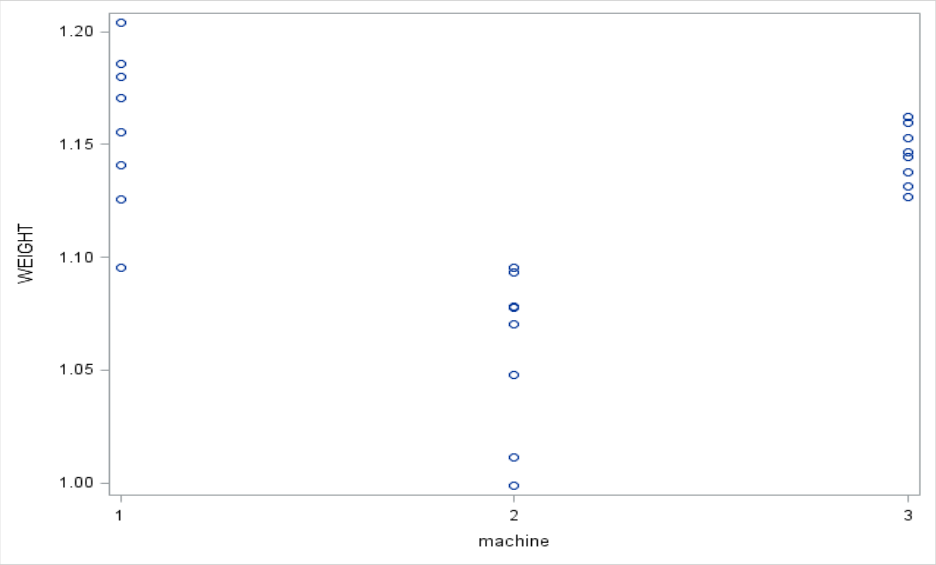


Figure: *Figure 5: A stratified plot of the weights of bread from 3 machines.*

The plot is called stratified because the data are broken into groups called **strata** and the distributions for the different strata are compared. The strata in this example are the machines.

The plot is called stratified because the data are broken into groups called **strata** and the distributions for the different strata are compared. The strata in this example are the machines.

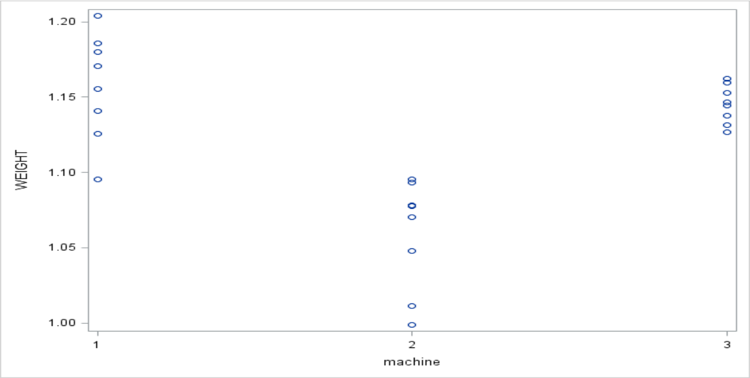
The things to look for in a stratified plot are the variation or spread within strata (**within variation**) and the differences in the locations or centers of the data in the different strata (**between variation**).

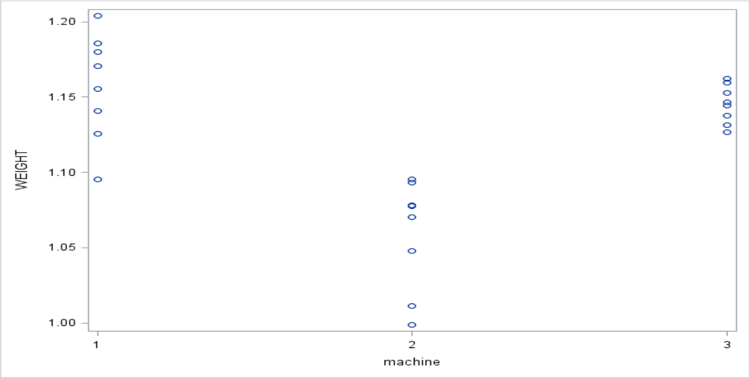
The plot is called stratified because the data are broken into groups called **strata** and the distributions for the different strata are compared. The strata in this example are the machines.

