Statistics of Illumination

Beth Chance
bchance@calpoly.edu

Roxy Peck
rpeck@calpoly.edu

Activity: Matching Variables to Graphs **(adapted from *Activity-Based Statistics*)**

After the warm-up example, match the following variables with the histograms given below. *Hint:* Think about how each variable should behave.

(a) The height of people in this class

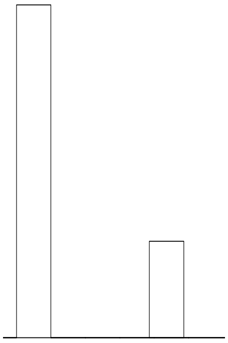
(b) Students' preference for coke vs. pepsi

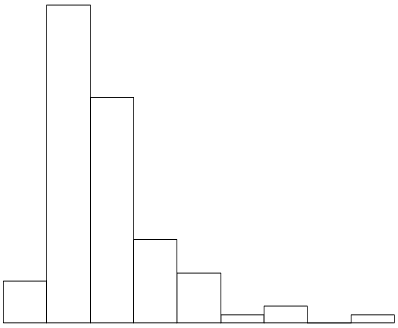
(c) Number of siblings of individuals in this class

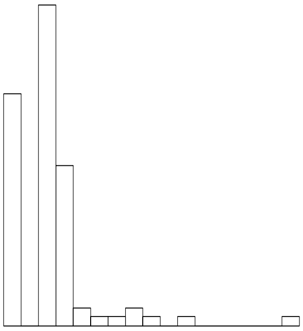
(d) Amount paid for last haircut by students in this class

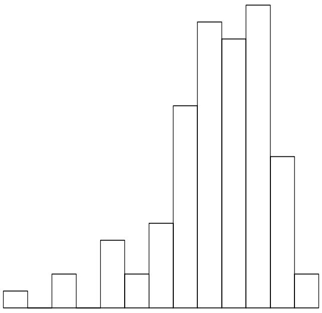
(e) Gender breakdown of students in this class

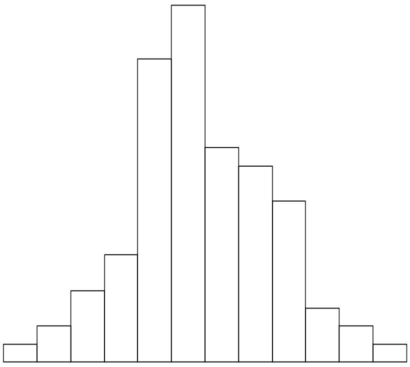
(f) Students’ guesses of my age.

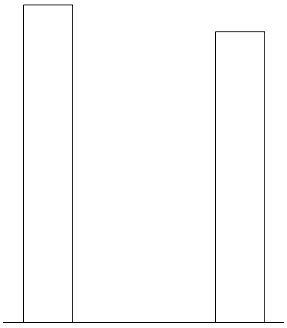










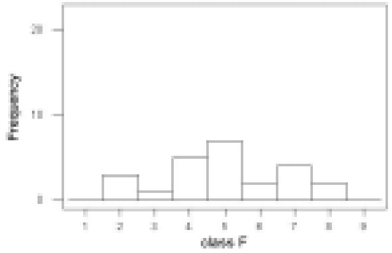


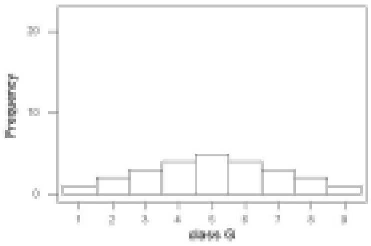
Write a paragraph explaining how you matched the graphs. For example, what features helped you decide?

