

Can Mice Learn and Transfer the Concept of *Middle*?

A Statistical Exploration

Authors: Floyd Bullard, Gretchen Davis, Julie Graves, and Janet Hassan.

Produced by members of the Statistics Leadership Institute at the North Carolina School of Science and Mathematics, July 2001.

Inspired by an experiment conducted by Tommy Miller at the North Carolina School of Science and Mathematics during the 2000-2001 school year.

About this Document

Tommy Miller, a recent graduate (2001) of the North Carolina School of Science and Mathematics, found in the biological literature¹ that chimpanzees are able to learn the concept of *middle*. Miller was unable to find any research involving how or whether any non-primate species understands this concept. He designed a study to test whether a single mouse was able to learn *middle* and then transfer that understanding to a new context. Because the appropriate analysis of Miller's data involves time-series analysis which is beyond the scope of the AP statistics curriculum and because his study involved only one mouse and therefore could lead to no inference on a larger population (it was essentially a preliminary study laying the groundwork for possible more in-depth research), the experiment and its results are *not* given here. However, Miller's study inspired the problem that follows. The data given in the problem are not actual data collected from mice; rather, they were simulated to reflect various observations made in Miller's experiment and to illustrate a number of principles that *are* in the AP statistics curriculum.

What follows are

- two scientific questions,
- the description of a pair of linked studies designed to answer the questions,
- (simulated) data from the studies,
- a set of questions suggested by the studies,
- a set of responses to those questions as a statistics student might answer them,
- and a commentary for teachers explaining the analysis and also alerting them to and correcting common errors that students make.

This document was prepared by members of the Statistics Leadership Institute at the North Carolina School of Science and Mathematics during the summer of 2001. We are fortunate to have the guidance this summer of four professional statisticians: Jackie Dietz of North Carolina State University, Roxy Peck of California Polytechnic State University, Jessica Utts of the University of California at Davis, and Linda Young of the University of Nebraska at Lincoln. We believe that other teachers will appreciate that the materials here were prepared by experienced high school teachers and critiqued by experienced statisticians. This document is in the public domain and may be copied.

¹ Rohles, F. H. and J. V. Devine. 1966. Chimpanzee performance on a problem involving the concept of middle. *Animal Behavior*, **14**, 159-162, and Rohles & Devine. 1967. Further studies of the middle concept with the chimpanzee. *Animal Behavior*, **15**, 107-112.

Can Mice Learn and Transfer the Concept of *Middle*?

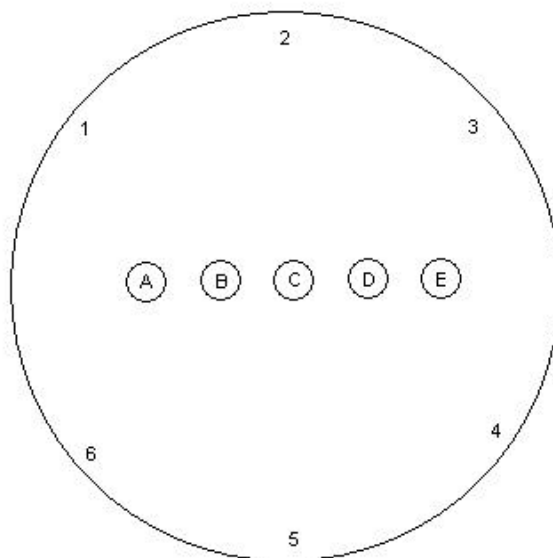
A Statistical Exploration

Introduction

Greg Martinez is a student taking a course in animal behavior research. He has learned that chimpanzees and human children have both been studied to see how they understand the concept of *middle*, but he is unable to find any such research on non-primate animals. His course includes an individual research project, so Greg decides to study whether mice can learn—and *transfer*—the concept of middle. (To *transfer* knowledge is to apply it to a new context.) After much thought and discussion with his biology instructor, he designs the following study.

After obtaining permission to use mice in his study, he will acquire 18 mice from a biological supply house. Then he will take his mice through several phases of his study:

- Phase One: “Introduction to Environment”. During this phase, Greg plans to take each mouse out of its cage and gently release it into a small wading pool (about five feet in diameter) filled 8 inches deep with water. After 30 seconds, the mouse will be lifted out of the water and returned to its cage. This will be repeated once each day for five days.
- Phase Two: “Pre-training”. Each mouse will again be released into the wading pool, but this time there will be 5 upright glass jars lined up in the pool and evenly spaced by 6 inches, with the middle jar in the exact middle of the pool. The jars are only 7 inches tall, so they do not reach above the surface of the water, but attached to each jar by a rubber band is a flexible drinking straw that extends above the surface and so is visible to the swimming mice (Previous research indicates that swimming mice will swim to objects that they see projecting from the surface of the water.) Additionally, the center jar has a glass dish placed on it just under the surface of the water. Greg intends for this dish to be invisible to a mouse swimming in the water, but easy for it to stand on. The location at which he releases the mouse into the pool will be determined independently for each



This time, rather than taking the mouse out of the water after thirty seconds, Greg plans to allow the mouse to continue swimming until it climbs onto the platform in the middle, at which time he will remove it from the pool and pet it gently before returning it to its cage (this is the mouse's *reward*). He will record three observations about the mouse's *swim*: the first drinking straw that the mouse visits (A, B, C, D, or E), how many *errors* (straws other than straw C, the one in the middle) the mouse makes before finally going to straw C, and how long (measured in seconds) it takes the mouse to get to straw C.

