

Blocking Activity: Can you see the trees for the forest?

This activity is a simulation of problem 4 on the 2001 AP Statistics exam. The simulation will ask you to assign trees of two varieties, a new hybrid and a standard variety, to the eight available plots in the field. You'll have the opportunity to assign the trees to the plots in three different ways: a design with no blocking, a design that uses blocking scheme A from the problem, and a design that uses blocking scheme B from the problem. Together with your classmates, your goal is to find the experimental design that has the best chance of correctly identifying which variety of tree performs best, on average.

Let's assume that the new variety really is better. In our simulation, trees of the new variety will have an average production of 53 pounds of fruit, while the old variety typically produces 47 pounds of fruit.

Ideally, designs that have a better chance of detecting the superiority of the new trees are preferred.

We'll also assume that trees planted close to the forest grow more poorly than those away. If a tree is planted toward the forest, its production is reduced by 10 pounds, and for those away the production will be increased by 10 pounds.

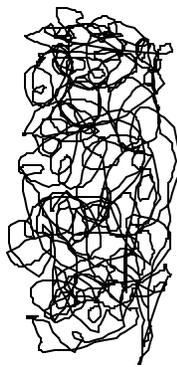
Even within a variety, there is variability in the production from tree to tree. Also, individual plots in the field will have differences due to water, fertility, or other factors, adding variability in the production from plot to plot. In this simulation the variability of trees and plots will be accounted for by adding or subtracting a random quantity for each tree. The total productivity will be calculated as a sum:

Variety (47 or 53) + Proximity to forest (-10 or +10) + Variation (Random)

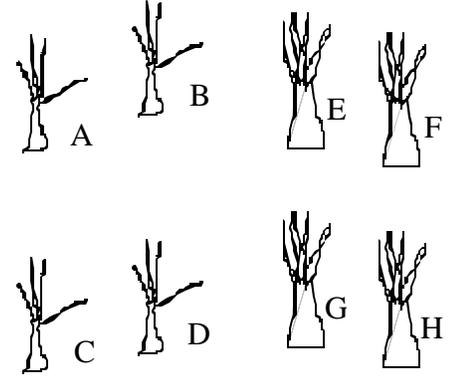
Here's one way to determine the random amount to add or subtract, using a coin and a die. First flip the coin: if it's heads, the tree is a little stronger than average, tails a little weaker. How much stronger or weaker is determined by rolling the die. For example, H 3 means add 3 to the productivity of that tree, T 5 means subtract 5.

There are other methods of accounting for the differences of individual trees, including using a table of random digits or a graphing calculator. Your teacher may want you to use one of these other strategies.

Simulation 1: A completely randomized design.



(Subtract 10 in this column)	(Add 10 in this column)
1	2
3	4
5	6
7	8



Old Variety New Variety

Ok let's plant the trees! The trees have been labeled A - H. The plots are numbered 1 - 8.

To assign the trees to the plots, you can either use a table of random digits or your calculator. Using a table of random digits, select a digit, which will be the plot in which tree A is planted. If the digit is 9 or 0, ignore it and choose the next digit. Then select a digit for where to plant tree B. Since only one tree can go in a plot, if you choose the same number you had for tree A, skip this and move on. Continue in this way until all trees are planted.

Now we'll have to wait while the trees grow and the fruit ripens.

Harvest time! We need to measure the productivity of each of the trees. You'll calculate the productivity as

- Tree Variety (new = 53, old = 47)
- + Forest position (-10 for near the forest, +10 for away)
- + Random variation (using coin and die, or other method)

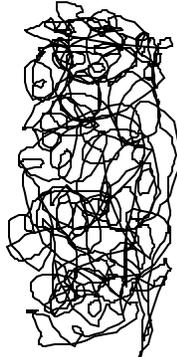
Calculate the productivity of each of your 8 trees and record this in the table above. Then calculate the average productivity of the new variety and the old variety. Record the results below, and indicate which variety was better. (As a check, these averages should be between 40 and 60.)

Old Average Productivity: _____ New Average Productivity: _____

Which was better? Difference (New - Old): _____

Combine your results with your classmates. How often did the new trees come out with a better result? Describe the distribution of differences.