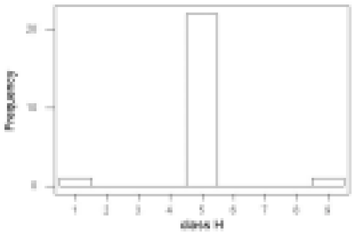
ACTIVITY: Describing Variability

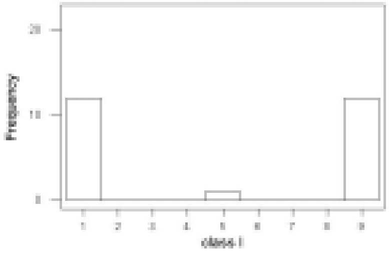
Consider students' rating the value of statistics in society on a numerical scale of 1 to 9. Below are the ratings of five hypothetical classes, where the data are given in the table and displayed in the following histograms:

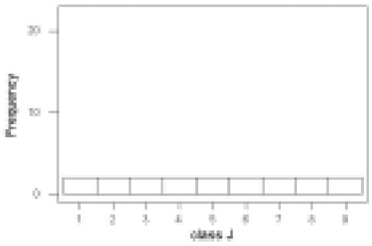
rating	1	2	3	4	5	6	7	8	9
class F count	0	3	1	5	7	2	4	2	0
class G count	1	2	3	4	5	4	3	2	1
class H count	1	0	0	0	22	0	0	0	1
class I count	12	0	0	0	1	0	0	0	12
class J count	2	2	2	2	2	2	2	2	2











a. Judging from the tables and histograms, take a guess as to which has more variability between classes F and G?

b. Judging from the tables and histograms, which would you say has the most variability among classes H, I, and J? Which would you say has the least variability?

most variability: least variability:

c. Below are the range, interquartile range, and standard deviation of the ratings for each class. :

	class F	class G	class H	class I	class J
range	6	8	8	8	8
interquartile range	2.75	3	0	8	5
standard deviation	1.769	2.041	1.180	4.000	2.657

Judging from these statistics, which measure spread, does class F or G have more variability? Was your expectation in (a) correct?

d. Judging from these statistics, which measure spread, which among classes H, I, and J has the most variability? Which has the least? Was your expectation in (b) correct?

e. Between classes F and G, which has more “bumpiness” or unevenness? Does that class have more or less variability than the other?

f. Among classes H, I, and J, which distribution has the most distinct values? Does that class have the most variability of the three?

g. Based on the previous two questions, does either “bumpiness” or “variety” relate directly to the concept of variability? Explain.

A common misconception about variability is to believe that a “bumpier” histogram indicates a more variable distribution, but this is not the case. Similarly, the number of distinct values represented in a histogram does not necessarily indicate greater variability.

Activity: Counts Versus Ratios

Consider the following statement:

In 1989, 5236 drivers age 65 and over were involved in fatal accidents, compared to only 2900 drivers aged 16 and 17, so young people are safer drivers.

Assessing the effect of motorcycle helmet laws in the U.S.:

25 states have mandatory helmet laws (called mandatory states)

22 states require helmet use only for riders under 18 years of age, and 3 states (Iowa, Illinois, and Colorado) have no helmet laws (called voluntary states)

Consider the following statements, used as arguments against helmet laws by The Motorcycle Industry Council:

- 65% of motorcycle fatalities in 1993 occurred in mandatory states.
- In 1993, accidents per 10,000 registrations was 222.1 for mandatory states and 194.02 for voluntary states.

What is the best way to assess the impact of helmet laws??

Activity: Some Simple Questions

On a separate sheet of paper you will be asked to answer three questions. We will pool the class results and discuss the lessons to be gleaned from them.

(a) What kind of study have we performed?

(b) For what type of study does this activity provide a cautionary moral?

(c) Summarize the moral of this activity in a sentence or two.

STOP!! Do not turn the page until instructed to do so.

For each question, check the one response that best captures your reaction to the scenario presented.

1. Suppose that you have decided to see a play for which the admission charge is $20 per ticket. As you prepare to enter the theater, you discover that you have lost a $20 bill. Would you still pay $20 for a ticket to see the play?

Yes ____ No ____

2. Social science researchers have conducted extensive empirical studies and concluded that the expression "absence makes the heart grow fonder " is generally true. Do you find this result surprising or not surprising?

Surprising ____ Not surprising ____

3. Suppose that the United States is preparing for the outbreak of an unusual Asian disease which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows:

- If Program A is adopted, 200 people will be saved.
- If Program B is adopted, there is a 1/3 probability that 600 people will be saved and a 2/3 probability that nobody will be saved.

Which of the two programs would you favor?

Program A ____ Program B ____

For each question, check the one response that best captures your reaction to the scenario presented.

1. Suppose that you have decided to see a play for which the admission charge is $20 per ticket. As you enter the theater, you discover that you have lost the ticket. Would you pay $20 for another ticket?