- Phase Three: "Training". It is during this phase that the mice will be "taught" the concept of *middle*. During this phase, Greg will release each mouse into the pool five consecutive times each day. The pool arrangement will be identical to the one in phase two, and the release location will be randomized in the same way as before. Greg will repeat this for each mouse for ten consecutive days, always removing the mouse after it finds the center platform, but he will not record any data during this period. When he removes the mouse, he will dry it and warm it in his hand, which previous research suggests is a reward for mice.
- Phase Four: "Post-training". This phase will be essentially identical to phase two. Each mouse will be released exactly once into the pool at a random location, and the same three variables will be recorded once more: first straw visited, number of errors before going to the middle straw, and time before getting to the middle straw.
- Phase Five: "New Context". This "phase" is really not so much a new *phase* as an altogether new *study* that happens to use the same mice as before. In this study, Greg will again release the mice into random locations around the edge of the pool, only this time there will be two differences. First, he will remove the outer jars and straws labeled A and E, and second, he will record only the first straw visited.

After collecting all of these data, Greg Martinez hopes to determine whether mice can learn and transfer the concept of *middle*.

Happily, Greg's study is approved by his college's biology department and by his instructor, and he begins his work. Everything goes as Greg planned², and the data that Greg collects are shown in the table on the following page.

² This exploration is inspired by a real experiment conducted by high school student Tommy Miller at the North Carolina School of Science and Mathematics in 2001, but as it is described here, it was never actually conducted. However, great effort has been made to make the study and simulated data realistic, and continuing that effort, it must be pointed out that an experiment with 18 mice is almost *certain* not to go exactly as planned. Some mice, unfortunately, will die before or during the experiment. Others may not cooperate in some way (*e.g.*, they may simply float motionless in the water rather than swimming, or they may endlessly circle the edge of the pool without ever going to any drinking straws.) These complications may make correct statistical analysis extremely difficult, which is why they are not included in Greg Martinez's study.

Mouse ID		Pre-training (5 stations ABCDE)			Post-training (5 stations ABCDE)			New context (3 stations BCD)
No.	Name	First visit	No. errors	Time (sec)	First visit	No. errors	Time (sec)	First visit
1	Scipio	B	7	21	A	4	18	C
2	Zelda	E	4	18	D	5	25	B
3	Biggs	B	5	24	D	8	19	C
4	Tiger	A	3	12	C	0	10	C
5	Jackie	B	3	12	C	0	14	C
6	Linda	B	9	39	D	6	46	C
7	Big Hombre	E	8	25	B	14	29	D
8	Beanie	E	8	17	E	2	11	D
9	Malcolm	A	8	20	C	0	16	D
10	Kipling	B	3	24	B	5	20	D
11	Roxy	A	5	24	E	1	27	B
12	Lara Croft	B	5	19	E	5	22	C
13	Dan	A	6	26	D	5	18	B
14	Fry	E	12	39	D	4	46	B
15	Wiggles	D	3	27	A	7	29	C
16	Monster	E	17	44	D	18	57	D
17	Jessica	E	11	28	B	3	25	C
18	Miller	D	2	18	C	0	14	C

Questions

Greg Martinez's overall research questions are: "Can mice learn the concept of middle?" and "Can mice transfer the concept of middle to a new context? These "big questions" should be kept in mind as you consider the following series of questions.

1. Before the training, do the mice appear predisposed to swim directly to the center of the pool? Why might this question be an important component of answering one of the "big questions"?
2. Before the training, how many errors do the mice typically make before finding the center platform? Describe the distribution of this number of errors among the mice. (*i.e.*, give an appropriate plot and describe in words the shape, center, and spread of the data.)
3. Before the training, how long does it typically take the mice to get to the center? Describe the distribution of this time among the mice.
4. Before the training, does it appear that the time that a mouse takes to get to the center and the number of errors it makes are related to one another?
5. *After* the training, do the mice appear to have a tendency to swim directly to the center of the pool?
6. After the training, how many errors do the mice typically make before finding the center platform? Discuss the distribution of this number of errors among the mice.
7. After the training, how long does it typically take the mice to get to the center? Describe the distribution of this time among the mice.
8. After the training, does it appear that the time that a mouse takes to get to the center and the number of errors it makes are related to one another?
9. Make a scatterplot of "number of errors after training" *versus* "number of errors before training". What, if anything, does this graph communicate?
10. Make a scatterplot of "time after training" *versus* "time before training". What, if anything, does this graph communicate?
11. Does it appear that the number of errors after training is lower than the number made before training?
12. Does it appear that the time the mice take to find the center is shorter after training than before training?
13. You should still be thinking about the "big questions", one of which is, "Can mice learn the concept of middle?" After answering the previous questions, what do you think is the answer to this "big question"? Justify your answer. Do you have any reservations?
14. In the third phase of the study, Greg imposed a slightly different treatment (3 straws instead of 5) upon his mice to see how they responded. If the mice had not learned anything about

middle but were essentially equivalent to “fresh”, untrained mice, how would you expect them to behave? Justify your answer.

15. What evidence is there, if any, that the mice have transferred knowledge of *middle* to the new context?

The remaining questions do not deal with the data, but with the design of Greg’s study.