The things to look for in a stratified plot are the variation or spread within strata (**within variation**) and the differences in the locations or centers of the data in the different strata (**between variation**).

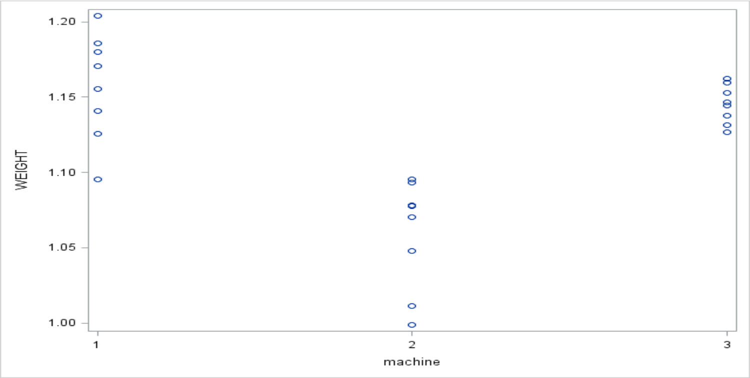
Have another look at the stratified plot of the bread weights and compare the within and between variation.

$\approx$


1.20
1.15
1.10
1.05
1.00
WEIGHT
1
2
3
machine

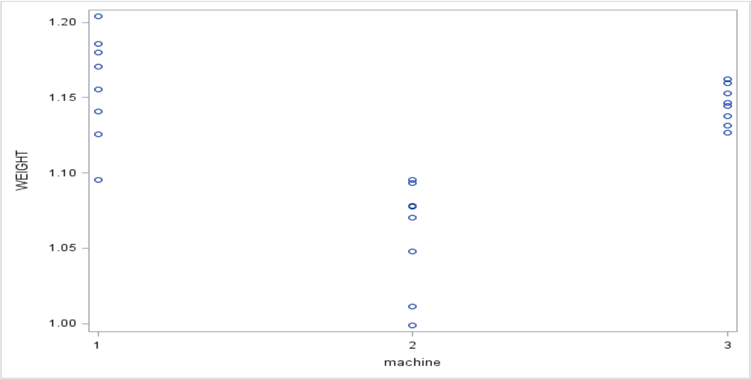


The plot seems to show substantial between and within variation.

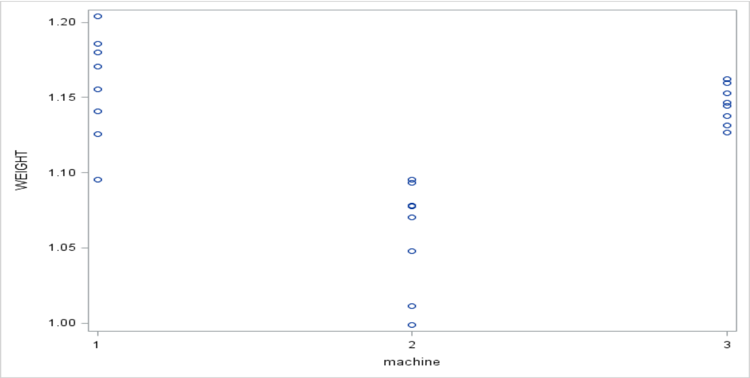


The plot seems to show substantial between and within variation.

The principal within variation is in machines 1 and 2: the variation in the weights produced by both machines is about the same and machine 3 produces bread with more consistent weights than either.



The between variation derives from the central value of the weights of bread produced by machine 2 being less than the central values in the weights produced by either machine 1 or 3. The central values for machines 1 and 3 are about equal.



The between variation derives from the central value of the weights of bread produced by machine 2 being less than the central values in the weights produced by either machine 1 or 3. The central values for machines 1 and 3 are about equal.

But how much of the variation is the result of differences in machines?

Measuring the Variation

We can measure the between, within, and total variation in the bakery data as follows. First, look at the data.