Simulation 2: Blocking Scheme A.



	Block 1, trees ABEF (Subtract 10)	Block 2, trees CDGH (Add 10)
1		2
3		4
5		6
7		8










Old Variety New Variety

As before, the trees have been labeled A – H and the plots are numbered 1 - 8. In this scheme, however, you need to make sure that two of each type of tree are planted in each block (column). So, make the assignments of trees A and B to the first block (keep picking digits until 1, 3, 5, or 7 is chosen), with C and D placed in the second block (must be 2, 4, 6, or 8). Do the same with the new trees: make sure that trees E and F are in block 1, G and H in block 2.

Grow, grow, grow, grow, ...

Time to harvest. You'll calculate the productivity of each tree as before:

- Tree Variety (new = 53, old = 47)
- + Forest position (-10 for near the forest, +10 for away)
- + Random variation (using coin and die, or other method)

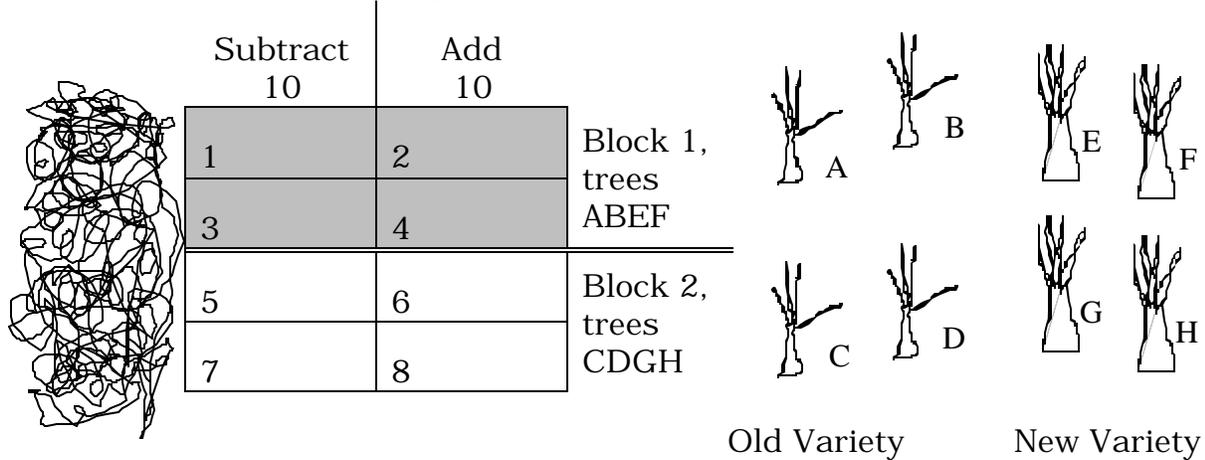
Calculate the productivity of each of your 8 trees, then calculate the average productivity of the new variety and the old variety. Record the results below, and indicate which variety was better. (As a check, these averages should be between 40 and 60.)

Old Average Productivity: _____ New Average Productivity: _____

Which was better? Difference (New – Old): _____

Combine your results with your classmates. How often did the new trees come out with a better result? Describe the distribution of differences.

Simulation 3: Blocking Scheme B.



As in the previous blocking example, you'll first assign the trees A, B, E, and F to the first block, but in this case these are the top 4 plots (shaded above). Make sure that these are assigned to plots 1 - 4. Then assign trees C, D, G, and H to the second block, plots 5 - 8.

Grow, grow, grow, grow, ...

Time to harvest. You'll calculate the productivity of each tree as before. Be careful to add or subtract the 10 pounds according to whether the tree is next to the forest or away - it doesn't vary by block in this case.

- Tree Variety (new = 53, old = 47)
- + Forest position (-10 for near the forest, +10 for away)
- + Random variation (using coin and die, or other method)

Calculate the productivity of each of your 8 trees, then calculate the average productivity of the new variety and the old variety. Record the results below, and indicate which variety was better. (As a check, these averages should be between 40 and 60.)

Old Average Productivity: _____ New Average Productivity: _____

Which was better? _____ Difference (New - Old): _____

Combine your results with your classmates. How often did the new trees come out with a better result? Describe the distribution of differences.