Yes ____ No ____

2. Social science researchers have conducted extensive empirical studies and concluded that the expression "out of sight, out of mind " is generally true. Do you find this result surprising or not surprising?

Surprising ____ Not surprising ____

3. Suppose that the United States is preparing for the outbreak of an unusual Asian disease which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows:

- If Program A is adopted, 400 people will die.
- If Program B is adopted, there is a 1/3 probability that nobody will die and a 2/3 probability that 600 people will die.

Which of the two programs would you favor?

Program A ____ Program B ____

Activity: Readability of Cancer Pamphlets (*Workshop Statistics, Activity 4-4*)

Researchers in Philadelphia investigated whether pamphlets containing information for cancer patients are written at a level that the cancer patients can comprehend. They applied tests to measure the reading levels of 63 cancer patients and also the readability levels of 30 cancer pamphlets (based on such factors as the lengths of sentences and number of polysyllabic words). These numbers correspond to grade levels, but patient reading levels of under grade 3 and above grade 12 are not determined exactly.

The tallies in the following table indicate the number of patients at each reading level and the number of pamphlets at each readability level.

patients' reading level	tally	pamphlets' readability level	tally
under 3	6	6	3
3	4	7	3
4	4	8	8
5	3	9	4
6	3	10	1
7	2	11	1
8	6	12	4
9	5	13	2
10	4	14	1
11	7	15	2
12	2	16	1
above 12	17	TOTAL	30
TOTAL	63		

(a) Explain why the form of the data does not allow one to calculate the mean reading skill level of a patient.

(b) Determine the median reading level of a patient and the median readability level of a pamphlet.

(c) How do these medians compare? Are they fairly close?

(d) Does the closeness of these medians indicate that the pamphlets are well-matched to the patients' reading levels?

(e) What proportion of the patients do not have the reading skill level necessary to read even the simplest pamphlet in the study? Do you want to re-think your answer to (d) in light of this question?

This activity serves as a reminder that considering the variability of a distribution is often more important than simply measuring its center. It also illustrates that inference procedures do not always address the most relevant question when analyzing a set of data and that simpler methods can sometimes be more helpful.

Activity: Draft Lottery

The following data are the draft numbers (1-366) assigned to birthdates in the 1970 draft lottery. Men born on the date assigned a draft number of 1 were the first to be drafted, followed by those born on the date assigned draft number 2, and so on.

date	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	305	86	108	32	330	249	93	111	225	359	19	129
2	159	144	29	271	298	228	350	45	161	125	34	328
3	251	297	267	83	40	301	115	261	49	244	348	157
4	215	210	275	81	276	20	279	145	232	202	266	165
5	101	214	293	269	364	28	188	54	82	24	310	56
6	224	347	139	253	155	110	327	114	6	87	76	10
7	306	91	122	147	35	85	50	168	8	234	51	12
8	199	181	213	312	321	366	13	48	184	283	97	105
9	194	338	317	219	197	335	277	106	263	342	80	43
10	325	216	323	218	65	206	284	21	71	220	282	41
11	329	150	136	14	37	134	248	324	158	237	46	39
12	221	68	300	346	133	272	15	142	242	72	66	314
13	318	152	259	124	295	69	42	307	175	138	126	163
14	238	4	354	231	178	356	331	198	1	294	127	26
15	17	89	169	273	130	180	322	102	113	171	131	320
16	121	212	166	148	55	274	120	44	207	254	107	96
17	235	189	33	260	112	73	98	154	255	288	143	304
18	140	292	332	90	278	341	190	141	246	5	146	128
19	58	25	200	336	75	104	227	311	177	241	203	240
20	280	302	239	345	183	360	187	344	63	192	185	135
21	186	363	334	62	250	60	27	291	204	243	156	70
22	337	290	265	316	326	247	153	339	160	117	9	53
23	118	57	256	252	319	109	172	116	119	201	182	162
24	59	236	258	2	31	358	23	36	195	196	230	95
25	52	179	343	351	361	137	67	286	149	176	132	84
26	92	365	170	340	357	22	303	245	18	7	309	173
27	355	205	268	74	296	64	289	352	233	264	47	78
28	77	299	223	262	308	222	88	167	257	94	281	123
29	349	285	362	191	226	353	270	61	151	229	99	16
30	164		217	208	103	209	287	333	315	38	174	3
31	211		30		313		193	11		79		100

(a) What draft number was assigned to your birthday? Is this draft number in the top third, middle third, or last third of the draft order?