	Machine		
Loaf	1	2	3
1	1.1799	0.9987	1.1269
2	1.1409	1.0783	1.1595
3	1.1256	1.0778	1.1442
4	1.2039	1.0706	1.1620
5	1.1703	1.0956	1.1378
6	1.1858	1.0113	1.1269
7	1.1553	1.0476	1.1316
8	1.0952	1.0935	1.1466
Mean	1.1571	1.0592	1.1452

The overall mean is 1.1205

To measure the **total variation**, subtract the overall mean from each loaf's weight, then square the results and add up all those squares:

	Machine		
Loaf	1	2	3
1	$(1.1799 - 1.1205)^2$	$(0.9987 - 1.1205)^2$	$(1.1269 - 1.1205)^2$
2	$(1.1409 - 1.1205)^2$	$(1.0783 - 1.1205)^2$	$(1.1595 - 1.1205)^2$
3	$(1.1256 - 1.1205)^2$	$(1.0778 - 1.1205)^2$	$(1.1442 - 1.1205)^2$
4	$(1.2039 - 1.1205)^2$	$(1.0706 - 1.1205)^2$	$(1.1620 - 1.1205)^2$
5	$(1.1703 - 1.1205)^2$	$(1.0956 - 1.1205)^2$	$(1.1378 - 1.1205)^2$
6	$(1.1858 - 1.1205)^2$	$(1.0113 - 1.1205)^2$	$(1.1269 - 1.1205)^2$
7	$(1.1553 - 1.1205)^2$	$(1.0476 - 1.1205)^2$	$(1.1316 - 1.1205)^2$
8	$(1.0952 - 1.1205)^2$	$(1.0935 - 1.1205)^2$	$(1.1466 - 1.1205)^2$

To measure the **total variation**, subtract the overall mean from each loaf's weight, then square the results and add up all those squares:

Loaf	Machine 1	Machine 2	Machine 3
1	$(1.1799 - 1.1205)^2$	$(0.9987 - 1.1205)^2$	$(1.1269 - 1.1205)^2$
2	$(1.1409 - 1.1205)^2$	$(1.0783 - 1.1205)^2$	$(1.1595 - 1.1205)^2$
3	$(1.1256 - 1.1205)^2$	$(1.0778 - 1.1205)^2$	$(1.1442 - 1.1205)^2$
4	$(1.2039 - 1.1205)^2$	$(1.0706 - 1.1205)^2$	$(1.1620 - 1.1205)^2$
5	$(1.1703 - 1.1205)^2$	$(1.0956 - 1.1205)^2$	$(1.1378 - 1.1205)^2$
6	$(1.1858 - 1.1205)^2$	$(1.0113 - 1.1205)^2$	$(1.1269 - 1.1205)^2$
7	$(1.1553 - 1.1205)^2$	$(1.0476 - 1.1205)^2$	$(1.1316 - 1.1205)^2$
8	$(1.0952 - 1.1205)^2$	$(1.0935 - 1.1205)^2$	$(1.1466 - 1.1205)^2$

The resulting value is 0.0650.

To measure the **total variation**, subtract the overall mean from each loaf's weight, then square the results and add up all those squares:

Loaf	Machine 1	Machine 2	Machine 3
1	$(1.1799 - 1.1205)^2$	$(0.9987 - 1.1205)^2$	$(1.1269 - 1.1205)^2$
2	$(1.1409 - 1.1205)^2$	$(1.0783 - 1.1205)^2$	$(1.1595 - 1.1205)^2$
3	$(1.1256 - 1.1205)^2$	$(1.0778 - 1.1205)^2$	$(1.1442 - 1.1205)^2$
4	$(1.2039 - 1.1205)^2$	$(1.0706 - 1.1205)^2$	$(1.1620 - 1.1205)^2$
5	$(1.1703 - 1.1205)^2$	$(1.0956 - 1.1205)^2$	$(1.1378 - 1.1205)^2$
6	$(1.1858 - 1.1205)^2$	$(1.0113 - 1.1205)^2$	$(1.1269 - 1.1205)^2$
7	$(1.1553 - 1.1205)^2$	$(1.0476 - 1.1205)^2$	$(1.1316 - 1.1205)^2$
8	$(1.0952 - 1.1205)^2$	$(1.0935 - 1.1205)^2$	$(1.1466 - 1.1205)^2$

The resulting value is 0.0650. This is called the **total sum of squares**.