Blocking Activity: Can you see the trees for the forest?

Summary and conclusion.

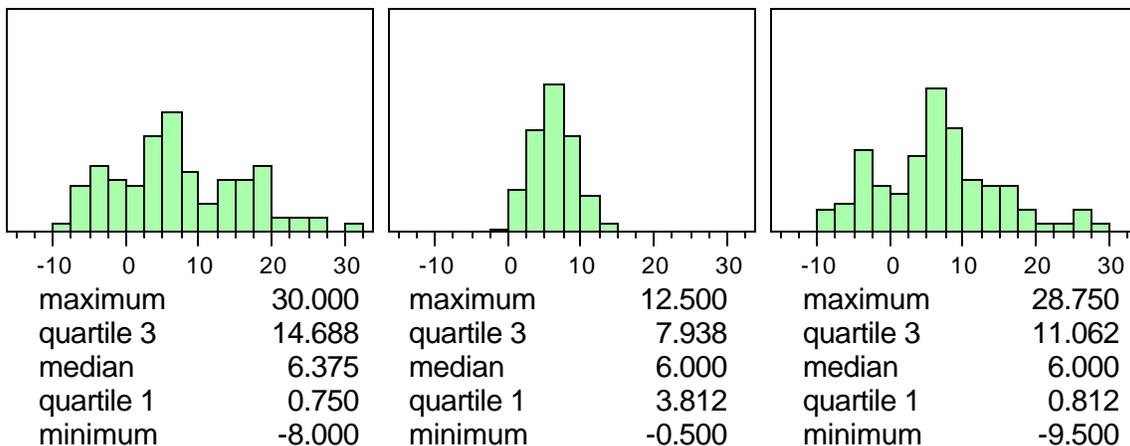
In your simulation, you and your classmates collected data from a number of runs of the simulation. You should have found that when blocking scheme A was used, the new trees had a higher average in almost every case. But for the completely randomized design or blocking with scheme B, the incorrect conclusion that the old variety of trees had a higher average was made in some of the runs.

Below is a graphical illustration for the results of 100 runs of the simulation for the completely randomized design and the designs with blocking schemes A and B. The histograms show the differences in productivity for these 100 runs. Notice that for all of these distributions, the center is approximately 6, indicating that none has any bias. However, the spread for the completely randomized design and blocking scheme B is much greater. Because of this larger spread, almost a quarter of the differences are negative. A decision based on these would find that the old variety is superior – which is incorrect. The narrower spread using blocking scheme A means that the results from run to run are much more consistent, and only very few run values are negative, making it much more likely that a single run, or single experiment in the real world, will produce the correct decision. Below the graphs are numerical summaries for the data.

Completely Randomized

Blocking Scheme A

Blocking Scheme B



This illustrates the purpose of blocking in experiments:

Blocking reduces variability in the estimated difference in treatment means.

Blocking Activity – Can you see the trees for the forest?

Teacher notes

Simulations

In the AP Statistics course we take on the task of helping students design and execute simulations, and analyze the results of such simulations. By simulations we mean an artificial chance experiment designed to mimic a real world problem in all relevant aspects. It is hoped that the distribution of outcomes of such a simulation will lead to an understanding of the corresponding real world problem. For additional discussion of the ideas and methodologies of simulations, see [Chris Olsen's paper].

Blocking and Randomization

The purpose of an experiment is to determine which of two or more treatments is superior in producing a desired outcome. However, sometimes other factors, called confounding variables, are present. These confounding variables make it difficult to correctly decide which treatment is better. This happens because the effects of these confounding variables can't be separated from the effects of the treatments in the outcome of the experiment. Two strategies are employed in experiments to overcome the effects of these: blocking and randomization. On the 2001 AP Statistics exam, a question appeared that asked students about these two important aspects of experiments. The purpose of this activity is to illustrate the concept of blocking through a simulation based on the AP problem. In particular, the goal of this simulation is to illustrate the main purpose of blocking: to reduce variability in the estimated difference in treatment means.

So that student's attention can be focused on the issues of blocking, the simulation doesn't address the issue of randomization.