16. What was the purpose of phase one (“Introduction to Environment”)? How might things have happened differently, or how might you interpret the data differently, if phase one had not been conducted?
17. What was the purpose of releasing the mice at randomized locations in the pool instead of releasing them all from the same location?
18. Physical, time, and budget constraints may prevent Greg from using different physical apparatus for different mice. What effect, if any, may using the same apparatus repeatedly have on conclusions drawn from Greg’s study? If these are negative effects, are there things Greg can do to overcome or to lessen them?
19. Is Greg’s study an experiment? If so, what are the treatments and how are they assigned to the experimental units? If not, explain why it is not an experiment and what the implications of that are for the conclusions drawn from the study.
20. Suggest ways that the study may be improved, considering both physical apparatus and design issues. Try to imagine what is reasonable for laboratory research. (*e.g.*, using 1000 mice may be possible, but would be so costly that it would not be reasonable unless the research questions being answered were extremely important.)

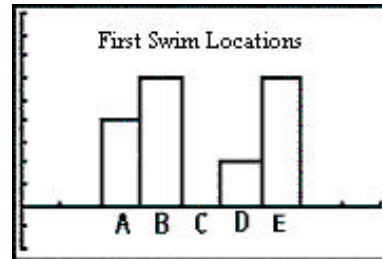
Summarize all of your observations about Greg’s study and his data into a short report that addresses the two research questions Greg set out to answer.

Sample responses and *commentary for teachers (in italics)*

- 1. Before the training, do the mice appear predisposed to swim directly to the center of the pool? Why might this question be an important component of answering one of the “big questions”?**

The following bar graph displays the locations for the first swims of the 18 mice that were involved in our study.

First Swim	Tally	Relative Frequencies
A	////	4/18
B	/////	6/18
C		0/18
D	//	2/18
E	/////	6/18



Since these mice never swam directly to the center location on their initial swims, we have no evidence that mice are predisposed to swim to the center. In fact, the data suggest the opposite. These mice seemed to avoid the center.

When we first heard about Tommy’s study, we did not realize that researchers often used swimming as an activity to measure the reactions of mice. In a brief Internet search, we found descriptions of projects that involved swimming tests for treated and control mice that included the study of Alzheimer’s vaccine in Scotland, schizophrenia treatments in China, and the use of performance enhancing drugs in the United States. Since mice are not natural swimmers and have a tendency to remain close to the walls, the platform is a refuge from this stressful activity. The ability of mice to improve their short-term memory, grow new nerve cells, and improve endurance can be measured during swimming activities.

Why might it be important to know if mice are predisposed to swim to the center?

If mice were predisposed to go to the center then Greg’s observation about post-training behavior would be hard to interpret. When mice went to the center, he wouldn’t know if it was because of training or because of a predisposition. The absence of such a predisposition makes the rest of the experimental design useful for detecting learning.

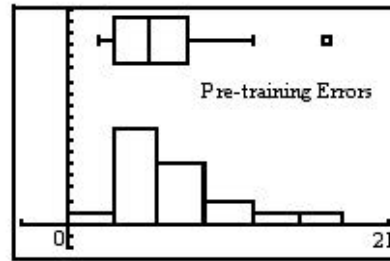
Some students may want to analyze these data using a Proportion Test. If there is no preference for the middle the proportion of visits to the middle should be about 1/5 on average. However, the conditions for approximating the normal distribution have not been met. ($18 * 1/5 = 3.6$ which violates $np > 10$ and $n(1-p) > 10$.) Using a Chi-squared Test would also be inappropriate since expected cell counts would be less than 5. In addition, Chi-squared would not address the hypothesis of mice swimming to the center, but the hypothesis that the mice were equally likely to swim to each of the five locations.

- 2. Before the training, how many errors do the mice typically make before finding the center platform? Describe the distribution of this number of errors among the mice. (i.e., give an appropriate plot and describe in words the shape, center, and spread of the data.)**

The following plots display the number of errors of the 18 mice during their first swim to the platform.

0	233334
0	555678889
1	12
1	7

1/7 means 17 errors



Summary Statistics

mean	sd	min	Q1	Median	Q3	max
6.61	3.88	2	3	5.5	8	17

The plots show us that the distribution for pre-training error data is skewed right with an outlier at 17 errors. On average, mice make about six before they reach the center. Half the mice made between three and eight errors.

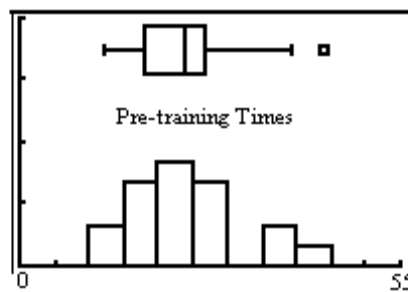
Since we have only 18 observations this is probably not enough to tell the shape of the population distribution. Since we are counting the number of errors, which are bounded on the left because we cannot have fewer than zero errors and infinite on the right, we would expect a right-skewed distribution. Since our distribution is skewed, reporting the interquartile range, rather than the standard deviation, is preferred

3. Before the training, how long does it typically take the mice to get to the center? Describe the distribution of this time among the mice.

Here are plots for the pre-training times for the 18 mice in our study.