(b) Look at a scatterplot of draft number vs. birthdate number (i.e., let January 1 be 1, January 31 be 31, February 1 be 32, and so on through December 31 as 366). Does the scatterplot reveal any association between draft number and birthdate?

The following table arranges in order the draft numbers for each month:

rank	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	17	4	29	2	31	20	13	11	1	5	9	3
2	52	25	30	14	35	22	15	21	6	7	19	10
3	58	57	33	32	37	28	23	36	8	24	34	12
4	59	68	108	62	40	60	27	44	18	38	46	16
5	77	86	122	74	55	64	42	45	49	72	47	26
6	92	89	136	81	65	69	50	48	63	79	51	39
7	101	91	139	83	75	73	67	54	71	87	66	41
8	118	144	166	90	103	85	88	61	82	94	76	43
9	121	150	169	124	112	104	93	102	113	117	80	53
10	140	152	170	147	130	109	98	106	119	125	97	56
11	159	179	200	148	133	110	115	111	149	138	99	70
12	164	181	213	191	155	134	120	114	151	171	107	78
13	186	189	217	208	178	137	153	116	158	176	126	84
14	194	205	223	218	183	180	172	141	160	192	127	95
15	199	210	239	219	197	206	187	142	161	196	131	96
16	211	212	256	231	226	209	188	145	175	201	132	100
17	215	214	258	252	250	222	190	154	177	202	143	105
18	221	216	259	253	276	228	193	167	184	220	146	123
19	224	236	265	260	278	247	227	168	195	229	156	128
20	235	285	267	262	295	249	248	198	204	234	174	129
21	238	290	268	269	296	272	270	245	207	237	182	135
22	251	292	275	271	298	274	277	261	225	241	185	157
23	280	297	293	273	308	301	279	286	232	243	203	162
24	305	299	300	312	313	335	284	291	233	244	230	163
25	306	302	317	316	319	341	287	307	242	254	266	165
26	318	338	323	336	321	353	289	311	246	264	281	173
27	325	347	332	340	326	356	303	324	255	283	282	240
28	329	363	334	345	330	358	322	333	257	288	309	304
29	337	365	343	346	357	360	327	339	263	294	310	314
30	349		354	351	361	366	331	344	315	342	348	320
31	355		362		364		350	352		359		328

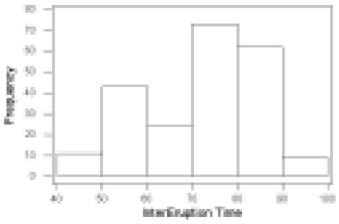
(c) Use this information to calculate the median draft number for your birth month. Is this number in the top half or bottom half of the draft order?

(d) Pool the findings of the class and record the median draft number for each month. Do you notice any tendency in these median draft numbers over time?

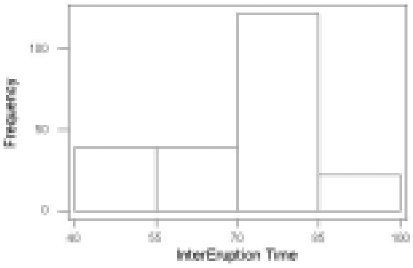
Activity: Geyser Eruptions (from *Handbook of Small Data Sets; A Casebook for a First Course in Statistics and Data Analysis; UCLA Website; Workshop Statistics, Activity 3-5) Data can be downloaded at www.rossmanchance.com/ws2/files.html*

The data set contains observations on the inter-eruption times (waiting time between the start of successive eruptions) for the Old Faithful geyser at Yellowstone National Park, WY between August 1st and August 15, 1985 (measured in minutes). There are a total of 222 observations.

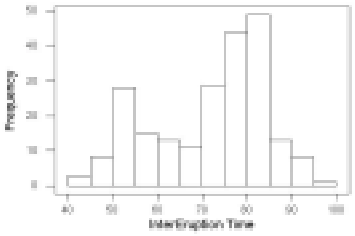
(a) Given below is a histogram for the intereruption times. Use this histogram to describe the distribution of intereruption times.