The AP Question is available at <http://collegeboard.org/ap/statistics>

Blocking

Blocking can be used to account for the effects of those confounding variables that we can anticipate as we set up the experiment. In this example, the presence of the forest on one side of the plots is an obvious source of variability, since trees next to the forest may systematically perform better or worse than those away. Blocking can be used to account for the effects of the location of the tree. By choosing blocking scheme A, and then assigning two trees of each variety to the four plots in each block, two trees of each variety will be adjacent to the forest and two trees of each variety will be away from the forest. This removes the

possibility that more of one variety than the other will be placed in the more favorable plots with respect to the forest, gaining an unfair advantage.

Generally, blocks should be constructed so that they include experimental units (plots, in this case) that are similar with regard to the blocking variable. Differences from block-to-block should be as great as possible.

Randomization

On the other hand, randomization is used to even out the effects of unanticipated confounding variables. In the present case, there may be differences such as water or fertility in the plots that are oriented in a north-south direction. If this is the case, with perhaps the southern plots producing better, a possible arrangement of the trees in the blocks might have all of one variety in the southern plots and the other variety in the northern plots. Again, this would give the variety of trees in the southern plots an unfair advantage.

Since we can't know how all these unanticipated effects are distributed in the plots, the best strategy is to place the trees randomly into the plots. On average, the effects of these confounding variables will be shared equally by the two varieties. (The benefit of randomization is not illustrated here because there are no other confounding variables.)

This simulation activity will ask students to assign trees of two varieties, a new hybrid and a standard variety, to the eight plots in the field. The productivity of each of the trees will be calculated from its variety (new or old), position in the grid (by the forest or away), and a random component that simulates variation among individual trees and individual plots. The average productivity of the two varieties will be compared to determine which performs better.

For this simulation, the new hybrid does perform better on average than the old variety. However, due to the effects of the placement in the plots and random variation, any particular run of the simulation may or may not show that the new variety is superior. The question of interest here is how often does it occur that the mean production of the new (assumed better) variety exceeds the mean production of the old. The goal of the simulation is to show that it is more likely that we will observe that the new trees are better if blocking scheme A is employed.

This simulation has avoided the issue of statistical significance, to keep the focus on blocking as clear as possible. Also, since experimental design is often introduced in a course before significance testing, students may not have been exposed to this idea.

The simulation described on the student worksheet pages is done without using technology. It is recommended that students carry out at least one run of each simulation in this way. However, repeated runs of the simulation become tedious very quickly. If your students have access to TI-83 calculators, there is a program listing attached as an appendix to this document that performs runs of the simulation using each of the designs. Particularly if your class is small, this is an effective way to gather a larger set of data, so that the points made will be more convincing. The program is also available to be downloaded at [webaddress].

For a further discussion of this problem, and other AP problems that deal with experimental design see [whatever Gloria, Beverly, Janice's project is called].

What to Expect

There are five pages for this student activity: an overview, then a worksheet page for each of the three schemes (completely randomized, blocking scheme A, blocking scheme B), then a summary page. It is recommended that you retain the summary page until the students have finished the activity. Here are some additional comments for each of these three approaches.

Simulation 1: Completely Randomized design

Task 1: Assigning trees to plots.

The trees have already been labeled (A – H) and the plots numbered (1 – 8). The suggested method of randomly assigning the trees to the plots uses a table of random digits, but you may want students to use a calculator to select a random digit from 1 to 8.

Task 2: Calculating the harvest.

Students need to be careful with the calculations. To better evaluate student work, you may want students to show the calculation (three terms) in each of the plots on the paper.

There are other methods for figuring out the variation of individual trees and plots. The method suggested on the first page of the student handout is to use a coin and a die. One alternative to this is to use a table of random digits. Use groups of two digits; the first to determine whether to add or subtract, the second digit how much to add or subtract. Another approach is to use a graphing calculator. This allows the most realistic simulation of the variation as a randomly selected number from a normal distribution. For example, on the TI-83 calculator, this can be done using `randNorm(0, 5)`, which selects a random number from a normal population with mean 0 and standard deviation 5.

Task 3: Reporting the results.