1	22
1	7889
2	01444
2	5678
3	
3	99
4	4

2/5 means 25 seconds



Summary Statistics

mean	sd	min	Q1	Median	Q3	max
24.28	8.87	12	18	24	27	44

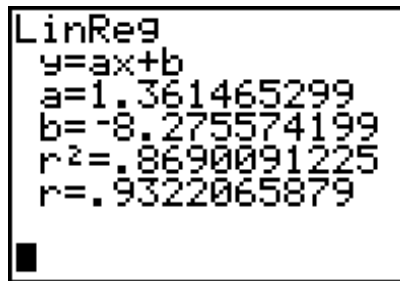
The distribution for pre-training times is roughly symmetric with a gap in the low 30 seconds and an outlier for #16 Monster Mouse at 44 seconds. (Monster was also the mouse that made 17 errors.) On average, the mice took about 24 seconds to reach the center. Half the

mice took between 18 and 27 seconds to reach the platform on their first swim. All but three of the mice were successful in less than 30 seconds.

Outliers that are of particular interest may be mistakes in measurement, elements from a different population, or just part of the natural variability of data. Perhaps we recorded Monster's errors and times incorrectly, or Monster may be a different kind of mouse, or Monster may be more careless and slower than other mice in our study.

4. Before the training, does it appear that the time that a mouse takes to get to the center and the number of errors it makes are related to one another?

The scatter plot of times versus errors for each mouse is shown below.

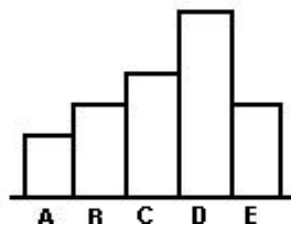


From our scatter plot and correlation coefficient calculation, it appears that there is a strong positive relationship between the number of errors and the time it took to reach the center platform.

5. After the training, do the mice appear to have a tendency to swim directly to the center of the pool?

The following is a frequency table and a histogram of the first visits of the mice after training

Location	Tally	Relative Frequencies
A	//	2 / 18
B	///	3 / 18
C	////	4 / 18
D	/////	6 / 18
E	///	3 / 18



After training it appears that the behavior of the mice changed. These mice swam to the middle more than they did in the pre-training sessions.

*Some students may want to analyze these data using a Proportion Test but the conditions for approximating the normal distribution have not been met. ($18 * 1/5 = 3.6$ which violates $np > 10$ and $n(1-p) > 10$.) Using a Chi-squared Test would also be inappropriate since the expected cell counts would be less than 5. In addition, Chi-squared would not address the*

hypothesis of mice swimming to the center, but the hypothesis that the mice were equally likely to swim to each of the five locations.

6. After the training, how many errors do the mice typically make before finding the center platform? Discuss the distribution of this number of errors among the mice.

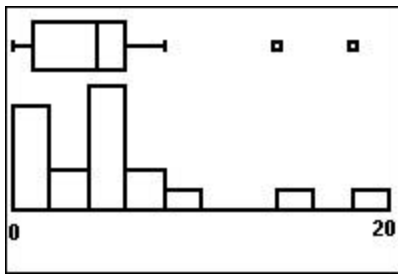
The stemplot, histogram and summary statistics of the after training error data are as follows.

```

0 000012344
0 5555678
1 4
1 8
  
```

Summary Statistics

mean	sd	min	Q1	Median	Q2	max
4.833	4.82	0	1	4.5	6	18



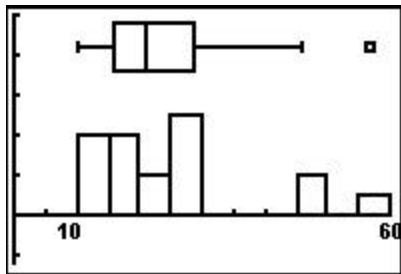
Post-training Errors

The distribution of post training error data appears slightly right-skewed due to the two outliers. The mean is 4.83 and the median 4.5 which are both less than the pre-training values. The interquartile range is 5 with outliers at 14 and 18. All but two of the mice made 8 or fewer errors.

Since we have only 18 observations this is probably not enough to tell the shape of the population distribution. Since we are counting the number of errors, which are bounded on the left and infinite on the right, we would expect a right-skewed distribution.

7. After the training, how long does it typically take the mice to get to the center? Describe the distribution of this time among the mice.

The following boxplot, histogram and summary statistics describe the post training time data.



Post-training Times

Summary Statistics

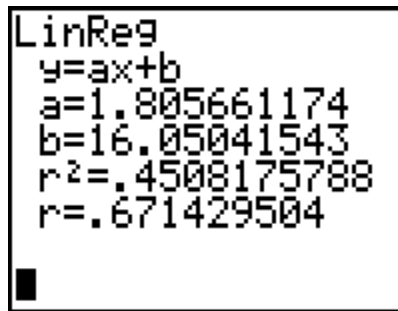
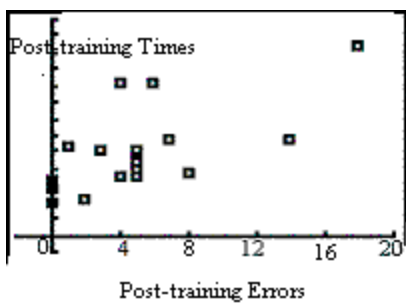
mean	sd	min	Q1	Median	Q2	max
24.77	12.95	10	16	21	29	57

The distribution of post training time data is also skewed right with “Monster” mouse floundering for 57 seconds before reaching the platform. The mean time was 24.77 seconds and median lower at 21 seconds. The IQR was 13 seconds. All but three of the mice took less than 30 seconds to reach the center.

Since we have only 18 observations this is probably not enough to tell the shape of the population distribution. Since we are counting the number of errors, which are bounded on the left and infinite on the right, we would expect a right-skewed distribution.