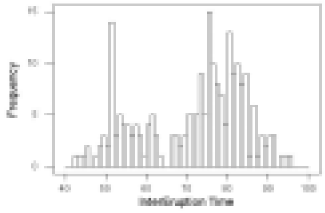
(b) Given below is a histogram that uses interval of width 15. Comment on how the appearance of the distribution has changed.



(c) The histogram below is based on intervals of width 5. Comment on how the appearance of the distribution has changed.



(d) Intervals of width 1 were used to create the histogram shown below. Explain why this picture is not the most useful display of the intereruption times.

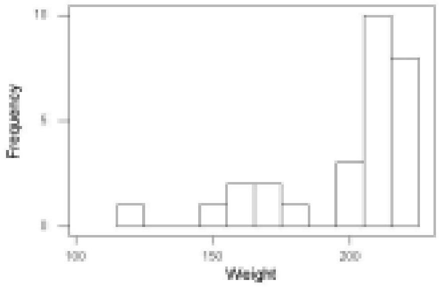


(f) How would you describe the shape of the intereruption times? If you arrive at the geyser just after an eruption, how long would you expect to wait on average for the next one?

Also works well on the computer or as a java applet, e.g. Histogram.html

Activity: Rowers' Weights (1996 data found in *JSE Datasets and Stories, Workshop Statistics)* The following are the weights of the 2000 Men's Olympic Rowing Team:

Name	Weight	Event	Name	Weight	Event
Auth	165	LW four	Mueller	220	four
Bea	205	pair	Murphy	220	pair
Cipollone	121	eight	Nuzum	210	double sculls
Collins	200	eight	Peterson	195	quad
Ferry	215	double sculls	Ruckman	160	LW four
Groom	168	LW double sculls	Schneider	160	LW four
Hall	195	quad	Simon	220	eight
Kaehler	220	eight	Smith	210	single sculls
Klepacki	212	eight	Teti	175	LW four
Koven	210	four	Tucker	153	LW double sculls
McGowan	215	quad	Volpenhein	205	eight
Miller	220	eight	Welsh	205	eight
Moser	210	four	Wetzel	205	quad
Mueller	220	four	Wherley	210	four



(a) Do any values stand out?

(b) Can you suggest an explanation for these unusual observations?

(c) The mean and median weights are

mean: 197.29 median: 207.50

If the coxswain is removed, what would happen to the mean and the median?

(d) What if the lightweight rowers are also removed?

With just the coxswain removed, the mean is 200.11 and the median is 210.00. With the lightweight rowers also removed, the mean becomes 210.57 and the median 210.00.

(e) What if the Kaehler weighs 320 pounds instead? What would happen to the mean and median?

(f) How do the mean and median values compare across these three data sets? Which measure of center (mean or median) is more *resistant* to the influence of extreme observations?

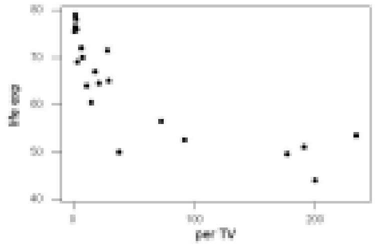
Activity 15: Televisions and Life Expectancy (from *Workshop Statistics, Activity 9-4*)

The following table provides information on life expectancies for a sample of 22 countries. It also lists the number of people per television set in each country.

Country	life exp	Per TV	country	life exp	per TV
Angola	44	200	Mexico	72	6.6
Australia	76.5	2	Morocco	64.5	21
Cambodia	49.5	177	Pakistan	56.5	73
Canada	76.5	1.7	Russia	69	3.2
China	70	8	South Africa	64	11
Egypt	60.5	15	Sri Lanka	71.5	28
France	78	2.6	Uganda	51	191
Haiti	53.5	234	United Kingdom	76	3
Iraq	67	18	United States	75.5	1.3
Japan	79	1.8	Vietnam	65	29
Madagascar	52.5	92	Yemen	50	38

(a) Which of the countries listed has the fewest people per television set? Which has the most? What are those numbers?

(b) A scatterplot of life expectancy vs. people per television set is shown. Does there appear to be an association between the two variables? Elaborate briefly.



(c) Since the association is so strongly negative, one might conclude that simply sending television sets to the countries with lower life expectancies would cause their inhabitants to live longer. Comment on this argument.