With the numbers that are included in the handout, and using the method of the die and coins, you should expect that about 75% of the runs with the completely randomized approach will correctly identify the new trees as having a higher average production.

The distribution of differences should be unbiased. That is, since the actual difference in productivity between the varieties is 6 pounds, on the average the difference should be 6 pounds. However, there can be

considerable variability. The distribution is quite spread out, to the extent that in about 25% of the runs, the difference is negative, meaning that the old trees would be judged to have performed better.

Having a large number of runs will help make the difference between the completely randomized design and the blocking design that follows more obvious. Depending on your class size, you may want to have students carry out several runs, perhaps as a homework assignment.

You can also use the calculator program at this point to generate a larger number of runs than is convenient using pencil and paper.

Simulation 2: Blocking Scheme A

Task 1: Assigning trees to plots.

In this scheme, the assignment to plots is a little more complicated. Make sure the students are ending up with 2 of each variety in the vertical blocks.

Task 2: Calculating the harvest.

The calculation of the harvest is very similar to the first case.

Task 3: Reporting results.

This method is generally superior to the completely randomized design. Almost all runs should result in the new trees having a higher average than the old trees (about 98% of the time).

The distribution of differences is also unbiased in this case (centered at 6), but there is much less variability from run to run. This explains why the proportion of runs in which the new trees have a higher average is so much larger – the differences are clustered much more tightly about 6, and only rarely is a difference negative (with the old trees having a higher average).

See the notes at the end of this document for some additional discussion and mathematical justification of why the variation is reduced by blocking scheme A.

Simulation 3: Blocking Scheme B

Task 1: Assigning trees to plots.

Make sure the students are ending up with 2 of each variety in the horizontal blocks.

Task 2: Calculating the harvest.

For these calculations, students should be reminded that the effect of the forest is still present. Thus, any tree that is adjacent to the forest gets the production penalty of -10 , and those away get a bonus of 10 . Since this doesn't correspond to the block structure, there is additional chance for confusion.

Task 3: Reporting results.

This method produces results similar to the completely randomized design, with about 75% of the runs having the new trees with the higher average. The distribution of differences is also similar to the completely randomized design, with no bias (center at about 6) but larger variability.

Activity Extensions

This activity can be redone with some changes to make the effects of the forest and the random variation more evident. Here are some suggestions:

1. Change the effect of the forest. Try adding/subtracting only 5. In this case, the difference between the completely randomized design and blocking scheme A will be less evident. Reducing this to 0 should result in having no difference between the completely randomized and blocking scheme A, reinforcing the idea that blocking is effective only if there is a confounding variable that affects the response in a systematic way. As it is set up in the original activity, the effect of the forest is very large, so making the effect even larger doesn't appreciably change the results.
2. Change the random variation component. If you make this larger (double the number on the die, or increase the standard deviation if you add this with a random normal), the differences between the varieties is harder to detect. Conversely, making this smaller makes it easier to see the differences between the varieties.
3. Change the difference between the varieties. In the original case, the difference is 6 on average. Making this smaller will make it harder to detect the difference between the varieties.

These changes can also be used to introduce the ideas of power in an informal way.

More on why blocking reduces variability.

Here is an example, including some extreme-case scenarios for the design schemes, that may clarify why the variation for blocking scheme A is much smaller.

First, let's consider the extreme cases in the **completely randomized design**. This happens when two things occur: trees of one variety are all planted in the plots away from the forest, and the random variability is large and positive for the one variety, but large and negative for the other. Here are some calculations to illustrate this:

$$\text{New variety maximum: } \frac{(53 + 10 + 6) + (53 + 10 + 6) + (53 + 10 + 6) + (53 + 10 + 6)}{4} = 69$$

$$\text{Old variety minimum: } \frac{(47 - 10 - 6) + (47 - 10 - 6) + (47 - 10 - 6) + (47 - 10 - 6)}{4} = 31$$

$$\text{New} - \text{Old} = 38.$$

$$\text{New variety minimum: } \frac{(53 - 10 - 6) + (53 - 10 - 6) + (53 - 10 - 6) + (53 - 10 - 6)}{4} = 37$$

$$\text{Old variety maximum: } \frac{(47 + 10 + 6) + (47 + 10 + 6) + (47 + 10 + 6) + (47 + 10 + 6)}{4} = 63$$

$$\text{New} - \text{Old} = -26.$$

Therefore the range of possible differences in treatment means is 64.