8. After the training, does it appear that the time that a mouse takes to get to the center and the number of errors it makes are related to one another?

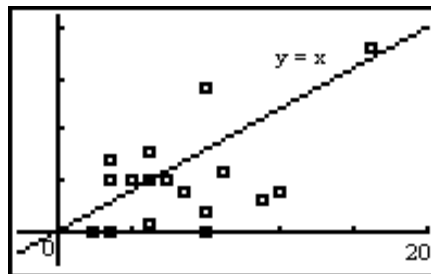
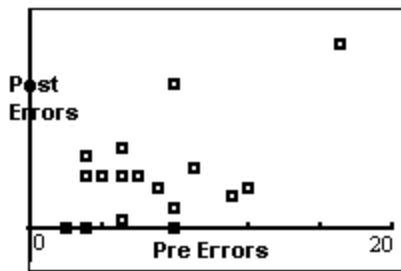
The scatter plot of times versus errors and regression calculations are shown below.



There appears to be a positive relationship between number of errors and time after training.

The relationship is highly dependent on one mouse, Monster, who may be influential.

9. Make a scatterplot of “number of errors after training” versus “number of errors before training”. What, if anything, does this graph communicate?

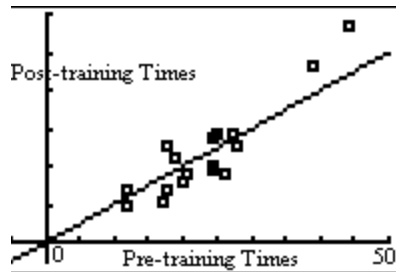


There is a weak positive relationship between errors before and after training. Mice who made many errors before also made many errors after.

Of particular interest is the upper right point. This mouse made a lot of errors both before and after training, which influences the correlation between the errors. Without this point the correlation would still be positive, but it would be weaker. When we add the line $y=x$ (which is not the regression line) to the graph, the points that lie below the line will show

a decrease in the number of errors. There are 11 mice below the line representing the mice that had fewer errors after training.

10. Make a scatterplot of “time after training” versus “time before training”. What, if anything, does this graph communicate?



Our scatter plot indicates a strong positive relationship between the Pre-training and Post-training Times. Mice who took a long time to reach the center on their initial swims also took a long time to on their Post-training swims.

Of particular interest are the two upper right points. These mice took a very long time both before and after training, which influences the correlation between points. Without these points, the correlation would still be positive, but it would be weaker. When we add the line $y = x$, we see that only half of the mice improved their times. After training, the two slowest mice spent more time floundering in the pool than they did before training.

11. Does it appear that the number of errors after training is lower than the number made before training?

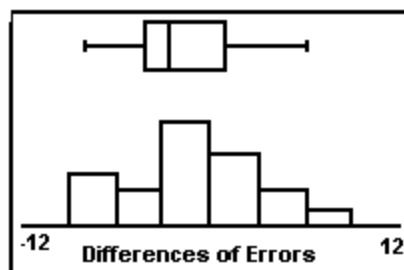
A histogram, stemplot and summary statistics for the distribution of errors (After-Before) data follows

-0	6888
-0	1233334
0	011234
0	6

Summary Statistics

mean	sd	min	Q1	Median	Q3	max
-1.78	4.17	-8	-4	-2.5	1	6

0/6 means 6 more errors



The distribution of the differences of the errors is nearly symmetric and there are no outliers. Since the t-test is robust, we will proceed with a hypothesis test.

We will define m_{diff} = the mean of the differences in post – pre errors (After – Before)

$H_o: m_{diff} = 0$ (On average mice would not improve.)

$H_a: m_{diff} < 0$ (On average mice would improve.)

```
T-Test
μ<0
t=-1.810278551
P=.043982331
x̄=-1.777777778
Sx=4.166470584
n=18
```

Since $p=.044$ we can reject the null hypothesis at a .05 level. There is evidence that the number of errors has decreased. On average the mice would get to the platform with fewer errors.

When using a paired design it is easier to detect a difference in the means because we have reduced variability in the estimated mean difference. A student who fails to understand that this is a paired design may try an independent two-sample t-test. The result of doing this would give $t = 1.22$ with a p-value of .12 which would not detect a difference. Pairing is done in order to reduce variability. In this case, the design of the study dictates that the data should be studied as paired differences. However, one can also see here why pairing should be done. Much of the variability in the number of errors after training is due to the variability among mice, explainable by the number of errors made before training. If you don't take paired differences, then any difference between the mean numbers of errors before and after training is "drowned out" by the variability among mice.

12. Does it appear that the time the mice take to find the center is shorter after training than before training?

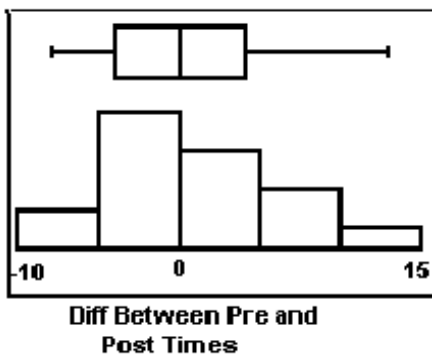
A histogram, stemplot and summary statistics for the distribution of errors data (After – Before) follows.