To measure the **within variation**, first subtract from each loaf's weight the mean weight for the loaves produced by the machine that manufactured it, then square the results and add up all those squares:

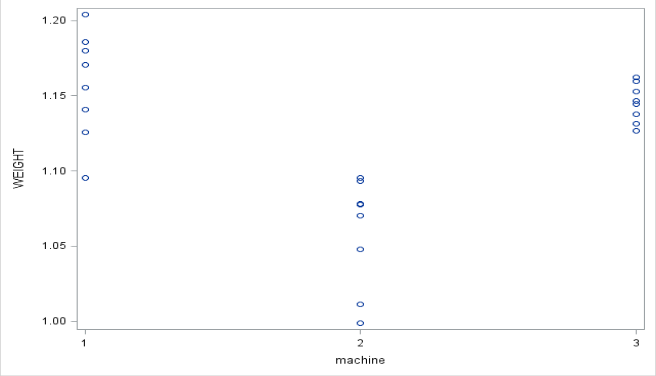
	Machine		
Loaf	1	2	3
1	$(1.1799 - 1.1571)^2$	$(0.9987 - 1.0592)^2$	$(1.1269 - 1.1452)^2$
2	$(1.1409 - 1.1571)^2$	$(1.0783 - 1.0592)^2$	$(1.1595 - 1.1452)^2$
3	$(1.1256 - 1.1571)^2$	$(1.0778 - 1.0592)^2$	$(1.1442 - 1.1452)^2$
4	$(1.2039 - 1.1571)^2$	$(1.0706 - 1.0592)^2$	$(1.1620 - 1.1452)^2$
5	$(1.1703 - 1.1571)^2$	$(1.0956 - 1.0592)^2$	$(1.1378 - 1.1452)^2$
6	$(1.1858 - 1.1571)^2$	$(1.0113 - 1.0592)^2$	$(1.1269 - 1.1452)^2$
7	$(1.1553 - 1.1571)^2$	$(1.0476 - 1.0592)^2$	$(1.1316 - 1.1452)^2$
8	$(1.0952 - 1.1571)^2$	$(1.0935 - 1.0592)^2$	$(1.1466 - 1.1452)^2$
Total	0.0088	0.0094	0.0011

To measure the **within variation**, first subtract from each loaf's weight the mean weight for the loaves produced by the machine that manufactured it, then square the results and add up all those squares:

	Machine		
Loaf	1	2	3
1	$(1.1799 - 1.1571)^2$	$(0.9987 - 1.0592)^2$	$(1.1269 - 1.1452)^2$
2	$(1.1409 - 1.1571)^2$	$(1.0783 - 1.0592)^2$	$(1.1595 - 1.1452)^2$
3	$(1.1256 - 1.1571)^2$	$(1.0778 - 1.0592)^2$	$(1.1442 - 1.1452)^2$
4	$(1.2039 - 1.1571)^2$	$(1.0706 - 1.0592)^2$	$(1.1620 - 1.1452)^2$
5	$(1.1703 - 1.1571)^2$	$(1.0956 - 1.0592)^2$	$(1.1378 - 1.1452)^2$
6	$(1.1858 - 1.1571)^2$	$(1.0113 - 1.0592)^2$	$(1.1269 - 1.1452)^2$
7	$(1.1553 - 1.1571)^2$	$(1.0476 - 1.0592)^2$	$(1.1316 - 1.1452)^2$
8	$(1.0952 - 1.1571)^2$	$(1.0935 - 1.0592)^2$	$(1.1466 - 1.1452)^2$
Total	0.0088	0.0094	0.0011

These give a measure of within variation for each machine. How does this jive with the stratified plot?

Within SS: 0.0088 0.0094 0.0011



$\approx$

To get an overall measure of within variation, compute the **within sum of squares** by adding the sums of squares for the three machines:

$$0.0088 + 0.0094 + 0.0011 = 0.0193.$$

To get an overall measure of within variation, compute the **within sum of squares** by adding the sums of squares for the three machines:

$$0.0088 + 0.0094 + 0.0011 = 0.0193.$$

The measure of **between variation**, called the **between sum of squares**, is the difference between the total sum of squares and the within sum of squares:

$$0.0650 - 0.0193 = 0.0457.$$

$\approx$

Thus, we have

Source	Sum of Squares
Between	0.0457
Within	0.0193
Total	0.0650

Thus, we have

Source	Sum of Squares
Between	0.0457
Within	0.0193
Total	0.0650

From these results, we see that between variation is the principal source of variation for the bakery data, and specifically, that it accounts for

$$100 \times \frac{0.0457}{0.0650} = 70.25\%$$

of the total variation in the data.