Using **blocking scheme A**, the extreme cases are:

$$\text{New variety maximum: } \frac{(53 + 10 + 6) + (53 + 10 + 6) + (53 - 10 + 6) + (53 - 10 + 6)}{4} = 59$$

$$\text{Old variety minimum: } \frac{(47 - 10 - 6) + (47 - 10 - 6) + (47 + 10 - 6) + (47 + 10 - 6)}{4} = 41$$

$$\text{New} - \text{Old} = 18.$$

$$\text{New variety minimum: } \frac{(53 + 10 - 6) + (53 + 10 - 6) + (53 - 10 - 6) + (53 - 10 - 6)}{4} = 47$$

$$\text{Old variety maximum: } \frac{(47 - 10 + 6) + (47 - 10 + 6) + (47 + 10 + 6) + (47 + 10 + 6)}{4} = 53$$

$$\text{New} - \text{Old} = -6.$$

Therefore the range of possible differences in treatment means is 24.

This result for blocking scheme A is much less than the extreme case of 64 for the completely randomized design. The reason for this is that the effect of the forest is distributed equally in the blocked design, essentially canceling out the 10's that are added or subtracted.

Importantly, this cancellation of the forest effect happens in all cases with blocking scheme A, no matter what the results of the random variation. As illustrated in the histograms on student page 5, blocking reduces variability in the estimated difference in treatment means.


```

N+1 N
Else
22 L>(B)
O+1 O
End

If N=2
Then
For(C,N+O+1,4,1)
22 L>(C)
End
Goto 8
End

If O=2
Then
For(C,N+O+1,4,1)
11 L>(C)
End
Goto 8
End

End

Lbl 8

0 N
0 O
0 Z

For(B,5,8,1)
Z+1 Z
rand R

If R=.5
Then
11 L>(B)
N+1 N
Else
22 L>(B)
O+1 O
End

If N=2
Then
For(C,N+O+5,8,1)
22 L>(C)
End
Goto 6
End

If O=2
Then
For(C,N+O+5,8,1)
11 L>(C)
End
Goto 6
End

End

Lbl 6

"CALCULATE AND DRAW THE
YIELD"

ClrDraw
RecallPic 9

If K=1
Text(3,68,Str1)

If K=2
Text(3,67,Str2)
If K=3
Text(3,67,Str3)

0 M
0 S

For(D,1,7,2)
If L>(D)=11
Then
53+randNorm(0,5)-
10 L?(D)
Text(L/(D),L^D)-
12,"NEW"
L?(D)+M M
End
If L>(D)=22
Then
47+randNorm(0,5)-
10 L?(D)
Text(L/(D),L^D)-
12,"OLD"
L?(D)+S S
End
Text(L/(D),L^D)+3,round
(L?(D),1)
End

For(D,2,8,2)
If L>(D)=11
Then
53+randNorm(0,5)+10 L?(D)
)
Text(L/(D),L^D)-
12,"NEW"
L?(D)+M M
End
If L>(D)=22
Then
47+randNorm(0,5)+10 L?(D)
)
Text(L/(D),L^D)-
12,"OLD"
L?(D)+S S
End
Text(L/(D),L^D)+3,round
(L?(D),1)
End

"BORDER AT END"

Line(-10,10,10,10)
Line(-10,-10,10,-10)
Line(10,-10,10,10)
Line(10/3,6.5,10,6.5)

"DISPLAY THE TWO MEANS"

Text(17,68,"E (NEW) ")
Text(27,73,round(M/4,1))
Text(39,68,"E (OLD) ")
Text(49,73,round(S/4,1))

Pause

0 N
0 O
0 Z

Menu("****OPTIONS****","
REPEAT SAME",51,"DIFF.
DESIGN",52,"QUIT",53)

Lbl 51
If K=1
Goto 1
If K=2
Goto 2
If K=3
Goto 3

Lbl 53

```