-0	568
-0	233444
0	22334
0	777
1	3

Summary Statistics

mean	sd	min	Q1	Median	Q3	max
.5	5.68	-8	-4	0	4	13

1/3 means 13 seconds



```
T-Test
μ<0
t=.3734585457
P=.643288244
x̄=.5
Sx=5.680202979
n=18
```

The distribution for the differences for the times has no outliers. We will proceed with a hypothesis test.

H_0 : The average difference in times is be zero (mice would not reach the platform faster).

$$m_{diff} = 0$$

H_a : The average difference in times is greater than zero (mice would reach the platform faster).

$$m_{diff} < 0$$

Since $p=.643$ we cannot reject the null hypothesis at any reasonable level. There is not enough evidence to conclude that times would decrease. On average mice are not getting to the platform any faster.

A student who fails to understand that this is a paired design may try an independent two-sample t-test. The result of doing this would give $t= .135$ with a p-value of .553.

13. You should still be thinking about the “big questions”, one of which is, “Can mice learn the concept of middle?” After answering the previous questions, what do you think is the answer to this “big question”? Justify your answer. Do you have any reservations?

The analysis done in question 11 shows that the number of mice errors decreased after training; on average the decrease was 1.78 errors. This decrease was shown to be statistically significant, and I think this shows that the mice learned to identify the middle straw. On the other hand, the analysis done in question 12 shows that the time it took the mice to find the middle *increased* by 0.5 seconds on average, which not only failed to indicate a statistically significant decrease in mean swim time, but in fact was, for the observed mice, an increase in mean swim time. Based on this criterion, one might think that mice did not learn middle, since they were not significantly faster at finding middle after training than they were before training. The two variables used to measure how well mice can identify middle, number of errors and time, do not “speak with one voice”, but rather lead to conclusions about learning that are somewhat contradictory.

This question calls on students to interpret and compare the results of two statistical tests. Based on number of errors, the data indicate that mice can learn to find the middle straw after training. Based on time, the data indicate that mice may not be able to learn to find the middle straw after training. Although students do not have the advantage of having observed the mice swimming, they can make some educated guesses about why the mice exhibited a significant decrease in errors but not in time. Perhaps near the end of the study the mice didn't dislike the water as much as they had at the beginning of the study. If mice were not highly motivated to get out of the water, they would not necessarily find the middle quickly. Another way to think about the two variables is in terms of errors per unit time. If the number of errors went down but the time did not go down, then the mice made fewer errors per unit time after training. It is possible that the mice spent enough time in the water during training that they were no longer “panicked” and thus did not make lots of errors in quick successions. It is important that students recognize that these suggestions are simply informed speculation and do not constitute firm knowledge about mouse behavior. Students' ability to interpret the data gathered during Greg's study remains limited by the fact that they did not observe the behavior of the mice as the data were collected.

- 14. In the “new context” phase of the study, Greg imposed a slightly different treatment (3 straws instead of 5) upon his mice to see how they responded. If the mice had not learned anything about middle but were essentially equivalent to “fresh”, untrained mice, how would you expect them to behave? Justify your answer.**

During phase 2 of the study, untrained mice presented with 5 straws never went to the middle first. This clearly shows that mice are not predisposed to go to the middle. I would expect that untrained mice would rarely go the middle of 3 straws. Even if they had learned nothing during the earlier phases of the study, we expect that mice would go to the middle straw first no more than 1/3 of the time.

- 15. What evidence is there, if any, that the mice have transferred knowledge of *middle* to the new context?**

Before training, mice choosing among 5 straws never chose the middle first. After training, mice choosing among 3 straws chose the middle first 9 times out of 18. If mice had not transferred knowledge of middle to this new context, I would expect them to choose the middle straw first at most one-third of the time. The test of $H_0 : p = \frac{1}{3}$ versus

$H_a : p > \frac{1}{3}$ produces a *p-value* of 0.0668. While this *p-value* does not cause me to reject the null hypothesis at the $\alpha = 0.05$ significance level, it does give some evidence that mice can learn. I would like to repeat the experiment with a larger sample size, since it may be that the effect as I measured it is real, but too small to detect with significance using only 18 mice.

This student used a proportions test of significance, which is appropriate in this case. However, he or she implicitly assumed that the normal approximation to the distribution of \hat{p} was a good one without checking. (Or, more likely, he or she used a TI-83 calculator to perform the test and forgot that the calculator uses a normal approximation that may or may not be reasonable.) In fact, $np = 18 \times \frac{1}{3} = 6$, which is less than 10, indicating that the normal approximation is not very reasonable and should not be used.

*There are two different things a student might do at this point if he or she wanted to compute the *P-value* of the test. One is a binomial test. Under H_0 we have $X \sim \text{Bin}(18, \frac{1}{3})$, and the observed value of X is 9. So the *P-value* would be $\Pr(X \geq 9) \approx 0.11$. Alternately, the student might simulate the scenario using a random digits table or a deck of cards or dice, etc. He or she will not be able to pin down the *P-value* exactly, but with a lot of runs of a simulation, he or she should be able to see that it is greater than 0.05.*

However, be sure that students do not attach too much importance to the number 0.05 (or 0.10) as a significance level. The behavior of the mice clearly changed after training, and 9 selections of middle among 18 mice is very different from 0 selections of middle among 18 mice. While 9 out of 18 is not formally statistically significant when testing $p = \frac{1}{3}$ vs. $p > \frac{1}{3}$, it certainly arouses suspicions that some learning is going on and should make Greg want to gather more data.