Thus, we have

Source	Sum of Squares
Between	0.0457
Within	0.0193
Total	0.0650

From these results, we see that between variation is the principal source of variation for the bakery data, and specifically, that it accounts for

$$100 \times \frac{0.0457}{0.0650} = 70.25\%$$

of the total variation in the data.

SAS code to produce the stratified plot and compute the sums of squares is found here.

$\approx$

Gage R&R

Gage R&R

In a measurement (or gage-ing) process, two sources of variation are often considered:

Gage R&R

In a measurement (or gage-ing) process, two sources of variation are often considered:

- **Repeatability**: consistency of the gage (measuring tool) in repeated measurements by the same operator on the same part.

Gage R&R

In a measurement (or gage-ing) process, two sources of variation are often considered:

- **Repeatability**: consistency of the gage (measuring tool) in repeated measurements by the same operator on the same part.
- **Reproducibility**: consistency of measurements between different operators.

$\approx$

Example:

In a gage R&R study of a laser ranging device, four operators each took fifteen measurements of the same distance. For a gage R&R study, we stratify by operator.

Example:

In a gage R&R study of a laser ranging device, four operators each took fifteen measurements of the same distance. For a gage R&R study, we stratify by operator.

Figure 6 (code here) shows four time series plots, one for each operator, on the same graph. We say the plots are **stratified** by operator.

$\approx$

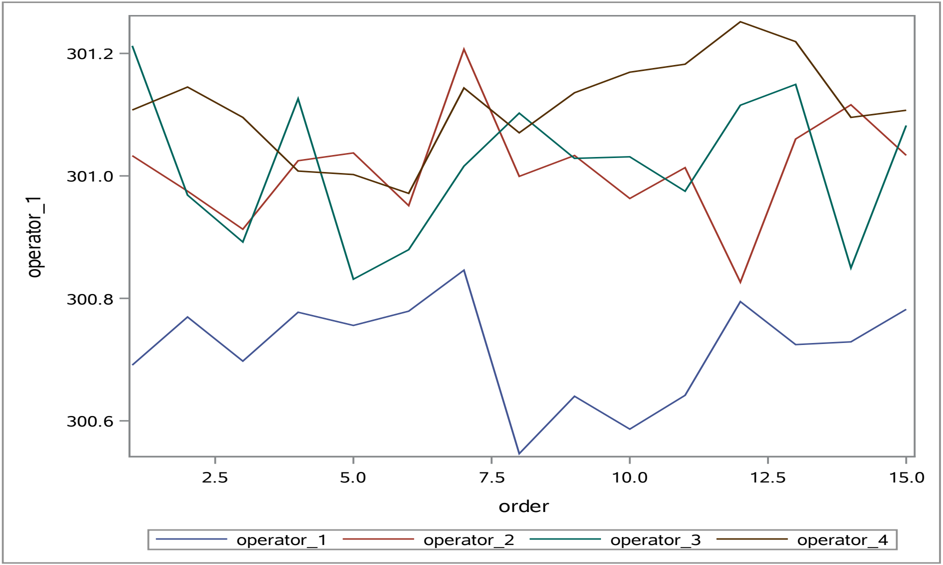


Figure: *Figure 6: Time series plots of laser ranging device readings, stratified by operator.*

None of these plots shows evidence of nonstationarity, which means we can analyze the pattern of variation without regard to order.

None of these plots shows evidence of nonstationarity, which means we can analyze the pattern of variation without regard to order.

Figure 7 (code here) is a stratified plot of the measurements stratified by operator but ignoring order. This is the kind of plot we considered in the bakery example.