(e) If two variables have a strong relationship between them, does it follow that there must be a cause-and-effect relationship between them?

(f) In the case of life expectancy and television sets, suggest a confounding variable that is associated both with a country's life expectancy and with the prevalence of televisions in the country.

Activity: Hospital Recovery Rates (*Workshop Statistics, Activity 7-4*)

The following two-way table classifies hypothetical hospital patients according to the hospital that treated them and whether they survived or died:

	survived	died	total
hospital A	800	200	1000
hospital B	900	100	1000

(a) Calculate the proportion of hospital A's patients who survived and the proportion of hospital B's patients who survived. Which hospital saved the higher percentage of its patients?

Suppose that when we further categorize each patient according to whether they were in fair condition or poor condition prior to treatment we obtain the following two-way tables:

fair condition:

	survived	died	total
hospital A	590	10	600
hospital B	870	30	900

poor condition:

	survived	died	total
hospital A	210	190	400
hospital B	30	70	100

(b) Convince yourself that when the "fair" and "poor" condition patients are combined, the totals are indeed those given in the table above.

(c) Among those who were in fair condition, compare the recovery rates for the two hospitals. Which hospital saved the greater percentage of its patients who had been in fair condition?

(d) Among those who were in poor condition, compare the recovery rates for the two hospitals. Which hospital saved the greater percentage of its patients who had been in poor condition?

(e) Write a few sentences explaining (arguing from the data given) how it happens that hospital B has the higher recovery rate overall, yet hospital A has the higher recovery rate for each type of patient. (Hints: Do fair or poor patients tend to survive more often? Does one type of hospital tend to treat most of one type of patient? Is there any connection here?)

(f) Which hospital would you rather go to if you were ill? Explain.

(g) Create your own fictional data to illustrate Simpson's paradox in the following context. Show that it's possible for one softball player (Amy) to have a higher batting average (proportion of hits) than another (Barb) in June and in July but for Barb to have the higher batting average for the two months combined.

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

The simulation will ask you to assign trees of two varieties, a new hybrid and a standard variety, to the eight available plots in a field. You'll have the opportunity to assign the trees to the plots in three different ways: a design with no blocking, and two designs that use blocking. Your goal is to find the experimental design that has the best chance of correctly identifying which variety of tree performs best, on average.

One design is preferred over another if it provides a better chance of detecting the superior variety. 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.

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

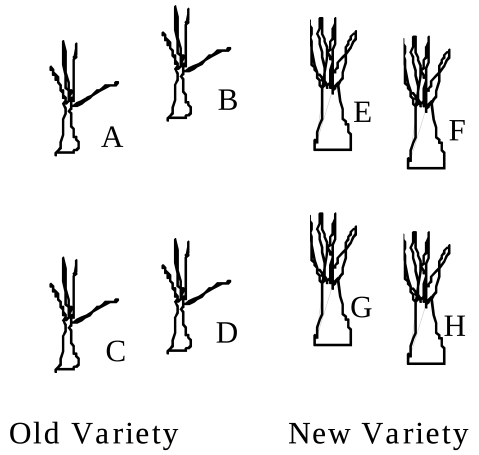
Even within a variety, there is variability in the production from tree to tree. In addition, individual plots in the field may have slight differences due to water, fertility, or other factors, adding variability in the production from plot to plot. In this simulation the variability of trees and plots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

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

One way to determine the random amount is to add or subtract a random amount, 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.

Simulation 1: A completely randomized design.

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



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 Mean (new = 53, old = 47)
+ Forest effect (-10 for near the forest, +10 for away)
+ Random variation (determined using coin and die)

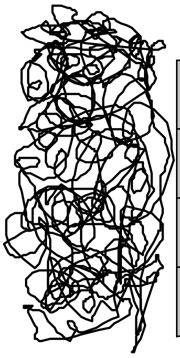
Calculate the productivity of each of your 8 trees and record this in the table above. Then calculate the average productivity of the new variety and the old variety. Record the results below, and indicate which variety was better. (As a check, these averages should be between 40 and 60.)