16. What was the purpose of phase one (“Introduction to Environment”)? How might things have happened differently, or how might you interpret the data differently, if phase one had not been conducted?

The purpose of phase one is to get the mouse used to swimming around in water. It may be that the first time a mouse is exposed to the wading pool environment, it behaves unusually (*e.g.* by swimming around the edge for a long time, *etc.*). If this phase were ignored, then we would not know whether any differences we observed in mouse behavior between the pre-training phase and the post-training phase were due to the training or to getting over the shock of the new environment.

The response above is essentially correct. A further question students might be encouraged to consider relating to this is “how does an experimenter know how long to continue phase one?” This is not at all an easy question. A scientist’s experience with mice will help him or her “get a feel” for when mouse behavior has “settled down” to a regular pattern, indicating that the mouse has “gotten used to” the new environment. Isn’t this somewhat unscientific? Yes. And subjective? Definitely. The implicit assumption being made when conclusions about learning are drawn from the study is that the time allotted to phase one was sufficient to eliminate the effect of “environment shock”. Some may disagree with this assumption, believing perhaps that the increased tendency of Greg’s mice to swim towards the center is not due to learning, but to the tendency of all mice to swim to the center of a pool of water once they get used to being in water regularly—and that his phase one was not long enough to eliminate this effect. As it is presented in these pages, Greg’s study provides no way to refute this claim. The response to question 19 below addresses this problem further.

17. What was the purpose of releasing the mice at randomized locations in the pool instead of releasing them all from the same location?

If the mice were all released at the same point in the pool, then there could be a lurking variable that was confounded with “middle”. For example, suppose that all the mice were released at location 2. An increased tendency to swim to the middle might not really be due to the mice learning “middle”, but rather to an increased tendency of the mice to swim to the nearest object they see.

This response is correct, but many students will find this a difficult question because they will confuse this randomization of release locations with the randomization of treatments, which it is not. (This idea is discussed further in question 19.) Question 17 could lead to a discussion of confounding variables. If, for example, the mouse were always released at location 2, then middle would be perfectly confounded with nearest. In fact, at any fixed release location, there may conceivably be a visual clue on the edge of the pool or in the room that the mice are keying on instead of middle. This visual clue (say, a light), would be confounded with middle in the experiment if the mice were all released from the same location. To reduce that possibility, release location is randomized.

There could be another experiment in which release location was the treatment of interest, in which case this randomization would indeed be the randomization of mice to treatments (although a die roll would not guarantee an equal number of mice per treatment, and therefore would not maximize the power of statistical tests). In such an experiment, release location would be randomized and recorded, and a relationship would be sought

between release location and some response variable. Greg Martinez was not interested in this option.

- 18. Physical, time, and budget constraints may prevent Greg from using different physical apparatus for different mice. What effect, if any, may using the same apparatus repeatedly have on conclusions drawn from Greg's study? If these are negative effects, are there things Greg can do to overcome or to lessen them?**

The experimenter should probably replace the drinking straws after each mouse's swim during all stages of the experiment. This wouldn't cost much, and it prevents the possibility that the mice leave scents behind on the straws that either attract or repel other mice.

This student response addresses what is probably the most important aspect of the physical apparatus that needs to be controlled. The mice could be leaving scents behind in the water, but it is difficult to change the water between every mouse all the time, and at any rate, it doesn't seem likely that scents in the water would be associated with one particular station. Similarly, an issue relating to population of inference is that if we only use one wading pool, then we cannot infer what mouse behavior would be like in any other situation, not even another wading pool. Again, however, space limitations probably prevent private pools for every mouse, and it is perhaps not an unreasonable assumption that all wading pools are essentially the same from a mouse's point of view. But the drinking straws are another matter. Suppose that in the pre-training phase, no mice left unusual scents behind, but at some point during the training phase or even early in the post-training phase, a single mouse left behind a scent on the middle straw that was detectable to other mice even from a distance and that attracted them. Clearly such a situation is plausible and would cause the mice to appear to have learned middle when in fact they were just following their noses.

Replacing the drinking straws would greatly reduce this problem and should be done in any case, but see question 19 for a discussion of experimental design that addresses this problem more thoroughly.

- 19. Is Greg's study an experiment? If so, what are the treatments and how are they assigned to the experimental units? If not, explain why it is not an experiment and what the implications of that are for the conclusions drawn from the study.**

Greg's study is *not* an experiment. All of the mice received the same treatment, which consisted of being rewarded for finding the middle. A true experiment must make a comparison among groups, either by having a control group, or else by having groups that are exposed to different treatments. In Greg's study *all* of the mice were rewarded for finding the middle, so Greg cannot conclude that this reward resulted in changed behavior. Without a control group, Greg cannot be sure that the effects he observed in the mice are due to the treatment. An improvement in the mice's ability to find the middle could be due to age/maturity, to having gotten used to the water, to seeing the platform, or to some scent or marker that mice left on the straws.

Even though Greg's study has many good design features, it is not a true experiment. Greg cannot reach any conclusions about causation without having a control group for comparison. This is a good opportunity to remind students that the randomization of starting locations discussed in question 17 does not make this a randomized experiment. The starting location is not the treatment of interest in this study. The treatment whose effect Greg cares about is the reward for reaching the middle straw.