$\approx$

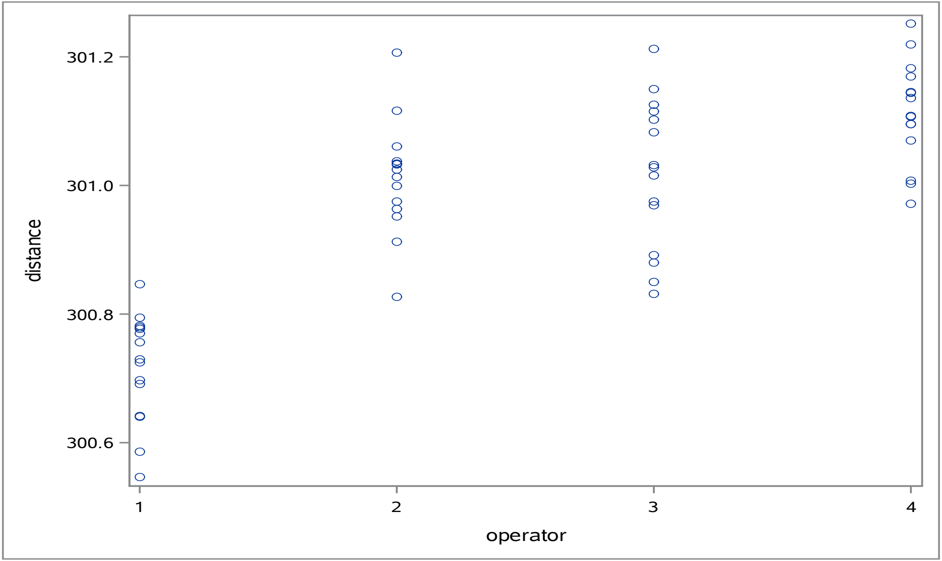


Figure: *Figure 7: Stratified plot of laser ranging device readings, stratified by operator.*

- The within variation measures the **repeatability** of the gage: the consistency of the gage in repeated measurements. Since larger within variation corresponds to the measurement process being less repeatable, it is more accurate to say the within variation measures the non-repeatability.
- The between variation measures the **reproducibility** of the measurement process: the consistency of measurements taken by different operators. As above, it is more accurate to say that the between variation measures the non-reproducibility.

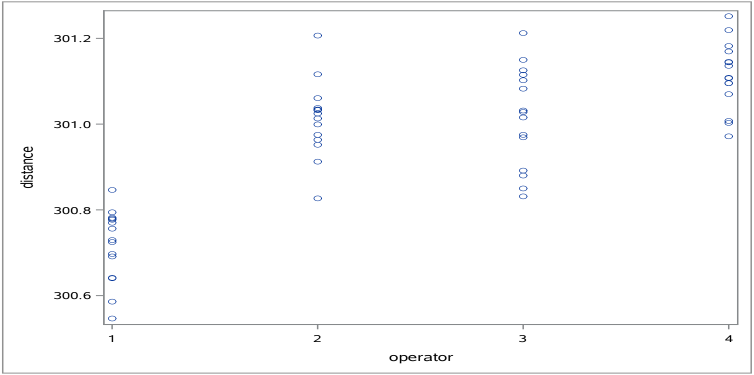
$\approx$

Consideration of the stratified plot in Figure 7 shows that all operators exhibit roughly the same repeatability, since the spreads of the measurements are roughly the same order of magnitude. In fact the within sums of squares for operators 1 to 4 are 0.0976, 0.1051, 0.1903, and 0.0869, respectively.

Whether the repeatability is adequate for the tasks for which the laser ranging device is used is not answered by this analysis. It can only be answered by the users.

$\approx$

Also, while the measurements of operators 2-4 are centered at roughly the same value, those of operator 1 are centered about a substantially lower value than the other three. This means there is a reproducibility problem with the measuring process.



To quantify this, we note that the measure of between variation is 1.3258, and that of within variation is 0.4799, giving a total of 1.8058. Thus the percentage of total variation attributable to between variation (or non-reproducibility) is

$$100 \times \frac{1.3258}{1.8058} = 73.42\%.$$

$\approx$

What's the IDEA?

What's the IDEA?

Variation can sometimes be broken into pieces which identify different sources of the variation and the amount of variation each source contributes. One example is breaking variation into between and within components. Plots of data stratified by these sources can be helpful in analyzing the structure of variation.

$\approx$

Recap:

Recap:

- Data and its science, statistics

Recap:

- Data and its science, statistics
- Stationary and nonstationary processes; displaying data from each.

Recap:

- Data and its science, statistics
- Stationary and nonstationary processes; displaying data from each.
- Assessing between and within variation.

$\approx$