Old Average Productivity: __________ New Average Productivity: __________

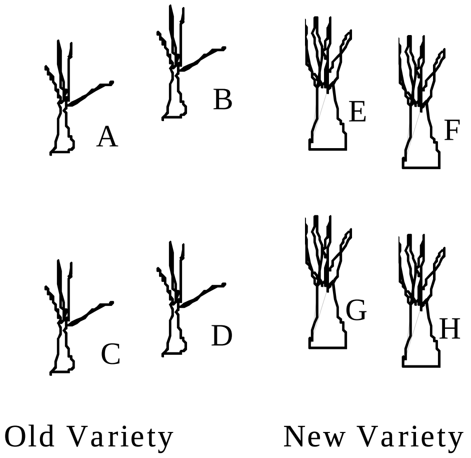
Which was better? Difference (New – Old): _________

Combine your results with those of 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



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 Mean (new = 53, old = 47)
+ Forest effect (-10 for near the forest, +10 for away)
+ Random variation (determined using coin and die)

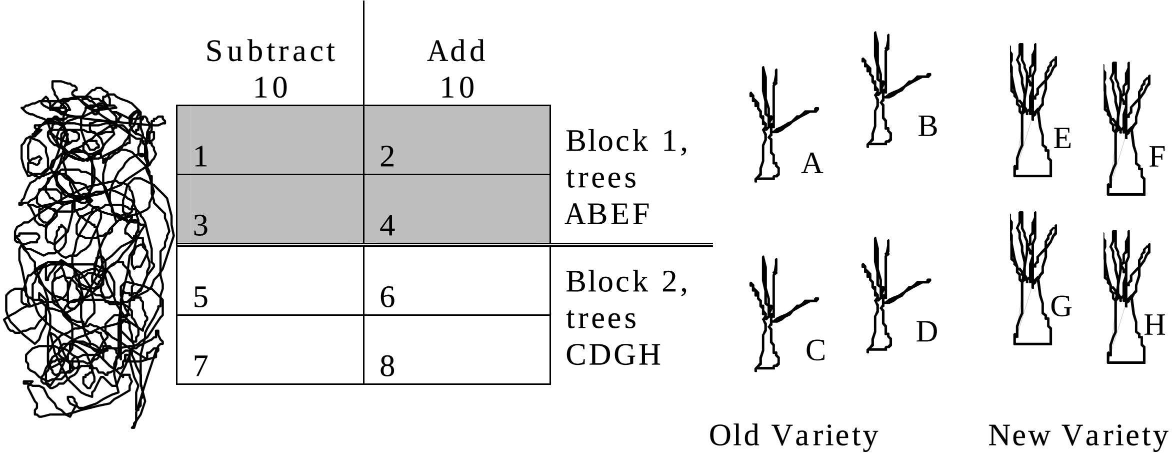
Calculate the productivity of each of your 8 trees, then calculate the average productivity of the new variety and the old variety. Record the results below, and indicate which variety was better. (As a check, these averages should be between 40 and 60.)

Old Average Productivity: __________ New Average Productivity: __________

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

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

Simulation 3: Blocking Scheme B.



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

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 Mean (new = 53, old = 47)
+ Forest effect (-10 for near the forest, +10 for away)
+ Random variation (determined using coin and die)

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

Activity: Sampling Logs

Question: Do you think that the procedure described would produce a random sample of logs from the day's shipment? Why or why not?

Activity: Sampling to Estimate Mean String Length

Does this sampling procedure produce a random sample of strings from the bag?

Do you think that the mean of a sample selected in the manner described would tend to consistently tend to under or over estimate the true mean string length? Explain.

Activity : Colors of Reese's Pieces (from *Workshop Statistics, Activity 16-2*)

Consider the population of the Reese's Pieces candies manufactured by Hershey. Suppose that you want to learn about the distribution of colors of these candies but that you can only afford to take a sample of 25 candies.

(a) Take a random sample of 25 candies and record the number and proportion of each color in your sample.

	orange	yellow	brown
number			
proportion			

(b) Is the proportion of orange candies among the 25 that you selected a parameter or a statistic?

(c) Is the proportion of orange candies manufactured by Hershey's process a parameter or a statistic? What symbol represents it?

(d) Do you know the value of the proportion of orange candies manufactured by Hershey?

(e) Do you know the value of the proportion of orange candies among the 25 that you selected?