A controlled experiment with the same basic structure of Greg Martinez's study would be to begin with, say, 36 mice, and to allocate them randomly to two groups of 18. One group, called the treatment group, would go through exactly the process described in Greg's study—all five phases. The other group, called the control group, would also go through the entire process, only with one exception: during the training phase, there would be no platform, and the mice would not be rewarded for going to the middle. As an alternative, they could, say, be removed from the water after a fixed time of 30 seconds, and given the same reward at that time that the mice in the treatment group receive for going to the middle straw. At the end of the experiment, the improvements in numbers of errors (pre-training minus post-training) would be the data of interest for the mice, and the appropriate comparison would be between the means of two independent samples: is the mean decrease in errors significantly greater for the treatment group than for the control group? A similar question could also be asked for the mice's swim times.

One further improvement could be made on this controlled experiment. It seems plausible that the extent to which a mouse improves may be correlated with how many errors it made originally. If so, then conducting a paired differences experiment could reduce the variability. Note that this means pairing mice, not pairing pre-training and post-training for individual mice. In this new experimental design, all 36 mice would go through the pre-training phase, and the number of errors they made would be recorded. Only then would they be assigned to treatment groups. They would be ordered from most errors to fewest errors, and then paired, with the two mice having made the most errors constituting a pair, etc., down to the two mice with the fewest errors constituting a pair. Within each pair, the two mice would be randomly assigned to the treatment group and the control group (one mouse to each group), say, by a coin flip. Then the experiment would be conducted as described in the previous paragraph, with the only difference between the two groups being platform/no-platform, and reward-for-middle/reward-after-30-seconds. Finally, when analyzing the data, the scientist should look at the 18 paired differences (decrease in errors made by control mouse minus decrease in errors made by treatment mouse) to see whether the mean difference was significantly greater than zero.

20. Suggest ways that the study may be improved, considering both physical apparatus and design issues. Try to imagine what is reasonable for laboratory research. (e.g., using 1000 mice may be possible, but would be so costly that it would not be reasonable unless the research questions being answered were extremely important.)

The study should be a controlled experiment with random allocation of mice to treatment groups (reward-for-middle and no-reward-for-middle). Pairing mice by initial middle-finding-ability would probably reduce variability in the response variables of error-reduction and time-decrease. Additionally, drinking straws should be replaced after each mouse. Using more mice would make the experiment better able to detect significant differences in behavior. Since the experiment raises the question of whether mice especially dislike being in the water, it might be worthwhile thinking of another reward to give mice for finding the middle (e.g., food, etc.). Although it is not mentioned in the study design, the experimenter should randomize the order in which the mice are put in the water each day, and he or she should either be consistent with the times of day that the swims occur, or else the times should be randomized.

This response is pretty comprehensive in the improvements it suggests. Interestingly, the study that inspired this one began as one in which mice were presented with five (dry)

tunnels, only the middle of which led to food. The experimenter found that the mouse either didn't like the food he was providing or else was sufficiently satiated in a very short time that food ceased to be a reward. He did research on the Internet and found that removal-from-water had been used successfully before as a reward for lab mice. However, his data, like those presented here, indicated that although the number of errors his mouse made tended to decrease over a period of several days, the time the mouse spent swimming did not.

Students' first reaction when asked to suggest improvements to a study tends to be "more data". They should understand clearly that no amount of data can overcome the inherent problem with lack of control in this study as it is designed. Furthermore, statistics students should be reminded that in practice, collecting data comes with a cost (in time, money, resources, etc.), and that at some point, the increase in power of statistical tests is not worth the increased cost of conducting the experiment.

The suggestion by the student above to randomize the order in which the mice swim each day is an excellent one. As with randomizing release locations, this would not be considered randomization of treatments (which is altogether impossible without more than one treatment group). But it does prevent the mice from consistently receiving different treatment. (e.g., early in the session the experimenter's hands are cooler and gentler than later, and this makes mice #1, #2, and #3 more relaxed, etc.)

Students, of course, may come up with other improvements to the study, but the most important one is having some kind of control and random allocation of mice to at least two treatment groups.

Summarize all of your observations about Greg's study and his data into a short report that addresses the two research questions Greg set out to answer.

No summary is given here. There are different ways this exploration may be shared with students, appropriate for students at different levels or at different points in the course (e.g., before or after formal inference). Some possibilities are:

- Give all of these questions to the students early in the year and have them do exploratory data analysis without formal inference.
- Give all of these questions to the students later in the year and have them do formal inference procedures where appropriate, checking requirements, etc.
- Select only some of the questions to give to students, focusing on particular topics of interest at appropriate points in the year.
- To develop student confidence and creative thinking, give only the study design and the data to students without any of the guiding questions and ask them how well the experiment addresses the research questions and what answers the data provide to those questions. Have the students provide a written report of all of their observations and conclusions.
- Have the students work on all the questions in a guided way but summarize what they've learned in a report in the end.
- Use this study as a model for designing a similar study or experiment with similar related questions.

However you use this document, we hope you will find that it helps you to grow as a statistics teacher as it helps your students to grow. We, the authors, all learned a lot while writing it by discussing and debating ideas with one another and with the professional statisticians that were present to guide us and critique our work.

This document is available in pdf format from the web by following the links to the math department and teachers' resources at www.ncssm.edu. It is in the public domain and may be copied.

Floyd Bullard, NCSSM
Gretchen Davis, Santa Monica, CA
Julie Graves, NCSSM
Janet Hassan, Science Academy of South Texas

The Statistics Leadership Institute, July 2001
The North Carolina School of Science and Mathematics