These simple questions point out the important fact that one typically knows (or can easily calculate) the value of a sample statistic, but only in very rare cases does one know the value of a population parameter. Indeed, a primary goal of sampling is to estimate the value of the parameter based on the statistic.

(f) Do you suspect that every student in the class obtained the same proportion of orange candies in his/her sample? By what term did we refer to this phenomenon in the previous topic?

(g) Use the axis below to construct a dotplot of the sample proportions of orange candies obtained by the students in the class.

```
  +---------+---------+---------+---------+---------+-
0.00      0.20      0.40      0.60      0.80      1.00
```

(h) Did everyone obtain the same number of orange candies in their samples?

(i) If every student was to estimate the population proportion of orange candies by the proportion of orange candies in his/her sample, would everyone arrive at the same estimate?

(j) Based on what you learned about random sampling and having the benefit of seeing the sample results of the entire class, take a guess concerning the population proportion of orange candies.

(k) Again assuming that each student had access only to her/his sample, would most estimates be reasonably close to the true parameter value? Would some estimates be way off? Explain.

(l) Remembering what you learned earlier, in what way would the dotplot have looked different if each student had taken a sample of ten candies instead of 25?

(m) Remembering what you learned earlier, in what way would the dotplot have looked different if each student had taken a sample of 75 candies instead of 25?

Our class results suggest that even though sample values vary depending on which sample you happen to pick, there seems to be a pattern to this variation. We need more samples to investigate this pattern more thoroughly, however. Since it is time-consuming (and possibly fattening) to literally sample candies, we will use the computer to simulate the process.

To perform these simulations we need to suppose that we know the actual value of the parameter. Let us suppose that 45% of the population is orange.

(n) Use Minitab to simulate drawing 500 samples of 25 candies each. (Pretend that this is really 500 students, each taking 25 candies and counting the number of orange ones.) Then look at a display of the sample proportions of orange obtained.

MINITAB: File > Other Files > Run an Exec
Number of times to execute: 1
Select File: `reeses.mtb`

See also ReesesPieces.html applet.

(o) Do you notice any pattern in the way that the resulting 500 sample proportions vary? Explain.

(p) Record the mean and standard deviation of these sample proportions.

(q) Roughly speaking, are there more sample proportions close to the population proportion (which, you will recall, is .45) than there are far from it?

(r) Let us quantify the previous question. Use Minitab to count how many of the 500 sample proportions are within ±.10 of .45 (i.e., between .35 and .55). (Ask for 3 counts.) Then repeat for within ±.20 and for within ±.30. Record the results below:

	number of the 500 sample proportions	percentage of these sample proportions
within ± .10 of .45		
within ± .20 of .45		
within ± .30 of .45		

(s) Forget for the moment that you have designated that the population proportion of orange candies be .45. Suppose that each of the 500 imaginary students was to estimate the population proportion of orange candies by going a distance of .20 on either side of her/his sample proportion. What percentage of the 500 students would capture the actual population proportion (.45) within this interval?

(t) Still forgetting that you actually know the population proportion of orange candies to be .45, suppose that you were one of those 500 imaginary students. Would you have any way of knowing definitively whether your sample proportion was within .20 of the population proportion? Would you be reasonably "confident" that your sample proportion was within .20 of the population proportion? Explain.

The pattern displayed by the variation of the sample proportions from sample to sample is called the sampling distribution of the sample proportion. Even though the sample proportion of orange candies varies from sample to sample, there is a recognizable long-term pattern to that variation. Thus, while one cannot use a sample proportion to estimate a population proportion exactly, one can be reasonably confident that the population proportion is within a certain distance of the sample proportion. This "distance" depends primarily on how confident one wants to be and on the size of the sample.

(u) Use Minitab (`reeses.mtb`) to simulate drawing 500 samples of 75 candies each (so these samples are three times larger than the ones you gathered in class and simulated earlier). Use Minitab to look at a display of the sample proportions and to calculate their mean and standard deviation.

(v) How has the sampling distribution changed from when the sample size was only 25 candies?

(w) Use Minitab to count how many of these 500 sample proportions are within ±.10 of .45 (1 count). Record this number and the percentage below.

(x) How do the percentages of sample proportions falling within ±.10 of .45 compare between sample sizes of 25 and 75?

(y) In general, is a sample proportion more likely to be close to the population proportion with a larger sample size or with a smaller sample size?